



# NIERELACYJNE ROZWIĄZANIA BAZODANOWE

WYKŁAD 5

# AGENDA

- Importowanie danych
- Zastosowanie baz nierelacyjnych w Big Data
- Wstęp do uczenia maszynowego – nieustrukturyzowane zbiory danych



# IMPORTOWANIE DANYCH

# IMPORTOWANIE DANYCH DO MONGODB W PYTHON

- `client = MongoClient("mongodb://localhost:27017/")`
- `db = client['baza']`
- `kolekcja = db['kolekcja']`
- `df = pd.read_csv('plik.csv')`
- `data = df.to_dict(orient='records')`
- `kolekcja.insert_many(data)`

# MONGOIMPORT (DOKUMENTACJA MONGODB)

- `mongoimport --uri 'mongodb+srv://MYUSERNAME:SECRETPASSWORD@mycluster-ABCDE.azure.mongodb.net/test?retryWrites=true&w=majority'`
- `mongoimport --uri 'mongodb+srv://mycluster-ABCDE.azure.mongodb.net/test?retryWrites=true&w=majority' \`
- `--username='MYUSERNAME' \`
- `--password='SECRETPASSWORD'`

# WIELE PLIKÓW JSON (DOKUMENTACJA MONGODB)

- `mongoimport --collection='mycollectionname' --file='file_per_document/ride_00001.json'`
- `cat *.json | mongoimport --collection='mycollectionname'`

- `{`
- `"tripduration": 602, "starttime": "2019-12-01 00:00:05.5640", "stoptime": "2019-12-01 00:10:07.8180", "start station id": 3382,`
- `"start station name": "Carroll St & Smith St", "start station latitude": 40.680611, "start station longitude": -73.99475825, "end station id": 3304,`
- `"end station name": "6 Ave & 9 St", "end station latitude": 40.668127, "end station longitude": -73.98377641, "bikeid": 41932,`
- `"usertype": "Subscriber", "birth year": 1970, "gender": "male"`
- `}`

# KOLEKCJA DOKUMENTÓW (DOKUMENTACJA MONGODB)

- [
  - { title: "Document 1", data: "document 1 value"},
  - { title: "Document 2", data: "document 2 value"}]
- `mongoimport --collection='from_array_file' --file='one_big_list.json' --jsonArray`

# ŚRODOWISKO GRAFICZNE – MONGODB COMPASS

## (DOKUMENTACJA MONGODB)

<input checked="" type="checkbox"/> _id	<input checked="" type="checkbox"/> airline	<input checked="" type="checkbox"/> name	<input checked="" type="checkbox"/> alias	<input checked="" type="checkbox"/> ia	
<div>Objectid</div>	<div>Int32</div>	<div>Mixed</div>	<div>Mixed</div>	<div>Mix</div>	
1	56e9b497732b6122f8790280	4	2 Sqn No 1 Elementary Flying Training Sch...	empty string	WYT
2	56e9b497732b6122f8790281	2	135 Airways	empty string	GNL
3	56e9b497732b6122f8790282	5	213 Flight Unit	empty string	TFU
4	56e9b497732b6122f8790283	3	1Time Airline	1T	RNX
5	56e9b497732b6122f8790284	6	223 Flight Unit State Airline	empty string	CHD
6	56e9b497732b6122f8790285	1	Private flight	-	N/A
7	56e9b497732b6122f8790286	7	224th Flight Unit	empty string	TTF
8	56e9b497732b6122f8790287	8	247 Jet Ltd	empty string	TWF
9	56e9b497732b6122f8790288	9	3D Aviation	empty string	SEC
10	56e9b497732b6122f8790289	10	40-Mile Air	Q5	MLA

Cancel

Import



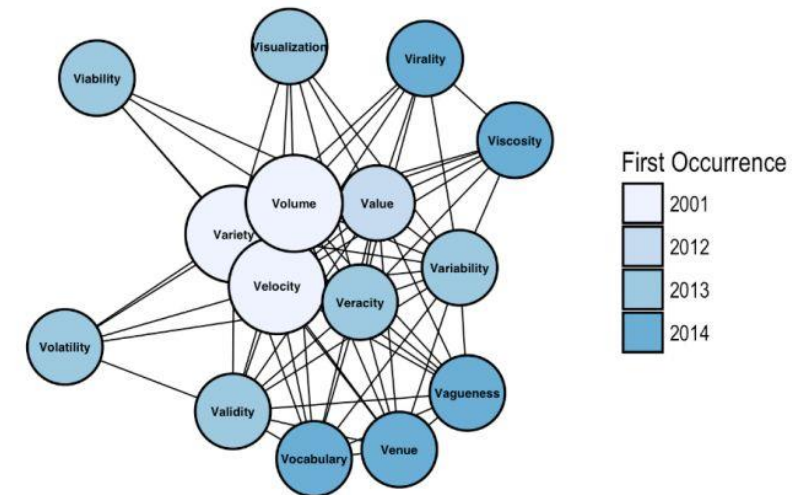
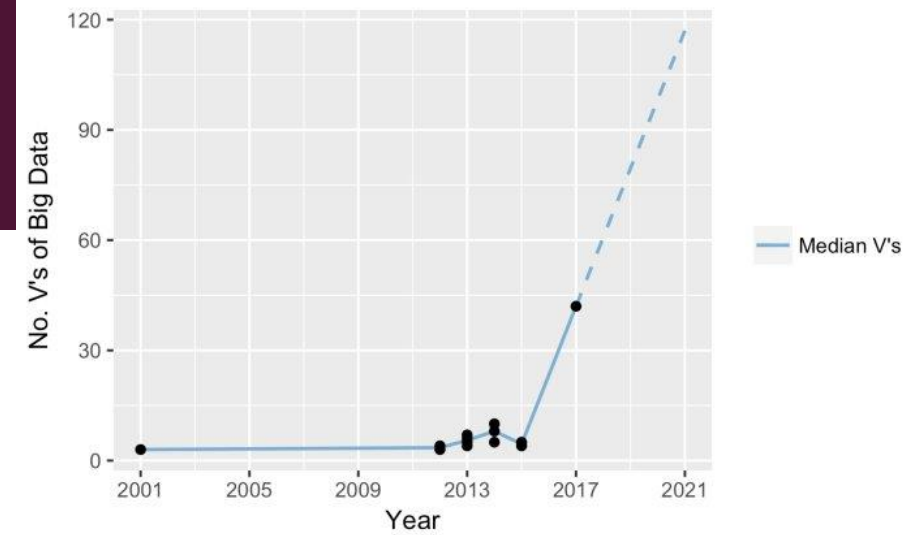


# ZASTOSOWANIE BAZ NIERELACYJNYCH – BIG DATA

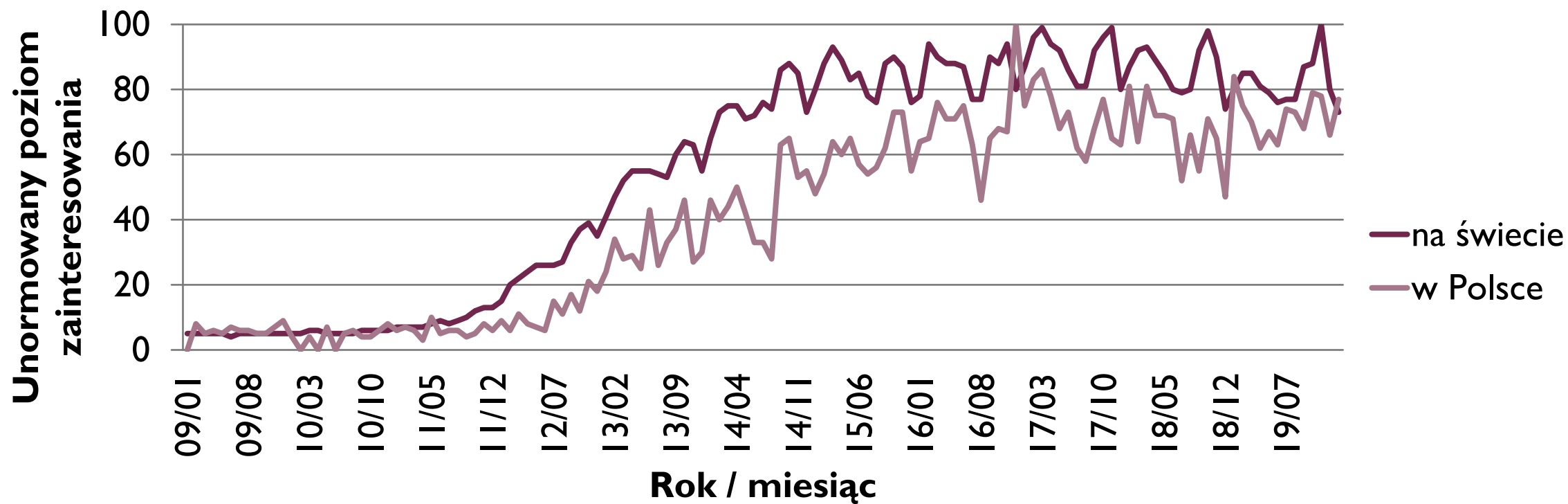


# BIG DATA

- Doug Laney, Gartner (2001)
- Volume (duży wolumen danych)
- Variety (duża różnorodność danych)
- Velocity (duża intensywność napływu danych)
- Duże zbiory danych, których przetwarzanie nie jest możliwe tradycyjnymi metodami.
- Zestaw metod i technik przetwarzania danych, m.in. algorytmy MapReduce w Apache Hadoop (2005)
- Big Data a Small Data



# ZAINTERESOWANIE TERMINEM BIG DATA WEDŁUG GOOGLE TRENDS



# RODZAJE DANYCH (ONZ)

Wyszczególnienie	Dane tworzone przez ludzi	Dane powstające w procesach biznesowych	Dane generowane przez urządzenia
Rodzaje źródeł	<b><u>I 100. Sieci społecznościowe:</u></b> Facebook, X, Tumblr itp. <b><u>I 200. Blogi i komentarze, strony internetowe</u></b> I 300. Dokumenty osobiste I 400. Zdjęcia: Instagram, Flickr, Picasa itp. I 500. Filmy: YouTube itp. <b><u>I 600. Wyszukiwanie w Internecie</u></b> I 700. Treść danych mobilnych: wiadomości tekstowe I 800. Mapy generowane przez użytkowników I 900. E-mail	21. Dane wytwarzane przez agencje publiczne 2110. Dokumentacja medyczna 22. Dane wytwarzane przez firmy 2210. Transakcje handlowe 2220. Ewidencja bankowa / zapasowa 2230. Handel elektroniczny 2240. Karty kredytowe	31. Dane z czujników 311. Naprawiono czujniki 3111. Automatyka domowa <b><u>3112. Czujniki pogody / zanieczyszczenia</u></b> 3113. Czujniki ruchu / kamera internetowa 3114. Czujniki naukowe 3115. Filmy / obrazy z zakresu bezpieczeństwa / nadzoru 312. Czujniki mobilne (śledzenie) <b><u>3121. Lokalizacja telefonu komórkowego</u></b> 3122. Samochody <b><u>3123. Zdjęcia satelitarne</u></b> 32. Dane z systemów komputerowych <b><u>3210. Dzienniki</u></b> <b><u>3220. Dzienniki sieciowe</u></b>

# ALTERNATYWNE ŹRÓDŁA DANYCH DO POZYSKANIA DANYCH

Źródło danych	Obszar potencjalnego wykorzystania
Dane ze skanerów kodów kreskowych	Statystyka cen, statystyki ekonomiczne
Dane nt. lokalizacji telefonów komórkowych	Statystyka turystyki, statystyka ludności i migracji
Dane z sensorów drogowych	Statystyka transportu
Dane z mierników zużycia energii	Statystyka ludności, statystyka gospodarstw domowych
Zdjęcia satelitarne, dane zdalnych sensorów	Statystyka rolnictwa, leśnictwa, rybołówstwa oraz statystyka środowiska naturalnego
Dane z serwisów społecznościowych, dane z Internetu	Statystyka rynku pracy, statystyka ludności i migracji, statystyka dochodów i konsumpcji gospodarstw domowych, statystyka cen, statystyka zdrowia, statystyka społeczna
Strony WWW z ofertami pracy	Statystyka rynku pracy
Ruch samolotów	Statystyka transportu, statystyka ochrony środowiska
Strony WWW: nieruchomości, działalność e-commerce	Statystyka cen

# KRYTYKA METOD I DANYCH BIG DATA

- „Nie tylko w podejmowaniu decyzji, ale także w empirycznych badaniach ekonomicznych Big Data ma do spełnienia funkcję raczej **komplementarną, a nie substytucyjną** (...) Big Data koncentruje się głównie na powiązaniach między zmiennymi, rejestrując duże liczby cech, które są w stanie opisywać najistotniejsze współzależności między nimi. Te współzależności mogą mieć **charakter sztuczny** (pozorny) lub przyczynowo-skutkowy.” (Szreder, 2017).
- Analiza danych Big Data w dużej części zbiorów **nie może zostać przeprowadzona z wykorzystaniem standardowych rozwiązań statystycznych** (Domański, Jędrzejczak, 2015).
- Ostatnie lata pokazują jednak, że dane uzyskiwane metodą **Big Data nie pomagają w osiągnięciu lepszych i bardziej wiarygodnych wyników** (Szreder, 2019; Gezgin, 2018).

## Bibliografia

- Szreder, M., (2017). Nowe źródła informacji i ich wykorzystywanie w podejmowaniu decyzji. Wiadomości Statystyczne. The Polish Statistician, (7)4.
- Domański, Cz., Jędrzejczak, A. (2015). Statistical Computing in Information Society. Folia Oeconomica Stetinensia. 15. 10.1515/fofi-2015-0041.
- Szreder, M. (2019). Istotność statystyczna w czasach big data. Wiadomości Statystyczne, 64(11), 42–57.
- Gezgin, U. B. (2018). An invitation to critical social science of big data: from critical theory and critical research to omni-resistance. AI & Society, 35(1), 187–195. <https://doi.org/10.1007/s00146-018-0868-y>

# JAKOŚĆ DANYCH BIG DATA

Trzy etapy:

- wejście (źródło danych – pozyskiwanie, wstępna ocena danych)
- przetwarzanie (obszar przejściowy – przekształcanie, modyfikowanie, analiza danych)
- wyjście (raporty – wyniki analiz lub przetwarzania)

Trzy obiekty:

- źródło – typ danych, charakterystyka obiektów oraz encji, uwarunkowania prawne oraz kwestie związane z dostępem
- metadane – odnoszą się do aspektów związanych z opisem danych, przede wszystkim zastosowane standardy i klasyfikacje czy stopień pokrycia badanej populacji
- dane – ocena jakości danych zawartych w źródłach



# WSTĘP DO UCZENIA MASZYNOWEGO

PRZETWARZANIE NIEUSTRUKTURYZOWANYCH ZBIORÓW DANYCH



# RODZAJE UCZENIA MASZYNOWEGO

- uczenie nadzorowane
- uczenie nienadzorowane
- uczenie przez wzmacnianie

# SUPERVISED MACHINE LEARNING

- Zadanie oszacowania wartości wyjściowej ze zbioru wartości wejściowych nazywane jest regresją w statystyce; dla modelu liniowego mamy regresję liniową. W uczeniu maszynowym regresja jest jednym z rodzajów **uczenia nadzorowanego**.

Ethem Alpaydin. (2016). Machine Learning :The New AI.The MIT Press.

## UCZENIE NADZOROWANE (I/2)

- Etykiety → Dane uczące się → Algorytmy uczenia maszynowego
- Dane → Model predykcyjny → Prognozy

## UCZENIE NADZOROWANE (2/2)

- Klasyfikacja — etykiety klas
  - Klasyfikacja binarna, np. SPAM (pozytywny, negatywny)
  - Klasyfikacja wieloklasowa (np. rozpoznawanie obrazów, liczb)
- Wyniki ciągłe — regresja
  - Dane zmienne objaśniające — prognozujące
  - Ciągła zmienna objaśniana — prognozowana

# REINFORCEMENT LEARNING

- W **uczeniu ze wzmacnianiem** należy podejmować decyzje sekwencyjne, a nie podejmować decyzje jednorazowo, co w niektórych przypadkach utrudnia trenowanie modeli.

Dangeti, P. (2017). Statistics for Machine Learning. Packt Publishing.

# UCZENIE PRZEZ WZMACNIANIE

- Podobnie jak uczenie nadzorowane, jednak przewiduje się nagrody za poprawne wyniki (wzmacnianie)
- System – regulator lub agent
- Poprawianie skuteczności poprzez interakcję
- Uczenie przez wzmacnianie to przykład problemów interaktywnych

# UNSUPERVISED MACHINE LEARNING

- Celem **uczenia nienadzorowanego** jest odkrycie ukrytych wzorców lub struktur danych, w których nie ma zmiennej docelowej, do wykonania metod klasyfikacji lub regresji.

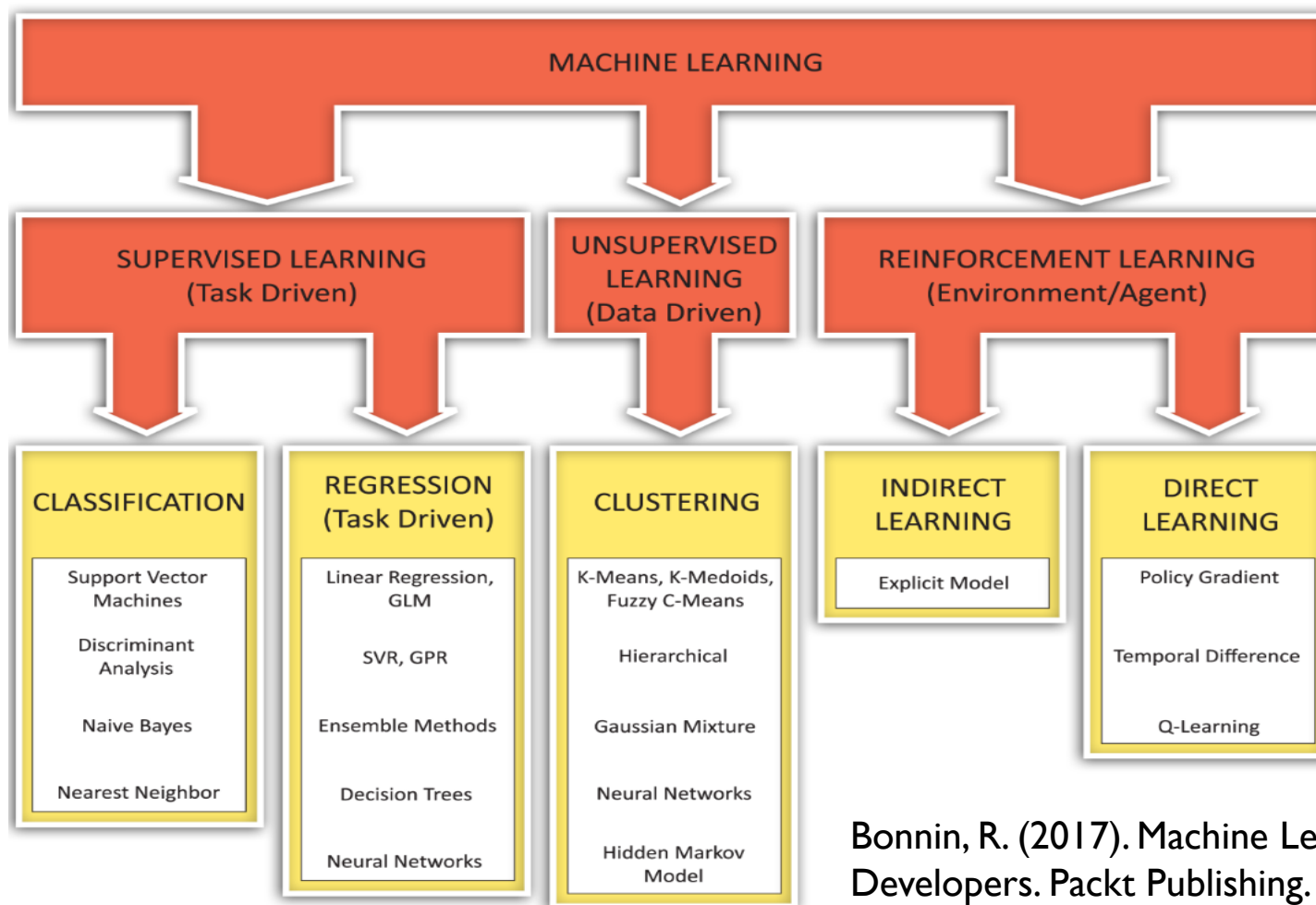
Dangeti, P. (2017). Statistics for Machine Learning. Packt Publishing.

# UCZENIE NIENADZOROWANE

- Dane o nieznanej strukturze – tzw. ukryte struktury
- Grupowanie – klasteryzacja, analiza skupień, analiza głównych składowych (PCA)
- Oddzielna dziedzina – redukowanie nadmiarowości – szumu z danych



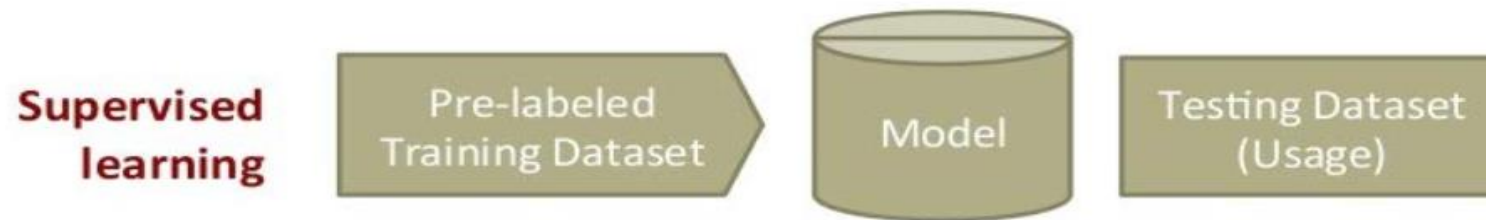
# ALGORYTMY UCZENIA MASZYNOWEGO





# SUPERVISED MACHINE LEARNING

# SUPERVISED LEARNING – ZBIÓR TRENINGOWY

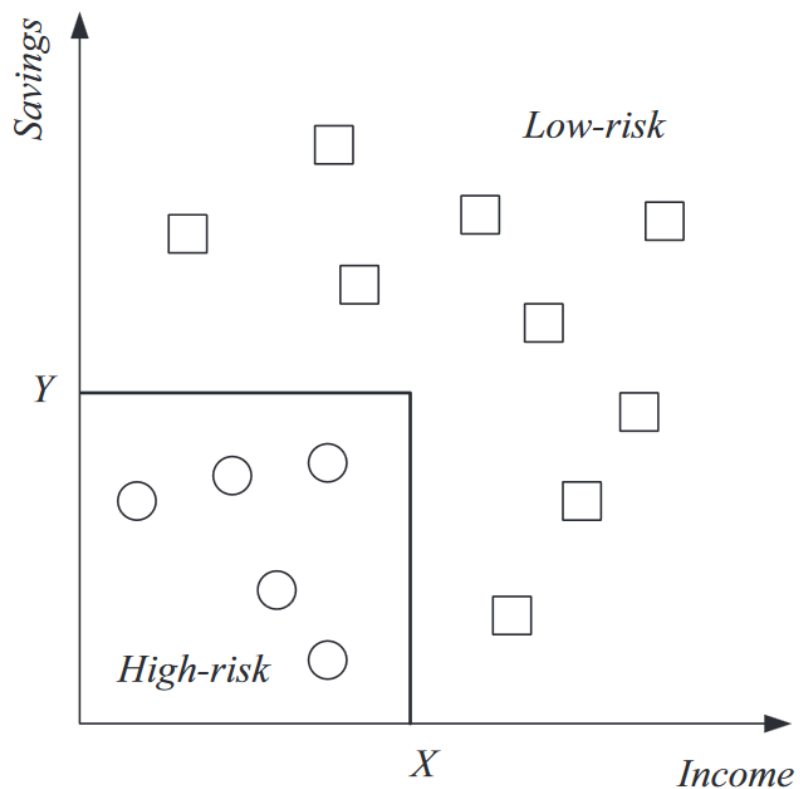


$$y = f(x)$$

Gollapudi, S., & Laxmikanth, V. (2016). Practical Machine Learning. Packt Publishing.

- I feel so bad today -> **smutny**
- Tomorrow I start the vacation and feel so exhilarated -> **szczęśliwy**
- I am so busy and tired -> **smutny**
- We are going for holidays and I feel so lovely -> **szczęśliwy**

# PRZYKŁAD: RYZYKO KREDYTOWE



- Supervised learning


## CZEGO MOŻNA NAUCZYĆ?

- Tekst, np. analiza nastrojów
- Zdjęcia, np. rodzaje upraw, ludzie (rozpoznawanie twarzy), płeć
- Liczby, np. rodzaj kwiatu na podstawie wymiarów, koloru itp.

# PRZYKŁAD ANALIZY SENTYMENTU

Happy



((KatherlNe Burke))  @KathyBurke · 31 min.  
I met my wife on Twitter. Don't ask me how, but it's made us deliriously **happy** |  
Matt Owen [theguardian.com/commentisfree/...](https://theguardian.com/commentisfree/...) This is a cheer up 🥰

Sad



**Faith in Humanity** @TheWorldImages · 7 u  
A polar bear is being kept inside a mall for selfies. This makes me so **sad**..

Scared



**Pat Kane** @thoughtland · 1 u  
Haven't been that **scared** at unlikely prospect of Trump victory. Am now. FFS.

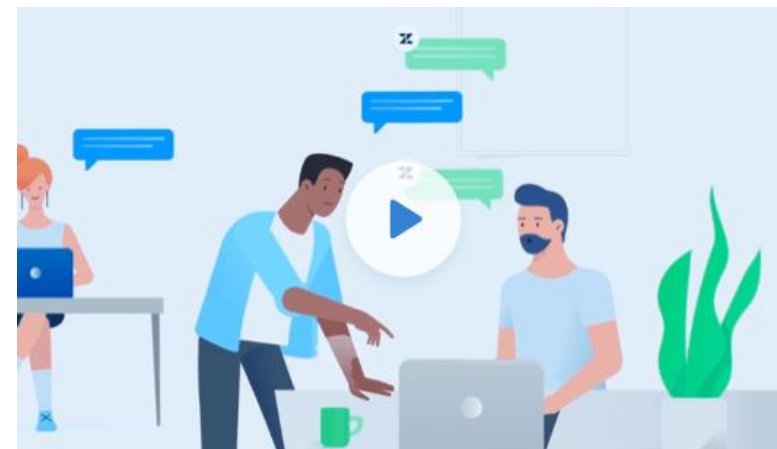
# PRZYKŁAD: UCZENIE NADZOROWANE



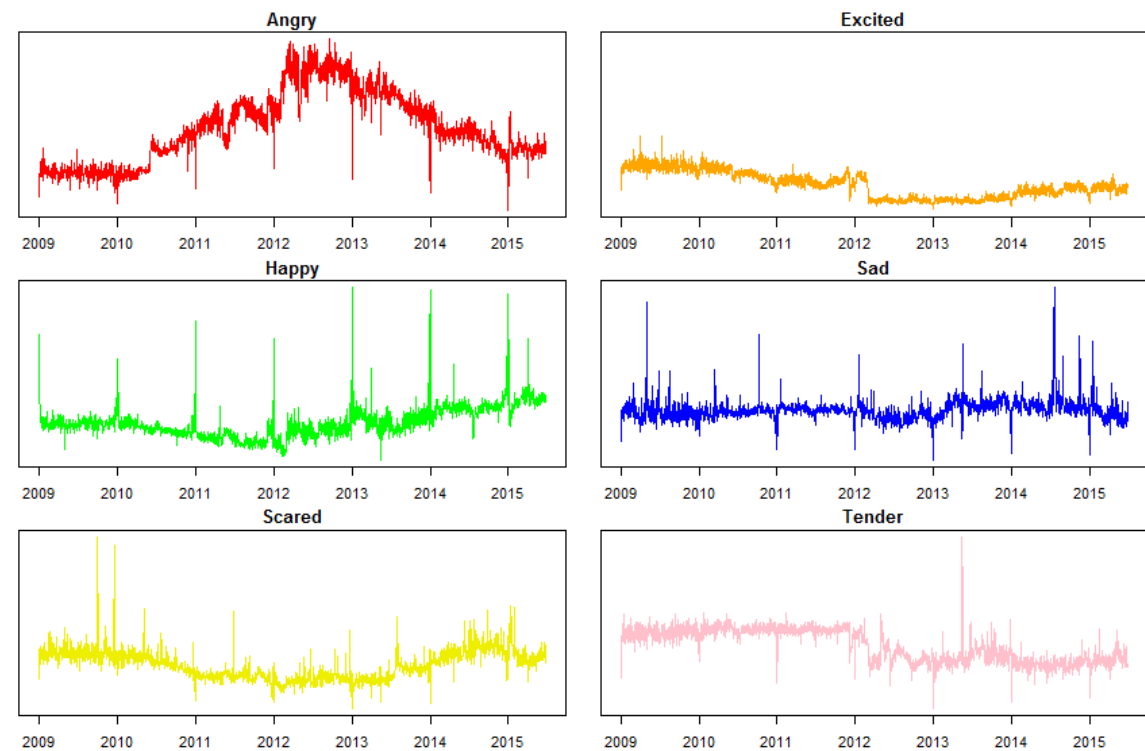
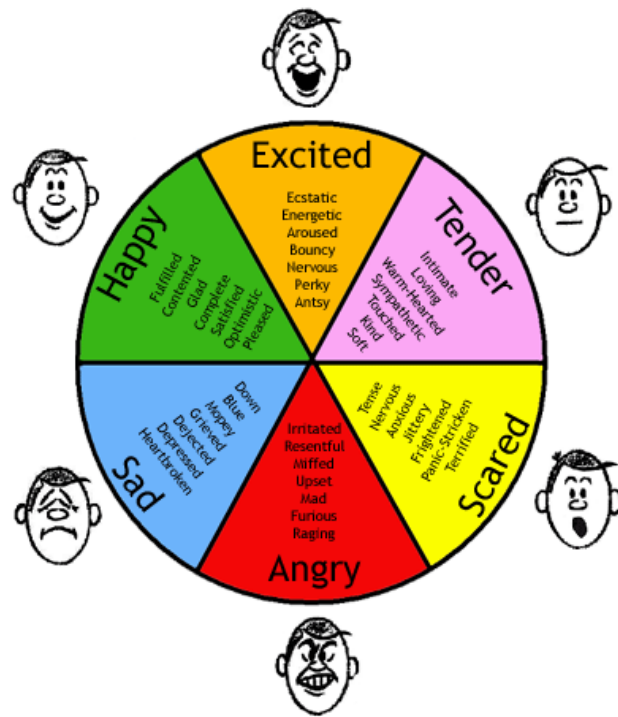
## Text Analysis with Machine Learning

Turn tweets, emails, documents, webpages and more into actionable data. Automate business processes and save hours of manual data processing.

- Positive
- Negative
- Neutral



# PODSTAWOWE EMOCJE W MEDIACH SPOŁECZNOŚCIOWYCH





# PRZYKŁAD: RODZAJE AKTYWNOŚCI PRZEDSIĘBIORSTW W SOCIAL MEDIA

## (1) Web scraping

- HTML File
- Extracting all links to find social media links
- Next iteration if the link is not present

## (2) Twitter API

- Scrap the tweet
- Process the post/tweet by Machine Learning algorithm

## (3) Machine Learning

- Classify the tweet:
  - based on C11 (ICT 2015)

```
(1) with open ('wp2_social.csv','a') as plikcsv:  
    kolumny=['URL','Facebook','Twitter','Youtube','LinkedIn','Instagram','GooglePlus']  
    zapis=csv.DictWriter(plikcsv,delimiter=';',dialect=csv.excel,fieldnames=kolumny)
```

```
training.target_names=  
['others',  
'recruitment',  
'marketing',  
'enterprise image',  
'commercials']
```

(2)

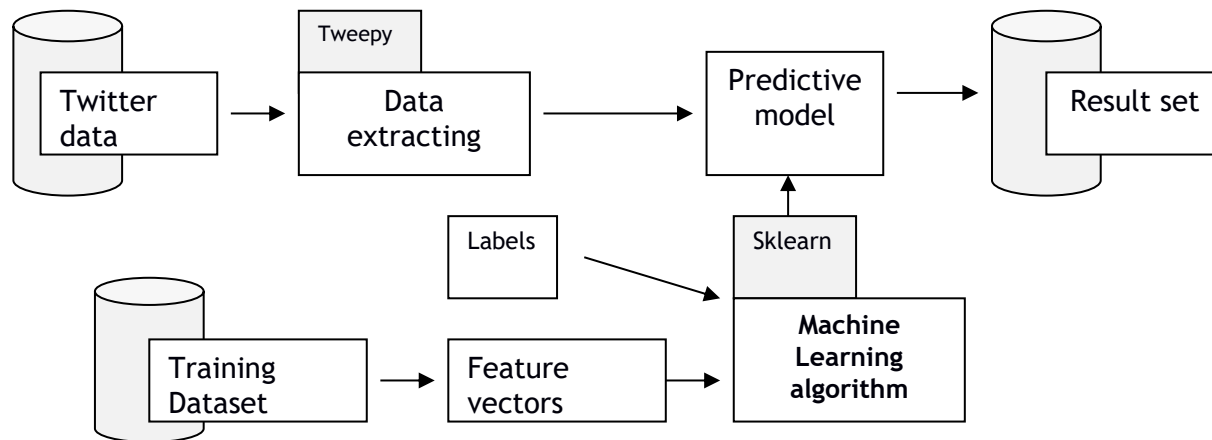
Use of Social Media		
Enterprises <u>using</u> social media are considered those that have a user profile, an account or a user licence depending on the requirements and the type of the social media.		
C10. Does your enterprise use any of the following social media? (not solely used for paid adverts) (add national examples; replace existing examples if necessary)	Yes	No
a) Social networks (e.g. Facebook, LinkedIn, Xing, Viadeo, Yammer, etc.)	<input type="checkbox"/>	<input type="checkbox"/>
b) Enterprise's blog or microblogs (e.g. Twitter, Present.ly, etc.)	<input type="checkbox"/>	<input type="checkbox"/>
c) Multimedia content sharing websites (e.g. YouTube, Flickr, Picasa, SlideShare, etc.)	<input type="checkbox"/>	<input type="checkbox"/>
d) Wiki based knowledge sharing tools	<input type="checkbox"/>	<input type="checkbox"/>
The following question ( C11 ) should only be answered if any of the above social media is used (i.e. C10 has at least one "Yes").		
C11. Does your enterprise use any of the above mentioned social media to:	Yes	No
a) Develop the enterprise's image or market products (e.g. advertising or launching products, etc)	<input type="checkbox"/>	<input type="checkbox"/>
b) Obtain or respond to customer opinions, reviews, questions	<input type="checkbox"/>	<input type="checkbox"/>
c) Involve customers in development or innovation of goods or services	<input type="checkbox"/>	<input type="checkbox"/>
d) Collaborate with business partners (e.g. suppliers, etc.) or other organisations (e.g. public authorities, non governmental organisations, etc.)	<input type="checkbox"/>	<input type="checkbox"/>
e) Recruit employees	<input type="checkbox"/>	<input type="checkbox"/>
f) Exchange views, opinions or knowledge within the enterprise	<input type="checkbox"/>	<input type="checkbox"/>

C10. Does the Website have any of the following?	Yes	No
a) Description of goods or services, price lists	<input type="checkbox"/>	<input type="checkbox"/>
*b) Online ordering or reservation or booking, e.g. shopping cart	<input type="checkbox"/>	<input type="checkbox"/>
c) Possibility for visitors to customise or design online goods or services	<input type="checkbox"/>	<input type="checkbox"/>
d) Tracking or status of orders placed	<input type="checkbox"/>	<input type="checkbox"/>
e) Personalised content in the website for regular/recurrent visitors	<input type="checkbox"/>	<input type="checkbox"/>
f) Links or references to the enterprise's social media profiles	<input type="checkbox"/>	<input type="checkbox"/>
g) Advertisement of open job positions or online job application	<input type="checkbox"/>	<input type="checkbox"/>
Optional		
C11. Does your enterprise use any of the following social media? (not solely used for paid adverts) (add national examples; replace existing examples if necessary)	Yes	No
a) Social networks (e.g. Facebook, LinkedIn, Xing, Viadeo, Yammer, etc.)	<input type="checkbox"/>	<input type="checkbox"/>
b) Enterprise's blog or microblogs (e.g. Twitter, Present.ly, etc.)	<input type="checkbox"/>	<input type="checkbox"/>
c) Multimedia content sharing websites (e.g. YouTube, Flickr, Picasa, SlideShare, etc.)	<input type="checkbox"/>	<input type="checkbox"/>
d) Wiki based knowledge sharing tools	<input type="checkbox"/>	<input type="checkbox"/>

<https://circabc.europa.eu/sd/a/a39ae859-8a16-4306-8020-ae06d3df3c91/Questionnaire%20ENT%202016.pdf>

<https://circabc.europa.eu/sd/a/7956316e-50f6-4f14-a144-055cb8af4901/Questionnaire%20ENT2015.pdf>

# PRZYKŁAD: SATYSFAKCJA Z ŻYCIA NA PODSTAWIE ANALIZY MEDIÓW SPOŁECZNOŚCIOWYCH



## EMOCJE

- happy,
- neutral,
- calm,
- upset,
- depressed,
- discouraged,
- indeterminate.



• Tweets' collecting

- Data cleaning
- repetitions (REtweet)
  - username
  - links

• Manual classification

• Verification

# PRZYKŁAD: ROZPOZNAWANIE UPRAW

**Data sources:** Satellite images, administrative data, in situ surveys.

**Methodology:**

combining data – data fusion on radar and optical remote sensing data;

data comparison with traditional surveys e.g. FSS;

combining data – administrative data sources with satellite data.

**The goal of the case study:** Crop type: look at the types of crops being grown and see if we can tell this accurately from the imagery; analysis of possibilities of using satellite images.

**Plan of Combining Datasets:** Data fusion – combining data sources by spatial reference.



# CO NALEŻY ROZWAŻYĆ PODCZAS ANALIZY TEKSTU?

- Identyfikacja języka
- Dzielenie zdań
- Tokenizacja
- Lematyzacja
- Stemming
- Analiza anaforyzmów
- Wyrażenia regularne
- POS (Part-Of-Speech)
- Rozpoznawanie encji
- Parsowanie
- Stop words

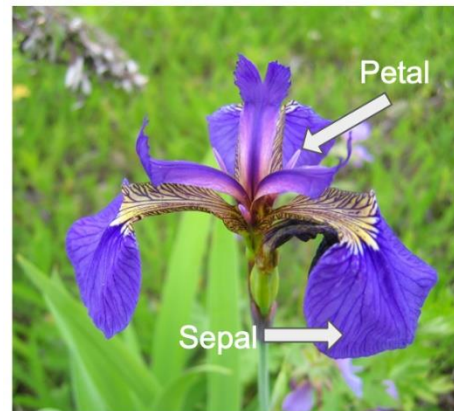
# ANALIZA TEKSTU I TEXT MINING

- Liczba słów:
  - Good morning, the weather is awful today and I don't feel good.
  - I feel not bad today, really happy.

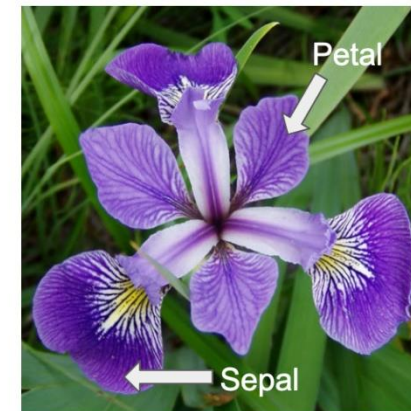
# PRZYKŁAD: WYKRYWANIE RODZAJÓW KWIATÓW NA PODSTAWIE ICH ROZMIARÓW

- Logistic Regression (LR)
- Linear Discriminant Analysis (LDA)
- K-Nearest Neighbors (KNN)
- Classification and Regression Trees (CART)
- Gaussian Naive Bayes (NB)
- Support Vector Machines (SVM)

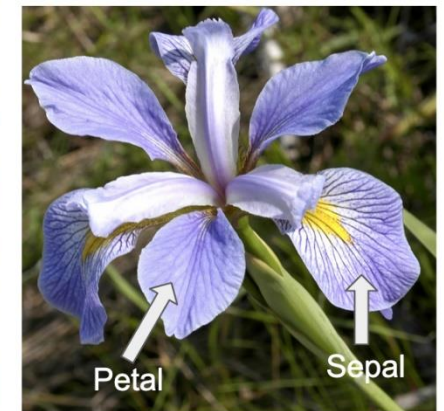
*Iris setosa*



*Iris versicolor*



*Iris virginica*



Source: [towardsdatascience.com](https://towardsdatascience.com)



# UNSUPERVISED MACHINE LEARNING



# UNSUPERVISED LEARNING – ANALIZA SKUPIEŃ, KLASTRY



$f(x)$

Gollapudi, S., & Laxmikanth, V. (2016). Practical Machine Learning. Packt Publishing.

- I feel so bad today -> **Klaster 1**
- Tomorrow I start the vacation and feel so exhilarated -> **Klaster 2**
- I am so busy and tired -> **Klaster 1**
- We are going for holidays and I feel so lovely -> **Klaster 2**



# PRZYKŁAD. UNSUPERVISED MACHINE LEARNING

- KLASTRY NA PODSTAWIE TEKSTU.

# PRZYKŁAD: NIENADZOROWANE UCZENIE DO SEGMENTACJI KLIENTÓW (NA PODSTAWIE TOWARDSDATASCIENCE.COM)

- 1. Pobierz dane
- 2. Normalizuj zbiór danych
- 3. Znajdź oczekiwaną wartość klastrow (sylwetka, łokieć)
- 4. Określ segment klientów

```
1 df_customers.head()
```

	TotalSales	OrderCount	AvgOrderValue
CustomerID			
12346.0	0.00	2	0.000000
12347.0	4310.00	7	615.714286
12348.0	1797.24	4	449.310000
12349.0	1757.55	1	1757.550000
12350.0	334.40	1	334.400000

# SILHOUETTE – OCZEKIWANA LICZBA KLASTRÓW

- Współczynnik Silhouette jest miernikiem używanym do szacowania jakości techniki grupowania.

$$S = \frac{b - a}{\max(a, b)}$$

- gdzie  $b$  jest średnią odległości między punktem a jego najbliższym klastrem,  $a$  jest średnią odległością między punktami danych w tym samym klastrze. Współczynnik sylwetki waha się od -1 do 1, gdzie im bliżej wartości są 1, tym lepsze.

# ELBOW – WALIDACJA LICZBY KLASTRÓW

- *Metoda łokcia służy do walidacji liczby skupień w klastrach k-średnich. Istotą metody łokcia jest uruchomienie grupowania k-średnich na zbiorze danych dla zakresu wartości k (np., k od 1 do 10) i dla każdej wartości k obliczane są:*
  - *Zniekształcenie (distortion), które jest średnią kwadratów odległości od centrów klastrów odpowiednich klastrów (używana jest metryka odległości euklidesowej)*
  - *Inertia (bezwładność), która jest sumą kwadratów odległości próbek do ich najbliższego środka klastru.*



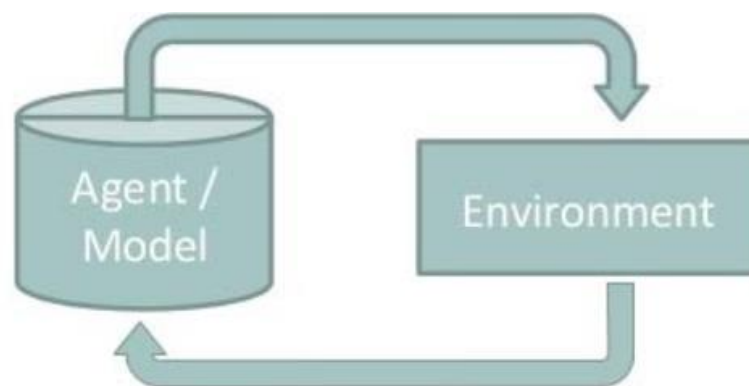
# REINFORCEMENT MACHINE LEARNING

# REINFORCEMENT LEARNING – ETYKIETY Z NAGRODAMI

$$y = f(x) \text{ z podanym } z.$$

Gollapudi, S., & Laxmikanth, V. (2016). Practical Machine Learning. Packt Publishing.

Reinforcement  
learning



- I feel so bad today -> **smutny**
- Tomorrow I start the vacation and feel so exhilarated -> **szczęśliwy**
- I am so busy and tired -> **smutny**
- We are going for holidays and I feel so lovely -> **szczęśliwy**

# PRZYKŁADY: UCZENIE ZE WZMACNIANIEM

- **Zachowanie robotów mobilnych**

- Robot mobilny musi zdecydować, czy dotrze do punktu ładowania, czy do następnego punktu ładowania, w zależności od tego, jak szybko był w stanie znaleźć punkt ładowania w przeszłości.

- **Harmonogram wind**

- Kluczowym wymogiem optymalizacji jest tutaj wybór, która winda ma być wysłana na które piętro i jest ona sklasyfikowana jako problem ze sterowaniem. Dane wejściowe to zestaw przycisków wciśniętych (wewnątrz i na zewnątrz windy) na piętrach, lokalizacjach wind i zestawie pięter. Nagrodą w tym przypadku jest najmniejszy czas oczekiwania osób, które chcą skorzystać z windy.

- **Gra w szachy**

- **Routing pakietów sieciowych**

Gollapudi, S., & Laxmikanth, V. (2016). Practical Machine Learning. Packt Publishing.

# CZYM JEST NLP?

- Przetwarzanie języka naturalnego (NLP) jest istotną dziedziną uczenia maszynowego, która zajmuje się interakcjami między językami maszynowymi (komputerowymi) i ludzkimi (naturalnymi).
- Języki naturalne nie ograniczają się do mowy i konwersacji.
- Mogą występować również w języku pisemnym i migowym.
- Dane do zadań NLP mogą mieć różne formy, na przykład tekst z postów w mediach społecznościowych, stron internetowych, a nawet recept lecarskich, dźwięk z poczty głosowej, polecenia do systemów sterowania, a nawet ulubiona muzyka lub film.

Liu, Y. (Hayden). (2017). Python Machine Learning By Example. Packt Publishing.



# CZĘŚCI MOWY – PRZYKŁADY

- **Noun** David, machine
- **Pronoun** Them, her
- **Adjective** Awesome, amazing
- **Verb** Read, write
- **Adverb** Very, quite
- **Preposition** Out, at
- **Conjunction** And, but
- **Interjection** Unfortunately, luckily
- **Article** A, the

Liu, Y. (Hayden). (2017). Python Machine Learning By Example. Packt Publishing.

# PRZYKŁAD: CZĘŚCI MOWY

- `import nltk`
- `nltk.download()`
- `from nltk.corpus import names`
- `names.words()[:10]`
- `len(names.words())`

# TOKENIZACJA

Input:

Machine learning is awesome, right?

Unigram:

Machine

learning

is

awesome

right

Bigram:

Machine learning

learning is

is awesome

awesome right

**... n-gram**

Liu, Y. (Hayden). (2017). Python Machine Learning By Example. Packt Publishing.

# PRZYKŁAD: STEMMING

Stemming to proces przywracania odmienionego lub pochodnego słowa do jego pierwotnej formy.

Liu, Y. (Hayden). (2017). Python Machine Learning By Example. Packt Publishing.

- `from nltk.stem.porter import PorterStemmer`
- `porter_stemmer = PorterStemmer()`
- `porter_stemmer.stem('lectures')`

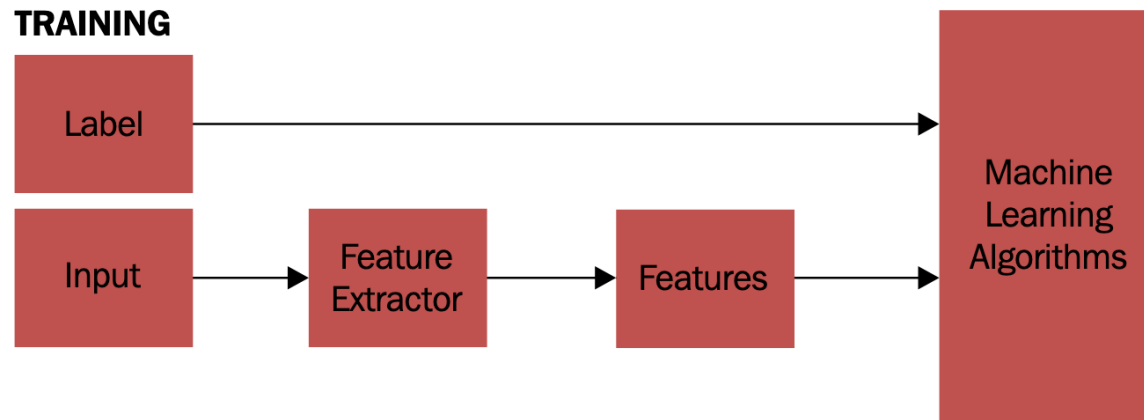
## PRZYKŁAD: LEMATYZACJA

Lemmatyzacja to ostrożna wersja stemmingu. Uwzględnia części mowy dla danego słowa.

- `from nltk.stem import WordNetLemmatizer`
  - `lemmatizer = WordNetLemmatizer()`
  - `lemmatizer.lemmatize('classes')`
- Liu, Y. (Hayden). (2017). Python Machine Learning By Example. Packt Publishing.

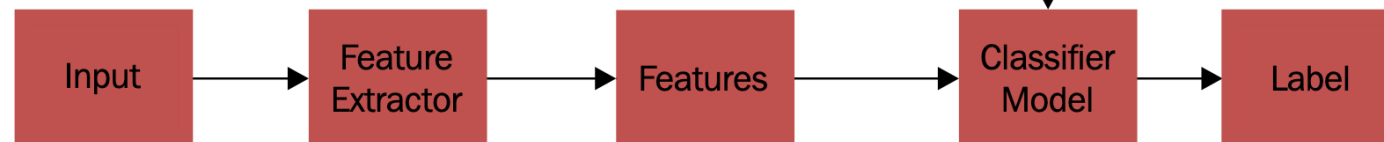
# MACHINE LEARNING ORAZ NLP

## TRAINING



Chopra, D., Mathur, I., & Joshi, N. (2016). Mastering Natural Language Processing with Python. Packt Publishing.

## PREDICTION



# PRZYKŁADY W BIZNESIE

- Klasyfikacja
  - charakterystyka przedsiębiorstwa na podstawie strony internetowej (np. e-commerce, rodzaj działalności)
  - opinie o produktach na podstawie komentarzy w sieci
  - produkty na podstawie ich opisu
  - potencjalnych klientów na podstawie ich aktywności w Internecie

# BIZNESOWE ZASTOSOWANIA UCZENIA MASZYNOWEGO

- 1. Modelowanie klienta**
- 2. Modelowanie churnu**
- 3. Dynamiczne ceny**
- 4. Segmentacja klientów**
- 5. Klasyfikacja obrazów**
- 6. Systemy rekomendacyjne**



# ISTOTA PROFILOWANIA DANYCH

- Istotnym zagadnieniem jest przygotowanie liczby słów tzw. „stop words”, które nie mają znaczenia w identyfikacji treści wiadomości.
- Są to słowa takie to np. „lub, oraz, jak, i, więc, a itd.”.
- Więcej: <http://www.ranks.nl/stopwords/polish>

## PRZYKŁAD BRAKU STOP WORDS

```
In [3]: from collections import Counter
word_freq = Counter(text_project_words).most_common()
print(word_freq[:10])
```

```
[('w', 99), ('się', 62), ('na', 56), ('z', 47), ('i', 45), ('nie', 38), ('do', 35), ('REKLAMA', 31), ('to', 25), ('-', 24)]
```



# ANALIZA SENTYMENTU

- Analiza sentymentu to określenie na podstawie zawartych słów, czy treść wiadomości jest pozytywna, negatywna lub neutralna.
- Może być wykorzystywana do badania nastrojów społecznych.
- Mogą być zbierane opinie dotyczące danego wydarzenia.
- Należy uważać na analizę komentarzy na typowych portalach.
- Przykładem dobrego wykorzystania jest badanie popularności danych tematów, jak np. w statystyce niderlandzkiej.

# ANALIZA SENTYMENTU

- Można określić podstawowe emocje – smutny, szczęśliwy, zły itp.
- Polaryzacja może być przypisana do słów pozytywnych i negatywnych, np. pozytywny: szczęśliwy, niezły, cieszę się, pozytywnie; negatywny: smutny, depresja, zdołowany, negatywnie.
- W takich przypadkach sarkazm jest niemal niemożliwy do wykrycia.

## PYTANIE

- W uczeniu maszynowym nadzorowanym wyróżnia się zbiory:
  - wzmocniony
  - treningowy
  - testowy
  - klastrowy