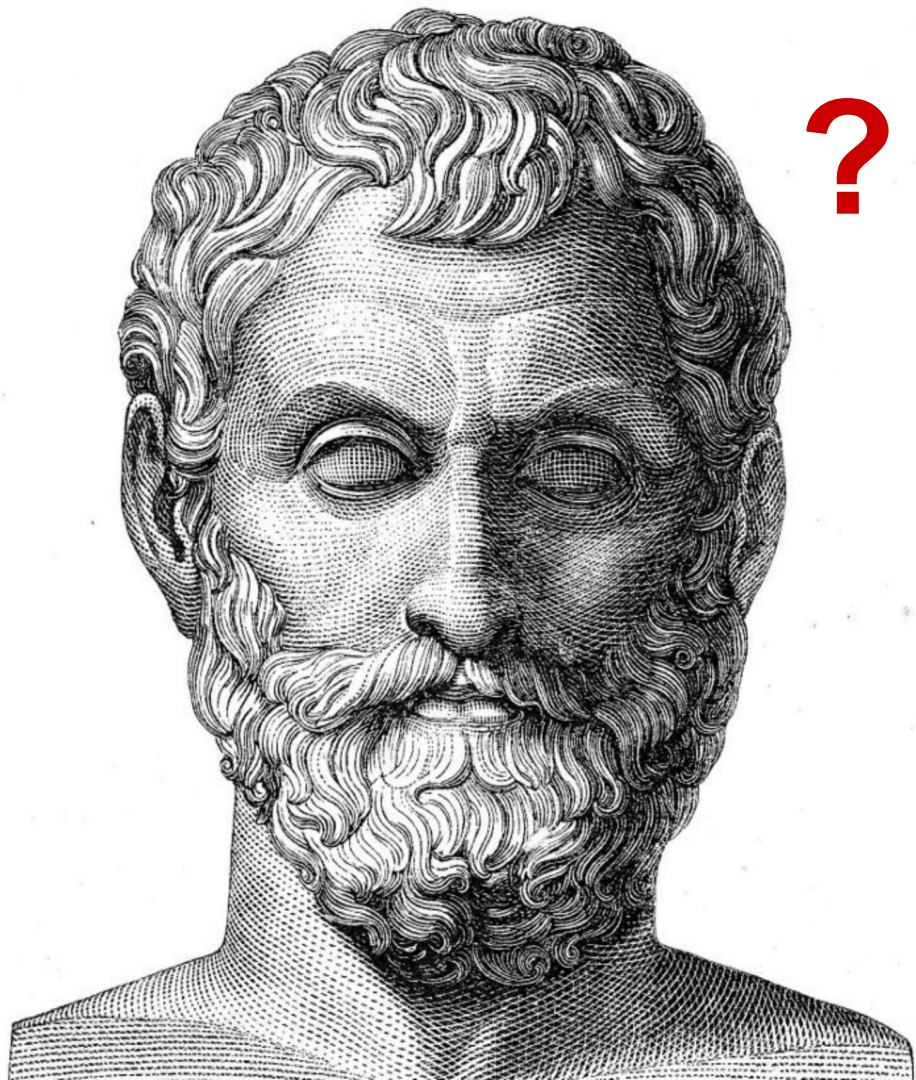


CDS Education

We explore, learn, and educate big minds.

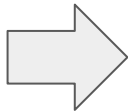
2017

624 BC.

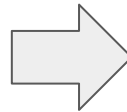




Olive Farm



Olive Press



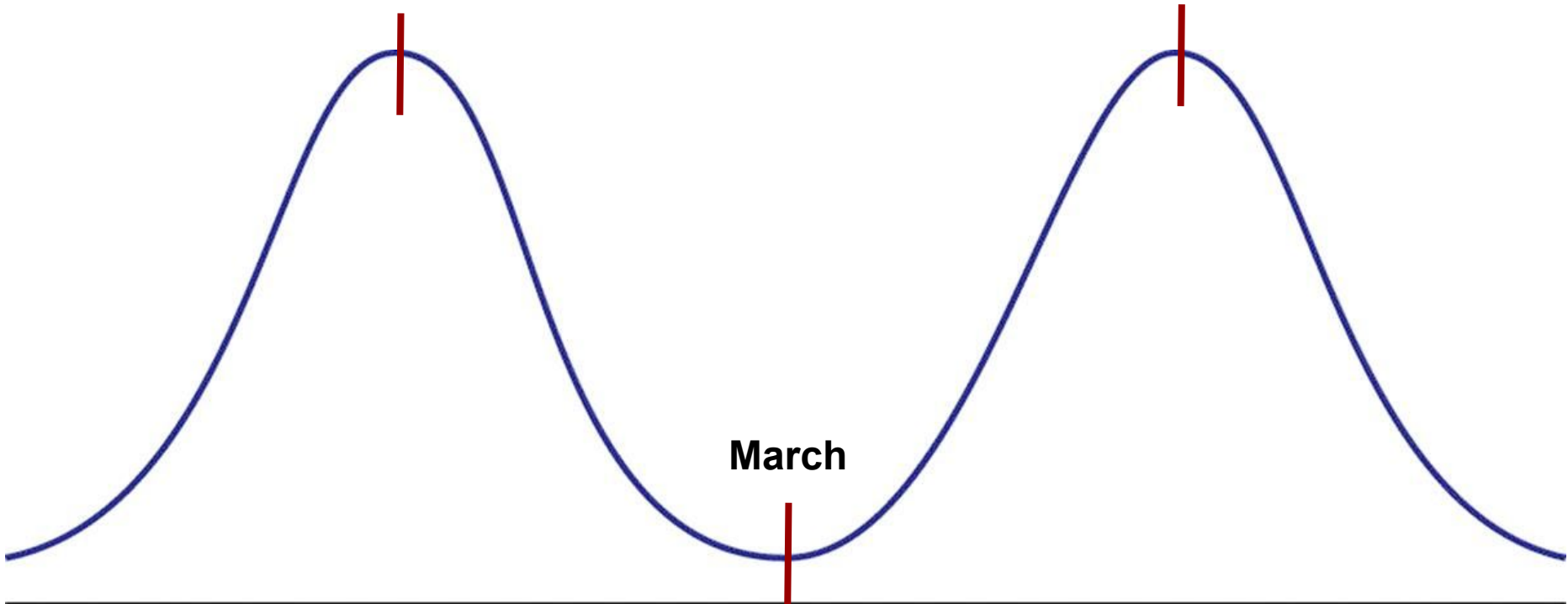
Storage

How to get rich?

September

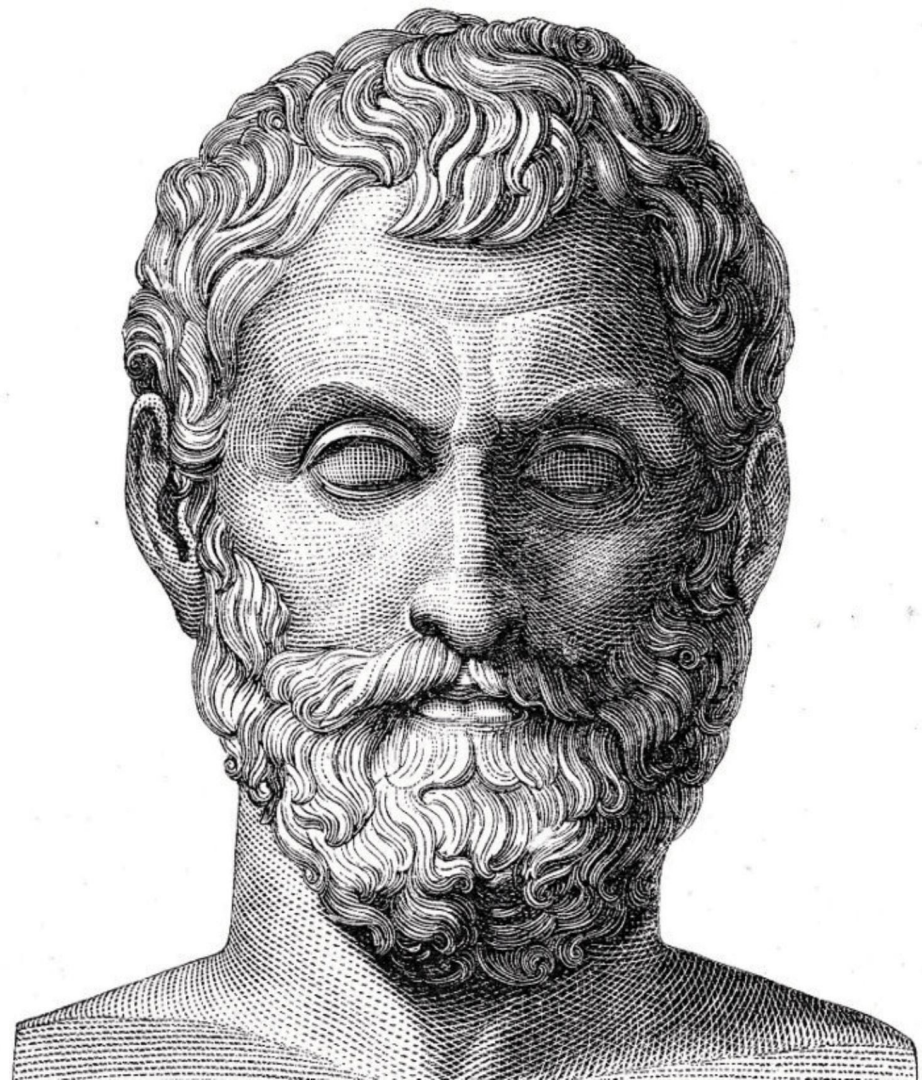
September

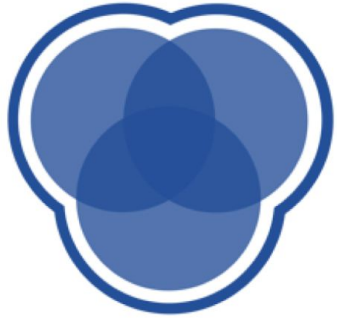
March



If I get all the oil press machines during March, I can buy them all with the minimum price but will be able to earn a lot of money back in September...

Too Obvious?!





CDS Education

We explore, learn, and educate big minds.

Data Science

Cornell Data Science

Cornell Data Science

Project Team

Student Organization

DL

DE

DV

Education

Business

Kaggle

Research

Algo

Courses

Events

ML

DL

DE

Career

Academics



History of Data Science and Machine Learning

- **1950, Alan Turing** creates “Turing Test” to determine if a computer has real intelligence by trying to fool a human that the program is human.
- **1952, Arthur Samuel** wrote first “Computer Learning Program” that played checkers and improved its strategy the more it played.
- **1967, The Nearest Neighbor Algorithm** was written, allowing computers to begin using pattern recognition.



- **1985, Terry Sejnowski** invents NetTalk, which learns how to pronounce words the same way a human baby does.
- **1990's, Machine Learning** shifts from knowledge based approach to a data driven approach. Computers can analyze large amounts of data and draw conclusions and learn from results.
- **1997, IBM's Deep Blue** beats the world champion at chess.
- **2006, Geoffrey Hilton** coins the term Deep Learning to explain new algorithms that let computers “see” and distinguish objects and text in images.



- **2009, Hal Varian - Google Chief Economist**

“The sexy job in the next 10 years will be statisticians. The ability to take data, understand it, process it, extract value from it, visualize it, and communicate it. That’s going to be a hugely important skill in the next decades.”

- **2011, IBM Watson** beats human competitors in Jeopardy.
- **2016, Google AI** called AlphaGo beats professional players at Go, which is considered by many to be the most complicated board game that needs the most “human strategy”.



Instructor[0]

Jared Junyoung Lim

Education Lead, CDS

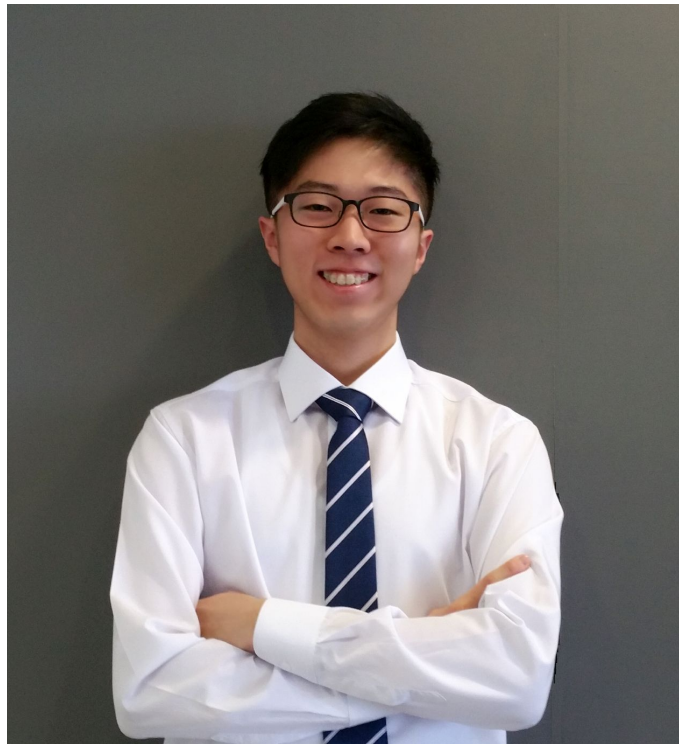
Instructor, INFO 1998

Computer Science '20

Fun Facts:

- 1) **No fun fact**
- 2) Does **not** tolerate **fun** and **facts**
- 3) There will be **no fun** in this class
- 4) #3 is a **fact**

jl3248@cornell.edu



Instructor[1]

Abby Beeler

Education Associate, CDS

Computer Science '20

Biometry & Statistics Minor

arb379@cornell.edu



Course Staffs

Piazza Team

Abby Beeler

Jared Lim

Shubhom Bhattacharya

Office Hour Team

Ann Zhang

Ethan Cohen

Ryan Kannanaikal

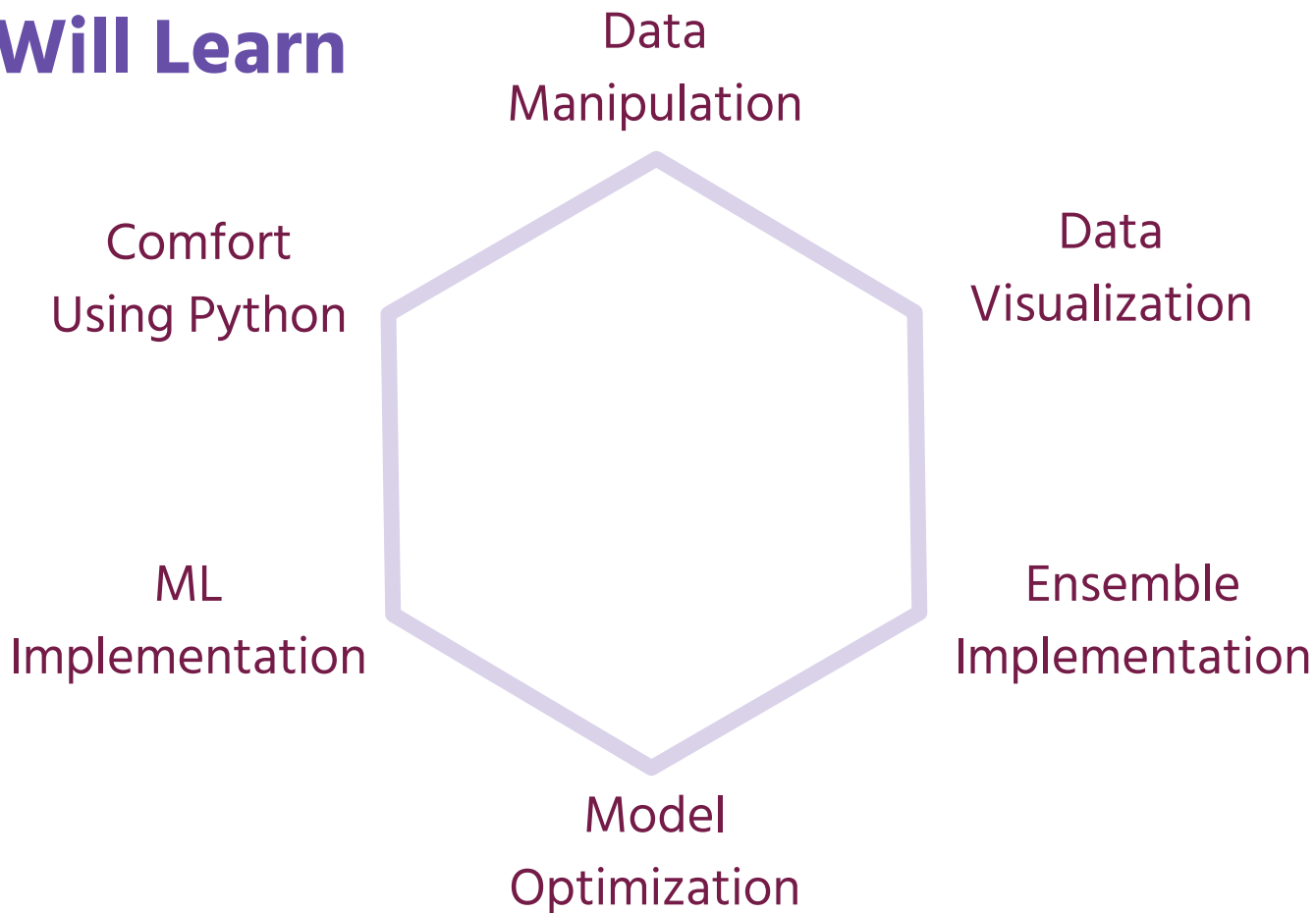


What Is This Class?

- Focus on application
- Data scientist starter pack
- Learning to speak data science
- Understanding those buzzwords
- A gateway to becoming a CDS member



What You Will Learn



Course Logistics

9-Week Course

Leaf 1: **Data Analysis** (1-2)

Leaf 2: **Machine Learning** (3-9)

One Big Project

Divided into **5 parts**

+ Mini **quiz** for **lecture 1**



Form a
GROUP of
3-4 people
ASAP

Course Logistics

Grading

10% Take-home Quiz

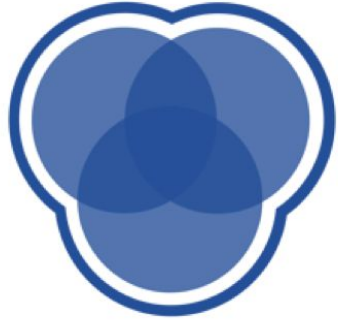
16% Each of Project part A, B, C, D

26% Project part E

70%

Every Assignment due **Tuesday at
Midnight**





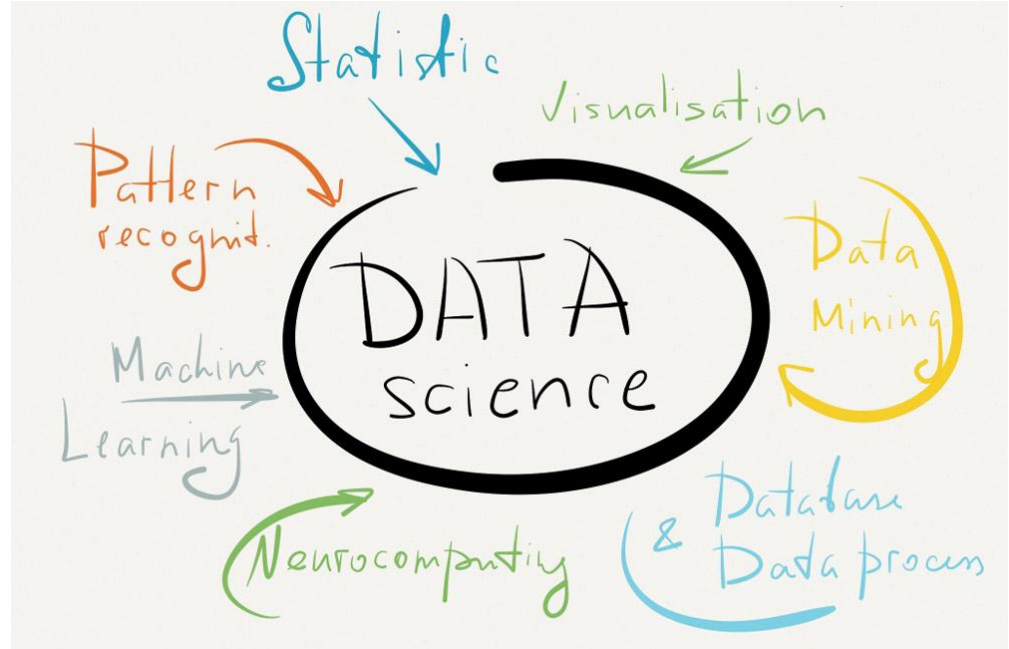
CDS Education

Introduction to Machine Learning for Python

Introduction and Data Manipulation

What is Data Science?

- Empirical Research
- Predictive Analytics
- Preventive Analytics
- Real-time Analysis
- Automation



Data can be...

LARGE

fast

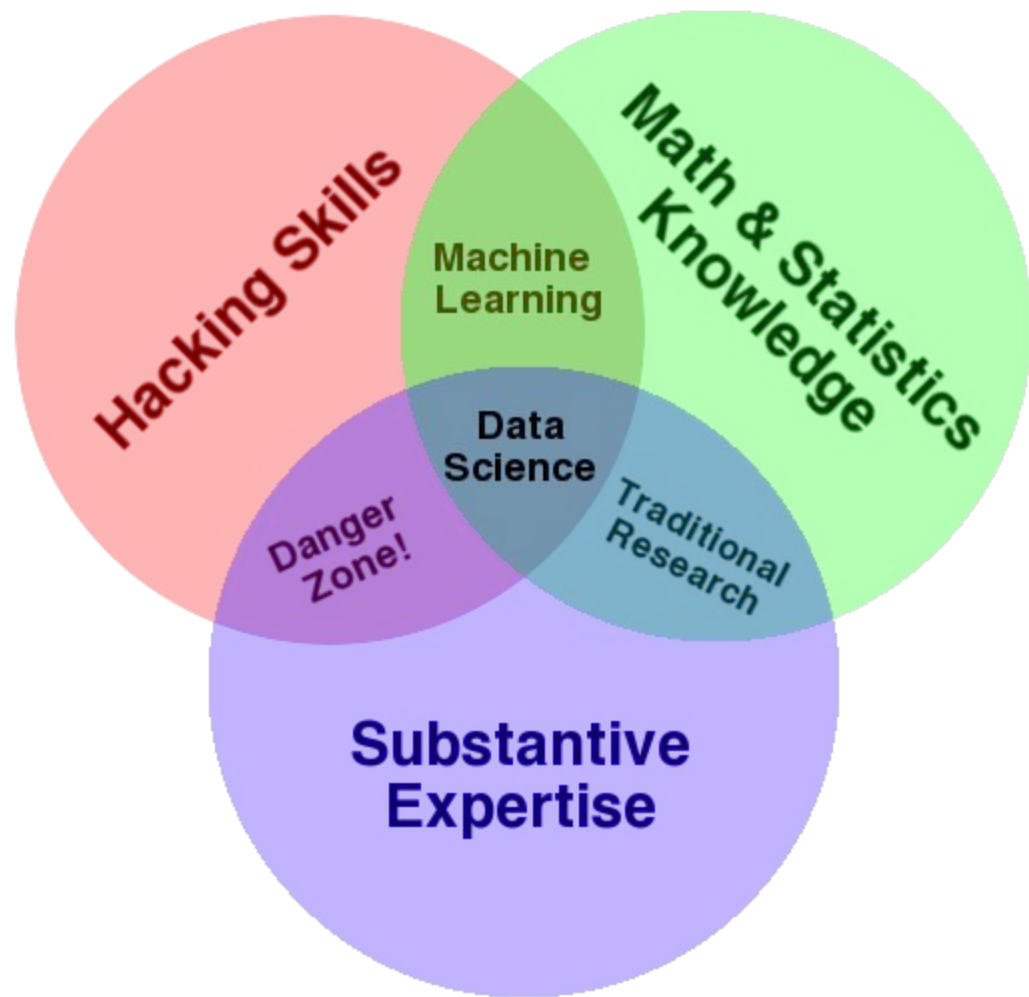
unStRUcTUReD

Volume

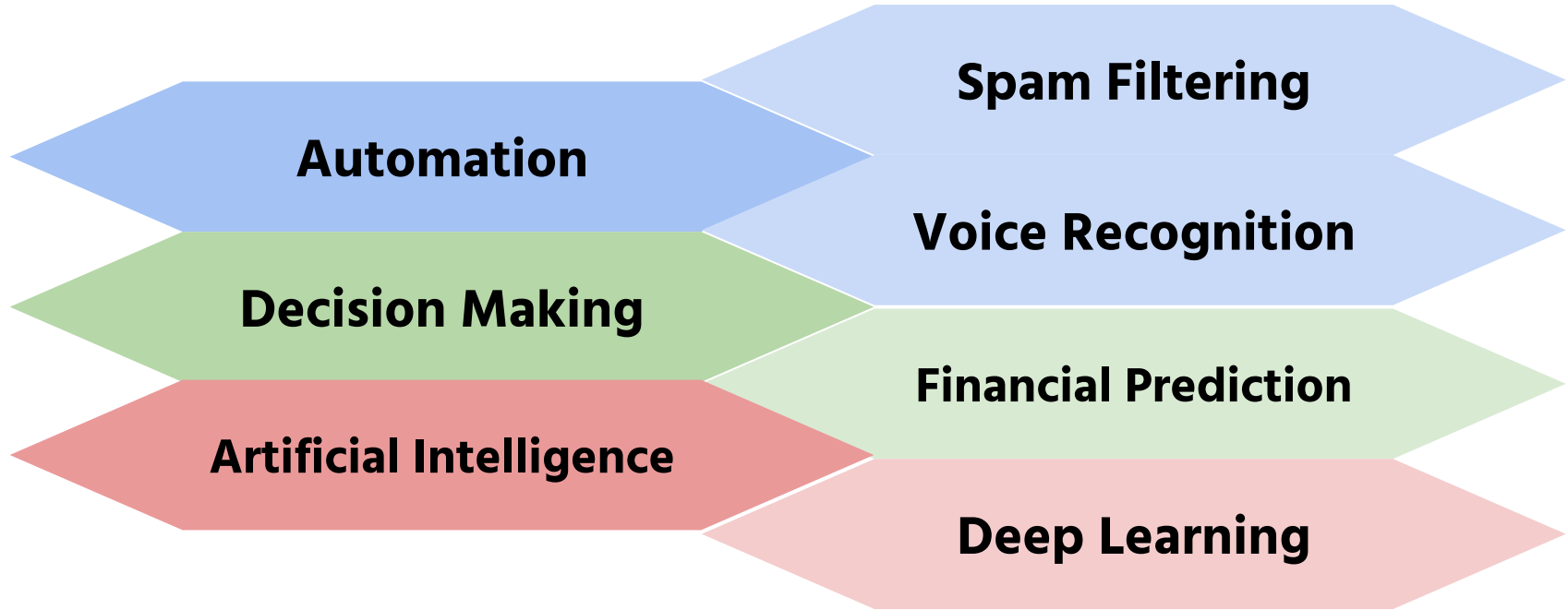
Velocity

Variety

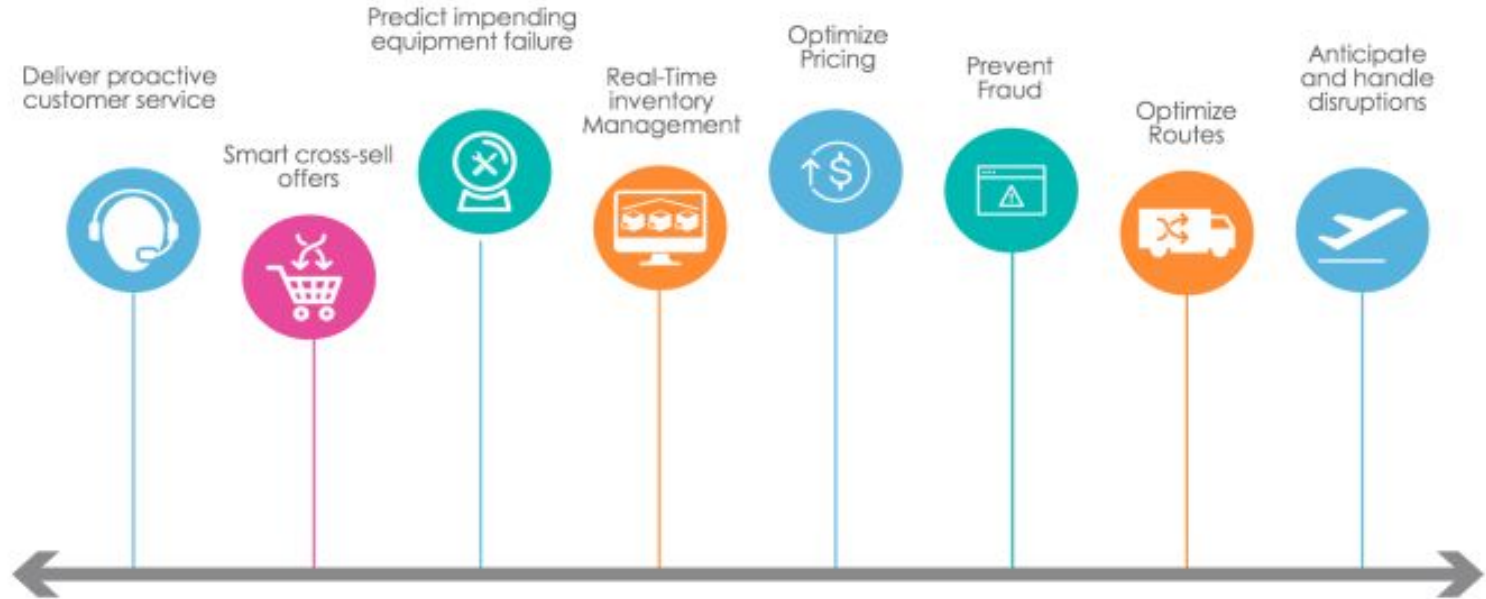




Applications



Applications



Why Jupyter Notebooks?

- Document the process
 - Code
 - Visuals
- Intuitive
 - Supports Python, R, Julia, etc.
- Easy to share



Lecture 2: Data Transformation

Now that we've picked up some basic tools for doing data science, we're ready to sharpen our data handling skills. As you might have already observed, data rarely comes in a neatly packaged "ready-to-use" format. We need to be able to manipulate datasets and shape them as we please so that we can run machine learning algorithms on them. Let's start with getting a little bit more comfortable with R.

Type *Markdown* and LaTeX: α^2

Writing Fast R

R is an excellent language for data science. However, R behaves very differently from commonly used object-oriented languages like Java and Python. Such differences can cause huge inefficiencies to unsuspecting beginners of R. Let's take a look at one of the most misunderstood concepts in R: the inefficiency of using explicit for-loops, as indicated below.

```
In [15]: # Process time comparison of explicit for-loop with implicit loops.
vec <- c(1:1000000)
```

```
# explicit version
system.time({for(i in 1:1000000) {
  vec[i] <- vec[i] * 2
}})
```

```
# implicit version
system.time({vec <- vec * 2})
```

```
user system elapsed
0.872 0.003 0.876
```

```
user system elapsed
0.003 0.000 0.003
```



Language Wars



Why Python?

Easy to learn and **readable**.

Extendable and **compatible**.

Open source with a large
community.



Python Packages Overview

Python

NumPy

Pandas

Matplotlib

SciPy

scikit-learn

statsmodel



NumPy Overview

NumPy

Arrays Improve
Speed

Vectorization

Built-in
Functions



\$\$ Golden Rules of Vectorization \$\$

Whatever you're trying to do, there's probably a NumPy function

Replace explicit Python loops with whole array NumPy operations



Array Operations

Operations



```
>> a + b  # same as np.add(a, b)
```



```
>> a - b  # same as np.subtract(a, b)
```



```
>> a * b  # same as np.multiply(a, b)
```



```
>> np.sqrt(a)
```

...

And more!



Data Frames

- Pandas offers **DataFrame** objects to help manage data in an orderly way
- Similar to Excel spreadsheet or SQL table
- Each column is one feature variable
- Each row is one sample or observation
- DataFrames facilitate selection and manipulation of data



Data Frame Example

A table of data

- Student, Sat Score, # Extracurriculars, etc.
- House Price, # Cars, # Rooms, etc.

```
In [6]: df
```

```
Out[6]:
```

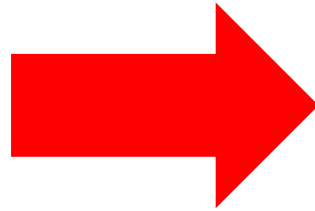
	city	humidity	maxtempi	meantempi	mintempi
2015/08/25	London, United Kingdom	85	66	59	52
2015/08/26	London, United Kingdom	85	67	63	59
2015/08/27	London, United Kingdom	79	67	62	56
2015/08/28	London, United Kingdom	70	70	60	51
2015/08/29	London, United Kingdom	77	72	64	57
2015/08/30	London, United Kingdom	81	69	64	60
2015/08/25	Birmingham, United Kingdom	87	62	56	51
2015/08/26	Birmingham, United Kingdom	78	69	63	57
2015/08/27	Birmingham, United Kingdom	78	64	58	51
2015/08/28	Birmingham, United Kingdom	76	66	57	48
2015/08/29	Birmingham, United Kingdom	69	69	60	51
2015/08/30	Birmingham, United Kingdom	81	64	60	55
2015/08/25	Lyon, France	45	76	66	55



Data Manipulation



Drunken Datasets Out There

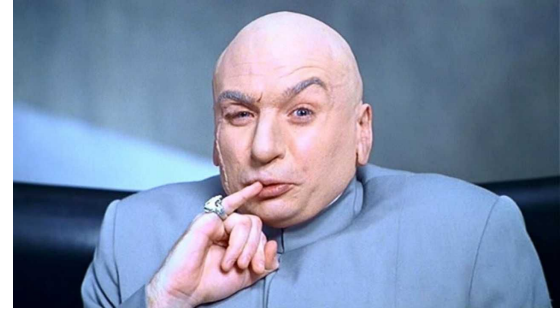


Question:

What are some ways in which data can be “messy”?



Why Do We Manipulate



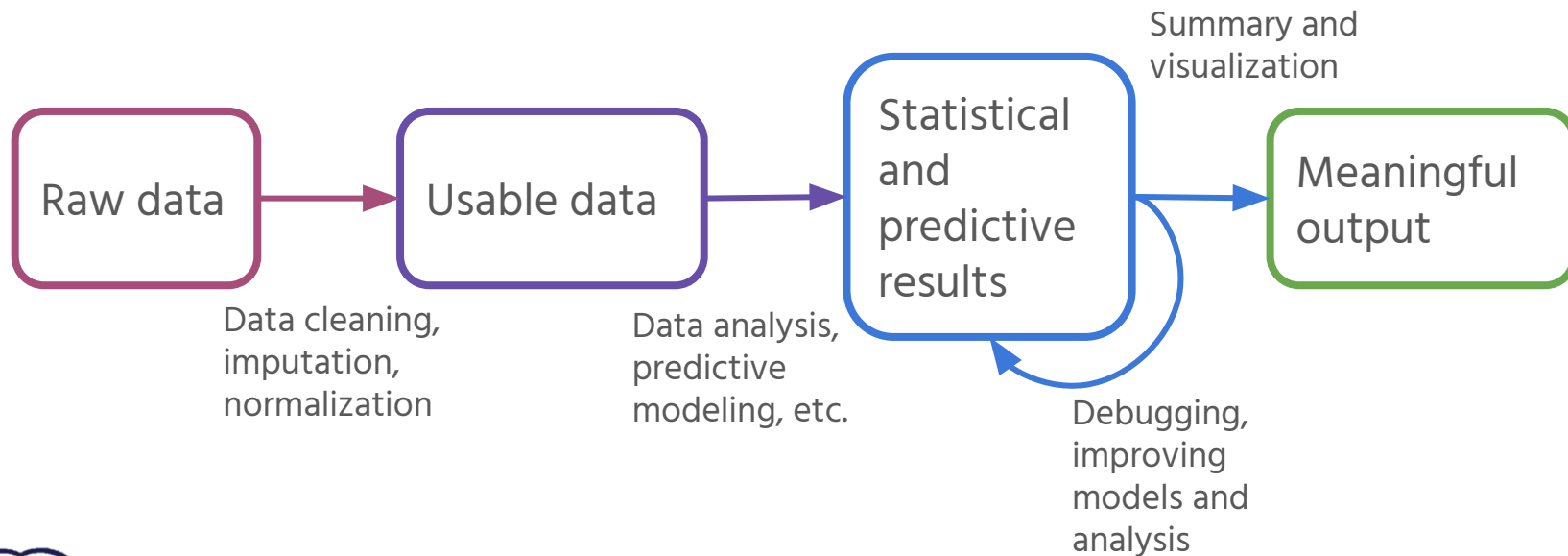
Increase clarity
and usability

Prevent
calculation errors

Improve memory
efficiency



The Data Pipeline



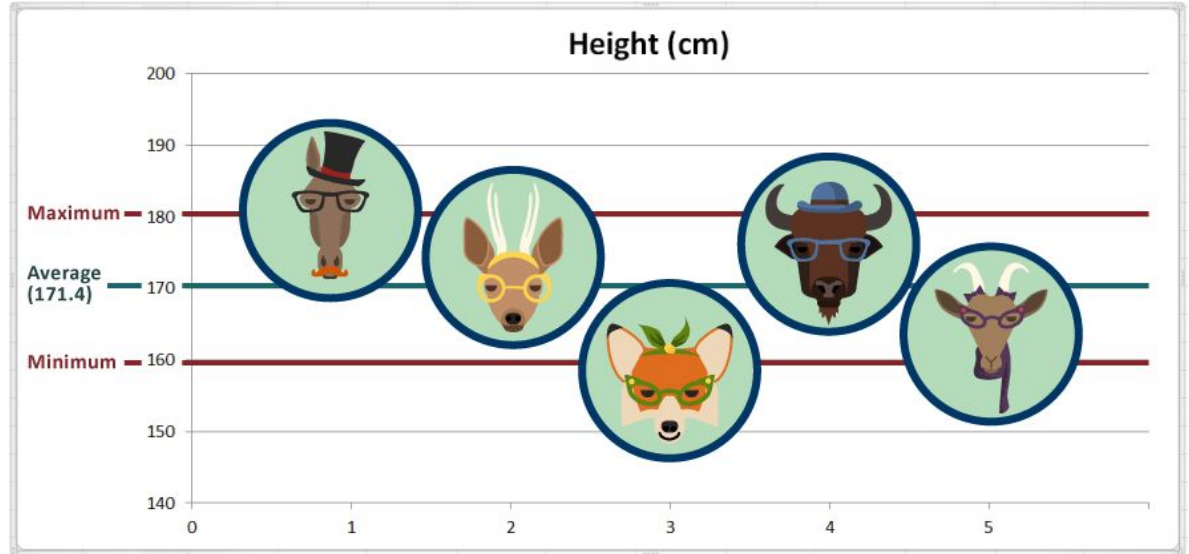
Summarizing

What it does

Gives a general overview of the dataset

Why?

To understand and explore the dataset!



Statistical Methods

mean()

```
>> an_array.mean(axis=1) # computes means for each row
```

median()

```
>> an_array.median()
```

sum()

```
>> an_array.sum(axis=0) # computes sum of each column
```



Filtering and Subsetting

What it does

Grab a subset in a data frame with a condition. **Filtering** grabs rows and **subsetting** grabs columns.

Why?

Decreasing data size or examining subgroups closer

Name	Age	Major
Amit	19	Computer Science
Dae Won	24	ORIE
Chase	19	Information Science
Jared	19	Computer Science

Filtering

Name	Age	Major
Amit	19	Computer Science
Dae Won	24	ORIE
Chase	19	Information Science
Jared	19	Computer Science

Subsetting



Combining

What it does

Joins together two data frames, either row-wise (horizontally) or column-wise (vertically)

concat!

Name	Age	Major
Amit	19	Computer Science
Dae Won	24	ORIE

Name	Age	Major
Jared	19	Computer Science
Kenta	20	Computer Science



Name	Age	Major
Amit	19	Computer Science
Dae Won	24	ORIE
Jared	19	Computer Science
Kenta	20	Computer Science



Combining (continued)

	Name
0	Amit
1	Dae Won
2	Chase
3	Jared
4	Kenta

	Age	Major
0	19	Computer Science
1	24	ORIE
2	19	Information Science



	Name	Age	Major
0	Amit	19	Computer Science
1	Dae Won	24	ORIE
2	Chase	19	Information Science
3	Jared	NaN	NaN
4	Kenta	NaN	NaN



Joining

What it does

Joins together two data frames, combining rows that have the same value for a column

How to do it

Pandas has **join** and **merge** functions. When we use **merge**, we want to set a column to *key on*, using *on=('key_name')*



But why would we get a dataset in pieces?

Name	Major	Age	Computer	Purchased
Dae Won	ORIE	31	Linux	Nvidia Titan X
Dae Won	ORIE	31	Linux	Nvidia Titan X
Dae Won	ORIE	31	Linux	CRT Monitor
Dae Won	ORIE	31	Linux	48GB RAM
Jared	CS	19	Mac	Big Book of Trivia
Jared	CS	19	Mac	“Help I don’t know fun facts” - A Life Story
Jared	CS	19	Mac	“10,000 Facts to Impress Your Friends”
Dae Two	ORIE	31	Linux	Friends



This is wasteful...

But why would we get a dataset in pieces?

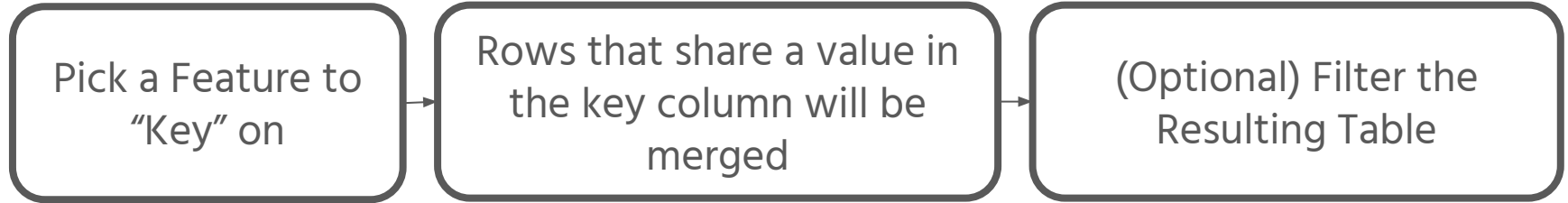
ID	Name	Major	Age	Computer
0001	Dae Won	ORIE	31	Linux
0002	Jared	CS	19	Mac

There's a lot less redundant data!

ID	Purchased
0001	Nvidia Titan X
0001	Nvidia Titan X
0001	CRT Monitor
0001	48GB RAM
0002	Big Book of Trivia
0002	"I don't know fun facts - My Life Story"
0002	"10,000 Facts to Impress Your Friends"
0001	Friends



A Join in Action



ID	Name	Major	Age	Computer	Purchased
0001	Jared	CS	19	Mac	Big Book of Trivia
0001	Jared	CS	19	Mac	"I don't know fun facts - My Life Story"
0001	Jared	CS	19	Mac	"10,000 Facts to Impress Your Friends"



Coming Up

Your assignment: Jupyter Setup & Take-home Quiz (released tonight)

Due: February 25th (Sunday) at Midnight

Submit Through: CMS

Next week: LECTURE 2 - Data Manipulation and Visualization

