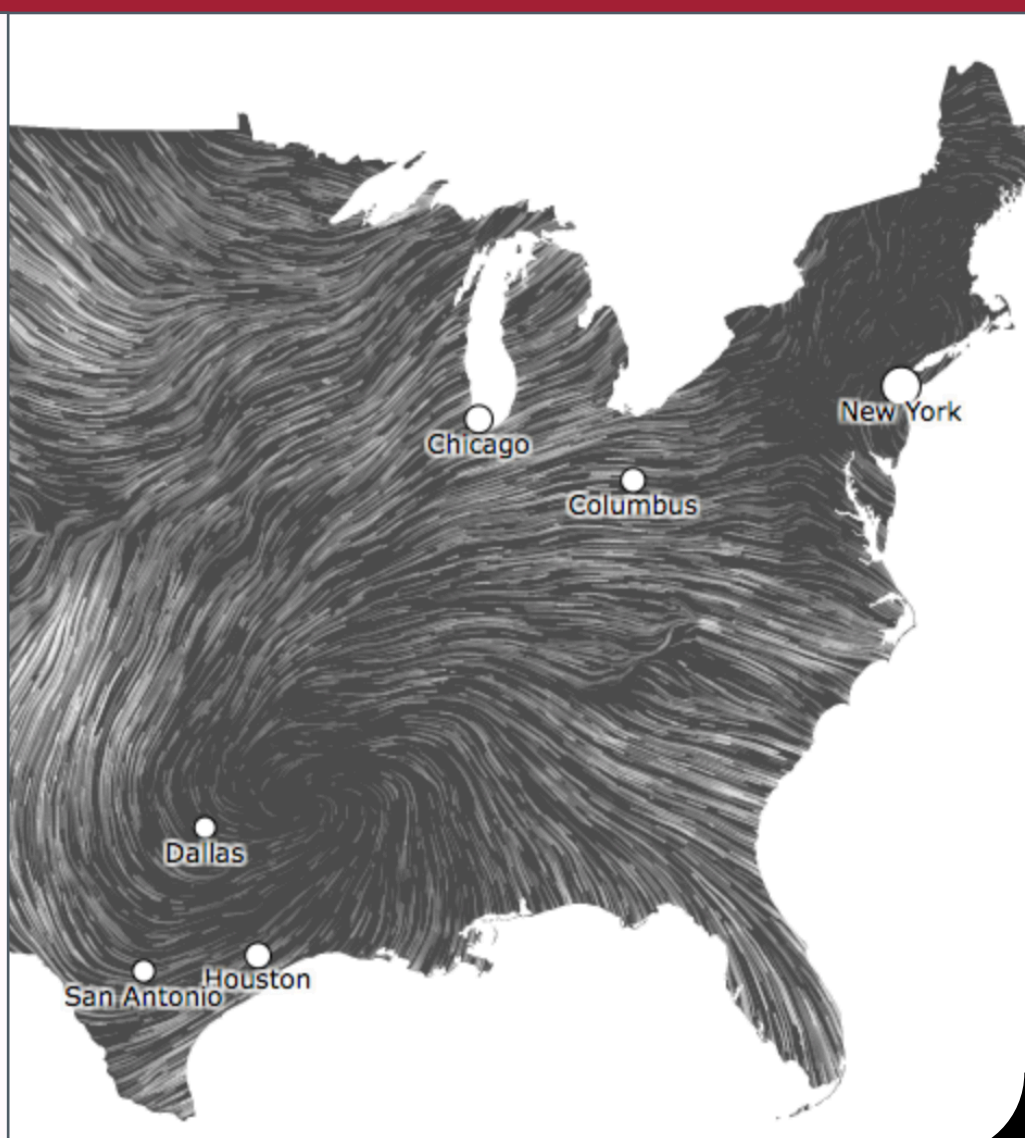
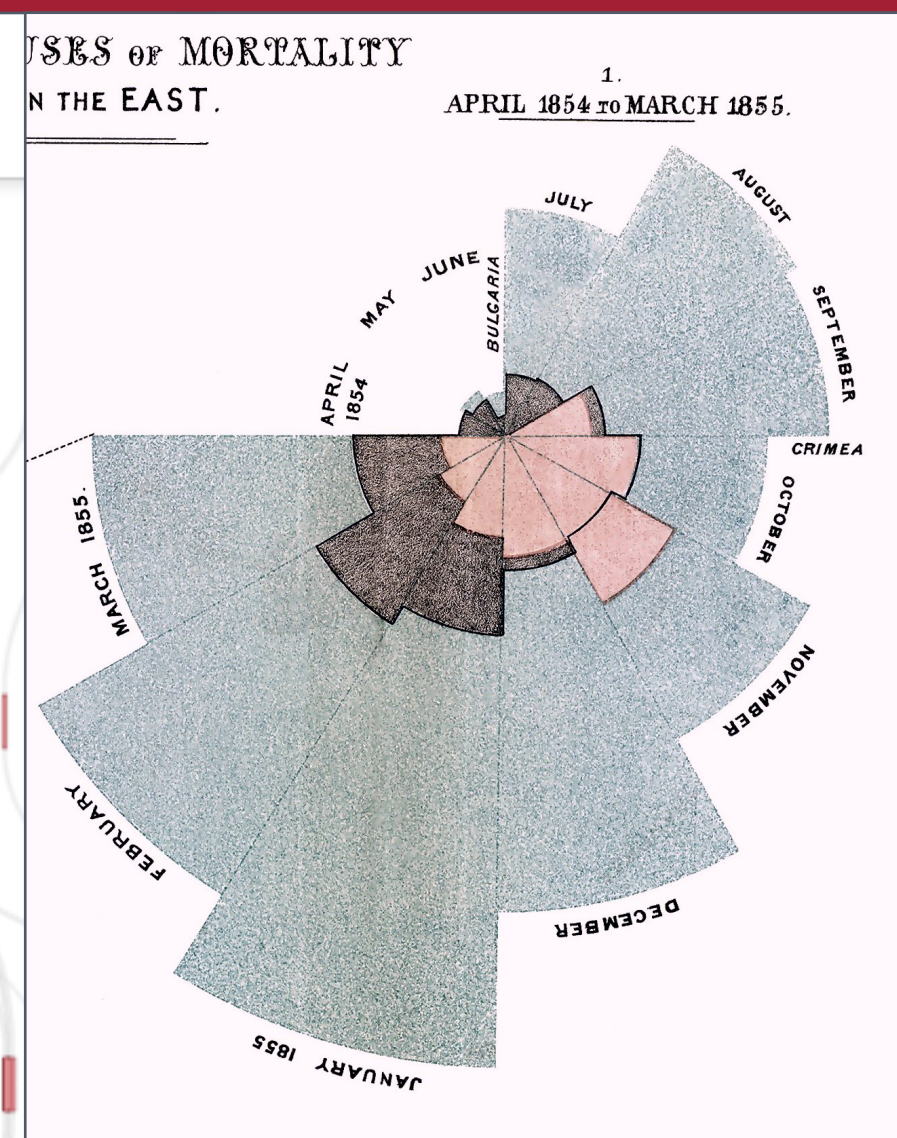
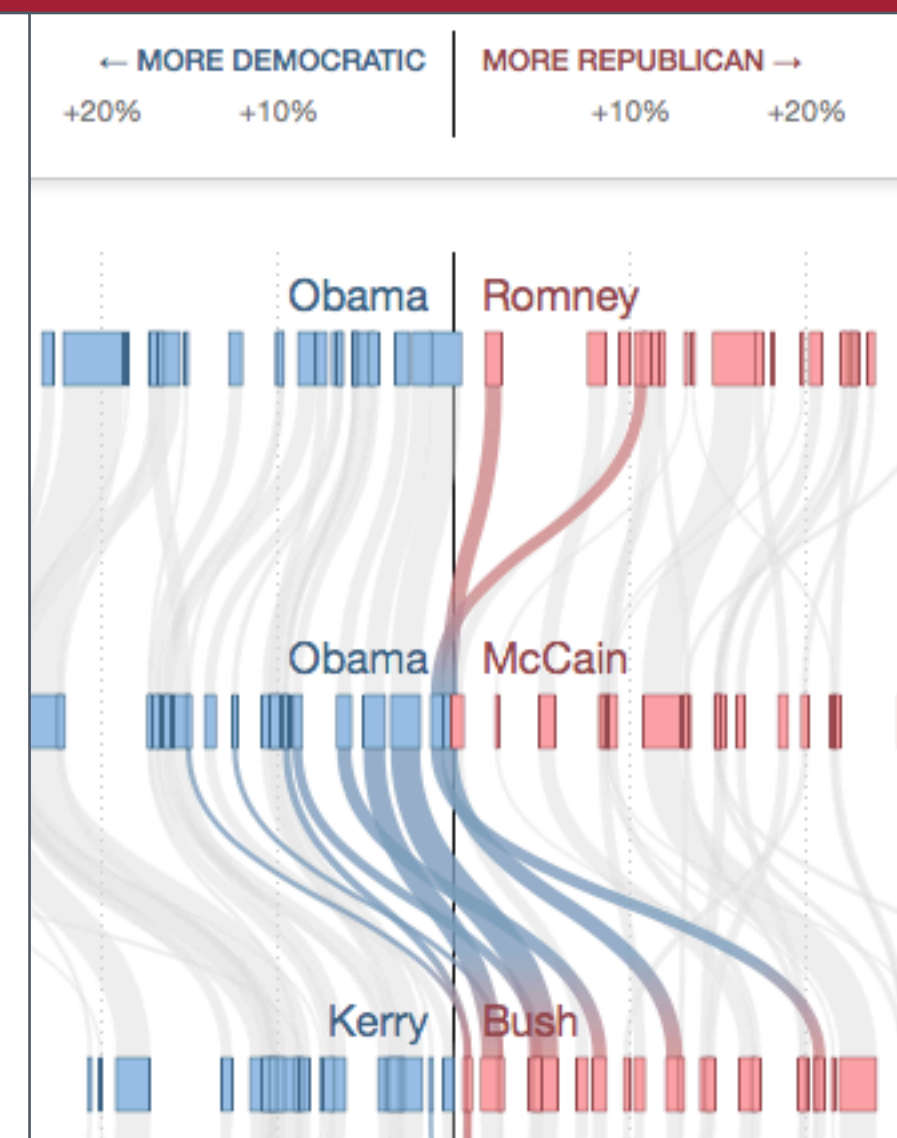
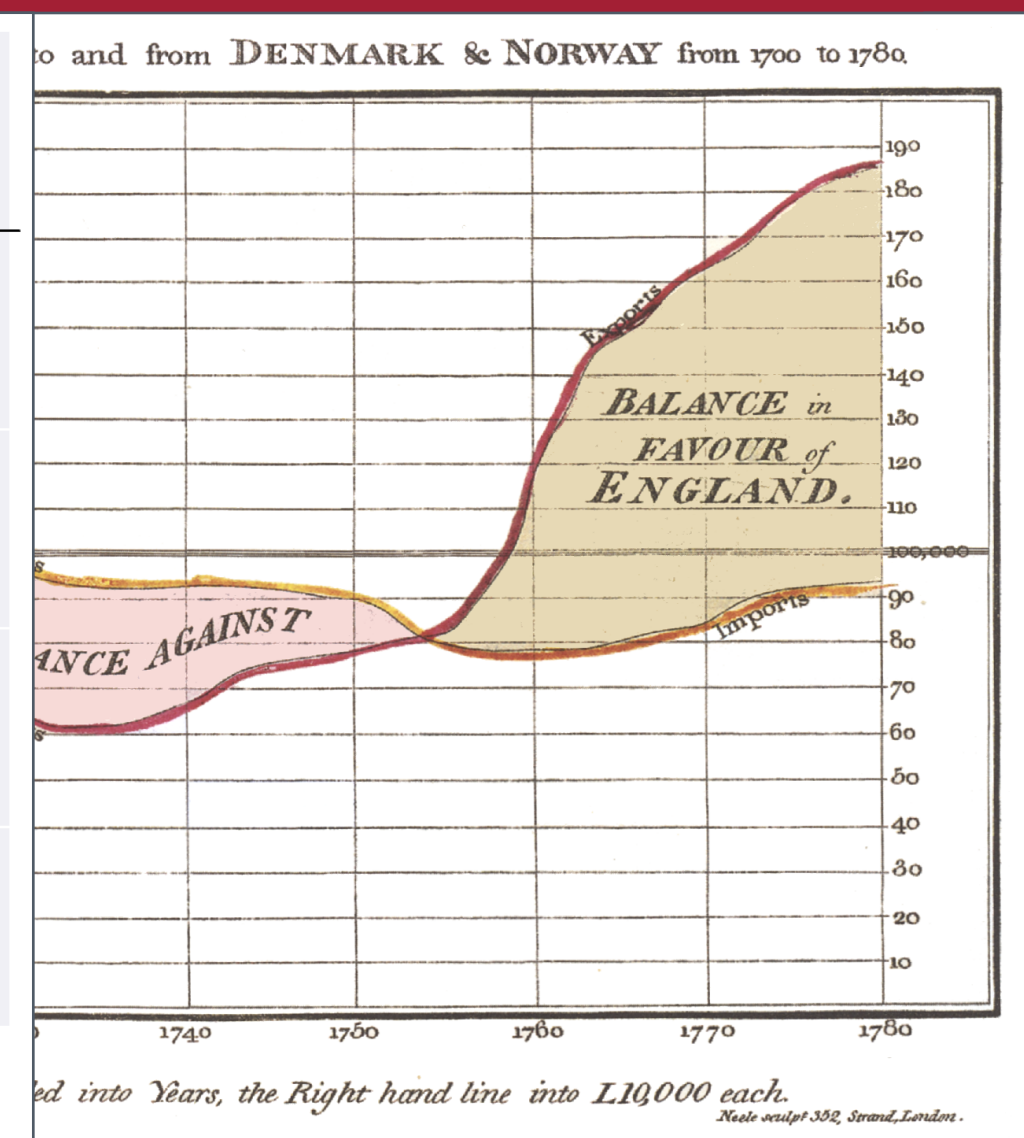
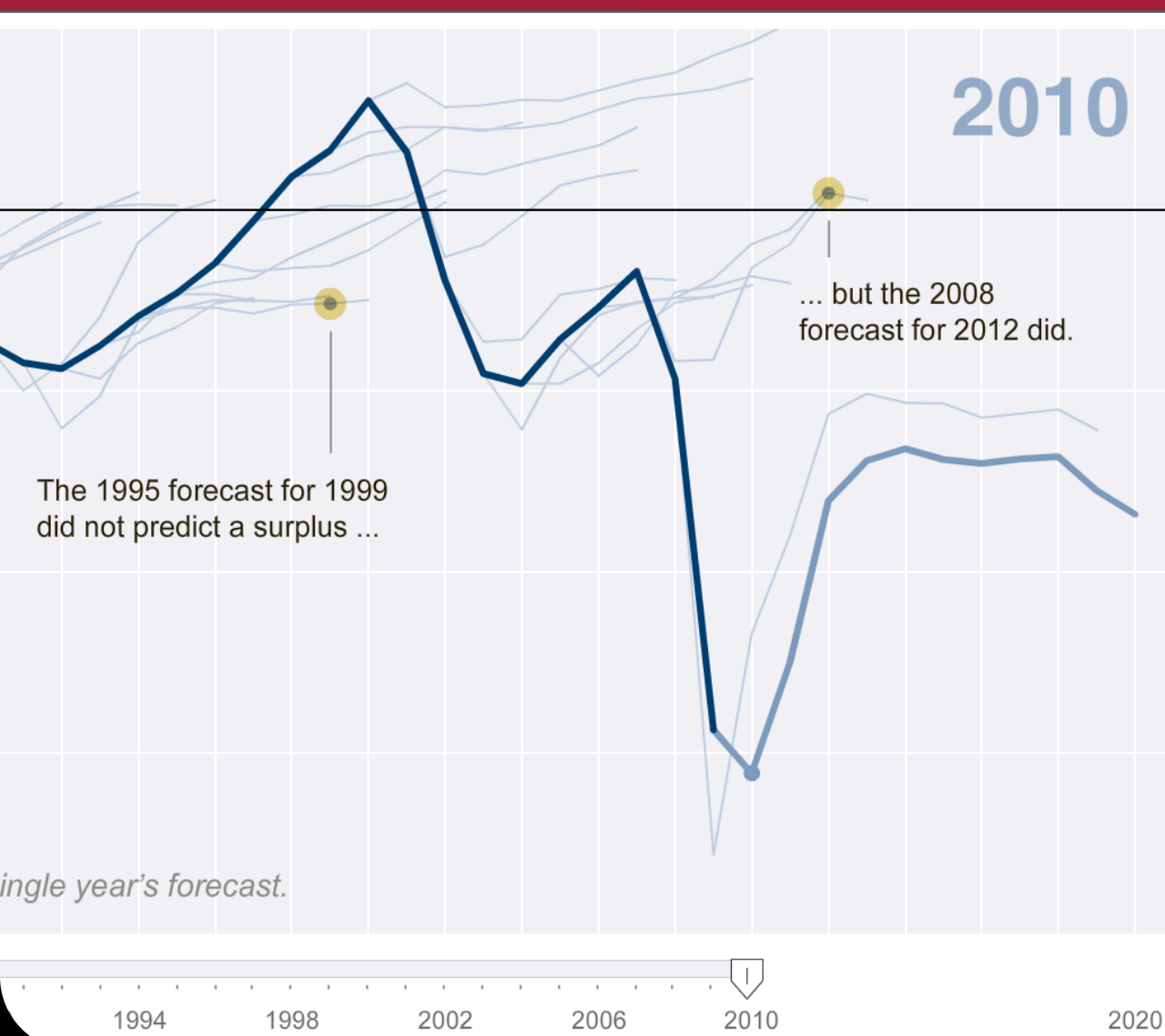


6.859: Interactive Data Visualization

The Value of Visualization

Arvind Satyanarayan



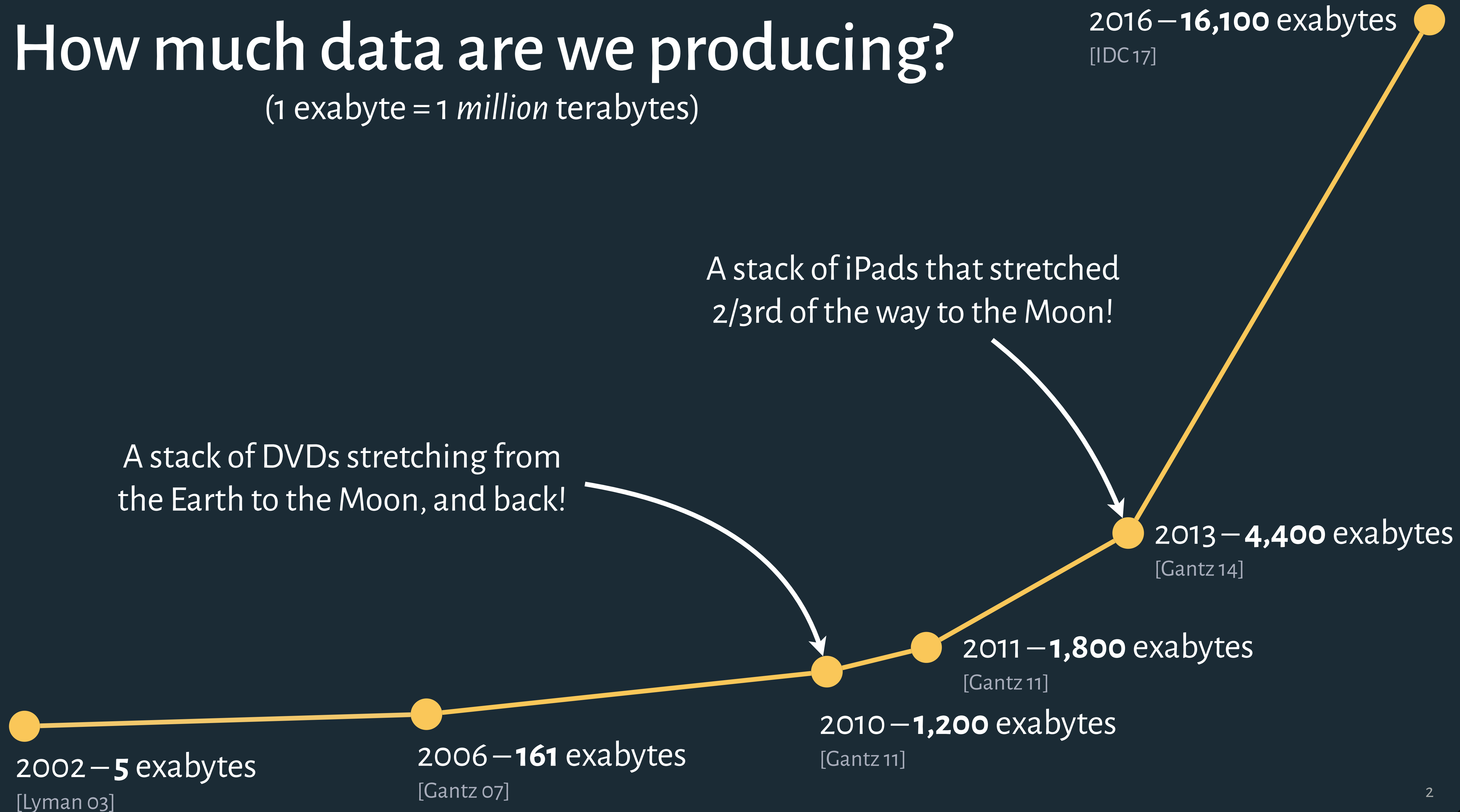
How much data are we producing?

(1 exabyte = 1 *million* terabytes)

2016 – **16,100** exabytes
[IDC 17]

A stack of iPads that stretched
2/3rd of the way to the Moon!

A stack of DVDs stretching from
the Earth to the Moon, and back!

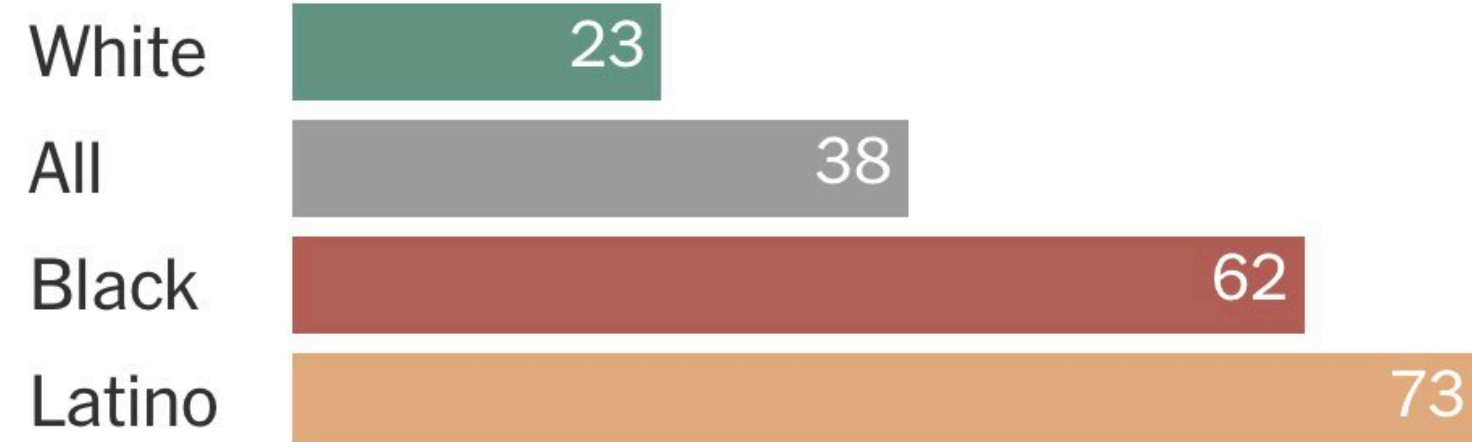


Health & Medicine

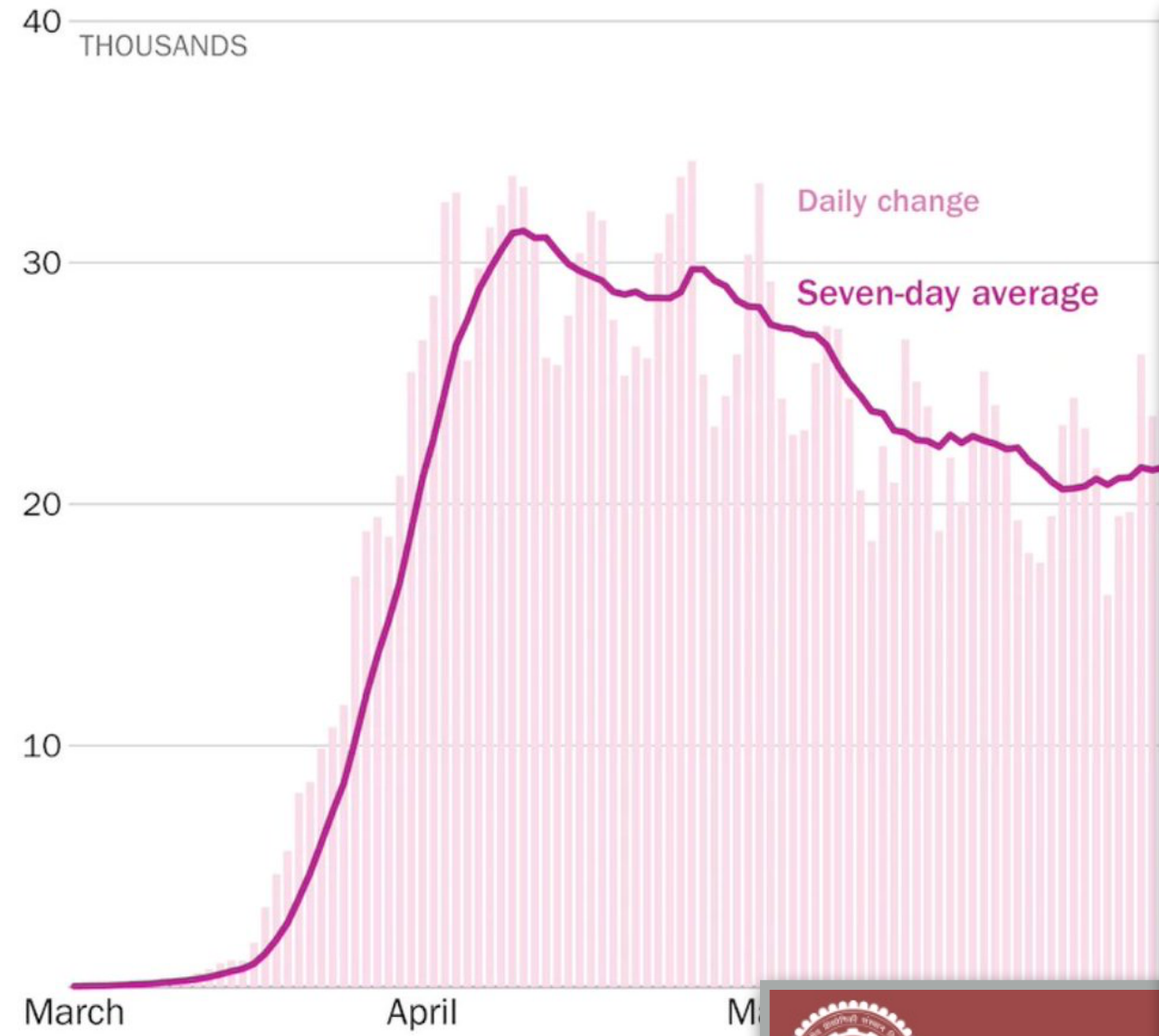


COVID-19

Coronavirus cases per 10,000 people

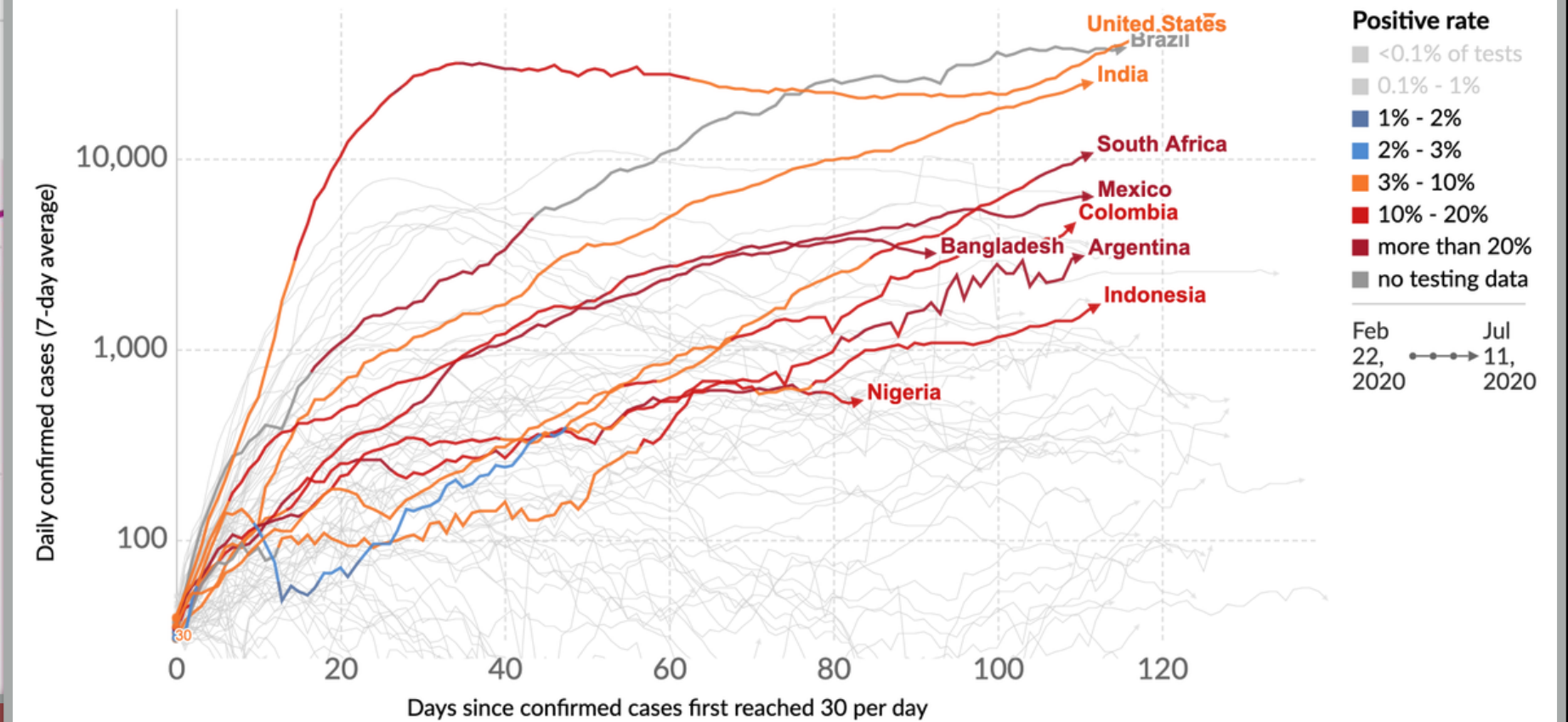


New confirmed coronavirus cases by day



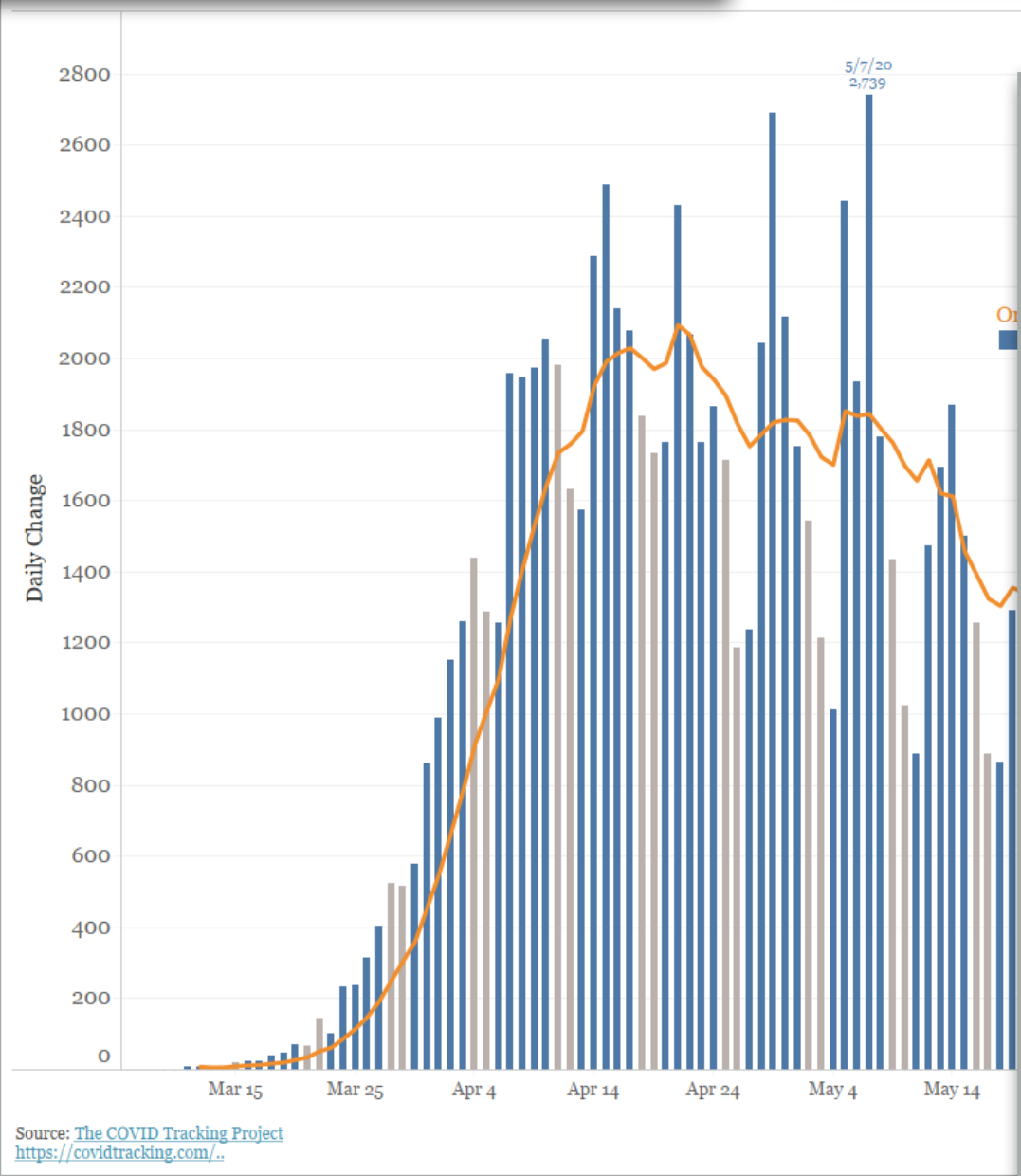
Daily new confirmed cases of COVID-19

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.

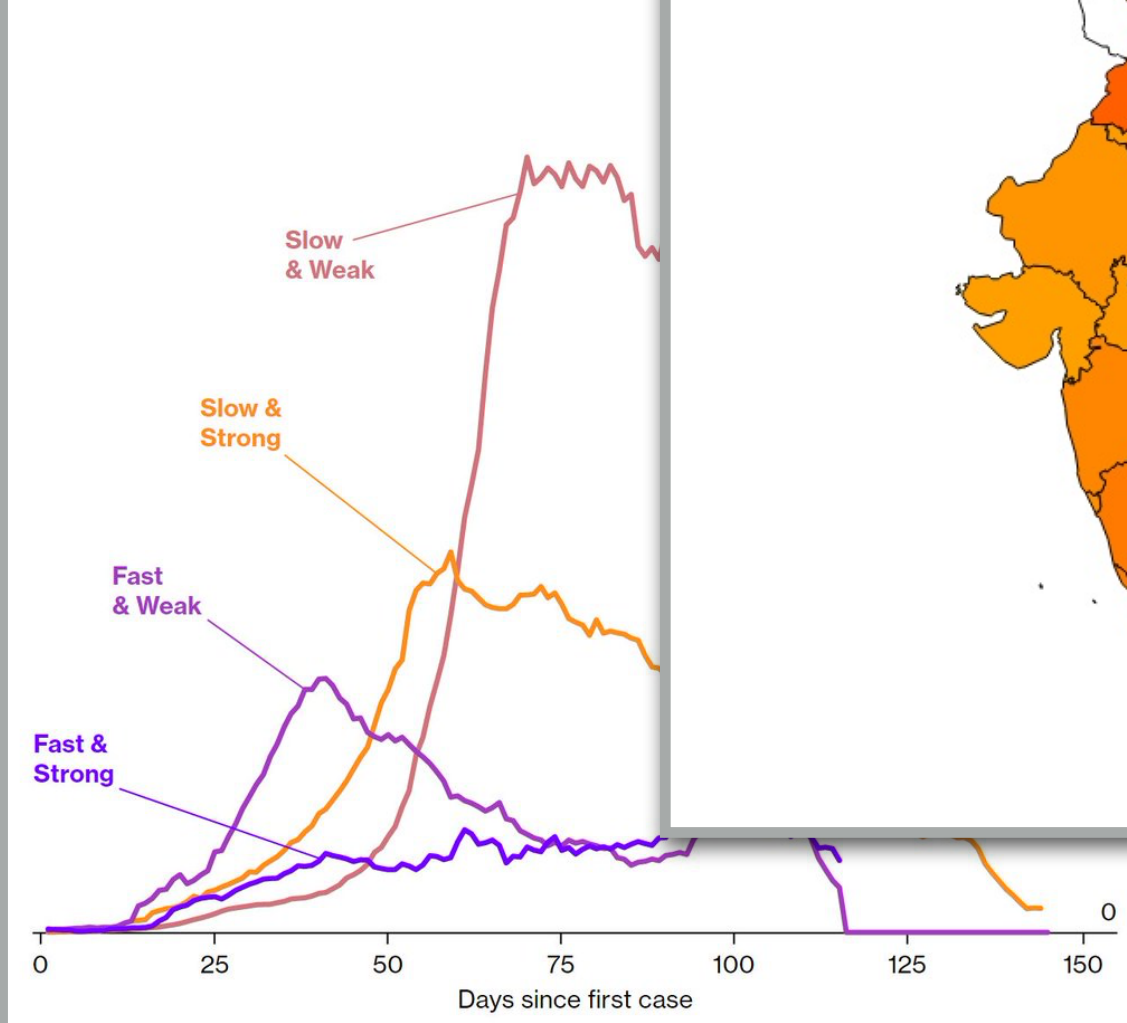


Source: European CDC – Situation Update Worldwide – Last updated 11 July, 10:47 (London time), Official data collated by Our World in Data
 Note: Only countries for which testing data is available are included. Details about this data can be found at [OurWorldInData.org/coronavirus-tests](https://ourworldindata.org/coronavirus-tests)
[OurWorldInData.org/coronavirus-tests](https://ourworldindata.org/coronavirus-tests)

19 Deaths - United States

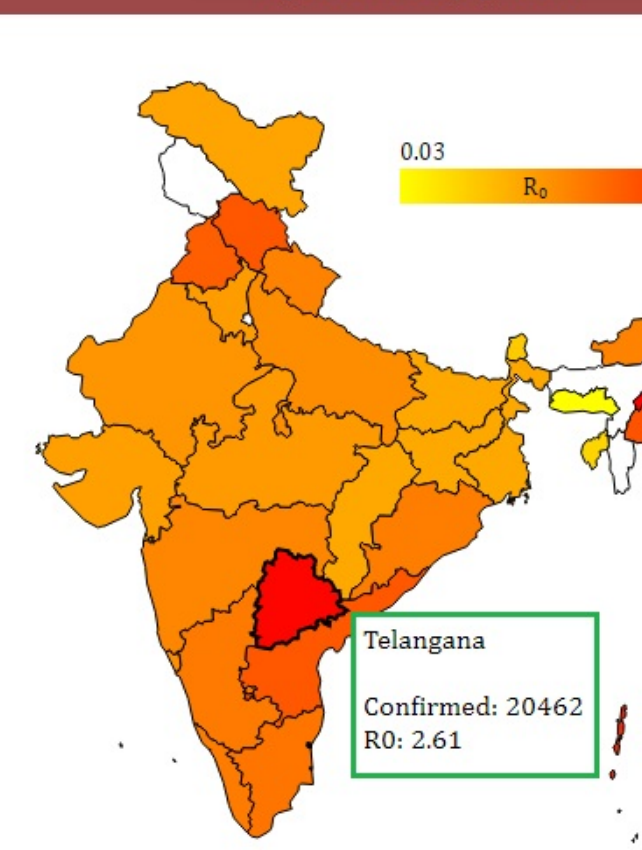


Every Day Countries that reacted slower had high regardless of how tough their



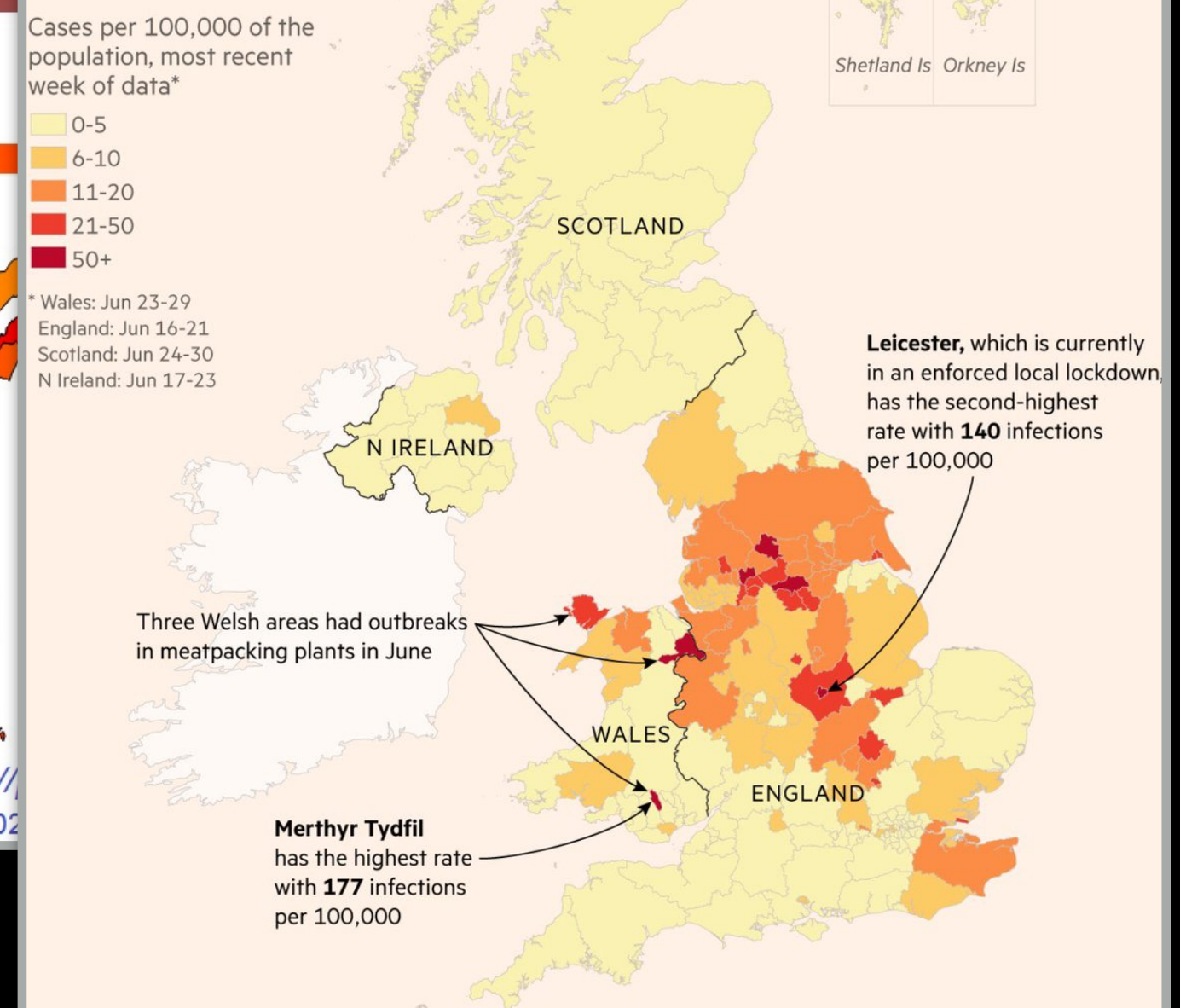
Note: Data as of June 14. Deaths are shown as a seven-day rolling average.
 Sources: Johns Hopkins University Center for Systems Science and Engineering, Oxford COVID-19 Government Response Tracker, Blavatnik School of Government, analysis by Bloomberg

PRediction



Source: <http://...>
 As on 4th July, 2020

Central and northern England report high rates of Covid-19 infections

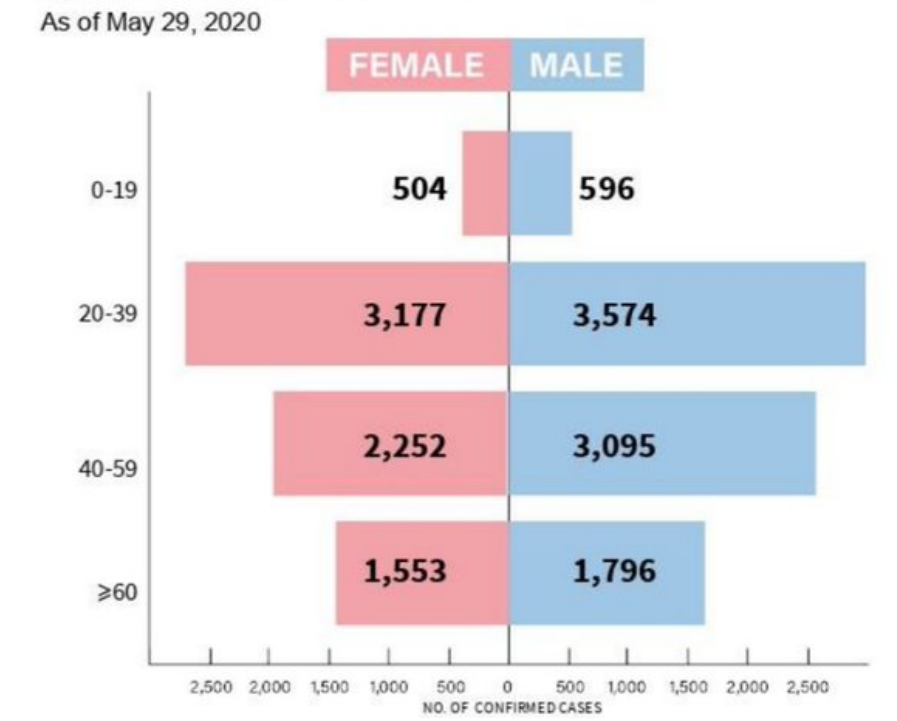


Source: Public Health England; Scottish Government; Public Health Wales; Northern Irish Department of Health
 Cartography: Steven Bernard / @sdbernard
 © FT

COVID-19 PHILIPPINE SITUATIONER #033 (MAY 30, 2020)

Goal: Shortening Duration of Infection

Age and Sex Distribution of confirmed cases



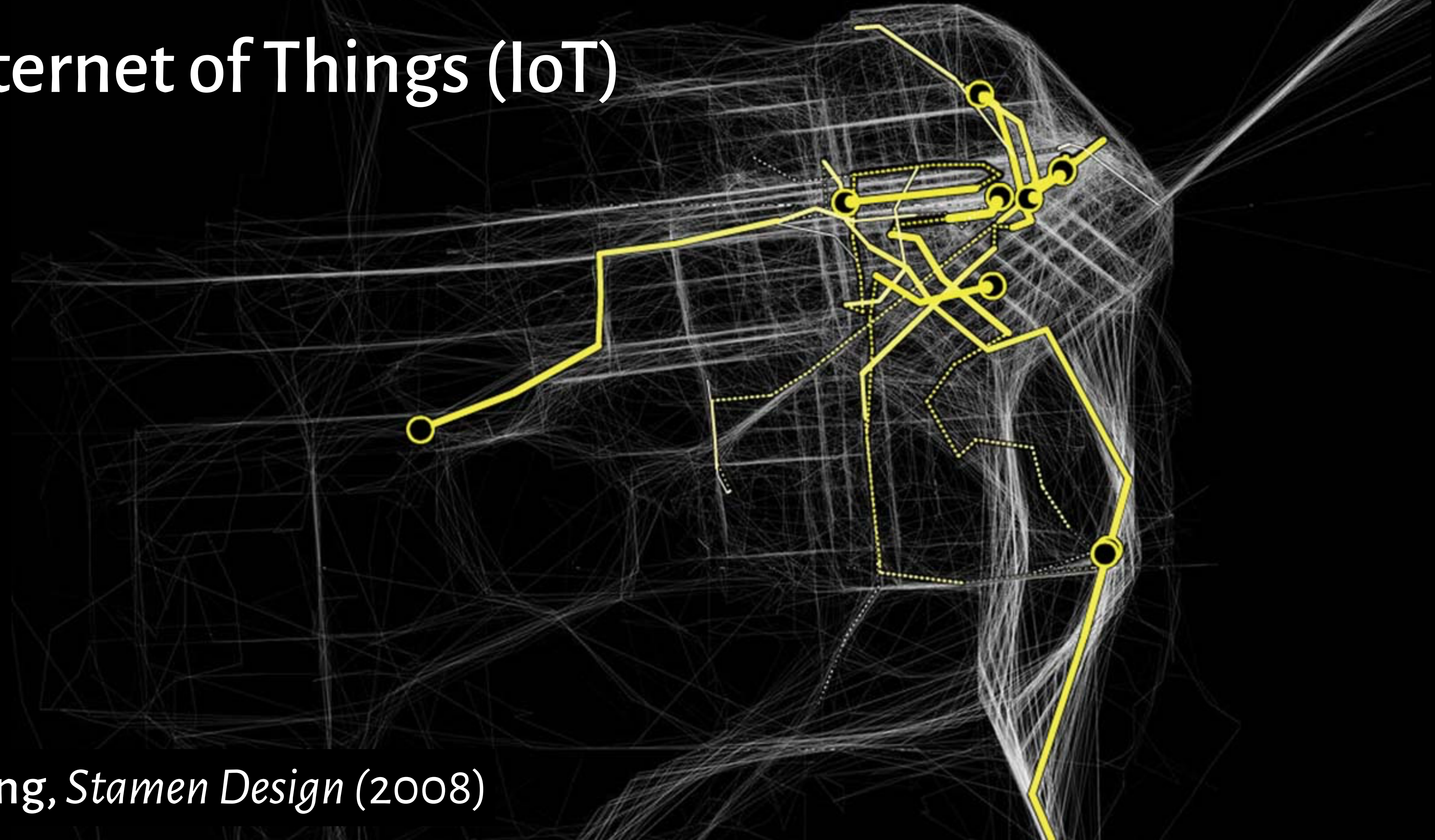
Confirmed COVID-19 Cases by Date of Onset of Illness (N=16,634)

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.



Physical Sensors + The Internet of Things (IoT)



Cabspotting, *Stamen Design* (2008)

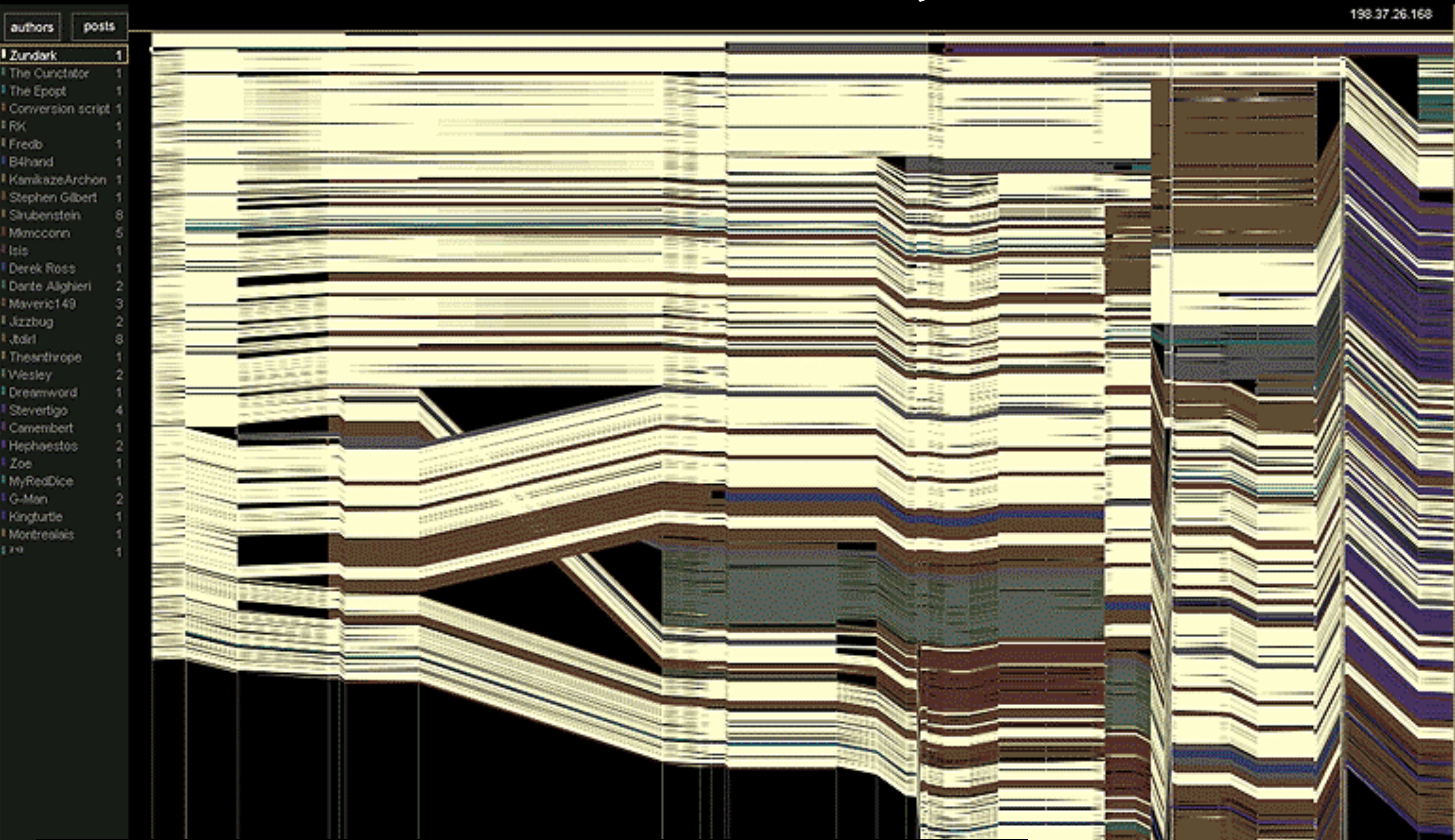
Records of Human Activity



facebook

December 2010

Records of Human Activity



Abortion
(Revision as of 22:56 4 Jun 2003)

"Abortion," in its most commonly used sense, refers to the deliberate early termination of pregnancy, resulting in the death of the embryo or fetus. [1] Medically, the term also refers to the early termination of a pregnancy by natural causes ("spontaneous abortion" or miscarriage, which ends 1 in 5 of all pregnancies, usually within the first 13 weeks) or to the cessation of normal growth of a body part or organ. What follows is a discussion of the issues related to deliberate or "induced" abortion.

Methods

Depending on the stage of pregnancy an abortion is performed by a number of different methods. For the earliest terminations (before nine weeks or so) a chemical abortion is the usual method, the drug **mifepristone** is usually the only legal method although research has uncovered similar effects from **methotrexate** and **misoprostol**. Concurrent with chemical abortion and extending up until around the fifteenth week suction-aspiration or vacuum abortion is the most common approach, replacing the more risky dilation and curettage (D & C). From the fifteenth week up until around the eighteenth week a surgical dilation and evacuation (D & E) is used.

As the fetus size increases other techniques must be used to secure abortion in the third trimester, premature expulsion of the fetus can be induced with prostaglandin, this can be coupled with injecting the amniotic fluid with saline or urea solution. Very late abortions can be brought about by the controversial intact dilation and extraction (D & X) or a hysterotomy abortion, similar to a caesarian section.

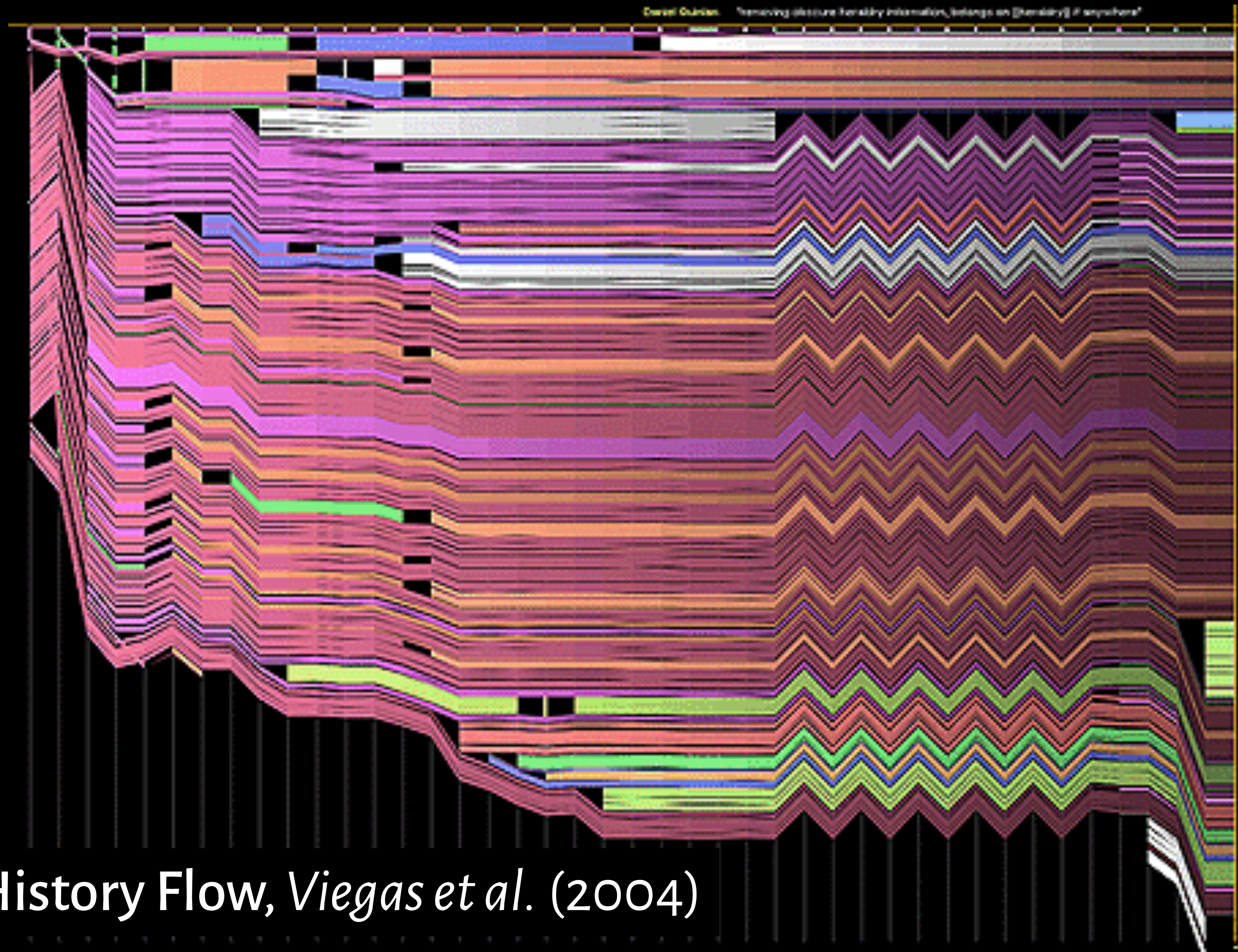
The controversy

The morality and legality of abortion is a large and important topic in applied ethics and is also discussed by legal scholars and religious people. Important facts about abortion are also researched by sociologists and historians.

Abortion has been common in most societies, although it has often been opposed by some institutionalized religions and governments. In 20th century politics in the United States and Europe, abortion became commonly accepted by the end of the 20th century. Additionally, abortion is legal and accepted in China, India and other populous countries. The Catholic Church remains opposed to the procedure, however, and in other countries, notably the United States and the predominantly Catholic Republic of Ireland, the controversy is still

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

Machine Learning!

we really do have essentially free and ubiquitous data. So the complementary 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

Machine Learning?

we really do have essentially free and ubiquitous data. So the complementary 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

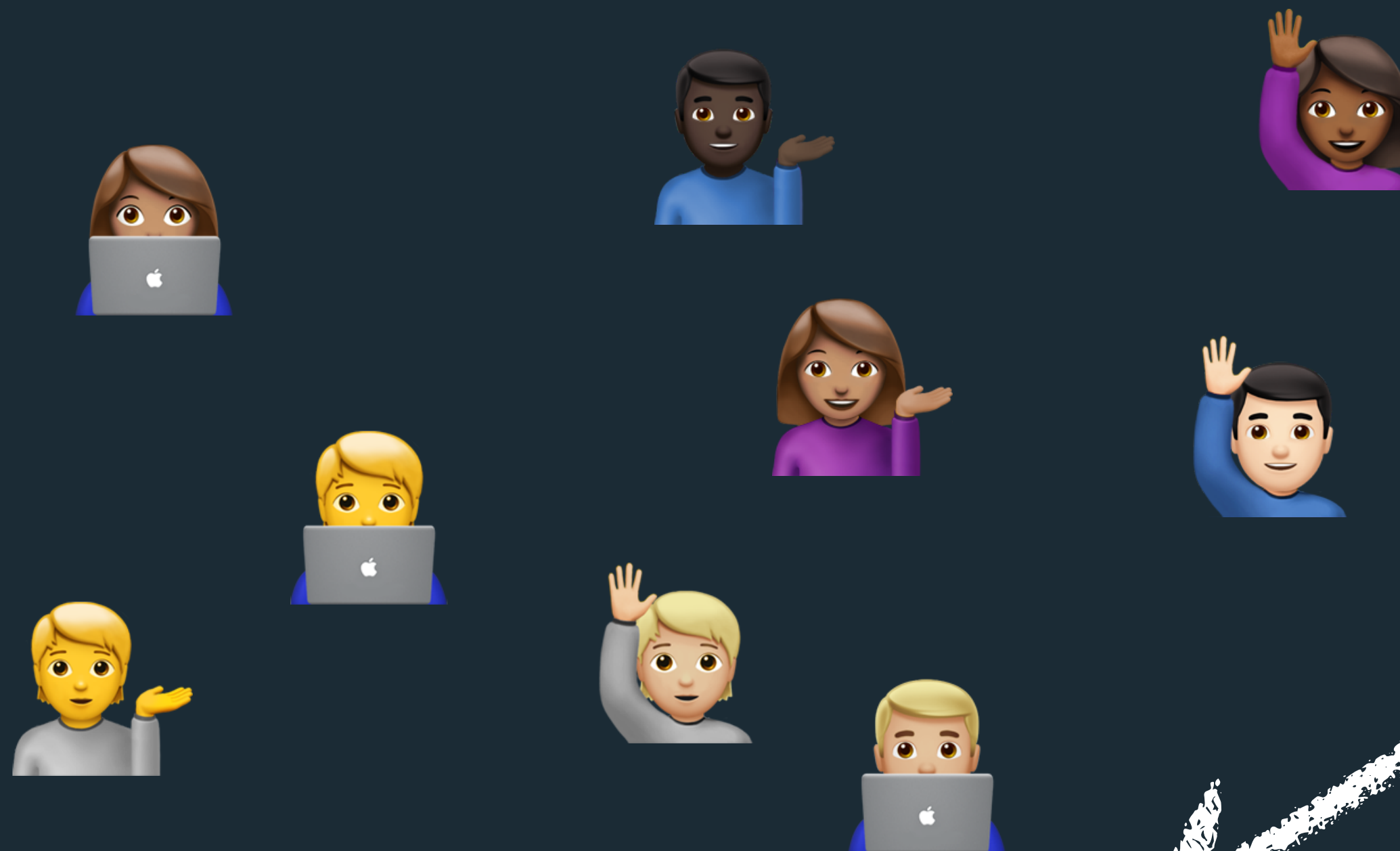
Machine Learning?

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?



Post in the chat



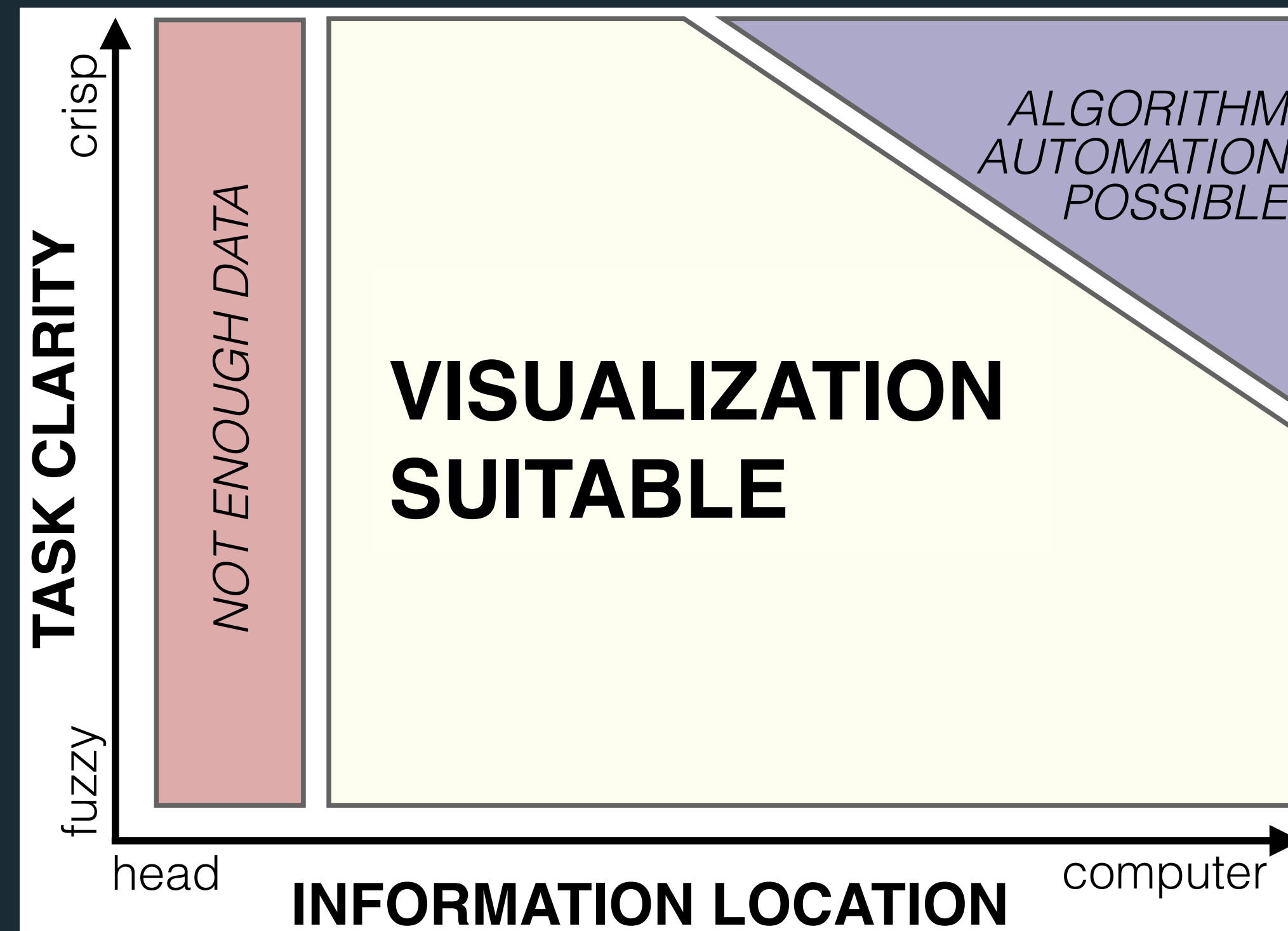
Raise your hand



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

X	Y
10	8.04
8	6.95
13	7.58
9	8.81
11	8.33
14	9.96
6	7.24
4	4.26
12	10.84
7	4.82
5	5.68

Set B

X	Y
10	9.14
8	8.14
13	8.74
9	8.77
11	9.26
14	8.1
6	6.13
4	3.1
12	9.11
7	7.26
5	4.74

Set C

X	Y
10	7.46
8	6.77
13	12.74
9	7.11
11	7.81
14	8.84
6	6.08
4	5.39
12	8.15
7	6.42
5	5.73

Set D

X	Y
8	6.58
8	5.76
8	7.71
8	8.84
8	8.47
8	7.04
8	5.25
19	12.5
8	5.56
8	7.91
8	6.89

Summary Statistics

$$\mu_X = 9.0 \quad \sigma_X = 3.317$$

$$\mu_Y = 7.5 \quad \sigma_Y = 2.03$$

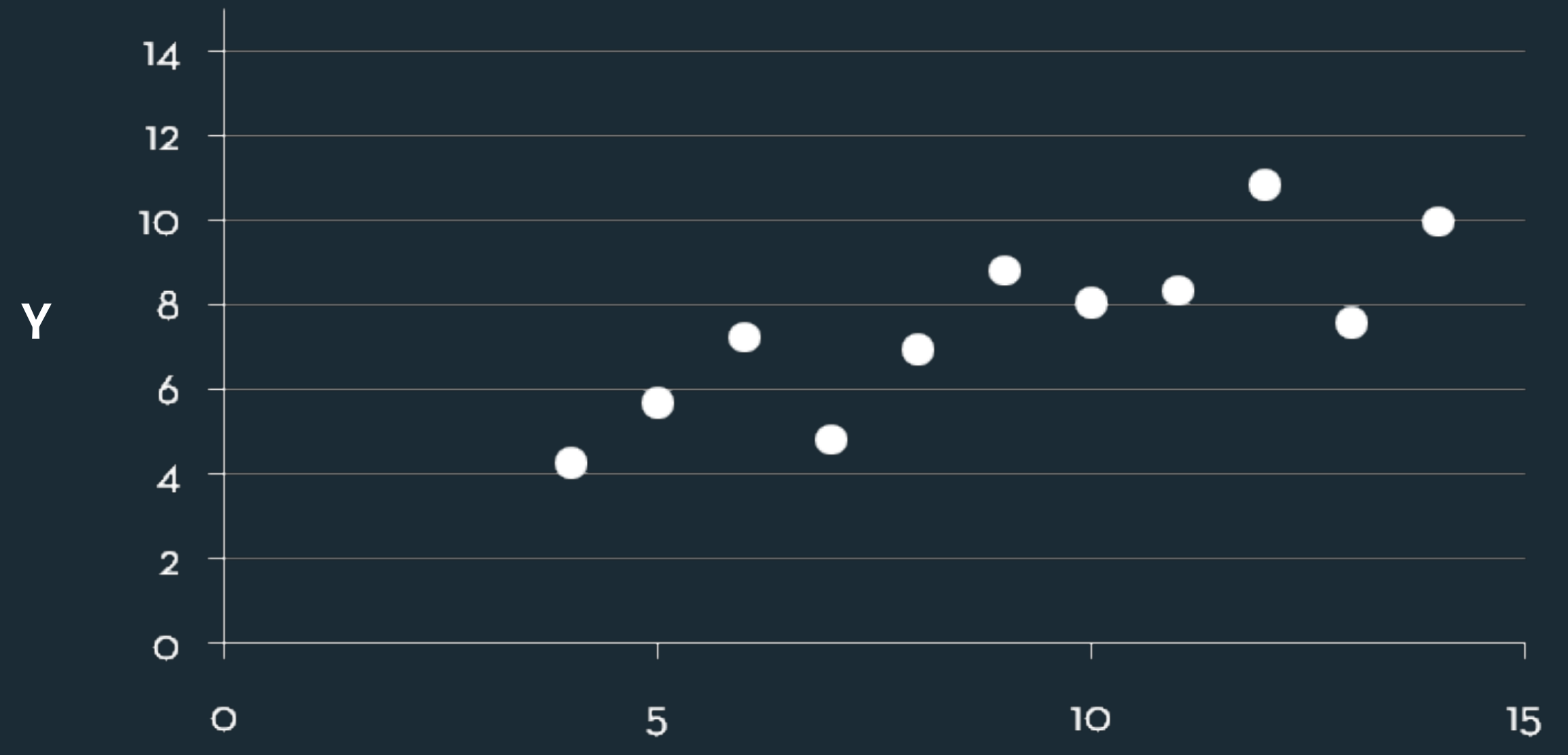
Linear Regression

$$Y = 3 + 0.5 X$$

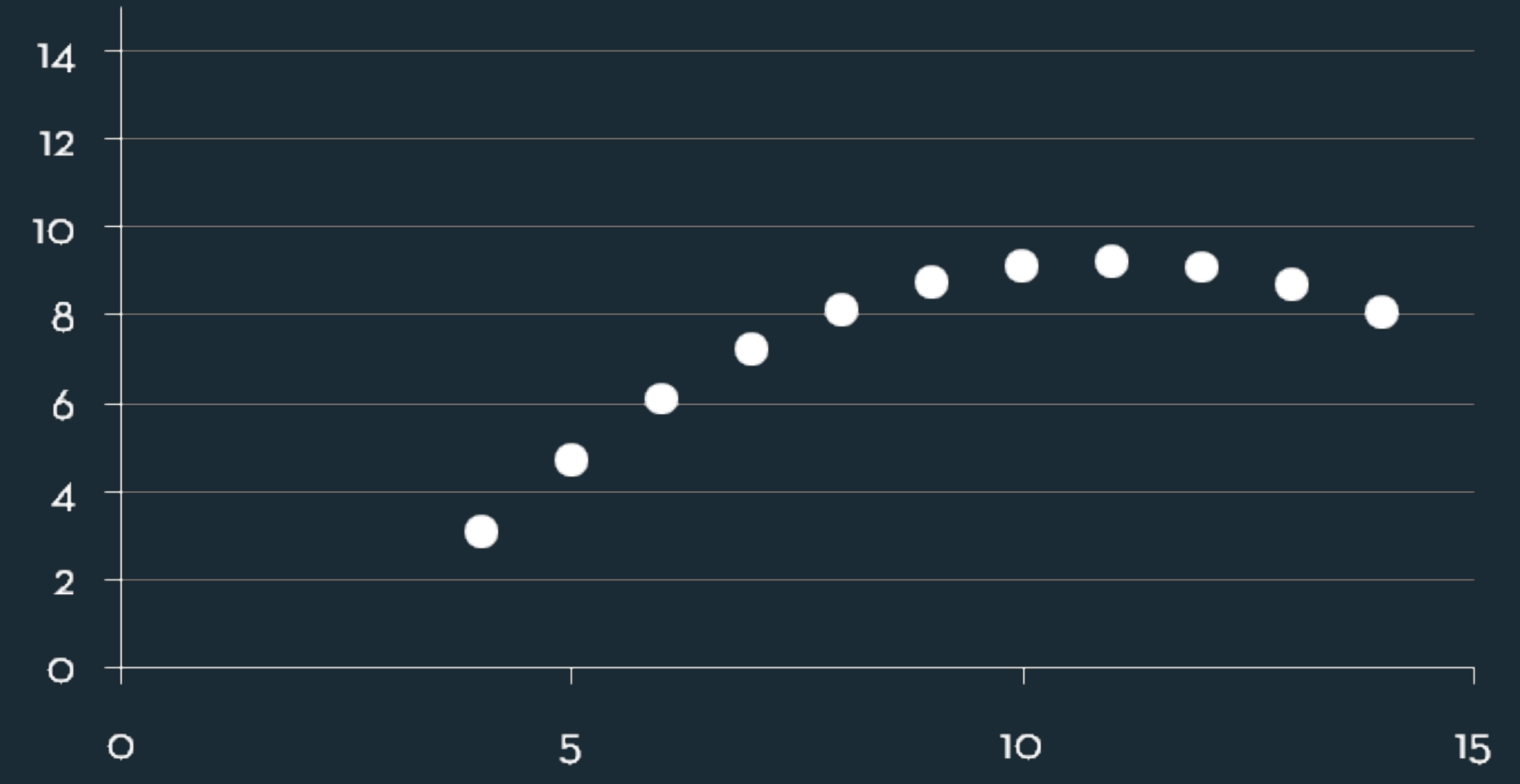
$$R^2 = 0.67$$

[Anscombe 1973]

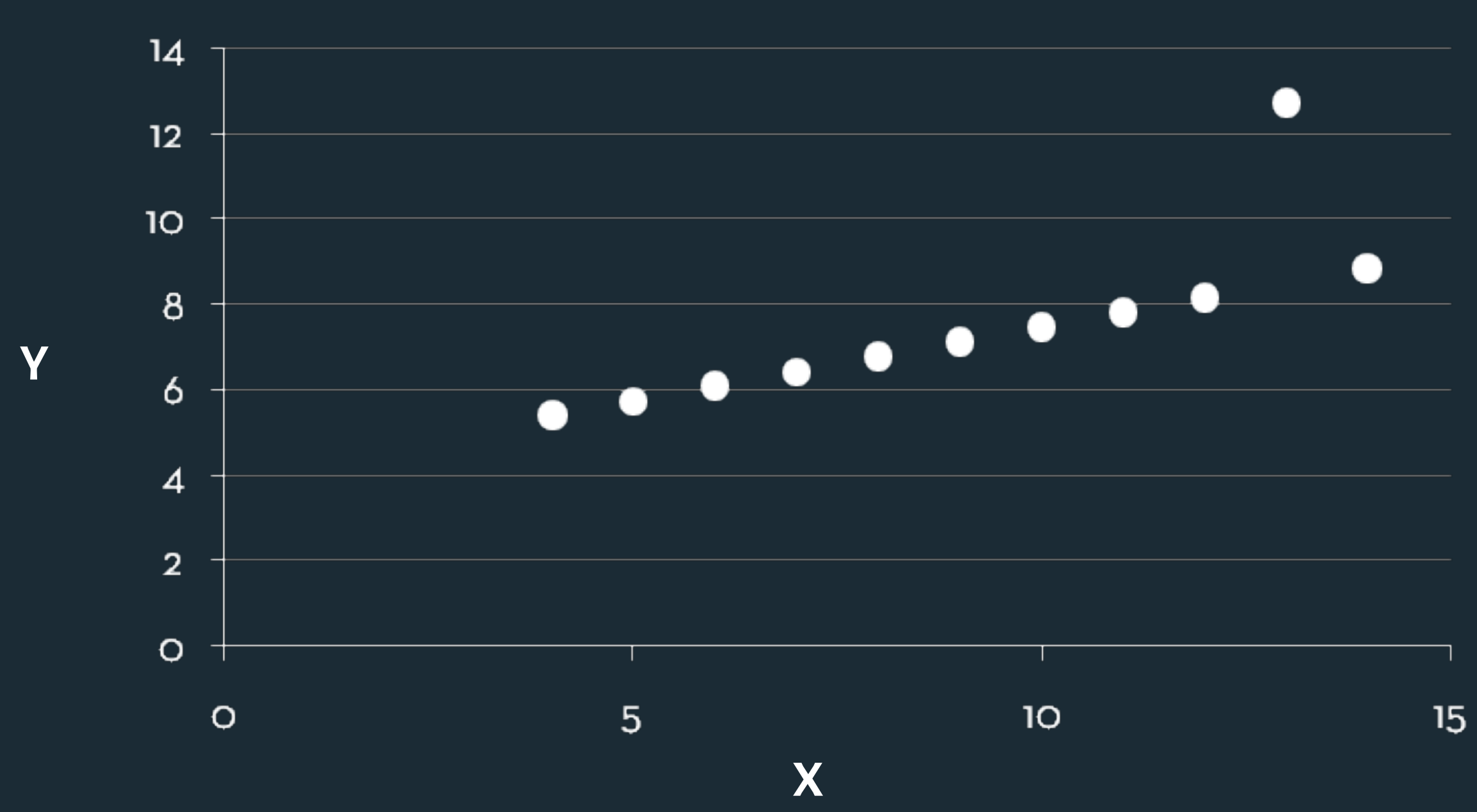
Set A



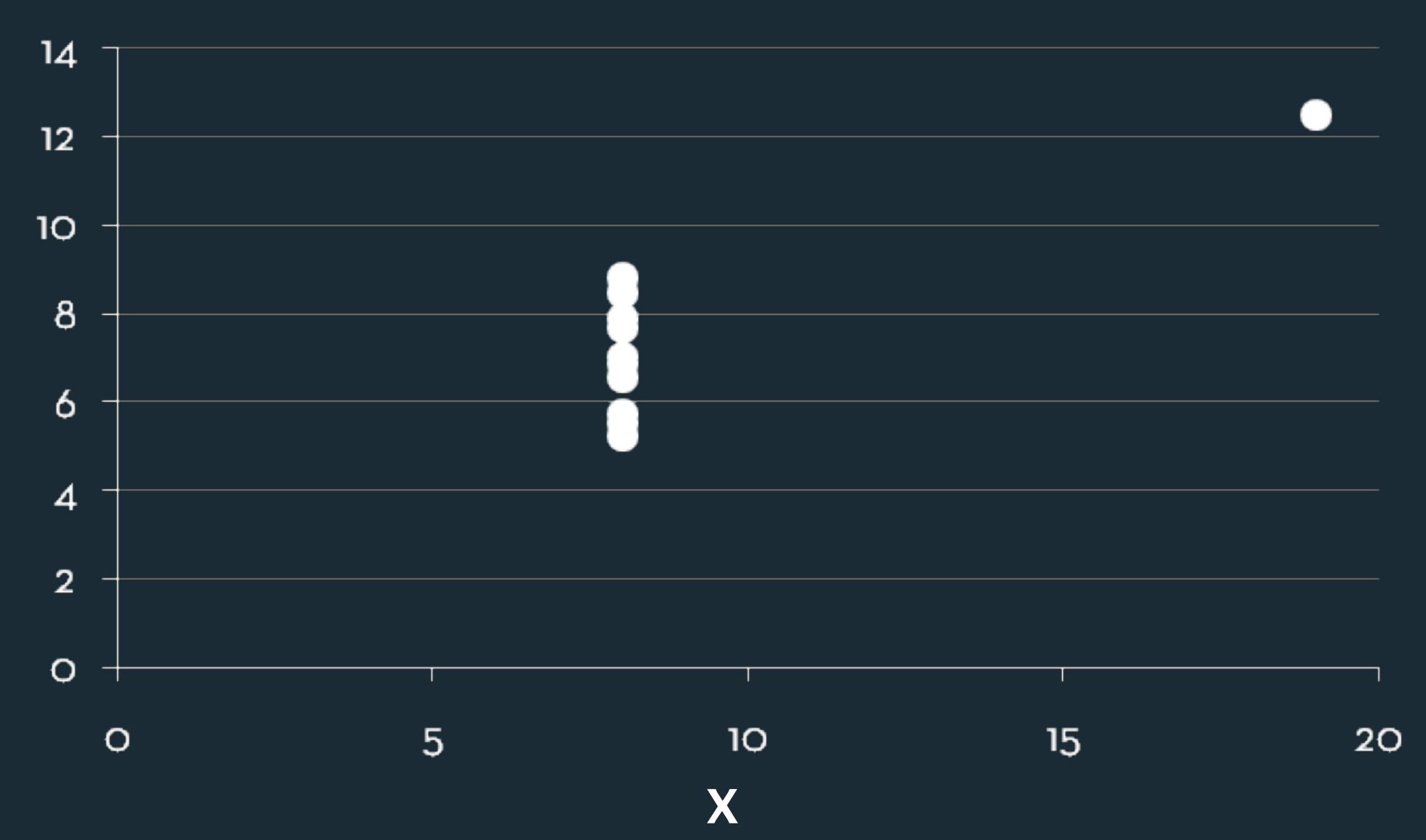
Set B



Set C



Set D





Alberto Cairo ✓

@albertocairo

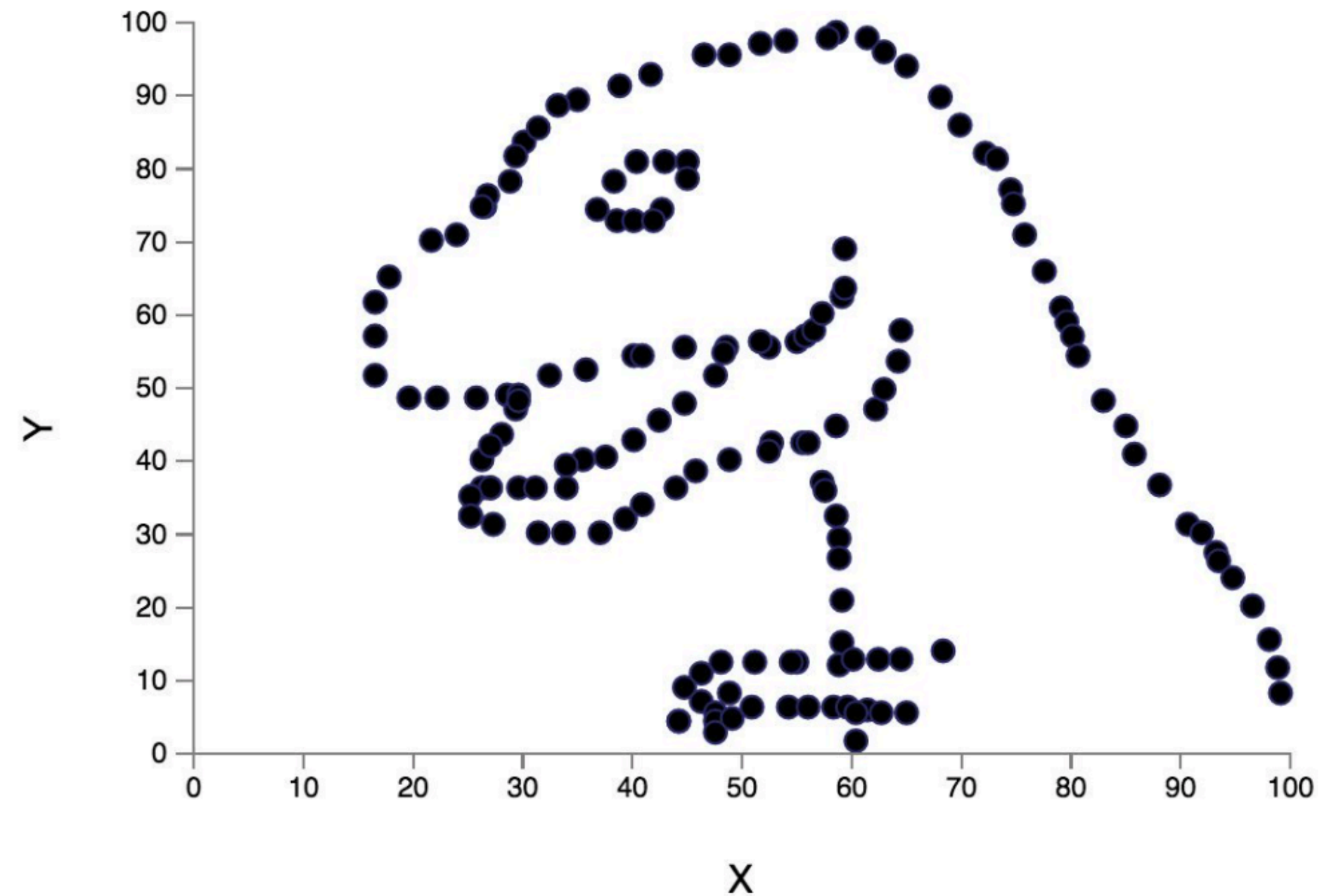
Following



Don't trust summary statistics. Always visualize your data first

robertgrantstats.co.uk/drawmydata.html

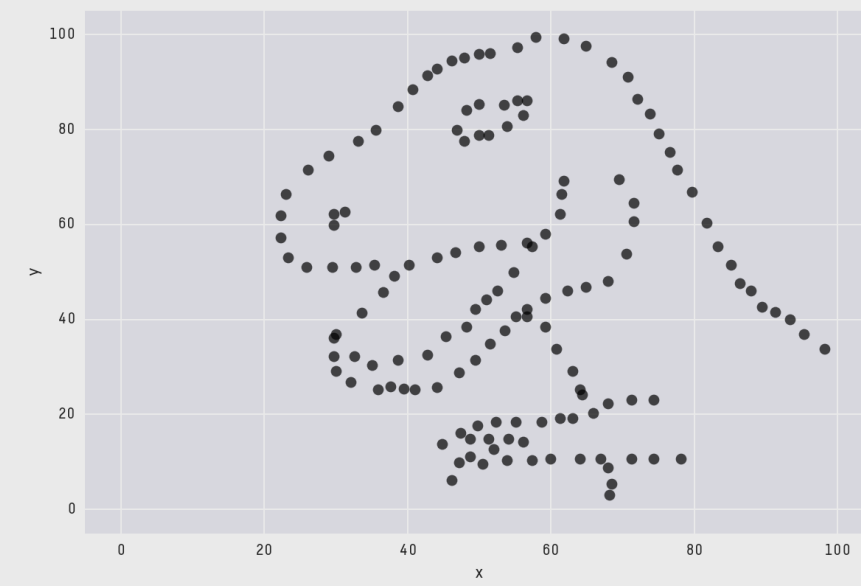
N = 157 ; X mean = 50.7333 ; X SD = 19.5661 ; Y mean = 46.495 ; Y SD = 27.2828 ;
Pearson correlation = -0.1772



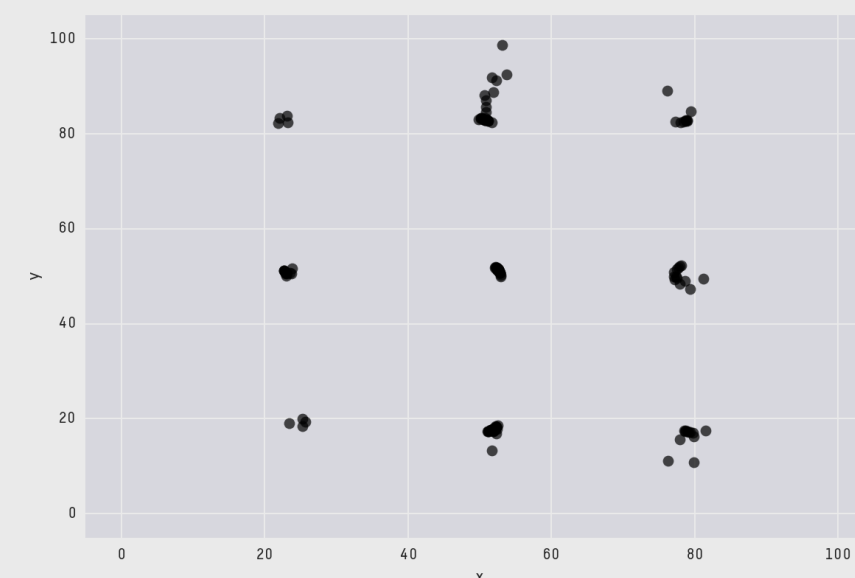
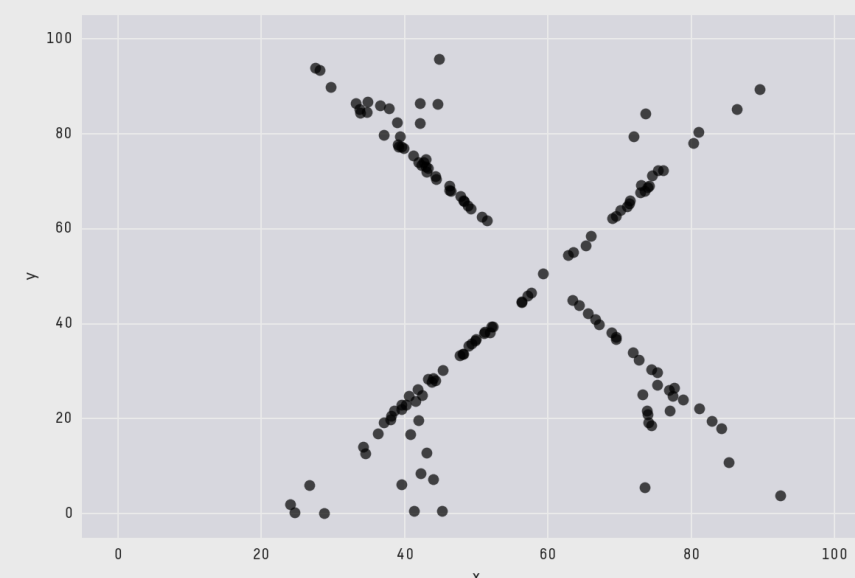
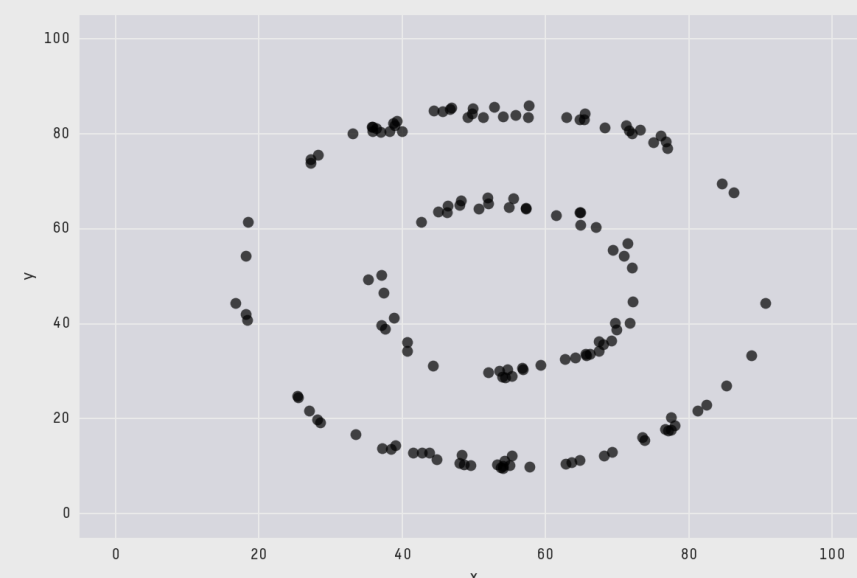
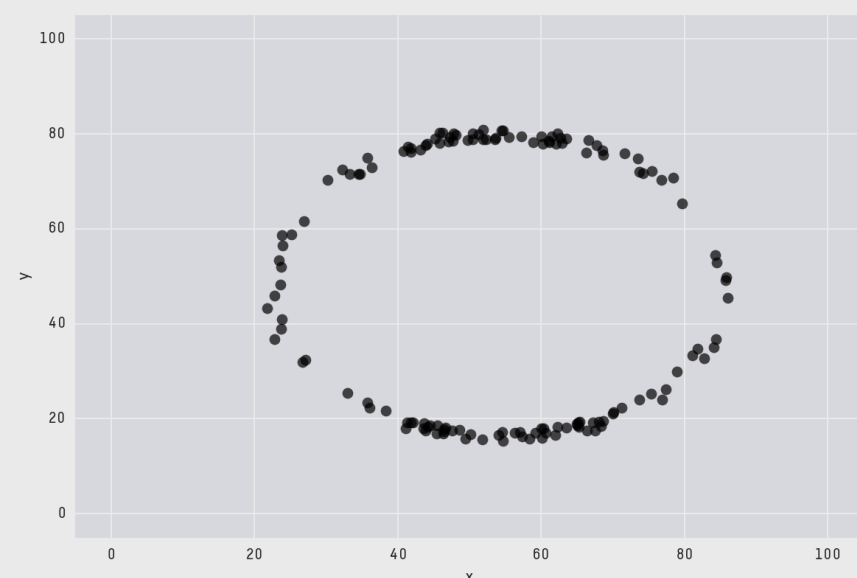
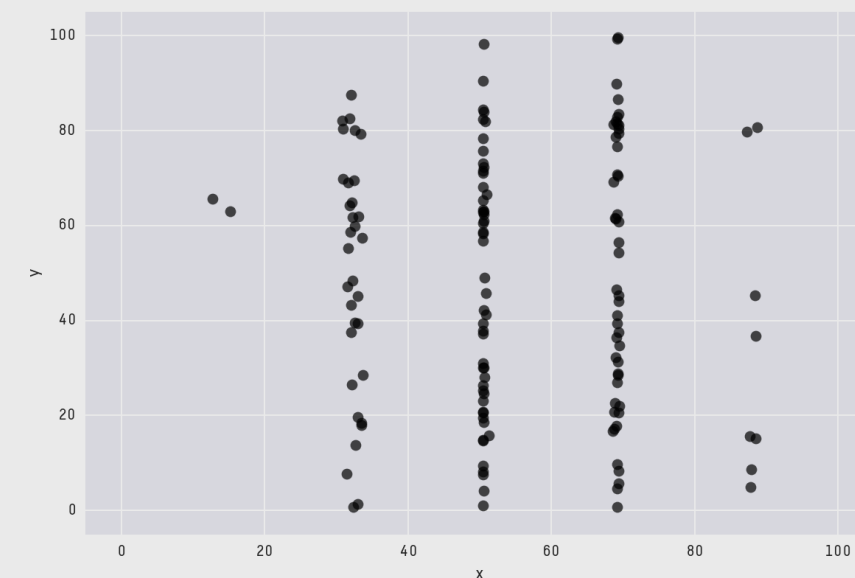
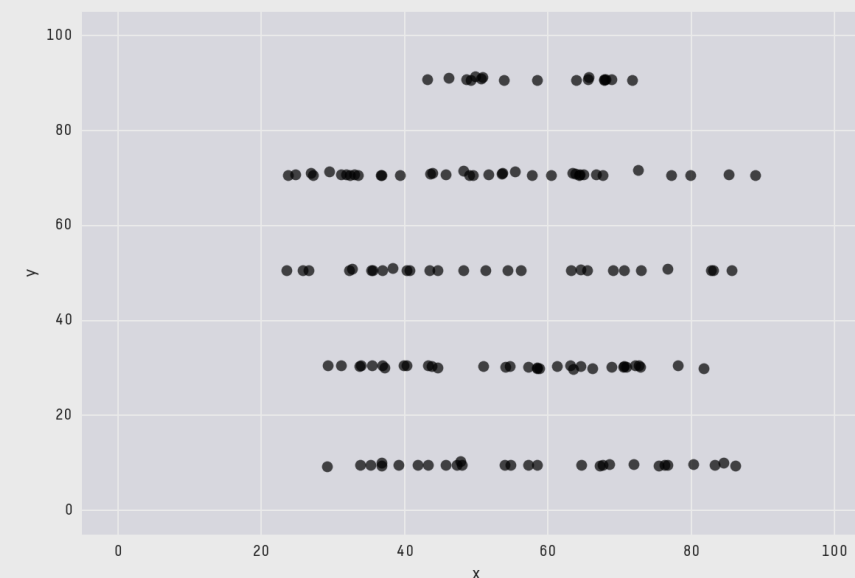
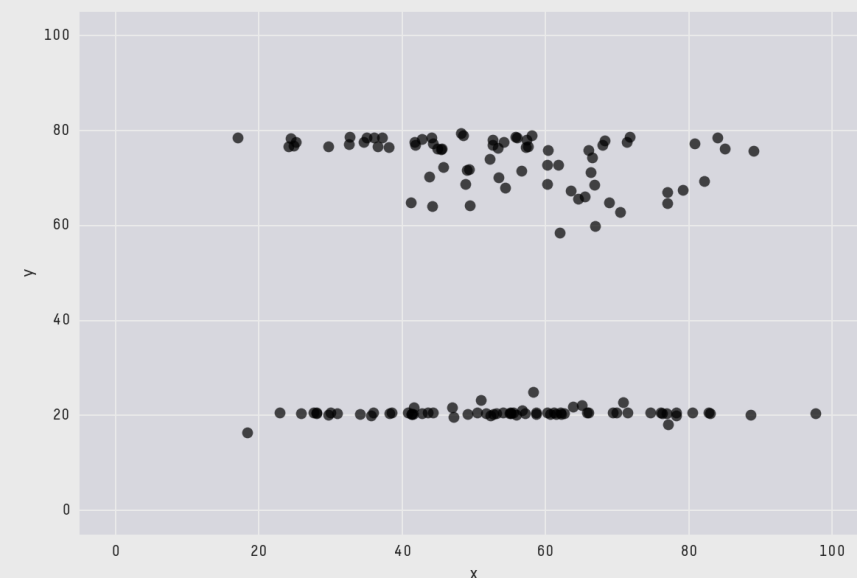
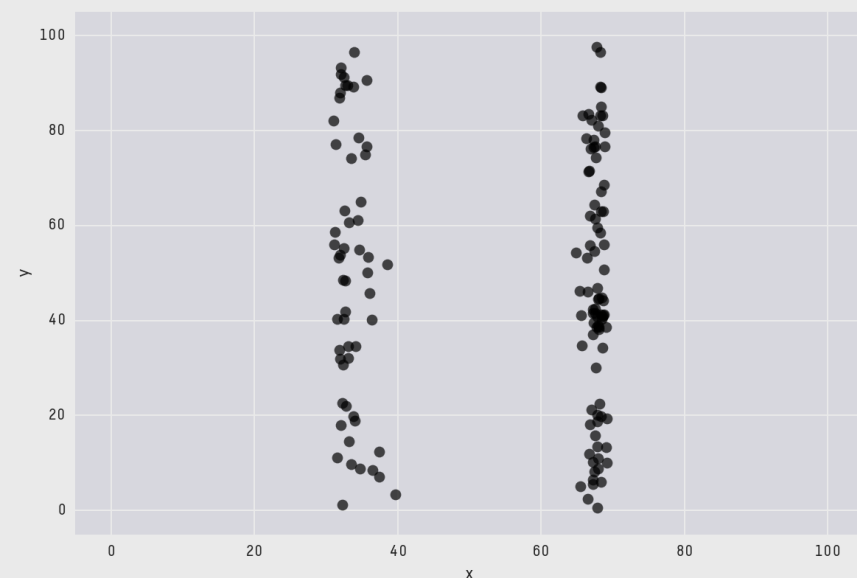
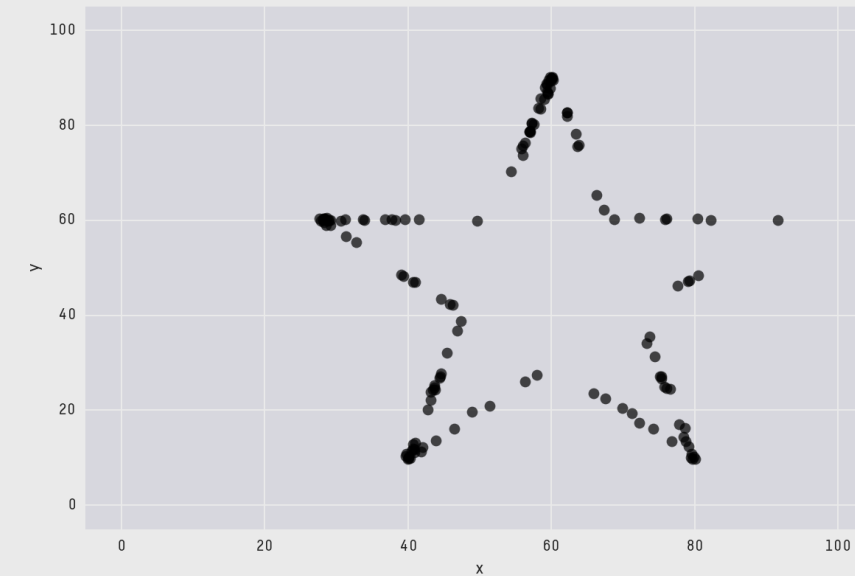
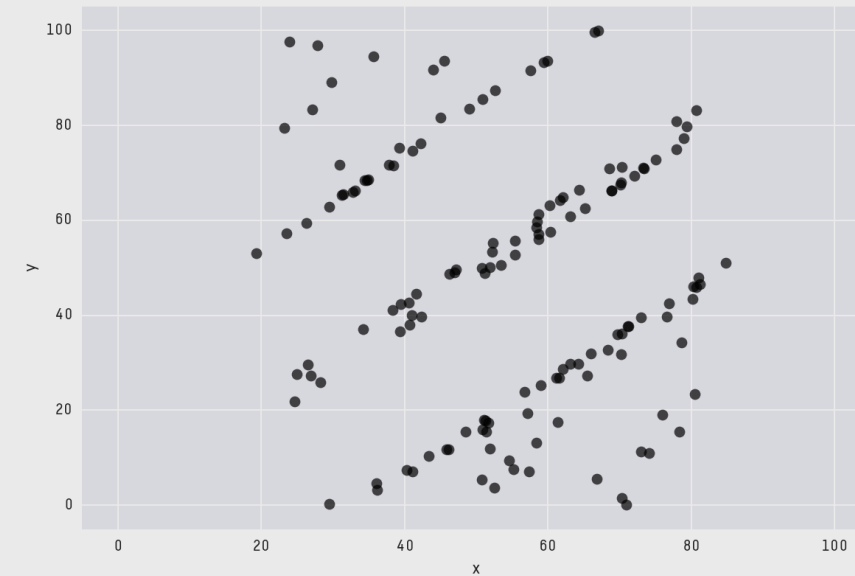
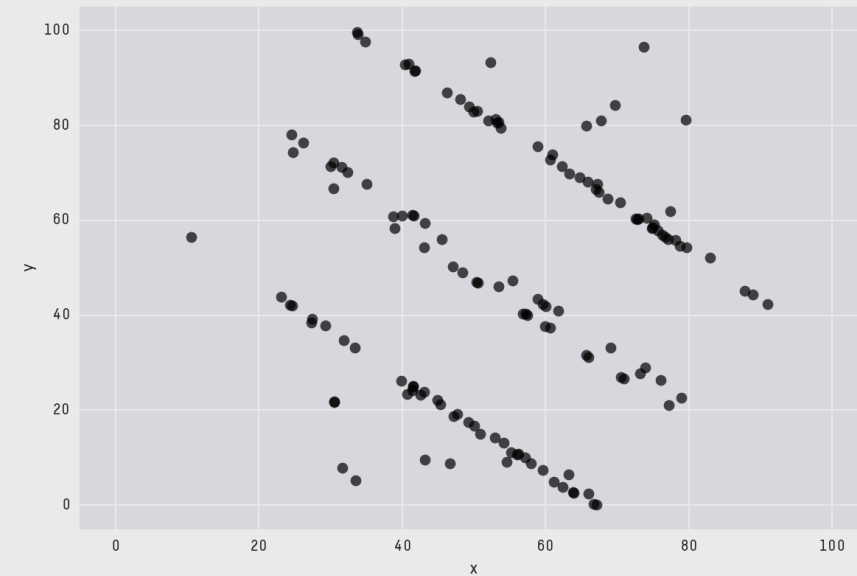
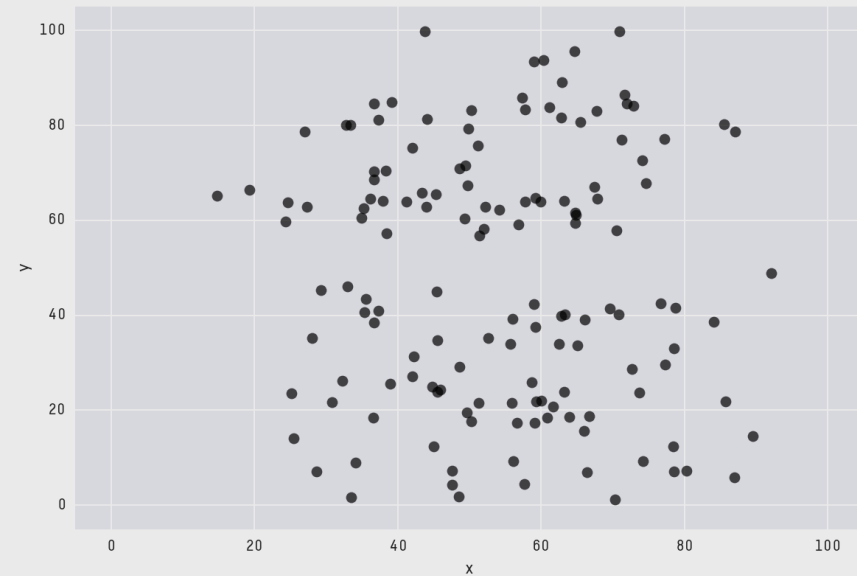
5:47 AM - 15 Aug 2016

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

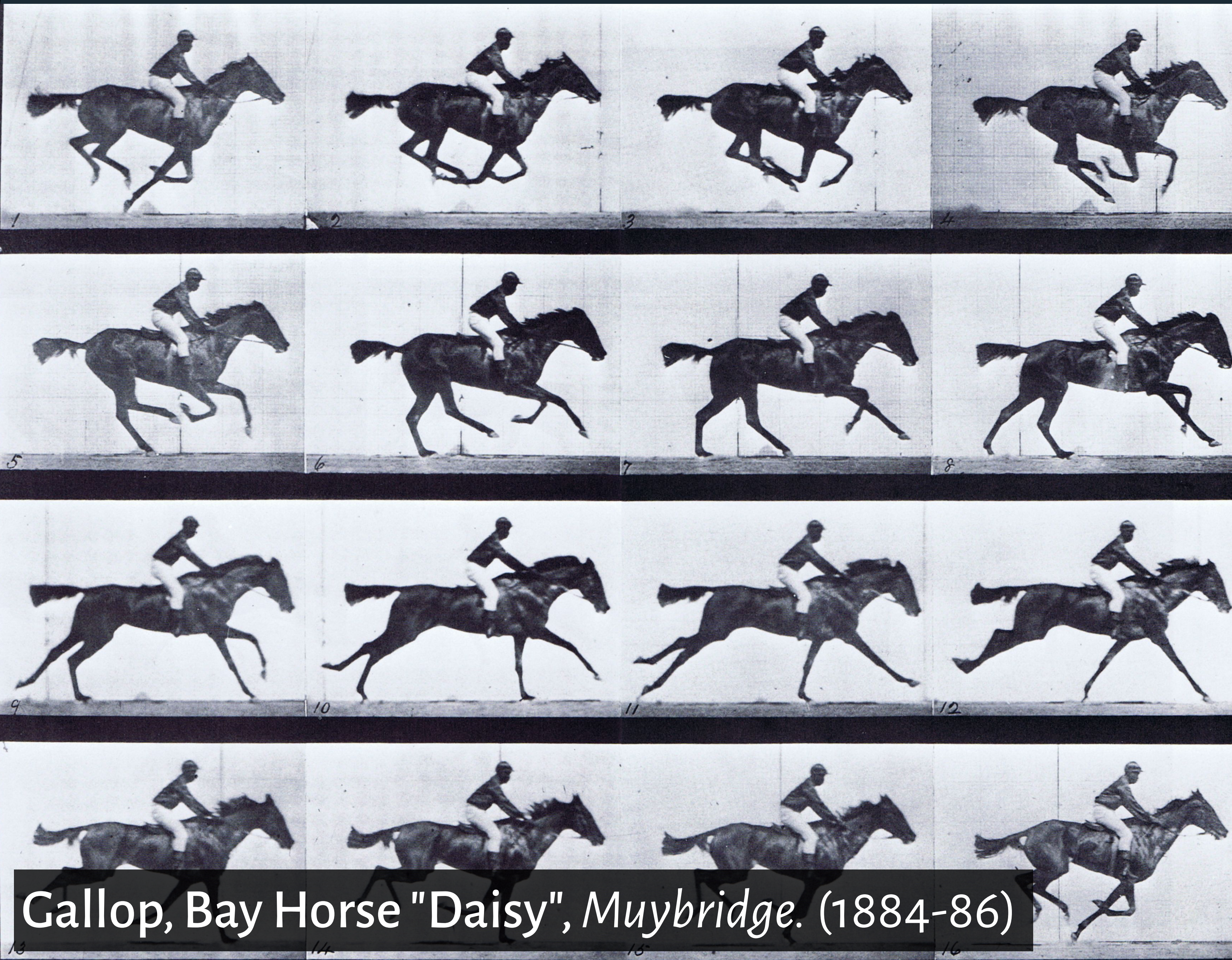
Communicate information to others (explanatory visualization)

Share and persuade

Collaborate and revise

Record Info

To answer a question:
Do all 4 hooves leave
the ground when a
horse gallops?



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

Record Info

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

Directly recorded pulse
as a waveform.

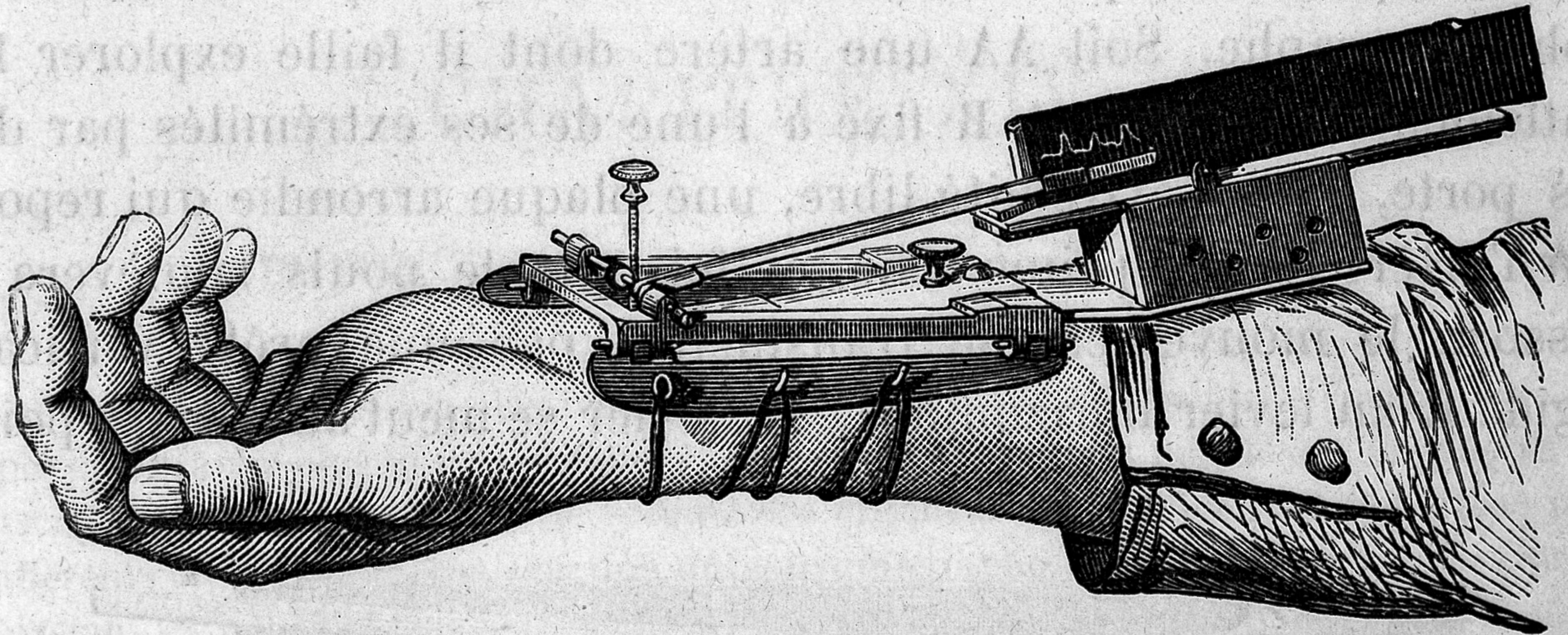
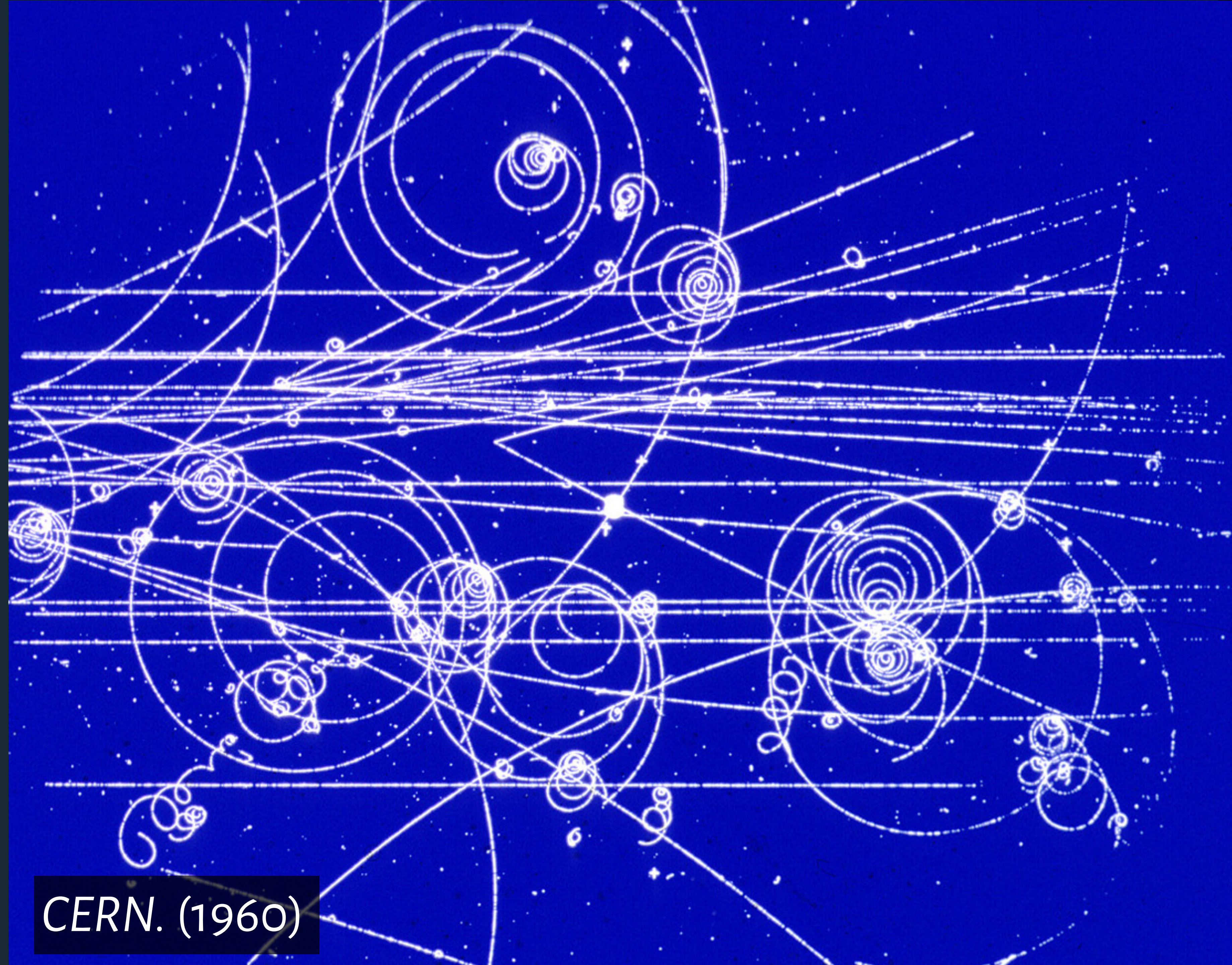


Fig. 109. Sphygmographe direct.

Record Info

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



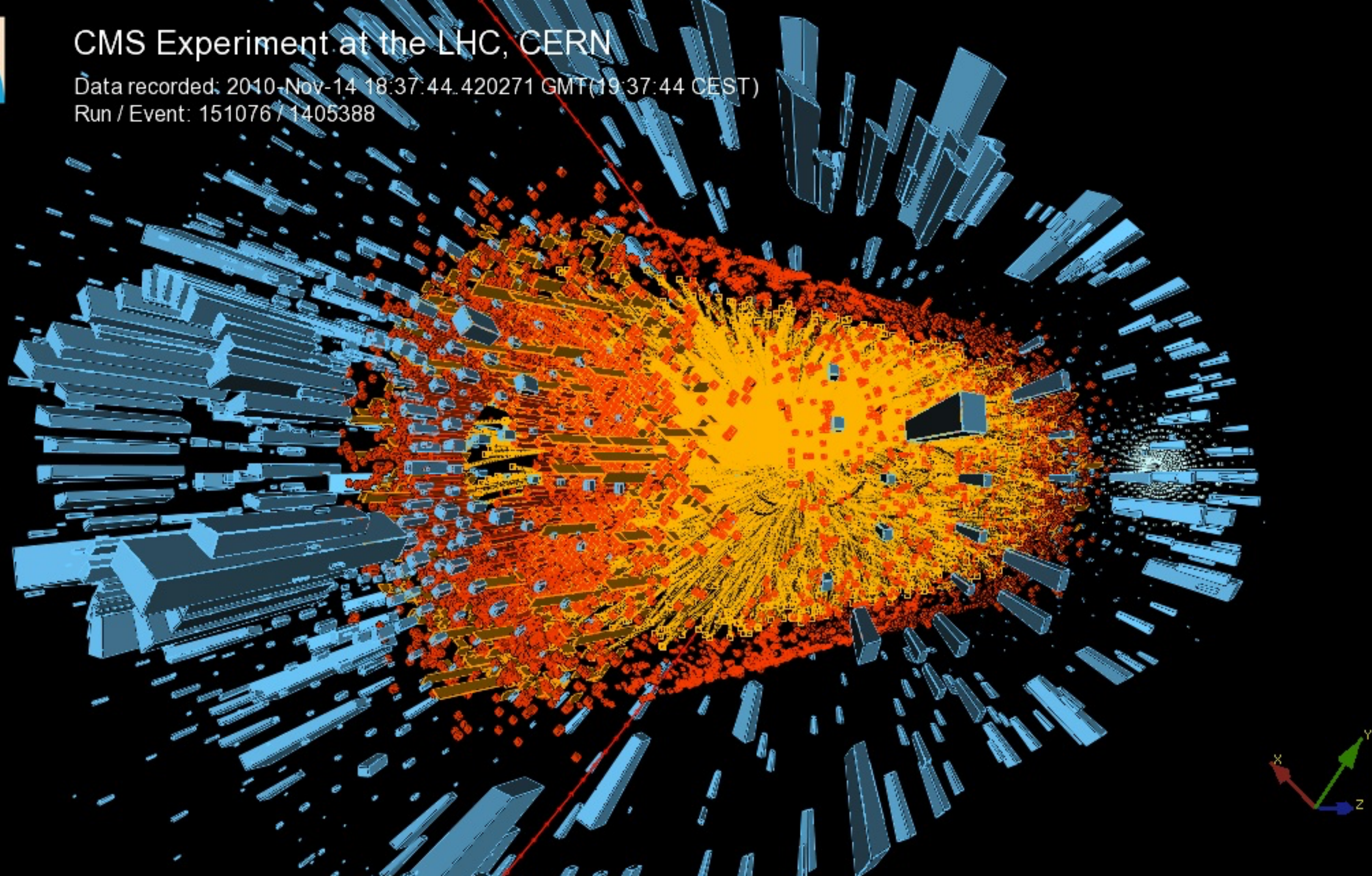
CERN. (1960)



CMS Experiment at the LHC, CERN

Data recorded: 2010-Nov-14 18:37:44.420271 GMT(19:37:44 CEST)

Run / Event: 151076 / 1405388



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

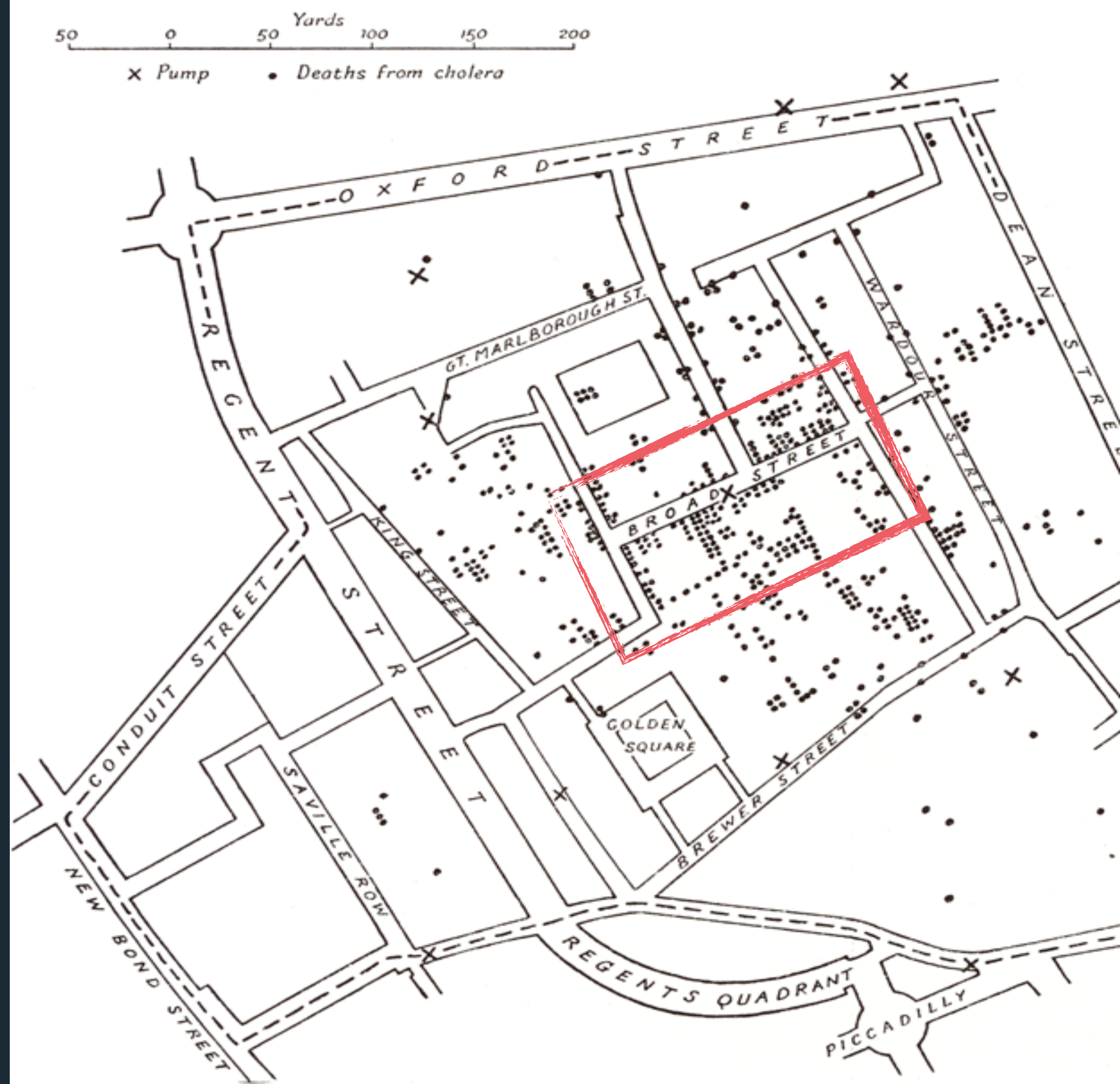
Collaborate and revise

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.

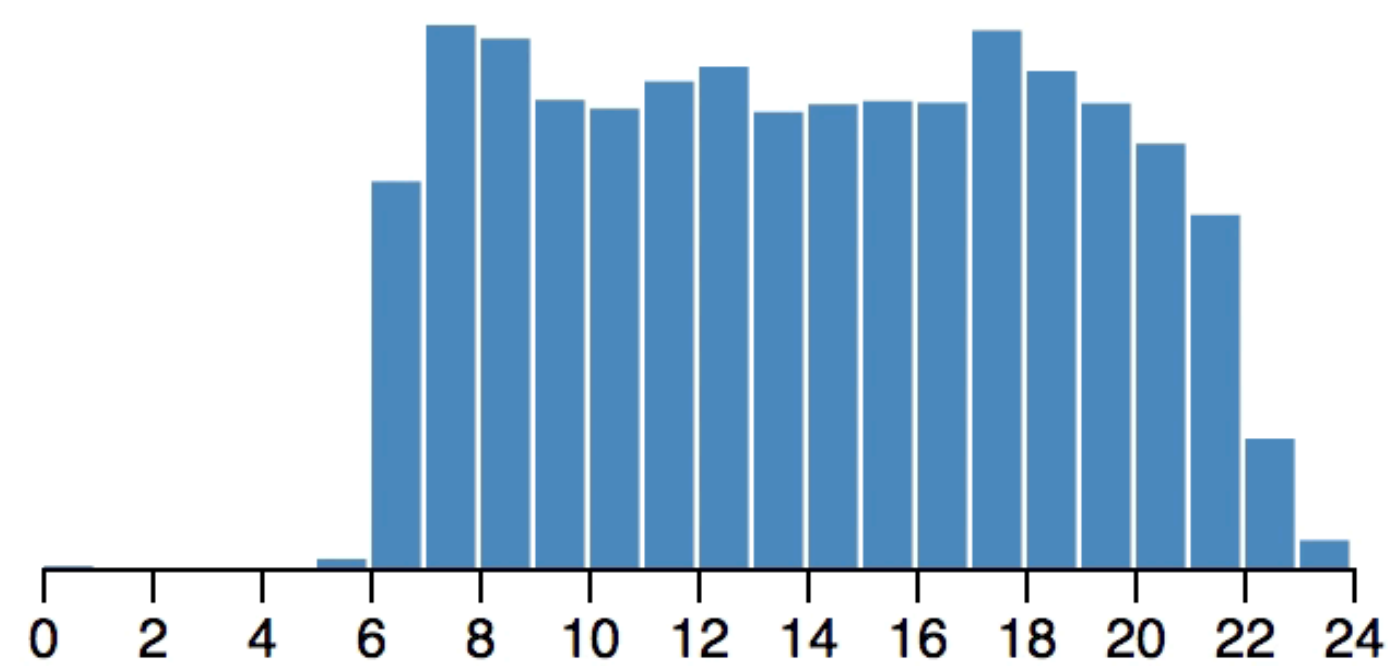
Used to support hypothesis that Broad St. pump was the cause.



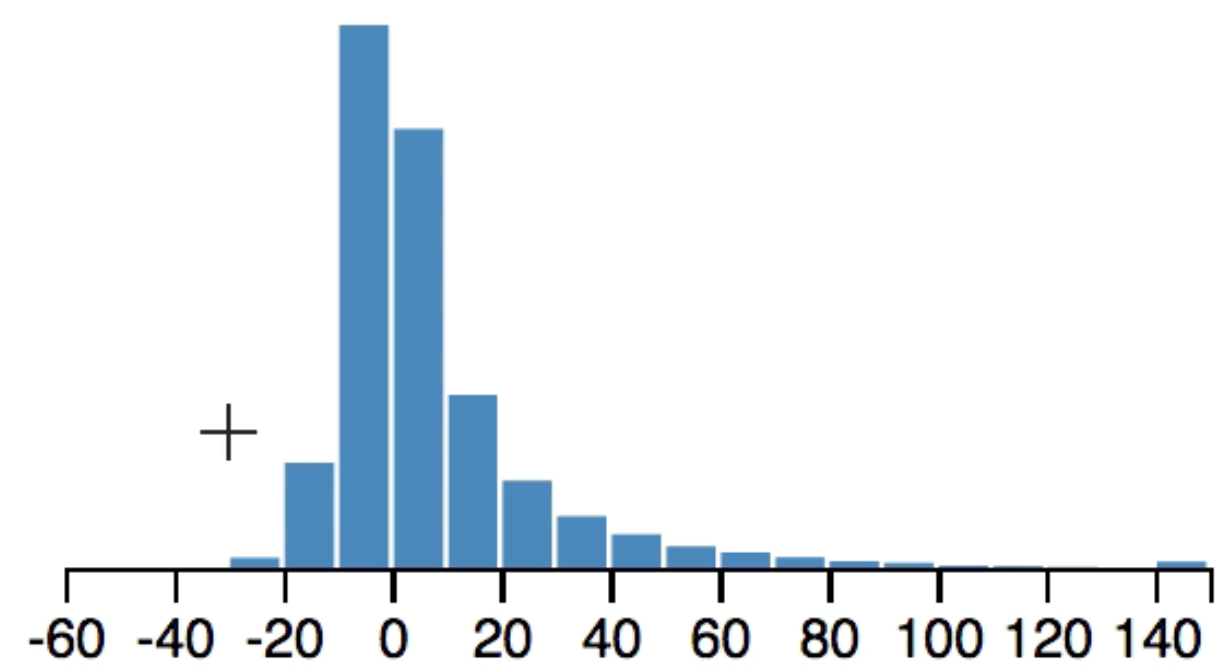
Support Reasoning

Interactively generate and evaluate hypotheses.

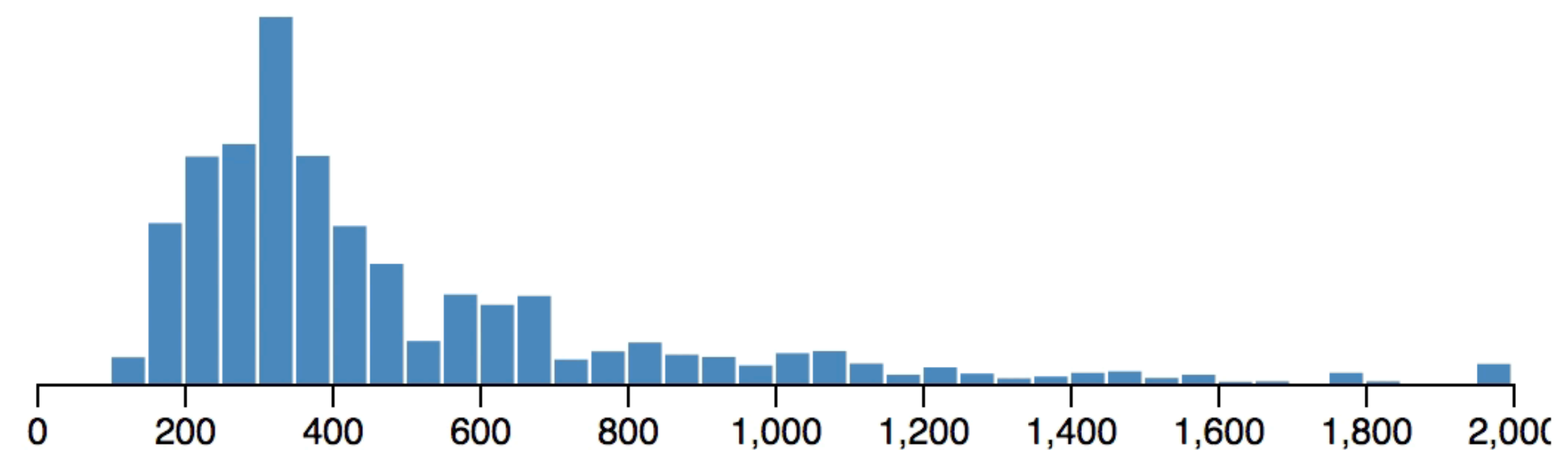
Time of Day



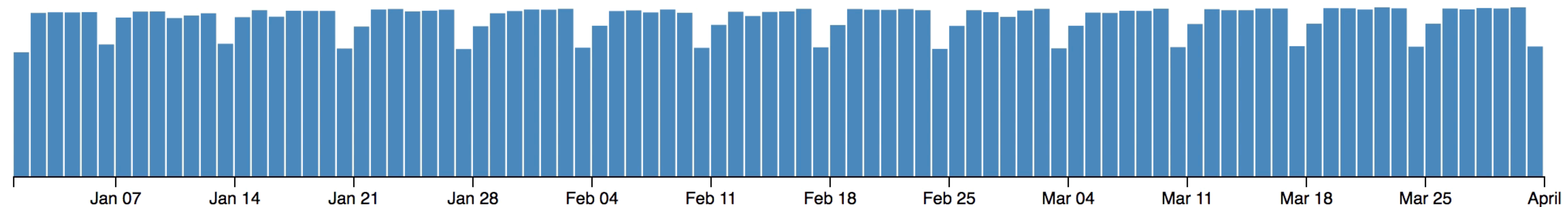
Arrival Delay (min.)



Distance (mi.)

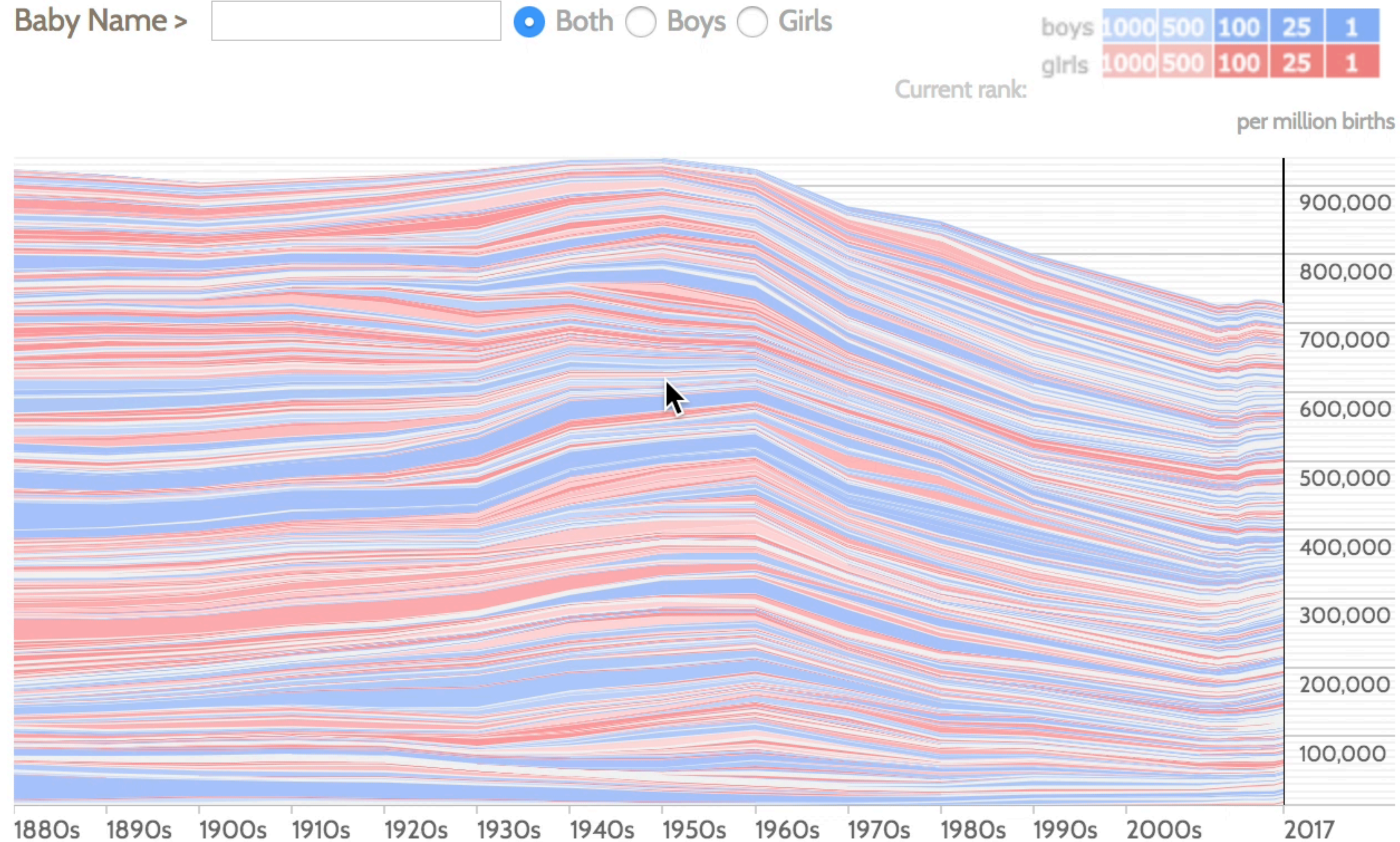


Date



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

[enlarge](#)

Support Reasoning

Class Exercise!

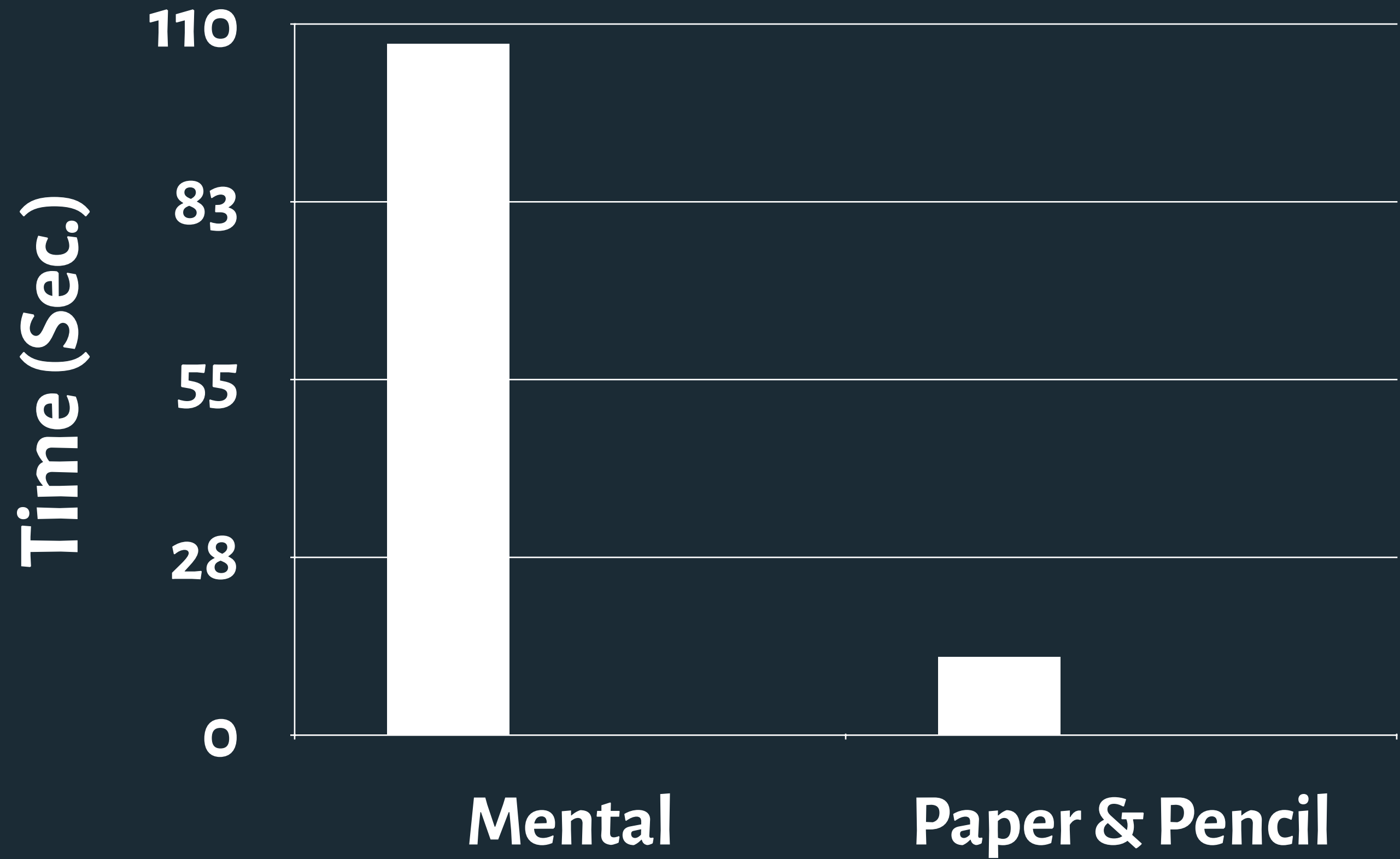
$$34 \times 72$$

Support Reasoning

$$\begin{array}{r} 34 \\ \times 72 \\ \hline \end{array}$$

Support Reasoning

$$\begin{array}{r} 34 \\ \times 72 \\ \hline 68 \\ 2380 \\ \hline 2448 \end{array}$$



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

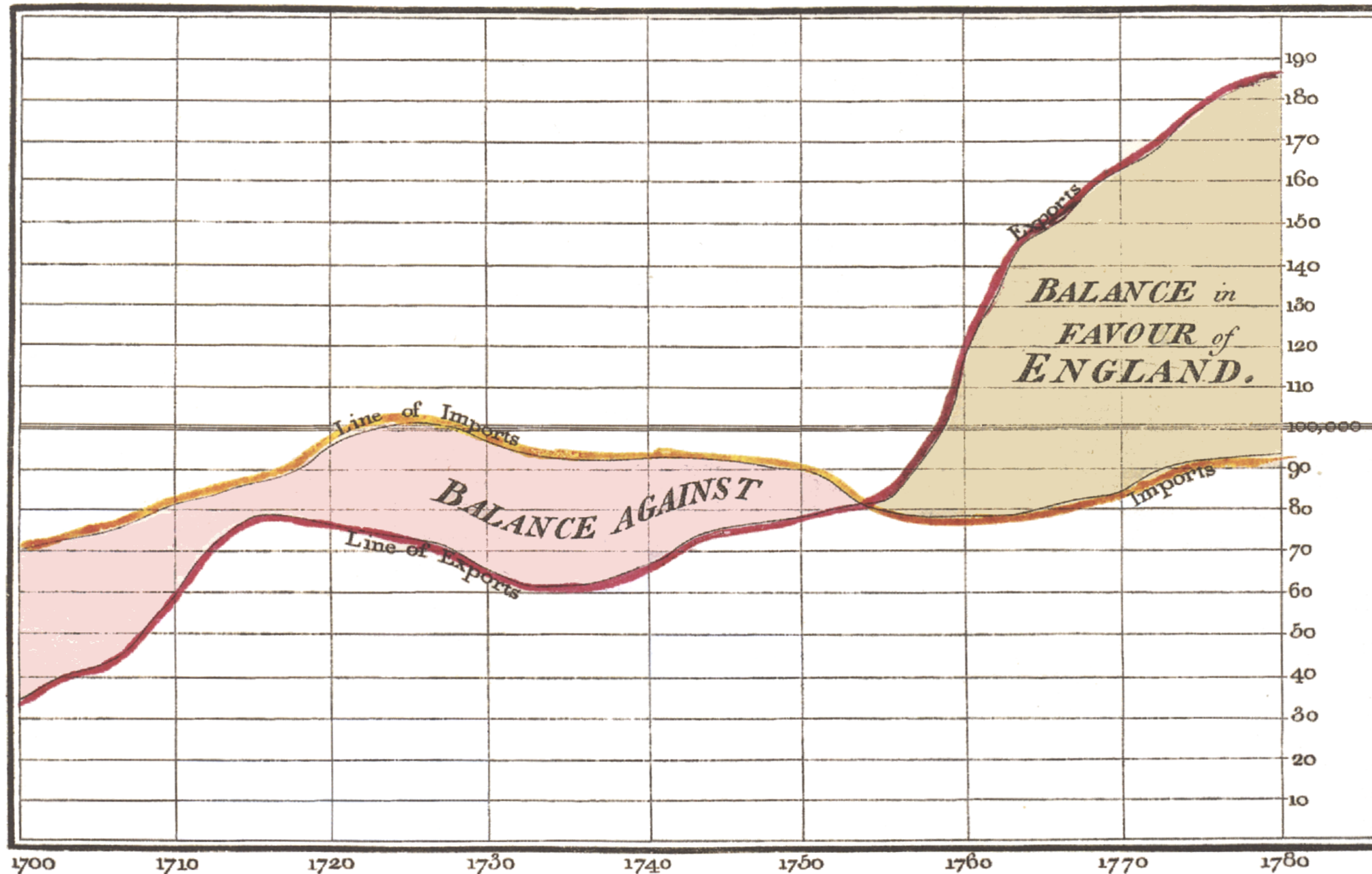
Communicate information to others (explanatory visualization)

Share and persuade

Collaborate and revise

Communicate Info

Exports and Imports to and from DENMARK & NORWAY from 1700 to 1780.



The Bottom line is divided into Years, the Right hand line into £10,000 each.

Published as the Act directs, 14th May 1786. by W^m Playfair

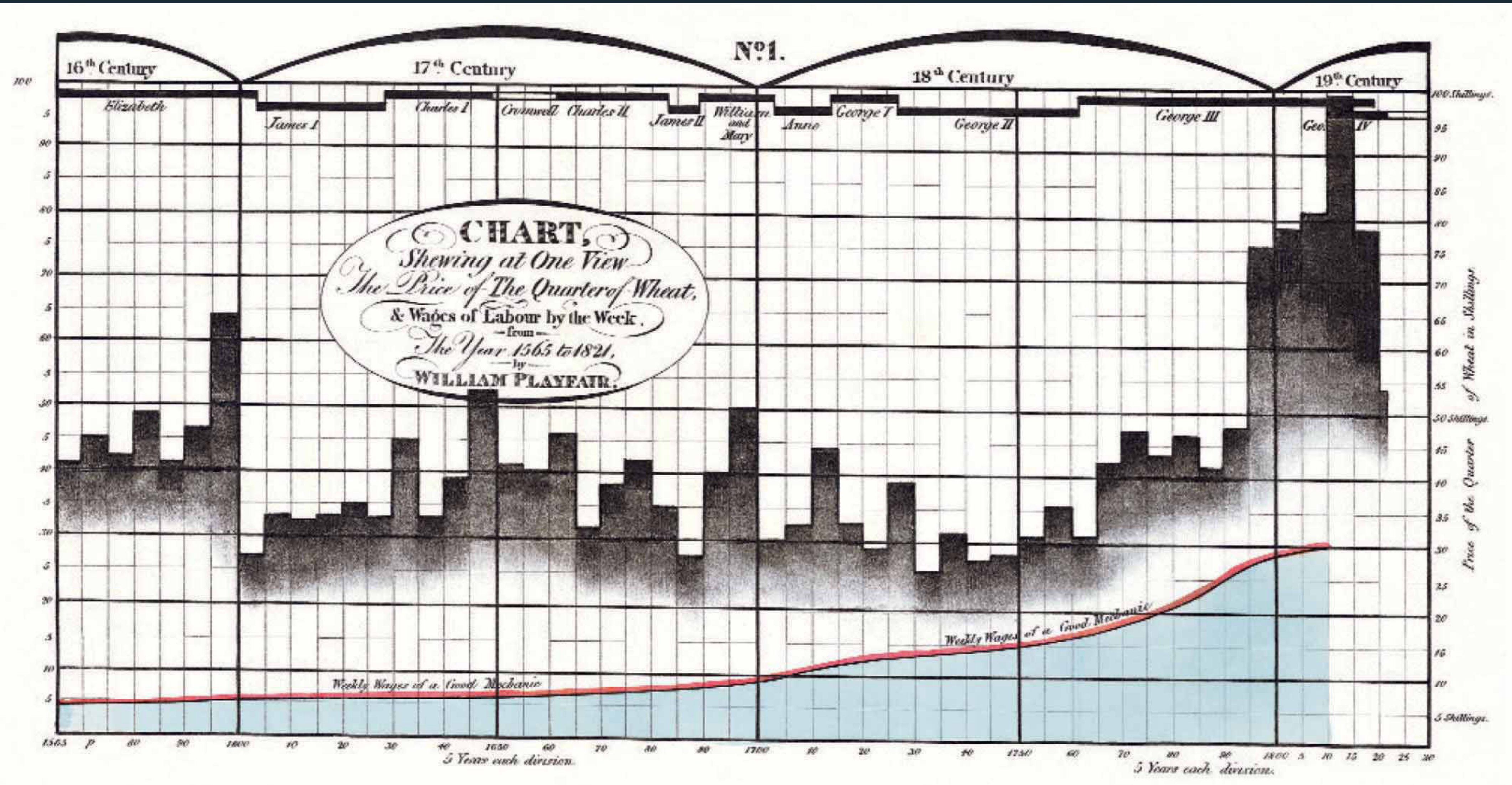
Needle sculpt 352, Strand, London.

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.

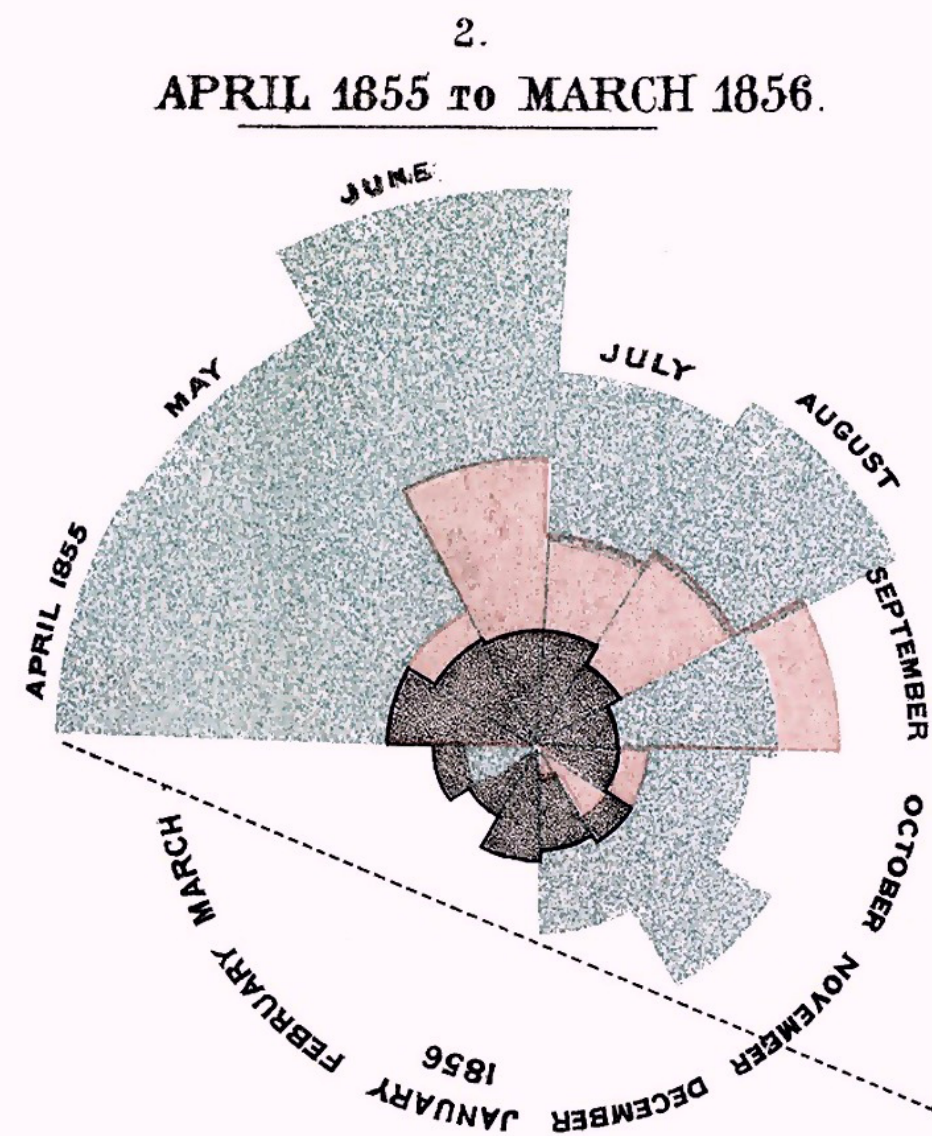


Communicate Info

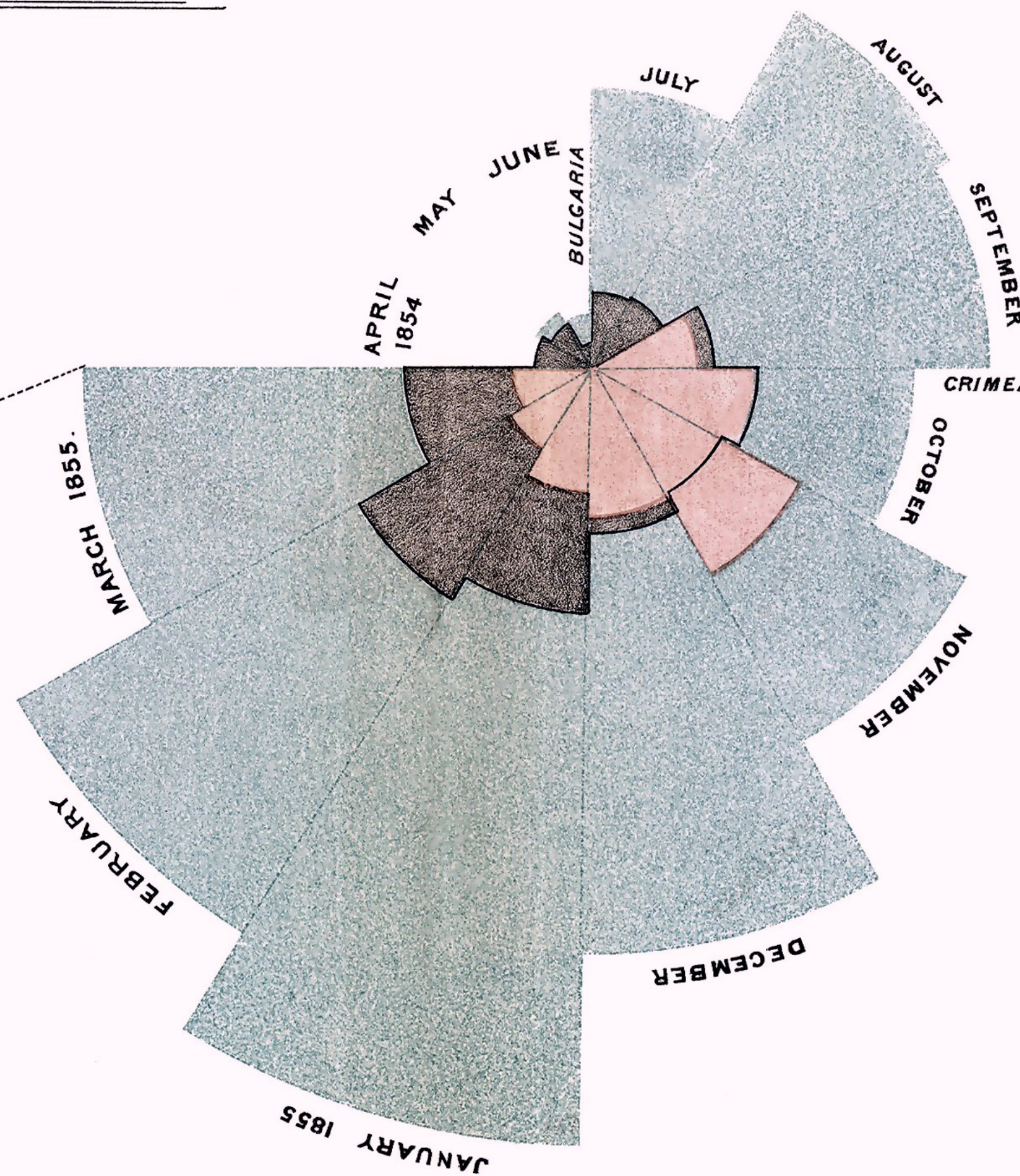
"to affect thro' the Eyes what we fail to convey to the public through their word-proof ears"

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

DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.



1.
APRIL 1854 TO MARCH 1855.



The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventable or Mitigable Zymotic diseases; the red wedges measured from the centre the deaths from wounds; & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during the month.
In October 1854, & April 1855, the black area coincides with the red; in January & February 1856, the blue coincides with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

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).

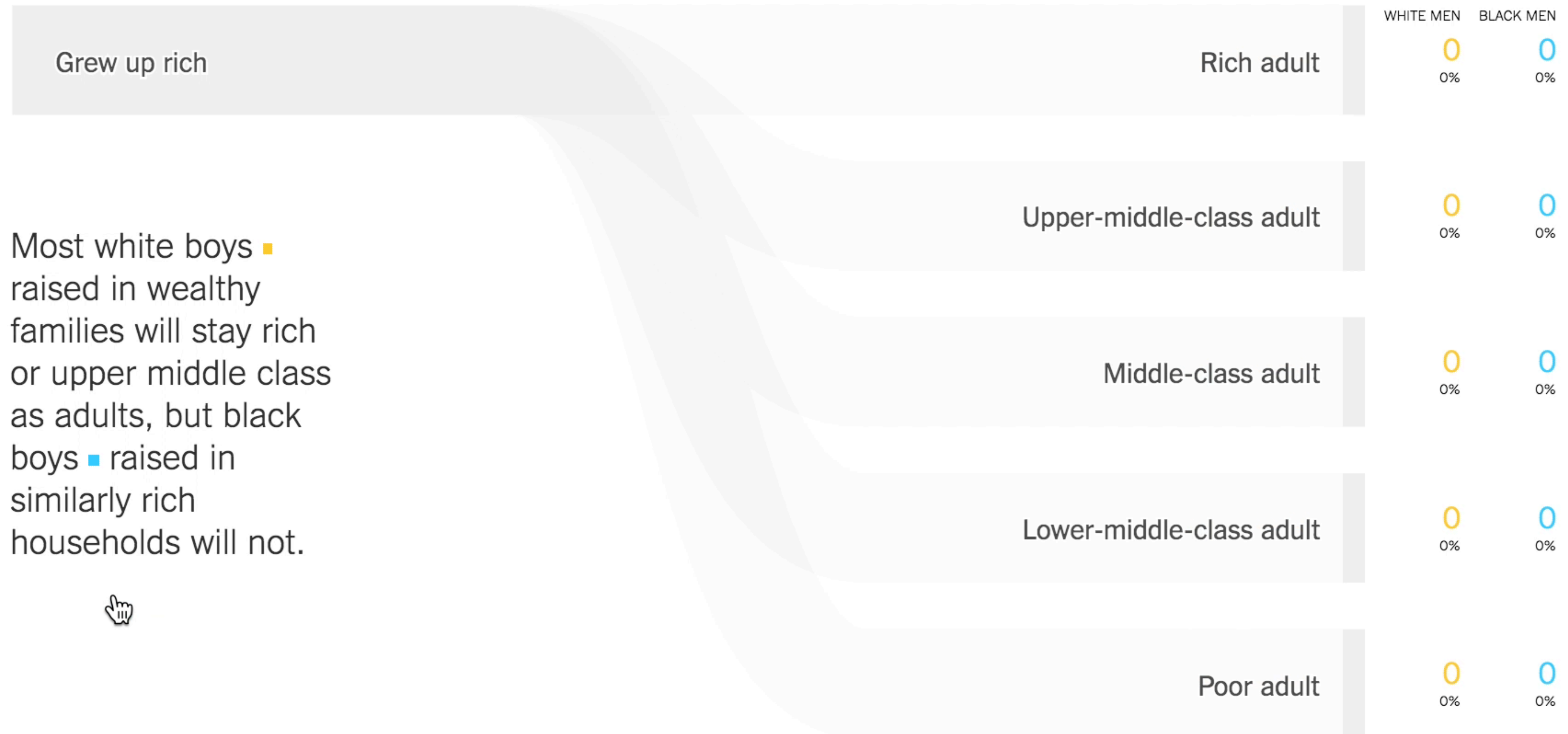


Extensive Data Shows Punishing Reach of Racism for Black Boys

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

Follow the lives of 0 boys who grew up in rich families ...

...and see where they end up as adults:



Most white boys ■ raised in wealthy families will stay rich or upper middle class as adults, but black boys ■ raised in similarly rich households will not.



Adult outcomes reflect household incomes in 2014 and 2015.

How the popular vote for the House translates into seats

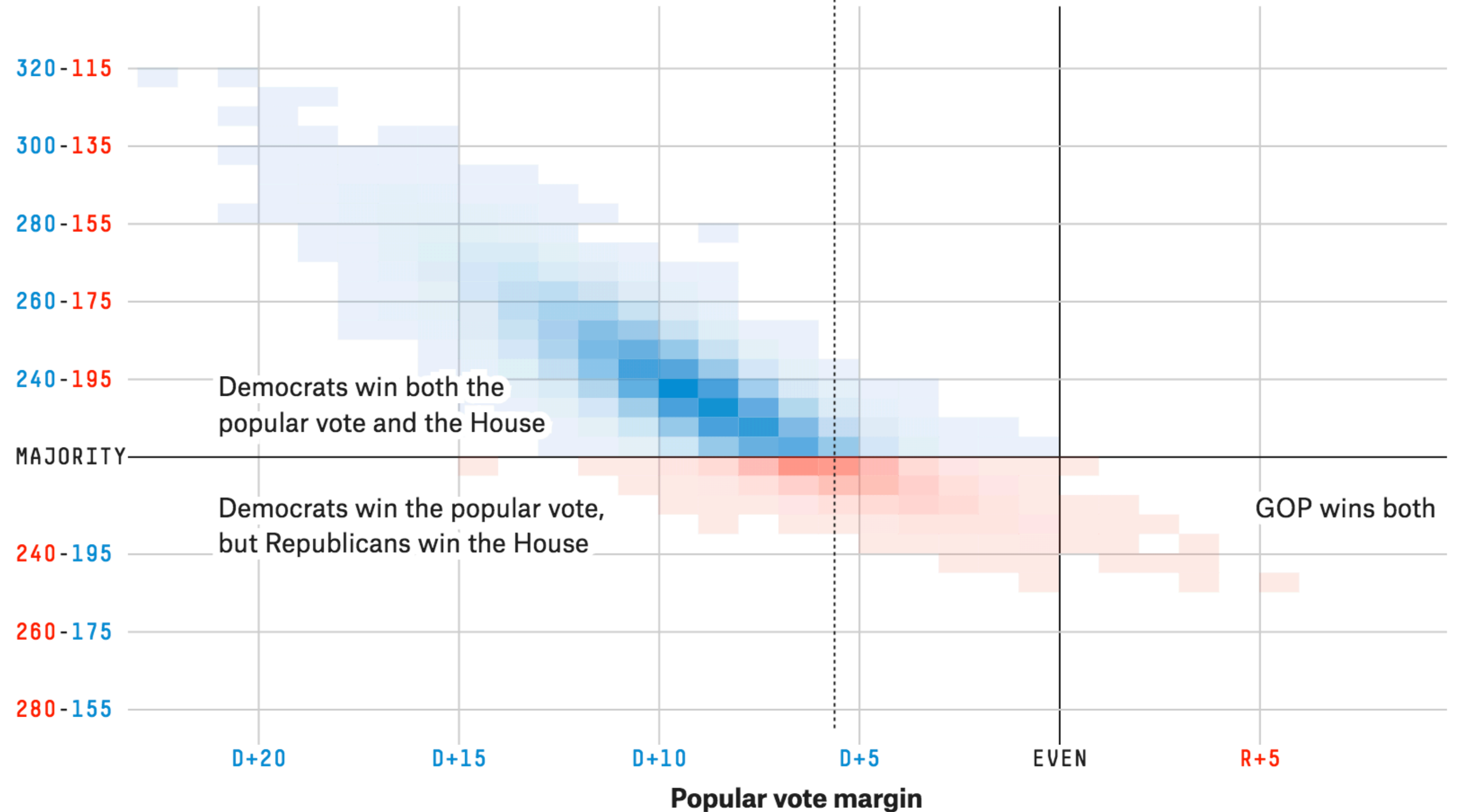
How various breakdowns in the national popular vote correspond to the most likely distributions of House seats by party, according to our forecast

HIGHER PROBABILITY →



Democrats are favored to win a majority of seats if they win the popular vote by at least 5.6 points

Party breakdown



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

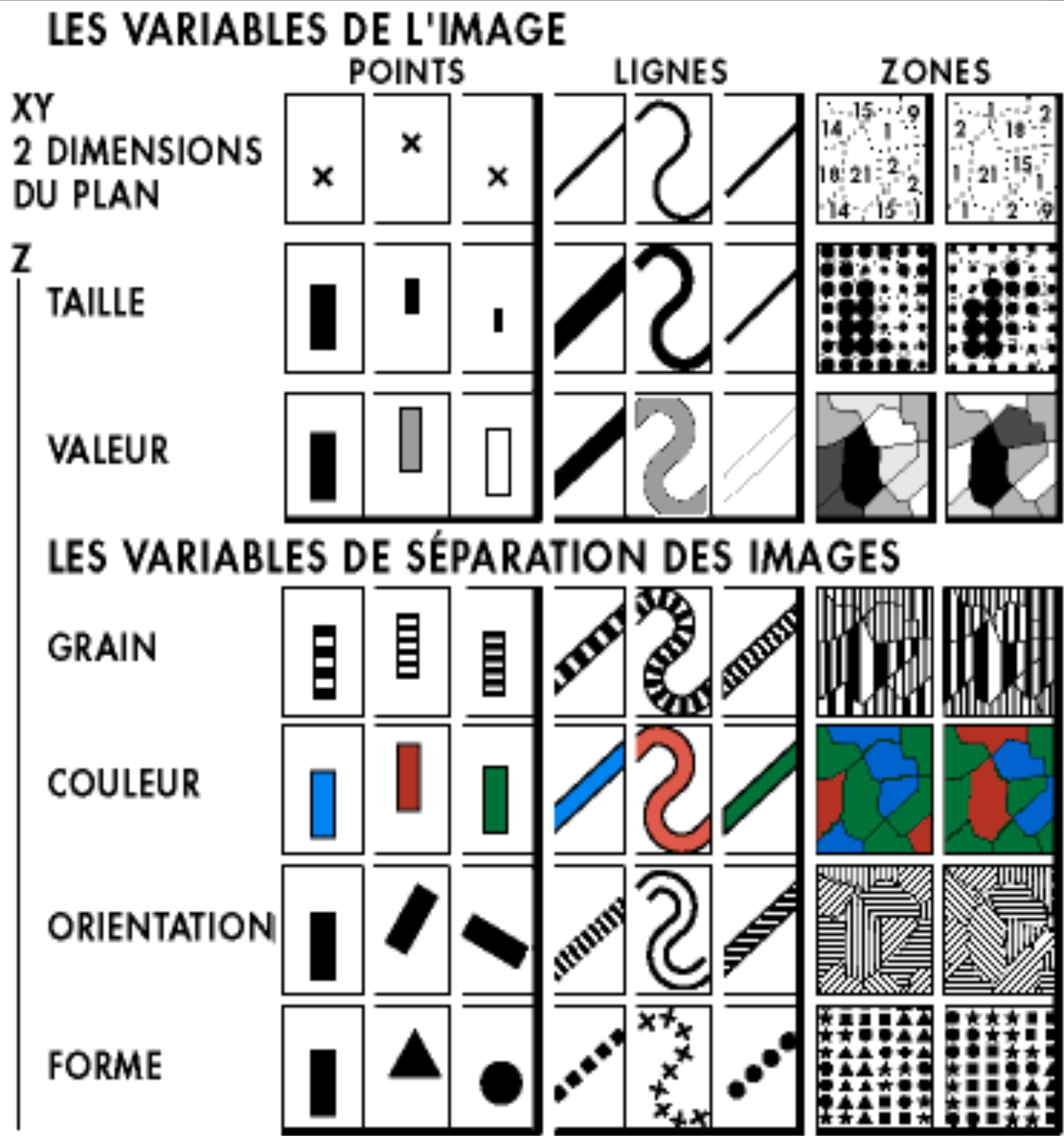
By the end of the course, you should expect to be able to:

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

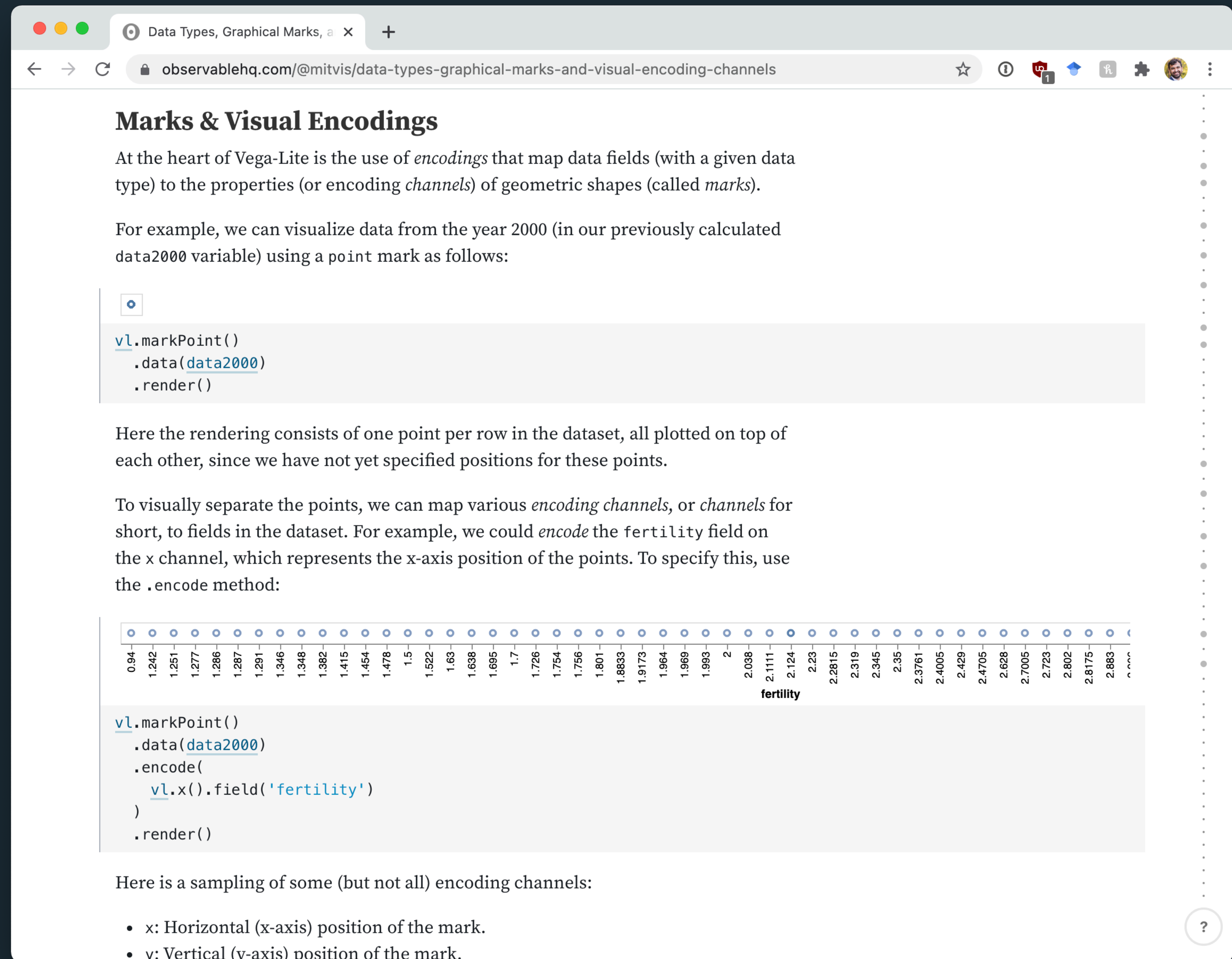
Course Topics

Data & Image Models



Course Topics

Visual Encoding with Vega-Lite



Marks & Visual Encodings

At the heart of Vega-Lite is the use of *encodings* that map data fields (with a given data type) to the properties (or encoding *channels*) of geometric shapes (called *marks*).

For example, we can visualize data from the year 2000 (in our previously calculated `data2000` variable) using a point mark as follows:

```
vl.markPoint()  
  .data(data2000)  
  .render()
```

Here the rendering consists of one point per row in the dataset, all plotted on top of each other, since we have not yet specified positions for these points.

To visually separate the points, we can map various *encoding channels*, or *channels* for short, to fields in the dataset. For example, we could *encode* the `fertility` field on the `x` channel, which represents the `x`-axis position of the points. To specify this, use the `.encode` method:

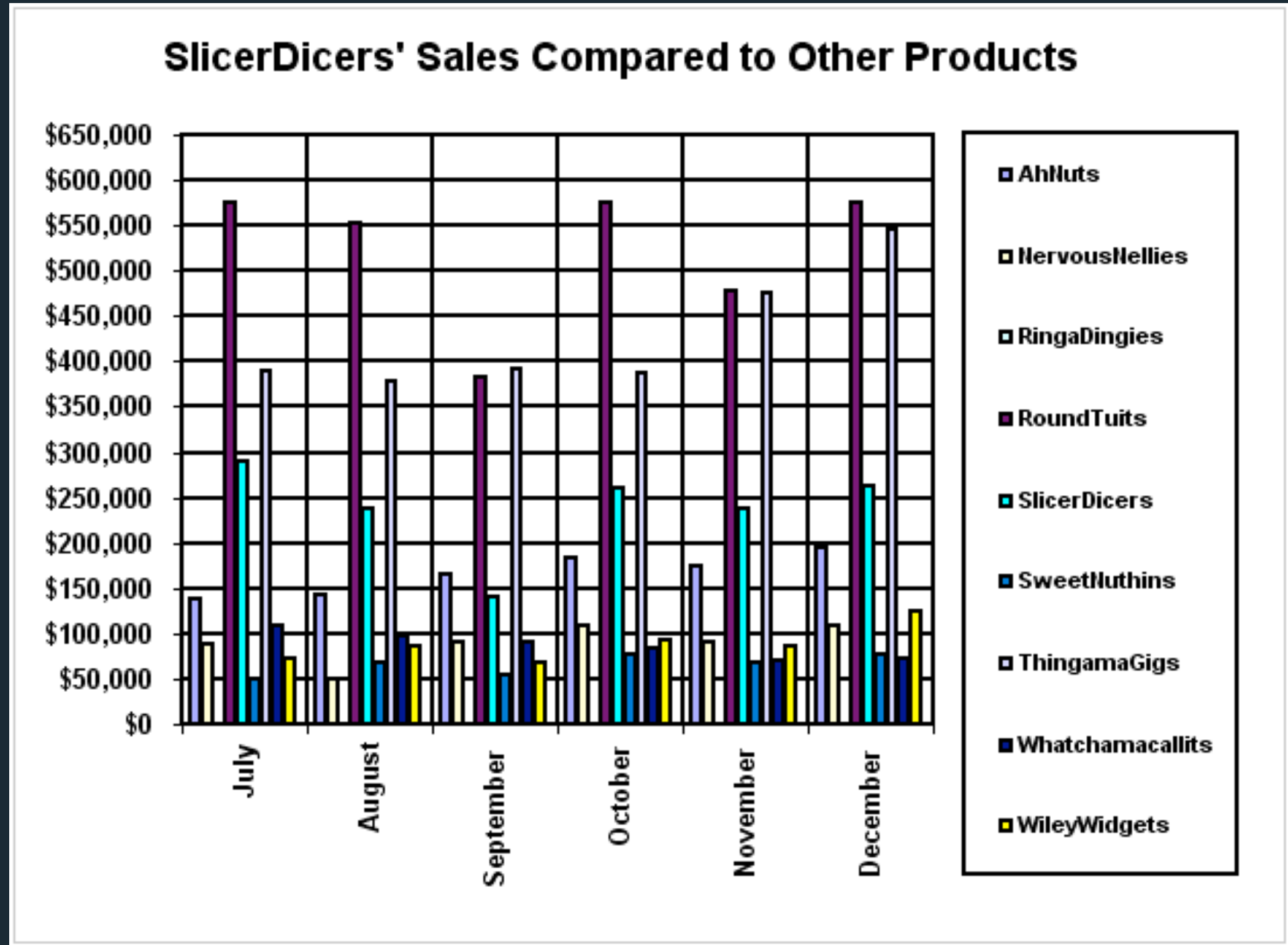
```
vl.markPoint()  
  .data(data2000)  
  .encode(  
    vl.x().field('fertility')  
  )  
  .render()
```

Here is a sampling of some (but not all) encoding channels:

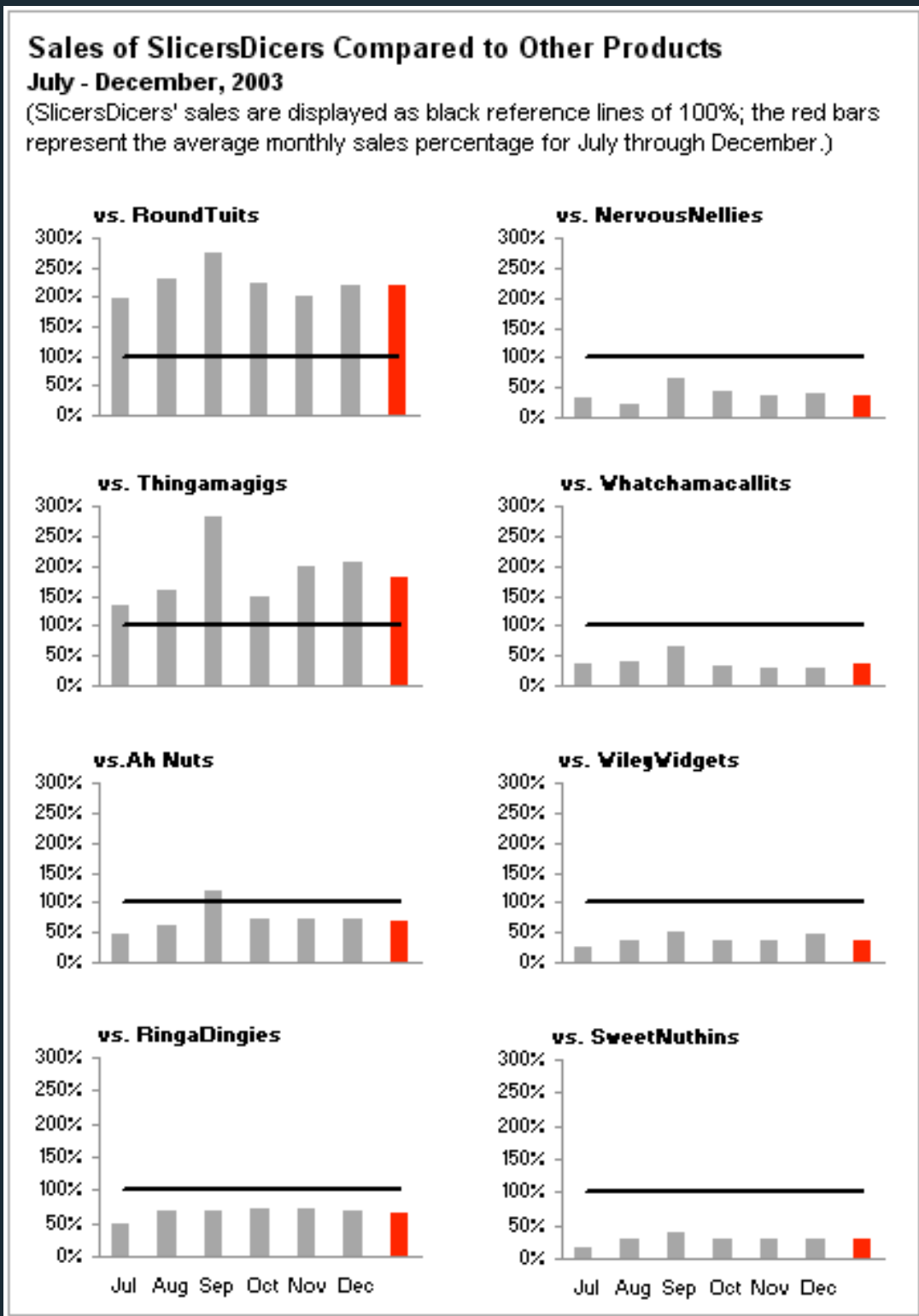
- `x`: Horizontal (`x`-axis) position of the mark.
- `y`: Vertical (`y`-axis) position of the mark.

Course Topics

(Re-)Design



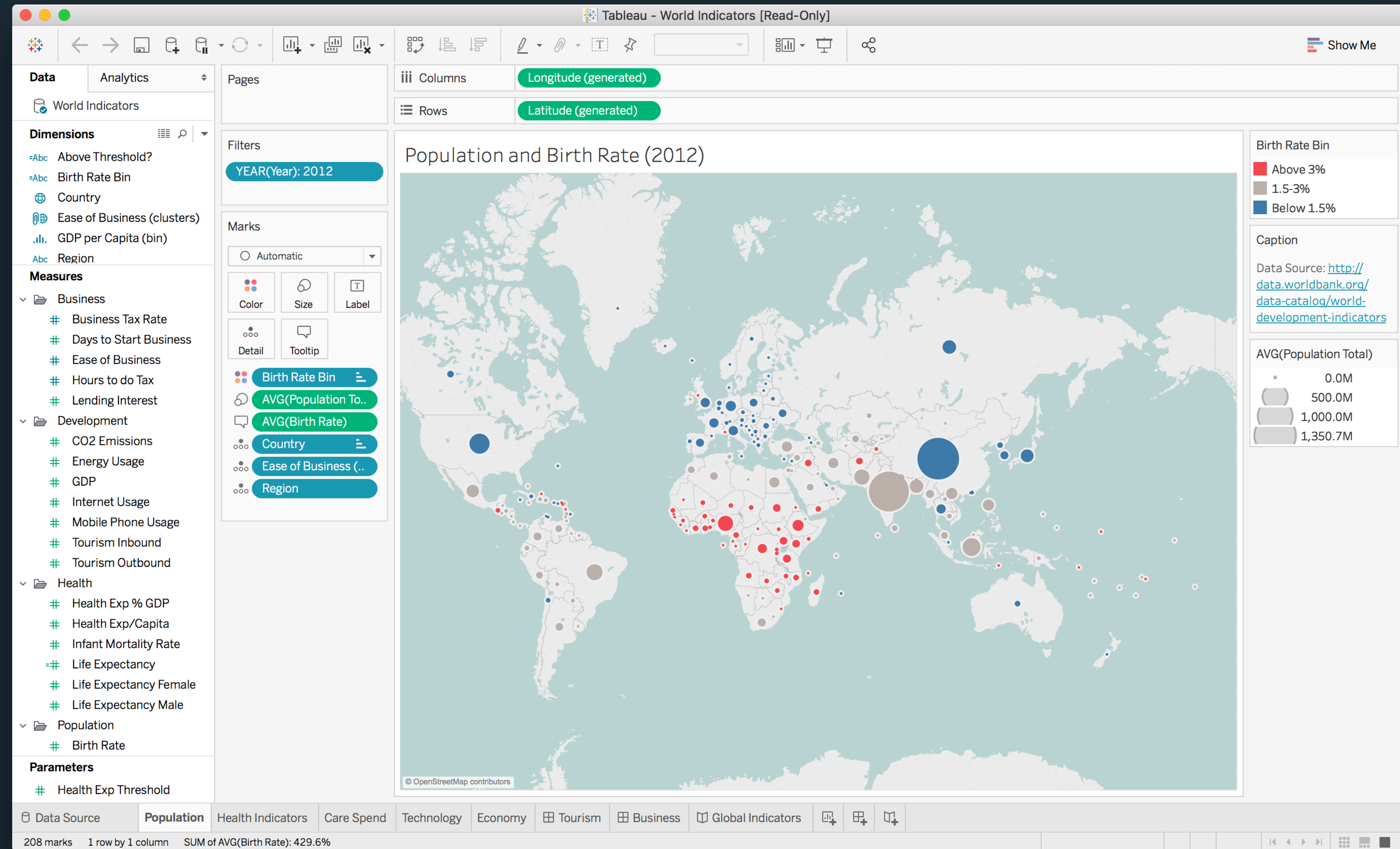
Problematic Design



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

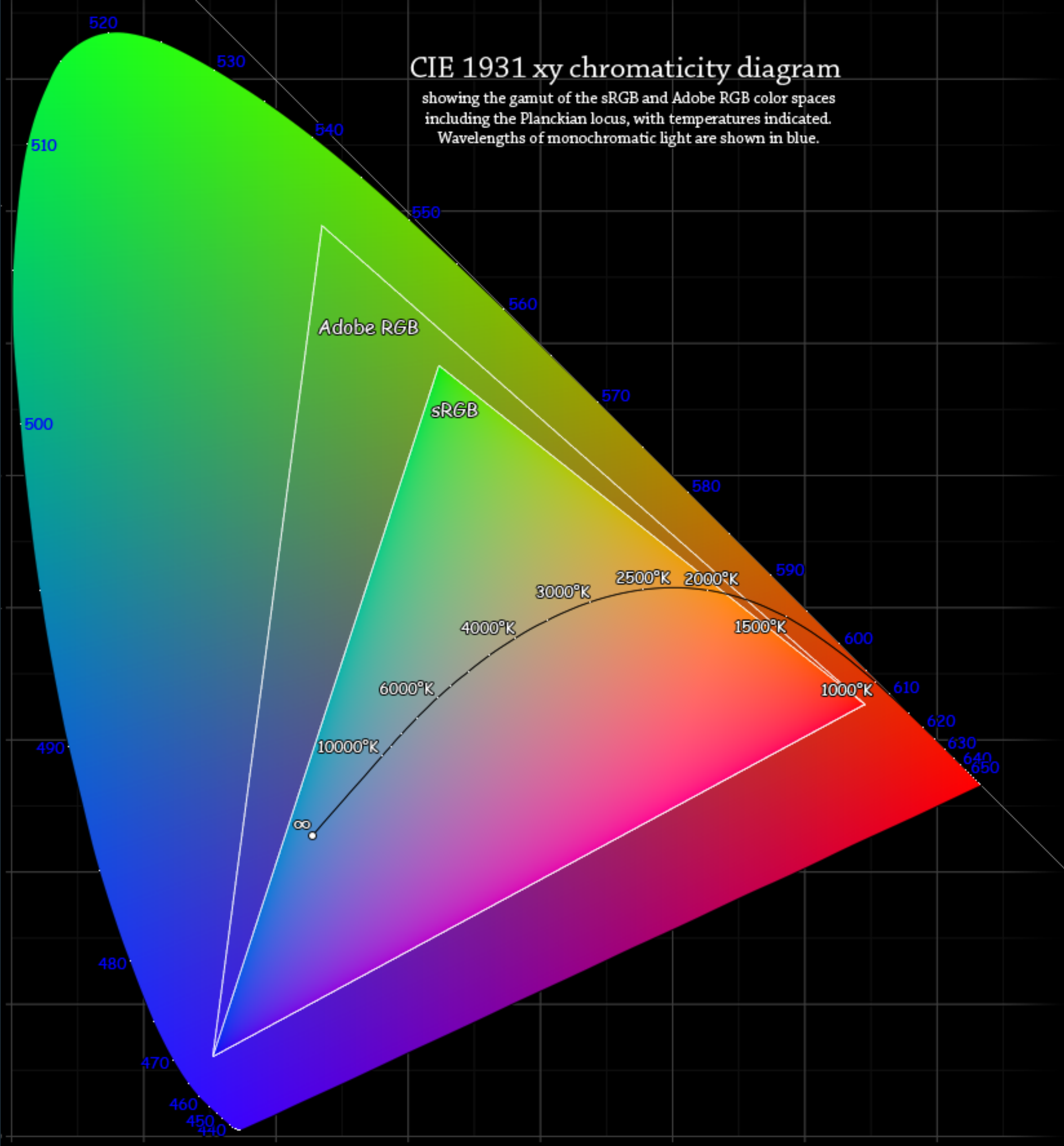
green

orange

purple

Course Topics

Color

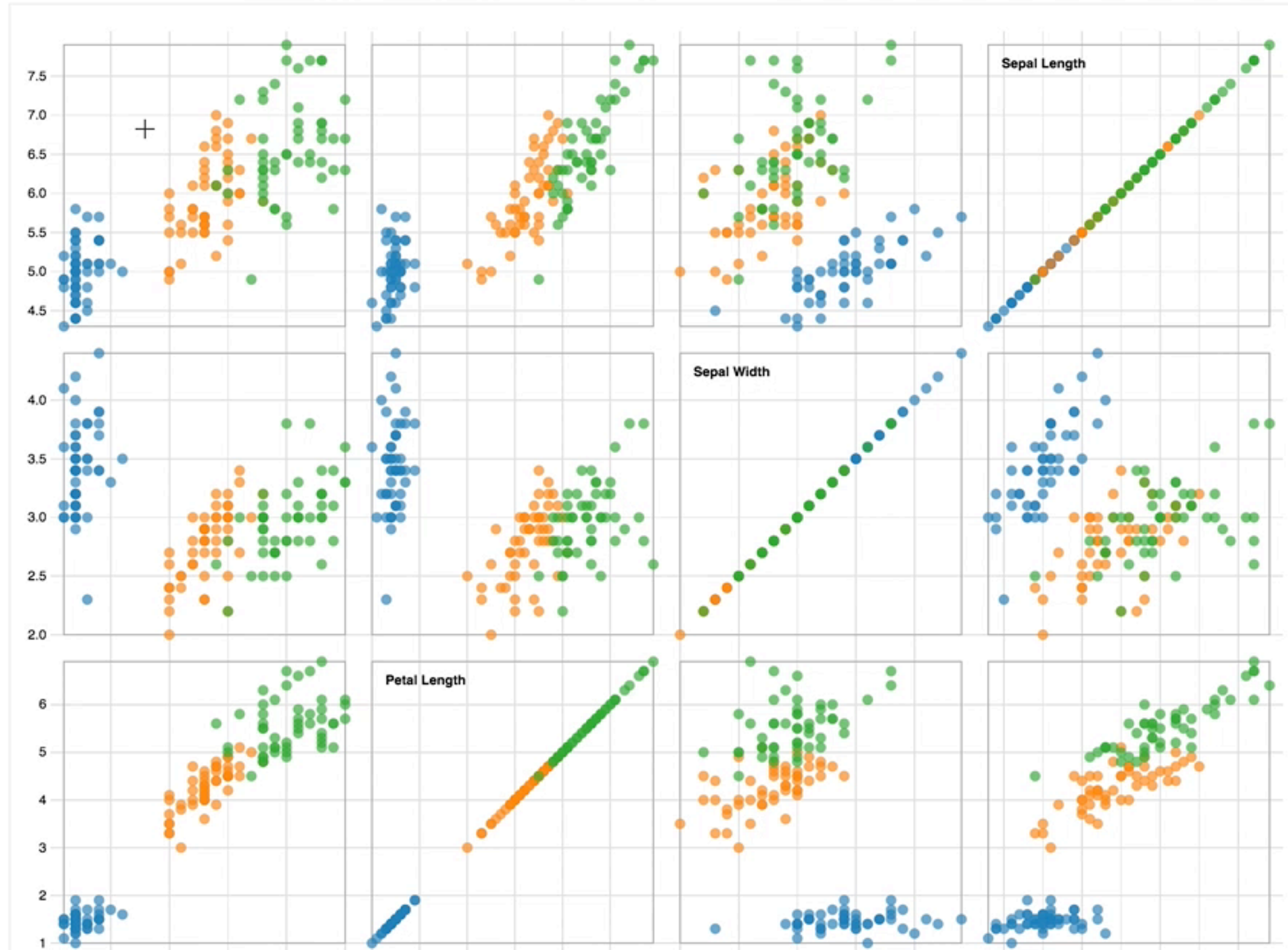


Course Topics

Interactivity

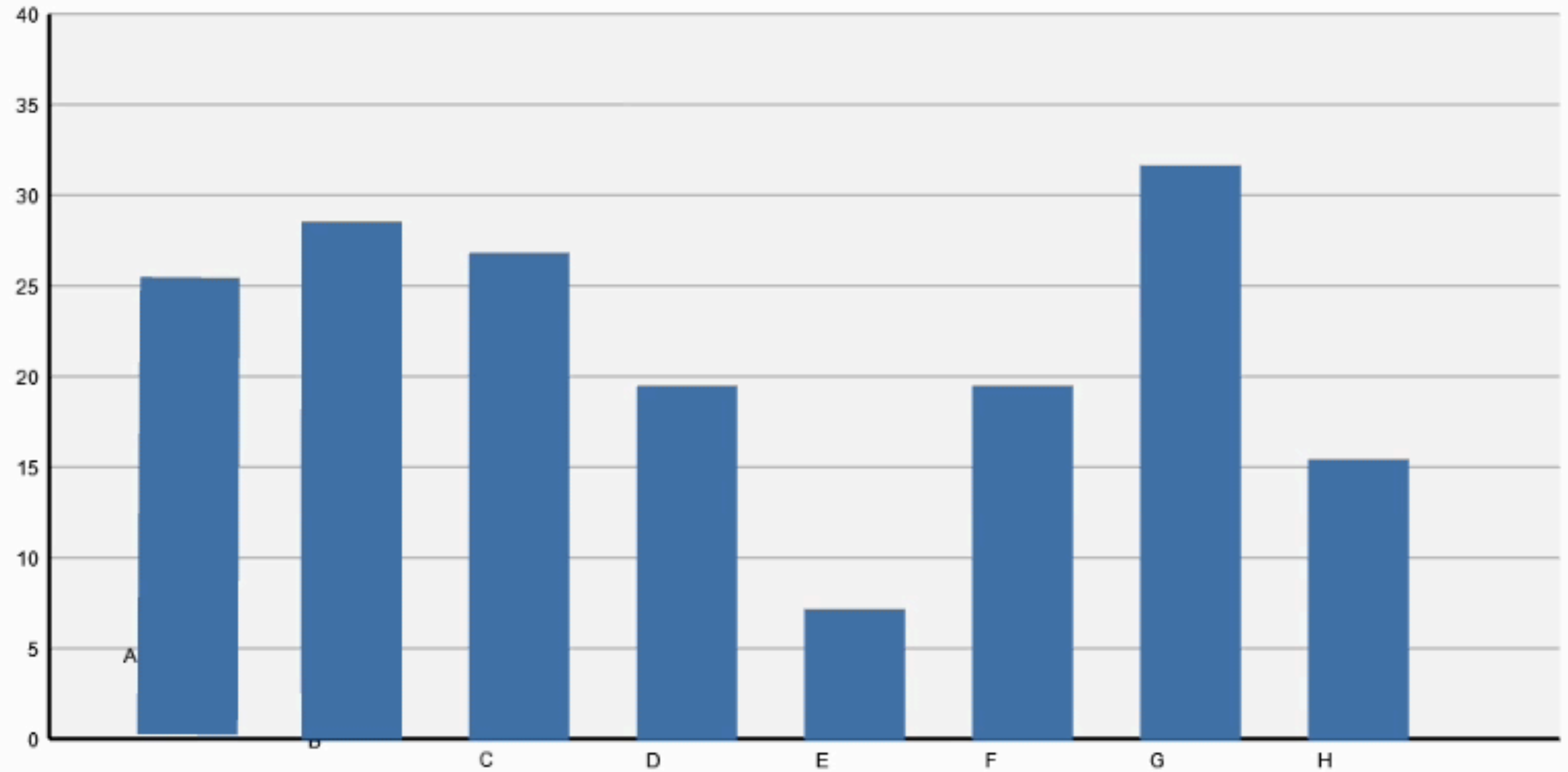


Scatterplot Matrix Brushing



Course Topics

Animation



Heer & Robertson. (2007).

Course Topics

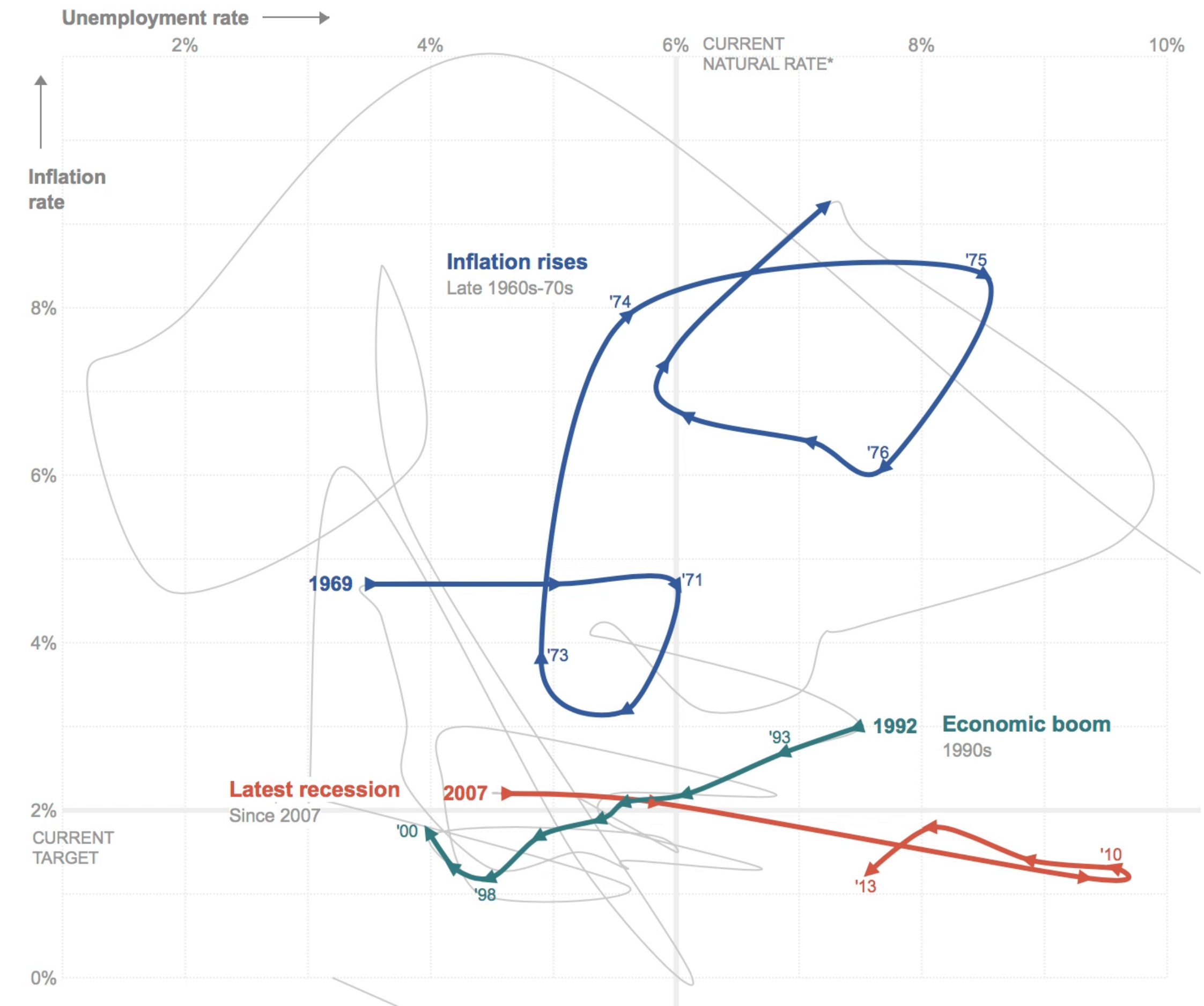
Narrative

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.



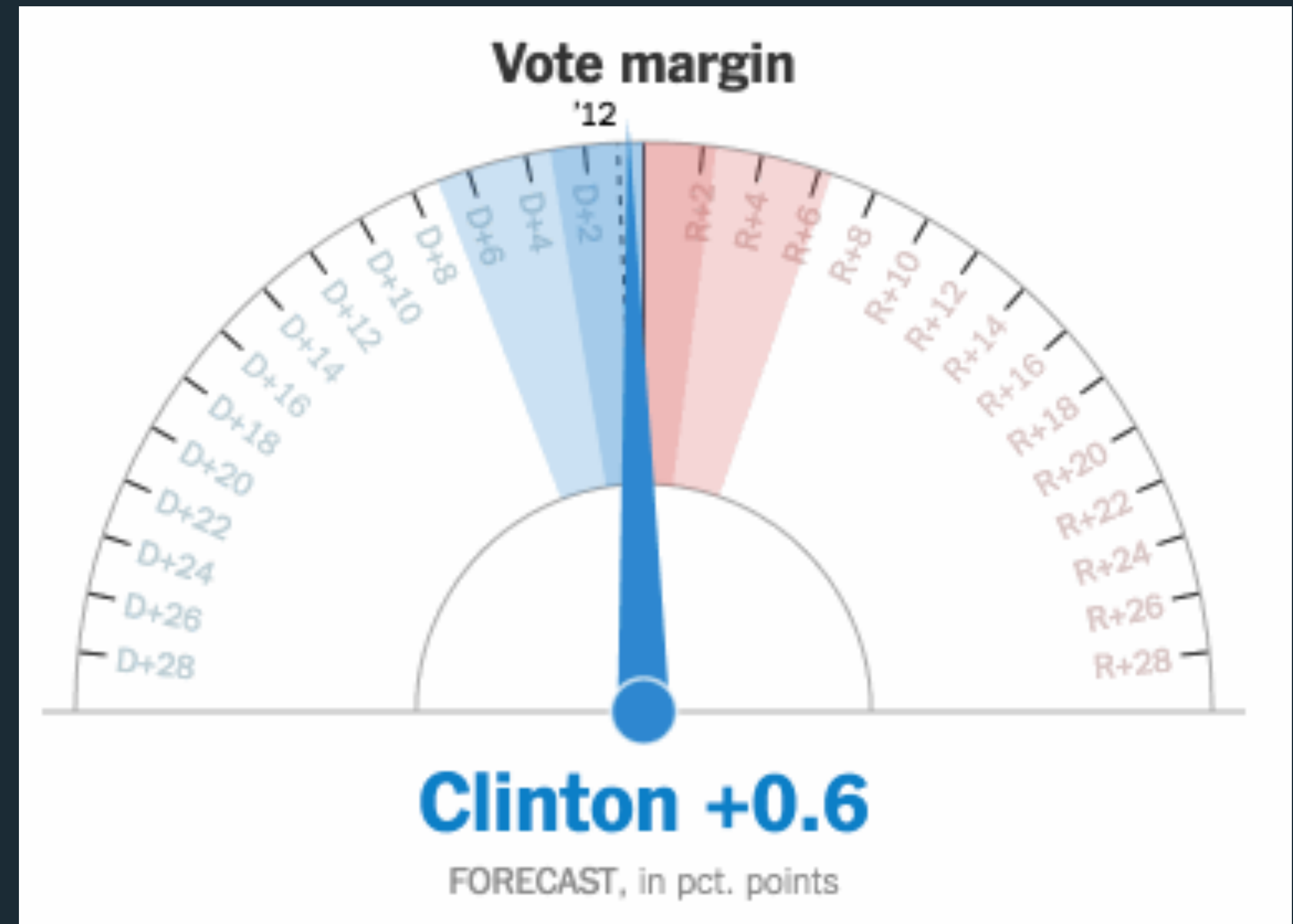
*The natural rate of unemployment is defined as the lowest sustainable level of unemployment over the long term. If the rate is pushed any lower than the natural level, wages and prices would rise.

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 → *Newish* 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



ECF MIT European Career Fair

Candidate Statistics Employer Statistics

European Career Fair Statistics

The MIT European Career Fair (ECF) is an annual recruiting event in the Greater Boston Area, organized by members of the European Club at the Massachusetts Institute of Technology (MIT).

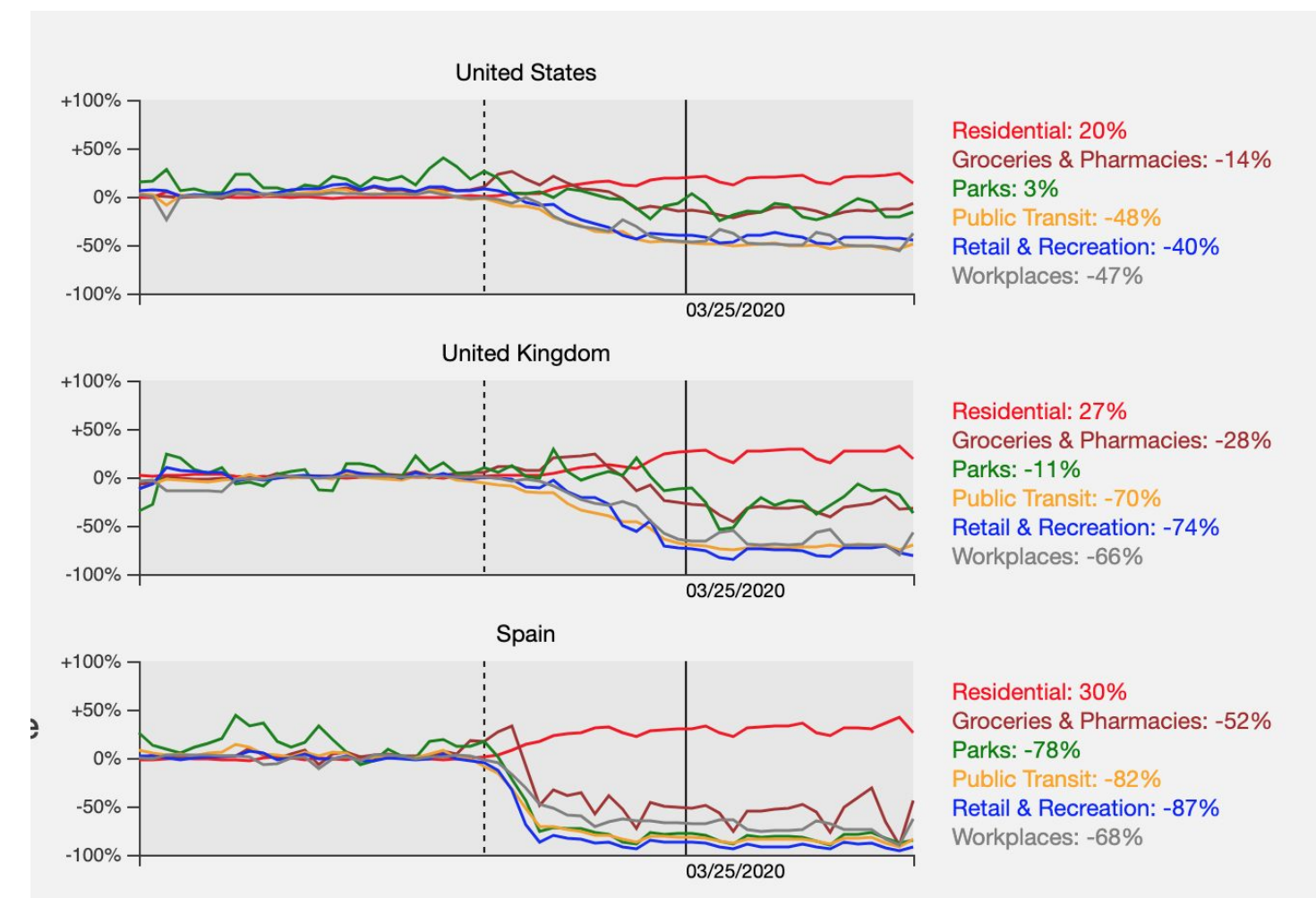
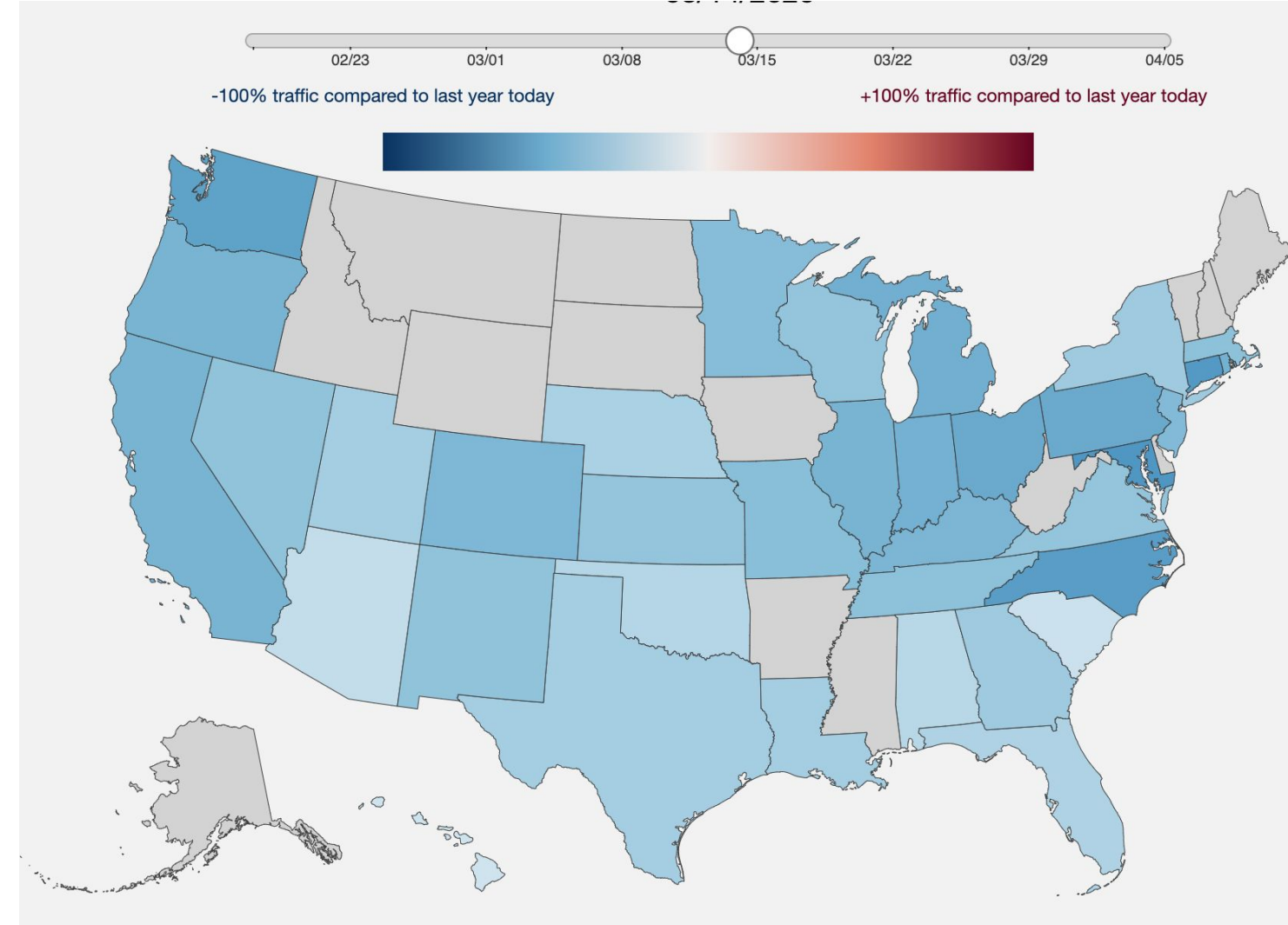
[GO TO THE ECF WEBSITE](#) [CONTINUE TO THE STATISTICS](#)

- Our mission**
Our mission is to connect top-tier US students and young professionals with European companies, universities, and non-profit organizations and help them learn about career opportunities in various fields, including robotics, pharmaceuticals, finance, aviation.
- Our legacy**
Organized since 1997, the ECF is the largest of its kind in the United States and has a proven track record providing domestic, foreign, and multinational companies access to some of the highest caliber applicants worldwide.
- Our participants**
Our candidates are students and recent alumni and renowned European and American universities. Our company participants are diverse, ranging from renowned multinationals to smaller, innovative firms and startups.



Saadiyah (she/her) MEng

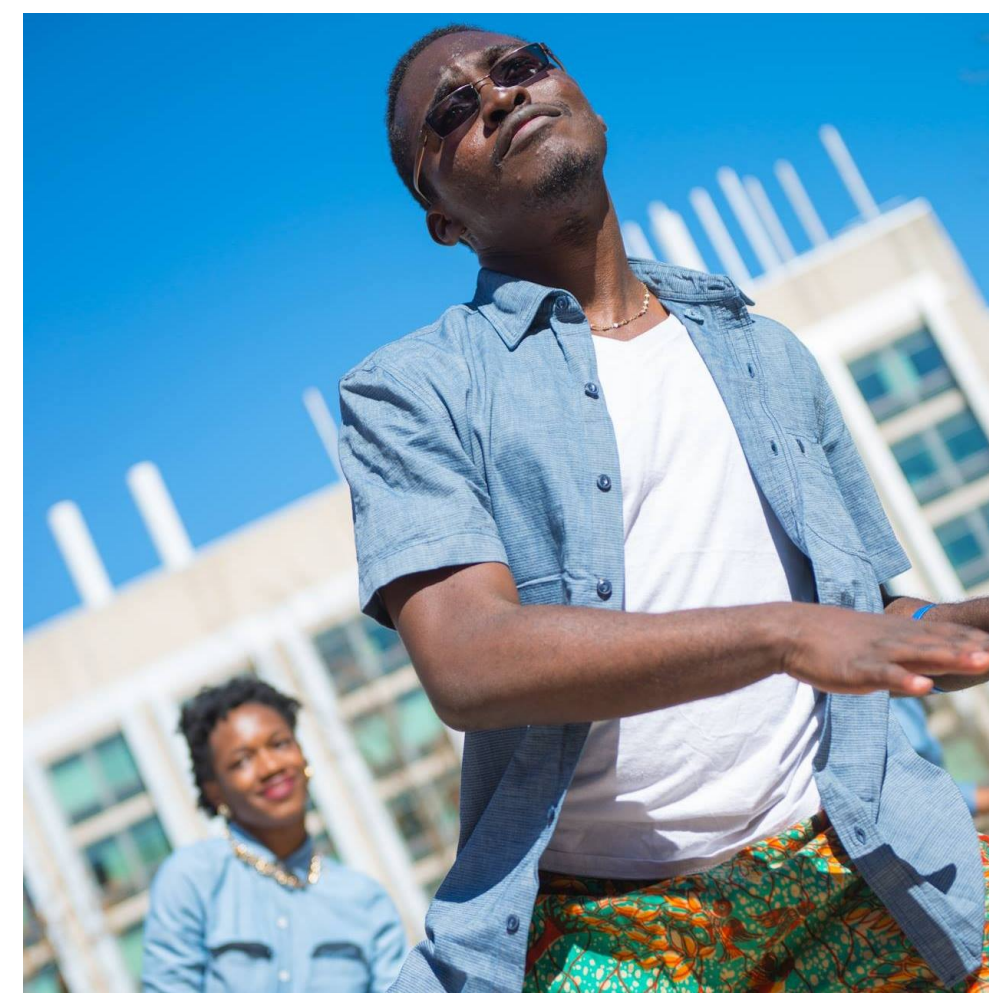
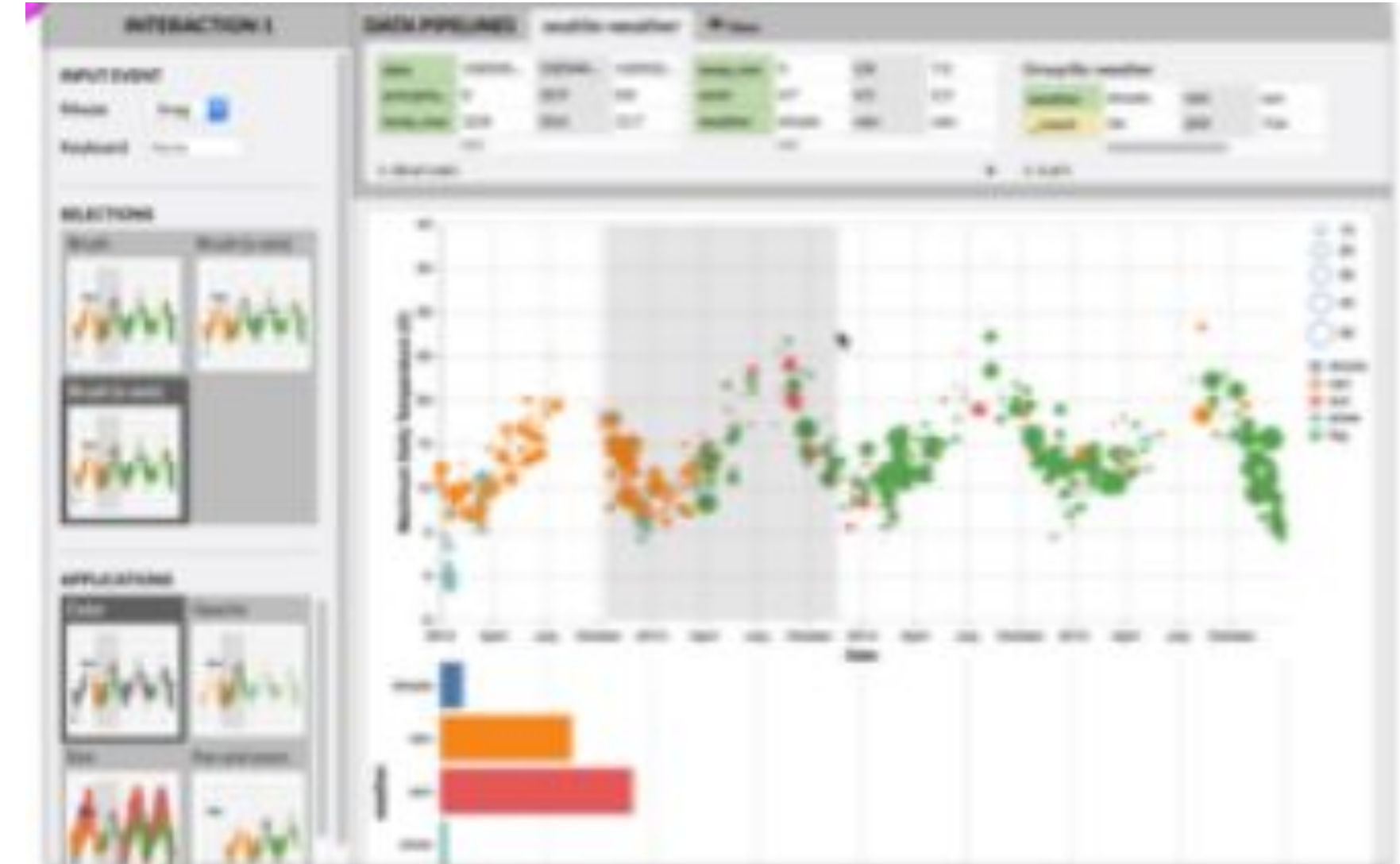
MEng @ D-Lab



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!



Running!



Hiking!

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%	Major visualization project on topic/ dataset of your choice.
Reading Commentaries	5%	
A0: Sketching Visualizations	2%	Second half of the course.
A1: Visualization Design	3%	
A2: Exploratory Data Analysis	10%	Teams of 1–4 people with a 1 page project proposal.
A3: White/Black Hat Visualization	15%	
A4: Interactive Narratives	20%	In-class presentations of minimal viable product (MVP). Peer review/critique.
Final Project	40%	Final presentations during the last class session (format TBD).
Proposal		
MVP + Presentations		
Poster Session + Final Deliverables		

Questions?

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