Thinking Like A Data Scientist

Em Grasmeder
ThoughtWorks Data Witch
@emgrasmeder

About Me

- Pronoun is they/them
- ThoughtWorks consultant
- Graduate research in economics
  - I flunked out
- Generalist within data space

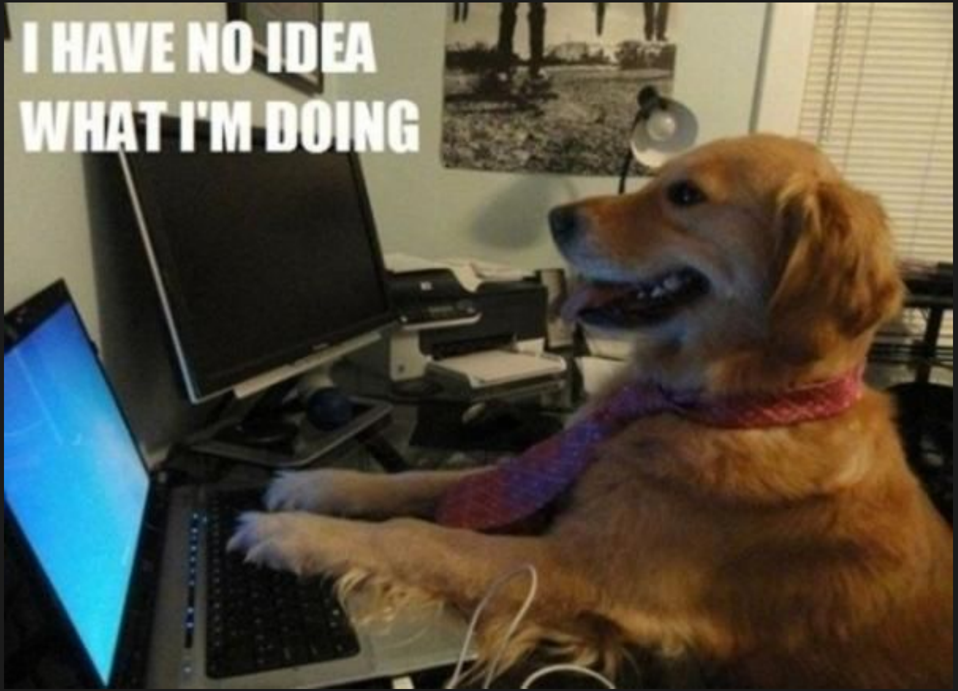
<Data Science Identity Crisis>

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

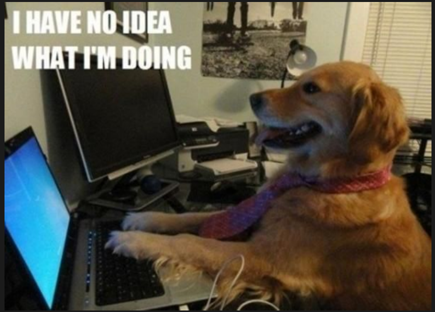
- Predictions, categorization, clustering
- Write software
- Visualizations

What Does a Data Scientist Do

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

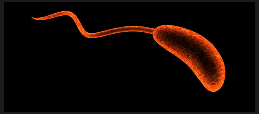
  If not...
- They might not be very useful

So how do data scientists actually think?

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

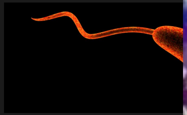
Cholera Facts (yay!)

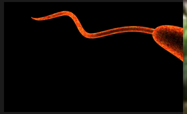
Deadly bacteria that can kill within hours



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.

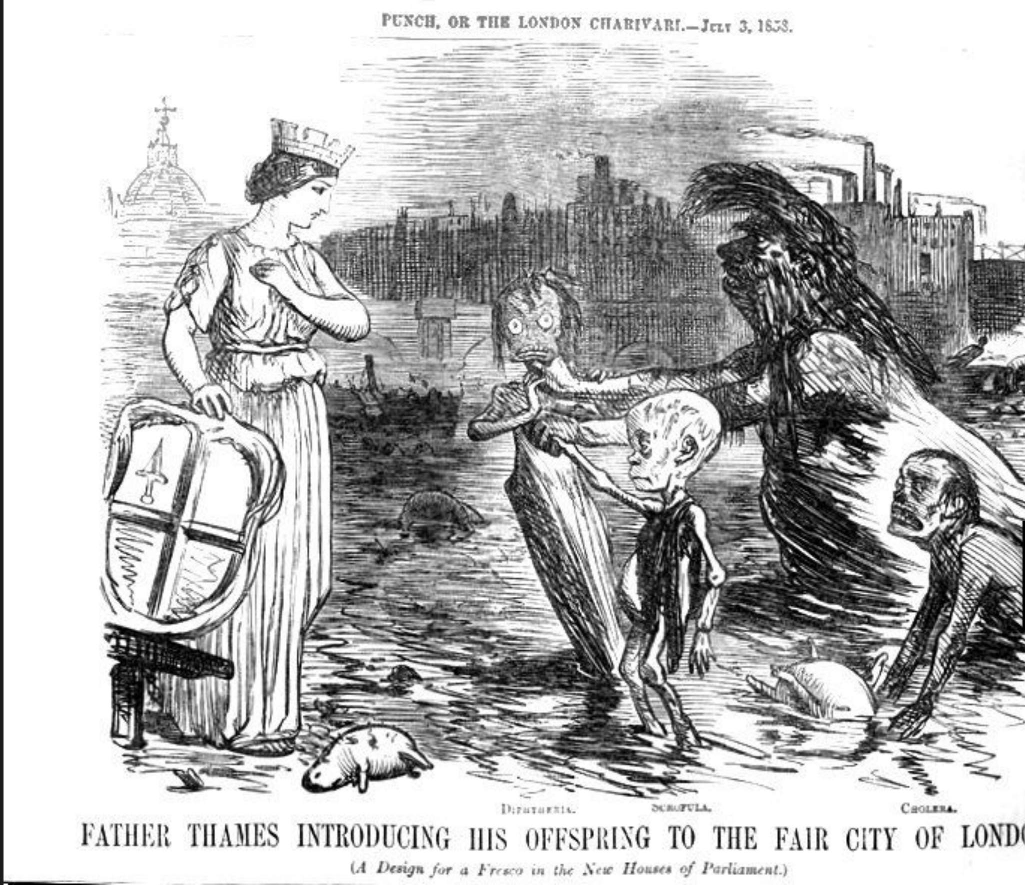
They thought it was "miasma" literally, bad air

“The Great Stink”

Deadly, exploding cesspits

Waste from houses, slaughterhouses and factories dumped in the Thames

“The Great Stink”



PUNCH, OR THE LONDON CHARIVARI.—July 3, 1858.

FATHER THAMES INTRODUCING HIS OFFSPRING TO THE FAIR CITY OF LONDON.

(A Design for a Fresco in the New Houses of Parliament.)

“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)
$+ \varepsilon = y$

The Broad Street Cholera Outbreak of 1854

Formally write your hypothesis

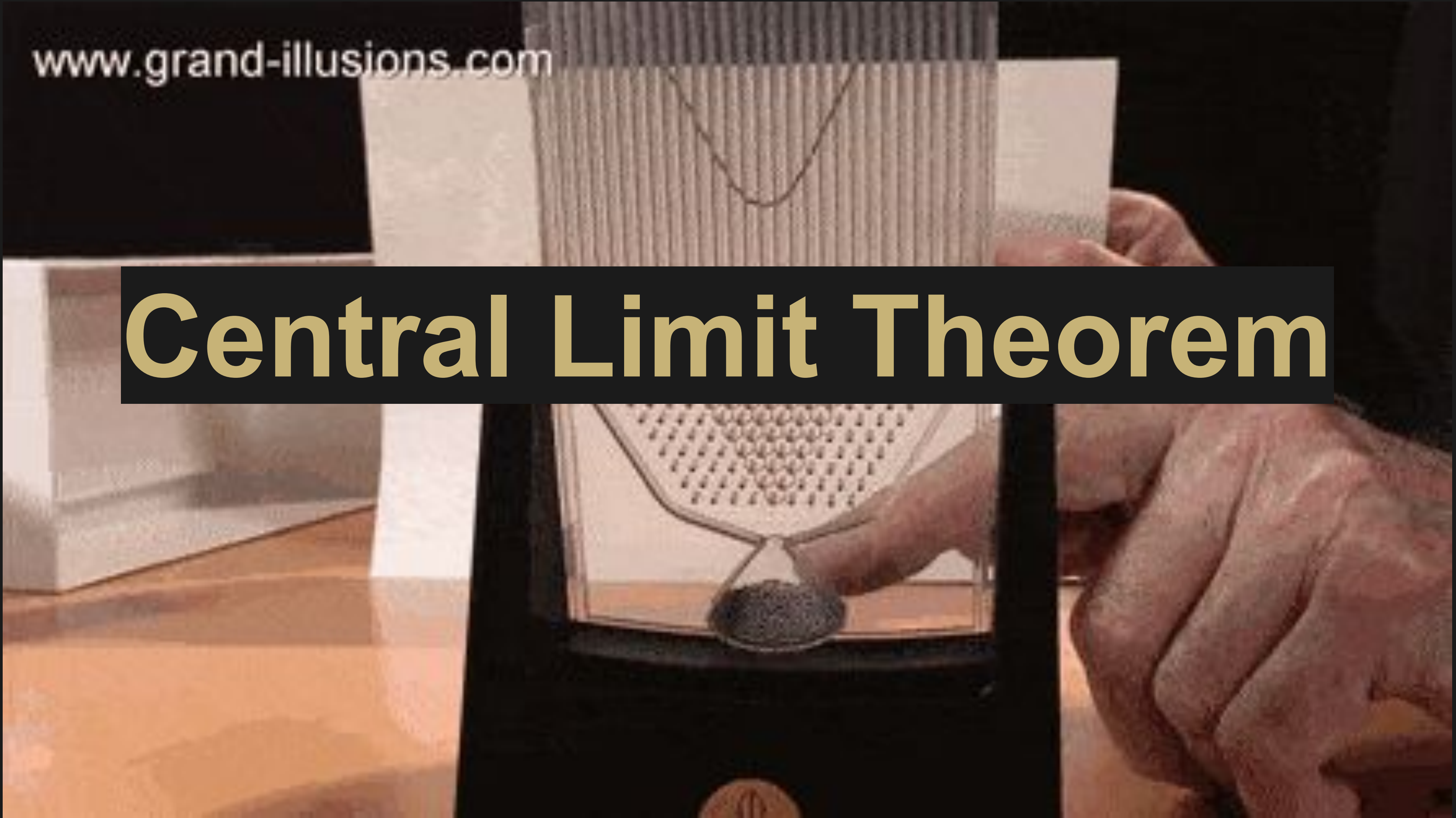
- H_0 is called the Null Hypothesis
- In 1850s England, the Null Hypothesis is “bad airs”

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

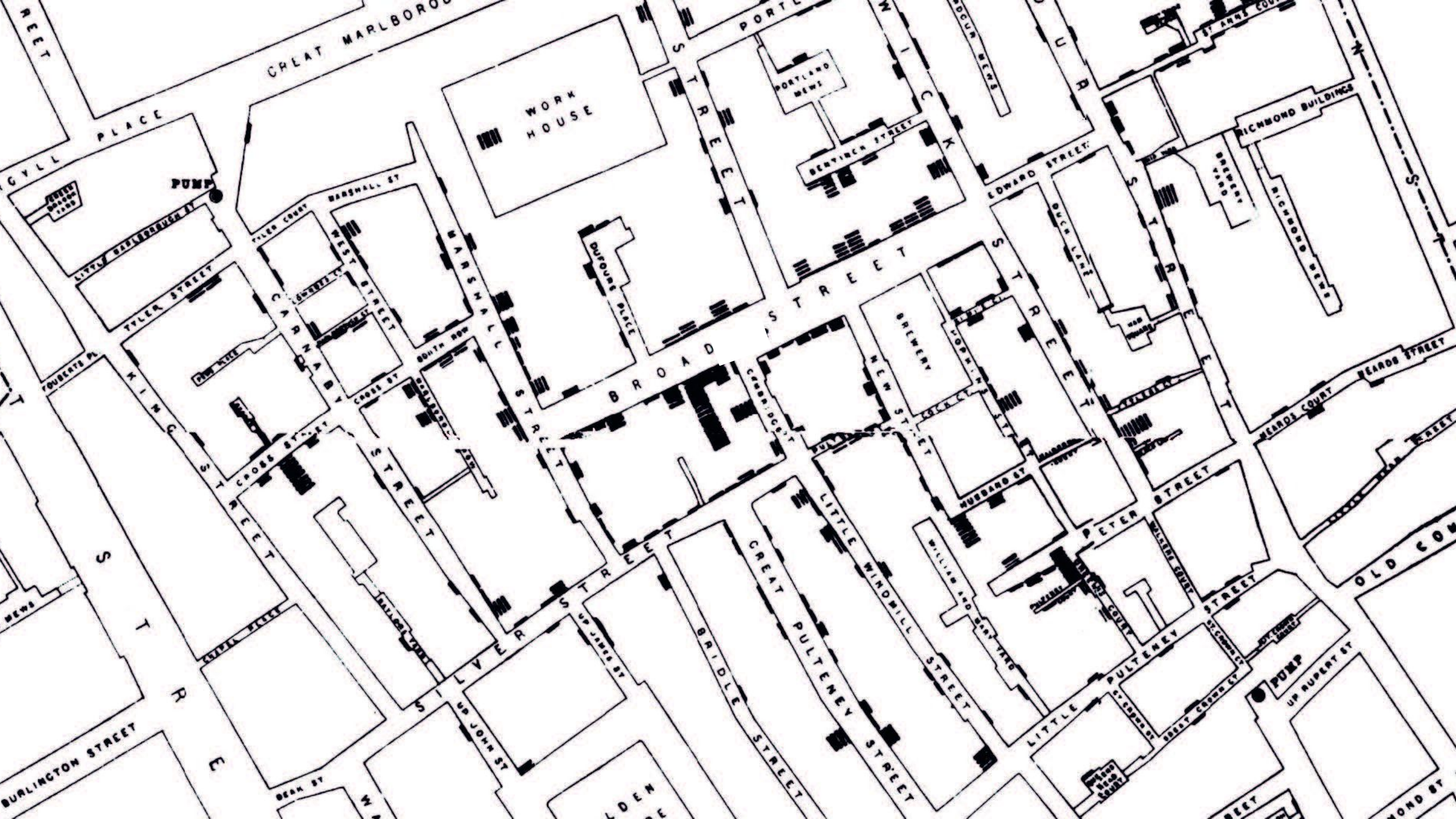
For
hyp

- H_0
oc
ha
of

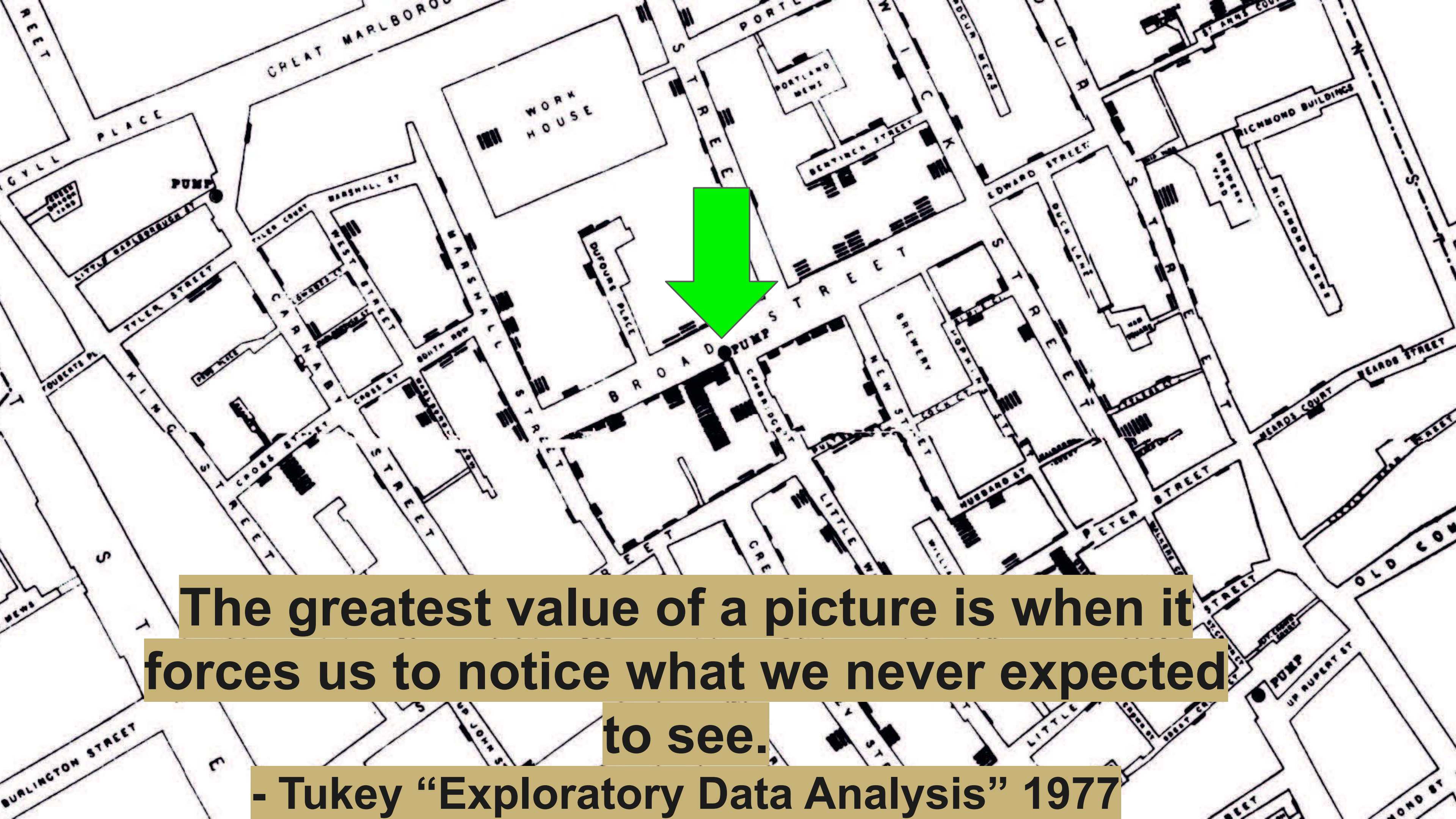
- H_A
ch
wa

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



CREAT MARLBOROU
PORTL
PLACE
RGYLL
PUMP
LITTLE MARLBOROUGH ST
TYLER STREET
TYLER COURT
MARSHALL ST
WORK HOUSE
PORTLAND MEWS
BERWICK STREET
S T R E E T
RICHMOND BUILDINGS
RICHMOND MEWS
BREWERY YARD
EDWARD STREET
DUCK LANE
HAM YARD
DUFOURS PLACE
B R O A D S T R E E T
WEST STREET
CARNABY
MARSHALL STREET
KING STREET
CROSS STREET
CROSS ST
SOUTH ROW
FOUBERTS PL
BREWERY
NEW STREET
CAMBRIDGE ST
HOPKINS ST
HUSBAND ST
PETER STREET
HEARDS COURT
BEARDS STREET
OLD COM
S I L V E R S T R E E T
UP JAMES ST
UP JOHN ST
GREAT PULTENEY STREET
LITTLE WINDMILL STREET
BRIDLE STREET
WILLIAM AND MARY YARD
LITTLE PULTENEY STREET
GREAT CROWN CT
PUMP
UP RUPERT ST
CHAPEL PLACE
NEWS
BURLINGTON STREET
S T R E
LDEN



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

- Tukey "Exploratory Data Analysis" 1977

Fun fact! When the word “statistics” first came about in the 18th century, it meant the "science dealing with data about the condition of a state or community"

Less fun fact! Decades after the cause and prevention of Cholera were known, states knowingly sacrificed thousands of people’s lives for the sake of protecting businesses

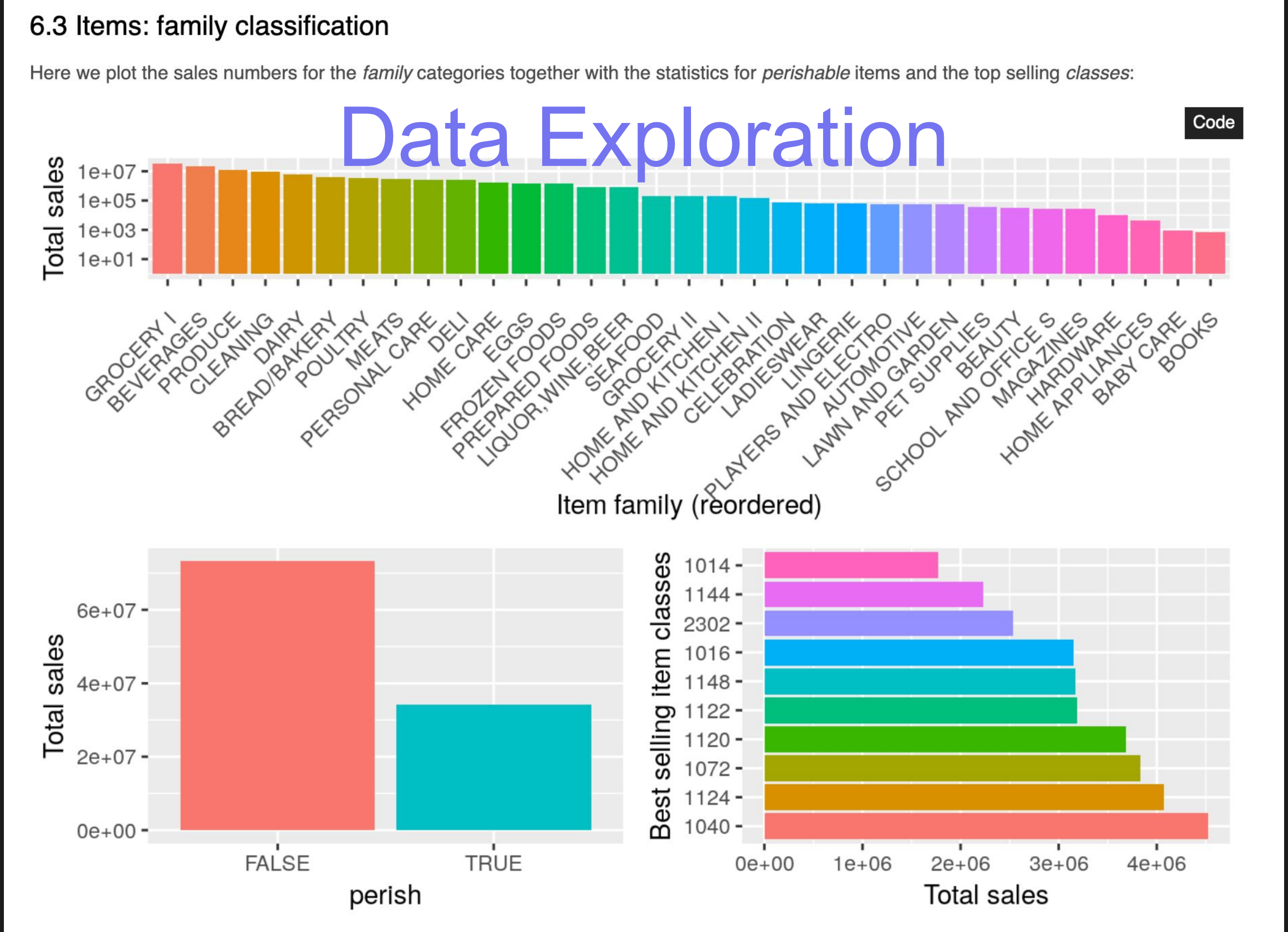
(for instance, Hamburg in 1892)

So what are the lessons?

- Data is good. More data is better
- Visualize your data!
- Do we really need machine learning for this?
- (Maybe states don’t have the people’s best interests at heart)

Data Exploration: Refining your mental model

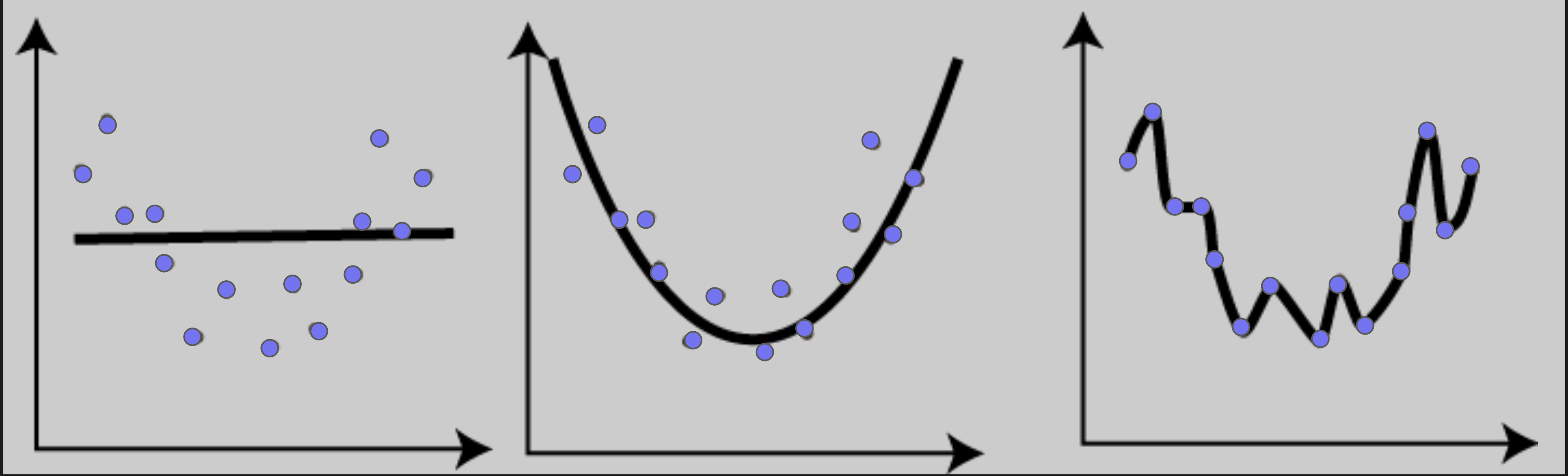
Data Exploration



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*:

Let’s talk about models



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

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

- 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

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

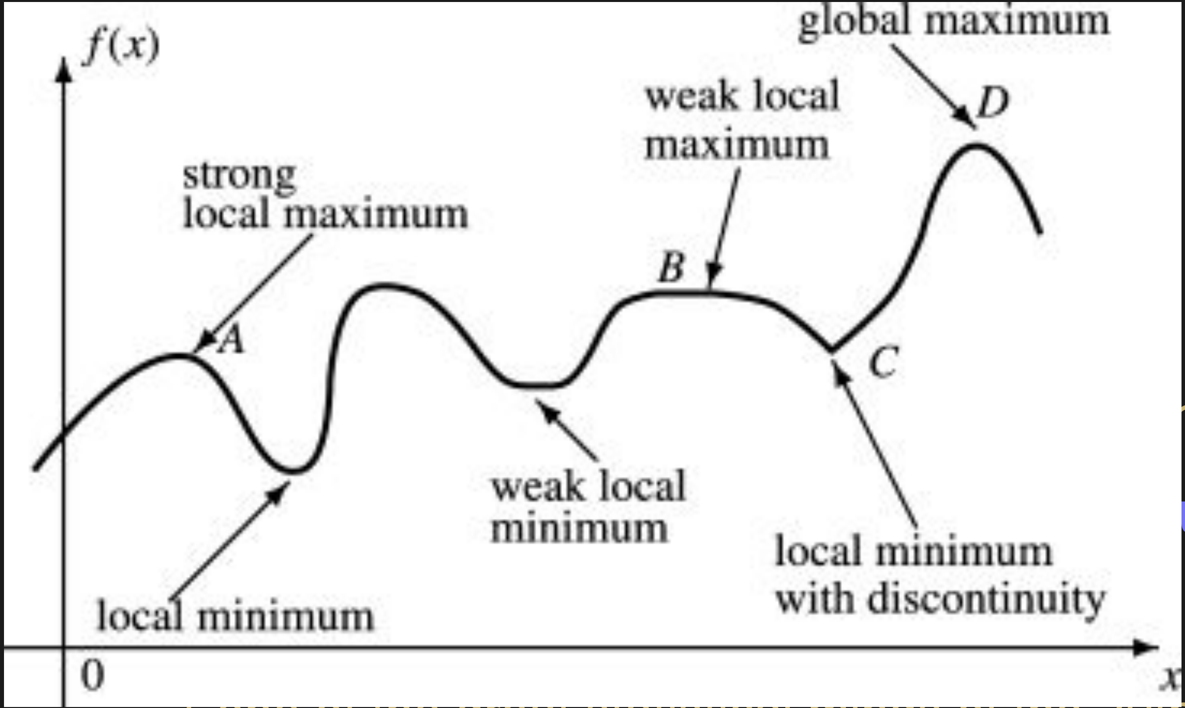
The hard part of making a model useful is not choosing a model, or even hyper-parameterization

Like in Hamburg, it’s overcoming bureaucracy and politics to use the model, even if it’s imperfect, to help your users

Thinking about the business
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.”

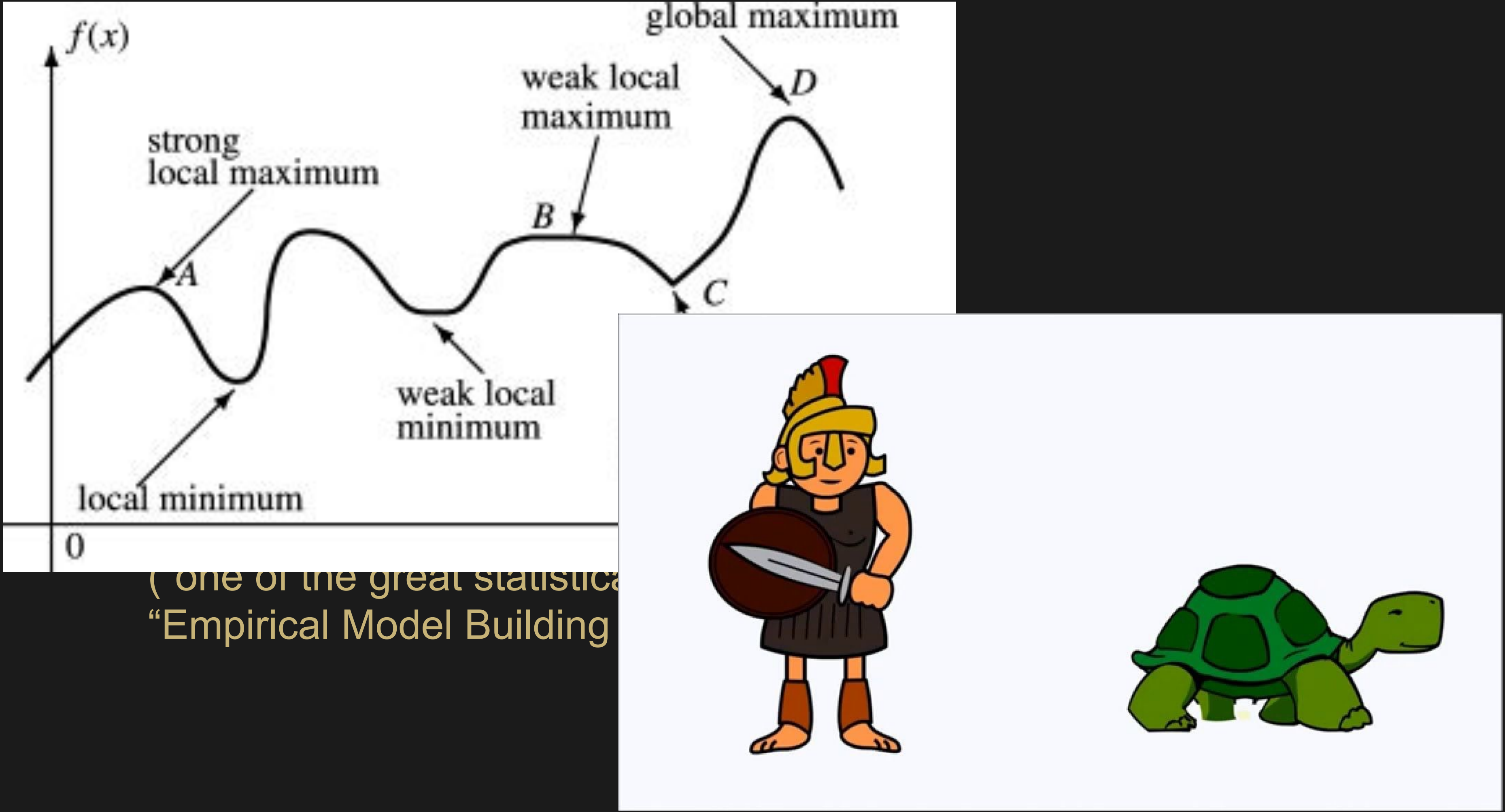
- George Box
(“one of the great statistical minds of the 20th century”)
“Empirical Model Building and Response Surfaces”, 1987



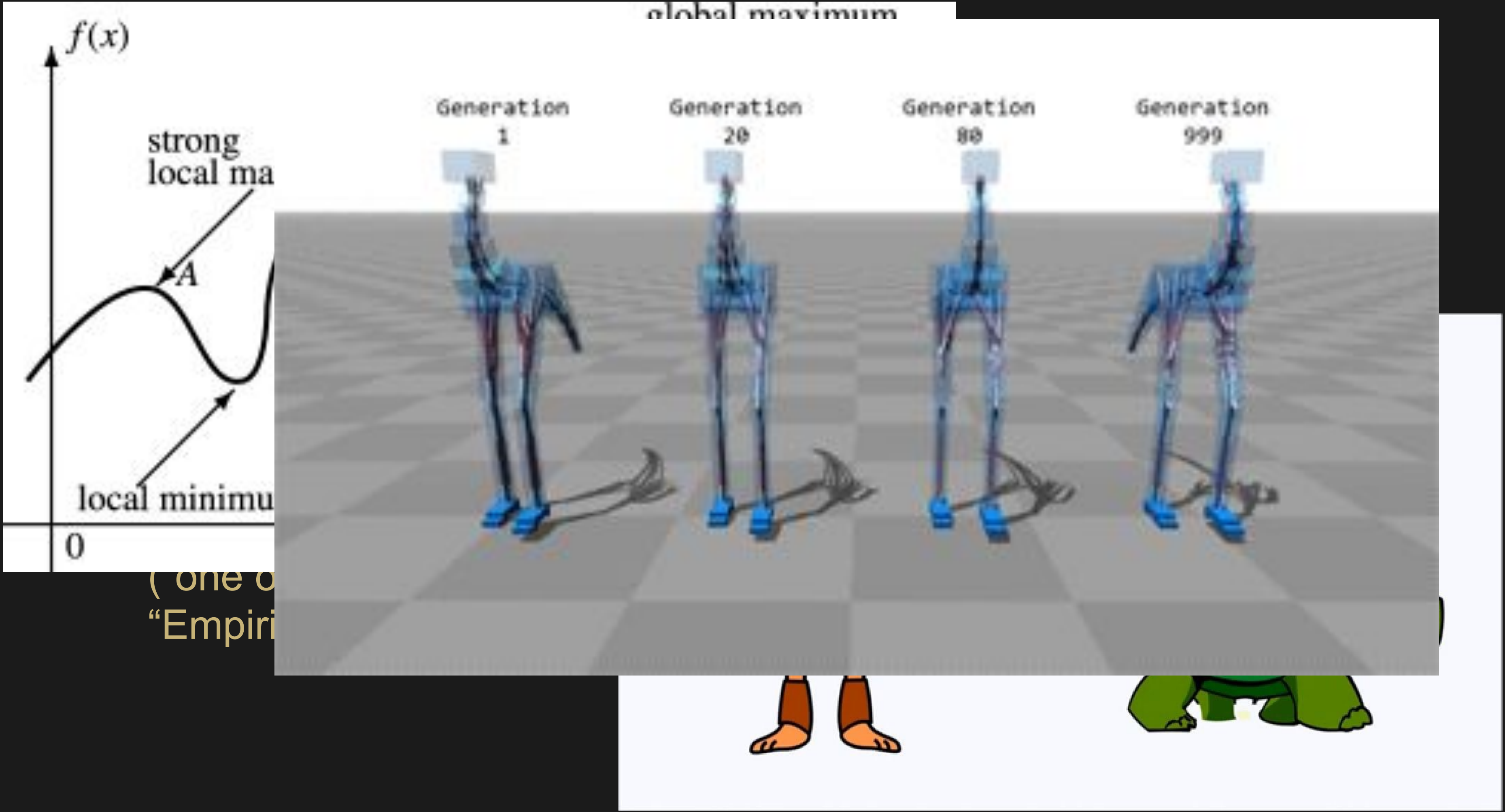
ne practical question is

useful.”

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

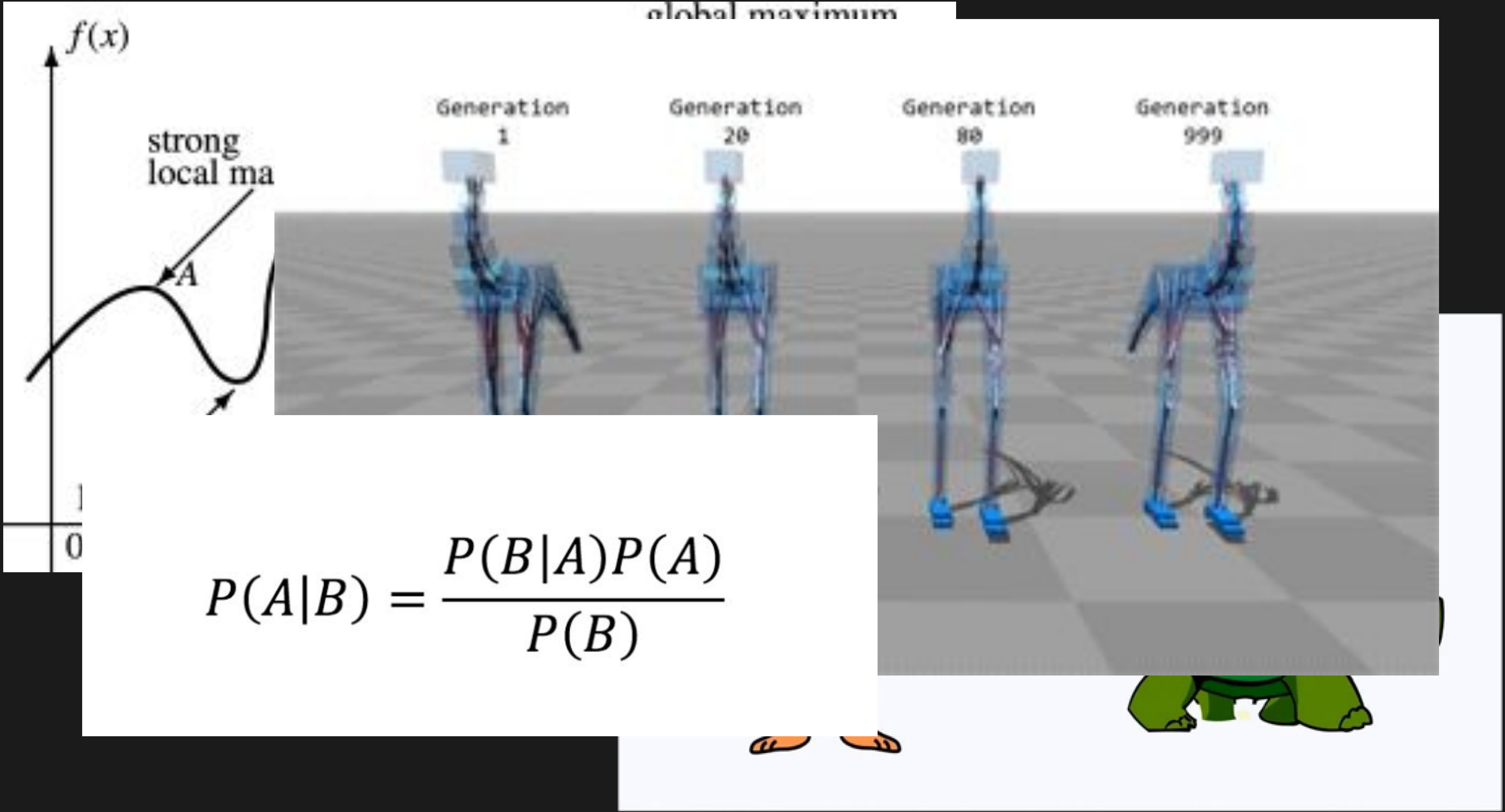


(“one of the great statistica
“Empirical Model Building


global maximum
f(x)
strong
local ma
A
local minimu
0
Generation 1
Generation 20
Generation 80
Generation 999

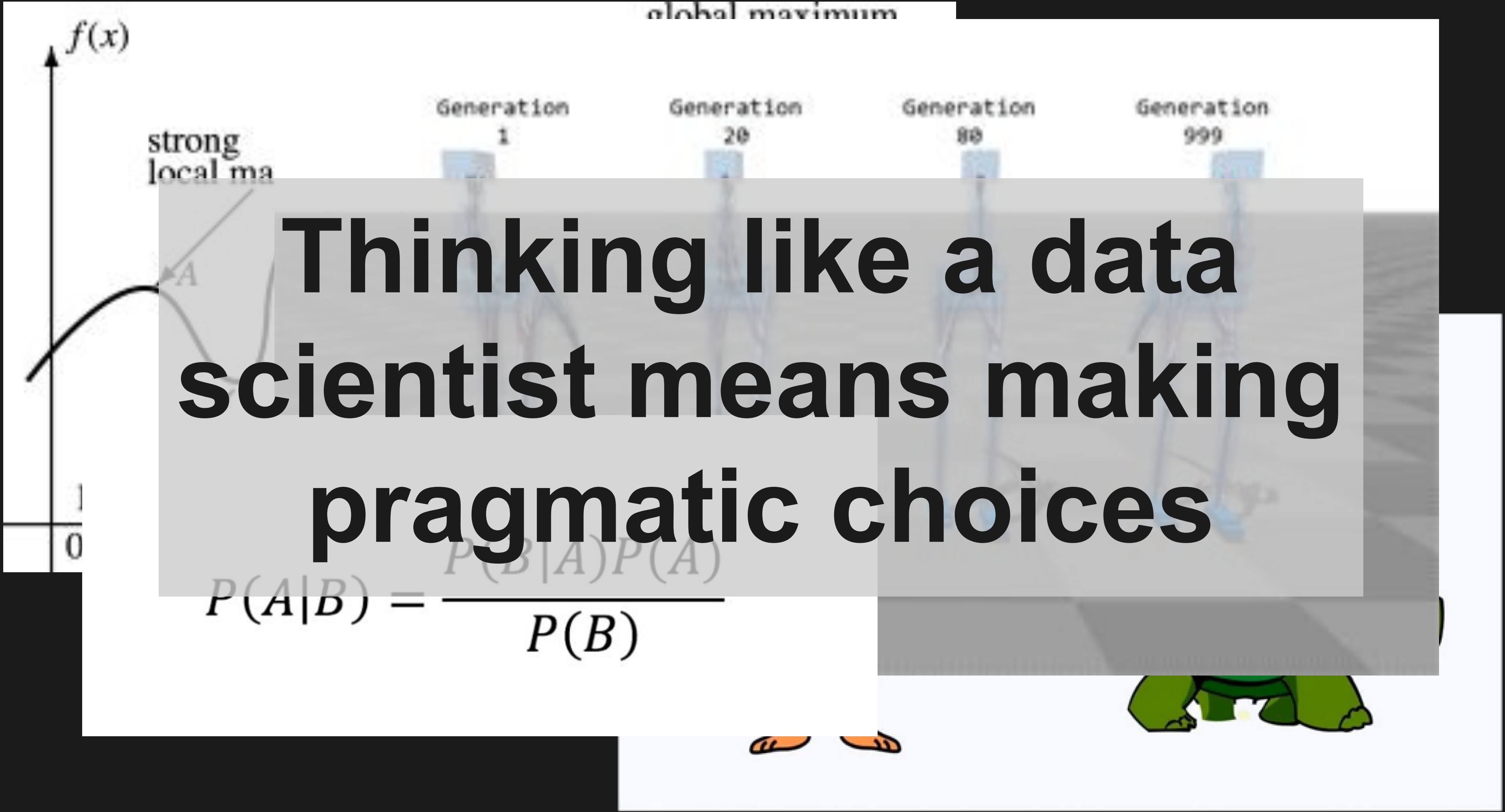
(one o

“Empiri

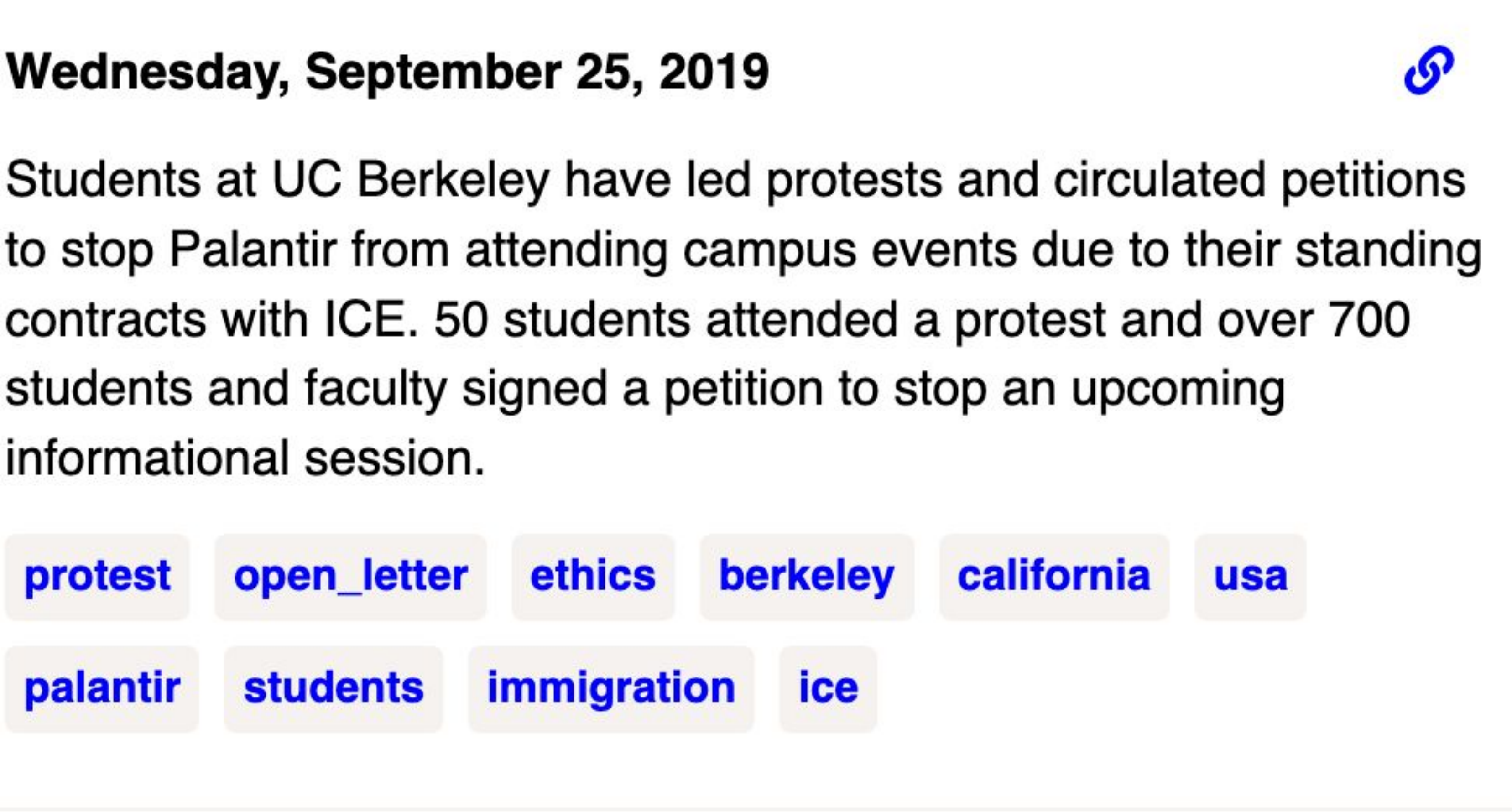


$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Thinking like a data scientist means making pragmatic choices



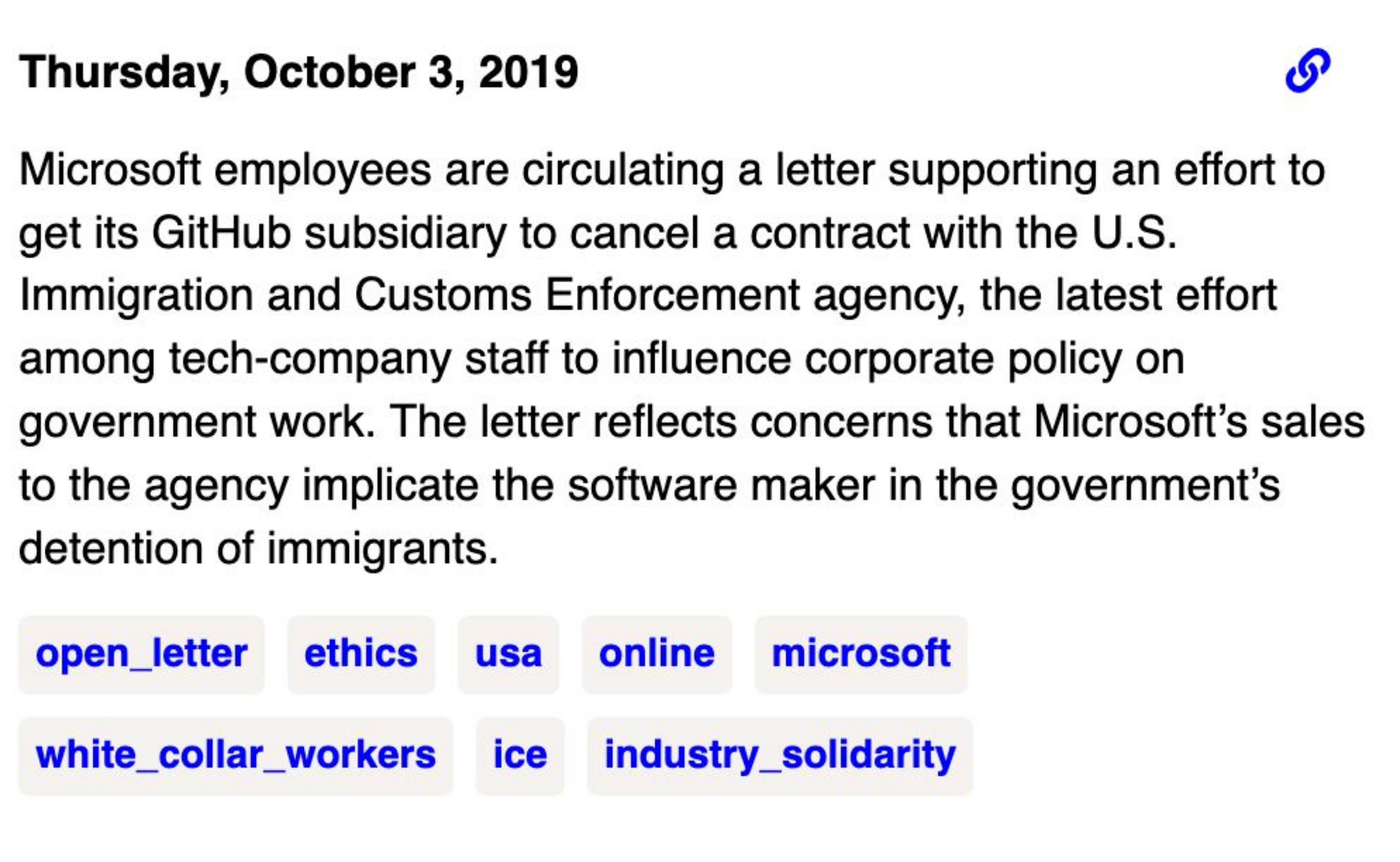
THERE'S NO ETHICAL SOFTWARE DEVELOPMENT UNDER CAPITALISM



Wednesday, September 25, 2019

Students at UC Berkeley have led protests and circulated petitions to stop Palantir from attending campus events due to their standing contracts with ICE. 50 students attended a protest and over 700 students and faculty signed a petition to stop an upcoming informational session.

https://collectiveactions.tech/



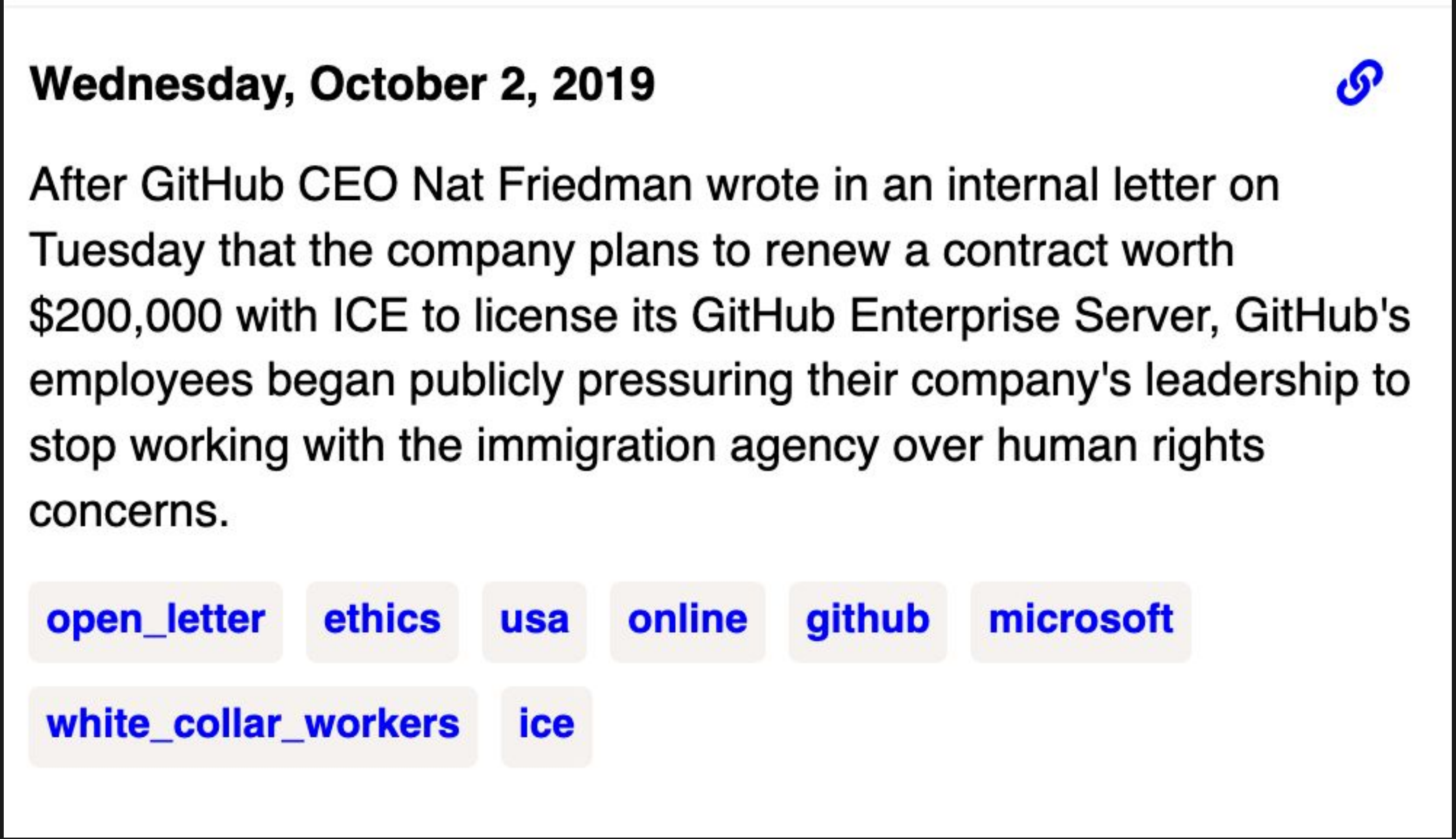
Thursday, October 3, 2019

Microsoft employees are circulating a letter supporting an effort to get its GitHub subsidiary to cancel a contract with the U.S. Immigration and Customs Enforcement agency, the latest effort among tech-company staff to influence corporate policy on government work. The letter reflects concerns that Microsoft's sales to the agency implicate the software maker in the government's detention of immigrants.

open_letter ethics usa online microsoft

white_collar_workers ice industry_solidarity

https://collectiveactions.tech/

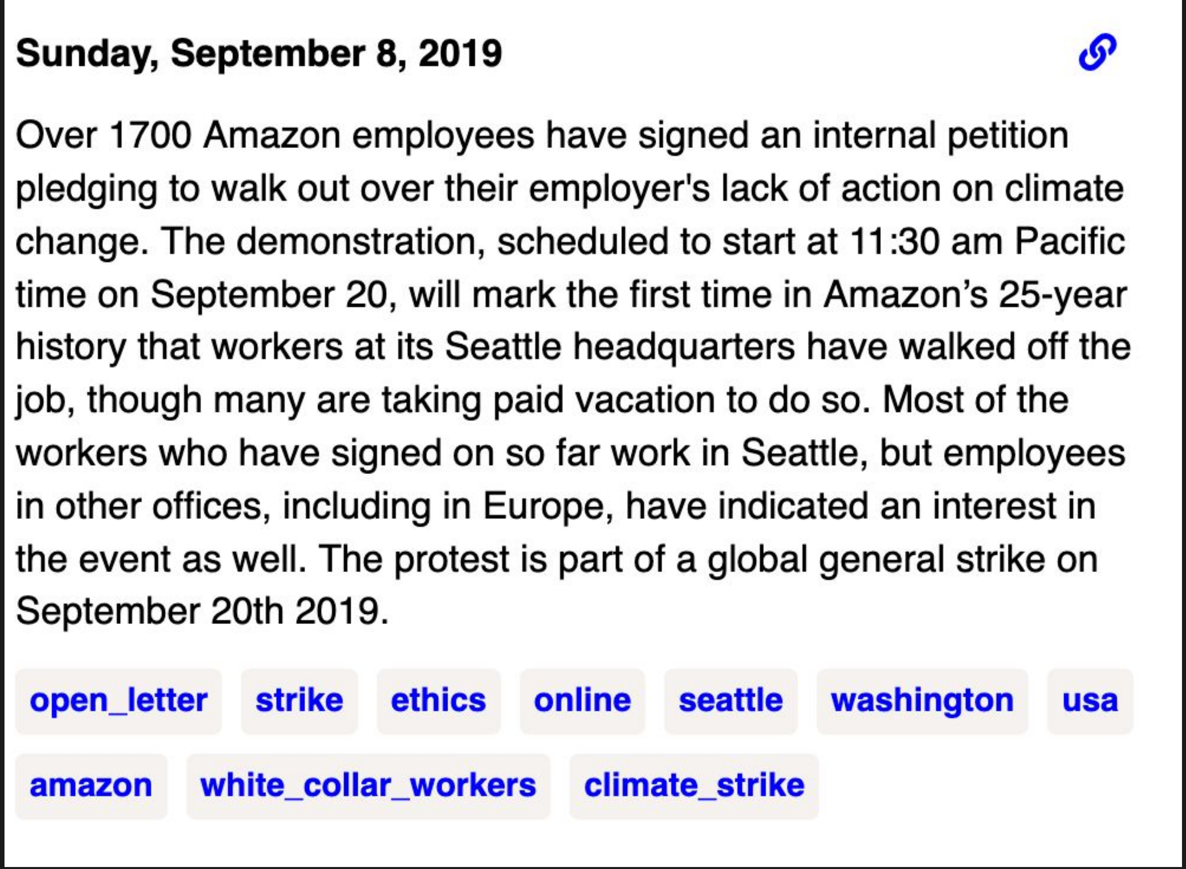


Wednesday, October 2, 2019

After GitHub CEO Nat Friedman wrote in an internal letter on Tuesday that the company plans to renew a contract worth $200,000 with ICE to license its GitHub Enterprise Server, GitHub's employees began publicly pressuring their company's leadership to stop working with the immigration agency over human rights concerns.

open_letter ethics usa online github microsoft white_collar_workers ice

https://collectiveactions.tech/



Sunday, September 8, 2019

Over 1700 Amazon employees have signed an internal petition pledging to walk out over their employer's lack of action on climate change. The demonstration, scheduled to start at 11:30 am Pacific time on September 20, will mark the first time in Amazon's 25-year history that workers at its Seattle headquarters have walked off the job, though many are taking paid vacation to do so. Most of the workers who have signed on so far work in Seattle, but employees in other offices, including in Europe, have indicated an interest in the event as well. The protest is part of a global general strike on September 20th 2019.

open_letter strike ethics online seattle washington usa amazon white_collar_workers climate_strike

https://collectiveactions.tech/

♪REVOLUTION♪

Thank you!

I’m Em Grasmeder

the ThoughtWorks Data Witch

@emgrasmeder on Twitter