

3. Pozyskiwanie danych. Anotacja danych (miary zgodności).  
Integracja różnych źródeł danych. Pomiar jakości danych.  
Transformacja danych (one-hot, kodowanie, standaryzacja,  
normalizacja). Czyszczenie danych (eliminacja, imputacja, ocena  
cech).

## Przetwarzanie danych i odkrywanie wiedzy

Tomasz Kajdanowicz, Krzysztof Rajda

# Plan wykładu

1. Pozyskiwanie danych.
2. Anotacja danych (miary zgodności).
3. Integracja różnych źródeł danych.
4. Pomiar jakości danych.
5. Transformacja danych (one-hot, kodowanie, standaryzacja, normalizacja).
6. Czyszczenie danych (eliminacja, imputacja, ocena cech).

Pozyskiwanie  
danych.

# Pozyskiwanie danych

**Akwizycja danych** to proces przenoszenia danych, które zostały stworzone przez źródło, z lub spoza organizacji, do użytku produkcyjnego w organizacji.

- wraz z pojawieniem data science organizacje doszły do wniosku, że dane przedsiębiorstwa muszą być łączone z danymi zewnętrznymi (dane wewnętrzne nie wystarczą)
- potencjalne źródła danych:
  - sprzedawcy danych
  - dane przygotowane (gotowe zbiory danych - np. Google Cloud Public Datasets)
  - dostęp do otwartych danych
    - API
    - scrapping

# Zadania procesu pozyskiwania danych

1. identyfikacja potrzeby danych
2. poszukiwanie wymaganych danych
3. budowa listy źródeł kwalifikujących się jako potencjalne
4. weryfikacja licencji i ewentualne działania prawno-finansowe (np. zakup)
5. weryfikacja przykładowego zestawu danych (jeszcze przed płatnością)
6. analiza semantyczna zbiorów danych
7. ocena danych pod kątem pierwotnie ustalonych przypadków użycia
8. sprawdzenie kwestii prawnych, dotyczących prywatności i zgodności, w szczególności w odniesieniu do dozwolonego wykorzystania danych
9. negocjacje ze sprzedawcą (i zakup)
10. opracowanie specyfikacji wdrożenia z uwzględnieniem wszystkich potrzebnych operacji na danych
11. in jest źródła, pozyskiwanie jest technicznie zakończone

Anotacja danych  
(miary zgodności).

# Anotacja danych

- tworzenie modelu AI lub ML, który działa jak człowiek, wymaga dużej ilości danych uczących
- anotacja danych to oznaczanie danych w wybranej przestrzeni do zastosowań sztucznej inteligencji (np. kategoryzacja, oznaczanie lokalizacji obiektów, zaznaczanie fragmentu tekstu)
- anotacja danych zależy od rodzaju danych (np. tekst, dźwięk, obraz i wideo) i jest tak dobra jak ograniczenia poznawcze i analityczne annotatorów
- anotacja może być wykonywana w kontrolowanym eksperymencie:
  - na tzw. zamówienie
  - z wykorzystaniem nieświadomych uczestników eksperymentu

## Kluczowe elementy narzędzi do anotacji danych

Typy danych

Text

Obraz

Wideo

Dźwięk

Serie czasowe

Sensory

Wsparcie dla  
anotacji

2-D

3-D

Wideo

Dźwięk

Transkrypcja

Wizja komputerowa

Dźwięk

Transkrypcja

Text

Przetwarzanie języka naturalnego

Budowa czy zakup

Zakup komercyjnego  
narzędzia

Narzędzia darmowe  
i o otwartych źródłach

Budowa własnego  
narzędzia

Model wdrożenia

Lokalnie

Kontener

SaaS (Chmura)

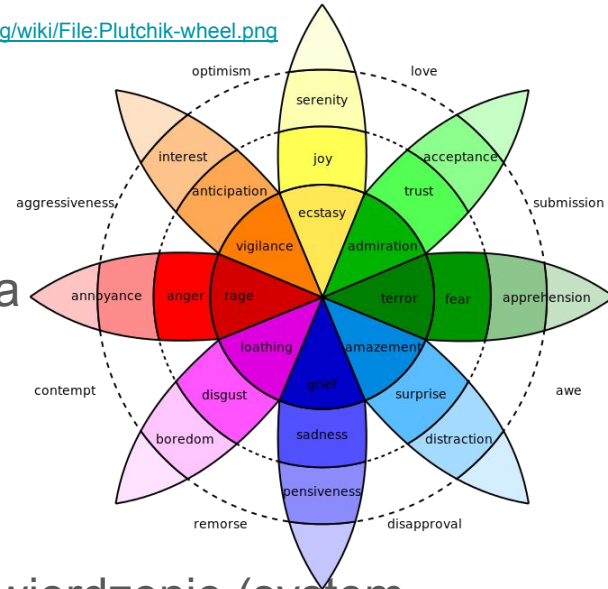


# Anotacja tekstu

szacuje się, że 70% anotowanej zawartości jest tekstowa

Przykładowe rodzaje anotacji:

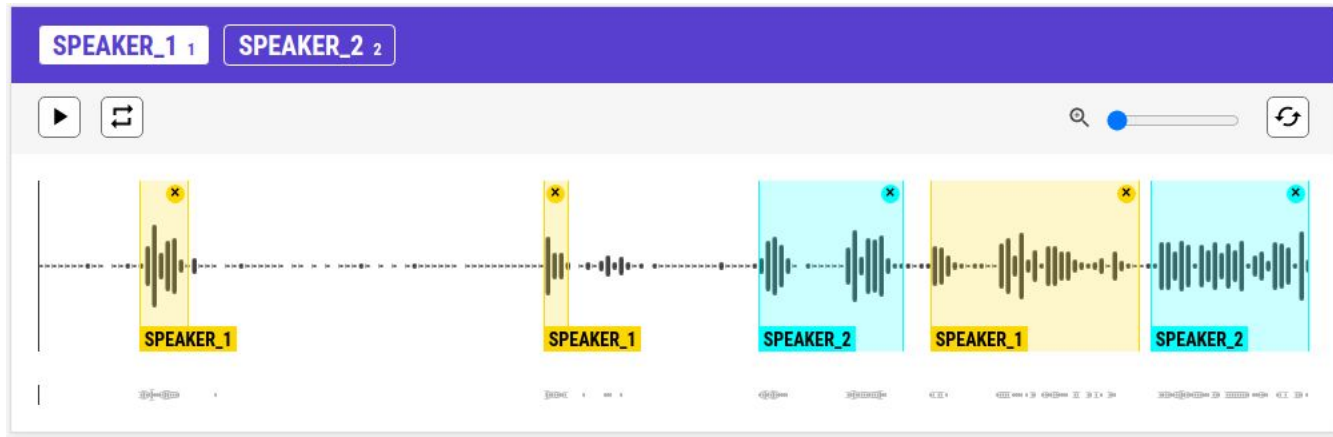
- wydźwięk - pozytywny, negatywny, neutralny
- emocje - skala Ekmana, skala Plutchika
- zamiar, intencja (intent) - np. prośba, polecenie, potwierdzenie (system dialogowy)
- semantyczna - dołączanie znaczników do pojęć i jednostek, np. ludzie, organizacje, produkty, miejsca, tematy
- nazwy własne
- opis relacji - zależności i koreferencje w dokumentach



# Anotacja dźwięku

Przykładowe rodzaje anotacji:

- transkrypcja i oznaczanie czasu danych mowy (wymowa, intonacja, dialekt, demografia)
- anotacja hałasów związanych z bezpieczeństwem
- anotacja emocji



źródło: prodi.gy -  
przykładowa  
funkcjonalność  
rozwiązania,  
<https://prodi.gy/docs/audio-video>  
Licencja otwarta.

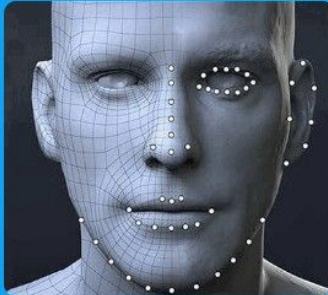
# Anotacja obrazów



 2D Bounding Boxes



 Cuboid



 Point & Landmark



 Lines & Splines



 Text Annotation



 Polygons



 Semantic Segmentation




 Video Annotation

# Anotacja wideo

SPEAKER\_1 1 SPEAKER\_2 2

▶ ⏮ 🔍 🔊 🔁



prodi.gy - przykładowa funkcjonalność rozwiązań, źródło <https://prodi.gy/docs/audio-video>  
Licencja otwarta

SPEAKER\_1 SPEAKER\_2

# Jak anotować?

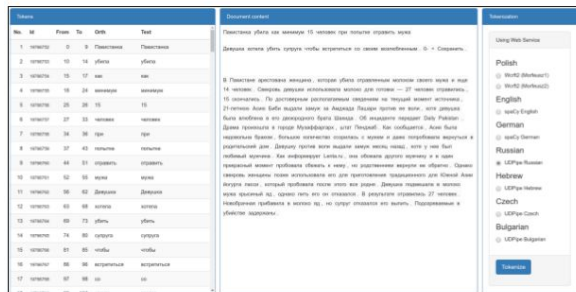
- firma zewnętrzna
- wewnątrz organizacji
  - narzędzia open source
  - własne narzędzie

## O co zadbać?

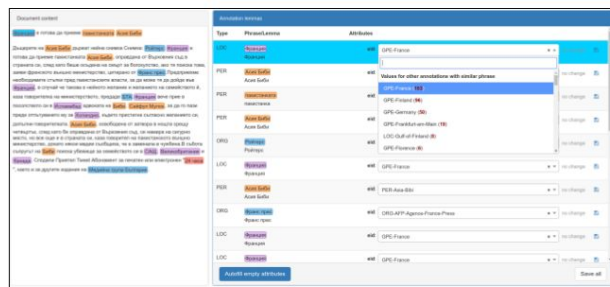
- zarządzanie zbiorami danych
- adekwatne metody anotacji
- kontrola jakości danych
- zarządzanie ludźmi i zadaniami
- bezpieczeństwo
- możliwości integracji



## Multilingual on-line text tagging



## Auto fill of annotation attributes



"2+1" annotation (morphology, NER, etc.)



## Auto annotation



# Narzędzia do anotacji danych

	Budowa własnego	Zakup
<b>Zalety</b>	<ul style="list-style-type: none"><li>• pełna kontrola nad procesem i narzędziami</li><li>• możliwość szybkiej reakcja na zmieniające się potrzeby</li></ul>	<ul style="list-style-type: none"><li>• szybszy start dzięki gotowym narzędziom</li><li>• regularne aktualizacje o najnowsze technologie i najlepsze praktyki branżowe</li></ul>
<b>Wady</b>	<ul style="list-style-type: none"><li>• czas i nakłady inwestycyjne na rozwój</li><li>• koszty bieżącej konserwacji</li></ul>	<ul style="list-style-type: none"><li>• mimo konfigurowalności, narzędzia nie są tworzone dla konkretnego przypadku użycia</li><li>• zmieniające się potrzeby projektu mogą nie być wspierane</li></ul>

# Problemy związane z anotacją danych

- stronniczość anotacji - każdy anotator ma swój własny sposób oznaczania danych które może wpływać na wydajność modelu i stronniczość w opisie zjawiska
  - uczenie na anotowanym zbiorze -> reinżynieria nawyków anotatorów
  - nieozwierciedlone wszystkie zawłośc zbioru danych (źle dobrana próbka)
  - brak udziału anotatorów w wyborze próbki
  - anotatorzy powinni opisywać zarówno zbiór treningowy i testowy
- zgodność anotacji
  - czy przy wielokrotnym oznakowaniu tego samego obiektu otrzymałem dobre anotacje?

# Słuszność a wiarygodność anotacji

- Interesuje nas ważność i słuszność anotacji
  - np. czy kategorie z anotacjami są poprawne
- Ale, w ogólności, nie ma „podstawowej prawdy” na temat anotowanych danych
  - Kategorie językowe są określane przez ludzki osąd
  - Konsekwencja: nie możemy bezpośrednio zmierzyć poprawności
- Zamiast tego mierzymy **wiarygodność i rzetelność (reliability)** anotacji
  - czy ludzcy anotatorzy konsekwentnie podejmują te same decyzje?
- Założenie: wysoka **wiarygodność i rzetelność** implikuje słuszność
- Jak można określić **wiarygodność i rzetelność**?



# Jak osiągnąć rzetelność i wiarygodność?

Podejścia:

- każdy element jest oznaczony jednym anotatorem, z losową kontrolą drugiego anotatora
  - niektóre elementy są opatrzone anotacjami przez co najmniej dwóch anotatorów
  - każdy element jest oznaczony dwoma lub więcej anotatorami - po których następuje uwspólnienie różnic
  - każdy element jest oznaczony dwoma lub więcej anotatorami - po których następuje ostateczna decyzja superanotatora (eksperta) - 2+1
- 
- we wszystkich przypadkach liczymy miary zgodności

# Podstawy mierzenia dobroci anotacji

Czułość (recall): mierzy ilość znalezionych anotacji

$$\text{Czułość} = \frac{\text{Liczba **poprawnych znalezionych** anotacji}}{\text{Liczba **poprawnych oczekiwanych** anotacji}}$$

Precyzja: mierzy jakość znalezionych anotacji

$$\text{Precyzja} = \frac{\text{Liczba **poprawnych znalezionych** anotacji}}{\text{Liczba **wszystkich znalezionych** anotacji}}$$

F-miara:

$$F = 2 \times \frac{\text{precyzja} * \text{czułość}}{\text{precyzja} + \text{czułość}}$$

# Miary zgodności

- Cohen's Kappa (1960)
- Fleiss's Kappa
- Scott's  $\pi$  (1955)
- Krippendorff's  $\alpha$  (1980)
- Rosenberg and Binkowski (2004)

# Kappa Cohena

Mierzy zgodność między dwoma anotatorami, biorąc pod uwagę możliwość przypadkowej zgodności.

$$\kappa \equiv \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

$p_o$  - faktyczna zgodność

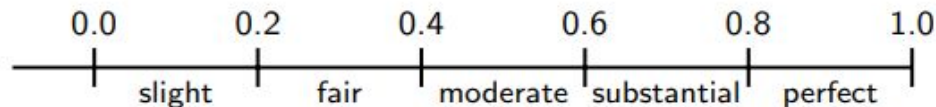
$p_e$  - oczekiwana zgodność

$$p_e = \frac{1}{N^2} \sum_k n_{k1} n_{k2}$$

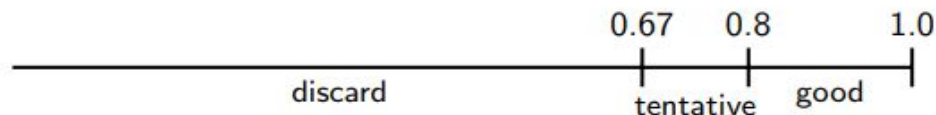
ilość przypadków

ilość wyborów w kategorii  $k$  anotatora 1

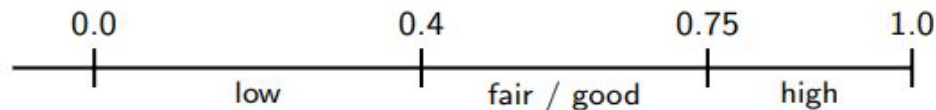
Landis and Koch, 1977



Krippendorff, 1980



Green, 1997



Integracja  
różnych źródeł  
danych.

# Integracja danych

- duża liczba źródeł danych
  - np. big data, Internet rzeczy (IoT), oprogramowanie jako usługa (SaaS), aktywność w chmurze
  - eksplozja liczby źródeł danych
  - ogromna ilość danych
  - większość danych jest zebranych i przechowywanych w samodzielnych lub oddzielnych magazynach danych

**Integracja danych** to proces, który łączy oddzielne zbiory danych w celu generowania wyższej wartości danych, możliwości realizacji procesów i z głębszej analizy.

# Integracja danych

**Integracja danych** to proces łączenia danych z różnych źródeł w celu uzyskania ujednoliconego i bardziej wartościowego obrazu, który pozwoli podejmować szybsze i lepsze decyzje.

**Integracja danych** może skonsolidować wszystkie rodzaje danych:

- ustrukturyzowane,
- nieustrukturyzowane,
- wsadowe
- przesyłane strumieniowo

w celu wykonania wszelkich operacji, od podstawowych zapytań do baz danych po złożone analizy predykcyjne.

# Wyzwania związane z integracją danych

## brak kadr

- brak doświadczonych specjalistów ds. (integracji), kosztowni
- opóźnienia analityków biznesowych
- opóźnienia analityków danych

## koszty integracji danych

- zarządzanie integracją danych to koszt zakupu technologii, wdrażania, utrzymywania i zarządzania infrastrukturą
- możliwość integracji chmurowej

## zaszłości w architekturze rozwiązań

- dane ściśle powiązane z aplikacjami
- konieczność rozdzielania warstw aplikacji i danych

## brak zarządzania zagadnieniami semantyki danych

- wiele wersji danych oznaczających to samo
- niezgodności formatów: data jako dd / mm / rr lub jako miesiąc, dzień, rok
- konieczność zastosowania „transformacji” ETL i narzędzi do zarządzania danymi podstawowymi



# Narzędzia integracji danych

- **ETL** (extract, transform, load) - najpopularniejsza metoda integracji danych
- **Data catalogs** (katalogi danych) inwentaryzowanie zasobów danych rozproszonych w wielu magazynach danych
- **Data governance tools** (narzędzia do zarządzania danymi) zarządzanie dostępnością, bezpieczeństwem, użytecznością i integralnością danych
- **Data cleansing tools** (narzędzia do czyszczenia danych)
- **Narzędzia do migracji danych** przenoszenie danych między komputerami, systemami pamięci masowej lub formatami w aplikacjach
- **Narzędzia do zarządzania danymi podstawowymi** - przestrzeganie wspólnych definicji danych
- **Konektory** - niskopoziomowe narzędzia do przenoszenia danych z jednej bazy danych do drugiej

Pomiar jakości  
danych.

# Dlaczego jakość danych jest tak ważna

złe dane + terminy = chaos i złe zarządzania

- to od jakości danych zależy główny sukces modelu, rozwiązania, biznesu
- słabe, niepoprawne, niekompletne i niewiarygodne dane nie powinny być analizowane
  - Pracownicy umysłowi marnują do 50% swojego czasu zajmując się przyziemnymi problemami z jakością danych. W przypadku analityków danych liczba ta może sięgać nawet 80%. (Sloan Management Review)
- roczny koszt słabej jakości danych w samych Stanach Zjednoczonych w 2016 r. wyniósł 3,1 bln USD (IBM)
- obsługa błędnych danych w obliczu napiętych terminów jest wyczerpujące i prawie nie rozwiązuje głównego problemu

# Zarządzanie jakością danych

**Zarządzanie jakością danych** (Data Quality Management) to zbiór praktyk, których celem jest utrzymanie wysokiej jakości informacji. DQM przechodzi całą drogę od pozyskiwania danych, przez wdrażanie zaawansowanych procesów przetwarzania danych, aż po efektywną dystrybucję danych i wspieranie decyzji.

**Jakość danych** odnosi się do oceny posiadanych informacji w odniesieniu do ich celu i zdolności do osiągnięcia tego celu.

# 5 filarów zarządzania jakością danych (w korporacjach)

## 1. Ludzie

- a. Menedżer programu - lider wysokiego szczebla, nadzór nad inicjatywami analitycznymi i czynnościami obejmującymi zakres danych, wizja dotycząca jakości danych i zwrotu z inwestycji
- b. Menedżer zmian w organizacji - organizuje transformacje danymi, zapewniając przejrzystość i wgląd w zaawansowane rozwiązania w zakresie technologii danych, określa i komunikuje jakość danych
- c. Analityk biznesowy / danych - definiuje potrzeby jakościowe i ilościowe w modelach danych w celu ich osiągnięcia

## 2. Profilowanie danych

- a. szczegółowy przegląd danych
- b. rozpoznanie struktur, zawartości, relacji
- c. podsumowanie statystyczne
  - i. np:

[https://pandas-profiling.github.io/pandas-profiling/examples/master/census/census\\_report.html](https://pandas-profiling.github.io/pandas-profiling/examples/master/census/census_report.html)

# 5 filarów zarządzania jakością danych (w korporacjach)

## 3. Określenie miar jakości danych

- a. utworzone „zasady jakości” - reguły biznesowe / techniczne, z którymi dane muszą być zgodne
- b. reguły wykorzystują metody pomiaru jakości danych

## 4. Raportowanie jakości danych

- a. proces rejestrowania i egzekwowania zasad dotyczących jakości danych
- b. raportowanie stanu jakości i wyjątków
- c. sedno kontroli jakości danych - zapewnia wgląd w stan danych w dowolnym momencie w czasie rzeczywistym

## 5. Naprawa danych

- a. dwuetapowy proces określania: najlepszego sposobu na naprawę danych, najbardziej efektywnego sposób wprowadzenia tej zmiany
- b. naprawa pierwotnej przyczyny

# Pomiar jakości danych (1/2)

- ACCIT - Accuracy, Consistency, Completeness, Integrity, and Timeliness (dokładność, spójność, kompletność, integralność i terminowość)
- pomiar jakości za pomocą zestawu reguł i procedur

## Miary:

- **dokładność** - stosunek ilości znanych błędów danych do wszystkich danych, (np. brakujący, niepełny lub nadmiarowy wpis)
  - wskaźnik powinien z czasem rosnąć, im wyższy, tym lepiej, np. oczekiwane min 95% dokładności
- **spójność** - określa, czy dwie wartości pobrane z oddzielnych zestawów danych nie kolidują ze sobą (nie oznacza automatycznie poprawności)
  - np. reguła, która sprawdzi, czy suma pracowników w każdym dziale firmy nie przekracza całkowitej liczby pracowników w tej organizacji
- **kompletność** - czy każdy wpis danych jest „pełnym” wpisem? zestawy rekordów danych nie powinny zawierać żadnych istotnych luk informacji
  - np. liczba pustych wartości w zestawie danych

# Pomiar jakości danych (2/2)

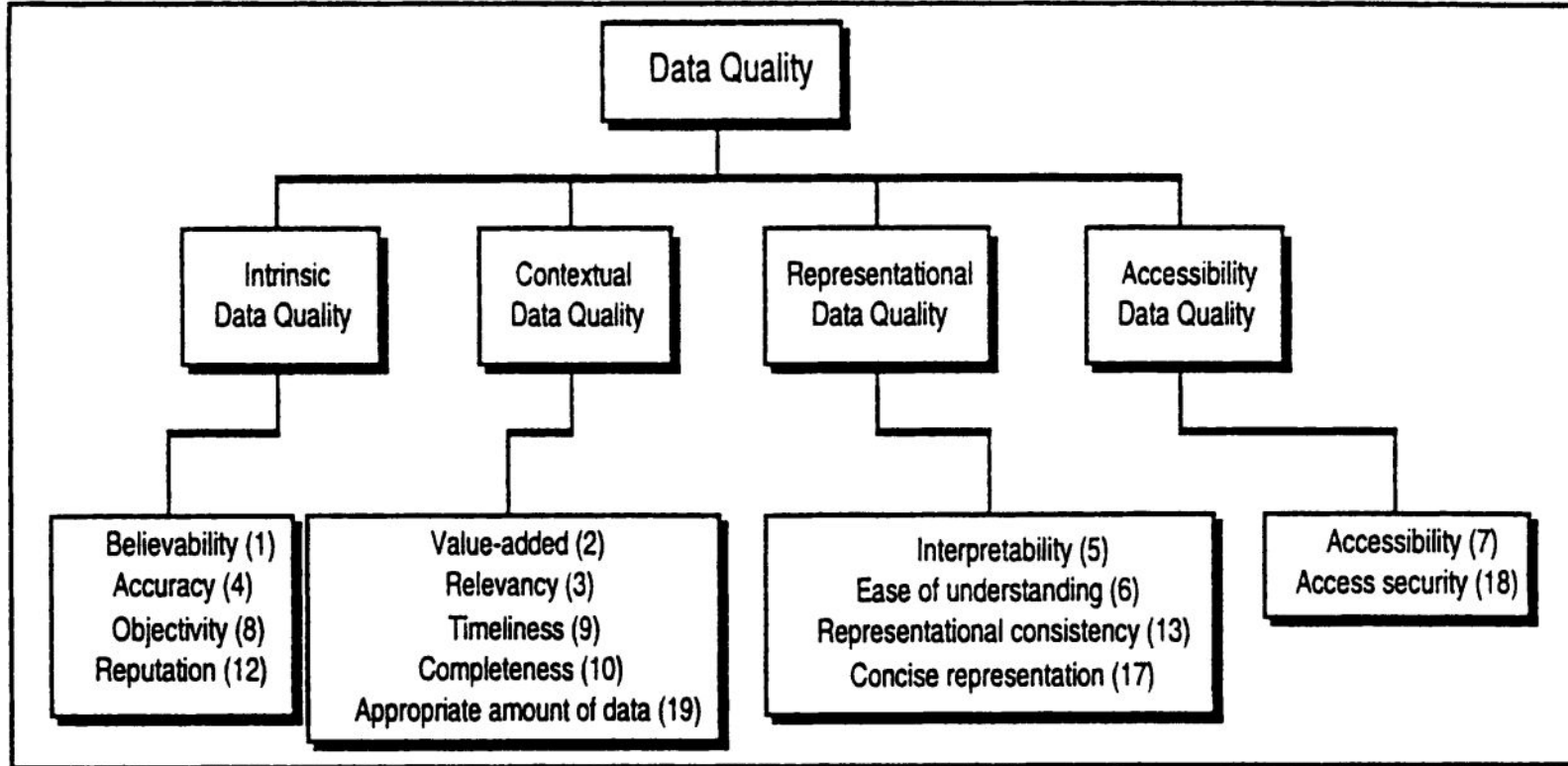
- ACCIT - Accuracy, Consistency, Completeness, Integrity, and Timeliness (dokładność, spójność, kompletność, integralność i terminowość)
- pomiar jakości za pomocą zestawu reguł i procedur

Miary:

- **integralność** - na ile jest zachowana strukturalna zgodność pomimo niezamierzonych błędów
  - np. współczynnik błędów transformacji danych - ile operacji transformacji danych kończy się niepowodzeniem w stosunku do całości; błędy powodują “rozjazdy” w danych
- **terminowość** - mierzy czas oczekiwania na dane do momentu gdy są one gotowe do użycia
  - niezbędne, aby zmierzyć i zoptymalizować ten czas zgodnie z wymaganiami biznesowymi



# A Conceptual Framework of Data Quality



Transformacja  
danych (one-hot,  
kodowanie,  
standaryzacja,  
normalizacja).

# Kodowanie zmiennych

- wiele algorytmów uczenia maszynowego nie może działać bezpośrednio na etykietach (np. czerwony, zielony, żółty) - **wszystkie zmienne muszą być numeryczne**
- jest to głównie ograniczenie wydajnej implementacji algorytmów uczenia maszynowego, a nie twarde ograniczenia samych algorytmów
- dane jakościowe muszą zostać przekonwertowane na postać liczbową

Możliwe kodowanie:

1. Kodowanie liczbami całkowitymi
2. Kodowanie one-hot (“na gorąco”)

# Kodowanie liczbami całkowitymi

- każdej unikalnej wartości kategorii jest przypisywana wartość całkowita
- np. „czerwony” to 1, „zielony” to 2, a „niebieski” to 3.
- jest łatwo odwracalne
- w przypadku niektórych zmiennych wystarcza
- wartości całkowite mają naturalny uporządkowany związek między sobą, a algorytmy uczenia maszynowego mogą być w stanie zrozumieć i wykorzystać tę zależność (zaleta i jednocześnie wada)

# Kodowanie one-hot

w przypadku zmiennych dyskretnych w których nie istnieje zależność porządkowa, kodowanie liczbami całkowitymi nie wystarcza

przyjęcia nienaturalnej kolejności między kategoriami może skutkować słabą wydajnością lub nieoczekiwanymi wynikami (np. przewidywania w połowie drogi między kategoriami)

one-hot: usuwana jest zmienna zakodowana w postaci liczby całkowitej i dodawana jest nowa zmienna binarna dla każdej unikalnej wartości całkowitej

kodowanie kolorów to 3 kategorie i dlatego potrzebne są 3 zmienne binarne

wartość „1” jest umieszczana w zmiennej binarnej dla koloru, a wartości „0” dla innych kolorów

np.

czerwony	zielony	niebieski
1	0	0
0	1	0
0	0	1

# Skalowanie

## Normalizacja:

Polega na przeskalowaniu wartości tak, żeby ich zakres mieścił się w przedziale  $<0,1>$ . Transformacja wymaga dokładnego oszacowania wartości minimalnych i maksymalnych.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

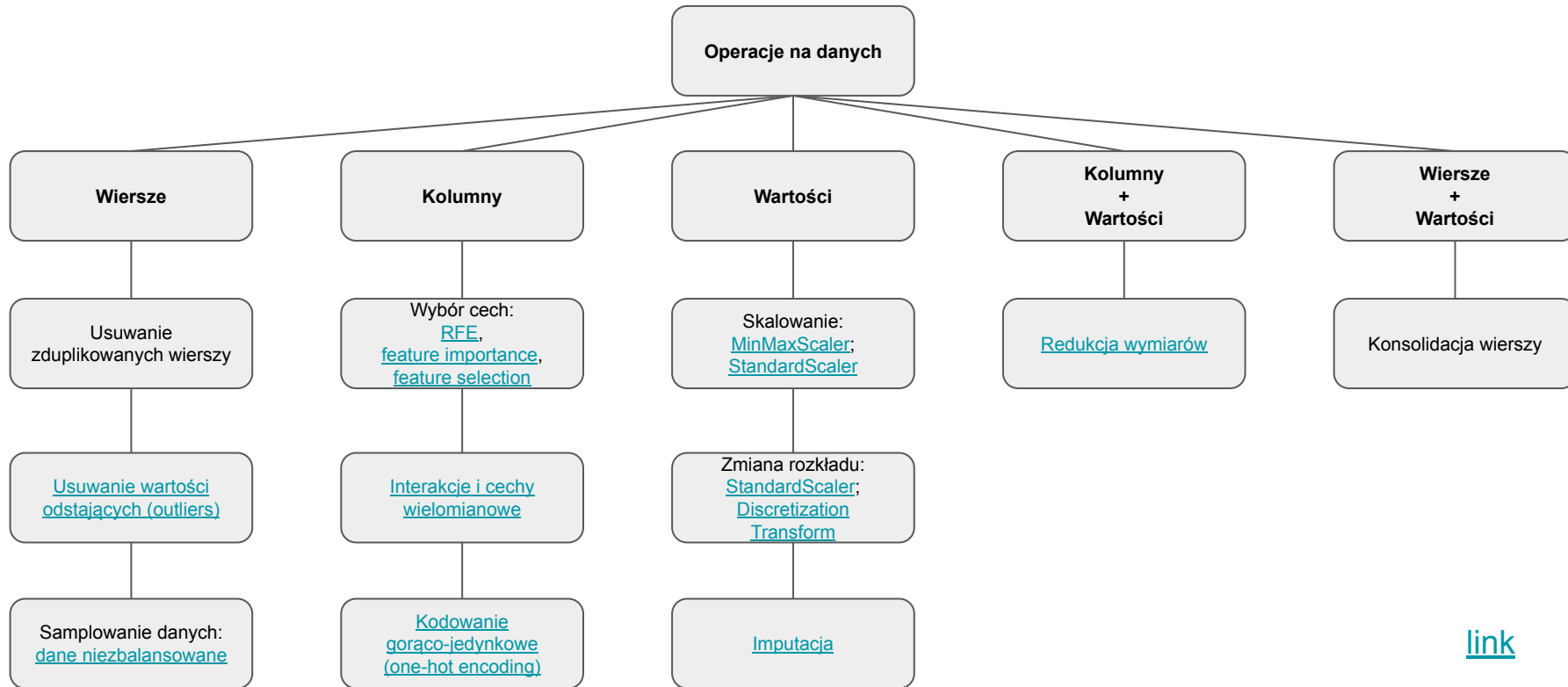
## Standaryzacja:

Polega na przeskalowaniu wartości tak, żeby średnia wynosiła 0, a odchylenie standardowe 1. Dane powinny tworzyć rozkład normalny, choć i bez tego warunku w pewnych sytuacjach możemy poprawić efektywność uczenia maszynowego.

$$X' = \frac{X - \mu}{\sigma}$$

Czyszczenie  
danych  
(eliminacja,  
imputacja, ocena  
cech).

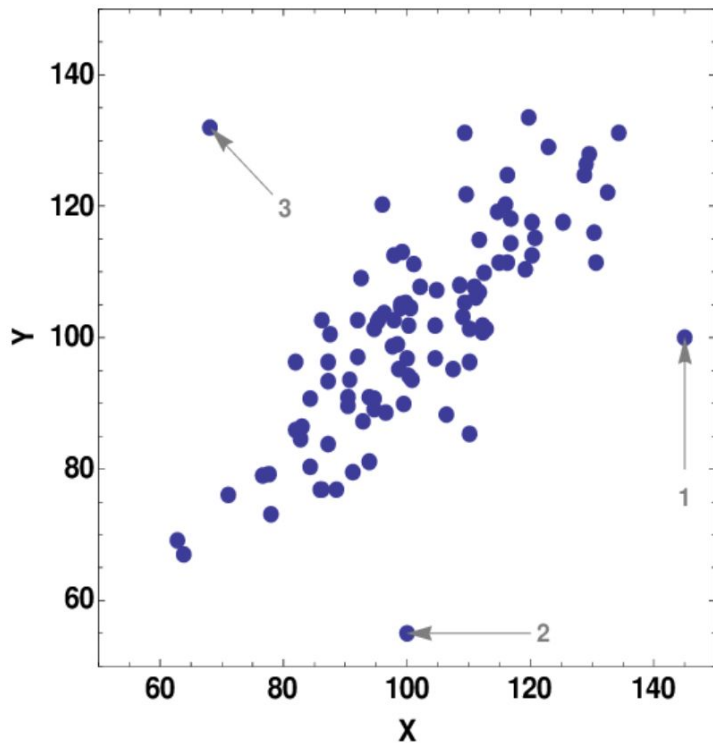
# Czyszczenie i transformacje danych





# Wartość odstająca

## Outlier



Definition of Hawkins (1980):

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”

[link](#)

# Usuwanie wartości odstających (outliers)

“Even with a thorough understanding of the data, outliers can be hard to define. [...] Great care should be taken not to hastily remove or change values, especially if the sample size is small.” [page 33, Applied Predictive Modeling, 2013]

Metoda oparta na odchyleniu standardowym:

Dane o rozkładzie normalnym

1 sigma od średniej: 68%  
2 sigmy od średniej: 95%  
3 sigmy od średniej: 99.7%

Dla małych zbiorów danych: 2 sigmy  
Dla dużych zbiorów danych: 4 sigmy

Isolation Forest [iForest]:

“... our proposed method takes advantage of two anomalies' quantitative properties: i) they are the minority consisting of fewer instances and ii) they have attribute-values that are very different from those of normal instances.”

— Isolation Forest, 2008.

Interquartile Range [IQR]:

Dane niemające rozkładu normalnego

Percentyle oblicza się sortując dane rosnąco i wybierając wartości z odpowiednich indeksów: 50. percentyl jest wartością na środku, 25 percentyl jest wartością z indeksem będącym w  $\frac{1}{4}$  drogi od minimum do maximum, a 75 percentyl analogicznie w  $\frac{3}{4}$  drogi.

kwartył 1 = percentyl 25  
kwartył 2 = percentyl 50  
kwartył 3 = percentyl 75

$IQR = \text{percentyl } 75 - \text{percentyl } 25$

Usuń wartości spoza przedziału  $k * IQR$ ,  $k=1.5 \dots 3$

Klasyfikacja jedno-klasowa (one class classification, OCC):

Uczenie z nadzorem na rozkładzie normalnym. Następnie wyrzucenie danych niezaklasyfikowanych jako dane z rozkładu normalnego

Local Outlier Factor [LOF]:

wykorzystanie techniki nearest neighbor do przypisania każdej wartości oceny - zbyt mało danych w bliskim sąsiedztwie to potencjalny outlier

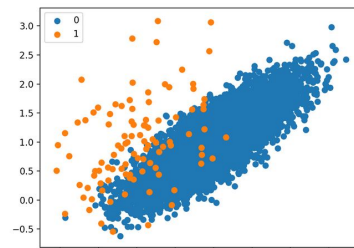
# Samplowanie danych

Synthetic Minority Oversampling Technique (SMOTE) - Nitesh Chawla i inni, rok 2002, artykuł zatytułowany “SMOTE: Synthetic Minority Over-sampling Technique.”

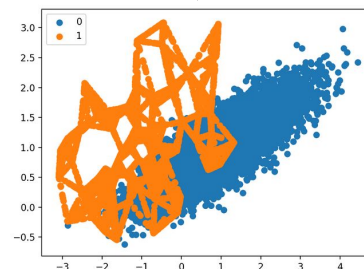
Wybiera się jeden element z klasy znacząco mniej licznej. Następnie znajduje się  $k$  (z reguły  $k=5$ ) sąsiednich punktów i wybiera losowo jeden z nich. Nowy, syntetyczny punkt jest kreowany pomiędzy nimi w przestrzeni cech.

Powyższe kroki powtarza się iteracyjnie (np. tak długo, żeby zrównać liczbę punktów obu klas).

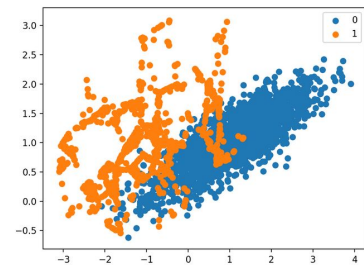
Na koniec stosuje się undersampling na obu klasach tak, żeby zmienić pierwotny stosunek (np. 1:100) na bliższy równowadze (np. 1:2).



oversampling



undersampling



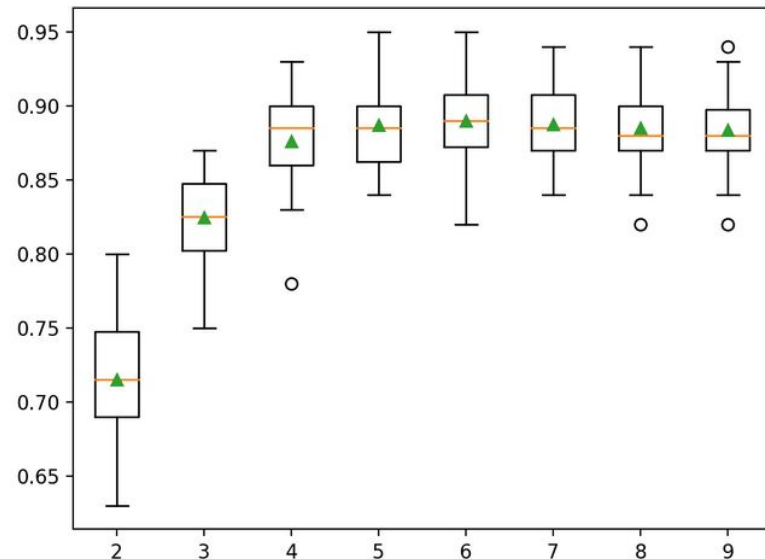
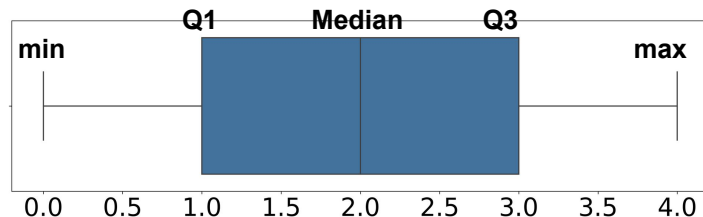
# Recursive Feature Elimination (RFE)

Rekursywna eliminacja cech pozwala wybrać cechy (kolumny) mające związek ze zmienną przewidywaną Y.

Po wybraniu algorytmu przeprowadzana jest klasyfikacja. Najpierw na całym zestawie cech, a następnie na ich podzbiorach. Na podstawie poszczególnych wyników, w kolejnych krokach usuwane są najmniej znaczące cechy.

“When the full model is created, a measure of variable importance is computed that ranks the predictors from most important to least. [...] At each stage of the search, the least important predictors are iteratively eliminated prior to rebuilding the model.”

— Pages 494-495, Applied Predictive Modeling, 2013.



Box Plot of RFE Number of Selected Features vs. Classification Accuracy

Źródło:

<https://machinelearningmastery.com/rfe-feature-selection-in-python/>

# Interakcje i cechy wielomianowe

Interakcje i cechy wielomianowe pozwalają dobrze modelować dane wykorzystując proste algorytmy. Czasami poprawa wyniku jest na tyle duża, że warto dodać kilka nowych kolumn danych.

## Interakcje

Nowe kolumny, w których  
przechowujemy iloczyny  
poszczególnych kolumn  
 $x_1 * x_2, x_1 * x_3, \dots$

## Cechy wielomianowe

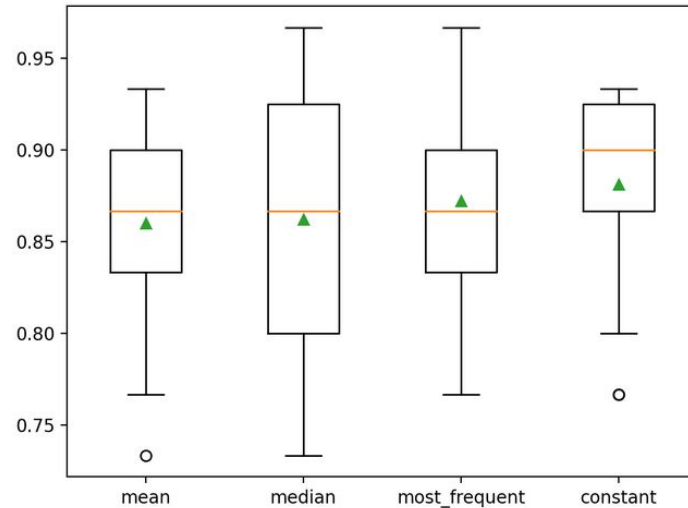
Nowe kolumny, w których  
przechowujemy potęgi  
poszczególnych kolumn  
 $x_1^1, x_1^2, x_1^3, \dots$

# Imputacja

Zamiast usuwać wiersze z brakującymi danymi, można dane uzupełniać w obrębie poszczególnych kolumn stosując różne podejścia. Dla danej kolumny oblicza się:

1. średnią
2. medianę
3. modę

lub decyduje się na jakąś wartość stałą i taką wpisuje się w miejsca, gdzie brakuje wartości.



Box and Whisker Plot of Statistical Imputation Strategies Applied to the Horse Colic Dataset

Źródło: <https://machinelearningmastery.com/statistical-imputation-for-missing-values-in-machine-learning/>

Dziękuję za uwagę.