CS/ENGR M148 L1: What is Data Science?

Sandra Batista

Introduction

Instructor: Sandra Batista, sandra@cs.ucla.edu

Administrative Information

Meetings: TTh 4-5:50 pm MGYOUNG CS50

Office hours M 3-4 ENGR VI 282 and W 7-8 on Zoom link in BruinLearn.

A little bit about my professional trajectory:

A little less formal

Less formal still…

Introduction

What excites you most about CS M148?

Sharing how to make challenging material more simple and accessible and supporting others in their goals.

What are you doing this fall aside from teaching M148?
Teaching TA Seminar and working on computational genetics research

Please share something that has nothing to do work.
I love Nordic Walking, hot yoga, gardening, Heartfulness meditation (there is a group here at UCLA), and all things baby Yoda (and Disney+ Star Wars series).

What about you?

What's your name and pronouns?

Where excites you most about M148?

What are you doing this quarter aside from M148?

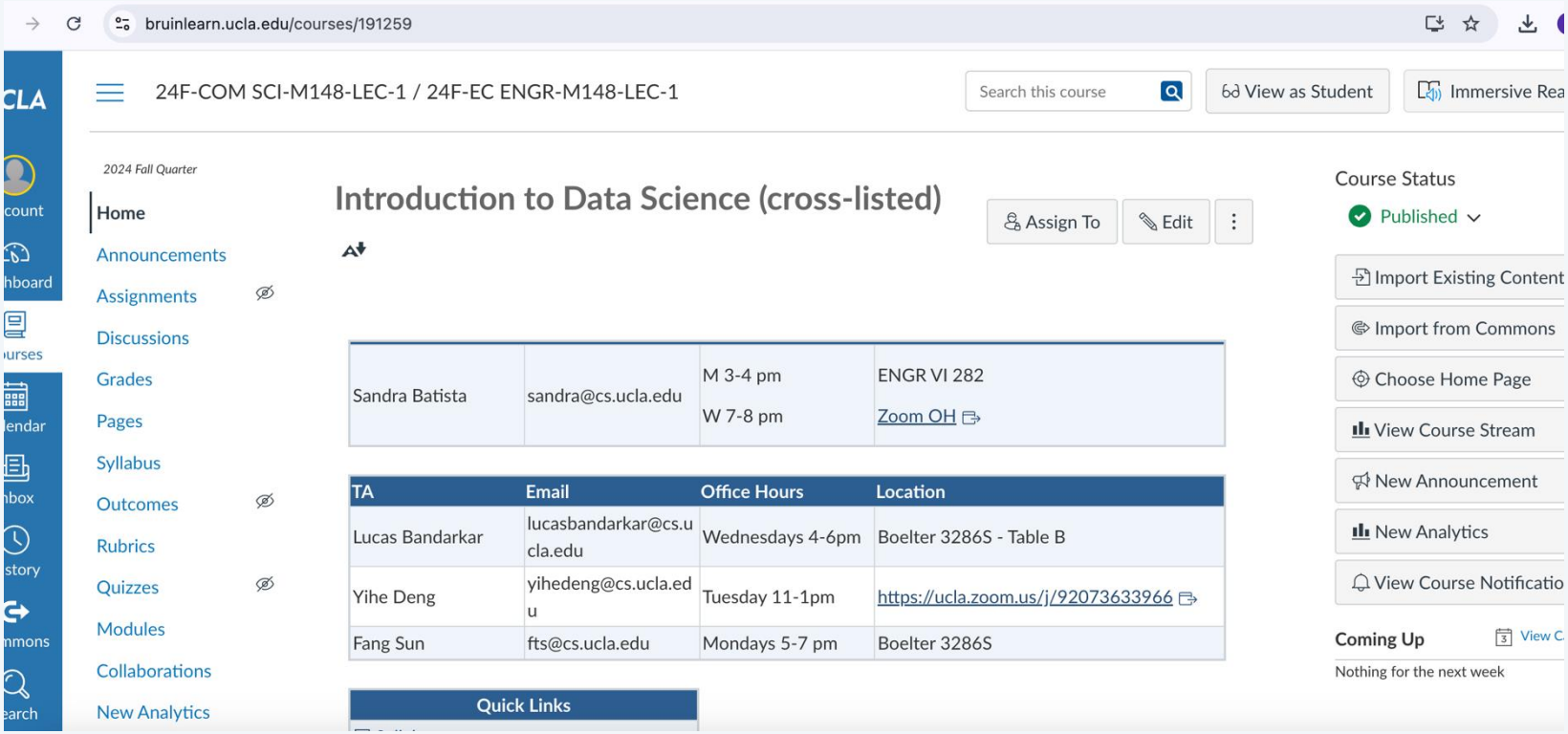
Please share one thing about you that has nothing to do with school or work.

Introduction

Course Administration

All course content will be on the course BruinLearn page:

https://bruinlearn.ucla.edu/courses/191259



Course Policies

The syllabus is on BruinLearn.

Please take a moment to access and review it now.

Grading:

Midterm 20%, November 5

Final 20%, December 10

Quizzes 20%, 4x 5%

Problem Sets 20%, 4 x 5%

Team Project 20%

Course Policies

Collaboration Policy:

You may collaborate on ungraded labs and team projects. You may not collaborate on problems sets, quizzes or exams.

Generative AI usage Policy:

You may use generative AI tools on ungraded labs and team projects. You may not use generative AI tools on problems sets, quizzes or exams.

Late Work Policy: No late work accepted. However, we will work to create fair accommodations for extenuating circumstances.

Assignments

1. Problem Sets 1 and 3 will focus on coding exercises. Problem sets 2 and 4 will be written exercises.
2. Labs are not graded.
3. Quizzes will be quick multiple choice and T/F quizzes during class.
4. The final and midterm will be more traditional exams with open-ended questions. *Exams are open-course materials.*
5. The team projects are your opportunity to explore a domain and problem of your choice

Projects

1. Projects will be graded on how well they demonstrate mastery of the methods taught in class and discussions.
2. You may choose your own data set or a data set supported by the course staff.
3. Team contract 5%
4. Project discussion check-ins: 30%, 6x5%
5. Final project code: 25%
6. Final project report: 40%

Course Overview

Course Outline

Weeks in syllabus

Tentative Schedule

Week	Day	Topic	Assignments
0	9/26	Python Essentials	Lab 0: Intro to Python
	9/27	Discussion: Lab 0	numpy, graphing, Jupyter Notebooks
1	10/1	Introduction	Lab 1: sklearn, matplotlib
	10/3	Data Collection and Cleaning	
	10/4	Discussion: Python and Project Teams	Team contracts Due
2	10/8	Linear Regression	Lab 2 : Scikit-learn Linear Regression
	10/10	Multiple and Polynomial Regression	
	10/11	Discussion: Lab 2 and Project Data	Project Data Check-in Due, PS1 posted
3	10/15	Model Selection and Cross Validation (CV)	Lab 3: Regression and CV
	10/17	Regularization and Hypothesis Testing	
	10/18	Discussion: Lab 3 and Project Regression	Project Regression Check-in Due
4	10/22	Logistic Regression (LR)	Lab 4: Logistic Regression, PS1 due, PS2 posted
	10/24	Classification	**PS1 Quiz in class**
	10/25	Discussion: Lab 4	Project LR Check-in due
5	10/29	Decision Trees and Random Forests	Lab 5: KNN Classification, PS2 due
	10/31	Midterm Review	**PS2 Quiz in class**
	11/1	Discussion: Lab 5, review	Project Classification Check-in Due

Course Outline

Weeks in syllabus

5	10/29 10/31 11/1	Decision Trees and Random Forests Midterm Review Discussion: Lab 5, review	Lab 5: KNN Classification, PS2 due **PS2 Quiz in class** Project Classification Check-in Due
6	11/5 11/7 11/8	**Midterm Exam** Dimensionality Reduction Discussion: Lab 6	 Lab 6: PCA PS3 posted
7	11/12 11/14 11/15	Clustering Hidden Markov Models and Expectation Maximization Discussion: Unsupervised Learning (UL)	 Project UL Check-in due
8	11/19 11/21 11/22	Neural Networks (NN) NN and Backpropagation Discussion: NN	Lab 7: NN, PS3 due **PS3 Quiz in class** PS4 posted
9	11/26 11/28	NN regularization and interpretation **No class: Thanksgiving Holiday**	
10	12/3 12/5 12/6	Explainability and Interpretability Final Exam Review Discussion: Project NN	PS4 due **PS4 Quiz in class** Project NN Check-In Due
	12/10	**Final exam 3-6 pm**	
	12/13	**Final projects due by 6 pm**	

Today's Learning Objectives

Students will be able to:

- ✓ Define **data science vs. data analytics**
- ✗ Define **veridical data science**
- ✗ Identify stages of the **data science life cycle**

What is data science?

According to Chirag Shah…

Professor in Information School,

University of Washington in Seattle

Shah, C. (2020). *A hands-on introduction to data science*. Cambridge University Press.

Datum, data, and science

Data = plural of datum, but we will use 'data' for both singular and plural versions

Information = meaningful data

Science = systematic study of the structure and behavior of the physical and natural world through observation and experimentation (Oxford English Dictionary)

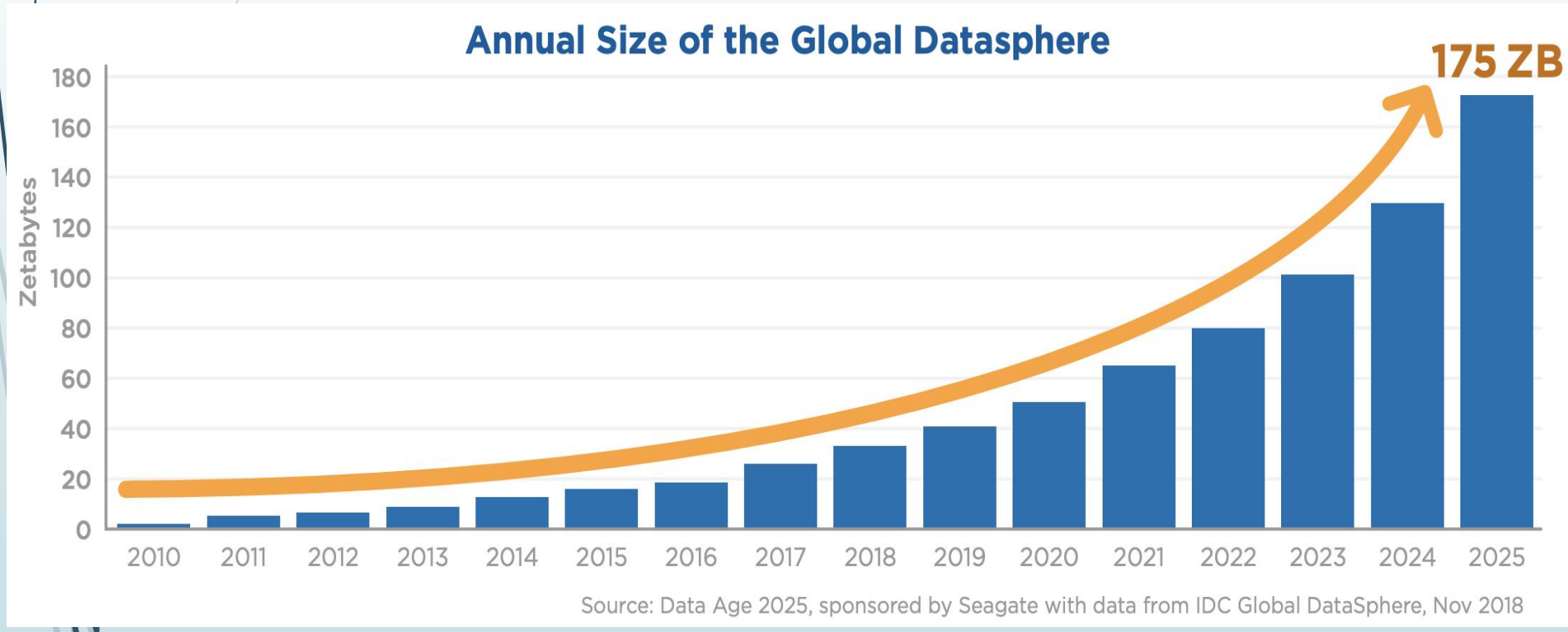
What is Data Science?

A field of study and practice that involves collection, storage, and processing of data in order to derive important insights into a problem or a phenomenon.

Data may be generated by **humans** (surveys, logs, etc.) or **machines** (weather data, road vision, etc.).

Data may be in different **formats** (text, audio, video, augmented or virtual reality, etc.).

Let's talk about data!



3V model of data

Velocity: The speed in which data is accumulated increased

Volume: The size and scope of the data increased

Variety: The massive array of data and types (structured and unstructured) increased

Data Science applications in different domains

- Finance
- Public policy
- Politics
- Healthcare
- Urban planning
- Education
- Libraries
- …

Relation of Data Science with other fields

Statistics

Computer Science

Engineering

Business Analytics

(Computational) Social Science

Data science is at intersection of CS, Statistics and ML

What skills do we need for Data Science?

Willing to experiment

Proficiency in mathematical reasoning

Data literacy

Computational Thinking!! Abstraction, programming, running analyses…

Hands-on exercise

OBSERVATION	HEIGHT (INCHES)	WEIGHT (LBS)
1	58	115
2	59	117
3	60	120
4	61	123
5	62	126
6	63	129
7	64	132
8	65	135
9	66	139
10	67	142
11	68	146
12	69	150
13	70	154
14	71	159
15	72	164

1. How much increase can we expect in weight with an increase of one inch in height?
2. What would you expect the weight to be of a person who is 57" tall?
3. What would you expect the weight of someone who is 73" tall?

https://colab.research.google.com/drive/142d3t82zlAQg9IeXKKYd5E7C8sJJeUsi?usp=sharing

Summary

- Data science is the field of study and practice that involves collection, storage, and processing of data in order to derive important insights into a problem or a phenomenon.
- It calls for computational thinking and understanding of statistics.
- Any tool that can represent and process data can be useful for doing data science, including spreadsheet programs (e.g., Excel), databases (e.g., MySQL), Python, R, and even the back of an envelop!

But is this all there is?

Data analysis vs. data analytics

Data analysis = hands-on data exploration and evaluation

Data analytics = broader term that includes data analysis as a necessary part

Analysis – typically looks at what happened in order to explain

Analytics – models the future or predicts a result

Types of analysis/ analytics (that we care about here)

- Descriptive
- Diagnostic
- Predictive
- Prescriptive
- Exploratory
- Mechanistic

1. Descriptive analysis

Typically it is the first kind of data analysis performed on a data set.

Usually it is applied to large volumes of data, such as census data.

Description and interpretation processes are different steps.

Core concepts/techniques :

Frequency distribution: histogram

Measures of centrality: mean, median, mode

Dispersion of distribution: range, interquartile range, variance,

2. Diagnostic analytics

Used for discovery, or to determine why something happened

Most commonly used technique: correlation

3. Predictive analytics

Obtain *insight* from *hindsight* as we identify patterns from existing data

Then use such *insights* to predict the future: *foresight*

4. Prescriptive analytics

An area of business analytics dedicated to finding the best course of action for a given situation

Typical steps:

Start by first analyzing the situation (using descriptive analysis)	Move toward finding connections among various parameters/variables, and their relation to each other	Use this to address a specific problem, more likely that of prediction

5. Exploratory analysis

An approach to analyzing data sets to find previously unknown relationships

Often involves visualization techniques

Useful when we lack a clear question or a hypothesis

6. Mechanistic analysis

Involves understanding the exact changes in variables that lead to changes in other variables for individual objects

Most common (and powerful) technique: regression

Regression analysis is a process for estimating relationships among variables: typically from predictors to outcome

Summary

- **Data analysis**: hands-on data exploration and evaluation. Analysis looks backwards, providing marketers with a historical view of what has happened.
- **Data analytics**: defines the science behind the analysis. The science means understanding the cognitive processes an analyst uses to understand problems and explore data in meaningful ways. It is used to models the future or predicts a result.
- Analysis/analytics types
  - Descriptive analysis
  - Diagnostic analytics
  - Predictive analytics
  - Prescriptive analytics
  - Exploratory analysis
  - Mechanistic analysis

Today's Learning Objectives

Students will be able to:

- ✓ Define **data science vs. data analytics**
- ✗ Define **veridical data science**
- ✗ Identify stages of the **data science life cycle**

Veridical Data Science

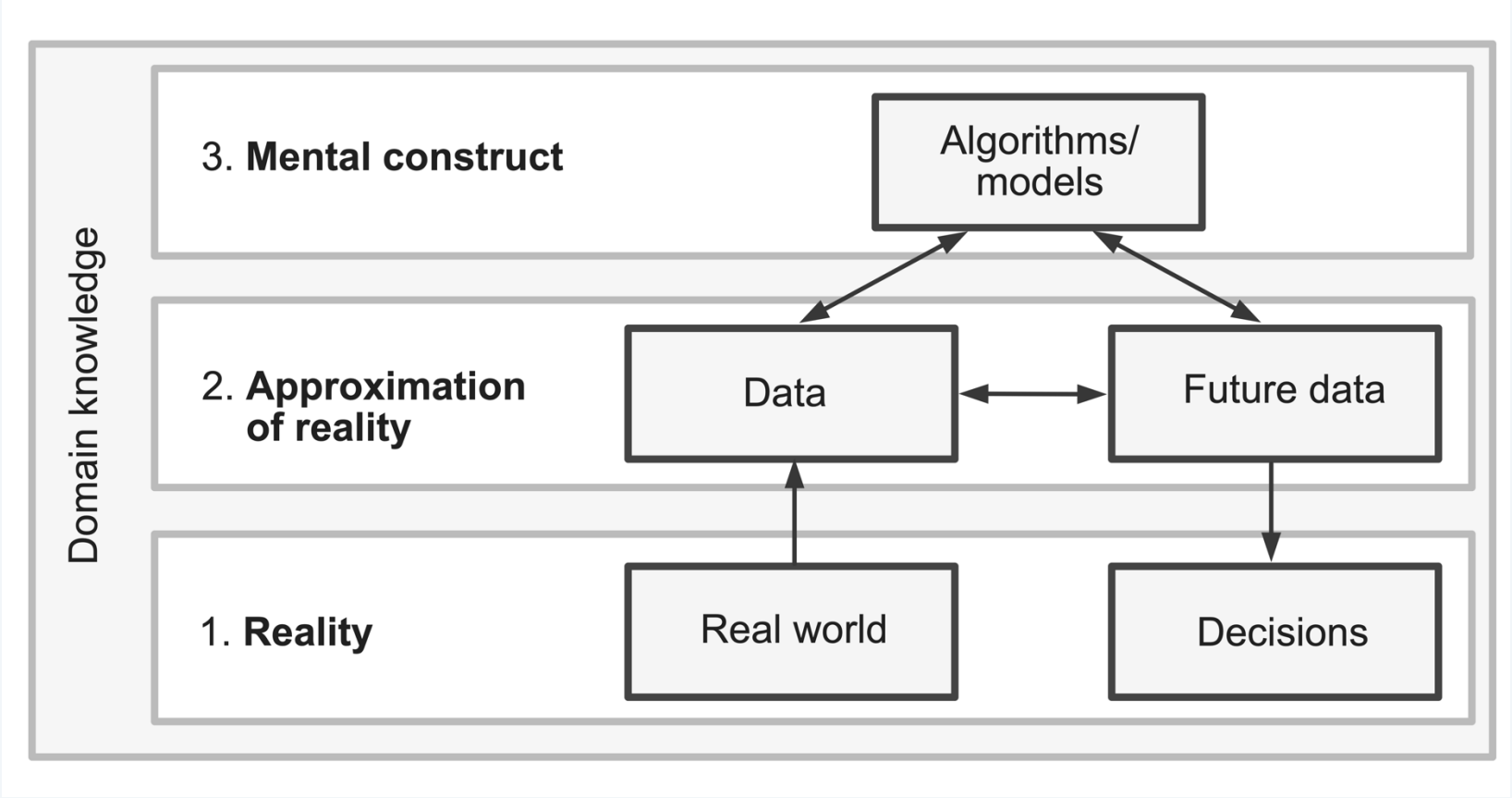
Veridical data science is the practice of conducting data analysis while making human judgment calls and using domain knowledge to extract and communicate useful and **trustworthy** information from data to solve a real-world domain problem

-Bin Yu, Professor of Statistics

University of California, Berkeley

Yu, B., & Barter, R. L. (2024). Veridical data science: The practice of responsible data analysis and decision making. The MIT Press.

How do we use algorithms in data science?



Critical Thinking and Domain Knowledge are Essential!

- Ask questions about the **domain problem**
- Ask questions about the **data collection**
- Ask questions about the **analysis, algorithms, and results**

- **It is always fair to ask if a step in a data science project is ethical.**

- Potential pitfall: **confirmation bias** – finding what you expect to find in the analysis
- Potential pitfall: **desirability bias** – finding what others expect to find in the analysis

Critical Thinking and Domain Knowledge are Essential!

- Ask questions about the **domain problem**
- Ask questions about the **data collection**
- Ask questions about the **analysis, algorithms, and results**

- **It is always fair to ask if a step in a data science project is ethical.**

- Potential pitfall: **confirmation bias** – finding what you expect to find in the analysis
- Potential pitfall: **desirability bias** – finding what others expect to find in the analysis

2019 Australian Bushfires Claim

Claim from news article:

"Professor Chris Dickman has revised his estimate of the number of animals killed in bushfires in NSW to more than 800 million animals, with a national impact of more than one billion animals."

Do you believe this claim? What will convince you?

https://www.sydney.edu.au/news-opinion/news/2020/01/08/australian-bushfires-more-than-one-billion-animals-impacted.html

2019 Australian Bushfires Claim

Skepticism:

Data was used from 2007 study to estimate animal Population densities in small range of NSW and then multiplied to scale for regions affected by the fires.

Johnson, C, H Cogger, Christopher Dickman, and H Ford. 2007. *Impacts of Landclearing: The Impacts of the Approved Clearing of Native Vegetation on Australian Wildlife in New South Wales*. World Wide Fund for Nature.

How do we conduct trustworthy data science

- **Predictability**: when results confirmed in future data or domain knowledge
- **Computability**: Using computation to successfully complete analyses
- **Stability**: Results remain consistent with data perturbations and algorithmic choices

- **PCS Framework** combines traditional statistical inference that measures how results change with perturbation to data generation process with modern machine learning that aims to predict results on future data sets.

Predictability

"Data-driven results are **predictable** if they can be shown to reemerge in (i.e., can be generalized to) new, relevant scenarios"

- This can apply to separate or future data sets

- A single data set can be partitioned into a **training set (60%), validation set (20%), test set(20%)**

Sets can be partitioned on **time-based splits, group-based splits,** or **randomly.**

Stability

"Data-driven results are **stable** if they tend not to change across reasonable alternative perturbations throughout the data science life cycle (DSLC)."

There is no single correct way to conduct a data science project.

Uncertainty comes from many different choices made in DSLC such as in

1. Data collection choices
2. Data cleaning and preprocessing choices
3. Algorithmic choices

Computability

Here we mean efficient, feasible, scalable analyses

How do we use our knowledge of algorithms, complexity, programming, data management, and computational environments to complete analyses?

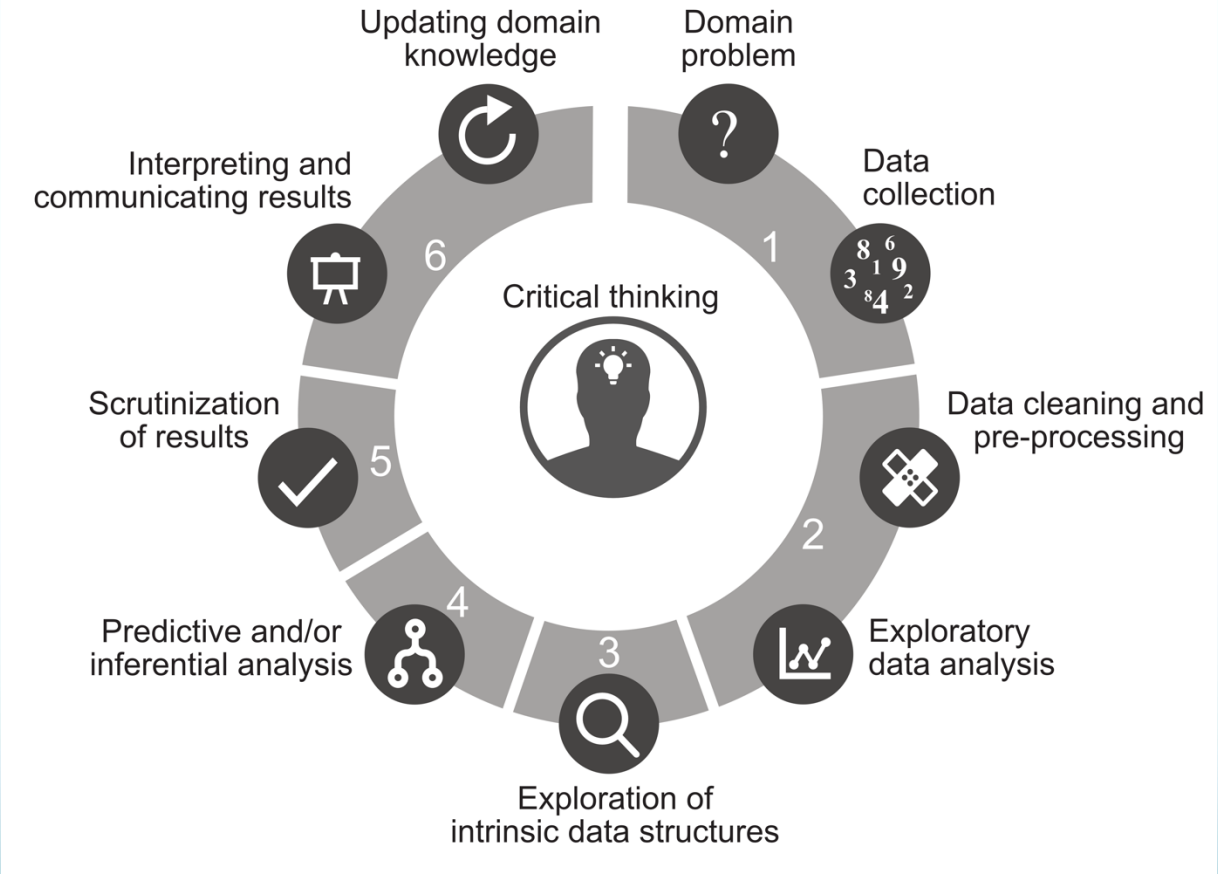
For this class we will use Python and Jupyter notebooks, but there are many other environments and languages we could use.

Today's Learning Objectives

Students will be able to:

- ✓ Define **data science vs. data analytics**
- ✓ Define **veridical data science**
- ✗ Identify stages of the **data science life cycle**

Data Science Life Cycle (DSLC)



Tabular Data

Features, variables, attributes, or covariates are columns

Dimension is number of columns.

Table 2.1: US government research and development budget and spending (reported in millions USD). The columns correspond to the year, the budget, the total spending, the spending on climate, the spending on energy, and the political party in office

Year	Budget	Total	Climate	Energy	Party
2000	142,299	1,789,000	2,312	13,350	Democrat
2001	153,197	1,862,800	2,313	14,511	Republican
2002	170,354	2,010,900	2,195	14,718	Republican
2003	192,010	2,159,900	2,689	15,043	Republican
2004	199,104	2,292,800	2,484	15,343	Republican
2005	200,099	2,472,000	2,284	14,717	Republican
2006	199,429	2,655,000	2,004	14,194	Republican
2007	201,827	2,728,700	2,044	14,656	Republican
2008	200,857	2,982,500	2,069	15,298	Republican
2009	201,275	3,517,700	2,346	16,492	Democrat

Tabular Data

Feature data types:

Numeric, categorical, dates and times,

Structured (short) text, Unstructured (long) text

Table 2.1: US government research and development budget and spending (reported in millions USD). The columns correspond to the year, the budget, the total spending, the spending on climate, the spending on energy, and the political party in office

Year	Budget	Total	Climate	Energy	Party
2000	142,299	1,789,000	2,312	13,350	Democrat
2001	153,197	1,862,800	2,313	14,511	Republican
2002	170,354	2,010,900	2,195	14,718	Republican
2003	192,010	2,159,900	2,689	15,043	Republican
2004	199,104	2,292,800	2,484	15,343	Republican
2005	200,099	2,472,000	2,284	14,717	Republican
2006	199,429	2,655,000	2,004	14,194	Republican
2007	201,827	2,728,700	2,044	14,656	Republican
2008	200,857	2,982,500	2,069	15,298	Republican
2009	201,275	3,517,700	2,346	16,492	Democrat

Tabular Data

Each row is an **observation**, **observational unit**, **data unit**, or **data point**

Table 2.1: US government research and development budget and spending (reported in millions USD). The columns correspond to the year, the budget, the total spending, the spending on climate, the spending on energy, and the political party in office

Year	Budget	Total	Climate	Energy	Party
2000	142,299	1,789,000	2,312	13,350	Democrat
2001	153,197	1,862,800	2,313	14,511	Republican
2002	170,354	2,010,900	2,195	14,718	Republican
2003	192,010	2,159,900	2,689	15,043	Republican
2004	199,104	2,292,800	2,484	15,343	Republican
2005	200,099	2,472,000	2,284	14,717	Republican
2006	199,429	2,655,000	2,004	14,194	Republican
2007	201,827	2,728,700	2,044	14,656	Republican
2008	200,857	2,982,500	2,069	15,298	Republican
2009	201,275	3,517,700	2,346	16,492	Democrat

DSLC Step 1: Problem Formulation and Data Collection

- Formulate the question

If hospital is concerned with reducing hospital-acquired infections (HAI), do they want to

i) Minimize risks to patients

ii) Improve practices to prevent HAI

DSLC Step 1: Problem Formulation and Data Collection

- Collect data to address questions

Sometimes data scientists are not involved with data collection. This requires knowledge of experimental design and data collection protocols.

Using existing data sets is common: **live data** is still being updated but **dead date** is no longer collected and maintained

Figure out data splits for predictability at this stage.

Your turn:
Air Quality Study

Imagine that you live in an area regularly affected by wildfires. You decide that you want to develop an algorithm that will predict the next day's Air Quality Index (AQI) in your town. In this project, we'll collectively brainstorm through stages of DSLC:

Go to:
https://docs.google.com/document/d/1a_bZczjeKCtLRNRjgSEiIKoZkxcOIimszuW7AqI1V40/edit?usp=sharing

We will edit this document together…

DSLC Step 2: Data Cleaning and Exploratory Data Analysis (EDA)

Data cleaning is the process of modifying a dataset so that it is tidy, appropriately formatted, and unambiguous.

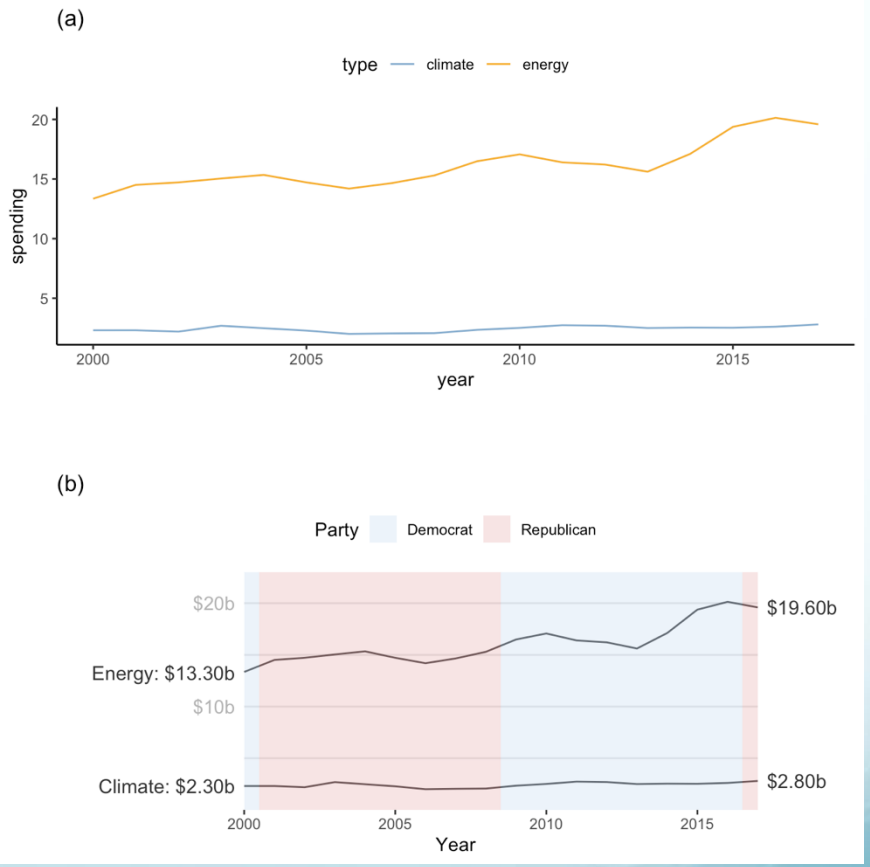
Preprocessing is the process of modifying a dataset so that it satisfies the formatting requirements of a particular analysis or algorithm.

This stage deals with errors, missing values, transforming data for algorithms (such as scaling), and **featurization** or constructing new features from data.

DSLC Step 2: Data Cleaning and Exploratory Data Analysis (EDA)

Exploratory data analysis (EDA) creates numeric and visual summaries of the data for understanding patterns in it

Explanatory data analysis polishes the most informative exploratory tables and graphs to communicate them to external audiences



Your turn:
Air Quality Study

What about step 2 for our air quality study?

Go to:
https://docs.google.com/document/d/1a_bZczjeKCtLRNRjgSEilKoZkxcOlimszuW7Aql1V40/edit?usp=sharing

We will edit this document together…

DSLC Step 3: Uncovering Intrinsic Data structures

Dimensionality reduction analysis helps us to reduce the number of features we consider to make the analyses more scalable.

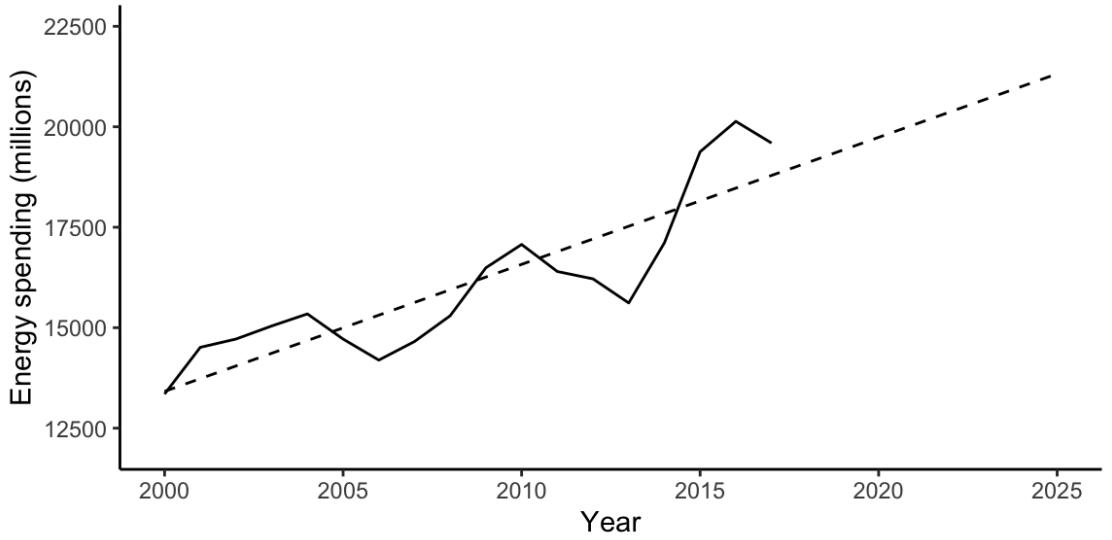
Cluster analysis helps us to identify groups of samples or observations that may be related and naturally form groups

The techniques used are often **unsupervised learning algorithms.**

DSLC Step 4: Predictive Analysis

In **prediction problems** our goal is to use past or current observable data to predict something about future unseen data.

Machine learning methods for prediction include **classification and regression.**



The techniques used are **supervised learning algorithms.**

DSLC Step 5 & 6: Evaluating and Communicating Results

Evaluating results should occur at every stage and we'll discuss specific techniques. At this stage external evaluation by **domain experts** is often needed.

Communicating your results should be catered to your **domain and audience.** Formats can include apps, software, slide presentations, and research papers among many.

We'll help you learn skills such as data visualization for this.

Your turn:
Air Quality Study

How would you complete your study?

Go to:
https://docs.google.com/document/d/1a_bZczjeKCtLRNRjgSEilKoZkxcOlimszuW7Aql1V40/edit?usp=sharing

We will edit this document together…

Today's Learning Objectives

Students will be able to:

- ✔ Define **data science vs. data analytics**
- ✔ Define **veridical data science**
- ✔ Identify stages of the **data science life cycle**