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

13/10/2020

WFAiS UJ, Informatyka Stosowana

I stopień studiów

What is Data Science?

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?

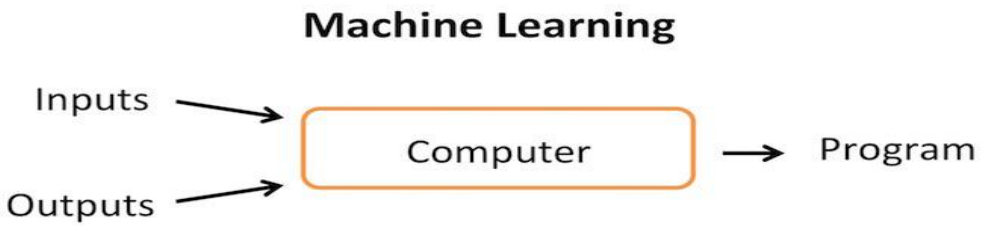
Machine learning algorithms are algorithms that learn (often predictive) models from data. I.e., instead of formulating "rules" manually, a machine learning algorithm will learn the model for you.

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

Machine Learning is the field of study that gives computers the ability to learn without being explicitly programmed
– Arthur Samuel (1959)



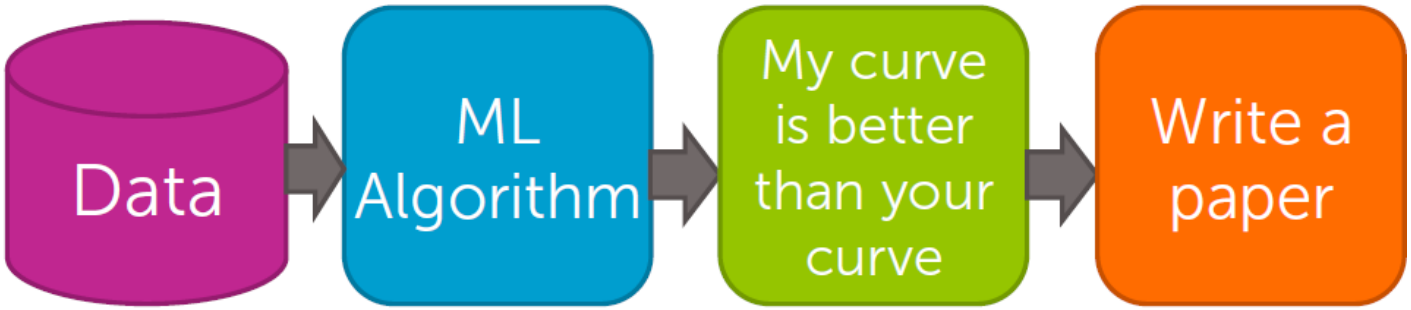
Sebastian Raschka, 2016

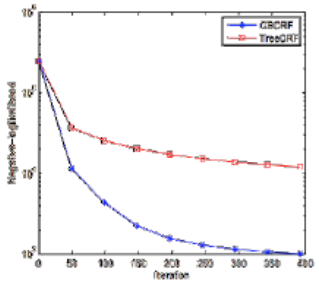
Outline of the course

- **Exploratory Data Analysis:** introduction
  → today
- **Data Analysis** with **Machine Learning** algorithms:
  → from next week till mid December 2020
  - Regression,
  - Classification,
  - Retrieval &Clustering
- **Other topics:**
  → weeks in January 2021
  - Model building and Monte Carlo methods
  - Statistical Inference and Data Analysis
  - Multivariate techniques and Artificial Neural Networks

Analyse data with Machine Learning

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



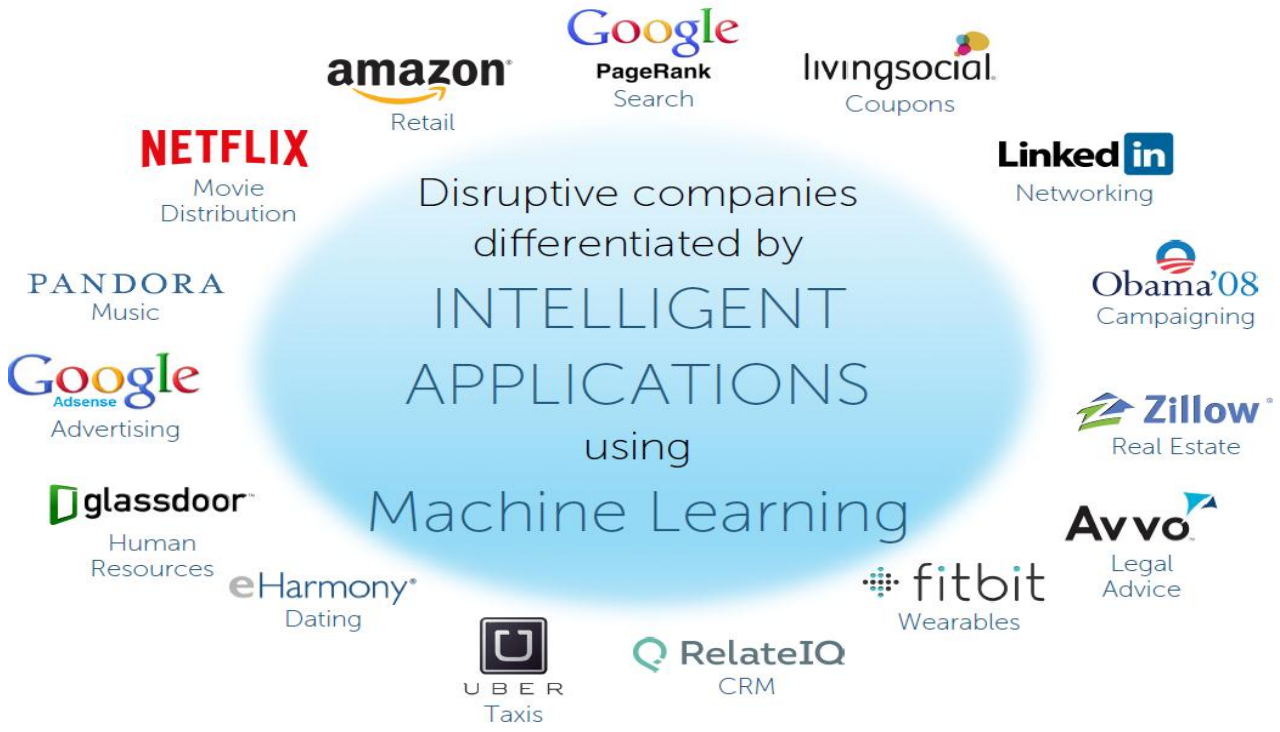


Neural Information Processing Systems Foundation

ICML

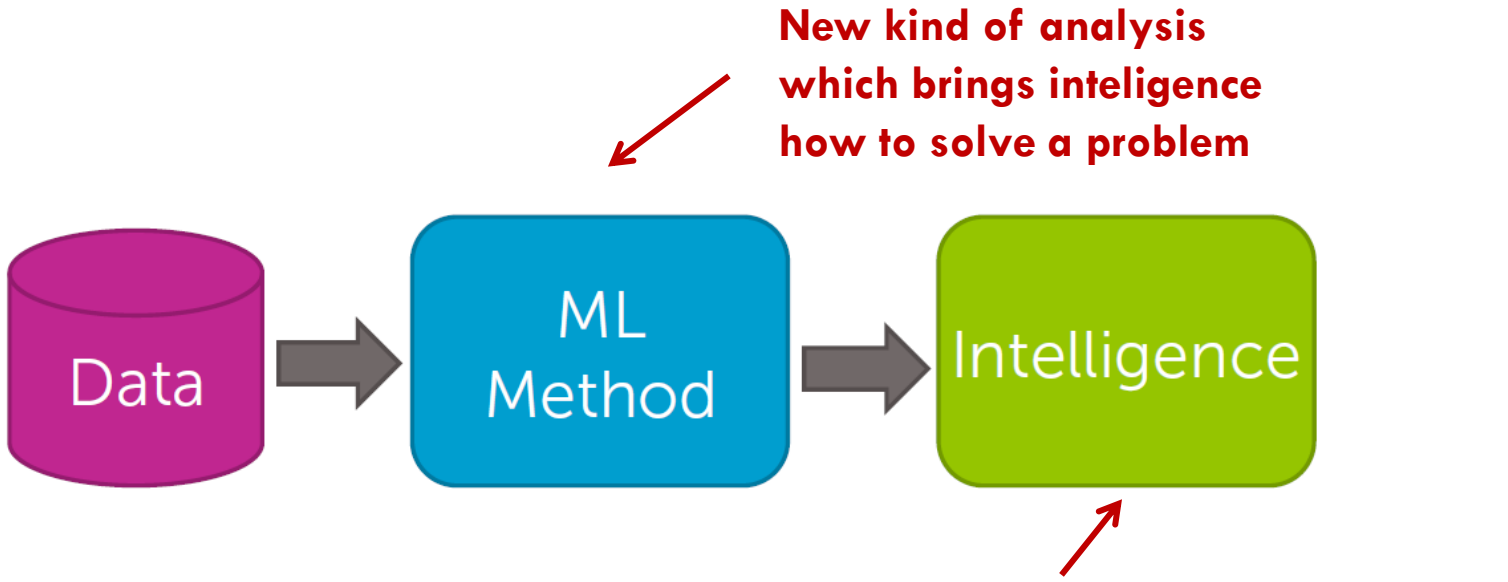
Machine learning is changing the world

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

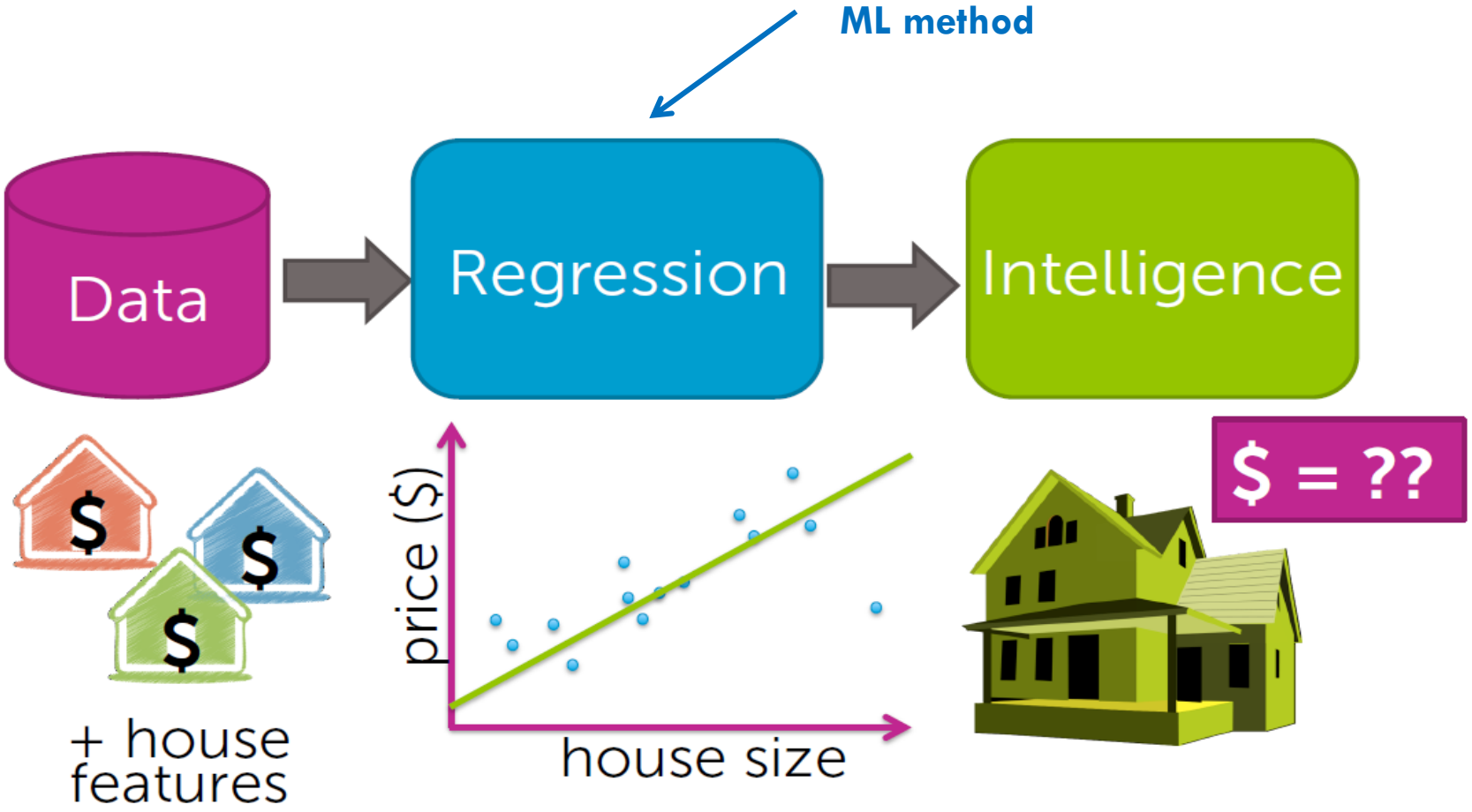


Machine learning

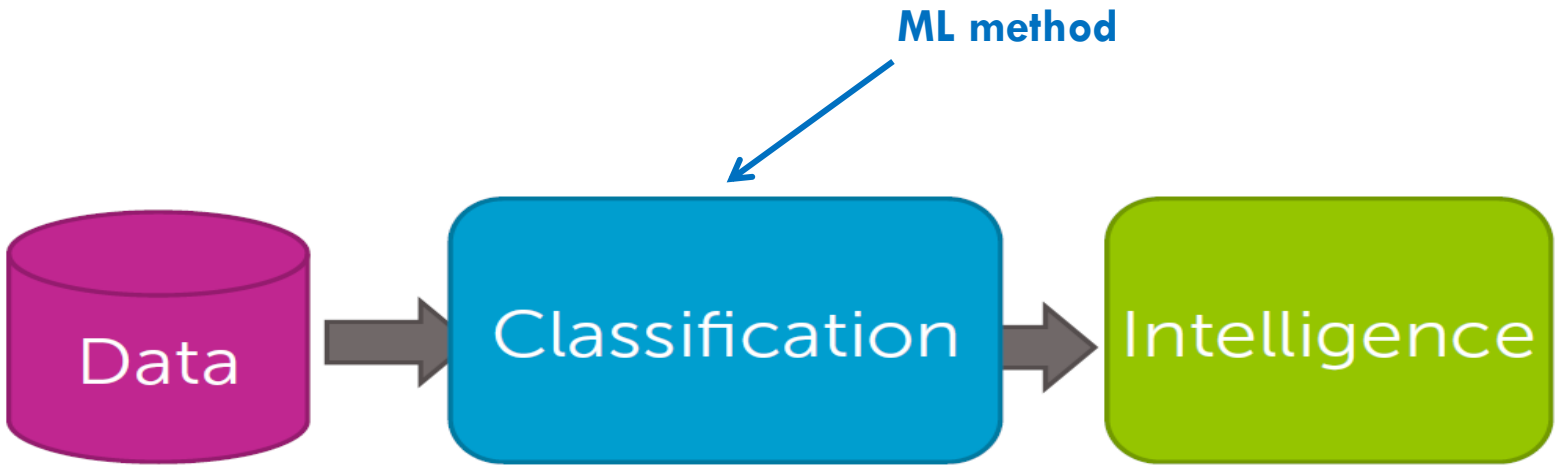
- **Data → inteligence pipeline**



Case study 1: Prediction

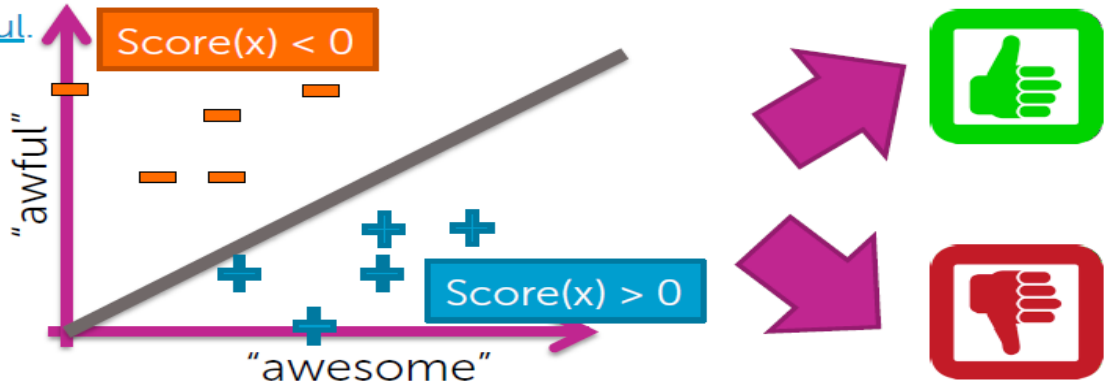


Case study 2: Classification

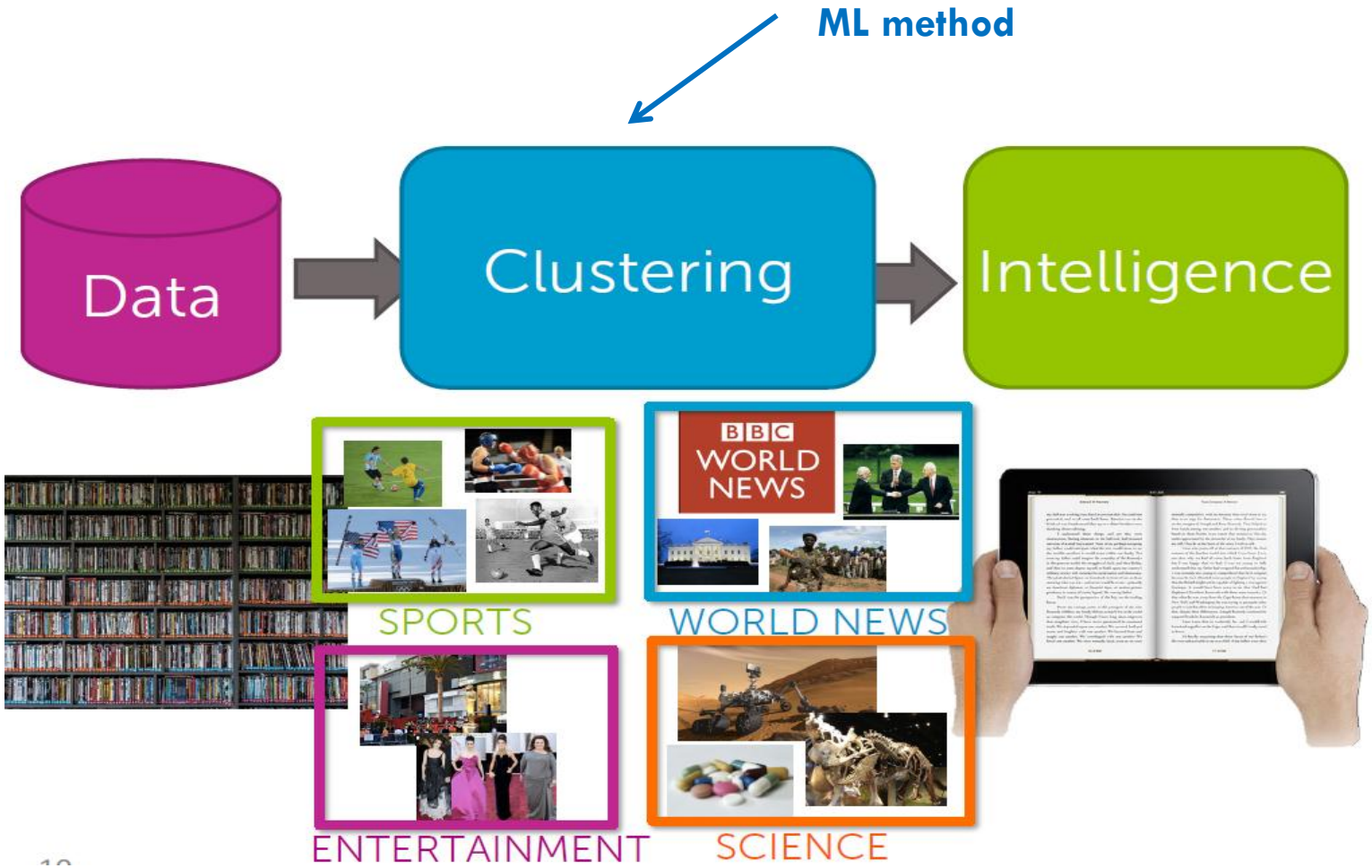


Sushi was awesome,
the food was awesome,
but the service was awful.

All reviews:

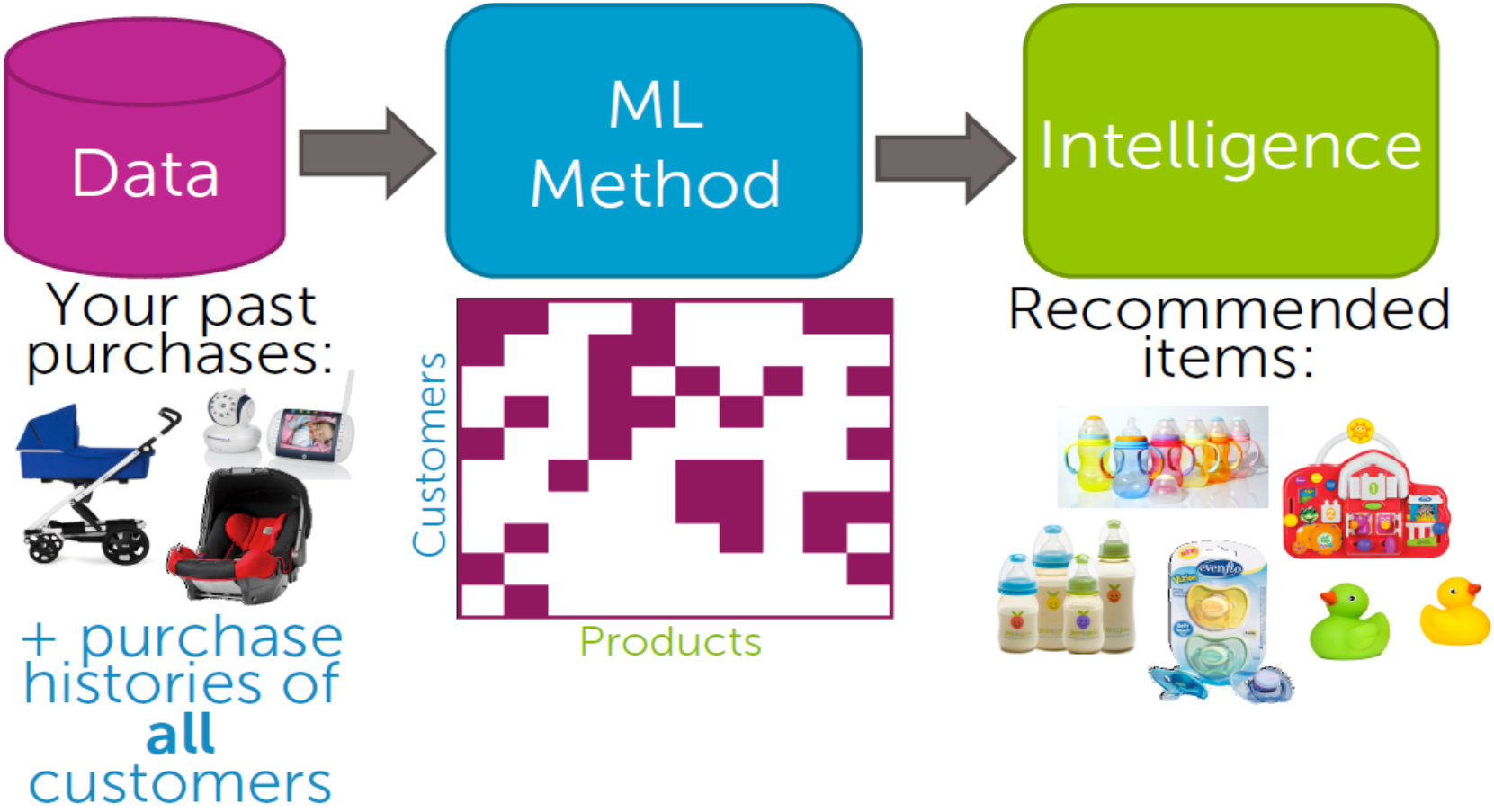


Case study 3: Clustering

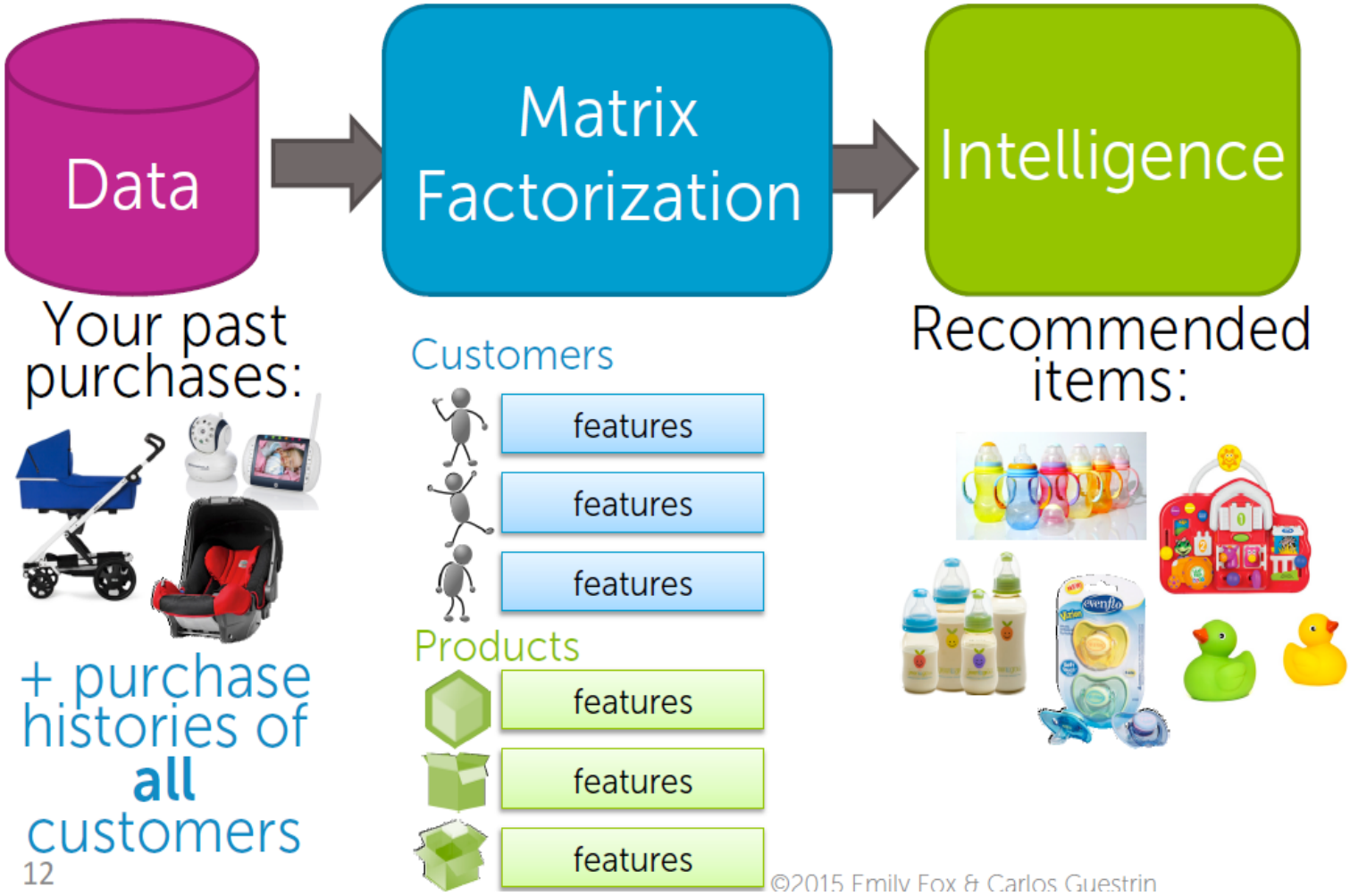


Case study: Product recommendation

(not covered here)

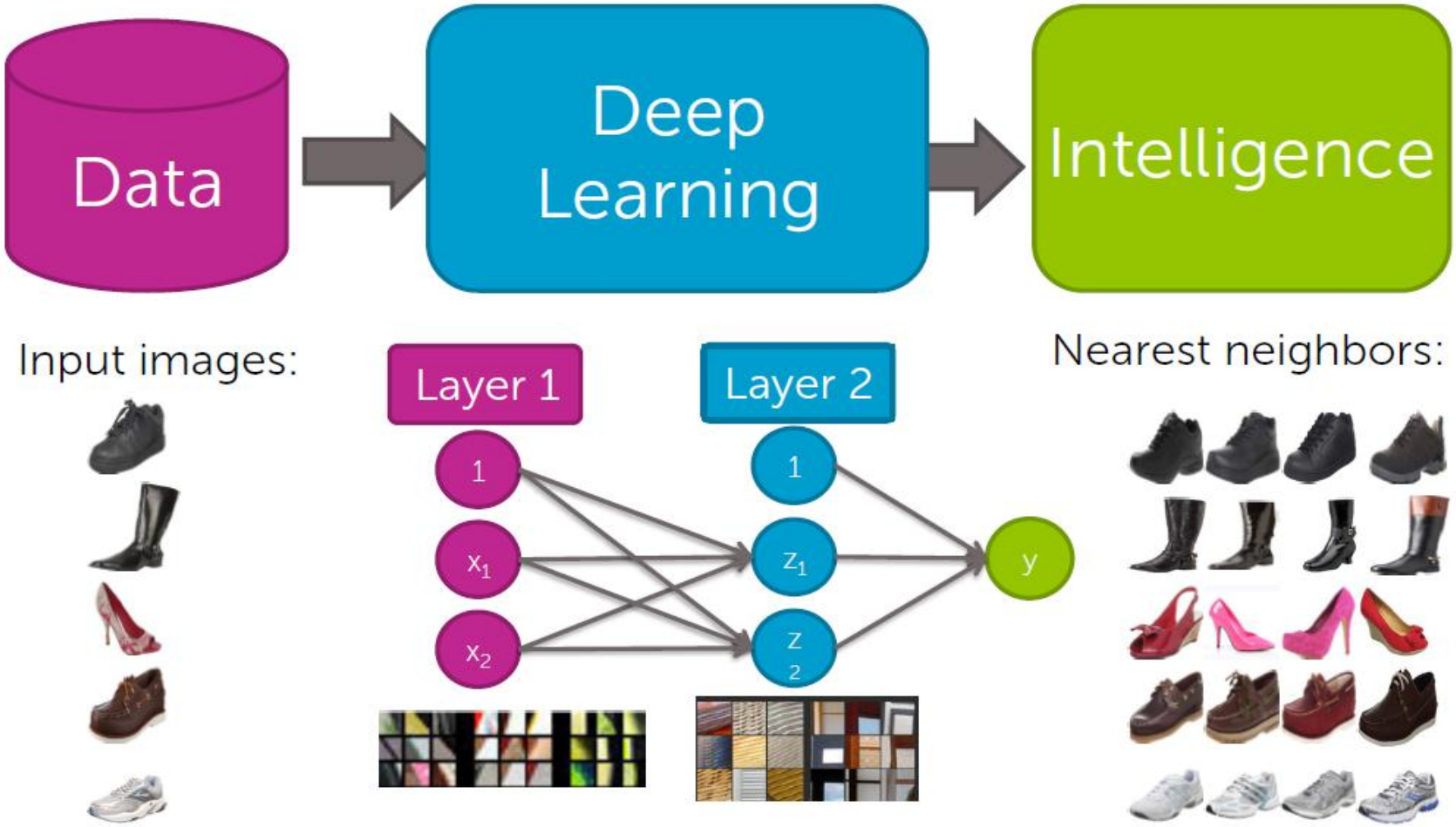


Case study: Product recommendation (not covered here)



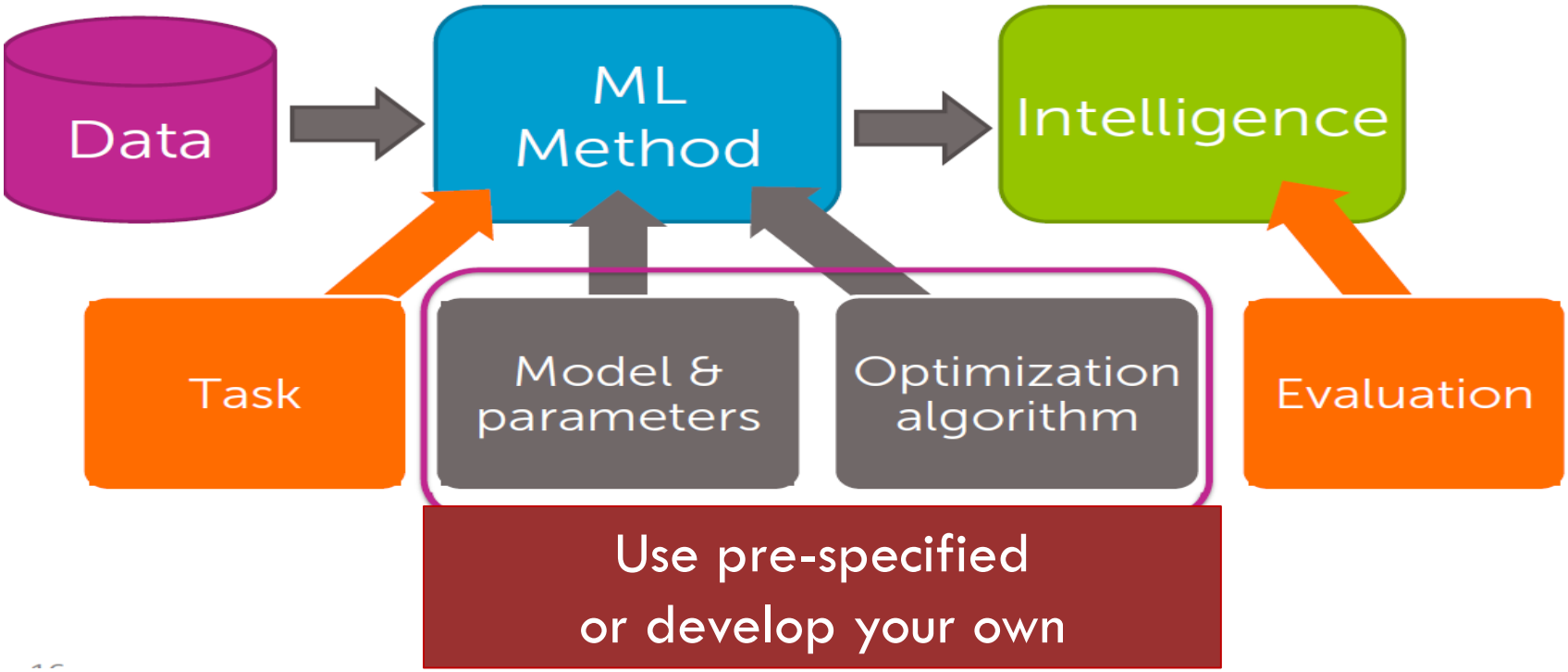
©2015 Emily Fox & Carlos Guestrin

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