CSCI 1070: Taming Big Data

Dr. Abby Stylianou

Logistics

- Are you on the waitlist? Talk to me after class!

- Class meetings: Tu/Th, 11AM–12:15PM, Ritter 117
  - Lecture, discussions, in class coding
  - Occasional quizzes about readings
  - Attendance not mandatory, but you're gonna have a bad time if you don't show up

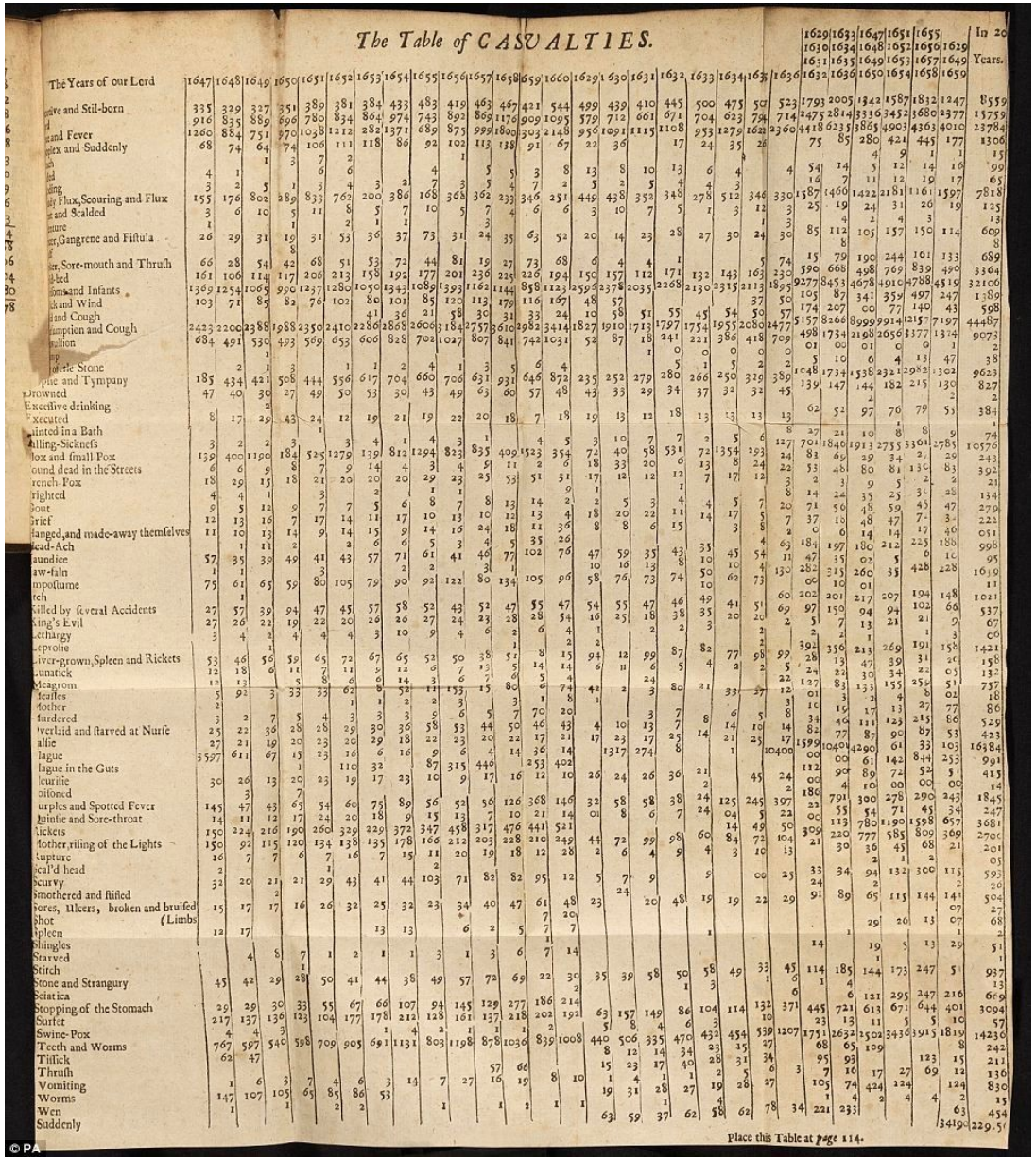
- Office Hours: Tues, 2:30-3:30PM, Ritter 107

Logistics

- Website/schedule/materials: https://cs.slu.edu/~stylianou/1070_sp2020/
- Blackboard for turning in assignments and seeing grades
- Technologies: Python (w/ data science + scientific libraries), Jupyter Notebooks
- Class communications: Piazza! *(not email)*

What is big data?

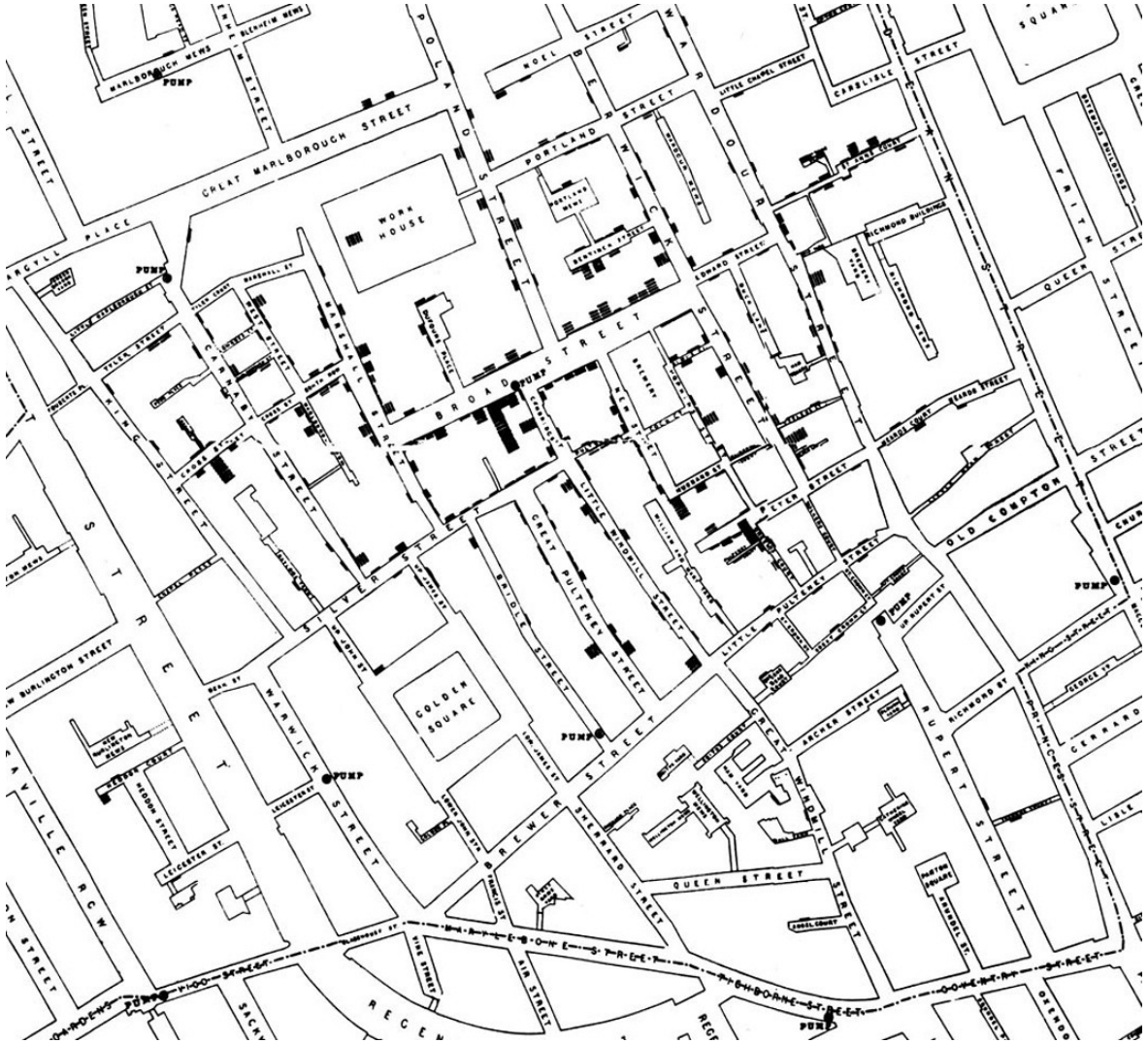
- **1663**: John Graunt is the first person credited w/ statistical data analysis in his studies of the bubonic plague in Europe, dealing with what he referred to as “overwhelming amounts of information”



https://www.dataversity.net/brief-history-big-data/

What is big data?

- **1854**: John Snow maps London Cholera outbreaks and finds that they are clustered around a single pump; it was found that a cesspit was leaking into that pump



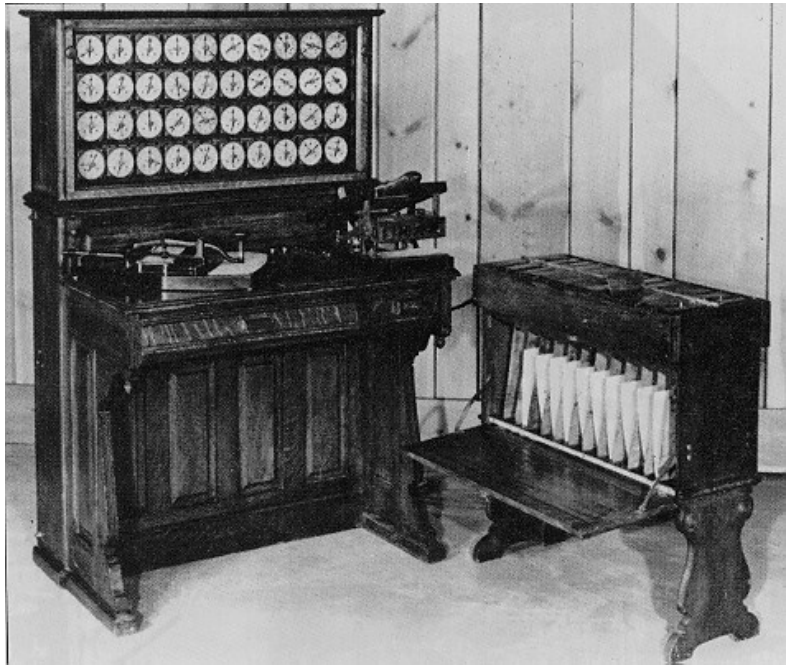
https://www.inferentialthinking.com/chapters/02/1/observation-and-visualization-john-snow-and-the-broad-street-pump

What is big data?

- **1880**: US Census Bureau estimates it will take eight years to process the data collected in the 1880 census, and over 10 years to process 1890 census data

- **Hollerith Tabulating Machine**

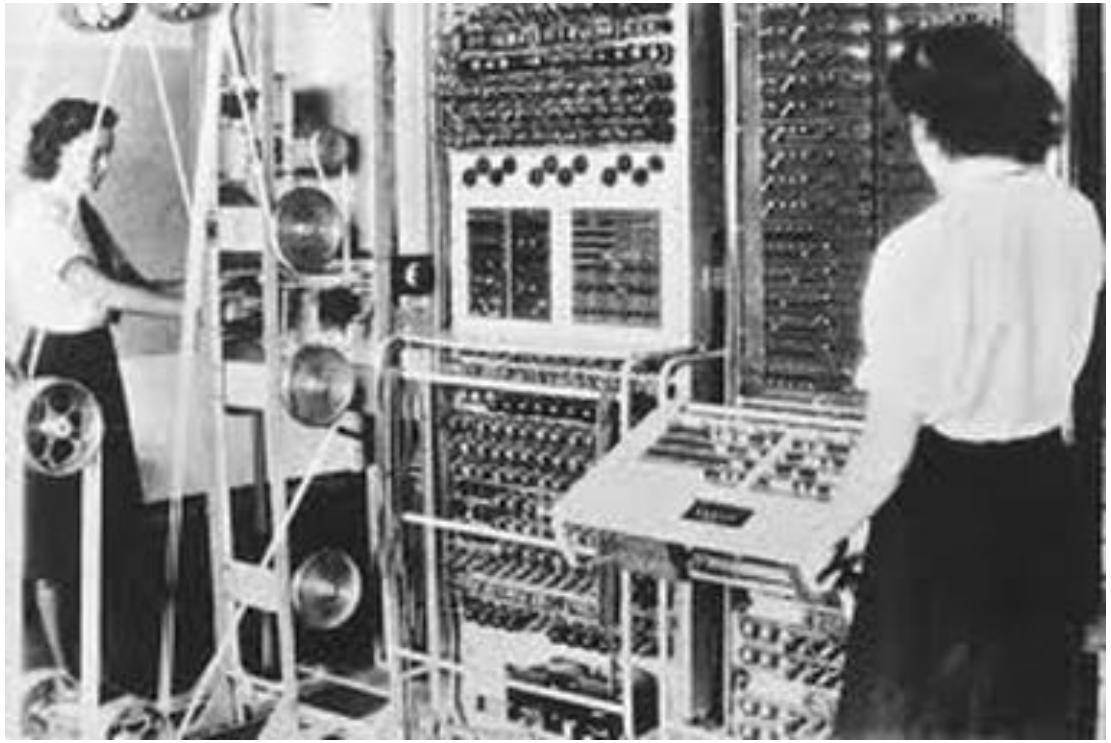
(punch card tabulation) reduces to ~3 months



https://www.dataversity.net/brief-history-big-data/

What is big data?

- **WW2**: British invent the Colossus machine to scan for patterns in intercepted Nazi codes. Scans 5,000 characters a second, reducing workload from weeks to hours



https://www.dataversity.net/brief-history-big-data/

What is big data?

- 1944: Librarian Fremont Rider @ Wesleyan estimates American university libraries doubling in size every 16 years:

"the Yale Library in 2040 will have approximately 200,000,000 volumes,
which will occupy over 6,000 miles of shelves…
[requiring] a cataloging staff of over six thousand persons"

https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#1fed7fb865a1

What is big data?

- 1961: Derek Price shows # of new scientific journals growing exponentially rather than linearly

"each advance generates a new series of advances at a reasonably constant birth rate, so that the number of births is strictly proportional to the size of the population of discoveries at any given time"

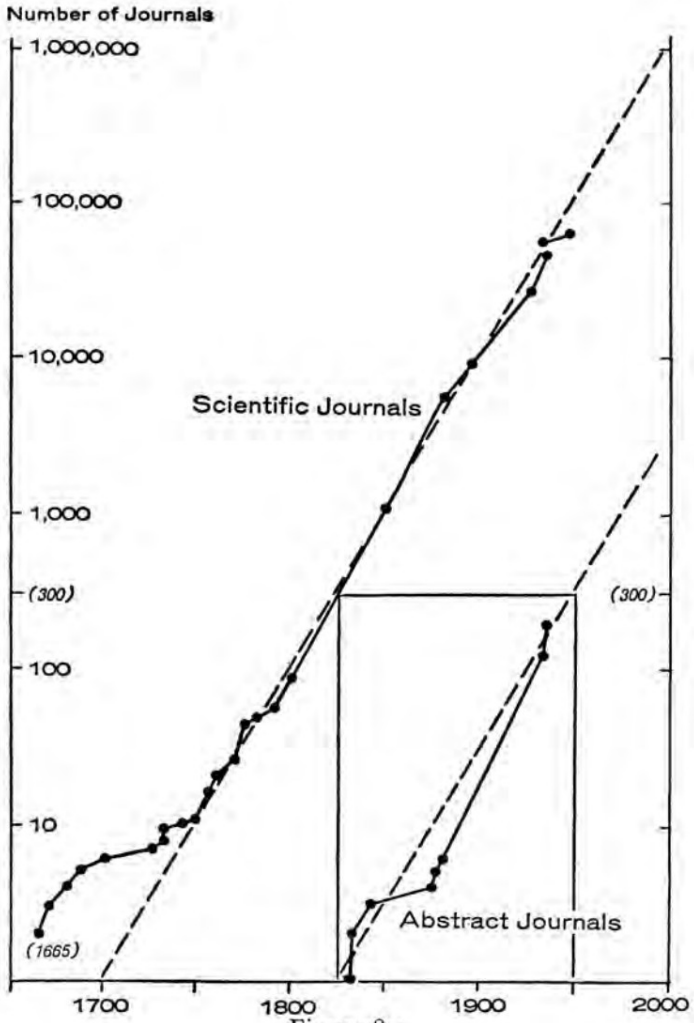


Figure 8.1

Number of journals founded (*not* surviving) as a function of date. The two uppermost points are taken from a slightly differently based list.

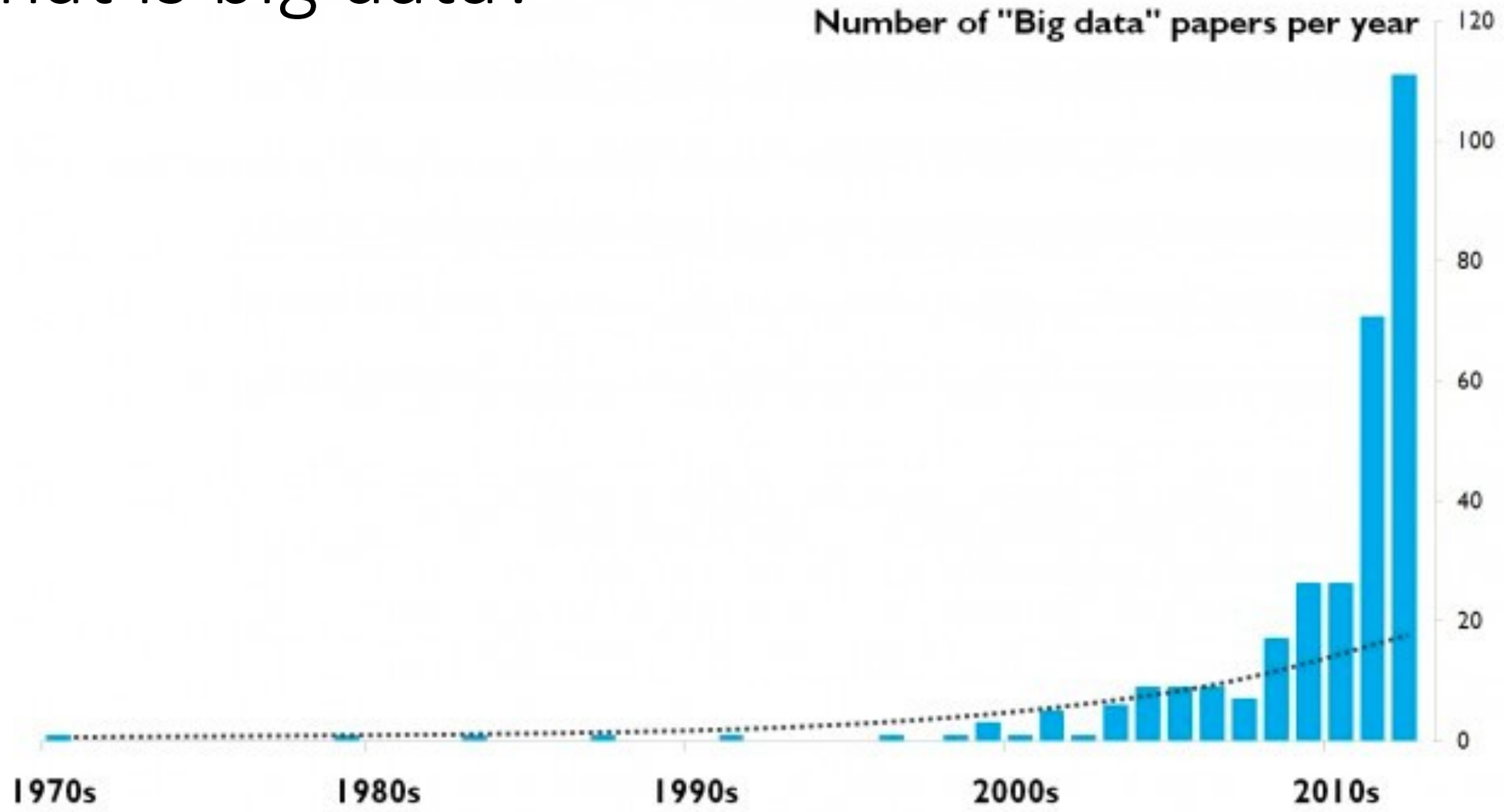
https://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#1fed7fb865a1

What is big data?

- **1969**: ARPANET kicks off the Internet
- **1977**: First personal computers come on the market
- **1989:** Tim Berners-Lee introduces the concept of the World Wide Web and the underlying protocols that support it (HTML, URL, HTTP)
- **1993:** CERN announces WWW will be free for everyone to develop and use
- **1990s– 2000s:** The explosion of the internet
- **2005:** Roger Mougalas coins term ‘Big Data’, referring to the scale of data that is nearly impossible to manage and process w/ available tools

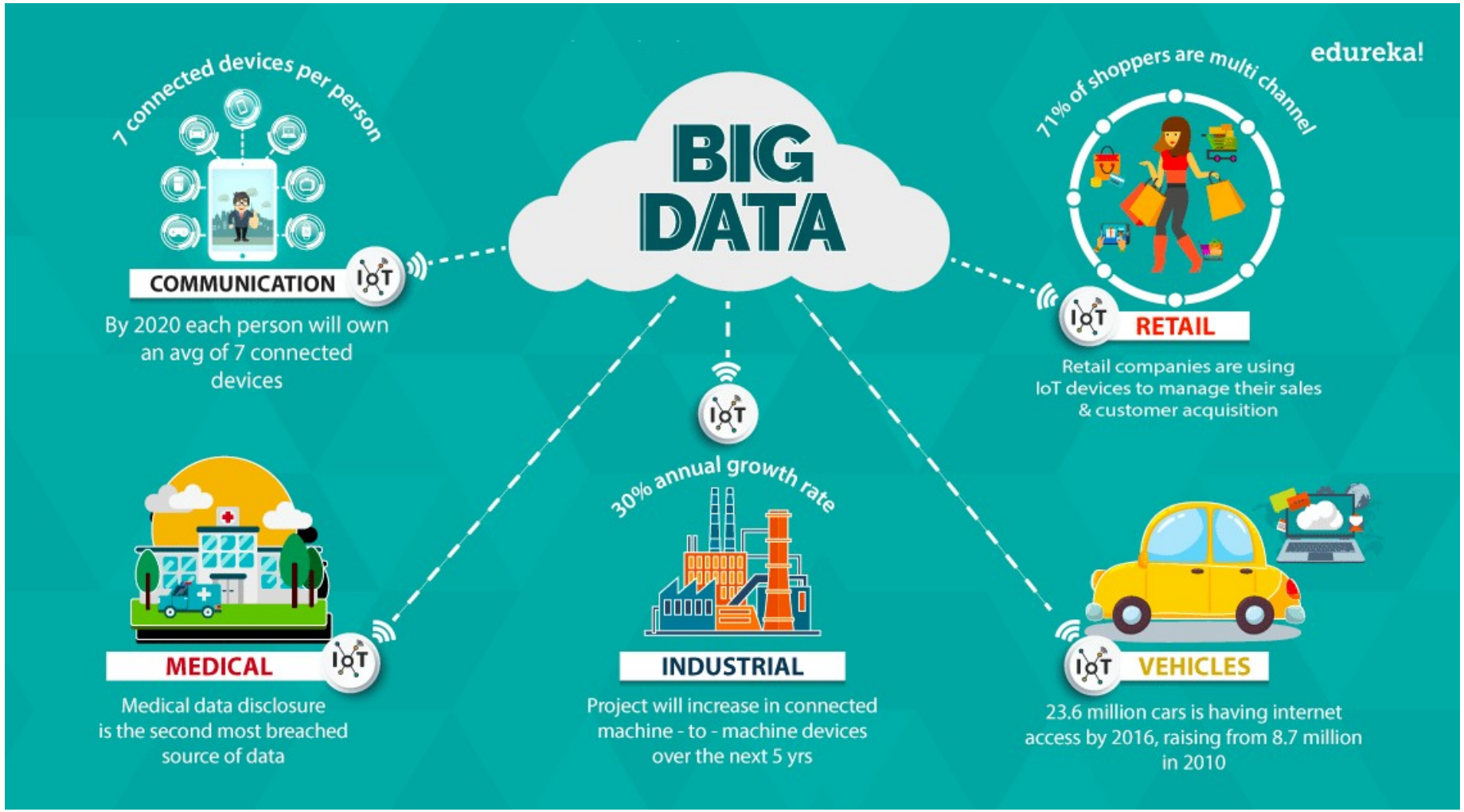
https://www.dataversity.net/brief-history-big-data/

What is big data?



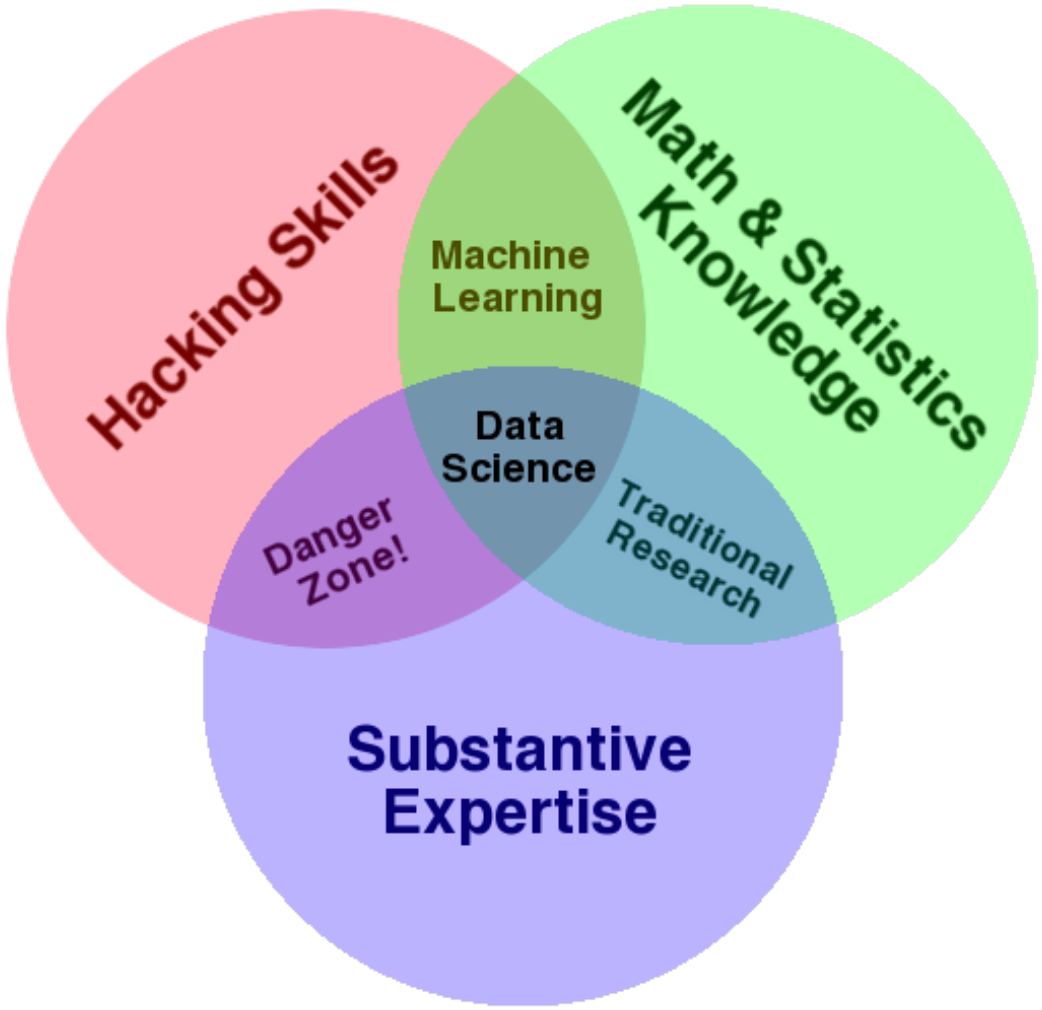
https://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/

Where is big data?



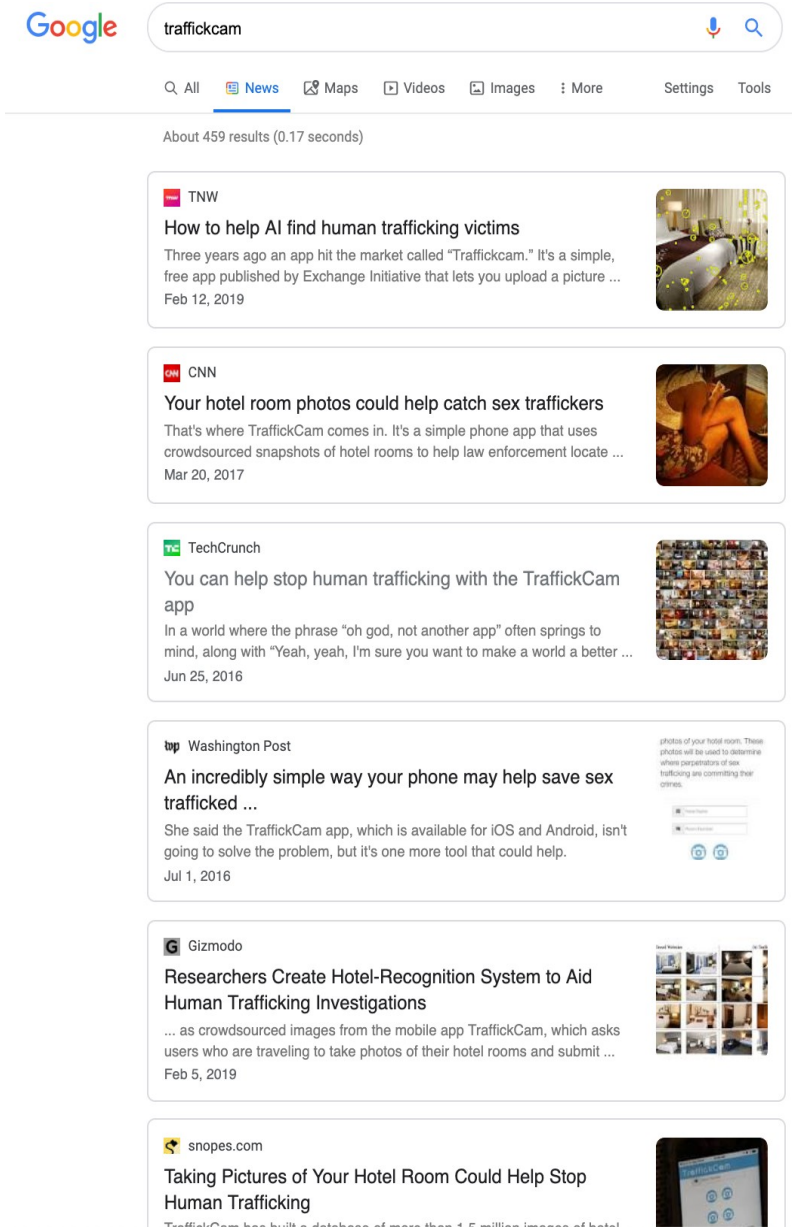
https://www.edureka.co/blog/big-data-applications-revolutionizing-various-domains/

What is data science?



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

The Big Data that I Work With



What’s the plan for this class?

- Learn the data science tools to collect, clean, understand, manipulate big data
- Exploration, Prediction and Inference

What’s the plan for this class?

- Learn the data science tools to collect, clean, understand, manipulate big data

- Exploration, Prediction and Inference
  - Exploration: Identify patterns in information
    - Tools: Visualization + Descriptive Statistics

What’s the plan for this class?

- Learn the data science tools to collect, clean, understand, manipulate big data

- Exploration, Prediction and Inference
  - Exploration: Identify patterns in information
    - Tools: Visualization + Descriptive Statistics

  - Prediction: make informed guesses about what we want to know, based on the patterns that we identified
    - Tools: Machine Learning + Optimization

What’s the plan for this class?

- Learn the data science tools to collect, clean, understand, manipulate big data

- Exploration, Prediction and Inference
  - Exploration: Identify patterns in information
    - Tools: Visualization + Descriptive Statistics
  - Prediction: make informed guesses about what we want to know, based on the patterns that we identified
    - Tools: Machine Learning + Optimization
  - Inference: quantify our certainty about our predictions
    - Tools: Statisticals Tests + Models

Intro to Python, Data Types, Sequences

Melissa Kline
@melissaekline

My strongest memory of learning to code is sitting thru my very first lecture silently panicking because while the FOR loop concept made perfect sense, I had no idea what I was supposed to do with it. I knew 'type it into Word' was wrong, but didn't know how to ask the Q...

Programming / Coding / Hacking

- Instruct a computer to carry out tasks
  - “Add 2 + 4”
  - “Round 3.14159265”

Programming / Coding / Hacking

- Instruct a computer to carry out tasks
  - “Add 2 + 4”
  - “Round 3.14159265”

- Combine multiple tasks into programs
  - “Add 2 + 4, then round 3.14159265, add the outputs together”
  - “Read in a file with the stock prices for a stock over the last 5 years and compute the average return”

Programming / Coding / Hacking

- Instruct a computer to carry out tasks
  - “Add 2 + 4”
  - “Round 3.14159265”

- Combine multiple tasks into programs
  - “Add 2 + 4, then round 3.14159265, add the outputs together”
  - “Read in a file with the stock prices for a stock over the last 5 years and compute the average return”

- Write these instructions in a language the computer can understand

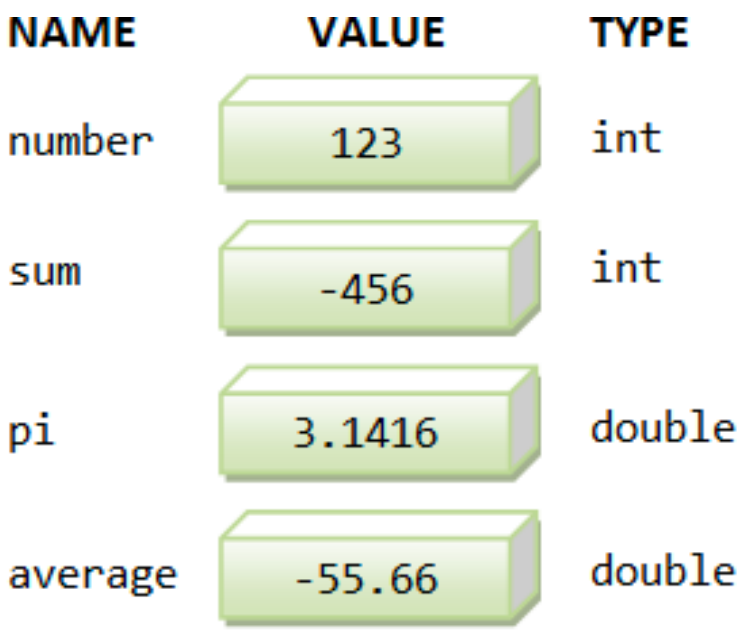
Parts of a Program

- Expressions

Expression Type	Operator	Example	Value
Addition	+	2 + 3	5
Subtraction	-	2 - 3	-1
Multiplication	*	2 * 3	6
Division	/	7 / 3	2.66667
Remainder	%	7 % 3	1
Exponentiation	**	2 ** 0.5	1.41421

Parts of a Program

- Variables



A variable has a *name*, stores a *value* of the declared *type*

https://www.ntu.edu.sg/home/ehchua/programming/cpp/c0_Introduction.html

Parts of a Program

- Operations

```
a = 1
b = 2
c = a + b
```

Parts of a Program

- Operations

```
a = 1
b = 2
c = a + b
```

- Functions (named operations)

```
max(2, 2 + 3, 4)
```

```
give_everyone_in_class_an_a(class_list)
```

Python

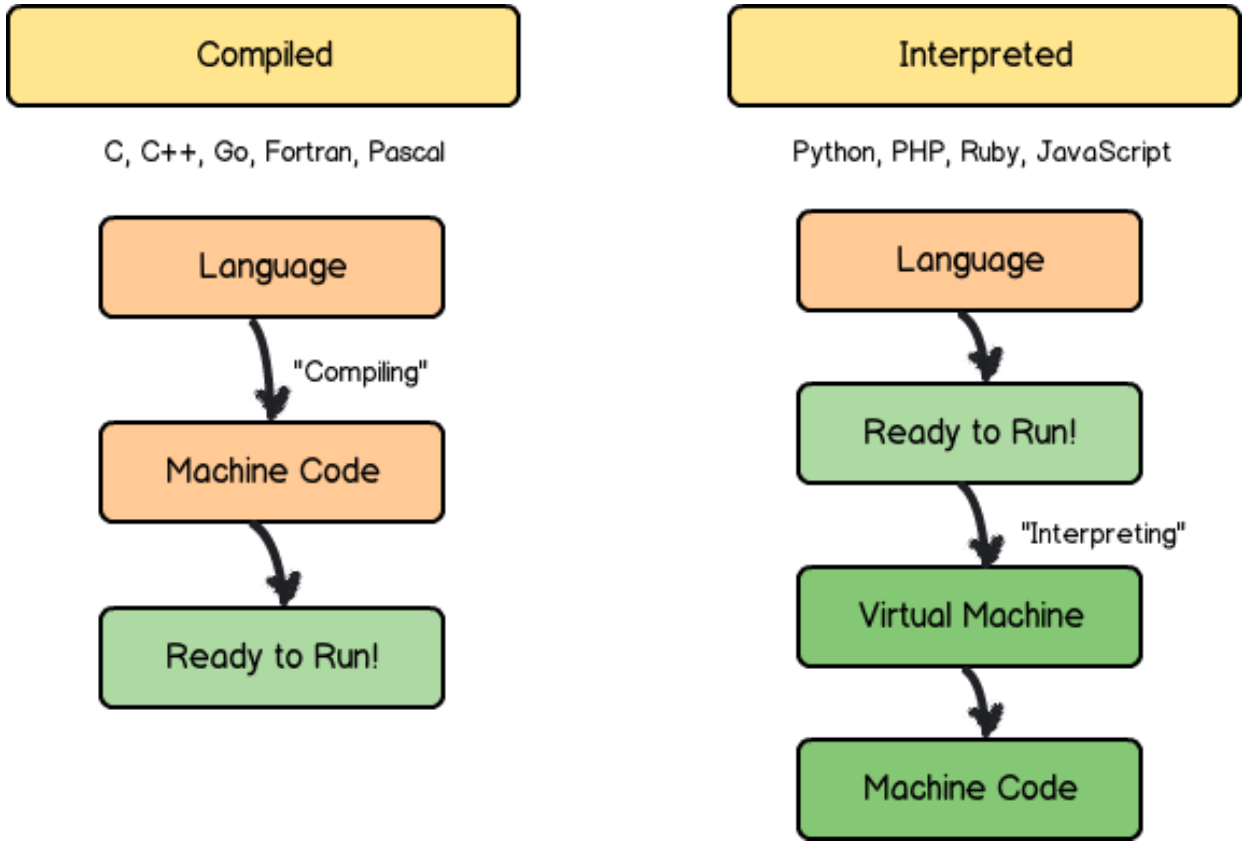
- Free & widely used (including in CSCI 1300)

Python

- Free & widely used (including in CSCI 1300)
- Compatible w/ different systems

Python

- Free & widely used (including in CSCI 1300)
- Compatible w/ different systems
- Interpreted rather than compiled



https://medium.com/@astermanuelg/blurred-lines-is-ruby-an-interpreted-language-2d3d6bca3d37

Python

- Free & widely used (including in CSCI 1300)
- Compatible w/ different systems
- Interpreted rather than compiled
- Dynamically typed

Static typing:

`String name;`	Variables have types
`name = "John";`	Values have types
~~`name = 34;`~~	Variables cannot change type

Dynamic typing:

`var name;`	Variables have no types
`name = "John";`	Values have types
`name = 34;`	Variables change type dynamically

http://techtales.co/2017/09/05/dynamic-typing-vs-static-typing-mean/

Python

- Free & widely used (including in CSCI 1300)
- Compatible w/ different systems
- Interpreted rather than compiled
- Dynamically typed
- Built in memory management

Python

- Free & widely used (including in CSCI 1300)
- Compatible w/ different systems
- Interpreted rather than compiled
- Dynamically typed
- Built in memory management
- Lots of built in modules + tons of very useful, well developed external libraries (NumPy, Pandas)

Python

- Free & widely used (including in CSCI 1300)
- Compatible w/ different systems
- Interpreted rather than compiled
- Dynamically typed
- Built in memory management
- Lots of built in modules + tons of very useful, well developed external libraries (NumPy, Pandas)
- Readable!

Hello World: Python vs. Java

A **"Hello, World!" program** generally is a computer program that outputs or displays the message "Hello, World!". Such a program is very simple in most programming languages, and is often used to illustrate the basic syntax of a programming language. It is often the first program written by people learning to code.[1][2]

https://en.wikipedia.org/wiki/%22Hello,_World!%22_program

```python
def hello_world():
    print("Hello, world!")

hello_world()
```

```java
public class HelloWorld {
    public static void main (String[] args) {
        System.out.println("Hello, world!");
    }
}
```

Whitespace in Python

```
def hello_world():
    print("Hello, world!")

hello_world()
```

```
def hello_world()
print("Hello, world!")

hello_world()
```

```
public class HelloWorld {
   public static void main (String[] args) {
      System.out.println("Hello, world!");
   }
}
```

```
public class HelloWorld {
public static void main (String[] args) {
System.out.println("Hello, world!");
}
}
```

```
public class HelloWorld { public static void main (String[]
args) {System.out.println("Hello, world!"); }}
```

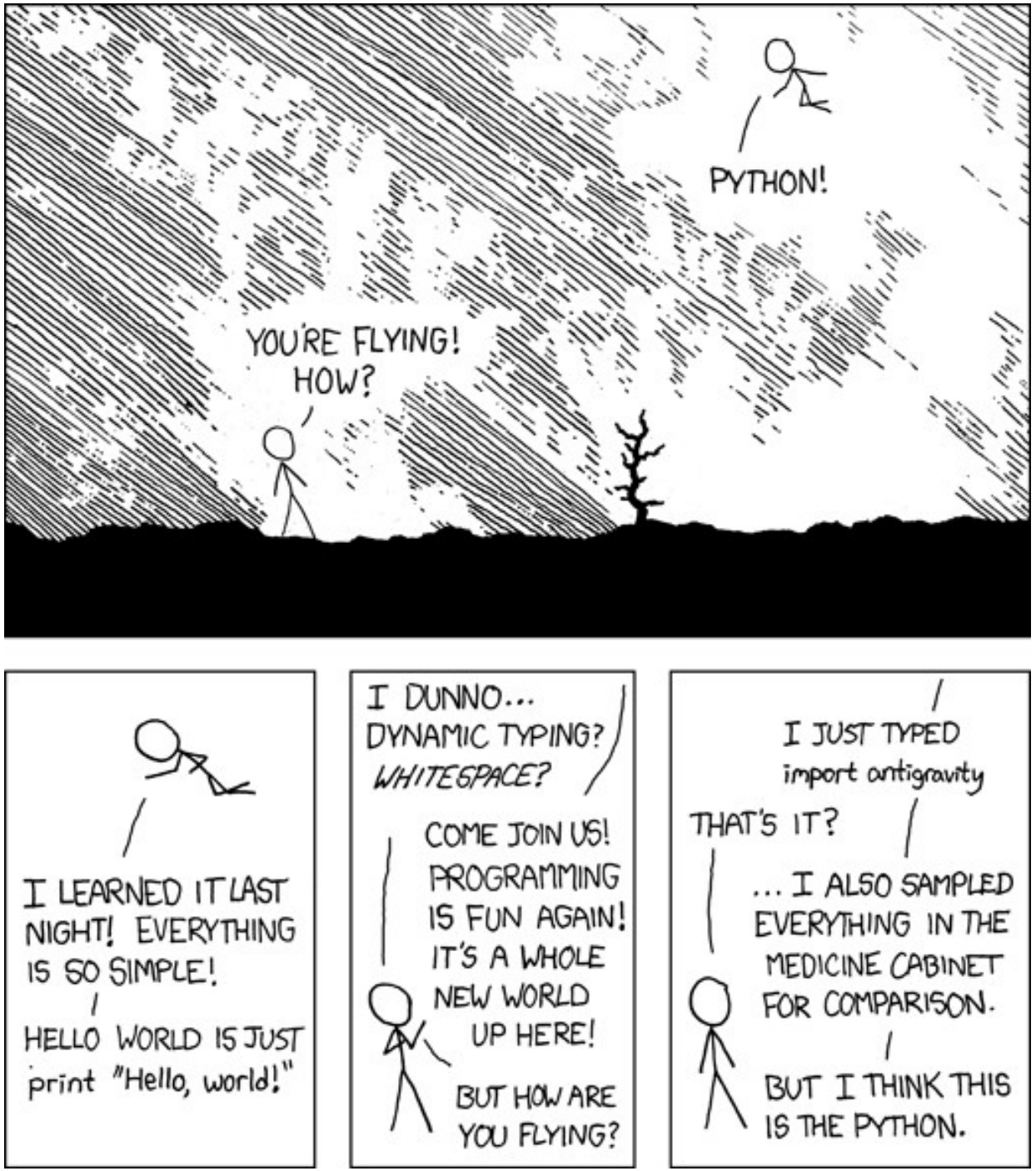
Whitespace in Python

```
def hello_world():
    print("Hello, world!")

hello_world()
```

```
def hello_world()
print("Hello, world!")

hello_world()
```

```
>>> def hello_world():
...     print("Hello, world!")
...
>>> hello_world()
Hello, world!
>>>
>>>
>>> def hello_world():
... print("Hello, world!")
  File "<stdin>", line 2
    print("Hello, world!")
        ^
IndentationError: expected an indented block
>>>
```

Python



https://xkcd.com/353/