Introduction to Machine Learning

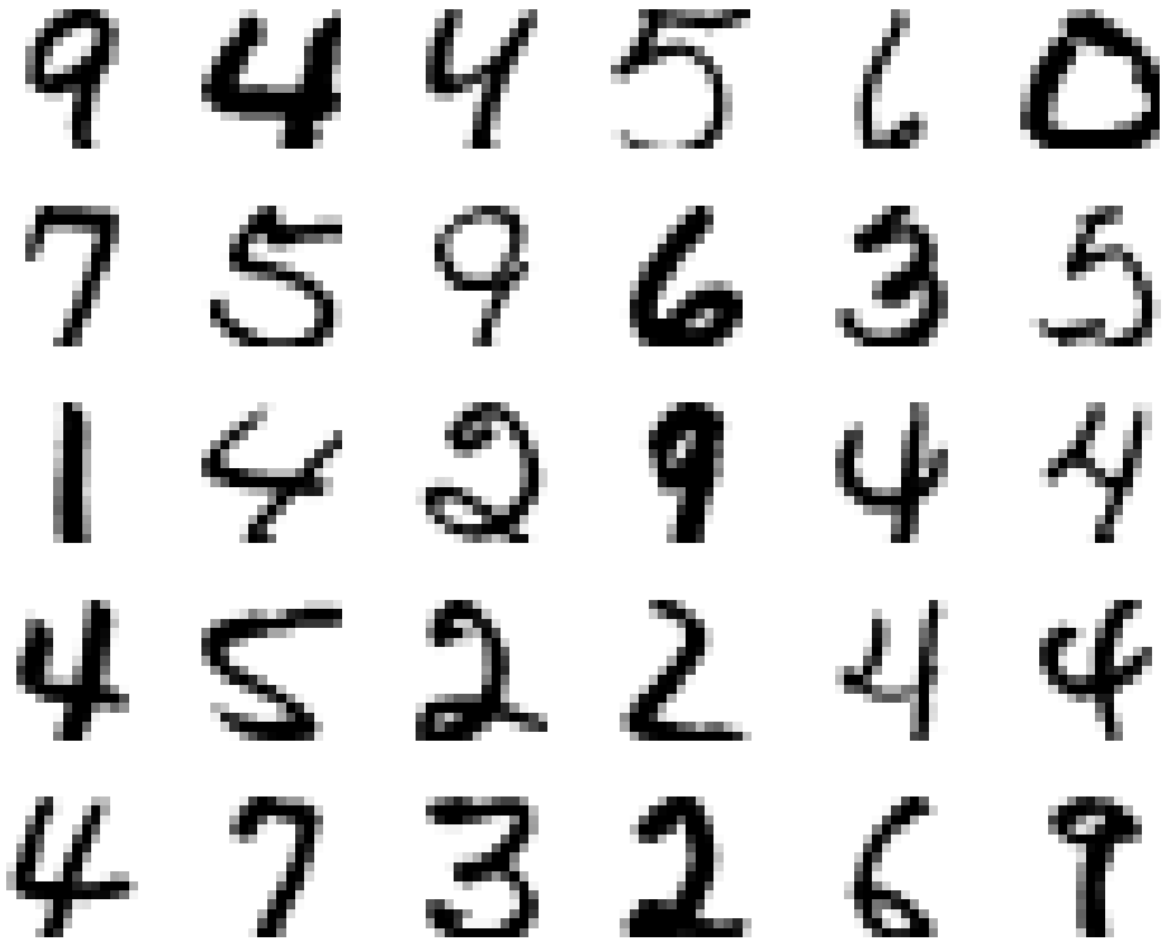
Klaus-Robert Müller, Mikio L. Braun
TU Berlin

October 15, 2008

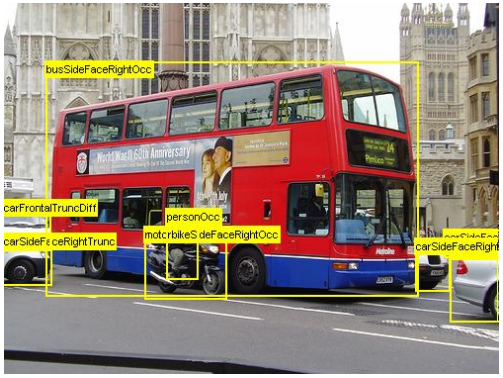
What is Machine Learning?

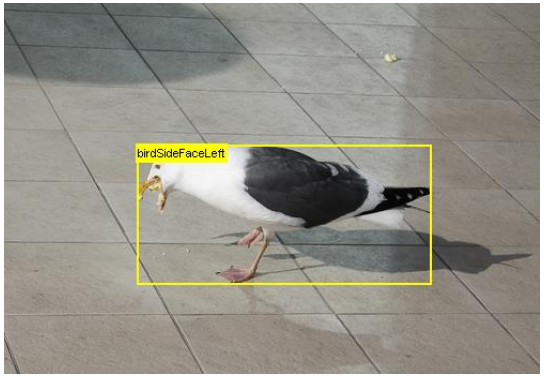
- Solving complex information processing tasks, for example,
  - hand-written character recognition,
  - image understanding
  - Network Intrusion Detection
  - Brain Signal Analysis
  - Chemical Compound Analysis
  - Genome Sequence Analysis
  - Document Organization/Search
  - etc.
- Learn to perform some task from examples, instead of solving the problem "by hand".

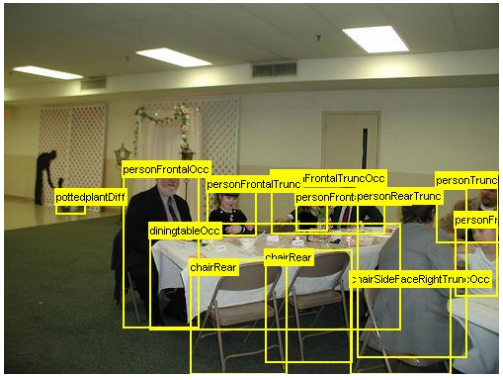
Example: Handwritten Character Recognition



Example: Image Understanding



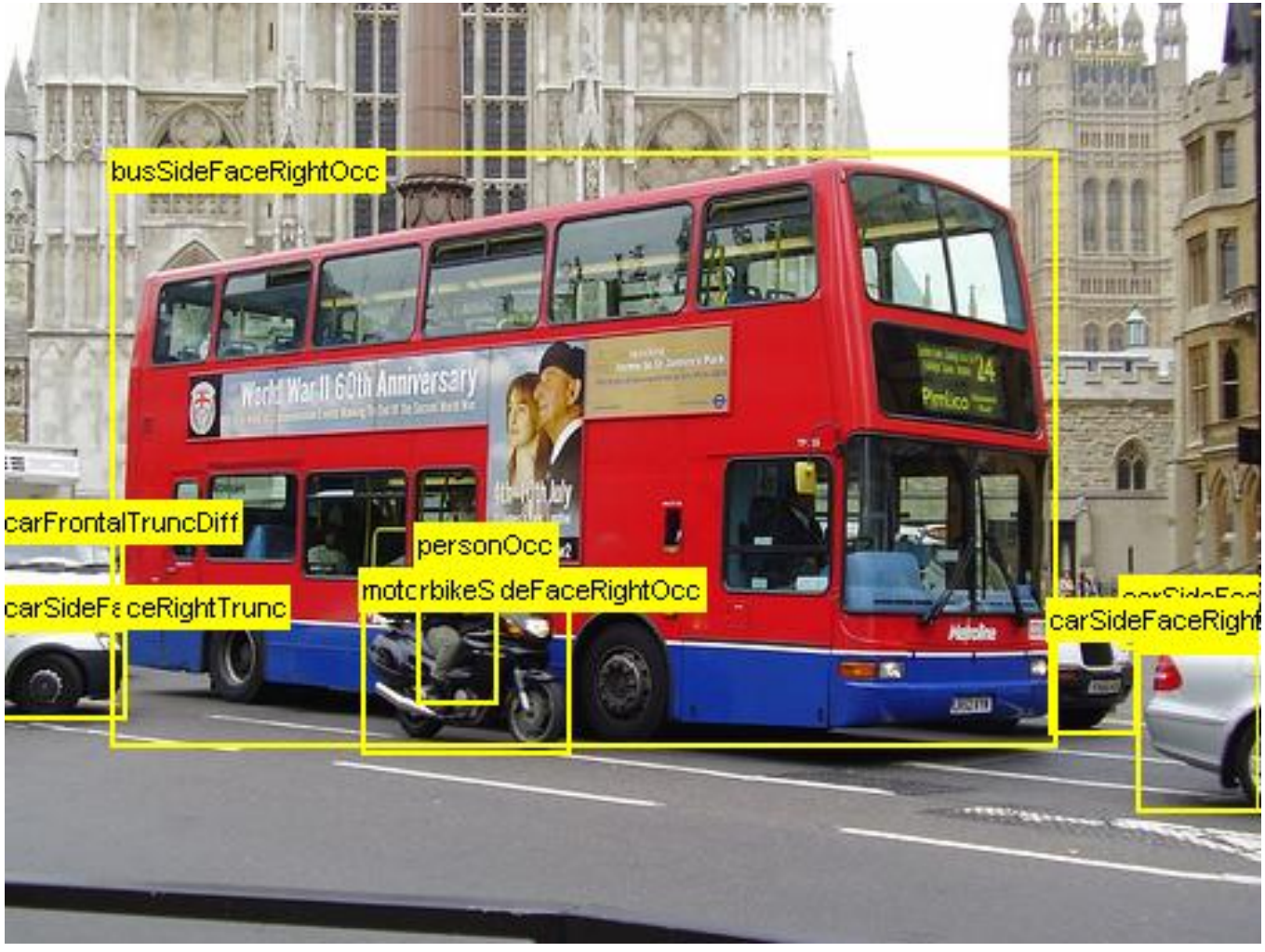




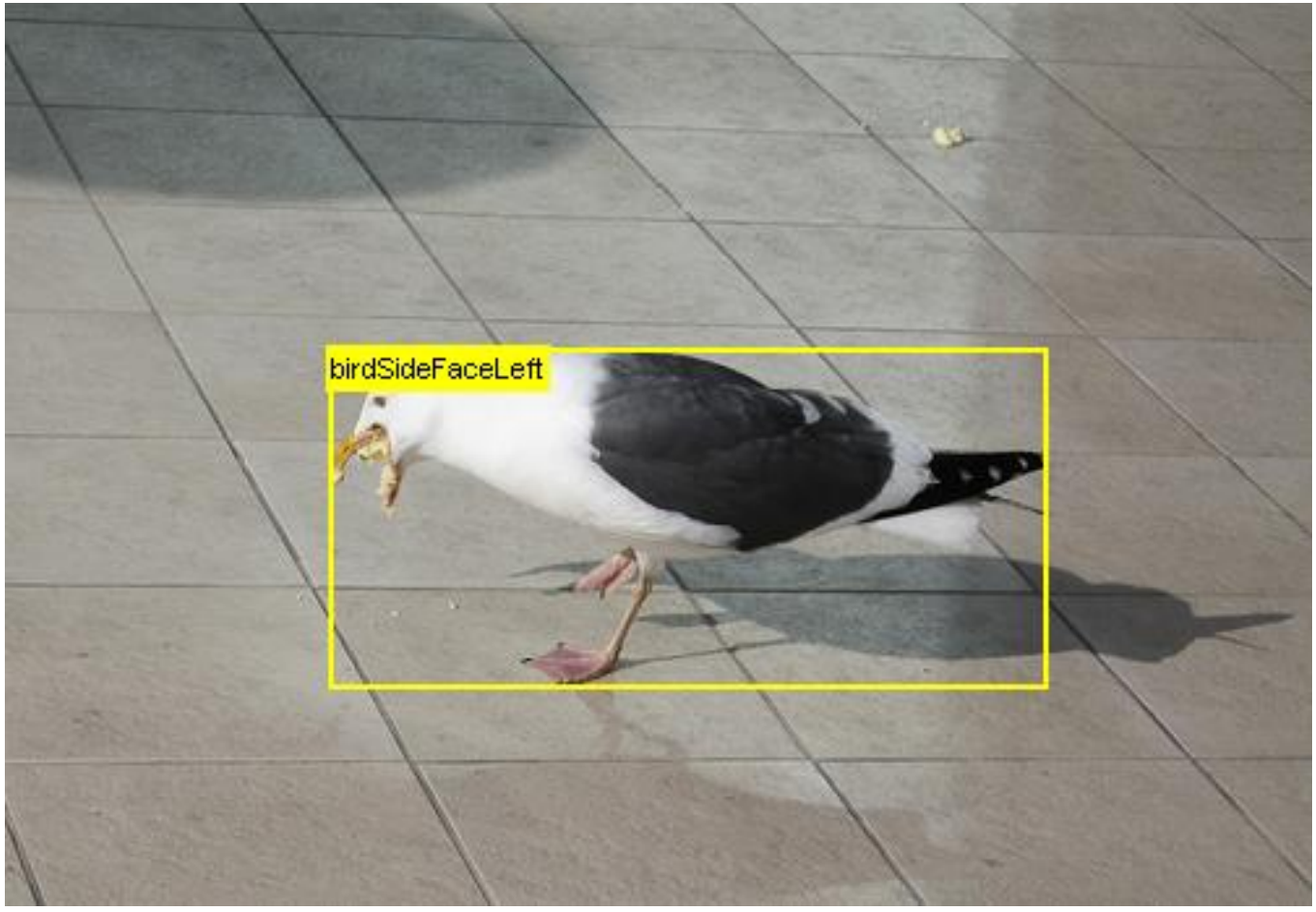
Example: Image Understanding

Example: Image Understanding

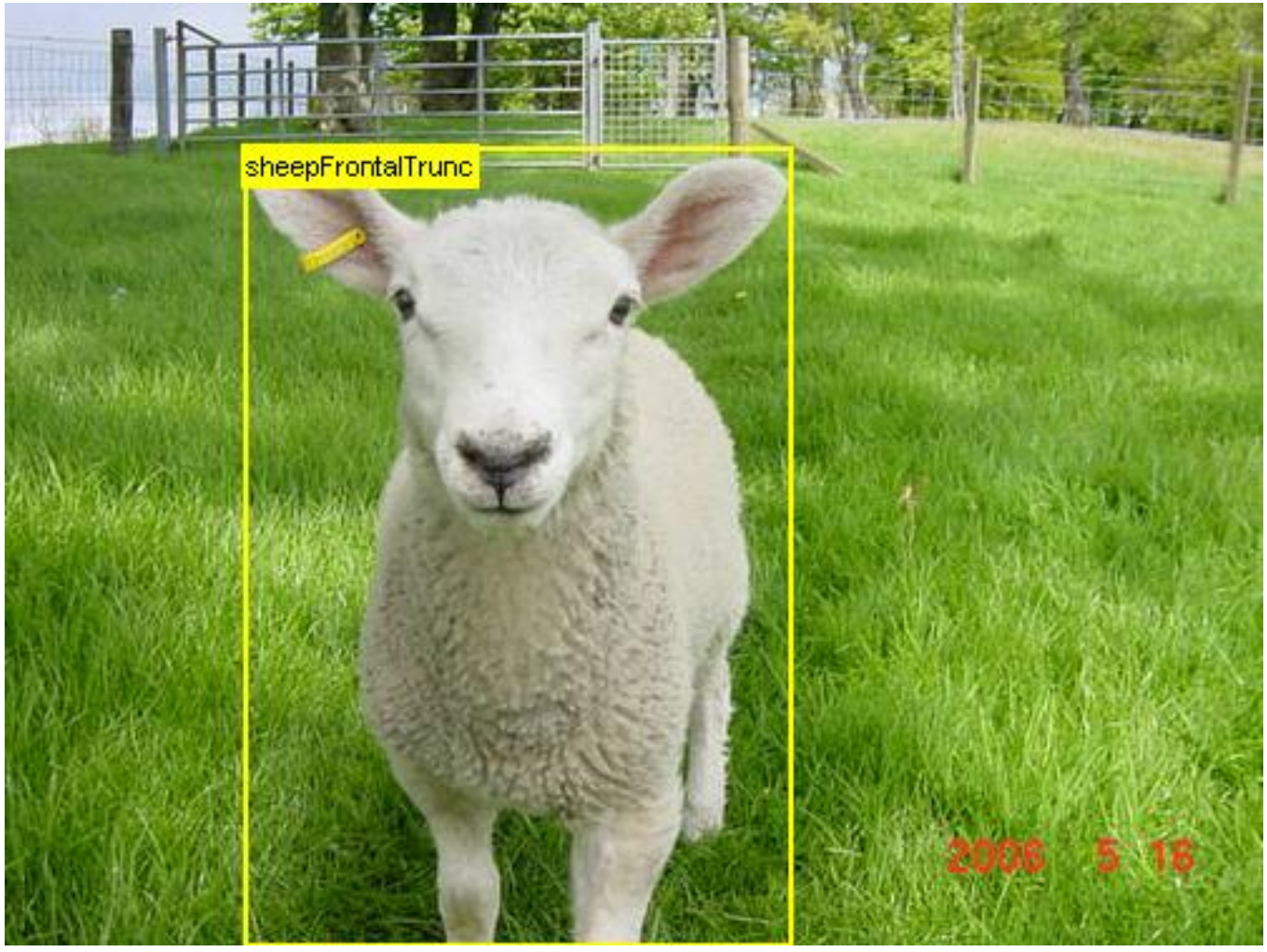
Example: Image Understanding



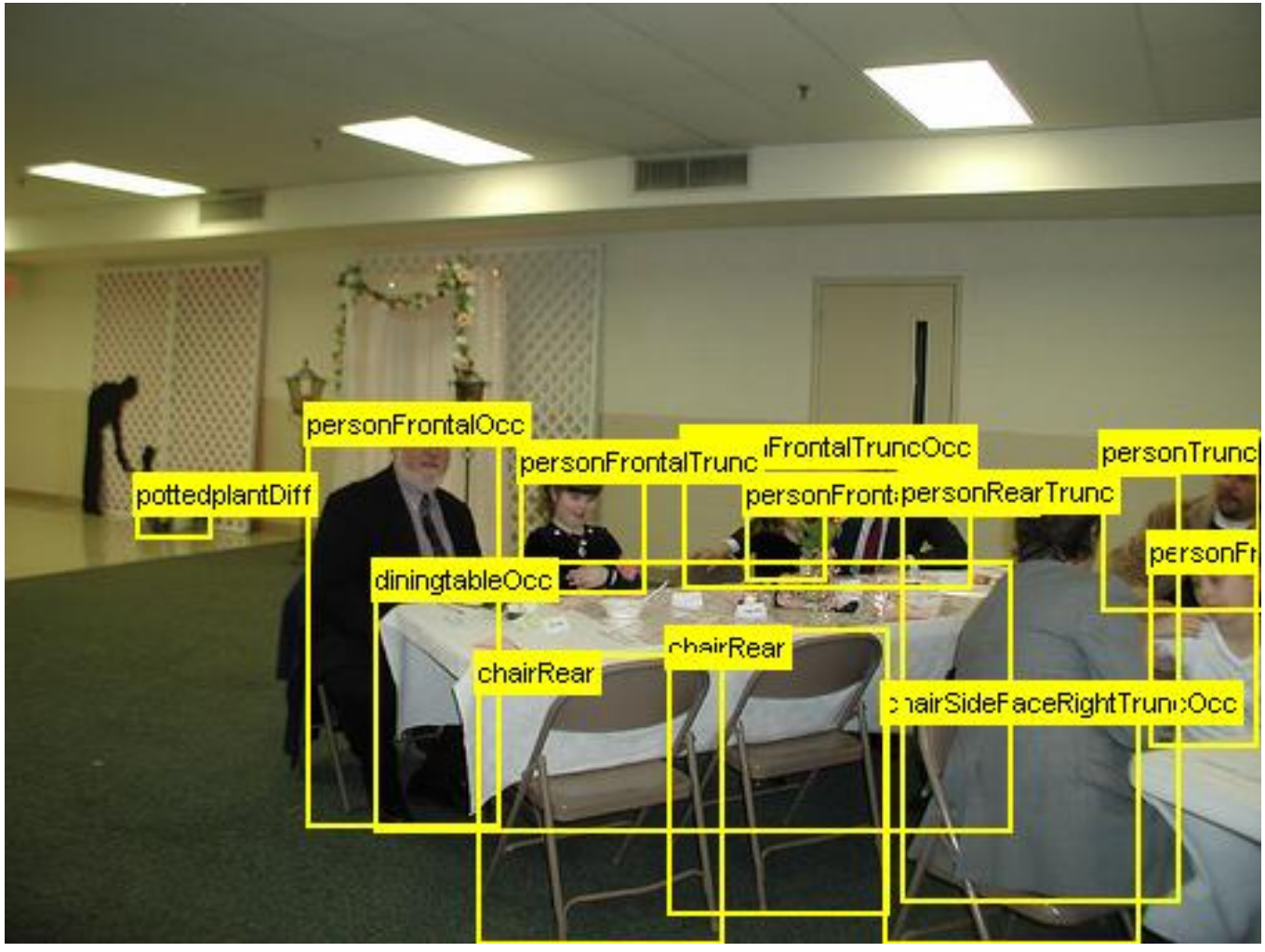
Example: Image Understanding



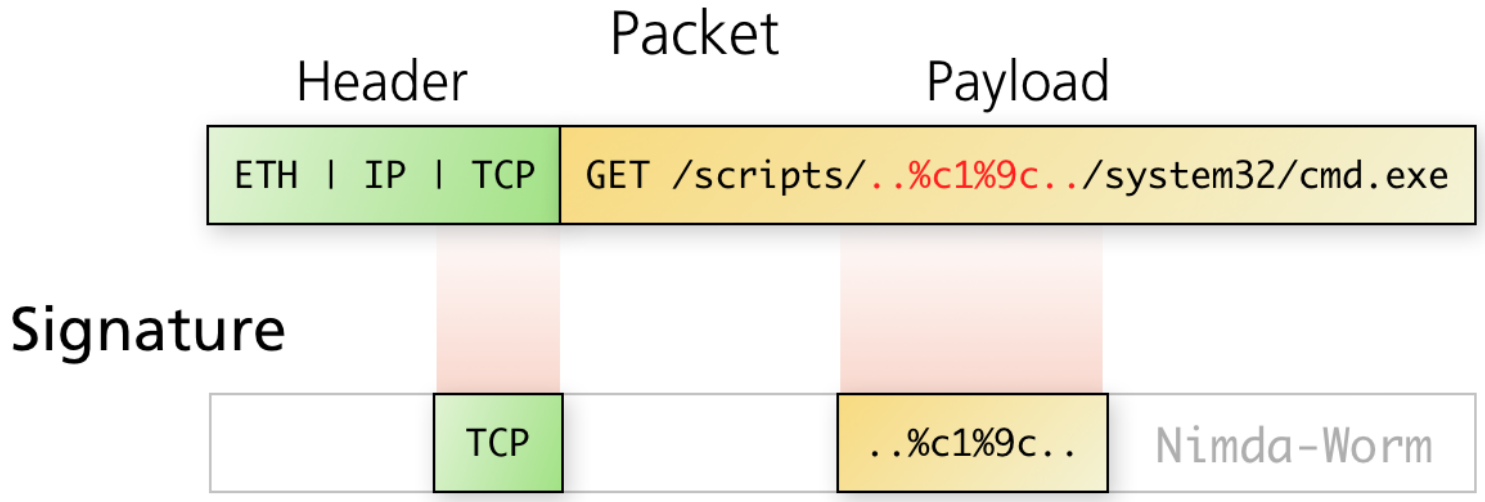
Example: Image Understanding



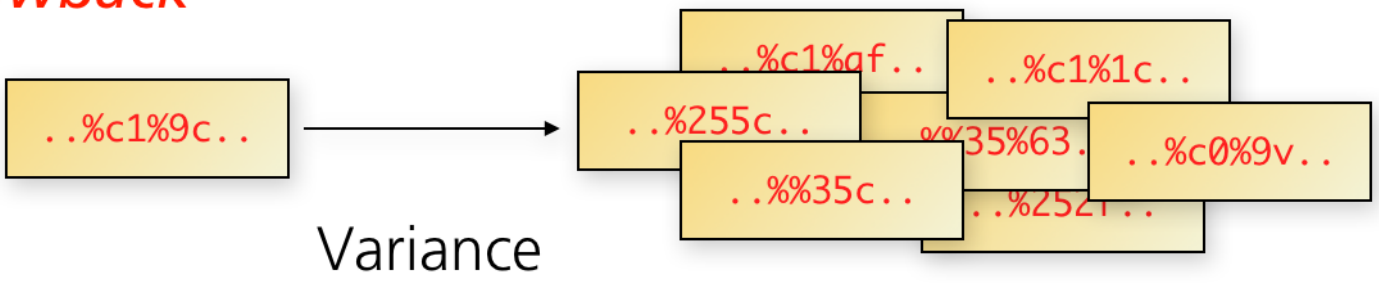
Example: Image Understanding



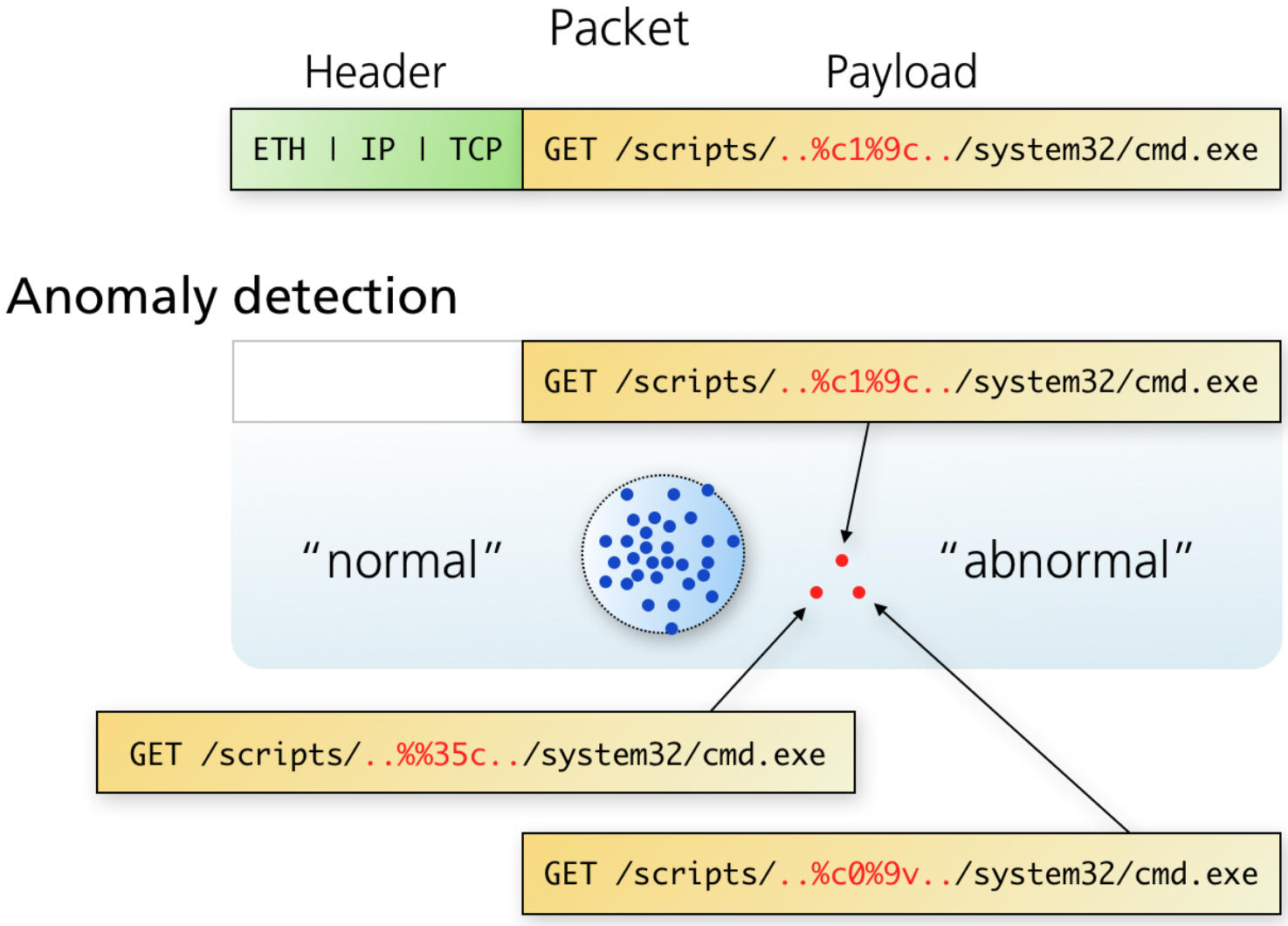
Example: Network Intrusion Detection



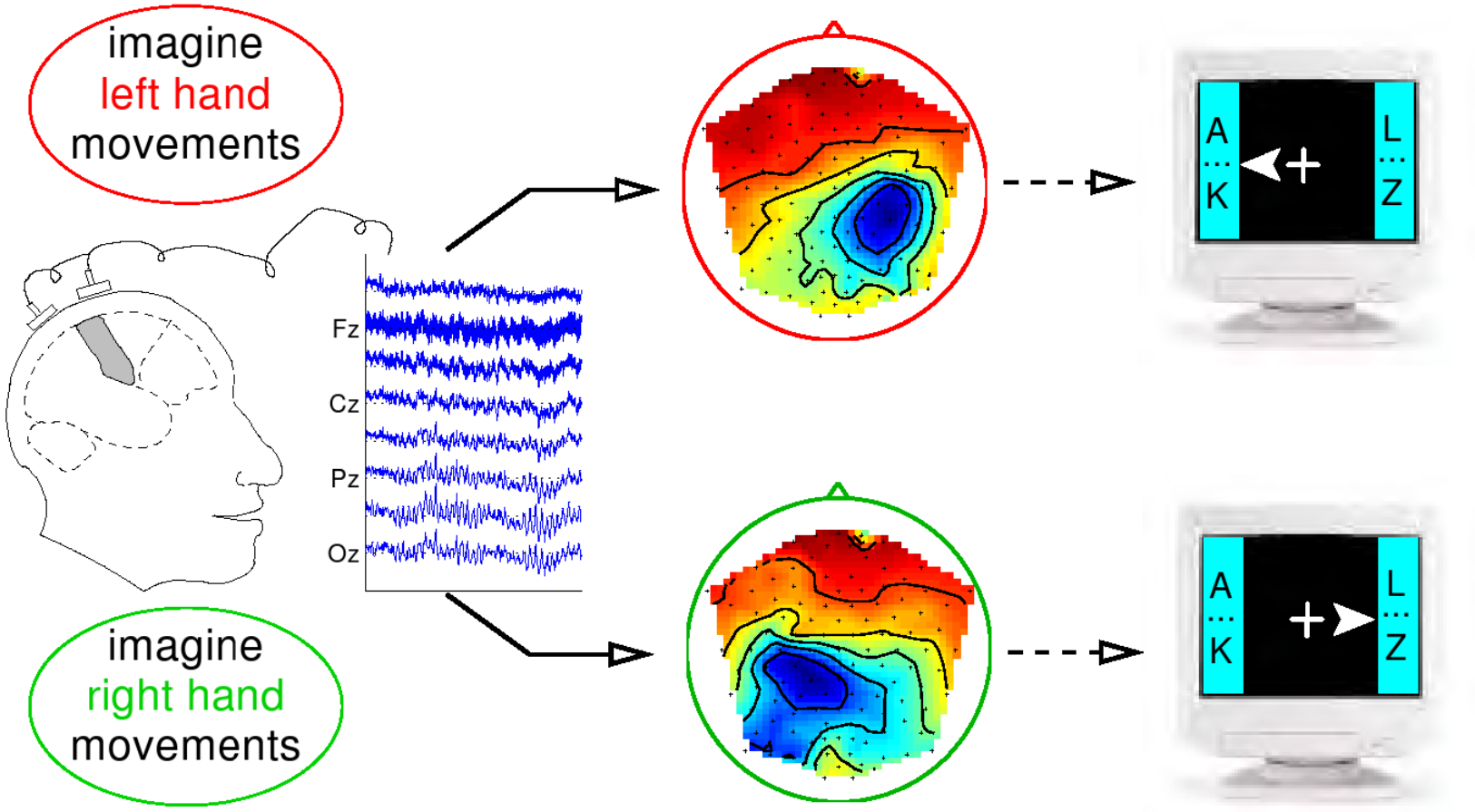
Drawback



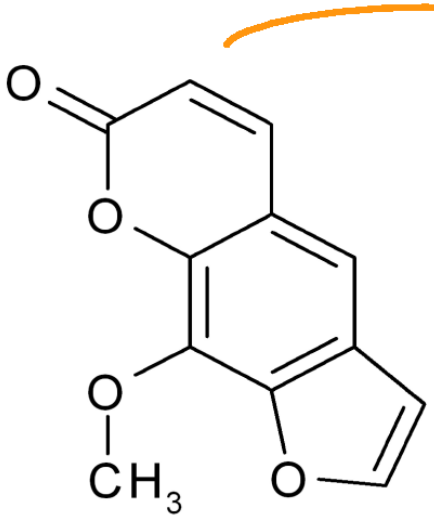
Example: Network Intrusion Detection



Example: Brain Signal Analysis



Example: Chemical Compound Analysis



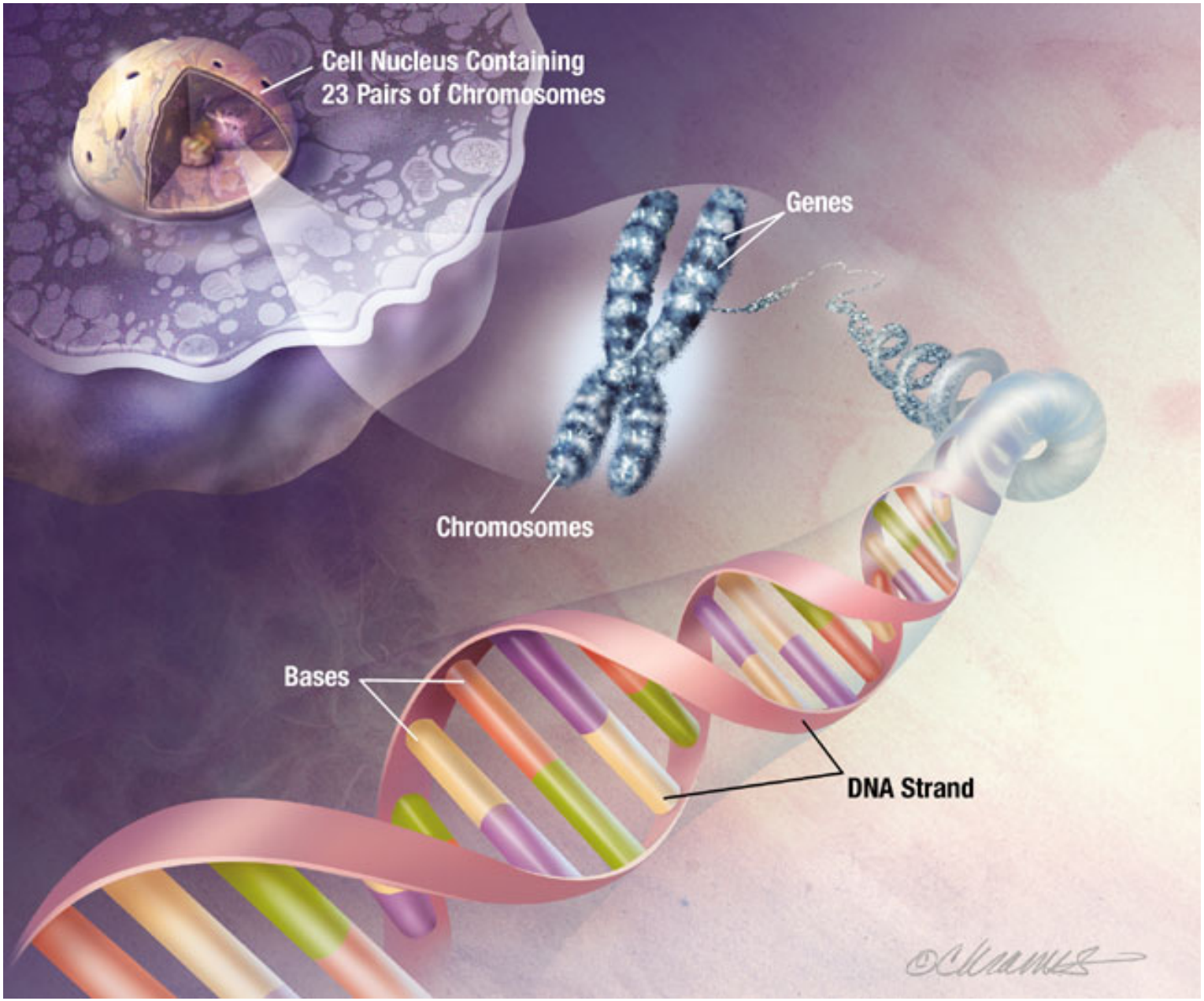
Descriptors (up to 1000)
and
Measurements

Features

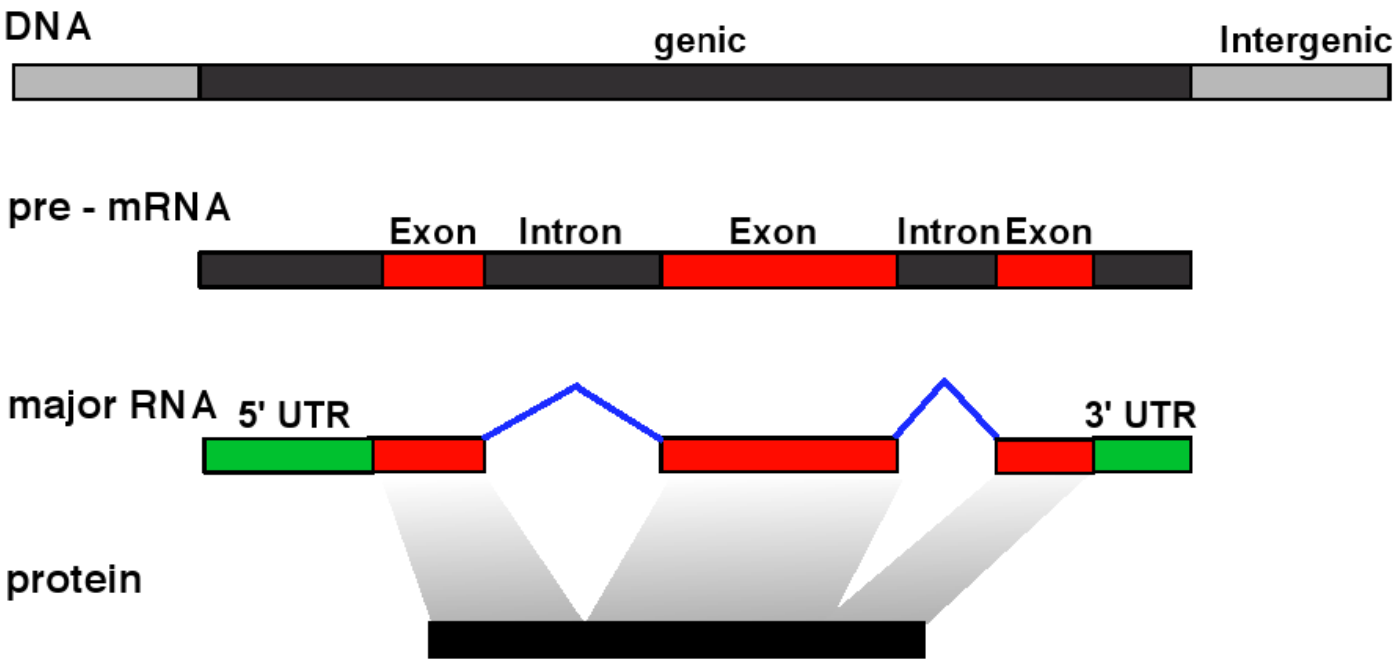
Metabolic Stability,
Target Interaction,
Toxicity, ...

Labels

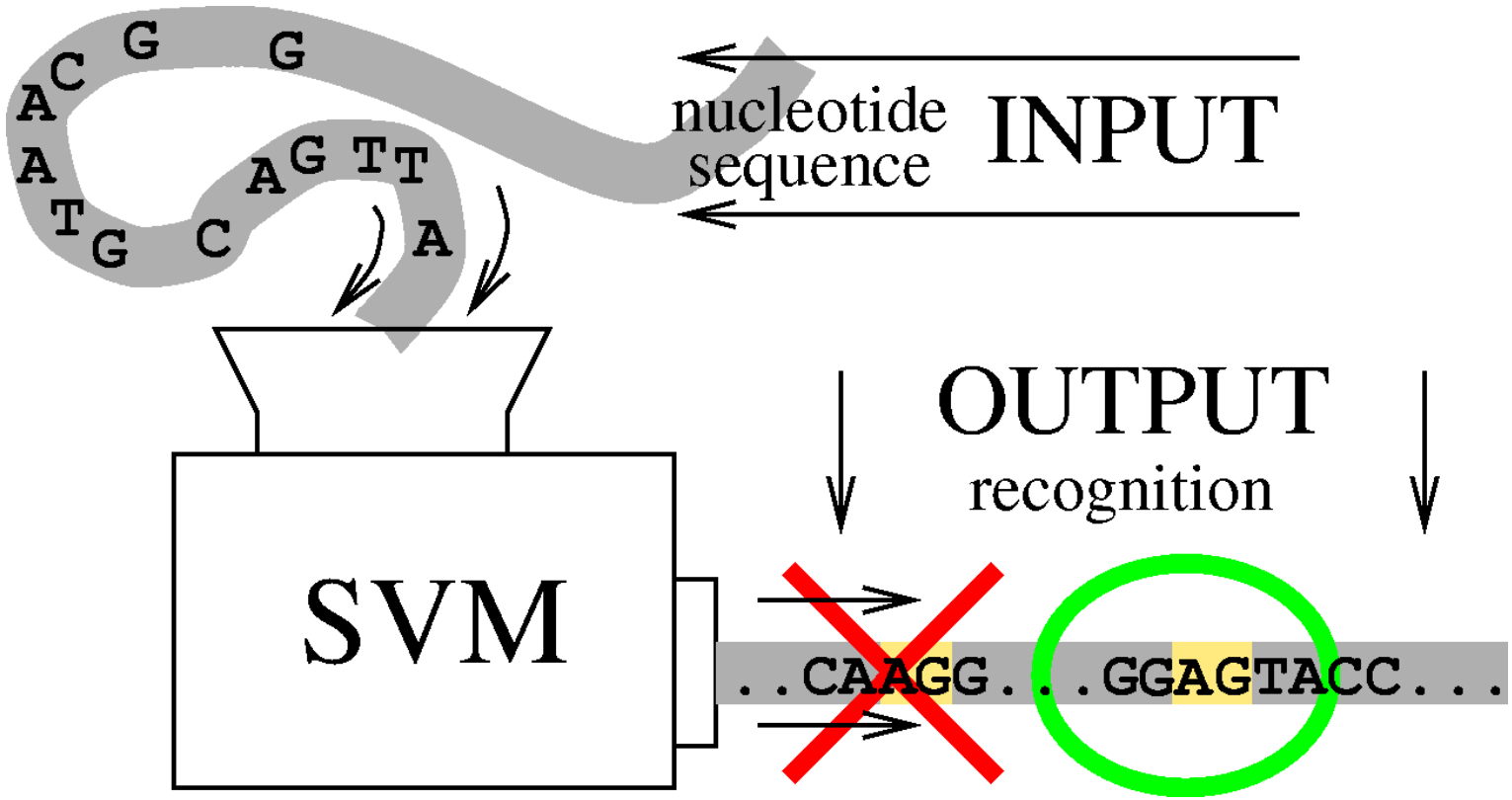
Example: Genome Sequence Analysis



Example: Genome Sequence Analysis



Example: Genome Sequence Analysis



Example: Document Organization/Search

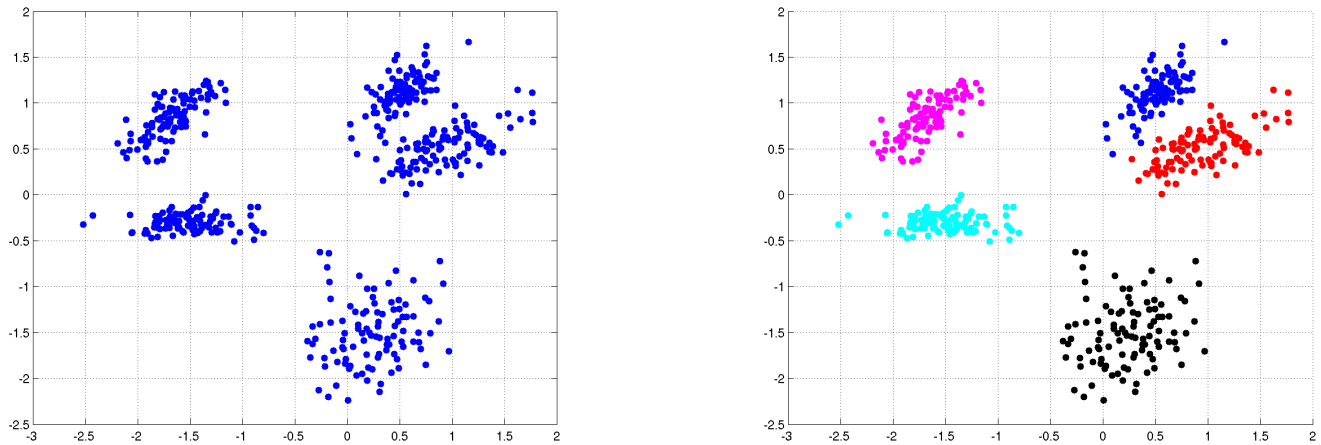
- Textual information sources
  - WWW, news archives, linked document archives, pdf files
- Information extraction
  - Relation and event extraction.
  - Find entities like names, date, time, location.
- Information retrieval.
  - Web search.
  - Find related (news) articles.
- Applications based on text mining:
  - Search engines (e.g., Google, Yahoo).
  - Recommender Systems (e.g., Amazon).
  - Machine translation (e.g., babelfish).

Types of Learning

Unsupervised Learning:

No labels given, data is $X_1, \ldots, X_n$, often $X_i \in \mathbb{R}^d$

- Clustering
  Partition data set into K classes

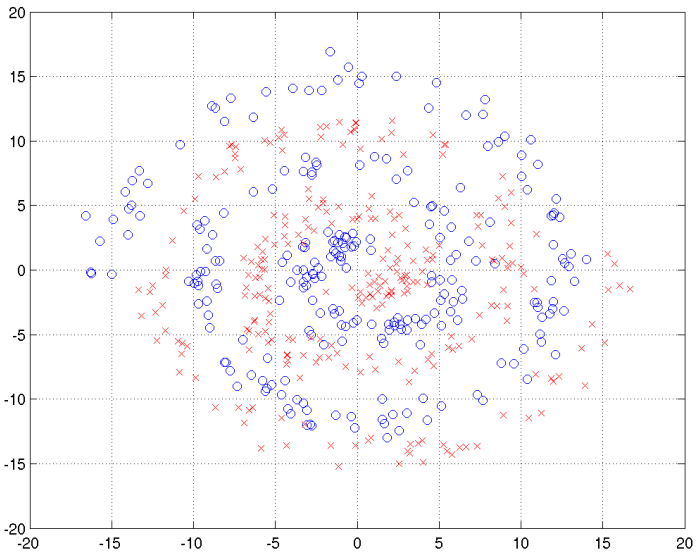


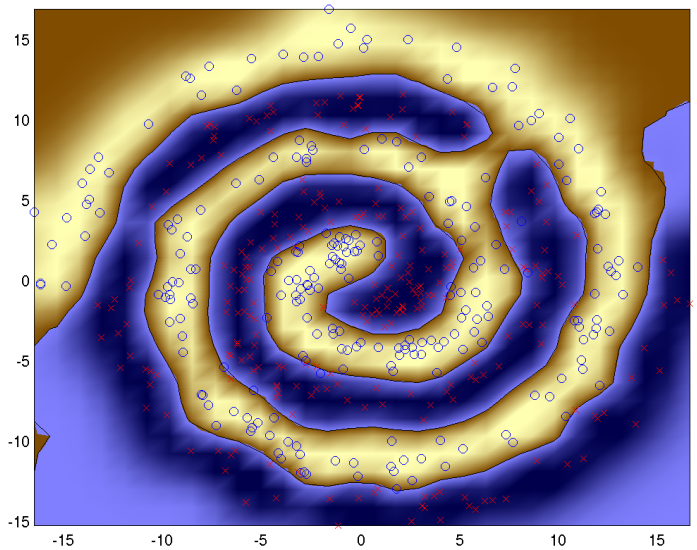
- Dimensionality reduction
  Reduce number of features by some criterion

Types of Learning

Supervised Learning:
Learn function $f : X \to Y$
from examples $X_1, Y_1, \ldots, X_n, Y_n$ with

- Classification: $Y \in \{\pm 1\}$, or $\{1, \ldots, K\}$.





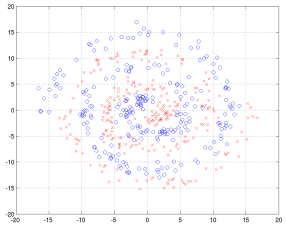
Types of Learning

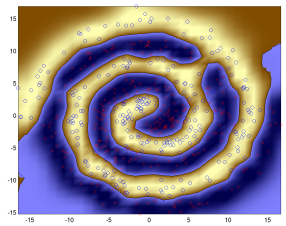
Supervised Learning:

Learn function $f : X \to Y$

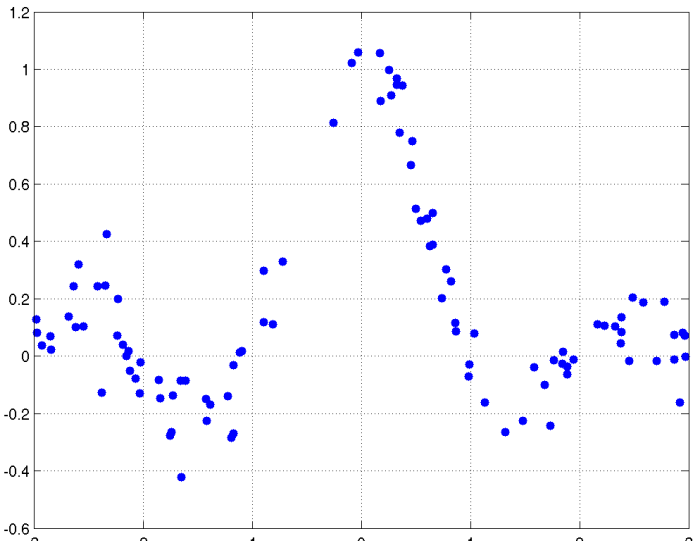
from examples $X_1, Y_1, \ldots, X_n, Y_n$ with

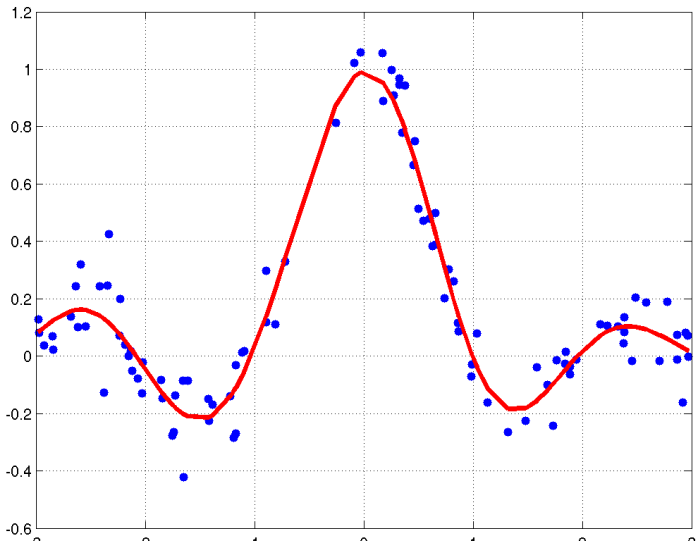
- Classification: $Y \in \{\pm 1\}$, or $\{1, \ldots, K\}$.





- Regression: $Y \in \mathbb{R}$

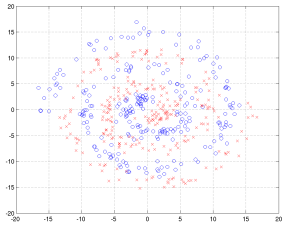
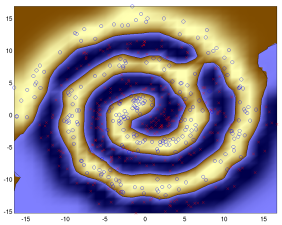




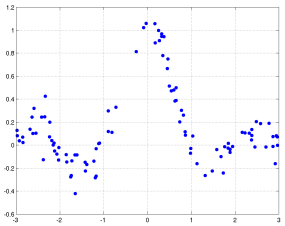
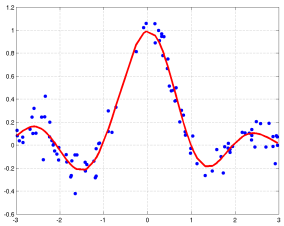
Types of Learning

Supervised Learning:
Learn function $f : X \to Y$
from examples $X_1, Y_1, \ldots, X_n, Y_n$ with

- Classification: $Y \in \{\pm 1\}$, or $\{1, \ldots, K\}$.

- Regression: $Y \in \mathbb{R}$

- Structured Output: Y is a graph, parse tree, etc.

Types of Learning

Other types of learning

- **Reinforcement Learning:** Learn an optimal policy from a reward function
- **Semi-supervised Learning:** Learn with partially unlabeled data
- **Transductive Learning:** Test points are already known while learning, no prediction to other points is necessary.
- **Covariate Shift:** Data distribution changes between training and testing.

Approaches

- Statistics — estimators, laws of large numbers, ...
- Physics — statistical physics, dynamical systems, ...
- Biology — neuronal networks, ...
- Bayes theory — Bayes inference, graphical models, ...
- Geometry — Separating hyperplanes, ...
- Graph theory — Graph laplacian, ...
- Optimization theory — convex programs, quadratic programs, linear programs ...

Solving Formally Underspecified Problems

Unlike "normal computer science", ML problems are often hard to specify formally:

easy to formalize	hard to formalize
shortest path in a graph	hand-written character recognition
sorting a set	detecting general objects in images
searching in a graph	finding related documents
parsing a program	machine translation
...	...

Instead of a formal problem definition, ML often uses "problem definition by example".

Proper Validation

- Methods need to perform well on *unseen* data.
- Often a fixed data set is all we have.
- Repeatedly split data set into training and testing part.
- Data which has been used for training cannot be used for *validation*.