Announcements

1. Readings:
   - https://github.com/gquer/dsc-96_winter19/blob/master/04_mapping_python/readings.md

2. Assignment: Wed 6PM

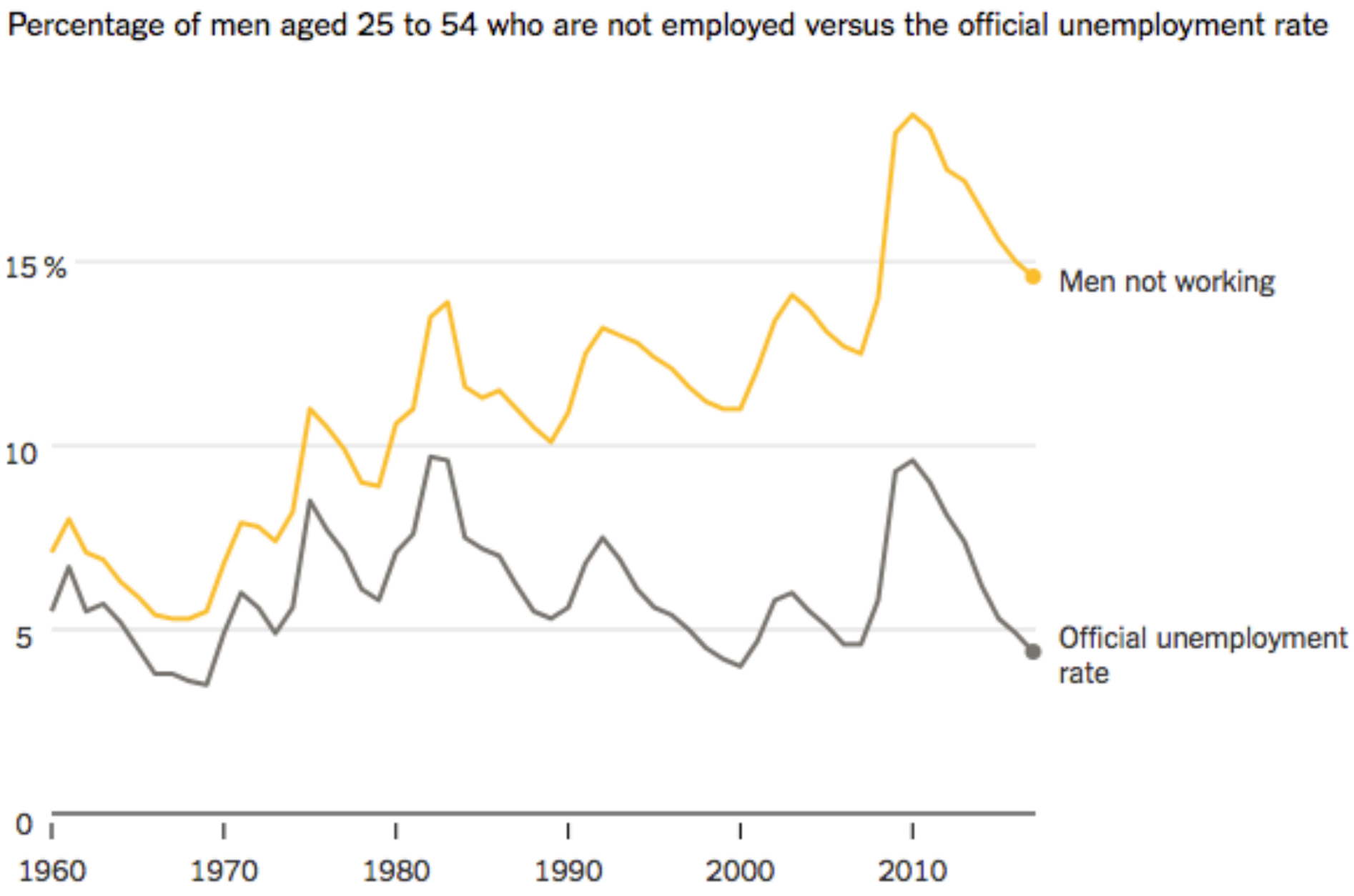
3. Guest lecturer: Ethics Research in Data Science
   - Thursday, 5PM

4. Jupyterhub !
   - Pull: click on https://datahub.ucsd.edu/hub/user-redirect/git-pull?repo=https%3A%2F%2Fgithub.com%2Fgquer%2Fdsc-96_winter1
   - * DO NOT:
     - modify any file in the folder
   - * DO
     - duplicate a file, call it with a proper name, and modify it as you wish in your jupyterhub
   - * IF YOU ACCIDENTALLY MODIFIED A FILE:
     - delete the file
     - click on the link once more

Questions, Metrics and Data Science

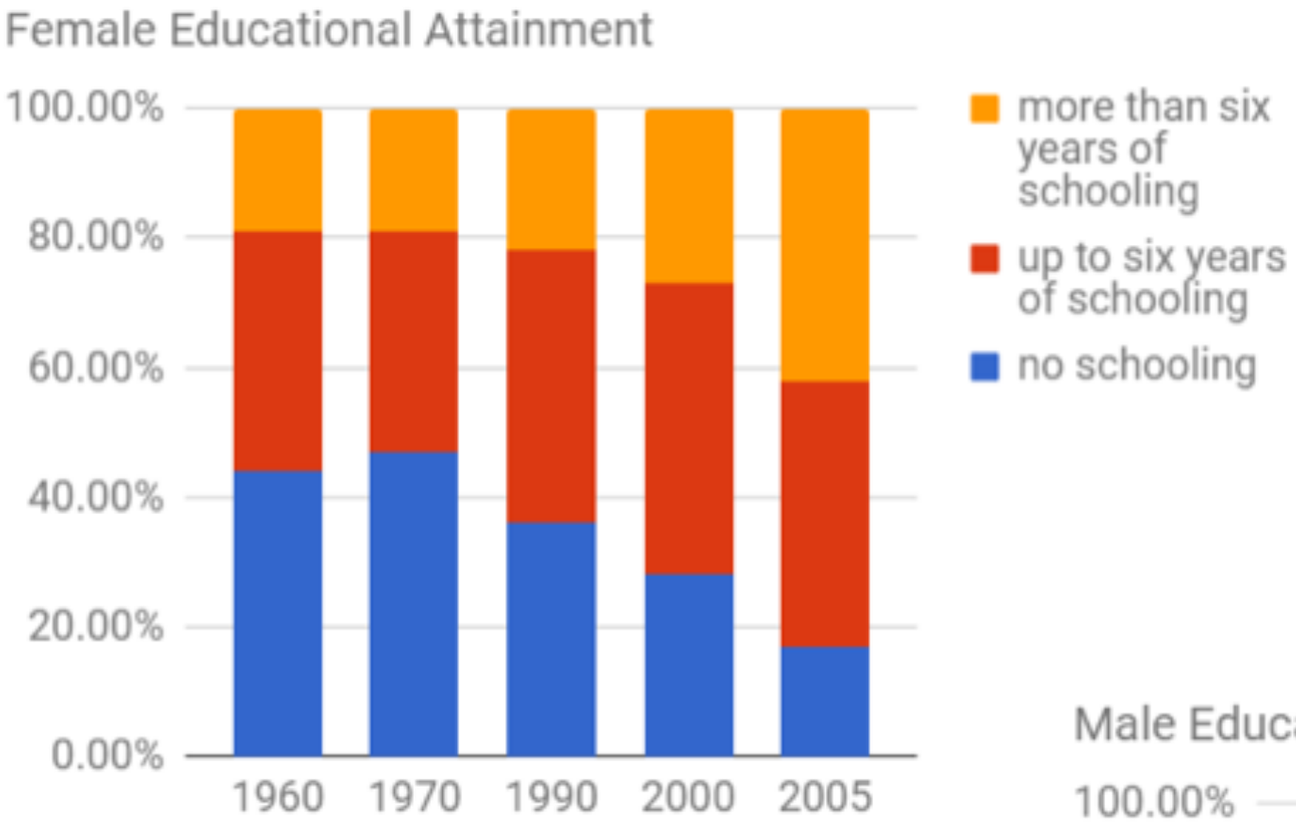
Giorgio Quer

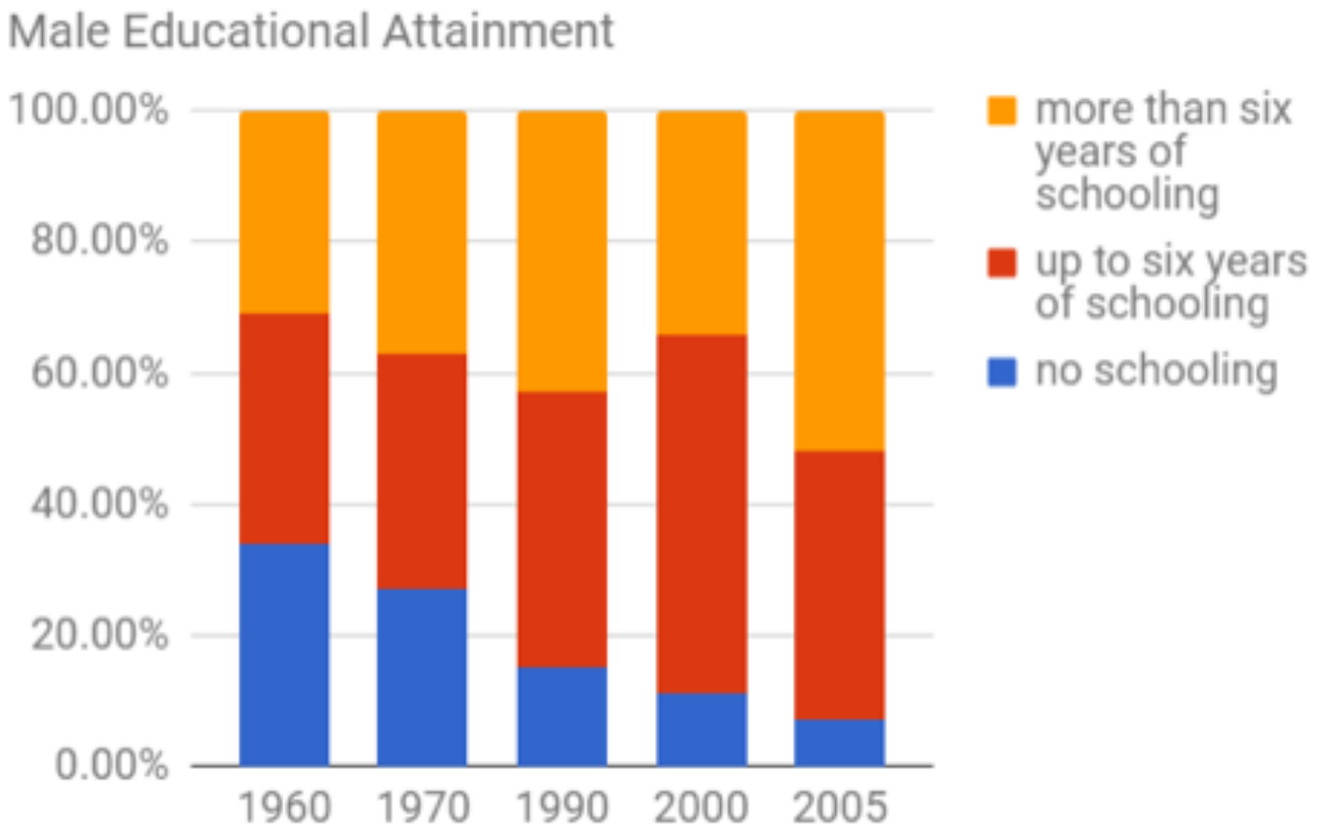
Talking about numbers



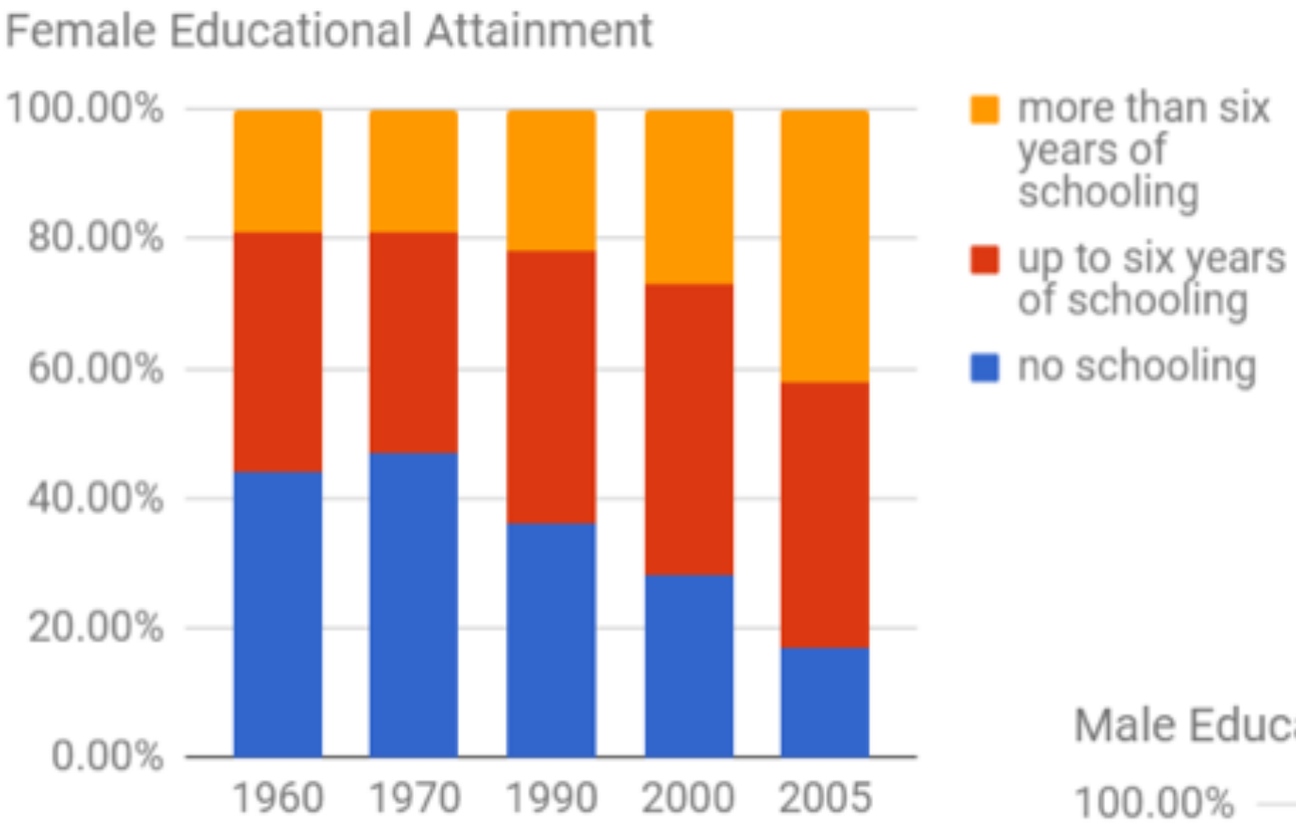
By The New York Times | Source: St. Louis Federal Reserve

Talking about numbers

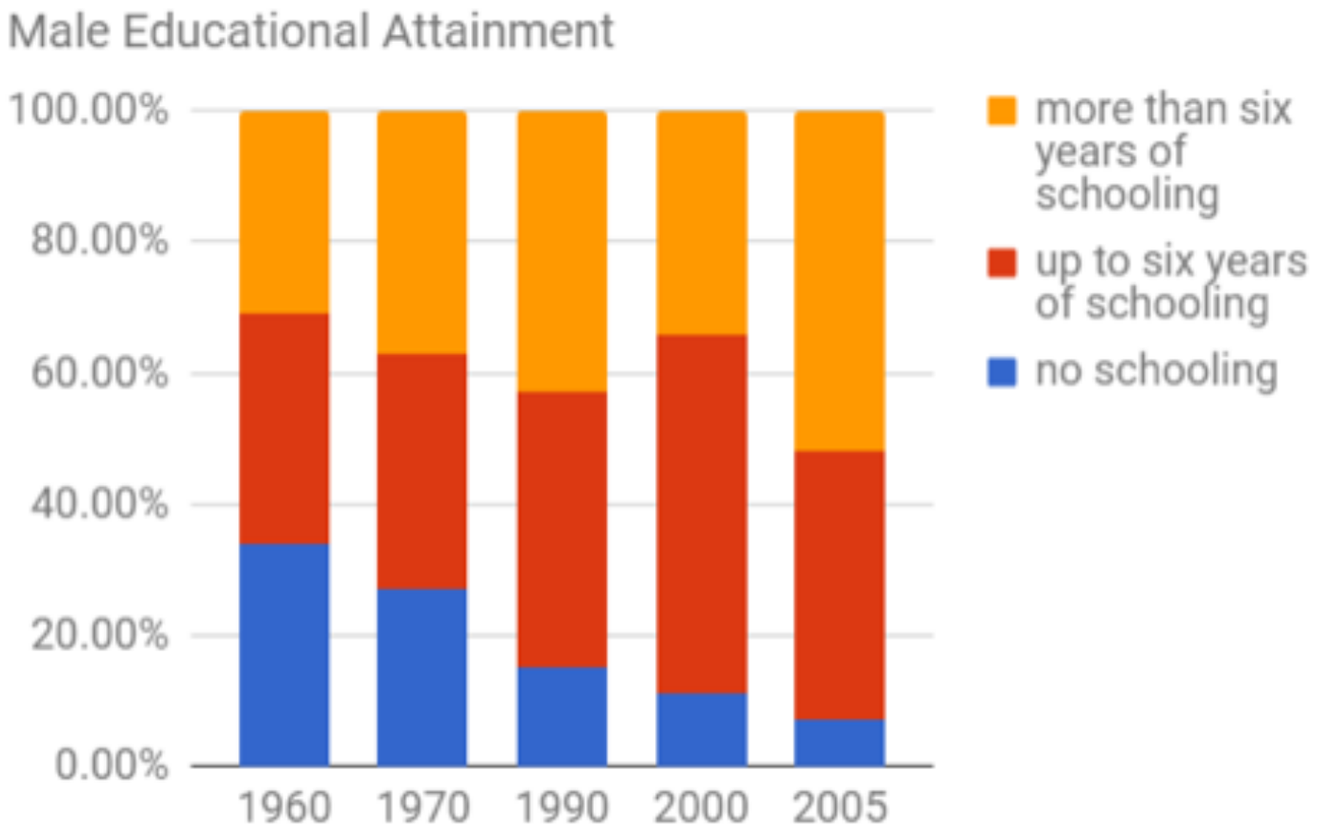




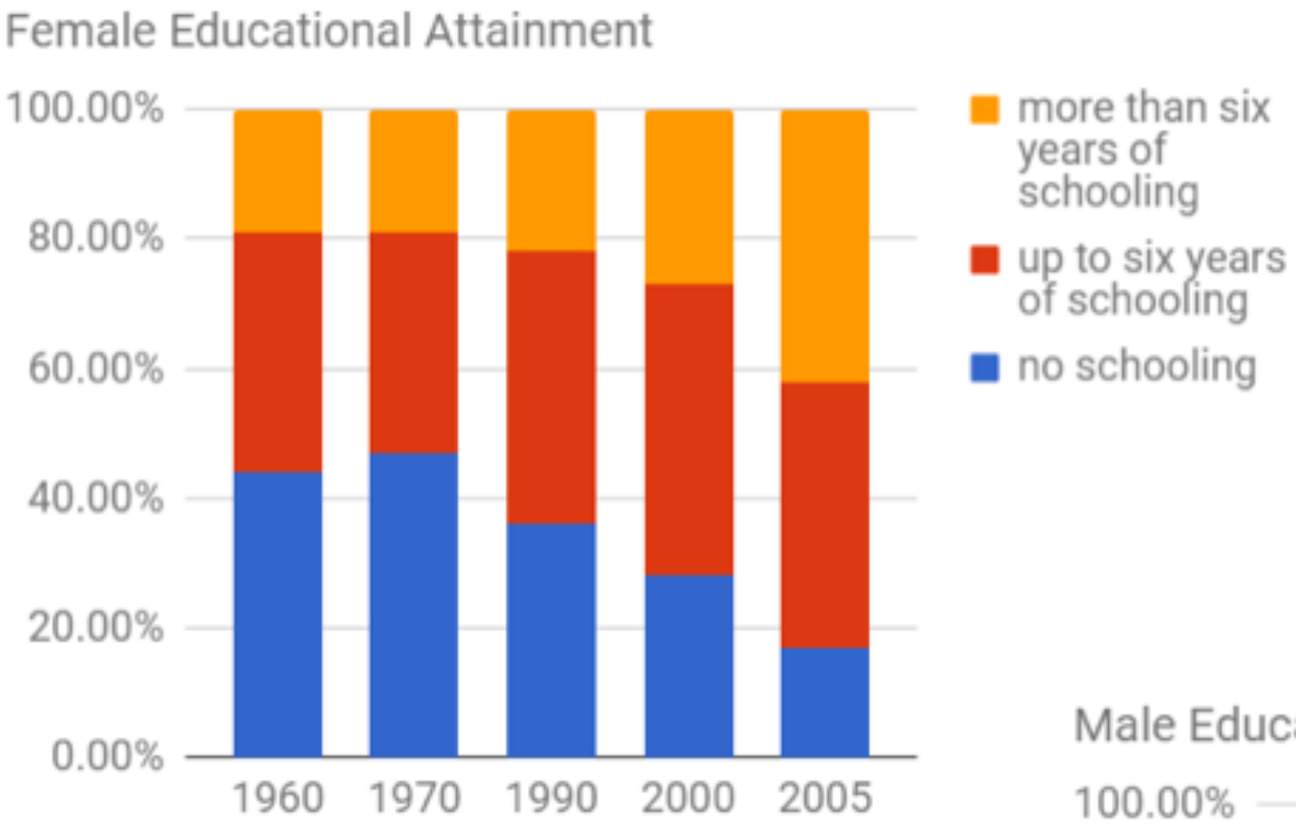
Talking about numbers



- More than 6 years of schooling in 2005
  1. 0-30%
  2. 30-40%
  3. 40-50%
  4. 50-60%
  5. 60-70%
  6. 70-100%

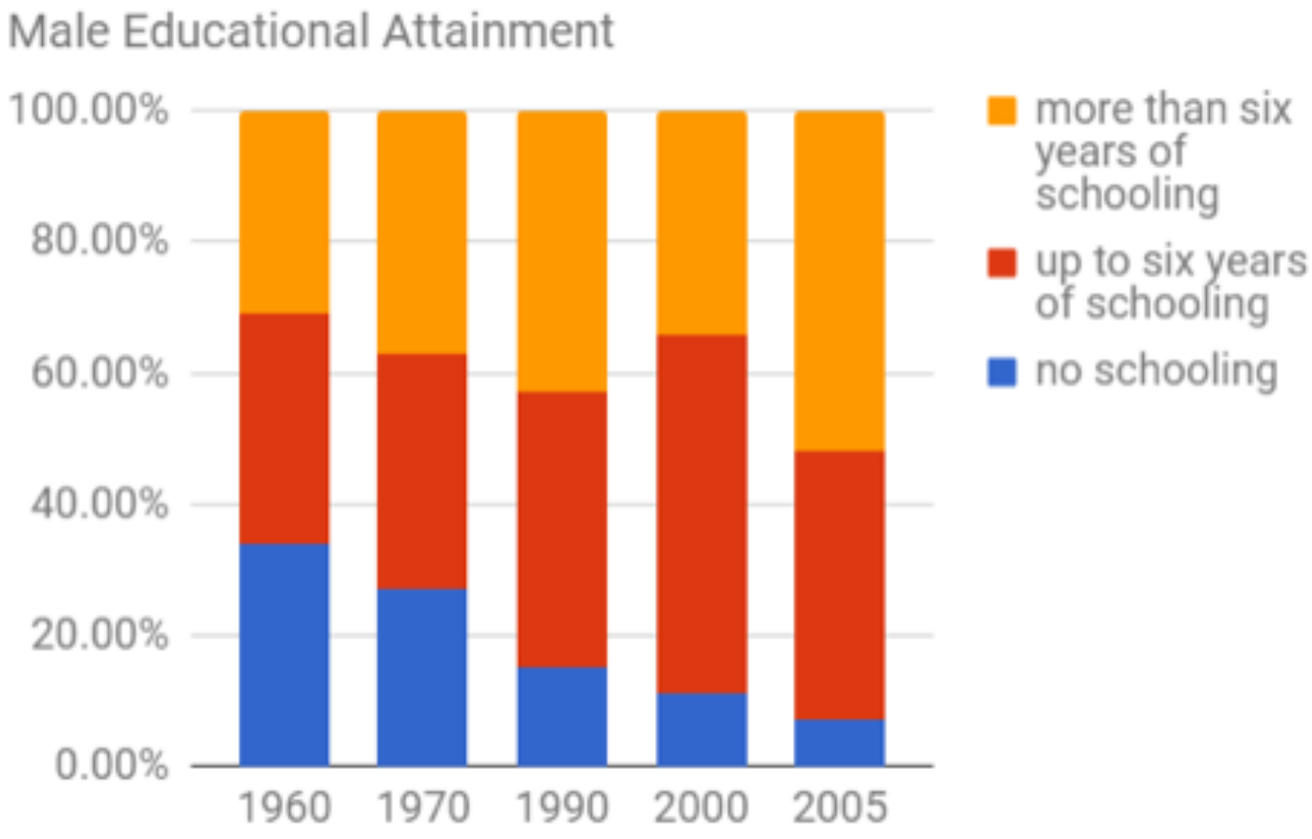


Talking about numbers

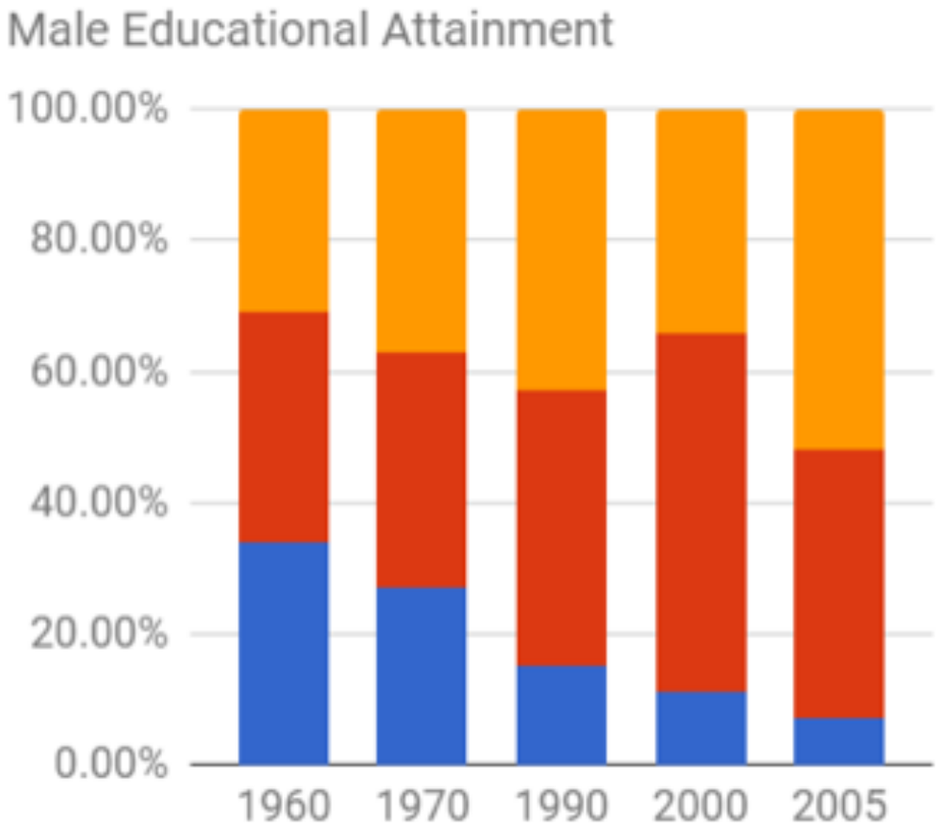


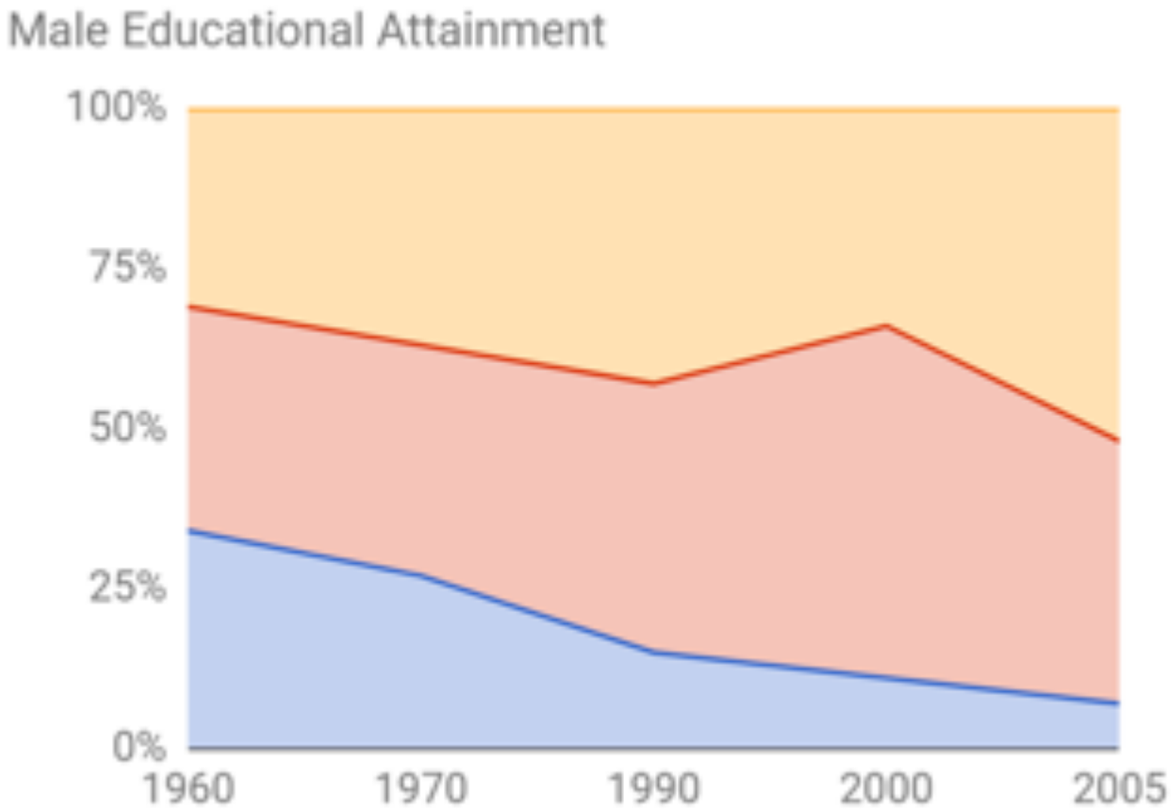
- Is this graph correct?
  - Doing analysis right
  - Providing the right answer
- **Misleading means lying!**

- More than 6 years of schooling in 2005
  1. 0-30%
  2. 30-40%
  3. 40-50%
  4. 50-60%
  5. 60-70%
  6. 70-100%



Talking about numbers





Year	no schooling	up to six years of schooling	more than six years of schooling
1960	34.00%	35.00%	31.00%
1970	27.00%	36.00%	37.00%
1990	15.00%	42.00%	43.00%
2000	11.00%	55.00%	34.00%
2005	7.00%	41.00%	52.00%

Talking about numbers

- Every man in my family has heart disease. **I want to be the last.**

Audience	Impressions	Clicks	Click rate
General	255,349	6425	2.5%
Heart disease	165,952	2055	1.2%

Talking about numbers

- Every man in my family has heart disease. **I want to be the last.**

- Is this message appealing for people who had a heart disease?
  - Yes, but
  - Is this meaningful?
  - What about the population?

Audience	Impressions	Clicks	Click rate
General	255,349	6425	2.5%
Heart disease	165,952	2055	1.2%

Talking about numbers

- Every man in my family has heart disease. **I want to be the last.**

Audience	Impressions	Clicks	Click rate
General	255,349	6425	2.5%
Heart disease	165,952	2055	1.2%

- Is this message appealing for people who had a heart disease?
  - Yes, but
  - Is this meaningful?
  - What about the population?

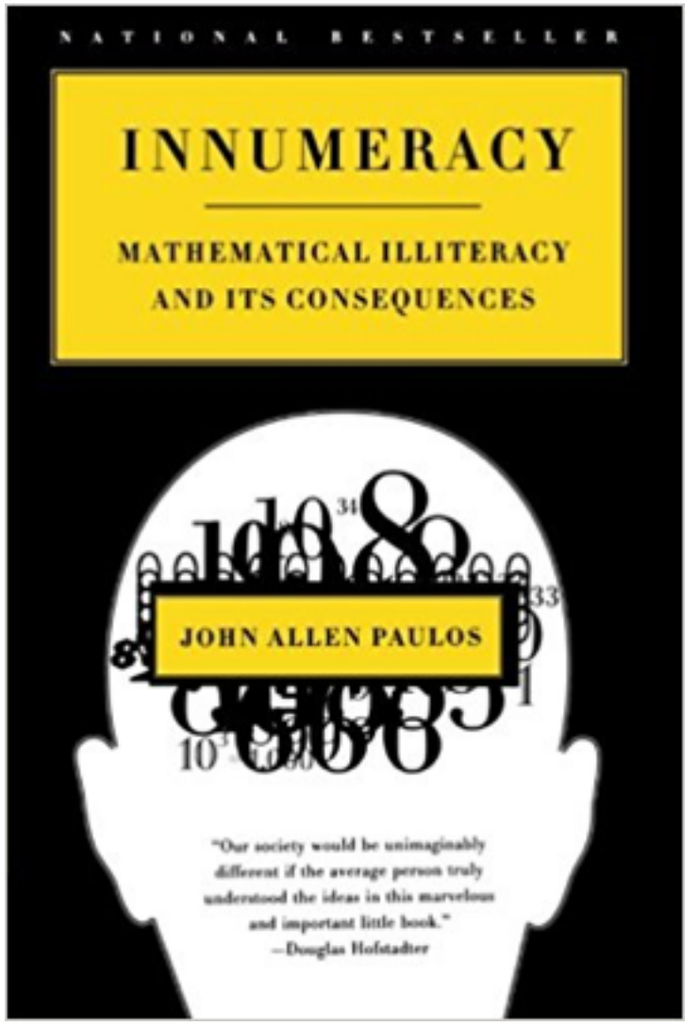
- % over 65 age
  - General
    - 40 %
  - Heart disease
    - 80 %
- **Are we still sure?**

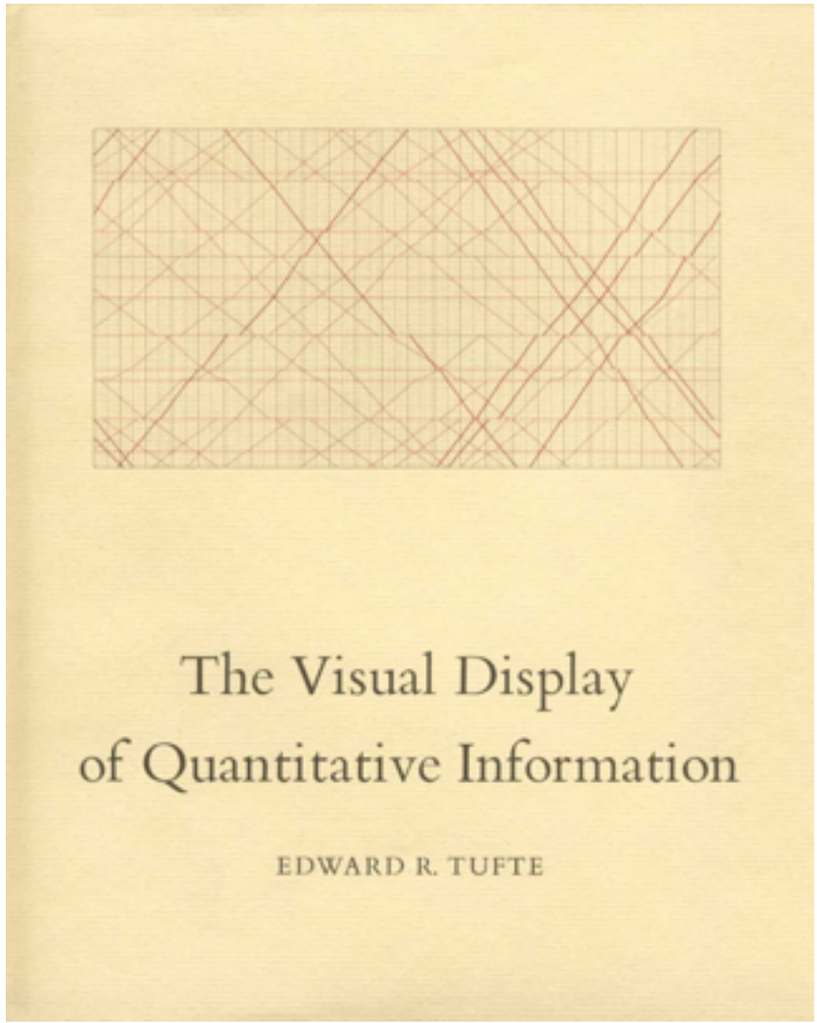
Talking about numbers

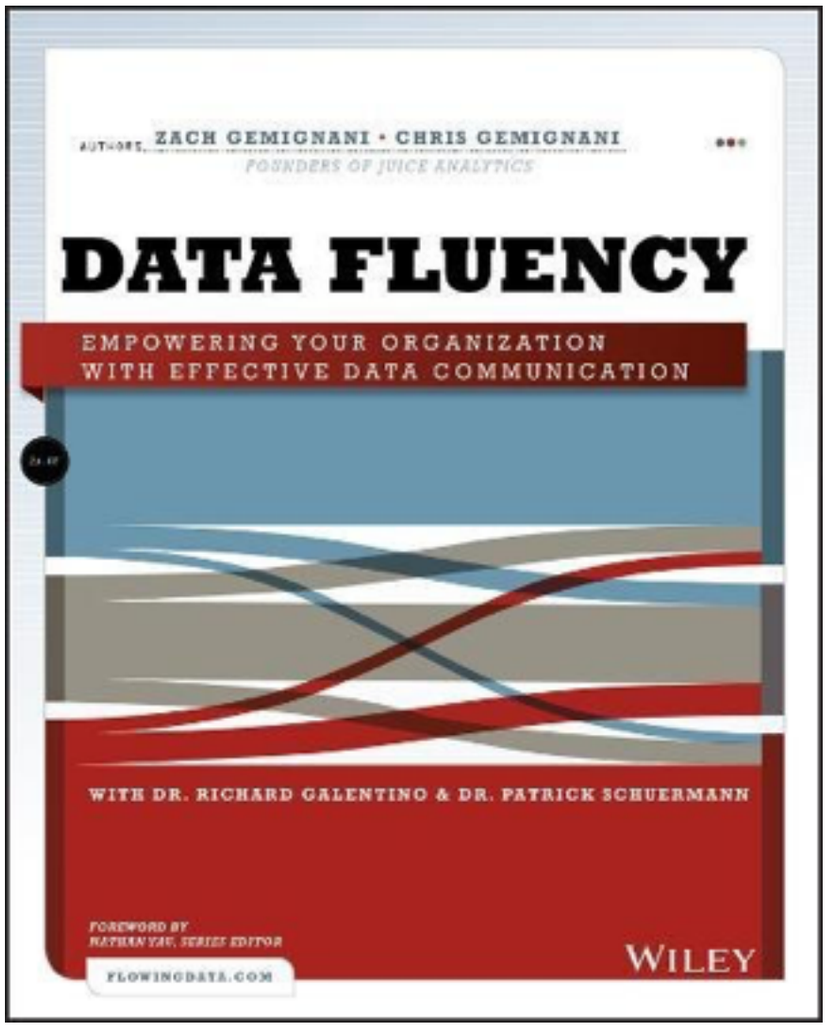
- Data driven culture
  - Data
    - Significance
    - Trust
  - Audience
    - Literacy
    - Decision making

- Data scientist
  - Answer one question
  - Experiment
  - Present your data
  - Get feedback
    - Iterate

Talking about numbers

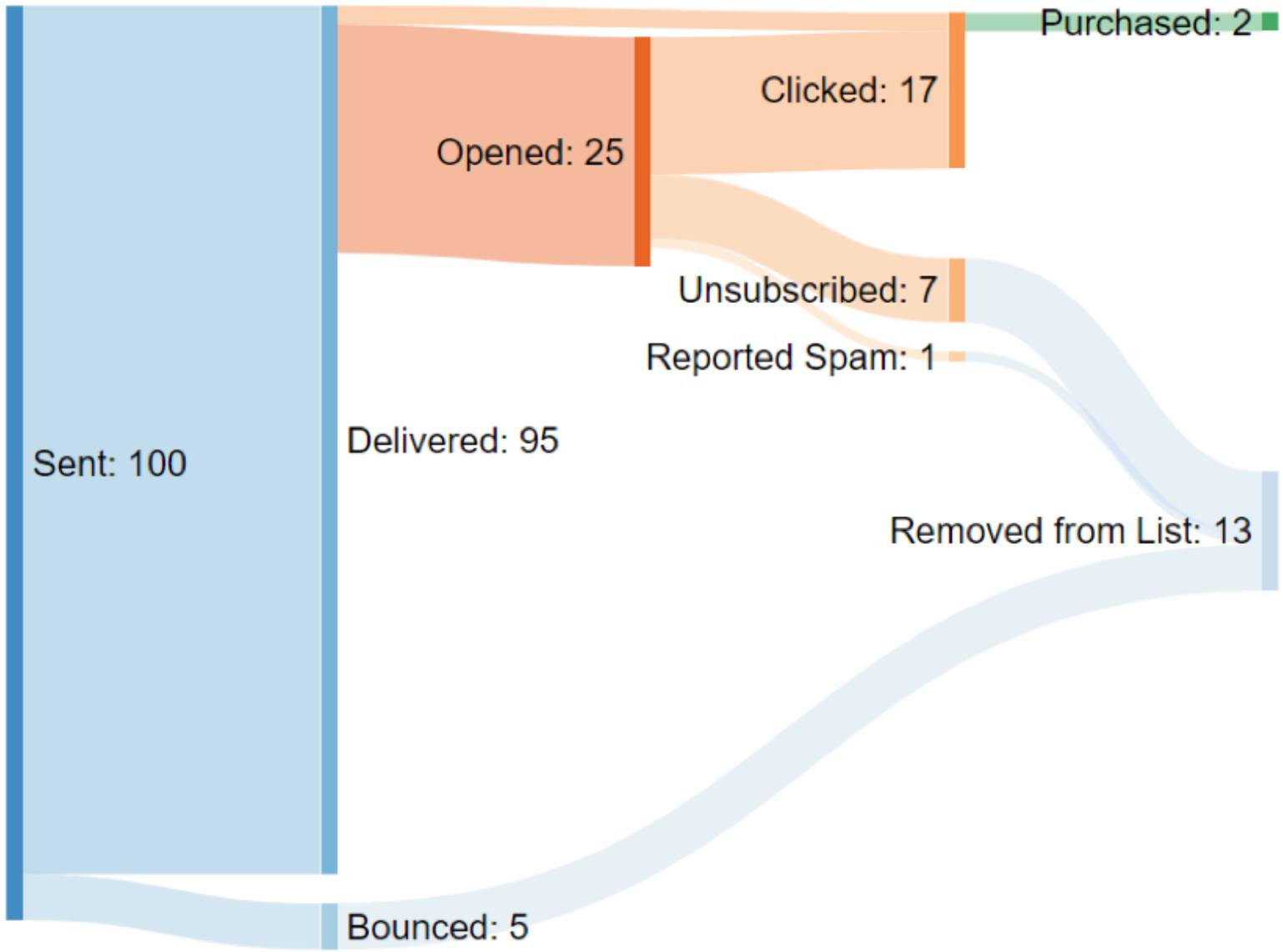






Metrics

- You are a Data Scientist
  - In a research program, with email marketing
- We want to understand people engagement with new emails we send: **when is user engagement down?**
- You need to **design a metric** to track it
- You have access to a real-time flow of events
  - Design a metric to alert if something goes wrong



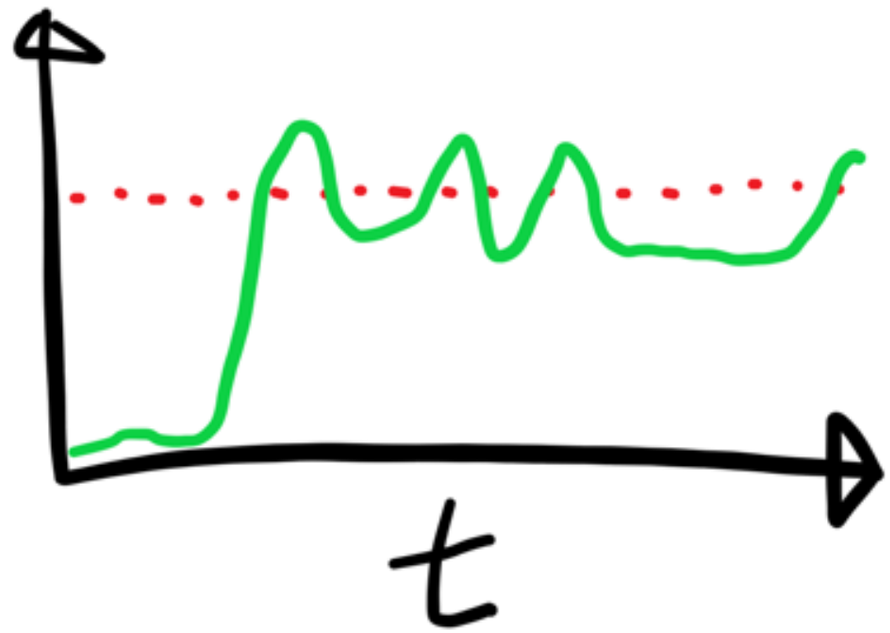
Metrics

Run hourly:

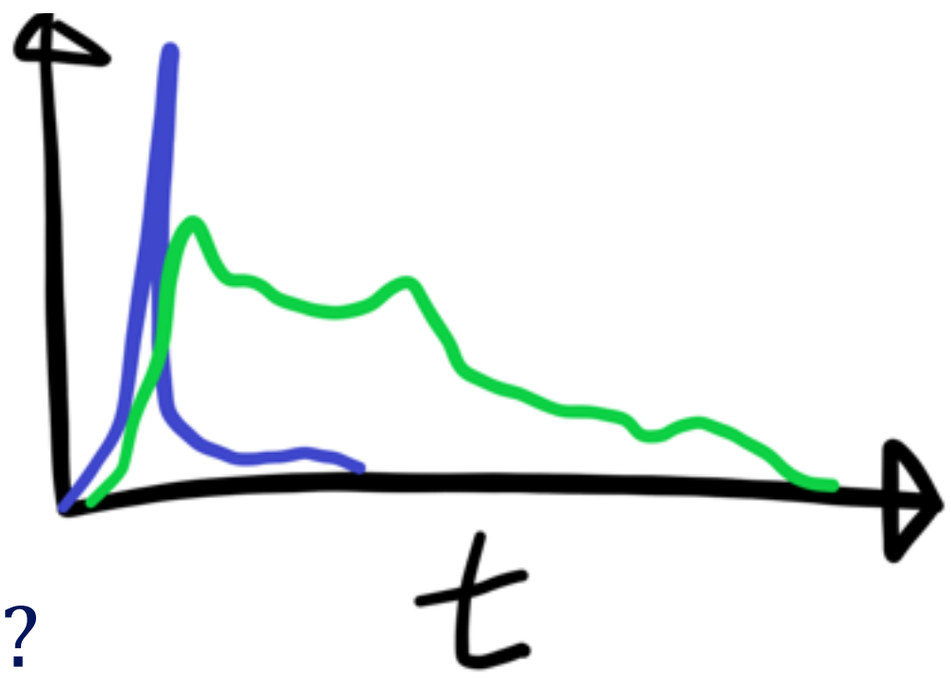
```
# Count emails sent in last 24 hours
emailCount = COUNT(event) WHERE
                        event["actionType"] == "Sent" AND
                        event["occuredAt"] < ( TODAY()-1 )

# Count click events in the last 24 hours
clickCount = COUNT(event) WHERE
                        event["actionType"] == "Click" AND
                        event["occuredAt"] < ( TODAY()-1 )

# Calculate the click rate
clickRate = clickCount / emailCount

# Compare to threshold
threshold = 0.17
IF clickRate > threshold:
      alertState = True
```

Metrics

- What can go wrong?
  - Small numbers
    - frequent threshold crossing



  - Clicks are delayed!
    - Clicks may not correspond to the email sent in the previous hour



  - Unique vs total clicks
  - Click per send or click per open?
  - What time window is appropriate?

Metrics

- Consumers of data science products are making data-driven decisions
- If a data consumer is mislead:
  - They may make important business or life decisions that are based on falsehoods
  - They may quickly lose trust that you may not be able to recover
- To maintain this:
  - **Never knowingly ship bad data** or analysis
  - Acknowledge and quickly **fix mistakes** that are reported
  - **Check** in with users to make sure they actually **understand** what is being presented

Right question

- Ask a **sharp** question
  - a sharp question must be answered with numbers, which is what you extract from data
  - "What's going to happen with my stock?" ---> "The price will change"
  - "What will my stock's sale price be next week?" ---> specific price!
- Make sure your data can answer the question!

Right question

- Ask a **sharp** question
  - a sharp question must be answered with numbers, which is what you extract from data
  - "What's going to happen with my stock?" ---> "The price will change"
  - "What will my stock's sale price be next week?" ---> specific price!
- Make sure your data can answer the question!
- Reformulate your question
  - insight from data
  - can they be generalized
  - can they be used for future prediction
- Questions we can answer now:
  - Is the police pulling over car at the right moment?
    - What time are cars usually pulled over?
    - What time are crashing usually happening?
    - Day of the week
    - Geographical area

https://docs.microsoft.com/en-us/azure/machine-learning/studio/data-science-for-beginners-ask-a-question-you-can-answer-with-data