Thinking Like A Data Scientist

Em Grasmeder
ThoughtWorks Data Witch
@emilyagras

About Me

- ThoughtWorks consultant
- Graduate research in economics
  - I flunked out
- Generalist within data space

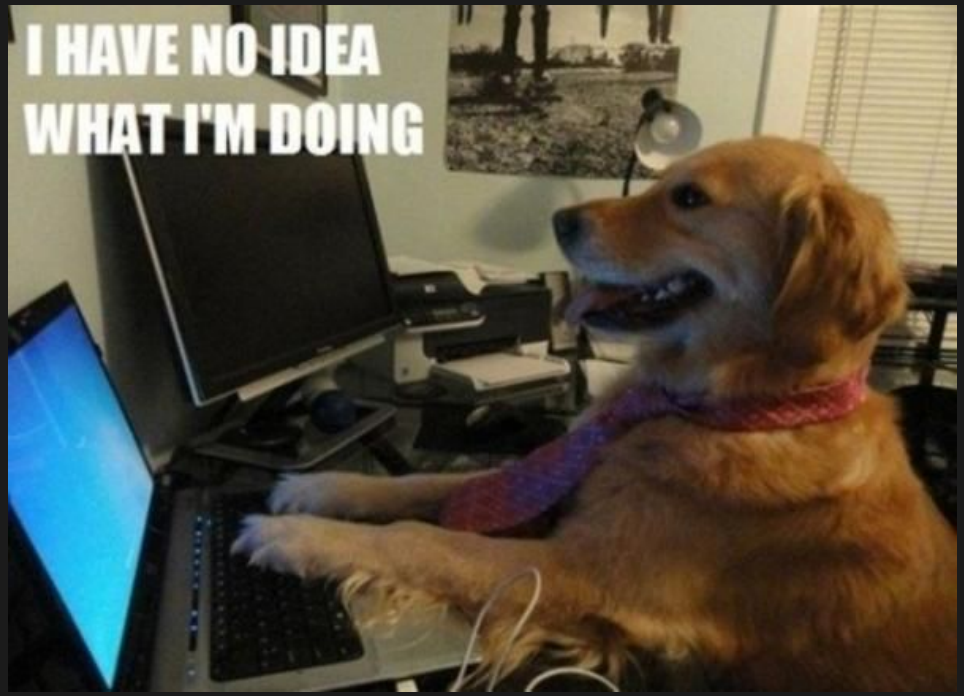
<Data Science Identity Crisis>

What Does a Data Scientist Do

- Predictions, categorization, clustering

What Does a Data Scientist Do

- Predictions, categorization, clustering
- Write software



What Does a Data Scientist Do

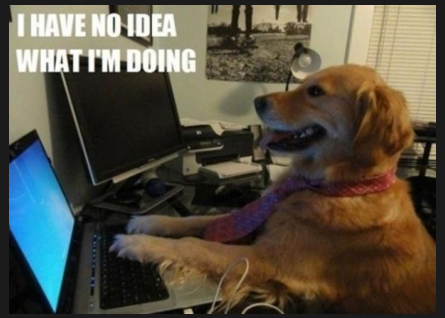
- Predictions, categorization, clustering
- Write software
- Visualizations

What Does a Data Scientist Do

- Predictions, categorization, clustering
- Write software
- Visualizations
- magically fix your business model??

How is a Data Scientist Useful

Code



+

Models

+

Visualization

+

=

Business Value

?

Controversial opinions about data scientists

- They should be good software and API developers
- They should be competent at continuous delivery, making and managing pipelines, and writing infrastructure as code
- They should speak the language of the business and be involved in conversations about KPIs

  If not...
- They might not be very useful

So how do data scientists actually think?

Let’s answer that question with a story about Cholera!

Cho

m

Cholera Facts (yay!)

Deadly bacteria that can kill within hours

The water in your body just comes out from everywhere

Pretty much curable (90% of cases) with salty, sugary water that costs $0.10

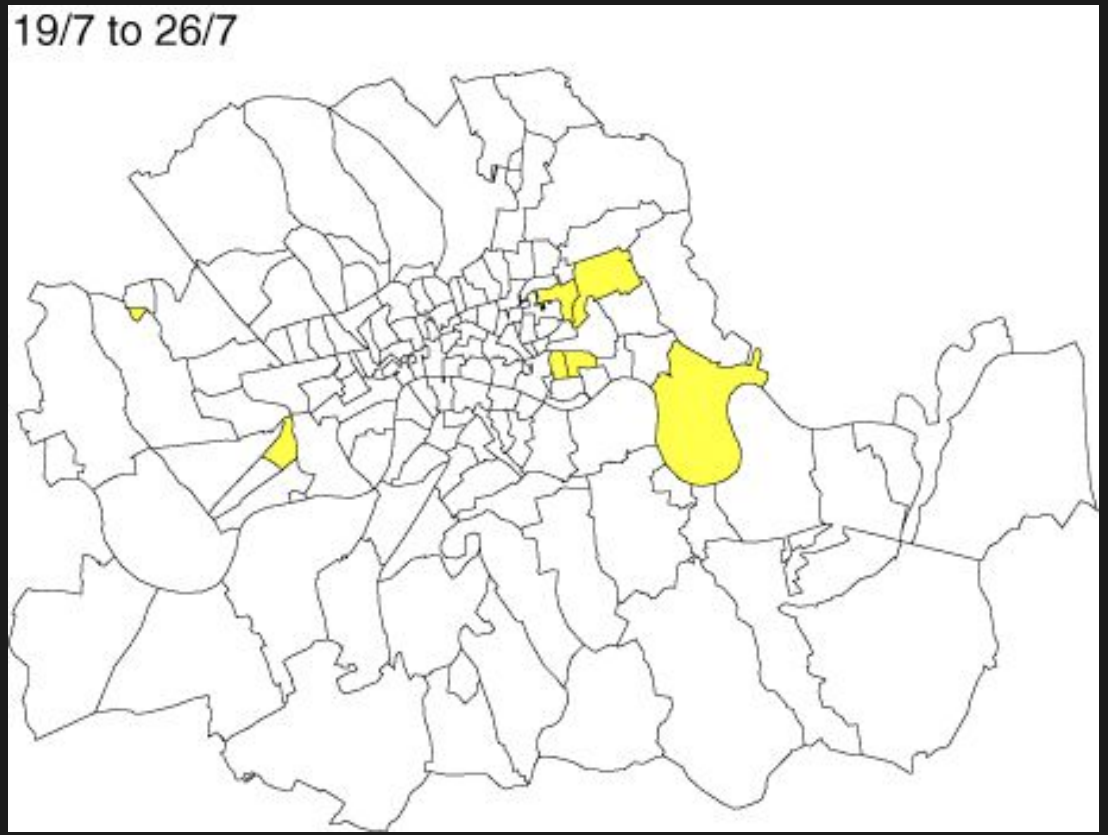
Used to be a problem in London, is still a problem in some places

The cause of cholera and how it spreads was unknown 1854.

They thought it was “miasma” literally, bad air

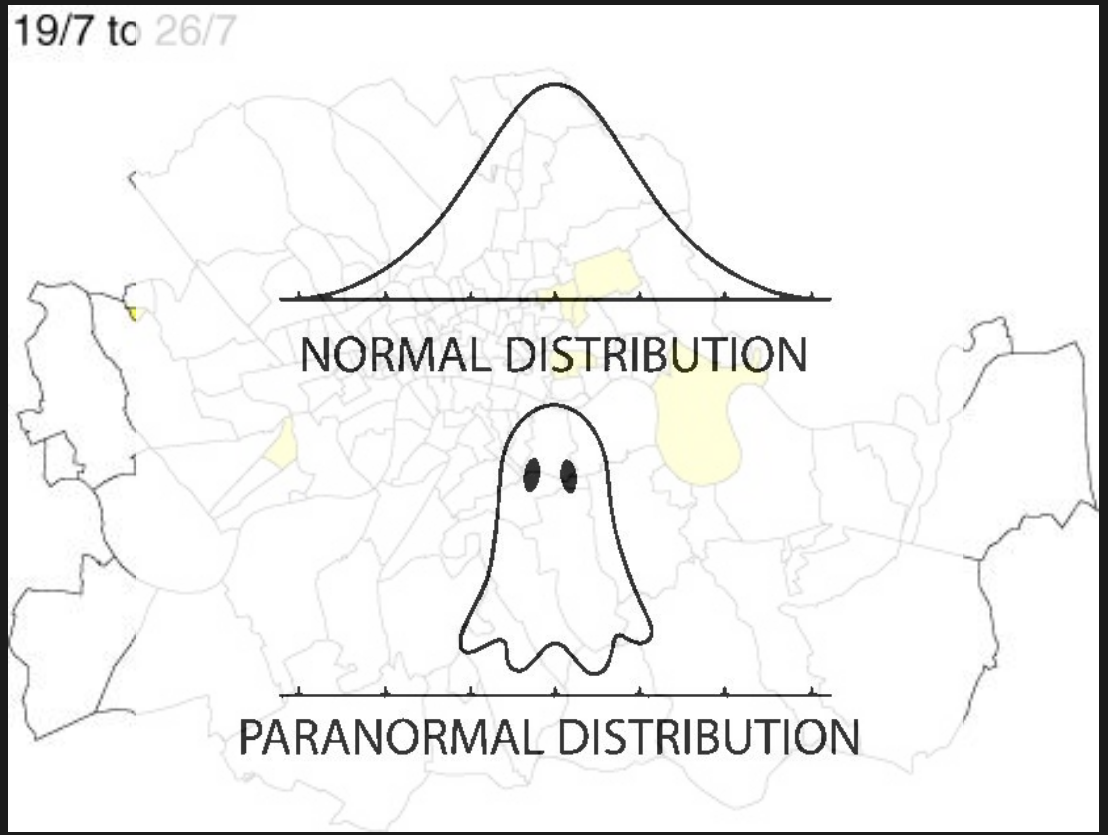
The Broad Street Cholera Outbreak of 1854

London’s cholera outbreak



<Awkward Switch To Jupyter Notebook>

London’s cholera outbreak



London’s cholera outbreak

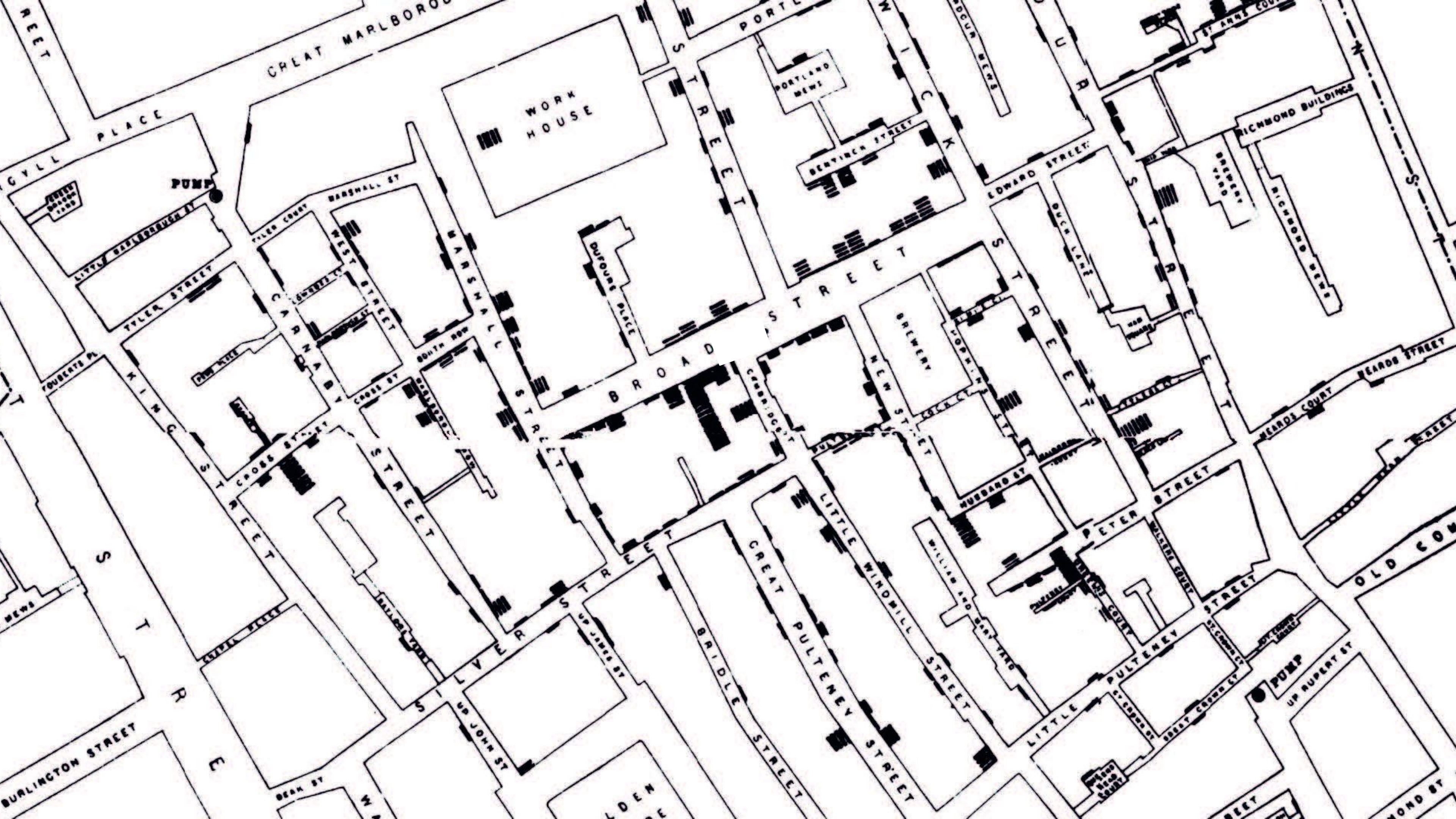
Formally write your hypothesis

- H_0 is called the Null Hypothesis
- In 1850s England, the Null Hypothesis is “bad airs”

Formally write your hypothesis

- H_0: Thing is normally distributed
  OR
- H_0: Thing is uniformly distributed
- H_1: Thing is distributed differently because reason

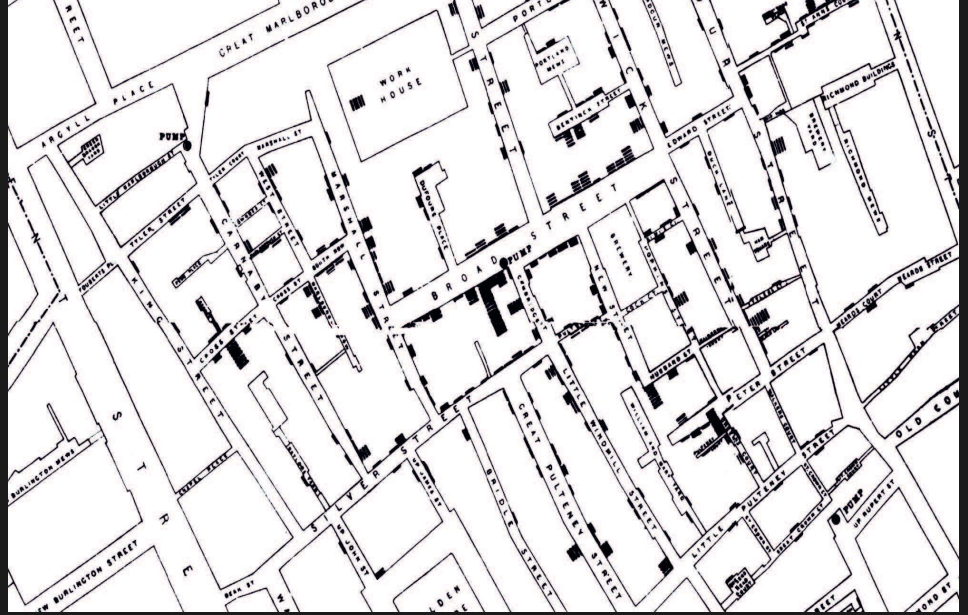
* H_1 is the same thing as H_A



GREAT MARLBOROUGH
WORK HOUSE
PUMP
PORTLAND MEWS
BERWICK STREET
RICHMOND BUILDINGS
RICHMOND MEWS
EDWARD STREET
DUFOURS PLACE
MARSHALL ST
MARSHALL STREET
WEST STREET
TYLER STREET
CARNABY STREET
KING STREET
BROAD STREET
BREWERY
NEW STREET
CAMBRIDGE ST
CROSS ST
CROSS STREET
SOUTH ROW
DUCK LANE
COCK CT
HUSBAND ST
HEARDS COURT
PETER STREET
SILVER STREET
UP JAMES ST
UP JOHN ST
BRIDLE STREET
GREAT PULTENEY STREET
LITTLE WINDMILL STREET
WILLIAM AND MARY YARD
LITTLE PULTENEY STREET
UP RUPERT ST
PUMP
BURLINGTON STREET
OLD COM

Refining hypotheses

- Assume normal or uniform distributions across population
- We know population is not uniformly distributed
- What we might consider: Map infection-per-capita



Formally write your hypothesis

- H_0: People living in equally odorous parts of town will have a uniform likelihood of contracting cholera


What if we actually talked to the poor people?


Preposterous!

Fun fact! The word statistics, in 1770 meant the "science dealing with data about the condition of a state or community"

Collecting more data

Workers at brewery were unaffected while their families died

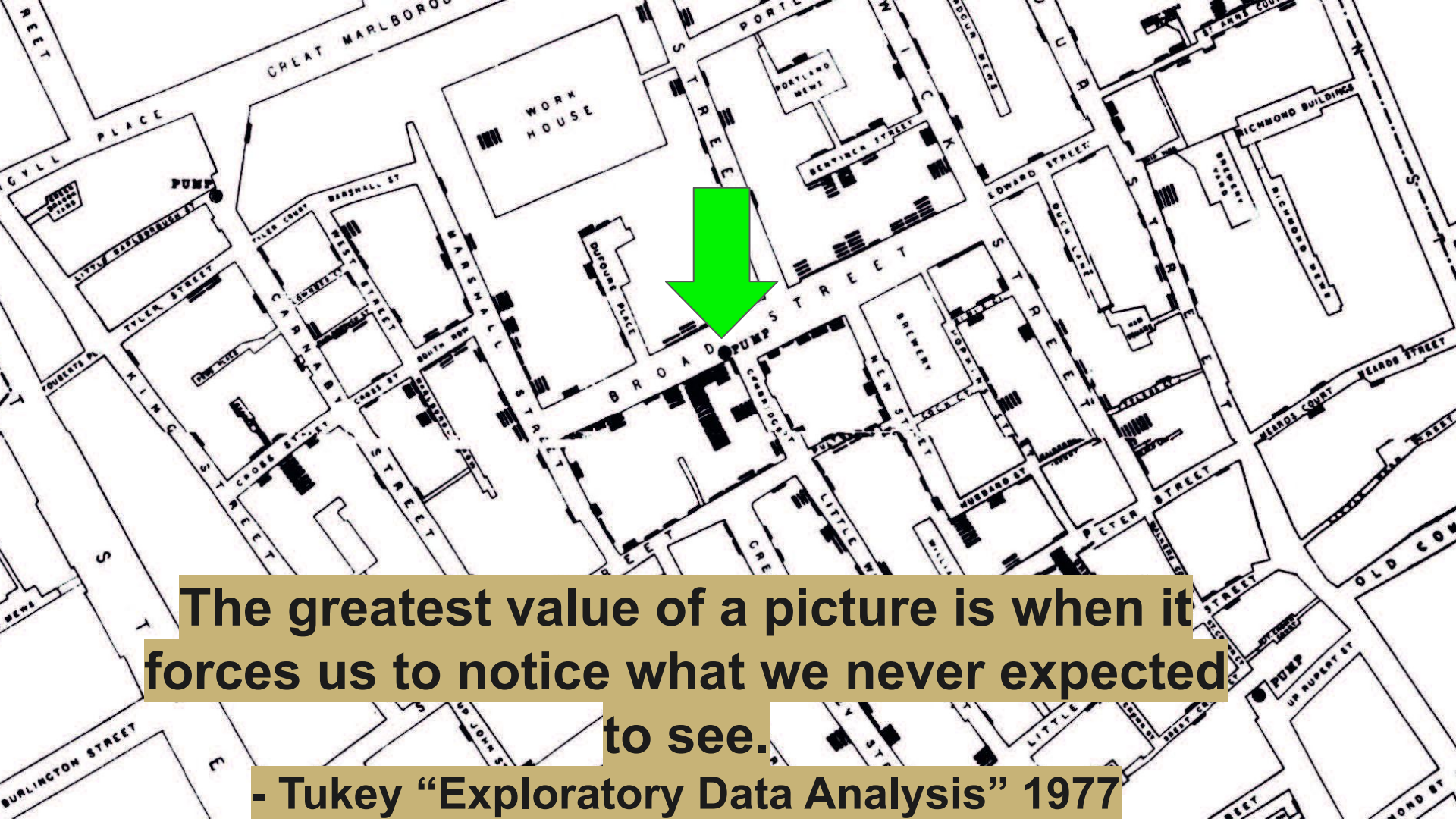
Some children died while their families lived

One woman was a complete outlier and the only person in her neighborhood to die

For
hyp

- H
oc
ha
of

- H
ch
wa



The greatest value of a picture is when it forces us to notice what we never expected to see.

- Tukey “Exploratory Data Analysis” 1977

So what are the lessons?

Data is good. More data is better

Visualize your data!

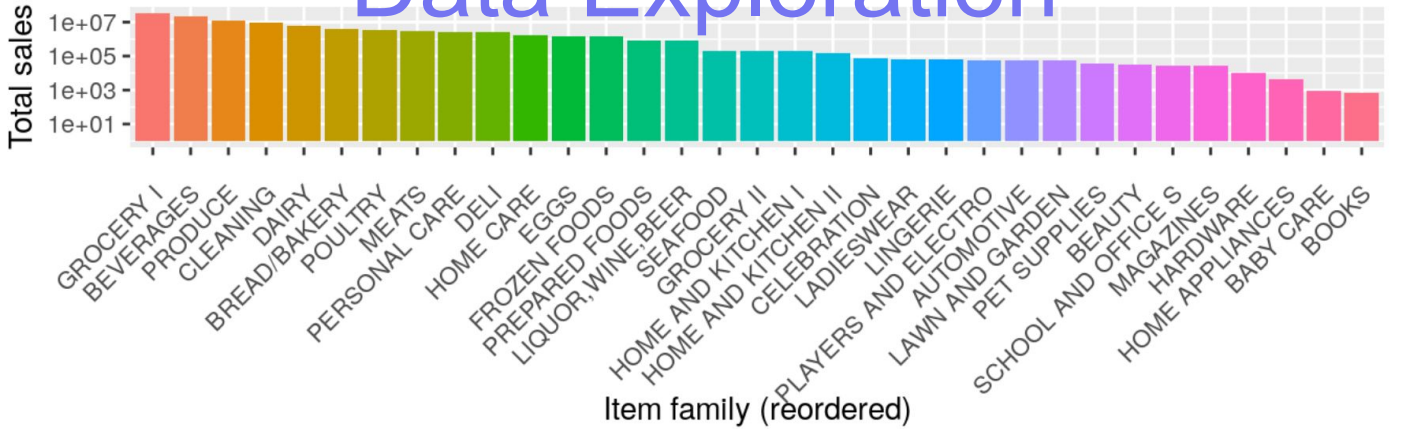
Do we really need machine learning for this?

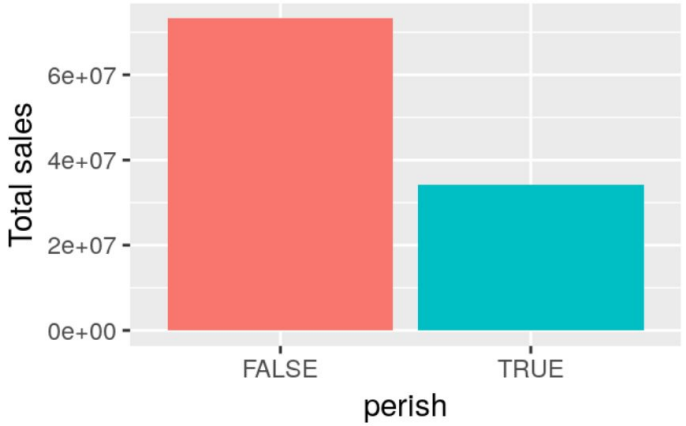
Data Exploration: The first step towards using a model

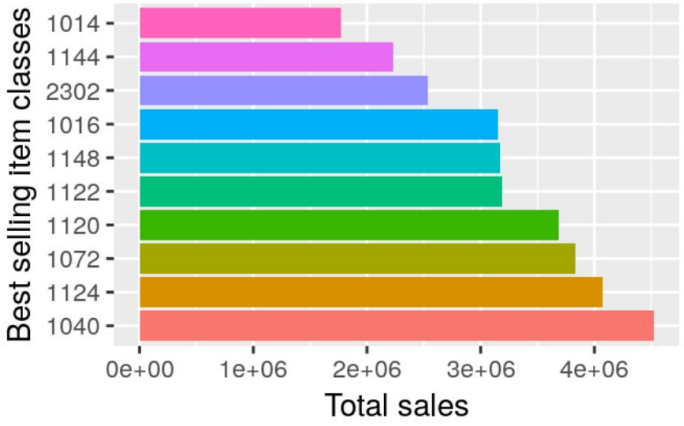
6.3 Items: family classification

Here we plot the sales numbers for the *family* categories together with the statistics for *perishable* items and the top selling *classes*:

Code





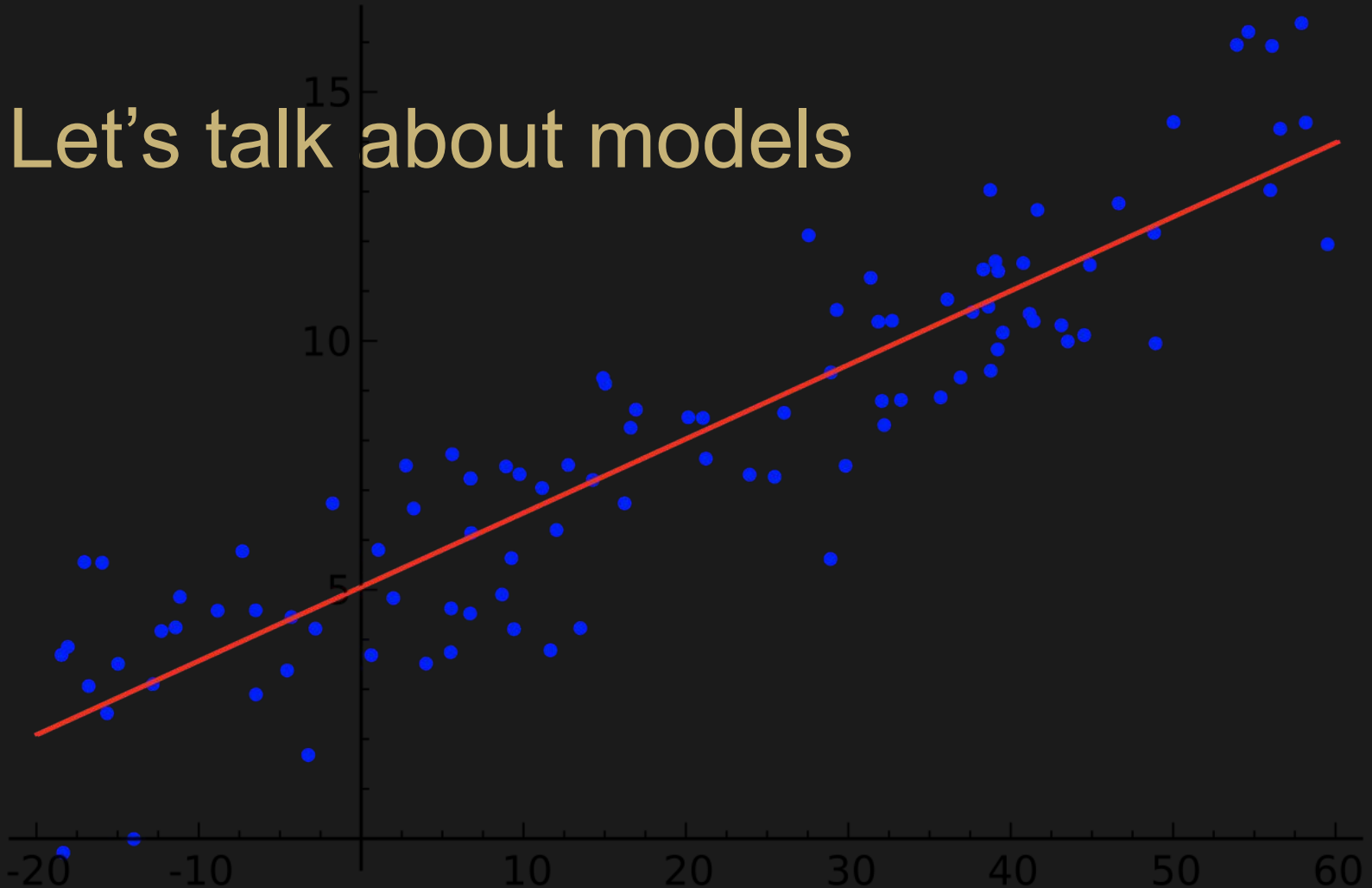



Thinking about the business
It’s, you… you’re the scientist in this metaphor now
The data
Your model ->

1. You have the data, you know its shape, if feels natural in your hands.

2. You know the business value you want to deliver

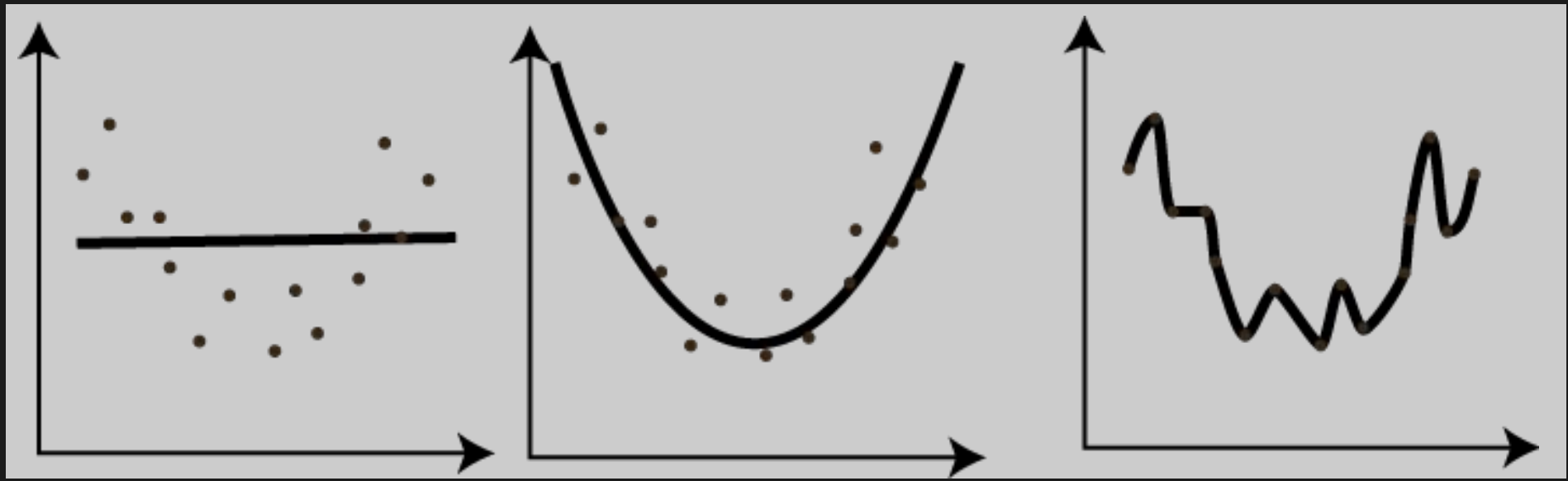
Let’s talk about models



“In nearly every detective novel... there comes a time when the investigator has collected all the facts he needs for at least some phase of his problem. These facts often seem quite strange, incoherent, and wholly unrelated. The great detective, however, realizes that no further investigation is needed at the moment, and that only pure thinking will lead to a correlation of the facts collected. So he plays his violin, or lounges in his armchair enjoying a pipe, when suddenly, by Jove, he has it! Not only does he have an explanation for the clues at hand but he knows that certain other events must have happened. Since he now knows exactly where to look for it, he may go out, if he likes, to collect further confirmation for his theory.”

- Einstein & Infeld, The Evolution of Physics 1938

Let’s talk about models



$$f(a, b, c, \ldots) + \varepsilon = y$$

f(a, b, c, ...) + ε = y

Data is good. More data is better

Try to move as much as possible from the ε into the function

Maybe b comes from an external API

Maybe c is too complicated and needs to be split into d and e

Maybe g is derived from a function/calculation based on other records or parameters a and b

$f(a, b, c, \ldots) + \varepsilon = y$

Data is good. More data is better.
Unless it's not

Sometimes b and c are just confusing the algorithm

Methods of dimensionality reduction or principle component analysis help extract a signal from noise, and help prevent overfitting

$$f(a, b, c, \ldots) + \varepsilon = y$$

In any case, you have to understand how you want to get value out of your $f(a, b, c, \ldots) + \varepsilon = y$

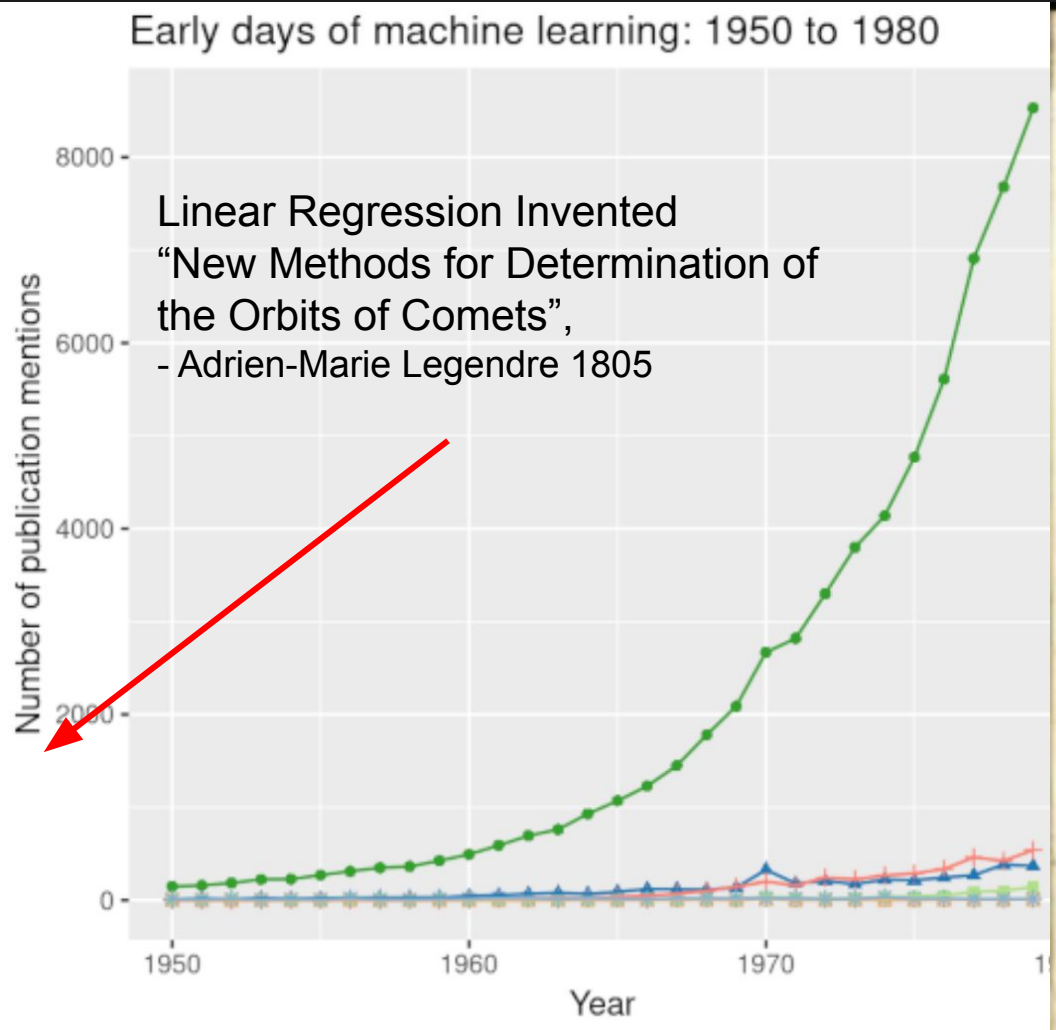
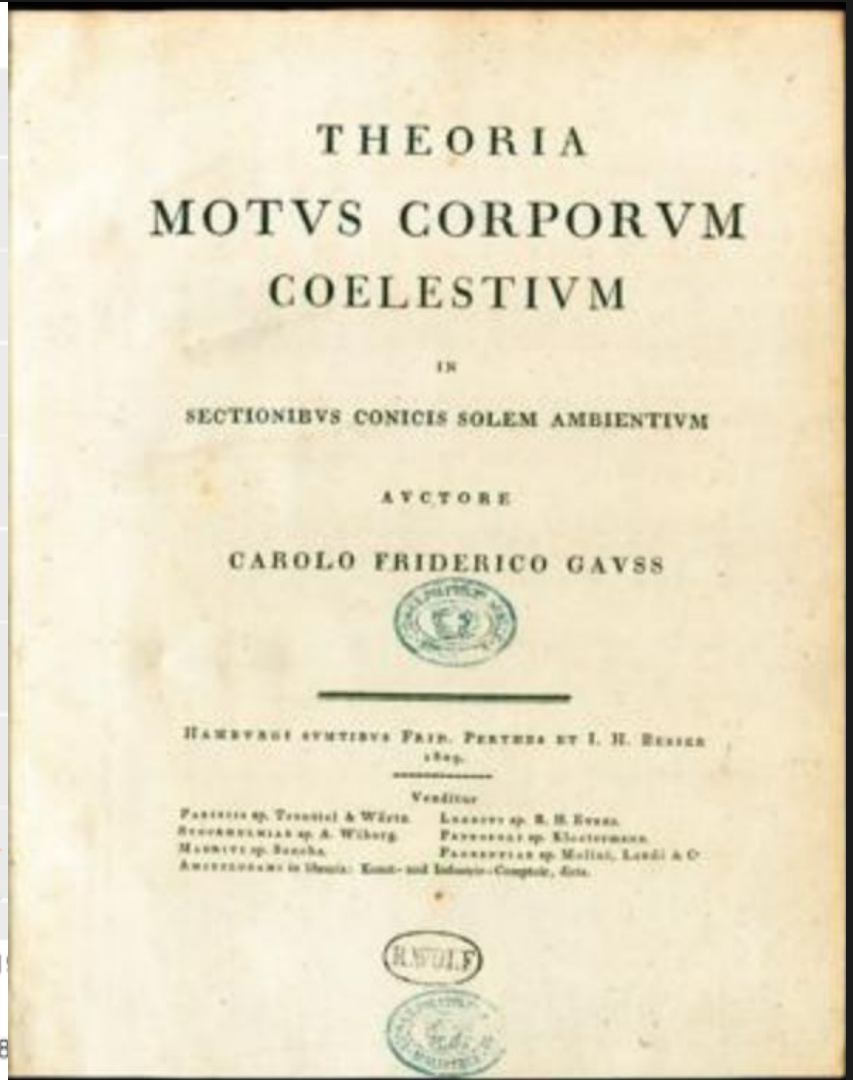


Figure 1: Overall publication rates of machine learning models from 1950 to 198



THEORIA

MOTVS CORPORVM

COELESTIVM

IN

SECTIONIBVS CONICIS SOLEM AMBIENTIVM

AVCTORE

CAROLO FRIDERICO GAVSS

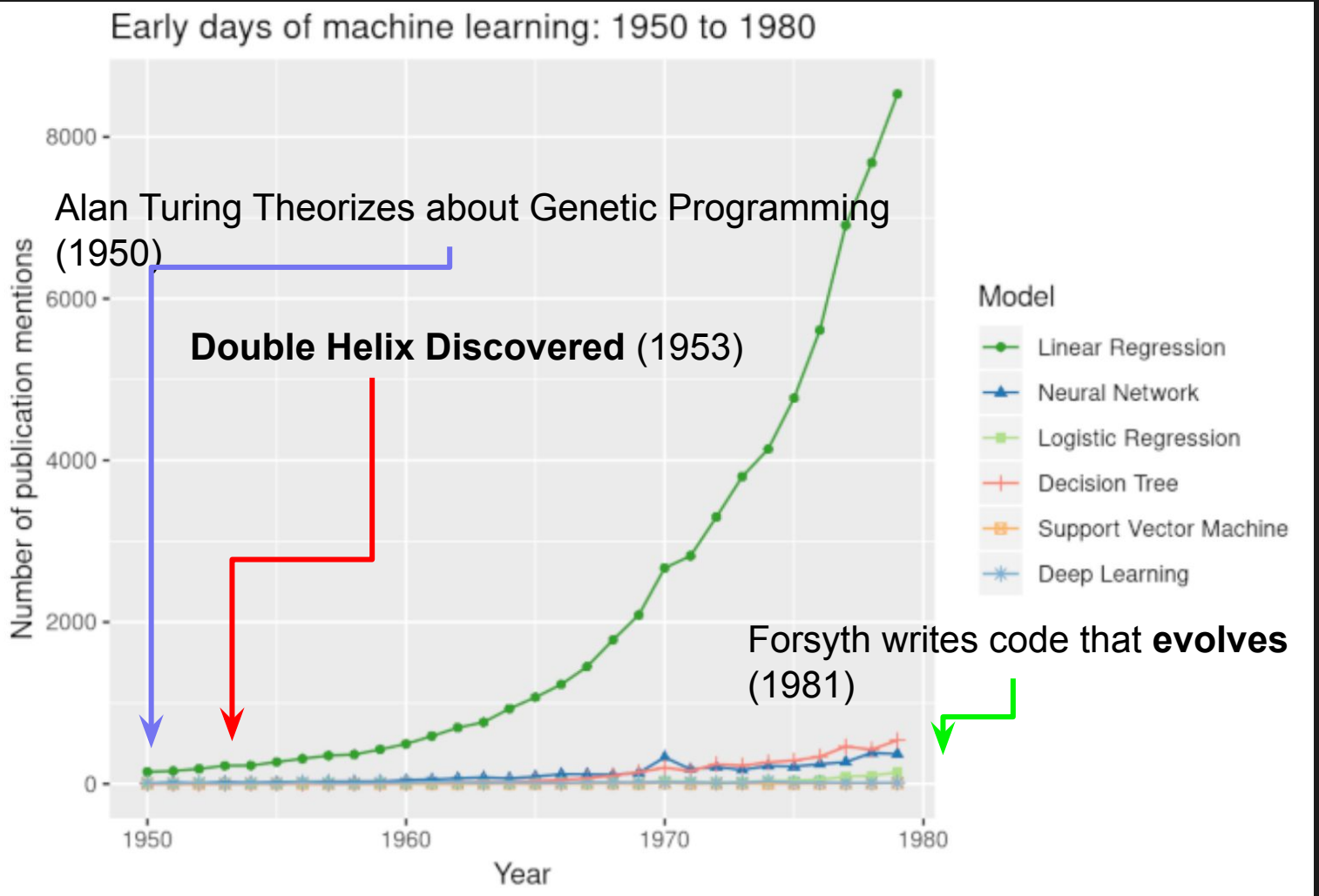


Figure 1: Overall publication rates of machine learning models from 1950 to 1980.

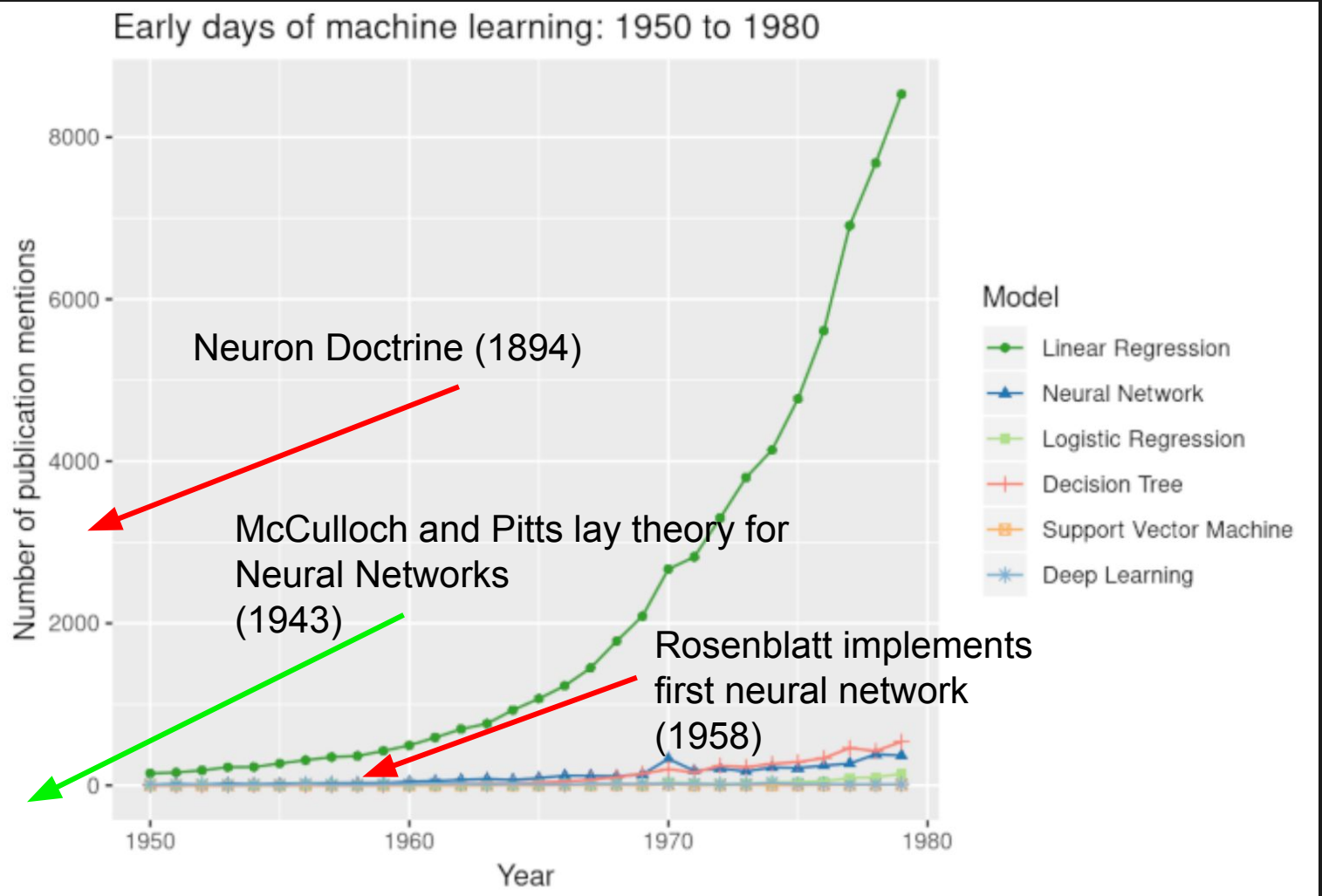


Figure 1: Overall publication rates of machine learning models from 1950 to 1980.

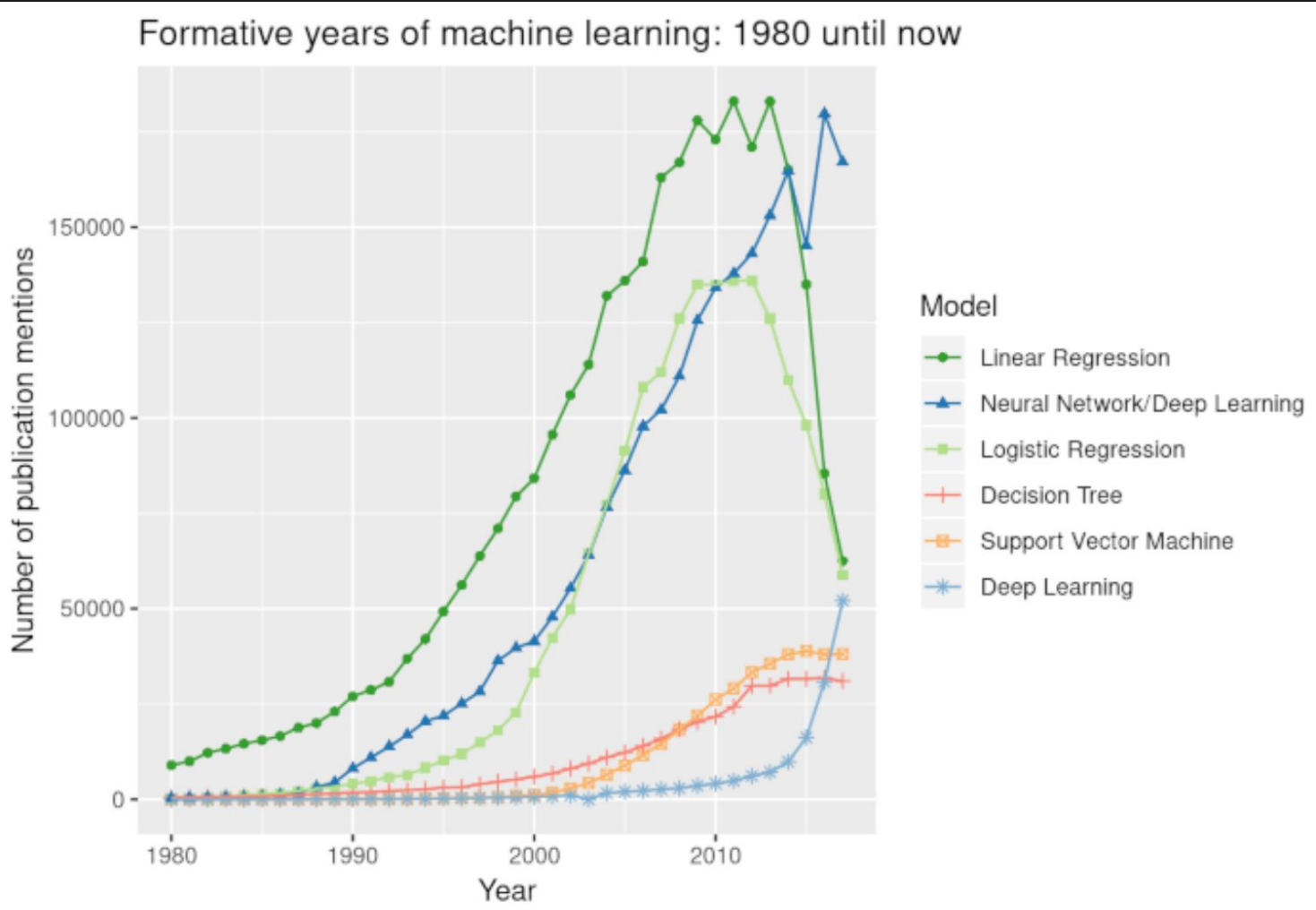
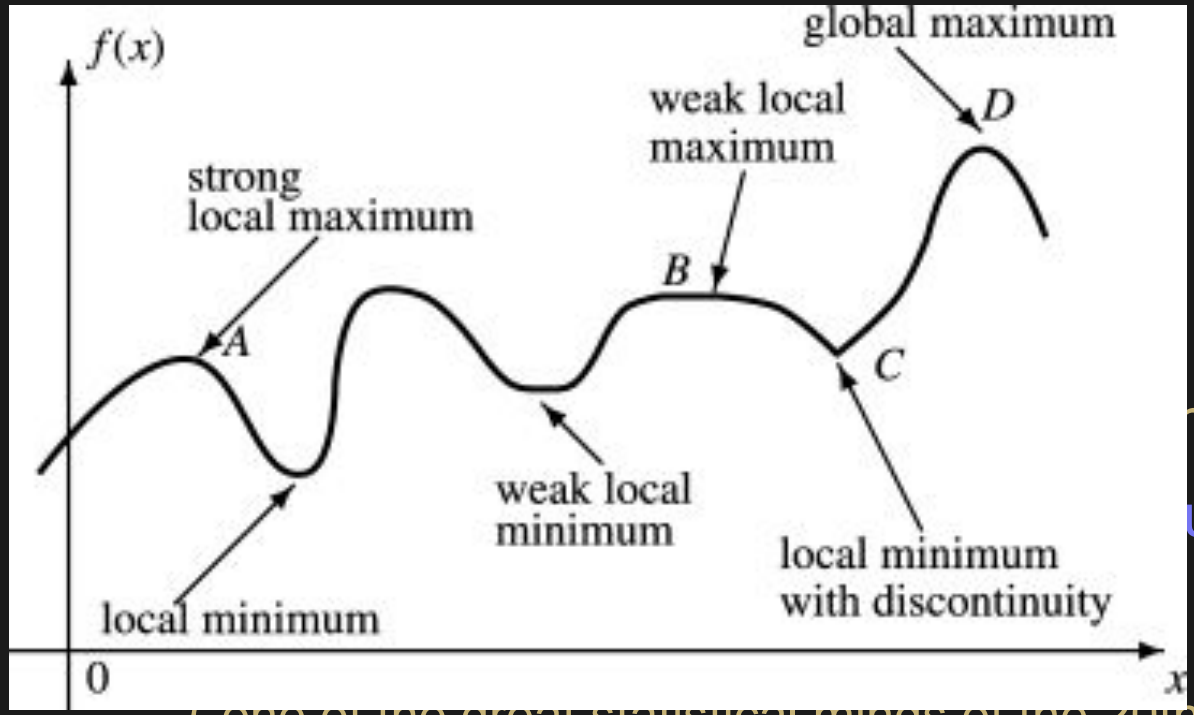


Figure 2: Overall publication rates for machine learning models from 1980 until now.

“Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.”

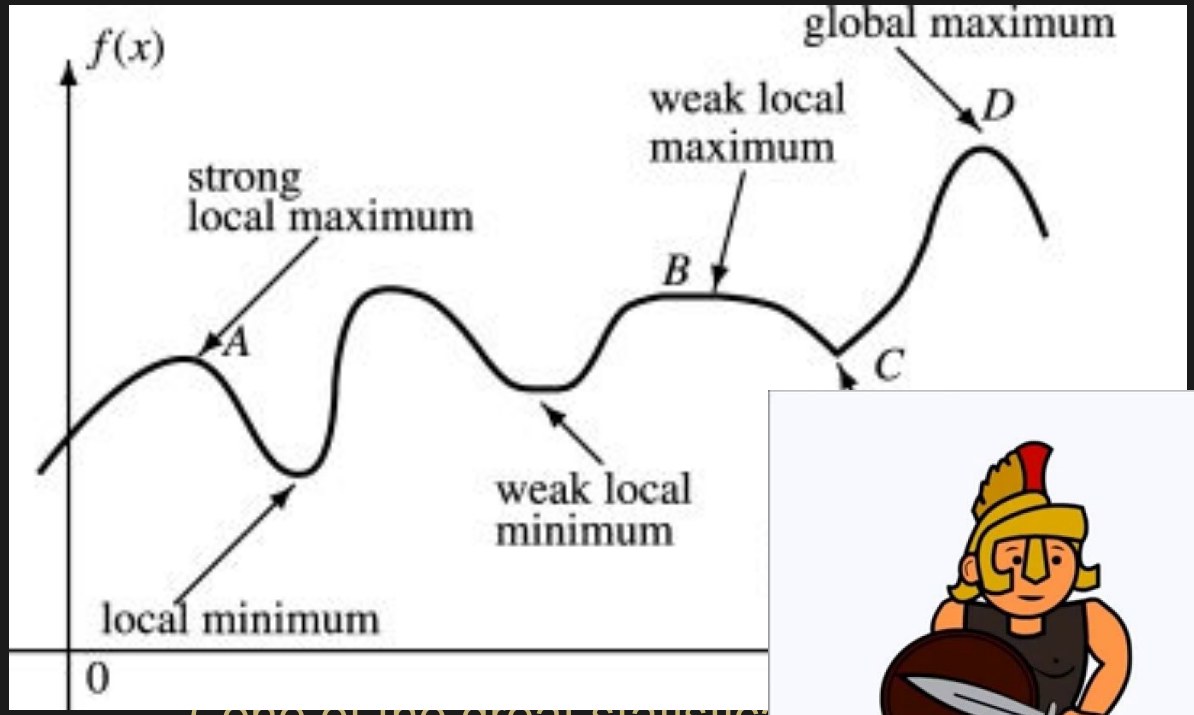
- George Box
(“one of the great statistical minds of the 20th century”)
“Empirical Model Building and Response Surfaces”, 1987



…he practical question is

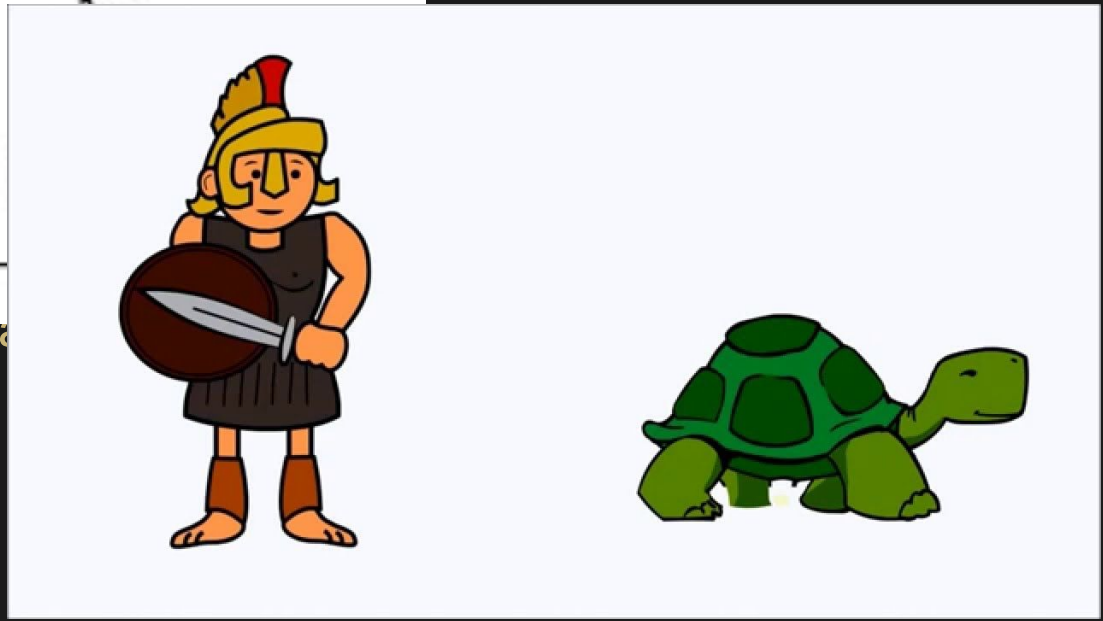
…useful.”

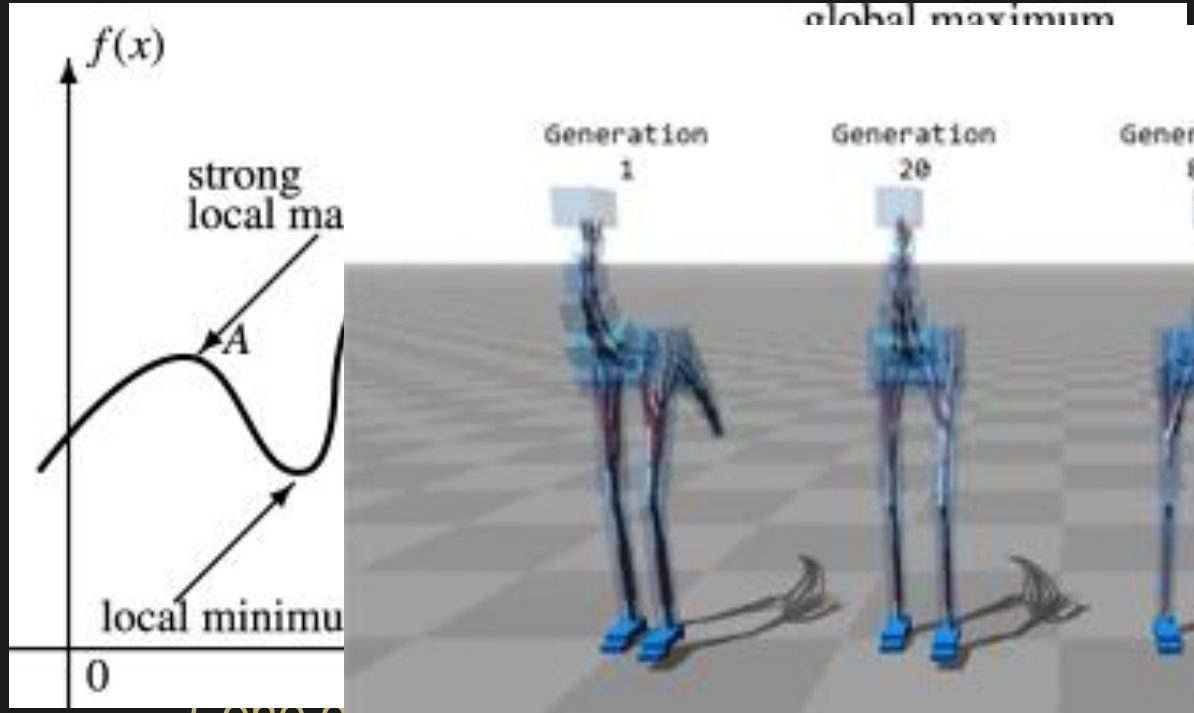
(“one of the great statistical minds of the 20th century”)
“Empirical Model Building and Response Surfaces”, 1987



(one of the great statistica

“Empirical Model Building

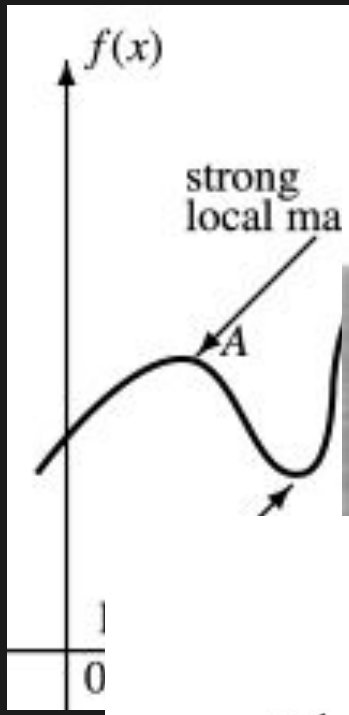


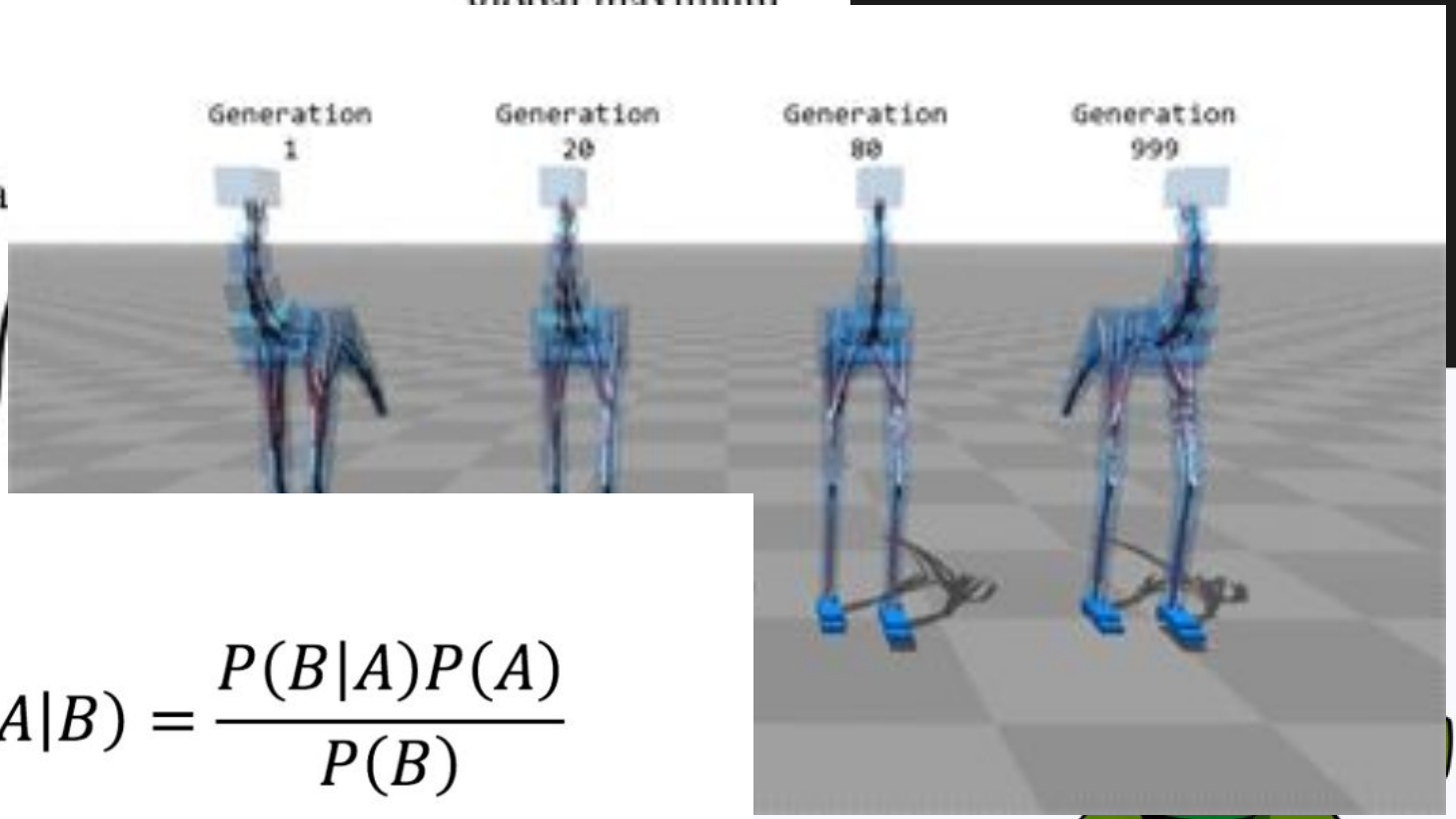


global maximum

(one o

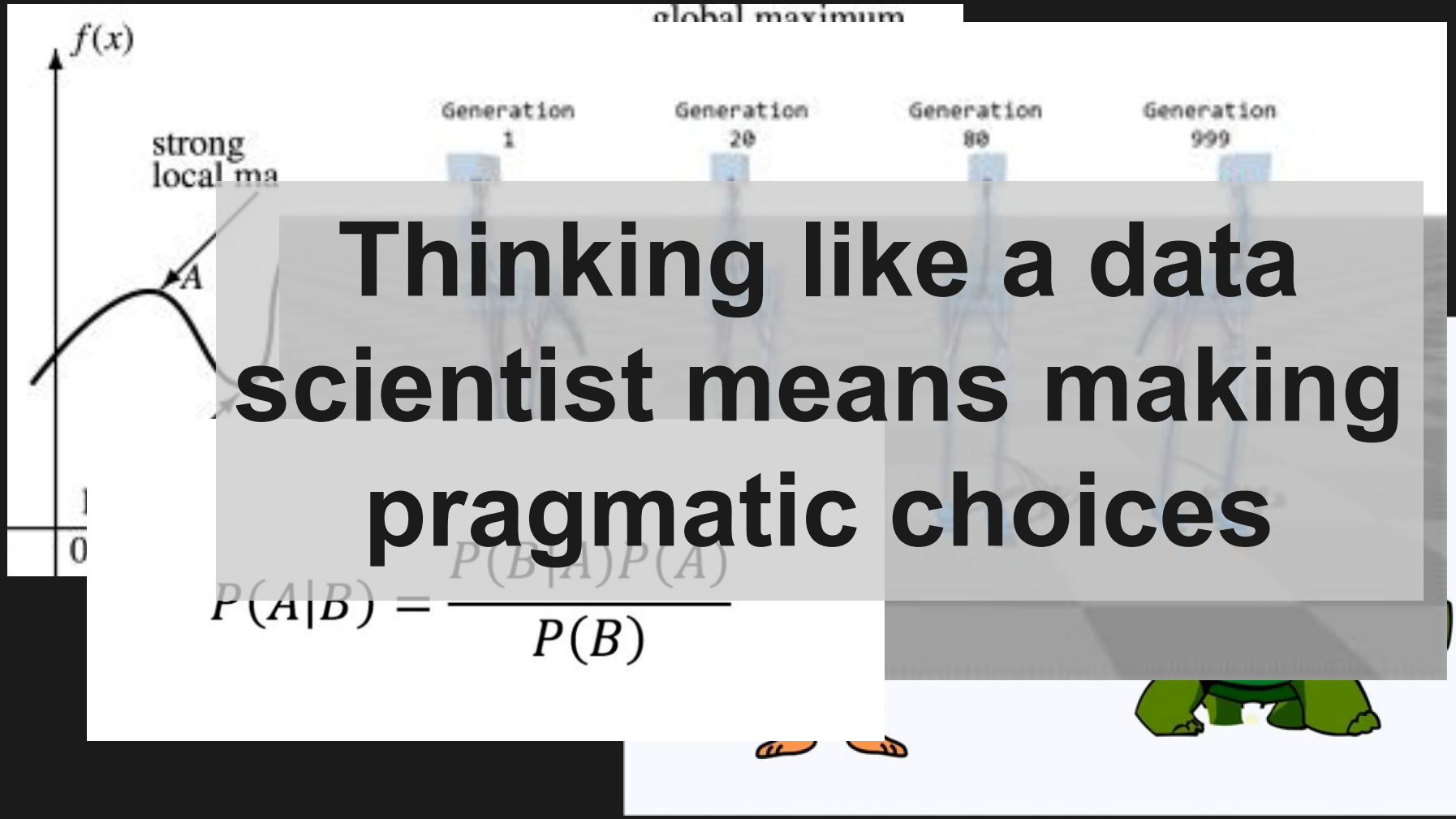
“Empiri





$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Thinking like a data scientist means making pragmatic choices



Algorithmic accountability BINGO

Trolley problem	Isaac Asimov's 3 laws of robotics	Apocalyptic doomsday scenario	Technological solutionism
Bias as something that exists outside of society	Namedropping of ethical theories (deontology, utilitarianism, virtue ethics)	"We need more oversight!" (But we don't know how...)	Algorithmic accountability not knowing on accountability theory
Self-driving cars	GDPR (the savior)	Call for ethical classes for data scientists/engineers	AI efforts
Hippocratic oath for data scientists	Robots everywhere!!?!	Right to explanation (and the belief that that will fix everything)	Predictive policing/parole recommendation

By Maranke Wieringa

Thinking like a data scientist means making pragmatic choices

To conform to Trump's policies, Reuters has learned, ICE modified a tool officers have been using since 2013 when deciding whether an immigrant should be detained or released on bond. The computer-based Risk Classification Assessment uses statistics to determine an immigrant's flight risk and danger to society.

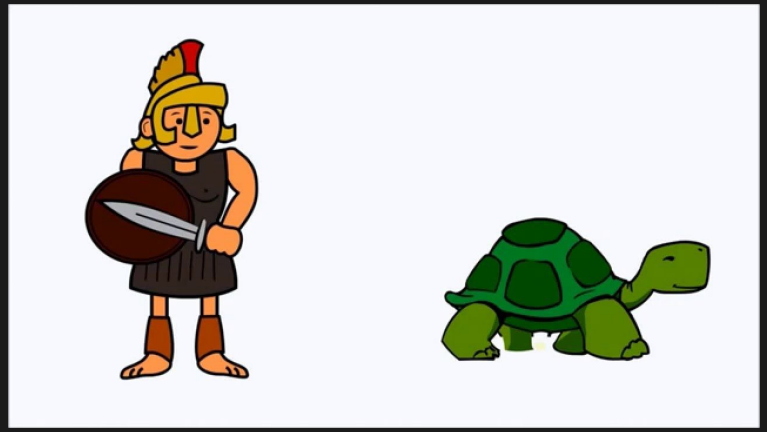
Previously, the tool automatically recommended either "detain" or "release." Last year, ICE spokesman Bourke said, the agency removed the "release" recommendation, but he


♪REVOLUTION♪

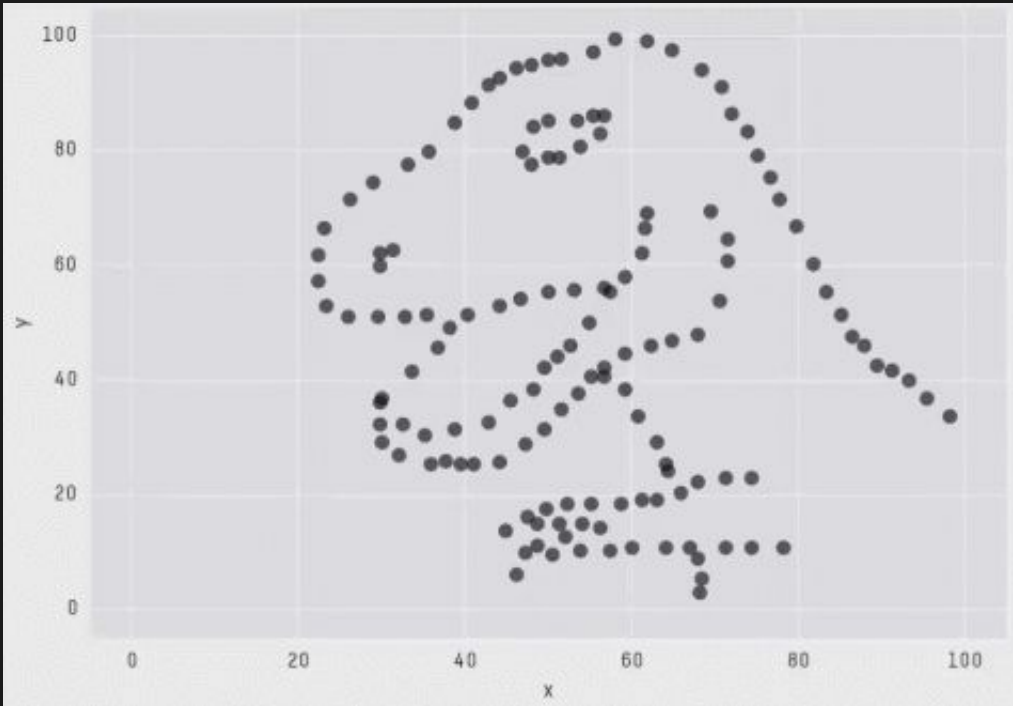
Thank you!

I’m Em Grasmeder
(@emilyagras)

Let’s talk about models



Models The first step towards using a model



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Linear Regression

Bivariate analysis using
linear least square opitmisation

Machine learning model trained with gradient descent using the partial direvative of the square error cost function

Deep learning model trained with Bayesian approach by minimizing the Kullback–Leibler divergence between the true posterior and the approximated variational distributions.