2 Podstawowe techniki eksploracji danych.

Wartości atrybutów.

• Wartości atrybutów to liczby lub symbole przypisane do atrybutu

• Rozróżnienie między atrybutami i wartościami atrybutów

- Ten sam atrybut można zamapować na różne wartości atrybutów

• Przykład: wysokość można mierzyć w stopach lub metrach

- Różne atrybuty można przypisać do zestawu wartości

• Przykład: Wartości atrybutów dla identyfikatora i wieku to liczby całkowite

• Ale właściwości wartości atrybutów mogą być różne

- ID nie ma ograniczeń, ale wiek ma wartość maksymalną i minimalną

Typy atrybutów.

• Istnieją różne typy atrybutów

- Nominalna

• Przykłady: numery identyfikacyjne, kolor oczu, kody pocztowe

- Porządkowy

• Przykłady: rankingi (np. Smak chipsów ziemniaczanych w skali od 1 do 10), oceny, wzrost w {wysoki, średni, niski}

- Interwał

• Przykłady: daty kalendarzowe, temperatury w stopniach Celsjusza lub Fahrenheita.

- Stosunek

• Przykłady: temperatura w kelwinach, długość, czas, liczba

Własności wartości atrybutów.

• Typ atrybutu zależy od tego, które z następujących właściwoąśi posiada:

- Oddzielanie: $= \neq$

- Porządek: $< >$

- Dodawanie: $+ -$

- Mnożenie: $\cdot \div$

- Atrybut nominalny: oddzielanie

- Atrybut porządkowy: oddzielanie i porządek

- Atrybut przedziału: oddzielanie, porządek i dodawanie

- Atrybut proporcji: wszystkie 4 właściwości

Attribute Type	Description	Examples	Operations
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, χ^2 test
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current	geometric mean, harmonic mean, percent variation

Attribute Level	Transformation	Comments
Nominal	Any permutation of values	If all employee ID numbers were reassigned, would it make any difference?
Ordinal	An order preserving change of values, i.e., *new_value = f(old_value)* where *f* is a monotonic function.	An attribute encompassing the notion of good, better best can be represented equally well by the values {1, 2, 3} or by {0.5, 1, 10}.
Interval	*new_value =a * old_value + b* where a and b are constants	Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).
Ratio	*new_value = a * old_value*	Length can be measured in meters or feet.

Dyskretne i ciągłe atrybuty.

• Dyskretny atrybut

- ma tylko skończony lub policzalnie nieskończony zbiór wartości

- Przykłady: kody pocztowe, liczby lub zestaw słów w pliku, zbiór dokumentów

- Często przedstawiane jako zmienne całkowite.

- Uwaga: atrybuty binarne są specjalnym przypadkiem dyskretnych

• Ciągły atrybut

- Ma liczby rzeczywiste jako wartości atrybutów

- Przykłady: temperatura, wzrost lub waga

- W praktyce rzeczywiste wartości można tylko mierzyć i przedstawiać przy użyciu skończonej liczby cyfr

- Atrybuty ciągłe są zwykle reprezentowane jako zmienne zmiennoprzecinkowe

Rodzaje zbiorów danych.

• Rekordy

- DataMatrix

- DocumentData

- TransactionData

• Grafy

- WorldWideWeb

- Struktury molekularne

• Uporządkowane

- Dane przestrzenne

- Dane czasowe

- Dane sekwencyjne

- Dane sekwencji genetycznej

Charakterystyki danych ustrukturowionych.

• Wymiarowość

- Przekleństwo wymiarowości

• Rzadkość

- Liczy się tylko obecność

• Rozdzielczość

- Wzory zależą od skali

Dane rekordowe.

Dane, które składają się ze zbioru rekordów, z których każdy składa się z ustalonego zestaw atrybutów.

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Macierz danych.

• Jeśli obiekty danych mają ten sam ustalony zestaw atrybutów liczbowych, wówczas obiekty danych można traktować jako punkty w przestrzeni wielowymiarowej, gdzie każdy wymiar reprezentuje odrębny atrybut

• Taki zestaw danych może być reprezentowany przez macierz m na n, w której istnieje m wierszy, po jednym dla każdego obiektu i n kolumn, po jednym dla każdego atrybutu

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Dane dokumentów.

• Każdy dokument staje się wektorem „termów”,

- każdy term jest składnikiem (atrybutem) wektora,

- wartością każdego składnika jest liczba razy ile odpowiedni term występuje w dokumencie.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Dane transakcyjne.

• Specjalny typ danych rekordu, gdzie

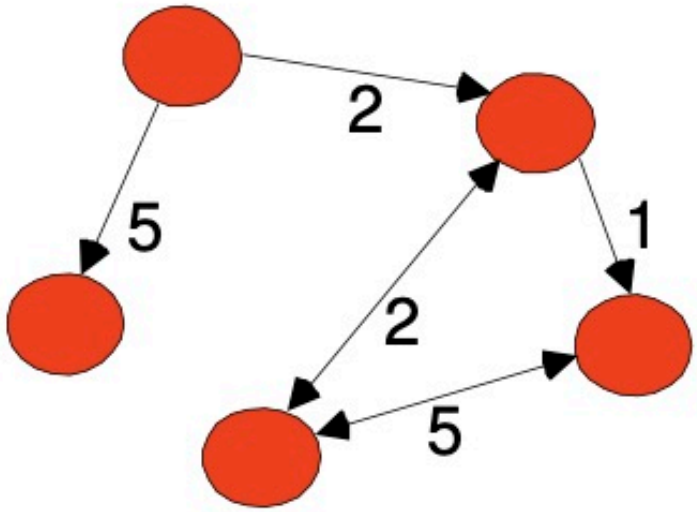
- każdy rekord (transakcja) obejmuje zbiór pozycji.

- Weźmy na przykład pod uwagę sklep spożywczy. Zestaw produktów zakupionych przez klienta podczas jednego wyjazdu na zakupy stanowi transakcję, a poszczególne produkty, które zostały zakupione, są przedmiotami.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Dane grafu.

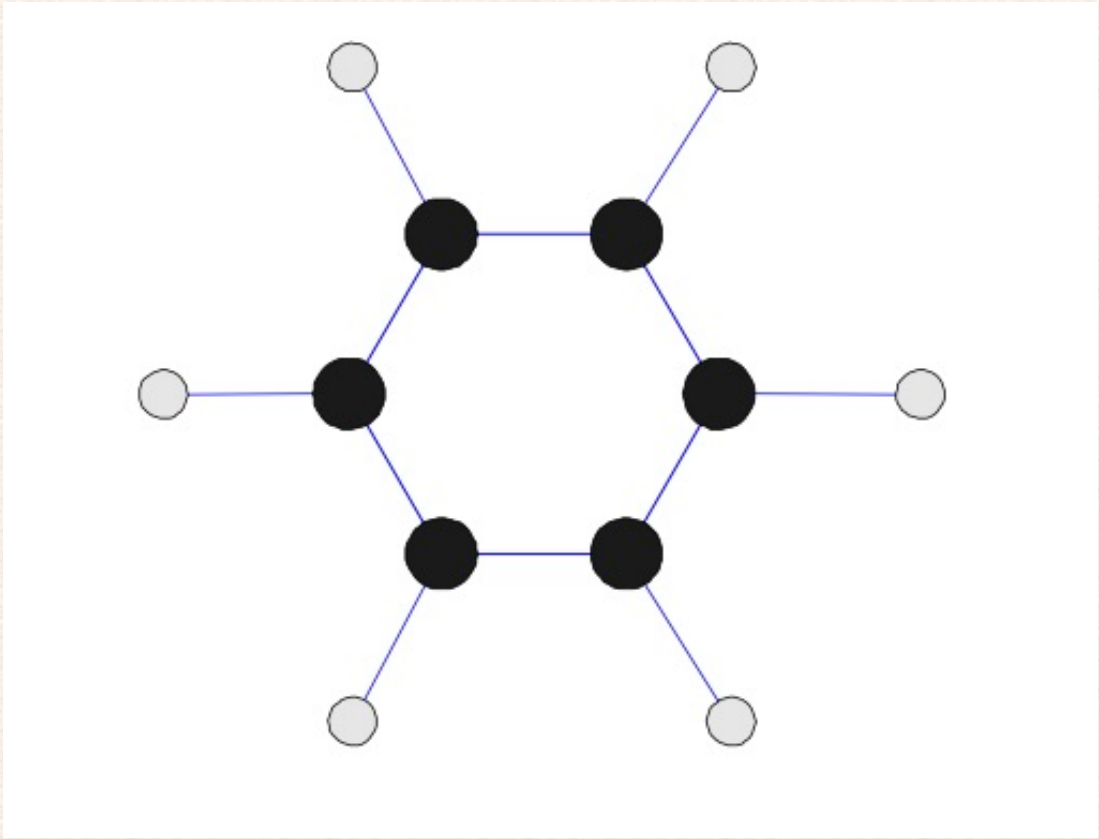
Przykłady: ogólny wykres i linki HTML




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

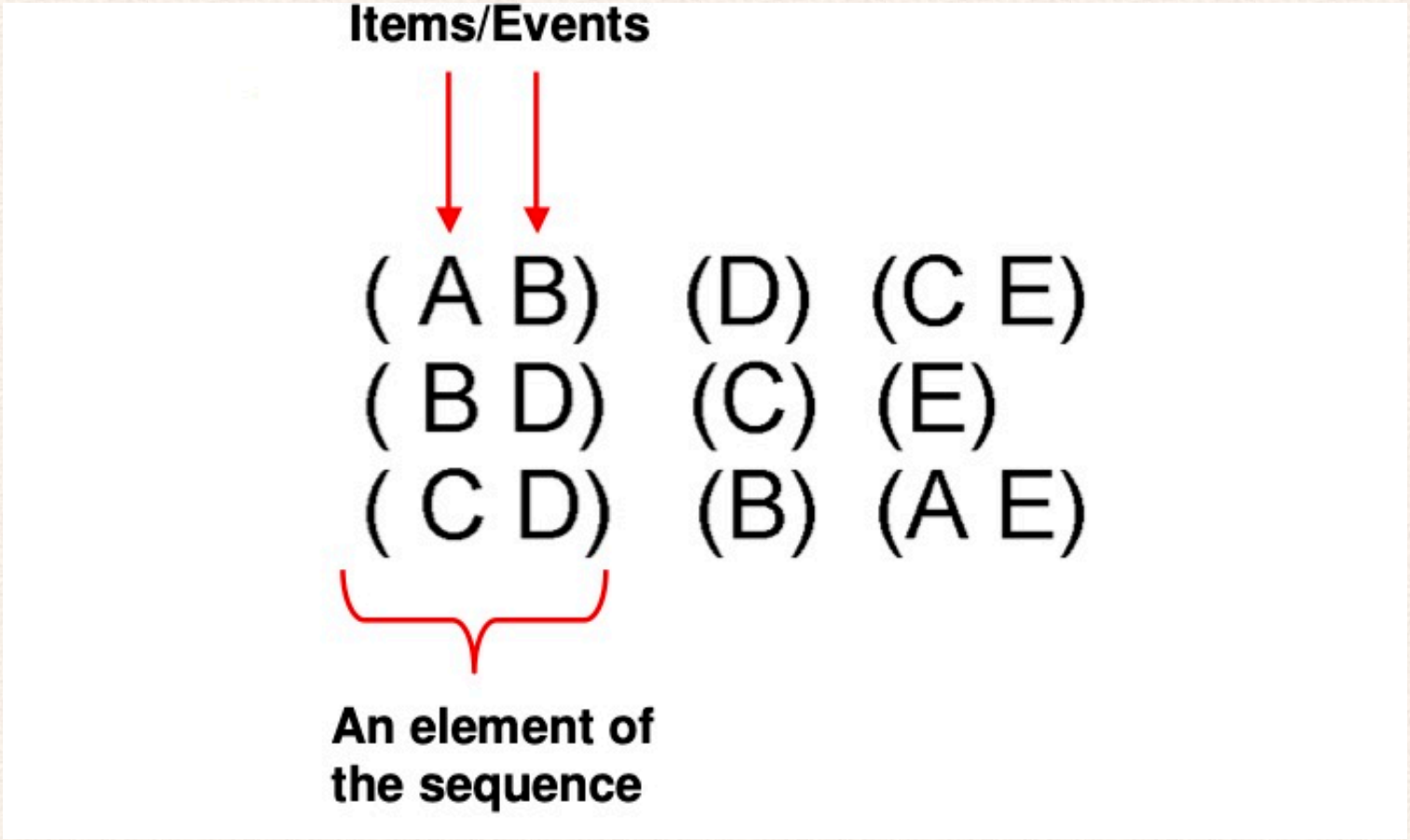
Dane chemiczne.

Cząsteczka benzenu C_6H_6



Dane uporządkowane.

Ciągi transakcji



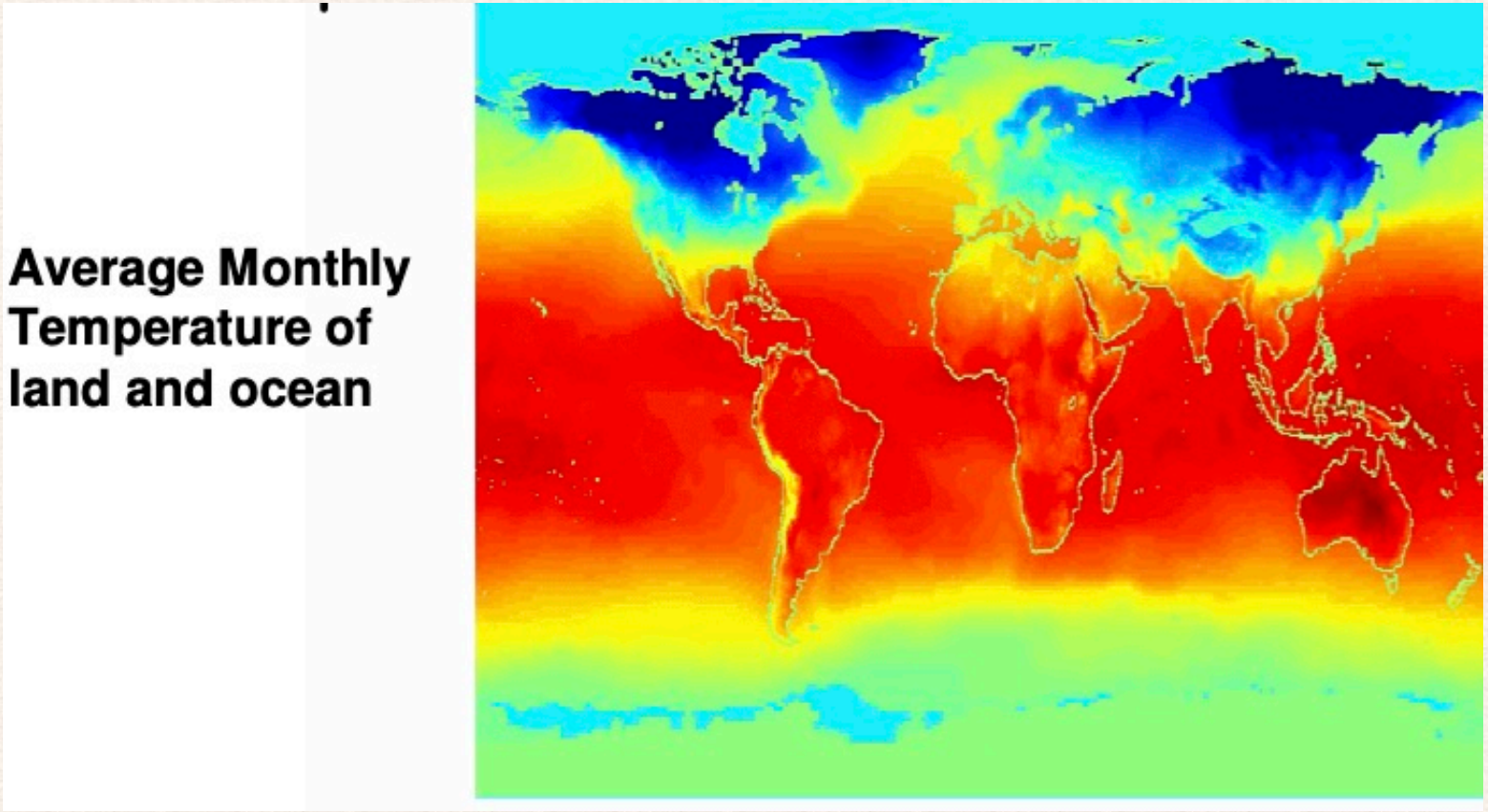
Dane uporządkowane.

Dane sekwencji genomu

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Dane uporządkowane.

Dane czasoprzestrzenne.



Jakość danych.

- Jakie rodzaje problemów z jakością danych?
- Jak możemy wykryć problemy z danymi?
- Co możemy zrobić z tymi problemami?
- Przykłady problemów z jakością danych:

- szum i wartości odstające

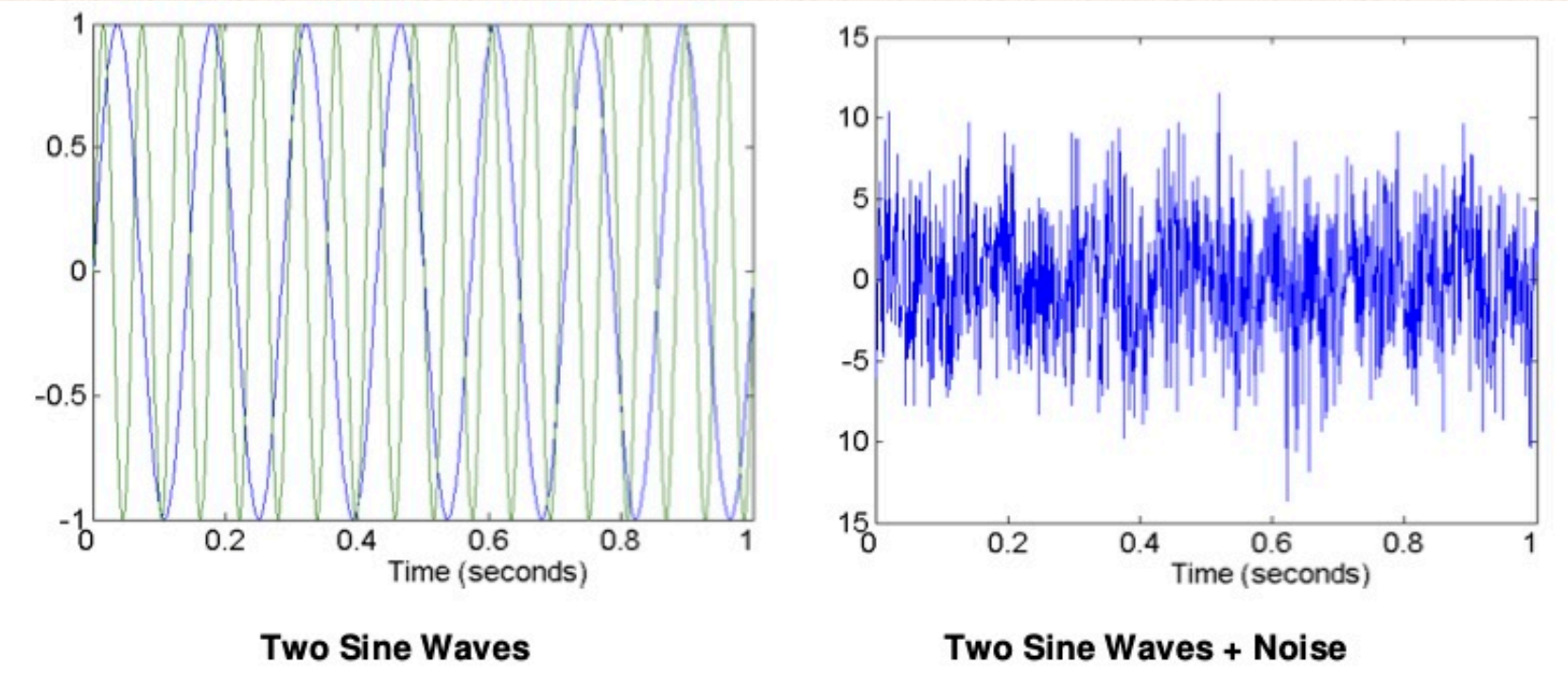
- brakujące wartości

- zduplikowane dane

Szum.

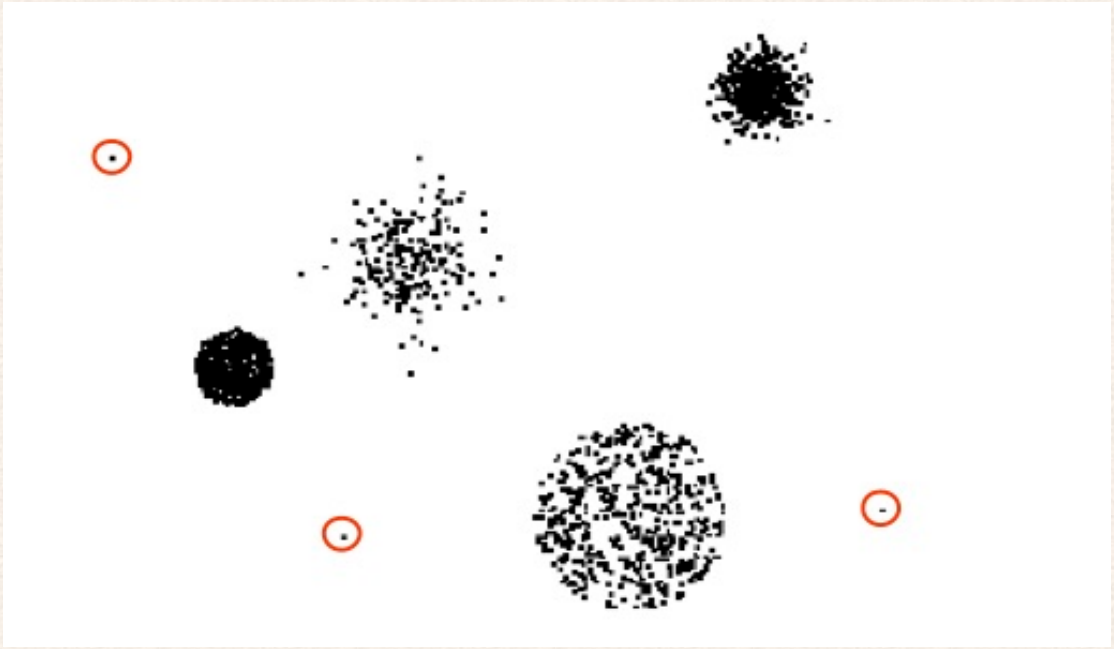
• Szum odnosi się do modyfikacji oryginalnych wartości

- Przykłady: zniekształcenie głosu osoby podczas rozmowy przez kiepski telefon i „śnieg" w telewizji



Wartości odstające.

• Wartości odstające to obiekty danych, których właściwości znacznie różnią się od większości innych obiektów danych w zestawie danych



Brakujące wartości.

• Przyczyny brakujących wartości

- Informacje nie zostały zebrane

(np. ludzie odmawiają podania swojego wieku i wagi)

- Atrybuty mogą nie mieć zastosowania w każdym przypadku (np. Roczny dochód nie dotyczy dzieci)

• Obsługa braków danych

- Eliminacja obiektów danych

- Oszacuj brakujące wartości

- Zignoruj brakującą wartość podczas analizy

- Zastąp wszystkimi możliwymi wartościami (ważonymi według ich prawdopodobieństw)

Powtarzające się dane.

• Zestaw danych może zawierać obiekty danych, które są duplikatami lub prawie duplikatami siebie

- poważny problem podczas łączenia danych z niejednorodnych źródeł

• Przykłady:

- Ta sama osoba z wieloma adresami e-mail

• Czyszczenie danych

- Proces rozwiązywania problemów z podwójnymi danymi

Wstępna obróbka danych.

- Agregacja
- Pobieranie próbek
- Redukcja wymiarowości
- Wybór podzbioru funkcji
- Tworzenie funkcji
- Dyskretyzacja i binaryzacja
- Transformacja atrybutów

Agregacja.

- Łączenie dwóch lub więcej atrybutów (lub obiektów) w jeden atrybut (lub obiekt)
- Cel

- Redukcja danych

  $\rightarrow$ Zmniejsz liczbę atrybutów lub obiektów

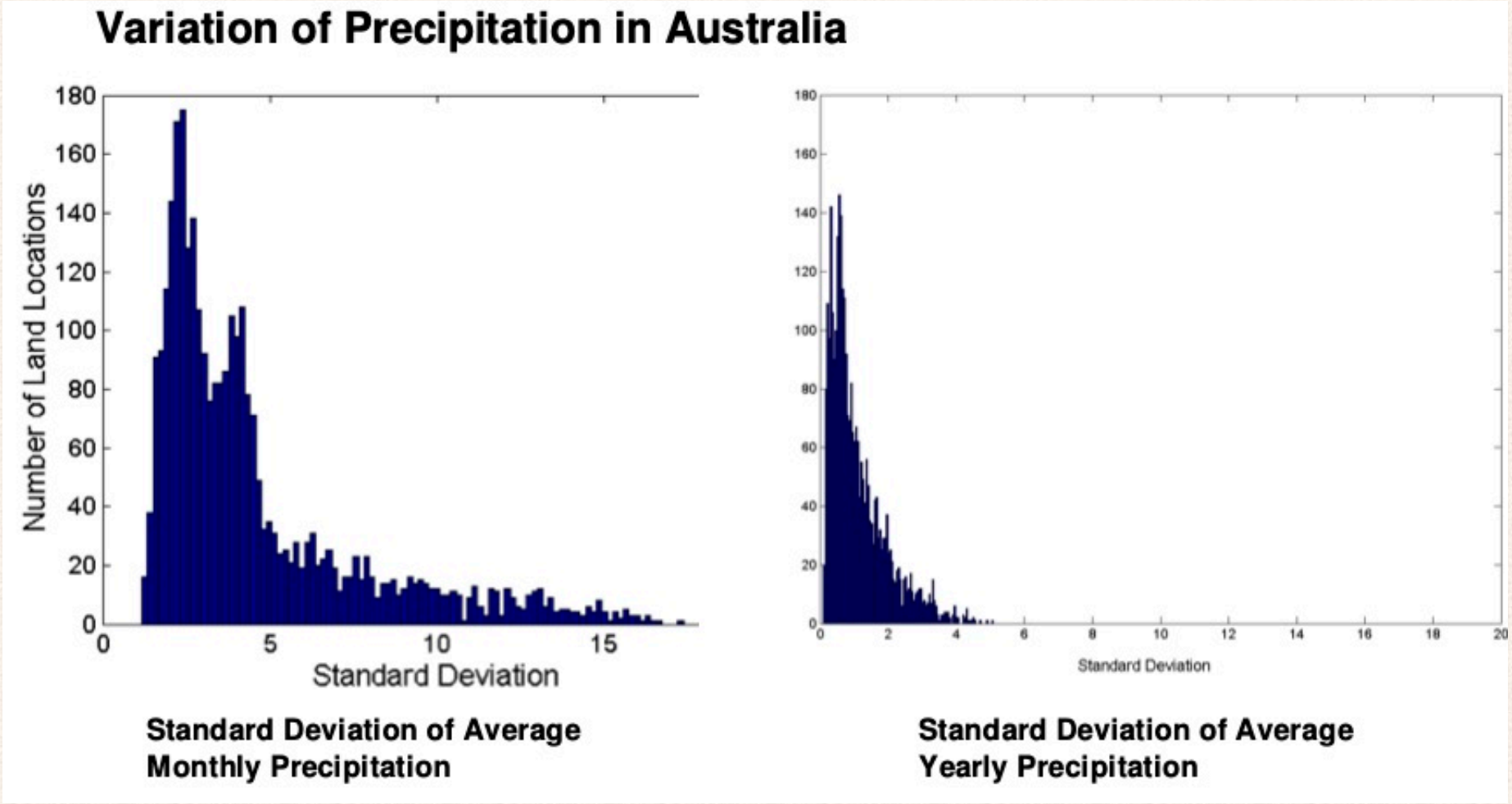
- Zmiana skali

  $\rightarrow$ Miasta zagregowane według regionów, stanów, krajów itp.

- Bardziej „stabilne" dane

  $\rightarrow$ Dane zagregowane mają zwykle mniejszą zmienność

Agregacja.



Standard Deviation of Average Monthly Precipitation

Standard Deviation of Average Yearly Precipitation

Próbkowanie.

• Próbkowanie jest główną techniką selekcji danych.

- Jest często używane zarówno do wstępnego badania danych, jak i do końcowej analizy danych.

• Statystycy próbkują, ponieważ uzyskanie całego zestawu danych jest zbyt kosztowne lub czasochłonne.

• Próbkowanie jest używane w eksploracji danych, ponieważ przetwarzanie całego zestawu danych jest zbyt kosztowne lub czasochłonne.

Próbkowanie.

• Kluczową zasadą skutecznego pobierania próbek jest:

- użycie próbki będzie działało prawie tak samo dobrze, jak wykorzystanie całych zbiorów danych, jeśli próbka jest reprezentatywna

- próbka jest reprezentatywna, jeśli ma w przybliżeniu tę samą właściwość (będącą przedmiotem zainteresowania), jak oryginalny zestaw danych

Rodzaje próbkowania.

• Proste próbkowanie losowe

- Istnieje równe prawdopodobieństwo wybrania określonej pozycji

• Pobieranie próbek bez zwracania

- Po wybraniu każdego elementu jest on usuwany z populacji

• Pobieranie próbek ze zwracaniem

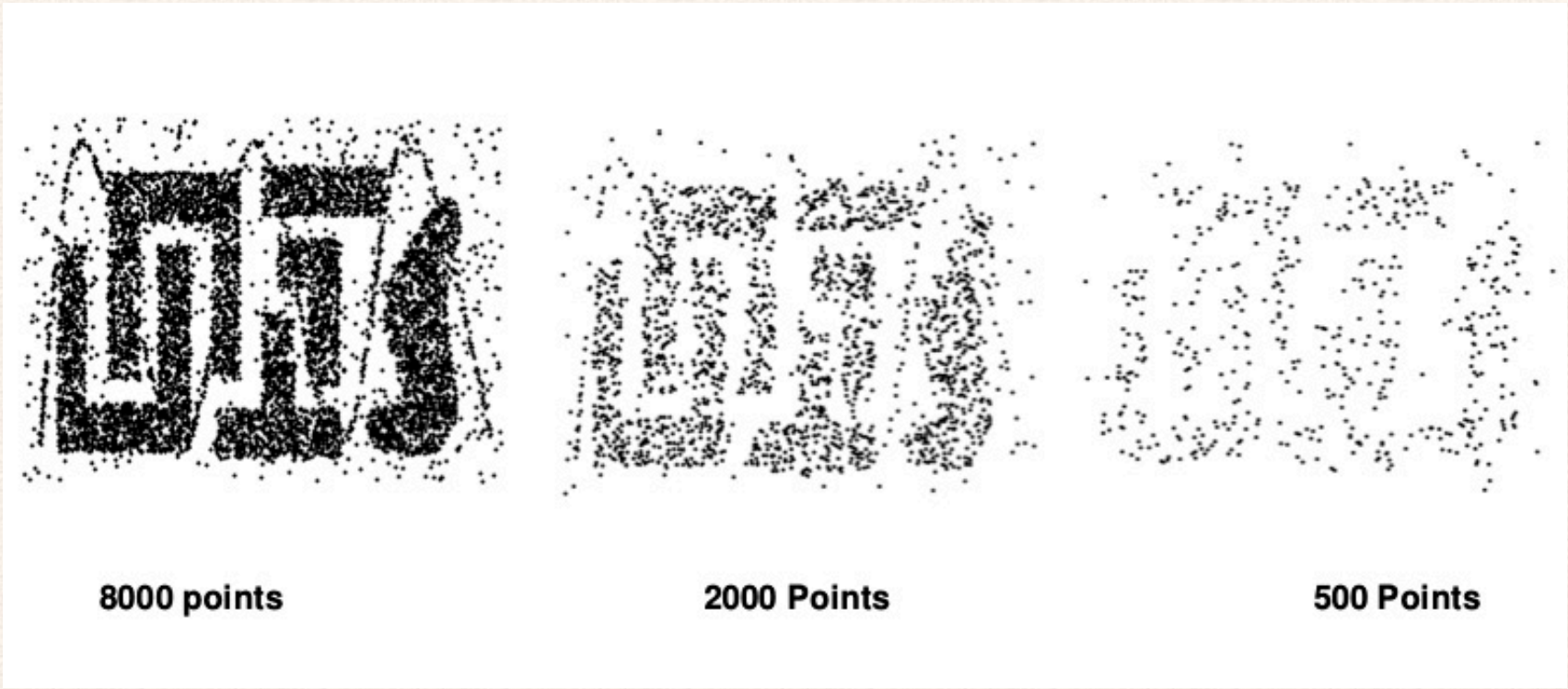
- Obiekty nie są usuwane z populacji w takiej postaci, w jakiej są pobrane do próbki.

• W przypadku pobierania próbek ze zwracaniem ten sam przedmiot może zostać podniesiony więcej niż raz

• Próbkowanie warstwowe

- Podziel dane na kilka partycji; następnie losuj próbki z każdej partycji

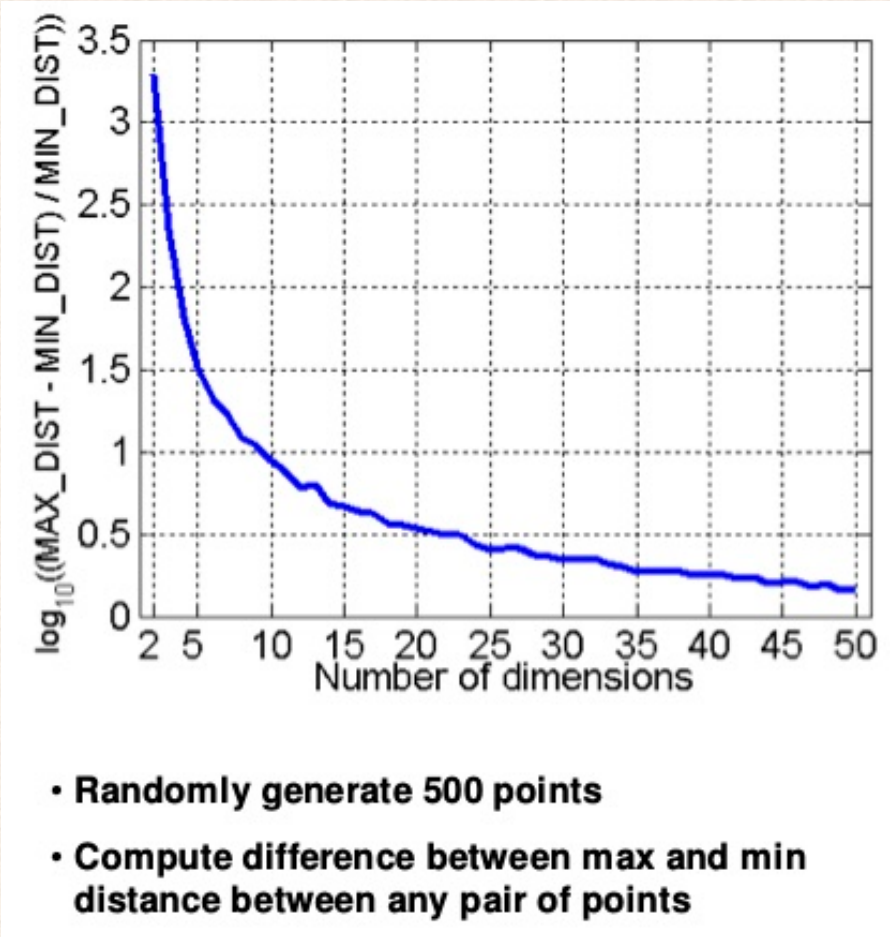
Wielkość próbki.



Przekleństwo wymiarowości.

• Wraz ze wzrostem wymiarowości dane w zajmowanej przez nie przestrzeni stają się coraz rzadsze

• Definicje gęstości i odległości między punktami, które są krytyczne dla grupowania i wykrywania wartoąci odstających, tracą na znaczeniu



Redukcja wymiarowości.

• Cel, powód:

- Unikaj przekleństwa wymiarowości

- Zmniejsz ilość czasu i pamięci wymaganą przez algorytmy eksploracji danych

- Pozwól na łatwiejszą wizualizację danych

- Może pomóc w wyeliminowaniu nieistotnych funkcji lub ograniczeniu szumu

• Techniki

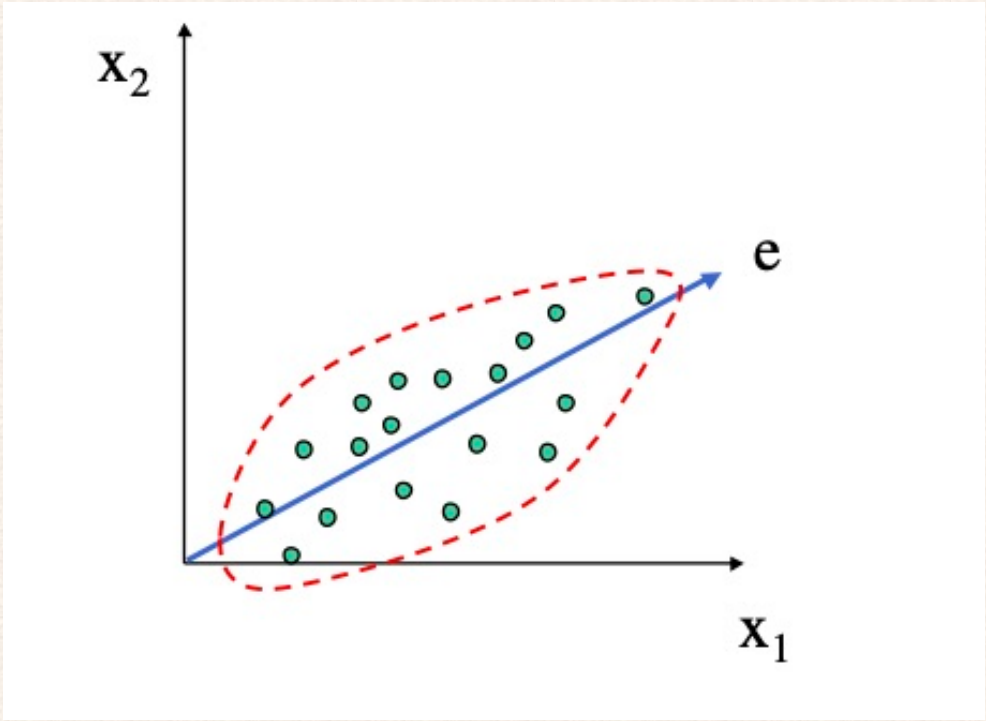
- Analiza podstawowych komponentów

- Rozkład według wartości osobliwych

- Inne: techniki nadzorowane i nieliniowe

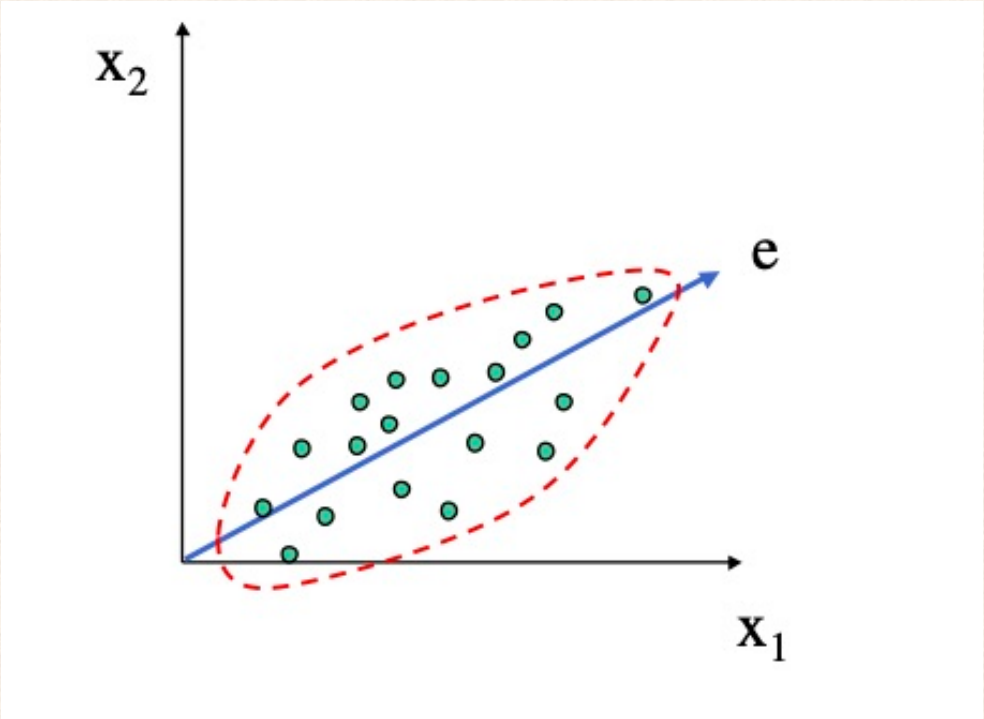
Redukcja wymiarowości: PCA.

- Celem jest znalezienie prognozy, która uchwyci największą zmienność danych



Redukcja wymiarowości: PCA.

- Znajdź wektory własne macierzy kowariancji
- Wektory własne definiują nową przestrzeń



Zbiory rozmyte i logika rozmyta.

Zbiór rozmyty: zbiór, gdzie funkcja przynależności zbioru jest funkcją o wartościach rzeczywistych z zakresu [0,1].

- $f(x)$: Prawdopodobieństwo x jest w F.

- $1 - f(x)$: Prawdopodobieństwo x nie znajduje się w F.

Przykład

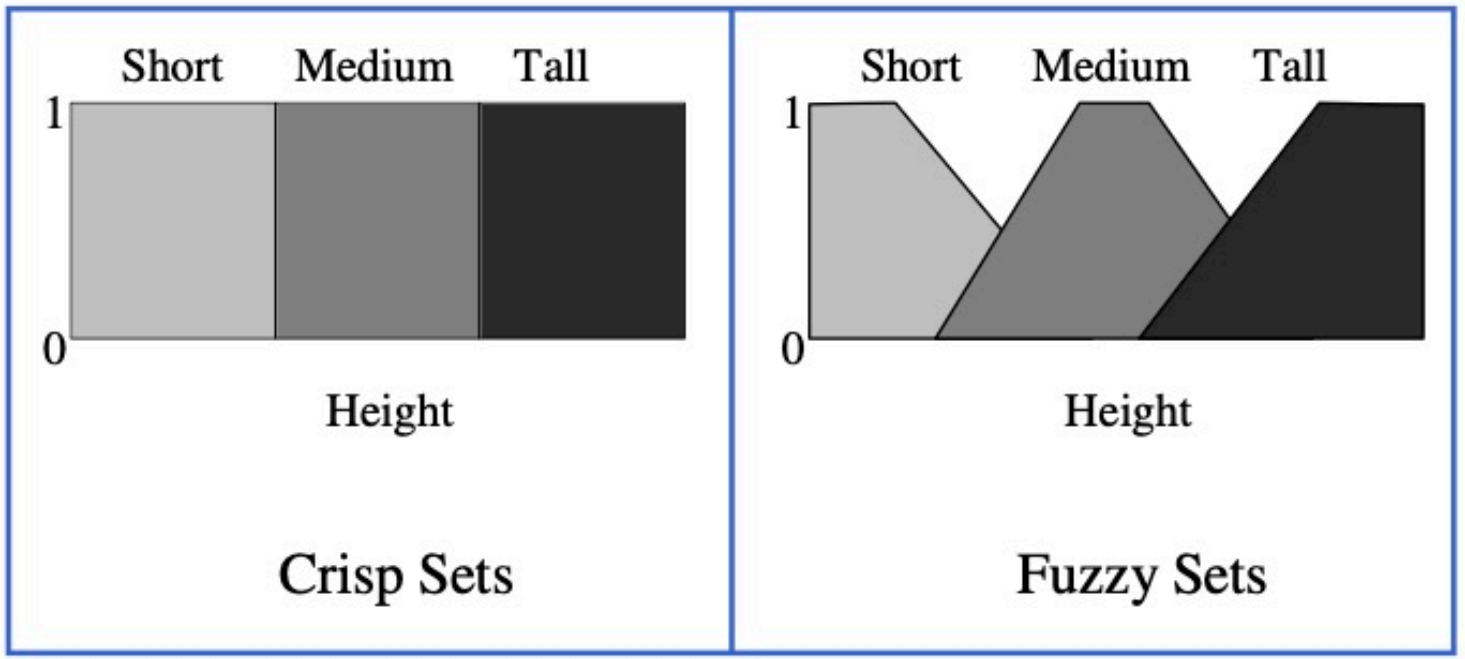
- $T = \{x \,|\, x \,\text{jest osobą i jest wysoki}\}$

- Niech $f(x)$ będzie prawdopodobieństwem tego, że x jest wysoki

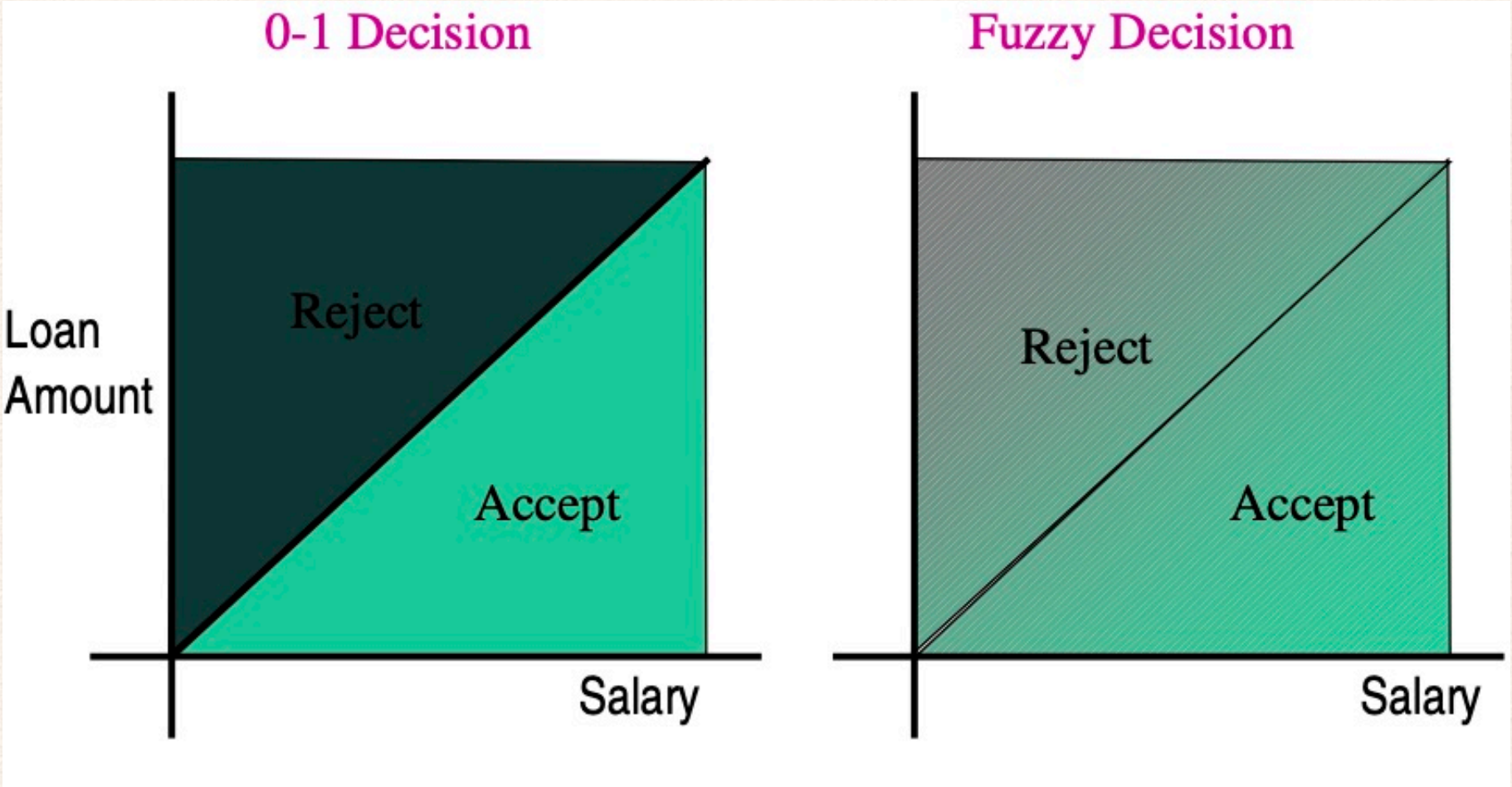
- Tutaj f będzie funkcją przynależności.

Eksploracja danych: Prognozy i klasyfikacja są często rozmyte.

Zbiory rozmyte.



Klasyfikacja i przewidywanie często są rozmyte.



Pobieranie informacji.

Pobieranie informacji (IR): pobieranie żądanych informacji z danych tekstowych

- Bibliotekoznawstwo

- Biblioteki cyfrowe

- Wyszukiwarki internetowe

- Tradycyjnie było oparte na słowach kluczowych

- Przykładowe zapytanie:

• Znajdź wszystkie dokumenty dotyczące „eksploracji danych".

Eksploracja danych: miary podobieństwa; eksploruj tekst lub dane internetowe

Pobieranie informacji c.d.

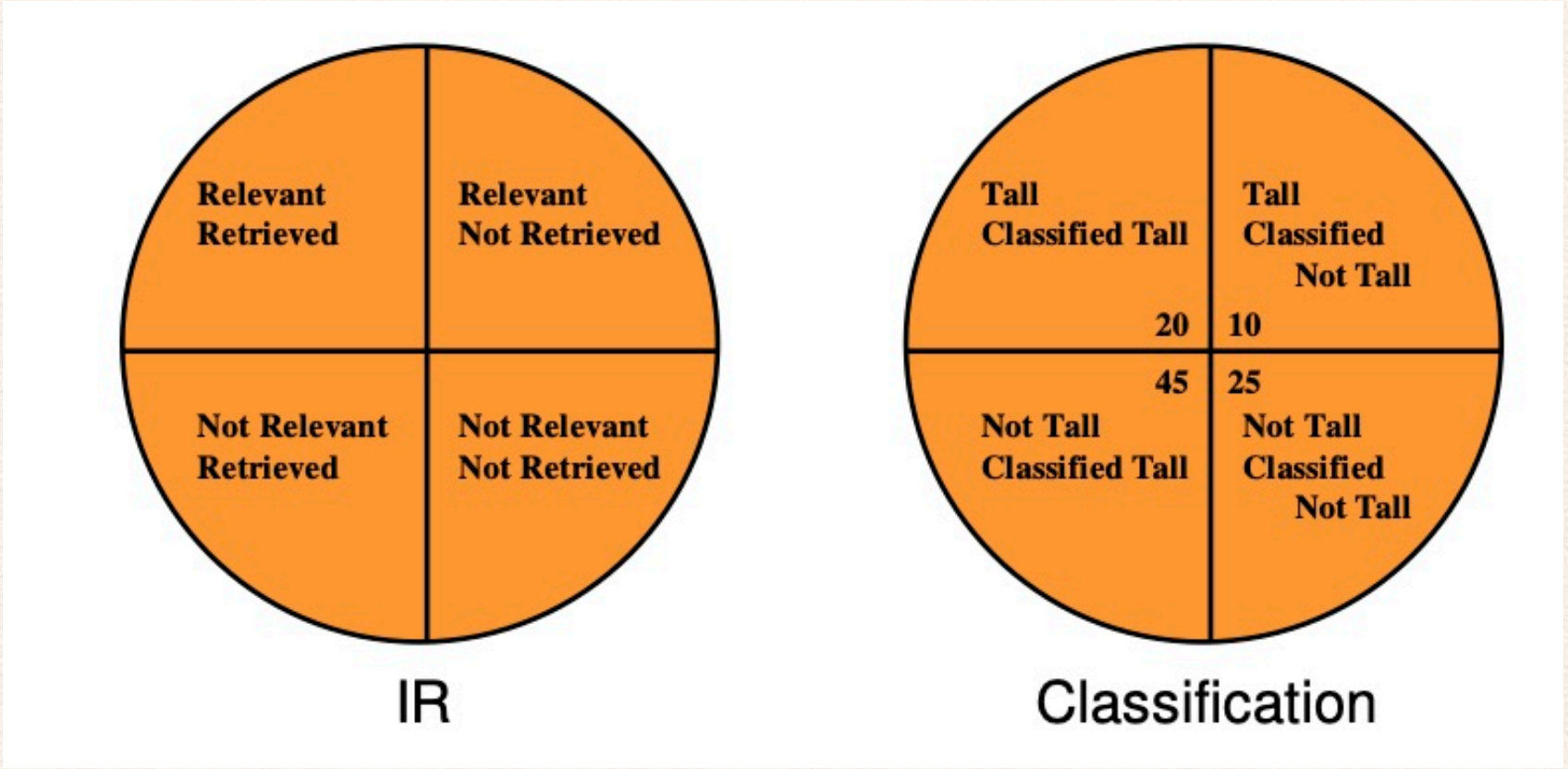
Podobieństwo: miara tego, jak blisko jest zapytanie do dokumentu.

- Pobierane są dokumenty, które są „wystarczająco blisko".
- Metryka:

$$\text{Dokładność} = \frac{|\text{istotne i pobrane}|}{|\text{pobrane}|}$$

$$\text{Przywołanie} = \frac{|\text{istotne i pobrane}|}{|\text{istotne}|}$$

Miary wyników zapytań IR i klasyfikacja.



Uczenie maszynowe.

• Uczenie maszynowe (ML): obszar sztucznej inteligencji, który bada, jak opracować algorytmy, które mogą się uczyć.

• Techniki z ML są często używane w klasyfikacji i prognozowaniu.

• Uczenie nadzorowane: uczy się na przykładzie.

• Uczenie się bez nadzoru: uczy się bez znajomości poprawności odpowiedzi.

• Uczenie maszynowe często dotyczy małych lub statycznych zbiorów danych.

Eksploracja danych: Wykorzystuje wiele technik uczenia maszynowego.

Statystyka.

• Zwykle tworzy proste modele opisowe.

• Wnioskowanie statystyczne: uogólnianie modelu utworzonego z próbki danych na cały zbiór danych.

• Analiza danych rozpoznawczych:

- Dane mogą w rzeczywistoąci wpływać na tworzenie modelu.

- W przeciwieństwie do tradycyjnego statystycznego punktu widzenia.

• Eksploracja danych skierowana do użytkowników biznesowych.

Eksploracja danych: Wiele metod eksploracji danych opiera się na technikach statystycznych.

Estymacja punktowa.

Oszacowanie punktowe: oszacowanie parametru populacji.

• Można to zrobić obliczając parametr dla próby.

• Może być używane do przewidywania wartości brakujących danych.

Przykład:

- R zawiera 100 pracowników

- 99 ma informacje o wynagrodzeniu

- Średnia pensja to 50 000 dolarów

- Użyj 50 000 $ jako wartości pensji pozostałych pracowników.

Czy to dobry pomysł?

Błąd estymacji.

Odchylenie (balans): różnica między wartością oczekiwaną a rzeczywistą wartością:

$$B = E(\hat{\Theta}) - \Theta$$

Błąd średniokwadratowy (MSE): oczekiwana wartość kwadratu różnicy między oszacowaną a rzeczywistą wartością:

$$\mathrm{MSE}\,(\hat{\Theta}) = E(\hat{\Theta} - \Theta)^2$$

Oszacowanie metodą „ostrza noża”.

• Oszacowanie metodą ostrza noża: oszacowanie parametru uzyskuje się poprzez pominięcie jednej wartości ze zbioru obserwowanych wartości.

• Np .: oszacowanie średniej dla $X = \{x_1, ..., x_n\}$:

$$\hat{\theta}(i) = \frac{1}{n}\left(\sum_{j=1}^{i-1} x_j + \sum_{j=i+1}^{n} x_j\right)$$

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}(i)$$

Oszacowanie największego prawdopodobieństwa (MLE).

• Uzyskaj oszacowania parametrów, które maksymalizują prawdopodobieństwo, że przykładowe dane pojawią się dla określonego modelu.

• Wspólne prawdopodobieństwo obserwacji danych z próby poprzez pomnożenie indywidualnych prawdopodobieństw. Funkcja prawdopodobieństwa:

$$L(\theta|\, x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\, \theta)$$

• Maksymalizuj L.

Przykład MLE

- Pięciokrotny rzut monetą: $\{R, R, R, R, O\}$
- Zakładając, że dla idealnej monety prawdopodobieństwo wyrzucenia orła i reszki jest takie samo, prawdopodobieństwo wyrzucenia tego, co powyżej, jest równe:

$$L(p|\,1,1,1,1,0) = \prod_{i=1}^{5} 0.5 = 0.03$$

- Gdyby prawdopodobieństwo wyrzucenia reszki wynosiło 0.8, to:

$$L(p|\,1,1,1,1,0) = 0.8 \cdot 0.8 \cdot 0.8 \cdot 0.8 \cdot 0.2 = 0.08$$