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

31/10/2017

WFAiS UJ, Informatyka Stosowana

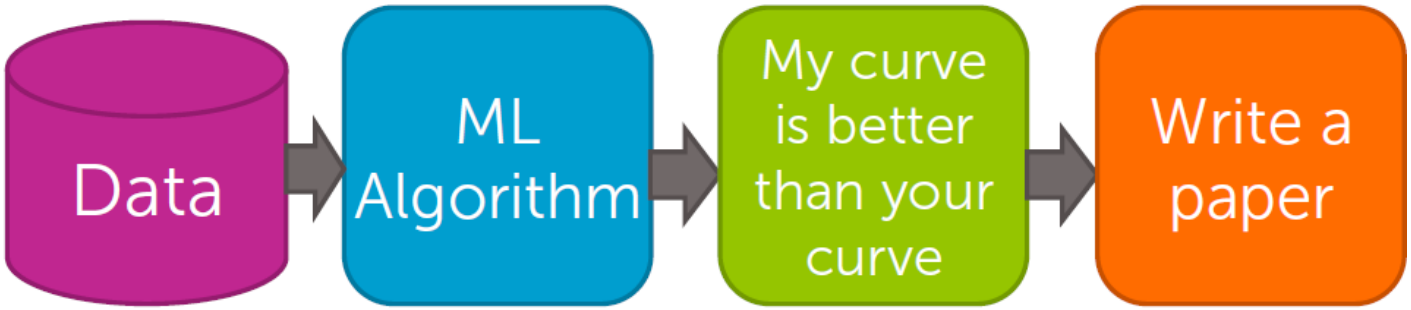
II stopień studiów

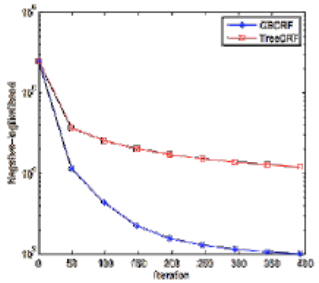
What I will cover

- Case studies for **Machine Learning** applications in data analysis
  - Should take us 6 weeks, more details follow
- Case studies for **Inference from Statistics** application in data analysis
  - Should take us 2 weeks, mor details latter

Analyse data with Machine Learning

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



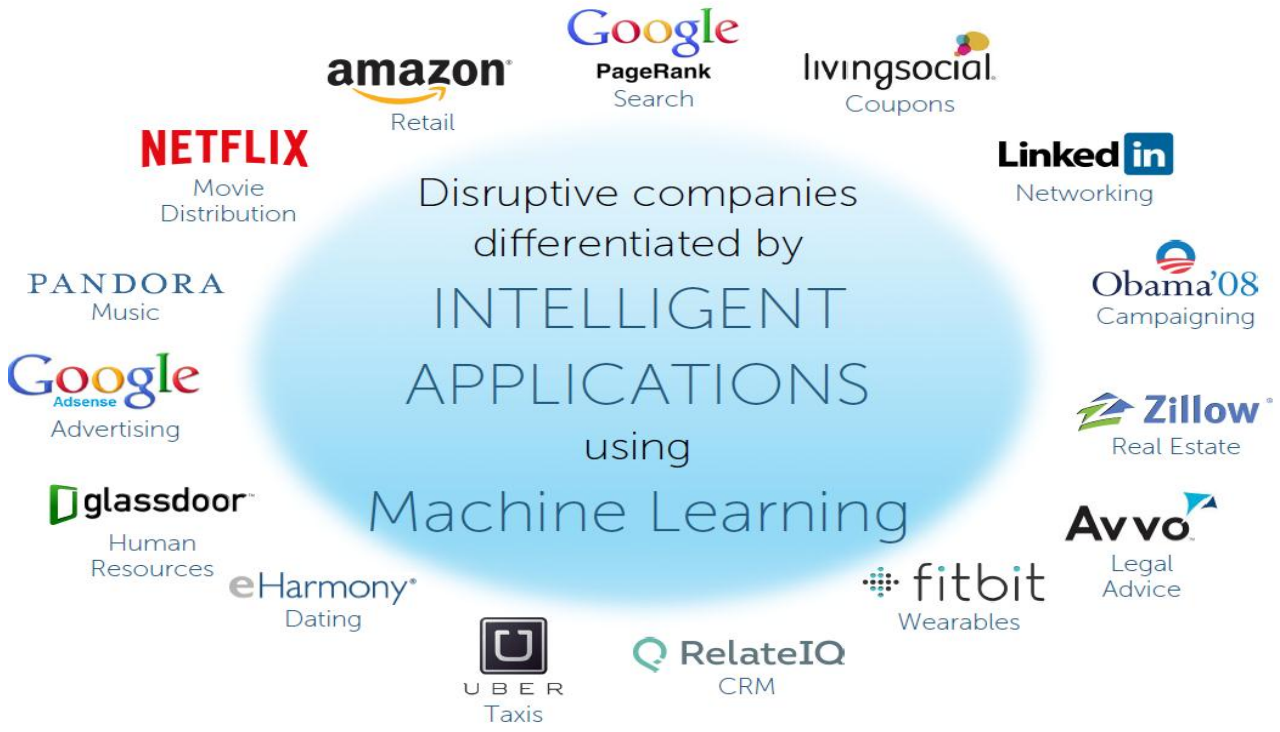


Neural Information Processing Systems Foundation

ICML

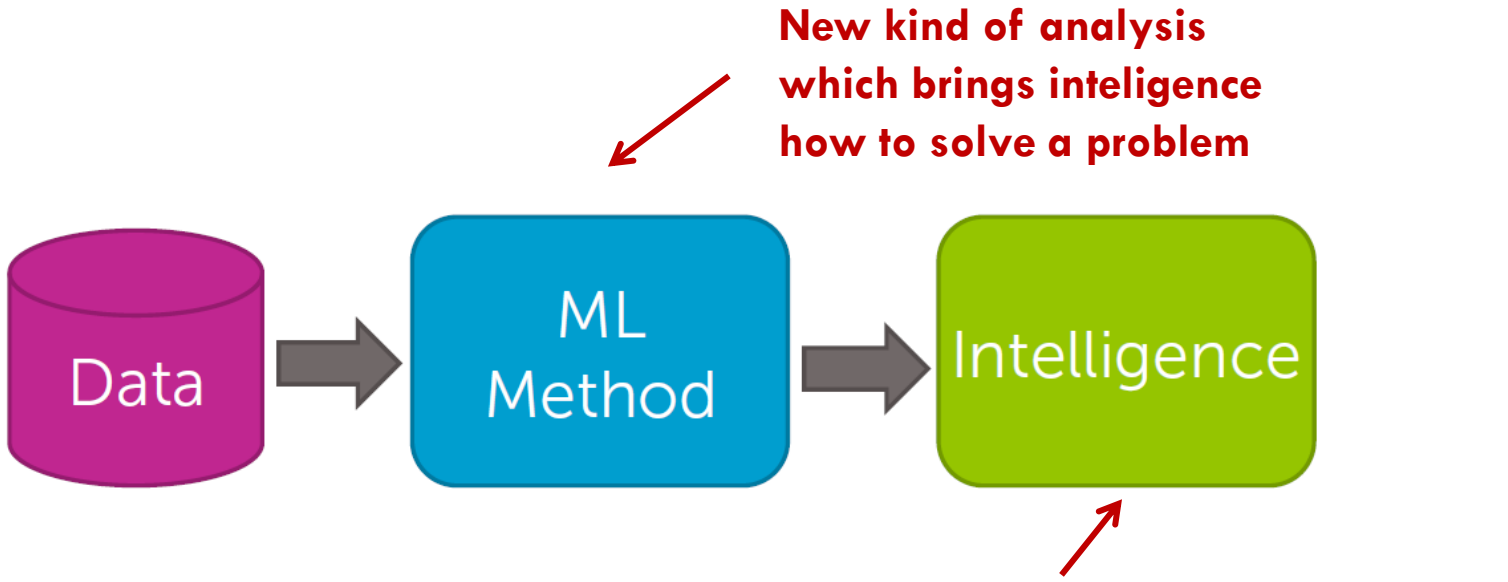
Machine learning is changing the world

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

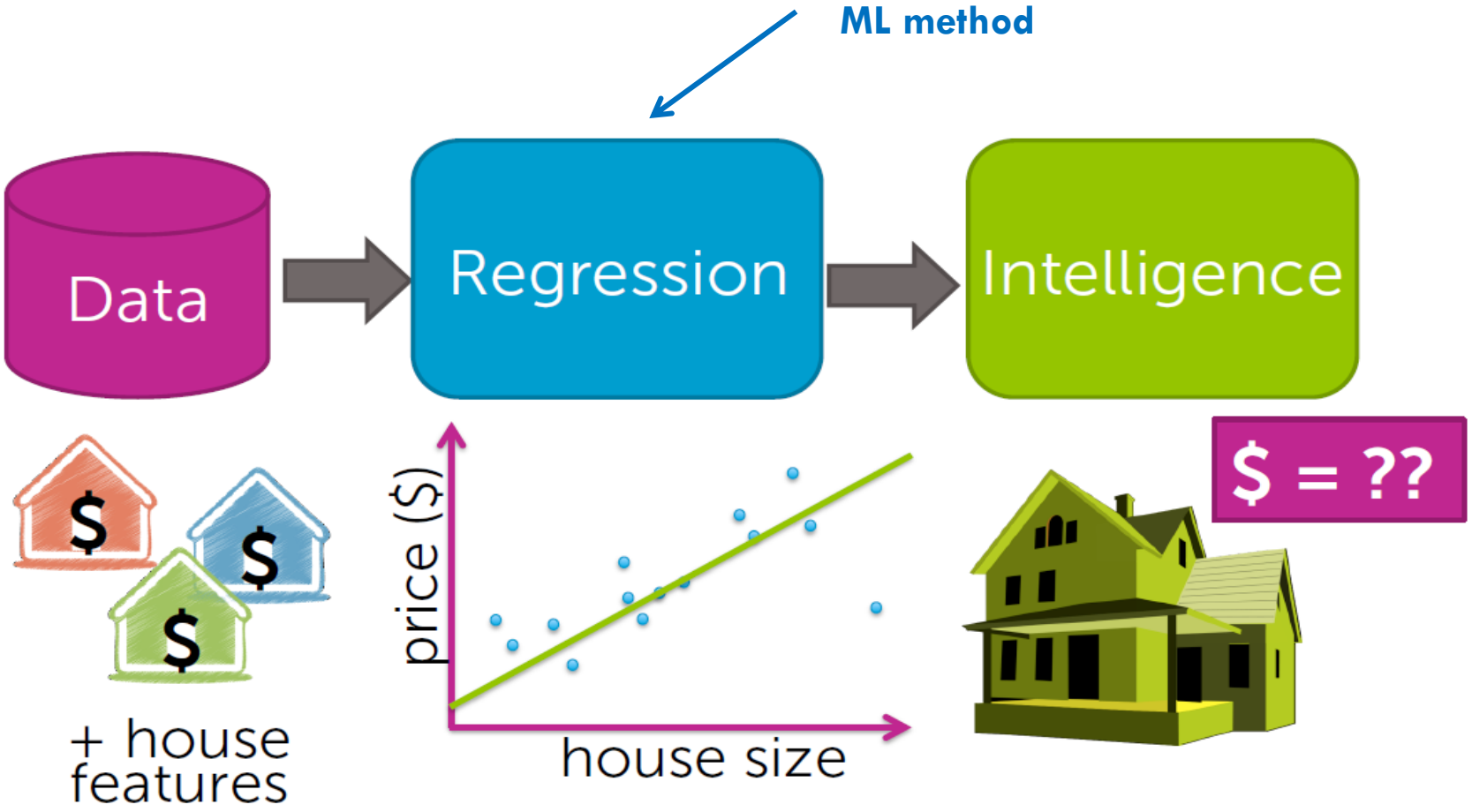


Machine learning

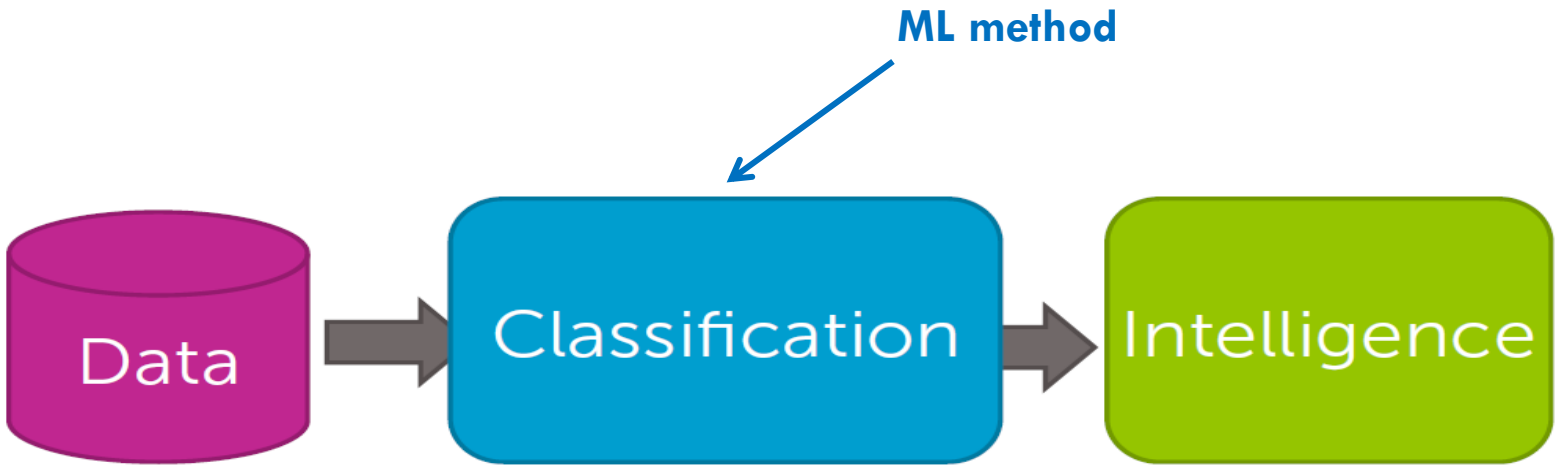
- **Data → inteligence pipeline**



Case study 1: Prediction

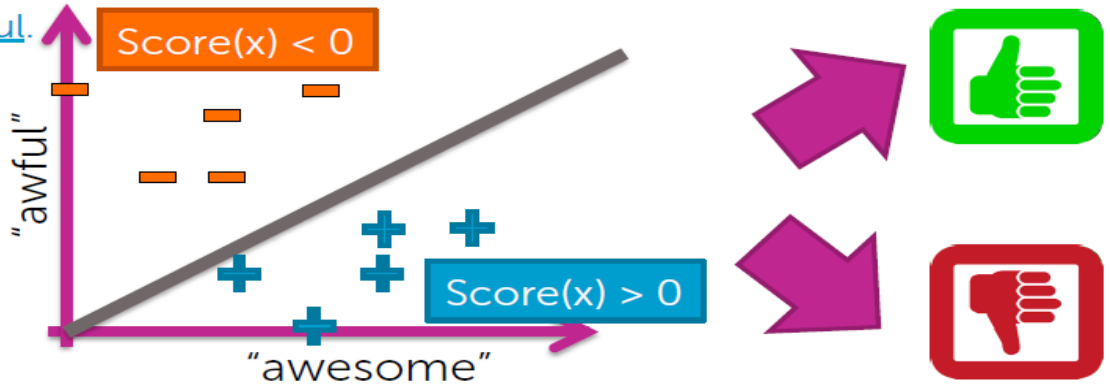


Case study 2: Classification

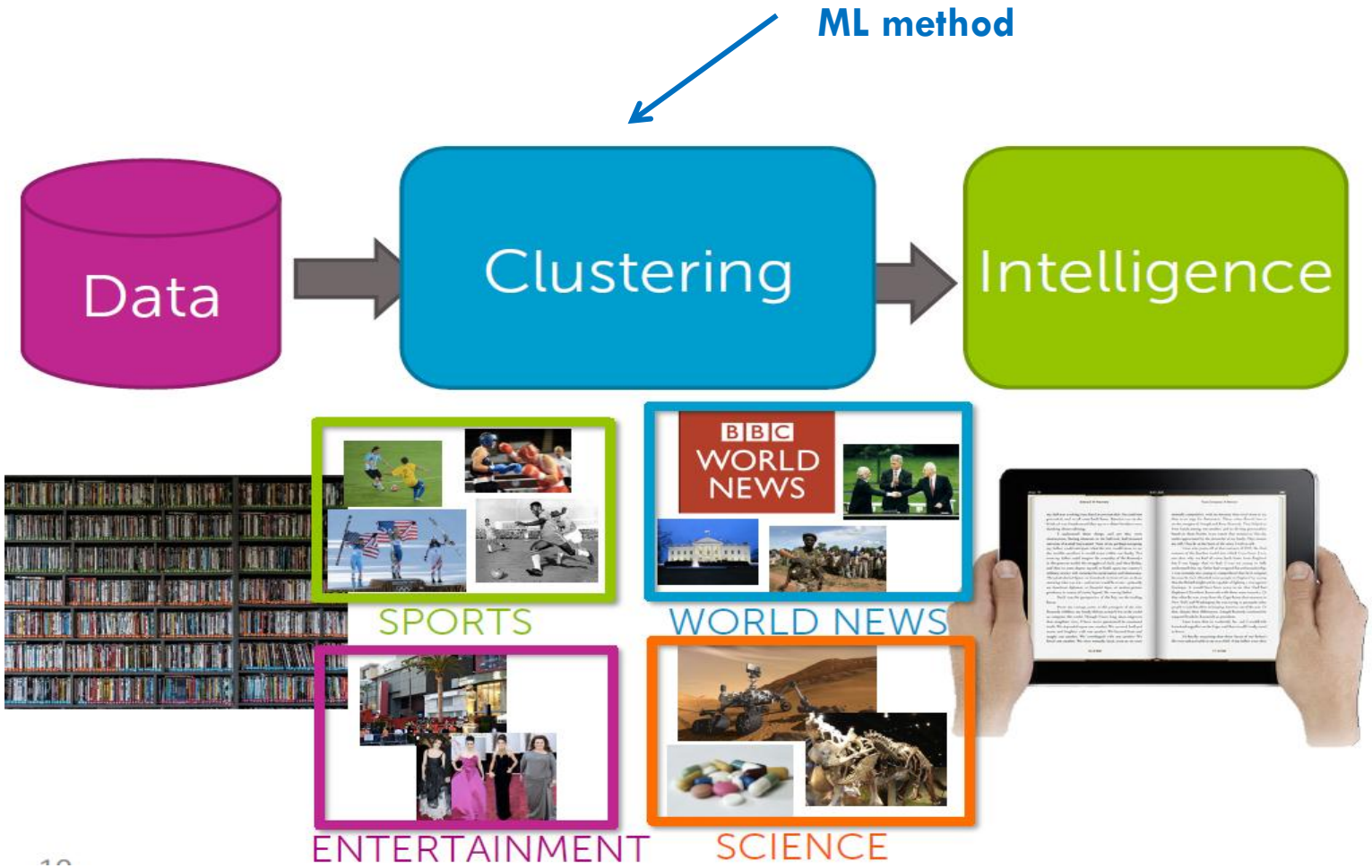


Sushi was awesome,
the food was awesome,
but the service was awful.

All reviews:

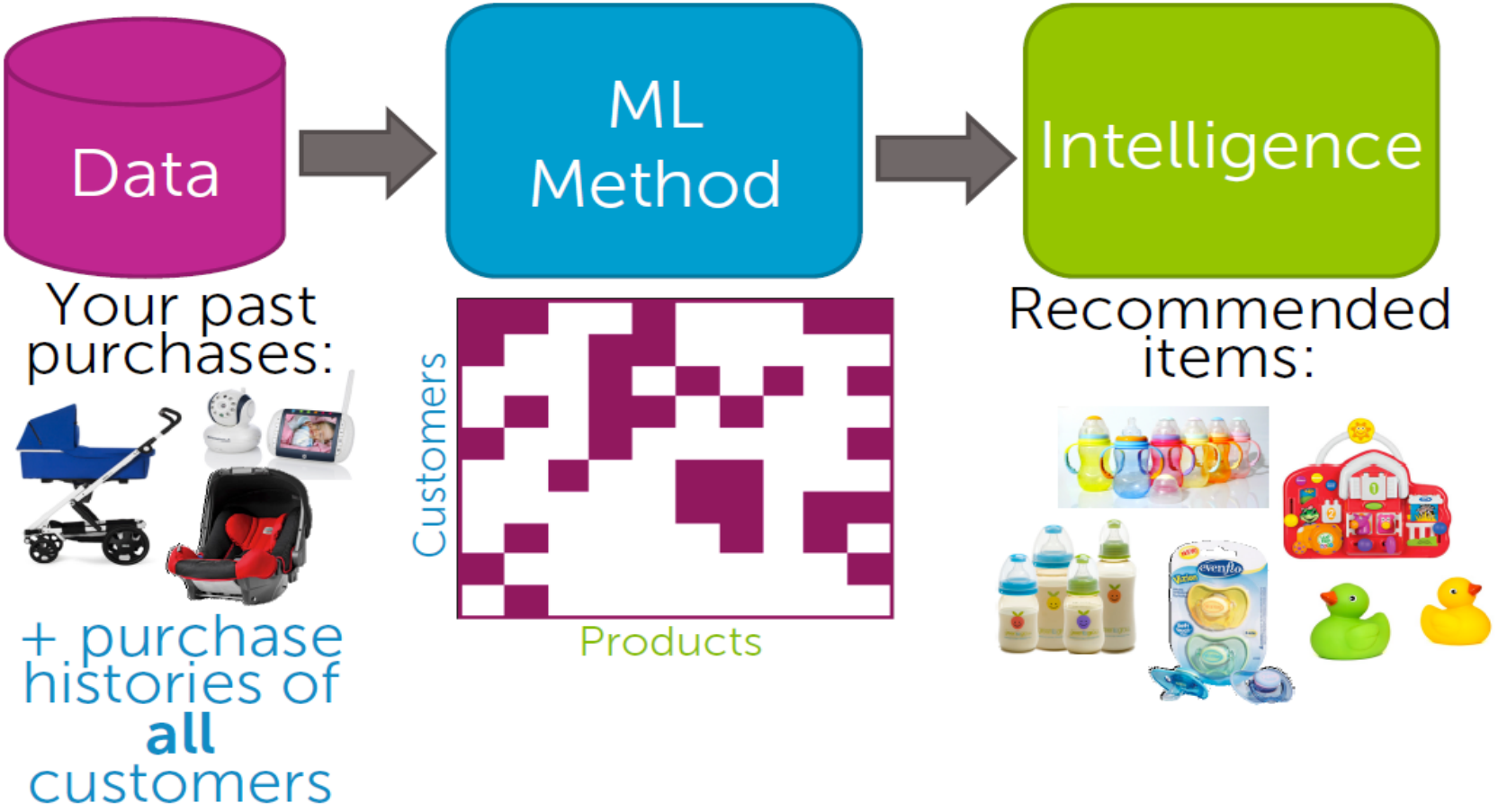


Case study 3: Clustering

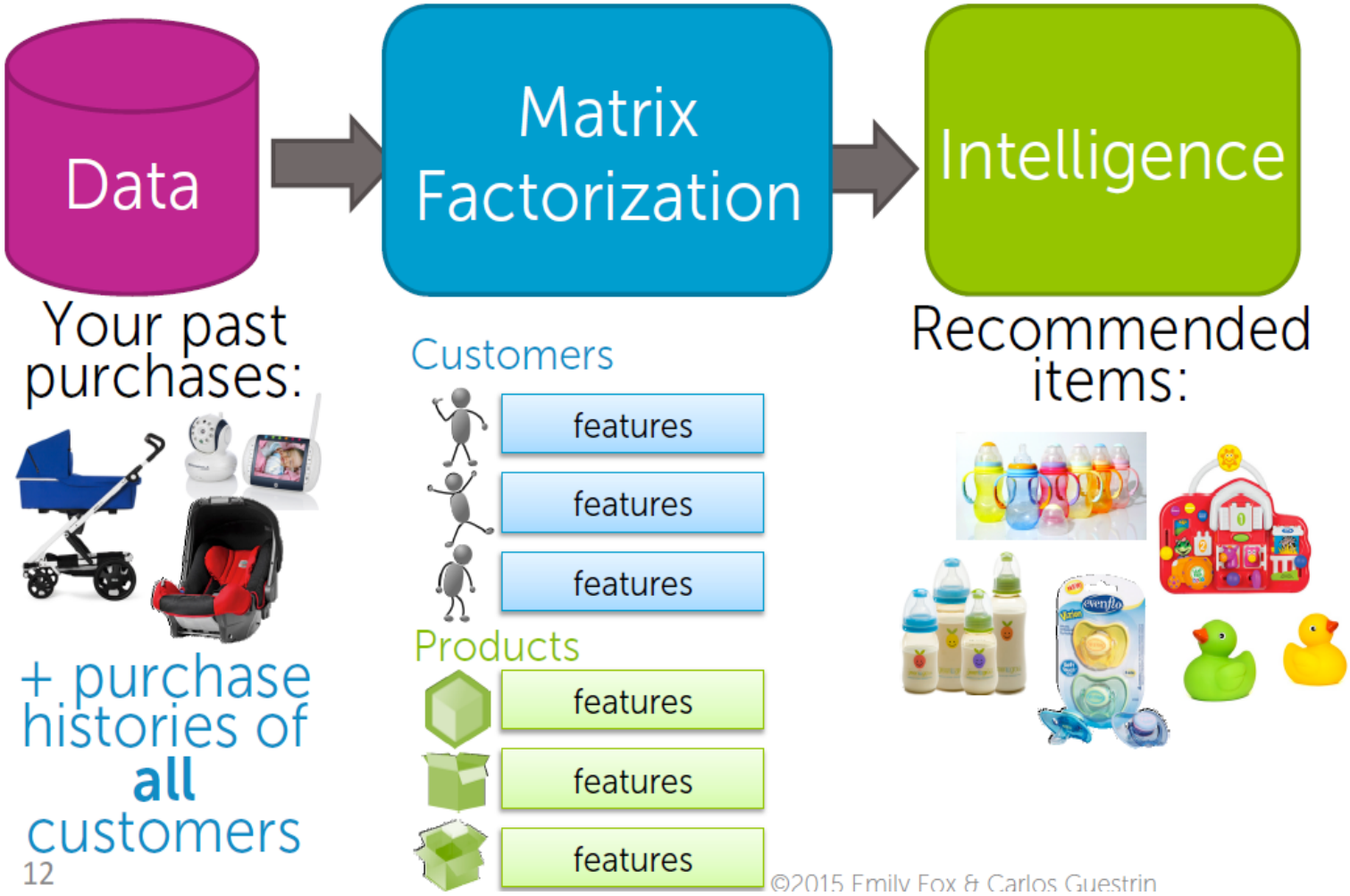


Case study: Product recommendation

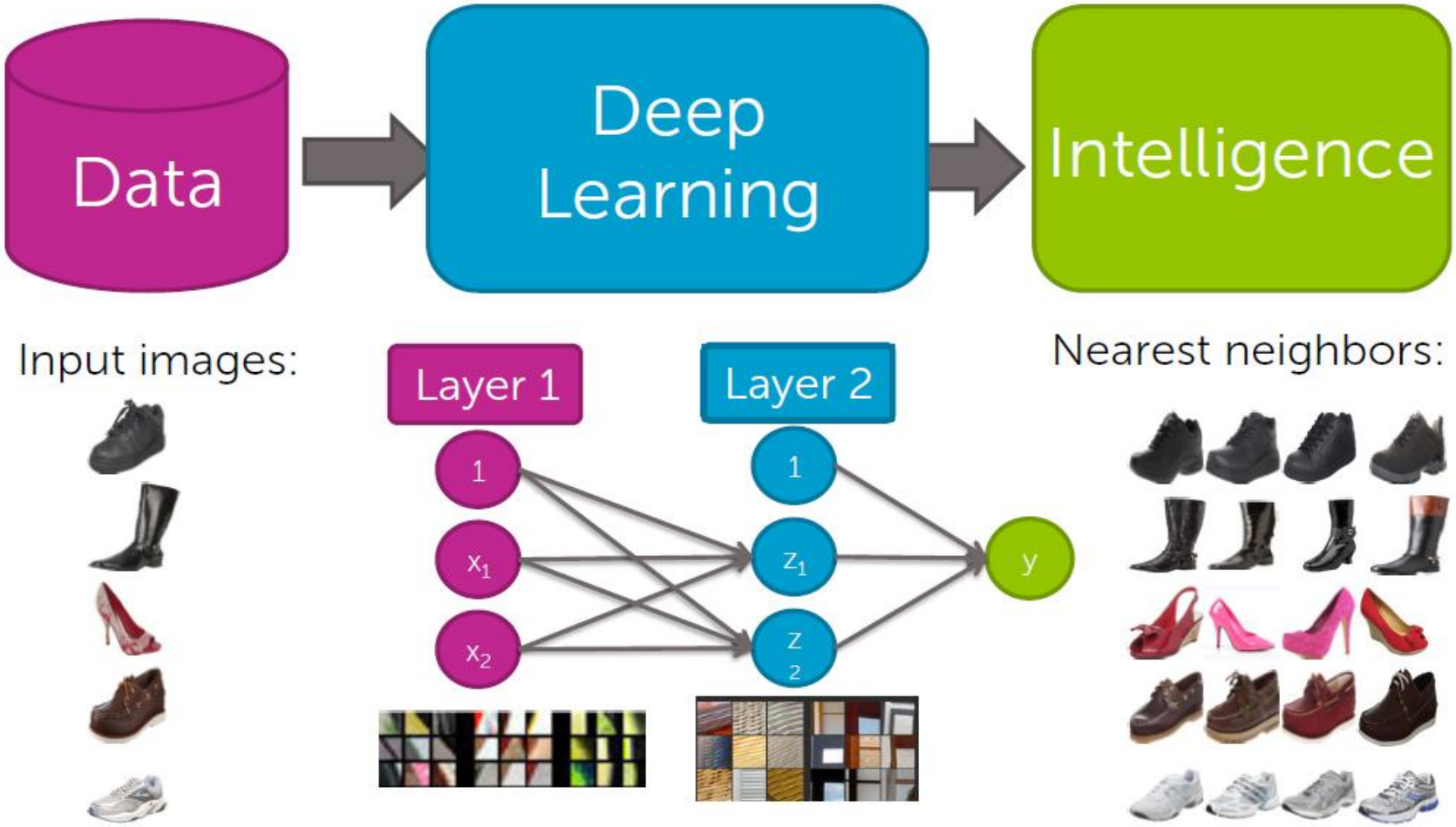
(not covered here)



Case study: Product recommendation (not covered here)

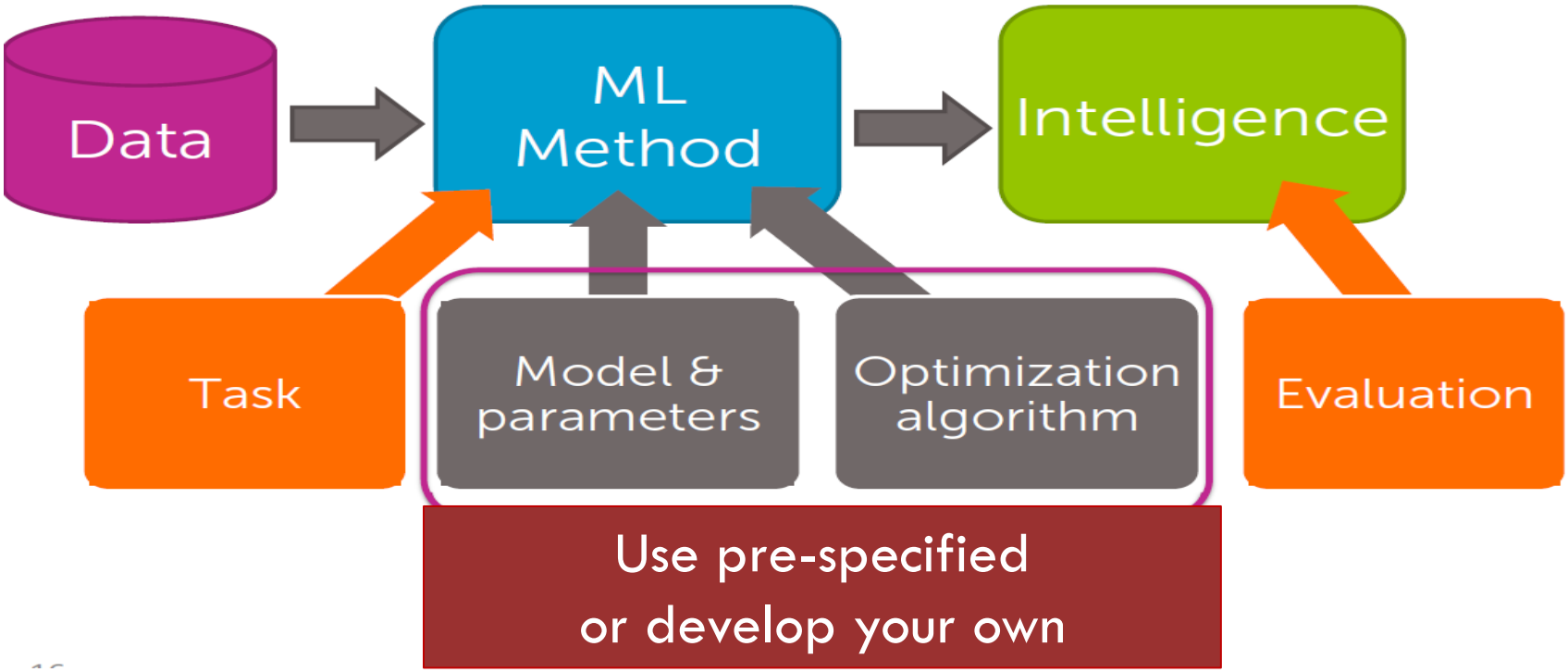


Case study: Visual product recommender (not covered here)



Deploing inteligence module

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



Lectures for each case study

- **Start with „Primer" level**
  - **Each group prepares simple analysis at this level**
- **Continue with „Advanced" level**
  - **Each group selects only one advanced level project and dive into it, maybe even beyond the scope of the lectures.**

Each case study will take us 2 weeks of lectures.

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