

Introduction to Data Science

WFAiS UJ, Informatyka Stosowana
I stopień studiów

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

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?

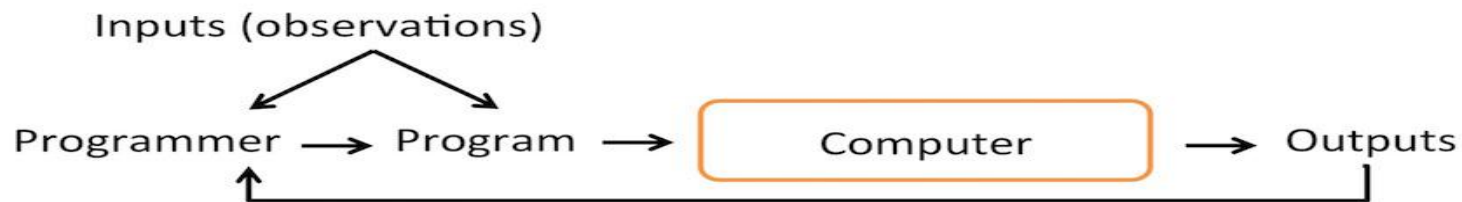
Machine learning algorithms are algorithms that learn (often predictive) models from data. I.e., instead of formulating "rules" manually, a machine learning algorithm will learn the model for you.

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

The Traditional Programming Paradigm



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

Machine Learning

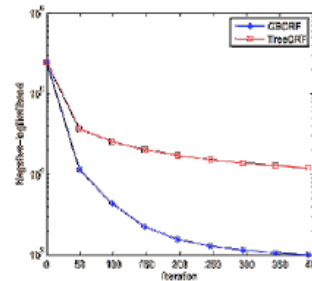


Outline of the course

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

Analyse data with Machine Learning

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

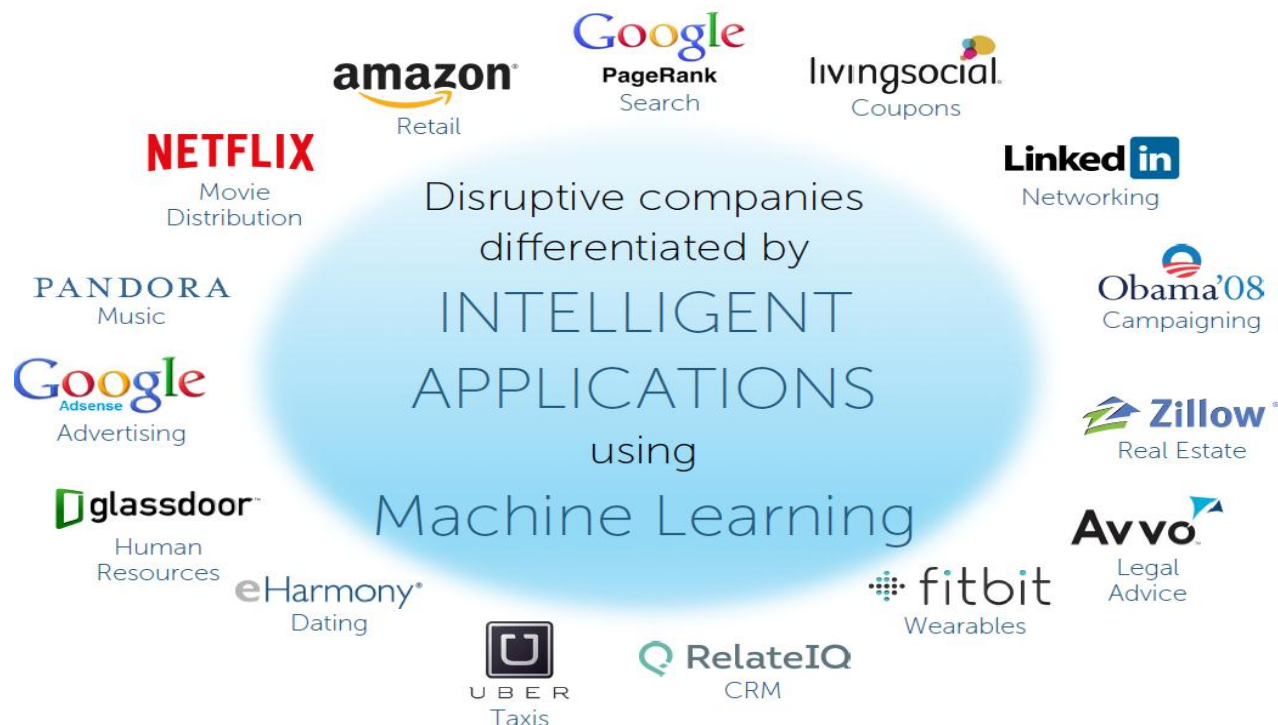


 Neural Information
Processing Systems
Foundation

ICML

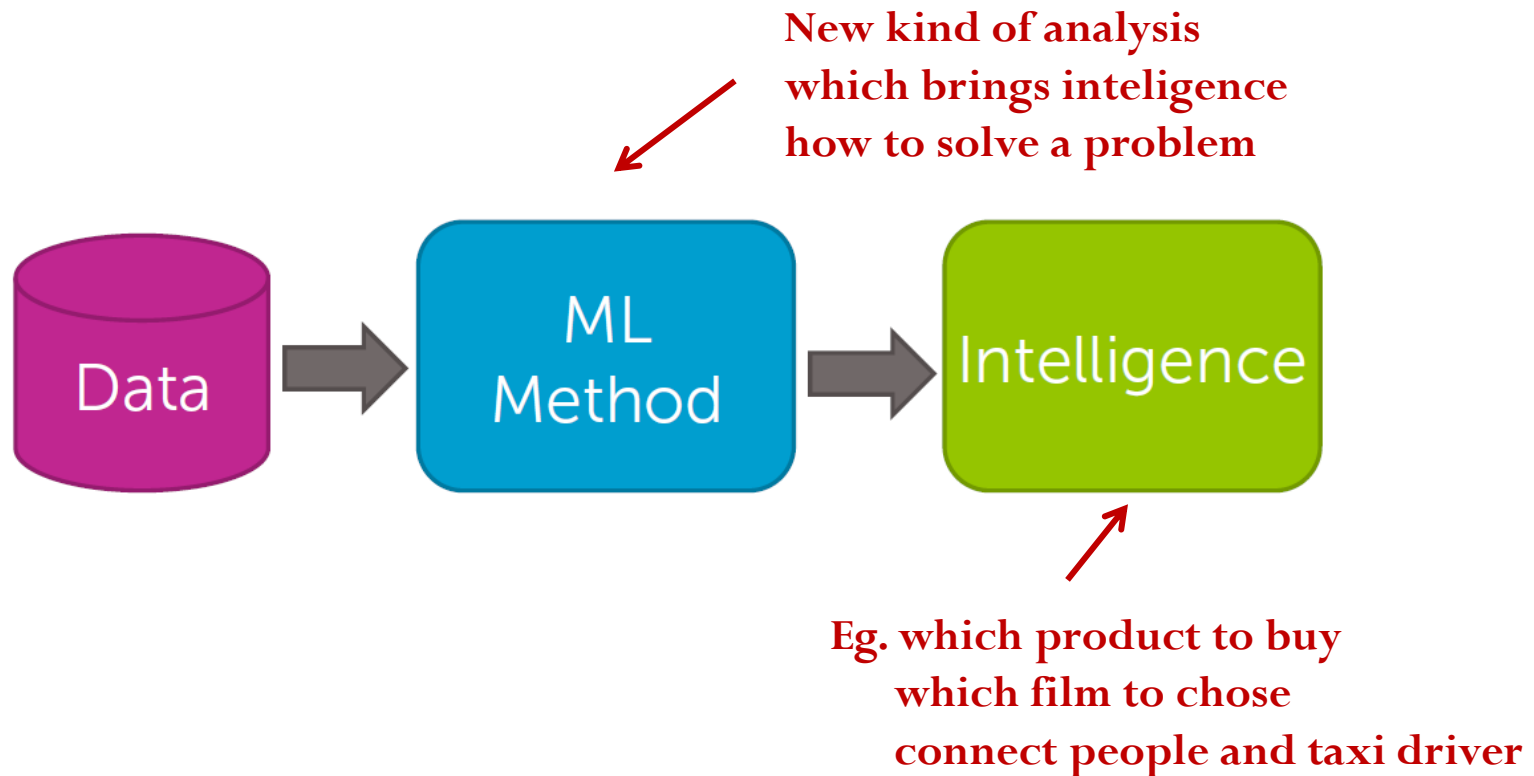
Machine learning is changing the world

- **Current view: disruptive intelligent applications are used by leading commercial companies**

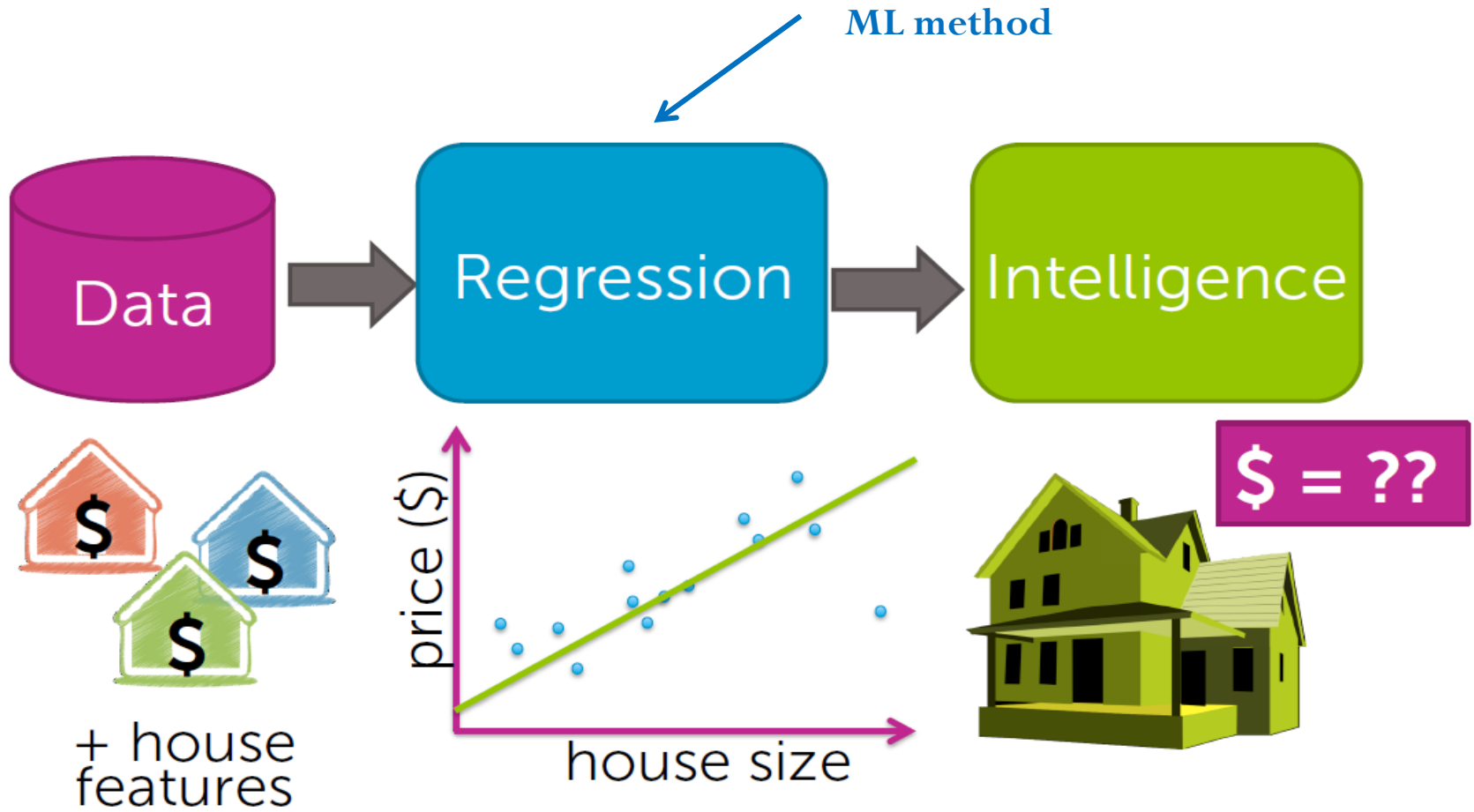


Machine learning

- Data → intelligence 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

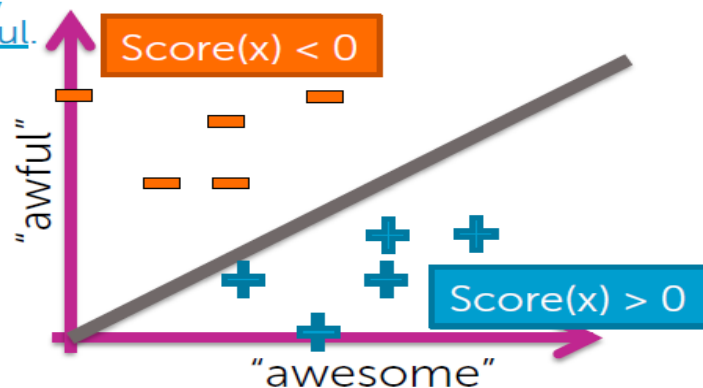
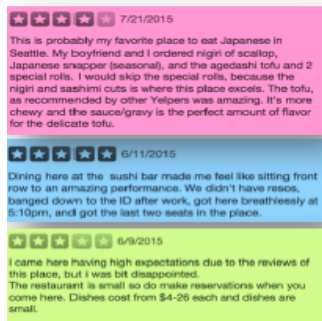
- Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection

Case study 2: Classification



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

All reviews:



Classification

Spam filtering

Osman Khan to Carlos show details Jan 7 11 days ago Re: Reply

sounds good
OK

Carlos: Good to hear!
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos

Welcome to New Media Installation: Art that Learns

Carlos: Questions to 10013-announce, Camen, Michel show details 3:13 PM (8 hours ago) Re: Reply

Hi everyone,

Welcome to New Media Installation Art that Learns

The class will start tomorrow.

Make sure you attend the first class, even if you are on the Wait List

The classes are held in Corbally Hall C305, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list. 10013-announce@osu.edu

You can contact the instructors by emailing: 10013-course-staff@osu.edu

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey. Free Trial 1 bottle, pay only \$5.95 for shipping mfw rik show details

Jacquelyn Hickey to rhenish, bcc: thehomey, bcc: ang show details 9:52 PM (1 hour ago) Re: Reply

==== Natural WeightLOSS Solution =====

Vital Acai is a natural WeightLOSS product that Enables people to lose weight and cleanse their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in their striving that they never thought they could.

- * Rapid WeightLOSS
- * Increased Metabolism - Burn Fat & calories easily!
- * Better Mood and Attitude
- * More Self-Confidence
- * Cleanse and Detoxify Your Body
- * Much More Energy

Text of email,
sender, IP,...

Not spam

Spam

Input: x

Output: y

Multiclass classifier

Output y has more than 2 categories



Input: x
Webpage

Output: y

Classification

Image classification



Input: x
Image pixels

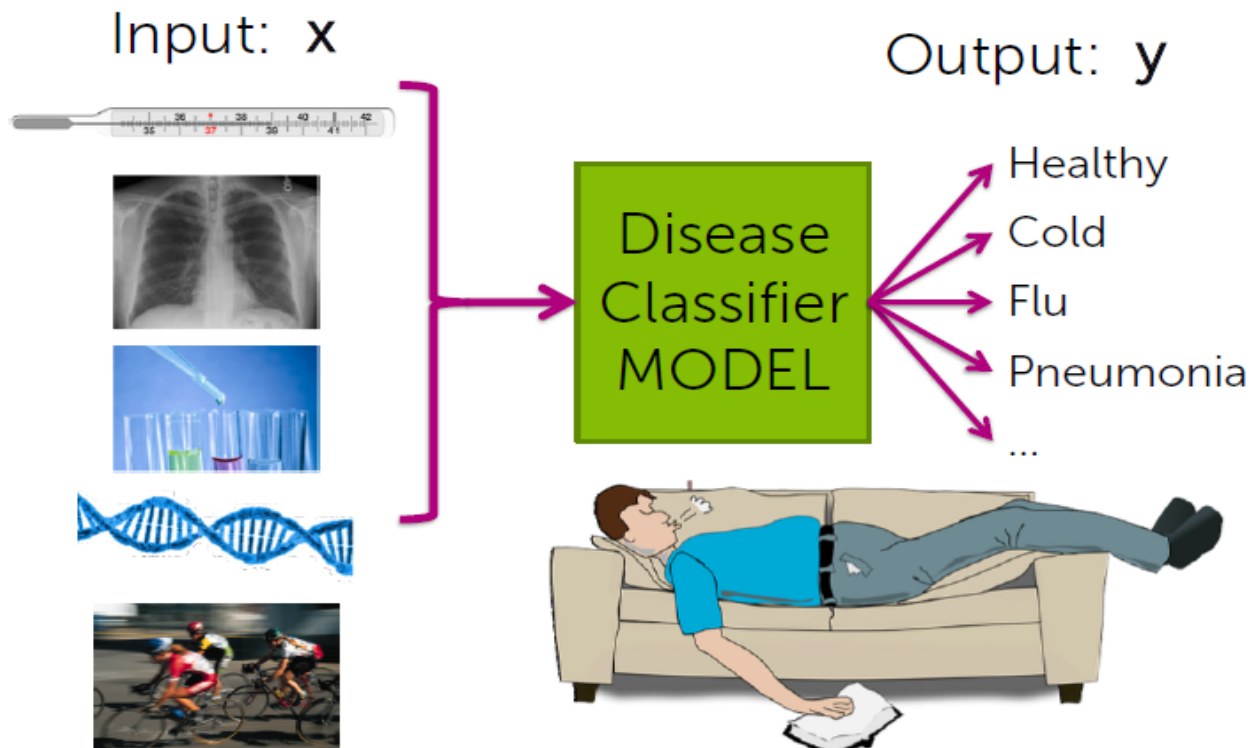


Output: y
Predicted object

Soft prediction

Classification

Personalized medical diagnosis



Classification:

Models

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

Algorithms

- Stochastic gradient descent
- Boosting

Concepts

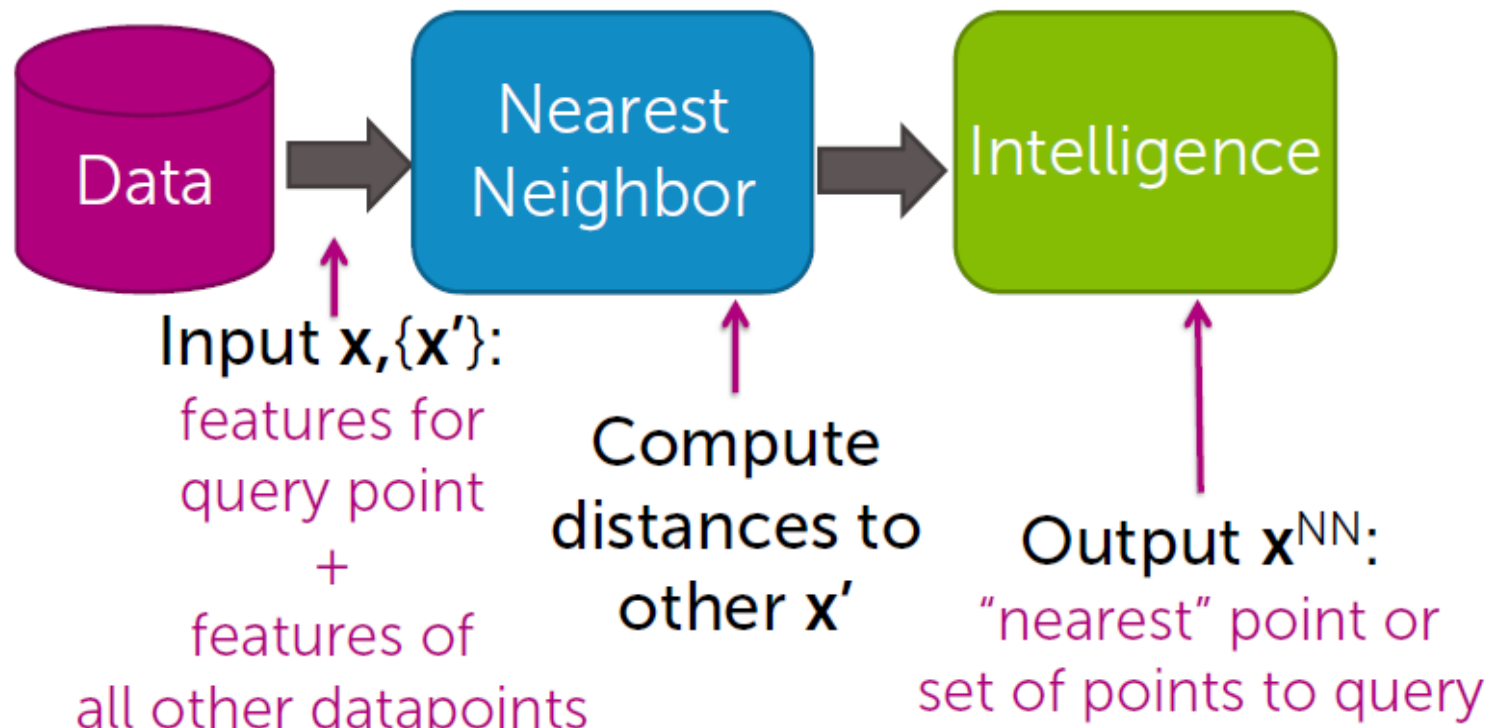
- Decision boundaries, MLE, ensemble methods, random forests, CART, online learning

Case Study 3: Document retrieval



Retrieval

Search for related items



Retrieval

Retrieve “nearest neighbor” article

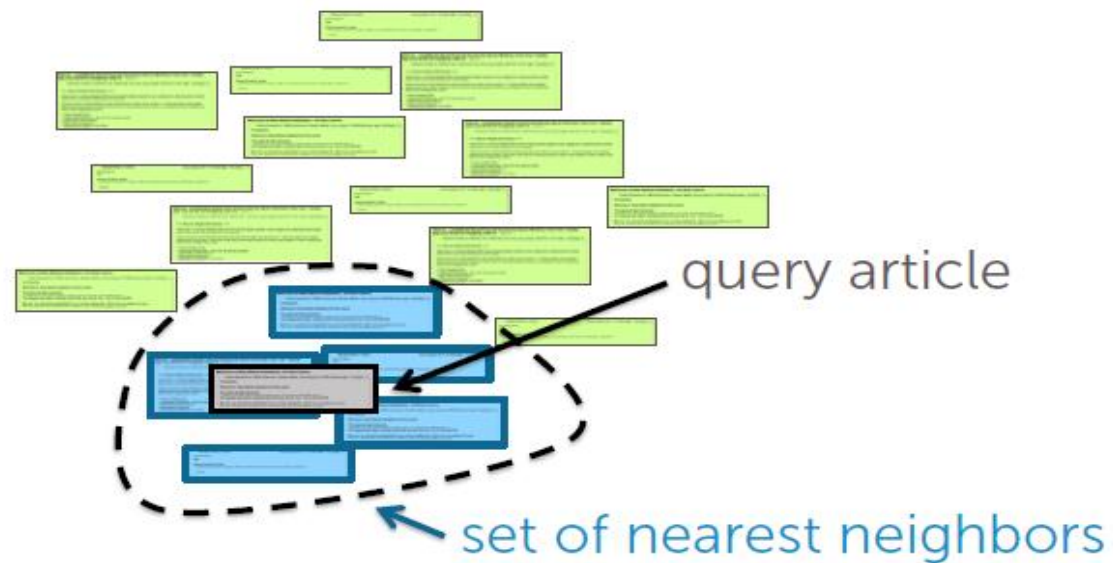
Space of all articles,
organized by similarity of text



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

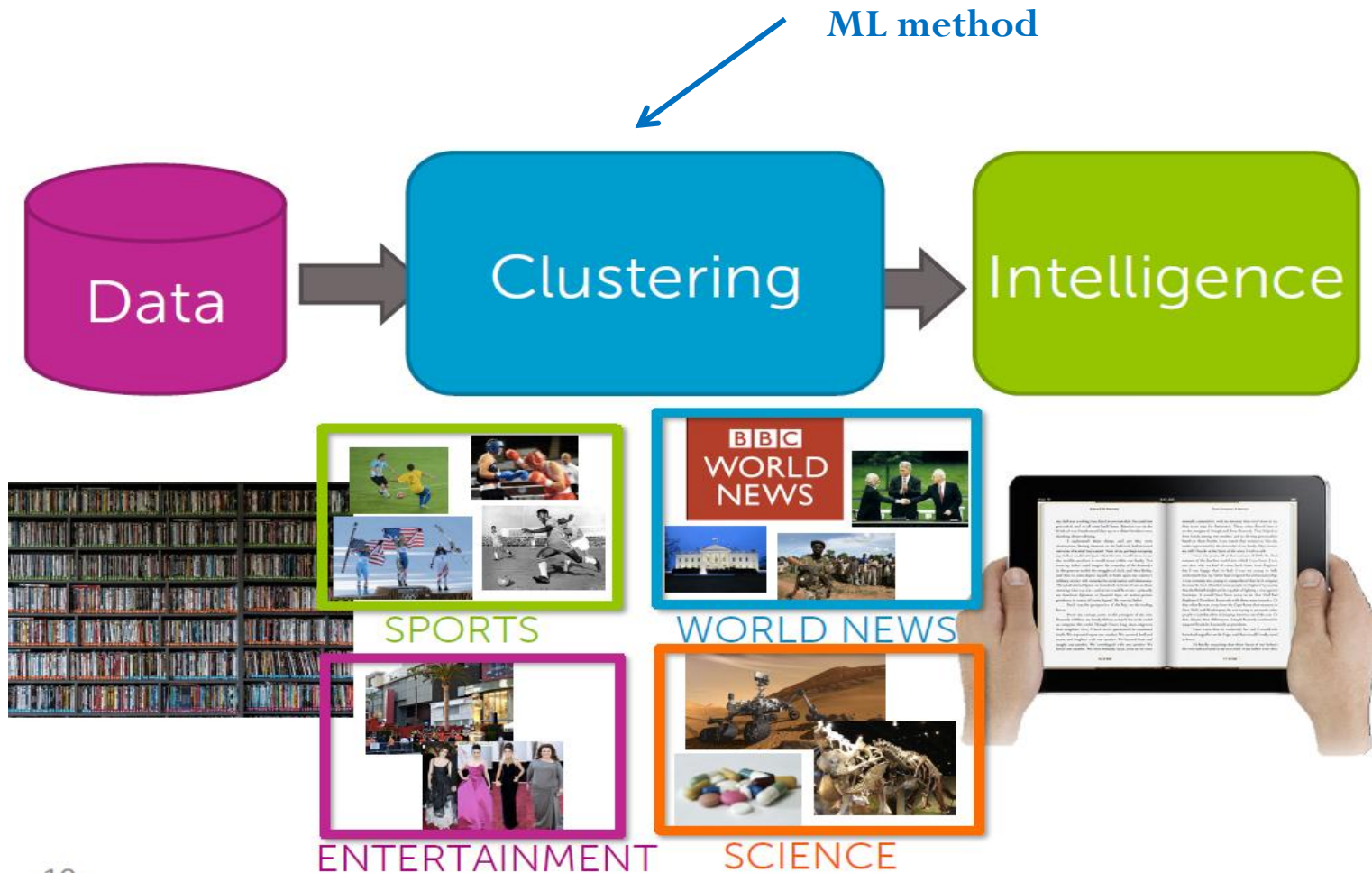


Social networks

(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

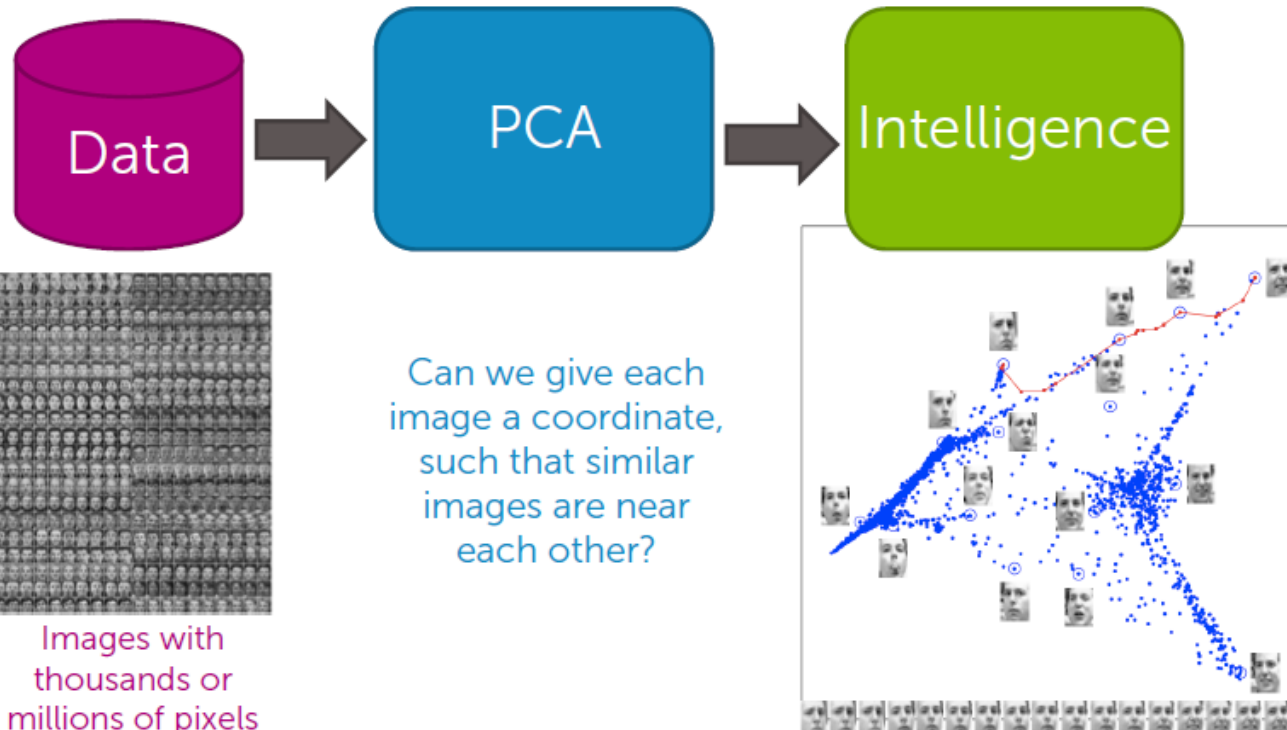
Or users on websites...

Discover groups of users for better targeting of content



Embedding

Example: Embedding images to visualize data



Clustering: Finding documents

Models

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

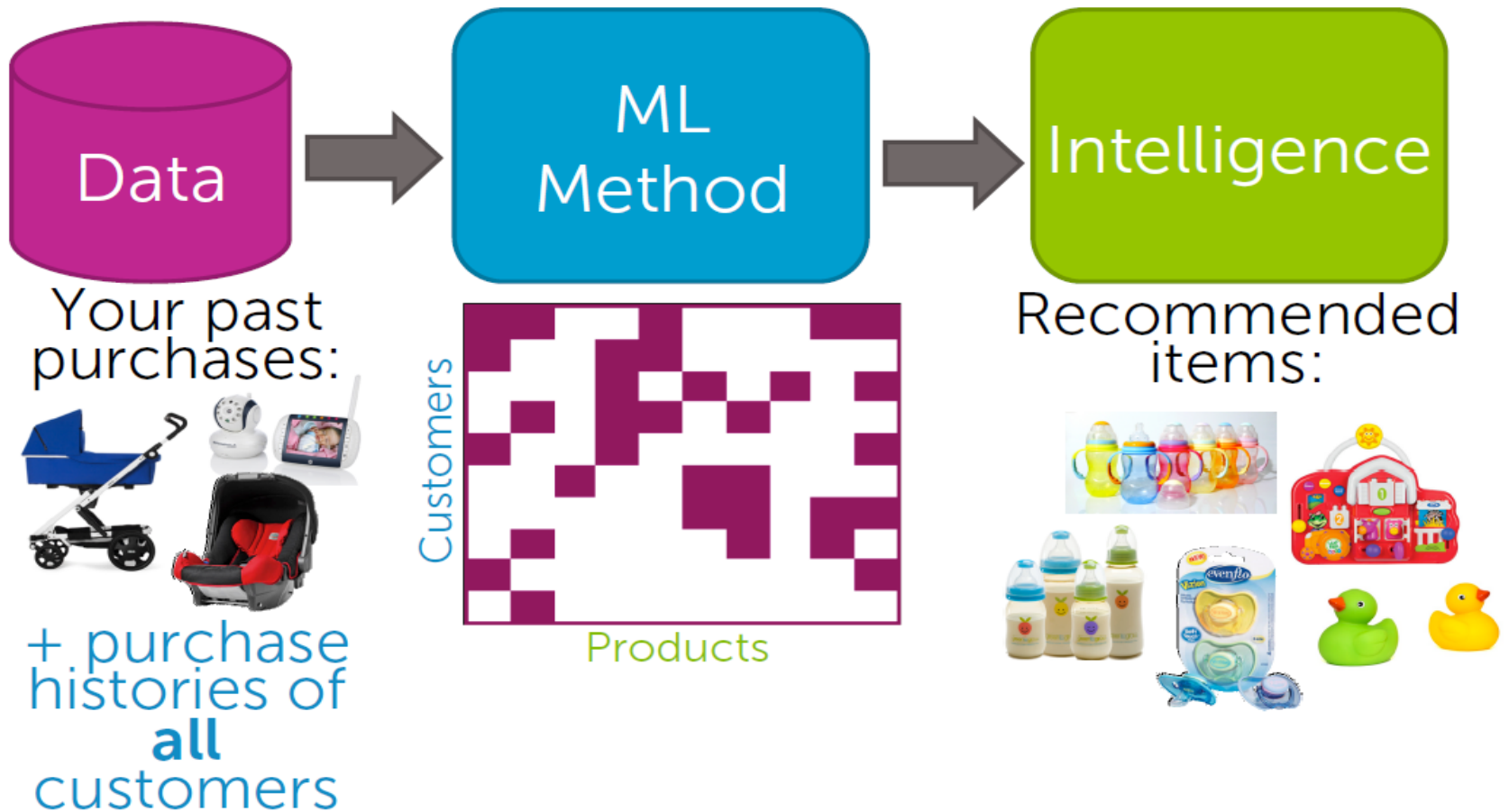
Algorithms

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

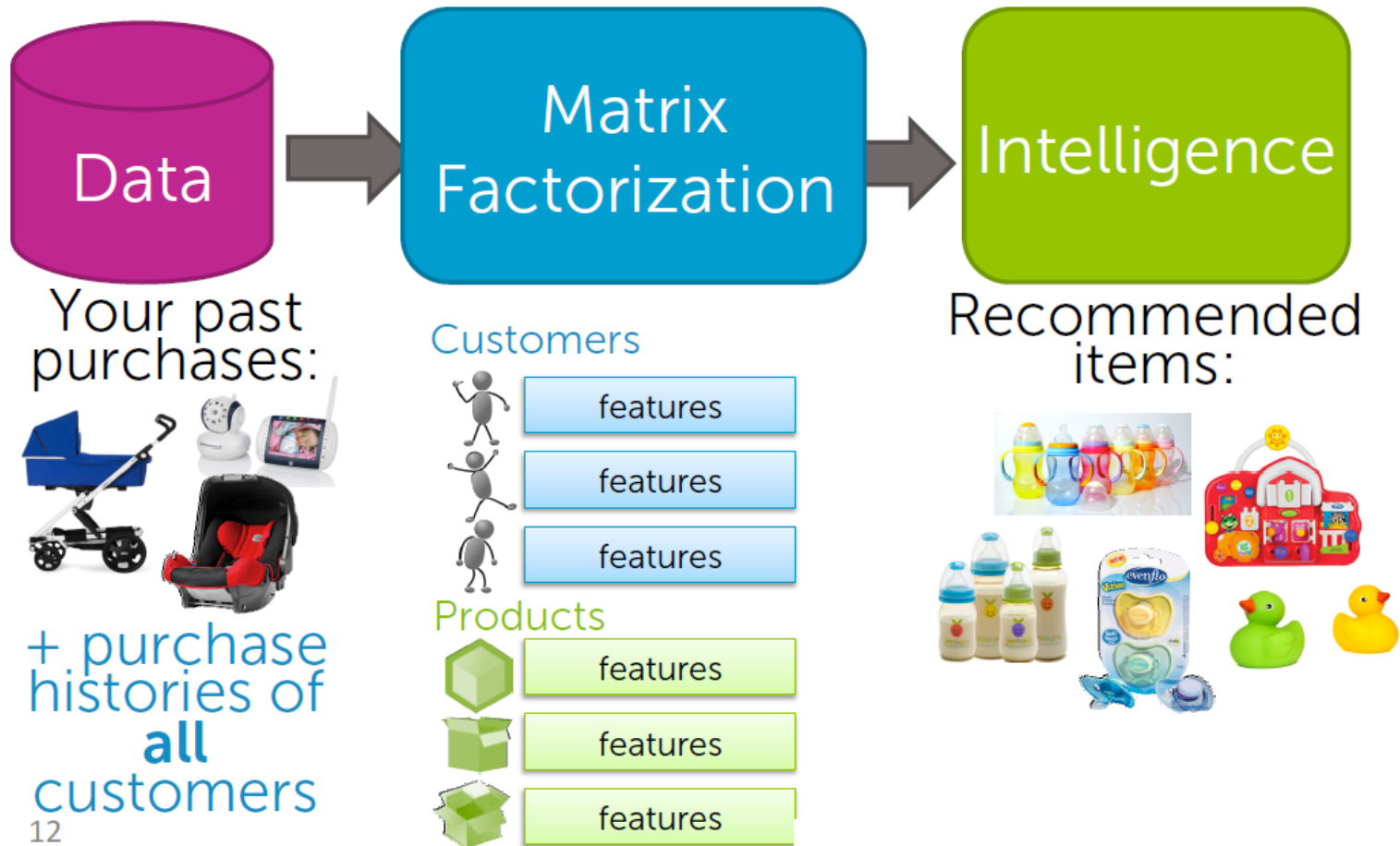
Concepts

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

Case study 5: Recommender system



Product recommendation



Visual product recommender



Input images:



Layer 1

1

x_1

x_2



Layer 2

1

z_1

z_2



Nearest neighbors:



y

Recomender systems applications



Movies



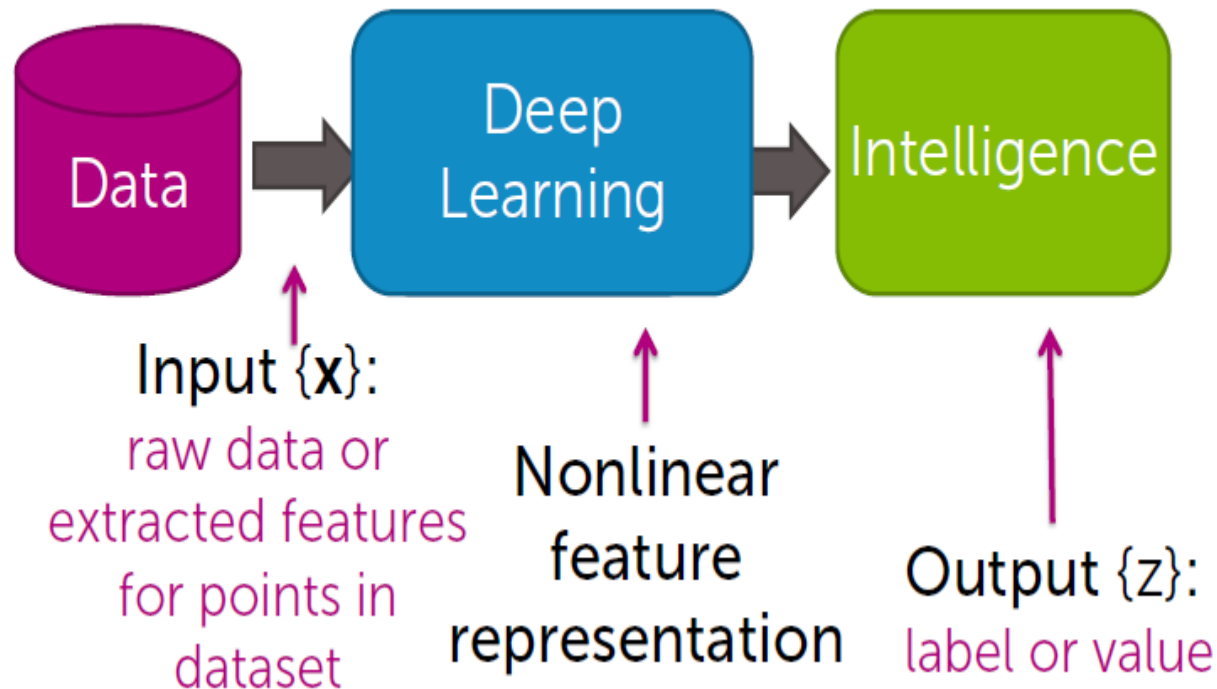
Songs



Friends, apps, ...

What is supervised (deep) learning

Flexible method for performing classification or regression



Examples of deep learning success stories

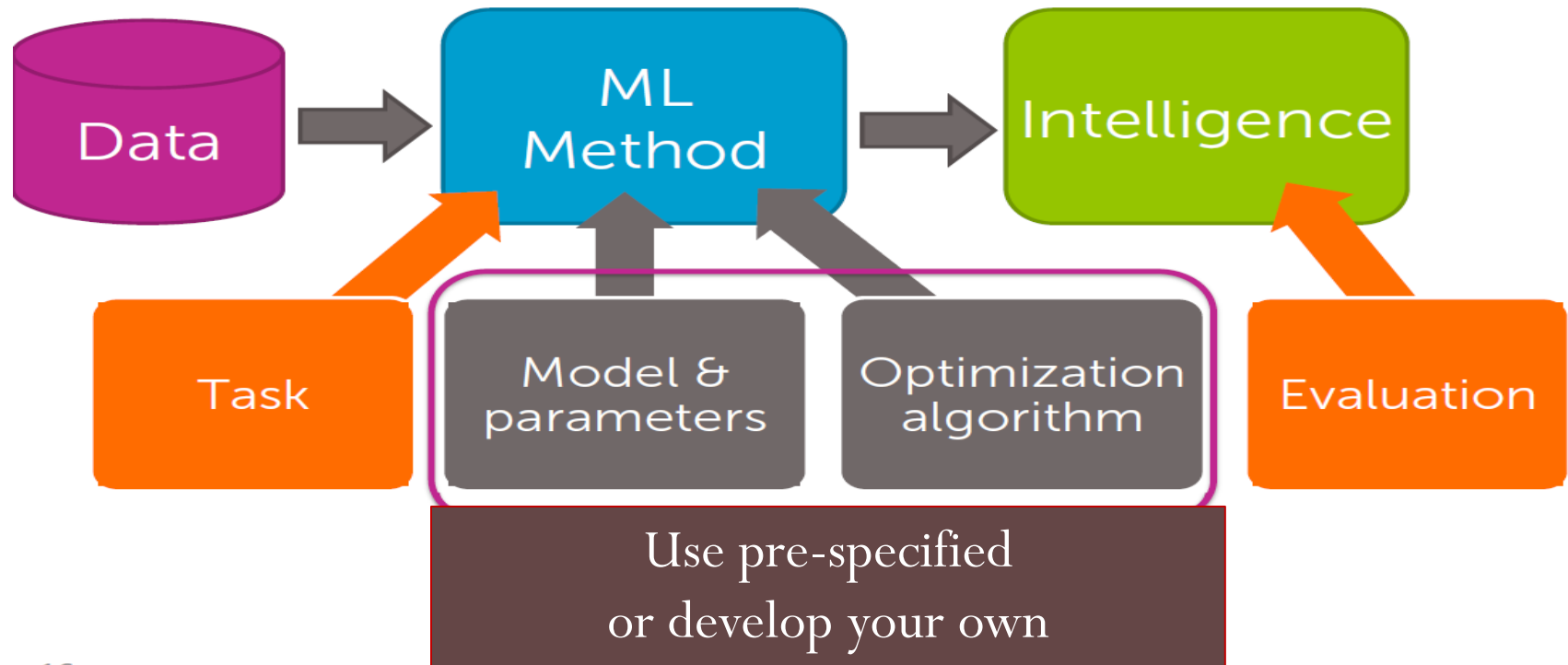
- Image classification
- Image segmentation
- Image captioning
- Object detection
- Speech recognition
- Speech synthesis
- Machine translation
- Handwriting recognition
- ...

Other ML methods

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

Deploying intelligence module

Case studied are about building, evaluating, deploying intelligence 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 developed 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 quantifying 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

- **Probability and distributions**
- **Hypothesis testing with randomisation**
- **Confidence intervals with bootstrapping**
- **Inference with mathematical models**