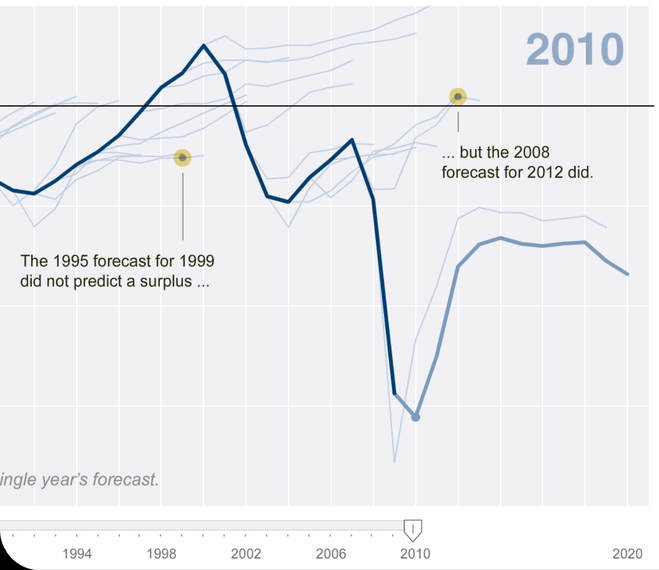
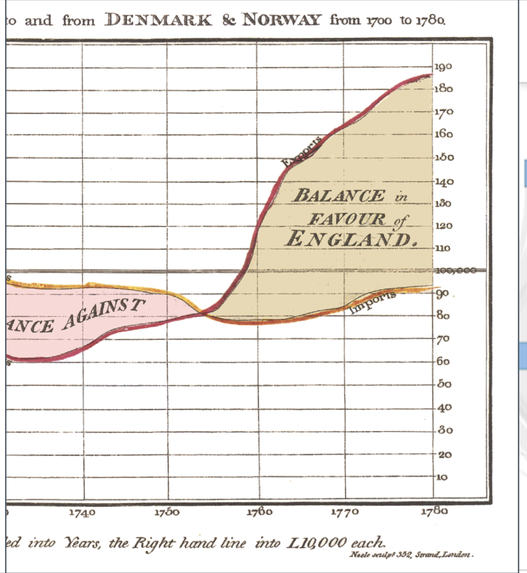
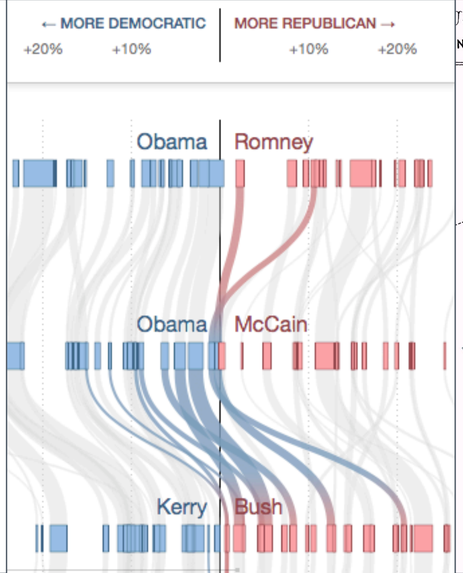
6.859: Interactive Data Visualization

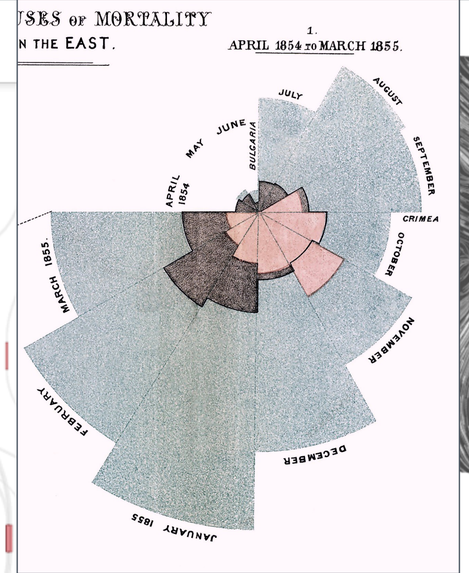
The Value of Visualization

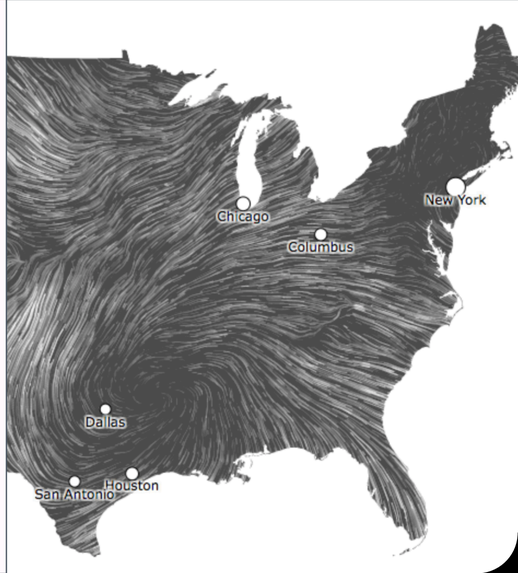
Arvind Satyanarayan





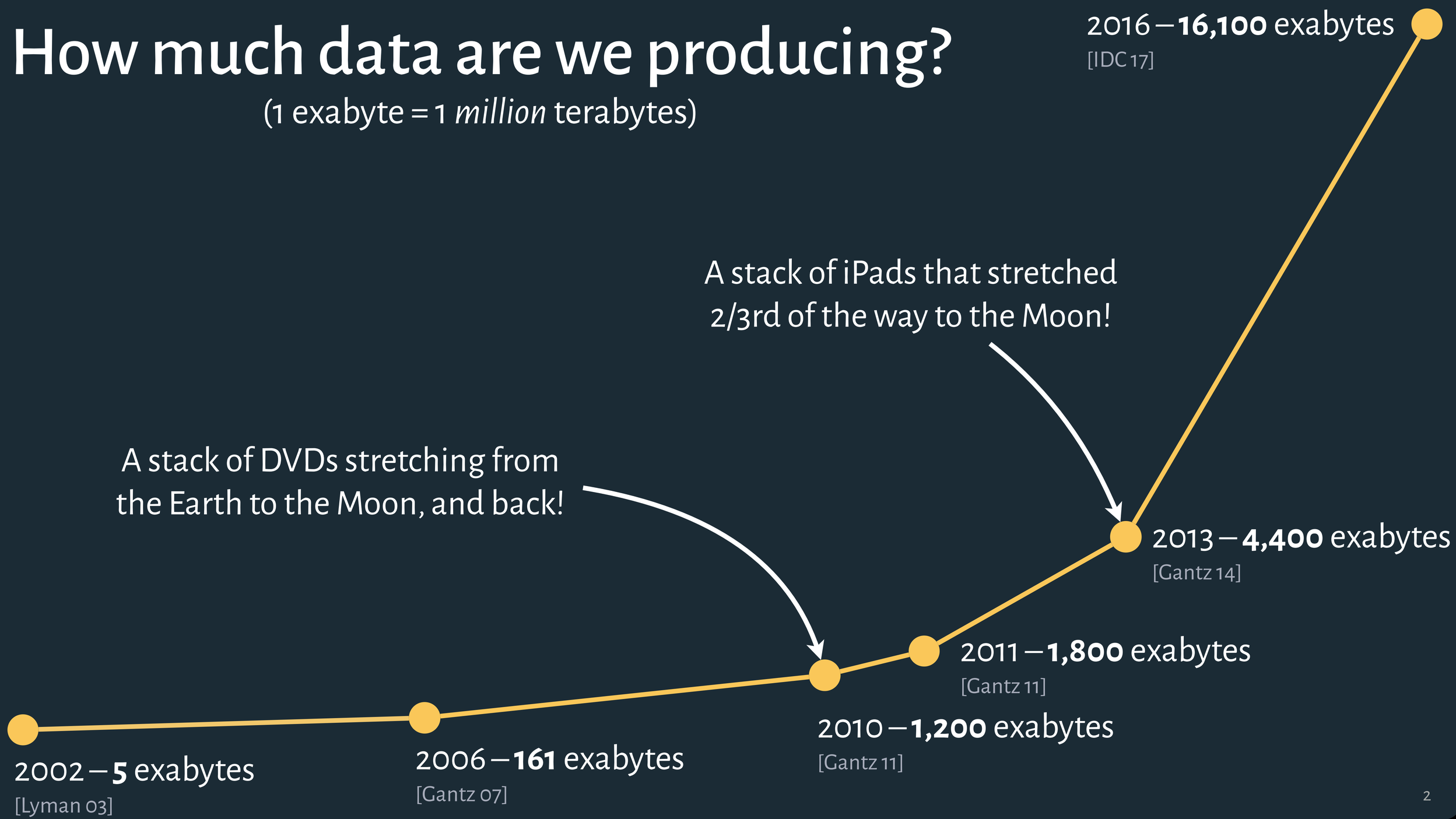






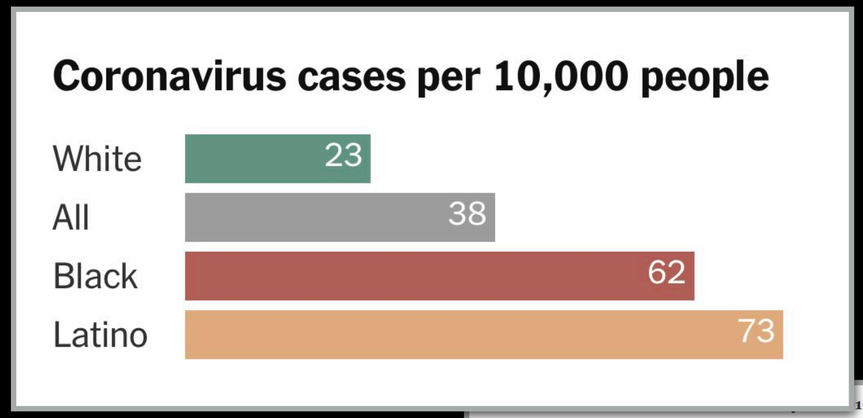
How much data are we producing?

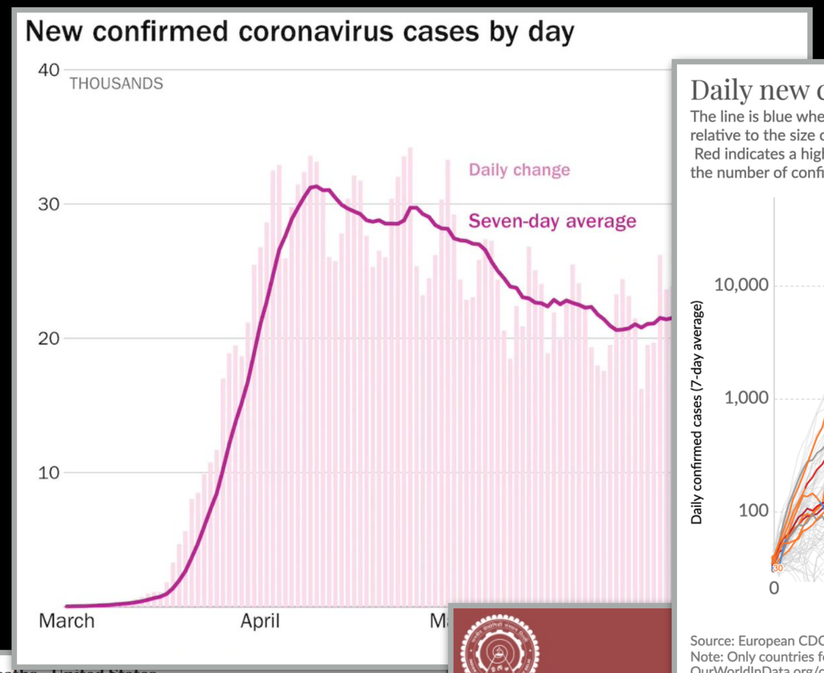
(1 exabyte = 1 *million* terabytes)

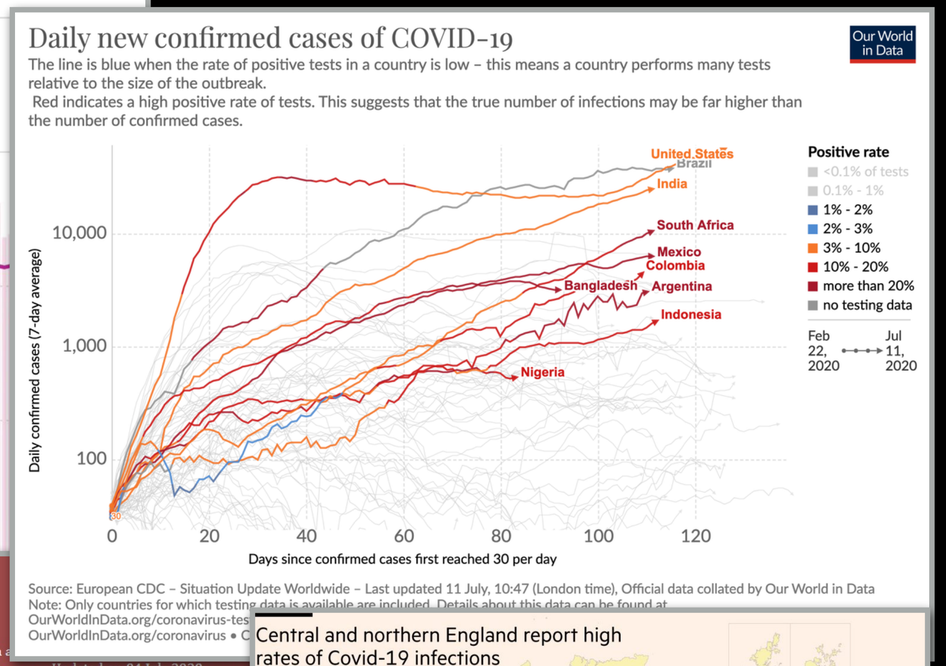


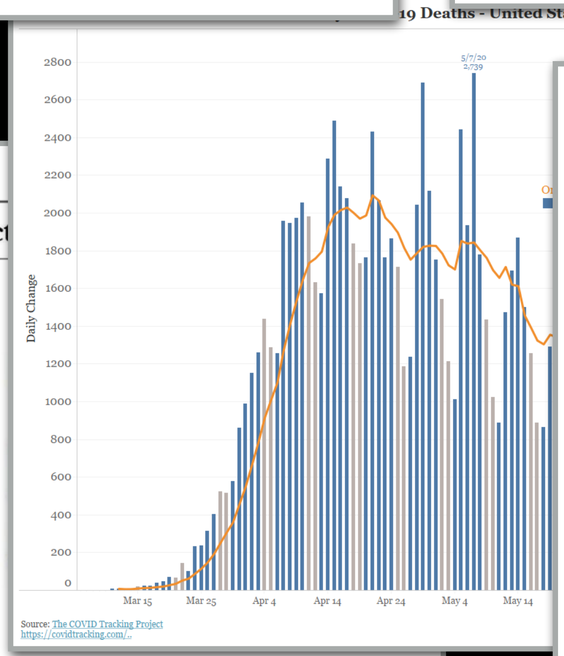
Health & Medicine

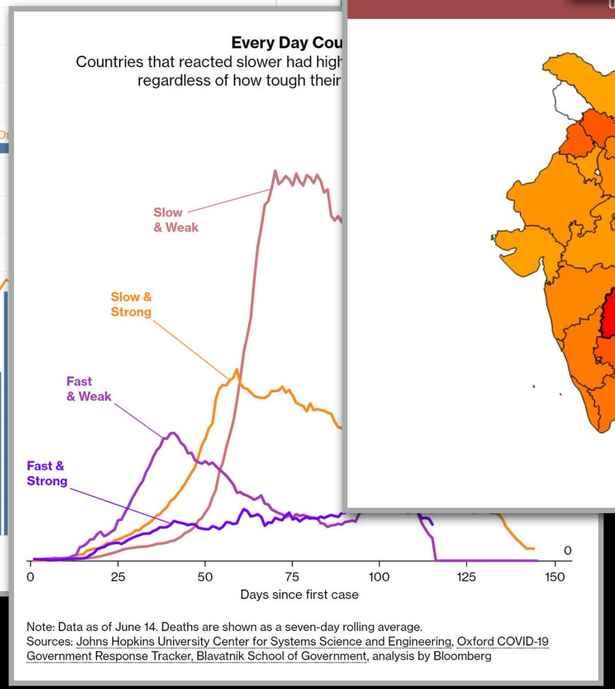
COVID-19

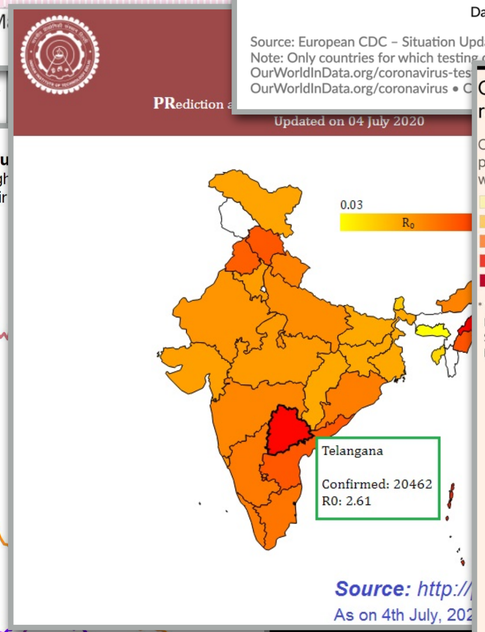

Coronavirus cases per 10,000 people
White
23
All
38
Black
62
Latino
73

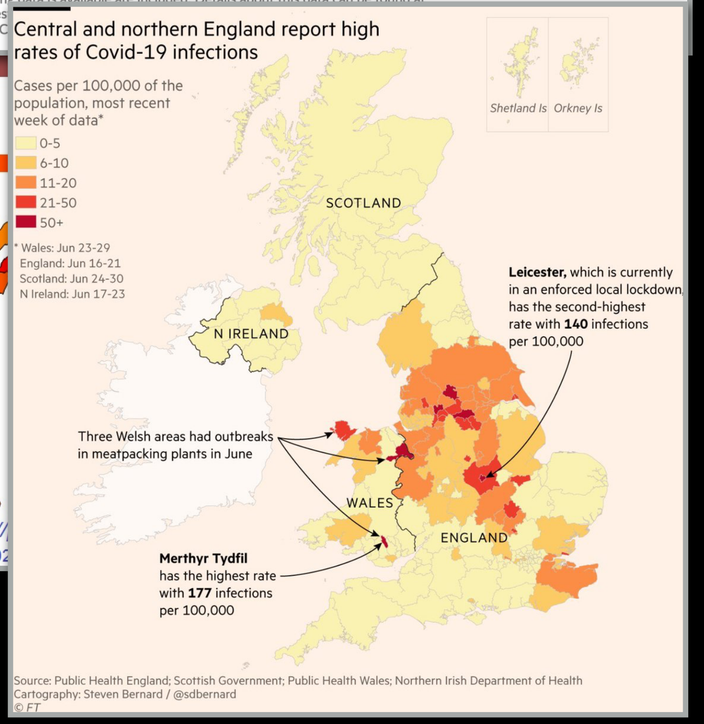

New confirmed coronavirus cases by day
40
THOUSANDS
30
20
10
Daily change
Seven-day average
March
April

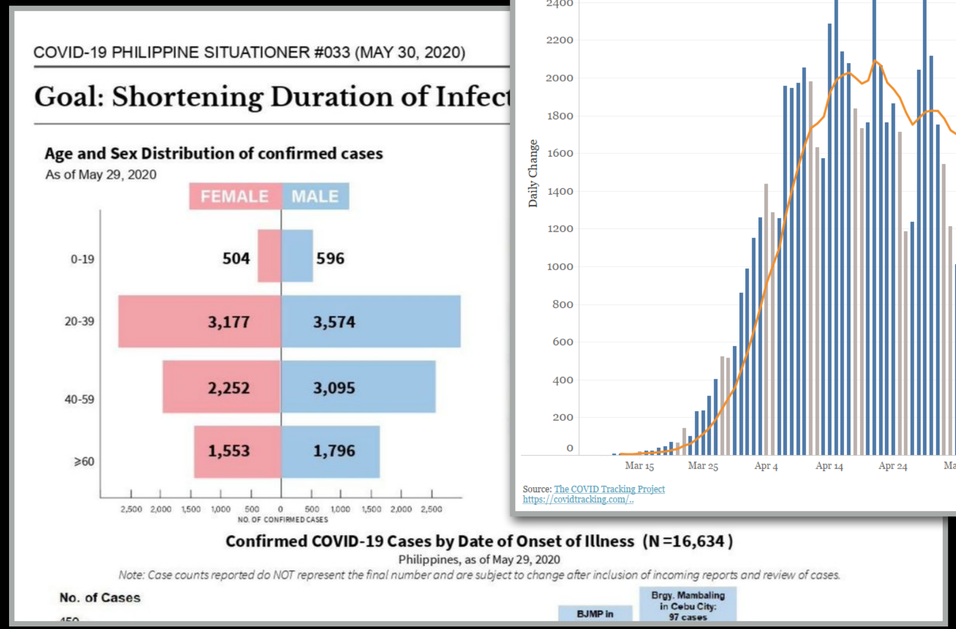

Daily new confirmed cases of COVID-19
The line is blue when the rate of positive tests in a country is low – this means a country performs many tests relative to the size of the outbreak.
Red indicates a high positive rate of tests. This suggests that the true number of infections may be far higher than the number of confirmed cases.
Our World in Data
Daily confirmed cases (7-day average)
10,000
1,000
100
United States
Brazil
India
South Africa
Mexico
Colombia
Bangladesh
Argentina
Indonesia
Nigeria
0
20
40
60
80
100
120
Days since confirmed cases first reached 30 per day
Positive rate
<0.1% of tests
0.1% - 1%
1% - 2%
2% - 3%
3% - 10%
10% - 20%
more than 20%
no testing data
Feb 22, 2020
Jul 11, 2020
Source: European CDC – Situation Update Worldwide – Last updated 11 July, 10:47 (London time), Official data collated by Our World in Data
Note: Only countries for which testing data is available are included. Details about this data can be found at
OurWorldInData.org/coronavirus-tes
OurWorldInData.org/coronavirus • C


19 Deaths - United States
Daily Change
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
5/7/20
2,739
Mar 15
Mar 25
Apr 4
Apr 14
Apr 24
May 4
May 14
Source: The COVID Tracking Project
https://covidtracking.com/...

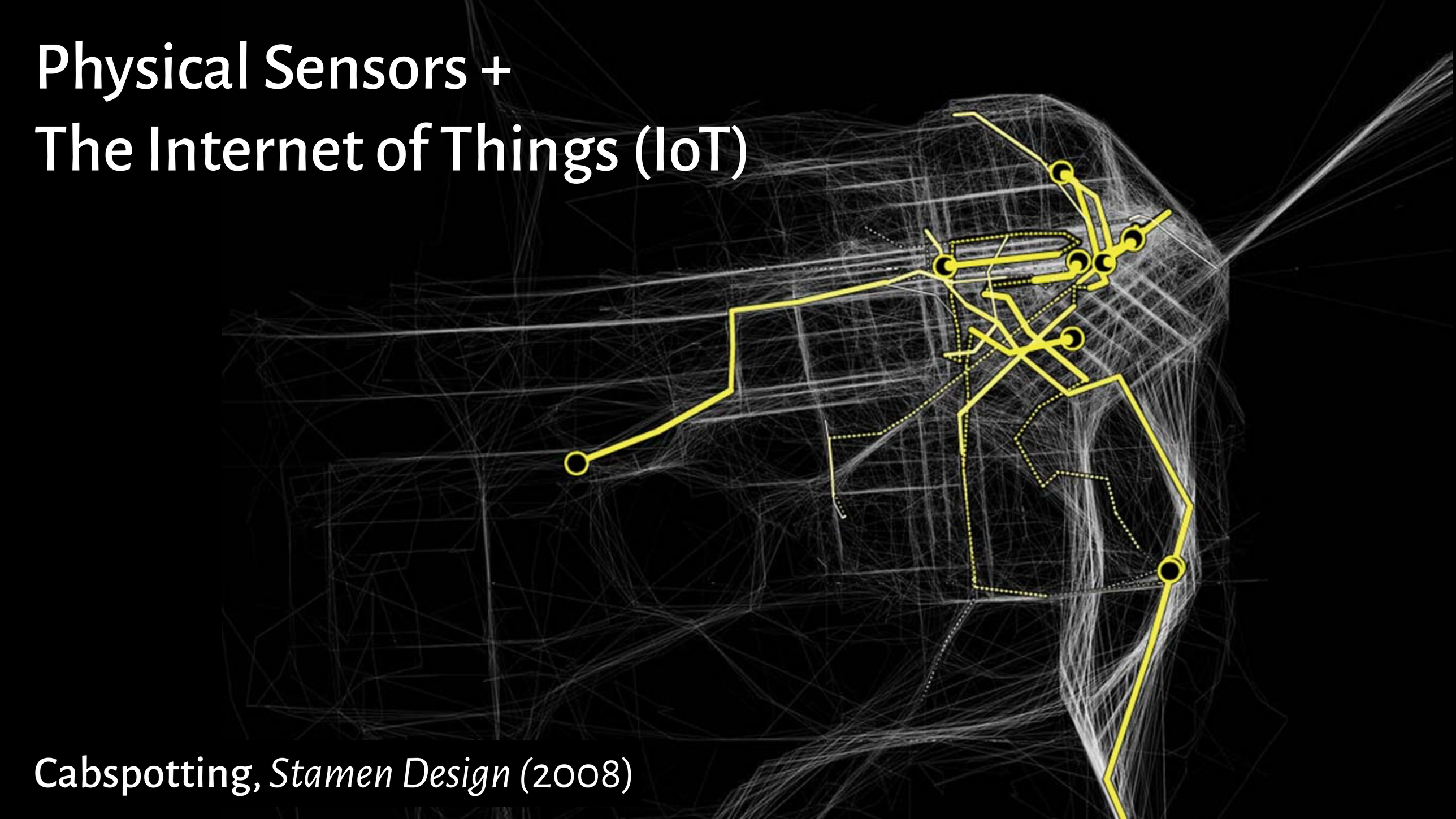

Every Day Cou
Countries that reacted slower had high
regardless of how tough their
Slow & Weak
Slow & Strong
Fast & Weak
Fast & Strong
0
25
50
75
100
125
150
Days since first case
Note: Data as of June 14. Deaths are shown as a seven-day rolling average.
Sources: Johns Hopkins University Center for Systems Science and Engineering, Oxford COVID-19 Government Response Tracker, Blavatnik School of Government, analysis by Bloomberg


PRediction
Updated on 04 July 2020
0.03
R0
Telangana
Confirmed: 20462
R0: 2.61
Source: http://
As on 4th July, 202


Central and northern England report high rates of Covid-19 infections
Cases per 100,000 of the population, most recent week of data*
0-5
6-10
11-20
21-50
50+
* Wales: Jun 23-29
England: Jun 16-21
Scotland: Jun 24-30
N Ireland: Jun 17-23
Shetland Is
Orkney Is
SCOTLAND
N IRELAND
WALES
ENGLAND
Leicester, which is currently in an enforced local lockdown, has the second-highest rate with 140 infections per 100,000
Three Welsh areas had outbreaks in meatpacking plants in June
Merthyr Tydfil has the highest rate with 177 infections per 100,000
Source: Public Health England; Scottish Government; Public Health Wales; Northern Irish Department of Health
Cartography: Steven Bernard / @sdbernard
© FT


COVID-19 PHILIPPINE SITUATIONER #033 (MAY 30, 2020)
Goal: Shortening Duration of Infect
Age and Sex Distribution of confirmed cases
As of May 29, 2020
FEMALE
MALE
0-19
504
596
20-39
3,177
3,574
40-59
2,252
3,095
≥60
1,553
1,796
2,500 2,000 1,500 1,000 500 0 500 1,000 1,500 2,000 2,500
NO. OF CONFIRMED CASES
Confirmed COVID-19 Cases by Date of Onset of Illness (N=16,634)
Philippines, as of May 29, 2020
Note: Case counts reported do NOT represent the final number and are subject to change after inclusion of incoming reports and review of cases.
No. of Cases
Brgy. Mambaling in Cebu City: 97 cases
BJMP in

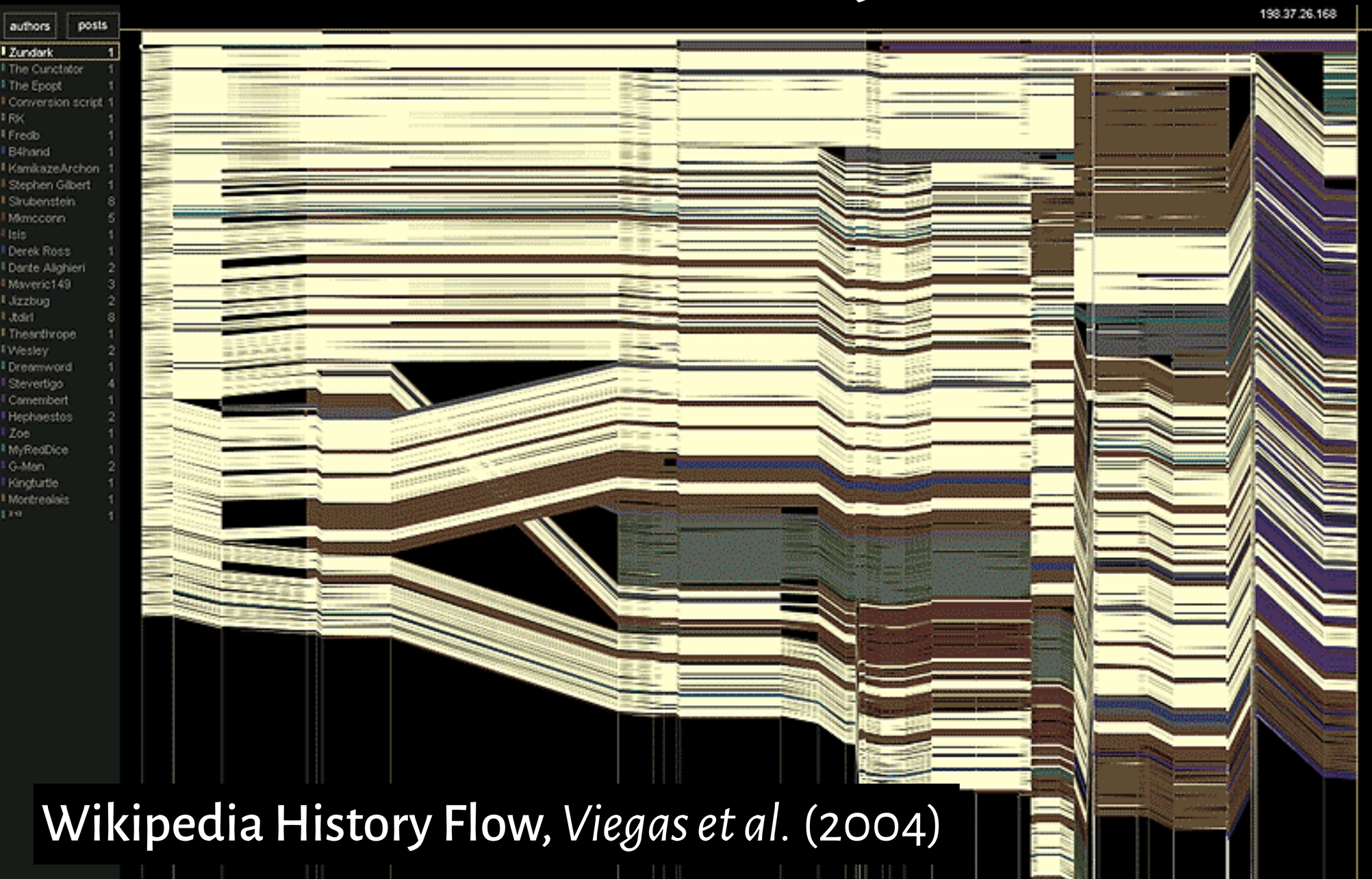
Physical Sensors + The Internet of Things (IoT)



Cabspotting, *Stamen Design* (2008)

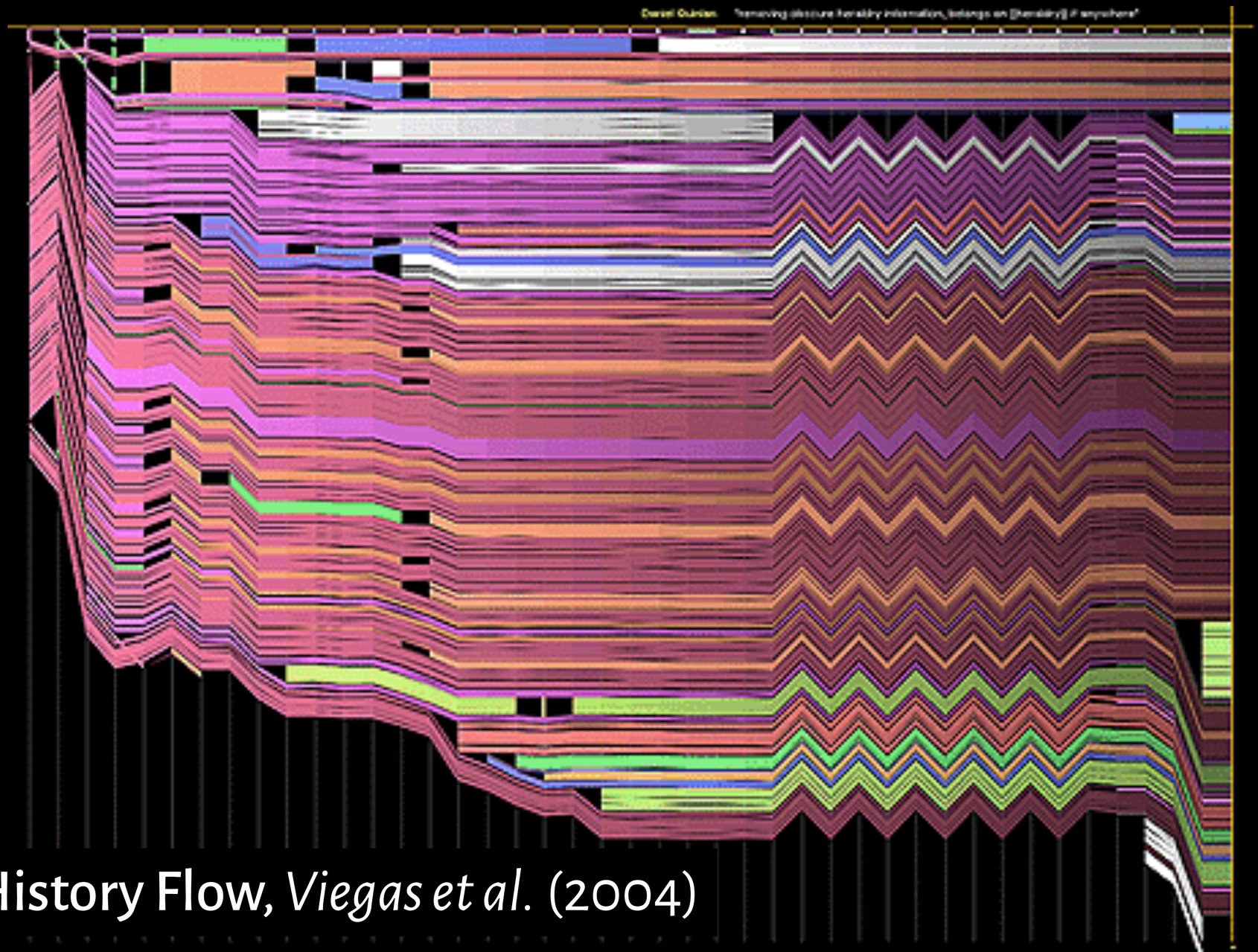
Records of Human Activity

Records of Human Activity



Wikipedia History Flow, *Viegas et al.* (2004)

Records of Human Activity



Wikipedia History Flow, *Viegas et al.* (2004)

"The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, [...] because now we really do have **essentially free and ubiquitous data**. So the complimentary scarce factor is the ability to understand that data and extract value from it."

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

"The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, [...] because now we really do have **essentially free and ubiquitous data**. So the scarce factor is the ability to understand that data and extract value from it."

Machine Learning!

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

"The ability to take data—to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it—that's going to be a hugely important skill in the next decades, [...] because now we really do have **essentially free and ubiquitous data**. So the complimentary scarce factor is the ability to understand that data and extract value from it."

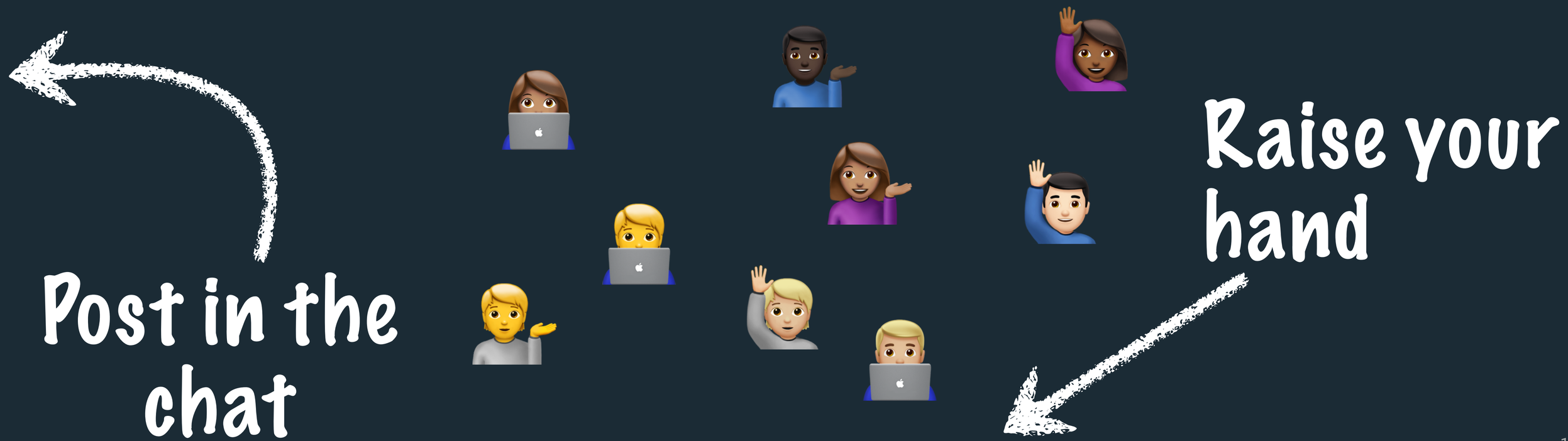
Machine Learning?

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

Machine Learning?

Imagine a system to analyze large amounts of data.

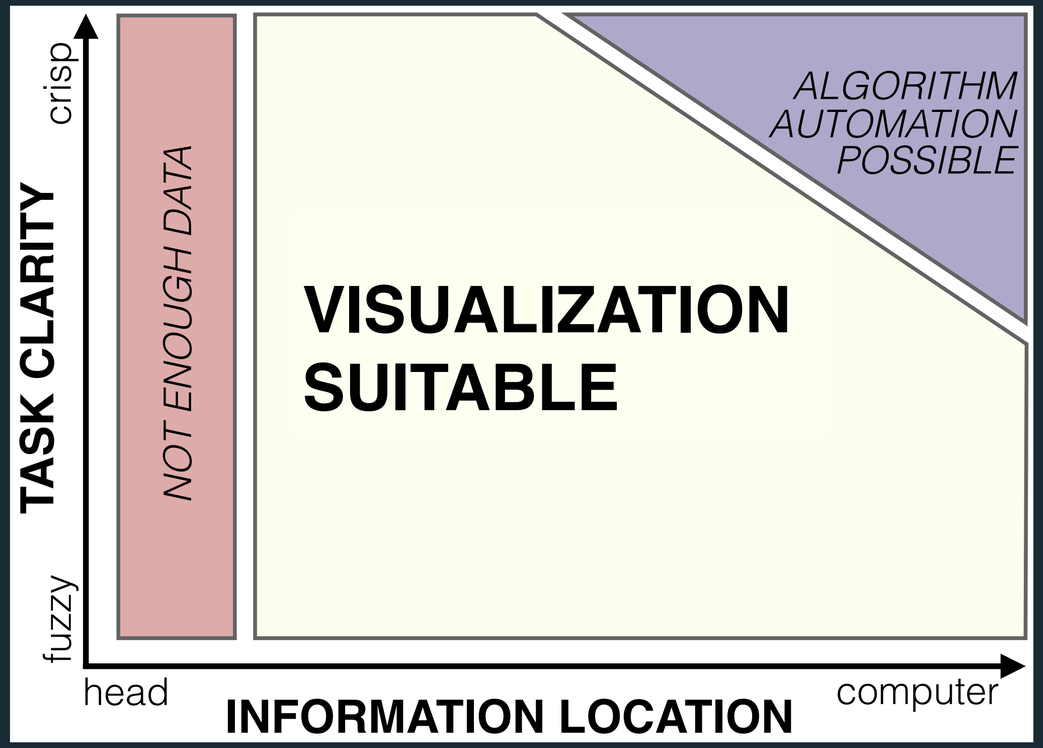
1. Why would you have a human in-the-loop?
2. Why would you have a computer in-the-loop?



Machine Learning?

Imagine a system to analyze large amounts of data.

1. Why would you have a human in-the-loop?
2. Why would you have a computer in-the-loop?



[Sedlmair, Meyer, and Munzner 2012]

Set A		Set B		Set C		Set D	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.11	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Summary Statistics

$u_X = 9.0$ $\sigma_X = 3.317$

$u_Y = 7.5$ $\sigma_Y = 2.03$

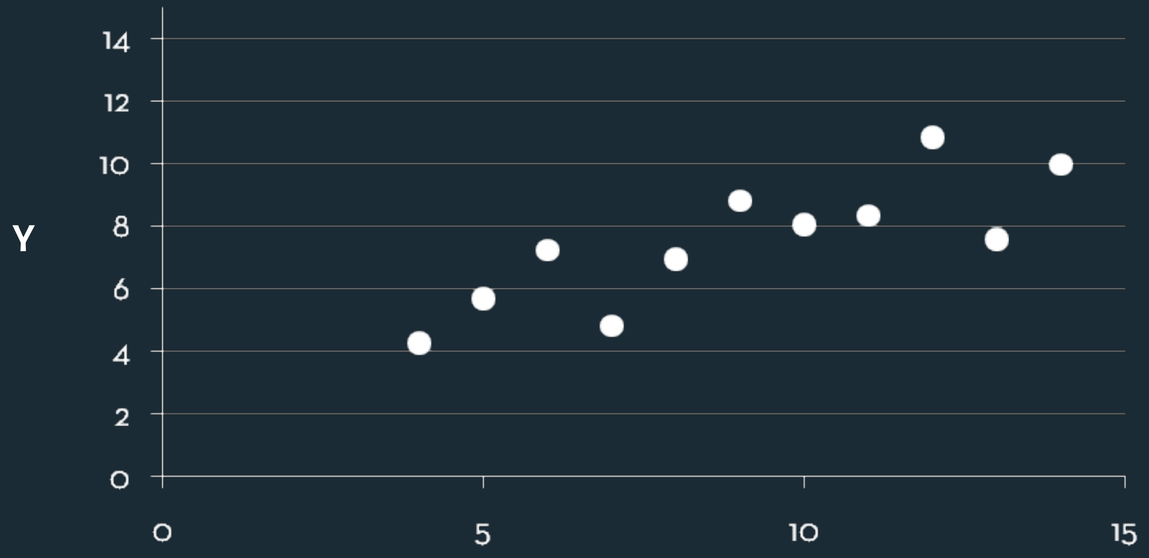
Linear Regression

$Y = 3 + 0.5\,X$

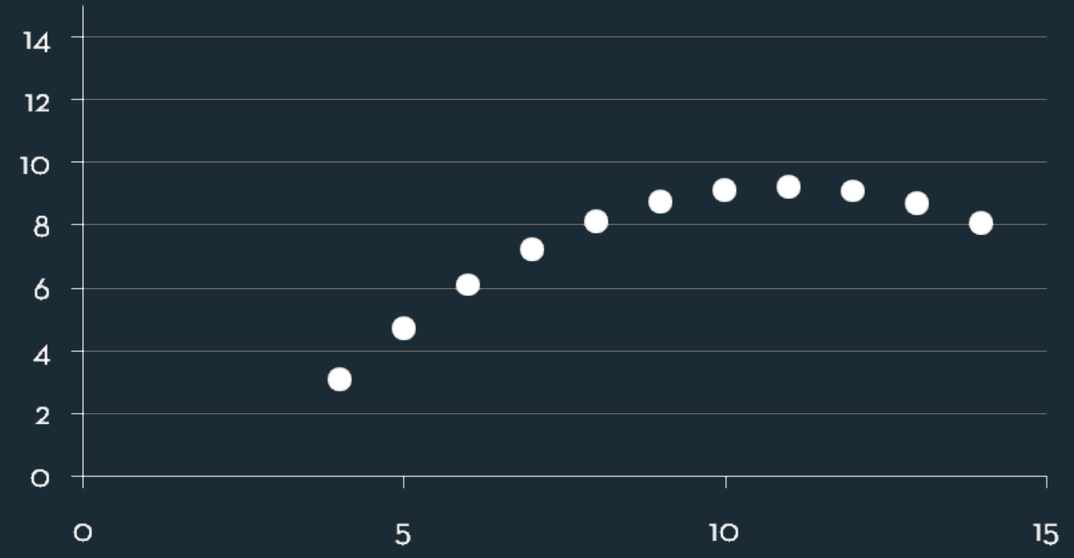
$R^2 = 0.67$

[Anscombe 1973]

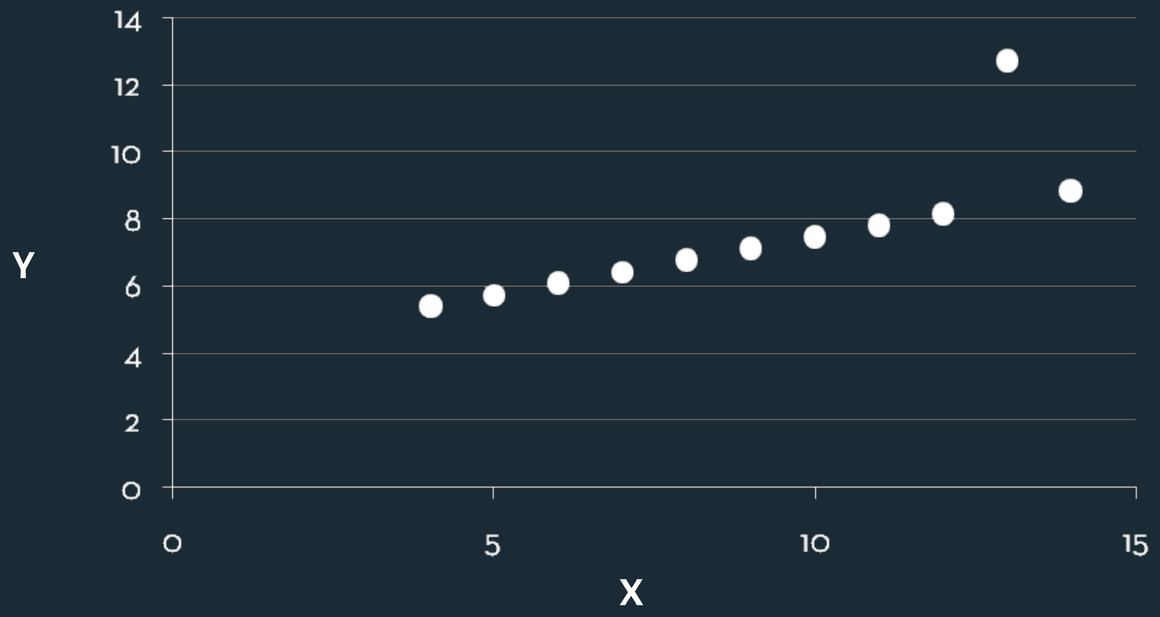
Set A



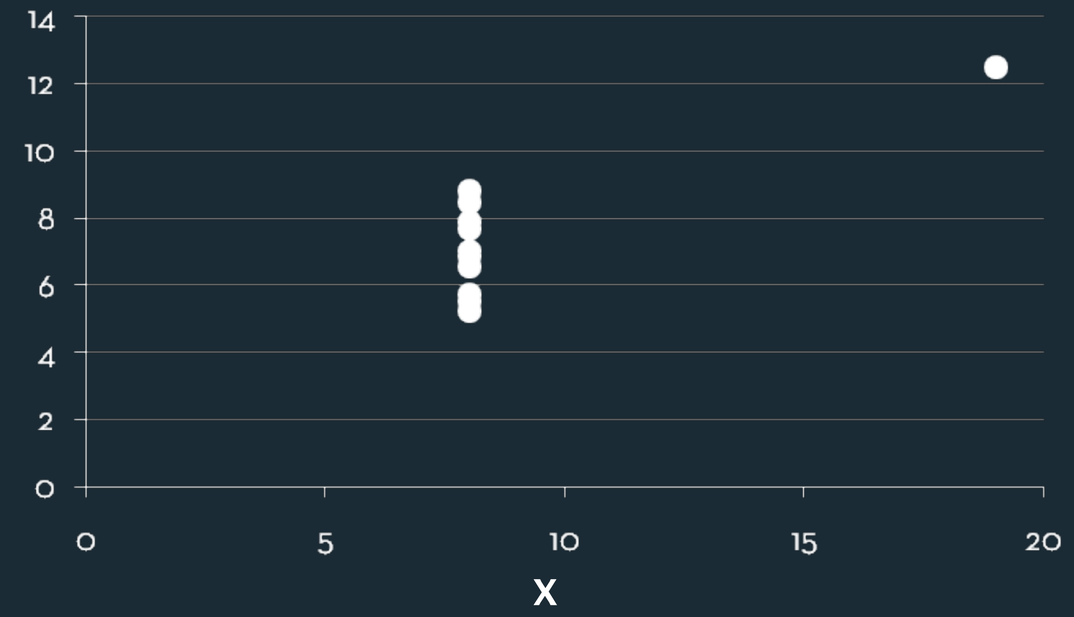
Set B



Set C



Set D

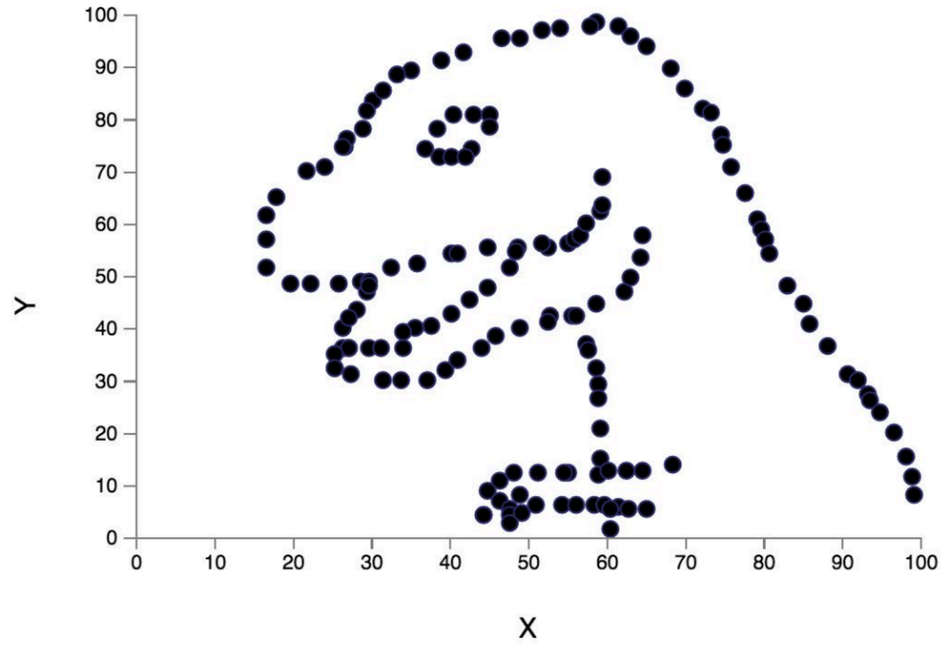


Alberto Cairo
@albertocairo

Following

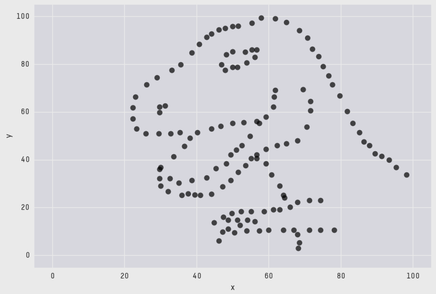
Don't trust summary statistics. Always visualize your data first
robertgrantstats.co.uk/drawmydata.html

N = 157 ; X mean = 50.7333 ; X SD = 19.5661 ; Y mean = 46.495 ; Y SD = 27.2828 ;
Pearson correlation = -0.1772

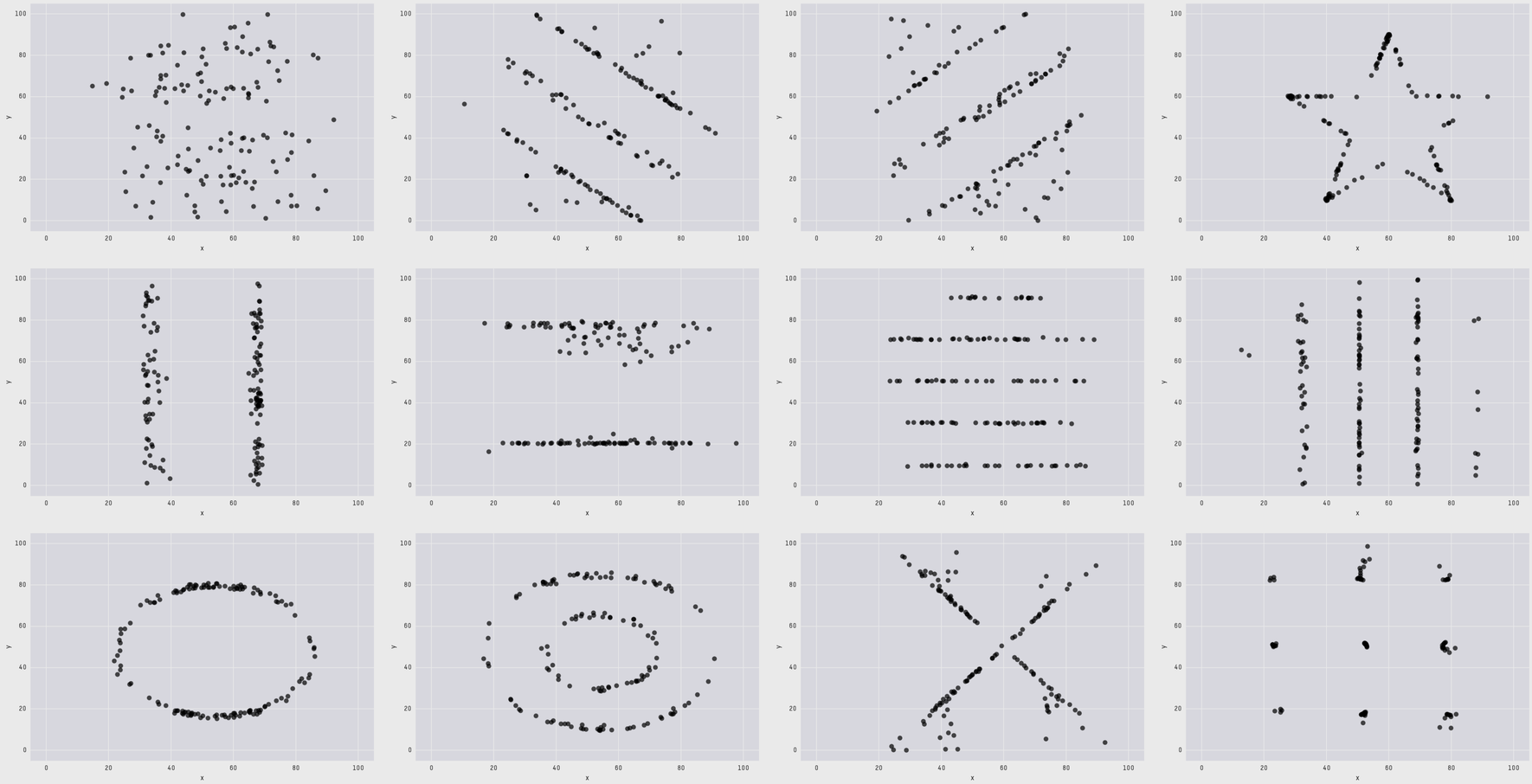


5:47 AM - 15 Aug 2016

952 Retweets **1,023** Likes



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06



The Datasaurus Dozen, *Matejka & Fitzmaurice* (2008)

The Value of Visualization

Aka why create visualizations?

Record information

- Blueprints, photographs, seismographs, ...

Analyze data to support reasoning (exploratory visualization)

- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

Communicate information to others (explanatory visualization)

- Share and persuade
- Collaborate and revise

The Value of Visualization

Aka why create visualizations?

Record information

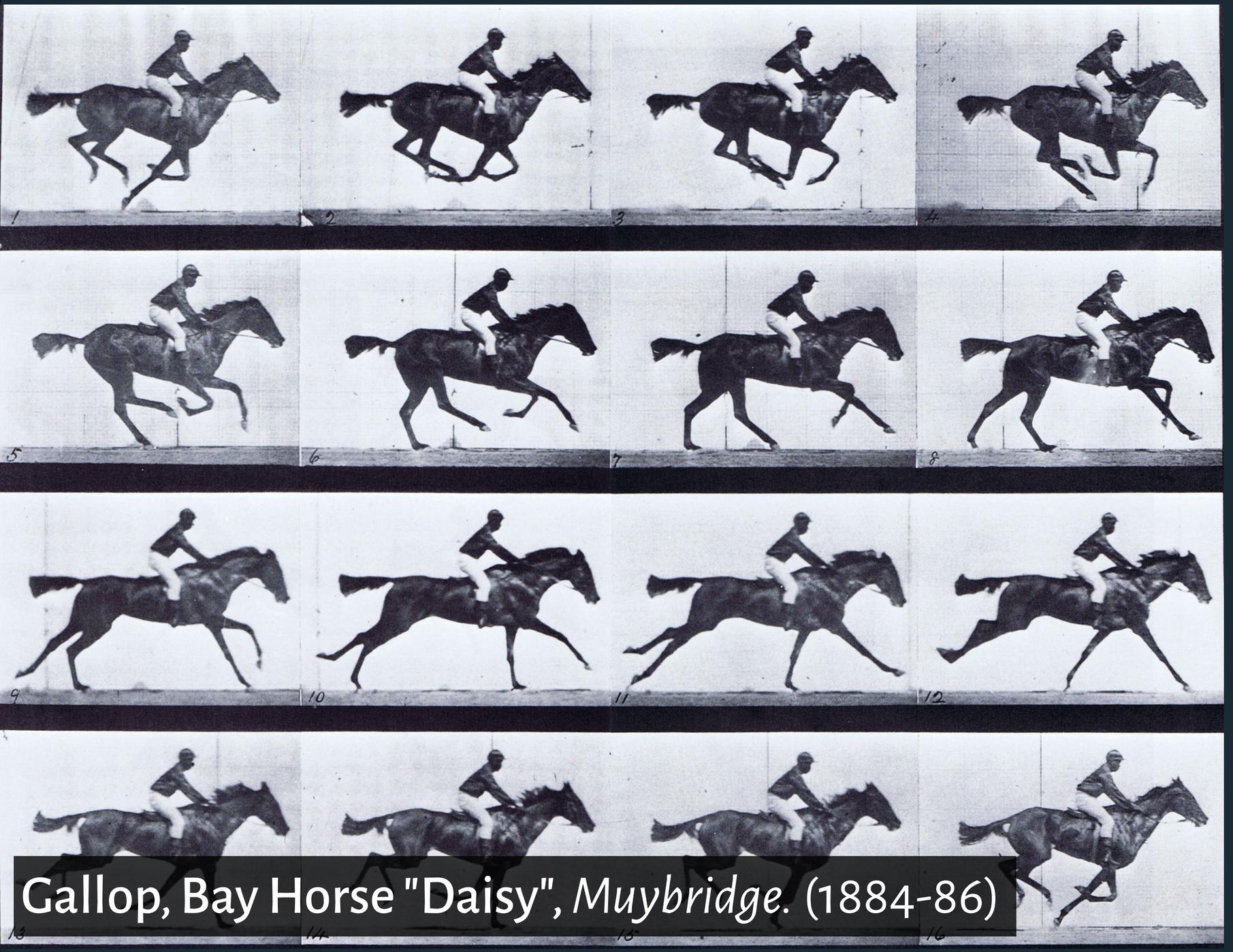
- Blueprints, photographs, seismographs, ...

Analyze data to support reasoning (exploratory visualization)

- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

Communicate information to others (explanatory visualization)

- Share and persuade
- Collaborate and revise



Gallop, Bay Horse "Daisy", *Muybridge.* (1884-86)

Record Info

To answer a question:

Do all 4 hooves leave the ground when a horse gallops?

Record Info

E.J. Marey's **sphygmograph** (1863)
First external, non-intrusive way to measure blood pressure.

Directly recorded pulse as a waveform.

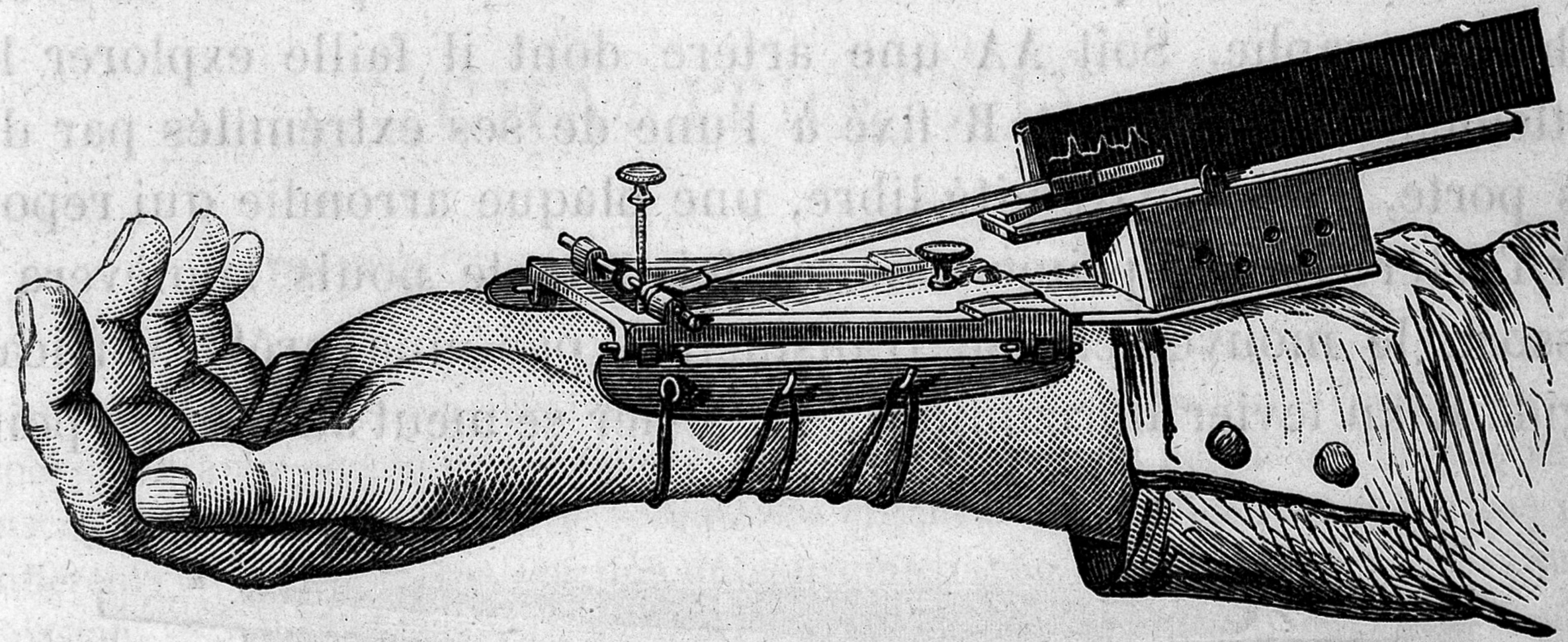
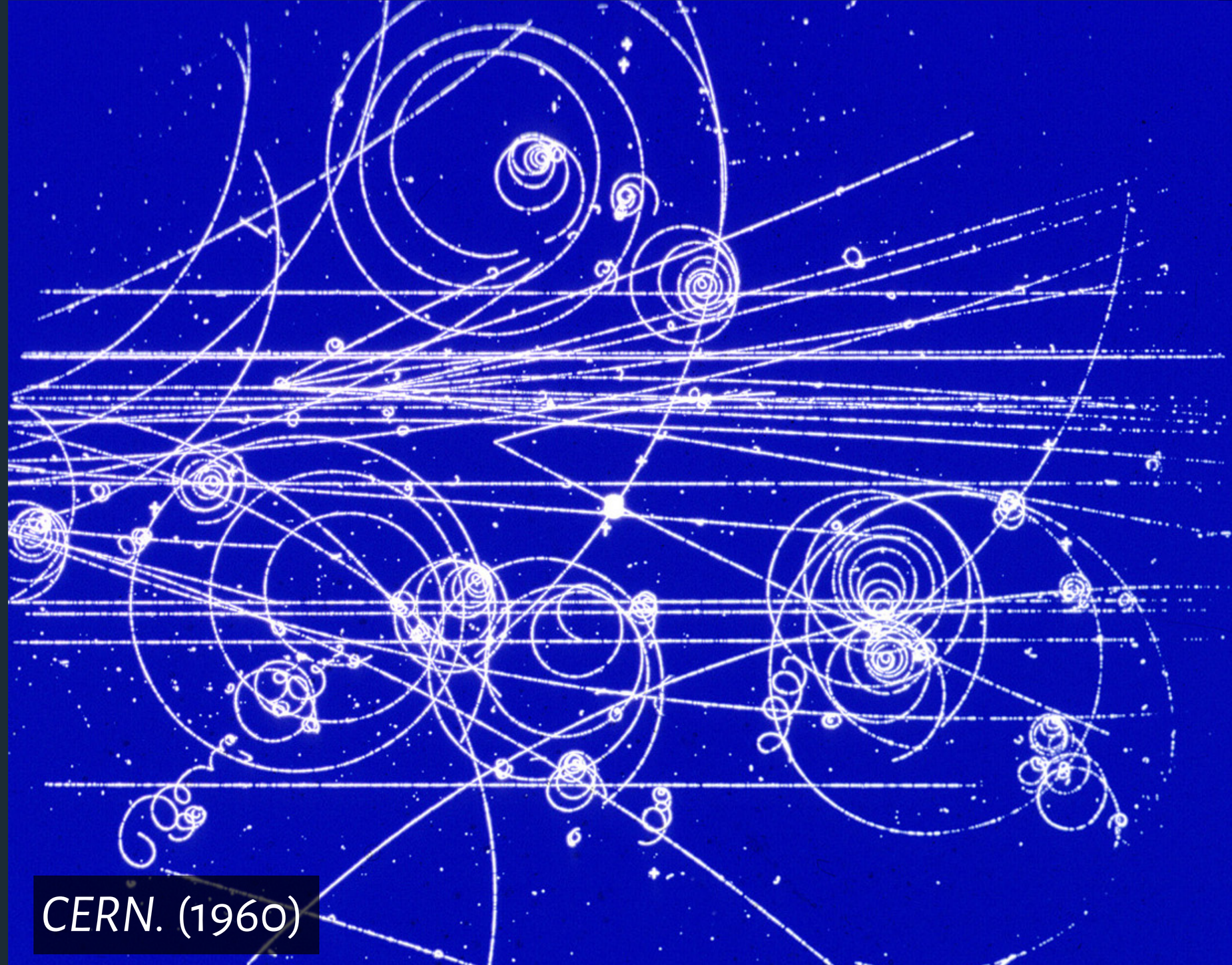


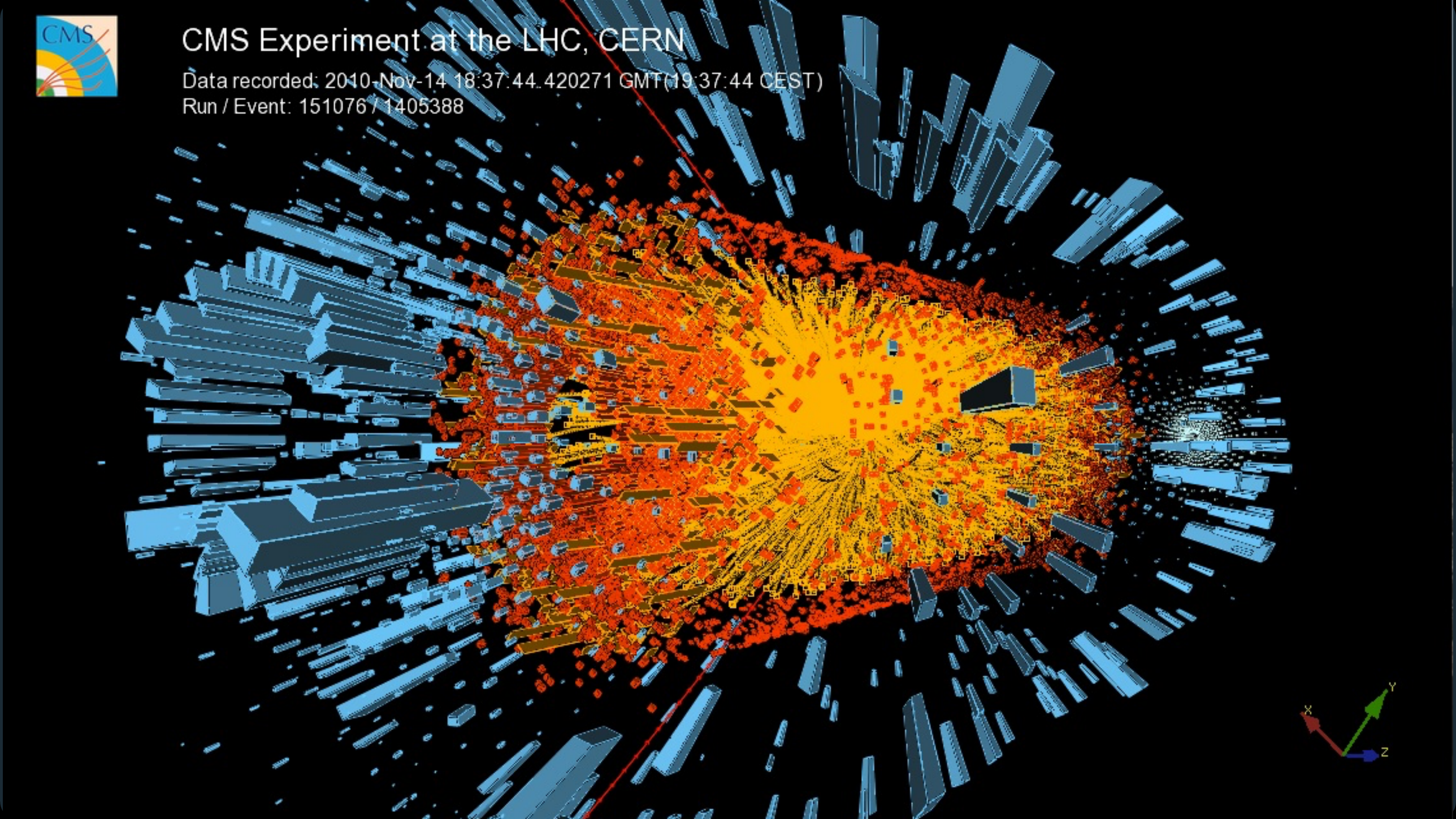
Fig. 109. Sphygmographe direct.

Record Info

Cloud and **bubble chambers** reveal properties of subatomic particles by making their tracks visible.



CERN. (1960)


CMS Experiment at the LHC, CERN
Data recorded: 2010-Nov-14 18:37:44.420271 GMT(19:37:44 CEST)
Run / Event: 151076 / 1405388
Y
X
Z

The Value of Visualization

Aka why create visualizations?

Record information

- Blueprints, photographs, seismographs, ...

Analyze data to support reasoning (exploratory visualization)

- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

Communicate information to others (explanatory visualization)

- Share and persuade
- Collaborate and revise

The Value of Visualization

Aka why create visualizations?

Record information

- Blueprints, photographs, seismographs, ...

Analyze data to support reasoning (exploratory visualization)

- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

Communicate information to others (explanatory visualization)

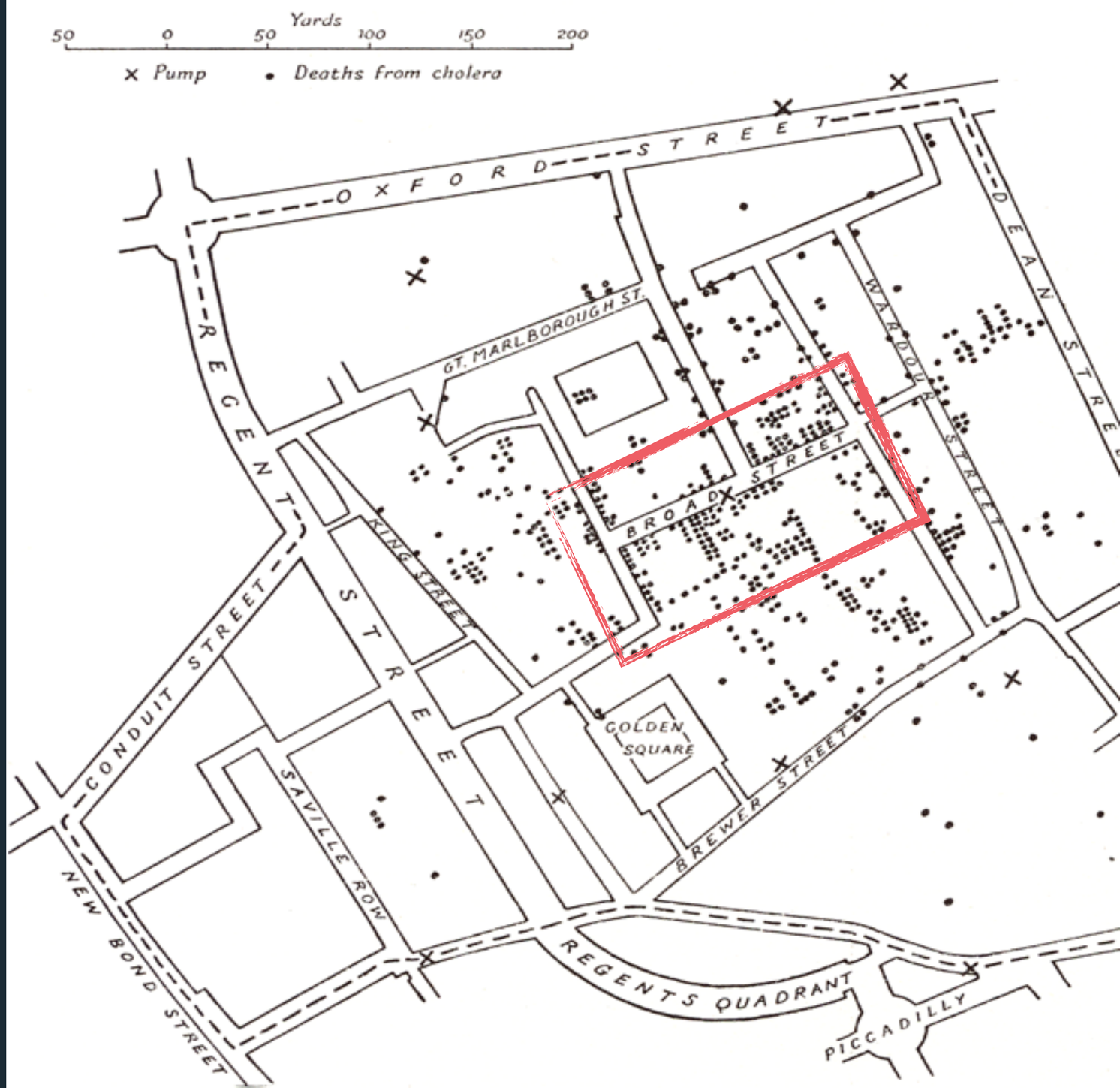
- Share and persuade
- Collaborate and revise

Support Reasoning

To investigate London's 1854 cholera epidemic, **John Snow** plotted position of each case on a map.

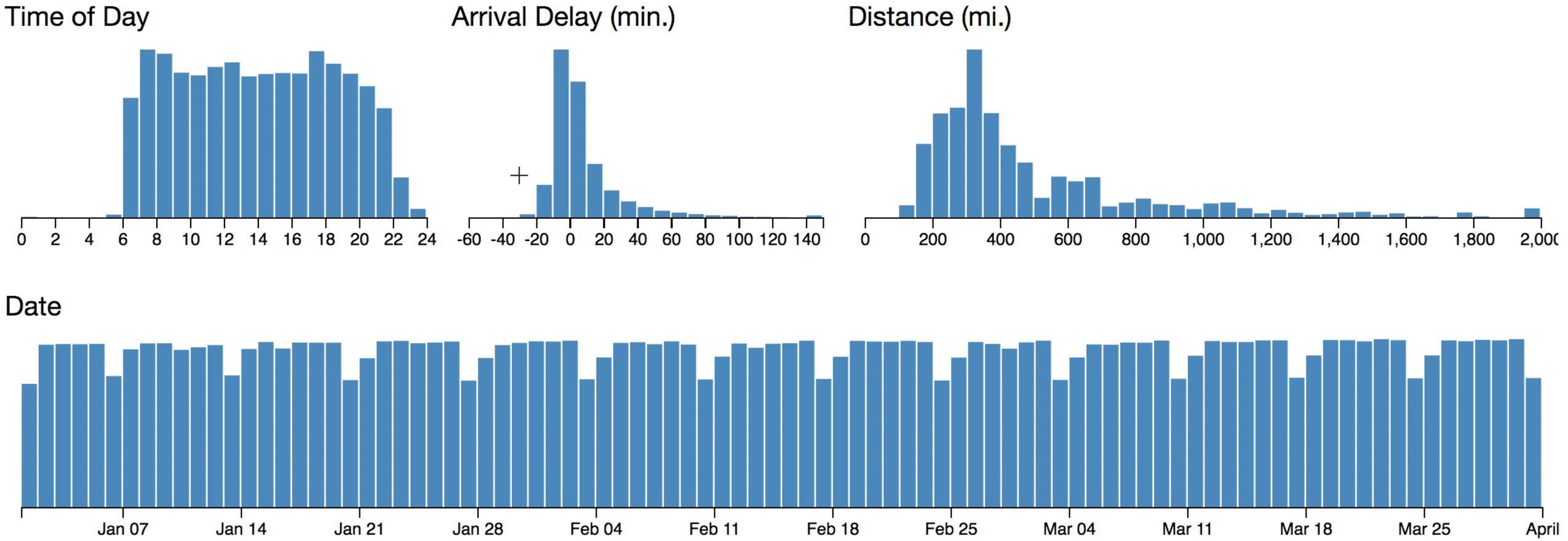
Map put the data in context.

Used to support hypothesis that Broad St. pump was the cause.



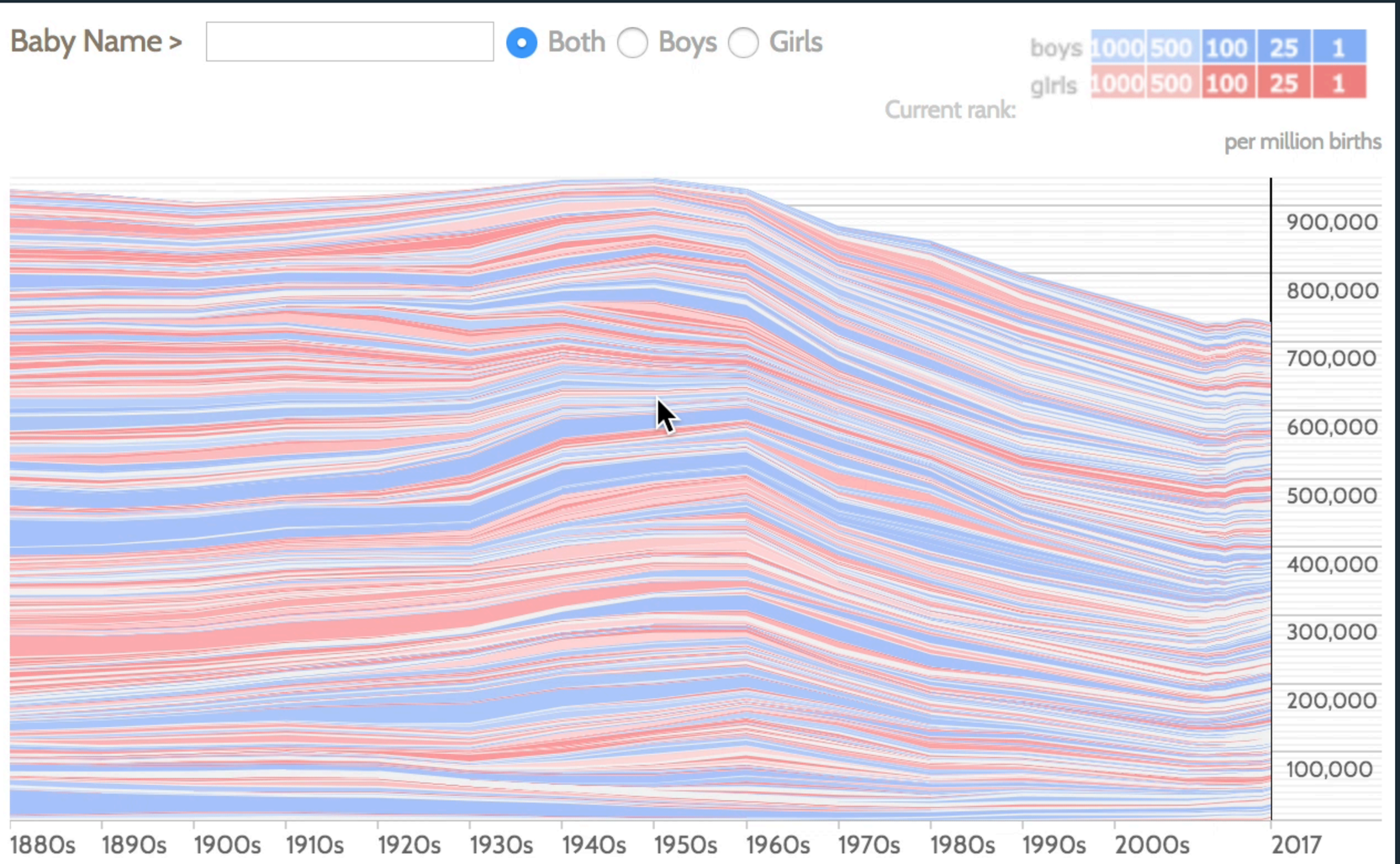
Support Reasoning

Interactively generate and evaluate hypotheses.



Crossfilter.js, *Bostock*. (2015)

Support Reasoning



NameVoyager, *Wattenberg*. (2005)

One of the first interactive visualizations that enabled **social data analysis**.

Engaging design + data fueled collaborative exploration.

- > *"Which letter has gone down most consistently? W? Observation: Note the recent upsurge in Y; basically all due to Hispanic (and some Middle Eastern) names"*
- < *"You're right, W has gone most consistently down, although F is pretty close (if it weren't for Faith...)"*

Try it out at:
babynamewizard.com/voyager

Support Reasoning

Class Exercise!

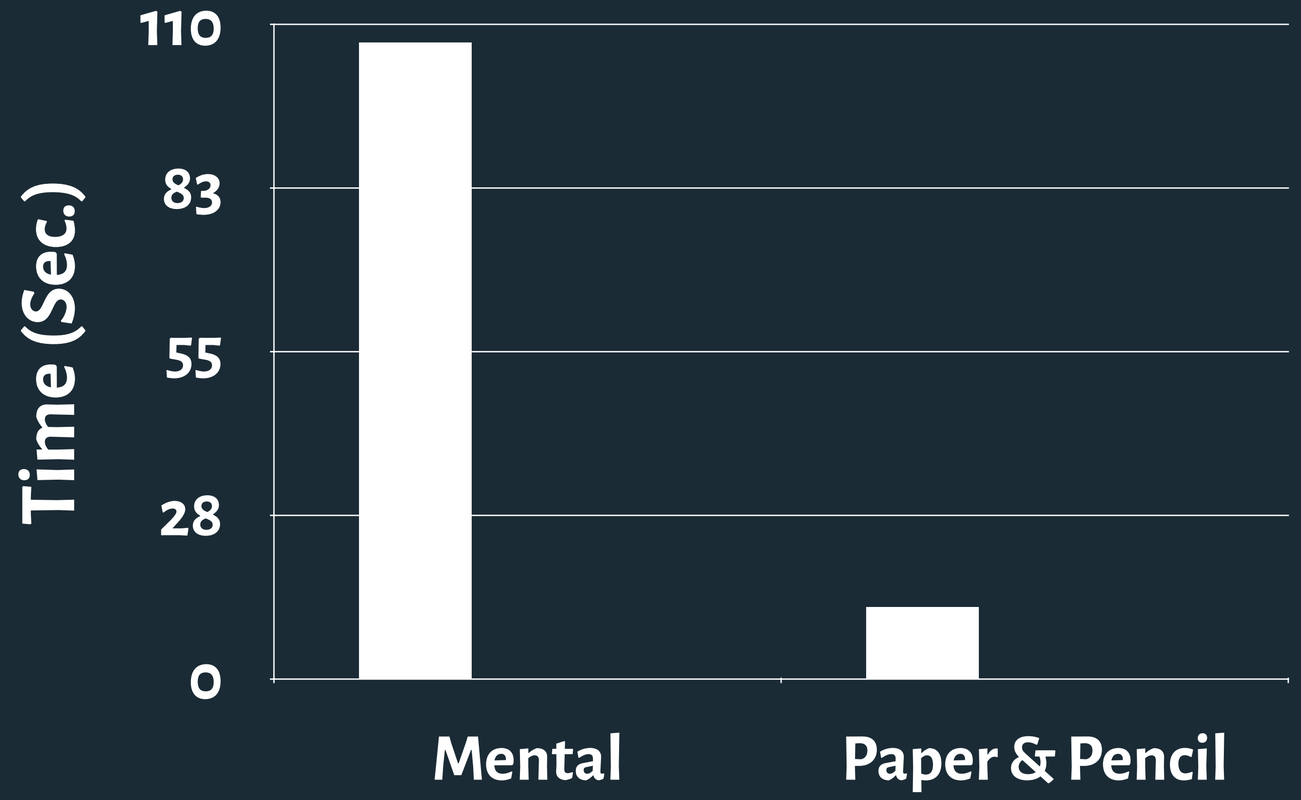
34 x 72

Support Reasoning

$$\begin{array}{rr} & 34 \\ \times & 72 \end{array}$$

Support Reasoning

```
   34
x  72
-----
   68
 2380
-----
 2448
```



The Value of Visualization

Aka why create visualizations?

Record information

- Blueprints, photographs, seismographs, ...

Analyze data to support reasoning (exploratory visualization)

- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

Communicate information to others (explanatory visualization)

- Share and persuade
- Collaborate and revise

The Value of Visualization

Aka why create visualizations?

Record information

- Blueprints, photographs, seismographs, ...

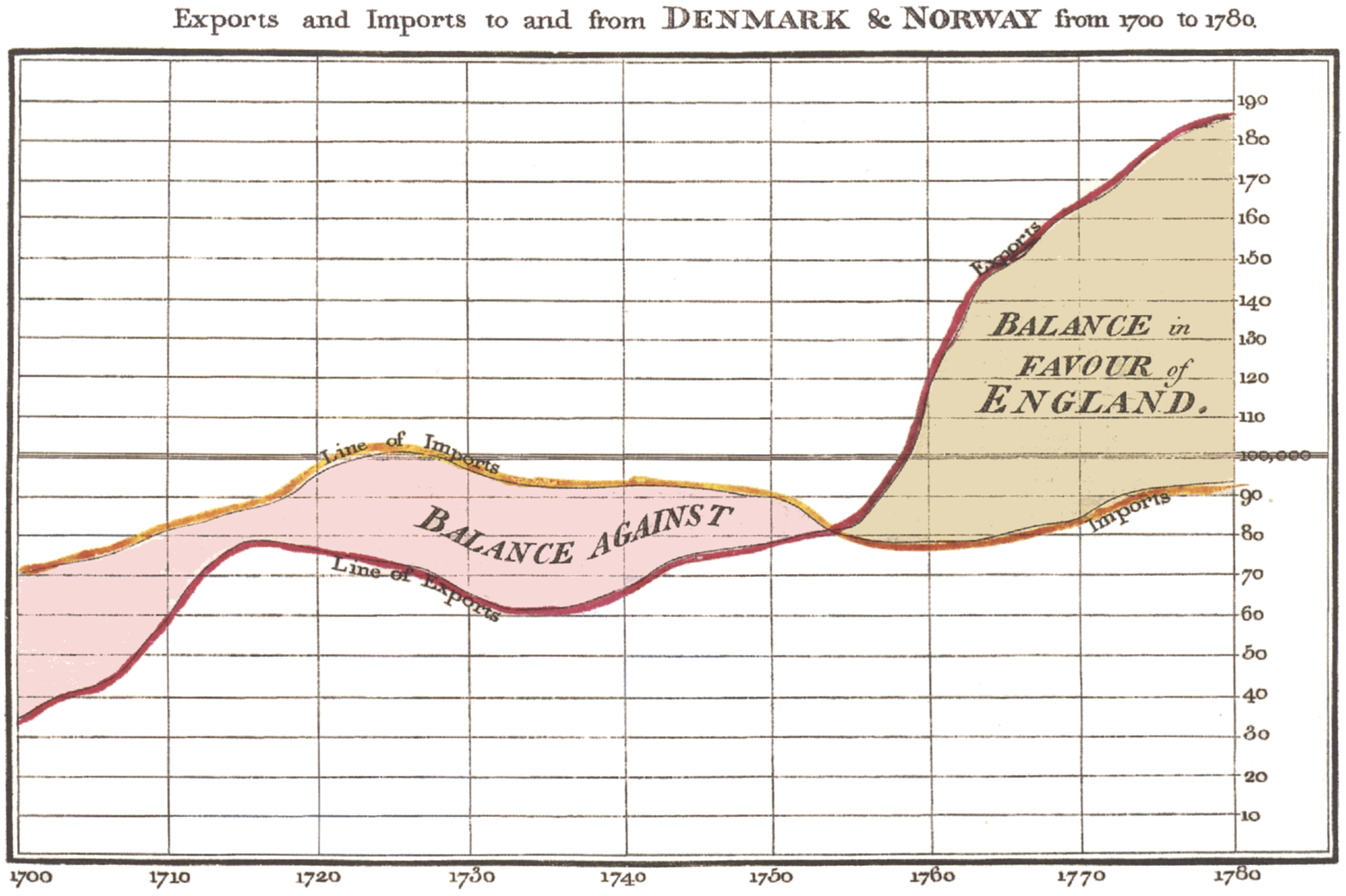
Analyze data to support reasoning (exploratory visualization)

- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

Communicate information to others (explanatory visualization)

- Share and persuade
- Collaborate and revise

Communicate Info

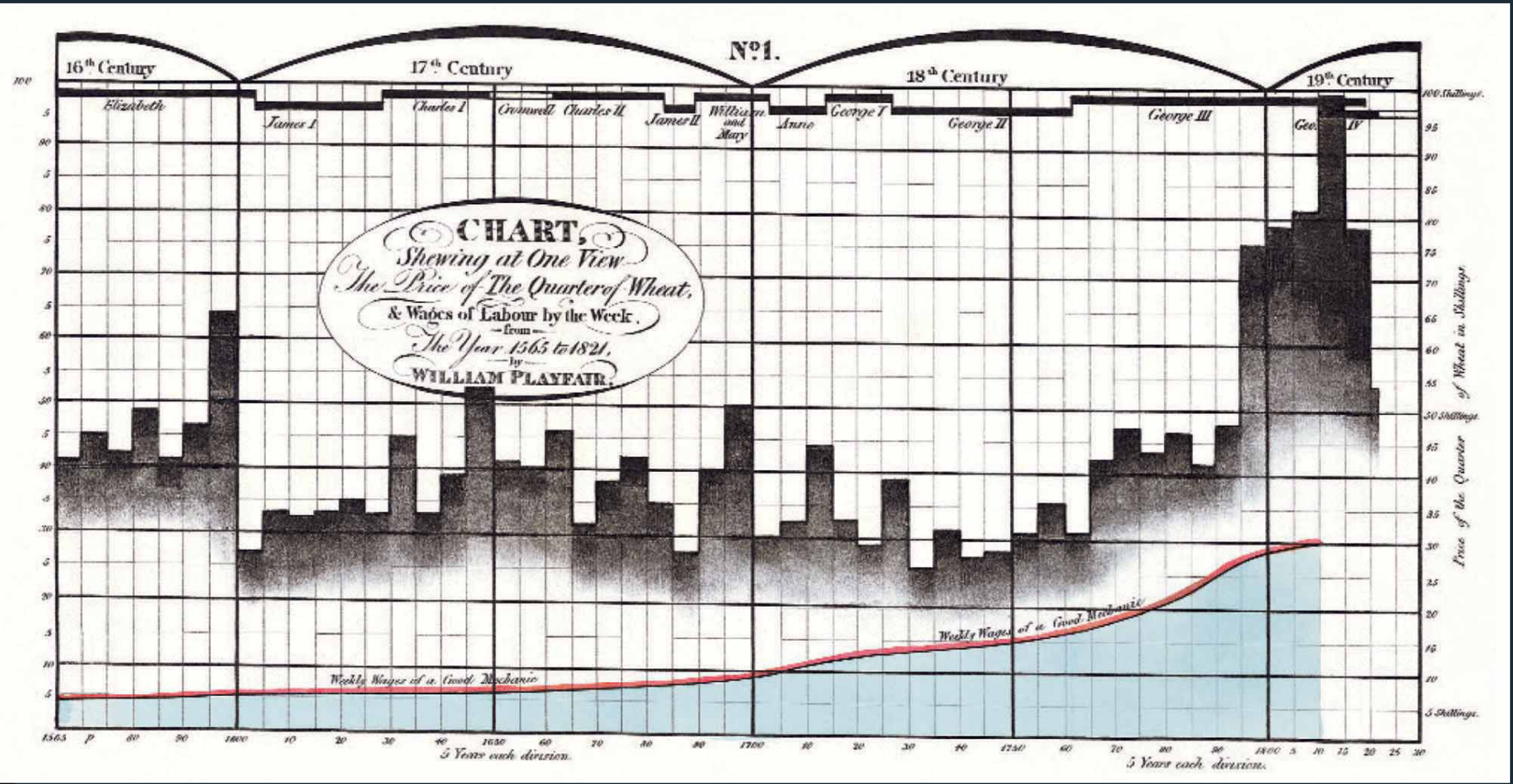


William Playfair, a Scottish engineer and economist, is credited with inventing modern graphical methods.

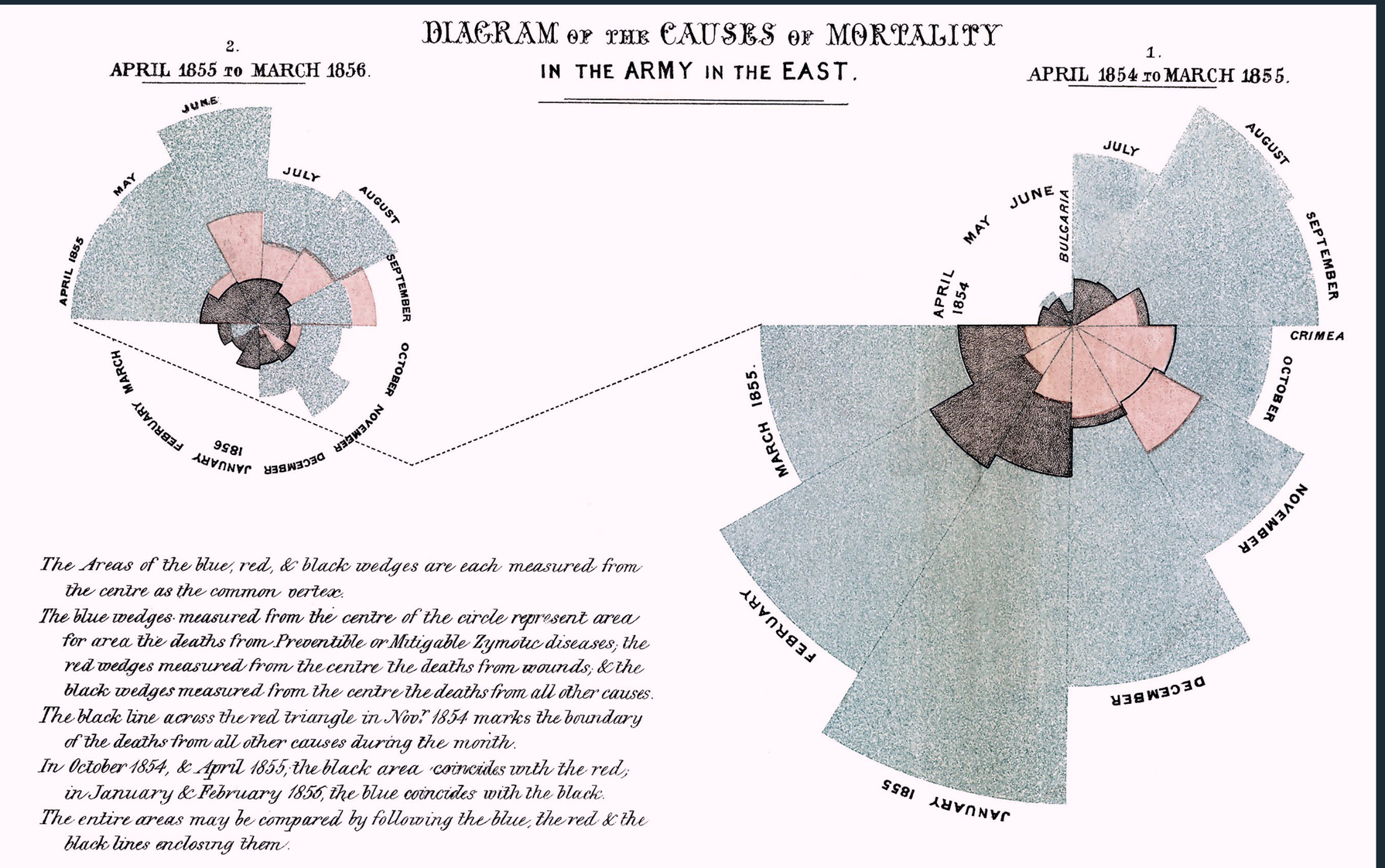
In 1786, published *The Commercial and Political Atlas* which contained the first time-series and bar charts.

Communicate Info

"You have before you, my Lords and Gentlemen, a chart of the prices of wheat for 250 years [...] the main fact deserving of consideration is, that never at any former period was wheat so cheap, in proportion to mechanical labour, as it is at the present time" — William Playfair, 1822 letter to Parliament.



Communicate Info



"to affect thro' the Eyes what we fail to convey to the public through their word-proof ears"

— Florence Nightingale on her "coxcomb" of Crimean War Deaths (1856).

Chart vividly depicts that the main cause of deaths was not war wounds but unsanitary conditions.

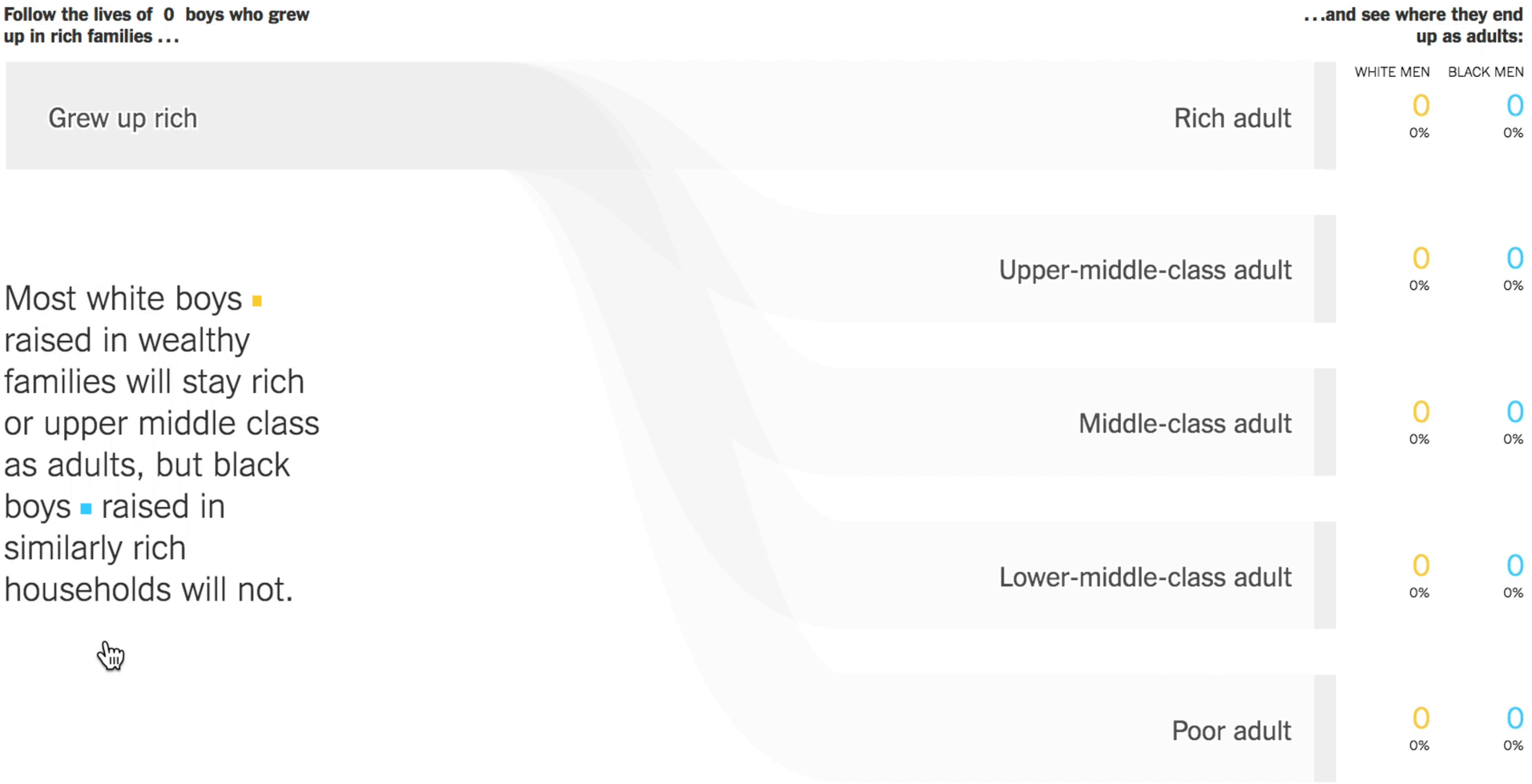
Returned to Britain and led a successful campaign for better conditions in barracks and hospitals.

Communicate Info

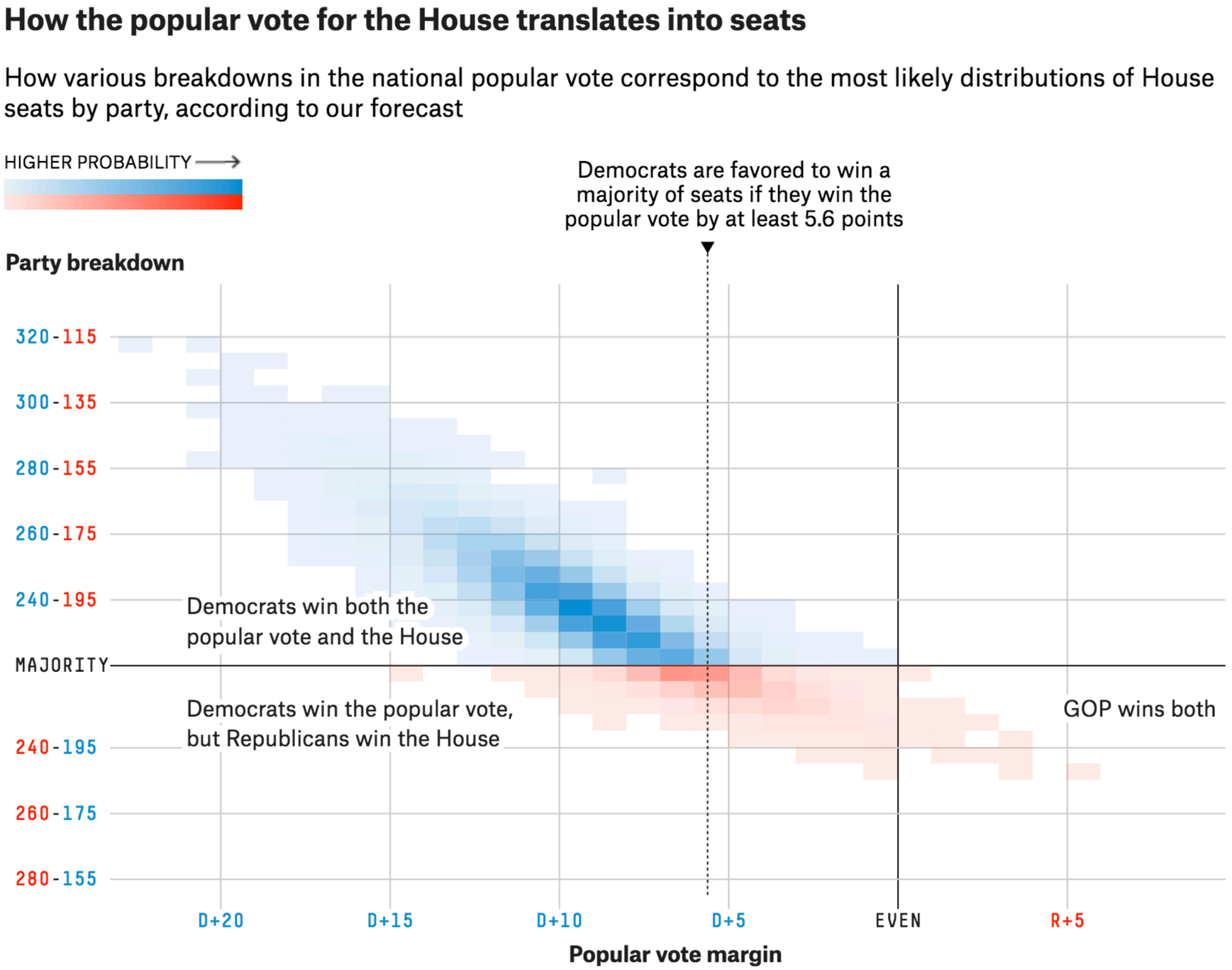
The Best Stats You've Ever Seen, *Hans Rosling* (2006).

Extensive Data Shows Punishing Reach of Racism for Black Boys

By **EMILY BADGER, CLAIRE CAIN MILLER, ADAM PEARCE** and **KEVIN QUEALY** MARCH 19, 2018



Adult outcomes reflect household incomes in 2014 and 2015.



House Forecast, *FiveThirtyEight* (2018).

The Value of Visualization

Aka why create visualizations?

Record information

- Blueprints, photographs, seismographs, ...

Analyze data to support reasoning (exploratory visualization)

- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

Communicate information to others (explanatory visualization)

- Share and persuade
- Collaborate and revise

The Value of Visualization

Aka why create visualizations?

Record information

- Blueprints, photographs, seismographs, ...

Analyze data to support reasoning (exploratory visualization)

- Develop and assess hypotheses
- Find patterns / Discover errors in data
- Expand memory

Communicate information to others (explanatory visualization)

- Share and persuade
- Collaborate and revise

Course Mechanics

http://vis.csail.mit.edu/classes/6.859/

Course Goals

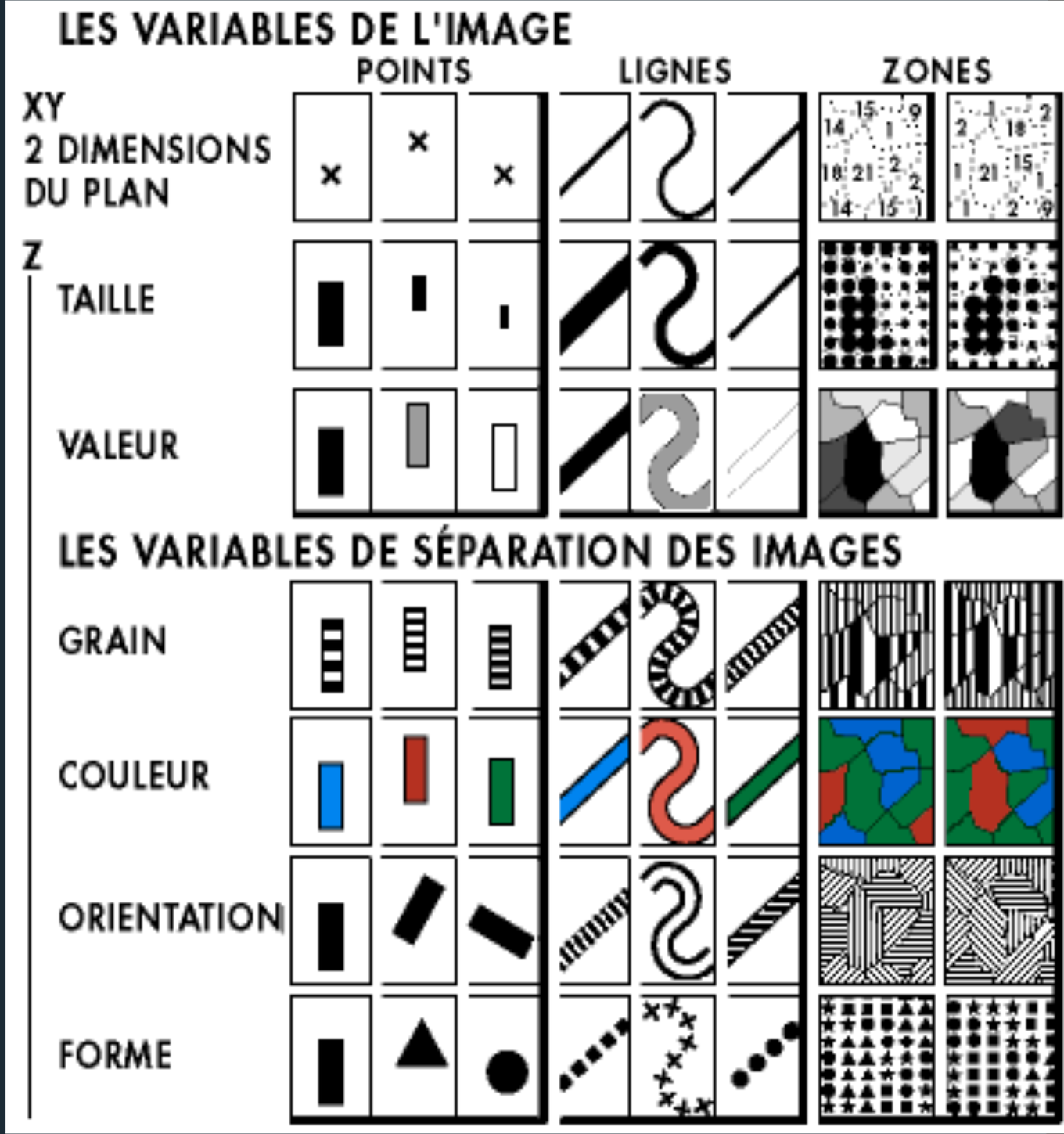
By the end of the course, you should expect to be able to:

1. *Understand* key visualization techniques and theory.
2. *Design*, *evaluate*, and *critique* visualizations.
3. *Wrangle*, *explore*, and *explain* datasets using visualizations.
4. *Implement* interactive data visualizations.
5. *Develop* a substantial visualization project.
6. *Read* and *discuss* visualization research.

Course Topics

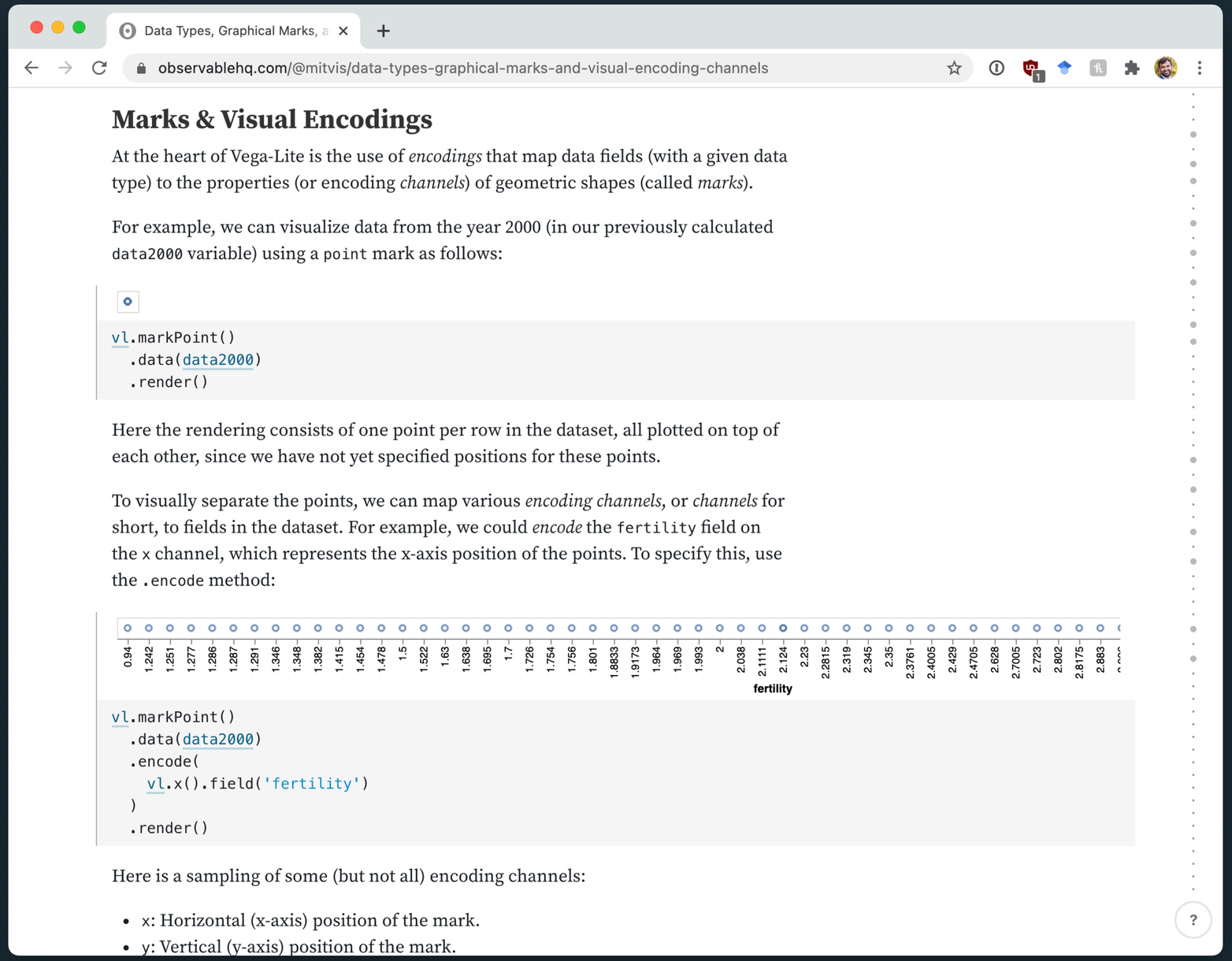
Course Topics

Data & Image Models



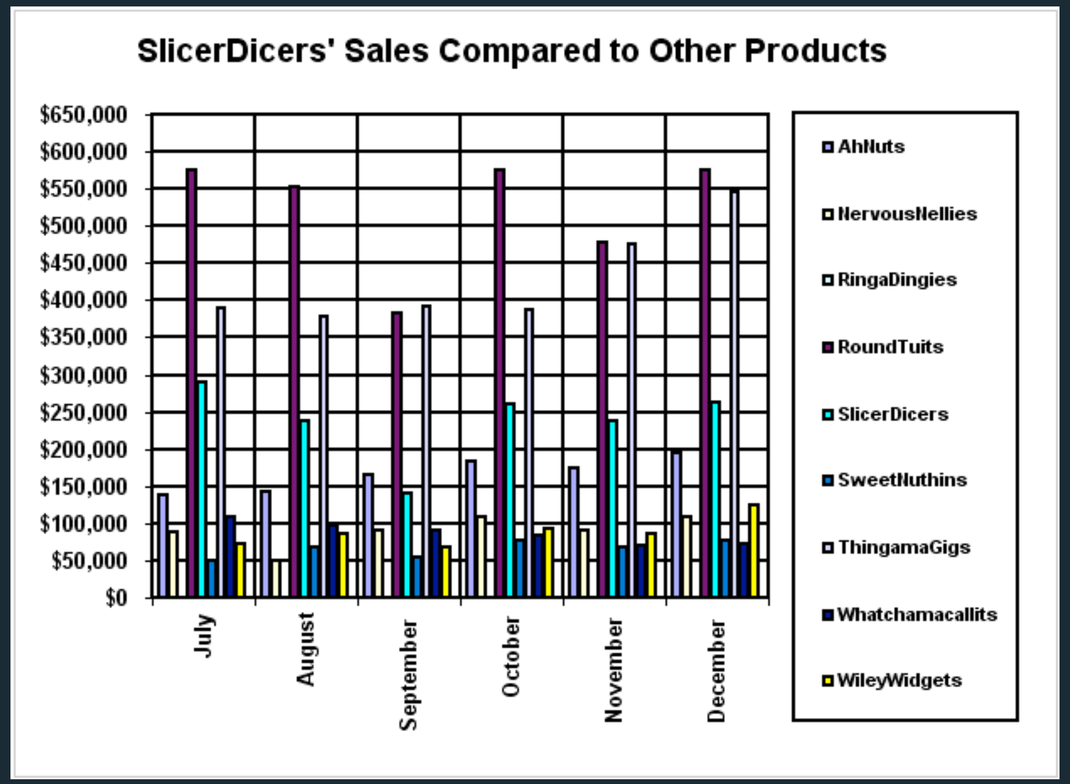
Course Topics

Visual Encoding with Vega-Lite

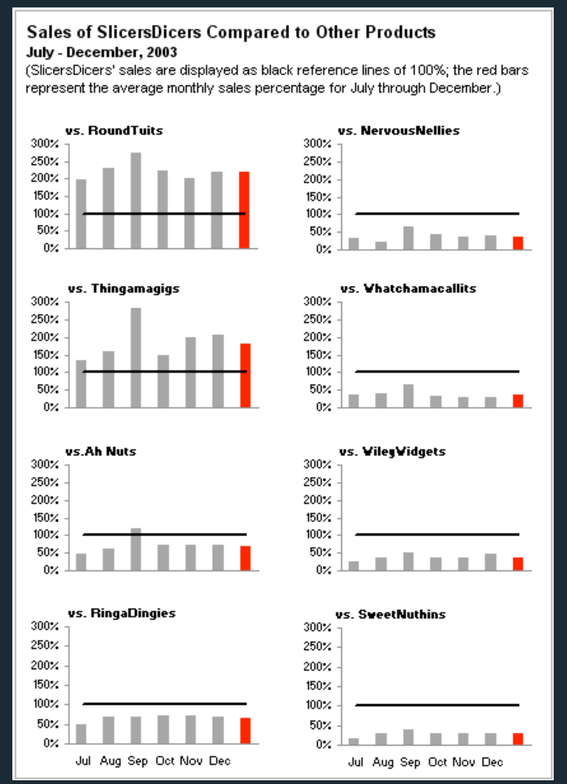


Course Topics

(Re-)Design



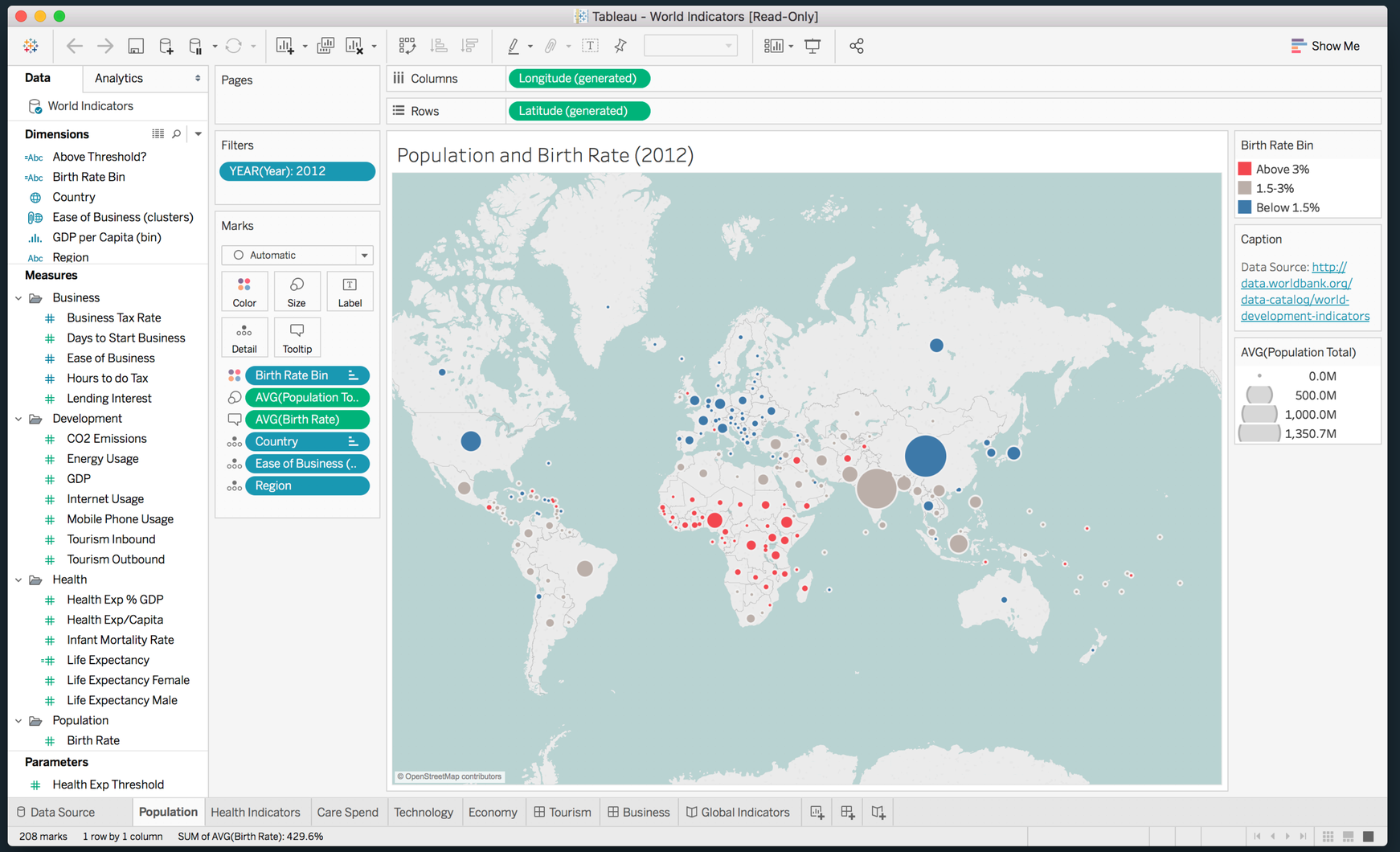
Problematic Design



Redesign

Course Topics

Exploratory Data Analysis (EDA)



Course Topics

Perception

blue

yellow

red

green

orange

purple

Course Topics

Perception

blue

yellow

red

green

orange

purple

Course Topics

Perception

blue

yellow

red

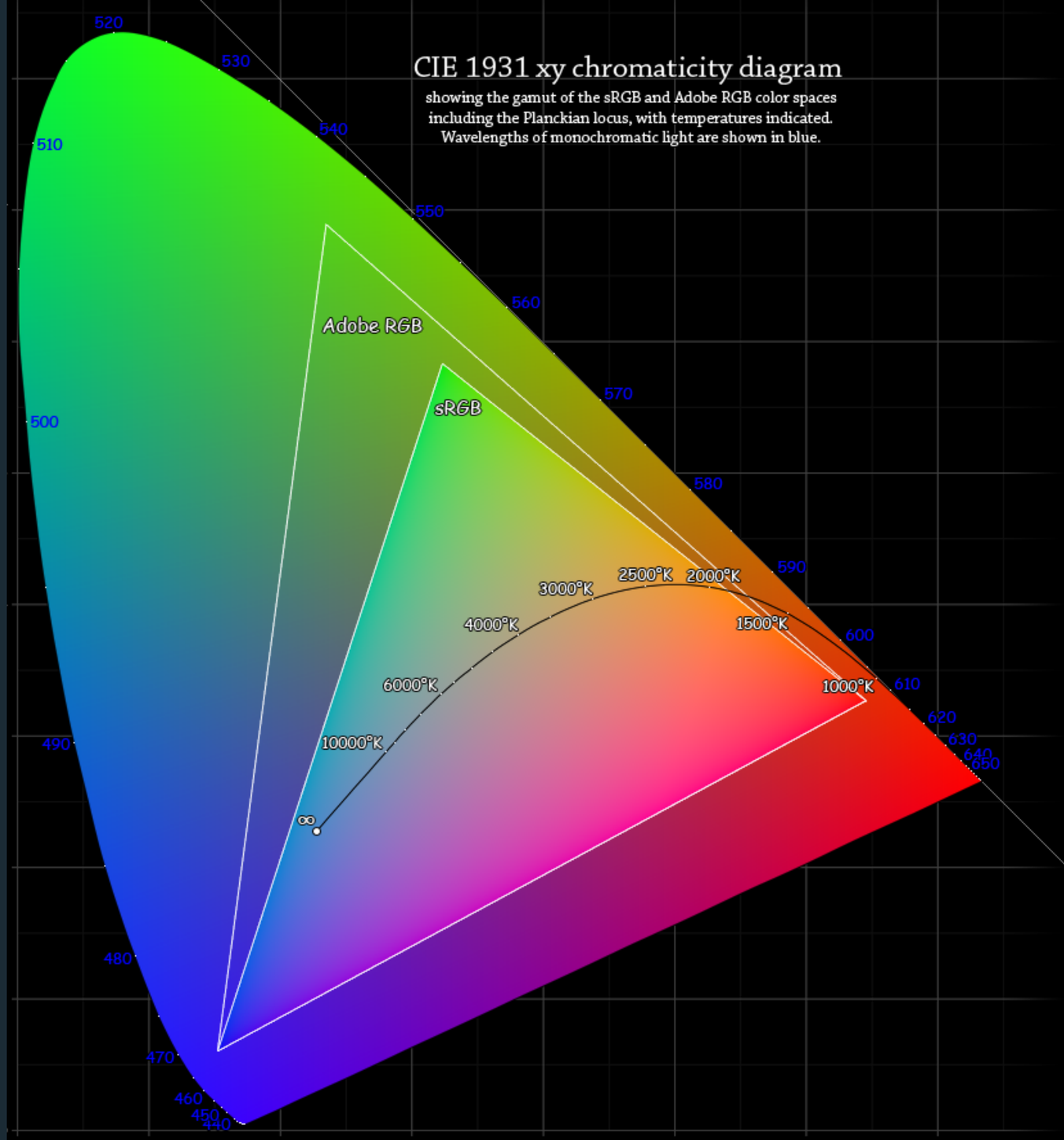
green

orange

purple

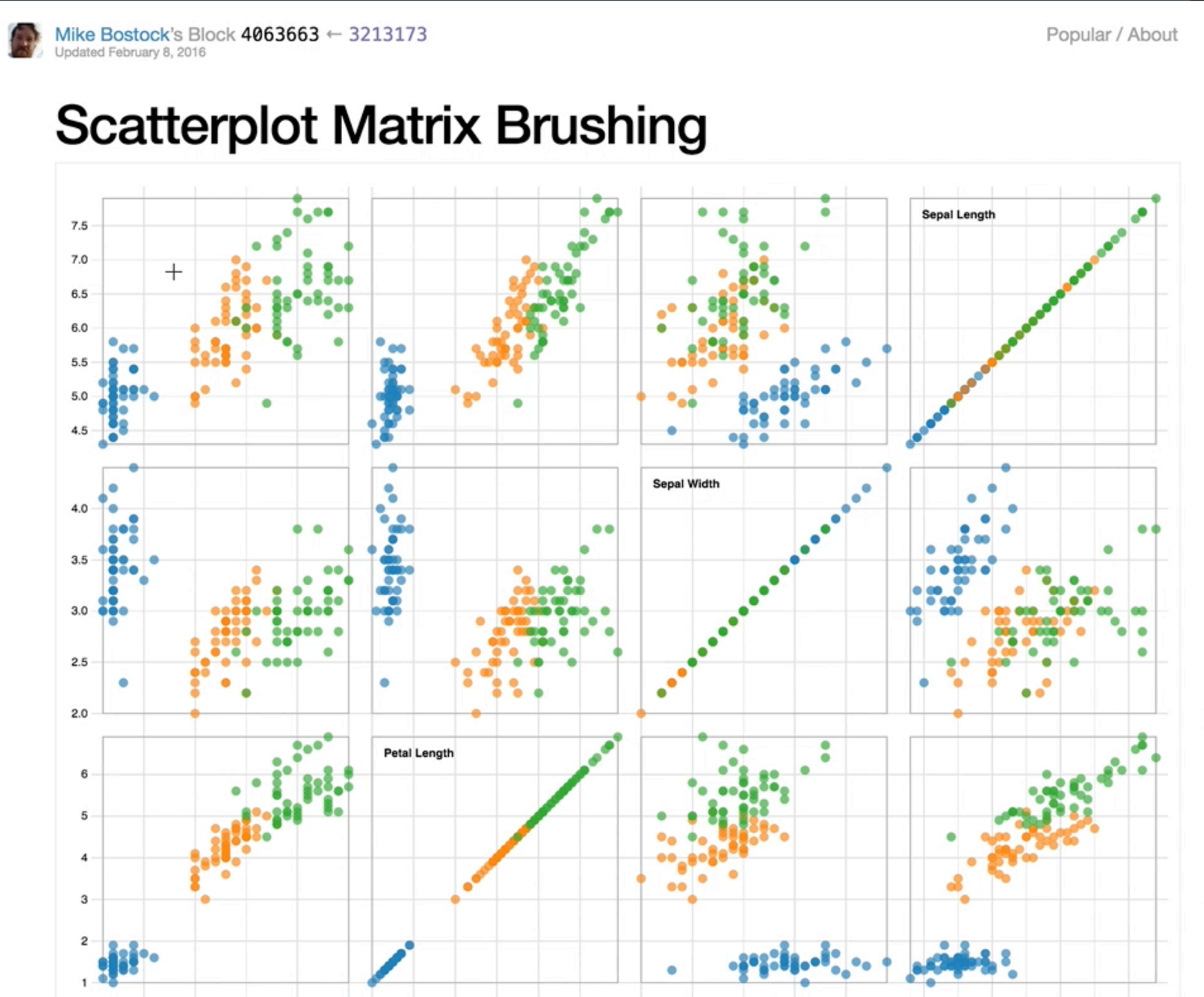
Course Topics

Color



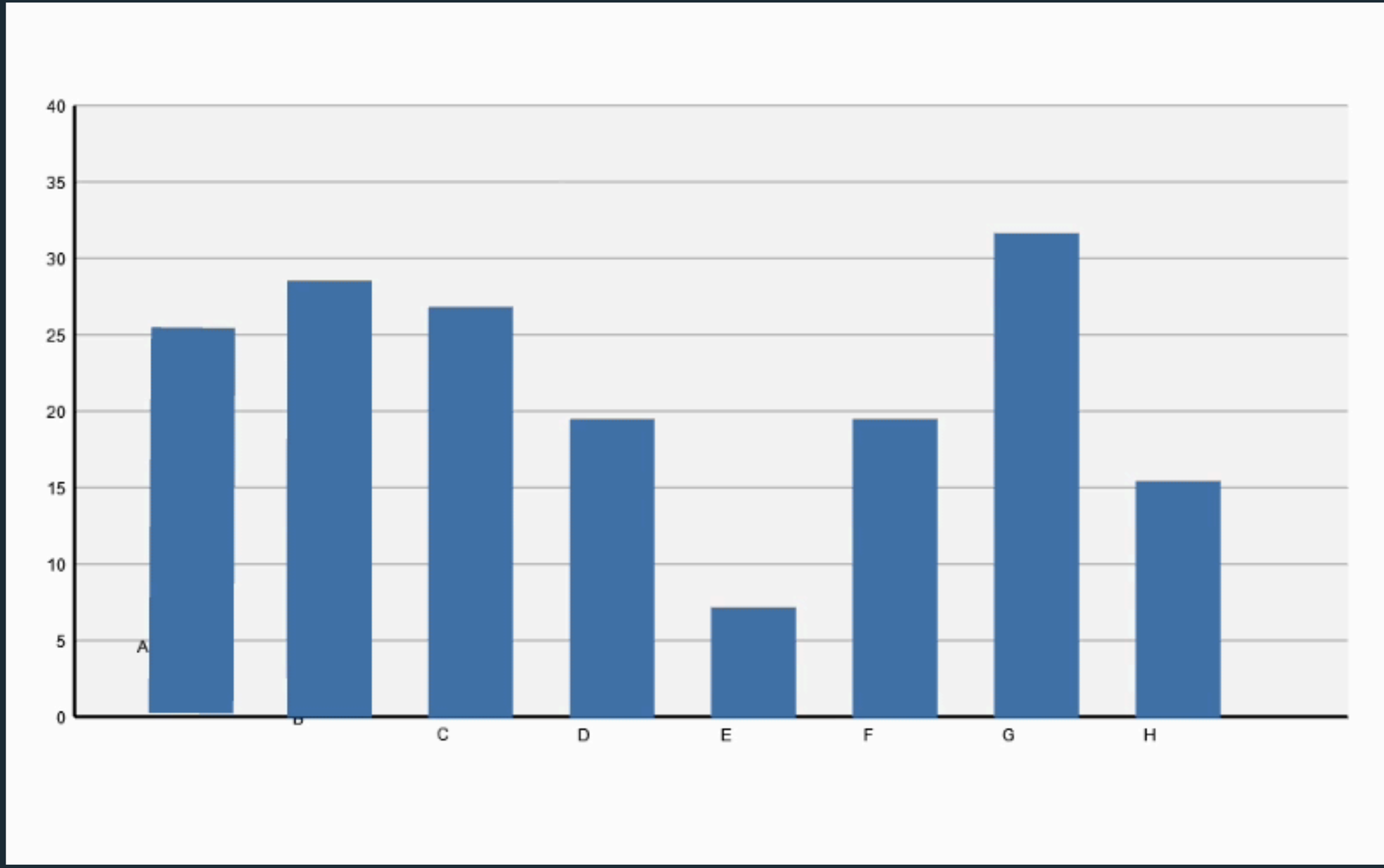
Course Topics

Interactivity



Course Topics

Animation



Heer & Robertson. (2007).

Course Topics

Narrative

archive.nytimes.com/www.nytimes.com/interactive/2013/10/09/us

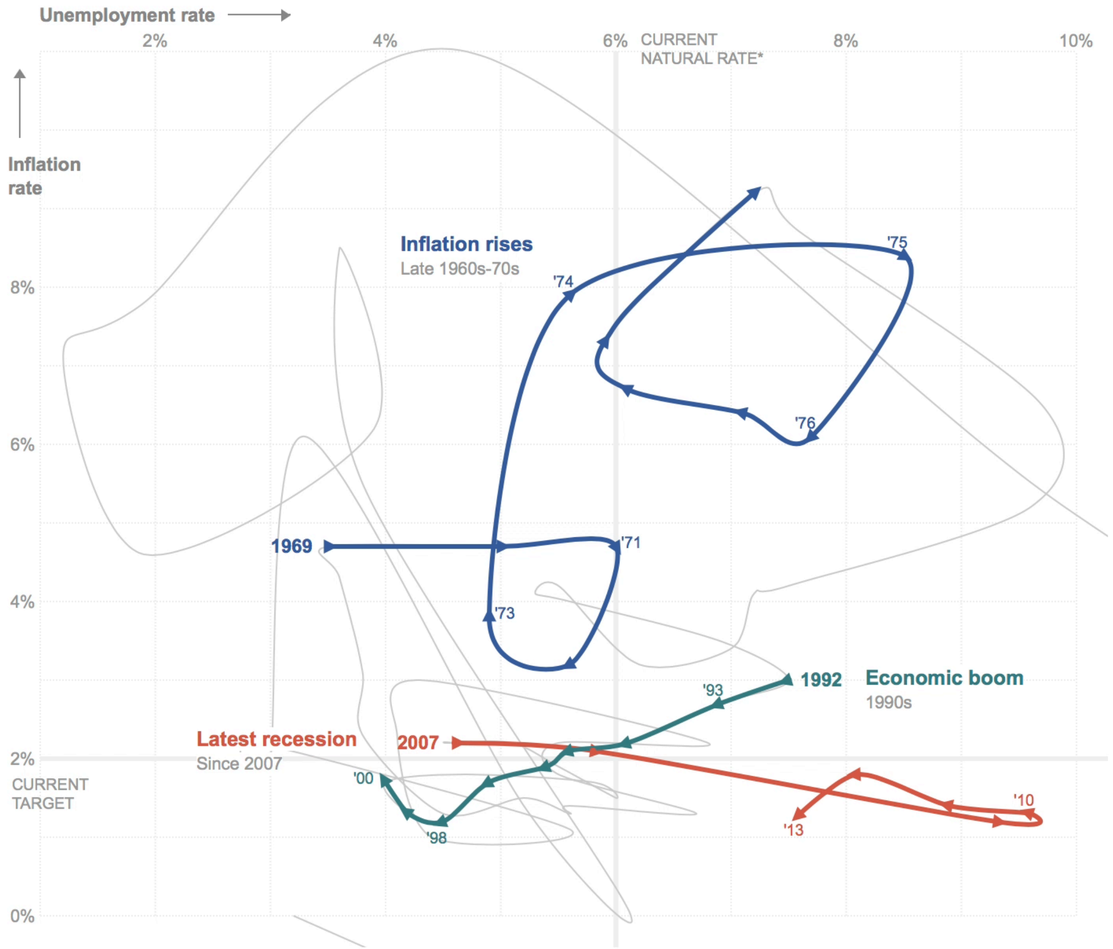
Janet L. Yellen, on the Economy's Twists and Turns

1 2 3 4 5 6 7 NEXT

Inflation and unemployment

The Federal Reserve is said to have a "dual mandate": keeping inflation in check and the unemployment rate low. These measures, which tend to change cyclically and in concert with each other, are charted for every year since the Great Depression.

In speeches and in meetings, Ms. Yellen, the nominee for the next Fed leader, has commented on the Fed's actions during significant periods, providing a window into her views and priorities.



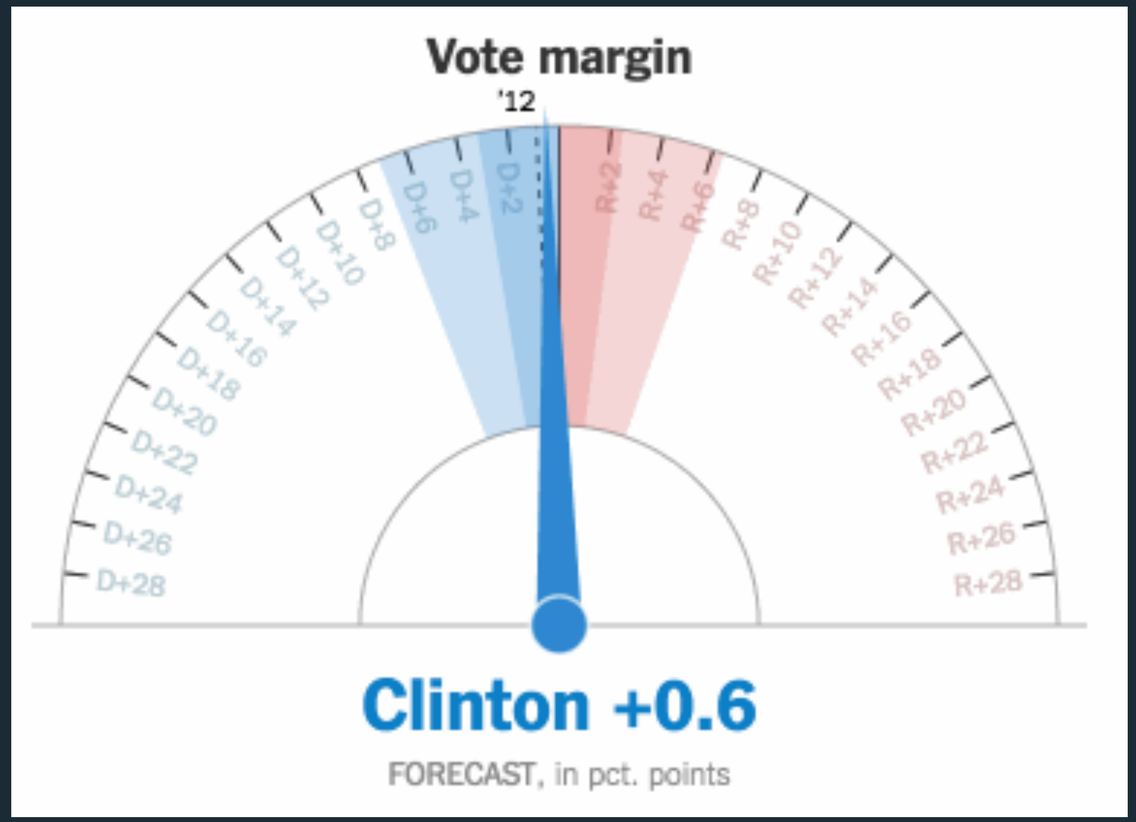
**The natural rate of unemployment is defined as the lowest sustainable level of unemployment over the long term. If the rate is pushed any lower than the natural level, wages and prices would rise.*

By TOM GIRATIKANON and ALICIA PARLAPIANO

Sources: Federal Reserve Bank of St. Louis (inflation, measured by annual change in core personal consumption expenditures); Bureau of Labor Statistics (unemployment rate, annual average); National Bureau of Economic Research (unemployment rate before 1947)

Course Topics

Error & Uncertainty



Course Topics

Mapping & Cartography

Dymaxion Maps, *Fuller.* (1946)

Course Staff

Course Staff

My career:

BS, UC San Diego → PhD, Stanford (+UW) →
Visiting Researcher, Google Brain → New*ish* faculty

Research Interests:

Interactive Data Visualization (!)

Machine Learning Interpretability

For fun:

Run along the esplanade

Cook (but not bake!)

Arvind Satyanarayan

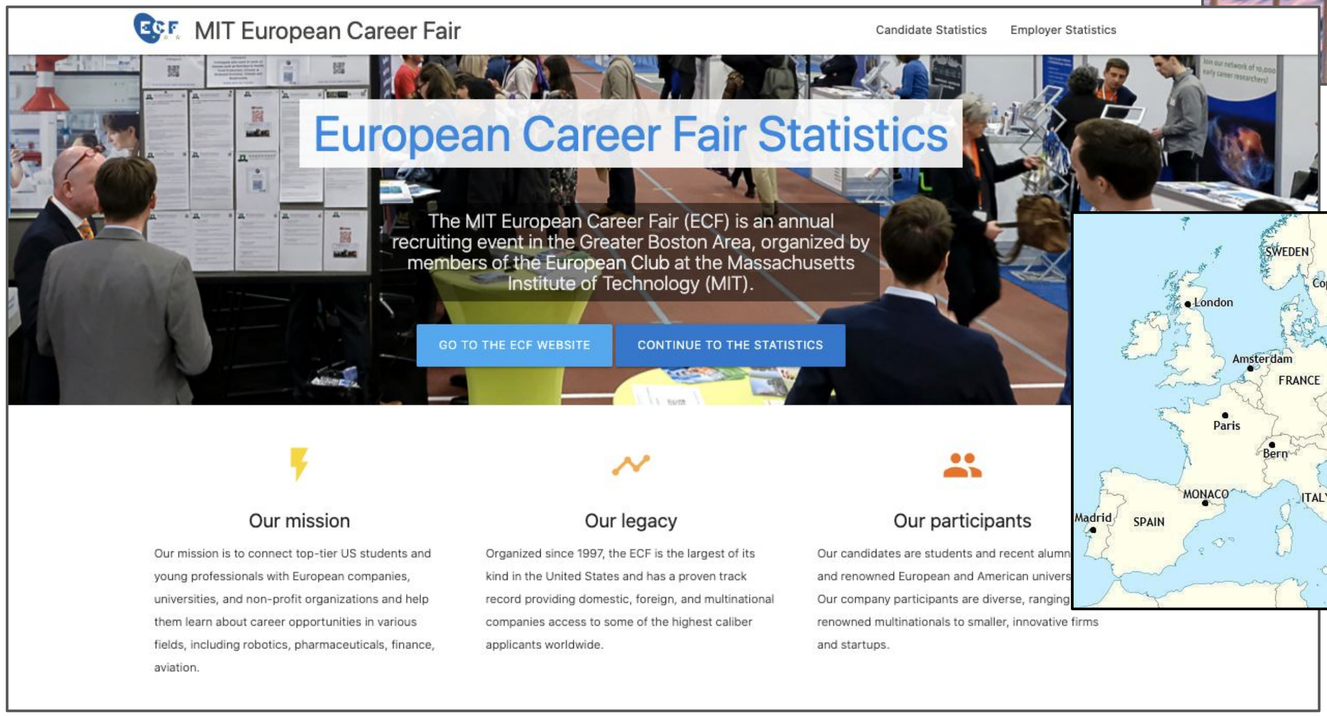
he/him

arvindsatya@mit.edu

Doğa (he/him)

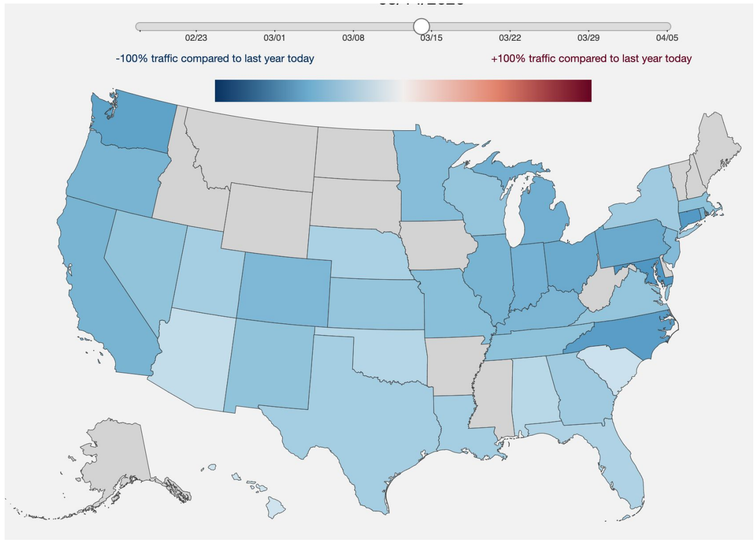
3rd-year PhD @ HCI Engineering Group, CSAIL
focus: *unobtrusive tags & digital fabrication*

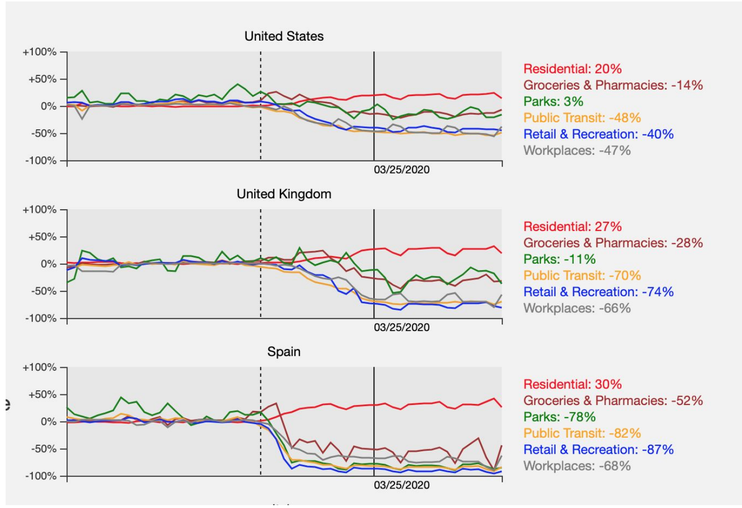
What I do in my free time: singing & playing the guitar, design, diversity & inclusion volunteering, Chinese, MIT European Club & European Career Fair



Saadiyah (she/her) *MEng*

MEng @ D-Lab





What I like to do in my free time: swim, cook, paint & learn new things!

EJ (he/him)

MEng @ Visualization Group - CSAIL

Katie *(she/her)*

MEng @ Visualization Group

Arts
& Crafting!

Course Grading

Class Participation	5%	
Reading Commentaries	5%	
A0: Sketching Visualizations	2%	Due 2/22
A1: Visualization Design	3%	Due 3/1
A2: Exploratory Data Analysis	10%	Due 3/9
A3: White/Black Hat Visualization	15%	Due 3/24
A4: Interactive Narratives	20%	Due 4/5, 4/12
Final Project	40%	
Proposal		Due 4/16
MVP + Presentations		Due 5/3
Poster Session + Final Deliverables		Due 5/11

Course Grading

Class Participation	5%
Reading Commentaries	5%
A0: Sketching Visualizations	2%
A1: Visualization Design	3%
A2: Exploratory Data Analysis	10%
A3: White/Black Hat Visualization	15%
A4: Interactive Narratives	20%
Final Project	40%
Proposal	
MVP + Presentations	
Poster Session + Final Deliverables	

Lectures will be recorded and posted to Canvas.

You may attend asynchronously but we encourage synchronous attendance if you're able to.

Class Participation grade will be primarily determined by activity on **Slack**:

- Introduce yourself in #introductions
- Ask and answer questions
- Post links to + critique interesting visualizations you find online.
- Share your work!!

Course Grading

Class Participation	5%
Reading Commentaries	5%
A0: Sketching Visualizations	2%
A1: Visualization Design	3%
A2: Exploratory Data Analysis	10%
A3: White/Black Hat Visualization	15%
A4: Interactive Narratives	20%
Final Project	40%
Proposal	
MVP + Presentations	
Poster Session + Final Deliverables	

Readings posted on nb2.csail.mit.edu – sign up invitations emailed, contact staff if you haven't received them.

1-2 readings per week – a mix of research papers, articles, interactive exercises.

On nb, post 1 paragraph per reading.

- Should *not* be a summary.
- Start a new thread, respond to an existing thread, etc.
- We'll discuss readings in class, so have commentaries posted **by noon**.
- You have 2 "passes" for the semester.

Course Grading

Class Participation	5%
Reading Commentaries	5%
A0: Sketching Visualizations	2%
A1: Visualization Design	3%
A2: Exploratory Data Analysis	10%
A3: White/Black Hat Visualization	15%
A4: Interactive Narratives	20%
Final Project	40%
Proposal	
MVP + Presentations	
Poster Session + Final Deliverables	

5 slack days which can be used as you wish for assignments.

Slack days should cover minor illnesses, special occasions (including religious holidays).

Additional extensions only granted for serious issues with a written note of support from S3 or GradSupport @ OGE.

Course Grading

Class Participation	5%
Reading Commentaries	5%
A0: Sketching Visualizations	2%
A1: Visualization Design	3%
A2: Exploratory Data Analysis	10%
A3: White/Black Hat Visualization	15%
A4: Interactive Narratives	20%
Final Project	40%
Proposal	
MVP + Presentations	
Poster Session + Final Deliverables	

Major visualization project on topic/ dataset of your choice.

Second half of the course.

Teams of 1–4 people with a 1 page project proposal.

In-class presentations of minimal viable product (MVP). Peer review/critique.

Final presentations during the last class session (format TBD).

Questions?

http://vis.csail.mit.edu/classes/6.859/