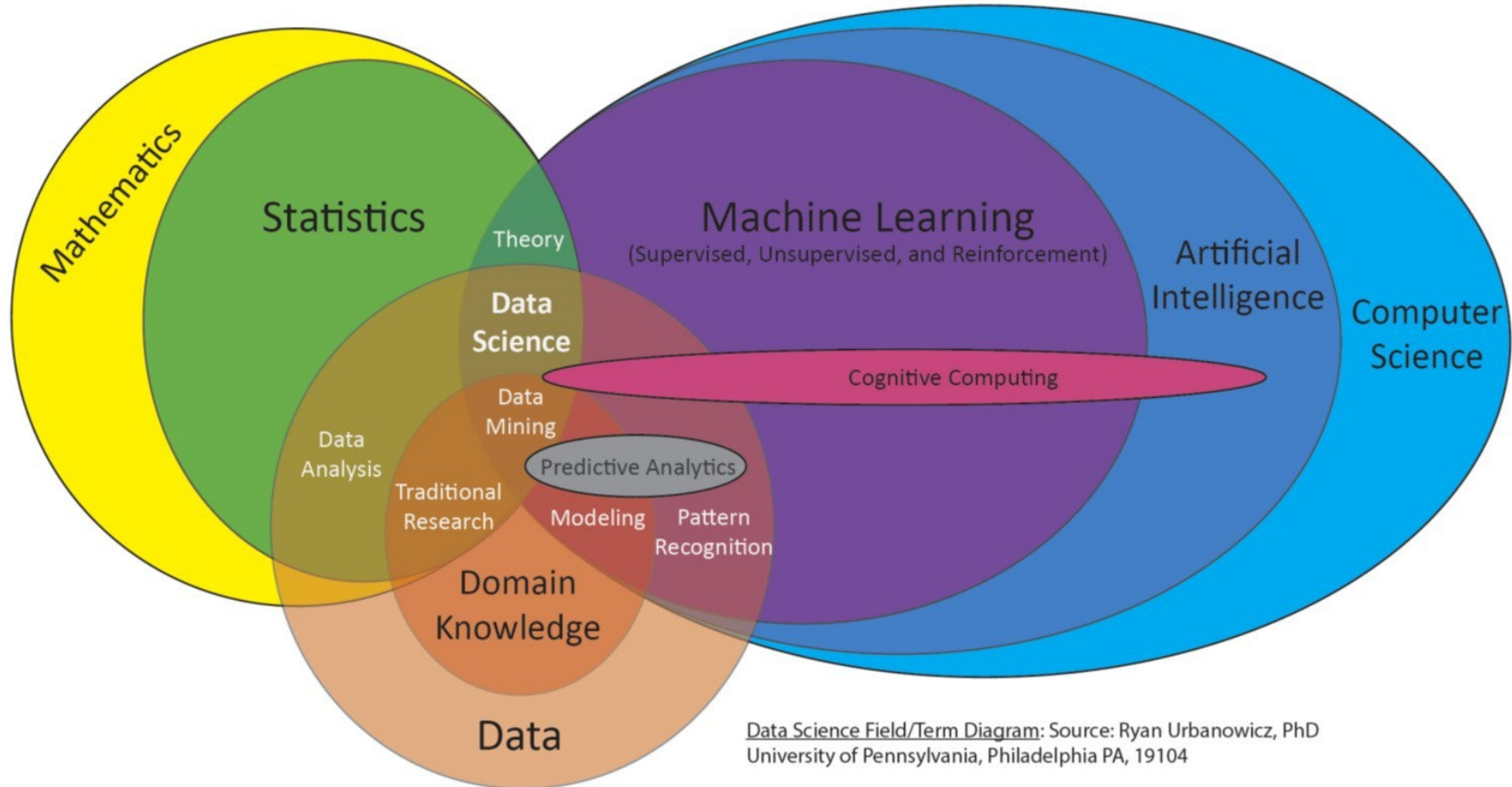# Artificial Intelligence
&
Data Science

Venn diagram

Data Science Field/Term Diagram: Source: Ryan Urbanowicz, PhD
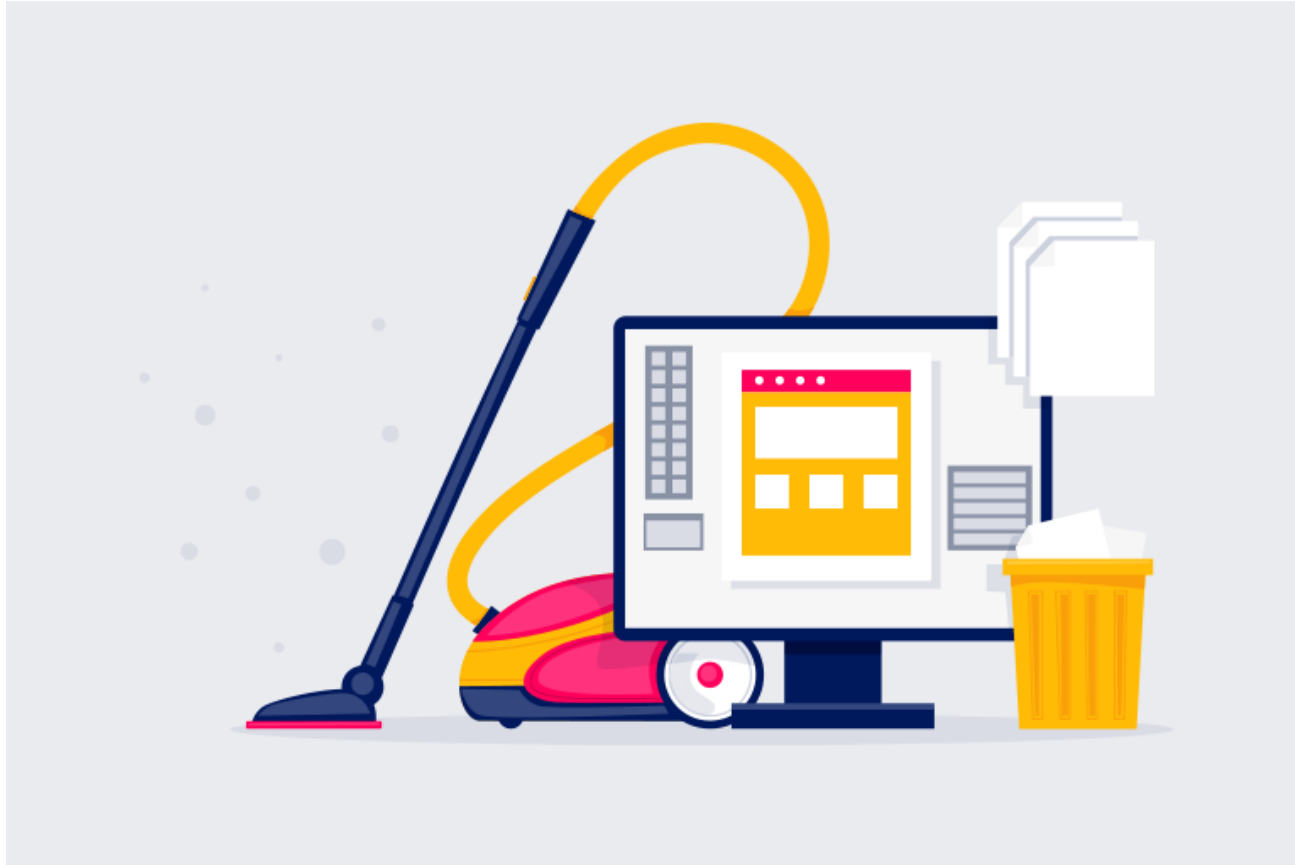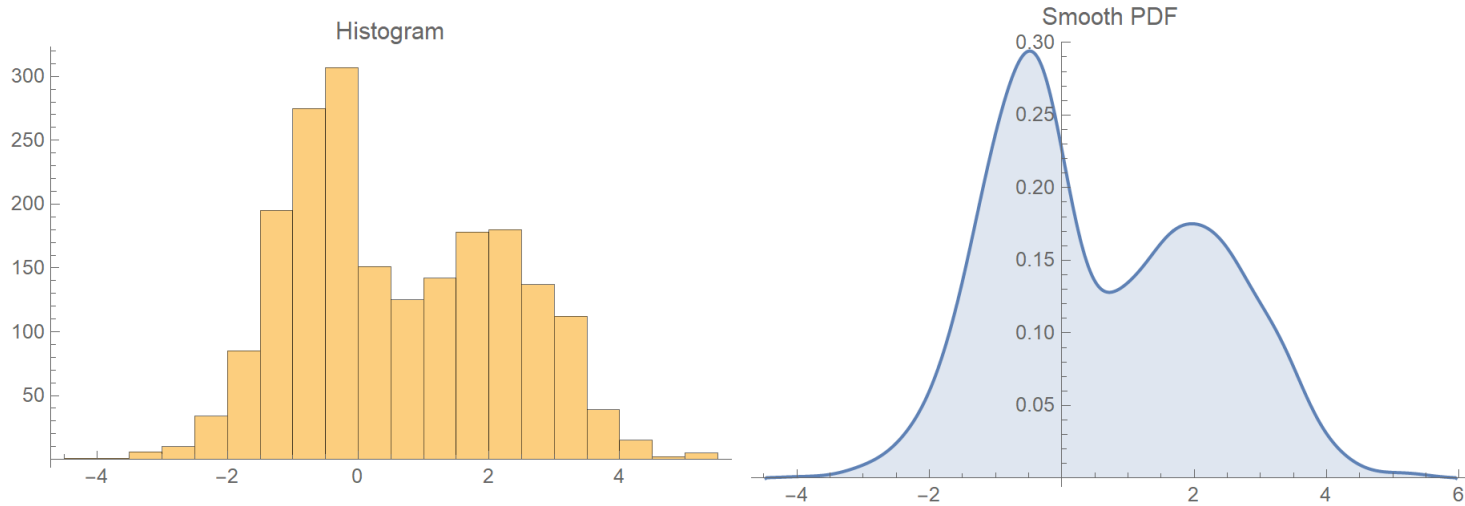University of Pennsylvania, Philadelphia PA, 19104

Roadmap

Author: Swami Chandrasekaran

9. Data Munging — Denoising, Feature Extraction, Binning Sparse Values, Unbiased Estimators, Handling Missing Values, Data Scrubbing, Normalization, Dimensionality & Numerosity Reduction

80%

Sampling — Using ETL, How much Data?, Google OpenRefine, Data Survey

Stratified Sampling, Principal Component Analysis

8. Data Ingestion — Transformation & Enrichment, Data Fusion, Data Integration, Data Sources & Acquisition, Data Discovery, Summary of Data Formats

5. Text Mining / NLP — Corpus, Named Entity Recognition, Text Analysis, Term Document Matrix, UIMA, Term Document Matrix, Term Frequency & Weight, Support Vector Machines, Association Rules, Market Based Analysis, Feature Extraction, Using Mahout, Using Weka, Using NLTK

Clustering

Hierarchical Clustering, K-means Clustering, Neural Networks, Sentiment Analysis, Collaborative Filtering, Tagging, Vocabulary Mapping, Classify Text

6. Visualization — Data Exploration in R (Hist, Boxplot etc), Uni, Bi & Multivariate Viz, ggplot2, Histogram & Pie (Uni), Tree & Tree Map, Scatter Plot (Bi), Line Charts (Bi), Spatial Charts, Survey Plot, Timeline, Decision Tree

Regression — Perceptron, Linear Regression, Ranking, Logistic Regression

Classification — K-Nearest Neighbor, Naïve Bayes Classifiers, Boosting, Decision Trees, Classification Rate, Trees & Classification, Bias & Variance, Overfitting, Lift, Prediction, Classifier, Training & Test Data, Concepts, Inputs & Attributes, Unsupervised Learning, Supervised Learning, Categorical Var, Numerical Var, What is ML?

50%

30%

Euclidean Distance, Least' Fit, Causation, Pearson Coeff, Correlation, Covariance, Regression, Kernel Density Estimate, MLE, Confid Int (CI), Estimation, Chi² Test, P-Value, Hypothesis Testing, Monte Carlo Method, Central Limit Theorem

4. Machine Learning

1. Fundamentals — Matrices & Linear Algebra Fundamentals, Hash Functions, Binary Tree, O(n), Relational Algebra, DB Basics, Inner, Outer, Cross, Theta Join, CAP Theorem, Tabular Data, Data Frames & Series, Sharding, OLAP, Multidimensional Data Model, ETL, Reporting Vs BI Vs Analytics, JSON & XML, NoSQL, Regex, Vendor Landscape, Env Setup

Entropy

Prob Den Fn (PDF), ANOVA, Skewness, Continuos Distributions (Normal, Poisson, Gaussian), Cumul Dist Fn (CDF), Random Variables, Bayes Theorem, Probability Theory, Percentiles & Outliers, Histograms, Exploratory Data Analysis

Tableau, IBM ManyEyes, InfoVis, D3.js

40%

10. Toolbox — MS Excel w/ Analysis ToolPak, Java, Python, R, R-Studio, Rattle, Weka, Knime, RapidMiner, Hadoop Dist of Choice, Spark, Storm, Flume, Scibe, Chukwa, Nutch, Talend, Scraperwiki, Webscraper, Flume, Sqoop, tm, RWeka, NLTK, RHIPE, D3.js, ggplot2, Shiny

Data Frames, Reading CSV Data, Reading Raw Data, Subsetting Data, Manipulate Data Frames, Functions

Lists, Factors, Arrays, Matrices, Vectors, Variables, Expressions, R Basics, R Setup R Studio

15%

Factor Analysis, Install Pkgs

Rapid Miner, IBM SPSS

Descriptive Statistics (mean, median, range, SD, Var), Pick a Dataset (UCI Repo)

Job & Task Tracker, M/R Programming, Sqoop: Loading Data in HDFS, Flume, Scribe: For Unstruct Data, SQL with Pig, DW/H with Hive, Scribe, Chukwa For Weblog, Using Mahout

Name & Data Nodes, Setup Hadoop (IBM / Cloudera / HortonWorks), Data Replication Principles, HDFS, Hadoop Components, Map Reduce Fundamentals

Zookeeper Avro, Storm: Hadoop Realtime, Rhadoop, RHIPE, rmr, Cassandra, MongoDB, Neo4j

Cassandra, MongoDB

IBM Languageware

100%

60%

2. Statistics

3. Programming

Working in Excel, Python Basics

7. Big Data

5%

Author: Swami Chandrasekaran

Data-Mining

Data Cleaning

Statistik

Histogram

Smooth PDF

AI

Reasoning, problem-solving

Knowledge representation

Planning

Learning

Natural language processing

Perception

Motion and manipulation

Social intelligence

General intelligence

Search and optimization

Logic

Probabilistic methods

Learning methods

Artificial neural networks
Deep learning

Languages and hardware

Path finding

© QGIS 2010
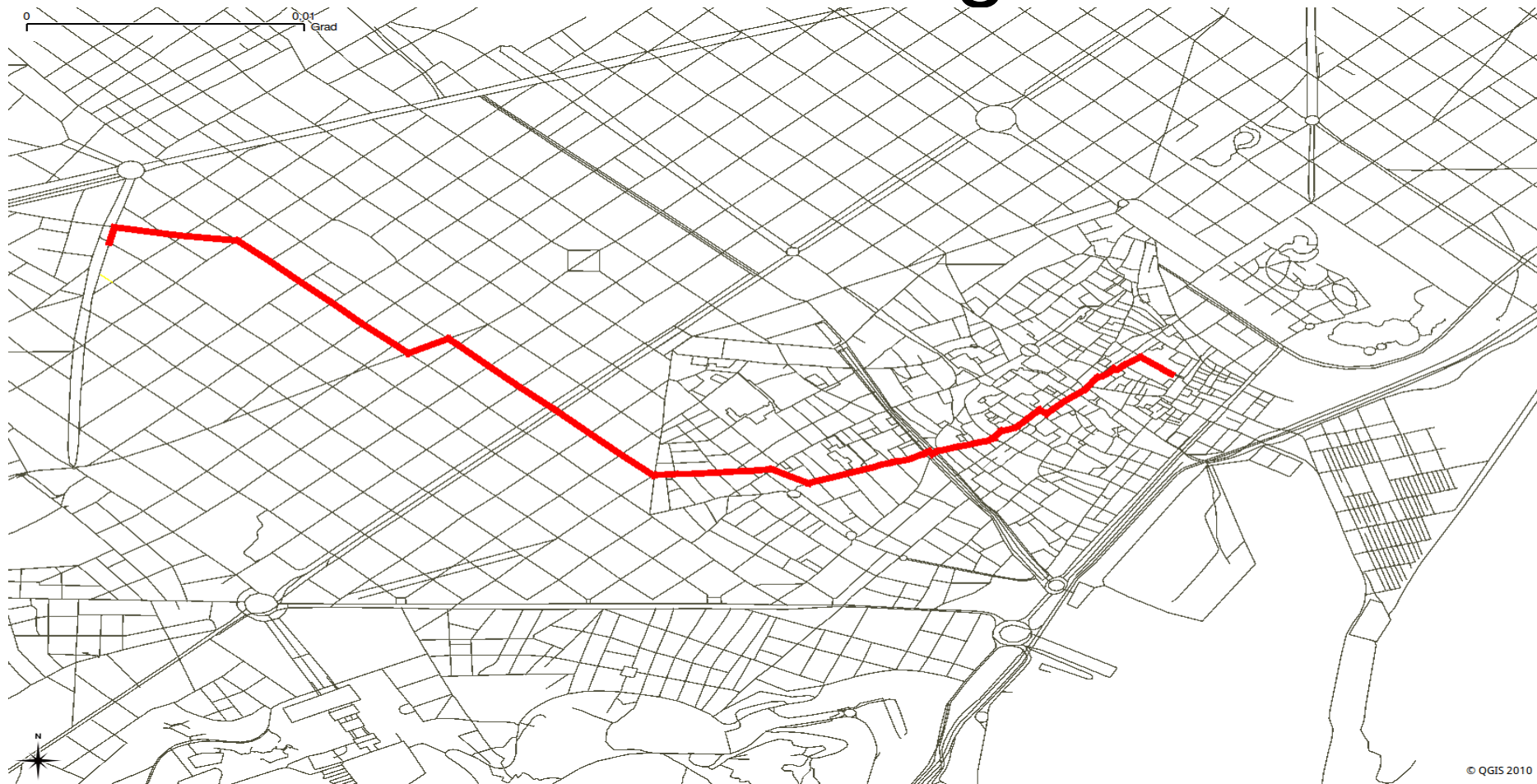
Tourplanning

Traveling sales person (TSP)

tour planing

multi vehicle tour planing

Constraints multi vehicle tour planing
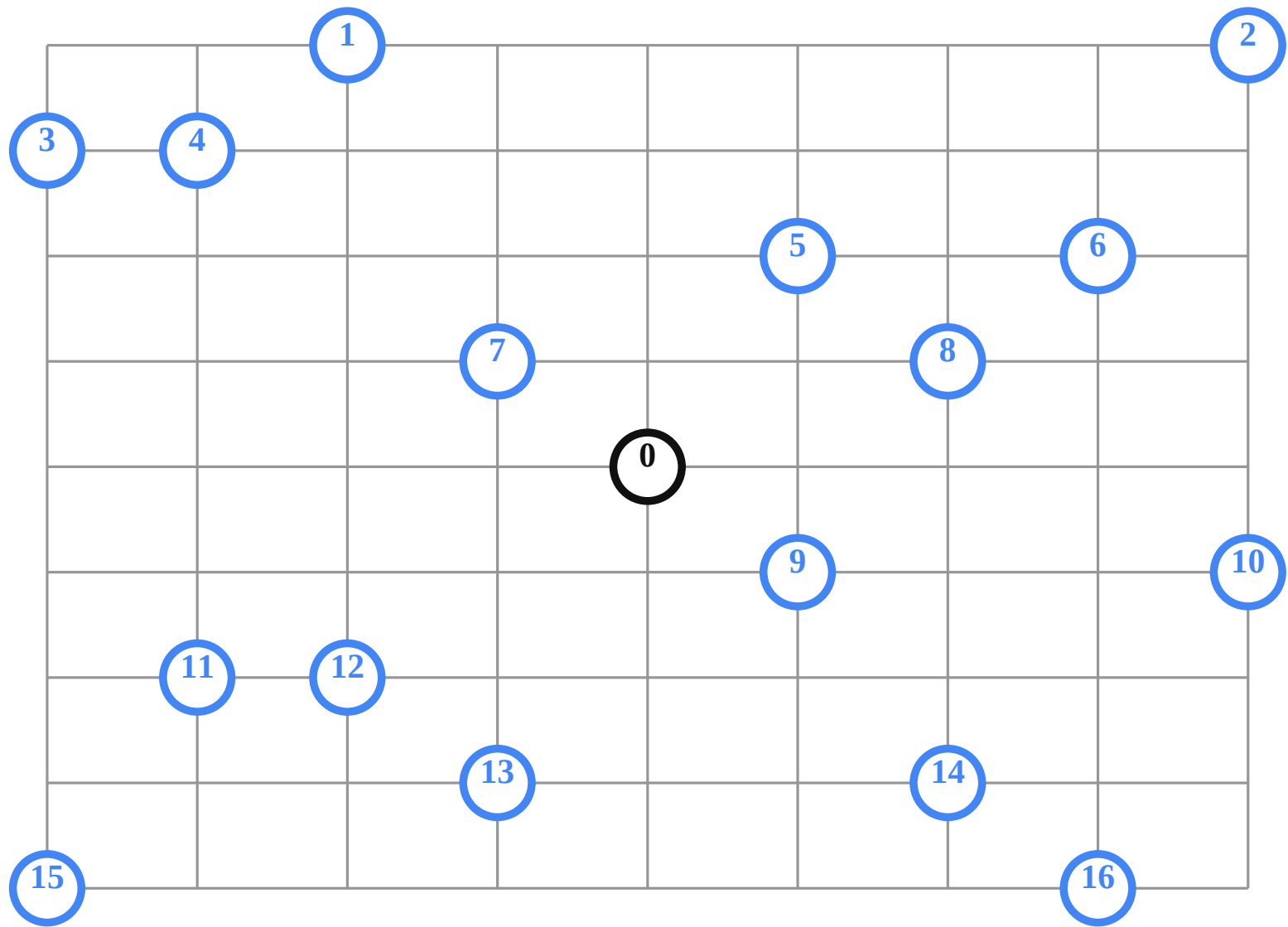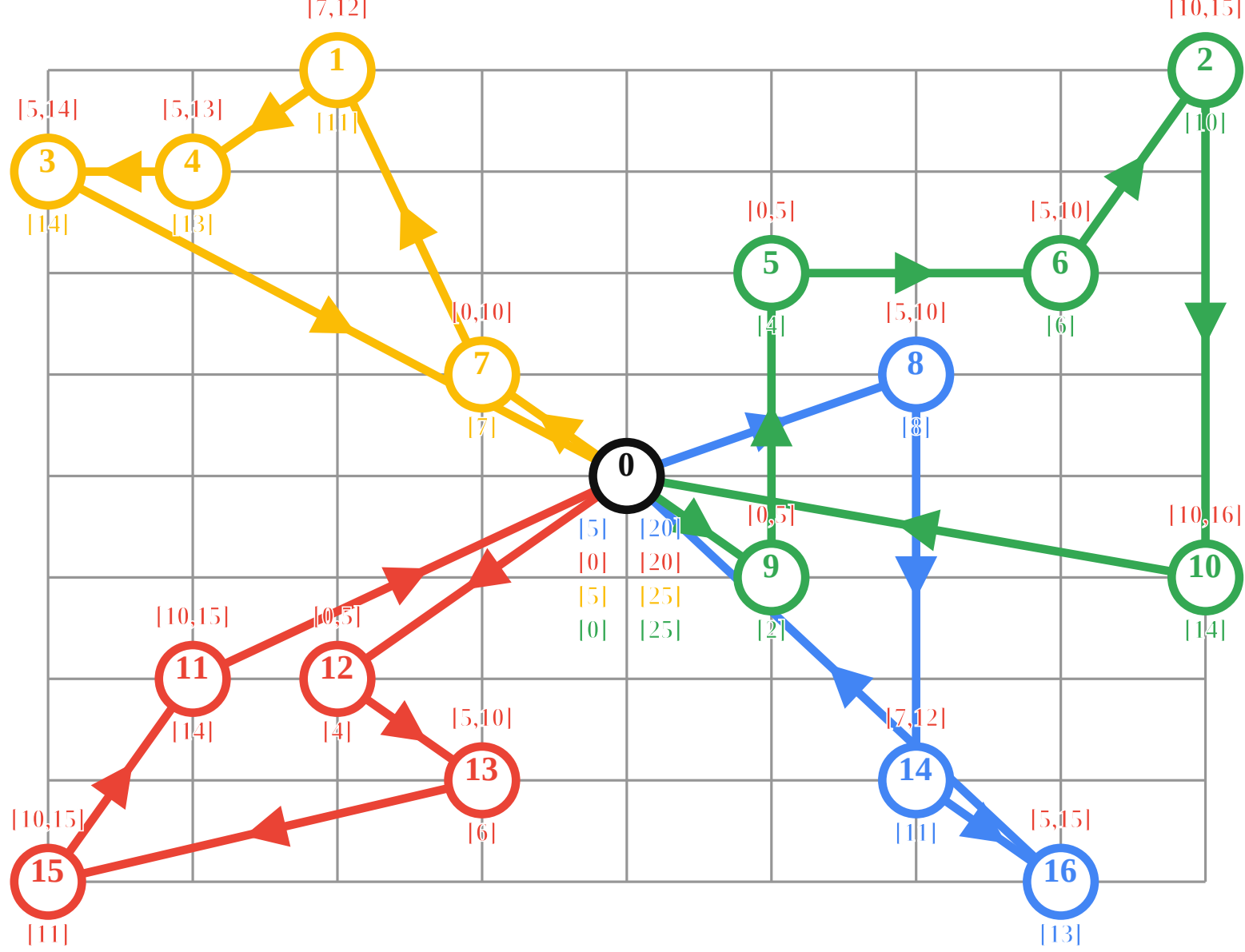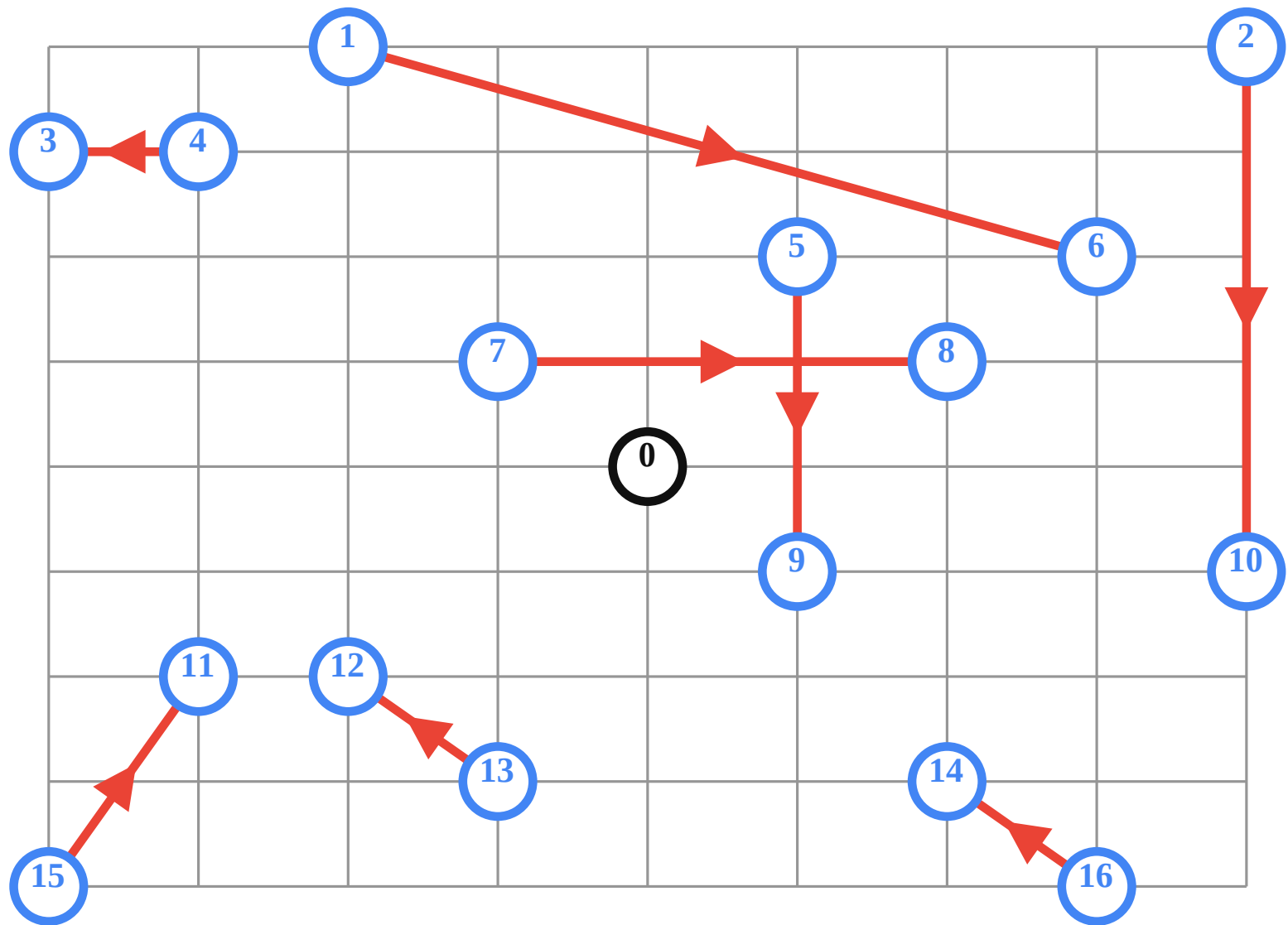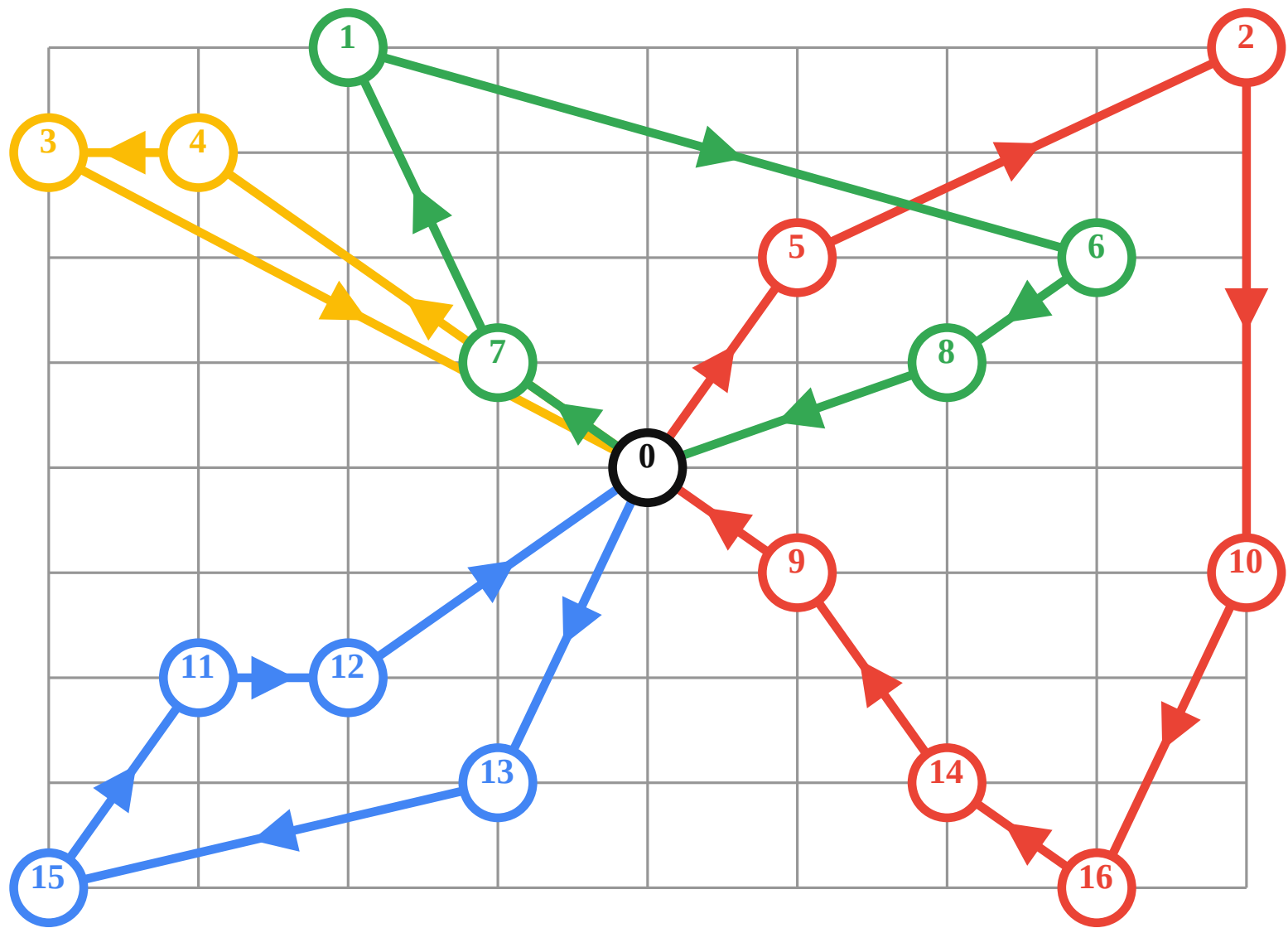
Genetic Algorithms

Evolution und Natur imitieren um Probleme zu lösen
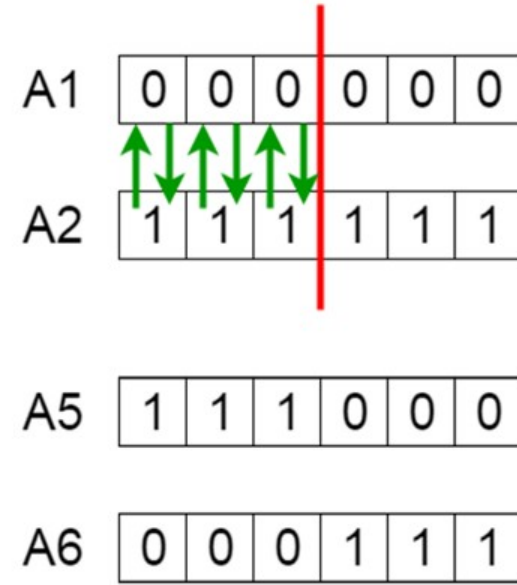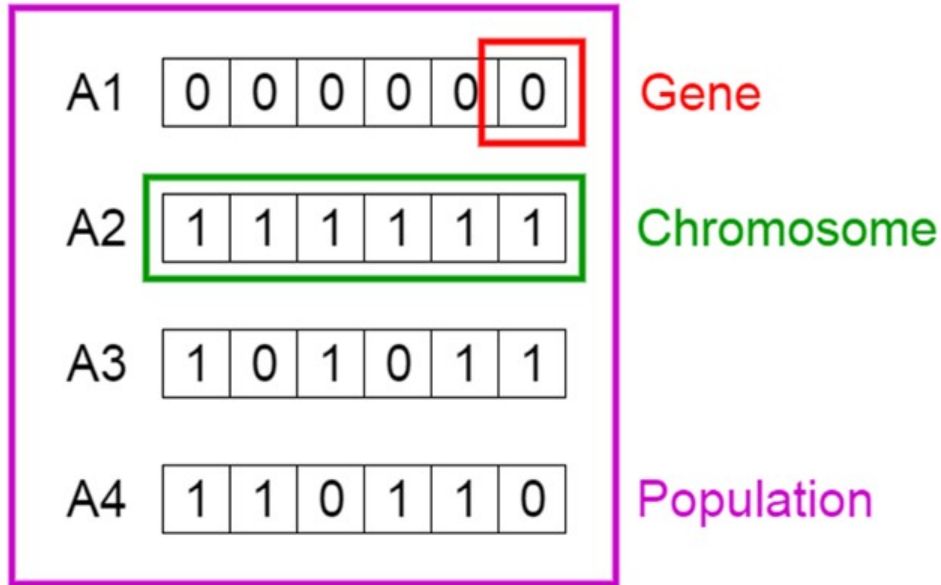
Ameisenkolonie
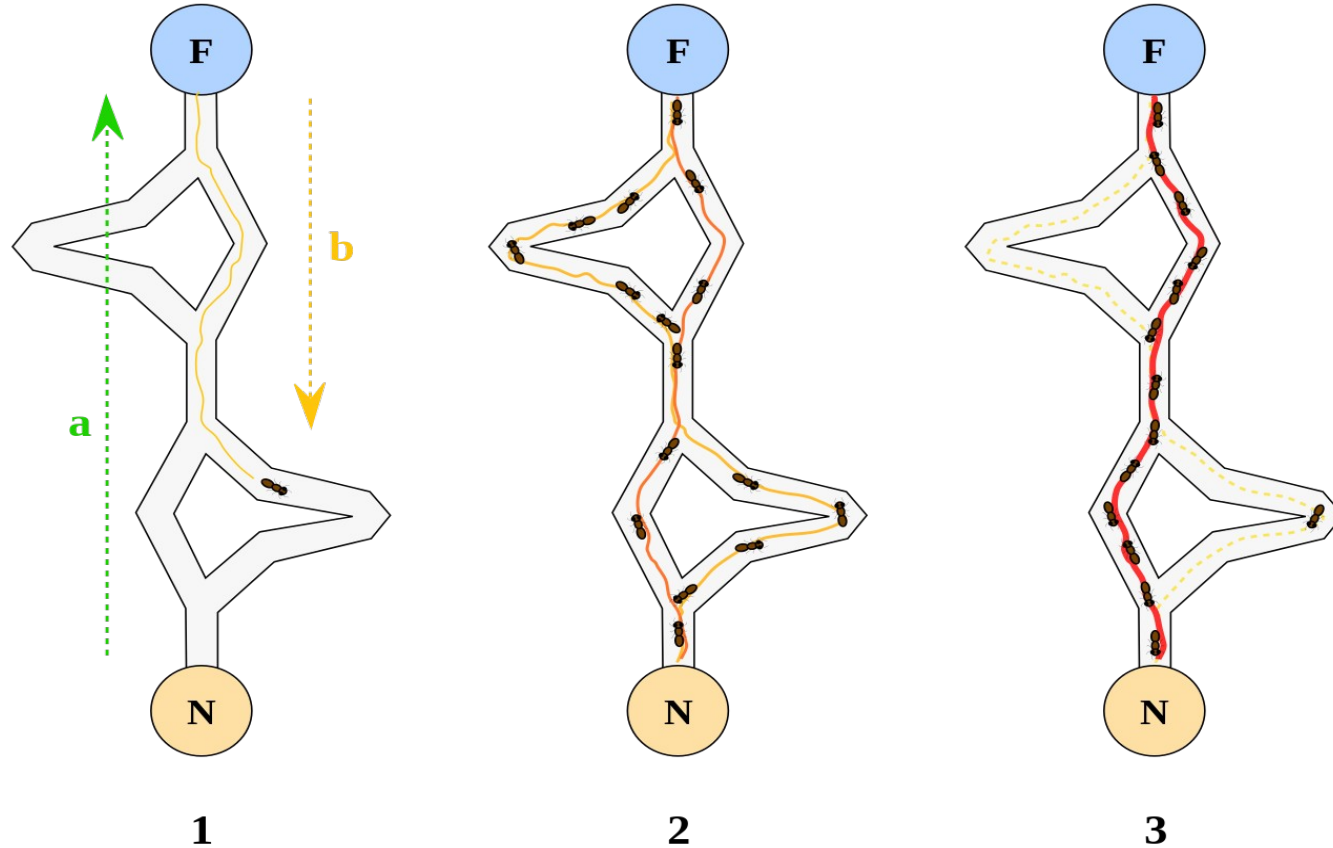
Optimierung
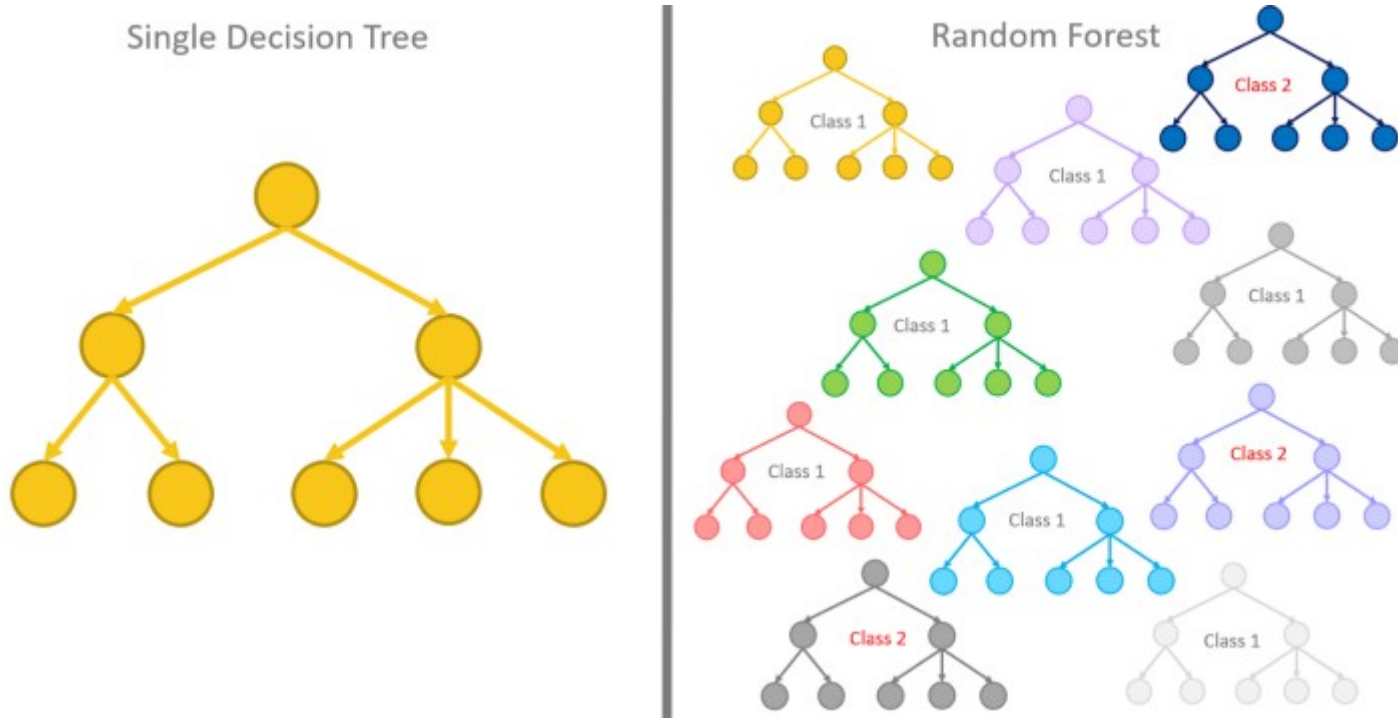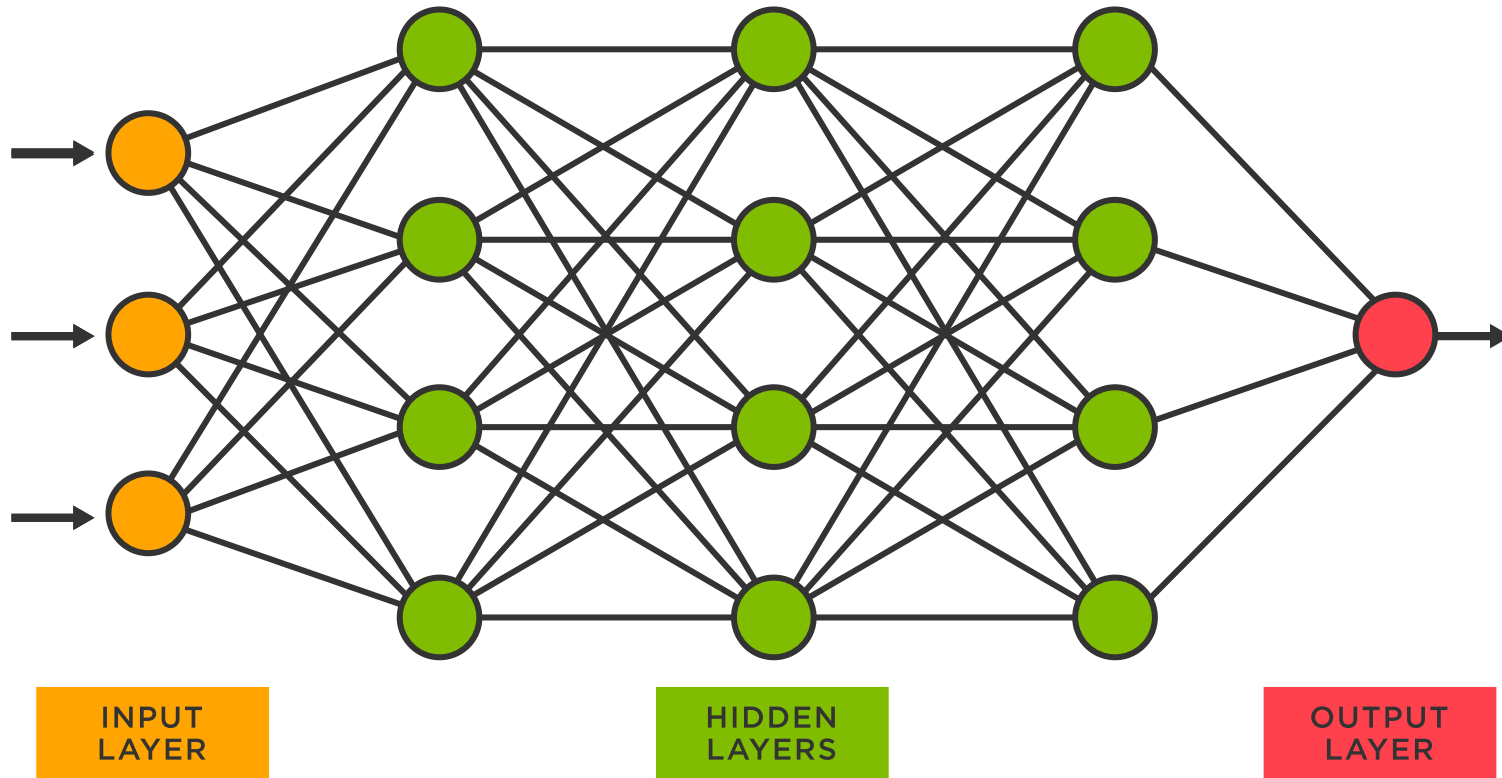
Clustering
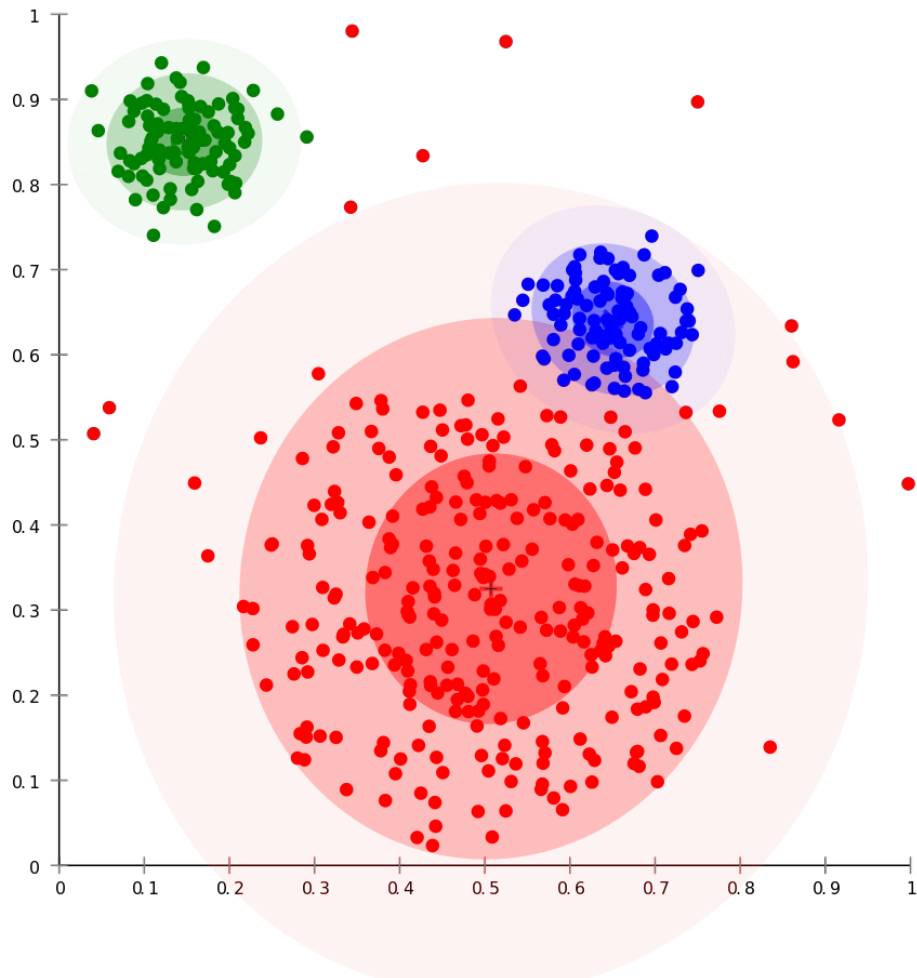
Machine Learning
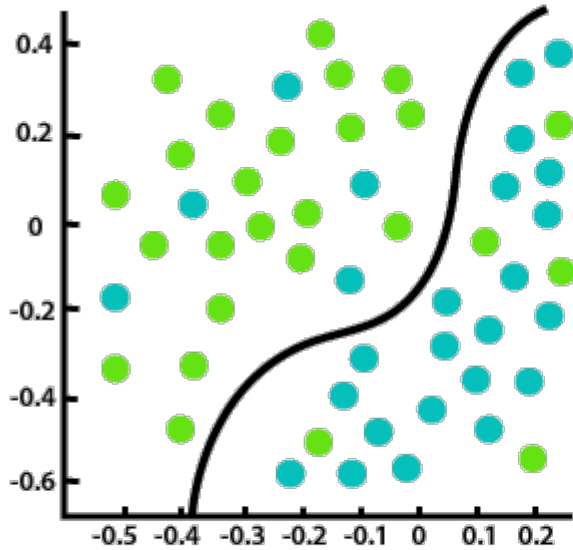
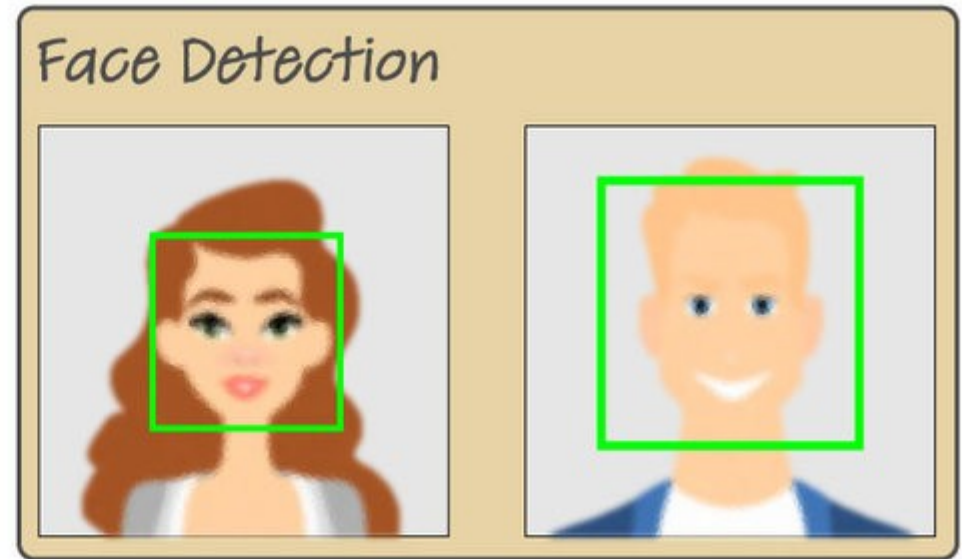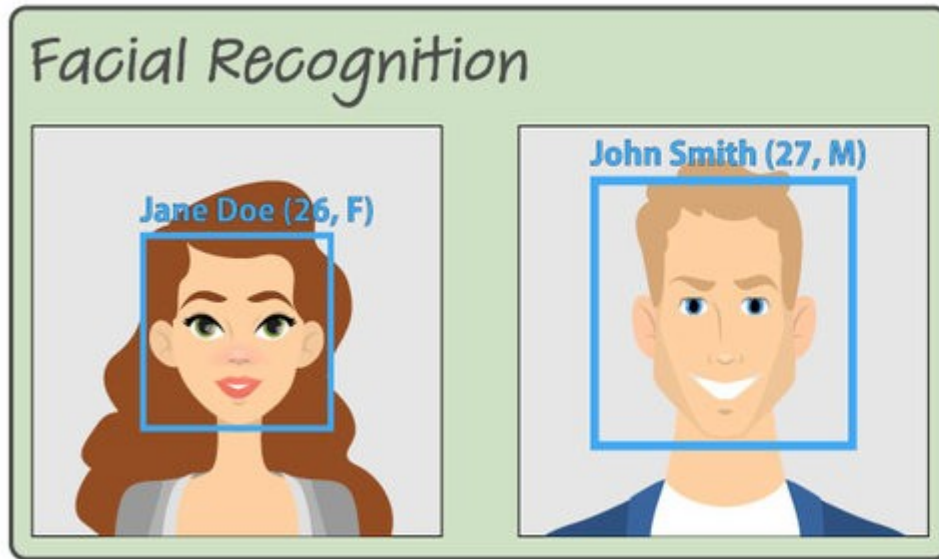Classification
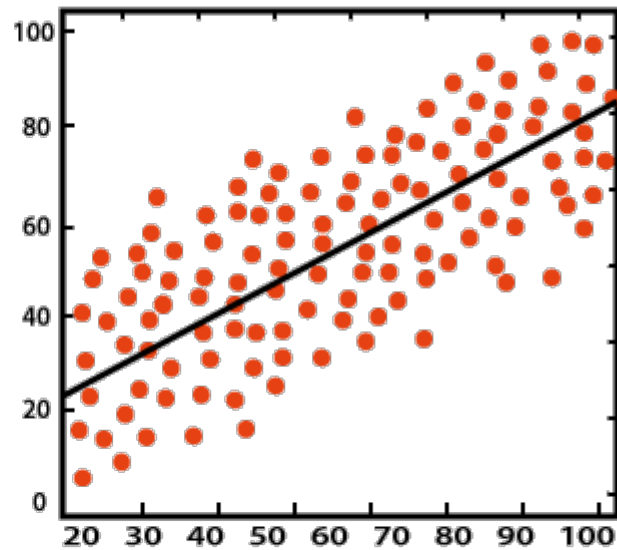
Regression

Decision Trees
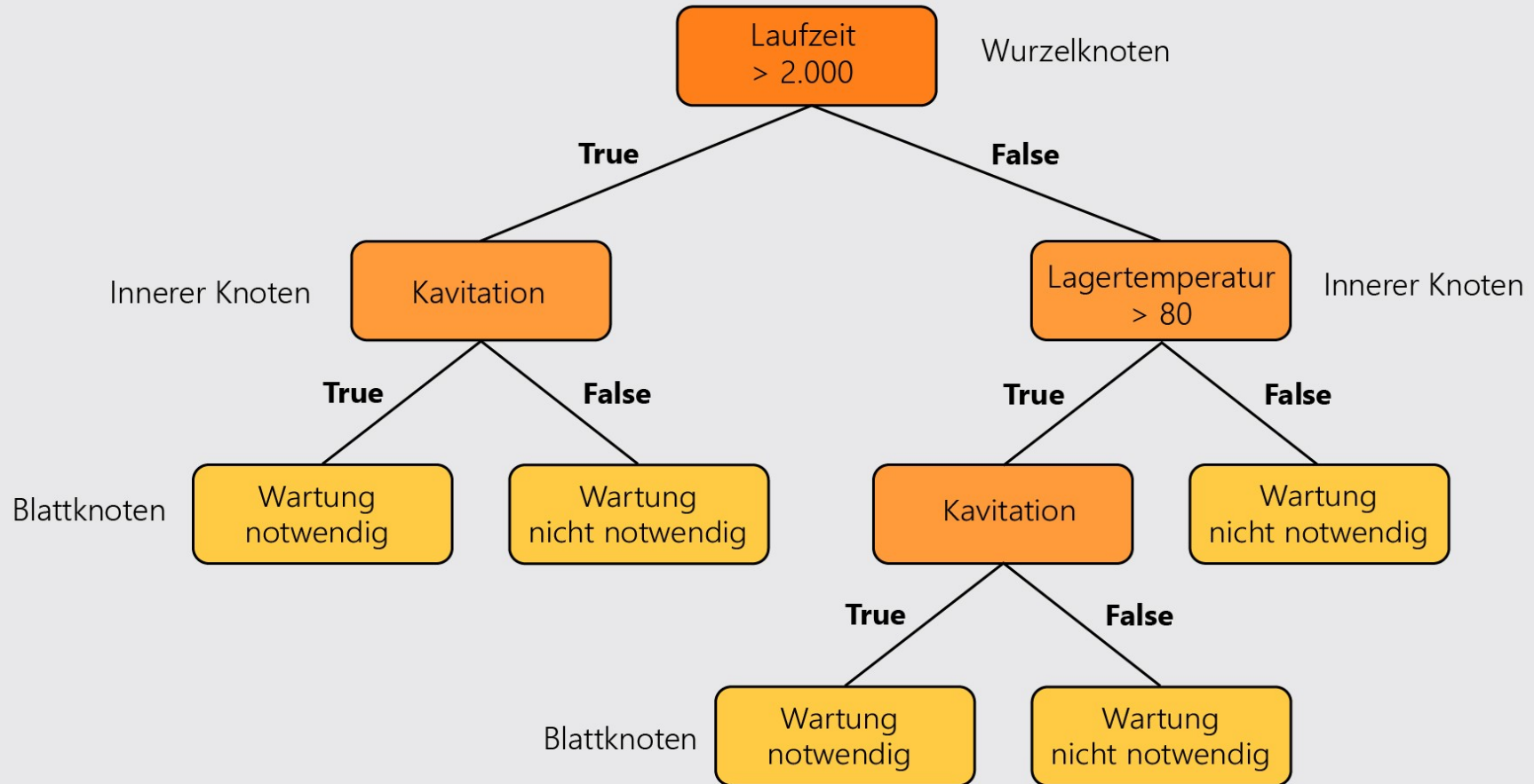
Random Forest, Ensemble Methods

Artificial Neural Network

INPUT
LAYER

HIDDEN
LAYERS

OUTPUT
LAYER

Face Detection

<Puppy> or <No Puppy>