How Experiment Gets its Own Life

Deborah G Mayo

ABSTRACT

A framework of experimental inquiry includes the process of planning, generating and analyzing data, and using them to design and test hypotheses. It enables questions of interest to be asked in relation to some aspect of a proposed data generating mechanism. It succeeds by providing a platform to link individual experimental inferences to substantive phenomena, circumventing threats from partial and erroneous perspectives: *it succeeds by getting a life of its own.*

How Experiment Gets its Own Life

Deborah G. Mayo

Lyrics of that silly song kept coming into my head as I wrote this, I don't even like it….

"**My Life" (Billy Joel)**
Don't need you to worry for me cause I'm alright
I don't *want you to tell me I can't live alone*,
I don't care what you say anymore, *this is my life*,
Go ahead with your own life, *I'm fine on my own*.

Note: Tired of rag tag slides hanging around, this time I wrote a paper….these are just notes for presentation…

“Experiment lives a life of its own!”

This familiar slogan is thought to capture the epistemic importance and special powers of experiments.

But decades after the slogan became popular in philosophical circles due to Hacking (1983, 1992a, 1992b) and others, it seems we’ve just taken a few steps toward *building philosophy of experiment.*

- **How should we understand this well-worn slogan?**
- **Where does the separate life of experiment take place?**
- **How does it manage to get a life of its own?**
- **Why should it want its very own life?**

I begin with 3 interconnected glosses of the “lives its own life” slogan:

Experimental Aims: Apart From Theory Testing

Local goals: how to obtain, model and learn from experimental data: check instruments, rule out extraneous factors, distinguish real effect from artifact, signal from noise.

Self-improvement: getting better at interacting, controlling, and distinguishing effects, and designing better experiments to learn more and catch flaws and biases (learn how to learn).

Theory building: Is there even something to research further? (exploratory)
If so, how can we build a theory?

There may be no theory yet; just trying to find things out (even the domain may be unclear).

Even in testing theories, experiments enter.

To bridge the data-theory gap
Experiments often concern experimental phenomena the theory does not even talk about….part of their power!

Theories don't tell us how to test them.

Stability and Stubbornness: Experimental Knowledge Remains

Claims that pass a stringent "test of experiment" are generally stable through changes of theory, and despite needing to reinterpret the significance of experimental inferences.

Where experiments estimate parameters of a theory, the crux of good experiments is that the inferred estimates may be detached "clean estimates"— not sullied by unknowns.

(e.g., deflection of light parameter in experimental GTR)

Experimental knowledge grows, as do methods for its acquisition: models, instruments, computational, and self-checking tools.

This points to a crucial kind of progress overlooked when measures of growth are sought in terms of large-scale theory change, in updating probability assignments to them, or other favored macromethodology schemes…

This leads to the most significant reading.

Independent Warrant for Experimental Inference

- There is an independent justification of experimental data and inference

  Independent of what?

- To say "theory" or "high level" theory invites the criticism that theory always enters.

  But that mistakes what the "own life" is all about.

- What really matters: freedom from whatever could be a threat to what the researchers are trying to find out about.

- The interesting thesis is that experimental evidence and inference need not be theory-*laden.*

  They are not dragged down by anything that could invalidate their roles in a given inquiry.

- An inference from experiment E is vouchsafed apart from what is so far unknown.

Some philosophers of experiment stress the independent grounding afforded by instruments; others stress manipulation.

But neither are necessary or sufficient for experimental learning, they only point to some of the strategies involved in experimental learning.

Any means to reliable experimental learning may do as well or better, including computation, simulation, and statistical methods and modeling.

"Reliability" is notoriously ambiguous—put to one side here, but it does not suffice that the experiment will get it right in the long run, with high probability, asymptotically or the like….it has to be capable of *controlling misinterpretations in the case at hand…*

(at least within the time of a typical inquiry)

The goal is not error avoidance, but *error control.*

Error Control

A particular experimental inference may be in error, but so long as errors are sufficiently understood and controlled we may discover and avoid being misled by them.

In speaking of “errors”, I am not referring to “observational” errors, systematic or non-systematic, but mistaken *understandings of an aspect of a phenomenon* (theoretical or observational).

We might call a context “experimental” (whether literal manipulation is present) insofar as error probing capacities are able to be controlled and assessed.

Co-opting a term from (frequentist) statistics, where the probabilities that methods discern errors are called *error probabilities*, I sometimes refer to an *error statistical* approach (though the account is certainly not limited to formal statistical experiments).

A typology of mistakes about:

- real vs. spurious effects
- parameter values
- causes
- assumptions of experimental data,
- links from experimental claims to substantive claims

But these are just ways to categorize strategies for much more context-dependent claims.

We are interested in

- erroneous inferences
- erroneous interpretations of data
- erroneous understanding of phenomena

and whatever can hinder discovering these…

Paying deliberate attention to errors, I claim, is at the heart of getting correct inferences, interpretations, etc.

Even granting that all models are wrong, all theories strictly false, that does not prevent getting a correct understanding of experimental stabilities and effects

They have their own life!

When is it Bad to Be Dependent on (or a slave to) theory?

Assuming a theory may blind us to discovering new things.

An extreme case would be if the very hypothesis or theory under test, *H*, is implicitly assumed.

If data ***x*** *is generating by a method or procedure with little or no capability of finding (uncovering, admitting) the falsity of H, even if H is false, then* ***x*** *is poor evidence for H.*

The procedure has maximal error probability, *H* "passes" a test with minimal stringency or severity.

However, it may not be obvious that a method is guilty of such pre-judging.

Whether data-dependent inferences (data-mining, non-novel data) lead to high error probabilities needs to be scrutinized on a case by case basis,

e.g., adjusting free parameters of a theory to get it to fit data **x**

—the whole of issue of non-novel data, data mining, etc. is a tricky issue I talk (a lot) about elsewhere.

Experiment will have to do better than steering clear of some of the worst cases to be of interest.

Yes, before inferring *H* we want the experiment to have had a fairly high capability of unearthing flaws in *H* (stringency, severity).

Aside from stringency we want informativeness.

The goal is to learn about naturally occurring phenomena by probing a deliberately triggered experimental phenomenon.

Kuru

I have been reading about a disorder known as Kuru (which means “shaking”) found mainly among the Fore people of New Guinea in the 1960s.

In around 3-6 months, victims go from an unsteady gait to severe tremors, outbursts of laughter to inability to swallow and death.

Kuru, and (what we now know to be) related diseases, e.g., Mad Cow, Crutzfield Jacobs, scrapie) are “spongiform” diseases because the brains of their victims have small holes giving them a spongy appearance.

(TSEs: Transmissible Spongiform Encephalopathies)

Carlton Gajdasuk probed the disease experimentally without a clue as to what a satisfactory theory of Kuru might be.

He asked: *What causes Kuru? Is it transmitted through genetics? Infection? How can it be controlled or irradicated?*

- Considerable data in the 1950s showed Kuru clusters within families, in particular among women and their children, or elderly parents.

- They began to suspect transmission was through mortuary cannibalism by the maternal kin (this was a main source of meat permitted women, and was also a way of honoring the dead).

  It seems that men also took part, but got first dibs on eating the muscle.

- Ending these cannibalistic practice all but eradicated the disease which had been of epidemic proportions.

  (others remained convinced it was witchcraft)

Experimental Kuru

But to start experimenting they had to dream up method that would allow delimiting and probing possible causes…

Could kuru be transmitted to animals if their brains are inoculated with infected tissue from kuru victims?

Only after years of very long experiments did they show that *kuru was experimentally transmissible to chimpanzees*, monkeys and other animals by injecting them with infected brain extracts (from Kuru victims).

They studied experimental kuru to study actual kuru.

From the chimp data **x**, experimenters did not pretend to have established more than evidence as to Kuru's transmission; they called it a "slow virus" (given the lengthy incubation).

No one expected revolutionary implications: the discovery Kuru involved a novel type of infectious particle: the "prion" (with no nucleic acid) (Stanley Prusiner).

Experiment & Revolutionary Science

Experimental knowledge with a life of its own offers a counter to theory-laden data.

But also to the supposition that experimental results cannot change whole paradigms (including the paradigm they are “in”).

(**paradigm**: cluster of aims, methods, theories)

Experimental anomalies may be shown to be unevadable, and may overthrow (so-called) “hard cores” with the same (“normal scientific”) methods.

Any future “theory” of the phenomenon, to be adequate, must accommodate it.

Big shake-ups can result from local experimental effects.

Experimental Anomalies of the “Central Dogma” of Molecular Biology

- Whatever is causing kuru (also scrapie in sheep) is not iradicated with techniques known to kill viruses and bacteria; and they are weakened with those that weaken protein.

  (Prions are resistant to inactivation by UV irradiation, boiling, standard gravity autoclaving at 121°C, hospital disinfectants, propriolactone, hydrogen peroxide, iodophors, peracetic acid, chaotropes…)

  Also, victims show no antibodies as are produced by infectious elements with nucleic acid.

- So if it were a mistake to construe Kuru as having no nucleic acid, at least one of these known techniques would have iradicated it.

Experimental Argument from Error

We argue that *H* is a correct construal of data **x**, when the procedure would have unearthed/signaled the misinterpretation, but instead regularly passes *H*.

- They did not know what this non-virus was, and said as much ("the transmission of experimental Kuru, became well established, but its mode of action remained puzzling")

- Experiments may let us ask a single question, nucleic acid or not? Protein or not?

x is generally a vector of outcomes possibly from several subexperiments.

It’s All in the Planning and Design of Experiments

Experiment’s living “its own life”, circumventing obstacles to finding things out, is a consequence of something that has received too little attention in philosophies of experiment:

Experimental design and planning

Deliberate planning for collecting, “treating”, modeling, and analyzing data.

Experiments are distinguished from passive observation precisely because of the role of deliberate design in affording control, and delimiting the factors that would otherwise interfere with learning about effects of interest.

(Even with “fortuitous” observations, researchers may try to mimic what experiment offers)

This brings up a question of what kinds of examples philosophers of experiment might fruitfully consider.

If perfect controls are attainable then we would not be in a very interesting domain of learning, so I am most interested in more challenging kinds of cases.

For the same reason we should look beyond cases where the goals are limited to a particular data generating mechanism (DGM).

(I thank Isabelle Peschard for highlighting, in discussing this conference, a tendency to blur the distinction between some primary phenomenon of interest and a DGM).

(In econometrics some say the main question is “embedded” in a statistical DGM, which then emphasizes the need to connect back up)

Homes for Experimental Life

Experiments have lives of their own, but it should be a real life, not a life in the street, with their own parameters, models, theories.

An account that begins with given statements of evidence and hypotheses will not be relevant to the practice of experimental science.
(An important way experimental philosophers correct evidential-relation "logics")

That an adequate account of experiment must explicitly incorporate the methods and models to obtain and analyze data is well accepted by now.

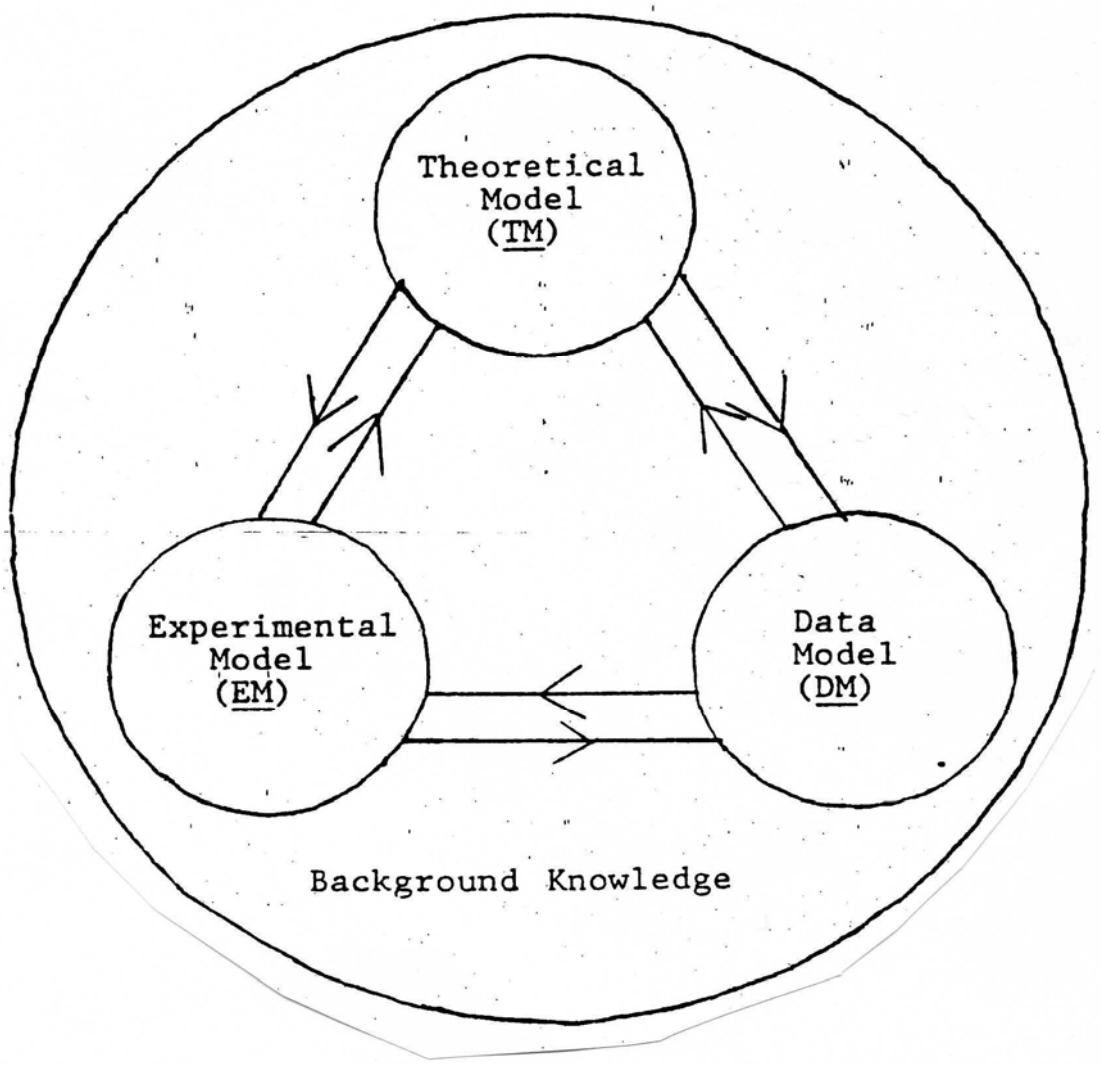
What about the work that goes into designing hypotheses to test or infer?

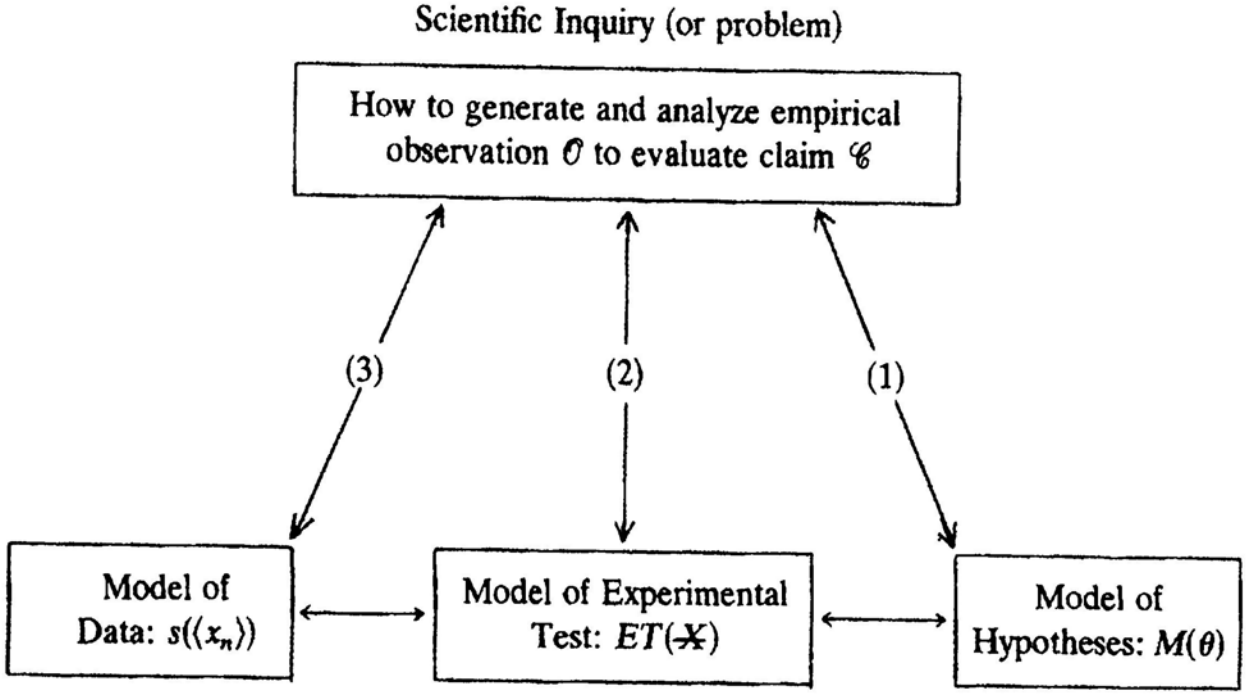
I want to highlight a triad of components in planning, running and interpreting experiments in practice:

question, hypotheses
data collection
data analysis and interpretation

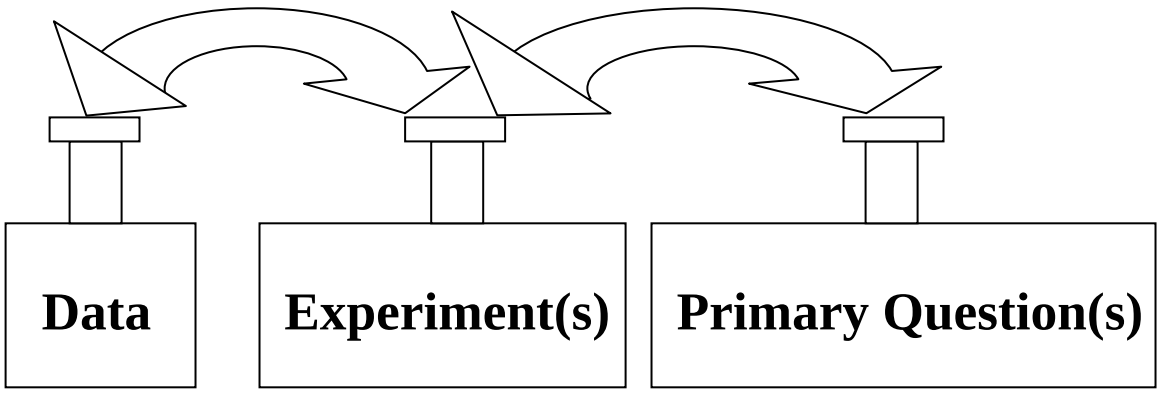
There may be just a vague question or speculation (pre-data) which may be very different from a hypothesis or inference arrived at (post-data).

To capture a single unit of experimentation (perhaps as big as a given inquiry or published paper) I may revive the simple circle from my Ph.D thesis.


Theoretical
Model
(TM)
Experimental
Model
(EM)
Data
Model
(DM)
Background Knowledge



Statistical Inquiry: Testing Hypotheses



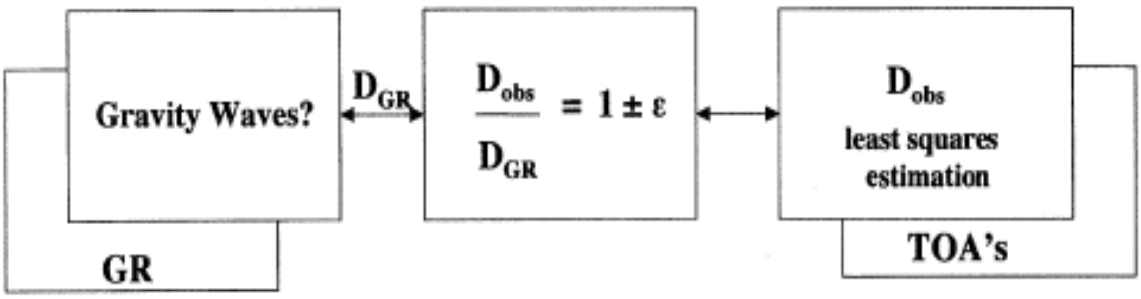
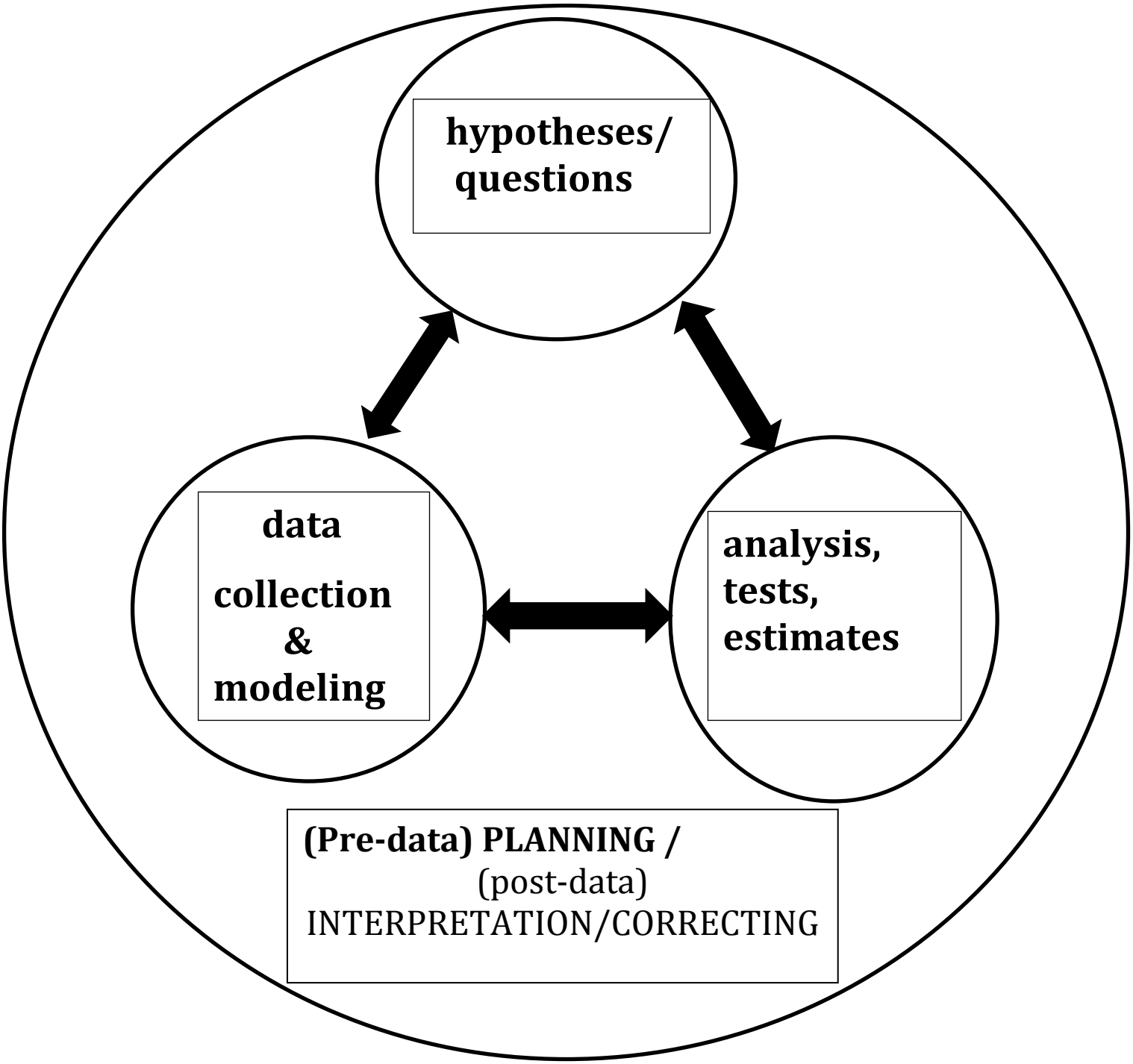


Figure 2. Binary pulsar: 1913 + 16. D = orbital decay.

EXPERIMENTAL TRIAD E_i



However, the three are actually intimately connected, and the planning concerns all three at once.

What is inferred at one stage may well be data for the next.

The deliberate use of raw data, differently modeled, is a key to the question of testing assumptions for a given experiment.

Again, I am imagining here a unit that might correspond to a paper reporting a chunk of results…to be combined with others later on…

C.S. Peirce

It is perhaps unsurprising to find support for this conception in Charles Peirce, one of my heroes:

- He coins the term “quasi-experimentation” to include the entire process of searching, generating and analyzing the data, and using them to test a hypothesis and “this whole proceeding I term induction” (CP 7.1115)

- Further, for Peirce, the “true and worthy” task of logic is to “tell you how to proceed to form a plan of experimentation” (7.59)

Goals of Planning:

- split off questions that can be probed in terms of experimental data generating procedures so as to afford adequate control of interfering factors.

- try to ensure (pre-data) that experimental analysis of the data will be possible so that something is very likely to be learned (even if it is botched).

- Make the data talk via “custom made” effects that might not even exist in nature.

Design specifications alter the error probing capacities of experiments, but far from sullying results.

Making use of this information is the key to an objective interpretation of results.

Estimate, Subtract out, Distinguish, Amplify

An experiment need not actually exclude interfering factors to be "free" of them.

Enough to *estimate (or simulate)* their influence, "knock out" or *subtract* out effects, or cleverly *distinguish* disentangle the effects of different factors.

This gives rise to a variety of strategies:

Amplifying effects

enhance the means to listen, amplify the whispers to distinguish the source of data

magnify the effects of distortions, such deviations can be made to speak volumes;

Self-correcting and/ or show robustness

Violating background assumptions might render them less efficient, the inference is not vitiated

Any particular experimental trian E_i uses and builds on <u>a repertoire of errors</u>: those ruled out and those remaining.

Experiments can "use" while but still not depend on theories.

Some claim experimental philosophers "throw out the baby with the bathwater" (Musgrave, Chalmers, Laudan).

We overlook, they say, how we invariably use background theories.

They overlook how experiment may "use" theories while not being threatened by them:

- *In hypothetically drawing out their consequences*
- *When they have passed severe tests of their own;*
- *When the only aspects being relied on are known to hold sufficiently*

Finally, an experiment can always free the interpretation from assumptions, possible sources of error:

State it in the sum-up (~ conditional proof in logic)

Experimental inferences should indicate both what has and what has not been adequately shown.

These Strategies Lead Experiment to be Piecemeal

"Getting small": we specify a question that will restrict or control erroneous interpretations of the kind of data we are actually in a position to collect.

This circumvents the familiar problem philosophical accounts face: "alternative hypothesis objections".

Unable to exhaust the space of alternatives, they ask: How can you rule out ways *H* can be false when there are always members in the "catchall hypothesis" — claims not even considered.

They then often turn to comparative accounts

But experiment can deliberately delimit factors that could be responsible for effects.

If a certain gene is successfully knocked out of a mouse, it can't be responsible for the effect.

Rather than trying to distinguish a hypothesis or theory from its rivals experiment sets out to distinguish and rule out a specific erroneous interpretation of the data from *this* experiment

One could rationally reconstruct experimental inquiry using models of large-scale theory change, of Bayesian updating, of decision theory with specified losses—once an episode is neat and tidy.

The ease of doing so, some think, is one of the weaknesses of such reconstructions.

Like a paint-by-number algorithm for art, they do not capture how the learning took place.

They are backwards-looking, not forward-looking, and fail to do justice to actual experiments, where pre-data specifications of an exhaustive set of hypotheses are atypical…

Even in evidence-based policy (or other subsequent decisions) should rest on a valid experimental knowledge base[i].

Arguing From Coincidence

The powerful form of argument experiments provide is often described as an "arguments from coincidence":

There's no way all of these well-known instruments and independent manipulations could consistently produce certain effects or data, were they all artifacts of instruments.

The inference to a non-artifact (or "real effect") is an instance of my **general argument from error**:

We argue that *H* correctly interprets (or is warranted by) data **x**, when the procedure would have (with high probability) unearthed the misinterpretation, but instead passes *H*.

We give the "artifact" interpretation a good chance to show itself…

Knowing an effect is "real" (non-artifactual) is one of the weakest kinds of experiment knowledge

Still it is important.

It was the first error on my list.

By leaving arguments from coincidence at a vague level, they are often appealed to as sustaining more than they actually allow.

Yet many say they do: It would be a preposterous coincidence if all these different experiments E_1, E_2, E_3, E_4, …. E_n, yield agreement if a theory *T* were false—

Would it? That's what we would need to figure out…

The error that needs avoiding to inferring merely that there is evidence of a real (not spurious) effect does not directly warrant a theory *T* that might explain the real effect.

Put epistemology of experiment at the level of experiment

A pattern of argument may work for ruling out just one error (erroneous interpretation) because we can exhaust the possibilities.

Moreover, the varied experiments E_1, E_2, E_3, E_4, …., E_n, must be shown to check the others; whatever might threaten one experiment must not also be able to cause an error in the others.

Let's go back to experimental prions…

An Experimental Platform for Understanding Prion Diseases

The use of “animal models”, mice and hamsters to probe various aspects of prion transmission.

- Prions, which contain single protein PrP, *prion protein*, is found in normal animals! So it does not always cause disease…
- They designed experiments to inform them if they had made a terrible mistake (that PrP had nothing to do with it).
- Mice with PrP deliberately knocked out were free of infection with PrP-Sc.
- The common form, PrP-c; while the pathogenic form, PrP-Sc.

Combining transgenic approaches and computer modeling methods, with mutations they could produce mice susceptible or resistant to prion disease in predictable ways.

- Transgenic mice with a hamster PrP gene were created: when inoculated with mouse prions, the made more mouse prions, then inoculated with hamster prions, more hamster prions.

- Crucial for manipulating transmission rates and for understanding the “species barrier” (which at least makes it less likely to transmit disease).

- Prion transmission between species had long been modeled as a stochastic process.

- Experimental manipulation in the 1980s enabled them to predict and control transmission, so it is (“it becomes a nonstochastic process”).

  Incubation is sped up from 600 days to 90 by continual injections of pathogenic brain tissue.

  (also relevant for whether humans could get Mad Cow)

  Yeast was a great model organism for biologists—a growth cycle of 80 minutes.

- But only by connecting these transgenic and normal animal models, and both these to yeast, can one serve as a check on the other.

The Only Correct Interpretation of the Data Experimentally Refuted the Reigning Paradigm

- Mixing (synthetic versions of) the two proteins together in a testube converted common Prp (PrP-C) into scrapie PrP (PrP-Sc) [in vitro].
- The infectious form has the same amino acid sequence as the normal type, nor were they chemically modified.
- Studying the amino acid sequence would have been unable to discern what made the difference.
- Accepting the “central dogma” of molecular biology would have been an obstacle: it assumes nucleic acid directs replication of pathogens.
- They hypothesized that the scrapie protein propagates itself by them to flip from their usual shape to a scrapie shape.
- To understanding the mechanism of pathological folding required knowing the structures of each.

A big experimental obstacle was not being able to discern the prion’s 3-D structure at the atomic level.

Magic Angle Spinning: Exploiting an Obstacle

Exploiting the obstacle provides the key…

The central difference between normal and pathogenic prions permits the normal but not the abnormal prion to have its structure discerned by known techniques, e.g., NMR for solutions.

The normal form, Pr-C, is soluble; Pr-SC is not.

NMR Nuclear Magnetic Resonance spectroscopy provides an image of molecular structure.

Put a material inside a very high magnetic field, hit with targeted radio waves, and its particles react to reveal structure; but it won't work for clumpy Pr-Sc.

While we need trillions of molecules to get a signal--**-amplification**
But this also amplifies interference of neighboring molecules in the non-soluble PrP-Sc.

We want to find out what it would be like if we were able to make it soluble, if we could subtract out neighboring molecules.

The Magic: erase the influence of these neighboring molecules.

If the sample, crushed into a powder form, spins within a magnetic field at a special angle to that field – 54.7 degrees – the influence of a molecule's environment is cancelled out.

("The effect on the spectrum from the magnetic interactions between the molecules vanishes.")

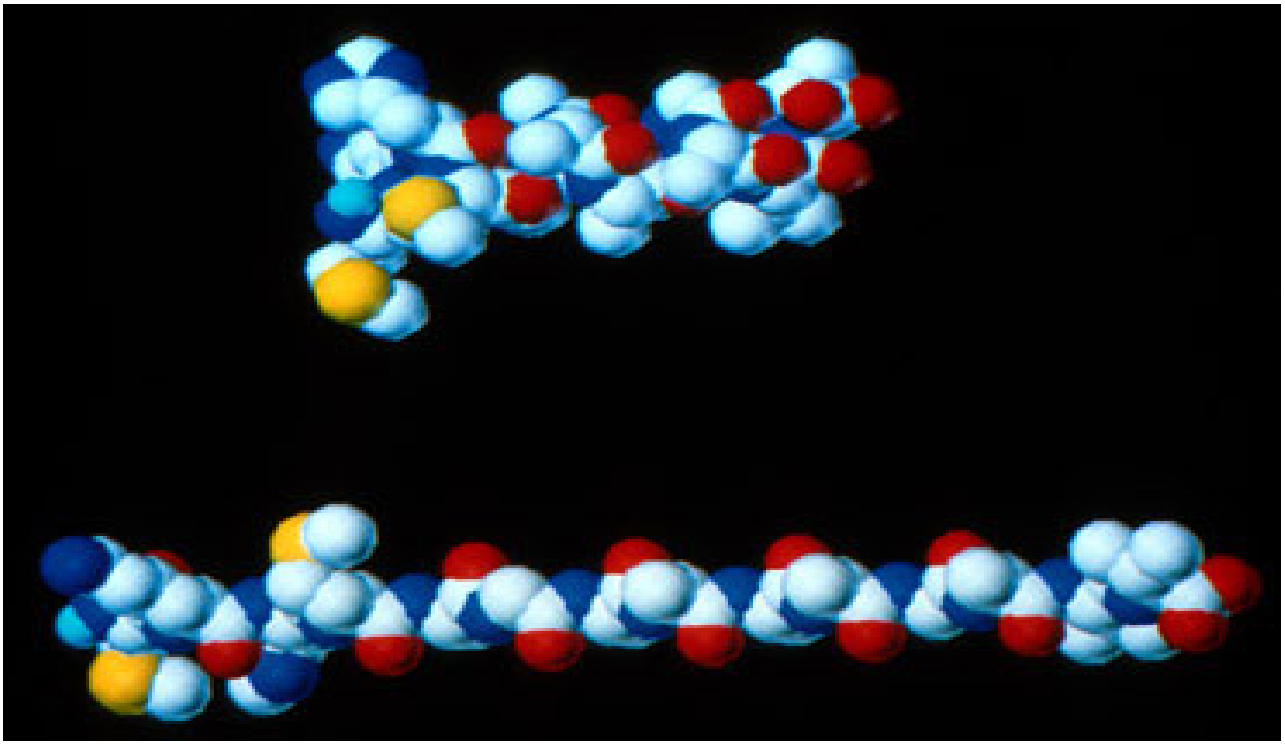
(Stems from quantum mechanics but that it works is shown with known molecular structures)

But the molecules cannot all be lined up at the magic angle, they can <u>on average</u> if they are spun fast enough.

(with respect to every other molecule in the sample)

In practice, this often means the sample needs to spin 25,000-50,000 revolutions per second.
(get all molecules in sample to sing in same key)

—It can take weeks of listening even in a 4 mm rotor



Prusiner's “prion only” theory is that prions target normal PrP and turn it into a form that folds and clumps, ultimately causing brain cells to rupture.

The abnormal prion moves on to normal prion proteins, pinning and flattening their spirals, akin to a deadly Virginia reel in the brain[ii].

As is typical with powerful experimental tools, we have very general instruments (e.g., solid state NMR) where we can check if we get it right for known sample.

By deliberately turning the NPR turbine too slowly for the magic angle, we get expected distortions, showing we are interacting as predicted.

The known distortions and limits of each experimental analysis is the key in order to link to the phenomenon of interest).

Generalizing into much wider spectrum of diseases involving pathological folding (e.g. Alzheimer’s) give more stringent, interconnected experimental probes.

CONCLUDING REMARKS

A model of experimental inquiry (a triad making up a data-inference-hypothesis grouping) enables questions of interest to be asked in relation to some aspect of the proposed data generating procedure

It is successful to the extent it is free of threats from obstacles to finding things out**: own life.**

It gives a general platform that is not domain-specific, enabling standard checks to interlink individual experimental inferences, known to be limited, partial, distorting….

With transgenic mice designed to produce tons of the normal prion, synthetic prions, protein folding in vitro..we have amplified, controlled, etc…

E_1, E_2, E_3, E_4, …., E_n

The immediate goal in each is not to rule out rival theories but a mistaken interpretation of results from this experiment.

But what do the pieces say about actual (real-world) prion disease?

Might these be relevant just for Frankenstein mice, yeast, synthetic prions, mimicked scrapie in a test tube?

To combine these pieces requires understanding the errors or distortions that remain, and those being avoided or subtracted out.

(What's being amplified? What's being quieted?)

We can investigate and learn about these limits and distortions by deliberately amplifying and controlling them in known or canonical cases.

The key stems from experimental benchmarks, calibration standards, extrapolation models.

By the time magic angle spinning is used on prions, it is a well understood, and reliable instrument.

Imagining an instrument works only when used on a known sample would be to imagine it reads our minds, and conspires to trick us when we are faced with unknown samples.

(As were I to claim all the scales are conspiring to show I’ve gained weight, despite their working fine with objects of known weight)

That would be a radical obstacle to learning!

The powdered PrP-Sc spun around in magic spinning turbines can tell us of the 3-D structure of PrP-Sc because we know how they work with known solid state specimans.

We only discern the mere structural relationship but that’s all we need for this piece of the puzzle.

Other interlocking hypothesis that make them relevant to what we want to know.

To understand specifically how and why this works we should consider how each checks, unearths, amplifies, erases threats or shortcomings of others.

Scientists are rarely that explicit about what makes their experiments work so well.

It is an important role for a future philosophy of experiment.

[i] I also reject the idea that evidential appraisal as unable to live a life apart from cost benefit analysis It has done much damage in evidence-based policy: If disagreements about interpreting are indistinguishable from disagreements about matters of subjective opinions and values, then each side gets its own scientific experts; we get “junk science” on both sides!

Should data on the frequency of Mad Cow be interpreted according to the economic values of the beef industry?

Obtaining sound inferences needs to be distinct from, and not inextricably intertwined with, subsequent policy decisions.

But that is a subject for a different discussion (Mayo 1991).

[ii] Nowadays, they also know how to perform cyclical amplification: lop off the ends and you get huge amounts of PrP-SC

-both affirming a correct understanding of the misfolding mechanism (at the ends) and giving an important new tool to detect it in living animals