BCB 731:

Defense Against the Dark Arts

Everything About Data

October 4th, 2023

Let's talk about everything

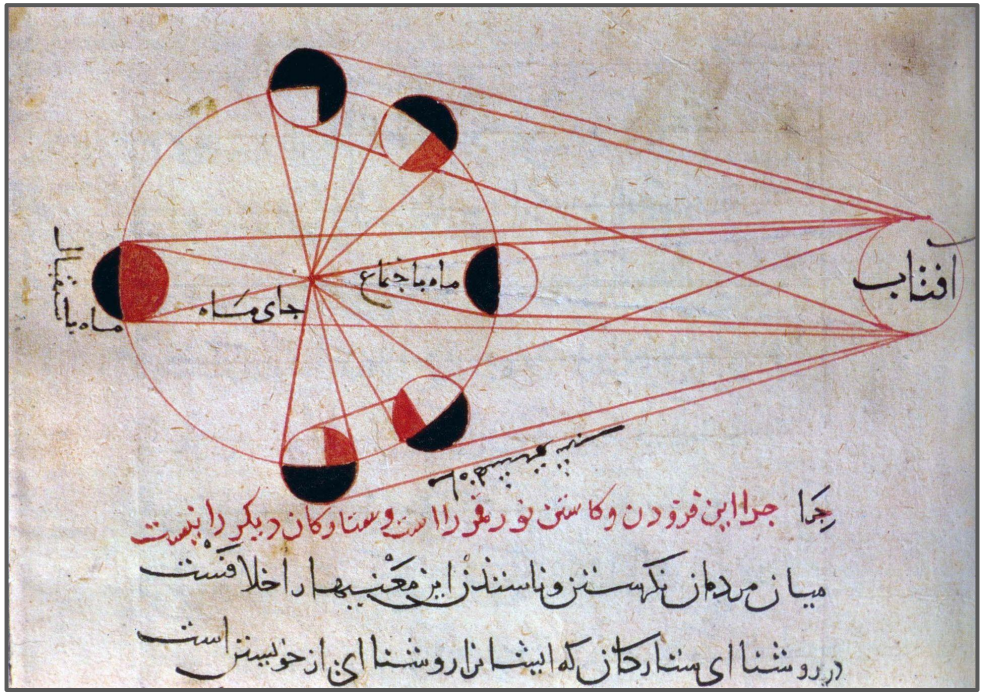
- Classical Statistics
- Exploratory Data Analysis
- Bayesian Statistics
- Classical Machine Learning
- Deep Learning
- Artificial Intelligence

...and more!

- What is knowledge?
- How do we come to know something about the world?
- What role does data play in knowledge formation?
- Do we even need statistics?

Before the “scientific revolution”

- Interplay between **observation** and strong **inductive** priors
- Abstract models more often descriptive than mathematical
- Validated by agreement with observation but also:
  - Elegance / aesthetics
  - Great “masters”
  - Theology & philosophy



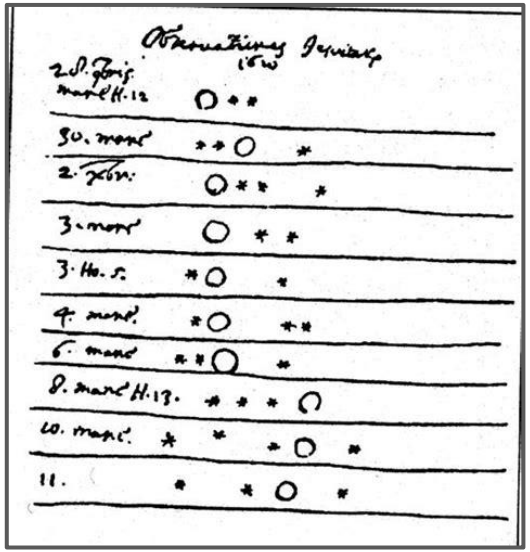
Abu-Rayhan al-Biruni’s *Al-Tafhim li Awa'il Sana'at al-Tanjim* (Book on the Elements of Astrology)

Modern science = empiricism

- 1600s science narrowed “natural philosophy” to:
  - Collect data
  - Build mathematical models (repeat)

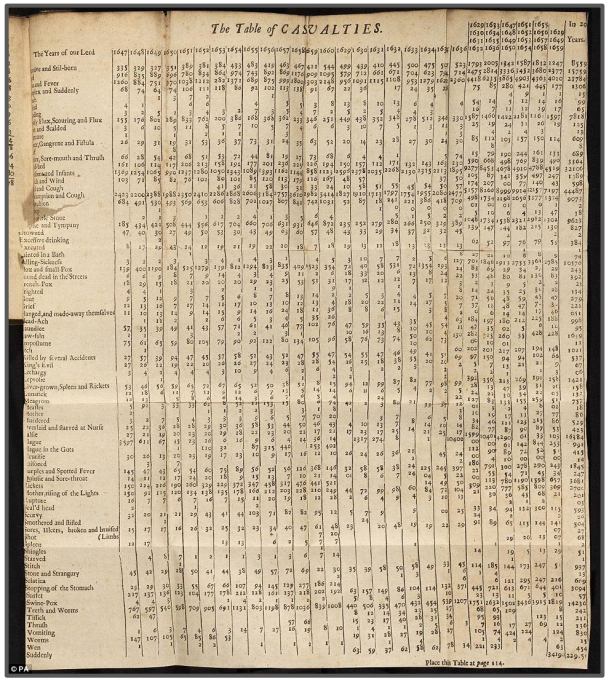
“the Universe — which stands continually open to our gaze, but it cannot be understood unless one first learns to comprehend the language and interpret the characters in which it is written. It is written in the language of mathematics” – Galileo in The Assayer

Galileo Gallilei’s “The Assayer”



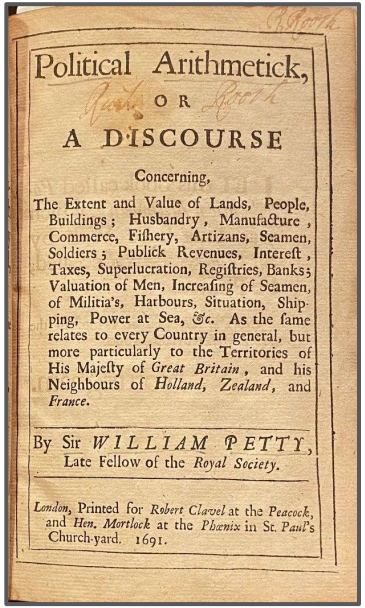
Galileo’s notebook observing moons of Jupiter (interpreted as evidence for refuting geocentrism)

Statistics = seeing like a state



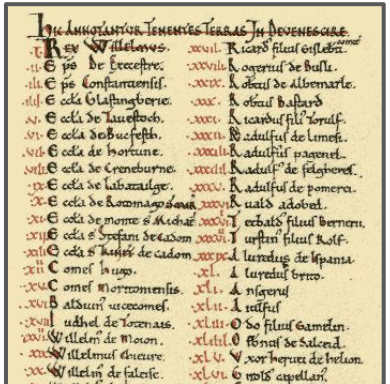
The Table of CASUALTIES.

Table from John Graunt's *Natural and Political Observations Made Upon the Bills of Mortality* (1662)

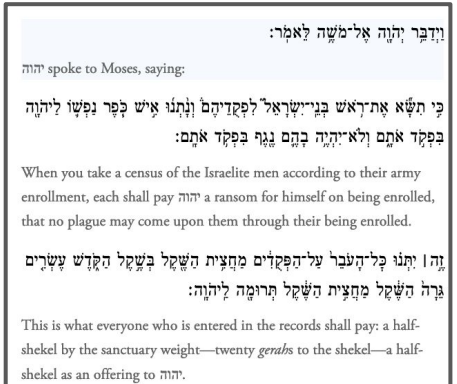


Political Arithmetick,
OR
A DISCOURSE
Concerning,
The Extent and Value of Lands, People, Buildings; Husbandry, Manufacture, Commerce, Fishery, Artizans, Seamen, Soldiers; Publick Revenues, Intereſt, Taxes, Superlucration, Regiſtries, Banks; Valuation of Men, Increaſing of Seamen, of Militia's, Harbours, Situation, Shipping, Power at Sea, &c. As the ſame relates to every Country in general, but more particularly to the Territories of His Majeſty of *Great Britain*, and his Neighbours of *Holland*, *Zealand*, and *France*.
By Sir *WILLIAM PETTY*,
Late Fellow of the *Royal Society*.
London, Printed for *Robert Clavel* at the *Peacock*, and *Hen. Mortlock* at the *Phœnix* in St. *Paul's* Church-yard. 1691.

William Petty's *Political Arithmetic* (1691)



William The Conqueror's Domesday Book of 1086



וַיְדַבֵּר יְהוָה אֶל־מֹשֶׁה לֵאמֹר׃

יהוה spoke to Moses, saying:

כִּי תִשָּׂא אֶת־רֹאשׁ בְּנֵי־יִשְׂרָאֵל לִפְקֻדֵיהֶם וְנָתְנוּ אִישׁ כֹּפֶר נַפְשׁוֹ לַיהוָה
בִּפְקֹד אֹתָם וְלֹא־יִהְיֶה בָהֶם נֶגֶף בִּפְקֹד אֹתָם׃

When you take a census of the Israelite men according to their army enrollment, each shall pay יהוה a ransom for himself on being enrolled, that no plague may come upon them through their being enrolled.

זֶה | יִתְּנוּ כָּל־הָעֹבֵר עַל־הַפְּקֻדִים מַחֲצִית הַשֶּׁקֶל בְּשֶׁקֶל הַקֹּדֶשׁ עֶשְׂרִים
גֵּרָה הַשֶּׁקֶל מַחֲצִית הַשֶּׁקֶל תְּרוּמָה לַיהוָה׃

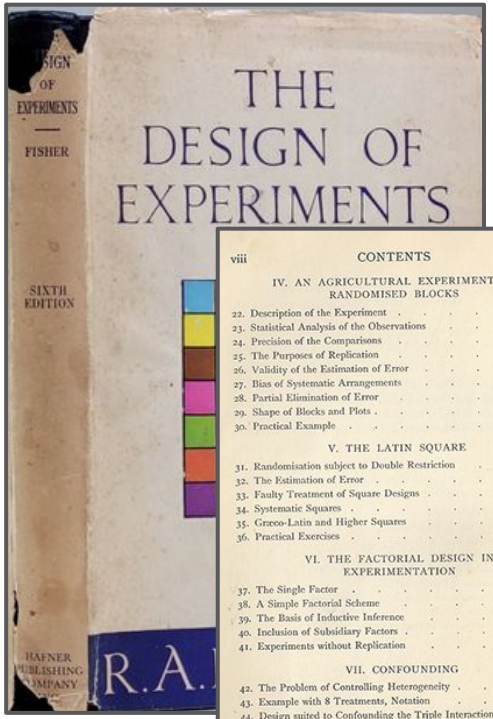
This is what everyone who is entered in the records shall pay: a half-shekel by the sanctuary weight—twenty *gerahs* to the shekel—a half-shekel as an offering to יהוה.

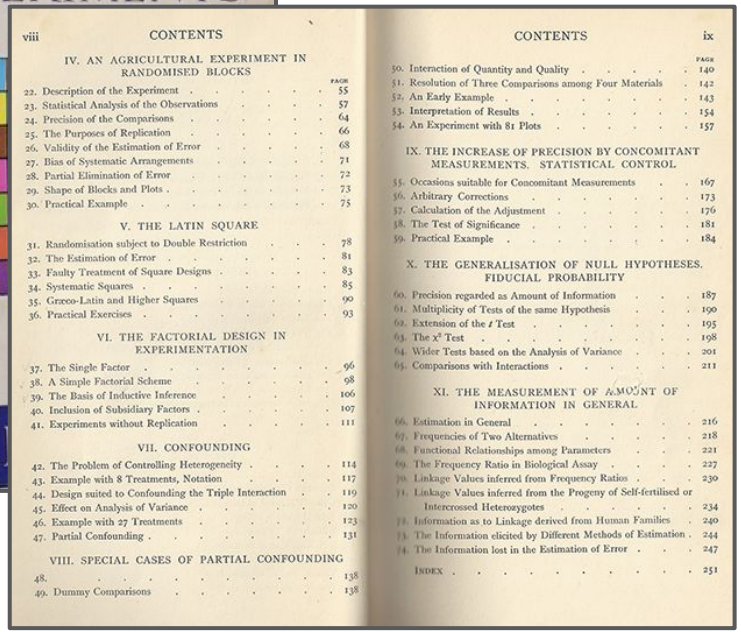
כֹּל הָעֹבֵר עַל־הַפְּקֻדִים מִבֶּן עֶשְׂרִים שָׁנָה וָמָעְלָה יִתֵּן תְּרוּמַת יְהוָה׃

Everyone who is entered in the records, from the age of twenty years up, shall give יהוה's offering:

Beginning of the Torah portion *Ki Tisa*, Moses is asked to take a census and collect a tax to build a tabernacle

1920s/30s: statistics infiltrates science





viii CONTENTS

IV. AN AGRICULTURAL EXPERIMENT IN RANDOMISED BLOCKS

V. THE LATIN SQUARE

VI. THE FACTORIAL DESIGN IN EXPERIMENTATION

VII. CONFOUNDING

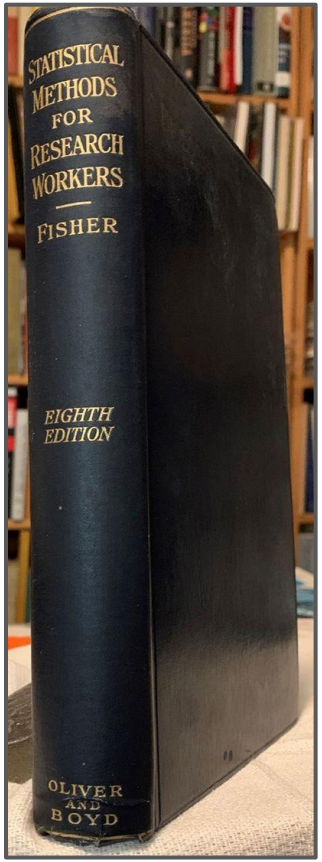
VIII. SPECIAL CASES OF PARTIAL CONFOUNDING

CONTENTS ix

IX. THE INCREASE OF PRECISION BY CONCOMITANT MEASUREMENTS. STATISTICAL CONTROL

X. THE GENERALISATION OF NULL HYPOTHESES. FIDUCIAL PROBABILITY

XI. THE MEASUREMENT OF AMOUNT OF INFORMATION IN GENERAL



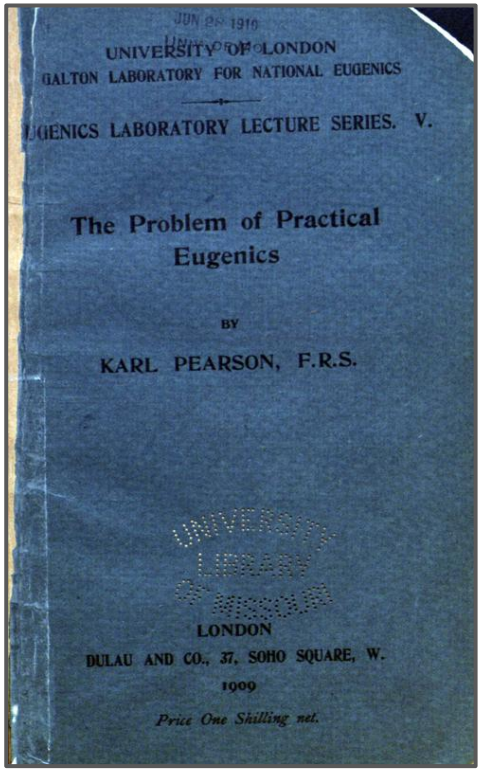
CONTENTS

TABLES

Side note: 2/4 “fathers” of statistics were very into eugenics

“*...eugenics urges us to simplify our lives, and to simplify our needs; the only luxury worth having is that of a worthy human environment. We must be ready to sacrifice social success, at the call of nobler instincts.*” -R. A. Fisher

“History shows me one way, and one way only, in which a high state of civilization has been produced, namely, the struggle of race with race, and the survival of the physically and mentally fitter race.” -Karl Pearson



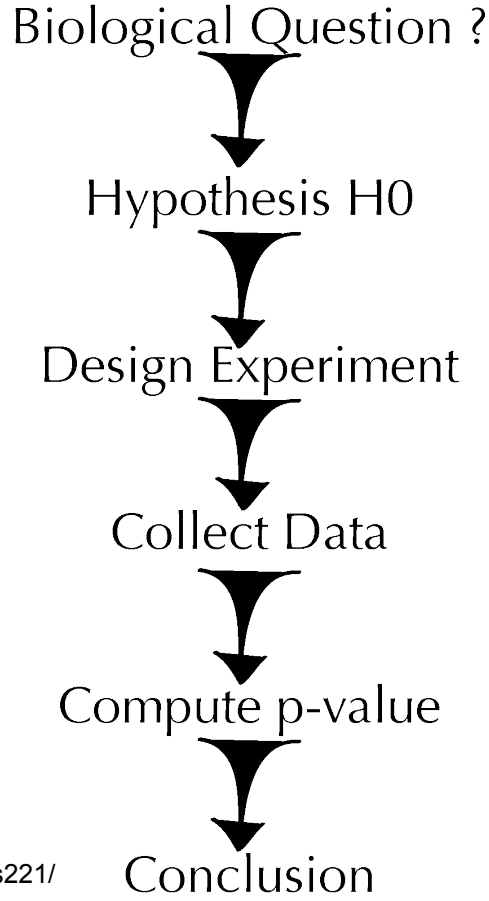
So, what is classical statistics?

- There is some real world quantity:
  - …how do we finite noisy measurements into a robust estimate of the “true” value?

- We have a mathematical model of reality:
  - …how do rigorously we use finite noisy measurements to reject (or conditionally accept) the model?

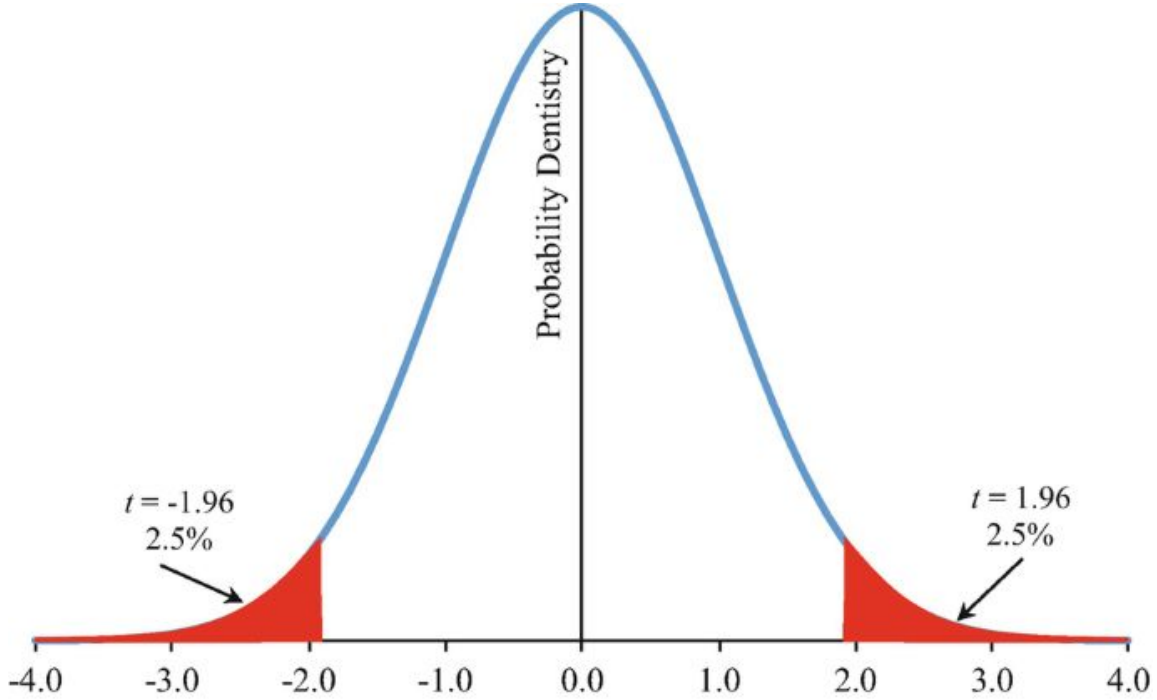
Tools of classical statistics

- Experimental design
- Inference & estimators
  - consistent, unbiased, efficient
- Confidence intervals
- Statistical hypothesis testing (Neyman-Pearson)
- Null hypothesis models (Fisher)
- Their unholy marriage: NHST



https://web.stanford.edu/class/bios221/

Null Hypothesis Significance Testing



Teaching Null Hypothesis Significance Testing (NHST) in the Health Sciences: The Significance of Significance

Exploratory Data Analysis: You're allowed to look at your data

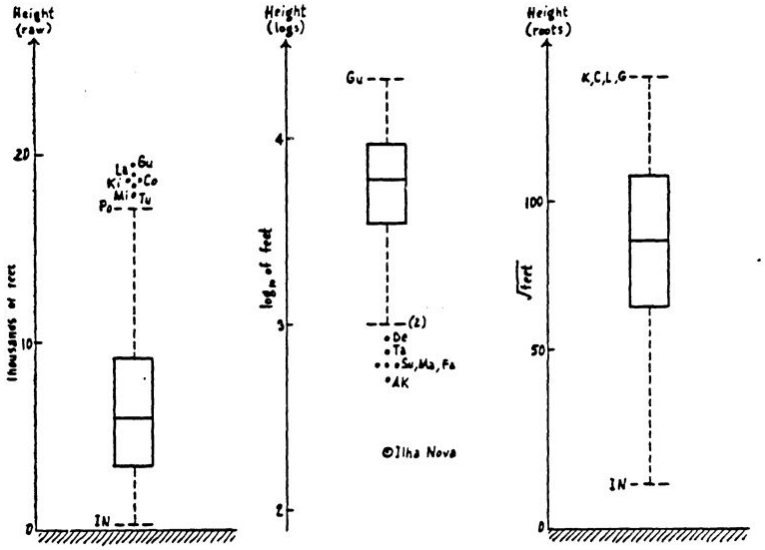
Exploratory Data Analysis: Past, Present, and Future

John W. Tukey[1]

Technical Report No. 302
Princeton University, 408 Fine Hall, Washington Road, Princeton, NJ 08544-1000

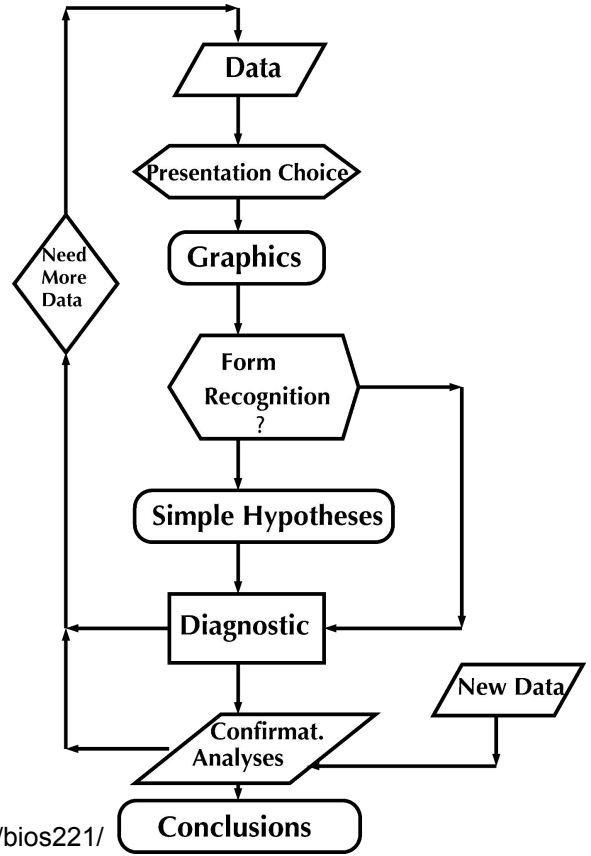
Abstract

The 1971-1977 early formulation of Exploratory Data Analysis, in terms of (a) results of some of its techniques and considerations which underlay, at various depths, the choices realized in the books. The 1991–1995 development of Exploratory Analysis of Variance, described in its simplest (two–way table) form and barely sketched in general. Discussion of the changes in apparent philosophy caused by the need to communicate more complicated things, notches, hints, the likely impact on a revised edition of Exploratory Data Analysis 1977. Dreams and targets for what might happen in 1996–2005, with emphasis on Exploratory Regression and the combined use of multiple description.



Exploratory Data Analysis: workflow

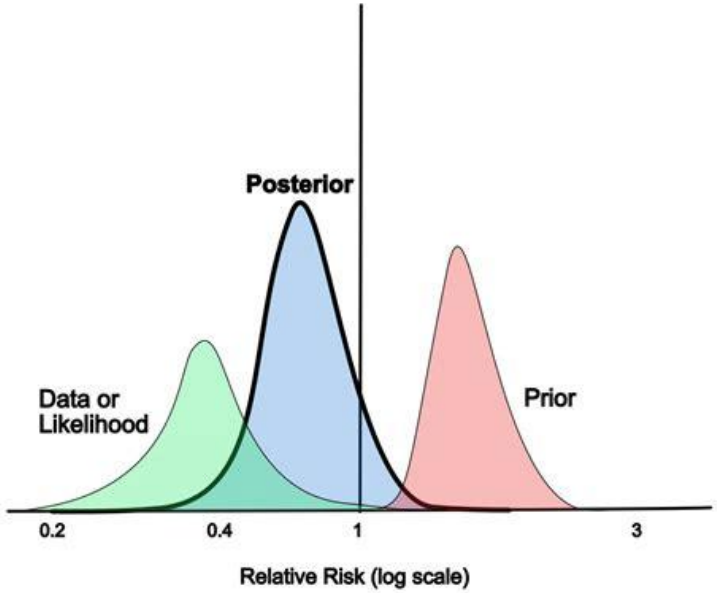
- Emphasis on visualization
- Look at residuals of your model
  - …you can even try different models!
- Does it make sense?
- …classical stats left for "confirmatory analyses"



https://web.stanford.edu/class/bios221/

Bayesian Statistics

- Why are we accepting or rejecting a hypothesis?
- Why do we think there's a single "true" value to parameters?
- Inference should give us a full probability distribution



https://www.thebottomline.org.uk/blog/ebm/bayesian-statistics/

Machine Learning

- Forget models of reality!
- …let's just do function approximation instead.
- Data is some set of vectors x_i (can have labels y_i)
- Learn functions $f_w(x)$ parameterized by weights w
  - Clustering: X -> {1, …, k}
  - Regression / Classification: X -> Y

TO THE WHITEBOARD!

quick tour of ML / loss minimization / gradient descent / evaluation metrics / deep learning / AI


Fin