Week 1 Lecture 1: What is Data Analytics?

Daryl DeFord

Data 115: Introduction to Data Analytics
Fall 2020

Outline

1. Introduction and Welcome
2. Course Overview
3. Break
4. What is Data Analytics?
5. Initial Examples
6. The Data Science Process

Course Overview

- Introduction to methods and tools for modern data analysis
- Lots of hands-on experience with exposure to Python and other common data analytics tools
- Critical analysis of published data projects
- Ethics and professional development
- Read the full syllabus (on Blackboard) for more complete details.

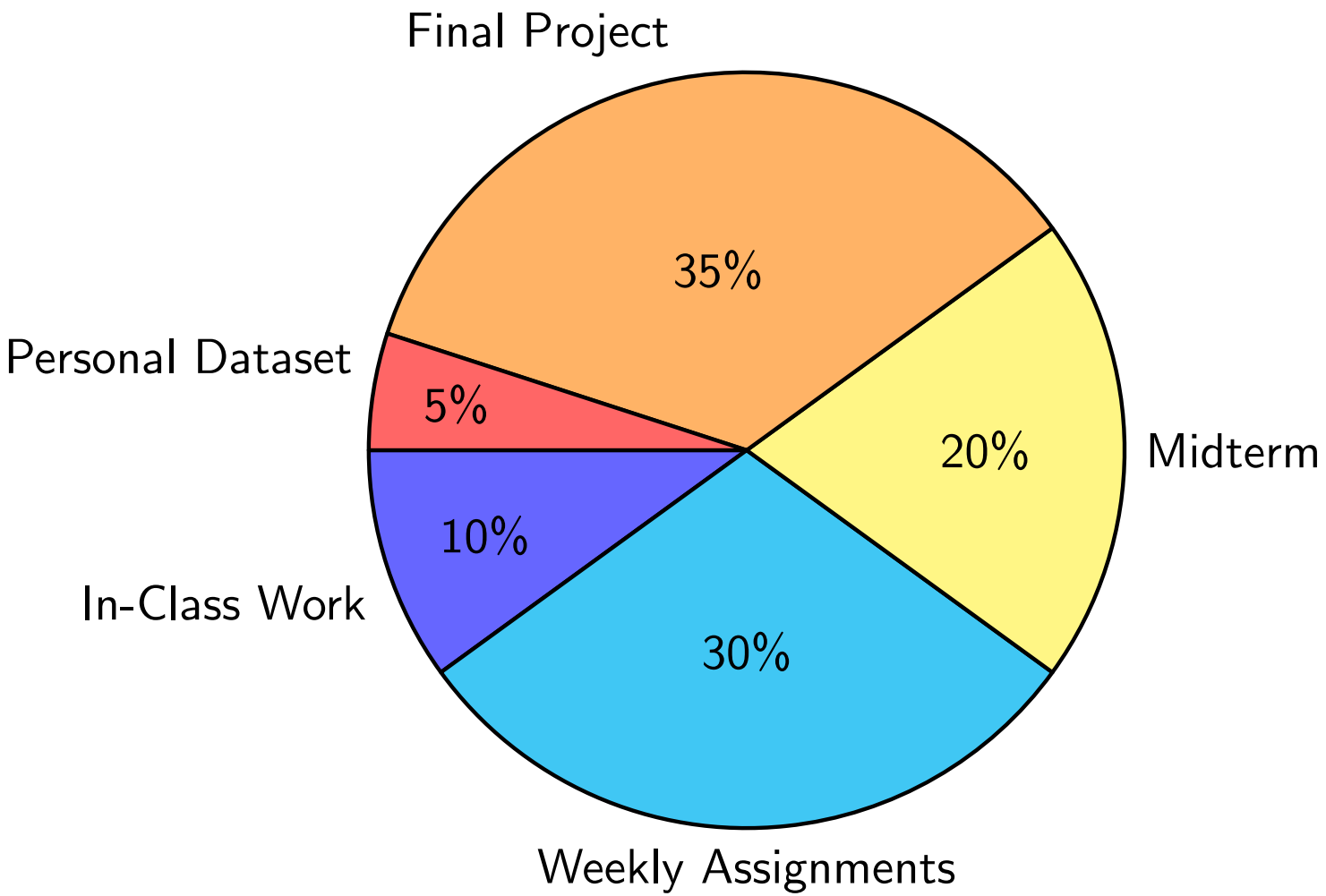
Course Resources

- Blackboard will be the main repository for material related to the course.
- Lecture videos will be uploaded weekly
- Readings and discussion materials will also be posted weekly
- Data and computational examples will be handled through the CoCalc platform

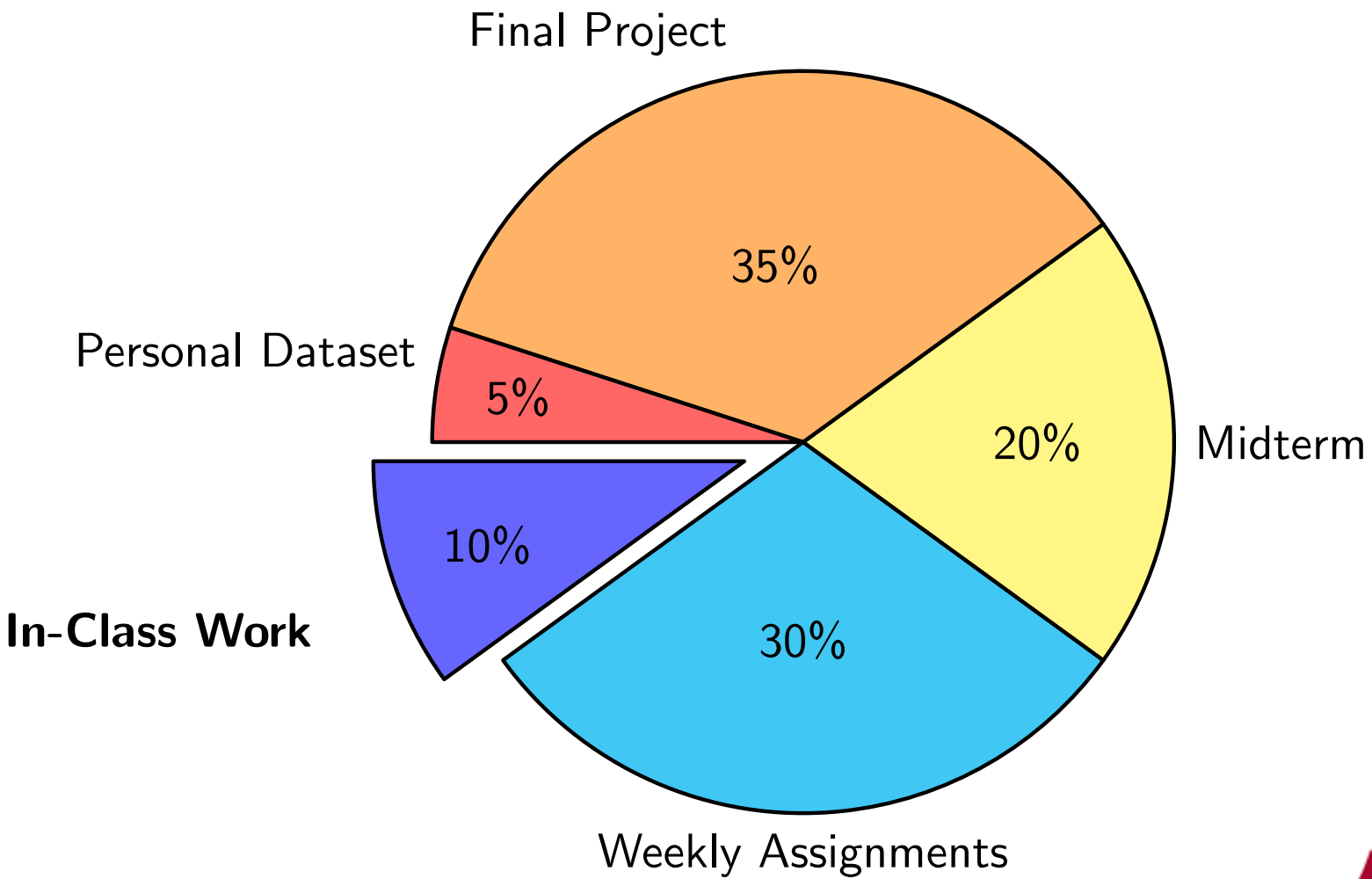
Course Material

- New content will be communicated during the Tuesday course meeting
- Group discussions and labs will occur on Thursdays
- Weekly topics are listed in the syllabus
- Discussions through Blackboard forums

Assignments and Grades



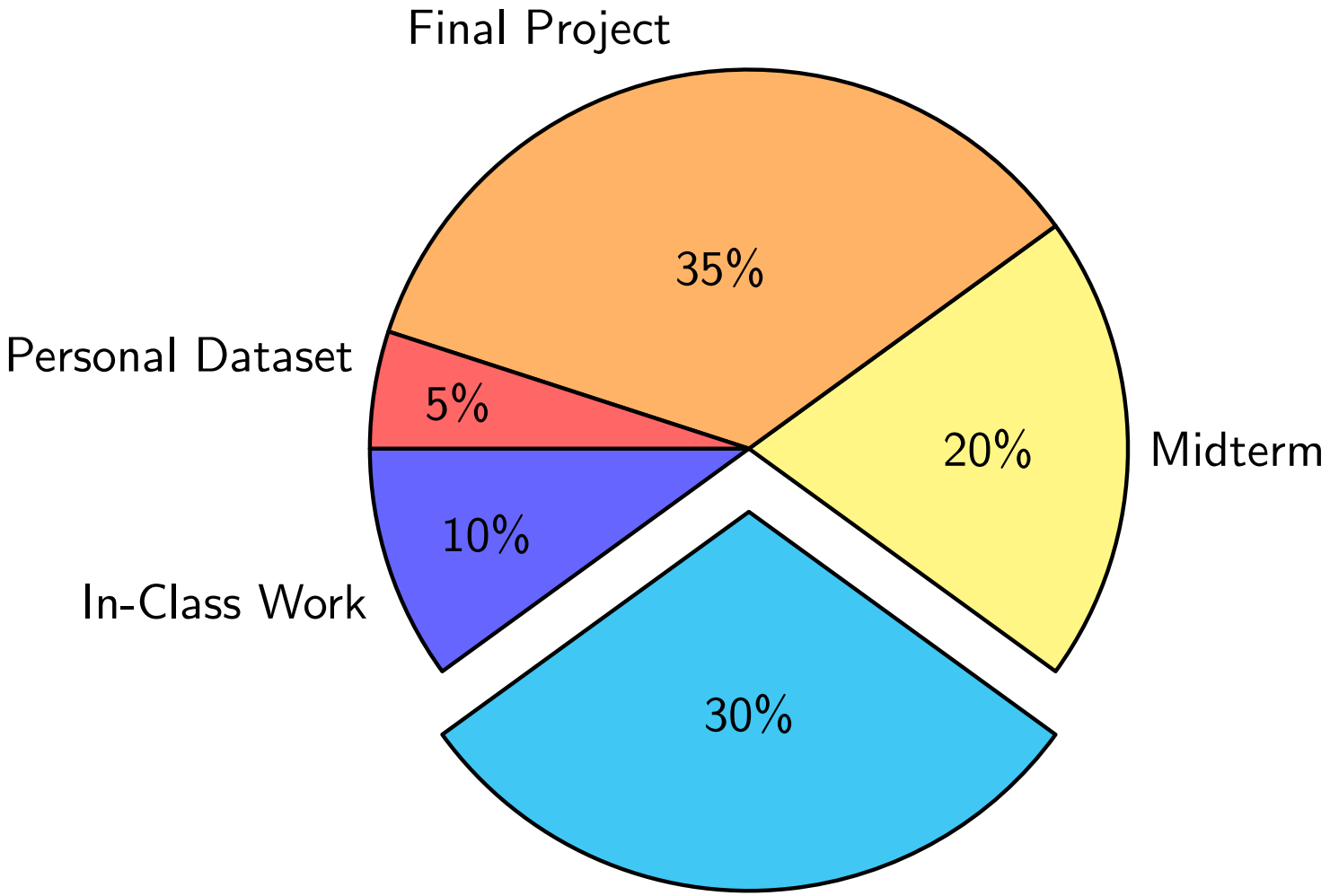
Assignments and Grades



In-Class Work

- Reading discussion questions (groups in breakout rooms)
- Responses need not be unanimous but should be well-justified
- Data lab examples (groups in breakout rooms)
- 15 minute quizzes (individual)

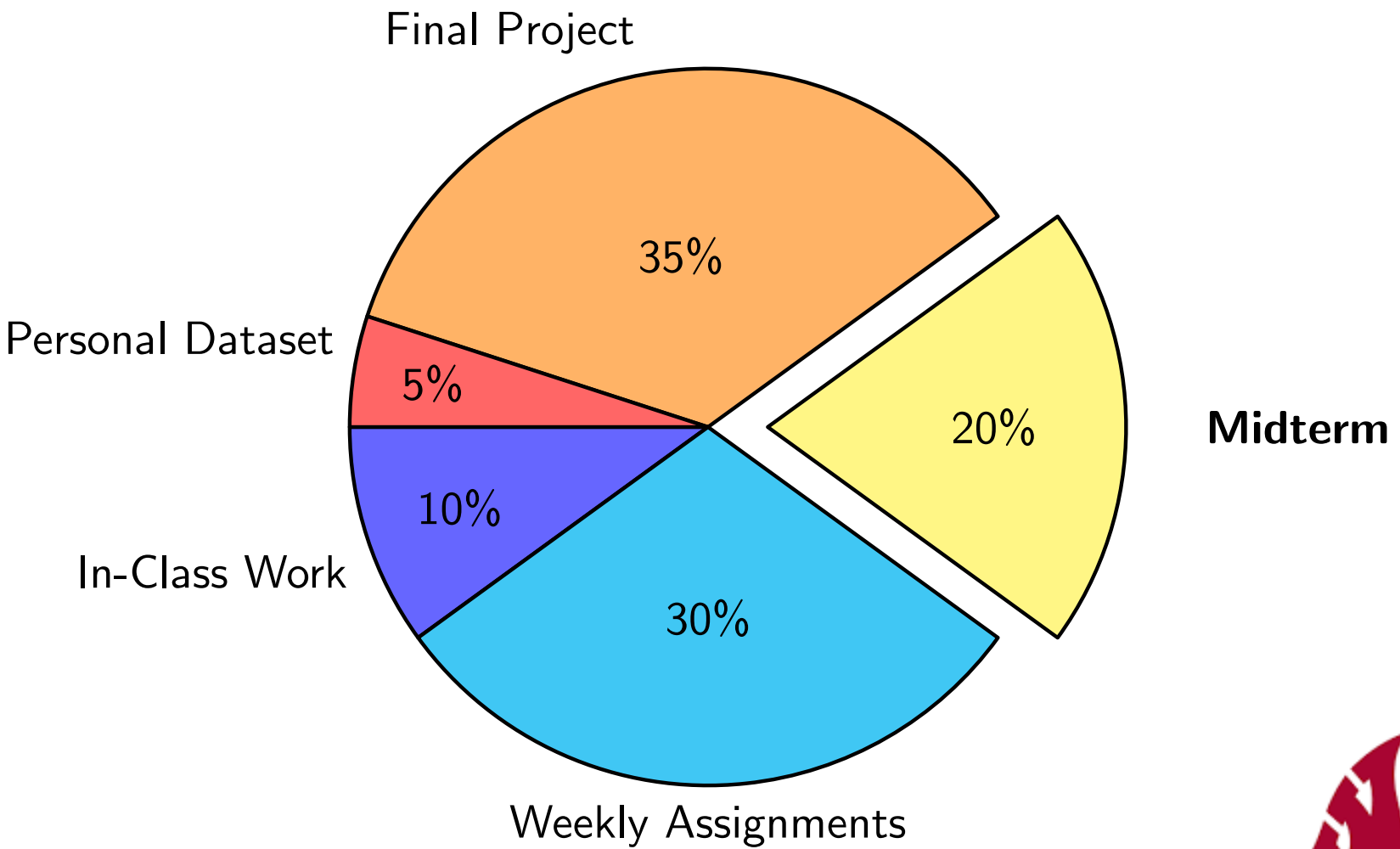
Assignments and Grades



Weekly Assignments

- Posted to Blackboard by Tuesday each week
- Written problems and data analysis tasks
- Due Wednesday of the next week at midnight
- No late homework accepted (lowest two assignments dropped)

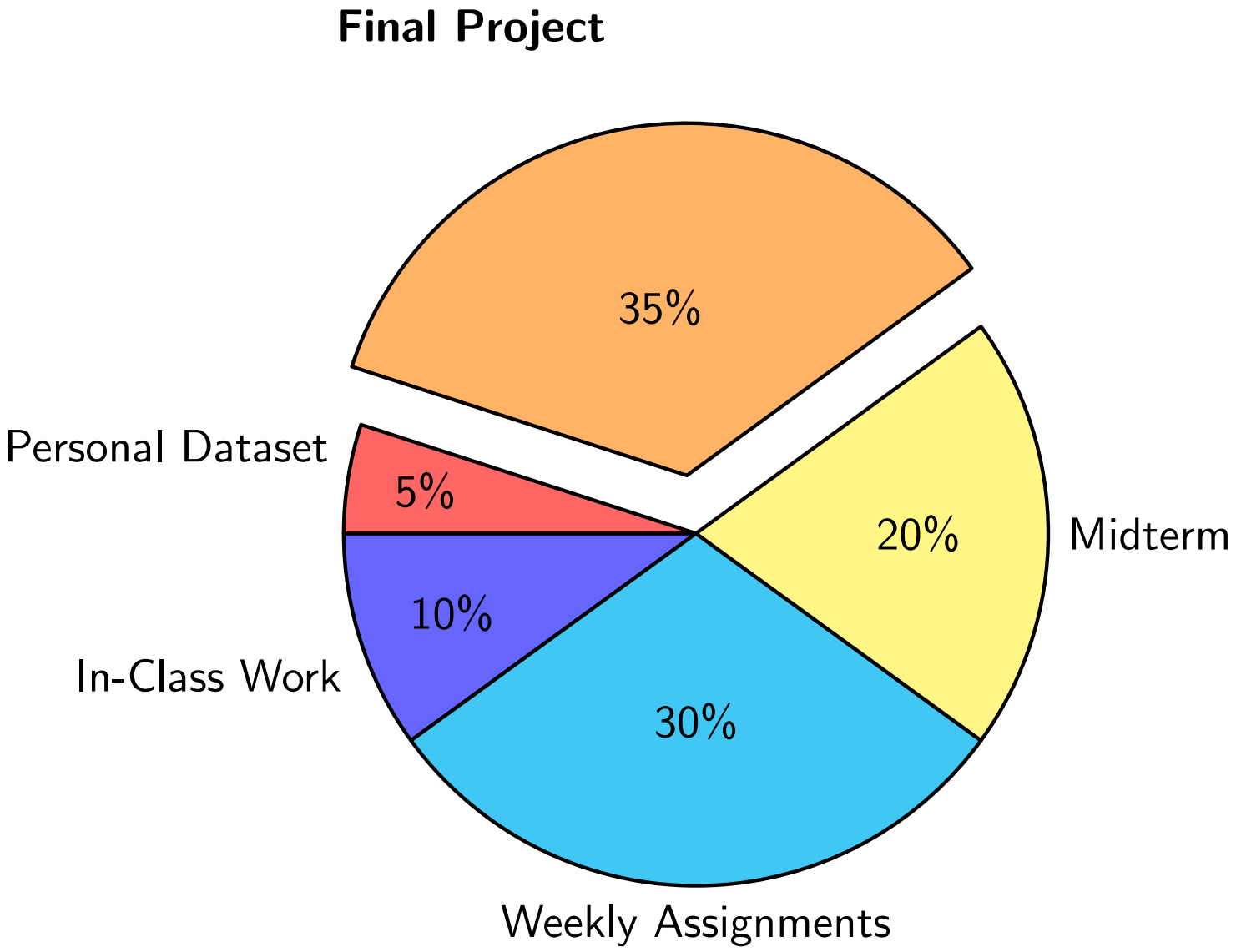
Assignments and Grades



Midterm Exam

- In-class on Thursday (10/15)
- Five 15-minute questions
- Covers material from first seven weeks of class

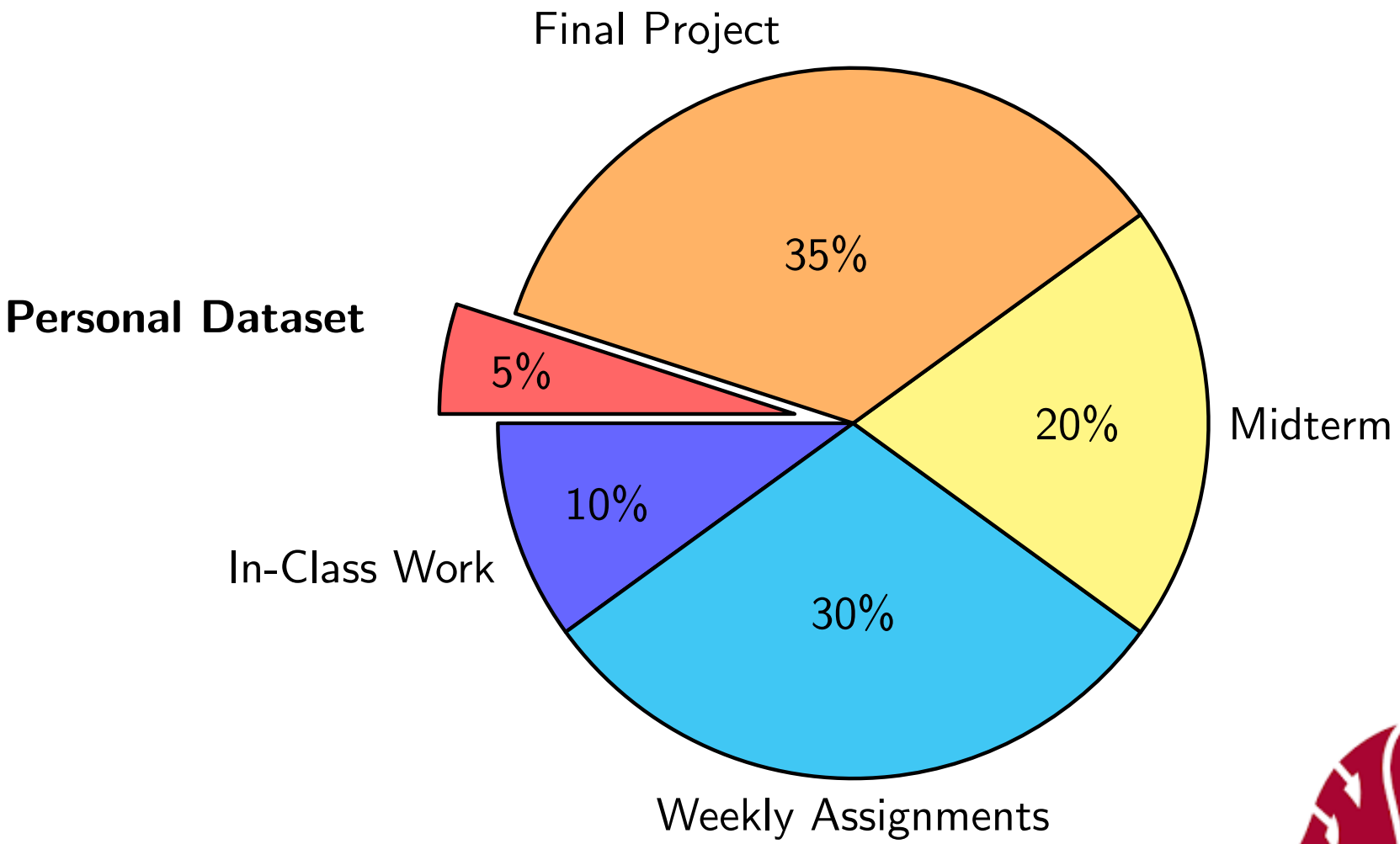
Assignments and Grades



Final Project

- Groups of four(ish)
- End-to-end analysis project on the topic of your choice
- More details week 12
- Writeup describing data collection, cleaning, and analysis
- 12-15 minute presentation

Assignments and Grades



Personal Dataset

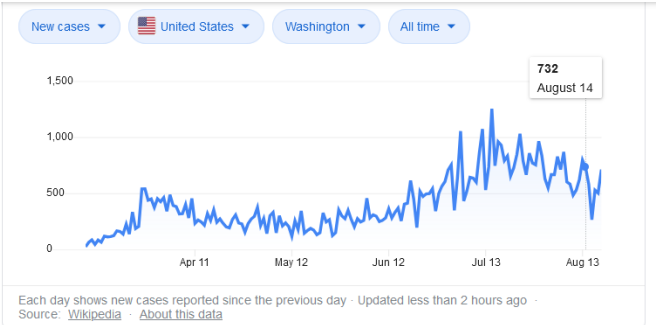
- Collect, prepare, and present a personally meaningful dataset
- Useful for later courses
- Also an excellent job interview topic

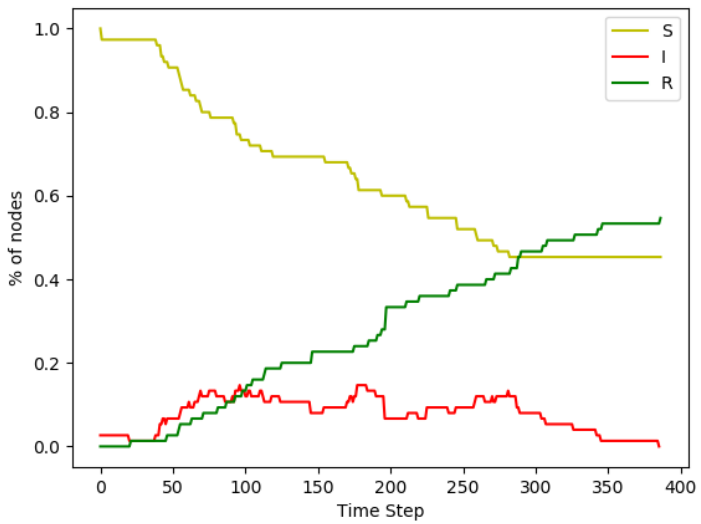
Keys to success

- Engage with the course and material
  - Attend class
  - Do the reading
  - Participate in discussions
  - Come to office hours
- Communicate effectively
  - Your written work should be complete and grammatical
  - Discussions should be respectful
- Professional behavior
  - Be resourceful
  - Follow the Academic Integrity Policy

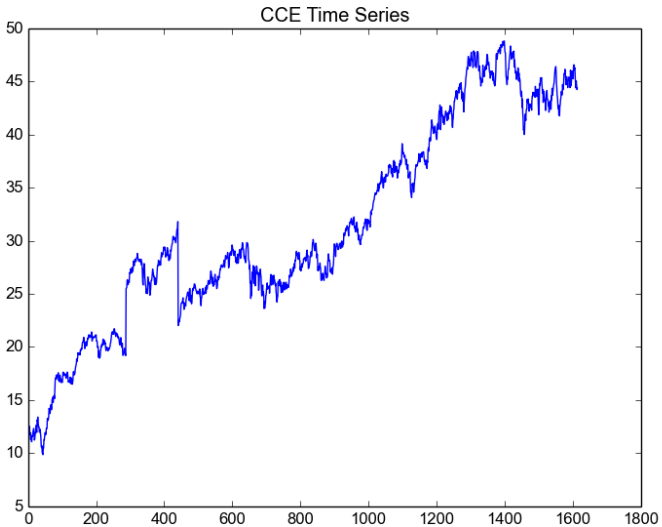
Big Data

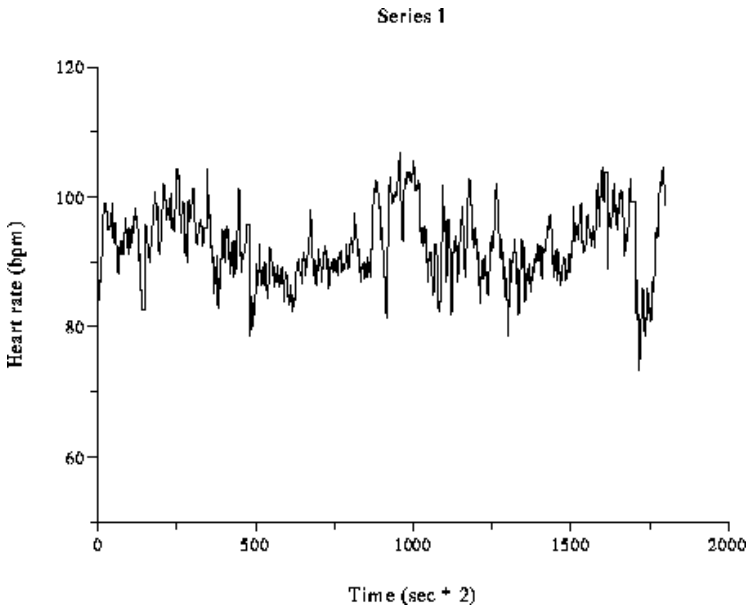
Epidemiology



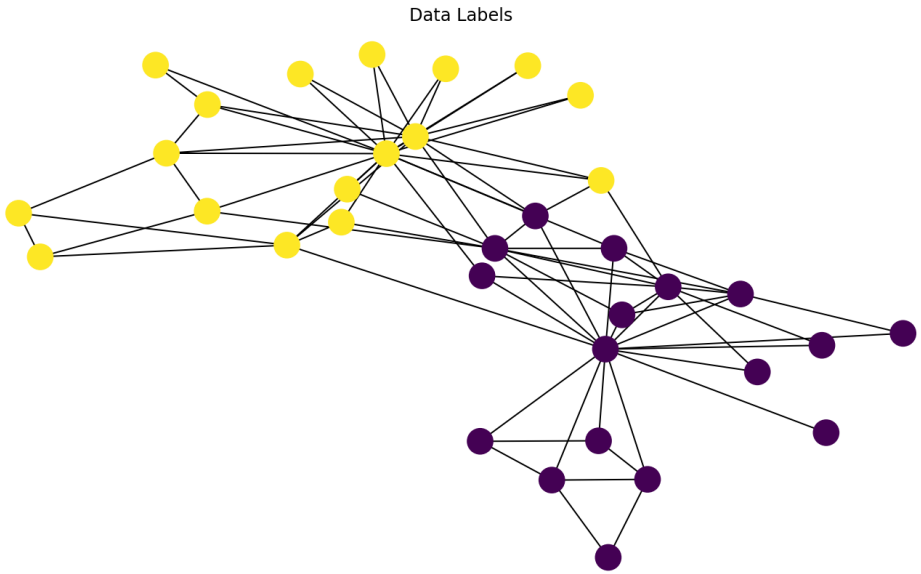


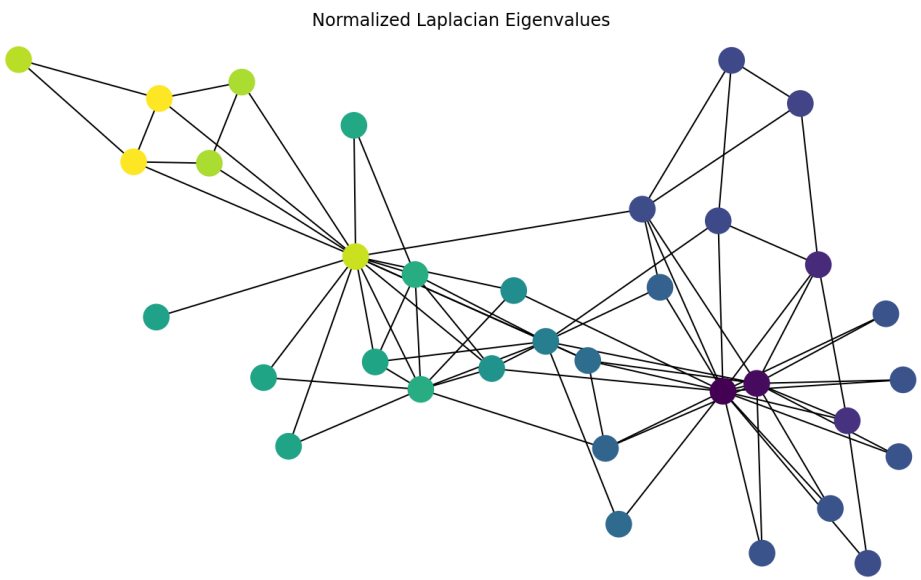
Time Series


CCE Time Series
50
45
40
35
30
25
20
15
10
5
0
200
400
600
800
1000
1200
1400
1600
1800


Series 1
120
100
80
60
Heart rate (bpm)
0
500
1000
1500
2000
Time (sec * 2)

Community Detection





Classification

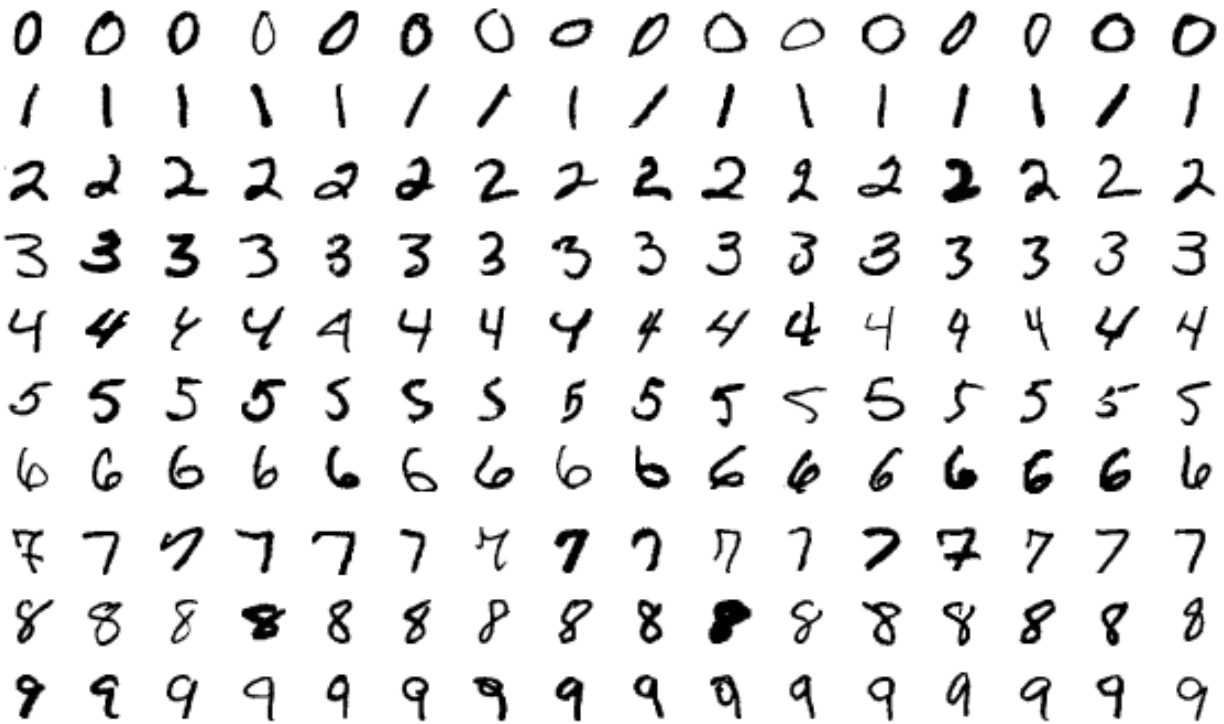
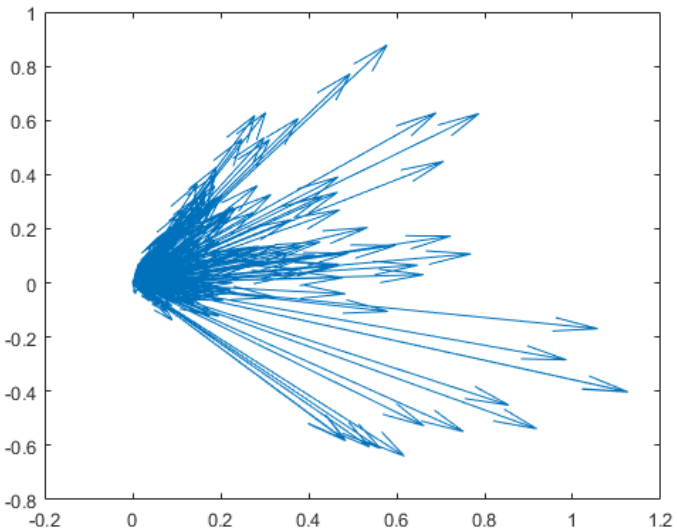
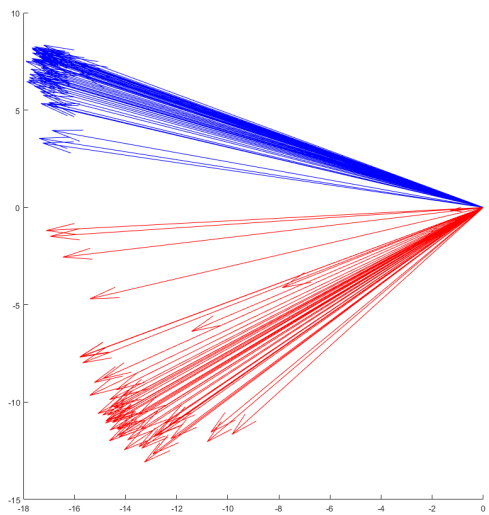


Figure Source Wikipedia: https://en.wikipedia.org/wiki/MNIST-database

Dimension Reduction





Recommendation Systems

'userid'	'01ec1a320ffded6b2dd47833f2c8e4fb',
'timestamp'	'2013-12-28',
'reviewsentences'	[[0, 'First, be aware that this book is not for ...,
'rating'	5,
'hasspoiler'	False,
'bookid'	'18398089',
'reviewid'	'4b3ffeaf14310ac6854f140188e191cd'

From: **Fine-grained spoiler detection from large-scale review corpora**, Mengting Wan, Rishabh Misra, Ndapa Nakashole, Julian McAuley ACL, 2019.

Common Features

- Storytelling
- Modeling/Abstraction
- Simplification
- Metrics

'The Process'

Raw Data

'Clean' Data

Exploratory Data Analysis

Algorithms, Models, and Analysis

Production Code

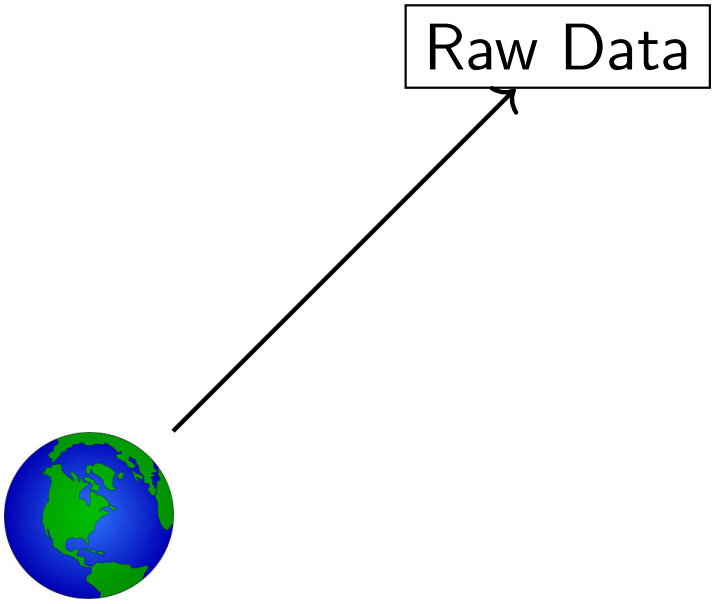
Vizualize and Communicate

Based on: Doing Data Science (O'Neil and Schutt)

Other Processes

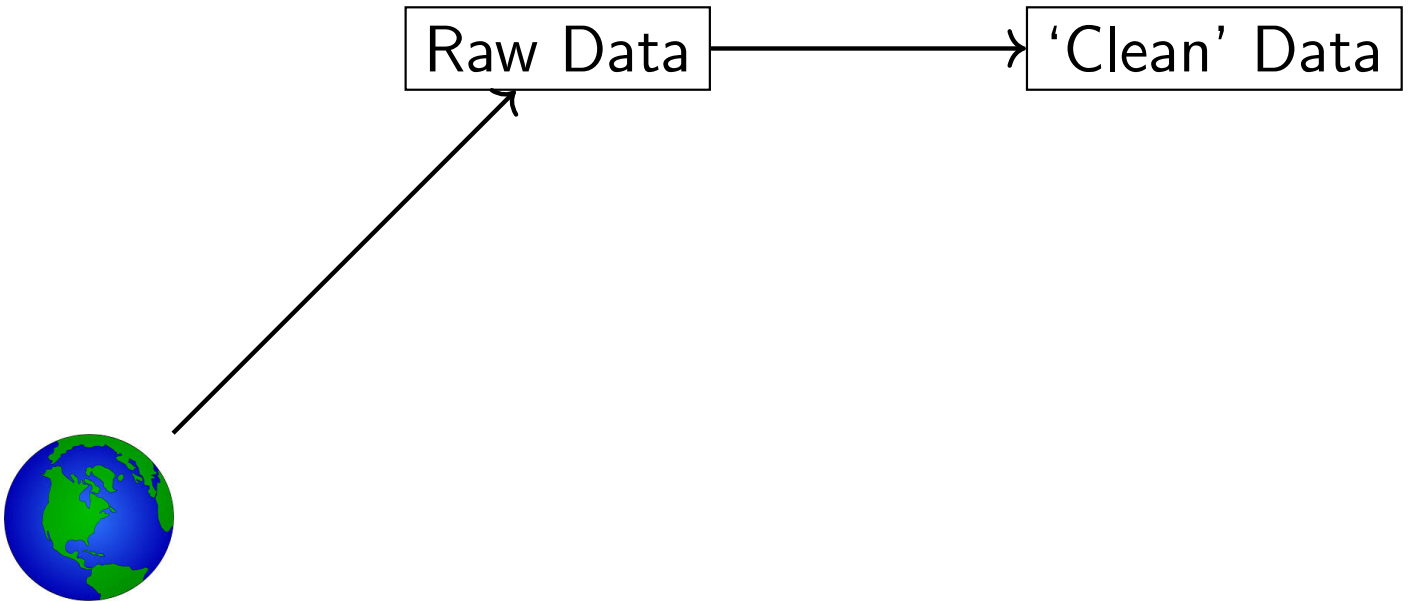
`https://lmgtfy.com/?q=the+data+science+process`

'The Process'



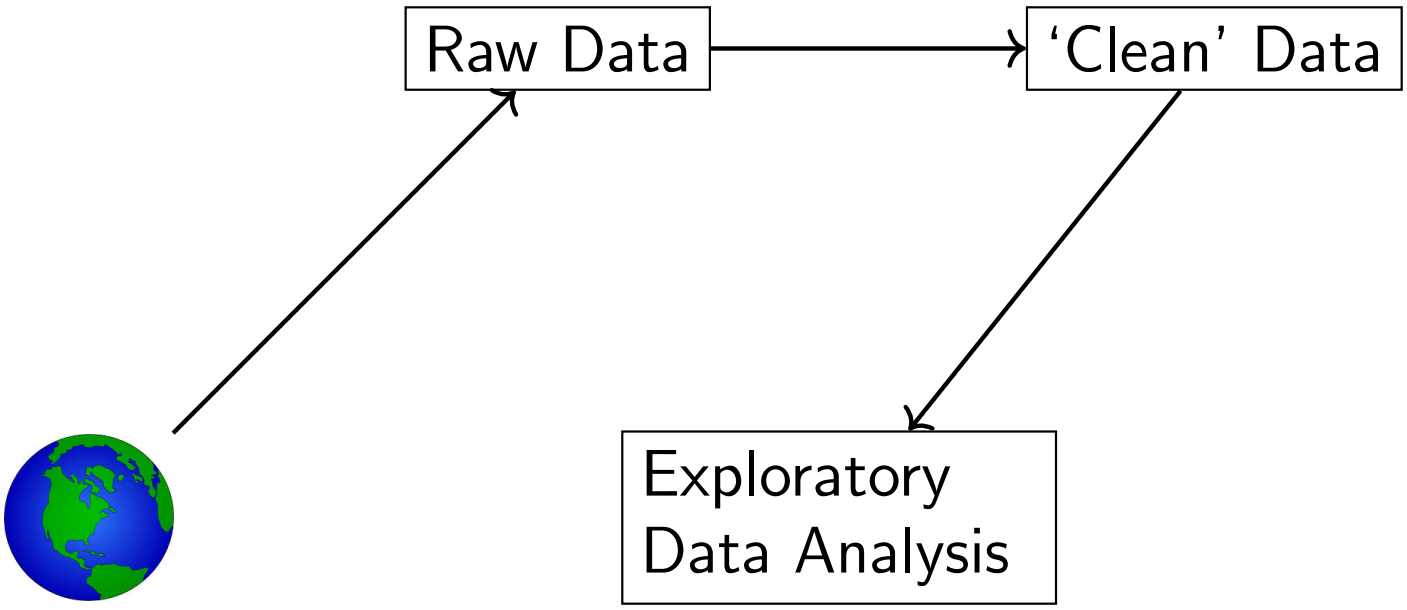
Based on: Doing Data Science (O'Neil and Schutt)

'The Process'



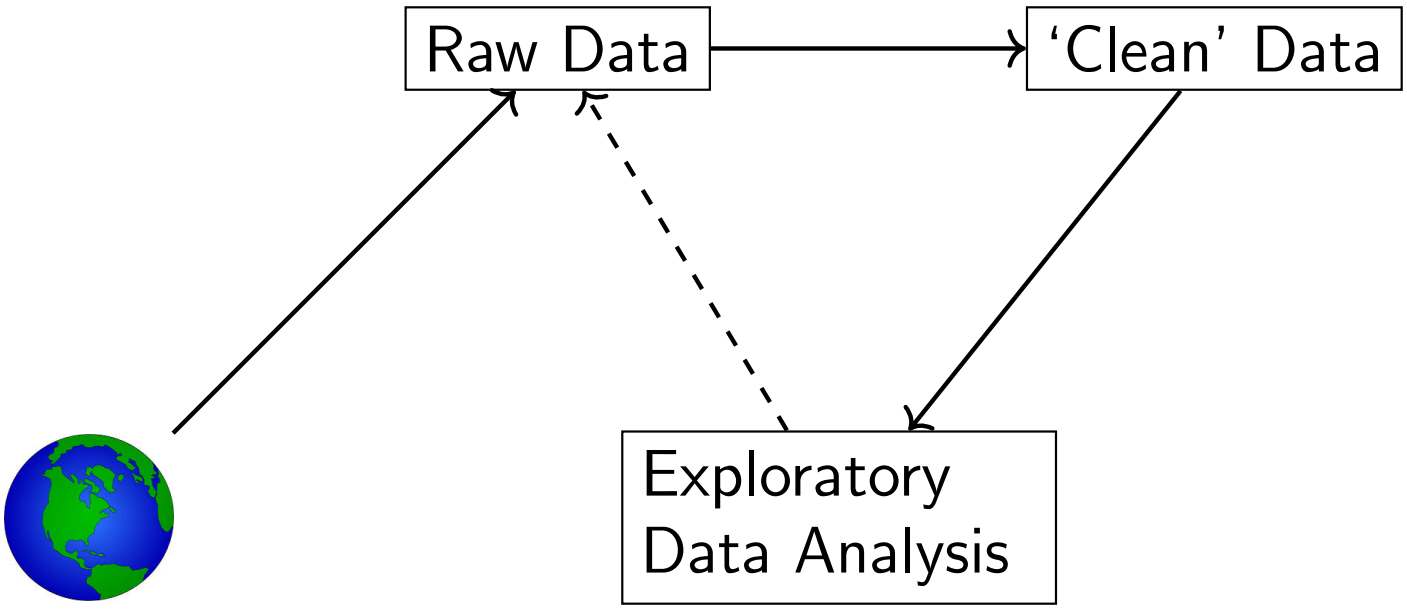
Based on: Doing Data Science (O'Neil and Schutt)

'The Process'



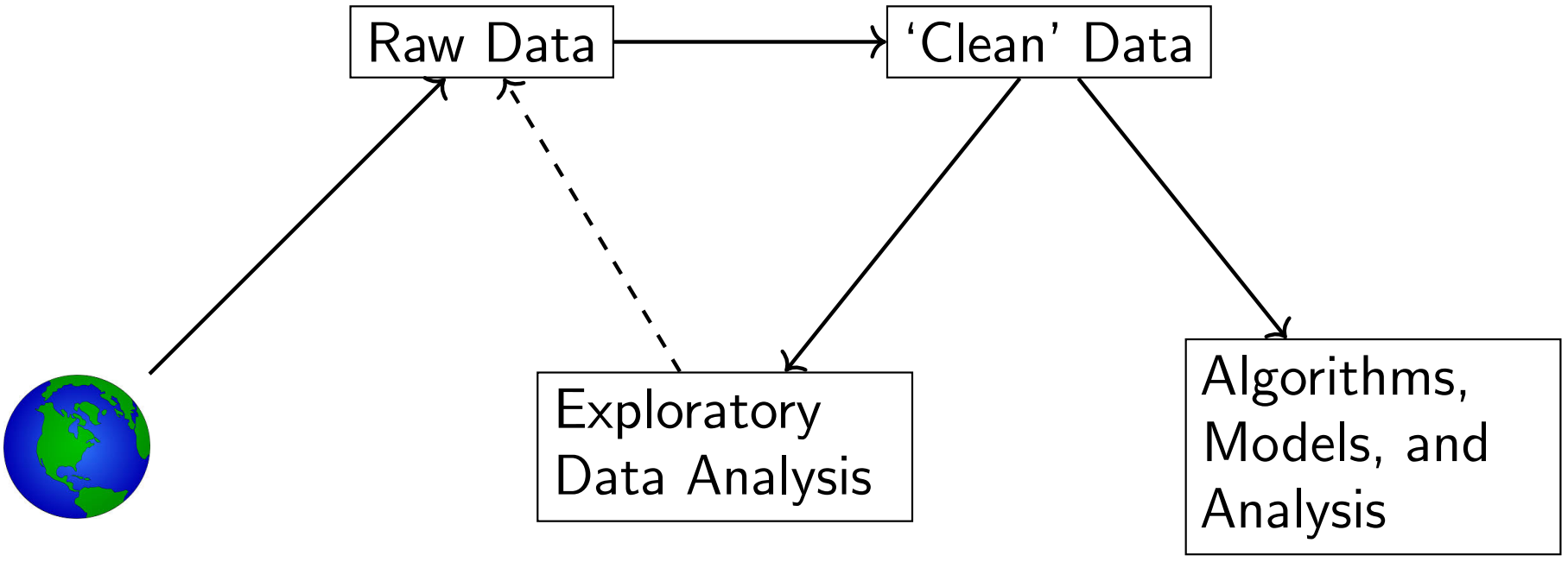
Based on: Doing Data Science (O'Neil and Schutt)

'The Process'



Based on: Doing Data Science (O'Neil and Schutt)

'The Process'



Based on: Doing Data Science (O'Neil and Schutt)

'The Process'

Raw Data

'Clean' Data

Exploratory Data Analysis

Algorithms, Models, and Analysis

Vizualize and Communicate

Based on: Doing Data Science (O'Neil and Schutt)

'The Process'

Raw Data

'Clean' Data

Exploratory Data Analysis

Algorithms, Models, and Analysis

Vizualize and Communicate

Based on: Doing Data Science (O'Neil and Schutt)

'The Process'

Raw Data

'Clean' Data

Exploratory Data Analysis

Algorithms, Models, and Analysis

Production Code

Vizualize and Communicate

Based on: Doing Data Science (O'Neil and Schutt)

'The Process'

Raw Data

'Clean' Data

Exploratory Data Analysis

Algorithms, Models, and Analysis

Production Code

Vizualize and Communicate

Based on: Doing Data Science (O'Neil and Schutt)