INTRODUCTION TO DATA SCIENCE

Lectures based on:

- E. Fox and C. Guestrin, „Machine Learning and Data Analysis”, Univ. of Washington
- M. Cetinkays-Rundel, „Data Analysis and Statistical Inference”, Univ. of Duke

9/10/2025

WFAiS UJ, Informatyka Stosowana

I stopień studiów

What is Data Science?

Is mainly about extracting knowledge from data (terms "data mining" or "Knowledge Discovery in Databases" are highly related). It can be about analyzing trends, building predictive models, … etc.

Is an agglomerate of data collection, data modeling and analysis, a decision making, and everything you need to know to accomplish your goals. Eventually, it boils down to the following fields/skills:

- **Computer science:**

Algorithms, programming (patterns, languages etc.), understanding hardware & operating systems, high-performance computing'

- **Mathematical aspects:**

Linear algebra, differential equations for optimization problems, statistics

- **Few others:**

Machine learning, domain knowledge, and data visualization & communication skills

Data Science and Machine Learning?

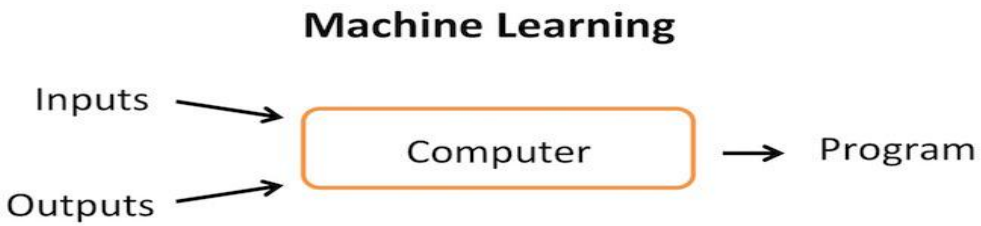
Machine learning algorithms are algorithms that learn (often predictive) models from data. I.e., instead of formulating "rules" manually, a machine learning algorithm will learn the model for you.

Machine learning - at its core - is about the use and development of these learning algorithms. **Data science** is more about the extraction of knowledge from data to answer particular question or solve particular problems.

Machine learning is often a big part of a "data science" project, e.g., it is often heavily used for exploratory analysis and discovery (clustering algorithms) and building predictive models (supervised learning algorithms). However, in **data science**, you often also worry about the collection, wrangling, and cleaning of your data (i.e., data engineering), and eventually, you want to draw conclusions from your data that help you solve a particular problem.

Traditional programming paradigm and Machine Learning

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)



Sebastian Raschka, 2016

Outline of the course

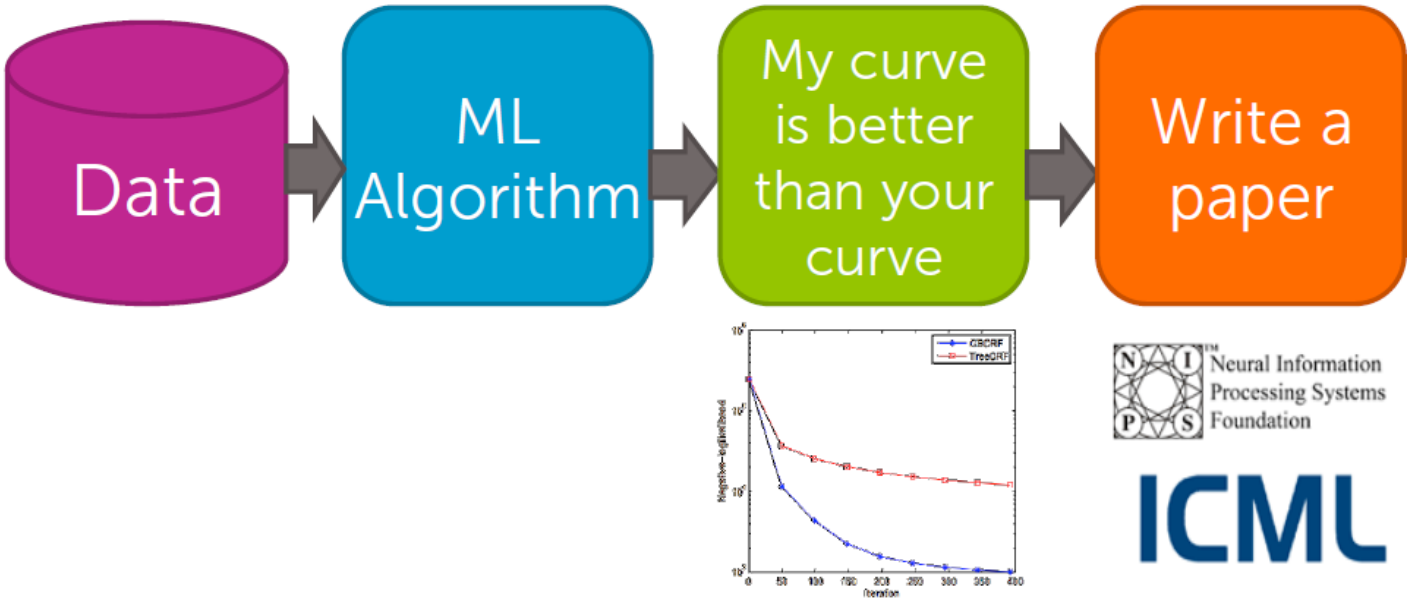
- **Exploratory Data Analysis:** introduction

  → today

- **Data Analysis** with **Machine Learning** algorithms:
  - Regression (October)
  - Classification (November)
  - Retrieval & Clustering (December)
  - Recomender system (January)
  - Statistical inference (January)
  - MC methods, ML methods (January)

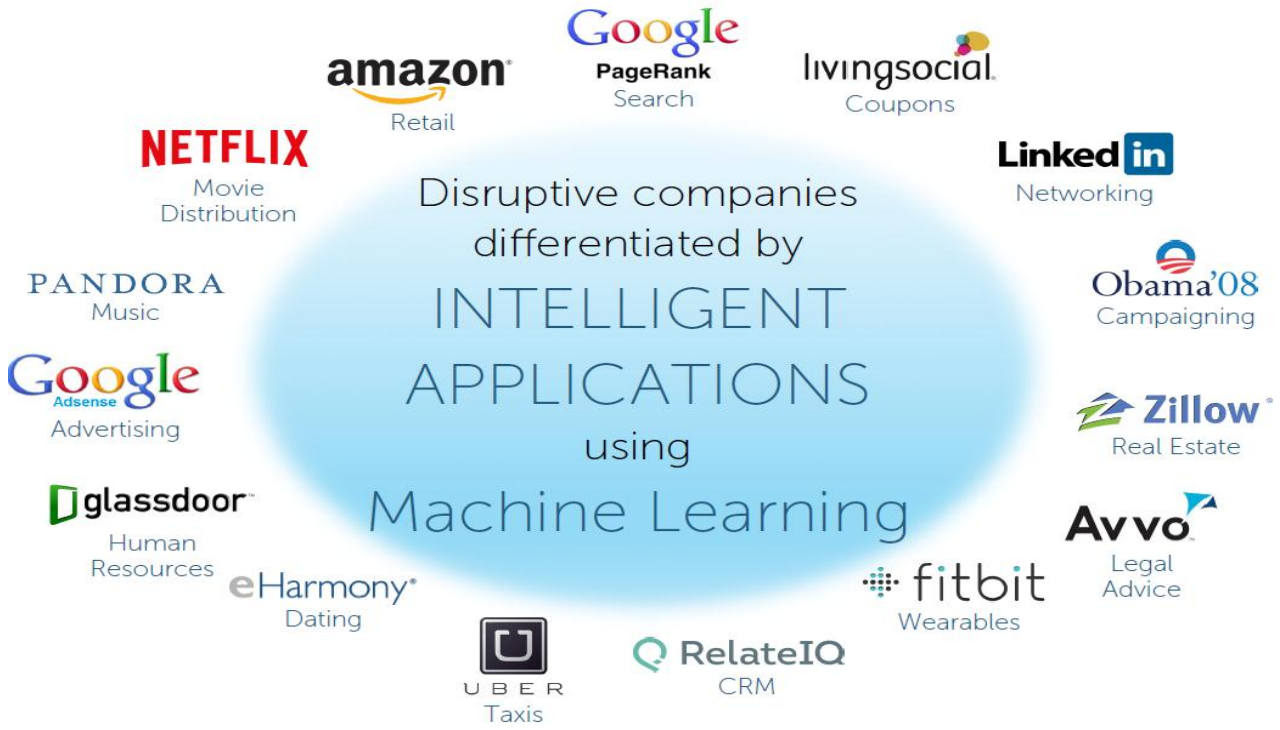
Analyse data with Machine Learning

- **Machine learning is changing the world.**
- **Old view of ML**



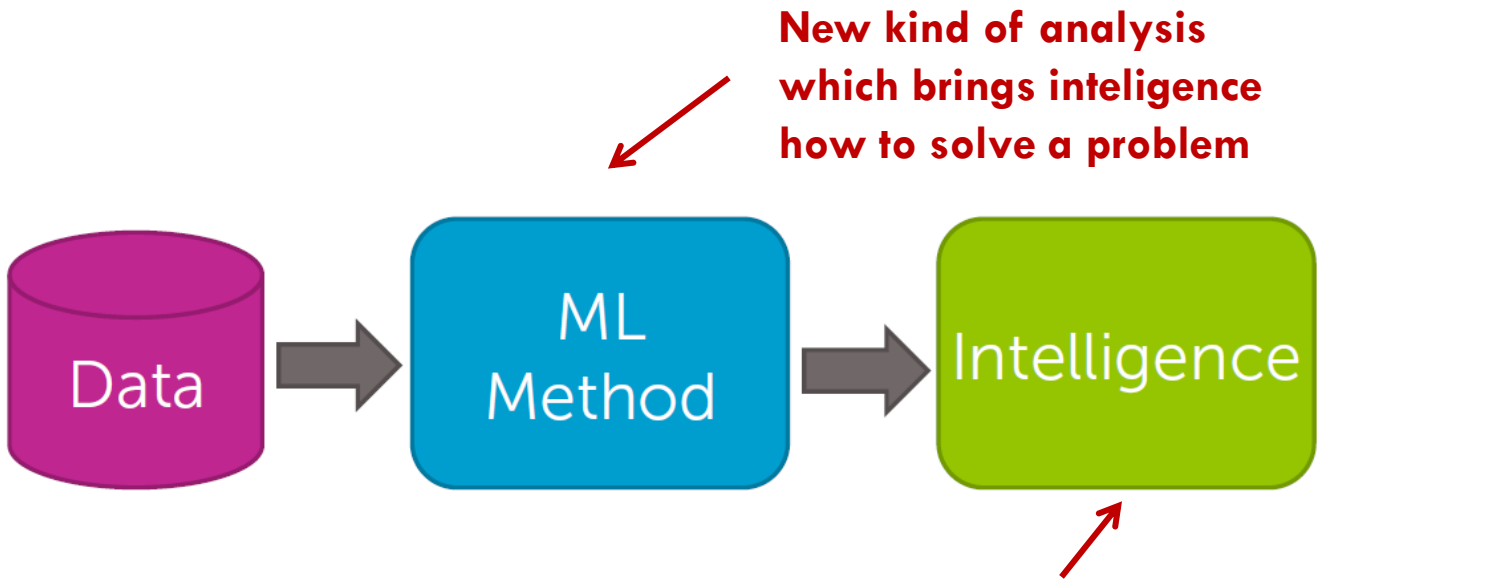
Machine learning is changing the world

- **Current view: disruptive inteligent applications are used by leading comercial companies**

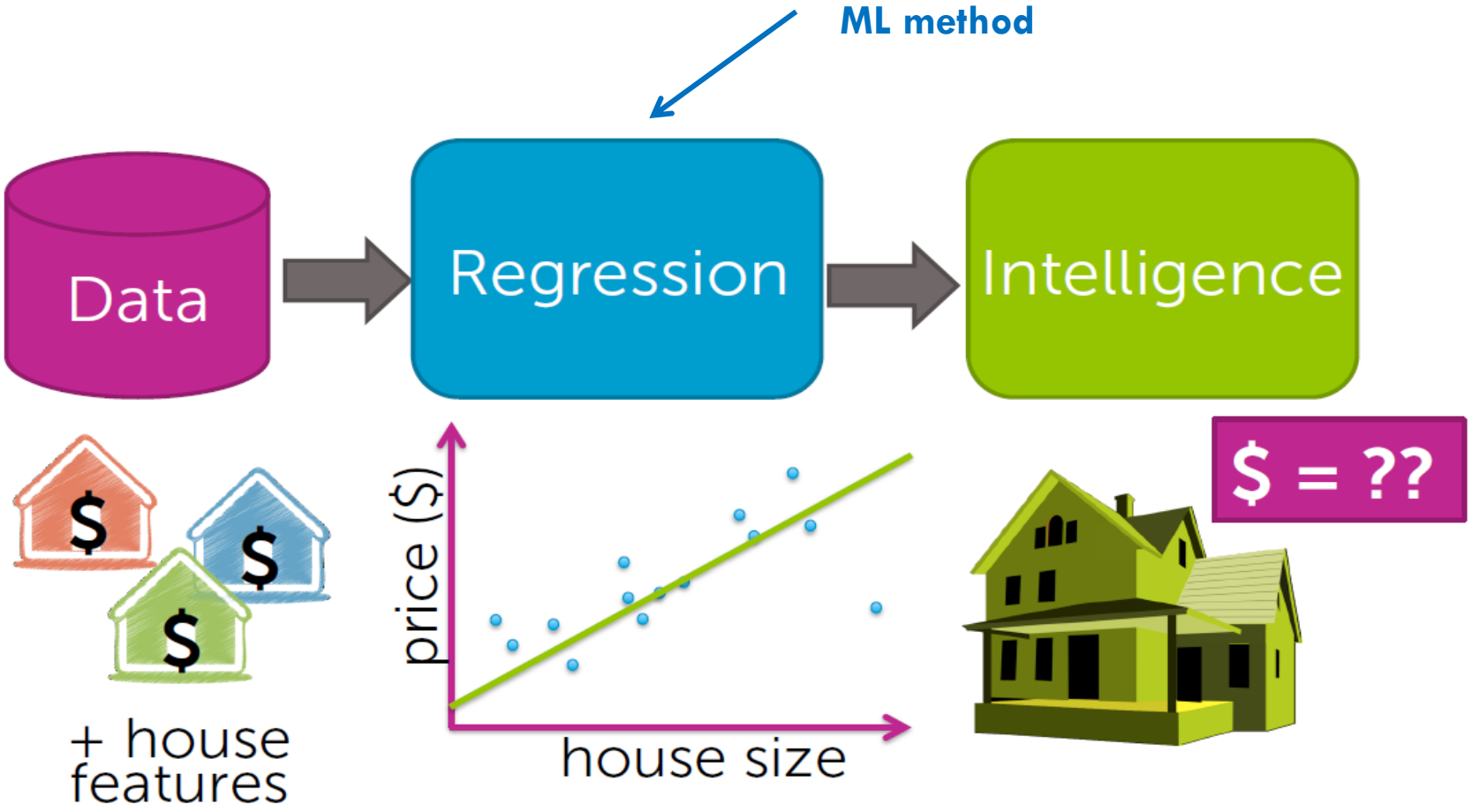


Machine learning

- **Data → inteligence pipeline**



Case study 1: Prediction



Prediction

hard work

- How much will your salary be? (y = $$)
- Depends on x = performance in courses, quality of programming assignments, # of discussion responses, …

Prediction

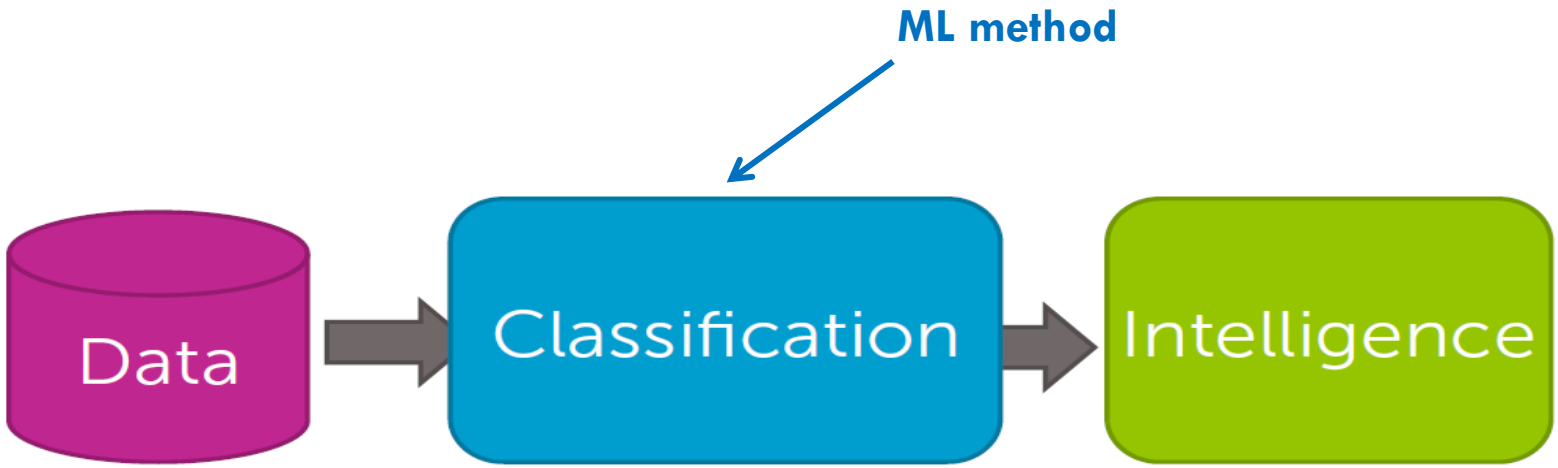
Tweet popularity

- How many people will retweet your tweet?
- Depends on # followers, # of followers of followers, features of text tweeted, popularity of hashtag, # of past retweets,...

Prediction:

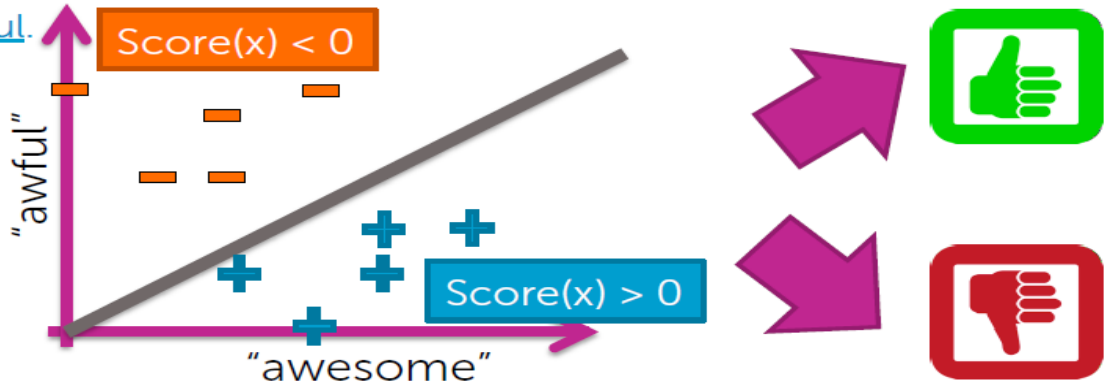
Models	• Linear regression • Regularization: Ridge (L2), Lasso (L1)
Algorithms	• Gradient descent • Coordinate descent
Concepts	• Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection

Case study 2: Classification



Sushi was awesome,
the food was awesome,
but the service was awful.

All reviews:



Classification

Spam filtering

Text of email, sender, IP,...

Not spam

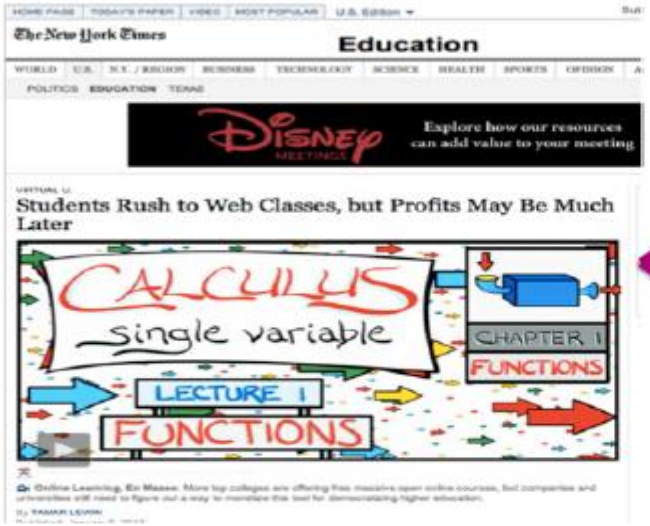
Spam

Input: **x**

Output: y

Multiclass classifier

Output y has more than 2 categories



Education

Finance

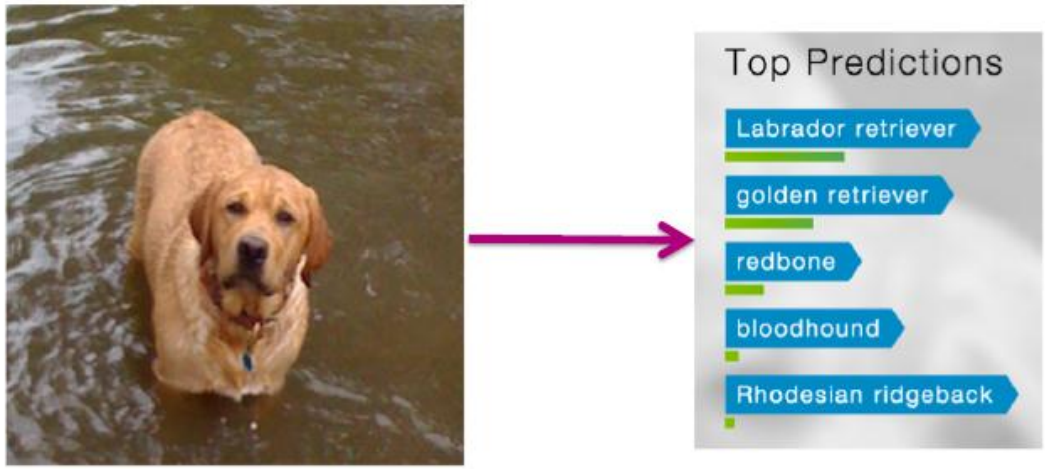
Technology

Input: x
Webpage

Output: y

Classification

Image classification



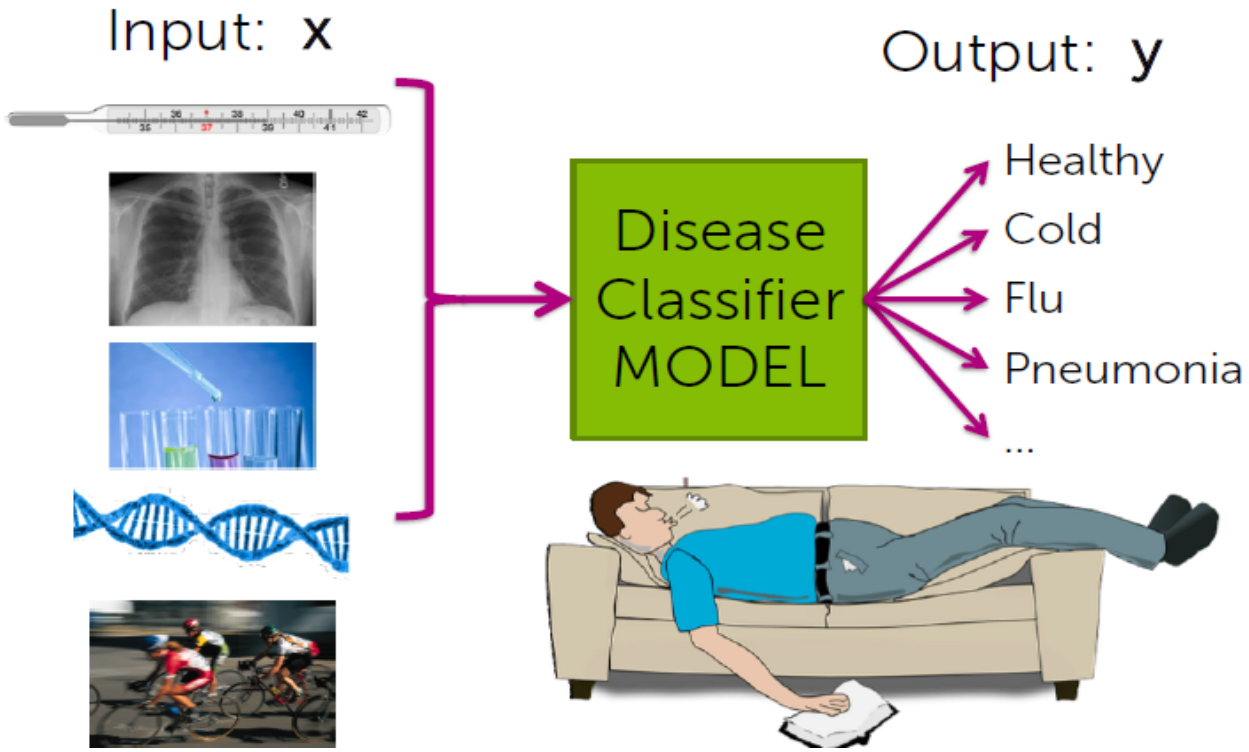
soft prediction

Input: x
Image pixels

Output: y
Predicted object

Classification

Personalized medical diagnosis



Classification:

Models

- Linear classifiers (logistic regression, SVMs, perceptron)
- Kernels
- Decision trees

Algorithms

- Stochastic gradient descent
- Boosting

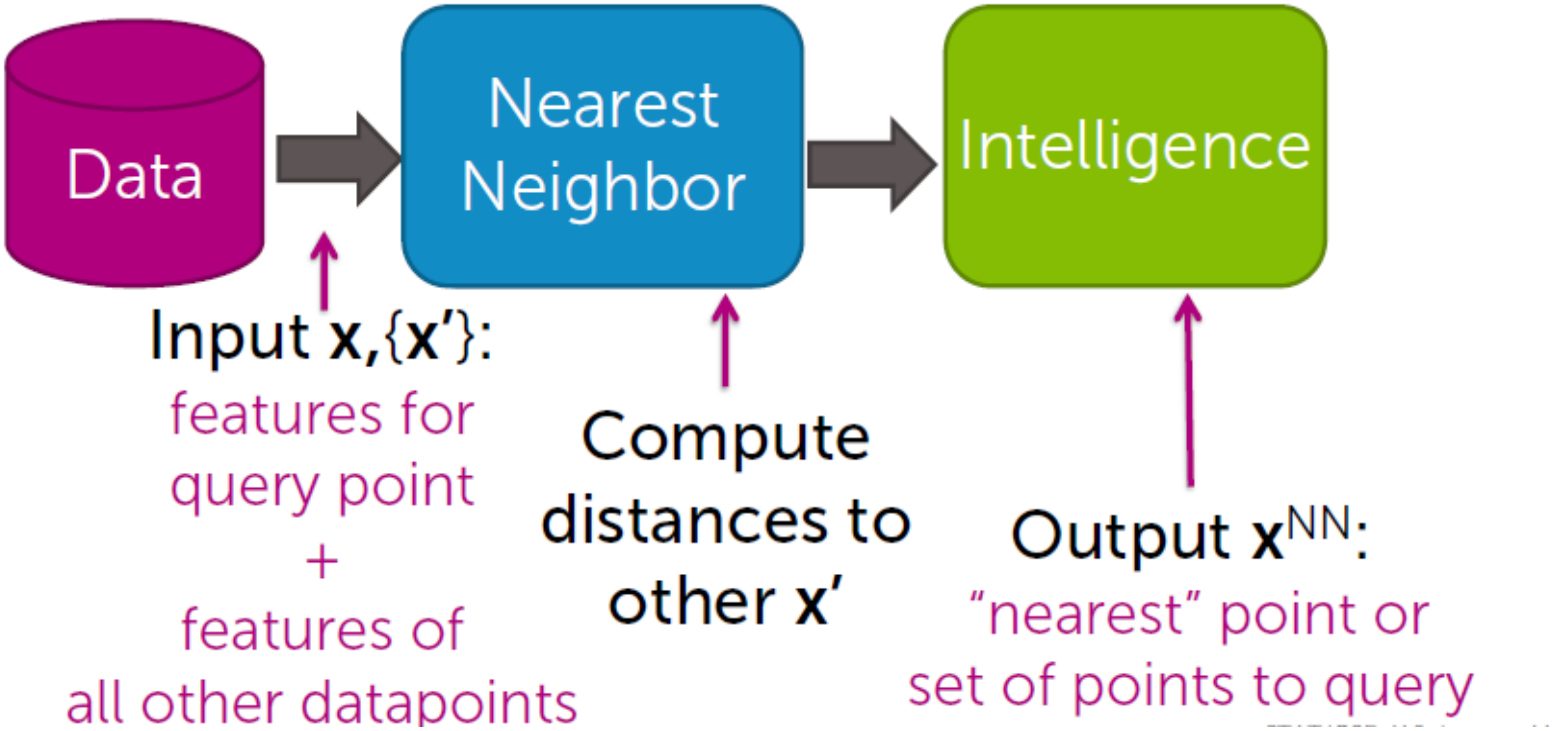
Concepts

- Decision boundaries, MLE, ensemble methods, random forests, CART, online learning

Case Study 3: document retrieval

Retrieval

Search for related items



Retrieval

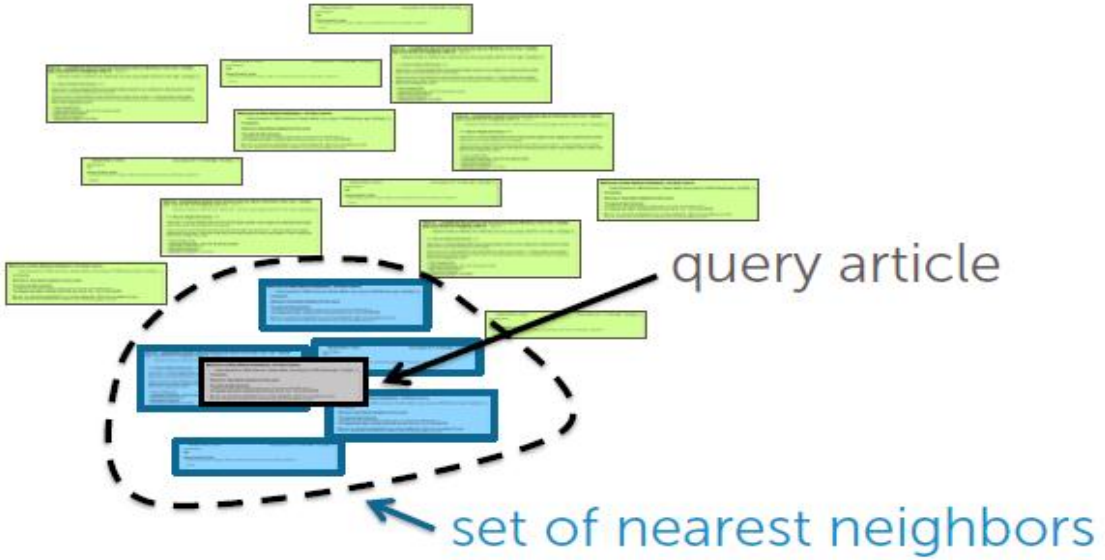
Retrieve “nearest neighbor” article

Space of all articles,
organized by similarity of text

Retrieval

Or set of nearest neighbors

Space of all articles,
organized by similarity of text



Retrieval

Retrieval applications

Just about everything…

Images

Products

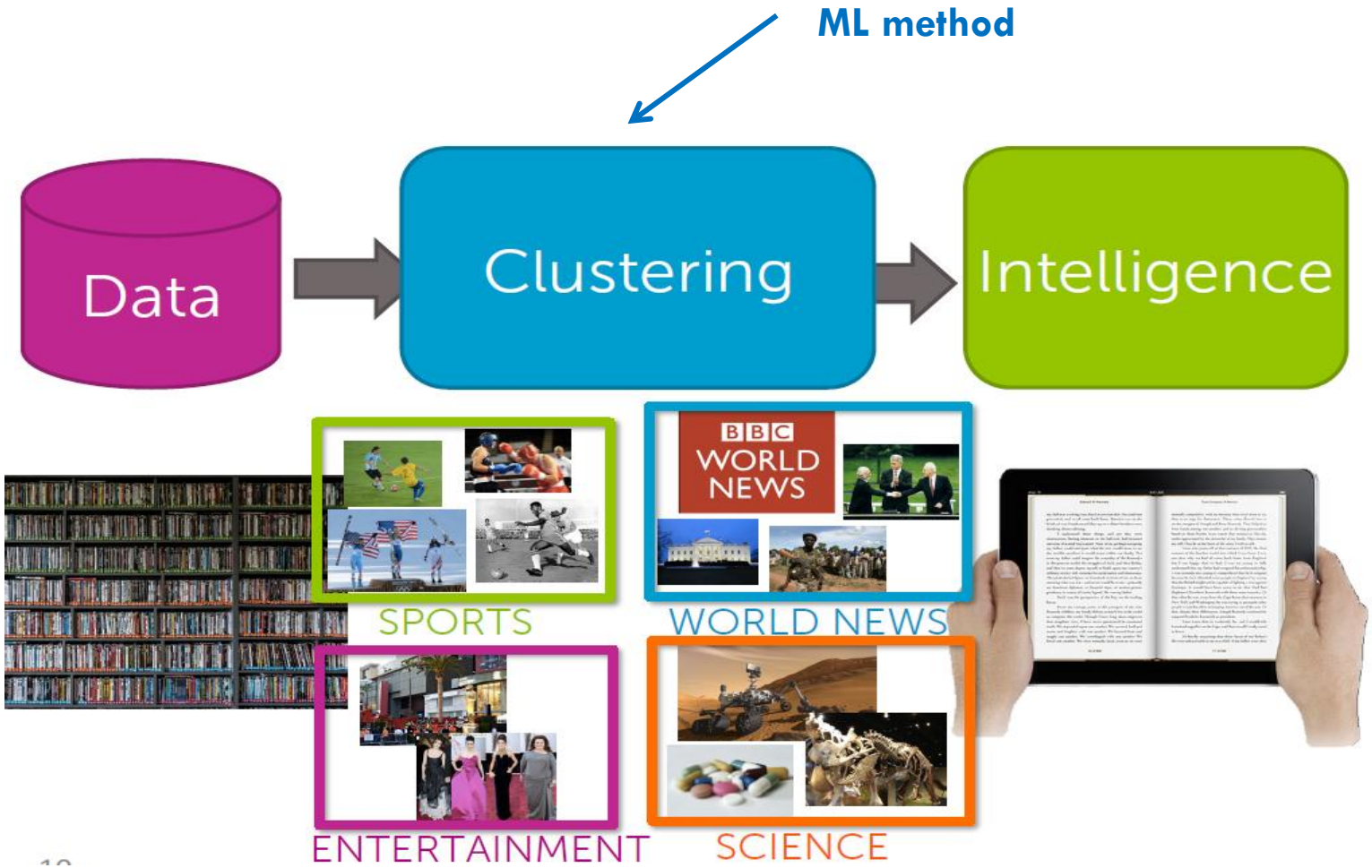
Streaming content:

- Songs
- Movies
- TV shows
- …

News articles

Social networks
(people you might want to connect with)

Case study 4: Clustering

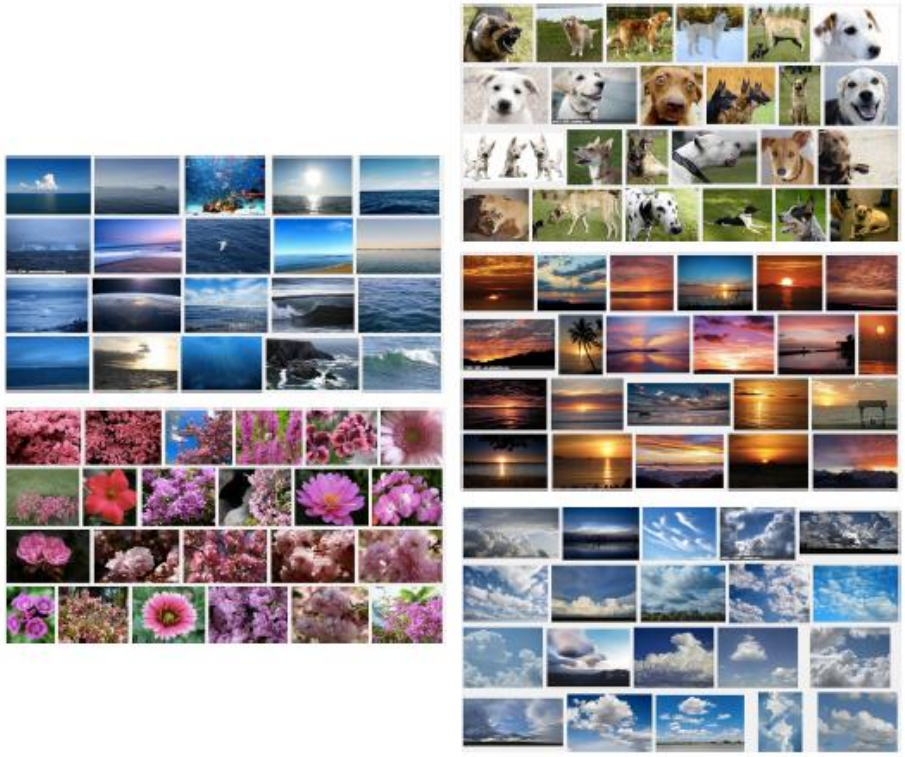


Clustering

Clustering images

For search, group as:

- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- …



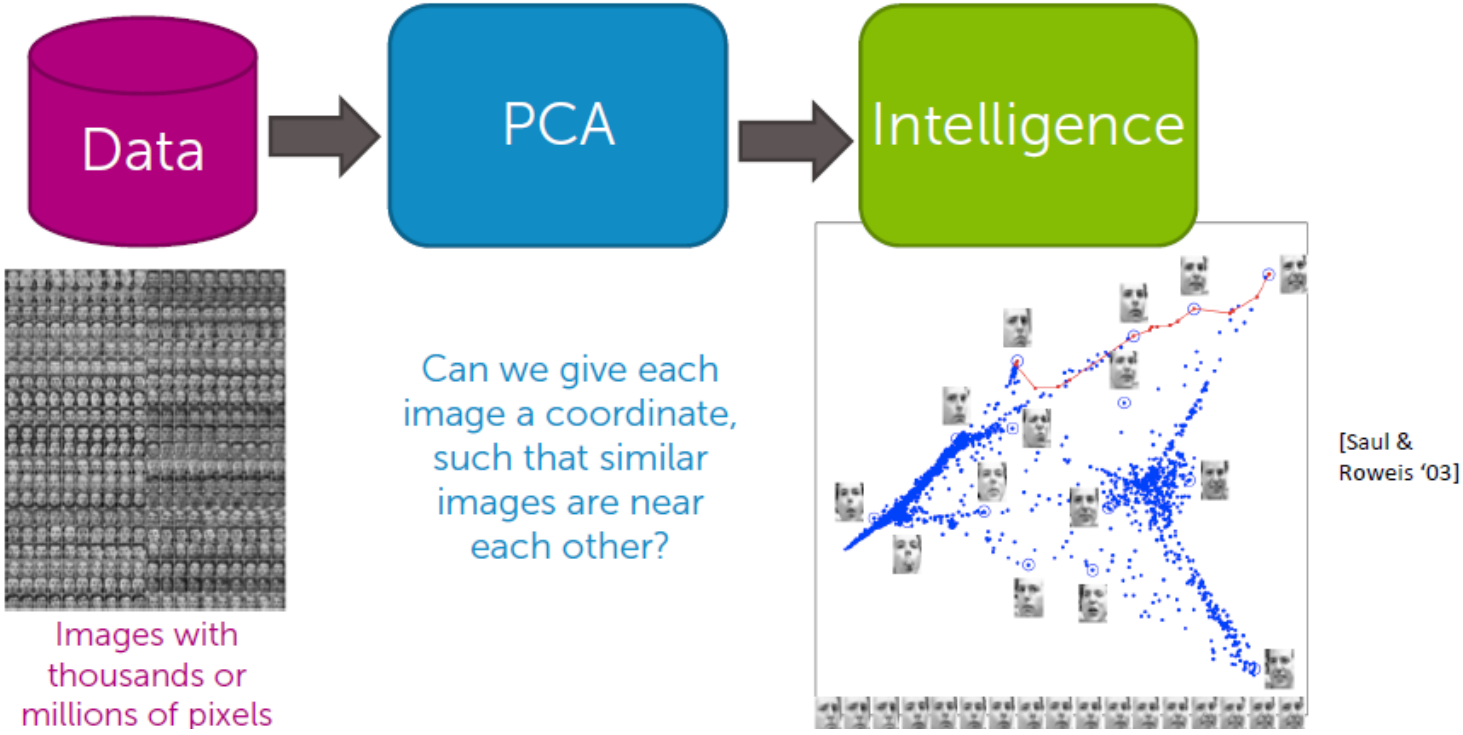
Clustering

Or users on websites…

Discover groups of users for better targeting of content

Embeding

Example: Embedding images to visualize data



Clustering: Finding documents

Models

- Nearest neighbors
- Clustering, mixtures of Gaussians
- Latent Dirichlet allocation (LDA)

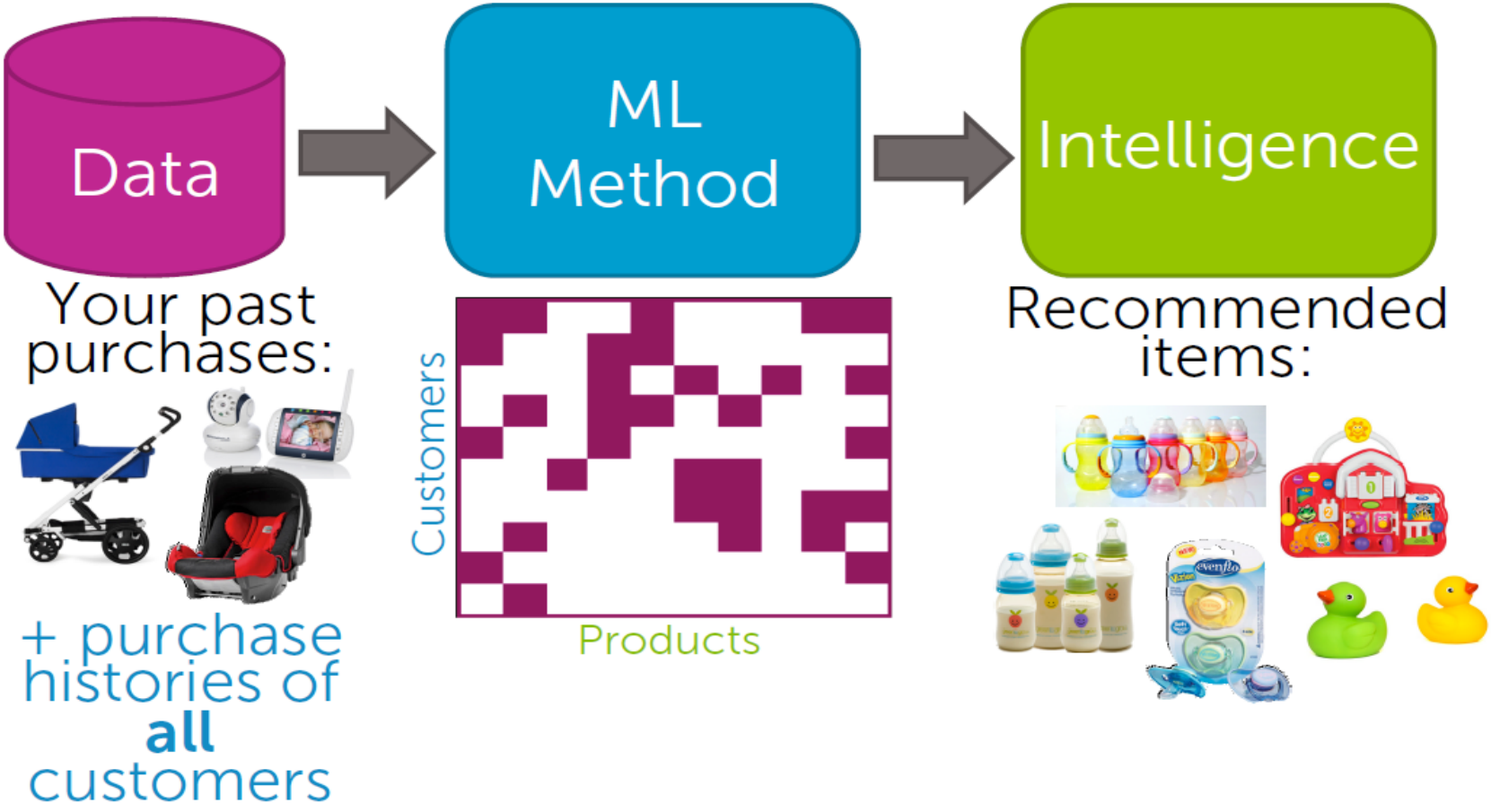
Algorithms

- KD-trees, locality-sensitive hashing (LSH)
- K-means
- Expectation-maximization (EM)

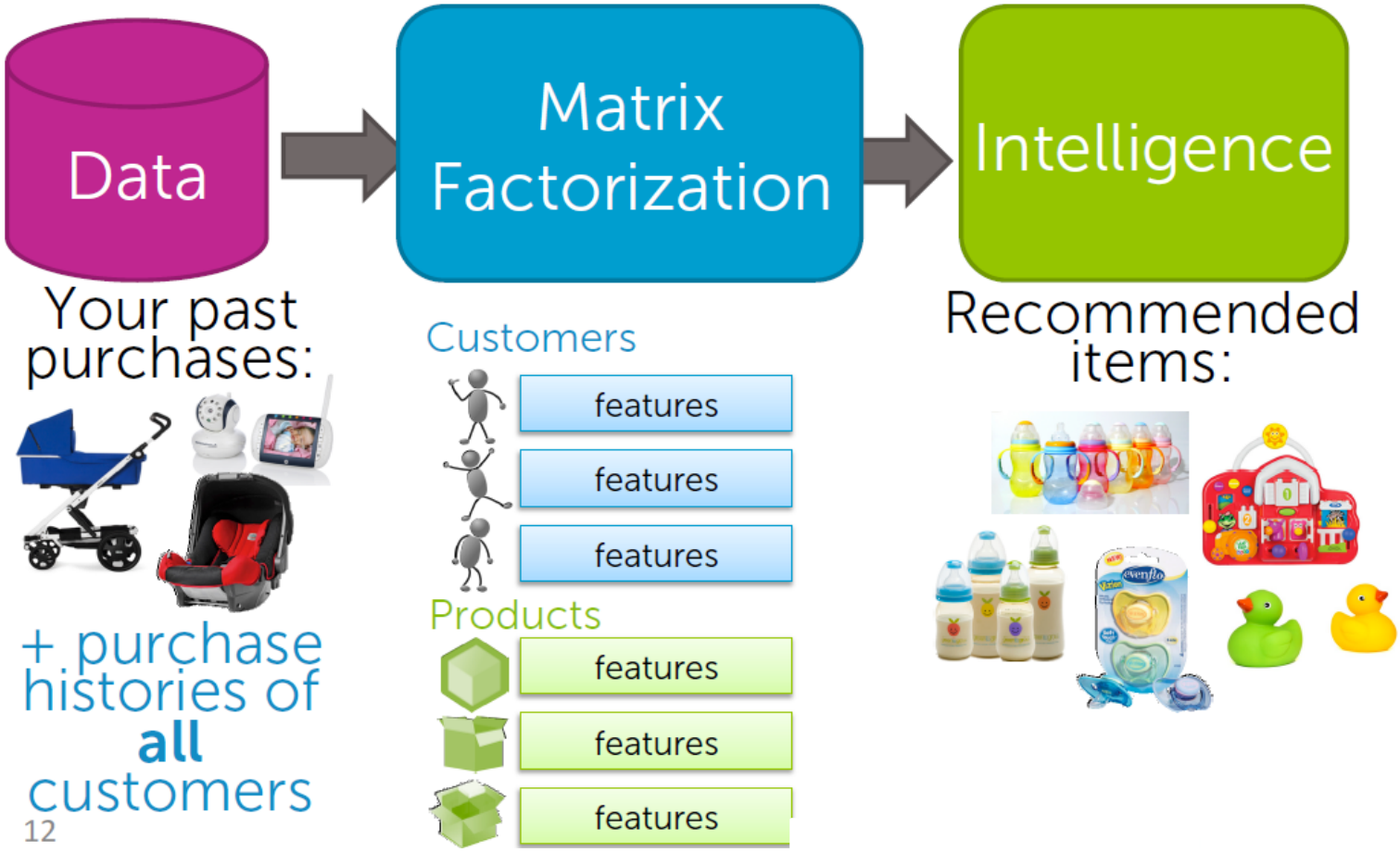
Concepts

- Distance metrics, approximation algorithms, hashing, sampling algorithms, scaling up with map-reduce

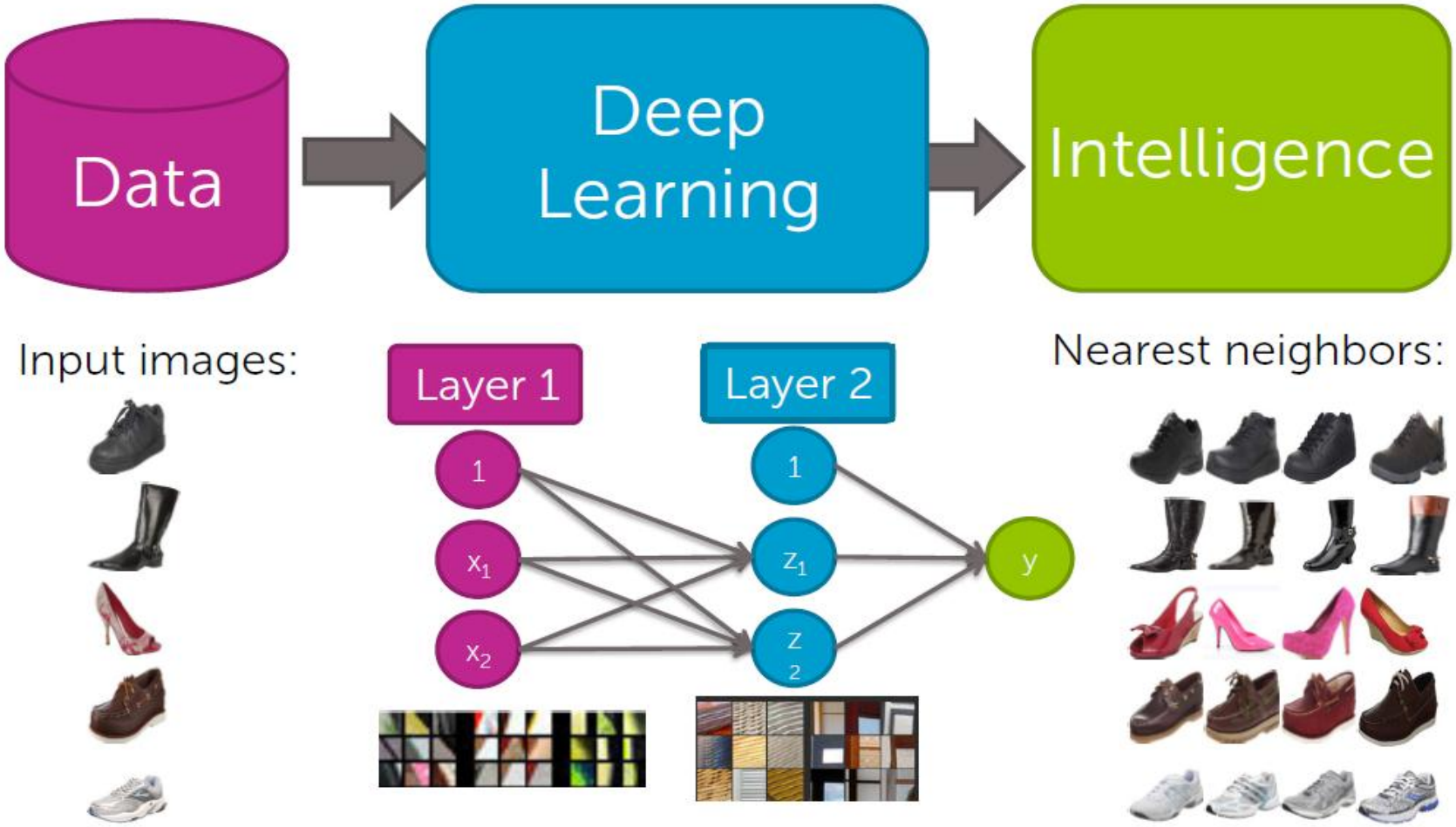
Case study 5: Recomender system



Product recommendation



Visual product recommender



Recomender systems applications

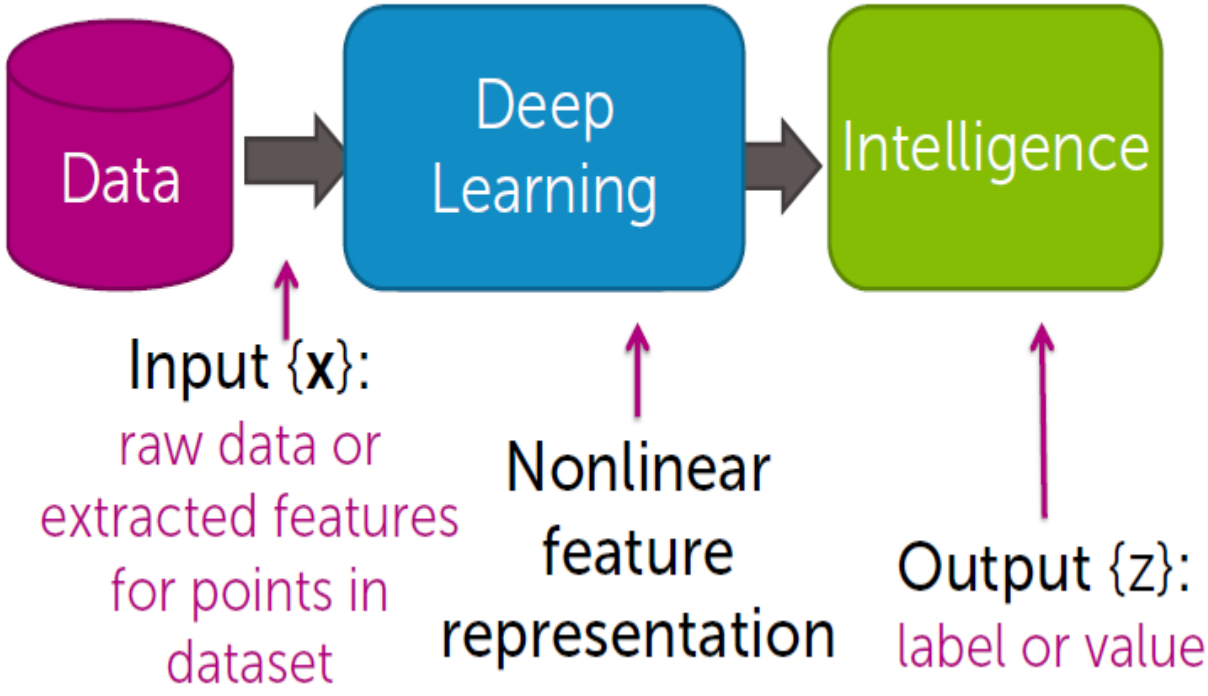
Movies

Songs

Friends, apps, ...

What is (supervised) deep learning?

Flexible method for performing classification or regression



Examples of deep learning success stories

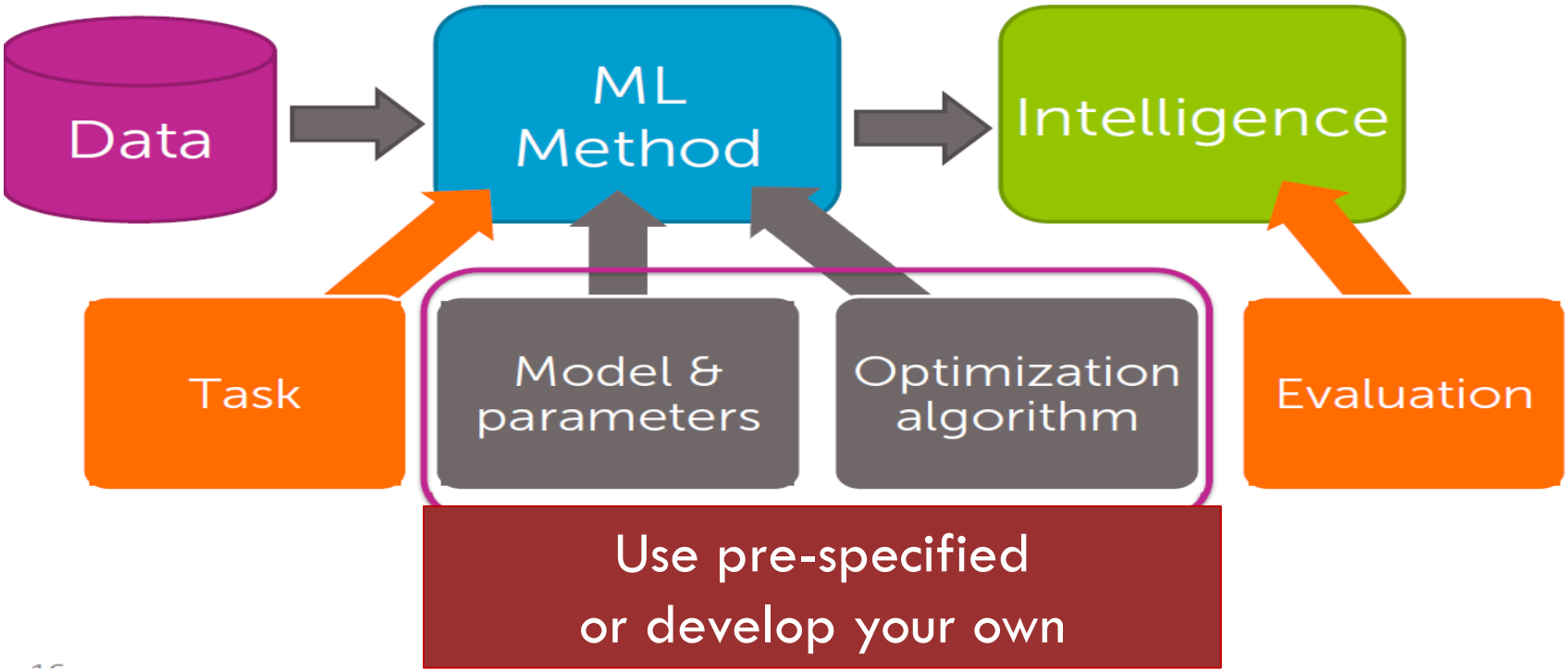
- Image classification
- Image segmentation
- Image captioning
- Object detection
- Speech recognition
- Speech synthesis
- Machine translation
- Handwriting recognition
- ...

Other ML methods

- Reinforcement learning
- Learning theory
- Active learning
- Multi-task and transfer learning
- Spectral methods
- …

Deploing inteligence module

Case studied are about building, evaluating, deploying inteligence in data analysis.



Statistical inference

- **The key concept in statistics is making conclusions about the population using information in a sample; the process is called statistical inference.**
- **By using computational methods as well as well develloped mathematical theory we can understand how one dataset differs from a different dataset –even if two dataset were collected under identical settings.**
- **Statistical inference is primarily concerned with qunatifying and understanding the uncertainty of parameter estimates. While the equations and details changes depending on the setting, the foundations for inference are the same through all the statistics.**

Foundation for inference

Hypothesis testing with randomisation

Confidence intervals with bootstrapping

Inference with mathematical models

Probability and distrbutions