Python for Data Science

Introduction to Machine Learning

Dr. Ilkay Altintas and Dr. Leo Porter

Twitter: #UCSDpython4DS

Python for Data Science

By the end of this video, you should be able to:

- Explain what machine learning is
- List three applications of machine learning encountered in everyday life

Machine Learning is…

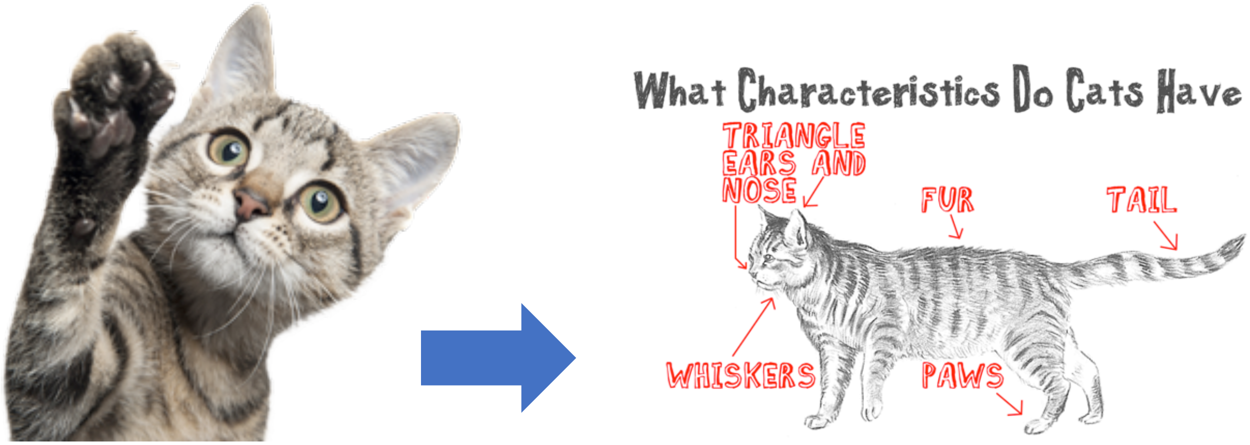
… learning from data

Python for Data Science

Machine Learning is…

… learning from data

… on its own



Python for Data Science

Machine Learning is…

… learning from data

… on its own

… discovering hidden patterns

Machine Learning is…

… learning from data

… on its own

… discovering hidden patterns

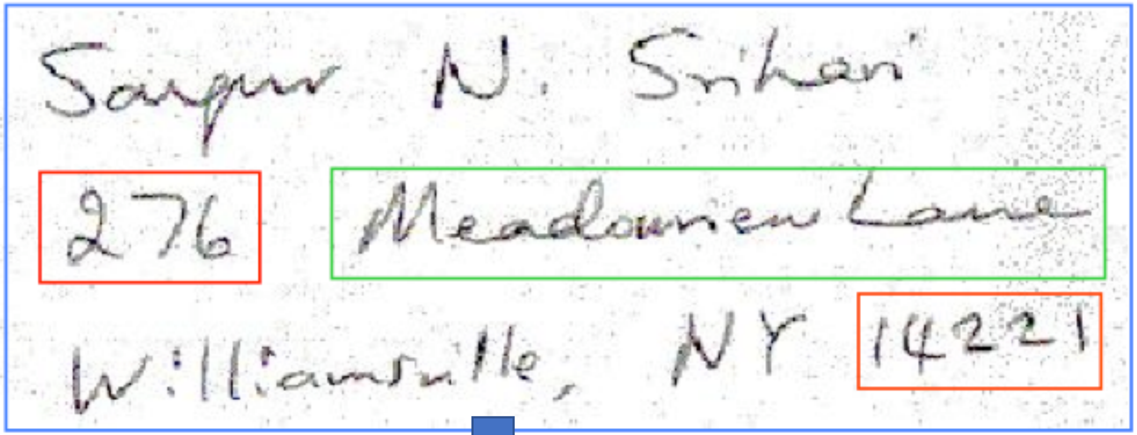
… data-driven decisions

Applications of Machine Learning

Python for Data Science

Credit Card Fraud Detection

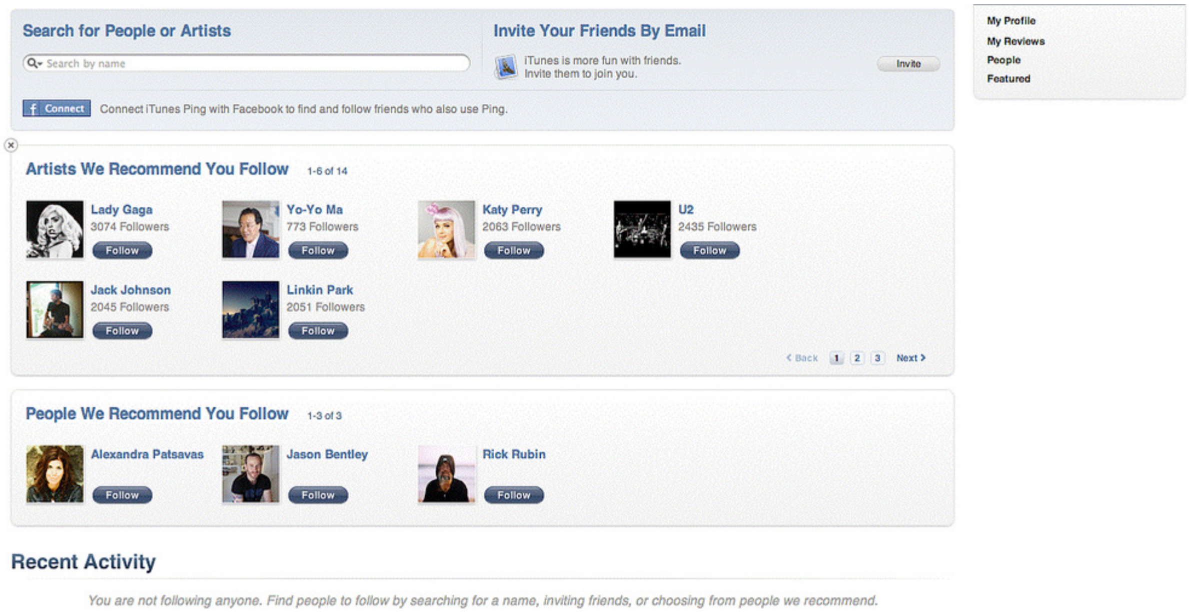
Handwritten Digit Recognition



Sargur N. Srihari
276 Meadowview Lane
Williamsville, NY 14221

ZIP Code: 14221
Primary number: 276

Recommendations on Websites



Search for People or Artists

Search by name

Invite Your Friends By Email

iTunes is more fun with friends.
Invite them to join you.

Connect iTunes Ping with Facebook to find and follow friends who also use Ping.

My Profile
My Reviews
People
Featured

Artists We Recommend You Follow 1-6 of 14

Lady Gaga 3074 Followers Follow	Yo-Yo Ma 773 Followers Follow	Katy Perry 2063 Followers Follow
U2 2435 Followers Follow	Jack Johnson 2045 Followers Follow	Linkin Park 2051 Followers Follow

People We Recommend You Follow 1-3 of 3

Alexandra Patsavas Follow	Jason Bentley Follow	Rick Rubin Follow

Recent Activity

You are not following anyone. Find people to follow by searching for a name, inviting friends, or choosing from people we recommend.

Machine Learning and Data Science

- Data mining
- Predictive analytics
- Big Data

Machine Learning Models

- Learn from data
- Discover patterns and trends
- Allow for data-driven decisions
- Used in many different applications

Python for Data Science

Categories of Machine Learning

Dr. Ilkay Altintas and Dr. Leo Porter

Twitter: #UCSDpython4DS

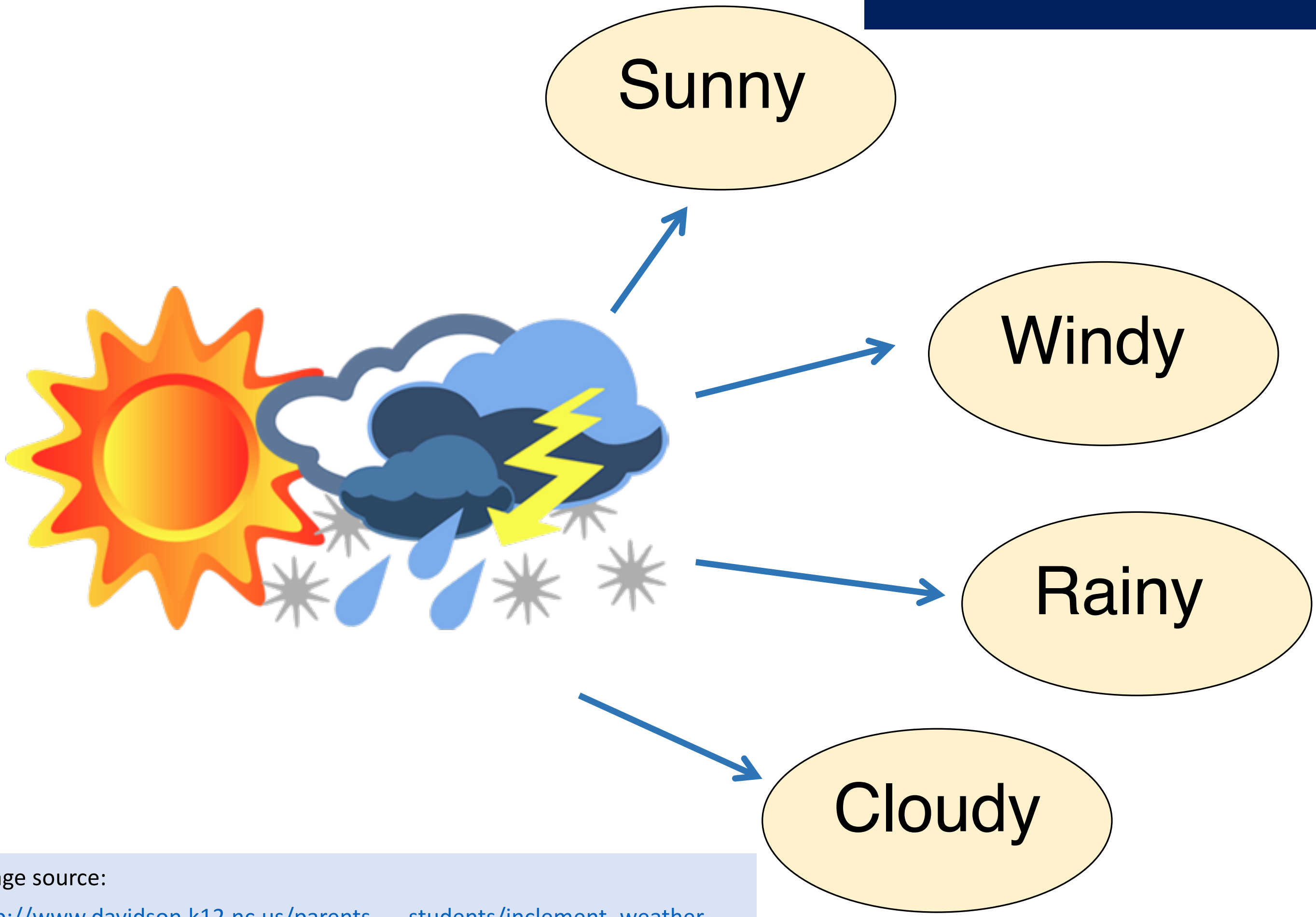
By the end of this video, you should be able to:

- Describe the main categories of machine learning techniques
- Summarize how supervised learning differs from unsupervised learning

Python for Data Science

Classification

Goal: Predict category



Sunny

Windy

Rainy

Cloudy

Image source:
http://www.davidson.k12.nc.us/parents___students/inclement_weather

Python for Data Science

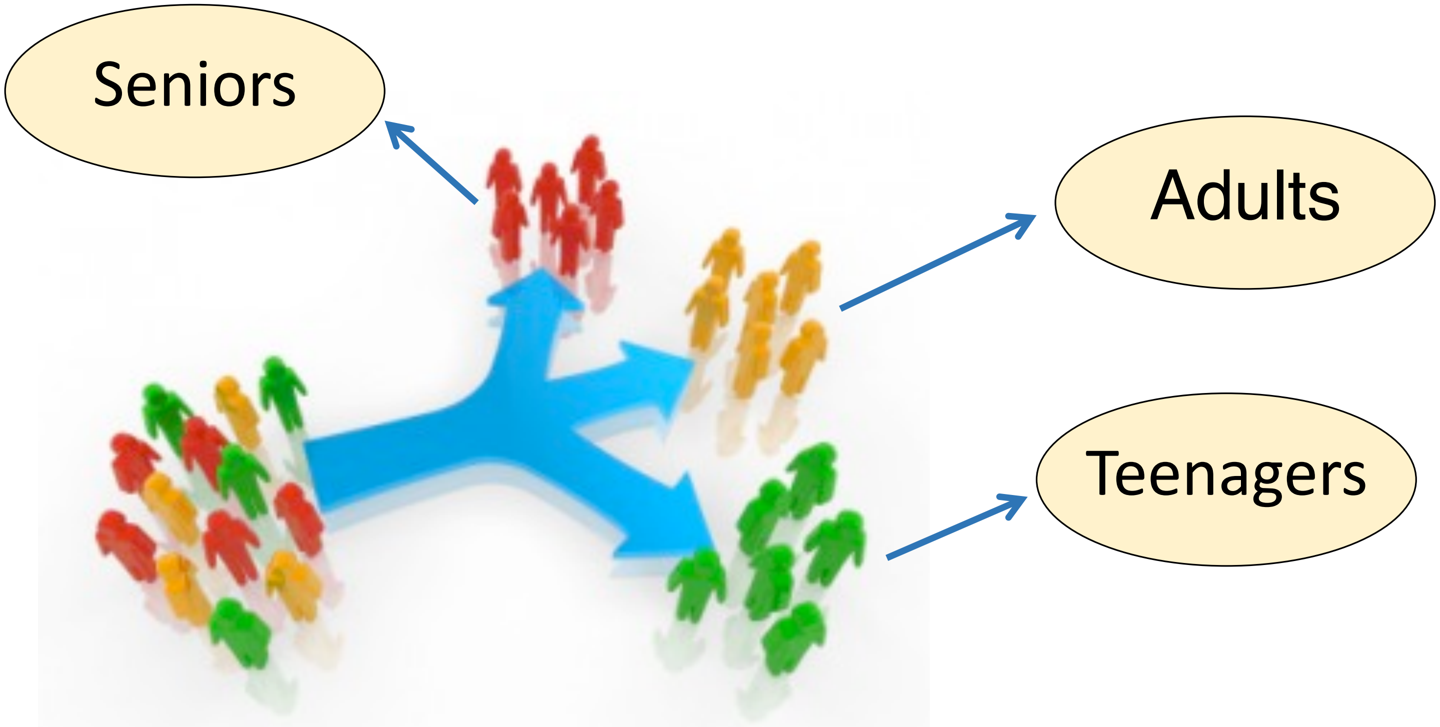
Regression

Goal: Predict numeric value

Python for Data Science

Cluster Analysis

Goal: Organize similar items into groups.



Seniors

Adults

Teenagers

Image source: http://www.monetate.com/blog/the-intrinsic-value-of-customer-segmentation

Association Analysis

Goal: Find rules to capture associations between items.

Categories of Machine Learning Techniques

- Classification
- Cluster Analysis
- Regression
- Association Analysis

Supervised vs. Unsupervised

- Supervised Approaches
  - Target (what model is predicting) is provided
  - ‘Labeled’ data
  - Classification & regression are supervised.

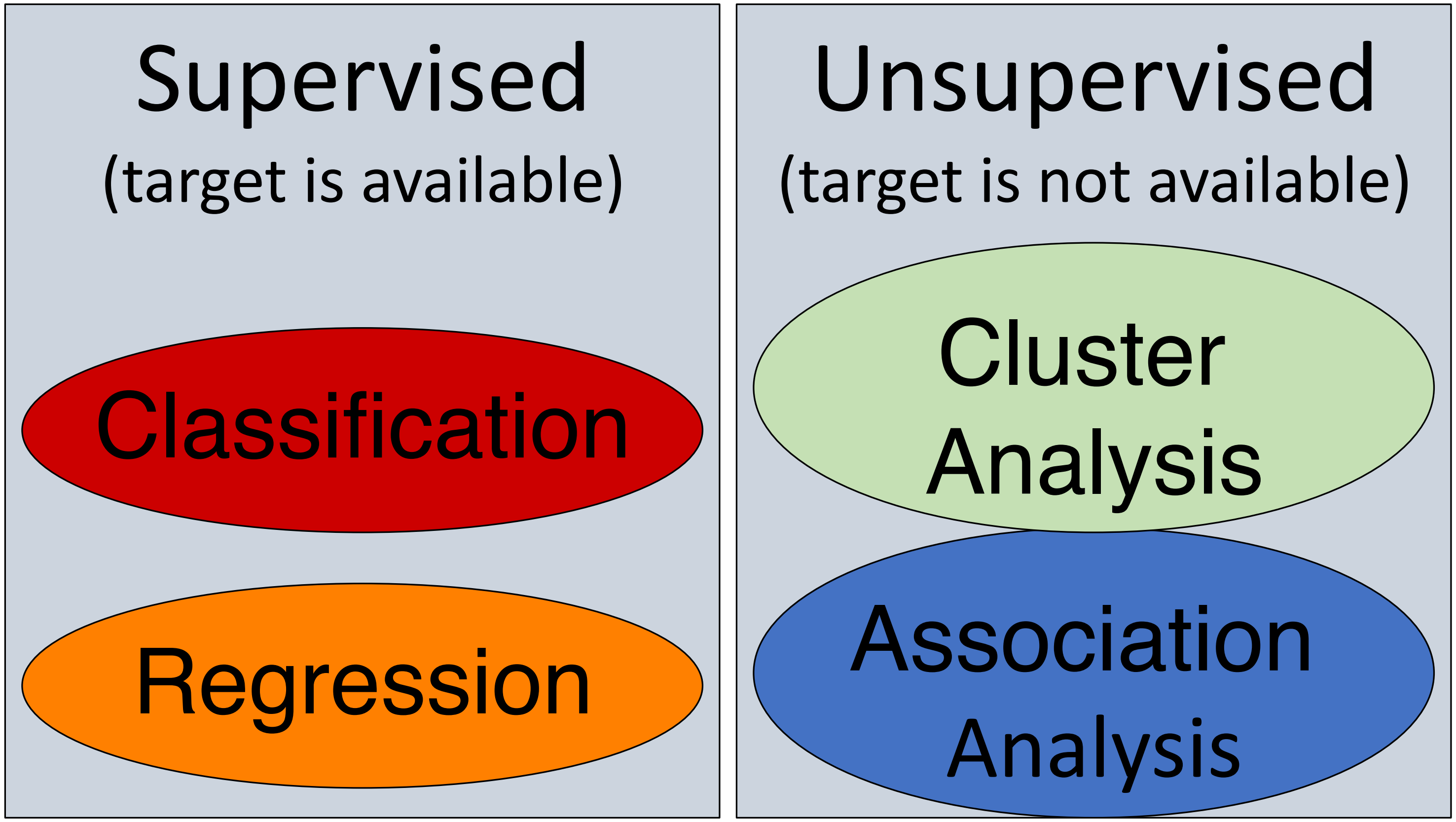
Supervised vs. Unsupervised

- Supervised Approaches
  - Target (what model is predicting) is provided
  - ‘Labeled’ data
  - Classification & regression are supervised.

- Unsupervised Approaches
  - Target is unknown or unavailable
  - ‘unlabeled’ data
  - Cluster analysis & association analysis are unsupervised.

Python for Data Science

Categories of Machine Learning Techniques



Supervised

(target is available)

Classification

Regression

Unsupervised

(target is not available)

Cluster Analysis

Association Analysis

Terminology Related to Machine Learning

Dr. Ilkay Altintas and Dr. Leo Porter

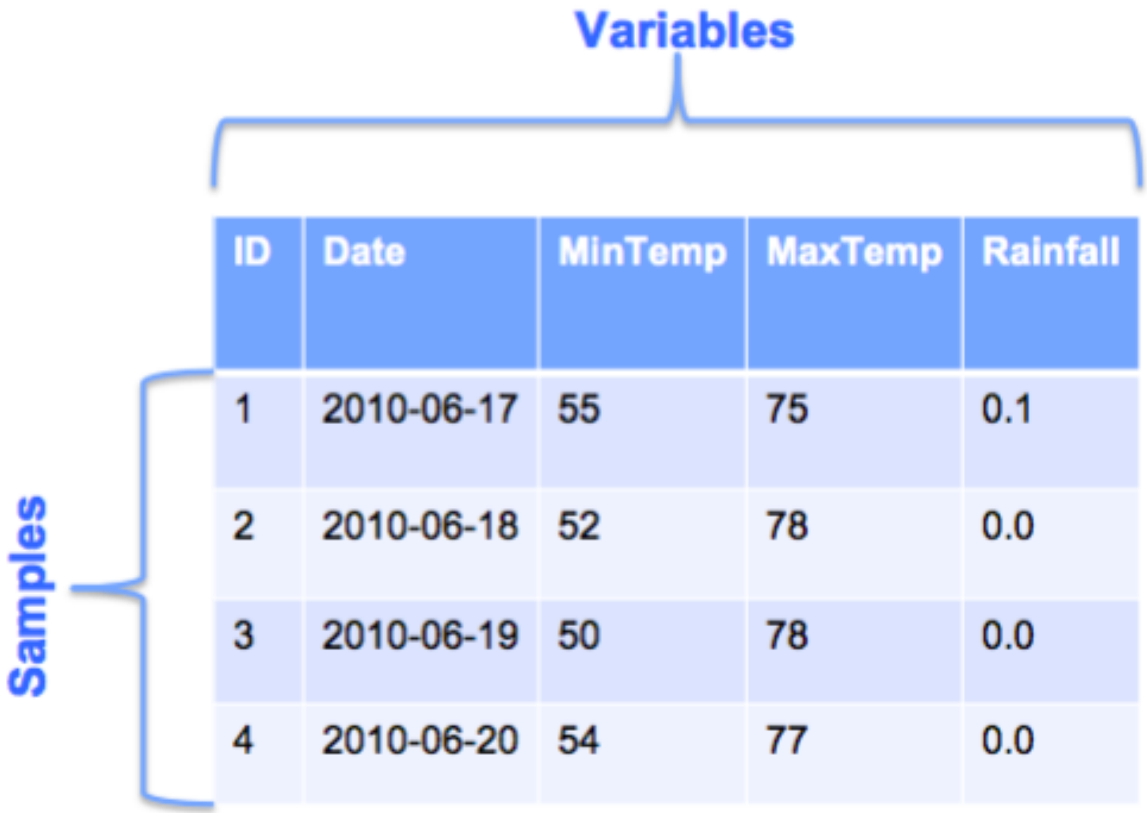
Twitter: #UCSDpython4DS

By the end of this video, you should be able to:

- Describe what a feature is and how it relates to a sample
- Name some alternative terms for ‘feature’
- Summarize how a categorical feature differs from a numerical feature

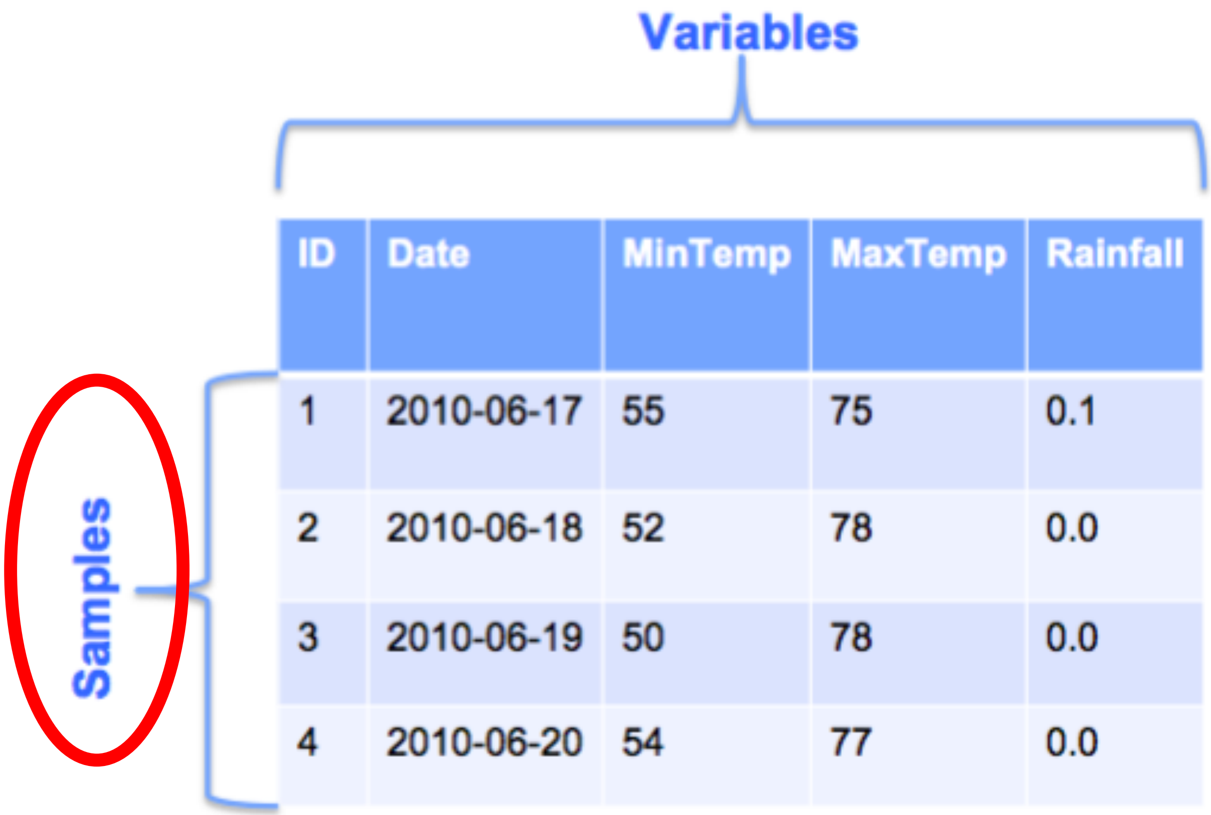
Python for Data Science

Terms to Describe Data


Variables
Samples

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

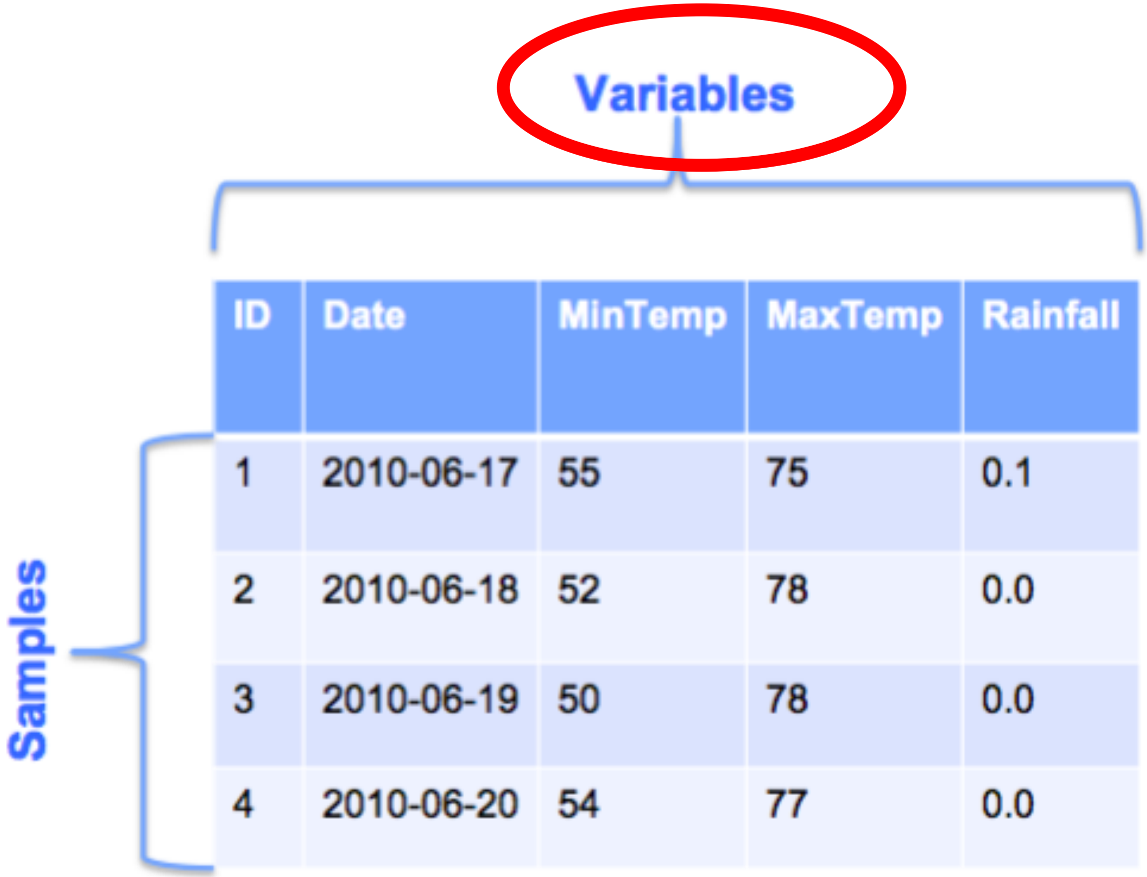
Terms to Describe Data: SAMPLE


Variables
Samples

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Python for Data Science

Terms to Describe Data: VARIABLE


Variables
Samples

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Other Names for ‘Sample’

sample

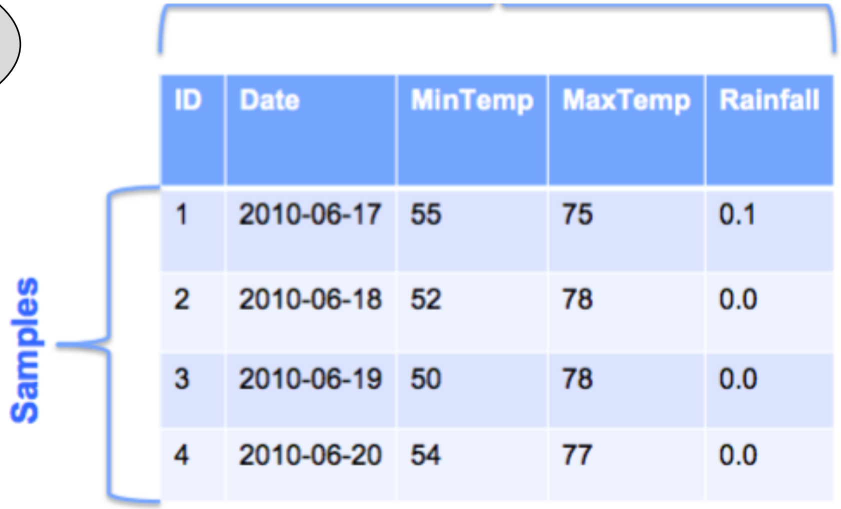
instance

row

observation

record

example



ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Other Names for ‘Variable’

feature

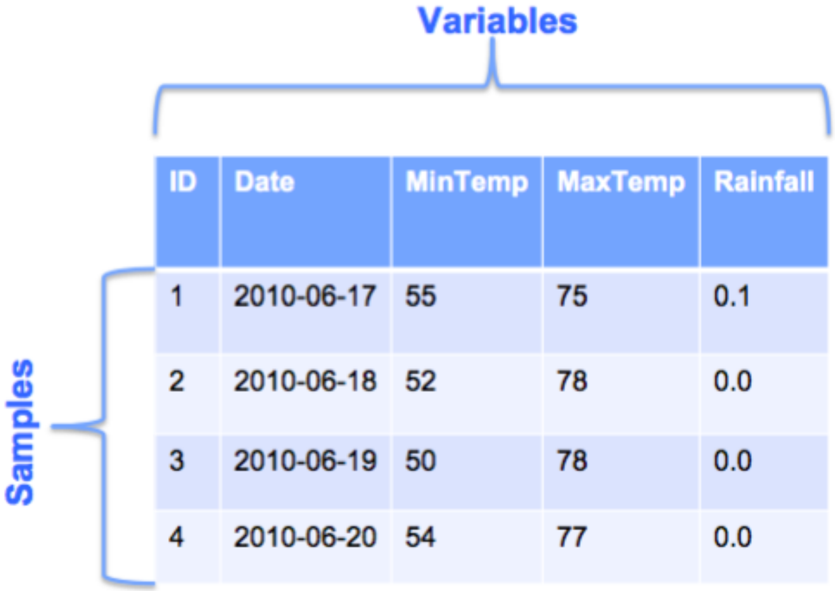
dimension

variable

column

attribute

field


Variables
Samples

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Python for Data Science

Data Types

- Most common

Numeric

Categorical

- Others

String

Date

…

Numeric Variables

- Values are numbers
- Also called ‘quantitative’

$7\text{x}10^5$

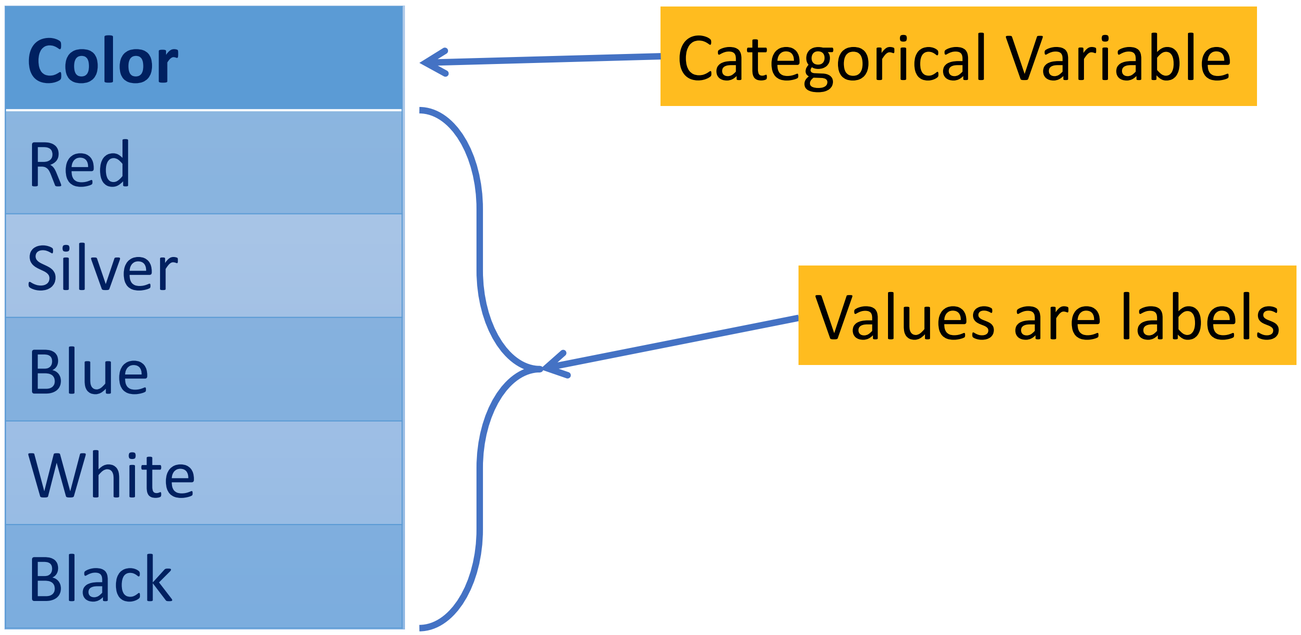
1

-0.4902

163.92

Categorical Variables

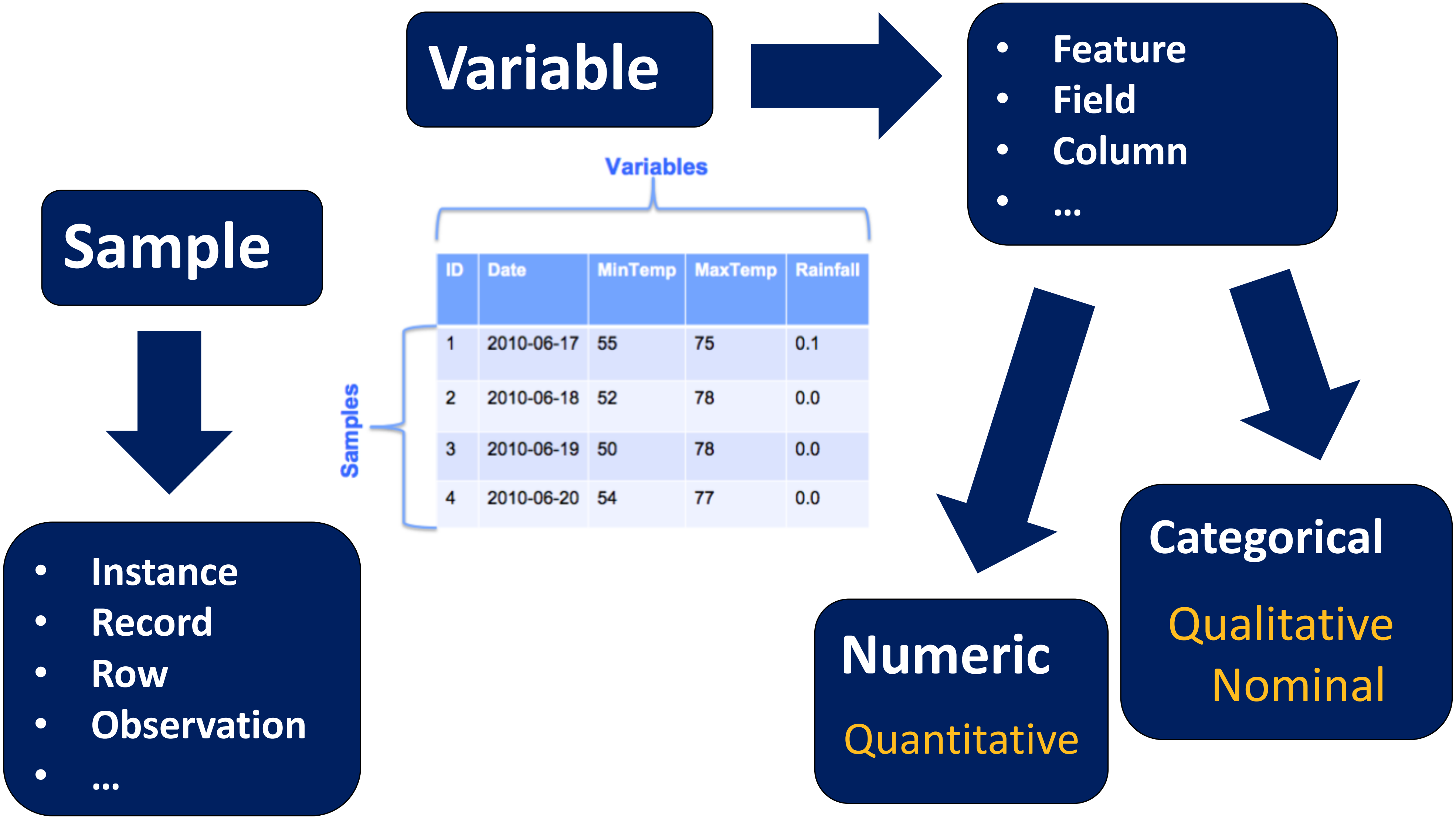
- Values are labels, names, or categories
- Also called ‘qualitative’ or ‘nominal’



Color
Red
Silver
Blue
White
Black

Categorical Variable

Values are labels



Variable

- Feature
- Field
- Column
- …

Sample

- Instance
- Record
- Row
- Observation
- …

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	55	75	0.1
2	2010-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

Numeric

Quantitative

Categorical

Qualitative

Nominal

Python for Data Science

scikit-learn

Machine Learning in Python

Dr. Ilkay Altintas and Dr. Leo Porter

Twitter: #UCSDpython4DS

By the end of this video, you should be able to:

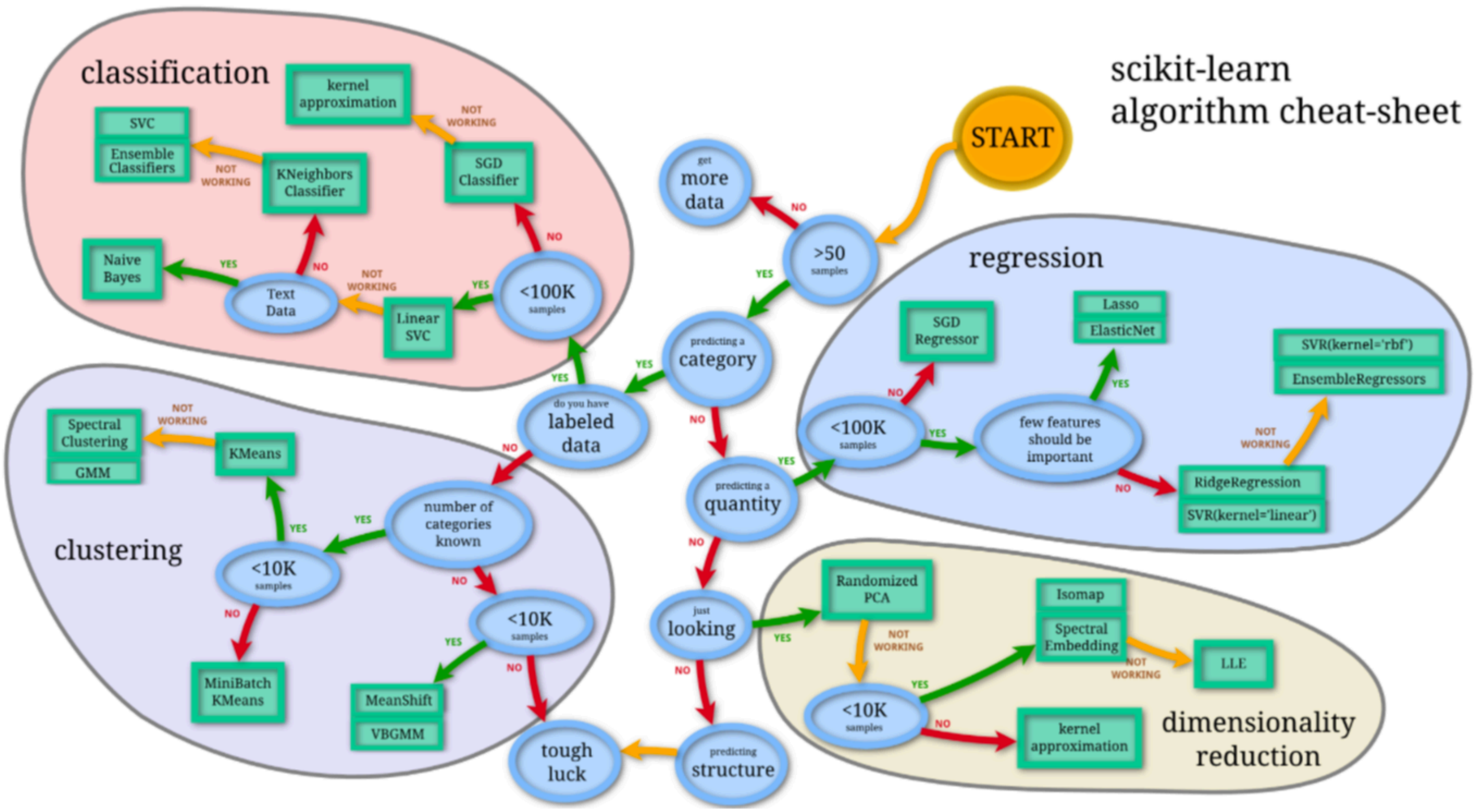
- Identify key strengths of scikit-learn
- Explain why it is a leading library for Machine Learning
- Navigate your way to find the right tool in scikit-learn
- Search for tutorials that provide problem specific examples using library functions

scikit-learn

- Open source library for Machine Learning in Python
- Built on top of NumPy, SciPy, matplotlib
- Active community for development
- Improved continuously by developers

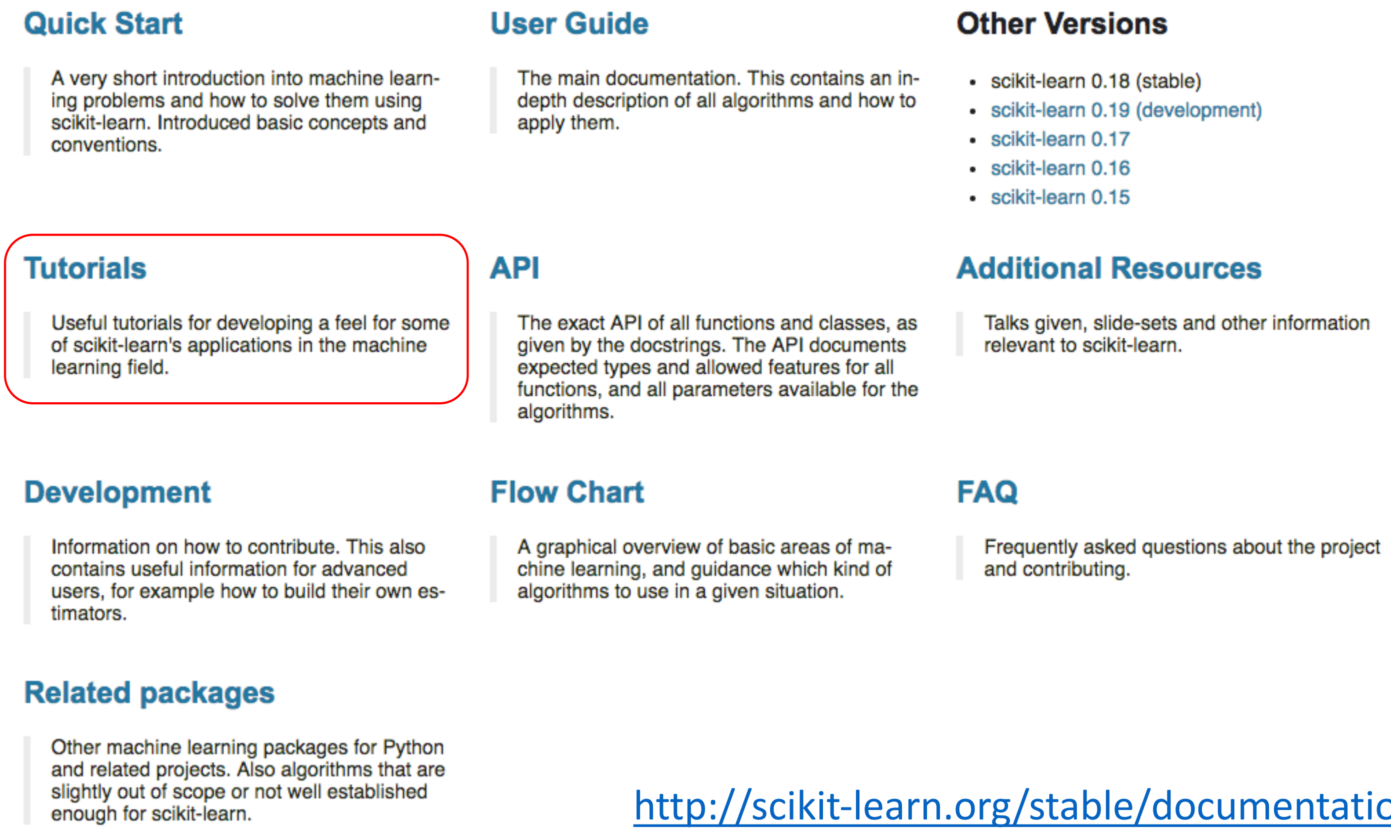
Preprocessing Tools

- Utility Functions for
  - Transforming raw feature vectors to suitable format
- Provides API for
  - Scaling of features: remove mean and keep unit variance
  - Normalization to have unit norm
  - Binarization to turn data into 0 or 1 format
  - One Hot Encoding for categorical features
  - Handling of missing values
  - Generating higher order features
  - Build custom transformations


scikit-learn
algorithm cheat-sheet
START
>50 samples
NO
get more data
YES
predicting a category
YES
do you have labeled data
YES
<100K samples
YES
Linear SVC
NOT WORKING
Text Data
YES
Naive Bayes
NO
KNeighbors Classifier
NOT WORKING
SVC
Ensemble Classifiers
NO
SGD Classifier
NOT WORKING
kernel approximation
classification
NO
number of categories known
YES
<10K samples
YES
KMeans
NOT WORKING
Spectral Clustering
GMM
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
clustering
tough luck
NO
predicting a quantity
YES
<100K samples
NO
SGD Regressor
YES
few features should be important
YES
Lasso
ElasticNet
NO
RidgeRegression
SVR(kernel='linear')
NOT WORKING
SVR(kernel='rbf')
EnsembleRegressors
regression
NO
just looking
YES
Randomized PCA
NOT WORKING
<10K samples
YES
Isomap
Spectral Embedding
NOT WORKING
LLE
NO
kernel approximation
dimensionality reduction
NO
predicting structure

Python for Data Science

Provides organized tutorials with specifics.



Quick Start

A very short introduction into machine learning problems and how to solve them using scikit-learn. Introduced basic concepts and conventions.

User Guide

The main documentation. This contains an in-depth description of all algorithms and how to apply them.

Other Versions

- scikit-learn 0.18 (stable)
- scikit-learn 0.19 (development)
- scikit-learn 0.17
- scikit-learn 0.16
- scikit-learn 0.15

Tutorials

Useful tutorials for developing a feel for some of scikit-learn's applications in the machine learning field.

API

The exact API of all functions and classes, as given by the docstrings. The API documents expected types and allowed features for all functions, and all parameters available for the algorithms.

Additional Resources

Talks given, slide-sets and other information relevant to scikit-learn.

Development

Information on how to contribute. This also contains useful information for advanced users, for example how to build their own estimators.

Flow Chart

A graphical overview of basic areas of machine learning, and guidance which kind of algorithms to use in a given situation.

FAQ

Frequently asked questions about the project and contributing.

Related packages

Other machine learning packages for Python and related projects. Also algorithms that are slightly out of scope or not well established enough for scikit-learn.

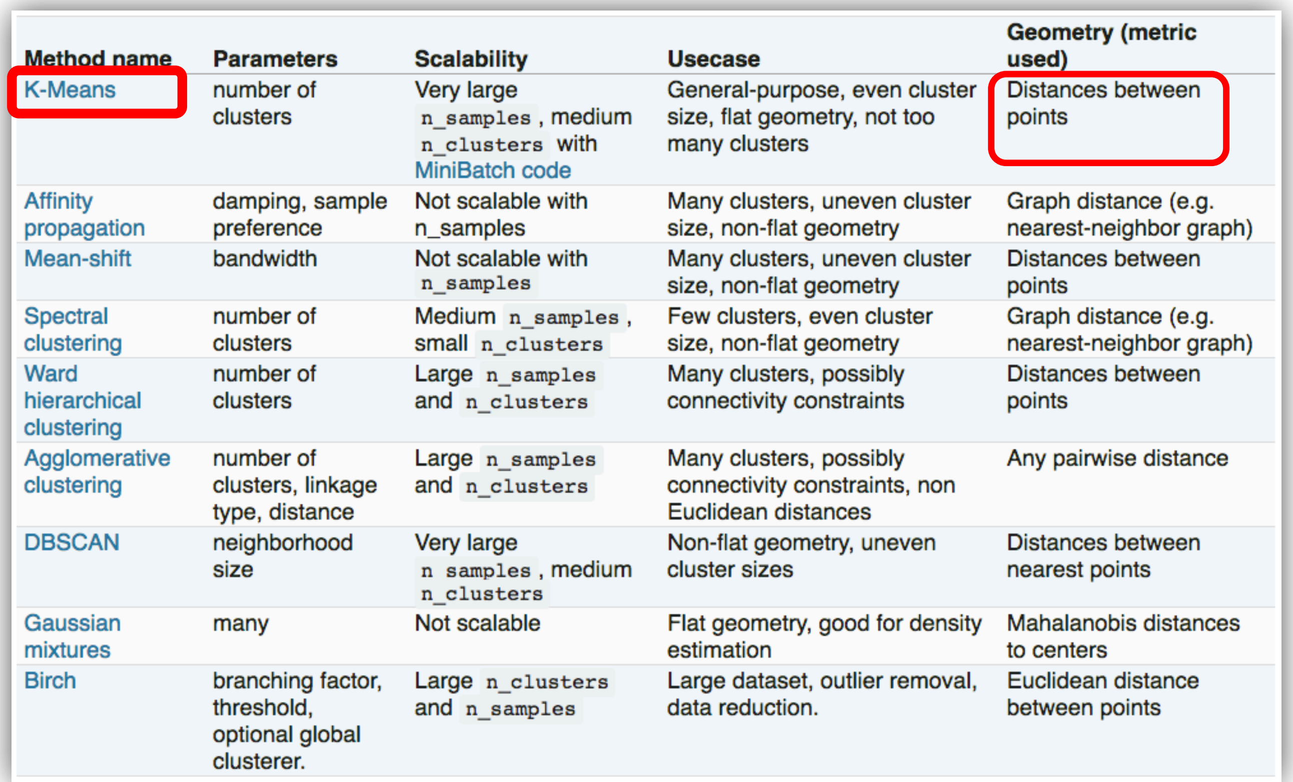
http://scikit-learn.org/stable/documentation.html

Python for Data Science

Clustering

http://scikit-learn.org/stable/modules/clustering.html#clustering

- sklearn.cluster gives algorithms for grouping of unlabeled data



Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large `n_clusters` and `n_samples`	Large dataset, outlier removal, data reduction.	Euclidean distance between points

Dimensionality Reduction

- Enables you to reduce features while preserving variance
- scikit-learn has capabilities for:
  - Principal Component Analysis (PCA)
  - Singular Value Decomposition
  - Factor Analysis
  - Independent Component Analysis
  - Matrix Factorization
  - Latent Dirichlet Allocation

Model Selection

- Provides methods for Cross Validation
- Library functions for tuning hyper parameters
- Model Evaluation mechanisms to measure model performance
- Plotting methods for visualizing scores to evaluate models

Summary of scikit-learn

- Extensive set of tools for full pipeline in Machine Learning
- Dependable due to community support
- Provides easy to use API for training, and making predictions
- Collection of the best, most popular, algorithms in one place