

DATA SCIENCE :::::
DEVELOPMENT SERIES



MACHINE LEARNING 101

Wednesday, 5/27

by **JINGJING CANNON**
Data Scientist & Six Sigma Green Belt

► REGISTER TODAY

brought to you by:



DATA SCIENCE DEVELOPMENT SERIES

brought to you by:



Session 1: MACHINE LEARNING 101
by Jingjing Cannon, 5/27

Session 2: COMPUTER VISION 101
by Matthew Hagen, 6/3

Session 3: DEEP LEARNING 101
by Nikolaos Vasiloglou, 6/10

Session 4: NLP 101
by Brian O'Connor, 6/17

Session 5: SEARCH & RELEVANCY 101
by Trey Grainger, 6/24

Session 6: RECOMMENDATIONS 101
by Khalifeh Al Jadda, 7/1

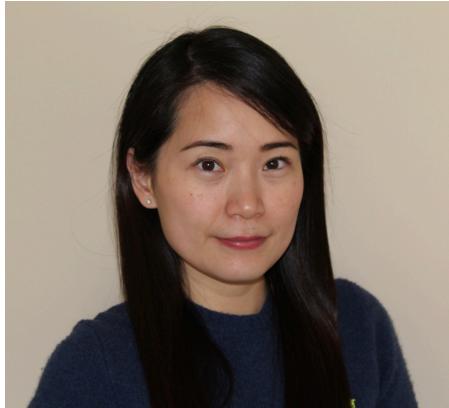
For more information, please follow:



Southern Data Science Conference

Information Technology & Services · Atlanta, GA

About the presenter



- PHD in Computational Neuroscience at Georgia State University
- Previous working experience:
 - Credit risk analyst in finance.
 - Jr. Data Scientist in Telecom.
 - Data Scientist at The Home Depot Online Data Science Team.

LinkedIn : <https://www.linkedin.com/in/jingjing-cannon-23a56524/>

Agenda

https://github.com/jingjingsky120/MachineLearning_101

- **Part I Basics**
 - General introduction
 - Data pre-processing methods
 - Machine Learning model development and validation
 - Basic Data visualization
-
- **Part II Supervised Machine Learning**
 - Algorithms introduction: Regression, Tree based, Boosting and Artificial Neural Networks.
 - A Case study of a RecSys Conference Challenge to compare above algorithms.
-
- **Part III Unsupervised Machine Learning**
 - Algorithms introduction: K-means and hierarchical clustering
 - High-dimensional data visualization examples in Python

Part I

Machine Learning Basics

Who is Data Scientist?

MODERN DATA SCIENTIST



MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

New to Data Science field

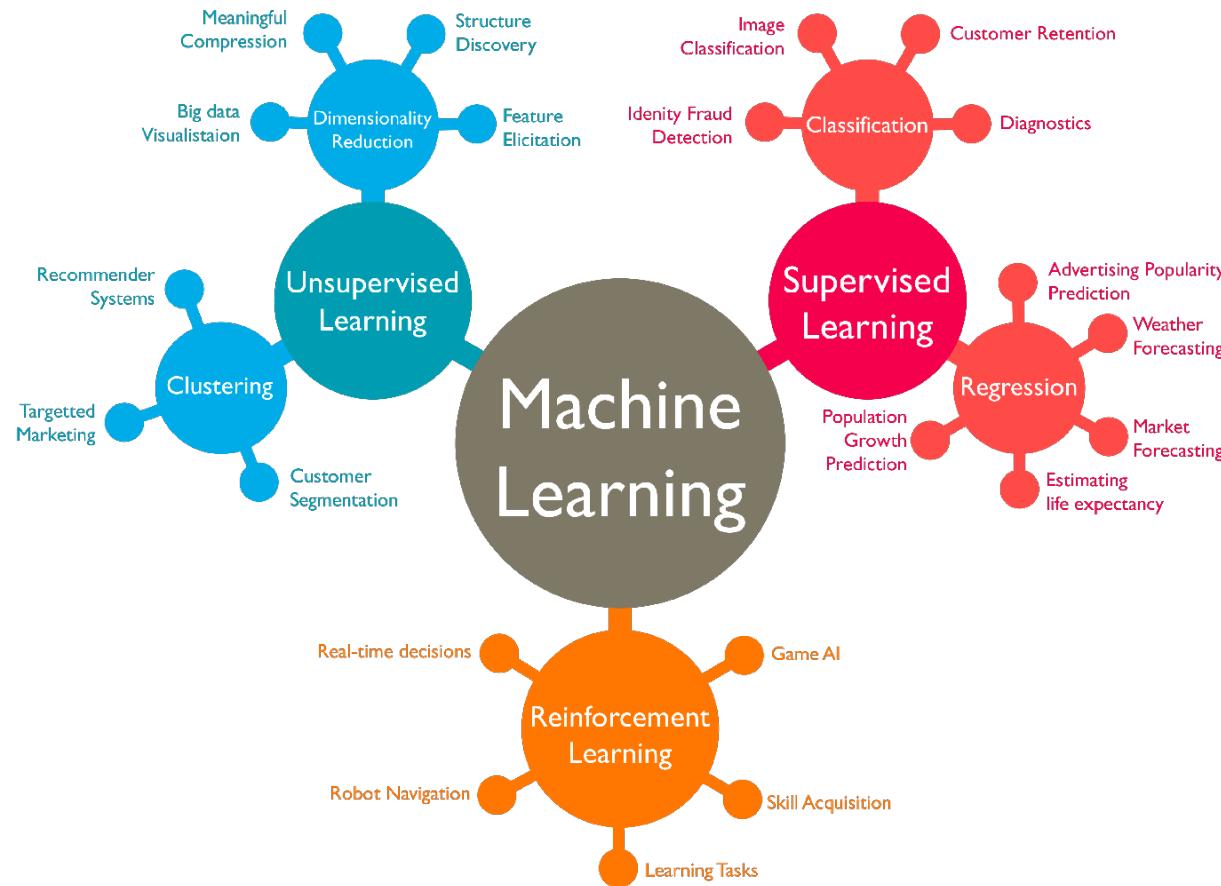
- **Skill sets required for entry level analytics jobs:**

- Query Language (Microsoft SQL, Teradata, Hadoop, Cloud Computing etc.)
- Data Analysis and Modeling (Python, R, SAS, etc)
- Visualization (matplotlib, ggplot, Tableau, etc)

- **Data Science mindset**

- Data Collection
 - Data query from structured database
 - Online Data Scraping etc
- Data Pre-Processing
 - Data Cleansing
 - Exploratory Data Analysis (Missing Data, Data Type, distribution, etc)
 - Feature Engineering or Feature Selection
- Data Modeling
 - Statistical Algorithms (Regression, Tree, Neuronal Network, K-means etc)
- Result Visualization

What is Machine Learning



Data Science Toolkit



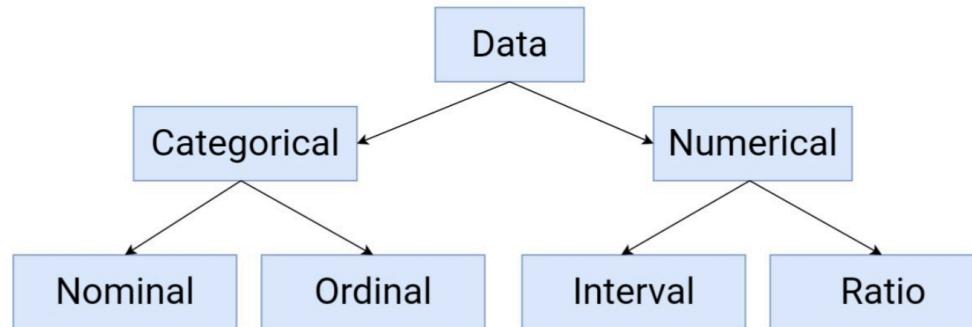
<https://www.anaconda.com/products/individual>



https://jupyterlab.readthedocs.io/en/stable/getting_started/installation.html

Data Science Basics

Structured Data Types



Data Pre-processing

- Data Quality Assessment
- Feature Aggregation
- Feature Sampling
- Dimensionality Reduction
- Feature Encoding

<https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825>

Data Science Basics

Feature preparation <https://innovation.alteryx.com/feature-engineering-vs-feature-selection/>

- Feature selection <https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2>
Correlation, Feature Importance, VIF, Logistic regression or Tree based model
- Feature engineering



RecSys Challenge 2015

WE RECOMMEND YOOCHOOSE SE

HOME

CHALLENGE

PRIZES

SUBMISSION

PAPER SUBMISSION

RULES

ORGANIZERS

FORUM

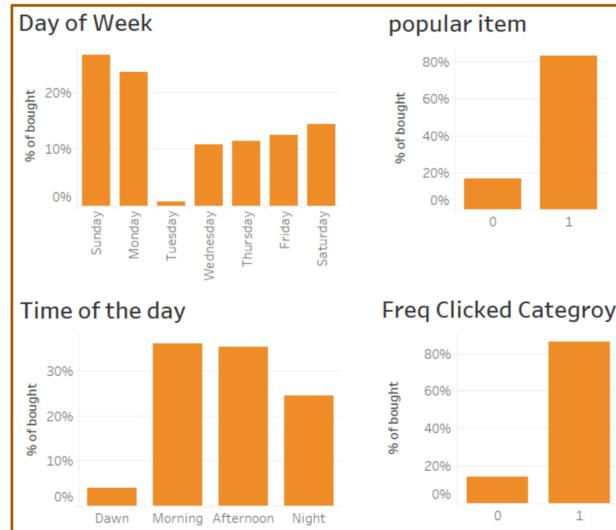


<https://2015.recsyschallenge.com/challenge.html>

Data Science Basics

Feature Engineering example:

Session ID
Timestamp
Item ID
Category
Price
Quantity



Model Validation

Confusion Matrix

		Actual class	
		P	N
Predicted class	P	TP	FP
	N	FN	TN

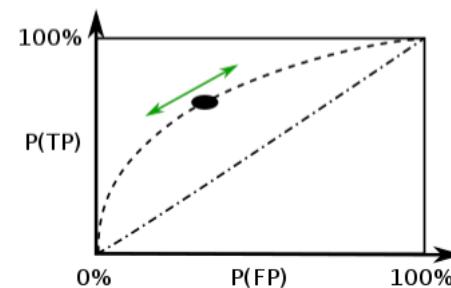
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

ROC and AUC

- Receiver Operating Characteristic
- Area Under the Curve



https://en.wikipedia.org/wiki/Confusion_matrix

Data Visualization

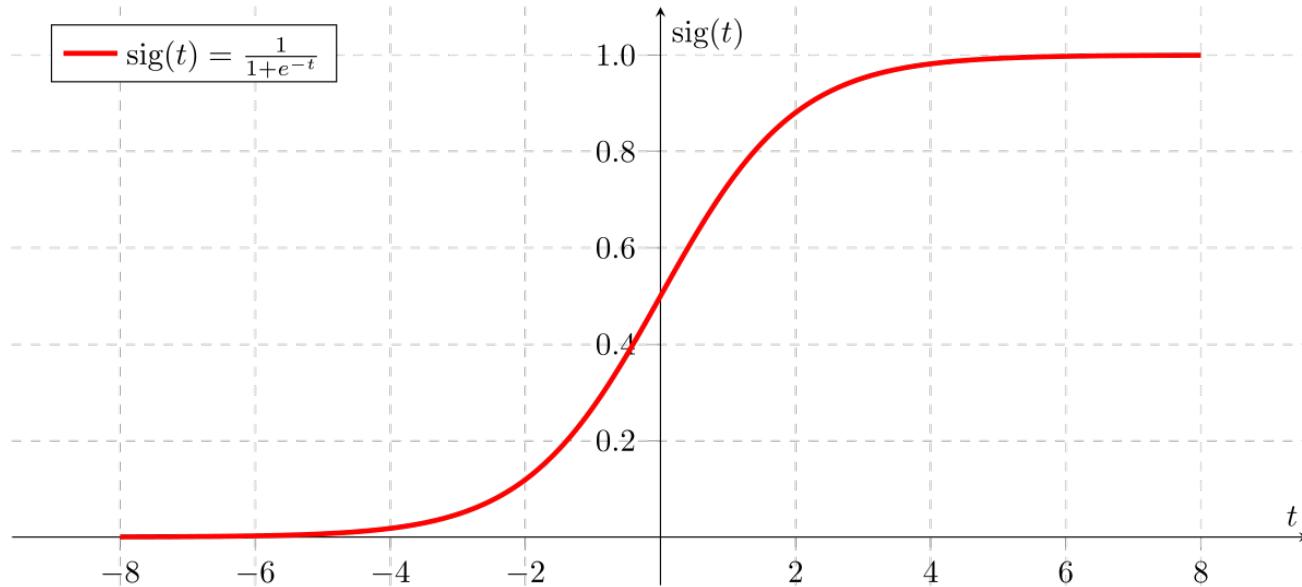
- Python library Seaborn
 - Scatter plot
 - Line chart
 - Histogram
 - Bar Chart
 - Box plots
 - Heatmap
- T-SNE for high dimensional data

Part II

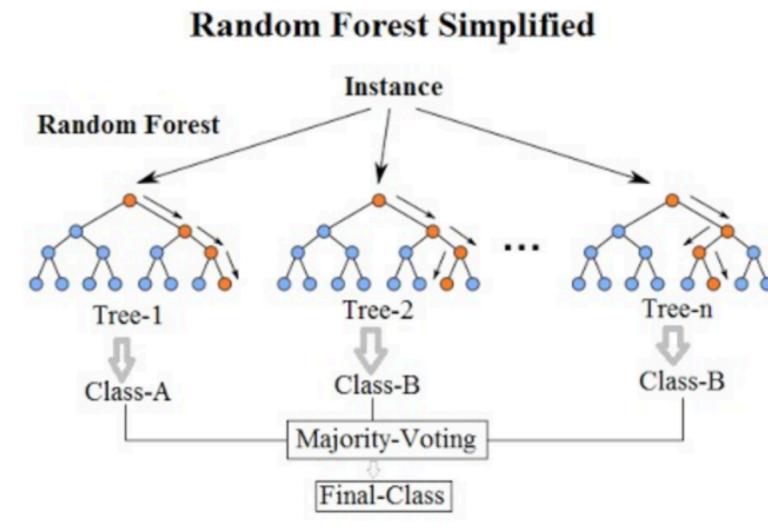
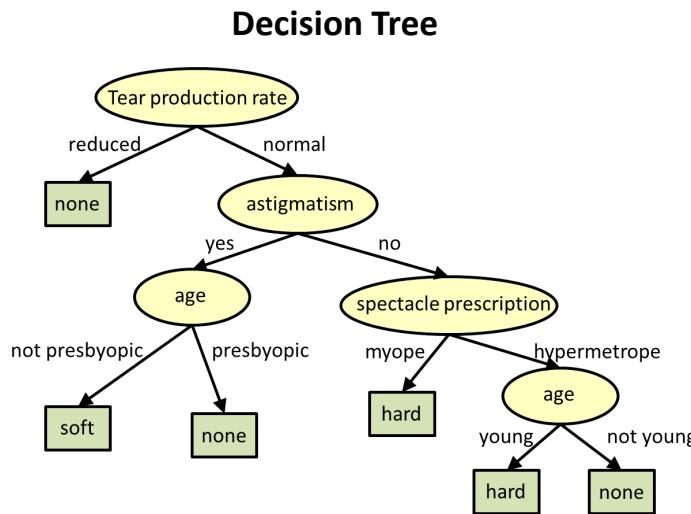
Supervised Machine Learning

Logistic Regression

Relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables

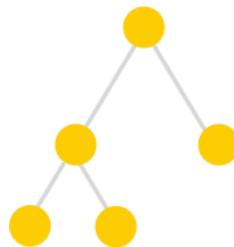


Decision Tree and Random Forest

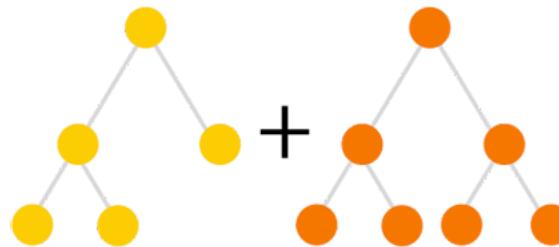




Yandex
CatBoost



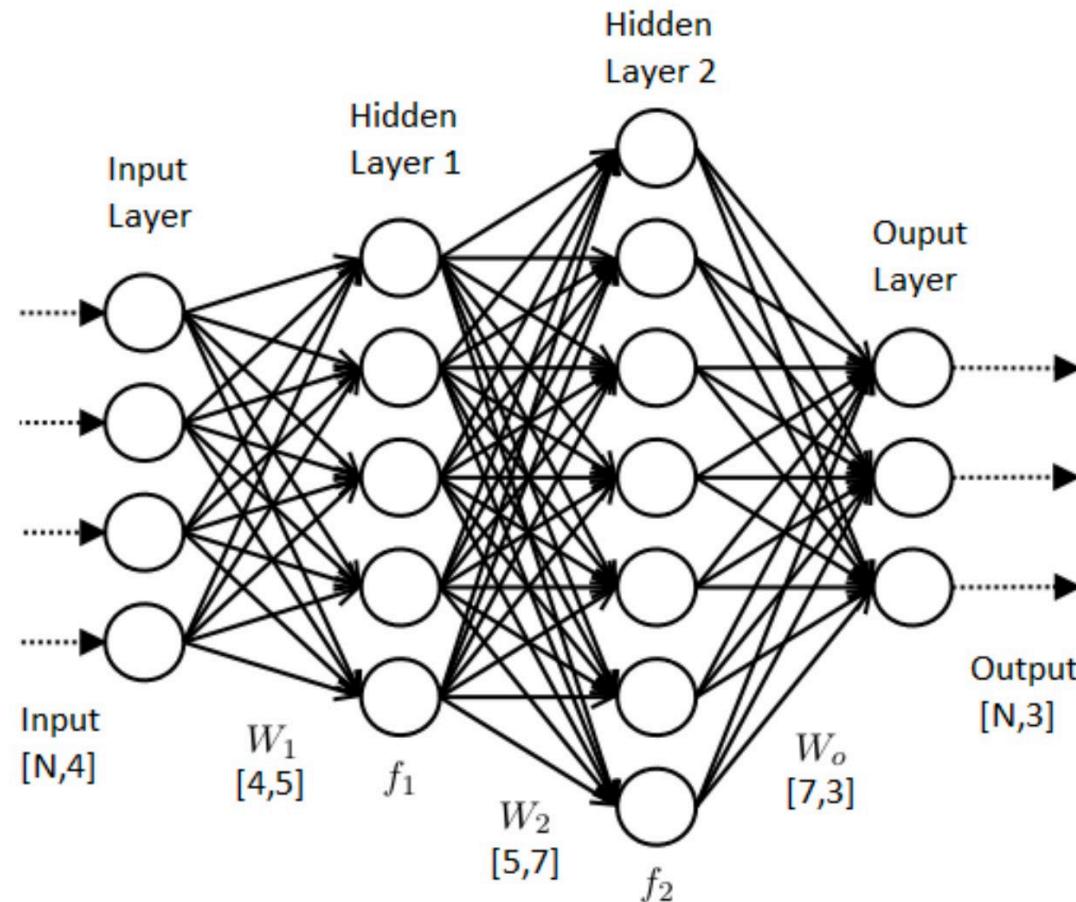
First Tree



Second Tree

<https://catboost.ai/news/catboost-enables-fast-gradient-boosting-on-decision-trees-using-gpus>

Artificial Neural Network



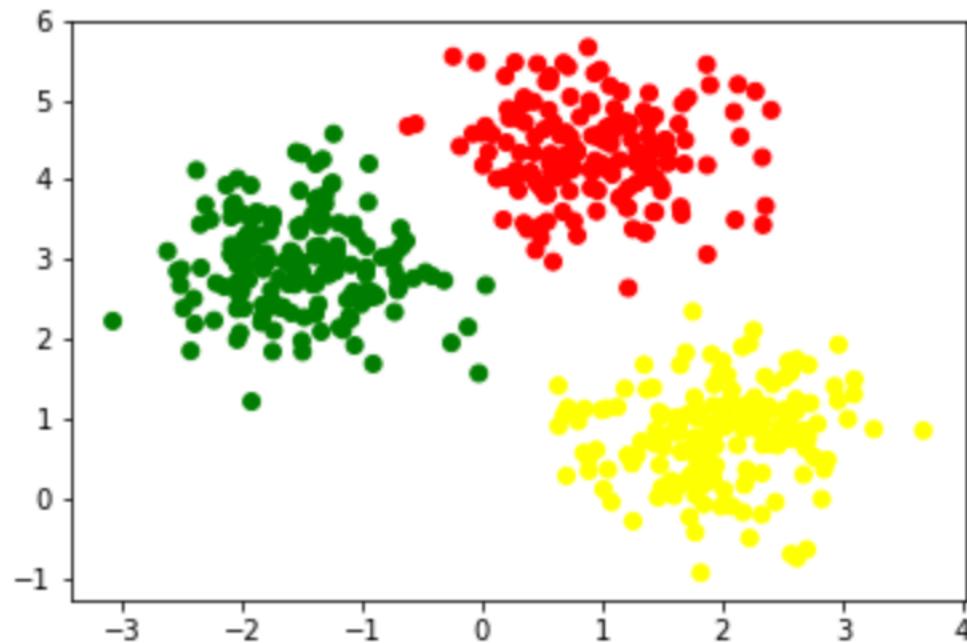
Algorithm Comparison

Algorithm	ROC AUC Score	
	Train	Validation
MatrixNet(Regressor)	74.62%	75.56%
MatrixNet(Classifier)	69.13%	69.48%
Random Forest	67.20%	67.51%
Decision Tree	99.78%	65.43%
Logistic Regression	61.18%	62.84%
Neural Network	54.08%	50%

Part III

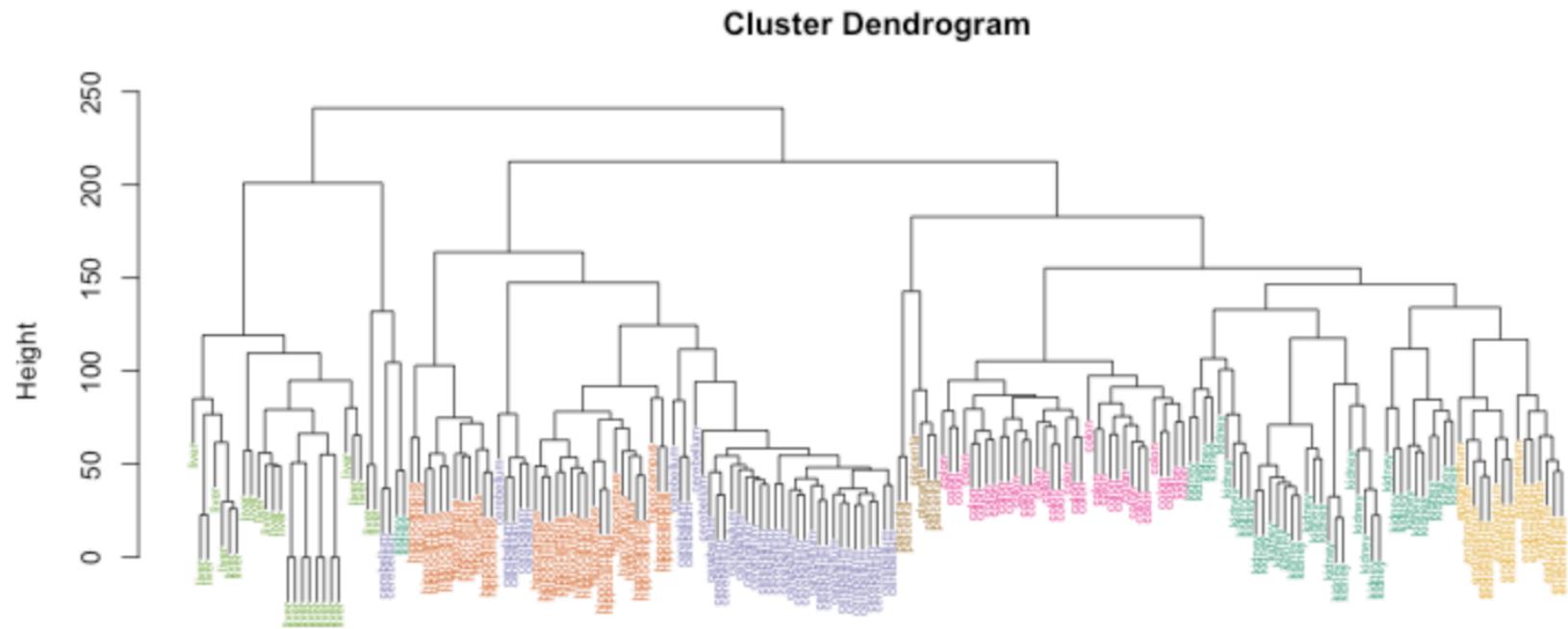
Unsupervised Machine Learning

K-means Clustering



<https://cmdlinetips.com/2019/05/k-means-clustering-in-python/>

Hierarchical Clustering



Thank you !