Getting (a bit) familiar with Data Science

JHUG October 2018

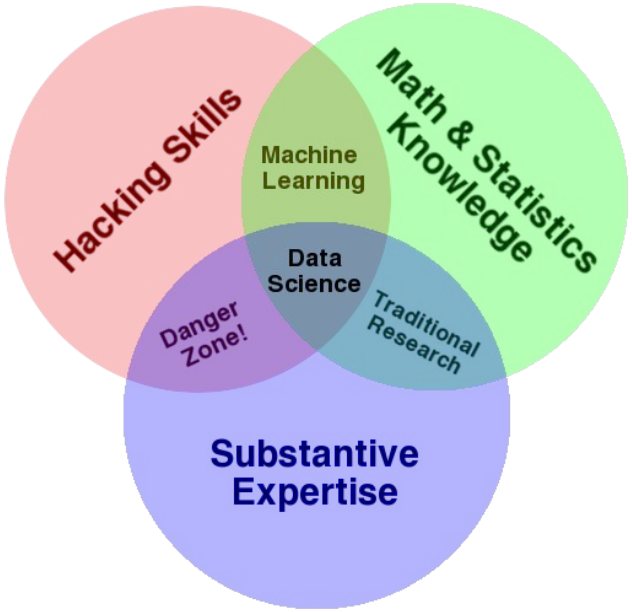
Rule-based systems

- If X then Y else if P then Q …
- Start with 100 scenarios, write 100 rules
- Exceptions kick in
- More rules
- Rule management

What is data science?

- Problem formulation
- Collect & Process Data
- Machine Learning
- Insights & action

The DS Venn Diagram



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

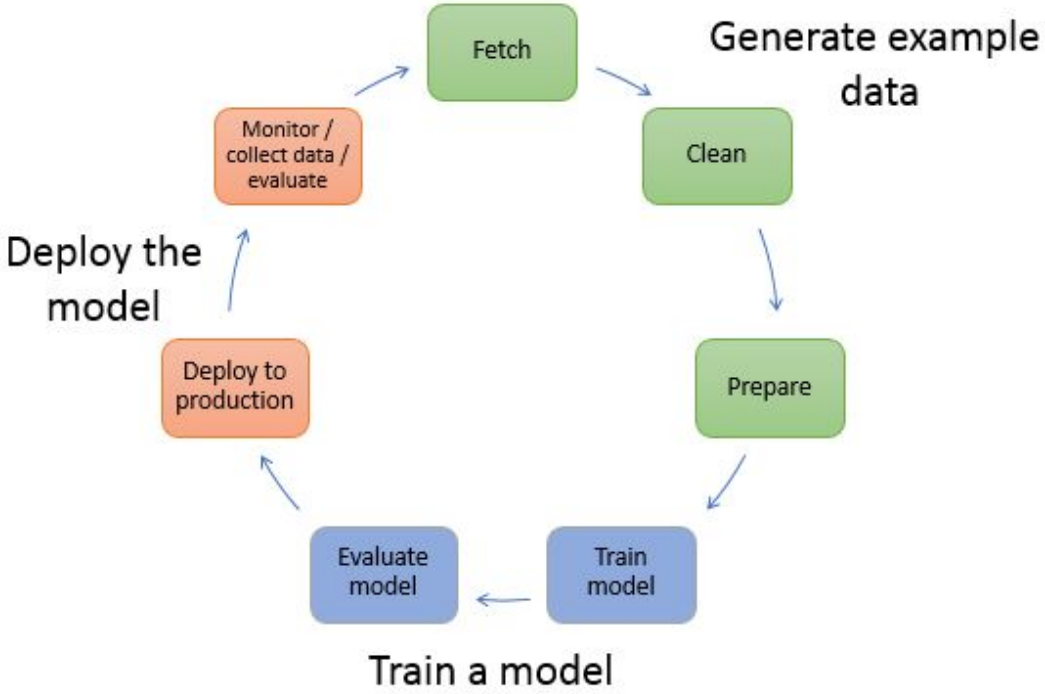
Data Scientist skill set

- Software engineering
  - Algorithms, R, Python, databases
- Business acumen
- Distributed Computing
  - MapReduce, Hadoop, Spark, Pig, AWS
- Communication
  - Senior management, storytelling, visualization
- Machine Learning
  - Statistical modeling, experiment design, algorithm
- Statistics
- Domain Knowledge
  - Curiosity, problem solver

Machine learning

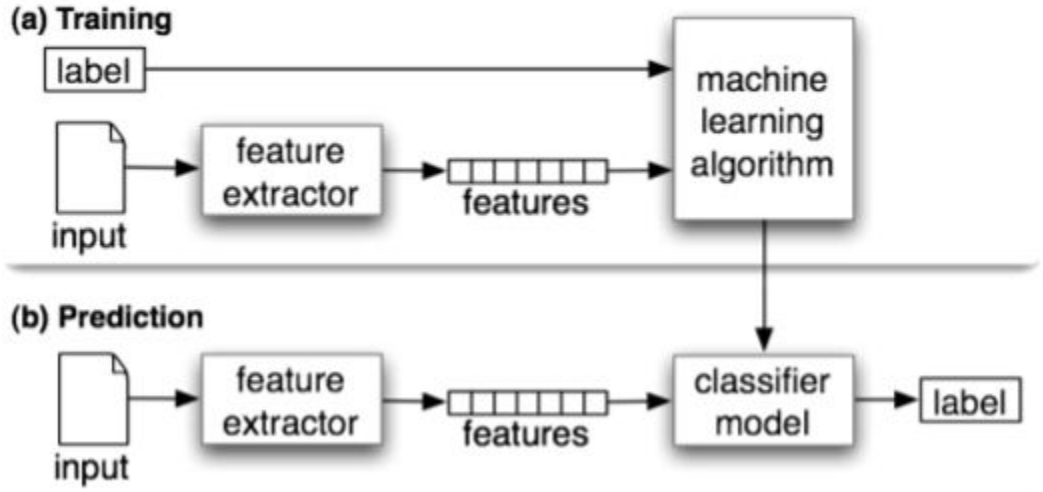
The ability of an AI system to acquire its own knowledge by extracting patterns from raw data.

Machine Learning Workflow



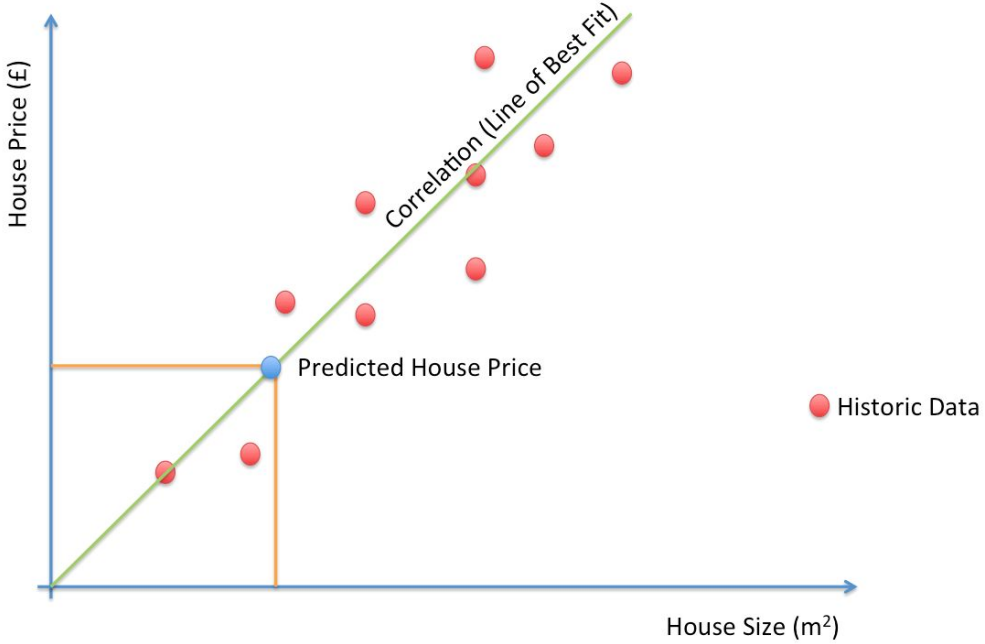
https://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-mlconcepts.html

Machine Learning Workflow



Source

Example: House Price prediction



$$\mathbf{a} \cdot x + \mathbf{b} \cdot y + \mathbf{c} = 0$$

Feature engineering

- Number of rooms
- Year built
- Year of last renovation
- Location
- Number of rooms, ...
- Family size

Example: Spam detection

- Problem: Decide if an email is spam or ham
- Dataset: ENRON (1st file only)
  - 1501 spam
  - 3673 ham
- Source

Data - Sample

Subject: re [8] : dear friend -
size = 1 > order confirmation . your order should be shipped by january , via fedex .
your federal express tracking number is 45954036 .
thank you for registering . your userid is : 56075519
learn to make a fortune with ebay !
complete turnkey system software - videos - turorials
clk here for information
clilings .

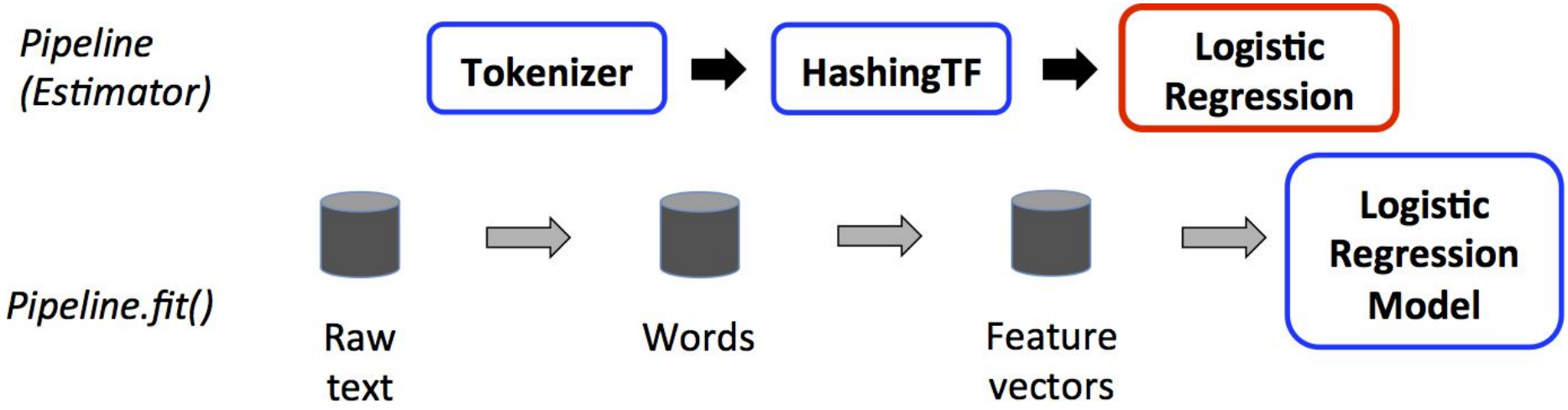
Step 1: Load data

```
JavaRDD<Email> spam = sc.textFile("data/spam/*.txt").map(x -> new Email(x, 1));
JavaRDD<Email> ham = sc.textFile("data/ham/*.txt").map(x -> new Email(x, 0));
JavaRDD<Email> emails = spam.union(ham);

Dataset<Email> dataset = this.spark.createDataset(emails.rdd(),
Encoders.bean(Email.class));
// Split to train-test dataset (80%/20%)
Dataset<Email>[] datasets = dataset.randomSplit(new double[]{0.8, 0.2});
```

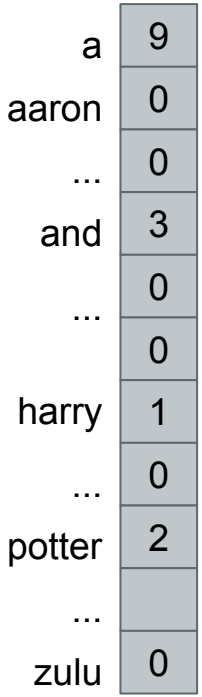
Step 2: Create the pipeline & train

```
Tokenizer tokenizer = new Tokenizer().setInputCol("body").setOutputCol("words");
HashingTF tf = new HashingTF().setInputCol(tokenizer.getOutputCol()).setOutputCol("features");
LogisticRegression lr = new LogisticRegression().setMaxIter(10).setRegParam(0.01);
Pipeline pipeline = new Pipeline().setStages(new PipelineStage[] {tokenizer, tf, lr});

PipelineModel model = pipeline.fit(datasets[0]);
```

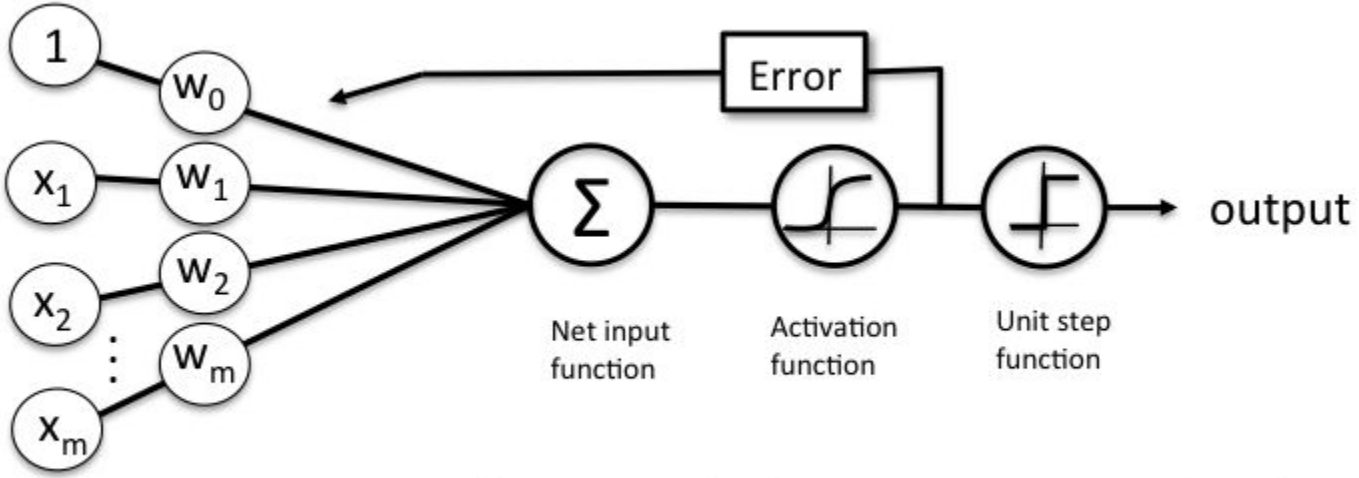
What just happened?



Feature vector



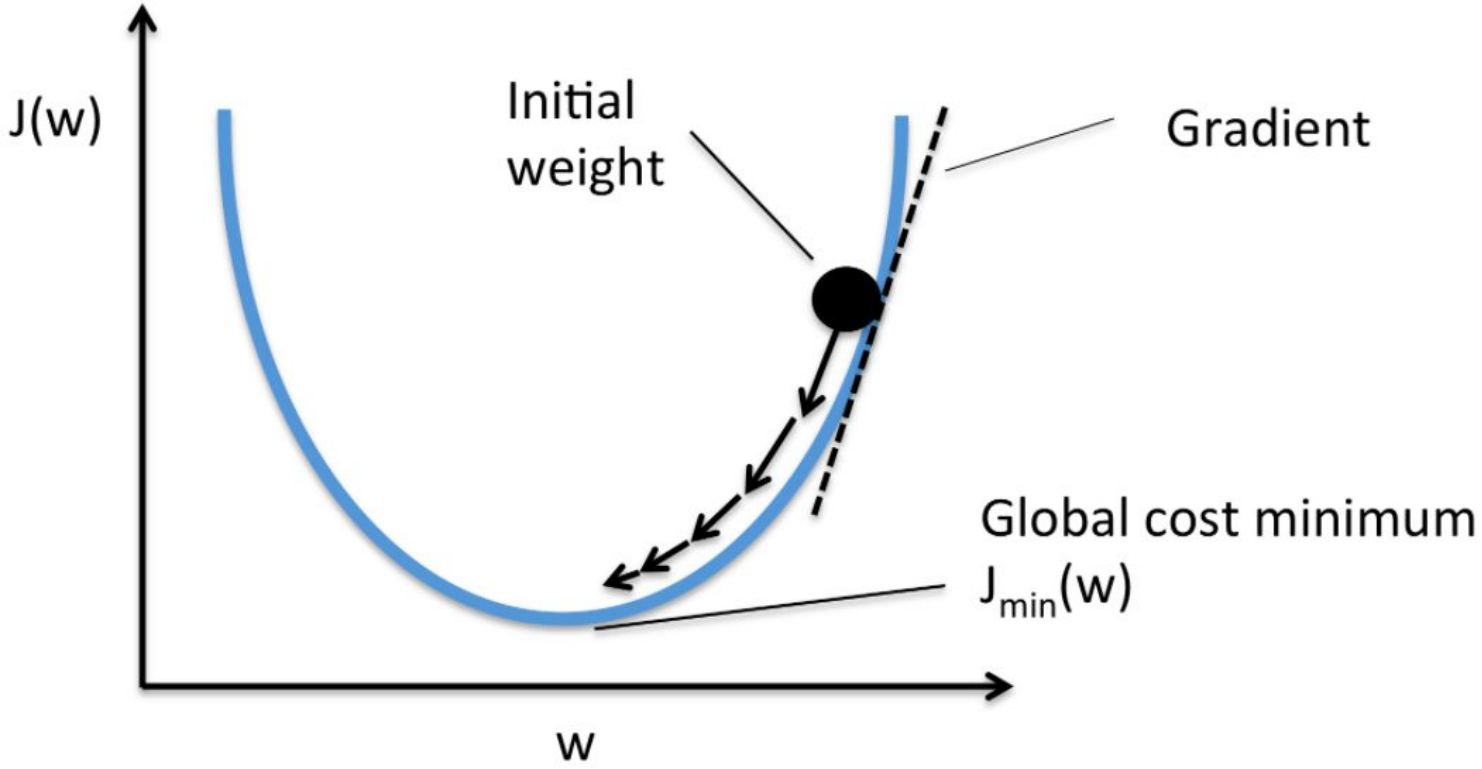
Logistic Regression



Schematic of a logistic regression classifier.

Source

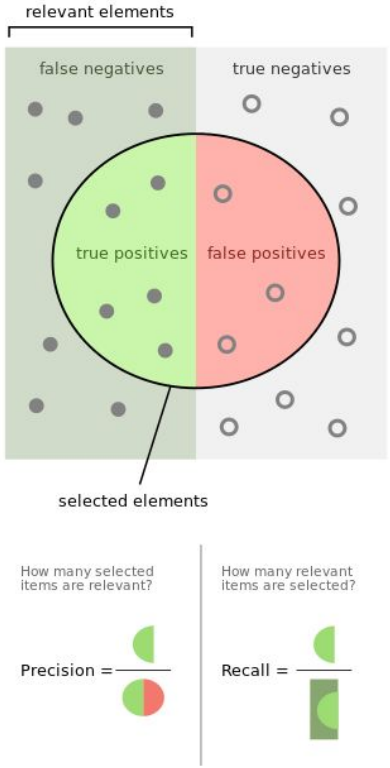
Gradient descent



Step 3: Evaluate

```
Dataset<Row> prediction = model.transform(datasets[1]);

MulticlassClassificationEvaluator eval = new MulticlassClassificationEvaluator()
                .setLabelCol("label")
                .setPredictionCol("prediction")
                .setMetricName("accuracy");

System.out.println(String.format("Accuracy = %s", eval.evaluate(prediction)));
```

Evaluating ML algorithms (example)

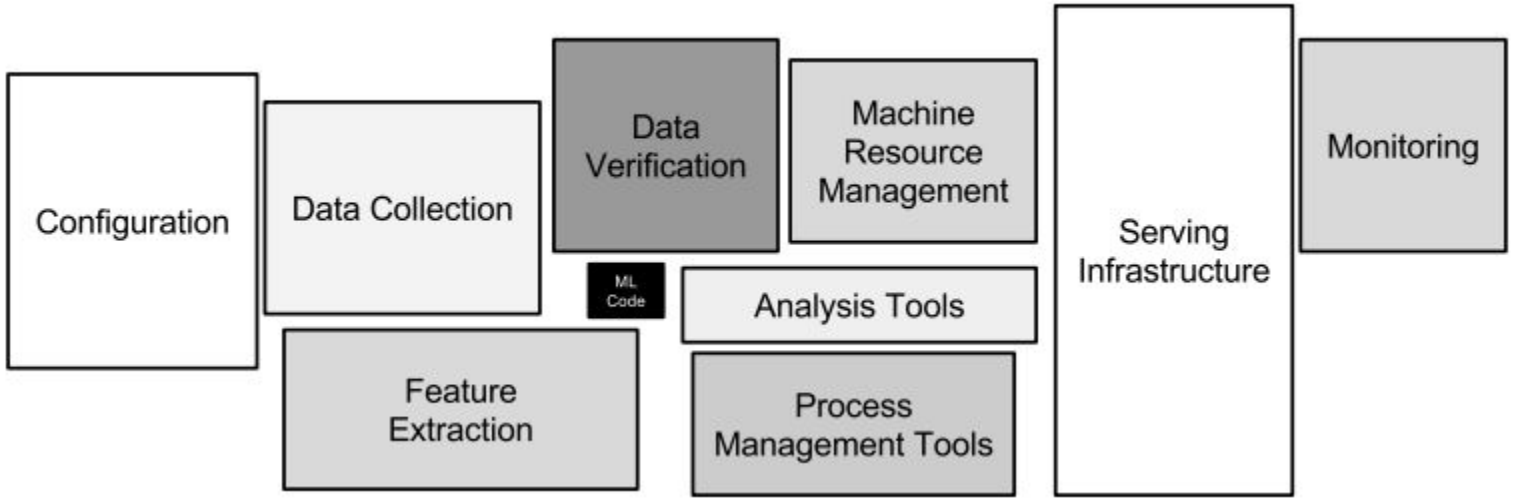


Source

Results

- Accuracy = 0.9101056408212079
- Precision = 0.9105886930557663
- Recall = 0.9101056408212078

Building an ML system



https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf

Kinds of ML

- Supervised learning
- Unsupervised Learning

Supervised Learning

- Train a model on a pre-defined dataset
- Accurately predict label on new data based on previous observations

- Applications:
  - Classification
  - Regression

- Examples
  - Classifying Twitter sentiments
  - Recommender systems

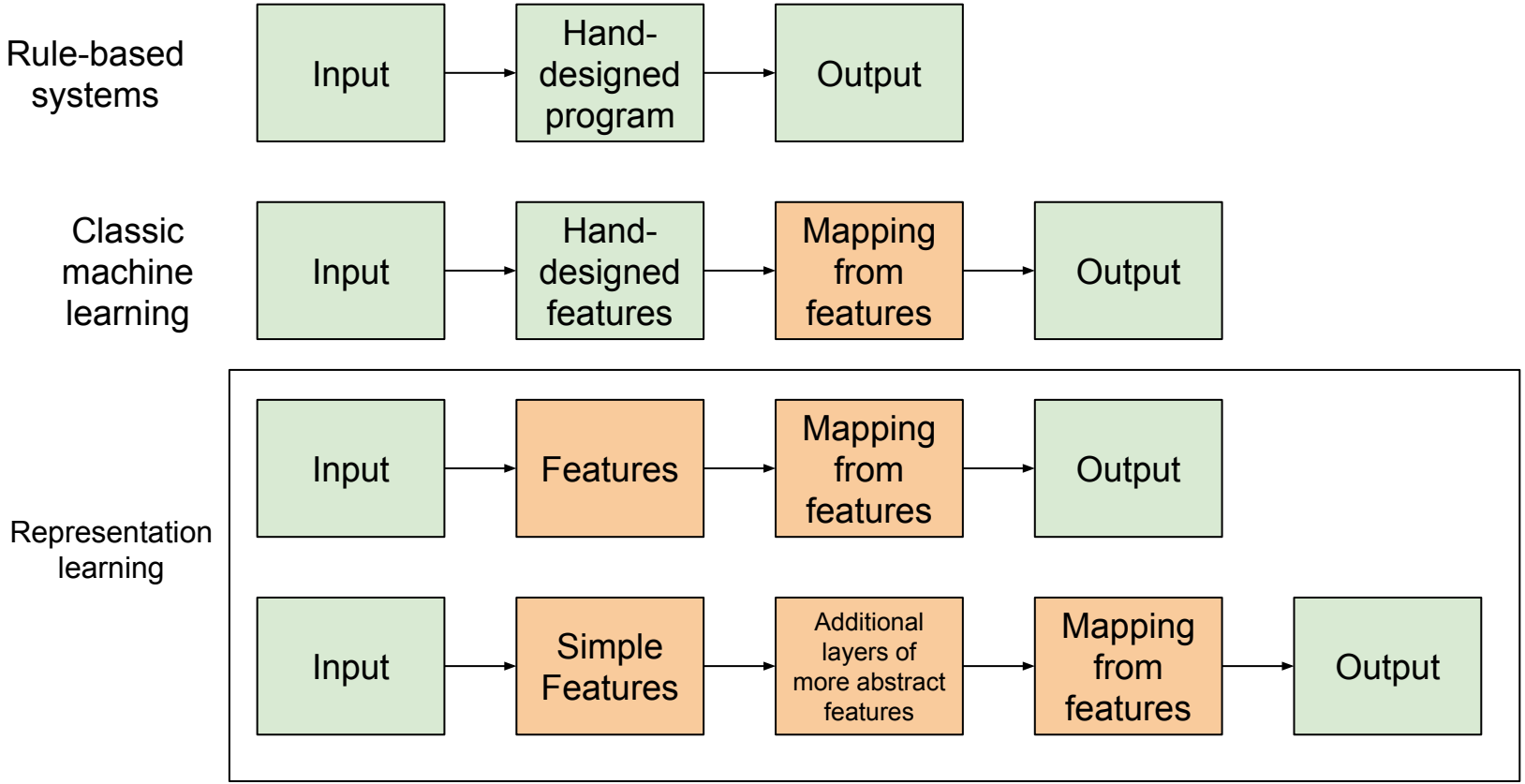
Unsupervised Learning

- Given a dataset, find patterns and relationships
- Accurately predict label on new data based on previous observations

- Applications:
  - Association
  - Clustering

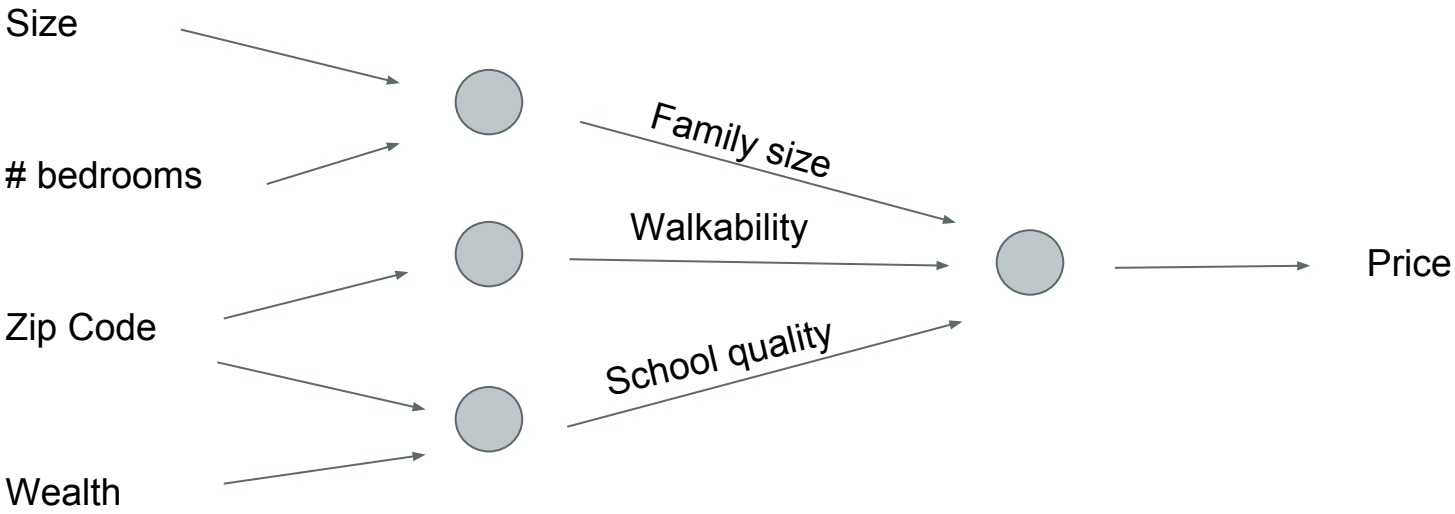
- Examples:
  - Customer segmentation
  - Similar items (autocomplete)

Different AI disciplines

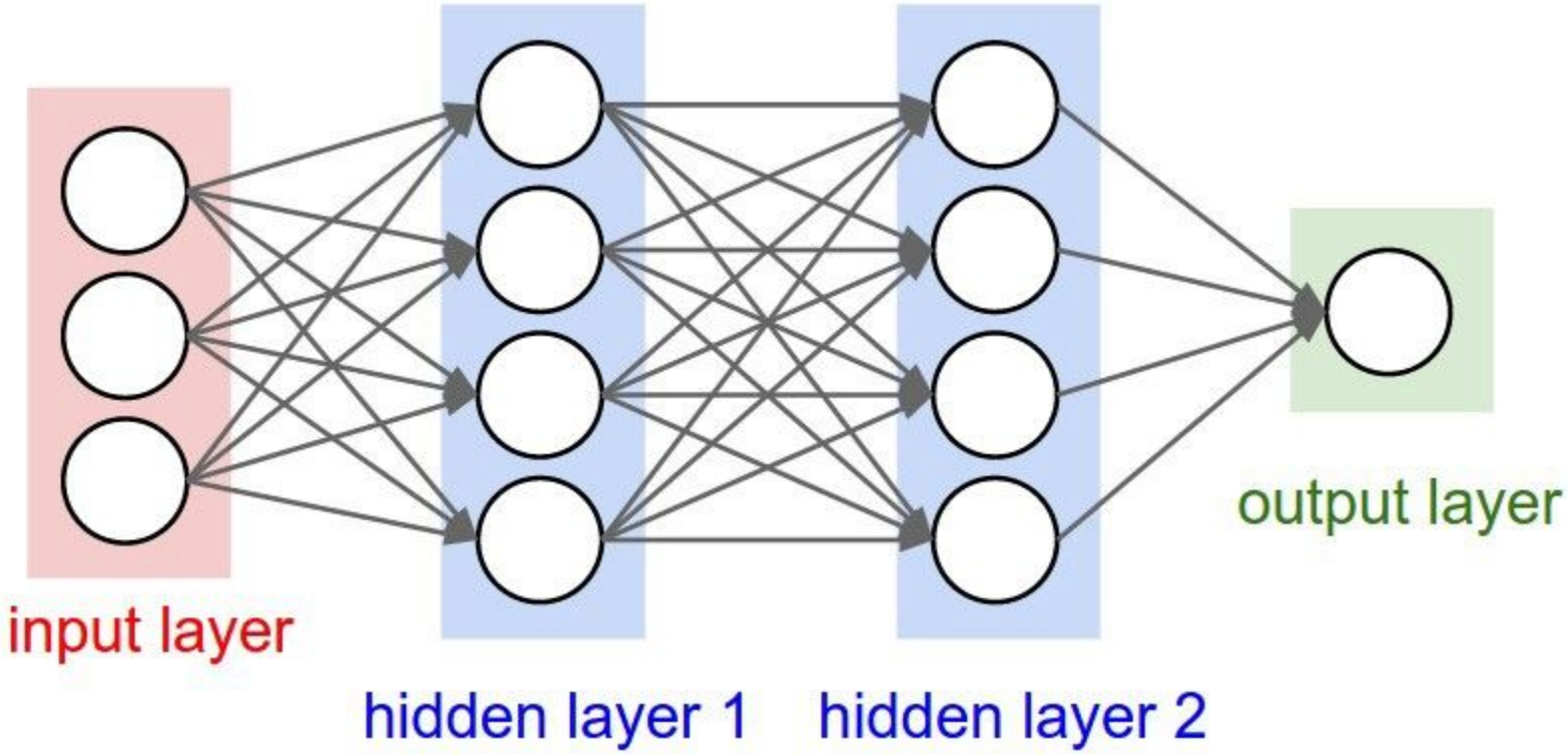


Source

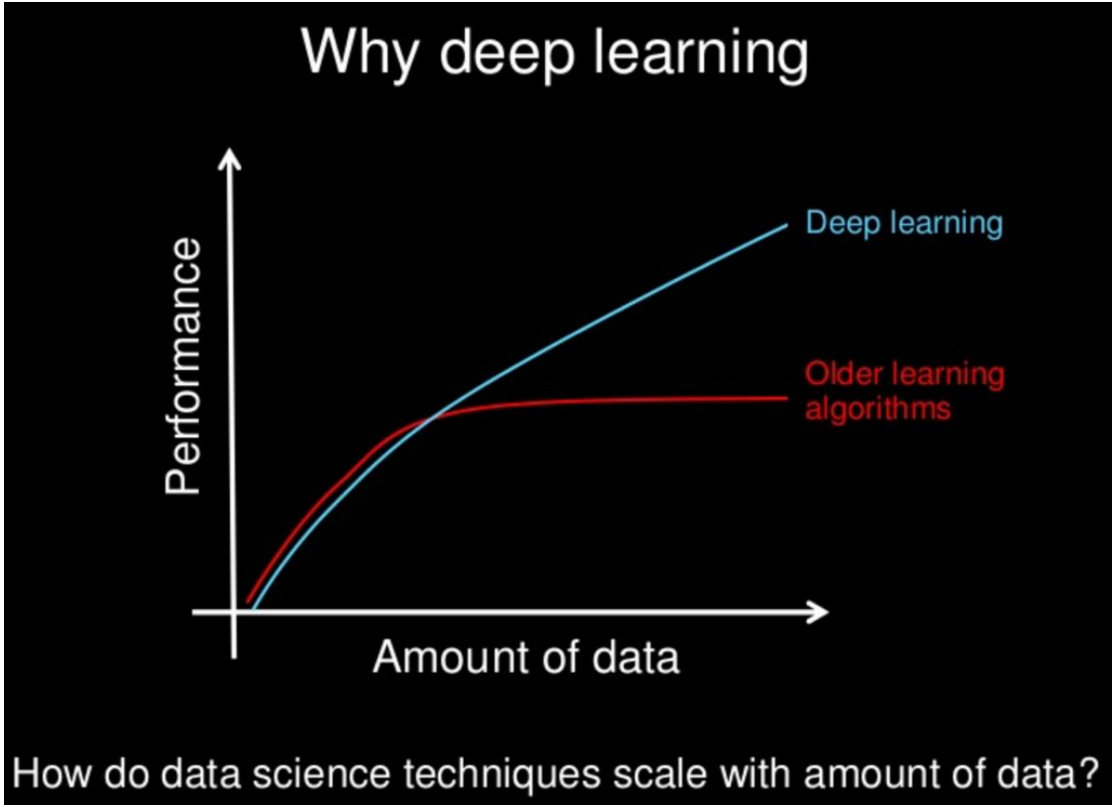
More complex features



Neural Network



Why Deep Learning?



Resources

- Web
  - KDnuggets
  - Towards Data Science
- Courses
  - Machine learning (Coursera)
  - Deep Learning
  - fast.ai
- Mailing lists
  - Data Machina
- Conferences
  - NIPS
  - ICML
  - KDD

Come join us!