TEORETYCZNE PODSTAWY INFORMATYKI

27/01/2020

WFAiS UJ, Informatyka Stosowana
I rok studiów, I stopień

Wykład 13c

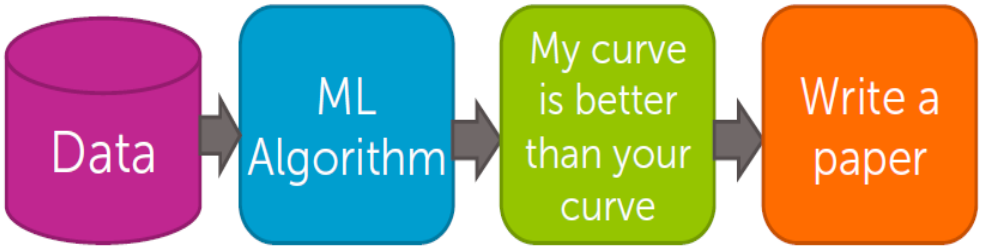
Data Science: Uczenie maszynowe

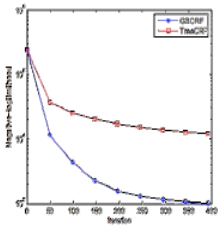
- Uczenie maszynowe: co to znaczy?
- Metody
  - Regresja
  - Klasyfikacja
  - Klastering i wybór podzbioru
  - System do rekomendacji

Wykład na podstawie materiałów:
E. Fox & C. Guestrin, Univeristy of Waschington, „Machine Learning Specialization”
http://www.coursera.org./learn/ml-foundations

Uczenie maszynowe

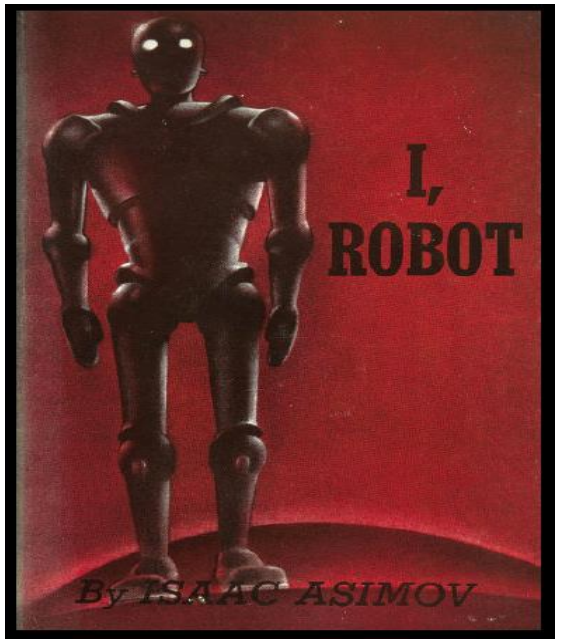
- **Metody uczenia maszynowego (machine learning) rewolucjonizują obecnie podejście do różnych problemów związanych z analizą danych.**
- **Jeszcze kilka lat temu był to bardziej „akademicki” problem z zakresu numeryki i algorytmiki.**





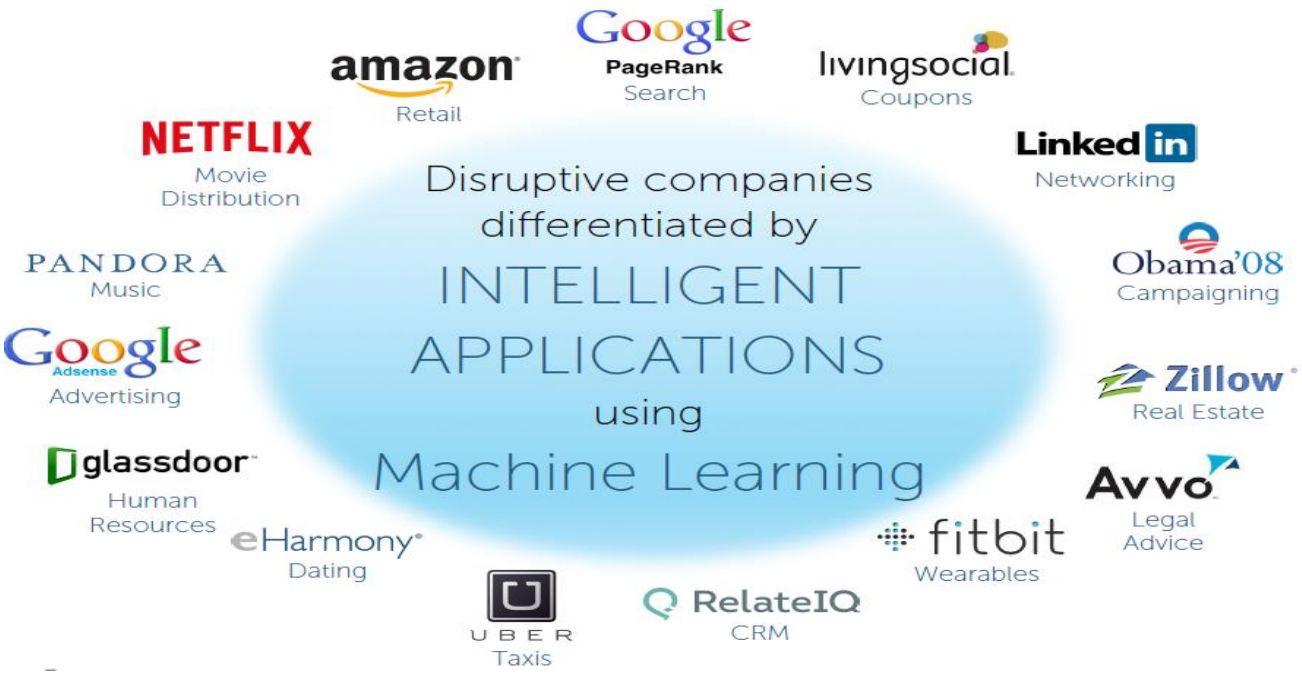
Neural Information Processing Systems Foundation

ICML



Uczenie maszynowe

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



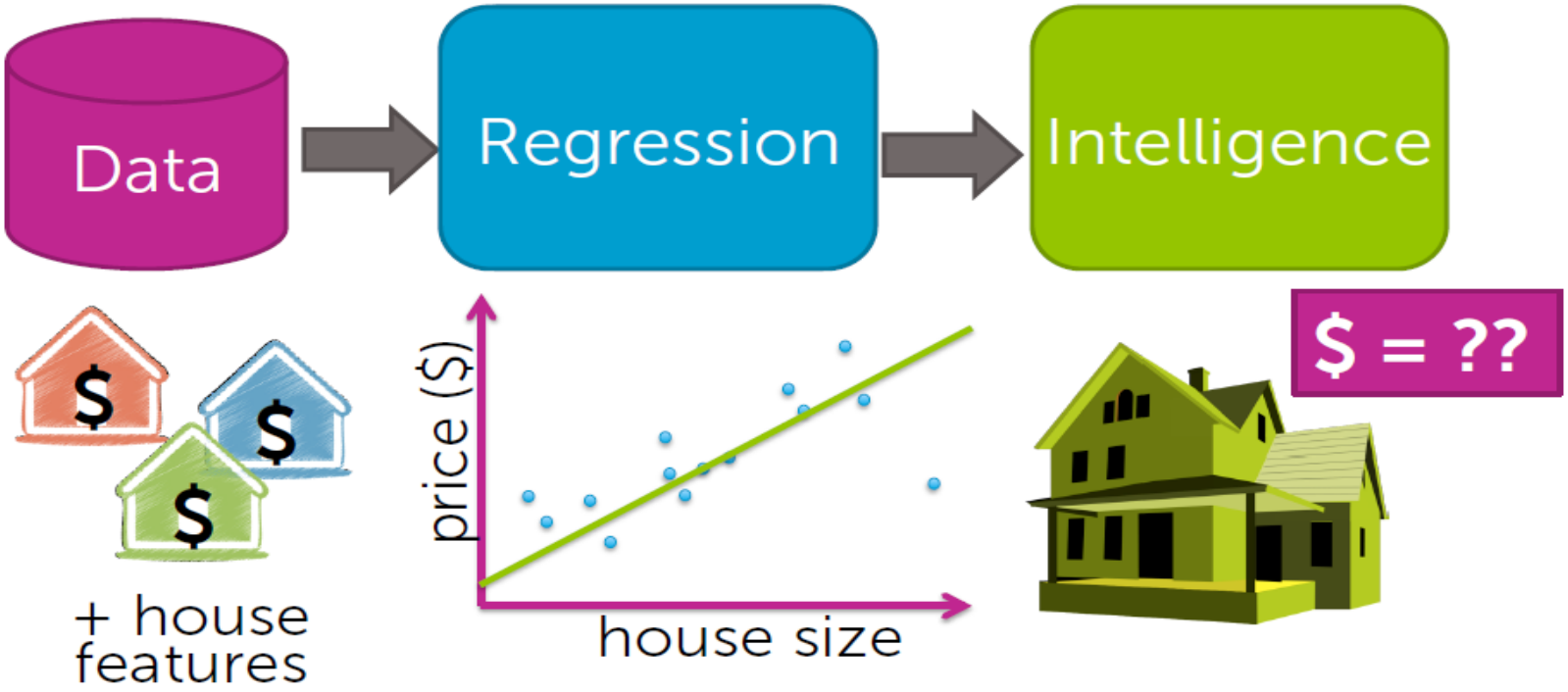
Uczenie maszynowe: cel

- 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

- Przewidywanie ceny domu na podstawie zebranych danych dotyczących ceny innych

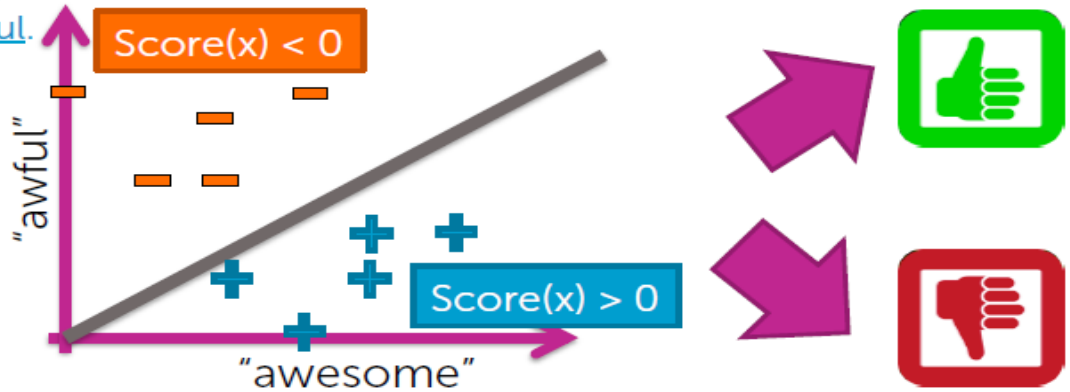


Uczenie maszynowe: przykład

- Ranking restauracji

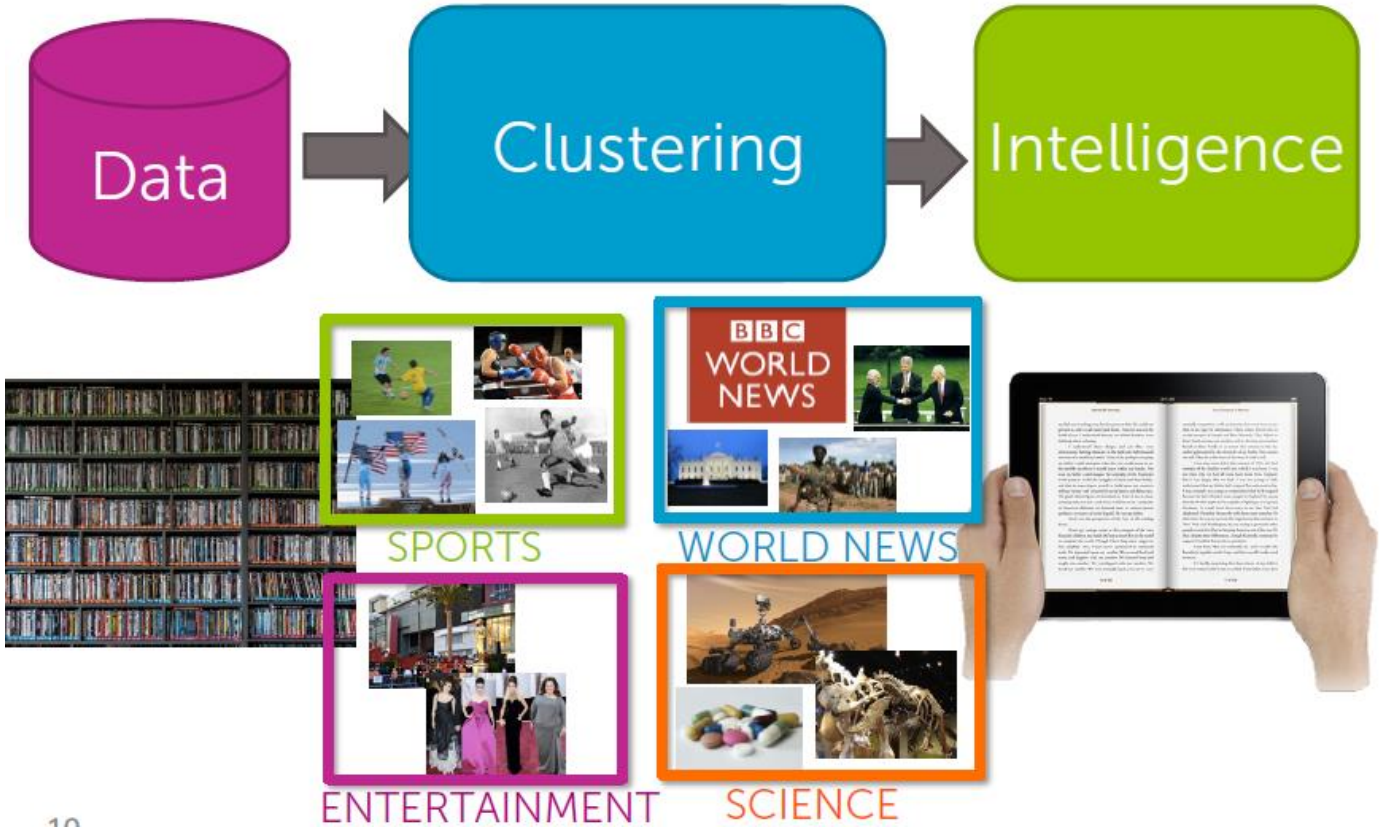
Sushi was awesome,
the food was awesome,
but the service was awful.

All reviews:



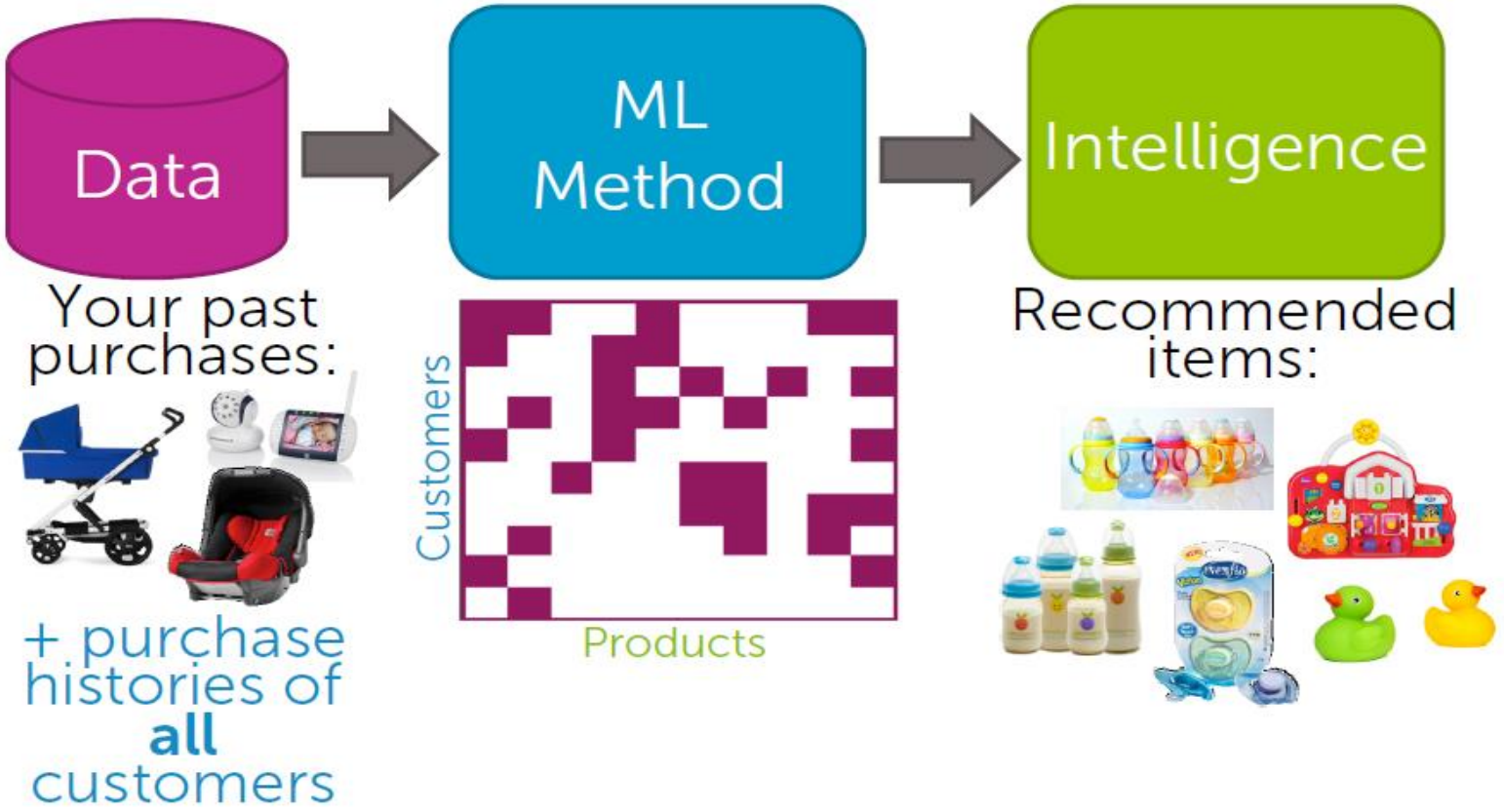
Uczenie maszynowe: przykład

- Znajdowanie podobnych dokumentów



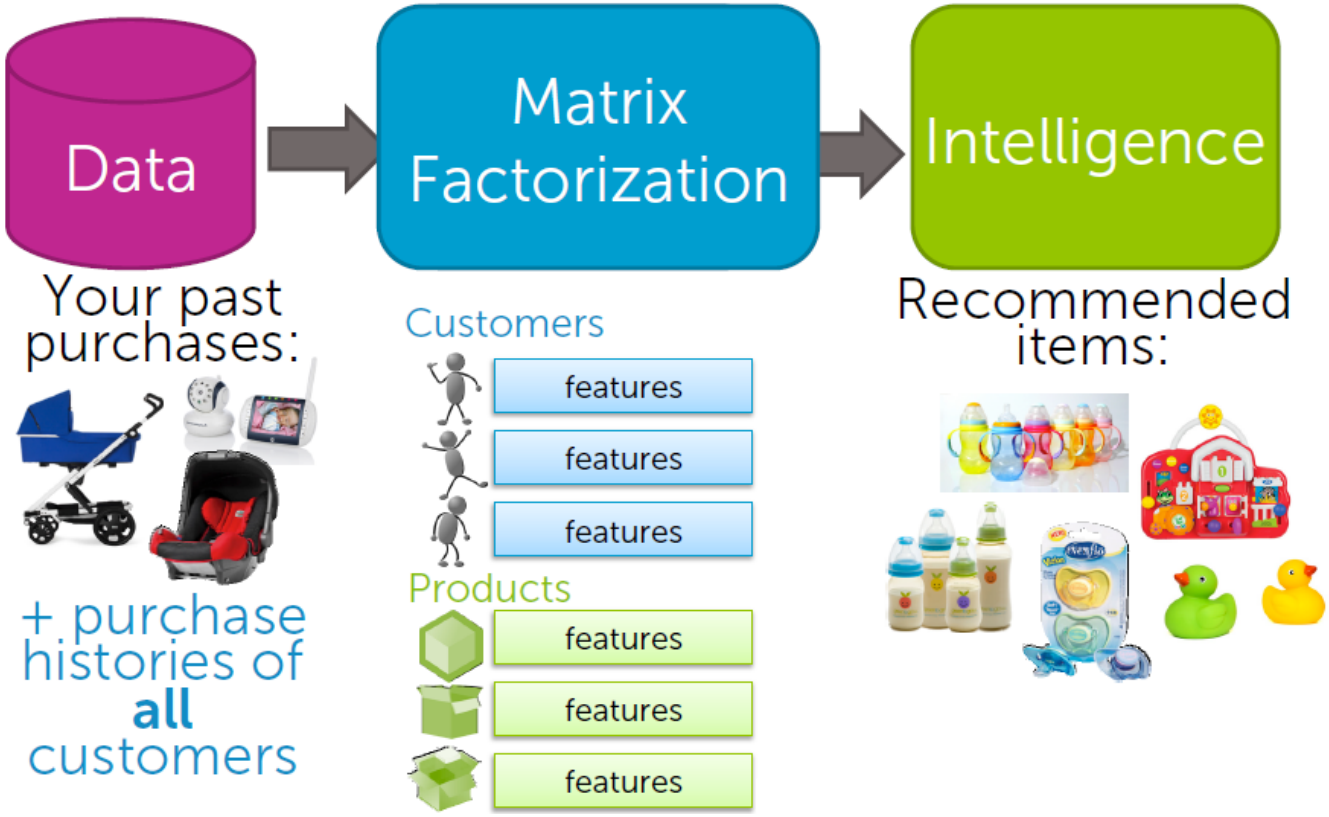
Uczenie maszynowe: przykład

- Rekomendowanie podobnego produktu



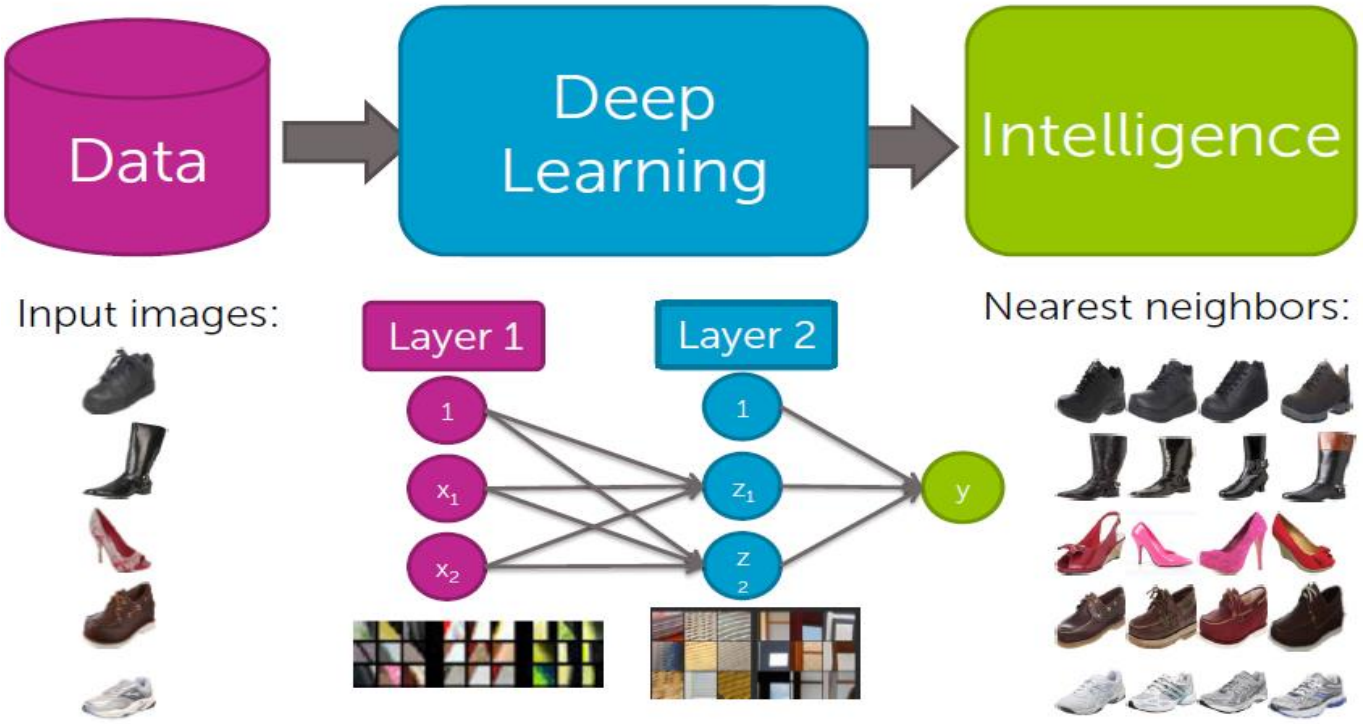
Uczenie maszynowe: przykład

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



Uczenie maszynowe: przykład

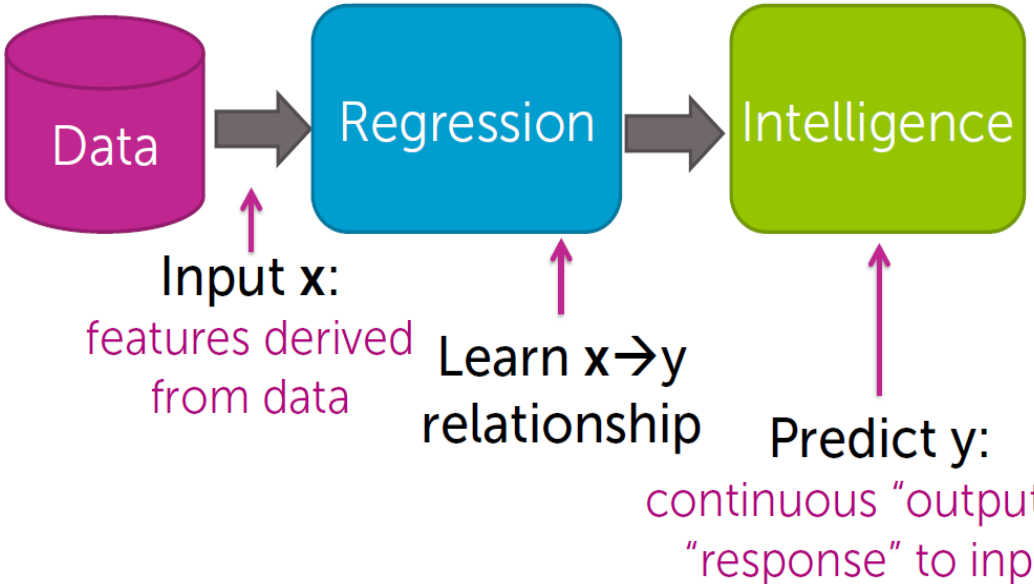
- **Rekomendowanie podobnego produktu: na podstawie jego charakterystyki graficznej**



Regresja

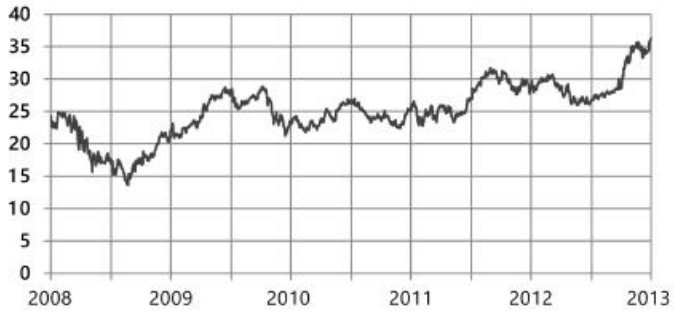
Uczenie maszynowe: Regresja

- Przewidywanie odpowiedzi na podstawie informacji wejściowej



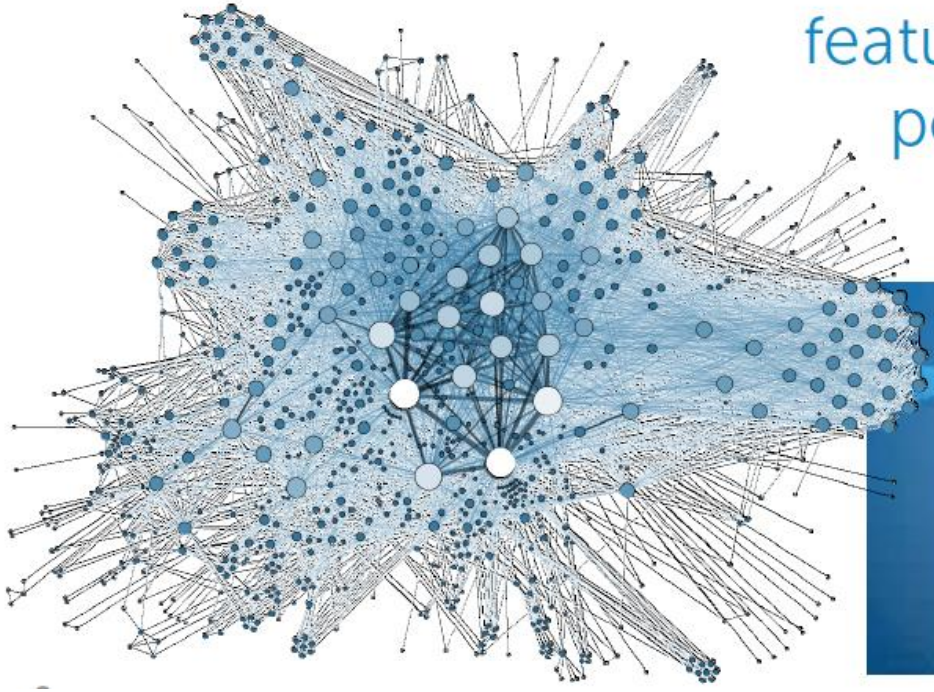
Cena akcji na giełdzie

- Predict the price of a stock (y)
- Depends on **x** =
  - Recent history of stock price
  - News events
  - Related commodities

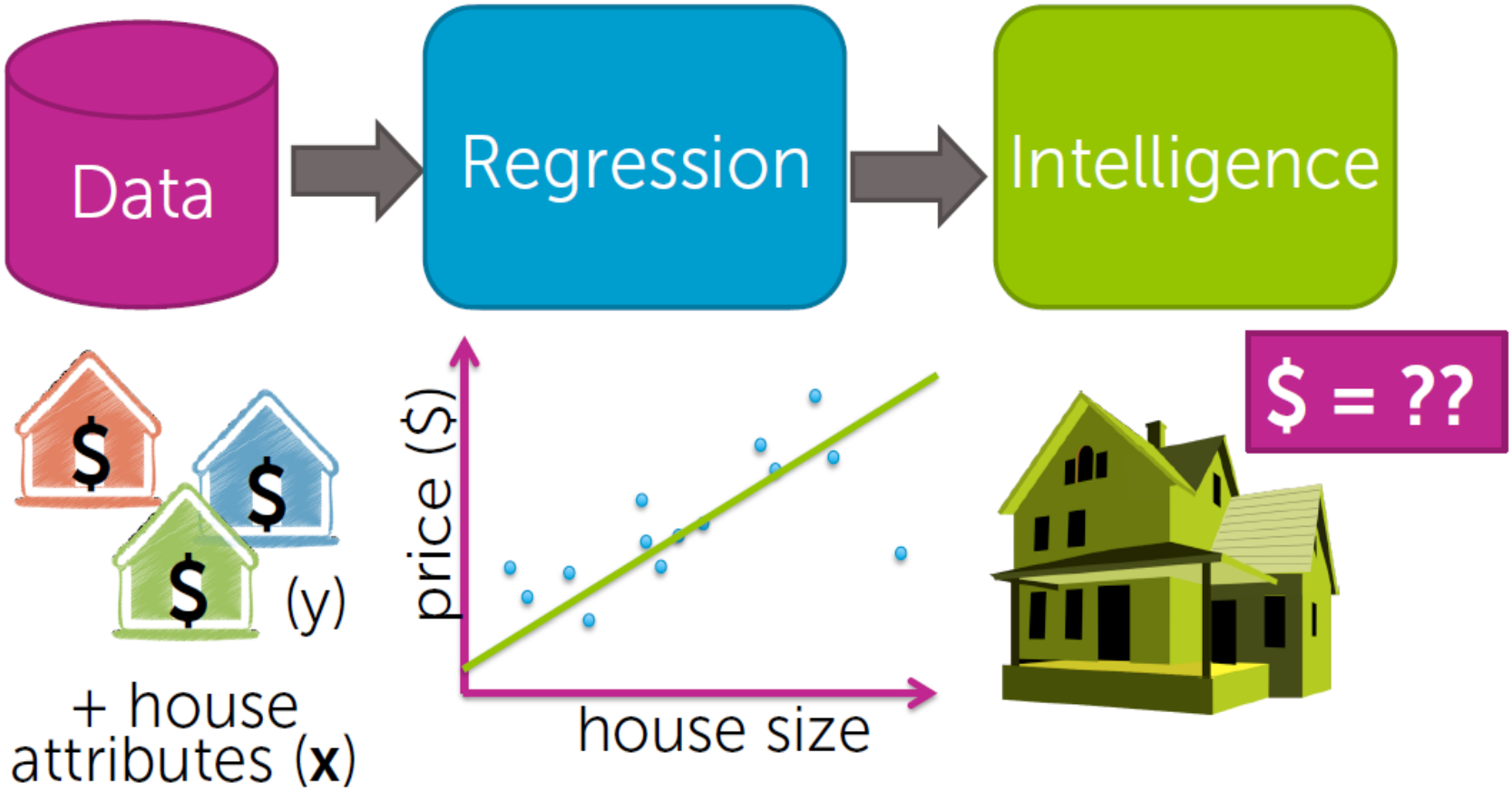


Tweet popularność

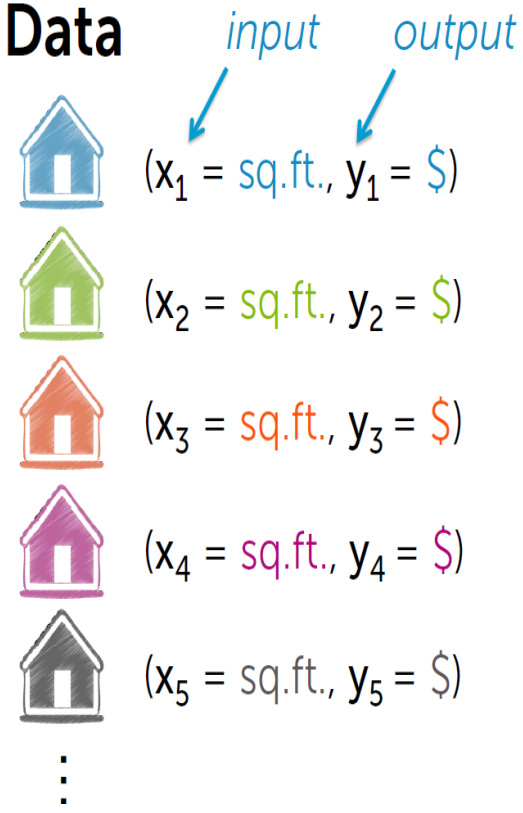
- How many people will retweet your tweet? (y)
- Depends on **x** = # followers,
  # of followers of followers,
  features of text tweeted,
  popularity of hashtag,
  # of past retweets,...

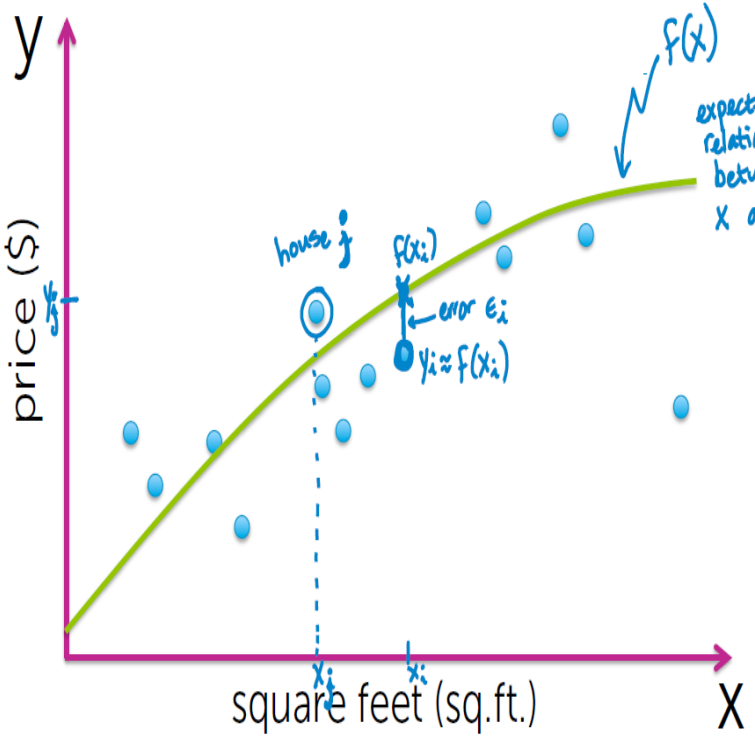


Przykład: przewidywana cena domu



Data, model





Regression model:

$y_i = f(x_i) + \epsilon_i$

$E[\epsilon_i] = 0$ ← equally likely that error is + or −

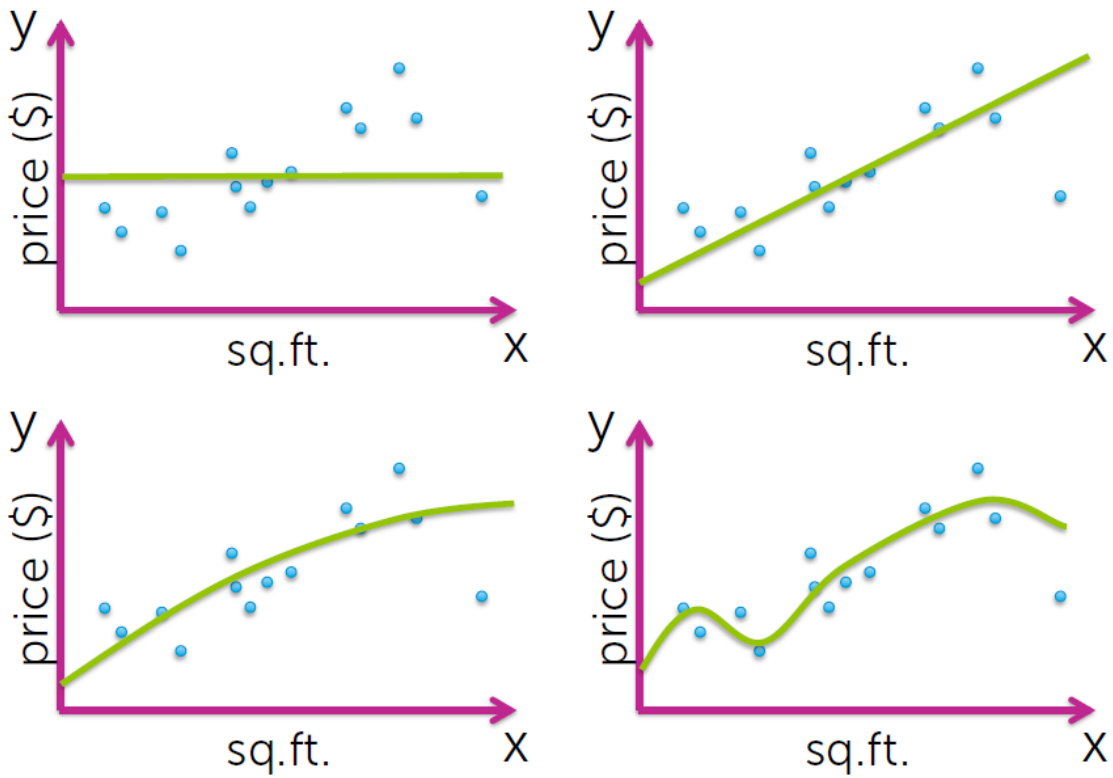
↑ expected value

⇓

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

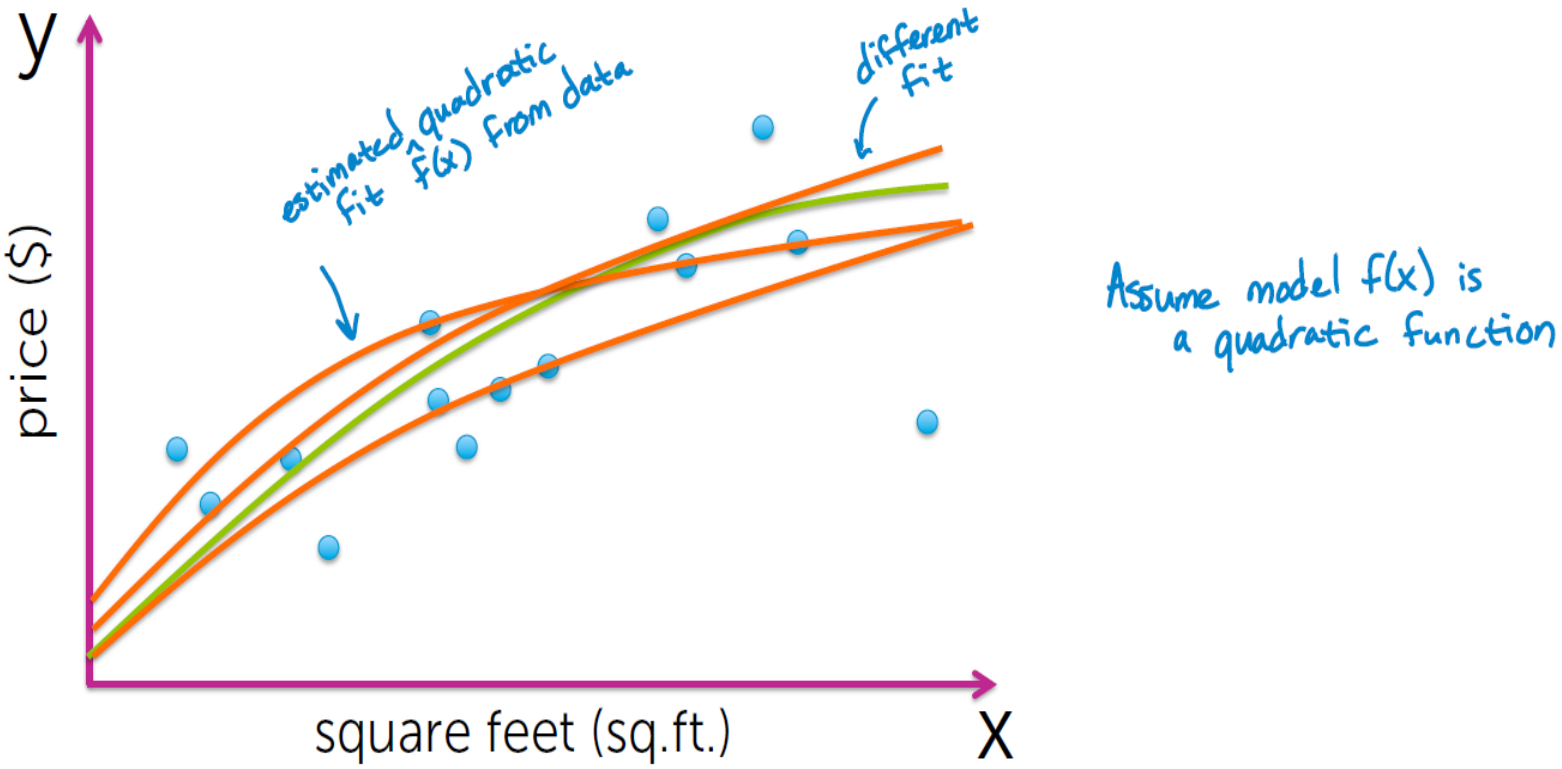
Data, model

- Jaki model dla f(x)?

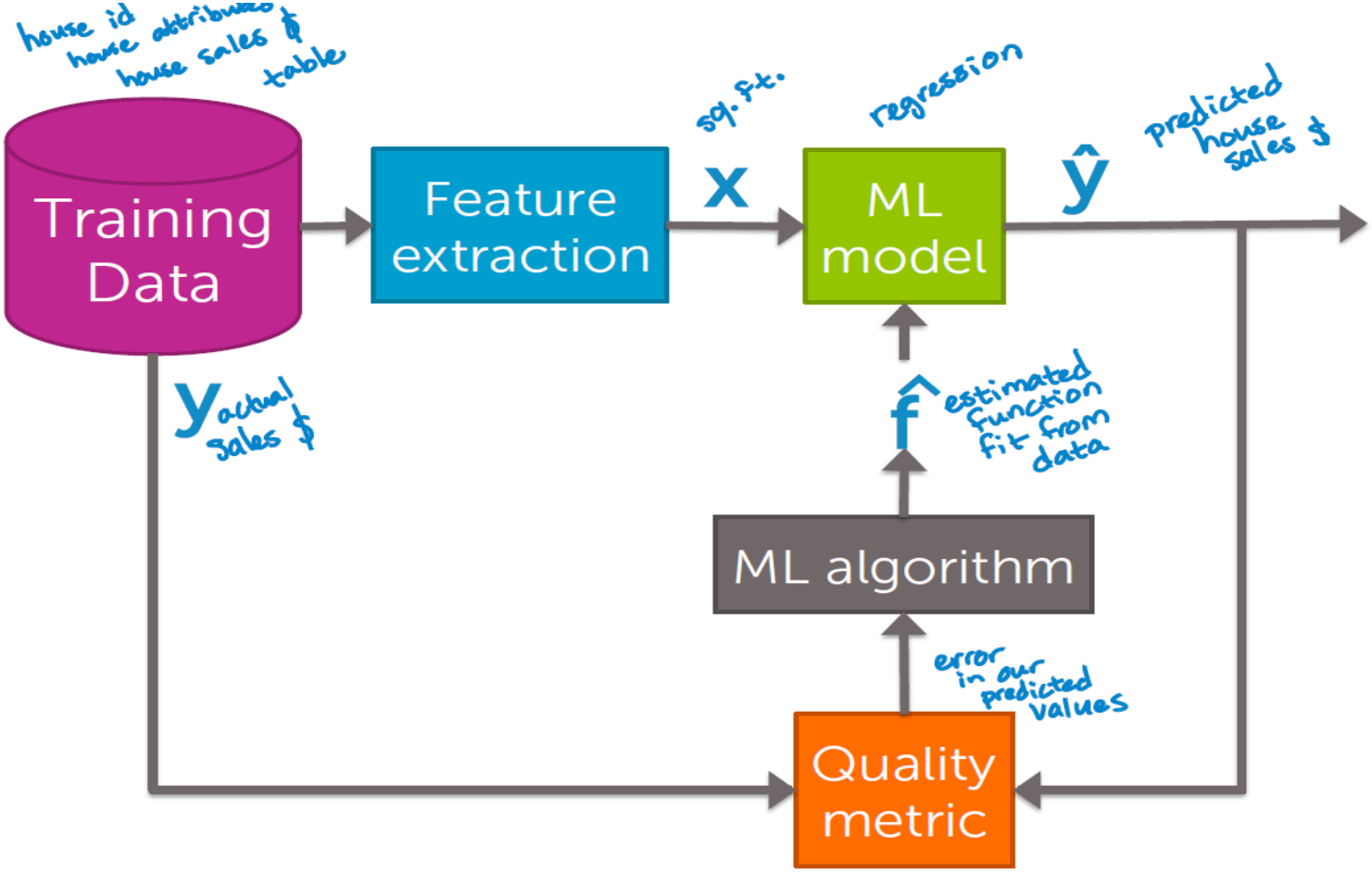


Przewidywanie

- A potem chcielibyśmy przewidzieć odpowiedź



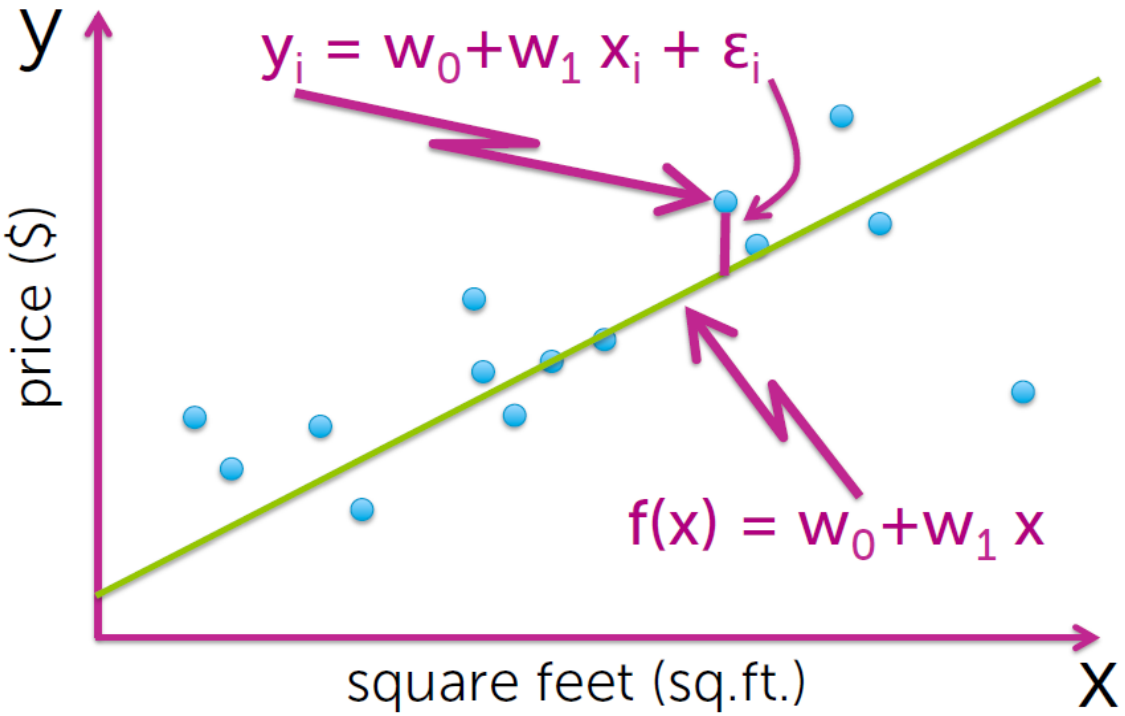
Pętla iteracyjna?



Simple linear regression model

Co to znaczy „simple”?

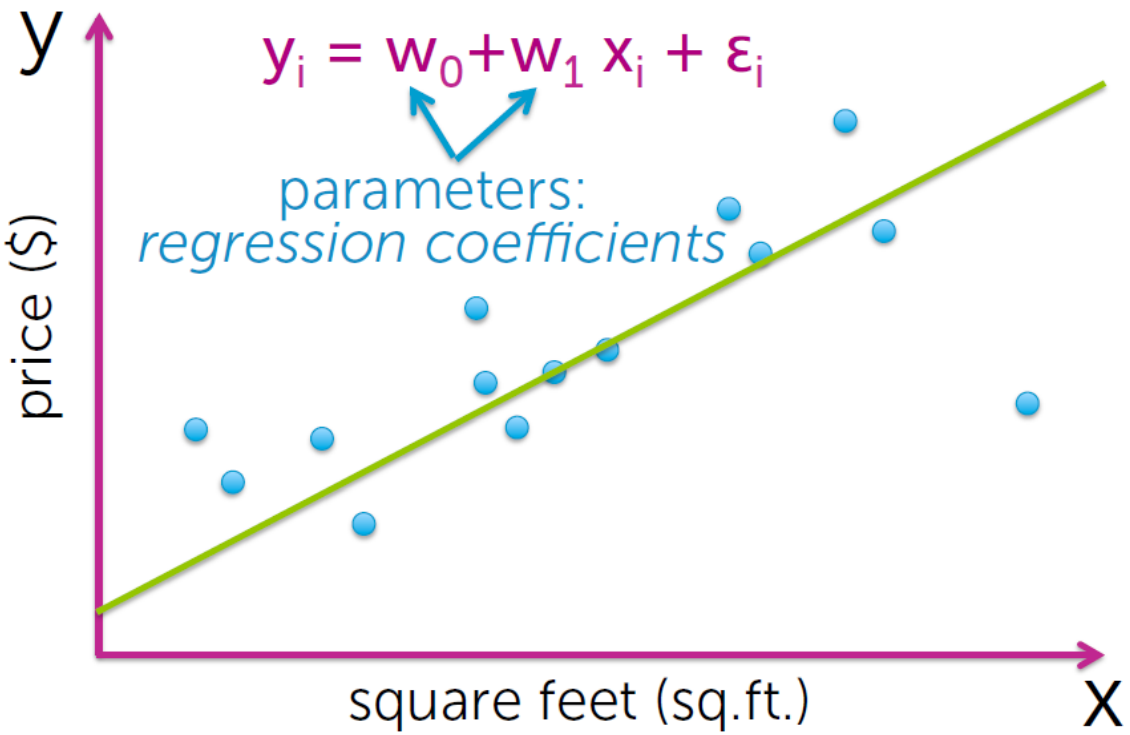
1 input x, fitujemy zależność liniową do danych



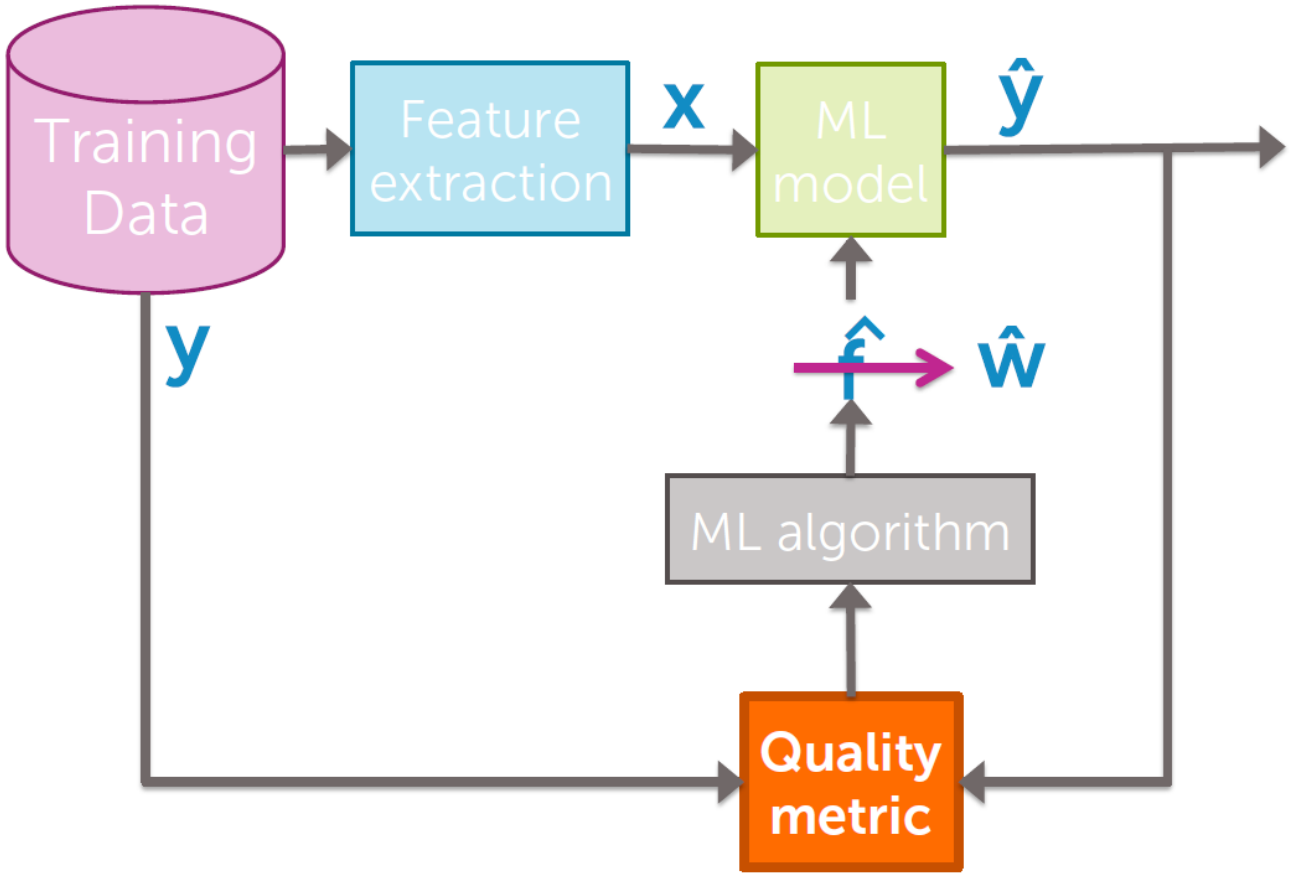
Simple linear regression model

Co to znaczy „simple”?

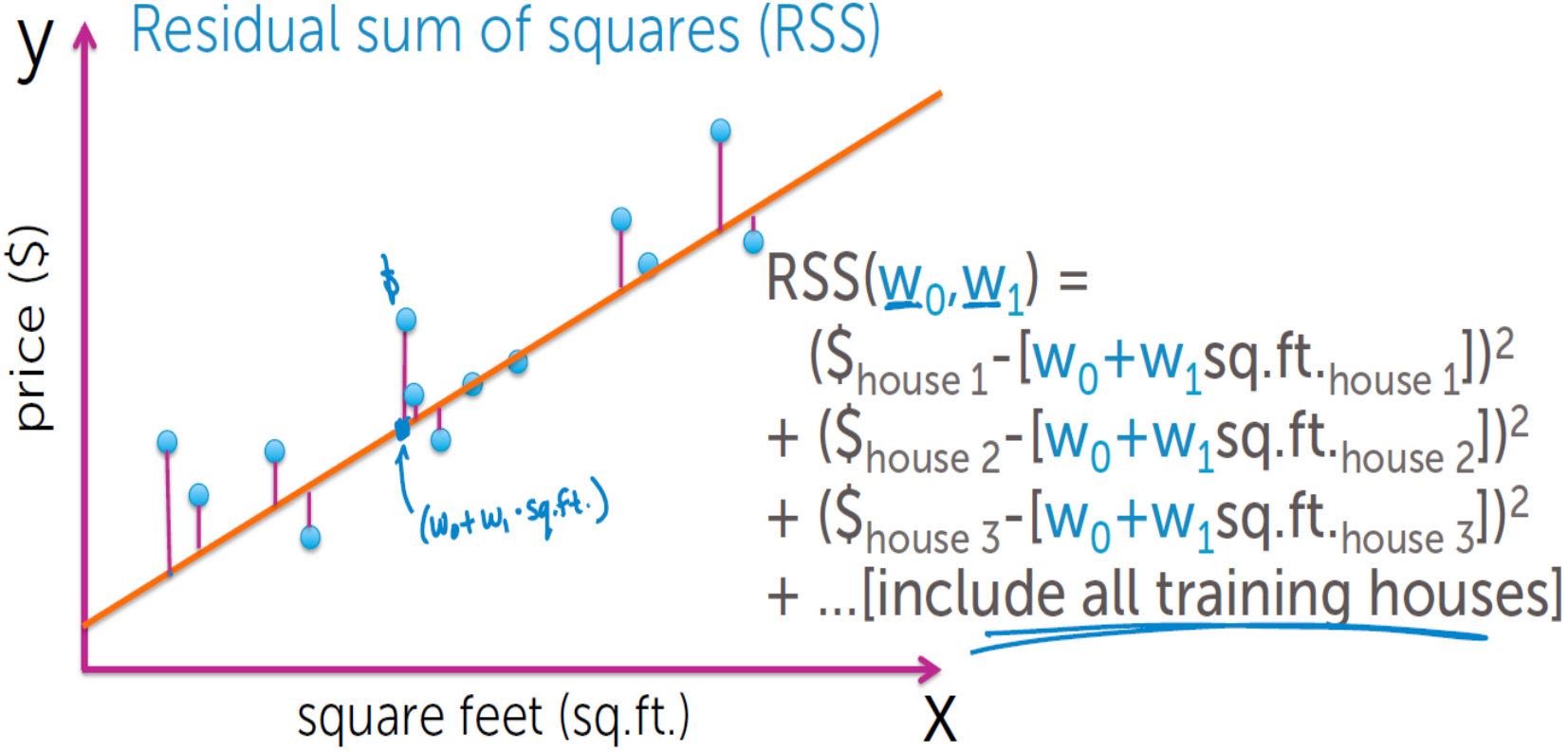
1 input x, fitujemy zależność liniową do danych



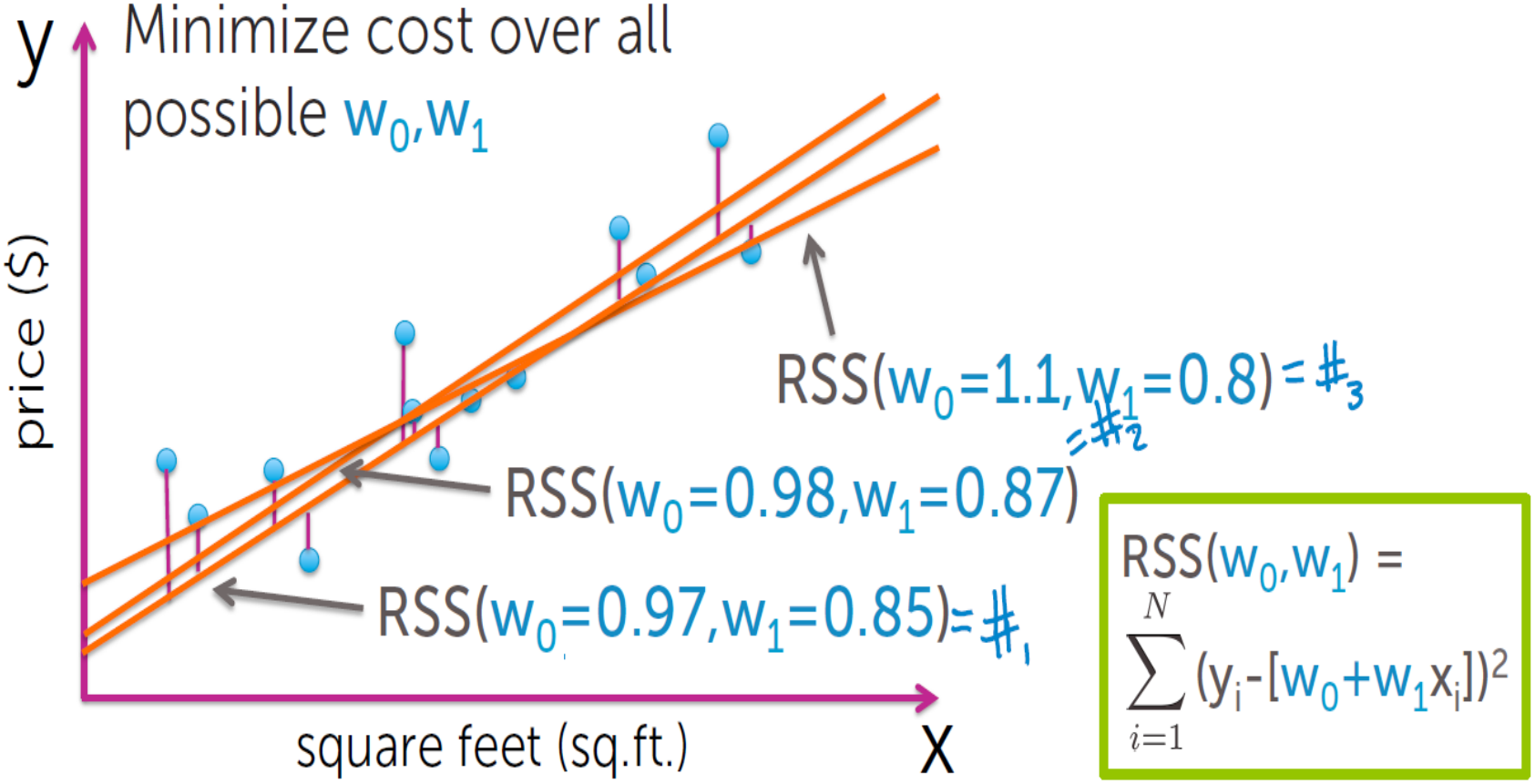
Pętla iteracyjna



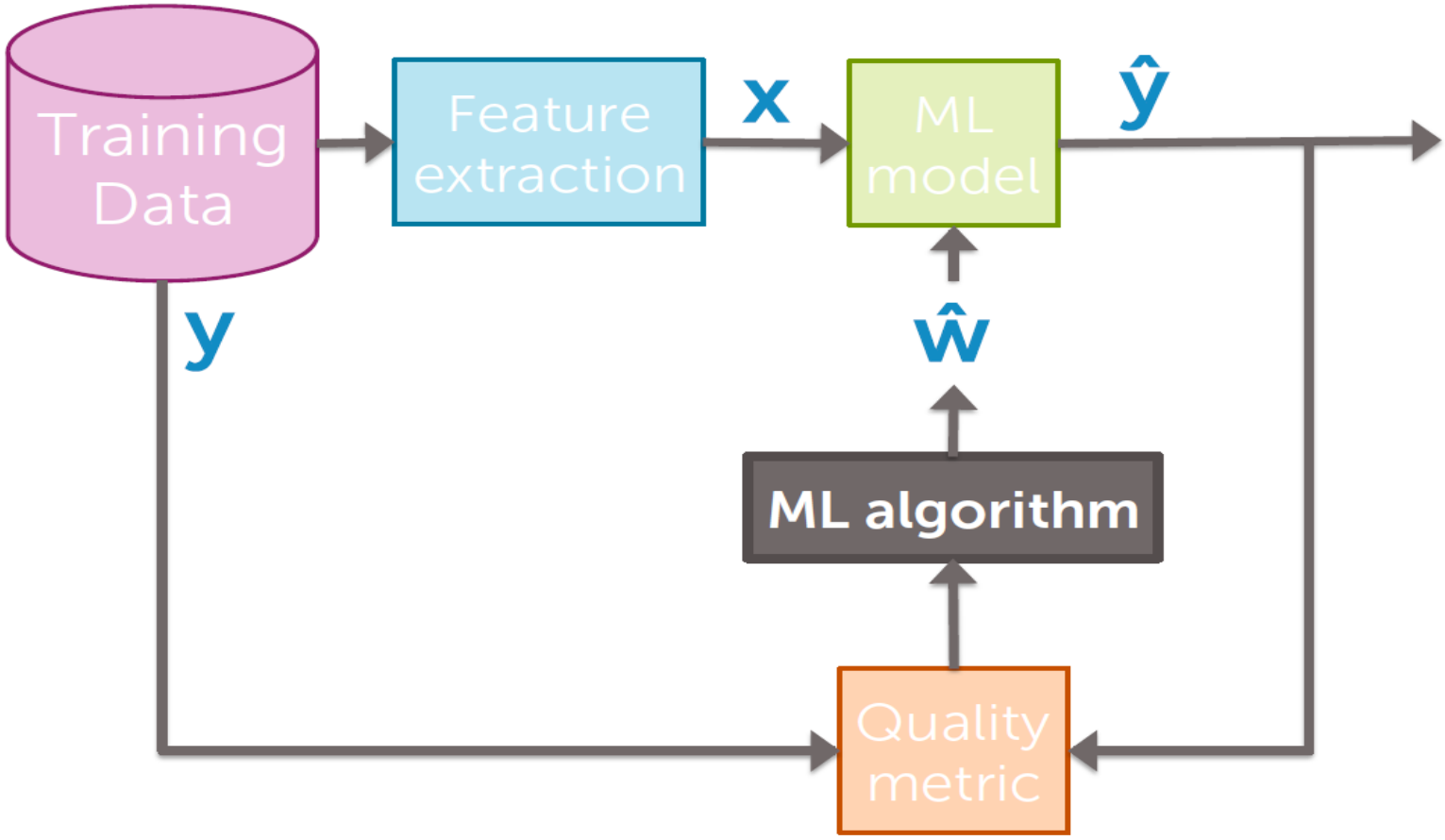
Funkcja „kosztu”



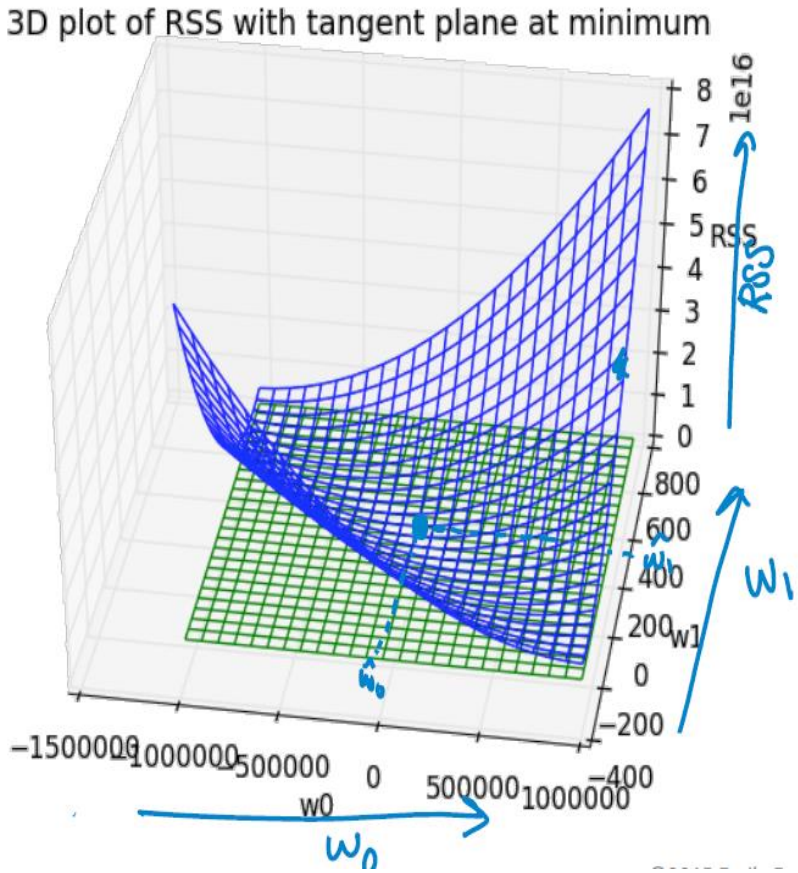
Minimalizacja funkcji „kosztu”



Pętla iteracyjna



Minimalizacja funkcji „kosztu”

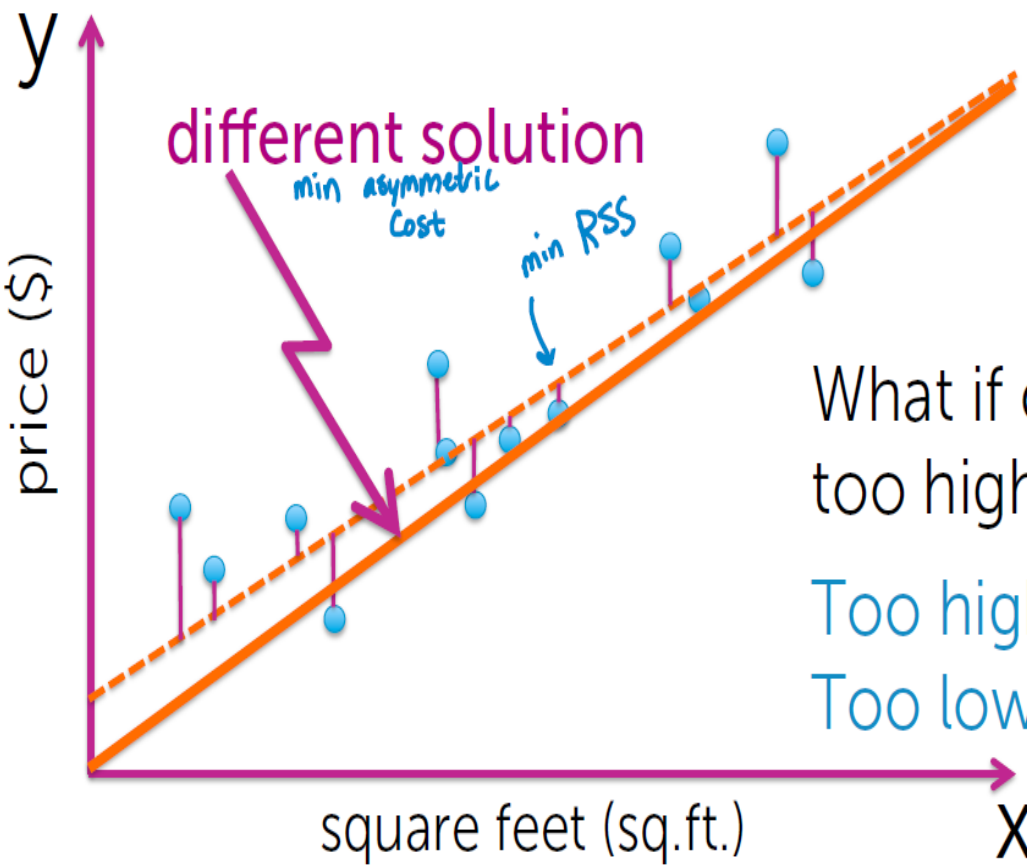


Minimize function over all possible w_0, w_1

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

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

Asymetryczny błąd kosztu

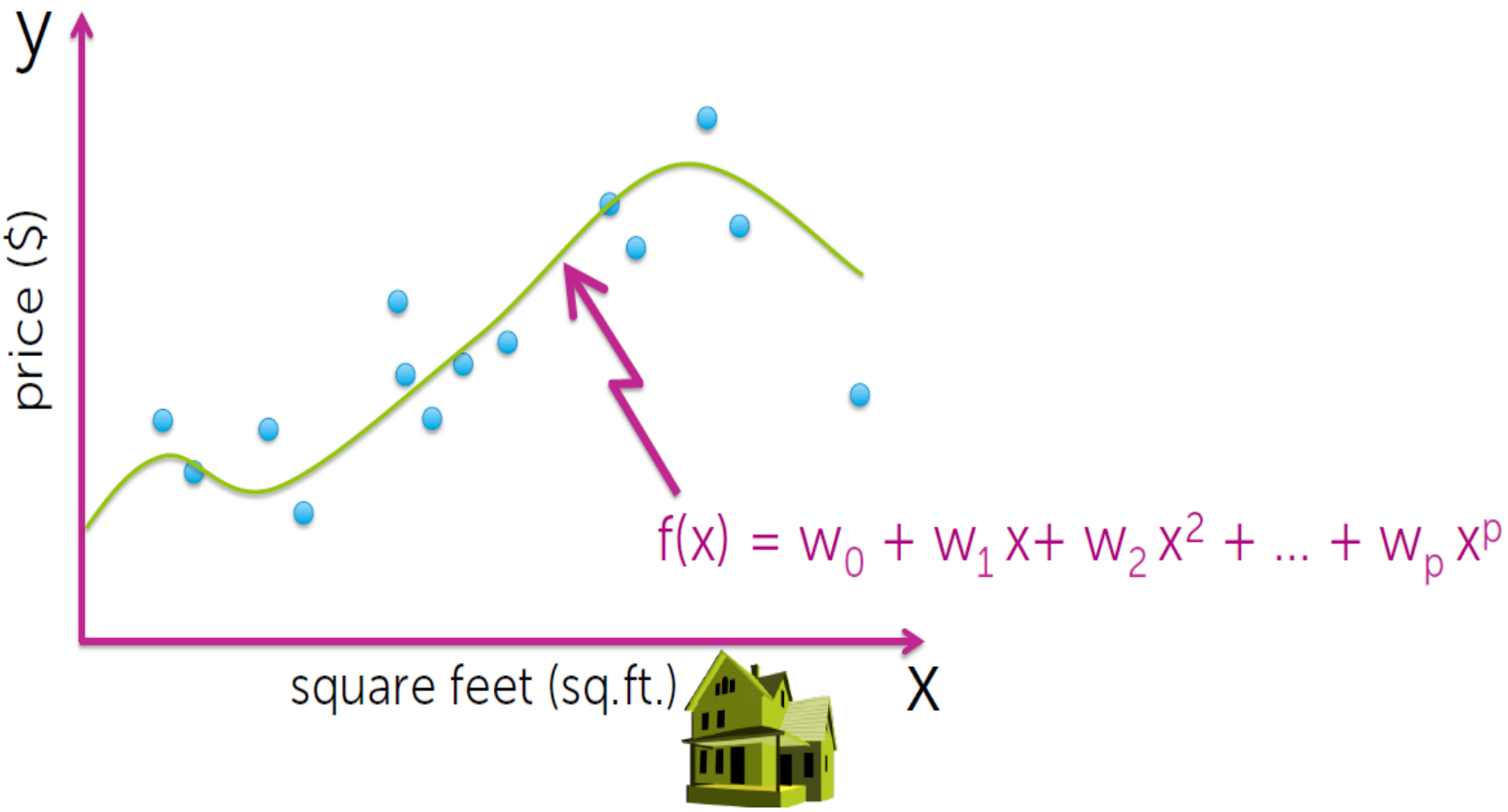


What if cost of listing house too high has **bigger cost**?

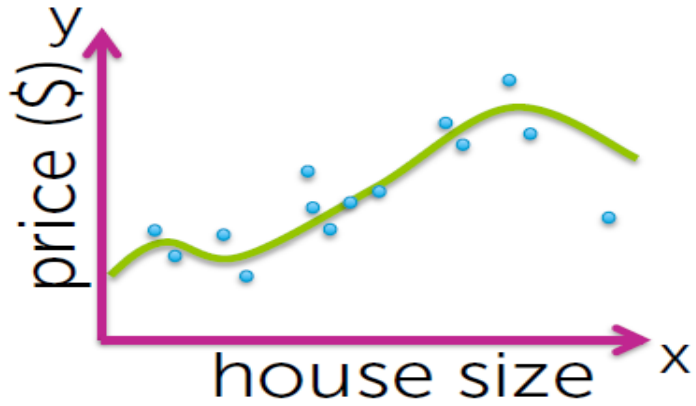
Too high → no offers ($=0)

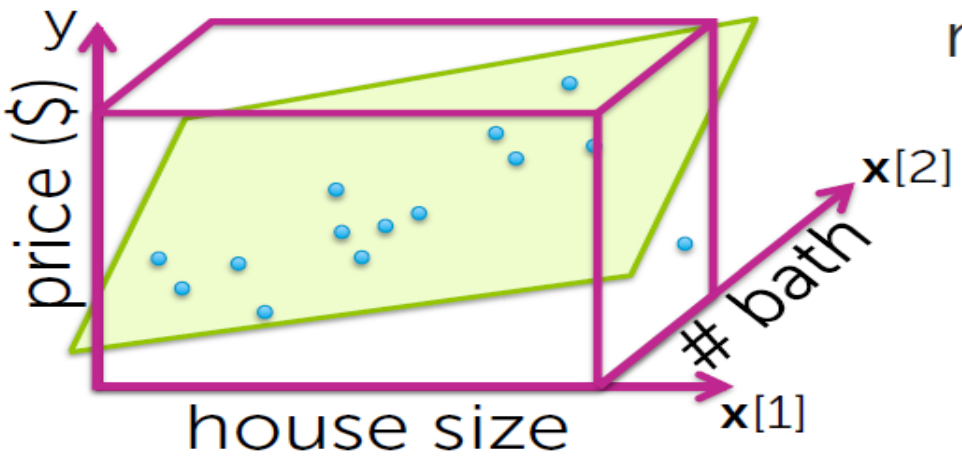
Too low → offers for lower $

Polynomial regression



Multiple regression





Fit more complex relationships than just a line

Incorporate more inputs

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

Regresja: przewidywanie na podstawie danych

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

Uczenie maszynowe: Klasyfikacja

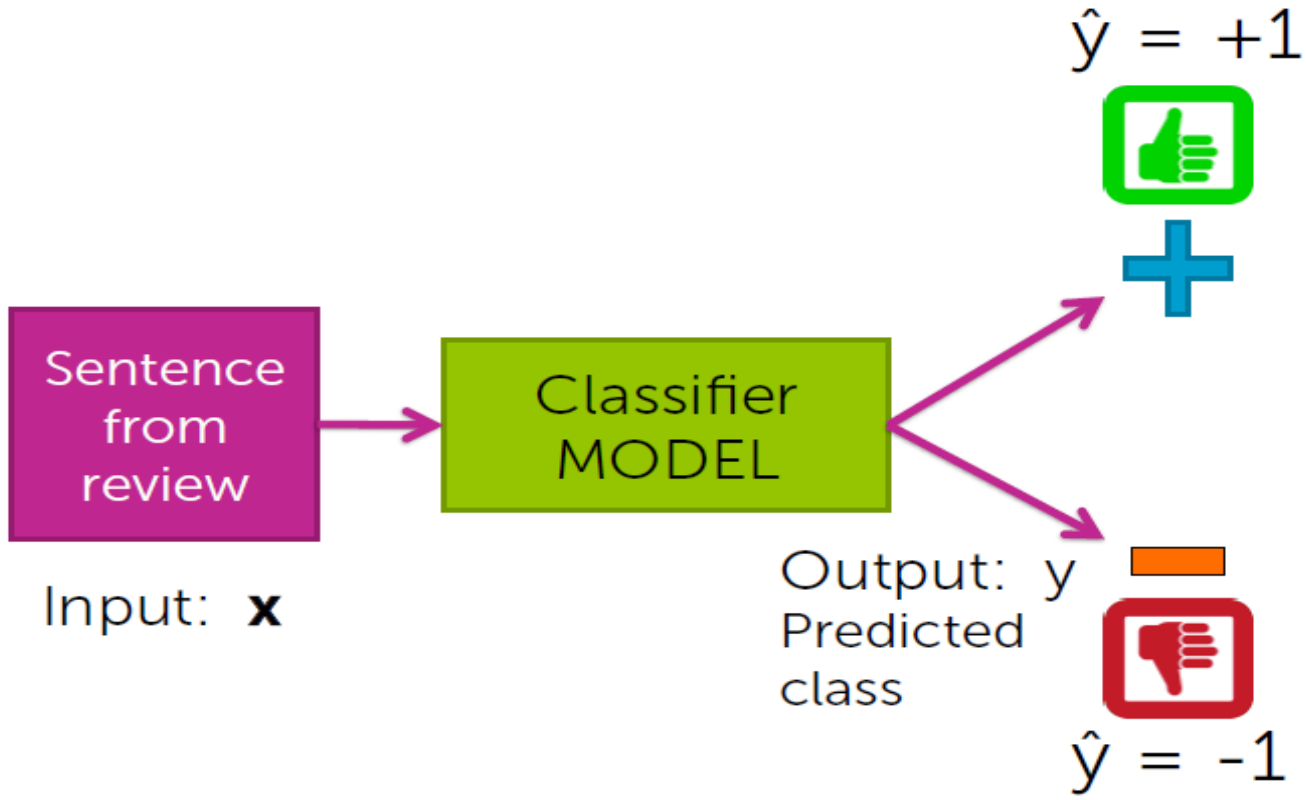
- Inteligentny system rankingu restauracji

Klasyfikacja

Uczenie maszynowe: Klasyfikacja

- Inteligentny system rankingu restauracji

Prosta klasyfikacja



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

Dane trenujące

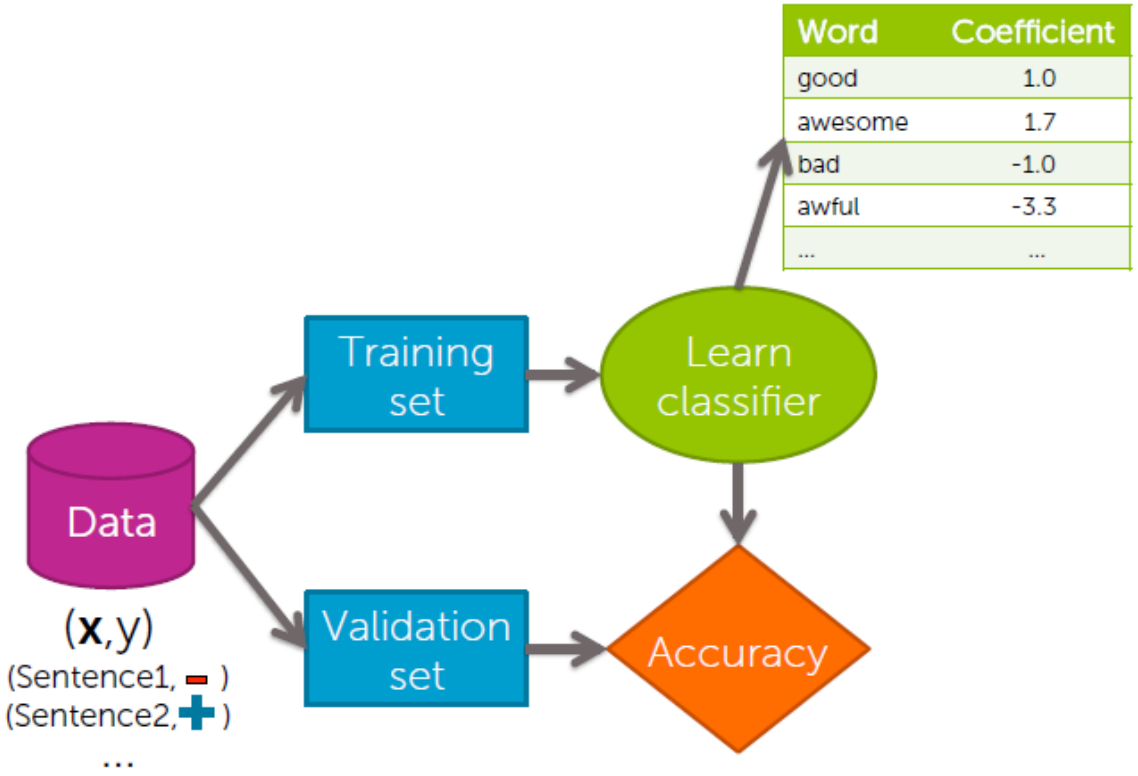
- Używamy danych trenujących aby przypisać prawdopodobieństwo (wagę) dla każdego słowa

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
...	...

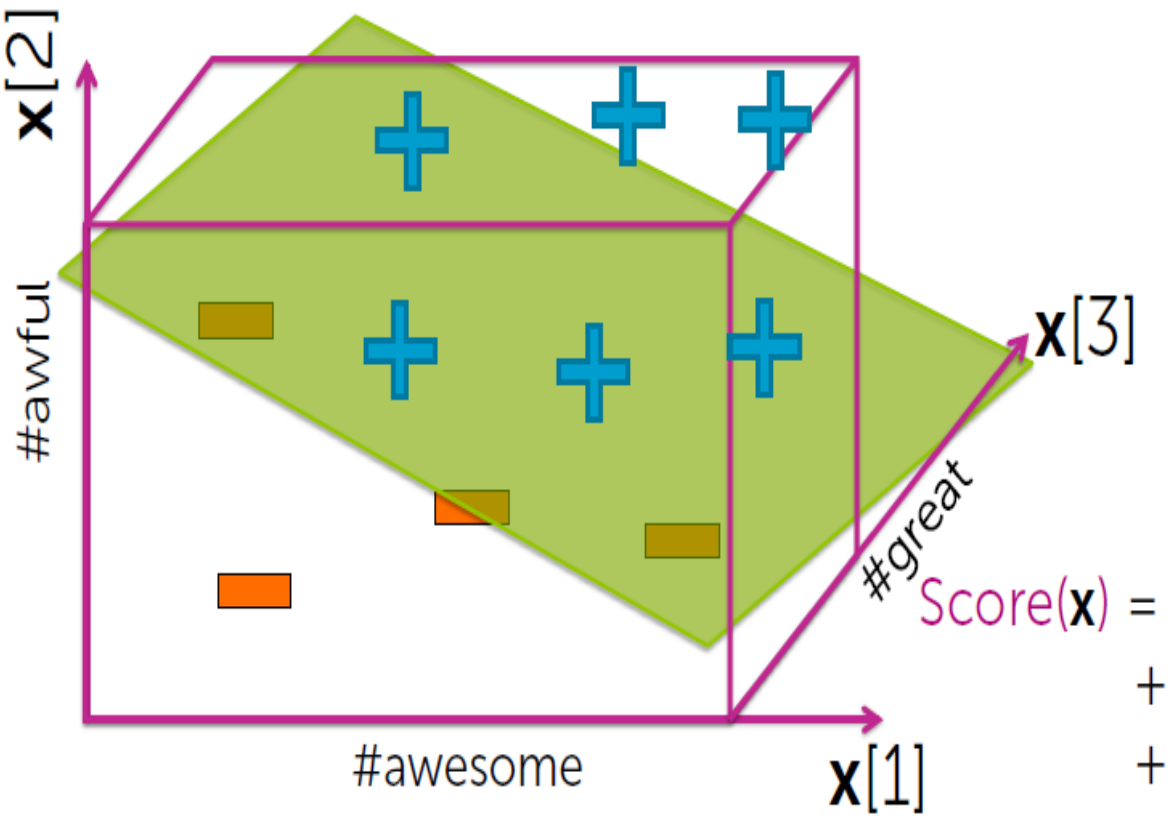
- Waga całego zdania (score) to będzie prosta suma tych wag

Uczenie klasyfikatora

Training a classifier = Learning the coefficients



Jak wygląda nasz model

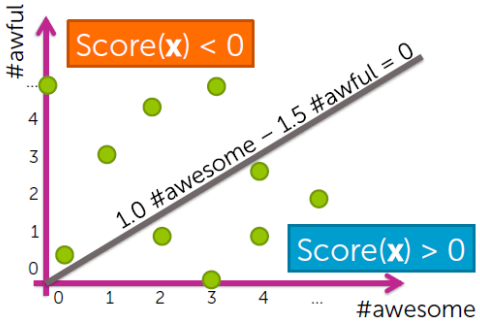


$$\text{Score}(\mathbf{x}) = w_0 + w_1 \#\text{awesome} + w_2 \#\text{awful} + w_3 \#\text{great}$$

Zmiana współczynników

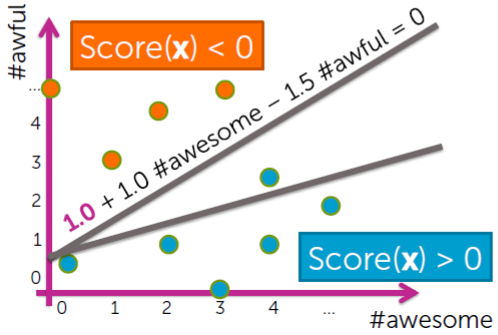
Input	Coefficient	Value
	w_0	0.0
#awesome	w_1	1.0
#awful	w_2	-1.5

Score(x) = 1.0 #awesome – 1.5 #awful

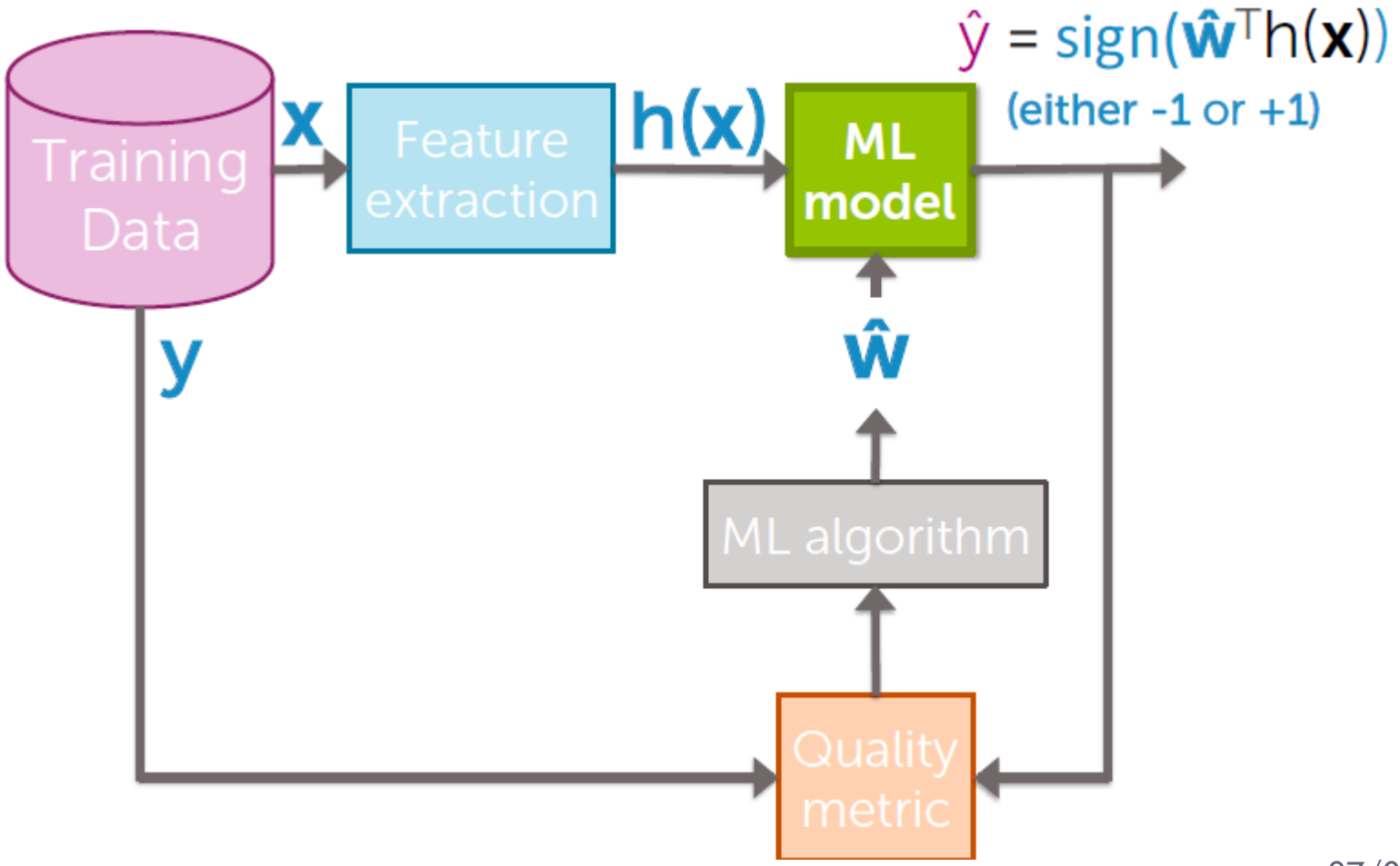


Input	Coefficient	Value
	w_0	1.0
#awesome	w_1	1.0
#awful	w_2	–3.0

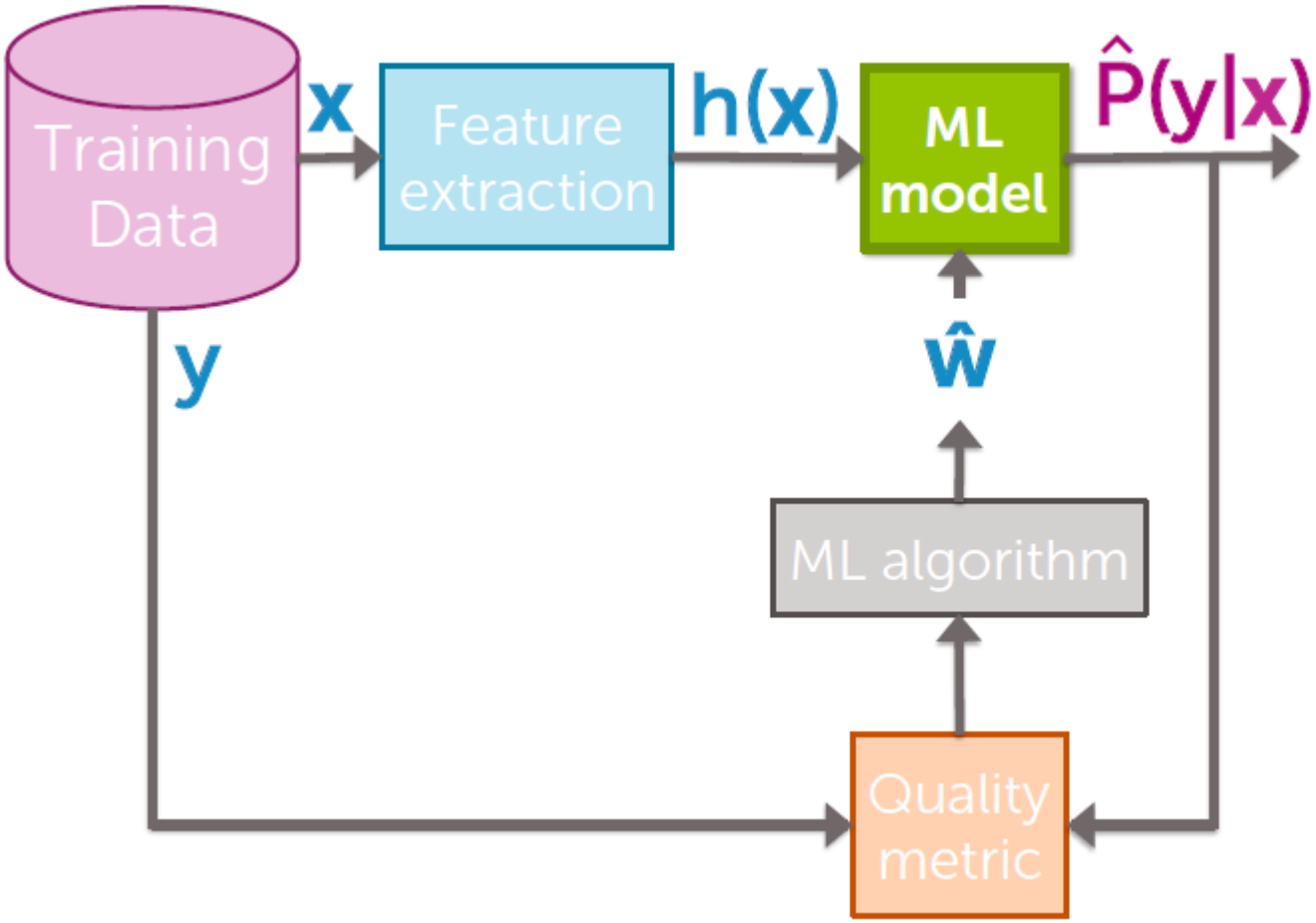
Score(x) = 1.0 + 1.0 #awesome – 3.0 #awful



Jak wygląda nasz model?

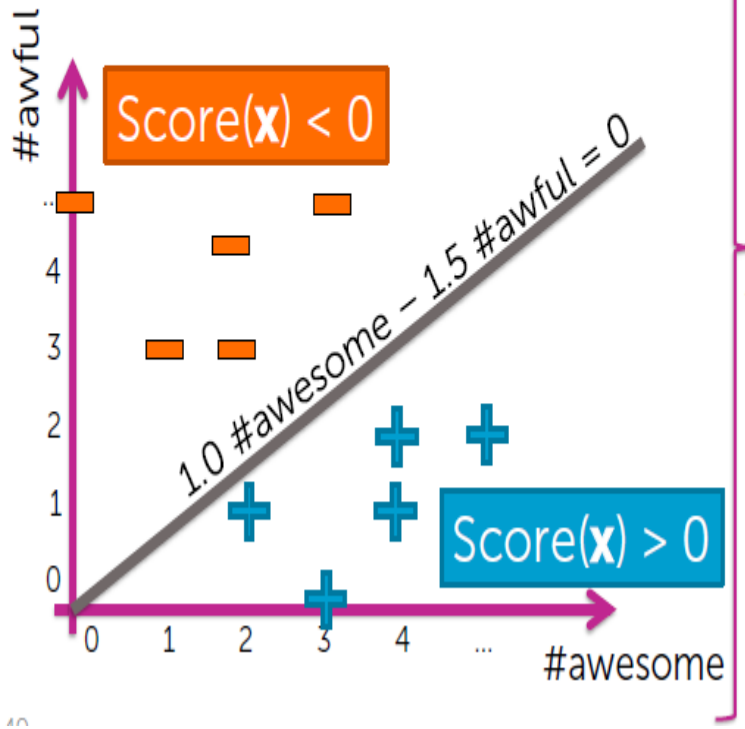


Czy jesteś pewny swojej klasyfikacji

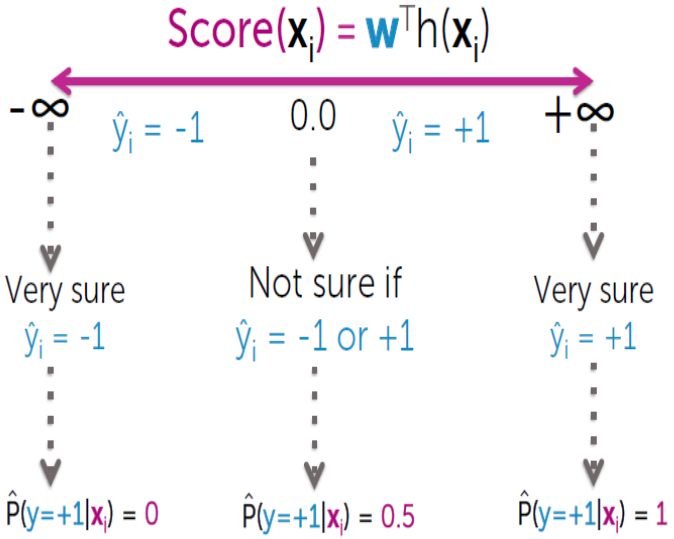


Interpretacja

$$\text{Score}(\mathbf{x}_i) = w_0 h_0(\mathbf{x}_i) + w_1 h_1(\mathbf{x}_i) + \ldots + w_D h_D(\mathbf{x}_i) = \mathbf{w}^T h(\mathbf{x}_i)$$

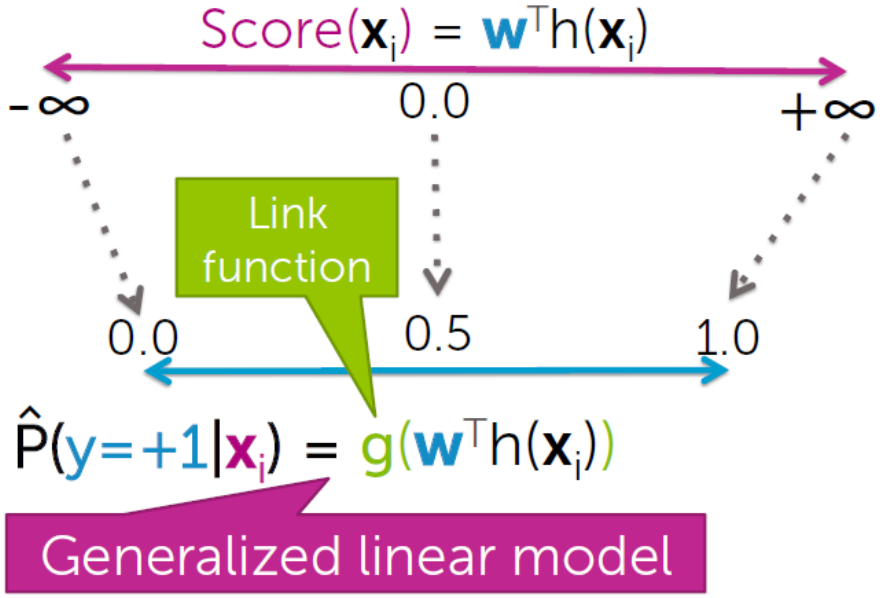


Relate $\text{Score}(\mathbf{x}_i)$ to $\hat{P}(y=+1|\mathbf{x},\hat{\mathbf{w}})$?



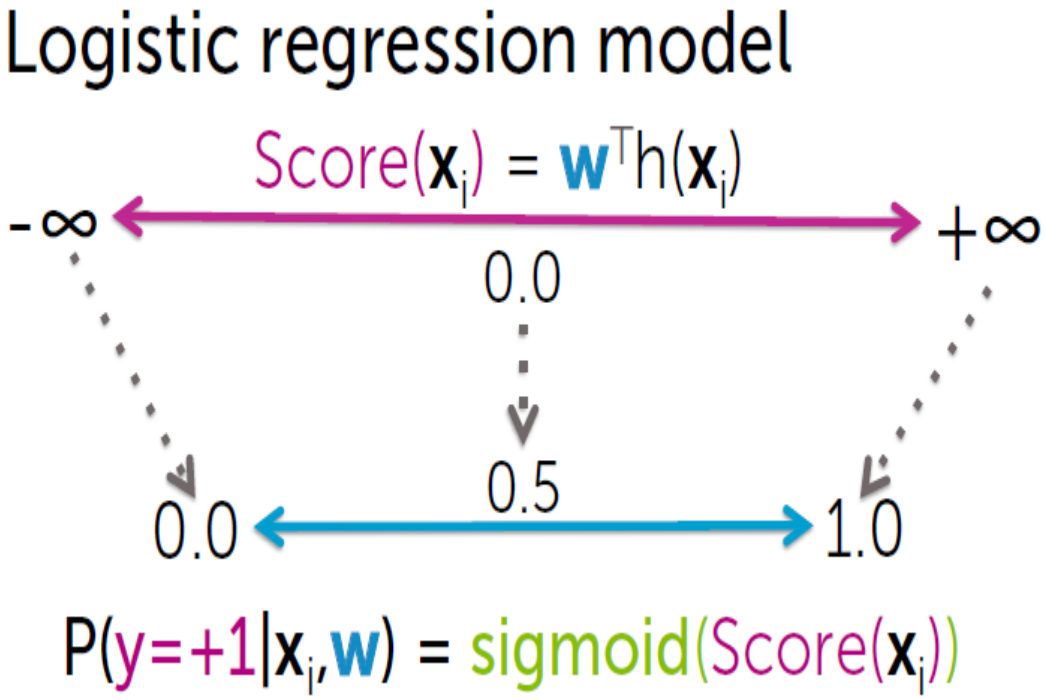
Interpretacja

- Link funkcja

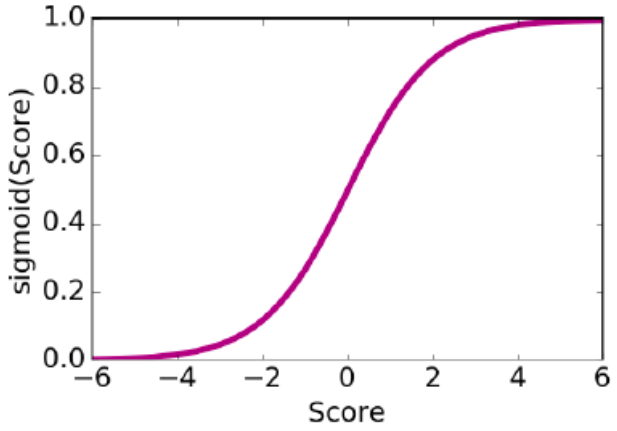


Logistic regression model

- Sigmoid funkcja



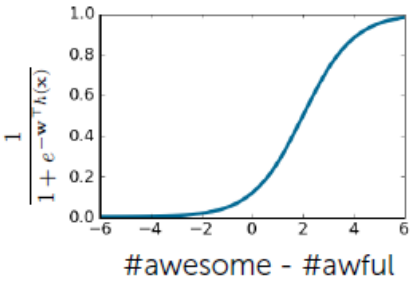
$$sigmoid(\text{Score}) = \frac{1}{1 + e^{-\text{Score}}}$$



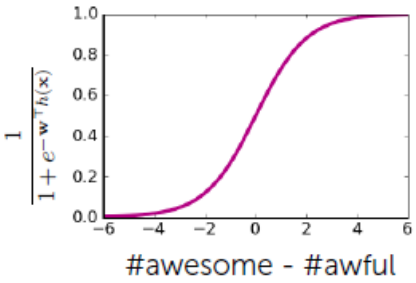
Logistic regression model

- Effekt współczynników na kształt funkcji sigmoidalnej

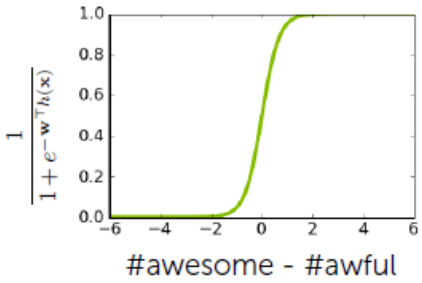
w_0	-2
$w_{\#awesome}$	+1
$w_{\#awful}$	-1



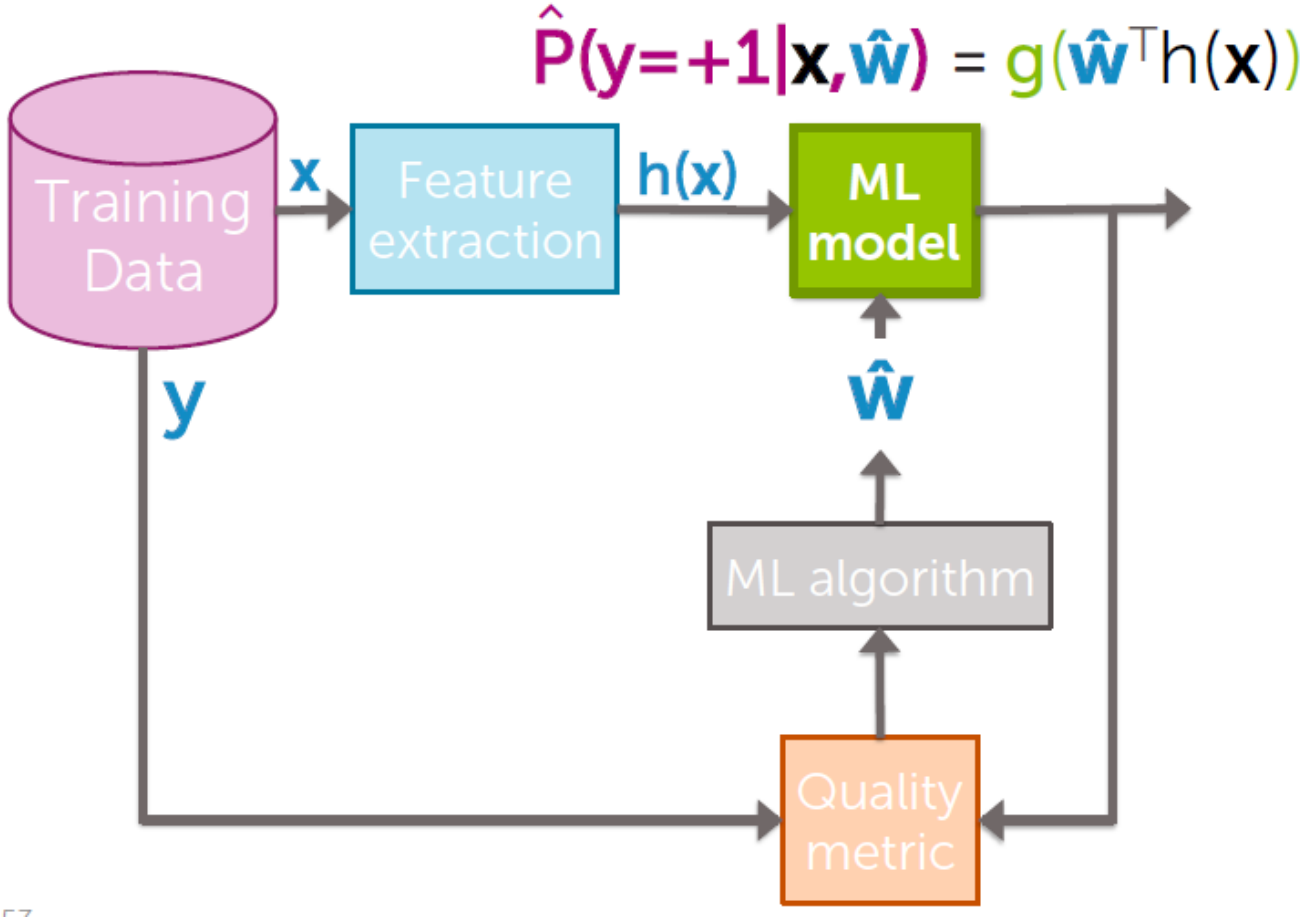
w_0	0
$w_{\#awesome}$	+1
$w_{\#awful}$	-1



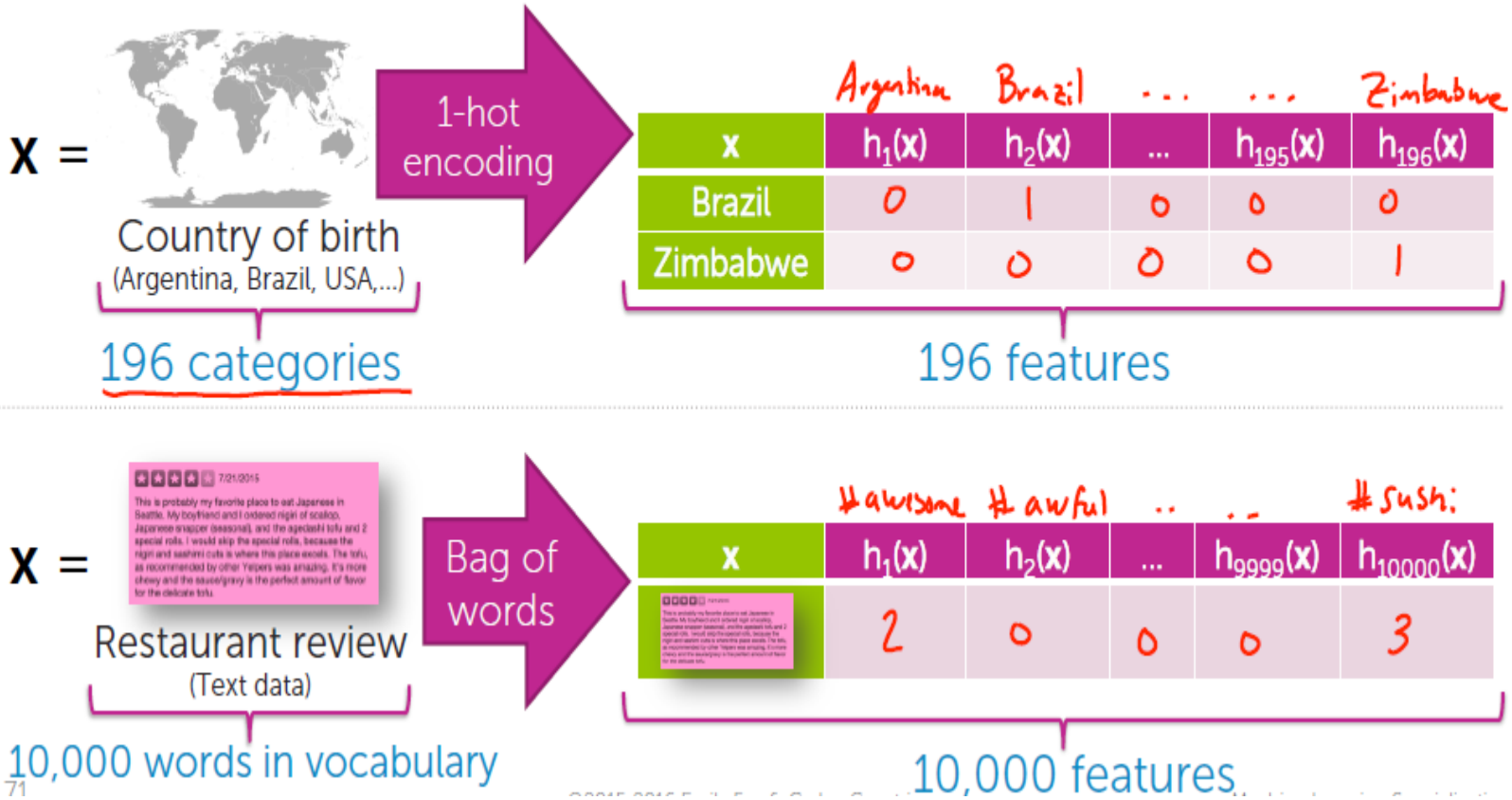
w_0	0
$w_{\#awesome}$	+3
$w_{\#awful}$	-3



Jak pewny jesteś swojej klasyfikacji



Zmienne opisowe



Klasyfikacja: ranking restauracji

Models

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

Algorithms

- Stochastic gradient descent
- Boosting

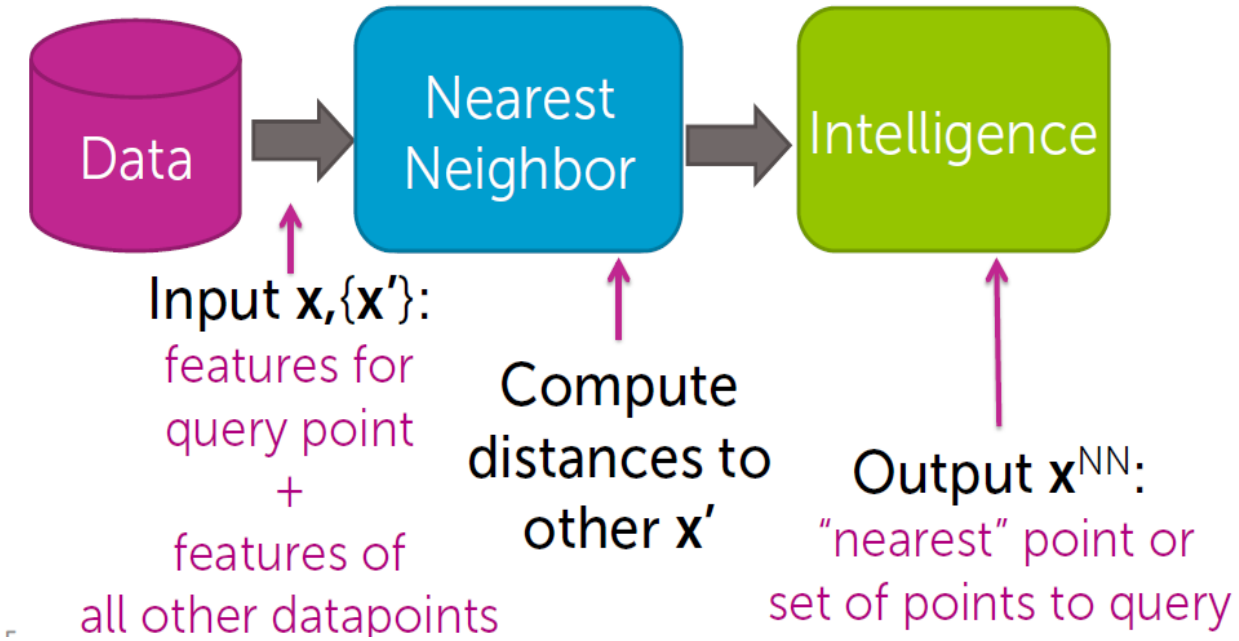
Concepts

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

Grupowanie i wybór podobnych

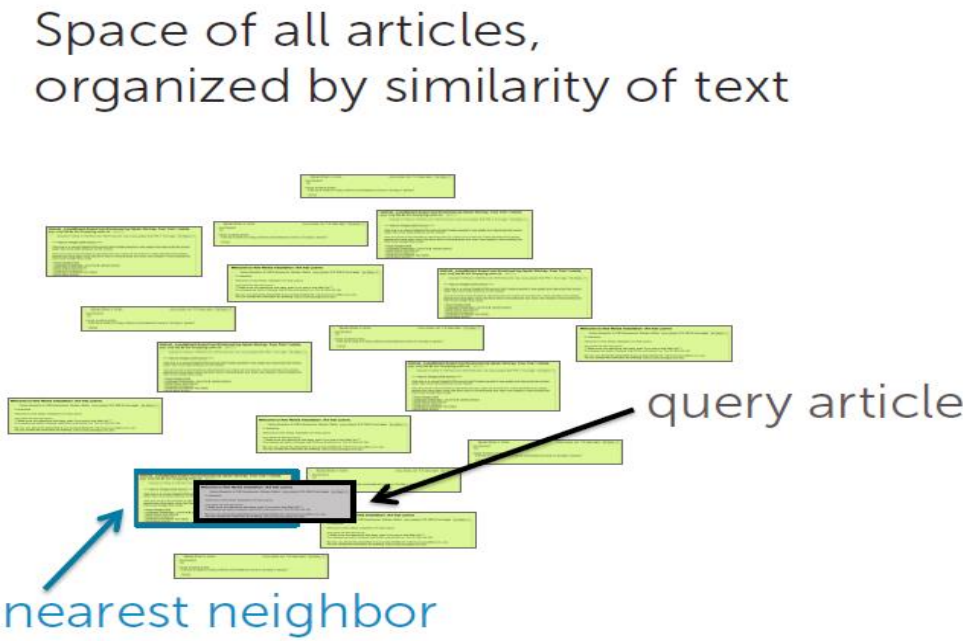
Uczenie maszynowe: Grupowanie i wyszukiwanie podobnych obiektów

- Poszukiwanie podobnych obiektów



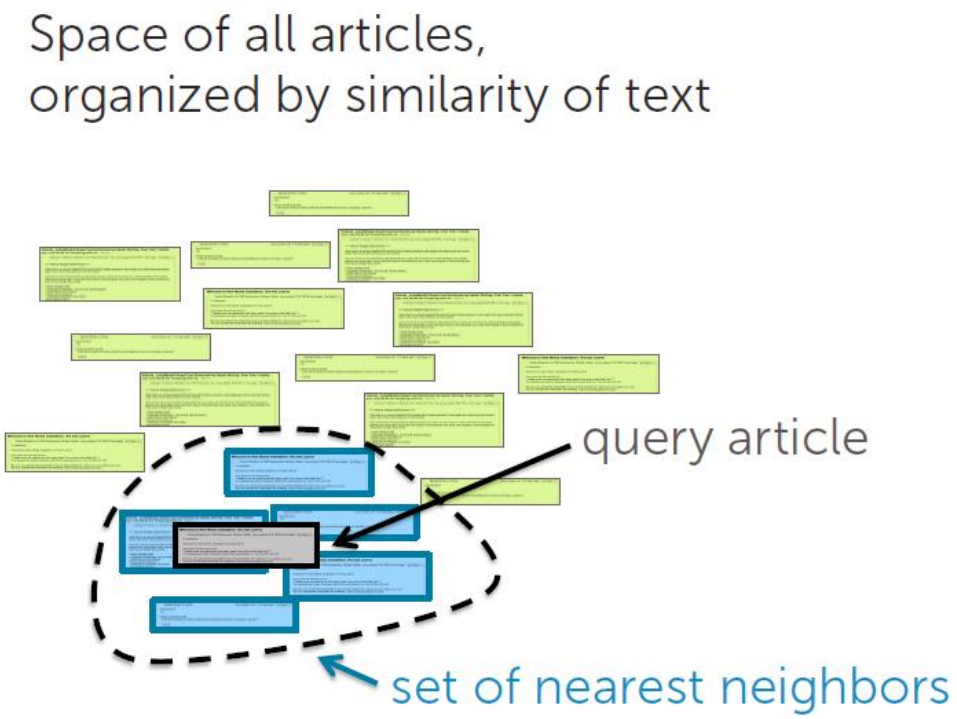
Wyszukaj podobny tekst

- Musimy zdefiniować co to znaczy „podobny”



Wyszukaj podobny tekst

- Musimy zdefiniować co to znaczy „podobny”



Technika może być stosowana do wielu zagadnień

Images

Products

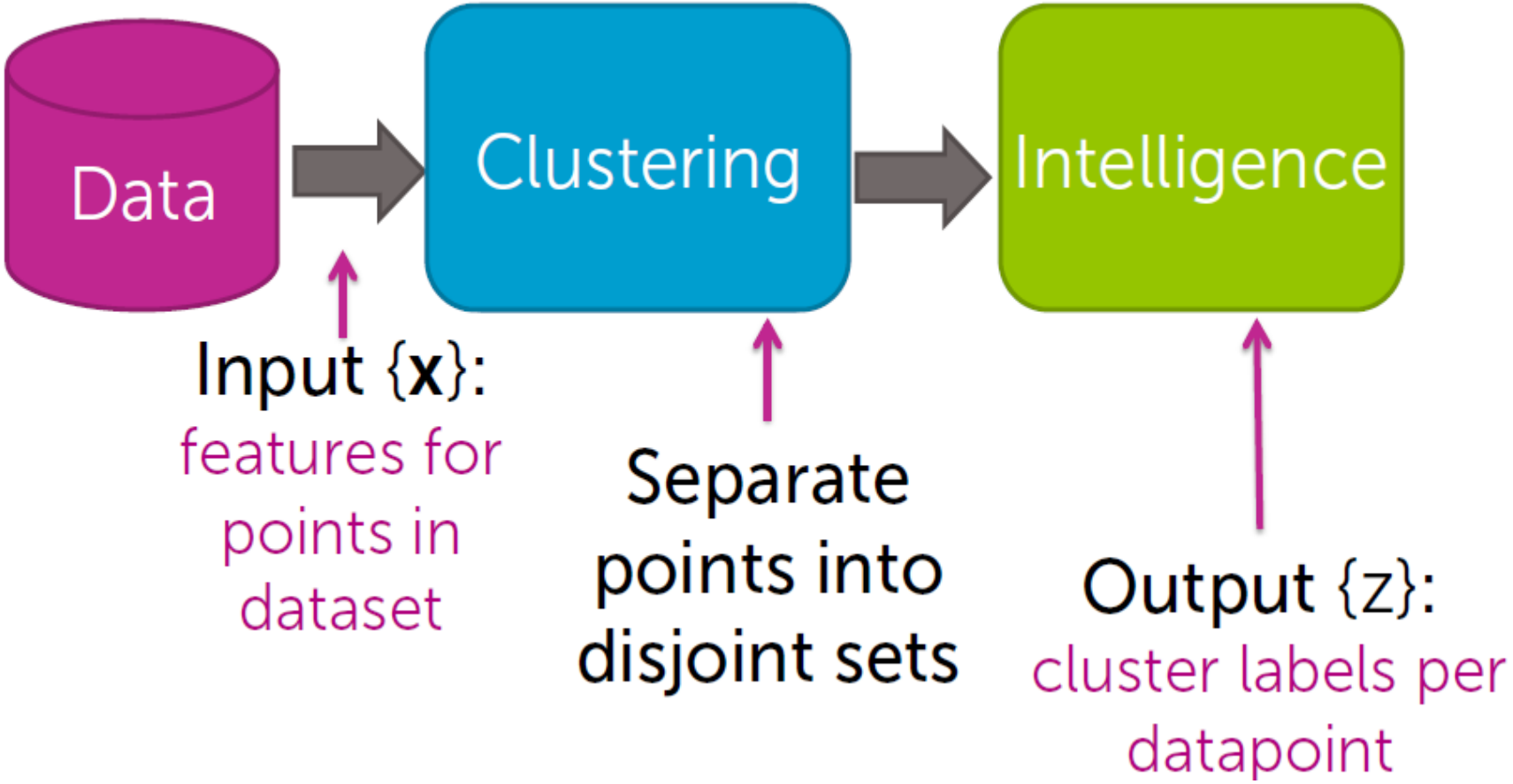
Streaming content:

- Songs
- Movies
- TV shows
- ...

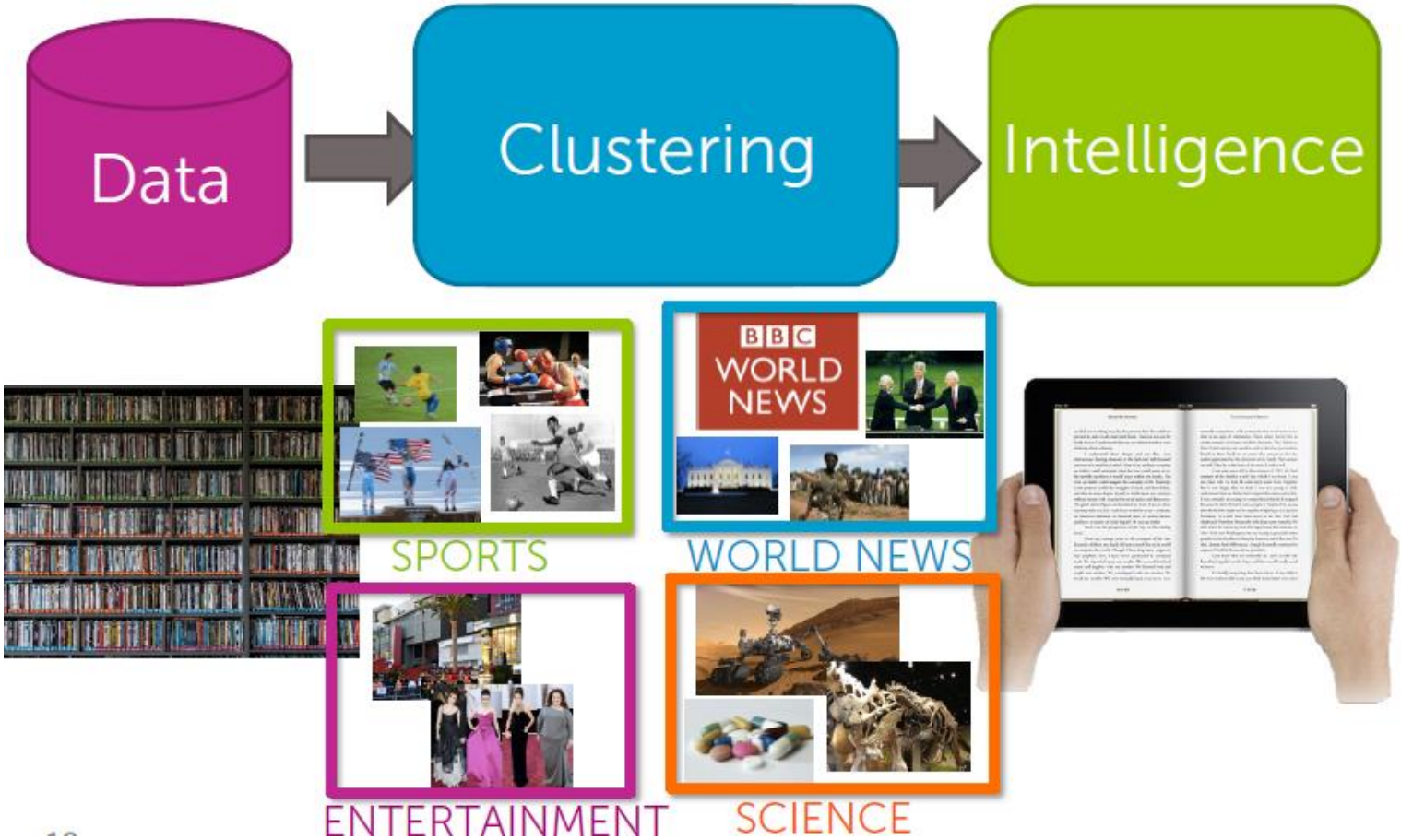
News articles

Social networks
(people you might want to connect with)

„Odkryj” grupę podobnych obiektów



Pogrupuj wg. tematów dokumenty



Pogrupuj obrazki

For search, group as:

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

Zastosowanie: grupowanie obrazków

Simple image representation

Consider average red, green, blue pixel intensities

[R = 0.05, G = 0.7, B = 0.9]

[R = 0.85, G = 0.05, B = 0.35]

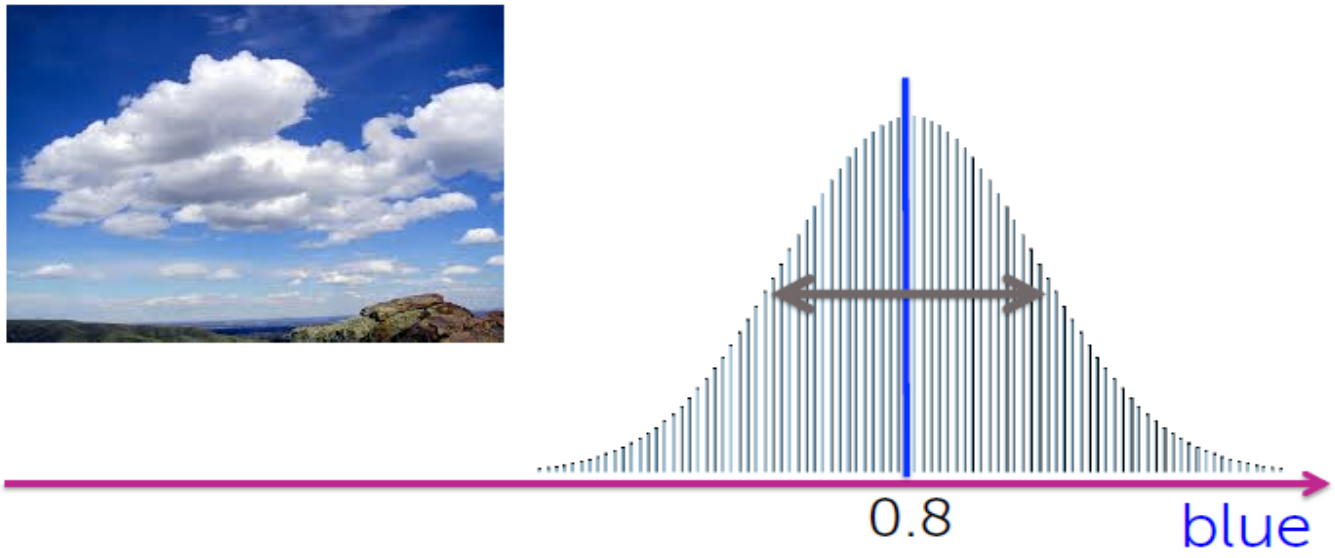
[R = 0.02, G = 0.95, B = 0.4]

Single RGB vector per image

Zastosowanie: grupowanie obrazków

Distribution over all cloud images

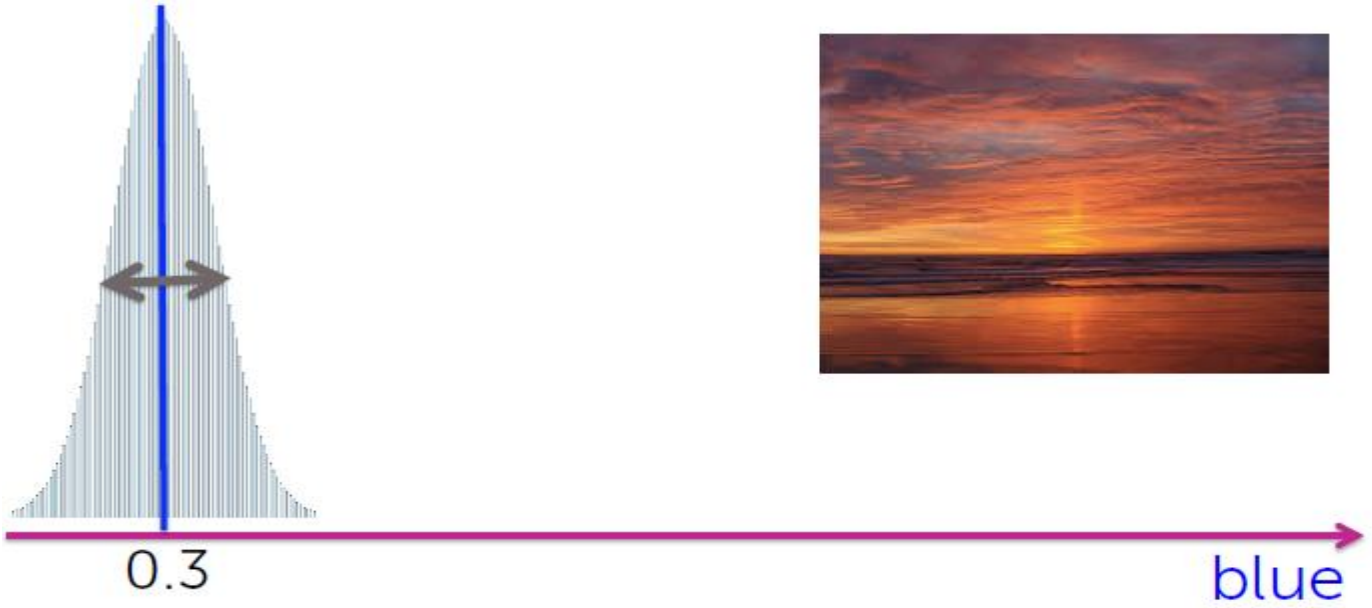
Let's look at just the blue dimension



Zastosowanie: grupowanie zdjęć

Distribution over all sunset images

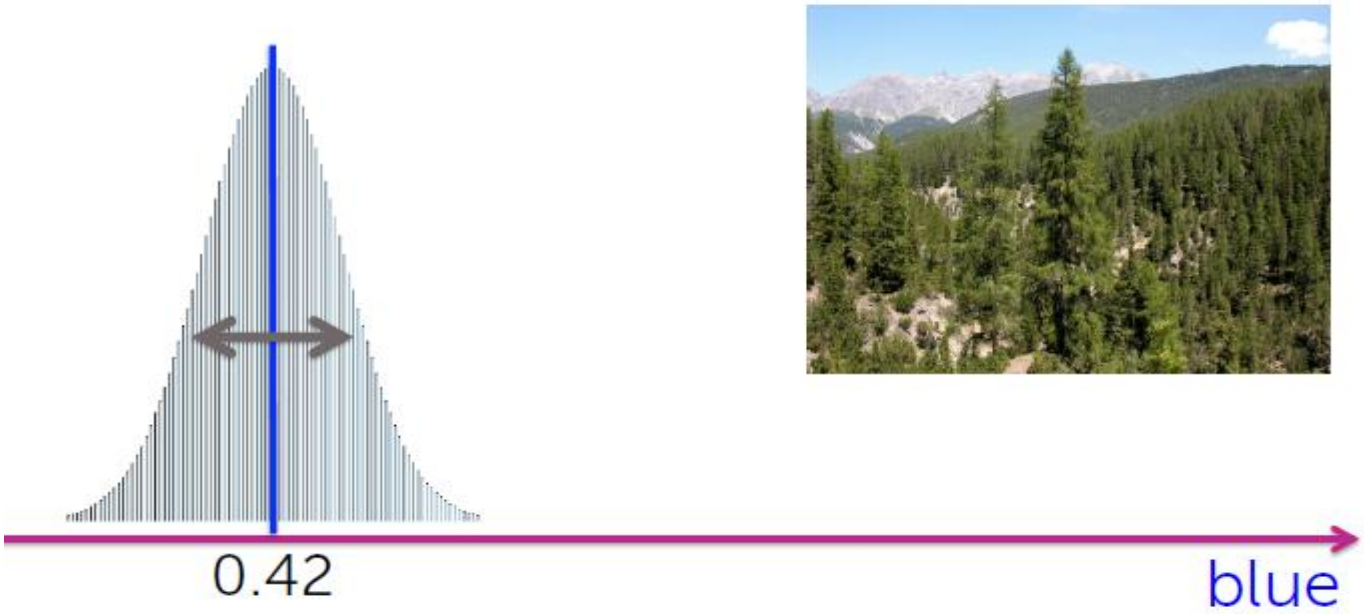
Let's look at just the blue dimension



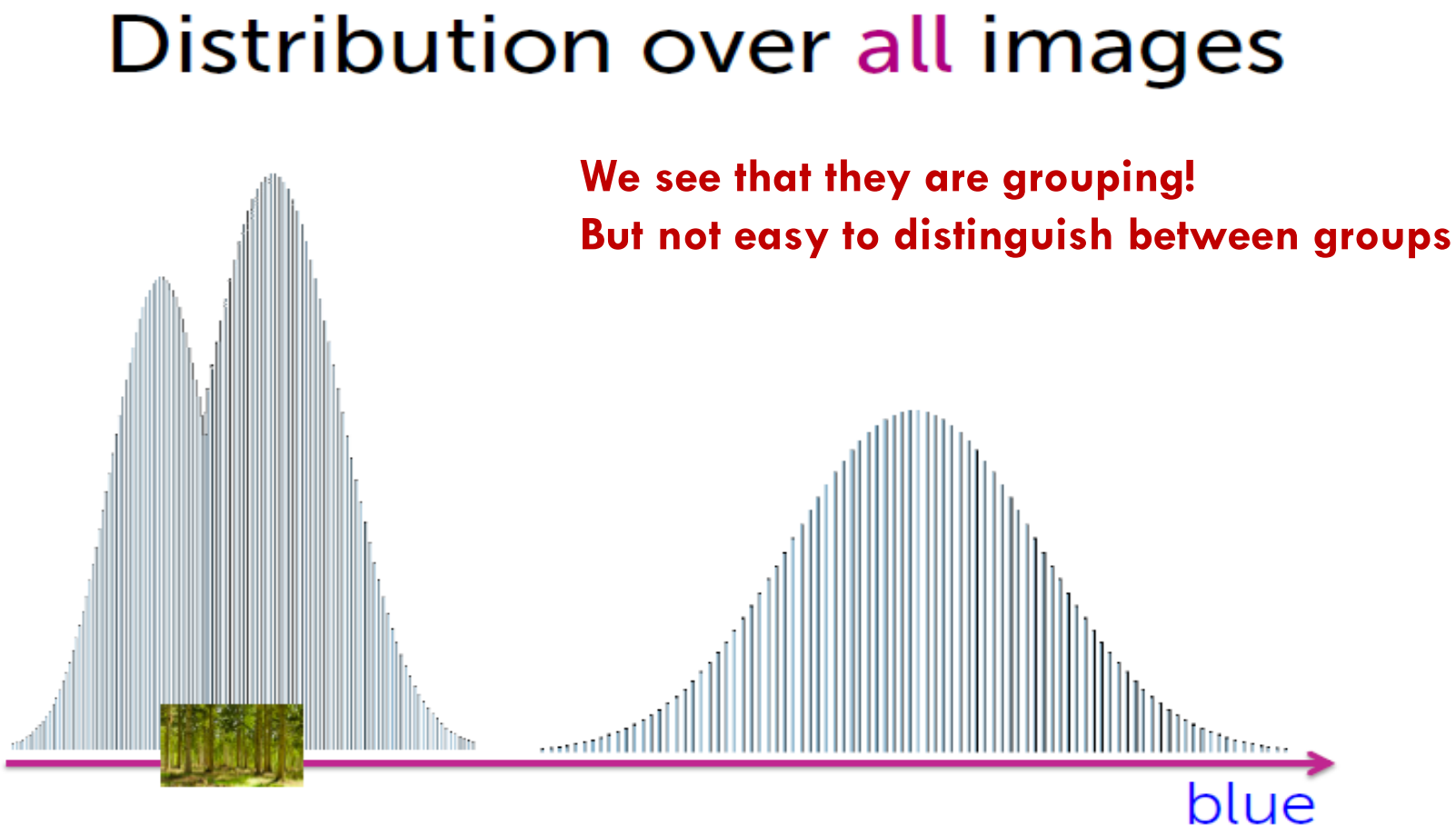
Zastosowanie: grupowanie zdjęć

Distribution over all forest images

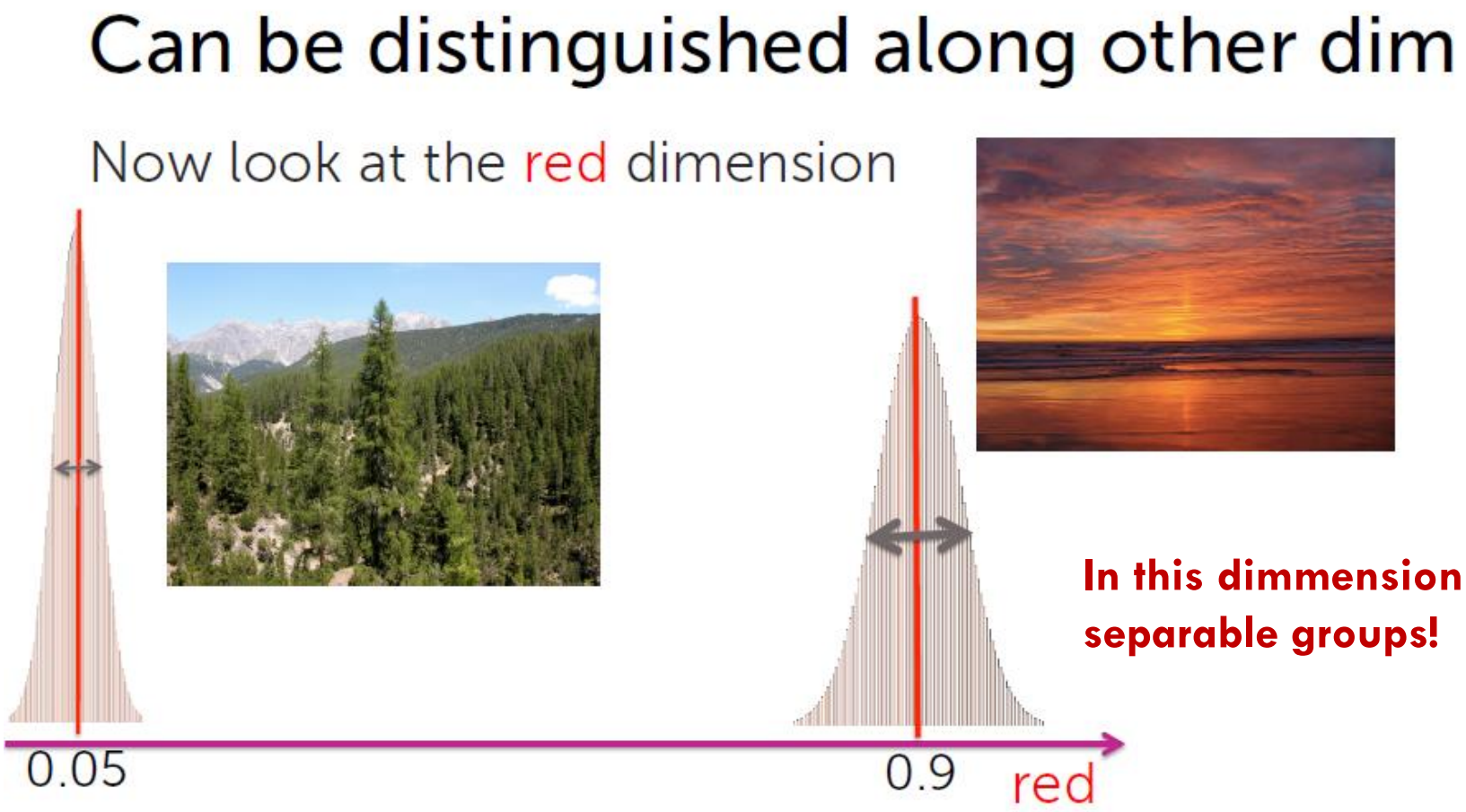
Let's look at just the blue dimension



Zastosowanie: grupowanie zdjęć



Zastosowanie: grupowanie zdjęć



Przykład: dokument na podobny temat

- Jak mierzymy podobieństwo ?
- Jak szukamy podobieństw ?

Najbliższy sąsiad

- **Input:** Query article  : $\mathbf{x}_q$

  Corpus of documents (N docs)

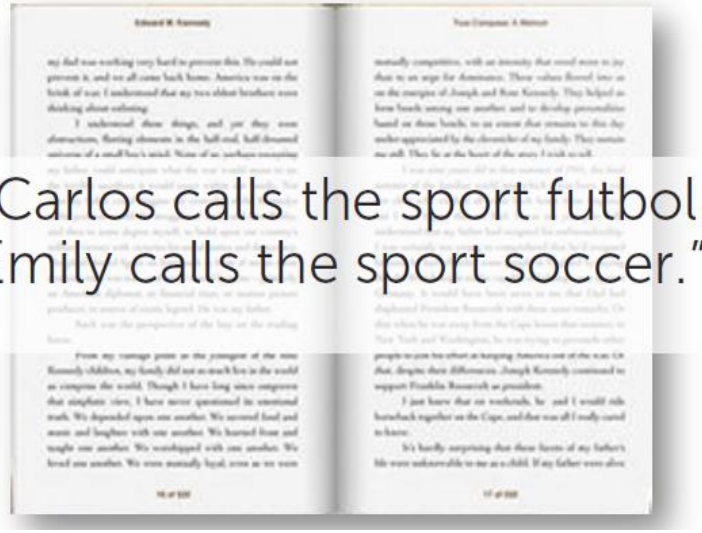
   : $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$

- **Output:** *Most* similar article  $\leftarrow x^{NN}$

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

Reprezentacja dokumentu

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



Skalowana Euklidesowa miara odległości

$$\text{distance}(\mathbf{x}_i, \mathbf{x}_q) = \sqrt{a_1(\mathbf{x}_i[1]-\mathbf{x}_q[1])^2 + \ldots + a_d(\mathbf{x}_i[d]-\mathbf{x}_q[d])^2}$$

weight on each feature
(defining relative importance)

Cosinusowa miara odległości

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

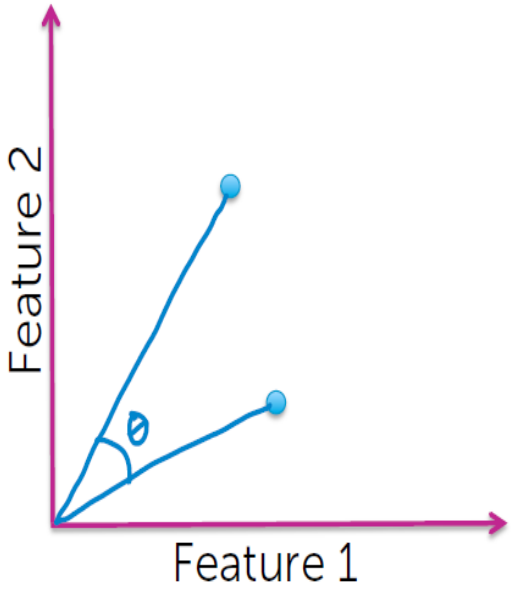
$a^T b = \|a\| \|b\| \cos(\theta)$

- 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



Naturalna miara odległości

1	0	0	0	5	3	0	0	1	0	0	0	0

Similarity = 0

0	0	1	0	0	0	9	0	0	6	0	4	0

Reprezentacja dokumentów : TF-IDF

- 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)

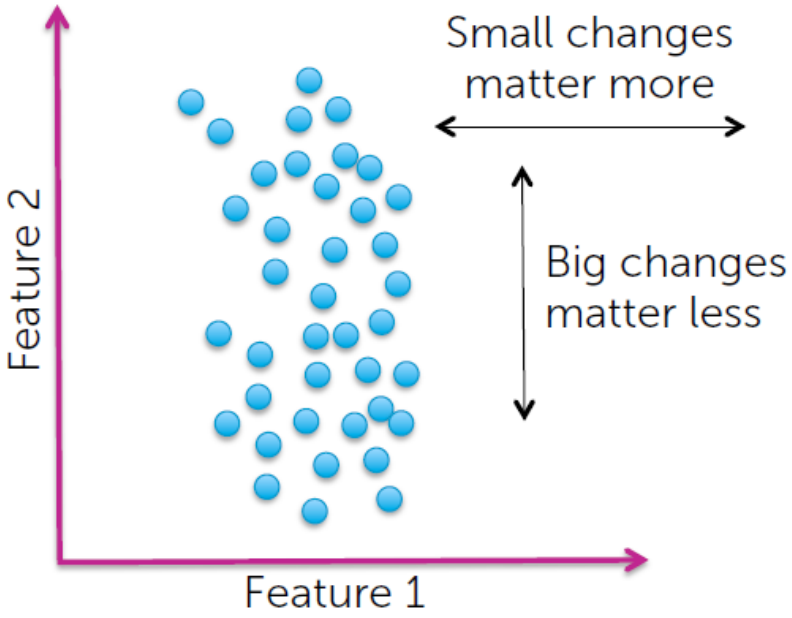
Inverse doc freq. = $\log \frac{\#\ \text{docs}}{1 + \#\ \text{docs using word}}$

Trade off: local frequency vs. global rarity

tf * idf

Ważymy cechy charakterystyczne

- 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:

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

Grupowanie:

Models

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

Algorithms

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

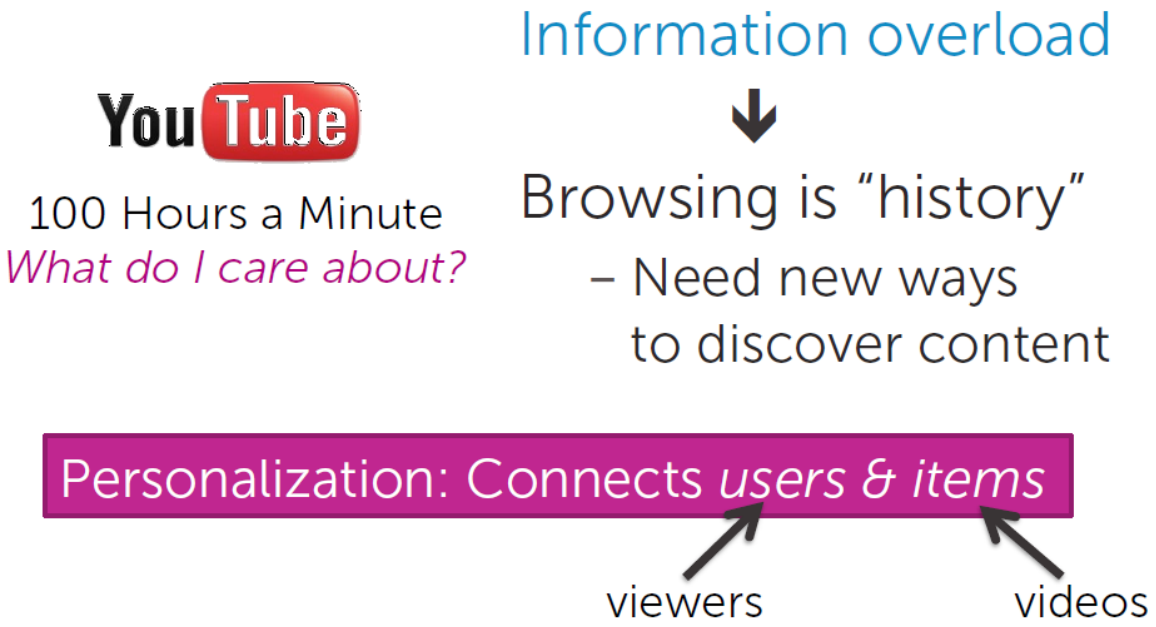
Concepts

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

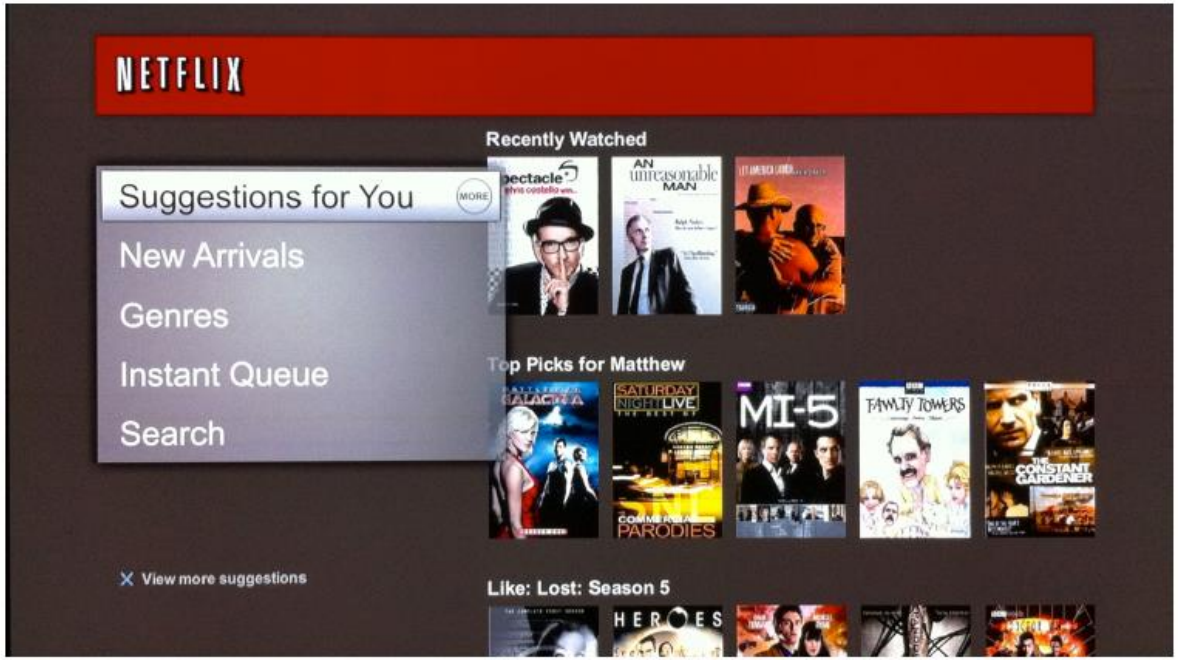
System rekomendujący

Uczenie maszynowe: system rekomendujący

- Personalizacja

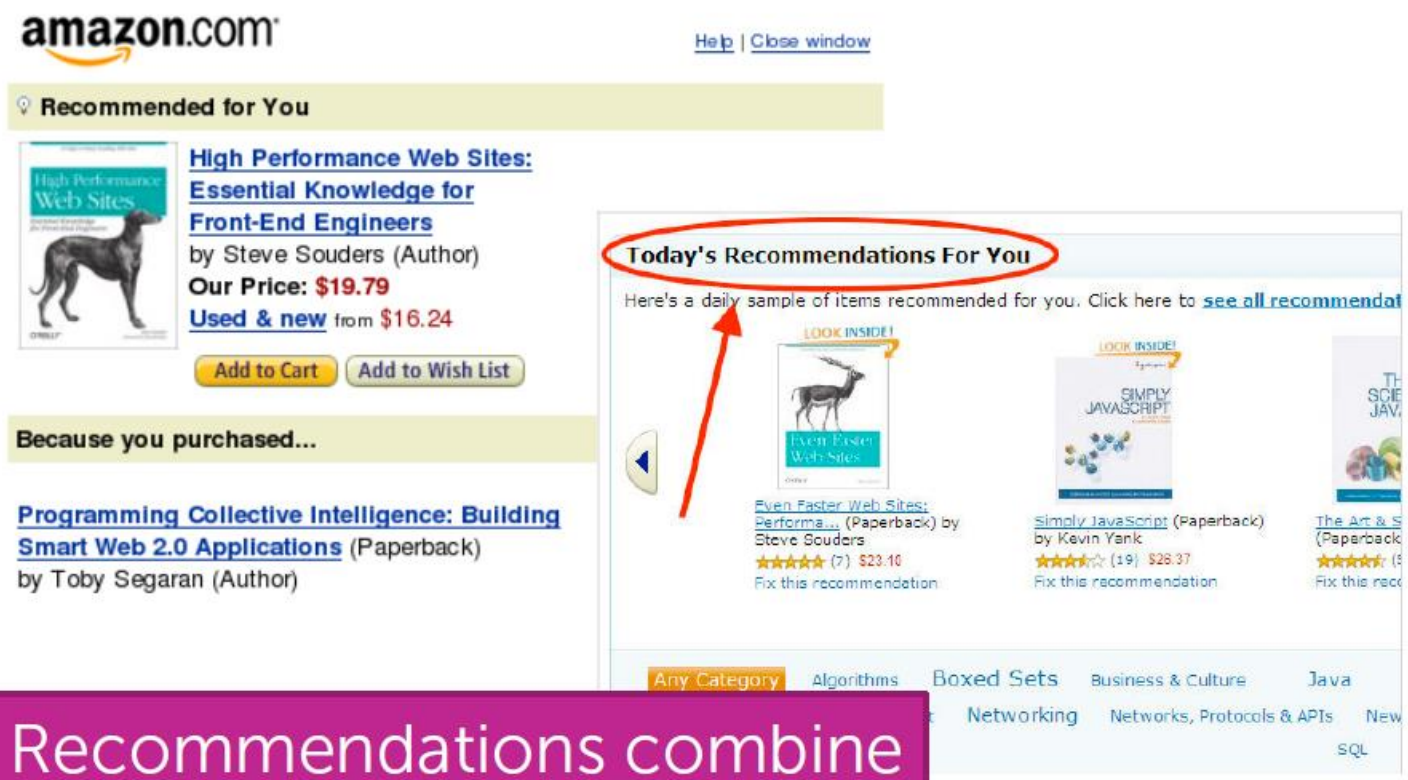


Rekomendacja: filmy



Connect users with movies they may want to watch

Rekomendacja: produkty



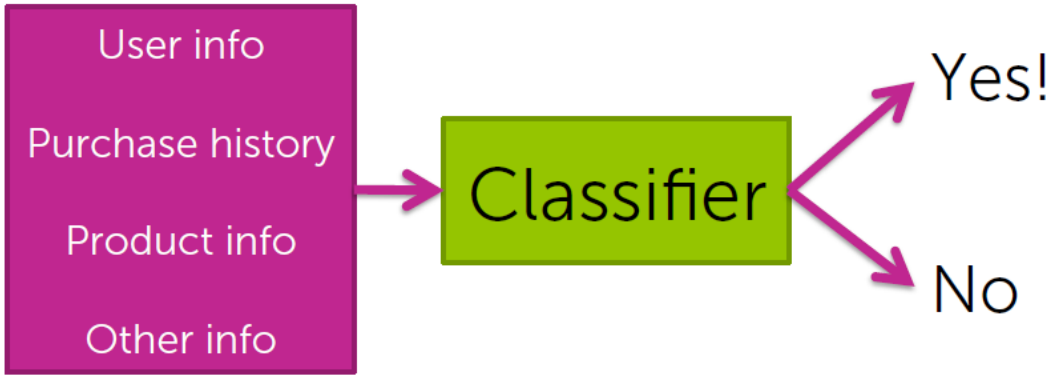
Recommendations combine global & session interests

System rekomendujący: popularność

- **Popularność?**
  - **Ranking wg. liczby wyświetlań**
  - **Nie ma personalizacji**

System rekomendujący: klasyfikacja

- **Klasyfikacja?**
  - **Jakie prawdopodobieństwo że kupię ten produkt.**
  - **Personalizacja: patrzy na historię zakupów, koreluje z porą roku, porą dnia, etc.**



System rekomendujący: korelacje

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

1. Look at *diapers* row of matrix

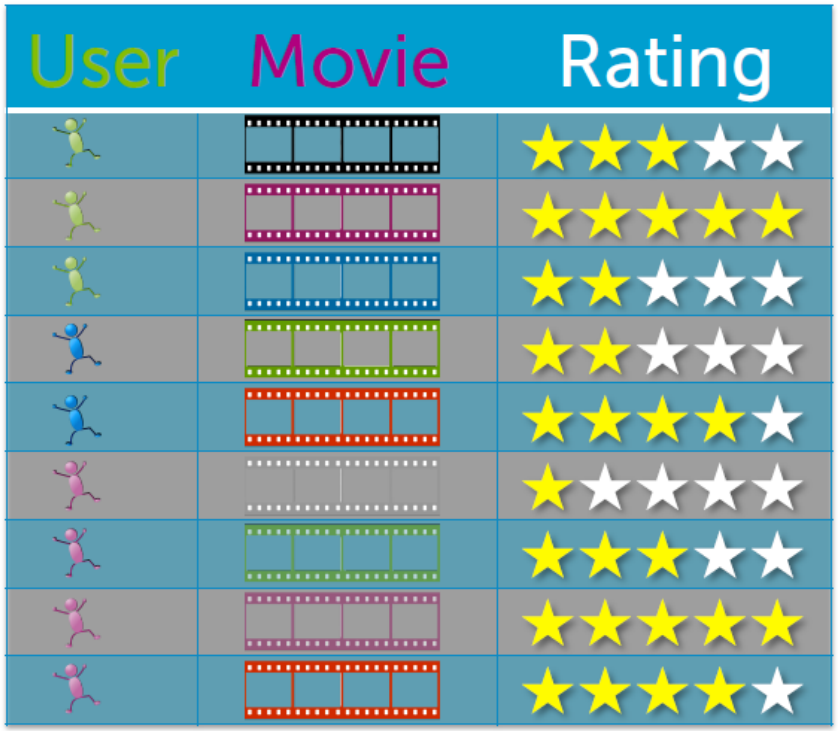
2. Recommend other items with largest counts
   - *baby wipes*, *milk*, *baby food*,...

System rekomendujący: korelacje

- **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?**

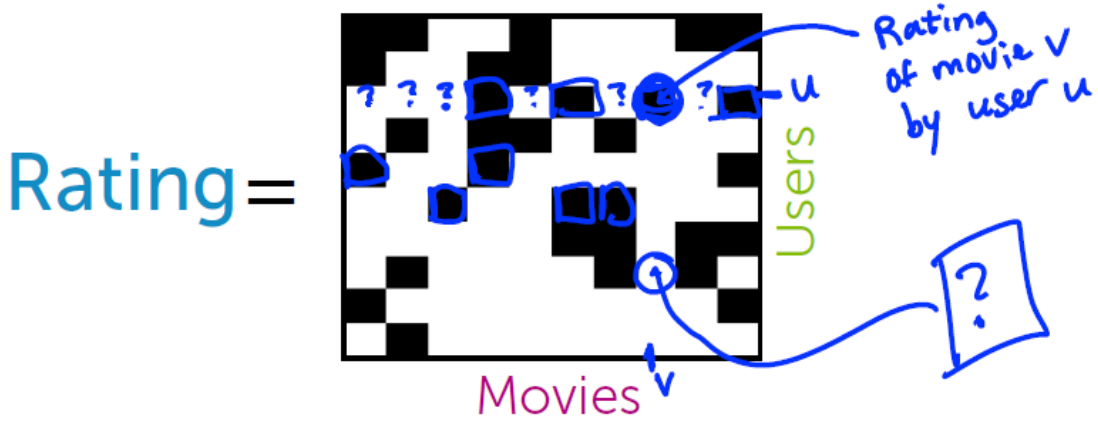
System rekomendujący: filmy

- Users watch movies and rate them



Each user only watches a few of the available movies

System rekomendujący: filmy



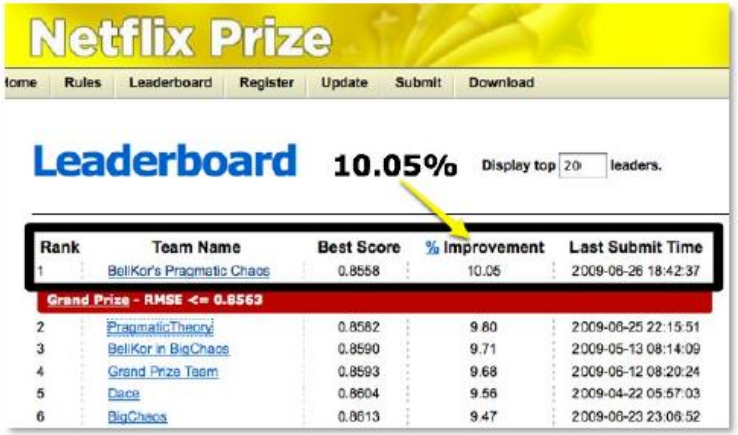
- **Data:** Users score some movies

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

- **Goal:** Filling missing data?

System rekomendujący: filmy

- 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



System rekomendujący: sprawność

System rekomendujący: sprawność

System rekomendujący: sprawność

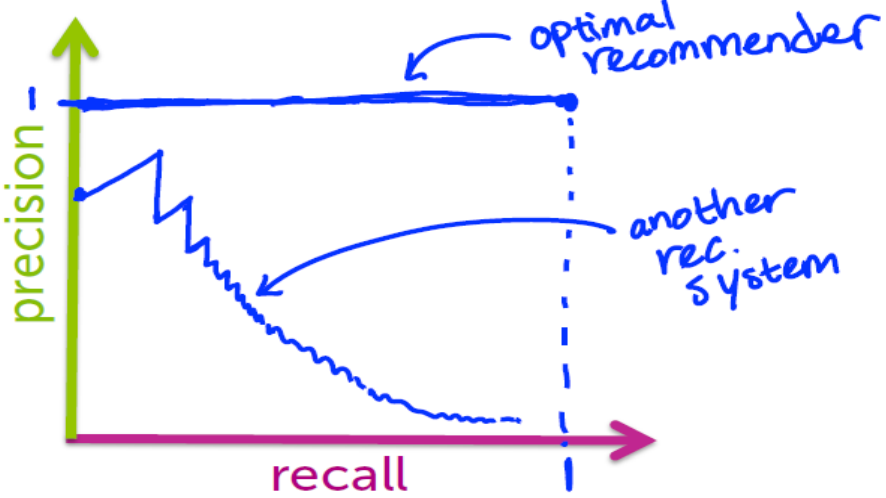
System rekomendujący: sprawność

System rekomendujący: sprawność

Precision-recall curve

- **Input:** A specific recommender system
- **Output:** Algorithm-specific precision-recall curve

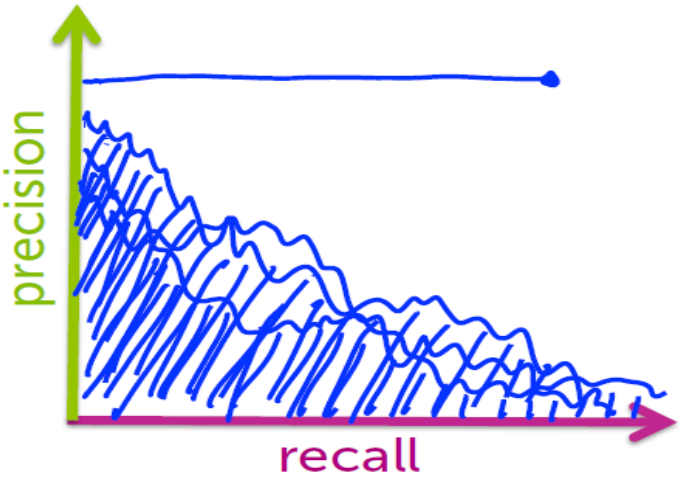
- To draw curve, vary threshold on # items recommended
  - For each setting, calculate the precision and recall



System rekomendujący: sprawność

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)



Rekomendacja produktu:

Models

- Collaborative filtering
- Matrix factorization
- PCA

Algorithms

- Coordinate descent
- Eigen decomposition
- SVD

Concepts

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