

TEORETYCZNE PODSTAWY INFORMATYKI

23/01/2017

WFAiS UJ, Informatyka Stosowana
I rok studiów, I stopień

Wykład 16

2

Data Science:
Uczenie
maszynowe

- Uczenie maszynowe: co to znaczy?
- Metody
 - ▣ Regresja
 - ▣ Klasyfikacja
 - ▣ Klastering i wybor podzbioru
 - ▣ System do rekomendacji

Wykład na podstawie materiałów:

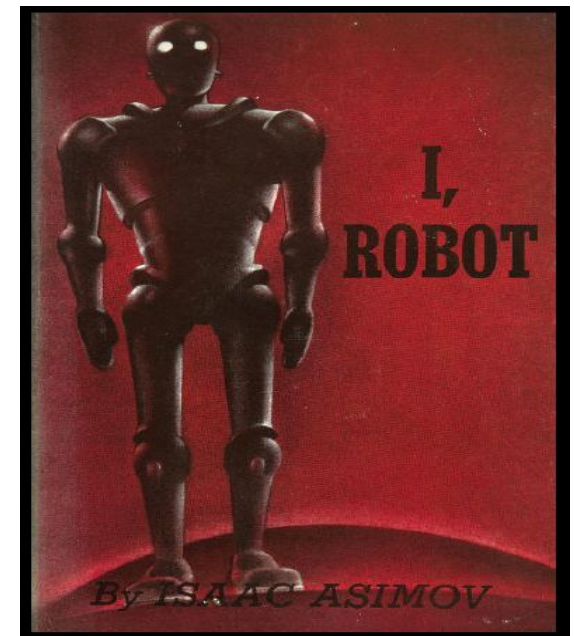
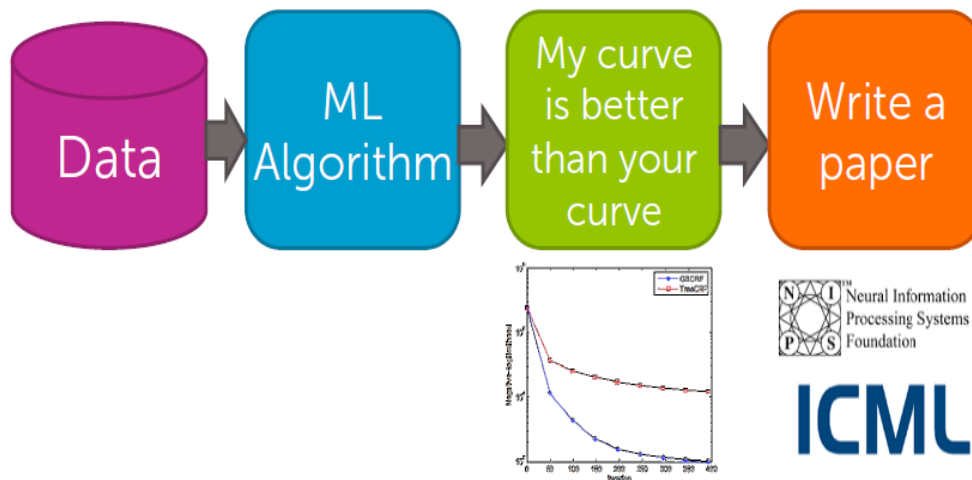
**E. Fox & C. Guestrin, Univeristy of Waschington,
„Machine Learning Specialization”**

<http://www.coursera.org/learn/ml-foundations>

Uczenie maszynowe

3

- **Metody uczenia maszynowego (machine learning) rewolucjonizują obecnie podejście do różnych problemów związanych z analizą danych.**
- **Jeszcze kilka lat temu był to bardziej „akademicki” problem z zakresu numeryki i algorytmiki.**



Uczenie maszynowe

4

- Obecnie, bardzo szybko staje się kluczową techniką dla wielu wiodących firm komercyjnych



Uczenie maszynowe: cel

5

- Uzyskać odpowiedź dla różnej klasy pytań na podstawie informacji zawartej w danych, bez potrzeby budowy „modelu zjawiska”.

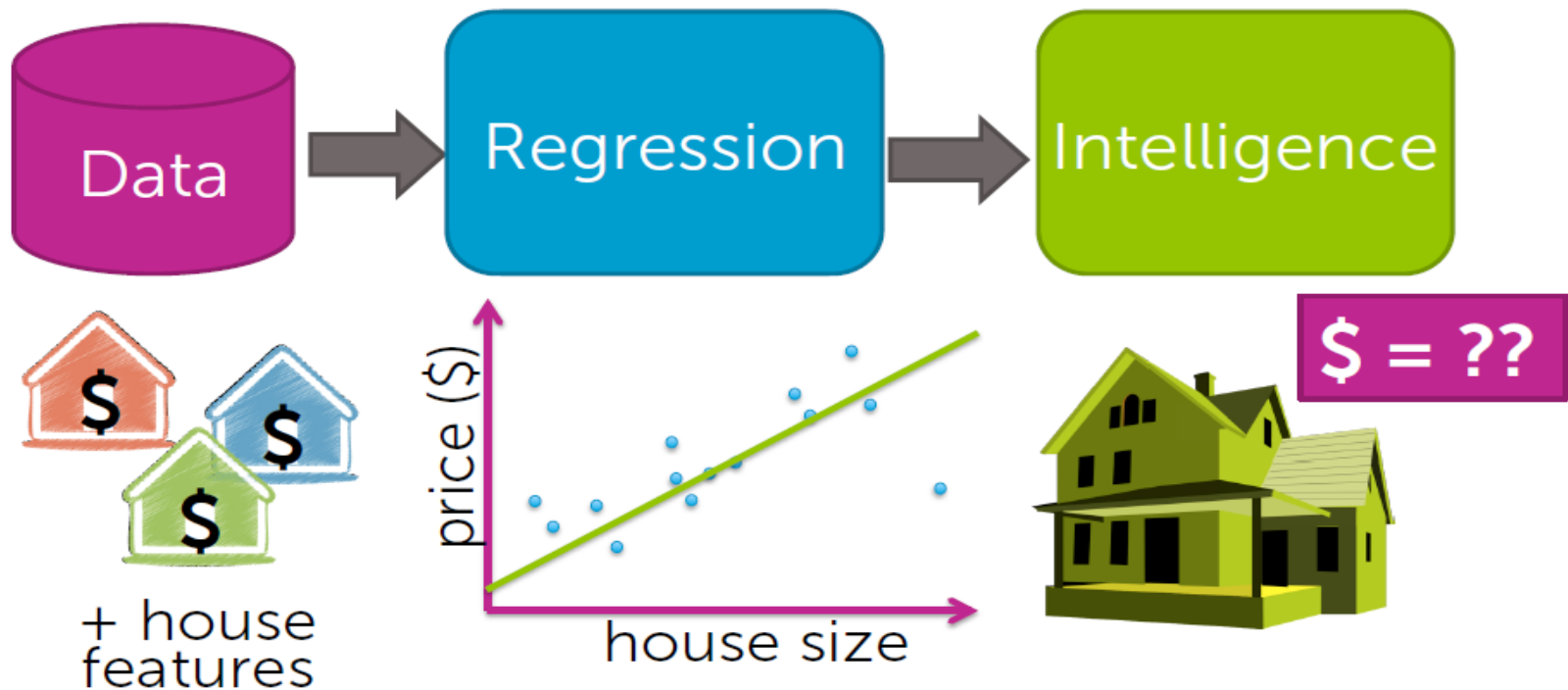


ML metod = np. klasyfikacja, regresja liniowa, sieci neuronowe

Uczenie maszynowe: przykład

6

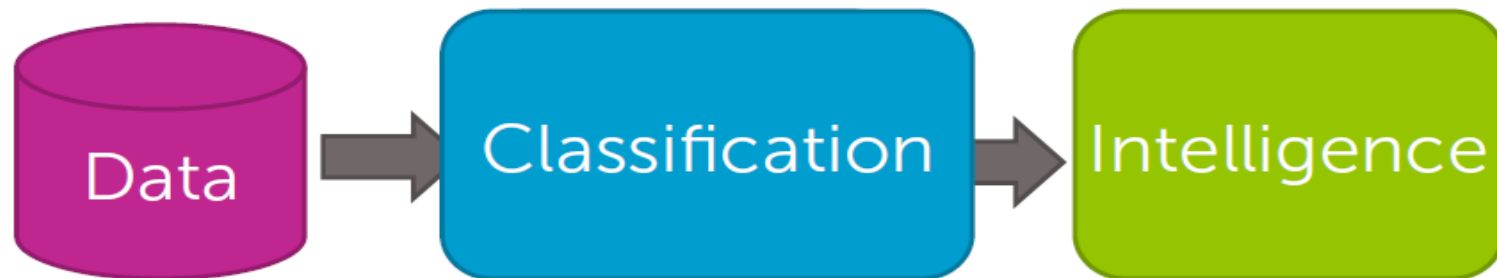
- Przewidywanie ceny domu na podstawie zebranych danych dotyczących ceny innych



Uczenie maszynowe: przykład

7

□ Ranking restauracji



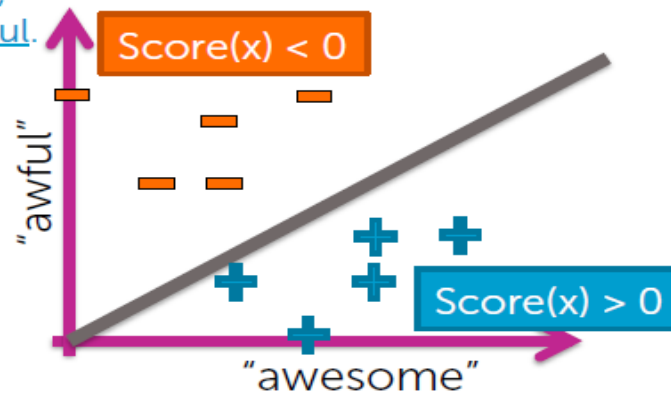
Sushi was awesome,
the food was awesome,
but the service was awful.

All reviews:

7/2/2015
This is probably my favorite place to eat Japanese in Seattle. My boyfriend and I ordered nigiri of scallop, Japanese wrapper (seasonal), and the agodashi 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.

8/11/2015
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.

8/9/2015
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 \$8-20 each and dishes are small.



Uczenie maszynowe: przykład

8

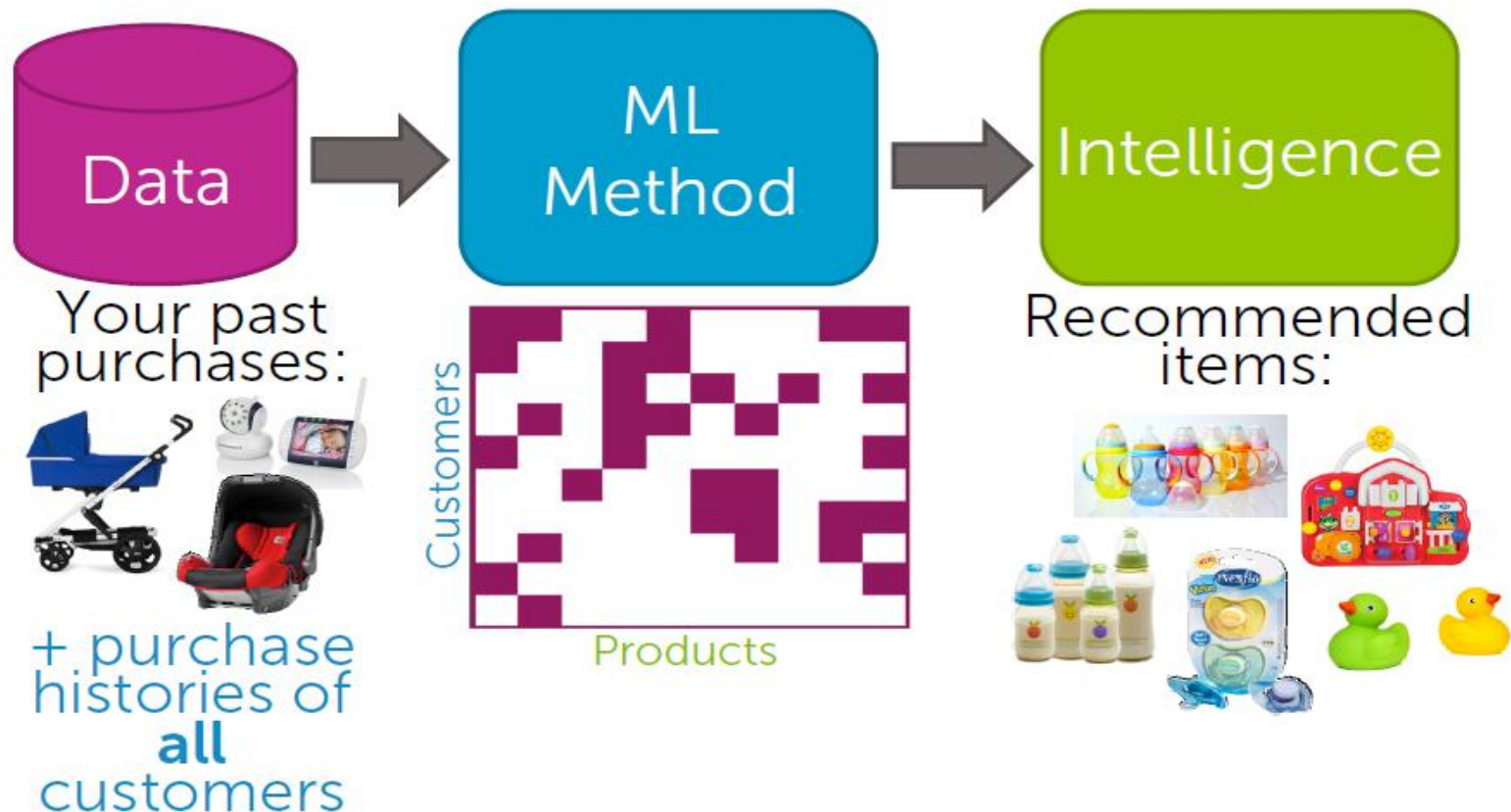
- Znajdowanie podobnych dokumentów



Uczenie maszynowe: przykład

9

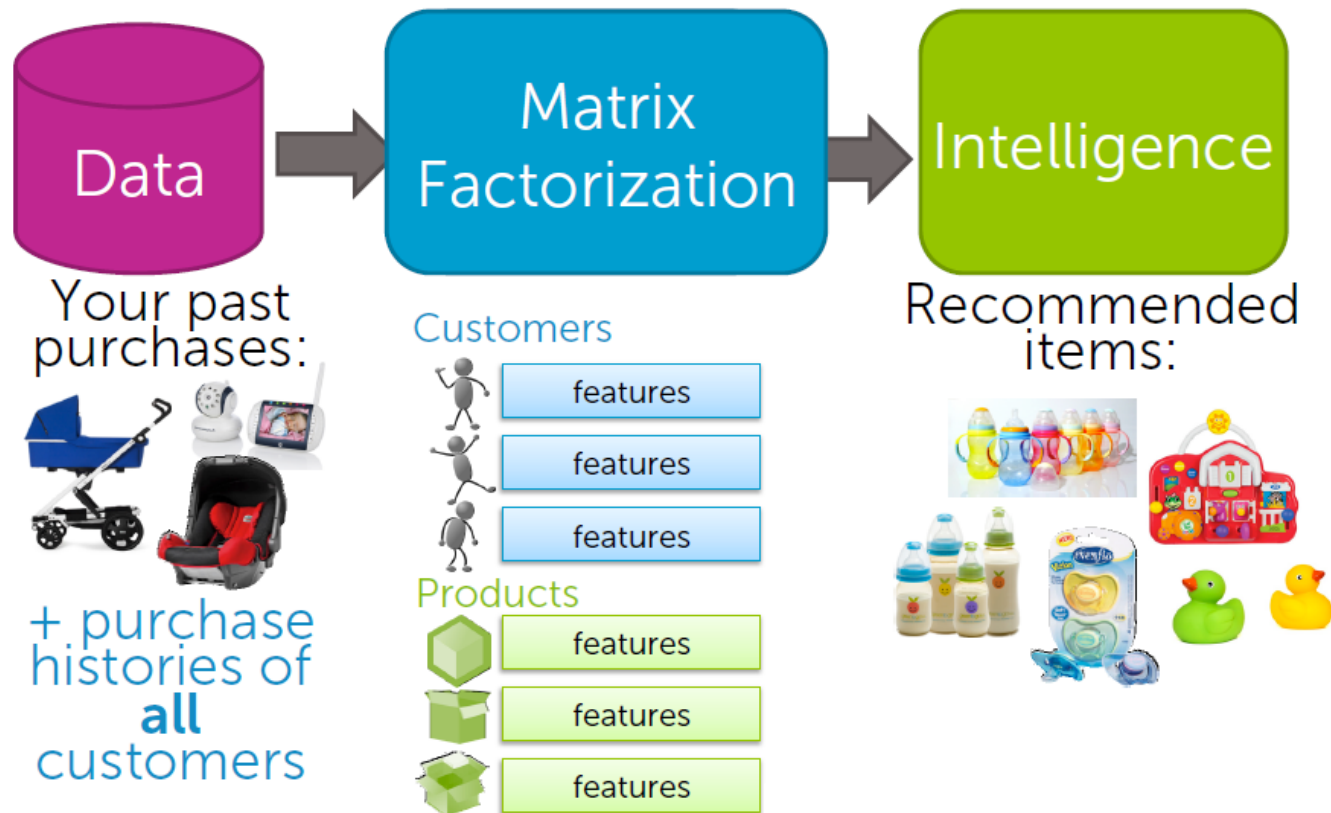
- Rekomendowanie podobnego produktu



Uczenie maszynowe: przykład

10

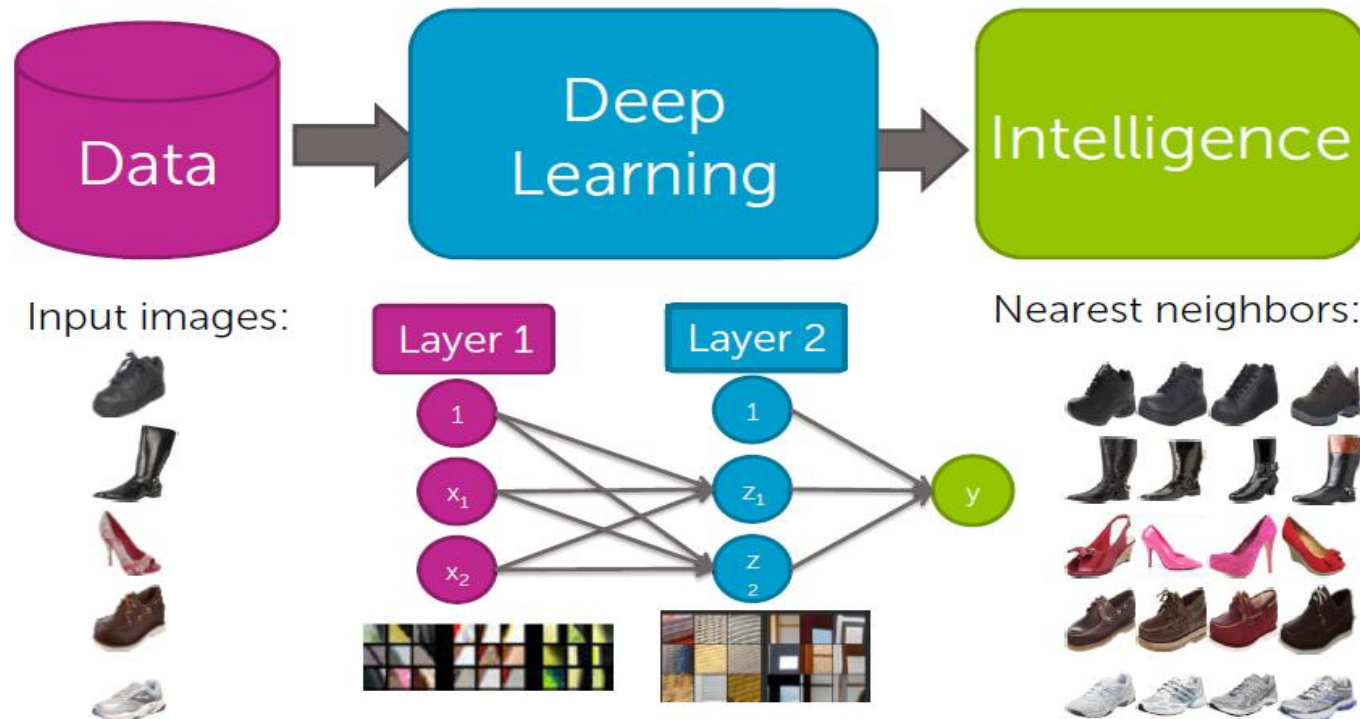
- **Rekomendowanie podobnego produktu na podstawie jego charakterystycznych cech**



Uczenie maszynowe: przykład

11

- **Rekomendowanie podobnego produktu: na podstawie jego charakterystyki graficznej**

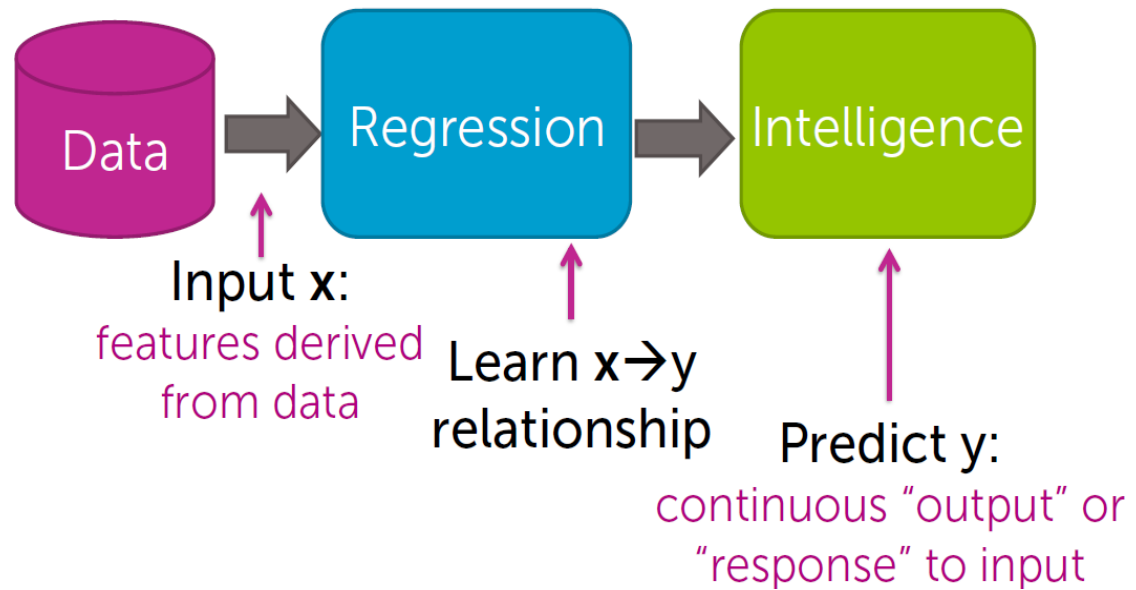


Regresja

12

Uczenie maszynowe:
Regresja

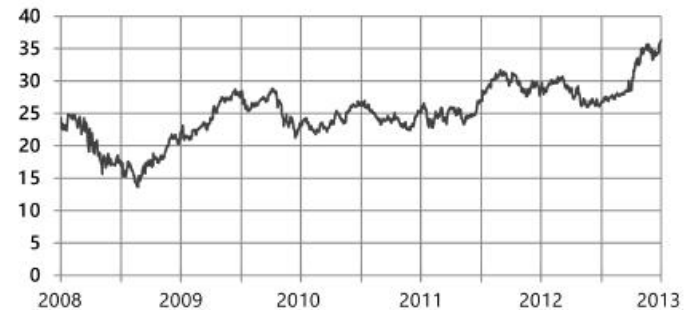
- Przewidywanie odpowiedzi na podstawie informacji wejściowej



Cena akcji na giełdzie

13

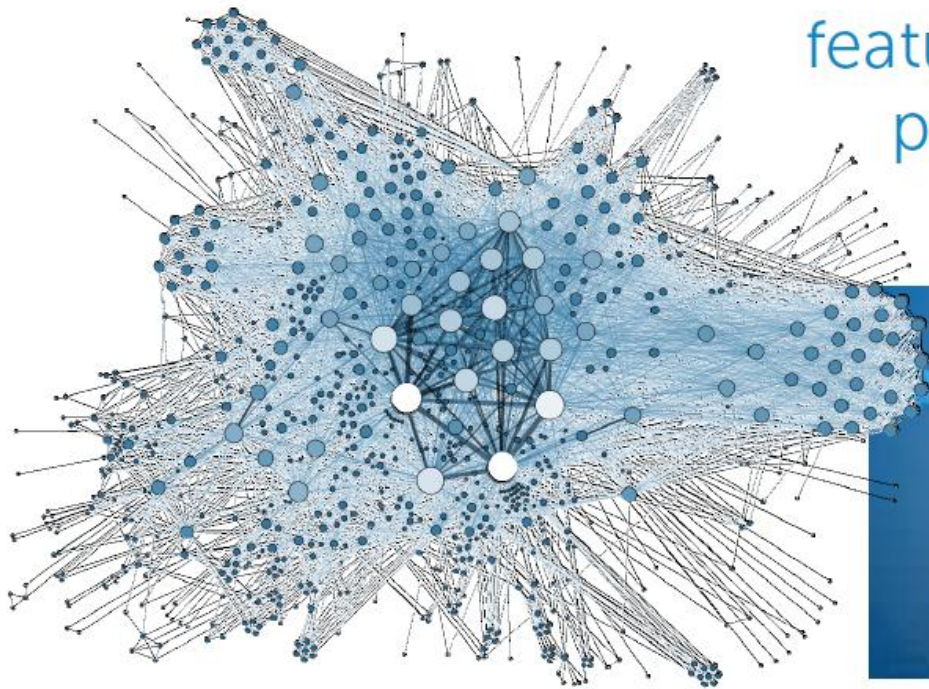
- Predict the price of a stock (y)
- Depends on $\mathbf{x} =$
 - Recent history of stock price
 - News events
 - Related commodities



Tweet popularność

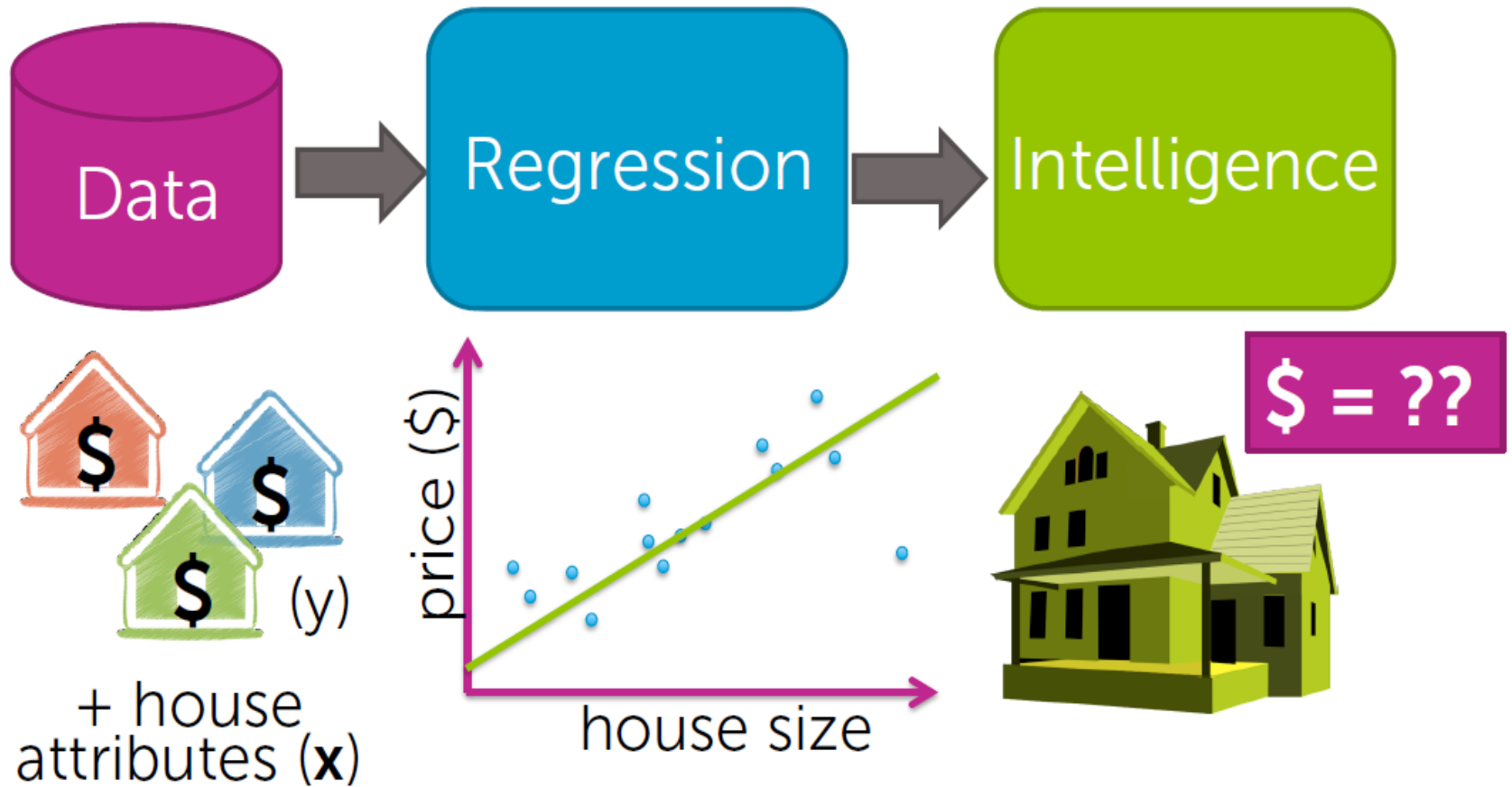
14

- How many people will retweet your tweet? (y)
- Depends on \mathbf{x} = # followers,
of followers of followers,
features of text tweeted,
popularity of hashtag,
of past retweets,...



Przykład: przewidywana cena domu

15



Data, model

16

Data

input output



$$(x_1 = \text{sq.ft.}, y_1 = \$)$$



$$(x_2 = \text{sq.ft.}, y_2 = \$)$$



$$(x_3 = \text{sq.ft.}, y_3 = \$)$$

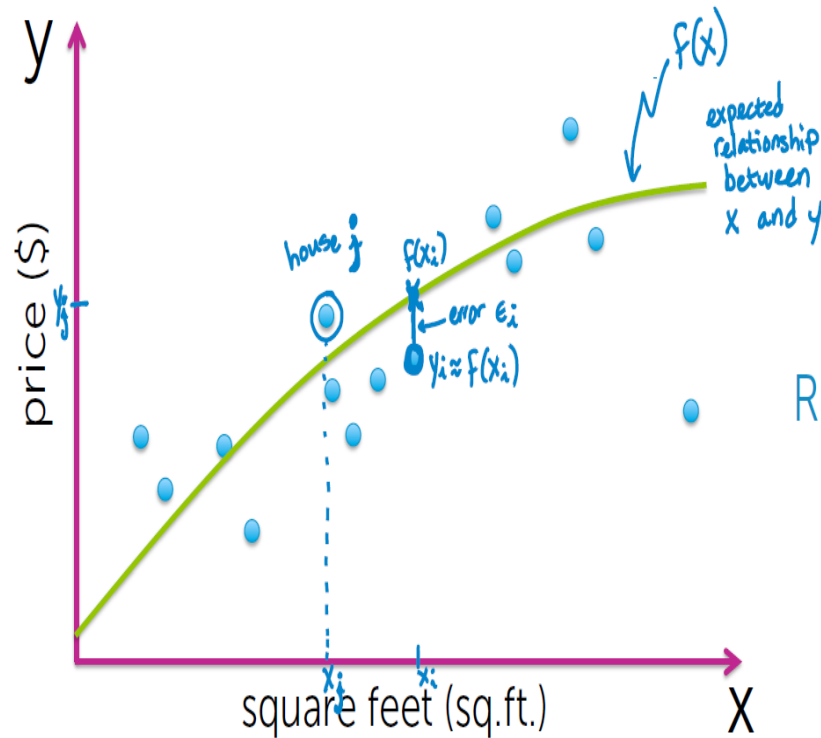


$$(x_4 = \text{sq.ft.}, y_4 = \$)$$



$$(x_5 = \text{sq.ft.}, y_5 = \$)$$

⋮



Regression model:

$$y_i = f(x_i) + \epsilon_i$$

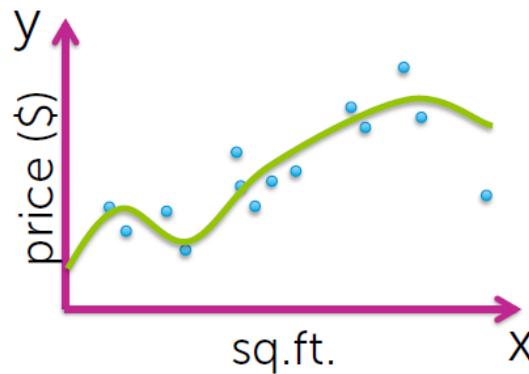
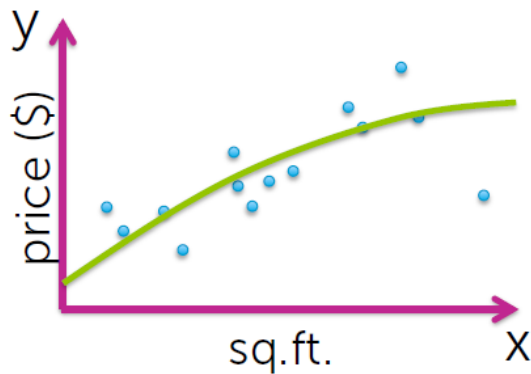
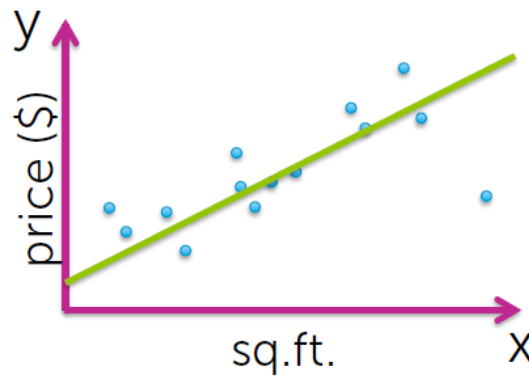
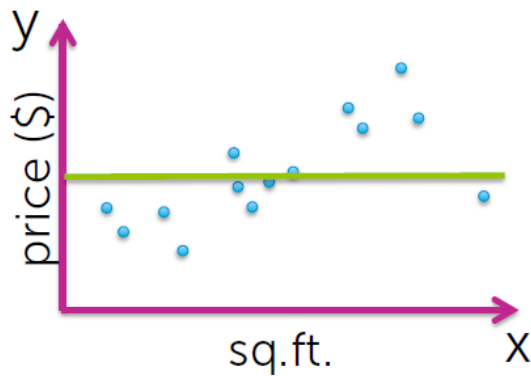
$E[\epsilon_i] = 0$ ← equally likely that error is + or -
↑ expected value

↓
 y_i is equally likely to be above or below $f(x_i)$

Data, model

17

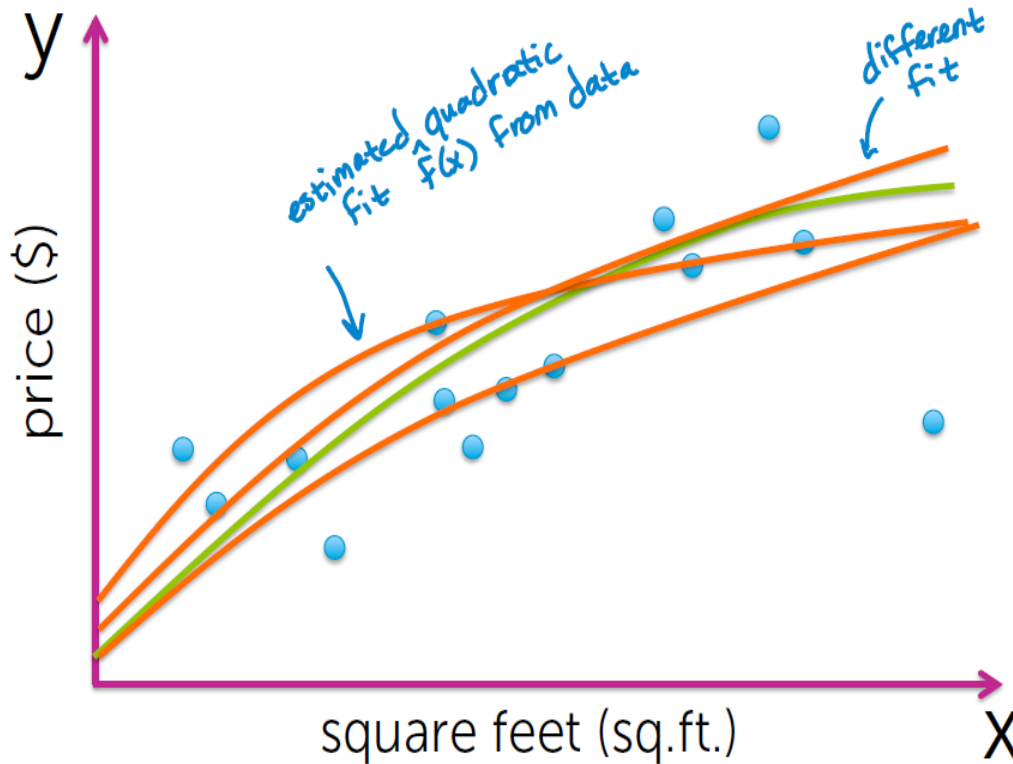
□ Jaki model dla $f(x)$?



Przewidywanie

18

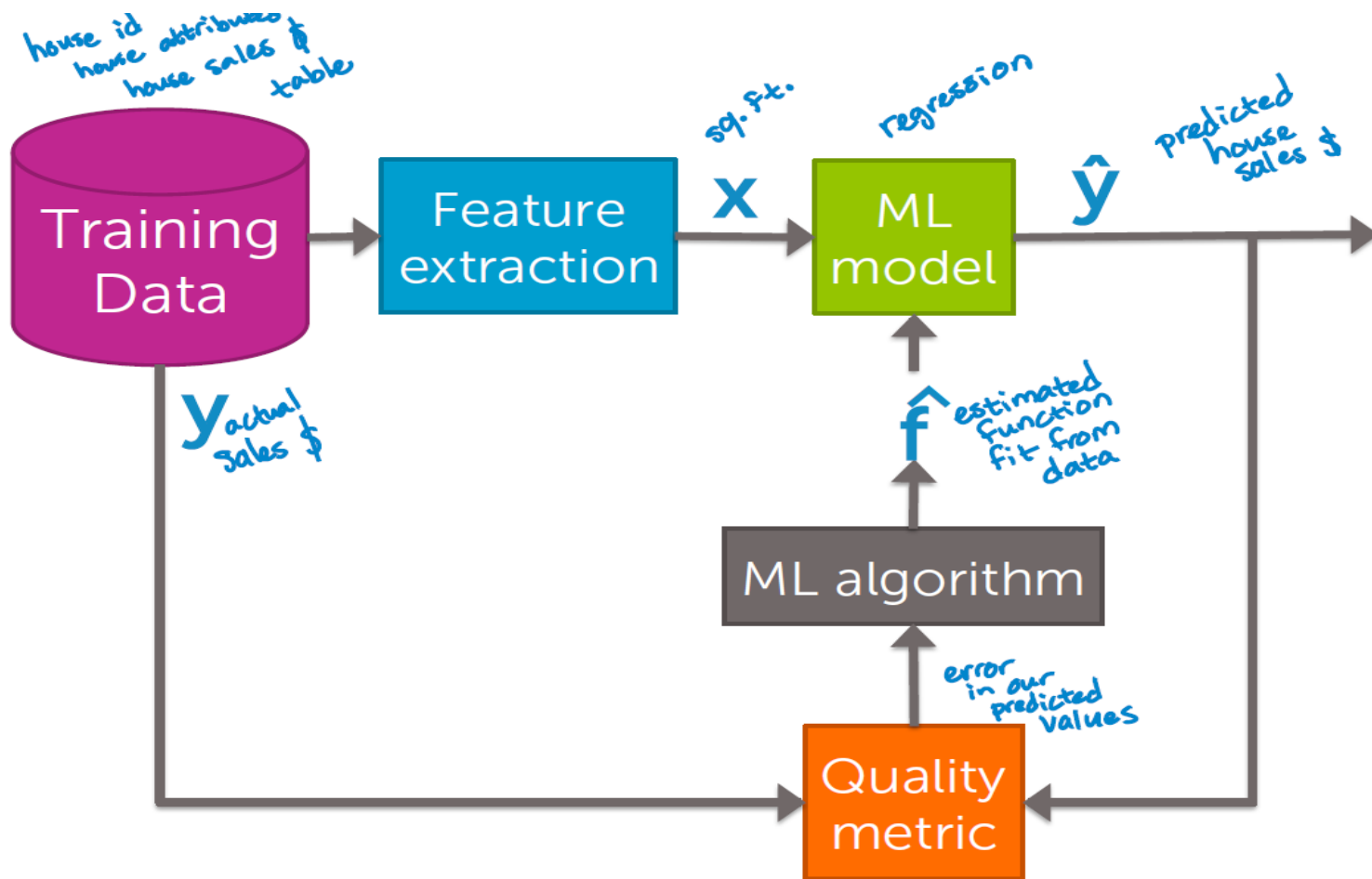
- A potem chcielibyśmy przewidzieć odpowiedź



Assume model $f(x)$ is a quadratic function

Pętla iteracyjna?

19

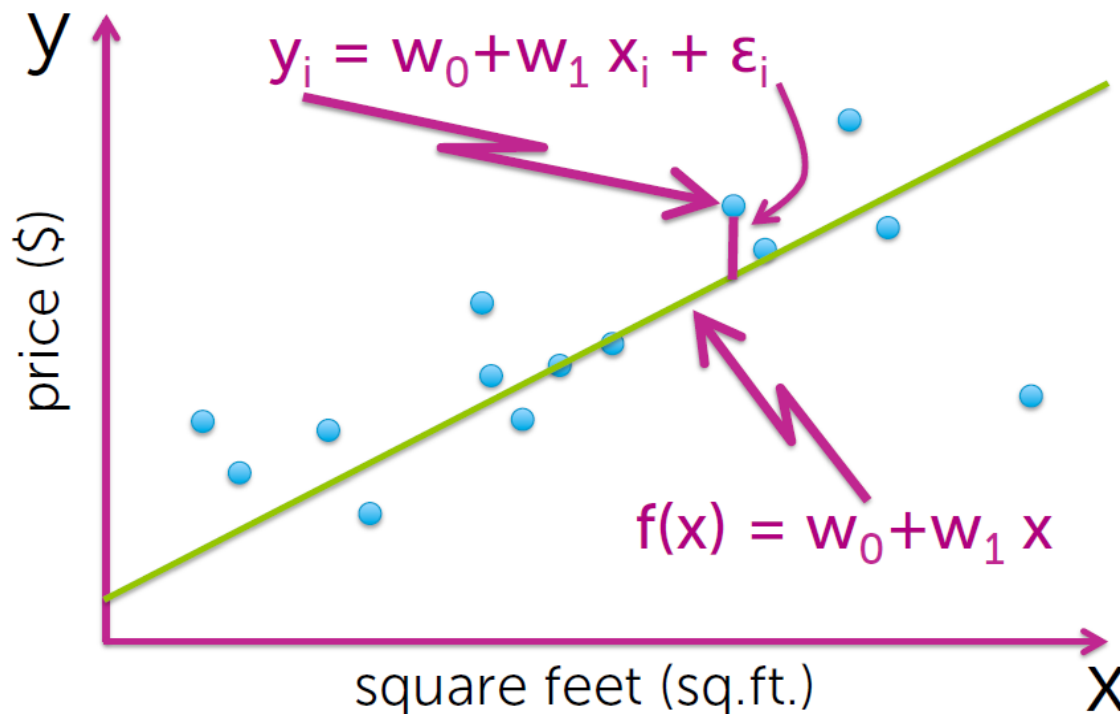


Simple linear regression model

20

Co to znaczy „simple”?

1 input x , fitujemy zależność liniową do danych

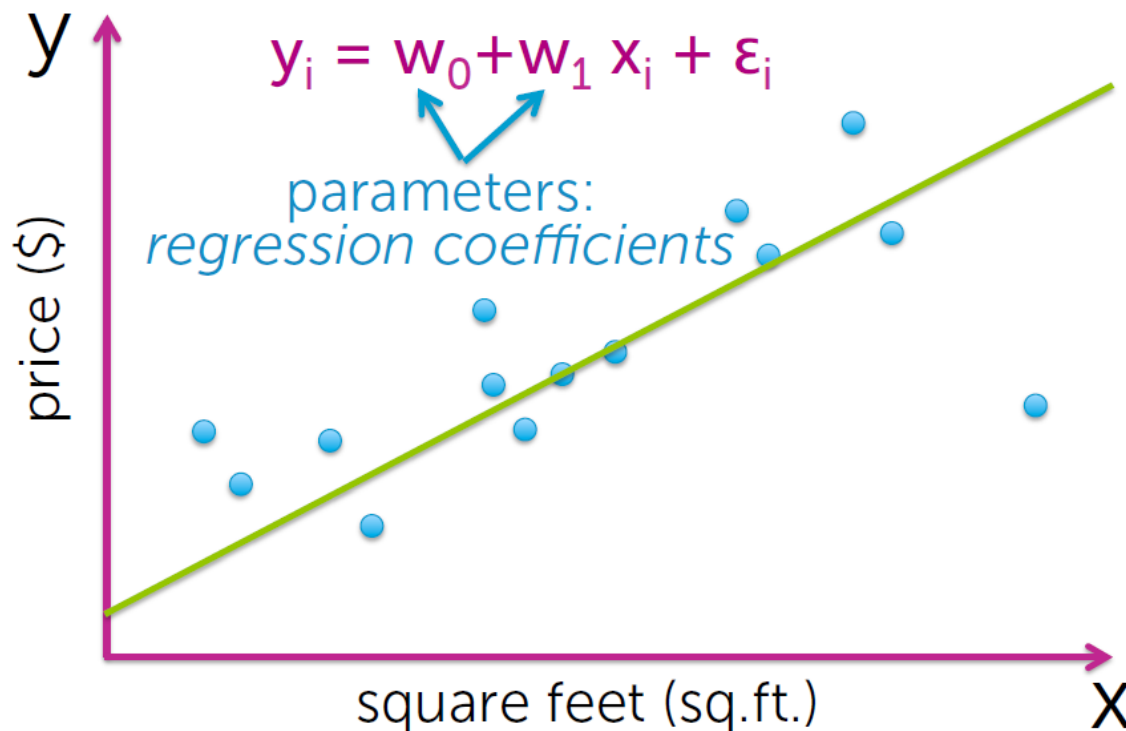


Simple linear regression model

21

Co to znaczy „simple”?

1 input x , fitujemy zależność liniową do danych

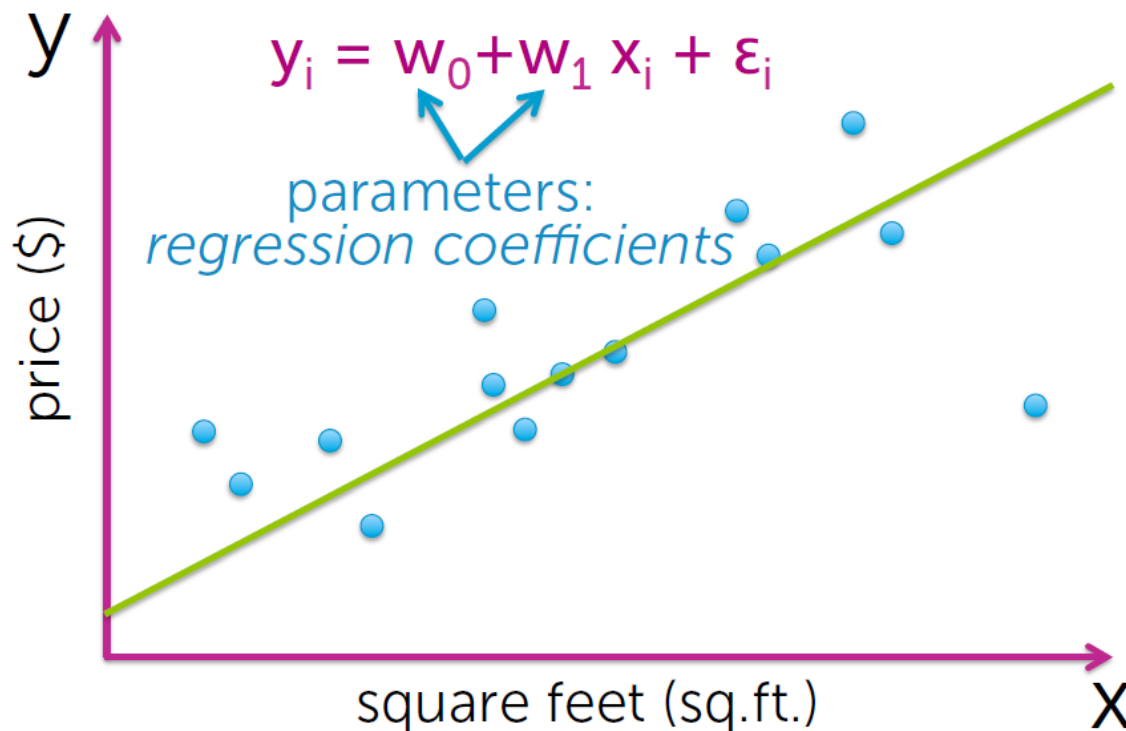


Simple linear regression model

22

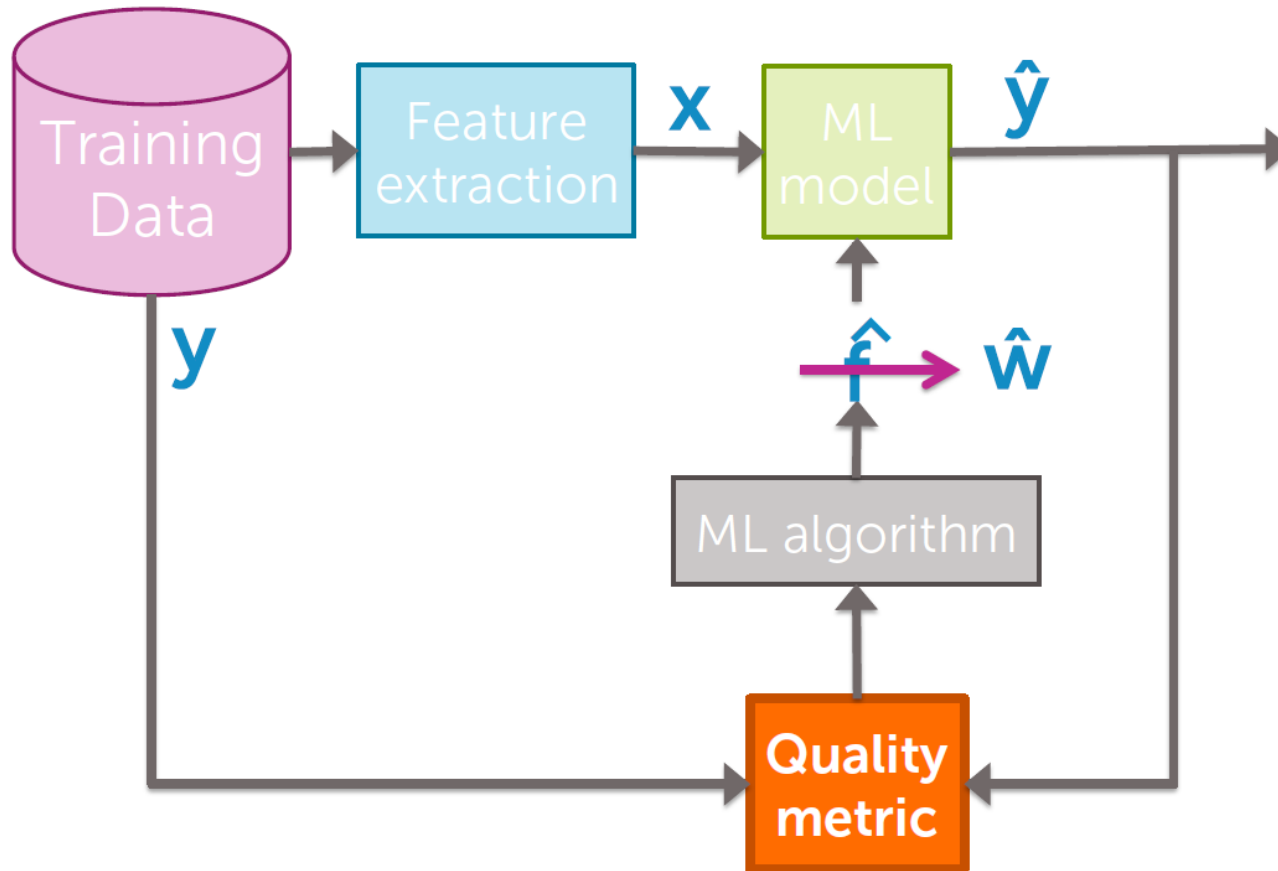
Co to znaczy „simple”?

1 input x , fitujemy zależność liniową do danych



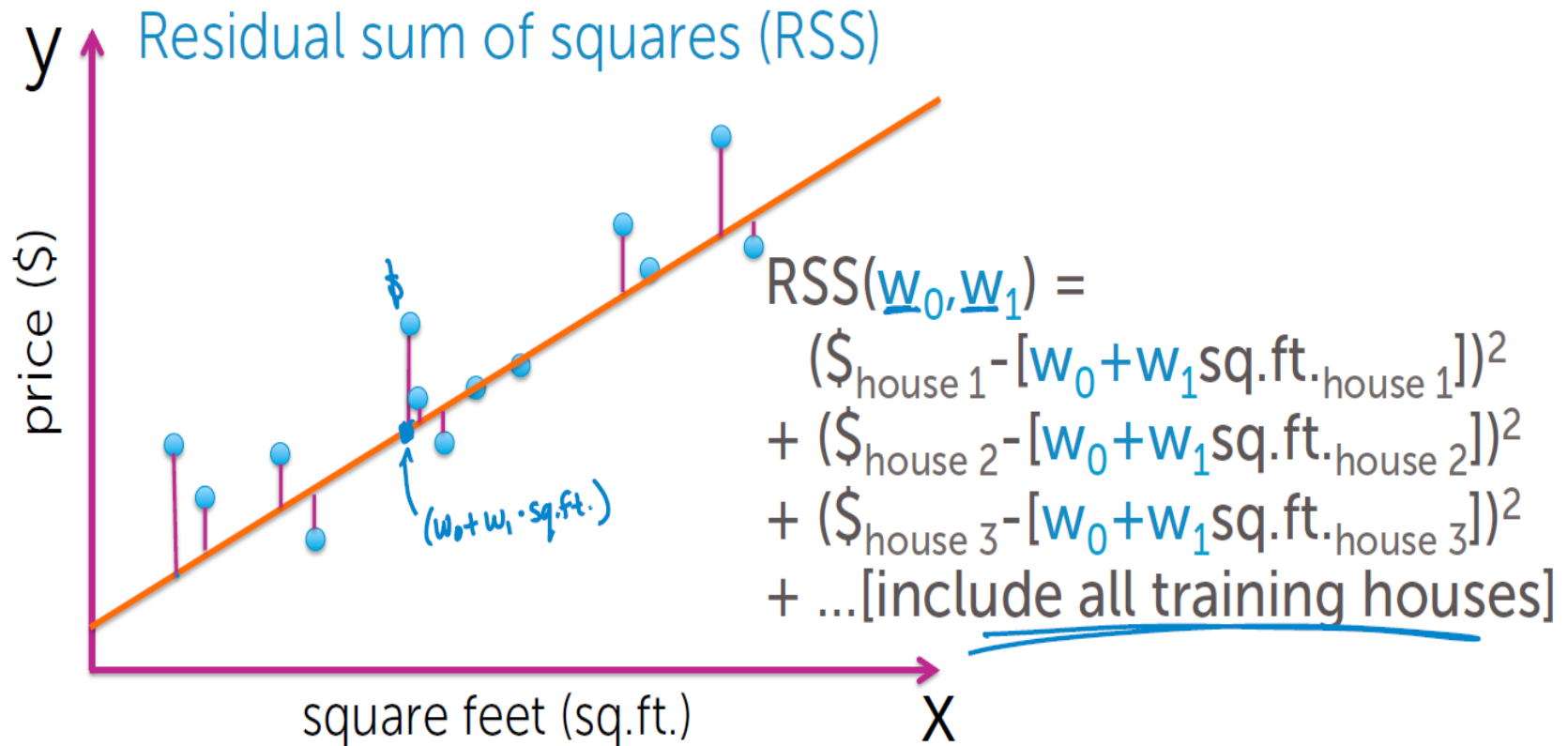
Pętla iteracyjna

23



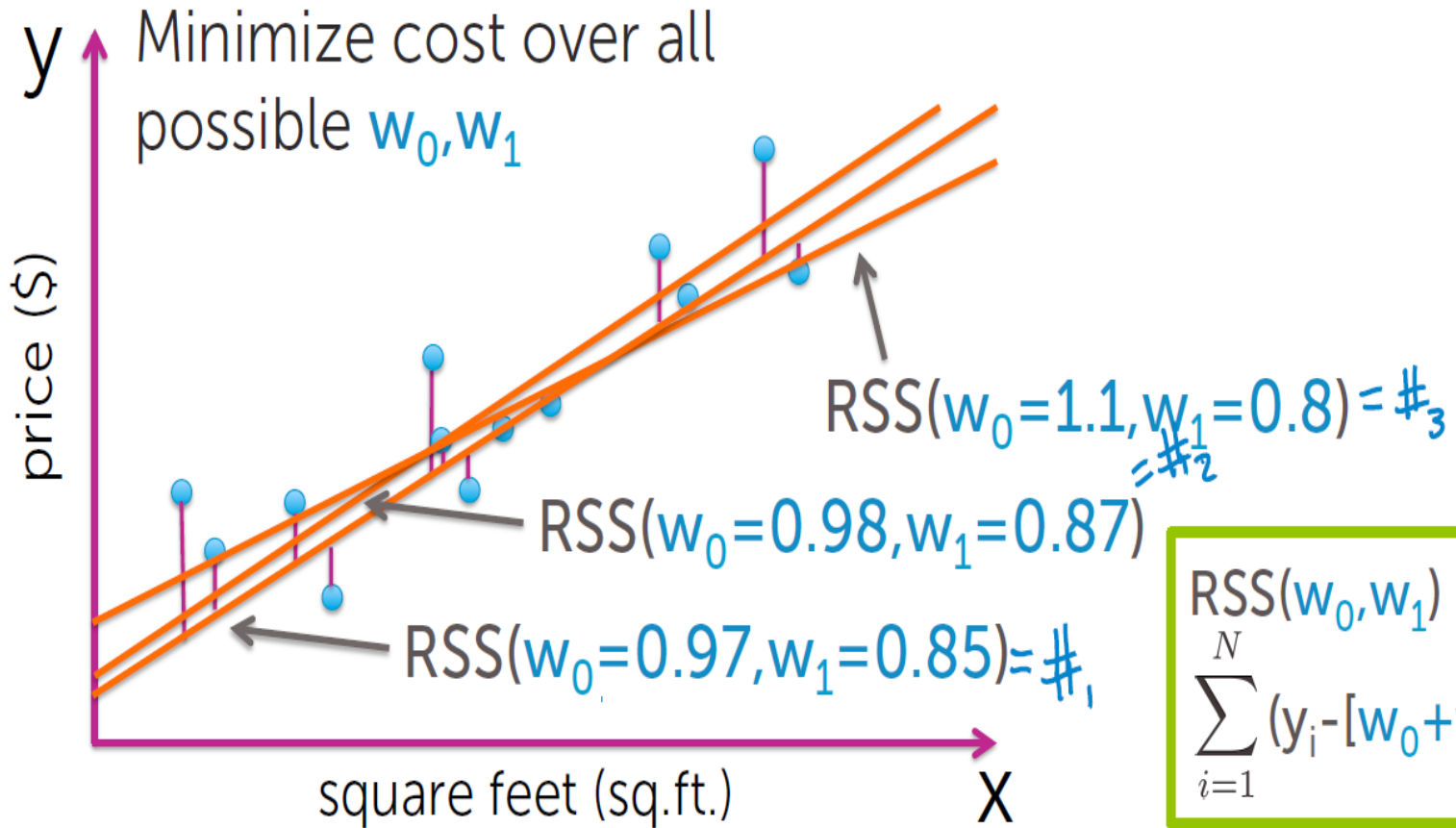
Funkcja „kosztu”

24



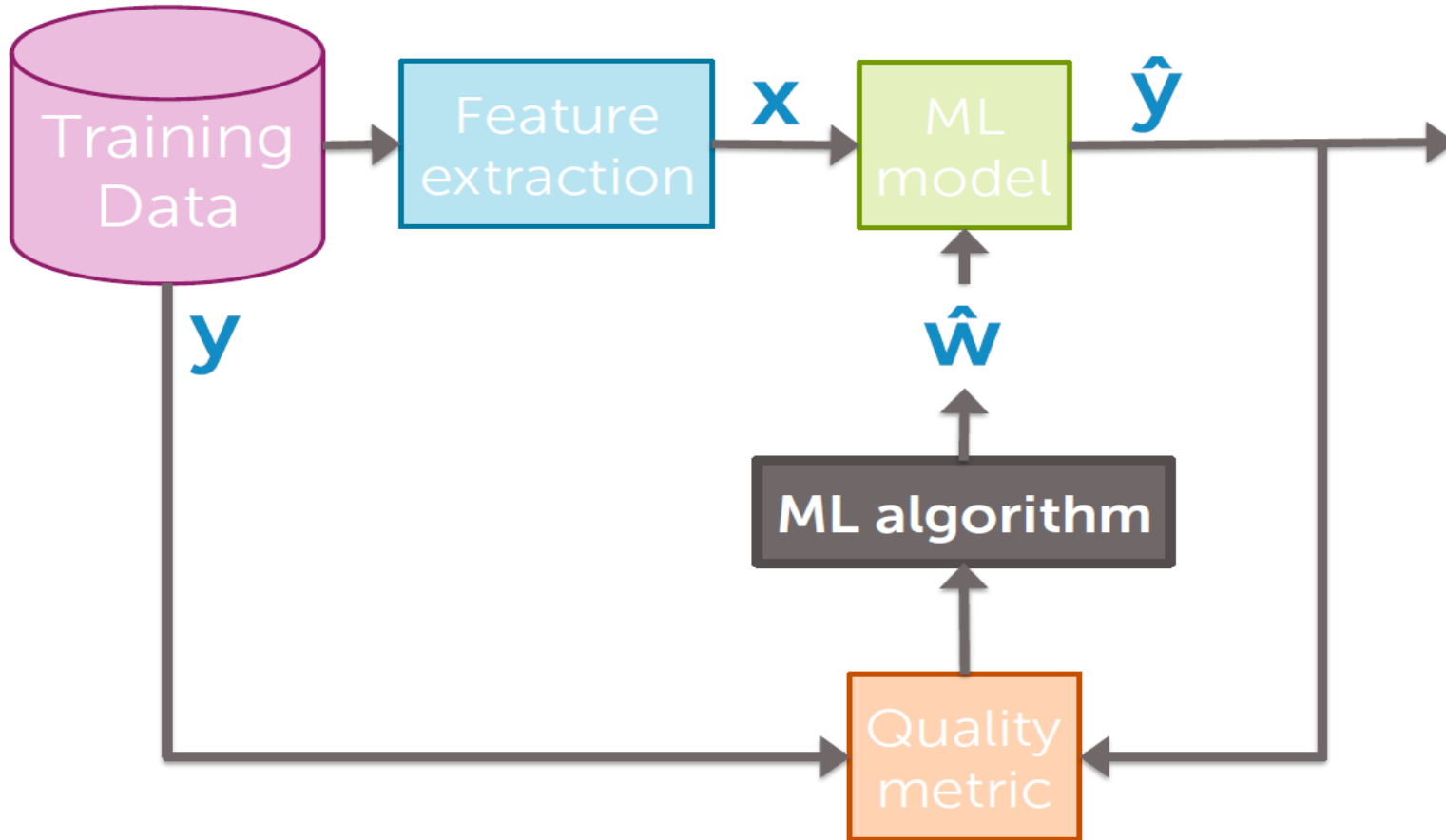
Minimalizacja funkcji „kosztu”

25



Pętla iteracyjna

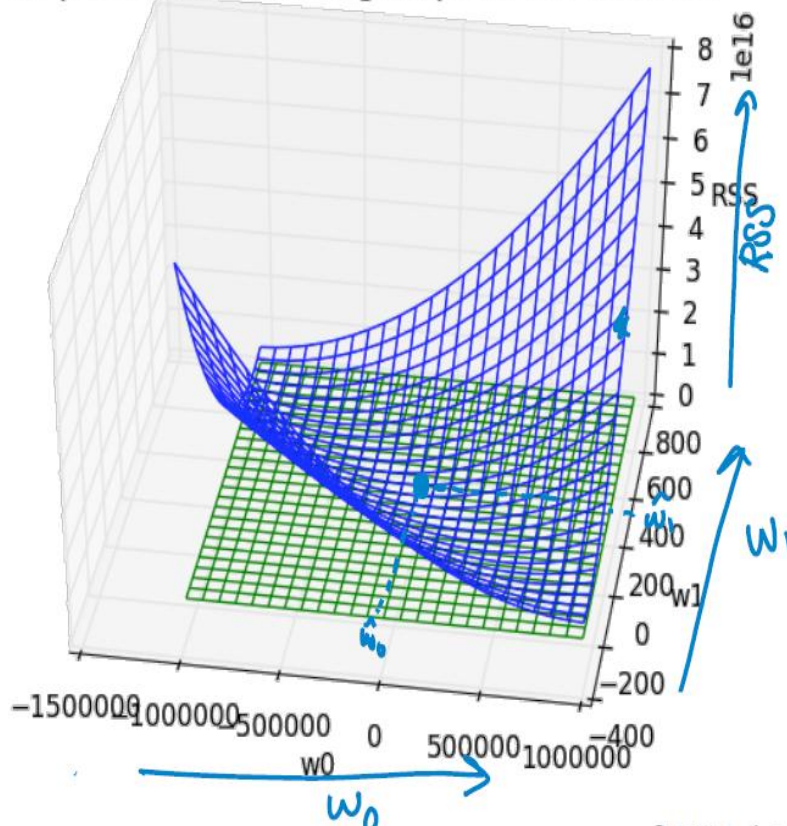
26



Minimalizacja funkcji „kosztu”

27

3D plot of RSS with tangent plane at minimum



Minimize function
over all possible w_0, w_1

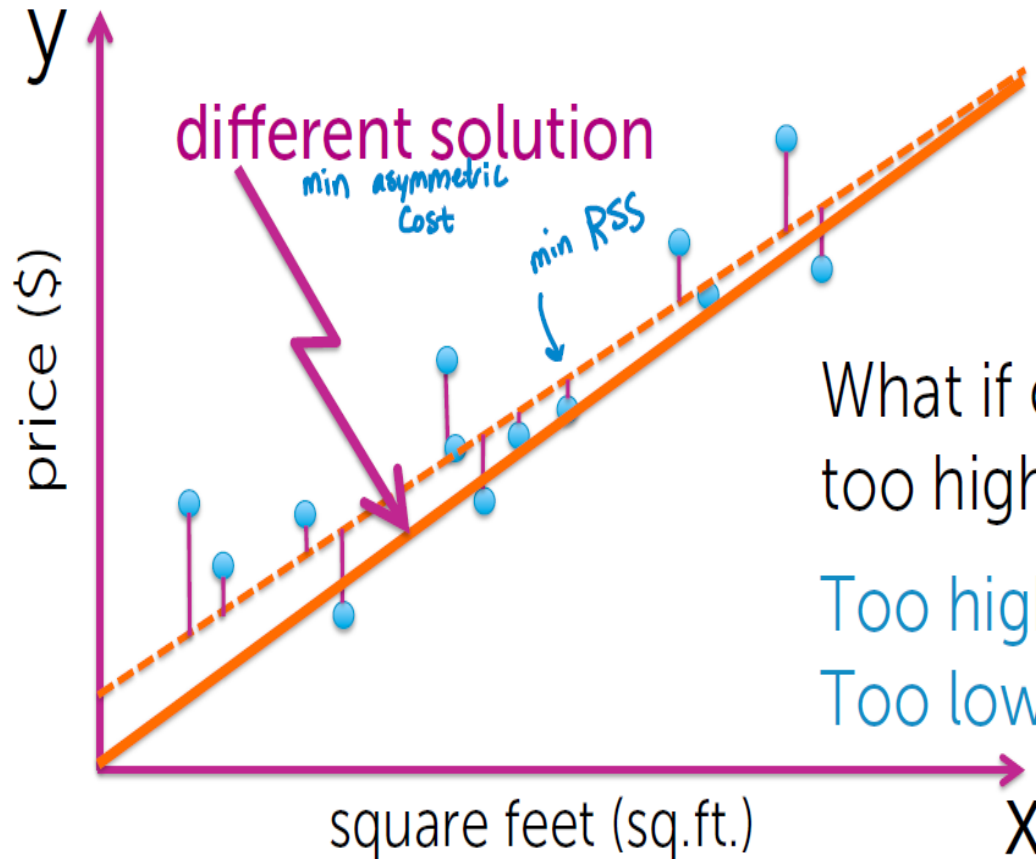
$$\min_{w_0, w_1} \sum_{i=1}^N (y_i - [w_0 + w_1 x_i])^2$$

RSS(w_0, w_1) is a function
of 2 variables = $g(w_0, w_1)$

15

Asymetryczny błąd kosztu

28



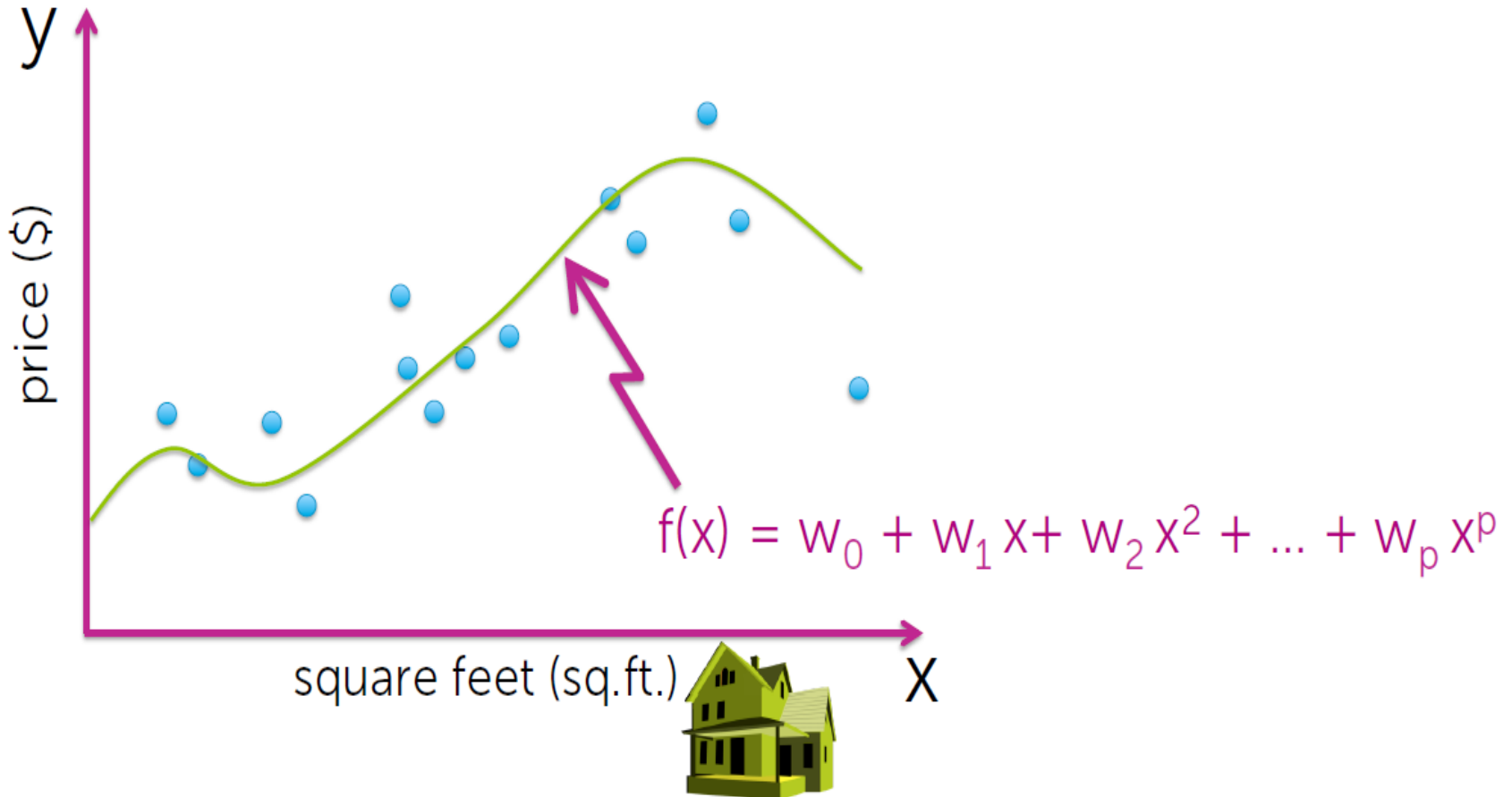
What if cost of listing house too high has **bigger cost**?

Too high \rightarrow no offers ($\$=0$)

Too low \rightarrow offers for lower \$

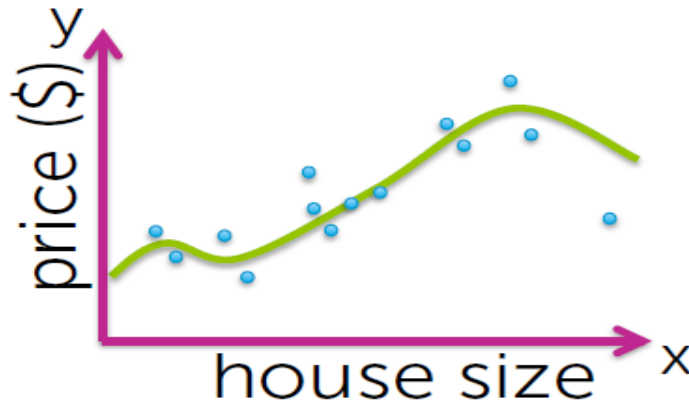
Polynomial regression

29

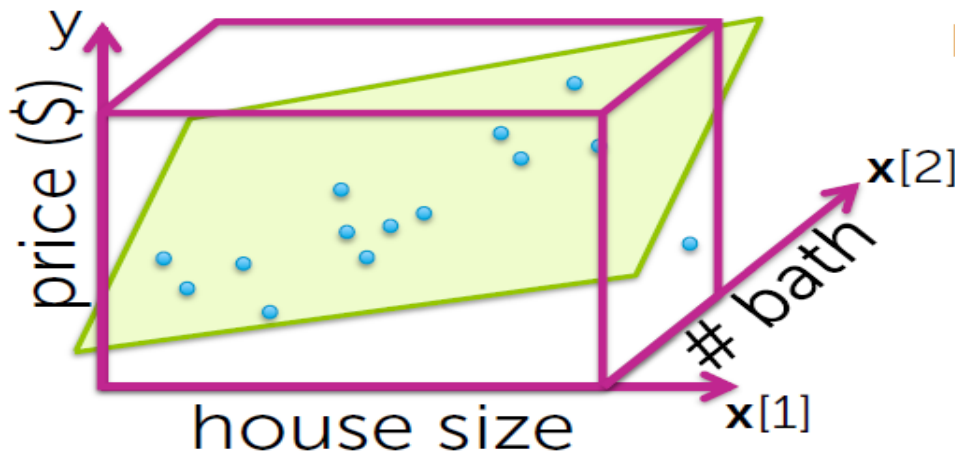


Multiple regression

30



Fit more complex relationships than just a line



Incorporate more inputs

- Square feet
- # bathrooms
- # bedrooms
- Lot size
- Year built
- ...

Regresja: przewidywanie na podstawie danych

31

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

Klasyfikacja

32

Uczenie maszynowe:
Klasyfikacja

□ Inteligentny system rankingu restauracji

It's a big day & I want to book a table at a nice Japanese restaurant



Klasyfikacja

33

Uczenie maszynowe:
Klasyfikacja

□ Inteligentny system rankingu restauracji



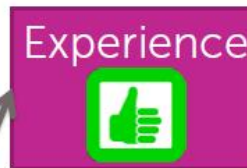
Positive reviews not positive about everything

Sample review:

Watching the chefs create incredible edible art made the experience very unique.

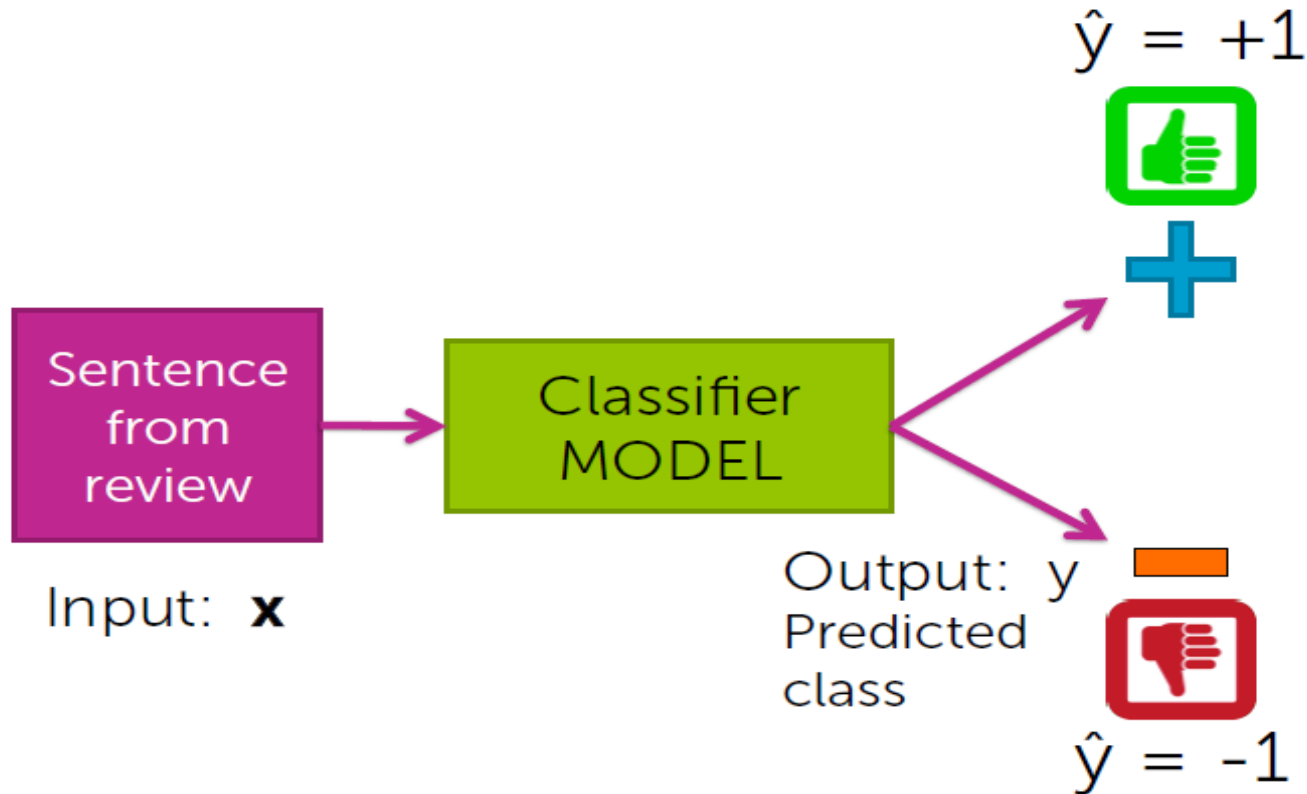
My wife tried their ramen and it was pretty forgettable.

All the sushi was delicious! Easily best sushi in Seattle.



Prosta klasyfikacja

34



Note: we'll start talking about 2 classes, and address multiclass later

Dane trenujące

35

- Używamy danych trenujących aby przypisać prawdopodobieństwo (wagę) dla każdego słowa

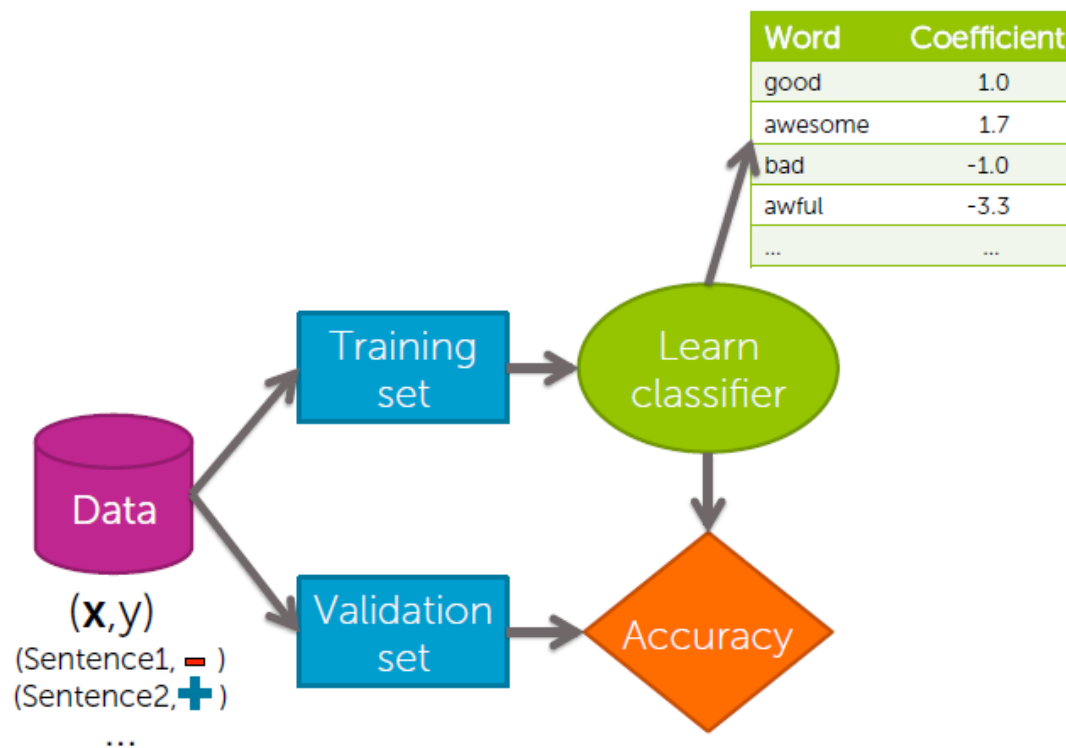
Word	Coefficient
good	1.0
great	1.5
awesome	2.7
bad	-1.0
terrible	-2.1
awful	-3.3
restaurant, the, we, where, ...	0.0
...	...

- Waga całego zdania (score) to będzie prosta suma tych wag

Uczenie klasyfikatora

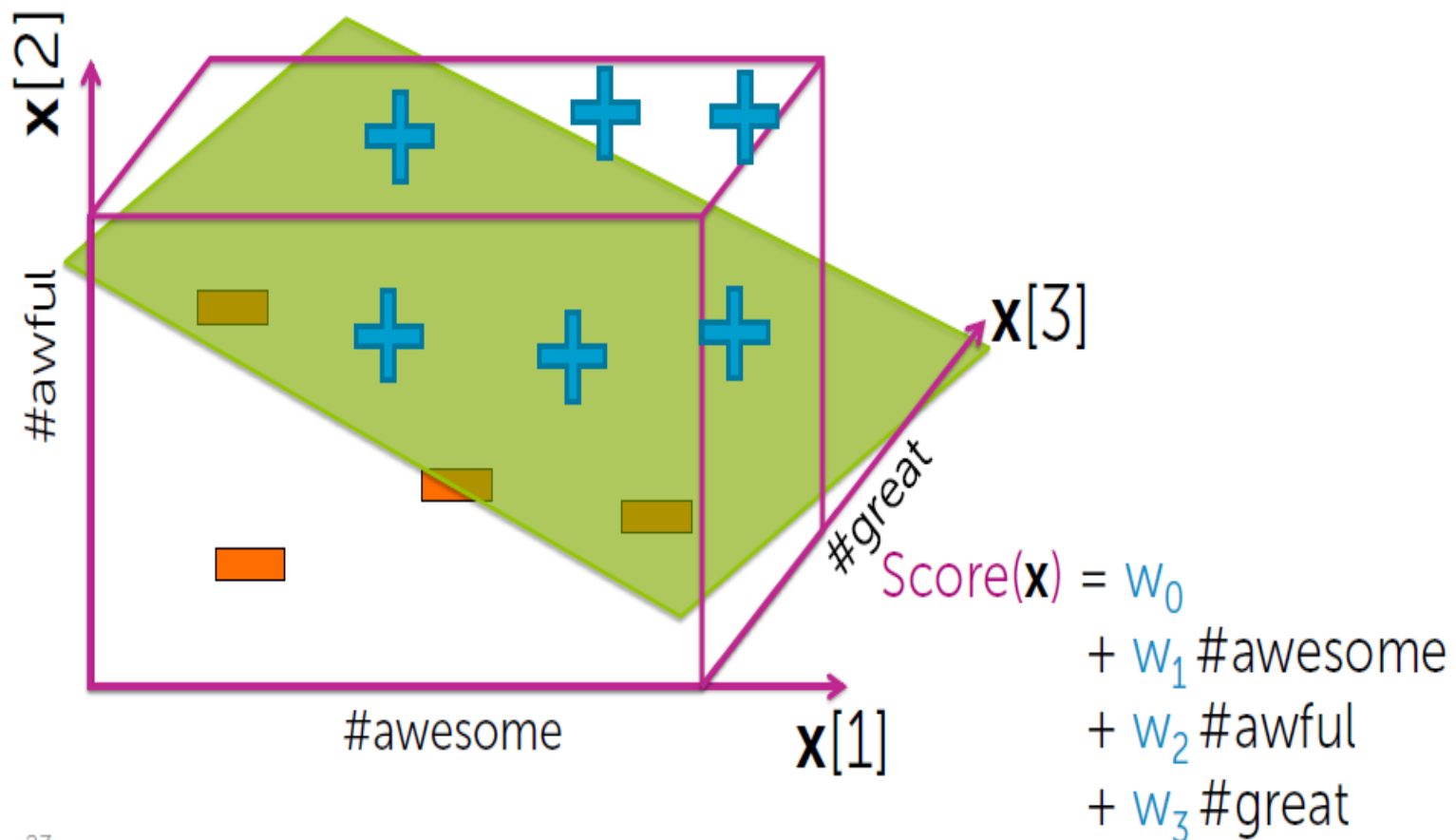
36

Training a classifier = Learning the coefficients



Jak wygląda nasz model

37



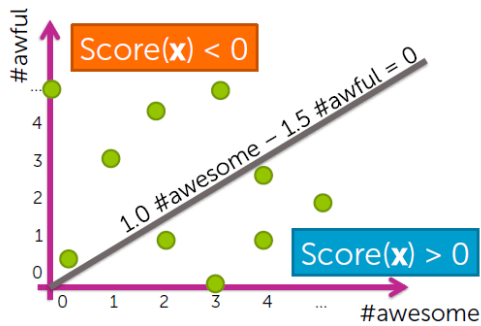
37

Zmiana współczynników

38

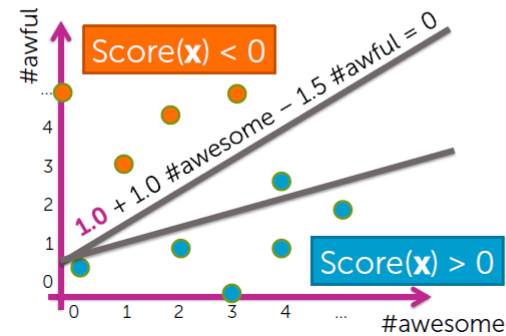
Input	Coefficient	Value
	w_0	0.0
#awesome	w_1	1.0
#awful	w_2	-1.5

→ $\text{Score}(x) = 1.0 \cdot \text{\#awesome} - 1.5 \cdot \text{\#awful}$



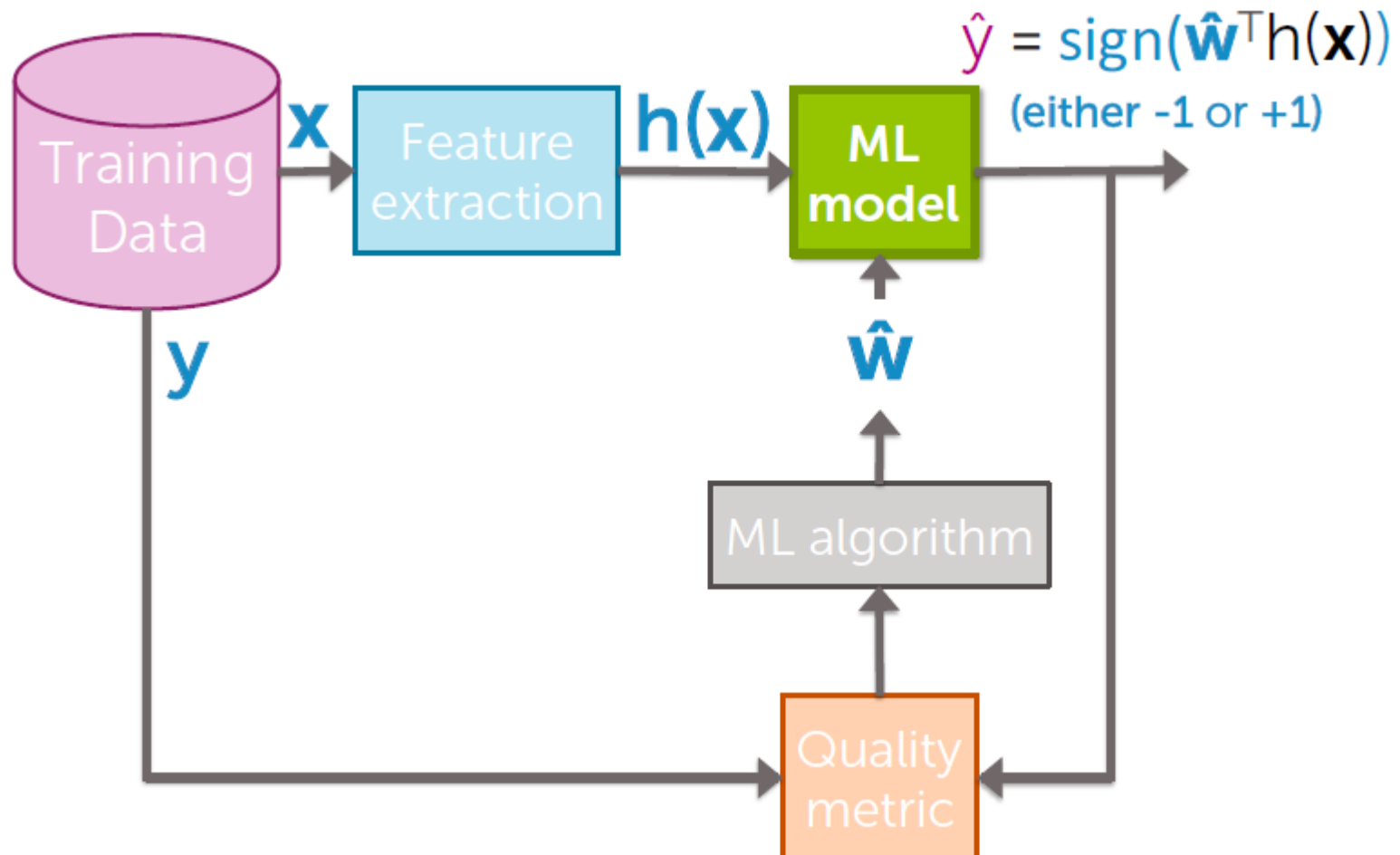
Input	Coefficient	Value
	w_0	1.0
#awesome	w_1	1.0
#awful	w_2	-3.0

→ $\text{Score}(x) = 1.0 + 1.0 \cdot \text{\#awesome} - 3.0 \cdot \text{\#awful}$



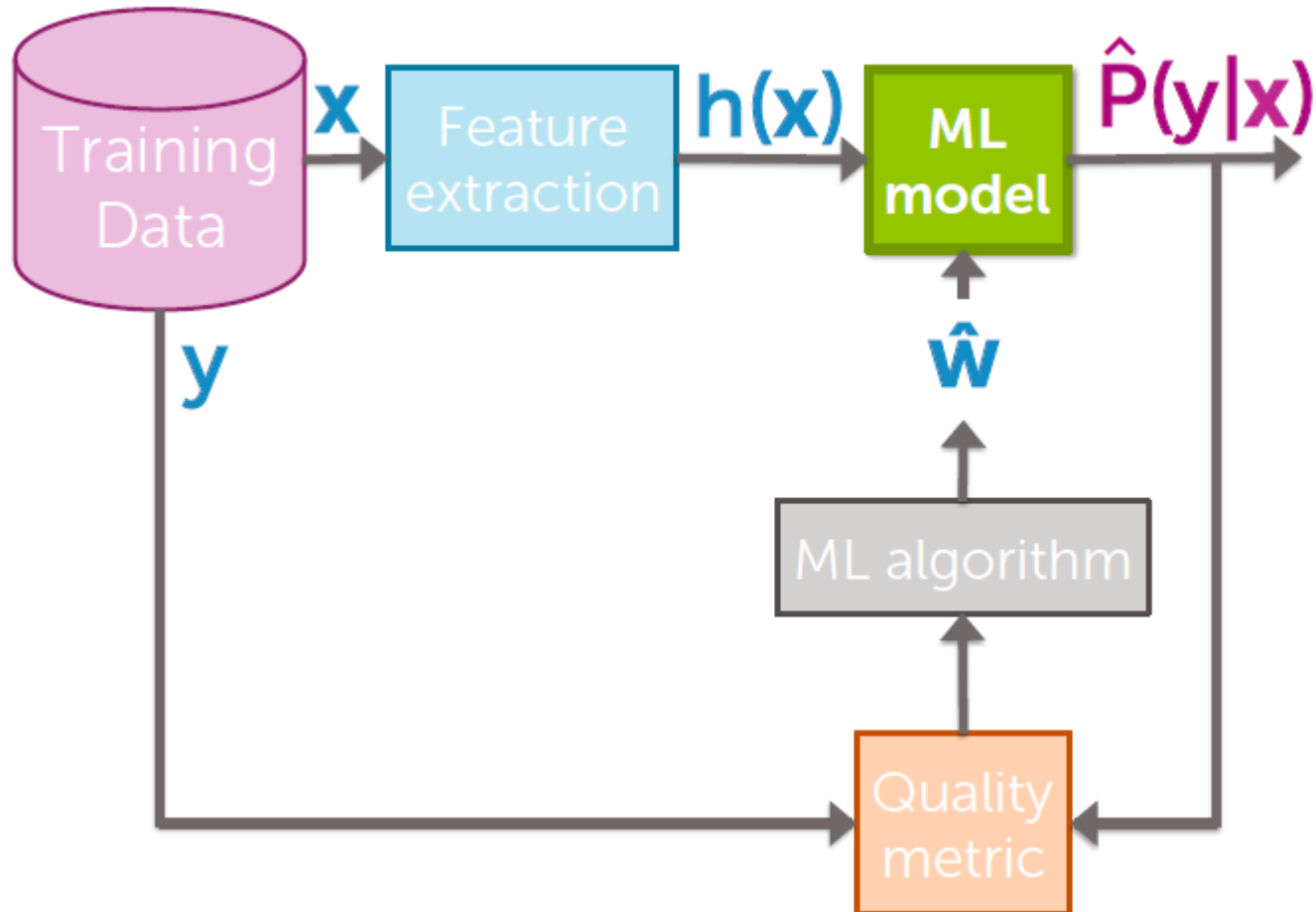
Jak wygląda nasz model?

39



Czy jesteś pewny swojej klasyfikacji

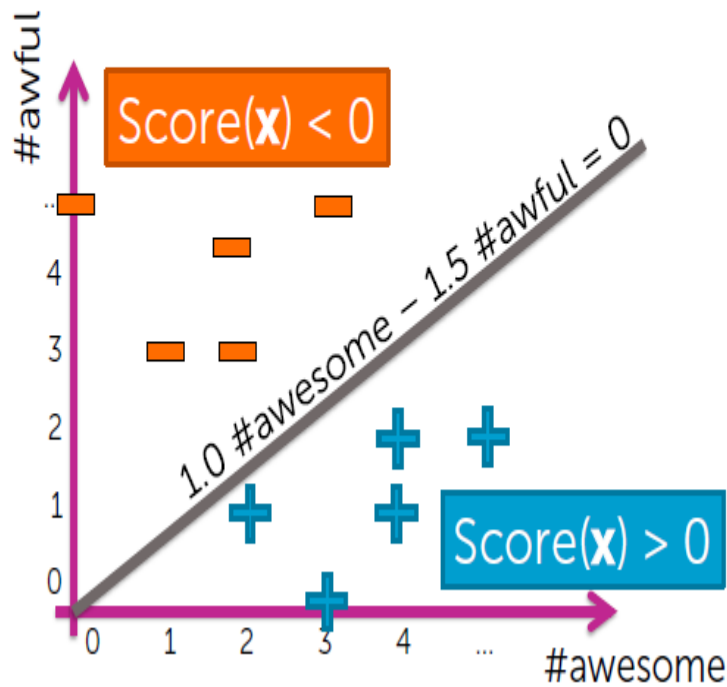
40



Interpretacja

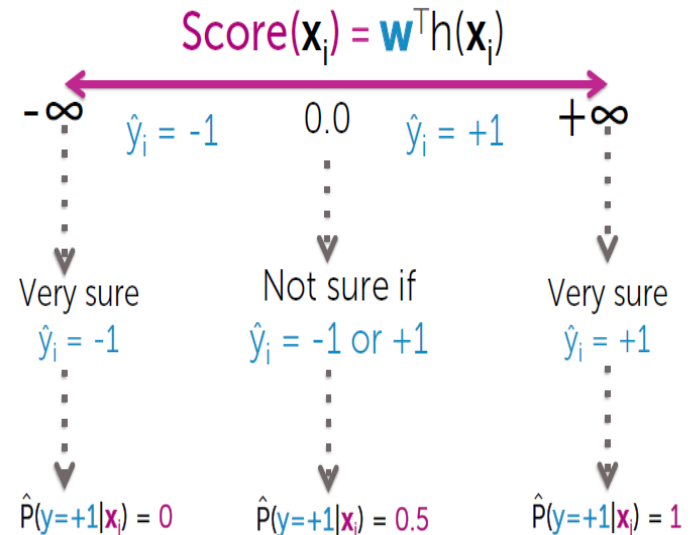
41

$$\begin{aligned} \text{Score}(\mathbf{x}_i) &= w_0 h_0(\mathbf{x}_i) + w_1 h_1(\mathbf{x}_i) + \dots + w_D h_D(\mathbf{x}_i) \\ &= \mathbf{w}^T \mathbf{h}(\mathbf{x}_i) \end{aligned}$$



Relate

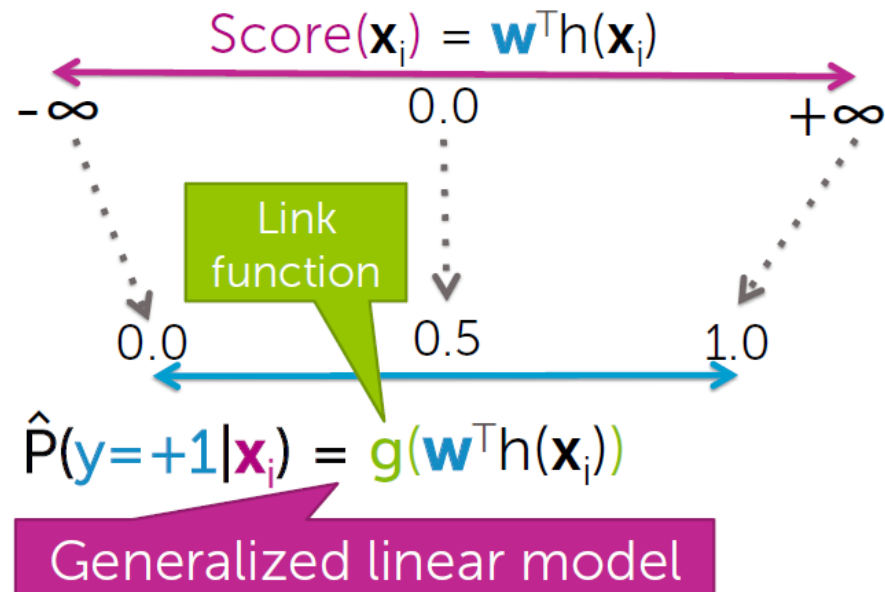
Score(\mathbf{x}_i) to $\hat{P}(y=+1|\mathbf{x}, \hat{\mathbf{w}})$?



Interpretacja

42

□ Link funkcja

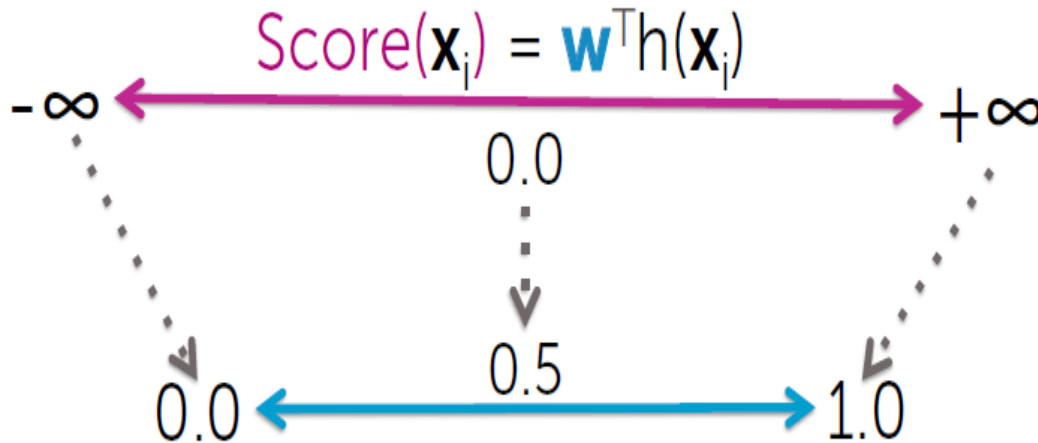


Logistic regression model

43

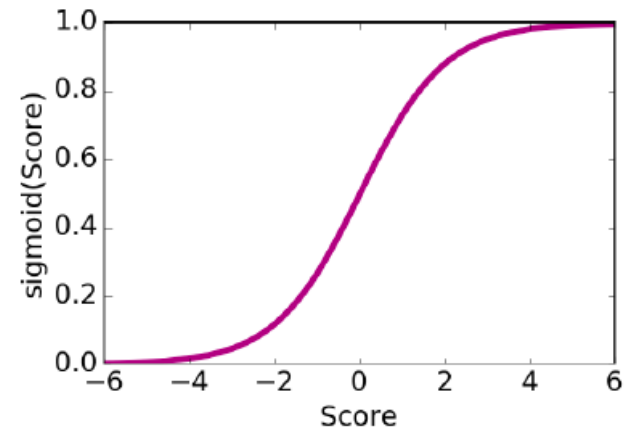
□ Sigmoid funkcija

Logistic regression model



$$P(y=+1|\mathbf{x}_i, \mathbf{w}) = \text{sigmoid}(\text{Score}(\mathbf{x}_i))$$

$$\text{sigmoid}(\text{Score}) = \frac{1}{1 + e^{-\text{Score}}}$$

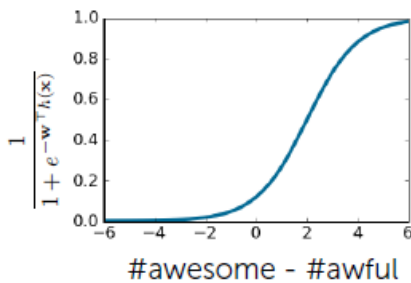


Logistic regression model

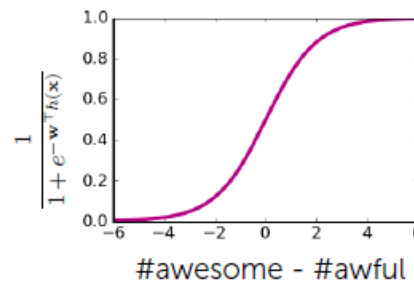
44

- Efekt współczynników na kształt funkcji sigmoidalnej

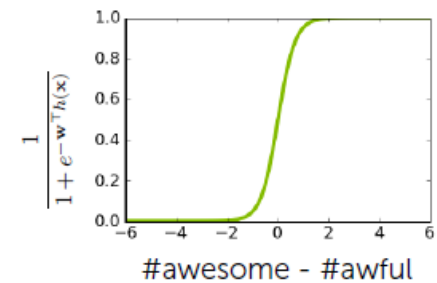
w_0	-2
$w_{\#awesome}$	+1
$w_{\#awful}$	-1



w_0	0
$w_{\#awesome}$	+1
$w_{\#awful}$	-1

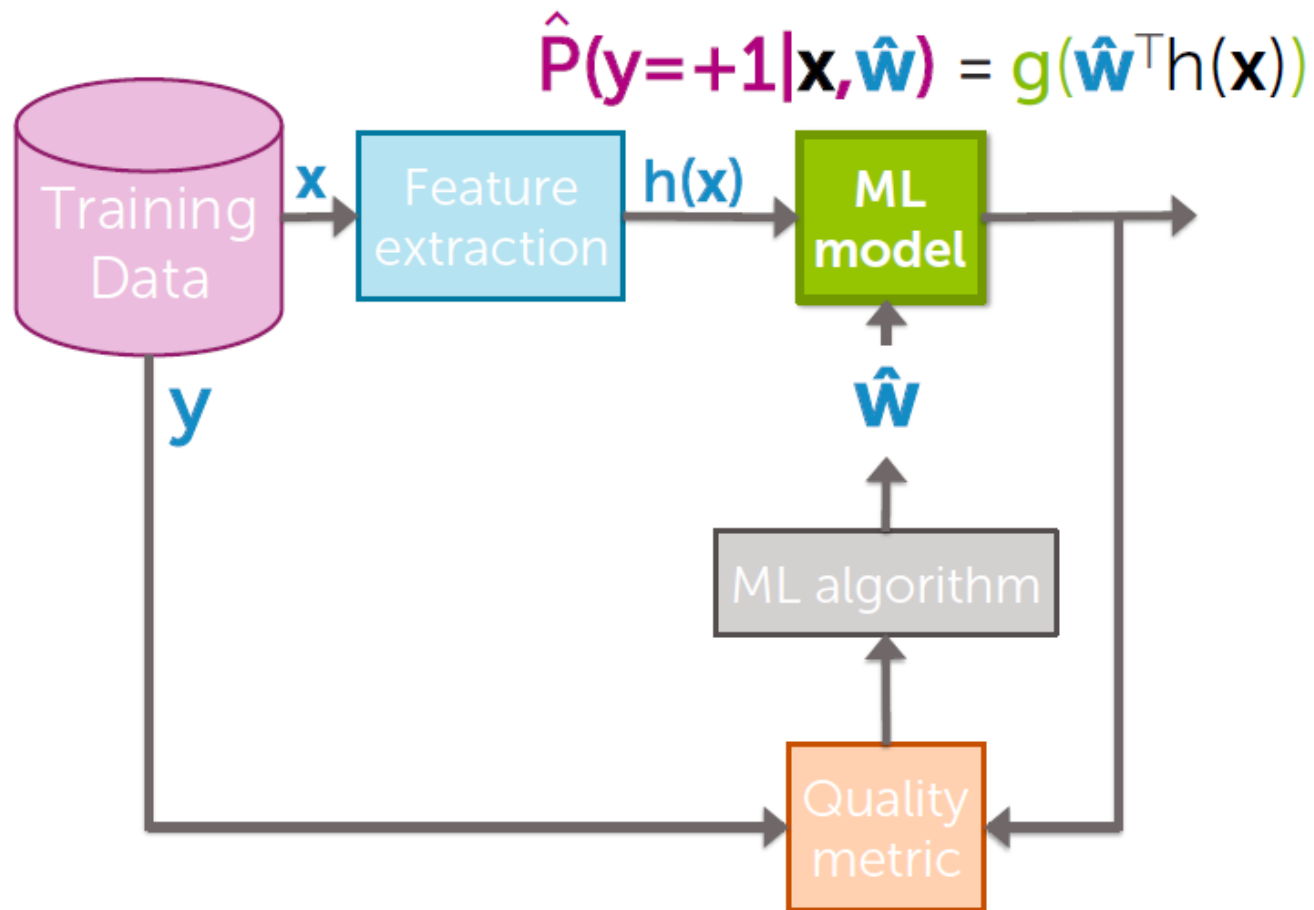


w_0	0
$w_{\#awesome}$	+3
$w_{\#awful}$	-3



Jak pewny jesteś swojej klasyfikacji

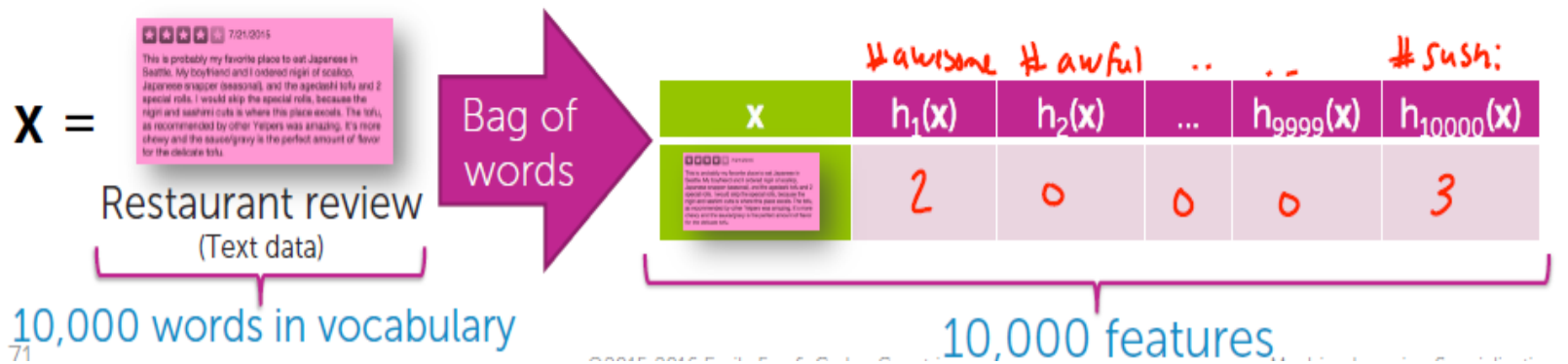
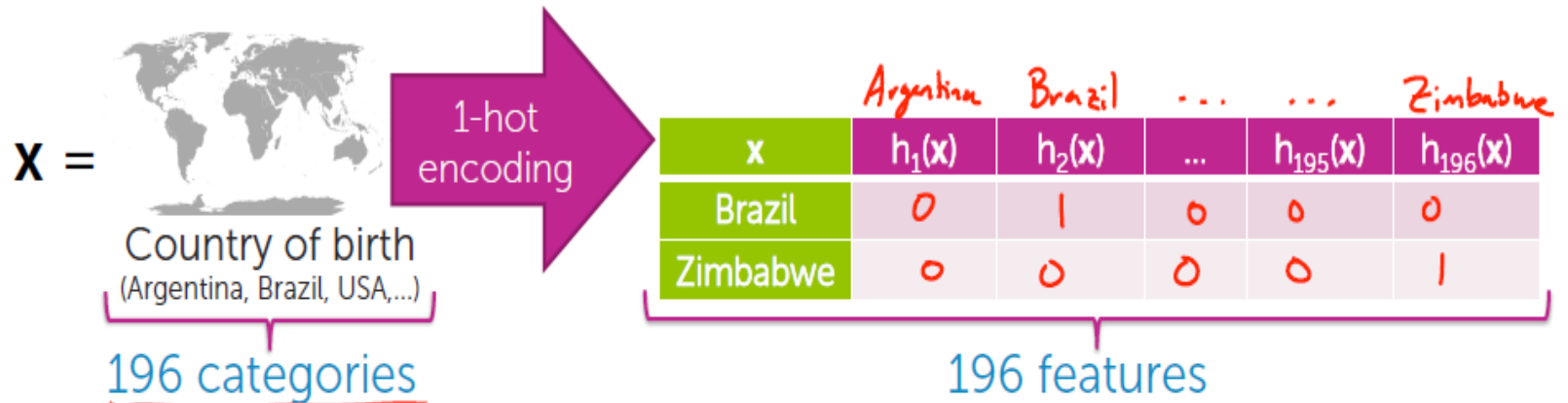
45



57

Zmienne opisowe

46



Klasyfikacja: ranking restauracji

47

Models

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

Algorithms

- Stochastic gradient descent
- Boosting

Concepts

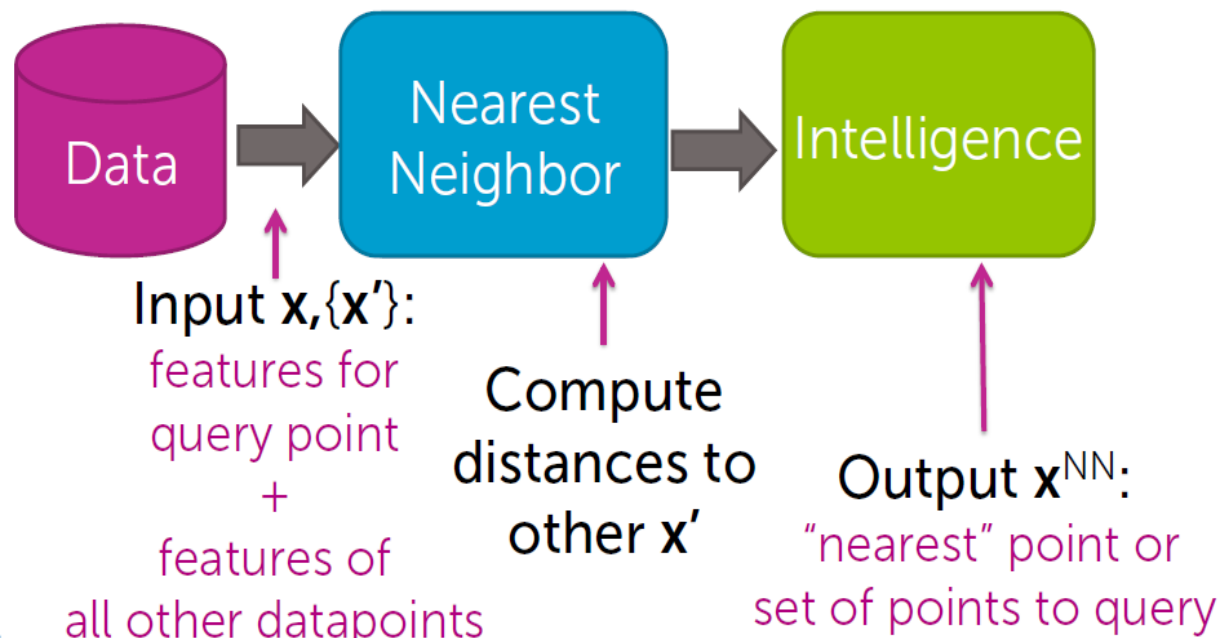
- Decision boundaries, MLE, ensemble methods, random forests, CART, online learning

Klastrowanie i wybór podobnych

48

Uczenie maszynowe:
Klasyfikacja i
wyszukiwanie
podobnych obiektów

□ Poszukiwanie podobnych obiektów



Wyszukaj podobny tekst

49

- Musimy zdefiniować co to znaczy „podobny”

Space of all articles,
organized by similarity of text

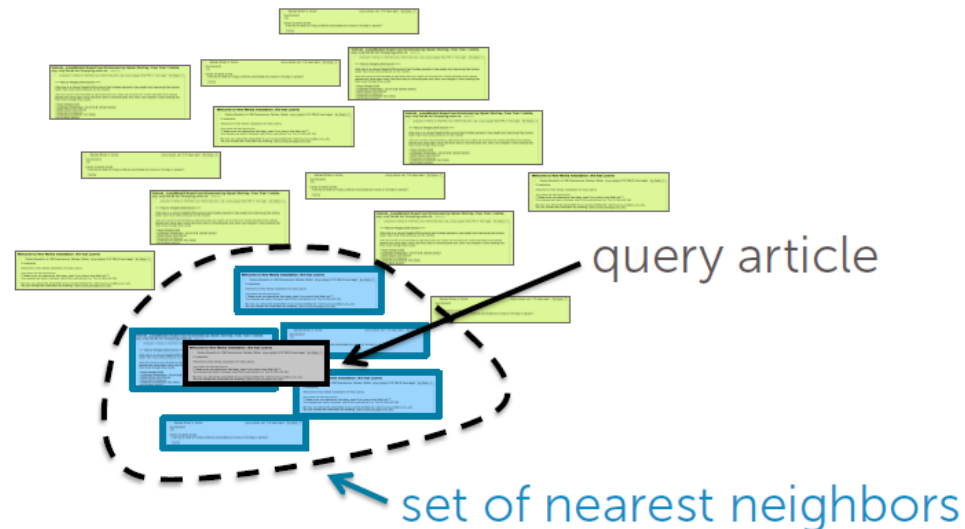


Wyszukaj podobny tekst

50

- Musimy zdefiniować co to znaczy „podobny”

Space of all articles,
organized by similarity of text



Technika może być stosowana do wielu zagadnień

51

Images



Products



Streaming content:

- Songs
- Movies
- TV shows
- ...

News articles



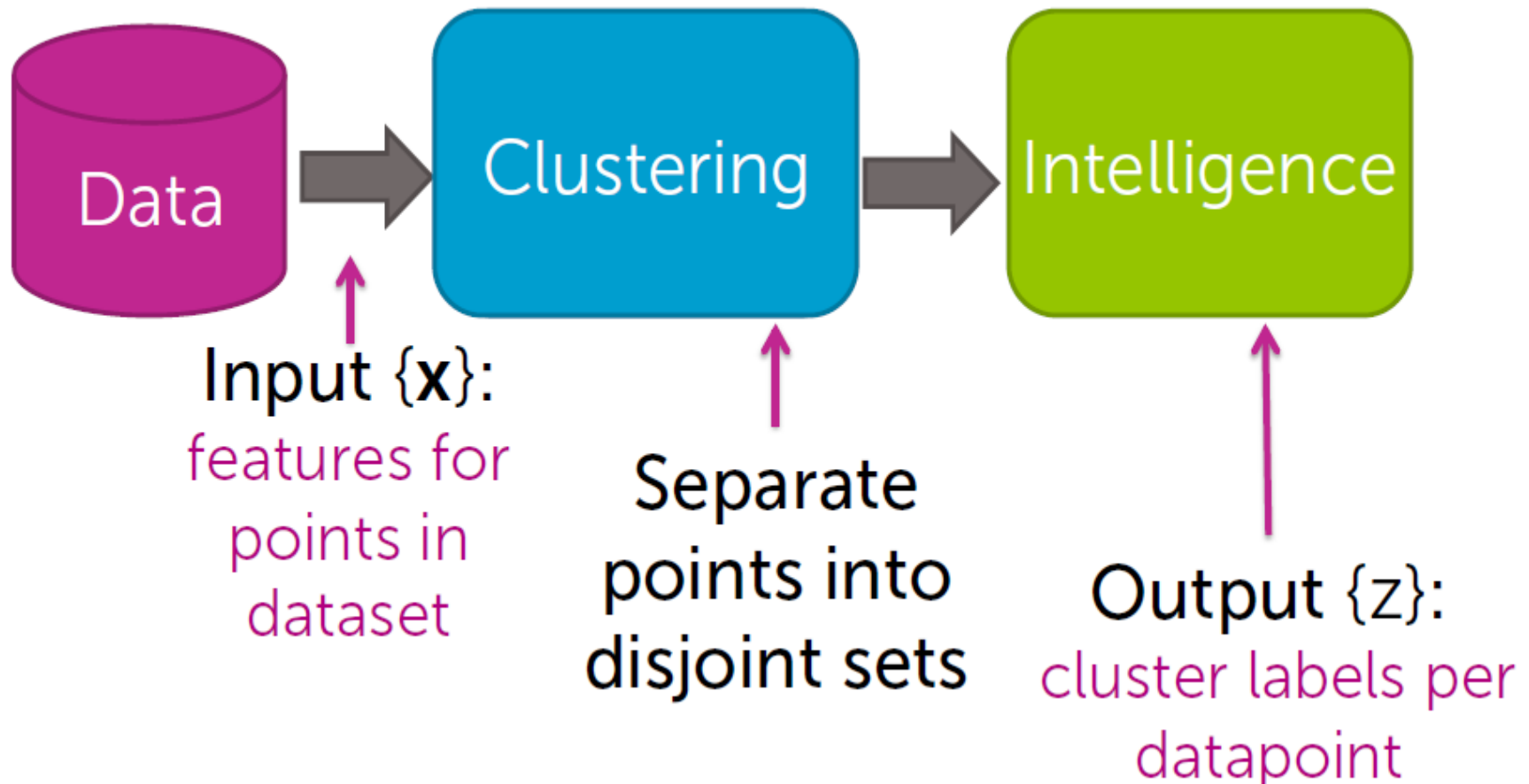
Social networks

(people you might want to connect with)



„Odkryj” grupę podobnych obiektów

52



Pogrupuj wg. tematów dokumenty

53

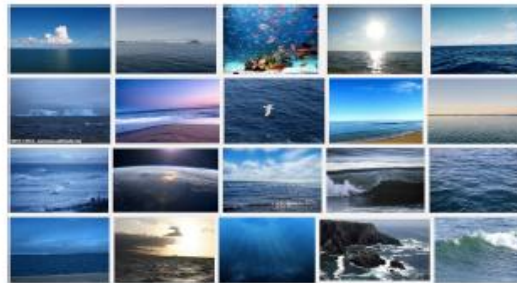


Pogrupuj obrazki

54

For search, group as:

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



Przykład: dokument na podobny temat

55

- Jak mierzymy podobieństwo ?
- Jak szukamy podobieństw ?



Najbliższy sąsiad

56

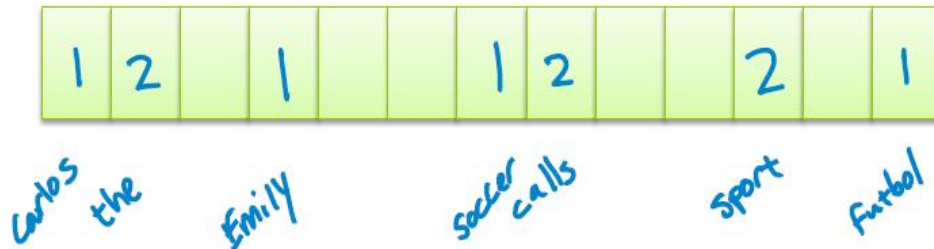
- **Input:** Query article  : \mathbf{x}_q
Corpus of documents (N docs)
 : $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$
- **Output:** *Most* similar article  $\leftarrow \mathbf{x}^{NN}$

Formally:
$$\mathbf{x}^{NN} = \min_{x_i} \text{distance}(x_q, x_i)$$

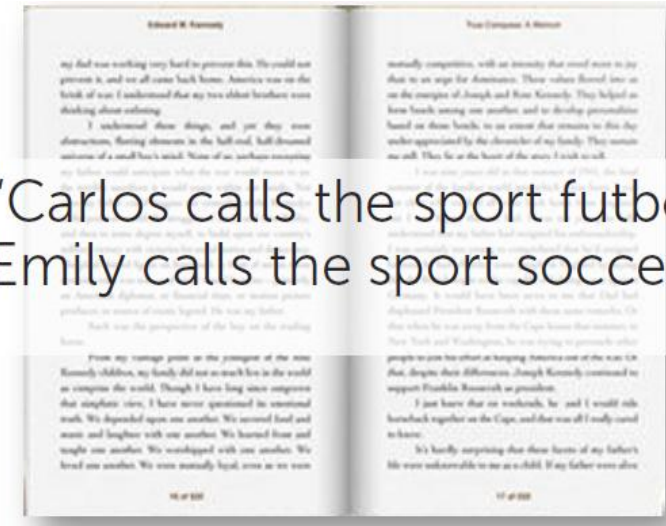
Reprezentacja dokumentu

57

- Dokument = zbiór słów
 - Nieistotna kolejność występowania
 - Zliczaj ilość wystąpień i zaznaczaj



“Carlos calls the sport futbol.
Emily calls the sport soccer.”



Reprezentacja dokumentów : TF-IDF

58


- Słowa mogą być rzadko występujące, te rzadko występujące są bardziej charakterystyczne.

Emphasizes **important words**

- Appears frequently in document (**common locally**)

Term frequency =  word counts

- Appears rarely in corpus (**rare globally**)

Inverse doc freq. = 
$$\log \frac{\# \text{ docs}}{1 + \# \text{ docs using word}}$$



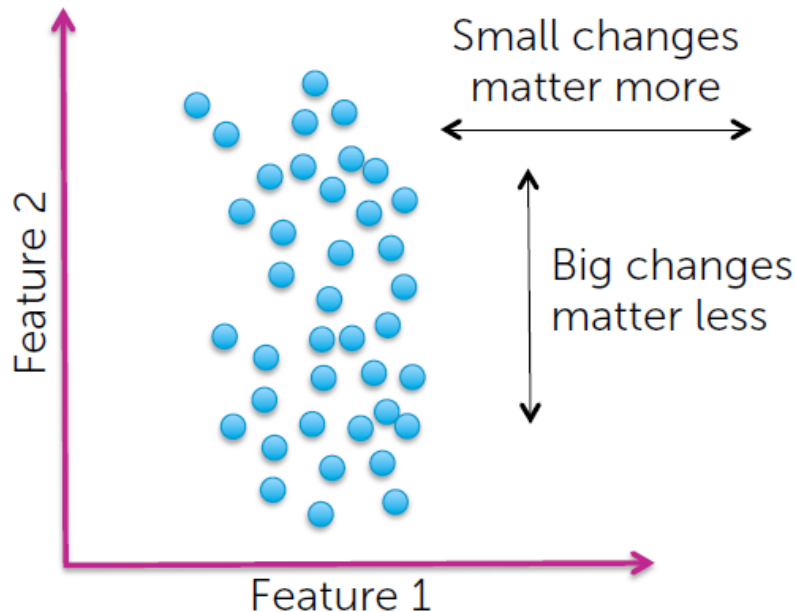
Trade off: **local frequency** vs. **global rarity**

tf * idf

Ważymy cechy charakterystyczne

59

- Niektóre cechy są bardziej ważne niż inne
- Niektóre różnice (absolutna wartość) są bardziej istotne niż inne. Liczy się rozmycie całego zbioru dla danej cechy.



Specify weights
as a function of
feature spread

For feature j :

$$\frac{1}{\max_i(\mathbf{x}_i[j]) - \min_i(\mathbf{x}_i[j])}$$

Skalowana Euclidesowa miara odległości

60

distance($\mathbf{x}_i, \mathbf{x}_q$) =

$$\sqrt{a_1(\mathbf{x}_i[1] - \mathbf{x}_q[1])^2 + \dots + a_d(\mathbf{x}_i[d] - \mathbf{x}_q[d])^2}$$

weight on each feature
(defining relative importance)

Naturalna miara odległości



Similarity

= 0



Cosine miara odległości

62

$$\text{Similarity} = \frac{\sum_{j=1}^d \mathbf{x}_i[j] \mathbf{x}_q[j]}{\sqrt{\sum_{j=1}^d (\mathbf{x}_i[j])^2} \sqrt{\sum_{j=1}^d (\mathbf{x}_q[j])^2}}$$

$$\frac{\sum_{j=1}^d \mathbf{x}_i[j] \mathbf{x}_q[j]}{\sqrt{\sum_{j=1}^d (\mathbf{x}_i[j])^2} \sqrt{\sum_{j=1}^d (\mathbf{x}_q[j])^2}}$$

$$a^T b = \|a\| \|b\| \cos(\theta)$$

- Not a proper distance metric

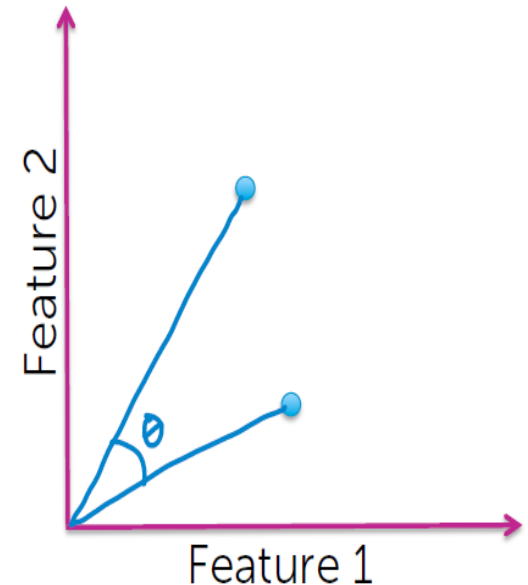
- Efficient to compute for sparse vecs

$$\mathbf{x}_i^T \mathbf{x}_q = \cos(\theta)$$

$$\frac{\mathbf{x}_i^T \mathbf{x}_q}{\|\mathbf{x}_i\| \|\mathbf{x}_q\|}$$

$$= \left(\frac{x_i}{\|x_i\|} \right)^T \left(\frac{x_q}{\|x_q\|} \right)$$

first normalize



Klastering: poszukiwanie dokumentu

63

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

System rekomendujący

64

Uczenie maszynowe:
system
rekomendujący

□ Personalizacja

You Tube

100 Hours a Minute
What do I care about?

Information overload



Browsing is "history"
– Need new ways
to discover content

Personalization: Connects *users & items*

viewers

videos

Rekomendacja: filmy

65



Connect users with movies
they may want to watch

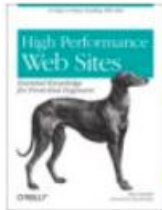
Rekomendacja: producty

66

amazon.com

[Help](#) | [Close window](#)

Recommended for You



**High Performance Web Sites:
Essential Knowledge for
Front-End Engineers**

by Steve Souders (Author)

Our Price: \$19.79

Used & new from \$16.24

[Add to Cart](#)

[Add to Wish List](#)

Because you purchased...

**Programming Collective Intelligence: Building
Smart Web 2.0 Applications** (Paperback)

by Toby Segaran (Author)



Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#)

Even Faster Web Sites: Performance... (Paperback) by Steve Souders
★★★★☆ (7) \$23.10
[Fix this recommendation](#)

Simply JavaScript (Paperback) by Kevin Yank
★★★★☆ (19) \$26.37
[Fix this recommendation](#)

The Art & Science of Java (Paperback)
★★★★☆ (5)
[Fix this recommendation](#)

[Any Category](#) Algorithms Boxed Sets Business & Culture Java
Networking Networks, Protocols & APIs SQL

Recommendations combine
global & session interests

System rekomendujący: popularność

67

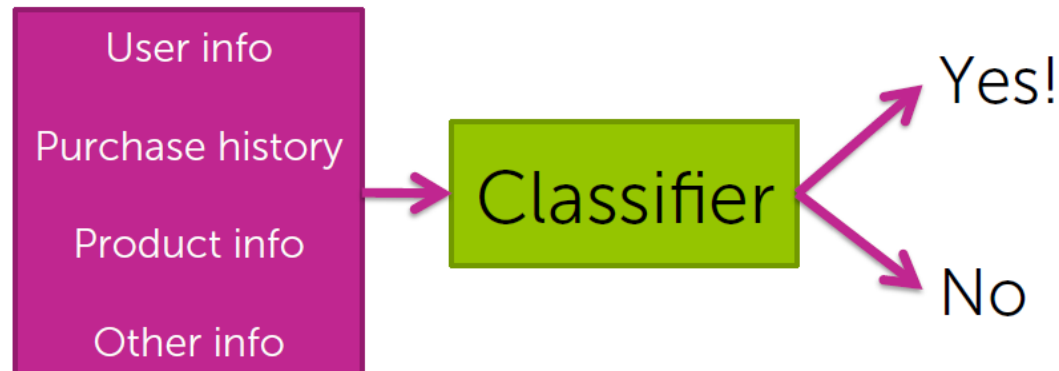
- **Popularność?**
 - ▣ **Ranking wg. liczba oglądań**
 - ▣ **Nie ma personalizacji**

System rekomendujący: klasyfikacja

68

□ Klasyfikacja?


- ▣ **Jakie prawdopodobieństwo że kupie ten product.**
- ▣ **Personalizacja: patrzy na historie zakupów, koreluje z pora roku, pora dnia, etc.**



System rekomendujący: korelacje

69

- **Patrzy na korelacje. Osoby które kupiły A kupiły również B**
 - ▣ **Utwórzmy macierz korelacji**

User  purchased *diapers*

1. Look at *diapers* row of matrix
2. Recommend other items with largest counts
 - *baby wipes, milk, baby food,...*

System rekomendujący: korelacje




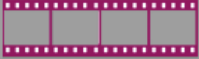














70

- **Patrzy na korelacje. Osoby które kupiły A kupiły również B**
 - ▣ **Macierz korelacji może należy znormalizować?**
 - ▣ **A może wprowadzić jakąś miarę „co znaczy podobne”?**
- **Ograniczenie:**
 - ▣ **Nie patrzy na historię w czasie**
 - ▣ **Co zrobić z nowym użytkownikiem systemu?**

System rekomendujący: filmy

71

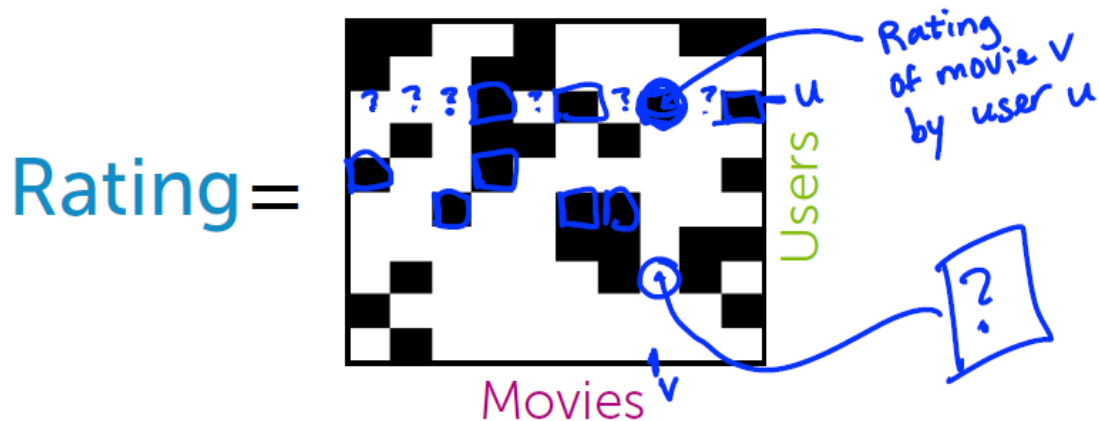
- Users watch movies and rate them

User	Movie	Rating
		★★★★☆
		★★★★★
		★★★☆☆
		★★★☆☆
		★★★★☆
		★★★☆☆
		★★★★☆
		★★★★★
		★★★★☆

Each user only watches a few of the available movies

System rekomendujący: filmy

72



- **Data:** Users score some movies

$Rating(u,v)$ known for black cells
 $Rating(u,v)$ unknown for white cells

- **Goal:** Filling missing data?



System rekomendujący: filmy

73

- Squeezing last bit of accuracy by blending models
- Netflix Prize 2006-2009
 - 100M ratings
 - 17,770 movies
 - 480,189 users
 - Predict 3 million ratings to highest accuracy
 - **Winning team blended over 100 models**



The screenshot shows the Netflix Prize Leaderboard interface. At the top, there is a yellow banner with the text "Netflix Prize". Below the banner, there is a navigation menu with links: Home, Rules, Leaderboard, Register, Update, Submit, and Download. The main heading is "Leaderboard" in blue, followed by "10.05%" and "Display top 20 leaders." A yellow arrow points to the "10.05%" value. Below this is a table with the following columns: Rank, Team Name, Best Score, % Improvement, and Last Submit Time. The table lists the top 6 teams, with the first team, BellKor's Pragmatic Chaos, highlighted in red. A red banner below the table indicates the Grand Prize - RMSE <= 0.8563.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	BellKor's Pragmatic Chaos	0.8558	10.05	2009-06-26 18:42:37
Grand Prize - RMSE <= 0.8563				
2	PragmaticTheory	0.8582	9.80	2009-06-25 22:15:51
3	BellKor in BigChaos	0.8590	9.71	2009-05-13 08:14:09
4	Grand Prize Team	0.8593	9.68	2009-06-12 08:20:24
5	Dace	0.8604	9.56	2009-04-22 05:57:03
6	BigChaos	0.8613	9.47	2009-06-23 23:08:52

System rekomendujący: sprawność

74

The world of all baby products



System rekomendujący: sprawność

75

User likes subset of items



System rekomendujący: sprawność

76

How many liked items were recommended?

Recall

$$\frac{\# \text{ liked \& shown}}{\# \text{ liked}} = \frac{3}{5}$$

System rekomendujący: sprawność

77

How many recommended items were liked?

Precision
$$\frac{\# \text{ liked \& shown}}{\# \text{ shown}}$$

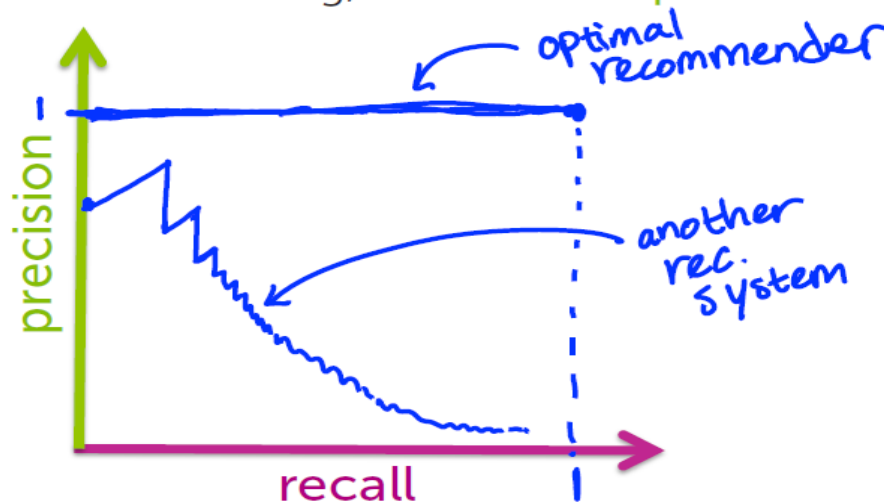
$$= \frac{3}{11}$$

System rekomendujący: sprawność

78

Precision-recall curve

- **Input:** A specific recommender system
- **Output:** Algorithm-specific precision-recall curve
- To draw curve, vary threshold on # items recommended
 - For each setting, calculate the precision and recall

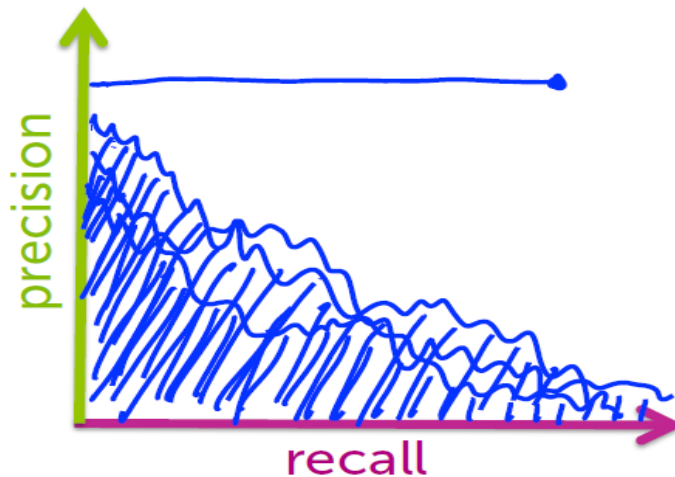


System rekomendujący: sprawność

79

Which Algorithm is Best?

- For a given **precision**, want **recall** as large as possible (or vice versa)
- One metric: largest **area under the curve (AUC)** ★
- Another: set desired recall and maximize precision (precision at k)



Macierz podobieństw: rekomendacja produktu

80

Models

- Collaborative filtering
- Matrix factorization
- PCA

Algorithms

- Coordinate descent
- Eigen decomposition
- SVD

Concepts

- Matrix completion, eigenvalues, random projections, cold-start problem, diversity, scaling up