Machine Learning

Python and Spark

- It is now time to begin with the Machine Learning Sections of the course!
- This introduction section will discuss a general introduction to machine learning and how Spark's MLlib library works for Machine Learning.

Python and Spark

Most Machine Learning Sections have:

- Suggested Reading Assignment
- Basic Theory Lecture
- Documentation Walkthrough
- More realistic custom code example
- Consulting Project
- Consulting Project Solutions

Python and Spark

Because our different particiapnts have different backgrounds in math, we will keep the mathematics behind the machine learning algorithms light.

Python and Spark

- If you are interested in reading more about the math behind the algorithms we discuss, we will be using Introduction to Statistical Learning by Gareth James as a companion book.
- It’s freely available online.
- I’ve also include that book as extra resources.

Companion Book

- Students who want the mathematical theory should do the suggested reading assignment that will appear for each machine learning section.
- Otherwise, feel free to watch the Intro Theory Lectures for the fundamentals.

Companion Book

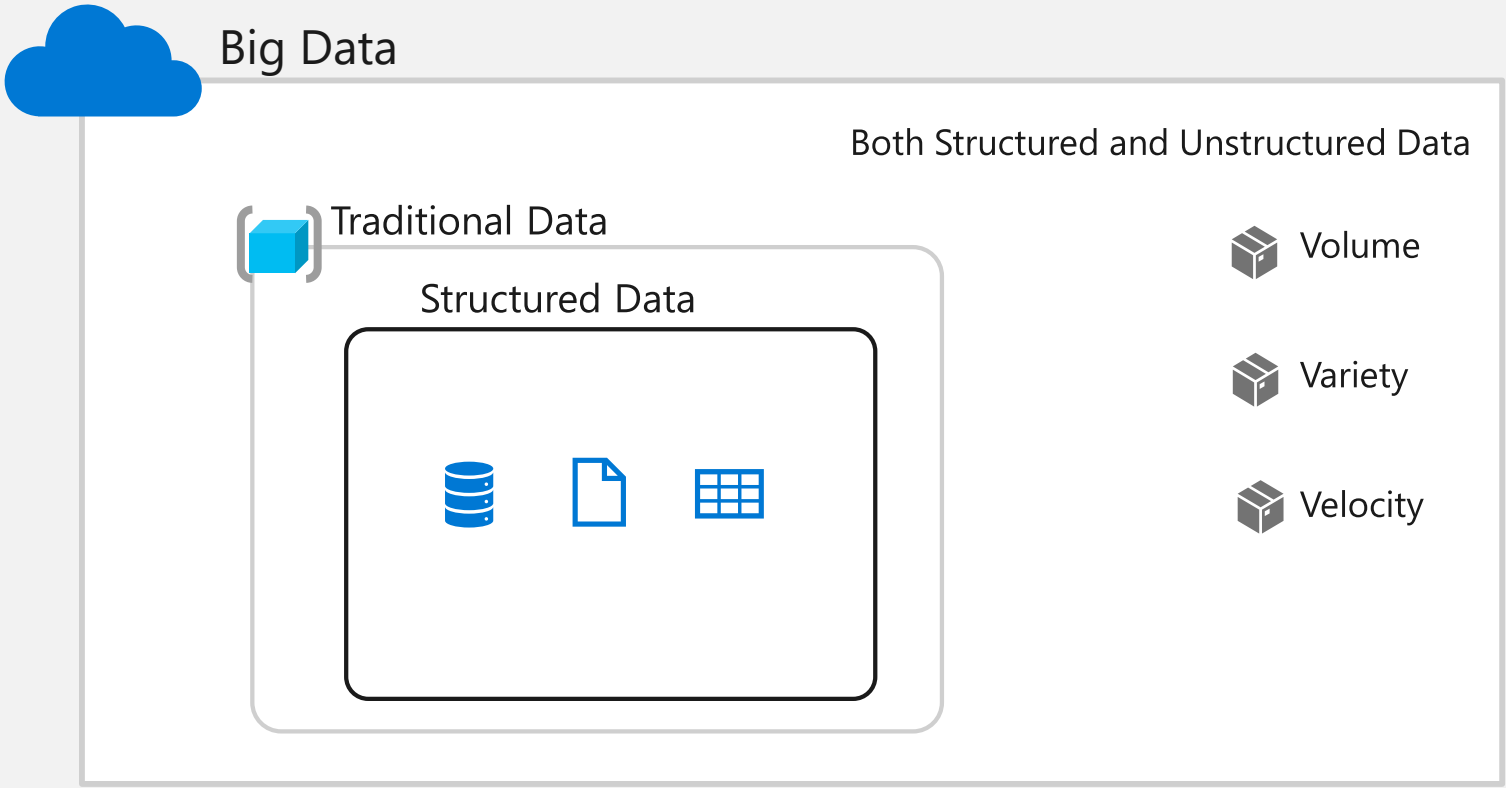
- First Suggested Reading Assignment:

  Read Chapters 1 & 2 to gain a background understanding before continuing to the Machine Learning Lectures.

What is Data Science?

Apply Scientific Methods to extract Knowledge from Data.

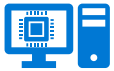
Data


Big Data
Both Structured and Unstructured Data
Traditional Data
Structured Data
Volume
Variety
Velocity

Scientific Methods

Statistics

Designed for inference about the relationships between variables



Machine Learning

Designed to make the most accurate predictions possible

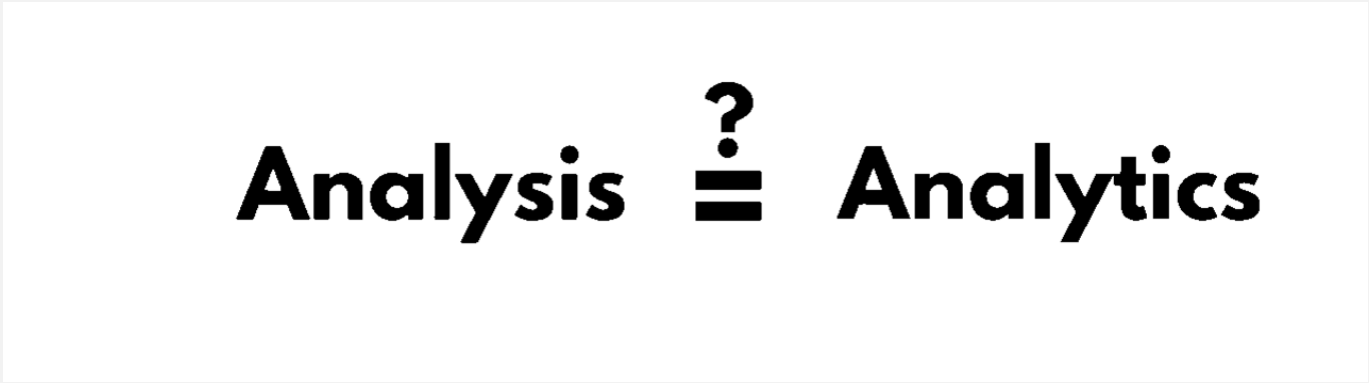
Artificial Intelligence

Designed to mimic human behavior using ML and Deep Learning

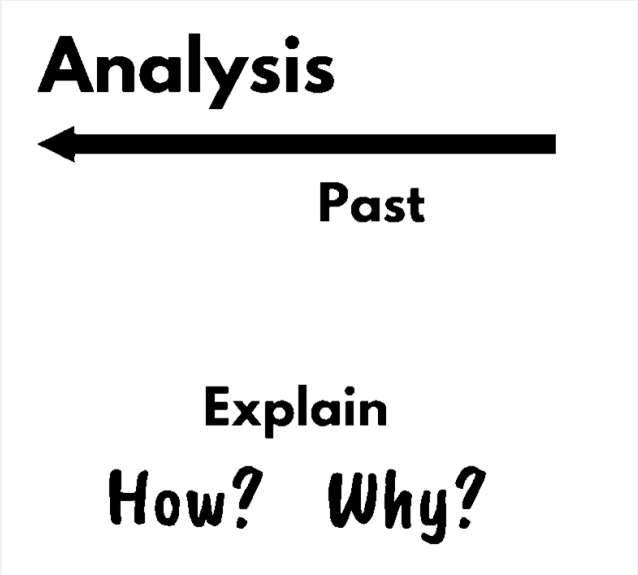
What is Machine Learning

Machine (computer) tries to find the pattern (self-learn) from the data without being explicitly programmed.

When we need to apply Machine Learning



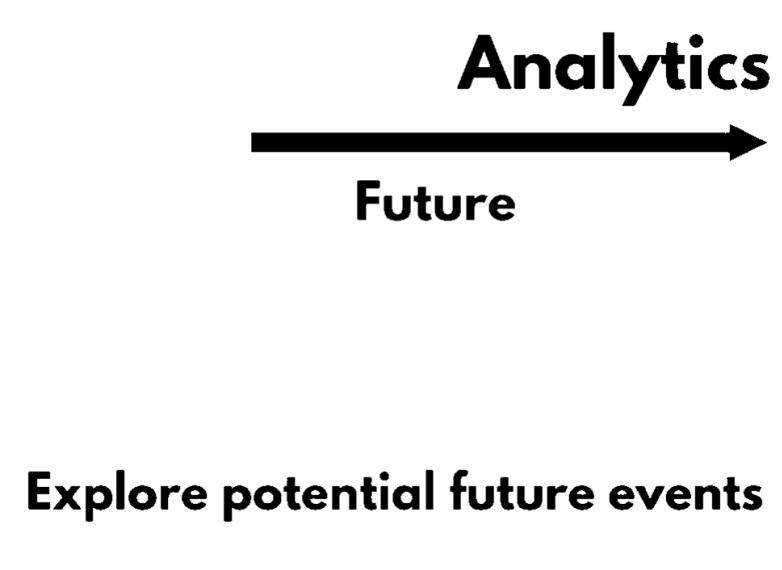
When we need to apply Machine Learning



BI

We can use different tools to explain the previous trends like, PowerBI, Tableau, Qlikview etc.

When we need to apply Machine Learning



ML/AI

We can use different language packages and framework to implement ML/AI model.

What is it used for?

- Fraud detection.
- Web search results.
- Real-time ads on web pages
- Credit scoring and next-best offers.
- Prediction of equipment failures.
- New pricing models.
- Network intrusion detection.
- Recommendation Engines
- Customer Segmentation
- Text Sentiment Analysis
- Predicting Customer Churn
- Pattern and image recognition.
- Email spam filtering.
- Financial Modeling

Supervised Learning

- Spark's MLlib is mainly designed for Supervised and Unsupervised Learning tasks, with most of its algorithms falling under those two categories.

- Let's discuss them in more detail and describe how they are different!

Supervised Learning

- **Supervised learning** algorithms are trained using **labeled** examples, such as an input where the desired output is known.

- For example, a piece of equipment could have data points labeled either “F” (failed) or “R” (runs).

Supervised Learning

Supervised learning, algorithms are trained using marked data, where the input and the output are known.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

labels

- Set of inputs ~ [Features] / [Independent Variables] / [X]
- Outputs ~ [Labels] / [Dependent Variables] / [Y]

Supervised Learning

- The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors which widely known as Model Evalution.

- It then modifies the model accordingly.

Supervised Learning

- Through methods like classification, regression, prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data.

- Supervised learning is commonly used in applications where historical data predicts likely future events.

Supervised Learning

User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION

Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

Unsupervised Learning

- **Unsupervised learning** is used against data that has no historical labels.
- The system is not told the "right answer." The algorithm must figure out what is being shown.
- The goal is to explore the data and find some structure within.

Types of Machine Learning

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

unlabeled

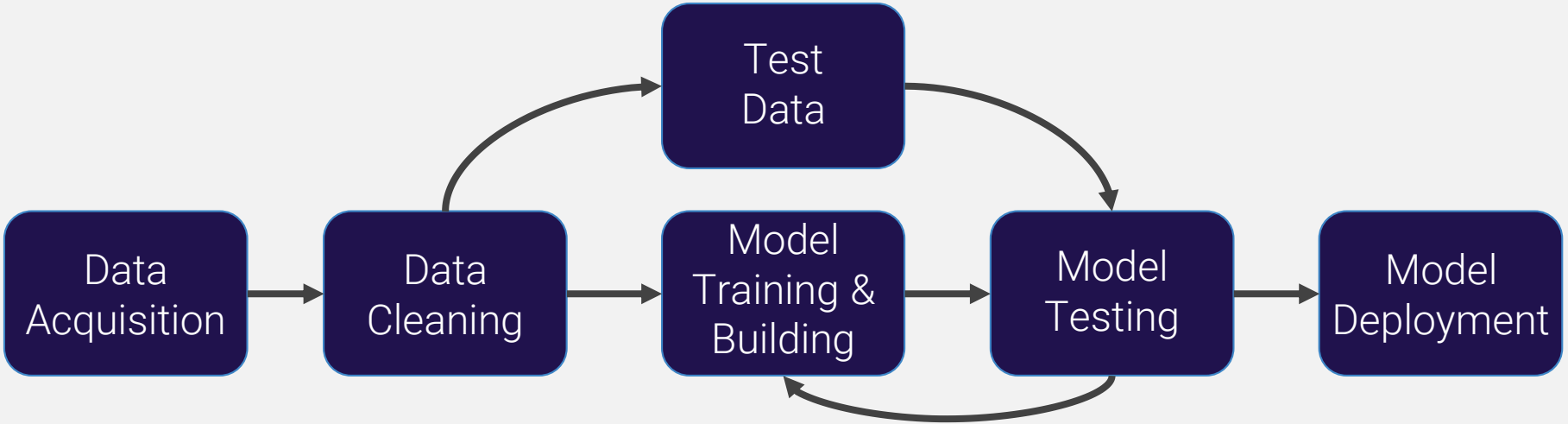
Unsupervised Learning

- For example, it can find the main attributes that separate customer segments from each other.

- Popular techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering and singular value decomposition.

- One issue is that it can be difficult to evaluate results of an unsupervised model!

Machine Learning Process



Machine Learning with Spark

Python and Spark

- Spark has its own MLlib for Machine Learning.
- The future of MLlib utilizes the Spark 2.0 DataFrame syntax.

Python and Spark

- One of the main “quirks” of using MLlib is that you need to format your data so that eventually it just has one or two columns:
    - Features, Labels (Supervised)
    - Features (Unsupervised)
- This requires a little more data processing work than some other machine learning libraries, but the big upside is that this exact same syntax works with distributed data, which is no small feat for what is going on “under the hood”!

Python and Spark

- When working with Python and Spark with MLlib, the documentation examples are always with nicely formatted data.
- However, we'll have our own custom examples that have messier, more realistic data!

Python and Spark

- A huge part of learning MLlib is getting comfortable with the documentation!
- Being able to master the skill of finding information (**not** memorization) is the key to becoming a great Spark and Python developer!

Python and Spark

Let’s jump to it now!