Introduction to Machine Learning Applications

Spring 2023

Minor Gordon

gordom6@rpi.edu

RENSSELAER POLYTECHNIC INSTITUTE • 1824 •

Rensselaer

Agenda

- About me
- About you (survey)
- Class logistics
- Class content
- Google Colab/Jupyter notebook demo

About me: Minor Gordon, PhD

- Grad research in distributed systems @ Berlin, Cambridge
- Consulting: social impact @ Asemio, data governance @ TopQuadrant
- Freelance: compilers, crowdsourcing, NLP, …
- R&D: High-performance computing @ NEC, IBM; AFRL, DARPA, IBM projects @ RPI
- Startups: Chartbeat, Birchbox, Onna; co-founded three marketing SaaS startups (two failed, one acquired)

About you

- https://forms.gle/FK189Mr7dnYvDvxG9

Logistics

When: Tuesdays and Fridays, 10 am to 11:20 am

Where: EATON 215

Web: Piazza (https://piazza.com/rpi/spring2023/mgmt496601656001) and Blackboard (submissions and grades)

Office hours: Tuesdays 12:30 to 2:00 pm over Webex

Email: gordom6@rpi.edu

Communications

- Announcements: Piazza
- Questions about the material and homework: Piazza
- Homework submission: Blackboard
- Questions about grades and other personal matters: Email
    - I will respond during the day.

Lecture notes and in-class exercises

- Lecture notes: will be posted on Piazza before lecture
- In-class exercises: will be posted on Piazza before the lecture

- **Participation in the class matters!**

Grading

MGMT 6560 (3 credits)

Component	Weight
Exams	40%
Assignments	30%
Project	15%
Quizzes	10%
Research Translation Exercise	5%

MGMT 4966 (4 credits)

Component	Weight
Exams	40%
Assignments	35%
Project	15%
Quizzes	10%

Tentative* A (93-100); A- (86-92); B+ (82-85); B (78-81); B- (74-77); C+ (70-73); C (66-69); C- (60-65); F (below 60)

Homework

- Homework and homework solutions will be posted on Piazza.
- It is OK to work in the same location as someone and ask questions. It is not OK to share code.
- You should produce everything that is submitted.
- Late submissions:
    - 10% per day for each late day
    - Please let me know if you are facing any difficulties with the material.

Quizzes

- Surprise quizzes through the semester
- Please arrive to class on time. You will receive a 0 if you arrive late.

After an evening lecture a man in the audience asked, “You say that Zen is everywhere. So why do we have to come to the Zen Center?”

“Zen is everywhere,” Suzuki Roshi agreed. “But for you, Zen is right here.”

From *Zen is Right Here: The Wisdom of Shunryu Suzuki*

Exams and project

- Three exams:
  - February 14th, 2023
  - March 21st, 2023
  - April 14th, 2023

- Project: initial report, final presentation and a final report.

Class content

Class goals

01

Prepare for advanced courses in analytics from across the RPI campus.

02

Enable you to gain skills necessary to begin careers as data scientists.

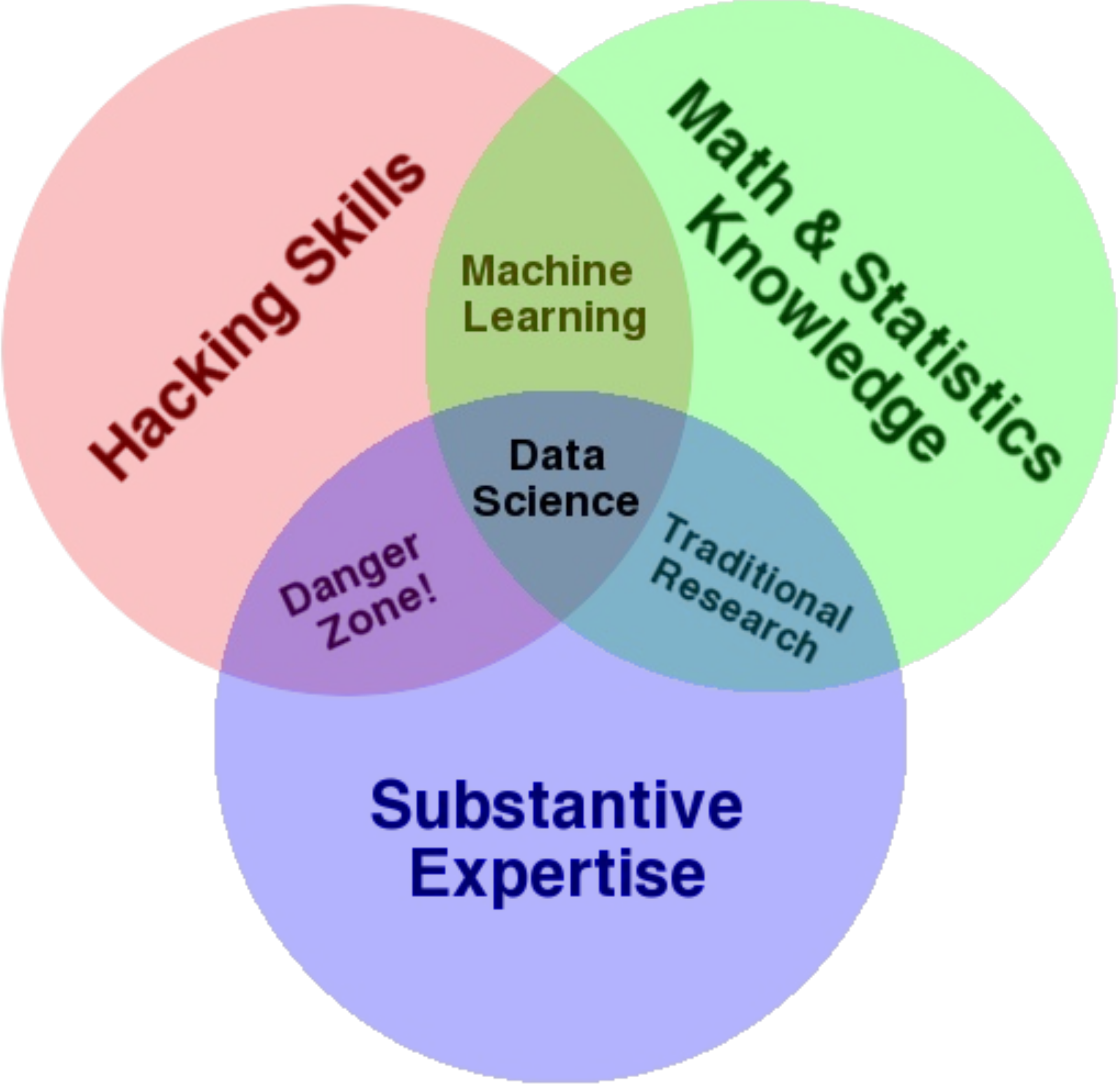
03

Empower you to apply analytics to solve real world problems.

Approach

- Emphasis on application rather than theory
  - Developing muscle memory
- Understand processes and objectives
  - No need to memorize every detail
- Select a project that’s meaningful to you
  - My interests: cultural heritage (museums, libraries), history (technology, fashion), flowers

Data Science



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

Some questions we can answer with data

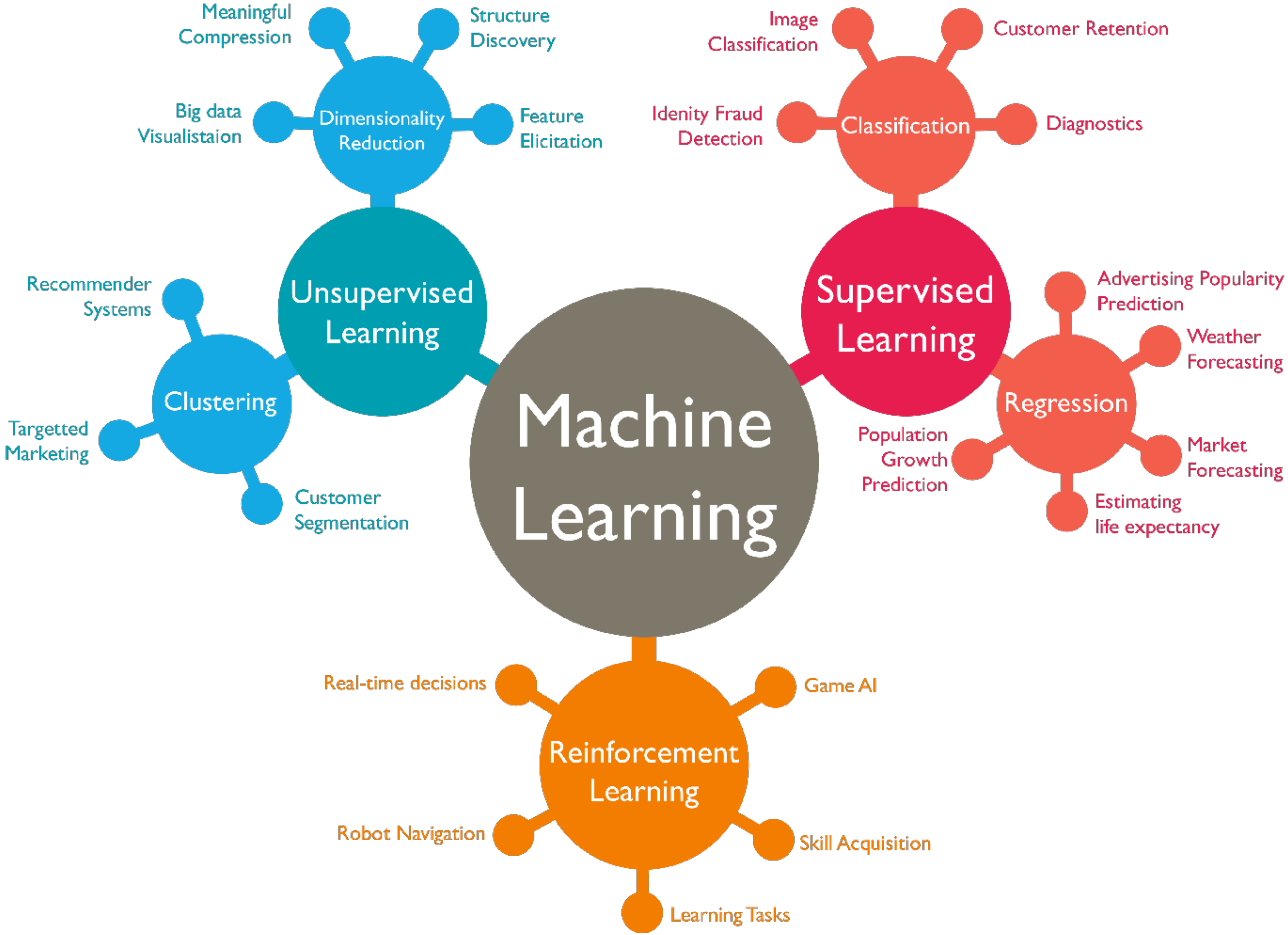
1. A/B Testing
   - If we propose new discounts, will it be better for our business?
2. Clustering patterns
   - What are the buying habits, age, location, etc., patterns of my customers?
3. Predictions
   - Will a customer purchase the product again within a few months of time?
4. Anomalies
   - Was there was a shift in the purchase pattern behavior of a customer?
5. Hypothesis testing
   - Is there a relationship between the purchase patterns of two customers buying similar products?

Concepts

- Models and their evaluation
- Features and dimensionality reduction
- Unsupervised learning
- Supervised learning
- Text and natural language processing
- Working with images
- Deep learning

Tools

- Python language
- Python data science libraries: NumPy, Pandas, Scikit-learn
- Python visualization: Matplotlib, Seaborn
- Google Colab and Jupyter Notebooks


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

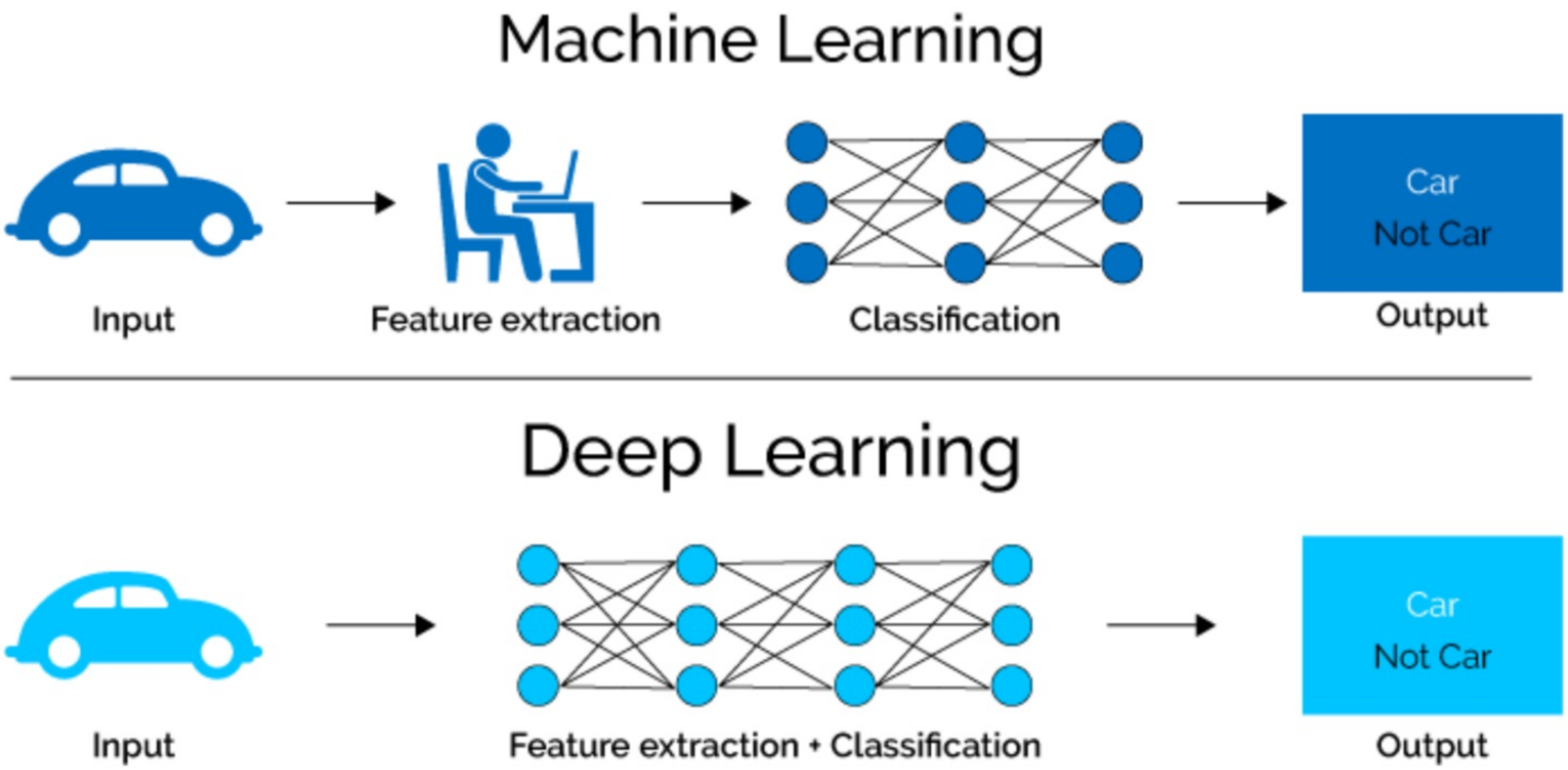
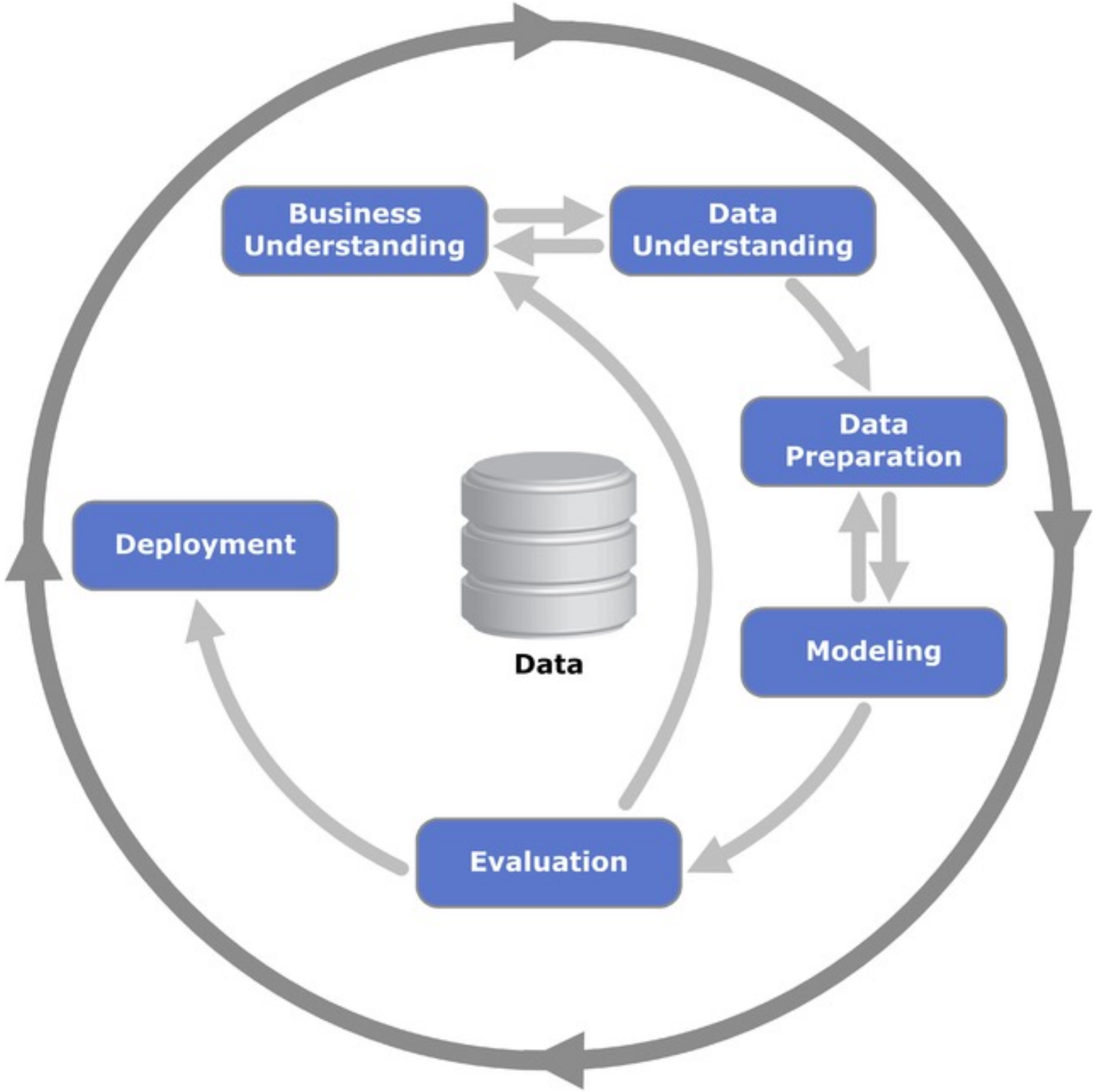


Figure 1: Machine Learning VS Deep Learning

Credits: towardsdatascience.com

Cross Industry Standard Process for Data Mining (CRISP-DM)

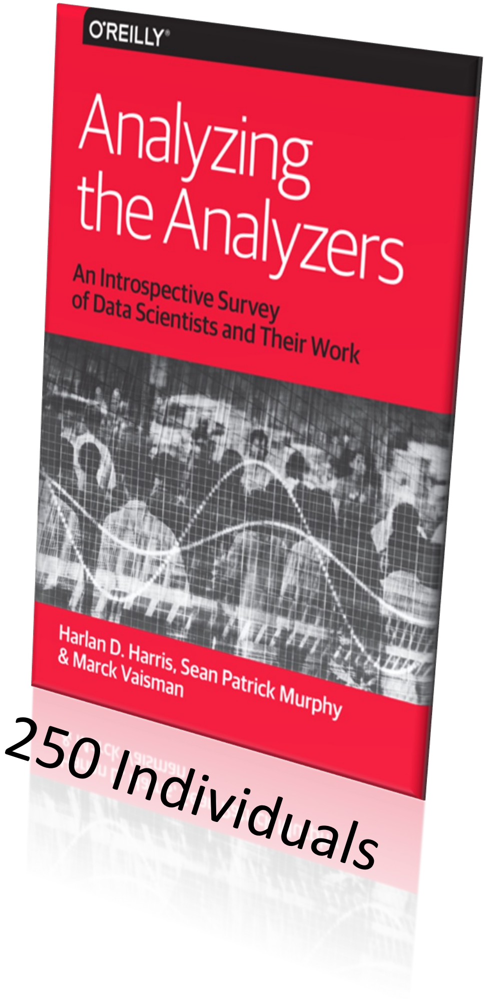


Credit [Kenneth Jensen (Own work) [CC BY-SA 3.0], via Wikimedia Commons]

O’Reilly Surveys

- Asked people involved in data science events to complete an online survey

983 Individuals



O'REILLY®

Analyzing the Analyzers

An Introspective Survey of Data Scientists and Their Work

Harlan D. Harris, Sean Patrick Murphy & Marck Vaisman

250 Individuals

Credit: DS100.org Class1 Presentation

What do they do?

How involved are you in task ___:
(a) Major, (b) Minor, (c) None

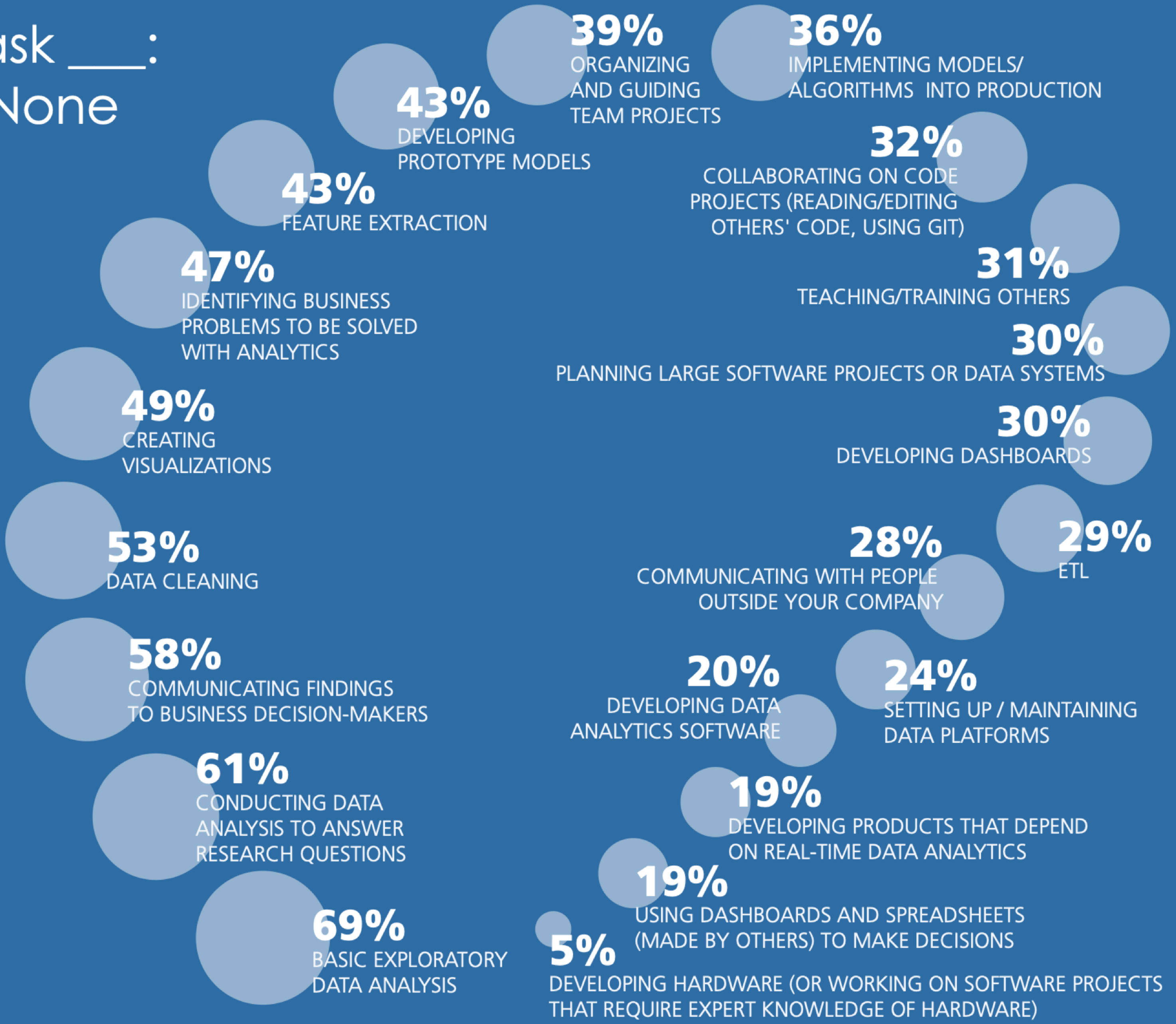
Developing Models
Implementing ML Algorithms
Visualization

Exploratory Data Analysis (EDA)
Researching Questions
Writing Reports,

…

Credit: DS100.org Class1 Presentation

How involved are you in task ___:
(a) Major, (b) Minor, (c) None



69% BASIC EXPLORATORY DATA ANALYSIS

61% CONDUCTING DATA ANALYSIS TO ANSWER RESEARCH QUESTIONS

58% COMMUNICATING FINDINGS TO BUSINESS DECISION-MAKERS

53% DATA CLEANING

49% CREATING VISUALIZATIONS

47% IDENTIFYING BUSINESS PROBLEMS TO BE SOLVED WITH ANALYTICS

43% FEATURE EXTRACTION

43% DEVELOPING PROTOTYPE MODELS

39% ORGANIZING AND GUIDING TEAM PROJECTS

36% IMPLEMENTING MODELS/ ALGORITHMS INTO PRODUCTION

32% COLLABORATING ON CODE PROJECTS (READING/EDITING OTHERS' CODE, USING GIT)

31% TEACHING/TRAINING OTHERS

30% PLANNING LARGE SOFTWARE PROJECTS OR DATA SYSTEMS

30% DEVELOPING DASHBOARDS

29% ETL

28% COMMUNICATING WITH PEOPLE OUTSIDE YOUR COMPANY

24% SETTING UP / MAINTAINING DATA PLATFORMS

20% DEVELOPING DATA ANALYTICS SOFTWARE

19% DEVELOPING PRODUCTS THAT DEPEND ON REAL-TIME DATA ANALYTICS

19% USING DASHBOARDS AND SPREADSHEETS (MADE BY OTHERS) TO MAKE DECISIONS

5% DEVELOPING HARDWARE (OR WORKING ON SOFTWARE PROJECTS THAT REQUIRE EXPERT KNOWLEDGE OF HARDWARE)

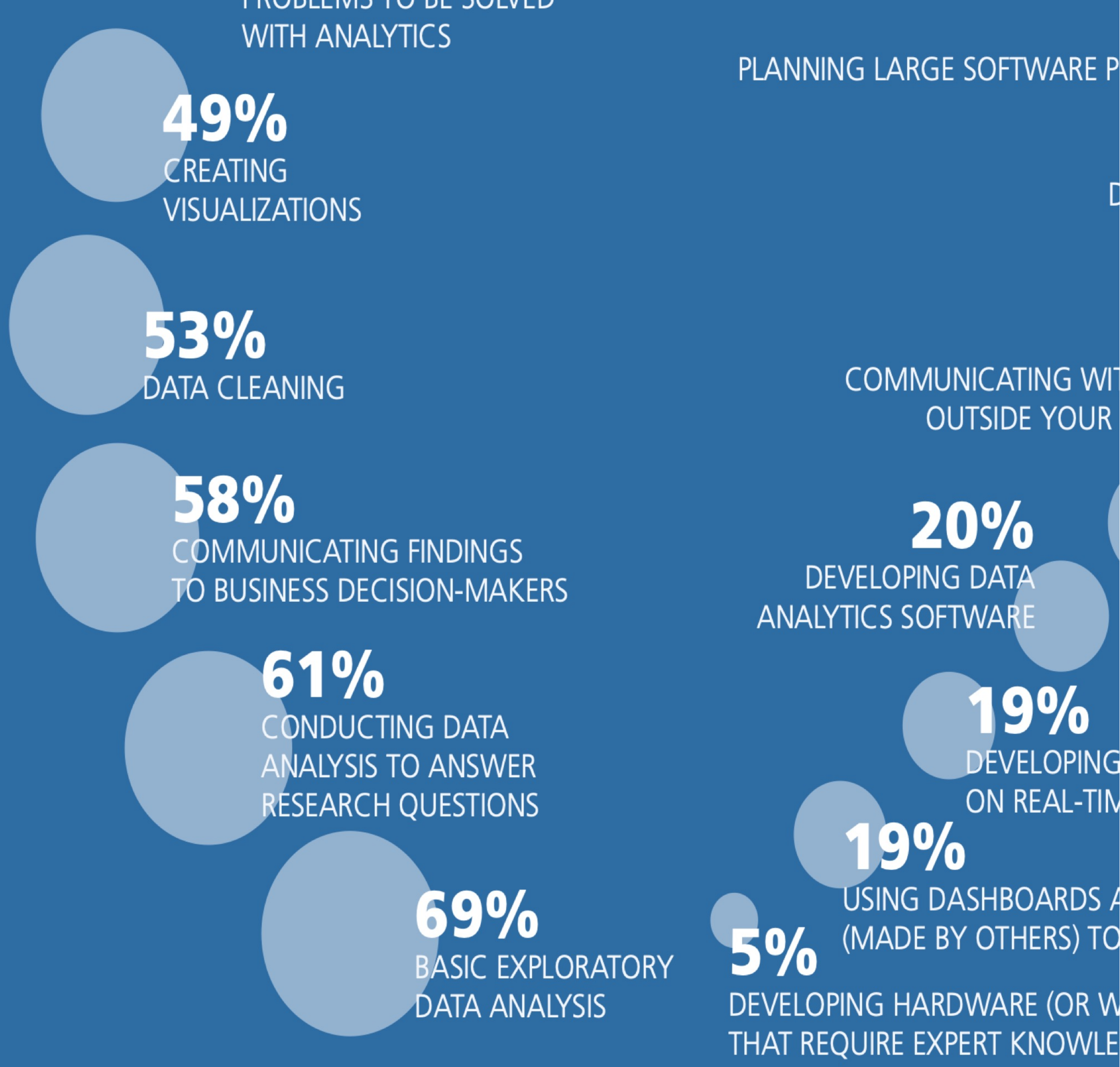
Credit: DS100.org Class1 Presentation

How involved are you in task ___:
(a) Major, (b) Minor, (c) None

Are the top items surprising?

Data Cleaning ☹

Where are Modeling / Prediction?

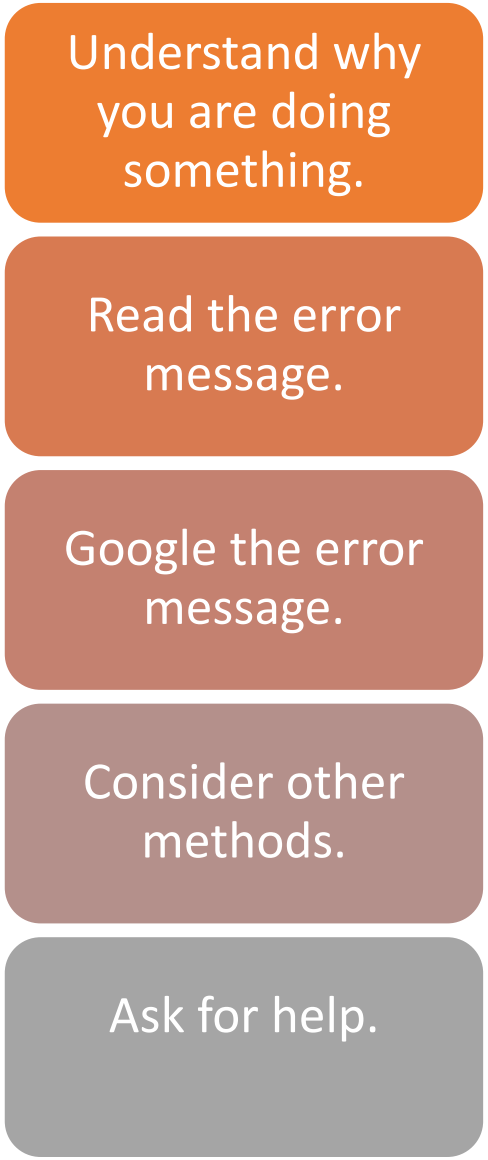


Credit: DS100.org Class1 Presentation

Working with real data

Real data has issues.

- Missing data, data at different scales, non-numeric data
- 80% of the work a data scientist does is collecting, cleaning and organizing data



Understand why you are doing something.

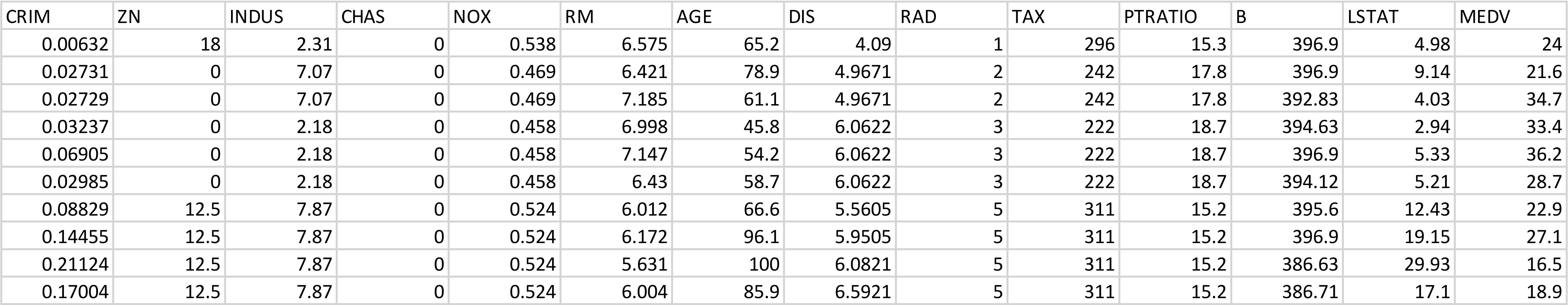
Read the error message.

Google the error message.

Consider other methods.

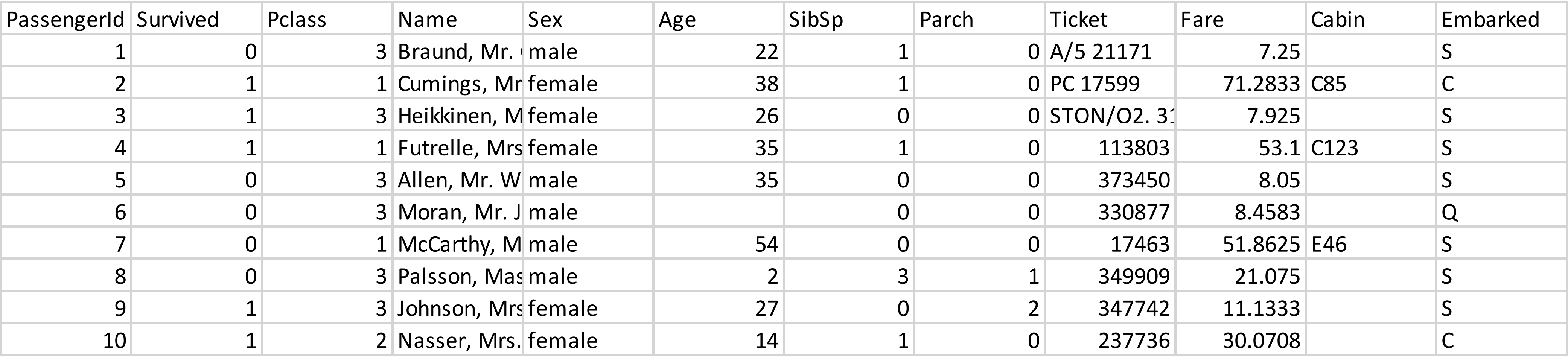
Ask for help.

Datasets: Boston housing



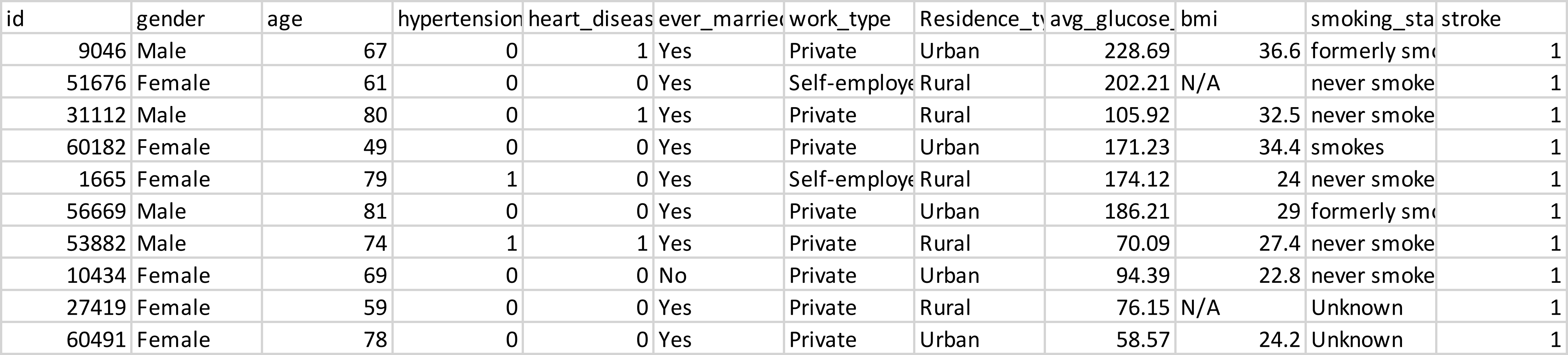
CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9

Datasets: Titanic survivors



PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. (	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mr	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, M	female	26	0	0	STON/O2. 31	7.925		S
4	1	1	Futrelle, Mrs	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. W	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. J	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, M	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Mas	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs.	female	14	1	0	237736	30.0708		C

Datasets: Stroke risk



id	gender	age	hypertension	heart_diseas	ever_marriec	work_type	Residence_ty	avg_glucose_	bmi	smoking_sta	stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smc	1
51676	Female	61	0	0	Yes	Self-employe	Rural	202.21	N/A	never smoke	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoke	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employe	Rural	174.12	24	never smoke	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smc	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoke	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoke	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1

Google Colab/Jupyter notebook demo

You can use either Google Colab or Jupyter notebooks installed on your local machine.