LINKTERA

Data Analytics

Topics

Topics to be covered in the workshop:

Section 1:

- ❑ Data EveryWhere!
- ❑ The 4V's of Big Data
- ❑ DIKW Hierarchy
- ❑ Structured Data vs Unstructured Data
- ❑ Data Preprocessing
- ❑ CRISP-DM
- ❑ Data Visualization
- ❑ Data Science Venn Diagram
- ❑ Model Basics & Machine Learning
- ❑ Why Use Machine Learning?
- ❑ Types of Machine Learning Systems
- ❑ Machine Learning Software
- ❑ Market Basket Analysis

Section 2:

- ❑ Pigeons as Art Experts
- ❑ From Biological to Artificial Neurons
- ❑ The Neural Network
- ❑ Deep Neural Network
- ❑ Deep Neural Network-Common Architectures
- ❑ Tensorflow&Keras
- ❑ Time Series Analysis with LSTM
- ❑ Further Readings

Data EveryWhere!

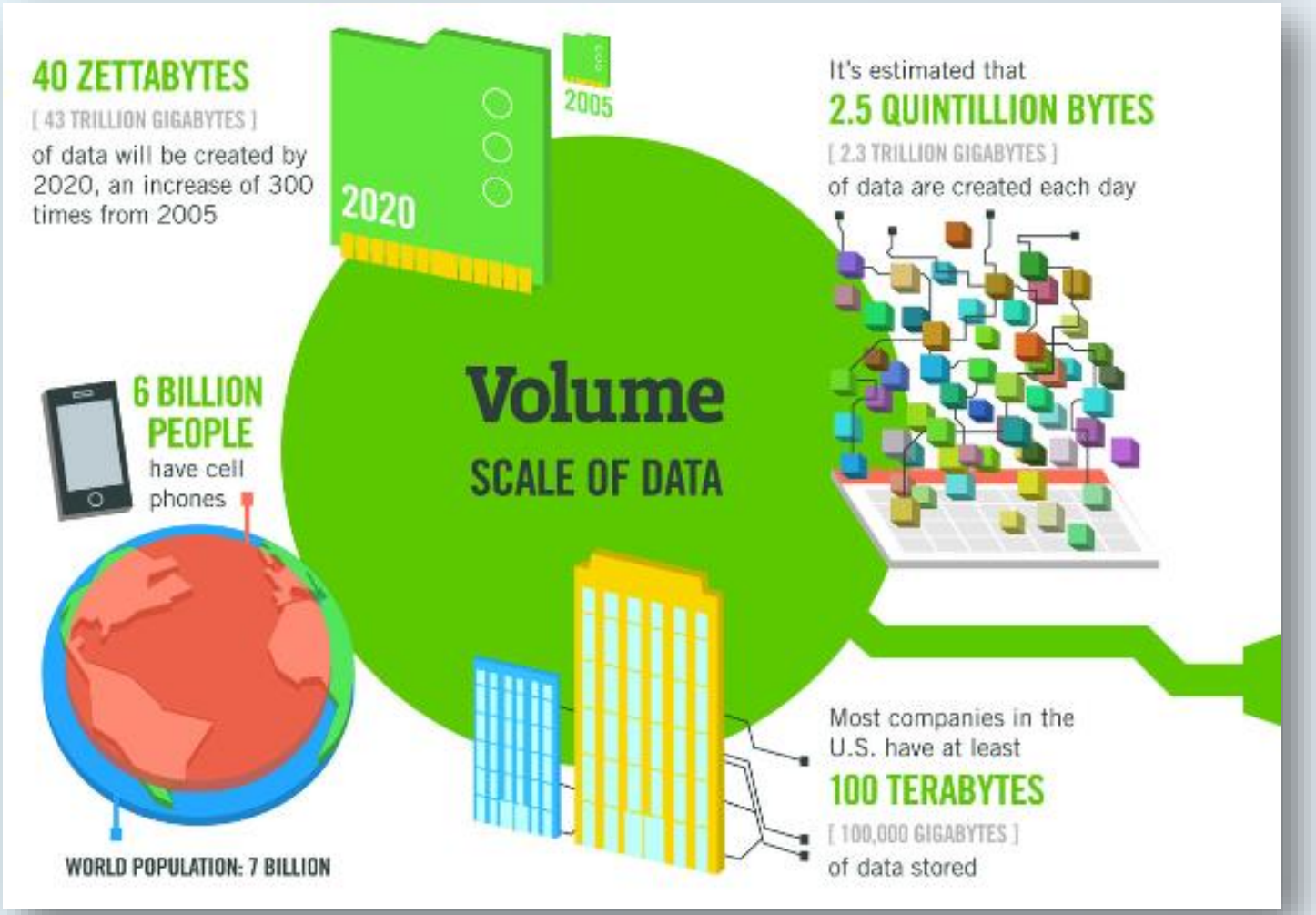
Lots of data is being collected and warehoused

- Web data, e-commerce
- Purchases at department / grocery stores
- Bank / Credit Card transactions
- Social Network

LINKTERA

The 4V's of Big Data

- Today it would take a person approximately 181 million years to download all the data from the internet. *(Source: Physics.org)*
- In 2012, only 0.5% of all data was analyzed. *(Source: The Guardian)*
- By 2020, every person will generate 1.7 megabytes in just a second. *(Source: Domo)*

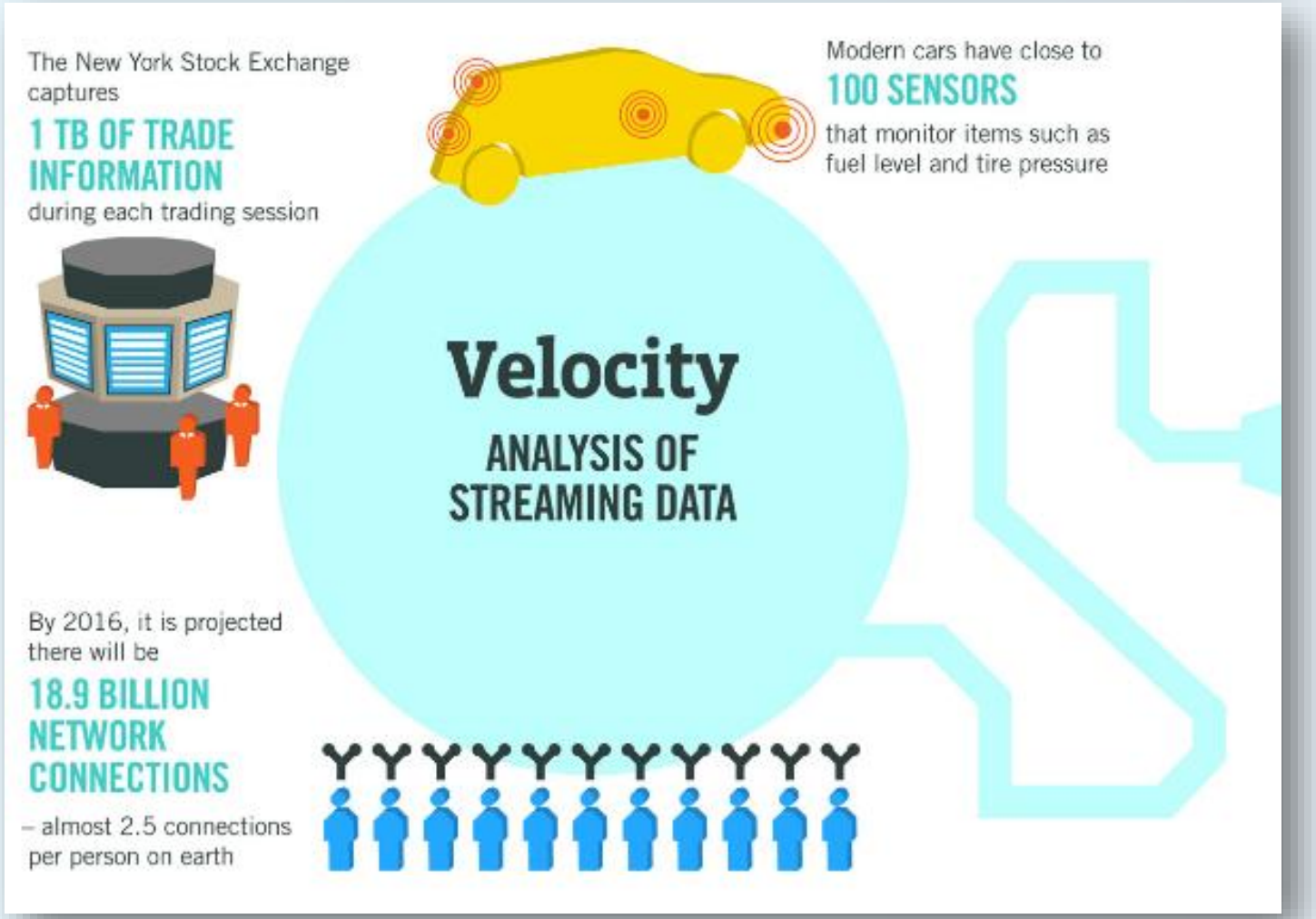


Source: https://techjury.net/stats-about/big-data-statistics/

LINKTERA

The 4V's of Big Data

- 90% of all data has been created in the last two years. *(Source: IBM)*
- Twitter users send nearly half a million tweets every minute. *(Source: Domo)*

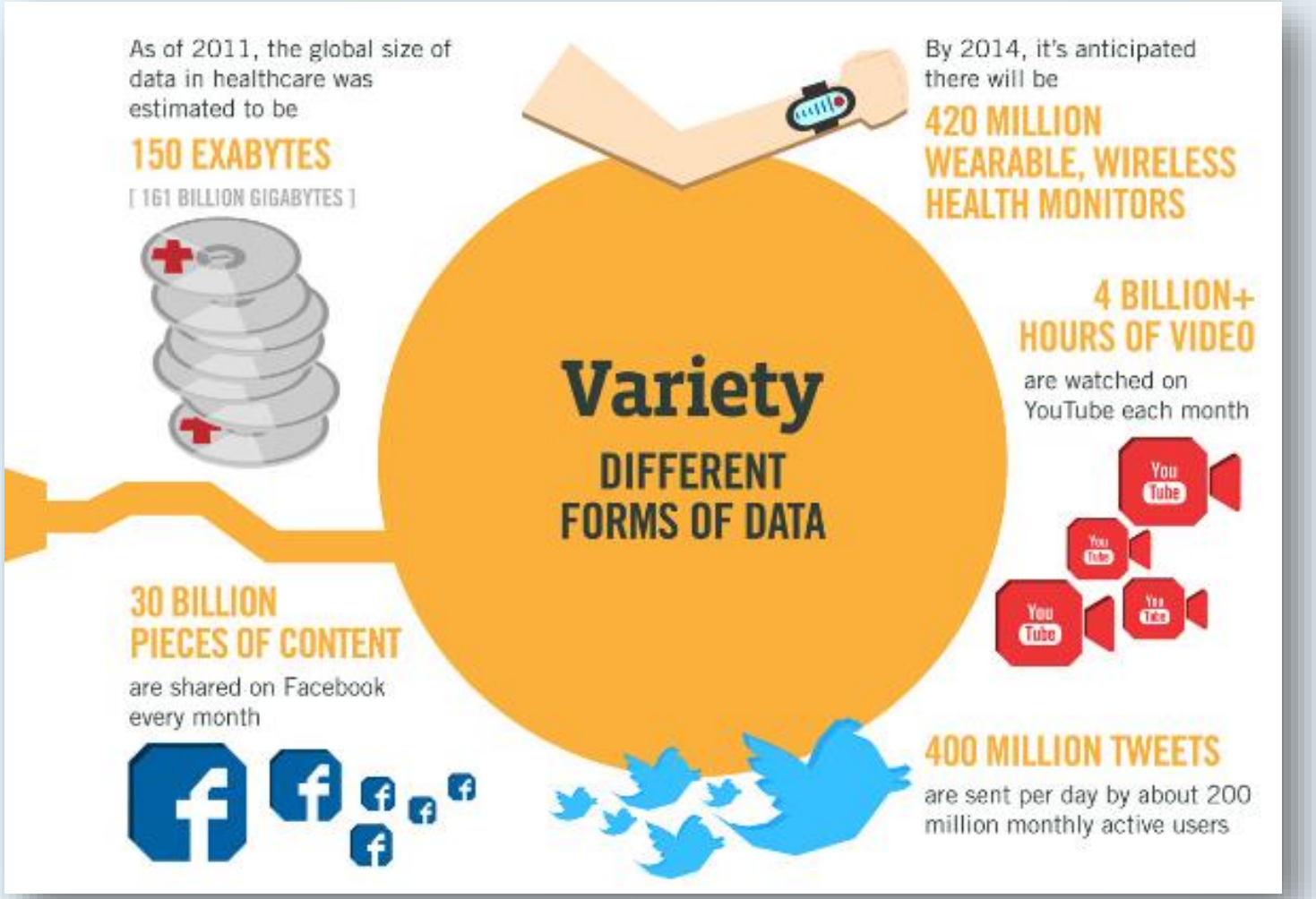


Source: https://techjury.net/stats-about/big-data-statistics/

LINKTERA

The 4V's of Big Data

- 30 billion pieces of content are shared on Facebook every month. *(Source: techjury.org)*
- By 2014, it's anticipated there will be 420 million wearable, wireless health monitors. *(Source: The Guardian)*
- By 2020, every person will generate 1.7 megabytes in just a second. *(Source: Domo)*

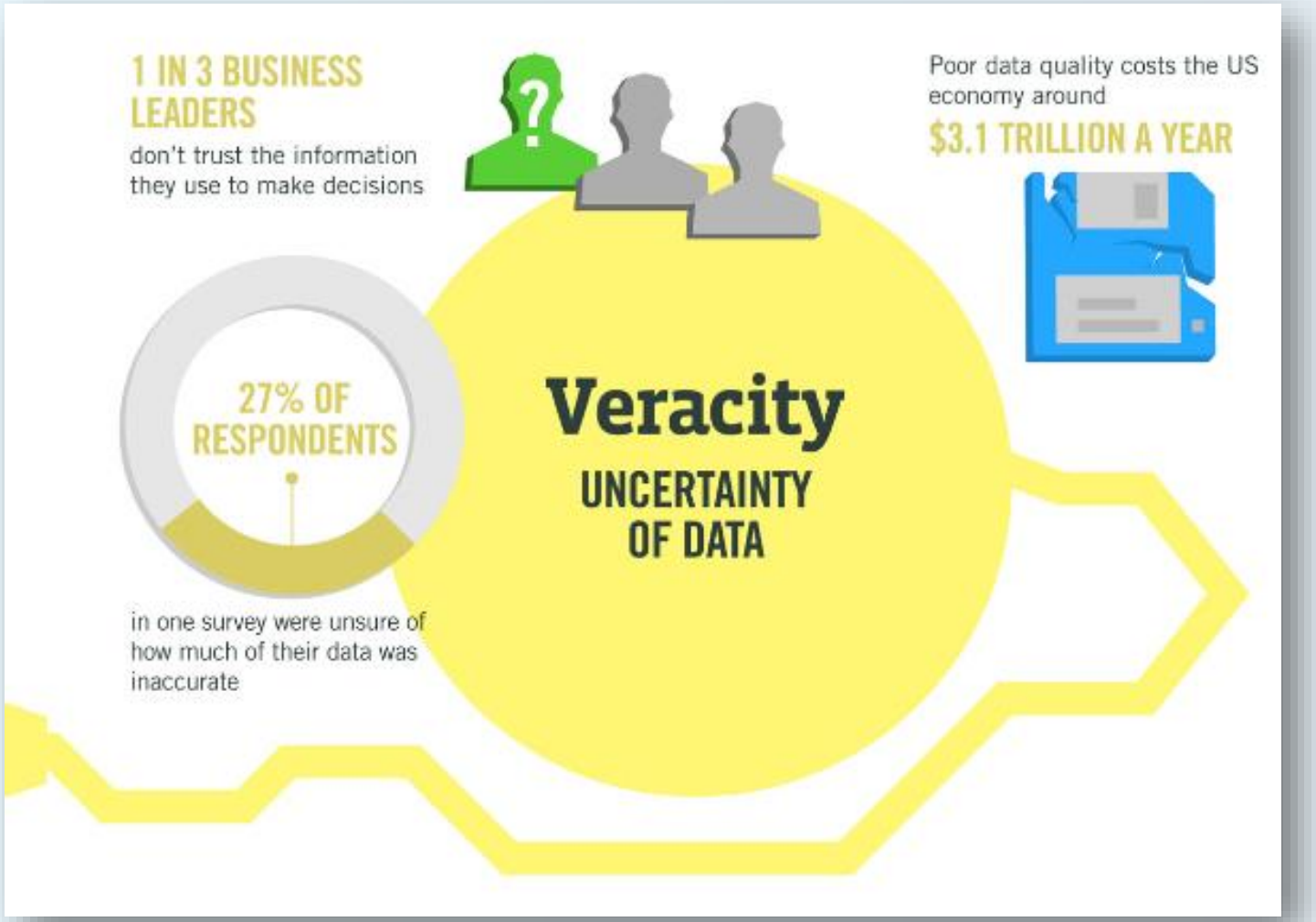


Source: https://techjury.net/stats-about/big-data-statistics/

LINKTERA

The 4V's of Big Data

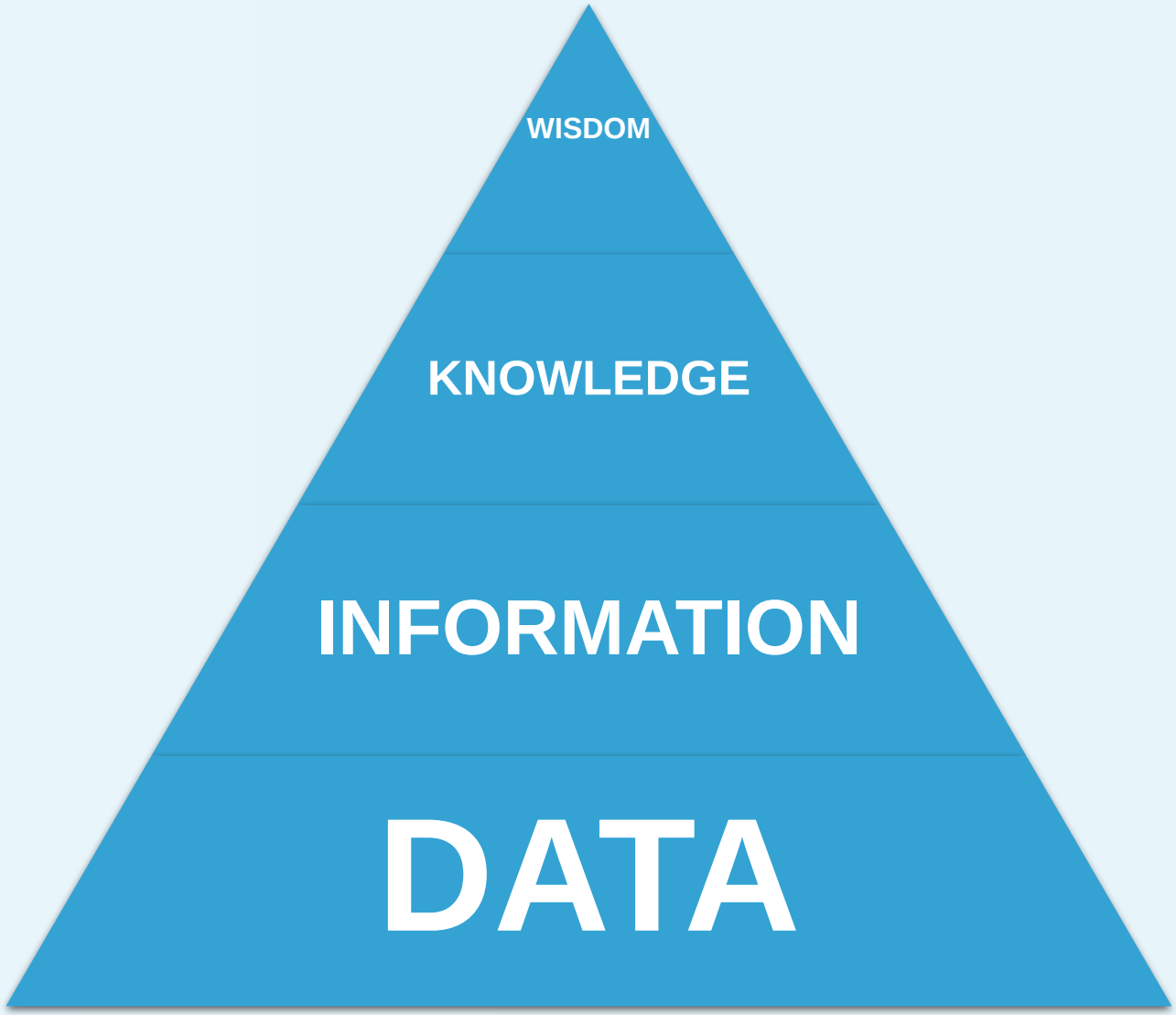
- %25 of US companies data is believed to be inaccurate. *(Source: The State of data quality. Experian,2013)*
- On average, poor data quality is estimated to cost of organizations a yearly sum of $3.1 Trillion a year. *(Source: The Guardian)*



Source: https://techjury.net/stats-about/big-data-statistics/

LINKTERA

DIKW Hierarchy



APPLIED

understanding, integrated, actionable

CONTEXT

contextual, synthesized, learning

MEANING

useful, organized, structured

RAW

signals, know nothing

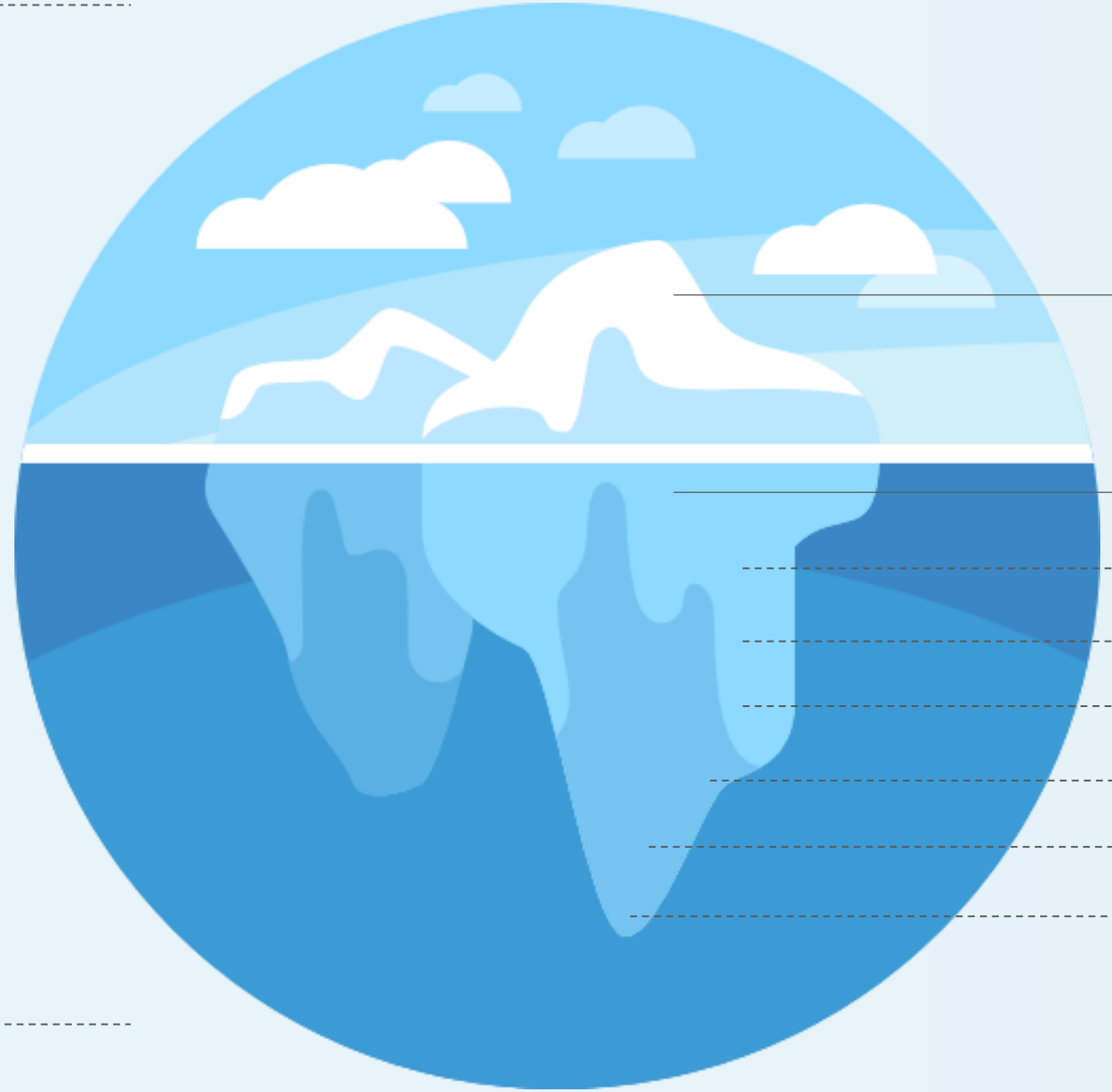
Source: Rowley, J. (2007). The wisdom hierarchy representations of the DIKW hierarchy. Journal of information science, 33(2), 163-180.

LINKTERA

STRUCTURED DATA vs UNSTRUCTERED DATA

%20

%80



structured data

unstructured data

- PDFS
- WORD DOCUMENTS
- SPREADSHEETS
- PRESENTATIONS
- SOCIAL MEDIA POSTS
- BOOKS

LINKTERA

Data Preprocessing

Although data increases exponentially, clean data does not increase at the same rate.

Why data processing important?

Data in the real world is dirty

Why Is Data Preprocessing Important?

No quality data, no quality mining results!

Garbage in garbage out!

Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse.

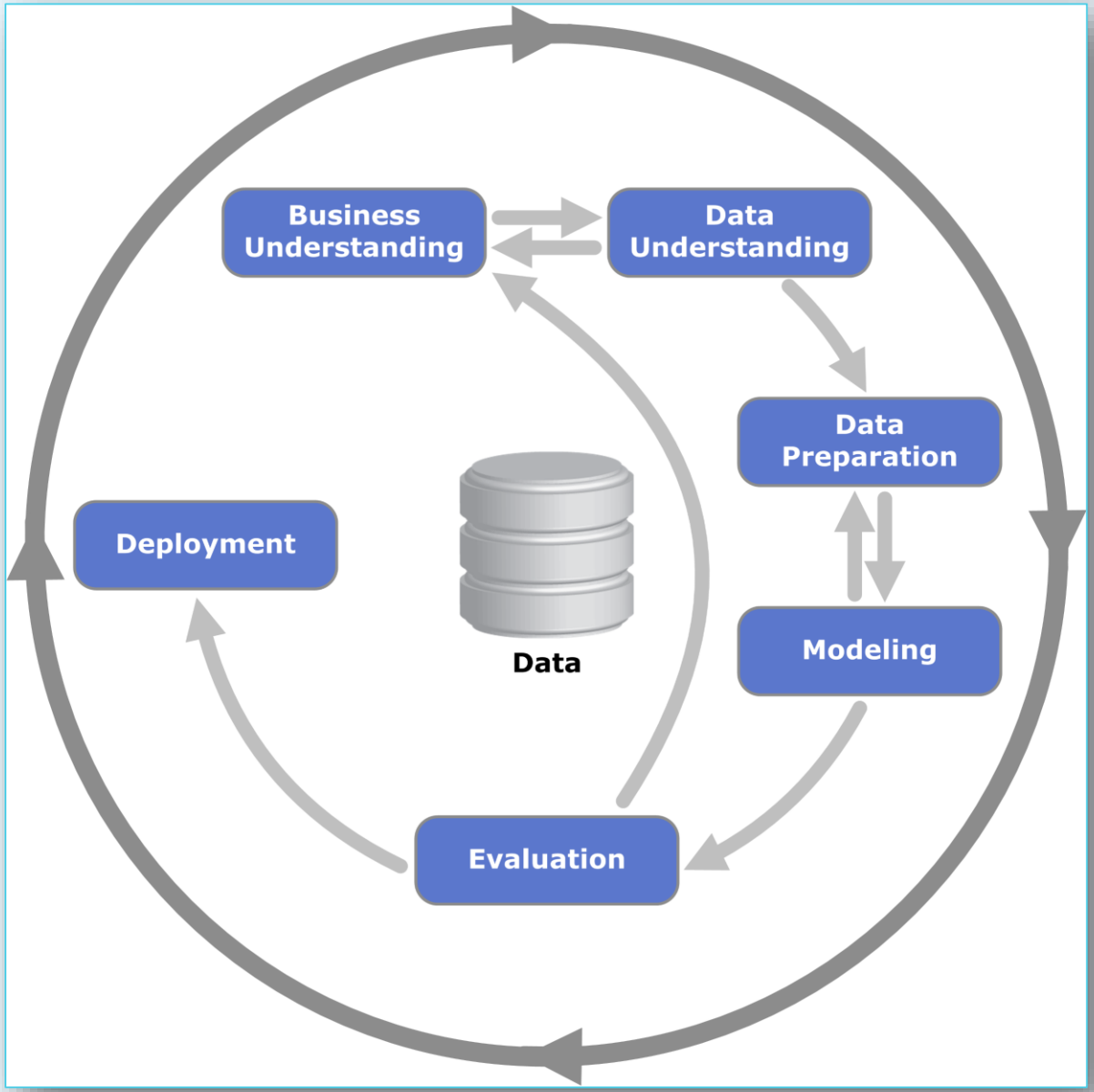
LINKTERA

CRISP-DM

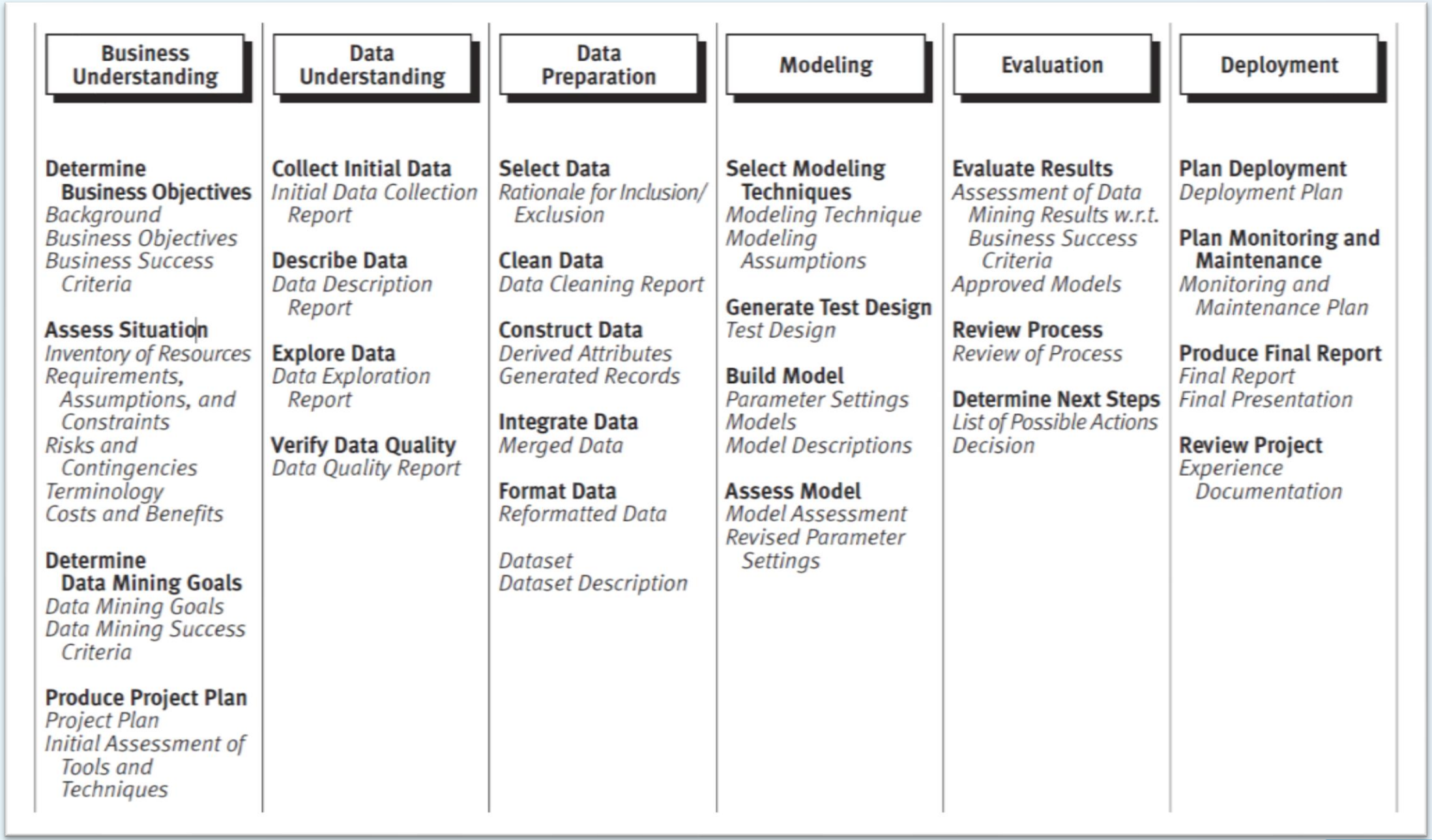
Cross-industry standard process for data mining, known as **CRISP-DM**, is an open standard process model that describes common approaches used by data mining experts. **CRISP-DM** breaks the process of data mining into six major phases.

The sequence of the phases is not strict and moving back and forth between different phases as it is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions, and subsequent data mining processes will benefit from the experiences of previous ones.

Source: Wikipedia

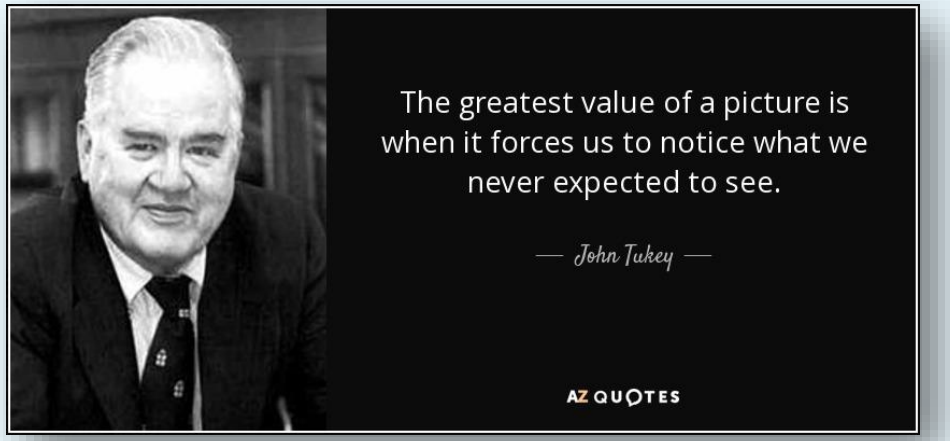


LINKTERA


Business Understanding
Determine Business Objectives
Background
Business Objectives
Business Success Criteria
Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits
Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria
Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques
Data Understanding
Collect Initial Data
Initial Data Collection Report
Describe Data
Data Description Report
Explore Data
Data Exploration Report
Verify Data Quality
Data Quality Report
Data Preparation
Select Data
Rationale for Inclusion/ Exclusion
Clean Data
Data Cleaning Report
Construct Data
Derived Attributes
Generated Records
Integrate Data
Merged Data
Format Data
Reformatted Data
Dataset
Dataset Description
Modeling
Select Modeling Techniques
Modeling Technique
Modeling Assumptions
Generate Test Design
Test Design
Build Model
Parameter Settings
Models
Model Descriptions
Assess Model
Model Assessment
Revised Parameter Settings
Evaluation
Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models
Review Process
Review of Process
Determine Next Steps
List of Possible Actions
Decision
Deployment
Plan Deployment
Deployment Plan
Plan Monitoring and Maintenance
Monitoring and Maintenance Plan
Produce Final Report
Final Report
Final Presentation
Review Project
Experience Documentation

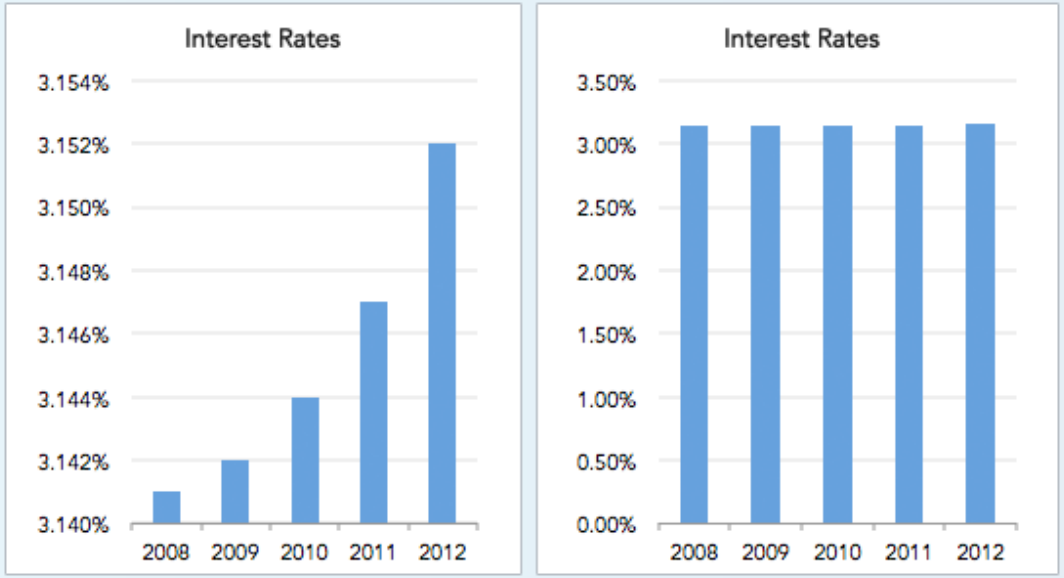
Visualization

Visualization is the conversion of data into a visual or tabular format.



Scale Distortions

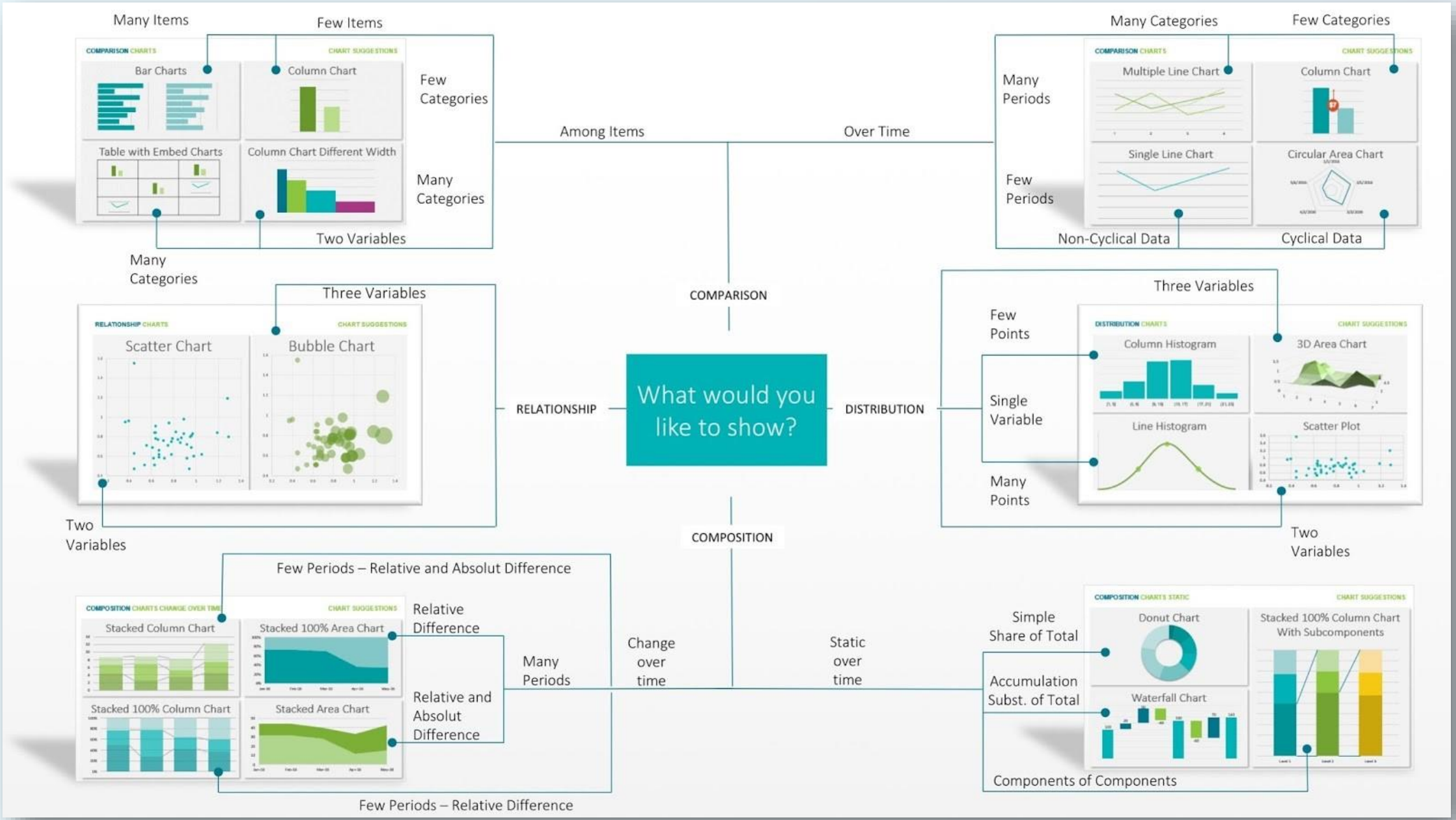
Same Data, Different Y-Axis



Some of good examples

https://www.gapminder.org/tools/#$chart-type=bubbles

LINKTERA



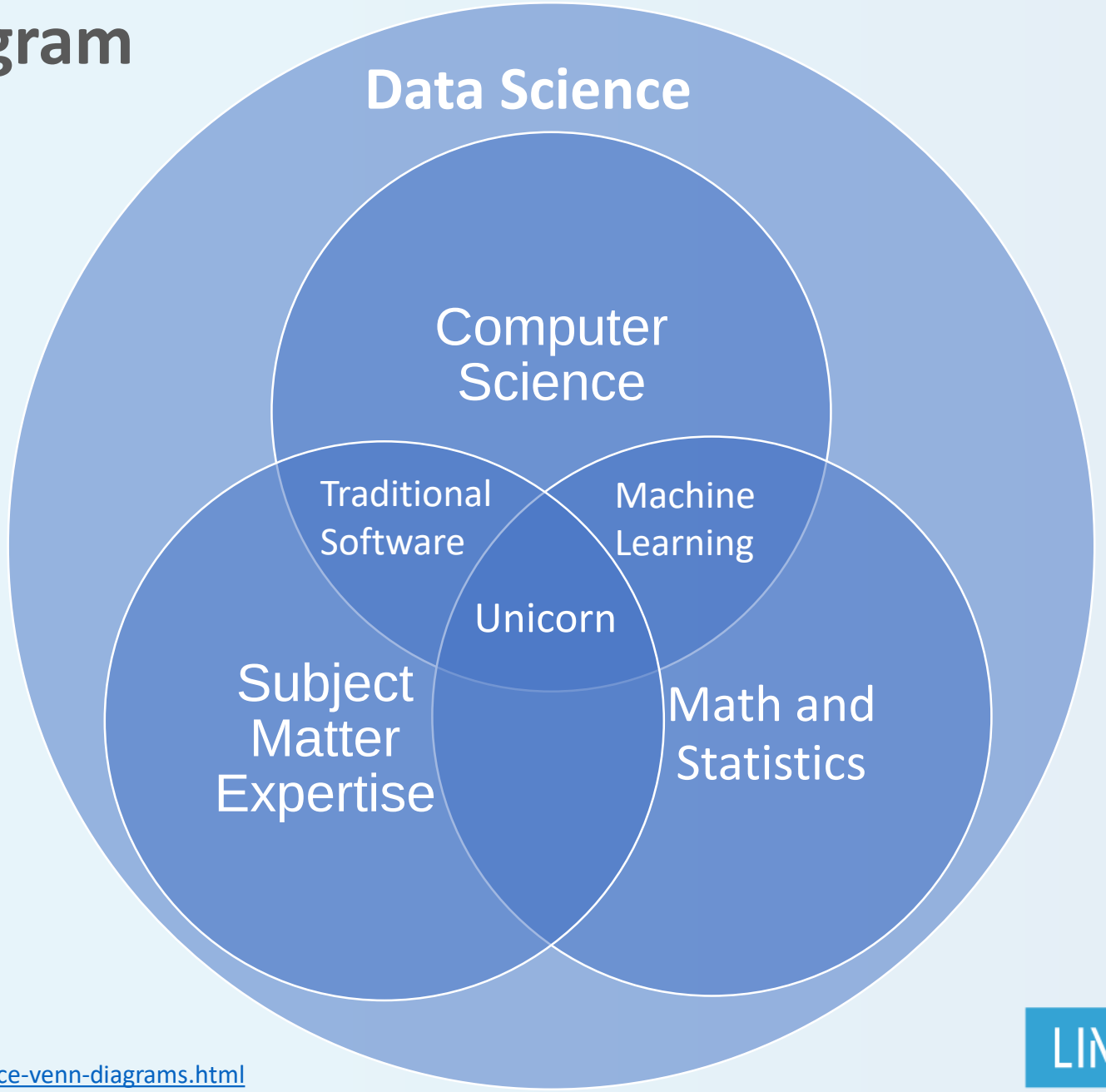
Source: https://www.techprevue.com/decision-tree-perfect-visualisation-data/

LINKTERA

Data Science Venn Diagram

The **unicorn** is a legendary creature that has been described since antiquity as a beast with a single large, pointed, spiraling horn projecting from its forehead.

Source: https://www.wikiwand.com/en/Unicorn



Source: https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html

LINKTERA

Model Basics & Machine Learning

What is a model?

- Models are typically thought of as an abstraction of (or an approximation to) reality.

Why do we need a model?

- Help understand the past, learn from it, and then predict the future
- Models allow us to separate data into predictable and unpredictable elements.

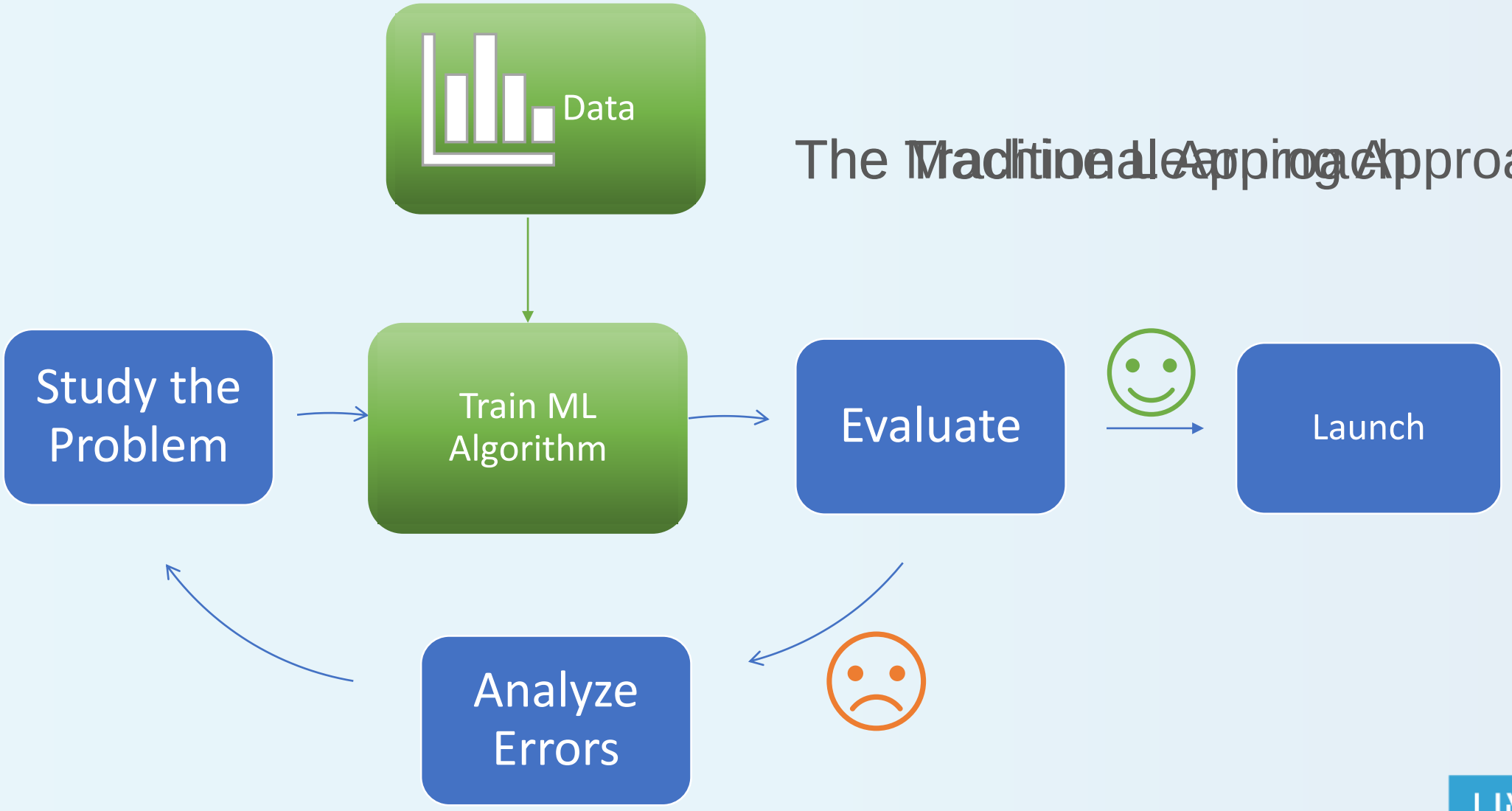
What Is Machine Learning?

- Machine Learning is the science (and art) of programming computers so they can learn from data.

What Is Data Mining?

- Applying ML techniques to dig into large amounts of data can help discover patterns that were not immediately apparent. This is called data mining.

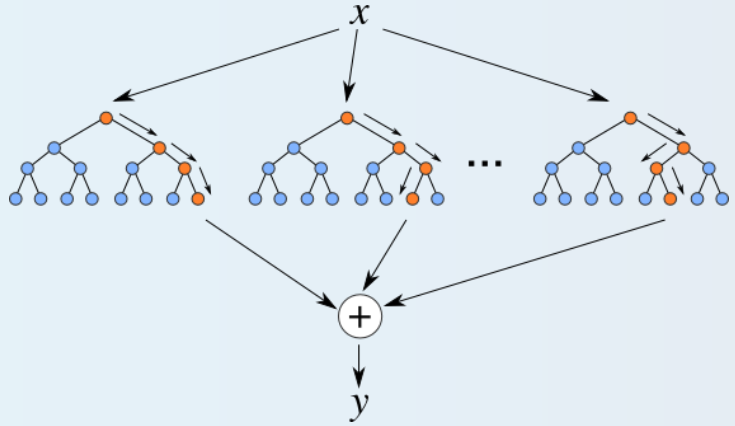
Why Use Machine Learning?



The Traditional Approach

The Machine Learning Approach

Types of Machine Learning Systems

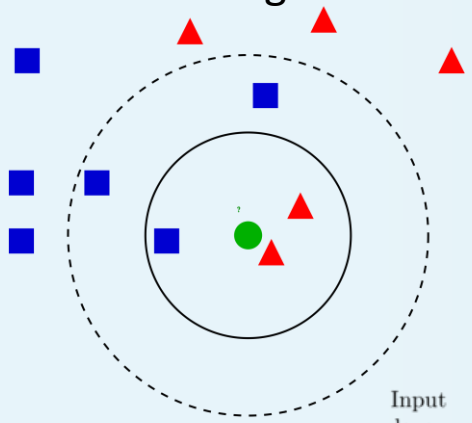


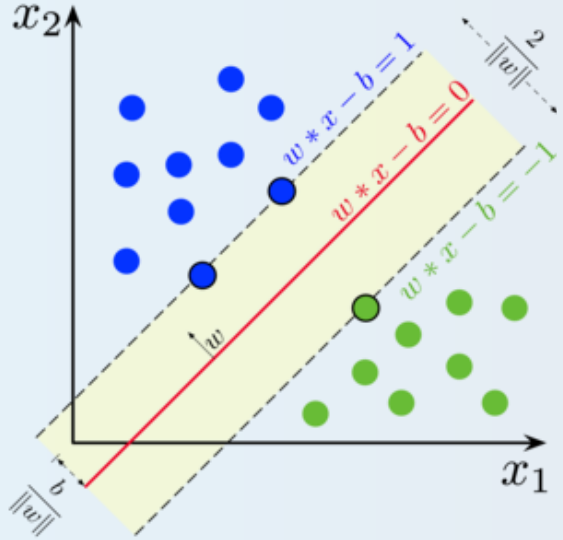
Machine Learning systems can be classified according to the amount and type of supervision they get during training.

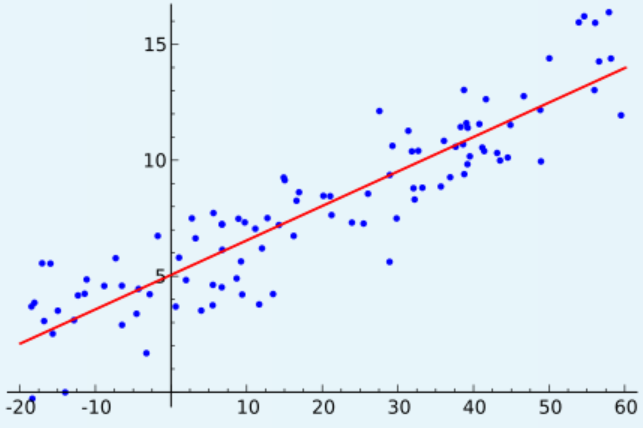
Supervised Learning

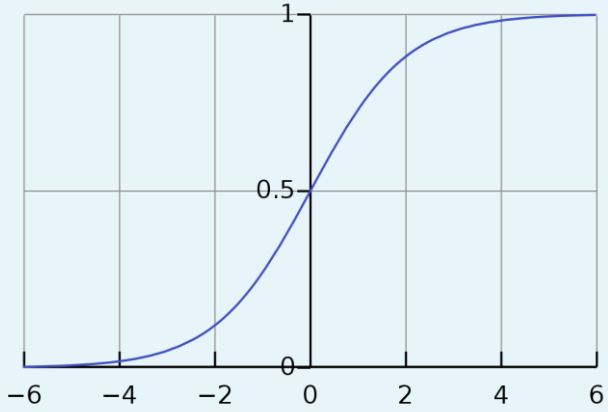
In supervised learning, the training data you feed to the algorithm includes the desired solutions, called labels

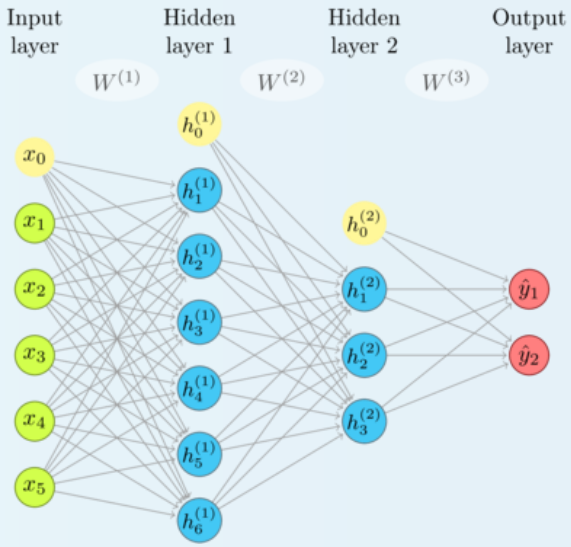
- ❑ k-Nearest Neighbors
- ❑ Linear Regression
- ❑ Logistic Regression
- ❑ Support Vector Machines (SVMs)
- ❑ Decision Trees and Random Forests
- ❑ Neural networks











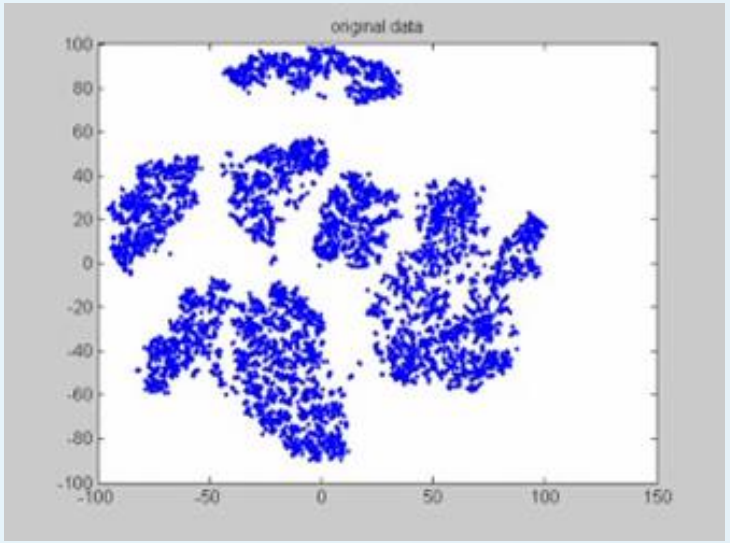
LINKTERA

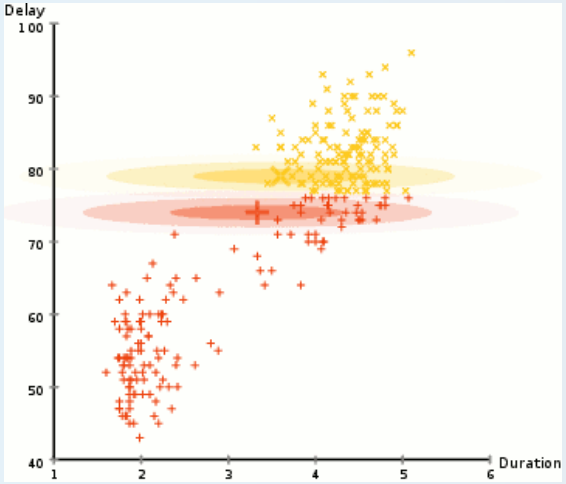
Types of Machine Learning Systems

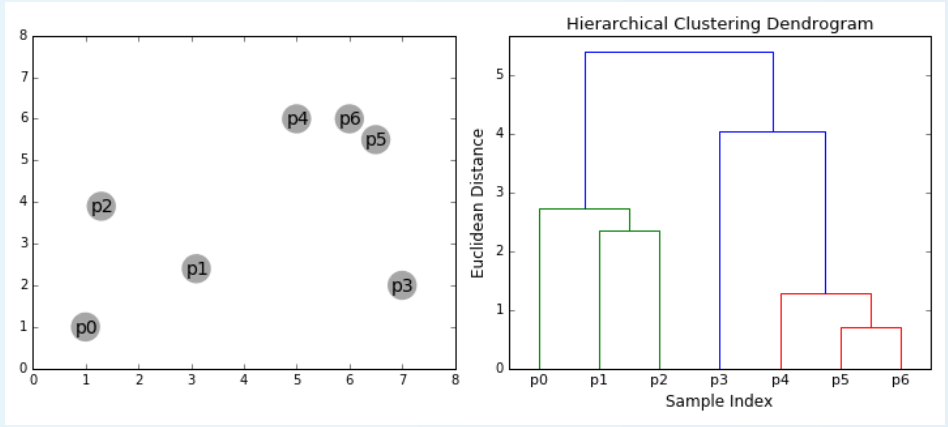
Unsupervised Learning

In unsupervised learning, the training data is unlabeled. The system tries to learn without a teacher.

- Clustering
  - k-Means
  - Hierarchical Cluster Analysis (HCA)
  - Expectation Maximization
- Visualization and dimensionality reduction
  - Principal Component Analysis (PCA)
  - Kernel PCA
  - Locally-Linear Embedding (LLE)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)
- Association rule learning
  - Apriori
  - Eclat







LINKTERA

Types of Machine Learning Systems

Reinforcement Learning

The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return (or penalties in the form of negative rewards, as in figure.

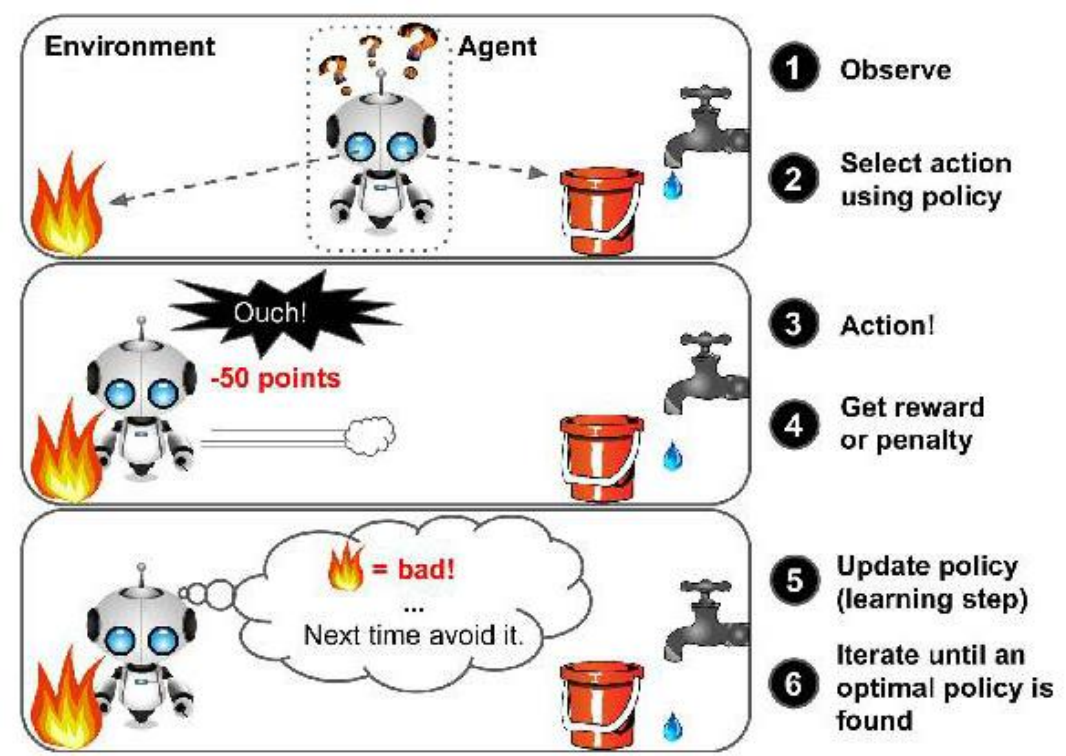
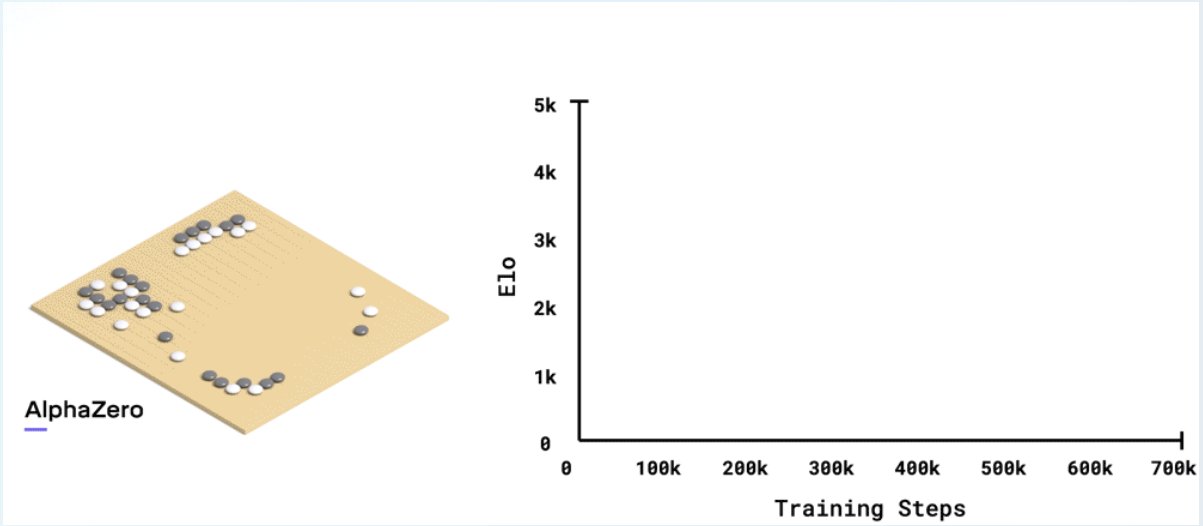


Figure 1-12. Reinforcement Learning



IN CHESS, ALPHAZERO FIRST OUTPERFORMED STOCKFISH AFTER JUST 4 HOURS; IN SHOGI, ALPHAZERO FIRST OUTPERFORMED ELMO AFTER 2 HOURS; AND IN GO, ALPHAZERO FIRST OUTPERFORMED THE VERSION OF ALPHAGO THAT BEAT THE LEGENDARY PLAYER LEE SEDOL IN 2016 AFTER 30 HOURS.

Source: https://deepmind.com/blog/article/alphazero-shedding-new-light-grand-games-chess-shogi-and-go

LINKTERA

Machine Learning Software

R vs Python

R is an open source programing language and environment for statistical computing and graphics.

ADVANTAGES

- **EXCELLENT GRAPHICS**
- **VIBRANT ONLINE COMMUNITY**
- **POWERFUL PACKAGE ECOSYSTEM**

PACKAGES

- **CARET**
- **GGPLOT2**
- **DPLYR/PLYR**
- **STRINGR**
- **FORECAST**

Python is a general-purpose programing language for data science that gained wide popularity because of it's syntax simplicity and operability on different systems.

ADVANTAGES

- **EXTENSIBLE**
- **GOOGLE & TENSORFLOW**
- **EASY TO LEARN**

PACKAGES

- **NUMPY**
- **PANDAS**
- **SCIKIT-LEARN**
- **STATMODELS**
- **MATPLOTLIB**

LINKTERA

Market Basket Analysis

Association Rule Mining

Frequent item example: {yogurt,fruit/vegetable juice,whole milk}

Rule example: {yogurt,fruit/vegetable juice} => {whole milk}

$$Support\ (X => Y) = \frac{(Frequency(X,Y)}{N}$$

Support value indicates the importance of the rule. It is the most commonly used criterion of the rules.

$$\text{Confidence (X => Y)} = \frac{(\text{Frequency (X, Y)}\ )}{(Frequency(X))}$$

The confidence value indicates the conditional probability of a rule. In an X => Y rule, how much % of those who prefer X is also preferred Y.

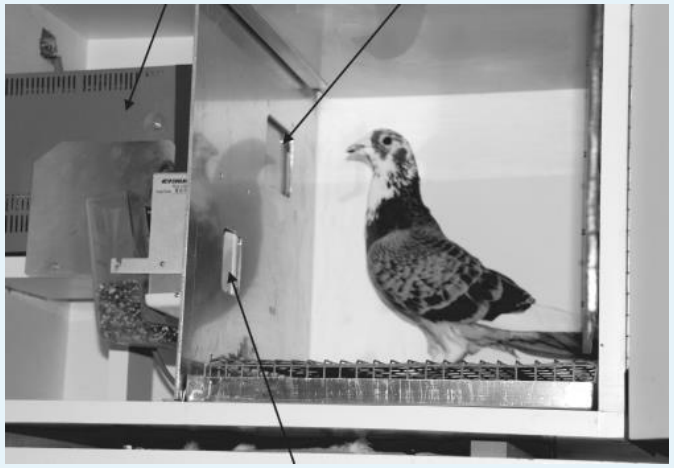
$$Lift\ (X => Y) = \frac{(Frequency\ (X,Y)\ )}{(Frequency(X) * Frequency(Y))}$$

The lift value is the correlation measure of the rules.

LINKTERA

Pigeons as art experts

- Pigeons were able to discriminate between Monet and Picasso with 95% accuracy (when presented with pictures they had been trained on)
- Discrimination still 85% successful for previously unseen paintings of the artists



- Pigeons do not simply memorise the pictures
- They can extract and recognise patterns (the ‘style’)
- They generalise from the already seen to make predictions

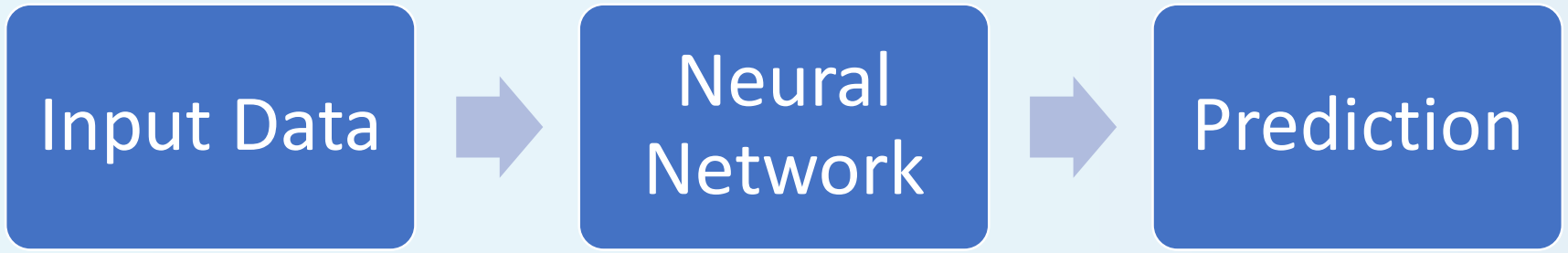
This is what neural networks (biological and artificial) are good at (unlike conventional computer)

Source: https://www.researchgate.net/publication/6971544_Pigeons'_discrimination_of_painting_by_Monet_and_Picasso

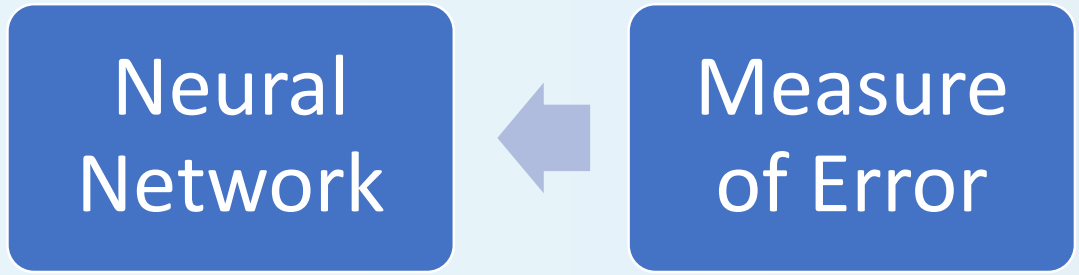
LINKTERA

How Neural Networks Learn: Backpropagation

Forward Pass:



Backward Pass(aka Backpropagation):

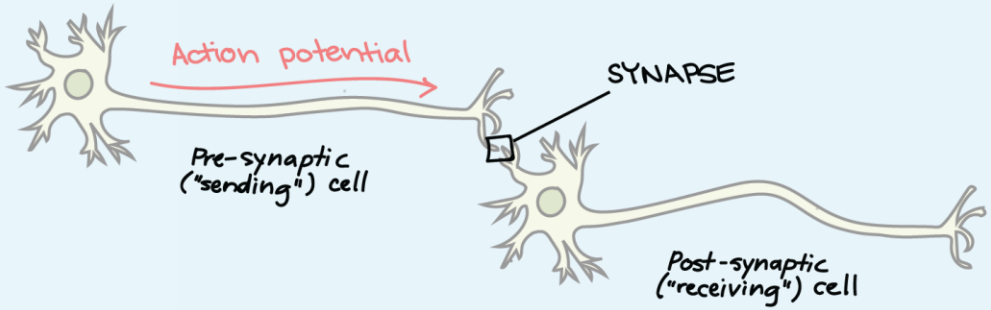


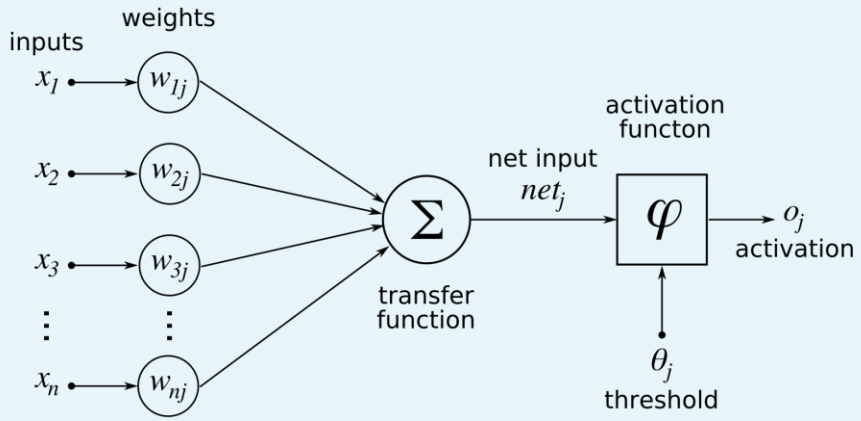
Adjust to Reduce Error

LINKTERA

From Biological to Artificial Neurons

Neuron:computational building block for the brain





(Artificial) Neuron:computational building block for the "neural network"

Differences (among others):

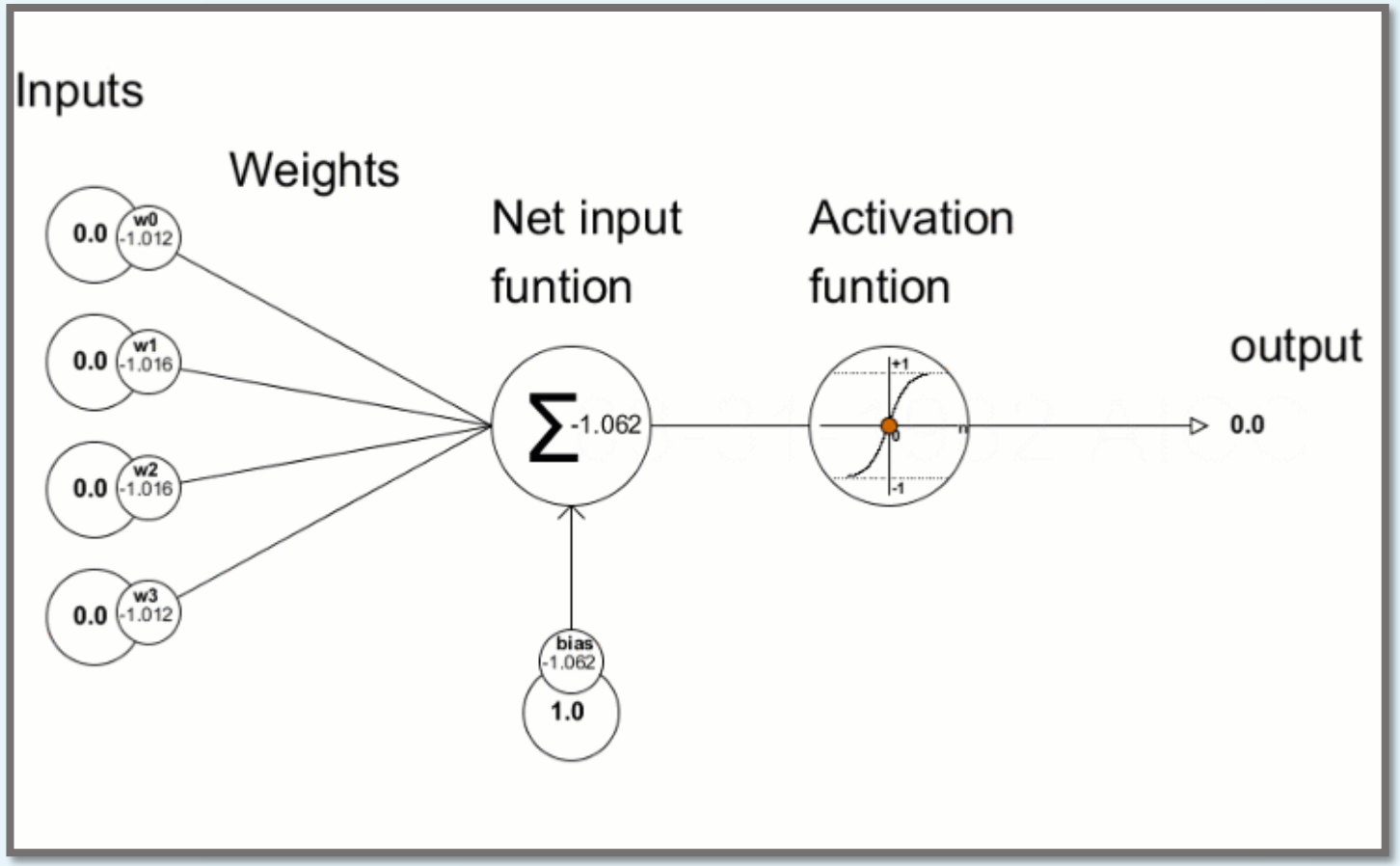
- **Parameters:**Human brains have ~10,000,000 times synapses than artificial neural networks.
- **Topology:**Human brains have no "layers". Topology is complicated.
- **Async:**The human brain works asynchronously, ANNs work synchronously.
- **Learning algorithm**: ANNs use gradient descent for learning. Human brains use … (we don't know)
- **Processing speed**: Single biological neurons are slow, while standard neurons in ANNs are fast.
- **Power consumption:** Biological neural networks use very little power compared to artificial networks
- **Stages:** Biological networks usually don't stop / start learning. ANNs have different fitting (train) and prediction (evaluate) phases.

Similarity (among others):

Distributed computation on a large scale.

LINKTERA

The Artificial Neuron



Inputs
The input is either an external trigger from the environment or comes from outputs of other artificial neurons; it is to be evaluated by the network. It serves as “food” for the neuron and passes through it, thereby becoming an output we can interpret due to the training we gave the neuron. They can be discrete values or real-valued numbers.

Weights
Weights are factors that are multiplied by the entries which correspond to them, increasing or decreasing their value, granting greater or lesser meaning to the input going inside the neuron and, therefore, to the output coming out. The goal of neural network training algorithms is to determine the "best" possible set of weight values for the problem to resolve.

Net Input Function
In this neuron part, the inputs and weights converge in a single-result product as the sum of the multiplication of each entry by its weight. This result or value is passed through the activation function, which then gives us the measures of influence that the input neuron has on the neural network output.

$$x = \sum_{i=1}^{N} A_i W_i + \theta.$$

Activation Function
The activation function leads to the output. There can be several types of activation function (Sigmoid, Tan-h, Softmax, ReLU, among others). It decides whether or not a neuron should be activated.

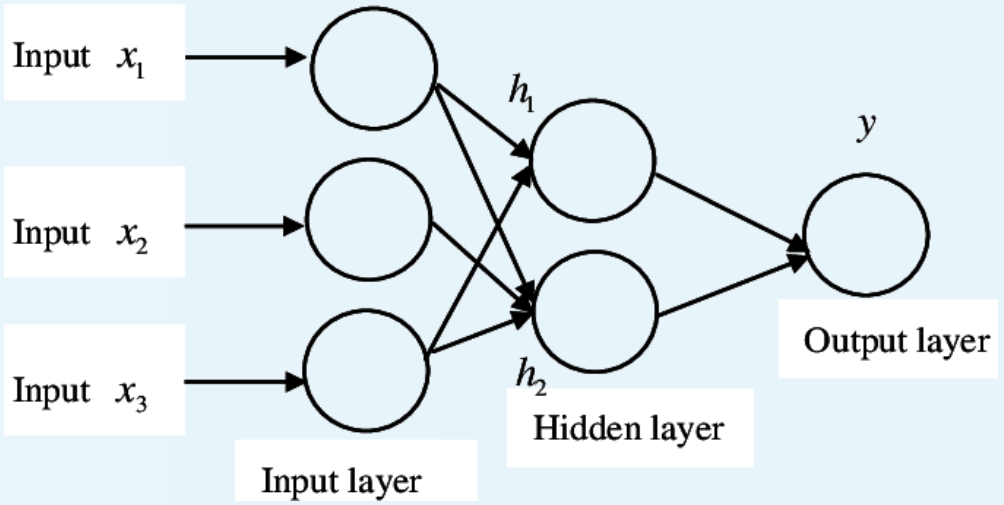
$$y = \frac{1}{1 + exp(-x)}.$$

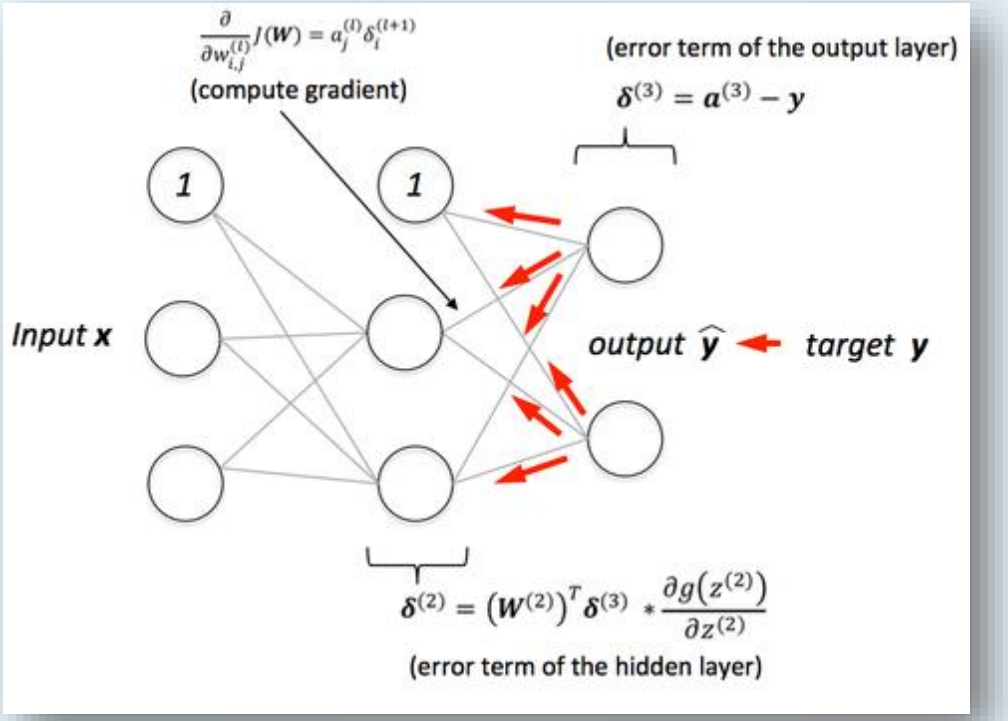
Output
Finally, we have the output. It can be passed to another neuron or sampled by the external environment. This value can be discrete or real, depending on the activation function used.

Source: https://www.mql5.com/en/blogs/post/724245

LINKTERA

Backpropagation





Task: Update the **weights** and **biases** to decrease **loss function**

Subtasks:

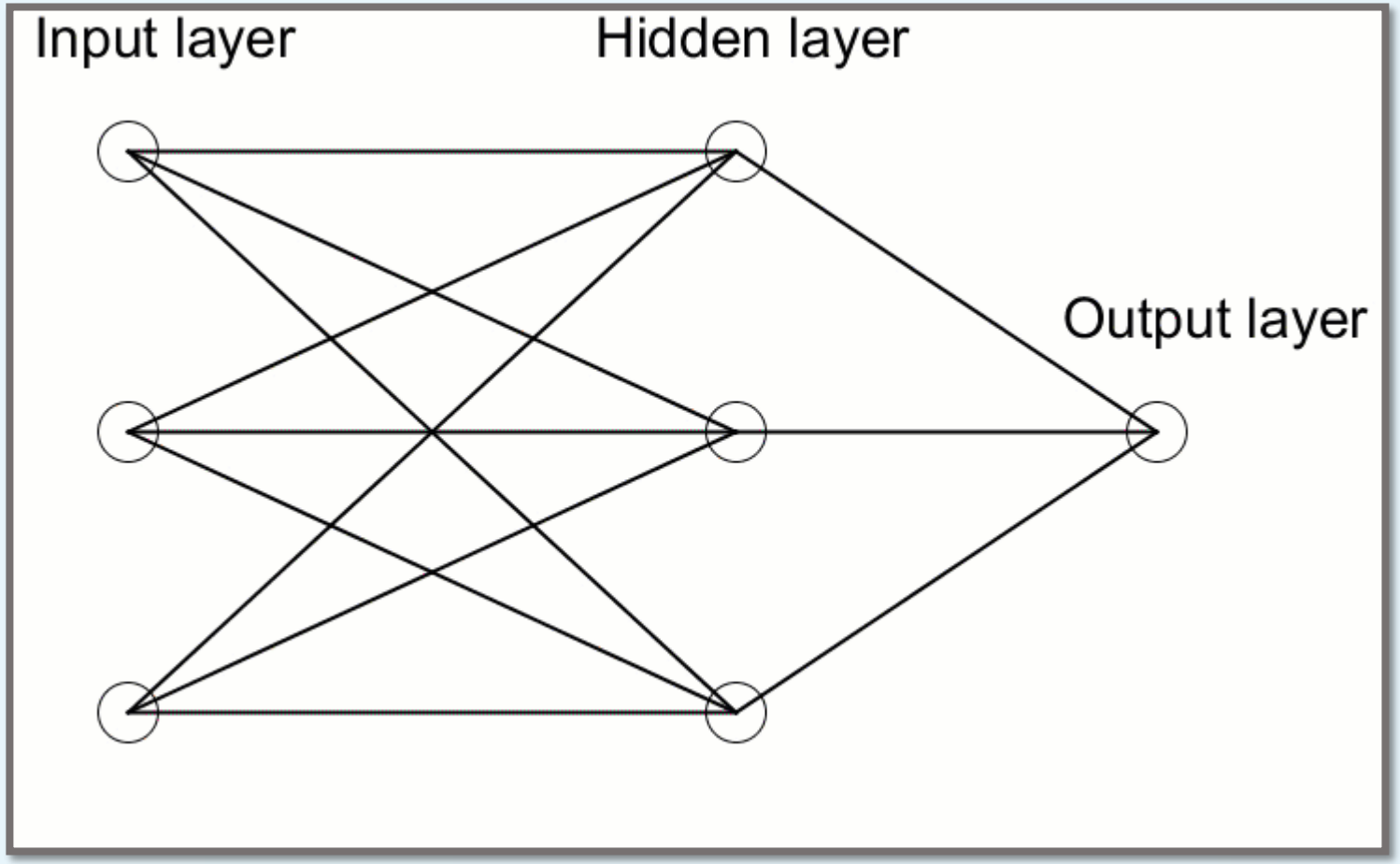
1. Forward pass to compute network output and "error"
2. Backward pass to compute gradients
3. A fraction of the weight's gradient is subtracted from the weight.

Learning Rate

Source: https://deeplearning.mit.edu

LINKTERA

The Neural Network

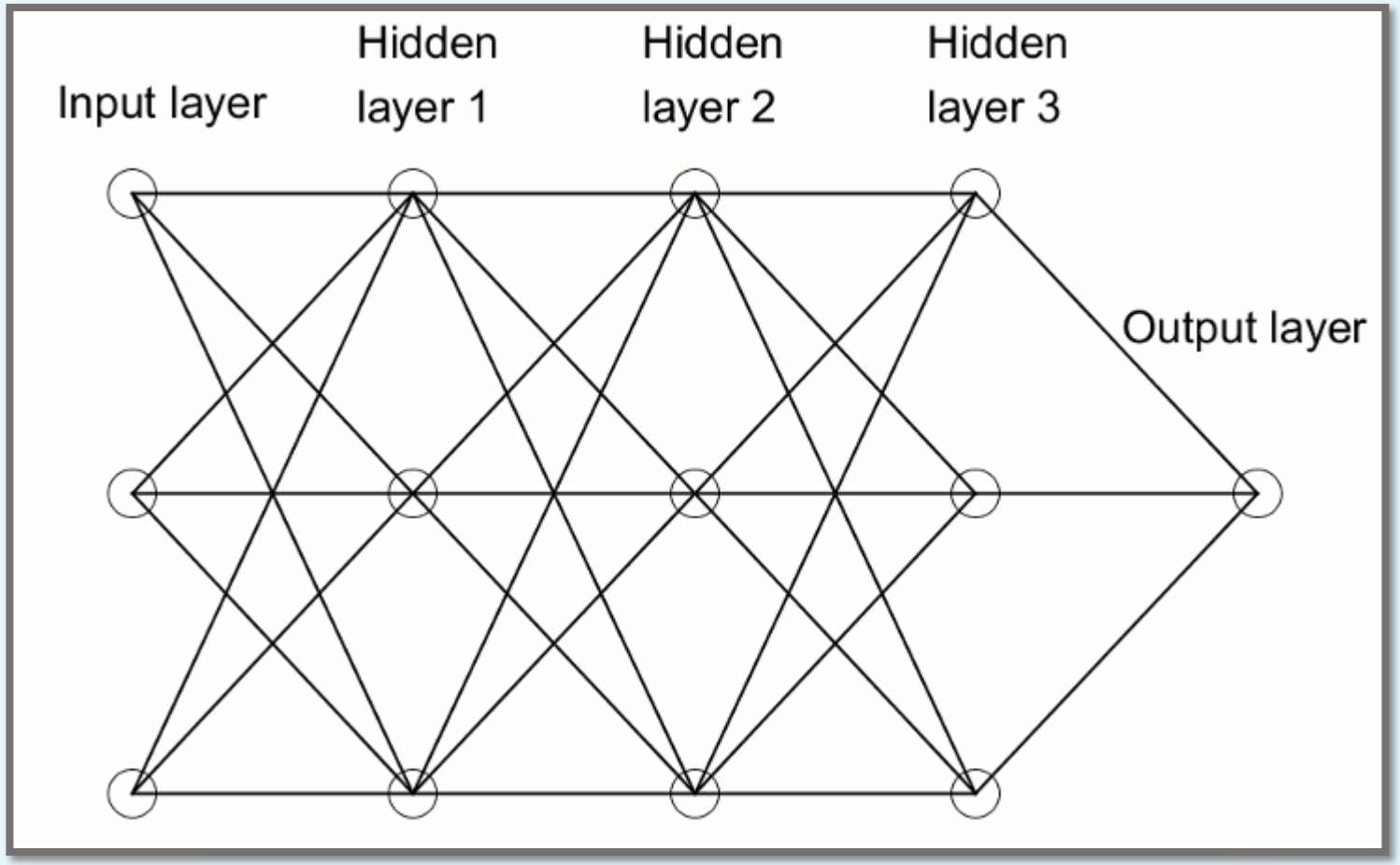


The neural network is inspired by the information processing methods of biological nervous systems, such as the brain.
It is composed of layers of artificial neurons, each layer connected to the next. Therefore, the previous layer acts as an input to the next layer, and so on to the output layer. The neural network's purpose could be clustering through unsupervised learning, classification through supervised learning or regression.

Source: https://www.mql5.com/en/blogs/post/724245

LINKTERA

Deep Neural Network



More than three layers (including input and output) qualifies as “deep” learning.

The further you advance into the neural net, the more complex features there are that can be recognized by your neurons, since they aggregate and recombine features from the previous layer.

Source: https://www.mql5.com/en/blogs/post/724245

LINKTERA

What's Deep Learning?

- **What is it:**
Extract useful patterns from data.

- **How:**
Neural network + optimization

- **How (Practical):**
Python + TensorFlow & friends

- **Hard Part:**
Good Questions + Good Data

- **Why now:**
Data, hardware, community, tools, investment

- **Where do we stand?**
Most big questions of intelligence have not been answered nor properly formulated

Exciting progress:

- Face recognition
- Image classification
- Speech recognition
- Text to speech generation
- Handwriting transcription
- Machine translation
- Medical diagnosis
- Cars: drivable area, lane keeping
- Digital assistants
- Ads, search, social recommendations
- Game playing with deep RL

Tensorflow

What is it: Deep Learning Library (and more)

Facts: Open Source, Python, Google

Community:

117,000+ GitHub stars

TensorFlow.org: Blogs, Documentation, DevSummit , YouTube talks

Ecosystem:

Keras high level API

TensorFlow.js: in the browser

TensorFlow Lite: on the phone

Colaboratory : in the cloud

TPU: optimized hardware

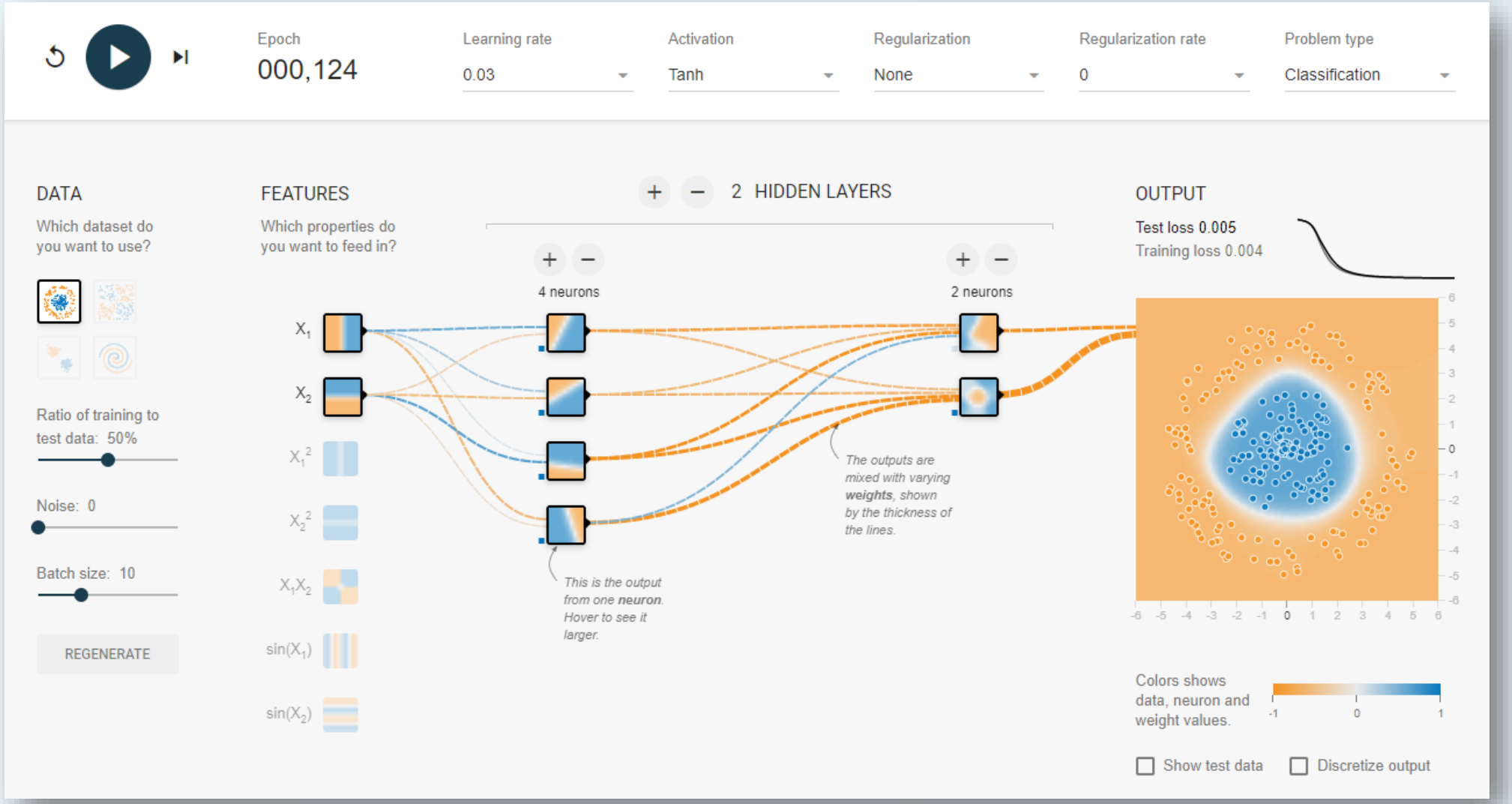
TensorBoard : visualization

TensorFlow Hub: graph modules

Alternatives: PyTorch , MXNet , CNTK

Tensorflow Playground

http://playground.tensorflow.org



LINKTERA

Deep Learning Project Example on Keras

We are going to use the Pima Indians diabetes dataset. This is a standard machine learning dataset from the UCI Machine Learning repository. It describes patient medical record data for Pima Indians and whether they had an onset of diabetes within five years.

As such, it is a binary classification problem (onset of diabetes as 1 or not as 0). All of the input variables that describe each patient are numerical. This makes it easy to use directly with neural networks that expect numerical input and output values, and ideal for our first neural network in Keras.

```
[6.0, 148.0, 72.0, 35.0, 0.0, 33.6, 0.627, 50.0] => 1 (expected 1)
[1.0, 85.0, 66.0, 29.0, 0.0, 26.6, 0.351, 31.0] => 0 (expected 0)
[8.0, 183.0, 64.0, 0.0, 0.0, 23.3, 0.672, 32.0] => 1 (expected 1)
[1.0, 89.0, 66.0, 23.0, 94.0, 28.1, 0.167, 21.0] => 0 (expected 0)
[0.0, 137.0, 40.0, 35.0, 168.0, 43.1, 2.288, 33.0] => 1 (expected 1)
[5.0, 116.0, 74.0, 0.0, 0.0, 25.6, 0.201, 30.0] => 0 (expected 0)
[3.0, 78.0, 50.0, 32.0, 88.0, 31.0, 0.248, 26.0] => 0 (expected 1)
[10.0, 115.0, 0.0, 0.0, 0.0, 35.3, 0.134, 29.0] => 1 (expected 0)
[2.0, 197.0, 70.0, 45.0, 543.0, 30.5, 0.158, 53.0] => 1 (expected 1)
[8.0, 125.0, 96.0, 0.0, 0.0, 0.0, 0.232, 54.0] => 0 (expected 1)
```

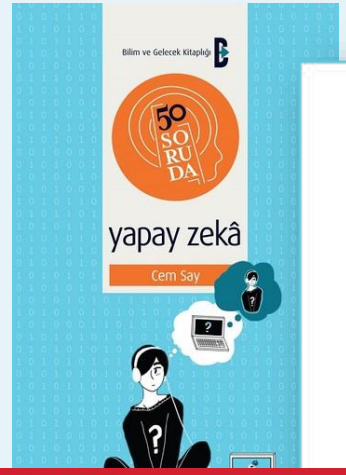
Source: https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

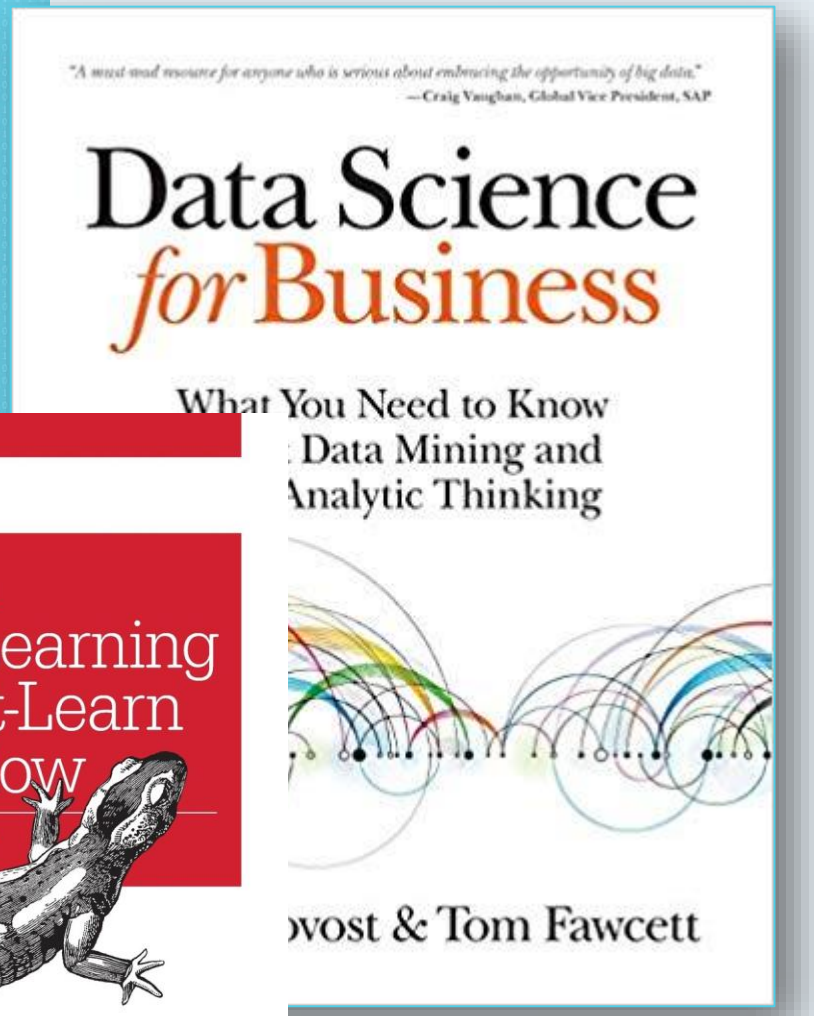
LINKTERA

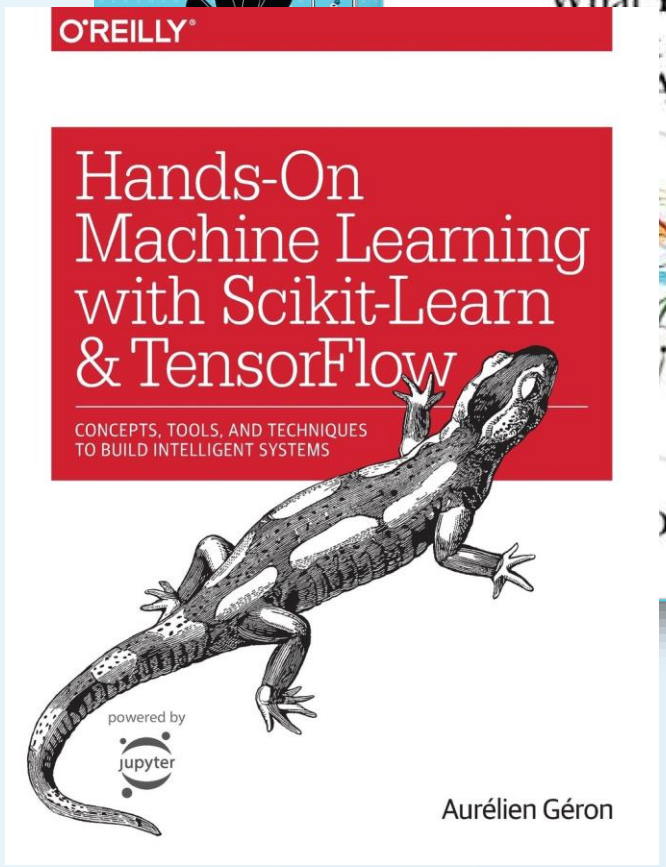
Further Readings

Books:

- 50 Soruda Yapay Zeka – Cem Say
- Introduction to Machine Learning – Ethem Alpaydın
- Hands-On Machine Learning with Scikit-Learn and TensorFlow… – Aurelien Geron
- Data Science for Business - Foster Provost, Tom Fawcett
- Practical Statistics for Data Scientists – Peter Bruce
- Python for Data Analysis – Wes McKinney







Web:

- https://datatau.net/
- https://www.kaggle.com/
- https://towardsdatascience.com/
- http://colah.github.io/
- https://www.datacamp.com/

LINKTERA

LINKTERA

Vogue Business Center, Küçükbakkalköy mah.
Merdivenköy Yolu Cad. Rüya sok. No.12, Kat 18
34746 İstanbul-Ataşehir