National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series

February 27, 2025

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

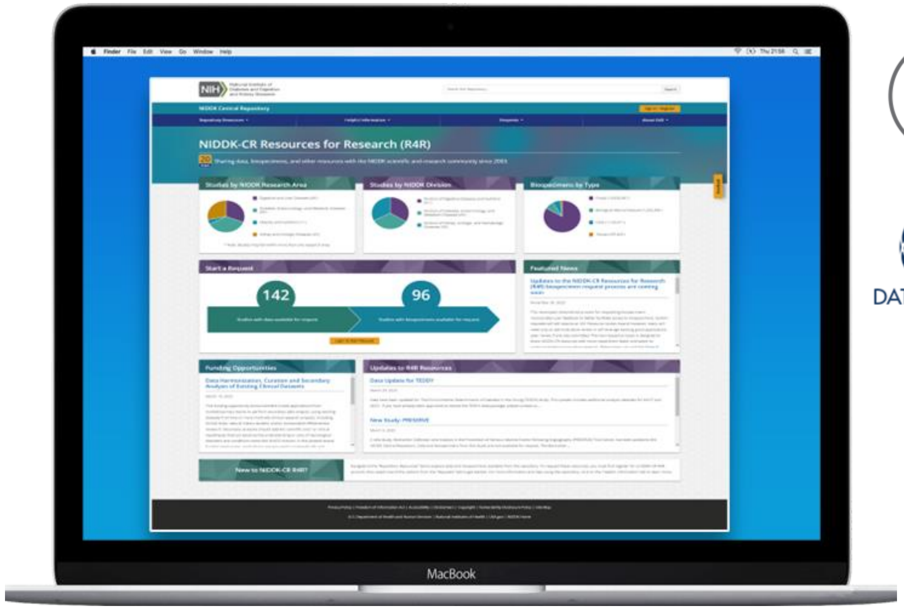
NIDDK Central Repository Overview

Mission

Established in 2003 to **facilitate the sharing of data, specimens, and other resources** generated from studies supported by NIDDK and within NIDDK's mission by making these **resources available for request to the broader scientific and research community**.

- Supports receipt and distribution of data and specimens in a manner that is ethical, equitable, and efficient
- Enables investigators not involved with the original work to test new hypotheses without the need to collect new resources
- Promotes FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles

Recorded past tutorials, webinars, and other educational resources can be found on the NIDDK-CR website



CORE TRUST SEAL

WORLD DATA SYSTEM

Imaging Data Files	Clinical Datasets	Biospecimens
15.8 M	>8,000 from 186 clinical studies	>16 M

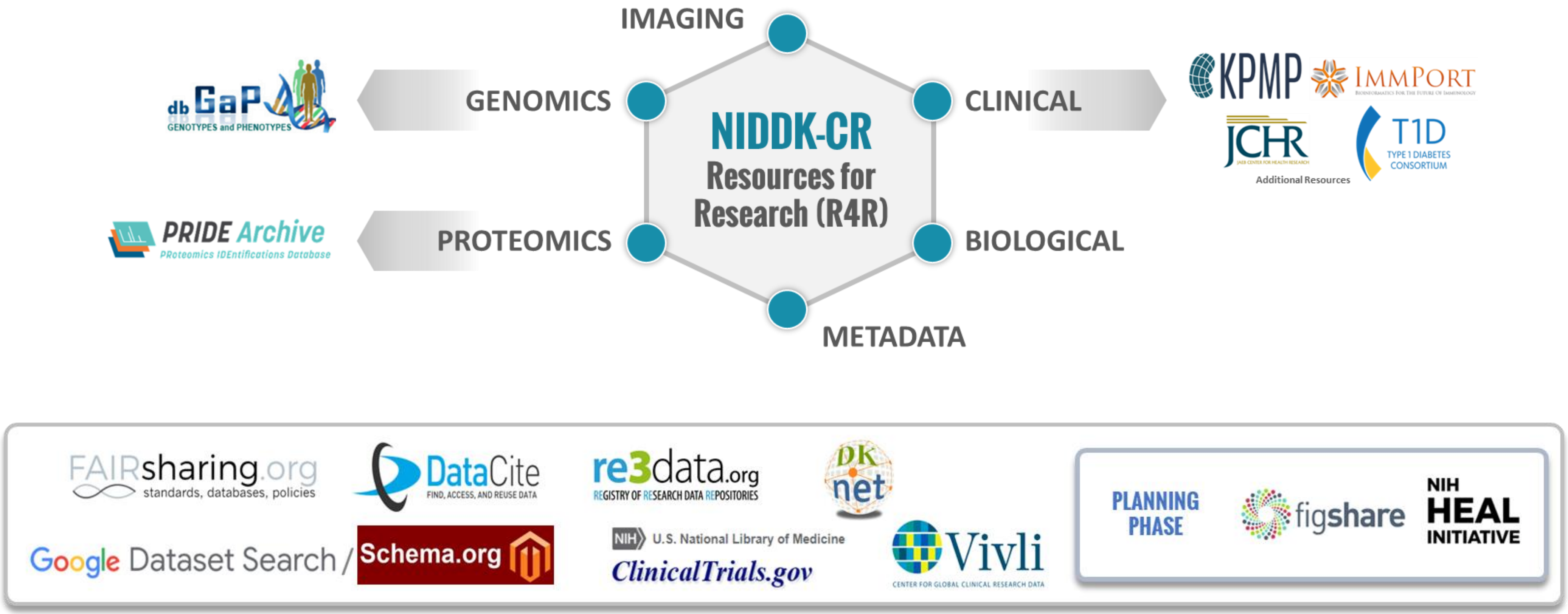
Registered Users	Weekly Users	Public Releases
6,785	>5,000	>800

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK Data Sharing Ecosystem

The NIDDK-CR is a part of the broader NIDDK-funded biomedical data ecosystem and plays a key role in NIH's FAIRness and TRUSTworthiness goals. The NIDDK-CR houses a broad range of data types for secondary research and provides access to specimens and direct links to other repositories with additional resources such as genomics data.



NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Future Functionality: Analytics Workbench

***Streamlining** end-to-end data science lifecycle and discovery of data-driven biomedical insights.*

Innovation and ease of use

A cloud-based analytics environment where researchers and data scientists can access a suite of integrated analytics tools and cloud computing resources to participate in data challenges and AI innovation.

Expected Benefits of Analytics Workbench:

Promote Collaboration	Support AI Innovation	Minimize Data Movement
Improve User Experience	Discover Data Insights	Advance NIDDK Research Mission

Analytics Workbench

Insights and Interactions

Search | Access Services | Communication Services | Collaborate & Share | Chatbot | Data Insights

Data Analytics and AI Services

Data Discovery
- Search
- Query
- Data Discovery and Exploration

Descriptive
- Usage Reports
- Dashboards
- Self Service BI

Predictive
- Predictive modeling with biomarkers
- Disease Forecasting
- Simulation

Cognitive AI
- Natural Language Processing
- Machine Learning
- Generative AI

User-Centric
- Personalization
- Collaboration workspace
- Communication

Prescriptive Data Analytics and AI Tools

jupyter | sas | GitHub Copilot

Data Services / API Connectors

Data and Cloud Infrastructure Foundation

Internal Data Sources | External Data Sources

Platform Provisioning | Deployment | Monitoring | Security | Access | User Management

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Data Science Centric Challenge Series

Goals of NIDDK-CR Data-science centric challenge series

- Develop tools, approaches, models and/or methods to increase data interoperability and usability for artificial intelligence (AI) and machine learning (ML) applications
- Augment and enhance existing data for future secondary research, including data-driven discovery by AI/ML researchers
- Discover innovative approaches to enhance the utility of datasets for AI/ML applications

Visit our website for more information on our data-centric movement and to learn more about our past data-challenges

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Data Science and Meet the Expert Webinar Series

About the Series

- Aims to accelerate data science and AI-driven biomedical research by fostering collaboration between biomedical researchers and experts in the field
- Monthly webinar held on the **last Thursday of each month**

Upcoming Webinars

- Data science fundamentals
- Artificial Intelligence fundamentals
- FAIR data sharing
- Privacy protections for sharing human research participants’ data
- Different privacy preserving techniques and implications for secondary researchers
- Challenges, opportunities, and considerations for secondary researchers using electronic health records and real-world data sources
- Impact and innovations realized

Learn more about the webinar series, register for future webinars, and access past webinars materials and recordings

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Meet the Experts

Arica Christensen is a Lead Associate Data Scientist at Booz Allen Hamilton, with a B.S. in Industrial and Systems Engineering from the University of San Diego. She specializes in natural language processing techniques and supervised machine learning. Arica has supported NAVWAR C4I PMW 130 on Project RAVEN applying predictive and proactive analytics for fleet readiness and cyber awareness. Currently Arica supports the Chief Digital Artificial Intelligence Office focusing on the development of dashboards and data pipelines measuring risk and resilience for all sailors at the individual and UIC level. Additionally, Arica leads the NAVWAR 4.0 Data Science Learning Program to create and facilitate trainings Navy wide on data science, machine learning, and artificial intelligence techniques.

Gordon Aiello is a Lead Scientist at Booz Allen Hamilton with a PhD in Applied Mathematical and Computational Sciences. He works full-time developing and delivering specialized data science, artificial intelligence, machine learning, and Python trainings for clients in the Navy and Intelligence Community. Prior to joining Booz Allen Hamilton, Dr. Aiello worked in the Office of Macroeconomic Affairs at the U.S. Department of State, using machine learning techniques to analyze developing and emerging market economies. Additionally, he has taught courses on data science and the R programming language for the Foundation for Advanced Education in the Sciences (FAES) at the NIH. He is passionate about working with others to expand their understanding of data science techniques and their applications.

Data Science Fundamentals

NIDDK-CR Data Science

Meet the Experts Webinar Series

Feb 27th, 2025

Presented by: Booz Allen Hamilton

Training Guidance

- Avoid CUI/PII/PHI conversations
- Questions in Teams Chat are encouraged
- Due to size of class, stay on mute until end of class

Office of Clinical Research Support

Instructor Introductions

Data Science Learning Program

If you're new to data science, start your learning journey with the **Foundations** courses. A more in-depth learning track starts with the **Data Science Fundamentals** course and continues to the **Data Science Labs**. Those interested in more specialized topics can explore courses in the **Select Topics** track.

INTRODUCTION

Foundations for Data Citizens

- Data Citizen best practices
- Data governance
- Data-driven organization

udemy

Foundations of Data Analytics

- For NAVWAR supervisors
- Data Science Overview
- Machine Learning and Artificial Intelligence

udemy

THEORY-TO-PRACTICE

Data Science Fundamentals

- Comprehensive intro to Data Science
- Python programming
- Statistics, Probability and Linear Algebra refresher
- Machine learning and Artificial Intelligence

Live Training

udemy

10.5 hours (3 sessions)

Data Science Project Lab*

- Theory-to-practice
- Case study format
- Hands-on exercises
- Tabular data cleansing and processing techniques
- Full-cycle analytics process

Live Training

JUPITER

12 hours (3 sessions)

Data Science NLP Lab*

- Theory-to-practice
- Case study format
- Hands-on exercises
- Natural Language Processing Techniques
- Large Language Models

Live Training

JUPITER

12 hours (3 sessions)

*Completion of the Introduction to Python course is recommended for those without programming experience.

SELECT TOPICS

Introduction to Data Visualization

- Telling a story with your data
- How to create more impactful briefings
- Not product specific

Live Training

udemy

3 hours

Python Fundamentals for Data Science

- Foundational Python syntax
- Develop essential analytic skills
- Machine Learning and Artificial Intelligence

Live Training

JUPITER

7 hours (2 sessions)

Artificial Intelligence Fundamentals

- AI initiatives and foundational AI
- AI ecosystems and AI operations
- Responsible and Ethical AI
- Neural Networks

Live Training

7 hours (2 sessions)

Data Science for Managers

Developed in partnership with NGA

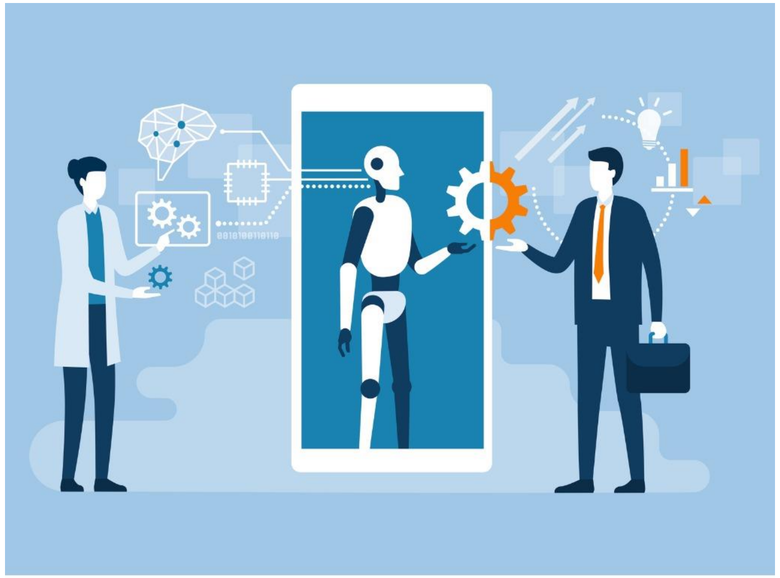
- Management responsibilities in Data Science Projects
- Ethical considerations in Data Science
- Data Science and AI Opportunities

In Person Training

8 hours

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Agenda



1. The Data Science Process
2. Supervised and Unsupervised Learning Techniques
3. Deep Learning
4. Specialized Data Science Topics
    1. Computer Vision
    2. Time Series
    3. Natural Language Processing

The Data Science Process

Data Science Defined

The goal of data science is to extract meaningful insights from data.

Data – any kind of qualitative or quantitative set of values

- Common examples in data science today:
  - Natural text: "I'm cold," "I'm not very cold"
  - Categories: "yellow," "green," "red"
  - Numbers: 1, 2.53, -4
  - Images:  
- Sometimes you have the data, sometimes you need to procure the data

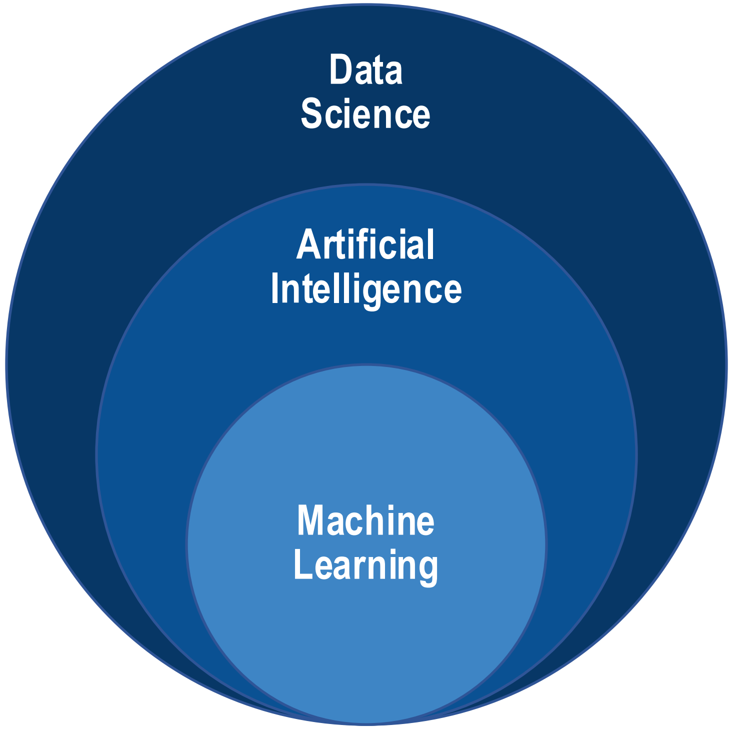
Science – a systematic approach to building knowledge by testing hypotheses

- Think Scientific Method:

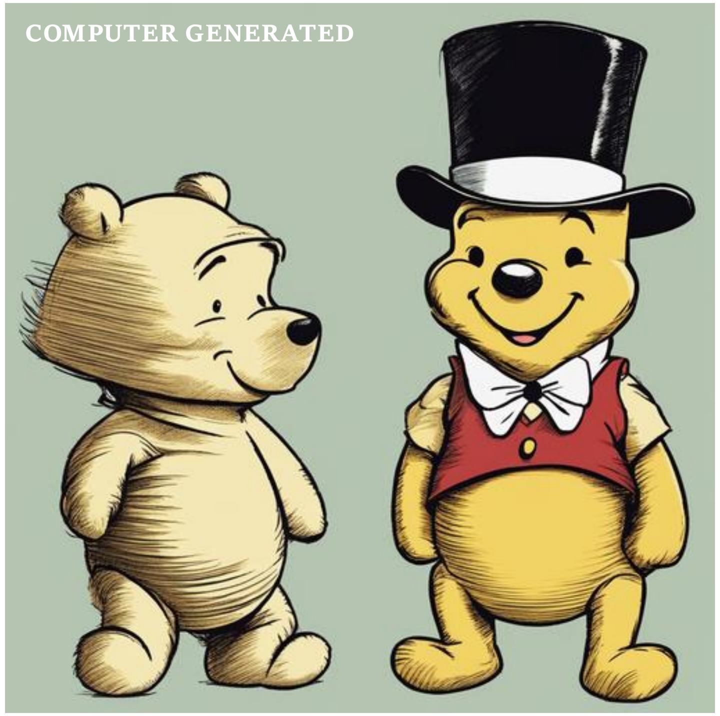
  Define a hypothesis → Collect the data → Analyze results → Draw conclusions
- Hypotheses must be testable, and experiments must be reproducible

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

AI Is a Subset of Data Science



Machine Learning is an application of Artificial Intelligence, and Machine Learning is part of Data Science by applying algorithms and statistics to extract knowledge and insights from data



Statistics Data Science

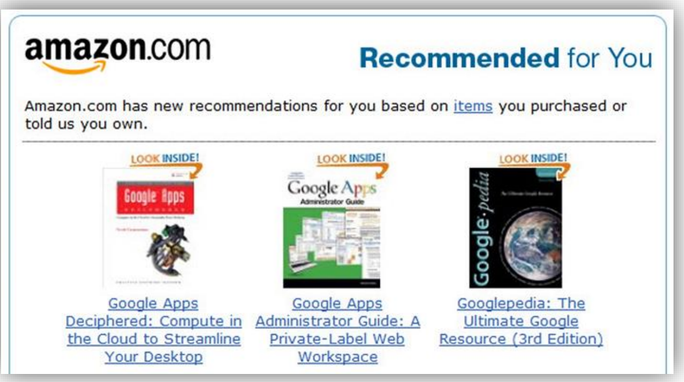
Artificial Intelligence (AI)

The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages

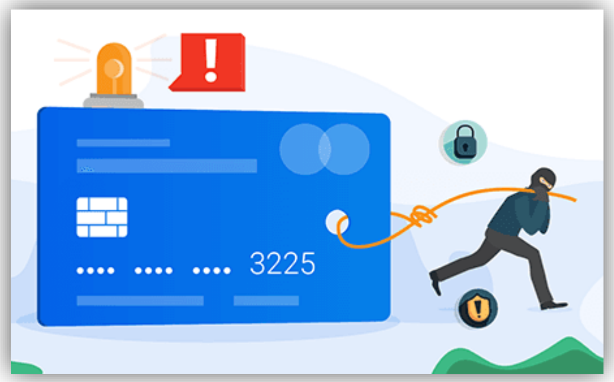
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

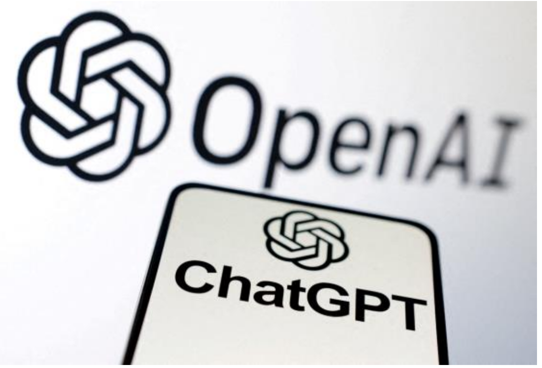
Data Science in the Commercial Space



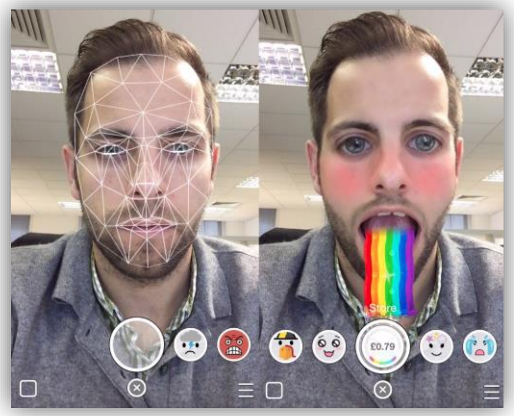
Amazon: Recommendation Systems



Credit Card Fraud Detection



ChatGPT



Snapchat: Computer Vision

Google Voice Recognition

UPS ORION (On-Road Integrated Optimization and Navigation)

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

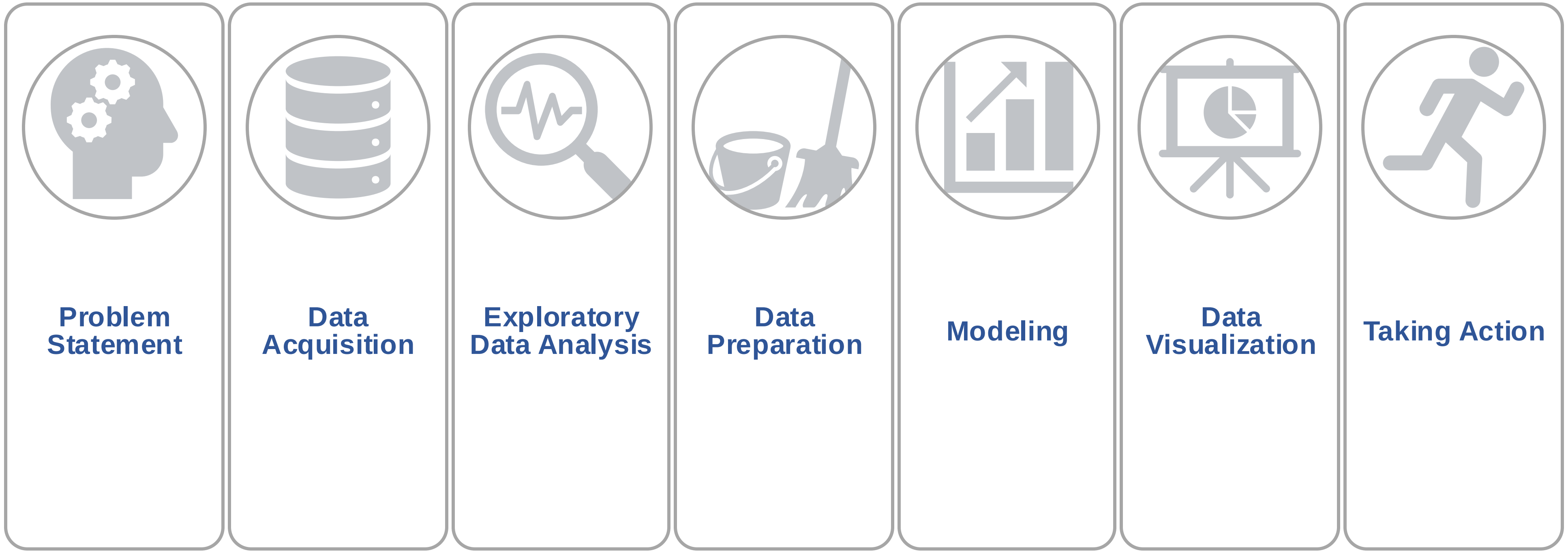
Data Science in Healthcare

- **Predictive Analytics for Early Diagnosis** – Data science enables early detection of diseases by analyzing patient data, identifying risk factors, and improving treatment outcomes.
- **Personalized Medicine** – Machine learning models help tailor treatments based on a patient's genetic profile, lifestyle, and medical history, leading to more effective therapies.
- **Early Detection and Progression Monitoring of Kidney Disease** – Data science helps analyze lab results (e.g., creatinine levels, eGFR) to detect kidney disease in its early stages and predict progression, allowing for timely intervention.
- **Diabetes Prediction and Management** – Machine learning models can analyze patient data, including glucose levels and lifestyle factors, to predict diabetes risk, personalize treatment plans, and optimize insulin management.

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Data Science Process

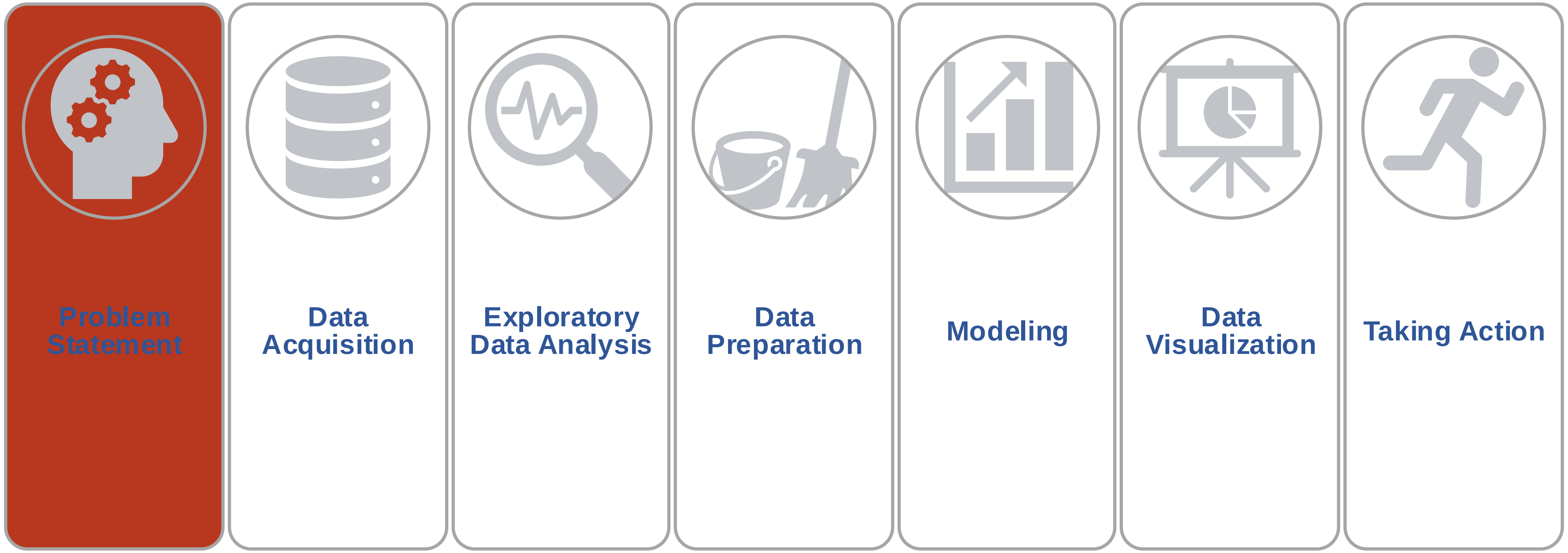


7 Step Data Science Process

Navigate through a data science project using the seven-step data science process:

1. Form a SMART problem statement, understanding what data science can and cannot do
2. Acquire useful data that can assist in solving the problem statement
3. Explore data and analyze preliminary findings to leverage initial insights from the data
4. Prepare data for use in machine learning pipelines
5. Understand basic machine learning model concepts
6. Render compelling visualizations to communicate data-driven narratives to your colleagues
7. Apply the insights gained from your data science project to your work

Problem Statement



What Questions Can Data Science Answer?

- Data science can't answer just any question
- Questions must be structured and attainable
- A few questions to ask yourself to help you get started:

What do you want to predict or estimate?

What is the goal?

What would be the value added to your work?

What would you do if you had *all* the data?

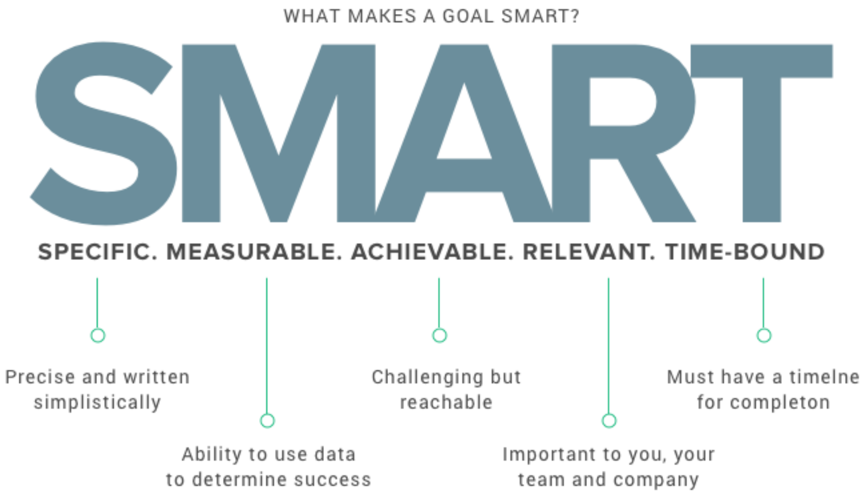
What patterns do you think are relevant?

Problem Statement | Data Acquisition | Exploratory Data Analysis | Data Preparation | Modeling | Data Visualization | Taking Action

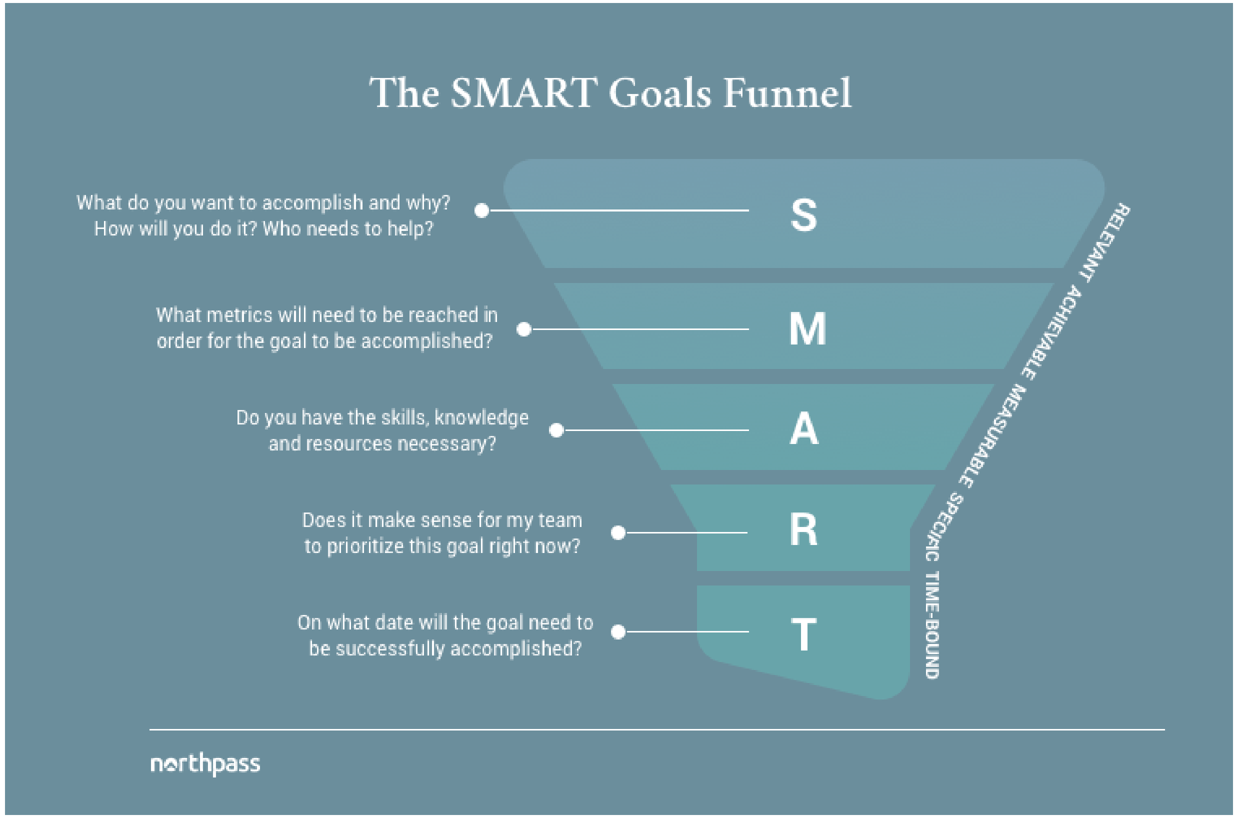
NIH National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Creating SMART Problem Statements

- When developing your problem statements, think through whether the question is SMART!
- Although SMART goals are not necessarily specific to data science, we can use this methodology to make sure we create attainable problem statements



The SMART Goals Funnel



Problem Statement | Data Acquisition | Exploratory Data Analysis | Data Preparation | Modeling | Data Visualization | Taking Action

Example – SMART Goals

Healthcare Objective

- Researchers seek a data-driven approach to better understand the factors that are most strongly associated with chronic kidney disease.
- The aim of this project is to develop models using clinical patient data to accurately predict chronic kidney disease.

Example – SMART Goals

S – We will train machine learning models using clinical patient data to predict who's at greatest risk for developing chronic kidney disease.

M – Success will be measured by the model's accuracy – targeting at least 90%.

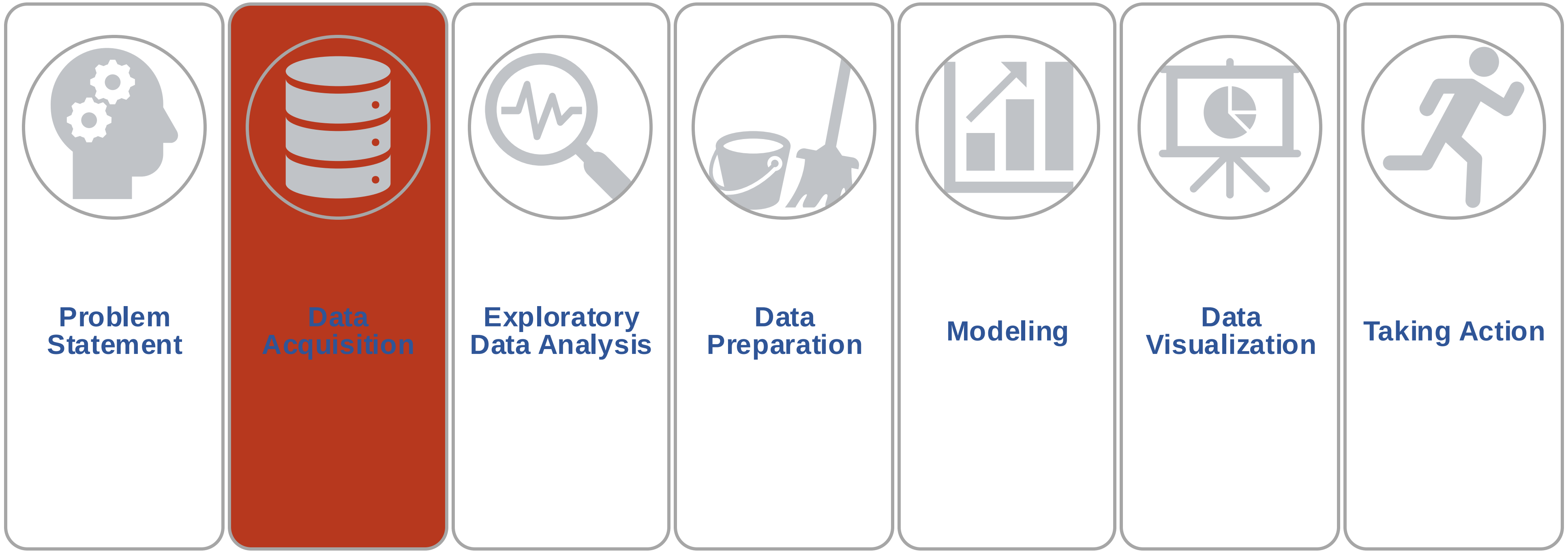
A – By leveraging existing data sets and proven analytics capabilities, we'll work with resources readily available to the NIH.

R – The models will help researchers and medical providers make objective, data-driven healthcare decisions by highlighting insights that may be currently overlooked.

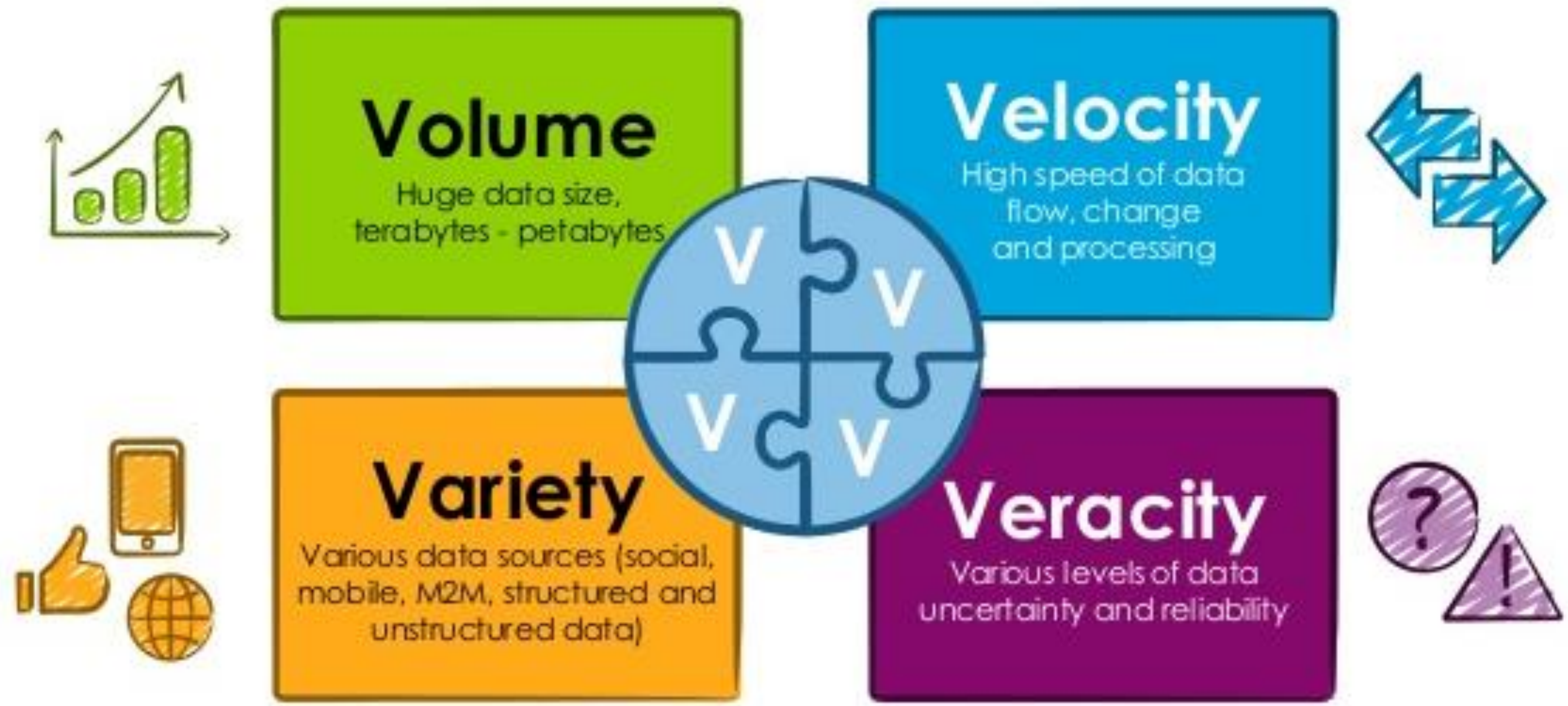
T – The models will be developed, validated, and ready for deployment within 6 months, with a prototype ready for review in 3 months.

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Data Acquisition



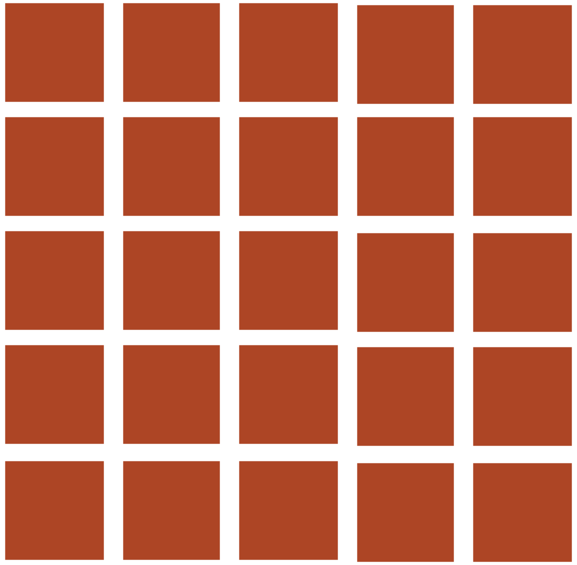
The Four Vs of Big Data



NIH National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Structured vs Unstructured Data

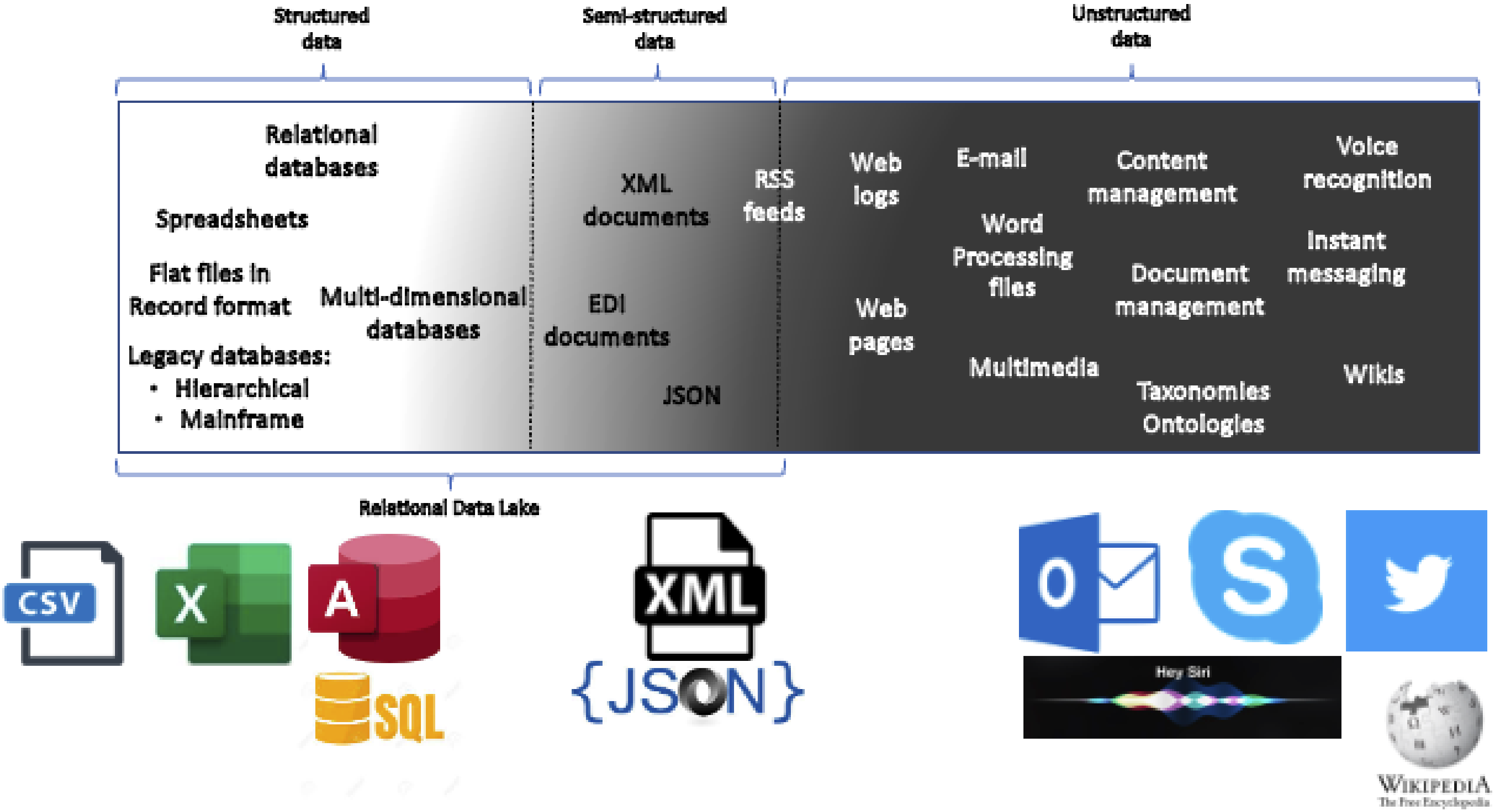
Structured Data



Unstructured Data

Problem Statement | Data Acquisition | Exploratory Data Analysis | Data Preparation | Modeling | Data Visualization | Taking Action

Data Acquisition: Data Structures



National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Example – Kidney Disease Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	id	age	bp	sg	al	su	rbc	pc	pcc	ba	bgr	bu	sc	sod	pot	hemo	pcv	wc	rc	htn	dm	cad	appet	pe	ane	classification
2	0	48	80	1.02	1	0		normal	notpresent	notpresent	121	36	1.2			15.4	44	7800	5.2	yes	yes	no	good	no	no	ckd
3	1	7	50	1.02	4	0		normal	notpresent	notpresent		18	0.8			11.3	38	6000		no	no	no	good	no	no	ckd
4	2	62	80	1.01	2	3	normal	normal	notpresent	notpresent	423	53	1.8			9.6	31	7500		no	yes	no	poor	no	yes	ckd
5	3	48	70	1.005	4	0	normal	abnormal	present	notpresent	117	56	3.8	111	2.5	11.2	32	6700	3.9	yes	no	no	poor	yes	yes	ckd
6	4	51	80	1.01	2	0	normal	normal	notpresent	notpresent	106	26	1.4			11.6	35	7300	4.6	no	no	no	good	no	no	ckd
7	5	60	90	1.015	3	0			notpresent	notpresent	74	25	1.1	142	3.2	12.2	39	7800	4.4	yes	yes	no	good	yes	no	ckd
8	6	68	70	1.01	0	0		normal	notpresent	notpresent	100	54	24	104	4	12.4	36			no	no	no	good	no	no	ckd
9	7	24		1.015	2	4	normal	abnormal	notpresent	notpresent	410	31	1.1			12.4	44	6900	5	no	yes	no	good	yes	no	ckd
10	8	52	100	1.015	3	0	normal	abnormal	present	notpresent	138	60	1.9			10.8	33	9600	4	yes	yes	no	good	no	yes	ckd
11	9	53	90	1.02	2	0	abnormal	abnormal	present	notpresent	70	107	7.2	114	3.7	9.5	29	12100	3.7	yes	yes	no	poor	no	yes	ckd
12	10	50	60	1.01	2	4		abnormal	present	notpresent	490	55	4			9.4	28			yes	yes	no	good	no	yes	ckd
13	11	63	70	1.01	3	0	abnormal	abnormal	present	notpresent	380	60	2.7	131	4.2	10.8	32	4500	3.8	yes	yes	no	poor	yes	no	ckd
14	12	68	70	1.015	3	1		normal	present	notpresent	208	72	2.1	138	5.8	9.7	28	12200	3.4	yes	yes	yes	poor	yes	no	ckd
15	13	68	70						notpresent	notpresent	98	86	4.6	135	3.4	9.8				yes	yes	yes	poor	yes	no	ckd
16	14	68	80	1.01	3	2	normal	abnormal	present	present	157	90	4.1	130	6.4	5.6	16	11000	2.6	yes	yes	yes	poor	yes	no	ckd
17	15	40	80	1.015	3	0		normal	notpresent	notpresent	76	162	9.6	141	4.9	7.6	24	3800	2.8	yes	no	no	good	no	yes	ckd
18	16	47	70	1.015	2	0		normal	notpresent	notpresent	99	46	2.2	138	4.1	12.6				no	no	no	good	no	no	ckd
19	17	47	80						notpresent	notpresent	114	87	5.2	139	3.7	12.1				yes	no	no	poor	no	no	ckd
20	18	60	100	1.025	0	3		normal	notpresent	notpresent	263	27	1.3	135	4.3	12.7	37	11400	4.3	yes	yes	yes	good	no	no	ckd
21	19	62	60	1.015	1	0		abnormal	present	notpresent	100	31	1.6			10.3	30	5300	3.7	yes	no	yes	good	no	no	ckd
22	20	61	80	1.015	2	0	abnormal	abnormal	notpresent	notpresent	173	148	3.9	135	5.2	7.7	24	9200	3.2	yes	yes	yes	poor	yes	yes	ckd
23	21	60	90						notpresent	notpresent		180	76	4.5		10.9	32	6200	3.6	yes	yes	yes	good	no	no	ckd
24	22	48	80	1.025	4	0	normal	abnormal	notpresent	notpresent	95	163	7.7	136	3.8	9.8	32	6900	3.4	yes	no	no	good	no	yes	ckd

Data Source: UC Irvine Machine Learning Repository

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Example – Kidney Disease Data

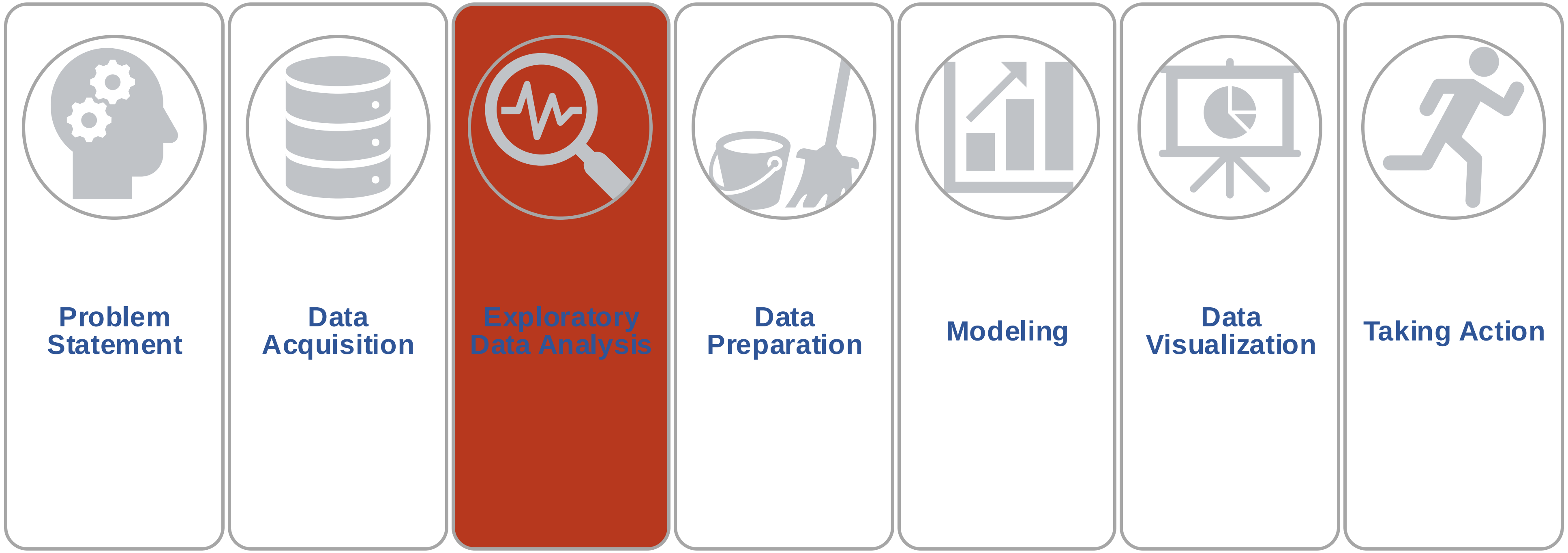
Column Name	Description	Data Type
age	Age	Numeric
bp	Blood Pressure	Numeric
sg	Specific Gravity	Numeric
al	Albumin	Numeric
su	Sugar	Numeric
rbc	Red Blood Cells	Categorical
sc	Serum Creatinine	Numeric
…	…	…
classification	Chronic Kidney Disease (yes/no)	Binary (Categorical)

Source: UC Irvine Machine Learning Repository

Problem Statement | Data Acquisition | Exploratory Data Analysis | Data Preparation | Modeling | Data Visualization | Taking Action

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Exploratory Data Analysis



Understand Your Data!

- It's time to sit down and analyze the many intricacies of the dataset
- It's important to look for different insights to understand your data as a whole

Arnon Rotem-Gal-Oz

Director of Technology Research, Amdocs

The blind men and the elephant. Poem by John Godfrey Saxe (Cartoon originally copyrighted by the authors; G. Renee Guzlas, artists http://www.nature.com/ki/journal/v62/n5/fig_tab/4493262f1.html

Types of Analytics

Four types of analytics:

- **Descriptive:** What happened?
- **Diagnostic:** Why did it happen?
- **Predictive:** What will happen?
- **Prescriptive:** How can we make this happen?

Types of Discovery

Class Discovery

- Find the categories of objects (population segments), events, and behaviors in your data

Outlier Discovery

- Find the new, surprising, unexpected one-in-a-million object, event, or behavior

2 4 6 9

Correlation Discovery

- Find trends, patterns, and dependencies in data that reveal the governing principles or behavioral patterns (the object's "DNA")

Association Discovery

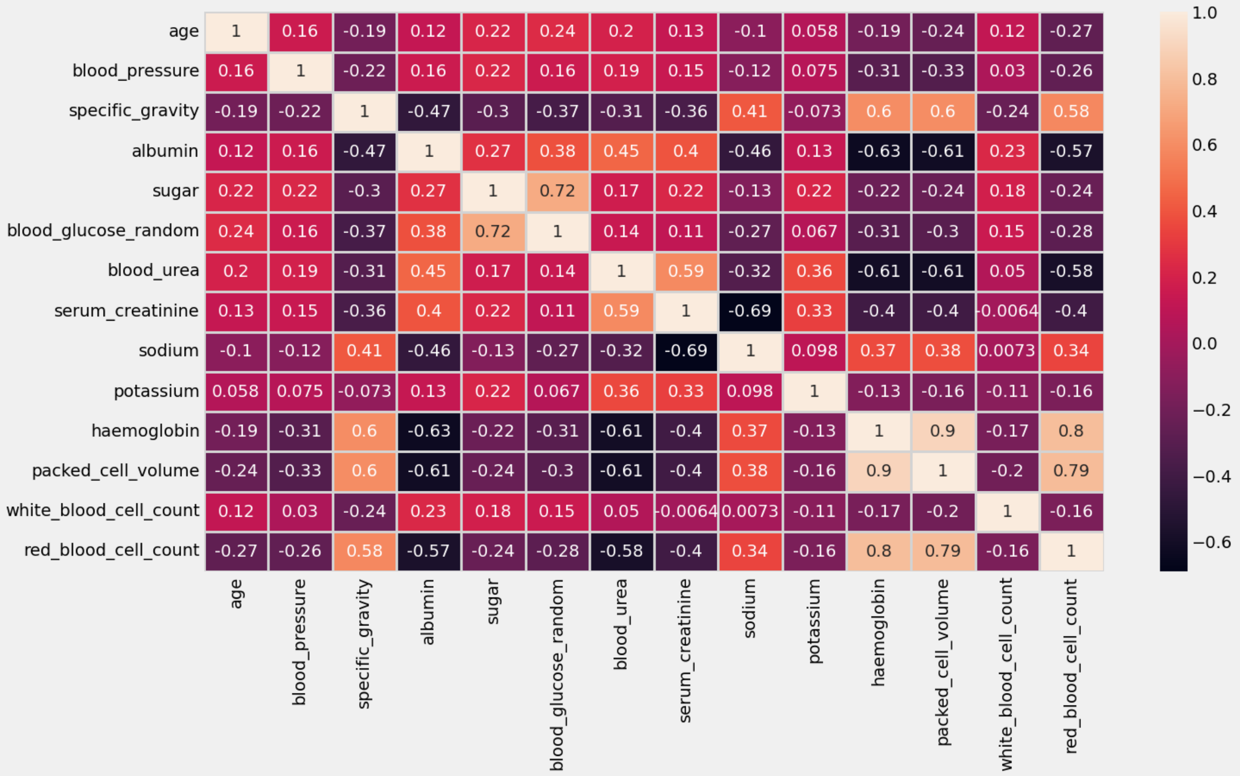
- Find both the typical (usual) and the atypical (unusual, interesting) data associations, links, or connections in your domain

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

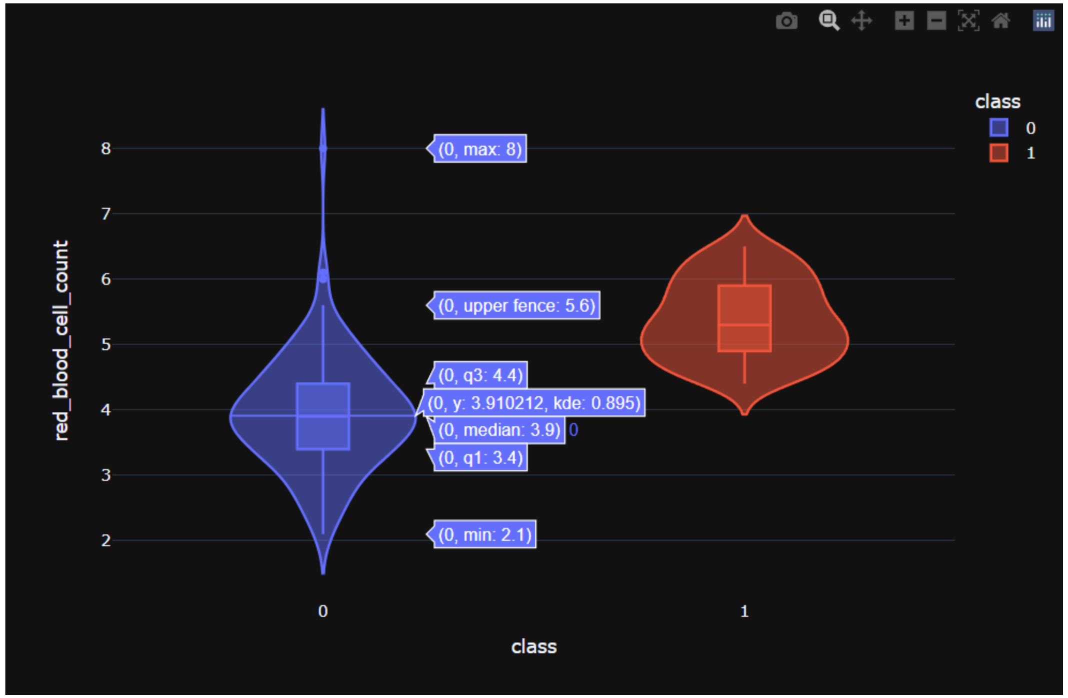
Example – Correlation Heatmap

```python
# heatmap of data

plt.figure(figsize = (15, 8))

sns.heatmap(df[num_cols].corr(), annot = True, linewidths = 2, linecolor = 'lightgrey')
plt.show()
```

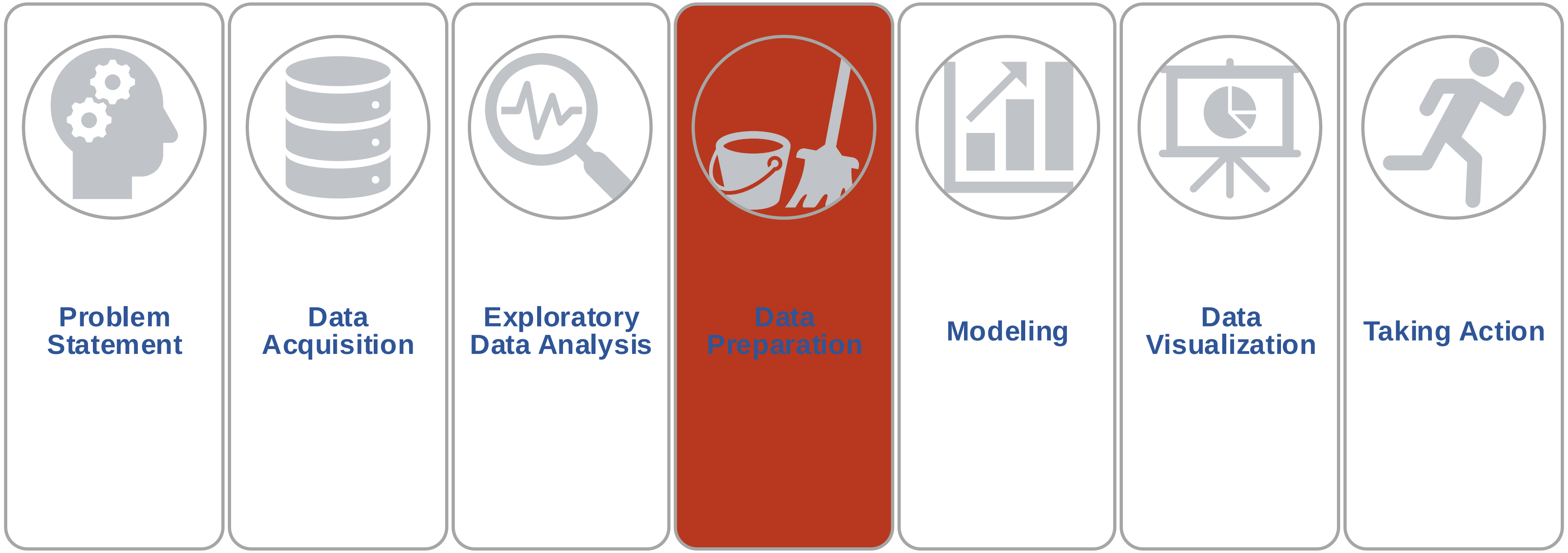


Example – Violin Distribution Plot

```
import plotly.express as px

fig = px.violin(df, y= 'red_blood_cell_count', x= 'class', color= 'class', box = True, template = 'plotly_dark')
fig.show()
```



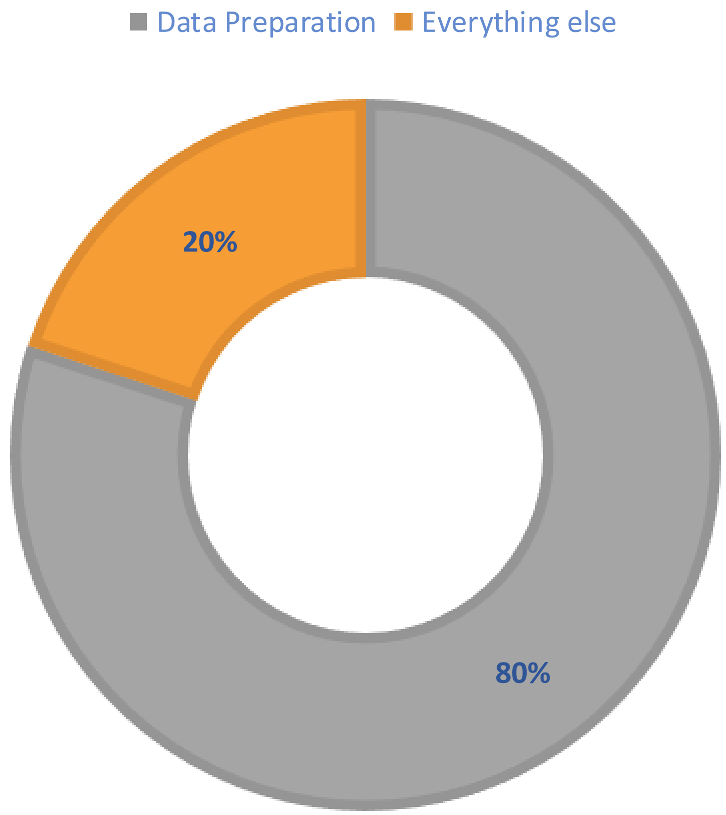
Data Preparation



Real World Data Is Messy

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

DATA SCIENTIST'S WORKLOAD



Dirty Data

- Dirty data is everywhere!
  - Any data with typos or errors
  - Missing data
  - Null fields
  - Different labels for the same item
  - Duplicate entries
  - Entries that don't match up with another dataset
  - So much more!
- These data entries can skew the outcome of your models

Example – Missing Values

```
# checking for null values

df.isna().sum().sort_values(ascending = False)
```

red_blood_cells	152
red_blood_cell_count	131
white_blood_cell_count	106
potassium	88
sodium	87
packed_cell_volume	71
pus_cell	65
haemoglobin	52
sugar	49
specific_gravity	47
albumin	46
blood_glucose_random	44
blood_urea	19
serum_creatinine	17
blood_pressure	12
age	9

Garbage In, Garbage Out

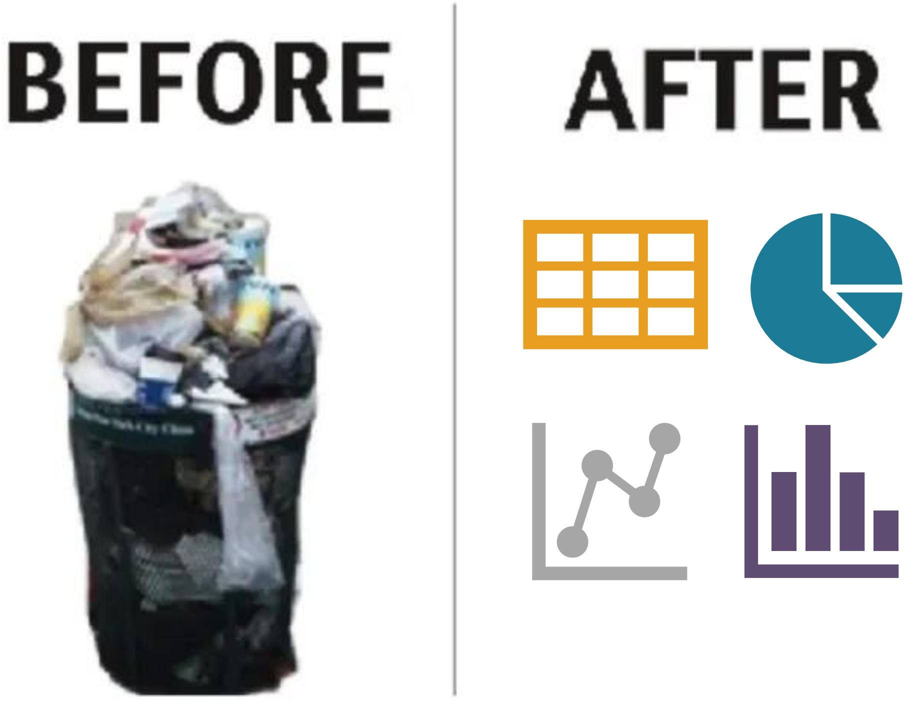
- Concept that the quality of information coming out can only be as good as the quality of information that went in

- In other words, the condition of the data going into a model is the ceiling of the condition of the outcoming data

NIH National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Feature Engineering

- Using domain knowledge to adjust the dataset and use it properly for the chosen model and question
- Applications of feature engineering:
    - Imputation
    - Handling Outliers
    - Binning
    - Scaling
    - Log Transformation
    - One-Hot Encoding
    - Grouping Operations
    - Feature Split
    - Extracting Date



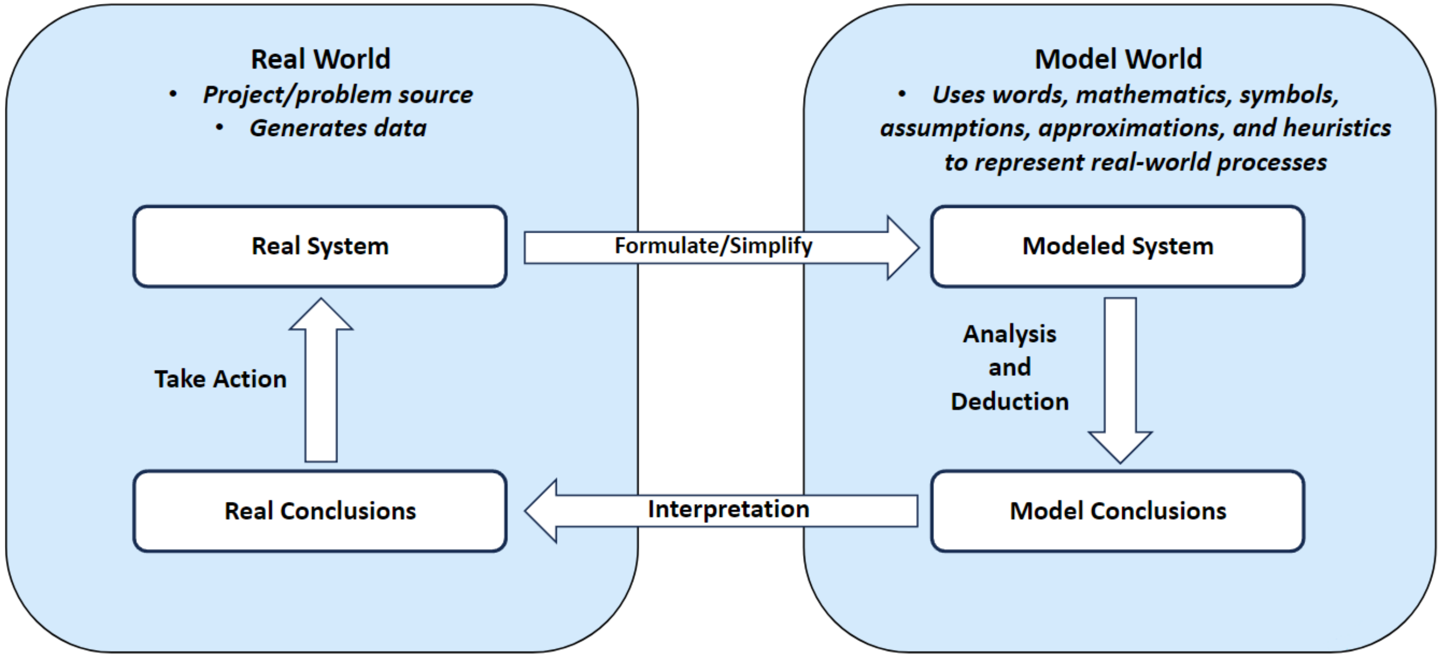
Problem Statement | Data Acquisition | Exploratory Data Analysis | Data Preparation | Modeling | Data Visualization | Taking Action

Modeling


Problem Statement
Data Acquisition
Exploratory Data Analysis
Data Preparation
Modeling
Data Visualization
Taking Action

What Is a Model?

- A model is a representation of a real-world process.
- Models use simplifying assumptions to make problems more tractable (e.g., for analyses and computational purposes).
- **Goal**: Balance representing the real world to a high fidelity with the level of simplification imposed.

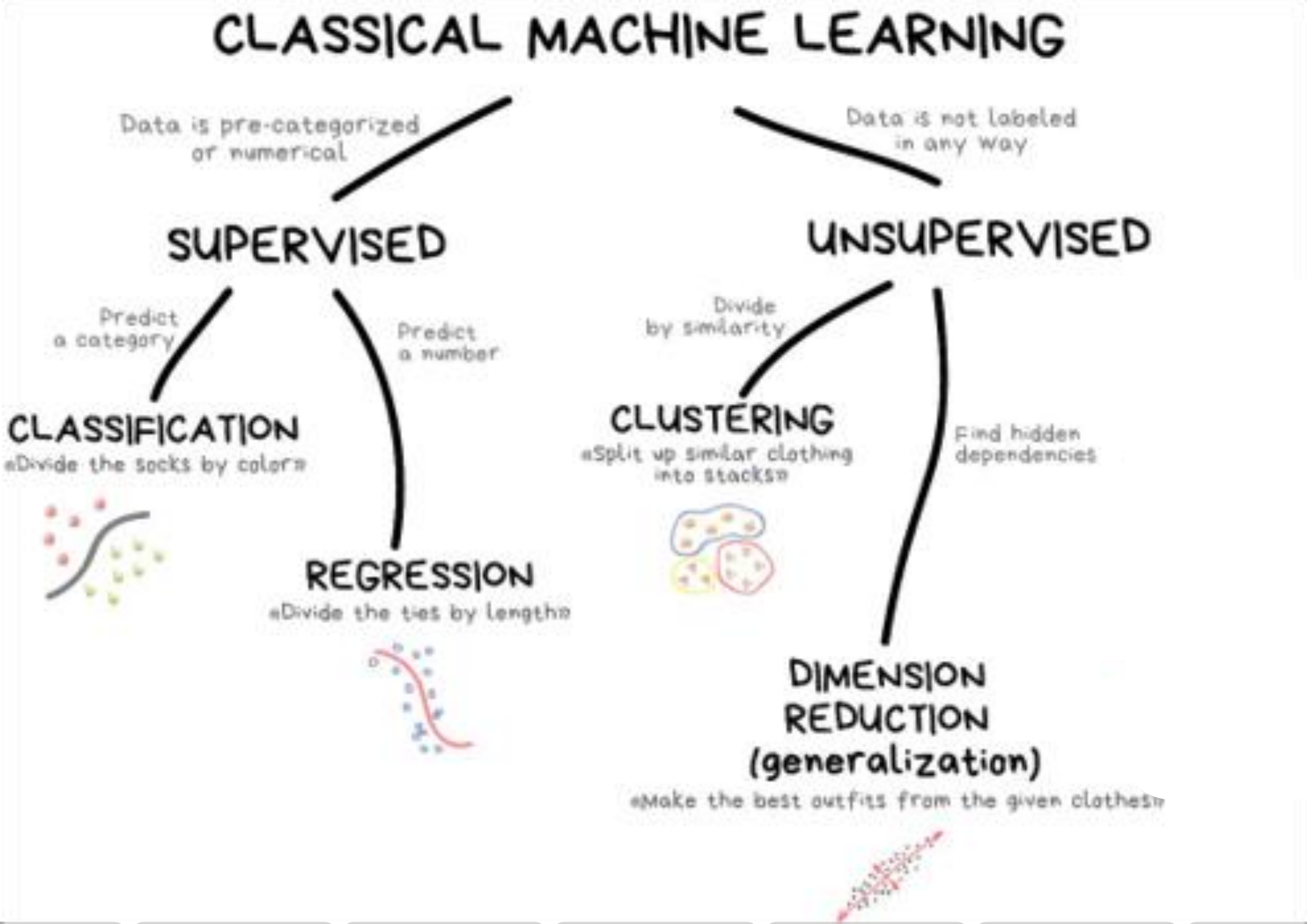


What Is a Machine Learning Model?

- A machine learning (ML) model is a program or system that learns patterns from data.
- ML models use discovered patterns to make predictions or decisions without being explicitly programmed for specific tasks.
- A helpful analogy is to think of ML as akin to baking, as illustrated in the correspondence table at right.
- The computer's goal is to determine the "best" way to mix data together to achieve a desired outcome.

Machine Learning	Baking
Data • Human provided	Ingredients to be mixed together (e.g., flour, sugar, butter, etc.)
Model parameters • Computer determines these	Quantities of ingredients used in recipe (e.g., 3 cups sugar, 4 tbsp butter, etc.)
Desired model output • Human provided	Tasty treat (e.g., cake, cookie, biscuit, etc.)

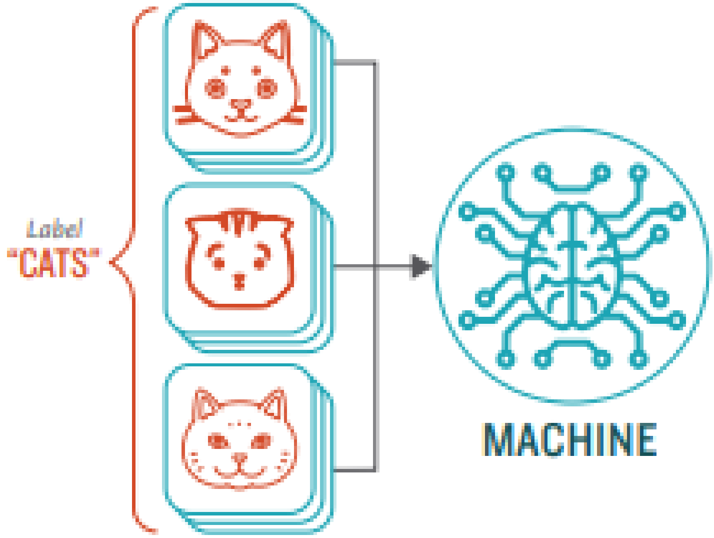
Types of Machine Learning



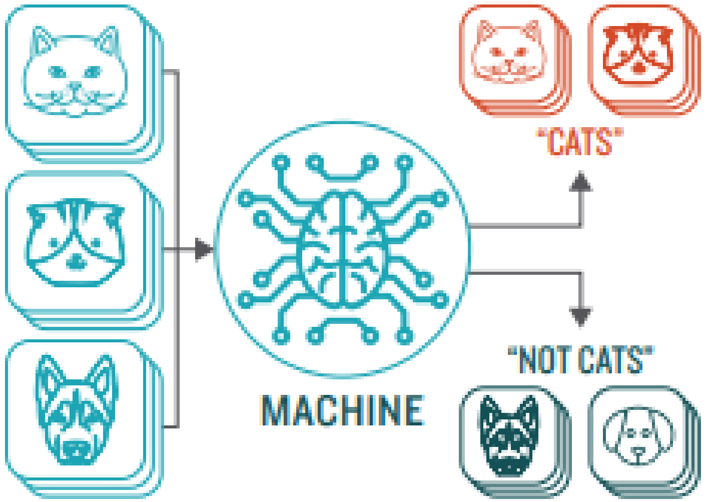
Problem Statement | Data Acquisition | Exploratory Data Analysis | Data Preparation | Modeling | Data Visualization | Taking Action

Supervised Learning

- Step 1: Provide the machine learning algorithm **categorized or “labeled” input** and output data to learn from

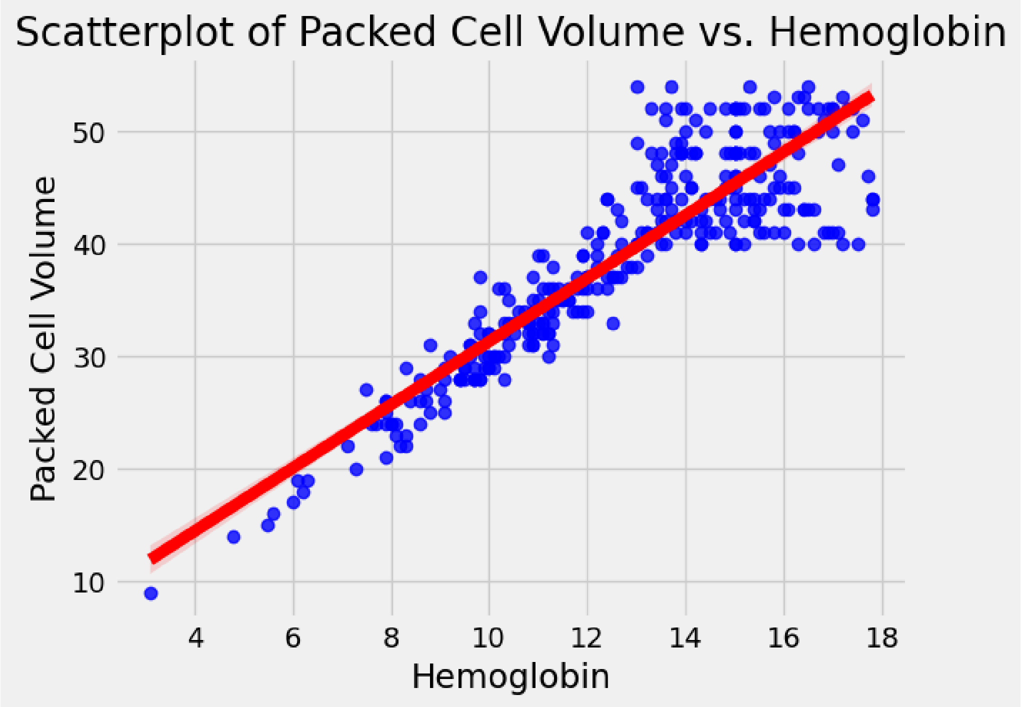


- Step 2: Feed the machine **new, unlabeled information** to see if it tags new data appropriately. If not, continue refining the algorithm



Regression

Regression is the task of predicting a continuous numeric value.



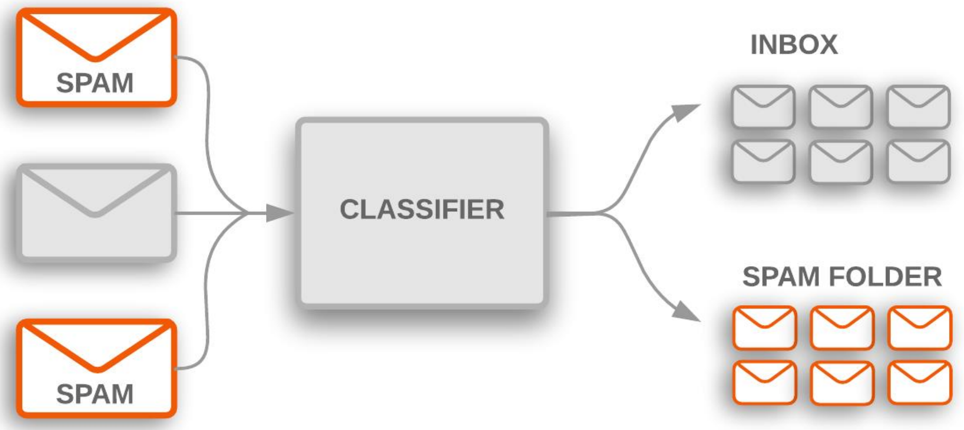
Examples:

- Predicting packed cell volume using hemoglobin measurements.
- Predicting house price using square footage.
- Forecasting the price of a stock.

Classification

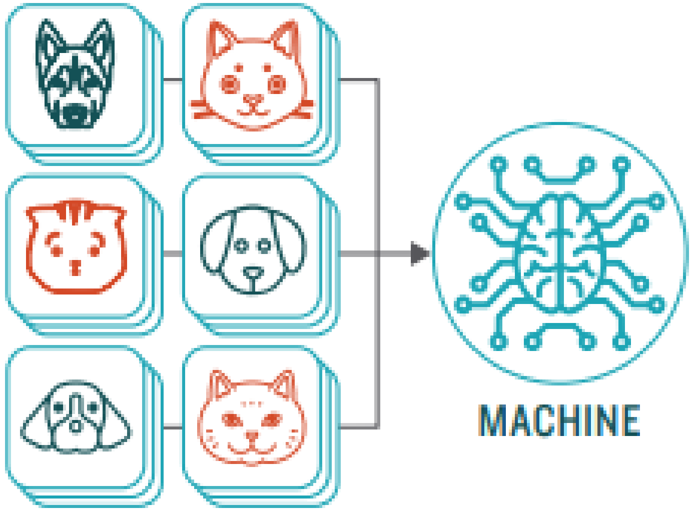
Classification is the task of predicting a discrete class label.

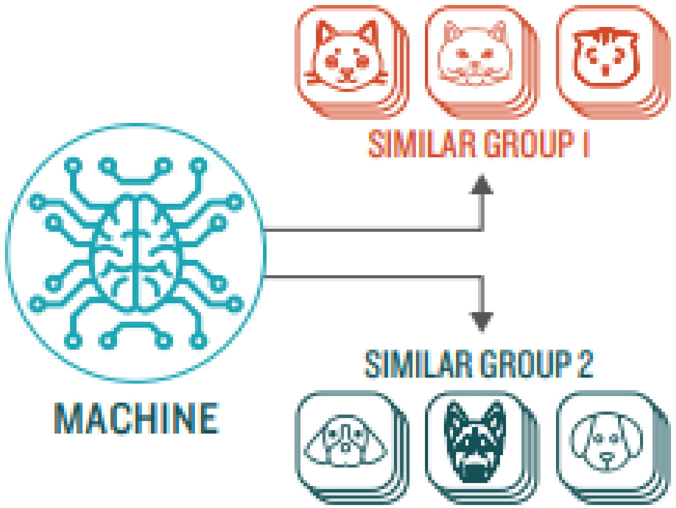
- Data is labeled into one of two or more classes.
- **Examples**:
  - Classifying a patient as at risk for chronic kidney disease or not
  - Labeling emails as spam or not spam.



Unsupervised Learning

- Step 1: Provide the machine learning algorithm **uncategorized, unlabeled input** data to see what pattern it finds
- Step 2: Observe and **learn from the patterns** the machine identifies



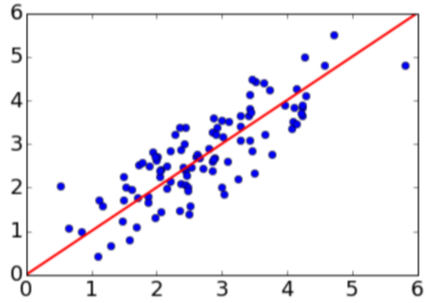


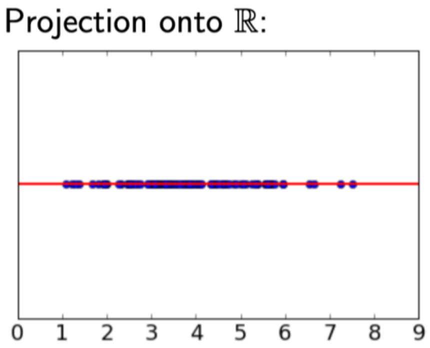
Dimensionality Reduction

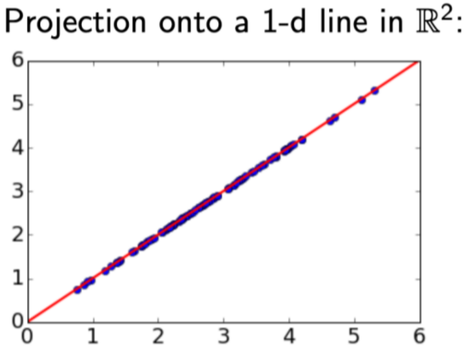
Dimensionality reduction is the process of reducing the number of input features in a dataset while retaining as much important information as possible.

- Dimensionality reduction often helps ML algorithms detect patterns in high-dimensional datasets.
- **Example**: A photograph reduces a 3-dimensional subject to a 2-dimensional representation while maintaining many important features.

Projecting 2-dimensional data onto a 1-dimensional line, preserving maximum variance in the data







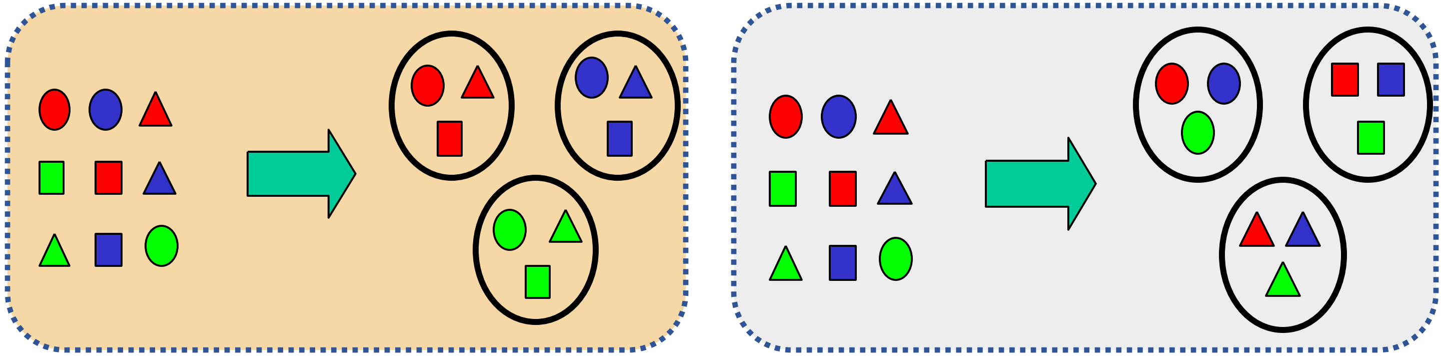
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Clustering

Clustering is the process of partitioning data into subsets (segments or clusters) such that data points most similar to one another are grouped together.

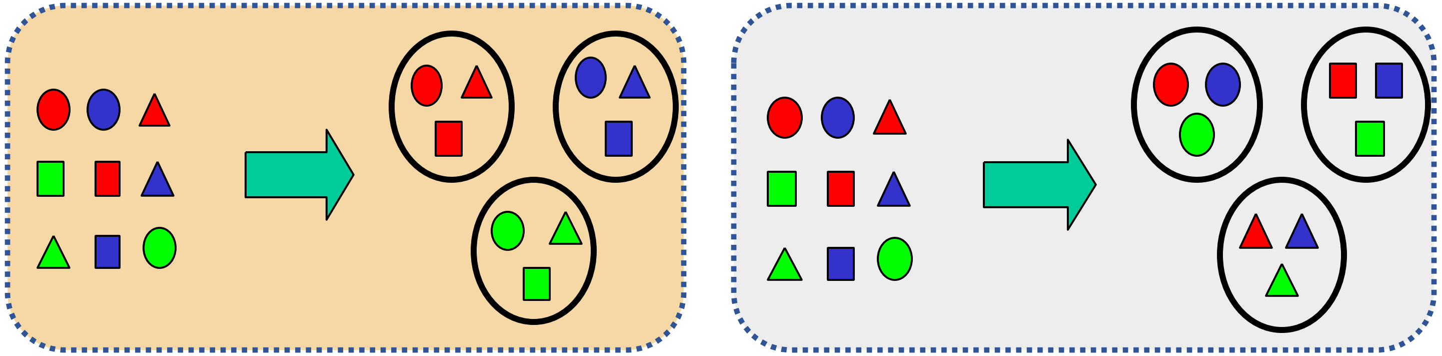
- The computer groups together data it sees as similar and separates dissimilar ones.
- Data scientists and SMEs work together to identify similar characteristics, patterns, or behaviors among the subsets identified by the algorithm.



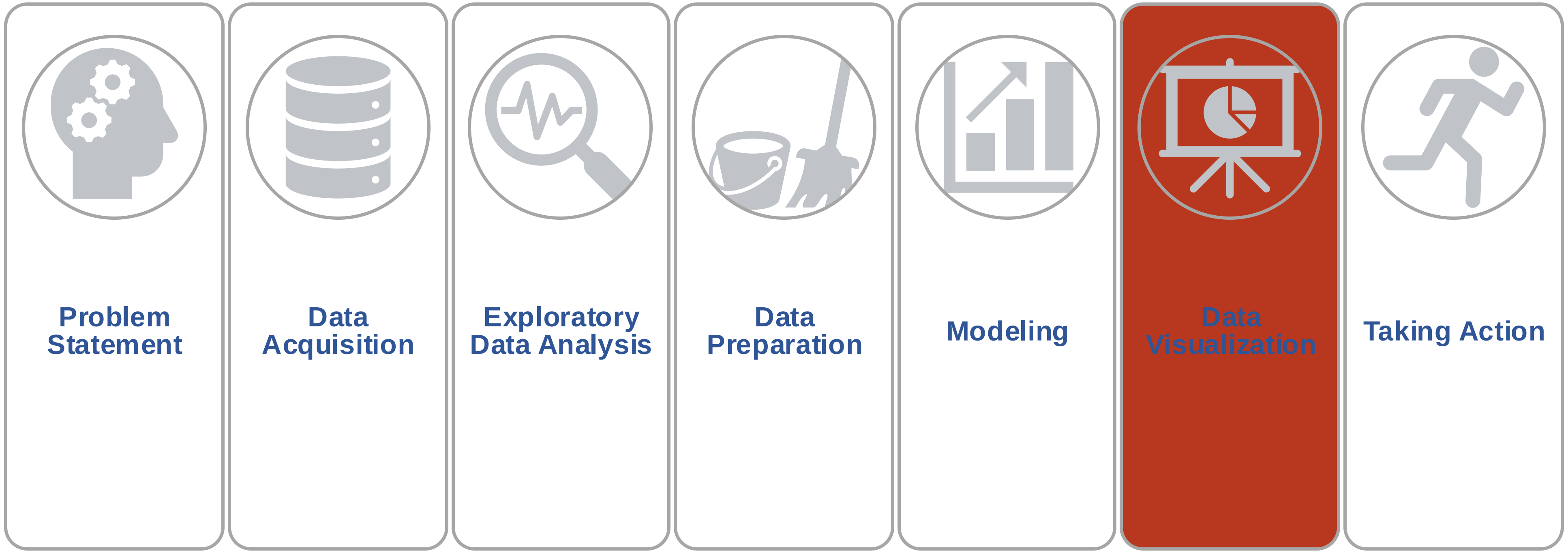
Problem Statement | Data Acquisition | Exploratory Data Analysis | Data Preparation | **Modeling** | Data Visualization | Taking Action

Clustering Challenges

- No prior knowledge of either the number or semantic meaning of the clusters.
- The same dataset can lead to different clusters.
  - Selecting different features can change the resulting clusters.



Visualizing and Communicating Results



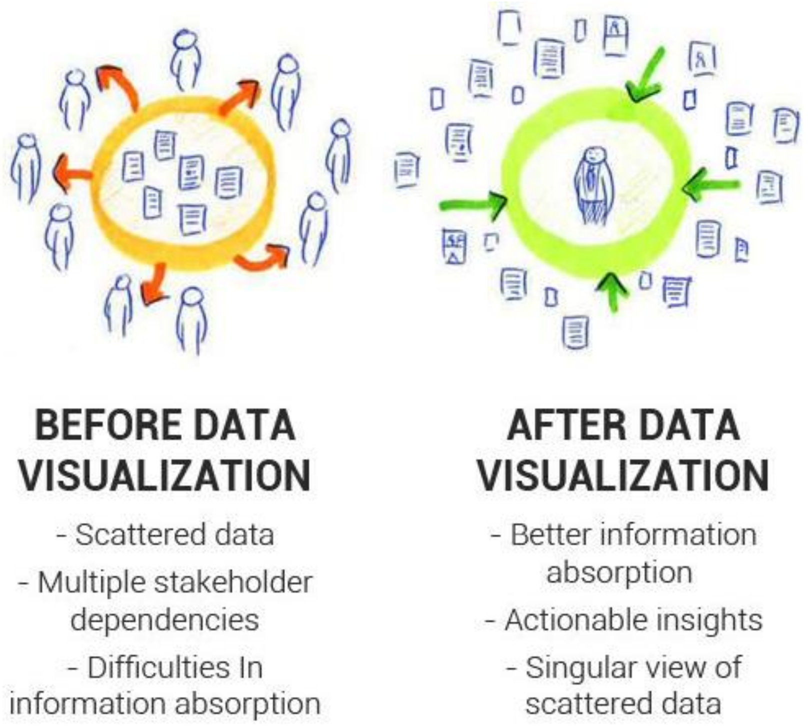
NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Why Visualization Is Important

- Visualizations can express aspects of the data that numbers alone cannot demonstrate

- They can tell a story about the results

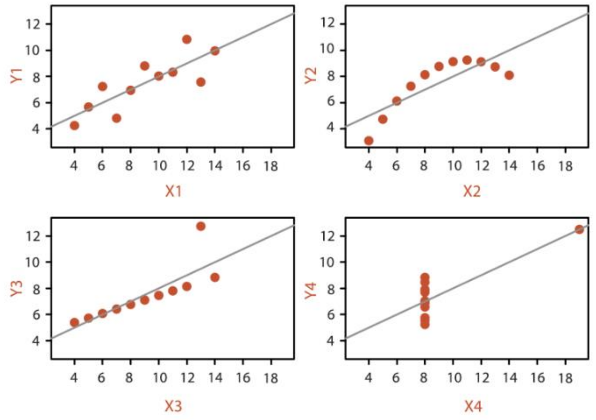


Aren’t Statistics Enough?

- All these plots have the same:
  - Mean
  - Variance
  - Correlation
- But looking at the visualization, you can see that they do not look anything alike
- Statistics can sometimes be misleading!
- Without effectively expressing the data, final results may be left up for interpretation

Anscombe's Quartet: Raw Data

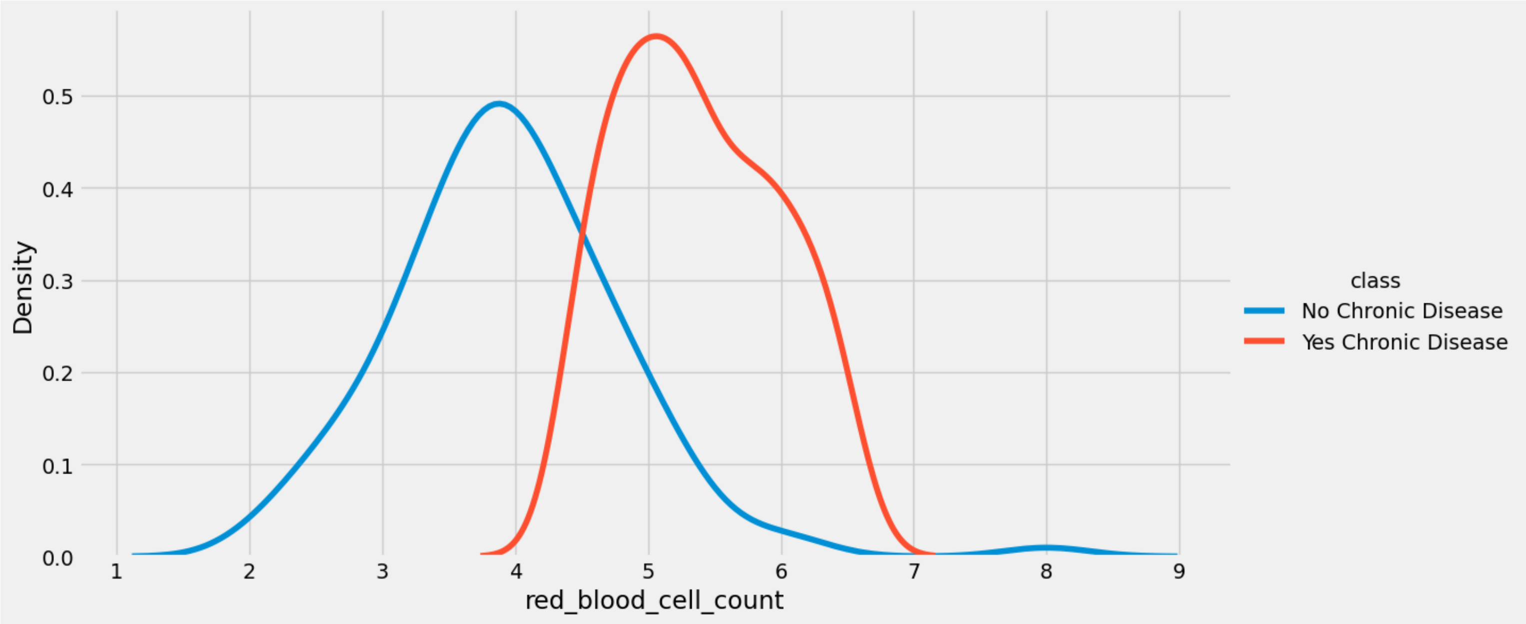
	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	



National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Example – Red Blood Cell Count Distribution

```python
import seaborn as sns

grid = sns.FacetGrid(df, hue="class", height = 6, aspect=2)
grid.map(sns.kdeplot, 'red_blood_cell_count')
grid.add_legend(labels = ['No Chronic Disease', 'Yes Chronic Disease'])
```



Problem Statement | Data Acquisition | Exploratory Data Analysis | Data Preparation | Modeling | **Data Visualization** | Taking Action

Visualization Tools: Open Source

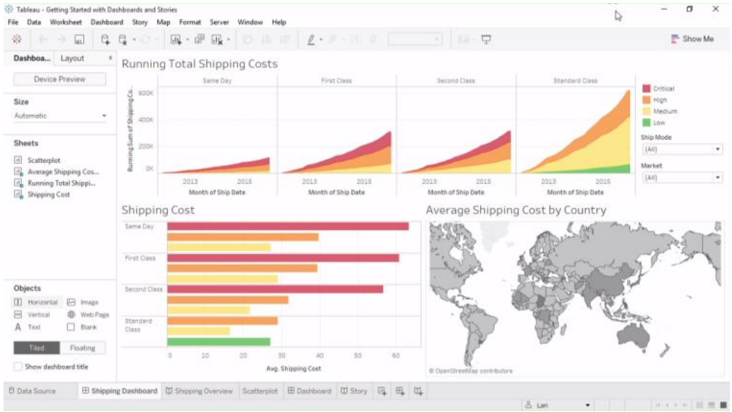
Language	R	python™	JS
Example of Library	ggplot2	matplotlib	D3

National Institute of Diabetes and Digestive and Kidney Diseases

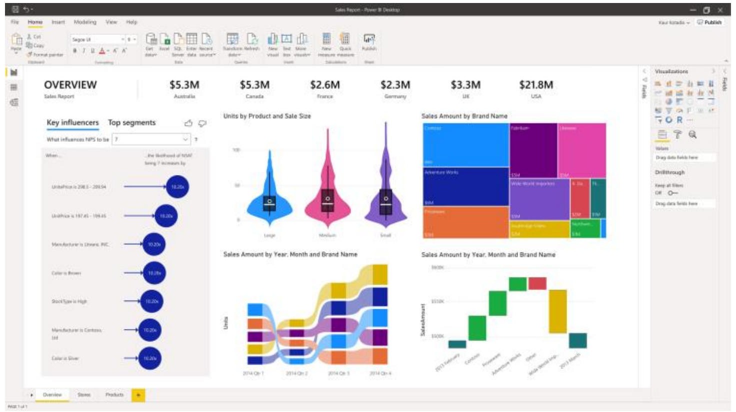
Central Repository

Visualization Tools: Premium

tableau



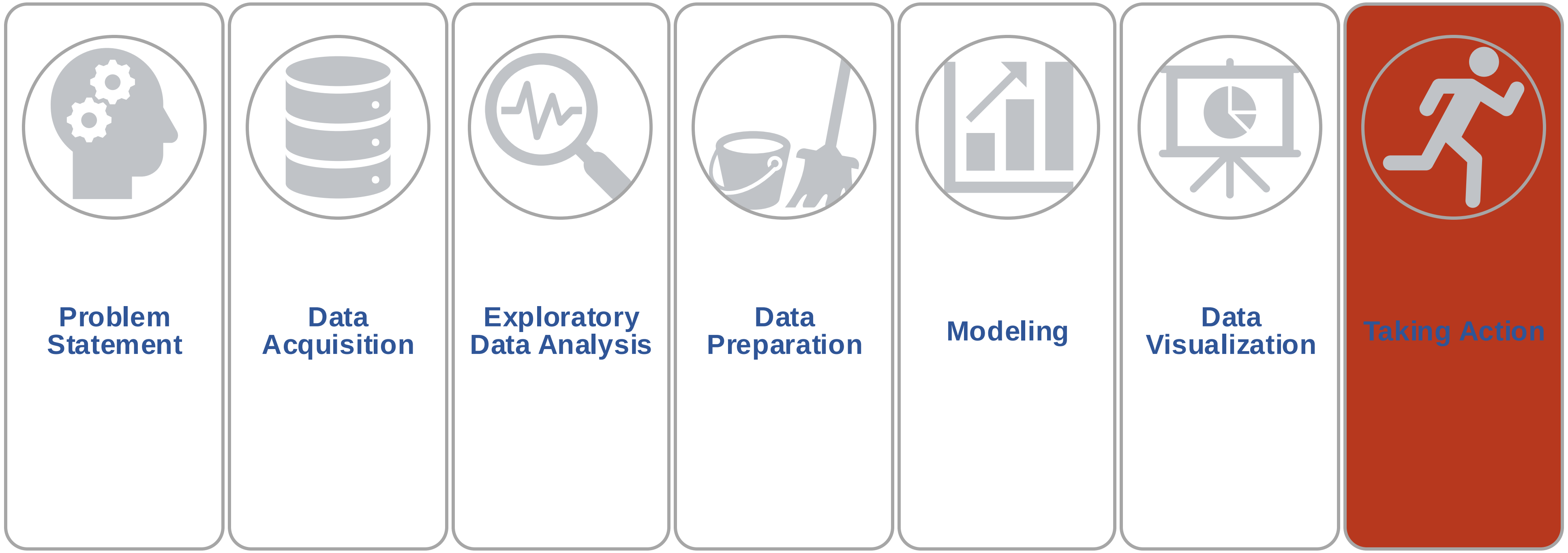
Power BI



Qlik

Problem Statement | Data Acquisition | Exploratory Data Analysis | Data Preparation | Modeling | Data Visualization | Taking Action

Taking Action



Measuring Model Performance

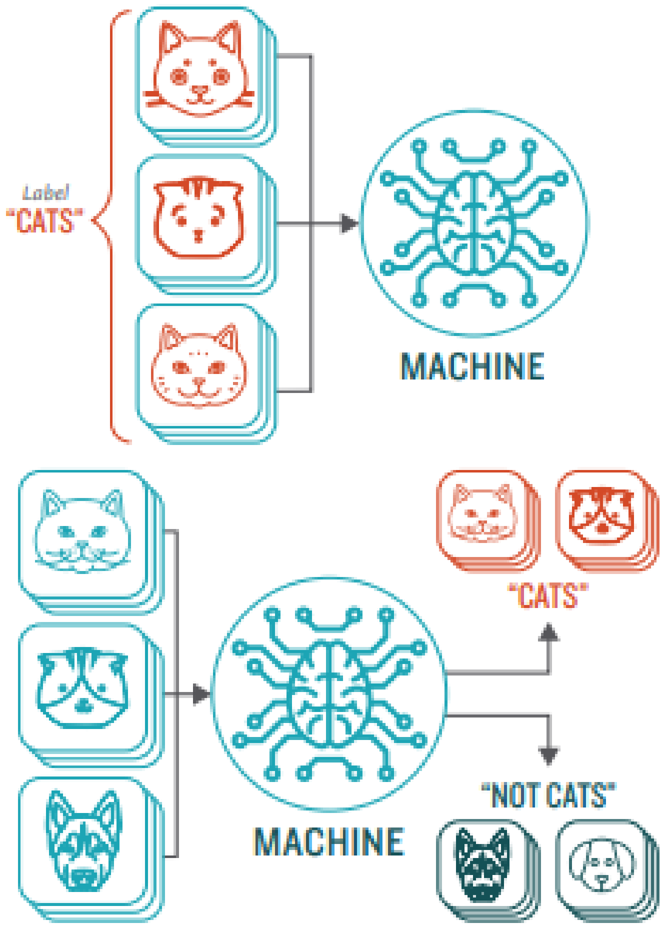
- Measuring a model's performance is important for users to be able to trust the model outputs
- Model performance not tracked over time can have direct and indirect adverse effects
- Ensure you are tracking appropriate metrics for the given model and dataset
  - Classification
    - Accuracy
    - Precision
    - Recall
  - Regression
    - Mean Absolute Error
    - Mean Squared Error
    - Root Mean Squared Error
    - R- squared

Machine Learning Deep Dive

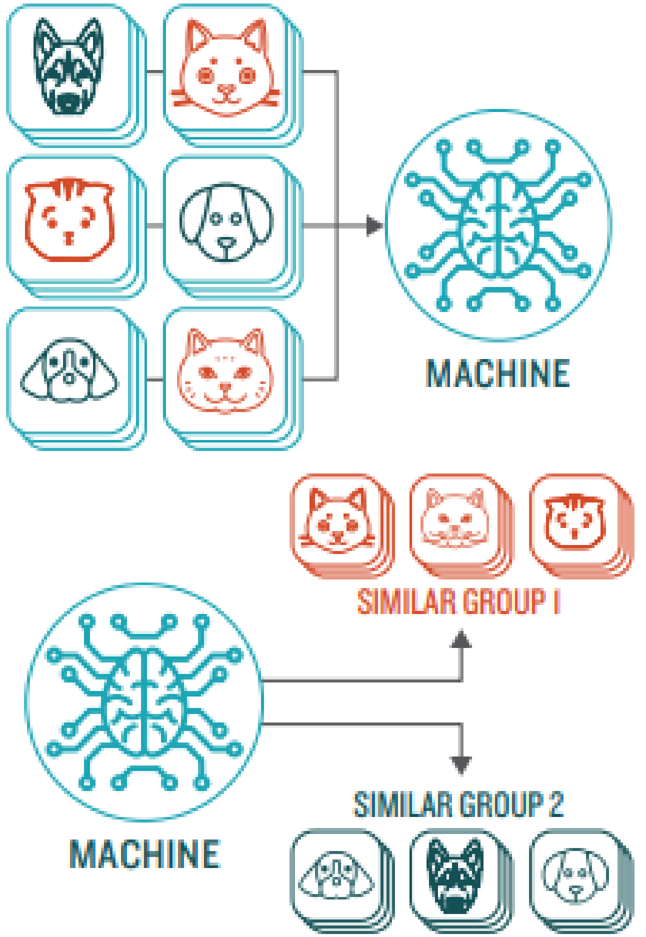
National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Supervised & Unsupervised Learning

Supervised: labeled data



Unsupervised: unlabeled data



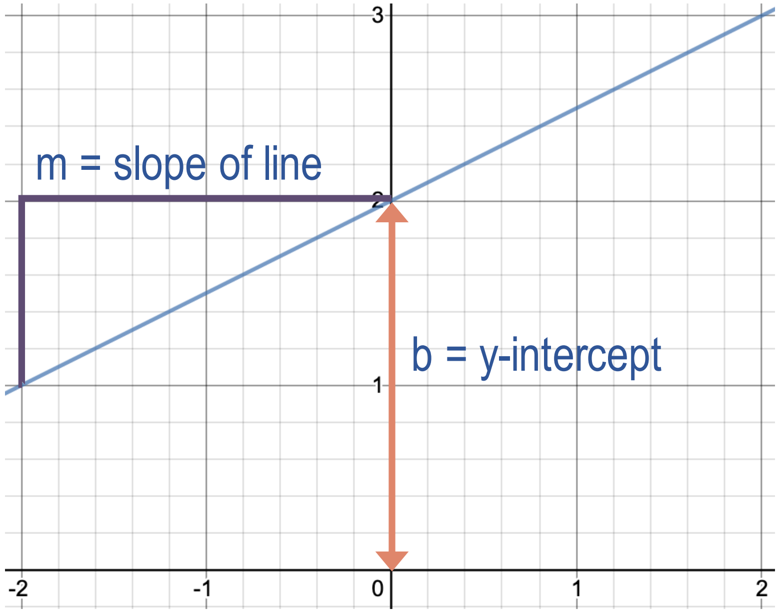
National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Linear Regression - Equation

$$y = \boxed{m}x + \boxed{b}$$

$$y = \boxed{\theta_1}X_1 + \boxed{\theta_0}$$

(linear regression)

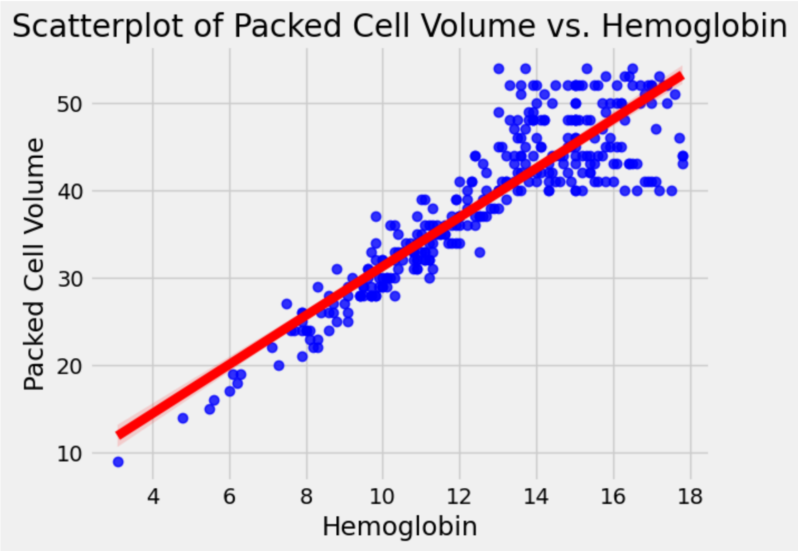


New notation to learn, but the same idea

Linear Regression - Parameters

How do you decide what your θ_1 and θ_0 (the coefficients/parameters) should be?

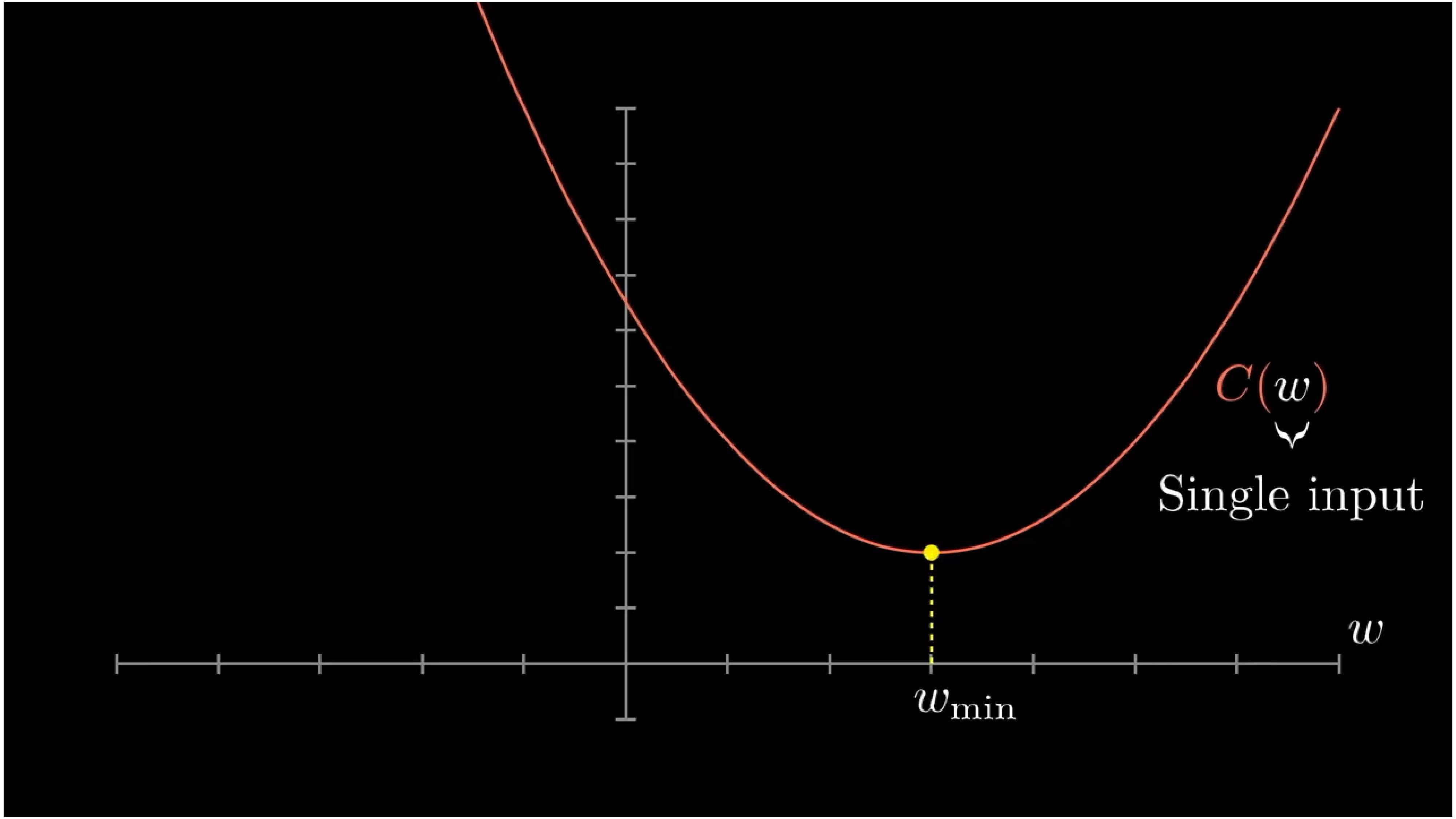
We could just draw a line that looks good to us…
But there's a better way to obtain the regression line of best fit.



Minimizing our cost function: **least squares estimation**

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Gradient Descent



Multiple Regression

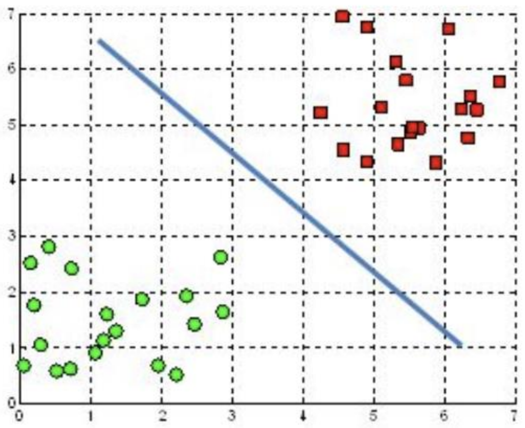
We can extend our regression model to include more features

$$y = \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_n X_n + \theta_0$$

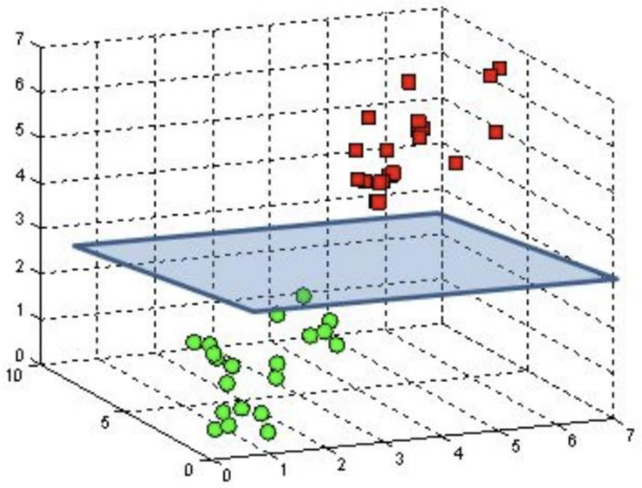
$$\theta_1 \times (\text{Red Blood Cell Count}) + \theta_2 \times (\text{Hemoglobin}) + \cdots$$

What would our model look like with more features?

A hyperplane in $\mathbb{R}^2$ is a line



A hyperplane in $\mathbb{R}^3$ is a plane



Source: deepai.org

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Summary: Linear Regression

Pros:

- Simple
- Easy to interpret
- Computationally inexpensive

Cons:

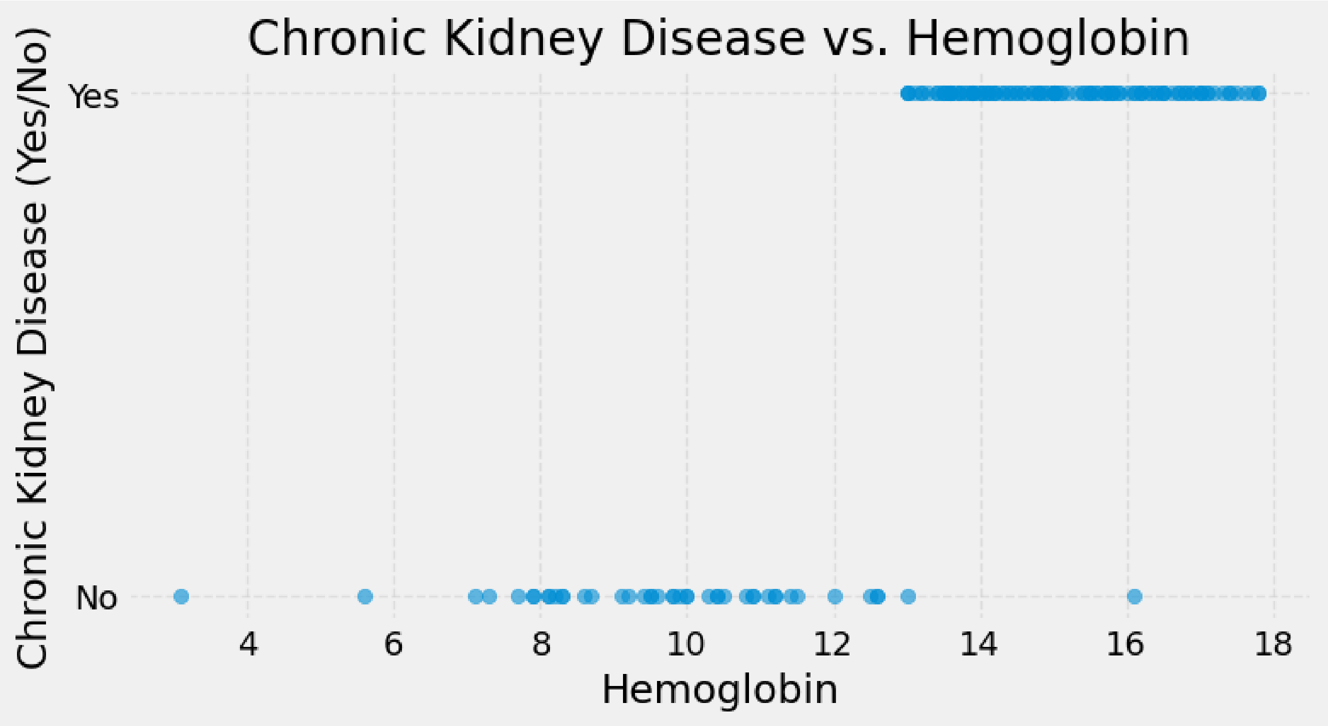
- Oversimplifies many real-world problems
- As the name implies, assumes a linear relationship between model parameters and dependent variables
- Sensitive to outliers

From Linear to Logistic Regression

- We used *linear regression* to predict on a quantitative continuous data type
- What if we want to predict a category of data?
  - Nominal – Named categories
  - Ordinal – Categories with implied order
  - Discrete – Finite values
- Using *logistic regression,* we can predict classes of objects.
- What type of data might be involved in predicting whether a patient has chronic kidney disease?

Example - Kidney Disease Detection

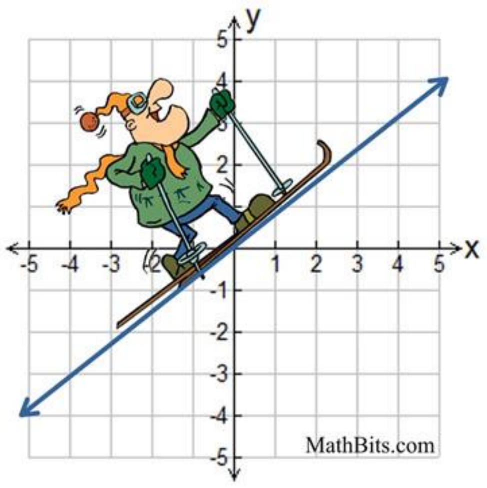
Your data might look something like this:

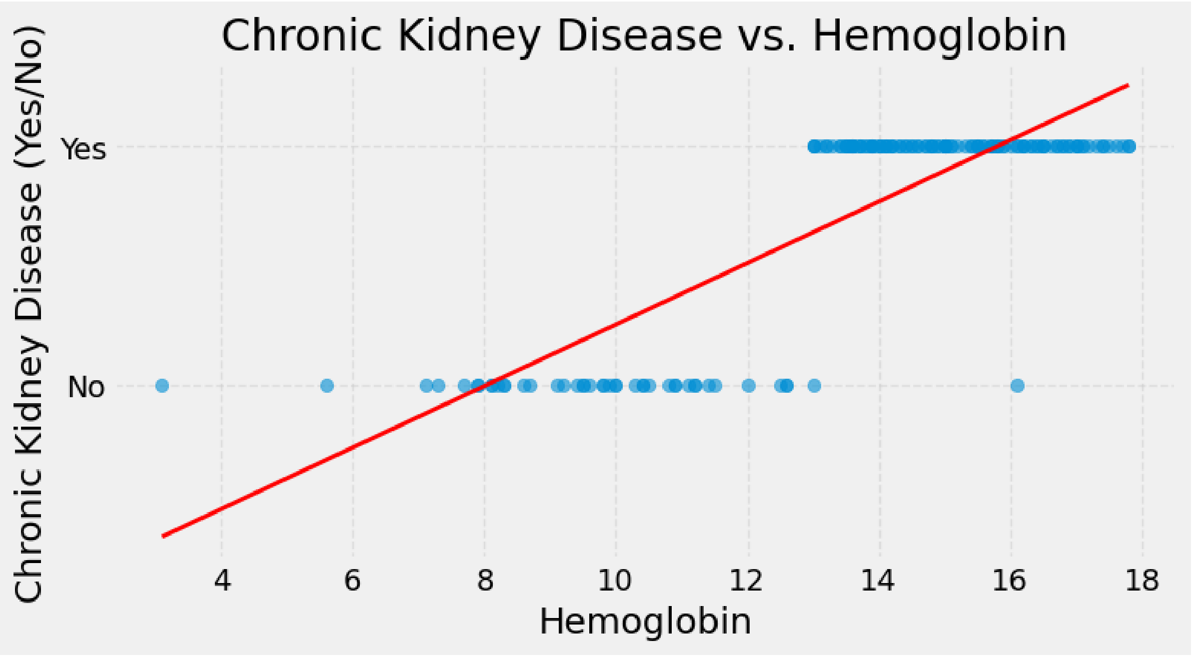


How would you define boundaries to model this data?

Maybe Something Like This?

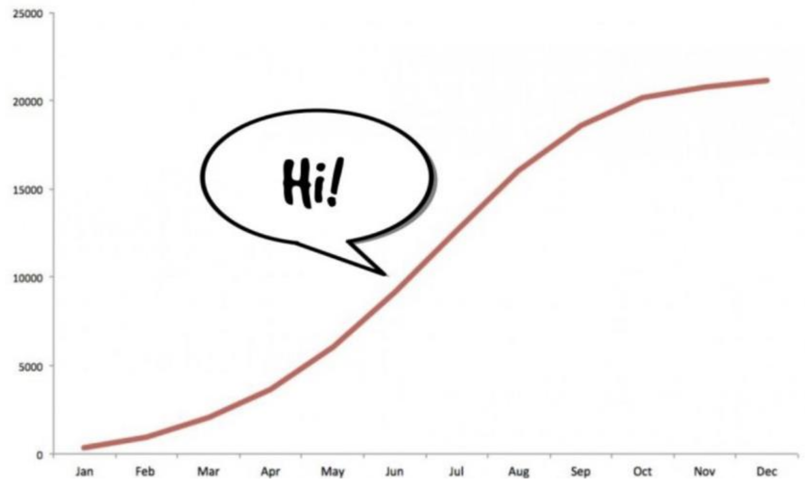
- Here is our line of best fit (linear regression):
- What's wrong with this model?

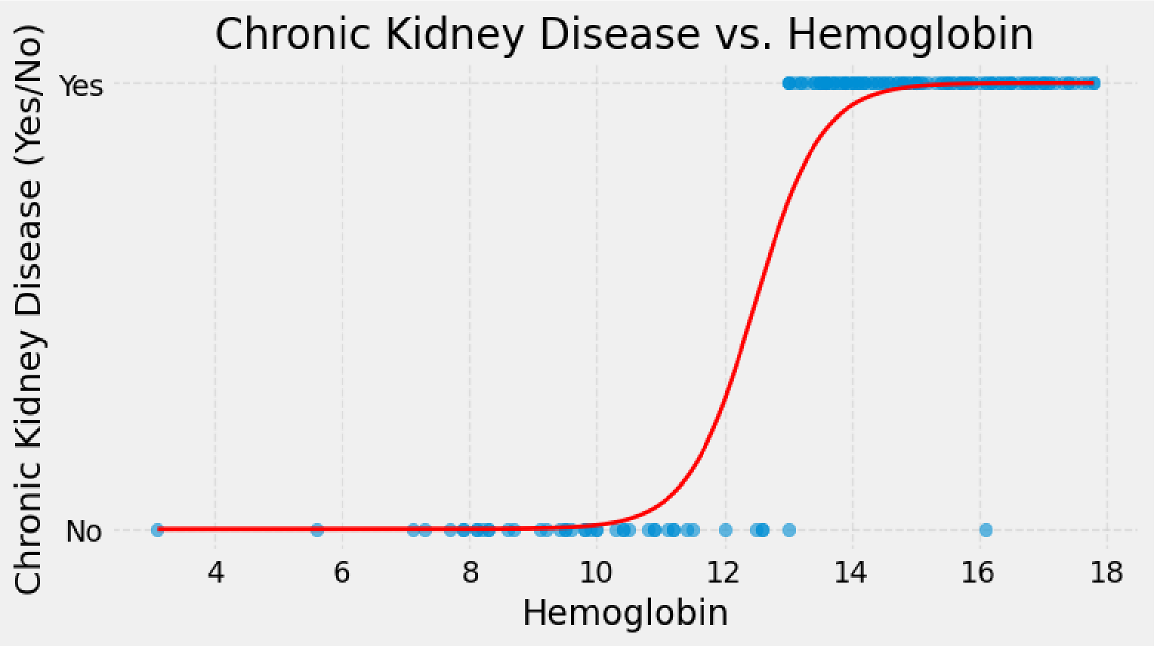




Or Something Like This?

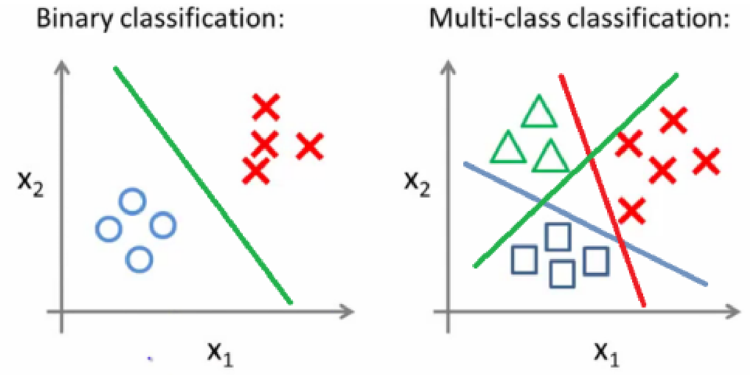
- As you may have ascertained, this is not a linear regression problem.
- An ideal boundary might look more like this:

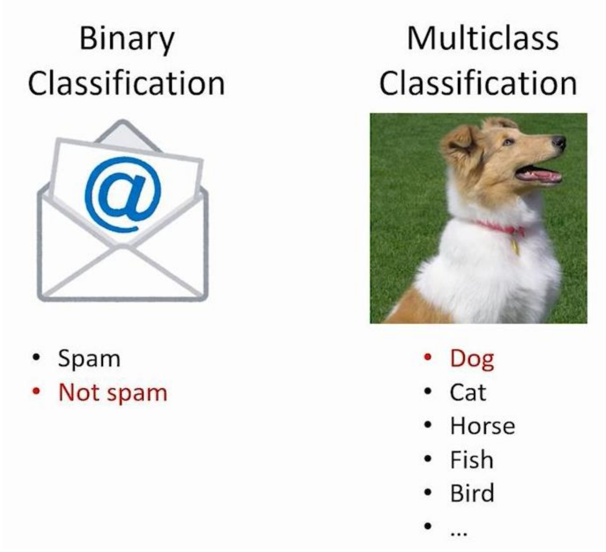




Logistic Regression

- A statistical method for analyzing a dataset that has one or more independent variables that determine an outcome
- Simplistic algorithm, but often makes a good baseline model
- Used to predict a binary outcome (1/0, Yes/No, True/False)
- Allows us to create models for classification problems:
  - What animals are in this image?
  - Is this email spam?
  - Disease predicted?



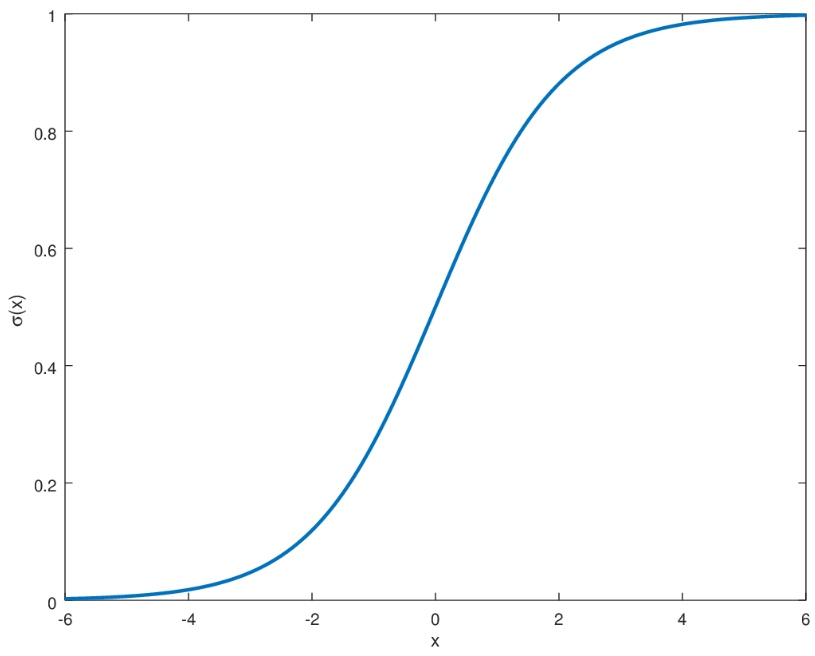


Logistic Regression - Equation

- Logistic regression is just linear regression with one additional step

$$y = \theta_1 X_1 + \theta_0$$

$$\sigma(y) = \frac{1}{1 + e^{-y}}$$



All values become constrained between 0 and 1

Summary: Logistic Regression

Pros:

- Easy to interpret
- Quick to train
- Provides probabilities as outputs

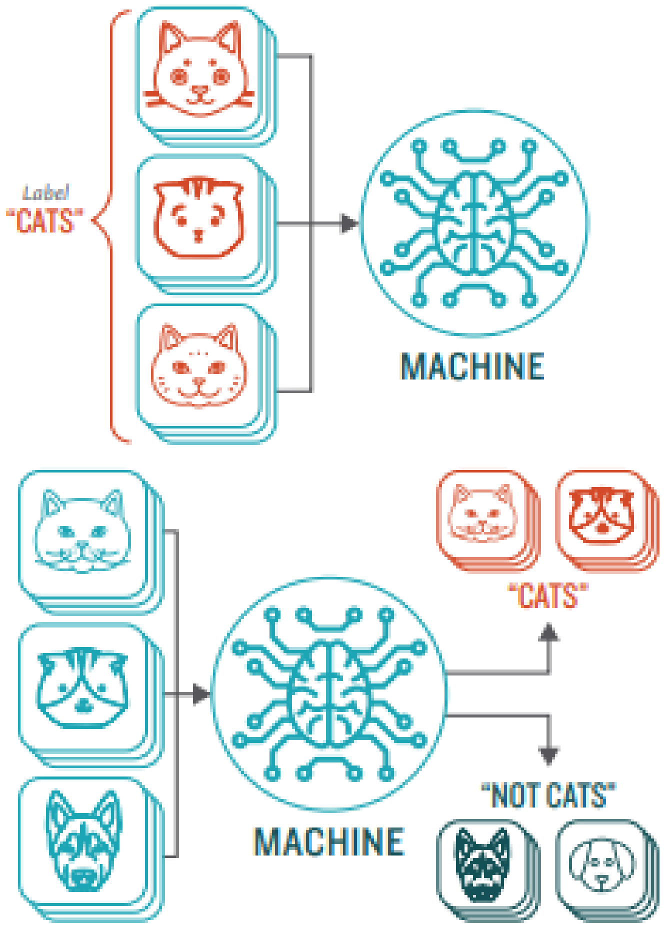
Cons:

- Poor performance in large feature spaces
- Cannot handle large amounts of categorical variables well
- In practice, it's typically only applied to binary classification outcomes

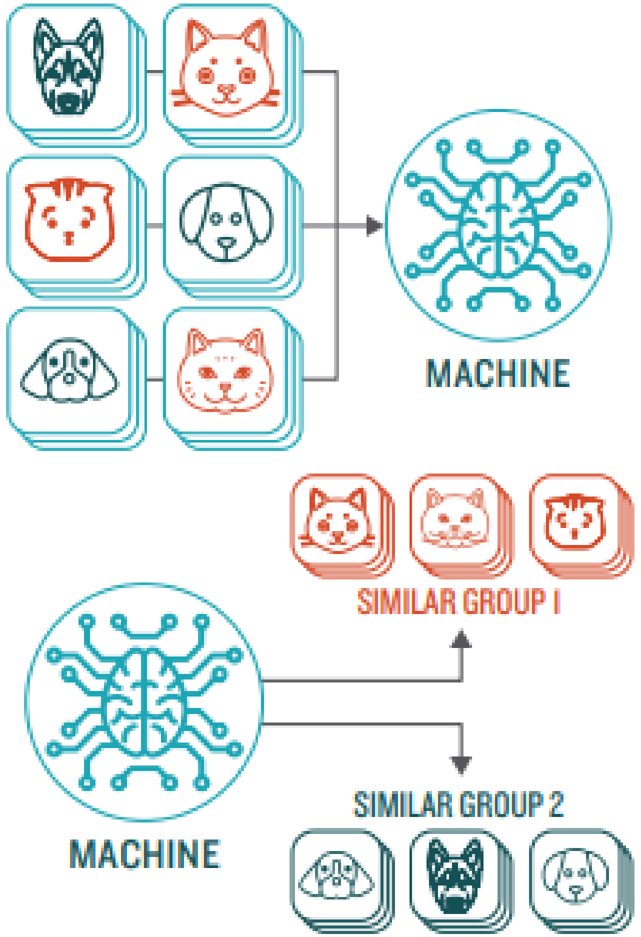
National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Supervised & Unsupervised Learning

Supervised: labeled data



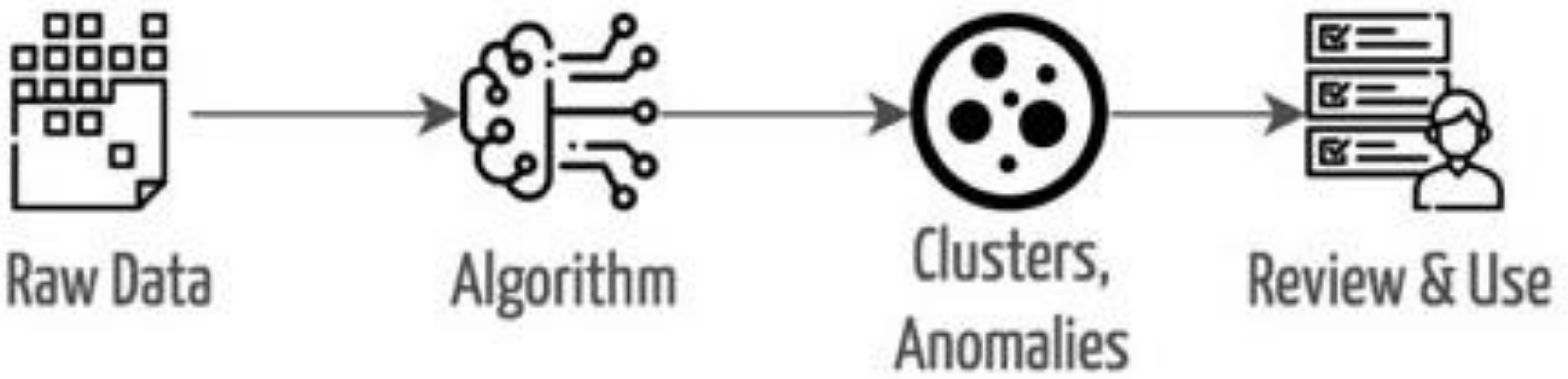
Unsupervised: unlabeled data



National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Unsupervised Learning - Overview

Uncovering inherent structures, patterns, and relationships hidden in collections of unlabeled data



What Do You Do with Unlabeled Data?

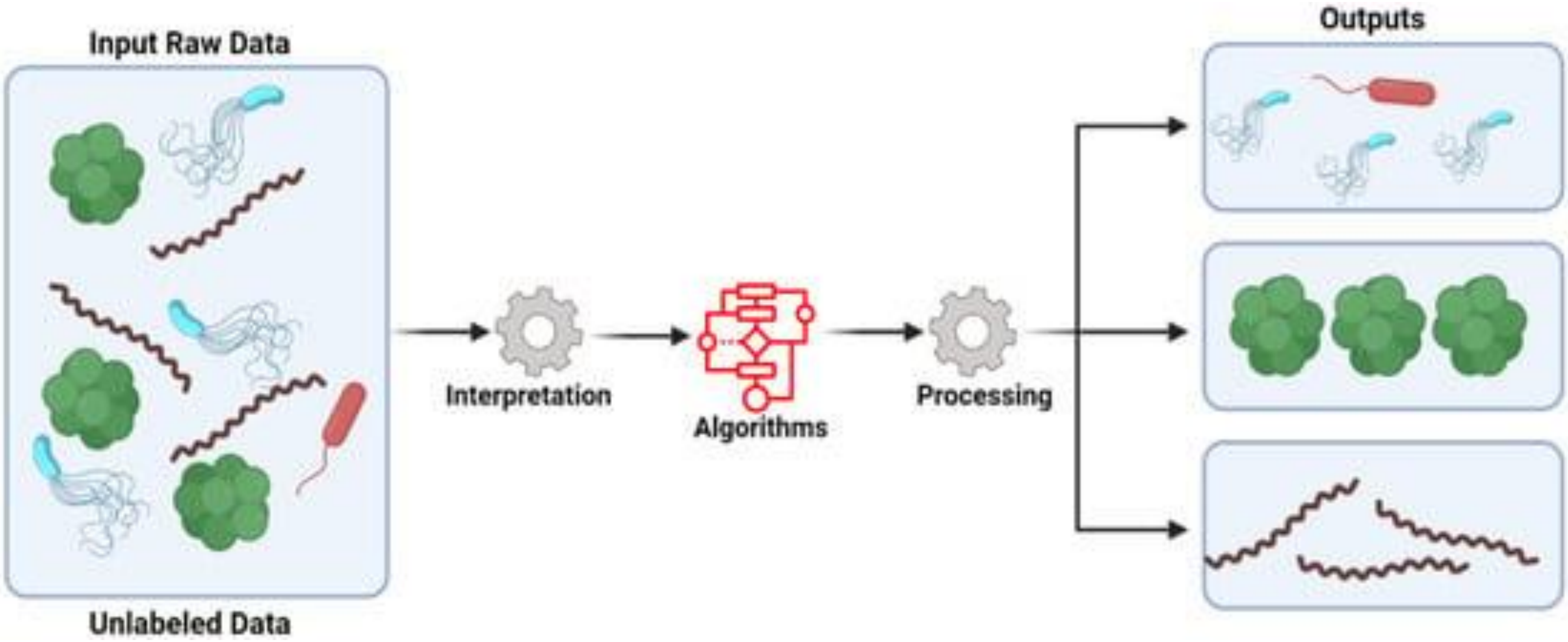
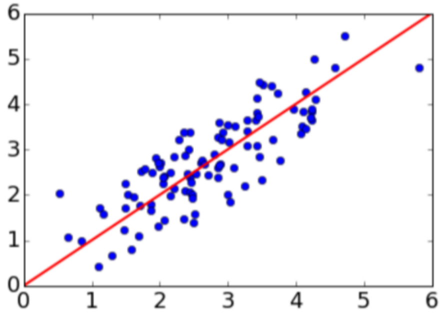


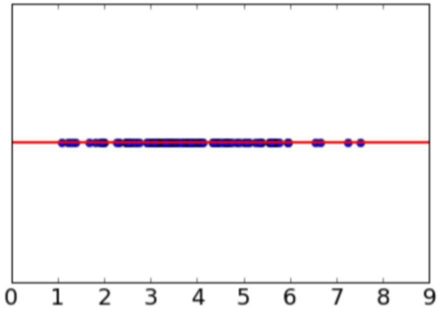
Image Source: [Unsupervised Learning in Precision Medicine | mdpi.com]

Refresher: Dimensionality Reduction

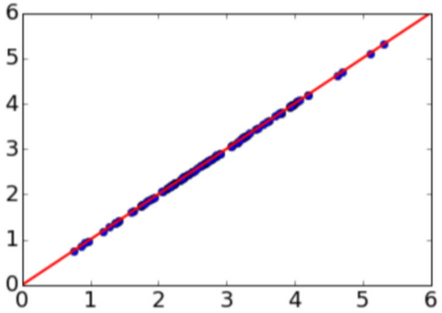
Dimensionality reduction is simply the process of reducing the dimensions of your feature set



Projection onto $\mathbb{R}$:



Projection onto a 1-d line in $\mathbb{R}^2$:



National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Example – Kidney Disease Data

For each patient you find, you have data on the following features:

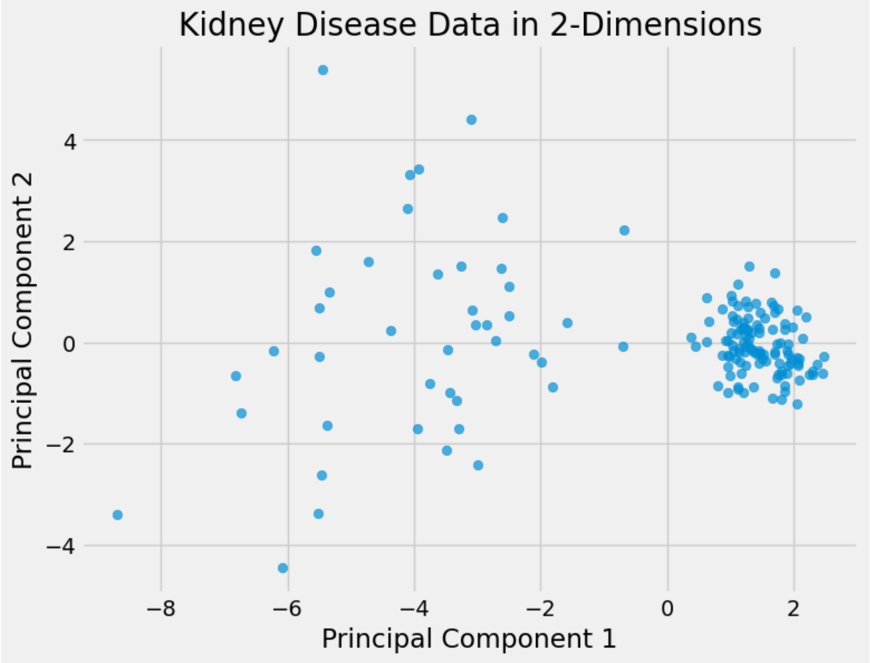
age = Age	sod = Sodium
pot = Potassium	hemo = Hemoglobin
pcv = Packed Cell Volume	wc = White Blood Cell Count
rc = Red Blood Cell Count	htn = Hypertension
dm = Diabetes Mellitus	cad = Coronary Artery Disease
appet = Appetite	pe = Pedal Edema
ane = Anemia	bp = Blood Pressure
sg = Specific Gravity	al = Albumin
su = Sugar	rbc = Red Blood Cells
pc = Pus Cell	pcc = Pus Cell Clumps
bgr = Blood Glucose Random	bu = Blood Urea
sc = Serum Creatinine	classification = Chronic Disease (Yes/No)

Principal Component Analysis (PCA)

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

We have a lot of features in our data, so it can be difficult to make sense of the data in this form. We can use principal component analysis (PCA) to reduce our data to two dimensions, which is a great way to visualize feature-rich data.

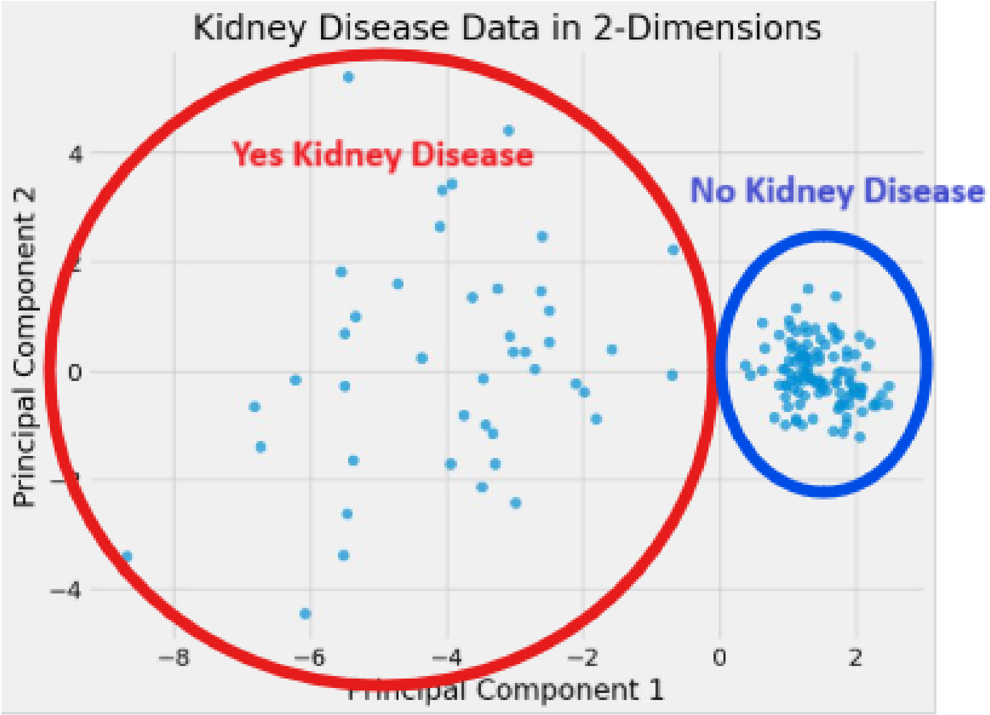
age = Age	sod = Sodium
pot = Potassium	hemo = Hemoglobin
pcv = Packed Cell Volume	wc = White Blood Cell Count
rc = Red Blood Cell Count	htn = Hypertension
dm = Diabetes Mellitus	cad = Coronary Artery Disease
appet = Appetite	pe = Pedal Edema
ane = Anemia	bp = Blood Pressure
sg = Specific Gravity	al = Albumin
su = Sugar	rbc = Red Blood Cells
pc = Pus Cell	pcc = Pus Cell Clumps
bgr = Blood Glucose Random	bu = Blood Urea
sc = Serum Creatinine	classification = Chronic Disease (Yes/No)



But what are these two dimensions now?
What are these PCA “components” on the X and Y axis?

Which Patients Have Kidney Disease?

- We need a way to assign labels (kidney disease yes/no) to our data
- Clustering techniques (like k-means) provide a possible solution



National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Summary: PCA

Pros:

- Prevents overfitting
- Removes correlated features
- Speeds up other machine learning algorithms
- Improves visualization

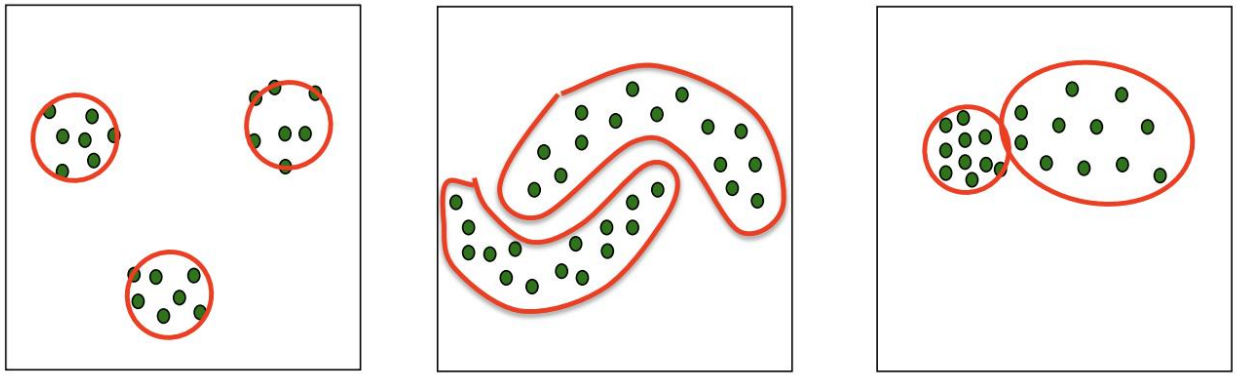
Cons:

- Difficult to interpret new components
- Can lead to losing information
- Computationally expensive

NIH National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

K-means Clustering

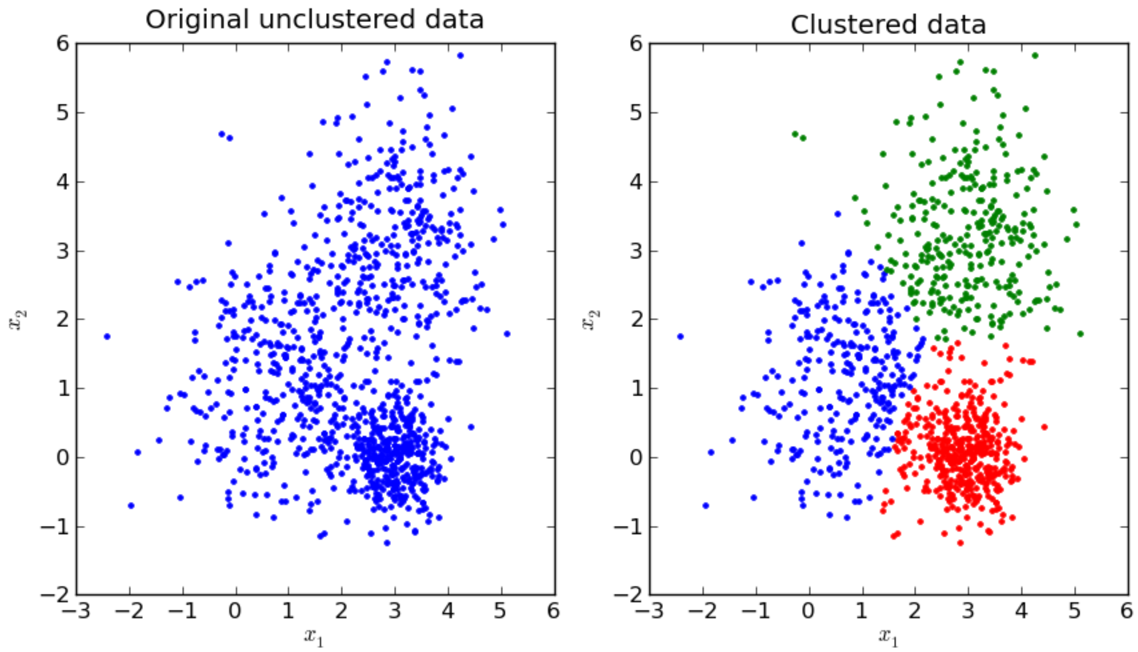
- Clustering describes the process of grouping data into shared characteristics
- The characteristics of a group may vary by data
- K-means takes a data sample as input and outputs the cluster that the new data point belongs to, according to the training that the model went through



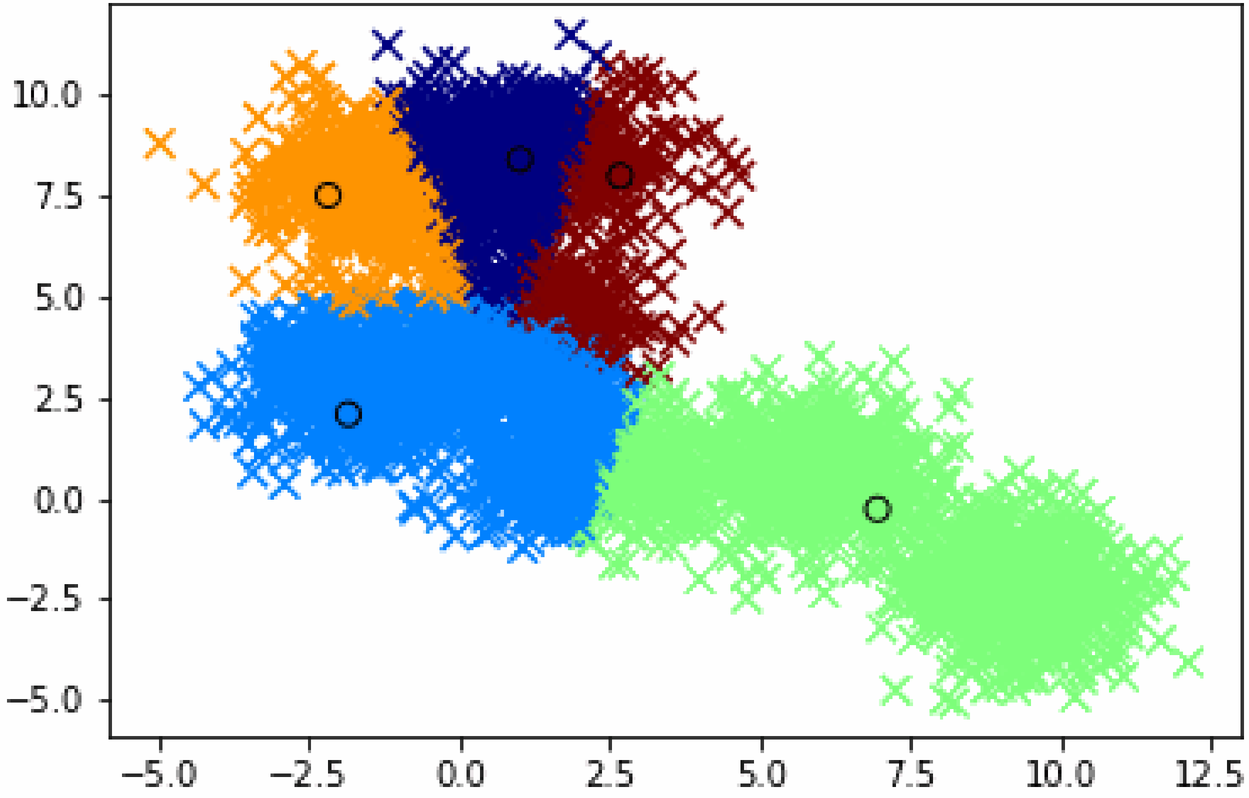
Location Shape Density

How Does K-means Work?

- An iterative process of clustering or finding groups of data in our dataset that are similar to one another
- Iterates until it reaches the best solution of clusters in our problem space



How Does K-means Work?

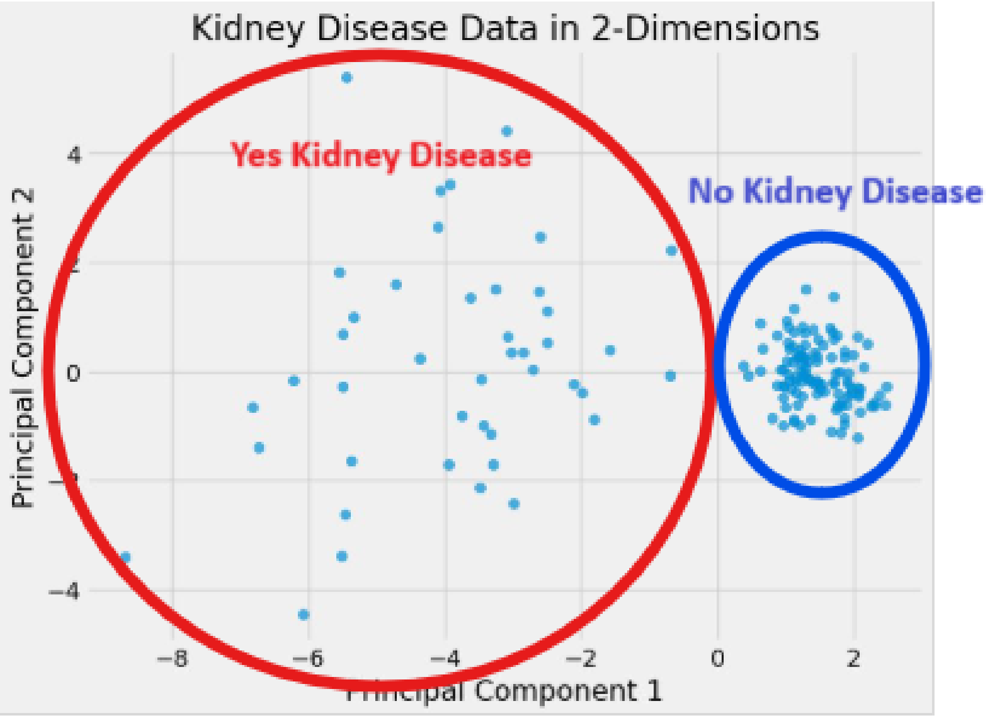


1. Choose *k* data points to be the initial centroids (cluster centers)
2. Assign each data point to the closest centroid
3. Re-calculate the centroids using the average of the assigned points
4. Iterate (repeat) over steps 2 & 3 until the centroids no longer move (converge)

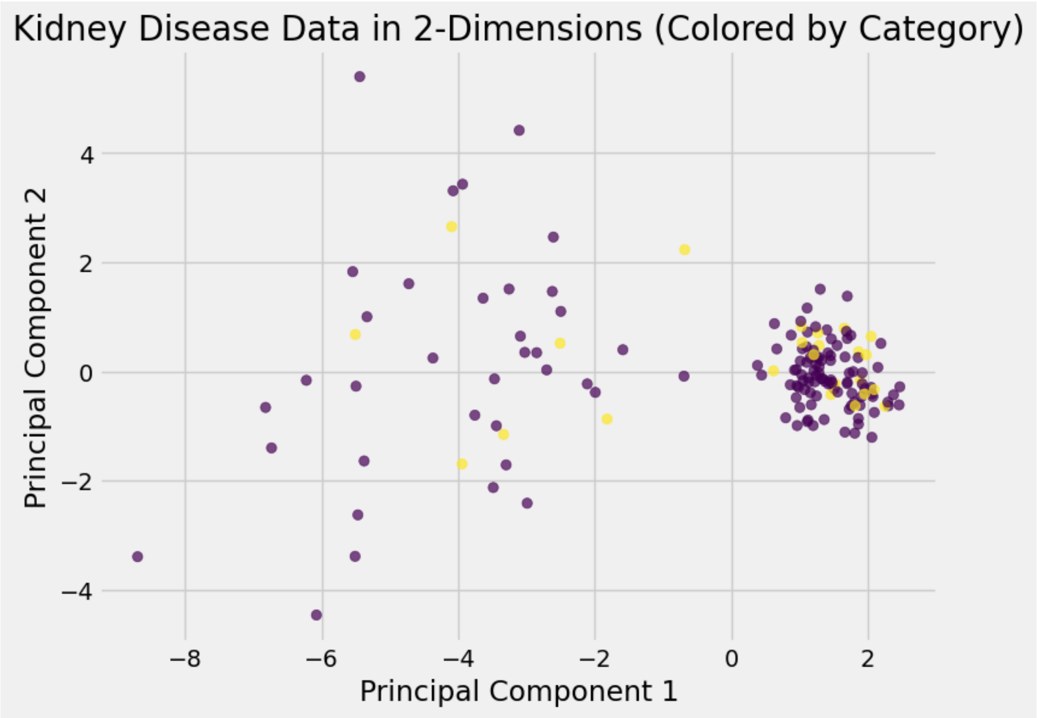
National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Not a Perfect Solution…

Predicted Unsupervised Labels



Actual Labels



Summary: K-means

Pros:

- Easy to understand and implement
- Computationally inexpensive
- Guarantees convergence

Cons:

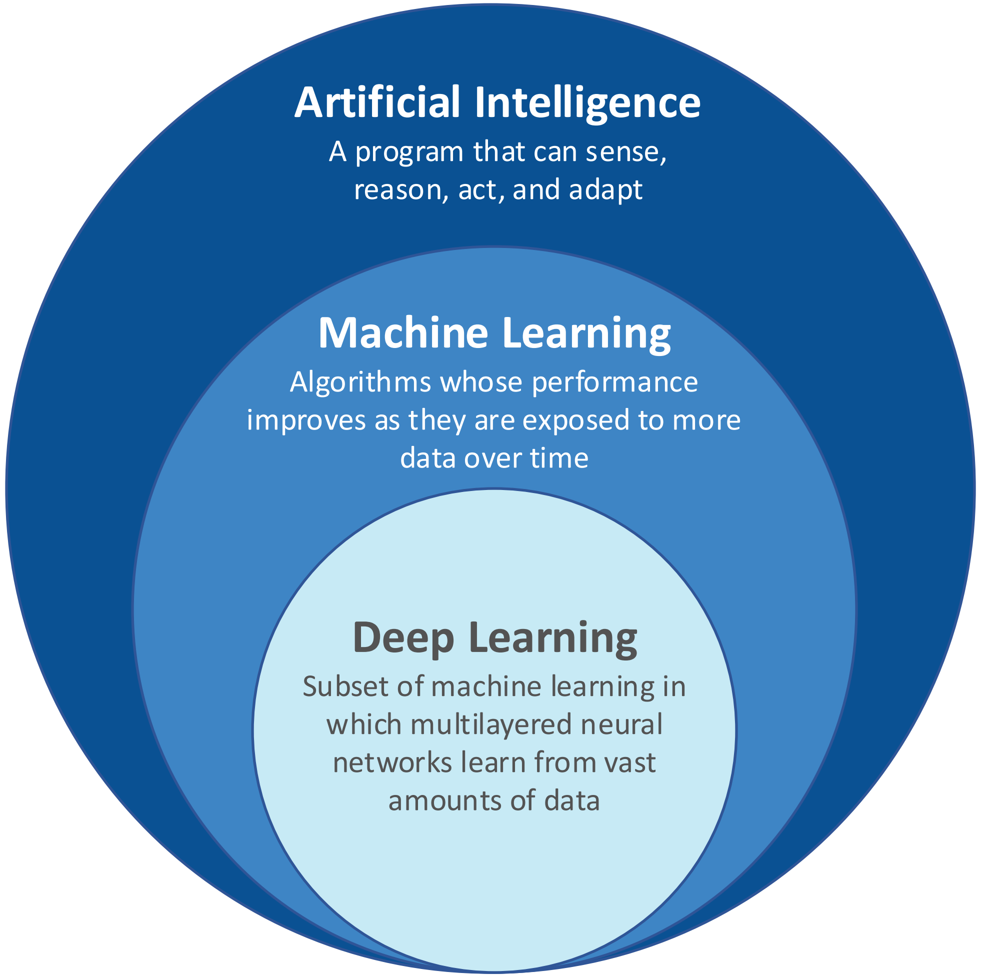
- Results are highly variable and dependent on initial values
- Sensitive to outliers
- Struggles with data of varying sizes and densities

Stretch Break

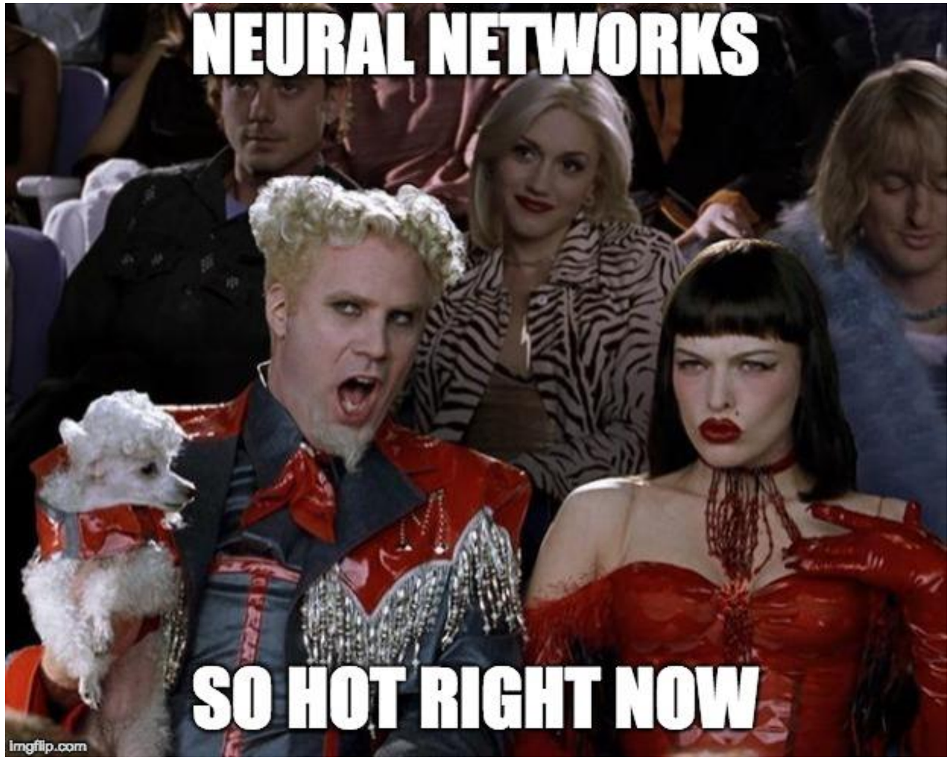
Office of Clinical Research Support

Deep Learning Overview

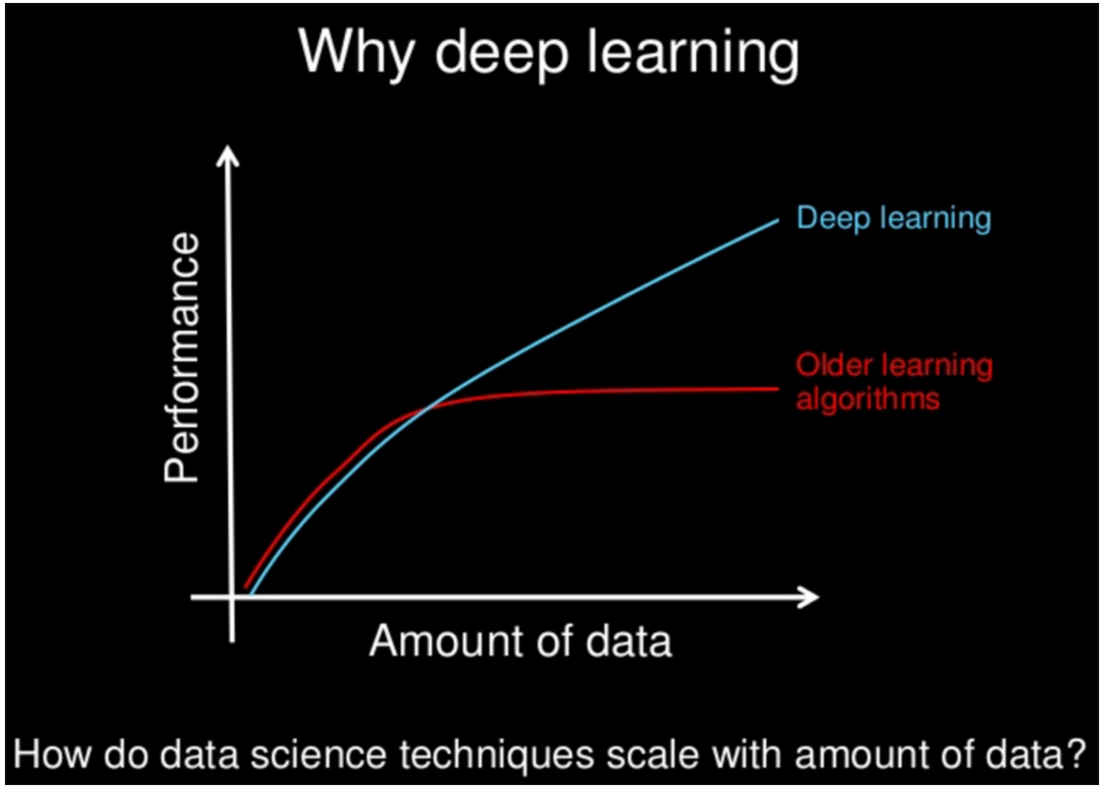
Deep Learning – Subset of Machine Learning



Neural Networks



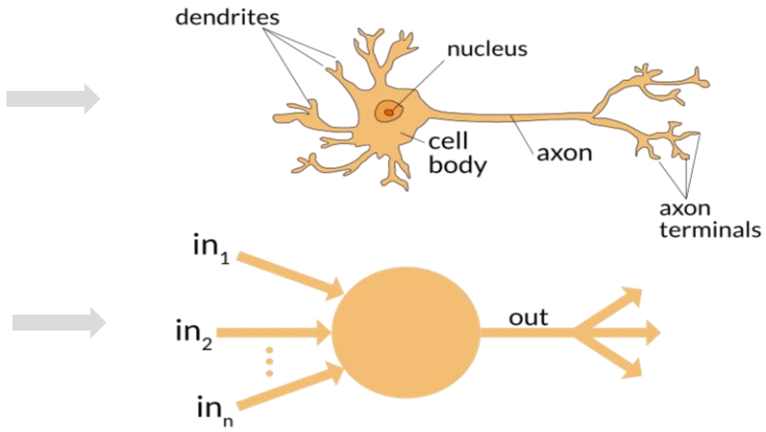
Why is Deep Learning So Hot?

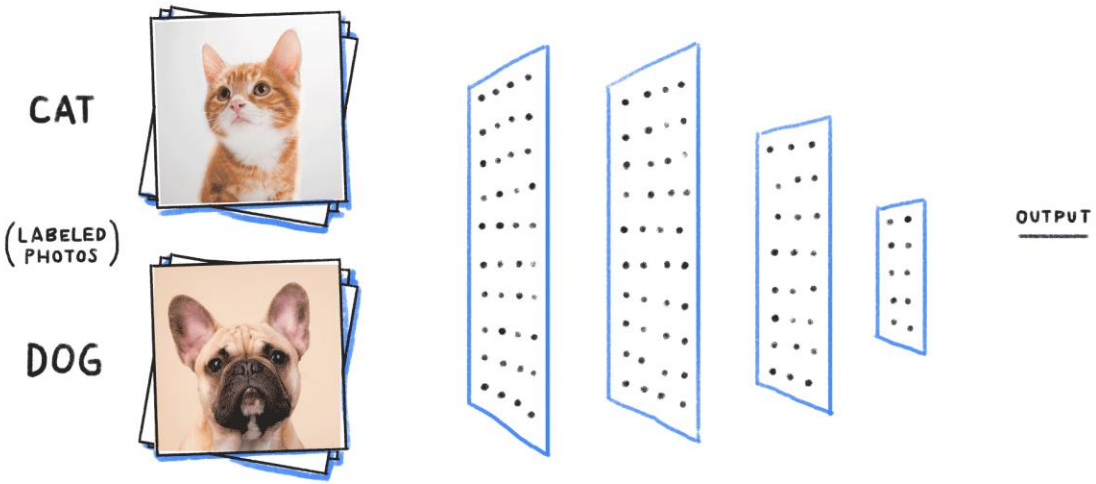


Deep Learning Inspired Party from Biology

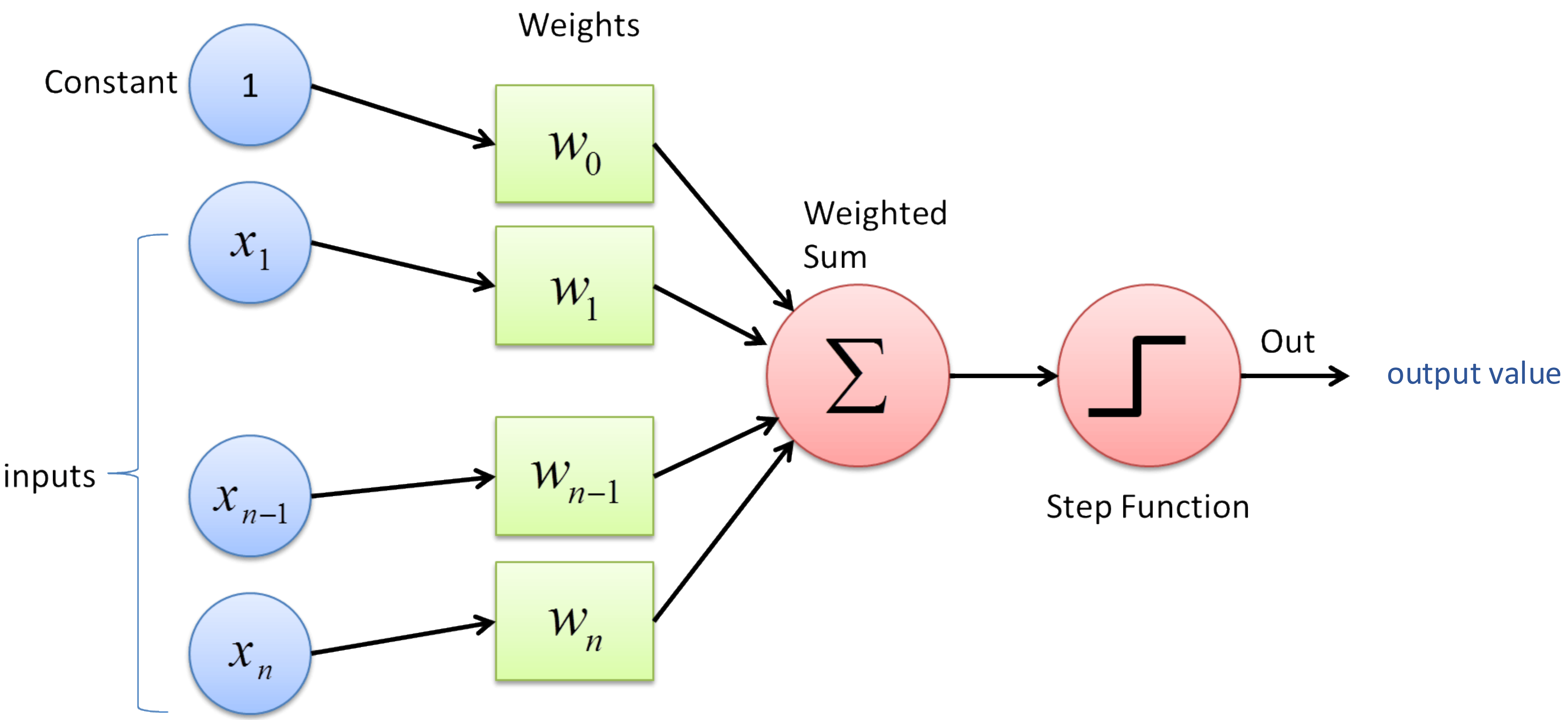
Human neuron

Neural network node



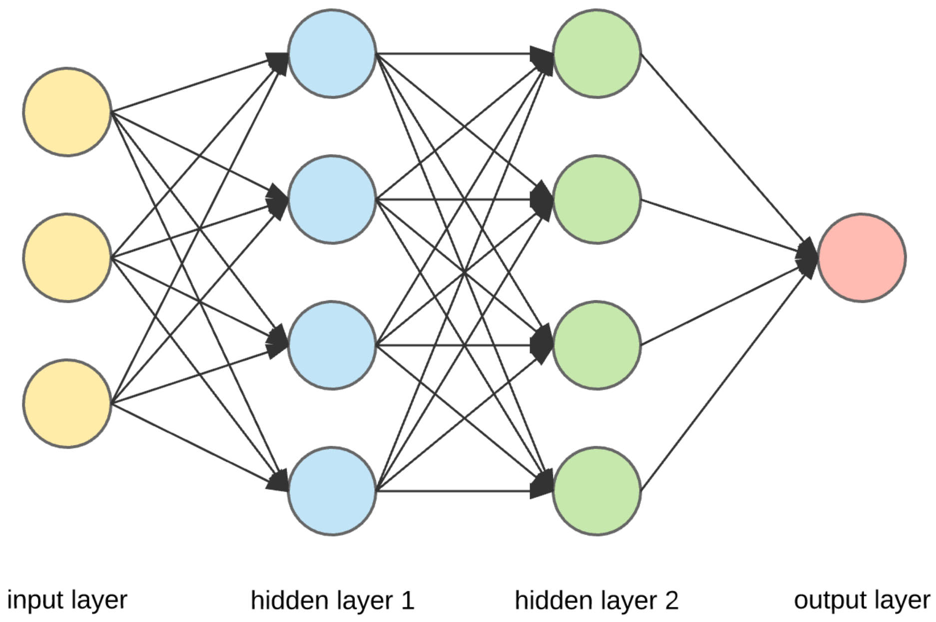


The Perceptron (A Single "Neuron")



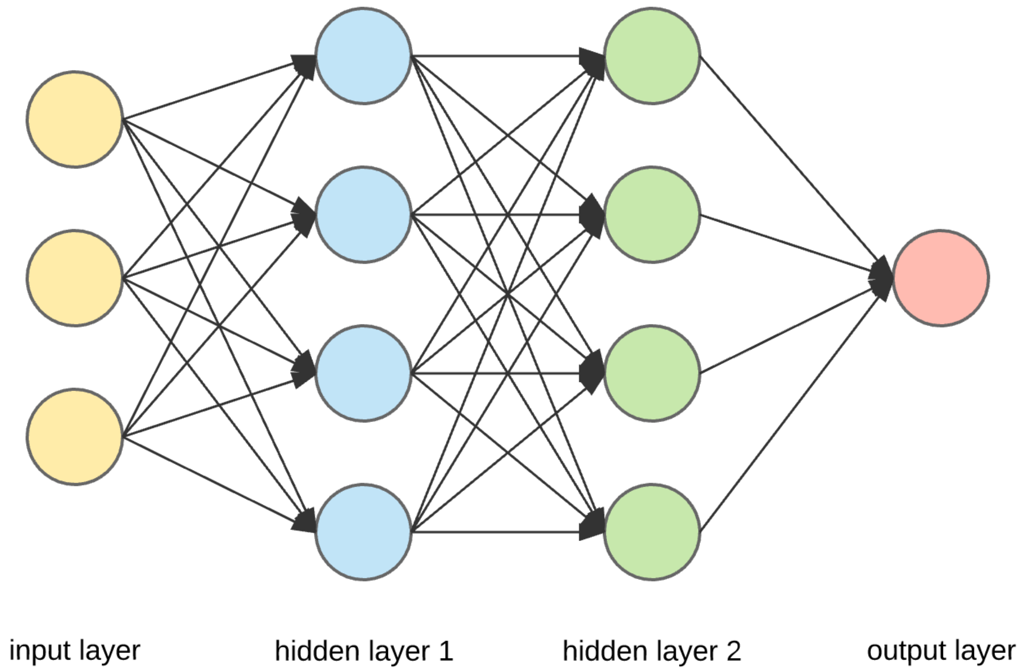
"Feed Forward" Neural Networks

- This is the most basic, vanilla form of neural network that all other neural networks use as a foundation.
- Number of layers and nodes/neurons per layer is a choice made by network's architect(s).
- Each node is essentially a perceptron.



How Neural Networks Learn

- Like we've seen previously, neural networks can use gradient descent to find ideal weights.
- But, they have a special trick called **backpropagation** to calculate the gradients.
- Backpropagation leverages the chain rule, though this goes beyond the scope of this class.



Neural Networks in Python

TensorFlow

PyTorch

Keras

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Summary: Neural Networks

Pros:

- Able to capture more complexity in a model
- Widely applicable to real-world business problems
- Once trained, predictions are fast

Cons:

- Computationally expensive to train
- Needs lots of data
- Can require lots of parameter tweaking and retraining
- Has a "black box" nature

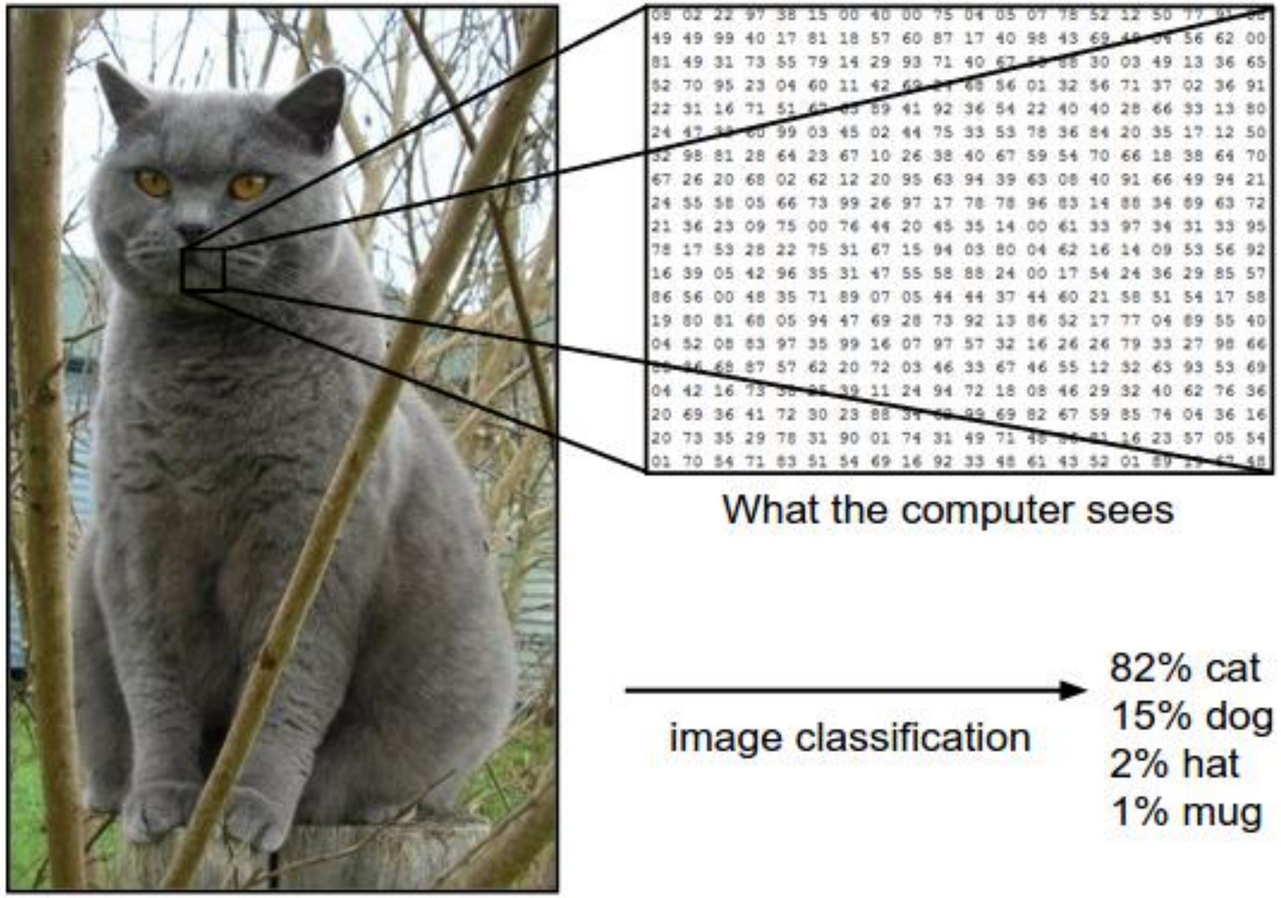
Specialized DS Topics

Specialized Topics in Data Science

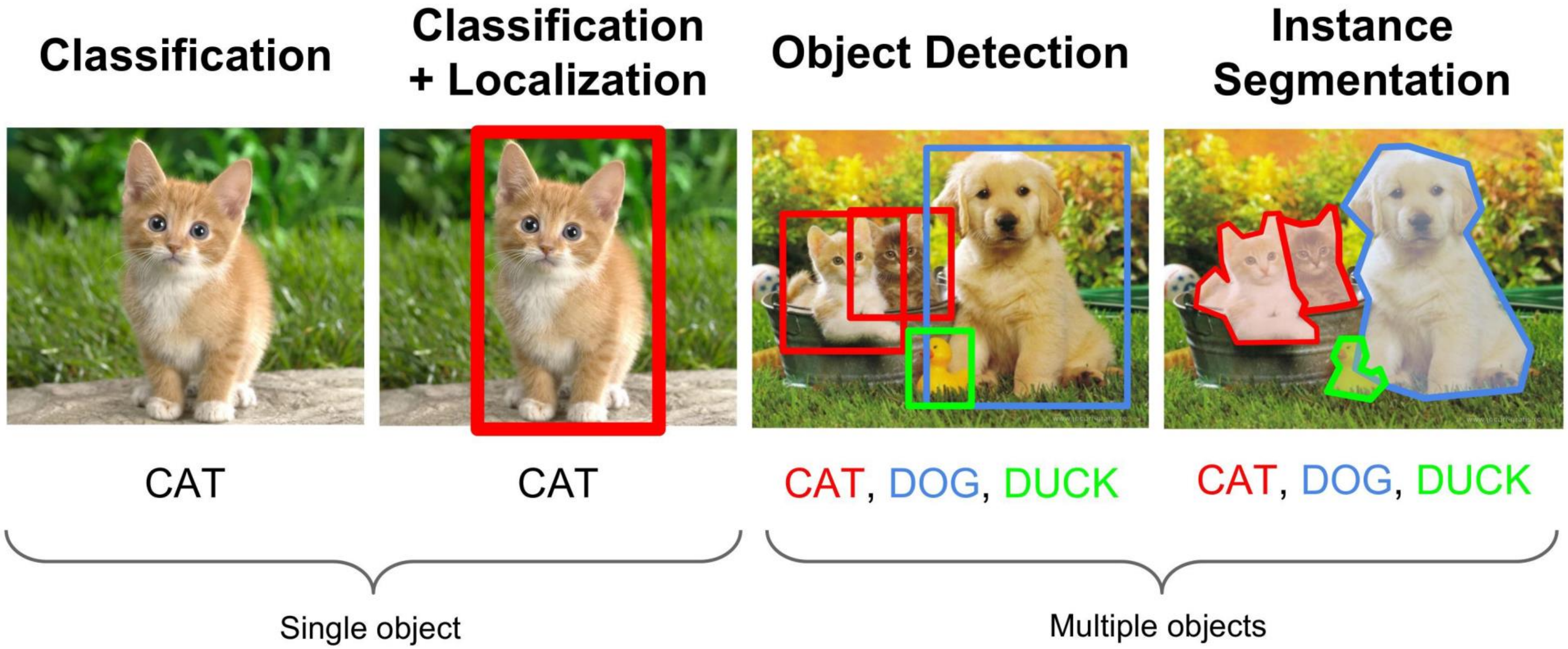
- **Computer Vision**: Interdisciplinary subfield of AI that enables interpretation and understanding of digital images or videos.
- **Time Series Analysis**: Modeling a sequence of data over an interval of time.
- **Natural Language Processing (NLP)**: Interdisciplinary subfield of AI that enables interpretation and understanding of natural language data (e.g., text and speech).

Computer Vision

How Does a Computer Understand Image Data?



Computer Vision Tasks



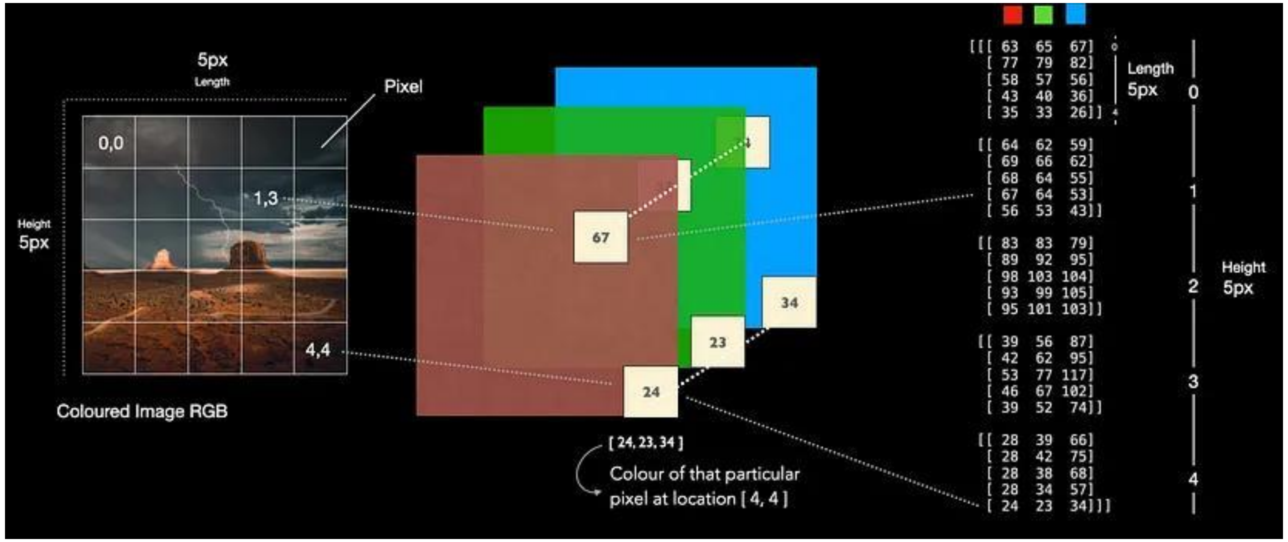
Computer Vision: MNIST Dataset

The problem:

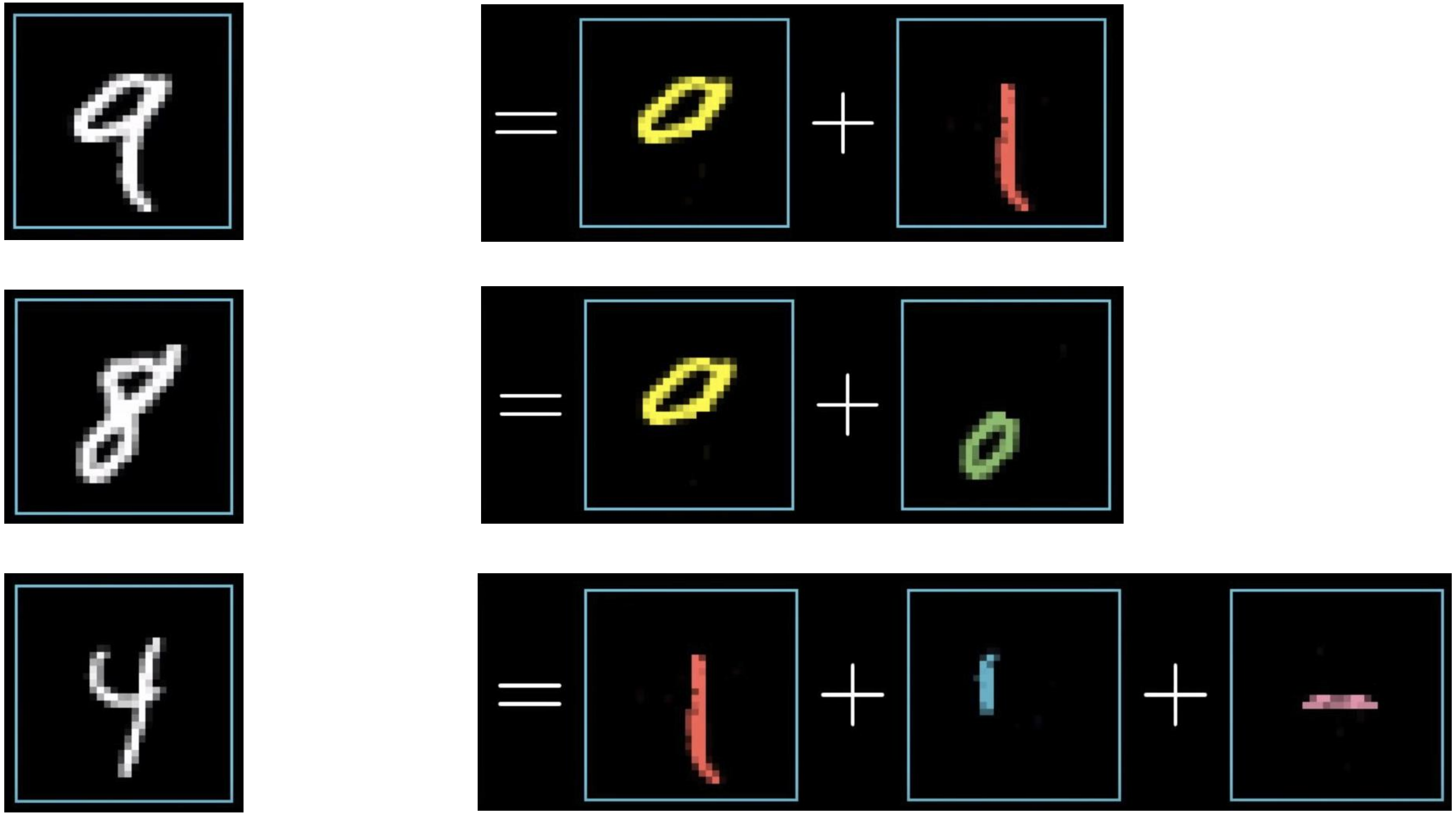
Given a collection of images of handwritten digits, determine which single-digit value was written.

Computer Vision: MNIST Dataset

- How do computers make sense of images?
- They convert them into a grid of pixel values.
- For example, in a color image, each pixel has a coordinate location on the image and an intensity value associated with the red-green-blue (RGB) color model.



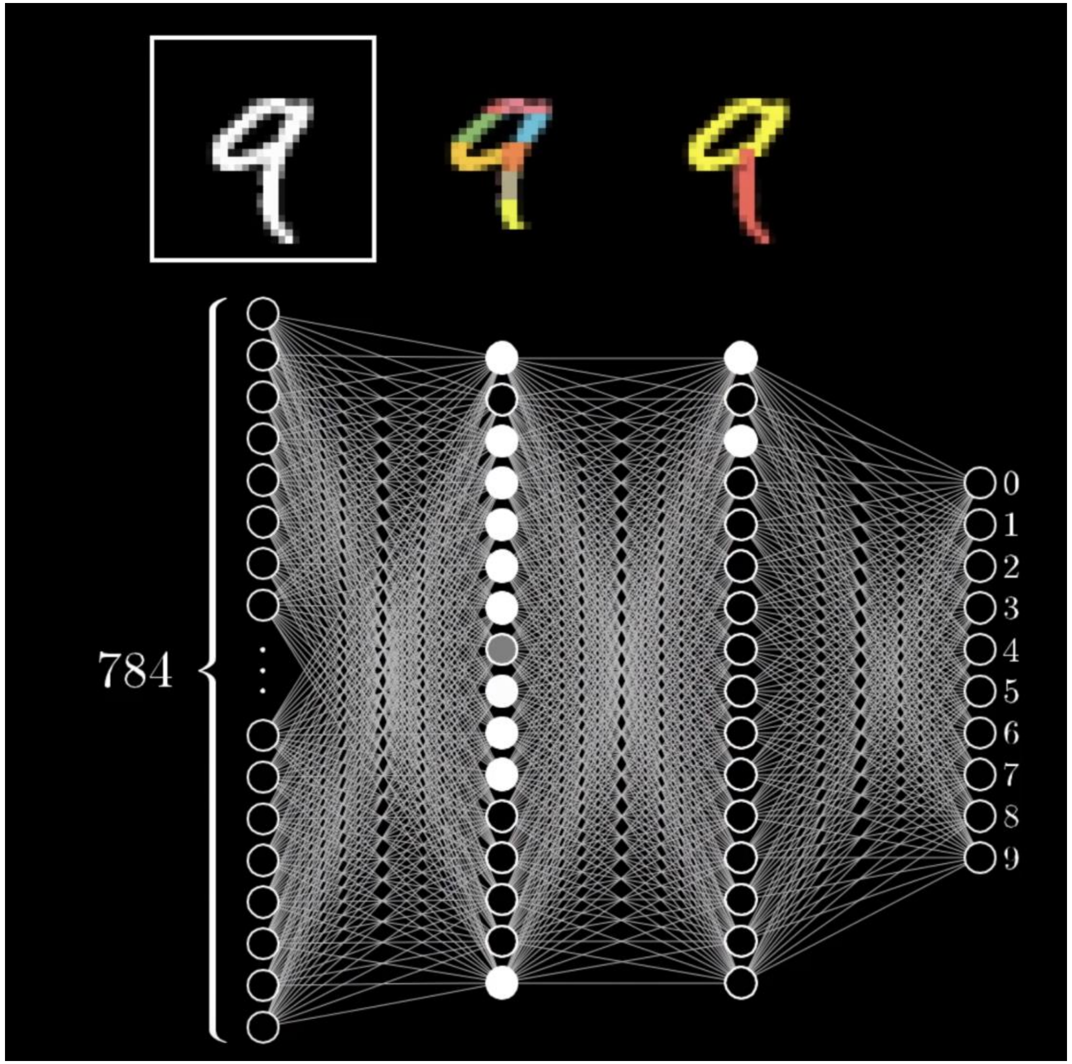
Computer Vision: MNIST Dataset



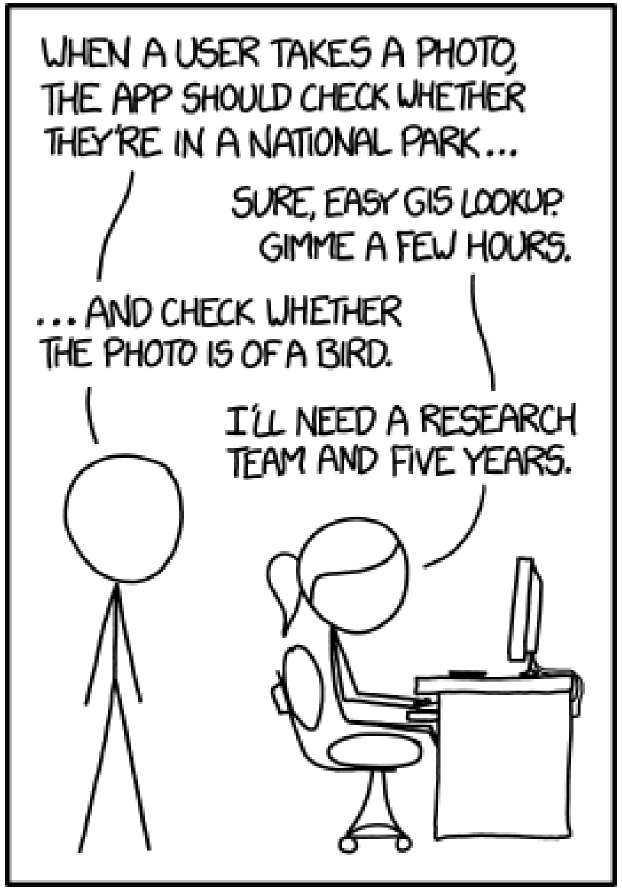
Computer Vision: MNIST Dataset

Conceptually, we can imagine each hidden layer of neurons acting to identify more and more complex features:

- 0^{th} layer (the input layer) is the numeric pixel data of our image.
- 1^{st} layer learns to look for vertical and horizontal lines.
- 2^{nd} layer learns to put the lines together to form loops.
- 3^{rd} layer (output layer) puts all of it together to decide what number the computer is "seeing."

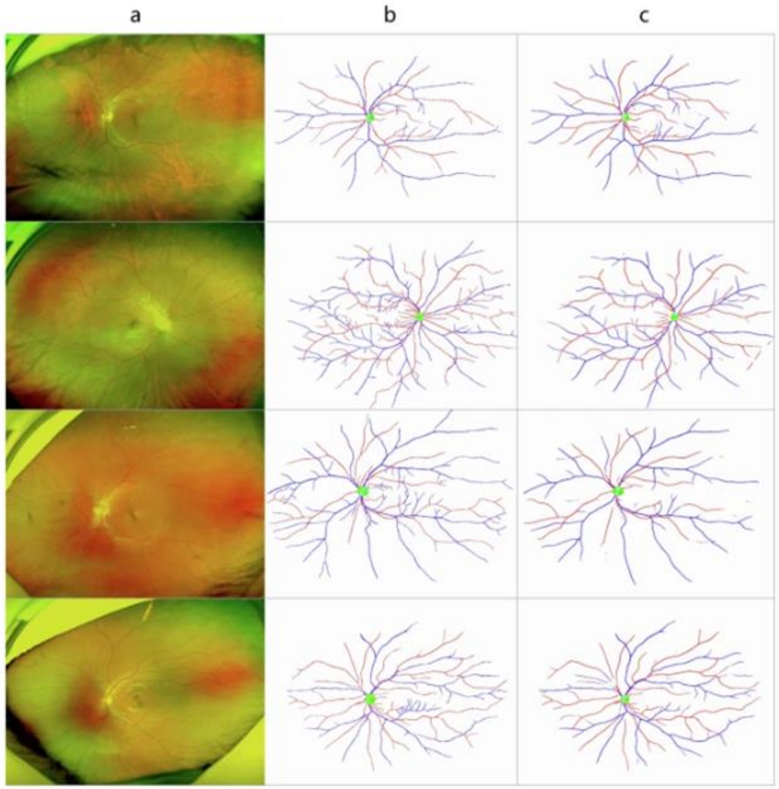


Computer Vision Isn't Perfect



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Example – Computer Vision for Renal Function

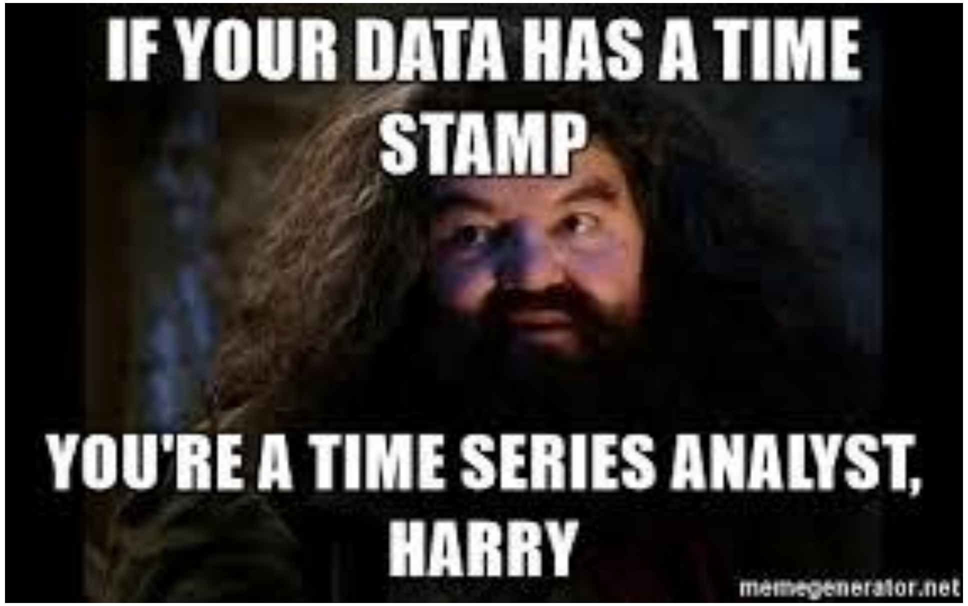


Four UWF images (**a**) and the segmentation results from experienced ophthalmology experts (**b**) and the segmentation model (**c**) were randomly selected for representation. The automatic segmentation of the optic disc and the vessels were very close to the doctor's annotation.

Source: [Screening CKD with Deep Learning | nature.com]()

Difficult Use Case: Chihuahua or Muffin?

Time Series Analysis

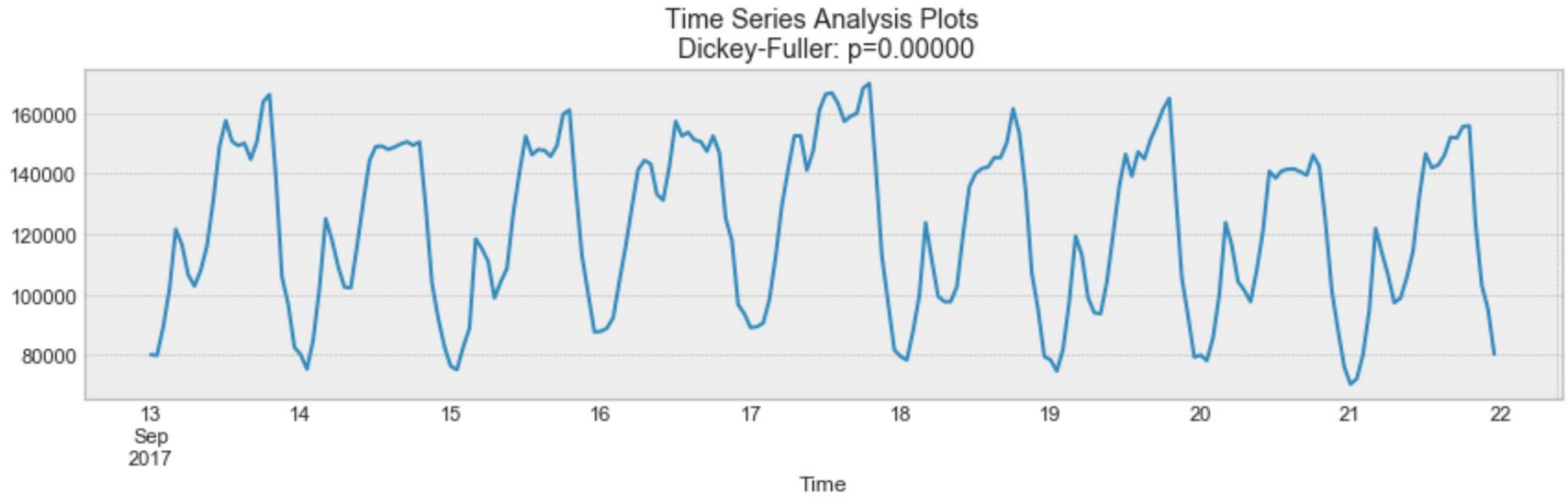


Time Series

- Time series data are commonly encountered in everyday life.
- Time series data is periodically captured for a given time period.
- Examples include financial prices, weather, home energy usage, height measured over time, etc.
- Stock prices and market indices are common examples.

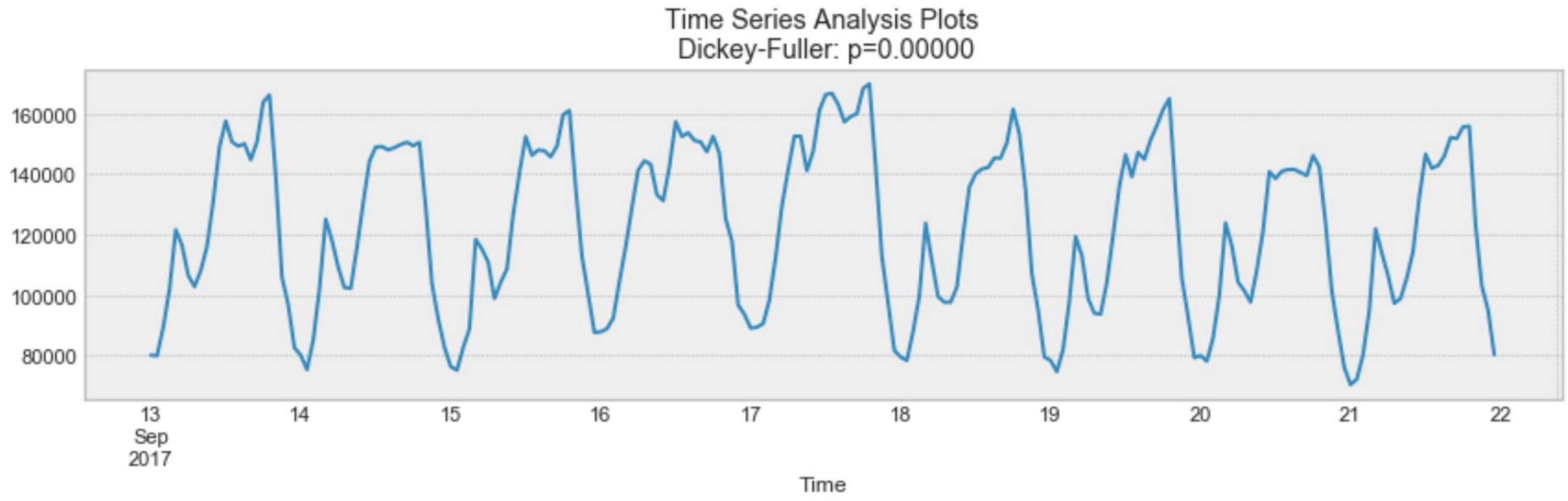
Seasonality in Time Series Data

- Periodic fluctuations in the graph.
- Trends that reoccur over time.
- Example: energy consumption is high during the day and low at night.



Stationarity

- A time series is stationary when its statistical properties do not change overtime (e.g., constant mean and variance).
- Stationary time series are ideal for modeling.
- The plot from the slide before is considered stationary.

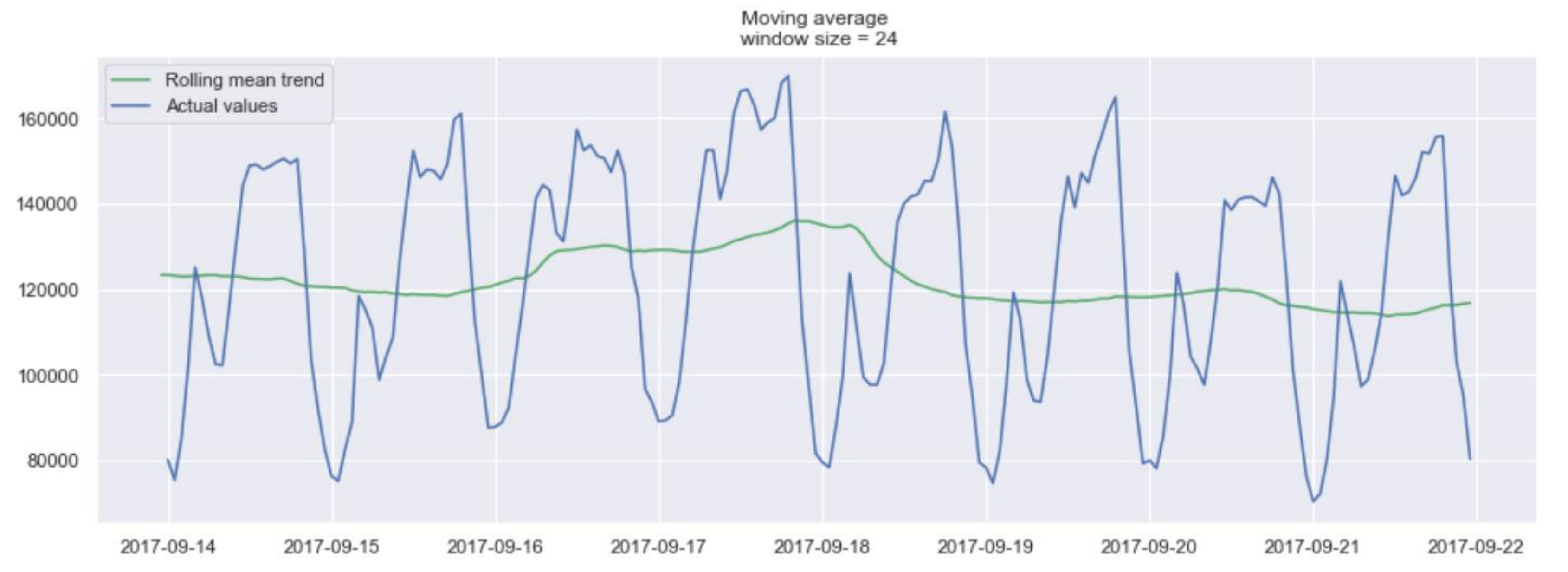


Smoothing Methods

- Smoothing methods reduce the effects of the random variation that comes from seasonality.
- These methods reveal the underlying trends in the data.
- Forecasts are weighted averages of past observations.
- There are two groups of smoothing methods:
  - Averaging Methods
  - Exponential Smoothing Methods

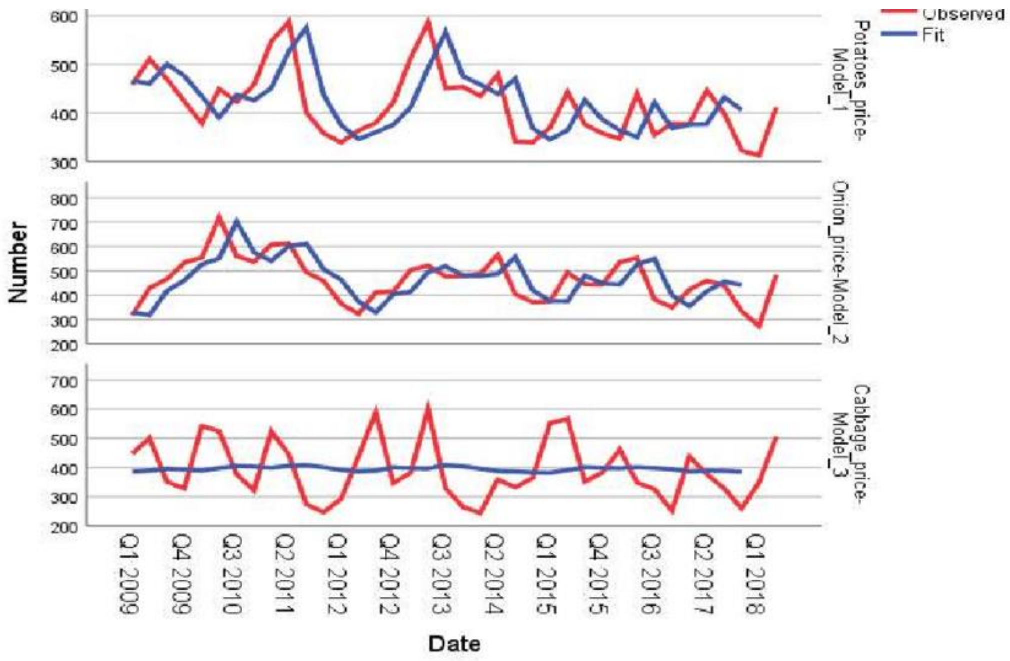
Moving Averages

- Also known as rolling means.
- A naïve way to evaluate the intricacies of the data.
- The next observation is the mean of a given window or all past observations.
- A window applies the moving average model to smooth the time series and highlight different trends.



Exponential Smoothing Methods

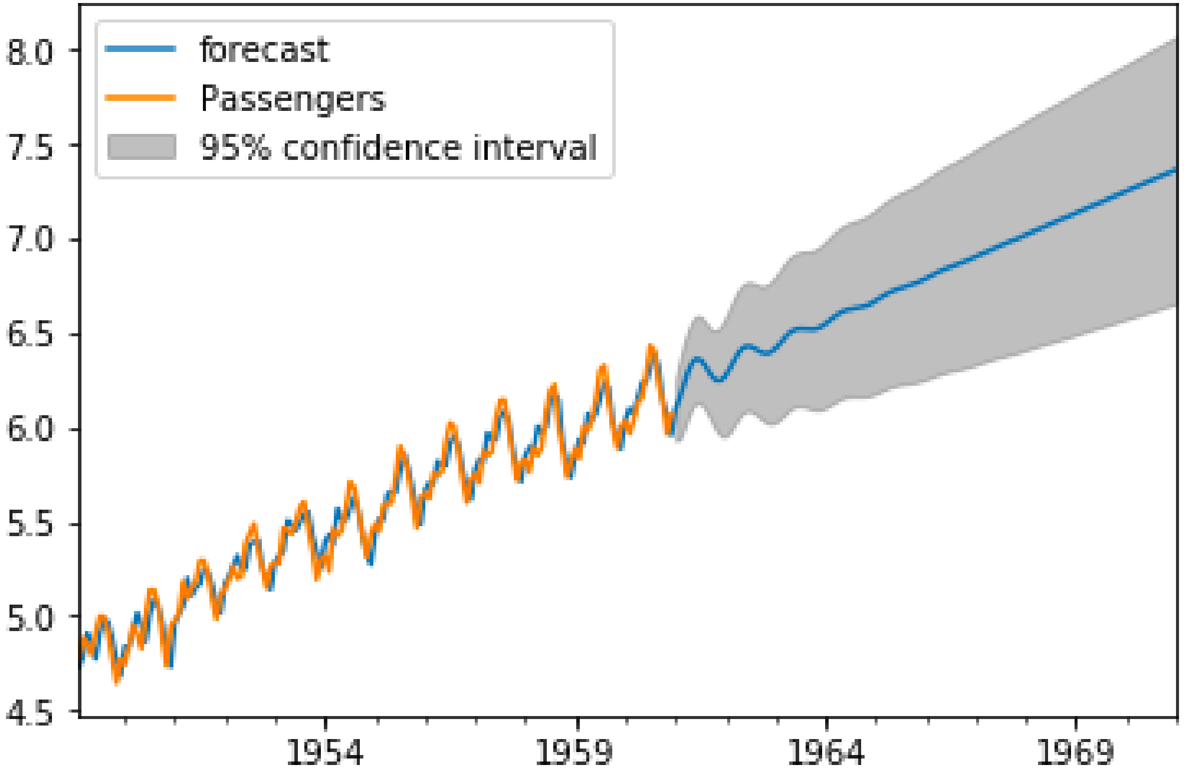
- Use this method for data sets that are more irregular where there is no seasonality or trends.
- Calculated as a weighted average from the previous level and the current observation.



ARIMA Model

ARIMA models in time series forecasting predict future values based on historical data and patterns.

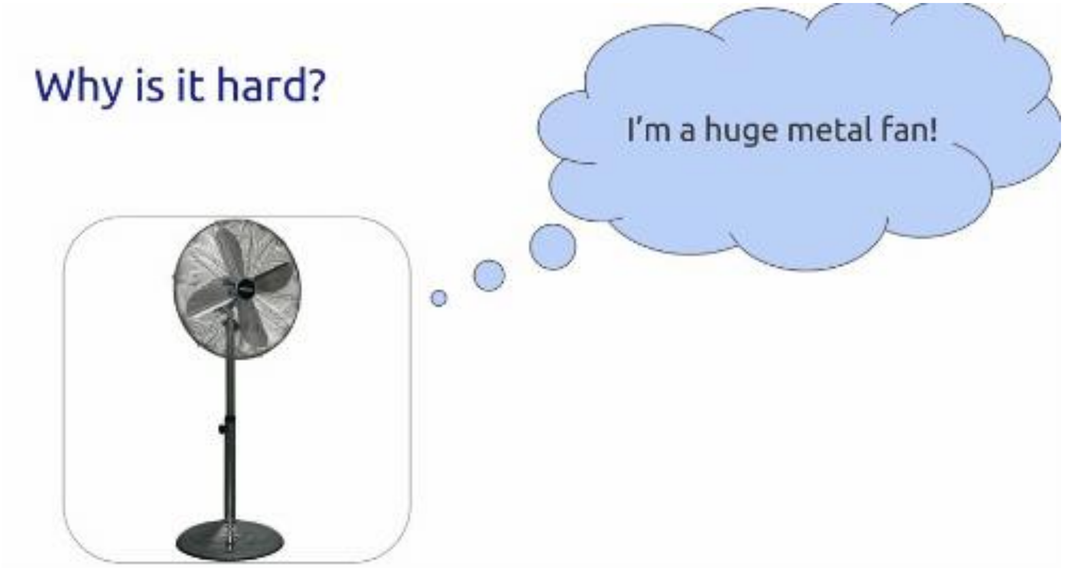
- AR: Autoregression
  - Linear relationship with previous data
  - lag observations – parameter p
- I: Integrated
  - making the time series stationary
  - differencing order – parameter d
- MA: Moving Average
  - uses the moving average for previous data
  - residual error window size – parameter q



The SARIMA model also accounts for seasonality patterns

Something to Remember about Time Series

Natural Language Processing (NLP)



NLP is hard.

Natural Language Processing

- How can we draw insights from our data when it has a lot of text?
- The focus of NLP is to program computers to process and analyze large amounts of natural language data.
- Many real-world use cases:
  - Machine translation
  - Chatbots
  - Resume filtering

NLP in Outlook

What do you look for when you read an e-mail?

From: Someone, David [USA] <Someone_David@bah.com>
Sent: Tuesday, January 8, 2019 4:28 PM
To: You, Silly [USA] <You_Silly@bah.com>
Subject: Quick chat about NLP

Hello,

I just wanted to take a moment of your time to tell you about NLP. I'm not talking about neuro-linguistic programming, though it does come up often when typed into a Google search. I'm expecting you as a human that understands context to get that I'm referring to natural language processing. I also expect that by the end of this email you will know my name and that I am an individual and not a puppy or corporation. Would you be free this weekend to grab coffee at the local Starbucks to talk more about NLP?

Best,
David

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NLP in Outlook

From: Someone, David [USA] <Someone_David@bah.com>
Sent: Tuesday, January 8, 2019 4:28 PM
To: You, Silly [USA] <You_Silly@bah.com>
Subject: Quick chat about NLP

Hello,

I just wanted to take a moment of your time to tell you about NLP. I'm not talking about neuro-linguistic programming, though it does come up often when typed into a Google search. I'm expecting you as a human that understands context to get that I'm referring to natural language processing. I also expect that by the end of this email you will know my name and that I am an individual and not a puppy or corporation. Would you be free this weekend to grab coffee at the local Starbucks to talk more about NLP?

Best,
David

Suggested Meetings

+ Get more add-ins

We think we've found an event

Would you be free this weekend to grab coffee at the local Starbucks to talk more about NLP?

When: 8:00 AM - 8:30 AM Saturday, January 12, 2019
Who:
Where: Enter location

NLP – Count Vectorizer

How do we convert human language to something a computer can understand?

Corpus

Document 1: "Today was a great, great day."

Document 2: "I like puppies."

All unique "tokens"

"Today", "was", "a", "great", "day", "I", "like", "puppies"

count vectorizer

DOC ID	day	great	a	puppies	I	like	was	Today
1	1	2	1	0	0	0	1	1
2	0	0	0	1	1	1	0	0

Word Embeddings / Word Vectors

DOC ID	day	great	a	puppies	I	like	was	Today
1	0	2	1	0	0	0	1	1
2	0	0	0	1	1	1	0	0

$$\left.\begin{array}{l}[0, 2, 1, 0, 0, 0, 1, 1] \\ [0, 0, 0, 1, 1, 1, 0, 0]\end{array}\right\}$$

word vectors
a.k.a.
"word embeddings"

Now, our text is in a representation that our machine learning models can understand.

Limitations of Count Vectorizers

What is less than ideal with our count vectorizer?

Limitations of Count Vectorizers

What is less than ideal with our count vectorizer?

Count vectorizers treat all occurrences of words equally, so common words (e.g., "the", "a", "of"…) dominate the signal of a vector.

Tf-idf Word Vectorizer

tf-idf = term frequency X inverse document frequency

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{N}{\mathrm{df}_i}\right)$$

$\mathrm{tf}_{i,j}$ = total number of occurences of i in j

df_i = total number of documents (speeches) containing i

N = total number of documents (speeches)

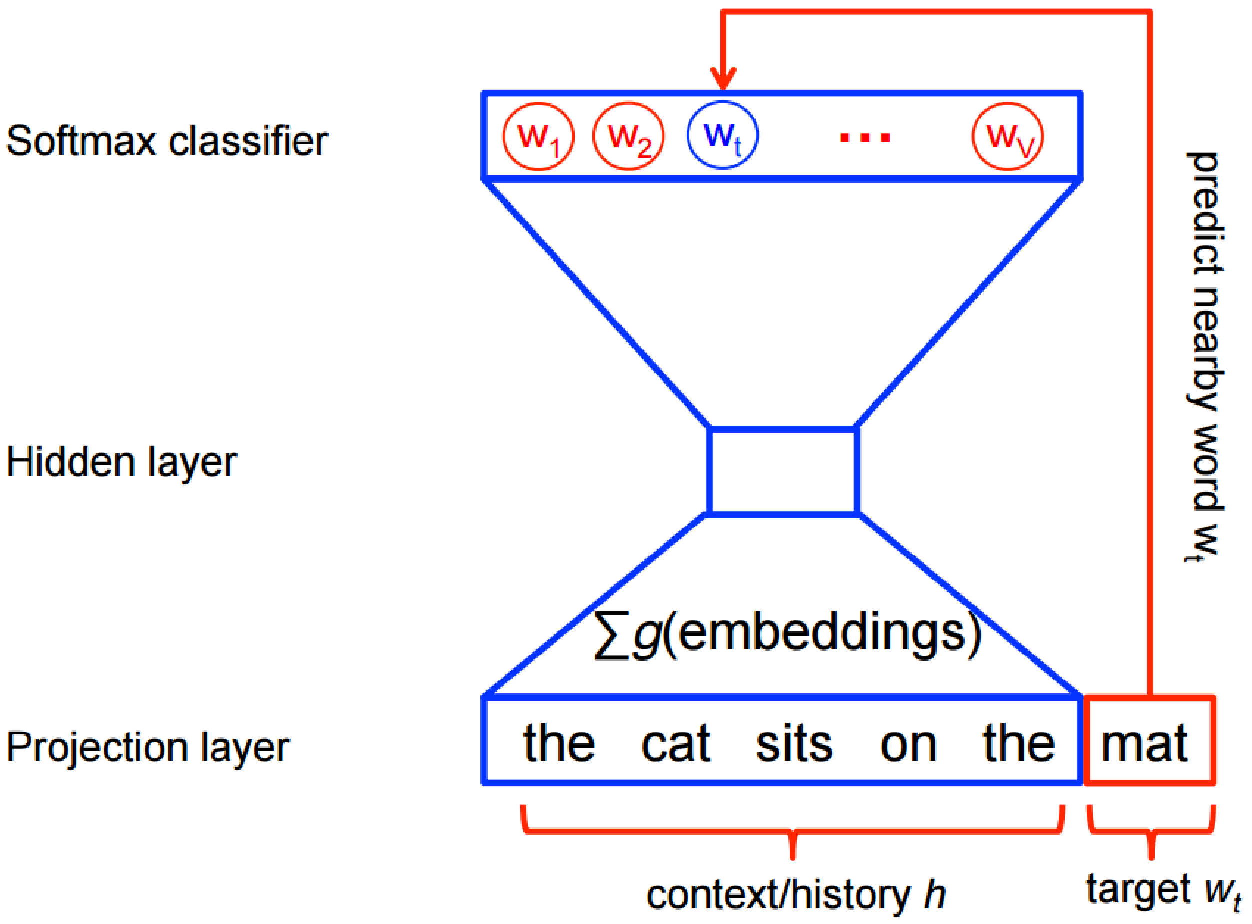
Word2Vec – A More Sophisticated Vector

What's different here?

Instead of creating a vector for each of our documents, we can create a vector for every word in our vocabulary.

How does it work?

Uses a neural network to predict what word comes next in a sequence, then adjusts the vector for the target word if it was wrong.

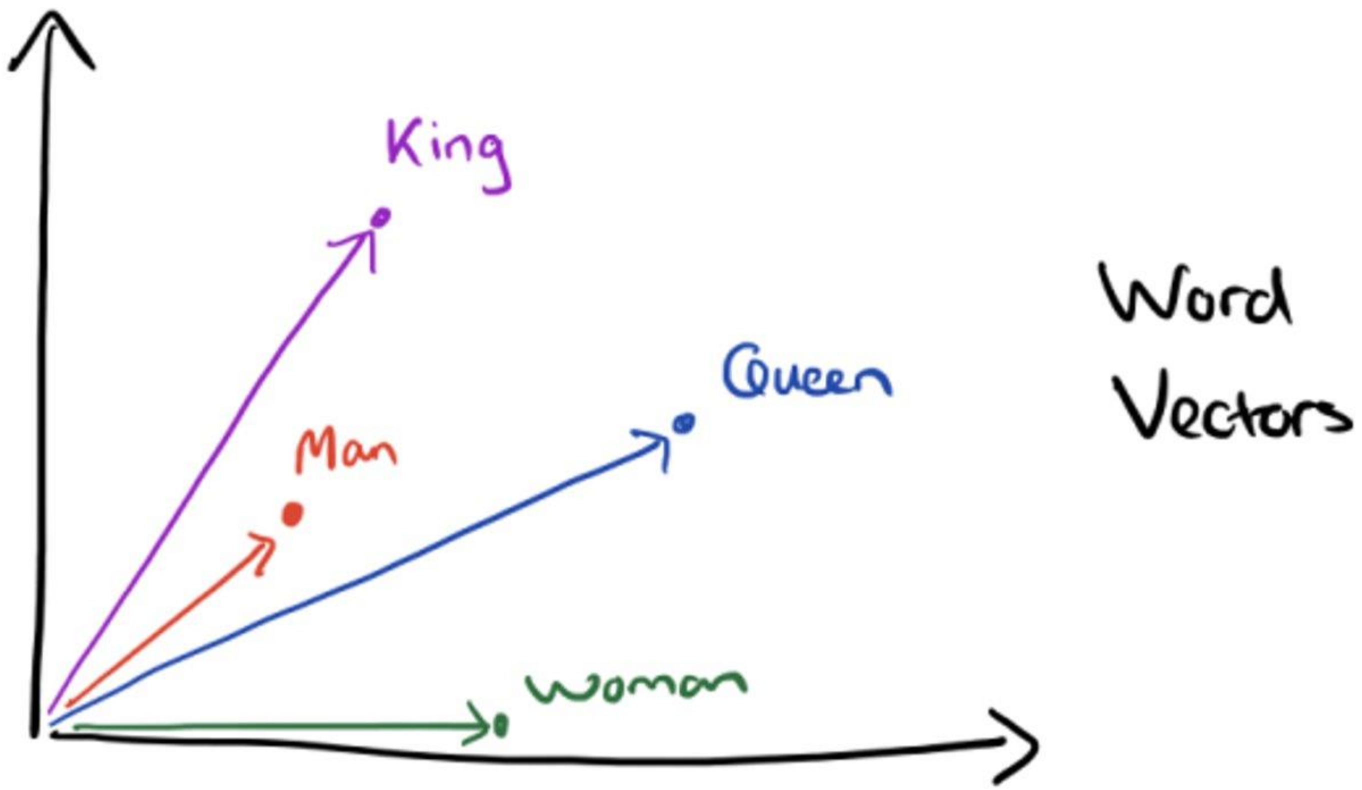


NIH National Institute of Diabetes and Digestive and Kidney Diseases

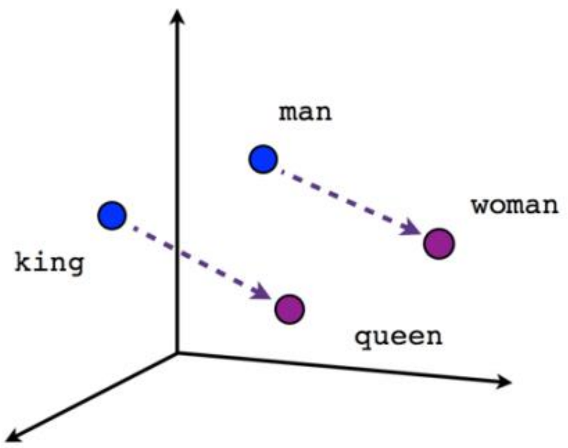
Central Repository

Word2Vec

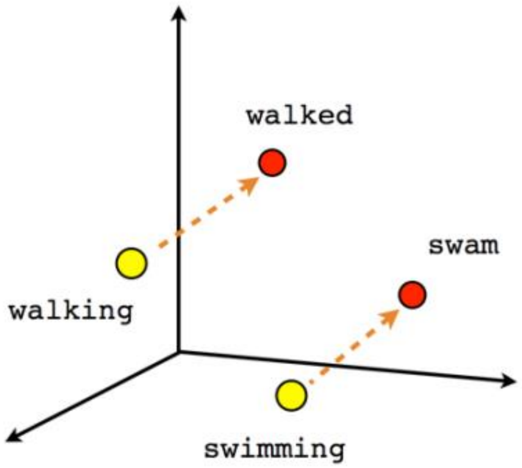
- The result of this strategy of vectorizing words means that individual words that are used in similar contexts are spatially close together.



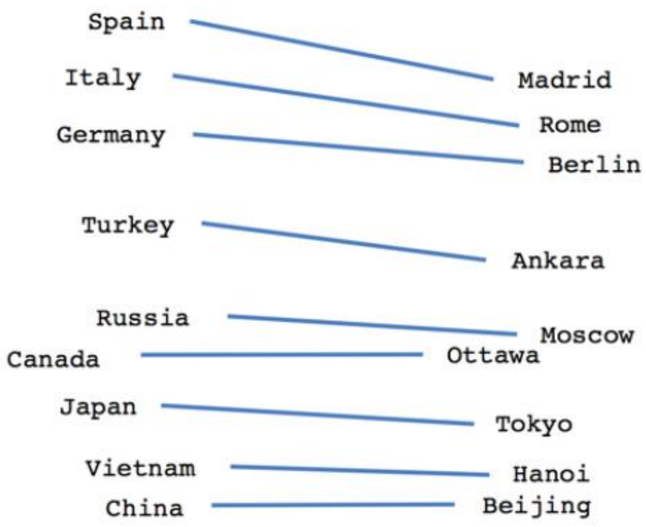
Word2Vec - Examples



Male-Female



Verb tense



Country-Capital

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Common Tasks in NLP

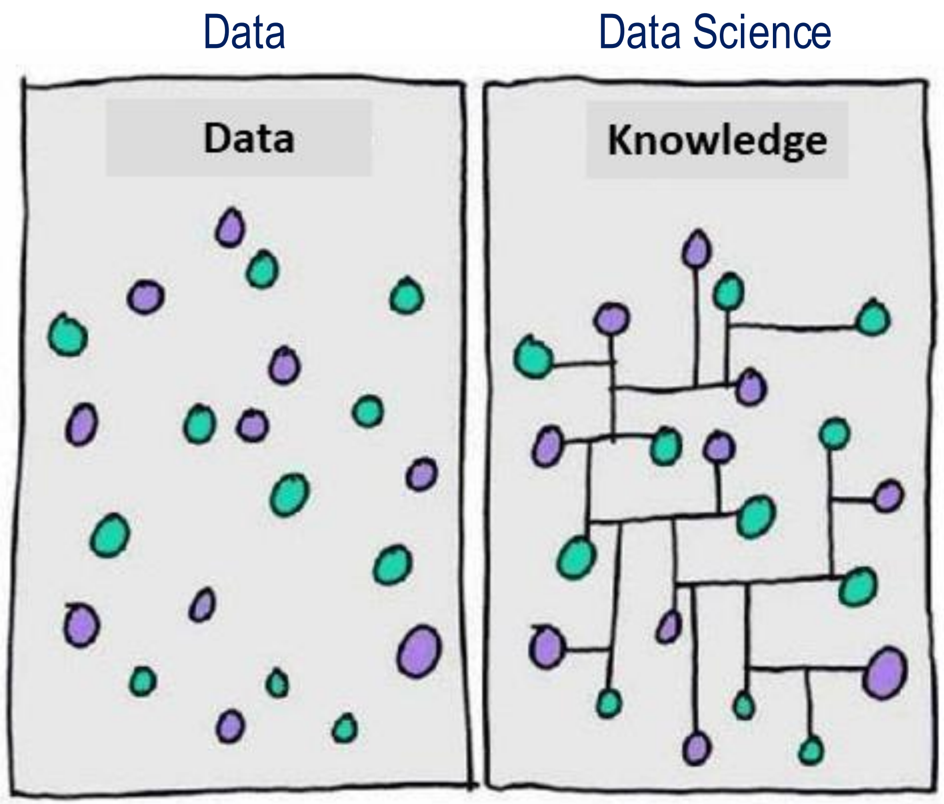
Name	Description
Tokenization	Segmenting text into words, punctuation marks, etc.
Part-of-speech (POS) Tagging	Assigning word types to tokens, like verb or noun.
Dependency Parsing	Assigning syntactic dependency labels, describing the relations between individual tokens, such as subject or object.
Lemmatization	Assigning the base forms of words. For example, the lemma of “was” is “be”, and the lemma of “rats” is “rat”.
Sentence Boundary Detection (SBD)	Finding and segmenting individual sentences.
Named Entity Recognition (NER)	Labelling named “real-world” objects, such as persons, companies, or locations.
Entity Linking (EL)	Disambiguating textual entities to unique identifiers in a Knowledge Base.
Similarity	Comparing words, text spans, and documents to determine how similar they are to each other.
Text Classification	Assigning categories or labels to a whole document or parts of a document.
Sentiment Analysis	Allows us to capture meaning or intent in document

Course Summary

End of Course Survey

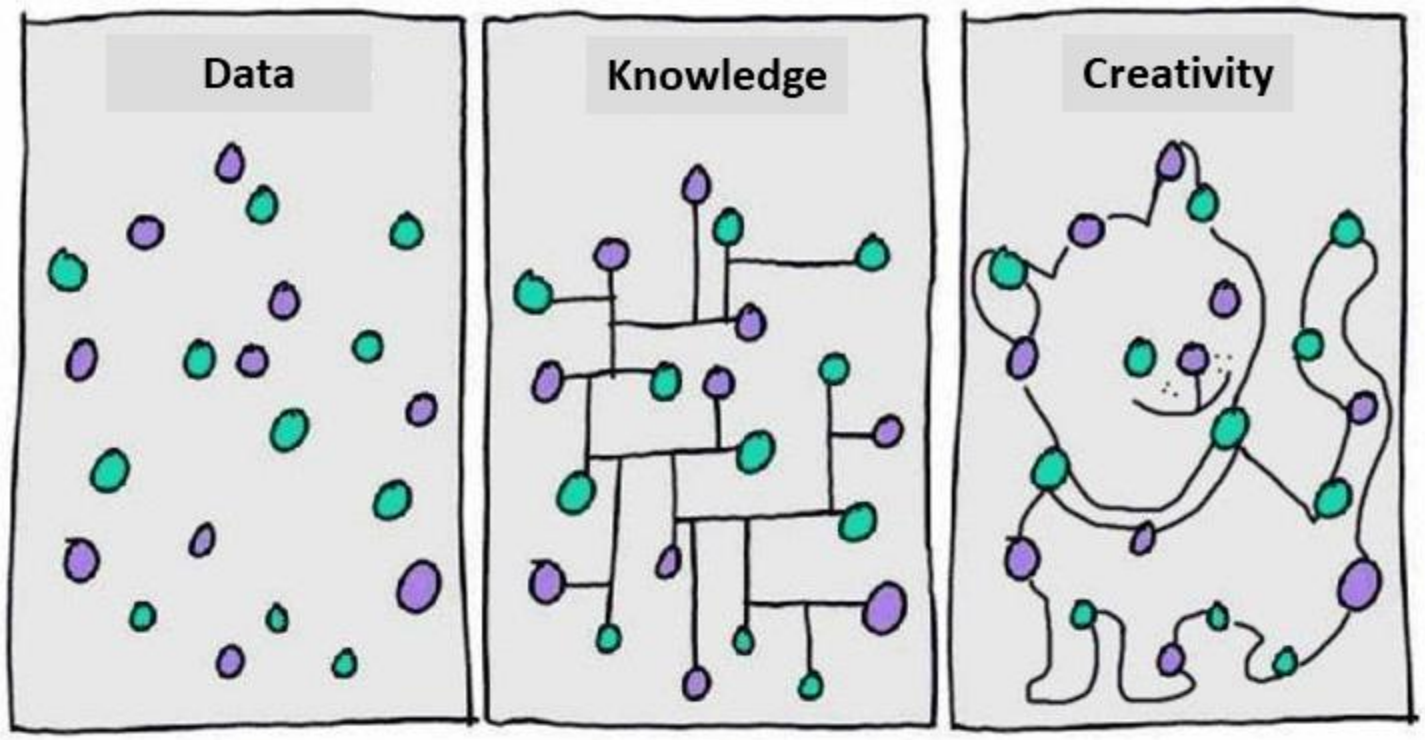
What Did We Just Learn?

In a Nutshell



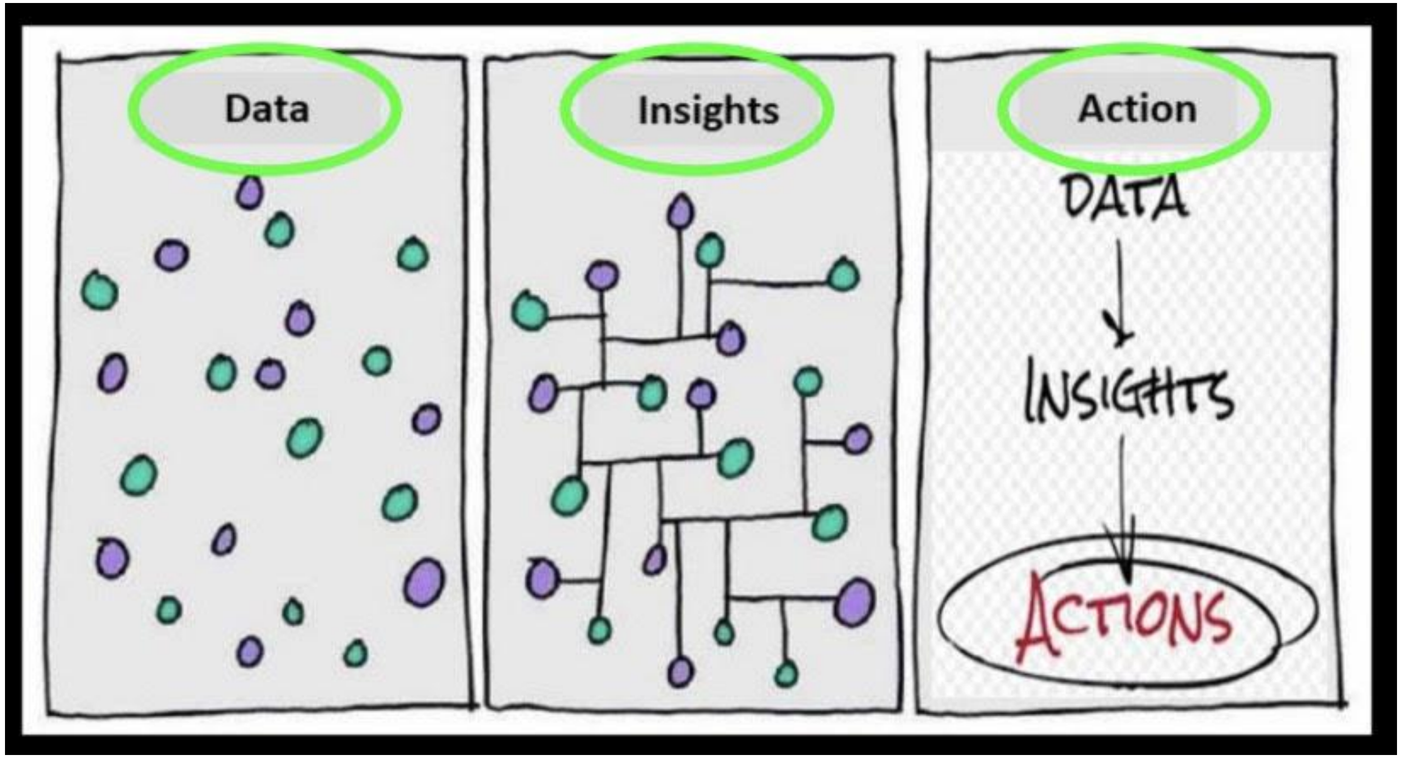
In a Nutshell

Find the pattern in the data



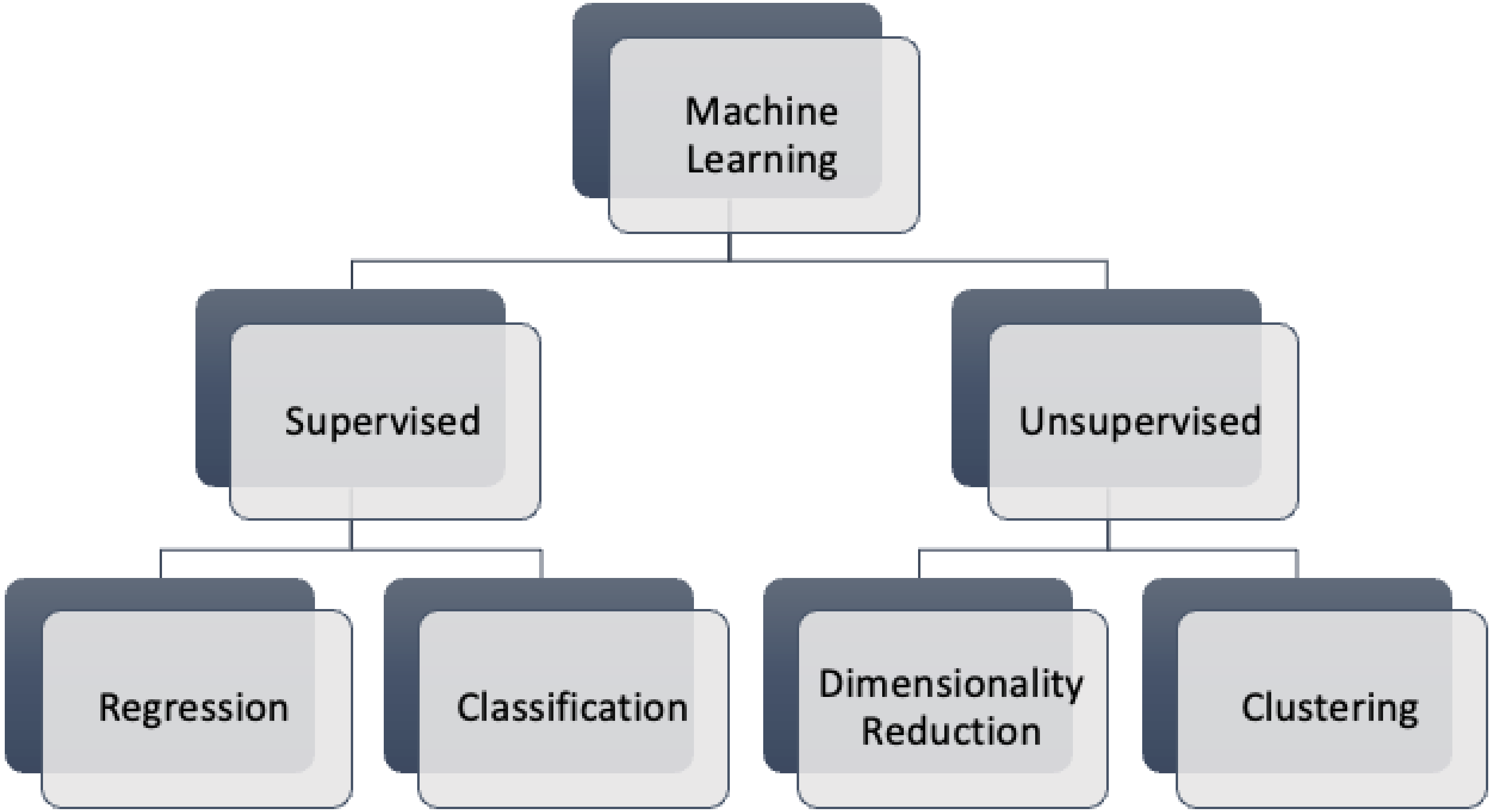
National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

From Data to Insights to Actions



National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Types of Learning



National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Summary of ML Techniques

Algorithm	Computationally expensive?	Requires lots of data?	Interpretable?
Linear Regression	No	No	Yes
Logistic Regression	No	No	Yes
PCA	Yes	Yes	No
K-means	No	No	Yes
Neural Networks	Yes	Yes	No

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Lessons Learned

- Navigate through a data science project using the seven-step data science process
    - Form a SMART problem statement, understanding what data science can and cannot do
    - Acquire useful data that can assist in solving the problem statement
    - Explore data and analyze preliminary findings to leverage initial insights from the data
    - Prepare data for use in machine learning pipelines
    - Develop models to represent relationships within the data
    - Render compelling visualizations to communicate data-driven narratives to your colleagues
    - Applying actionable insight to your problem statement
- Identify data opportunities within the organization to apply higher-level analytics, data science, and machine learning
    - Understand different machine learning algorithms and how to apply them
    - Recognize data science specialties, such as natural language processing and computer vision
- Identify tools that can assist in all parts of the data science process

Questions?


REPOSITORY
TECHNOLOGY ADVANCEMENT
BIOSTATISTICS
CLINICAL STUDIES
REGULATORY
Office of Clinical Research Support
NIH National Institute of Diabetes and Digestive and Kidney Diseases

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Arica Christensen – Christensen_arica@bah.com

Dr. Gordon Aiello – Aiello_Gordon@bah.com

March 27 - AI Fundamentals Part 1

April 24 – AI Fundamentals Part 2

Thank You!