

# 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

4/10/2023

WFAiS UJ, Informatyka Stosowana  
I stopień studiów

# What is Data Science?

2

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?

3

**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.

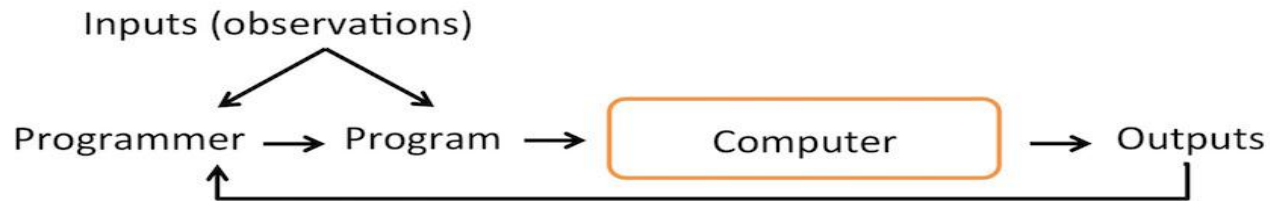
**Machine learning** - at its core - is about the use and development of these learning algorithms. **Data science** is more about the extraction of knowledge from data to answer particular question or solve particular problems.

**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

4

## 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

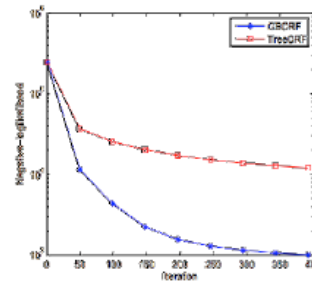
5

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

# Analyse data with Machine Learning

6

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



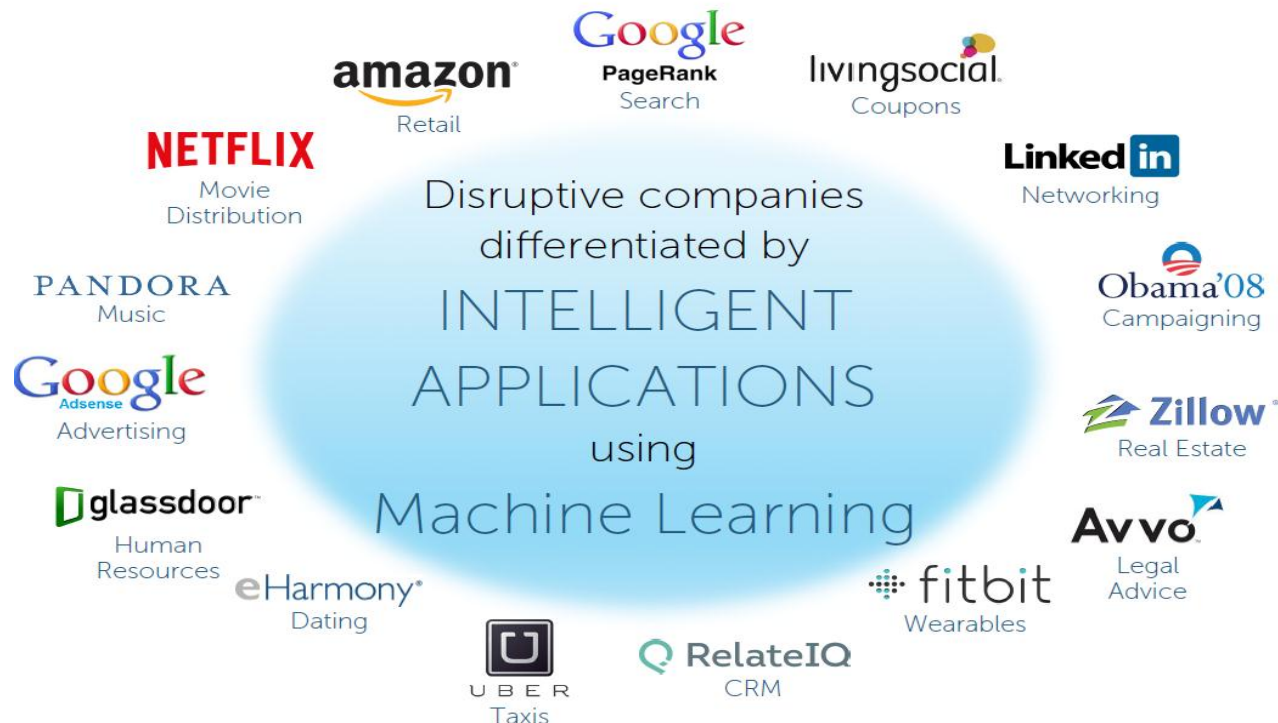
 Neural Information  
Processing Systems  
Foundation

**ICML**

# Machine learning is changing the world

7

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

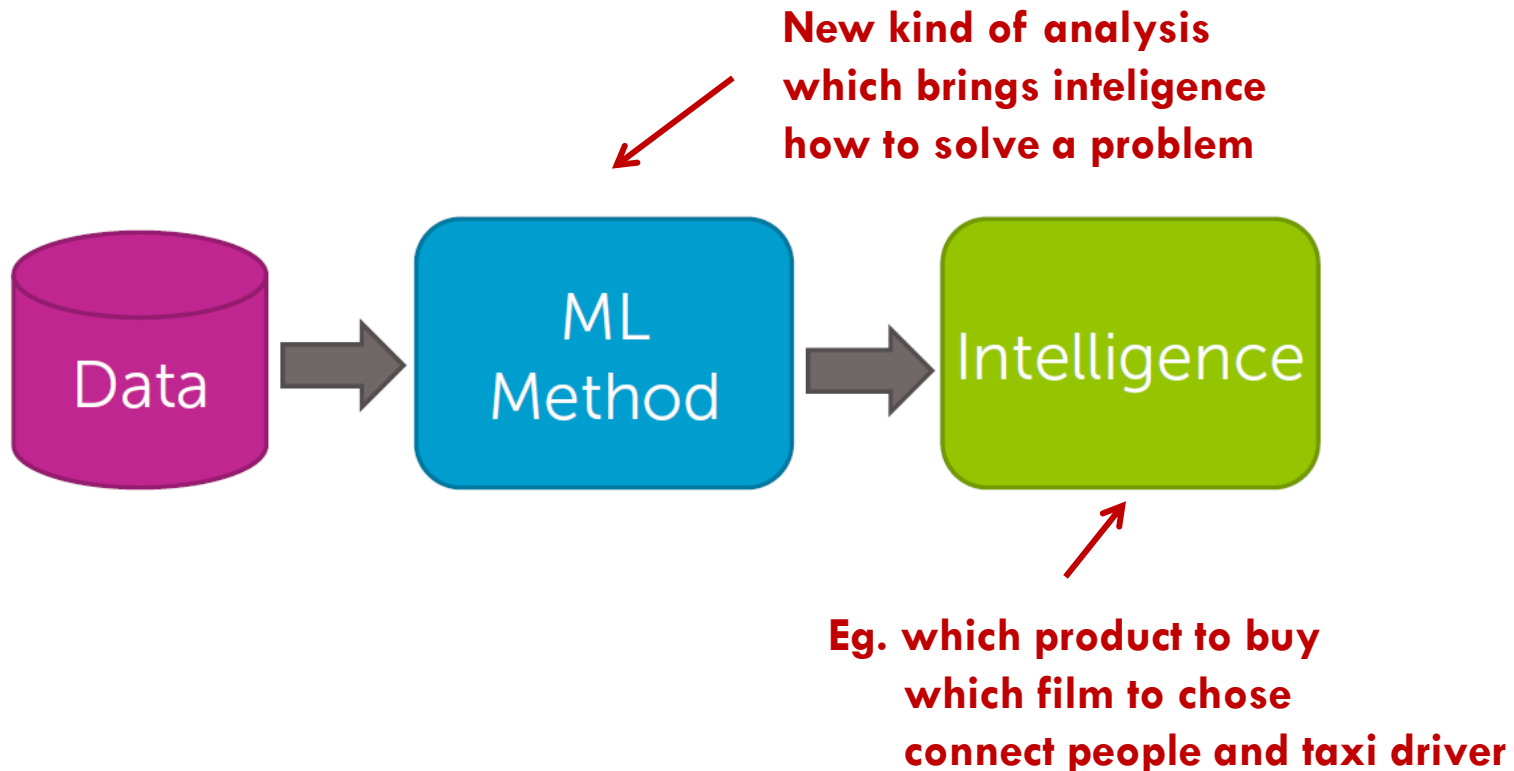


4/10/2023

# Machine learning

8

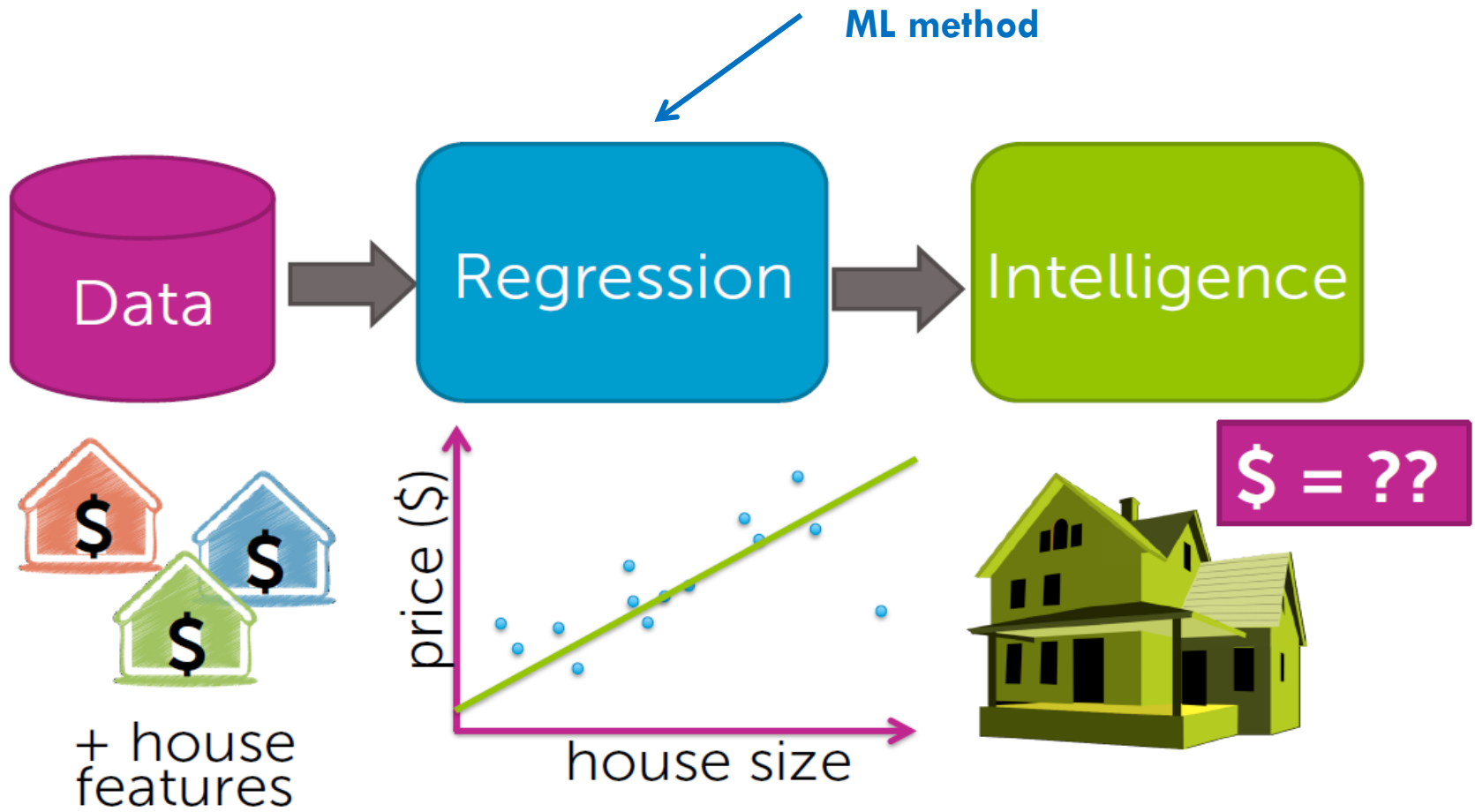
## □ Data → intelligence pipeline





# Case study 1: Prediction

9



# Prediction

10



hard work



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

# Prediction

11

## 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:

12

## Models

- Linear regression
- Regularization: Ridge (L2), Lasso (L1)

## Algorithms

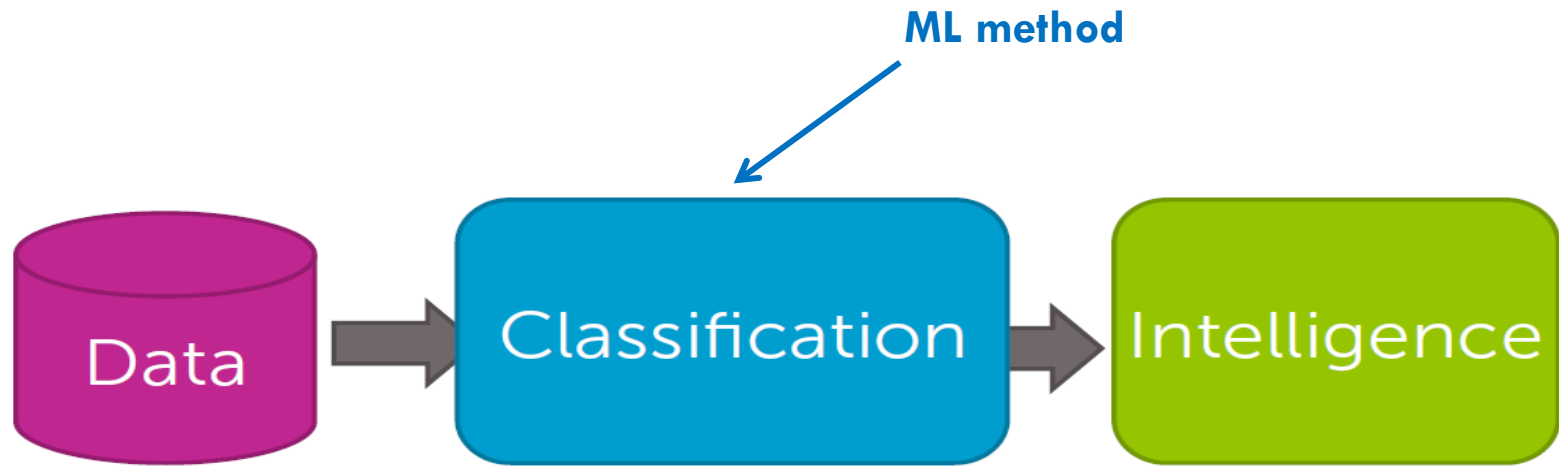
- Gradient descent
- Coordinate descent

## Concepts

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

# Case study 2: Classification

13



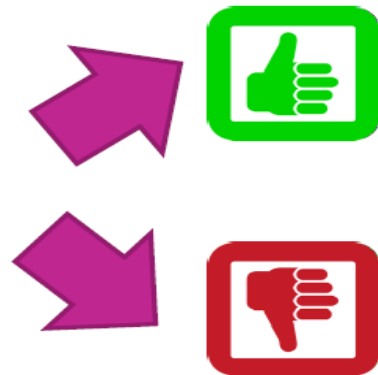
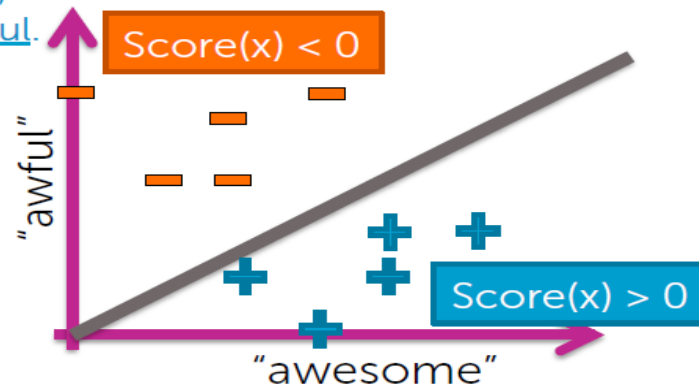
Sushi was awesome,  
the food was awesome,  
but the service was awful.

All reviews:

7/21/2015  
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese snapper (seasonal), and the agedashi tofu and 2 special rolls. I would skip the special rolls, because the nigiri and sashimi cuts is where this place excels. The tofu, as recommended by other Yelpers was amazing. It's more chewy and the sauce/gravy is the perfect amount of flavor for the delicate tofu.

5/11/2015  
Dining here at the sushi bar made me feel like sitting front row to an amazing performance. We didn't have reservations, banged down to the ID after work, got here breathlessly at 5:10pm, and got the last two seats in the place.

6/9/2015  
I came here having high expectations due to the reviews of this place, but I was bit disappointed. The restaurant is small so do make reservations when you come here. Dishes cost from \$4-26 each and dishes are small.



# Classification

14

## Spam filtering



Text of email,  
sender, IP, ...

Not spam

Spam

Input:  $x$

Output:  $y$

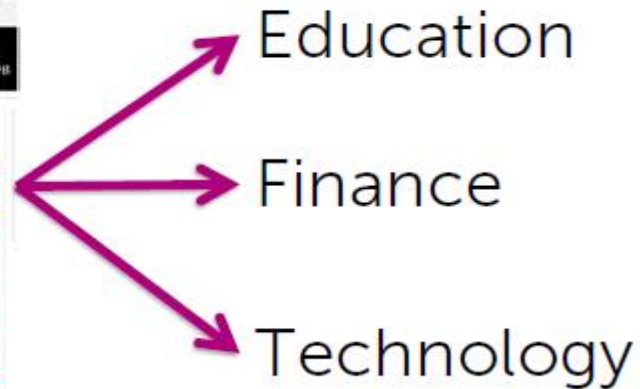
# Multiclass classifier

15

*Output  $y$  has more than 2 categories*



Input:  $x$   
Webpage



Output:  $y$

# Classification

16

## Image classification



Input:  $x$   
Image pixels



Output:  $y$   
Predicted object

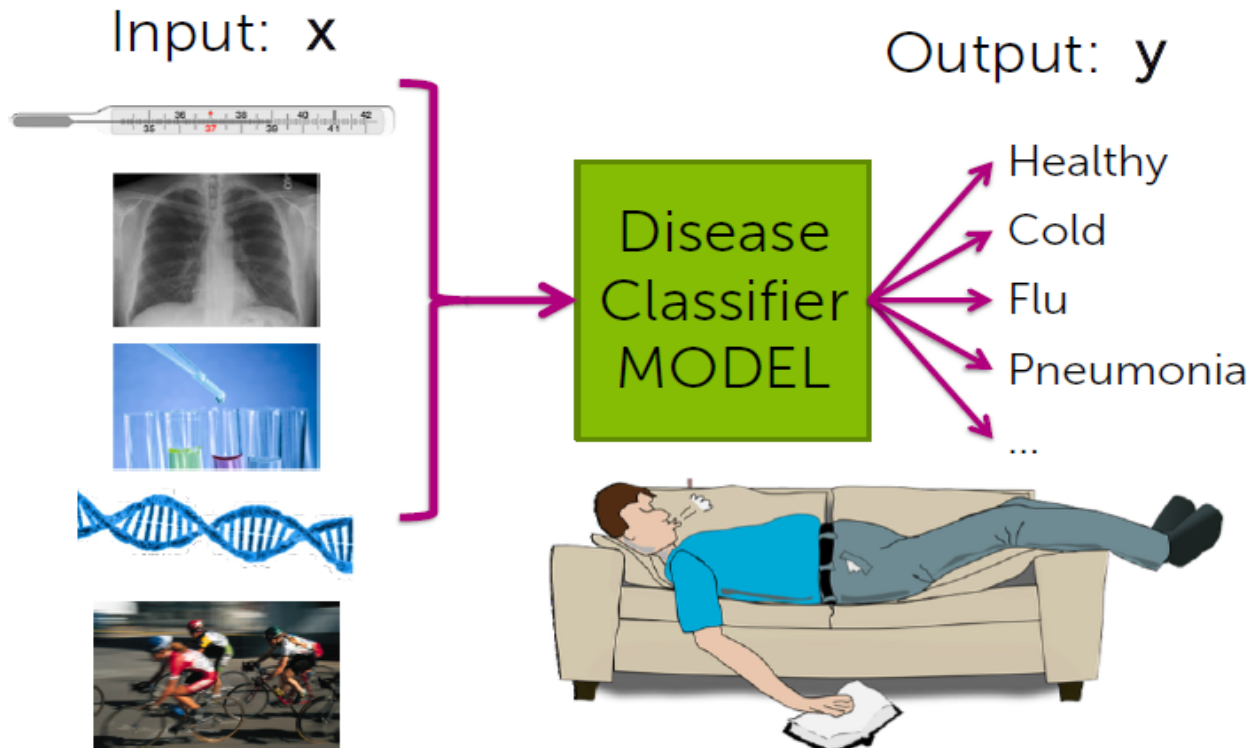
*Soft prediction*



# Classification

17

## Personalized medical diagnosis



# Classification:

18

## 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: document retrieval

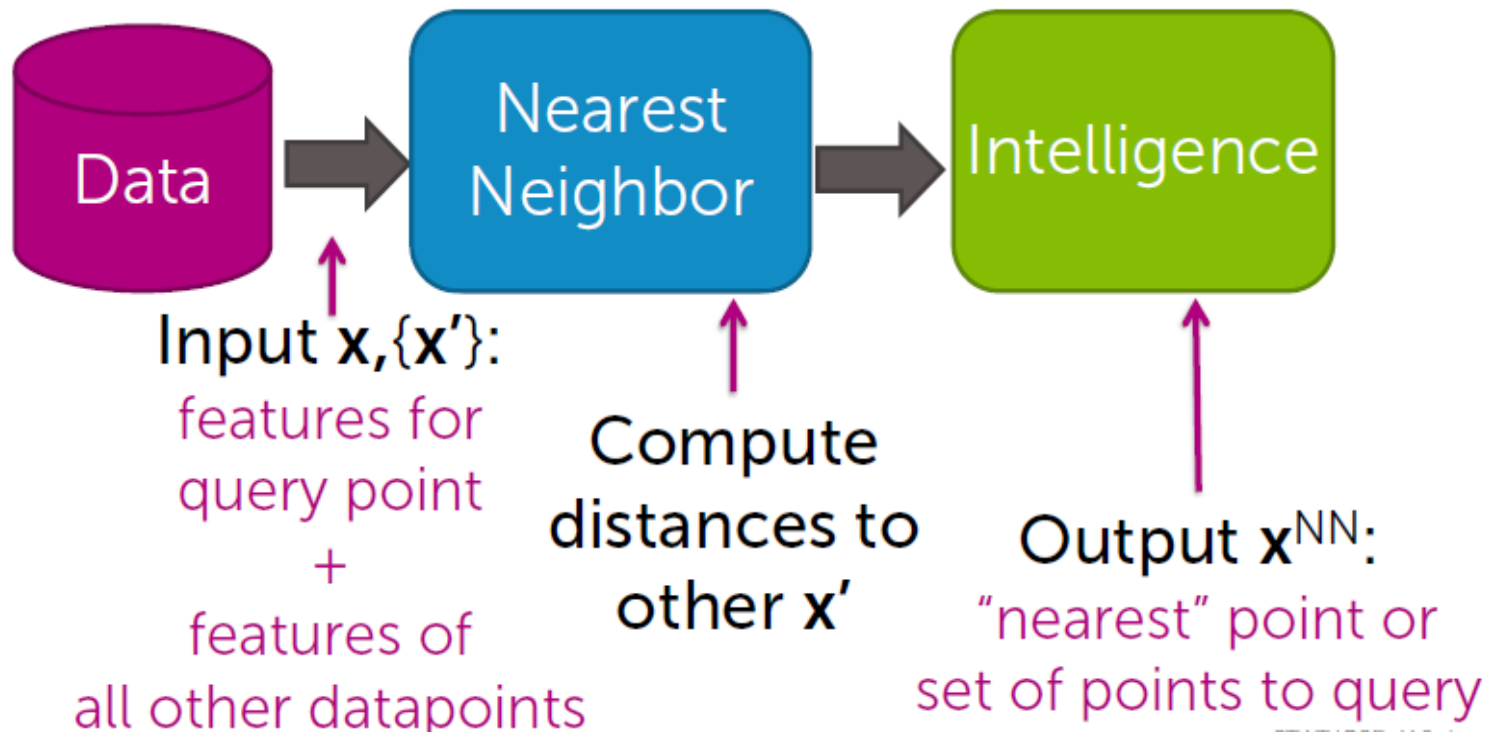
19



# Retrieval

20

Search for related items



# Retrieval

21

## Retrieve “nearest neighbor” article

Space of all articles,  
organized by similarity of text

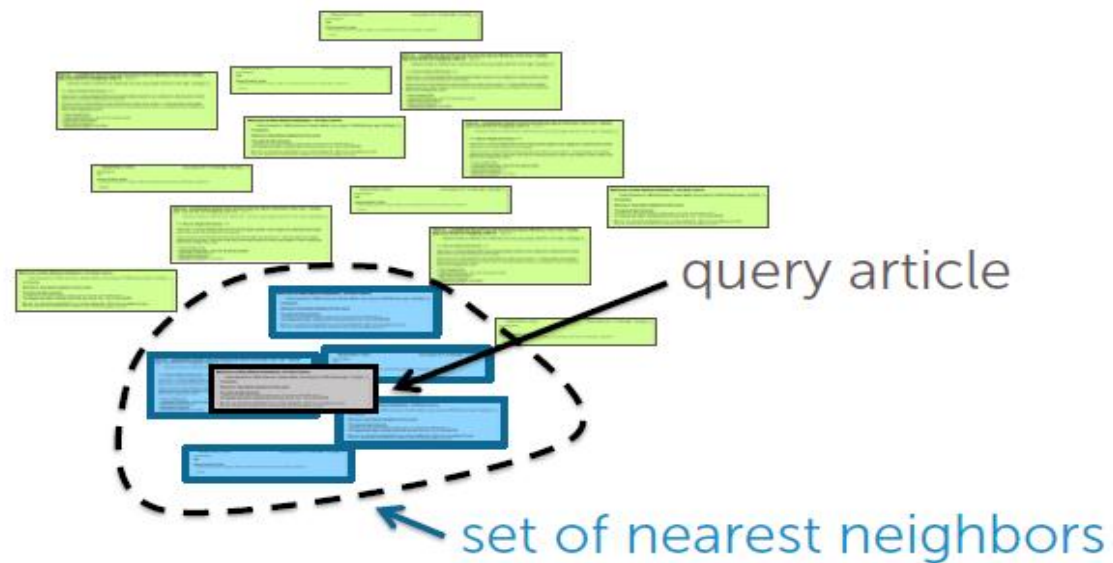


# Retrieval

22

## Or set of nearest neighbors

Space of all articles,  
organized by similarity of text



# Retrieval

23

## Retrieval applications

Just about everything...

Images



Products



Streaming content:

- Songs
- Movies
- TV shows
- ...

News articles



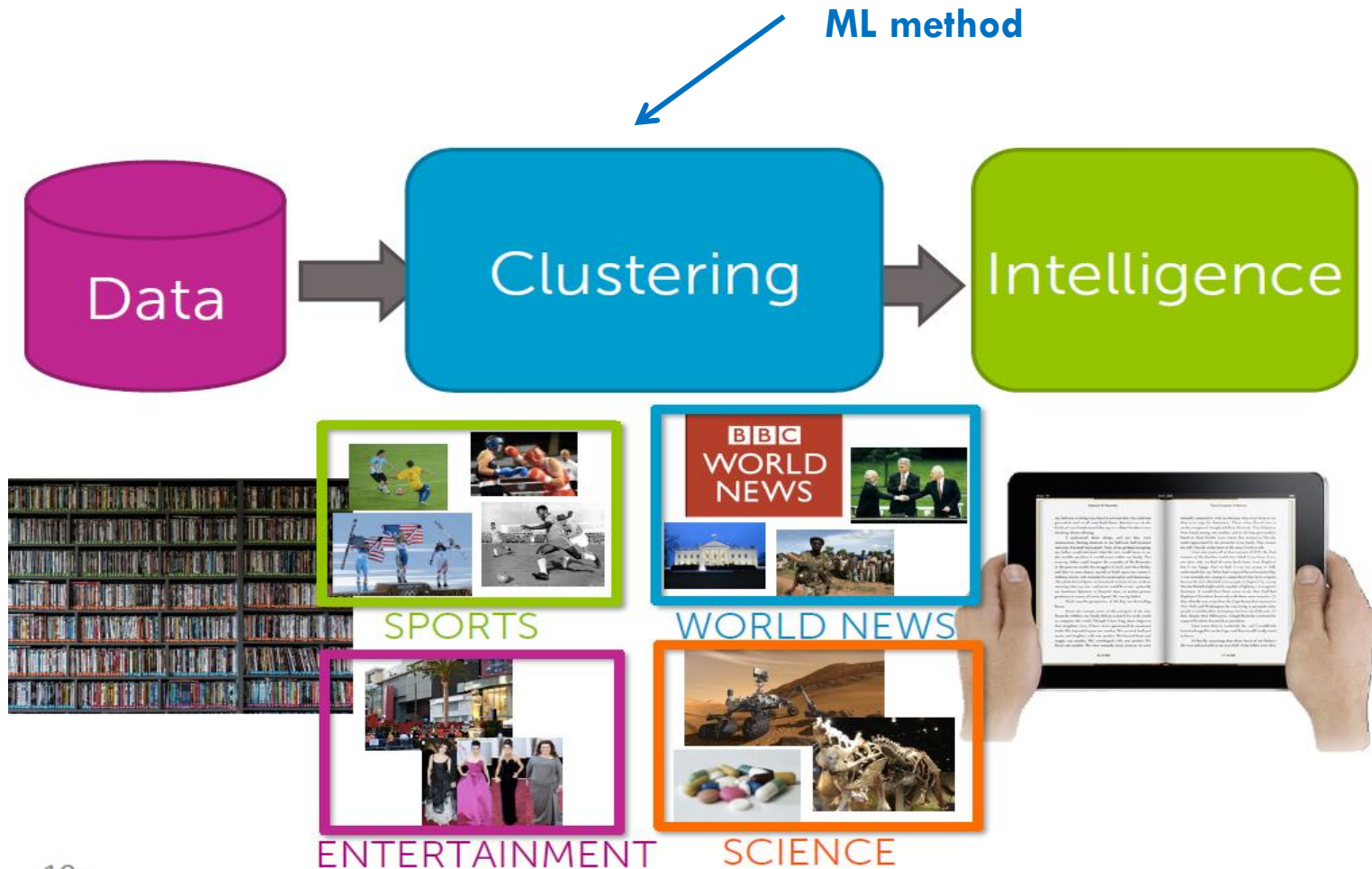
Social networks

(people you might want to connect with)



# Case study 3++: Document structuring for retrieval

24



10



# Clustering

25

## Clustering images

For search, group as:

- Ocean
- Pink flower
- Dog
- Sunset
- Clouds
- ...



# Clustering

26

## Or users on websites...

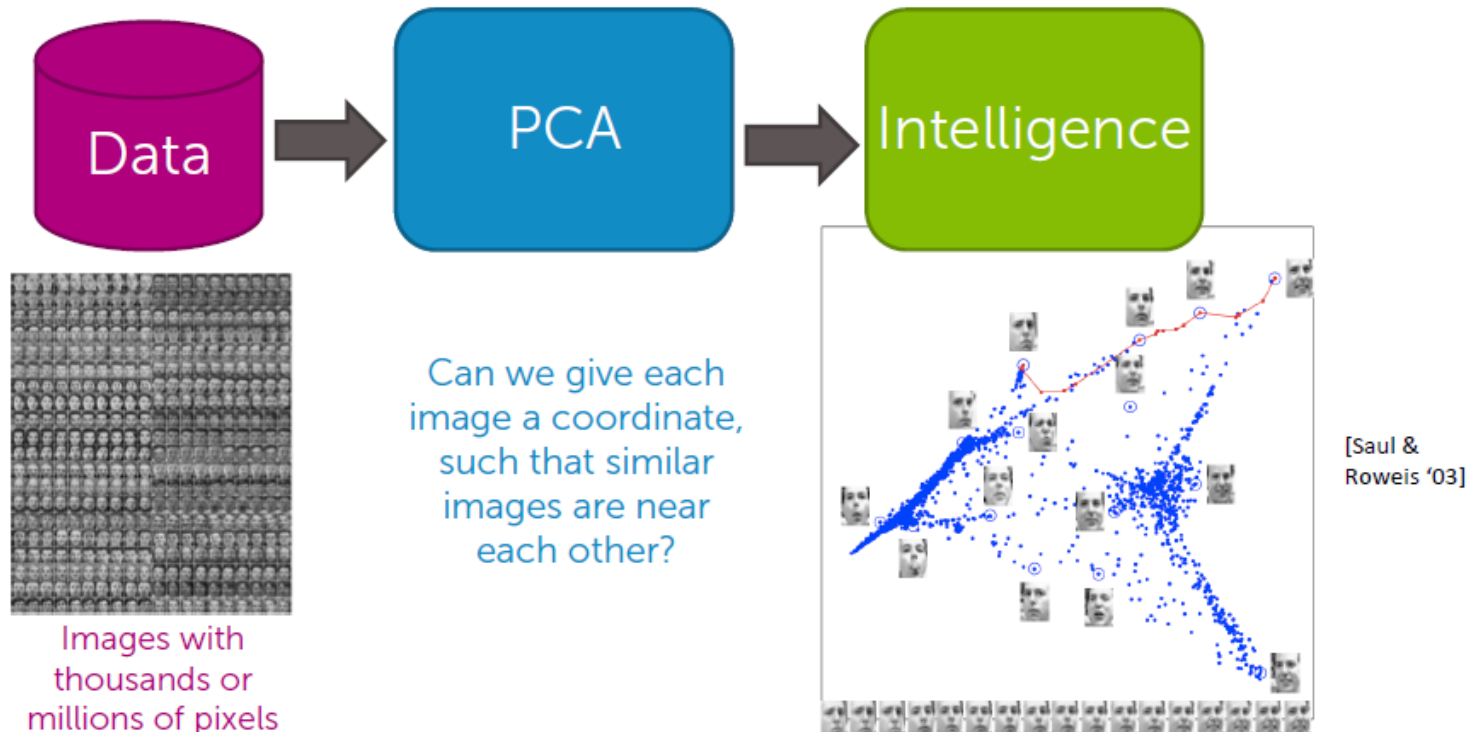
Discover groups of users for better targeting of content



# Embedding

27

Example: Embedding images to visualize data



# Clustering: Finding documents

28

## Models

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

## Algorithms

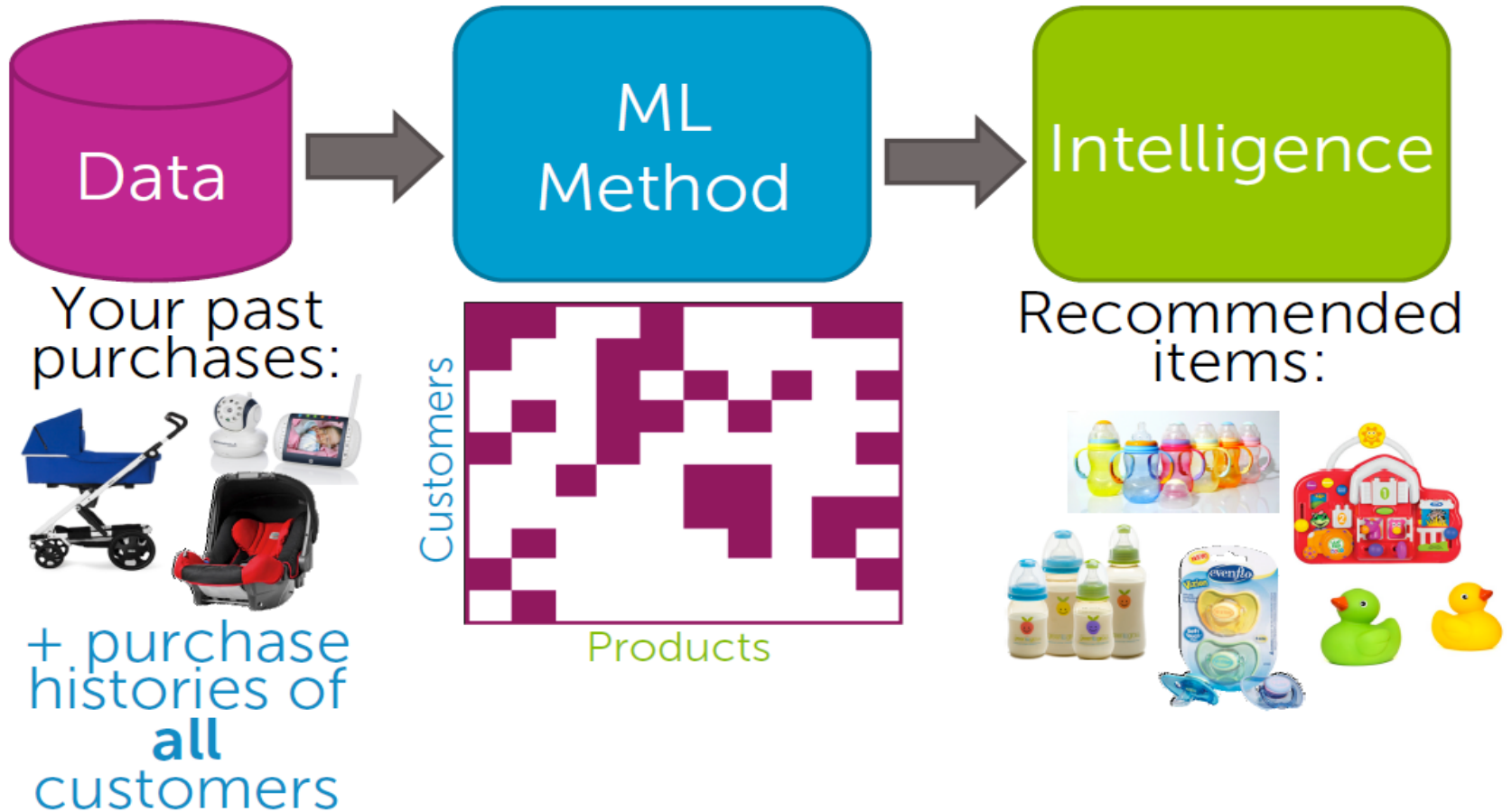
- KD-trees, locality-sensitive hashing (LSH)
- K-means
- Expectation-maximization (EM)

## Concepts

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

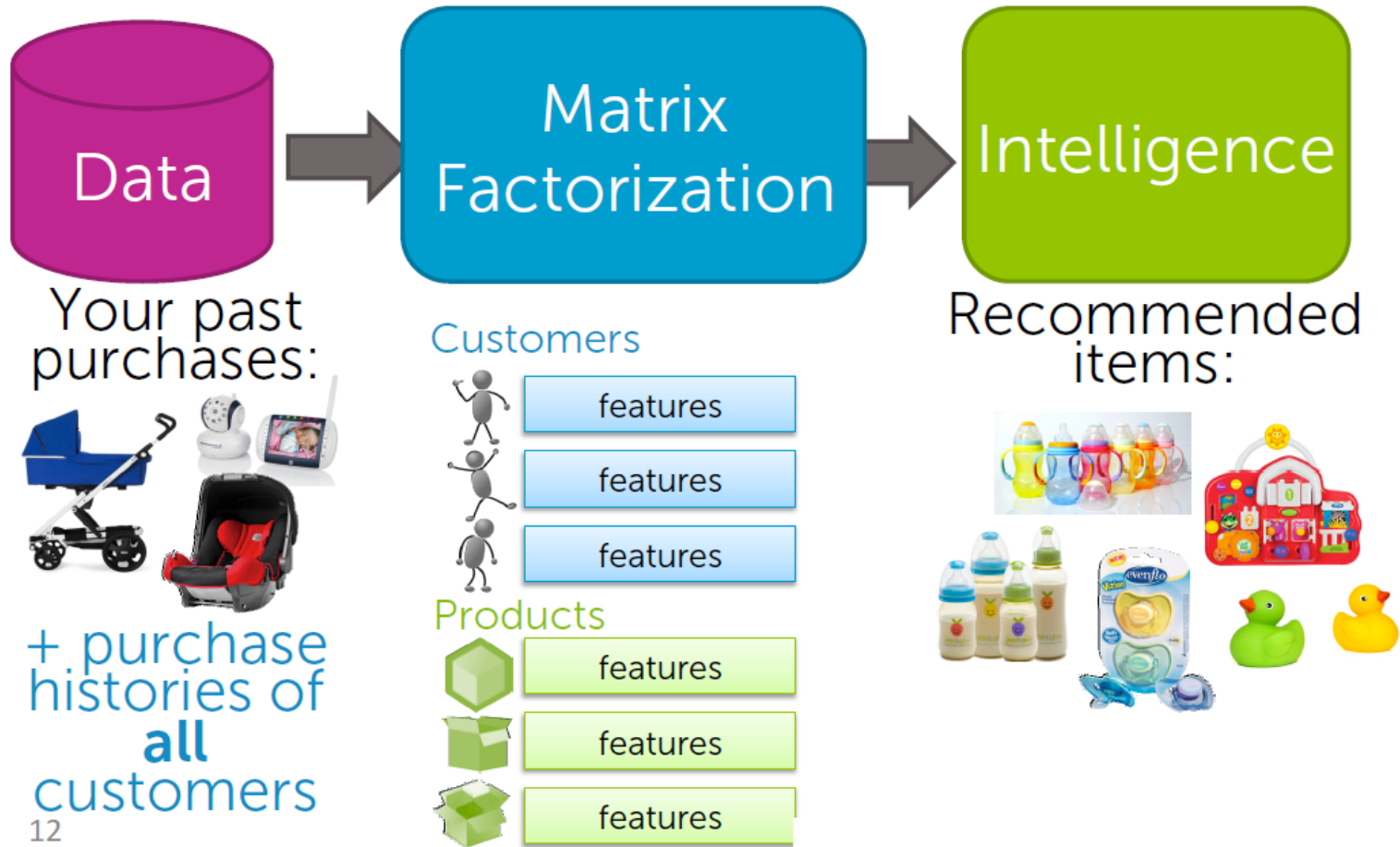
# Case study: Product recommendation

29



# Case study: Product recommendation

30



# Recomender systems applications

31



Movies



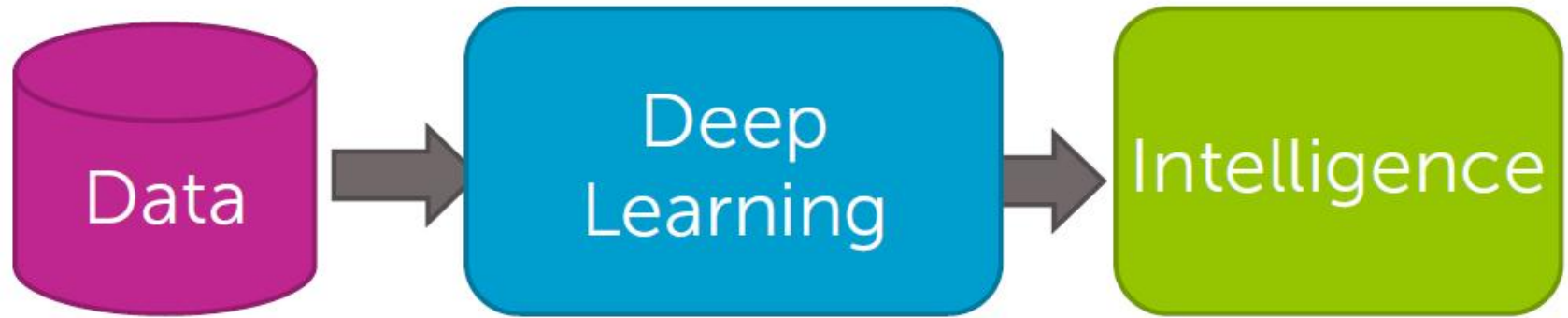
Songs



Friends, apps, ...

# Case study 5: Visual product recommender

32



Input images:



Layer 1



Layer 2



Nearest neighbors:

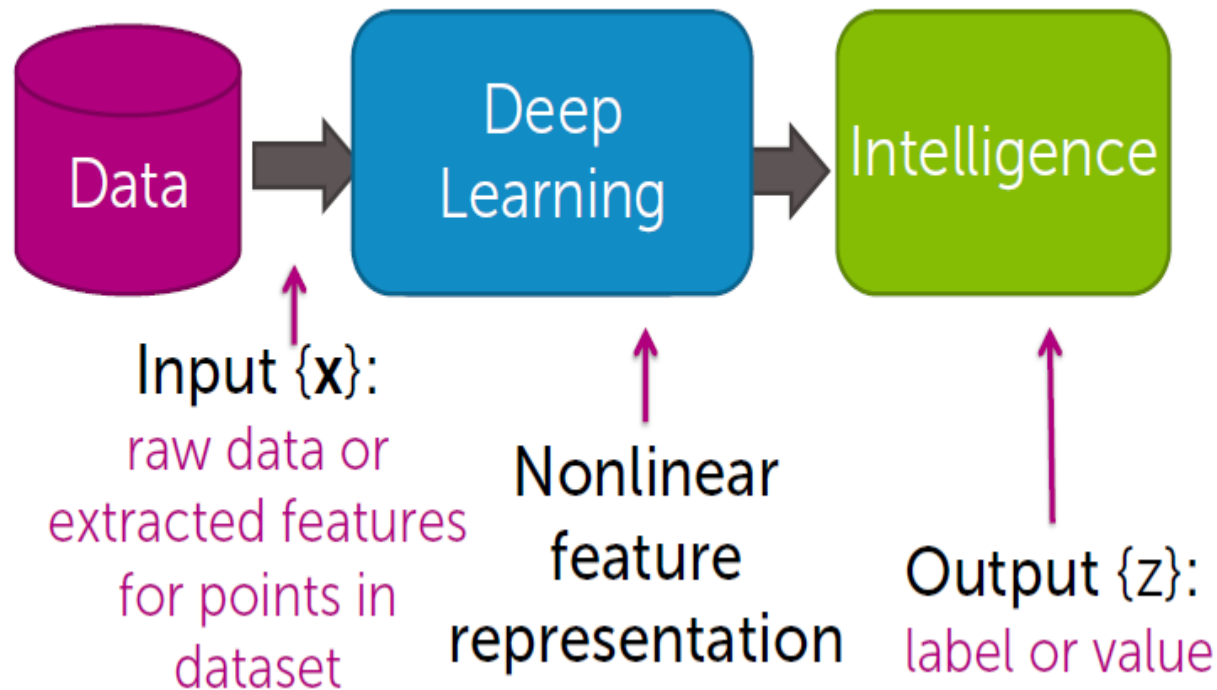




# What is (supervised) deep learning?

33

Flexible method for performing classification or regression



# Examples of deep learning success stories

34

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

# Other ML methods

35

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

# Deploying intelligence module

36

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

