EXPERIMENTAL RESEARCH METHODOLOGY

Can one always trust scientific results?

What is done to make them reliable?

This lecture:

- Choice of study designs
- Reproducibility
- Openness in science
- Examples from computer science
- Validity (statistical significance)
- and if enough time, serendipity in science

1. STUDY DESIGNS

In clinical medicine in particular, considerable effort goes into the design and and statistical analysis of experimental studies, and one can learn from this field. Let us explore the example from last lecture a bit further (Vitamin C treatment):

1. Survey
Do a web questionnaire. Problem: Secondary data, subject bias

2. Uncontrolled study
Gather some subjects who take a daily dosis during some time period. Problem: No baseline comparison

3. Subjects choose treatment
Gather some subjects. Each person decide if she wants to take a daily dosis. Problem: Subject bias, unbalanced groups?

4. Controlled trial
Find some subjects. Experiment leader decides who will have a daily dosis and who will not. Measure which group had fewer days of infection. Problem: Researcher bias, treating physician bias, selection procedure?

5. Randomized controlled trial
Find some subjects. Randomly decide who is going to have a daily dosis of vitamin and who will not. Measure which group had fewer days of infection. Problem: No estimate of placebo effect.

6. Blind randomized controlled trial
Find some subjects. Randomly decide who is going to have a daily dosis and who gets a placebo. Subjects do not know which group they belong to. Measure which group had fewer days of infection. Problem: Physician bias

1. STUDY DESIGNS

7. Double blind randomized controlled trial
Suggestion: Find some subjects. Randomly decide who is going to have a daily dosis of vitamin and who will not. Subjects *and the treating physician* do not know which group they belong to. Measure which group had fewer days of infection.
Problem: Researcher bias - choice of methods

8. Triple blind randomized controlled trial
Suggestion: Find some subjects. Randomly decide who is going to have a daily dosis of vitamin and who will not. Subjects, *treating physician, and the researcher* analyzing the data do not know which group they belong to. Measure which group had fewer days of infection.

9. Case control study
Identify patients with the infection, determine their history of taking the drug or not, and analyze the data.

10. Retrospective cohort study
Find a very large set of historical data that shows whether patients have taken the drug or not, together with other medical information, and apply advanced data analysis methods.

11. Meta-analysis. Analyze all existing studies of the treatment and try to use statistical concepts to combine the findings from individual studies.

1. STUDY DESIGNS

Double (or triple) blinded controlled trials are today consider the gold standard for medical trials, i.e., the highest standard of evidence.

First documented (single) blinded study from 1784 - French Royal Academy of Sciences investigated claims of ”animal magnetism” due to Franz Anton Mesmer, a very famous doctor/charlatan.

1. STUDY DESIGNS - COMMENTS

Issues with double blinded studies could be:

Ethical - patients are not completely informed, and may reasonably feel that they have a right to know what treatment they are receiving. This is known to limit participation.

Resource limitations - the complexity of arranging studies could limit size of studies leading to lower accuracy, and also limit the number of studies done.

Can you give any example where randomized studies could be used in experimental computer science? Blinded randomized studies?

1. STUDY DESIGNS - DATA SCIENCE, META-ANALYSES

- The increasing availability of large scale databases in medicine is rapidly increasing the importance based on retrospective analysis of historical data. The Swedish health care system reports data to more than 100 national quality registers.

- Many medical questions can only be studied in this way. Many factors cannot be manipulated at will by experimenters, say when studying the effects of smoking or other harmful habits. And some studies may involve changes to larger systems that seldom can be changed for the purposes of an experiment.

- Meta-analyses are statistical studies that integrate information from many different research publications on the same research questions. This is very common in medicine (e.g. Cochrane reports)

2. REPRODUCIBILITY

Why is it important to reproduce experiments?

- expose scientific fraud
- find errors and inaccuracy (statistical accidents)
- exploring same effect in a different way gives additional credibility to the findings
- new viewpoints can result in new discoveries

Why may one want to reproduce experiments?

- science as competition - find errors with competitor
- get new ideas on how to improve on the results
- check and calibrate new experimental methods and technology on known results
- use as benchmark

Why may one not want to reproduce experiments?

- unlikely to get get published
- little credit compared to more original discoveries
- low priority given resource limitations
- some experiments are impossible to reproduce

2. REPRODUCIBILITY

Three levels of replication

Repeatable
The original team of researchers can reliably produce the same result using the same setup

Replicable
A different team can produce the same result using the original setup

Reproducible
A different team can produce the same result using a different experimental setup

2. REPRODUCIBILITY - Examples: unrepetable, unreplicable and unreproducible experiments

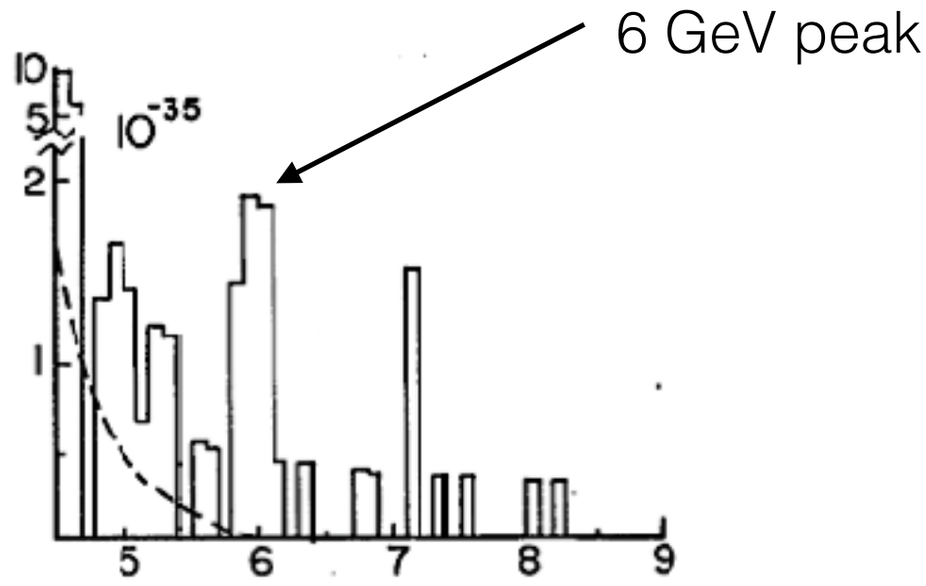
a. Unrepeatable

Rare events result in limited statistics - sometimes results may be communicated too early and disappear with more extensive measurements.

In 1976 at the Fermilab accelerator in the US, a team led by Leon Lederman (later physics Nobel prize 1988) announced the discovery of a new particle of mass of approximately 6 GeV decaying into an electron-positron pair. The probability of this being a chance occurrence was reported as less than 2%.

With additional measurements the experimental signal disappeared.

In particle physics today, events are required to be at least 5 standard deviations (5-sigma) above the random background to be reported, meaning a probability of a chance occurrence of 1 in 3.5 million (compare to p-values of 0.05 in many fields)



2. REPRODUCIBILITY

Examples - unreplicable experiments

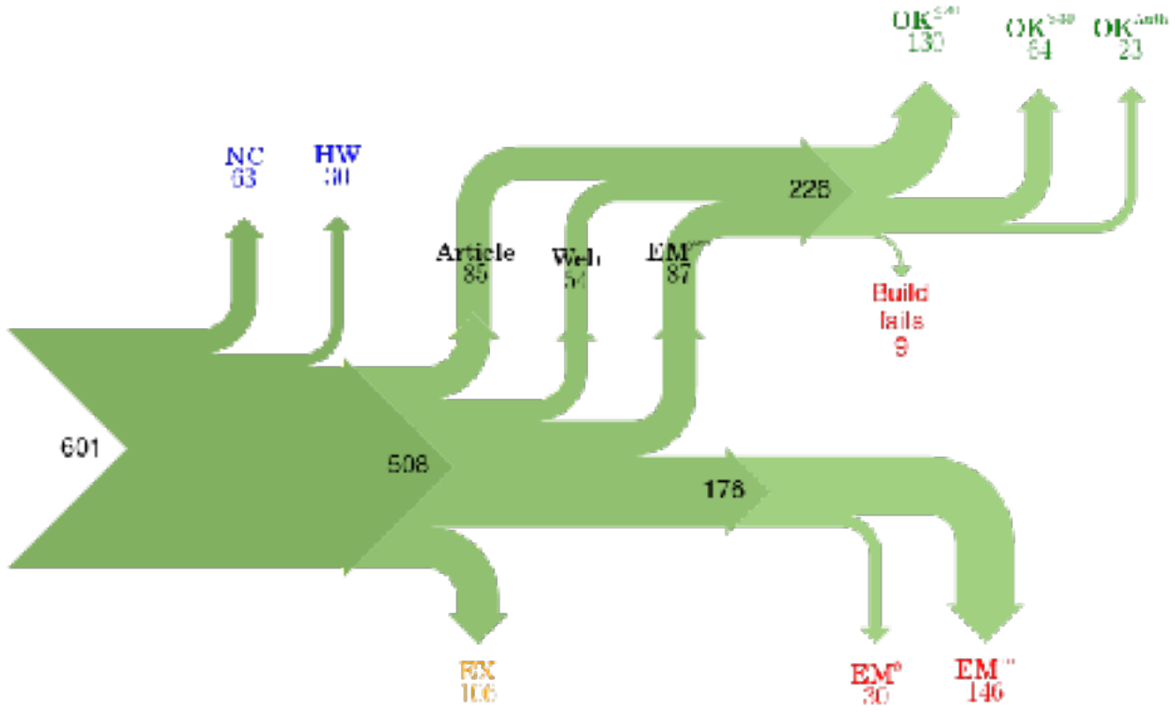
Computer science research deals with complex artifacts:
source code, datasets, sometimes esoteric hardware, etc

To repeat experiments under identical conditions, other researchers must have access to these, and be able to use them. In other words, for computer science experiments to be replicated, the following may need to be satisfied:

- source code and dataset must be available
- the code has to build
- the execution environment has to be replicated
- the code has to run to completion
- accurate measurement must be made

2. REPRODUCIBILITY

For example, in Collberg et al (2015), the authors of 601 papers from 13 leading ACM conferences and journals in areas such as computer systems, programming languages and embedded systems were asked for code and data to perform a replication of the result. Out of 402 papers, the code was made available and built with reasonable effort in 130 cases, and with more effort in an additional 64 cases.



Legend

Classification		Code Location		Build Results	
BC	Paper where the results are backed by code.	Article	Code is found from link in the article itself.	$OK^{\leq 30}$	We succeed in building the system in ≤30 minutes.
NC	Paper excluded due to results not being backed by code.	Web	Code is found from a web search.	$OK^{>30}$	We succeed in building the system in >30 minutes.
HW	Paper excluded due to replication requiring special hardware.	EM^{yes}	Code is provided by author after email request.	$OK^{>Author}$	We fail to build, but the author says the code builds with reasonable effort.
EX	Paper excluded due to overlapping author lists.	EM^{no}	Author responds that the code cannot be provided.	Fails	We fail to build, and the author doesn't respond to survey or says code may have problems building.
		$EM^{\emptyset}$	Author does not respond to email request within 2 months.		

Christian Collberg, Todd Proebsting, Alex M Warren (2015), Repeatability and Benefaction in Computer Systems Research, University of Arizona TR 14-04

Excuses (the dog ate it):

National Science Foundation's (NSF) Grant Policy Manual 12:
Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing.

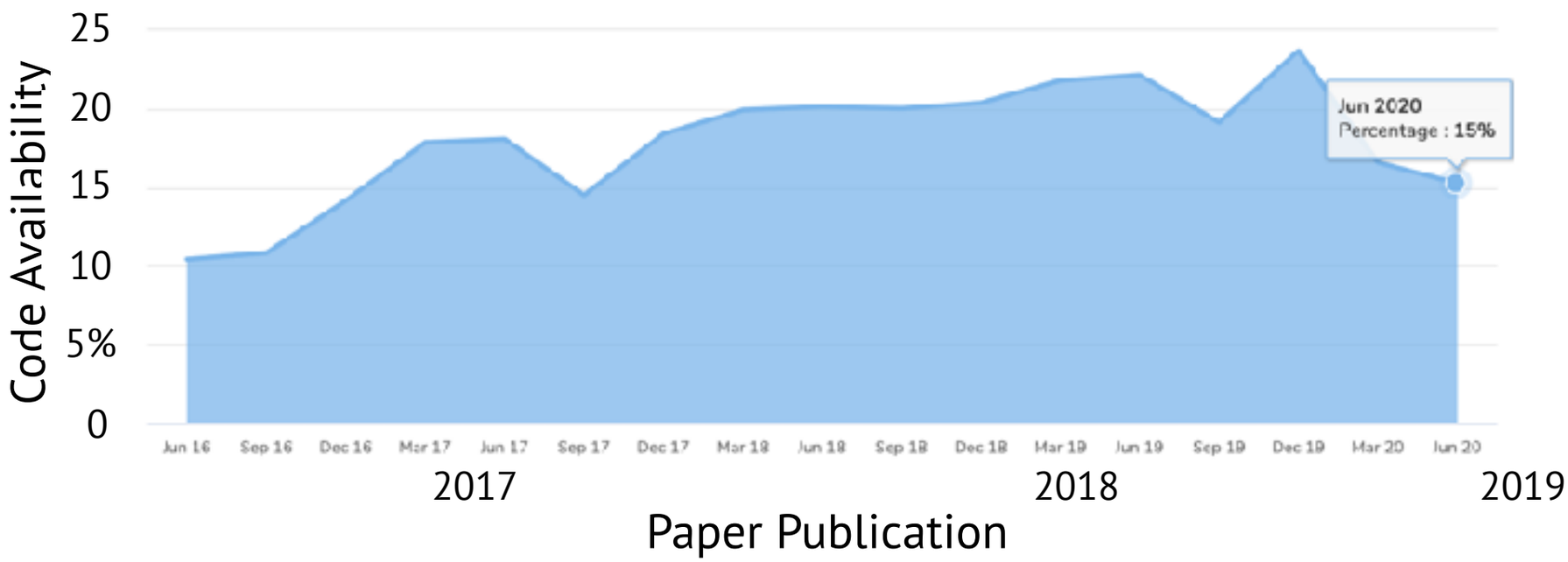
Some responses to requests for code from Collberg et al. (2015):

```
Thank you for your interest in our work. Unfortunately the current system is
not mature enough at the moment, so it's not yet publicly available. We are
actively working on a number of extensions and things are somewhat volatile.

I am afraid that the source code was never released. The code was never
intended to be released so is not in any shape for general use.

X was a graduate student in our program but he left a while back so I am
responding instead. For the paper we used a prototype that included many
moving pieces that only X knew how to operate and we did not have the time to
integrate them in a ready-to-share implementation before he left.

Thanks for your interests in this paper. Unfortunately, the server in which
my implementation was stored had a disk crash in April and three disks
crashed simultaneously.

The code owned by COMPANY, and AFAIK the code is not open-source. Your best
bet is to reimplement :( Sorry.
```

2. REPRODUCIBILITY

AI research is also lacking in openness: Only 15% of papers published their code June 2020 (increasing to 26% 2021)

Research paper code implementations are important for accountability, reproducibility and driving progress in AI. Academic groups are more likely to publish their code than industry groups. Notable organisation that don't publish all of their code are OpenAI and DeepMind. For the biggest tech companies, their code is usually intertwined with proprietary scaling infrastructure that cannot be released. Some efforts to improve rate of code publication, e.g., Papers with Code (Facebook AI), or ML Reproducibility Challenge.



Papers With Code

from stateof.ai 2020

From N. Benaich and I. Hogart, State of AI Report 2020

2. REPRODUCIBILITY

Irreproducible experiments

a. Neutrinos travelling faster than light

In March 2011, scientists reported first evidence that neutrinos produced at CERN in Geneva recorded at the OPERA detector at Gran Sasso, Italy, had traveled faster than light, arriving 60.7 nanoseconds sooner than light traversing the same distance in vacuum.

After six months of cross checking, in September 2011, researchers announced that neutrinos had been observed traveling faster than light, with 0.2 in a million chance of a false positive (significance of six sigma).

By June 2012, five different groups had reproduced the experiment without finding any anomaly, and the two leaders of the 170 member experimental group had resigned. The result was most likely the effect of a loose optical cable and an incorrectly calibrated clock.

2. REPRODUCIBILITY - a notorious irreproducible experiment

b. Cold fusion

In 1989, electrochemists Martin Fleischmann and Stanley Pons gave a press conference at the University of Utah announcing that they had tamed the power of nuclear fusion in an electrolysis cell.

They assumed that deuterium atoms from heavy water had penetrated into the palladium cathode and fused to form helium atoms. Palladium absorbs so much hydrogen or deuterium that some atoms could be pushed near enough to each other to cause fusion events. The excess energy was then dissipated as heat.

Fleischmann and Pons claimed that this could not be caused by any known chemical reaction, and invented the nuclear reaction term “cold fusion”.
If confirmed, the discovery could have transformed the global energy landscape.

Pons and Fleischmann, and others caught up in the large global excitement, broke several norms for responsible scientific conduct:

1. An editor at Journal of Electroanalytic Chemistry allowed the original article to be published with minimal peer review.

2. The University of Utah organized a press conference before the paper was published, where the scientists and university officials emphasized the amount of energy the fusion cells could produce in the future.

3. Pons and Fleischmann withheld experimental details from the community and tried to shield their ideas from testing.

4. They and others who claimed to reproduced cold fusion, only to later retract their results, failed to perform adequate and reasonable tests to evaluate their ideas.

5. The researchers’ behavior was dishonest in other ways, for example by breaking an agreement on publication together with another researcher, and submitting on their own without his knowledge.

NB - It is important to note that even with such unscientific behavior, the process of science still worked. Within a year, the scientific community had investigated Pons and Fleischmann's claims and come to the consensus that what had been observed was not cold fusion.

2. REPRODUCIBILITY - Irreproducible experiments?

Large scale AI projects - irreproducible or not?

Large scale AI projects are done at enormous costs by a few actors with enormous resources (Google etc). Can these projects be viewed as reproducible science?

1. **AlphaGo**

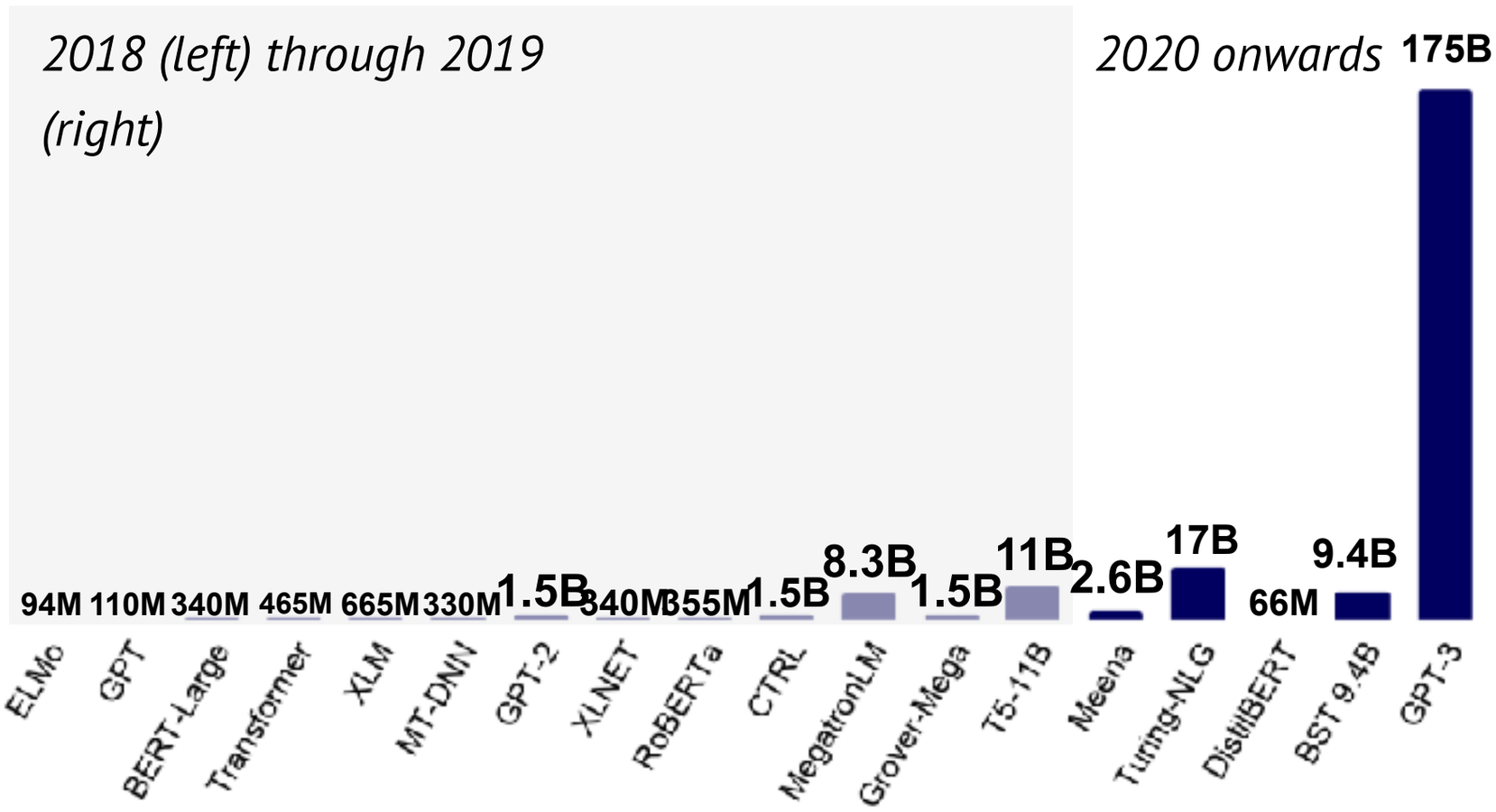
AlphaGo developed by DeepMind (Google) was the first program to achieve superhuman strength at the game of Go through a combination of self-play reinforcement learning and more traditional techniques. This was by most experts considered unlikely to happen anytime soon.

One version of AlphaGo was essentially reproduced by researchers at Facebook AI in 2019 in a large scale project (using 2000 GPUs for self-play for example).

2. REPRODUCIBILITY - Irreproducible experiments?

Large scale AI projects - irreproducible or not?

Huge models, large companies and massive training costs dominate the hottest area of AI today, natural language processing.



Number of parameters the algorithm optimizes during the training process.

The figure above is already obsolete - the largest models from 2021 can use more than 10^12 (tera-) parameters

From N. Benaich and I. Hogart, State of AI Report 2020

from stateof.ai 2020

2. REPRODUCIBILITY - Irreproducible experiments?
Large scale AI projects - irreproducible or not?

Tuning billions (or trillions!) of model parameters costs millions of dollars

Based on variables released by Google et al., you're paying circa $1 per 1,000 parameters. This means that OpenAI's 175B parameter GPT-3 could have cost tens of millions to train. Experts suggest the likely budget was $10M.

Just how much does it cost to train a model? Two correct answers are "depends" and "a lot". More quantitatively, here are current ballpark list-price costs of training differently sized BERT [4] models on the Wikipedia and Book corpora (15 GB). For each setting we report two numbers - the cost of one training run, and a typical fully-loaded cost (see discussion of "hidden costs" below) with hyper-parameter tuning and multiple runs per setting (here we look at a somewhat modest upper bound of two configurations and ten runs per configuration).[4]

- $2.5k - $50k (110 million parameter model)
- $10k - $200k (340 million parameter model)
- $80k - $1.6m (1.5 billion parameter model)

For example, based on information released by Google, we estimate that, at list-price, training the 11B-parameter variant[5] of T5 [5] cost well above $1.3 million for a single run. Assuming 2-3 runs of the large model and hundreds of the small ones, the (list-)price tag for the entire project may have been $10 million[6].

Not many companies – certainly not many startups – can afford this cost. Some argue that this is not a severe issue; let the Googles of the world pre-train and publish the large language models, and let the rest of the world fine-tune them (a much cheaper endeavor) to specific tasks. Others (e.g., Etchemendy and Li [6]) are not as sanguine.

from stateof.ai

2020

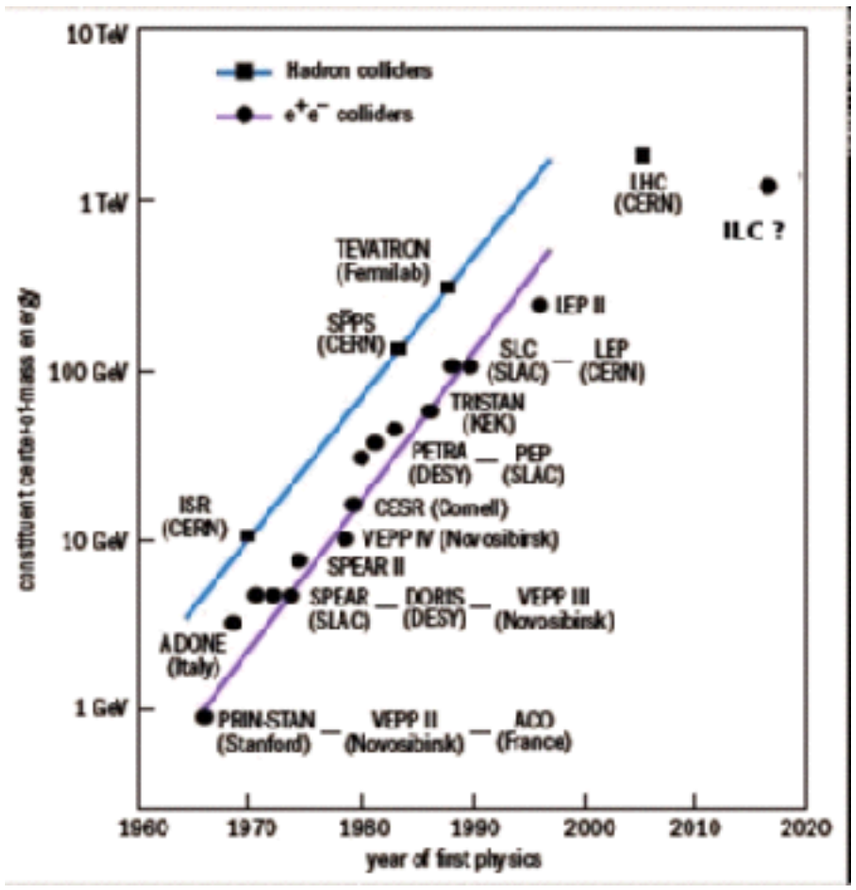
From N. Benaich and I. Hogart, State of AI Report 2020

2. REPRODUCIBILITY - Irreproducible experiments?

The cost of studies limits reproducibility, or even the progress of science in general.

a. particle physics - stagnating science last 40 years

b. no manned exploration of moon or space beyond earth orbit since 1972 (Artemis III probably next mission in 2025)



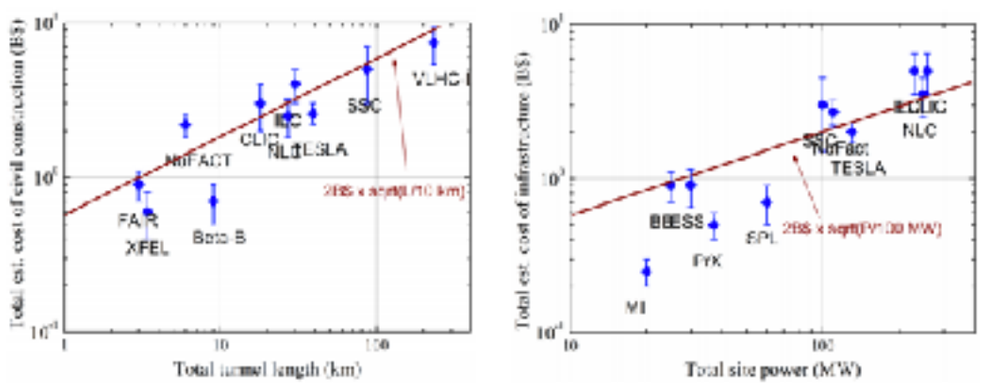


Figure 1. a) left — total estimated cost of the civil construction for accelerator facilities vs cumulative length of their tunnels; b) right — total estimated infrastructure cost of the accelerator facilities vs their electric power consumption.

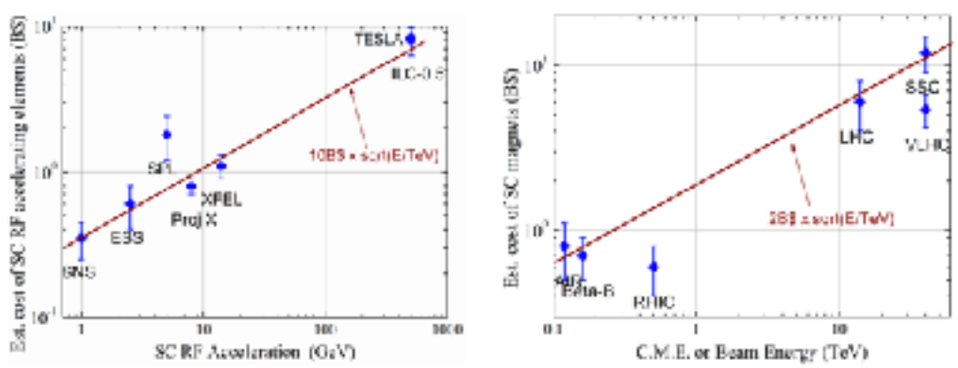


Figure 2. a) left — estimated cost of the accelerating elements for accelerator facilities based on SC RF vs total beam acceleration; b) right — estimated cost of the SC magnets and associated elements vs collider center of mass energy / single beam energy.

2. REPRODUCIBILITY CRISIS

Reproducibility has recently become a hotly debated issue in many other scientific fields

- psychology
- medicine
- empirical economics
- and many others (computer science?)

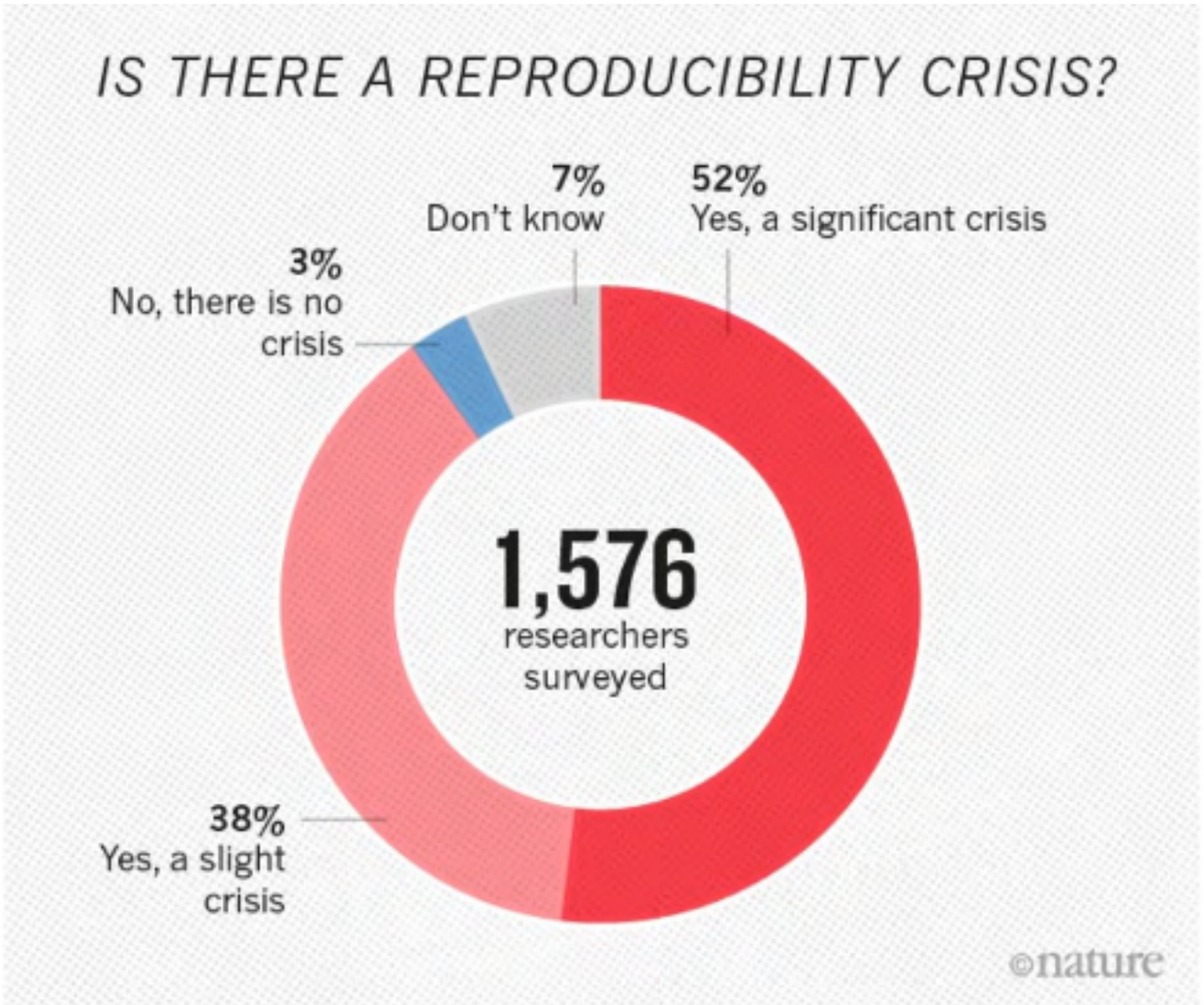
A number of larger systematic studies aimed at repeating well-known experiments published in the literature have recently been carried out.

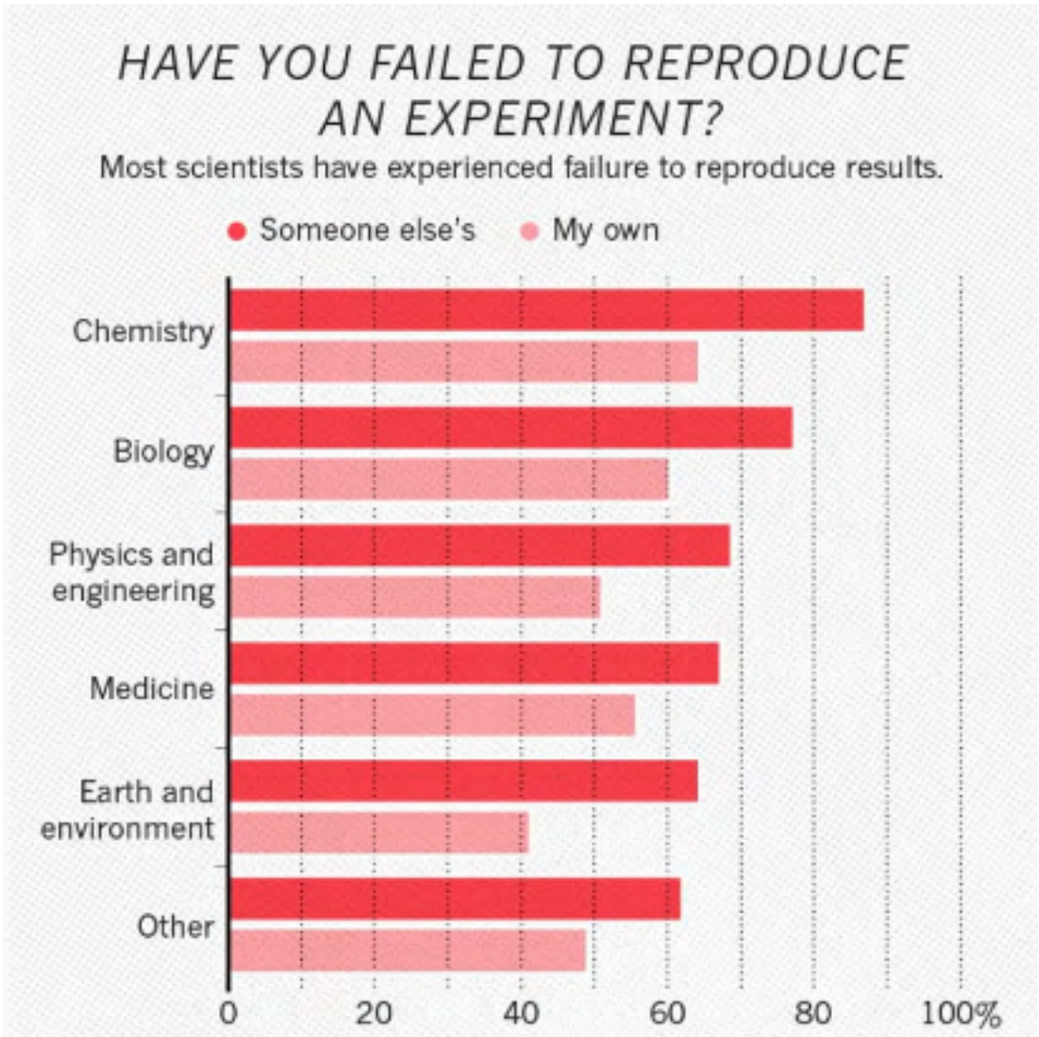
The results have often been less than satisfactory - in some studies less than half of the studies could be replicated.

Some call this the *replication crisis (or reproducibility crisis)* in science

2. REPRODUCIBILITY

A survey in Nature in 2016 indicated that many scientists in different fields are concerned about the reproducibility crisis:





From M. Baker, 1500 scientists lift the lid on reproducibility, Nature 25 May 2016.

2. REPRODUCIBILITY

Some of the reasons for the replication crisis lie in the **researcher's degrees of freedom** which may result in **questionable research practices**:

- sample size (often too small)
- treatment of outliers
- inclusion of more data
- HARK-ing
- p-hacking
- data fishing
- mid-experiment adjustments
- incomplete reporting of experimental conditions
- lacking statistical power and confidence limits

2. REPRODUCIBILITY - replication crisis in psychology

Two examples of large replication studies:

a. Many Labs 2 experiment (2018) - around 200 psychology researchers worldwide attempted to reproduce 28 well-known and highly cited published experiments. 15305 subjects from 36 countries participated. Examples of experiments included:

- whether moral violations can induce a desire for cleansing (Lady Macbeth)
- higher number of siblings make people more cooperative
- being in a consumer mind-set would reduce trust in other people
- the thinking of Westerns is more rule based than that of Asians
- etc

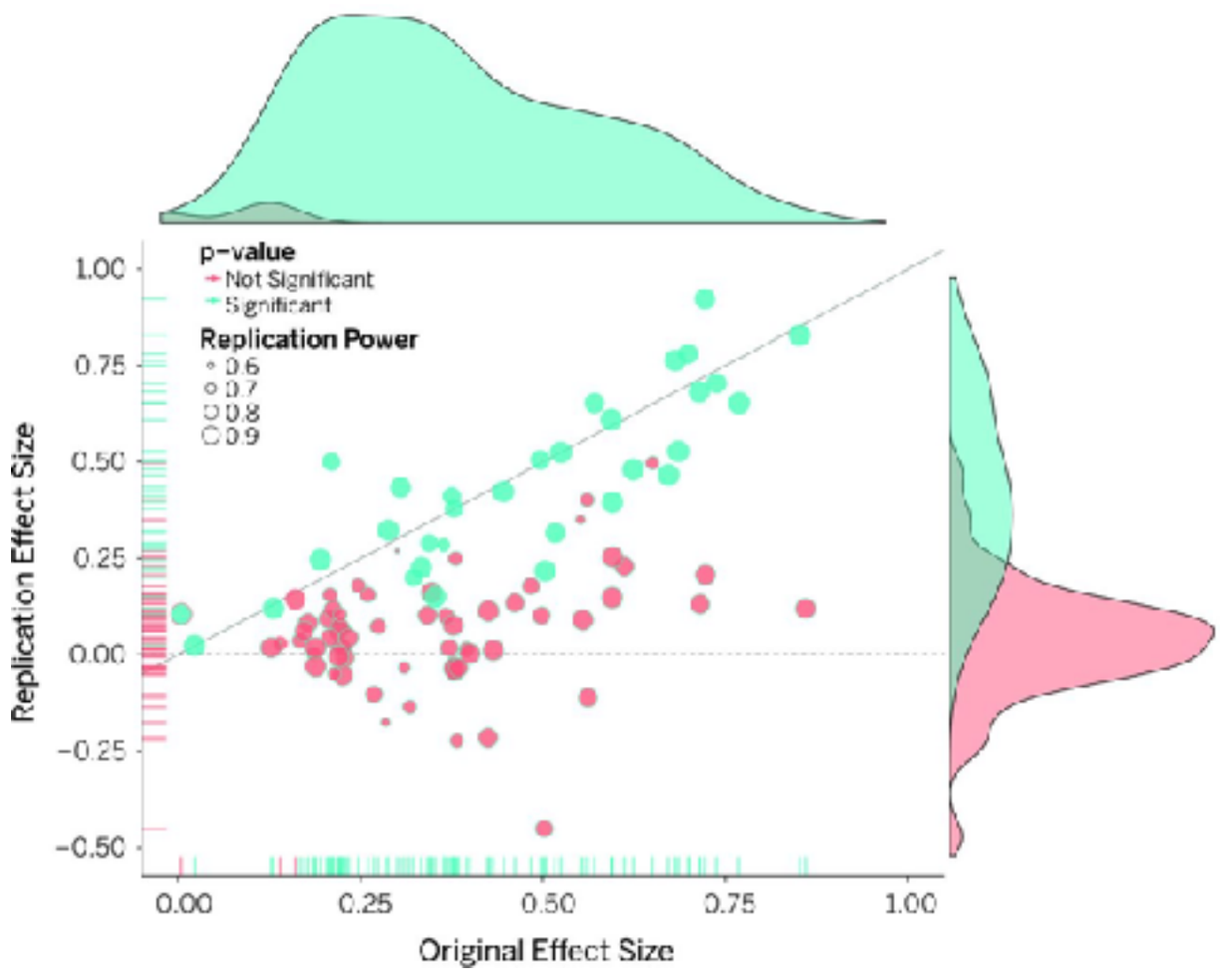
Only 14 out of 28 experiments could be replicated successfully, or depending on significance criterion, 15/28 if $p<0.05$ and 14/28 if $p < 0.0001$.

2. REPRODUCIBILITY - replication crisis in psychology

b. Open Science Collaboration 2015 (Aarts et al., Science, 2015)

Replications of 100 experimental studies published in three psychology journals in 2008 using high-powered designs and original materials when available. Replication effects were half the magnitude of original effects, representing a substantial decline.

Ninety-seven percent of original studies had statistically significant results. Thirty-six percent of replications had statistically significant results; 47% of original effect sizes were in the 95% confidence interval of the replication effect size; 39% of effects were subjectively rated to have replicated the original



2. REPRODUCIBILITY - replication crisis in psychology

An example - the facial feedback hypothesis

William James (1842-1910) — a famous American psychologist and philosopher — proposed that bodily sensations are not a consequence of our emotions, but rather the *cause* of the emotions.

Example - you meet a large bear in the forest, your pulse races and you run. Then it is the pulse and the running that causes the fear.

We feel sorry because we cry, angry because we strike, afraid because we tremble (W James, 1884)

He who gives way to violent gestures will increase his rage; he who does not control the signs of fear will experience fear in a greater degree,..
(Charles Darwin, The Expression of the Emotions in Man and Animals, 1872)

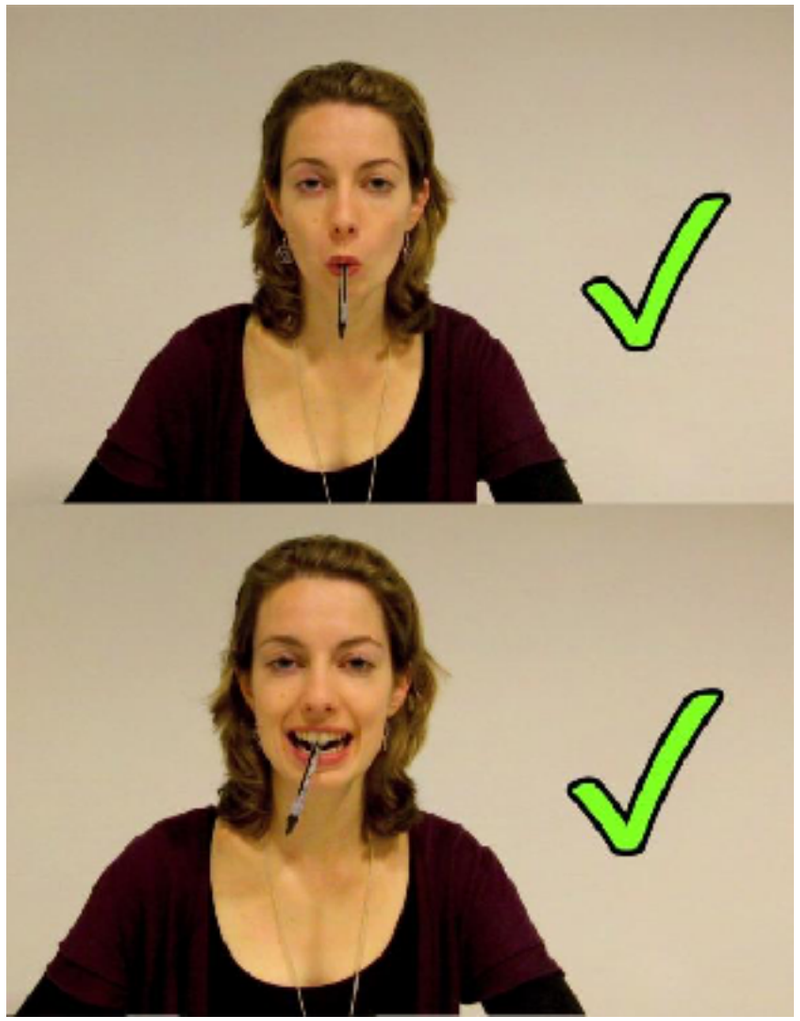
Facial feedback hypothesis - making a smile makes you feel happier.

Highly cited paper by Strack et al. (1988). Participants looking at cartoons while holding a pen between their teeth, forcing them to smile, or between their lips, forcing them to pout. This helped in deveceiving the participating students about the purpose of the experiment, and improved on earlier studied in that way.

Those in the smile condition said they found the cartoons funnier (scores 5.1 vs 4.3).

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*(5), 768–777. (2327 citations)

Replicated by 17 labs, with a total of 1784 students. The original study reported a rating difference of 0.82 units on a 10-point Likert scale (1-10). The meta-analysis revealed a rating difference of 0.03 units with a 95% confidence interval of [−0.11,0.16].



2. REPRODUCIBILITY - replication crisis in medicine

Reproducibilty has also been questioned in many areas of medicine.

Some examples:

Well-known paper - John P. A. Ioannidis, Why Most Published Research Findings Are False (2005)

In 2012, researchers at the biotechnology firm Amgen in Thousand Oaks, California, announced that they had failed to replicate 47 of 53 landmark cancer papers (commercial study, not published).

The Reproducibility Project: Cancer Biology - replication of 18 well-known studies from 2010-12.
5 fully reproduced, 6 partially, 6 non-replicable so far.

2. REPRODUCIBILITY - attempts to address reproducibility in CS (machine learning)

NeurIPS (major conference) - appointed Chair of Reproducibility in 2019 (Joelle Pineau, McGill)

Submissions with code have gone from 50% 2018 to 75%. Availability of code actually had a strong correlation to reviewers recommending the paper for submission.

1. ML Reproducibility Challenge (since 2017)

2. Machine Learning Reproducibility Checklist (Joelle Pineau, McGill)

The Machine Learning Reproducibility Checklist (v2.0, Apr.7 2020)

For all **models** and **algorithms** presented, check if you include:

- ❑ A clear description of the mathematical setting, algorithm, and/or model.
- ❑ A clear explanation of any assumptions.
- ❑ An analysis of the complexity (time, space, sample size) of any algorithm.

For any **theoretical claim**, check if you include:

- ❑ A clear statement of the claim.
- ❑ A complete proof of the claim.

For all **datasets** used, check if you include:

- ❑ The relevant statistics, such as number of examples.
- ❑ The details of train / validation / test splits.
- ❑ An explanation of any data that were excluded, and all pre-processing step.
- ❑ A link to a downloadable version of the dataset or simulation environment.
- ❑ For new data collected, a complete description of the data collection process, such as instructions to annotators and methods for quality control.

For all shared **code** related to this work, check if you include:

- ❑ Specification of dependencies.
- ❑ Training code.
- ❑ Evaluation code.
- ❑ (Pre-)trained model(s).
- ❑ README file includes table of results accompanied by precise command to run to produce those results.

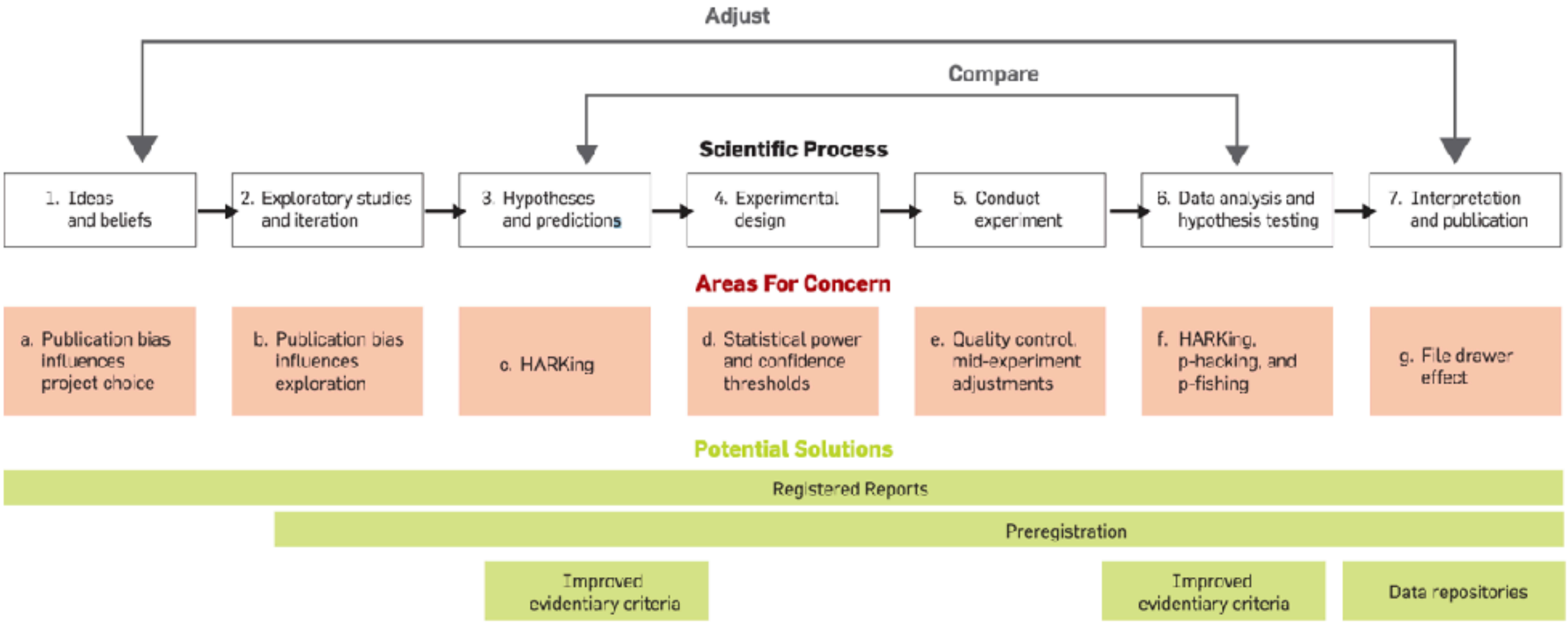
For all reported **experimental results**, check if you include:

- ❑ The range of hyper-parameters considered, method to select the best hyper-parameter configuration, and specification of all hyper-parameters used to generate results.
- ❑ The exact number of training and evaluation runs.
- ❑ A clear definition of the specific measure or statistics used to report results.
- ❑ A description of results with central tendency (e.g. mean) & variation (e.g. error bars).
- ❑ The average runtime for each result, or estimated energy cost.
- ❑ A description of the computing infrastructure used.

Reproduced from: ***www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist-v2.0.pdf***

2. REPRODUCIBILITY - what causes irreproducibility and what can be done?

Figure 1. Stages of a typical experimental process (top, adapted from Gundersen[18]), prevalent concerns at each stage (middle), and potential solutions (bottom).



From A. Cockburn et al, Communications of the ACM, August 2020, vol 63, no 8.

2. REPRODUCIBILITY - Publication bias and file drawer effect

Publication bias means:

Papers support their hypothesis are accepted for publication at a much higher rather than those who do not.

File drawer effect means:

Null findings tend to be unpublished and therefore hidden from the research community

Questionable research practices and results of publication bias

a. Publication bias negatively influences project selection

Publication bias is likely to draw researchers towards safer topics in which outcomes are more certain.

Publication bias discourages replication. A successful replication is likely to be rejected because it replicates what is already ‘known’

b. Publication bias discourages exploratory research

Not all research is alike. Exploratory studies play a very important role in the scientific process. If all studies are expected to obey the same standards important developments may be hindered.

.

c. HARK-ing: Hypothesising After the Results are Known.

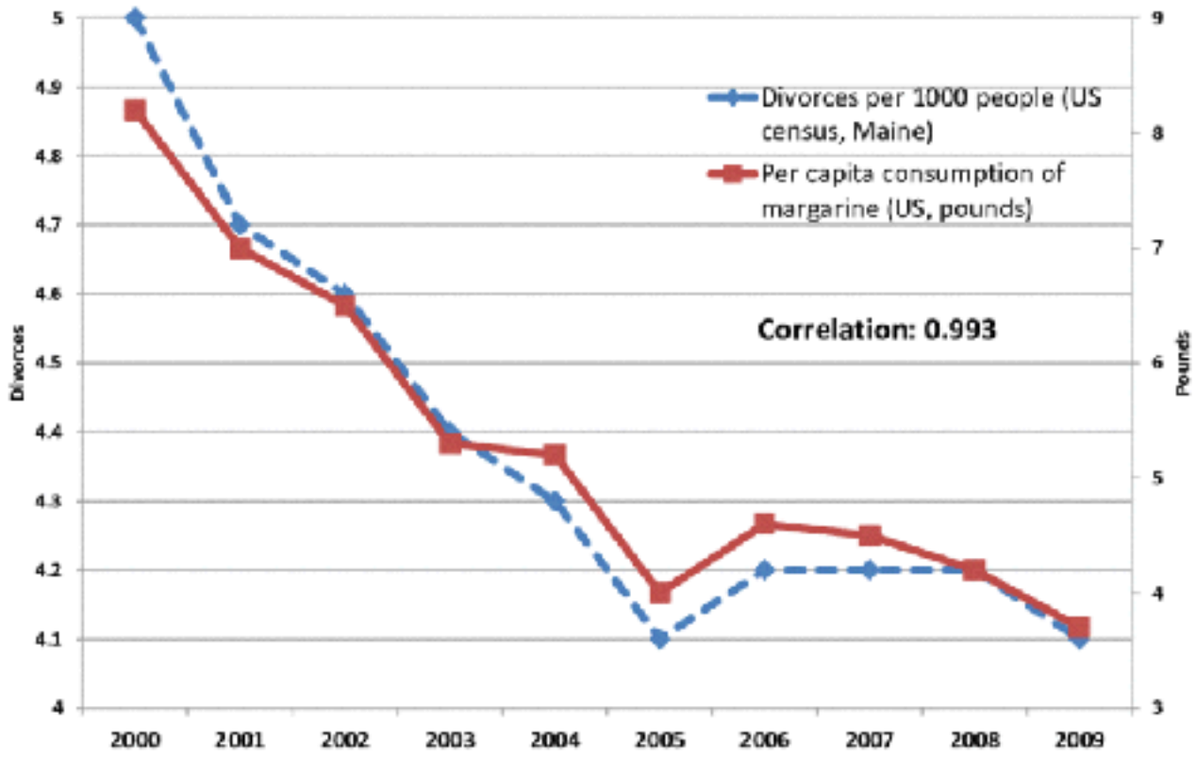
Explore hypotheses that are different than those that they originally set out to test (outcome switching).

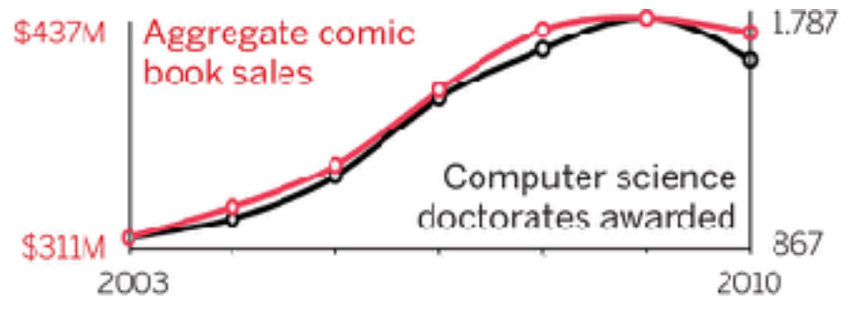
Experimenters normally record a wide set of experimental data beyond that required to test the hypothesis - this is normal in order to understand and interpret the experiment.

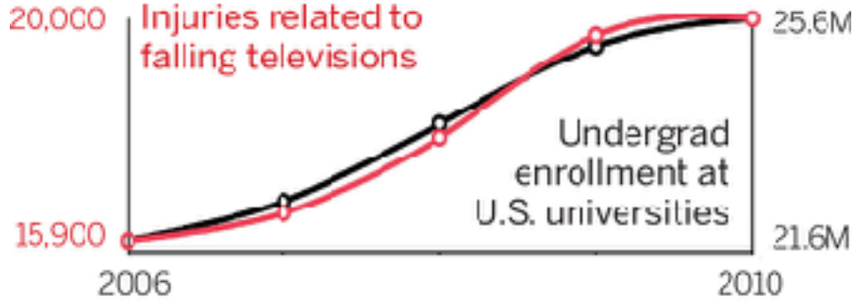
It is then tempting to adjust the hypothesis according to the outcome (consciously or unconsciously).

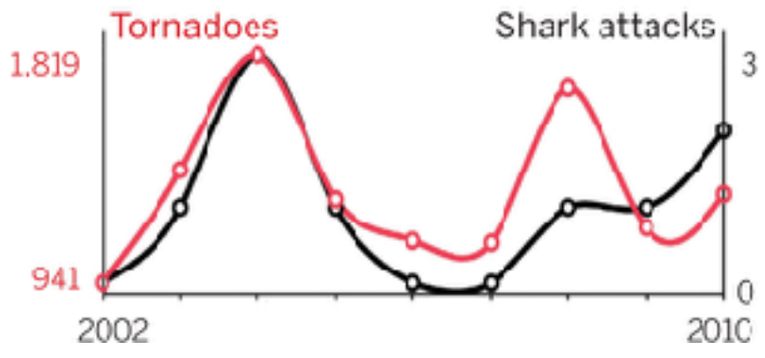
d. Data fishing (exploring multiple hypotheses)

Example - Spurious correlations





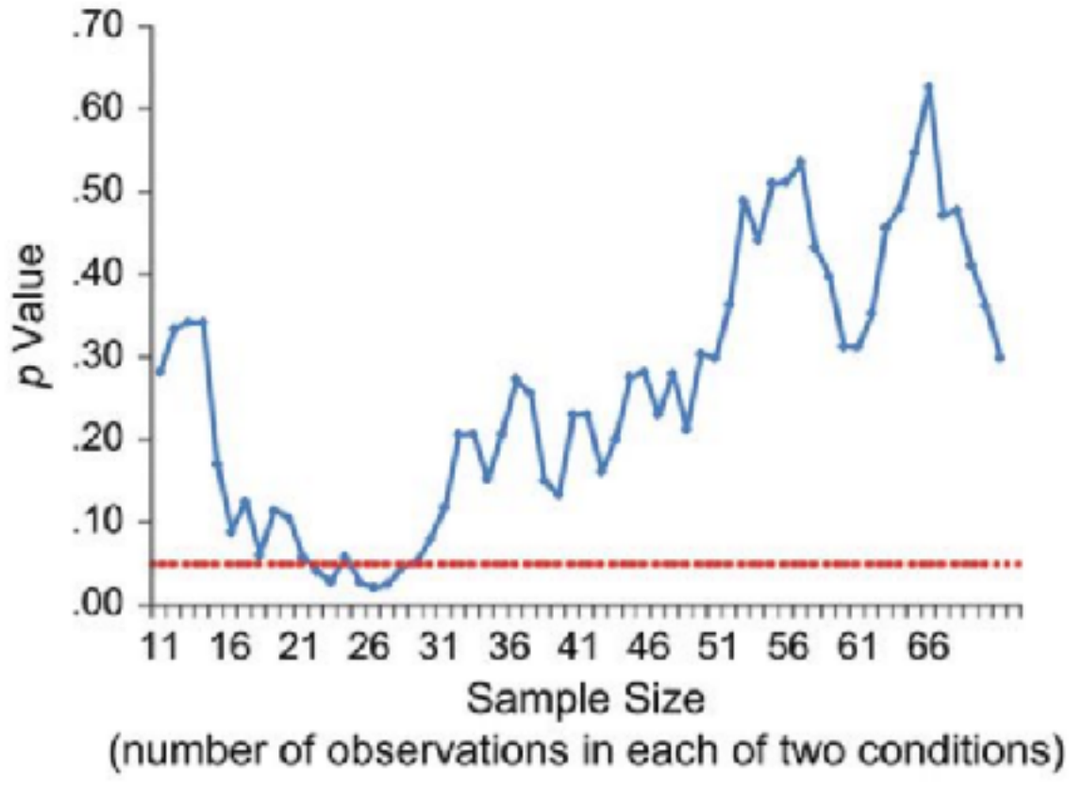




e. P-hacking

P-hacking is a term for when researchers collect or select data or statistical analyses during an experiment, and stop when nonsignificant results become significant (e.g., p-value < 0.05).

Gradually adding more data during an experiment can lead to spurious significant results:



Source: Simmons, Nelson and Simonsohn (2011)

f. Statistical power and confidence limits

Sample sizes may be too small, and effects of sample size sometimes not explored enough.

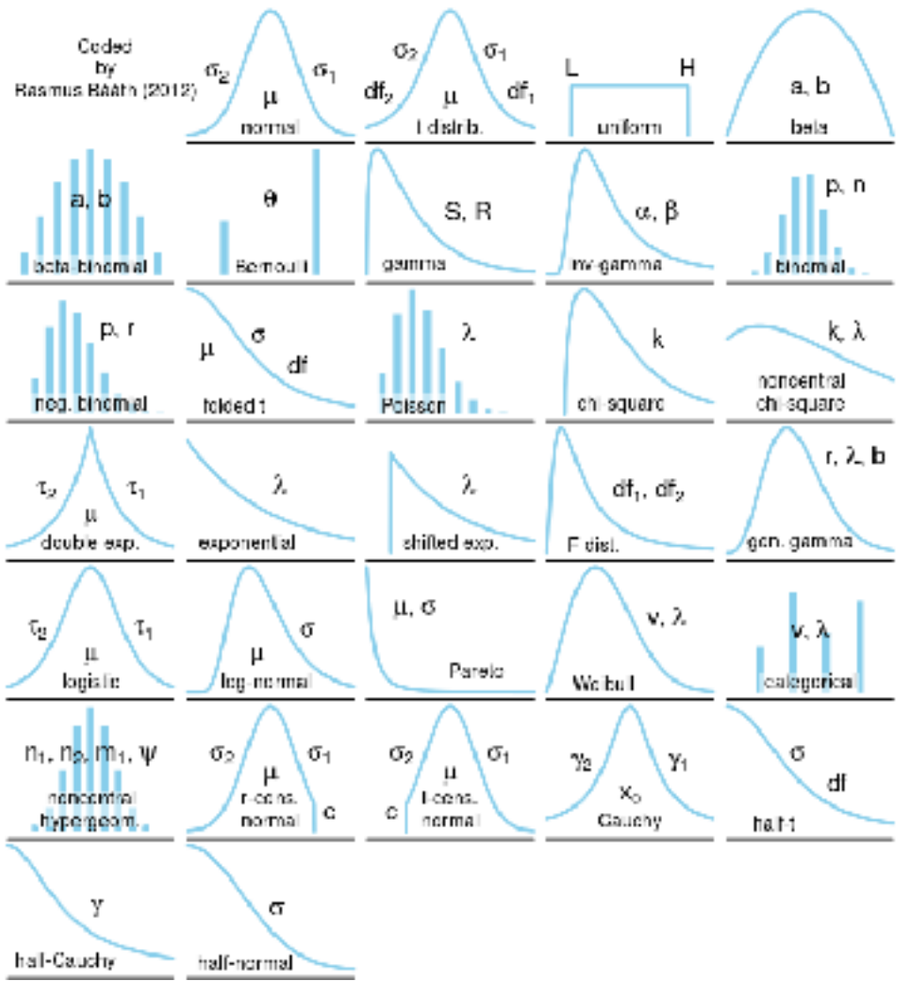
Statistical power estimates that determine minimal sample size have not been performed (or cannot be done).

In many cases, statistical significance should be defined more stringently. Routinely assuming that $p<0.05$ means statistical significance may be misleading. Recent suggested solutions in CS:

- Change definition of statistical significance to $p<0.005$
- Abandon dichotomous definition completely
- Adopt Bayesian statistics (prior?)
- Let readers form their own conclusion on significance

Statistics for experiments

Many different probability distributions occur regularly in data from real world applications. Some theory relating to using these in experiments, e.g., in hypothesis testing, was given in Lecture 7. Here we will expand this with some more examples in order to connect to real world applications. When writing a master's thesis (for the large majority who include some form of experiment in the thesis), be prepared to go back to the literature from your course in mathematical statistics for further detail!



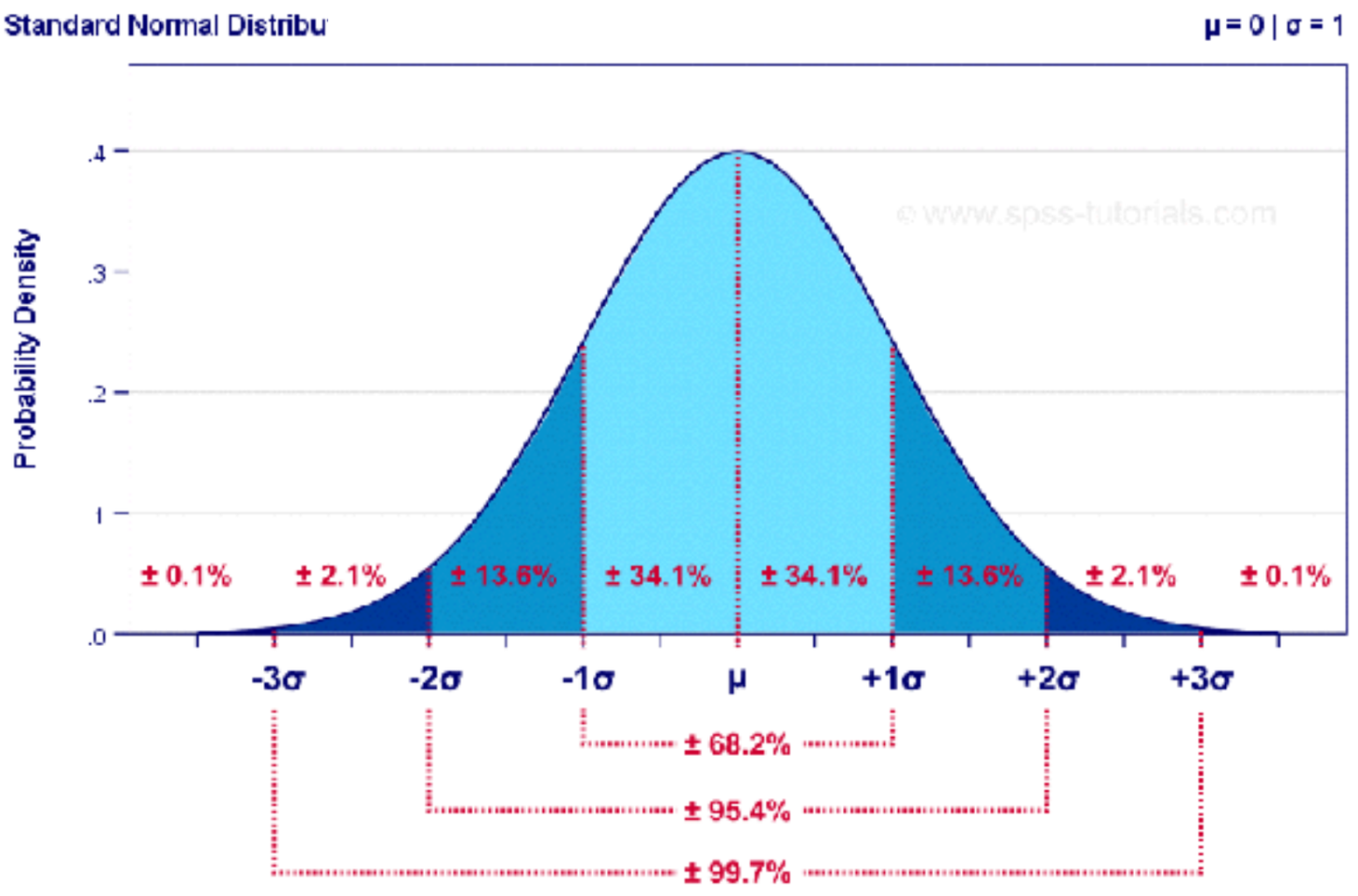
Probability distributions

Experimental data can follow many different distributions!

1. The normal distribution

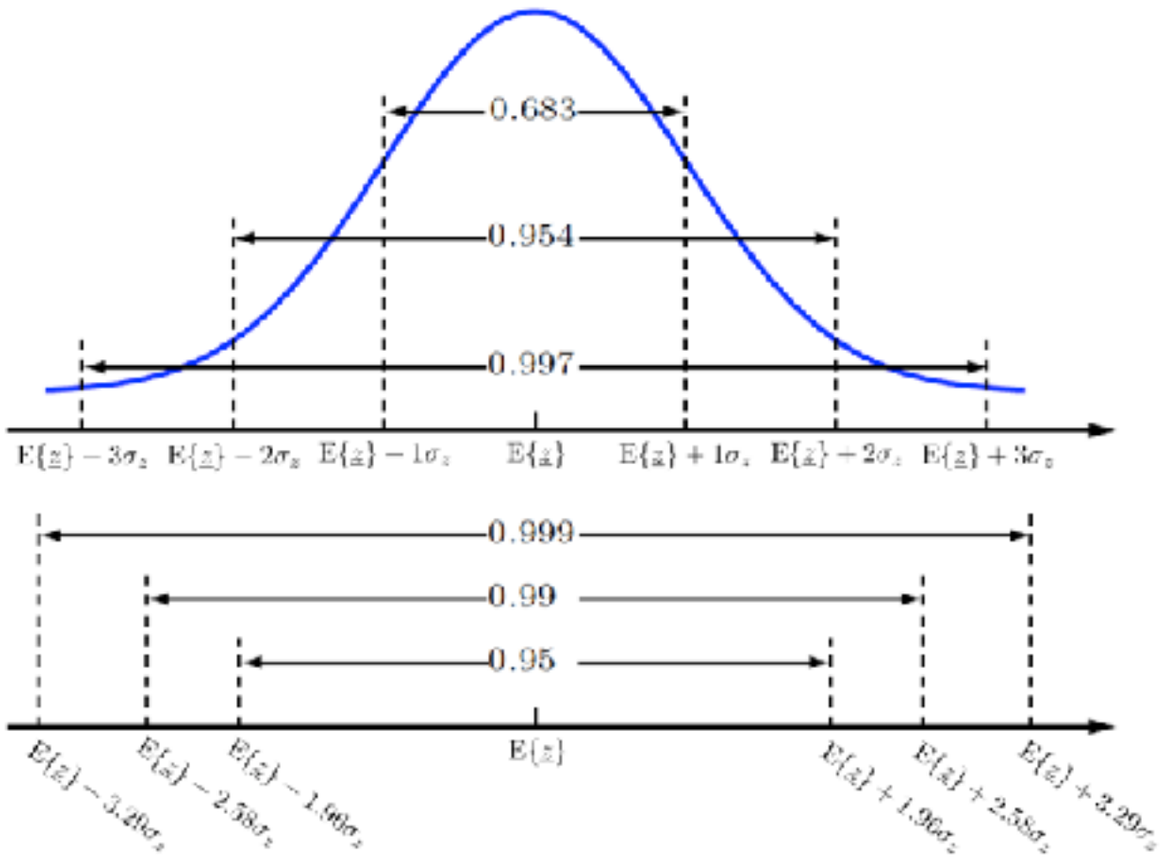
The normal distribution often appears in a situation where something results as a *sum of many random contributions*.

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



The normal distribution:

Confidence intervals:



Example: A 95% confidence interval covers 1.96 standard deviations on both sides of the mean.

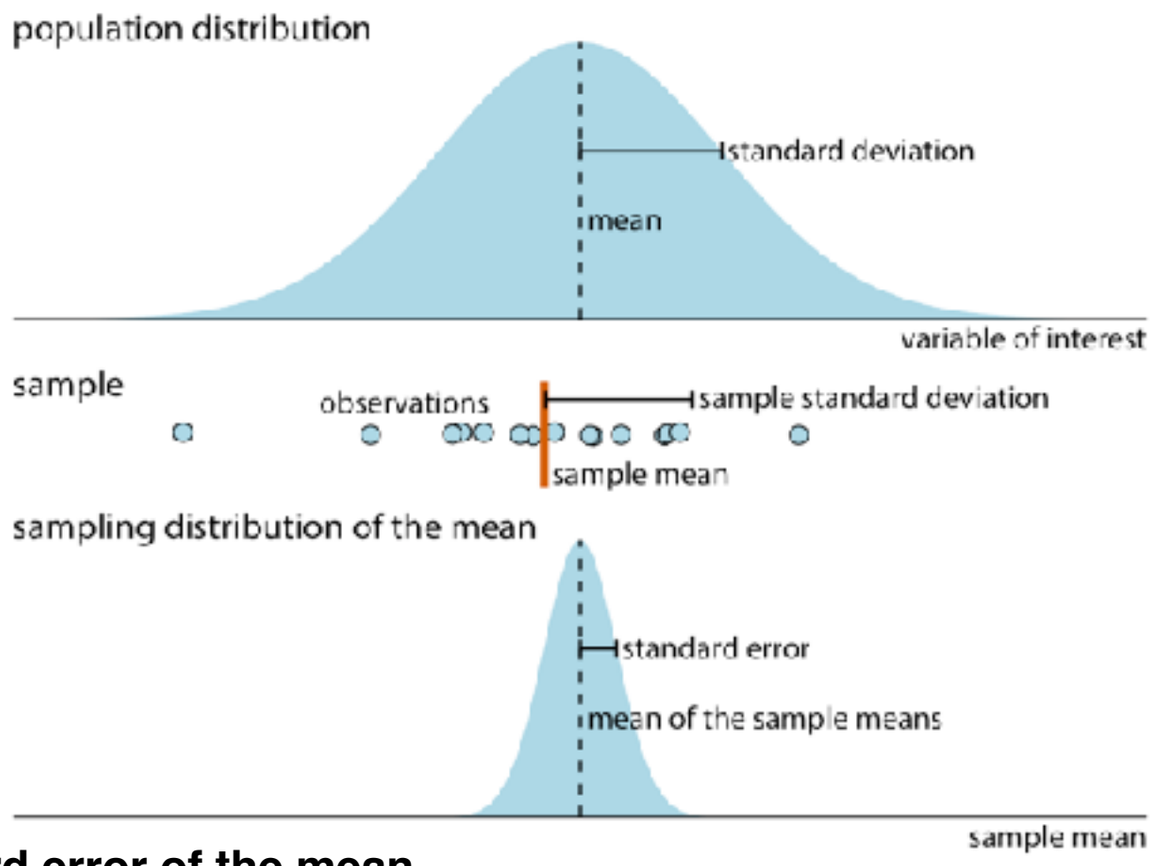
Central limit theorem:

If $X_1, X_2, \ldots, X_n$ are n random samples drawn from a population with overall mean μ and finite variance σ^2, and $\bar{X}_n$ is the sample mean, then the limiting form of the distribution

$Z = \lim_{n\to\infty} \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$ is a standard normal distribution (mean 0 and standard deviation 1)

Sampling a probability distribution

In an experiment, the result is normally a sample of values where the distribution is not known with certainty. Sometimes one may have reasonable ground to assume a certain distribution, for example for the mean value of a large number of observations due to the central limit theorem.



Standard error of the mean

- estimate of the accuracy of a result calculated as a mean value (see HW10)
- applies to estimated mean of values from any distribution if the number of values is large enough

$$SE_x = \frac{s}{\sqrt{n}} \qquad s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Other probability distributions

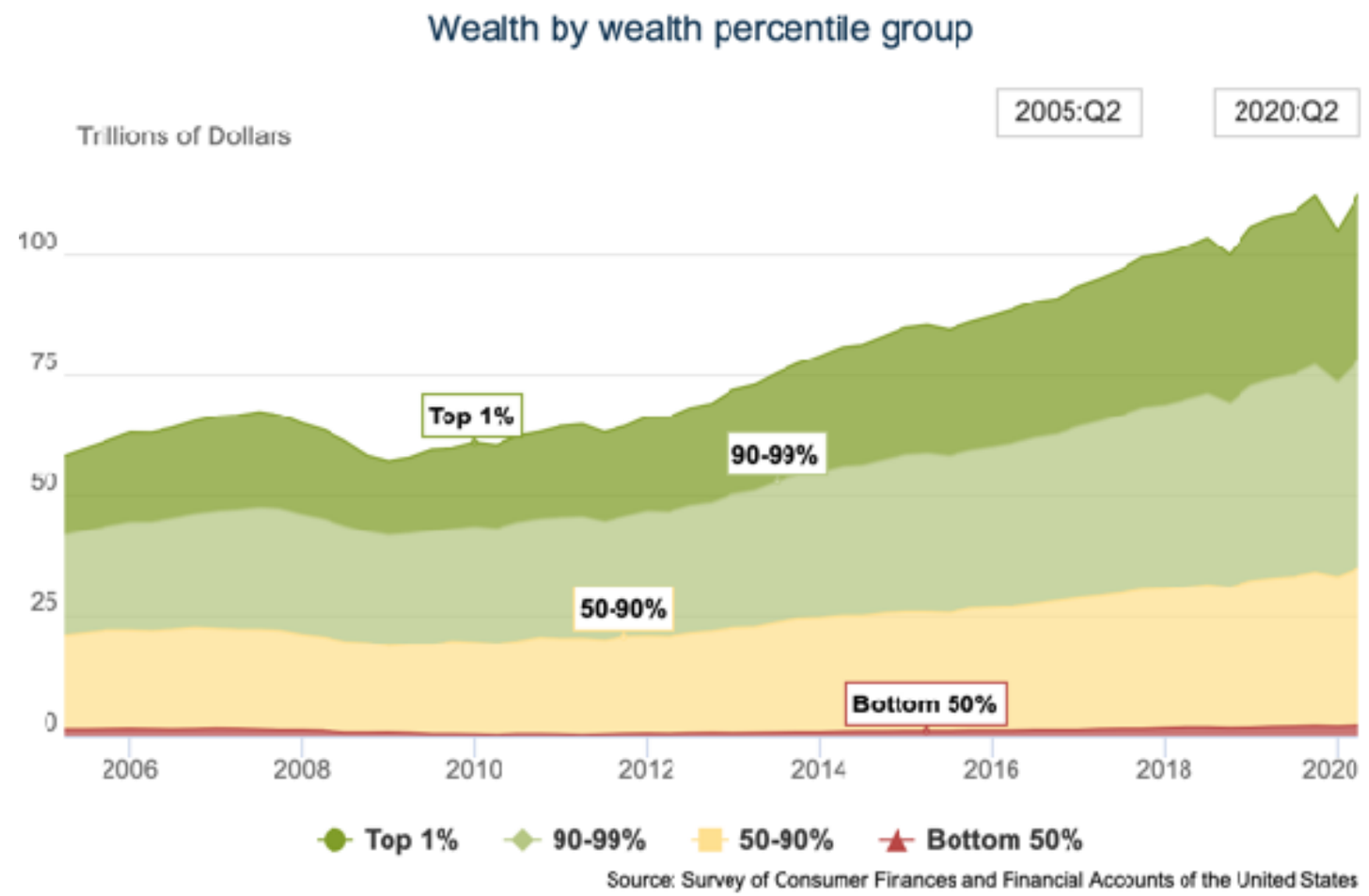
Many other probability distributions also occur regularly in data from real world applications:

- Pareto distribution - power law distribution (related to Zipf's law from HW)
- Product of random contributions? (e.g., investments) - lognormal distribution
- Survival processes (how long until my car breaks down) - e.g., Weibull distribution
- Discrete distributions - e.g., binomial distribution
- Exponential distribution - time intervals in arrival process, queueing theory
- Etc

Skewed distributions

Many distributions are very far from normal e.g. US wealth distribution (many other countries similar). The mean value may not be very representative: median net worth of US household: $97,300 average net worth of $692,100!

Median values are often more suitable for skewed distributions, but less suited to analytic calculations



How much do these individuals shift the average?
(9/10 studied at least some CS, physics or math)

Examples of skewed distributions

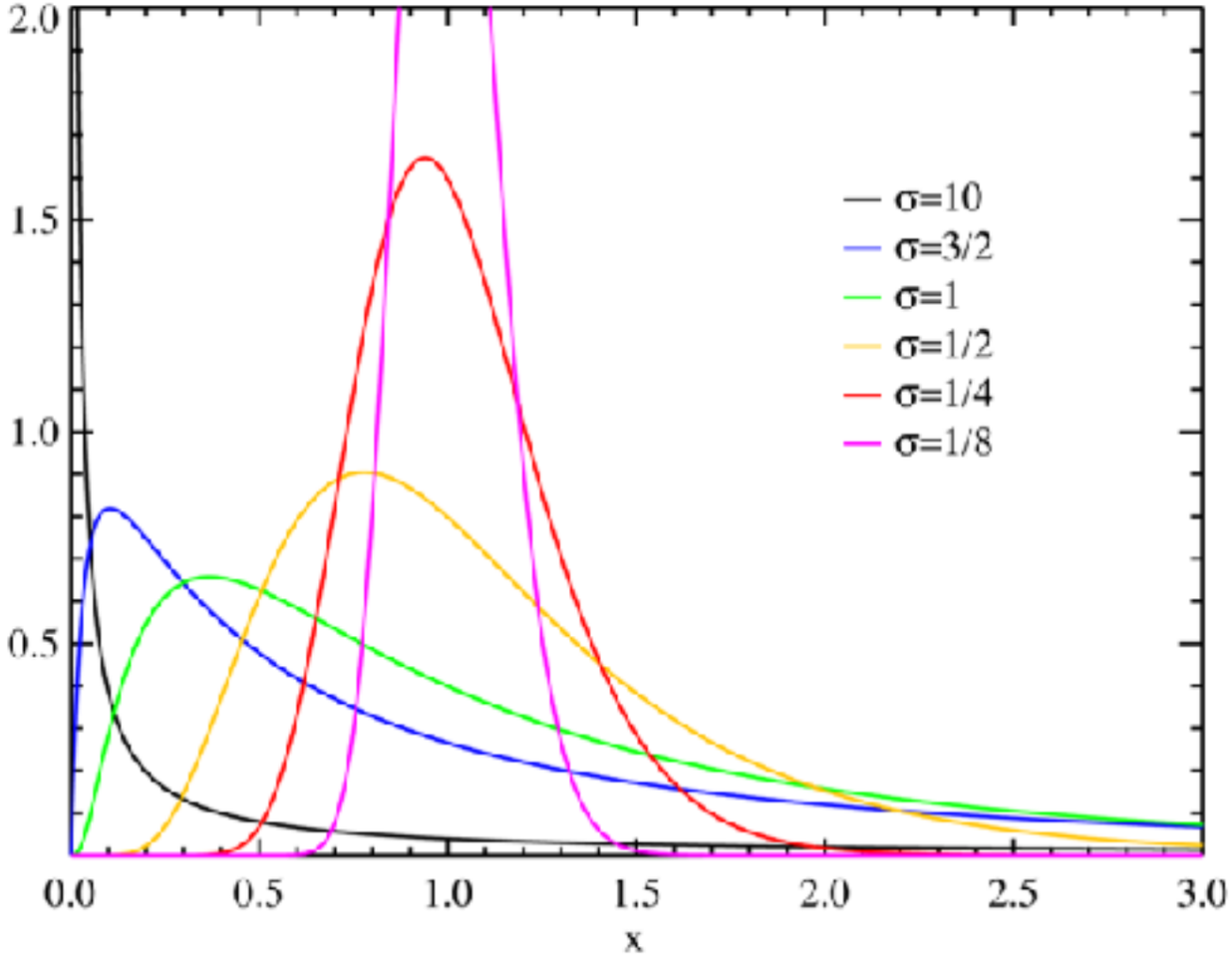
a. The lognormal distribution:

The logarithm of X (rather than X itself) follows a normal distribution, or in other words, this is the probability distribution of exp(X), where X follows a normal distribution

Defined for positive X only.

Can arise as a *product of many random values*, for example in the gradual growth of an investment, leading to a lognormal wealth distribution.

The figure shows the distribution of exp(N(0,sigma) for different sigma.



Examples of skewed distributions

b. The Weibull distribution:

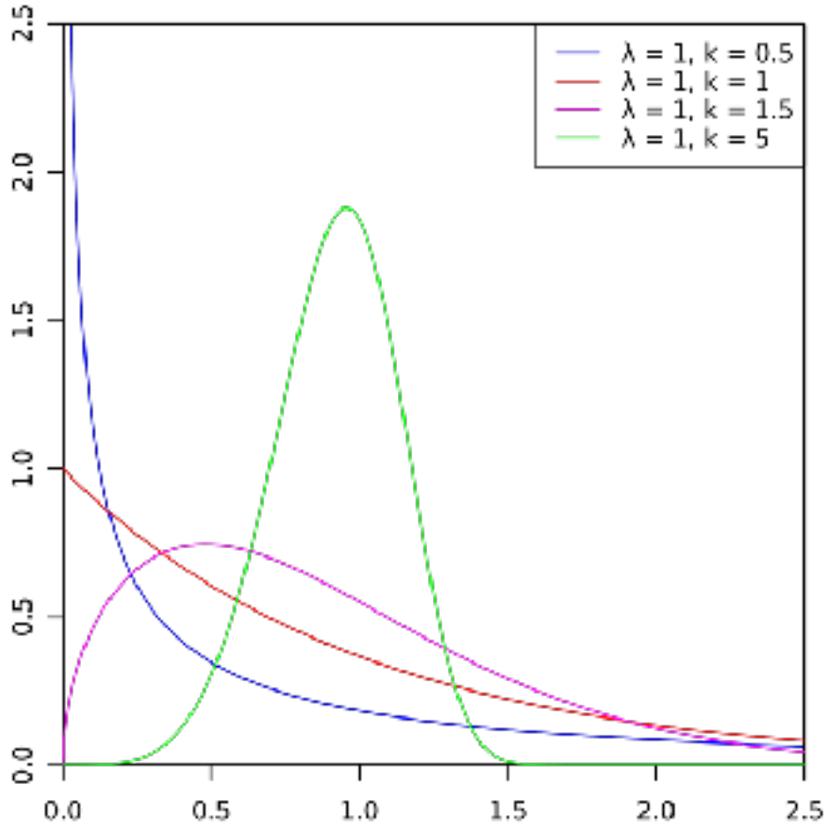
A two parameter distribution:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

Defined for positive X only.

Often used in analysis of reliability to described the distribution of time to failure
(e.g., failure of components in a complex engineering system such as a truck or an airplane).

k= 1 is the exponential distribution. If the parameter k<1, the failure rate decreases with time, if k=1 it is constant, if k>1 it increases.



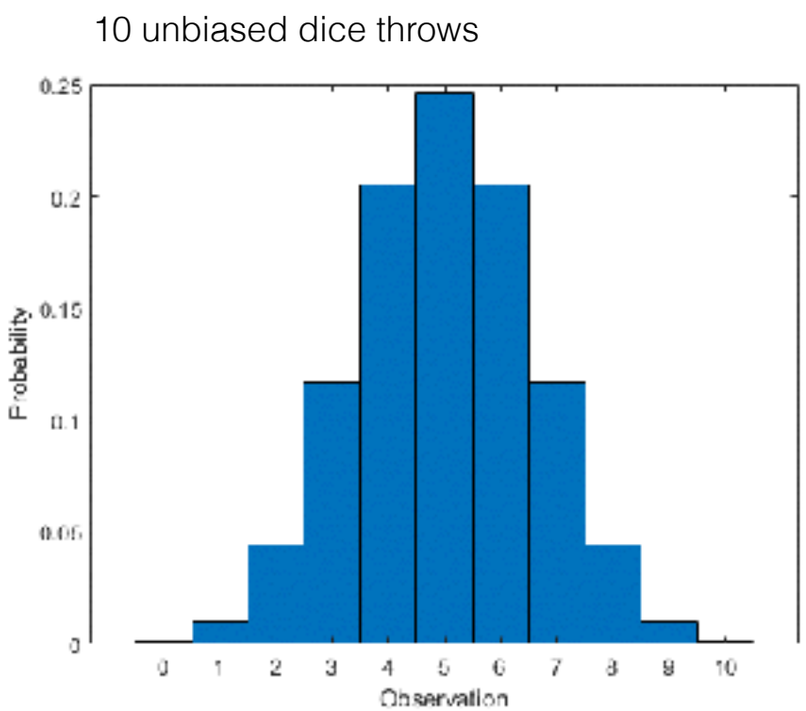
The binomial distribution

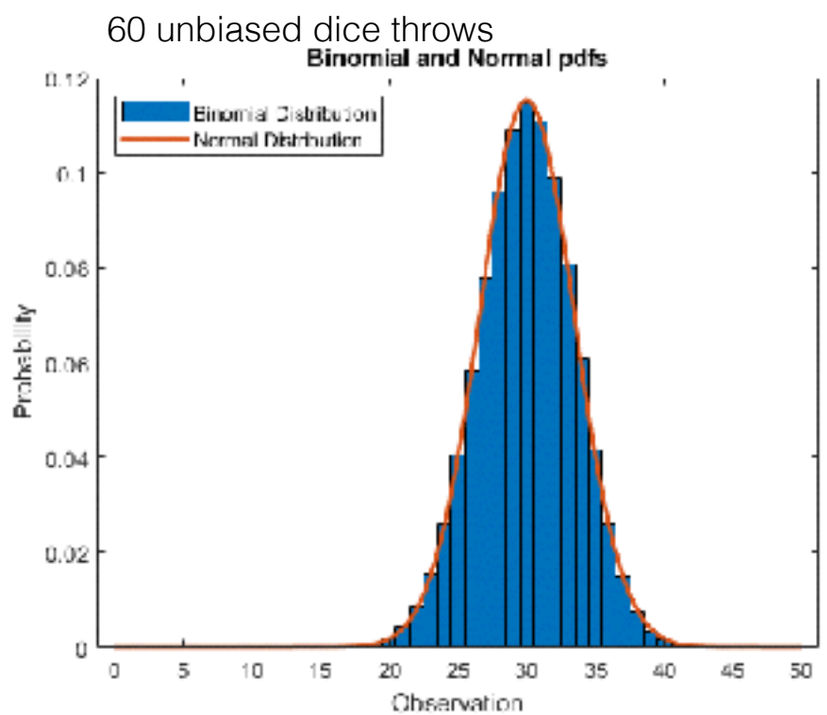
$$X \in \mathrm{Bin}(n, p)$$

Binomial distribution: two alternatives (say success/failure), with p(success) = p

The probability of k successes among n outcomes is given by

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$





For N large and p not close to 0 or 1, the binomial distribution is well approximated by a normal distribution:

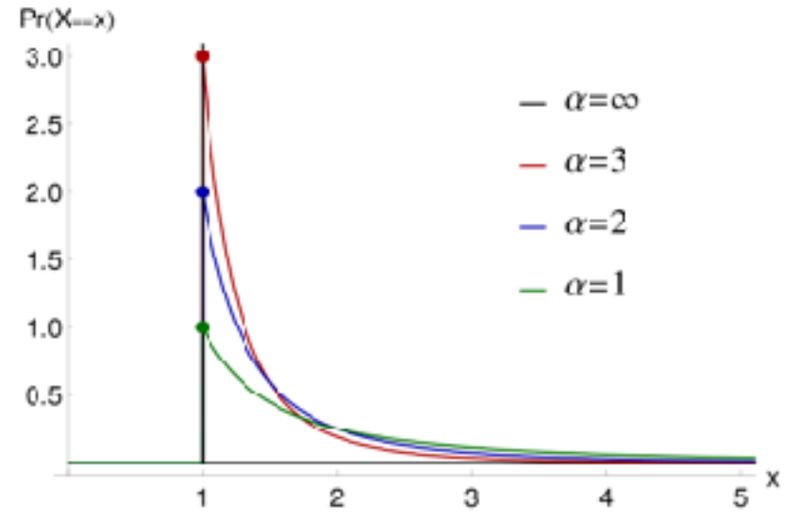
$$N(np, \sqrt{np(1-p)})$$

Pareto distributions

Power law probability distribution (alpha >=1 for convergence)

Long tail compared to normal distribution!

$$f_X(x) = \begin{cases} \frac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}} & x \geq x_{\mathrm{m}}, \\ 0 & x < x_{\mathrm{m}}. \end{cases}$$



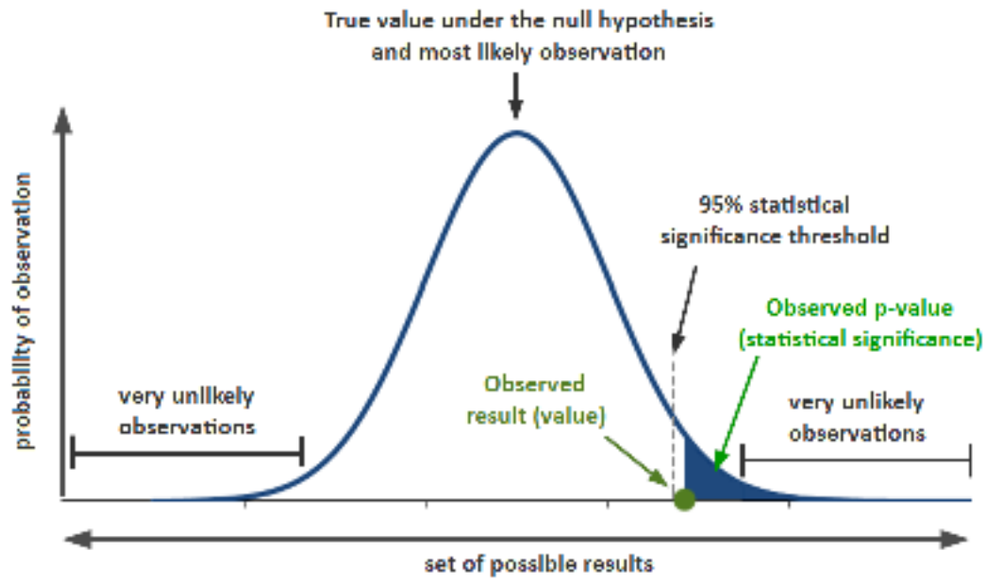
Not uncommon in real data from complex natural or artificial systems:

- city population distribution
- file size distribution in Internet traffic
- one day rainfall
- many other examples from the Zipf's law homework

(note that alpha = 1 not required here)

Hypothesis testing - statistical significance

p-value = the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.



The null hypothesis could be that an effect does not exist, and that an effect is generated by chance.

The same reasoning could be applied when comparing to a benchmark, with the benchmark as null hypothesis (e.g., in machine learning)

Statististical significance is in many fields defined as $p<0.05$. This is a weak criterion - particle physics uses 5 sigma

Statistical tests

To check statistical significance in different experimental situations, statisticians have developed a large range of statistical tests. When analyzing an experiment, you need to decide what form of testing to use.

One difference - can the underlying distributions be assumed to be known (e.g., normal distributions)?

1. Parametric vs non-parametric tests

Parametric - known distributions (most often normal)

Non-parametric - no assumption on distributions, but less discriminating

Examples:

Student t test (parametric)

Statistical tests

Student t test (parametric)

Example - one-sample t-test, test if mean different from null hypothesis mu

Calulate *test statistic t* $t = \frac{Z}{s} = \frac{\bar{X} - \mu}{\widehat{\sigma}/\sqrt{n}}$

for a sample of n values from the distribution, where the formula involves the sample mean and the estimated standard deviation of the sample (see previous slide). The quantity t can be shown to follow Student's t-distribution with n-1 degrees of freedom:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

For all reasonably large values of n (say n>30) this is well approximated by a normal distribution N(0,1) which can be used to calculate p-values.

Statistical tests

Non-parametric tests

Non-parametric tests can be used instead when you cannot be sure that the distribution is a normal distribution.

Example - Mann-Whitney U-test
(non-parametric equivalent of t-test)

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(X_i, Y_j),$$

with

$$S(X,Y) = \begin{cases} 1, & \text{if } X > Y, \\ \frac{1}{2}, & \text{if } Y = X, \\ 0, & \text{if } X < Y. \end{cases}$$

Multiple comparisons

Comparing more than 2 sample distribution

(Example - comparing multiple methods for a problem in machine learning)

Statistical significance may occur by chance given enough comparisons
(p-hacking or HARK-ing?)

Simple solution - Bonferroni correction changes significance limit
by dividing by the number of hypotheses tested.
(simple but often unnecessarily strict).

Other more sophisticated methods exist, such as ANOVA,
which can be combined with ad hoc pairwise comparisons.

An exercise

In a quiz question given an earlier year, 113 out of 210 students gave the correct answer, and 97 answered incorrectly.

One faculty member suggests that all students tossed a coin to answer. Another faculty member believes that students must have learnt something from the lectures.

Test the hypothesis of the second faculty member. What is the probability that 113 or more students answer correctly if the answers are generated randomly with equal probability?

An exercise cont.

p-value = the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.

The null hypothesis is that p=0.5 for both answers

What is the probability that 113 or more students answer correctly if the answers are generated randomly with equal probability?

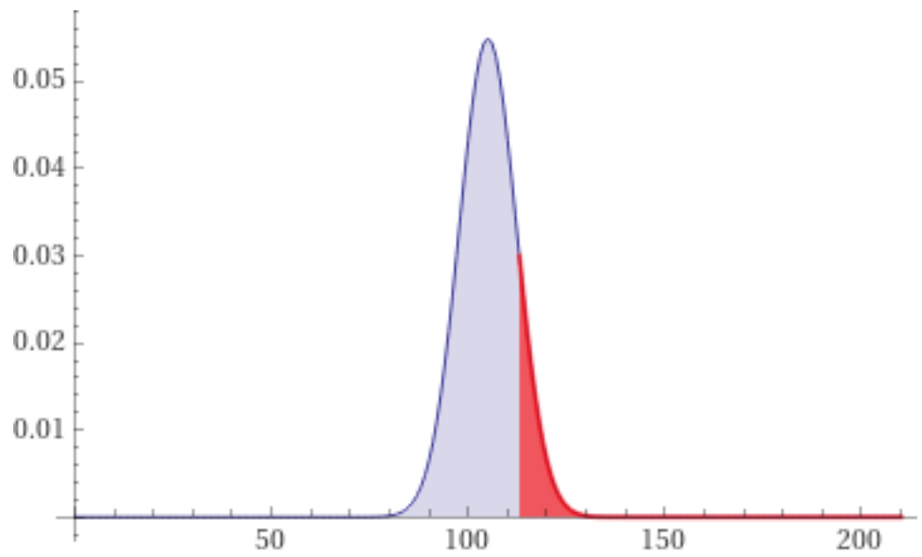
Binomial distribution: two alternatives (say success/failure), with p(success) = p

The probability of k successes among n outcomes is given by

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

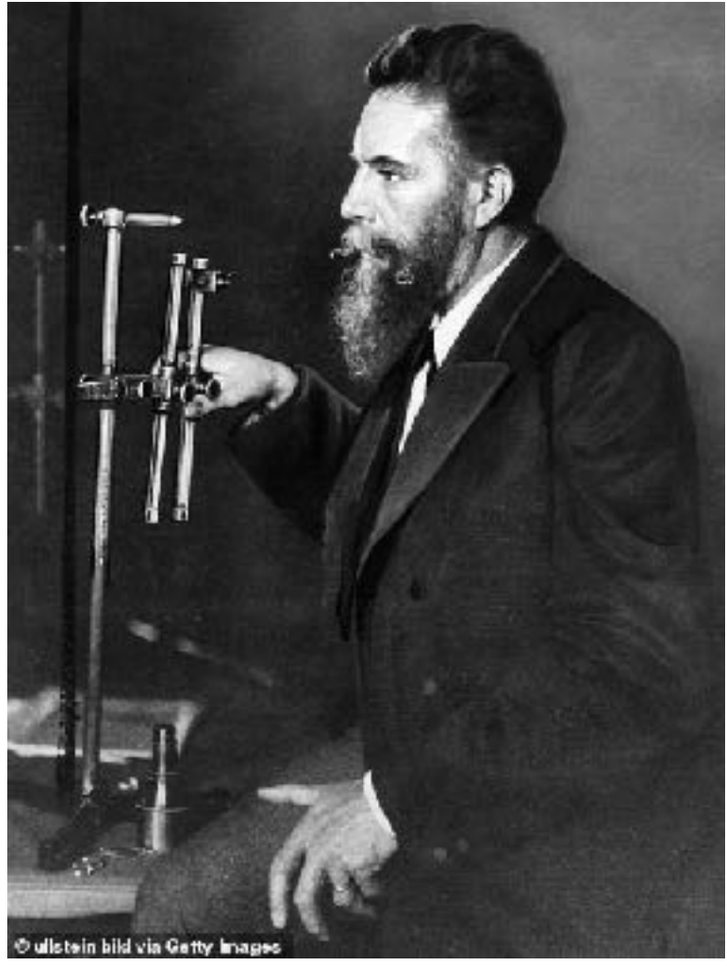
One can simply sum this with n=210, p=0.5, and k running from 113 to 210, or use the normal distribution approximation, and find that the p value is approximately 0.15:

$$\frac{676\,914\,546\,625\,125}{4\,503\,599\,627\,370\,496} \approx 0.150305$$



Serendipity

Wilhelm Conrad Röntgen (1895) experimented with an electrical discharge tube (Crookes tube) with partial vacuum. When he put a black cover over the tube without turning the apparatus off he discovered that the rays emerging still cause a fluoresence (a green glow) on a plate 9 feet away.



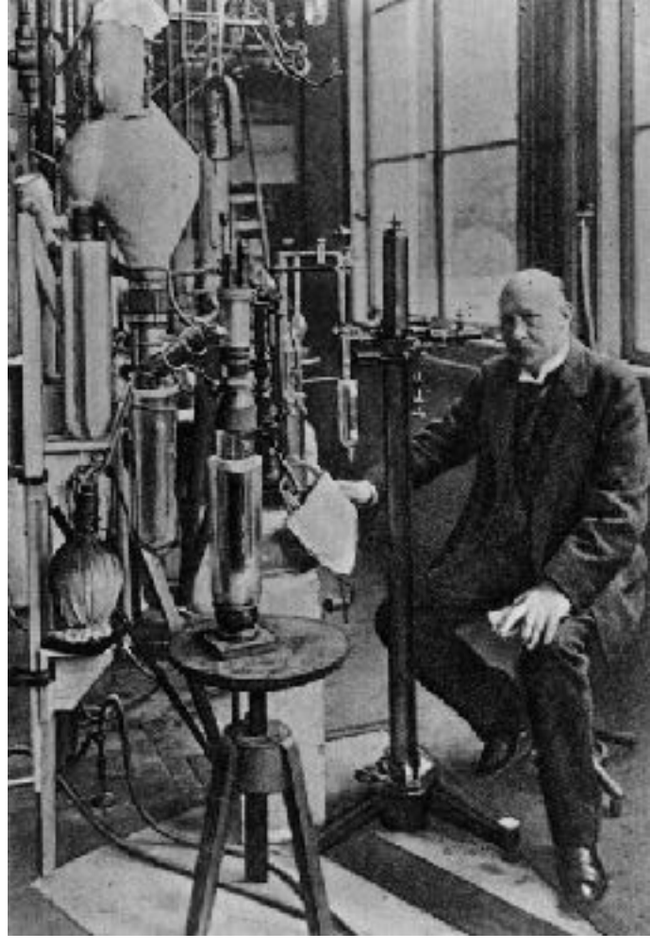
Serendipity

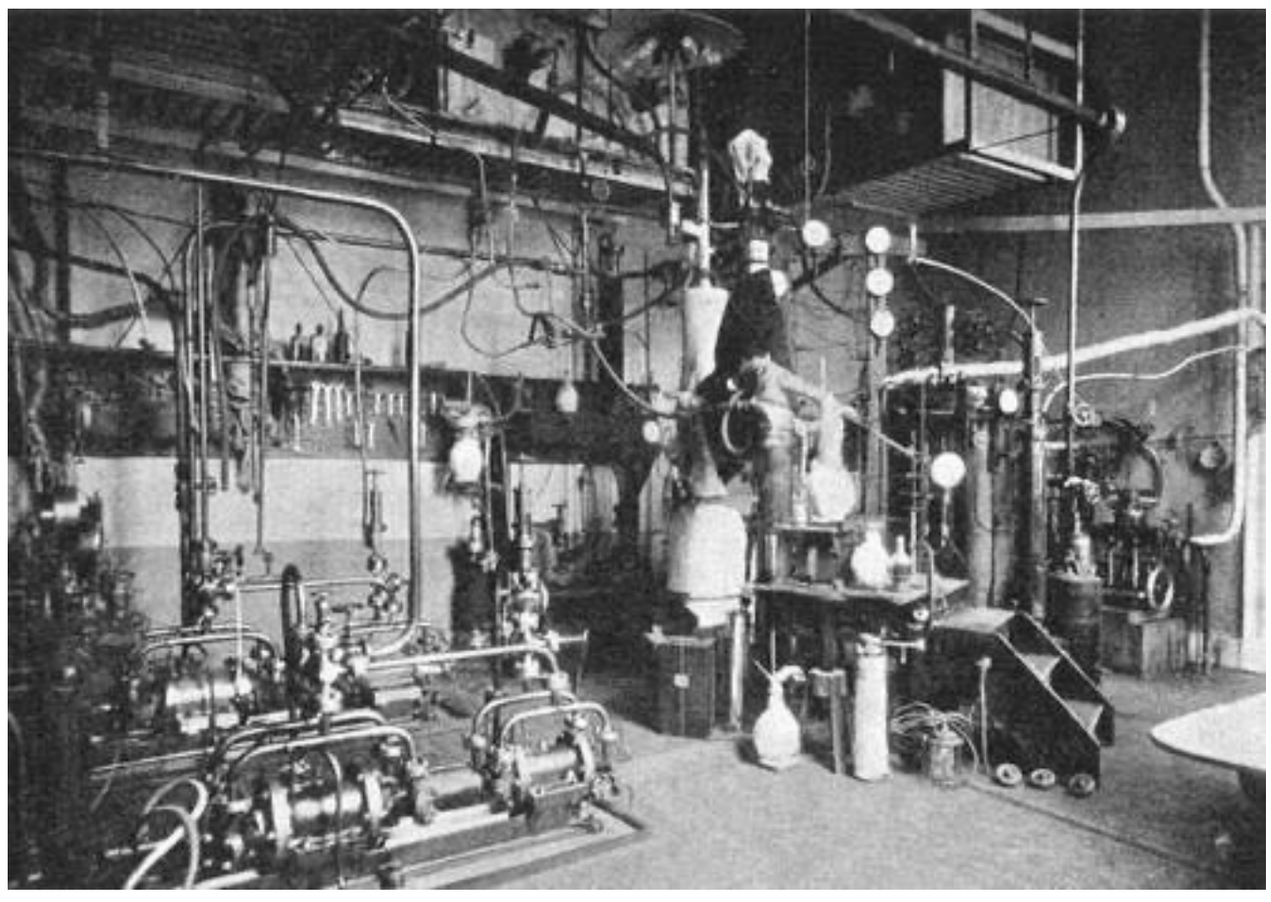
Or the role of chance rather than planning in scientific discovery (voluntary extra material)

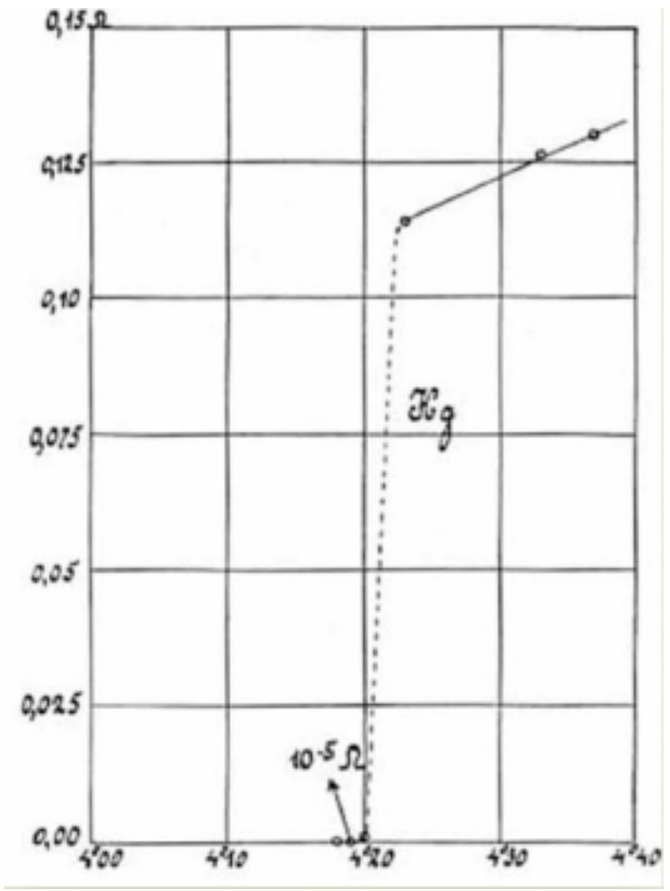
An example: Heike Kamerlingh Onnes (Leiden, Netherlands)

First to create liquid helium (1908) -269 degrees C (4.2K)

In 1911 happened to observe abruptly vanishing resistance of mercury below approximately 4.2K (superconductivity) - entirely unexpected.







Serendipity

Alexander Fleming

Discovery of penicillin (1928)

Spores of mold may perhaps have drifted across the street from the Fountains Abbey pub in through the open window of Alexander Fleming's lab at St Mary's Hospital on the other side of Praed St. (near Paddington Station).

Serendipity

Discovery of the muon

Heavy lepton, similar to an electron but heavier - mass 105.7MeV

Discovered by Carl Anderson and Seth Neddermeyer at Caltech (193
cosmic ray experiments (PikesPeak,4300 m elevation in Colorado
(Carl Anderson also discovered the positron in 1931)

Cloud chamber (Charles Wilson,1911), bubble chamber (Donald Gla
1952) - new technology drives new experiments and creates new
knowledge

”who ordered that?!!”
I.I.Rabi

esipap

Anderson & Neddermeyer I

Phys. Rev. 50 (1936) 263

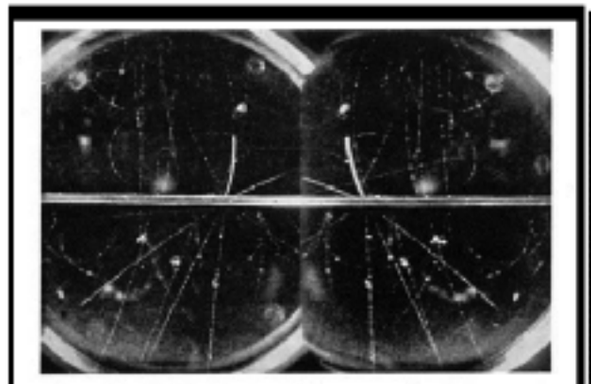


FIG. 12. Pike's Peak, 7900 gauss. A disintegration produced by a nonionizing ray occurs at a point in the 0.35 cm lead plate, from which six particles are ejected. One of the particles (strongly ionizing) ejected nearly vertically upward has the range of a 1.5 MEV proton. Its energy (given by its range) corresponds to an $H\rho = 1.7 \times 10^6$, or a radius of 20 cm, which is three times the observed value. If the observed curvature were produced entirely by magnetic deflection it would be necessary to conclude that this track represents a massive particle with an e/w much greater than that of a proton or any other known nucleus. As there are no experimental data available on the multiple scattering of low energy protons in argon it is difficult to

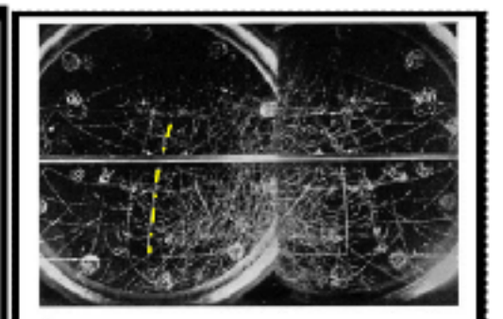
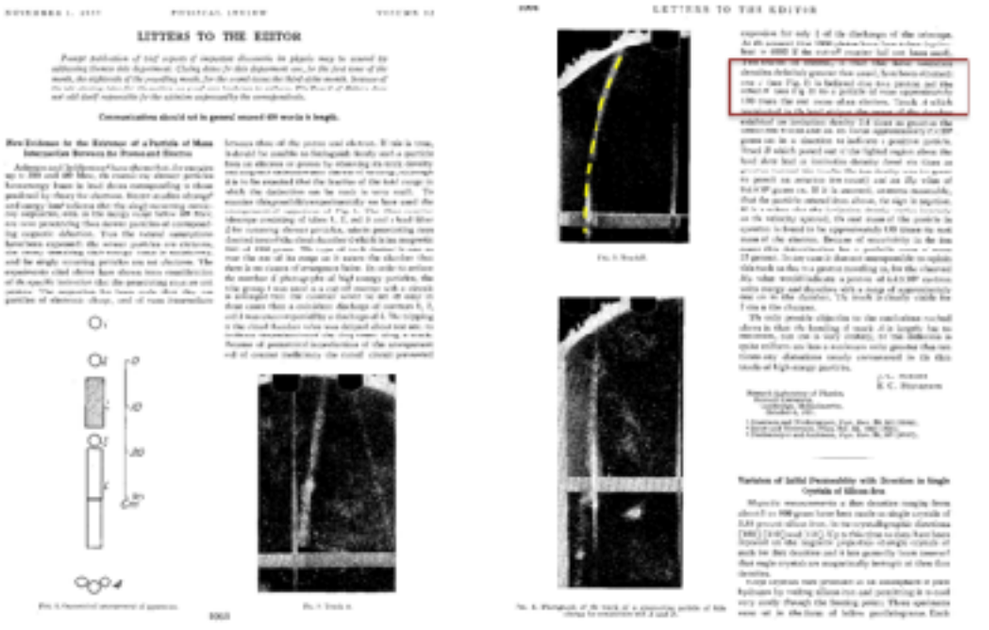


FIG. 13. Pasadena, 4500 gauss. A complex electron shower not clearly defined in direction, and three heavy particles with specific ionizations definitely greater than that of electrons. The sign of charge of two of these heavy particles represented by short tracks cannot be determined, but the assumption that they represent protons is consistent with the information supplied by the photograph. The third heavy track appears above the 0.35 cm lead plate where it has a specific ionization not noticeably different from that of an electron. It penetrates the lead plate and appears in the lower half of the chamber as a nearly vertical track near the middle. Below the plate it shows a greater ionization than an electron, and is deviated in the magnetic field to indicate a positively charged particle. Its $H\rho$ is apparently at most 1.4×10^6 gauss cm, which corresponds to a proton energy of 1 MEV and a range of only 2 cm in the chamber, whereas the observed range is greater than 5 cm. A difficulty of the same nature was discussed in the description of the previous photograph.

sipap

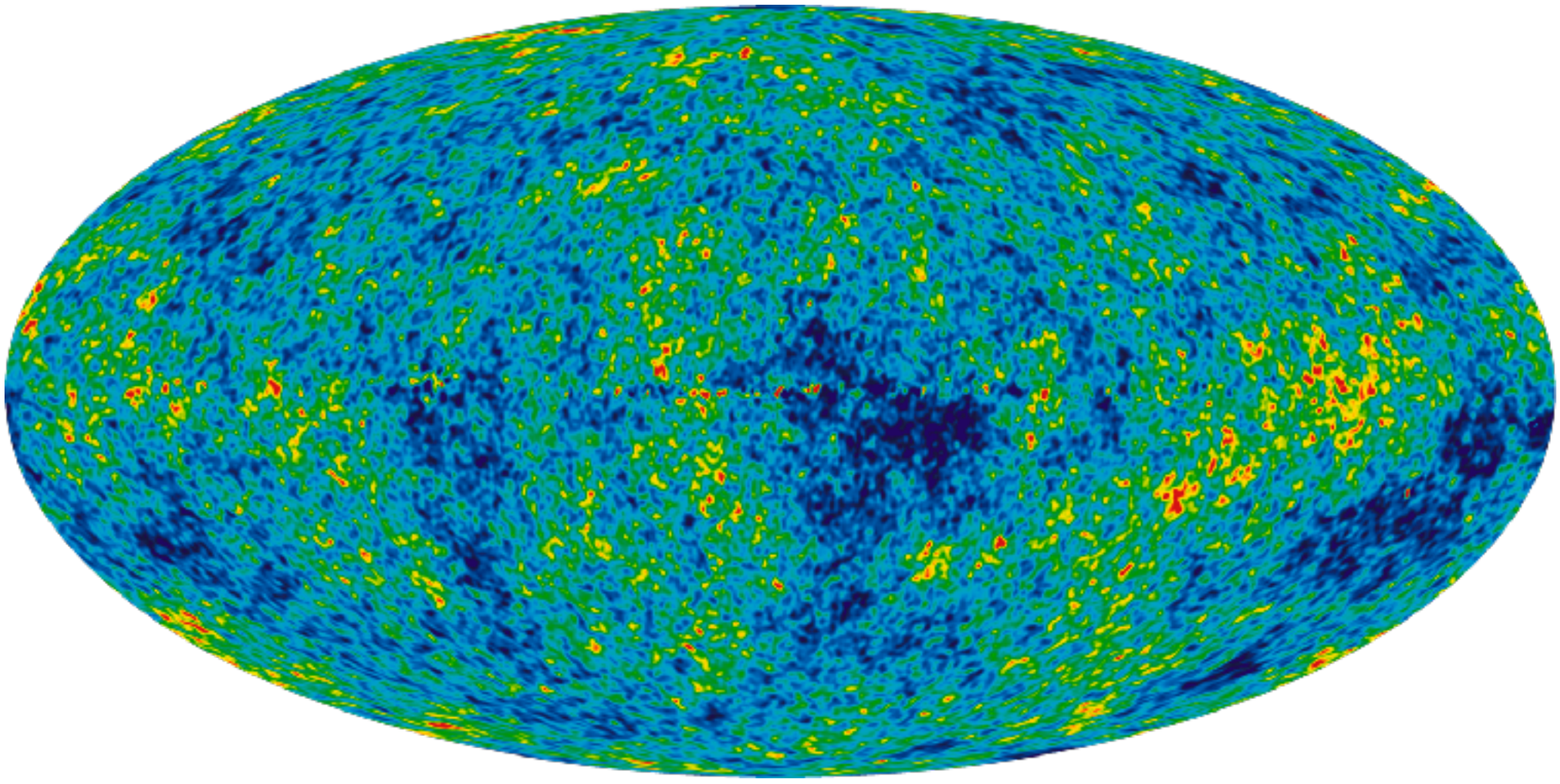
Street & Stevenson (cloud chamber)



Serendipity

Cosmic microwave background radiation

Arno Penzias and Robert Woodrow Wilson (1964)
Bell Telephone Laboratories, Crawford Hill, NJ

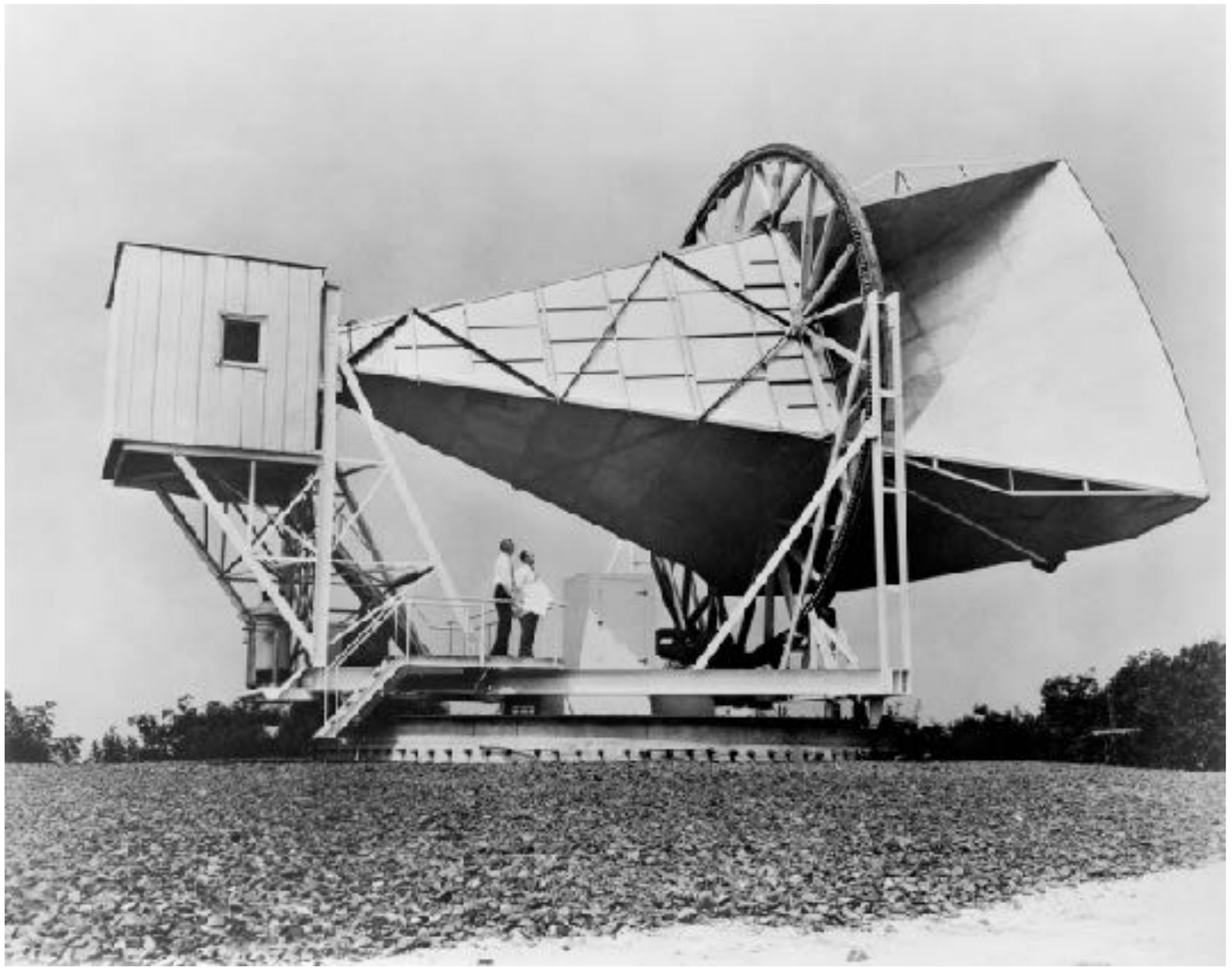


Serendipity

Cosmic microwave background radiation discovered by chance in another experiment.

Arno Penzias and Robert Woodrow Wilson (1964)

Holmdel Horn Antenna. Built 1959 for communication project - bouncing radio waves off passive satellites (balloons) from one point on earth to another.



Serendipity

Edward Lorenz, MIT (1961)

Observation of deterministic chaos (unpredictability) in simulations.
12 ordinary differential equations realted to weather simulations.

Internal representation 6 decimals
Restarting from printout (3 decimals only)

Librascope LGP-30, Memory size: 4096 word Technology: 113 vacuum tubes and 1350 diodes. Clock rate: 120 kHz