

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

12/10/2022

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.

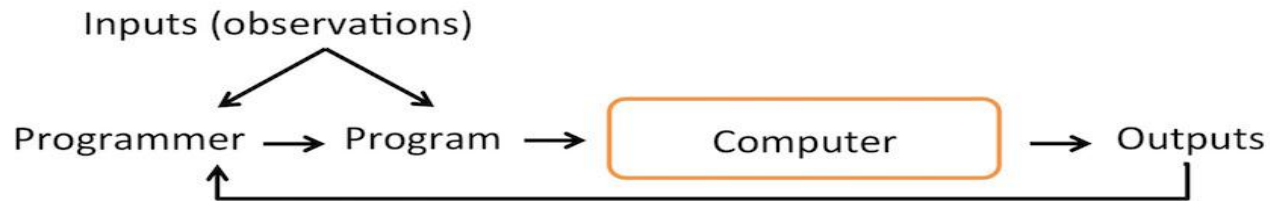
Machine learning - at its core - is about the use and development of these learning algorithms. **Data science** is more about the extraction of knowledge from data to answer particular question or solve particular problems.

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

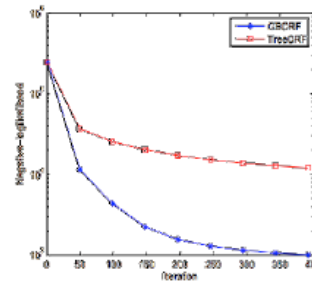
5

- **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**



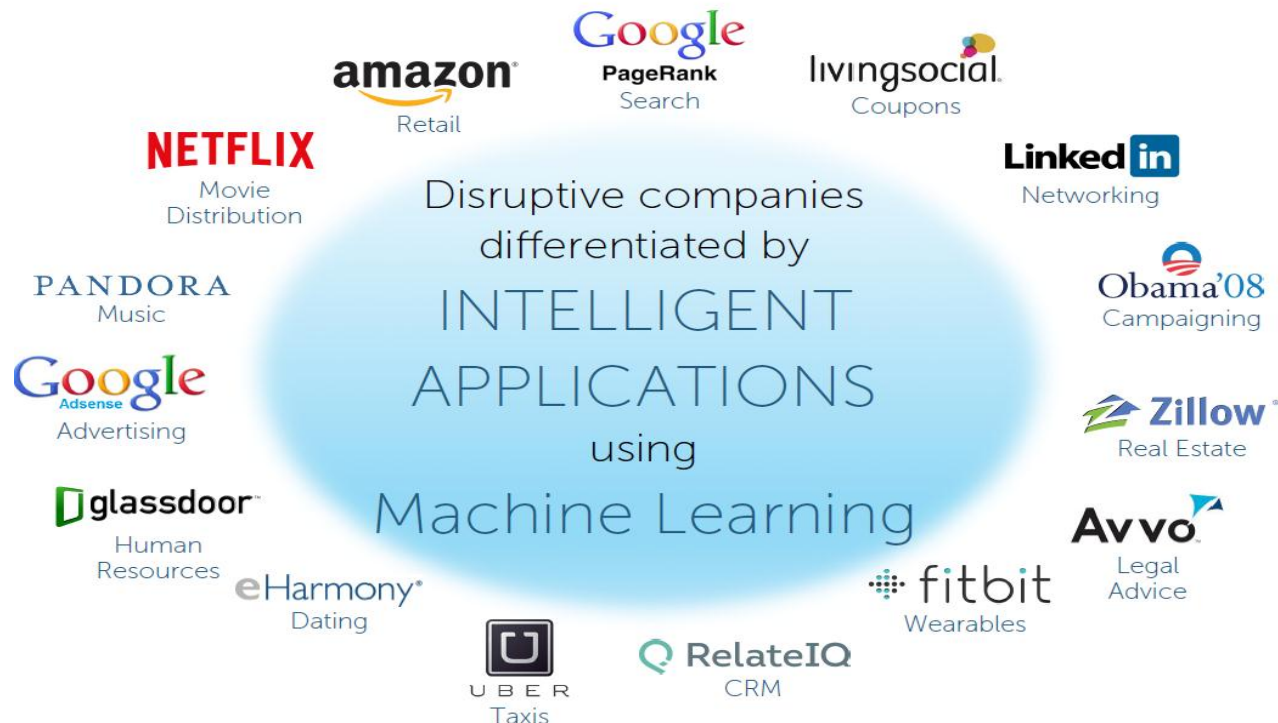
 Neural Information
Processing Systems
Foundation

ICML

Machine learning is changing the world

7

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

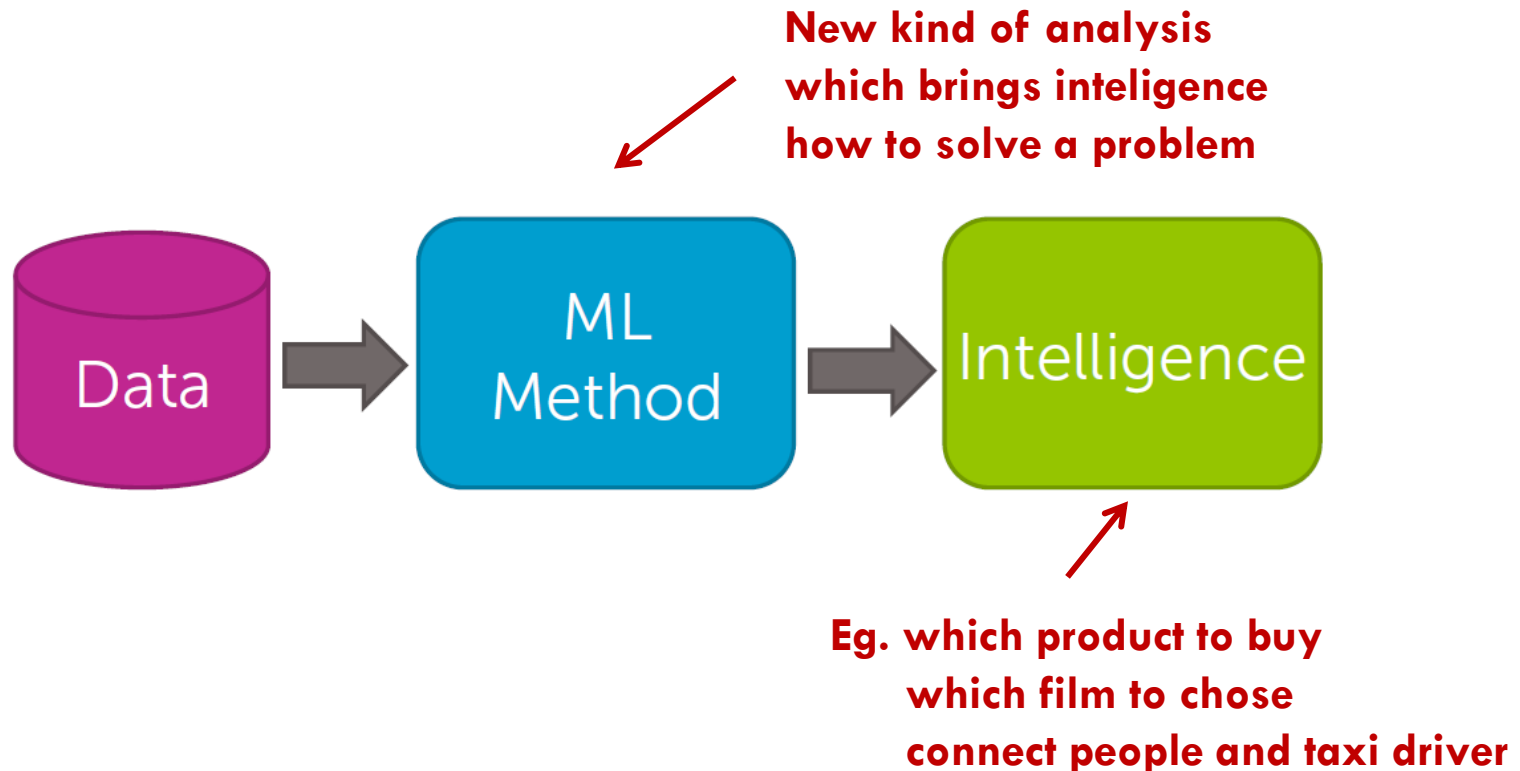


12/10/2022

Machine learning

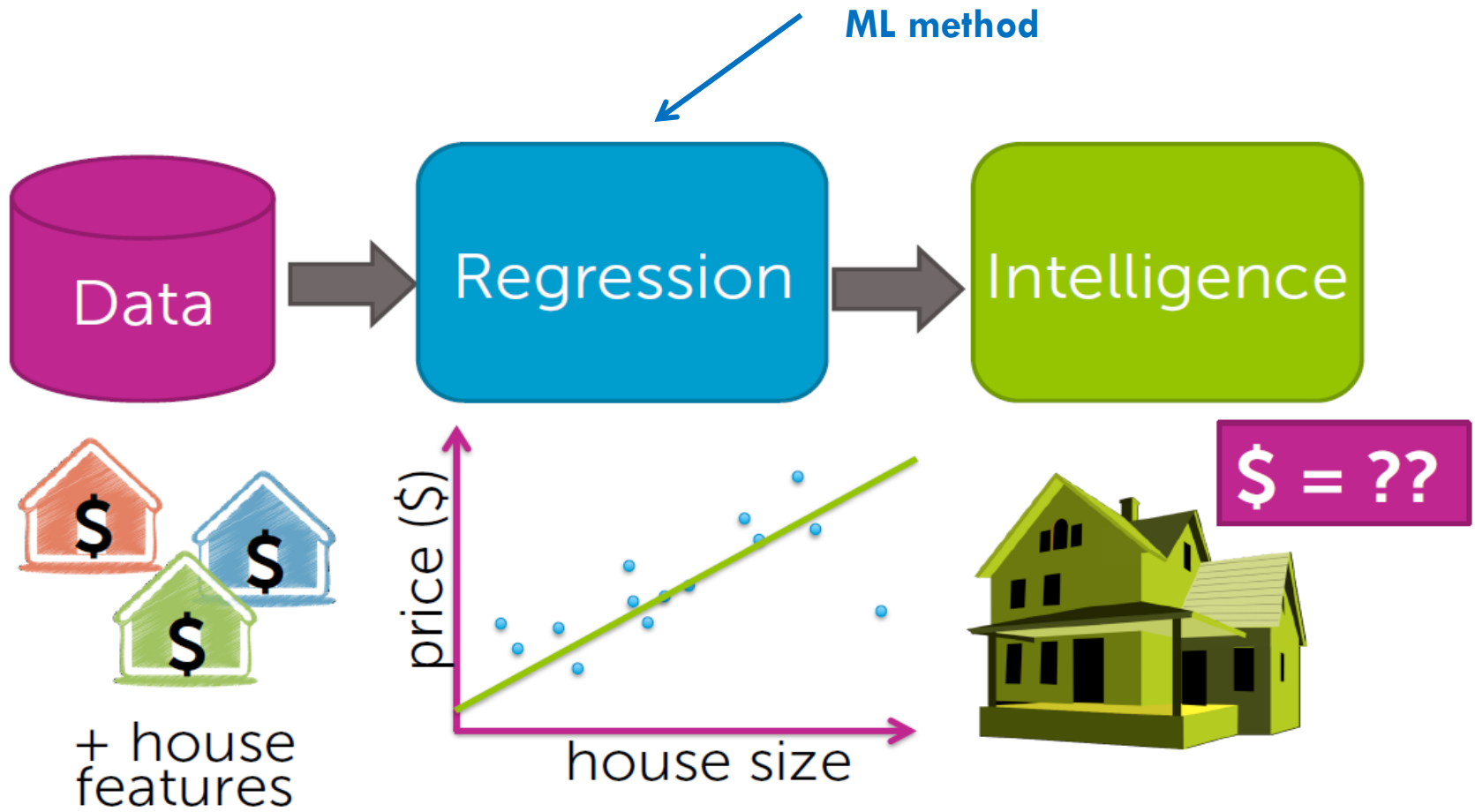
8

□ Data → intelligence pipeline



Case study 1: Prediction

9



Prediction: Predicting house prices

10

Models

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

Algorithms

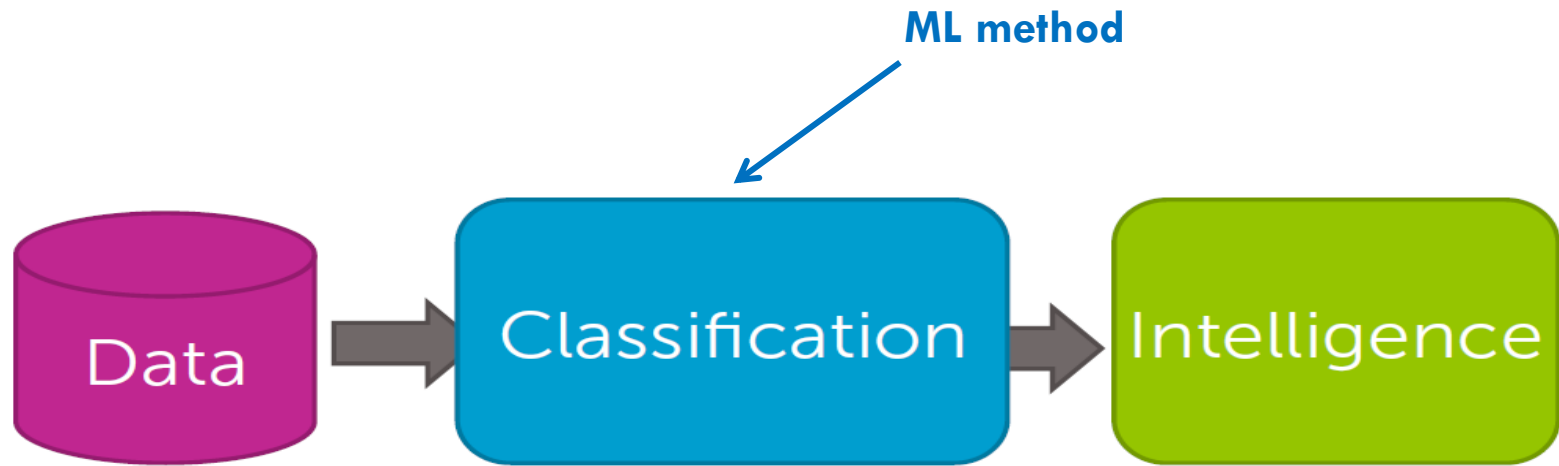
- Gradient descent
- Coordinate descent

Concepts

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

Case study 2: Classification

11



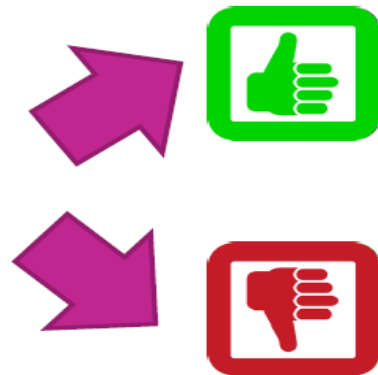
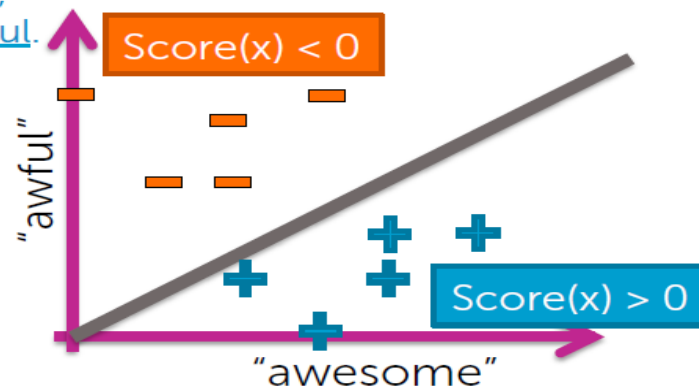
Sushi was awesome,
the food was awesome,
but the service was awful.

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: Sentiment analysis

12

Models

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

Algorithms

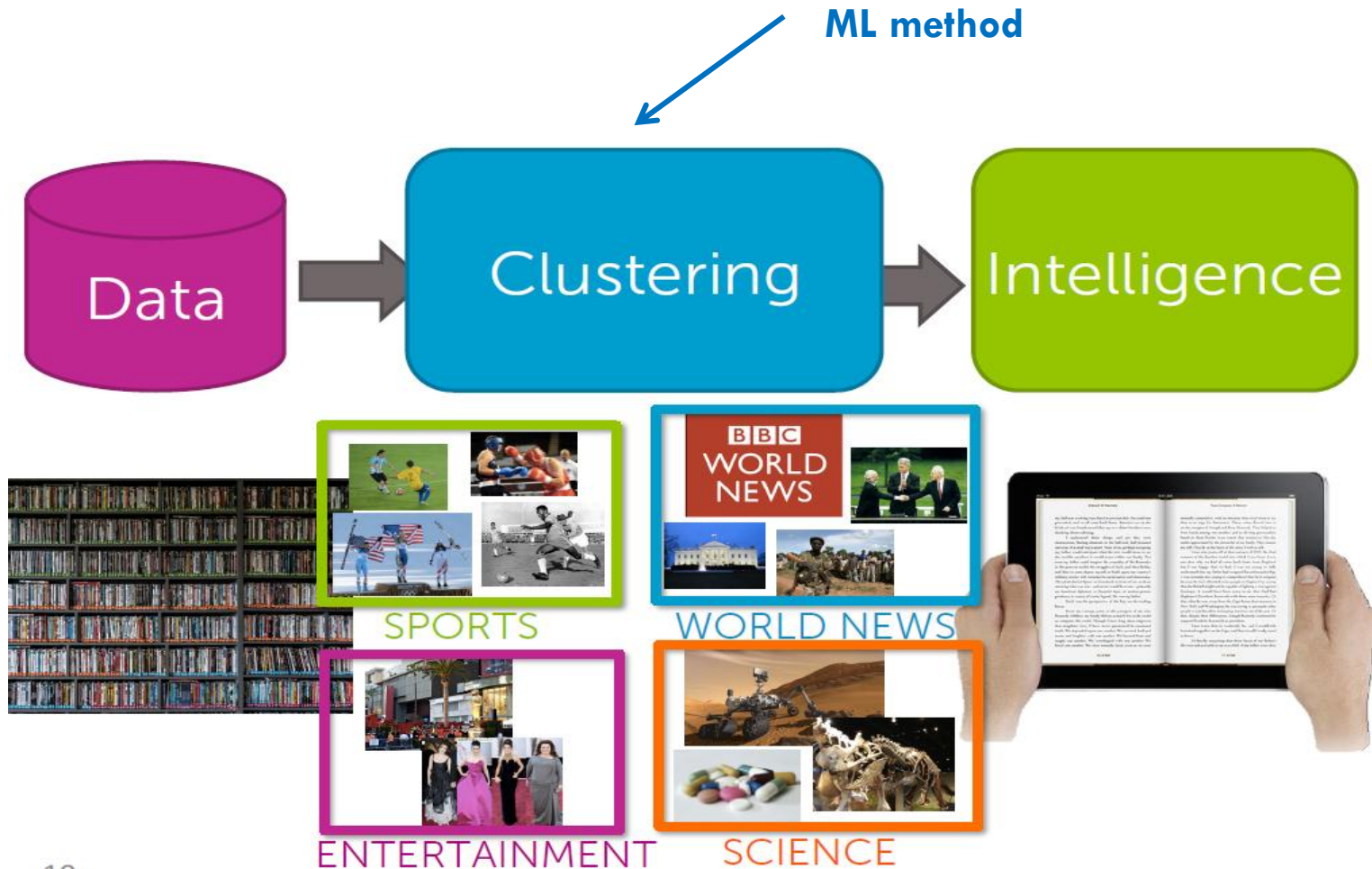
- Stochastic gradient descent
- Boosting

Concepts

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

Case study 3: Clustering

13



10

Clustering: Finding documents

14

Models

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

Algorithms

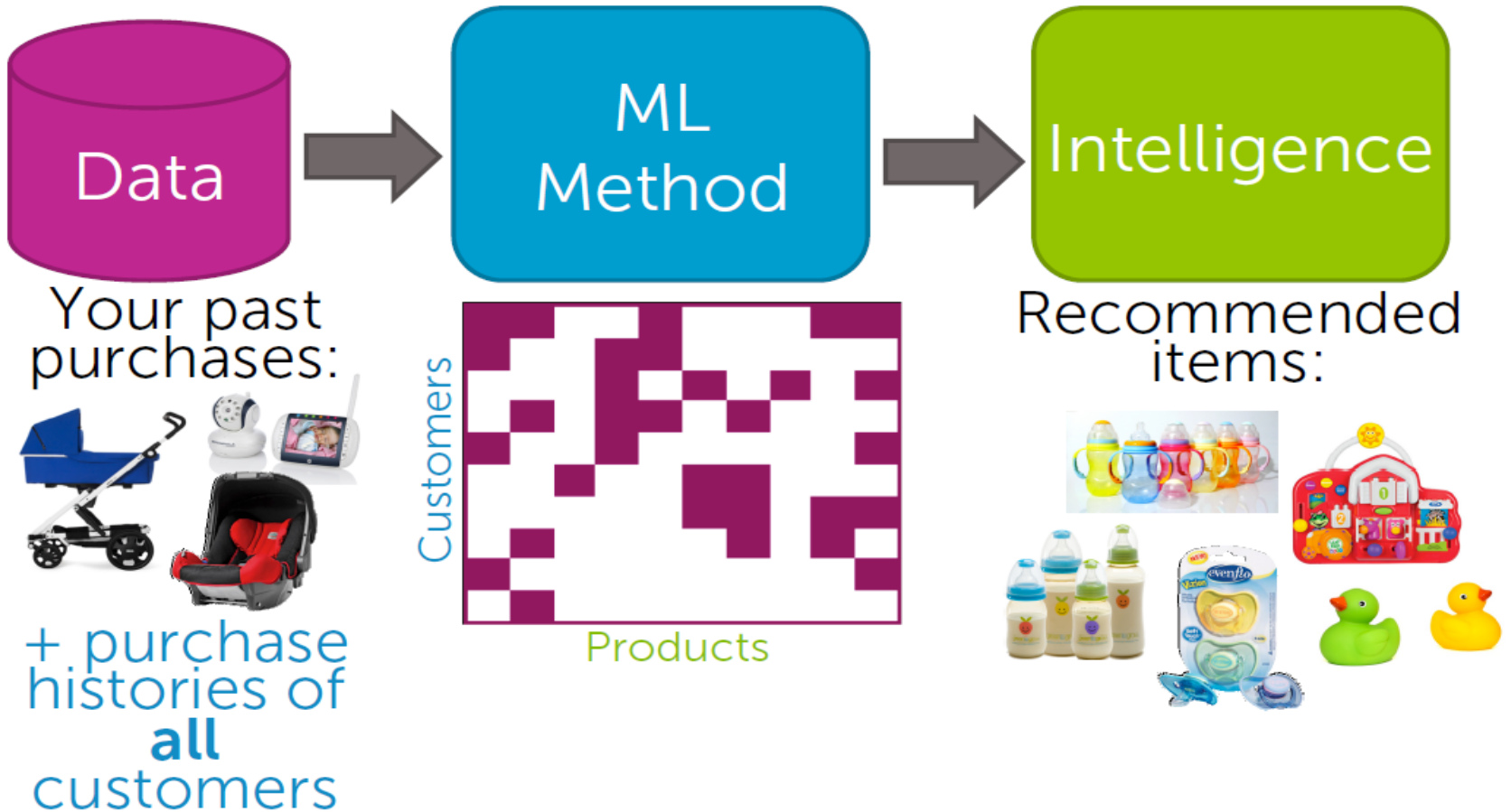
- KD-trees, locality-sensitive hashing (LSH)
- K-means
- Expectation-maximization (EM)

Concepts

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

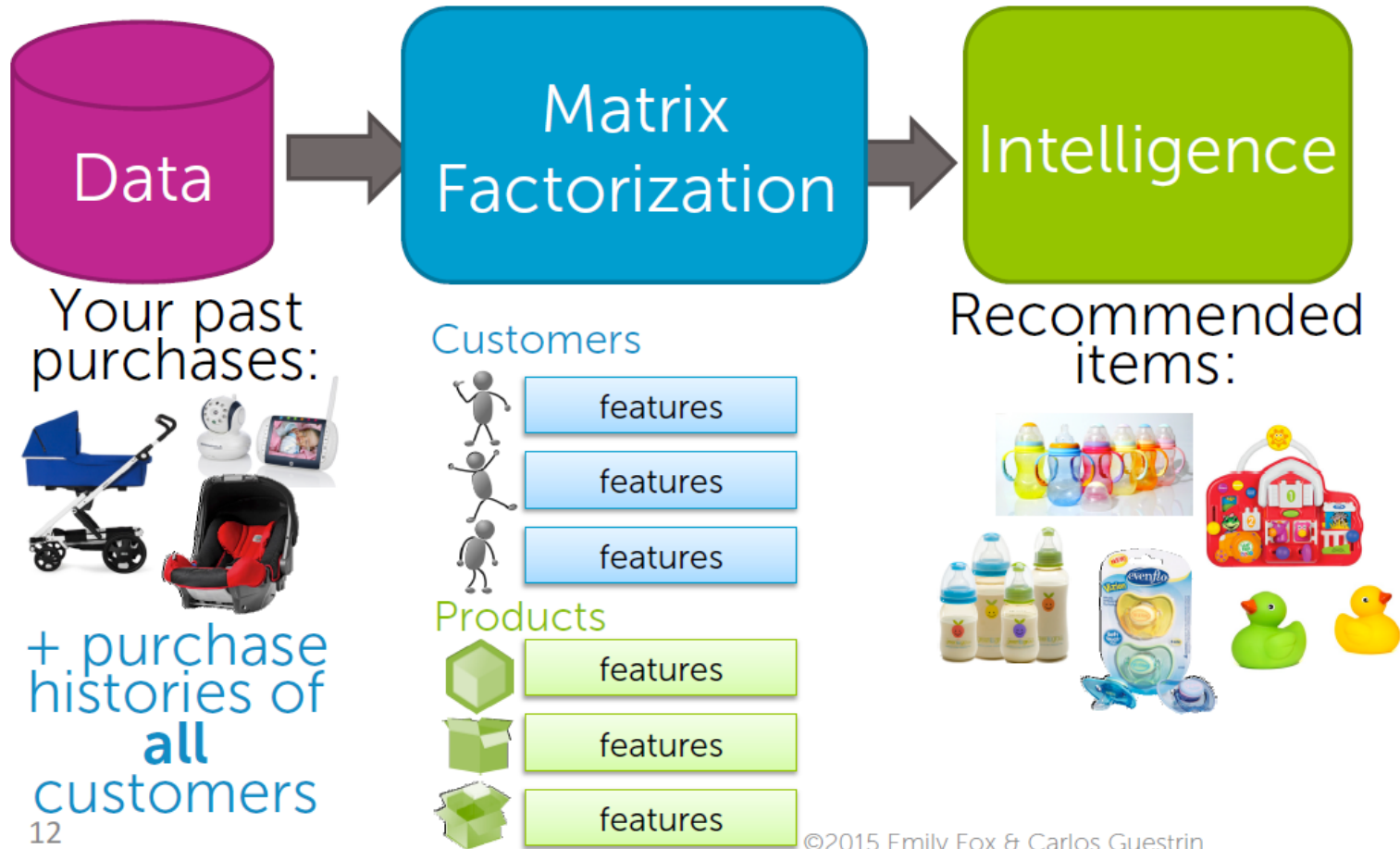
Case study: Product recommendation (not covered in this course)

15



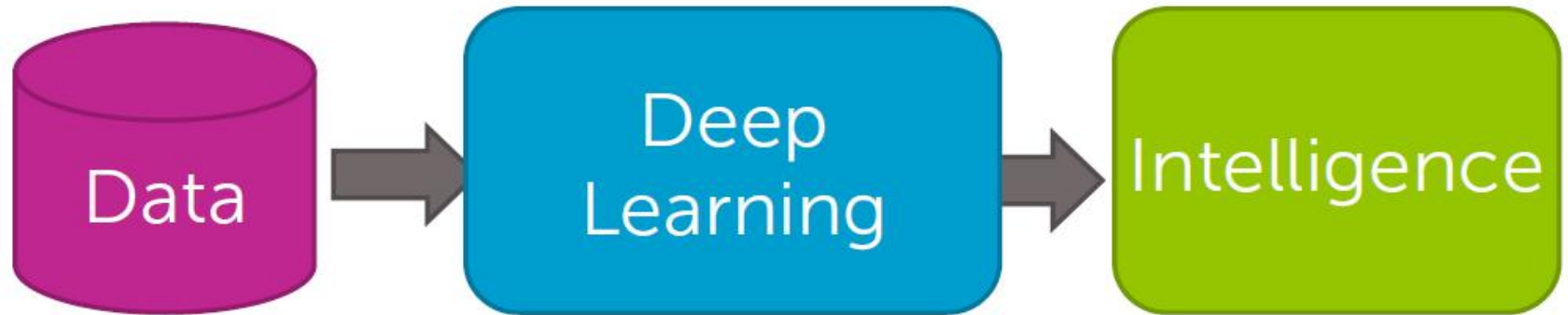
Case study: Product recommendation (not covered here)

16



Case study: Visual product recommender (not covered in this course)

17



Input images:



Layer 1



Layer 2



Nearest neighbors:



Deploying intelligence module

18

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

