# INNOVATE

ONLINE CONFERENCE

## MACHINE LEARNING AND AI EDITION

aws | intel

INNOVATE
ONLINE CONFERENCE

MACHINE LEARNING
AND AI EDITION

aws | intel

# Introduction to machine learning with Python and scikit-learn

Alex Casalboni
Technical Evangelist
Amazon Web Services

@alex_casalboni

# Agenda

Machine learning in 5 minutes
Scikit-learn
Algos & demos
Resources

# Machine learning in 5 minutes

INNOVATE
ONLINE CONFERENCE

MACHINE LEARNING
AND AI EDITION

Artificial intelligence: Design software applications that exhibit humanlike behavior, e.g., speech, natural language processing, reasoning, and intuition

Machine learning: Using statistical algorithms, teach machines to learn from featurized data without being explicitly programmed

Deep learning: Using neural networks, teach machines to learn from complex data where features cannot be explicitly expressed

# Types of machine learning

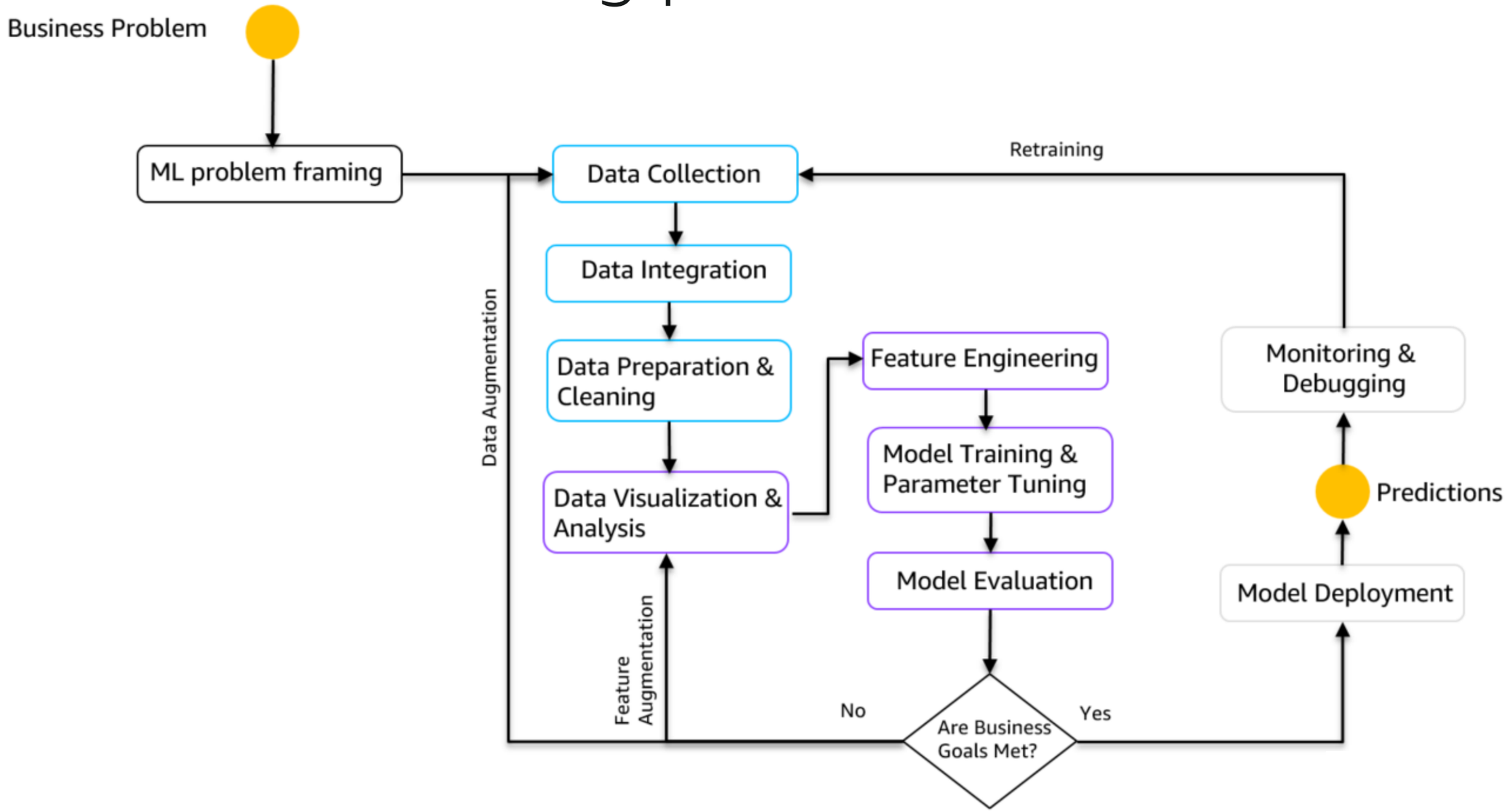
## Supervised learning

- Run an algorithm on a labeled dataset
- The model learns how to correctly predict the right answer
- Regression and classification are examples of supervised learning

## Unsupervised learning

- Run an algorithm on an unlabeled dataset
- The model learns patterns and organizes samples accordingly
- Clustering and topic modeling are examples of unsupervised learning

# The machine learning process



# Scikit-learn

INNOVATE
ONLINE CONFERENCE

MACHINE LEARNING
AND AI EDITION

# Scikit-learn


Open-source library in Python released in February 2010

Built on NumPy, SciPy, and Matplotlib

Simple tools for data analysis and machine learning

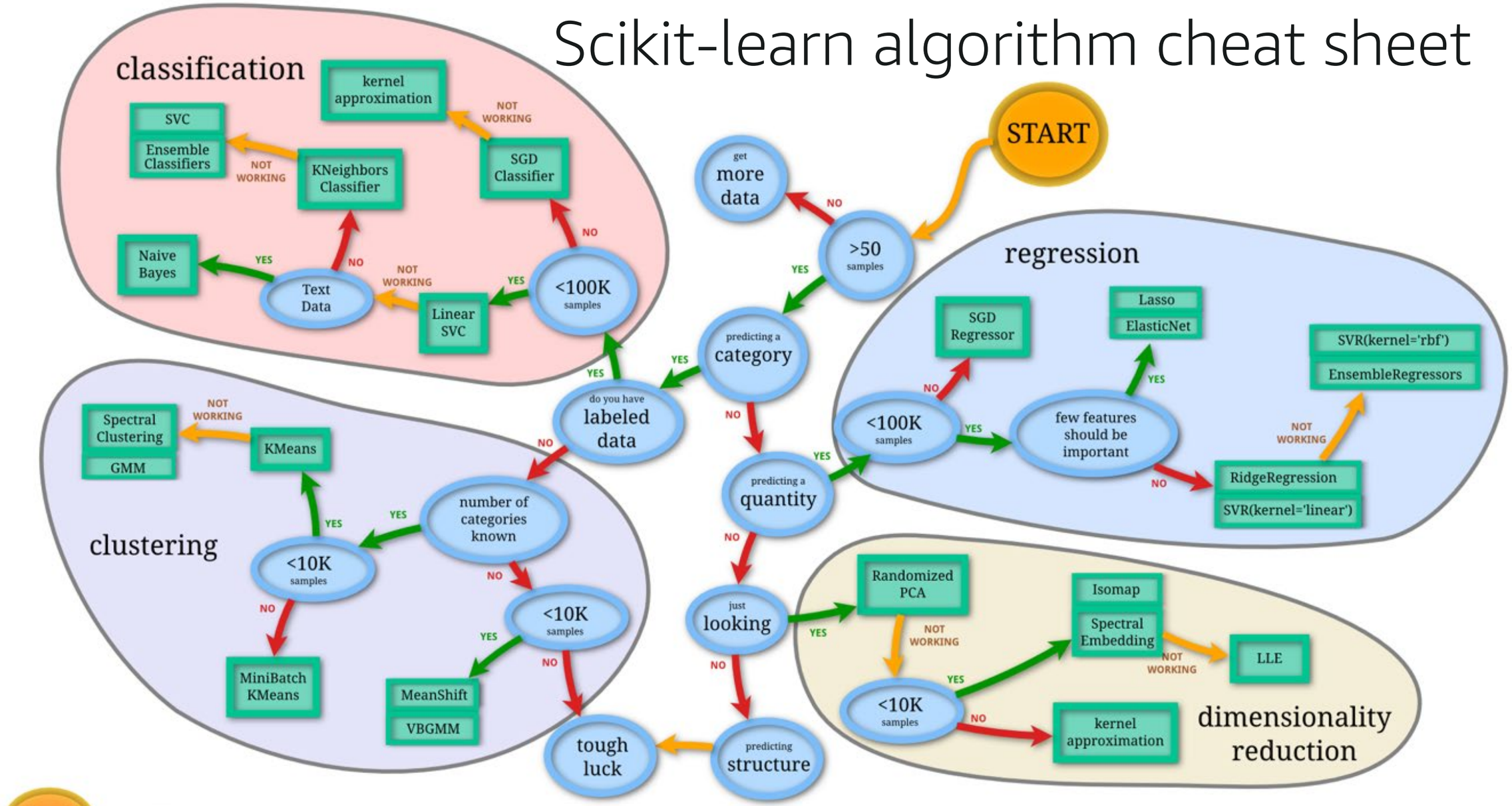
Excellent collection of algorithms

Very good documentation, tons of tutorials

*Limited scalability for datasets that don't fit in RAM*

*Not appropriate for deep learning (no GPU support)*

# Scikit-learn algorithm cheat sheet




http://scikit-learn.org/stable/tutorial/machine_learning_map/

Back

# Linear regression

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

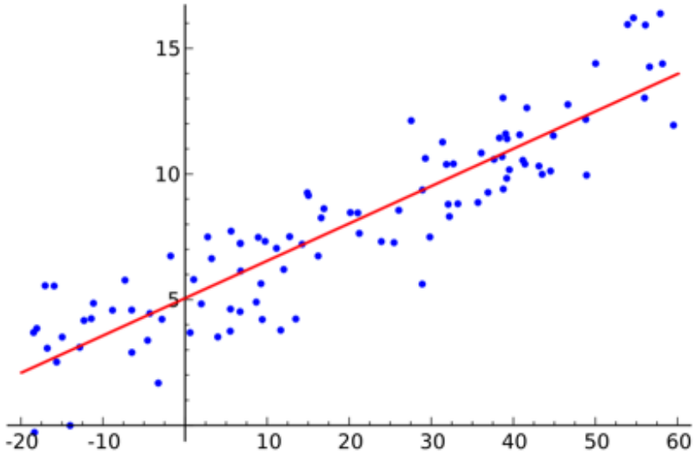
Supervised learning algorithm

Goal: Fit data to a linear function in order to predict numerical values

Dataset: Features + target (scalar or scalar vector)

1 feature → line, 2 features → plane, etc.

Intuition: Minimize the “distance” between data points and the linear function




$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

*This can also be used for binary classification: A sample is either “above” or “below” the linear function*

# Logistic regression (1958)

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

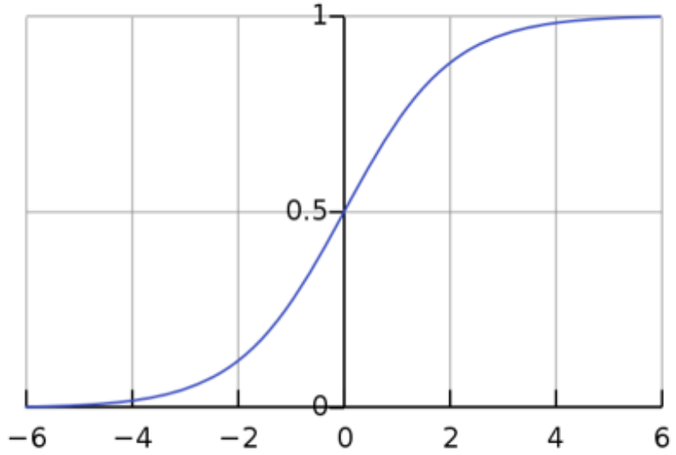
Supervised learning algorithm

Goal: Fit data to a linear function in order to predict the class of a sample

Dataset: Features + binary label (yes/no, true/false, etc.)

Can be extended to more than two classes

Intuition: Find a function computing a score between 0 and 1
and set a threshold separating both classes




$$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

# Decision trees

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

Supervised learning algorithm

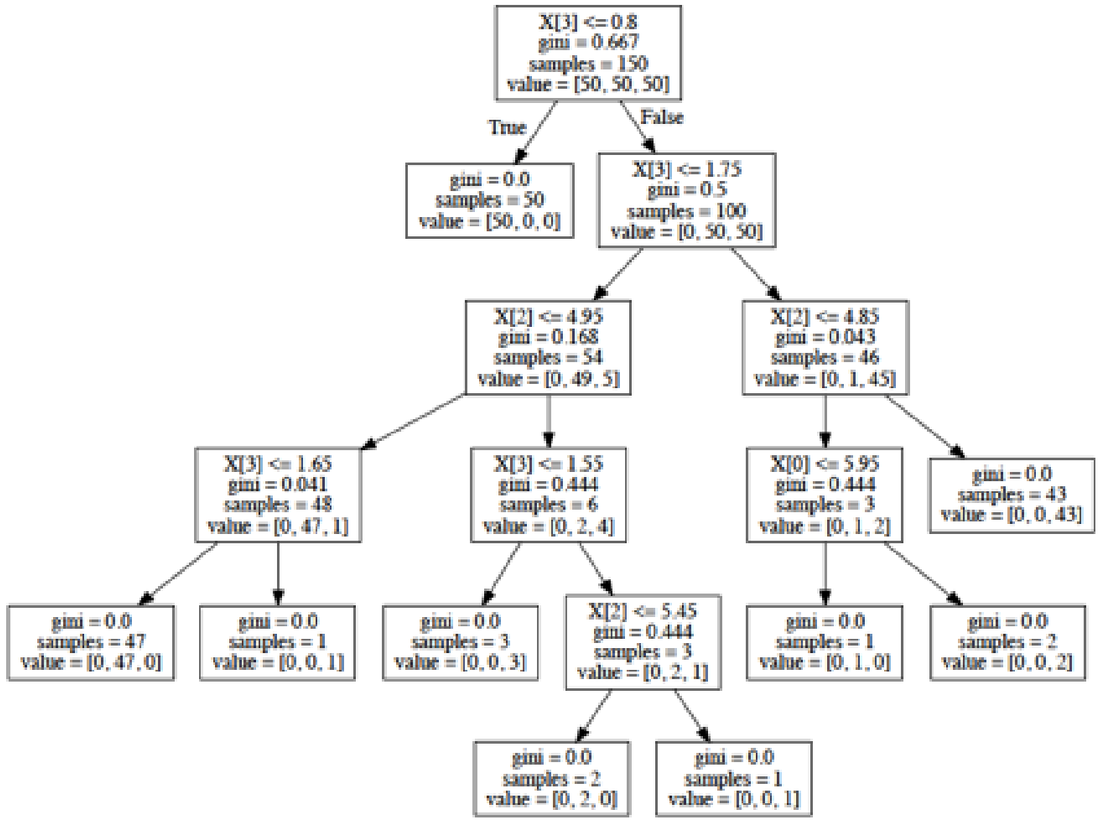
Goal: Build a decision tree for regression or classification

Dataset: Features + target value/class

Intuition: Find the “best” feature thresholds to go left or right

“Easy” to interpret, but prone to overfitting

Plenty of advanced variants with multiple trees: Random forests, XGBoost (2016), etc.



# K-means (1957)

https://en.wikipedia.org/wiki/K-means_clustering

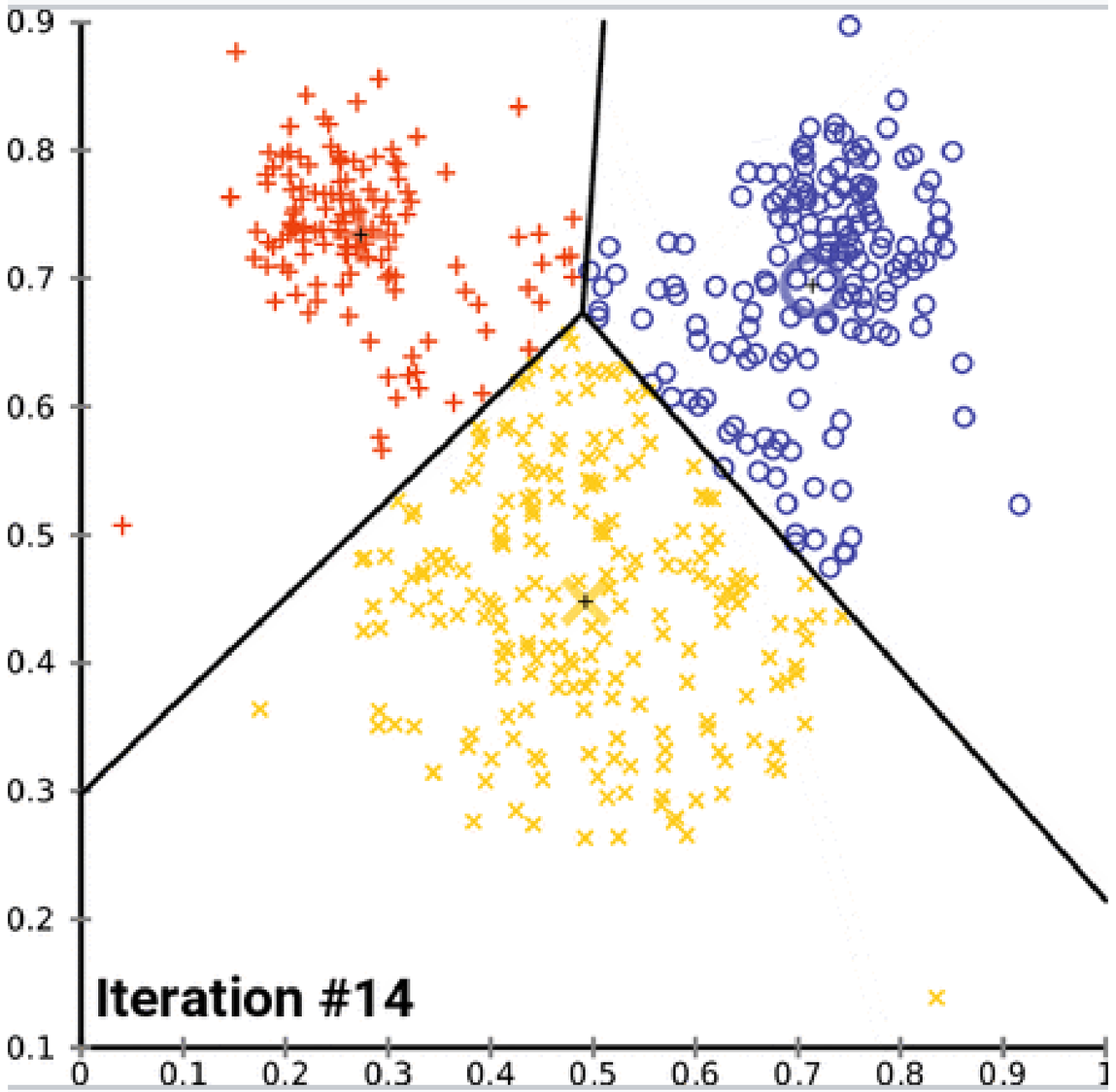
Unsupervised learning algorithm

Goal: Group samples in ‘k’ clusters

Dataset: Features only

Intuition: Find ‘k’ cluster centers that minimize the “distance” to their respective samples

This assumes “spherical” clusters of similar “radius”: Maybe, maybe not!



# Principal component analysis (aka PCA, 1901!)

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

Unsupervised learning algorithm

Goal: Build a new dataset with a smaller number of uncorrelated features (aka dimensionality reduction)

…keeping as much variance as the number of new features will allow

Dataset: Features only

Sample use cases:

- Visualize high-dimension datasets in 2D or 3D
- Remove correlation in high-dimension datasets
- Preliminary step to building linear models

# Demos

gitlab.com/alexcasalboni/aws: ML/scikit folders

# Scaling scikit-learn

Scikit-learn runs on a single machine, loading the full dataset in memory

Scaling options are quite limited: http://scikit-learn.org/stable/modules/computing.html

- Some algorithms can leverage multi-core (*joblib)*
- Some algorithms support incremental training

Amazon SageMaker can help

- Use ML-optimized multi-core instances (C5)
- Use Pipe mode, i.e., the ability to stream data from Amazon S3

# Beyond scikit-learn

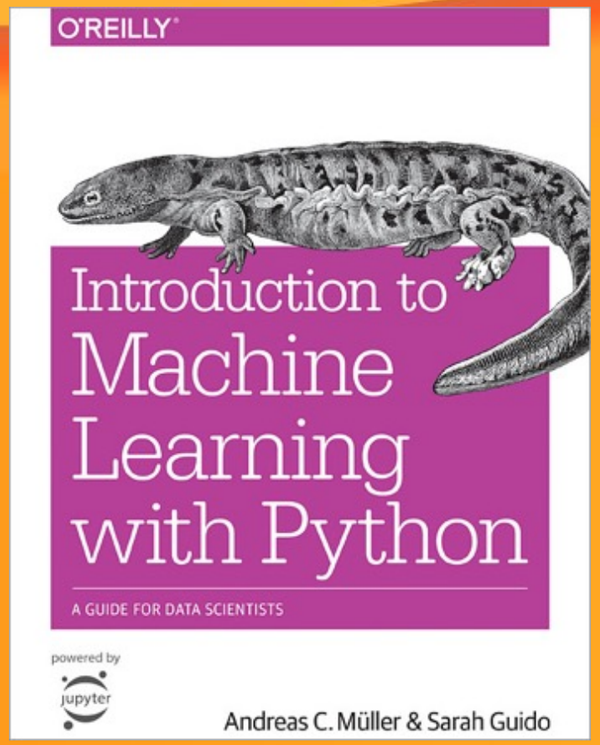
## Amazon SageMaker

- Train models on fully managed infrastructure at any scale
- Built-in algorithms (17) for regression, classification, etc.
- Built-in environments for deep learning

## Apache Spark MLlib

- Available in Amazon EMR
- Distributed processing by design
- Nice collection of machine learning algorithms
- Seamless integration with Amazon SageMaker (Scala/PySpark SDK)

# Resources




https://scikit-learn.org
https://www.numpy.org

https://ml.aws
https://aws.amazon.com/sagemaker

https://machinelearningmastery.com

https://gitlab.com/alexcasalboni/aws

# Thank you!

Alex Casalboni
Technical Evangelist
Amazon Web Services

@alex_casalboni

INNOVATE
ONLINE CONFERENCE
MACHINE LEARNING
AND AI EDITION