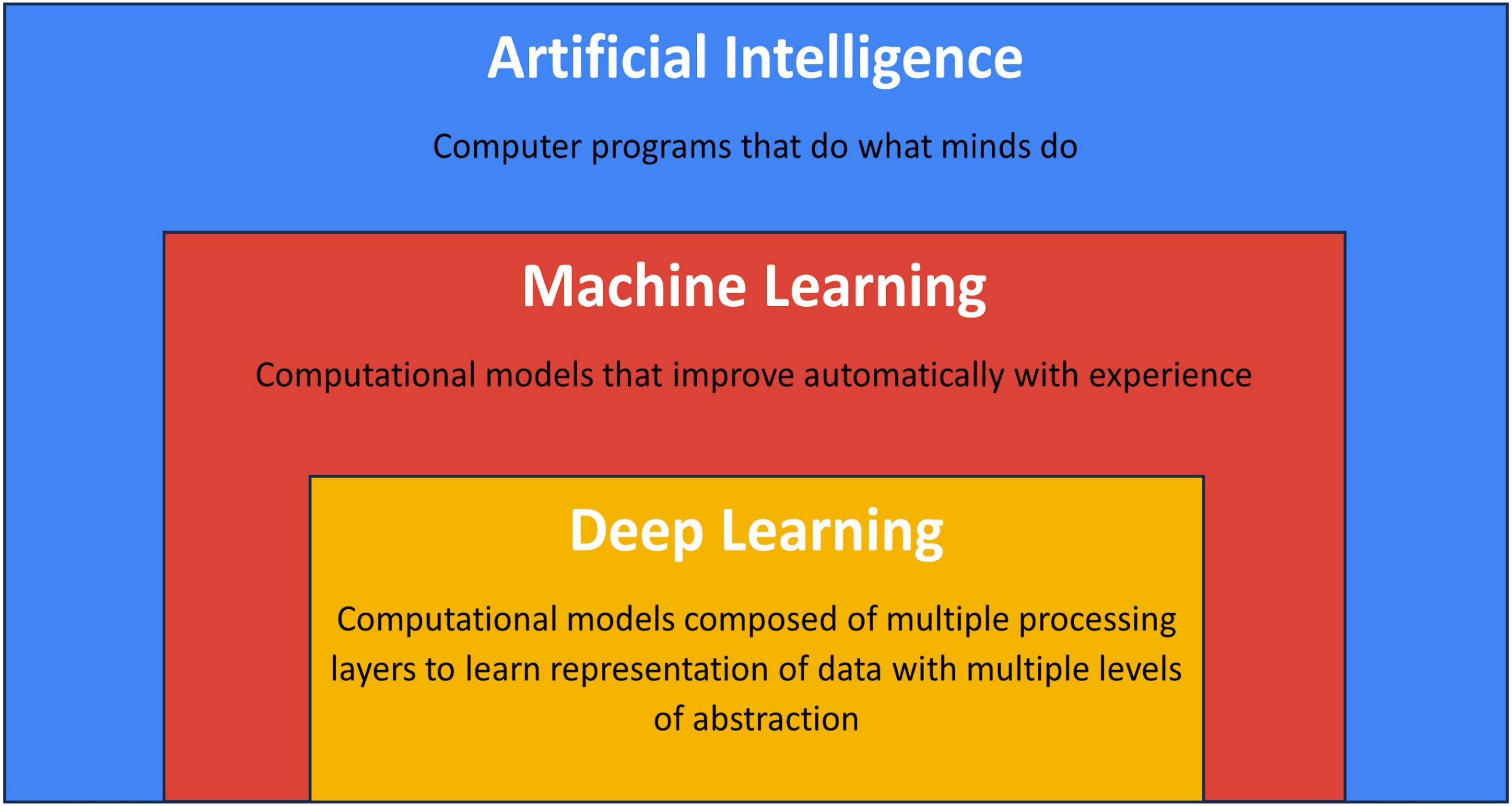
INTRODUCTION TO

Machine Learning


Artificial Intelligence
Computer programs that do what minds do
Machine Learning
Computational models that improve automatically with experience
Deep Learning
Computational models composed of multiple processing layers to learn representation of data with multiple levels of abstraction

What is Learning?

- Herbert Simon: “Learning is any process by which a system improves performance from experience.”
- What is the task?
  - Classification
  - Categorization/clustering
  - Problem solving / planning / control
  - Prediction
  - others

Why "Learn" ?

- Machine learning is programming computers to optimize a performance criterion using example data or past experience.
- There is no need to "learn" to calculate payroll
- Learning is used when:
  - Human expertise does not exist (navigating on Mars),
  - Humans are unable to explain their expertise (speech recognition)
  - Solution changes in time (routing on a computer network)
  - Solution needs to be adapted to particular cases (user biometrics)

What is Machine Learning?

Aspect of AI: creates knowledge

Definition:

“changes in [a] system that ... enable [it] to do the same task or tasks drawn from the same population more efficiently and more effectively <u>the next time</u>." (Simon 1983)

There are two ways that a system can improve:

1. By acquiring new knowledge
   - acquiring new facts
   - acquiring new skills
2. By adapting its behavior
   - solving problems more accurately
   - solving problems more efficiently

Workings in Machine Learning?

- Optimize a performance criterion using example data or past experience.
- Role of Statistics: Inference from a sample
- Role of Computer science: Efficient algorithms to
  - Solve the optimization problem
  - Representing and evaluating the model for inference

What We Talk About When We Talk About"Learning"

- Learning general models from a data of particular examples
- Data is cheap and abundant (data warehouses, data marts); knowledge is expensive and scarce.
- Example in retail: Customer transactions to consumer behavior:

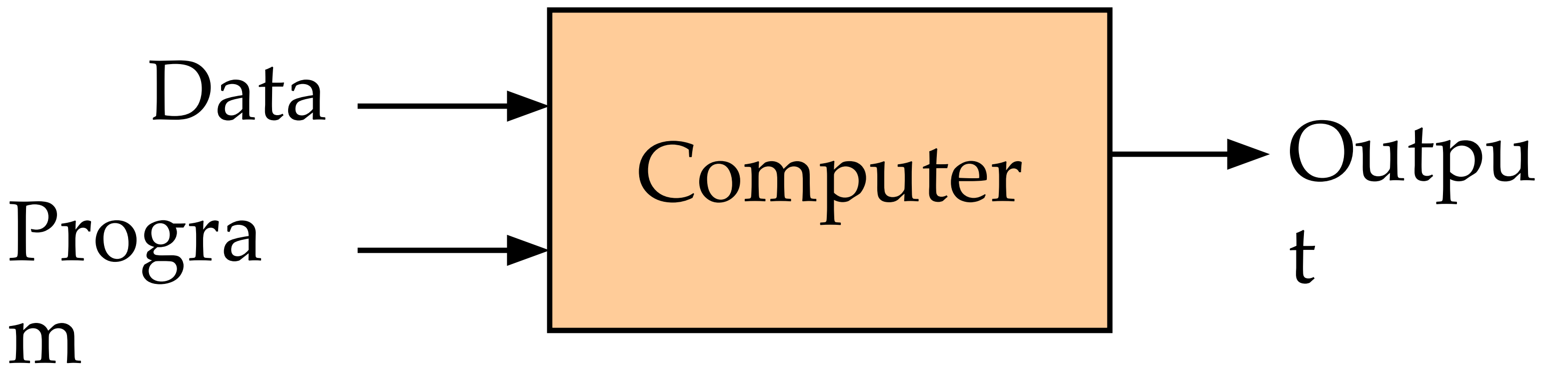
  People who bought "Da Vinci Code" also bought "The Five People You Meet in Heaven" (www.amazon.com)

- Build a model that is *a good and useful approximation* to the data.

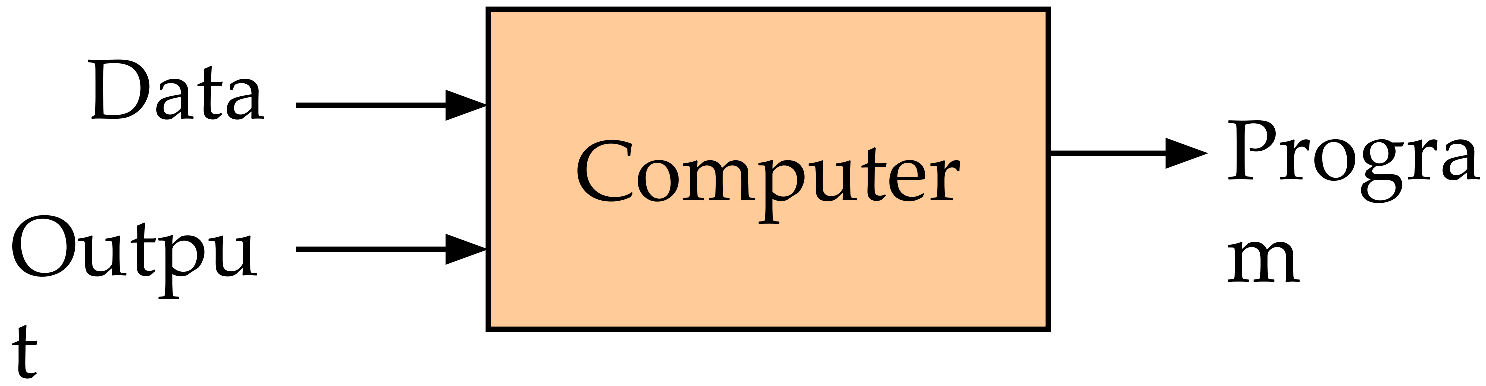
Data Mining

- Retail: Market basket analysis, Customer relationship management (CRM)
- Finance: Credit scoring, fraud detection
- Manufacturing: Optimization, troubleshooting
- Medicine: Medical diagnosis
- Telecommunications: Quality of service optimization
- Bioinformatics: Motifs, alignment
- Web mining: Search engines
- ...

Traditional Programming



Machine Learning



Areas of Influence for Machine Learning

1. **Statistics:** How best to use samples drawn from unknown probability distributions to help decide from which distribution some new sample is drawn?
2. **Brain Models:** Non-linear elements with weighted inputs (Artificial Neural Networks) have been suggested as simple models of biological neurons.
3. **Adaptive Control Theory:** How to deal with controlling a process having unknown parameters that must be estimated during operation?
4. **Psychology:** How to model human performance on various learning tasks?
5. **Artificial Intelligence:** How to write algorithms to acquire the knowledge humans are able to acquire, at least, as well as humans?
6. **Evolutionary Models:** How to model certain aspects of biological evolution to improve the performance of computer programs?

Related Disciplines

- Artificial Intelligence
- Data Mining
- Probability and Statistics
- Information theory
- Numerical optimization
- Computational complexity theory
- Control theory (adaptive)
- Psychology (developmental, cognitive)
- Neurobiology
- Linguistics
- Philosophy

Types of Learning

- **Association**
- **Supervised (inductive) learning**
  - Training data includes desired outputs
- **Unsupervised learning**
  - Training data does not include desired outputs
- **Semi-supervised learning**
  - Training data includes a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

Learning Associations

- Basket analysis:

  $P(Y \mid X)$ probability that somebody who buys X also buys Y where X and Y are products/services.

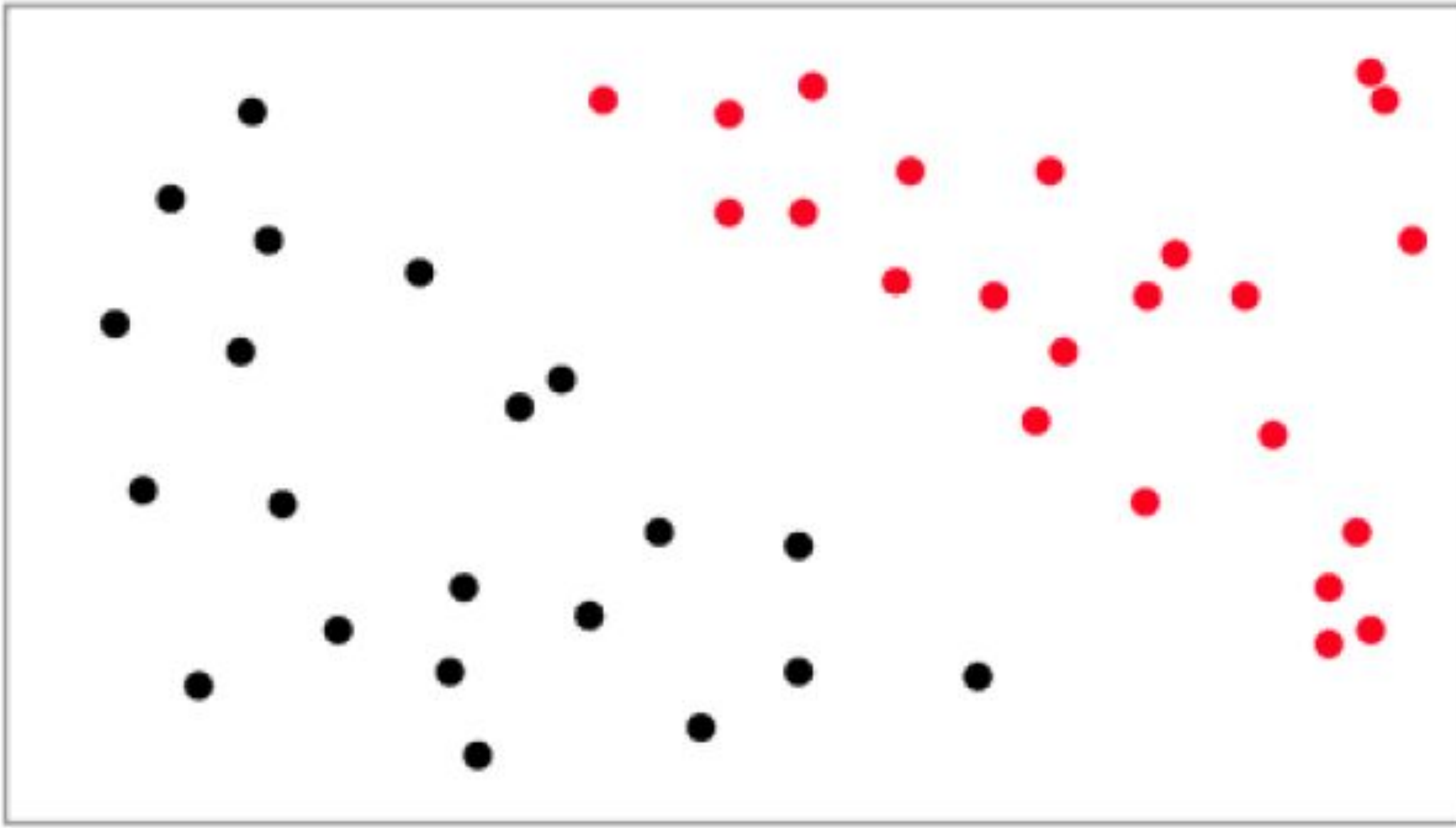
  Example: P (chips | beer) = 0.7

Supervised Learning

- **Given** examples of a function *(X, F(X))*
- **Predict** function *F(X)* for new examples *X*
  - Discrete *F(X)*: Classification
  - Continuous *F(X)*: Regression
  - *F(X)* = Probability(*X*): Probability estimation

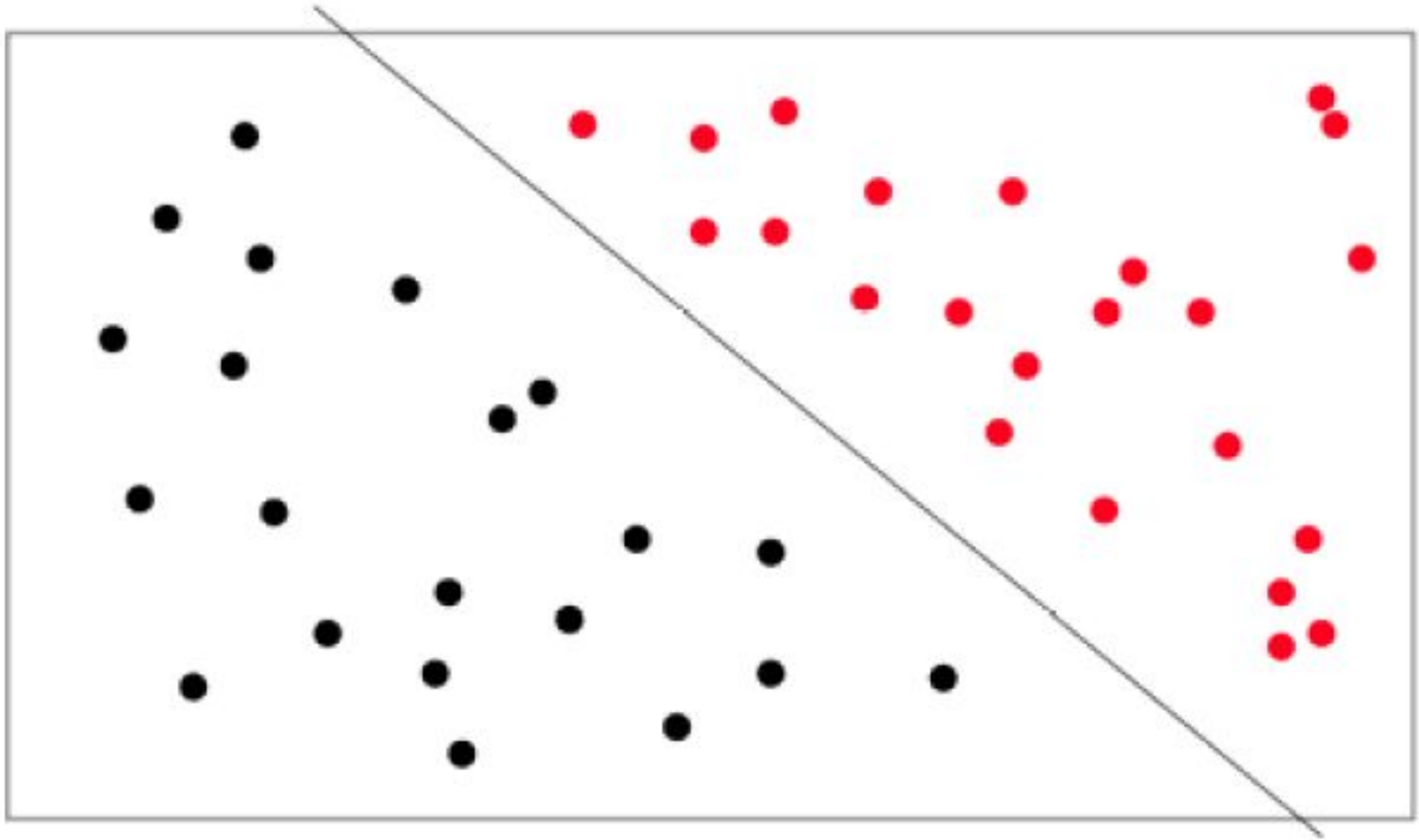
Supervised Learning:

What is the right Hypothesis?



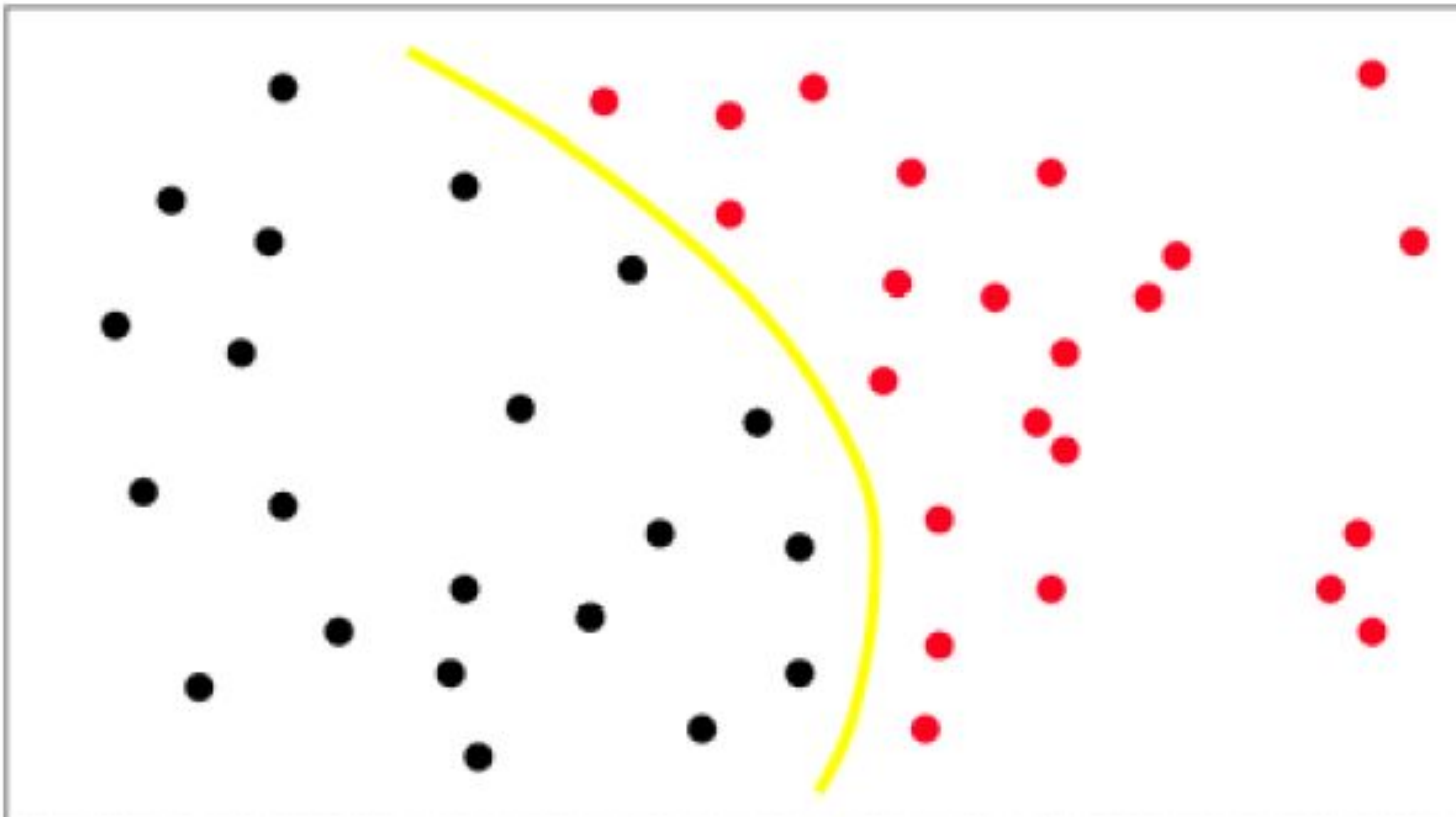
Supervised Learning:

Hypothesis – Linear Separation



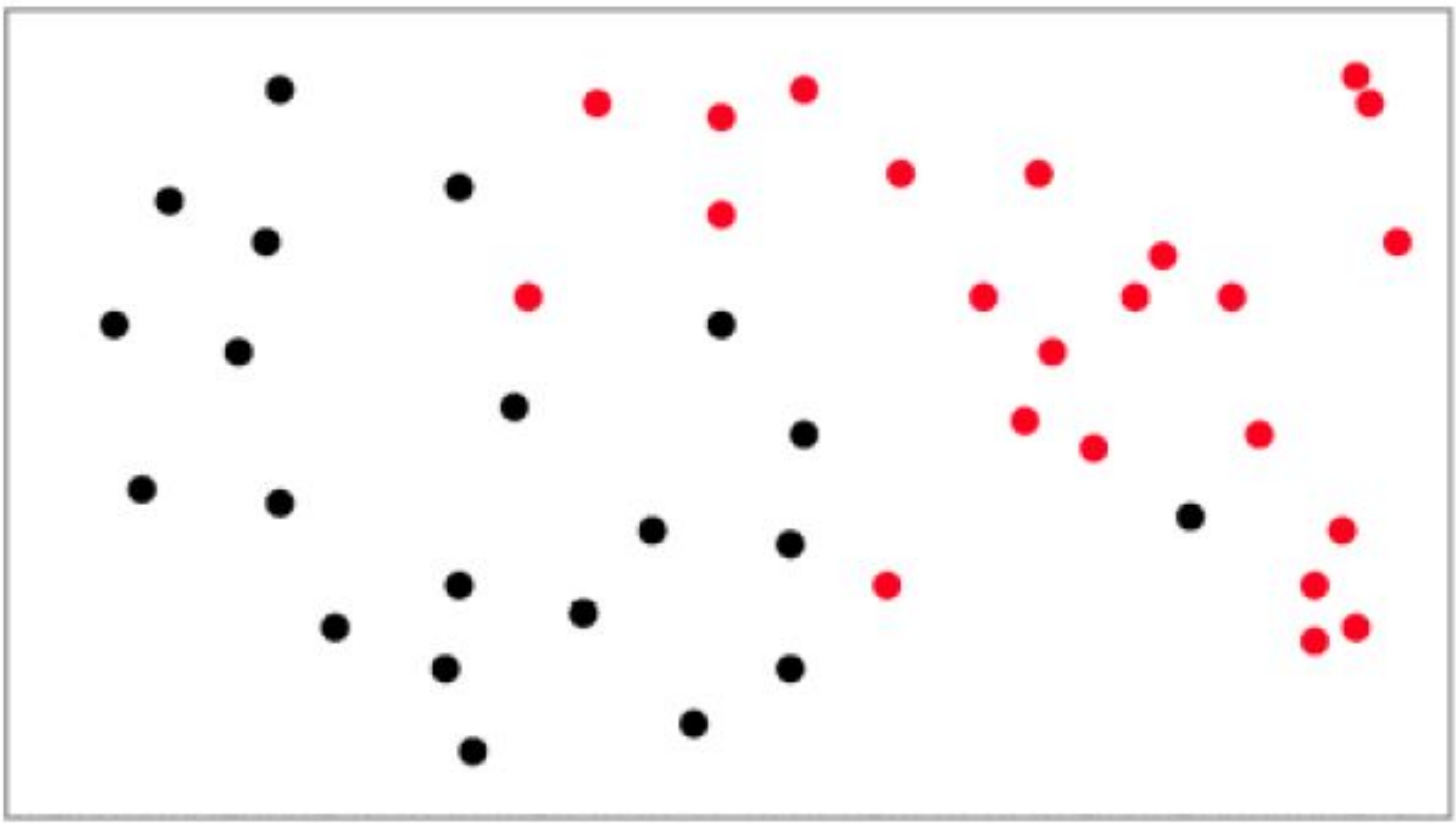
Supervised Learning:

Hypothesis – Quadratic Separation



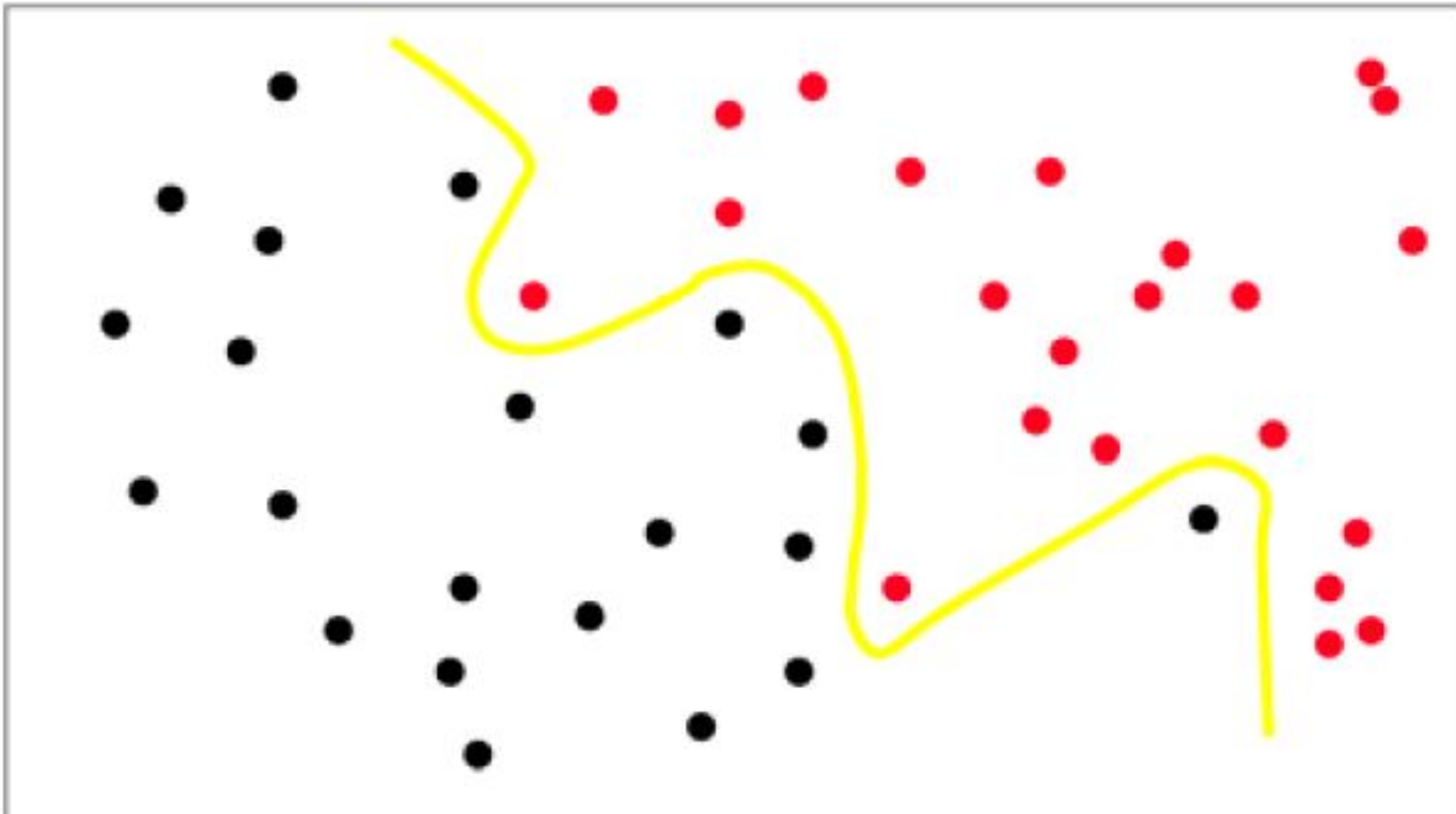
Supervised Learning:

Hypothesis – Noisy/Mislabeled Data



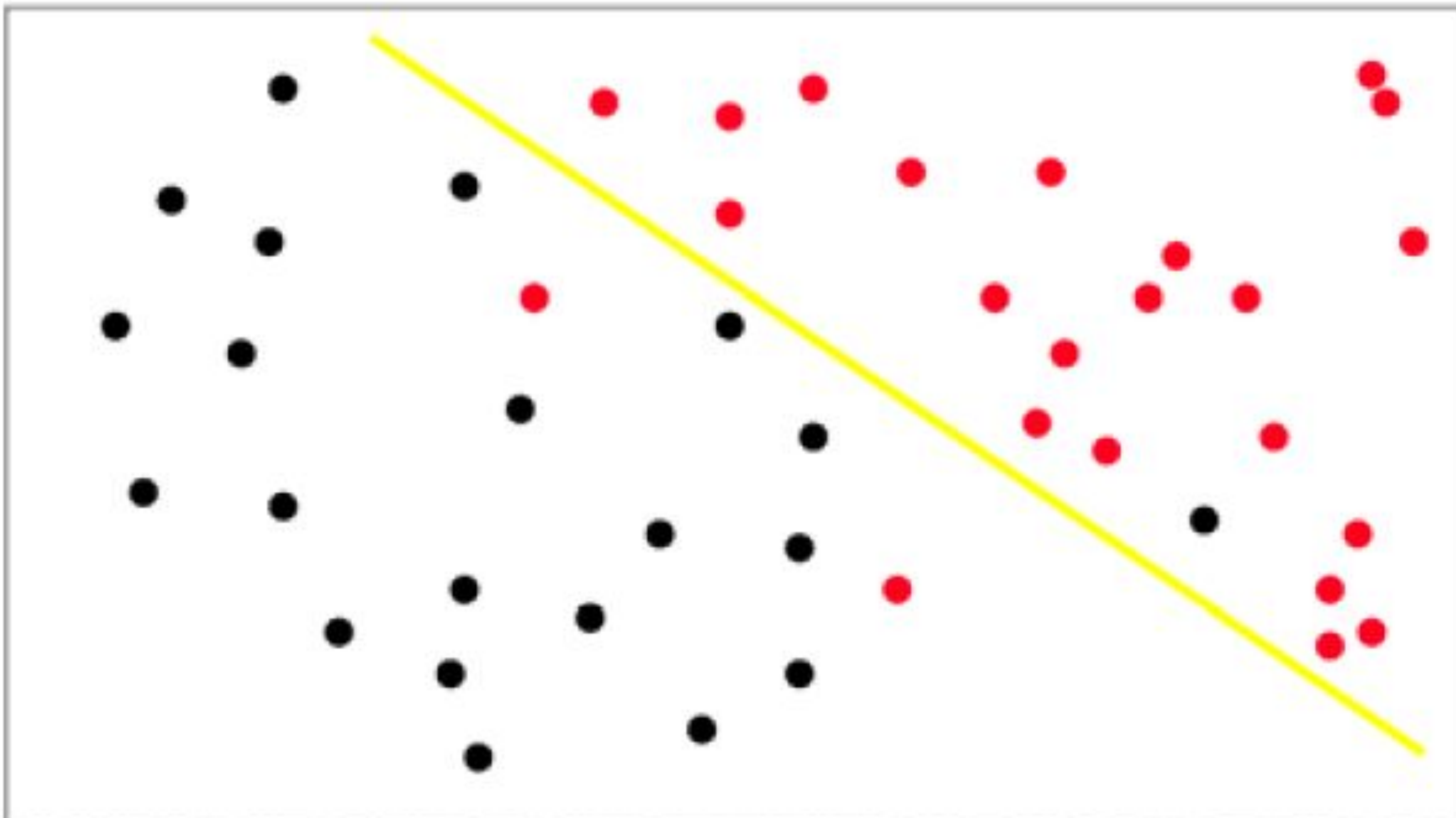
Supervised Learning:

Hypothesis – Overfitting



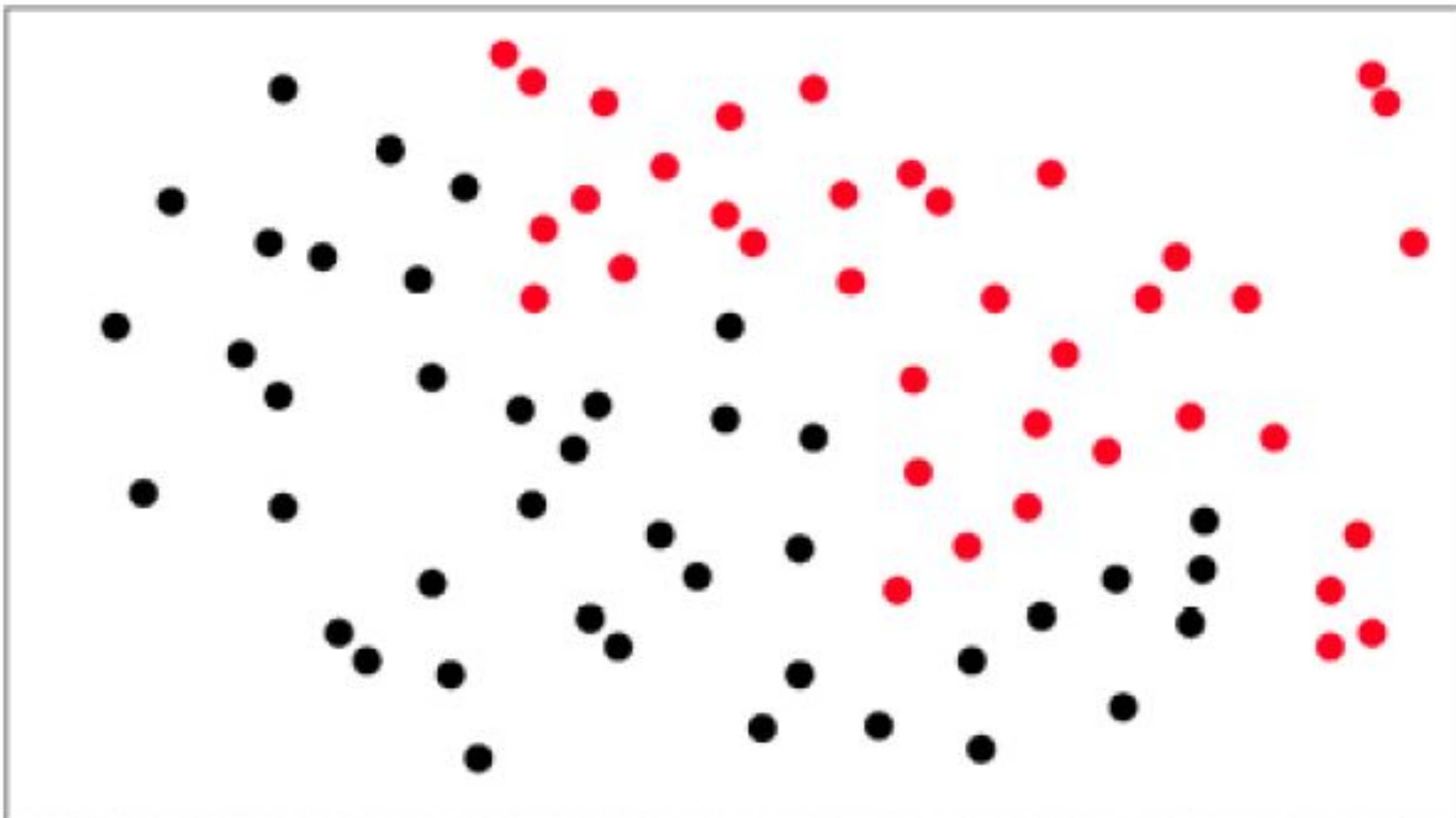
Supervised Learning:

Hypothesis – Underfitting?



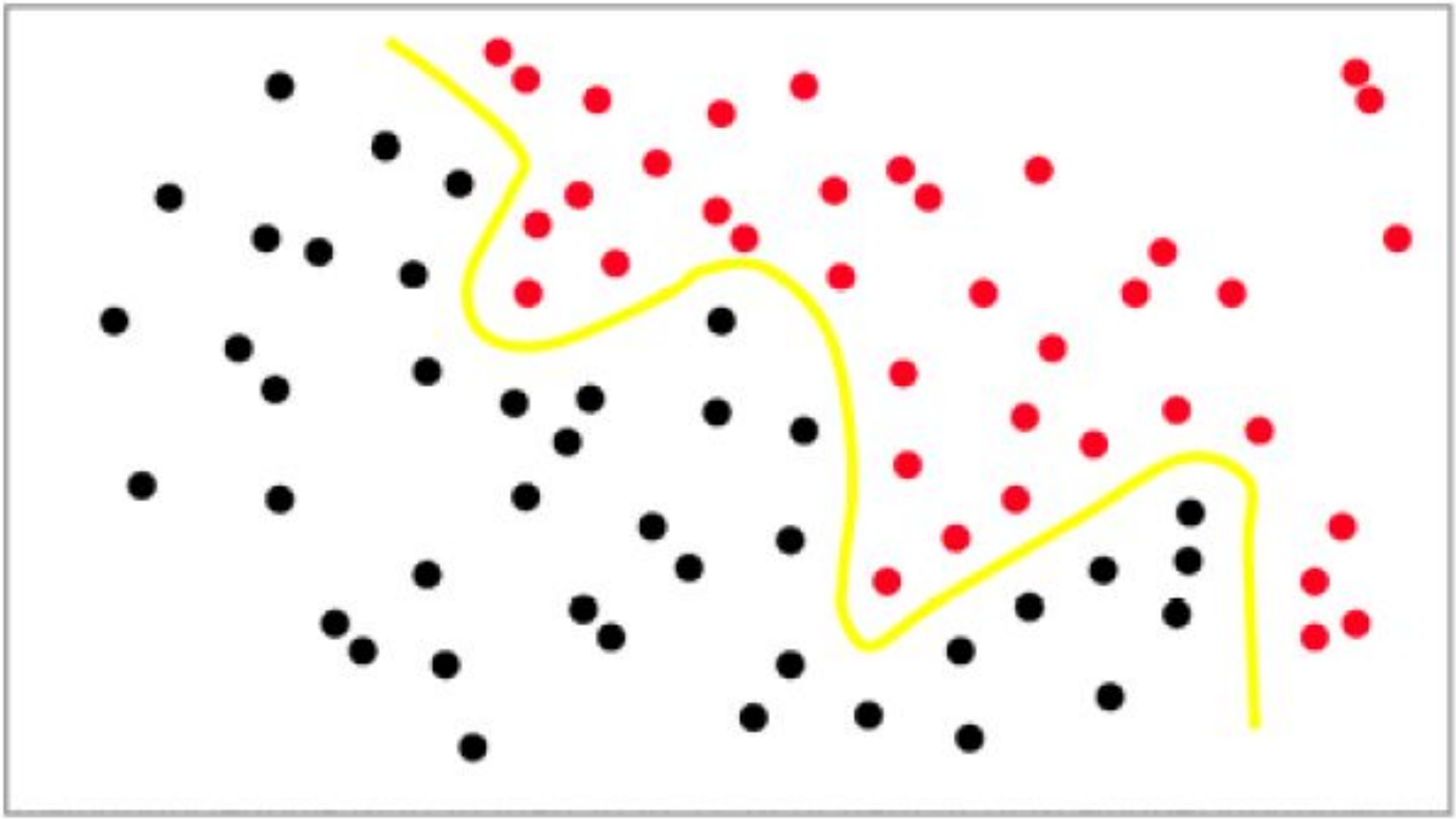
Supervised Learning:

Hypothesis – More data



Supervised Learning:

Hypothesis – More complex

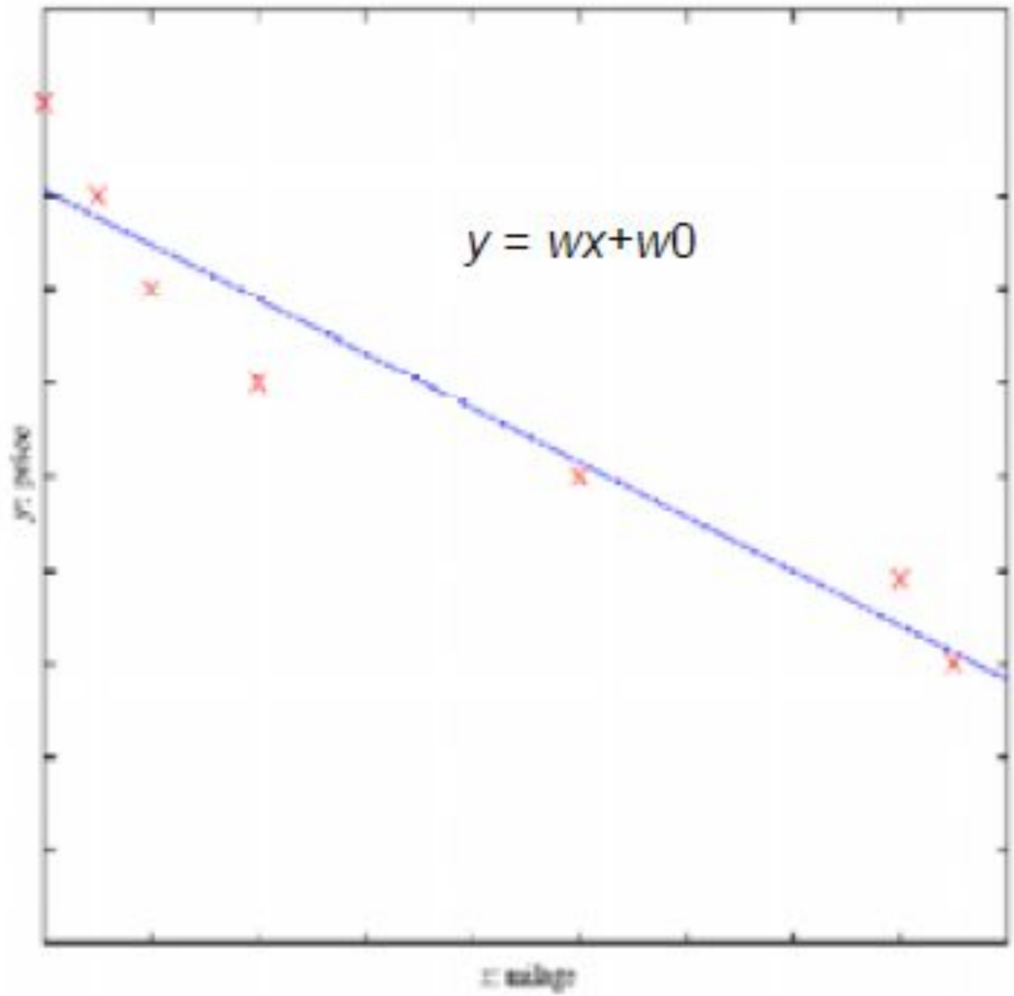


Supervised Learning:

Linear Regression

- Example:
  Price of a used car
  x : car attribute
  y : price

- $y = g(x \mid \theta)$
  model:
  $g(\ )$
  parameters:
  $\theta = (w, w_0)$



Regression

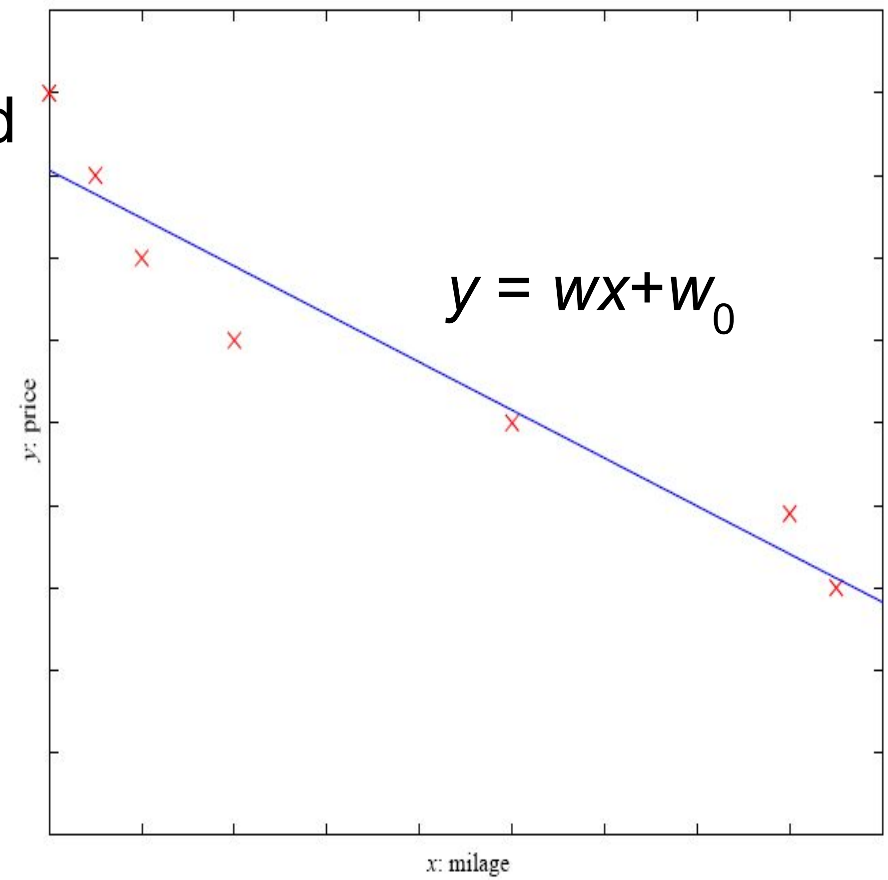
- Example: Price of a used car
- x : car attributes

  y : price

  $y = g(x \mid \theta)$

  $g(\,)$ model,

  θ parameters



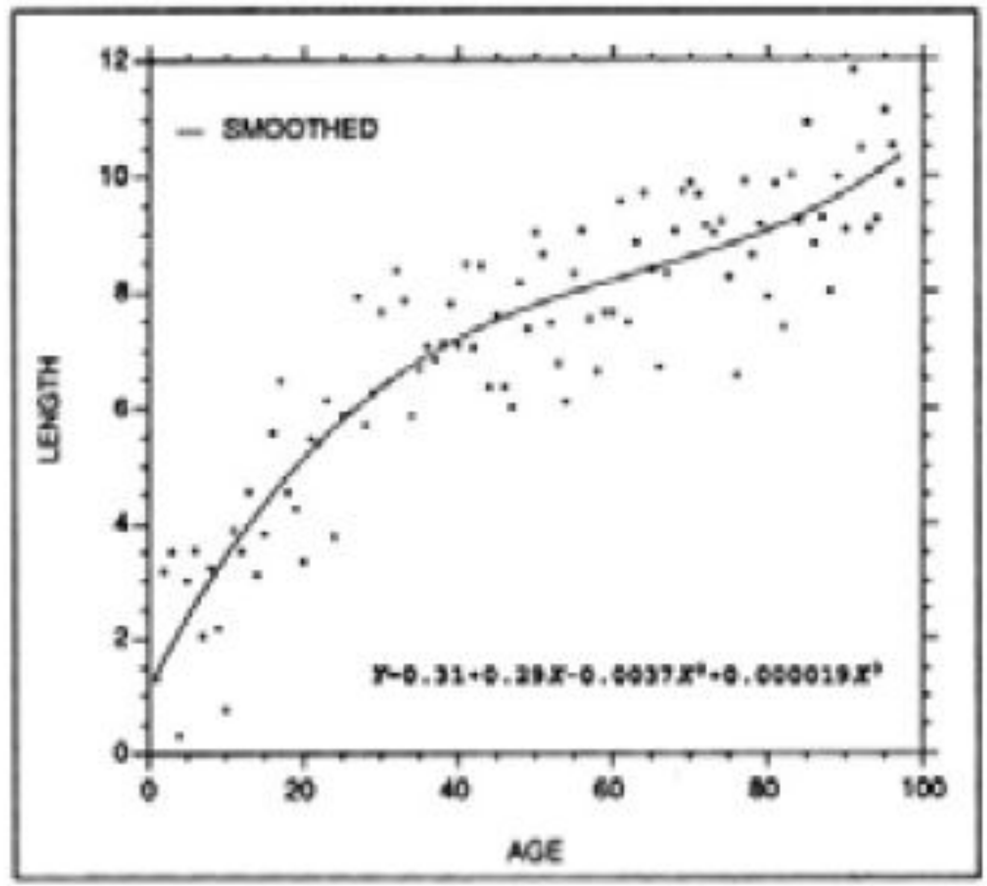
$y = wx+w_0$

Supervised Learning:

Polynomial Regression

- Example:
  Growth of a species
  x : age
  y : length

- $y = g\,(x \mid \theta)$
  model:
  $g\,(\,)$
  parameters:
  $\theta = (w_3, w_2, w_1, w_0)$



Supervised Learning:

Some Regression Applications

- Cost estimation
  - Energy consumption
- Control
  - Angle of steering wheel for robot car
  - Kinematics of a robot arm
- Predicted response
  - Surface materials

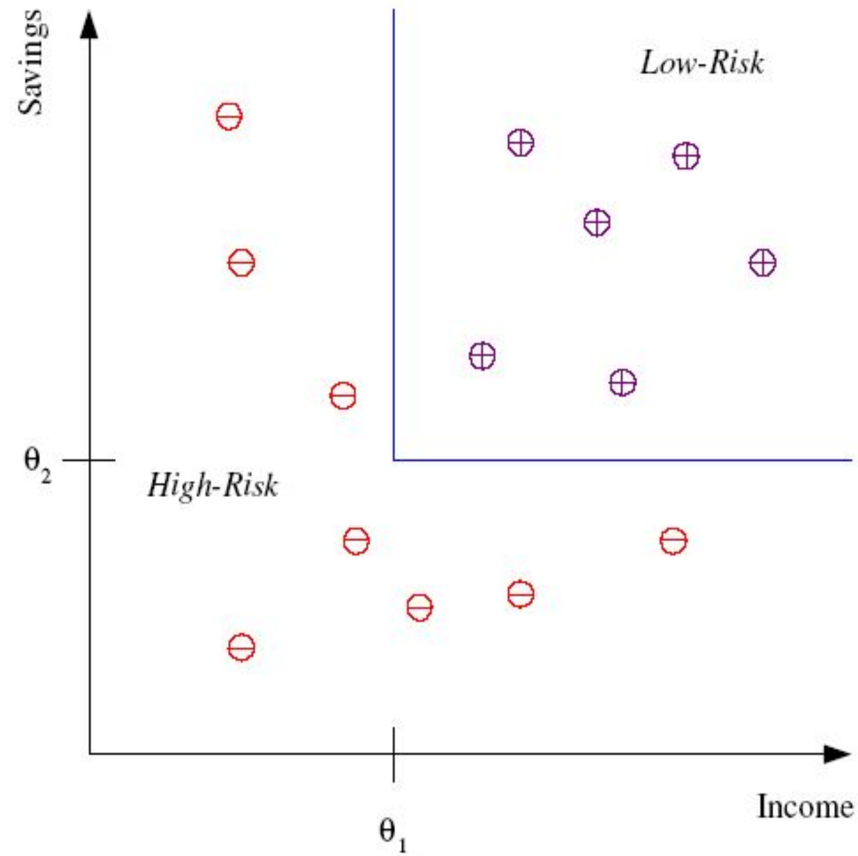
Supervised Learning:

Range of Methods

- Methods differ in terms of
  - The form of hypothesis space
  - The way to find best hypothesis given data
- There are many successful approaches
  - Decision trees
  - Support vector machines
  - Neural networks
  - Case-based reasoning
  - ...

Classification

- Example: Credit scoring
- Differentiating between low-risk and high-risk customers from their *income* and *savings*



Discriminant: IF *income* > θ_1 AND *savings* > θ_2
THEN low-risk ELSE high-risk

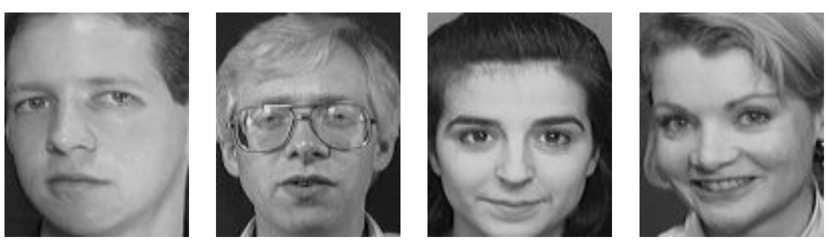
Classification: Applications

- Aka Pattern recognition
- Face recognition: Pose, lighting, occlusion (glasses, beard), make-up, hair style
- Character recognition: Different handwriting styles.
- Speech recognition: Temporal dependency.
  - Use of a dictionary or the syntax of the language.
  - Sensor fusion: Combine multiple modalities; eg, visual (lip image) and acoustic for speech
- Medical diagnosis: From symptoms to illnesses
- ...

Face Recognition

Training examples of a person

Test images



AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

Supervised Learning: Uses

- Prediction of future cases: Use the rule to predict the output for future inputs
- Knowledge extraction: The rule is easy to understand
- Compression: The rule is simpler than the data it explains
- Outlier detection: Exceptions that are not covered by the rule, e.g., fraud

Unsupervised Learning

- Learning “what normally happens”
- No output
- Clustering: Grouping similar instances
- Example applications
  - Customer segmentation in CRM
  - Image compression: Color quantization
  - Bioinformatics: Learning motifs

Unsupervised Learning:

Image Clustering

Reinforcement Learning

- Learning a policy: A sequence of outputs
- No supervised output but delayed reward
- Credit assignment problem
  - Which action led me to winning the game?

- Examples
  - Game playing
  - Robot in a maze
  - Multiple agents, partial observability, ...

Reinforcement Learning:
Overview

- Characteristics
  - Learning a Policy: A sequence of outputs
  - No supervised output, but a delayed reward
  - Credit assignment problem:
    - Which action led me to winning the game?
- Examples
  - Elevator scheduling
  - Backgammon and Chess
  - Robot control

Some more examples of tasks that are best solved by using a learning algorithm

1. **Recognizing patterns:**
   - Facial identities or facial expressions
   - Handwritten or spoken words
   - Medical images
2. **Generating patterns:**
   - Generating images or motion sequences
3. **Recognizing anomalies:**
   - Unusual sequences of credit card transactions
   - Unusual patterns of sensor readings in a nuclear power plant or unusual sound in your car engine.
4. **Prediction:**
   - Future stock prices or currency exchange rates

ML problems

1. **The web contains a lot of data.**
2. **Tasks with very big datasets often use machine learning**
   - especially if the data is noisy or non-stationary.
3. **Spam filtering, fraud detection:**
   - The enemy adapts so we must adapt too.
4. **Recommendation systems:**
   - Lots of noisy data. Million dollar prize!
5. **Information retrieval:**
   - Find documents or images with similar content.
6. **Data Visualization:**
   - Display a huge database in a revealing way

Sample Applications

- Web search
- Computational biology
- Finance
- E-commerce
- Space exploration
- Robotics
- Information extraction
- Social networks
- Debugging
- [Your favorite area]

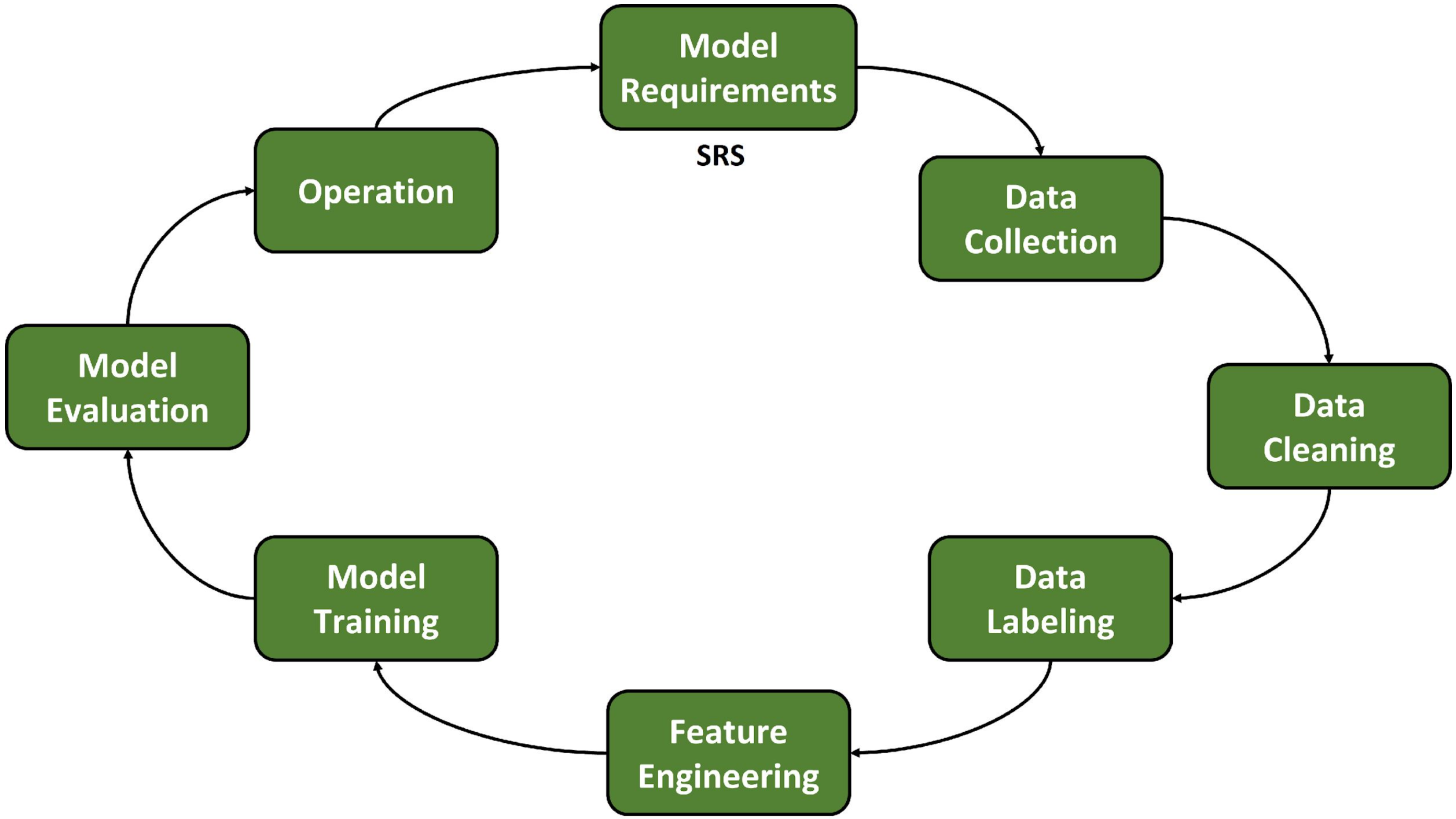
ML in a Nutshell

- Tens of thousands of machine learning algorithms
- Hundreds new every year
- Every machine learning algorithm has three components:
  - **Representation**
  - **Evaluation**
  - **Optimization**

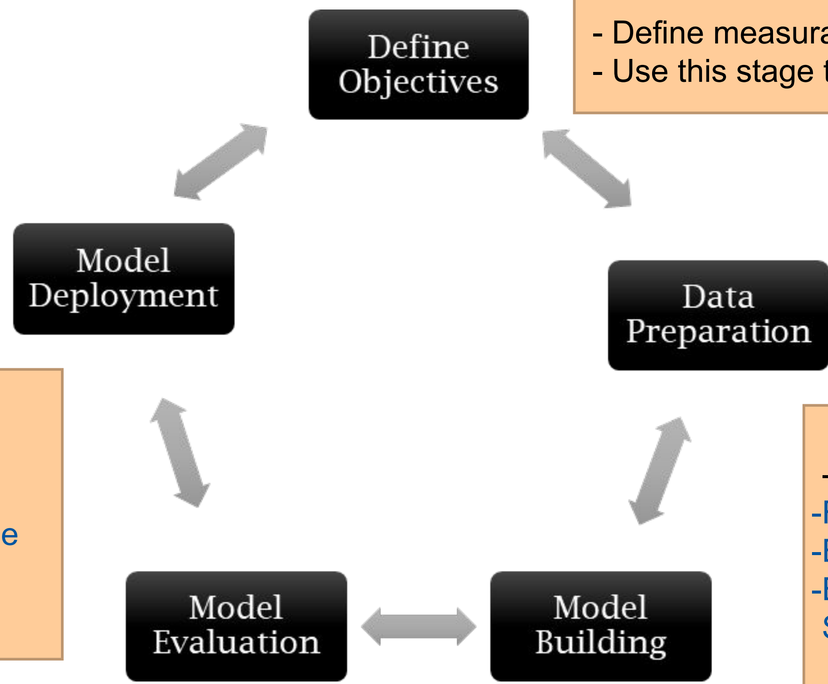
Usual ML stages

- Hypothesis, data
- Training or learning
- Testing or generalization

ML lifecyle


Model Requirements
SRS
Data Collection
Data Cleaning
Data Labeling
Feature Engineering
Model Training
Model Evaluation
Operation

Machine Learning as a Process



- Define measurable and quantifiable goals
- Use this stage to learn about the problem

- Normalization
- Transformation
- Missing Values
- Outliers

- Data Splitting
- Features Engineering
- Estimating Performance
- Evaluation and Model Selection

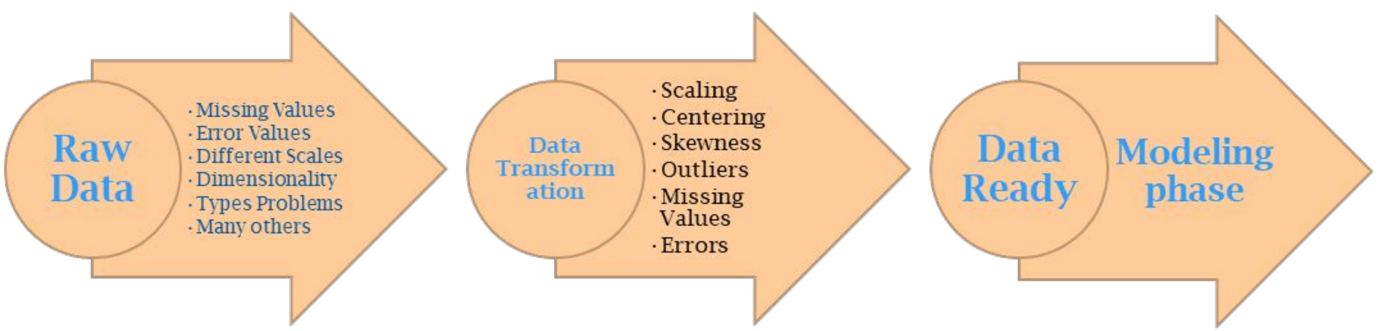
- Study models accuracy
  - Work better than the naïve approach or previous system
- Do the results make sense in the context of the problem

Representation

- Decision trees
- Sets of rules / Logic programs
- Instances
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles
- Etc.

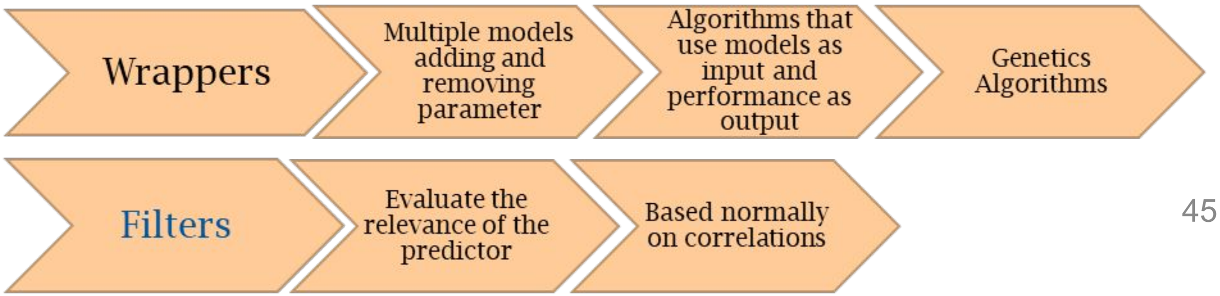
ML as a Process: Data Preparation

- Needed for several reasons
  - Some Models have strict data requirements
    - Scale of the data, data point intervals, etc
  - Some characteristics of the data may impact dramatically on the model performance
- Time on data preparation should not be underestimated



ML as a Process: Feature engineering

- Determine the predictors (features) to be used is one of the most critical questions
- Sometimes we need to add predictors
- Reduce Number:
  - Fewer predictors more interpretable model and less costly
  - Most of the models are affected by high dimensionality, specially for non-informative predictors



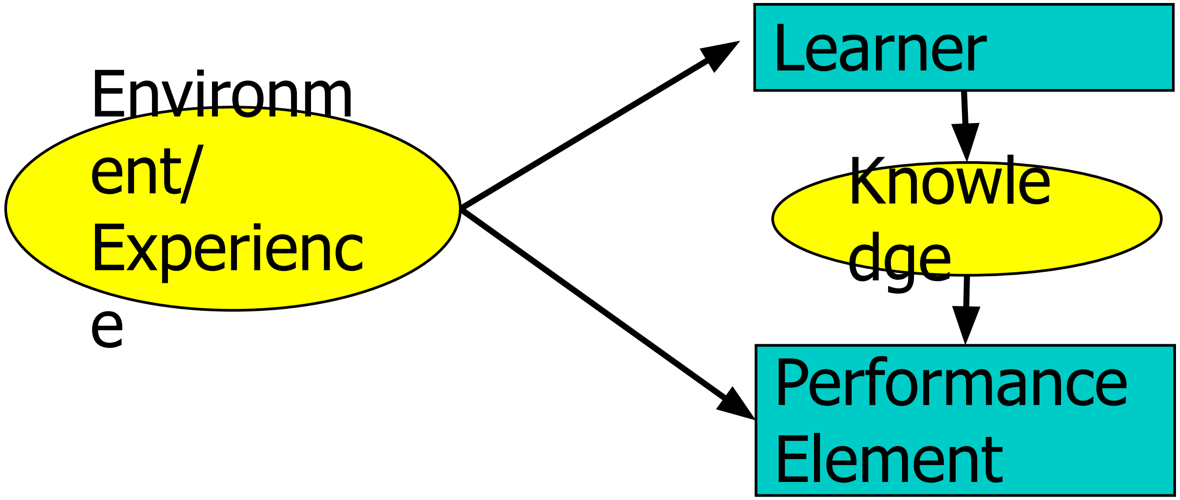
  - Binning predictors

ML as a Process: Model Building

- Data Splitting
  - Allocate data to different tasks
    - model training
    - performance evaluation
  - Define Training, Validation and Test sets
- Feature Selection (Review the decision made previously)
- Estimating Performance
  - Visualization of results – discovery interesting areas of the problem space
  - Statistics and performance measures
- Evaluation and Model selection
  - The ‘no free lunch’ theorem no a priory assumptions can be made
  - Avoid use of favorite models if NEEDED

Designing a Learning System

- Choose the training experience
- Choose exactly what is too be learned, i.e. the ***target function***.
- Choose how to represent the target function.
- Choose a learning algorithm to infer the target function from the experience.



Types of testing

- Evaluate performance by testing on data NOT used for testing (both should be randomly sampled)
- Cross validation methods for small data sets
- The more (relevant) data the better.

Testing

- How well the learned system work?
- Generalization
  - Performance on unseen or unknown scenarios or data
  - Brittle vs robust performance

Training vs. Test Distribution

- Generally assume that the training and test examples are independently drawn from the same overall distribution of data.
  - IID: Independently and identically distributed
- If examples are not independent, requires ***collective classification***.
- If test distribution is different, requires ***transfer learning***.

Evaluation

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- Etc.

Measuring Performance

- Generalization accuracy
- Solution correctness
- Solution quality (length, efficiency)
- Speed of performance

Scaling issues in ML

- Number of
  - Inputs
  - Outputs
  - Batch vs realtime
  - Training vs testing

Optimization

- Combinatorial optimization
  - E.g.: Greedy search
- Convex optimization
  - E.g.: Gradient descent
- Constrained optimization
  - E.g.: Linear programming

ML in Practice

- Understanding domain, prior knowledge, and goals
- Data integration, selection, cleaning,
  pre-processing, etc.
- Learning models
- Interpreting results
- Consolidating and deploying discovered knowledge
- Loop

Machine Learning versus Human Learning

- Some ML behavior can challenge the performance of human experts (e.g., playing chess)
- Although ML sometimes matches human learning capabilities, it is not able to learn as well as humans or in the same way that humans do
- There is no claim that machine learning can be applied in a truly creative way
- Formal theories of ML systems exist but are often lacking (why a method succeeds or fails is not clear)
- ML success is often attributed to manipulation of symbols (rather than mere numeric information)

Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive: http://kdd.ics.uci.edu/summary.data.application.html
- Statlib: http://lib.stat.cmu.edu/
- Delve: http://www.cs.utoronto.ca/~delve/

Resources: Journals

- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
  - ICML05: http://icml.ais.fraunhofer.de/
- European Conference on Machine Learning (ECML)
  - ECML05: http://ecmlpkdd05.liacc.up.pt/
- Neural Information Processing Systems (NIPS)
  - NIPS05: http://nips.cc/
- Uncertainty in Artificial Intelligence (UAI)
  - UAI05: http://www.cs.toronto.edu/uai2005/
- Computational Learning Theory (COLT)
  - COLT05: http://learningtheory.org/colt2005/
- International Joint Conference on Artificial Intelligence (IJCAI)
  - IJCAI05: http://ijcai05.csd.abdn.ac.uk/
- International Conference on Neural Networks (Europe)
  - ICANN05: http://www.ibspan.waw.pl/ICANN-2005/
- ...