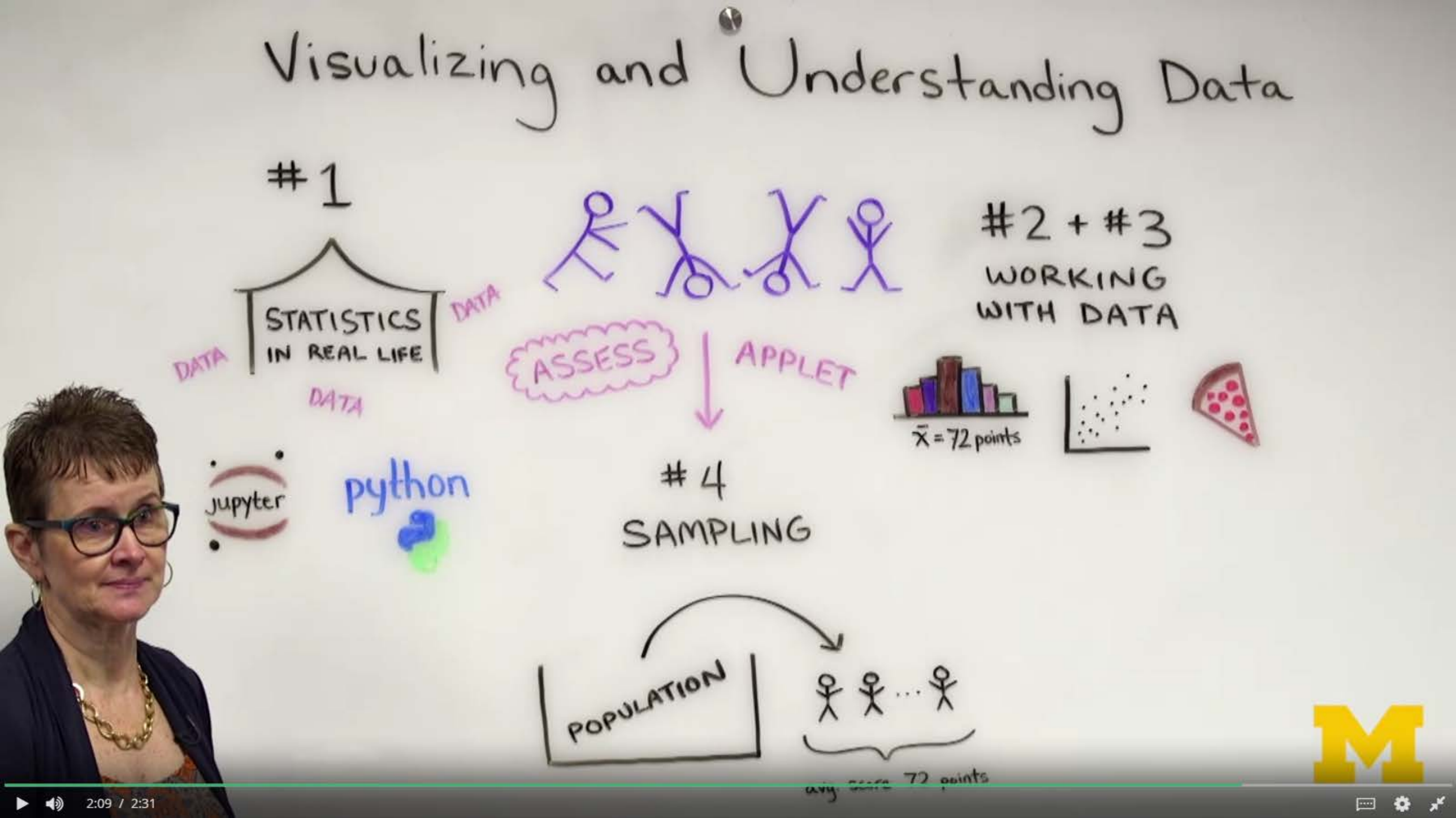
Visualizing and Understanding Data

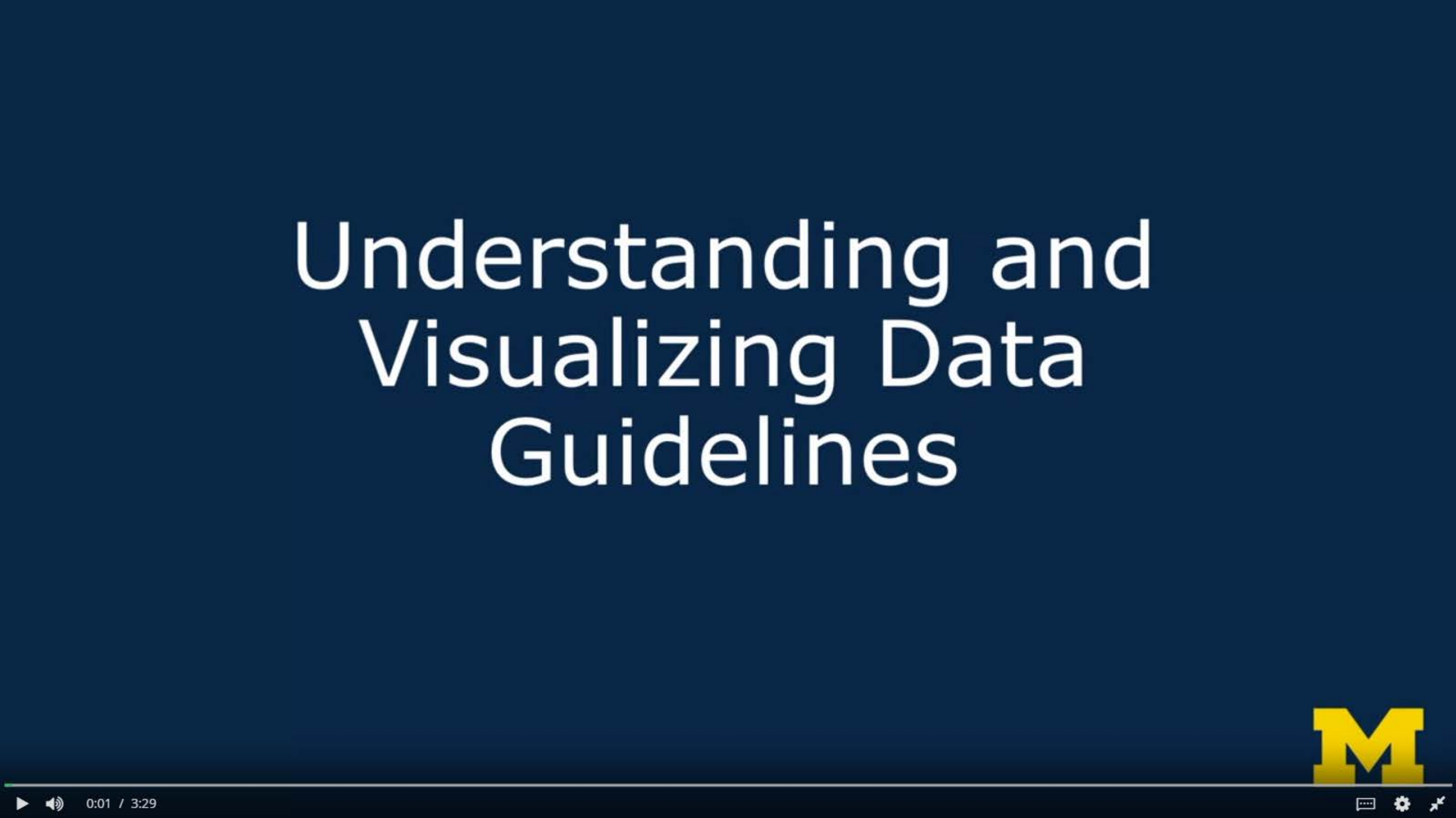
Dr. Brenda Gunderson

Lecturer IV in Statistics and Research Fellow, Statistics,
College of Literature, Science, and the Arts

Visualizing and Understanding Data



Understanding and Visualizing Data Guidelines



Guideline

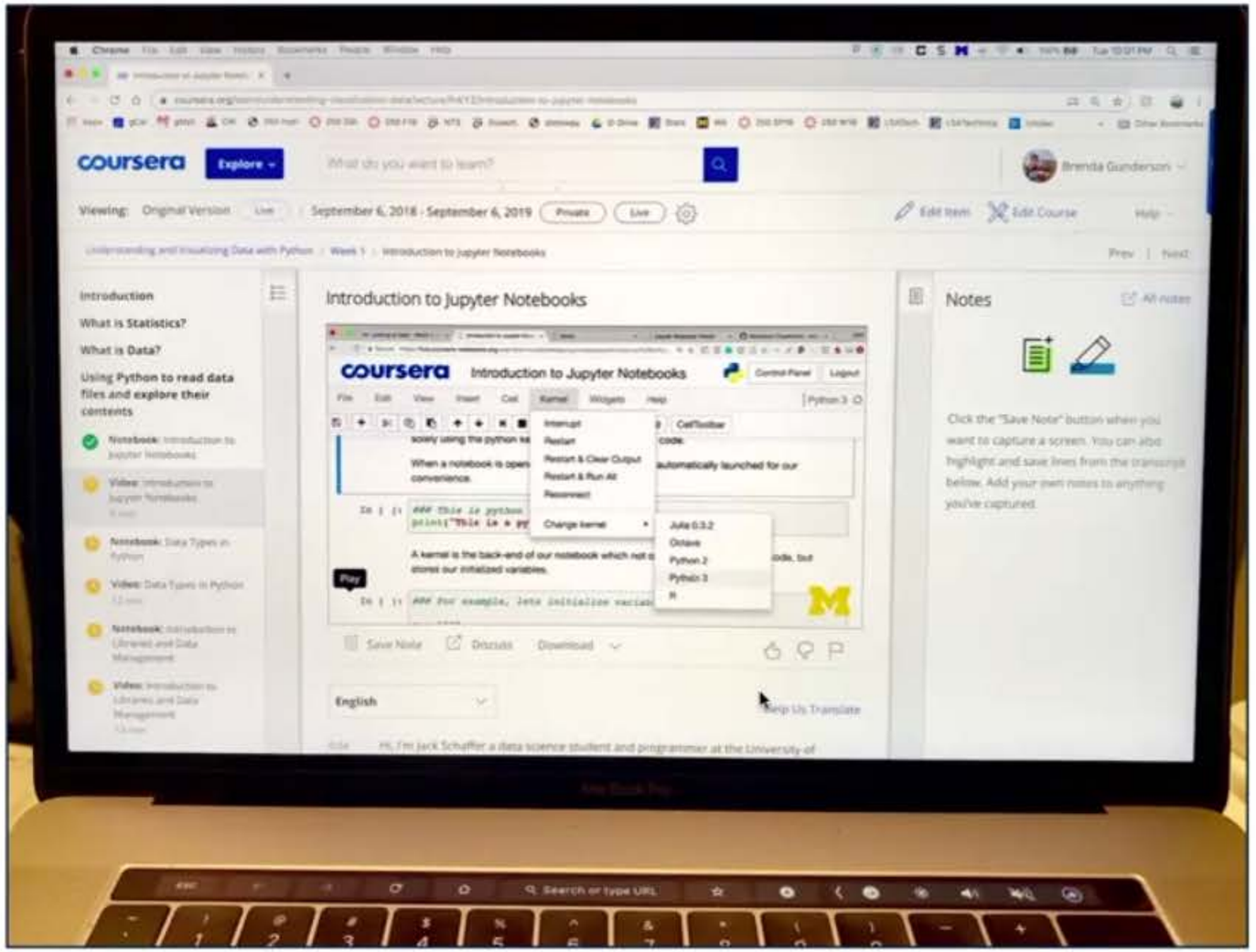
#1 Don't Wait to open Notebooks

Keep open simultaneously
as you go through various lectures
Pause lectures ⟷ **Try it** in Notebook

Larger or Multiple Monitors Can Help

go through the various lectures.

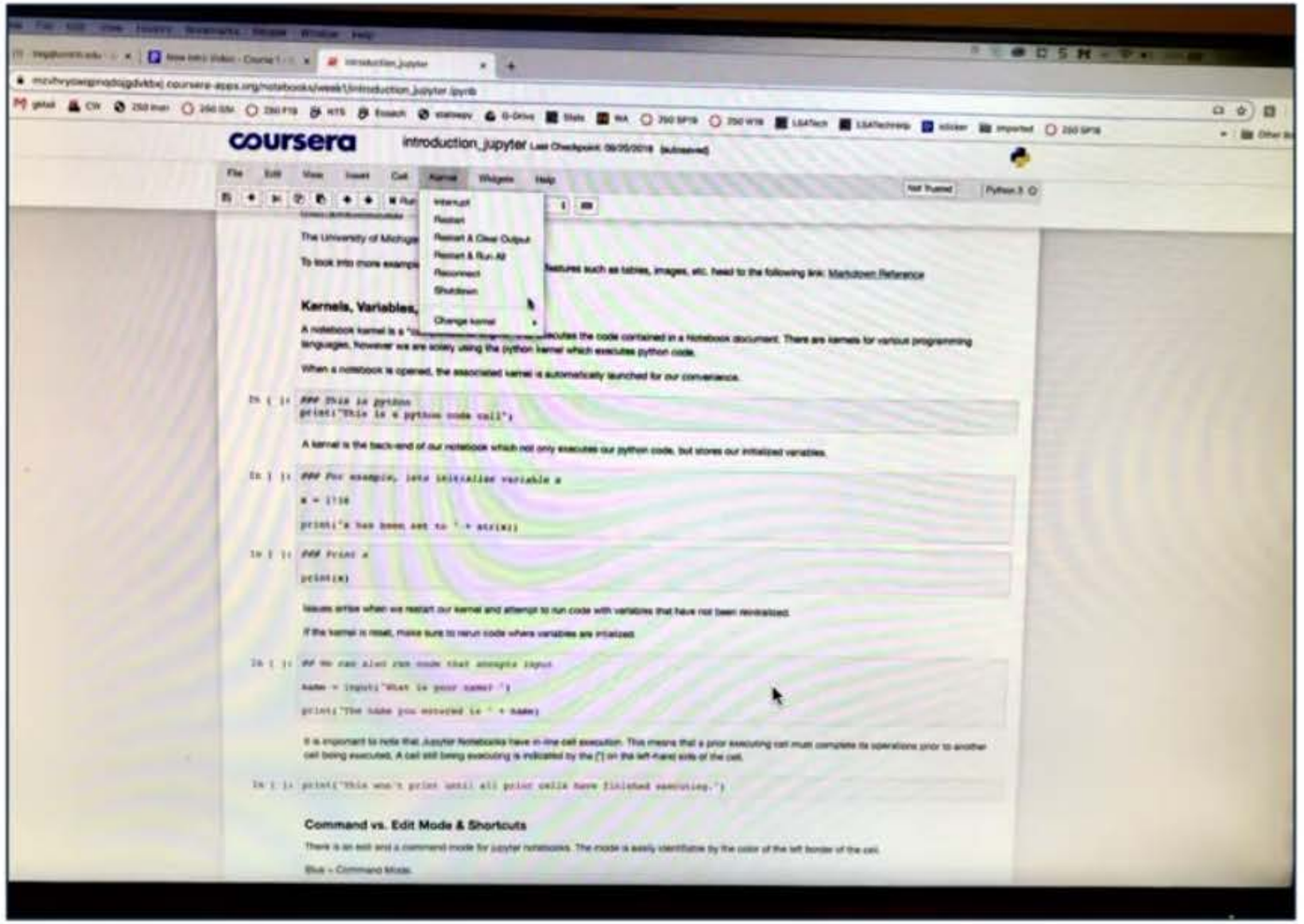
Video:
guiding me through what is a Notebook Kernel and how to change the Kernel through dropdown menu



a notebook kernel is and how to change the kernel through a drop-down menu.

1:41 / 3:29

My Notebook:
Following along and trying out the Kernel dropdown menu



clicking and
following along trying out the same steps.

Week 4

- **Focus:** understand where data come from when you prepare for data analysis

Guideline

#2 Make Good Use of In Video Questions

Variety added to make ideas more concrete

of in video questions that were added
to make these ideas more concrete.

Questions are Welcomed

Guideline

#3 Check archived FAQs and Discussion Forums

feel free to ask the more complex question
too, it will be routed to the instructors.

More Difficult Topics

Guideline

#4 Check out Additional Readings under Course Resources

So be sure to check out
the Additional Reading section

About Our Datasets

National Health and Nutrition Examination Survey (NHANES)

The National Health and Nutrition Examination Survey (NHANES) is a program of studies designed to assess the health and nutritional status of adults and children in the United States. The survey is unique in that it combines interviews and physical examinations.

For two-year cycles (e.g., 2015-2016), cross-sectional national samples of individuals living in the United States are invited to participate in both aspects of the data collection. The data produced are widely considered by the research community as among the most important scientific indicators of the health and well-being of the U.S. population.

For this specialization, we will be analyzing data collected from a national sample of individuals during the 2015-2016 cycle.

The NHANES dataset that we will be using can be downloaded from the Resources section in the left column of the course layout.

The Cartwheel Dataset

A simpler and smaller set of data, the Cartwheel Dataset was collected by our very own course team at the University of Michigan. It includes the following information: age, gender, glasses-wearing or not, height, weight, wingspan (arm length), completion, cartwheel distance, and overall cartwheel score.

The Cartwheel dataset that we will be using can be downloaded from the Resources section in the left column of the course layout.

Seaborn Tips Dataset

The Seaborn package comes with a number of packages that one can use for analysis. One of these pre-loaded datasets is the Tips Dataset which contains data on the meal tipping amounts of various individuals depending on the size of their party, whether they were a smoker or not, their gender, the day of the week, the time of the day, and a variety of other variables. This dataset is provided to demonstrate statistical concepts and is not meant to provide insight into the tipping behavior of any particular group of people.

What is Statistics?

Dr. Brenda Gunderson

Lecturer IV in Statistics and Research Fellow, Statistics,
College of Literature, Science, and the Arts

What is Statistics?

- **Methodological** subject encompassing all aspects of **learning from data**.

  tools and methods
  for working with and understanding data

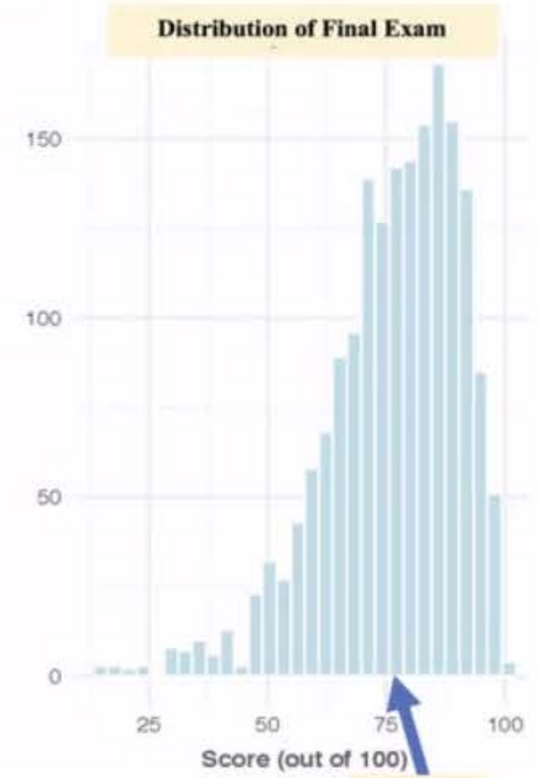
- **Statisticians** apply and develop data analysis methods,
  seek to understand their properties…

  …when do these tools provide ***insight?***
  …when are they ***possibly misleading?***

- **Researchers** and **workers** apply and extend statistical methodology,
  and contribute new ideas and methods for conducting data analysis.

A "Statistic" and the field of "Statistics"

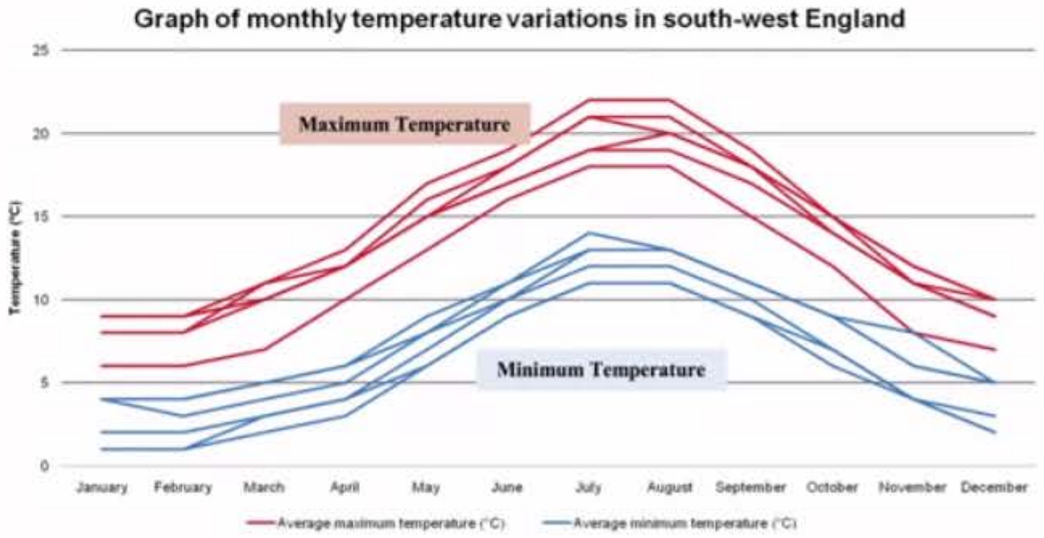
- A **statistic** ~ numerical or graphical summary of a collection of data.
  - Average score on final exam



This could be the average score on the final exam

A "Statistic" and the field of "Statistics"

- A **statistic** ~ numerical or graphical summary of a collection of data.
  - Average score on final exam
  - Minimum temperature at a location over year



A "Statistic" and the field of "Statistics"

- A **statistic** ~ numerical or graphical summary of a collection of data.
  - Average score on final exam
  - Minimum temperature at a location over year
  - Proportion of people who are retired

our survey that might allow us to extend to what that might be in the city.

A "Statistic" and the field of "Statistics"

- A **statistic** ~ numerical or graphical summary of a collection of data.
  - Average score on final exam
  - Minimum temperature at a location over year
  - Proportion of people who are retired
- **Statistics** ~ academic discipline focusing on research methodology. Statisticians develop new statistical tools, calculate statistics from data, and collaborate with subject-matter experts to interpret them.

is that academic discipline that's focusing on research methodology.

The Landscape of Statistics

Evolving and dynamic field ~ Emerging **challenges** and **opportunities**

- **Properties** of statistical methods are under **continuing study**
- New application areas → **development** of new analytic methods
- New types of sensors → **new types of data**
- Advances in **computing** → sophisticated analyses on Big Data

Of course, we're relying often on those advances in computing.

Brain by juliana takeuchi from the Noun Project
Computer by David from the Noun Project

Perspectives on Statistical Science

Statistics is a "**big tent**" discipline ~ incorporates new ideas from theory, practice, allied fields.

Different Perspectives:

- "art of summarizing data"
- "science of uncertainty"
- "science of decisions"
- "science of variation"
- "art of forecasting"
- "science of measurement"
- "basis for principled data collection"

statistics and how people who work with data view that field.

Statistics as the "art of summarizing data"

- Data can be **overwhelming**
- Making sense of data usually involves **reduction** and **summarization**

make a dataset **comprehensible** to human observer

always **depends primarily on goals of "data consumer"** to be meaningful -- many approaches

rigorous and effective methods for summarizing data.

Statistics as the "science of uncertainty"

- Data can be **misleading**
- Statistics provides framework for **assessing whether claims based on data are meaningful**
- Uncertainty is inevitable, but it is highly desirable to **quantify how far our reported findings may fall from "the truth"**

Many public opinion polls report **± margin of error**

→ potential discrepancy between reported and actual states of public opinion

how far away reported findings may be from the truth.

Statistics as the "science of decisions"

- Understanding data is important →
  only consequential if we act on what we have learned
- **Decision-making =** ultimate goal of any statistical analysis
- We **make decisions in face of uncertainty**!
  What are costs and benefits of different approaches?

 → at higher than average risk for cancer…
should they undergo preventative procedure?

For example, a person who finds that they might be at

Statistics as the "science of variation"

- Often focus on most typical or "**central**" value
- Great emphasis on understanding **variation** in data!

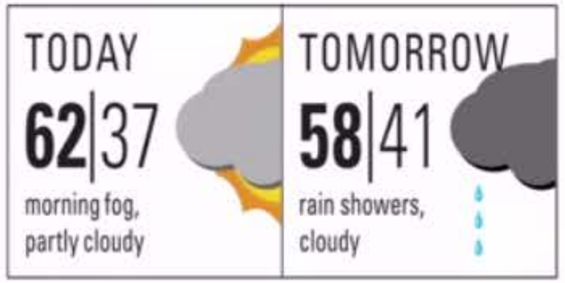
Average American has around $6000 of credit card debt
→ central value of credit card debt in US population.

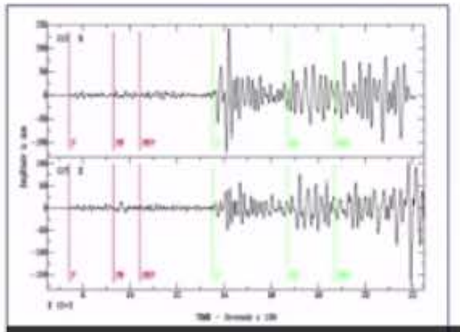
10% of Americans have more than $30,000 in credit card debt
→ variation of credit card debt in US population.

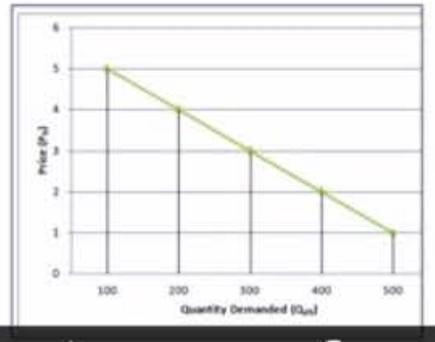
the variability in credit card debt for our population.

Statistics as the "art of forecasting"

- Forecasting or prediction = central tasks in statistics
- ***Cannot*** know future with absolute certainty, but efficient use of available data
  → ***can*** sometimes make accurate predictions about future







Predicting the outcome of an election,

Statistics as the "science of measurement"

- **High accuracy**: person's age or height
- **More difficult**: blood pressure (*varies minute to minute*)
- **Harder**: "mood", "political ideology", "personality"

Statistics: major role in **constructing and evaluating rigorous approaches for measuring difficult-to-define concepts** and in **assessing quality**.

much harder to define and then quantify.

Statistics as the "basis for principled data collection"

- Data often expensive and difficult to collect
- Resource limitations → collect least data possible

Statistics: provides a rational way to manage this trade-off

wanting more data, but knowing and allowing those resource limitations.

History of Statistics Milestones

Ancient Times:

Data Collection on harvests floods population sizes

1700's:

Probability Theory
→ randomness and variation

19th Century:

Modern Statistics emerges, via genetics demography economics

20th Century:

Statistical Theory advances, new application areas, **computers**

21st Century:

"massive data", "data science" "machine learning"

massive data, data science, and machine learning.

Statistics and its Allied Fields

Computer science: algorithms, data structures for working with data, programming languages for manipulating data.

Mathematics: language and notation for expressing statistical concepts concisely, tools for understanding properties of statistical methods.

Probability theory: branch of mathematics ~ crucial part of foundations of statistics – to express ideas about randomness and uncertainty.

Data Science: database management, machine learning, computational infrastructure to carry out data analysis.

that infrastructure to be able to carry out data analysis.

Resource: This is Statistics

A great resource that you can explore is the "This is Statistics" website, created by the American Statistical Association. This insightful and motivating campaign has countless links, videos, and resources to raise awareness of the wide variety of fascinating careers within statistics.

Discover how you can change the world in this ever-growing profession while having fun and earning great money. We highly recommend you investigate everything this website has to offer.

A great starting point is this compelling interview by Roger Peng, a statistician and professor at Johns Hopkins University. He delves into a few of the many exciting components that make the field of statistics so desirable.

Cool Stuff in Data

Julie Deeke

Statistics with Python Course Developer

Data can be Numbers

National Health and Nutrition Examination Survey

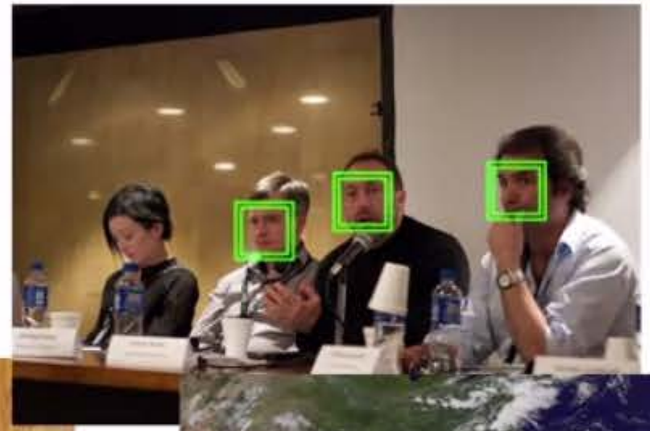
seqn	ridstatr	riagendr	RIDRETH1	dmdmartl	WTINT2YR	WTMEC2YR
62161	2	1	3	5	102641.406	104236.583
62162	2	2	1	NA	15457.737	16116.354
62163	2	1	5	NA	7397.685	7869.485
62164	2	2	3	1	127351.373	127965.226
62165	2	2	4	NA	12209.745	13384.042
62166	2	1	3	NA	60593.637	64068.123
62167	2	1	5	NA	5024.465	5303.683
62168	2	1	5	NA	5897.025	6245.044
62169	2	1	5	5	14391.778	14783.601
62170	2	1	5	NA	7794.527	8291.637

and has numbers representing whether the person is female or male.

https://www.cdc.gov/nchs/nhanes/index.htm

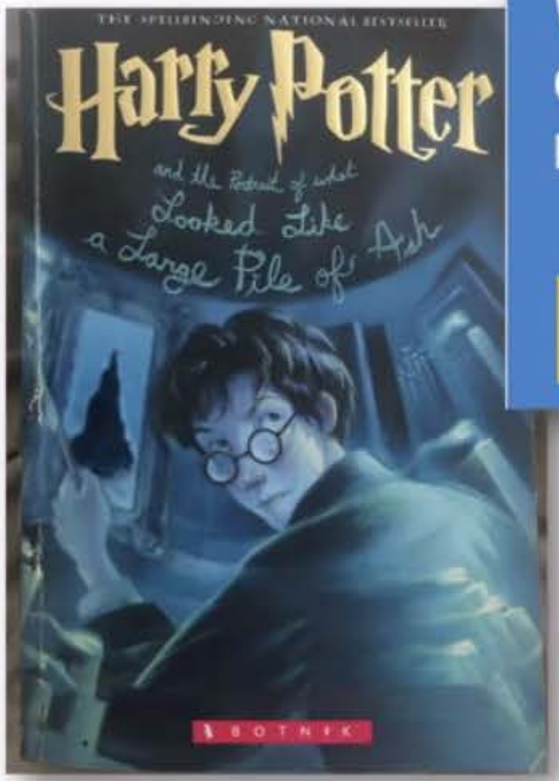
Data can be images

https://quickdraw.withgoogle.com/



One quick thing that I'll show you is this Quick Draw! with Google.

Data can be Words



What is an electronic health record (EHR)?

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]
[1]Boston University, 8 Saint Mary's Street, Boston, MA
[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	EDUCATION
MUSIC	BUDGET	CHILD	TEACHERS
MOVIE	BILLION	YEARS	HIGH
PLAY	FEDERAL	FAMILIES	PUBLIC
MUSICAL	YEAR	WORK	TEACHER
BEST	SPENDING	PARENTS	BENNETT
ACTOR	NEW	SAYS	MANIGAT
FIRST	STATE	FAMILY	NAMPHY
YORK	PLAN	WELFARE	STATE
OPERA	MONEY	MEN	PRESIDENT
THEATER	PROGRAMS	PERCENT	ELEMENTARY
ACTRESS	GOVERNMENT	CARE	HAITI
LOVE	CONGRESS	LIFE	

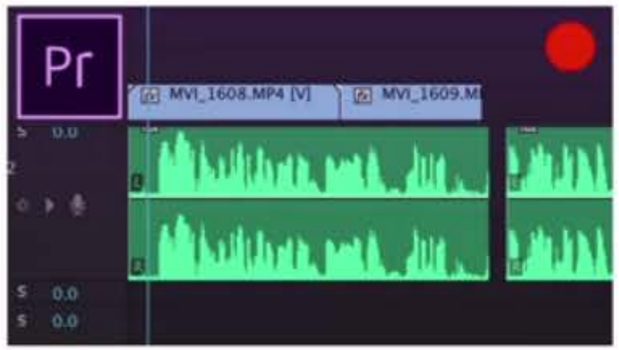
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

We want to try to make those word embeddings potentially

Image source: https://www.healthit.gov/topic/health-it-and-health-information-exchange-basics/health-it-and-health-information-exchange
Image source: http://botnik.org/content/harry-potter.html
Image source: Page 17 http://www.jmlr.org/papers/volume3/bei03a/bei03a.pdf
Image source: https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf

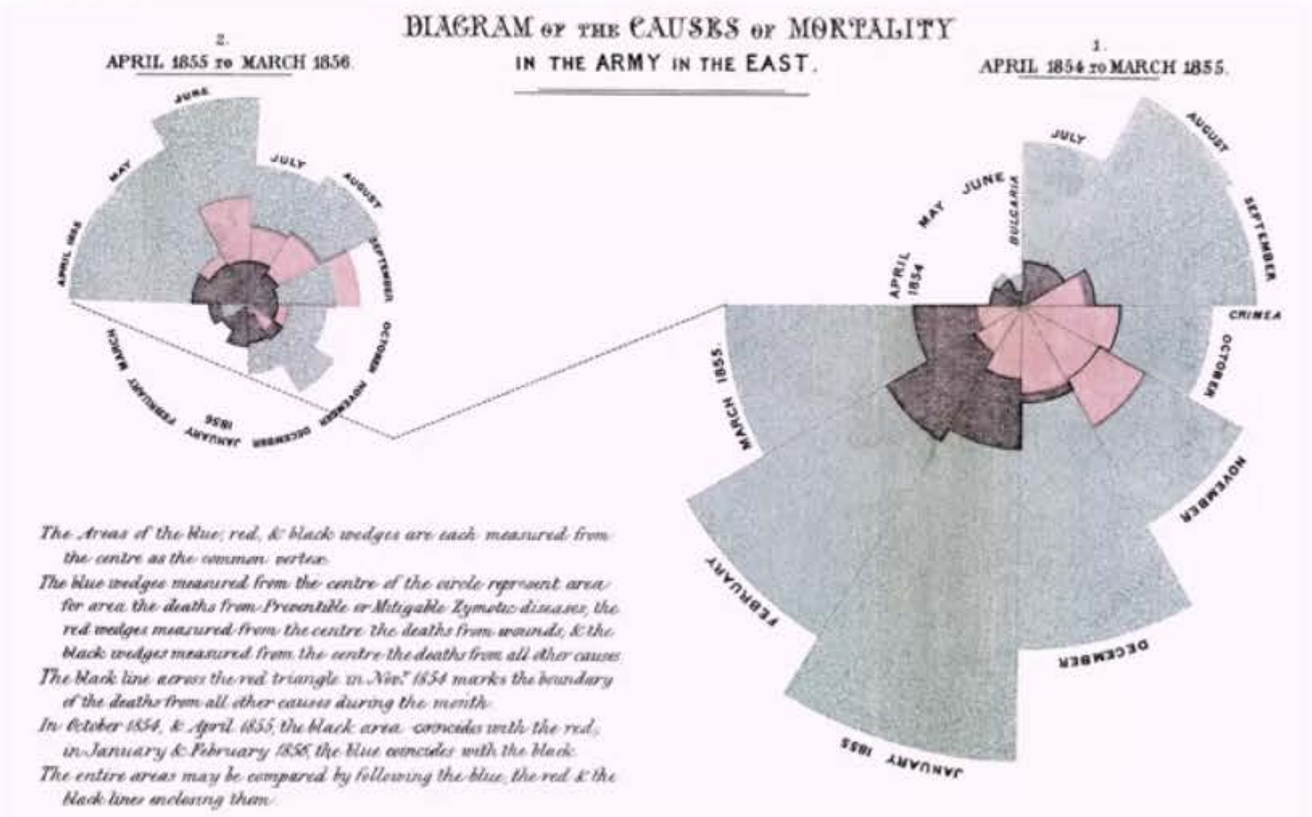
Data can be Audio



their voice around so it sounds like somebody is saying something when they're not.

Image source: https://www.youtube.com/watch?v=QGmzll4mmlY

Historical Example



and how these different death rates changed based on the season.

Let's Play with Data!

To get us started, here are some sites you can try out to play with data!

- Want to see how different Americans spend their days? Check out this cool website to see some interesting ways to visualize data: https://flowingdata.com/2015/12/15/a-day-in-the-life-of-americans/. *In addition to reading the article, you can interact with the data visualization. Try changing the speed from slow to fast to speed up past the morning times.*
- You can also compare different occupations over time here: https://flowingdata.com/2017/05/17/american-workday/
- You can also break down the data into different subpopulations using this interactive tool here: https://flowingdata.com/2015/11/30/most-common-use-of-time-by-age-and-sex/. *With this visualization, you can change the time, sex, and age. This can be useful to answer a questions like: How do older females (65+) spend their mornings (10:00-10:29 am) compared to younger females (ages 15-24)?*