

DATA SCIENCE

Brian Chung

WHO ARE WE?



BRIAN CHUNG, INSTRUCTOR

Brian is a researcher in the field of quantitative finance. He has worked at Citadel, LLC researching trading signals and building prediction models.

He graduated with a BS in Electrical Engineering from University of Illinois-Urbana Champaign and an MS from Stanford University. When not in front of a computer, he enjoys motorcycling, CrossFit, and cooking with various gadgets.

WHO ARE WE?



SCOTT LITTLE, EXPERT IN RESIDENCE

Scott Little is a data scientist who likes working with physical sensor data. Recently, he completed a project that predicts solar power from satellite imagery and ground photometer sensors. He has a PhD in Physics from the University of Toledo, where he specialized in thin-film photovoltaic solar cells. For fun he enjoys cycling, dreaming, electronics, quadcopters, neurohacking and making things at Pumping Station: One, the local hackerspace.

WHO ARE YOU?

3 minutes:

- ▶ Turn to a person next to you and share your answers
- ▶ You will introduce them to the class 😊

Questions:

- ▶ What is your name?
- ▶ What industry do you work in or what field do you study?
- ▶ What are you most excited to learn in this class?
- ▶ What is a hobby or interest of yours?

AGENDA

- Logistics
- Course Philosophy
- What is Data Science?
- Machine Learning taxonomy
- Project Discussion

LOGISTICS

EXERCISE #1: BOOKMARK THIS PAGE

[HTTPS://GITHUB.COM/BRIANCHANDBOUND/GA-DS](https://github.com/Brianchandbound/GA-DS)

The course website has all the information regarding logistics. If you have a course question not answered, please email gadschicago@gmail.com

Website Topics:

Course logistics

Schedule

Project

ADDITIONAL COURSE EXPECTATIONS

Attendance / late policy

Computer / Phone Use

Participation before, during, and outside of class

Requesting help & Getting feedback

Treating other students with respect and helping others

COURSE PHILOSOPHY

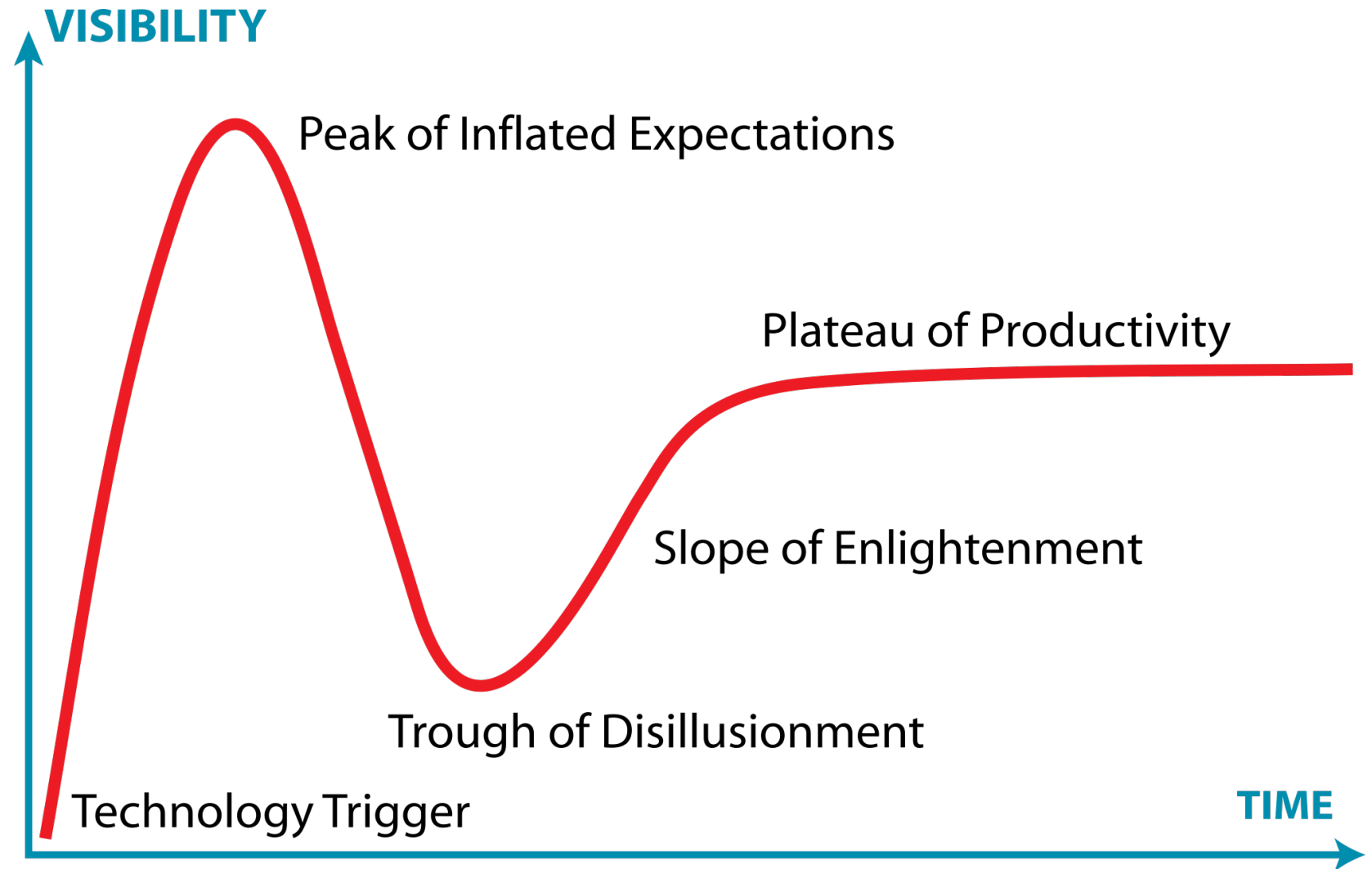
THIS IS NOT THE END



COURSE PHILOSOPHY

THIS IS NOT THE END

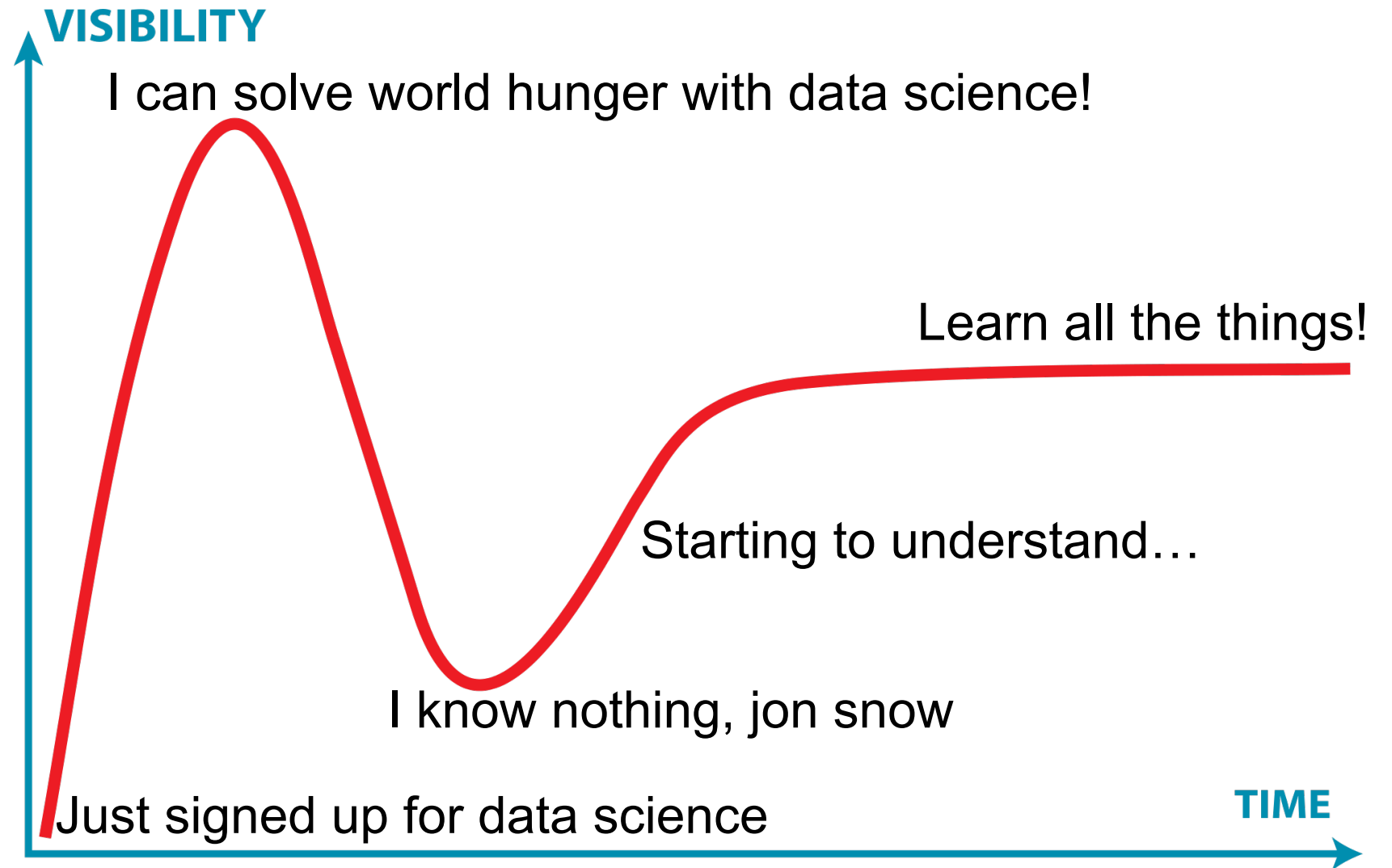
DATA SCIENCE IS HARD



COURSE PHILOSOPHY

THIS IS NOT THE END

DATA SCIENCE IS HARD



COURSE PHILOSOPHY

THIS IS NOT THE END

DATA SCIENCE IS HARD

**SEEK AND YE SHALL FIND
(HELP)**



COURSE PHILOSOPHY

THIS IS NOT THE END

DATA SCIENCE IS HARD

**SEEK AND YE SHALL FIND
(HELP)**

LEARN BY DOING



WHAT IS DATA SCIENCE?

WHAT IS DATA SCIENCE?

A set of tools and techniques used to extract useful information from data

WHAT IS DATA SCIENCE?

A set of tools and techniques used to extract useful information from data

An interdisciplinary, problem-solving oriented subject

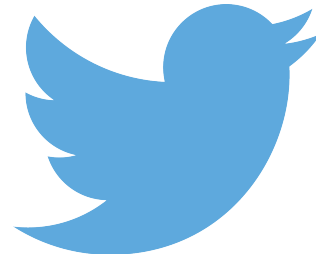
WHAT IS DATA SCIENCE?

A set of tools and techniques used to extract useful information from data

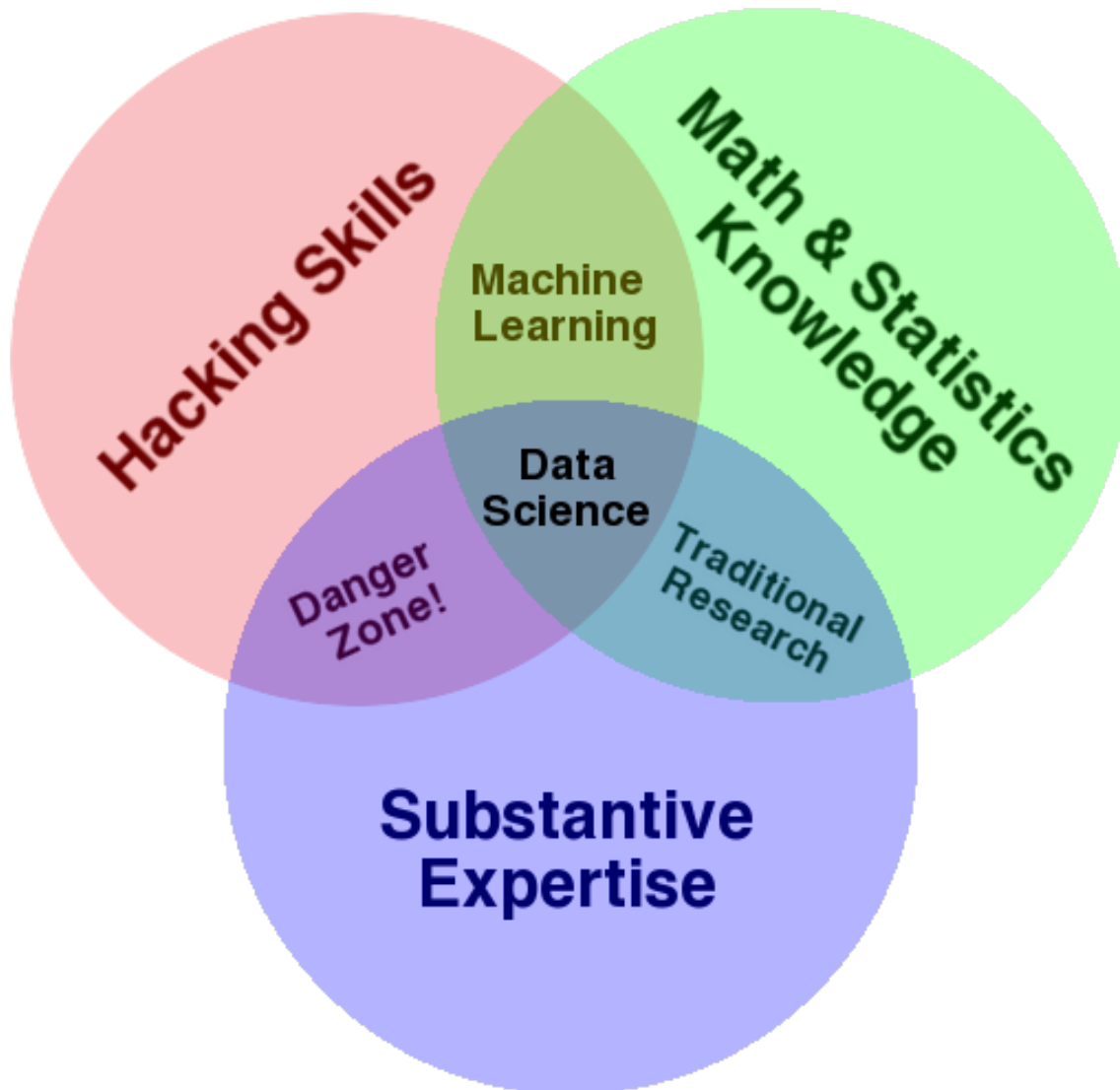
An interdisciplinary, problem-solving oriented subject

The application of statistical techniques to model practical problems

WHO USES DATA SCIENCE? TL;DR EVERYONE

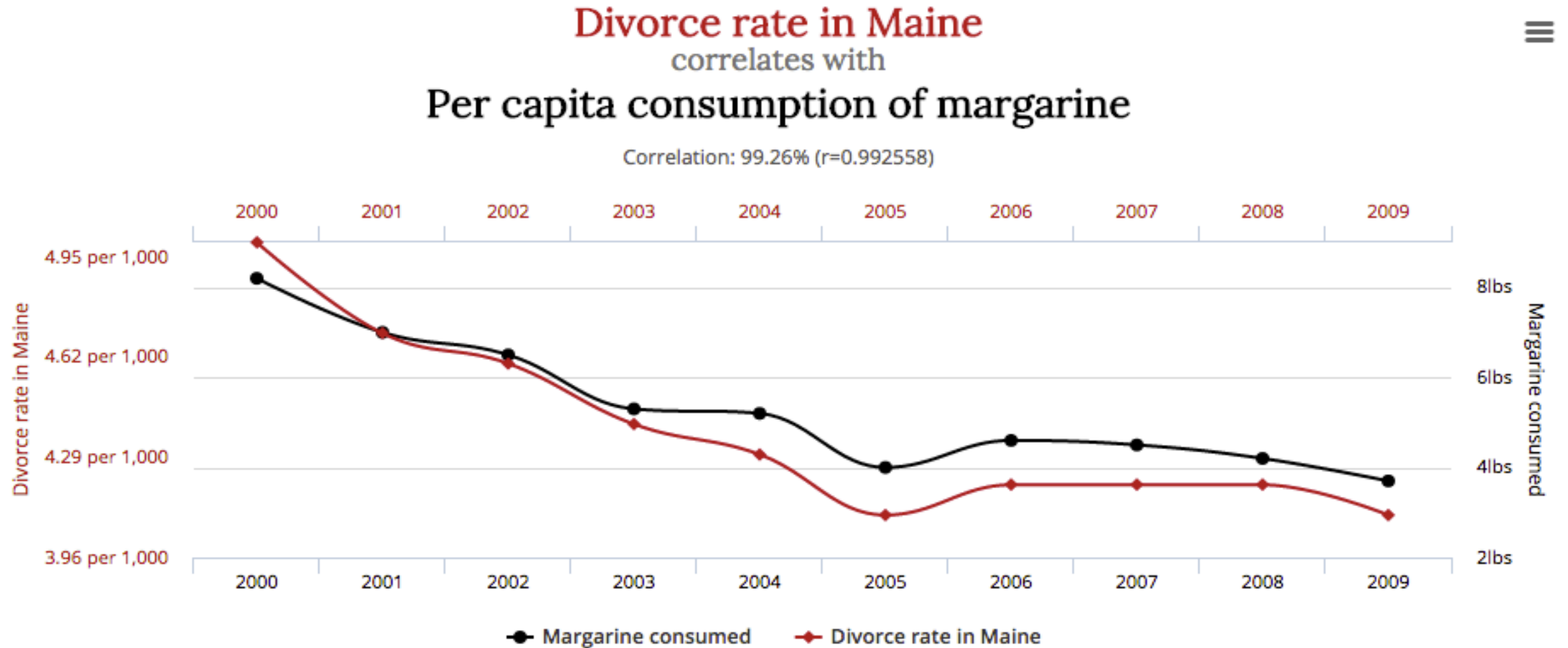
The Netflix logo, featuring the word "NETFLIX" in white, bold, sans-serif capital letters on a red rectangular background.The Google logo, consisting of the word "Google" in its multi-colored, rounded sans-serif font.The LinkedIn logo, featuring the word "Linked" in black and "in" in white inside a blue square, followed by a registered trademark symbol.The Facebook logo, featuring the word "facebook" in white, lowercase, sans-serif letters on a blue rectangular background.The Microsoft logo, consisting of four colored squares (red, green, blue, yellow) arranged in a 2x2 grid, followed by the word "Microsoft" in a grey sans-serif font.The CIVIS ANALYTICS logo, featuring the word "CIVIS" in large, bold, orange sans-serif capital letters, with the word "ANALYTICS" in smaller, bold, grey sans-serif capital letters below it.The Nest logo, featuring the word "nest" in a blue, lowercase, rounded sans-serif font, followed by a small trademark symbol.

WHAT QUALITIES MAKE UP A DATA SCIENTIST?



- Hacking skills
- Math and Stats knowledge
- Substantive expertise

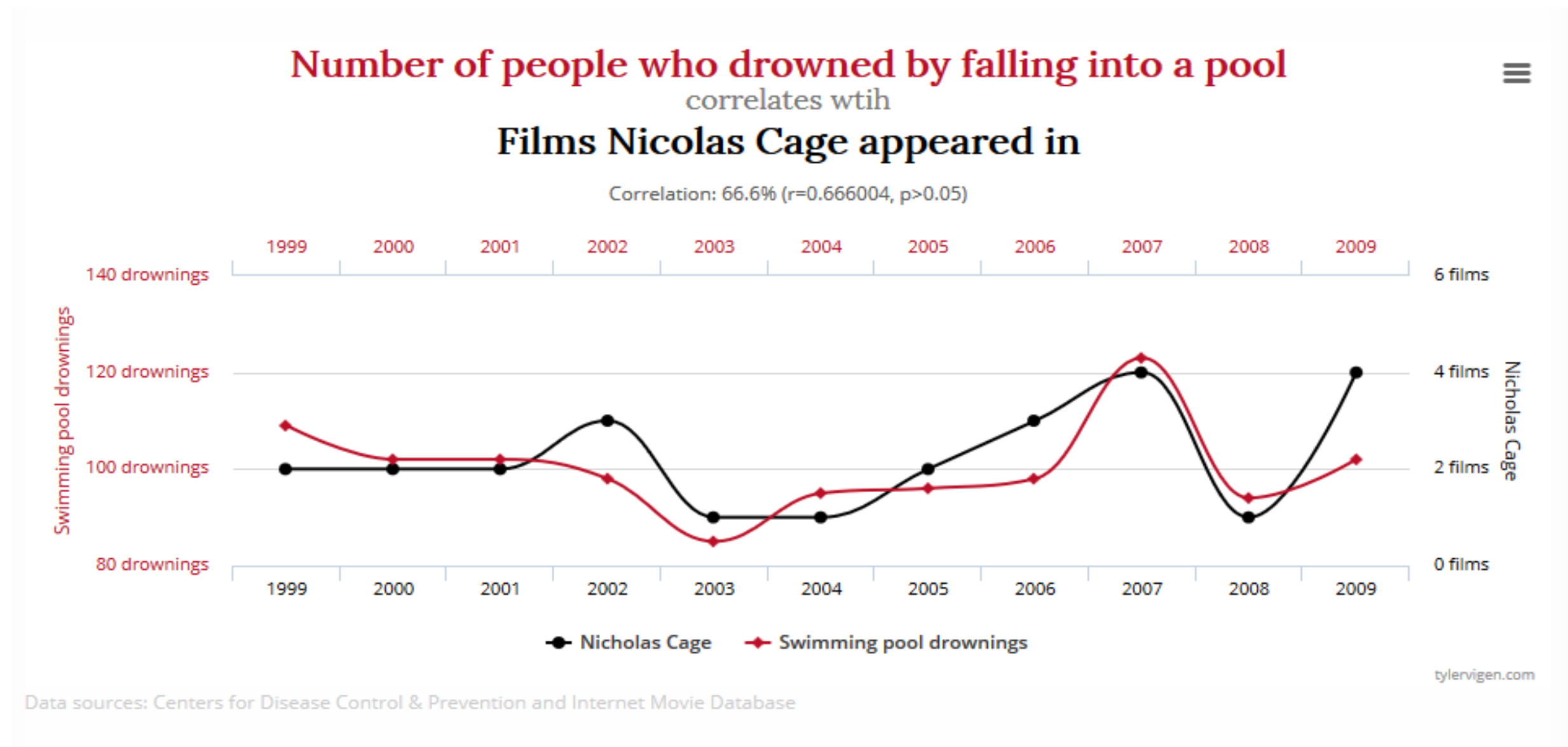
WHAT QUALITIES MAKE UP A DATA SCIENTIST?



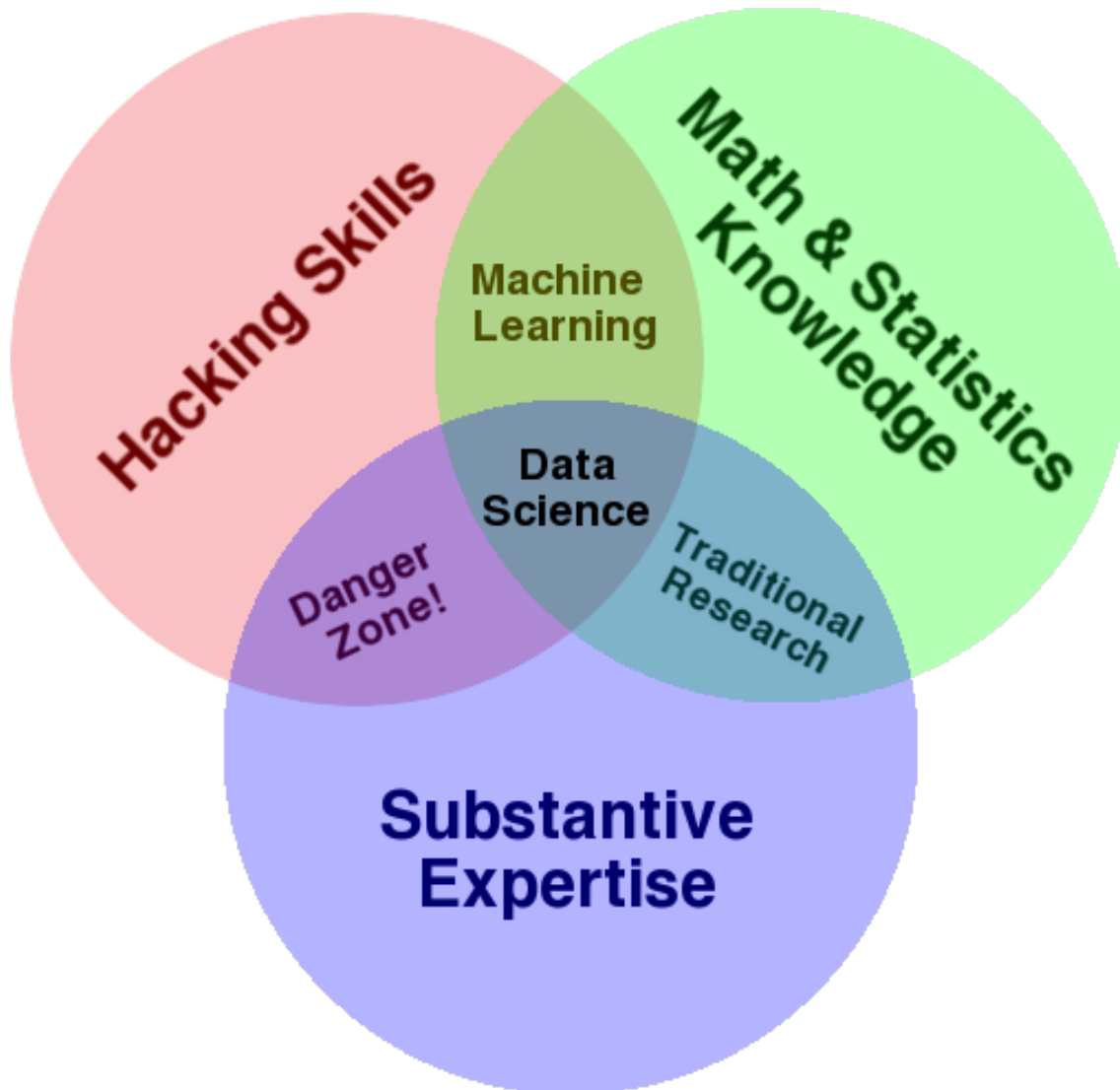
tylervigen.com

Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

WHAT QUALITIES MAKE UP A DATA SCIENTIST?

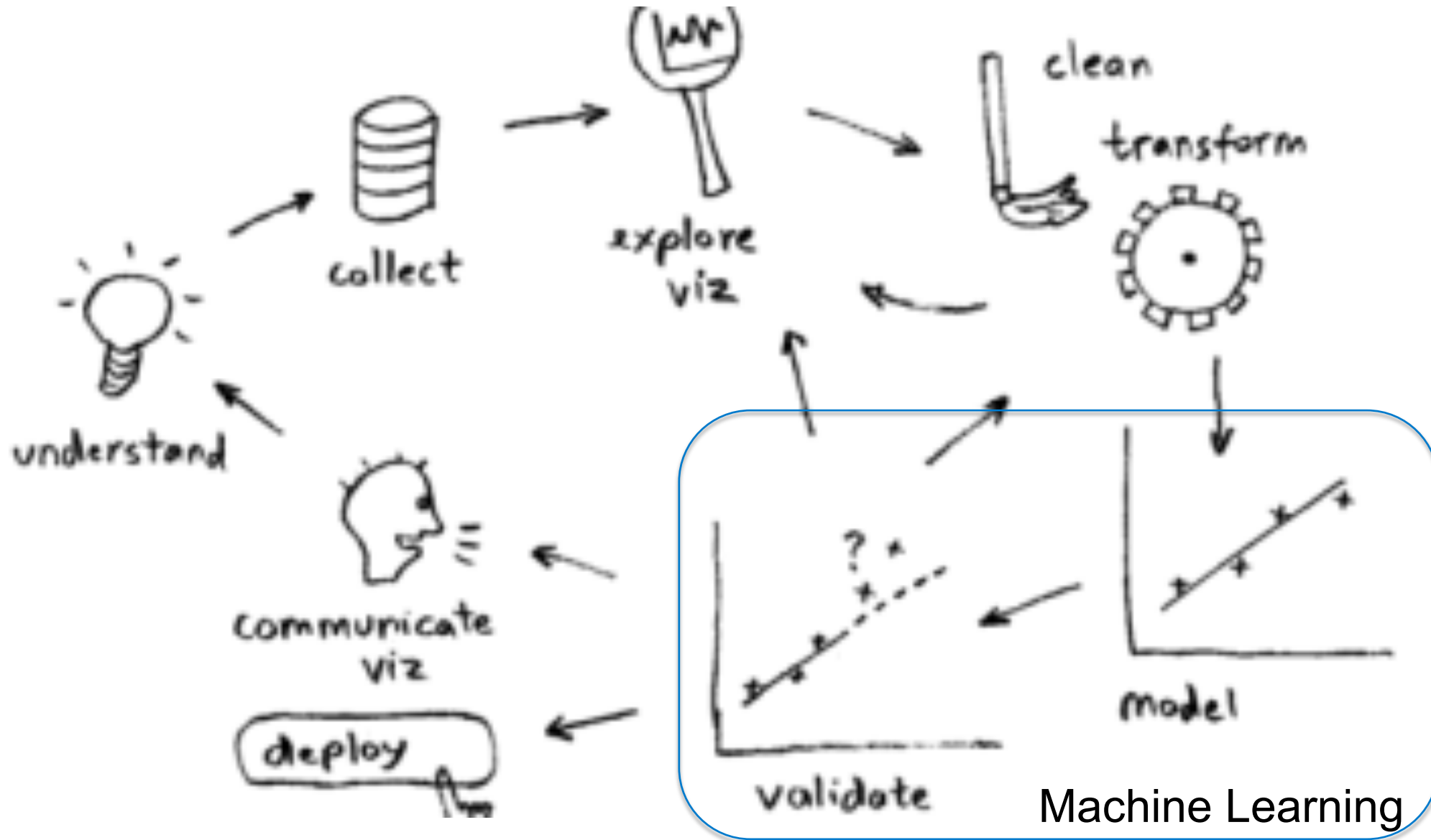


WHAT QUALITIES MAKE UP A DATA SCIENTIST?



- Hacking skills
 - Math and Stats knowledge
 - Substantive expertise
-
- Lastly.....Communication skills!

DATA SCIENCE WORKFLOW



DATA SCIENCE WORKFLOW EXAMPLE

PROBLEM: HOW MUCH SHOULD I CHARGE FOR A NEW CPU?

Understand: Can my previous CPU sales help predict future \$ sales? I would like to predict \$ Sales based on known quantities

DATA SCIENCE WORKFLOW EXAMPLE

PROBLEM: HOW MUCH SHOULD I CHARGE FOR A NEW CPU?

Understand: Can my previous CPU sales help predict future \$ sales? I would like to predict \$ Sales based on known quantities

Collect: What pieces of data might be important in my problem—requires expertise!
i.e. number of cores, clock speed, L1 and L2 cache sizes, number of competing chips, \$ sales

DATA SCIENCE WORKFLOW EXAMPLE

PROBLEM: HOW MUCH SHOULD I CHARGE FOR A NEW CPU?

Understand: Can my previous CPU sales help predict future \$ sales? I would like to predict \$ Sales based on known quantities

Collect: What pieces of data might be important in my problem—requires expertise!

i.e. number of cores, clock speed, L1 and L2 cache sizes, number of competing chips, \$ sales

Explore/Vis: Check the data. Are there frequently missing bits of information? Can it be used?

DATA SCIENCE WORKFLOW EXAMPLE

PROBLEM: HOW MUCH SHOULD I CHARGE FOR A NEW CPU?

Understand: Can my previous CPU sales help predict future \$ sales? I would like to predict \$ Sales based on known quantities

Collect: What pieces of data might be important in my problem—requires expertise!

i.e. number of cores, clock speed, L1 and L2 cache sizes, number of competing chips, \$ sales

Explore/Vis: Check the data. Are there frequently missing bits of information? Can it be used?

Clean/Transform: Maybe consumers don't want to pay 2x the price for 2x the clock speed. Maybe this is a logarithmic relationship? Solution: $\log(\text{CPU clock})$

DATA SCIENCE WORKFLOW EXAMPLE

PROBLEM: HOW MUCH SHOULD I CHARGE FOR A NEW CPU?

Understand: Can my previous CPU sales help predict future \$ sales? I would like to predict \$ Sales based on known quantities

Collect: What pieces of data might be important in my problem—requires expertise!

i.e. number of cores, clock speed, L1 and L2 cache sizes, number of competing chips, \$ sales

Explore/Vis: Check the data. Are there frequently missing bits of information? Can it be used?

Clean/Transform: Maybe consumers don't want to pay 2x the price for 2x the clock speed. Maybe this is a logarithmic relationship? Solution: $\log(\text{CPU clock})$

Model: You'll learn how to do this 😊

DATA SCIENCE WORKFLOW EXAMPLE

PROBLEM: HOW MUCH SHOULD I CHARGE FOR A NEW CPU?

Understand: Can my previous CPU sales help predict future \$ sales? I would like to predict \$ Sales based on known quantities

Collect: What pieces of data might be important in my problem—requires expertise!

i.e. number of cores, clock speed, L1 and L2 cache sizes, number of competing chips, \$ sales

Explore/Vis: Check the data. Are there frequently missing bits of information? Can it be used?

Clean/Transform: Maybe consumers don't want to pay 2x the price for 2x the clock speed. Maybe this is a logarithmic relationship? Solution: $\log(\text{CPU clock})$

Model: You'll learn how to do this 😊

Validate: Does this model really work? For instance, let's try predicting sales on other previous chips. Does the model accurately predict the sales of those chips? If not, go back to the drawing board

DATA SCIENCE WORKFLOW EXAMPLE

PROBLEM: HOW MUCH SHOULD I CHARGE FOR A NEW CPU?

Understand: Can my previous CPU sales help predict future \$ sales? I would like to predict \$ Sales based on known quantities

Collect: What pieces of data might be important in my problem—requires expertise!

i.e. number of cores, clock speed, L1 and L2 cache sizes, number of competing chips, \$ sales

Explore/Vis: Check the data. Are there frequently missing bits of information? Can it be used?

Clean/Transform: Maybe consumers don't want to pay 2x the price for 2x the clock speed. Maybe this is a logarithmic relationship? Solution: $\log(\text{CPU clock})$.

Model: You'll learn how to do this 😊




Validate: Does this model really work? For instance, let's try predicting sales on other previous chips. Does the model accurately predict the sales of those chips? If not, go back to the drawing board

Communicate: Great! So the \$Sales of a new CPU can be predicted based on a mixture of Gaussian variables based on logarithmic cpu clock speed, $10.45 * \# \text{ of cores}$, $(\# \text{Cores})^2$, and $\exp(\# \text{ of competing chips})$.

Now, how do you communicate this to a non-technical audience?

PROBLEM: HOW WOULD YOU IMPLEMENT “MORE ITEMS TO CONSIDER” ON AMAZON.COM?

Frequently Bought Together


+

+



Total price: **\$154.18**

[Add all three to Cart](#)

[Add all three to List](#)

i These items are shipped from and sold by different sellers. [Show details](#)


- ☒ **This Item:** Beats Solo2 Wired On-Ear Headphones - Black **\$139.10**
- ☒ **Matte Zipper Earphones Carrying Case for Beats Monster by Dr.Dre Studio, Solo Wireless, Solo, Solo...** **\$7.99**
- ☒ **Original Replacement Cable/Wire For Beats By Dre Headphones Solo/Studio/Pro/Detox/Wireless...** **\$7.09**



V-MODA Crossfade M-100
3D Custom Headphones
★★★★☆ (1161)
\$310.00 **\$199.99**

[Add feedback](#)

Customers Who Bought This Item Also Bought




Matte Zipper Earphones Carrying Case for Beats Monster by Dr.Dre Studio, Solo Wireless, Solo, Solo...

★★★★☆ 95

#1 Best Seller in Headphone Cases


\$7.99



Original Replacement Cable/Wire For Beats By Dre Headphones Solo/Studio/Pro/Detox...

★★★★☆ 653


\$7.09 *Prime*



Beats Solo HD On-Ear Headphone (Discontinued by Manufacturer - Black) wired


★★★★★ 1,064

\$127.99 *Prime*



Black Menba Matte Zipper Earphones Carrying Case/ Pouch/ Box for Beats & Monster


★★★★★ 2



Beats Solo HD On-Ear Headphone - Light Blue (Certified Refurbished)

★★★★☆ 76


\$139.99 *Prime*



Bluecell Protection Carrying Hard Case/Bag for Monster Dr Dre Beats Solo/Studio Headphone

★★★★☆ 419

\$4.35



Official Monster Beats By Dre 3.5mm in ear/earbuds Stereo Headset for HTC Red...

★★★★☆ 334

\$54.90



MACHINE LEARNING

WHAT IS MACHINE LEARNING?

from Wikipedia:

Machine learning explores the study and construction of algorithms that can *learn from* and make predictions on data.

WHAT IS MACHINE LEARNING?

“A computer program is said to learn from experience **E** with respect to some set of tasks **T** and performance measure **P**, if its performance at tasks **T**, as measured by **P**, improves with experience **E**.”



Tom Mitchell,
Professor CMU

WHAT IS MACHINE LEARNING?

“A computer program is said to learn from experience **E** with respect to some set of tasks **T** and performance measure **P**, if its performance at tasks **T**, as measured by **P**, improves with experience **E**.”

“A student is said to learn from the General Assembly **Data Science Course** with respect to some set of **homeworks** and measured by **grades**, if its performance at **homeworks** as measured by **grades**, improves throughout the **course**”

WHAT IS MACHINE LEARNING?

from Wikipedia:

Machine learning explores the study and construction of algorithms that can *learn from* and make predictions on data.

“The core of machine learning deals with **representation** and **generalization**...”

Representation – extracting a mathematical structure from data

Generalization – making predictions from data

TAXONOMY OF MACHINE LEARNING PROBLEMS

Supervised

Labeled examples - Making Predictions (**generalization**)

Unsupervised

No labeled examples - Discovering patterns (**representation**)

TAXONOMY OF MACHINE LEARNING PROBLEMS

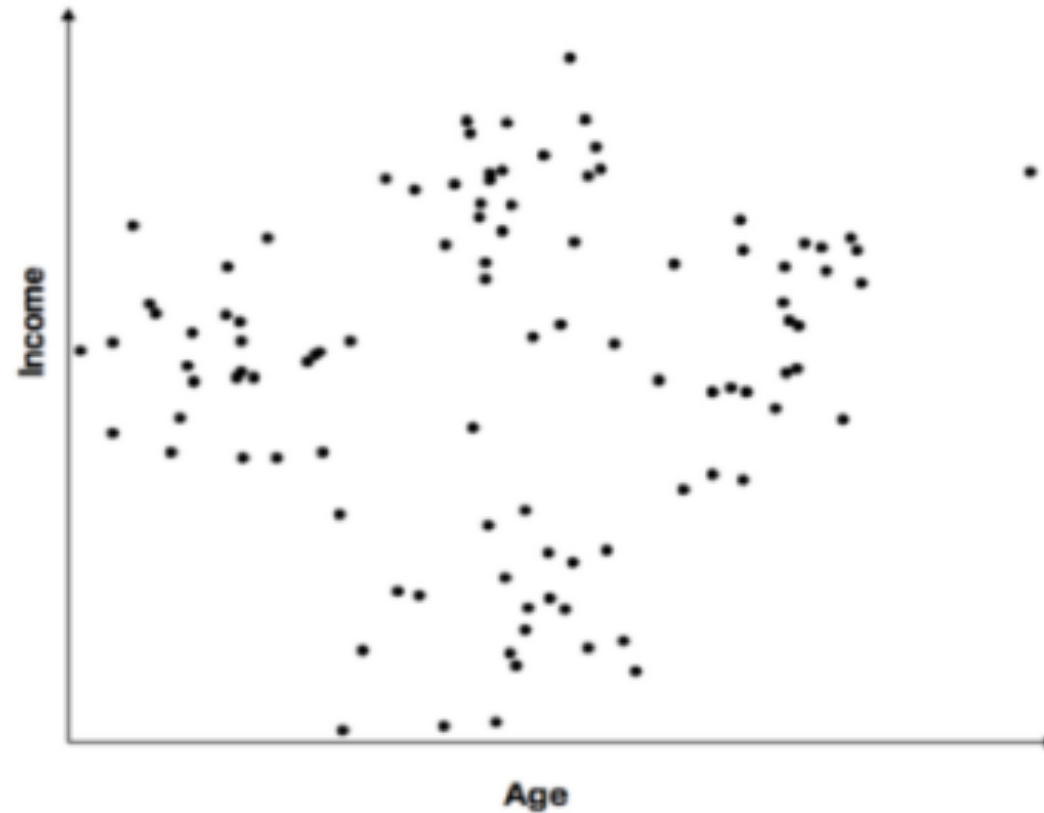
Supervised Example

Jim is 30 years old and can eat 4 donuts. Sally can eat 2 donuts and is 60 years old. Bobby is 15 years old. How many donuts can he probably eat?

TAXONOMY OF MACHINE LEARNING PROBLEMS

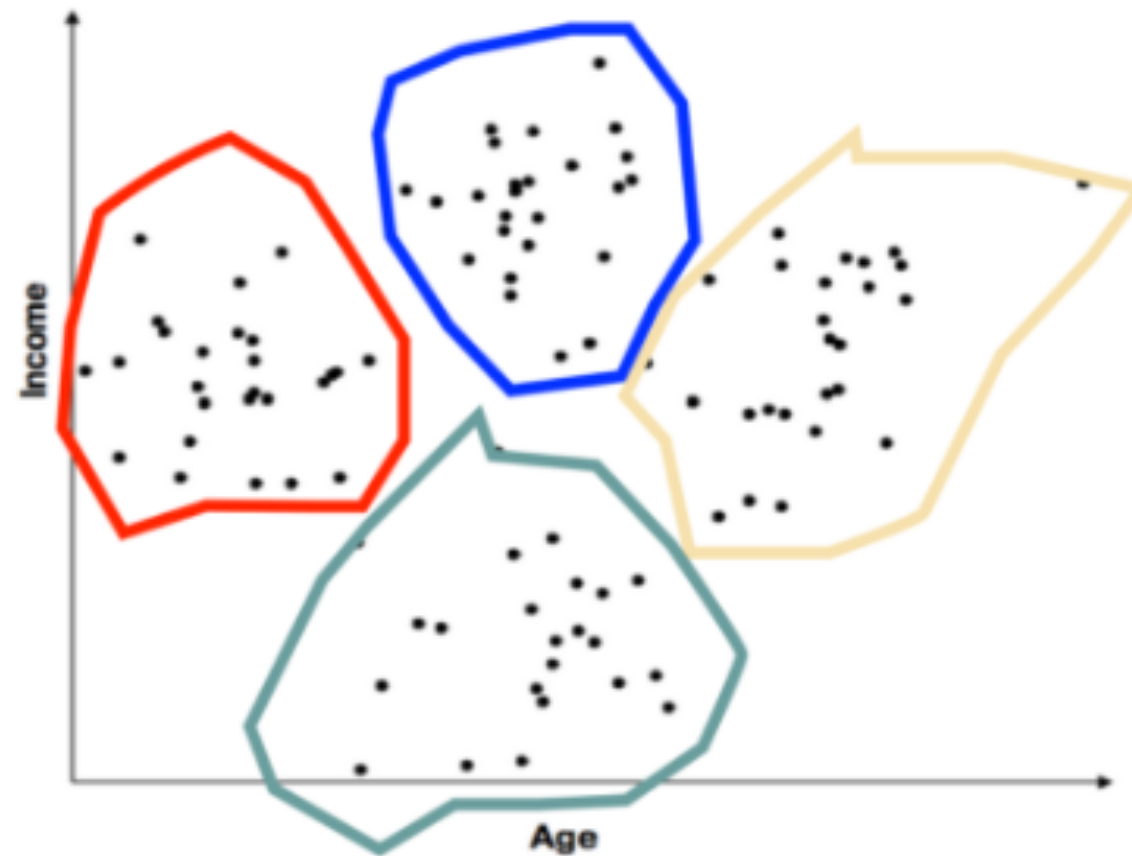
Unsupervised Example

Can we find structure to unlabeled data?



TAXONOMY OF MACHINE LEARNING PROBLEMS

Unsupervised Example



TAXONOMY OF MACHINE LEARNING PROBLEMS

Continuous

Categorical

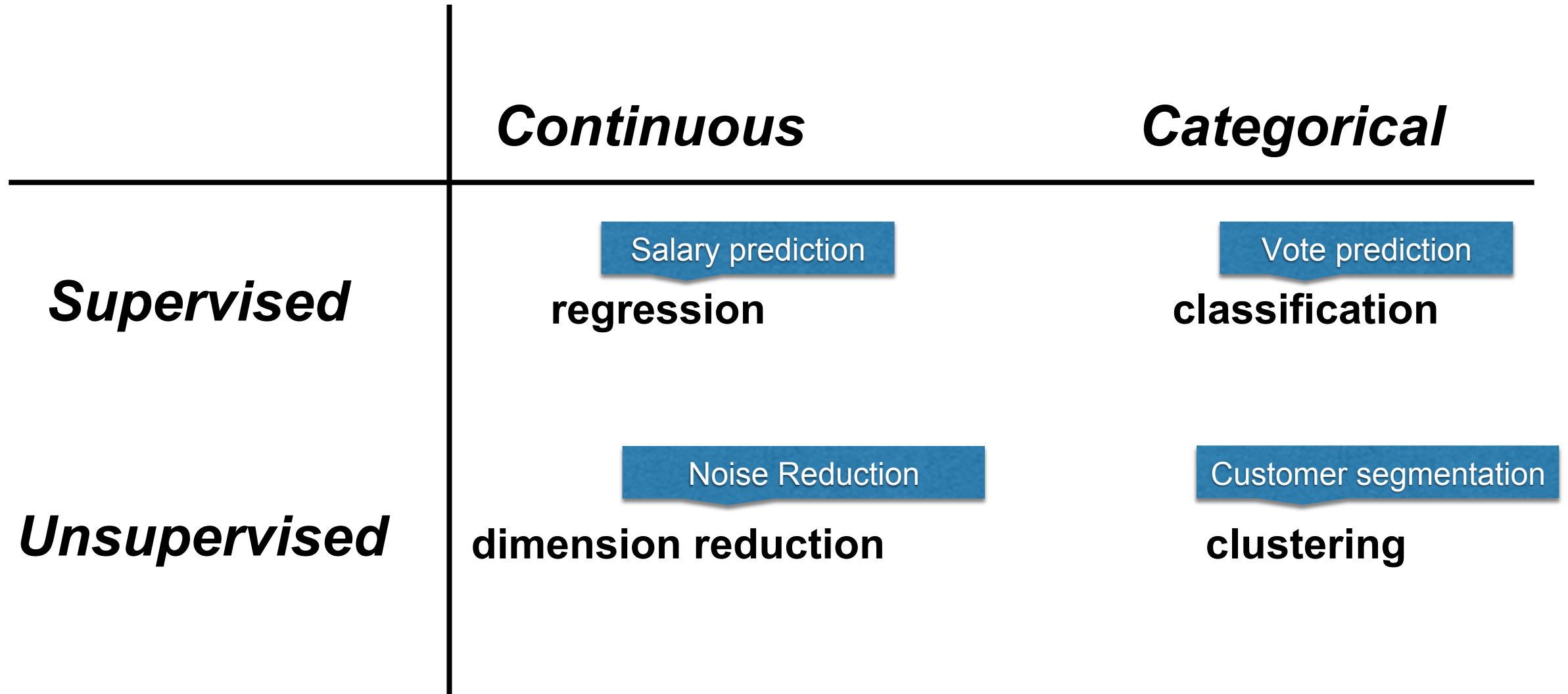
Quantitative
(ordered data, age,
Height, salary, etc.)

Qualitative
(sets, yes/no, vote, etc.)

TAXONOMY OF MACHINE LEARNING PROBLEMS

	<i>Continuous</i>	<i>Categorical</i>
<i>Supervised</i>	regression	classification
<i>Unsupervised</i>	dimension reduction	clustering

TAXONOMY OF MACHINE LEARNING PROBLEMS



SUPERVISED OR UNSUPERVISED?

You want to determine whether an email is spam or not

SUPERVISED OR UNSUPERVISED?

You want to group Amazon customers together based on their previous purchases, location, and number of visits to the website so you can advertise to them specifically

SUPERVISED OR UNSUPERVISED?

You want to predict the rating of a Netflix movie

SUPERVISED OR UNSUPERVISED EXERCISE

In a group, answer what kind of ML problems these can be classified as:

- Pandora Music Recommendation (i.e. What songs would you like)
- Digit recognition (i.e. post office performs digit recognition on mail)
- Predicting likelihood (i.e. probability) of a student passing high school
- You want to automatically reduce noise in your dataset
- You want to predict whether someone prefers Chevy or Ford based on their level of Car knowledge (1-10), age, and whether they like LS engines or Coyote engines

Homework 1 on Github – Due Dec 9 before class!

Python Survey

Exit tickets (This is Lesson 1)