Computational Analysis of Big Data

Week 1

Mostly an introduction, but also:

Coding with data in Python

Course overview

Course overview

Sessions

1. Coding with data in Python
2. A Data Scientist's most fundamental tools
3. Getting data—scraping and APIs
4. Machine learning 1
5. Machine learning 2
6. Networks
7. Natural language processing
8. Crunching Big Data with MapReduce
9. Lab work on project report
10. Lab work on project report
11. Project presentations

Course overview

Sessions

1. **Coding with data in Python**
2. A Data Scientist's most fundamental tools
3. Getting data—scraping and APIs
4. Machine learning 1
5. Machine learning 2
6. Networks
7. Natural language processing
8. Crunching Big Data with MapReduce
9. Lab work on project report
10. Lab work on project report
11. Project presentations

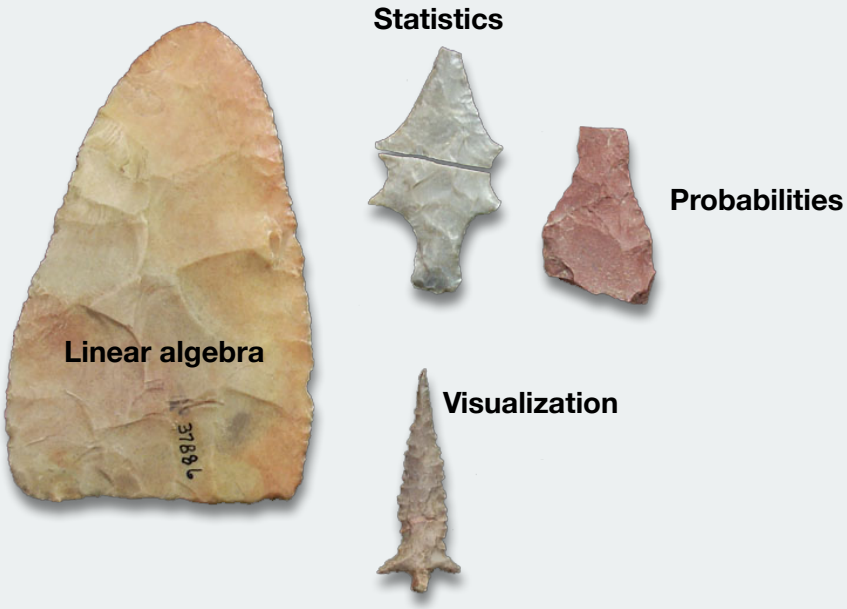
```python
def randomize_by_edge_swaps(G, num_iterations):
    """Randomizes the graph by swapping edges in such a way that
    preserves the degree distribution of the original graph.

    Source: https://gist.github.com/gotgenes/2770023
    """
    newgraph = G.copy()
    edge_list = newgraph.edges()
    num_edges = len(edge_list)
    total_iterations = num_edges * num_iterations

    for i in xrange(total_iterations):
        rand_index1 = int(round(random.random() * (num_edges - 1)))
        rand_index2 = int(round(random.random() * (num_edges - 1)))
        original_edge1 = edge_list[rand_index1]
        original_edge2 = edge_list[rand_index2]
        head1, tail1 = original_edge1
        head2, tail2 = original_edge2

        # Flip a coin to see if we should swap head1 and tail1 for
        # the connections
        if random.random() >= 0.5:
            head1, tail1 = tail1, head1

        if head1 == tail2 or head2 == tail1:
            continue

        if newgraph.has_edge(head1, tail2) or newgraph.has_edge(
                head2, tail1):
            continue

        # Suceeded checks, perform the swap
        original_edge1_data = newgraph[head1][tail1]
        original_edge2_data = newgraph[head2][tail2]

        newgraph.remove_edges_from((original_edge1, original_edge2))

        new_edge1 = (head1, tail2, original_edge1_data)
        new_edge2 = (head2, tail1, original_edge2_data)
        newgraph.add_edges_from((new_edge1, new_edge2))

        # Now update the entries at the indices randomly selected
        edge_list[rand_index1] = (head1, tail2)
        edge_list[rand_index2] = (head2, tail1)

    assert len(newgraph.edges()) == num_edges
    return newgraph
```

Aim: Get comfortable with basic Python

Course overview

Sessions

1. Coding with data in Python
2. **A Data Scientist's most fundamental tools**
3. Getting data—scraping and APIs
4. Machine learning 1
5. Machine learning 2
6. Networks
7. Natural language processing
8. Crunching Big Data with MapReduce
9. Lab work on project report
10. Lab work on project report
11. Project presentations

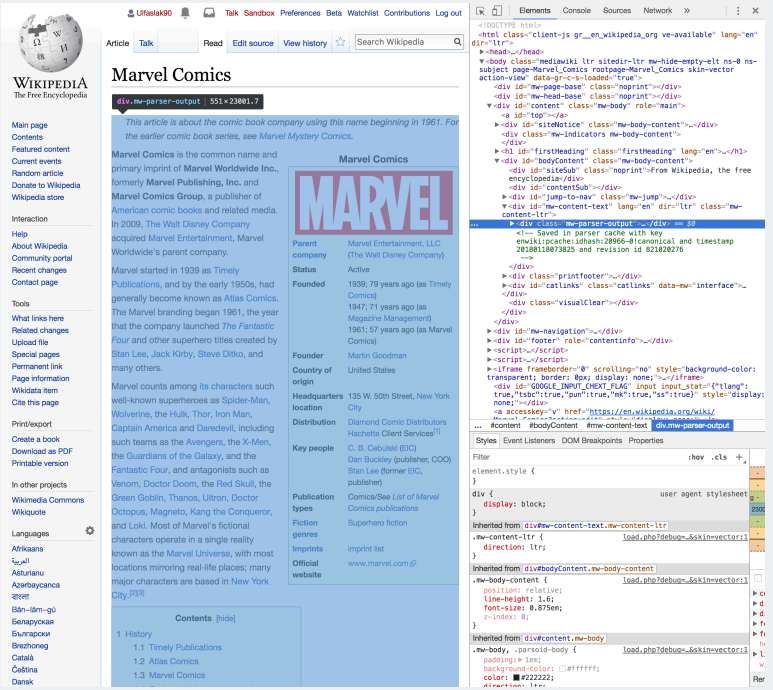


Aim: Refresh math skills and learn how to use them in Python

Course overview

Sessions

1. Coding with data in Python
2. A Data Scientist's most fundamental tools
3. **Getting data—scraping and APIs**
4. Machine learning 1
5. Machine learning 2
6. Networks
7. Natural language processing
8. Crunching Big Data with MapReduce
9. Lab work on project report
10. Lab work on project report
11. Project presentations

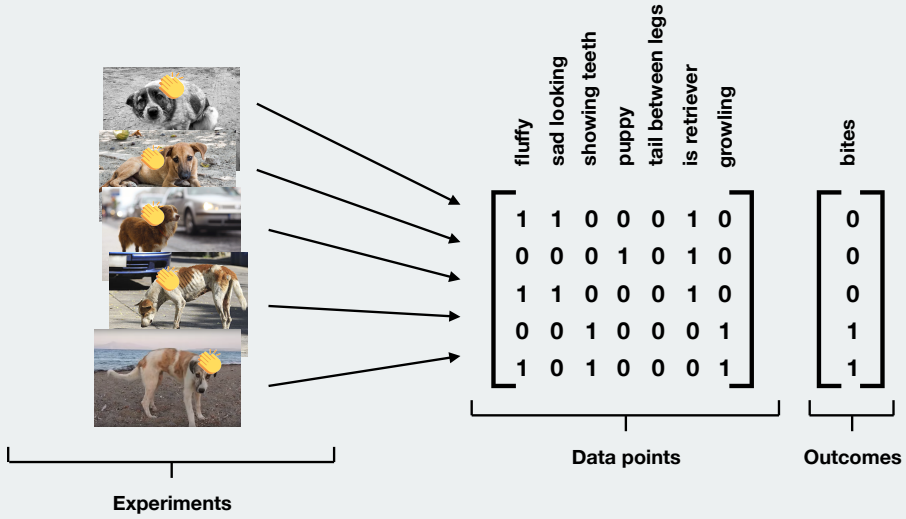


Aim: Learn how to get data from web and get some from the Wiki API

Course overview

Sessions

1. Coding with data in Python
2. A Data Scientist's most fundamental tools
3. Getting data—scraping and APIs
4. **Machine learning 1**
5. **Machine learning 2**
6. Networks
7. Natural language processing
8. Crunching Big Data with MapReduce
9. Lab work on project report
10. Project presentations

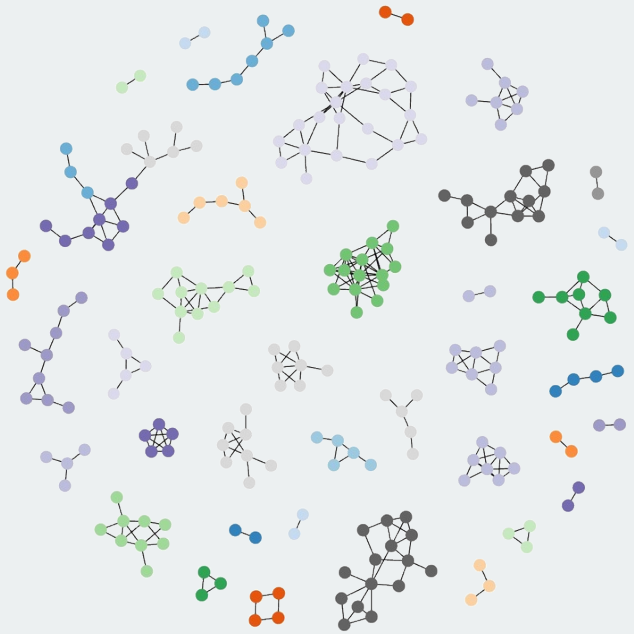


Aim: Understand the paradigm, learn how (some of) it works

Course overview

Sessions

1. Coding with data in Python
2. A Data Scientist's most fundamental tools
3. Getting data—scraping and APIs
4. Machine learning 1
5. Machine learning 2
6. **Networks**
7. Natural language processing
8. Crunching Big Data with MapReduce
9. Lab work on project report
10. Lab work on project report
11. Project presentations



Aim: Learn how to describe and visualize complex data as a network

Course overview

Sessions

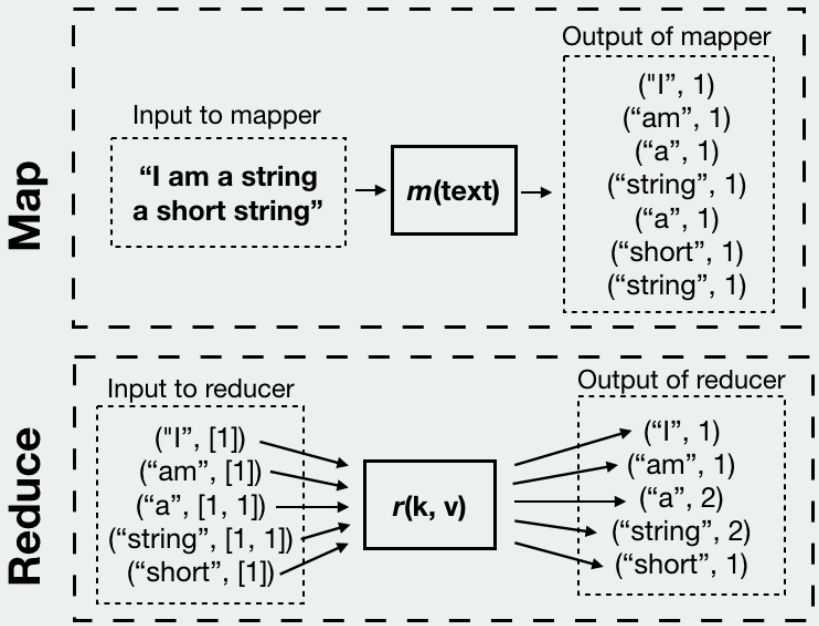
1. Coding with data in Python
2. A Data Scientist's most fundamental tools
3. Getting data—scraping and APIs
4. Machine learning 1
5. Machine learning 2
6. Networks
7. **Natural language processing**
8. Crunching Big Data with MapReduce
9. Lab work on project report
10. Lab work on project report
11. Project presentations

Aim: Get introduced to the (huge) field of NLP, and learn a few skills

Course overview

Sessions

1. Coding with data in Python
2. A Data Scientist's most fundamental tools
3. Getting data—scraping and APIs
4. Machine learning 1
5. Machine learning 2
6. Networks
7. Natural language processing
8. **Crunching Big Data with MapReduce**
9. Lab work on project report
10. Lab work on project report
11. Project presentations

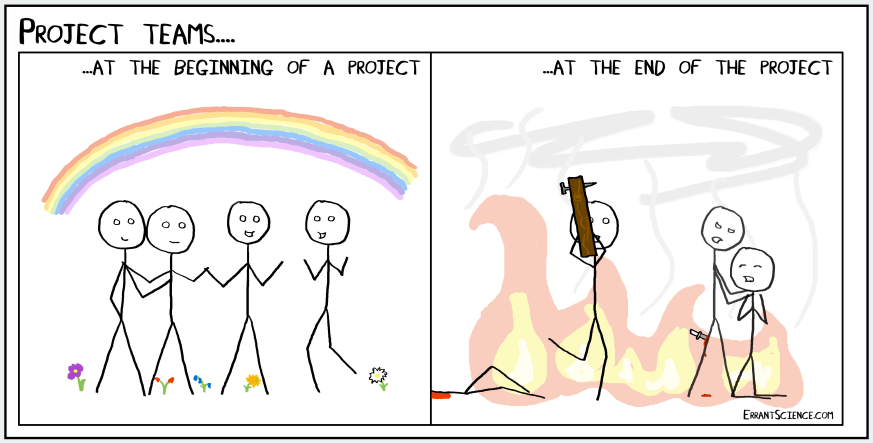


Aim: Process a massive dataset with an awesome method

Course overview

Sessions

1. Coding with data in Python
2. A Data Scientist's most fundamental tools
3. Getting data—scraping and APIs
4. Machine learning 1
5. Machine learning 2
6. Networks
7. Natural language processing
8. Crunching Big Data with MapReduce
9. **Lab work on project report**
10. **Lab work on project report**
11. **Project presentations**



Aim: Synthesize all the things you have learned and make a project for your portfolio

What will you learn?

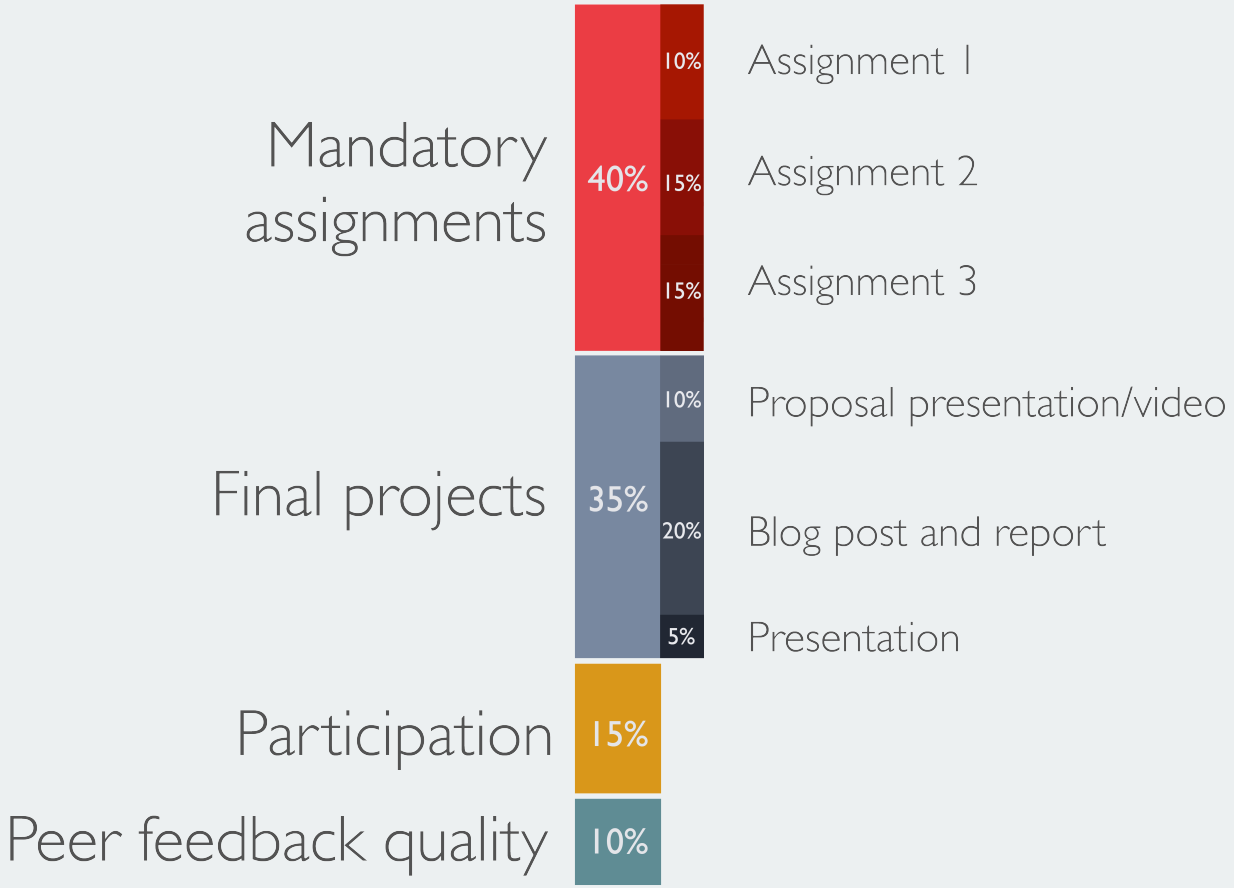
Knowledge and competences

- Get to know the landscape of problems and tools in Data Science
- Learn what Big Data is in this context
- Know where to start when you want to analyze something that requires lots of data

Concrete skills

- Juggle data in Python
- Predict outcomes from input data (machine learning)
- Visualize complex data
- Analyze Big Data using parallel methods (MapReduce)

How will you be graded?

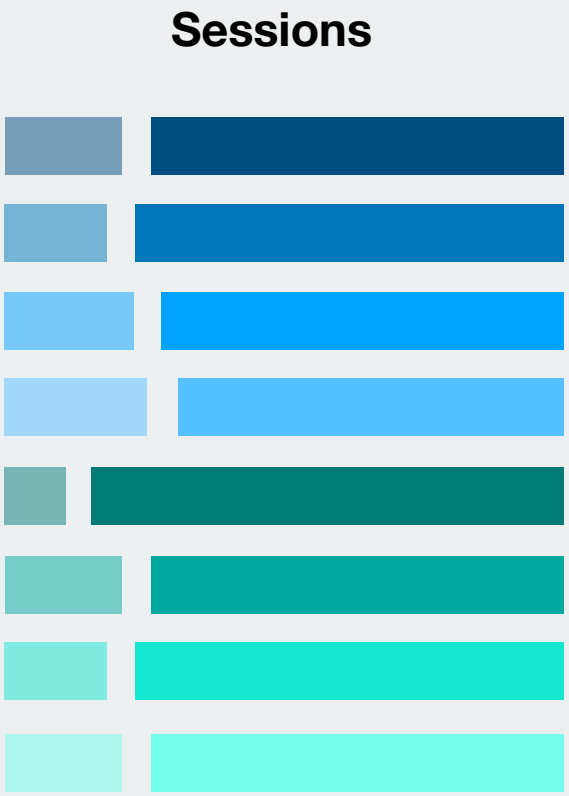


How assignments work

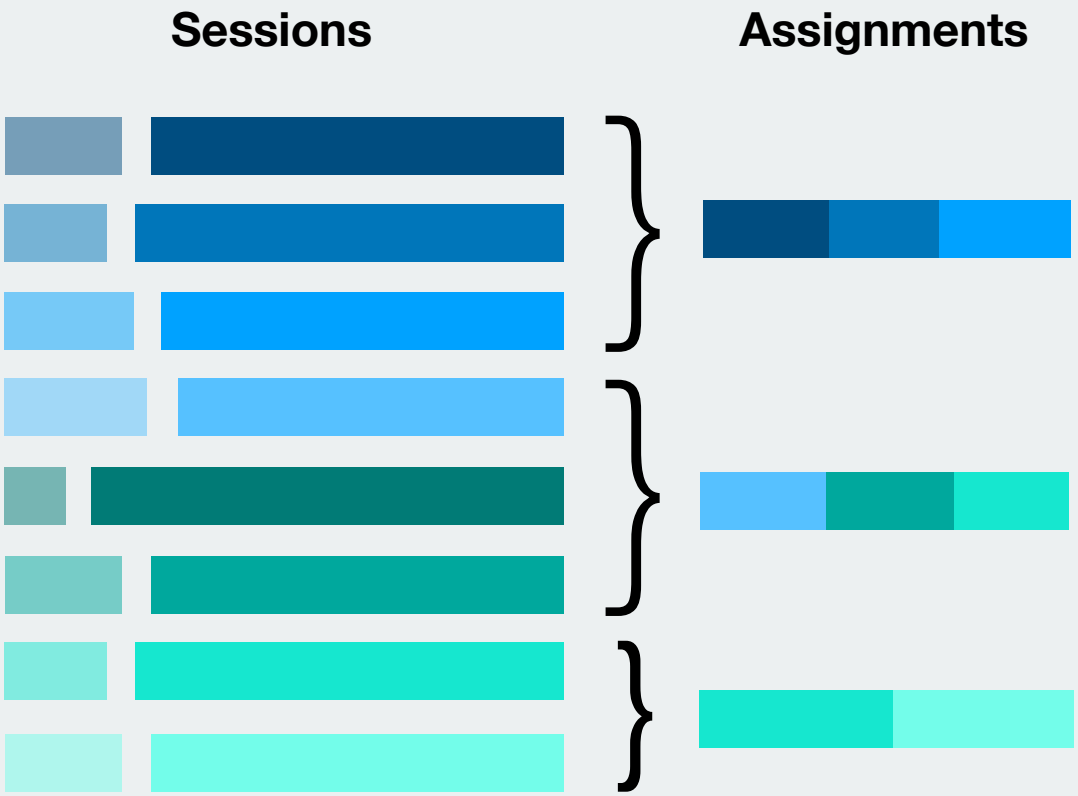
Sessions

Talk **Solve exercises**

How assignments work



How assignments work



The final project

Feb. 17

I validate your project idea

Part A Feb. 22

You present a proposal video

Deliverable

Part B March 18

You deliver a blog post and your code

Deliverable

March 19

You give a presentation

Previous students projects

“What makes us happy?”

“The secrets hidden in your Venmo and Instagram data”

"Pizza Makes the World go Round"

How to do well in this course

Best strategy:

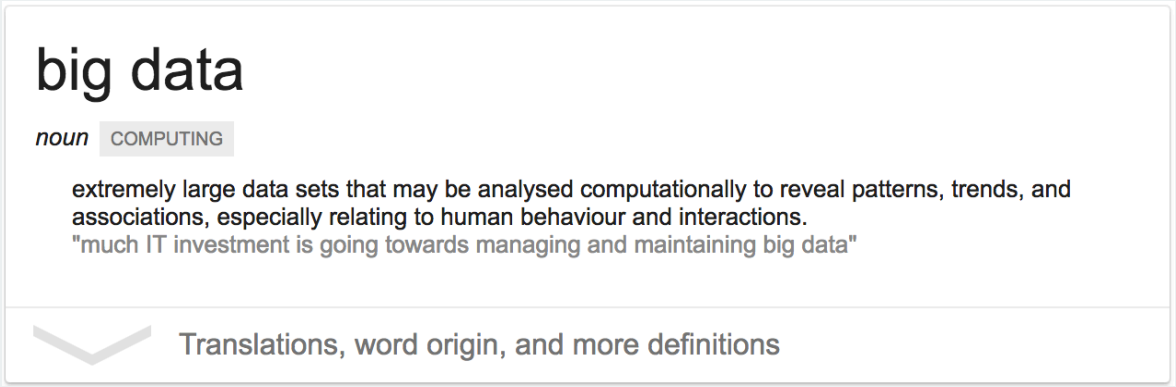
1. Complete the *preparation goals* for each session (see calendar on Canvas)

2. Be inquisitive. Ask lots of questions to your neighbors and me, and up your googling-game

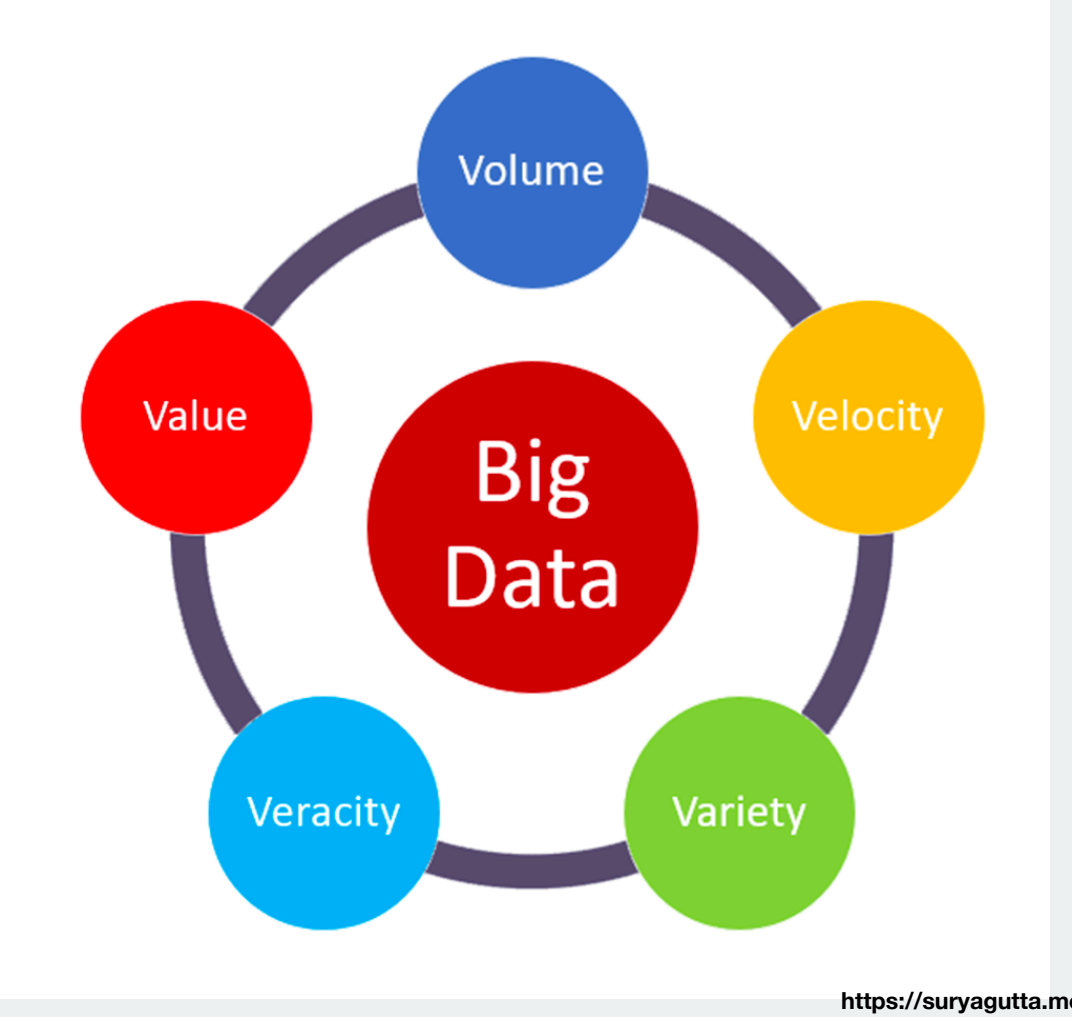
Everything else is on Canvas

What is Big Data?

What is Big Data?

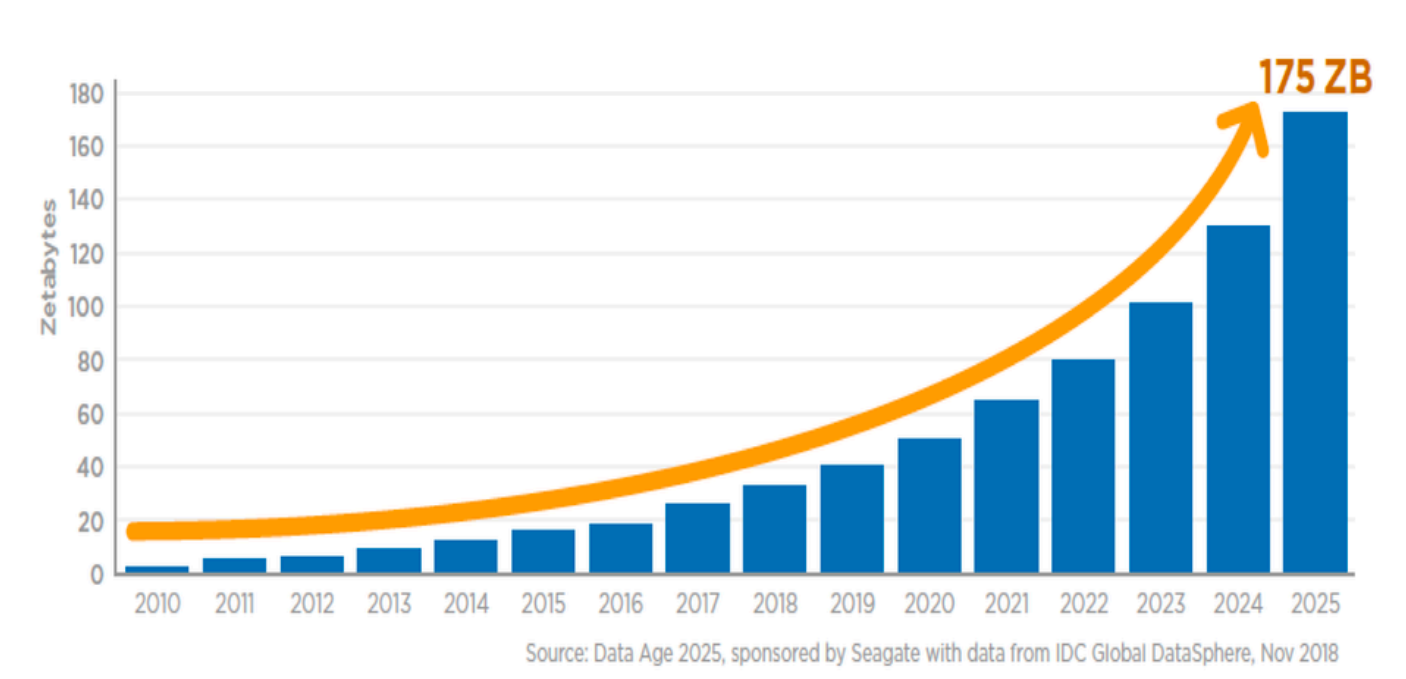


The 5Vs of Big Data



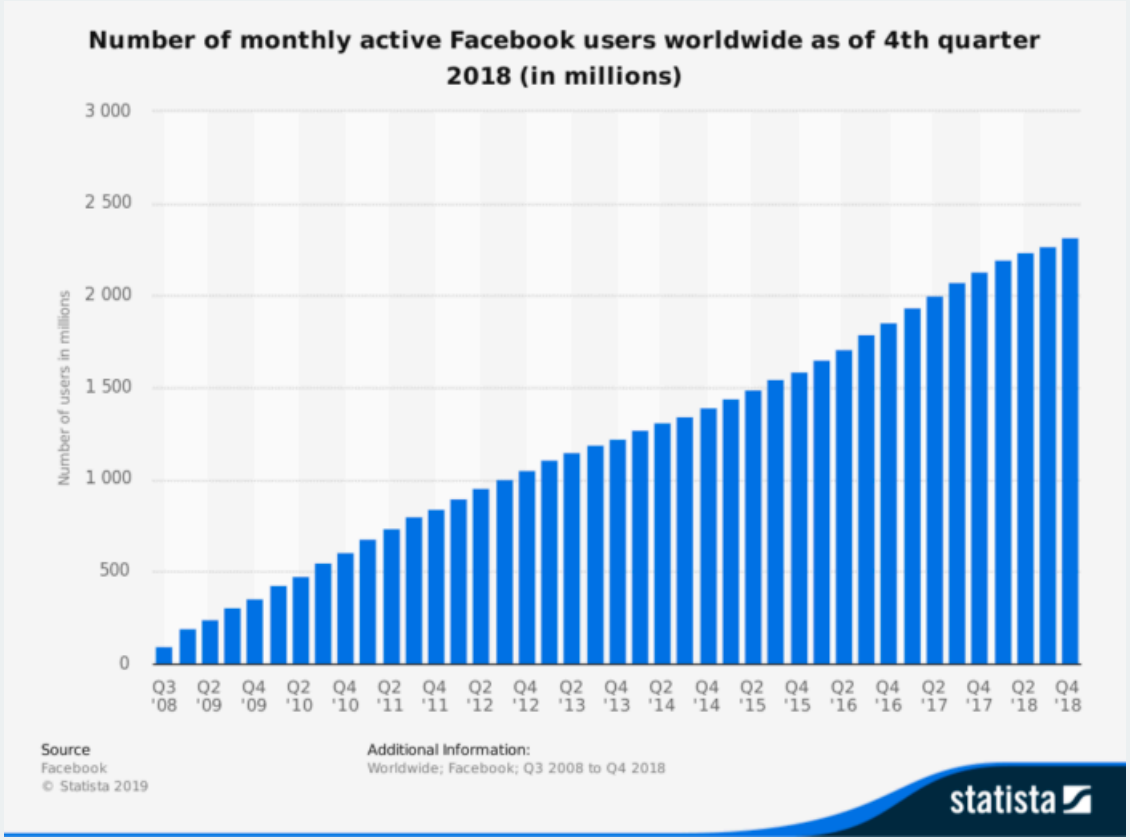
https://suryagutta.medium.com/the-5-vs-of-big-data-2758bfcc51d

The 5Vs of Big Data: Volume



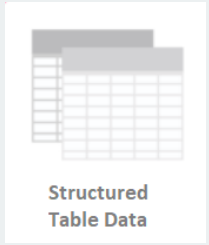
https://suryagutta.medium.com/the-5-vs-of-big-data-2758bfcc51d

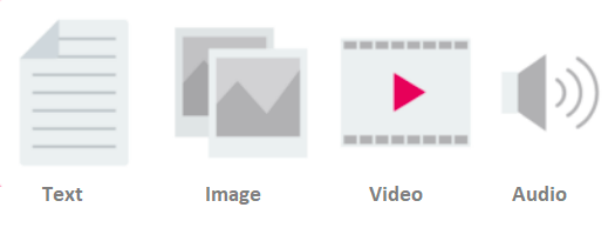
The 5Vs of Big Data: Velocity



https://suryagutta.medium.com/the-5-vs-of-big-data-2758bfcc51d

The 5Vs of Big Data: Variety





https://suryagutta.medium.com/the-5-vs-of-big-data-2758bfcc51d

The 5Vs of Big Data: Veracity

- Quality
- Integrity
- Credibility
- Accuracy

https://suryagutta.medium.com/the-5-vs-of-big-data-2758bfcc51d

The 5Vs of Big Data: Value

https://suryagutta.medium.com/the-5-vs-of-big-data-2758bfcc51d

Other Vs

- **Viscosity** (complexity or degree of correlation)
- **Variability** (inconsistency in data flow)
- **Volatility** (durability or how long time data is valid and how long it should be stored)
- **Viability** (capability to be live and active)
- **Validity** (understandable to find the hidden relationships)

https://suryagutta.medium.com/the-5-vs-of-big-data-2758bfcc51d

Where and for what is Big Data being used?

Where and for what is Big Data being used?

Ads

theTradeDesk

Omnicom WPP

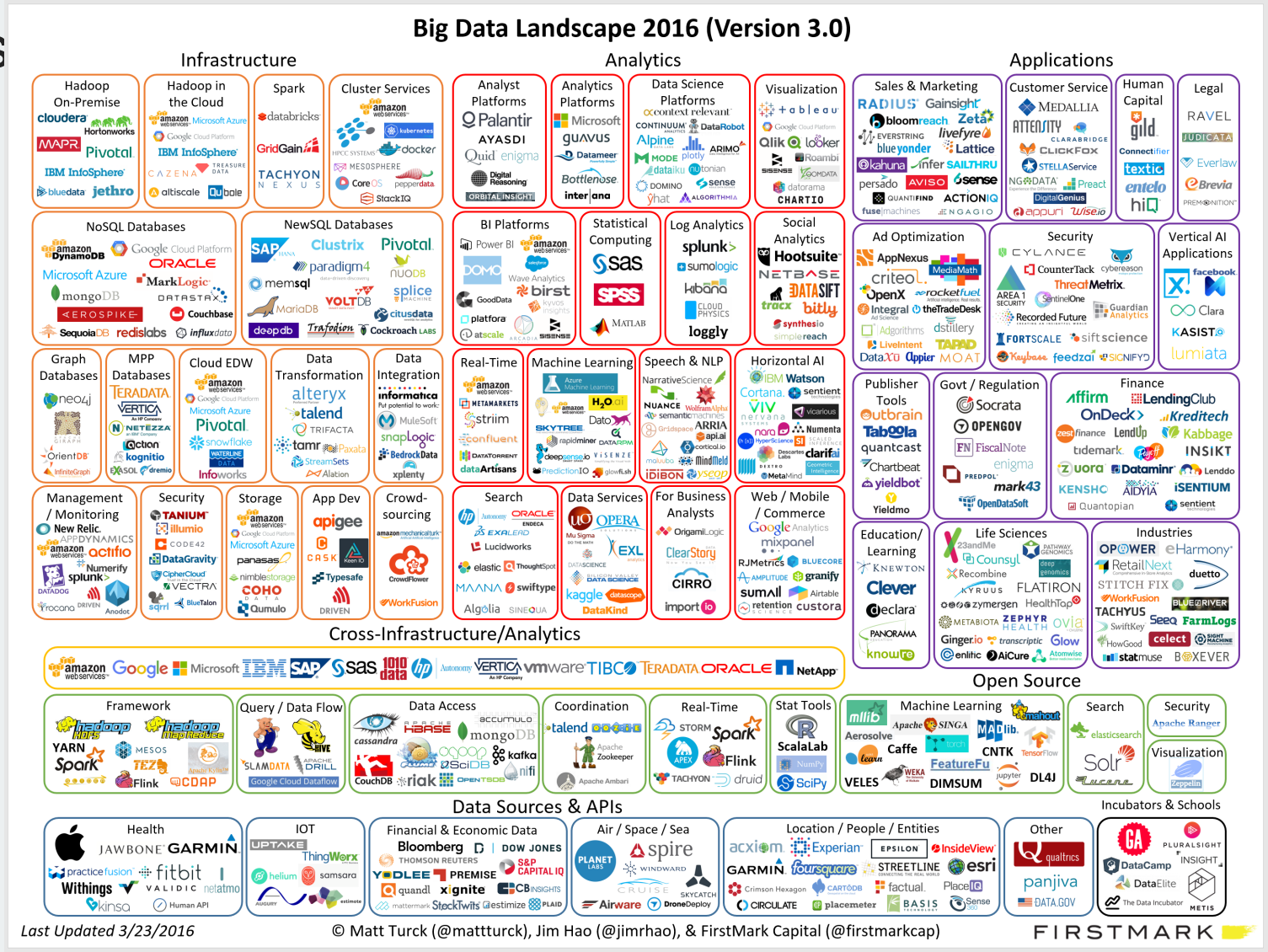
User customization

Research

Ted Talk: "Is Big Data Killing Creativity?"

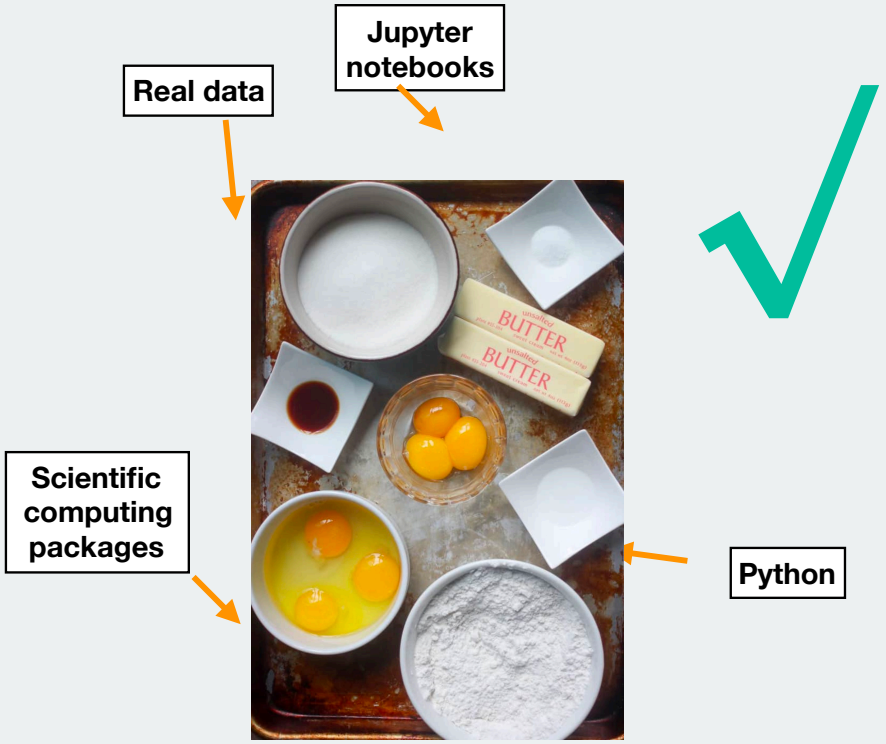
How does one work with Big Data?

How does


Big Data Landscape 2016 (Version 3.0)
Infrastructure
Analytics
Applications
Hadoop On-Premise
Hadoop in the Cloud
Spark
Cluster Services
Analyst Platforms
Analytics Platforms
Data Science Platforms
Visualization
Sales & Marketing
Customer Service
Human Capital
Legal
NoSQL Databases
NewSQL Databases
BI Platforms
Statistical Computing
Log Analytics
Social Analytics
Ad Optimization
Security
Vertical AI Applications
Graph Databases
MPP Databases
Cloud EDW
Data Transformation
Data Integration
Real-Time
Machine Learning
Speech & NLP
Horizontal AI
Publisher Tools
Govt / Regulation
Finance
Management / Monitoring
Security
Storage
App Dev
Crowd-sourcing
Search
Data Services
For Business Analysts
Web / Mobile / Commerce
Education / Learning
Life Sciences
Industries
Cross-Infrastructure/Analytics
Open Source
Framework
Query / Data Flow
Data Access
Coordination
Real-Time
Stat Tools
Machine Learning
Search
Security
Visualization
Data Sources & APIs
Health
IOT
Financial & Economic Data
Air / Space / Sea
Location / People / Entities
Other
Incubators & Schools
Last Updated 3/23/2016
© Matt Turck (@mattturck), Jim Hao (@jimrhao), & FirstMark Capital (@firstmarkcap)
FIRSTMARK

How does one work with Big Data?

In this course: Everything from scratch



Let's get started!

1. If you haven't already: download and install Anaconda (Python 3.8 version) https://www.continuum.io/downloads.

2. Make a folder for this course. Open a terminal (Mac) or a console (PC) and navigate to that folder.

3. Run the following command in your terminal/console:

   ```
   git clone https://github.com/lucian979/CarletonBD
   ```

4. In your terminal/console, navigate into the exercises folder inside of the newly created **CarletonBD** directory, and run the following command:

   ```
   jupyter notebook exercises_week1.ipynb
   ```