Thinking Like A Data Scientist

Em Grasmeder
ThoughtWorks Data Witch
@emgrasmeder

About Me

- Name: Em
- Pronoun: they/them
- Job: ThoughtWorks data witch
- Background: Graduate research in economics

- Generalist within data space

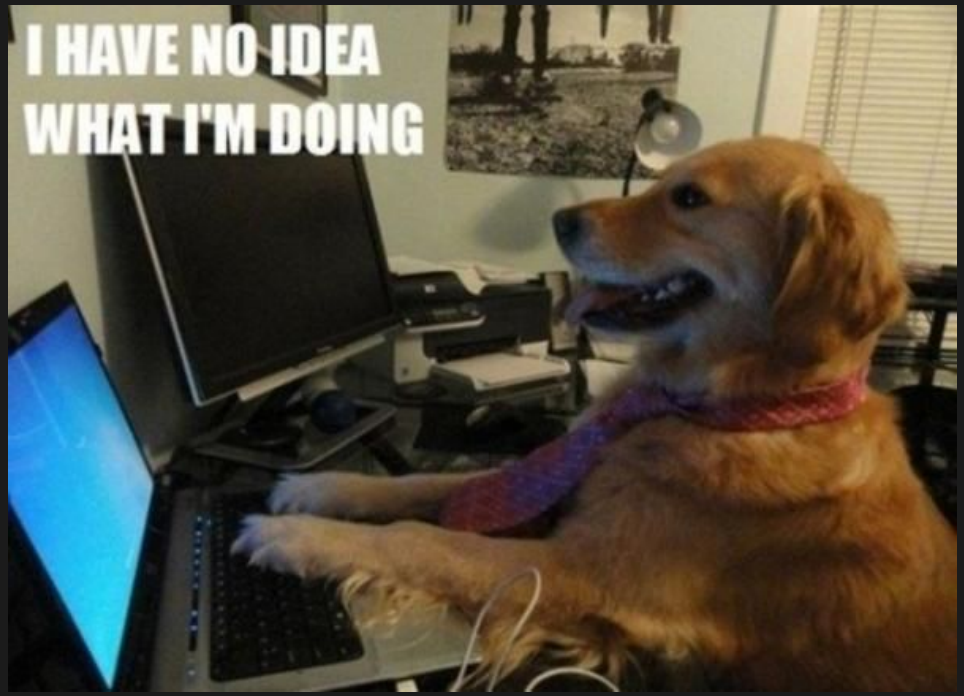
<Data Science Identity Crisis>

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

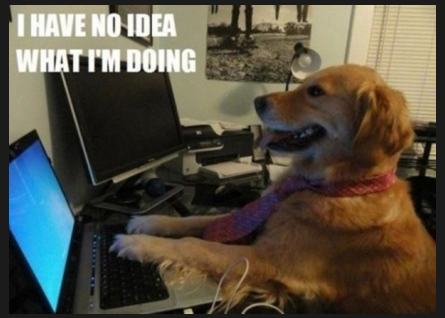
- Predictions, categorization, clustering
- Write software
- Visualizations

What Does a Data Scientist Do

- Predictions, categorization, clustering
- Write software
- Visualizations
- ...magically fix the 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!

Cholera Facts (yay!)

Deadly bacteria that can kill within hours

The water in your body just comes out from everywhere

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, for example, 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 Great Stink”

Deadly, exploding cesspits

Waste from houses, slaughterhouses and factories dumped in the Thames

“The Great Stink”

“The Great Stink”

I can certify that the offensive smells, even in that short whiff, have been of a most head-and-stomach-distending nature

Charles Dickens

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

f(proximity to bad air,
sinful,
too much blood,
other old fashioned belief)
+ ε = probability of contracting cholera

The Broad Street Cholera Outbreak of 1854

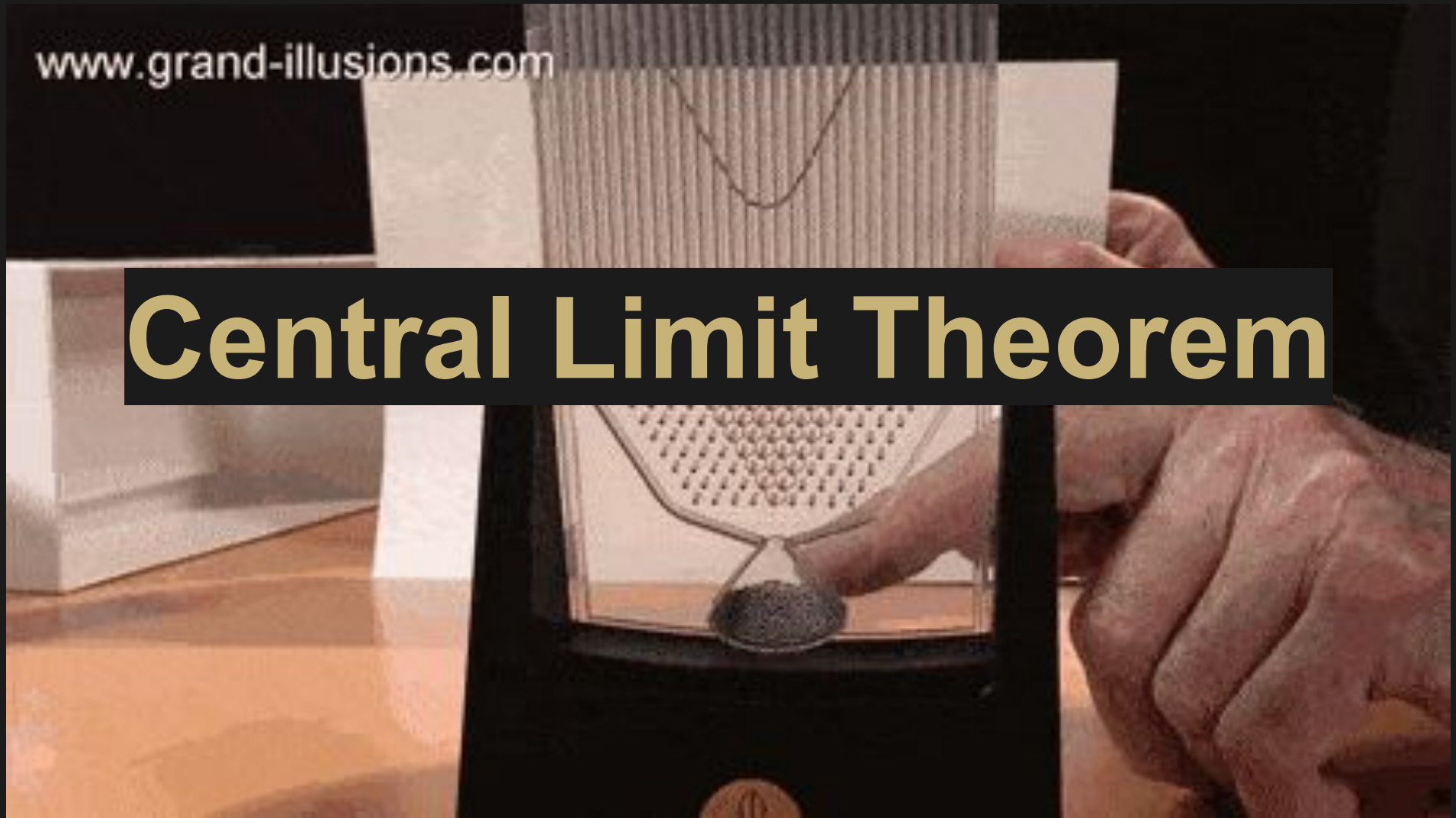
Formally write your hypothesis

- 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

* A model is another way of writing a hypothesis



Central Limit Theorem

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

* A model is another way of writing a hypothesis

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!

Collecting more data

Workers at brewery were unaffected while their families died

Children of some families died while their families lived

There was this one woman, a complete outlier, the only person in her neighborhood to die

Formally write your hypothesis

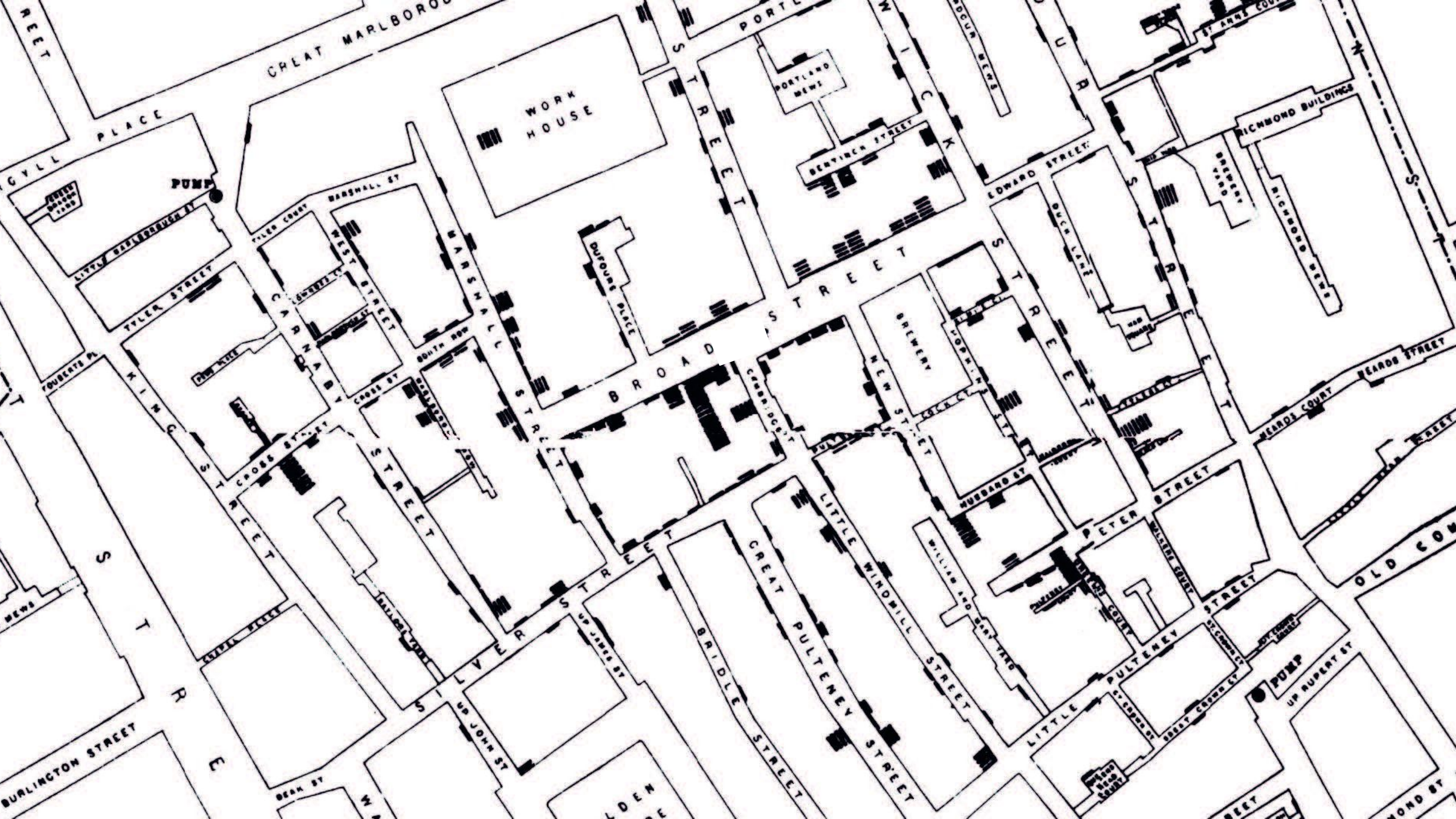
- H_0: People living in equally odorous parts of town will have a uniform likelihood of contracting cholera

Formally write your hypothesis

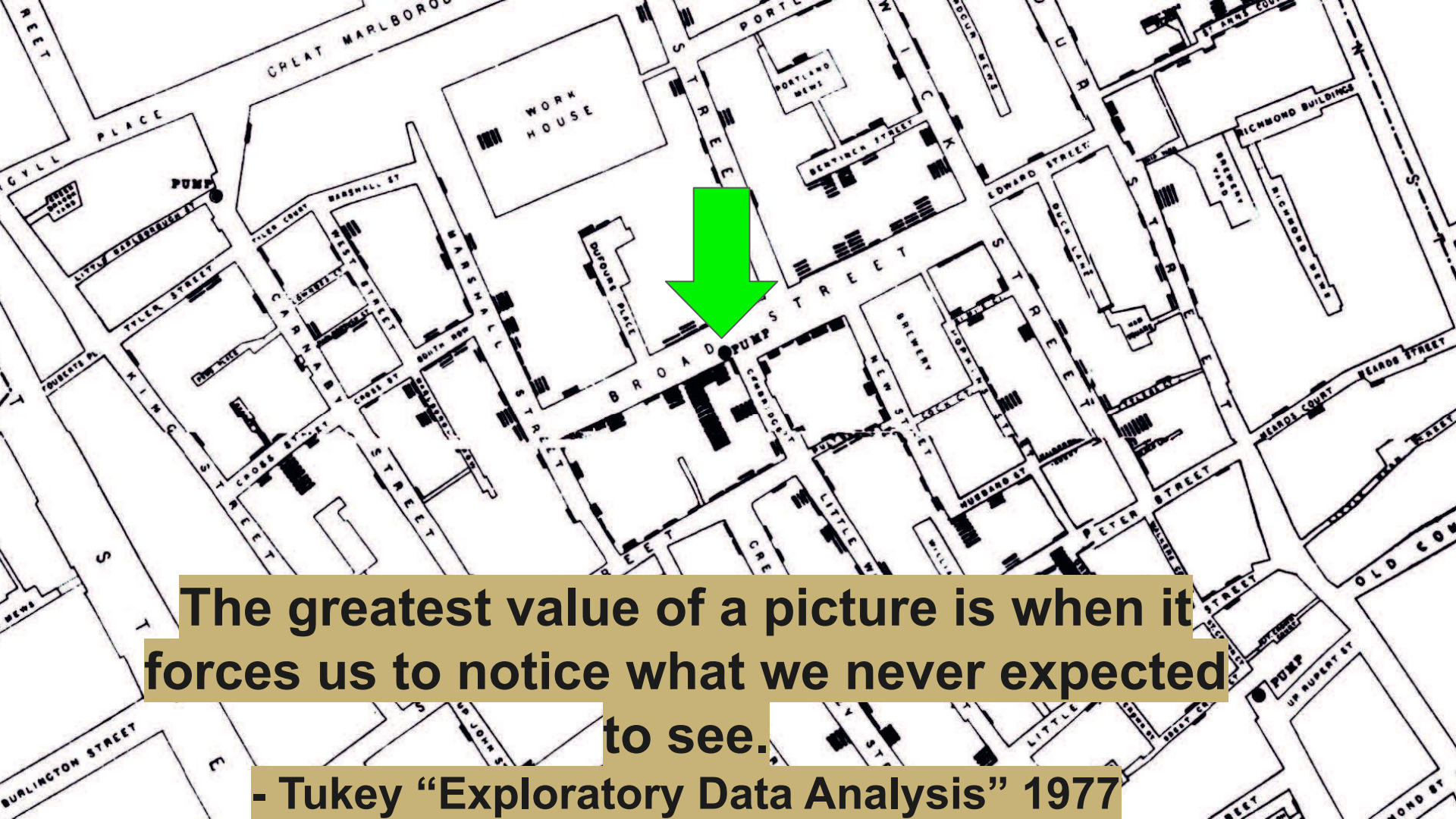
- H_0: People living in equally odorous parts of town will have a uniform likelihood of contracting cholera
- H_A: People who drink contaminated poo-water have a uniform likelihood of contracting cholera

For
hyp

- H
oc
ha
of
- H
co
ha
of



WORK HOUSE
PUMP
BROAD STREET
MARSHALL STREET
MARSHALL ST
CARNABY STREET
KING STREET
SILVER STREET
GREAT PULTENEY STREET
LITTLE WINDMILL STREET
BRIDLE STREET
PETER STREET
LITTLE PULTENEY STREET
BREWERY
PORTLAND MEWS
BERWICK STREET
RICHMOND BUILDINGS
RICHMOND MEWS
EDWARD STREET
DUFOURS PLACE
TYLER STREET
TYLER COURT
CROSS STREET
CROSS ST
WEST STREET
SOUTH ROW
NEW STREET
HEARDS COURT
HUSBAND ST
CAMBRIDGE ST
UP JAMES ST
UP JOHN ST
UP RUPERT ST
BURLINGTON STREET
PUMP



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

- Tukey “Exploratory Data Analysis” 1977

Collecting more data

Workers at brewery drank beer which required boiling the water

The children who died went to school near the infected well, far from their homes and family

That one woman, she just loved the flavor of the water from that poo-contaminated cholera well

So what are the lessons?

Data is good. More data is better

Visualize your data!

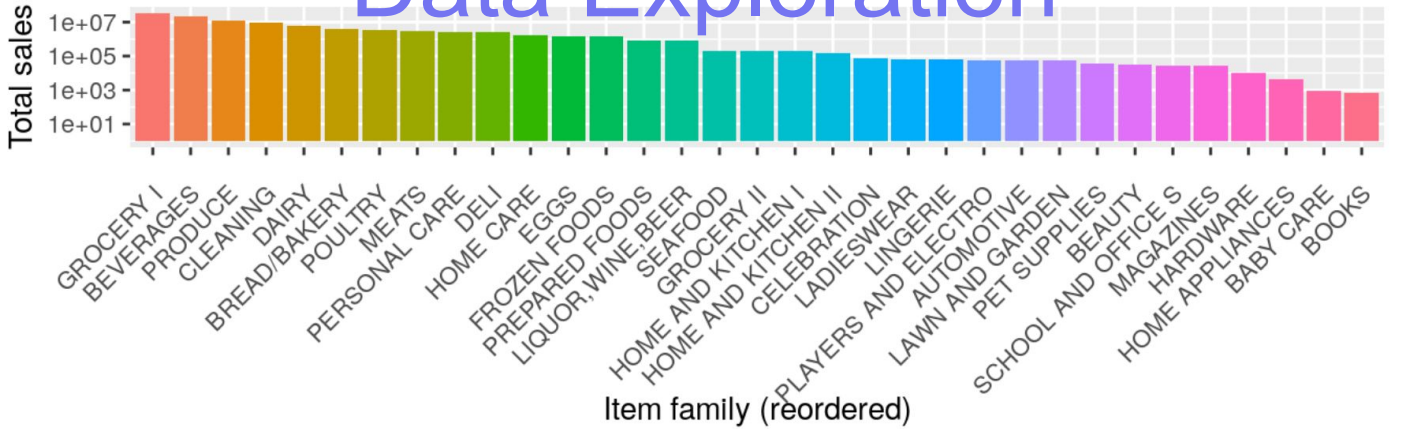
Do we really need machine learning for this?

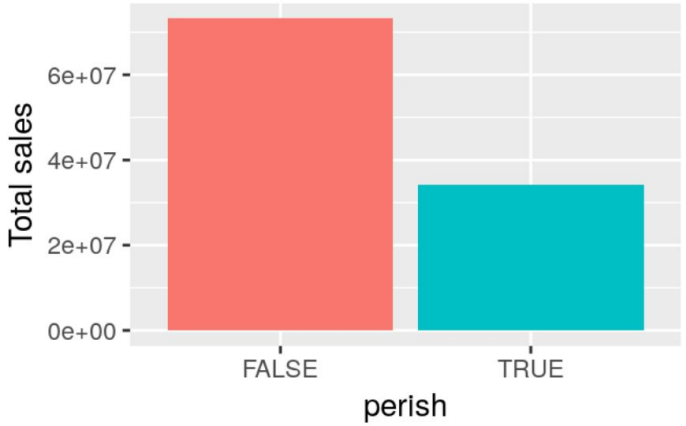
Data Exploration: Refining your mental model

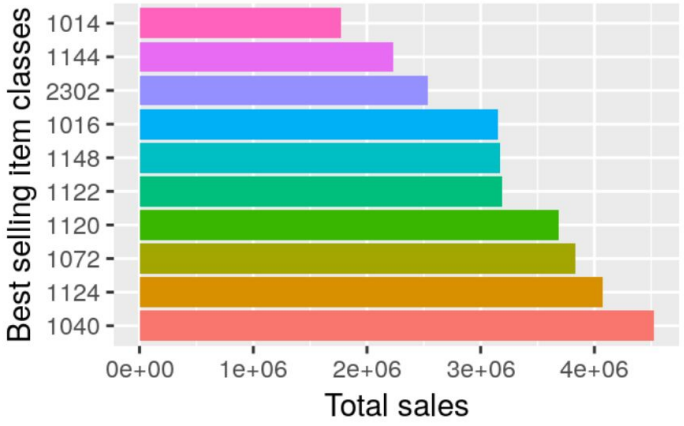
6.3 Items: family classification

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

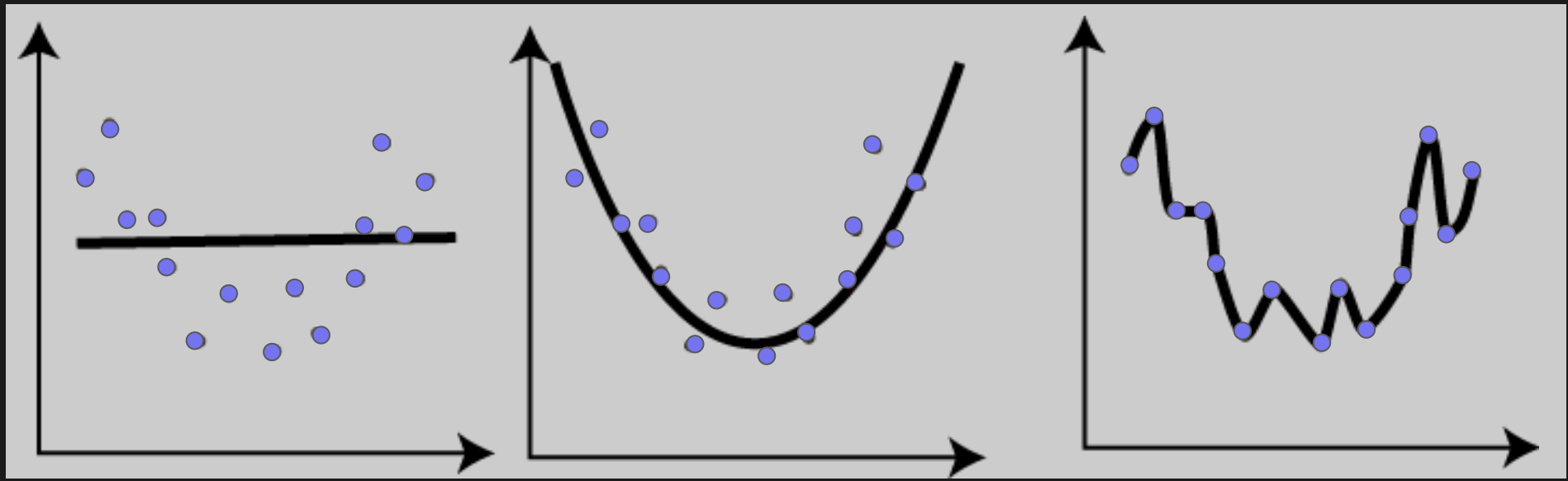
Code







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

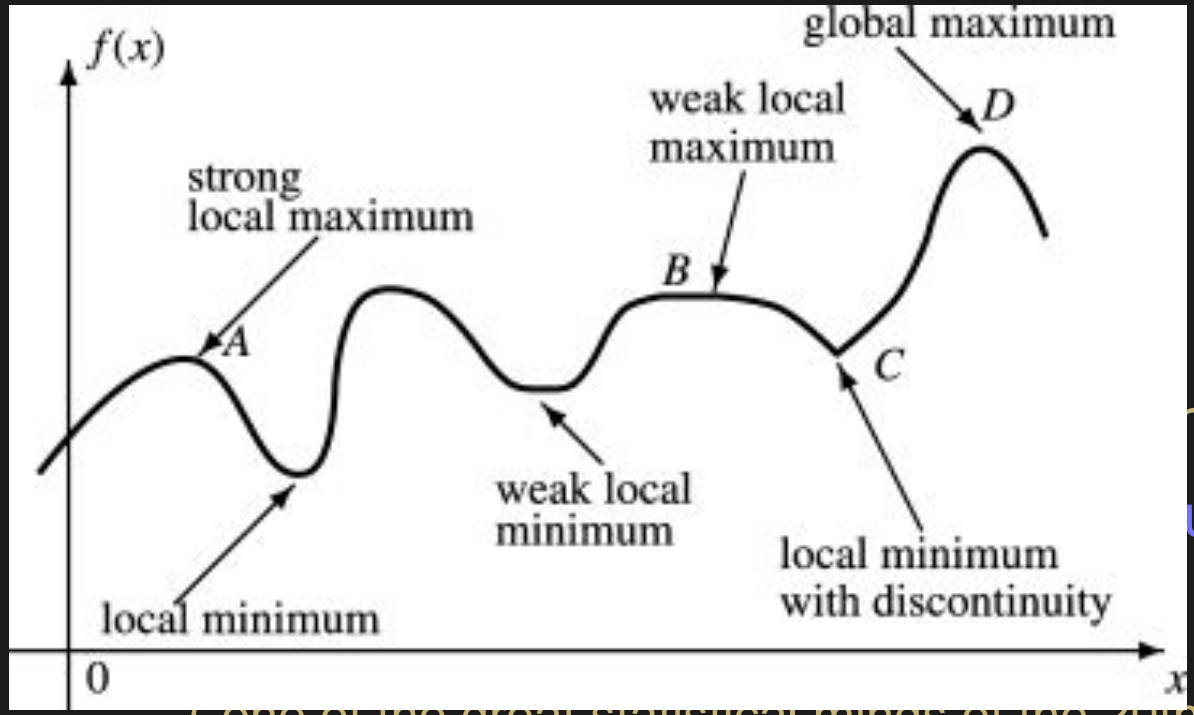
Sometimes b and c are just confusing the algorithm

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


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

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

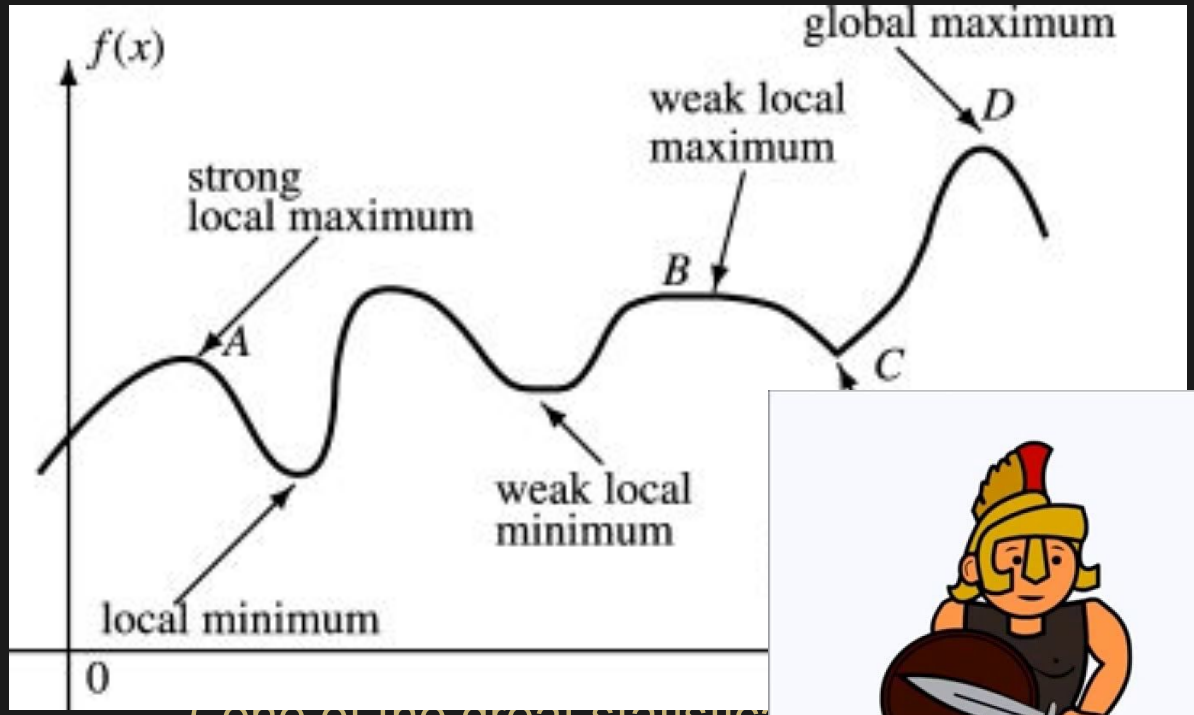
- George Box
(“one of the great statistical minds of the 20th century”)
“Empirical Model Building and Response Surfaces”, 1987



…he practical question is

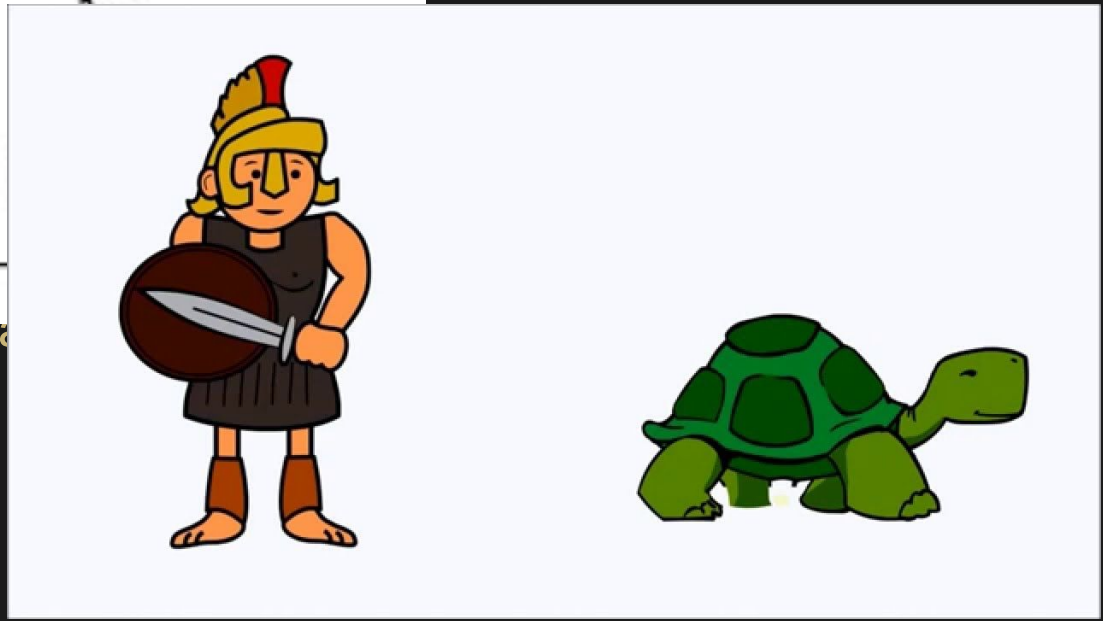
…useful.”

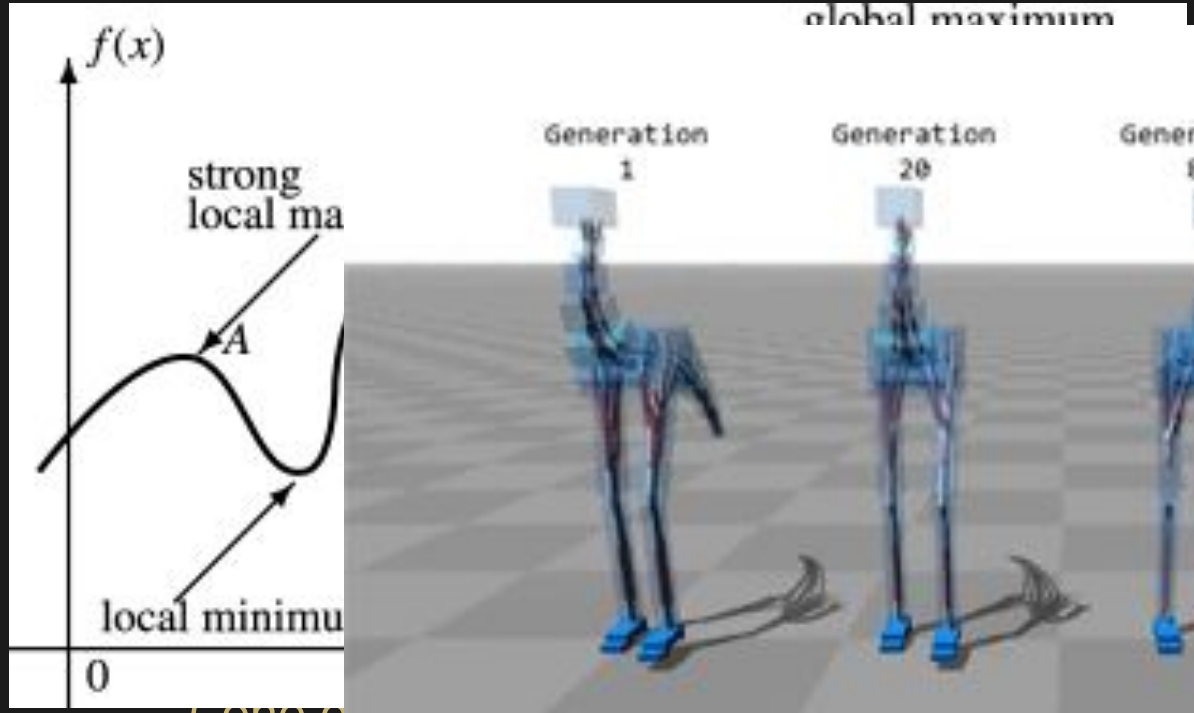
(“one of the great statistical minds of the 20th century”)
“Empirical Model Building and Response Surfaces”, 1987



(one of the great statistica

“Empirical Model Building

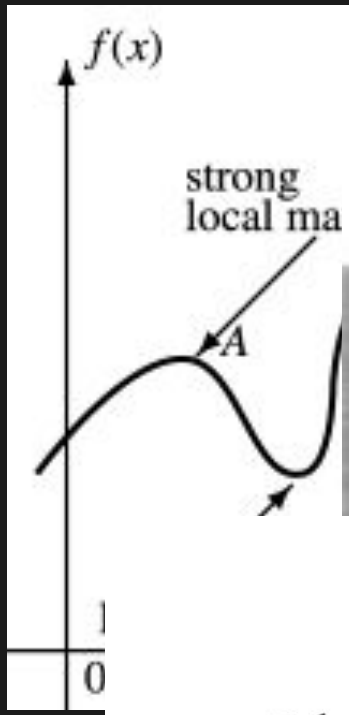


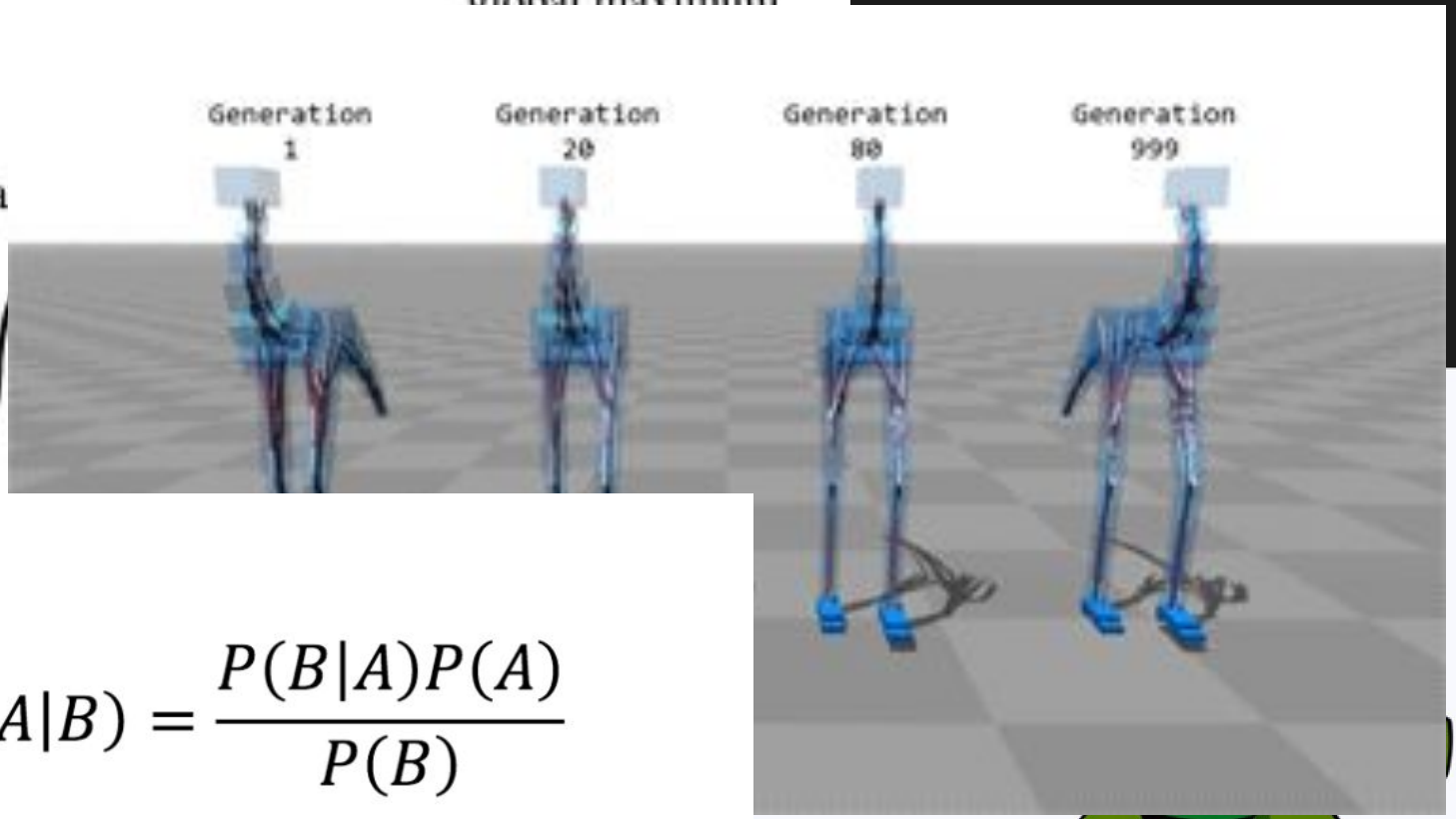


global maximum

(one o

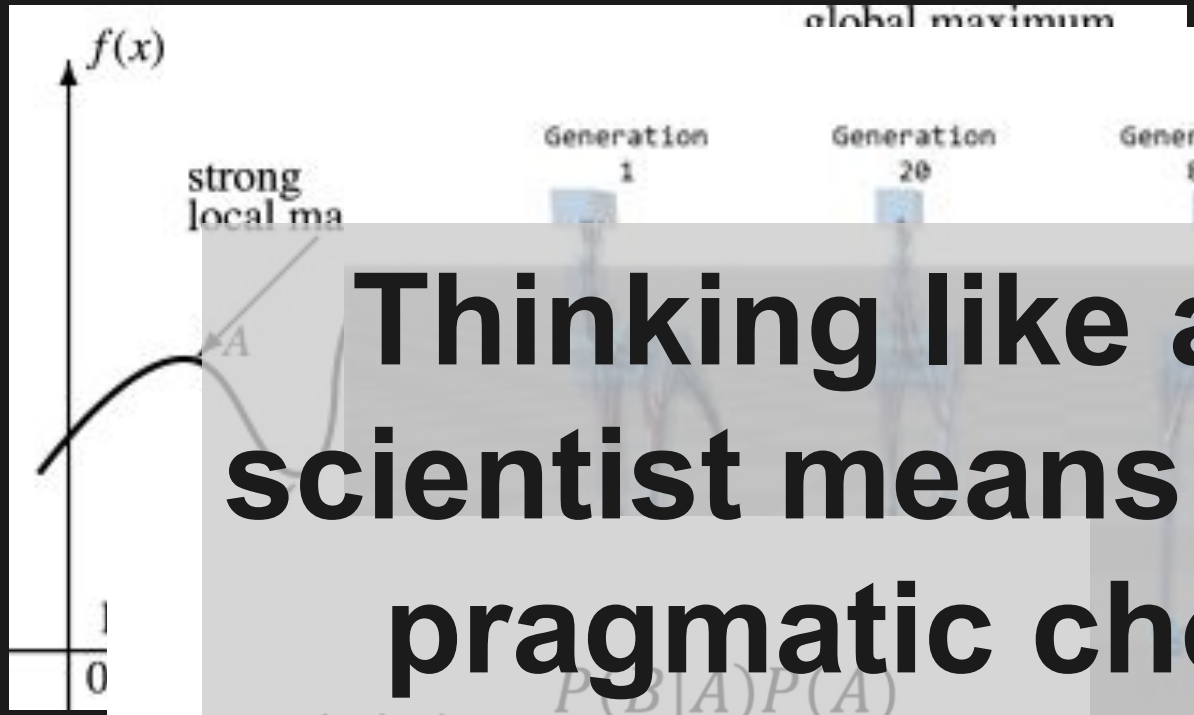
“Empiri



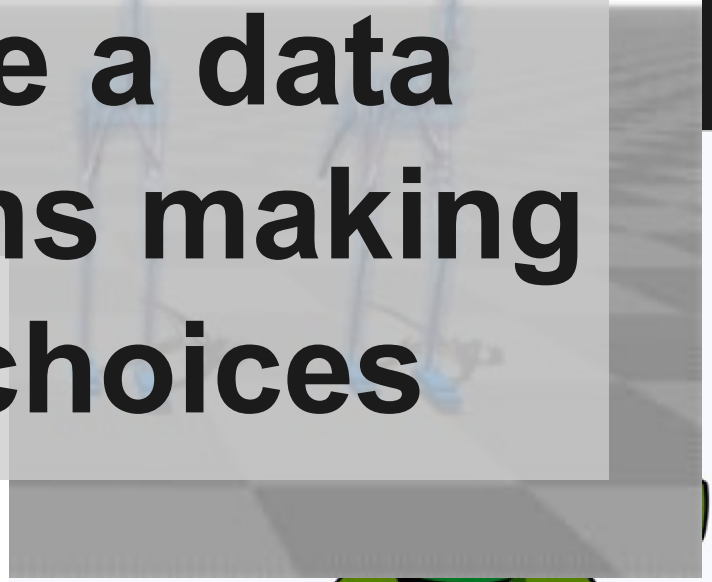


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Thinking like a data scientist means making pragmatic choices



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



Building Models

- Fun to do
- Difficult to master
- Not very different from science throughout the ages

- Not the hard part

Hamburg Cholera Outbreak of 1892

- Hamburg was an independent city-state
- They bought into "laissez-faire" economics
- Business would have been temporarily hurt by sanitation modernization

Hamburg Cholera Outbreak of 1892

- Other cities (e.g. nearby Altona) which invested in sanitation didn't have an outbreak
- 1.5% of Hamburg's population died
- People rioted

The Hard Part of doing Data Science

- Internal politics
- Overcoming bureaucracy
- Positioning you model to deliver value to users quickly
- Getting trust from internal stakeholders
- Building something meaningful, something good

Fun fact! The word “statistics” came about in the 18th century and meant "*science dealing with data about the condition of a state or community*"


THERE'S NO ETHICAL
SOFTWARE DEVELOPMENT
UNDER CAPITALISM


♪REVOLUTION♪

Thank you!

I’m Em Grasmeder

the ThoughtWorks Data Witch

@emgrasmeder on Twitter