National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

NIDDK-CR Resources for Research

Data Science and Meet the Expert Webinar Series

April 24, 2025

National Institute of Diabetes and Digestive and Kidney Diseases

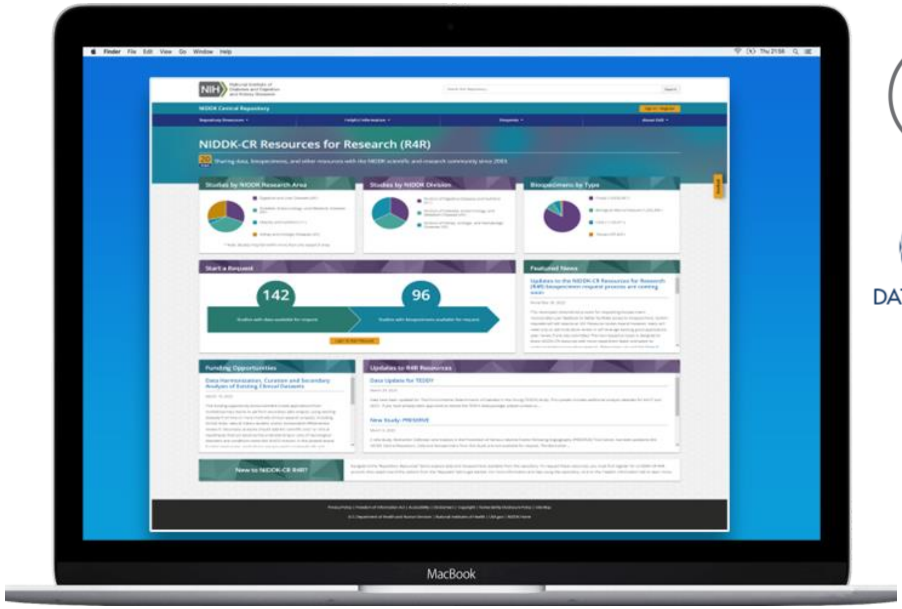
Central Repository

NIDDK Central Repository Overview

Mission

Established in 2003 to **facilitate sharing of data, biospecimens, 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 biospecimens 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 data or biospecimens
- Promotes FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST (Transparency, Responsibility, User focus, Sustainability, and Technology) principles



CORE TRUST SEAL

WORLD DATA SYSTEM

Imaging Data Files	Clinical Datasets	Biospecimens
15.8 M	>8,400 from 189 clinical studies	>16 M

Registered Users	Weekly Users	Public Releases
6,889	>5,000	>875

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

National Institute of Diabetes and Digestive and Kidney Diseases

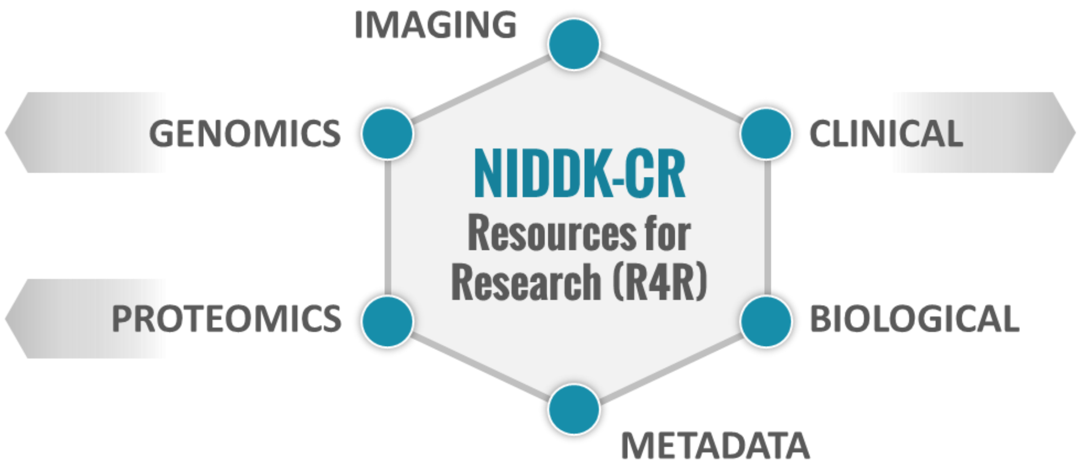
Central Repository

NIDDK Data Sharing Ecosystem

The NIDDK-CR is a part of the broader NIH-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, provides access to biospecimens, and direct links to other repositories with additional resources such as genomics data.

dbGaP GENOTYPES and PHENOTYPES

PRIDE Archive PRoteomics IDEntifications Database



KPMP

IMMPORT

JCHR

T1D TYPE 1 DIABETES CONSORTIUM

Additional Resources

FAIRsharing.org standards, databases, policies

DataCite FIND, ACCESS, AND REUSE DATA

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

DK net

Google Dataset Search / Schema.org

NIH U.S. National Library of Medicine ClinicalTrials.gov

Vivli CENTER FOR GLOBAL CLINICAL RESEARCH DATA

PLANNING PHASE

figshare

NIH HEAL INITIATIVE

NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Future Functionality: Analytics Workbench

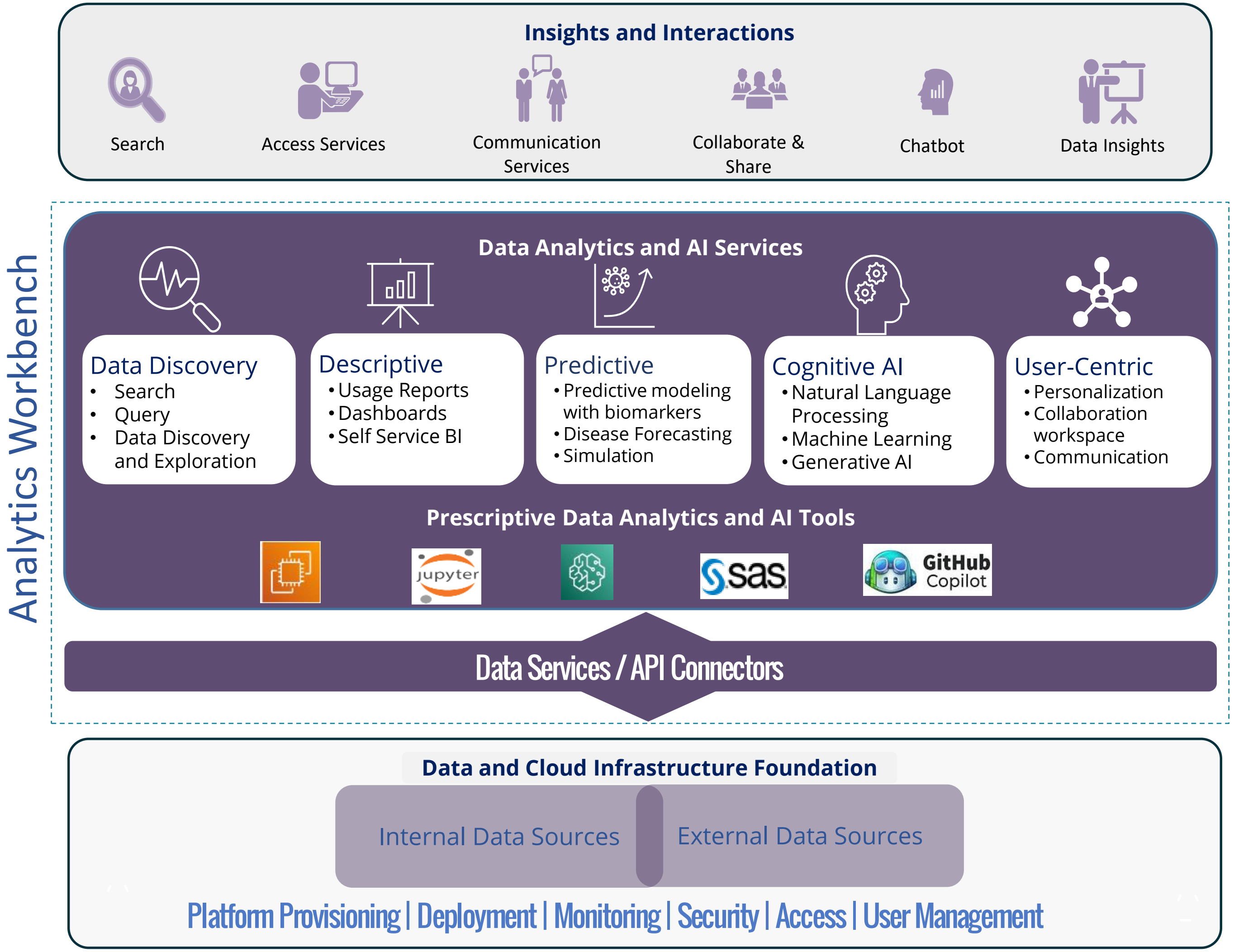
***Streamlining** end-to-end data science lifecycle and discovery of data-driven biomedical insights.*

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



National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

NIDDK-CR 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

Secondary 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

- Today – Artificial Intelligence fundamentals applications
- May 29 – FAIR and AI-ready data sharing
- June 26 – Different privacy preserving techniques and implications for researchers
- July 31 – Challenges, opportunities, and considerations for secondary researchers using electronic health records and real-world data sources
- August 28 – 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.

AI Fundamentals
Part 2: AI Applications

NIDDK-CR Data Science

Meet the Experts Webinar Series

April 24th, 2025

Presented by: Booz Allen Hamilton

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

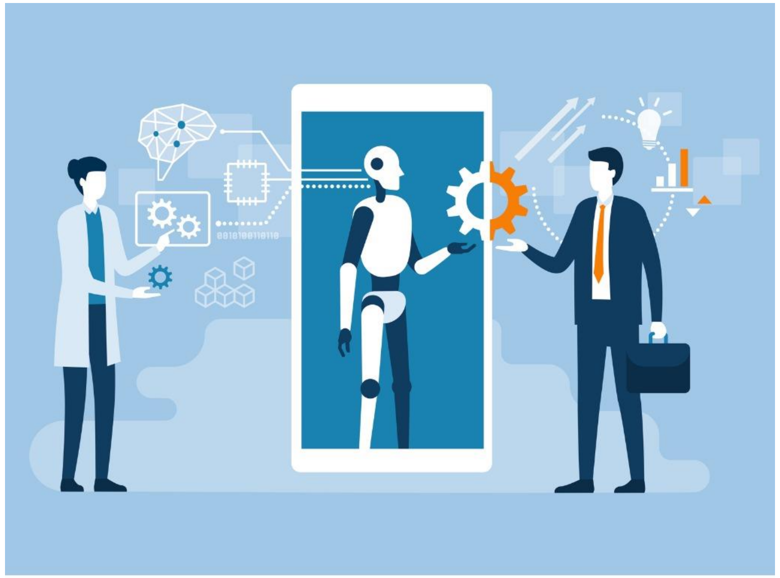
7 hours (2 sessions)

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

Agenda



1. Statistics Primer
2. Model Metrics
   1. Classification
   2. Regression
3. Neural Networks and Their Applications
   1. Feedforward
   2. Convolutional
   3. Recurrent
   4. Transformer
   5. Generative Adversarial

Office of Clinical Research Support

Statistics Primer

Statistics Primer

To motivate the discussion, we'll examine a sample kidney disease data set

- **Note**: This is the same data used during Data Science Fundamentals in February

	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]

Statistics Primer

For each patient we 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)

Source:

Statistics Primer

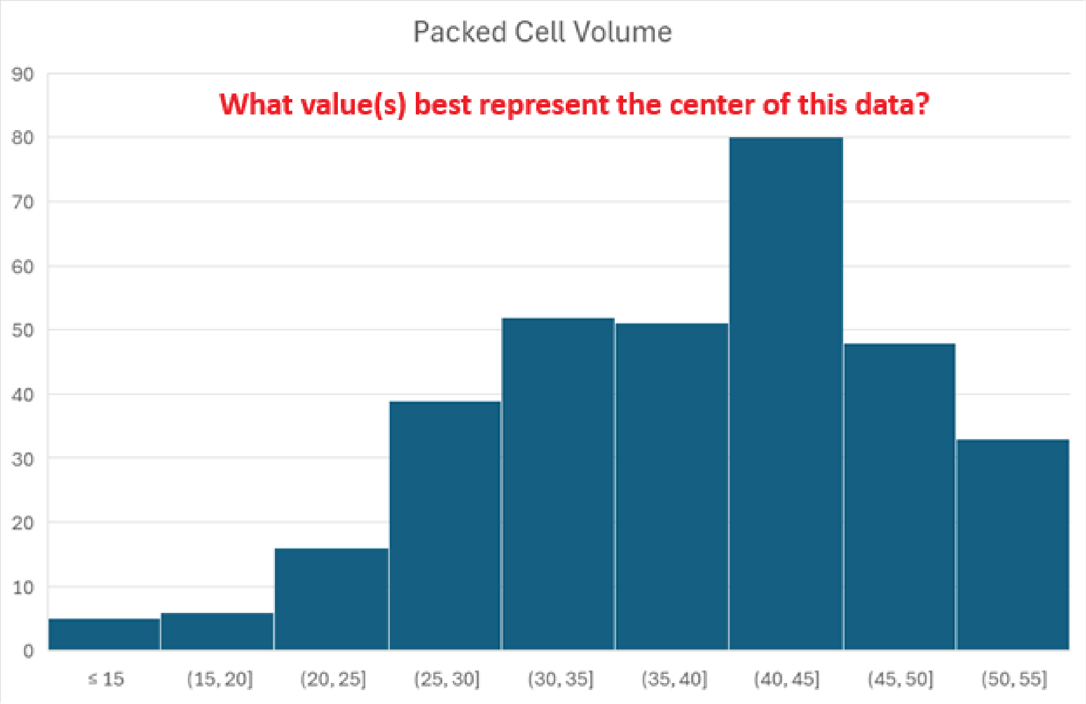
- The basics:
  - Mean
  - Median
  - Mode
  - Distributions
- What are descriptive statistics?
- What is correlation?

Source:

Statistics Basics

Measures of Central Tendency

- Measures of central tendency aim to quantify the central value of a dataset's distribution.
- The most common measures of central tendency are the mean, median, and mode.
- Each measure of central tendency provides a numeric value quantifying a representative point around which the data tends to cluster.
- The values produced by the mean, median, and mode can be different.
- Measures of central tendency facilitate comparing datasets, identifying trends over time, and data-driven decision making.



Source:

Statistics Basics – Mean

The Mean

- Quantifies the average value of a dataset.
- Calculated by summing all the values and dividing by the total number of values.
- Represents the balancing point of a dataset.
- Interactive visualization of the mean as a dataset's balancing point
- Is sensitive to outliers; i.e., extreme values can significantly influence the mean.

Source:

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Statistics Basics – Mean Example

Example

The values in the table below are from the kidney disease data and show the packed cell volume for a sample of nine patients.

1. Calculate the mean value for this data sample.

Employee	1	2	3	4	5	6	7	8	9
# of Years	22	24	37	44	45	45	47	51	54

Mean:

21

31

41

Statistics Basics – Mean Example

Example

The values in the table below are from the kidney disease data and show the packed cell volume for a sample of nine patients.

1. Calculate the mean value for this data sample.

Employee	1	2	3	4	5	6	7	8	9
# of Years	22	24	37	44	45	45	47	51	54

Mean:

21

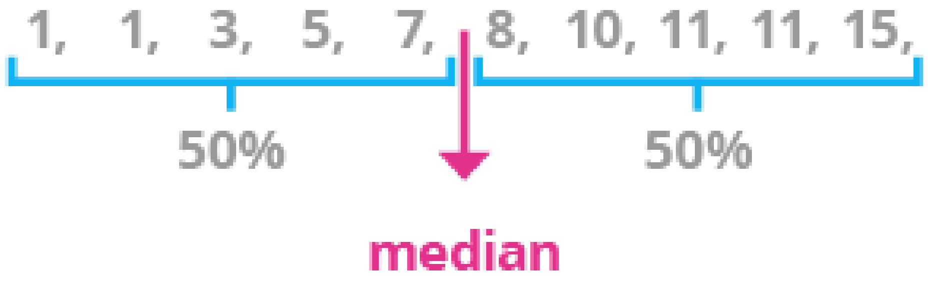
31

41

Statistics Basics – Median

The Median

- Is the middle value of a dataset when it is ordered from smallest to largest.
- Represents the point that divides the dataset into two equal halves.
- It is less affected by outliers than the mean, making it a more robust measure of central tendency in some applications.



Source:

Statistics Basics – Median Example

Example

The values in the table below are from the kidney disease data and show the packed cell volume for a sample of nine patients.

1. Make sure the data are ordered from least to greatest.
2. Determine the median value for this data sample.

Employee	1	2	3	4	5	6	7	8	9
# of Years	22	24	37	44	45	45	47	51	54

Median:

35

45

55

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Statistics Basics – Median Example

Example

The values in the table below are from the kidney disease data and show the packed cell volume for a sample of nine patients.

1. Make sure the data are ordered from least to greatest.
2. Determine the median value for this data sample.

Employee	1	2	3	4	5	6	7	8	9
# of Years	22	24	37	44	45	45	47	51	54

Median:

35

45

55

Statistics Basics – Mode

The Mode

- Is the value that occurs most frequently in a dataset.
- Identifies the value that is most common/popular in a dataset.
- There can be one mode (unimodal), multiple modes (multimodal), or no mode if no value occurs more frequently than others.

2, 2, 3, 4, 5, 5, 7, 8, 8, 8, 9

2 times   2 times   3 times

Source:

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Statistics Basics – Mode Example

Example

The values in the table below are from the kidney disease data and show the packed cell volume for a sample of nine patients.

1. Determine the mode for this data sample

Employee	1	2	3	4	5	6	7	8	9
# of Years	22	24	37	44	45	45	47	51	54

Mode:

NA

45

47

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Statistics Basics – Mode Example

Example

The values in the table below are from the kidney disease data and show the packed cell volume for a sample of nine patients.

1. Determine the mode for this data sample

Employee	1	2	3	4	5	6	7	8	9
# of Years	22	24	37	44	45	45	47	51	54

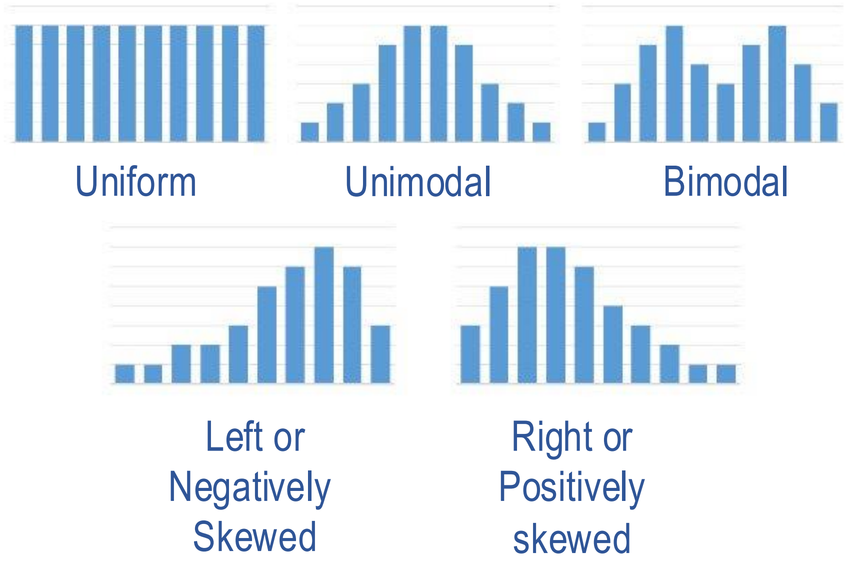
Mode:

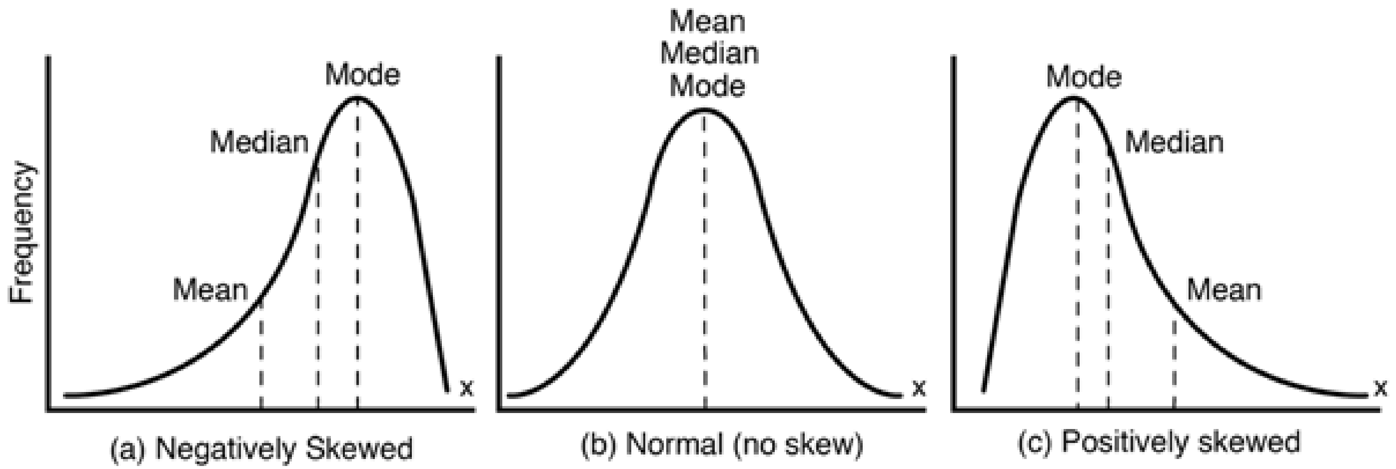
NA

45

47

Types of Distributions in Statistics

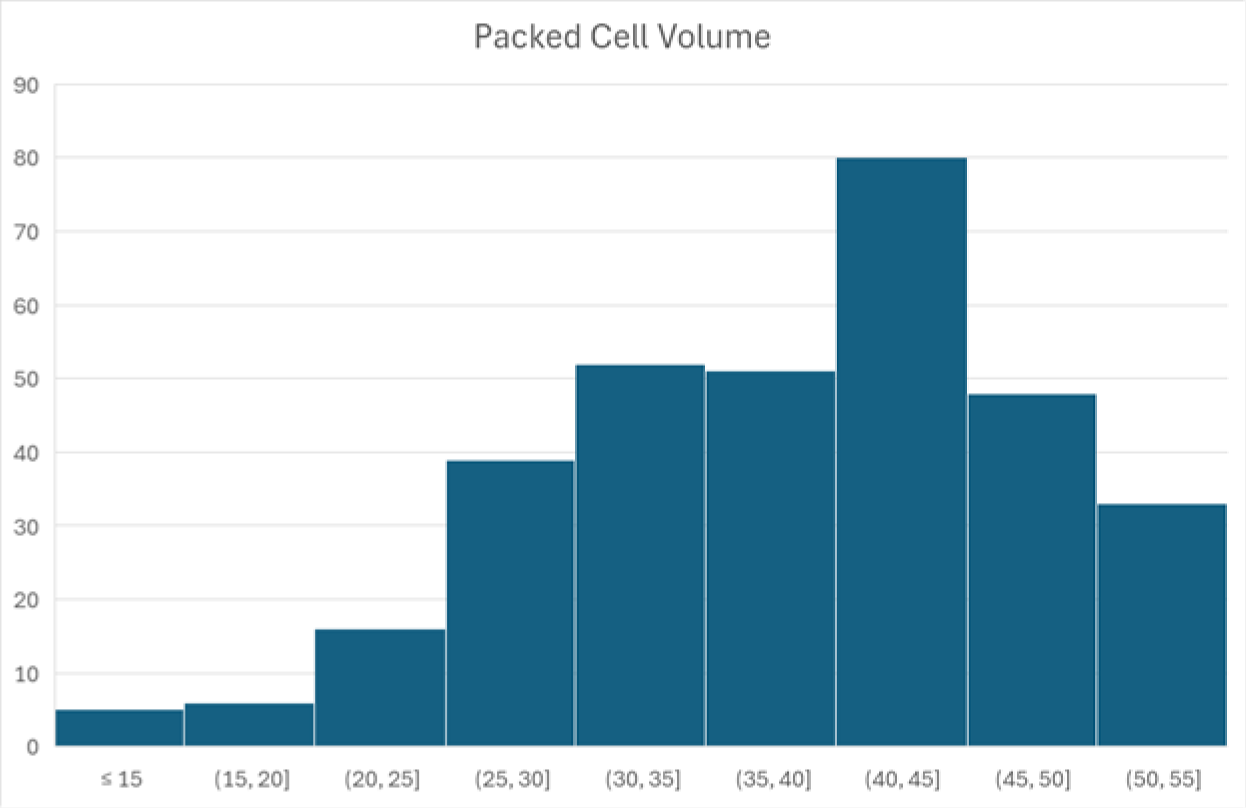




Source:

Types of Distributions in Statistics

What kind of distribution does the packed cell volume data have?



Source:

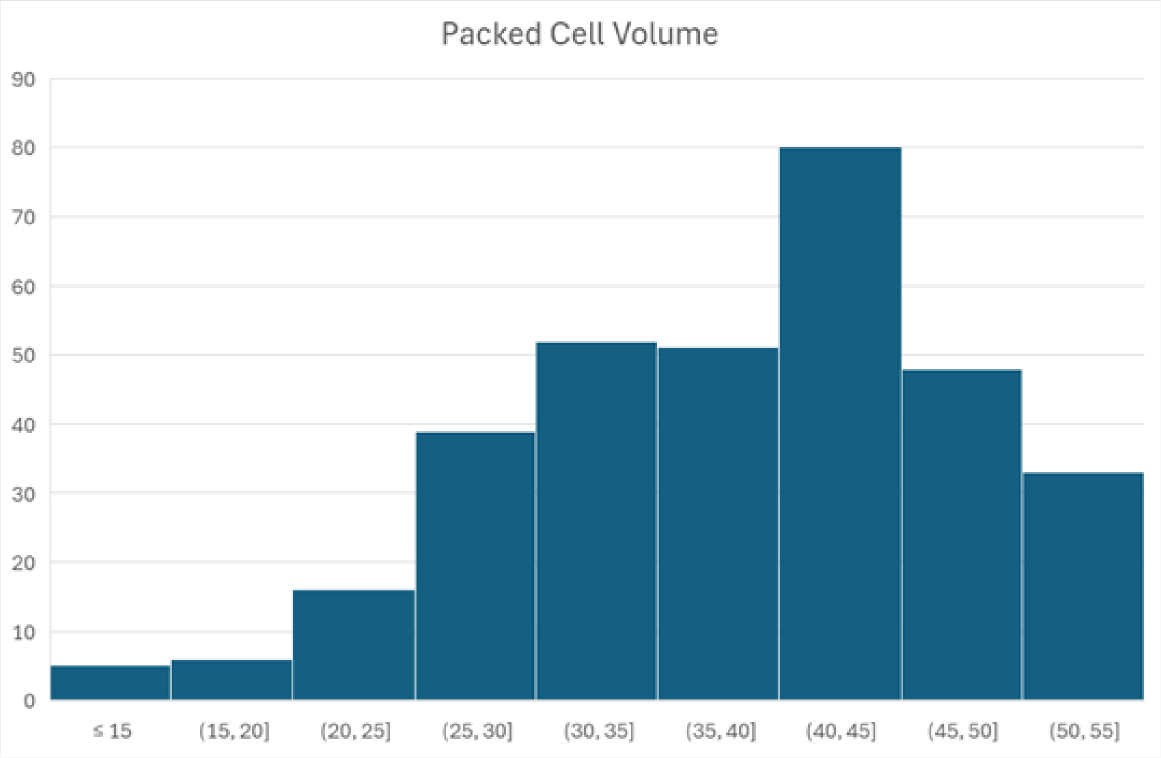
NIH National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Descriptive Statistics

- Descriptive statistics focuses on the procedures and methods used to informatively organize, summarize, and present data.

- **Question**: What inferences can be made about the packed cell volume data from the kidney disease table based on the distribution histogram shown at right?

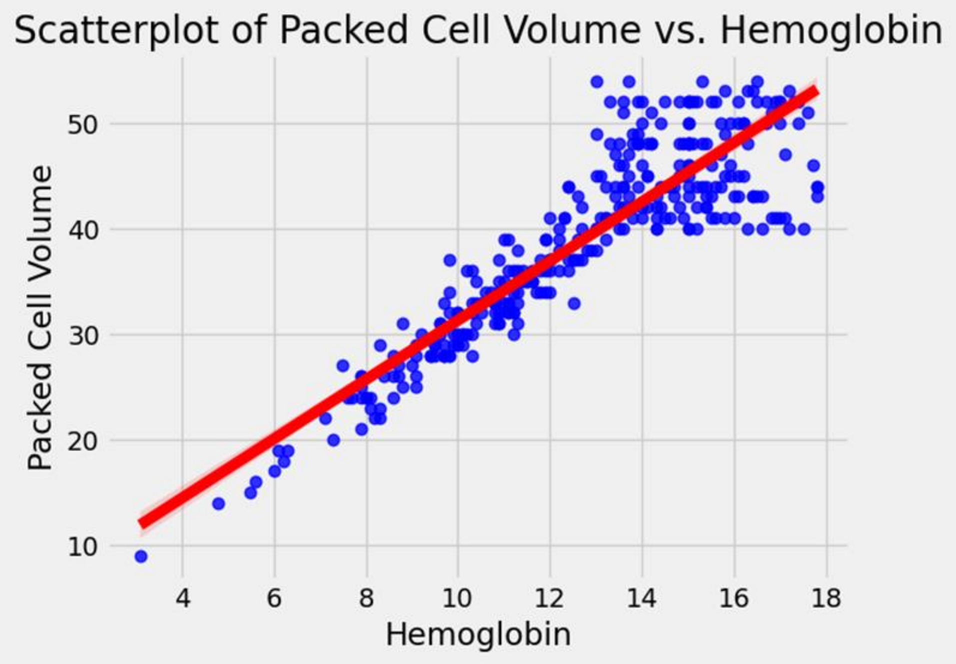


National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Packed Cell Volume vs Hemoglobin

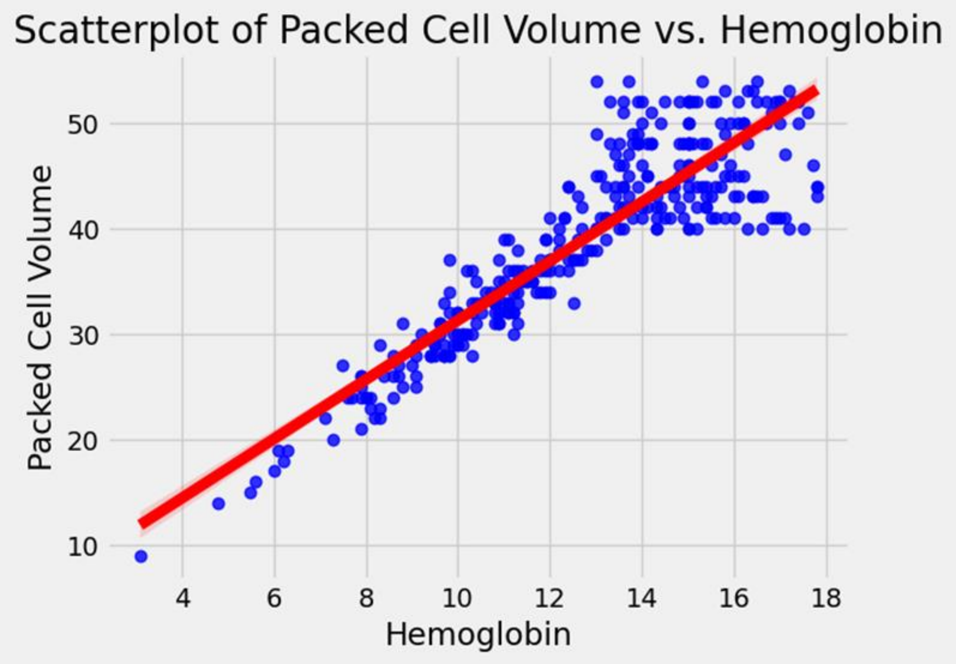
You are looking to understand the relationship between packed cell volume and hemoglobin data.

What did you learn after plotting?



Correlation

- Packed cell volume and hemoglobin data are correlated.
- Two or more measures are "correlated" when they have a mutual relationship or connection.

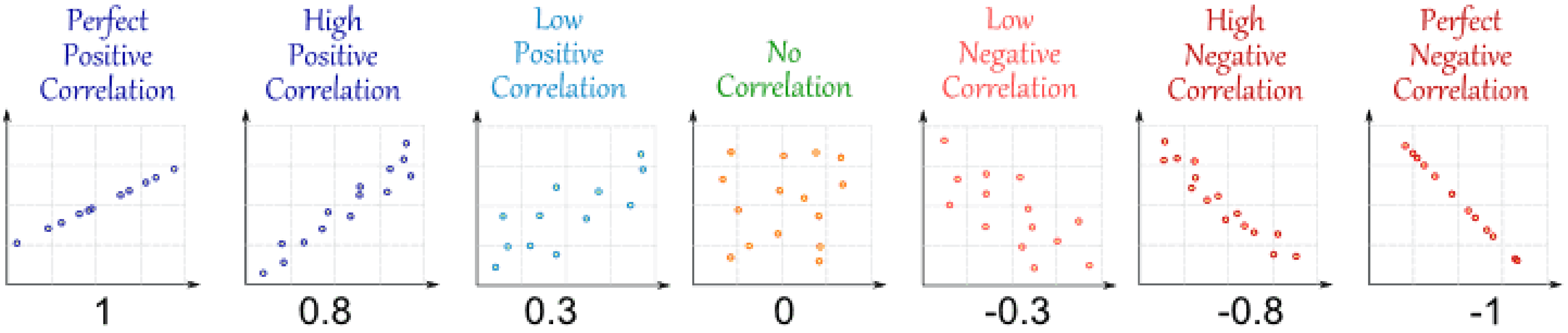


Source:

Correlation Coefficient

- Used to measure the strength of the relationship between two variables
- The most common coefficient used is r, which ranges from –1 to 1, and measures the linear correlation between variables

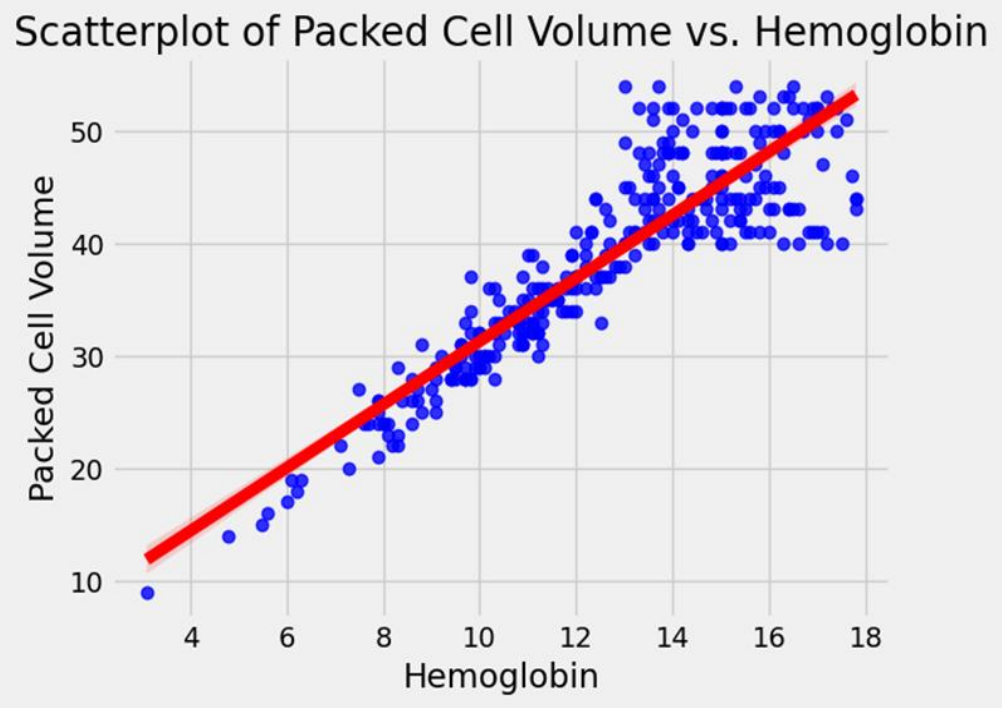
$$r = \frac{\operatorname{cov}(X, Y)}{S_X S_Y}$$



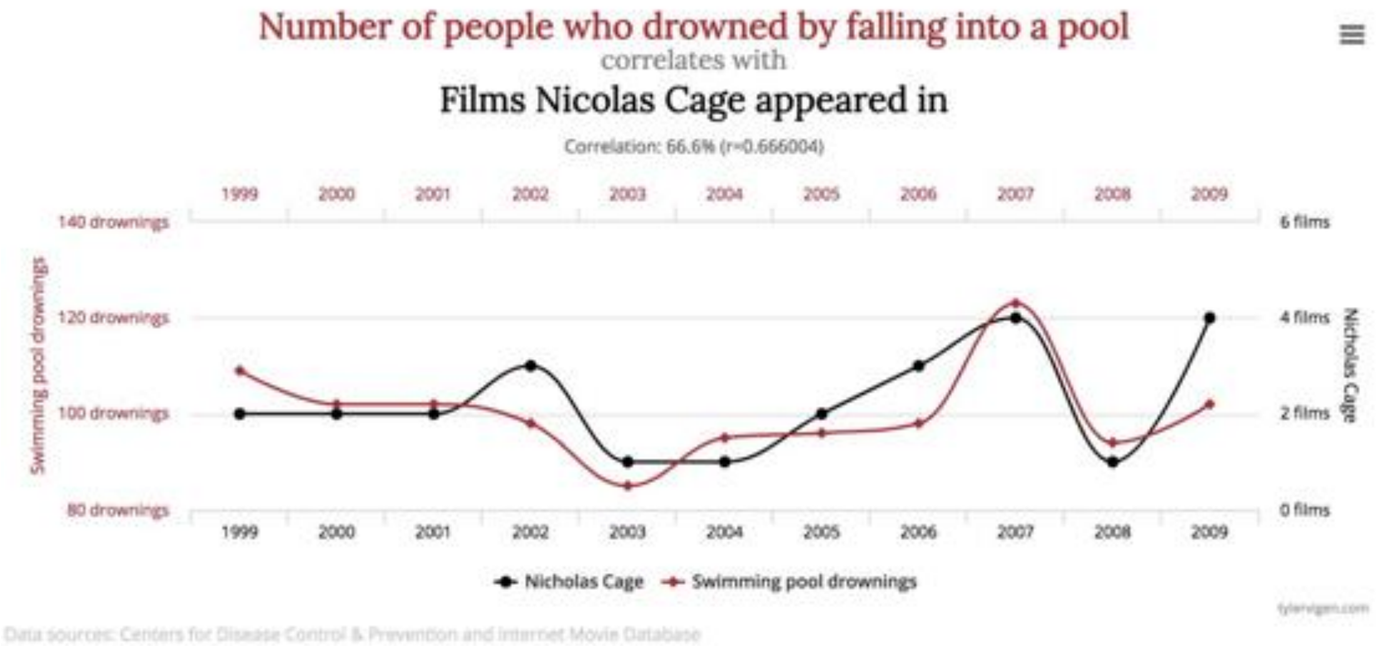
Source:

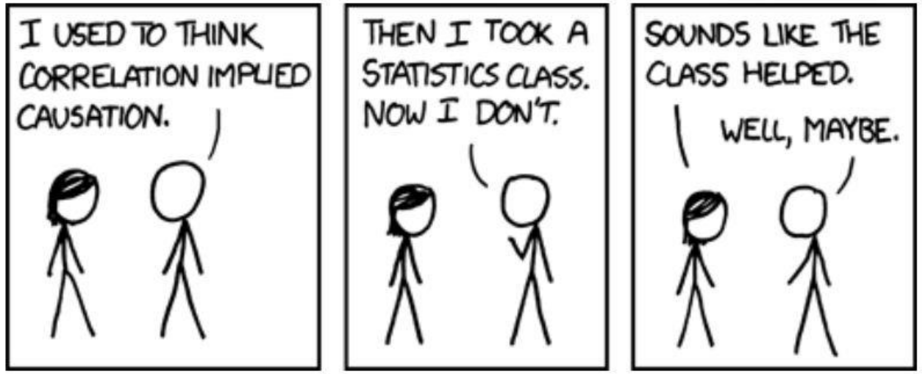
Correlation Coefficient – Kidney Disease Data

- Excel's CORREL function was used to calculate the correlation coefficient between pcv (packed cell volume) and hemo (hemoglobin).
- Correlation = 0.90



Correlation Does Not Imply Causation





Statistics vs Machine Learning

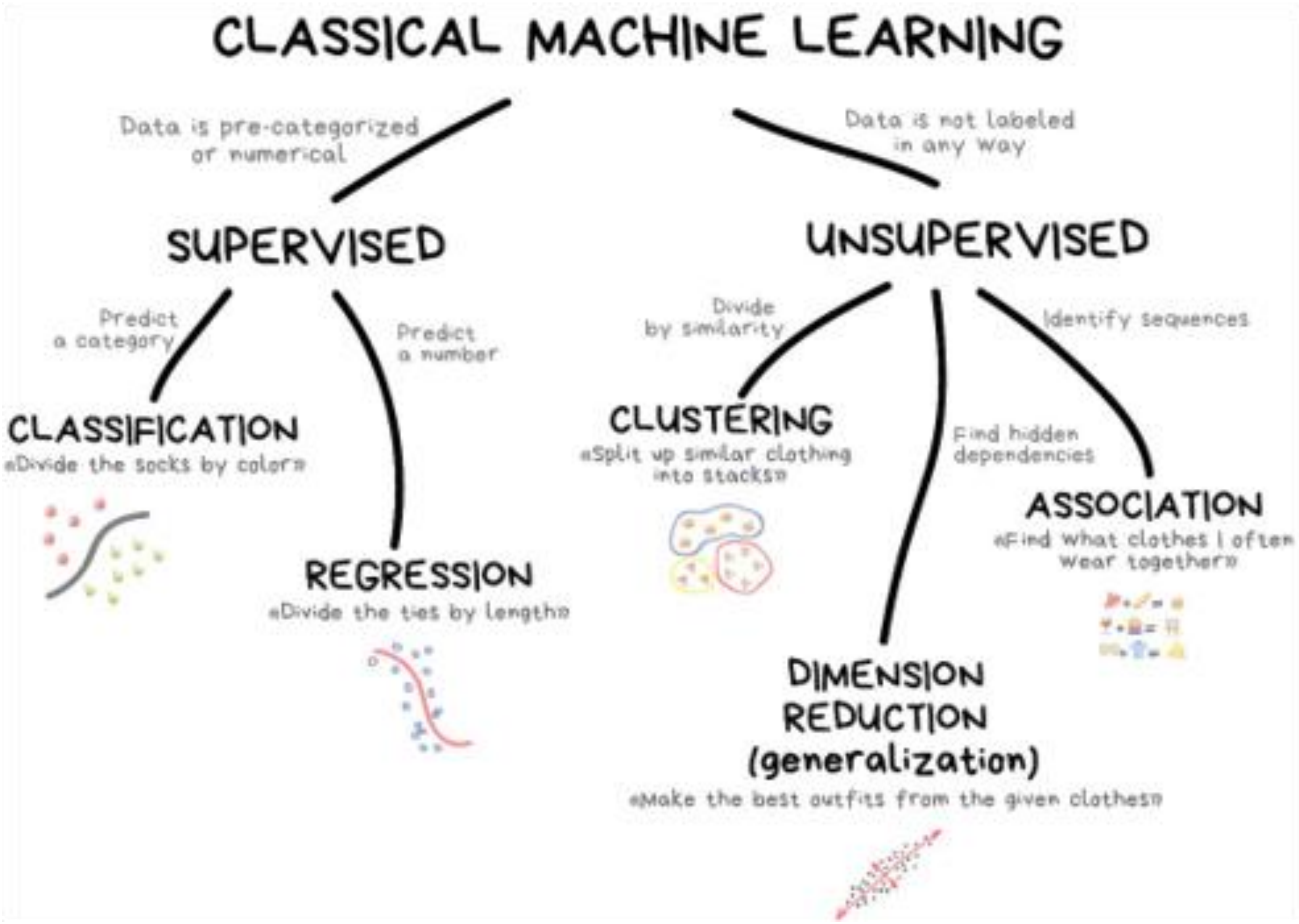
Aspect	Statistics	Machine Learning
Goal	Explains relationships in data	Make accurate predictions on data
Assumptions	Strong Assumptions (normality, linearity)	Often fewer assumptions; model learns patterns
Interpretability	High (regression coefficients)	Can be low (deep learning)
Data Size	Performs well with smaller datasets	Excels with larger datasets

Bridging Statistics and Machine Learning

- **Feature Engineering**: Use statistical insight to craft useful inputs
- **Model Evaluation**: Leverage statistical tests and metrics (confidence intervals, ROC/AUC)
- **Bias & Variance:** Concepts from statistics drive model tuning
- **Uncertainty Estimation**: Statistical thinking helps in probabilistic modeling
- **Explainability**: Statistics helps demystify complex models for stakeholders

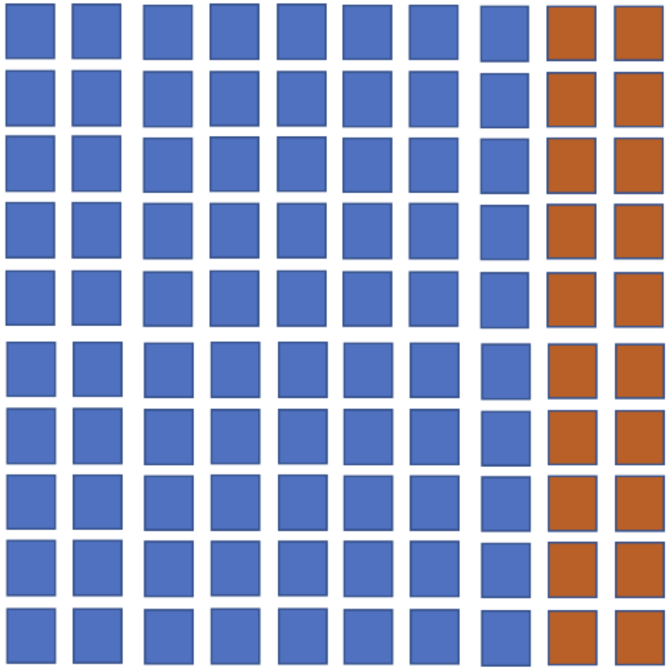
Model Metrics

Types of Models



Jumping to Machine Learning

80/20% train/test split



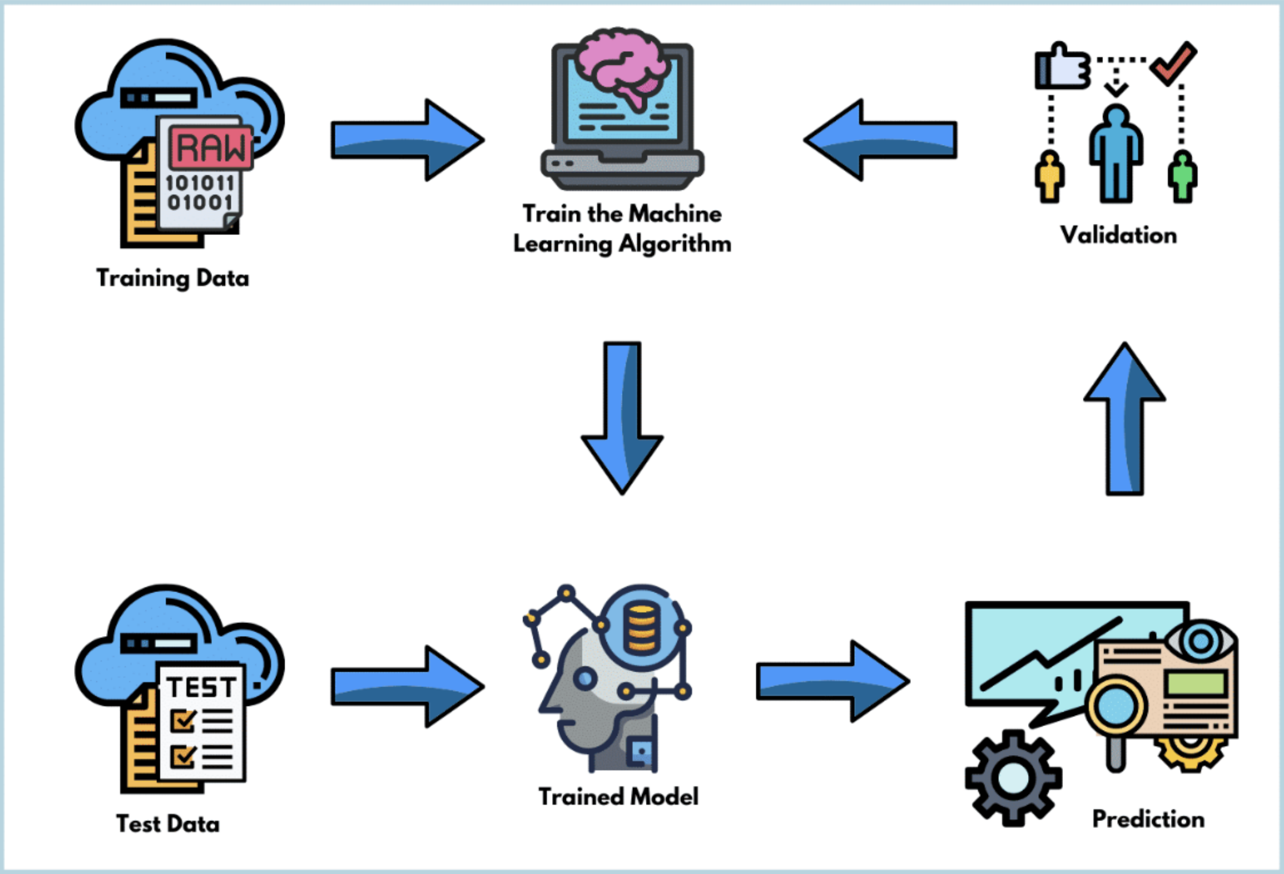
Training data is used to train a model (i.e., learn the best combination of parameters to minimize error)

Test data is used to assess the performance of a trained model (to determine if additional training is necessary)

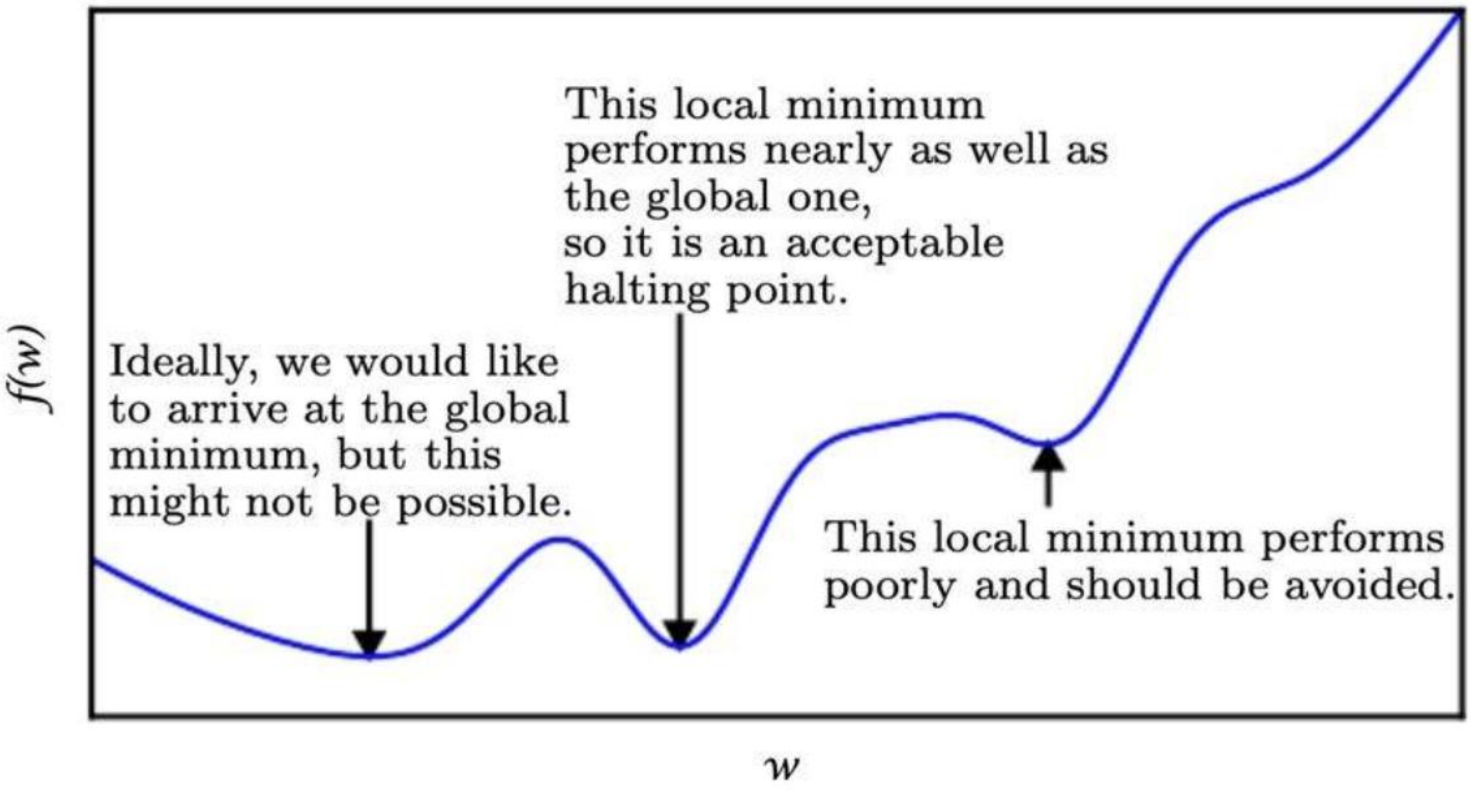
National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Machine Learning Model Fit



Loss Function

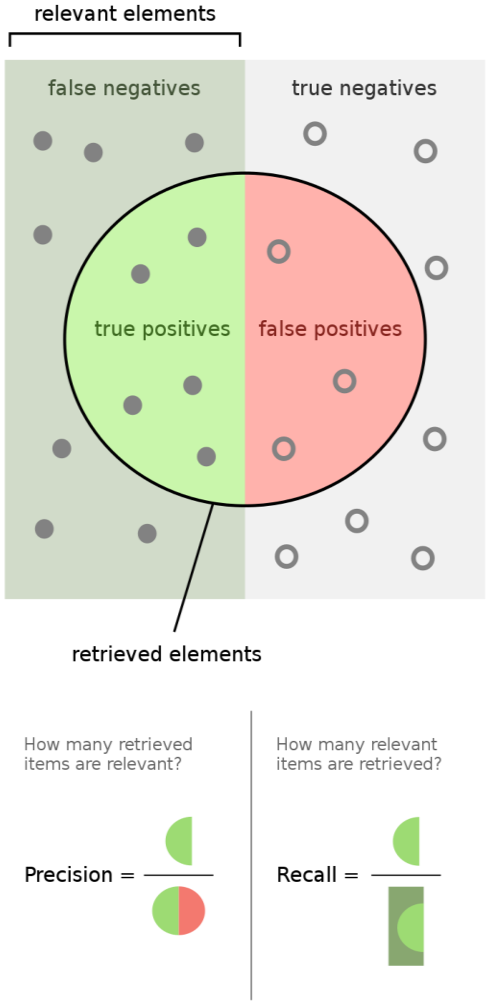


Measuring Model Performance

- 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

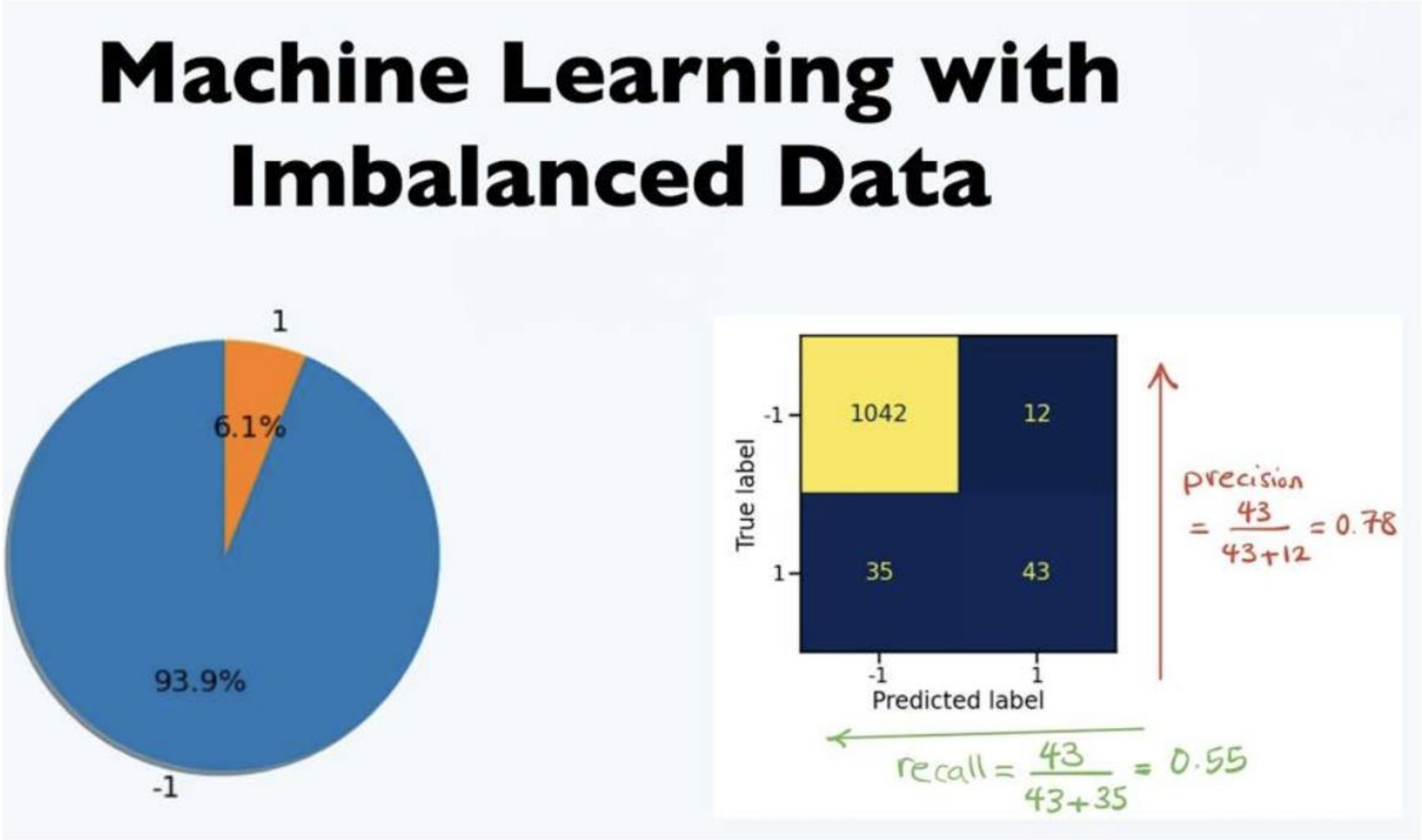
Classification Metrics

- **Accuracy** is a measure of how often a model gets a prediction right.
  - Accuracy can lead to misleading results in the case of an imbalanced dataset.
- **Precision** is the measure of how many relevant predictions were correct.
- **Recall** is the measure of how good the model was at identifying all relevant observations.
- Example
  - Precision: Of all the patients the model predicts to have kidney disease, how many actually had kidney disease?
  - Recall: Of all the patients that actually have kidney disease, how many of those did the model successfully predict to have kidney disease?



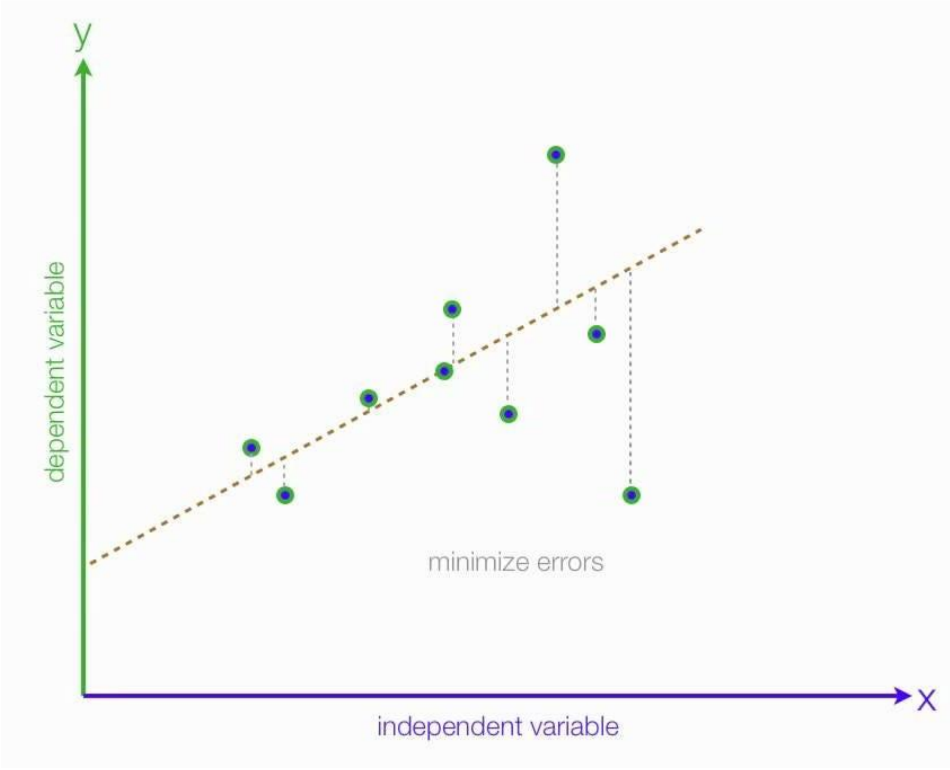
Source:

Machine Learning with Imbalanced Data



Regression Metrics

- **Mean Absolute Error (MAE)** represents the average of the absolute difference between actual and predicted values.
- **Mean Squared Error (MSE)** measures how close the fit line is to a set of data points, it will penalize larger errors more severely by squaring the difference.
- **Root Mean Squared Error (RMSE),** the square root of MSE, measures the standard deviation of residuals.
- **R squared** is the coefficient of determination representing the proportion of the variance in the dependent variable.



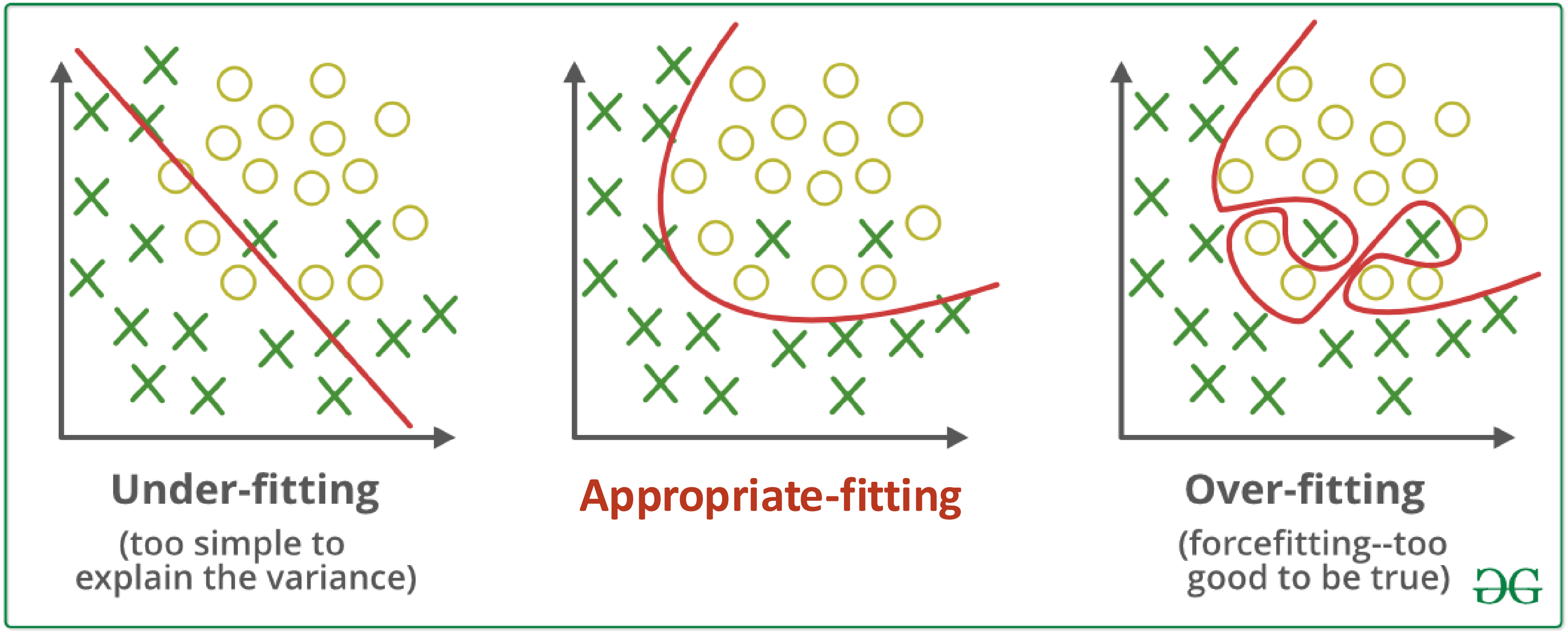
Source:

National Institute of Diabetes and Digestive and Kidney Diseases
Central Repository

Overfitting and Underfitting

Modeling

- Underfitting: biased sampling or biased models
- Overfitting: ignoring natural variance in the data
- The goal is to be somewhere in the middle



Source:

Overfitting Explained

Source:

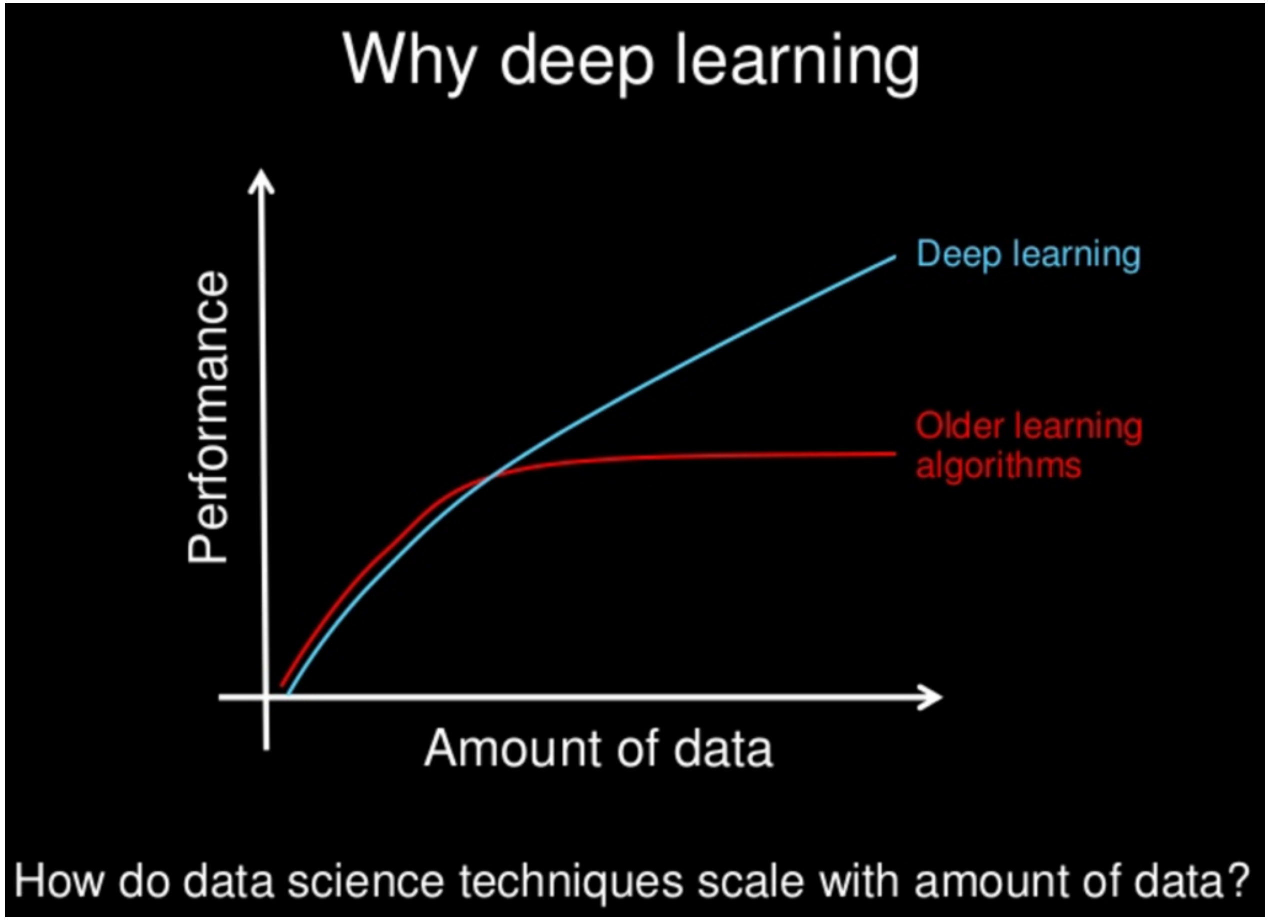
Session Break

Office of Clinical Research Support

REPOSITORY · TECHNOLOGY ADVANCEMENT · BIOSTATISTICS · CLINICAL STUDIES · REGULATORY

Neural Networks and Their Applications

Why Is Deep Learning So Hot Right Now?



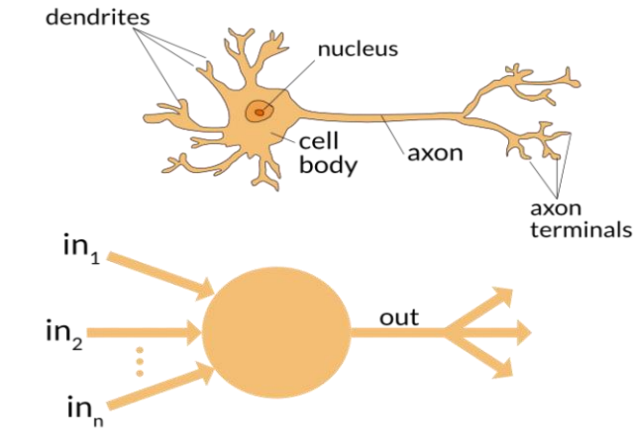
Deep Learning

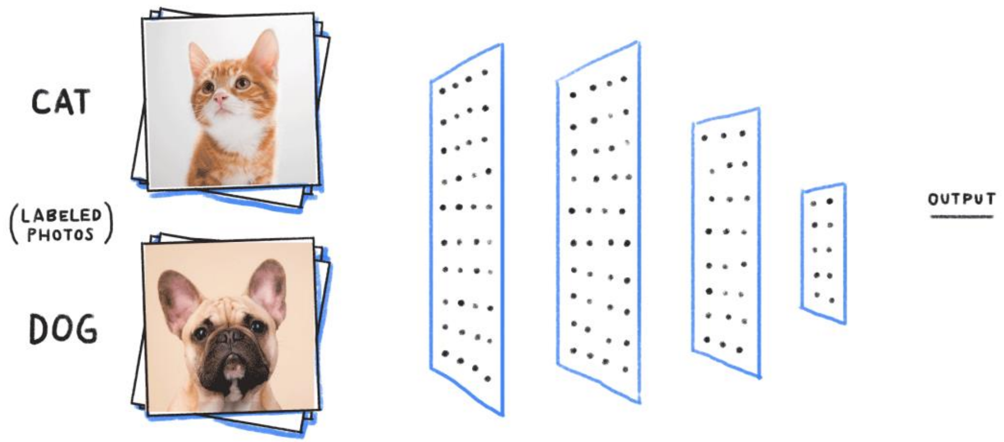
"…hierarchy of concepts allows the computer to learn complicated concepts by building them out of simpler ones."
– Ian Goodfellow

Deep Learning Inspired Partly from Biology

Human neuron

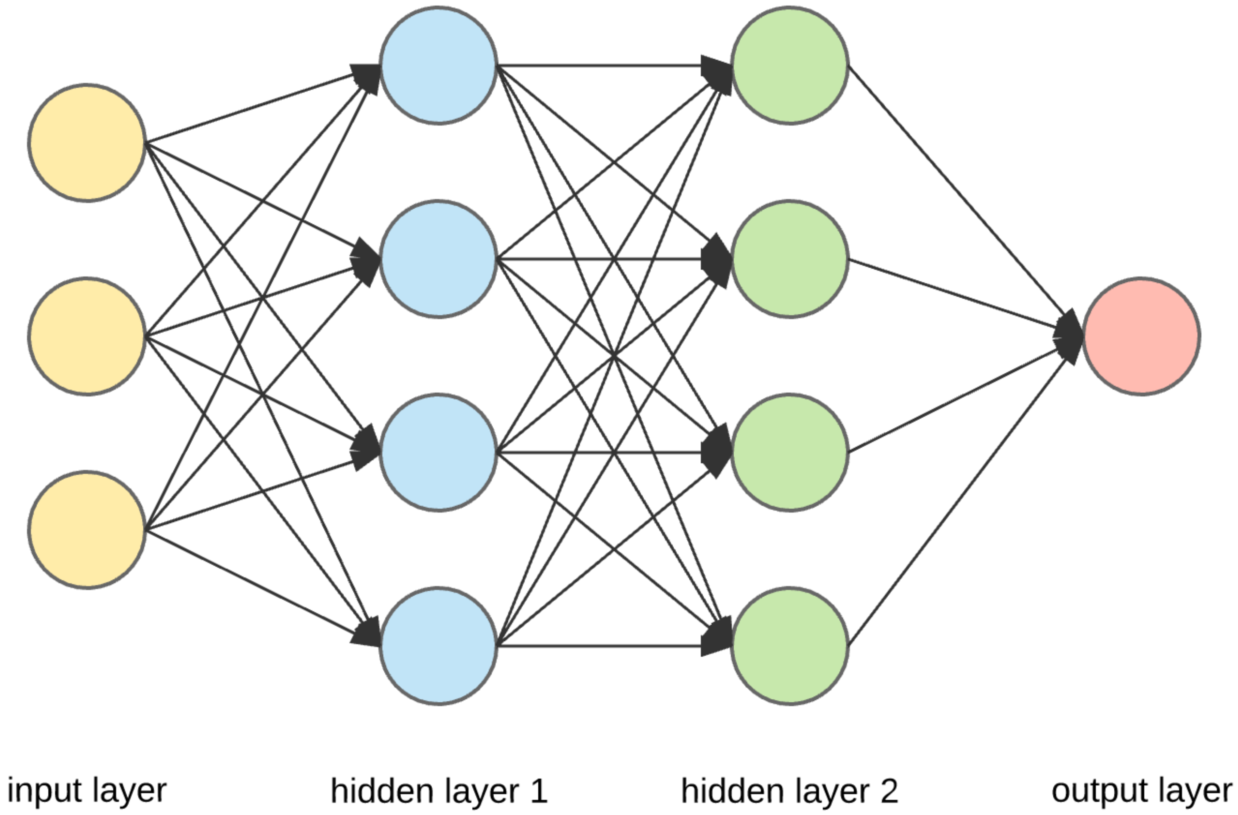
Neural network node





“Feed Forward” Neural Networks

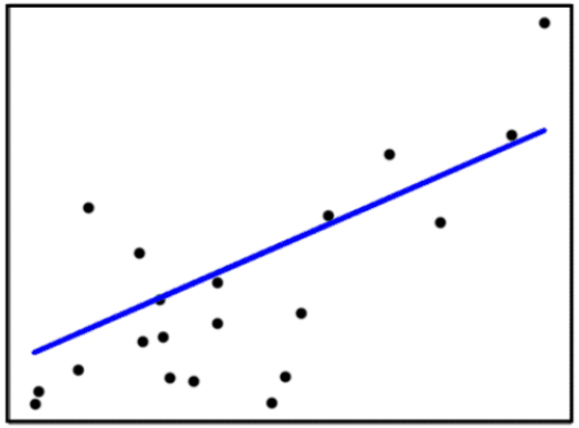
- This is the most basic, vanilla form of neural network that all other neural networks use as a foundation.
- Number of layers and nodes/neurons per layer is a choice made by network's architect(s).
- Each node is essentially a perceptron.



Source:

NN Increasing Complexity

Linear Regression

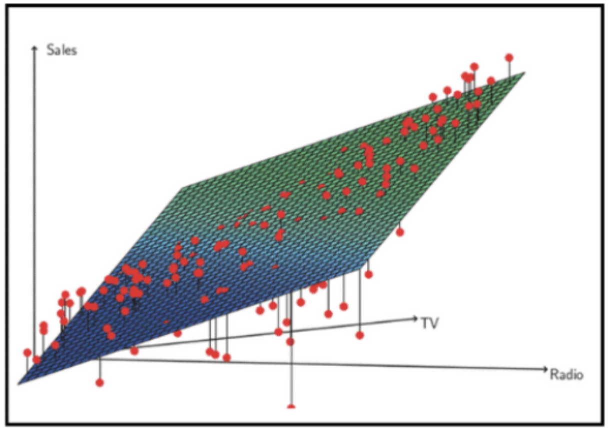


$$\hat{y} = \theta_1 X + \theta_0$$

Models *one* input (X) to *one* predicted output ($\hat{y}$) using *two* parameters (θ_1, θ_0).

NN Increasing Complexity

Multiple Linear Regression

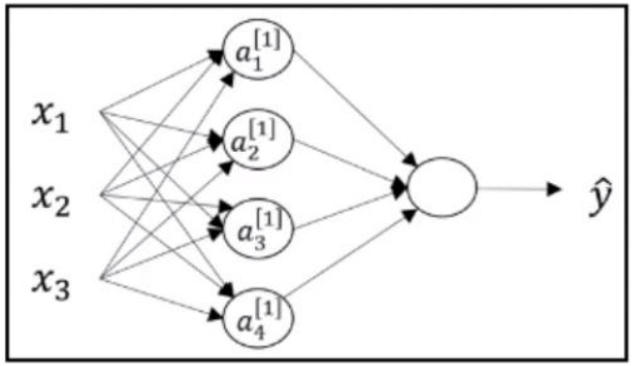


$$\hat{y} = \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_0$$

Models *multiple* inputs (X_1, X_2, X_3) to *one* predicted output ($\hat{y}$) using *several* parameters (θ_1, θ_2, θ_3,θ_0).

NN Increasing Complexity

Neural Network

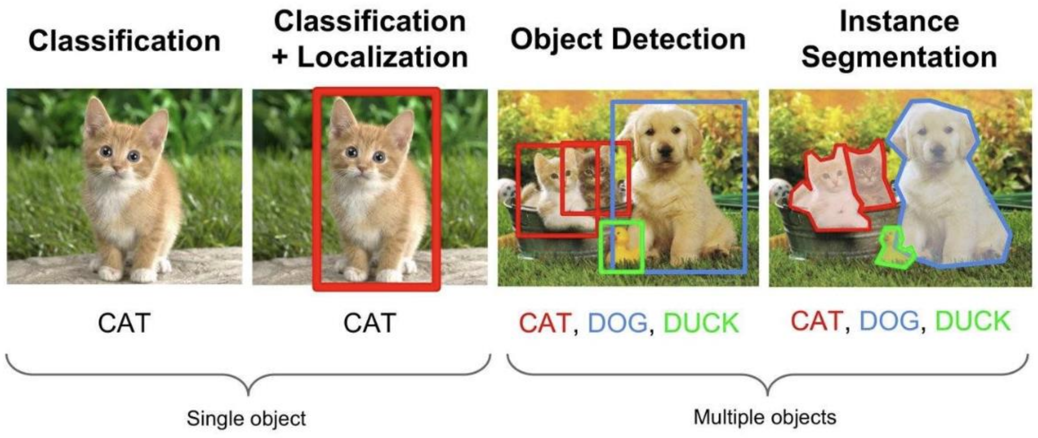


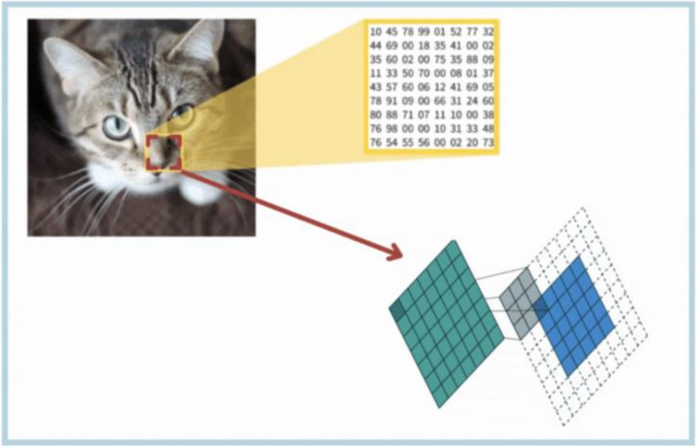
$z_\mathrm{i} = w_i X_i + b_i$ with $a_i = \sigma(z_i) \rightarrow \hat{y}$

Models *multiple* inputs (X_i) to *multiple* outputs ($\mathrm{z_i}, \hat{y}$) using *many* model parameters ($w_\mathrm{i}, \mathrm{b_i}, \mathrm{a_i}$), ultimately resulting in a prediction.

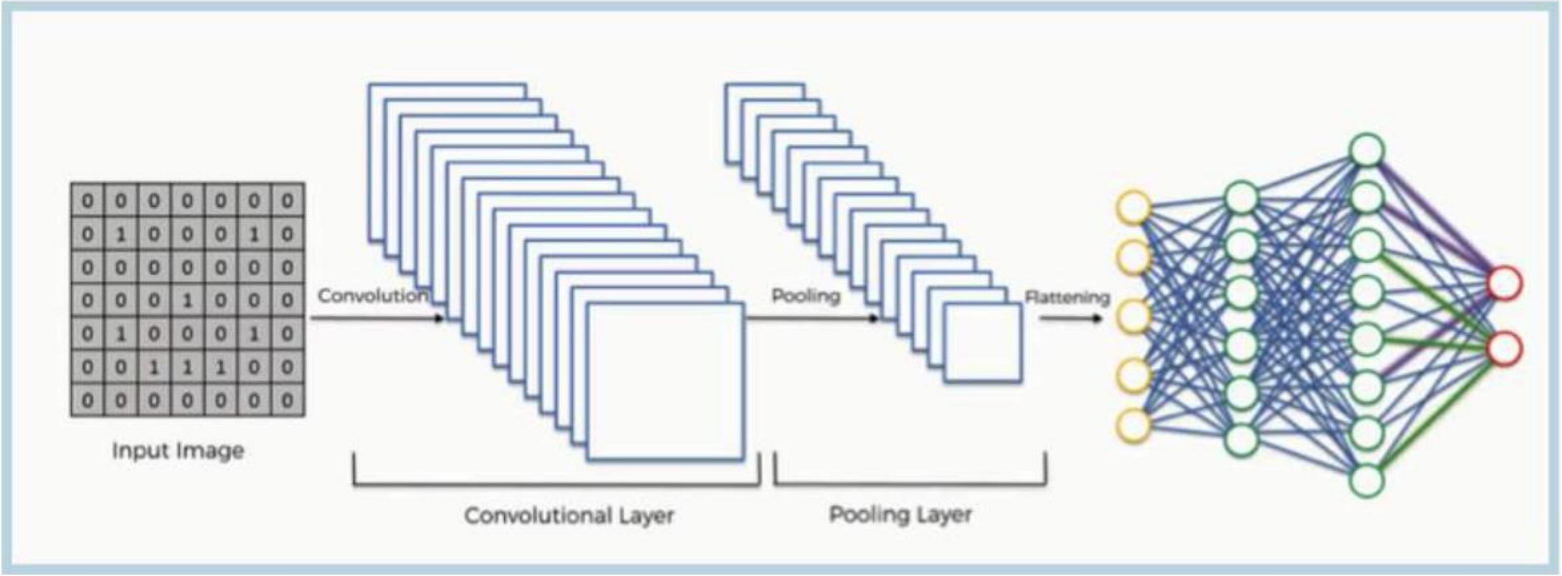
Convolutional Neural Networks

- A convolutional neural network (CNN) is a further advancement of the basic feed forward neural network.
- CNNs pass inputs through numerous connected layers and include additional layers that perform the "convolutional" operation to identify important visual elements in an image.



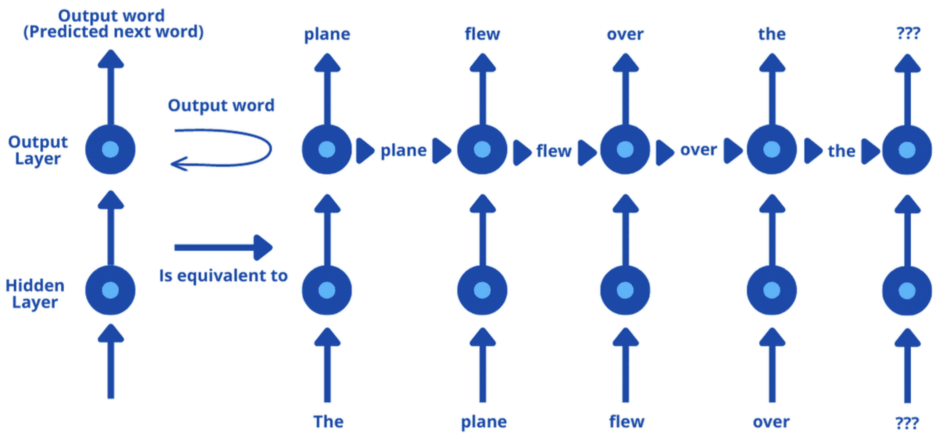


CNN Classification



Recurrent Neural Networks

- A recurrent neural network (RNN) is a neural network with a feedback mechanism (as opposed to a simple "feed forward" approach) .
- "Long Short-Term Memory" (LSTM) are a common component of RNNs
  - LSTM are capable of "remembering" events many steps in the past
  - Can aid in recognizing context in series type data (e.g., natural language)



Notice how the network uses a given word to predict the next word in the sequence. RNN's are a natural choice for sequential data, and language is a type of sequential data.

RNN Applications

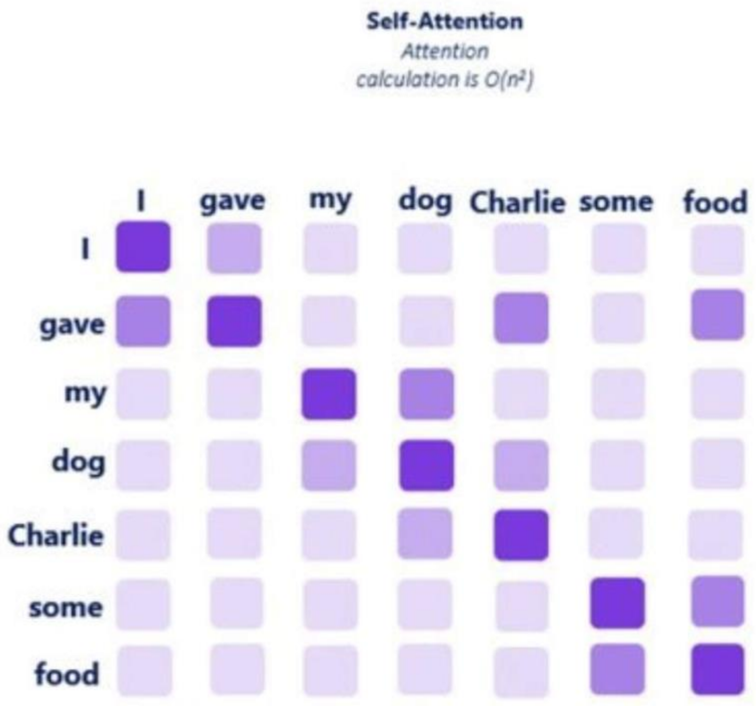
- RNNs are widely used for AI text and speech applications (NLP)
  - Next-word prediction
  - Translation
  - Speech-to-text and text-to-speech
  - Sentiment analysis
  - Summarization

Transformer Learning

- Transformer learning diverges from traditional feed forward mechanisms and embraces a self-attention mechanism to efficiently process input sequences. Many modern Large Language models (LLMs) use a transformer architecture.
- Introduced in the paper “Attention is All You Need” by Vaswani in 2017 and has revolutionized the field of NLP.
- Transformers are made up of billions or trillions of parameters

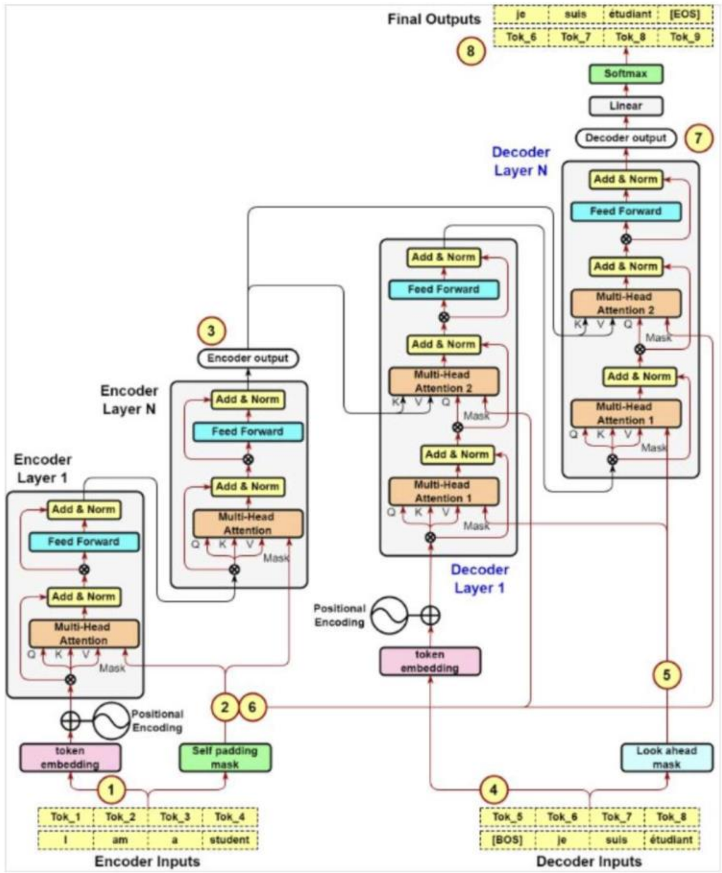
Self-Attention Mechanisms

- Self-attention mechanisms, allow the transformer network to weigh the importance of different parts of the input data independently
  - Can process entire sequences simultaneously, making them faster and more efficient



Transformer Applications

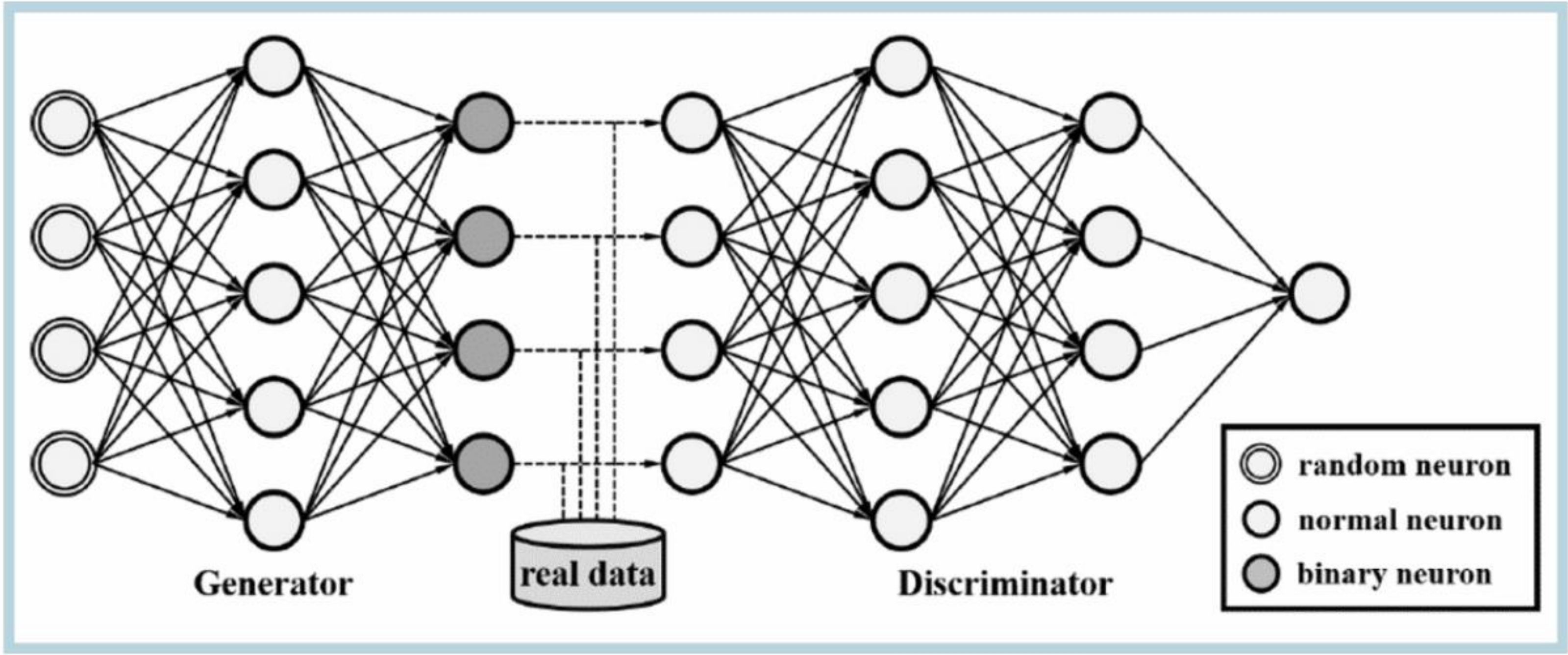
- Transformers are well suited for:
  - Machine Translation
  - Text Summarization
  - Language Modeling



Generative Adversarial Networks

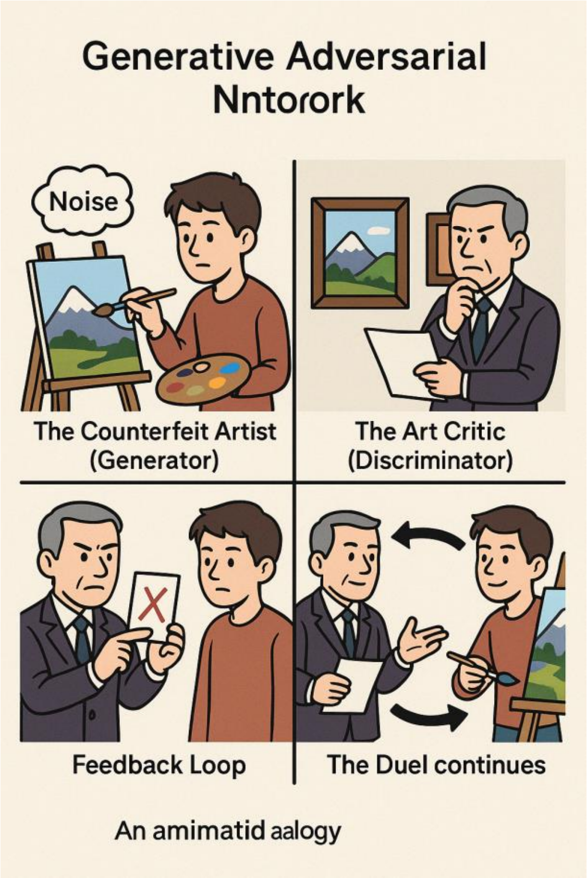
- Generative adversarial networks (GANs) consist of two networks, a “generator” and “discriminator”
- GANs can generate “new” data instances that resemble the dataset
- Training requires extra steps:
  - Discriminator is trained to distinguish between real and fake data (created by generator)
  - Generator is trained to fool the discriminator (how they are “adversarial”)
  - Output from generator or real data is fed into discriminator > training is preformed

GAN Setup


Generator
real data
Discriminator
random neuron
normal neuron
binary neuron

GAN Applications

- Transform data in useful ways:
  - Text-to-image generation
  - Face aging
  - Generate realistic photographs
  - Photograph editing
  - Semantic segmentation
  - 3D object generation
  - Video Prediction



Adversarial Purposes?

Potential consequences of GANs with respect to security?

Adversarial Purposes?

Potential consequences of GANs with respect to security?

- GAN- generated fake medical images
- Manipulated clinical data
- Privacy erosion & re-identification risks
- Erosion of trust in biomedical institutions
- Legal & regulatory challenges

Day 2 Summary

- Statistics Primer
- Model Metrics
- Neural Network Overview & Applications
  - Feed-forward
  - Convolutional
  - Recurrent
  - Transformer
  - Generative Adversarial

Questions?

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Contacts:

- Arica Christensen – christensen_arica@bah.com
- Dr. Gordon Aiello – aiello_gordon@bah.com

Upcoming Webinar: FAIR and AI-Ready Data Sharing

- **Date:** May 29th from 2-3pm ET
- **Experts:** Dr. Courtney Shelley, Anya Dabic
- **Scan the QR code register**

Thank You!