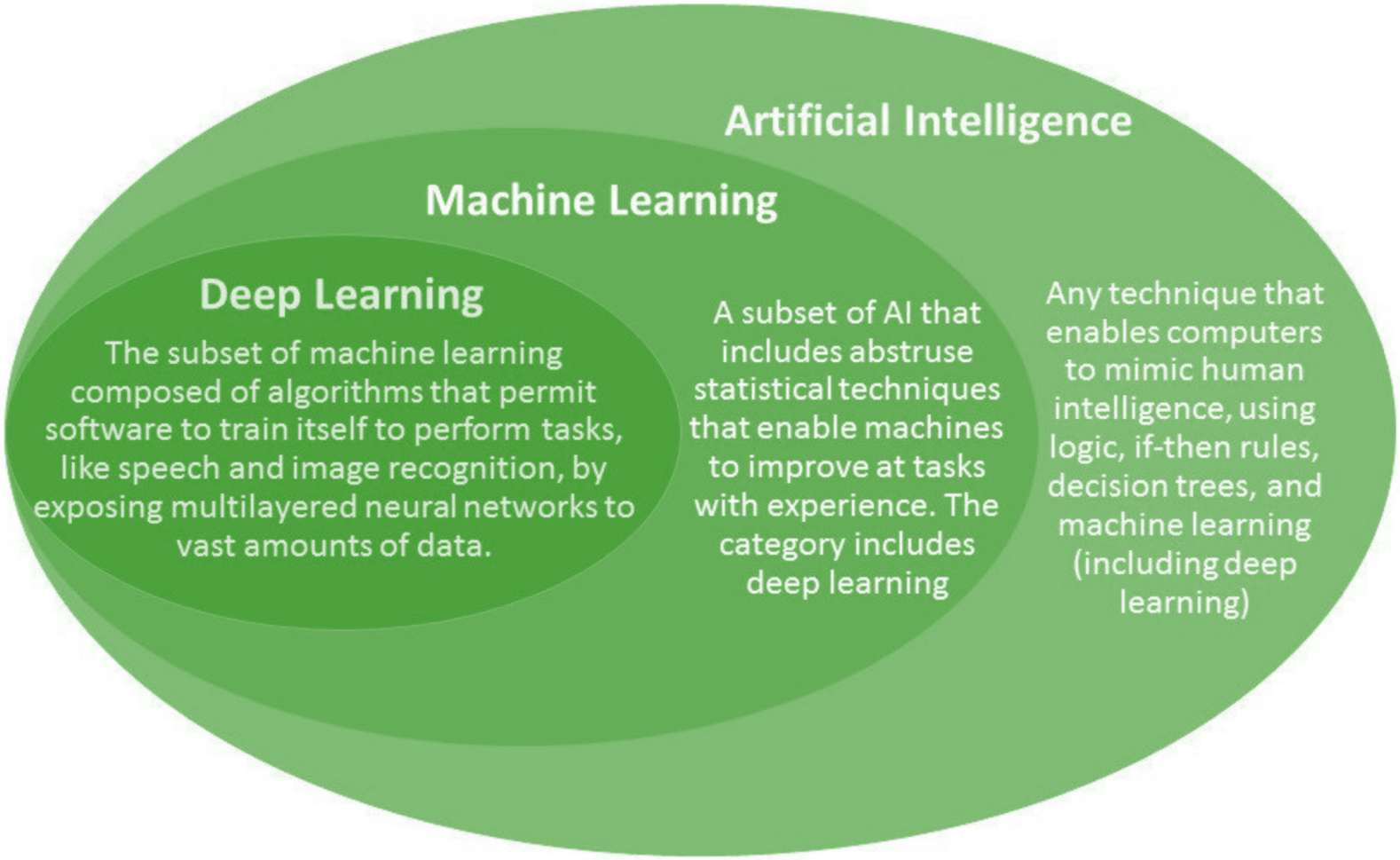
DATA SCIENCE CRASH COURSE BERLIN 2018

Robert Hryniewicz
Data Evangelist
@RobHryniewicz

HORTONWORKS

AI
MACHINE LEARNING
DEEP LEARNING
WHY NOW?


Artificial Intelligence
Any technique that enables computers to mimic human intelligence, using logic, if-then rules, decision trees, and machine learning (including deep learning)
Machine Learning
A subset of AI that includes abstruse statistical techniques that enable machines to improve at tasks with experience. The category includes deep learning
Deep Learning
The subset of machine learning composed of algorithms that permit software to train itself to perform tasks, like speech and image recognition, by exposing multilayered neural networks to vast amounts of data.

Key drivers behind AI Explosion

- **Exponential data growth**
  - And the ability to Process All Data – both Structured & Unstructured
- **Faster & open distributed systems**
  - Such as Hadoop, Spark, TensorFlow, …
- **Smarter algorithms**
  - Esp. in the Machine Learning and Deep Learning domains
  - More Accurate Models → Better ROI for Customers

Source: Deloitte Tech Trends 2017 report

HORTONWORKS

Healthcare

Predict diagnosis

Prioritize screenings

Reduce re-admittance rates

Financial services

Fraud Detection/prevention

Predict underwriting risk

New account risk screens

Public Sector

Analyze public sentiment

Optimize resource allocation

Law enforcement & security

Retail

Product recommendation

Inventory management

Price optimization

Telco/mobile

Predict customer churn

Predict equipment failure

Customer behavior analysis

Oil & Gas

Predictive maintenance

Seismic data management

Predict well production levels

HORTONWORKS

DATA

Google does not have better algorithms, only more data.

-- Peter Norvig, Dir of Research, Google

HORTONWORKS


50ZB+ in 2021

Data: The New Oil

Training Data: The *New* New Oil

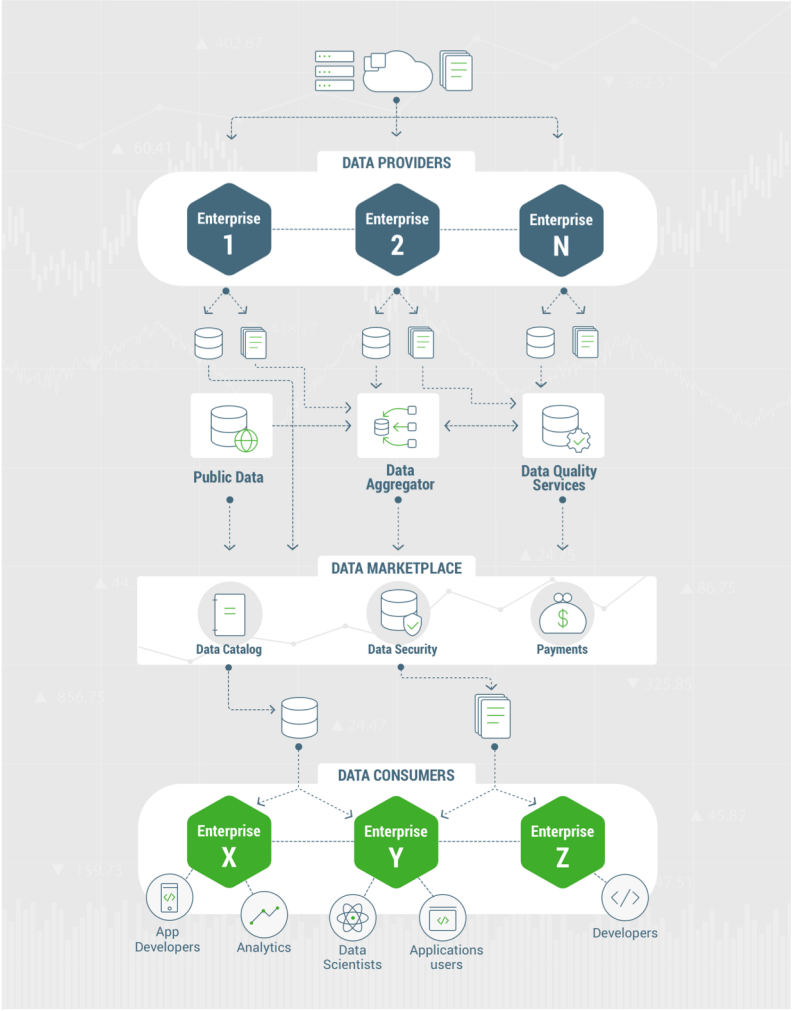
HORTONWORKS

MIT Sloan Management Review

"**Effectiveness of AI** technologies will be only **as good as the data they have access to**, and the most valuable data may exist beyond the borders of one's own organization."

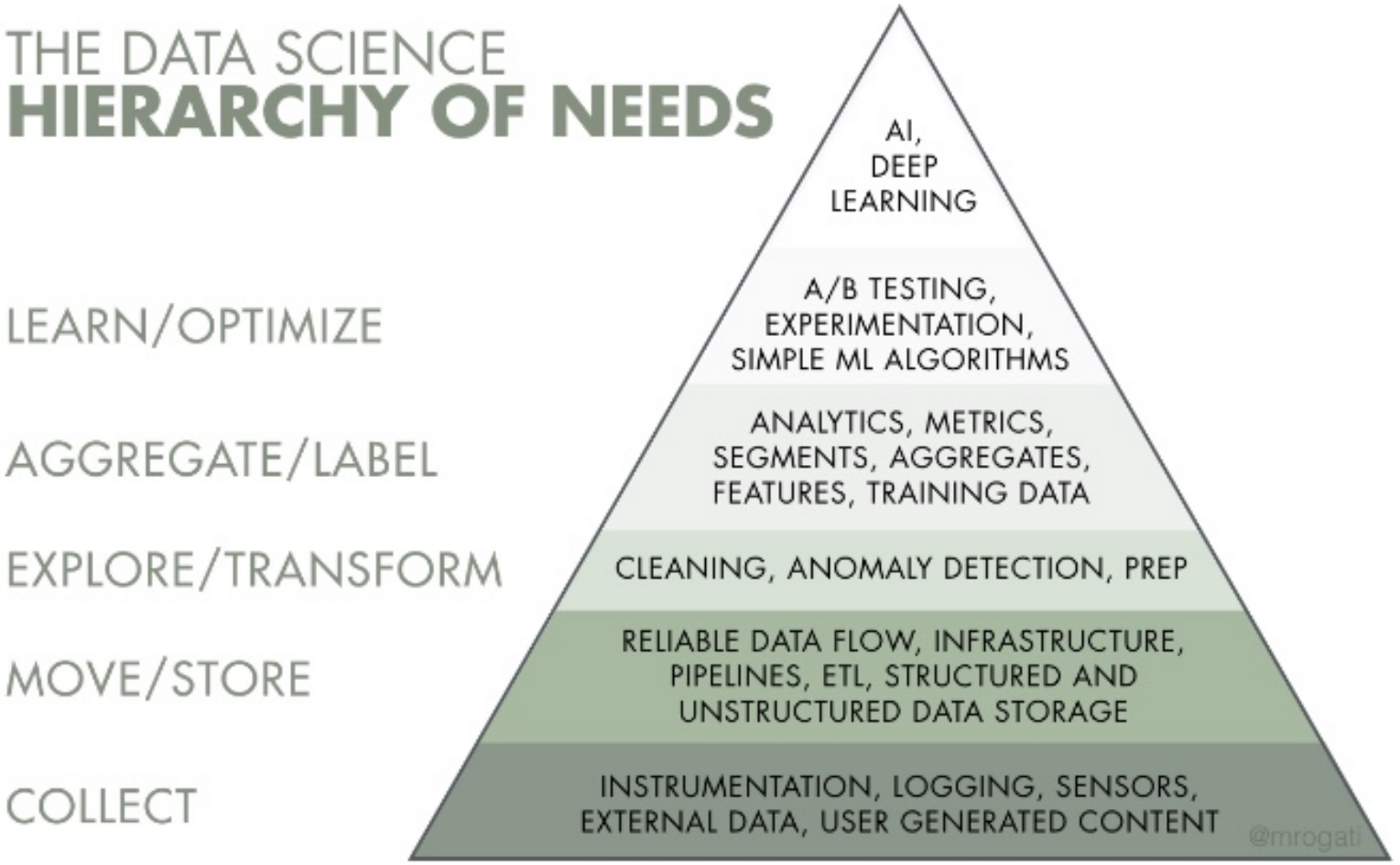
HORTONWORKS

DATA MARKETPLACES

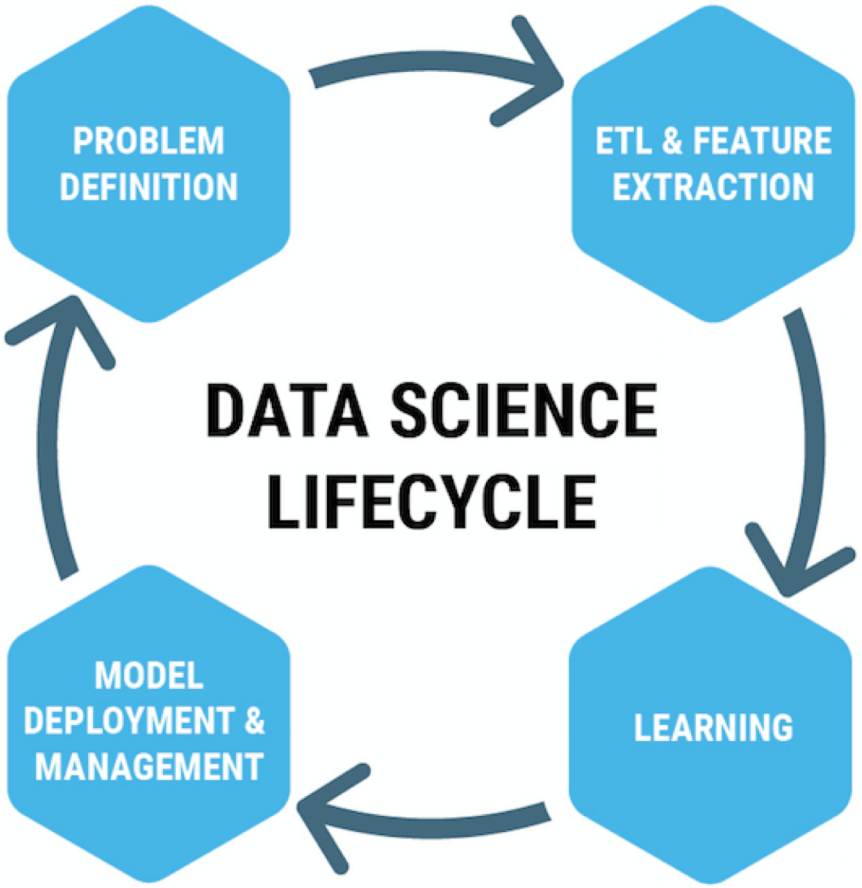


HORTONWORKS

DATA SCIENCE PREREQUISITES

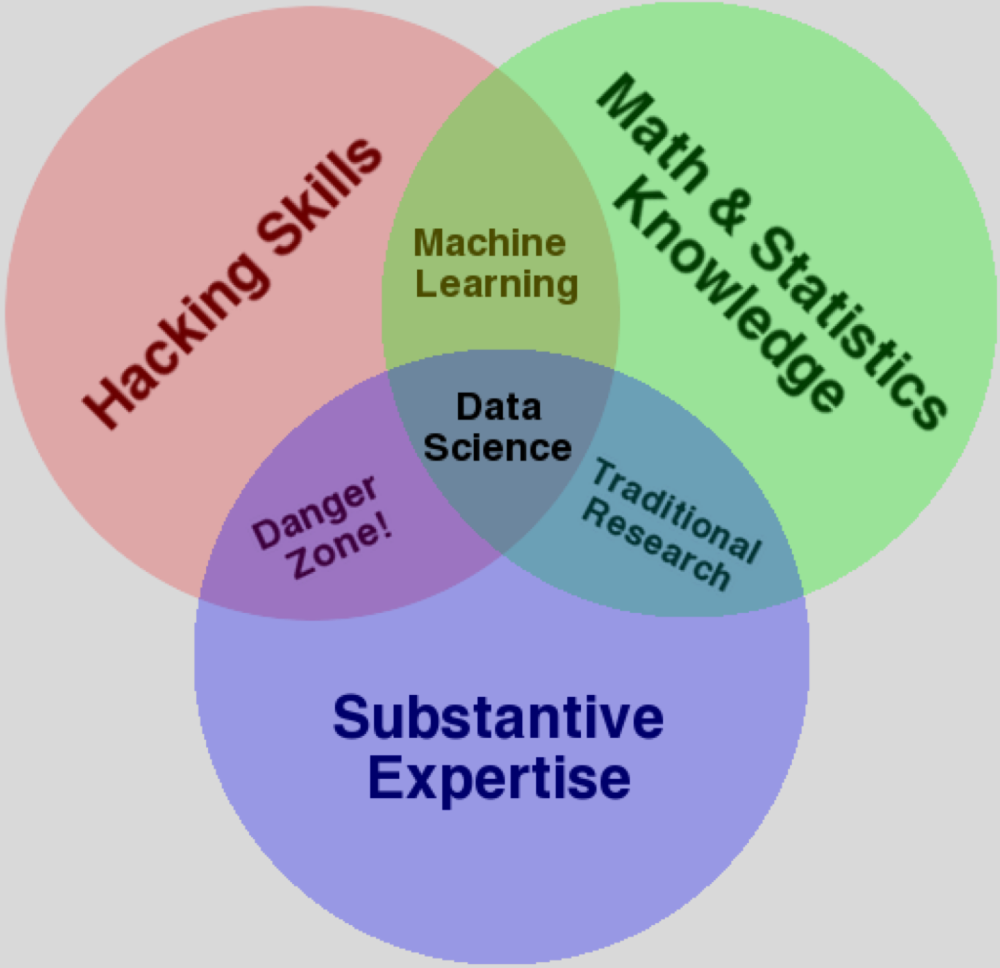


Source: hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007



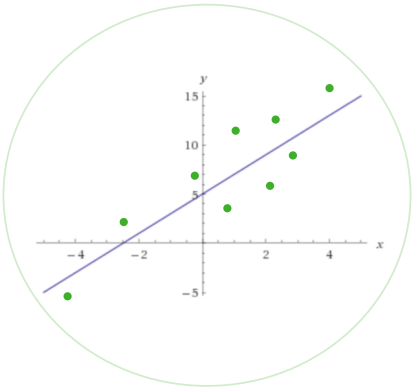
HORTONWORKS

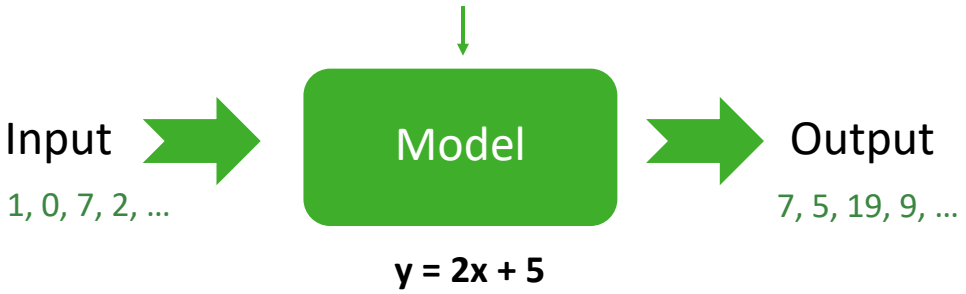
DATA SCIENCE & MACHINE LEARNING
WHAT IS A MODEL?


Hacking Skills
Math & Statistics Knowledge
Substantive Expertise
Machine Learning
Data Science
Danger Zone!
Traditional Research

What is a ML Model?

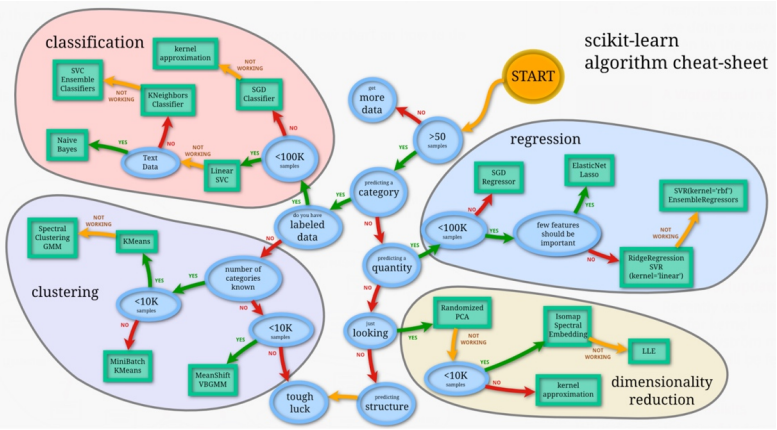
- **Mathematical formula** with a number of **parameters** that need to be **learned** from the **data**. **Fitting a model to the data** is a process known as **model training.**
- E.g. **linear regression**
  - **Goal**: fit a line **y = mx + c** to data points
  - After model training: **y = 2x + 5**

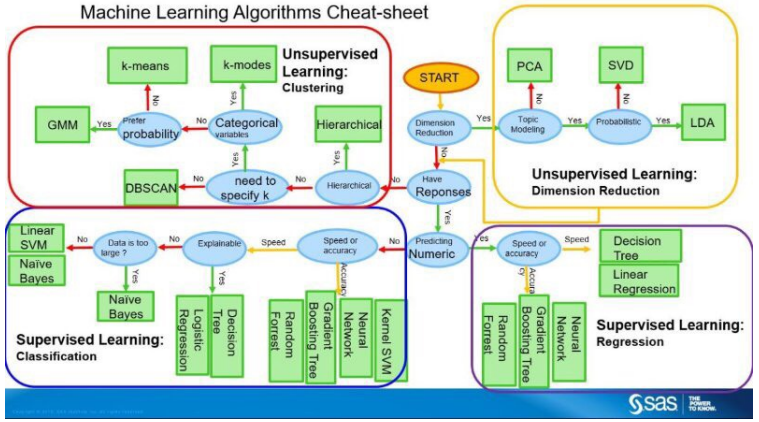


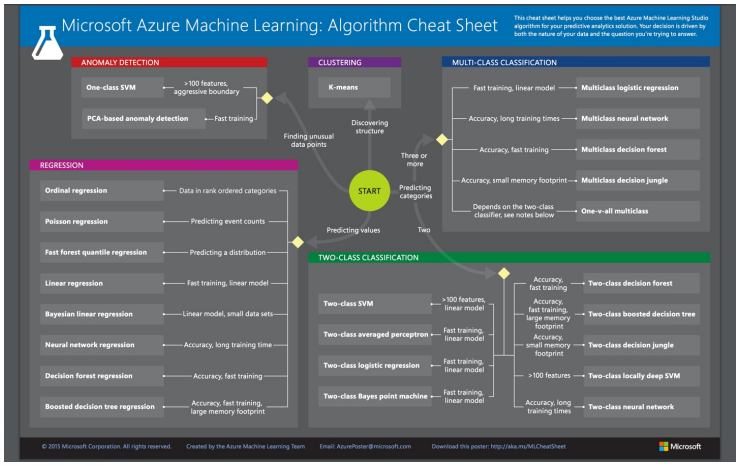


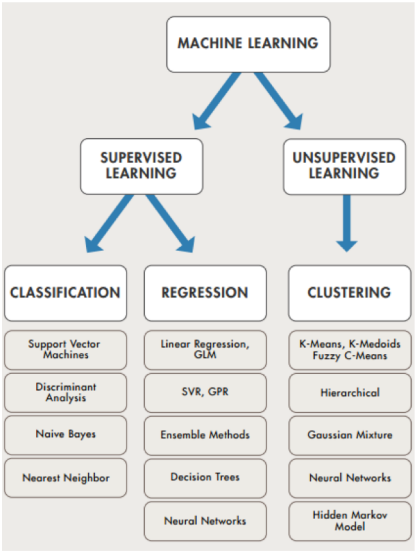
HORTONWORKS

ALGORITHMS

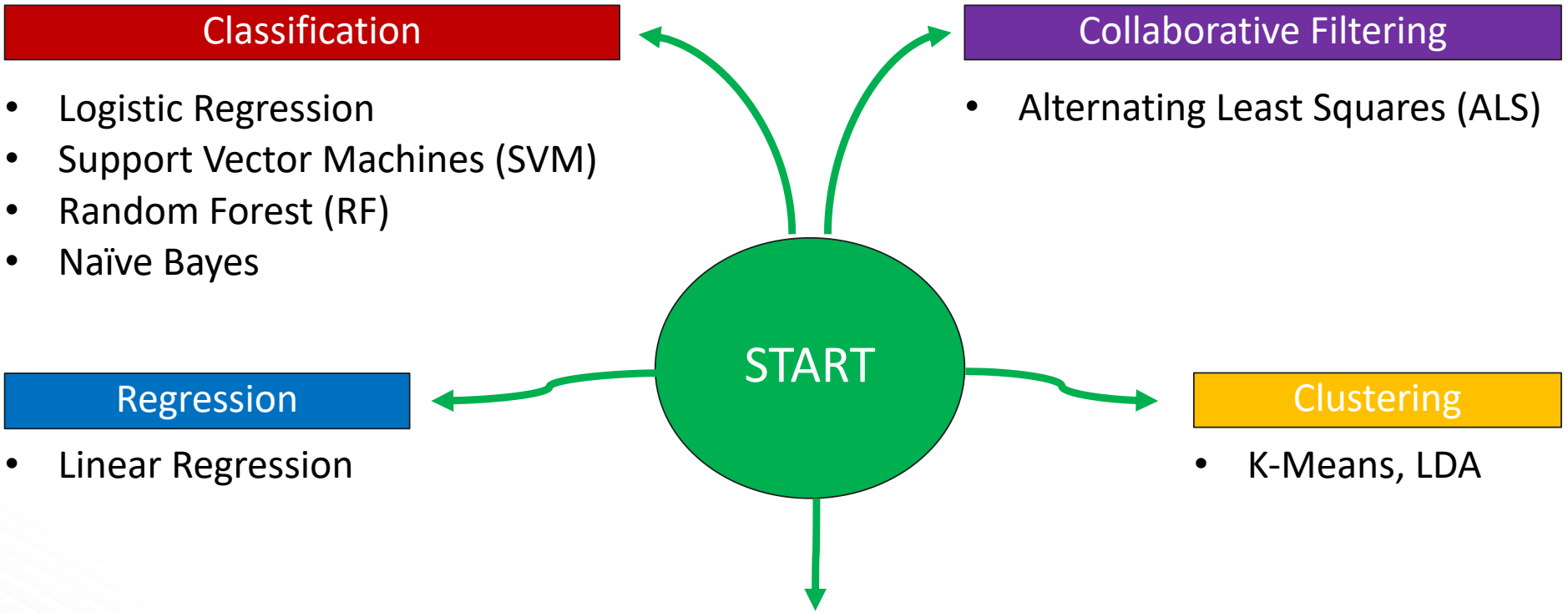








HORTONWORKS



HORTONWORKS

CLASSIFICATION

Identifying to which category an object belongs to

Examples: spam detection, diabetes diagnosis, text labeling

Algorithms:

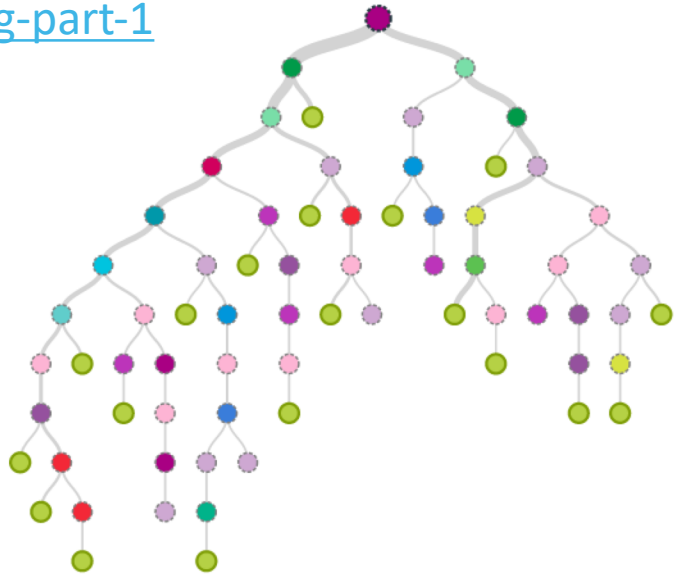
- Logistic Regression
  - Fast training, linear model
  - Classes expressed in probabilities
- Support Vector Machines (SVM)
  - "Best" supervised learning algorithm, effective
  - More robust to outliers than Log Regression
  - Handles non-linearity
- Random Forest
  - Fast training
  - Handles categorical features
  - Does not require feature scaling
  - Captures non-linearity and feature interaction
- Naïve Bayes
  - Good for text classification
  - Assumes independent variables

HORTONWORKS

CLASSIFICATION

Visual Intro to Decision Trees

- http://www.r2d3.us/visual-intro-to-machine-learning-part-1



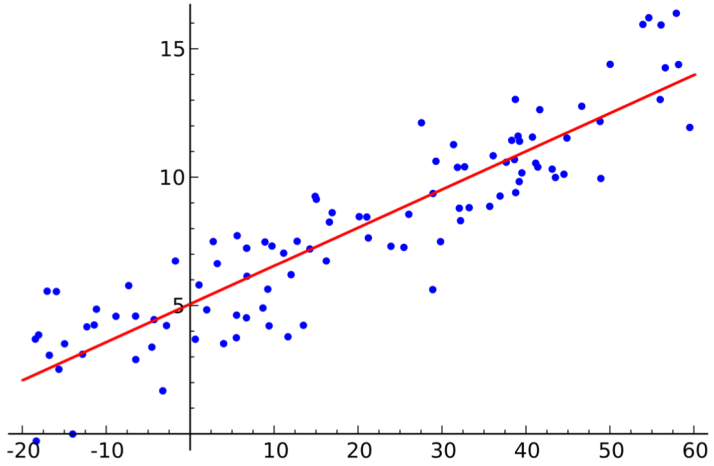
HORTONWORKS

REGRESSION

Predicting a continuous-valued output

Example: Predicting house prices based on number of bedrooms and square footage

Algorithms: Linear Regression



HORTONWORKS

CLUSTERING

Automatic grouping of similar objects into sets (clusters)

Example: market segmentation – auto group customers into different market segments

Algorithms: K-means, LDA

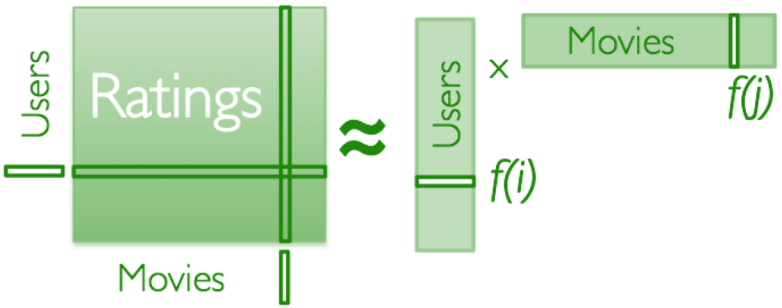
HORTONWORKS

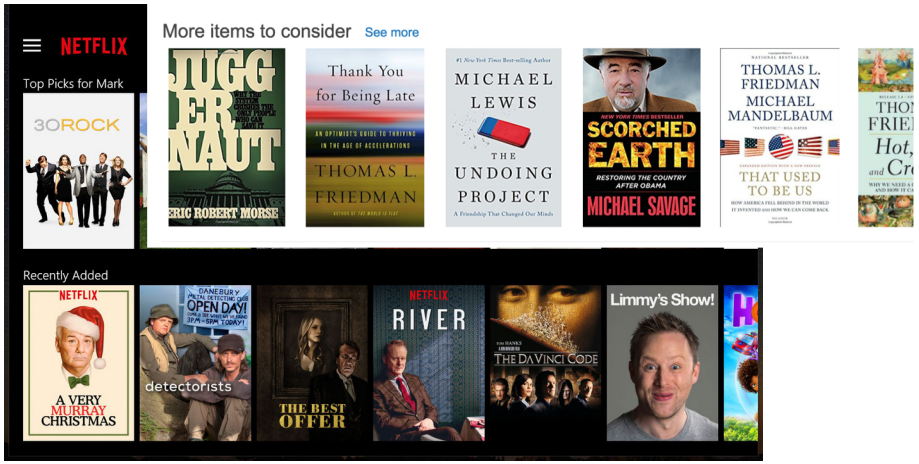
COLLABORATIVE FILTERING

Fill in the missing entries of a user-item association matrix

Applications: Product/movie recommendation

Algorithms: Alternating Least Squares (ALS)





HORTONWORKS

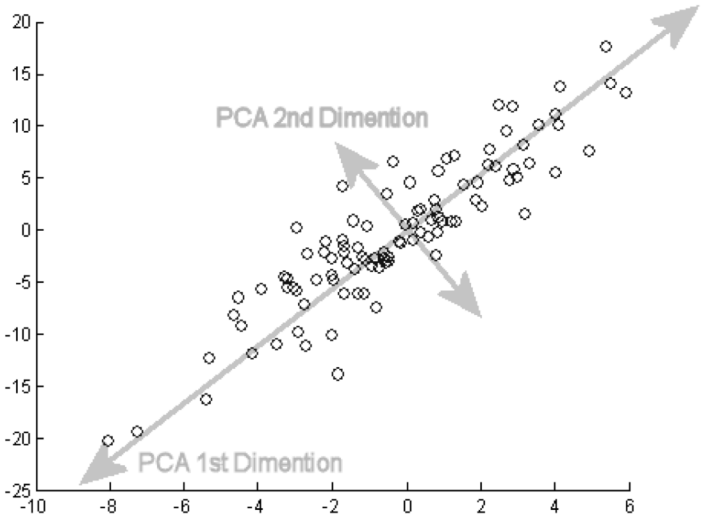
DIMENSIONALITY REDUCTION

Reducing the number of redundant features/variables

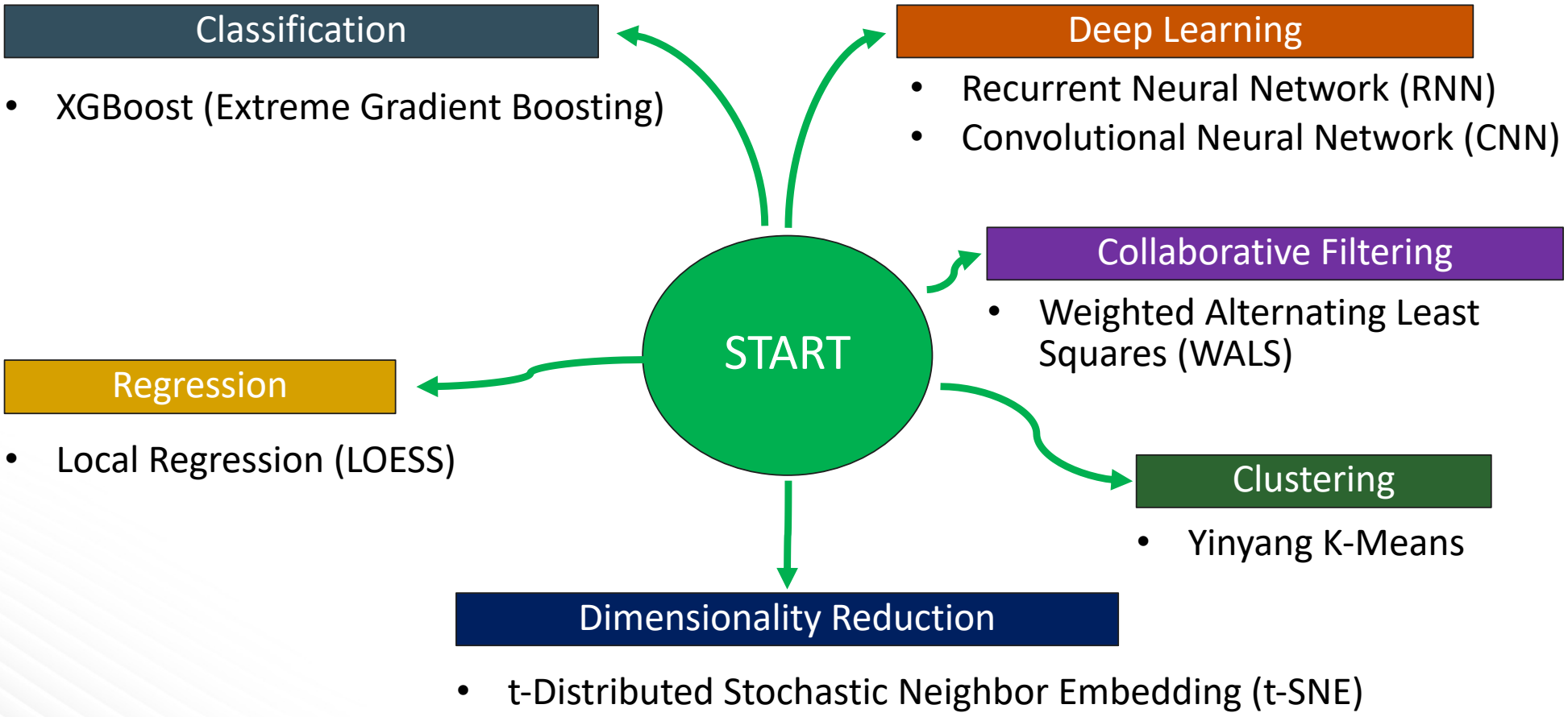
Applications:

- Removing noise in images by selecting only "important" features
- Removing redundant features, e.g. MPH & KPH are linearly dependent

Algorithms: Principal Component Analysis (PCA)

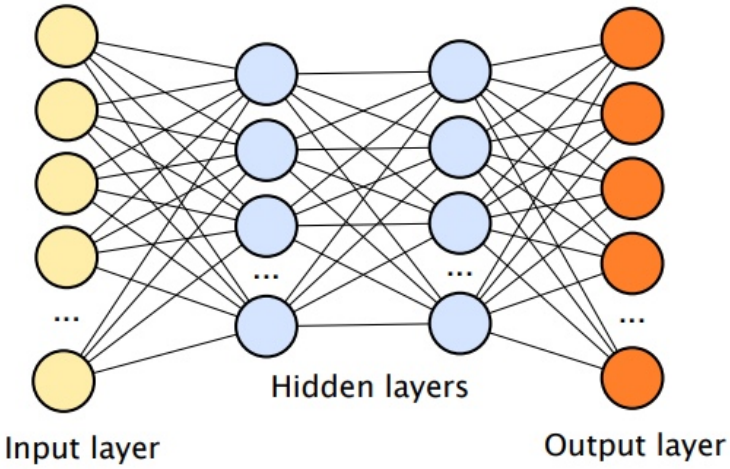


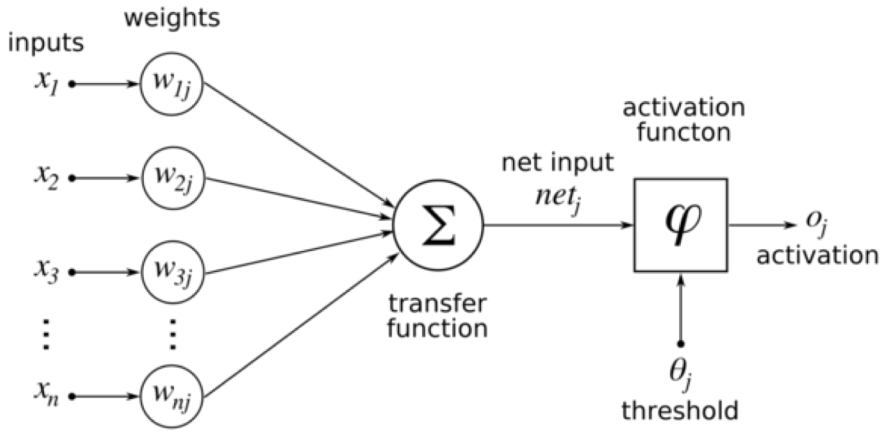
HORTONWORKS

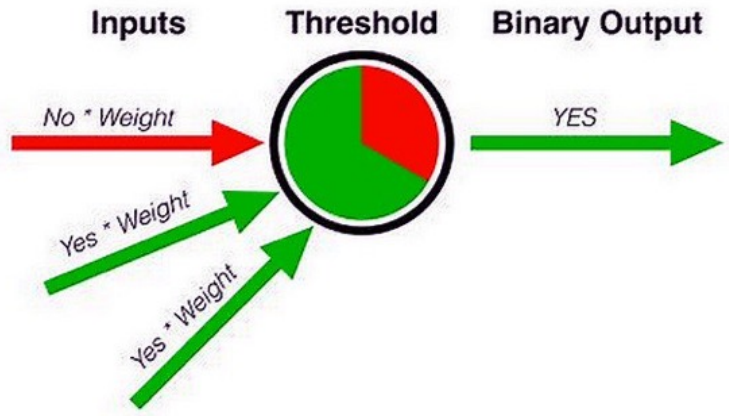


HORTONWORKS

DEEP LEARNING


Hidden layers
Input layer
Output layer


weights
inputs
x_1
x_2
x_3
x_n
w_{1j}
w_{2j}
w_{3j}
w_{nj}
Σ
transfer function
net input
net_j
activation functon
φ
o_j
activation
θ_j
threshold


Inputs
Threshold
Binary Output
No * Weight
Yes * Weight
Yes * Weight
YES

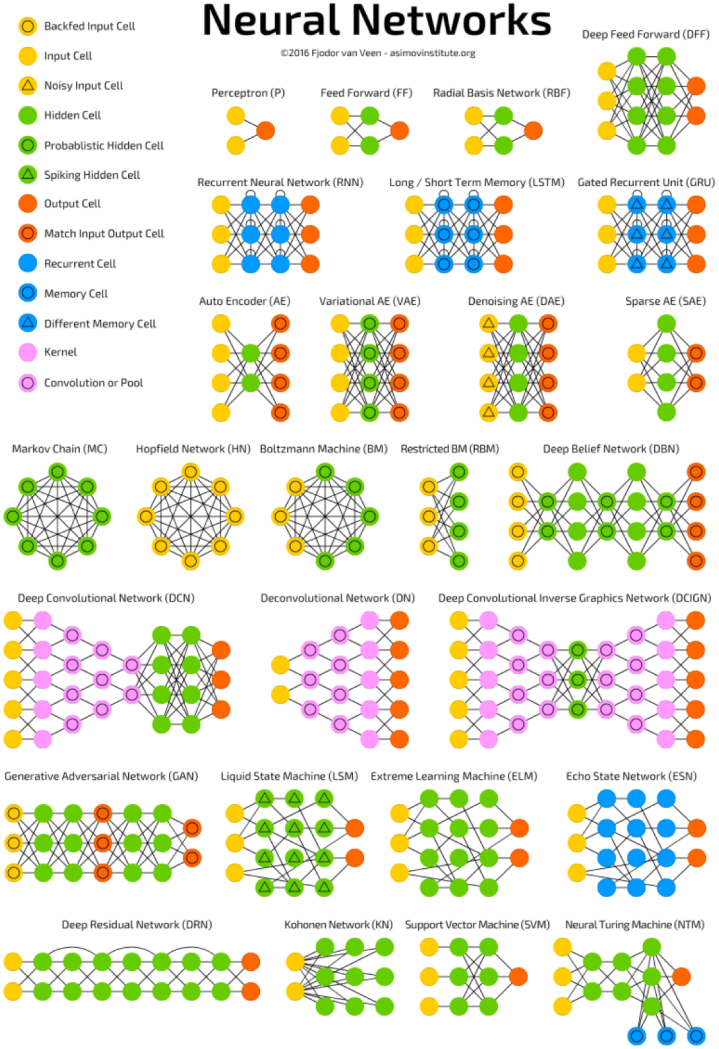
TensorFlow Playground

playground.tensorflow.org

HORTONWORKS

A mostly complete chart of

Neural Networks



Identify the right Deep Learning problems

- DL is terrific at language tasks, image classification, speech translation, machine translation, and game playing (i.e. Chess, Go, Starcraft).
- It is less performant at traditional Machine Learning tasks such as credit card fraud detection, asset pricing, and credit scoring.

Source: towardsdatascience.com/deep-misconceptions-about-deep-learning-f26c41faceec

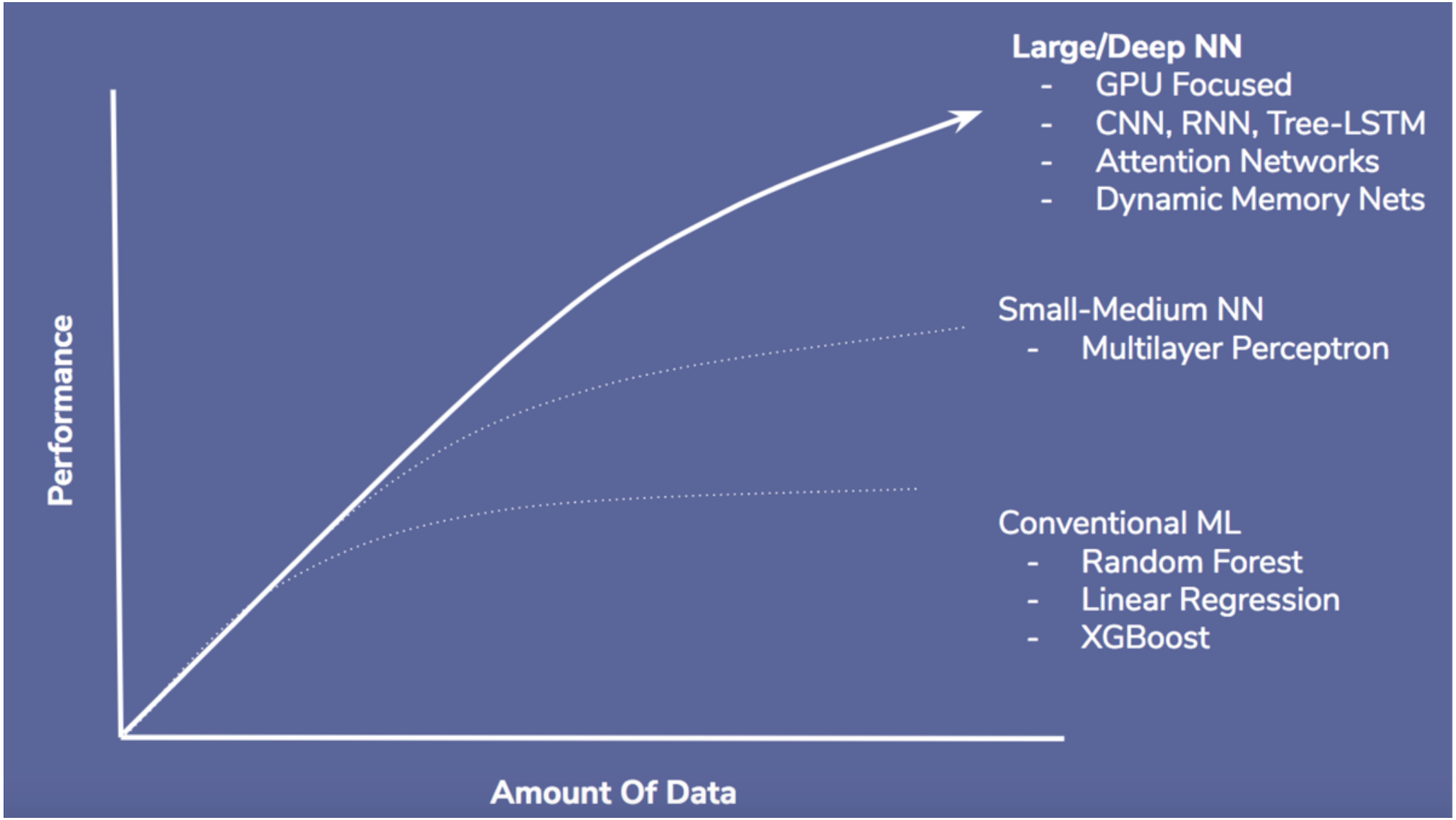
HORTONWORKS

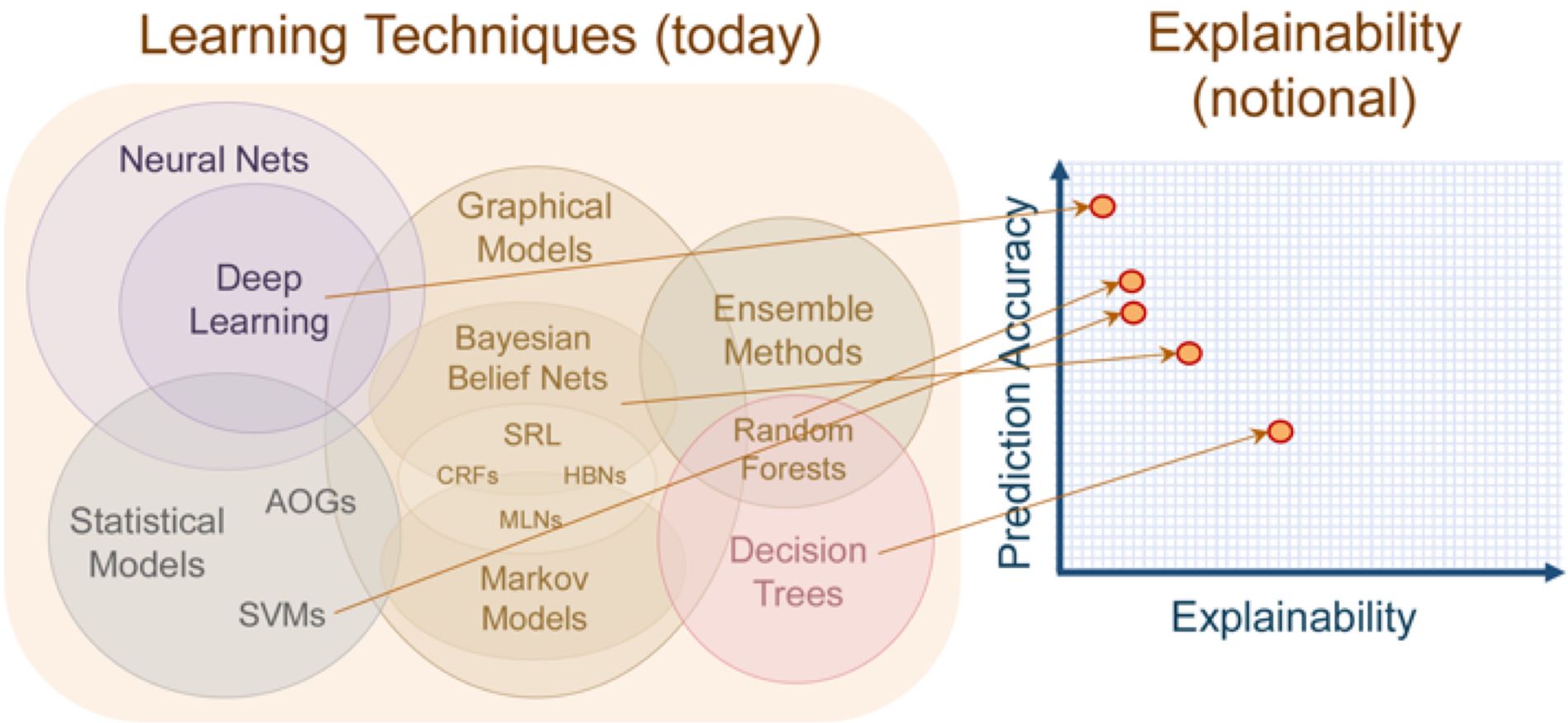
Limits to Deep Learning (DL)

- We don't have infinite datasets
  - DL is not great at generalizing
  - ImageNet: 9 layers and 60 mil parameters with 650,000 nodes from 1 mil examples with 1000 categories
- Top 10 challenges for Deep Learning
  1. Data hungry
  2. Shallow and limited capacity for transfer
  3. No natural way to deal w/ hierarchical structure
  4. Struggles w/ open-ended inference
  5. Not sufficiently transparent
  6. Now well integrated w/ prior knowledge
  7. Cannot distinguish causation from correlation
  8. Presumes stable world
  9. Works well as an approximation, but answers cannot be fully trusted
  10. Difficult to engineer with

Source: arxiv.org/pdf/1801.00631.pdf

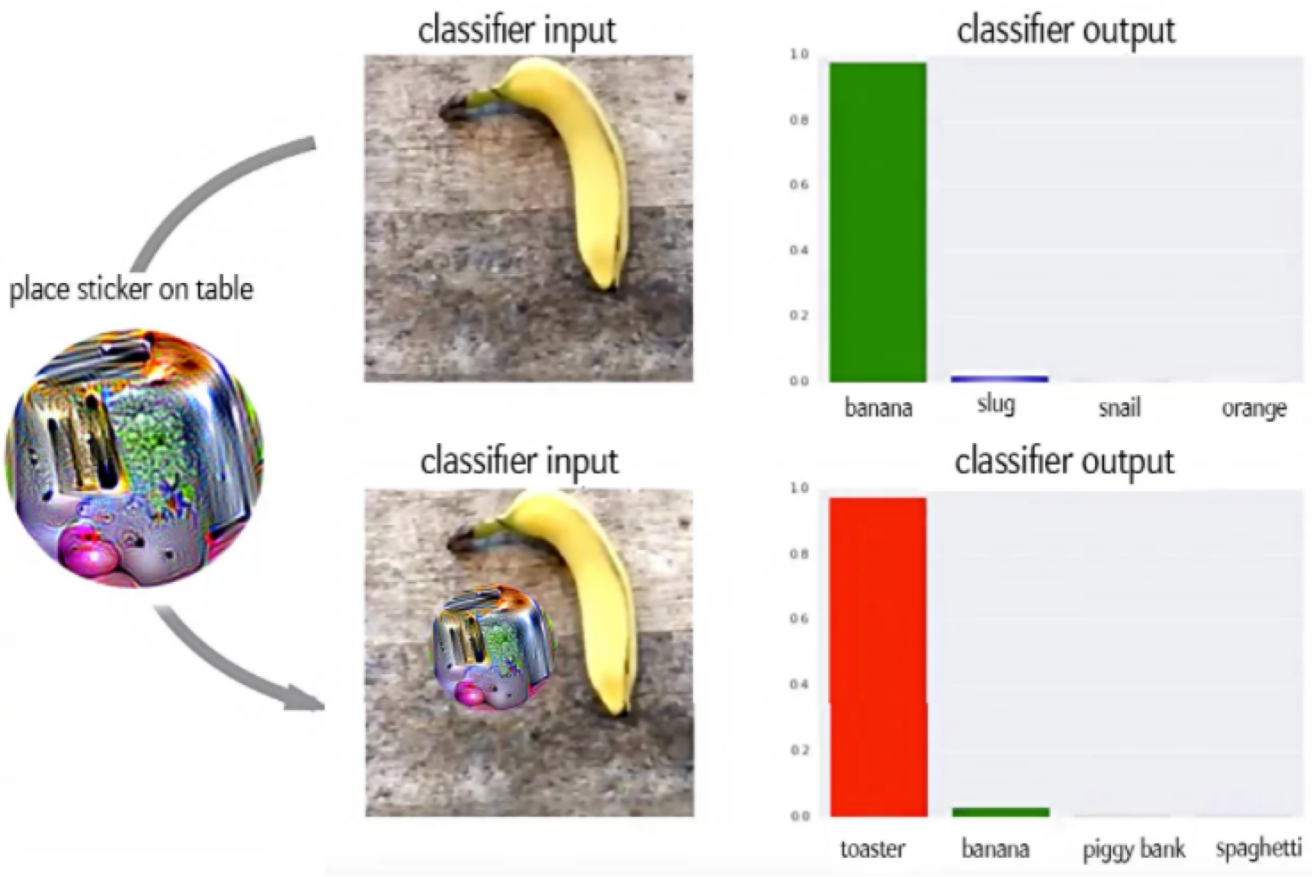
HORTONWORKS


Large/Deep NN
- GPU Focused
- CNN, RNN, Tree-LSTM
- Attention Networks
- Dynamic Memory Nets
Small-Medium NN
- Multilayer Perceptron
Conventional ML
- Random Forest
- Linear Regression
- XGBoost
Performance
Amount Of Data


Learning Techniques (today)
Neural Nets
Deep Learning
Graphical Models
Bayesian Belief Nets
SRL
CRFs
HBNs
MLNs
Markov Models
Ensemble Methods
Random Forests
Decision Trees
Statistical Models
AOGs
SVMs
Explainability (notional)
Prediction Accuracy
Explainability

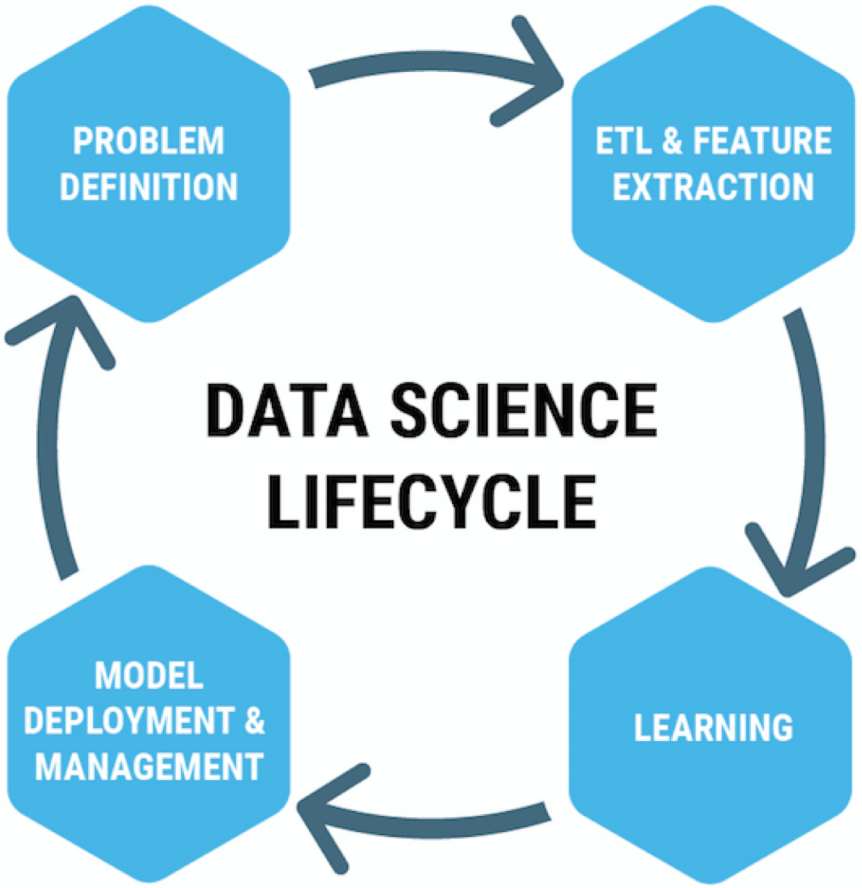
AI HACKING

Source: scientificamerican.com/article/how-to-hack-an-intelligent-machine



Source: scientificamerican.com/article/how-to-hack-an-intelligent-machine

DATA SCIENCE JOURNEY



HORTONWORKS

Start by Asking Relevant Questions

- **Specific** (can you think of a clear answer?)
- **Measurable** (quantifiable? data driven?)
- **Actionable** (if you had an answer, could you do something with it?)
- **Realistic** (can you get an answer with data you have?)
- **Timely** (answer in reasonable timeframe?)

HORTONWORKS

Data Preparation

Example of multiple values used for U.S. States ➔ California, CA, Cal**.**, Cal

1. **Data analysis** (audit for anomalies/errors)
2. **Creating an intuitive workflow** (formulate seq. of prep operations)
3. **Validation** (correctness evaluated against sample representative dataset)
4. **Transformation** (actual prep process takes place)
5. **Backflow of cleaned data** (replace original dirty data)

Approx. 80% of Data Analyst's job is Data Preparation!

HORTONWORKS

Feature Selection

Q: Which features should you use to create a predictive model?

- Also known as variable or attribute selection
- Why important?
  - simplification of models ➔ easier to interpret by researchers/users
  - shorter training times
  - enhanced generalization by reducing overfitting
- Dimensionality reduction vs feature selection
  - Dimensionality reduction: create new combinations of attributes
  - Feature selection: include/exclude attributes in data **without changing** them

HORTONWORKS

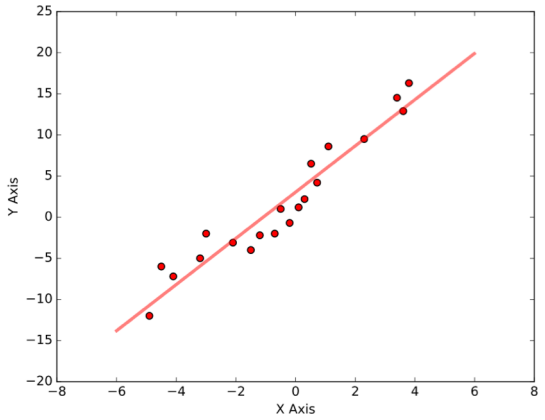
Hyperparameters

- Define higher-level model properties, e.g. complexity or learning rate
- Cannot be learned during training → need to be predefined
- Can be decided by
  - setting different values
  - training different models
  - choosing the values that test better
- Hyperparameter examples
  - Number of leaves or depth of a tree
  - Number of latent factors in a matrix factorization
  - Learning rate (in many models)
  - Number of hidden layers in a deep neural network
  - Number of clusters in a k-means clustering

HORTONWORKS

Residuals

- residual of an observed value is the **difference** between the **observed** value and the ***estimated*** value



R2 (R Squared) – Coefficient of Determination

- indicates a **goodness of fit**
- R2 of **1** means regression line **perfectly fits data**

RMSE (Root Mean Square Error)

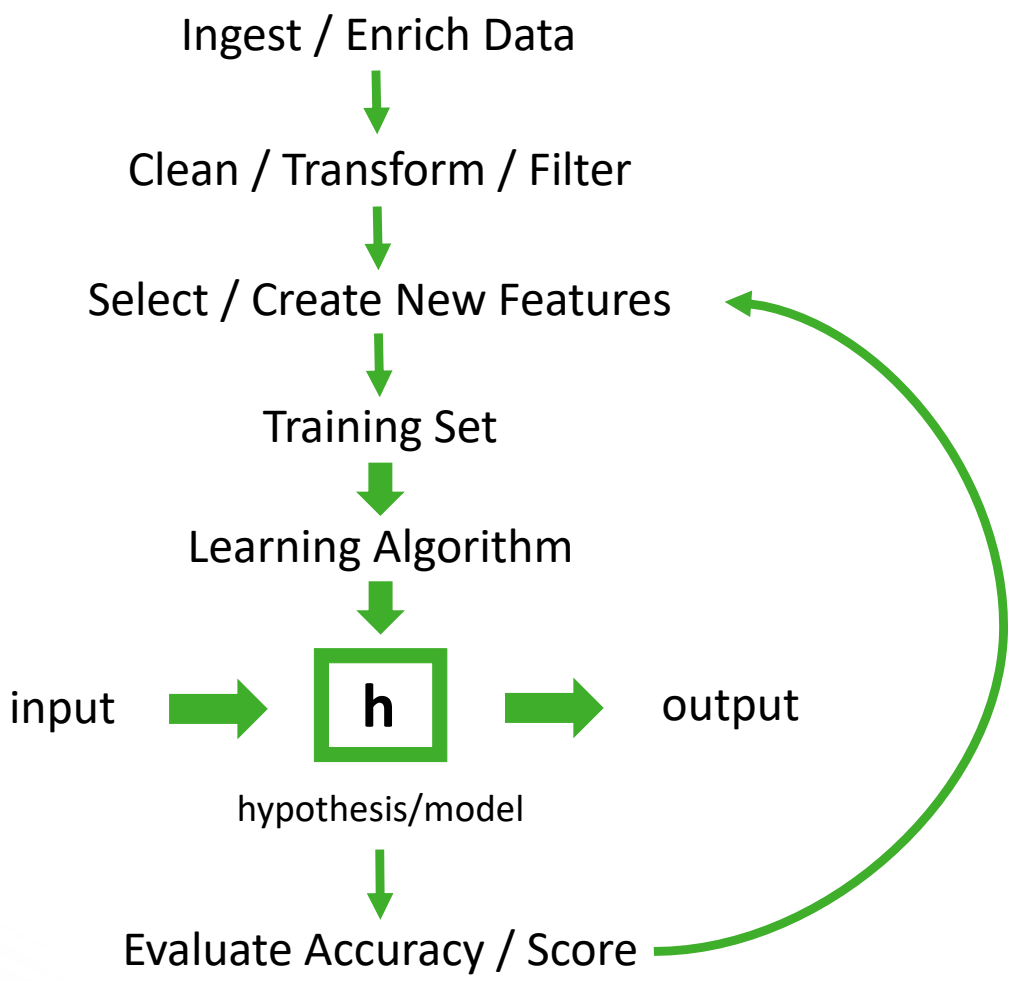
- measure of **differences** between **values predicted** by a model and **values** actually **observed**
- good measure of accuracy, but only to compare forecasting errors of different models (individual variables are scale-dependent)

HORTONWORKS

With that in mind…

- No simple formula for “good questions” only general guidelines
- The right data is better than lots of data
- Understanding relationships matters

HORTONWORKS

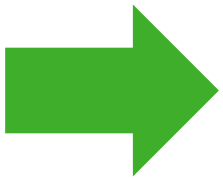


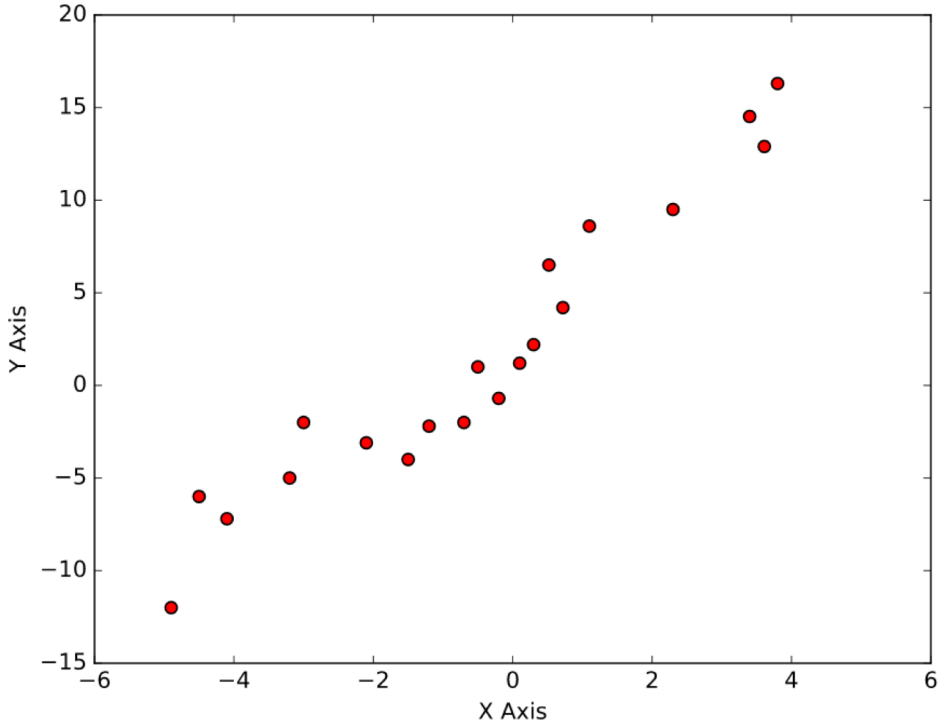
HORTONWORKS

MODEL TRAINING

Scatter Data

```
|label|features|
|-12.0| [-4.9]|
| -6.0| [-4.5]|
| -7.2| [-4.1]|
| -5.0| [-3.2]|
| -2.0| [-3.0]|
| -3.1| [-2.1]|
| -4.0| [-1.5]|
| -2.2| [-1.2]|
| -2.0| [-0.7]|
| 1.0|  [-0.5]|
| -0.7| [-0.2]|
...
...
...
```





HORTONWORKS

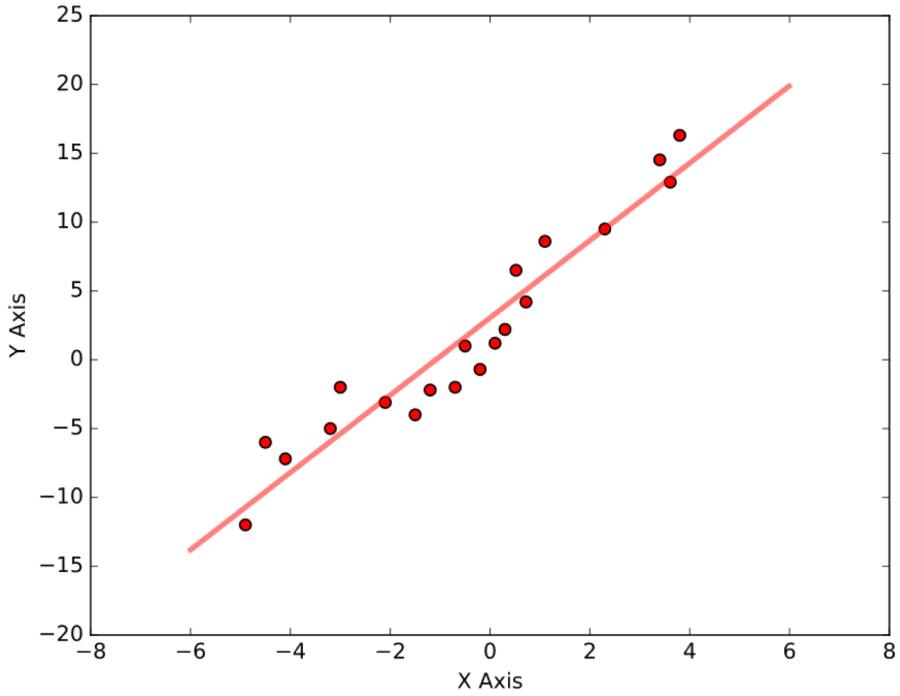
```
import org.apache.spark.ml.regression.LinearRegression

// Initialize model
val lr2D = new LinearRegression()

// Fit the model
val lrModel2D = lr1.fit(scatterData)
```

Training Result

Coefficients: 2.81 Intercept: 3.05

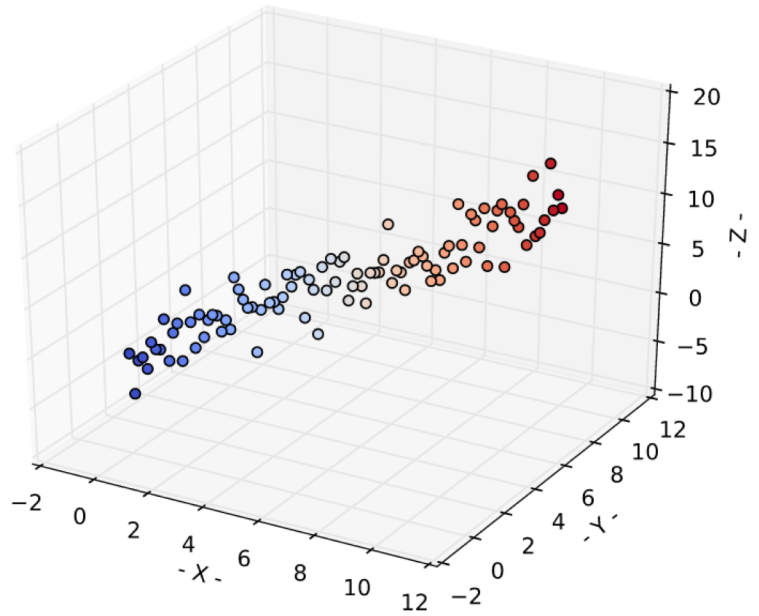


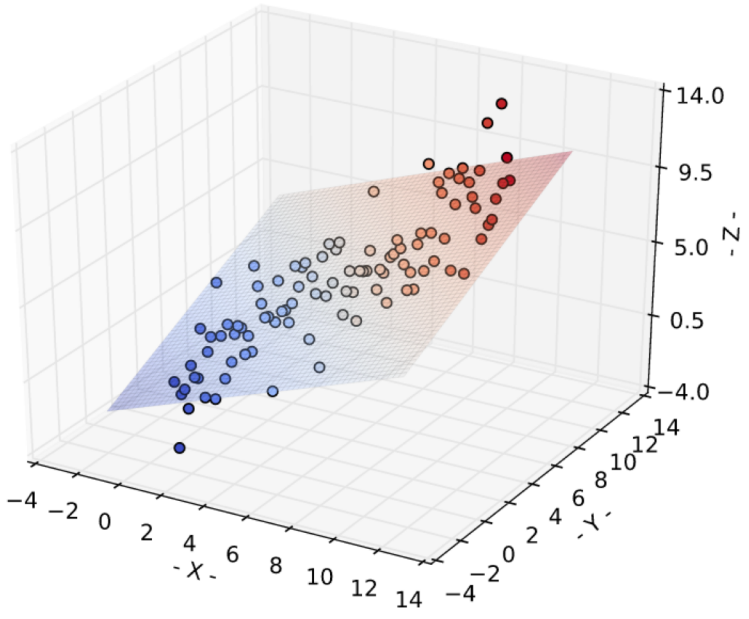
y = 2.81x + 3.05

HORTONWORKS

Linear Regression (two features)

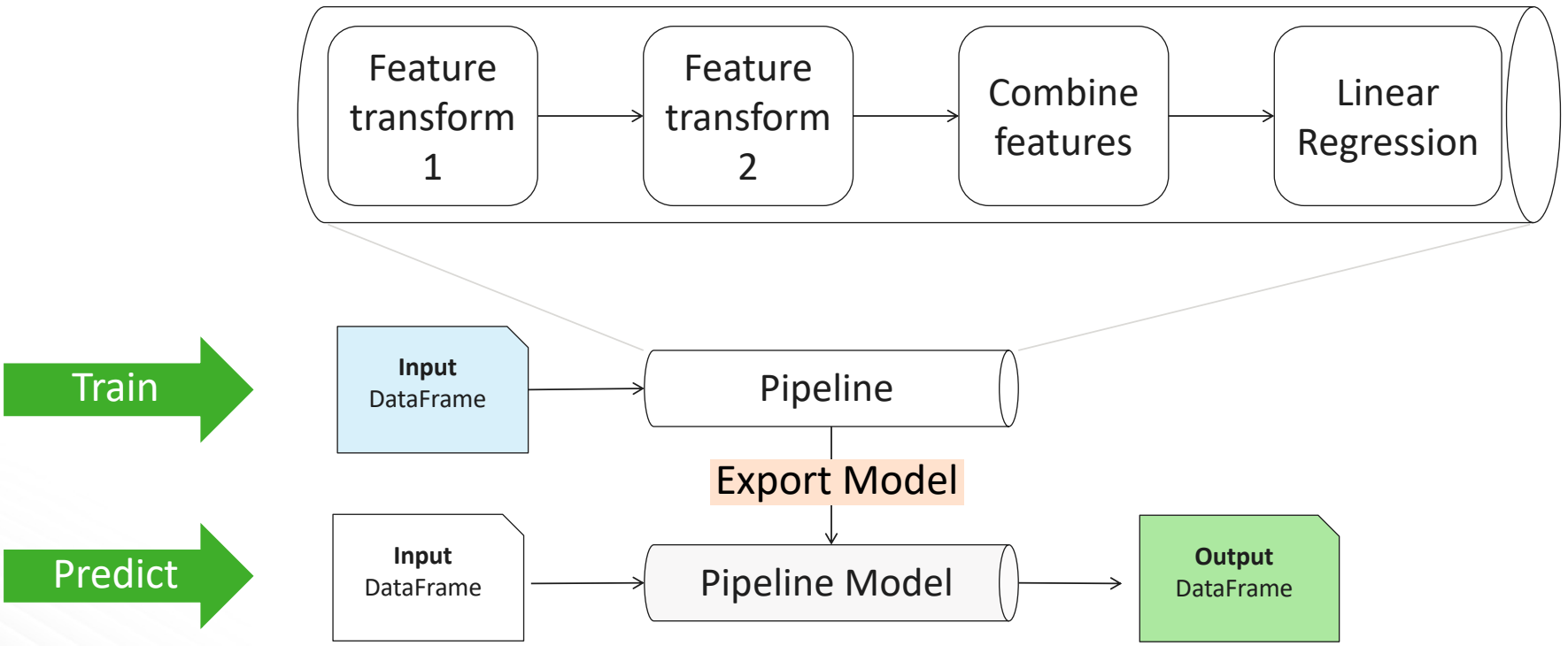
```
Coefficients: [0.464, 0.464]
Intercept: 0.0563
```





HORTONWORKS

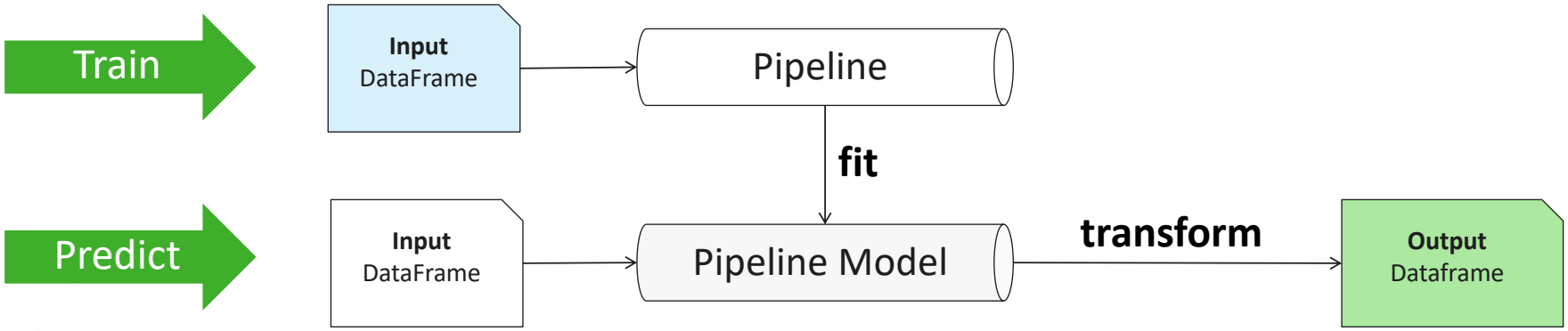
SPARK ML PIPELINES



HORTONWORKS

Spark ML Pipeline

- **fit()** is for **training**
- **transform()** is for **prediction**



HORTONWORKS

SAMPLE CODE

```
rf = RandomForestClassifier(numTrees=100)
pipe = Pipeline(stages=[indexer, parser, hashingTF, vecAssembler, rf])

model = pipe.fit(trainData)                # Train model
results = model.transform(testData)        # Test model
```

DSX + HDP

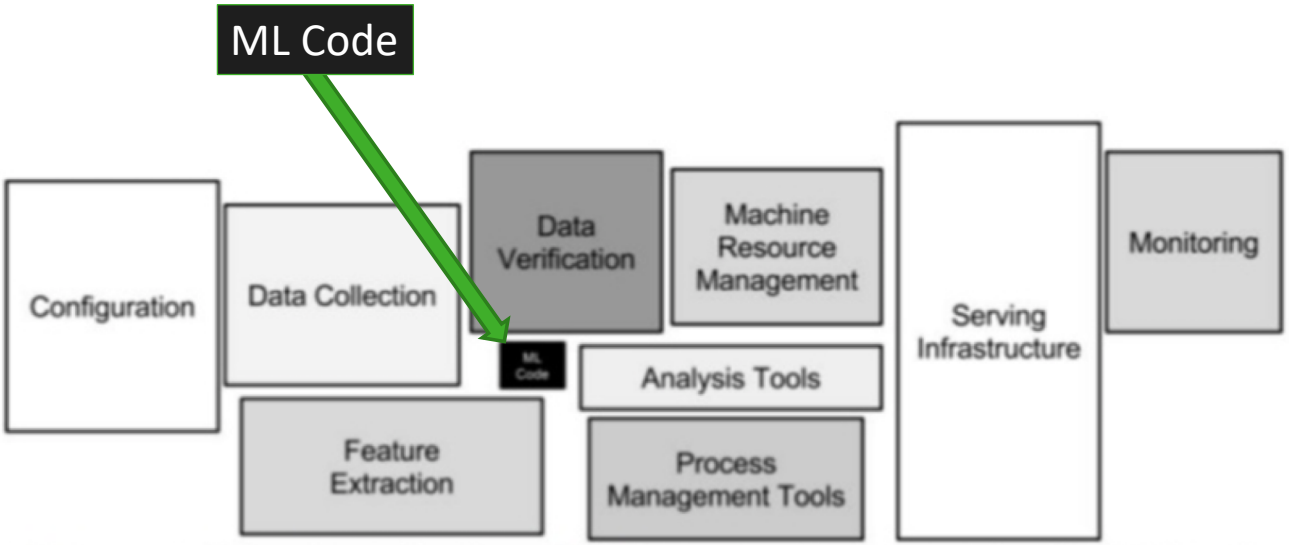
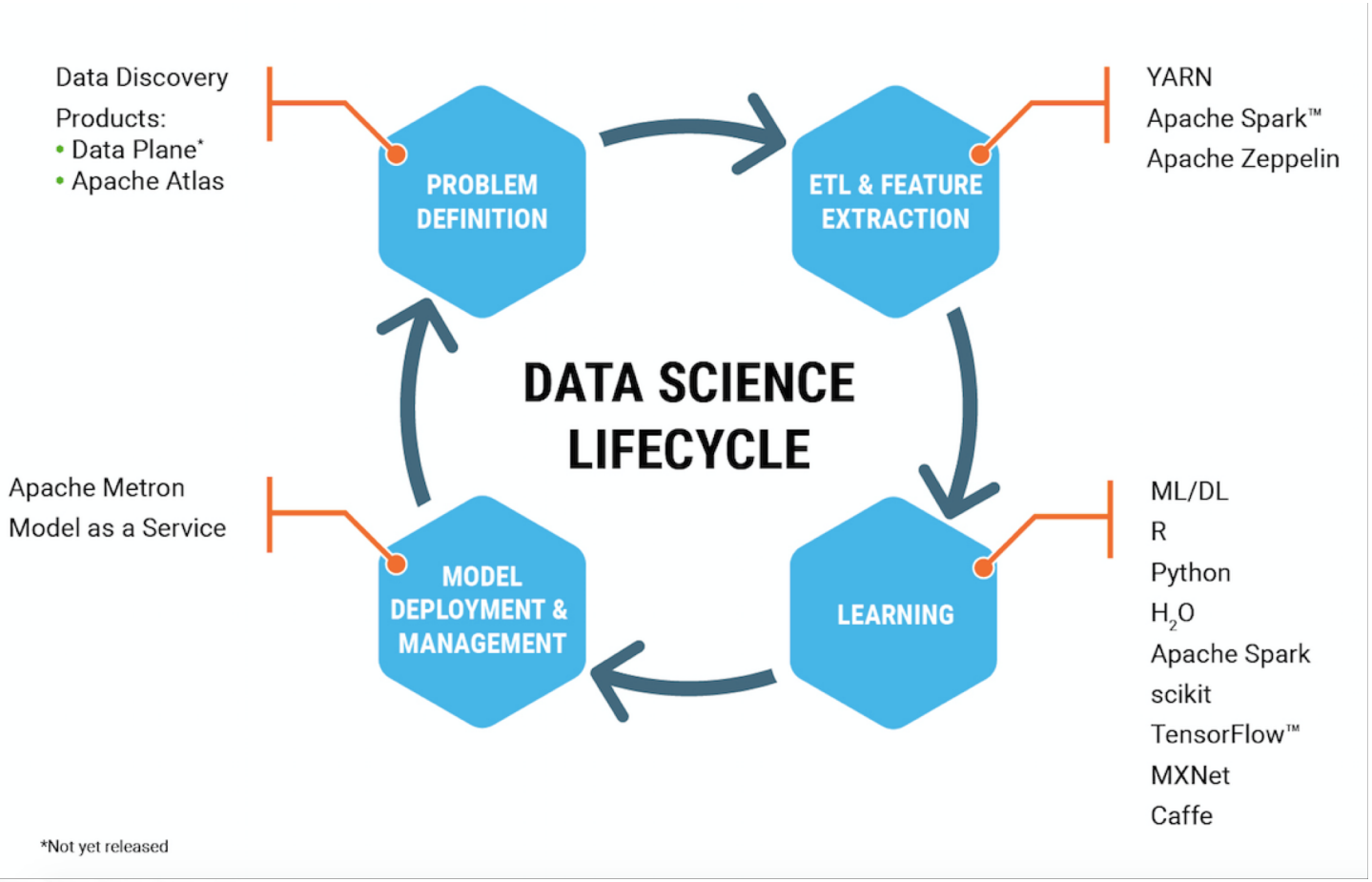


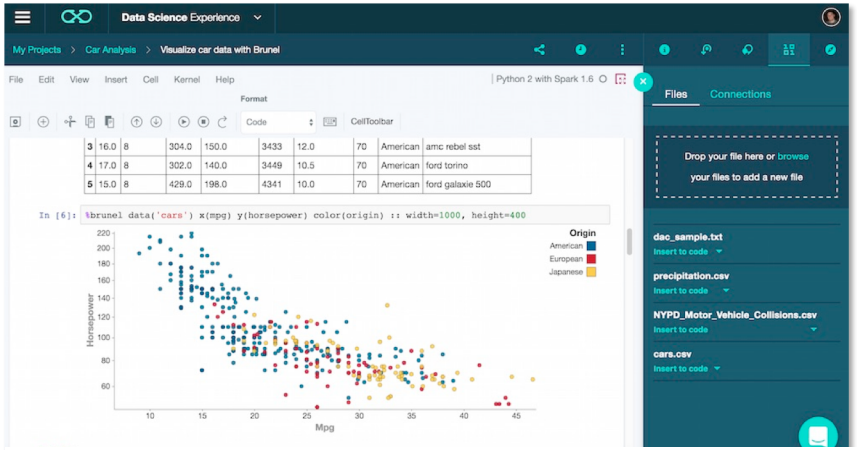
Figure 1: Only a small fraction of real-world ML systems is composed of the ML code, as shown by the small black box in the middle. The required surrounding infrastructure is vast and complex.

Source: Google NIPS



HORTONWORKS

Data Science Experience



SPSS Modeler for DSX

DO for DSX

DATA SCIENCE PLATFORM

Community

- Find tutorials and datasets
- Connect with Data Scientists
- Ask questions
- Read articles and papers
- Fork and share projects

Open Source

- Code in Scala/Python/R/SQL
- Zeppelin & Jupyter Notebooks
- RStudio IDE and Shiny
- Apache Spark
- Your favorite libraries

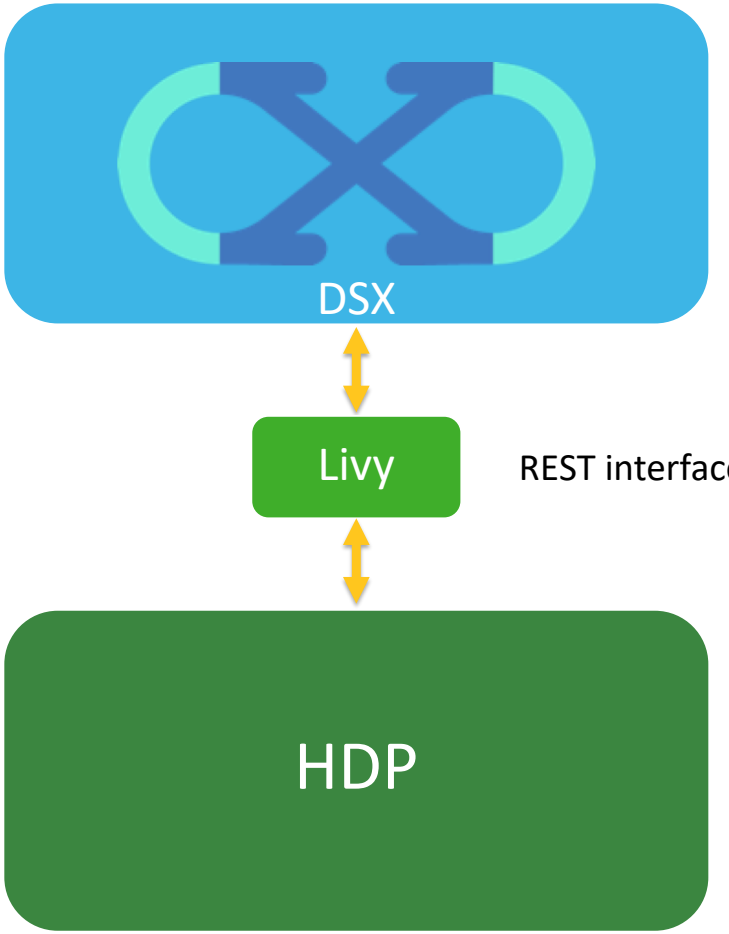
Model Management

- Data Shaping Pipeline UI
- Auto-data preparation & modeling
- Advanced Visualizations
- Model management & deployment
- Documented Model APIs

Scale & Enterprise Security

- Data Science at Scale
- Run Spark Jobs on HDP Cluster
- Secure Hadoop Support
- Ranger Atlas Support for Data
- Support for ABAC

HORTONWORKS



Data Science Experience (DSX) Local
Enterprise Data Science platform for teams

Hortonworks Data Platform (HDP)
Enterprise compute (Spark/Hive) & storage (HDFS/Ozone)

HORTONWORKS

FINAL NOTES

MIT Sloan Management Review

"The business **value of AI** consists of its **ability to lower the cost of prediction**, just as computers lowered the cost of arithmetic."

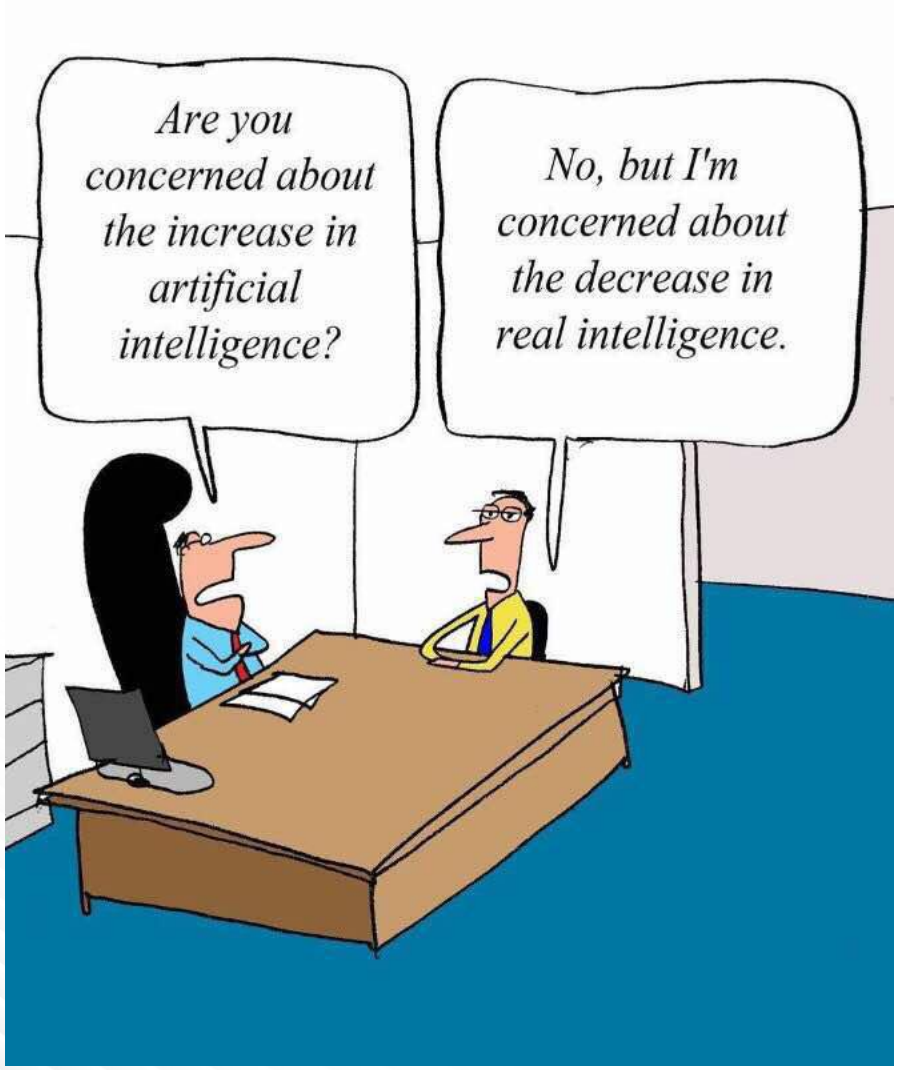
HORTONWORKS

Building a Business Around AI - HBR

1. Find and own valuable data no one else has
2. Take a systemic view of your business, and find data adjacencies
3. Package AI for the customer experience

"No single tool, even one as powerful as AI, determines the fate of a business. As much as the world changes, deep truths — around unearthing customer knowledge, capturing scarce goods, and finding profitable adjacencies — will matter greatly. As ever, the **technology works to the extent that its owners know what it can do, and know their market**."

HORTONWORKS



Thanks!

Robert Hryniewicz
@RobHryniewicz

HORTONWORKS