An Outline of the History of Methods of Discovering Causality

Discovery is what science is all about

--Norwood Russell Hanson

Apology

This historical sketch was prompted by the discovery that my (and others') students working on algorithms for discovering causal relations knew nothing—and I mean *nothing*—of the history of methodology, and also discovering that there was no source to which I could direct them. Knowing something of that history is unnecessary for doing good mathematics, but it can help prevent reinventing old wheels merely shined up with new formalism, and it can help in lighting up blind spots and in undermining current dogmas. Perhaps the most important things we can learn is that the scientists who created the basic sciences did so without the aid of statistics (even when it was available) and by defying our modern shibboleths against confirming hypotheses by the very data from which they were brought to mind, that is, by failing to recognize any distinction between "confirmatory" and "exploratory" research. And perhaps sentimentally, I think we owe it to those long dead who created the modern world, step by groping step, to try to remember a little of what they did.

This is not a social history. Except as occasional asides, it is not concerned with religion, economics, or politics, just with ideas about what causation consists in and how to discover it-- mostly the latter. Among the most important works are

illustrations of method in concrete cases rather than essays on methodology itself. Accordingly, the history of causation is substantially a history of scientific episodes where methodological problems were vivid in retrospect or where methodological advances were made. It is best to remember that most of this history is about a time when work was done by candle light or oil lamp, often in small private laboratories or hospital offices, scientific news was often communicated only by letter, and by our standards life was pretty hard.

This is not a scholarly history. Many of my sources are secondary and even when--as for example Newton's *Principia* or the *Compleate Works* of Robert Boyle--I have read the originals, I have long since forgotten them.

Preface

Anything that knows how to change the world in any way to get what it wants has some understanding of causation. Even casual observers should know that animals make causal inferences. I once watched a red angus bull work on a complex wooden lock to open a gate to free his corralled comrades. Angus the Cow Genius knew that cattle could not jump the corral fence. He had watched us do something with the lock followed by the gate swinging open. He worked for hours, pushing pieces of wood in one direction or another, one sequence or another, until the lock came free, the gate was pushed open, and the cattle escaped. Angus the Cow Genius had a goal, and having watched humans work the lock, perhaps a little prior knowledge. With that and trial and error he discovered a causal relation while effecting it.

Almost certainly our forgotten ancestors implemented causal understanding in similar if more diverse practical ways. In impractical ways they invented unseen causes, gods, spirits and jinn, with prayers and sacrifices and dances to manipulate them. Causation was everywhere. It is no surprise then that when civilization could afford the leisure of philosophy, causality was among the first things that philosophers tried to understand, to systematize, and to seek methods for discovering. Over the ages, those reflections on how to produce or prevent or explain things involved a variety of ideas about inductive inference more generally, logic, and much later, probability.

Excellent histories of probability and statistics scarcely touch on causal inference. (There is one exception, William Wallace's *Causality and Scientific Explanation*, in two volumes, the first of which surveys medieval science and the second of which I have been unable to obtain.) The essay that follows aims to fill the gap temporarily until something more sturdy and scholarly comes along. The focus is on methods proposed to discover causal relations, less on metaphysical accounts of what causation is, although metaphysical considerations are inevitably involved. From incapacity rather than disdain, I do not consider Arabian or Indian or Chinese or Persian ideas; I do not read the languages and, with the exception of Islamic optics, the secondary sources focus pretty exclusively on metaphysical and religious themes.

Causality in Ancient Science

Interventions made in the belief that what happens in the future can be influenced by manipulations of some kind--that the future can be controlled by what is done now--are as ancient as prayer and the sacrificial torture and slaughter of humans. Belief in causality is as old as humanity and as fundamental. What has changed is metaphysical stories and proposed methods for discovering what causes what.

The evidence that one thing causes another may come from rough experimentation—make interventions that alter conditions and note corresponding changes in features that are not directly manipulated--or it may come from regularities that are not in anyone's power to manipulate, or both. Ancient scientists and physicians practiced both kinds of inference, as well as postulating causal relations and mechanisms on purely theoretical grounds. There are two naïve experimental methods: trial and error to produce an outcome, and contrast.

Nebuchadnezzar, King of the Babylonians, is said to have tried vegetarian and omnivorous diets and found the former to be healthier. Early Christian propagandists claimed the Egyptian heathens experimented with vivisection on humans, but I don't trust their accounts (of much of anything). Roman physicians varied drug dosages to see effects.

Judged by duration of ideas, Aristotle, from the 5th century B.C., is possibly the most influential methodologist that ever lived. Aristotle is thought to have been Plato's student in Athens, but one could argue that Aristotle did not exist. The fragments of his work that survive, so impressive given that to our knowledge there was nothing like them previously, and so enduring (his formal logic is still taught to many college freshmen, especially in Catholic schools) that it is difficult to believe they were the product of one mind. He wrote on logic, physics, biology, astronomy, ethics, politics, scientific method, poetry and more. After his work was made available in translation for European Latin readers in the 12th and 13th centuries, five centuries passed before his conception of scientific method began to be widely repudiated. Even so, on reading Aristotle anyone educated in almost any part of modern science will find themselves in a very different and somewhat obscure milieu of mind, although not nearly so much as in reading some 19th century and 20th century philosophers, e.g., Hegel and Heidegger.

The abstract structure of the science Aristotle conceives is a hierarchy of kinds: humans are mammals, mammals are animals, animal are living beings, living beings are material things---and so on. Transitions in the hierarchy are constituted by a distinct "form" or essence differentiating each kind in the hierarchy from all other kinds. Forms are piled on. Humans, for example, are essentially *rational animals*. (One wishes!) Science has two kinds of tasks; one is classification, distinguishing kinds of things and placing them appropriately in the hierarchy of kinds. The other is establishing the causes of specific kinds and the phenomena that exhibit them, which is in part the same as identifying their essential properties. The essential properties, the forms, of kinds of entities and their parts

function to serve a purpose in the growth and mature action of things of that kind.

Roughly, the *form* of a kind distinguishes that kind from other kinds and gives the members of the kind the potential to reach a mature state and to respond to circumstances. Potentials are multifarious, but for each kind there is an essential potential: a chestnut could end up as food for squirrel, but its essence is to become an oak tree. (One framework in modern statistics—"potential outcomes" is a variation on Aristotle in which each entity has determinate "potentials"—a response it would manifest according to the individual cause it might be subjected to.) The forms have causal roles. So Aristotle's account of vision is that light transmits a form to the eye from the object perceived. (An alternative medieval account is that the eye projects something to the object.)

Nominally, Aristotelian scientific explanation has a logical structure. A general principle (e.g., all men are mammals) is given together with a "middle term" (e.g., all mammals are warm blooded) and a particular fact or cause (e.g., Socrates is a man) and a conclusion (e.g., Socrates is warm blooded). Often the middle term is treated as the cause.

In practice, Aristotle's science doesn't look entirely (or sometimes very much at all) like this theory. For example, his texts provide extensive accounts of biology. Among the most interesting is his theory of the generation of living beings. It is a mass of keen and unkeen observations, leaps to conclusions, overgeneralization, rumor and prejudice. In explaining the generation of animals, he posits general

principles, causal mechanisms, ad hoc exceptions to the general principles, and so on. Aristotle's Nature is not value neutral—some kinds of beings are just better than others--males for example than females, masters than slaves.

Here is a bit of what he wrote on conception:

in the individuals which are male and female. And as the proximate motive cause, to which belong the *logos* and the Form, is *better* and more divine in its nature than the Matter, it is *better* also that the superior one should be separate from the inferior one. That is why wherever possible and so far as possible the male is separate from the female, since it is something *better* and more divine in that it is the principle of movement for generated things, while the female serves as their matter. The male, however, comes together with the female and mingles with it for the business of generation, because this is something that concerns both of them. . . .

. . . Just as it sometimes happens that deformed offspring are produced by deformed parents, and sometimes not, so the offspring produced by a female are sometimes female, sometimes not, but male. The reason is that the female is as it were a deformed male; and the menstrual discharge is semen, though in an impure condition; *i.e.*, it lacks one constituent, and one only, the principle of Soul. . . . This principle has to be supplied by the semen of the male, and it is when the female's residue secures this principle that a fetation is formed. . . .

[illegible] and this part is one which is evident to the senses. . . .

. . . A woman is as it were an infertile male; the female, in fact, is female on account of inability of a sort, viz., it lacks the power to concoct semen out of the final state of the nourishment (this is either blood, or its counterpart in bloodless animals) because of the coldness of its nature. . . .

[illegible]

Males are hotter, sperm moves, cool ova apparently do not. (Aristotle did not have any way to measure the "heat" of ova and sperm, or of females and males

for that matter.) Females provide the material cause of the embryo. Males provide the form, which will be male except for when there is some unexplained deformation, which changes the male into its opposite, the female.

Aristotle's analysis of causation partitions causes into four kinds and his account of the generation of animals roughly follows that account. There is a material cause, the ovum; an efficient cause, the insertion of sperm into the ovum; the formal cause of their combination into an entity with potentials; and a final cause—the mature adult. Amidst it all there is another cause, heat, which allows for the production and transmission of sperm.

The causes of organs are explained by their functions in the life of the organism. Thus:

"Not all animals have a neck, but only those with the parts *for the sake of which* the neck is *naturally* present—these are the windpipe and the part known as the esophagus. Now the larynx is present *by nature for the sake of* breathing; for it is through this part that animals draw in and expel air when they inhale and exhale. *This is why* those without a lung have no neck, e.g. the kind consisting of the fish. The esophagus is the part through which nourishment proceeds to the gut; so that animals without necks manifestly do not have an esophagus. But *it is not necessary* to have the esophagus *for the sake of* nutrition; for it concocts nothing. And further, it is *possible* for the gut to be placed right next to the position of the mouth, while for the lung this is *impossible*. For there *needs* first to be something common like a conduit, which then divides in two and through

which the air is separated into passages—in this way the lung *may best* accomplish inhalation and exhalation."[1]

Among Aristotle's most noted biological observations are descriptions of the structural development of chicks in bird eggs. The observations were accompanied as usual for Aristotle with an interpretation of functions: the white of the egg is the source of the chick, the yolk its nutrients. He was half right.

It required millennia until science settled on a conception of causation that is free from purpose or goals, on the conviction that Nature just doesn't give a damn.

Aristotle describes interventions, generally made to aid observations, and he describes a particular kind of causal experiment, the determination of composition. To show that saltwater is composed of water and something else, and thus not a fundamental kind, he evaporated saltwater and condensed the vapor to pure water. But he did not typically report manipulating conditions and observing differences in results to infer that the manipulated condition is a causal variable. This is in part perhaps because he surely knew the result of some alternatives: take the newly laid egg away from the hen and the egg does not develop into a chick. Nor did Aristotle incline to quantitative measurements. He did not, for example, report the weights of chicks. Aristotle's writings on scientific method contain essentially nothing about experiments in the modern sense and how to conduct them. It is no surprise that where his work was the research manual for science, causal inference from experiment played little role.

[1] From J. Lennox, "Aristotle's Biology," Stanford Encyclopedia of Philosophy

Three centuries after Latinizing Aristotle, Europe recovered a quite different approach to science in the works of Archimedes. While Aristotle's work survived and was read for millennia, it did not take long for civilization to lose Archimedes. We have only parts of three books, and much of what we think we know of him was written centuries after his death. Archimedes science, containing (parts of) three of his works bound in a book (a "codex"), were hidden in a Greek monastery from Roman invaders. The monks wrote over the text to inscribe religious nonsense, and subsequently pages were painted over. Somehow, the codex survived, and Archimedes writing was recovered by technical means. I am told that Jeff Bezos now owns the thing, bought at auction. But by the 16th century Archimedes work on floating bodies was known to mathematicians, notably to Galileo.

Archimedes most famous work, "On Floating Bodies," is geometry as physics. Treating bodies of uniform density, Archimedes identifies volumes with weights, and weights with forces. He computes the shapes above and below water of various floating bodies from their geometry, but the explanations he gives are causal, the weights pressing up and down on objects, typically calculated for the center of gravity of an object. A key idea is equilibrium: when the forces pressing a thing in opposite directions cancel, the thing does not move in either direction. When the weight of displaced water equals the weight of the part of a ship displacing it, the ship neither sinks nor rises. There is no hint that he carried out experiments to confirm his calculations.

Archimedes is credited with a host of other discoveries, for example the compound pulley and the law of the lever, some of which were surely known before him. It should be noted that regularities about equilibria, as of floating bodies or balancing on sides of a fulcrum, are also laws about non-equilibria and implicitly about effects of changes, as in moving a balancing weight on one side of a fulcrum closer to it or farther away.

Ancient craftsmen and architects measured lengths and angles and weights and the passage of time, but the full application of mathematics—for the Greeks, number theory and geometry--to Nature came in astronomy and optics. Eudoxus proposed a system of spheres on which the planets move, which Aristotle took literally. Ancient mathematical astronomy reached a pinnacle in the *Almaghest* of Ptolemy, written in Greek sometime in the 2nd century, A.D. and translated into Arabic (hence the name; the original Greek title is, in English translation, "mathematical systematic treatise") and then in the 12th century into Latin. Until Copernicus work in the 16th Century it was, with modifications, the universal astronomical theory in Europe.

The *Almaghest* presents a mathematical theory of the causes of observations (by Ptolemy and before him) of the apparent positions and motion of the stars, the sun, the planets and the moon on the celestial sphere. The mathematics is all geometry. The theory at its time and as it developed in Arabic and European lands fits and predicts the positions of the planets with respect to the stars with great precision and explains and predicts solar and lunar eclipses.

Ptolemy begins the *Almaghest* by arguing that these apparent motions must be produced by light from the planets and stars as these bodies move through space on closed paths, not by motions in a straight line.

"For if one were to suppose that the stars' motion takes place in a straight line towards infinity, as some people have thought, what device could one conceive of which would cause each of them to appear to begin their motion from the same starting-point every day? How could the stars turn back, if their motion is towards infinity? Of, if they did turn back, how could this not be obvious? [On such a hypothesis], they must gradually diminish in size until they disappear, whereas, on the contrary, they are seen to be greater at the very moment of their disappearance, at which time they are gradually obstructed and cut off, as it were, by the earth's surface…"

Ptolemy wrote another book, only partially translated into Latin and long ignored, on the physics of planetary movement. An Arabic version of part of it was only discovered in the 1960s by Bernie Goldstein at the University of Pittsburgh. Many commentators on the *Almaghest* have taken it to be intended only as a calculating system for predicting the positions of the observable planets and the moon. Ptolemy seems quite clear that he meant it to be more than that. Even so, he worries that the circular paths he postulates would overlap, which posed a problem for their realizations by material spheres—but not necessarily for their physical orbits in space.

The basic Ptolemaic framework has the Earth as the center of the universe, with the stars and the sun rotating around it. The apparent motions of the planets and the moon are compositions of circular motions in a plane. A planet (for example) moves on a circle (the epicycle) around a center that moves on another circle (the deferent) around the Earth or offset from the Earth. What we see of the planet are the appearances of these motions as we look out from the Earth in this plane.

Ptolemy also wrote on optics, and he wrote another book, *Tetrabiblios,* on astrology. Astrology was a companion to positional astronomy at least from the time of the Babylonians, continued in ancient Egypt and thence to Greece and into the Renaissance. In the 16th century, Johannes Kepler made his living as an astrologer. Astrology was used for prediction—of the course of lives, of weather, of catastrophes, and so on, but at least for Ptolemy we know it was a causal subject. Ptolemy argues that since we know that the position of the sun affects matters on Earth, why should that not also be true of the planets?

There is of course no hint that the causes of patterns of motion in the night sky might be alterable, quite the contrary. Causal explanation in astronomy made no assumption that the causes of celestial are manipulable, and had no engagement with even informal experimentation except about methods of naked eye measurements comparing times and positions. No one, so far as I know, described praying to change the orbit of Mars, although in the book of Joshua, God makes the Sun and the Moon stand still. But surely informal experimentation was going on. Among our earliest potential examples are the works reported by Hero of Alexandrea in Egypt and Archimedes of Syracuse in Sicily. Hero describes a

variety of mechanisms that can be driven by steam or air but he does not claim that any of them are his own discovery. We have no idea whether or how Hero experimented, but surely some of his sources did.

Discussions of causality in the post Roman centuries were principally concerned with the powers of God and men. God created everything, of course, but did God cause every event? Were effects *logical* consequences of their causes? How could miracles happen? Some of these issues are still debated, albeit in secular framing. If God has foreordained every happening, can human will be a cause of anything? (If our brains determine what we do, does our will cause anything?) If God has prespecified what shall happen, are we responsible for what we will? (If our brains cause us to will as we do, are we responsible for what we do or what we will?). Attempts to influence God through prayer continued unabated, although no one seems to have been counting their sums of success and failure.

There were occasional medieval expressions of a nearly modern conception of causality and causal inference. William of Ockham claimed that when there is a known effect of a cause one can confidently assume that like circumstances will produce a like effect. In the 17th century Isaac Newton offered the same as a "Rule of Reasoning in Natural Philosophy," and the "uniformity of nature" was a common theme in 18th and 19th century methodology. The key thing, which none of them answered, is how to tell when circumstances are so alike that the evidence of a single case suffices. Never pass up a banality too good to pass up.

The great change we call The Scientific Revolution took place in the 17th century, and again in the 19th. By the end of the 19th century the basics of every science now taught in secondary school—with the exception of DNA--had been established. Much of it was established by violating one of the fundamental dogmas of modern methodology: don't derive your hypotheses from your data and take your data to support those same hypotheses. And, outside of astronomy, science was established almost entirely by eye-balling, without any formal statistics.

David Wooten. *The Invention of Science* (p. 18) gives a nice account of the change of mind of an educated Englishman in 1600 as against a century and a quarter later, 1733.

In 1600 "he believes in witchcraft and has perhaps read the Daemonologie (1597) by James VI of Scotland, the future James I of England, which paints an alarming and credulous picture of the threat posed by the devil's agents. He believes witches can summon up storms that sink ships at sea – James had almost lost his life in such a storm. He believes in werewolves, although there happen not to be any in England – he knows they are to be found in Belgium (Jean Bodin, the great sixteenth-century French philosopher, was the accepted authority on such matters). He believes Circe really did turn Odysseus's crew into pigs. He believes mice are spontaneously generated in piles of straw. He believes in contemporary magicians: he has heard of John Dee, and perhaps of Agrippa of Nettesheim (1486–1535), whose black dog, Monsieur, was thought to have been a demon in disguise. If he lives in London he may know people who have consulted the

medical practitioner and astrologer Simon Forman, who uses magic to help them recover stolen goods. He has seen a unicorn's horn, but not a unicorn. He believes that a murdered body will bleed in the presence of the murderer. He believes that there is an ointment which, if rubbed on a dagger which has caused a wound, will cure the wound. He believes that the shape, colour and texture of a plant can be a clue to how it will work as a medicine because God designed nature to be interpreted by mankind. He believes that it is possible to turn base metal into gold, although he doubts that anyone knows how to do it. He believes that nature abhors a vacuum. He believes the rainbow is a sign from God and that comets portend evil. He believes that dreams predict the future, if we know how to interpret them. He believes, of course, that the earth stands still and the sun and stars turn around the earth once every twenty-four hours – he has heard mention of Copernicus, but he does not imagine that he intended his sun-centred model of the cosmos to be taken literally. He believes in astrology, but as he does not know the exact time of his own birth he thinks that even the most expert astrologer would be able to tell him little that he could not find in books. He believes that Aristotle (fourth century BCE) is the greatest philosopher who has ever lived, and that Pliny (first century CE), Galen and Ptolemy (both second century CE) are the best authorities on natural history, medicine and astronomy. He knows that there are Jesuit missionaries in the country who are said to be performing miracles, but he suspects they are frauds. He owns a couple of dozen books. Within a few years change was in the air. In 1611 John Donne, referring to Galileo's discoveries with his telescope made the previous year, declared that 'new philosophy calls all in doubt'. 'New philosophy' was a catchphrase of William Gilbert, who had published the first major work of experimental science for 600 years in 1600; for

Donne, the ‘new philosophy’ was the new science of Gilbert and Galileo. His lines bring together many of the key elements which made up the new science of the day: the search for new worlds in the firmament, the destruction of the Aristotelian distinction between the heavens and the earth, Lucretian atomism.

“Let us take an educated Englishman a century and a quarter later, in 1733, the year of the publication of Voltaire’s Letters Concerning the English Nation (better known under the title they bore a year later when they appeared in French, Lettres philosophiques), the book which announced to a European audience some of the accomplishments of the new, and by now peculiarly English, science. The message of Voltaire’s book was that England had a distinctive scientific culture: what was true of an educated Englishman in 1733 would not be true of a Frenchman, an Italian, a German or even a Dutchman. Our Englishman has looked through a telescope and a microscope; he owns a pendulum clock and a stick barometer – and he knows there is a vacuum at the end of the tube. He does not know anyone (or at least not anyone educated and reasonably sophisticated) who believes in witches, werewolves, magic, alchemy or astrology; he thinks the Odyssey is fiction, not fact. He is confident that the unicorn is a mythical beast. He does not believe that the shape or colour of a plant has any significance for an understanding of its medical use. He believes that no creature large enough to be seen by the naked eye is generated spontaneously – not even a fly. He does not believe in the weapon salve or that murdered bodies bleed in the presence of the murderer. Like all educated people in Protestant countries, he believes that the Earth goes round the sun. He knows that the rainbow is produced by refracted light and that comets have no significance for our lives on earth. He believes the

future cannot be predicted. He knows that the heart is a pump. He has seen a steam engine at work. He believes that science is going to transform the world and that the moderns have outstripped the ancients in every possible respect. He has trouble believing in any miracles, even the ones in the Bible. He thinks that Locke is the greatest philosopher who has ever lived and Newton the greatest scientist. (He is encouraged to think this by the *Letters Concerning the English Nation*.) He owns a couple of hundred – perhaps even a couple of thousand – books."
--Wootton, David. The Invention of Science (p. 22). HarperCollins. Kindle Edition.

And in another 166 years our educated Englishman in 1899 is an atheist or agnostic but in any case believes natural phenomena are to be explained by natural causes, believes humans are animals and developed from them, that diseases are caused by germs and may be defeated by pharmaceuticals, that electricity may replace gaslight, that horseless automobiles and even flying carriages are possible, that materials are made of atoms of a finite number of elements that can combine only in definite ways.

All of these changes are due, of course to ingenuity and labor, but also and essentially to revolutions in methods of inquiry and criteria for considering a claim to be a discovery.

Let's start with Johannes Kepler. Kepler was employed as a mathematician by Tycho Brahe who maintained a (naked eye) observatory on a Danish island. Tycho had upset Aristotelian ideas of the unchangeability of the heavens by observing a

sudden change in the sky—what we now know was a supernova. Upon Tycho's death, Kepler took Tycho's extensive data and moved to the continent where he set upon improving the Copernican heliocentric theory of the solar system, resulting in, among other things, what we know as Kepler's three laws: the planets move in elliptical orbits with the sun as one focus; a line from the sun to a planet sweeps out equal areas in equal times; and the ratio of the square of the time for a planet to complete an orbit to the cube of the maximum distance of the sun and the planet is the same for all planets.

Kepler's Method

Kepler introduced a form of argument for laws relating unobserved quantities. The idea is quite simple:

1. Start with known regularities among observed quantities;
2. Identify the values of observed quantities with values of theoretical quantities;
3. Show that with that identification, the observed regularity is transformed into a known law or mathematical necessity.

One of Kepler's examples is vivid but a bit technical. In the Almaghest, Ptolemy had noted the following regularity for the "superior" planets: Mars, Jupiter and Saturn (superior in the sense that unlike Mercury and Venus, these planets could appear in the sky at any angle from the location of the sun):

In a long number of solar years, the number of revolutions of longitude of the planet plus the number of cycles of anomaly of the planet equals the number of solar years.

A solar year is the period of time between the appearance and reappearance of the sun in the same place with respect to the "fixed" stars. A revolution of longitude is the period of time between the appearance and reappearance of the planet in the same place with respect to the fixed stars. A cycle of anomaly is the period of time between the appearance and reappearance of the planet opposite the sun on the celestial sphere—it is also the period between two episodes of greatest brightness of the planet and also the period between two episodes in which the planet starts to move backwards on the celestial sphere. These are quantities that had been measured for millennia or could be extracted from historical measurements of positions on the celestial sphere.[2]

#Solar Years = #Revolutions of Longitude + #Cycles of Anomaly

Ptolemy said he had no explanation for this regularity—meaning that it does not follow from his general model of the solar system in which each of the sun and planets move on a separate deferent and epicycle around the Earth (or off center from the Earth). It can be accommodated in his theory by appropriate adjustment of parameters (essentially, by keeping the line from a superior planet to the

[2] If you are puzzled by how astronomers from ancient times measured such things as the position of the sun with respect to the fixed stars—given that when the sun is out the stars are invisible—I recommend reading Thomas Kuhn's *The Copernican Revolution*.

center of its epicycle always parallel to the Earth – Sun line), but Ptolemy clearly did not regard that as an explanation.

Kepler made these identifications:
Solar year= period of one complete orbit of the Earth around the Sun;
Revolution of Longitude = period of one complete orbit of the planet around the Sun.
Cycle of Anomaly = period between one colination (opposition on celestial sphere) of the Sun – Earth—Superior planet.

1. #Solar Years = #Revolutions of Longitude + #Cycles of Anomaly
   || || ||
2. #Earth orbits = #Superior planet orbits = #Oppositions

Given one further fact that is implied by these identifications and the observed facts, namely

The (constant) period of an Earth orbit is less than the (constant) period of a superior planet, i.e., the Earth moves faster around the Sun than do the superior planets.

RI leave it to the reader to prove that Regularity 2 is then a mathematical necessity, which was Kepler's point: An explanation of an empirical regularity is reduction to something necessary or at least certain.

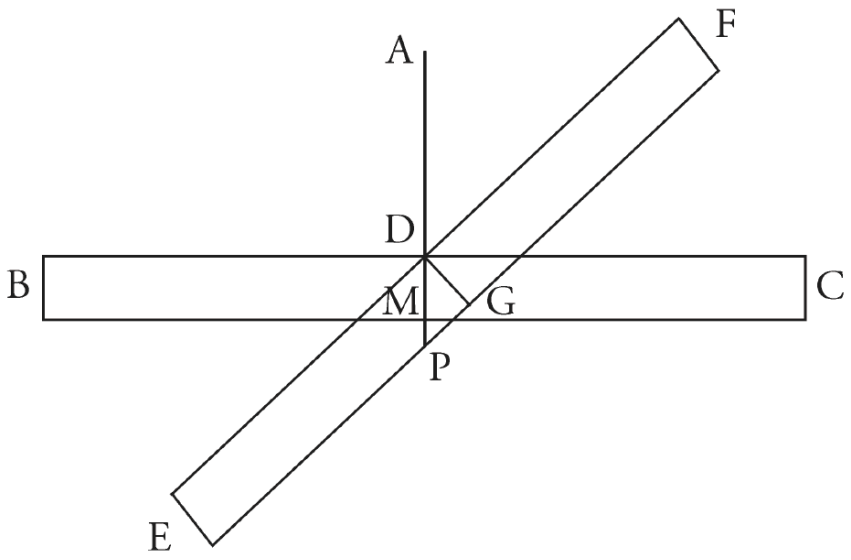
Medieval mechanics

Medieval mechanics had two sources, authentic Aristotelian texts and a treatise on Mechanical Problems, attributed to Aristotle but now thought not to have been produced by him. A central idea in the latter work is that all mechanics is based on the lever; the motion of the lever about its fulcrum is an arc of a circle, so mechanics is the mathematical description of machines in term of circles. It begins with this interesting remark:

"…the how is clear through mathematics, the what is clear through physics."

A simple illustration of the reasoning in the Mechanics is the second problem: why does a beam supported at the middle from above become or remain level, when one supported from below at the middle does not? The reason is that when supported from the bottom, more than half of the weight is on one (or the other) side of the support[3]:

[3] Winter, T. N. (2007). The mechanical problems in the corpus of Aristotle.



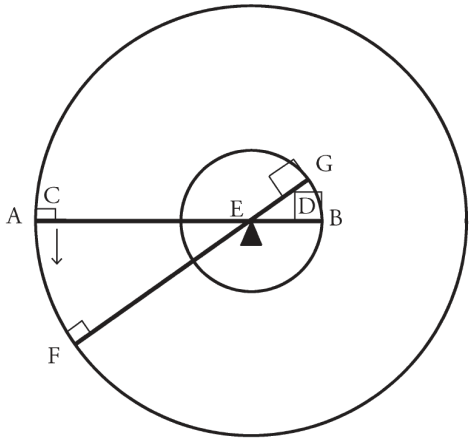
So if a slant is imposed on B, B will be at E and C will be at F. S
the dividing line, at the exact vertical DM before, will be at DG dui
ing the slant. So the part of beam EF outside the vertical, marked G
makes that side greater than half.

Illustrating that the issue is the proper geometrical representation of systems and interpretation of causal relations, Giovanni Benedetti, writing in 1553, took issue with this explanation, thinking of the beam as the top of a balance:

"But the true cause of why, if the support is form the above and one arm of the balance is depressed and is then let free, it returns to the horizontal position, is not the greater weight of the balance which has passed beyond the vertical line, but also the length of the raised arm found beyond the vertical line. Therefore the weight at this end is [effectively] greater in the ratio which I shall set forth…"[4]

[4] Drake, S., & Drabkin, I. E. (1968). Mechanics in sixteenth-century Italy: selections from Tartaglia, Benedetti, Guido Ubaldo & Galilei. *Mechanics in sixteenth-century Italy: selections from Tartaglia*., p. 182.

The principle of the lever is explained by circles in “Aristotle’s” mechanics. Why a smaller weight can via a lever move a greater weight is explained by motions through circles:



“When the part farther from the center gets moved more quickly by the same weight, there are three things about the lever: the fulcrum—string and center, and two weights, the one moving and the one getting moved. The weight getting moved to the weight moving is the opposite of length to length. And always, the farther from the fulcrum, the easier it will move. The reason is the aforesaid, that the more distant from the center scribes the larger circle. So by the same force, the mover will manage more the farther from the fulcrum.

“Let there be a lever AB, a weight on it C, the motive weight on it D, fulcrum E. The D, moving, goes to F, the weight being moved to G.”

There is a considerable discussion of how and where oars move boats.

Medieval mathematical mechanics, like Archimedes', is conceptually more complex than ancient astronomical mechanics because it attempts mathematical representations of cause-effect relations. The causes are the weights of objects and their positions, shapes and volumes, all of which (with the help of uniformity assumptions about composition and the force of gravity) were represented geometrically. For mathematical astronomy, causes were entirely supplementary and not represented by the mathematics. Stillman Drake, otherwise astute, claims that the "mathematicians" of the 16th and 17th centuries abandoned causes for mathematical relationships. I think that is incorrect, perhaps based on Drake's view—as in the definitions offered by Hume and Galileo--that causes are happenings, events. Instead, from at least the 14th century, mathematical philosophers had attempted to treat causes as variables that could be related mathematically—usually as relations of proportions. So a question that concerned some Oxfordians in the 14th century was whether a rigid rod falling toward its natural place according to Aristotle, i.e., the center of the world, could come to rest at that place. Relating distance from two sides of the center to the varying force of attraction to the center, Swineshead gave a mathematical argument that the center could not be reached, essentially a two-sided version of the Tortoise and the Hare.[5] And by the end of 16th century Galileo had fully mathematized falling bodies—but I doubt that even for a moment he thought, as Drake's view would imply, that the height of a body above the Earth when released was not a cause of the velocity of the body in free fall.[6] The scientific

[5] Hoskin, M. A., & Molland, A. G. (1966). Swineshead on Falling Bodies: An Example of Fourteenth-Century Physics. *The British Journal for the History of Science*, *3*(2), 150-182.

[6] Influenced by Jesuit teachers, the young Galileo was immersed in Aristotle's physics and metaphysics. See *Galileo's Early Notebooks.* I do not know when and why he began to reject the Arististotelian way of thinking.

conquest of the terrestrial world depended on the introduction of variables as causes with laws expressed in terms of ratios or proportions. Algebra reached Europe from North Africa in the 14th century but it took a while to be used as a language for causal laws.

The Experimental Example

As a student and for some time after, Galileo was immersed in Aristotelian metaphysics and explanations. Around 1604 he began to think in new ways, influenced perhaps by phenomena of the balance and the pendulum. Causal inference became critical in astronomy with Galileo's uses of his telescope. Galileo discovered four satellites of Jupiter, which he named after the Medici family, who played the real Game of Thrones. The motions he inferred from changes in the positions of these new "stars" provided a model of the heliocentric system, but Galileo did not make the analogy in his book announcing his finding. He confined himself to arguing that the Jovian system was evidence for the proposition that the moon orbits around the Earth. Second, Galileo discovered sunspots and their motion, which he attributed to the motion of the Sun. Galileo hotly debated with his critics the nature and genesis of sunspots, which he thought analogous to clouds on Earth, although not of course of the same composition. Others argued that they could be small planets orbiting inside the orbit of Mercury. The issue was important because sunspots as a feature of the Sun contradicted the received opinion of the time that the Sun did not change except in moving in its orbit around the Earth. And third, Galileo discovered that Venus has phases like those of the Moon, which can appear from reflection to

Earth of light from the Sun on Venus on a heliocentric model of the solar system but not on the Ptolemaic, geocentric model.

Galileo's systematic experiments with bodies moving on inclined planes involved careful observation of numerical values and relevant calculations with them. They have been faithfully reconstructed by historians from remaining manuscripts.[7]

In 1600 William Gilbert, who later became physician to Elizabeth the first published *On the Magnet and Magnetic Bodies, and on That Great Magnet the Earth* (ok, the Latin version of this title). Gilbert's work was mostly trial and error demonstration of magnetic phenomena, including induced magnetism. His remarkable inference was from the known changes in declination of compass needles at different longitudes and the similar declinations of freely moving needles at different positions with respect to a magnet to the conclusion that the Earth has a magnetic field.

Understanding of causal inference emerged in the 16th and 17th centuries more by example than by theory, but there was one influential theorist, Francis Bacon. Bacon was a British government administrator, left sufficient time to write a number of speculative projects for the improvement of science. His most lasting contribution was a method for causal discovery phrased as the discovery of "forms." For discovering the cause of a phenomenon, Bacon's method was to

[7] Notably in MacLachlan, J. (1973). A Test of an" Imaginary" Experiment of Galileo's. *Isis*, *64*(3), 374-379.and S. Drake Drake, S. (1973). Galileo's experimental confirmation of horizontal inertia: unpublished manuscripts (Galileo gleanings XXII). *Isis*, *64*(3), 291-305.

make and survey three lists: A list of circumstances in which the phenomenon occurs, a list of circumstances in which a variable phenomenon increases, and a list of circumstances similar to the first list but in which the phenomenon does not occur. The idea is to look for whatever feature is shared by members of the first list but does not occur in the third list and increases in intensity in accord with the second list. Bacon gave psychological guides but no algorithm. His notable example is that heat is caused by motion. Contrary to some modern popular accounts, his lists were not meant to be from passive observation alone. Bacon recommended manipulating things, or as he put it "torturing nature." The metaphor was apt for the time.

Bacon's conception of how to find causes incorporates the thought that causes are not just happenings, they are features that vary in degree, what we would now call variables. That conception is different in focus from the one formulated at about the same time by Galileo and later in the 18th century by David Hume, who (well, Hume did so once) define causes to be conditions or happenings such that were they absent the effect would not exist.

William Harvey, physician to Kings James I and Charles I of England, provided an early 17th century example of a system analysis—the circulatory system—with a causal theory and both quasi-experimental and observational support. Harvey observed venous valves and noted that their direction was to prevent blood from flowing away from the heart. He understood the heart as a kind of pump. Those measurements and observations he used to argue that the blood circulates in the entire body rather than, as was thought at the time, being entirely absorbed in

the body. He supplemented these observations with simple experiments, for example that blood accumulates in a limb when the superficial vessels (veins) are constricted. His theorizing was in an Aristotelian, not a mechanical, mode. Harvey opposed the "mechanical" explanations of natural phenomena that were becoming fashionable in the 17th century.

The ideal of the new "experimental philosophy" of the 16th and 17th centuries was an effort to find the composition and causes of things and phenomena from observation of natural phenomena or the production of new phenomena, and to formulate mathematical regularities from them. The causes and compositions were typically fanciful, and the observations were sometimes whimsical. Robert Hooke, Boyle's assistant and Newton's nemesis wrote a long book, *Micrographia*. Hooke had got hold of a microscope and that provided him with a world of wimsey. Observing that apparently solid materials had "pores," Hooke explained all sorts of things from the properties he supposed of pores. John Mayow, an Oxford contemporary, was enamored of potassium nitrate, or nitre as it was called at the time, a fascinating substance ever since gunpowder was introduced into Europe in the 13th century. The "elasticity" of air is due to its "nitro-aerial particles;" nitro-aerial particles produce sparks and transmit light, and more.

Interpreting experiments is not straightforward, and it was especially not in the 16th and 17th centuries. Robert Boyle, Nicholas Lemery, Evangelista Torricelli—Galileo's sometime assistant--and Blaise Pascal were rough contemporaries of Isaac Newton (who was born the year Galileo died) all of whom conducted and

interpreted experiments. Boyle even wrote a book on experimentation, on which he had less than a firm grip.

Boyle's tool was the recently invented (by Otto van Gericke) air pump which he used to evacuate vessels. They were not easy or cheap to construct at the time, but Boyle was wealthy and he had his own laboratory assistant, Robert Hooke. Boyle's famous law was probably formulated by Hooke. Boyle's experiments with the air pump largely consisted of noting that it was hard to pump further air into a vessel already filled with the stuff—air resisted, it has a "spring"--and in putting things in a closed glass vessel, pumping out the air and seeing what followed: candles went out, birds died. Boyle's experimental sequencing was not always apt. Upon burning metals in a sealed glass vessel, weighed before the burning, and weighed after, he concluded that calcining, as it was called, did not involve combination of the metal with air but rather particles penetrating the glass vessel and combining with the metal, evidenced by the increased weight of the system before and after burning the metal. Unfortunately, after burning the metal in the sealed container, he opened the container to the atmosphere before he weighed the system a second time. Boyle was an atomist, or in the terminology of the time a corpuscularian, and his theoretical commitments sometimes coated his observations. Thus he claimed to see corpuscles moving about when water boiled.

Notwithstanding, Pascal put store in Boyle's experiments and used the weight of the air to explain some of them—notably the difficulty of removing a plug from an evacuated vessel. Pascal has a puzzling account of Boyle's claim that after

removing the plug Boyle felt resistance on replacing it. Pascal seems to think that was from residual air left in the vessel after evacuating it. I suspect Boyle had let the atmosphere reenter the vessel after removing the plug and before reinserting it.

Nicholas Lemery in Paris and Evangelista Torricelli in Florence worked in the same decades of the 16th and 17th centuries and their work bears comparison. Torricelli is famous for inventing the barometer while Lemery is forgotten, but they had similar struggles in interpreting their experiments. Torricelli had the advantage of Pascal and of a less whimsical theory than Lemery's.

Lemery's theory, probably influenced by Mayow, was that air consists of a web or three dimensional net (thus accounting for its "spring"—i.e., resistance to compression) held together by "nitro-aerial" particles. Lemery reproduced an ancient experiment, first reported in the *Pneumatica* of Philo of Byzantium in the 3rd century B.C., in which a candle is burned under an inverted jar open at the bottom in a pool of water. The effect is that the candle burns out and the water rises inside the jar. Lemery's explanation was that the heat of the candle flame knocks loose the nitro-aerial particles of the air, causing the air to lose its springy resistance to the water. (The experiment is a commonplace of school science, commonly with false explanations, e.g., that by combining the oxygen in the glass with carbon in combustion, the total number of particles in the air in the glass, and hence the pressure, is reduced. Another is that as the system cools after burning, the air pressure is reduced in accord with Henry's law and the water rises. Hint to schoolteachers: water and carbon dioxide are the products of

burning hydrocarbons and carbon dioxide is about 22 times more soluble in water than is oxygen.)

Torricelli used long glass tubes closed at one end and opened at the other. After filling a tube with mercury and suspending the tube upright in a pool of mercury, he found the mercury fell in the tube to a height above the level of the mercury in the tube. He experimented with different shapes of bulbs at the closed end of the tubes and different angles of the tubes to the surface. The height the mercury stood in the tube was always the same.

Torricelli's attributed the phenomenon to the weight of the air pressing on the pool of mercury in which the tube stood. It was not surprising that a force in one direction, down, could produce a force in the opposite direction, up, since that kind of effect was familiar to everyone from the behavior of partially filled bladders when the sides or one part of the top are pressed. But there were Aristotelian objections. Perhaps the mercury communicated some attracting form to the air above it to keep the mercury raised. Perhaps there was in invisible string suspending the mercury, etc. A sensible objection was given by Ricci, a prelate and friend of Torricelli's, who pointed out that if a lid is put over the pool of mercury in which the tube sits, the weight of air would rest on the lid not the mercury in the pool--but the mercury still stands in the tube. Torricelli responded that either there would be no space between the lid and the mercury, in which case the weight would be on the mercury, or there would be a vacuum (whose possibility Aristotelians denied) or there would be a density of air in the space which would press against the Mercury. Moving from the weight of air to density

was progress but also a different explanation. Torricelli appears to have had no clear conception of pressure as force per unit area or of the relation between weight of a gas per unit volume (density) and pressure, nor did have a very satisfactory response to Ricci. Neither he nor Ricci could have satisfactorily explained why a paper box does not collapse when it is sealed up. It remained for Pascal to provide some clarity to the matter through his development of a theory of forces in fluids.

After measuring barometer levels using water and wine, Pascal, living near Paris, persuaded his brother-in-law along with a community group to carry a mercury barometer to the top of a small mountain, the Puy de Dome, and compare its reading on the mountain with its reading at the base and with another barometer left at the base. The level of mercury fell as they climbed and then rose to its original level when they returned. Assuming that the column of air above altitude weighs less than the same column above ground level, Pascal and others, although far from all, took this as confirmation of Torricelli's conclusion. The experimental reasoning implements Francis Bacon's idea of causes as variable quantities. The experiment did not take account of the effect of temperature differences between the base and the peak of the mountain—scientific results on the effect of temperature on pressure had to wait for another half century--but Pascal wrote to his brother-in-law about the issue, and the latter gathered data on temperatures at a few different sites around France.

In the 17th century, scientific leaders who were not Aristotelian were devoted to "mechanical" explanations, meaning that all fundamental causes are pushes or

pulls of rigid bodies on one another. Invisible "corpuscles" constitute matter. They have rigidity, weight, shape, and motion. Lemery's case is described above. Robert Boyle explained the "spring of the air" (resistance to compression) by supposing that fundamental air corpuscles are shaped like curly wood shavings—little springs. The accretions of corpuscles into larger materials was explained by corpuscular shapes fitting into one another, like three-dimensional jigsaw puzzles, or by supposing that corpuscles had hooks and eyes. Descartes provided an influential version of the "mechanical philosophy" in which the fundamental entities (besides souls) are vortices of various sizes that entirely fill space. There is no void or vacuum according to Descartes. Causes are propagated by the rotation of one vortex causing others with which it is in contact also to rotate by some sort of friction or tooths as in gears. Anyone who has worried about sand in gear boxes will recognize that Descartes vortices could not rotate at all. (I do not know that any 17th century writers made this objection.) Descartes denied that weight (what we would call mass) is a fundamental property.

Descartes method, as he described it, was to consider "clear and distinct" ideas, which he argued could not be false because God would not deceive us. Two properties are distinct only if one can be conceived clearly and distinctly without the other. So, for example, an unextended body cannot be conceived (clearly and distinctly), so body is the same property as spatial extension—or volume.

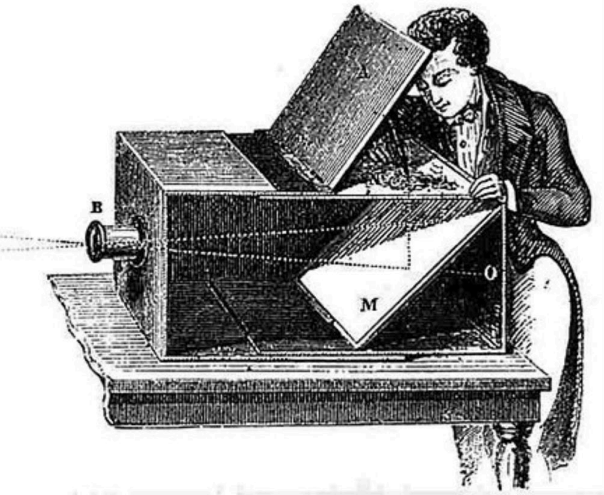
Descartes and 17th century Cartesians promulgated his framework to provide explanations of almost everything physical, from light transmission to the orbits of planets.

Descartes did extensive research on refraction and reflection, and tried to use his discoveries to explain the phenomena of the rainbow. His exposition of his research program is close to unfathomable in this case, but a general strategy can be abstracted. To explain a phenomenon seek whatever observable features will produce, alter or control it. If none can be found posit an unobserved mechanism that could, were it real, produce the phenomenon and test consequences of the mechanism. Unfortunately, his mechanism was that a ray of light consists of spheres rotating against one another.

Optics was one domain in which medieval and Renaissance investigators did sometimes do something akin to modern experimentation, although much of the work, for example on what generates the colors of the rainbow, was in some or large degree speculative. Much of the Renaissance work in Europe was derivative from the work of the Islamic author Al Hazen, which was in turn derivative from Ptolemy's optics. It was essentially geometric optics with speculations about how the eye worked. One topic, refraction, was pursued in original ways in the 17th century. The result was the sine law of refraction which we now know as Snell's law but which was known to several people at the time and first published by Descartes. The law is an empirical regularity not a mathematization of a causal relation. Causal explanations were soon produced, notably one by Descartes and one by Newton. Fermat criticized Descartes' explanation—and implicitly Newton's

as well. In 1657 Fermat produced his variational derivation based on the principle of least time, which like the sine law itself was not a causal explanation.[8]

Another optical concern of medieval and Renaissance researchers was the camera obscura, essentially a box with a pinhole on one side through which light from outside the box could pass and some means of observing the resulting image on a surface inside the box.



A camera obscura box with mirror, with an upright projected image at the top

Images were reversed and upside down in the camera obscura, and the question was why. The explanations given were from geometrical optics (light rays moving in straight lines) and reflection.

[8] For an interesting discussion of Newton's and Descartes theories and their connection with modern theories of light see https://www.mathpages.com/home/kmath721/kmath721.htm. e

Medieval and Renaissance empiricists favored looking but by and large did not know how to change observation into scientific knowledge. Systematic, sensible experimentation was rare—Galileo and Torricelli and Pascal stand out—but earlier, Roger Bacon and Mayow and Robert Grossteste were more typical: curious about everything, full of speculation, and unable to turn observations into a well-founded understanding of constituents and processes. Chemists like Boyle and Mayow had the most challenging task, since the "kinds"—the chemical elements—were largely unknown. For example, potassium nitrate—"nitre"—had an aerial part and an earthy part said Mayow. Aside from studies of the mechanical properties of air, the appeal to mechanical explanations only gave a scientific veneer and was largely a distraction.

Newton on Gravity

Isaac Newton provided what I count as the first more or less systematic method after Kepler's for establishing causal relations among "unobserved" or "theoretical" variables and features. He used the method both for his theory of gravitation and for his theory of light but it is clearest in the former.

Newton's theory of gravitation is given in pieces in his Principia.
"If the matter of two spheres mutually attracting each other shall be homogeneous on all sides in all directions, which are equally distant from the centres: the weight of the spheres of the one to the other shall be reciprocally as the square of the distance between the centres."

"All bodies are attracted by gravity to the individual planets, and the weights of these for whatever planet, for equal distances from the centre of the planet, are proportional to the quantity of matter in the individual planets."

"Gravitation happens in bodies universally, and that is to be proportional to the quantity of matter in the individual bodies."

.

Only the universality (and reciprocity) of gravitational force seems to have been original with Newton. The idea that the planets were subject to an inverse square force centered on the sun had been proposed by several writers, most notably Newton's nemesis, Robert Hooke.[9] What was dramatically original in the Principia was how Newton argued for his theory and how he drew consequences from it.

Newton's Principia has three parts:

- Book I proves a variety of consequences of his three laws of motion, which are:
    - when net forces are zero, bodies move in straight lines with constant velocity;
    - when a net force applies to a body, the body undergoes an acceleration in the direction of the sum of such forces inversely proportional to the body's mass and proportional to the force applied;

[9] Which may have been the motivation for one 21st century philosopher's claim that Newton's effort was only to estimate the gravitational constant, a reading which I think is remarkably stupid.

    - for any force acting on a body there is a force in the opposite direction and of the same strength (what we now call conservation of momentum).
- Book II applies the results to hydrostatics and dynamics
- Book III applies the results, and his "Rules of Reasoning in Natural Philosophy" to gravitation and the motions of the planets and the Moon.

"Force" was a natural enough notion, but Newton used only one quantitative measurable quantity related to force, weight, and used it in only one experiment--to establish that the period of a pendulum does not depend on the weight of the pendulum bob.

The Rules of Reasoning are essential to Newton's argument for his gravitational theory.

- Rule 1 We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.
- Rule 2 Therefore to the same natural effects we must, as far as possible, assign the same causes.
- Rule 3 The qualities of bodies, which admit neither intensification nor remission of degrees, and which are found to belong to all bodies within the reach of our experiments, are to be esteemed the universal qualities of all bodies whatsoever.
- Rule 4 In experimental philosophy we are to look upon propositions inferred by general induction from phenomena as accurately or very nearly true, notwithstanding any contrary hypothesis that may be imagined, till

such time as other phenomena occur, by which they may either be made more accurate, or liable to exceptions.

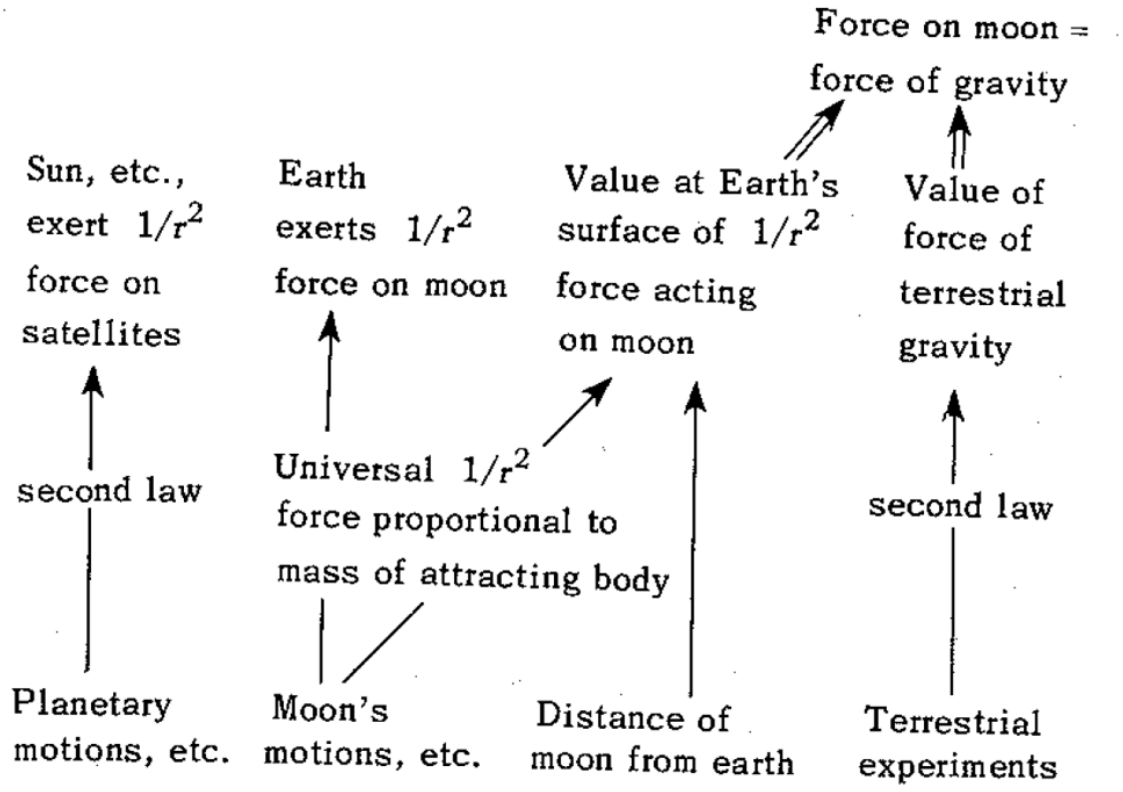
The first rule has been claimed to be involved in a circularity on the grounds that Newton used it to establish his theory, which would presume the truth of his theory. He did no such thing. Having established his theory to his satisfaction on other grounds, he uses the first rule to explain the tides, Kepler's laws, etc.[10] Rule 2 is a simplicity principle that might as well have been formulated by William of Ockham. Rule 3 is a rule of detachment for a generalization from series of positive instances, the fundamental principle of deterministic inductive inference. Rule 4 is to dismiss theories, notably Cartesian theories, that are not founded by arguments of the kind Newton was about to give. And what kind of arguments are those?

From his three laws of motion

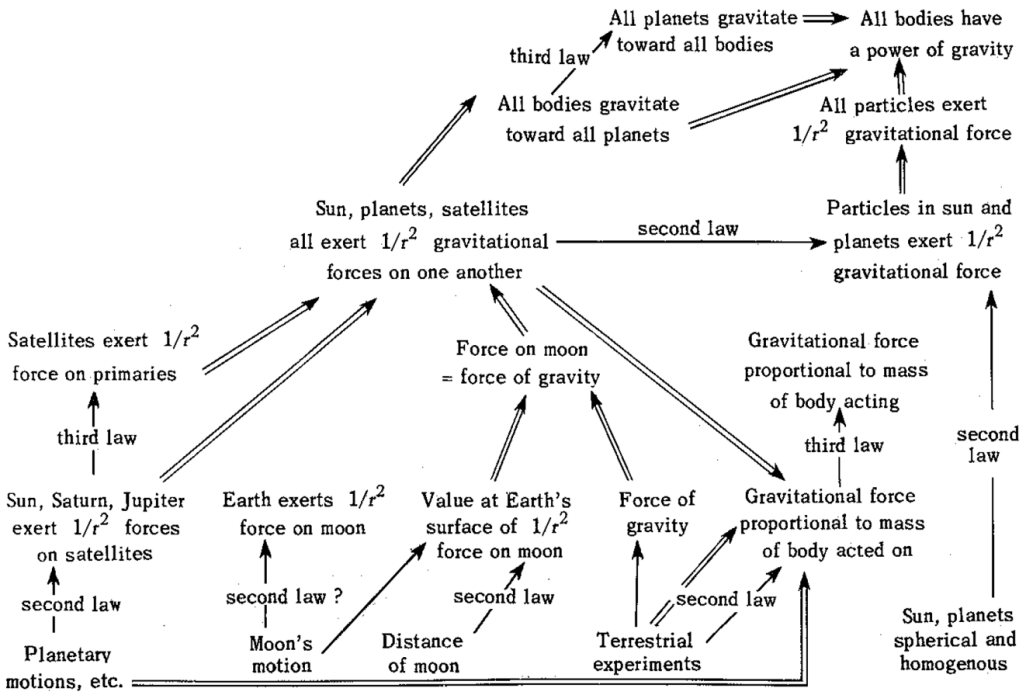
- Objects move with a constant velocity unless acted on by an external force.
- Objects accelerate directly as the force applied and inversely as their mass (or "quantity of matter")
- For every force one body applies to another there is a force directed on the first body of the same magnitude and opposite direction (conservation of momentum)

[10] Of course, Newton used Kepler's laws to obtain his inverse square force laws, as in the diagram. But he did not use Kepler's first law—elliptical orbits. Instead his empirical premise, which he knew to be false, was that the planets move on circular orbits. Having established his gravitational law he then deduces that the orbits are ellipses. I can think of no reason for the bit of sleight of hand other than vanity. Newton wrote elsewhere that Kepler merely conjectured that the orbits are ellipses; he, Newton, *proved* it.

Newton deduces relations between astronomical quantities--the orbits of planets and the stability of aphelia of orbits (the point at which the planet is furthest from the sun) which he proves are unique to inverse square forces; using these, and the experimental claim that the acceleration of terrestrial bodies is independent of their mass (argued from pendulums), Newton infers instances of the law of universal gravitation and generalizes using his rules of reasoning and "generalized induction." In outline, his argument looks something like the following. First, his argument that Earth exerts a force on the moon which is the same force as terrestrial gravity:



The full argument for universal gravitation incorporates the moon argument as a part:



Broadly, Newton's strategy is to use known generalizations and empirical phenomena to deduce *instances* of a purported law, and do so without logical circularity—(although what counts as "circularity" is a thorny problem.) The particular and perhaps peculiar feature is the moon argument, which invokes his rules of reasoning in the principle that if there are putatively two forces acting on a system, one of which cannot be measured in the actual circumstances, and the value of that force (the attraction of the moon to the Earth) would in some non-actual condition (the moon on the surface of the Earth) equal the value of another force (gravity) for that system in that counterfactual situation, then the two forces on the system are one and the same.

The strategy is different from Kepler's. Kepler identified measured quantities with theoretical quantities to transform *entire* empirical regularities into

mathematically provable relationships. Newton identified measured quantities with theoretical quantities to provide *instances* of mathematical features he then generalized and proved theorems about.

The same problem—identifying and measuring unobserved causes—persisted through the 19th century and was resolved in various ways as we will see. Kepler and Newton had the advantage that if their identifications were true, then their conclusions were either true or would be true in the long run under suitable assumptions about the uniformity of nature.

The Birth of Statistics

Something close to the modern theory of statistics began with Laplace's publications in the late 18th century. Beyond the Normal Distribution (which soon came to be called the "error curve") , and the Central Limit Theorem (to which De Moivre and Gauss contributed versions), and beyond Bayesian "inverse" inference (in which Bayes anticipated Laplace), Laplace provided methods for estimation and the finite sample error probabilities (based on the asymptotic limit, of course, but Laplace did not emphasize that), a concern for the multiplicity of reference sets (i.e., what to condition on), and hypothesis testing. His claims for statistics were gargantuan, but his examples were not. He applied his methods to mortality tables and to the longevity of marriages (which he assumed ended only with the death of one of the couple), and to a lengthy discussion of the reliability of testimony. In the latter case he assumed the probability of miracles was infinitesimal and concluded (as had Hume) that miracles are not to be believed.

He criticized Pascal's famous wager argument for the existence of God on the grounds that the probability of that existence is infinitely small. The only physical issue to which his methods were applied were the deviations of planetary orbits from the orbital plane, which he argued probably have a common cause. Gauss provided the close connections between least squares estimates and the Normal distribution.

Between them, Gauss and Laplace had by the late 18th century provided the basis for applications of statistics to scientific data and causal inference, and their work was not obscure. It had little effect. The principle 19th century applications were by two French physicians, Pierre Louis (1836) and Jules Gavarret (1840) who carried out statistical tests. Gavarret noted that establishing a difference between two groups did not determine the cause of the difference, for which in medical cases there might be several. Their advocacy of Laplacian methods met stiff (and contemptuous) resistance, chiefly on the grounds that every patient is a unique case and should be treated as such by the physician, with logic, not numbers. Perhaps the most influential opposition came from Claude Bernard, whose *Introduction to the Study of Experimental Medicine* abjured statistical methods altogether in favor of physiological experiments. Bernard's book is chiefly a diatribe against a priori reasoning in medicine (and elsewhere) and not a practical guidebook to the design of experiments.

Bernard had a curious combination of views. He advocated what today would be regarded as a Popperian philosophy of science—science does not establish causal claims, it only refutes them while keeping in memory those as yet unrefuted. But

he was boastfully proud of having distinguished the vasodilator and vasoconstrictor nerves. At about the same time Ralph Waldo Emerson observed that "a foolish consistency is the hobgoblin of little minds."

Bernard wrote: "The physiologist is no ordinary man. He is a learned man, a man possessed and absorbed by a scientific idea. He does not hear the animals' cries of pain. He is blind to the blood that flows. He sees nothing but his idea, and organisms which conceal from him the secrets he is resolved to discover."
"A great surgeon performs operations for stone by a single method; later he makes a statistical summary of deaths and recoveries, and he concludes from these statistics that the mortality law for this operation is two out of five. Well, I say that this ratio means literally nothing scientifically and gives us no certainty in performing the next operation; for we do not know whether the next case will be among the recoveries or the deaths. What really should be done, instead of gathering facts empirically, is to study them more accurately, each in its special determinism….to discover in them the cause of mortal accidents so as to master the cause and avoid the accidents."[11]

Laplace had remarked that his methods were most appropriate for astronomy and geodesy, and he was right. Aside from national mortality data and some epidemiological data sets they were the sciences with big data, Physics, Chemistry and Biology developed in the following century without help from statistics.

Chemistry in the 18th and 19th Centuries

[11] *An Introduction to the Study of Experimental Medicine*, 1865.

The problem of unobservables was most vivid in the 19th century in debates over atomic explanations in chemistry but chemistry also had other problems.

The substances of the world do not neatly divide into elements and compounds. Whatever the substances, until the end of the 18th century chemists had few ways of manipulating them, chiefly by heating and distilling or simply adding one liquid substance to another or to a solid, or by combining gases once it was learned how to capture gases. When heating a substance produces two products, does that mean the products are components of the substance, or does it mean the products are new substances that combine heat or air with the original substance or its loss? Elements came to be identified with substances that could not be decomposed by heating in isolation or by distilling or filtering.

Heat was thought in the 17th and much of the 18th centuries to be or to be produced by a substance; heating or cooling was the transfer of that substance. In the 18th century that substance was phlogiston. According to the theory, phlogiston and is present in varying degrees according to the substance and escapes from them when they burn. Phlogisticated air (e.g., hydrogen) has lots of it so combusts easily; dephogisticated air (which turned out to be oxygen) not so much. Plants absorb phlogiston from the air (which is why they burn easily and the atmosphere does not.) By the late 18th century the theory was in trouble because metals were found to gain, not lose, weight on burning. One solution among several was to postulate that phlogiston has negative mass—(theories are flexible). By the end of the 18th century, phlogiston began to be abandoned for an

alternative substance, caloric, introduced by Lavoisier in 1783. Caloric was supposed to consist of a fluid of very fine particles that could penetrate other materials (recall Hooke's "pores"). It was self-repulsive and therefore tended to expand. It was also a conserved substance that could not be created or destroyed. In burning, caloric combined with the burning material—the reverse of phlogiston. The theory of caloric was undone by experiments in the 19th century, but lasted until the development of the kinetic theory—so much that key ideas in the 19th century development of modern thermodynamics presumed it. After the development of atomic theory in chemistry early in the 19th century, individual atoms were supposed to be surrounded by an envelope of caloric.

The chemistry of "airs" and their combining weights was pursued by several researchers in the 18th century, with conflicting interpretations. Lavoisier's *Treatise of Chemistry* in 1783 changed the subject, providing a consistent, logical interpretation of a variety of results, and including demonstrations that water, hitherto thought an element, is a compound of oxygen and hydrogen, asserting the conservation of mass in chemical reactions, setting the "non-decomposability" standard for elements, revising chemical nomenclature in ways we still use, and much else. The 18th century history of chemistry is rife with causal reasoning but details would require nearly a course in chemistry, so I skip it. [12] One use of Newton's counterfactual principle with his rules of reasoning for identifying causes—applied by Newton to argue that the force that keeps the Moon in its orbit is the Earth's gravity-- is of special note. Lavoisier and Laplace confined a guinea pig for ten hours and measured the carbon dioxide ("fixed air"

[12] For that history, see A. Ihde, The History of Chemistry.

at the time) produced, and with their calorimeter measured the heat produced (measured as the weight of ice melted in the calorimeter.) They then burned a weighed amount of carbon and determined the amount of carbon dioxide generated. With these numbers they calculated the quantity of ice that would be melted in burning enough carbon to generate the amount of carbon dioxide the guinea pig generated in ten hours. The numbers were not close but close enough for them to conclude that respiration was a form of combustion—burning and respiration were instances of the same chemical process.

In chemistry, theory plaid havoc with experience. Supposing atoms combine to form stable molecules, what holds them together? Not heat, which drives things apart, certainly not gravity. The hooks and crannies of the 17th century were passe'.by the 18th century but two other forces seemed available: magnetism and electric charge. Magnetism, many of whose phenomena had been described in the 18th century by William Gilbert, was not plausible because if atoms had magnetic poles and aligned accordingly, all substances would be magnetic, which did not agree with experience. Electricity then? Static electricity had been known since forever. Coulomb's experiments in the 1780s argued that electrostatic force, whether attractive or repulsive is proportional to the inverse square of distance between the charged bodies and directly proportion to the product of their charges. So a common theory in the 19th century was that chemical bonding is entirely electrostatic (similar to what we would call ionic bonding). On the assumption that all atoms of the same kinds have the same charge, it follows that the separate atoms of the same kind in gases cannot combine: all elemental gases must be monoatomic. (That leaves a puzzle of course as to why the same is not

true of solids—why are there on this electro-atomic theory any macroscopic solid pure quantities of any element?) That assumption made the accurate reconstruction of relative atomic weights impossible and Guy-Lussac's law of combining volumes--at standard temperature and pressure, the ratios of volumes of reactant gases and product gases are simple whole numbers-- inexplicable. No surprise then, that John Dalton, who introduced something close to modern atomic theory in chemistry around 1810 and had done extensive experiments on the weights of gases, dismissed Gay-Lussac's data and conclusions. (Dalton's arguments were based chiefly on disagreement with Gay-Lussac's experimental numbers.)

Dalton's theory made chemistry a methodological mess. The theory is simple enough: atoms of all elements have the same weight. Pure compounds are made of molecules (or "compound atoms") each composed of a definite number of atoms of the component elements, the same numbers for all molecules. Proust's law, then known but controversial, follows: chemical reactants combine in definite proportions by weight. The macroscopic additivity of weights of reactants and products is explained by the supposed additivity of weights of their microscopic components and the conservation of mass, and so is the equality of the sum of weights of reactants and the sum of weights of products. But what were the weights of atoms? Dalton had an unfortunate addendum to his basic theory, the simplicity rules: If elements X and Y form a single compound, the molecules are XY; if they form two compounds, one is XY and the other is X_2Y or XY_2 and so on. From the empirical combining weights and the molecular formula, the (relative) atomic weights can be inferred—often wrongly using Dalton's rules.

The empirical combining weights of some pairwise combinations three elements—X,Y,Z contradicted the weights required by the rule.

At almost the same time that Dalton's theory appeared, Guy-Lussac published research arguing that gases combine in definite proportions, both by volume and by weight. In an appendix to his *New System of Chemical Philosophy*, Dalton disputed Gay-Lussac's experimental numbers and his conclusion. Avogadro soon published (1811) a more general account assuming that molecules of many substances in the gas state are dimers and that all gases at the same temperature and pressure have the same number of molecules. Referencing Gay-Lussac, he wrote:

"It must then be admitted that very simple relations also exist between the volumes of gaseous substances and the numbers of simple or compound molecules which form them. The first hypothesis to present itself in this connection, and apparently the only admissible one, is the supposition that the number of integral molecules in any gases is always the same for equal volumes, or always proportional to the volumes."[13]

Accepting Guy-Lussac's numbers, Avogadro compared estimates of molecular weights on his hypothesis (based on the weights of reactants and products) with those obtained with Dalton's, arguing that when they agreed it was because Dalton had made compensating errors, and arguing, as suggested above, that

[13] Avogadro, A. "Essay on a manner of determining the relative masses of the elementary molecules of bodies and the proportions in which they enter into these compounds." Alembic Club Reprints, no. 4.

there are series of chemical combinations that are flatly inconsistent with Dalton's rules.

Avogadro's essay is a cornucopia of examples and numbers that does not leave a clear impression. Gaudin soon published a clearer more systematic account of Avogardro's theory. His work was largely ignored.

There was no physical justification for Avogadro's hypothesis, or rather the debates about it turned on physical assumptions about Lavoisier's fictitious quantity, caloric. Dalton claimed all atoms and molecules in gases have the same quantity of caloric, making all particles with their caloric envelopes in gases of equal size. Avogadro suggested that different gases have different "condensations" of caloric, so that their volumes are the same and consequently "without the distances between the molecules varying; or, in other words, without the number of molecules contained in a given volume being different."

The (under) determination of atomic weights put Dalton's atomic theory in jeopardy. William Hide Wollaston, a prominent physicist of the day, said he could never endorse a theory whose fundamental quantities are indeterminate. By mid-19th century, Jean Baptiste Dumas, the most eminent French chemist of the day, said if he were "master" he would ban the word "atome" because "it goes beyond all experience and never in science should we go beyond experience." Apparently, he had never read or thought about the science of Kepler or Newton.

Dumas' attitude was realized in chemistry by the theory of chemical equivalents. Essentially, instead of estimating molecular formulas or relative atomic weights, chemists would simply record "equivalent weights"—the combining proportions of weights of reactants and weights of products in a chemical reaction of any type. For decades, politically astute chemists describing new reactions would give both conjectured molecular formulas and equivalent weights. They were enabled by a series of publications of tables of atomic weights, mostly from Berzelius. Berzelius' tables were guesswork, based on chemical analogies, similarities of crystallographic forms, and assumptions that are false—for example, that all metallic oxides are dioxides.

After the invention of the calorimeter by Laplace and Lavoisier in 1789, heat capacities and specific heats of materials could be measured. In 1817, Dulong and Petit used specific heats and one of Berzelius tables to announce a new law: the numerical product of specific heat and atomic weight is the same for all elements. Well, not quite. Some of Berzelius' atomic weights did not agree with that generalization. Where there were differences, Dulong and Petit simply substituted new atomic weights agreeing with their "law." It is no wonder that many chemists were skeptical of Dalton's "New System of Chemical Philosophy."

Various methods of estimating the weights of atoms were proposed in the first half of the 19th century. Their variety enhanced skepticism. By mid-century chemical nomenclature was a mess: atoms, compound atoms, elements, compounds, molecules, "oxygenated" compounds etc., and the fundamental issue of the time—whether relative weights of atoms could be consistently

estimated from empirical data—remained in doubt. In an attempt to resolve nomenclature, and so implicitly the physical ontology of chemicals, a congress was called at Karlsruhe, in Germany on the Rhine border with France, for 1860. 140 chemists attended, most from the UK, France and Germany, five from Russia, two from Poland, two from Italy and one from Mexico. Most of what happened there did not matter for the history of the atomic theory, but the attendance of two people did, Cannizzarro and Mendeleev. Stanislas Cannizarro was an Italian chemist from Genoa of no particular renown; Mendeleev was Russian, and likewise not famous among chemists.

Cannizzaro had looked through the history of measurements of the vapor density of elements and compounds. Assuming Avogadro's law, and setting the atomic weight of hydrogen gas at two, he estimated the molecular weights of a large collection of chemicals, finding the vapor densities of elements were multiples of half the weight of hydrogen and the vapor densities of compounds were always (nearly) sums of simple multiples of the estimated weights of their component elements. The results were not perfect, and there were anomalies, phosphorus for example (Phosphorus vaporizes as P_9). The results coincided (with some corrections) with the relations of atomic weights and specific heats of Dulong and Petit. Cannizzaro had copies of his essay on atomic weights printed and distributed to the attendees at Karlsruhe. Nine years later, using these weights and analogous chemical properties, Mendeleev published his periodic table, a version of which now hangs in every chemical laboratory.

Cannizzaro and Mendeleev convinced most physicists and chemists but not all. Ernst Mach and Wilhelm Ostwald remained prominent opponents of the atomic theory. Each of them preferred scientific explanation in terms of energy rather than particles. For Mach, "energy" was simply a shorthand for the facts about equivalences in quantities computed from heat, mechanics, and electricity. For Ostwald, it was something more, a quantity transferable from one system to another but always conserved, even constituting matter. The young Einstein applied for a research position with Ostwald and was refused despite a plea from Einstein's father. One can imagine that the history of physics might have been quite different had he been accepted.

Was There a Method?

With the exception of the guinea pig experiment, I can find no strategy like Kepler's or Newton's for the atomic theory in the 19th century better than Avogadro's: accepting Guy Lussac's conclusions for gases, the known definite proportions of combining weights of liquids and solids, and the assumption of corpuscular composition, and the many tests of simple proportions (in combination with Avogadro's hypothesis) that were implicit in Cannizzaro's analyses of vapor densities, it is hard to think of an alternative precise explanation. Several things seem remarkable and important:

1. Some of the argument turned on views about a hypothetical, and it turned out fictive, substance, caloric.
2. The measurements in chemistry of weights and volumes of gases were known to be inexact and expected to vary somewhat between

experiments. The conclusions drawn depended on which numbers were accepted.

3. Although by the middle of the 19th century error probability calculations had become common enough in astronomy, in chemistry there was no use of least squares or other statistical methods except occasionally an average.
4. The critical case was made by "ransacking" historical data. Novel predictions played no role in Cannizzaro's work, and none in Mendeleev's derivation of the periodic table. Mendeleev used his periodicities to make predictions about novel elements, some of them correct and some incorrect.

The Conservation of Energy

The notion of energy, without any quantitative measurements, had been around for a long time, sometimes associated with a quasi-mystical notion that energy, whatever it is, cannot be created or destroyed. In the 18th century, E'mile du Chatelet, a remarkable woman who wrote, among other things, on Newtonian physics, formulated kinetic energy as ½ mv^2, as we now do, and proposed that "energy" is a conserved quantity—although kinetic energy is of course not.

Experiments in the 18th and 19th centuries began to reveal remarkable associations. A definite quantity of mechanical work, measured by the heat produced by a falling definite weight falling from a definite height, was

always the same: same work, same heat. Conversely, heat (as in steam engines) produced work, in definite proportions. Even electrical current produced heat. Conversely, work could produce heat, for example through friction.

Energy became a central theoretical quantity in physics through the work of a Scottish engineer, William Rankine. Rankine defined "energy" in causal terms[14]:

(1.) IN this investigation the term *energy* is used to comprehend every affection of substances which constitutes or is commensurable with a power of producing change in opposition to resistance, and includes ordinary motion and mechanical power, chemical action, heat, light, electricity, magnetism, and all other powers, known or unknown, which are convertible or commensurable with these. All conceivable forms of energy may be distinguished into two kinds; actual or sensible, and potential or latent.

Actual energy is a measurable, transferable, and transformable affection of a substance, the presence of which causes the substance to tend to change its state in one or more respects; by the occurrence of which changes, actual energy disappears, and is replaced by

Potential energy, which is measured by the amount of a change in the condition of a substance, and that of the tendency or force whereby that change is produced (or, what is the same thing, of the resistance overcome in producing it), taken jointly.

If the change whereby potential energy has been developed be exactly reversed, then as the potential energy disappears, the actual energy which had previously disappeared is reproduced.

And

The law of the conservation of energy is already known, viz. that the sum of the actual and potential energies in the universe is unchangeable.

[14] William John Macquorn Rankine C.E. F.R.S.E. F.R.S.S.A. (1853) XVIII. On thegenerallawofthetransformationofenergy , The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 5:30, 106-117,

Of course, no such thing was known. The conservation of “energy” was simply a fiat by Rankine on behalf of a vague notion shared by the scientific and some of the literary communities. No one seems to have doubted it. “Potential energy” was essentially Rankine’s bookkeeping device to ensure that total energy is conserved. Rankine produced a differential equation which he claimed represented all possible transformations of energy. By judicious interpretations of its quantities and their derivatives, he was able to claim that known it reproduced known equivalences among the effects of electrical, mechanical and thermal processes. Nothing new was predicted. Energetics was the string theory of its day.

“Energetics” became a movement in speculative physics, an alternative to mechanical theories of fundamental processes. For some “energy” was just a calculating device; for others, it was a real property that Nature conserves through all processes in all of time.

Energetics was boosted by the development of Lagrangian mechanics which replaced Newton’s forces causing accelerations by the principle that motions minimize “action”--a function of kinetic and potential energies. Lagrangian physics was a source of what we now call field theory: a source causes a field potential distributed through space (or later, in general relativity, through spacetime) which causes motions of appropriate kinds of bodies located in the field.

Medicine and Epidemiology

Urbanizing European countries suffered a host of diseases in the 17th through the 19th centuries: smallpox, tetanus, typhus, pneumonia, cholera, puerperal fever, and more. Pneumonia was a common killer, smallpox was endemic, cholera outbreaks occurred in the 19th century in India, England, Germany and elsewhere. For new mothers childbirth was often a harbinger of death from infection—"puerperal fever" as it was known at them time. Where horses and their manure abounded, tetanus was sure to ride along. Henry David Thoreau's brother died of the disease, apparently contracted through a cut while shaving. Finding causes and preventions was a big deal in medical research. Much of the "scientific" work was back and forth debates citing cases, or generalizations from tiny samples.

Something like serious investigations of the causes of disease began in the 18th century. As with chemistry, distinguishing kinds—different diseases posed a problem. "Fever" was regarded as a disease, not a symptom, and various classifications were proposed, for example from the initial location of an outbreak or the name or location of whoever first described it (a practice we retain for many diseases, as in "German measles,""Spanish Flu" etc.), or from the location of an inflammation on the body. The standard "anti-phlogistic" treatment (as in removing heat—chemistry and medicine have long been entwined) was bloodletting on no better grounds than that Galen describes a case and some people whose blood was let have recovered. Francis Bacon had not got through to most medical practice.

The diseases that received quasi-epidemiological attention in the 18th century included pneumonia, smallpox, scurvy, dropsy (bloating), and "fever." There were lots of case collections from individual physicians and from diverse reports that conflated circumstances. Notably, some cases were collected from specific hospitals with selection of patients by objective categories which entailed some control on the sample. Inference was confounded by confounded diagnoses and treatments. A striking example is scurvy. Amazing at the time, James Cook circumnavigated the Earth without losing a single sailor (on his second trip around the globe, through arrogance he lost himself to an Hawaiian spear). His method was to require his crew to eat local vegetables and to give them regular drops of lemon juice. His achievement did not change practices in the British admiralty, partly because of a confused report from his ship's surgeon, and partly because his treatment was confounded with the requirement that sailors be given malt. Malt was the principal "medical" supplement in the navy to prevent scurvy, although it had no effect on that disorder. But malt did address vitamin B deficiency from which presumably sailors that returned ill to port often suffered. Even the recognition that lemon juice could help to prevent scurvy was foiled by its usual preparation for seafarers, as "rob"—lemon juice boiled with water--which assured that the ascorbic acid was destroyed.

The treatment for compound fractures-or worse from war—was amputation—without anesthesia—often followed by death. A notable opposition was expressed in a book by Johan Bilguer in 1761. Bilguer

recommended treatment of wounds by various medications, only one of which, wine, could have had any actual benefit, followed by amputation only if necessary. He claimed that of 6618 wounded soldiers treated by his methods in a medical hospital in the Seven Years War, 5,557 returned to fighting.

The study of medicines was, in retrospect, appalling, without systematic controls. Some medicines were appropriate—foxglove (digitalis) for dropsy—but others widely used—prussic acid (hydrogen cynanide)--not so much.

Physiology studies were pursued in the 18th century by vivisection and by poisoning animals to observed the symptoms preceding their deaths. Vivisection continued in the 19th century. Claude Bernard, who has been called one of the great men of science by a notable historian[15], was so avid a vivisectionist that his wife left him in disgust and became an anti-vivisectionist advocate.

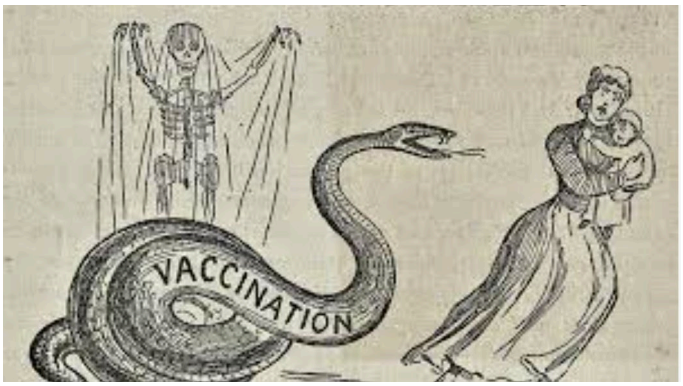
The disappointment of 18th century medical research is that sound methods were sometimes employed and published but not copied. Experiments were done with control groups (and very small samples), and in at least one hospital patients were assigned to alternative treatments in the order in which they came, and in at least one case an investigator "blinded" himself. There was a remarkable placebo experiment to test quack therapy, and at

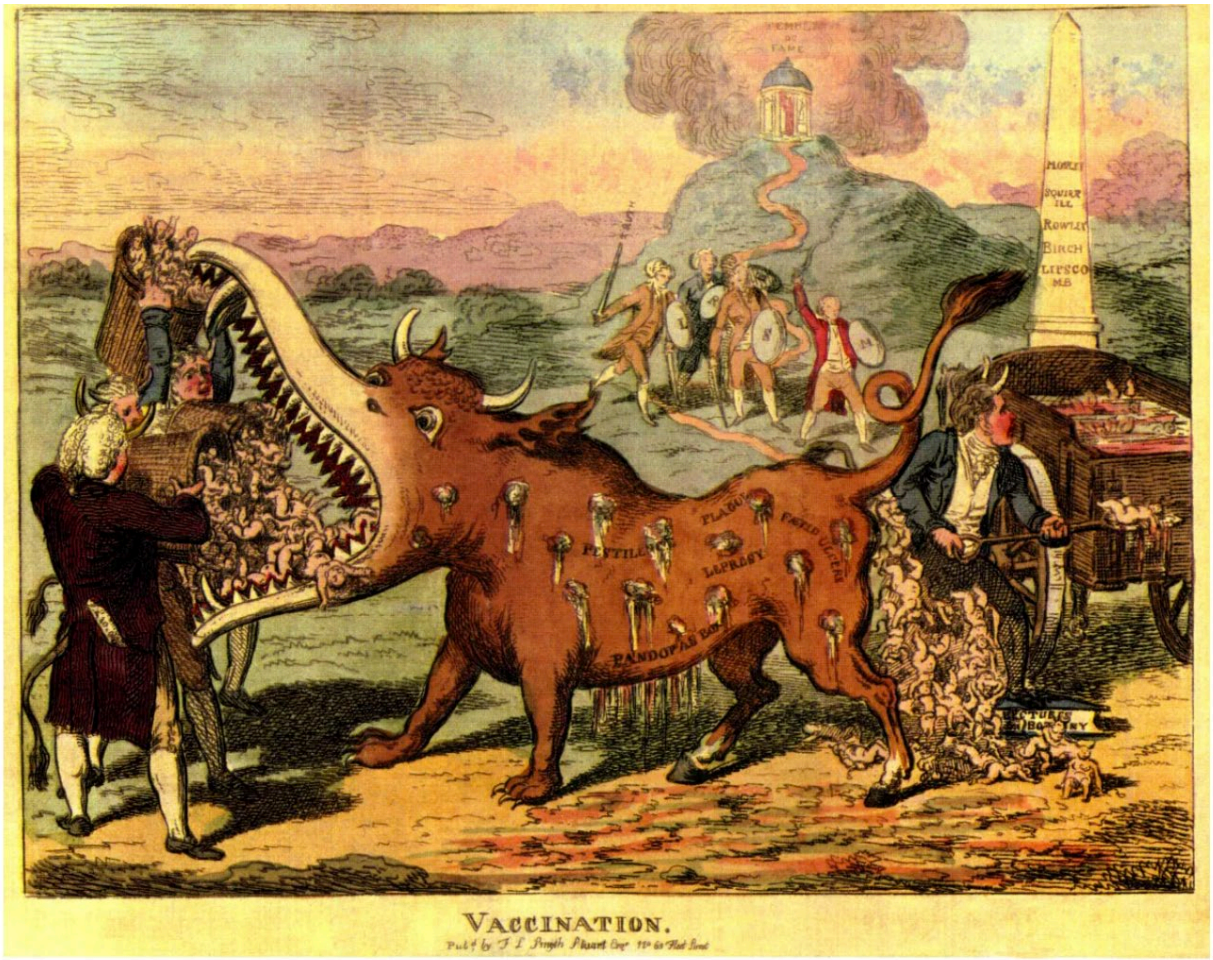
[15] I. B. Cohen, who was the premier historian of science in the middle of the last century.

least one effort at matching patients for alternative treatments. Rates rather than just absolute numbers were calculated. At least two researchers late in the century—mentioned above--attempted to compute the probability that differences in treatment outcomes were due to chance. These innovations did not spread much. The 19th century began with more experimentation but little improvement in methods of inquiry. We remember the lucky successes, not the failures.

Smallpox

Smallpox was endemic in the 17th, 18th and 19th centuries. The standard preventative treatment was "variolation," which was rubbing pus from an infected person onto an uninfected person. A frequent result was a new infection, but data were collected in the American colonies indicating a reduction in infection rate from .14 to .025 per cent of the population. Around 1796, Edward Jenner, an English physician who as a child had nearly died as a consequence of variolation, noted that a milkmaid infected with cowpox did not contract smallpox. He injected some of the pus from cowpox into the son of his gardener, who did not contract smallpox. After trials with a couple of dozen (the exact number is uncertain) subjects, including deliberately infecting at least one person with smallpox, Jenner published his results and conclusion: vaccination with cowpox prevented smallpox. Within a few decades, vaccination was required in several European nations, including England, in the face of considerable popular opposition illustrated in the following cartoons from the 19th century.


VACCINATION


PESTILENCE
PLAGUE
LEPROSY
PANDORA'S BOX
VACCINATION.

Here is what John Birch, the British physician most vociferously opposed to vaccination wrote in 1806[16]:

list, and what is the conclusion we are to draw? There is but one; namely, that Vaccination neither secures the patient from catching the Small-pox by variolous infection, nor when so caught, lessens the danger of disease. For my own part I tremble to think on the perils which await Society, from the prevalence of Vaccination. Unless it be stopped, we shall see Small-pox at no very distant period recur in all the terrors with which it was first surrounded; desolating cities like the plague, and sweeping thousands from the earth, who, lulled into a false security, will have fatally deprived themselves of the only proper means of defence.

[16] An examination of that part of the evidence relative to cow-pox, which was delivered to the Committee of the House of Commons / by two of the surgeons of St. Thomas's Hospital [J. Birch

Birch believed in case studies and opposed experimentation. So far as I know, there was no epidemiological survey of smallpox and vaccination before the 20th century. 19th century data are scattered and incomplete. This is not the last case of successful policy interventions based on what would now be regarded as insufficient evidence. Here is one testimonial:

"In1736 I lost one of my sons, a fine boy of four years old, by the small-pox, taken in the common way. I long regretted bitterly, and still regret that I had not given it to him by inoculation [had his son vaccinated]. This I mention for the sake of parents who omit that operation, on the supposition that they should never forgive themselves if a child died under it; my example showing that the regret may be the same either way, and that, therefore, the safer should be chosen." – Benjamin Franklin, *Autobiography*

Puerperal Fever

Puerperal fever—post-partum infections—was a common cause of death among women after childbirth. The connection with physicians and their behavior seems first to have been noted in publication by Alexander Gordon in 1795. Gordon's book influenced Oliver Wendell Holmes, Sr., who collected cases of post childbirth infection associated with particular physicians who had attended autopsies or previously treated infected women. Holmes published his argument and

---

and H. Cline]. To which is added, a letter to the author, from John Birch. https://iiif.wellcomecollection.org/pdf/b30381654. The American radical conservative movement of the 1950s and 1960s is named after a different John Birch.

conclusion repeatedly in the 1840s, to little effect. His views were vociferously opposed by Charles Miegs, a Philadelphia physician. Gentlemen, he maintained, have clean hands.

Another epidemiological case was made by Ignasz Semmelweis in the 1850s using data from the Vienna Hospital, where he eventually and briefly became director of the maternity wards. The birth ward clinics had notorious death rates, but they were not uniform. Admissions on some days were send to a ward of midwives, on other days a ward of physicians and students. The death rate from the latter was so famous that women made whatever effort they could to be admitted to the midwives', or second, clinic.

Table 1. Annual births, deaths, and mortality rates for all patients at the two clinics of the Vienna maternity hospital from 1841 to 1846.

	First Clinic			Second Clinic		
	Births	Deaths	Rate	Births	Deaths	Rate
1841	3036	237	7.7	2442	86	3.5
1842	3287	518	15.8	2659	202	7.5
1843	3060	274	8.9	2739	164	5.9
1844	3157	260	8.2	2956	68	2.3
1845	3492	241	6.8	3241	66	2.03
1846	4010	459	11.4	3754	105	2.7
Total	20 042	1989		17 791	691	
Avg.			9.92			3.38

Semmelweis collected data, shown above, on death rates in the two clinics. The notable difference in the two clinics was that only the physicians and medical students of Clinic 1 attended autopsies. Semmelweis suspected that their hands passed to mothers some infectious material acquired from autopsies. He recommended washing with chlorine water after attending autopsies, and once

he acquired the authority, he demanded it. Death rates fell and Semmelweis was fired. The doctors did not like washing their hands with chlorine water.

Semmelweis' argument was not a methodological breakthrough. As was done in other subjects he collected cases of outcomes under different treatment conditions and drew a causal conclusion. A difference was that his cases were in common circumstances in which the most obvious conceivable causal factors were the same. Francis Bacon would have approved.

Cholera

By the early 19th century, public health, especially among the poor, had become an issue of political concern in France and England, with special concern for sewage and water supply. Sewage was a major problem, not least from the numbers of dying horses in cities where the animals were the sole means of transporting large quantities of goods. Abundant anecdotes related disease to filthy water used for drinking or cooking. The result was "statistical" surveys of health conditions and water supplies and sewage, in France by Alexandre Parent du Chatelet (who also investigated health conditions in several other aspects, most famously prostitution in Paris) in 1824 and in Britain by Edwin Chadwick in 1842.

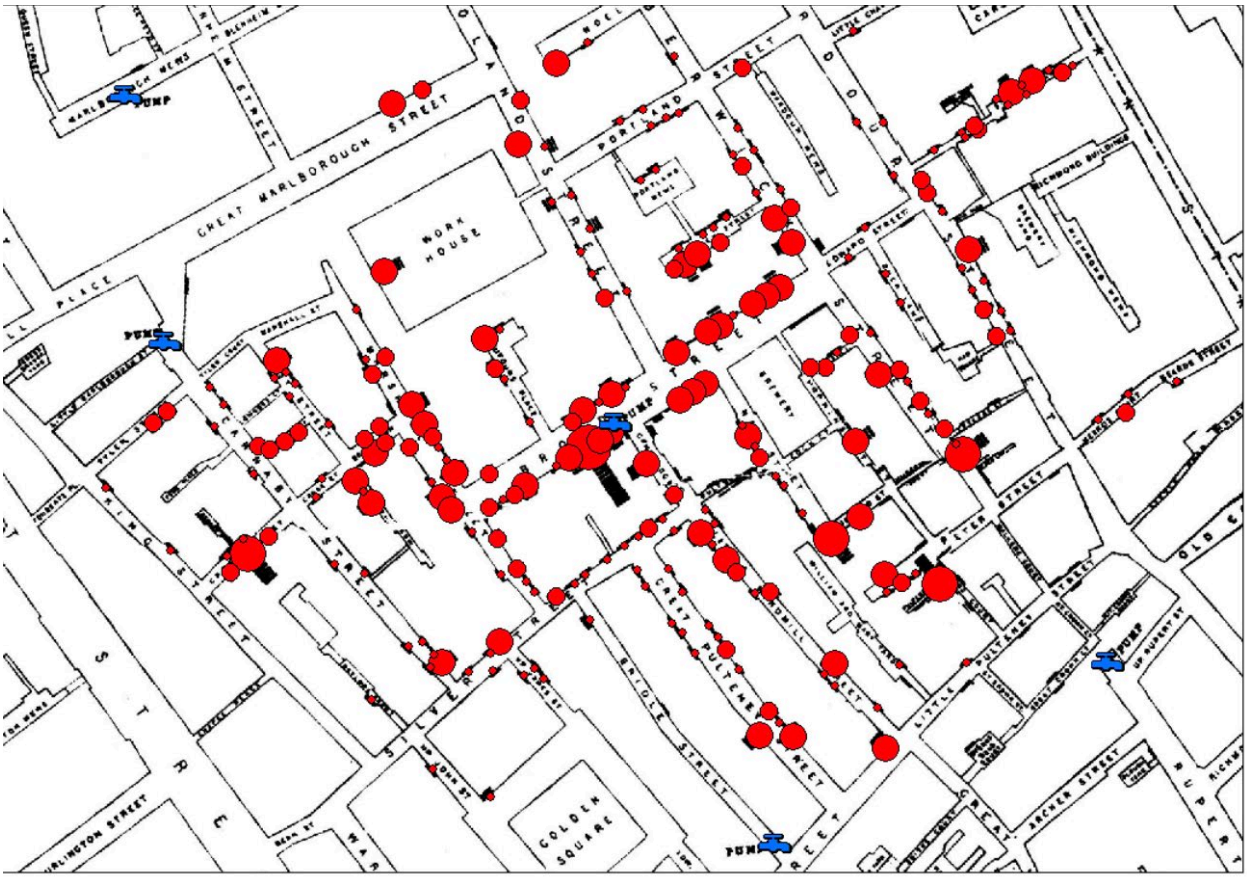
Chadwick, assigned the investigation by the government, was influenced by the doctrines of a British philosopher and legislator, Jeremy Bentham, who had advocated for public policies for "the greatest good for the greatest number", to

be measured by comparisons of sums of pleasures and pains. Previous reports had attributed disease among the poor to the living conditions and habits that accompanied poverty—for example, an aversion to cleanliness. These studies documented misery and attributed it to poverty but added little to methodology.

Cholera was a recurrent plague in Asia and Europe. In the 19th century it was the subject to two important investigations, one by John Snow which has been celebrated by recent popularists[17] and another by William Farr which is less recognized.

John Snow, a British physician, had some experience with cholera and a suspicion that it was associated with water consumed, when he undertook to investigate in detail a cholera outbreak in the Soho district of London in the 1850s. Snow had already published the theory that cholera is spread through water contaminated with feces, dead bodies (e.g., of horses) and other noxae. Snow investigated household by household, pub by pub, as to their sources of water. He found that occurrences of cholera were associated with water drawn from a pump on Broad street. Snow mapped the locations of occurrences, finding they tended to cluster around the Broad street pump and grew less frequent with distance from the pump. He submitted his report, which contained discussions of several other outbreaks; the government response was to remove the pump handle from the Broad street pump.

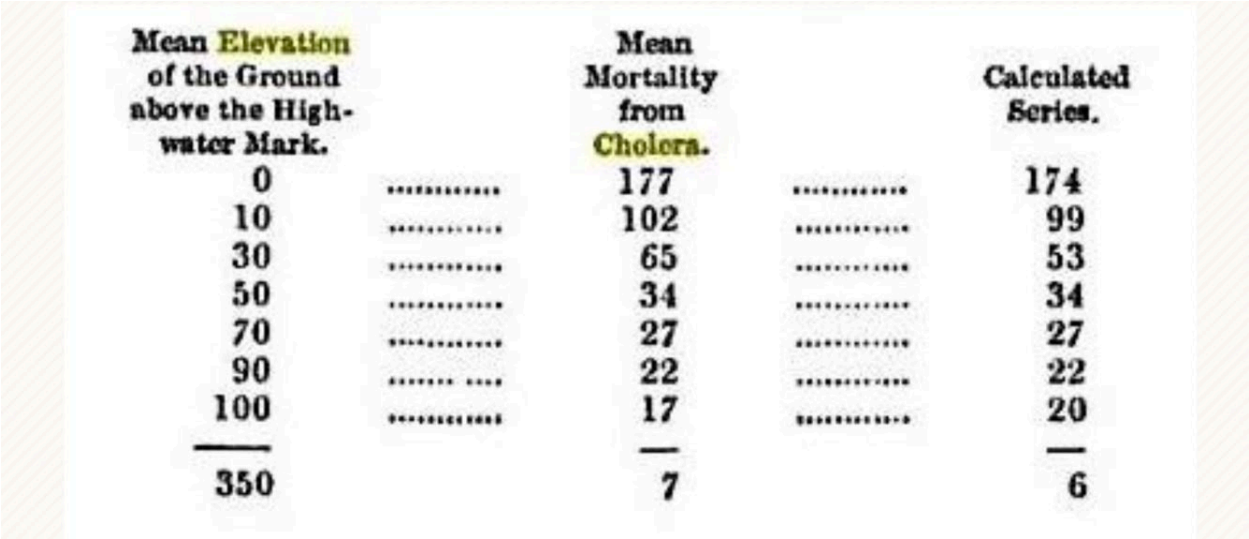
[17] For example, several papers by the late David Freedman celebrate Snow's work as a model for causal inference in epidemiology.



Snow's map of cholera occurrences in London. Blue circles are water pumps; the Broad Street pump is in the middle of the picture.

Almost as soon as it was published, Snow's book on the "Mode of Communication of Cholera" was subjected to devastating methodological criticism by E.A. Parkes, whose complaints were chiefly that Snow did not consider other possible factors (poverty, elevation, sewage lines, previous household flooding by sewage, etc.), in many cases relied on informal claims of residents about water sources and sewage connections, and, especially, included differences of numbers of cholera cases in different neighborhoods and circumstances but provided no basis for computing rates.

William Farr was an anesthesiologist and prominent “statistician” who undertook the study of cholera shortly after before Snow’s work. Farr undertook a study of the spread of disease (anthrax) among cattle and an outbreak of cholera in London in 1849 as well as several other cases. In the process he made a mathematical description of the frequency distribution of epidemics. Farr posited a nearly symmetric distribution, which he used to predict the peaks of epidemics. He attributed the symmetry to the diminution of the potency of the unknown causal factor as it was transmitted. The second novelty was the presentation of a rank order correlation—without the modern measure of that relation. The cholera and factor ordering were given for a variety of potential causes, notably elevation, poverty, and population density, but like Snow Farr failed to give death *rates*. Here is one of his tables. I have no idea what the numbers at the bottom of the columns signify.



Mean Elevation of the Ground above the High-water Mark.	Mean Mortality from Cholera.	Calculated Series.
0	177	174
10	102	99
30	65	53
50	34	34
70	27	27
90	22	22
100	17	20
350	7	6

In sum, the state of “field” of epidemiology through the majority of the 19th century was little, if any, different from what Francis Bacon would have

recommended. There was essentially no use of statistics and often a failure to estimate rates. The focus was on mechanisms of disease transmission—water or air—and concomitant social factors—poverty, elevation of residences, density of housing, etc. Implicit in these discussions was that unobserved sources of disease—sometimes called "viruses"-- were transmitted from person to person. More important advances in causal explanation of disease came from the search for these particles.

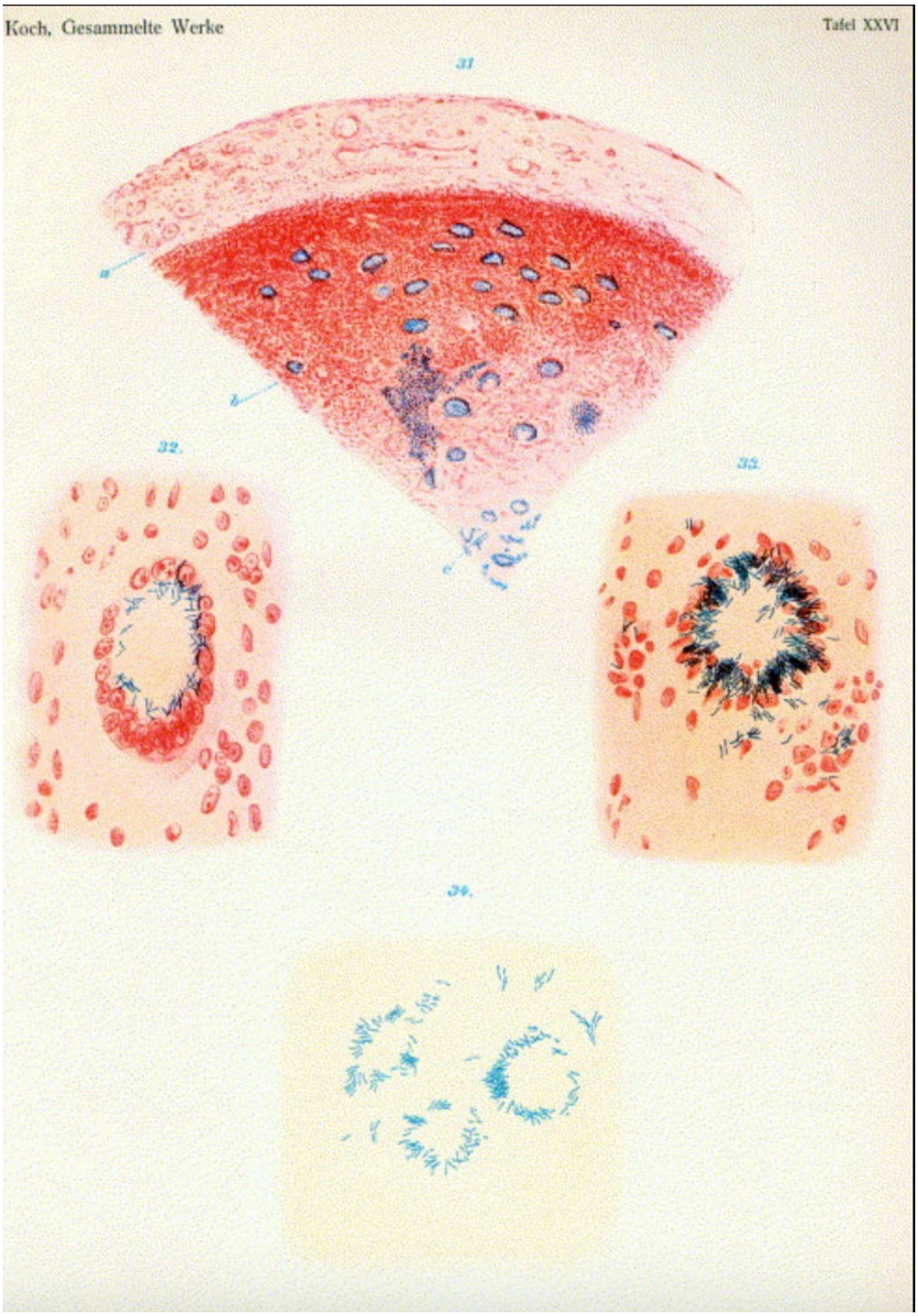
Criteria for Causation

The microscope, invented in the 16th century, presented epidemiologists with a tool for generating hypotheses about causation. Examining cases of cholera in India, Robert Koch found that distinguishable (by microscope) particles increased with progress of the disease, and eventually established that the microbe causes the disease. Experimenting on anthrax, and reading Pasteur's experiments arguing against "spontaneous generation" of living things and led Koch to formulate his four criteria for identifying the cause of a disease:

1. The microorganism must be found in abundance in all organisms suffering from the disease, but should not be found in healthy organisms.
2. The microorganism must be isolated from a diseased organism and grown in pure culture.
3. The cultured microorganism should cause disease when introduced into a healthy organism.

4. The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.

Koch realized that his criteria did not apply in all cases, that, for example people could be carriers of a disease without showing symptoms—i.e., could be healthy. Notwithstanding, his rules of thumb gained wide influence. Koch was among the first to photograph images through a microsope, with obvious improvement in demonstrating the presence of bacteria. He took images of tuberculosis bacteria in the course of identifying them as causes of the disease.



Koch went on to isolate the cholera bacillus and show by his criteria that it causes the disease. He claimed a “glycerol extract of pure culture of tubercle bacilli” would cure tuberculosis. Unfortunately, his good results with guinea pigs did not

extend to humans. The medical community had learned the misery of experimentation: what works with one population need not work with others.

Lister, Statistics, Causality and Reproducibility

Joseph Lister, a Scottish surgeon, introduced antiseptic surgery in the 1850s. Lister's method was to soak all surgical tools and clothing in carbolic acid, and for good measure to spray the air in the operating room with carbolic acid. Lister reported statistics comparing deaths (mostly from amputations) before and after the introduction of his methods. His principal form of argument, however, was by example, describing case after case in great detail. He also favored giving surgeries with an audience.

Lister's methods were repeatedly challenged. It was claimed that simple cleanliness produced better results. In a reversal of Bacon's third list, it was claimed that as ever more dilute solutions of carbolic acid were used, survival rates improved. It was objected, reasonably, that Lister would only publish a selection from his hospital outcomes and that a longer history might show periods in which surgical survival without Lister's methods was as good or better than with his. Since survival from surgery might be influenced by fluctuating periods of disease, that complaint was not unreasonable. There were proposals for controlled experiments but none seem to have been carried out at the time.

Statistical Developments

At Cambridge in 1833 the stated aim of the new statistical section was to collect 'facts relating to communities of men which are capable of being expressed by numbers, and which promise when sufficiently multiplied, to indicate general laws'.[18]

But in the course of the century it became apparent that collecting facts and numbering them was well insufficient to identify causes and effective policies. Missing was any reliable method to infer causes from collections of facts. Homeopathy, for example, began in Europe and spread to America, supposedly warranted by Baconian numbers. In 1855, in Louisiana, Holcome and David compared their homeopathic treatments for yellow fever with outcomes of "allopathic"—conventional medicine. 5.4 % deaths versus as against 25%. In Massachusetts doctors reported even more striking differences.

The philosophers were not wanting with tomes of advice and analysis. John Stuart Mill's *Methods of Logic* was the mid-century manual of scientific method, but it was (and is) little more than a rehash of Bacon pretending originality. (Mill did the same in ethics, rehashing and pretending originality from Jeremy Bentham.) Whewell, noted as an historian of science, offered only the vague advice of "consilience" of inductions. (Whewell posited Aristotlelion final causes). William Herschel, who had made serious contributions to astronomy, offered a restatement of a medieval principle:

[18] Goldman, Lawrence. *Victorians and Numbers*

'If the analogy of two phenomena be very close and striking, while, at the same time, the cause of one is very obvious, it becomes scarcely possible to refuse to admit the action of an analogous cause in the other though not so obvious in itself.' (1831)

German philosophy was much worse : Hegel turned Newtonian physics into obscure hash, claimed the whole of Newtonian gravitational theory was contained in Kepler and announced that, necessarily, there could not be more than six planets.

Darwin and the Hidden Hand

"Hidden Hand" explanations—processes that result in patterns or features from the action of multiple more or less independent causes without any design or intent—were introduced by Adam Smith in the 18th century in an argument for free markets in *The Wealth of Nations*. I doubt Charles Darwin ever read Smith (who was one of the worst writers of English I have ever read) but he likely knew of the doctrine. Darwin was prompted by several factors. One was his reading of Thomas Malthus, the 18th century writer who had argued that misery and disaster from overpopulation is inevitable because populations grow exponentially while the land needed to sustain them grows only quadratically. Malthus' verbose, intensely pessimistic essay On the *Principle of Population* criticized the most influential utopian theorists of his day for naivete', want of evidence, and ignoring the eventual effects of unchecked population growth.

To Darwin, the moral was that in every species more progeny are produced than can survive long enough to reproduce themselves. Varying circumstances, or what Darwin called “chance.” determine which survive to reproduce and which do not. Another source of Darwin’s views, and his evidence, was the history of domestic animal breeding and his own experience breeding pigeons. Deliberate selection causes desired forms to appear and be sustained: small dogs, floppy eared dogs, pigeons of a variety of sizes and colors, and so on. (Alas, no breeding led from dogs to cats), the ever-growing discovery of fossil life forms,

The final trigger for Darwin seems to have been his observations in the Gallapagos islands in the course of a British “voyage of exploration.” He noted the variations in beaks of finches on different islands which had plausibly transferred from a common source and thereafter remained relatively isolated populations. He noted the physical analogies between some species in Africa and South America (the latter he rode through on horseback). Another of his sources was the “uniformitarian” geology of James Hutton, who argued that the Earth had

changed slowly through processes that were still active. But fundamentally, Darwin's argument was an instance of Newton's rules of reasoning: the chance production of variation in progeny, the advantages of progeny that had developed novel characteristics that made them more likely to survive, and the inheritance of such characteristics, explained the variety of forms of life, the existence of fossils of extinct species. That, and the fact that deliberate breeding could dramatically change the forms of animals, offered a true cause of the variety of species and the only natural cause available.

As with the atomic theory in chemistry, Darwinian evolution was disturbed by erroneous physics, most prominently by Lord Kelvin, the most eminent physicist of the latter half of the 19th century. Kelvin estimated the age of the Earth to be at most 100 million years, assuming the energy of the sun was from gravitational pressure on its interior. Darwin's only response was the evolution must have happened faster than he had previously thought. Kelvin made a different error important in the history of geology, estimating the Earth was at most 400 million years old (Kelvin used empirical estimates of heat flow but assumed the Earth is solid.)

I will say little about physics because the issues in thermodynamics, electricity and magnetism and their histories are complicated without much innovation in methodology as distinct from experimental technique. Electricity and magnetism is largely a history of trial and error, improved measurement accuracy, and opportunistic recognition of effects—induced charge, electric conduction, electrical properties of materials, etc.—the identification of phenomena as of the

same kind (lighting and electrostatic sparks for example), the formation of a variety of laws from Ohm, Henry, Faraday and others—and their synthesis into a mathematical theory by Maxwell and Heaviside in the mid 19th century. Investigation of what we now call electrostatics prompted theories of an electric fluid emanating from charged bodies and speculations about its properties.

Amidst the many speculations about what entities and processes produce electrical and magnetic phenomena, one stands out. In 1846 "Thoughts on Ray Vibrations" Faraday speculated that all of the physical world is made up of centers of force and the "lines of force" they generate—magnetic, electric and gravitational. He said nothing about mass. Presciently, he speculated that light is transverse vibrations of electric lines of force and that vibrations of all sources of force produce waves in their lines of force—hence in principle magnetic waves and gravitational waves.

"The view which I am so bold as to put forth considers, therefore, radiation as a high species of vibration in the lines of force which are known to connect particles and also masses of matter together. It endeavours to dismiss the aether, but not the vibrations."

This aspect of his conjecture was required by the Special Theory of Relativity sixty years later. Arguably, he could have dealt with inertial mass by identifying it with the mass in Newton's law of gravitation, i.e., what Einstein called "gravitational mass."

So, throughout most of the 19th century we find a dramatic increase in “simple” experimentation in the physical sciences, albeit in increasingly complex circumstances, mathematical formula to capture empirical laws, a few broad generalizations as general mathematical theories (e.g. electrodynamics, thermodynamics, atomic weights), wide use of epidemiological data with little or no use of available statistical methods, improved but generally neglected experimental methodology employing controls, blinding, and large, homogenous samples. The role of statistics in scientific discovery began to change in the latter part of the 19th century, in consequence of the peculiar ingenuity of Francis Galton.

In 1890, Galton might have qualified as the most interesting man in the world. Wealthy by inheritance, he traveled through Africa. Apparently struck by the differences in physical features of Africans and Europeans, upon returning he created a laboratory for the study of physical anthropology and hired an able mathematician, Karl Pearson. Among its products were fingerprinting, eugenics, the correlation coefficient and the idea of a regression line, which in modern view is simply a least squares lines, but for Galton meant a natural phenomenon (he originally called it “reversion”) in which the mean of the distribution of features among grandchildren was closer to the grandparental mean than to the parental mean although both were Normally distributed. Galton gave evidence for his claims through a number of experiments on peas in the 1870s. Pearson provided the formal regression coefficient. More precise and informative work on inheritance had been done by Gregor Mendel and published in 1866 (" Experiments on Plant Hybridization " German: "Versuche über Pflanzen-

Hybriden") unfortunately in the *Proceedings of the Natural History Society of Brünn* . Mendel's work, and his "law of independent assortment" of traits was essentially unknown until the early 20th century.

Randomization

While there were many attempts to match treatment groups, and one (noted above) assigning patients by the order in which they came, randomization occurred only on occasions from the 17th century to the 20th. Random treatment assignments in experiments seem first to have been proposed by Jan Baptiste van Helmont, a Belgian chemist, early in the 17th century. It is not clear whether he carried out the procedure. Mesmer, in 1781, proposed a randomized test of his theories but so far as is known, it did not happen—Benjamin Franklin did the proposed experiment without randomization. A mechanically randomized experiment on homeopathy was carried out in 1835 (it found no homoepathic effect). Charles Sanders Pierce, the rather eccentric philosopher and mathematician, and Jastrow, a psychologist, introduced both randomization and blinding in a psychological experiment in 1885. They randomized treatments, not subjects, and blinded subjects to the treatment received.

In the early decades of the 20th century R.A. Fisher essentially created frequentist statistics, emphasizing randomization and experimental design, hypothesis testing, estimation procedures, likelihood, and introducing a variety of techniques still used: anova, linear discriminant analysis, and one that is not much used

today, his “pivot” method for obtaining probabilities of hypotheses, an alternative to Bayesian methods.

In Fisher’s view, establishing and estimating causal relations was the chief point of science, but he was cagey about both about causality and probability. Fisher was trained in astronomy, worked in a biological research center and wrote an influential work on evolution. But when in the 1950s and 1960s epidemiological research seemed to show an influence of smoking on cancer, Fisher, a lifelong smoker, was adamant that only randomized experiments could identify causal relations. As to the meaning of probability, Fisher never says—his section on “The Meaning of Probability” is about methods of estimation, not meanings. He says only in a footnote that the estimates are “justified” by limiting frequencies. Fisher was a mathematician, not a philosopher, and it is no surprise that he gives no account of how limiting frequencies justify inferences from finite samples.

Unobserved Common Causes and Marginal Constraints

The late 19th and early 20th centuries saw the introduction of intelligence tests The early tests, involving both verbal and physical motor measurements, were eagerly taken up by several psychologists, most notably by Charles Spearman. Spearman had two fundamental ideas. First that test scores have a single common cause which he called “general intelligence” and denoted by g. And second, that the existence of g could be established by patterns of correlations of intelligence scores and once established, estimated.

Spearman's argument for the existence of a single common cause was his claim that intelligence test correlations satisfy vanishing tetrad constraints on the correlations, rijrkl = rikrkj =rilrkj for four variables, i, j, k, l , which he claimed are characteristic of a single common cause. Spearman and others proposed statistical tests for single tetrad equations; it was not until the year 2000 that a statistical test for entire sets of tetrad equations was offered by Bollen and Ting. Spearman did not seem to realize that vanishing tetrad equations can be entailed by models with no unmeasured common causes; Bollen and Ting had the same oversight. The distinguishing feature is that models with a single common cause imply no vanishing partial correlations among the measured variables.

Spearman's method attracted followers in the World War I era, most notably Truman Kelley, a Stanford psychologist. In 1928, the same year Spearman published his fullest statement of his theory of intelligence, Kelley published the perplexingly titled *Crossroads in the Mind of Man*. Kelly had a rather vague theory that the brain has linkages ("Crossroads") that somehow generate linear common cause structures. The notable innovation was realizing that two linear common causes imply constraints among triples of correlations among five measured variables.

Kelley and other followers of Spearman found that single common cause models often did not fit their data and attempted to adjust the framework by introducing further unmeasured common causes of subsets of variables. Tetrad methods suffered from the computational demands of assessing the vast number of possible tetrad constraints among the many variables in intelligence tests.

Guilford's influential psychometric monograph gave that as the major reason for abandoning Speaman's methods in favor of factor analysis, introduced in the 1930s by Thurstone. Thurstone's method amounted to introducing a single common unmeasured cause, estimating the linear coefficients, calculating the residual correlations, and repeating until the residual correlations were acceptably small. The relations among the unmeasured variables were essentially arbitrary up to a linear transformation. Thurstone preferred a simplicity rule, effectively whatever linear transformation minimizes the number of edges (in a graphical representation which Thurstone did not give.). Aside from linear regression, Thurstone's was the first automated search procedure for causal models. Except it wasn't. Thurstone said his procedure was a method of "data reduction" and said nothing about causal interpretations—except that his "general intelligence" was supposed to be a cause of IQ test responses.

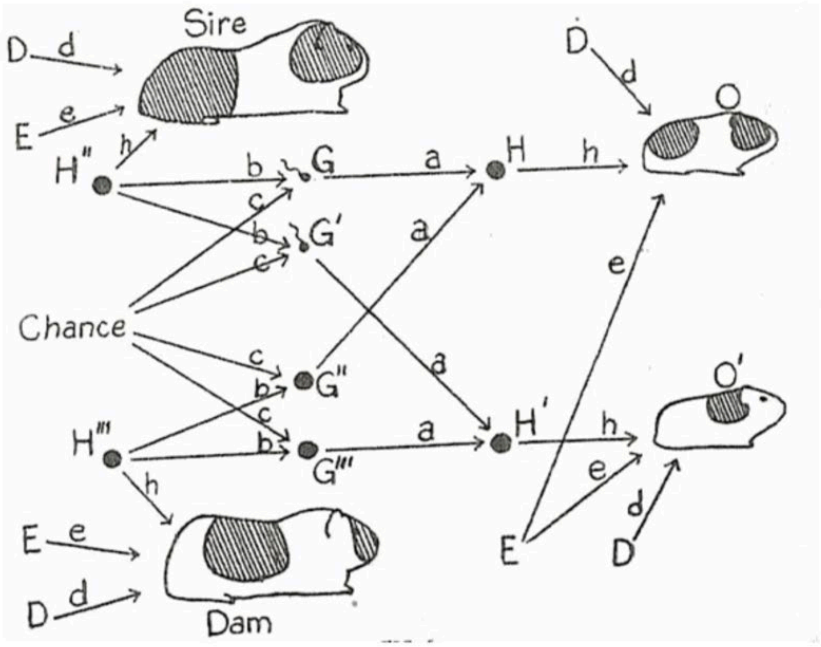
In opposition to psychometrics, John Watson and B.F. Skinner promoted behaviorism, the idea that psychology should confine itself to describing regularities between stimulus inputs to an organism and its behavior in response. Skinner at least seems to have been motivated by the dominance of the neural network model of brain function developed in the late 19th century by several physiogists and psychiatrists following the identification of brain nerve cells. Skinner's conclusion was that whatever was happening in the brain between sensory input and motor output is far too complex to be uncovered by anything that could be observed about human or animal behavior, and it would be unscientific and unfruitful to speculate. In particular, traditional psychological variables denoted nothing at all.

Path Analysis: Sewall Wright

In 1918, and in many publications thereafter, Sewall Wright introduced a representation of causal relations as directed acyclic graphs whose edges represented direct effects associated with (standardized) linear coefficients. Wright's work introduced a number of ideas and results that remain important. His aim was to provide a representation of causal ideas in linear models that would allow calculation of correlations between remote variables, so that established correlations could be pieced together into a coherent causal model and inferences made from the model. To that end he showed that his representation satisfied what are now called the "trek rules"": the correlation between two variables is the sum, over all directed paths between them and all pairs of paths from a common cause, of the products of the path coefficients on the several paths and pairs of paths. Notably, Wright points out that two independent variables will be dependent conditional on their common causal effect. This is the first statement I can find of what is now known as the "collider principle." Further, Wright described the feature we now call violation of faithfulness through canceling pathways. Wright was unsure how to estimate path coefficients. After some struggles over the years he eventually settled on multiple regression. There is no discussion of interventions or unobserved common causes in his 1918 paper—I have not checked his later publications.

Wright's first paper was severely criticized (with little comprehension) in 1920 by Henry Niles, which should serve as evidence that it is not just philosophers who cannot read. Here is Nile's complaint in a nutshell.

"There is no philosophical basis for giving it a wider meaning than partial or absolute association. In no case has it been proved that there is an inherent necessity in the laws of nature. Causation is correlation.
"Statistical methods, particularly multiple correlation, indicate causes when they are used with common sense and upon the data of critical experiments. But the method of path coefficients does not aid us because of the following three fallacies that appear to vitiate this theory. These are (1) the assumption that a correct system of the action of the variables upon each other can be set up from a priori knowledge; (2) the idea that causation implies an inherently necessary connection between things, or that in some other way it differs from correlation; (3) the necessity of breaking off the chain of causes at some comparatively near finite point."



Wright's first path model—on genetic inheritance.

In the first paragraph quoted above, Niles was following a view—causation is correlation—exposited by Karl Pearson in 1911 in his *The Grammar of Science*, a particularly incoherent book in which Pearson claims that *everything material* is ideas in our *brainss*. Wright's reply is in essence that Niles did not read and did not think. (The same can be said for a more recent defense of Niles' complaints.[19])

"..any careful reader of NILES'S own paper would immediately see the wide difference between what NILES says of his [Wright's] purpose and what it really is.

[19] I refer to an atrociously confused and uninformed essay by two psychologists: Denis, D., & Legerski, J. (2006). Causal modeling and the origins of path analysis. *Theory & Science*, *7*(1), 2-10.

Wright does not explicitly bring up counterfactual or intervention objections to Nile's "causation is correlation and anything more is unknowable" claim. But he does point out that there are lots and lots of mathematically described causal relations in established physics and chemistry. The most telling point Wright makes about Niles discussion (which the paper cited below claims is mathematically correct) is that in presenting alleged counterexamples Niles uses undirected graphs.

Wright emphasized that path analysis was not a method of inferring causal connections from sample data, but evidently he thought about the problem. In 1925 (in an essay on corn and hog price cycles) he introduced instrumental variable methods: find a cause X of Y that influences Z only through Y and use the trivial decomposition: rXZ = rXYrYZ to estimate the effect of Y on Z. In Wright's example, Y was exogenous to Z and so the method was strictly unnecessary. In 1928 in an appendix to his book *The Tariff on Animal and Vegetable Oils*. Sewall Wright's son, Philip Wright, extended the idea to its present use, to estimate the effect of Y on Z when there may be unmeasured common causes of Y and Z. In both works it was the assumed systems are linear.

Gilbert Walker and Climate Correlations

Gilbert Walker was a British professor of mathematics early in the 20th century. In 1910 he was appointed to a position in the government British ruled India—the Raj. The recurring problem in India was famine occasioned by the failure of annual rainstorms, the monsoons. Walker set about trying to use the new correlation

methods to identify the cause of monsoon failure. Thanks to the fact that much of the world was then part of the British empire, he was able to acquire time series of weather data for some years from many parts of the globe. His problem as he saw it was to determine which weather features, where, influenced the weather in India. With complicated sets of correlations, and multiple variables using Normal distribution assumptions for correlations (the correct sampling statistic for multiple correlations was not found until 1928 by Wishart), and applying hypothesis tests, Walker found a correlation between temperature pressure oscillations in the southern Pacific ocean and southeast Asia. He concluded that the monsoons, and their failures, are due to the southern oscillation—the high surface pressure vacillation later came to be referred to as ENSO.

Structural Equation Models

Perhaps the first structural equation model was proposed by a Danish statistician, .T. N. Thieles, in 1888. His presentation is verbal without explicit equations. Spearman's model was effectively a structural equation model with a single latent variable and individual "factors" for each measured variable—what eventually became "noise," "disturbance" or "error" terms. The very idea was developed and explored in the 1940s and 1950s by the Cowles Commission, whose members were economists several of whom later won Nobel prizes in economics. The commission members were chiefly concerned with methods for estimating parameters in models and with the interpretation of the error terms. "Error" terms are ambiguous. They can mean mismeasurement of a variable so that the measured value is a direct effect of the true value and unknown measurement

errors; they can mean unknown causes of the true value; they can mean unknown common causes of multiple variables. The first two were distinguished as “errors in variables” versus “errors in equations; the last was a befuddlement. A main concern was with time series, including how many lags to include. Correlations from time series had been insightfully explored by George Udny Yule early in the 20th century, notably the correlation of two jointly monotonically increasing or decreasing time series. The Cowles commission, consisting mostly of economists, was interested in time series with business cycles. The Cowles Commission also considered “simultaneous equation models” representing feedback systems. Two things were sorely missed in the Cowles’ literature: a methodology for finding structural equation models and a clear connection with using the models to predict the effects of policy interventions. Trygve Haavelmo, for example, distinguished one set of equations as “structural” and algebraically equivalent forms as not, without clearly explaining why. His only consideration of interventions was adding a term to the causal side of a structural equation to represent a government policy such as a tax.

Developments in the years following world war 2 included “full information” estimators for latent variable models and methods of model search due to the Norwegian statistician, Frisch. One of Frisch’s ideas was to suppose that one or more input variables were known, and try to eliminate in each dependent variables all variation due to other variables via regression. The result was supposed to capture the “true” noise variance of each dependent variable, which could be eliminated leaving a deterministic system whose linear coefficients could be estimated. Unsurprisingly, the results were not unique. Frisch introduced the

problem of multicollinearity and emphasized the pitfalls of using regression in simultaneous equation models representing feedback. Frisch, as with others, avoided writing of regression coefficients that measure causal dependencies, instead he wrote of "meaningful" coefficients. While he was concerned as an economist with policy, the causal concepts involved were not explicit and not explicitly connected with the "structural" equations—"structural" became a standard euphemism for "causal."

An unrelated development was Wold's proposal of "partial least squares" whose basic idea was to project correlated regressors onto uncorrelated variables and use those in regression. Partial least squares is still used but has largely been replaced by other methods—ridge regression, principal components and kernel regression. No case was made that partial least squares captures causal relations.

There were several innovations in the late 1960s and 1970s. In 1969, Clive Granger proposed estimating causal effects in time series by linear regression, treating each case as a tuple of values of variables and their lags. "Granger causality" as it was widely called (and by some allergic to any causal inference, "Granger Non-Causality") has been widely used. Since it was recognized by users (and by anyone who read Frisch) that the partial regression coefficient for a regressor could vary with whatever other regressors were used in a multiple regression, the problem of variable selection became a major issue. In the 1970s many issues of the journal *Technometrics* were devoted to the problem. A host of heuristics were proposed, none of them with proofs that they converged to the correct answers—largely because the statistical community at the time was loath

to talk about cause and effect, especially not when dealing with observational data.

Smoking and Lung Cancer

The confused state of empirical causal inference in the 1960s is illustrated by the debates over the effects of smoking. I excerpt a long section from *Causation, Prediction and Search*:

In the 1950s a retrospective study by Doll and Hill found a strong correlation between cigarette smoking and lung cancer. That initial research prompted a number of other studies, both retrospective and prospective, in the United States, the United Kingdom, and soon after in other nations, all of which found strong correlations between cigarette smoking and lung cancer, and more generally between cigarette smoking and cancer and between cigarette smoking and mortality. The correlations prompted health activists and some of the medical press to conclude that cigarette smoking causes death, cancer, and most particularly, lung cancer. Sir Ronald Fisher took very strong exception to the inference, preferring a theory in which smoking behavior and lung cancer are causally connected only through genetics. Fisher wrote letters, essays, and eventually a book against the inference from the statistical dependencies to the causal conclusion. Jerzy Neyman criticized the evidence from retrospective studies. The heavyweights of the statistical profession were thus allied against the methods of the medical community. A review of the evidence containing a response to Fisher and Neyman was published in 1959 by Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin, and Wynder. The Cornfield paper became part of the blueprint for the Report of the Surgeon General on Smoking and Health in

1964, which effectively established that as a political fact smoking would be treated as an unconfounded cause of lung cancer, and set in motion a public health campaign that is with us still. Brownlee (1965) reviewed the 1964 report in the Journal of the American Statistical Association and rejected its arguments as statistically unsound for many of the reasons one can imagine Fisher would have given. In 1979, the Surgeon General published a second report on smoking and health, repeating the arguments of the first report but with more extensive data, but offering no serious response to Brownlee's criticisms. The report made strong claims from the evidence, in particular that cigarette smoking was the largest preventable cause of death in the United States. The foreword to the report, by Joseph Califano, was downright vicious, and claimed that any criticism of the conclusions of the report was an attack on science itself. That did not stop P. Burch (1983), a physicist turned theoretical biologist turned statistician, from publishing a lengthy criticism of the second report, again on grounds that were detailed extensions of Fisher's criticisms, but buttressed as well by the first reports of randomized clinical trials of the effects of smoking intervention, all of which were either null or actually suggested that intervention programs aimed to cause people to cease or decrease smoking actually increased mortality. Burch's remarks brought a reply by A. Lilienfeld (1983), which began and ended with an ad hominem attack on Burch.

Fisher's criticisms were directed against the claim that uncontrolled observations of a correlation between smoking and cancer, no matter whether retrospective or prospective, provided evidence that smoking causes lung cancer, as against the alternative hypothesis that there are one or more common causes of smoking and

lung cancer. His strong views can be understood in the light of features of his career. Fisher had been largely responsible for the introduction of randomized experimental designs, one of the very points of which was to obtain statistical dependencies between a hypothetical cause and effect that could not be explained by the action of unmeasured common causes. Another point of randomization was to ensure a well-defined distribution for tests of hypotheses, something Fisher may have doubted was available in observational studies. Throughout his adult life Fisher's research interests had been in heredity, and he had been a strong advocate of the eugenics movement. He was therefore disposed to believe in genetic causes of very detailed features of human behavior and disease. Fisher thought a likely explanation of the correlation of lung cancer and smoking was that a substantial fraction of the population had a genetic predisposition both to smoke and to get lung cancer. One of Fisher's fundamental criticisms of these epidemiological arguments was that correlation underdetermines causation: besides smoking causing cancer, wrote Fisher "there are two classes of alternative theories which any statistical association, observed without the precautions of a definite experiment, always allows— namely, (1) that the supposed effect is really the cause, or in this case that incipient cancer, or a precancerous condition with chronic inflammation, is a factor in inducing the smoking of cigarettes, or (2) that cigarette smoking and lung cancer, though not mutually causative, are both influenced by a common cause, in this case the individual genotype." Not even Fisher took (1) seriously. To these must be added others Fisher did not mention, for example that smoking and lung cancer have several distinct unmeasured common causes, or that while smoking causes cancer, something unmeasured also causes both smoking and cancer. If we

interpret “statistical association” as statistical dependence, Fisher’s contention was that given observation only of a statistical dependence between smoking and lung cancer in an uncontrolled study, the possibility that smoking does not cause lung cancer cannot be ruled out. By the 1960s a number of personal and social factors associated with smoking had been identified, and several causes of lung cancer (principally associated with occupational hazards and radiation) potentially independent of smoking had been identified, but their potential bearing on questions of common causes of smoking and lung cancer seems to have gone unnoticed. The more difficult cases to distinguish are the hypotheses that smoking is an unconfounded cause of lung cancer versus the joint hypotheses that smoking causes cancer and that there is also an unmeasured common cause—or causes—of smoking and cancer. Fisher’s hypothesis that genotype causes both smoking behavior and cancer was speculative, but it wasn’t a will-o-the-wisp. Fisher obtained evidence that the smoking behavior of monozygotic twins was more alike than the smoking behavior of dizygotic twins. As his critics pointed out, the fact could be explained on the supposition that monozygotic twins are more encouraged by everyone about them to do things alike than are dizygotic twins, but Fisher was surely correct that it could also be explained by a genetic disposition to smoke. On the other side, Fisher could refer to evidence that some forms of cancer have genetic causes. The paper by Cornfield et al. (including Lilienfeld) argued that while lung cancer may well have other causes besides, cigarette smoking causes lung cancer. This view had already been announced by official study groups in the United States and Great Britain. Cornfield’s paper is of more scientific interest than the Surgeon General’s report five years later, in part because the former is not primarily a political document.

Cornfield et al. claimed the existing data showed several things: 1. Carcinomas of the lung found at autopsy had systematically increased since 1900, although different studies gave different rates of increase. Lung cancers are found to increase monotonically with the amount of cigarette smoking and to be higher in current than in former cigarette smokers. In large prospective studies diagnoses of lung cancer may have an unknown error rate, but the total death rate also increases monotonically with cigarette smoking. 2. Lung cancer mortality rates are higher in urban than in rural populations, and rural people smoke less than city people, but in both populations smokers have higher death rates from lung cancer than do nonsmokers. 3. Men have much higher death rates from lung cancer than women, especially among persons over 55, but women smoked much less and as a class had taken up the habit much later than men. 4. There are a host of causes of lung cancer, including a variety of industrial pollutants and unknown circumstances associated with socioeconomic class, with the poorer and less well off more likely than the better off to contract the disease, but no more likely to smoke. Cornfield et al. emphasize that "The population exposed to established industrial carcinogens is small, and these agents cannot account for the increasing lung-cancer risk in the remainder of the population. Also, the effects associated with socioeconomic class and related characteristics are smaller than those noted for smoking history, and the smoking class differences cannot be accounted for in terms of these other effects" (p. 179). This passage states that the difference in cancer rates for smokers and nonsmokers could not be explained by socioeconomic differences. While this claim was very likely true, no analysis was given in support of it, and the central question of whether smoking and lung cancer were independent or nearly independent conditional on

all subsets of the known risk factors that are not effects of smoking and cancer—area of residence, exposure to known carcinogens, socioeconomic class, and so on, was not considered. Instead, Cornfield et al. note that different studies measured different variables and "The important fact is that in all studies when other variables are held constant, cigarette smoking retains its high association with lung cancer." 5. Cigarette smoking is not associated with increased cancer of the upper respiratory tract, the mouth tissues or the fingers. Carcinoma of the trachea, for example, is a rarity. But, Cornfield et al. point out, "There is no a priori reason why a carcinogen that produces bronchogenic cancer in man should also produce neoplastic changes in the anspharynx or in other sites" 6. Experimental evidence shows that cigarette smoke inhibits the action of the cilia in cows, rats and rabbits. Inhibition of the cilia interferes with the removal of foreign material from the surface of the bronchia. Damage to ciliated cells is more frequent in smokers than in nonsmokers. 7. Application of cigarette tar directly to the bronchia of dogs produced changes in the cells, and in some but not other experiments applications of tobacco tar to the skin of mice produced cancers. Exposure of mice to cigarette smoke for up to 200 days produced cell changes but no cancers. 8. A number of aromatic polycyclic compounds have been isolated in tobacco smoke, and one of them, the form of benzopyrene, was known to be a carcinogen.

Perhaps the most original technical part of the argument was a kind of sensitivity analysis of the hypothesis that smoking causes lung cancer. Cornfield et al. considered a single hypothetical binary latent variable causing lung cancer and statistically dependent on smoking behavior. They argued such a latent cause

would have to be almost perfectly associated with lung cancer and strongly associated with smoking to account for the observed association. The argument neglected, however, the reasonable possibility of multiple common causes of smoking and lung cancer, and had no clear bearing on the hypothesis that the observed association of smoking and lung cancer is due both to a direct influence and to common causes. In sum, Cornfield et al. thought they could show a mechanism for smoking to cause cancer, and claimed evidence from animal studies, although their position in that regard tended to trip over itself (compare items 5 and 7). They didn't put the statistical case entirely clearly, but their position seems to have been that lung cancer is also caused by smoking and in some but not all other experiments applications of tobacco tar to the skin of animals cause cancer. A number of measurable factors that are not plausibly regarded as effects of smoking but which may cause smoking, and that smoking and cancer remain statistically dependent conditional on these factors. Against Fisher they argued as follows: The difficulties with the constitutional hypothesis include the following considerations: (a) changes in lung-cancer mortality over the last half century; (b) the carcinogenicity of tobacco tars for experimental animals; (c) the existence of a large effect from pipe and cigar tobacco on cancer of the buccal cavity and larynx but not on cancer of the lung; (d) the reduced lung-cancer mortality among discontinued cigarette smokers. No one of these considerations is perhaps sufficient by itself to counter the constitutional hypothesis, ad hoc modification of which can accommodate each additional piece of evidence. A point is reached, however, when a continuously modified hypothesis becomes difficult to entertain seriously. (p. 191) Logically, Cornfield et al. visited every part of the map. The evidence was supposed to be inconsistent

with a common cause of smoking and lung cancer, but also consistent with it. Objections that a study involved self-selection—as Fisher and company would object—was counted as an "ad hoc modification" of the common cause hypothesis. The same response was in effect given to the unstated but genuine objections that the time series argument ignored the combined effects of dramatic improvements in diagnosis of lung cancer, a tendency of physicians to bias diagnoses of lung cancer for heavy smokers and to overlook such a diagnosis for light smokers, and the systematic increase in the same period of other factors implicated in lung cancer, such as urbanization. The rhetoric of Cornfield et al. converted reasonable demands for sound study designs into ad hoc hypotheses. In fact none of the evidence adduced was inconsistent with the "constitutional hypothesis." A reading of the Cornfield paper suggests that their real objection to a genetic explanation was that it would require a very close correlation between genotypic differences and differences in smoking behavior and liability to various forms of cancer. Pipe and cigar smokers would have to differ genotypically from cigarette smokers; light cigarette smokers would have to differ genotypically from heavy cigarette smokers; those who quit cigarette smoking would have to differ genotypically from those who did not. Later the Surgeon General would add that Mormons would have to differ genotypically from non-Mormons and Seventh Day Adventists from nonSeventh Day Adventists. The physicians simply didn't believe it. Their skepticism was in keeping with the spirit of a time in which genetic explanations of behavioral differences were increasingly regarded as politically and morally incorrect, and the moribund eugenics movement was coming to be viewed in retrospect as an embarrassing bit of racism. In 1964 the Surgeon General's report reviewed many of the same studies and arguments as had

Cornfield, but it added a set of “Epidemiological Criteria for Causality,” said to be sufficient for establishing a causal connection and claimed that smoking and cancer met the criteria. The criteria were indefensible, and they did not promote any good scientific assessment of the case. The criteria were the “consistency” of the association, the “strength” of the association, the “specificity” of the association, the temporal relationship of the association and the “coherence” of the association. All of these criteria were left quite vague, but no way of making them precise would suffice for reliably discriminating causal from common causal structures. Consistency meant that separate studies should give the “same” results, but in what respects results should be the same was not specified. Different studies of the relative risk of cigarette smoking gave very different multipliers depending on the gender, age and nationality of the subjects. The results of most studies were the same in that they were all positive; they were plainly not nearly the same in the seriousness of the risk. Why stronger associations should be more likely to indicate causes than weaker associations was not made clear by the report. Specificity meant the putative cause, smoking, should be associated almost uniquely with the putative effect, lung cancer. Cornfield et al. had rejected this requirement on causes for good reason, and it was palpably violated in the smoking data presented by the Surgeon General’s report. “Coherence” in the jargon of the report meant that no other explanation of the data was possible, a criterion the observational data did not meet in this case. The temporal issue concerned the correlation between increase in cigarette smoking and increase in lung cancer, with a lag of many years. Critics pointed out that the time series were confounded with urbanization, diagnostic changes and other factors, and that the very criterion Cornfield et al. had used to avoid the

issue of the unreliability of diagnoses, namely total mortality, was, when age-adjusted, uncorrelated with cigarette consumption over the century. Brownlee (1965) made many of these points in his review of the report in the *Journal of the American Statistical Association.* His contempt for the level of argument in the report was plain, and his conclusion was that Fisher's alternative hypothesis had not been eliminated or even very seriously addressed. In Brownlee's view, the Surgeon General's report had only two arguments against a genetic common cause: (a) the genetic hypothesis would allegedly have to be very complicated to explain the dose/response data, and (b) the rapid historical rise in lung cancer following by about 20 years a rapid historical rise in cigarette smoking. Brownlee did not address (a), but he argued strongly that (b) is poor evidence because of changes in diagnostics, changes in other factors of known and unknown relevance, and because of changes in the survival rate of weak neonates whom, as adults, might be more prone to lung cancer. One of the more interesting aspects of the review was Brownlee's "very simplified" proposal for a statistical analysis of "E2 causes E1" which was that E1 and E2 be dependent conditional on every possible vector of values for all other variables of the system. Brownlee realized, of course, that his condition did not separate "E2 causes E1" from E1 causes E2," but that was not a problem with smoking and cancer. But even ignoring the direction of causation, Brownlee's condition—perhaps suggested to him by the fact that the same principle is used (erroneously) in regression—is quite wrong. It would be satisfied, for example, if, E1 and E2 had no causal connection whatsoever provided some measured variable Ej were a direct effect of both E1 and E2. Brownlee thought his way of considering the matter was important for prediction and intervention: If the inequality holds only for, say, one

particular subset Ej,..., Ek, and for all other subsets equality holds, and if the subset Ej,...,Ek occurs in the population with low probability, then Pr{E1|E2}, while not strictly equal to Pr{E1|E2 c }, will be numerically close to it, and then E2 as a cause of E1 may be of small practical importance. These considerations are related to the Committee's responsibility for assessment of the magnitude of the health hazard. Further complexities arise when we distinguish between cases in which one of the required secondary conditions Ej,...,Ek is, on the one hand, presumably controllable by the individual, e.g., the eating of parsnips, or on the other hand uncontrollable, e.g., the presence of some genetic property. In the latter case, it further makes a difference whether the genetic property is identifiable or nonidentifiable: for example it could be brown eyes which is the significant subsidiary condition Ej, and we could tell everybody with not-brown eyes it was safe for them to smoke. (p. 725) No one seems to have given any better thought than this to the question of how to predict the effects of public policy intervention against smoking. Brownlee regretted that the Surgeon General's report made no explicit attempt to estimate the expected increase in life expectancy from not smoking or from quitting after various histories. Fifteen years later, in 1979, the second Surgeon General's Report on Smoking and Health was able to report studies that showed a monotonic increase in mortality rates with virtually every feature of smoking practice that increased smoke in the lungs: number of cigarettes smoked per day, number of years of smoking, inhaling versus not inhaling, low tar and nicotine versus high tar and nicotine, length of cigarette habitually left unsmoked. The monotonic increase in mortality rates with cigarette smoking had been shown in England, the continental United States, Hawaii, Japan, Scandinavia and elsewhere, for whites and blacks, for men and

women. The report dismissed Fisher's hypothesis in a single paragraph by citing a Scandinavian study (Cederlof, Friberg, and Lundman 1977) that included monozygotic and dizygotic twins: When smokers and nonsmokers among the dizygotic pairs were compared, a mortality ratio of 1.45 for males and 1.21 for females was observed. Corresponding mortality ratios for the monozygotic pairs were 1.5 for males and 1.222 for females. Commenting on the constitutional hypothesis and lung cancer, the authors observed that "the constitutional hypothesis as advanced by Fisher and still supported by a few, has here been tested in twin studies. The results from the Swedish monozygotic twin series speak strongly against the constitutional hypothesis." The second Surgeon General's report claimed that tobacco smoking is responsible for 30% of all cancer deaths; cigarette smoking is responsible for 85% of all lung cancer deaths. A year before the report appeared, in a paper for the British Statistical Association P. Burch (1978) had used the example of smoking and lung cancer to illustrate the problems of distinguishing causes from common causes without experiment. In 1982 he published a full fledged assault on the second Surgeon General's report. The criticisms of the argument of the report were similar to Brownlee's criticisms of the 1964 report, but Burch was less restrained and his objections more pointed. His first criticism was that while all of the studies showed an increase in risk of mortality with cigarette smoking, the degree of increase varied widely from study to study. In some studies the age adjusted multiple regression of mortality on cigarettes, beer, wine and liquor consumption gave a smaller partial correlation with cigarettes than with beer drinking. Burch gave no explanation of why the regression model should be an even approximately correct account of the causal relations. Burch thought the fact that the apparent dose/response

curve for various culturally, geographically, and ethnically distinct groups were very different indicated that the effect of cigarettes was significantly confounded with environmental or genetic causes. He wanted the Surgeon General to produce a unified theory of the causes of lung cancer, with confidence intervals for any relevant parameter estimates: Where, he asked, did the 85% figure come from? Burch pointed out, correctly, that the cohort of 1487 dizygotic and 572 monozygotic twins in the Scandinavian study born between 1901 and 1925 gave no support at all to the claim that the constitutional explanation of the connection between smoking and lung cancer had been refuted, despite the announcements of the authors of that study. The study showed that of the dizygotes exactly 2 nonsmokers or infrequent smokers had died of lung cancer and 10 heavy smokers had died of lung cancer; of the monozygotes, 2 non smokers and 2 heavy smokers had died of the disease. The numbers were useless, but if they suggested anything, it was that if genetic variation was controlled there is no difference in lung cancer rates between smokers and nonsmokers. The Surgeon General's report of the conclusion of the Scandinavian study was accurate, but not the less misleading for that. Burch also gave a novel discussion of the time series data, arguing that it virtually refuted the causal hypothesis. The Surgeon General and others had used the time series in a direct way. In the U.K. for example, male cigarette consumption per capita had increased roughly a hundredfold between 1890 and 1960, with a slight decrease thereafter. The age-standardized male death rate from lung cancer began to increase steeply about 1920, suggesting a thirty-year lag, consistent with the fact that people often begin smoking in their twenties and typically present lung cancer in their fifties. According to Burch's data, the onset of cigarette smoking for women lagged

behind males by some years, and did not begin until the 1920s. The Surgeon General's report noted that the death rate from lung cancer for women had also increased dramatically between 1920 and 1980. Burch pointed out that the autocorrelations for the male series and female series didn't mesh: there was no lag in death rates for the women. Using U.K. data, Burch plotted the percentage change in the age-standardized death rate from lung cancer for both men and women from 1900 to 1980. The curves matched perfectly until 1960. Burch's conclusion is that whatever caused the increase in death rates from lung cancer affected both men and women at the same time, from the beginning of the century on, although whatever it is had a smaller absolute effect on women than on men. But then the whatever-it-was could not have been cigarette smoking, since increases in women's consumption of cigarettes lagged twenty to thirty years behind male increases. Burch was relentless. The Surgeon General's report had cited the low occurrence of lung cancer among Mormons. Burch pointed out that Mormon's in Utah not only have lower age-adjusted incidences of cancer than the general population, but also have higher incidences than non-Mormon nonsmokers in Utah. Evidently their lower lung cancer rates could not be simply attributed to their smoking habits. Abraham Lilienfeld, who only shortly before had written a textbook on epidemiology and who had been involved with the smoking and cancer issue for more than twenty years, published a reply to Burch that is of some interest. Lilienfeld gives the impression of being at once defensive and disdainful. His defense of the Surgeon General's report began with an ad hominem attack, suggesting that Burch was so out of fashion as to be a crank, and ended with another ad hominem, demanding that if Burch wanted to criticize others' inferences from their data he go get his own. The most substantive reply

Lilienfeld offered is that the detailed correlation of lung cancer with smoking habits in one subpopulation after another makes it seem very implausible that the association is due to a common cause. Lilienfeld said, citing himself, that the conclusion that 85% of lung cancer deaths are due to cigarettes is based on the relative risk for cigarette smokers and the frequency of cigarette smoking in the population, predicting, in effect, that if cigarette smoking ceased the death rate from lung cancer would decline by that percentage. (The prediction would only be correct, Burch pointed out in response, provided cigarette smoking is a completely unconfounded cause of lung cancer.) Lilienfeld challenged the source of Burch's data on female cigarette consumption early in the century, which Burch subsequently admitted were estimates. Both Burch and Lilienfeld discussed a then recent report by Rose et al. (1982) on a ten-year randomized smoking intervention study. The Rose study, and another that appeared at nearly the same time with virtually the same results, illustrates the hazards of prediction. Middle-aged male smokers were assigned randomly to a treatment or nontreatment group. The treatment group was encouraged to quit smoking and given counseling and support to that end. By self-report, a large proportion of the treatment group either quit or reduced cigarette smoking. The difference in self-reported smoking levels between the treatment and nontreatment groups was thus considerable, although the difference declined toward the end of the ten-year study. To most everyone's dismay, Rose found that there was no statistically significant difference in lung cancer between the two groups after ten years (or after five), but there was a difference in overall mortality—the group that had been encouraged to quit smoking, and had in part done so, suffered higher mortality. Fully ignoring their own evidence, the authors of the Rose study

concluded nonetheless that smokers should be encouraged to give up smoking, which makes one wonder why they bothered with a randomized trial. Burch found the Rose report unsurprising; Lilienfeld claimed the numbers of lung cancer deaths in the sample are too small to be reliable, although he did not fault the Surgeon General's report for using the Scandinavian data, where the numbers are even smaller, and he simply quoted the conclusion of the report, which seems almost disingenuous. To Burch's evident delight, as Lilienfeld's defense of the Surgeon General appeared so did yet further experimental evidence that intervening in smoker's behavior has no benign effect on lung cancer rates. The Multiple Risk Factor Intervention Trial Research Group (1982) reported the results after six years of a much larger randomized experimental intervention study producing roughly three times the number of lung cancer deaths as in the Rose study. But the intervention group showed more lung cancer deaths than the usual care group! The absolute numbers were small in both studies but there could be no doubt that nothing like the results expected by the epidemiological community had materialized. The results of the controlled intervention trials illustrate how naive it is to think that experimentation always produces unambiguous results, or frees one from requirements of prior knowledge. One possible explanation for the null effects of intervention on lung cancer, for example, is that the reduced smoking produced by intervention was concentrated among those whose lungs were already in poor health and who were most likely to get lung cancer in any case. (Rose et al. gave insufficient information for an analysis of the correlation of smoking behavior and lung cancer within the intervention group.) This possibility could have been tested by experiments using blocks more finely selected by health of the subjects.

In retrospect the general lines of the dispute were fairly simple. The statistical community focused on the want of a good scientific argument against a hypothesis given prestige by one of their own; the medical community acted like Bayesians who gave the "constitutional" hypothesis necessary to account for the dose/response data so low a prior that it did not merit serious consideration. Neither side understood what uncontrolled studies could and could not determine about causal relations and the effects of interventions. The statisticians pretended to an understanding of causality and correlation they did not have; the epidemiologists resorted to informal and often irrelevant criteria, appeals to plausibility, and in the worst case to ad hominem. Fisher's prestige as well as his arguments set the line for statisticians, and the line was that uncontrolled observations cannot distinguish among three cases: smoking causes cancer, something causes smoking and cancer, or something causes smoking and cancer and smoking causes cancer. The most likely candidate for the "something" was genotype. Fisher was wrong about the logic of the matter, but the issue never was satisfactorily clarified, even though some statisticians, notably Brownlee and Burch, tried unsuccessfully to characterize more precisely the connection between probability and causality. While the statisticians didn't get the connection between causality and probability right, the Surgeon General's "epidemiological criteria for causality" were an intellectual disgrace, and the level of argument in defense of the conclusions of the Surgeon General's Report was sometimes more worthy of literary critics than scientists. The real view of the medical community seems to have been that it was just too implausible to suppose that genotype strongly influenced how much one smoked, whether one smoked at all, whether one smoked cigarettes as against a cigar or pipe, whether

one was a Mormon or a Seventh day Adventist, and whether one quit smoking or not. After Cornfield's survey the medical and public health communities gave the common cause hypothesis more invective than serious consideration. And, finally, in contrast to Burch, who was an outsider and maverick, leading epidemiologists, such as Lilienfeld, seem simply not to have understood that if the relation between smoking and cancer is confounded by one or more common causes, the effects of abolishing smoking cannot be predicted from the "risk ratios," that is, from sample conditional probabilities.

Hans Reichenbach and the Direction of Time

In the 1950s, in the last years of his life, Reichenbach, a German philosopher who emigrated to the United States during the Nazi era, made the first substantial effort at a general linkage between directed acyclic graphs and probabilities. Unfortunately, his posthumously published book, *The Direction of Time*, was almost entirely ignored by everyone. Reichenbach described "screening off"—conditional independence—the "Principal of the Common Cause—that associations are to be explained (usually) by common causes—and recognized the collider principle (without naming it such). Unfortunately, the last in combination with the Principle of the Common Cause drove him to a silly conclusion. Convinced that causes must not be explained by their effects, and that associations must be explained by common causes, and that conditioning on common effects produces associations among the otherwise independent causes, Reichenbach concluded that whenever X and Y have a common effect Z, X and Y must have a common cause.

Steps Towards Graphical Causal Models and Search

In the 1960s, steps toward representation and search with graphical causal models were made by two mathematically inclined sociologists, Herbert Costner and Hubert Blalock. Blalock focused on models without unmeasured common causes while Costner focused on latent variable models but they often published jointly. The two gave a rather confused defense of causal modeling in the social sciences, arguing the "cause" is a primitive, undefined notion that does not require or allow useful definition, and that causal models are tested by their empirical consequences:

Like the notions of "probability," "point," "line," and "plane" - all of which are mathematical abstractions - causes can never be observed empirically, nor can they be defined theoretically except in terms of other concepts (some of which must be primitive) in the theory. But this does not make them useless constructs, nor does it prohibit their use in deductively formulated theories.

But then:

….following Simon and others, we have repeatedly emphasized that causal models and causal inferences refer only to our own thought processes.[20]

How a theory which is only about the theorist's thought processes can be tested by observational data or experiment is a mystery.

[20] Costner, H. L., & Blalock, H. M. (1972). Scientific fundamentalism and scientific utility: a reply to Gibbs. *Social Science Quarterly*, 827-844.

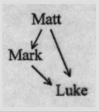
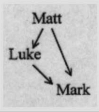
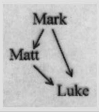
In parallel with views and techniques in statistics, there were developments in philosophy that echoed behaviorism. In the early 20th century Bertrand Russell proposed a model of the mind as a Fregean logical machine that constructs "concepts" from sensory data. He gave no details, but in Germany, Rudolf Carnap, a more literal guy, tried to describe the requisite logical operations. There were essentially early efforts at automated intelligence. Carnap's constructions didn't get very far, and arguably did not work as intended so far as they got. Logically inspired philosophers came quickly to reject the very idea that cognitive processes, and scientific inference in particular, could be captured or excelled by algorithms working on data, sensory or otherwise. By the 1950s, Carl Hempel, then perhaps the most eminent philosopher of science, announced that algorithms could never discover scientific theories. His argument was that genuine scientific theories introduce "theoretical concepts" that are not logically reducible to "observational data." Algorithms could never do that he claimed—even though psychometricians were doing so routinely with factor analysis. He was not alone. In physics, earlier in the century Percy Bridgeman argued for "operationalism"—distinct physical concepts are defined by the laboratory operations that are said to measure them. On Bridgeman's account, there could not possibility be two ways of measuring the same quantity.

There were a few outliers and ironies in philosophy. Norwood Russell Hanson, after a detailed study of Kepler's methods, claimed there is a "logic of discovery," which, unfortunately he was unable to describe with any specificity. Through the work of one of his graduate student's graduate student, Hempel's theory of scientific explanation as deduction of particular facts from laws and other

particular facts became part of the first semi-commercial artificial intelligence program. Herbert Simon, sometimes described as the father of artificial intelligence, studied with Carnap at the University of Chicago and even dedicated a book to him.

Statistical work related to causality was done throughout the 20th century but almost all of it was devoted to estimating parameters after a causal model is specified, not to discovering the causal relations except from experiment.

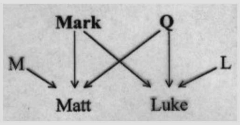
In the 1950s, Samuel Mason developed an algorithm for computing the correlations among variables in any feedback system with any number of interwined cycles represented in a finite cyclic graph. And at least one biblical scholar began using causal graphs to represent conjectures about the synoptic gospels.

The Traditional Theory (Augustinian Hypothesis)	**The Griesbach Theory** (Two-Gospel Hypothesis)	**The Farrer/Goulder Theory** (positing Markan Priority)
		

Note: Many other solutions have been proposed over the years, but most are variations of one of these three basic theories.

The Four-Source Theory (the solution accepted by most scholars today):

The Four-Source Theory
(a.k.a. Two-Document Hypothesis, from B.H. Streeter)



Mark = the oldest written Gospel, which provided the *narrative framework* for both Matt & Luke
Q = "Quelle" = a hypothetical written "Source" of some *sayings / teachings* of Jesus (now lost)
M = various other materials (mostly oral, some maybe written) found *only in Matthew*
L = various other materials (mostly oral, others probably written) found *only in Luke*
***Note*:** the arrows indicate direction of influence; older materials are above, later Gospels below
***Note*:** by definition, Q consists of materials found in Matthew *and* Luke, but *not* in Mark.

By the middle of the 20th century, multiple regression had become a common method for discovering causal relations. Textbooks often disclaimed any causal role for the method but then traded on examples and exercises with just such a purpose. The problems with the method for causal inference were well known: partial regression coefficients depend on which variables are chosen as regressors; omission of a common cause of prediction and outcome variables would break the interpretation of partial regression coefficients as estimates of effect. Less recognized was that a confounded prediction variable could produce mistaken estimates of effect for any other prediction variable with which it was associated. Attempts were made to address the first issue by variable selection methods; in the 1970s many issues of an entire journal, Technometrics, were largely devoted to such criteria. Arthur Dempster, for example, recommended eliminating variables corresponding the inverse covariance matrix—which is the wrong choice when predictors are associated and at least one of them is confounded with the outcome by an unobserved common cause. None of the statistical fitness measures bore logical connections to causal structure. Unmeasured confounding in observational studies was typically regarded as a hopeless problem, to be resolved by "judgement" or, later, by simulation studies of the sensitivity of regression estimates to confounders—which of itself can no information about how large the effects of confounders in any particular study actually were.

Although causal search methods such as regression were obviously being used as estimators of causal relations, they were never subjected to any of the usual criteria for statistical estimators. One can guess (and I do) that a principal reason

is that unlike the usual estimators, causal estimation had no mathematical object—no number—that was its object, except in the case of estimations of effect sizes. That changed in the late 1980s with the work of Harry Kiiveri and Terry Speed. Arthur Dempster had worked on "factorizations" of probability distributions: representations of joint distributions as products of conditional probabilities. Stefan Lauritzen and his collaborators developed the Markov factorization for distributions on variables over a directed acyclic graph. Kiiveri and Speed applied the idea to structural equation models around 1980.

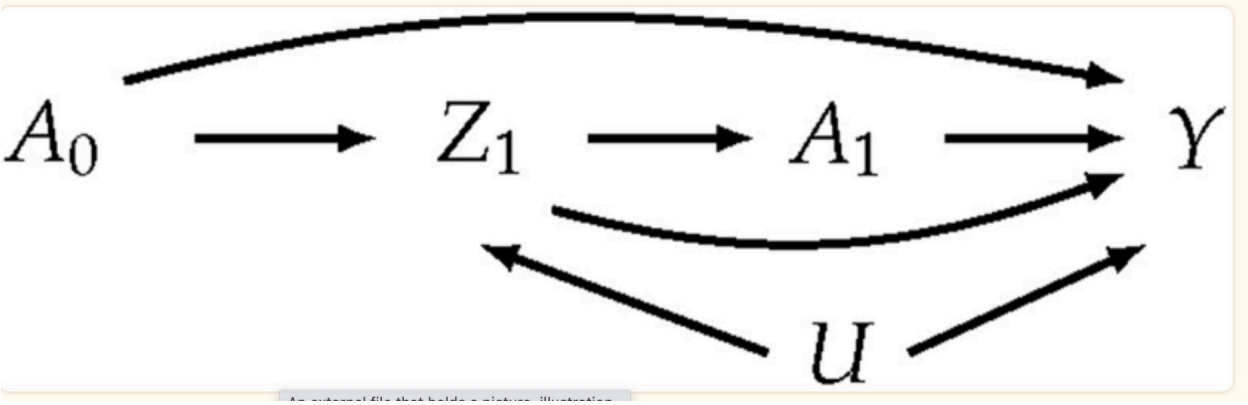
Almost a decade later, Stephan Lauritzen and Judea Pearl (and his students) independently proposed algorithms for computing whether a directed acyclic graph subject to the Markov factorization of an associated joint probability distribution implies that two variables are independent conditional on a subset of the variables in the graph. Pearl's version, known as d-separation, became the more widely used but the methods give equivalent results. Pearl, however, initially followed the tradition that claims "causality" is about what is in the investigator's head, not about what is in the world. His reason was that data could never provide a basis for distinguishing a causal relation from a common cause.

The connection between causal models and prediction of interventions remained largely opaque through the 20th century. Wold made general comments that a causal model implies claims about experiments but did not explain how. In 1974, Donald Rubin proposed what is now known as the "backdoor" criterion for estimating effect sizes. Rubin went on to develop and advocate the "potential outcomes" framework which has taken over professional statistics. In Aristotelian

spirit, the framework posits that for each value of a variable for each unit in a system its “direct” effects have a deterministic (in some versions, stochastic) response. This sounds like standard modeling except that in the framework, the response of the unit to any input is represented by the value of a separate variable, the “potential outcome.” For obvious reasons, the framework is usually applied when variables have only a finite range of values.Because in potential outcomes the causal ordering was assumed known, the adoption of the framework helped cement the antipathy of statisticians for causal search algorithms.

A Scandanavian group, including Stefan Lauritzen among others, introduced the Markov factorization for acyclic systems and arround 1980, Harry Kiverri and Terry Speed applied it to acyclic structural equation models. The Markov factorization immediately provided a method for predicting the effects of interventions in such systems, simply by plugging in the intervened value (or marginal distribution) of a variable or variables in the factorization. That left the issue of predicting the distributions resulting from interventions in systems with latent variables and feedback cycles, since no factorization was available for cyclic systems.

Jaime Robins in 1986 provided a partial solution for problems with latent variables as in the illustration:



U is unobserved, and the problem is to predict the effect on Y of a joint intervention (assumed binary) on Ao and A1.

Robins g-formula (which I have never fully understood) provided an answer. In 1993, Spirtes et al. laid the foundations for a general algorithm for predicting the effects of interventions in systems with latent variables, including what are now called the “back door” criterion (already described by Rubin two decades previously) and the “front door” criterion, which was novel. They described the procedure for partially oriented directed graphs in which some directions of edges are undetermined and there may be doubly directed edges representing the presence of unobserved common causes. That framework was subsequently specialized and elaborated for directed acyclic graphs by Pearl.

In 1991, Spirtes and Glymour introduced the PC algorithm for discovering Markov equivalence classes of causal structure from sample data and proved it correct in the large sample limit for i.i.d data from systems without unmeasured common causes. In 1992 Cooper and Herskovitz published a Bayesian algorithm to the same purpose but assuming information about causal order—their ideas had

circulate prior to publication and had prompted Spirtes and Glymour's work. By 1993, Spirtes had developed and proved correct the Fast Causal Inference algorithm which tolerated, and sometimes could identify, unobserved common causes, essentially refuting Hempel's argument. By 1996 Chris Meek had developed in his thesis the greedy equivalence search, a quasi Bayesian algorithm that did not require any prior knowledge of causal structure except acyclicity. Its correctness was subsequently proved with his colleague at Microsoft. Meanwhile, Cooper and others proposed a strictly Bayesian search. In 2006 Shohei Shimizu, JP Patrik O. Hoyer, Aapo Hyvarinen and Antti Kerminen introduced the LINGaM algorithm in 2006 for linear, non-Gaussian systems, which was soon generalized to tolerate unobserved confounders. Hyvarinen and Kun Zhang further developed linear non-gaussian search for a wide variety of distributions and Zhang has recently proposed a hierarchical search method, GIN, for similar systems. There have been a variety of contributions to time series search and its applications, including an algorithm by Biwei Huang in collaboration with Zhang that applies to non-linear time series.

After about the year 2000 the developments in automated causal search have been too many and too diverse to properly survey. Less impressive have been applications by users who were not developers. Even methods that would seem to be revolutionary for problems of variable selection, such as those of Steckhoven and collaborators, seem to have been rarely applied. This may be a pessimistic assessment since researchers who hit upon a causal model with the aid of a search algorithm may not report their source.