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

2/10/2024

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

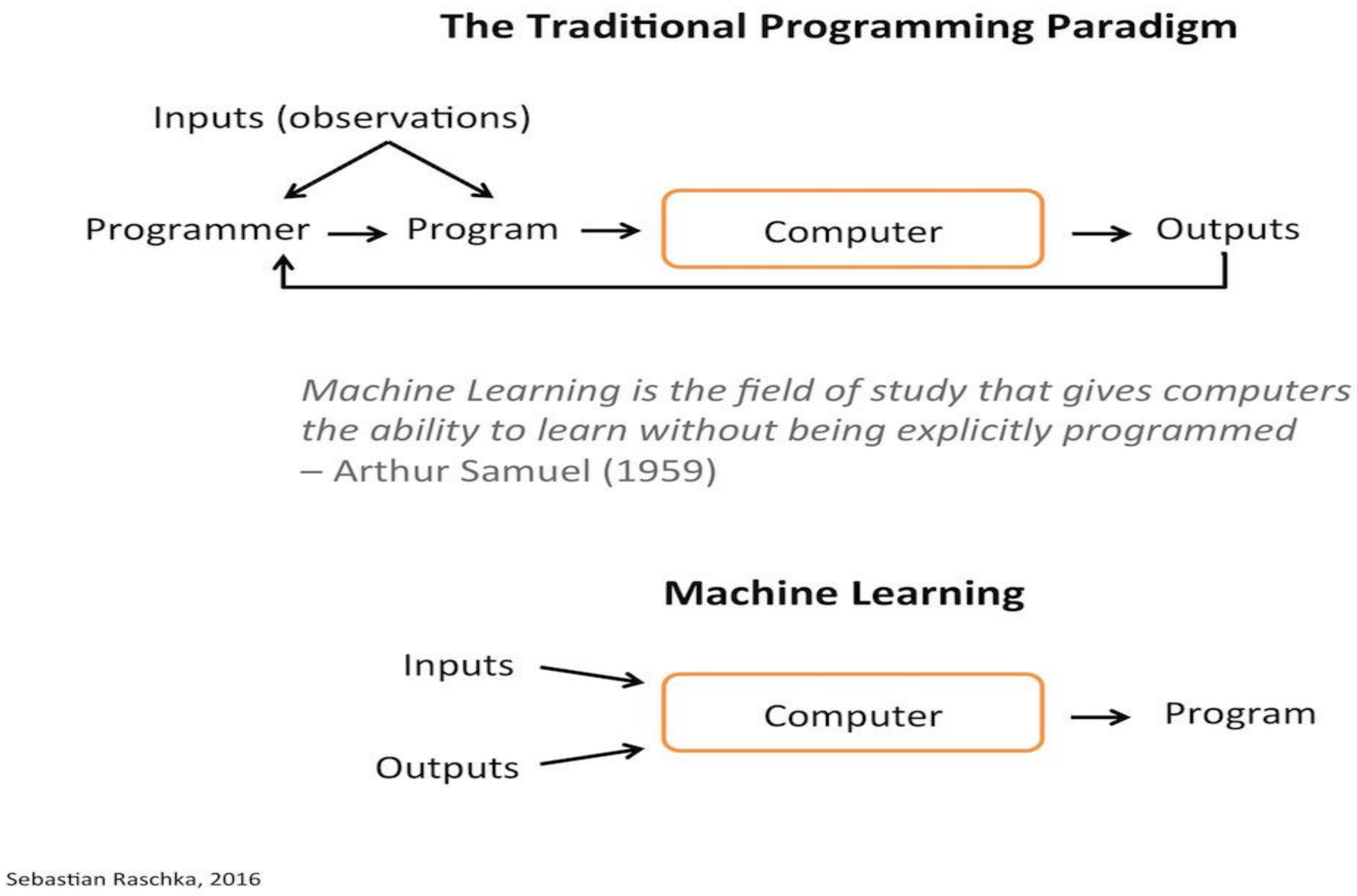
Data Science and Machine Learning?

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



The Traditional Programming Paradigm

Inputs (observations)

Programmer → Program → Computer → Outputs

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

Machine Learning

Inputs

Outputs

Computer → Program

Sebastian Raschka, 2016

Outline of the course

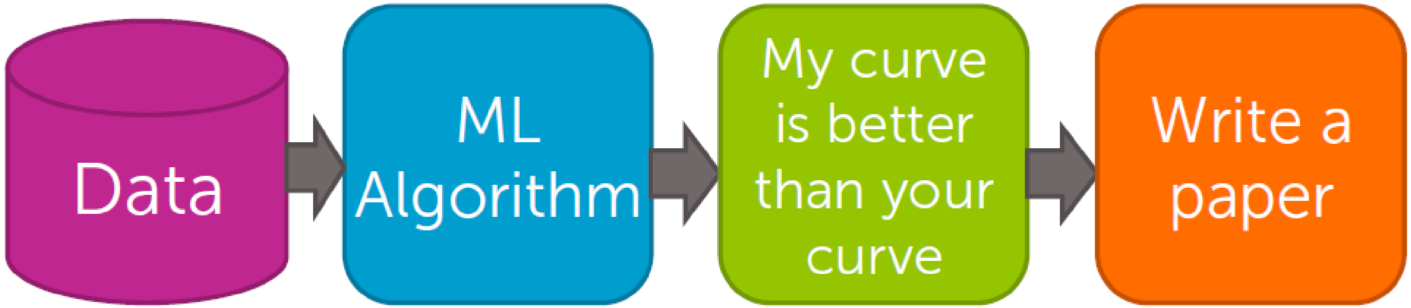
- **Exploratory Data Analysis: introduction**

  → **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

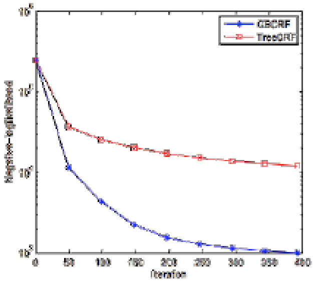
- **Machine learning is changing the world.**
- **Old view of ML**


Data

ML Algorithm

My curve is better than your curve

Write a paper

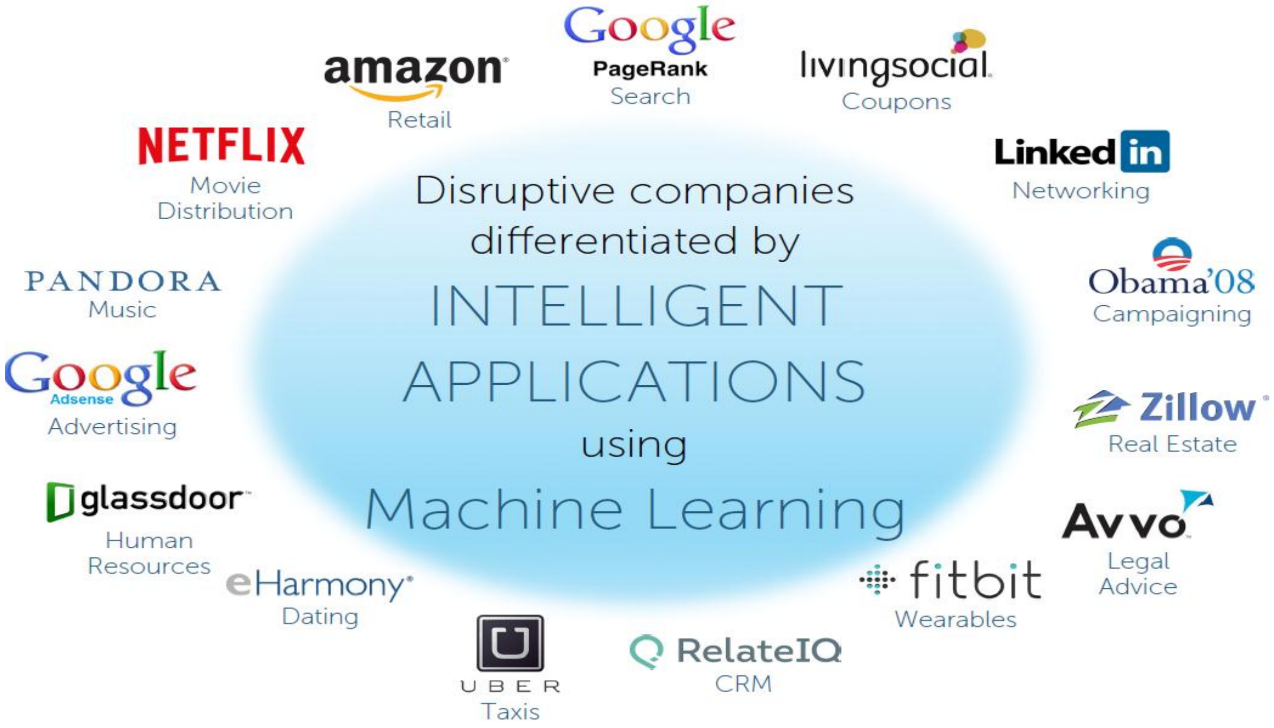


Neural Information Processing Systems Foundation

ICML

Machine learning is changing the world

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

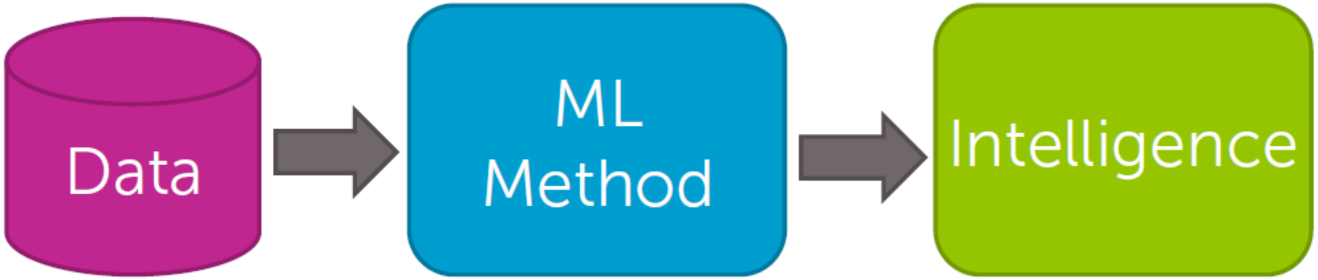


Disruptive companies differentiated by INTELLIGENT APPLICATIONS using Machine Learning

Machine learning

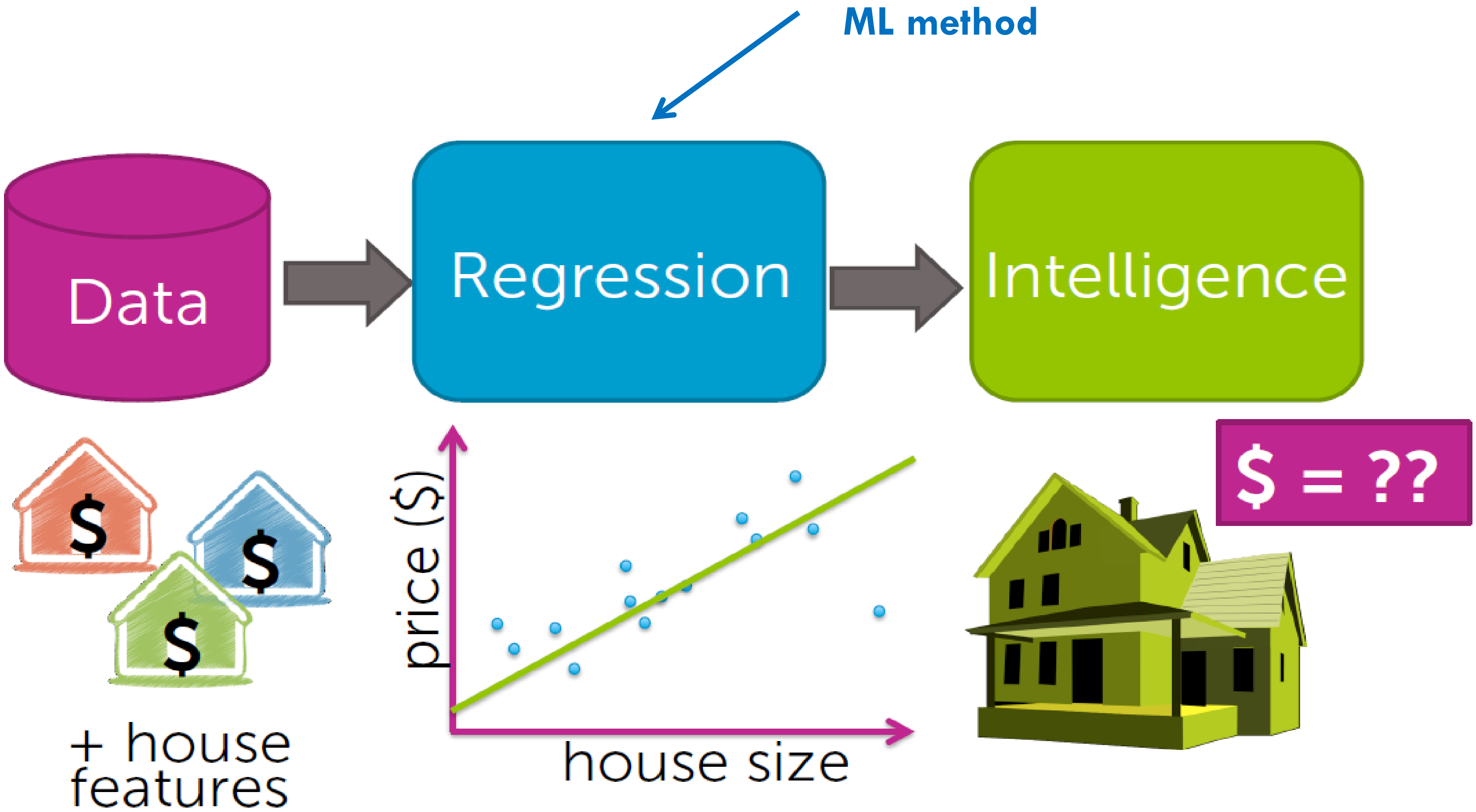
- **Data → inteligence pipeline**

New kind of analysis
which brings inteligence
how to solve a problem



Eg. which product to buy
which film to chose
connect people and taxi driver

Case study 1: Prediction



ML method

Data → Regression → Intelligence

+ house features

$ = ??

Prediction

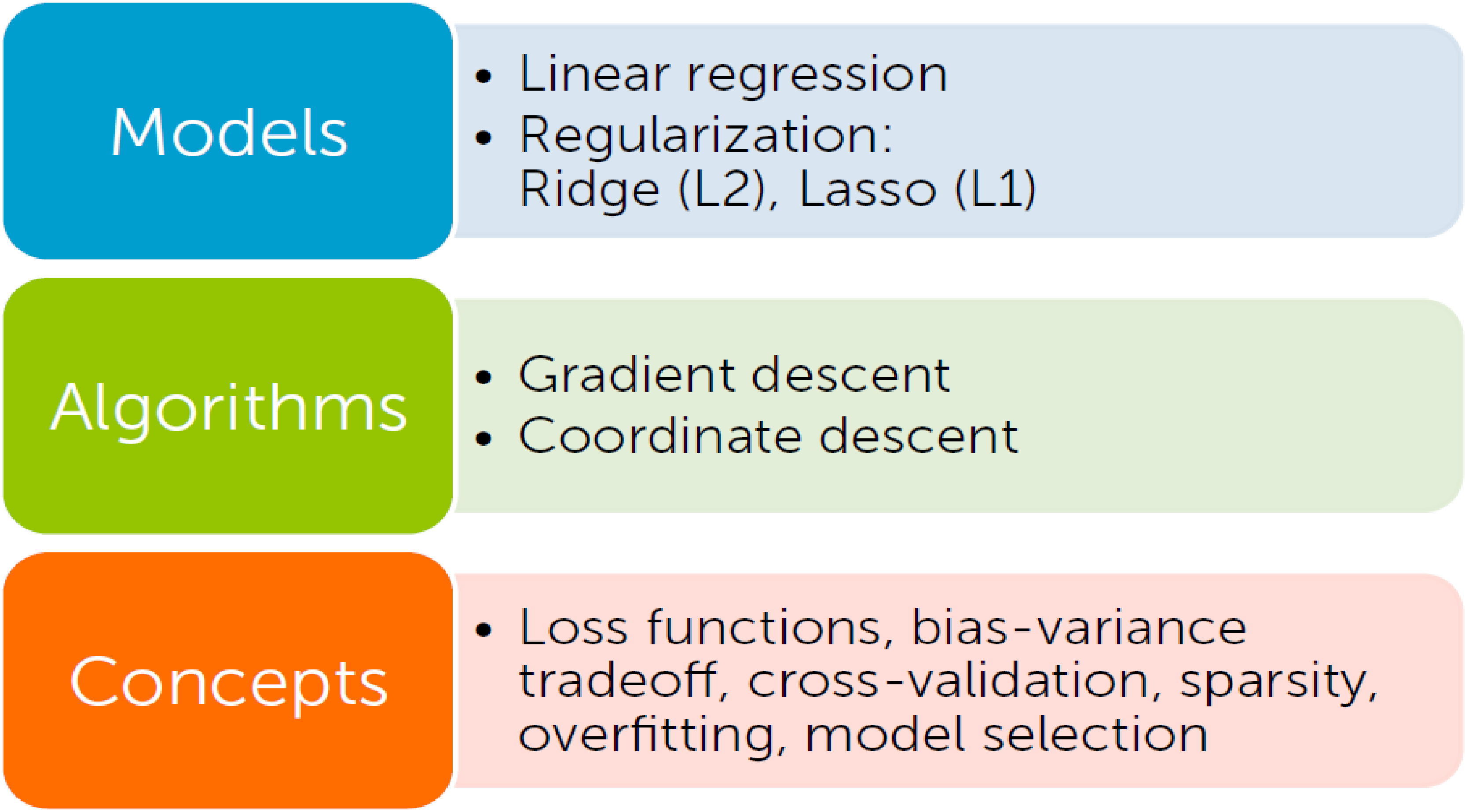
- How much will your salary be? (**y** = $$)
- Depends on **x** = performance in courses, quality of programming assignments, # of discussion responses, …

Prediction

Tweet popularity

- How many people will retweet your tweet?
- Depends on # followers, # of followers of followers, features of text tweeted, popularity of hashtag, # of past retweets,...

Prediction:



Models

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

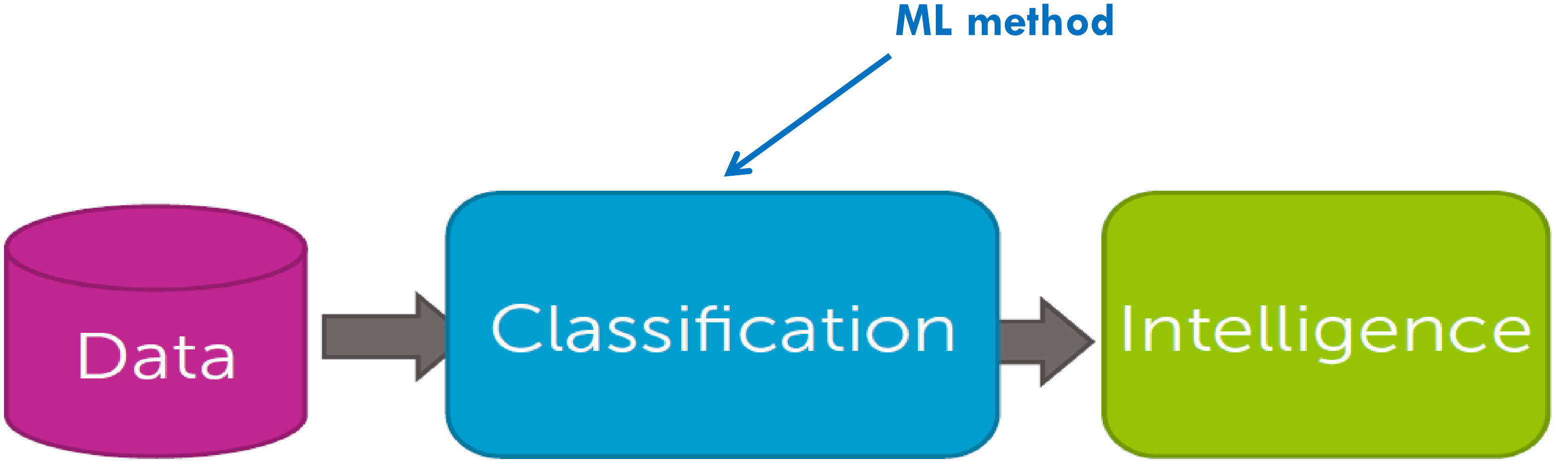
Algorithms

- Gradient descent
- Coordinate descent

Concepts

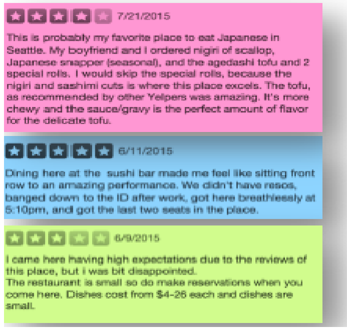
- Loss functions, bias-variance tradeoff, cross-validation, sparsity, overfitting, model selection

Case study 2: Classification



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

All reviews:



★★★★☆ 7/21/2015

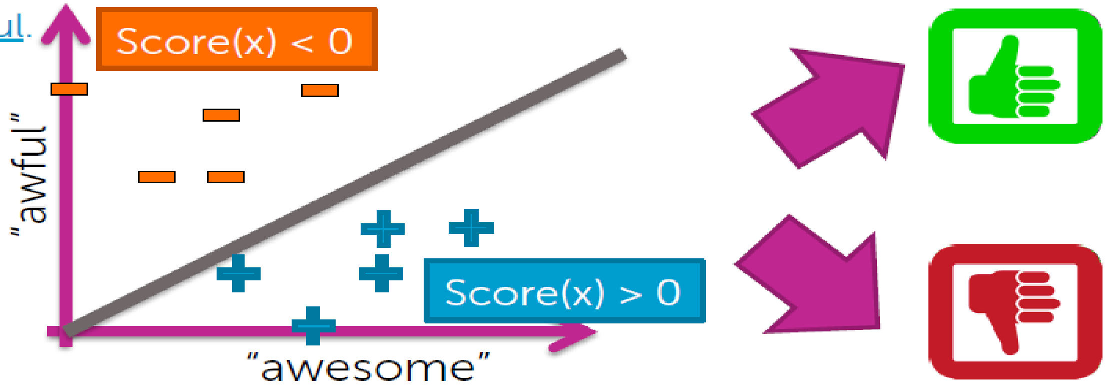
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese snapper (seasonal), and the agedashi tofu and 2 special rolls. I would skip the special rolls, because the nigiri and sashimi cuts is where this place excels. The tofu, as recommended by other Yelpers was amazing. It's more chewy and the sauce/gravy is the perfect amount of flavor for the delicate tofu.

★★★★★ 6/11/2015

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

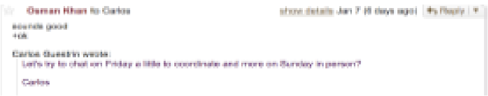
★★★☆☆ 6/9/2015

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

Spam filtering

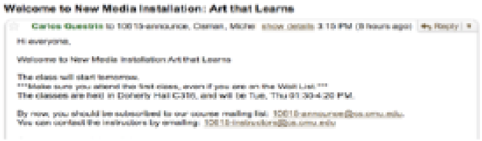


Osman Khan to Carlos show details Jan 7 (6 days ago) Reply

sounds good
+ok

Carlos Guestrin wrote:
Let's try to chat on Friday a little to coordinate and more on Sunday in person?

Carlos



Welcome to New Media Installation: Art that Learns

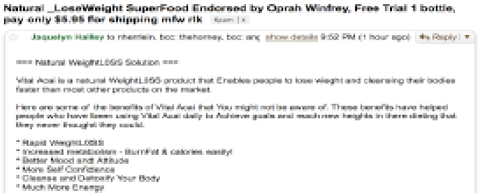
Carlos Guestrin to 10615-announce, Osman, Michel show details 3:15 PM (8 hours ago) Reply

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.
You can contact the instructors by emailing: 10615-instructors@cs.cmu.edu



Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | X

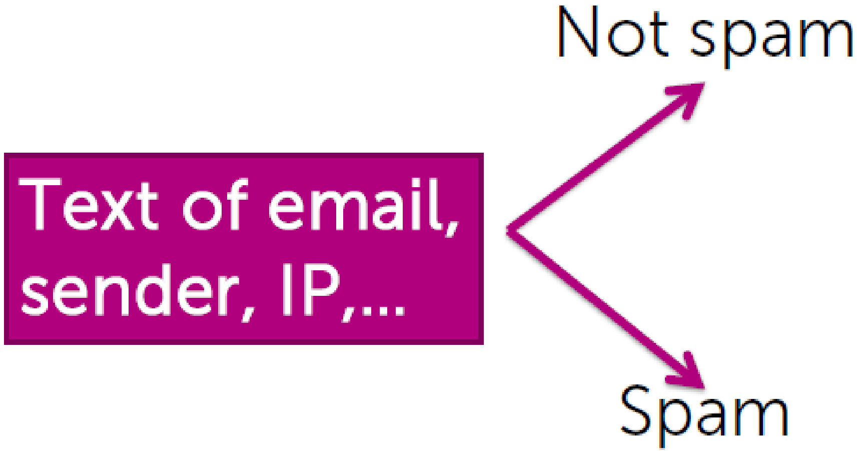
Jaquelyn Halley to nherrlein, bcc: thehorney, bcc: ang show details 9:52 PM (1 hour ago) Reply

=== Natural WeightLOSS Solution ===

Vital Acai is a natural WeightLOSS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightLOSS
* Increased metabolism - BurnFat & calories easily!
* Better Mood and Attitude
* More Self Confidence
* Cleanse and Detoxify Your Body
* Much More Energy



Text of email, sender, IP,…

Not spam

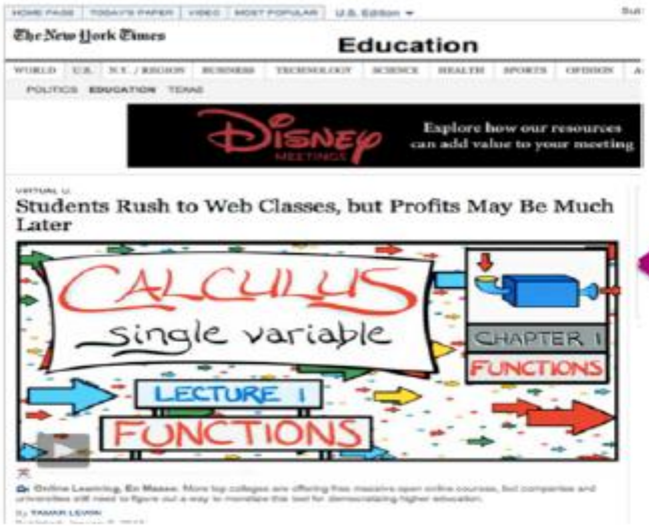
Spam

Input: x

Output: y

Multiclass classifier

Output y has more than 2 categories



HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition

The New York Times

Education

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

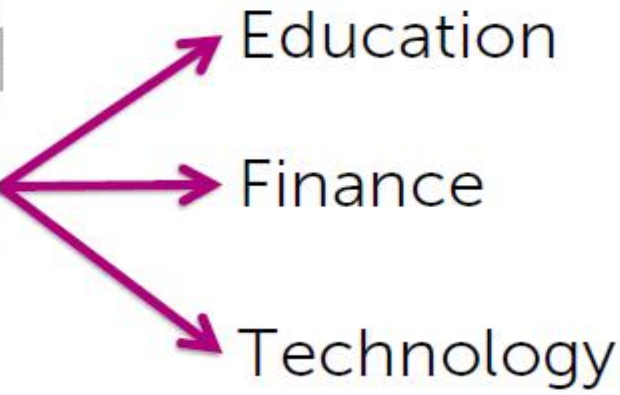
POLITICS EDUCATION TEXAS

VIRTUAL U.

Students Rush to Web Classes, but Profits May Be Much Later

Online Learning, En Masse: More top colleges are offering free massive open online courses, but companies and universities still need to figure out a way to monetize this tool for democratizing higher education.

By TAMAR LEWIN



Education

Finance

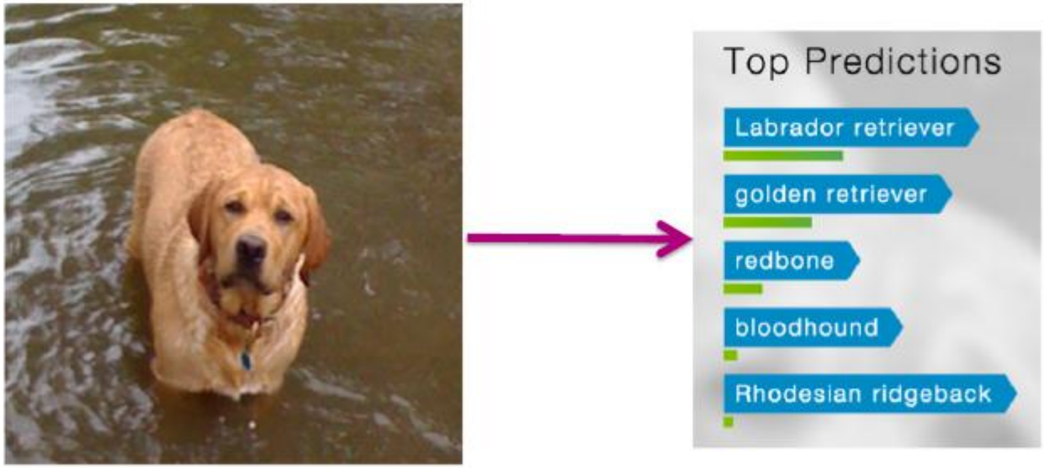
Technology

Input: x
Webpage

Output: y

Classification

Image classification



Top Predictions

- Labrador retriever
- golden retriever
- redbone
- bloodhound
- Rhodesian ridgeback

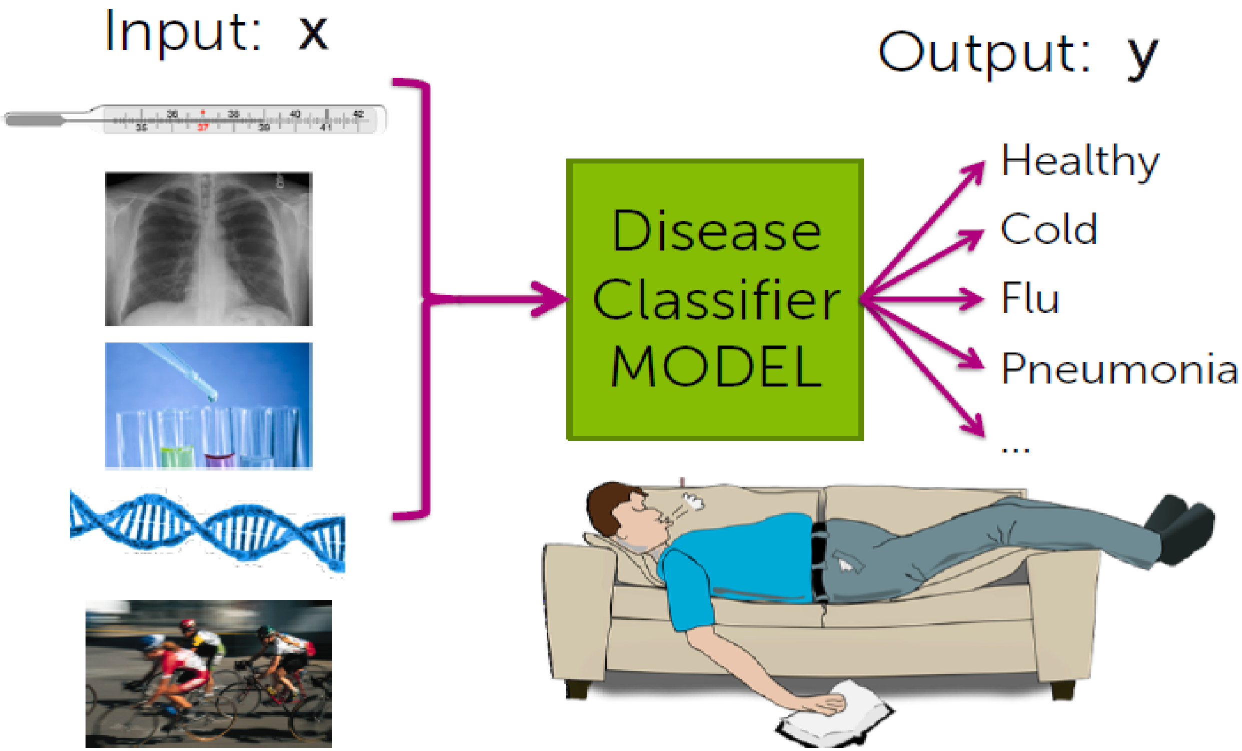
soft prediction

Input: x
Image pixels

Output: y
Predicted object

Classification

Personalized medical diagnosis



Input: x

Output: y

Disease Classifier MODEL

Healthy

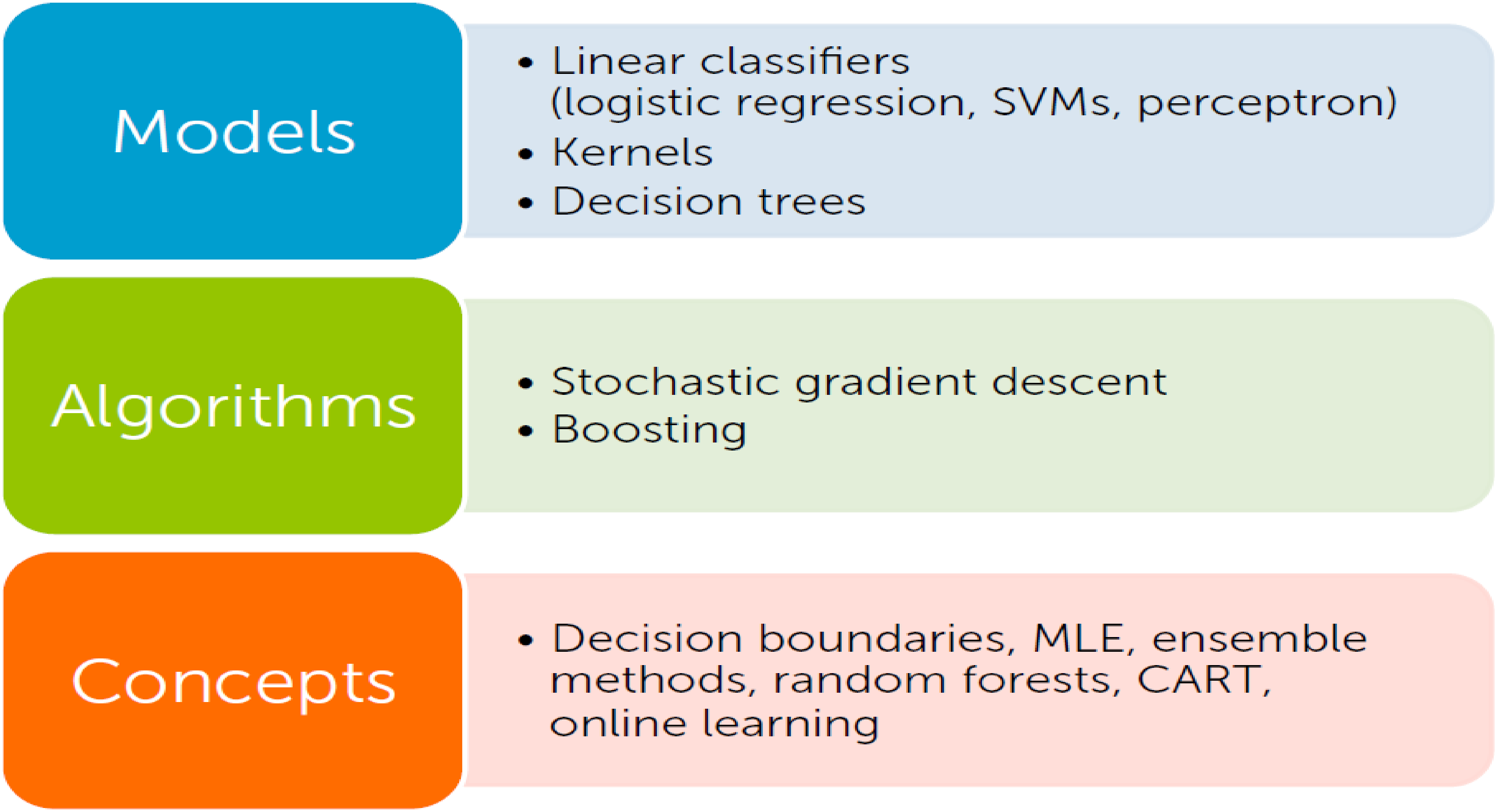
Cold

Flu

Pneumonia

...

Classification:



Models

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

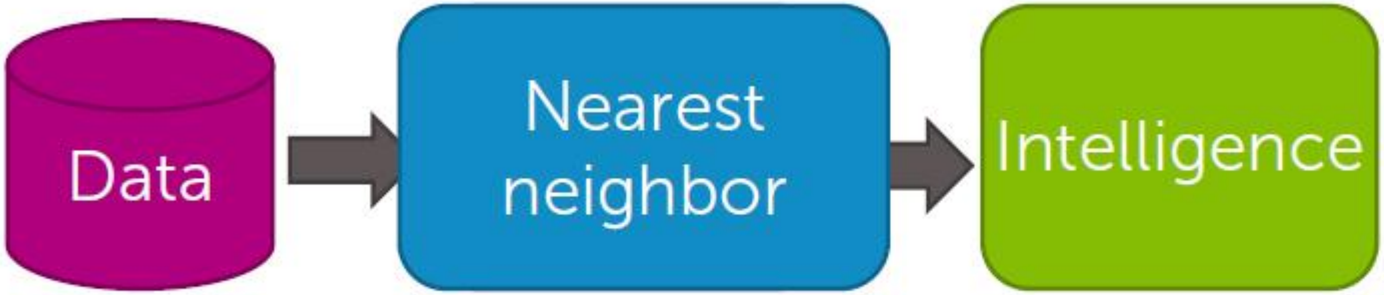
Algorithms

- Stochastic gradient descent
- Boosting

Concepts

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

Case Study: document retrieval



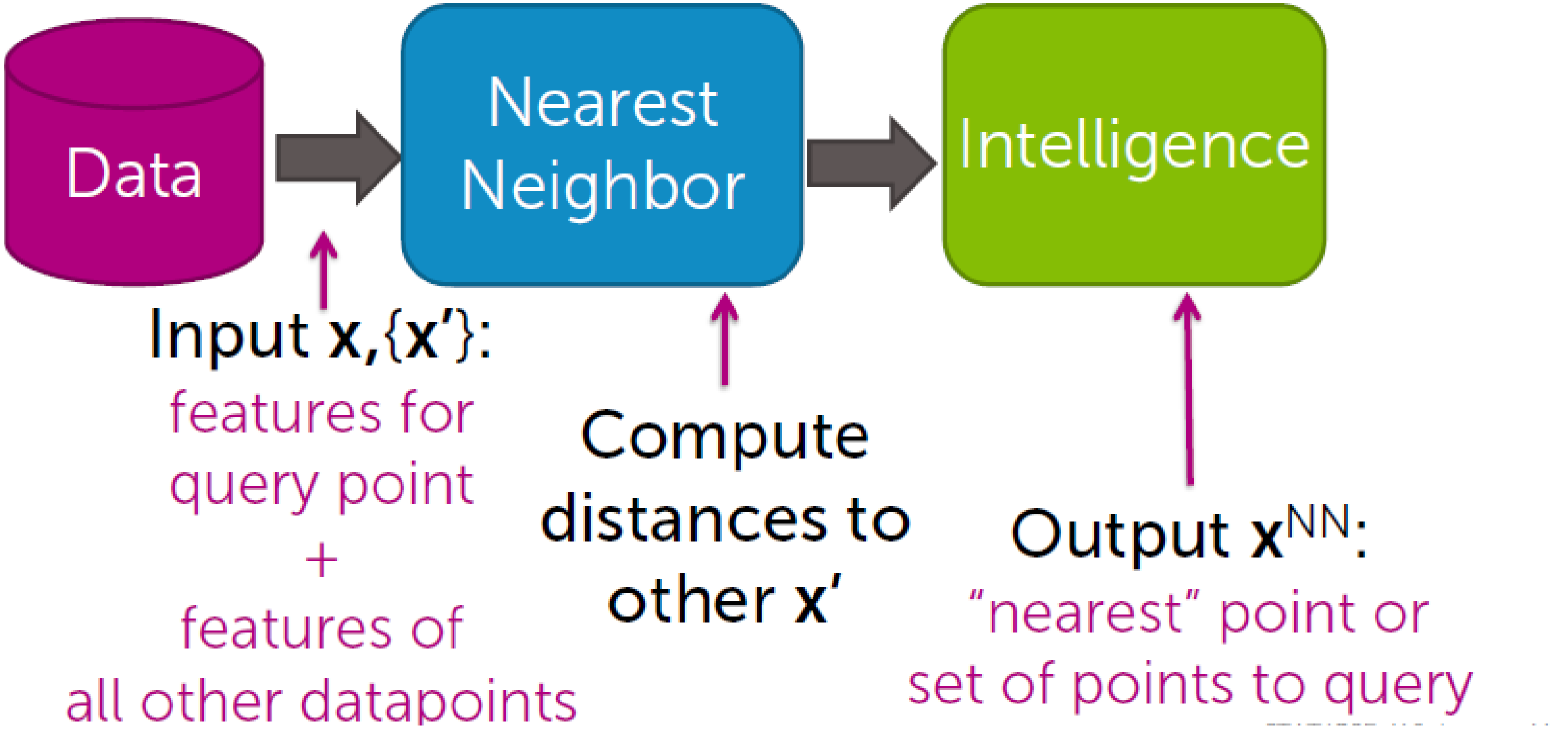
Data

Nearest neighbor

Intelligence

Retrieval

Search for related items



Data

Nearest Neighbor

Intelligence

Input $\mathbf{x}$,{$\mathbf{x}'$}:

features for query point + features of all other datapoints

Compute distances to other $\mathbf{x}'$

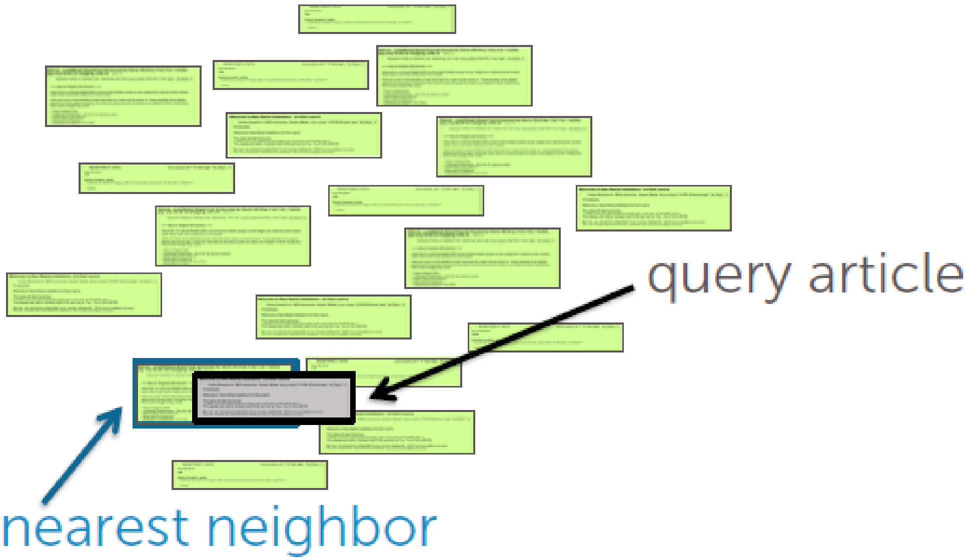
Output $\mathbf{x}^{NN}$:

"nearest" point or set of points to query

Retrieval

Retrieve “nearest neighbor” article

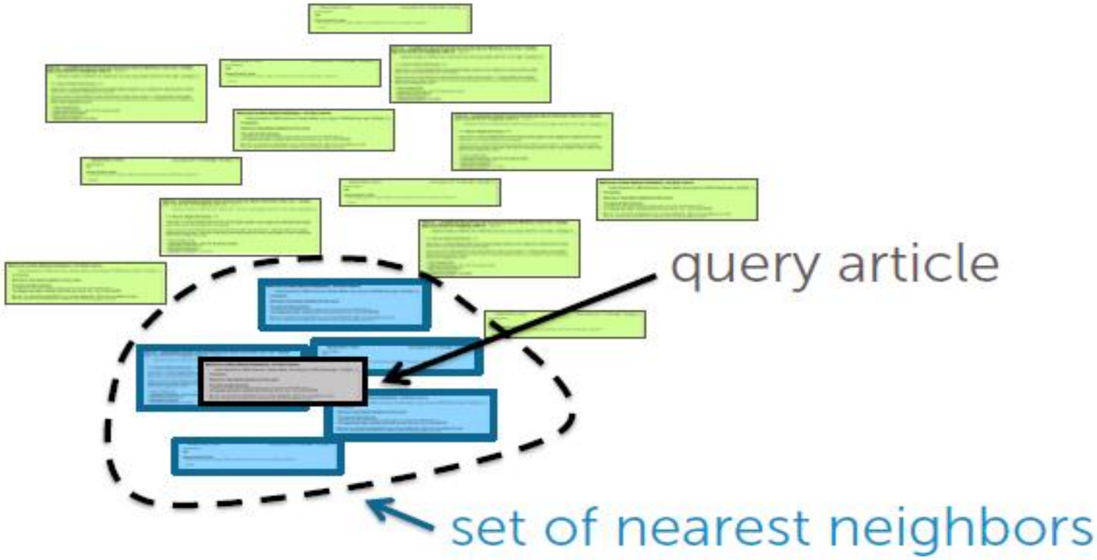
Space of all articles,
organized by similarity of text



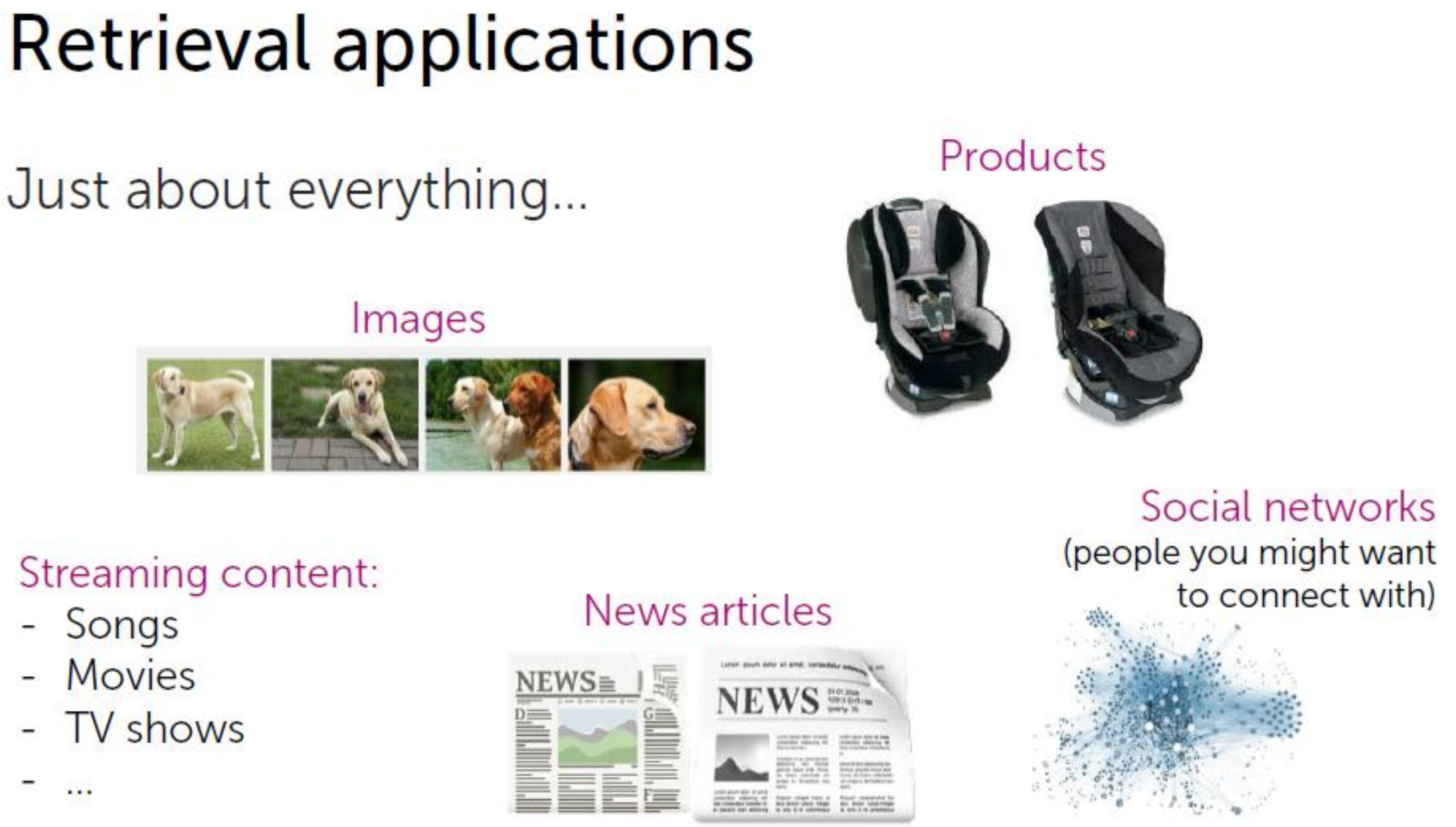
Retrieval

Or set of nearest neighbors

Space of all articles,
organized by similarity of text



Retrieval



Retrieval applications

Just about everything…

Images

Products

Social networks (people you might want to connect with)

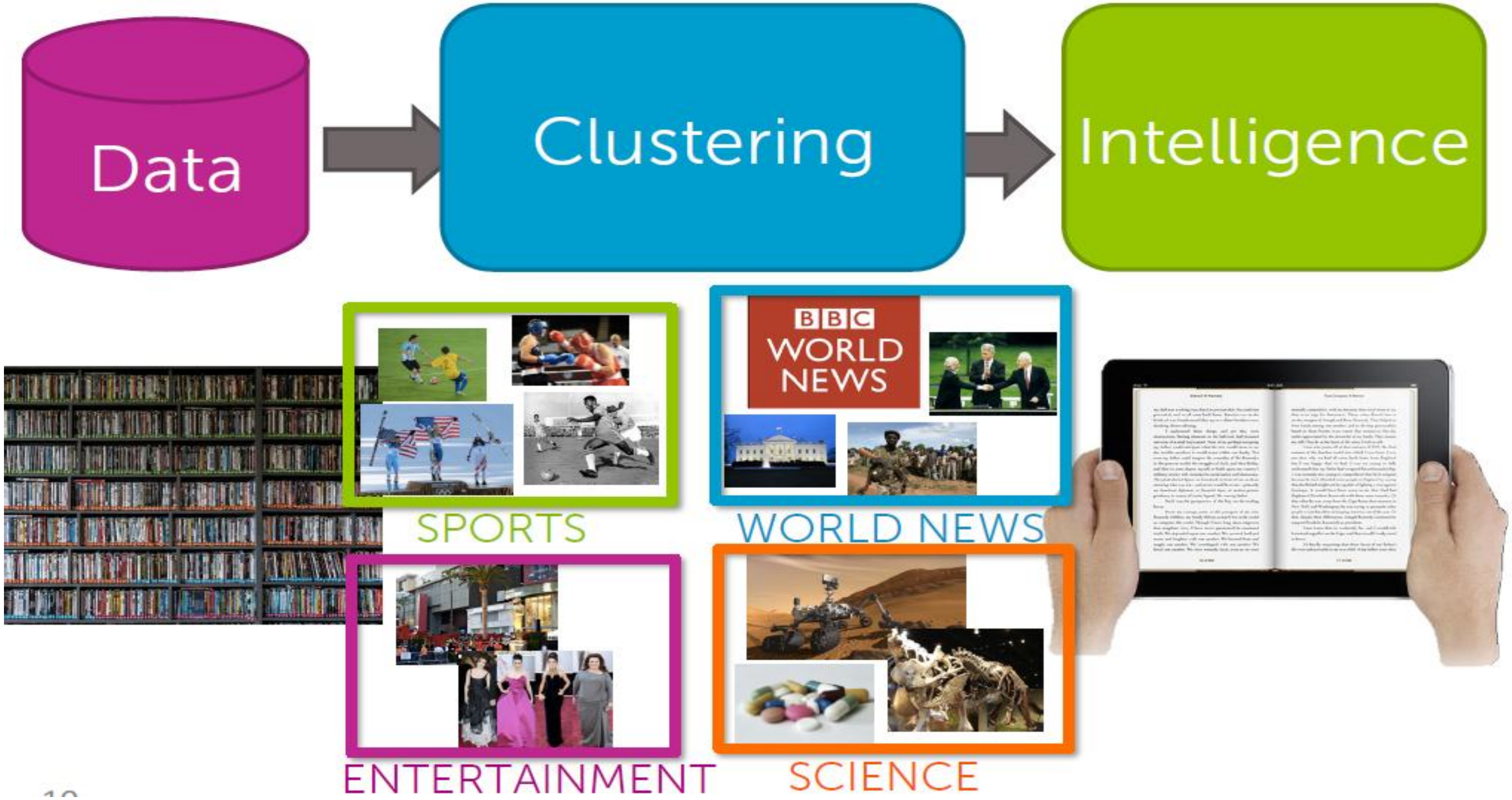
Streaming content:
- Songs
- Movies
- TV shows
- …

News articles

Case study 3++:

Document structuring for retrieval

ML method

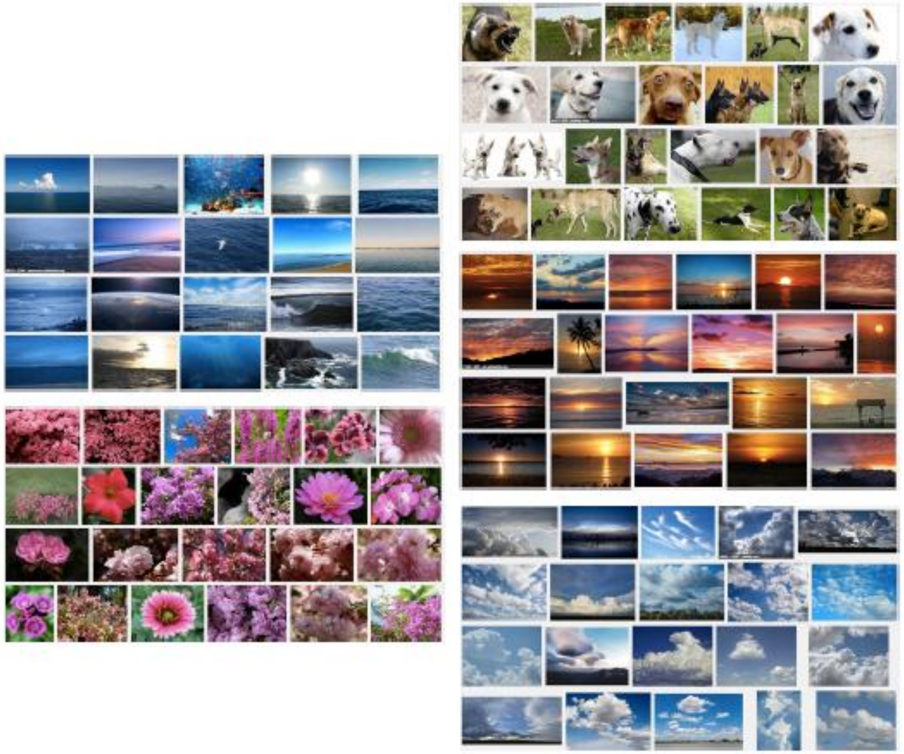


Clustering

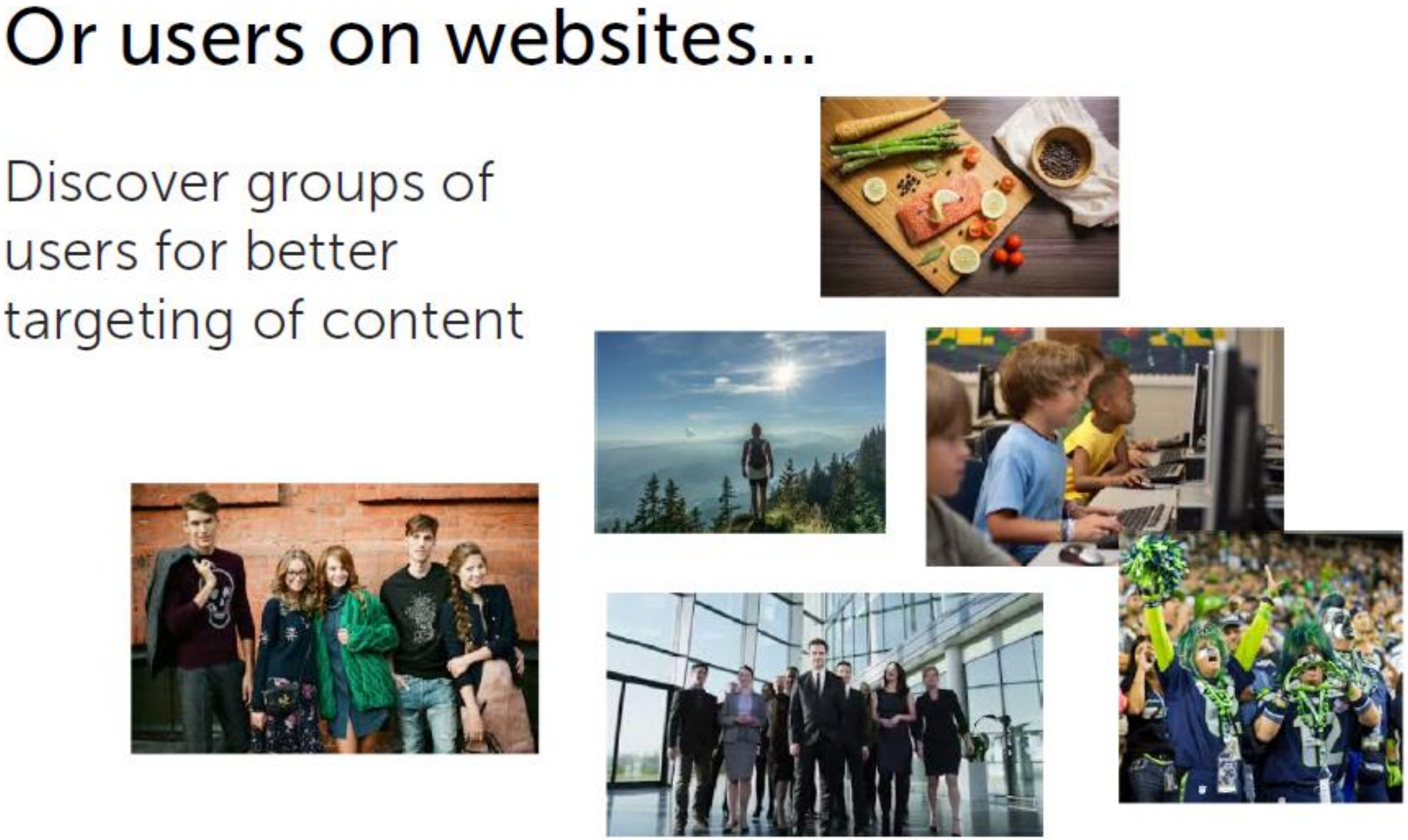
Clustering images

For search, group as:

- – Ocean
- – Pink flower
- – Dog
- – Sunset
- – Clouds
- – ...



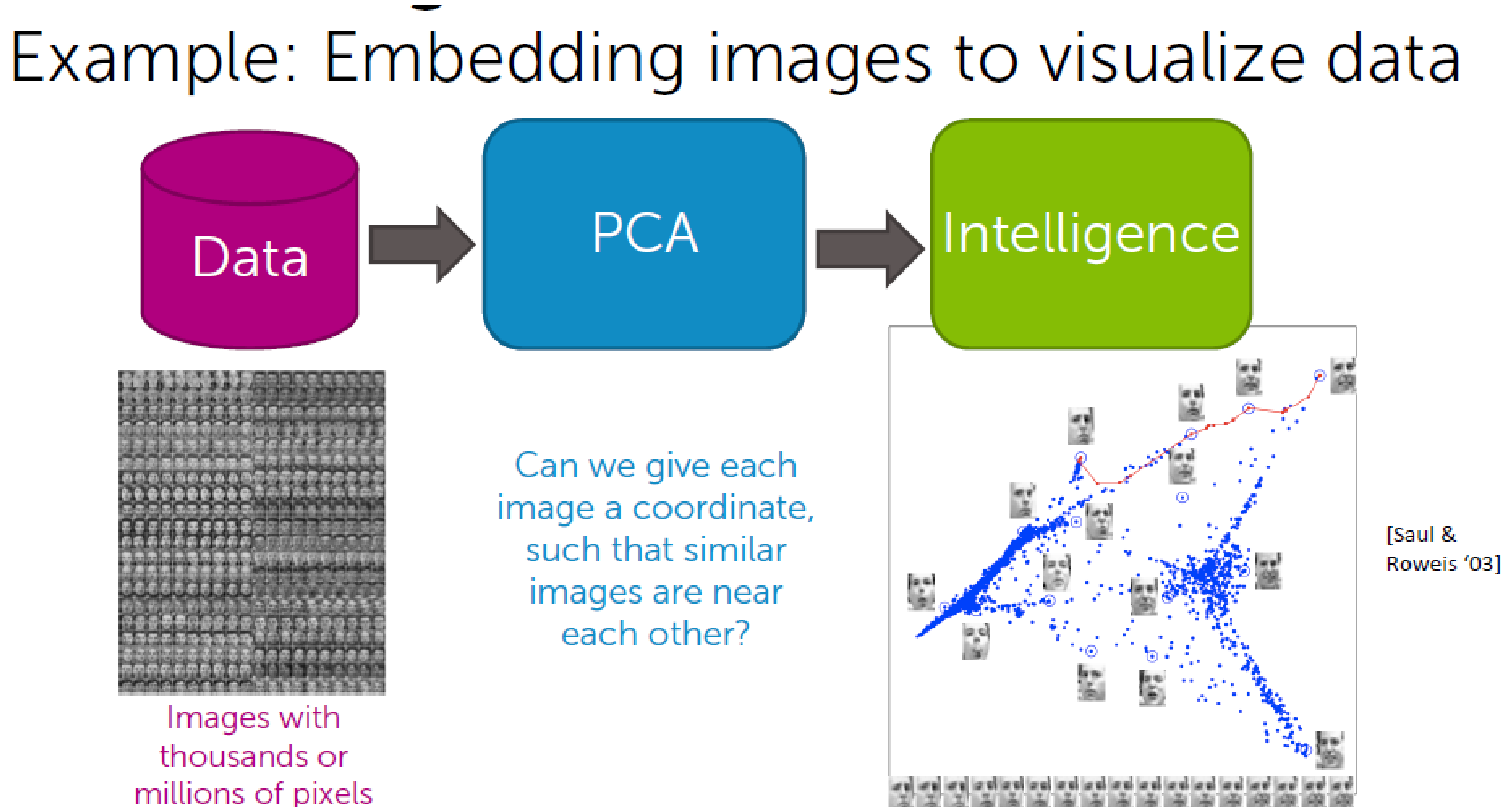
Clustering



Or users on websites...

Discover groups of users for better targeting of content

Embeding



Example: Embedding images to visualize data

Data

PCA

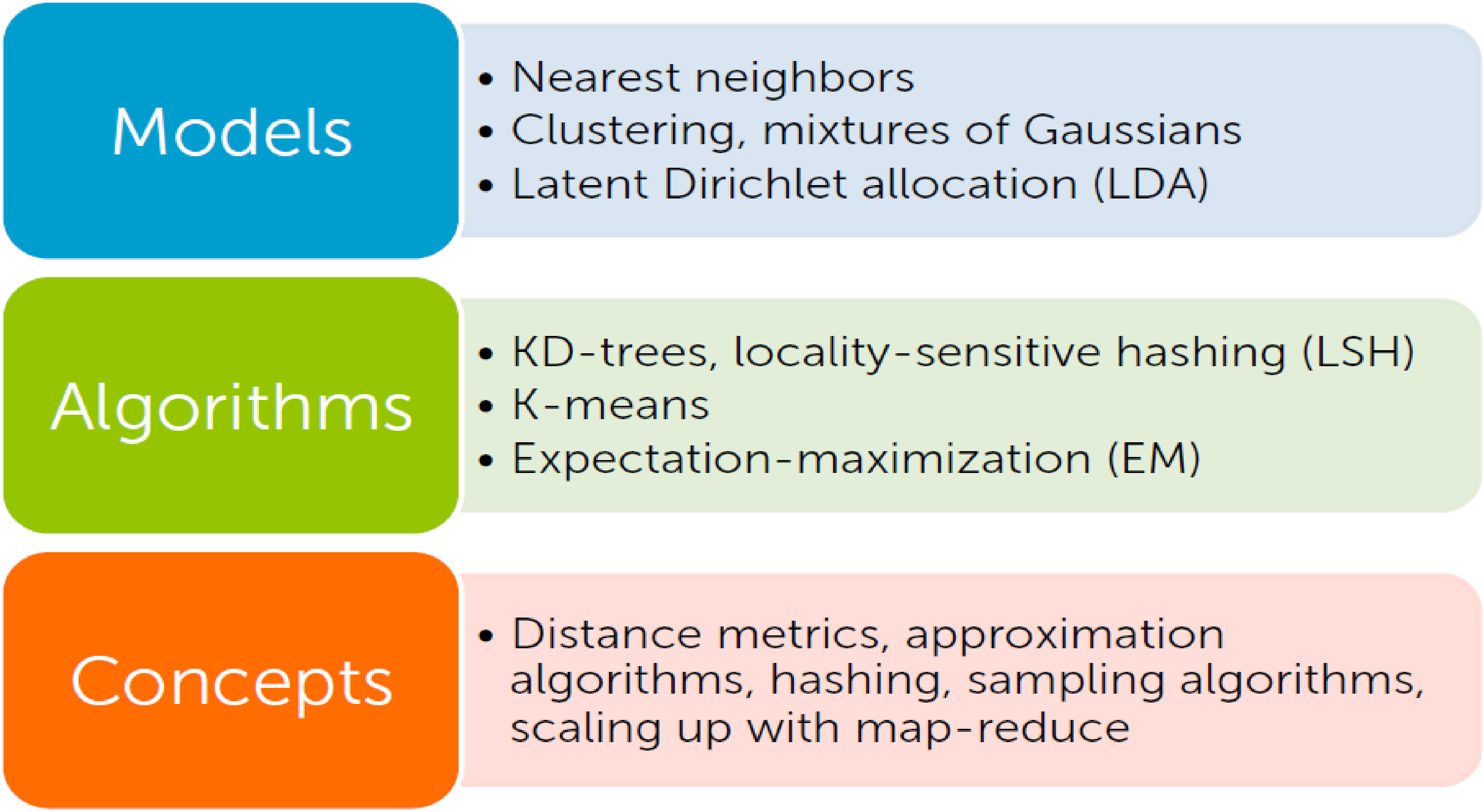
Intelligence

Images with thousands or millions of pixels

Can we give each image a coordinate, such that similar images are near each other?

[Saul & Roweis '03]

Clustering: Finding documents



Models

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

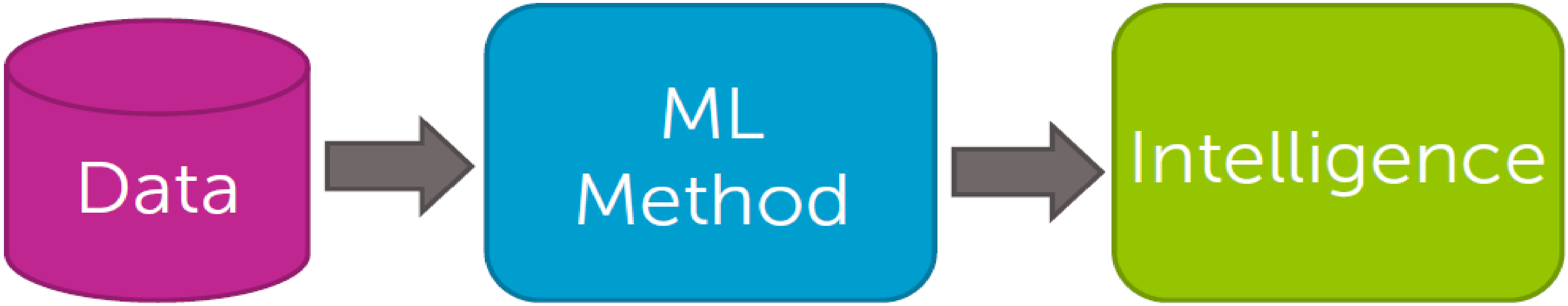
Algorithms

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

Concepts

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

Case study: Product recommendation



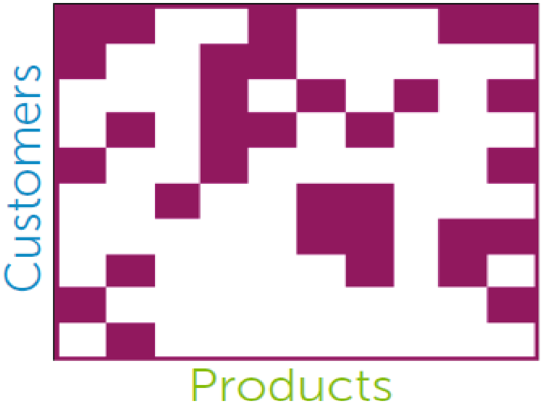
Data

ML Method

Intelligence

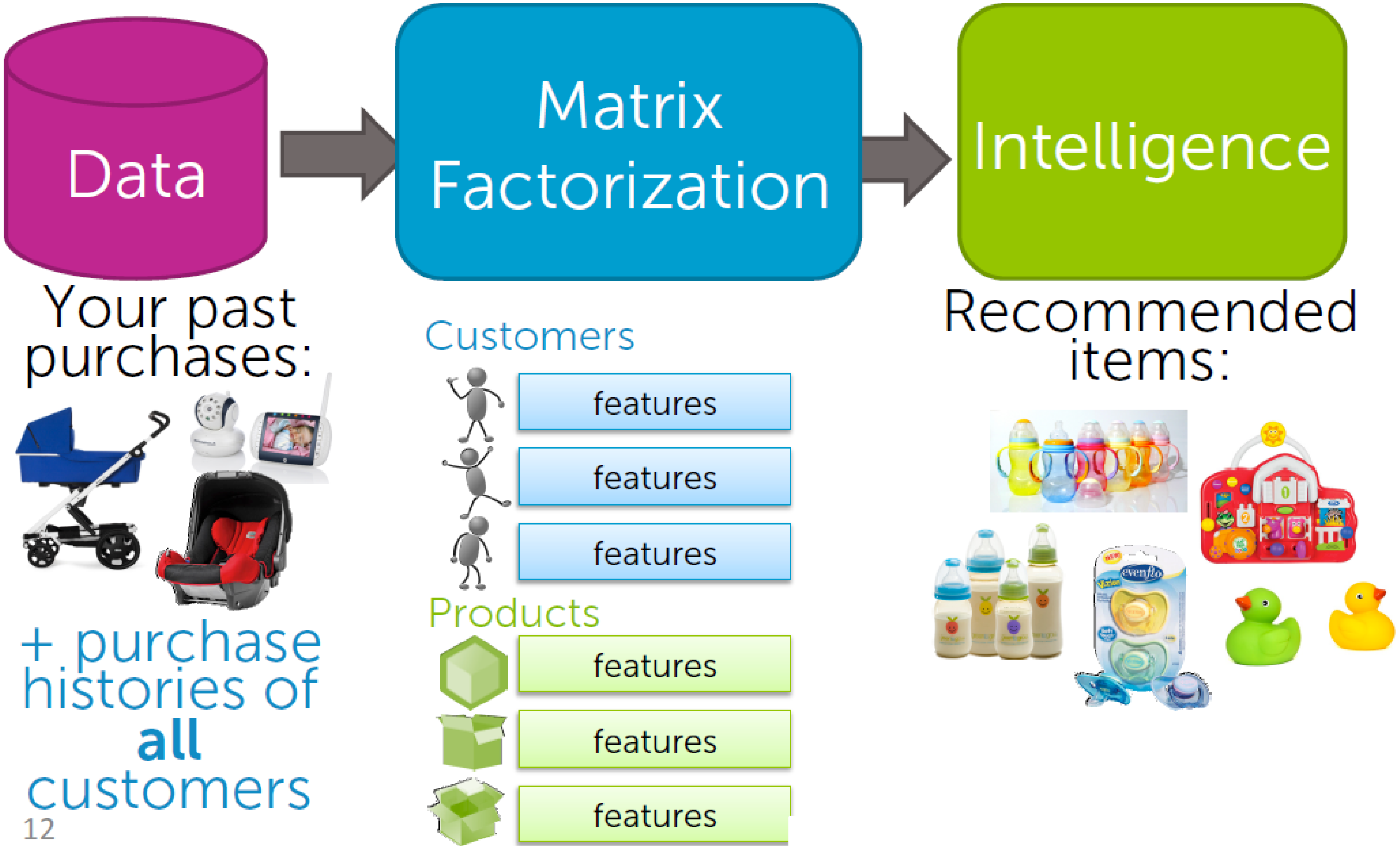
Your past purchases:

+ purchase histories of **all** customers



Recommended items:

Case study: Product recommendation



Data → Matrix Factorization → Intelligence

Your past purchases:

+ purchase histories of **all** customers

Customers

features

features

features

Products

features

features

features

Recommended items:

12

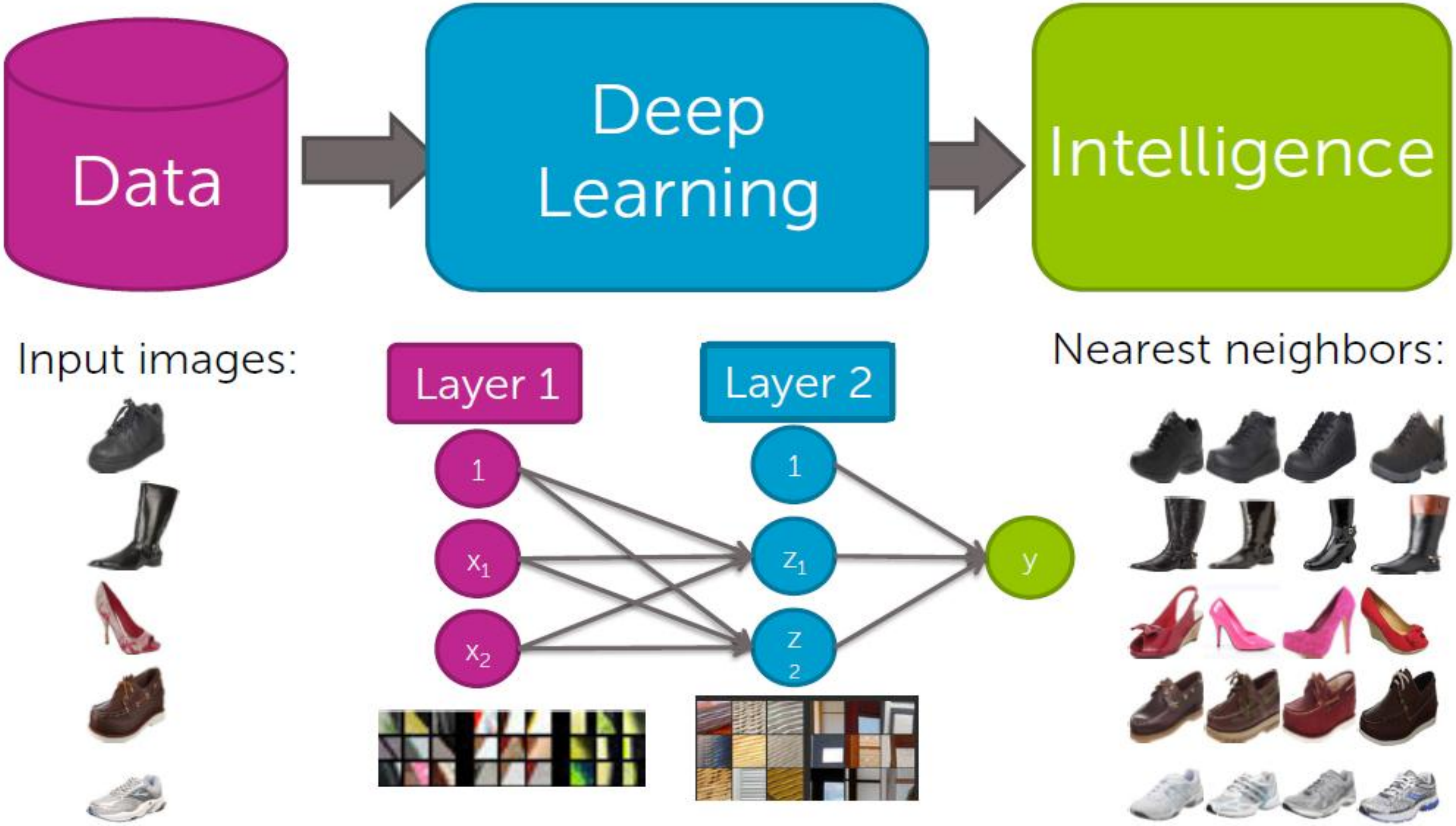
Recomender systems applications

Movies

Songs

Friends, apps, ...

Case study 5:
Visual product recommender



Data → Deep Learning → Intelligence

Input images:

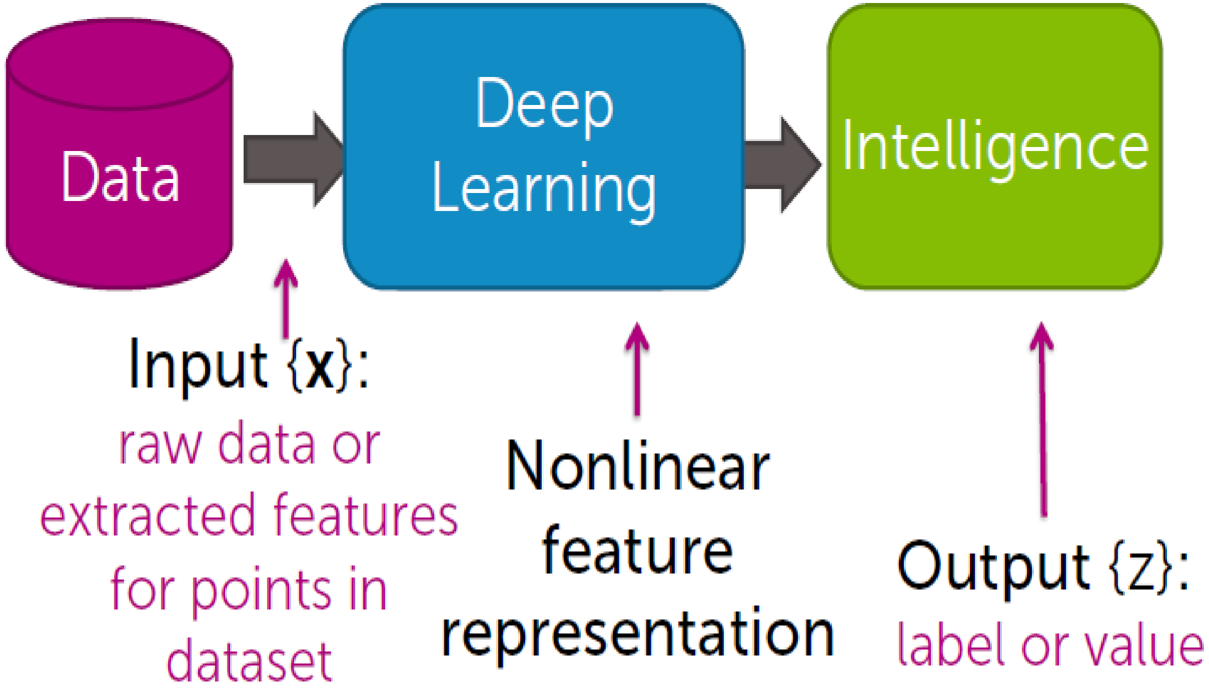
Layer 1

Layer 2

Nearest neighbors:

What is (supervised) deep learning?

Flexible method for performing classification or regression



Data

Deep Learning

Intelligence

Input {x}:

raw data or extracted features for points in dataset

Nonlinear feature representation

Output {z}:

label or value

Examples of deep learning success stories

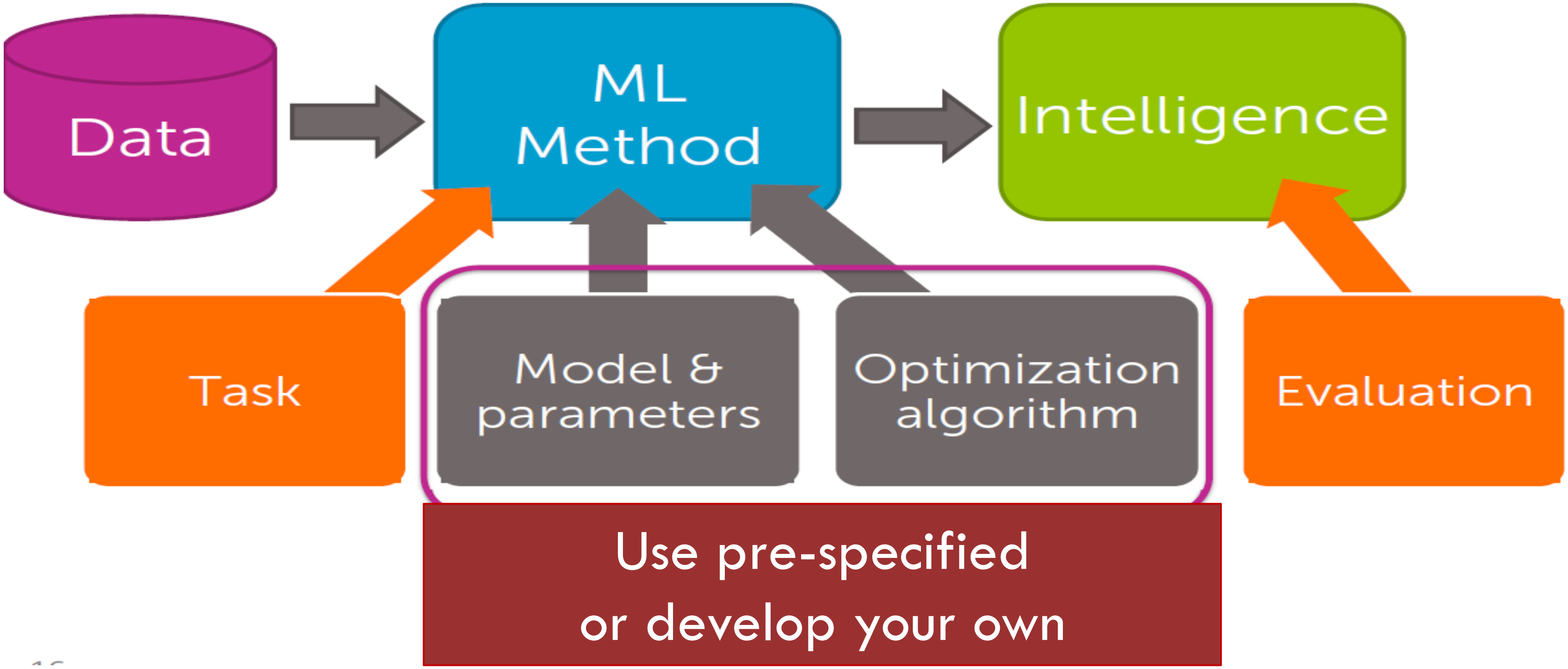
- Image classification
- Image segmentation
- Image captioning
- Object detection
- Speech recognition
- Speech synthesis
- Machine translation
- Handwriting recognition
- …

Other ML methods

- Reinforcement learning
- Learning theory
- Active learning
- Multi-task and transfer learning
- Spectral methods
- …

Deploing inteligence module

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



Use pre-specified
or develop your own