
Welcome to INFO 2950 (Intro to Data Science)!

Pick up 1 whiteboard, 1 marker, and a few tissues (erasers) on your way in.

Feel free to draw a cat while you wait for class to start.

(Make sure to return these at the end of class!)

INFO 2950: Intro to Data Science

Lecture 1
2022-08-21

Agenda

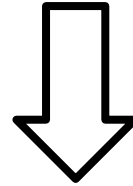
1. What is INFO 2950?
2. Staff intros
3. What is data science?
4. Course goals
5. DS project examples
6. Dataframes
7. Course logistics
8. HW0

What is INFO 2950?

- Computational tools + real data + [*data science skills*]

What is INFO 2950?

- Computational tools + real data + [*data science skills*]

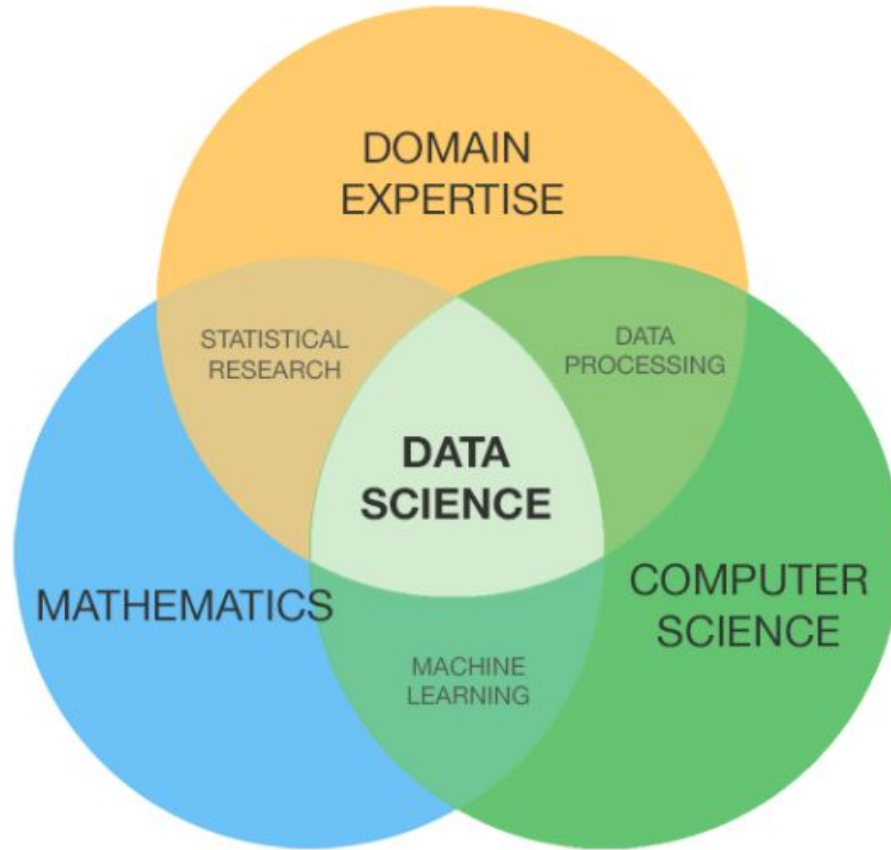


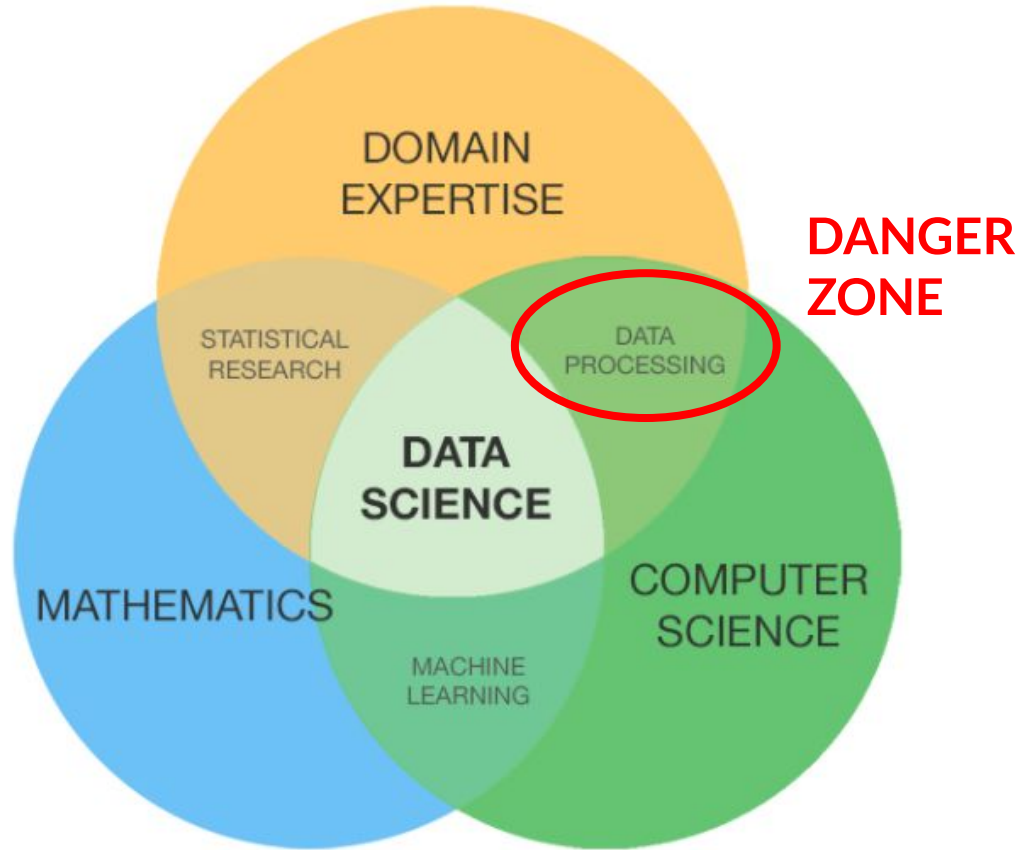
- Use data ethically to create evidence to support an argument

What will you get out of 2950?

- The ability to think critically about & generate your own data analyses
- Foundational knowledge for future DS, ML, AI courses
- The skills to ace a basic DS job interview in industry

**When I say “data science”,
what do you think of?**





On your **whiteboards: what skills
are you most comfortable with?**

- A. Math / Stats**
- B. Coding**
- C. Physical Sciences (Bio/Chem/Physics)**
- D. Finance / Econ / Business**
- E. Engineering**
- F. None of these, I'm here to learn!**



Staff Intros

- **Prof. Koenecke / Prof. K** (*she/her*)
 - **Academic:** MIT, Stanford (Computational & Mathematical Eng.)
 - **Industry:** big tech, consulting; research w/ non-profits
 - **Student hours:** Tues 11-noon (Gates 227)
 - **Toxic traits:** Talks too fast, thinks “data” is plural
- **Roz Thalken** (*she/her*)
 - **Academic:** U of Nebraska, Washington SU (Literature), now a Cornell PhD Student (Info Sci)
 - **Industry:** big tech, data science; tech startup
 - **Student hours:** Wed 11-noon (Rhodes 402)
 - **Toxic traits:** Makes every conversation about my cat, thinks “data” is singular

Staff

- **Graduate TAs**

- Andrea Wang
- Anna Choi
- Rejoice Hu
- Tangwuyou Su

- **Undergraduate TAs**

- Bella, Sydney, Karla, Arunabh, Alexia, Chiara, Ryan, Hao, Elliot, Zack, Gaby, Samhita, Jonathan, Anya, Ethan, Annie, Ahmed, Julius, Cassandra, Kevin, Sarah, Charlie



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Current events](#)
[Random article](#)
[About Wikipedia](#)

 Not logged in

Article

[Talk](#)

Read

[Edit](#)

[View history](#)

Data science

From Wikipedia, the free encyclopedia

Not to be confused with [information science](#).

Data science is an [interdisciplinary](#) field that uses [scientific methods](#), processes, [algorithms](#) and systems to extract [knowledge](#) and insights from noisy, structured and [unstructured data](#),^{[1][2]} and apply knowledge from data across a broad range of application domains. Data science is related to [data mining](#), [machine learning](#) and [big data](#).^[3]



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Current events](#)
[Random article](#)
[About Wikipedia](#)

 Not logged in

Article

[Talk](#)

Read

[Edit](#)

[View history](#)

Data science

From Wikipedia, the free encyclopedia

Not to be confused with [information science](#).

Data science is an [interdisciplinary](#) field that uses [scientific methods](#), processes, [algorithms](#) and systems to extract knowledge and insights from noisy, structured and unstructured data,^{[1][2]} and apply knowledge from data across a broad range of application domains. Data science is related to [data mining](#), [machine learning](#) and [big data](#).^[3]

Course Goals

- Use statistical methods, data processing, and real-world knowledge to make arguments using data

Course Goals

- Use statistical methods, data processing, and real-world knowledge to make arguments using data
- Course content:
 - Programming with Data
 - Regression and Linear Models
 - Evaluation, Probability, & Hypothesis Testing
 - Real World Applications (e.g. Machine Learning)

Course Goals

- Use statistical methods, data processing, and real-world knowledge to make arguments using data
- Ability to execute each phase of a typical DS project:
 - Data collection
 - Exploration and summarization
 - Model fitting
 - Hypothesis testing
 - Communication of findings

DS Project Examples

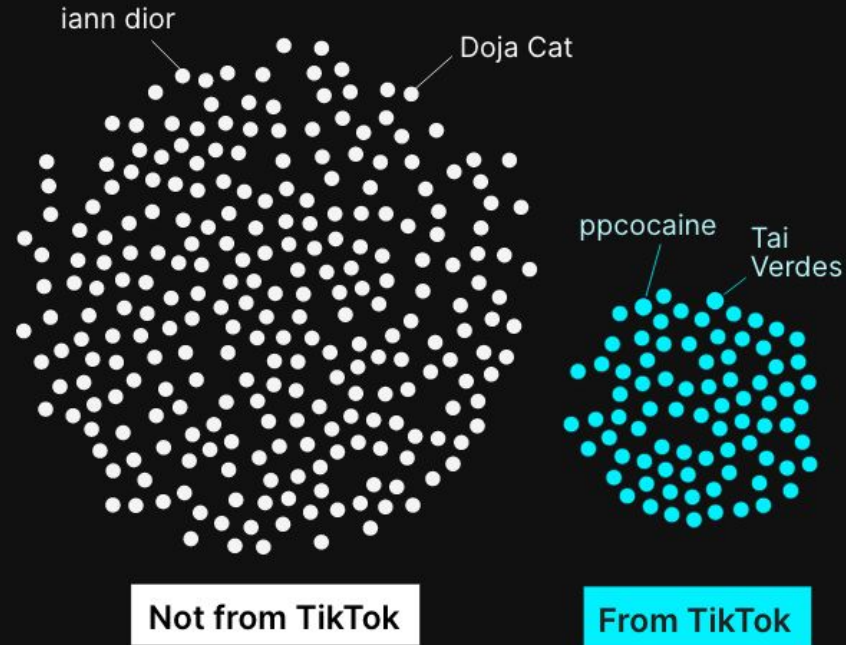
Successful projects often use data to...

1. Answer a question you have about the world
2. Ameliorate an existing problem you have
3. Make tools to help others solve problems they have

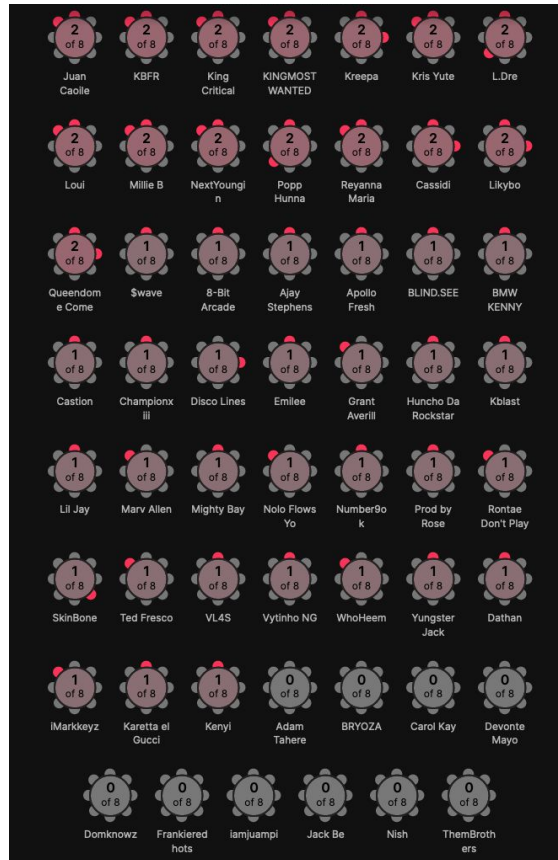
1. Question: TikTok virality → music careers?



Of the artists who charted on Spotify from January 2020 to December 2021, 332 had never charted before. 25% of them came from TikTok.







nt	Is this an established artist?	artist's big break?
	Yes	No

data type: **string?** **int?** **float?** **boolean?**

nt	Is this an established artist?	artist's big break?
	Yes	No
		<input type="text"/>

data type: **string?** int? float? boolean?

nt	Is this an established artist?	artist's big break?
	Yes	No
		<input data-bbox="1151 604 1559 699" type="text"/>

data type: **string?** **int?** **float?** **boolean?**

nt	Is this an established artist?	artist's big break?
	Yes	No
		<input data-bbox="1155 609 1568 707" type="text"/>

data type: **string?** **int?** **float?** **boolean?**

Cardi B

data type:

string?

int?

float?

boolean?

Cardi B

data type: **string?** **int?** **float?** **boolean?**

3.7

data type: string? int? float? boolean?

3.7

data type: **string?** **int?** **float?** **boolean?**

1

data type: string? **int?** float? boolean?

1

Python would interpret your inputting 1 as an int, but if you see a 1 in Python, it could be a float displayed with truncation, it could be a numeric display of a Boolean ($\text{True}+1=2$), or it could be the output of printed string "1".

data type: **string?** **int?** **float?** **boolean?**

one

data type: **string?** int? float? boolean?

one

data type: **string?** **int?** **float?** **boolean?**

2.0

data type: string? int? float? boolean?

2.0

data type: **string?** **int?** **float?** **boolean?**

“true”

data type: **string?** int? float? boolean?

“true”

data type: **string?** **int?** **float?** **boolean?**

True

data type: string? int? float? **boolean?**

True

(Same caveat that this could also be the printed display of a string variable with value "True")

data type: **string?** **int?** **float?** **boolean?**

yes

data type: **string?** int? float? boolean?

yes

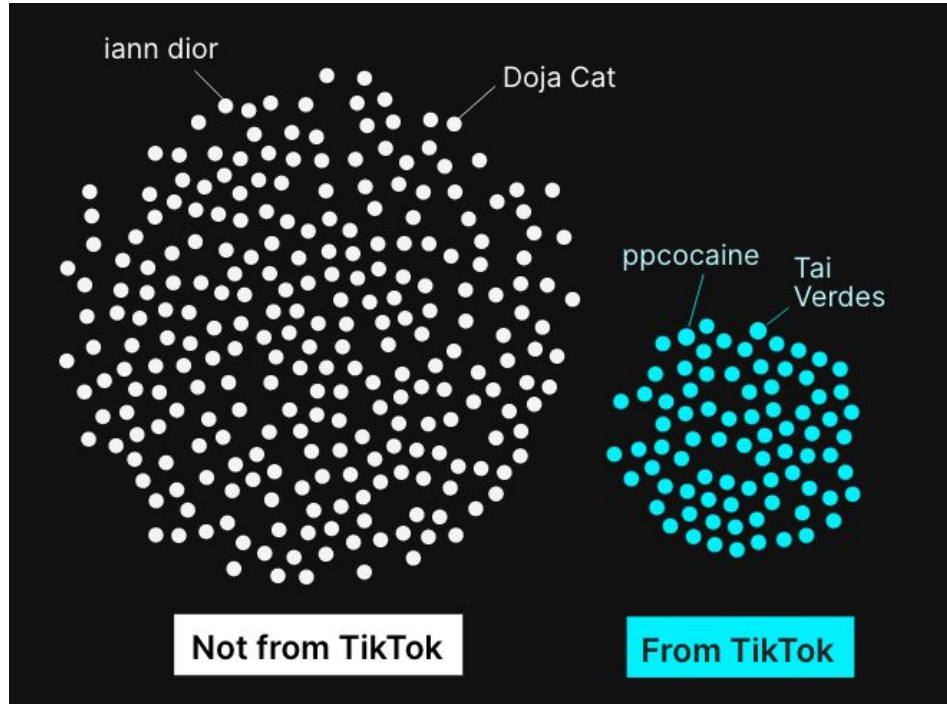
data type: **string?** **int?** **float?** **boolean?**

yasss

data type: **string?** int? float? boolean?

yasss

1. Question: TikTok virality → music careers?



Generate
question

Generate
question



Define metrics
(what does
"famous" or
"viral" mean?)

Generate
question



Define metrics
(what does
“famous” or
“viral” mean?)



Find
data

Generate
question



Define metrics
(what does
“famous” or
“viral” mean?)



Find
data



Clean
data

Generate
question



Define metrics
(what does
“famous” or
“viral” mean?)



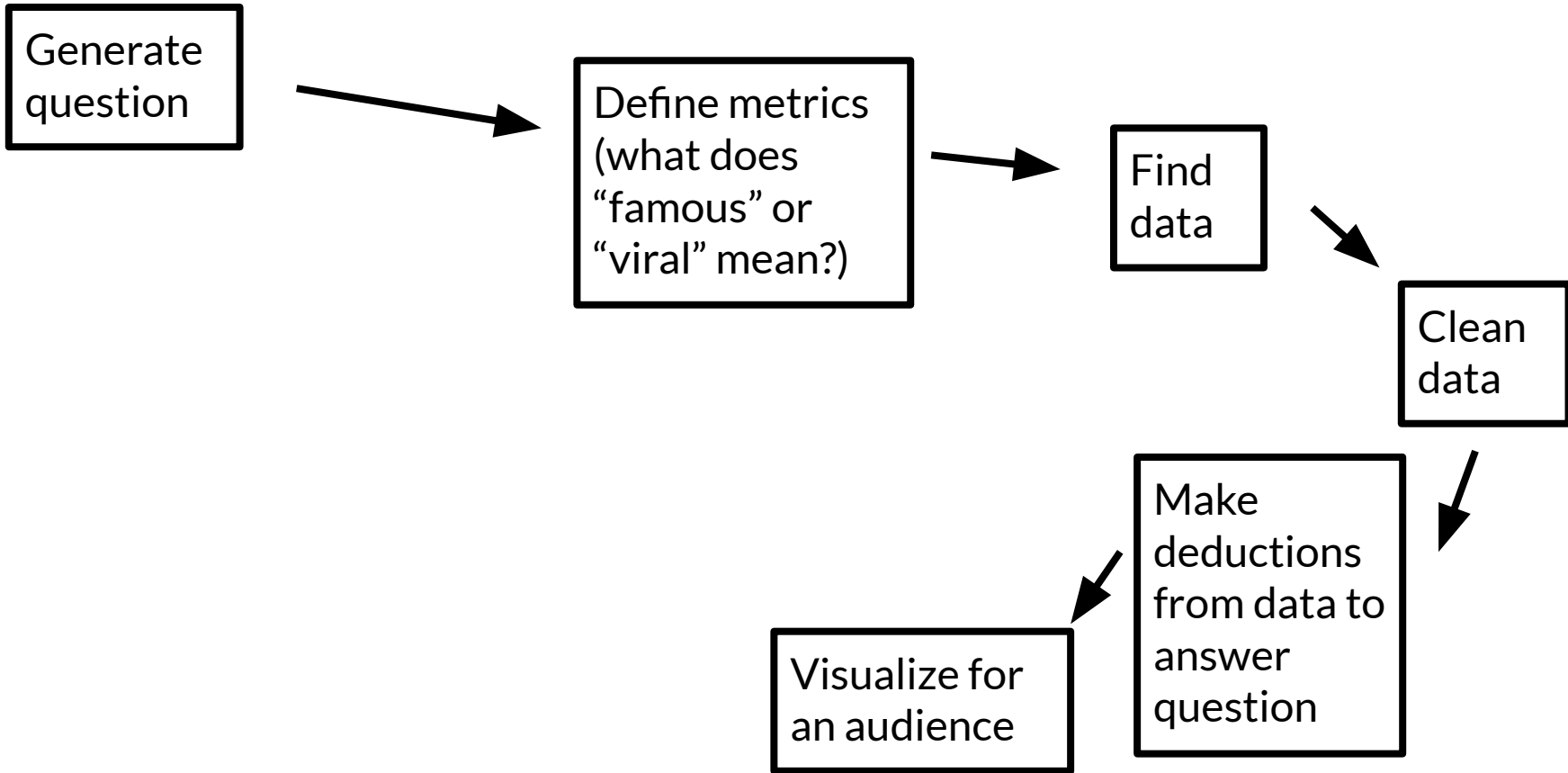
Find
data

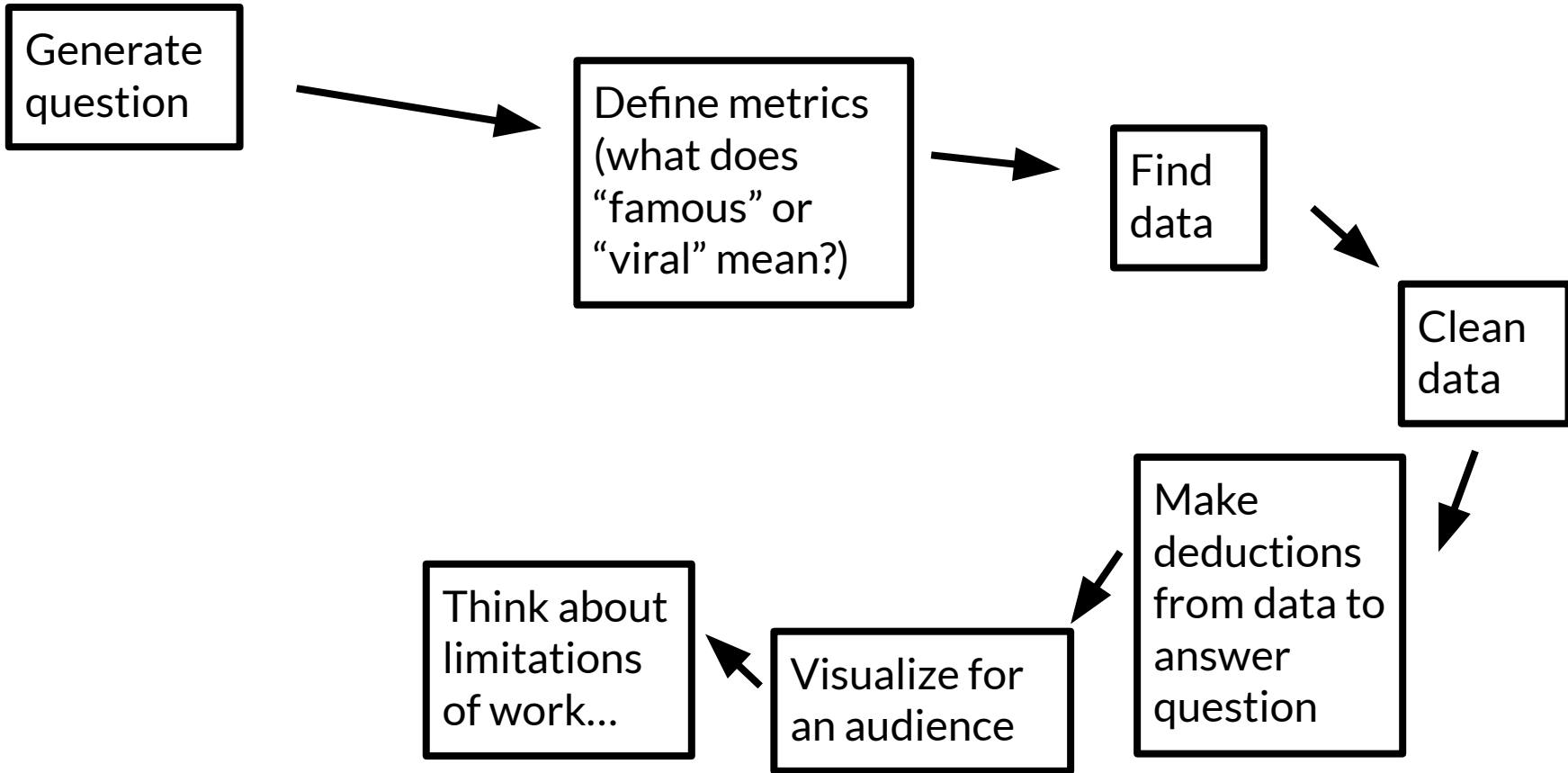


Clean
data



Make
deductions
from data to
answer
question





Generate
question



Define metrics
(what does
“famous” or
“viral” mean?)



Find
data



Clean
data



Make
deductions
from data to
answer
question



Visualize for
an audience



Think about
limitations
of work...



Refine
question



Refine
question

Think about
limitations
of work...

Visualize for
an audience

Make
deductions
from data to
answer
question

Clean
data

Find
data

Define metrics
(what does
“famous” or
“viral” mean?)

Generate
question

DS Project Examples

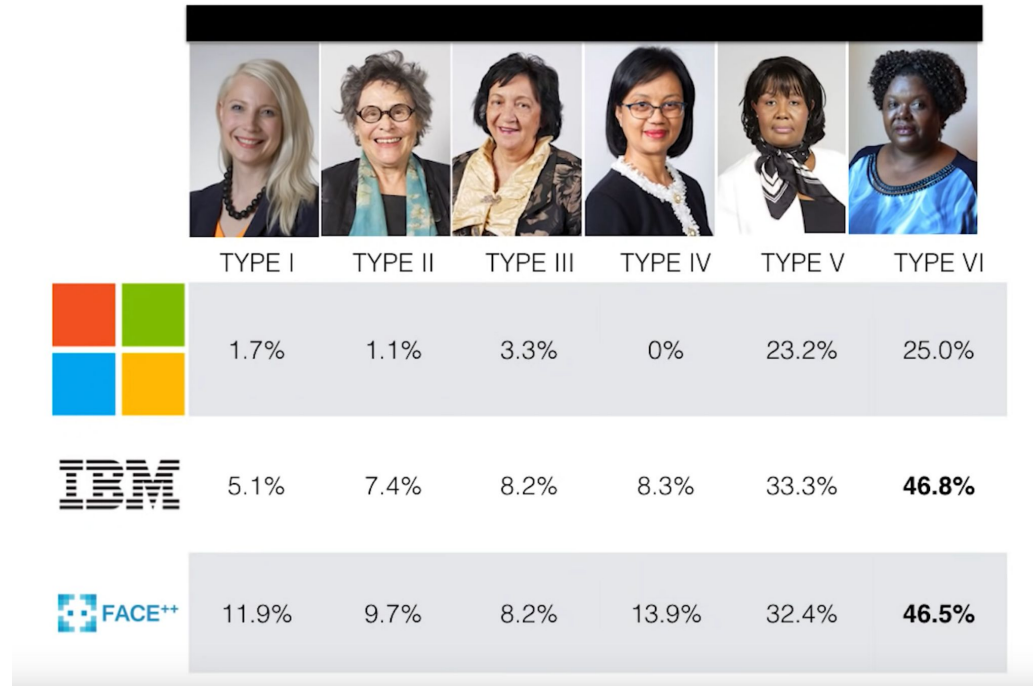
Successful projects often use data to...

1. Answer a question you have about the world
2. **Ameliorate an existing problem you have**
3. Make tools to help others solve problems they have

2. Problem-solving: computer vision



How does this chart make their argument?



What Type has highest accuracy for Microsoft?







TYPE I TYPE II TYPE III TYPE IV TYPE V TYPE VI



TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
1.7%	1.1%	3.3%	0%	23.2%	25.0%
5.1%	7.4%	8.2%	8.3%	33.3%	46.8%
11.9%	9.7%	8.2%	13.9%	32.4%	46.5%



What Type has highest accuracy for Microsoft?

						
	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
	1.7%	1.1%	3.3%	0%	23.2%	25.0%
	5.1%	7.4%	8.2%	8.3%	33.3%	46.8%
	11.9%	9.7%	8.2%	13.9%	32.4%	46.5%

Notice
problem

Notice
problem



Research and
understand
underlying issue

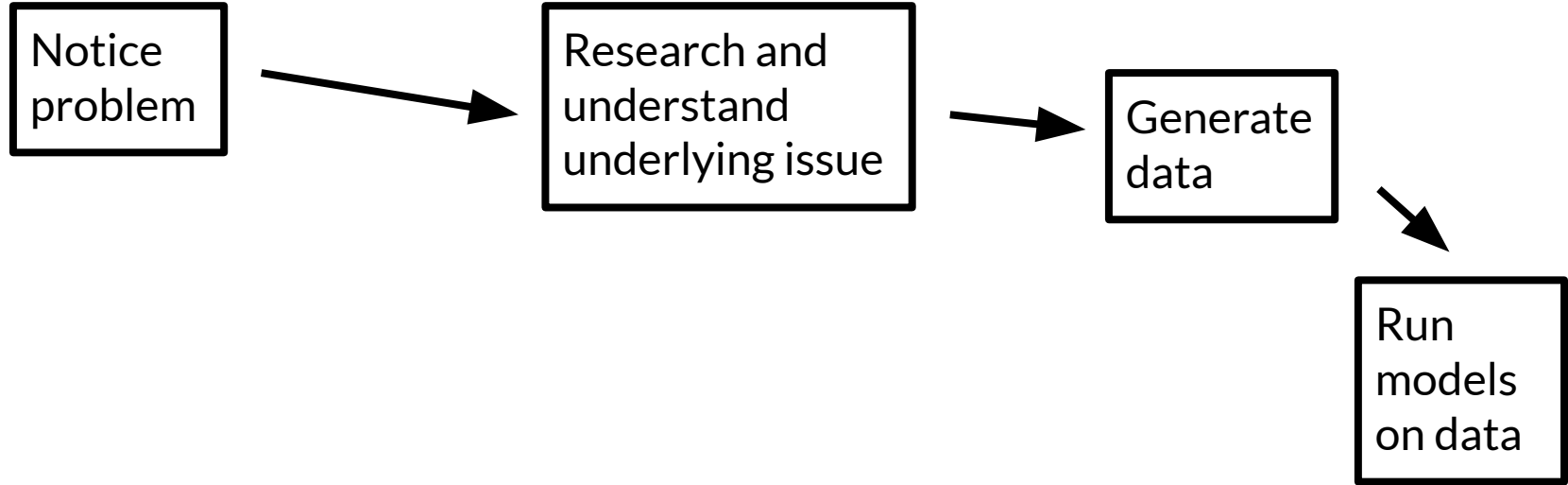
Notice
problem

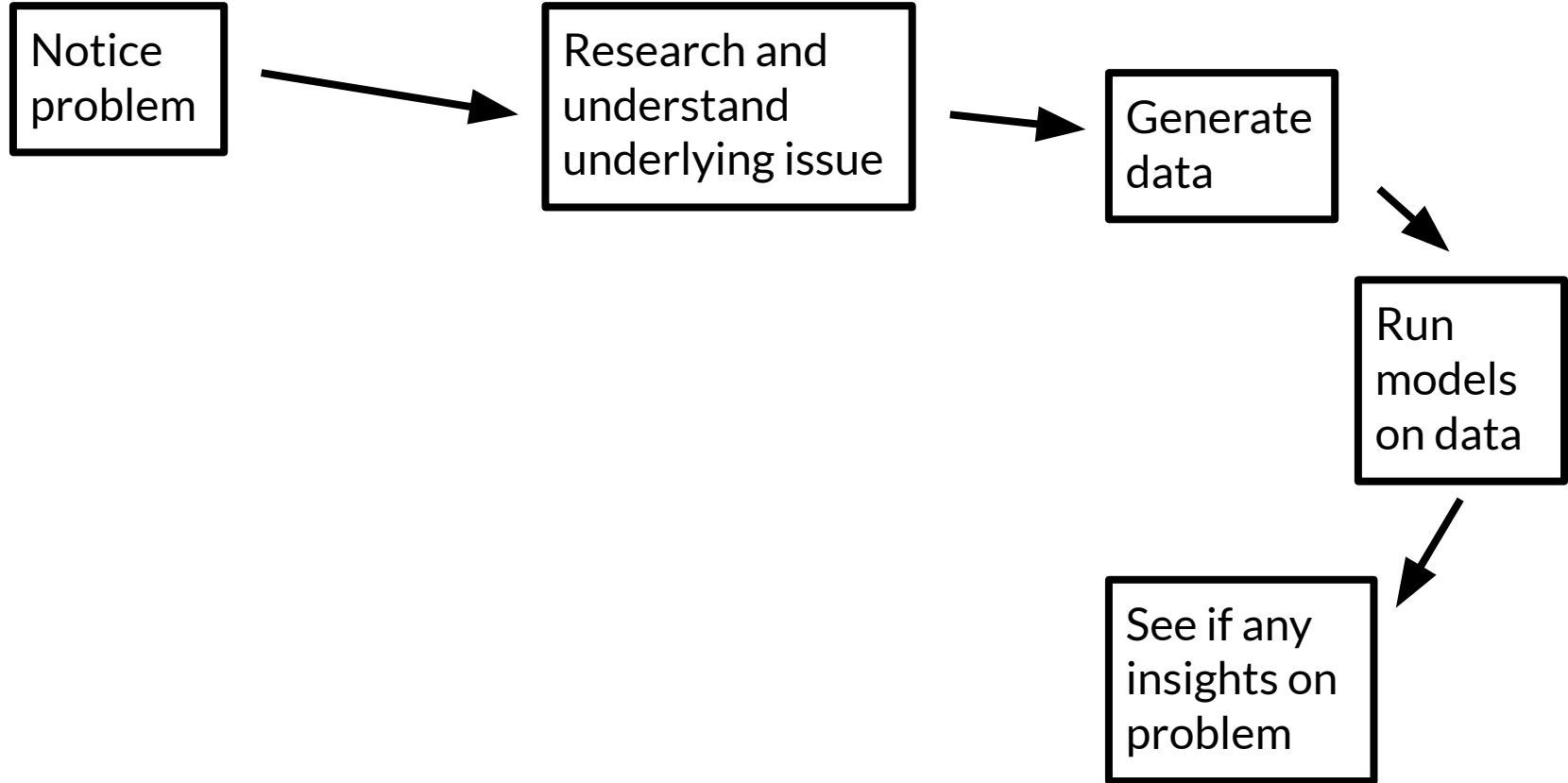


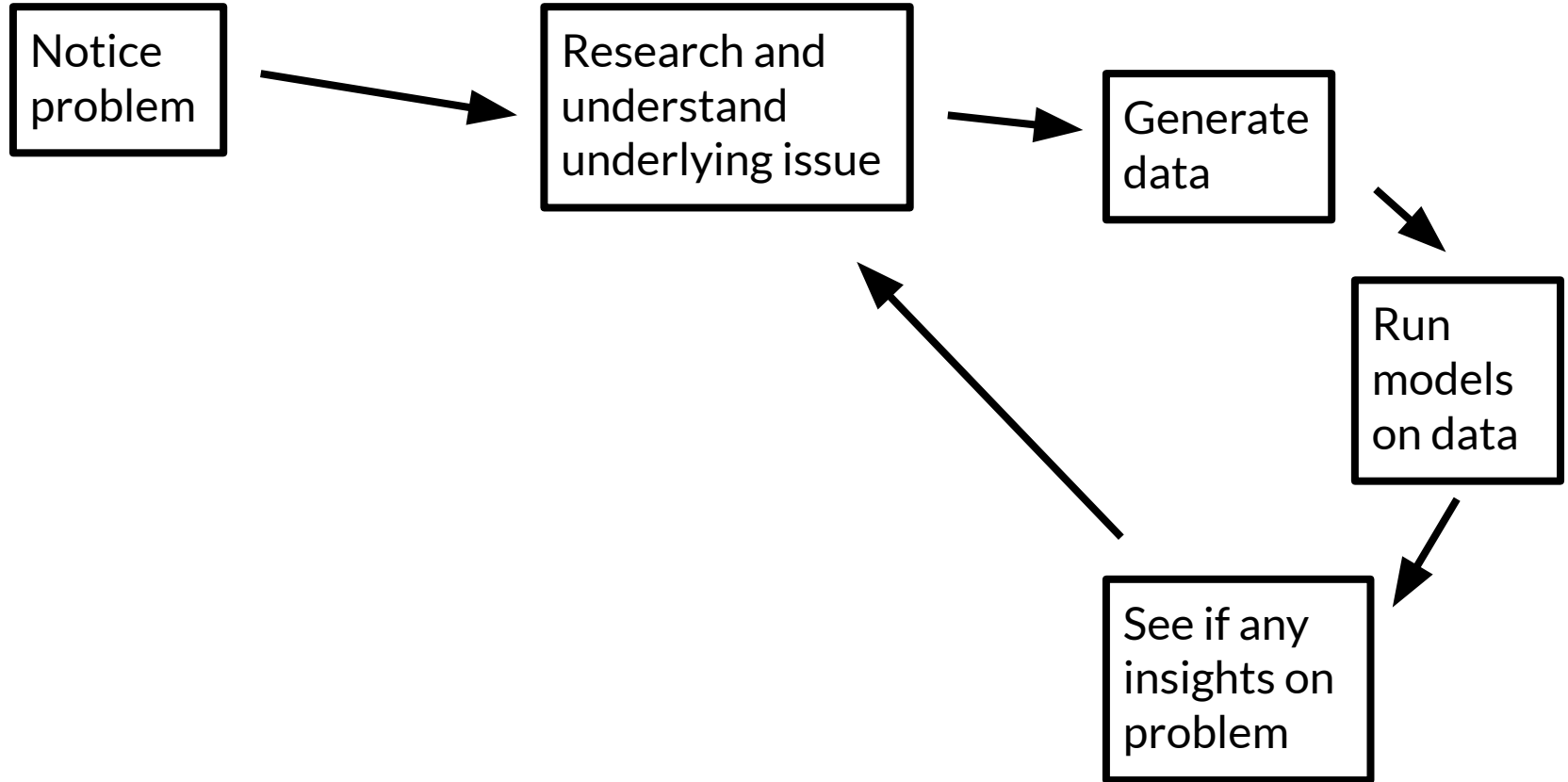
Research and
understand
underlying issue

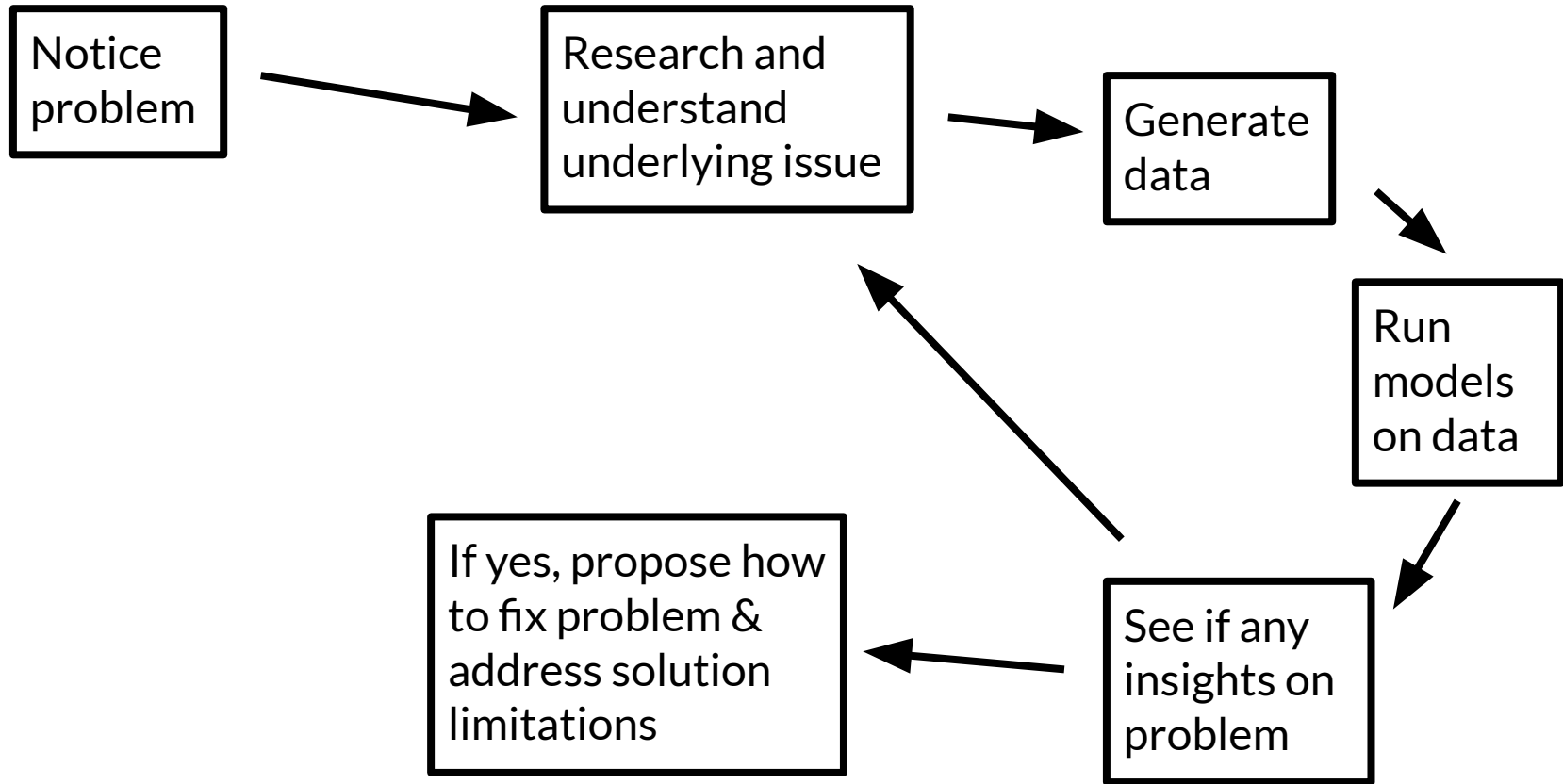


Generate
data







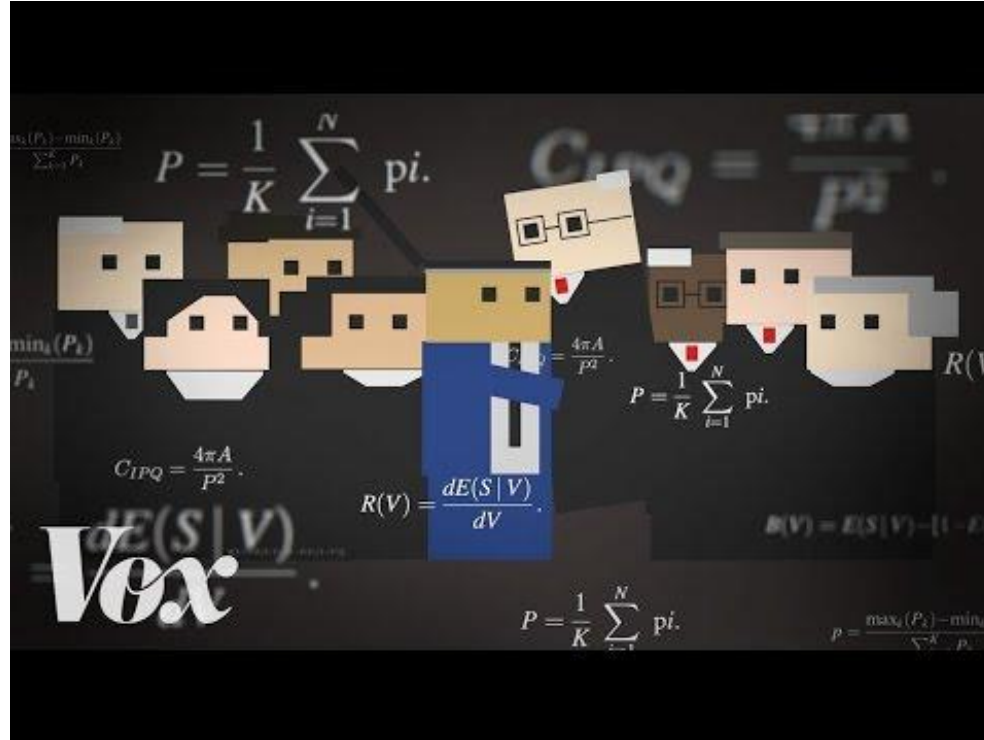


DS Project Examples

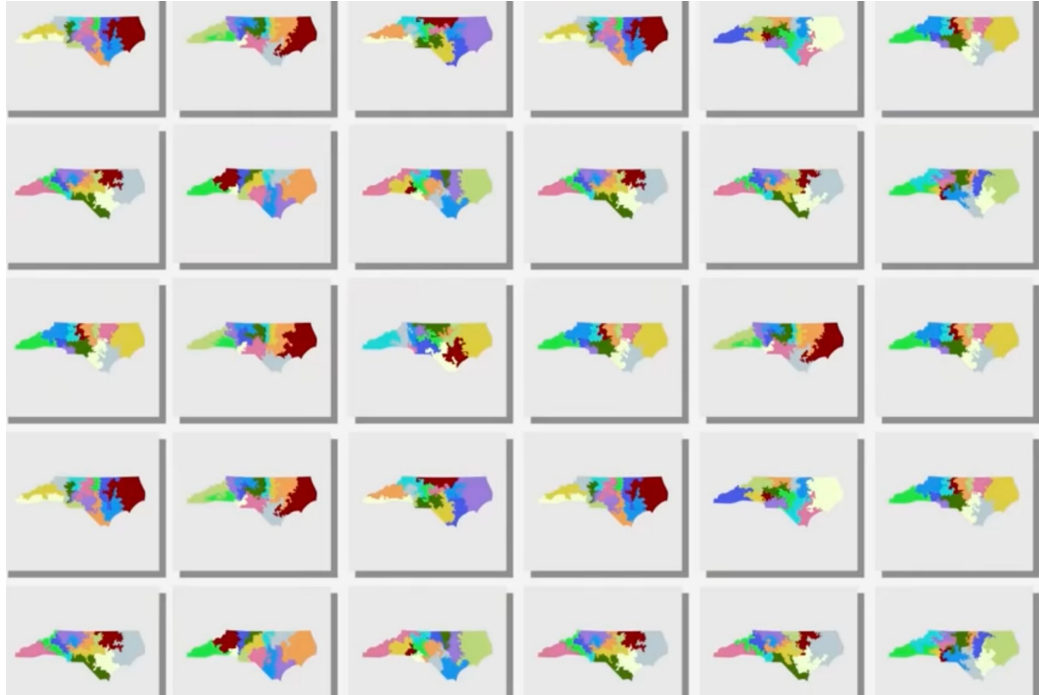
Successful projects often use data to...

1. Answer a question you have about the world
2. Ameliorate an existing problem you have
3. **Make tools to help others solve problems they have**

3. Tool-building: detect gerrymandering



How does this simulation relate to their argument?

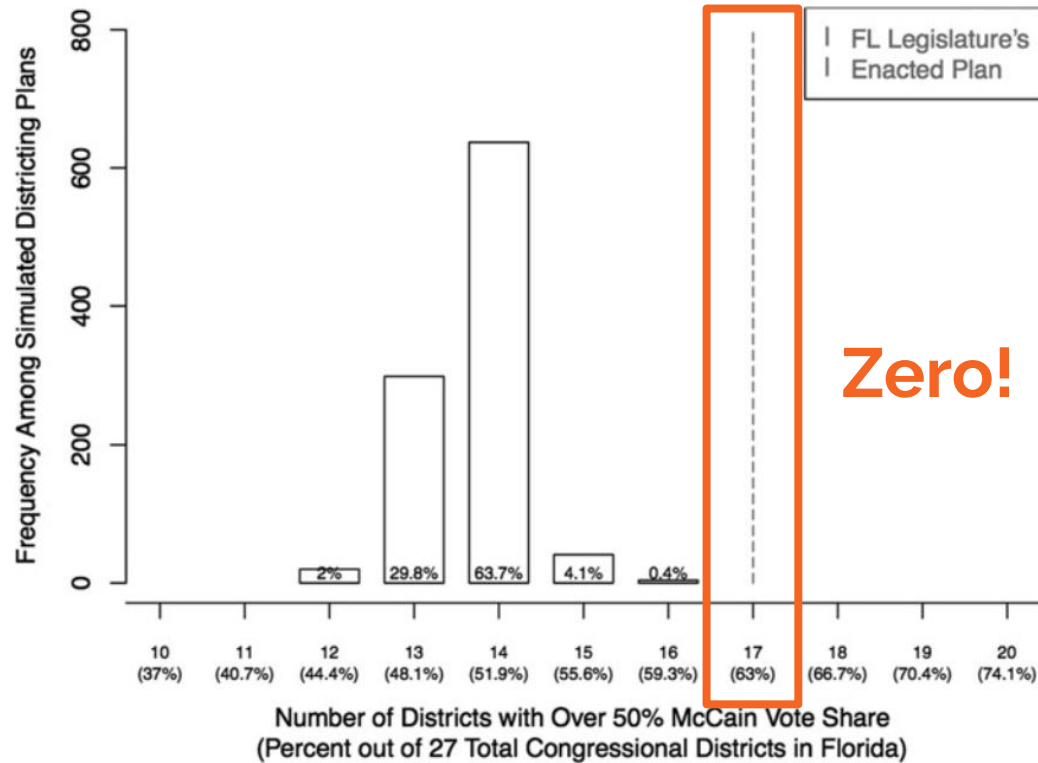


Florida has 27 Congressional districts.

In 2012, the Florida legislature's redistricting plan produced 17 Republican seats.

Academics ran 1,000 simulations.

Guess: How many of the 1,000 simulated maps also gave Republicans 17 wins?



Notice what
tool is
lacking

Notice what
tool is
lacking



Research and
understand
existing issues

Notice what
tool is
lacking



Research and
understand
existing issues



Develop
method

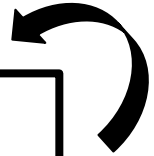
Notice what
tool is
lacking

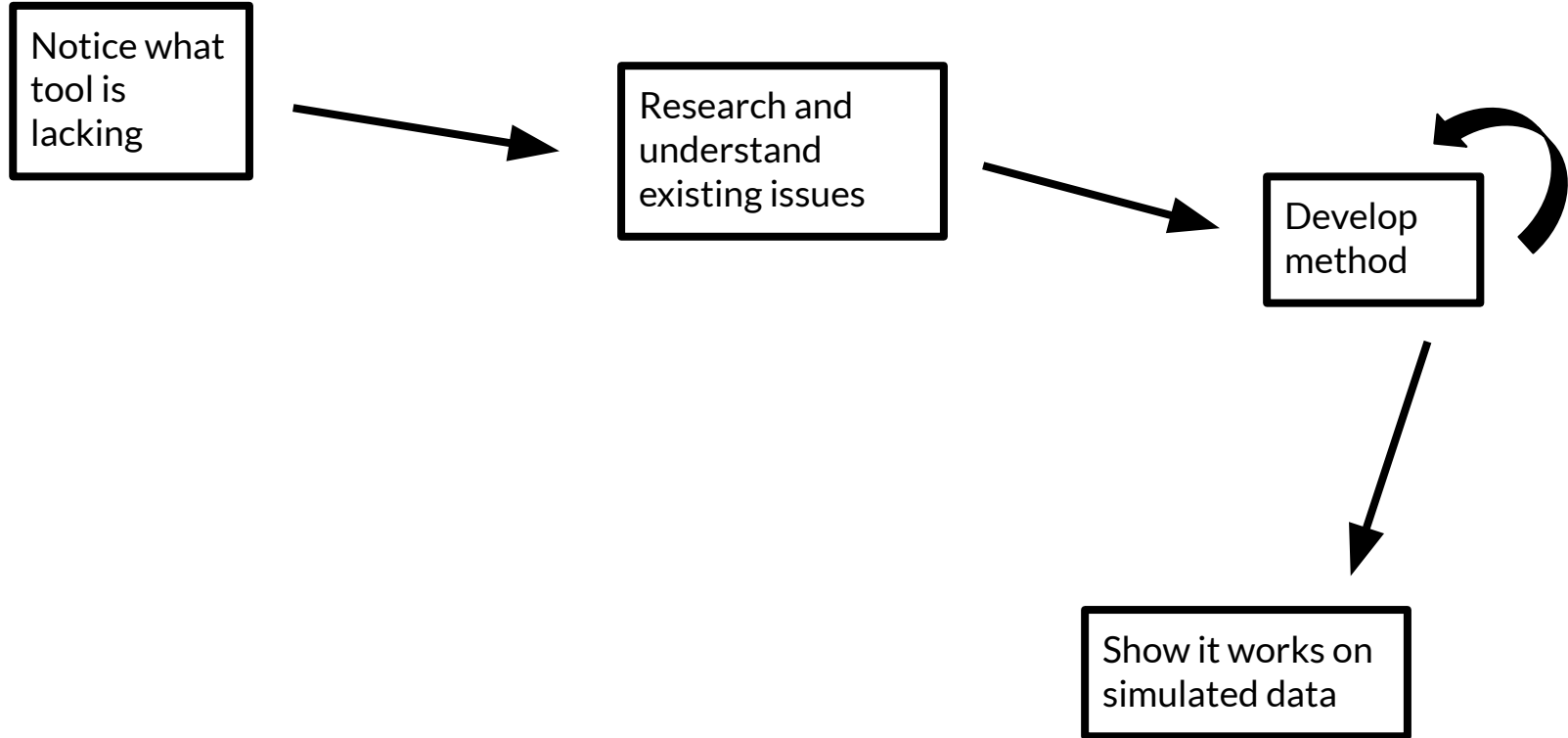


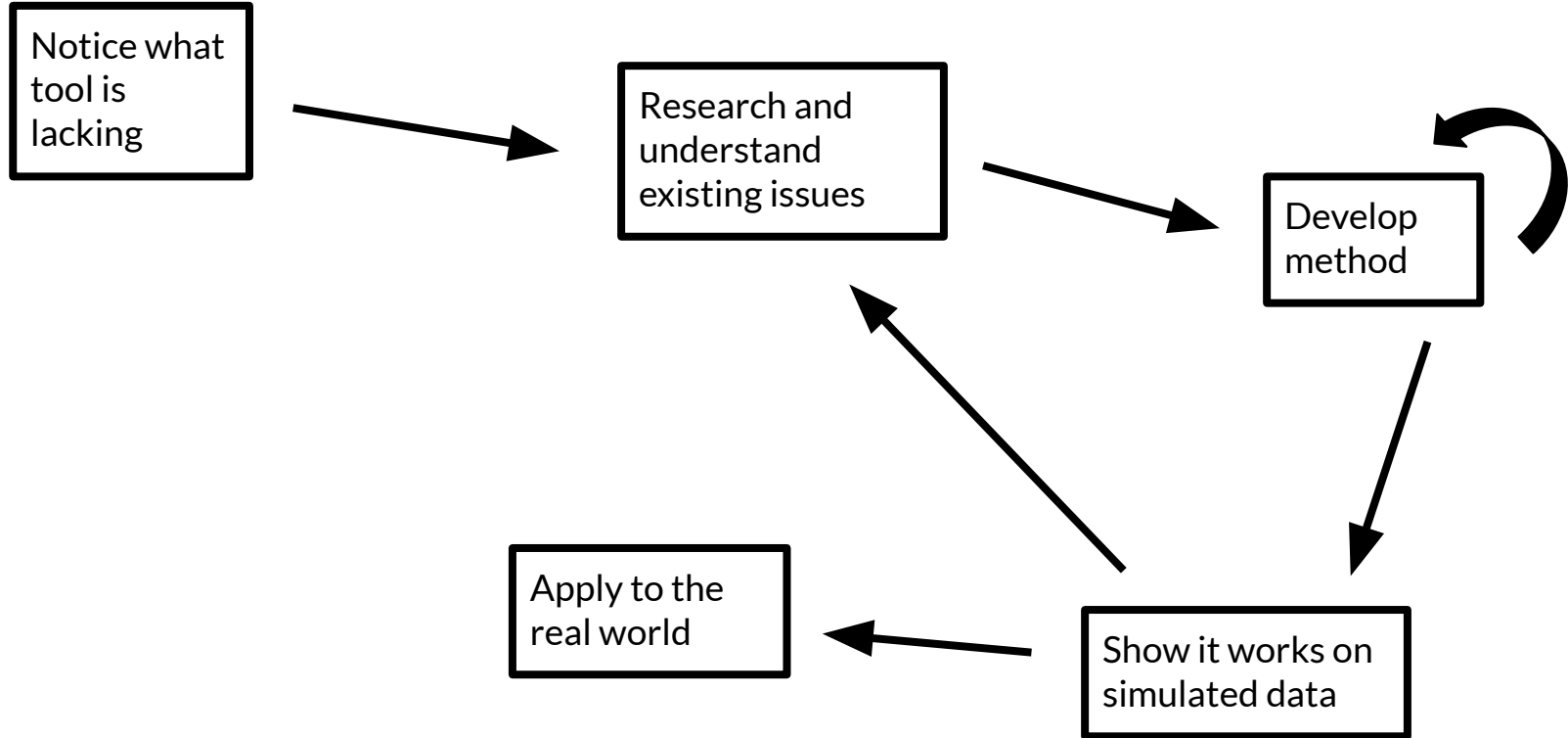
Research and
understand
existing issues



Develop
method







- Social media
- Big tech
- Politics
- Health
- Literature
- Migration →
- Climate Change
- Fashion
- Sports
- ...

Benjamin Q Huynh ¹, Sanjay Basu ² ³

1 min break & Think, Pair, Share

- What types of data projects interest you? Why?

Data, specifically

- What does “data” mean?
- How do we represent it so
 - Humans can understand it, AND
 - Computers can operate on it?

Data, specifically

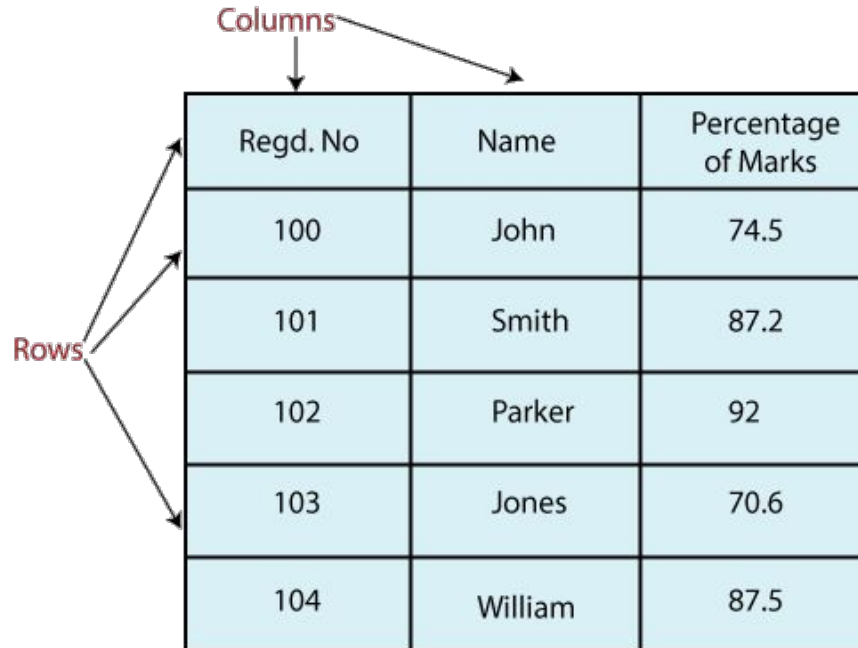
- What does “data” mean?
- How do we represent it so
 - Humans can understand it, AND
 - Computers can operate on it?

nt	Is this an established artist?	artist's big break?
	Yes	No

Data Frame (df)

- The foundation of data analysis!
- Data organized into a 2-dimensional table with rows & columns
- Within-column is always the same data type
- Across columns can be different data types

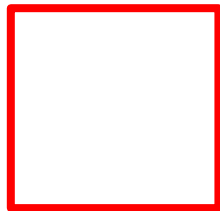
Data Frame (df)



The diagram illustrates a Data Frame (df) as a table. The word "Columns" is written in red above the table, with two arrows pointing to the first and second columns. The word "Rows" is written in red to the left of the table, with four arrows pointing to the first, second, fourth, and fifth rows. The table has three columns: "Regd. No", "Name", and "Percentage of Marks". It contains six rows of data, with the first row being the header.

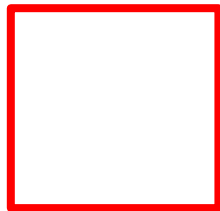
Regd. No	Name	Percentage of Marks
100	John	74.5
101	Smith	87.2
102	Parker	92
103	Jones	70.6
104	William	87.5

Data Frame vs. Spreadsheet



- All data frames can be spreadsheets
- Not all spreadsheets can be data frames
- Why?

Data Frame vs. Spreadsheet



- All data frames can be spreadsheets
- Not all spreadsheets can be data frames
- Why? **Cell colors, formatting, extra headers, total \$ summaries, formulas**

Course Logistics

- Attendance mandatory M/W/F
- **Canvas** for course materials
- **Ed Discussion** for questions
- Deliverables
 - Weekly **homework** (generally: Thursday release, Thursday due)
 - 1 **prelim** (in-class, Oct 2) & 1 **final** exam (determined by the registrar)
 - **Group project** (teams of 2-4, more info on Weds)

How do you learn?

- Go to class
- Read the course materials

How do you learn?

- Go to class & engage with the course material
- Read the course materials

How do you learn?

- Go to class & engage with the course material
- Read the course materials & practice coding!

How do you learn?

- Go to class & engage with the course material
- Read the course materials & practice coding!



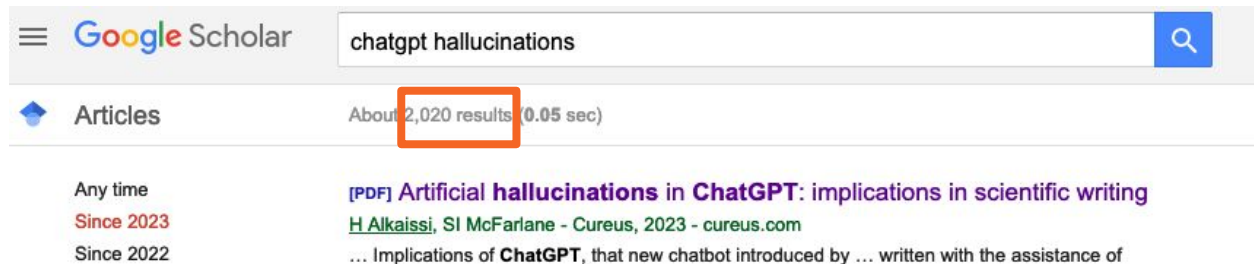
Pre-2950



Post-2950

ChatGPT != going to the gym

- ChatGPT is often confidently incorrect



- INFO 2950 will give you the skills to determine whether ChatGPT outputs are correct or not

ChatGPT won't solve your problems

- Data science plug-ins like Code Interpreter exist, and are *super dangerous*



Teresa Kubacka
@paniterka_ch

...

Code Interpreter really is a parrot. I tried it with the Boston Dataset. It doesn't understand a thing about data, it just regurgitates a statistically plausible tutorial. It performs poorly with variable understanding, not noticing a problem with "Bk" unless asked specifically



Teresa Kubacka @paniterka_ch · Jul 10

...

You can basically type "yes" and "what do you recommend" all the way through, and CI will gladly create a very sophisticated, very discriminatory model, even pretending it's all good and fair (unless you really insist it's not), and it will happily assist you in deploying it.

Supermarket AI Offers Recipe for Mom's Famous Mustard Gas

An AI from a New Zealand grocery chain gave one user a recipe for an "Aromatic Water Mix" that included bleach and ammonia as ingredients.

By **Kyle Barr** Updated Friday 10:53AM | Comments (30)



AI could provide you instructions how to deal with chlorine gas, it didn't give you precise instructions how to create it in the first place.

Illustration: Andrey_Popov (Shutterstock)

Final project

- Phase 0 due Sep 7
 - Form your team next week: 2-4 people
 - Agree on how you will work together
- Phase 1 due Sep 21
 - Research datasets & brainstorm ideas
- **Phase 2 due Oct 19**
 - Report about a datasets with description and viz
- Phase 3 due Nov 2
 - Register a hypothesis
- Phase 4 due Nov 16
 - Draft of complete project with statistical model
- **Phase 5 due Dec 4**
 - Revised draft of complete project

HW1

- Install Python 3 Anaconda by Friday 08/25/2022
 - Recommend VSCode for IDE
 - Instructions are on Canvas > Modules > Course Policies
- Be able to answer:
 - What version of Python did you install?
 - What is a conda environment?
 - Why might you use different conda environments?
 - What IDE are you using?

**What questions do you have
about INFO 2950?**

-
- 1. Cap your marker**
 - 2. Return marker & whiteboards**
 - 3. Throw your tissue in the trash**