

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
- M. Thomson course on Statistics in Physics Analyses, Cambridge

8/10/2019

WFAiS UJ, Fizyka, II stopień studiów

How this course is organised

2

Two block:

Data Scientist oriented:

- ▣ Introduction to **Exploratory Data Analysis**
- ▣ Case studies for **Machine Learning** applications in data analysis
 - Regression,
 - Classification
 - Clustering

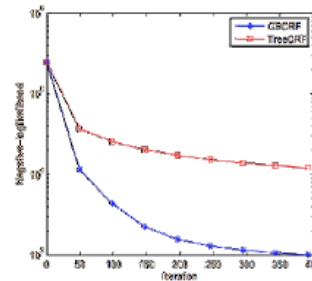
Physics analysis oriented:

- ▣ **Program to be defined**

Analyse data with Machine Learning

3

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



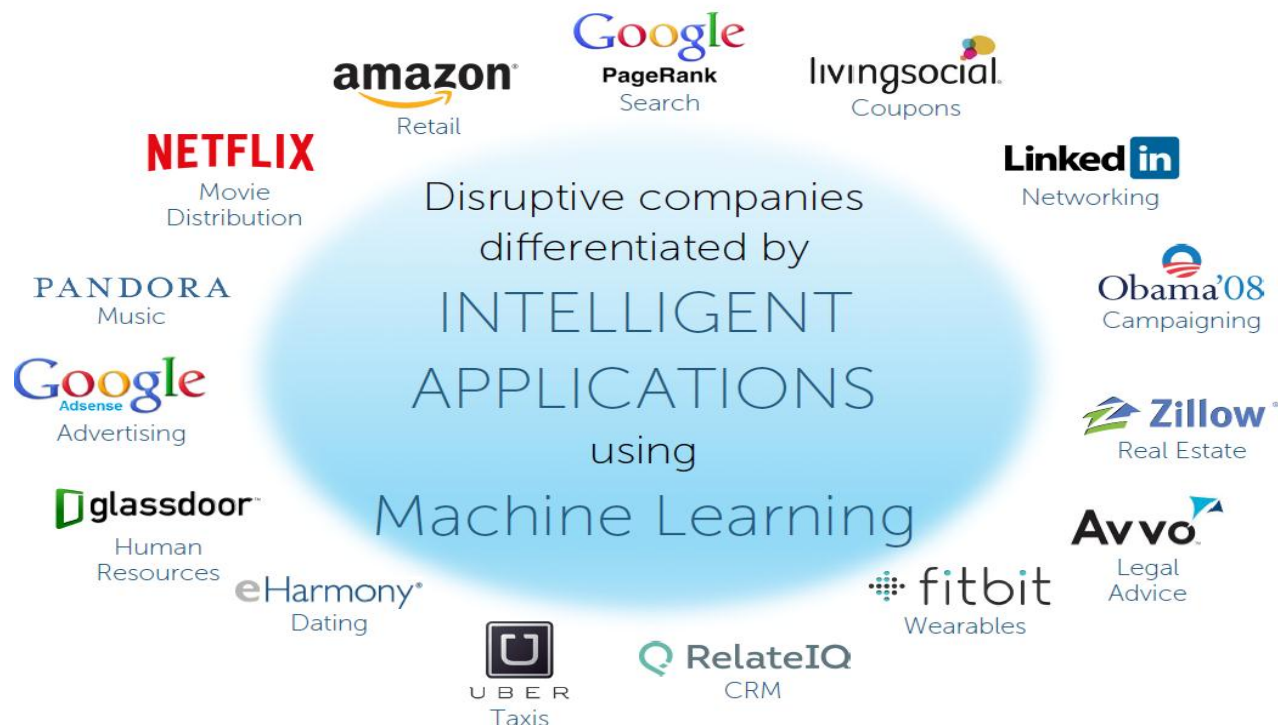
 Neural Information
Processing Systems
Foundation

ICML

Machine learning is changing the world

4

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

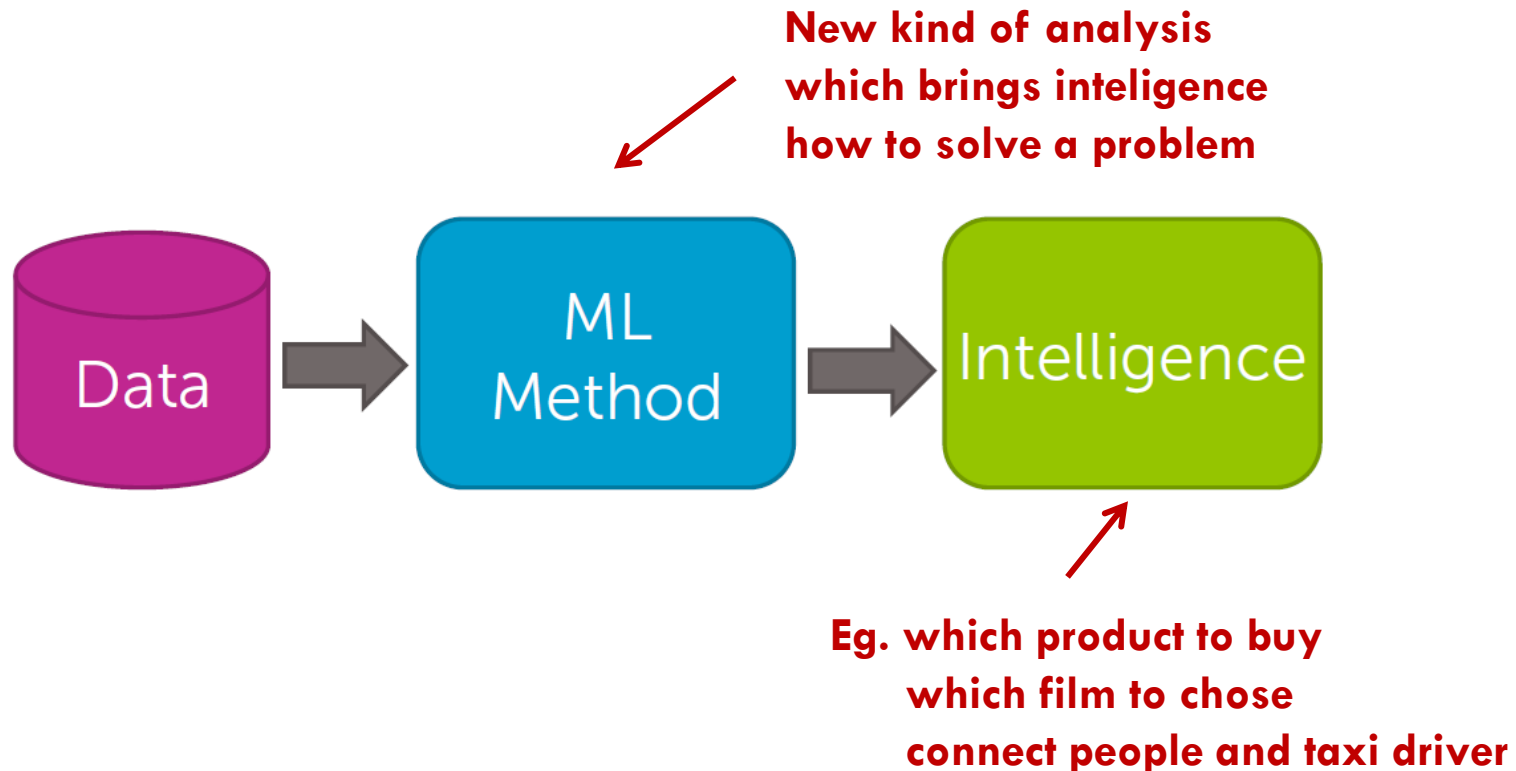


9/10/2019

Machine learning

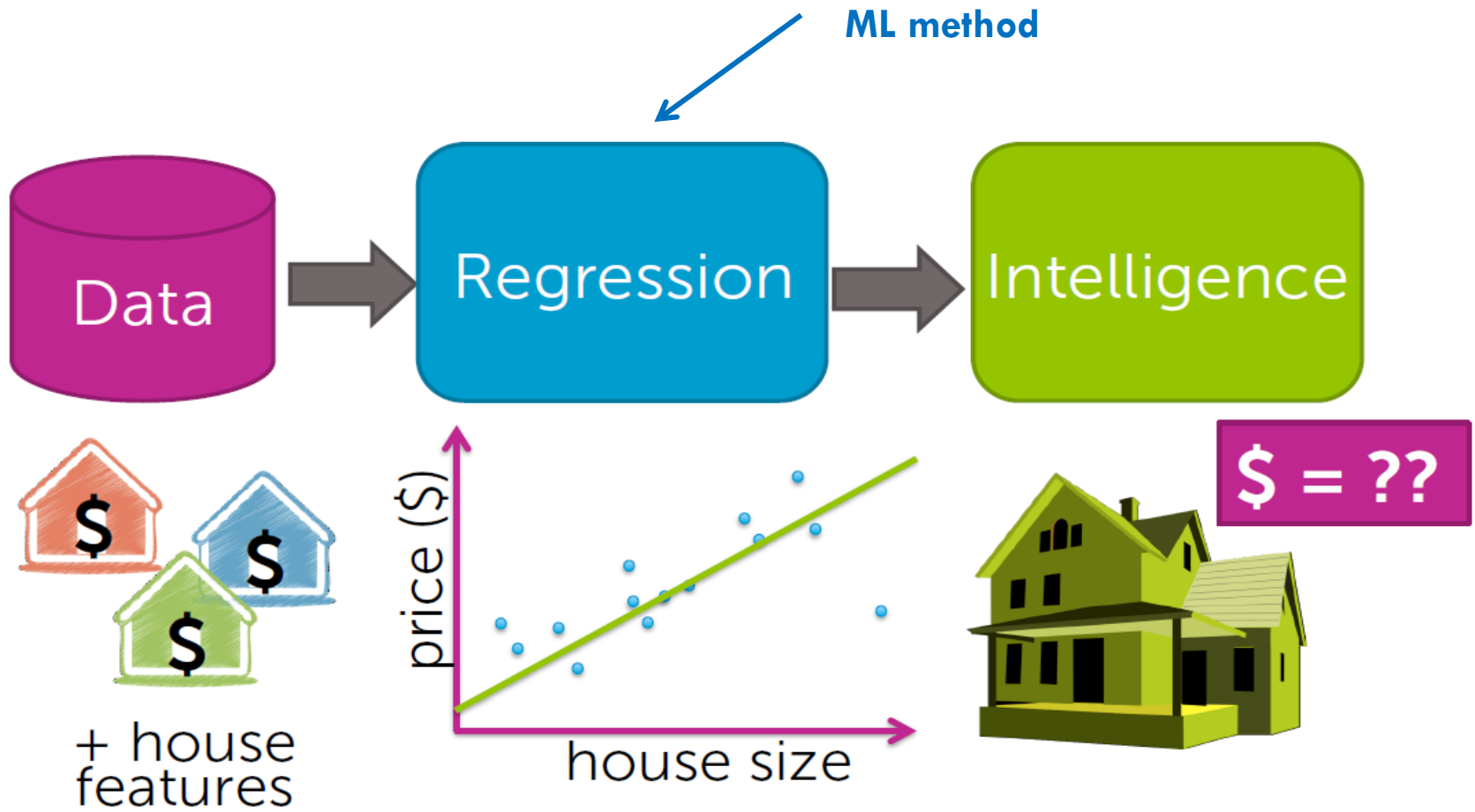
5

□ Data → intelligence pipeline



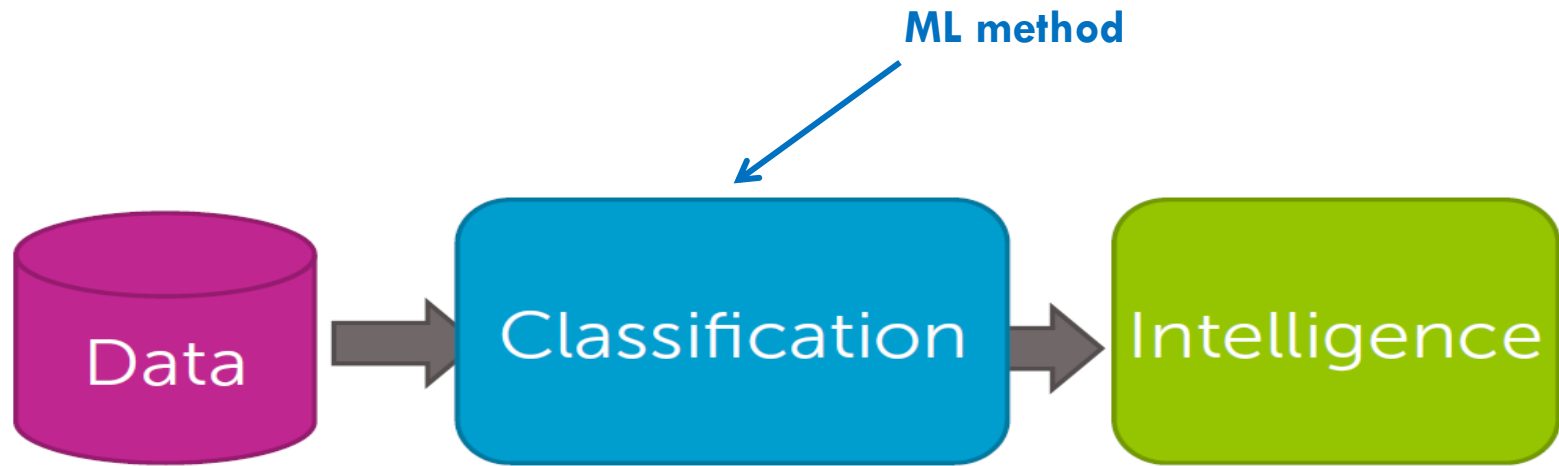
Case study 1: Prediction

6



Case study 2: Classification

7



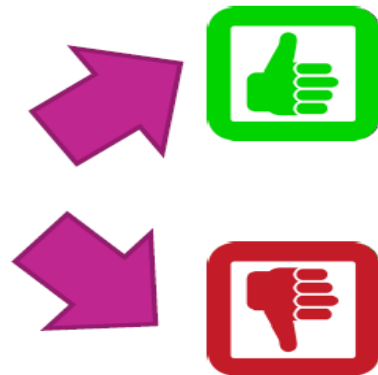
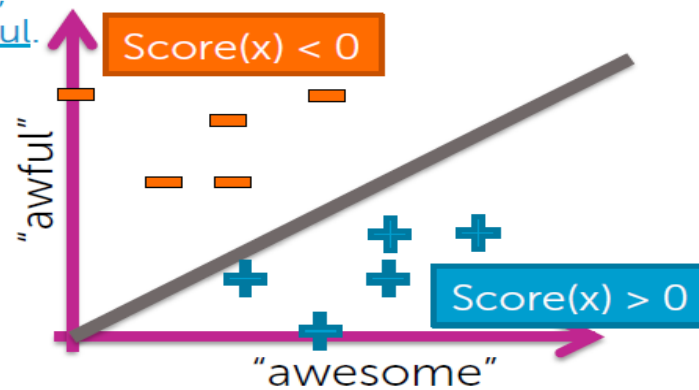
Sushi was awesome,
the food was awesome,
but the service was awful.

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.

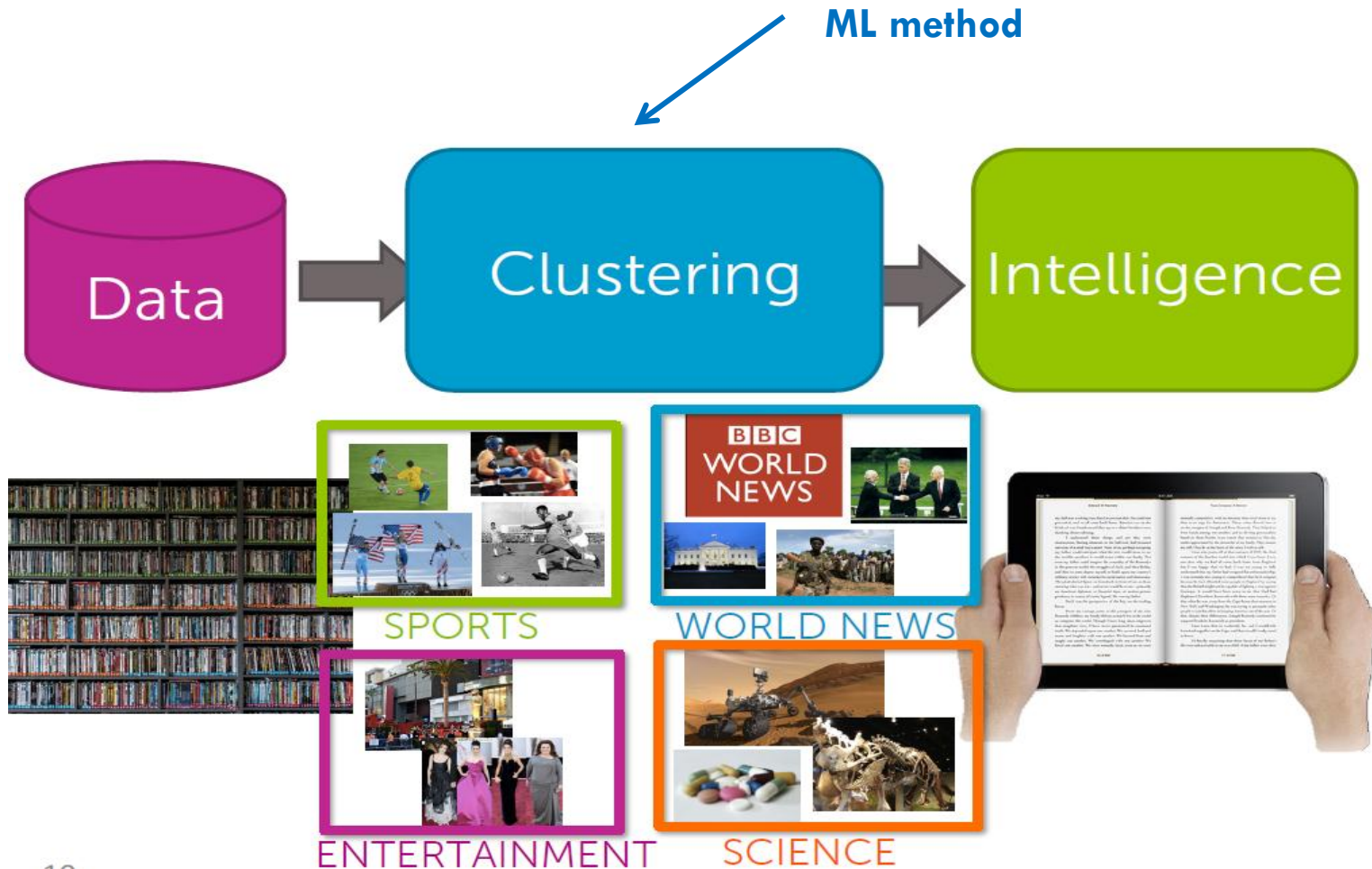
★★★★★ 5/11/2015
Dining here at the sushi bar made me feel like sitting front row to an amazing performance. We didn't have reservations, banged down to the ID after work, got here breathlessly at 5:10pm, and got the last two seats in the place.

★★★★☆ 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.



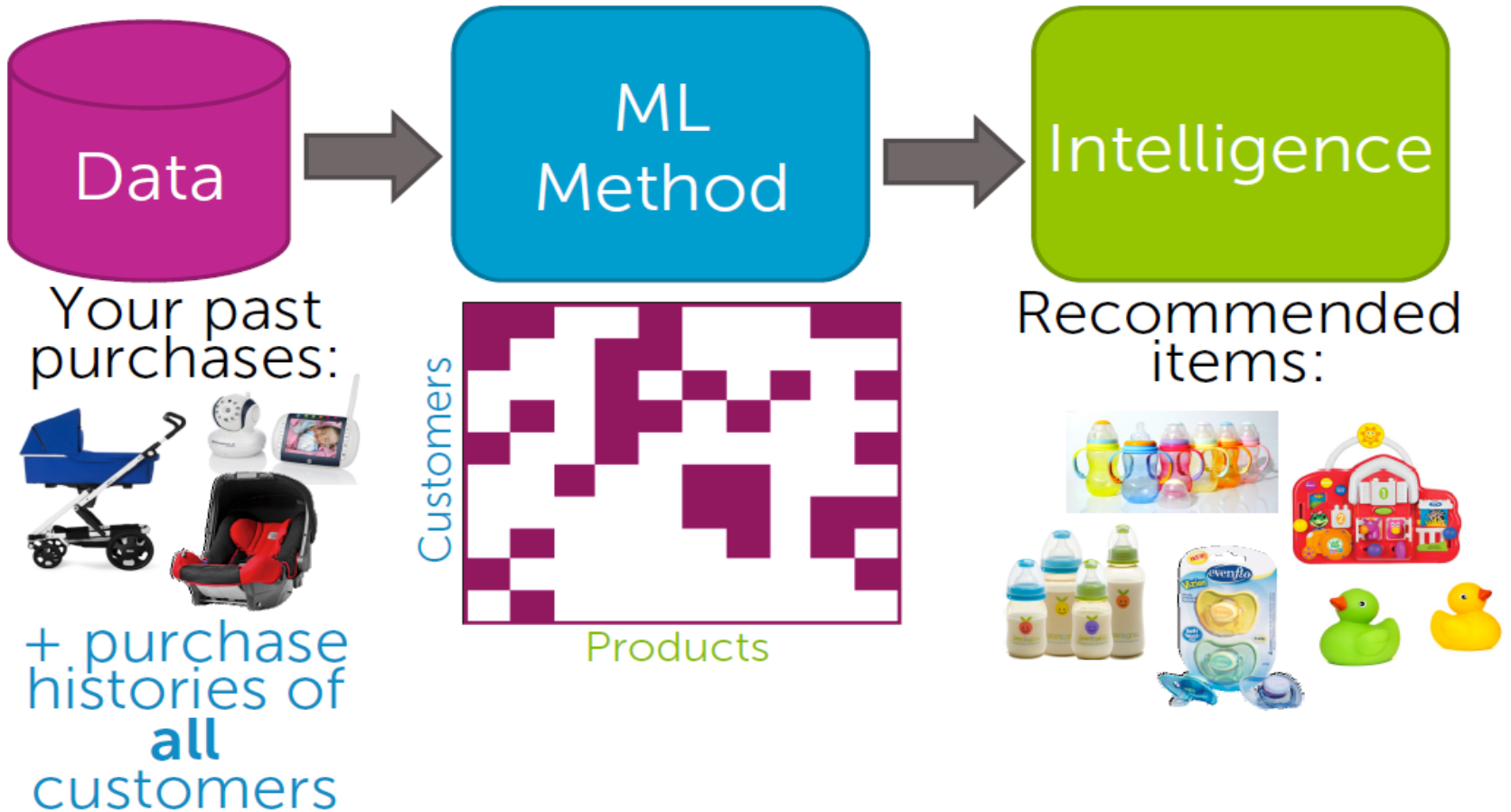
Case study 3: Clustering

8



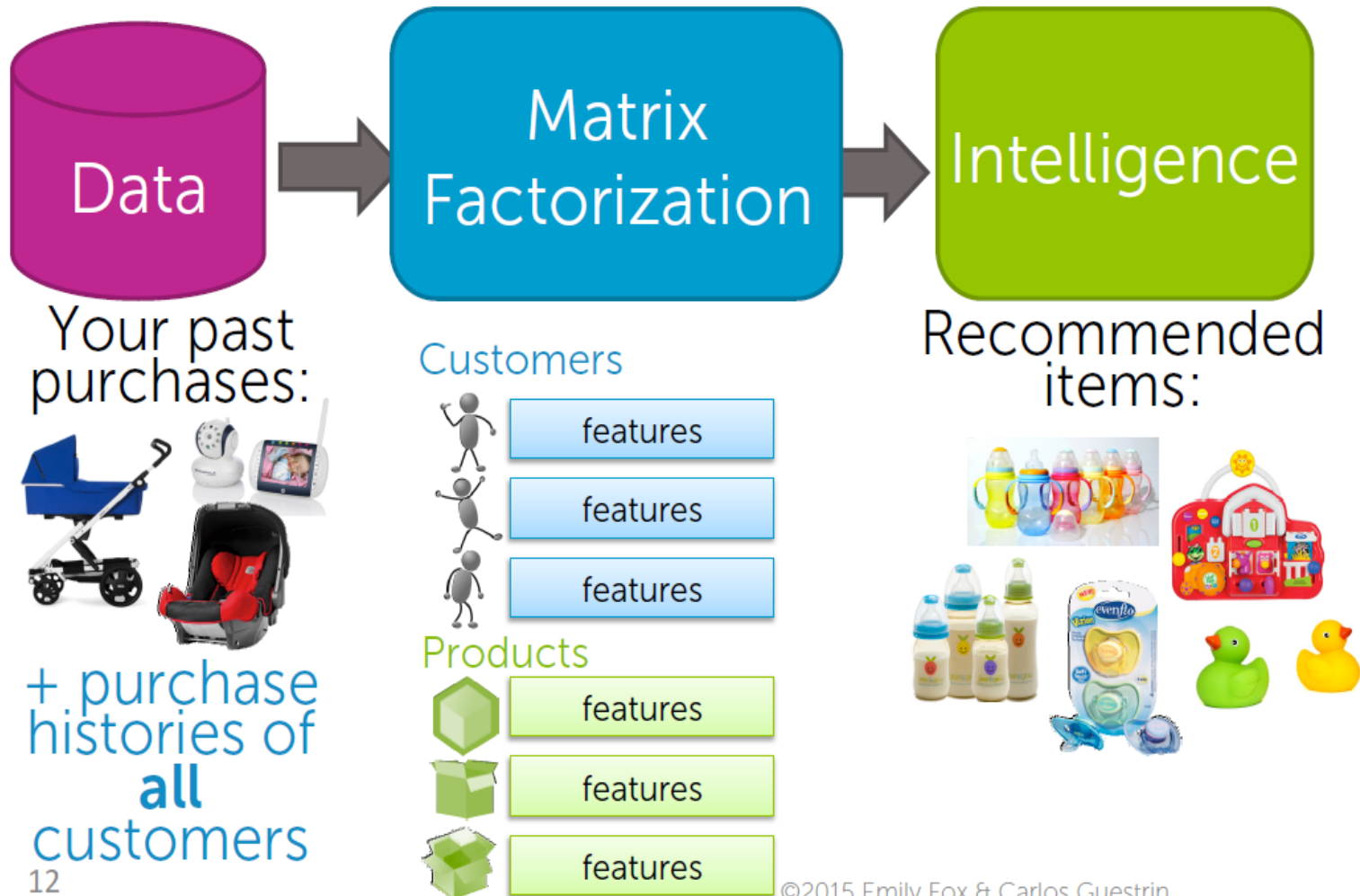
Case study: Product recommendation (not covered here)

9



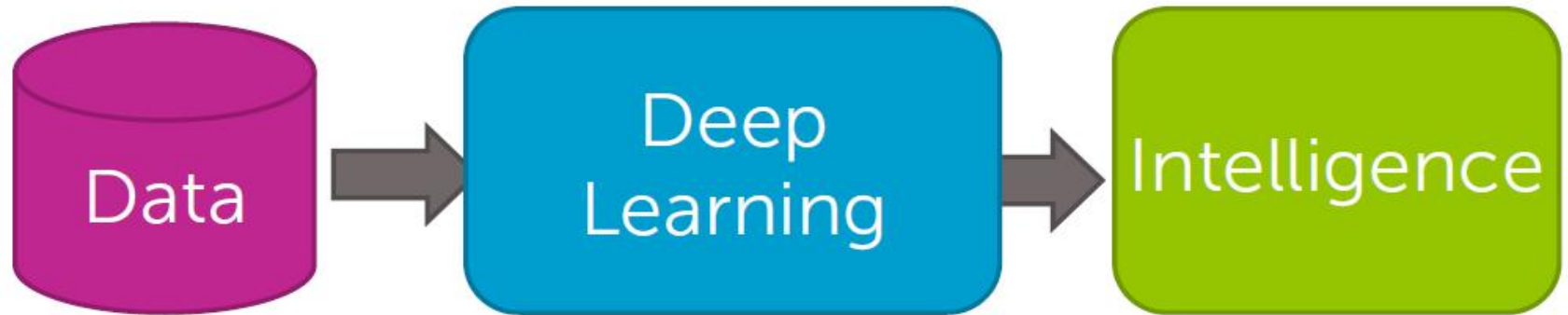
Case study: Product recommendation (not covered here)

10



Case study: Visual product recommender (not covered here)

11



Input images:



Layer 1



Layer 2



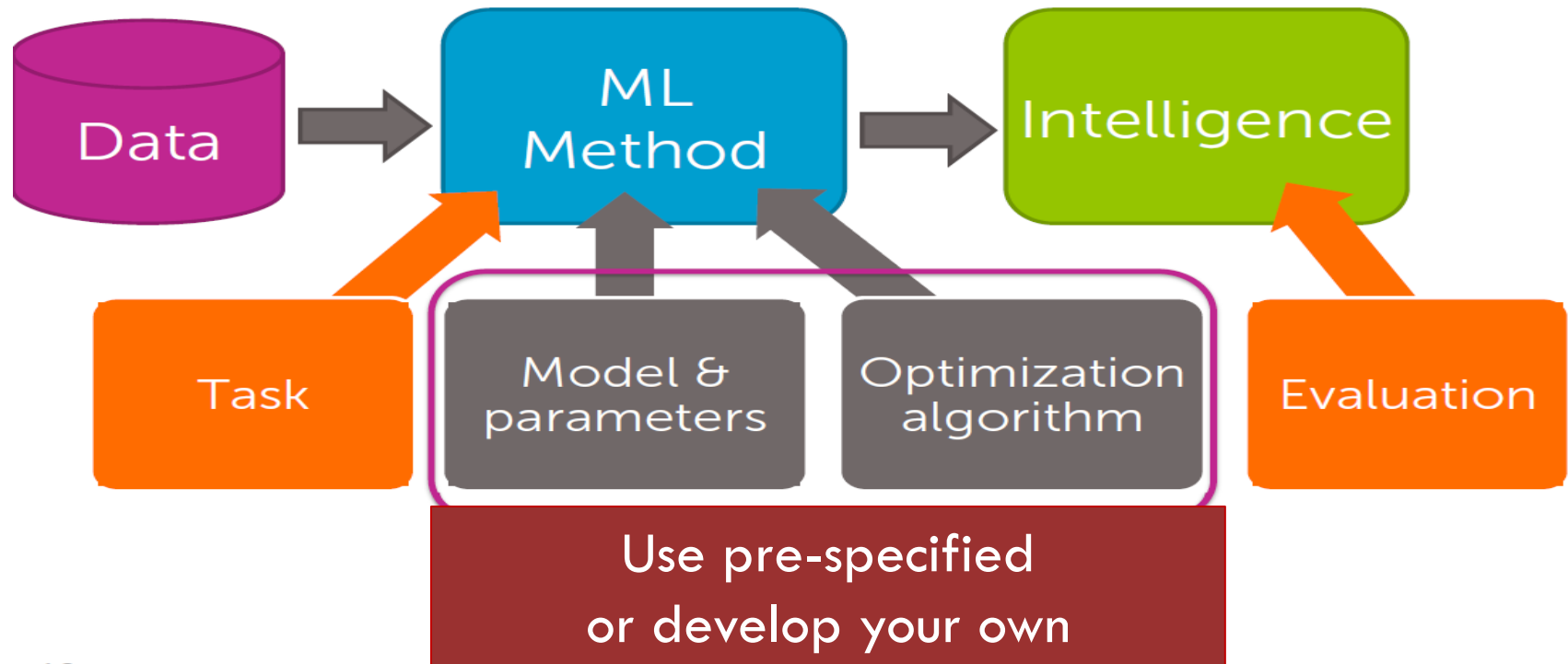
Nearest neighbors:



Deploying intelligence module

12

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



Prediction: Predicting house prices

13

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

Classification: Sentiment analysis

14

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

Clustering: Finding documents

15

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

Lectures for each case study

16

- **We will start with „Primer” level**
 - ▣ **LAB: 5 simple assignments realised individual projects**
- **Then continue with „Advanced” level**
 - ▣ **LAB: 1 advanced project, realised as individual one or in the group.**