2019 HEALTH HACK COMPETITION

INNOVATION4 HEALTH

Data Science Workshop

Data Science Workshop

Instructor - Philip Ciunkiewicz

Before We Start

Experience with data analysis

- Organizing data in Excel
- Visualizing data in Excel
- Computing basic statistics
  - Mean, median, mode
  - Standard deviation, uncertainty
- More advanced analysis
  - Transforming columns
  - Operations between columns

Experience with programming

- Exposure to programming basics
  - Concept of variables
  - Concept of functions
- Working in Python / R / Matlab
  - Defining variables
  - Defining functions
  - Loading and saving data
  - Using external libraries

Interactive Notebook

The interactive notebook for this workshop is available on GitHub

https://github.com/PCiunkiewicz/I4H_Workshop

Presentation slides will also be there

Presentation Flow

Data Science Basics

- What is data science?
- Applications of data science
- Statistics and analytics vs machine learning

Data Preparation

- Working with different types of data
- How data is structured
- How machines interpret data

Machine Learning

- Core principles of machine learning
- Machine learning project workflow
- Different machine learning tasks

Presentation Flow

Data Science Basics
- What is data science?
- Applications of data science
- Statistics and analytics vs machine learning

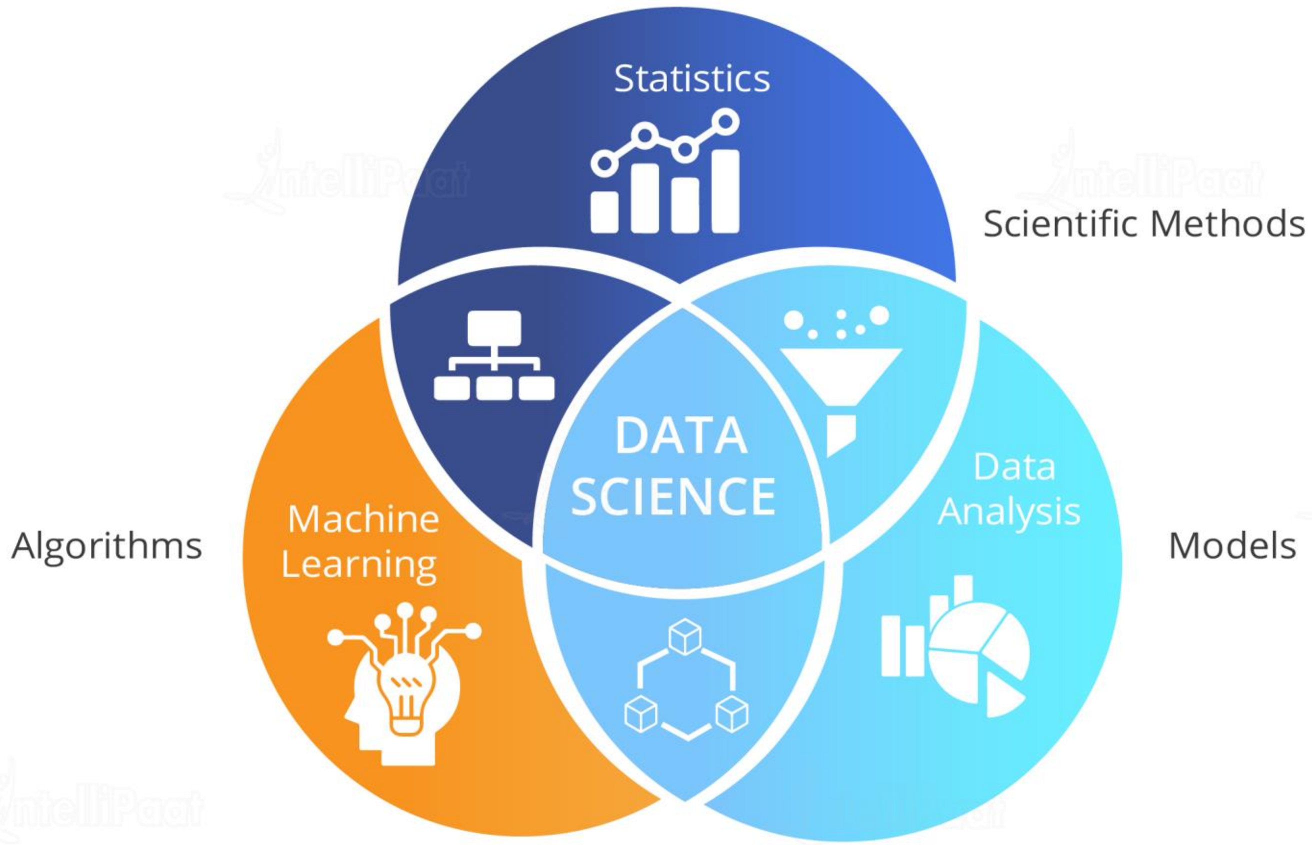
Data Preparation
- Working with different types of data
- How data is structured
- How machines interpret data

Machine Learning
- Core principles of machine learning
- Machine learning project workflow
- Different machine learning tasks

Data Science Basics

Defining and understanding the field

What is Data Science?



https://intellipaat.com/blog/what-is-data-science/

- No rigid / formal definition for data science

- Two popular ideas
  - Applying the scientific method to data
  - Transforming data into useful knowledge

Applications of Data Science

Types of Applications

- Classification
- Regression
- Clustering

- Image and audio processing
- Natural language processing

Fields of Use

- Science
  - Experimental and theoretical
  - Clinical applications

- Engineering and Industry
  - Research and development
  - Business intelligence

Applications of Data Science - Classification

- Classification problems focus on the prediction of a discrete variable

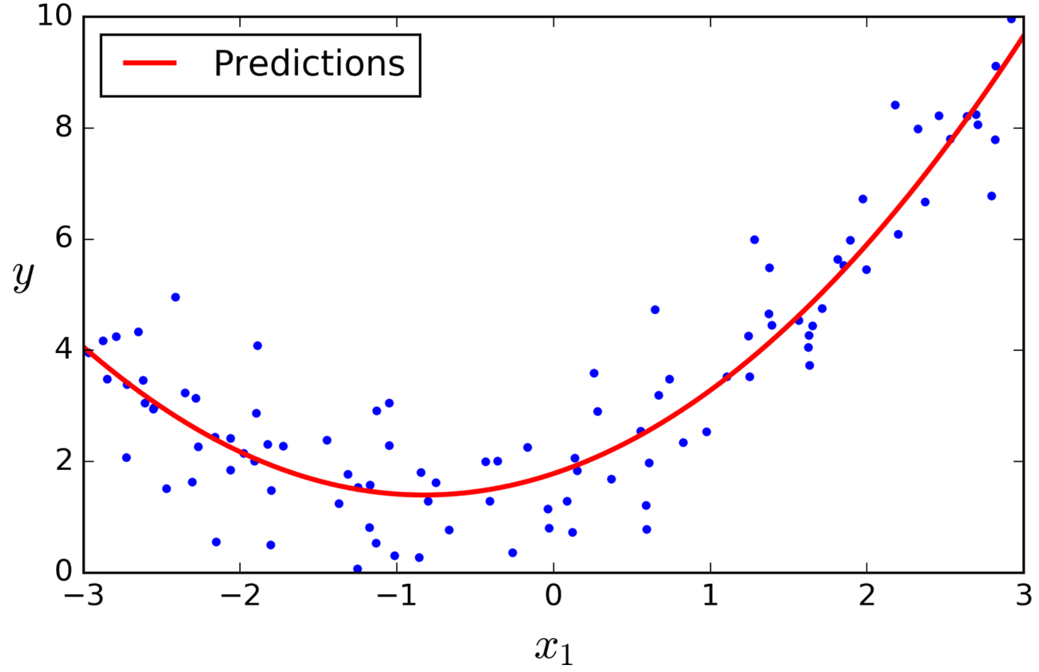
Dog

Cat

https://medium.com/@gnabr/machine-learning-c28daf3cf60a

Applications of Data Science - Regression

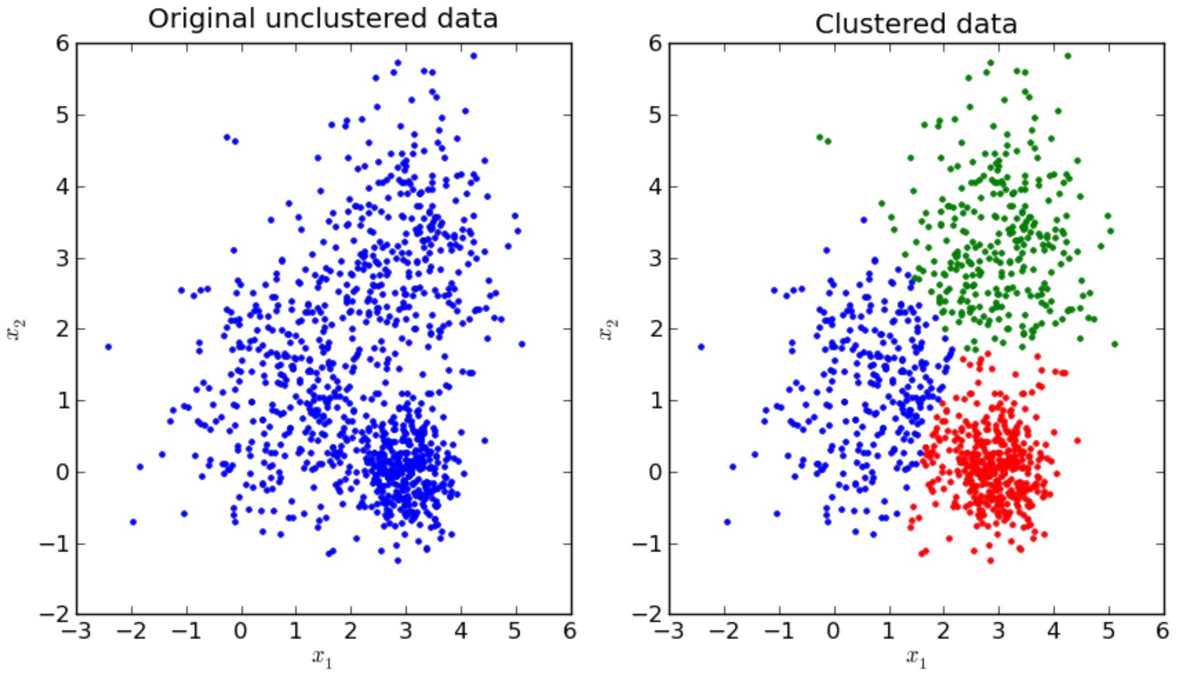
- Regression problems focus on the prediction of a continuous variable



https://learning.oreilly.com/library/view/hands-on-machine-learning/9781491962282/ch04.html

Applications of Data Science - Clustering

- Clustering problems focus on identifying similarities in populations



http://mathalytics.blogspot.com/2015/04/k-means-clustering-machine-learning.html

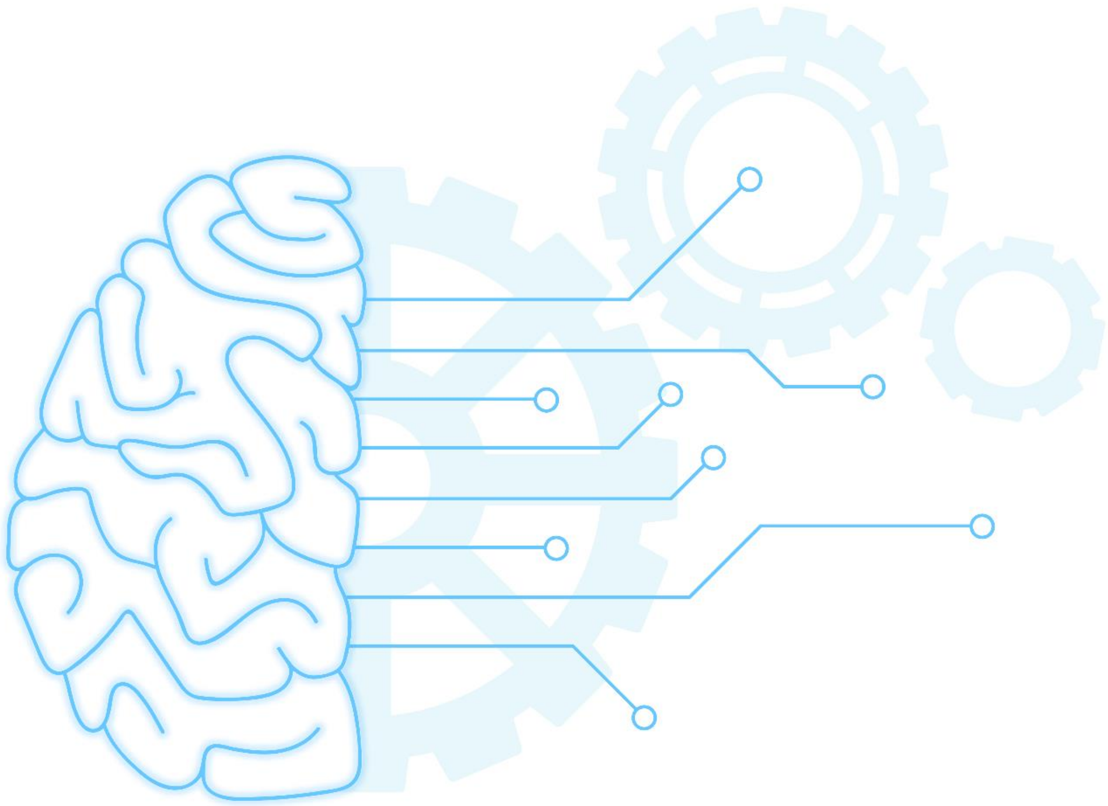
Statistics and Analytics vs Machine Learning

Statistics and Analytics

- Subfield of mathematics
- Offers more insight than ML
  - Formalizes existing relationships
  - Offers less predictive ability than ML
- Results based on formal definitions
  - Explicit mathematical relationships
- Strong results from minimal data

https://www.iconfinder.com/icons/1828922/

Statistics and Analytics vs Machine Learning



https://docs.microsoft.com/en-us/windows/ai/windows-ml/

Machine Learning

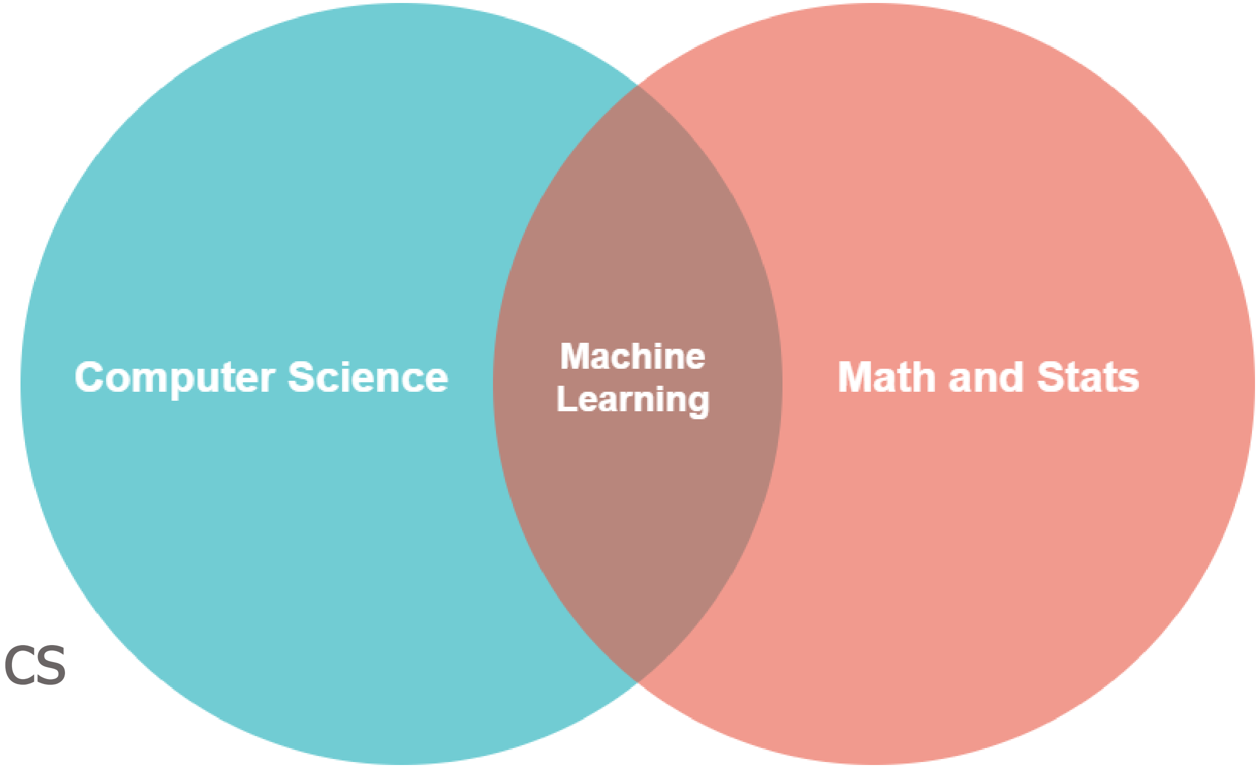
- Subfield of computer science
- Offers more predictive ability than stats
  - Predicts future relationships
  - Offers less insight than stats
- Results based on stochastic processes
  - Implicit relationships from optimization
- Requires lots of data for strong results

However…

There are also many similarities

- Applications
- Required knowledge
- Overall complexity

Machine learning is built on statistics



Presentation Flow

Data Science Basics

- What is data science?
- Applications of data science
- Statistics and analytics vs machine learning

Data Preparation

- Working with different types of data
- How data is structured
- How machines interpret data

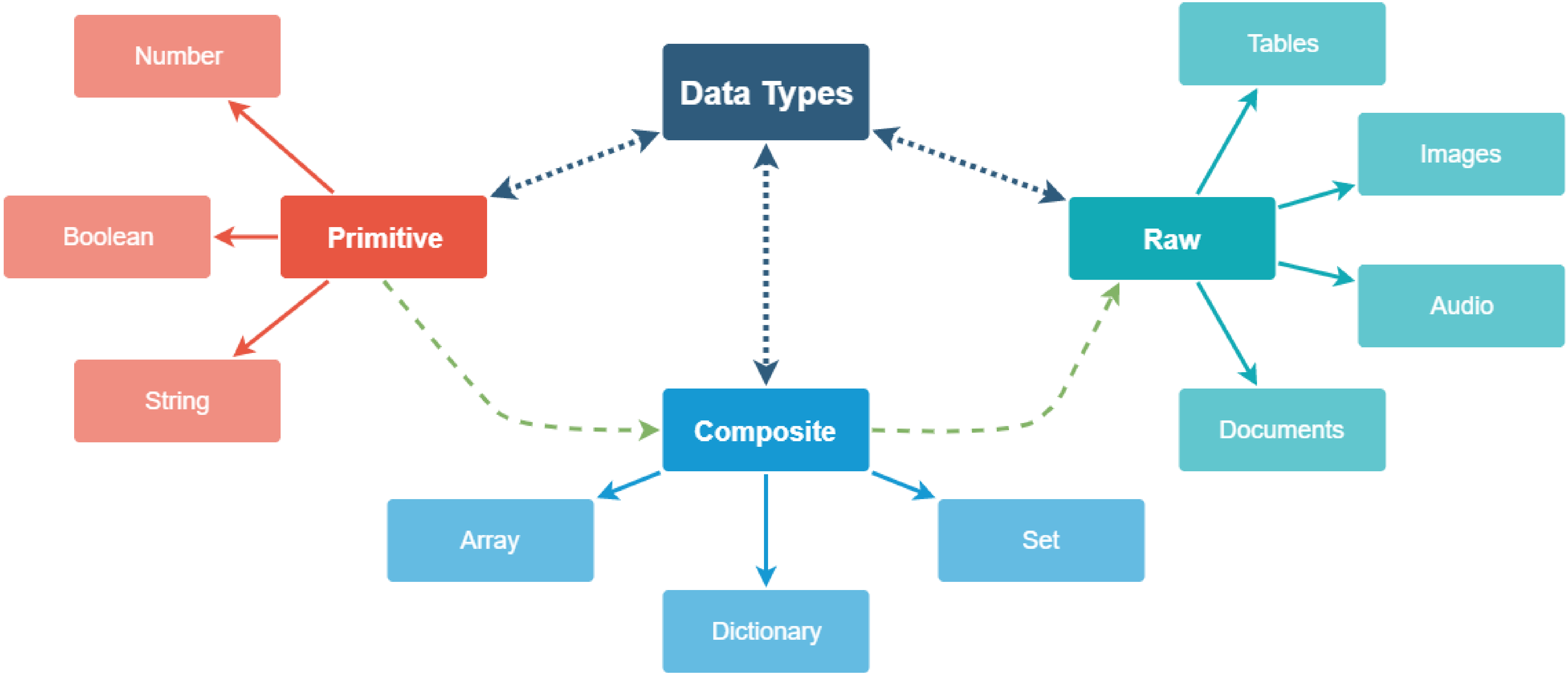
Machine Learning

- Core principles of machine learning
- Machine learning project workflow
- Different machine learning tasks

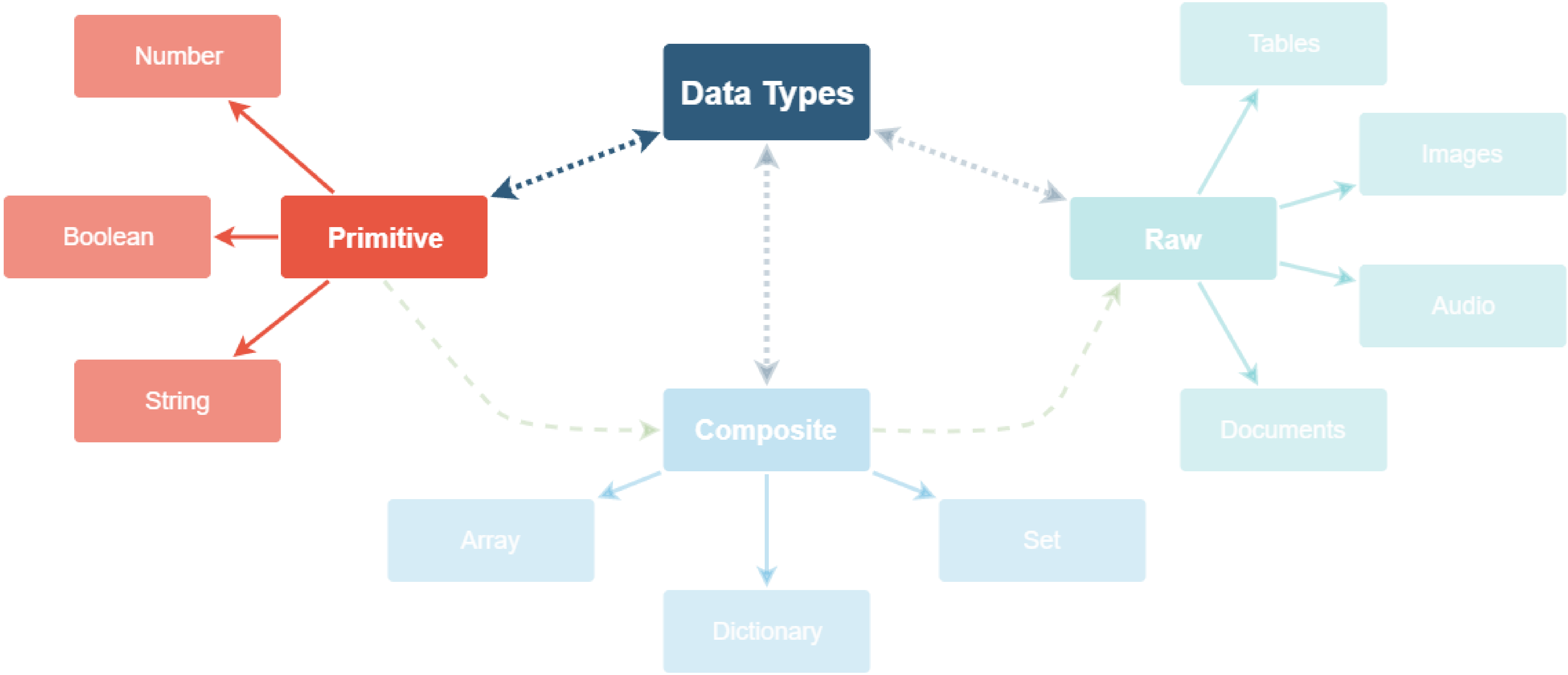
Data Preparation

Making data “digestible” for machines

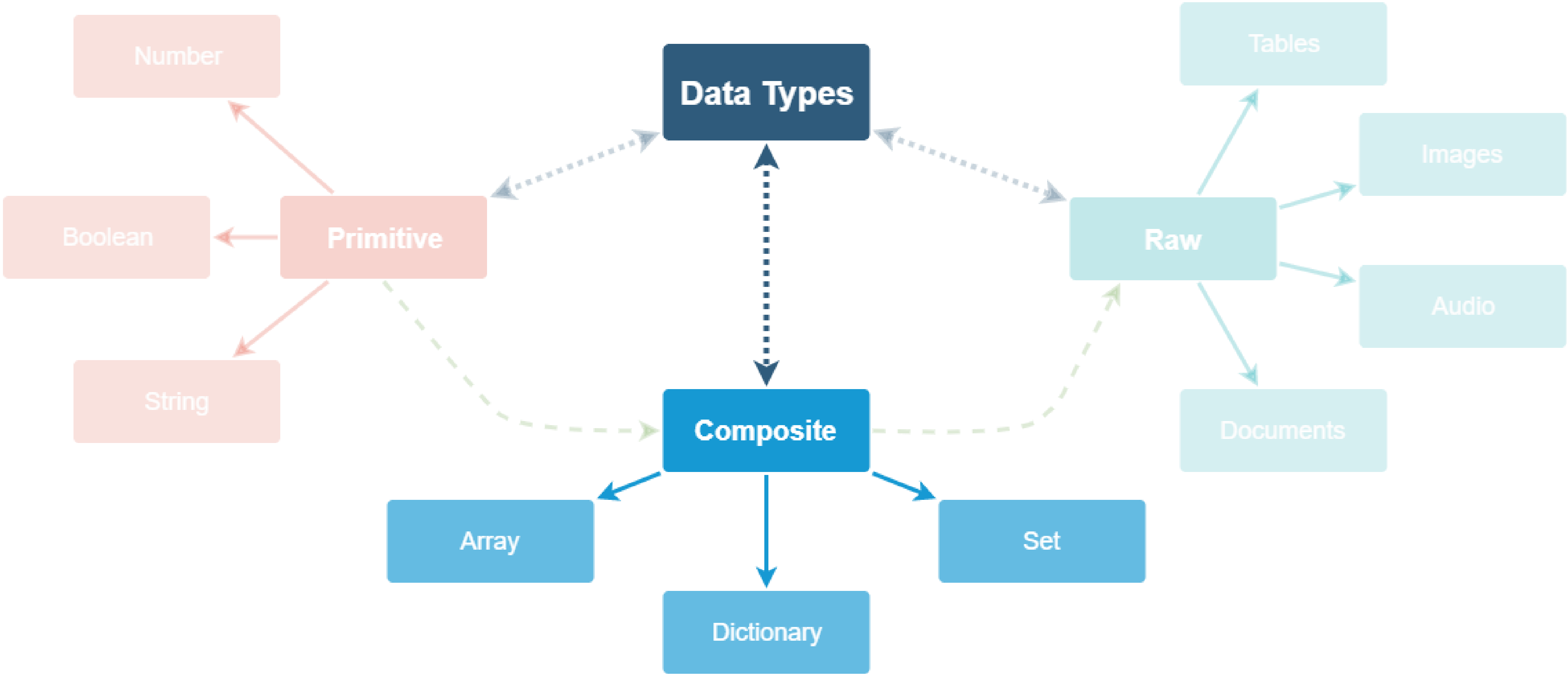
Working with Different Types of Data



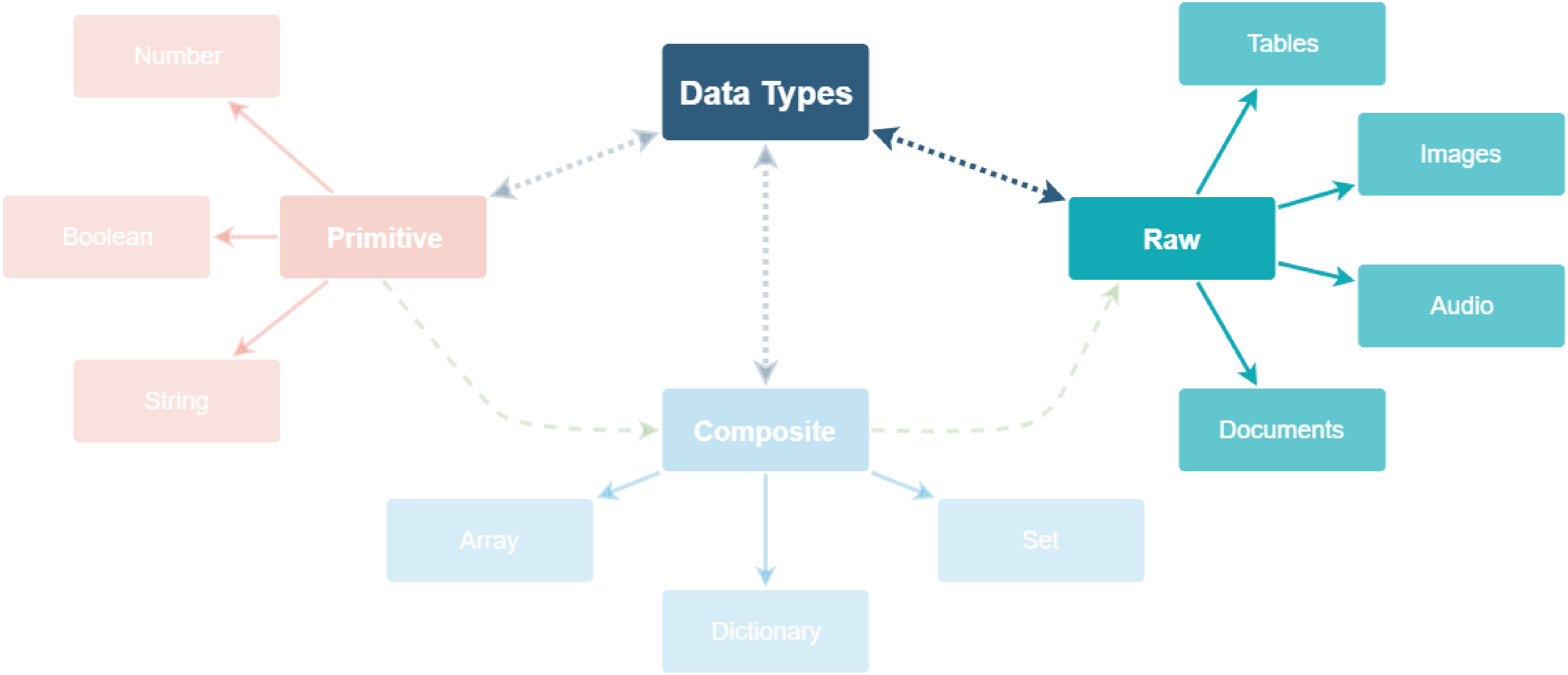
Working with Different Types of Data



Working with Different Types of Data



Working with Different Types of Data

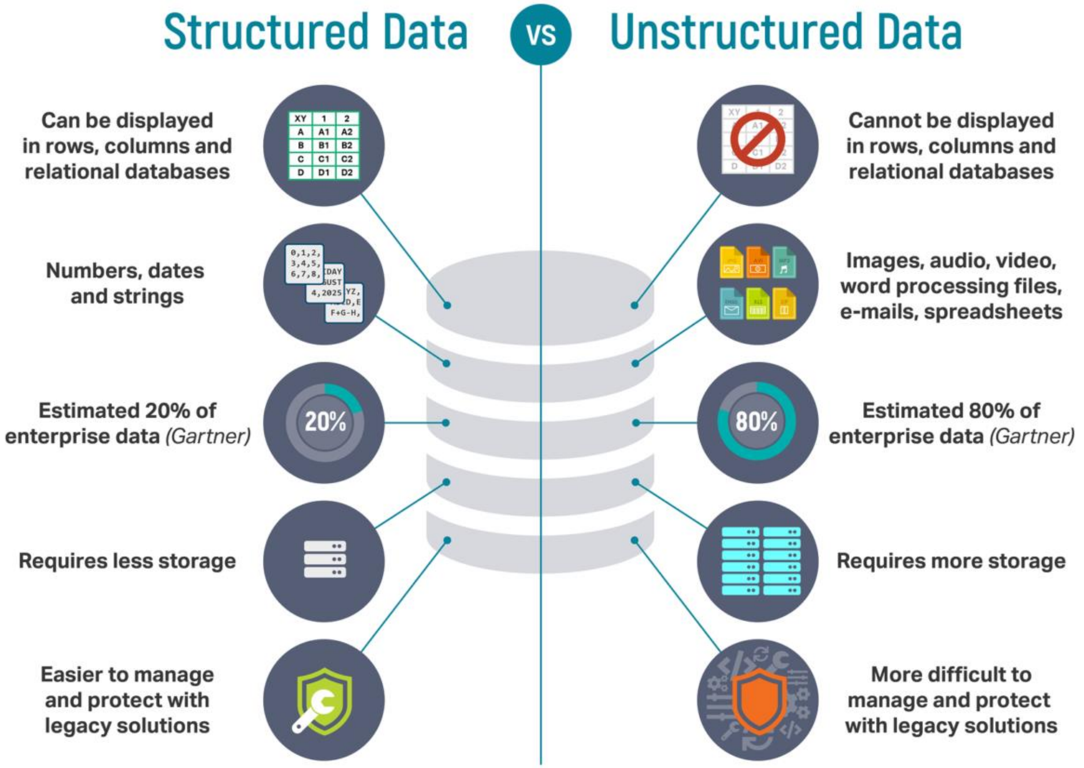


INNOVATION4 HEALTH

How is Data Structured?

Structured

- Structured data follows “Schema”
  - Highly organized and consistent
- Easily interpretable by machines
- Basis for many databases
  - Structured Query Language (SQL)

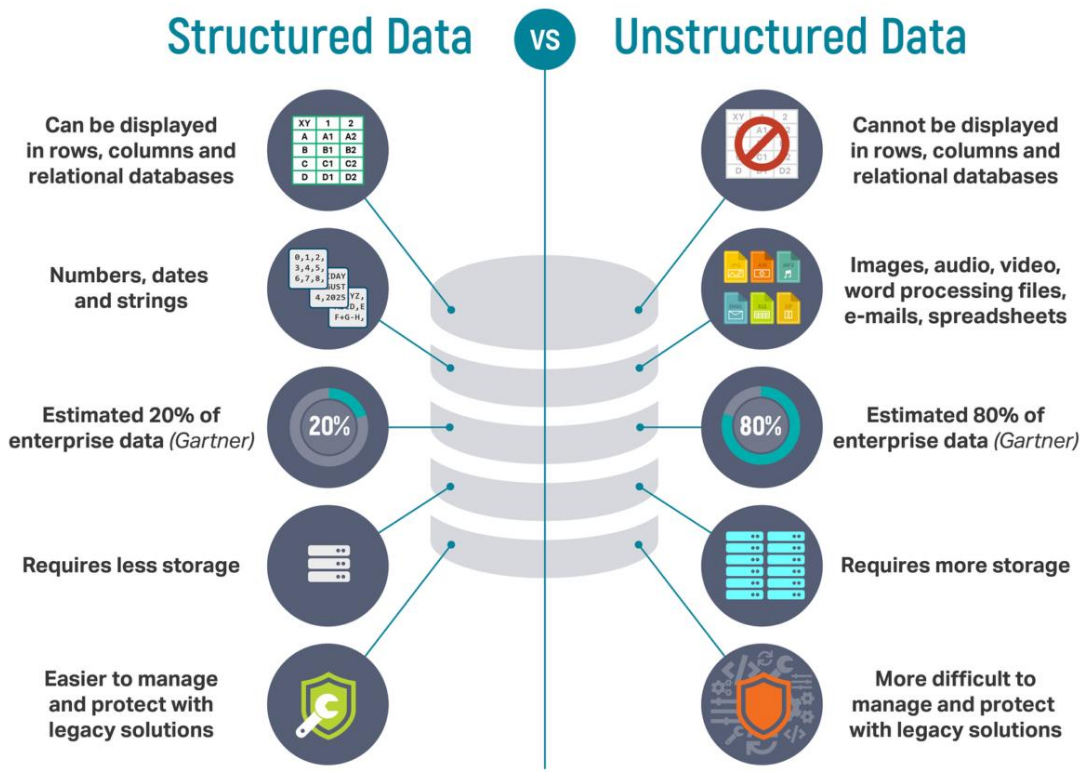


https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

How is Data Structured?

Unstructured

- Unstructured data can take any form
  - Potentially organized but inconsistent
- Not easily interpretable by machines
- The current state for >80% of data
  - Only expected to increase



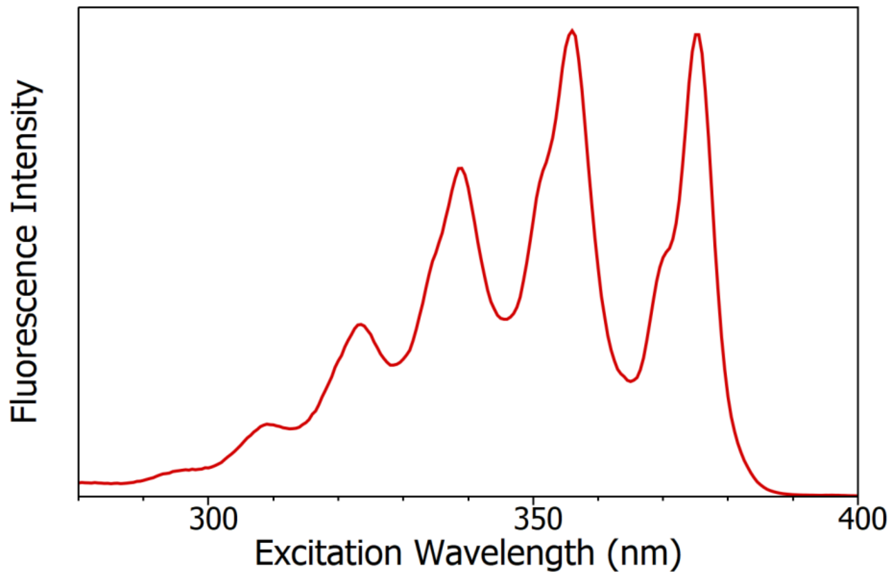
https://lawtomated.com/structured-data-vs-unstructured-data-what-are-they-and-why-care/

Common Examples of Structured Data

- Tabular Data

Feature 1	Feature 2	Feature 3
34	1.004	AAB
42	4.293	BTY
142	7.934	XYZ
23	4.143	PWX
98	0.391	HRQ
738	3.240	TMG
423	6.996	KLO

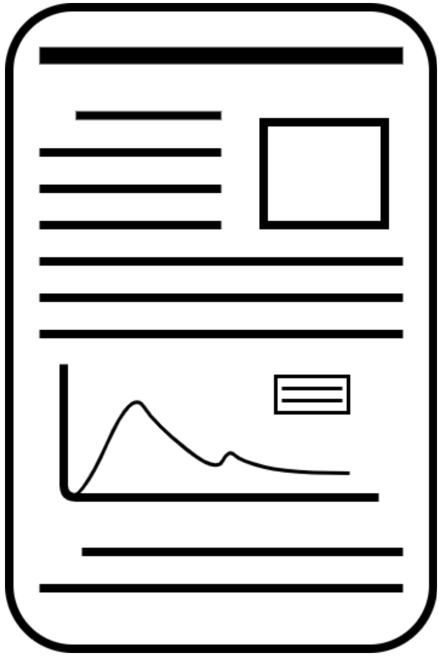
- Sensor Data



https://www.edinst.com/blog/what-are-absorption-excitation-and-emission-spectra/

Common Examples of Unstructured Data

- Reports and Documents



- Other Text / Audio / Video / Images

https://techblogwriter.co.uk/wp-content/uploads/2016/01/text-video-audio-and-images..png

How Machines Interpret Data

Computer Processors

- All data is interpreted as binary
- All operations are binary logic
- Requires consistent format

Structured vs. Unstructured Data

- Representation similar for both
  - Comprised of primitive data types
- Operations are very different
  - Structured data
    - Consistent and predictable results
  - Unstructured data
    - Inconsistent and often undefined results

Preparing Data for Machine Learning

- Raw data must be converted into a structured form
  - Most machine learning algorithms ingest arrays
  - This step is generally different for all unstructured data

- Missing values and inconsistencies must be addressed
  - Many methods for this
    - Dropping missing rows
    - Imputation or interpolation

- Scaling and other preprocessing techniques must be applied

Working with Missing Data

1 - Imputation

- Replace missing data
- Many methods
  - Mode/mean/median
  - Zero-fill
  - Random-fill
  - Interpolation
- Requires assumptions
  - Can introduce bias

Feature 1	Feature 2	Feature 3
34	1.004	AAB
	4.293	BTY
142	7.934	
23	4.143	PWX
		HRQ
738	3.240	TMG
423	6.996	KLO

→

Feature 1	Feature 2	Feature 3
34	1.004	AAB
0	4.293	BTY
142	7.934	0
23	4.143	PWX
0	0	HRQ
738	3.240	TMG
423	6.996	KLO

Working with Missing Data

2 - Interpolation

- Predict values
  - Within known range
    - Else extrapolation
- Multiple methods
  - Linear methods
  - Nonlinear methods
- Requires assumptions
  - Can introduce bias

Feature 1	Feature 2	Feature 3
34	1.004	AAB
	4.293	BTY
142	7.934	
23	4.143	PWX
		HRQ
738	3.240	TMG
423	6.996	KLO

→

Feature 1	Feature 2	Feature 3
34	1.004	AAB
88	4.293	BTY
142	7.934	???
23	4.143	PWX
357	3.624	HRQ
738	3.240	TMG
423	6.996	KLO

Working with Missing Data

3 - Removal

- Drop incomplete rows
- Will result in less data
  - But higher quality data
- No assumptions
  - Will not introduce bias

Feature 1	Feature 2	Feature 3
34	1.004	AAB
	4.293	BTY
142	7.934	
23	4.143	PWX
		HRQ
738	3.240	TMG
423	6.996	KLO

→

Feature 1	Feature 2	Feature 3
34	1.004	AAB
23	4.143	PWX
738	3.240	TMG
423	6.996	KLO

Scaling Input Features

- Scaling data is very important for certain models
  - Maintaining consistent data range equalizes all features

- Multiple available methods with various applications
  - Min-Max scaling
    - Rescaling all features to the same closed interval [min, max]
  - Standard scaling (z-score normalization)
    - Rescaling to zero-mean and unit-variance
  - Quantile scaling
    - Rescaling all features to match a target distribution

Presentation Flow

Data Science Basics

- What is data science?
- Applications of data science
- Statistics and analytics vs machine learning

Data Preparation

- Working with different types of data
- How data is structured
- How machines interpret data

Machine Learning

- Core principles of machine learning
- Machine learning project workflow
- Different machine learning tasks

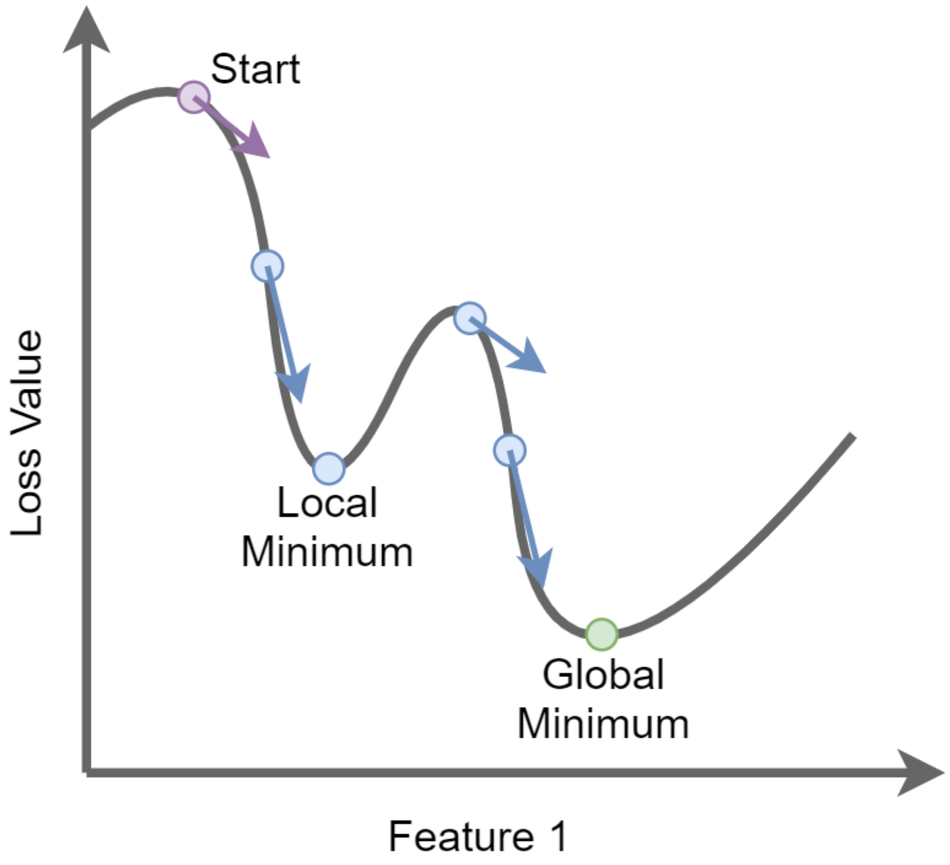
Machine Learning

Problem solving by inference

Core Principles of Machine Learning

- Minimizing a “loss” function
  - Mean squared error
  - Root mean squared error
  - Mean absolute error
  - Cross-entropy (log loss)

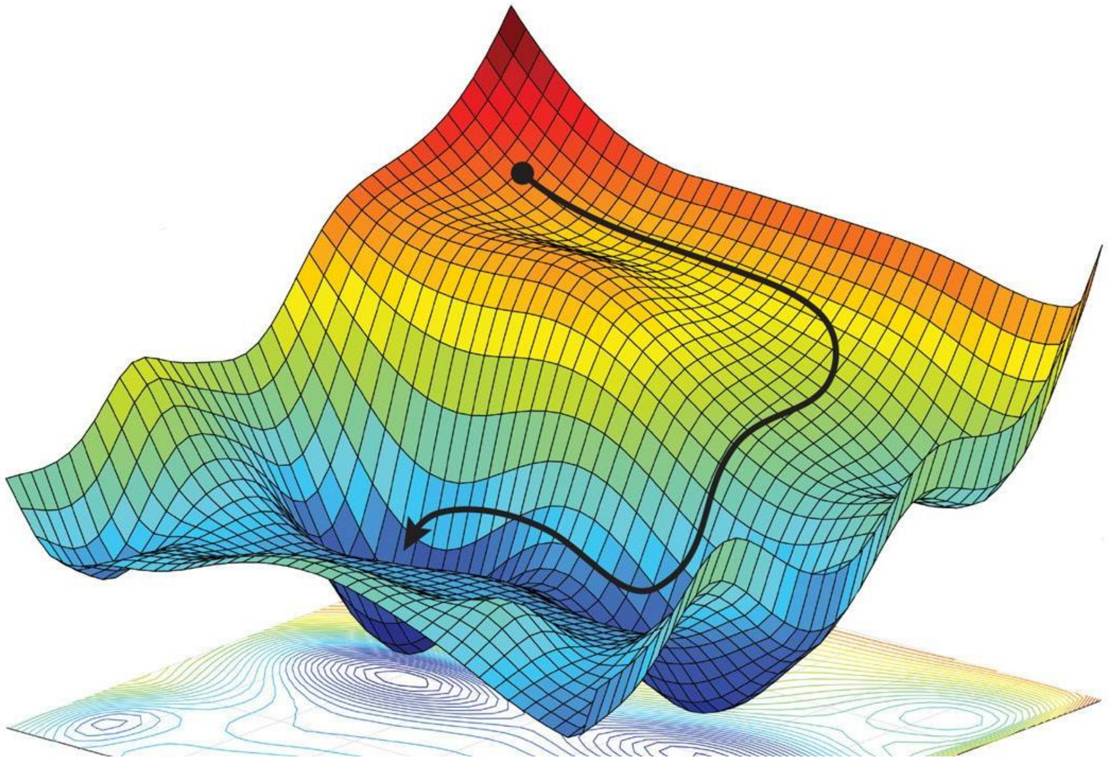
- Optimized via gradient descent
  - Analogy: rolling down a hill



Core Principles of Machine Learning

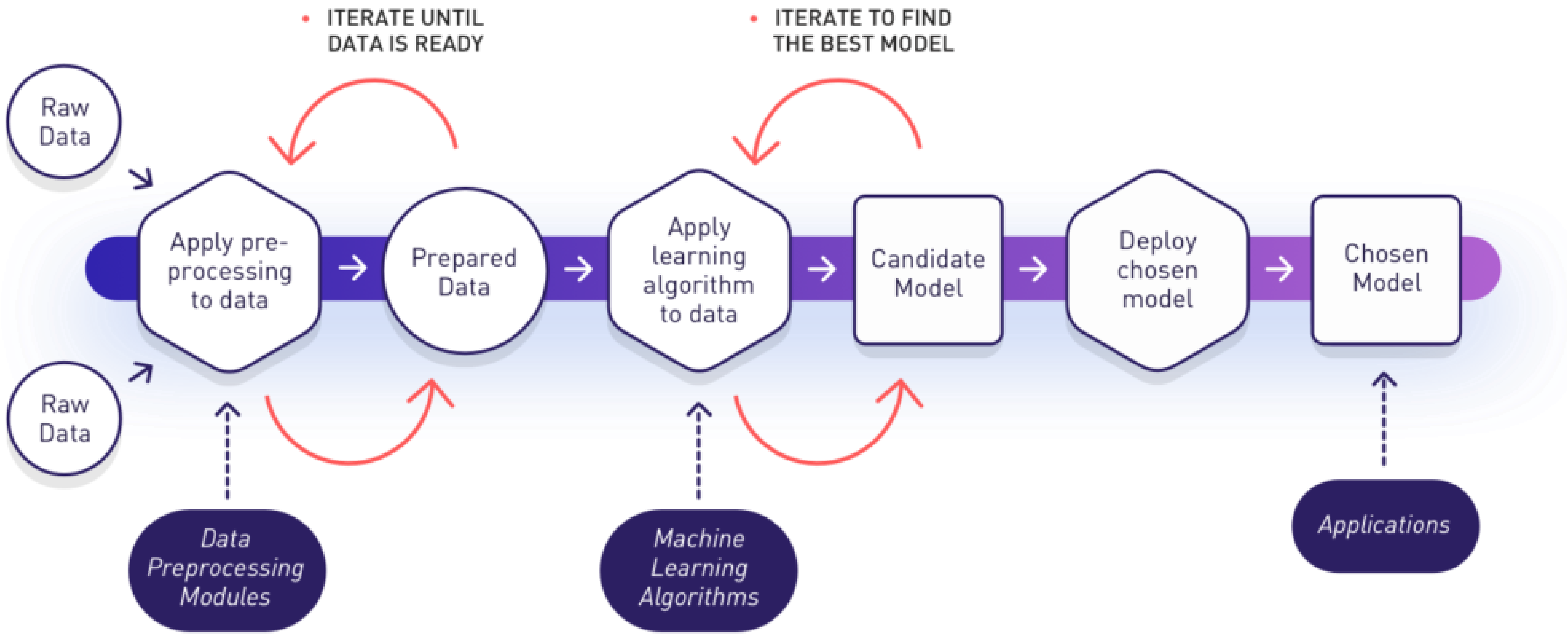
- Minimizing a “loss” function
  - Mean squared error
  - Root mean squared error
  - Mean absolute error
  - Cross-entropy (log loss)

- Optimized via gradient descent
  - Analogy: rolling down a hill



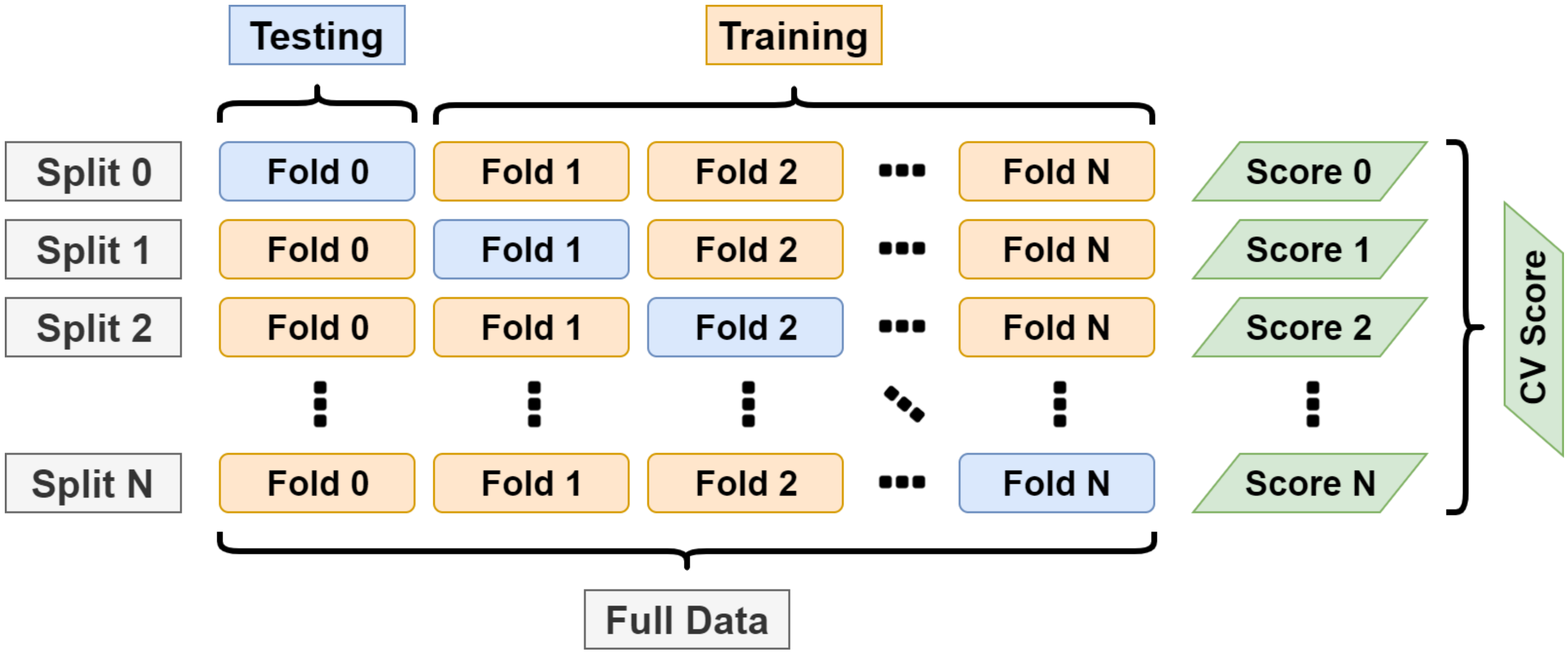
http://bioinformatics.org.au/ws18/wp-content/uploads//sites/22/2016/02/Marcus-Gallgher_2018-Winter-School.pdf

Typical ML Project Workflow



https://www.uruit.com/blog/wp-content/uploads/2018/02/Diagram-1-1024x435.png

Cross Validation



Interactive Notebook

The interactive notebook for this workshop is available on GitHub

https://github.com/PCiunkiewicz/I4H_Workshop

Presentation slides will also be there

2019 HEALTH HACK COMPETITION

INNOVATION4 HEALTH

Theme

CHILD HEALTH & WELLNESS

Data Science Workshop