# 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

8/10/2019

WFAiS UJ, Informatyka Stosowana

I stopień studiów

# How this course is organised



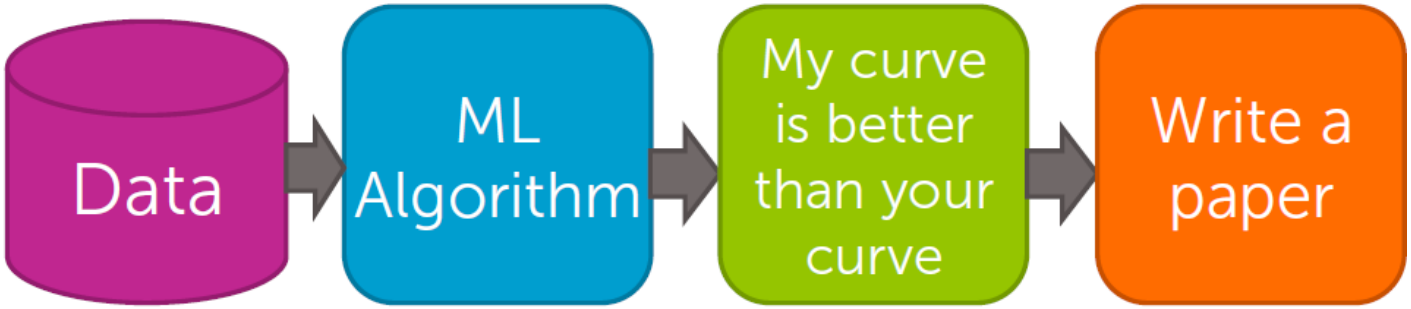
**Three block:**

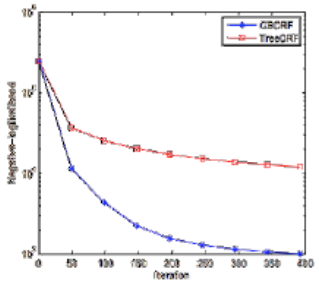
- Introduction to Exploratory Data Analysis
- Case studies for Machine Learning applications in data analysis
  - Regression,
  - Classification
  - Clustering
- Data Analysis and Statistical Inference

# Analyse data with Machine Learning



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



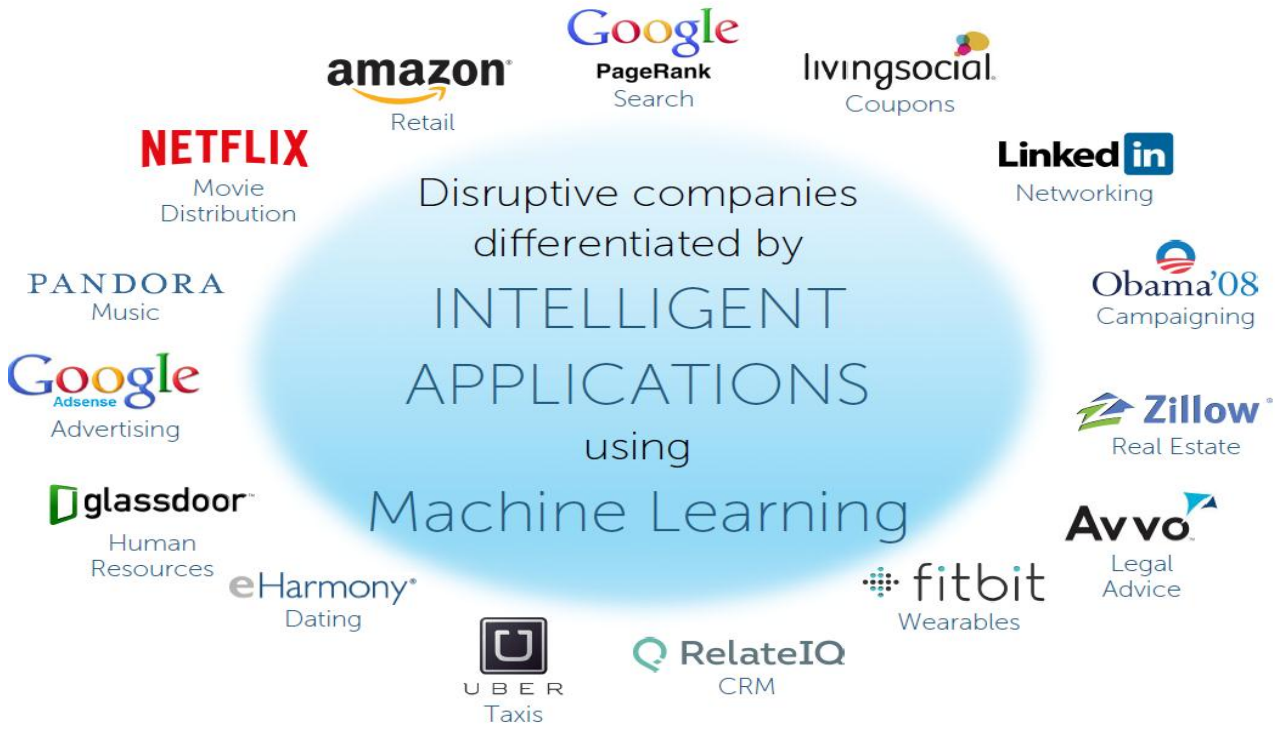
Neural Information Processing Systems Foundation

ICML

# Machine learning is changing the world



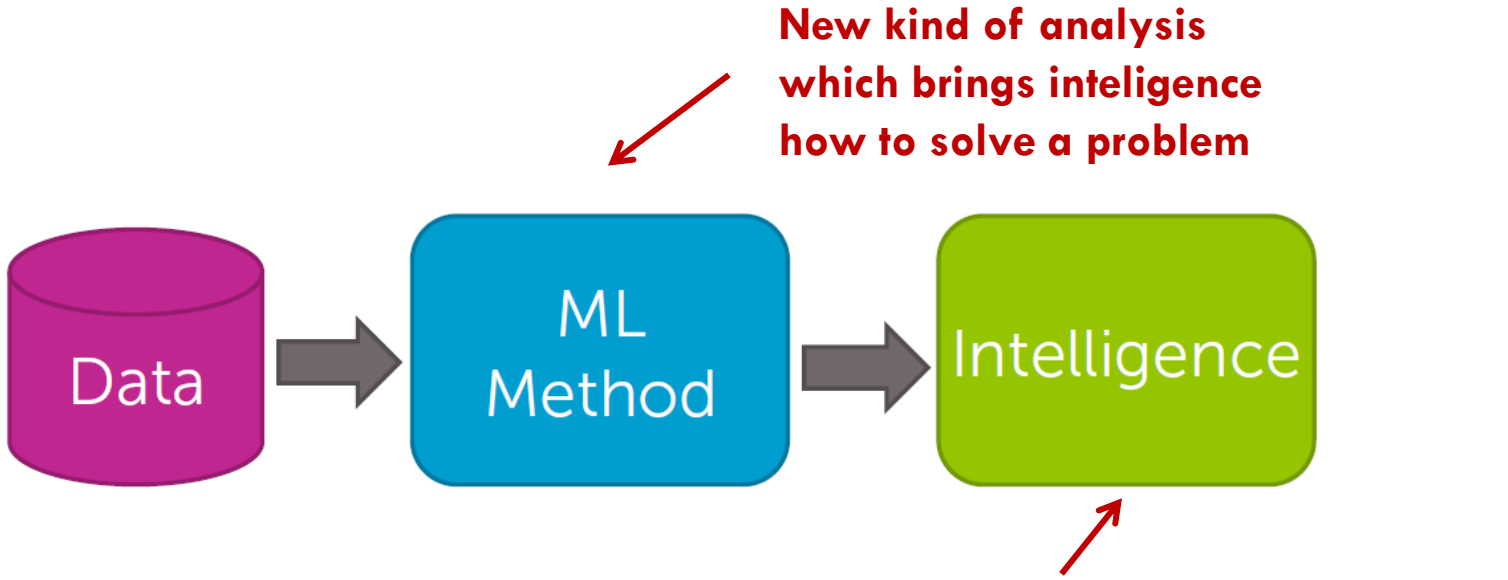
- **Current view: disruptive inteligent applications are used by leading comercial companies**



# Machine learning

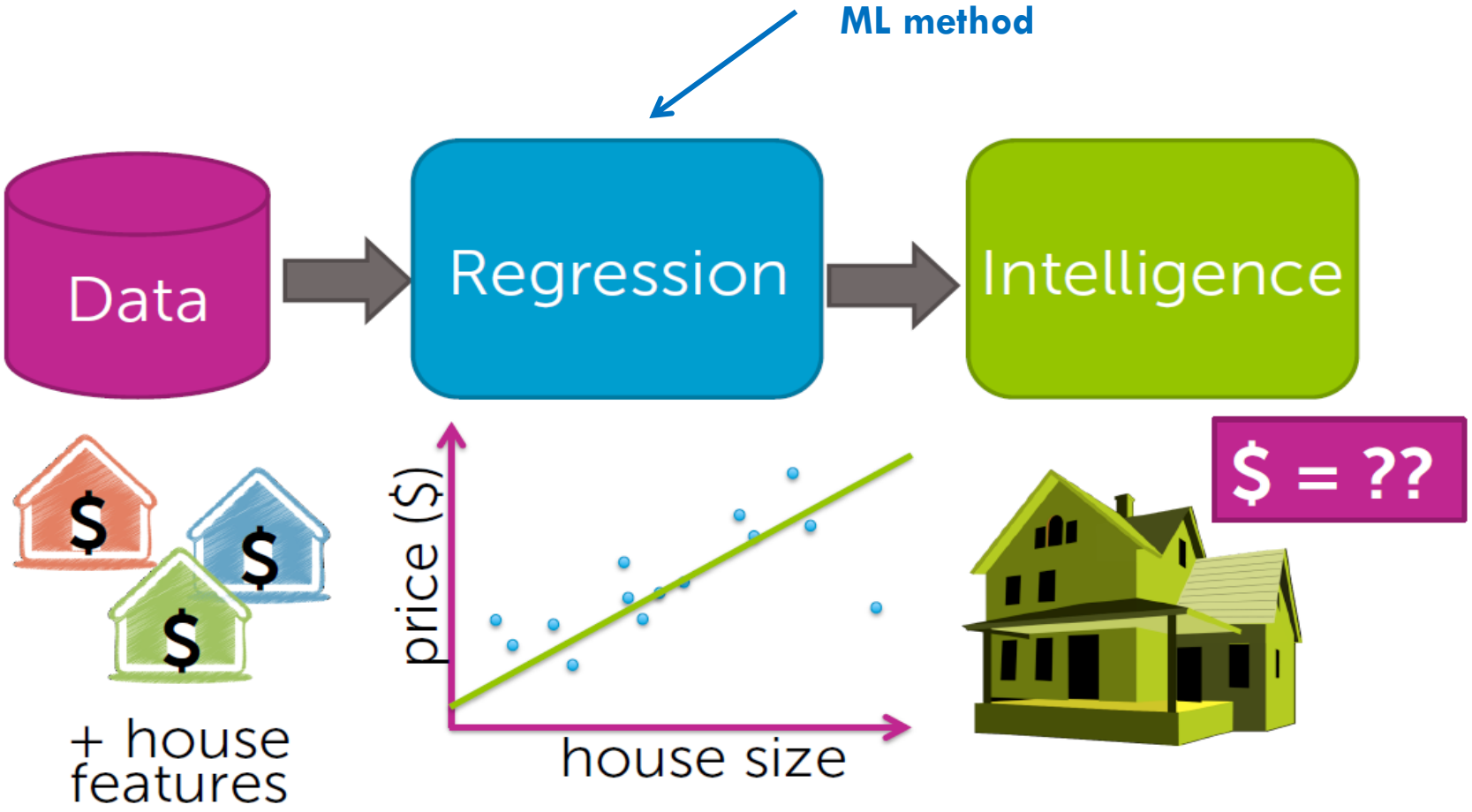


- **Data → inteligence pipeline**



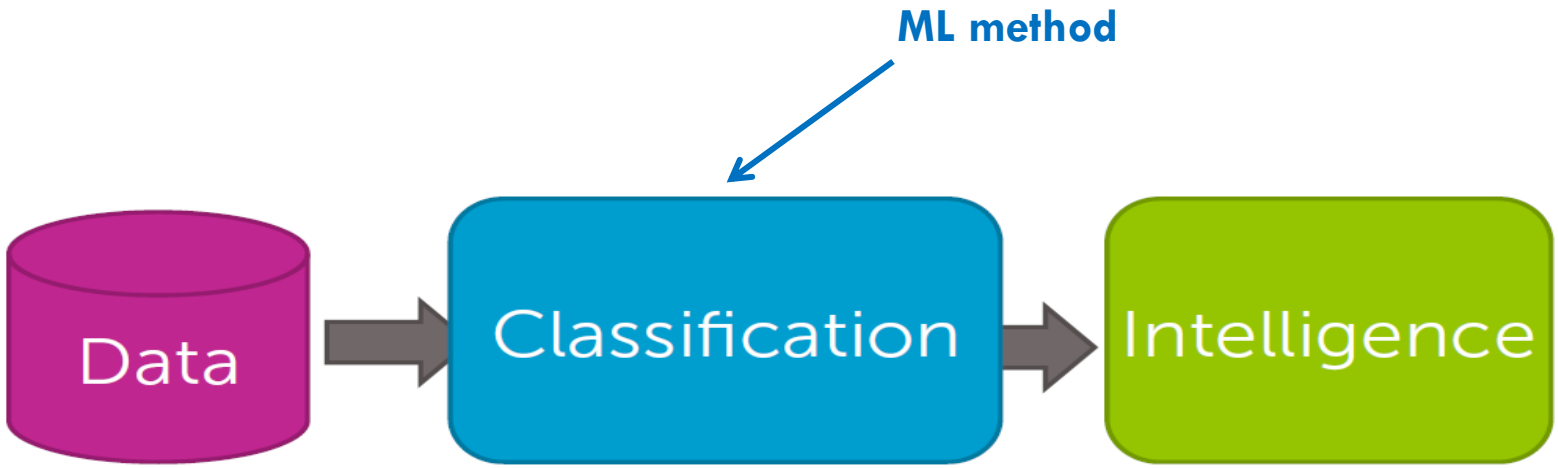
# Case study 1: Prediction





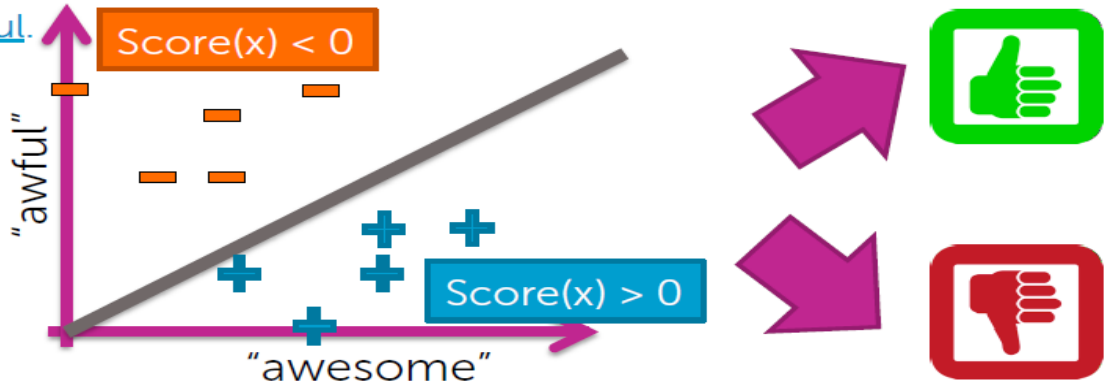
# Case study 2: Classification





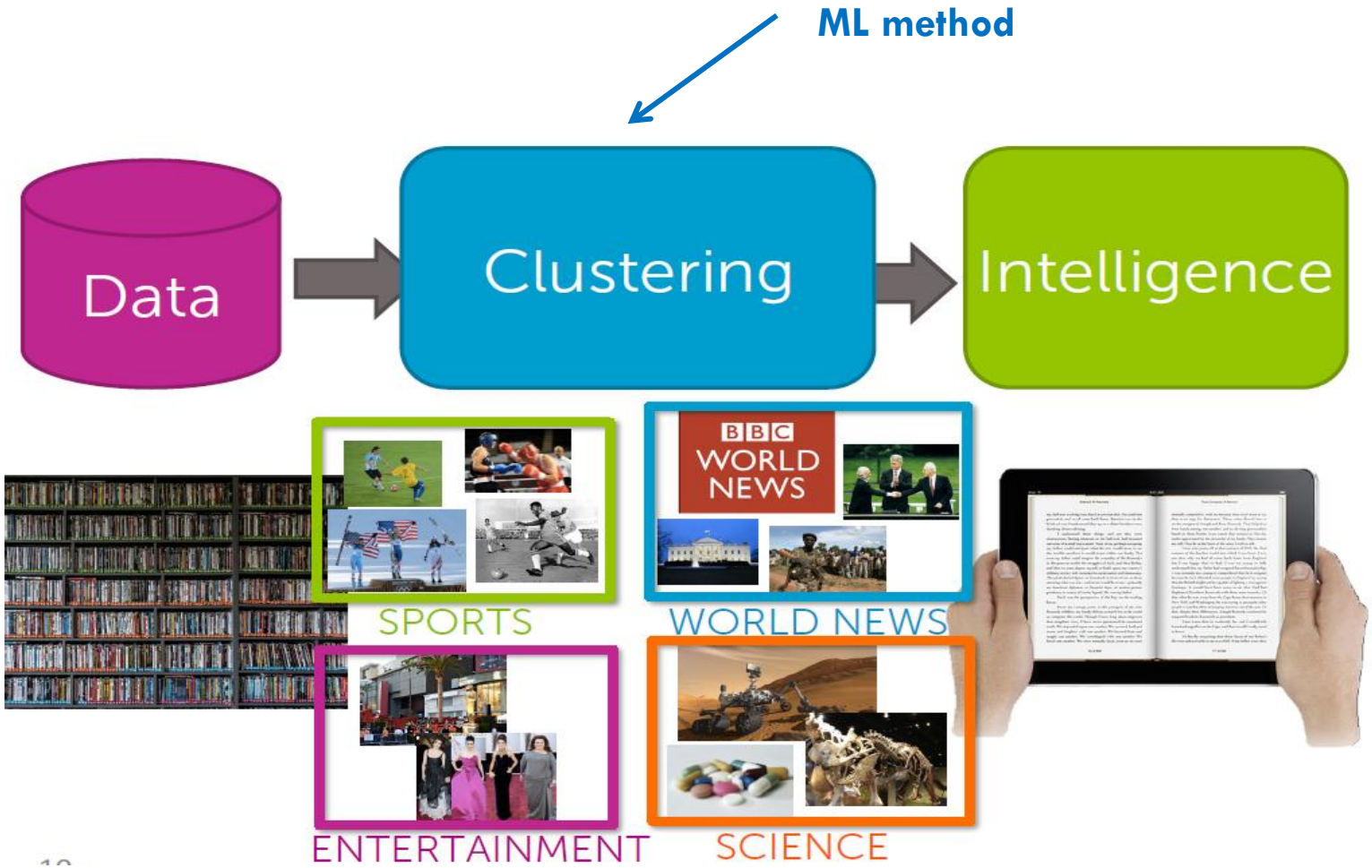

Sushi was awesome,
the food was awesome,
but the service was awful.

All reviews:



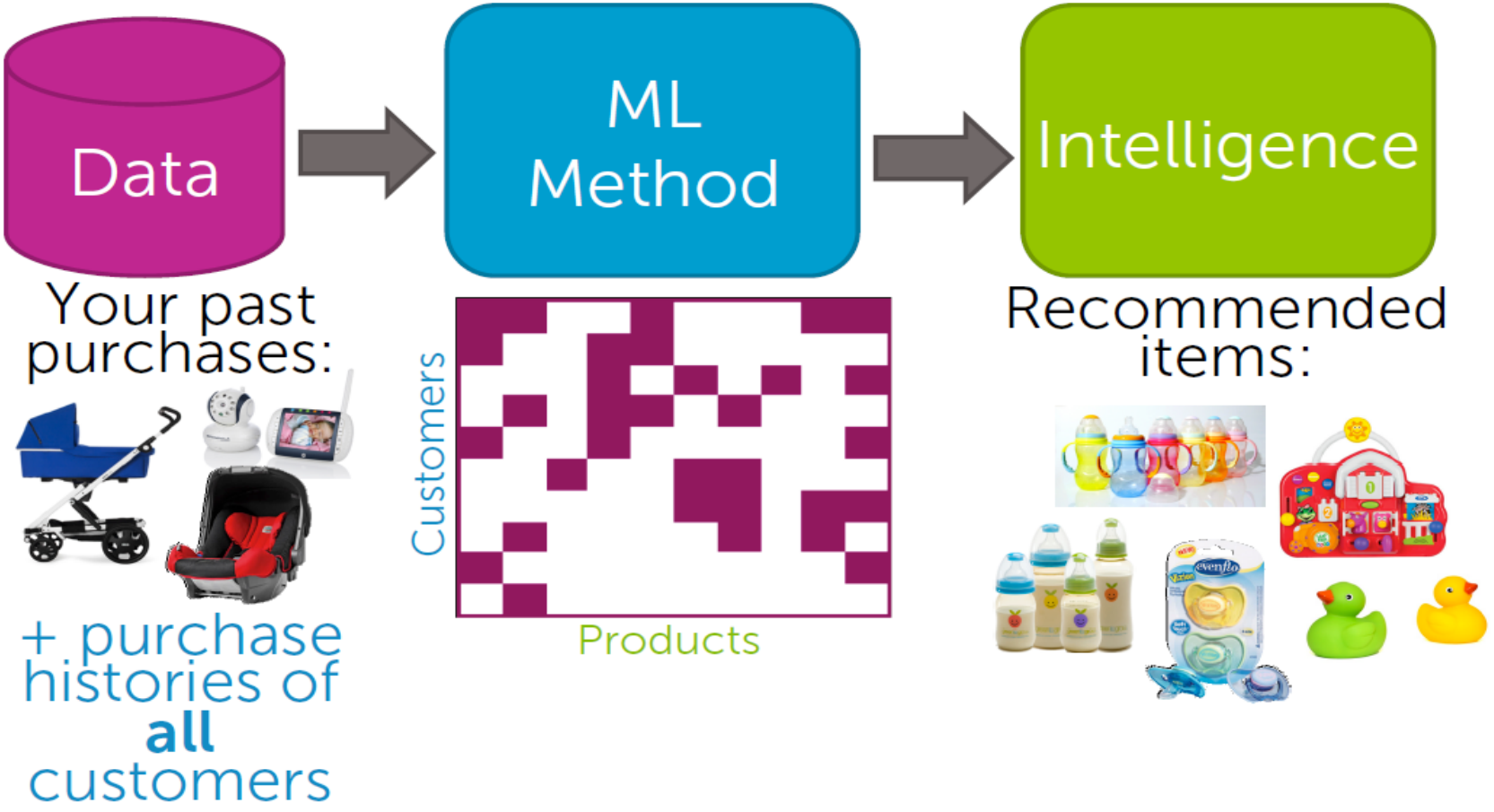
# Case study 3: Clustering





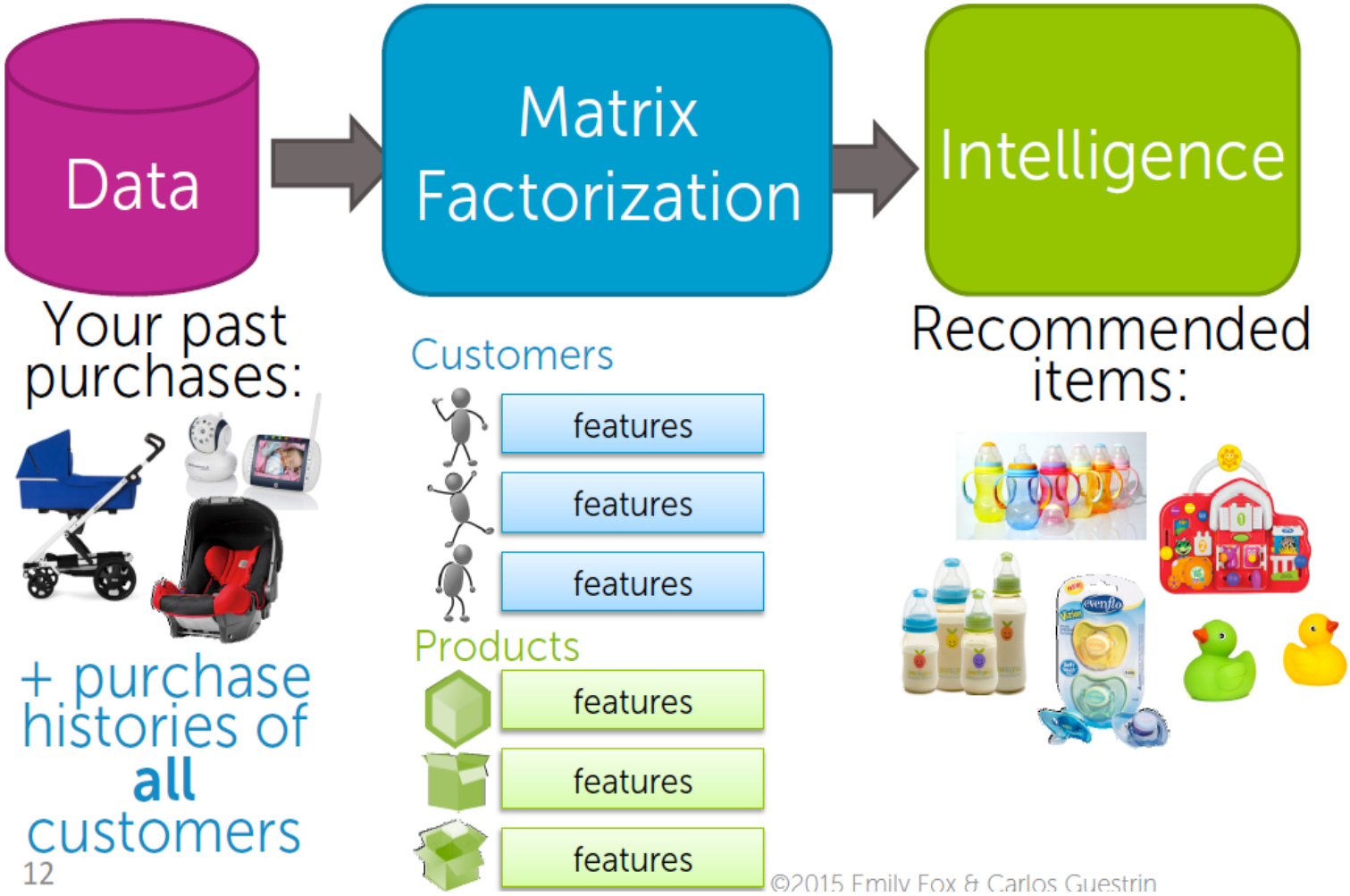
# Case study: Product recommendation (not covered here)





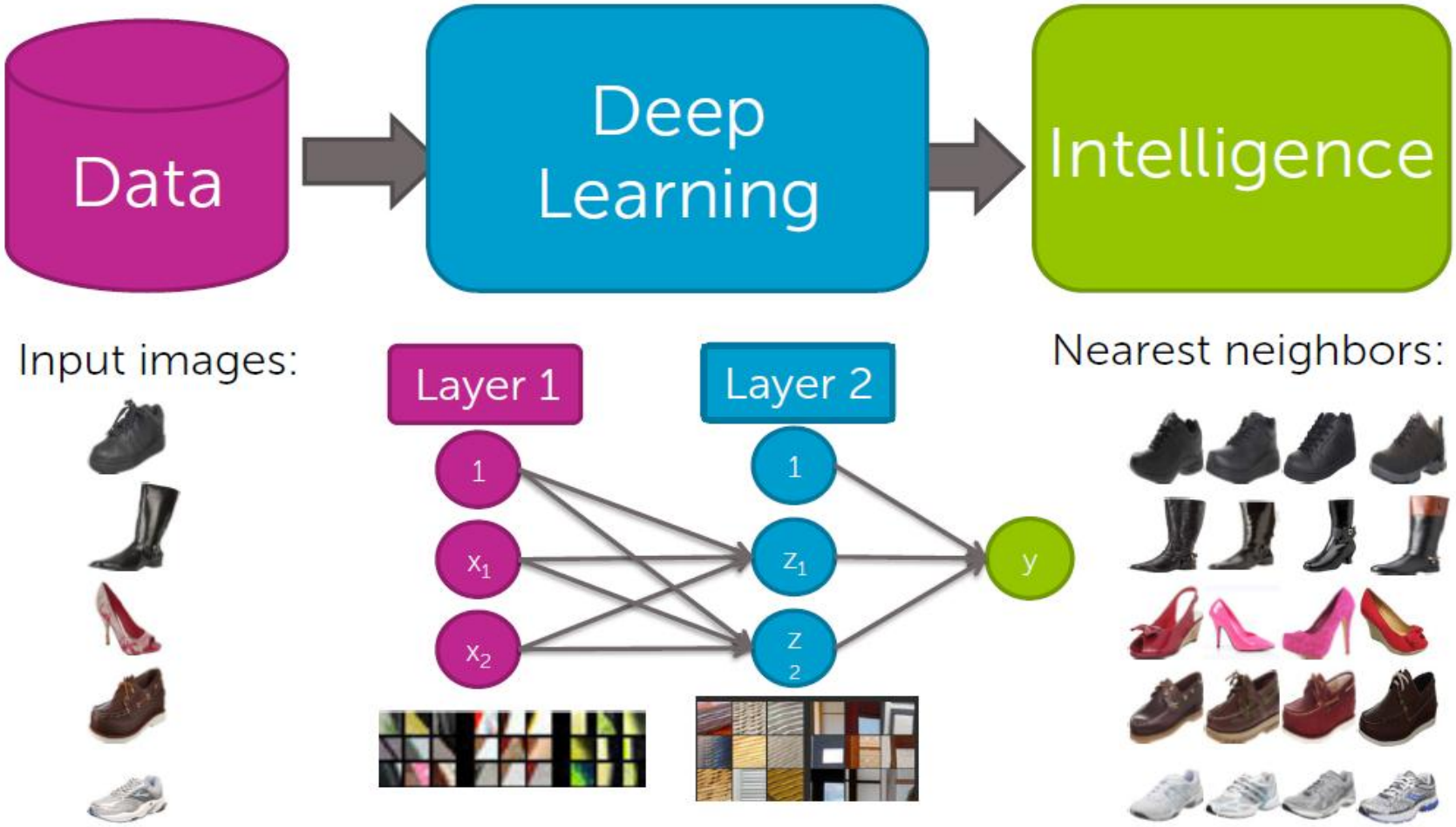
# Case study: Product recommendation (not covered here)





# Case study: Visual product recommender (not covered here)

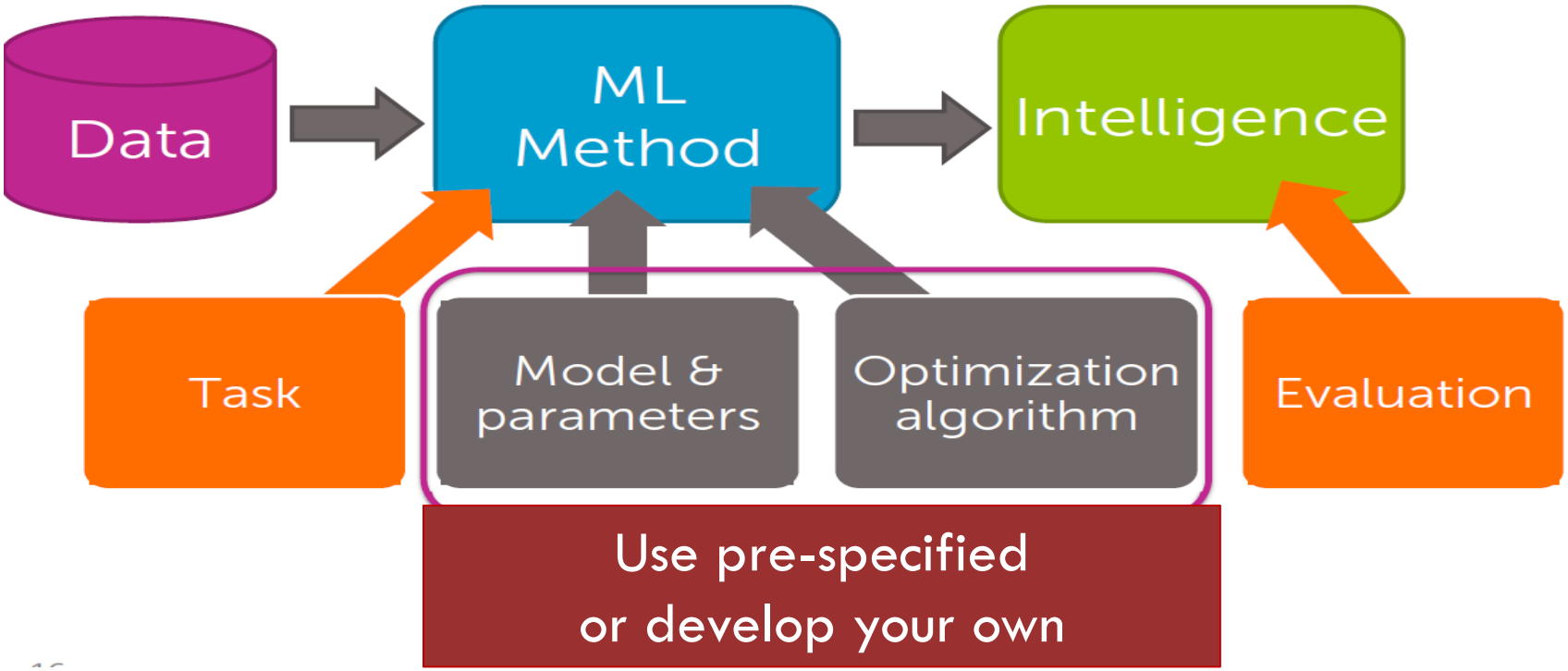




# Deploing inteligence module



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



# Prediction: Predicting house prices



**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



**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



**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



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