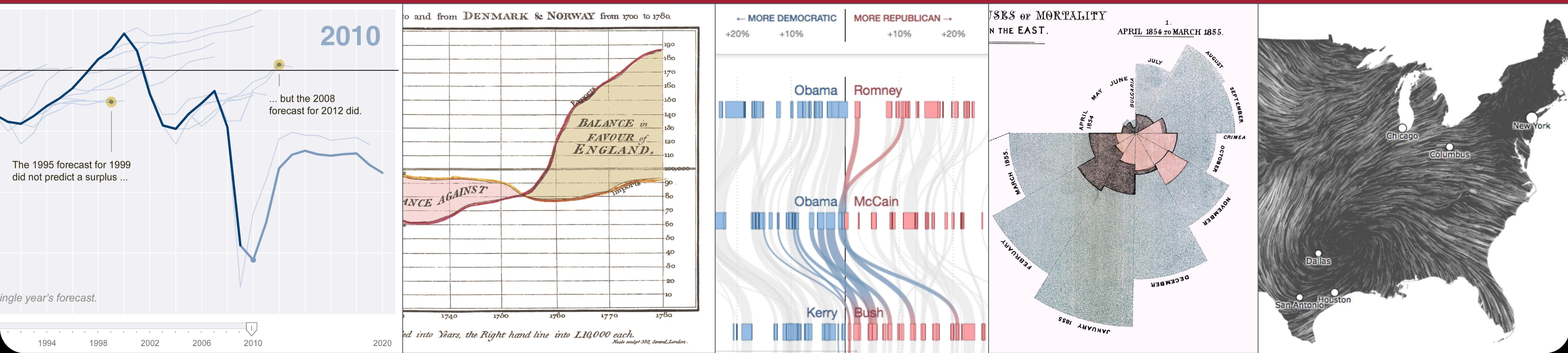


# 6.894: Interactive Data Visualization

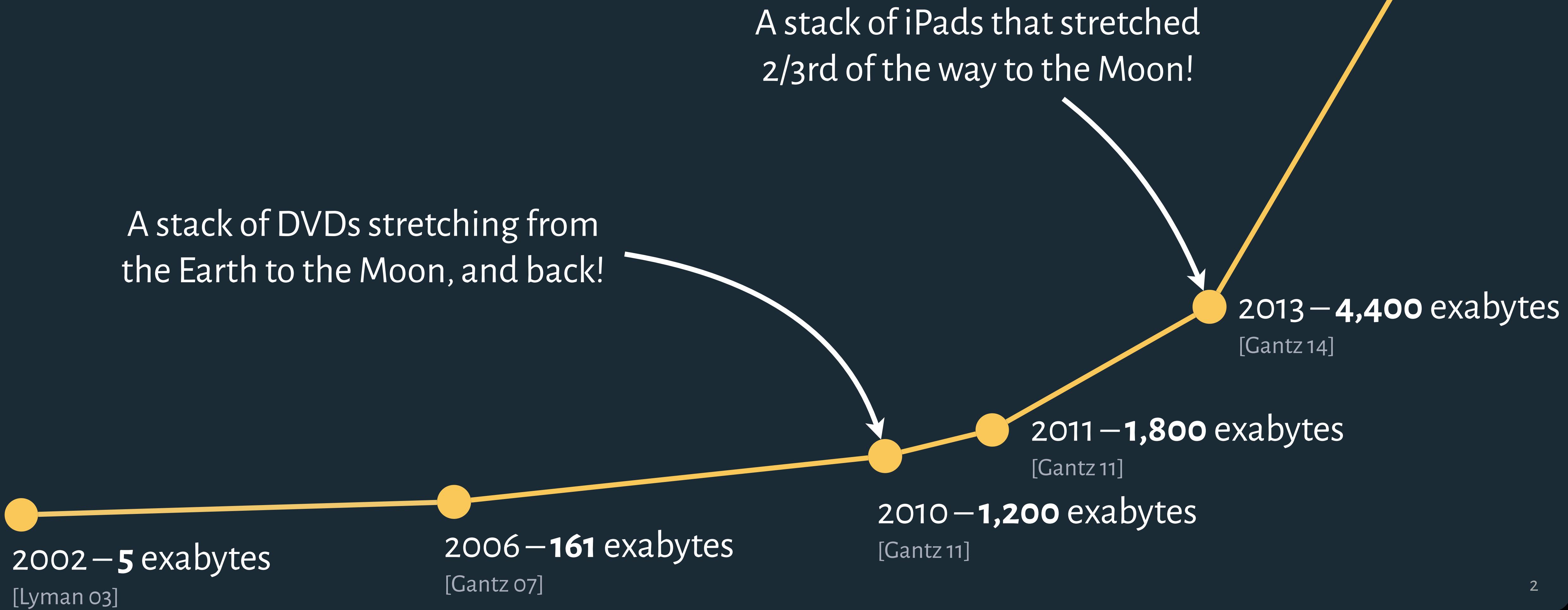
# The Value of Visualization

Arvind Satyanarayan

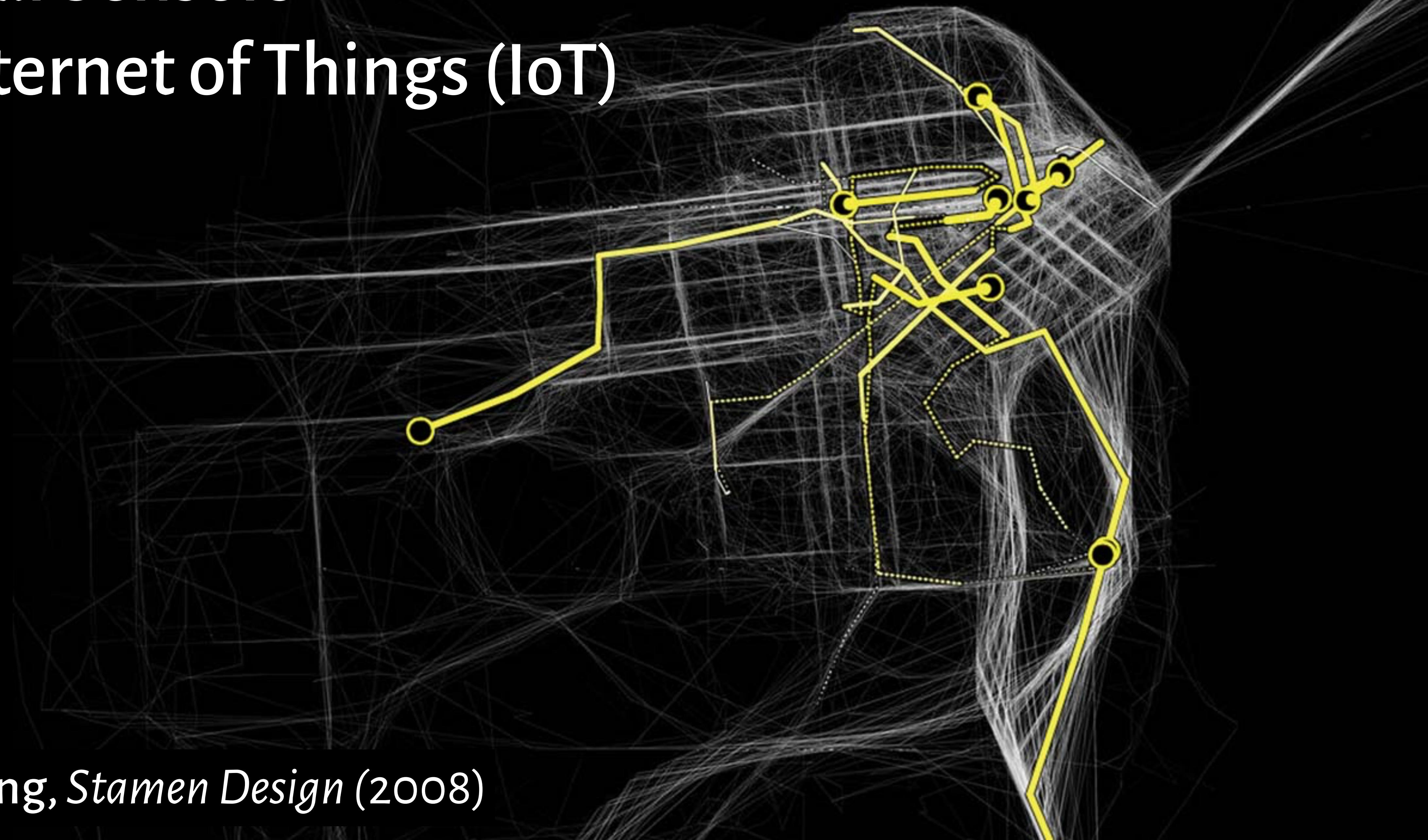


# How much data are we producing?

(1 exabyte = 1 *million* terabytes)



# Physical Sensors + The Internet of Things (IoT)



Cabspotting, Stamen Design (2008)

# Health & Medicine



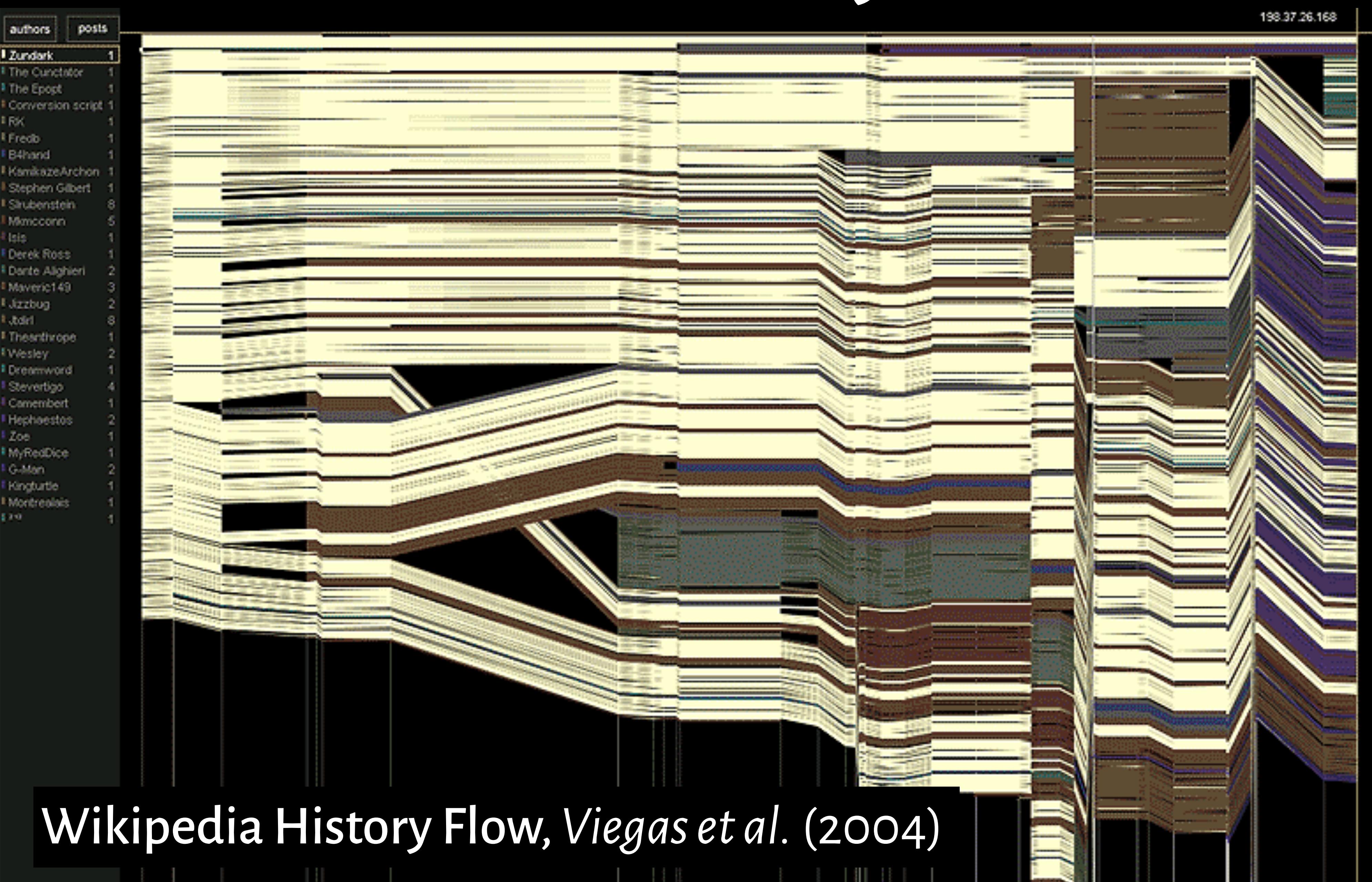
# Records of Human Activity



facebook

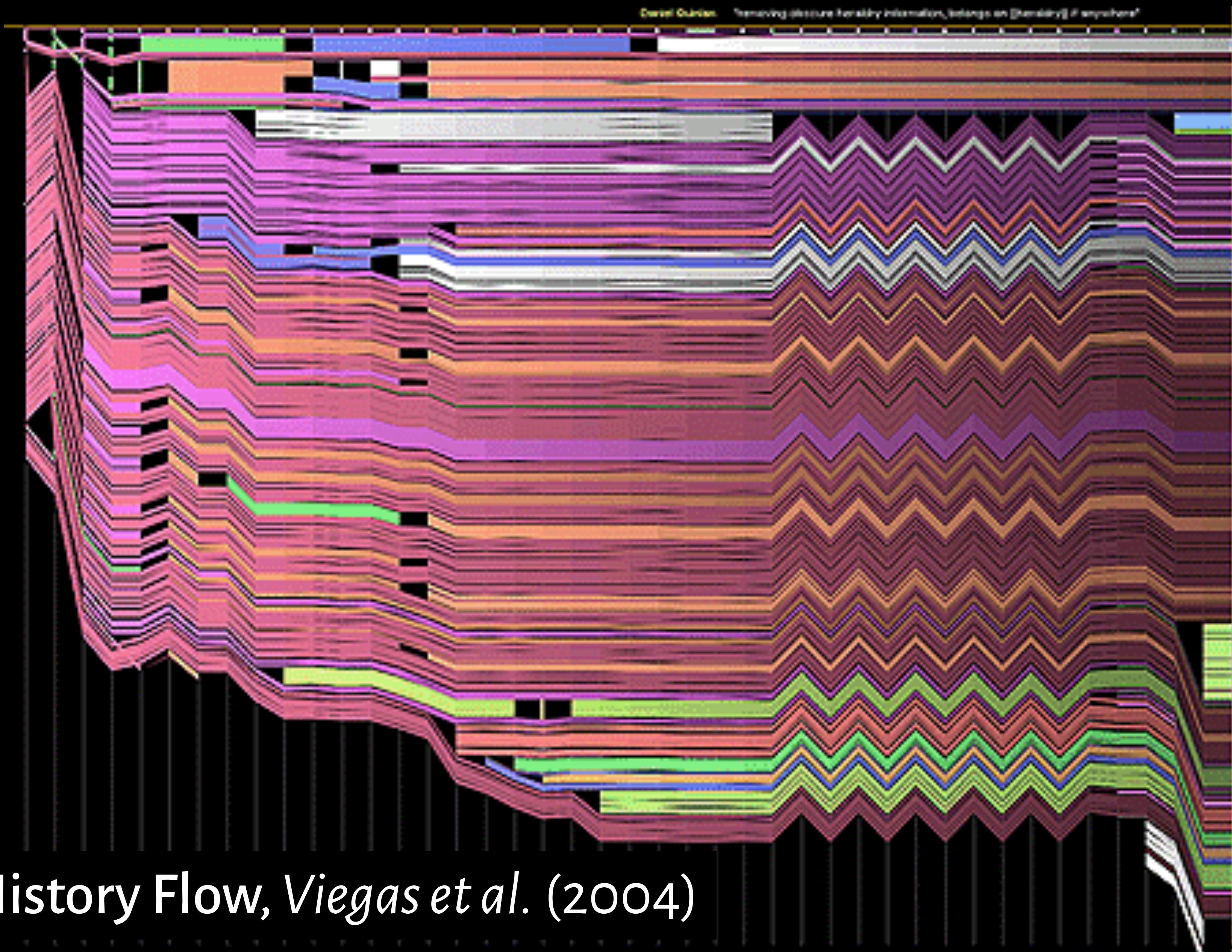
December 2010

# 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

# Machine Learning!

w really do have essentially free and ubiquitous data  
soft, cheap in many cases factors, 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?

understand that data and extract value from it."



Hal Varian, Google's Chief Economist

*The McKinsey Quarterly, Jan 2009*

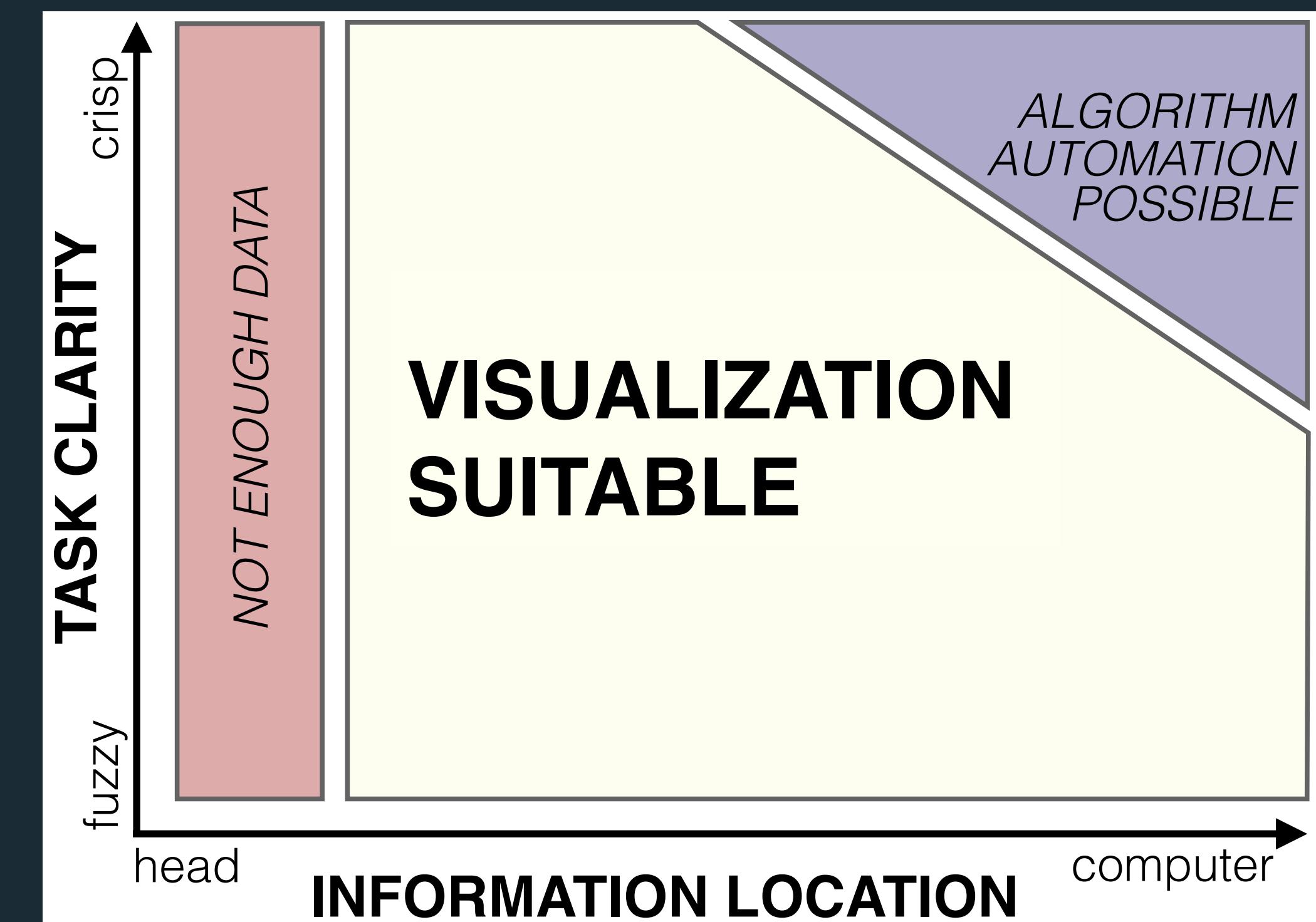
# Activity!

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?

## Process:

1. Think (2 minutes)
2. Pair (5 minutes)
3. Share



**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**

$$\bar{X} = 9.0 \quad \sigma_X = 3.317$$

$$\bar{Y} = 7.5 \quad \sigma_Y = 2.03$$

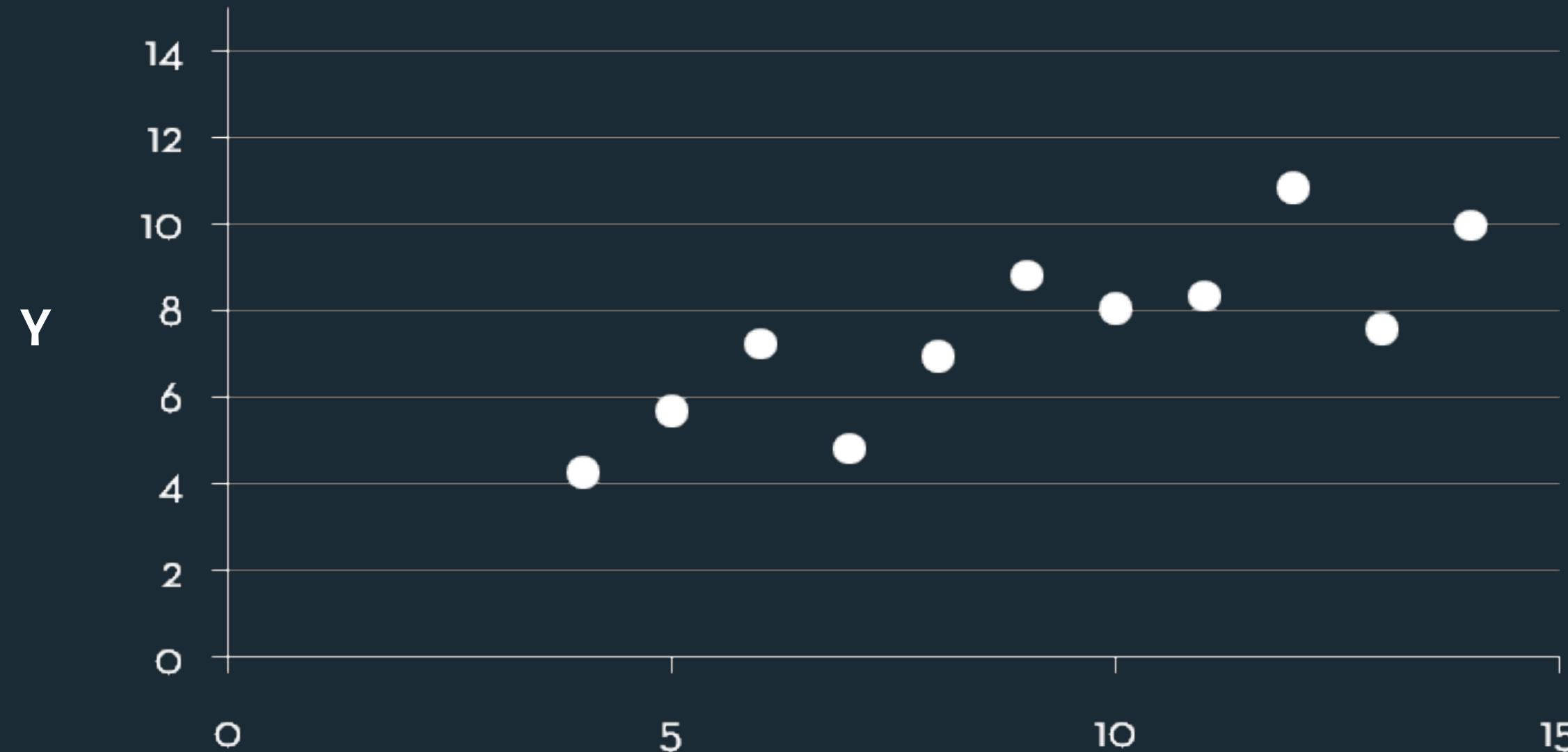
**Linear Regression**

$$Y = 3 + 0.5 X$$

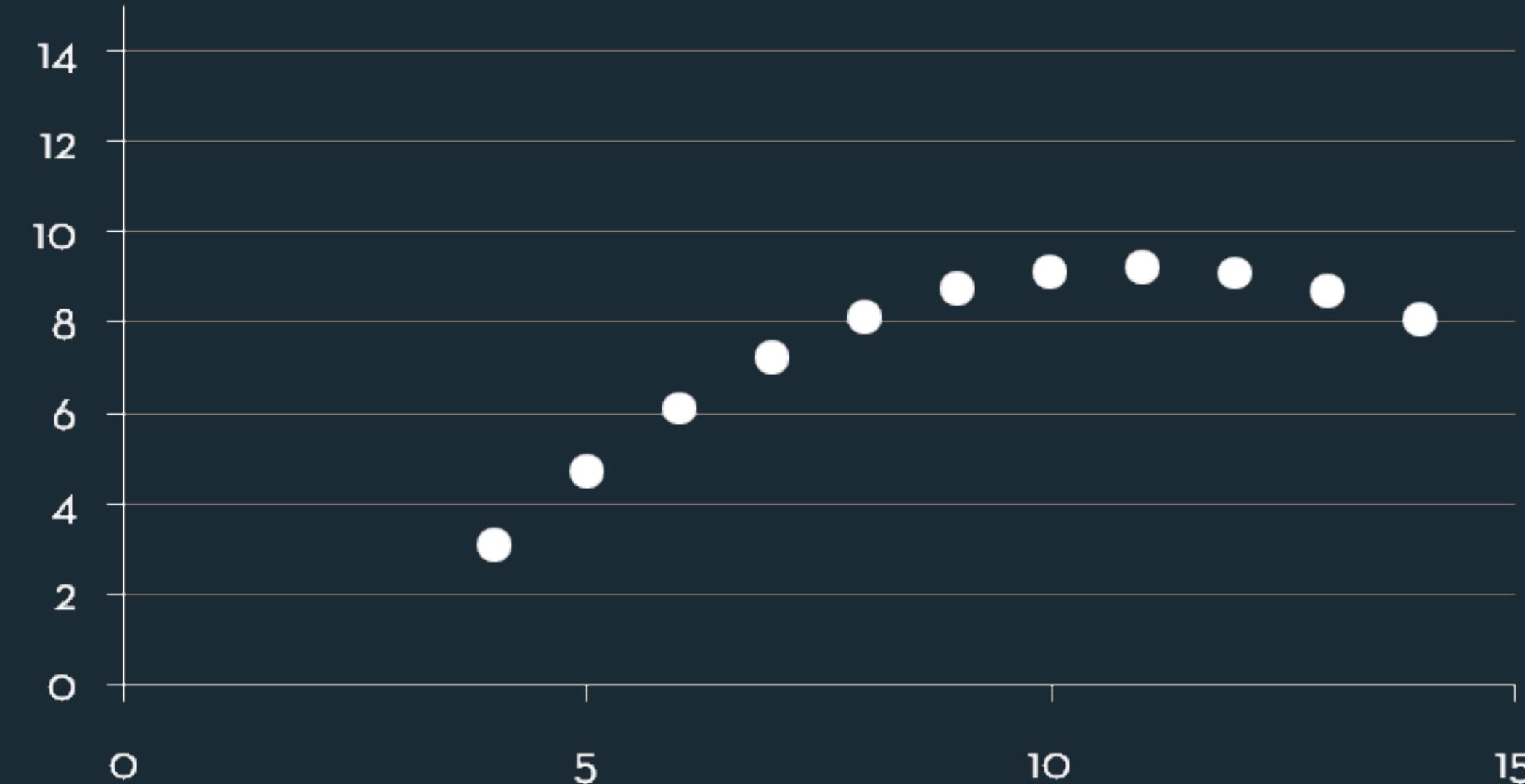
$$R^2 = 0.67$$

[Anscombe 1973]

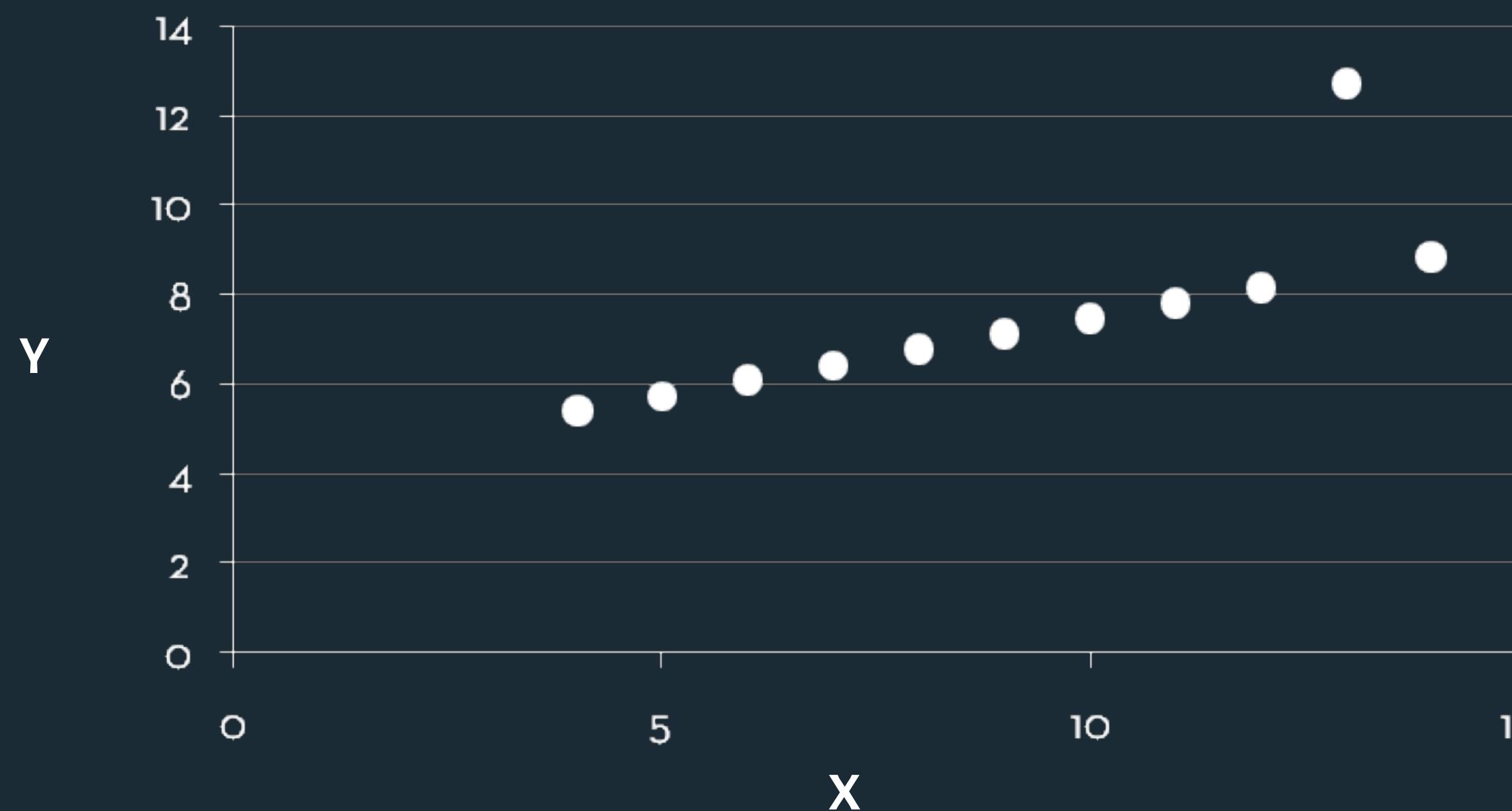
### Set A



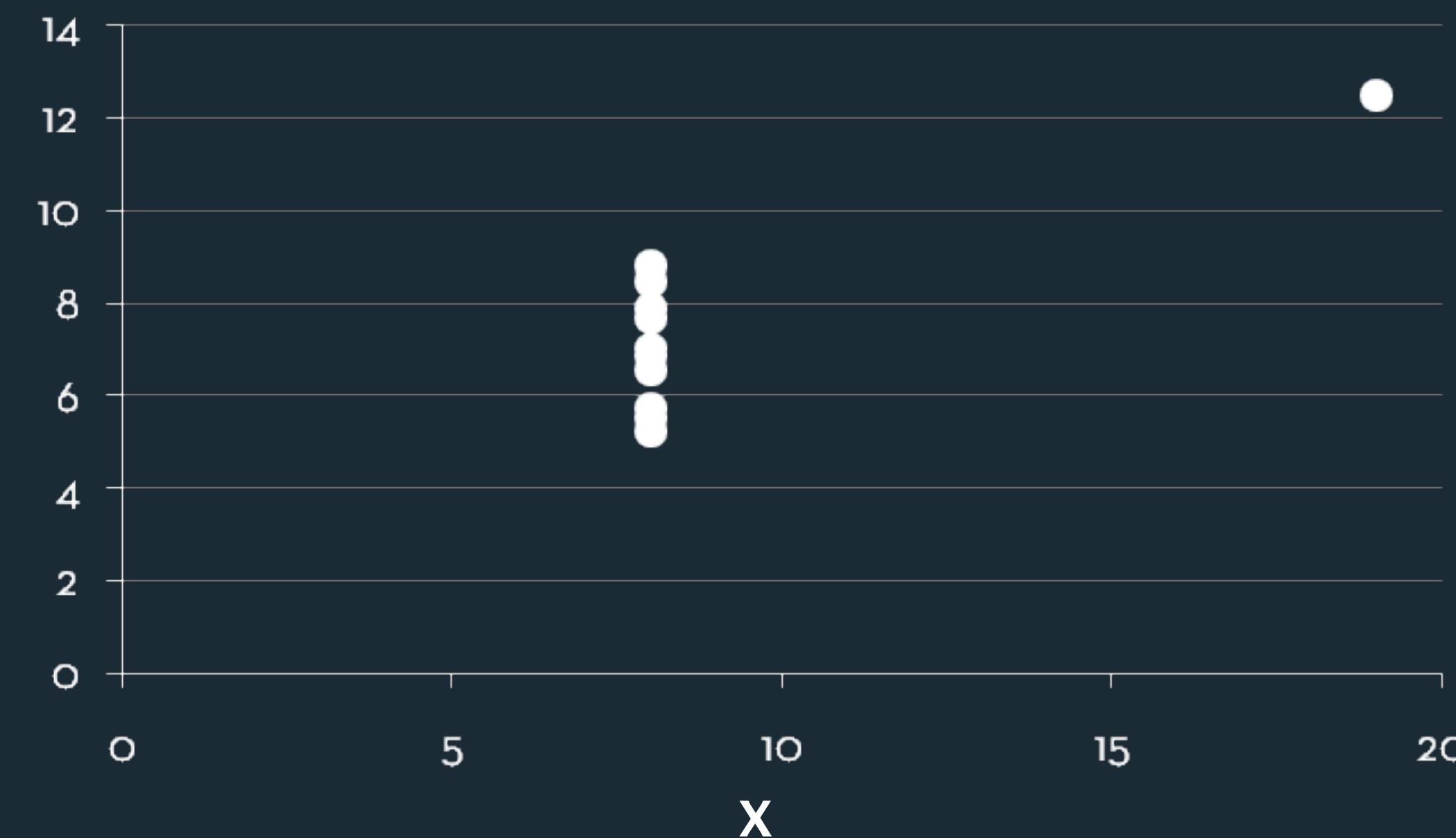
### Set B



### Set C



### Set D





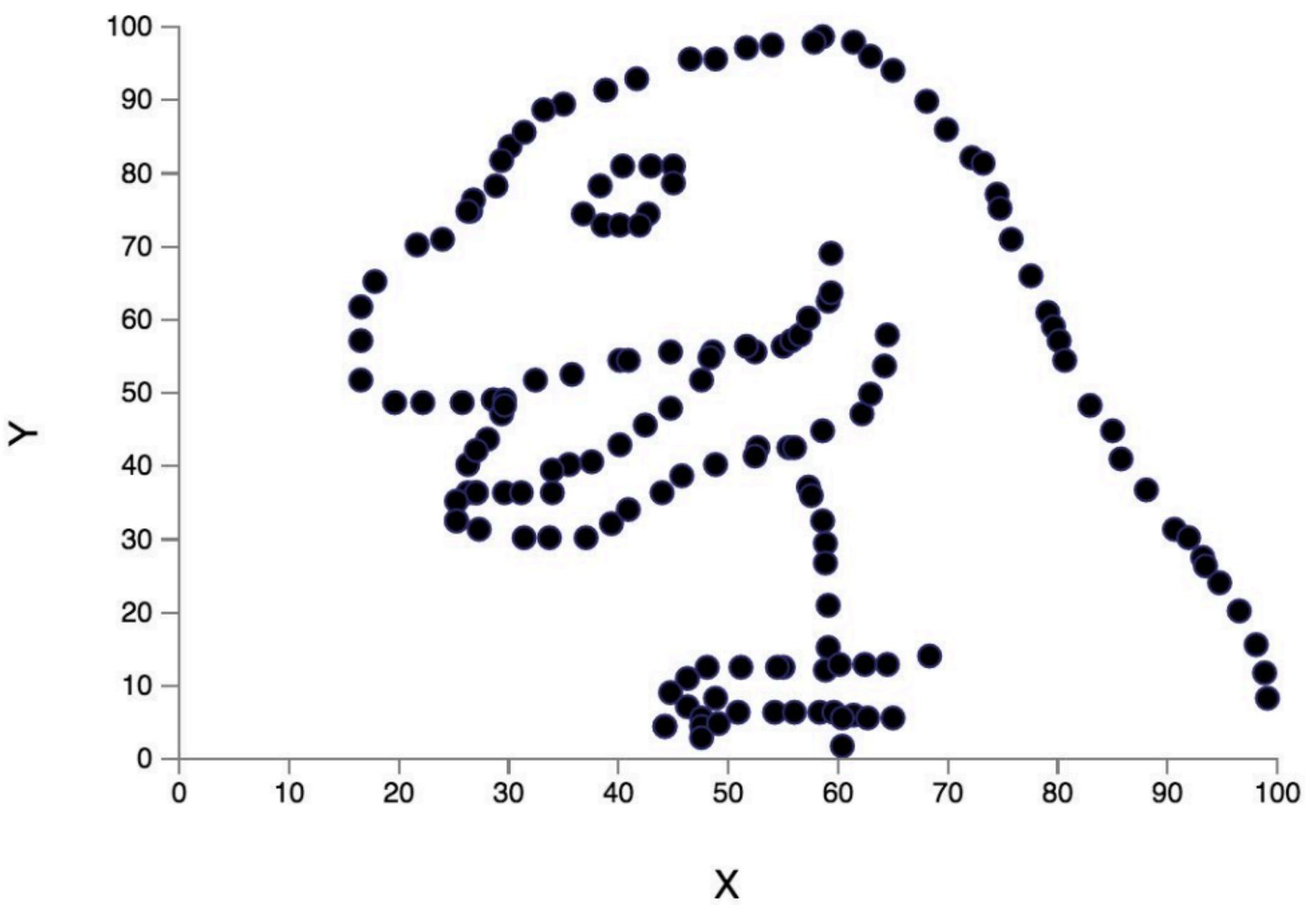
Alberto Cairo   
@albertocairo

Following



Don't trust summary statistics. Always  
visualize your data first  
[robertgrantstats.co.uk/drawmydata.html](http://robertgrantstats.co.uk/drawmydata.html)

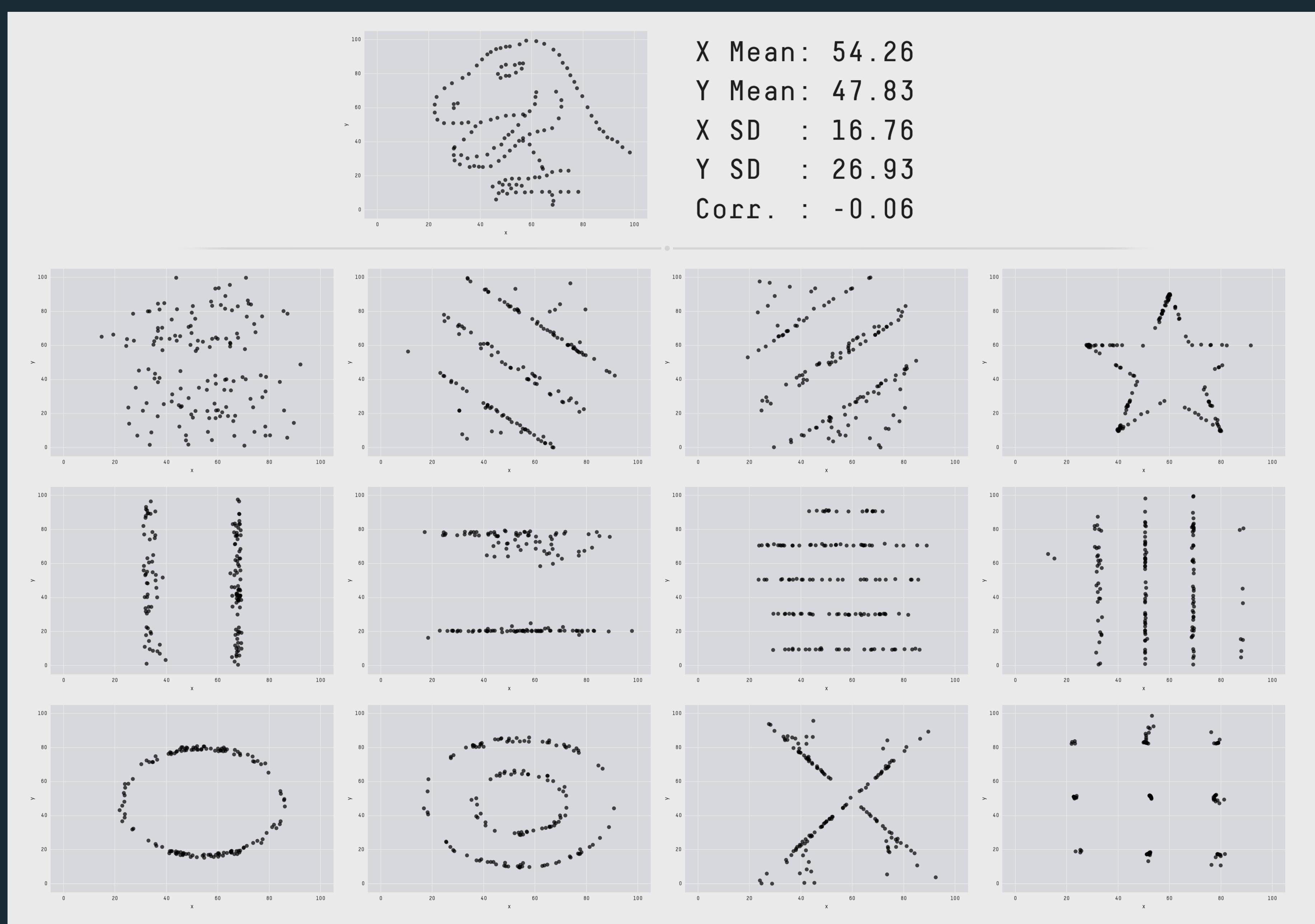
N = 157 ; X mean = 50.7333 ; X SD = 19.5661 ; Y mean = 46.495 ; Y SD = 27.2828 ;  
Pearson correlation = -0.1772



5:47 AM - 15 Aug 2016

952 Retweets 1,023 Likes





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

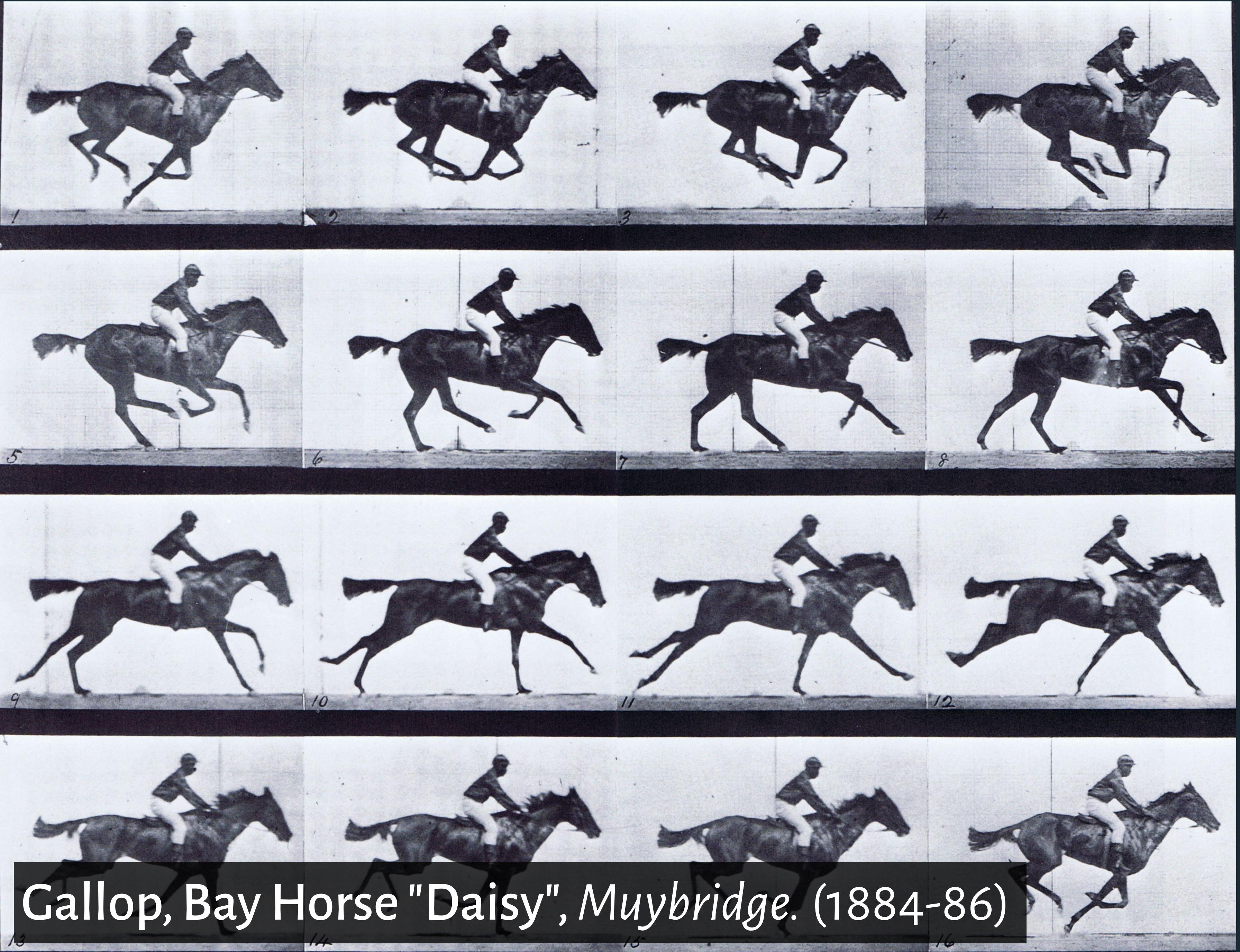
Expand memory

**Communicate** information to others (explanatory visualization)

Share and persuade

Collaborate and revise

# Record Info



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

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.

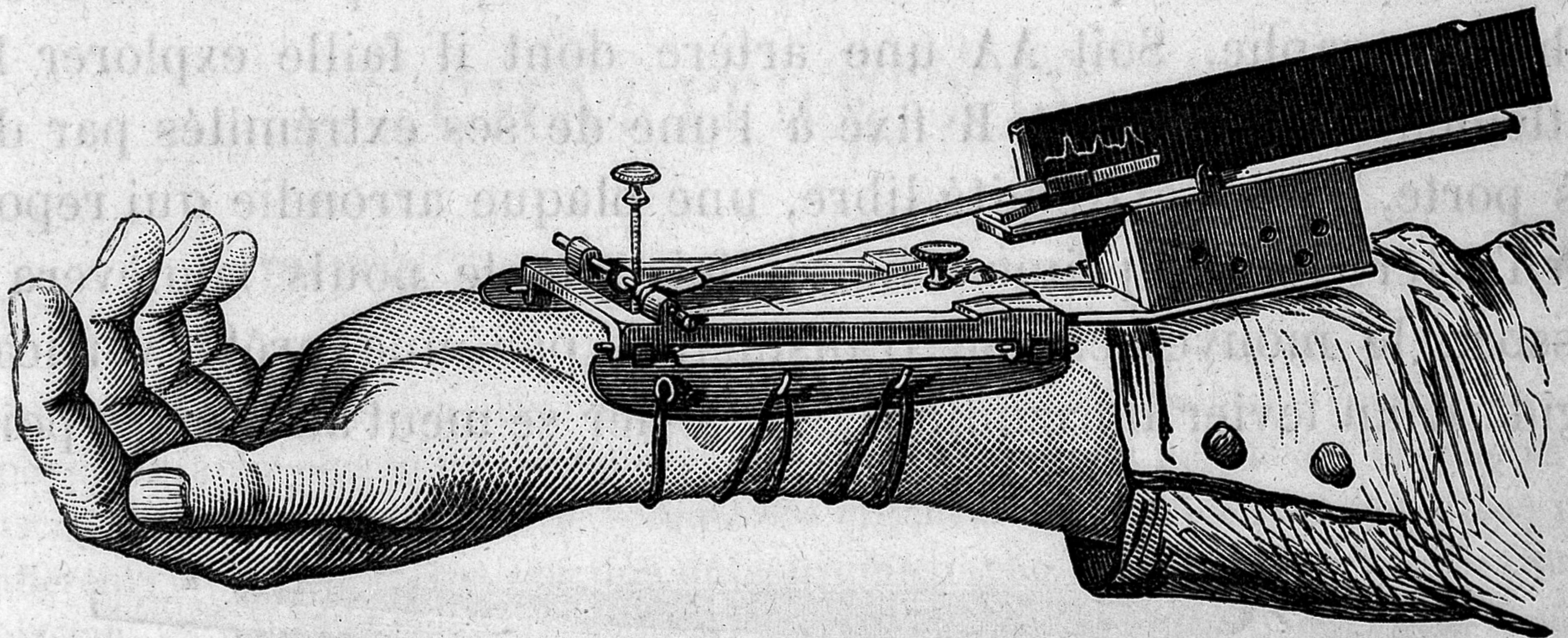
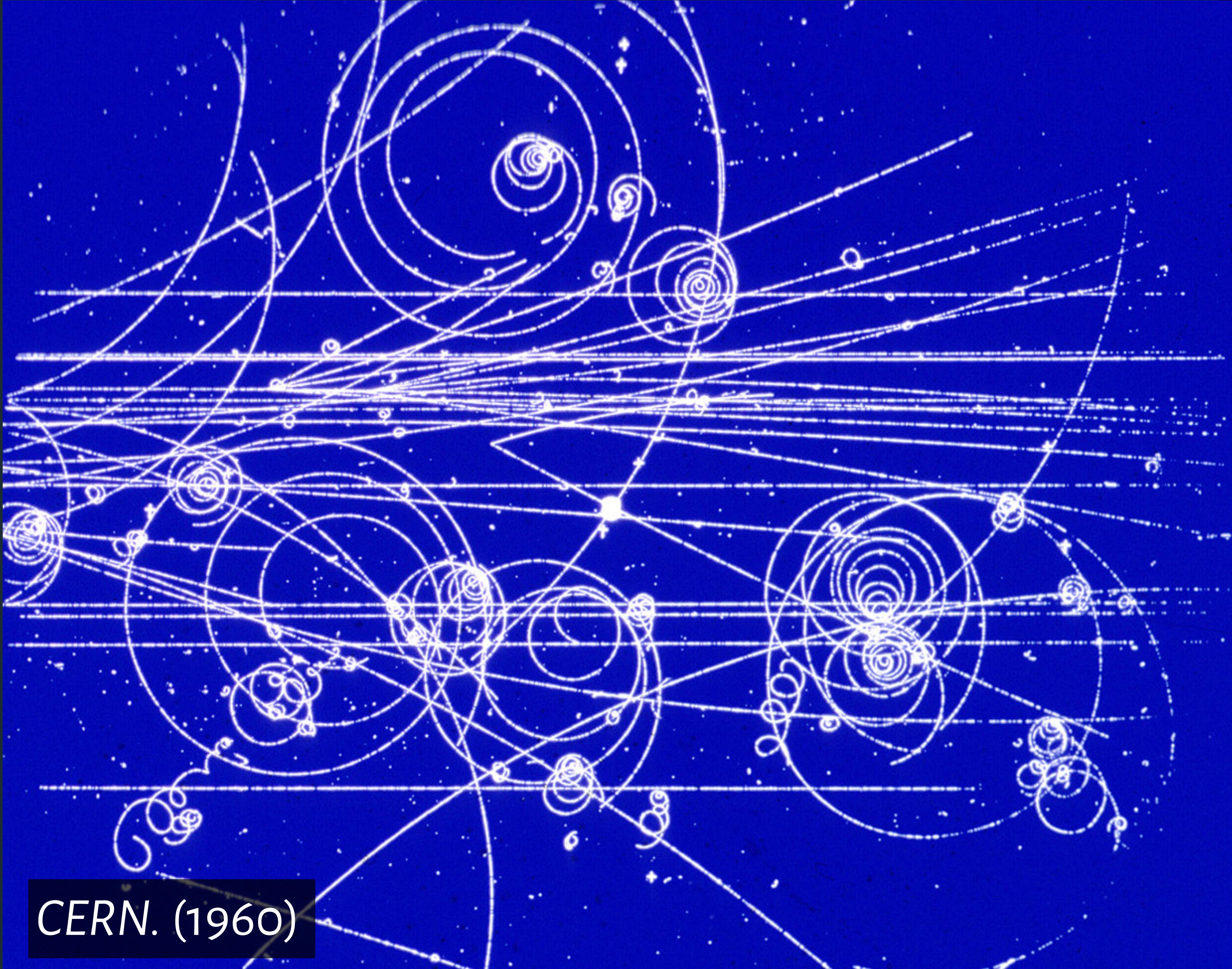


Fig. 109. Sphygmographe direct.

# Record Info

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



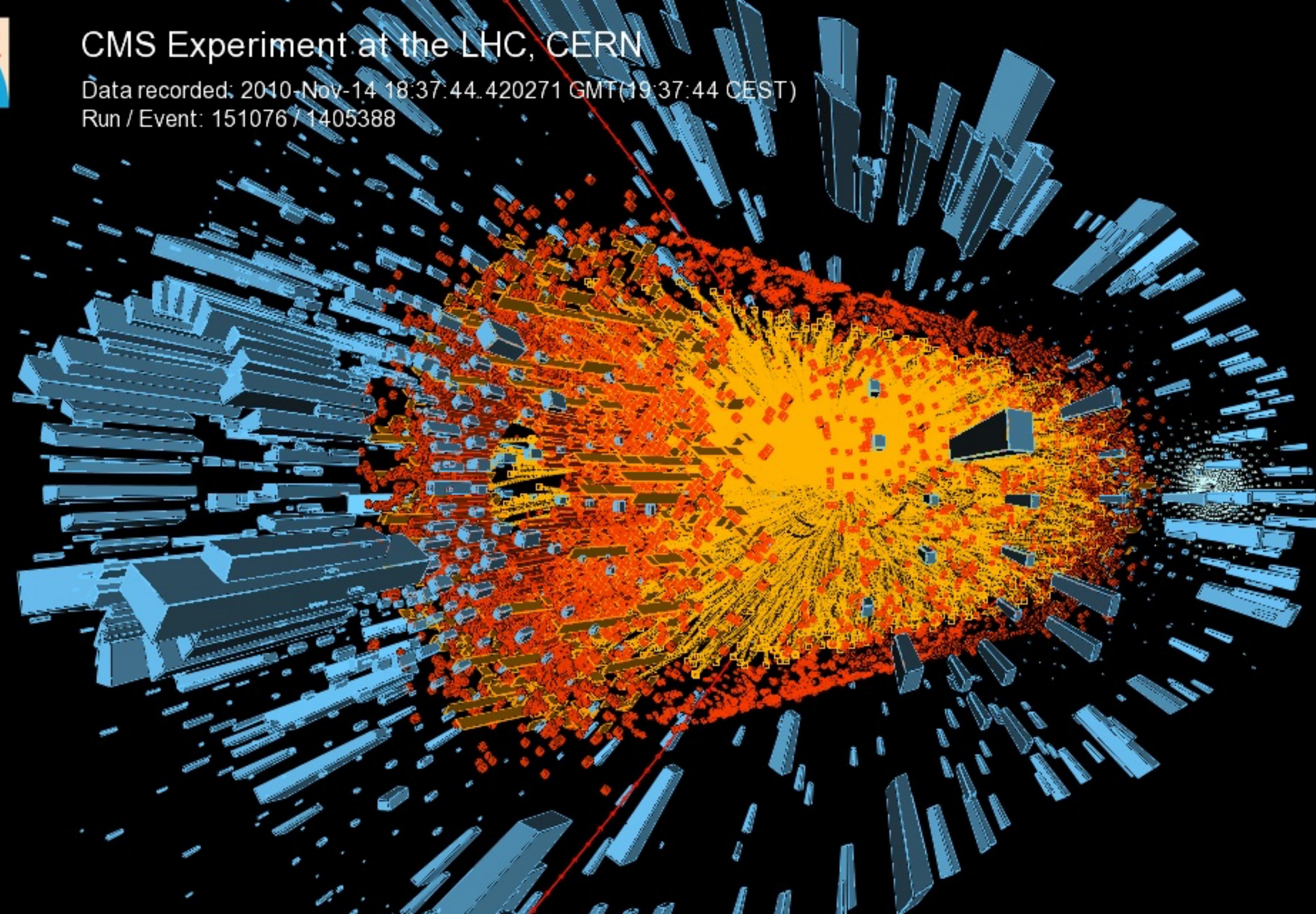
CERN. (1960)



# CMS Experiment at the LHC, CERN

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

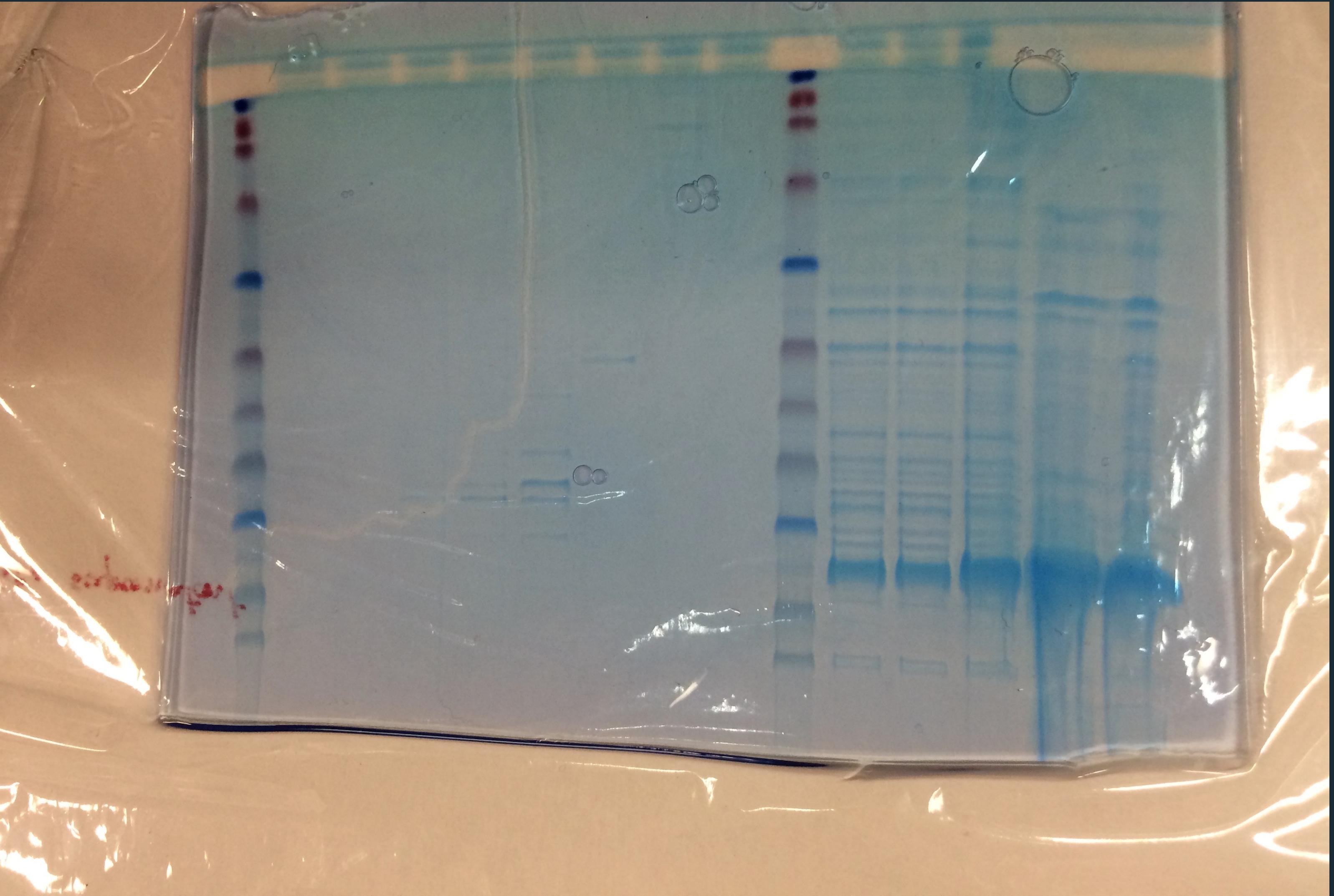
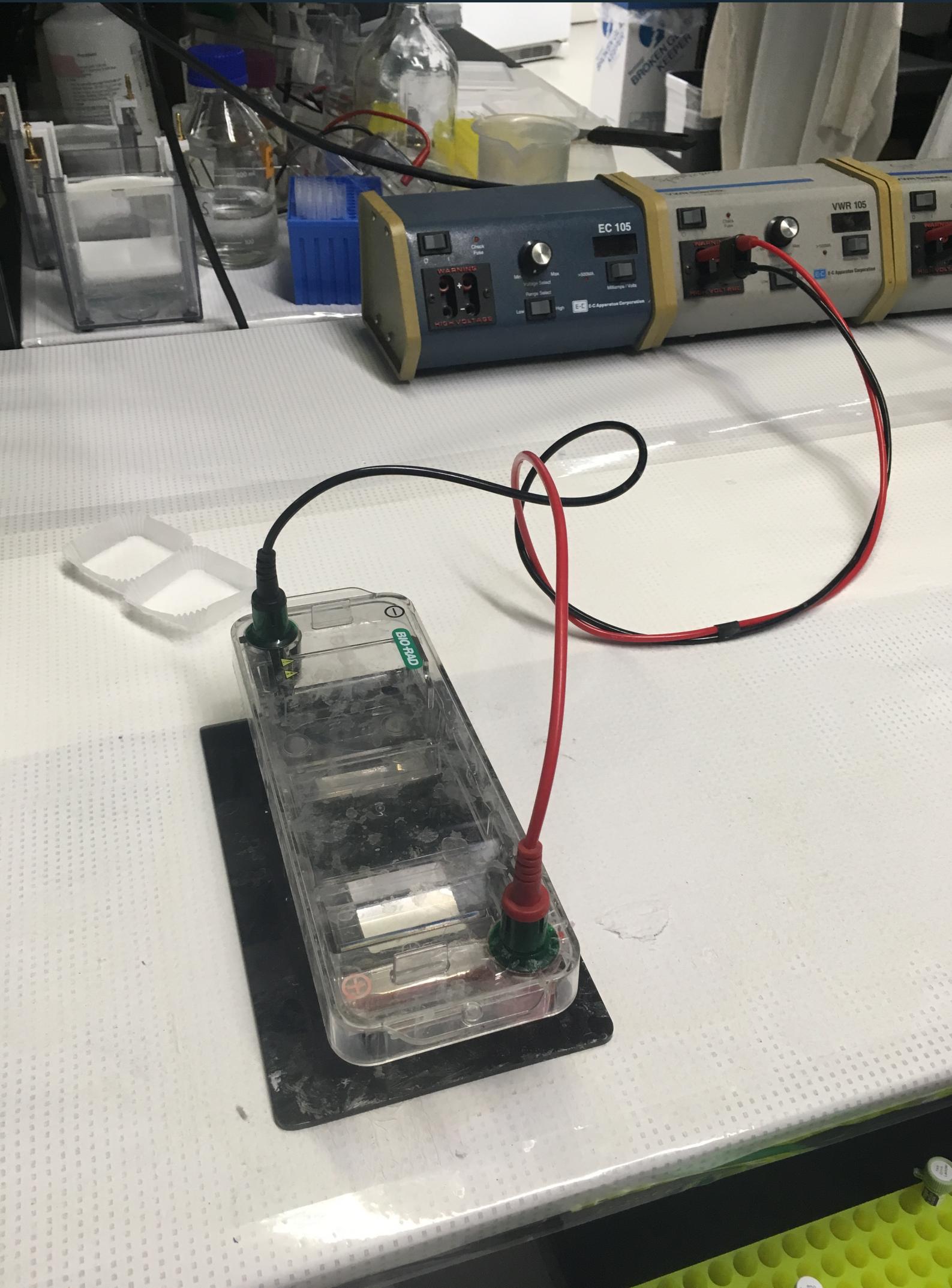
Run / Event: 151076 / 1405388



# Record Info

Gel electrophoresis is a technique to separate DNA, RNA, or proteins by their size.

The visualization *is* the data.



# 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

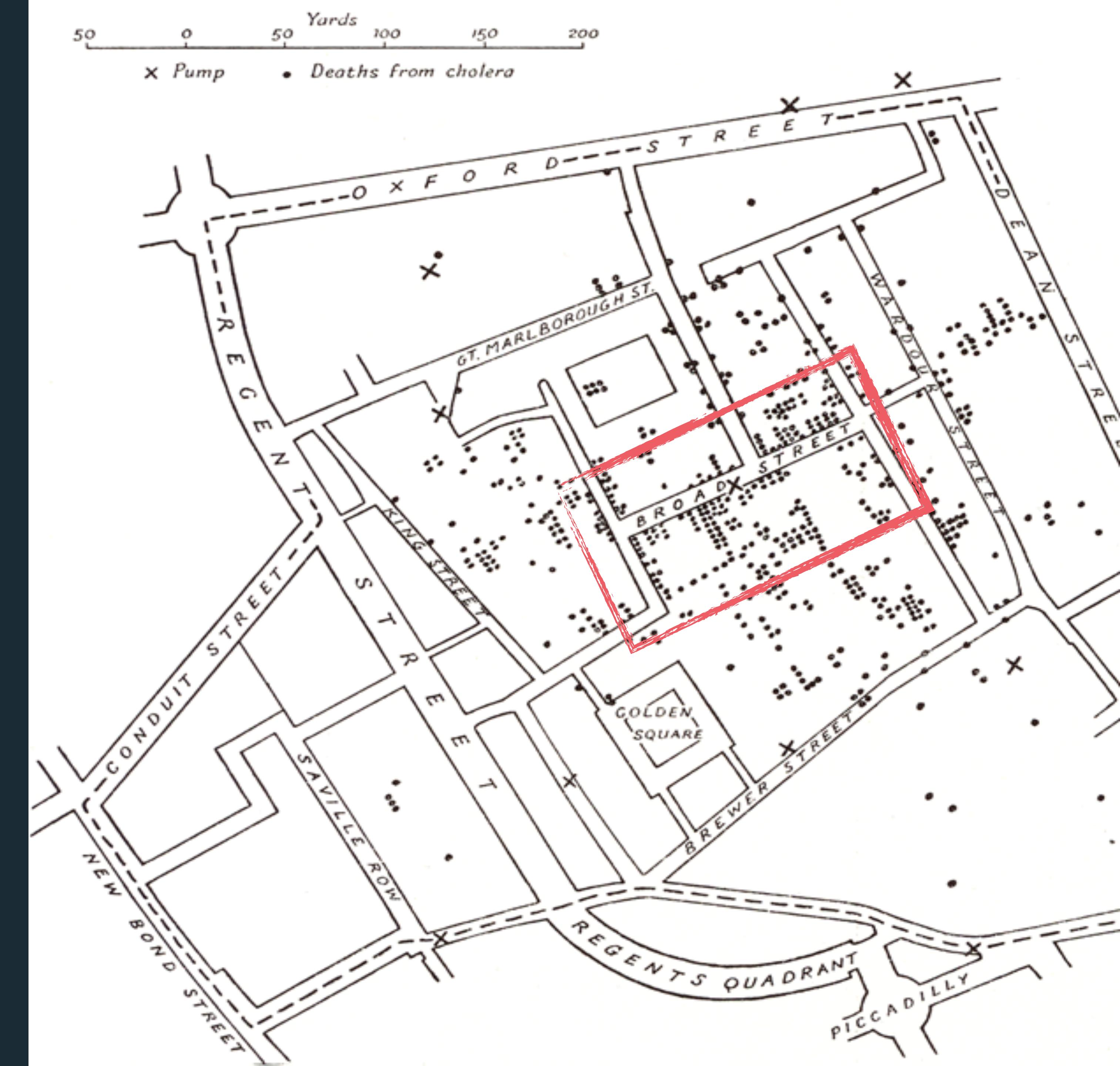
Collaborate and revise

# Support Reasoning

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

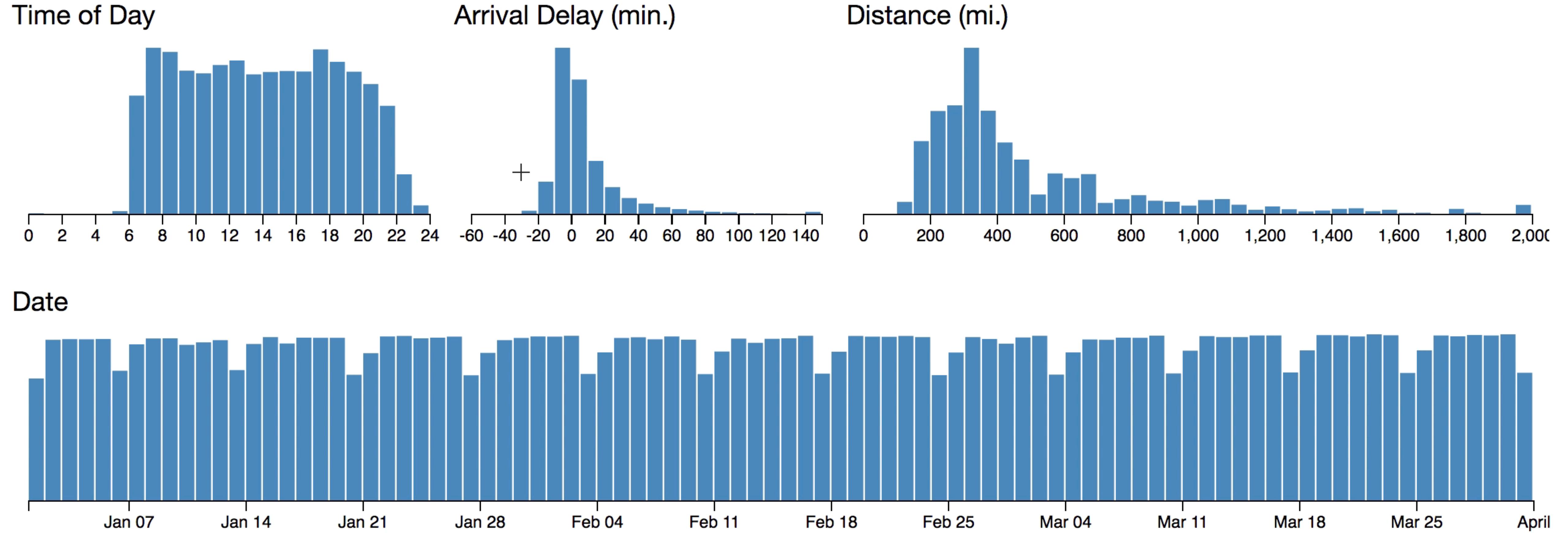
Map put the data in context.

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

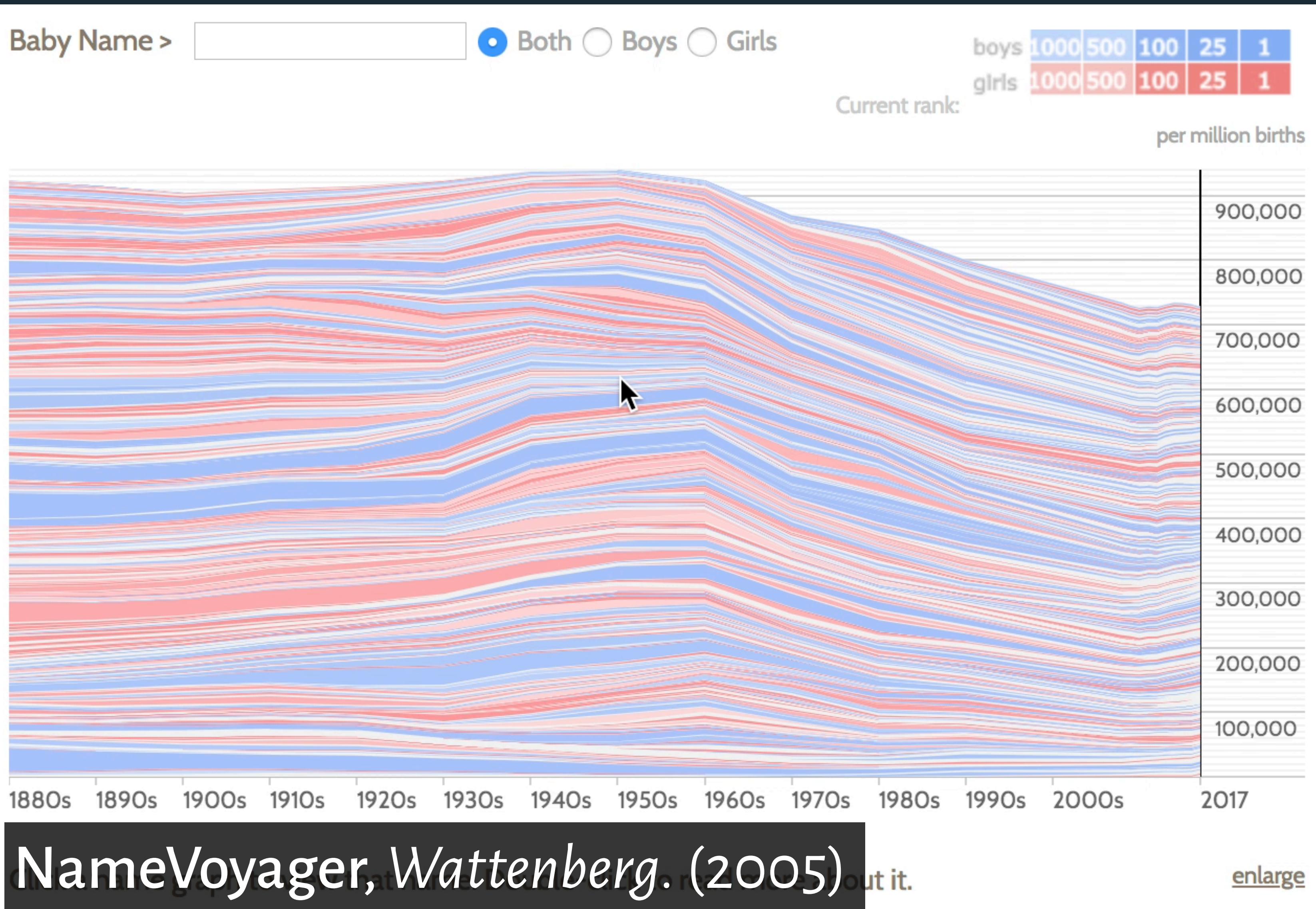


# Support Reasoning

Interactively generate and evaluate hypotheses.



# Support Reasoning



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](http://babynamewizard.com/voyager)

# Support Reasoning

## Class Exercise!

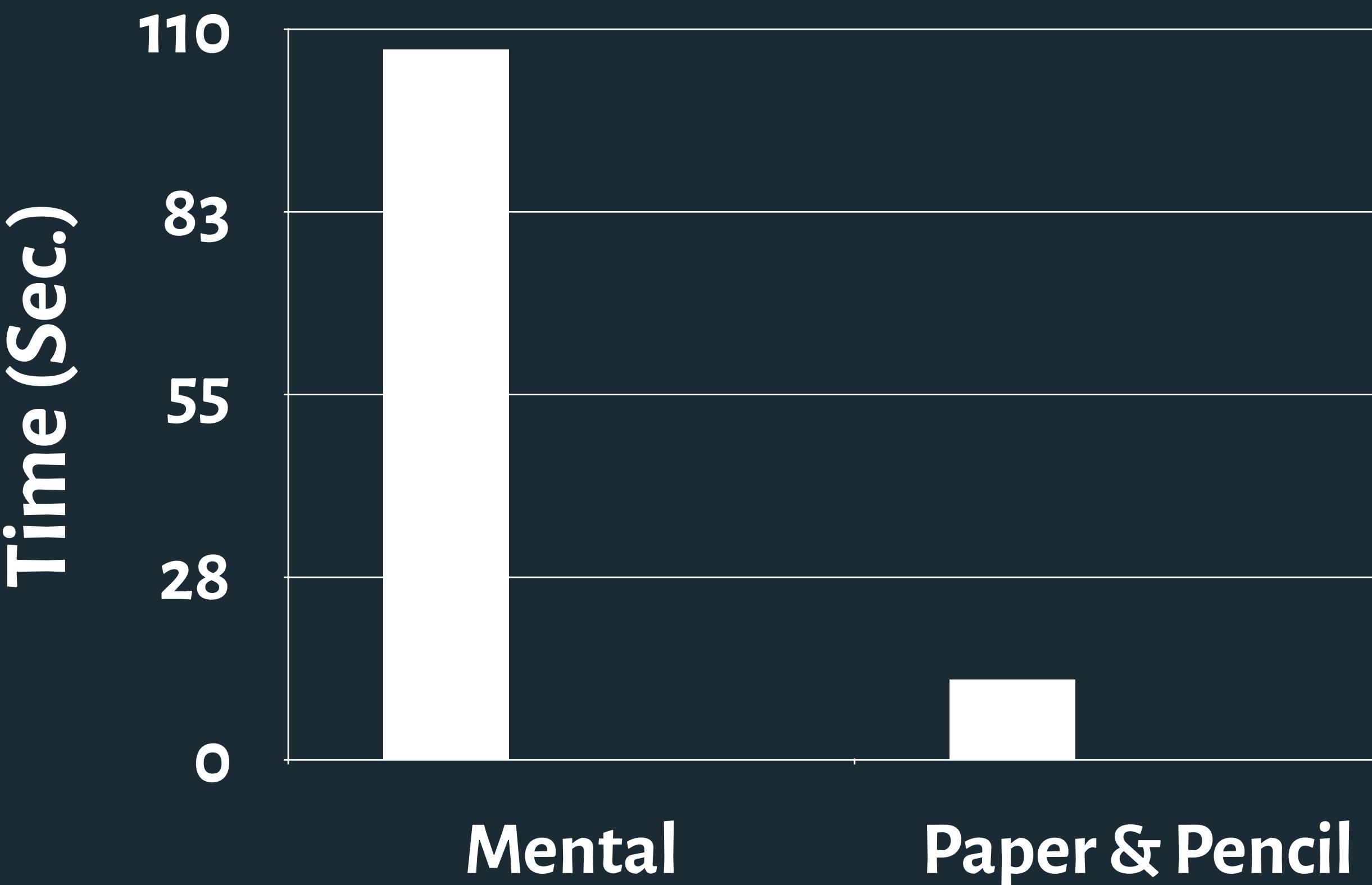
# Support Reasoning

34

x 72

# 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

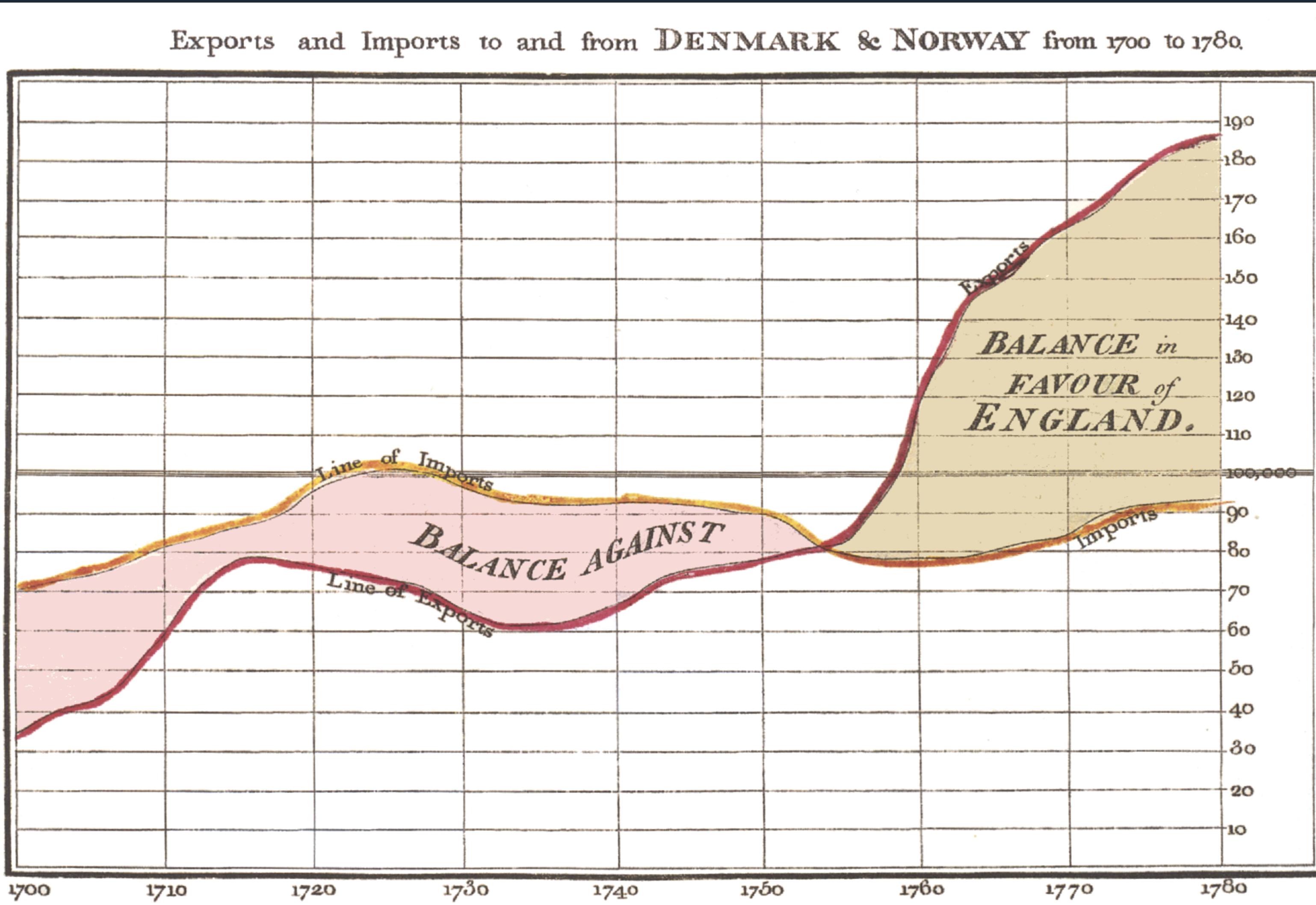
Expand memory

**Communicate** information to others (explanatory visualization)

Share and persuade

Collaborate and revise

# Communicate Info



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

*Published as the Act directs, 1<sup>st</sup> May 1786, by W<sup>m</sup> Playfair*

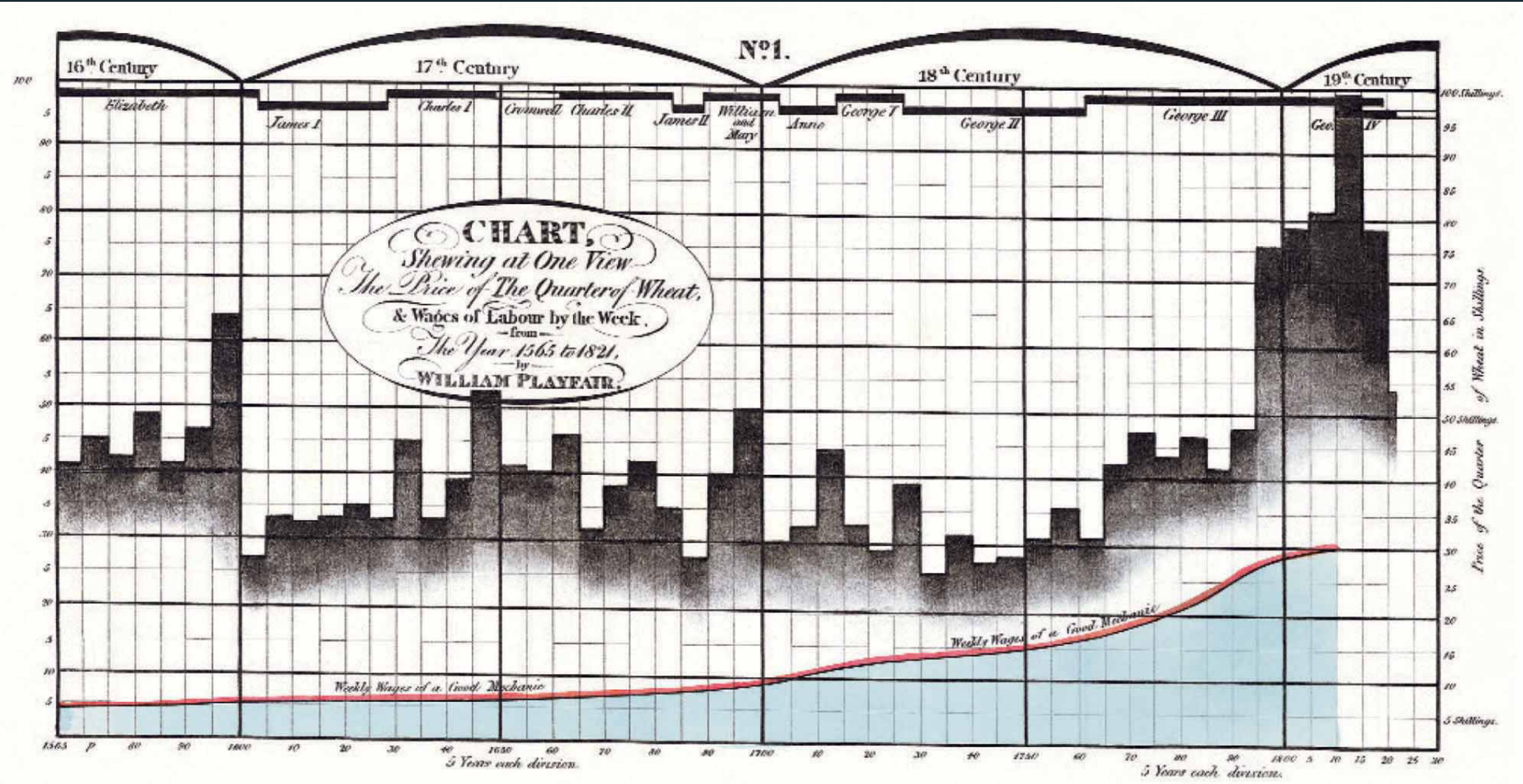
*Neale sculpt<sup>r</sup> 352, Strand, London.*

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

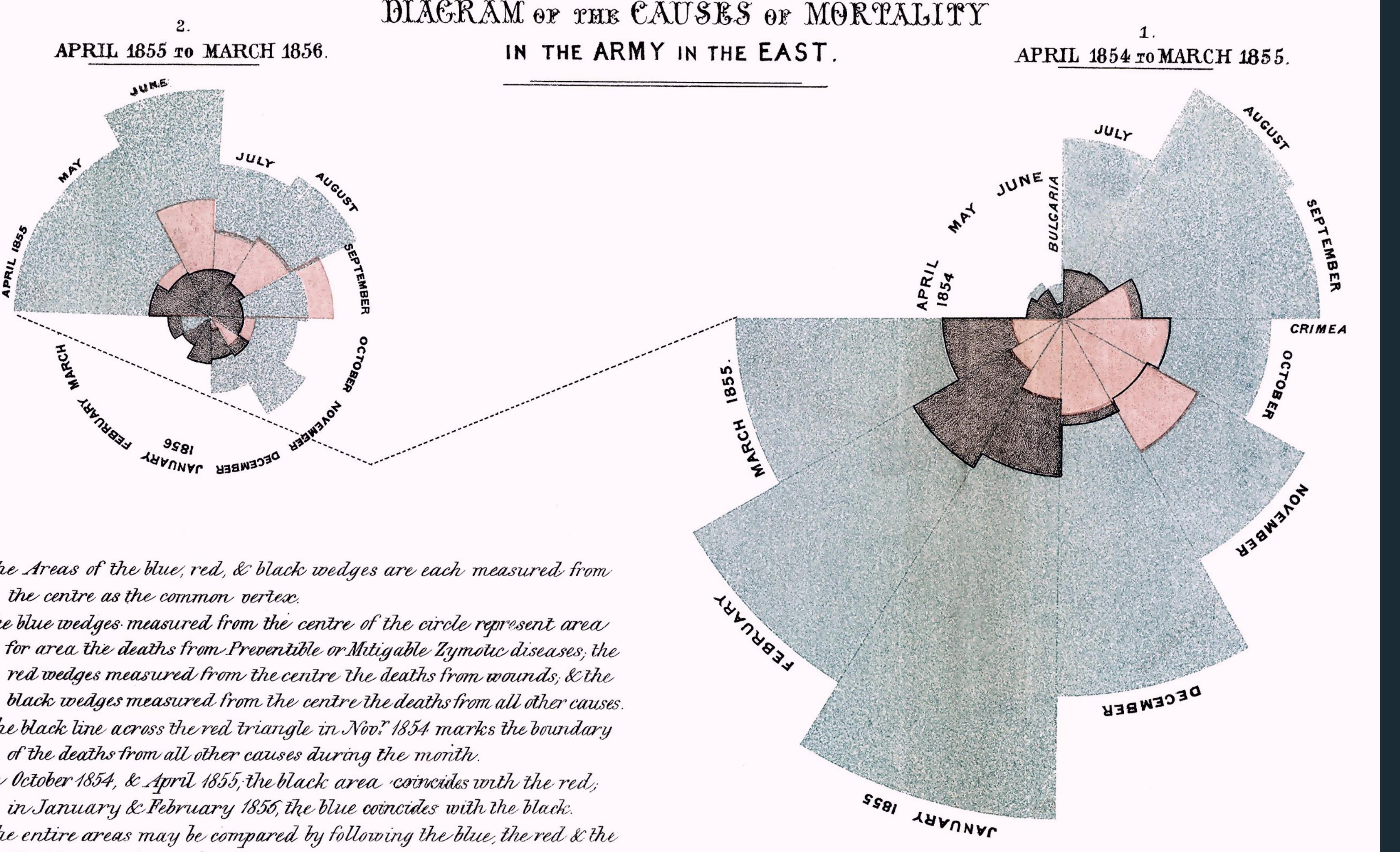
In 1786, published *The Commercial and Political Atlas* which contained the first time-series and bar charts.

# 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



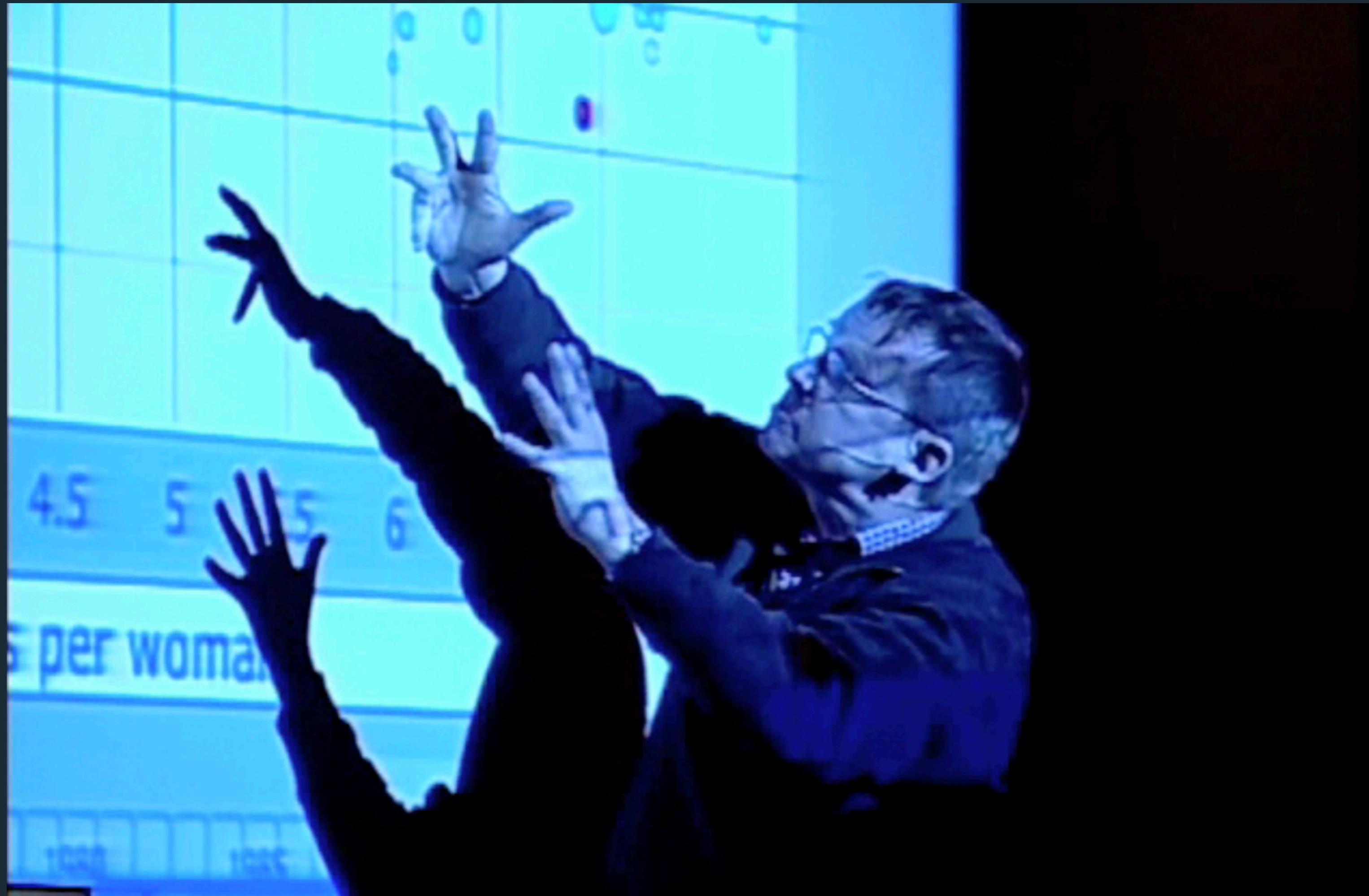
*"to affect thro' the Eyes what we fail to convey to the public through their word-proofears"*

— 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*

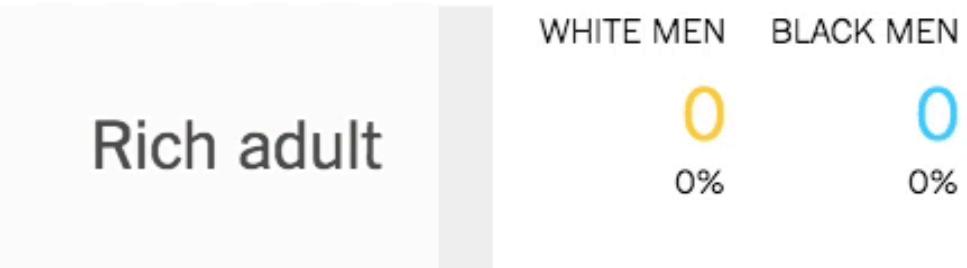
By EMILY BADGER, CLAIRE CAIN MILLER, ADAM PEARCE and KEVIN QUEALY MARCH 19, 2018

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

Grew up rich

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.

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



Upper-middle-class adult



Middle-class adult



Lower-middle-class adult



Poor adult



## 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

320-115

300-135

280-155

260-175

240-195

MAJORITY

240-195

260-175

280-155

Democrats win both the popular vote and the House

Democrats win the popular vote, but Republicans win the House

GOP wins both

D+20

D+15

D+10

D+5

EVEN

R+5

Popular vote margin

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.mit.edu/classes/6.894/>

# Course Goals

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

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

# Course Topics

# Course Topics

## Data & Image Models

### LES VARIABLES DE L'IMAGE

	POINTS	LIGNES	ZONES
XY 2 DIMENSIONS DU PLAN	x x x	/ \ / \ / \	15 9 14 1 18 21 2 14 15 1 2 16 2 1 21 1 1 2 9
Z TAILLE	■ ■ ■	— — —	■ ■ ■
VALEUR	■ ■ ■	— — —	■ ■ ■

### LES VARIABLES DE SÉPARATION DES IMAGES

	GRAIN	COULEUR	ORIENTATION	FORME
GRAIN	■ ■ ■	■ ■ ■	■ ■ ■	■ ■ ■
COULEUR	■ ■ ■	■ ■ ■	■ ■ ■	■ ■ ■
ORIENTATION	■ ■ ■	■ ■ ■	■ ■ ■	■ ■ ■
FORME	■ ■ ■	■ ■ ■	■ ■ ■	■ ■ ■

# Course Topics

## Visual Encoding with Vega-Lite & Altair

01\_marks\_encoding.ipynb - Colaboratory

File Edit View Insert Runtime Tools Help

CODE TEXT CELL CELL COPY TO DRIVE CONNECT EDITING

**X**

The x encoding channel sets a mark's horizontal position (x-coordinate). In addition, default choices of axis and title are made automatically. In the chart below, the choice of a quantitative data type results in a continuous linear axis scale:

```
[ ] alt.Chart(data2000).mark_point().encode(
    alt.X('fertility:Q')
)
```

**Y**

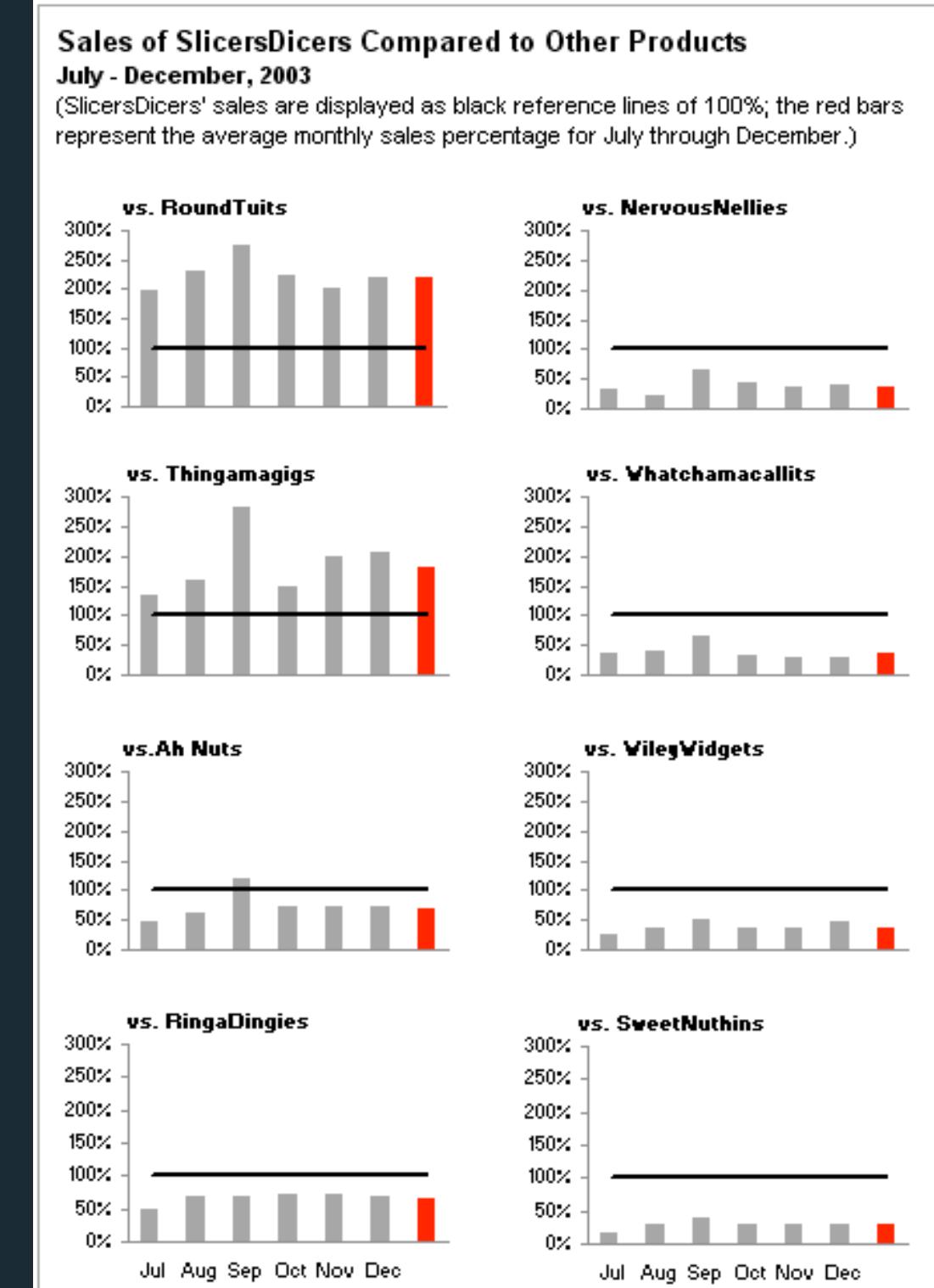
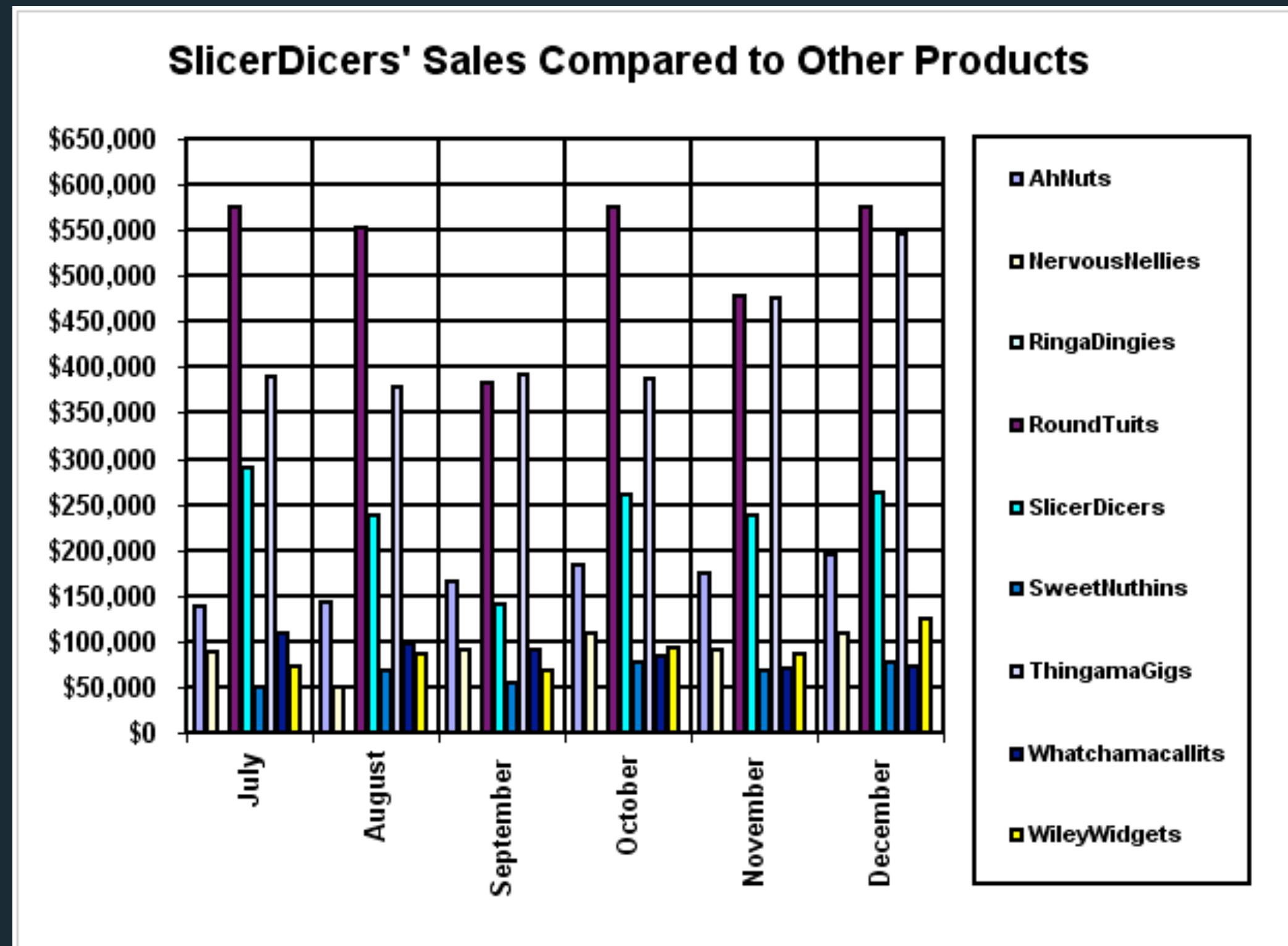
The y encoding channel sets a mark's vertical position (y-coordinate). Here we've added the cluster field using an ordinal (O) data type. The result is a discrete axis that includes a sized band, with a default step size, for each unique value:

```
[ ] alt.Chart(data2000).mark_point().encode(
    alt.X('fertility:Q'),
    alt.Y('cluster:O')
)
```

If we instead add the life\_expect field as a quantitative (Q) variable, the result is a scatter plot with linear scales for both axes:

# Course Topics

## (Re-)Design

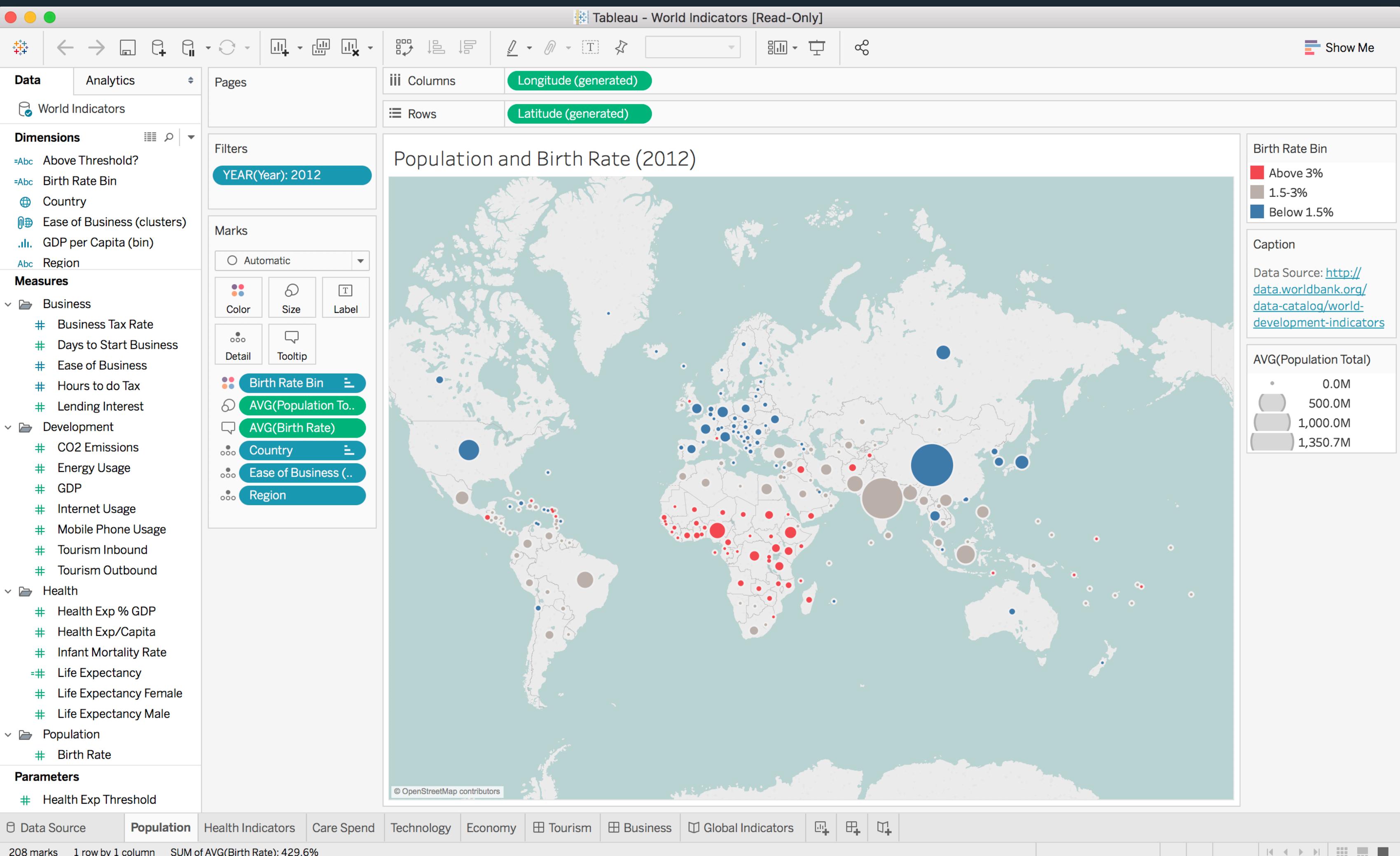


Problematic Design

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

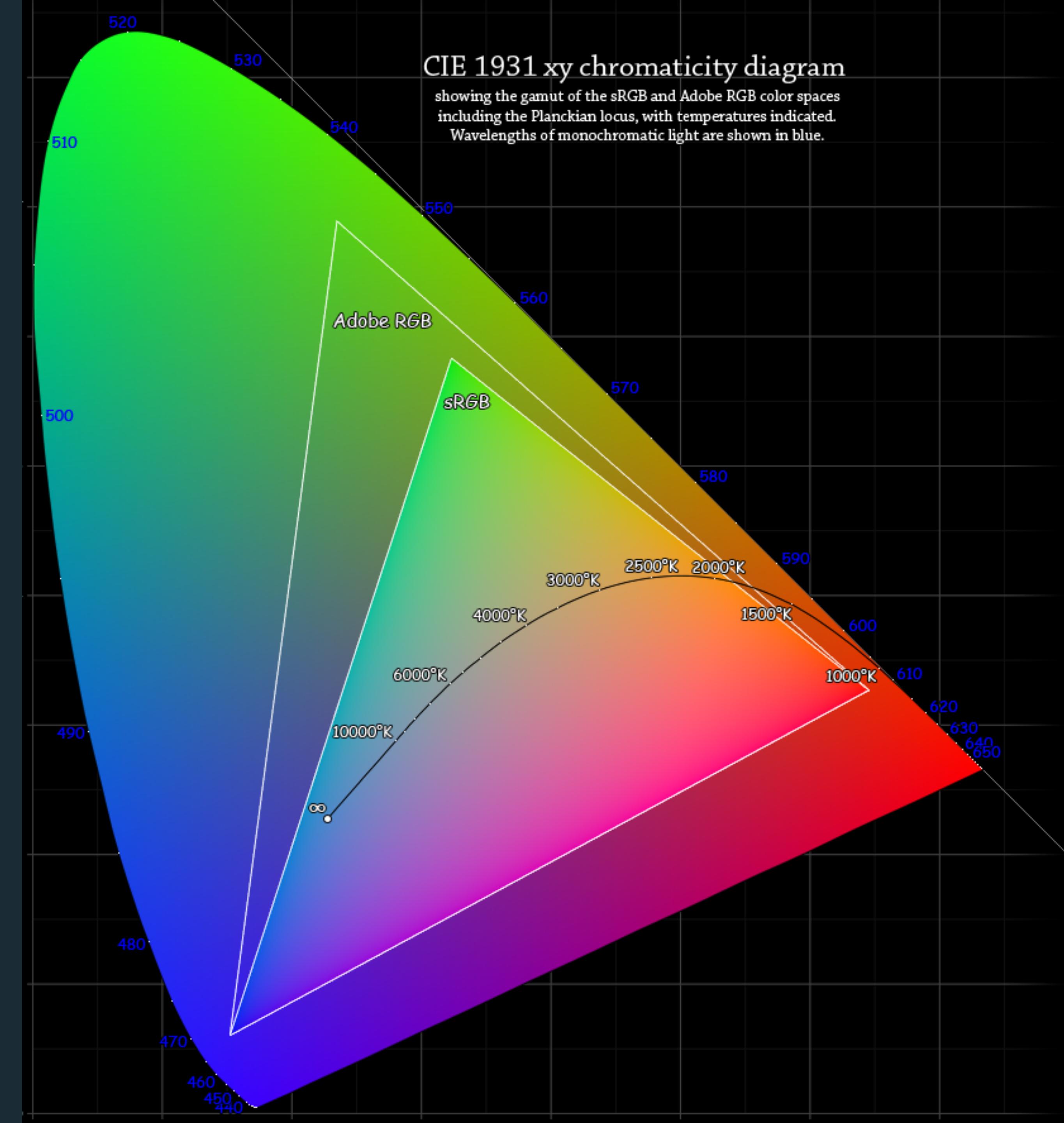
green

orange

purple

# Course Topics

## Color

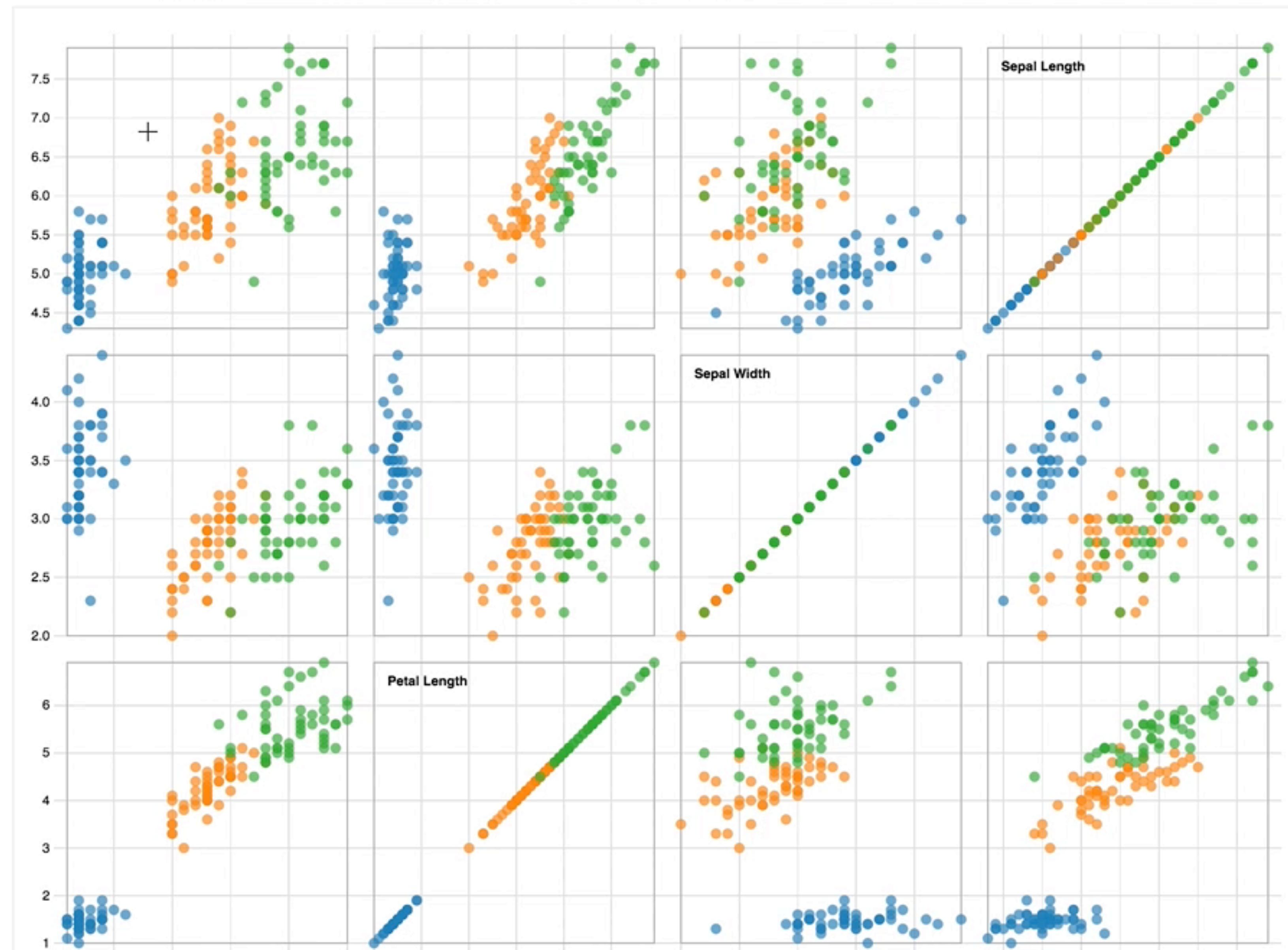




# Course Topics

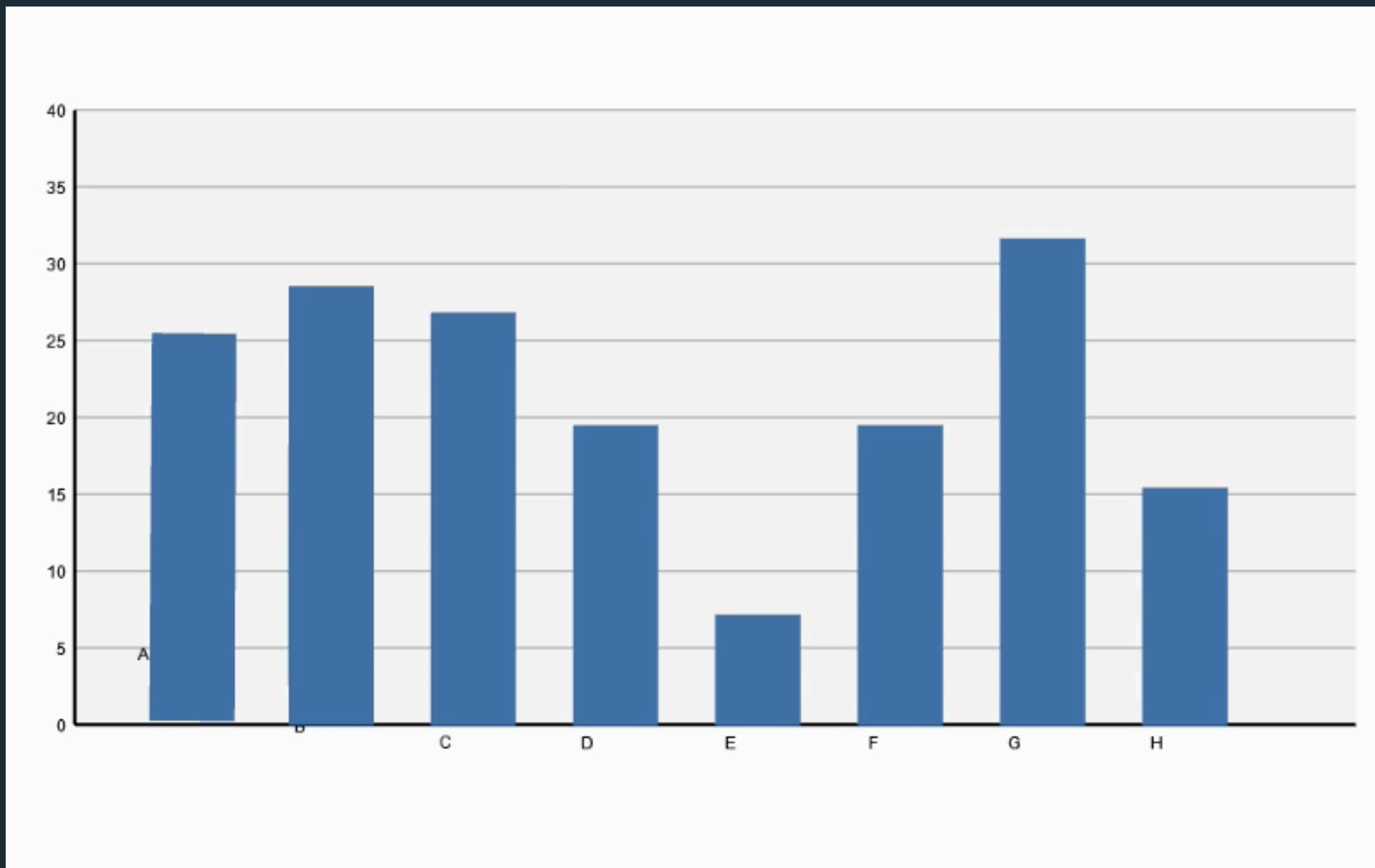
Interactivity

## Scatterplot Matrix Brushing



# Course Topics

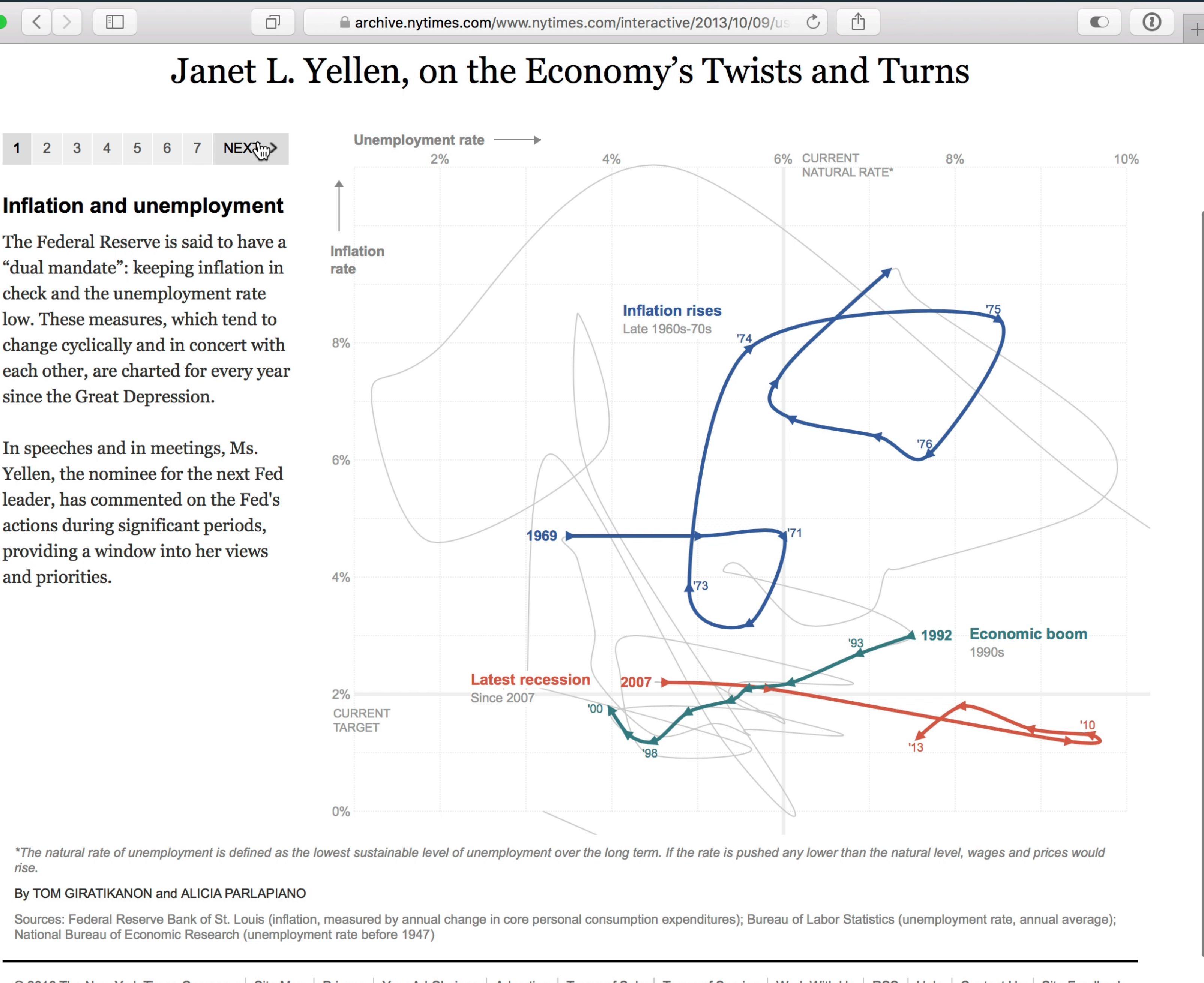
## Animation



*Heer & Robertson. (2007).*

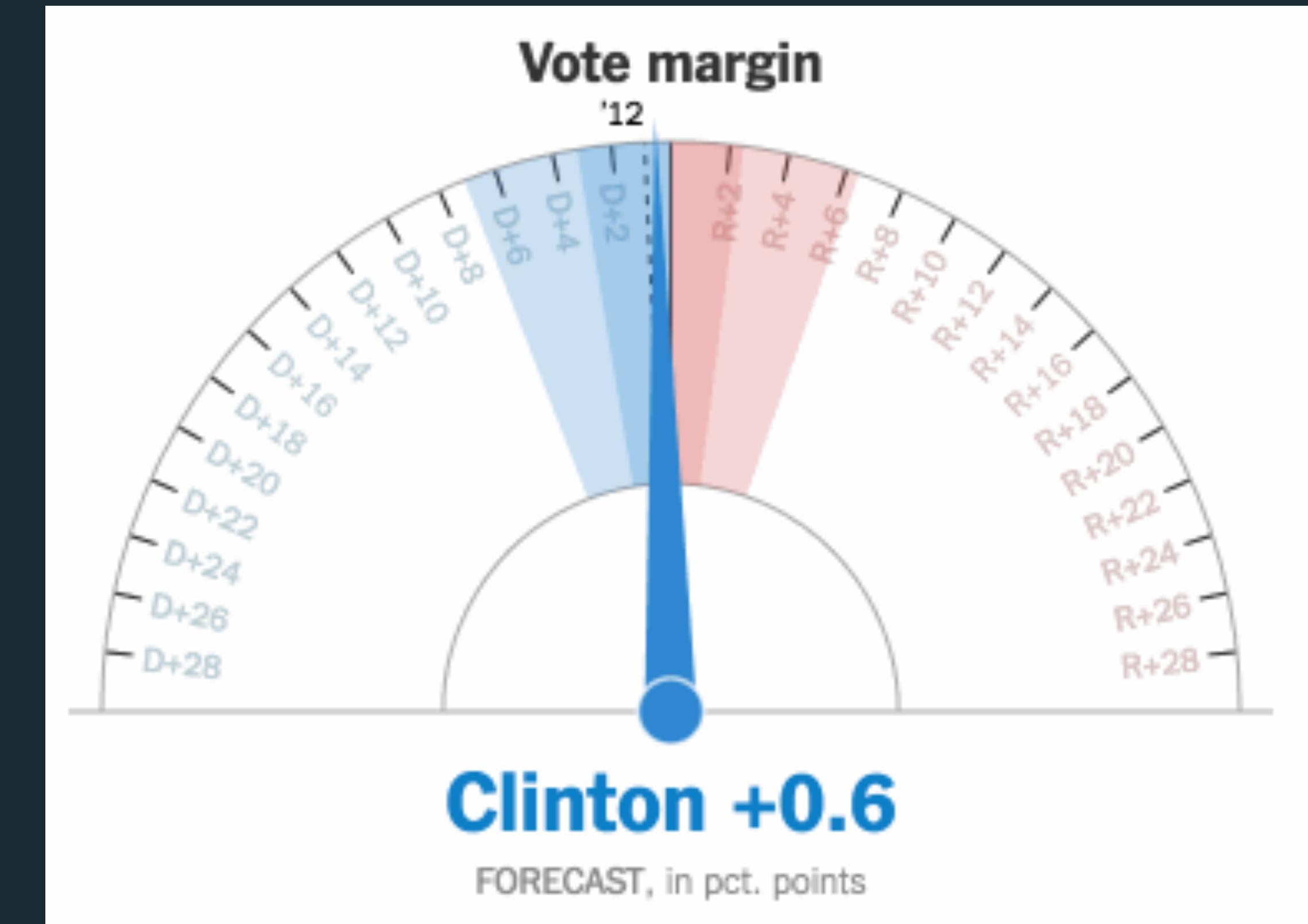
# Course Topics

## Narrative



# Course Topics

## Error & Uncertainty



# Course Topics

Mapping &  
Cartography



Dymaxion Maps, *Fuller*. (1946)

# Course Staff

# Course Staff

My career:

CS all the way through!

BS, UC San Diego

PhD, Stanford (+UW)

Visiting Researcher, Google Brain

New faculty @ MIT



Interests:

Interactive Data Visualization Tools

Machine Learning Interpretability

Arvind Satyanarayan

# Course Staff

2nd year PhD student advised by David Karger

Study online collaboration and build systems to help people get their work done.

#information retrieval #automation #task management

<http://people.csail.mit.edu/soya/>



Soya Park

# Course Grading

Class Participation	5%	
Reading Commentaries	10%	
A1: Visualization Design	7.5%	Due 2/13
A2: Exploratory Data Analysis	12.5%	Due 2/27
A3: Interactive Prototype	20%	Due 3/13
A3 Peer Evaluation	5%	Due 3/22
Final Project	40%	
Proposal		Due 4/3
MVP + Presentations		Due 4/22
Poster Session + Final Deliverables		Due 5/13

# Course Grading

Class Participation

5%

Reading Commentaries

10%

A1: Visualization Design

7.5%

A2: Exploratory Data Analysis

12.5%

A3: Interactive Prototype

20%

A3 Peer Evaluation

5%

Final Project

40%

Proposal

MVP + Presentations

Poster Session + Final Deliverables

Readings posted on nb.mit.edu – sign-up emails EOD today, sign-up link in LMOD.

2 readings per week – a mix of research papers, articles, textbook chapters.

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

10%

A1: Visualization Design

7.5%

A2: Exploratory Data Analysis

12.5%

A3: Interactive Prototype

20%

A3 Peer Evaluation

5%

Final Project

40%

Proposal

MVP + Presentations

Poster Session + Final Deliverables

Major visualization project on topic/  
dataset of your choice.

Second half of the course, starting after  
spring break.

Teams of 1–3 people with a 1 page project  
proposal.

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

Final poster presentation on May 13.

Due 4/22

Due 5/13

# Course Grading

Class Participation

5%

Reading Commentaries

10%

A1: Visualization Design

7.5%

A2: Exploratory Data Analysis

12.5%

A3: Interactive Prototype

20%

A3 Peer Evaluation

5%

Final Project

40%

Proposal

MVP + Presentations

Poster Session + Final Deliverables

**Design a static visualization for a dataset.**

Every 10 years, the census documents the demographic make-up of the U.S., influencing congressional districting and social services. This dataset contains a summary of census data for two years a century apart: 1900 and 2000.

Due 3/22

You must choose the message you want to convey. What question(s) do you want to answer? What insight do you want to communicate?

Due 5/13

# Course Grading

Class Participation	5%	Pick a <b>guiding question</b> , use it as your title.
Reading Commentaries	10%	Design a <b>static visualization</b> to answer it.
A1: Visualization Design	7.5%	You are free to use any tools (inc. pen & paper).
A2: Exploratory Data Analysis	12.5%	<small>Due 2/13</small>
A3: Interactive Prototype	20%	<b>Deliverables</b> (upload via LMOD; see A1 page).
A3 Peer Evaluation	5%	Image of your visualization (PNG or JPG format).
Final Project	40%	Short description + design rationale ( $\leq 4$ paragraphs).
Proposal		<small>Due 3/13</small>
MVP + Presentations		<small>Due 3/22</small>
Poster Session + Final Deliverables		Due by <b>noon, Wednesday, Feb 13</b> .
		<small>Due 5/13</small>

# Questions?

<http://vis.mit.edu/classes/6.894/>