Analiza statystyczna danych

dr Marcin Woźniak

woz@amu.edu.pl

konsultacje po wcześniejszym umówieniu!

Literatura

- *Sobczyk M. [red], 2000. Statystyka. Podstawy teoretyczne, przykłady, zadania, Wyd. UMCS Lublin.*
- *Rutkowski T., 1999. Statystyka. Zagadnienia wybrane. Wyd. WSB Poznań.*
- *Statystyczna analiza danych z wykorzystaniem programu R*, red. M. Walesiak, E. Gatnar, PWN 2012.
- B. Everitt, T. Hothorn, *A Handbook of Statistical Analyses Using R*, Taylor & Francis 2010.
- Robert Nisbet, Gary Miner and Ken Yale, *Handbook of Statistical Analysis and Data Mining Applications*, Elsevier, 2018.
- D. Griffith, C. Amrhein, J. Desloges, *Statistical Analysis for Geographers*, Prentice Hall, 1990.

Warunki zaliczenia

- Egzamin pisemny z **treści wykładów**
- Termin I: koniec stycznia 2025
- Termin poprawkowy: luty 2025 (druga połowa)
- Warunkiem podejścia do egzaminu jest zaliczenie ćwiczeń

Plan wykładów

- Czym jest statystyka i kontekst historyczny
- Dane statystyczne
- Metoda statystyczna
- Podstawowe miary opisu danych (miary rozrzutu i środka, miary zależności i bliskości, współczynnik Giniego i krzywa Lorenza)
- Projektowanie badań społecznych (dobór próby, pomiar)

Plan wykładów

- Dalsza eksploracja danych (podstawowe algorytmy klasyfikacji danych: k-means, DB-scan, metody hierarchiczne i niehierarchiczne)
- AI i uczenie maszynowe
- Predykcja i wnioskowanie statystyczne (modele regresji liniowej pojedyńczej i wielorakiej, regresja logistyczna, modele szeregów czasowych)

Czym jest statystyka?

https://www.youtube.com/watch?v=4DruxASC1kM

Czym jest statystyka?

Czym jest statystyka?

"*Statistics is the branch of scientific method which deals with the data obtained by counting or measuring the properties of populations of natural phenomena. In this definition 'natural phenomena' includes all the happenings of the external world, whether human or not*."

Professor Maurice Kendall, 1943

"*Statistics is: the fun of finding patterns in data; the pleasure of making discoveries; the import of deep philosophical questions; the power to shed light on important decisions, and the ability to guide decisions..... in business, science, government, medicine, industry…*"

Professor David Hand, 2009

Czym jest statystyka?

- Dziedzina matematyki zajmująca się zbieraniem, analizą, wizualizacją i interpretacją danych
- Dwa (główne) działy: statystyka opisowa i wnioskowanie statystyczne
- Dostarcza metod do analizy i interpretacji otaczającej nas rzeczywistości

Czym jest statystyka?

- Dziedzina matematyki zajmująca się zbieraniem, analizą, wizualizacją i interpretacją danych
- Dwa (główne) działy: statystyka opisowa i wnioskowanie statystyczne
- Dostarcza metod do analizy i interpretacji otaczającej nas rzeczywistości
- Termin “statystyka” może też odnosić się do konkretnych wskaźników
- Rozwój wielu nowych technik analitycznych opartych na statystyce

Statystyka – podejście analityczne

- Excell
- SPSS
- Statistica
- Stata
- MATLAB
- GAUSS
- E Views
- GRETL
- **R**

Środowisko R

R (programming language)

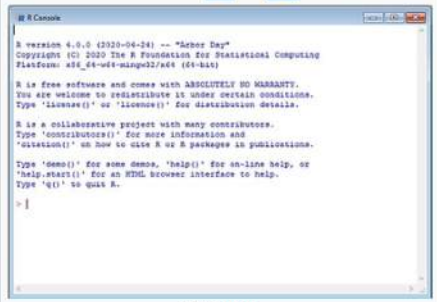
From Wikipedia, the free encyclopedia

R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. Created by statisticians Ross Ihaka and Robert Gentleman, R is used among data miners, bioinformaticians and statisticians for data analysis and developing statistical software.[7] Users have created packages to augment the functions of the R language.

According to user surveys and studies of scholarly literature databases, R is one of the most commonly used programming languages used in data mining.[8] As of December 2022, R ranks 11th in the TIOBE index, a measure of programming language popularity, in which the language peaked in 8th place in August 2020.[9][10]

The official R software environment is an open-source free software environment within the GNU package, available under the GNU General Public License. It is written primarily in C, Fortran, and R itself (partially self-hosting). Precompiled executables are provided for various operating systems. R has a command line interface.[11] Multiple third-party graphical user interfaces are also available, such as RStudio, an integrated development environment, and Jupyter, a notebook interface.

R



R terminal

Paradigms	Multi-paradigm: procedural, object-oriented, functional, reflective, imperative, array[1]
Designed by	Ross Ihaka and Robert Gentleman
Developer	R Core Team
First appeared	August 1993; 29 years ago
Stable release	4.2.2[2] / 31 October 2022; 2 months

Środowisko R

- Wywodzi się z języka S opracowanego na potrzeby analizy danych
- Własne czasopismo “R Journal”
- Używany na uniwersytetach (np. MIT, Cambridge)
- Używany w wielu firmach komercyjnych (np. Google, CitiBank, Novartis)

Zalety R

- Darmowy i dostępny na wszystkie platformy
- Otwarty kod źródłowy
- Szybki, elastyczny i stabilny
- Duże możliwości graficzne
- Duża liczba tutoriali i system pomocy
- Możliwość instalacji dodatkowych pakietów

Wady R

- Ograniczony interfejs graficzny
- Brak komercyjnego wsparcia technicznego
- Przechowuje dane w pamięci fizycznej
- Konieczność nauki języka (+)

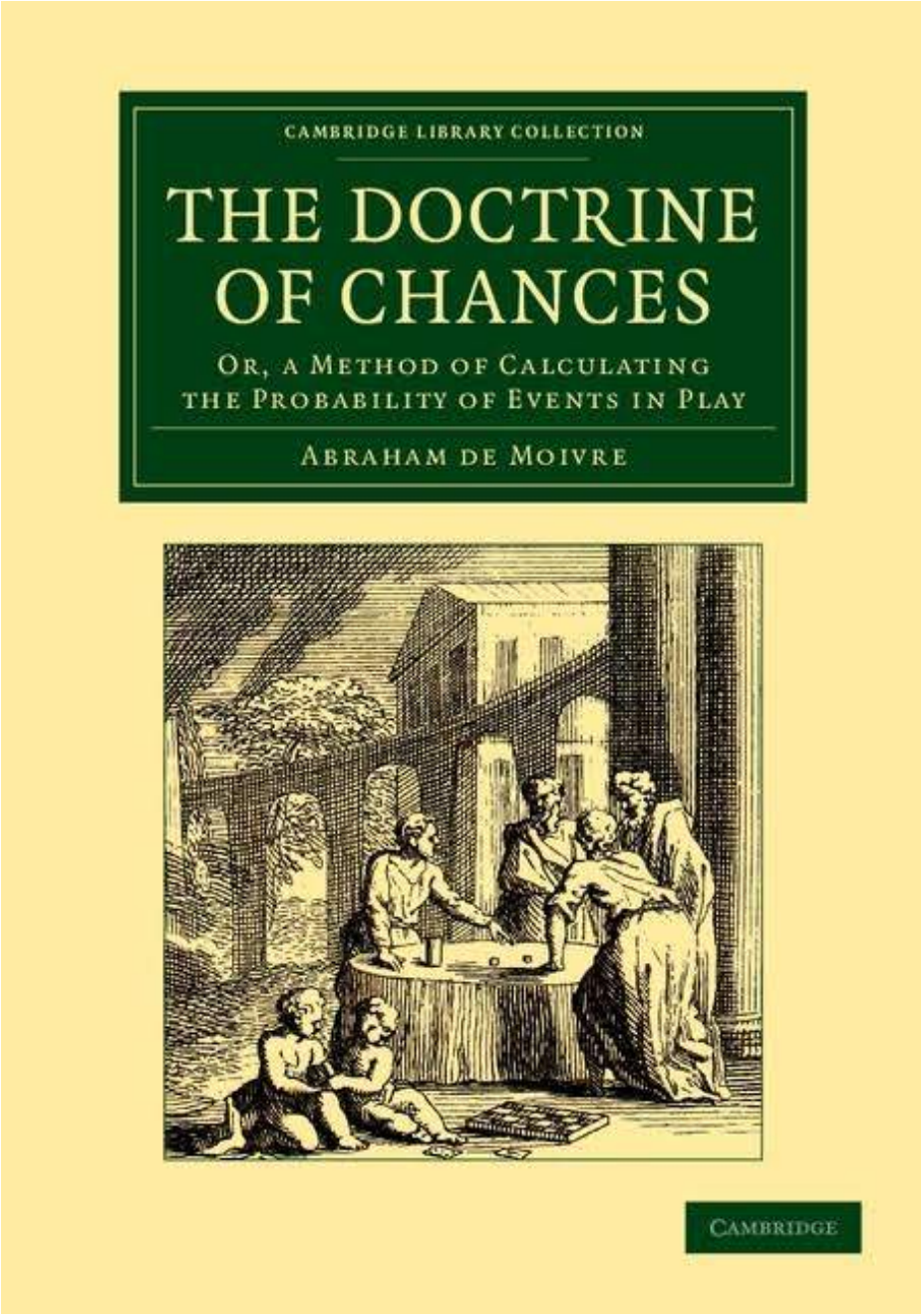
Statystyka - kontekst historyczny

Początki...

Przypuśćmy, że mamy stos 13 kart w jednym kolorze i inny stos 13 kart w innym kolorze.

Załóżmy, że każdy stos zawiera po jednym asie.

Jakie jest prawdopodobieństwo, że biorąc po jednej karcie z każdego stosu, wyciągnę dwa asy?



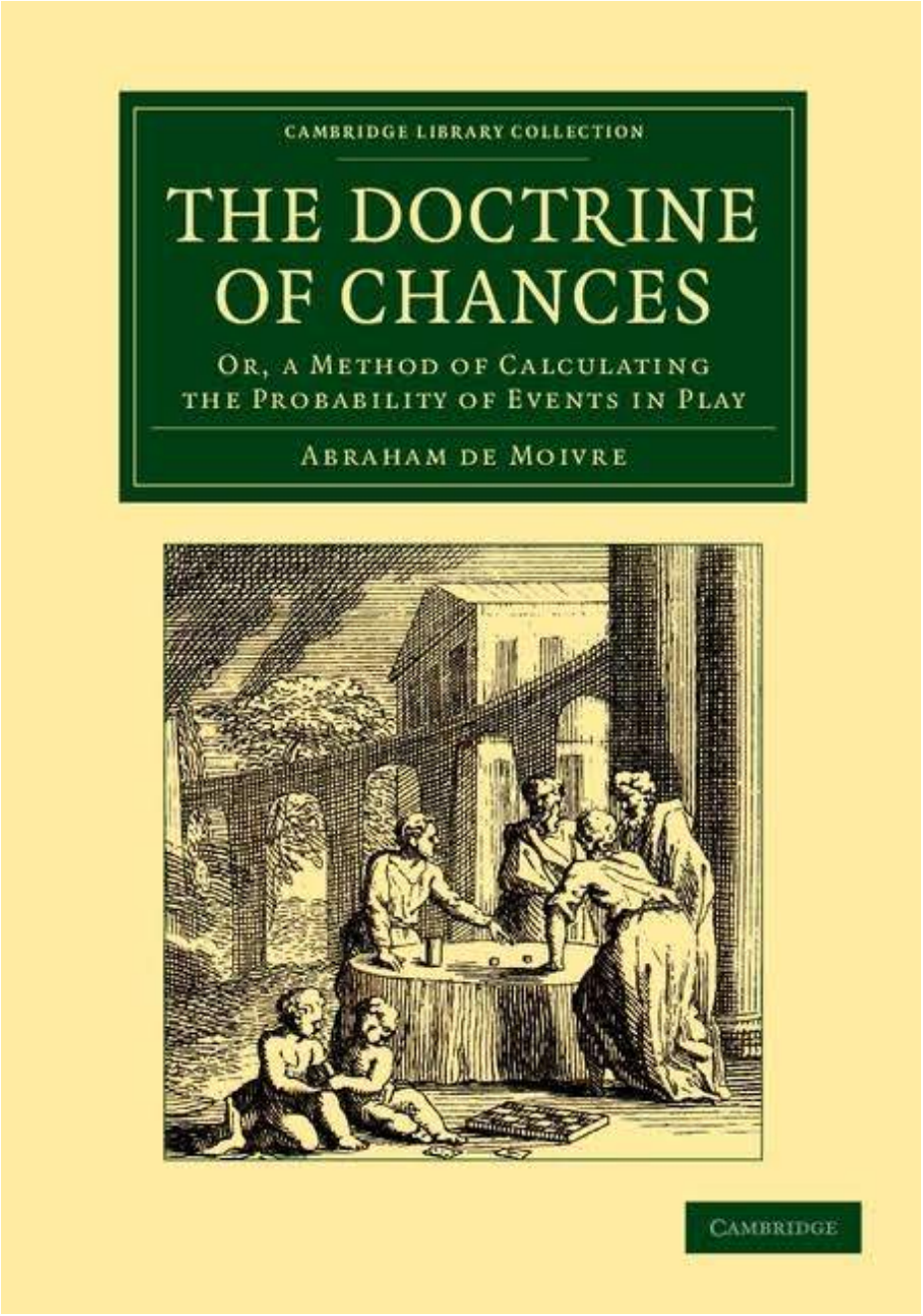
Początki...

Przypuśćmy, że mamy stos 13 kart w jednym kolorze i inny stos 13 kart w innym kolorze.

Załóżmy, że każdy stos zawiera po jednym asie.

Jakie jest prawdopodobieństwo, że biorąc po jednej karcie z każdego stosu, wyciągnę dwa asy?

$$1/13 * 1/13 = 1/169$$



Statystyka - kontekst historyczny

- Relatywnie nowa dyscyplina nauki
- Większość dorobku powstała w ostatnich 150 latach
- Początkowo głównym czynnikiem rozwoju był hazard
- Dualność współczesnej analizy statystycznej: ujęcie klasyczne vs ujęcie Bayesowskie

Dualizm analizy statystycznej

- Thomas Bayes (prawdopodobieństwo warunkowe) vs Ronald Fisher (model parametryczny)

Ujęcie Bayesa

- Prawdopodobieństwo wystąpienia zdarzenia w **przyszłości** jest równe prawdopodobieństwu jego wystąpienia w **przeszłości** podzielonemu przez prawdopodobieństwo wystąpienia wszystkich konkurencyjnych zdarzeń.

Thomas Bayes
1702-1761

Ujęcie Bayesa

Thomas Bayes
1702-1761

- Prawdopodobieństwo wystąpienia zdarzenia w **przyszłości** jest równe prawdopodobieństwu jego wystąpienia w **przeszłości** podzielonemu przez prawdopodobieństwo wystąpienia wszystkich konkurencyjnych zdarzeń.
- Analiza przebiega w oparciu o pojęcie **prawdopodobieństwa warunkowego:** prawdopodobieństwa wystąpienia zdarzenia przy założeniu, że inne zdarzenie już wystąpiło.
- Rozpoczyna się od kwantyfikacji istniejącego stanu wiedzy badacza, przekonań i założeń na temat przeszłych zdarzeń.

Twierdzenie Bayesa

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Jakie jest prawdopodobieństwo zachorowania na raka w wieku 65 lat?

Załóżmy, że ogólna częstość występowania raka wynosi 2% (wcześniejsze prawdopodobieństwo zachorowania na raka – *apriori*).

Następnie, załóżmy, że prawdopodobieństwo bycia w wieku 65 lat wynosi 0,3% i że prawdopodobieństwo, że ktoś, u kogo zdiagnozowano raka ma 65 lat, wynosi 0,4%.

Mając te dane, możemy obliczyć prawdopodobieństwo zachorowania na raka jako 65-latek.

Twierdzenie Bayesa

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Jakie jest prawdopodobieństwo zachorowania na raka w wieku 65 lat?

Załóżmy, że ogólna częstość występowania raka wynosi 2% (wcześniejsze prawdopodobieństwo zachorowania na raka – *apriori*).

Następnie, załóżmy, że prawdopodobieństwo bycia w wieku 65 lat wynosi 0,3% i że prawdopodobieństwo, że ktoś, u kogo zdiagnozowano raka ma 65 lat, wynosi 0,4%.

Mając te dane, możemy obliczyć prawdopodobieństwo zachorowania na raka jako 65-latek.

0.004*0.02/0.003 = 0.026 (2.6%)

Model parametryczny

Ronald Fisher
1890-1962

- Eliminacja pojęcia prawdopodobieństwa *apriori* na rzecz prawdopodobieństwa *rzeczywistego*
- Szereg założeń, m.in.:

  - dane wpasowują się w jeden ze znanych rozkładów prawdopodobieństwa (najczęściej rozkład normalny (tzw. rozkład Gaussa))

  - niezależność danych objaśniających

Model parametryczny

Ronald Fisher
1890-1962

- Eliminacja pojęcia prawdopodobieństwa *apriori* na rzecz prawdopodobieństwa *rzeczywistego*
- Szereg założeń, m.in.:

  - dane wpasowują się w jeden ze znanych rozkładów prawdopodobieństwa (najczęściej rozkład normalny (tzw. rozkład Gaussa))

  - niezależność danych objaśniających

  - efekty oddziaływania jednej zmiennej na drugą mają charakter liniowy

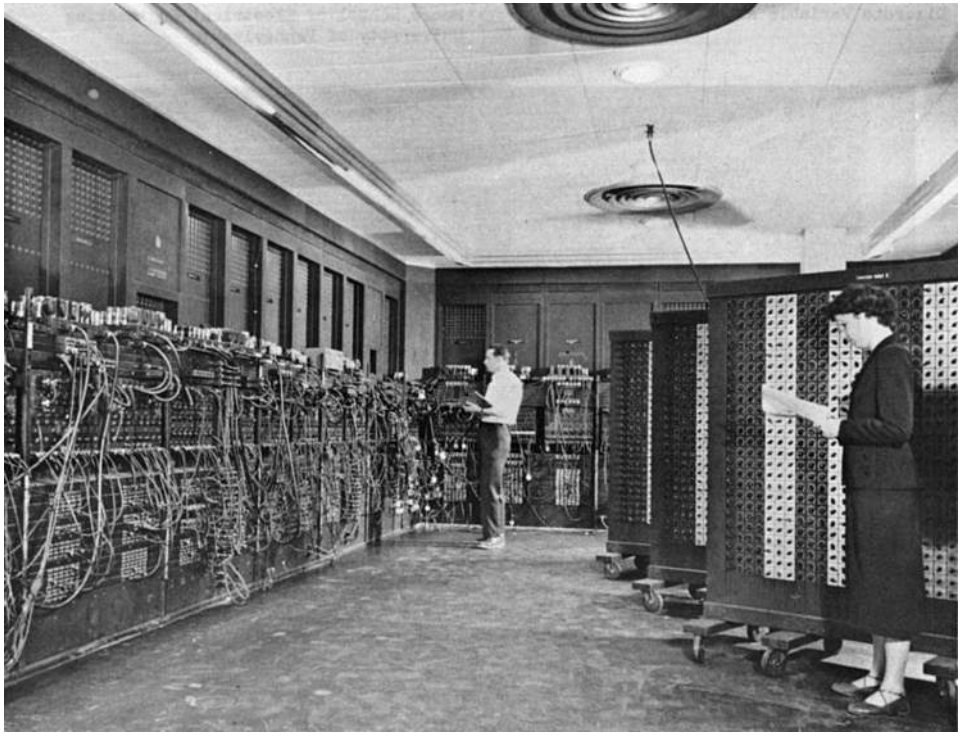
  - dane muszą być numeryczne i ciągłe
- Dylemat: warunki laboratoryjne a świat rzeczywisty? (Fisher czy Bayes?)

Dalszy rozwój analizy statystycznej

- I generacja (lata > 20 XX w.) - opis i wnioskowanie, analiza wariancji, metody skupione na modelach liniowych
- II generacja (lata > 80 XX w.) – metody skupione na modelach nieliniowych, zmiennych dyskretnych (np. modele logitowe i logistyczne, SEM)
- III generacja (lata > 2010 XXI w.) – metody oparte na uczeniu maszynowym (np. sztuczne sieci neuronowe (ANN), drzewa decyzyjne)

Rozwój analizy statystycznej w praktyce

ENIAC (1965)



- **12 m x 6 m**
- **30 ton**
- **150 000 W**
- **6 000 000 $**
- **0.05 MIPS**

INTEL Core i7 (2022)

- **38 cm x 26 cm**
- **2 kg**
- **60 W**
- **500 $**
- **161 173 MIPS**

WEŹ UDZIAŁ W BADANIU!

Wypełnij formularz kontaktowy

Celem badania jest identyfikacja sposobów doświadczania przestrzeni miejskiej przy wykorzystaniu technologii poszerzonej rzeczywistości (XR).

Udział w badaniu jest dobrowolny i anonimowy.

Czas trwania badania nie powinien przekroczyć 2h.

OTRZYMAJ GRATYFIKACJĘ:

250 ZŁ

BON DO ALLEGRO

Po więcej informacji zapraqszam do kontaktu:

macglo@amu.edu.pl

Dane statystyczne

Dane statystyczne

- Statystyka opiera się na wykorzystaniu danych
- Pozyskanie danych (zazwyczaj) wymaga pomiaru
- Różne rodzaje i typy danych
- Klasyczny podział na dane ilościowe i jakościowe (i badania ilościowe i jakościowe)

Rodzaje i typy danych

- **Dane ustrukturyzowane (np. dane tabelaryczne)**
- Dane nieustrukturyzowane (np. tekst, video, zdjęcia, audio)
- Dane częściowo ustrukturyzowane (np. logi, xml)

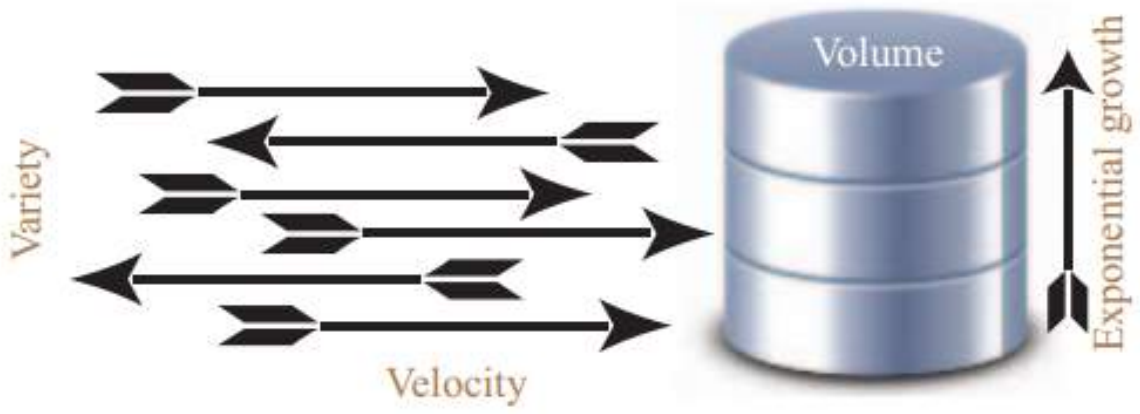
Rodzaje i typy danych

- **Dane ustrukturyzowane (np. dane tabelaryczne)**
- Dane nieustrukturyzowane (np. tekst, video, zdjęcia, audio)
- Dane częściowo ustrukturyzowane (np. logi, xml)
- Dane generowane w czasie rzeczywistym (np. streaming, transakcje bankowe)
- Data at Rest (dane w bibliotekach cyfrowych, np. dane sprzedażowe, dane z urządzeń mobilnych)
- Metadane (dane o danych)

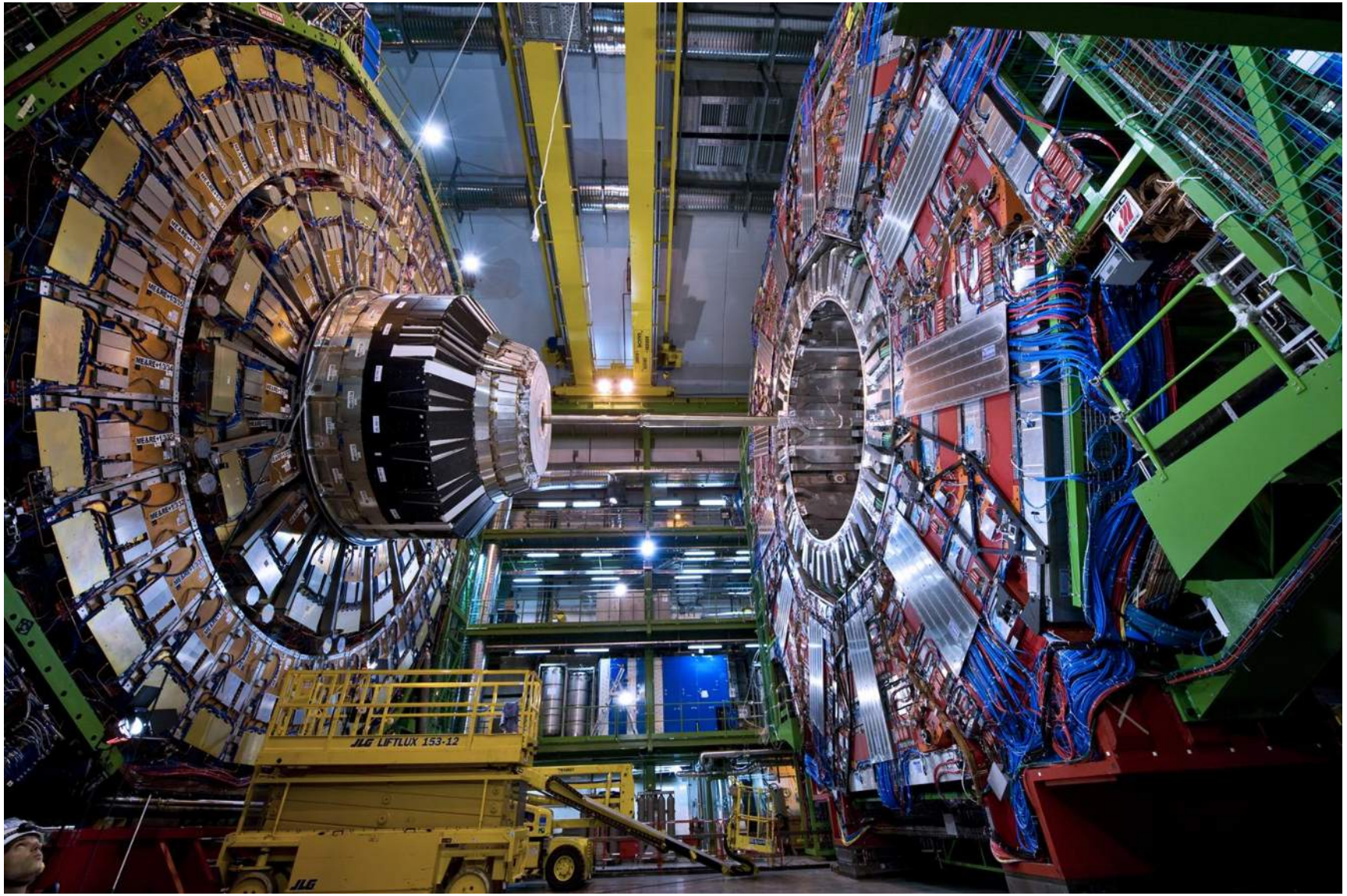
Big Data

Wg Doug'a Laney (2001) – 3V:

- **Volume** - duża (relatywnie) ilość n-wymiarowych danych
- **Velocity** - wysoka zmienność, dynamika
- **Variety** - różnorodność źródeł, typów, formatów



Volume



9 Petabajtów/ miesiąc

Volume

10 TB / silnik / 30 minut

Etapy pracy z danymi

1) Identyfikacja źródeł danych. Dane mogą być zebrane w jednej bazie albo rozproszone.

2) Pozyskanie danych (interfejs API,
tablice: .csv, .xlsx, .txt
GIS: geopackage, shapefile, sqlite).

Etapy pracy z danymi

1) Identyfikacja źródeł danych. Dane mogą być zebrane w jednej bazie albo rozproszone.

2) Pozyskanie danych (interfejs API,
tablice: .csv, .xlsx, .txt
GIS: geopackage, shapefile, sqlite).

3) Integracja (harmonizacja) danych: dane mogą być w różnych formatach lub występować na różnych poziomach agregacji lub są wyrażone w różnych jednostkach.

Skale pomiarowe

Skale pomiarowe

- **Nominalna:** przypisanie nazw do klas (np. czerwony, żółty, zielony). Skala jest nominalna, jeśli rozróżnia grupy. *Np. kolor może być użytecznym atrybutem, ale sam nie ma znaczenia numerycznego*.
- Operacje arytmetyczne nie mają sensu
- Inne przykłady?

Skale pomiarowe

- **Porządkowa:** kategorie danych, które można uporządkować (np. Stopień trudności)
- Zbiór pozornie uporządkowanych kategorii nie tworzy skali porządkowej. Cecha jest porządkowa, jeśli implikuje ranking (stopniowanie).
- Operacje arytmetyczne nie mają sensu.
- Przykłady?

Skale pomiarowe

- **Interwałowa:** dane liczbowe, które wykazują porządek, a także możliwość zmierzenia interwału (odległości) pomiędzy dowolną parą obiektów na skali (np data urodzenia).
- Dane są typu interwałowego, jeśli różnice mają sens.
- Niektóre operacje arytmetyczne są dozwolone
- Inne przykłady?

Skale pomiarowe

- **Ilorazowa:** interwałowa + naturalne pochodzenie (np. stopa bezrobocia, poziom płac, liczba wakatów).
- Pozwala określić rozmiar różnic pomiędzy analizowanymi cechami
- Stosunki między dwiema jej wartościami mają interpretację w świecie rzeczywistym.
- Nie nakłada ograniczeń w stosowaniu operacji matematycznych i metod statystycznych.
- Przykłady?

Skale pomiarowe - podsumowanie

Nazwa uniwersytetu (cecha nominalna)	**Poziom edukacji (cecha porządkowa)**	Data rozpoczęcia sesji letniej (cecha interwałowa)	Ilość studentów (cecha ilorazowa)
Szkoła Handlowa Główna	**Wysoki**	14. czerwca 2019	32300
Uniwersytet Ekonomiczny	**Bardzo wysoki**	21. Czerwca 2019	12760
Politechnika Techniczna	**Bardzo wysoki**	13. Czerwca 2019	18710
SGWG	**Dostateczny**	28. Czerwca 2019	21290

Statystyki opisowe

Statystyki opisowe - narzędzia

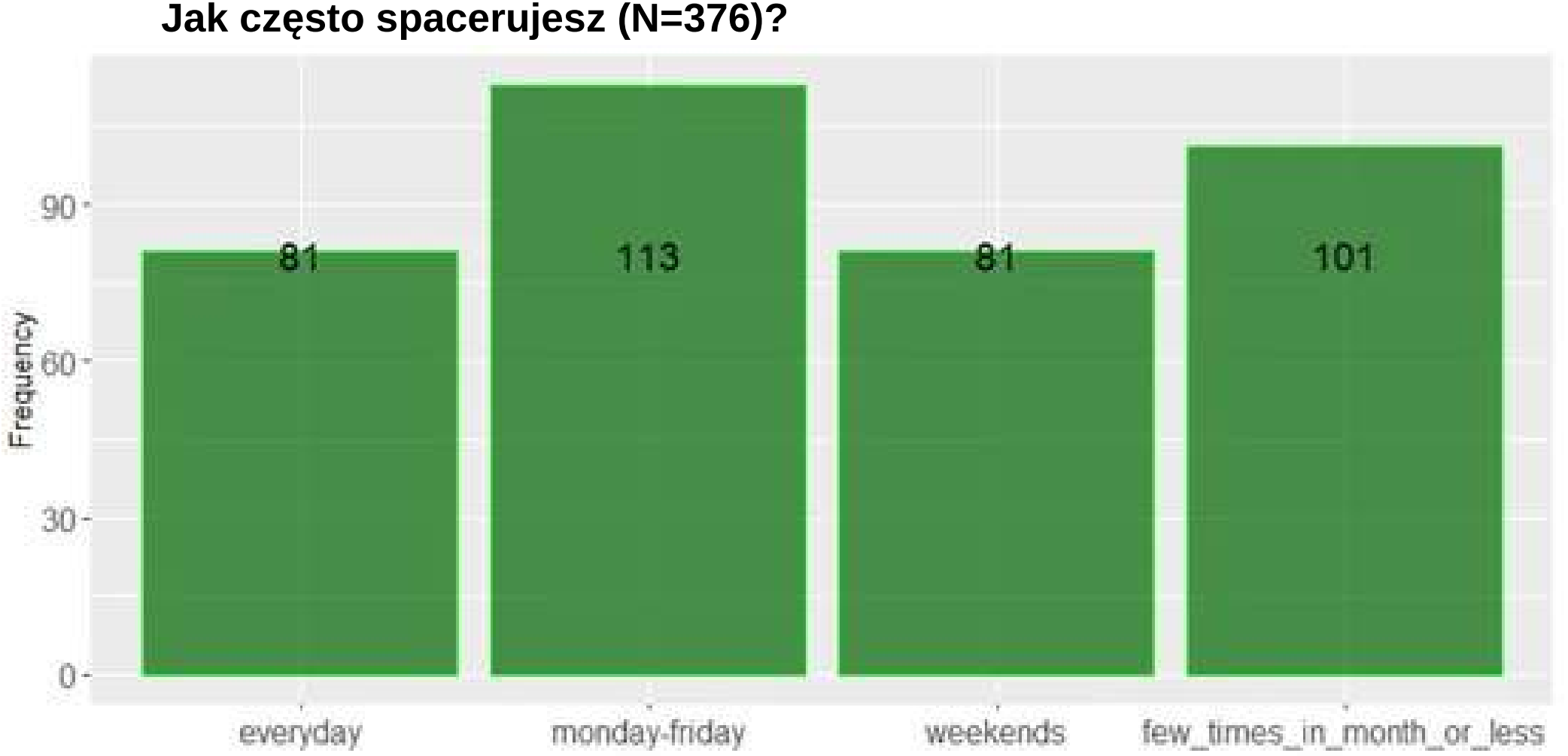
- Ilustracja graficzna (np. histogram, wykres pudełkowy, słupkowy, liniowy, kartogramy, kartodiagramy)
- Tabele częstości + tabele krzyżowe
- Miary środka
- Miary rozproszenia
- Miary koncentracji i asymetrii
- Miary zależności i bliskości

Podstawowe narzędzia graficznej prezentacji danych

Wykres słupkowy

- to wykres, który przedstawia dane za pomocą prostokątnych słupków o wysokości lub długości proporcjonalnej do wartości, które reprezentują
- pokazuje porównania pomiędzy kategoriami danych. Jedna oś wykresu pokazuje konkretne kategorie, a druga oś przedstawia mierzoną wartość.

Wykres słupkowy



Histogram

- Jeśli pomiary mają wartości liczbowe w skali interwałowej lub ilorazowej, mogą one być pogrupowane w klasy (przedziały)
- Liczby w każdej klasie można zwizualizować jako wykres słupkowy, w którym ważna jest kolejność na osi x
- Wykres słupkowy tego typu nazywany jest histogramem

Histogram

Przykład: mediana cen nieruchomości w stanie Teksas

Przedziały klasowe

- Jeśli dane są pomiarami zmiennej ciągłej (numerycznej), wtedy standardową procedurą w tworzeniu histogramu jest utworzenie przedziałów klasowych i policzenie częstotliwości występujących w każdym przedziale obserwacji.

Przedziały klasowe

- Jeśli dane są pomiarami zmiennej ciągłej (numerycznej), wtedy standardową procedurą w tworzeniu histogramu jest utworzenie przedziałów klasowych i policzenie częstotliwości występujących w każdym przedziale obserwacji.
- Wartości, determinujące przedziały są określane jako punkty odcięcia.
- Kluczowe jest ustalenie liczby klas/przedziałów.

Przedziały klasowe

$$k = \frac{max - min}{h} \qquad lub \qquad h = \frac{max - min}{k}$$

Gdzie k to liczba klas, a h szerokość przedziału

$$k = 3.5 * sd / n^{(1/3)} \quad (formuła\ Scotta, 1979)$$

Zgodnie z formułą Scotta, ile przedziałów utworzymy dla zbioru liczącego 1000 elementów, w którym odchylenie standardowe wynosi 25?

Wykres pudełkowy (boxplot)

- Wykres pudełkowy (przedstawiony pionowo) tworzy się odkładając na osi y wartości kluczowych parametrów rozkładu

Wykres pudełkowy (boxplot)

- Wykres pudełkowy (przedstawiony pionowo) tworzy się odkładając na osi y wartości kluczowych parametrów rozkładu
- To sposób wyświetlania zbioru danych oparty na podsumowaniu pięciu statystyk: *minimum, maksimum, mediana oraz pierwszy i trzeci kwartyl.*
- Składa się z prostokąta i przylegających do niego “wąsów”

Wykres pudełkowy (boxplot)

Przykład – liczba cylindrów a zasięg auta

Wykres liniowy

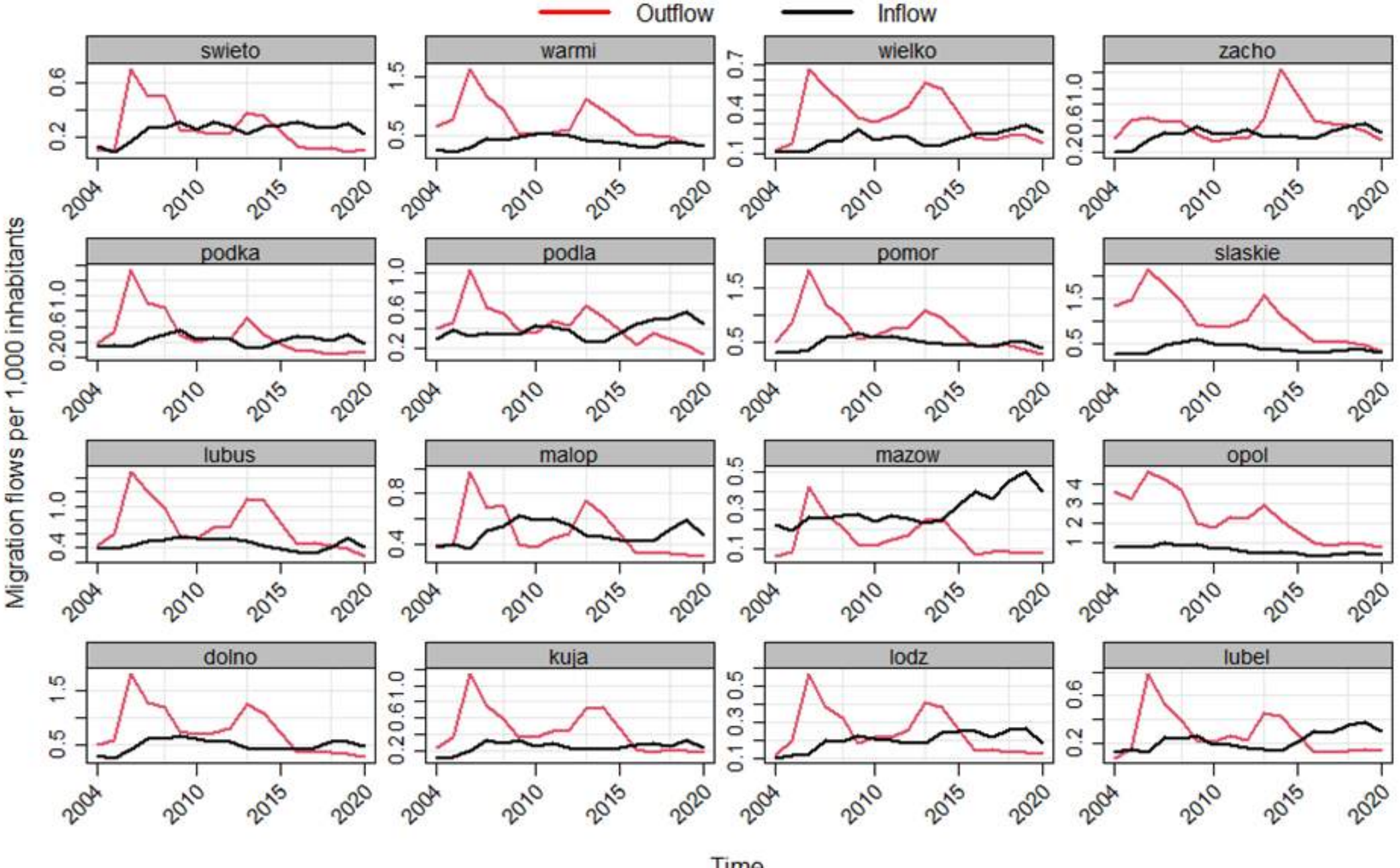
- To rodzaj wykresu, który wyświetla informacje w postaci serii punktów danych zwanych "markerami"
- Punkty te są połączone odcinakmi linii prostej
- Jest często używany do wizualizacji trendu w danych w odstępach czasu – tzw. szeregu czasowego.

Wykres liniowy

Przykład – popularność imion żeńskich

Wykres panelowy liniowy

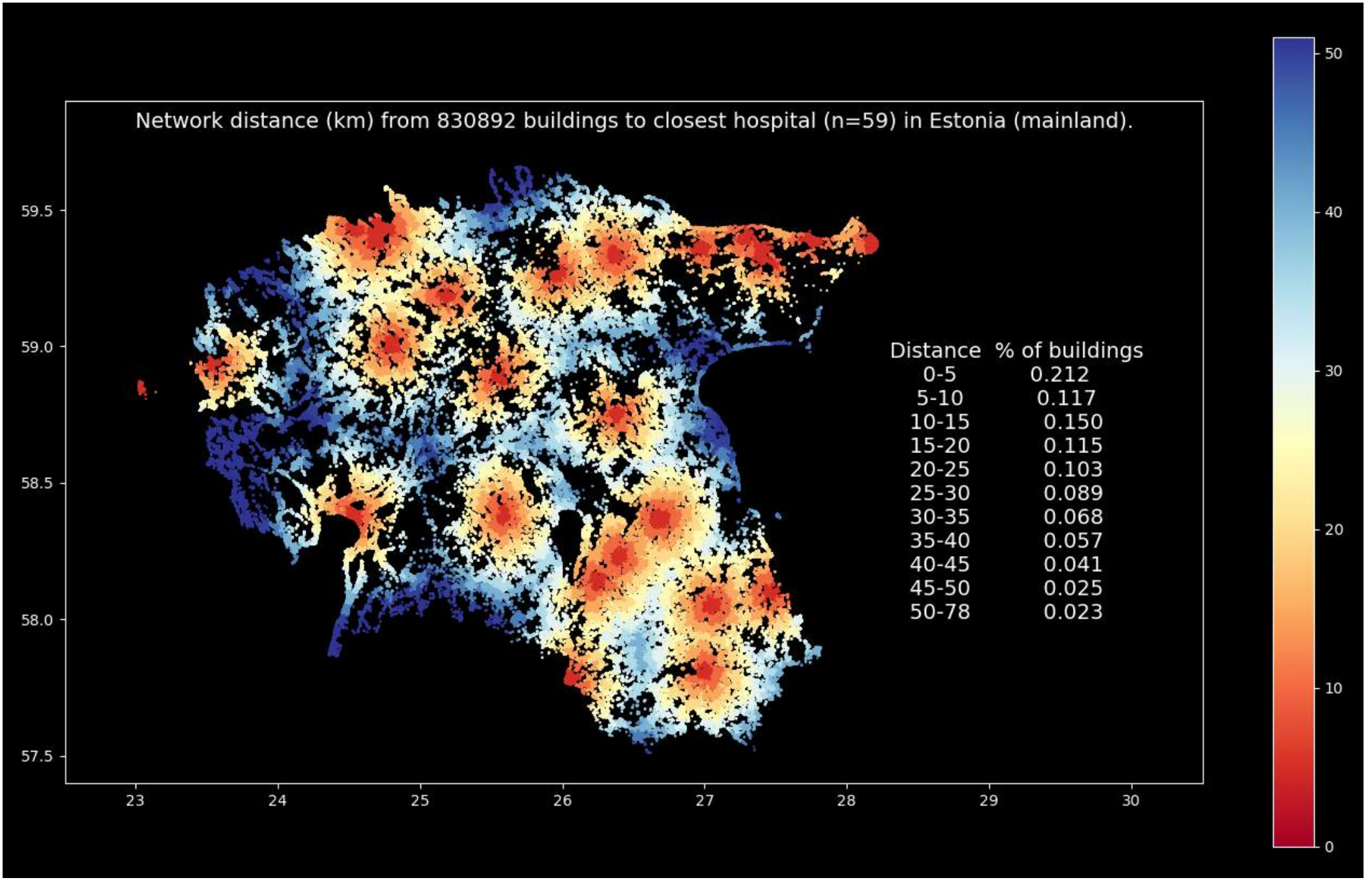
Wykres. Odpływy i napływy migracyjne do polskich regionów 2004-2020



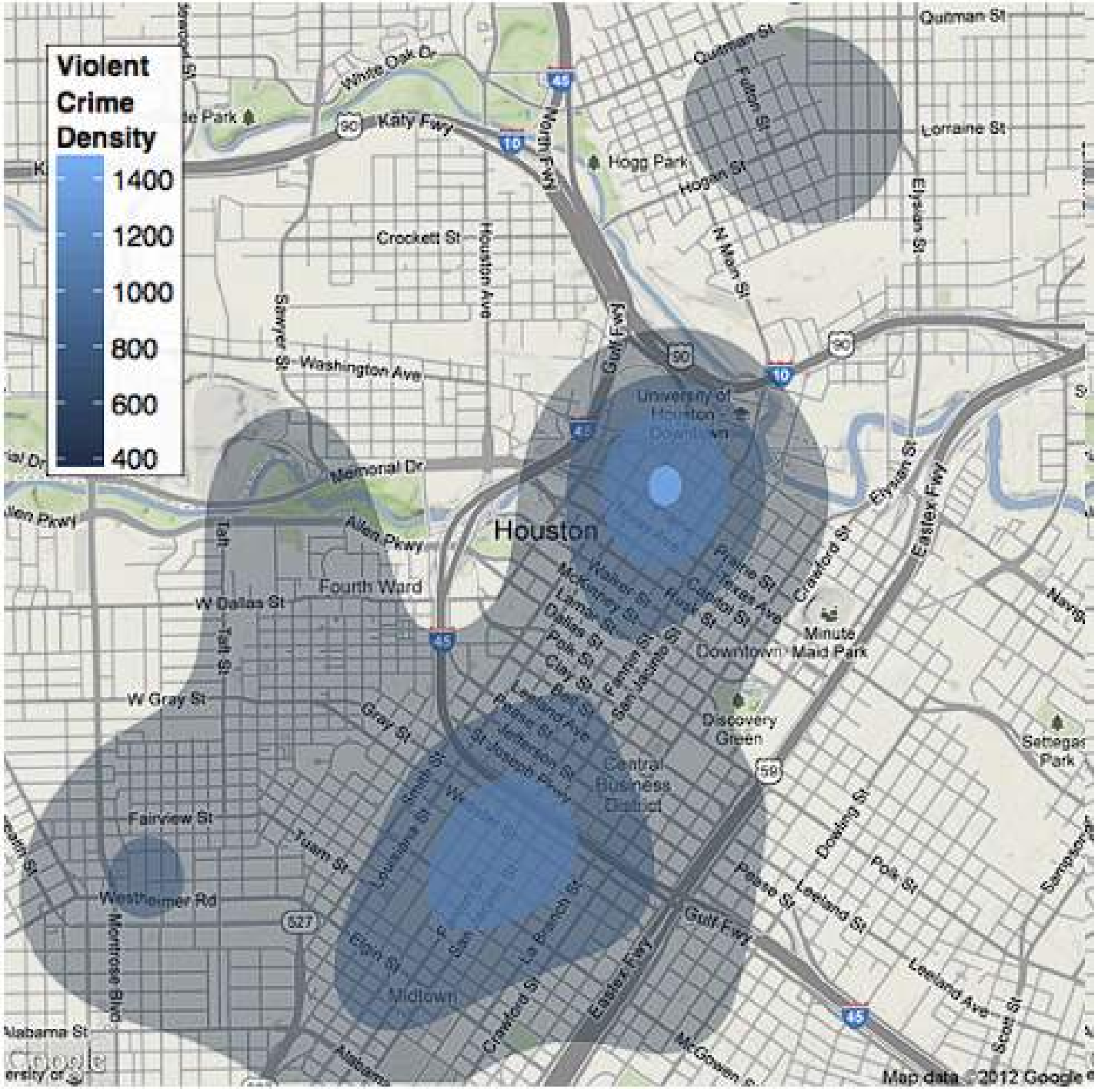
Inne przykłady

ggplot2 R library – wykres gętości

OSMNX Python library – kartogram - czas dostępności

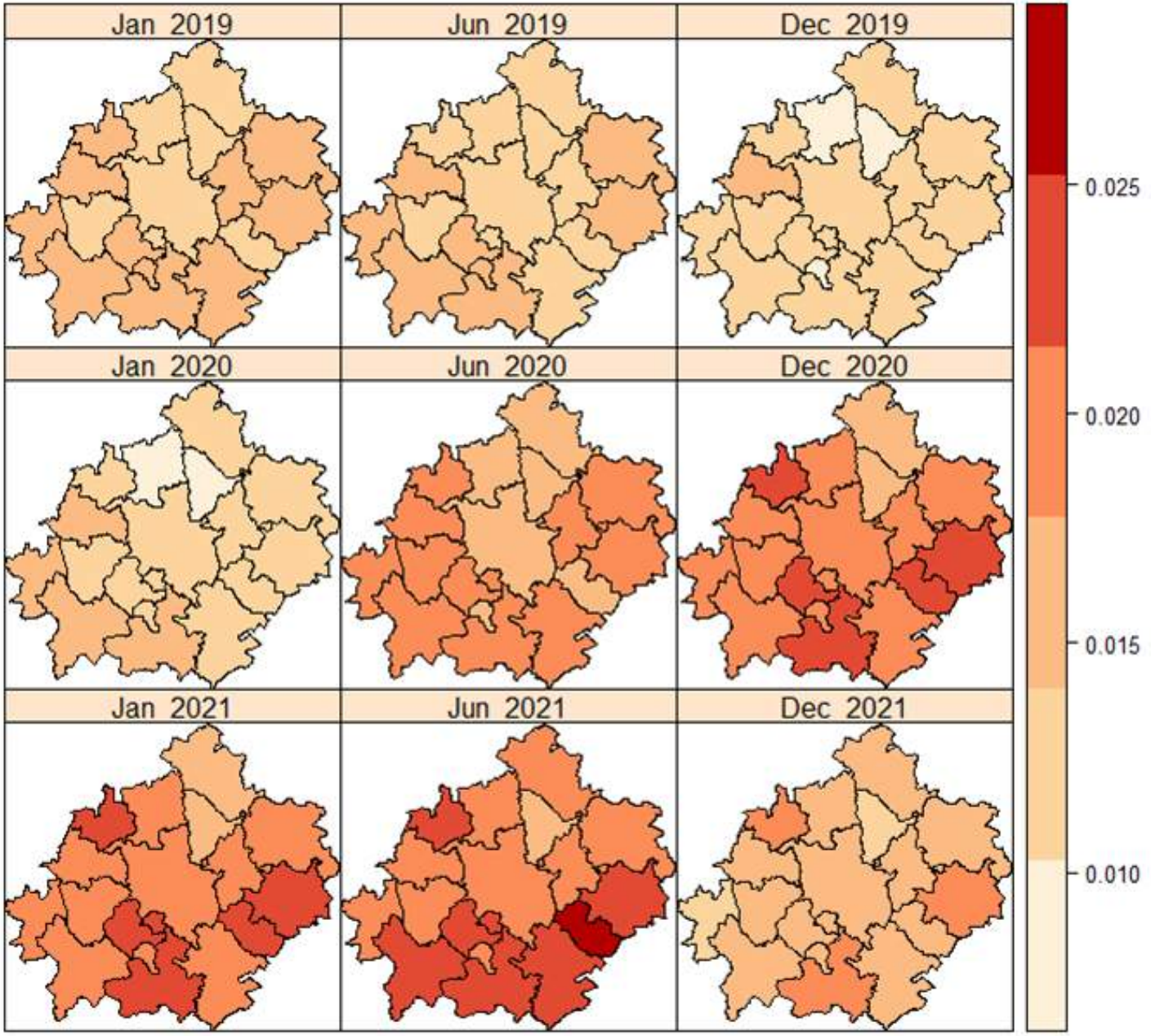


ggmap R library – kartogram gęstości

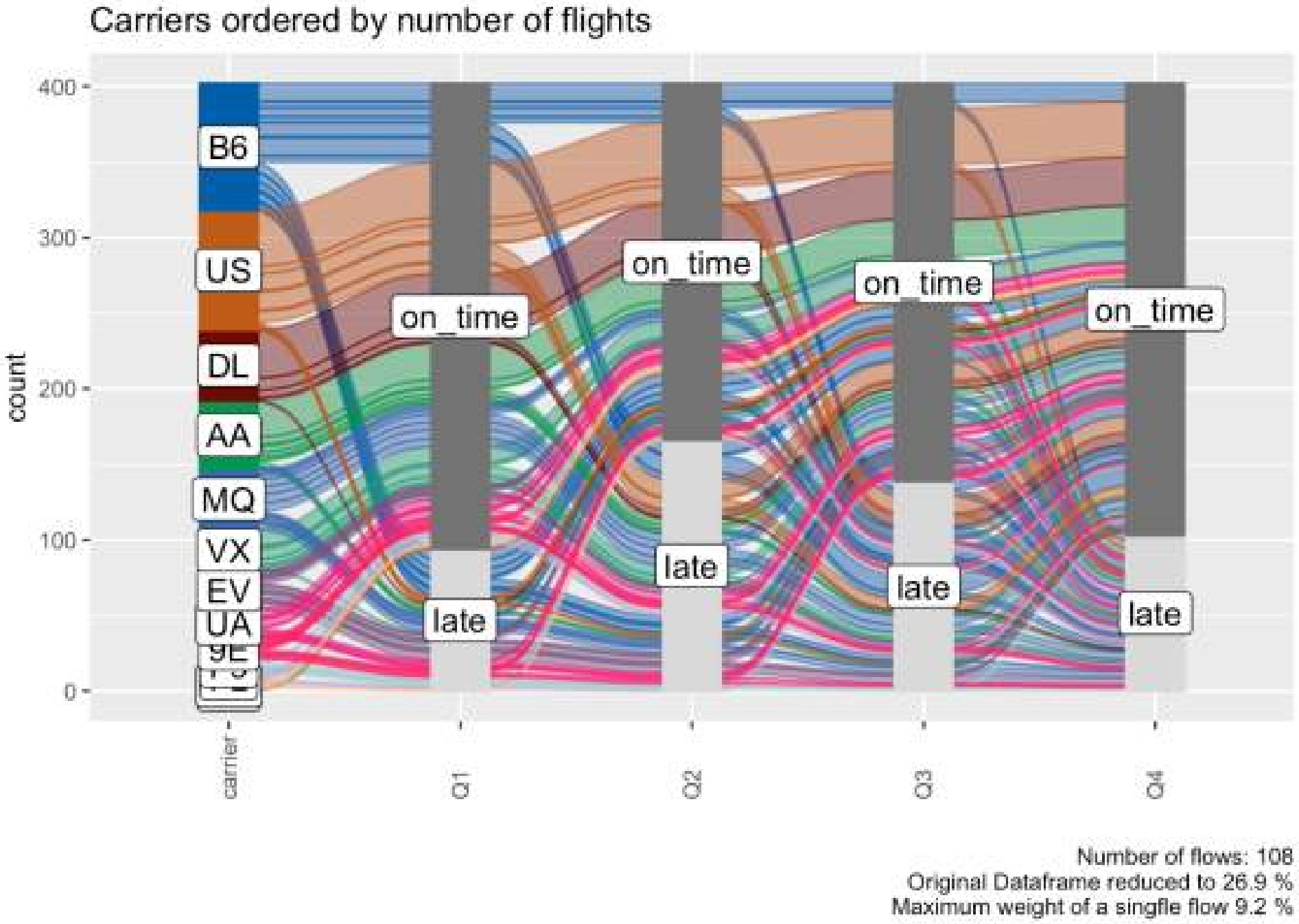


sp R library – kartogramy

Stopa bezrobocia w 17 gminach aglomeracji poznańskiej



Alluvial R library – wykres przepływów



Zalety (dobrej) grafiki:

- W porównaniu z innymi rodzajami prezentacji, dobrze zaprojektowane wykresy są bardziej efektywne i interesujące
- Relacje wizualne przedstawione przez wykresy są łatwiejsze do zapamiętania.

Zalety (dobrej) grafiki:

- W porównaniu z innymi rodzajami prezentacji, dobrze zaprojektowane wykresy są bardziej efektywne i interesujące
- Relacje wizualne przedstawione przez wykresy są łatwiejsze do zapamiętania.
- Wykorzystanie wykresów oszczędza czas - istotne rezultaty mogą być widoczne “na pierwszy rzut oka”.
- Dają całościowy obraz problemu

Podstawowe narzędzia prezentacji danych

- **Tablice częstości:** to tabelaryczny zbiór danych, pokazujący liczbę wystąpień poszczególnych obserwacji.
- Jest to wygodny sposób uniknięcia konieczności wymieniania każdej obserwacji osobno.
- Rozkłady częstości mogą często zapewnić lepszy wgląd we wzorce pojawiające się w zbiorach.

Tablice częstości

Number of Pets	Frequency
1-2	7
3-4	3
5-6	3
7-8	2

Jak mógłby wyglądać zbiór surowych danych na podstawie których powstała powyższa tablica?

Tablice częstości

Przykład – liczba cylindrów

Tabele krzyżowe

- Tabele krzyżowe (dwudzielcze i wielodzielcze): rozszerzenie tabeli częstości na przypadki wielowymiarowe.
- W przypadku dwuwymiarowym dane mogą być oddzielnymi miarami zastosowanymi do tych samych klas, lub mogą to być wspólne miary.

Tabele krzyżowe

Age	Laptop	Phone	Tablet	Digital Camera
20-25	38%	29%	31%	12%
25-30	19%	15%	24%	17%
30-35	23%	19%	11%	27%
35-40	19%	12%	9%	30%
above 40	12%	17%	5%	31%

Jak mogłoby brzmieć pytanie, które zadano respondentom?

Tabele krzyżowe - przykład

Miary środka i położenia

- Średnia arytmetyczna
- Mediana
- Dominanta
- Percentyle i kwartyle

Średnia arytmetyczna

- Suma wszystkich liczb w zbiorze podzielona przez ich liczbę (długość zbioru)

$$m_n = \frac{a_1 + a_2 + \ldots + a_n}{n}$$

- Jaka jest średnia arytmetyczna zbioru?

  2,2,3,3,4,4,5,5,6,6

Mediana

- wartość środkowa zbioru
- mediana wskazuje, że połowa naszych wyników ma wartość poniżej, a druga połowa ma wartość powyżej wartości mediany

Mediana

- wartość środkowa zbioru
- mediana wskazuje, że połowa naszych wyników ma wartość poniżej, a druga połowa ma wartość powyżej wartości mediany
- mediana jest odporna na przypadki odstające znajdujące się w zbiorze (np. w Polsce mediana wynagrodzeń w roku 2023 była o ok. 1000 PLN niższa niż średnia:)

Mediana

- wartość środkowa zbioru
- mediana wskazuje, że połowa naszych wyników ma wartość poniżej, a druga połowa ma wartość powyżej wartości mediany
- mediana jest odporna na przypadki odstające znajdujące się w zbiorze (np. w Polsce mediana wynagrodzeń w roku 2023 była o ok. 1000 PLN niższa niż średnia: ~7300 vs ~6300 brutto)

Dominanta

- Tzw. wartość modalna, wartość najczęstsza zbioru
- Jeżeli dwie wartości pojawiają się z równą i największą częstością, obie są dominantami
- W przypadku wynagrodzeń w 2023 roku dominantą była kwota:

Dominanta

- Tzw. wartość modalna, wartość najczęstsza zbioru
- Jeżeli dwie wartości pojawiają się z równą i największą częstością, obie są dominantami
- W przypadku wynagrodzeń w 2023 roku dominantą była kwota: ~3328 PLN brutto

Miary rozproszenia (rozrzutu)

- Rozstęp (różnica między największą a najmniejszą wartością w zbiorze)
- Wariancja
- Odchylenie standardowe
- Współczynnik Ginniego

Wariancja

- Jest obliczana przez zsumowanie średniej kwadratów odchyleń od średniej zbioru:

$$\sigma^2 = \frac{\Sigma\left(X - \bar{X}\right)^2}{n-1}$$

- Wariancja mówi o stopniu rozrzutu danych. Im bardziej rozłożone są dane, tym większa jest wariancja **w stosunku do średniej**.

Odchylenie standardowe

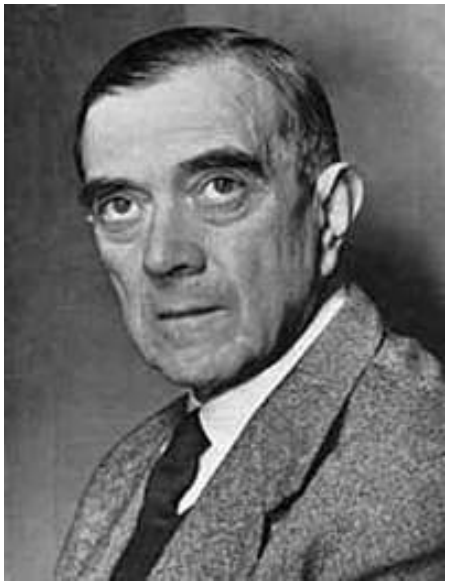
- Podobnie jak wariancja mierzy rozproszenie w zbiorze **w odniesieniu do średniej**
- Jest to pierwiastek kwadratowy z wariancji:

$$\sigma = \sqrt{\frac{\Sigma\left(X - \bar{X}\right)^2}{n-1}}$$

- Jest to również najczęściej pojawiająca się miara opisowa (oprócz średniej arytmetycznej)

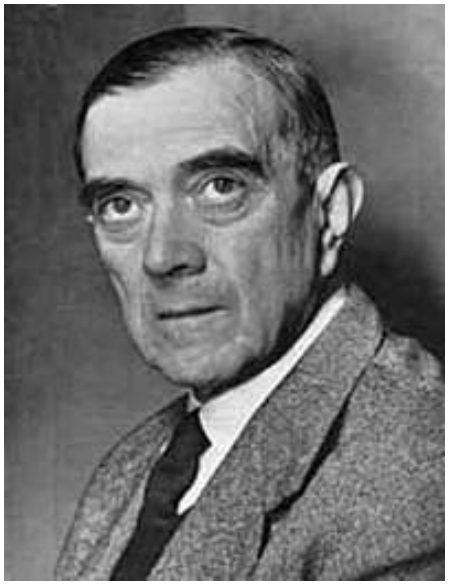
Odchylenie standardowe - przykład

Współczynnik (indeks) Giniego



Corrado Gini
1884 - 1965

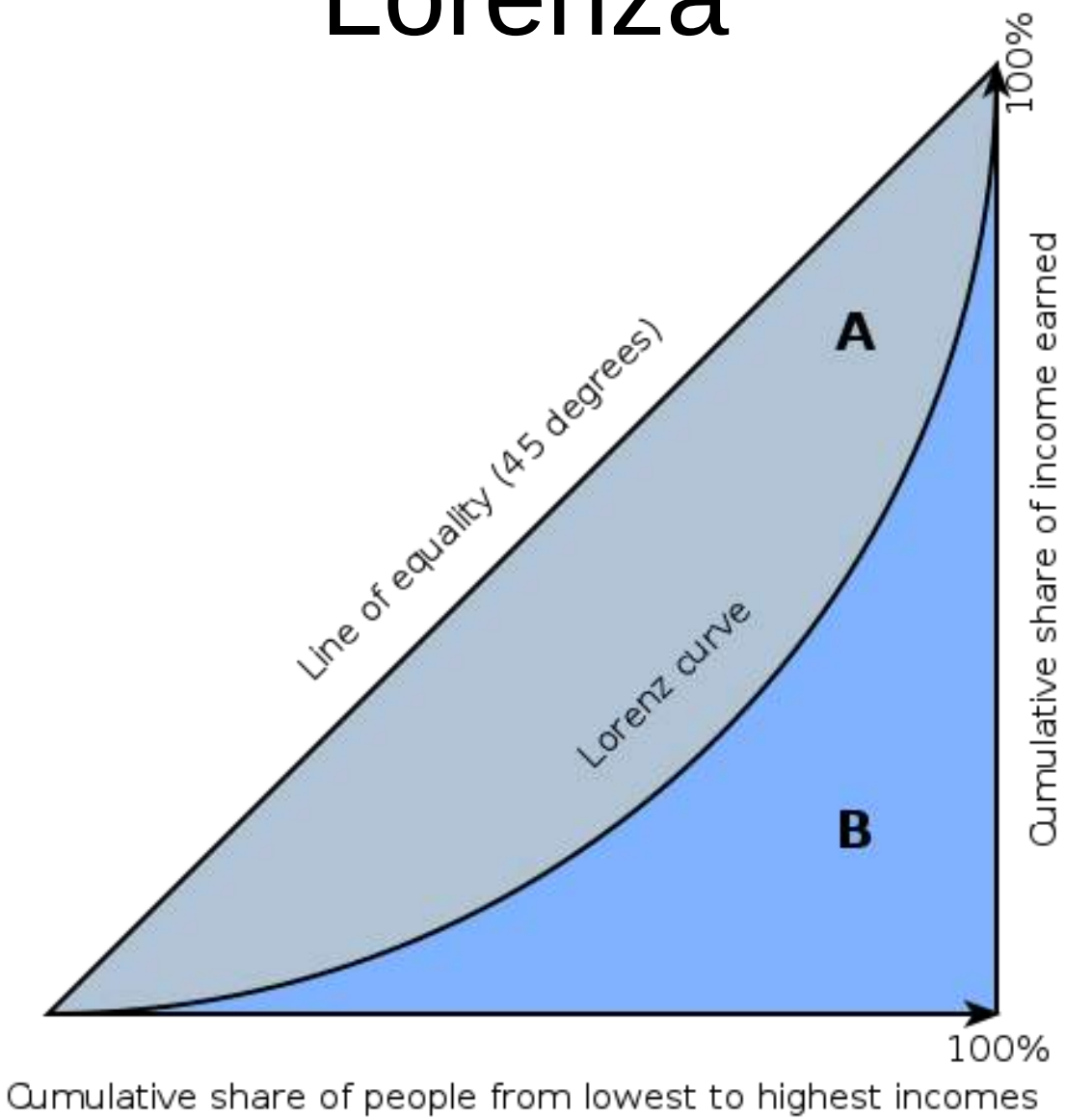
Współczynnik (indeks) Giniego



Corrado Gini
1884 - 1965

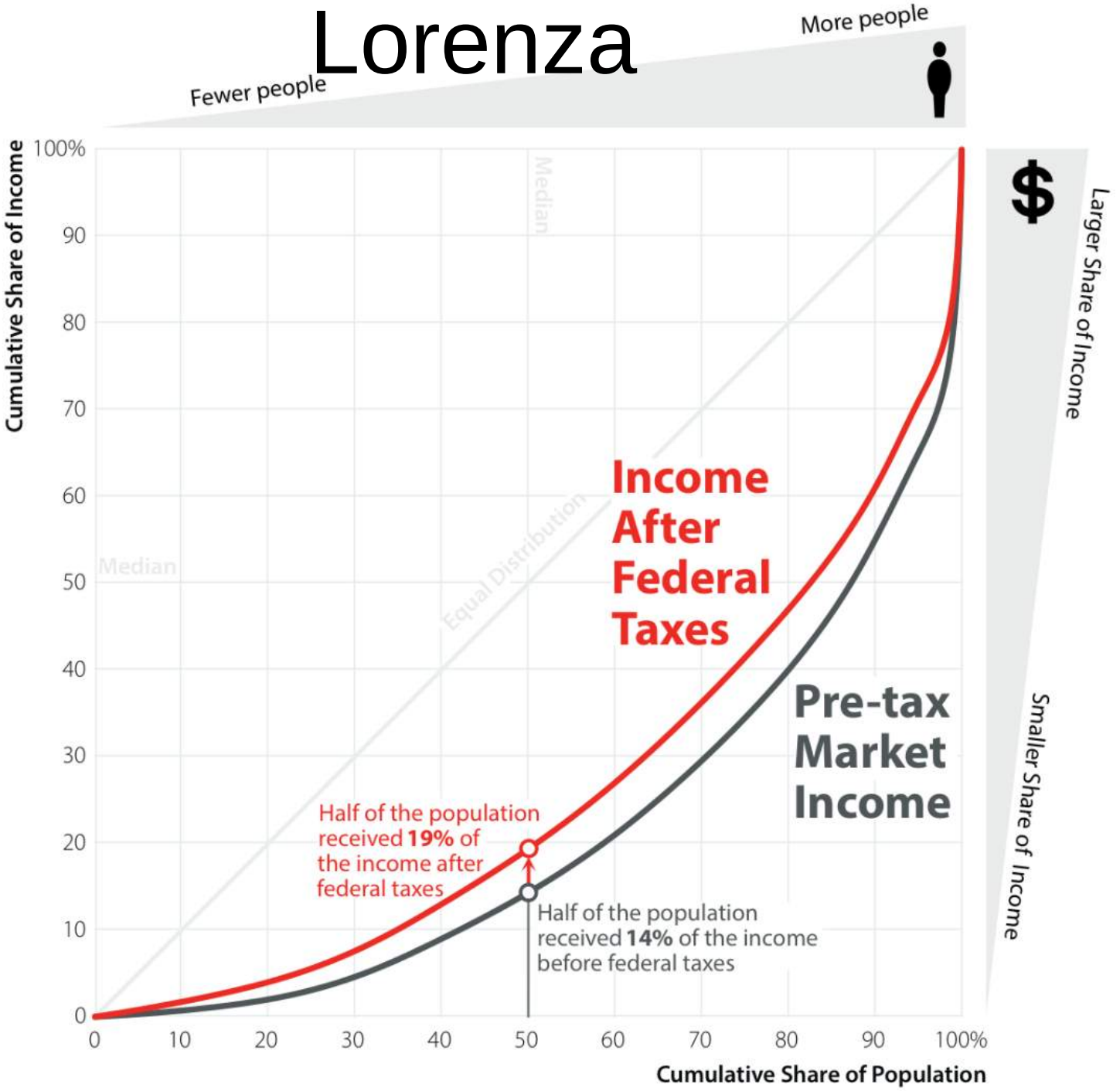
- Miara rozproszenia rozkładu zmiennej
- Zaprojektowana, aby mierzyć nierówności dochodowe, w rozkładzie bogactwa i konsumpcji
- Może mieścić się w przedziale od 0 (całkowita równość) do 1 (całkowita nierówność)
- Czasami jest wyrażany w procentach

Współczynnik Giniego i krzywa Lorenza

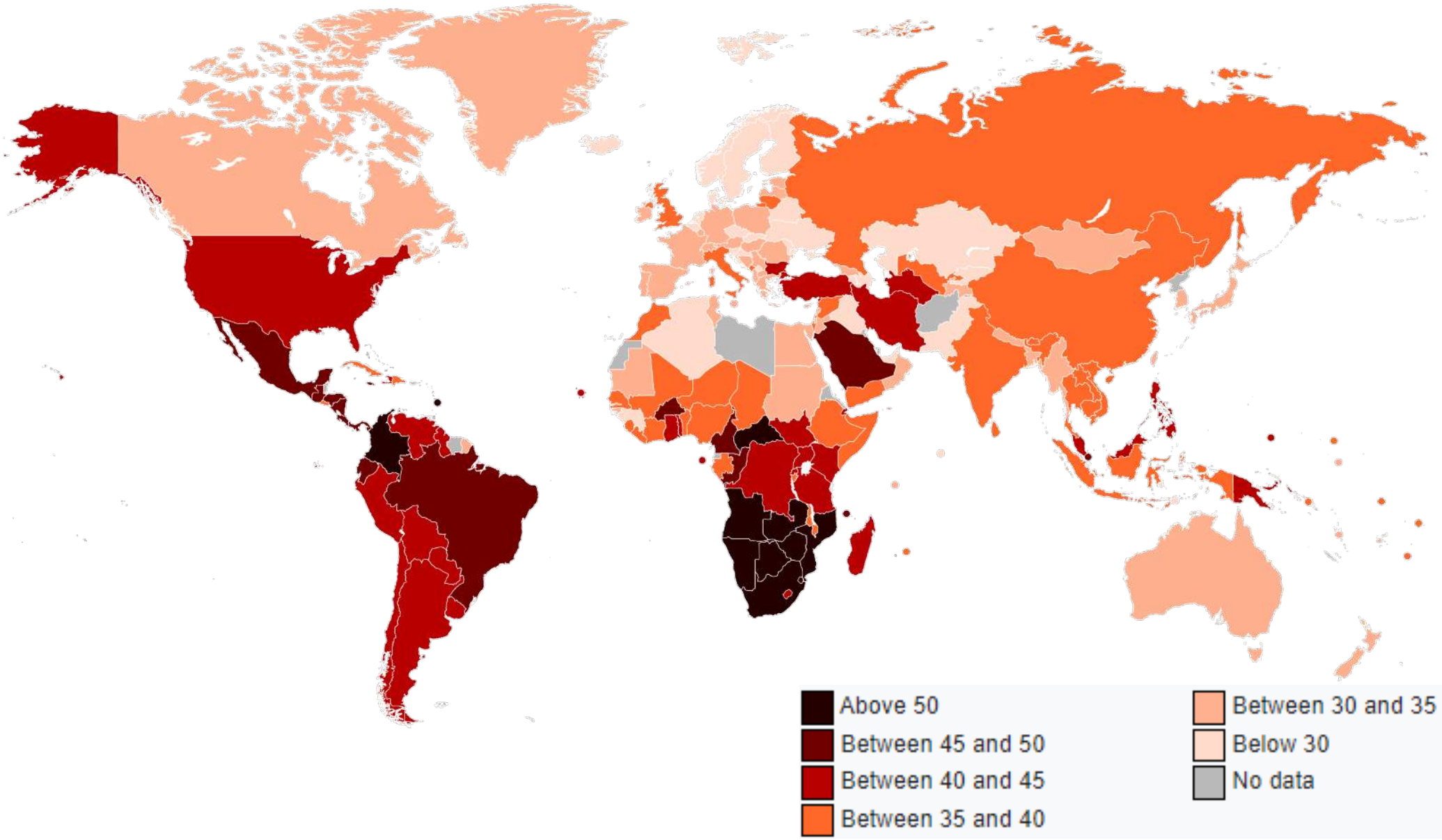


Gini = A/(A + B)

Współczynnik Giniego i krzywa Lorenza



Współczynnik Giniego



Współczynnik Giniego

Przykład.

Współczynnik Giniego

https://www.youtube.com/watch?v=y8y-gaNbe4U&ab_channel=KhanAcademy

Miary koncentracji i asymetrii

- Są sytuacje, w których badanie średniego poziomu zmiennej i rozproszenia jej wartości nie wskazuje na istnienie różnic między badanymi zbiorowościami.
- Pomocne wtedy mogą być dodatkowe wskaźniki pozwalające określić kształt rozkładu:
  - Percentyle
  - Skośność
  - Kurtoza

Percentyle

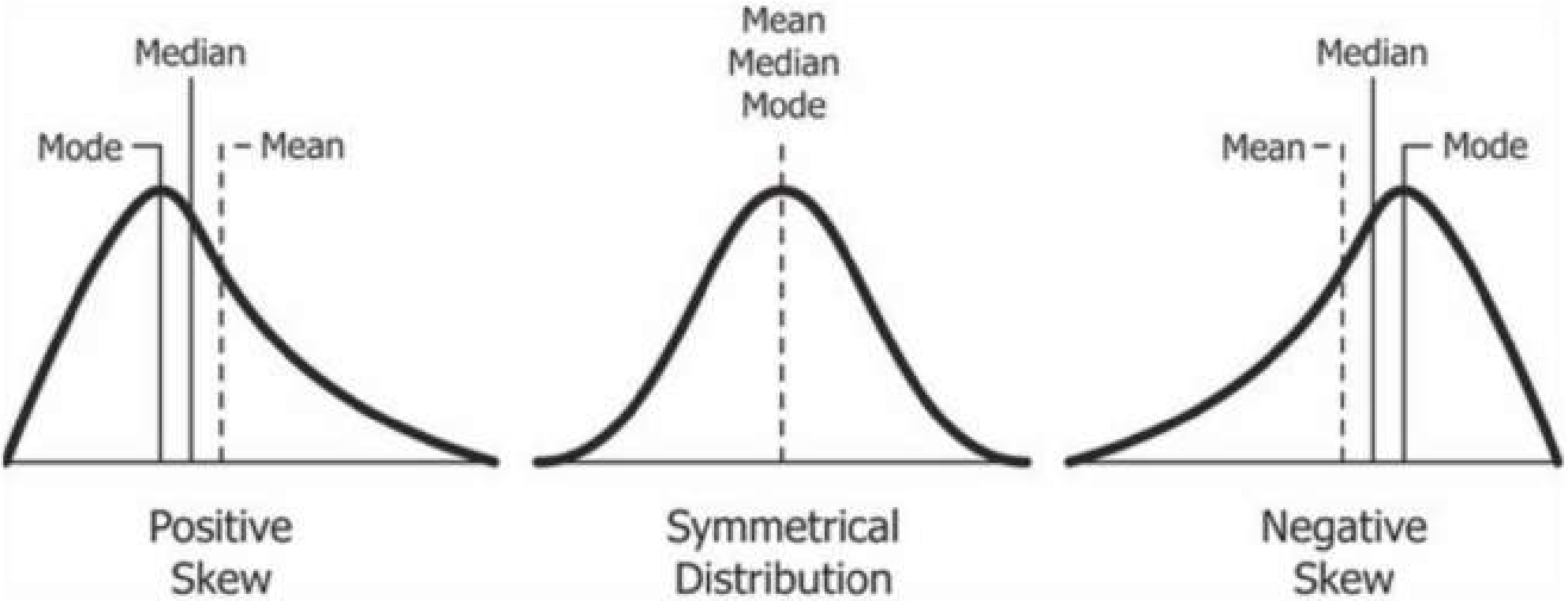
- *N-ty* percentyl zbioru danych to wartość, która odcina pierwsze n procent danych, gdy wszystkie wartości są posortowane od najmniejszej do największej
- Najczęściej wykorzystywane to mediana (50), I kwartyl (25) oraz III kwartyl (75)
- Przykład: *Jaki wynik musi uzyskać uczeń w danym teście, aby znaleźć się w 10% najlepszych wyników?*

Percentyle - przykład

Skośność

- Jest miarą asymetrii rozkładu. Wartość ta może być dodatnia lub ujemna.
- Negatywna skośność wskazuje, że ogon znajduje się po lewej stronie rozkładu
- Pozytywna skośność wskazuje, że ogon znajduje się po prawej stronie rozkładu
- Wartość zero oznacza, że w rozkładzie nie ma skośności, co oznacza, że rozkład jest idealnie symetryczny.

Skośność



Współczynnik skośności:

- Symetryczne: Wartości od -0,5 do 0,5
- Skośność średnia: Wartości pomiędzy -1 i -0,5 lub pomiędzy 0,5 i 1
- Skośność wysoka: Wartości mniejsze niż -1 lub większe niż 1

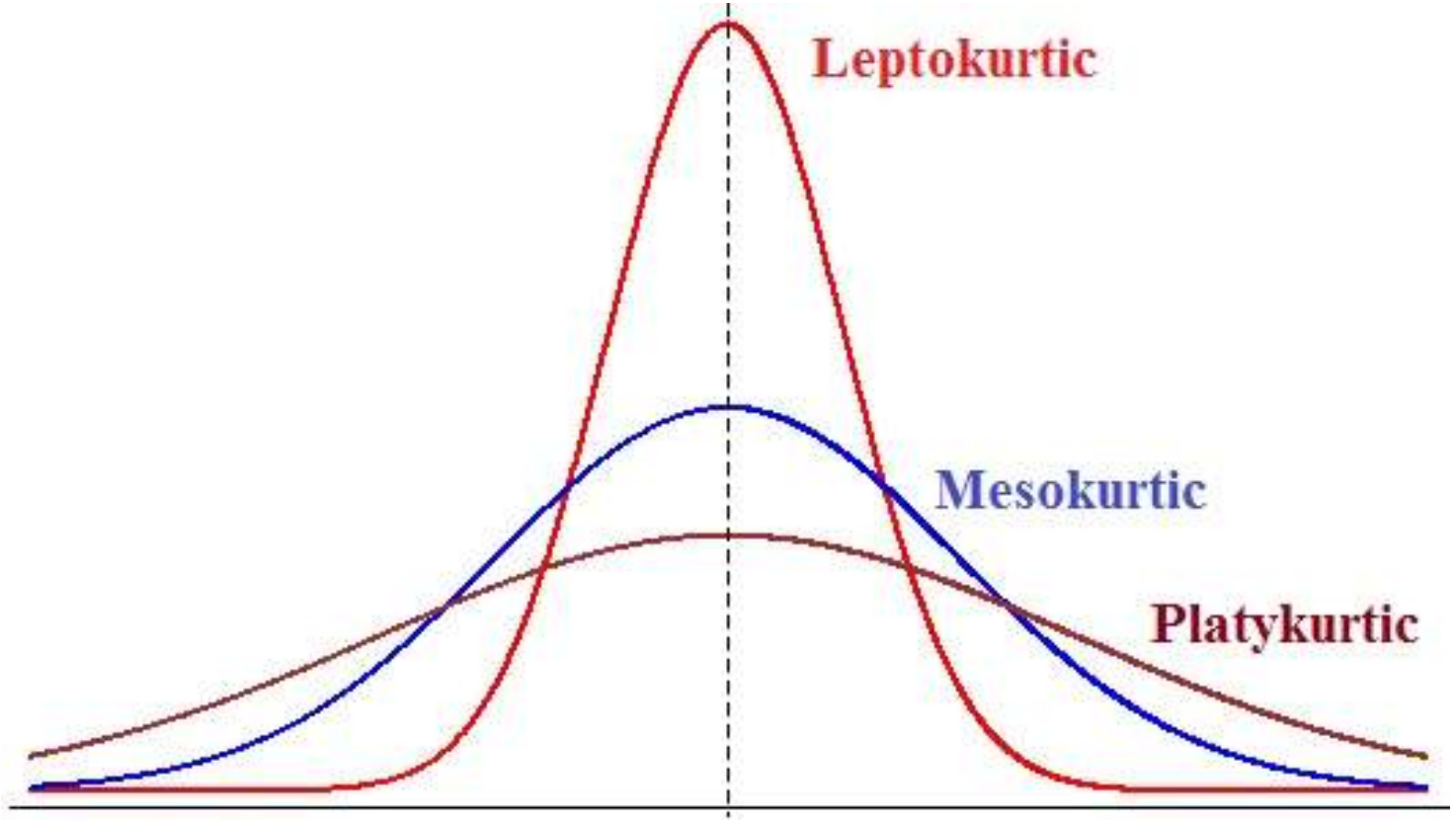
Skośność danych

- Przykład

Kurtoza

- Najpopularniejszą miarą skupienia obserwacji wokół średniej jest kurtoza
- Im wyższa jest jej wartość, tym bardziej wysmukła jest krzywa liczebności, a zatem większa koncentracja cechy wokół średniej

Kurtoza



Kurt = 3 - mezokurtyczny
Kurt > 3 - leptokurtyczny
Kurt < 3 - platykurtyczny

Miary zależności i bliskości

- Współczynniki korelacji Pearsona
- Test chi-kwadrat
- Współczynnik korelacji rang Spearmana
- Autokorelacja

- Korelacja przestrzenna

Korelacja (ogólnie)

- Mierzy (potencjalną) zależność pomiędzy dwoma zmiennymi: jak jedna zmienna ewoluuje wraz ze zmianą innej
- To, że zmienne są skorelowane **nie oznacza**, że jedna wpływa na/wywołuje drugą

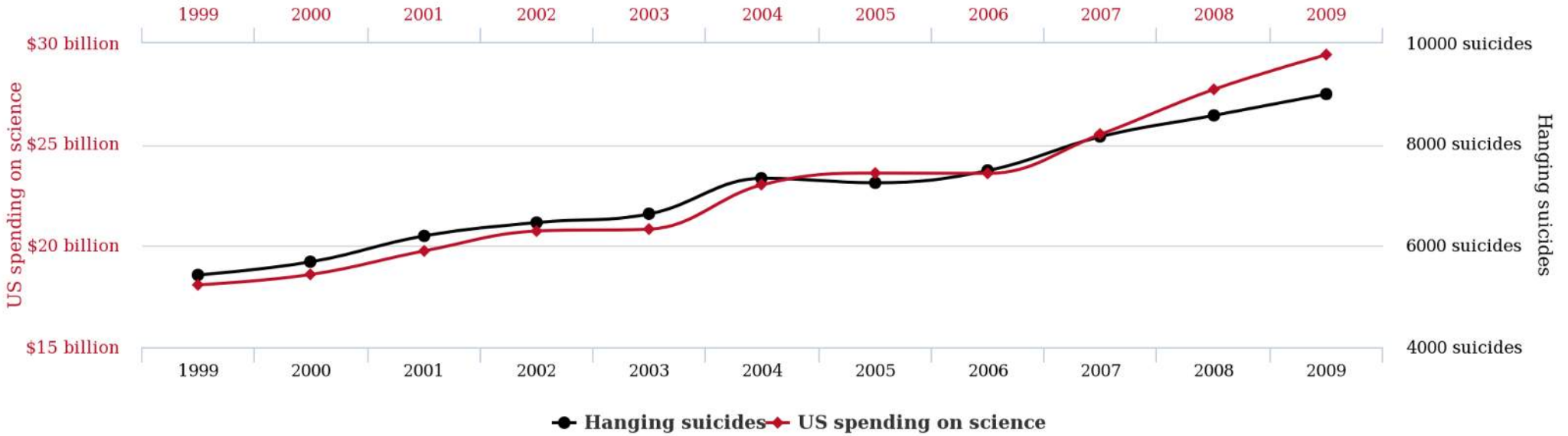
Korelacja (ogólnie)

- Mierzy (potencjalną) zależność pomiędzy dwoma zmiennymi: jak jedna zmienna ewoluuje wraz ze zmianą innej
- To, że zmienne są skorelowane **nie oznacza**, że jedna wpływa na/wywołuje drugą
- Pozwala stwierdzić, które zmienne ewoluują w tym samym kierunku, które w przeciwnym, a które są niezależne.
- Metody parametryczne i nieparametryczne obliczania korelacji

US spending on science, space, and technology

correlates with

Suicides by hanging, strangulation and suffocation



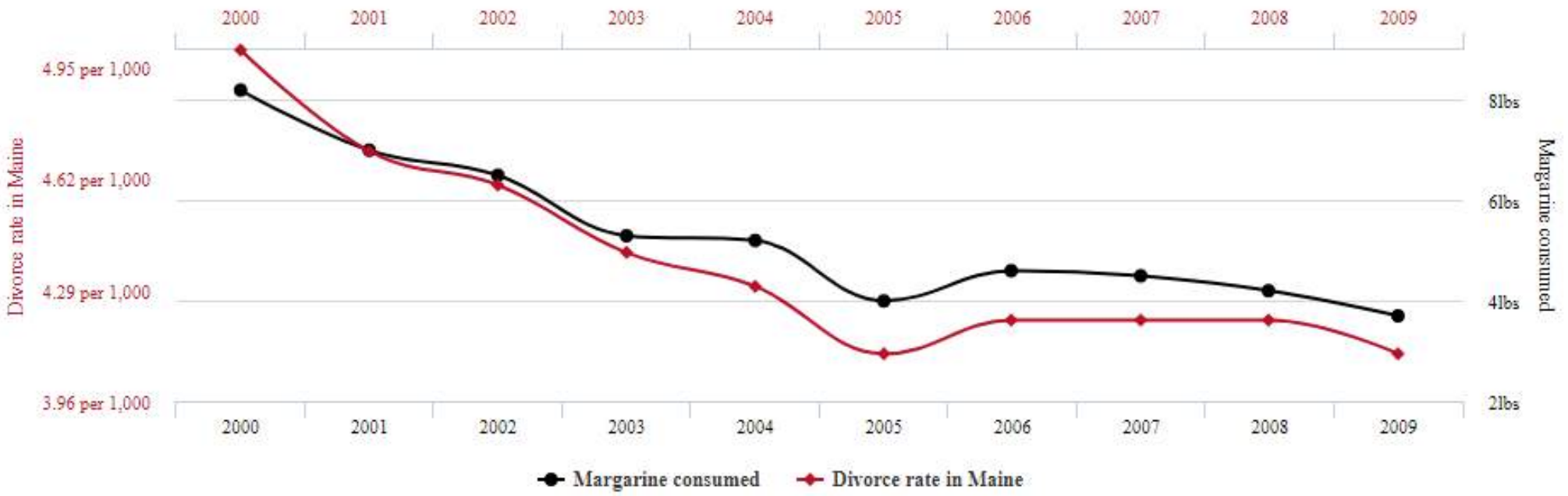
tylervigen.com

Divorce rate in Maine

correlates with

Per capita consumption of margarine

Correlation: 99.26% (r=0.992558)



tylervigen.com

Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

Współczynnik korelacji Pearsona

Karl Pearson
1857-1936

Współczynnik korelacji Pearsona

Karl Pearson
1857-1936

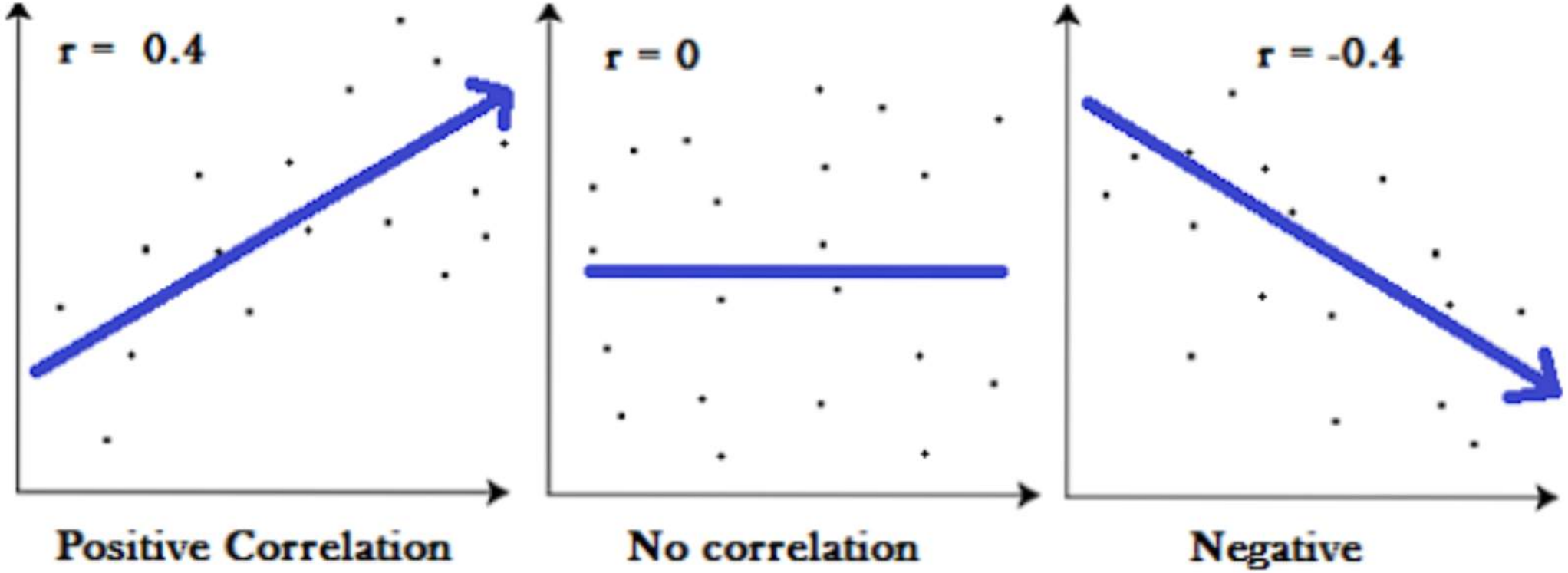
- Tzw. *r Pearsona*
- Mierzy korelację liniową pomiędzy dwoma seriami danych
- Ma zastosowanie tylko w przypadku zmiennych numerycznych
- Przyjmuje zawsze wartości z przedziału 0-1/(-1)

Współczynnik korelacji Pearsona

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2]\,[n\sum y^2 - (\sum y)^2]}}$$

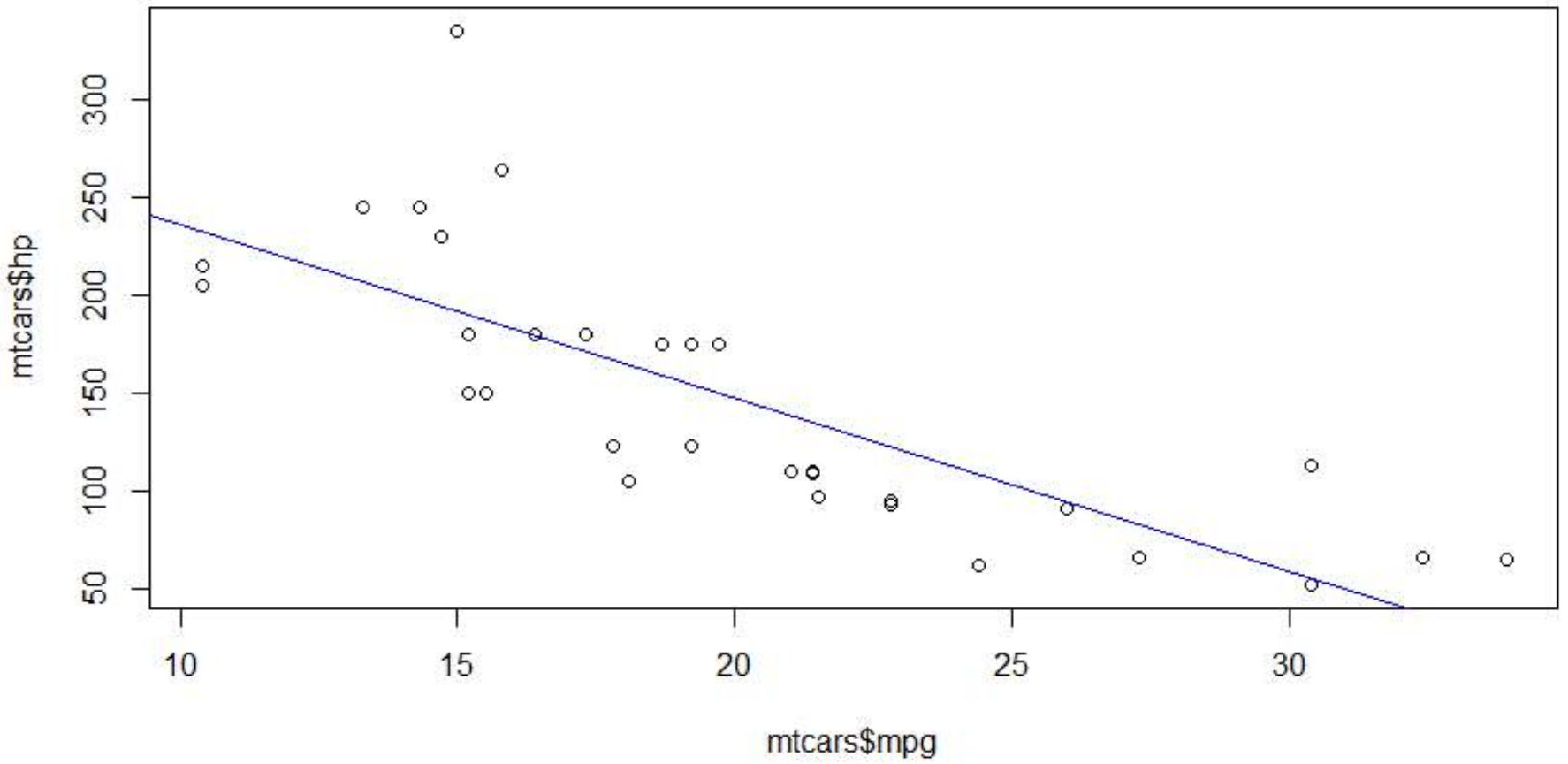
- r = 0 – brak zależności
- r <= 0.3 – korelacja słaba
- r = 0.4 - 0.6 – korelacja umiarkowana
- r = 0.7 - 0.9 – korelacja silna
- r = 1 – korelacja idealna

Współczynnik korelacji Pearsona



Przykład – korelacja Pearsona

moc auta a mile na galon



Test niezależności chi-kwadrat

- Jest metodą statystyczną, która służy do określenia, czy dwie zmienne nominalne są ze sobą powiązane (tablice częstości)
- Jak wszystkie testy statystyczne, tak i ten test zakłada hipotezę zerową i hipotezę alternatywną:

  H0: Zmienne są niezależne.

  H1: Zmienne są ze sobą powiązane (skorelowane).

Test niezależności chi-kwadrat

- Jest metodą statystyczną, która służy do określenia, czy dwie zmienne nominalne są ze sobą powiązane (tablice częstości)
- Jak wszystkie testy statystyczne, tak i ten test zakłada hipotezę zerową i hipotezę alternatywną:

  H0: Zmienne są niezależne.

  H1: Zmienne są ze sobą powiązane (skorelowane).
- Odrzucamy hipotezę zerową, jeśli tzw. ***wartość p***, która pojawia się w wyniku jest mniejsza od ustalonego wcześniej tzw. **poziomu istotności** (zazwyczaj 0.05)
- **Brak informacji o sile związku!**

Test chi-kwadrat - przykład

Współczynnik korelacji rang Spearmana (*rho*)

Charles Spearman
1863 - 1945

Współczynnik korelacji rang Spearmana (*rho*)

Charles Spearman
1863 - 1945

- Podobnie jak współczynnik korelacji Pearsona, tak i tzw. *rho* Spearmana pozwala określić siłę związku pomiędzy zmiennymi
- Jest to metoda nieparametryczna, którą można stosować dla **danych porządkowych**

Współczynnik korelacji rang Spearmana (*rho*)

Charles Spearman
1863 - 1945

- Podobnie jak współczynnik korelacji Pearsona, tak i tzw. *rho* Spearmana pozwala określić siłę związku pomiędzy zmiennymi
- Jest to metoda nieparametryczna, którą można stosować dla **danych porządkowych**
- Wartości *rho* Speramana intepretujemy podobnie jak w przypadku r Pearsona
- Dodatkowo otrzymujemy również ***wartość p*** (prawdopodobieństwo testowe)

rho Spearmana - przykład

Autokorelacja

Autokorelacja

- Jest to podobieństwo pomiędzy poszczególnymi obserwacjami zmiennej, a jej opóźnienieniami czasowymi.
- Opóźnienia czasowe to interwały w których obserwowana jest zmienna

Autokorelacja

- Jest to podobieństwo pomiędzy poszczególnymi obserwacjami zmiennej, a jej opóźnienieniami czasowymi.
- Opóźnienia czasowe to interwały w których obserwowana jest zmienna
- Gdy autokorelacja zmiennej jest wysoka, łatwe staje się przewidywanie jej przyszłych wartości poprzez odniesienie do wartości przeszłych.
- Wyniki i interpretacja → *r* Pearsona

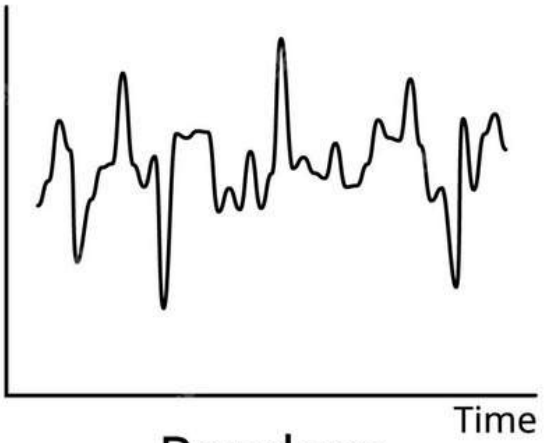
Autokorelacja

- W przypadku danych losowych autokorelacja dla wszystkich opóźnień jest bliska zeru (tzw. "biały szum").
- Dane, które nie są losowe, mają co najmniej jedno istotne opóźnienie

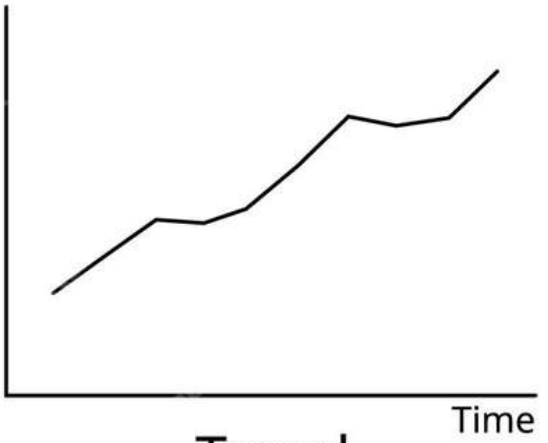
Autokorelacja

- W przypadku danych losowych autokorelacja dla wszystkich opóźnień jest bliska zeru (tzw. "biały szum").
- Dane, które nie są losowe, mają co najmniej jedno istotne opóźnienie
- Jeśli dane nie są losowe, to oznacza, że można je analizować/modelować
- 4 komponenty: trend, wahania cykliczne, wahania sezonowe, losowe fluktuacje

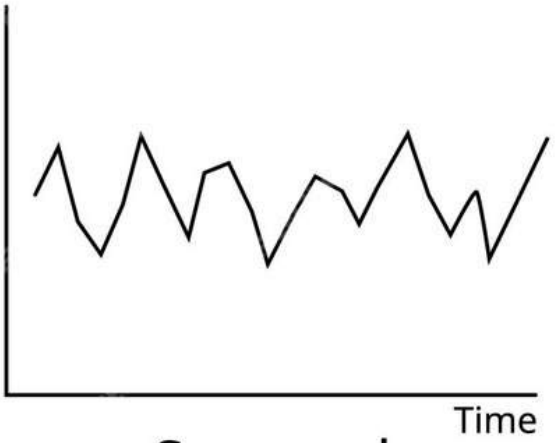
Time Series Components



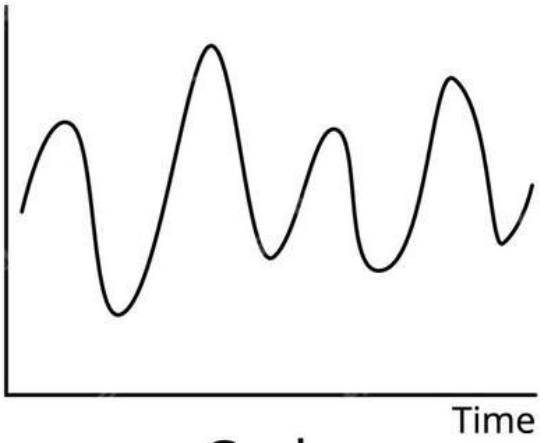
Random



Trend



Seasonal



Cycle

Autokorelacja - przykład

Autokorelacja przestrzenna

- Pierwsze Prawo Geografii:

  "everything is related to everything else, but near things are more related than distant things."

- Jest to fundamentalny koncept analizy przestrzennej w geografii (i nie tylko)

Waldo Tobler
1930-2018

Autokorelacja przestrzenna

- Pomaga zrozumieć, w jakim stopniu jeden obiekt jest podobny do innych pobliskich obiektów
- W przypadku gdy sąsiadujące ze sobą w przestrzeni obiekty mają podobne wartości danych, mamy dodatnią (pozytywną) autokorelację przestrzenną

Autokorelacja przestrzenna

- Pomaga zrozumieć, w jakim stopniu jeden obiekt jest podobny do innych pobliskich obiektów
- W przypadku gdy sąsiadujące ze sobą w przestrzeni obiekty mają podobne wartości danych, mamy dodatnią (pozytywną) autokorelację przestrzenną
- Wskaźnik “**I Morana”** jest najczęściej stosowaną miarą autokorelacji przestrzennej

Autokorelacja przestrzenna

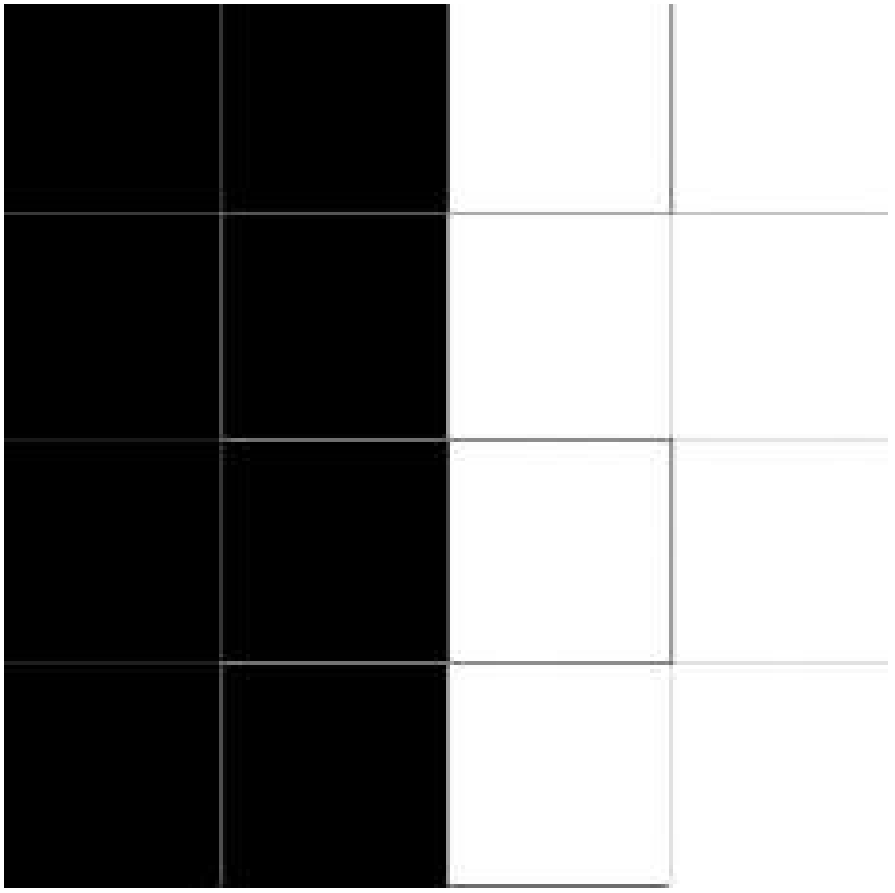
- I Morana przyjmuje wartości od -1 do 1
- Interpretacja jest nieco inna, niż w przypadku klasycznych miar korelacji:

  -1 to idealne skupienie niepodobnych wartości (doskonałe rozproszenie)

  0 to idealna losowość

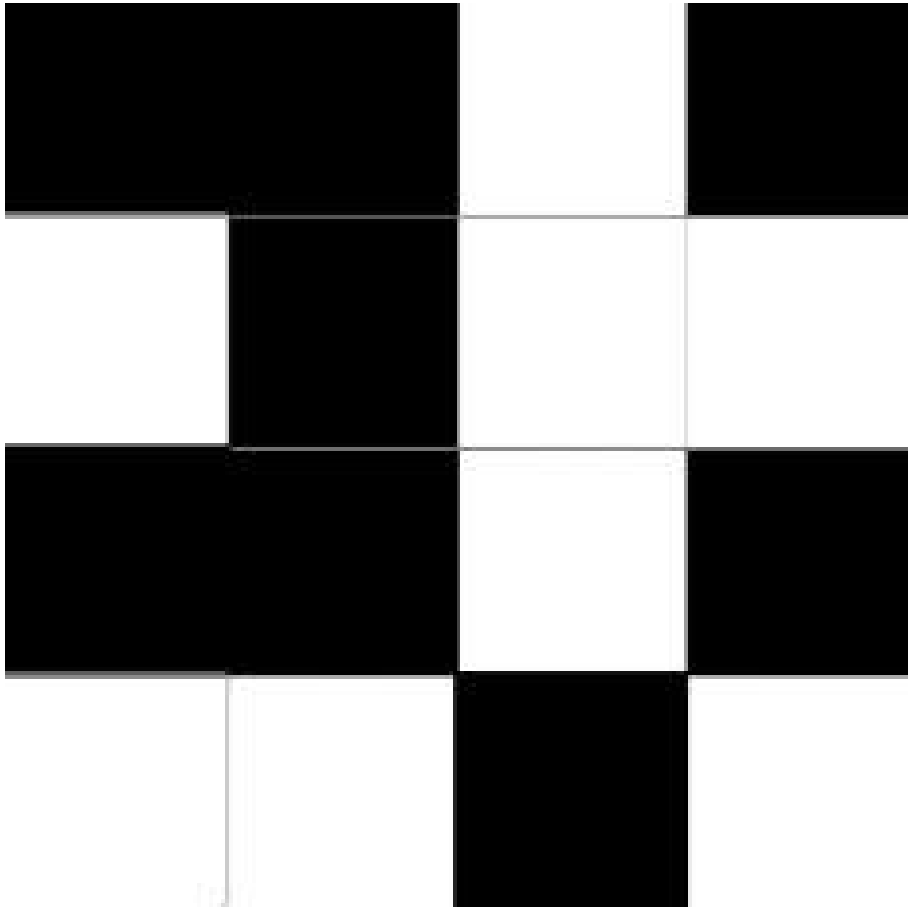
  1 to idelana pozytywna autokorelacja przestrzenna (idealna klasteryzacja)

Autokorelacja przestrzenna



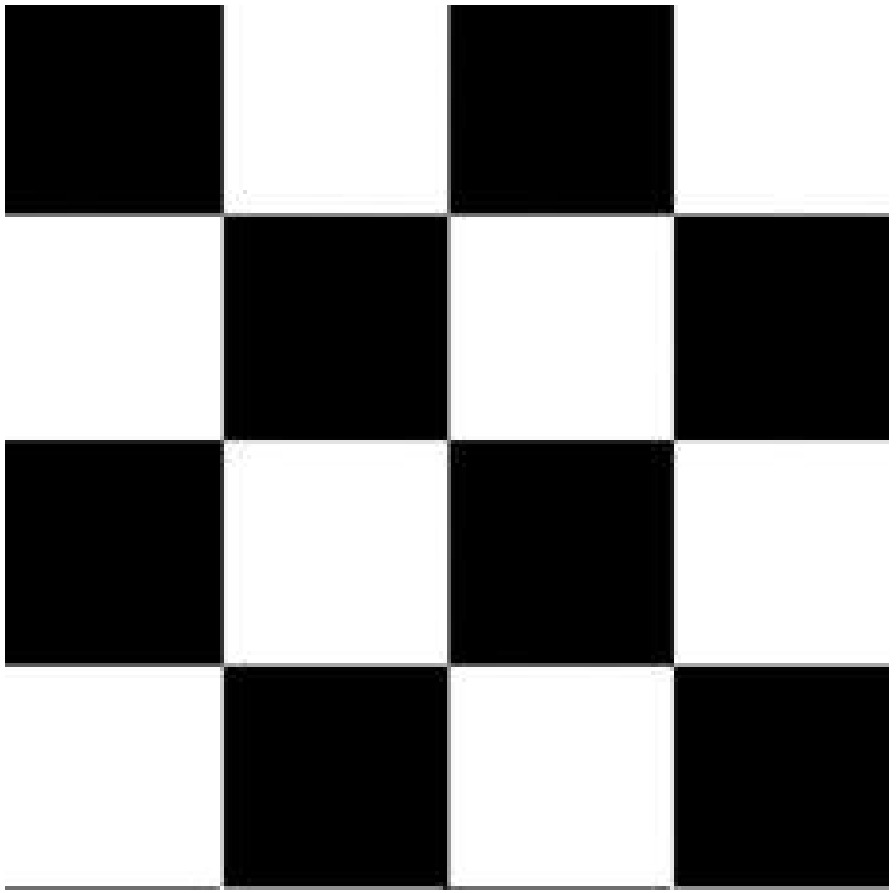
Idealna pozytywna korelacja przestrzenna; Moran I = 1

Autokorelacja przestrzenna



Losowe rozproszenie, Moran I = 0

Autokorelacja przestrzenna



Idealna dyspersja, Moran I = -1

Autokorelacja przestrzenna

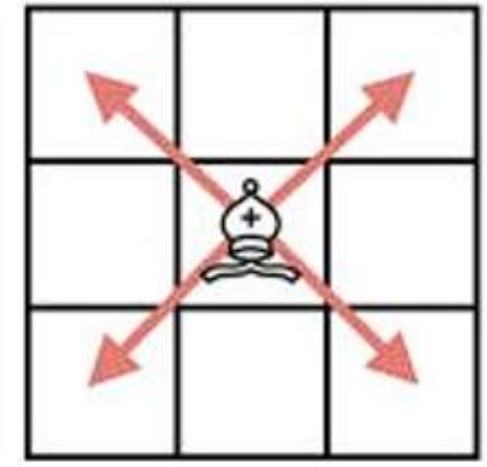
Macierze wag przestrzennych:

- Macierze wag przestrzennych bazują na relacjach przestrzennych obiektów.
- Istnieje wiele rodzajów macierzy wag przestrzennych (np. oparte na dystansie, na gęstości sieci drogowej, etc.)

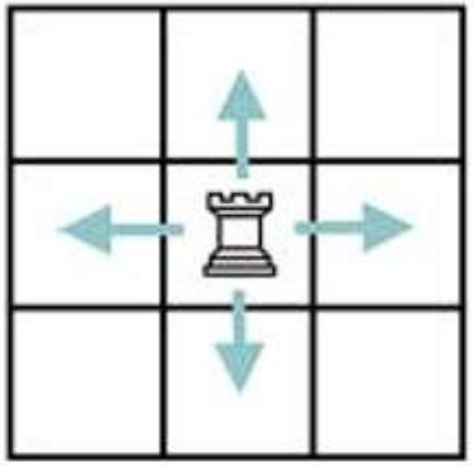
Autokorelacja przestrzenna

Macierze wag przestrzennych:

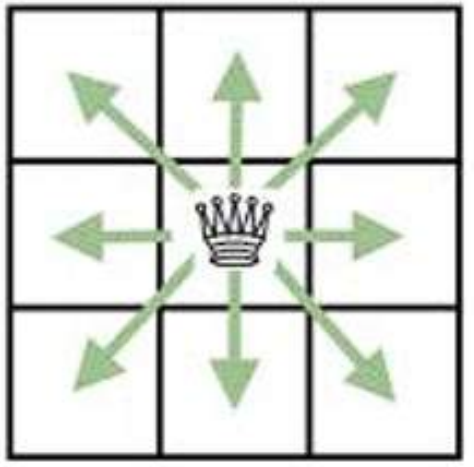
- Macierze wag przestrzennych bazują na relacjach przestrzennych obiektów.
- Istnieje wiele rodzajów macierzy wag przestrzennych (np. oparte na dystansie, na gęstości sieci drogowej, etc.)
- Podstawowe typy sąsiedztwa:



Sąsiedztwo typu Bishop

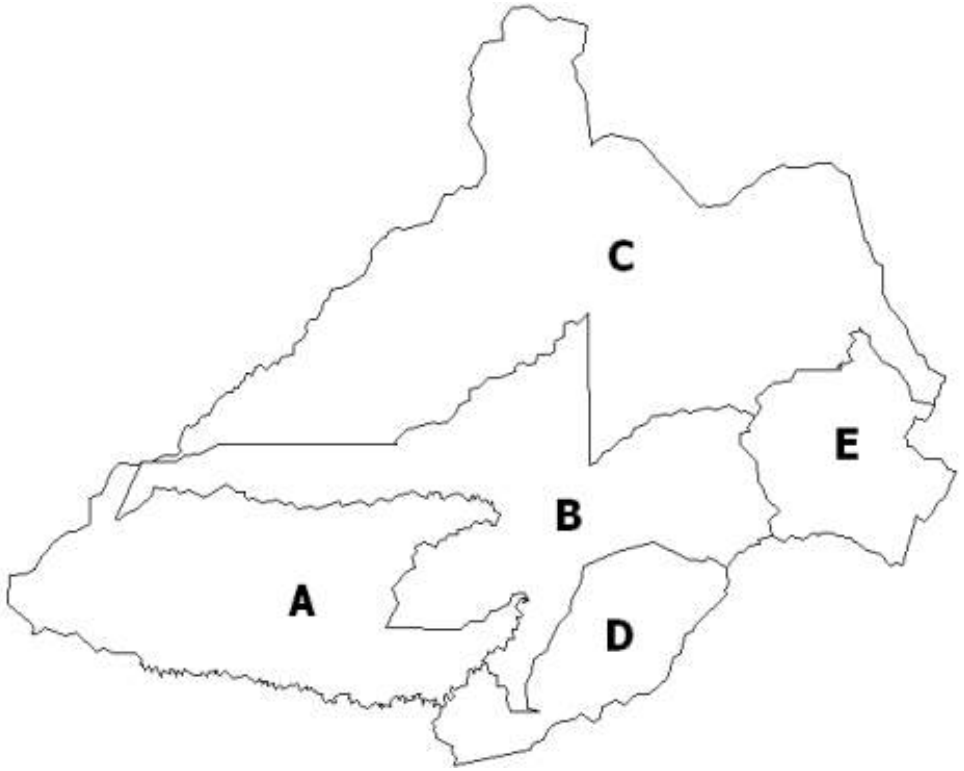


Sąsiedztwo typu Rook



Sąsiedztwo typu Queen

Przykład macierzy wag przestrzennych



$$\mathbf{W}_s = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 0.5 & 0 & 0.5 & 0 \\ 0.25 & 0 & 0.25 & 0.25 & 0.25 \\ 0 & 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 \end{pmatrix} \end{array}$$

Autokorelacja przestrzenna

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

N – liczba obserwacji
$\mathbf{X_i}$ – wartość obserwacji w obiekcie i
$\mathbf{X_j}$ – wartość obserwacji w obiekcie j
$\mathbf{w_{ij}}$ – matryca wag przestrzennych dla połączeń obiektów i oraz j

Autokorelacja przestrzenna

- Moran I procedura:

  1) dane geoprzestrzenne (*.shp, .geopackage, .PostGIS, .Sqlite,* etc.)

  2) zdefiniowanie “sąsiadów” za pomocą matrycy wag przestrzennych

Autokorelacja przestrzenna

- Moran I procedura:

  1) dane geoprzestrzenne (*.shp, .geopackage, .PostGIS, .Sqlite,* etc.)

  2) zdefiniowanie “sąsiadów” za pomocą matrycy wag przestrzennych

  3) przydzielenie wag poszczególnym "sąsiadom”

  4) obliczenie wartości statystyki testowej i weryfikacja hipotezy 0:

  H0: dane są losowo rozmieszczone w przestrzeni (brak korelacji)
  H1: dane przejawiają wzorce (istnieje korelacja)

Autokorelacja przestrzenna

- Przykład – dochód mieszkańców w hrabstwach stanu Maine i New Hampshire

Korelacja i zależność przyczynowo-skutkowa

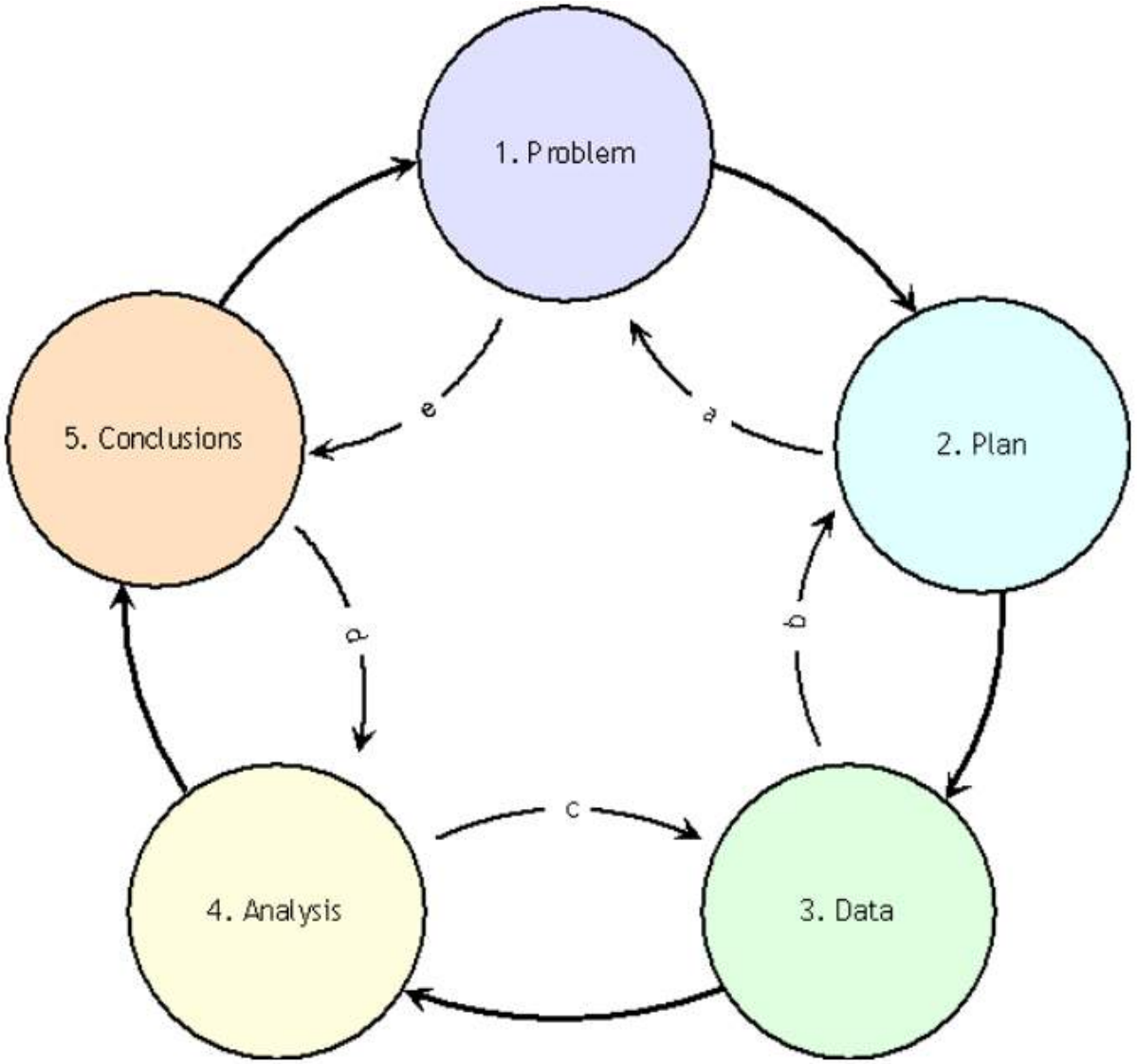
https://www.youtube.com/watch?v=Nre4cjz3U4A&ab_channel=KhanAcademyPoPolsku

Metoda statystyczna

Metoda statystyczna

- Analiza statystyczna nie jest czysto technicznym ćwiczeniem
- Powinna być realizowana w szerokim kontekście zarówno metodologicznym jak i teoretycznym
- Wymaga więc zarówno wiedzy tematycznej jak i technicznej
- Powinna opierać się na modelu: PPDAK

Metoda statystyczna (PPDAK)



Metoda statystyczna

Problem:

- Zrozumienie i zdefiniowanie problemu jest istotną częścią całego procesu analitycznego (O *czym są te badania? Czego chcę się dowiedzieć?*)
- Problem powinien wyczerpywać zakres badania i uwzględniać zależności pomiędzy zmiennymi

Metoda statystyczna

Problem:

- Zrozumienie i zdefiniowanie problemu jest istotną częścią całego procesu analitycznego (O *czym są te badania? Czego chcę się dowiedzieć?*)
- Problem powinien wyczerpywać zakres badania i uwzględniać zależności pomiędzy zmiennymi
- Im więcej interakcji i zmiennych, tym bardziej skomplikowane wnioskowanie
- Sformułowanie problemu powinna porzedzać faza "desk research"

Metoda statystyczna

Problem – przykłady (obszar edukacji):

1) Dyskryminacja w zatrudnieniu absolwentów

2) Czy koncepcje edukacji wielokulturowej powinny być wdrażane w większym stopniu?

Metoda statystyczna

Problem – przykłady (obszar edukacji):

1) Dyskryminacja w zatrudnieniu absolwentów

2) Czy koncepcje edukacji wielokulturowej powinny być wdrażane w większym stopniu?

3) Nadużywanie narkotyków i alkoholu na kampusach uniwersyteckich

4) Czy uniwersytety kształcą w zgodzie z potrzebami rynku pracy?

Metoda statystyczna

Plan:

- Następnym etapem jest sformułowanie podejścia, które ma największe szanse na rozwiązanie problemu i uzyskanie odpowiedzi
- W przypadku projektów, które mają charakter bardziej eksperymentalny wymaga opracowania szczegółowych kroków
- Produktem etapu jest szczegółowy plan badań zawierający czas, zasoby, zaangażowane osoby, sprzęt, etc.

Metoda statystyczna

Dane:

- Dane pierwotne, dane wtórne, mix
- Dylematy: jakość danych, koszt, uzgodnienia licencyjne, dostępność, kompletność, format, szczegółowość, kwestie etyczne

Metoda statystyczna

Dane:

- Dane pierwotne, dane wtórne, mix
- Dylematy: jakość danych, koszt, uzgodnienia licencyjne, dostępność, kompletność, format, szczegółowość, kwestie etyczne
- Jeżeli dane są nieodpowiednie/niemożliwe do zdobycia --> reformulacja problemu badawczego
- Nie ma idealnego zbioru danych

Metoda statystyczna

Dane:

- Dodatkowo błędy:
  - błąd I typu (wynik fałszywie pozytywny)
  - błąd II typu (wynik fałszywie negatywny)
  - błąd systematyczny
  - błąd przypadkowy

Metoda statystyczna

Analiza:

- Jest zazwyczaj czynnością wieloetapową
- Zaczyna się od przeglądu i przekształcania danych, aby otrzymać spójny zbiór (harmonizacja i transformacja)

Metoda statystyczna

Analiza:

- Jest zazwyczaj czynnością wieloetapową
- Zaczyna się od przeglądu i przekształcania danych, aby otrzymać spójny zbiór (harmonizacja i transformacja)
- Kolejne kroki to np.: analiza opisowa, eksploracja danych, modelowanie statystyczne
- Należy unikać stosowania pojedynczej techniki analitycznej

Metoda statystyczna

Analiza:

- Jest zazwyczaj czynnością wieloetapową
- Zaczyna się od przeglądu i przekształcania danych, aby otrzymać spójny zbiór (harmonizacja i transformacja)
- Kolejne kroki to np.: analiza opisowa, eksploracja danych, modelowanie statystyczne
- Należy unikać stosowania pojedynczej techniki analitycznej

"It is as well to remember the following truths about models: all models are wrong; some models are better than others; the correct model can never be known with certainty; and the simpler a model the better it is"!

George Box

Metoda statystyczna

Konkluzje (rezultaty):

- Etap ma na celu opracowanie wniosków w "języku problemu" w celu ich upowszechnienia
- Powinny zawierać ścisłe podsumowanie badań i prezentację graficzną
- Nie powinny zawierać szczegółowych detali technicznych
- Wskazują mocne i słabe strony badań

Rodzaje badań statystycznych

https://www.youtube.com/watch?v=H5pLC05ycoo&ab_channel=KhanAcademyPoPolsku

Nadużycia, nadinterpretacje, błędy

Nadużycia, nadinterpretacje, błędy

1) Nieadekwatne lub niereprezentatywne dane
2) Myląca wizualizacja rezultatów
3) Błędy we wnioskowaniu
4) Celowe fałszowanie danych

1) Nieadekwatne lub niereprezentatywne dane

- Problem doboru i liczebności próby:

  - zbyt mała liczebność próby

  - niedoreprezentowanie/ nadreprezentowanie warstw

  - wykluczenie pewnych grup społecznych

1) Nieadekwatne lub niereprezentatywne dane

- Problem doboru i liczebności próby:

  - zbyt mała liczebność próby

  - niedoreprezentowanie/ nadreprezentowanie warstw

  - wykluczenie pewnych grup społecznych

  - niepoprawne wykorzystanie technik badawczych

  - efekty czasowe i przestrzenne

  - efekt społecznych oczekiwań (przeszacowanie lub niedoszacowanie)

2) Myląca wizualizacja rezultatów

- Brak skali i opisu osi
- Brak informacji o początku skali (0 lub inna wartość)
- Wybiórcze punkty danych (*cherry picking*)
- Nieodpowiednia forma wizualizacji
- Brak porównywalności pomiędzy wykresami

2) Myląca wizualizacja rezultatów

NEWS Magazine

News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
England
Northern Ireland
Scotland
Wales
UK Politics
Education
Magazine
Business
Health
Science & Environment
Technology
Entertainment
Also in the news
Video and Audio
Programmes
Have Your Say
In Pictures
Country Profiles
Special Reports
Related BBC sites
Sport
Weather
On This Day
Editors' Blog
BBC World Service

Page last updated at 01:40 GMT, Friday, 29 January 2010

E-mail this to a friend Printable version

The bumps in a falling teenage pregnancy rate

GO FIGURE
Different ways of seeing stats

When teenage pregnancy rates in one community fell drastically, it looked like a policy of sex education had paid off. But, as Michael Blastland explains in his regular column, the bumps were still there... you just had to know where to look.

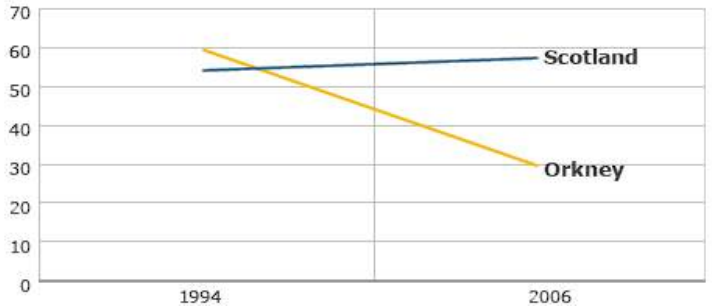
Can stories be bad for you? Let me tell you a true one.

Once upon a time, Orkney had one of the highest teenage pregnancy rates in Scotland. Scotland itself is high in the international league.

So health workers in Orkney tried something new. They began talking to young people about sex in terms of relationships, not only mechanics. They also made condoms easily available because in a small community the shopkeeper might just be your auntie.

Then came data showing that Orkney's teenage pregnancy rate had dramatically halved.

Teenage pregnancies
Per thousand women



COMPLETE MICHAEL BLASTLAND ARCHIVE
Measuring niceness GO

In today's Magazine

Big beasts
How elephants helped to shape human history, by David Cannadine

Change a-coming
Justin Webb on America's love affair with progress

Audience of one
Would you watch a play all on your own?

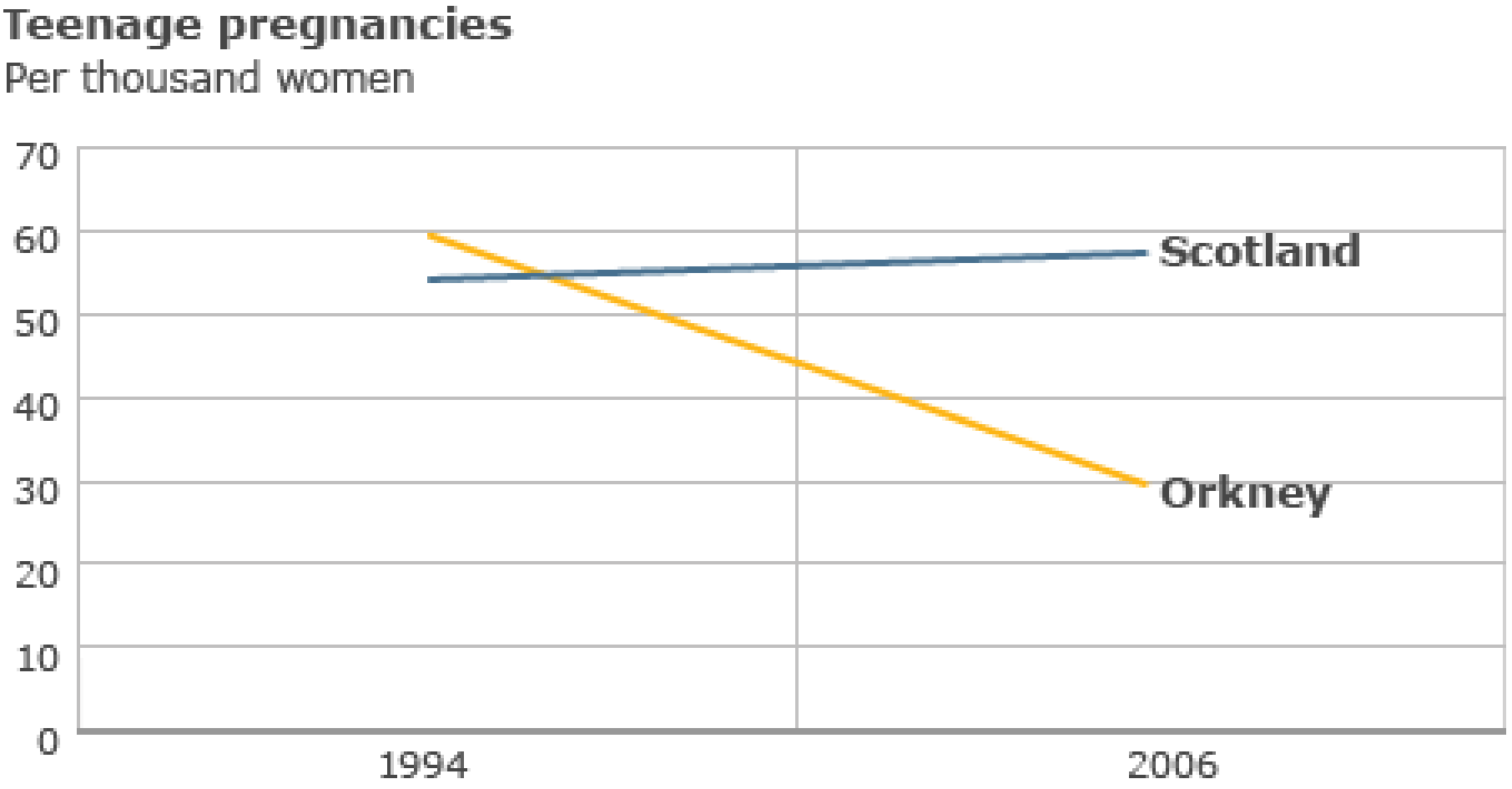
7 days quiz
What now for Paul the eight-limbed oracle?

Magazine regulars

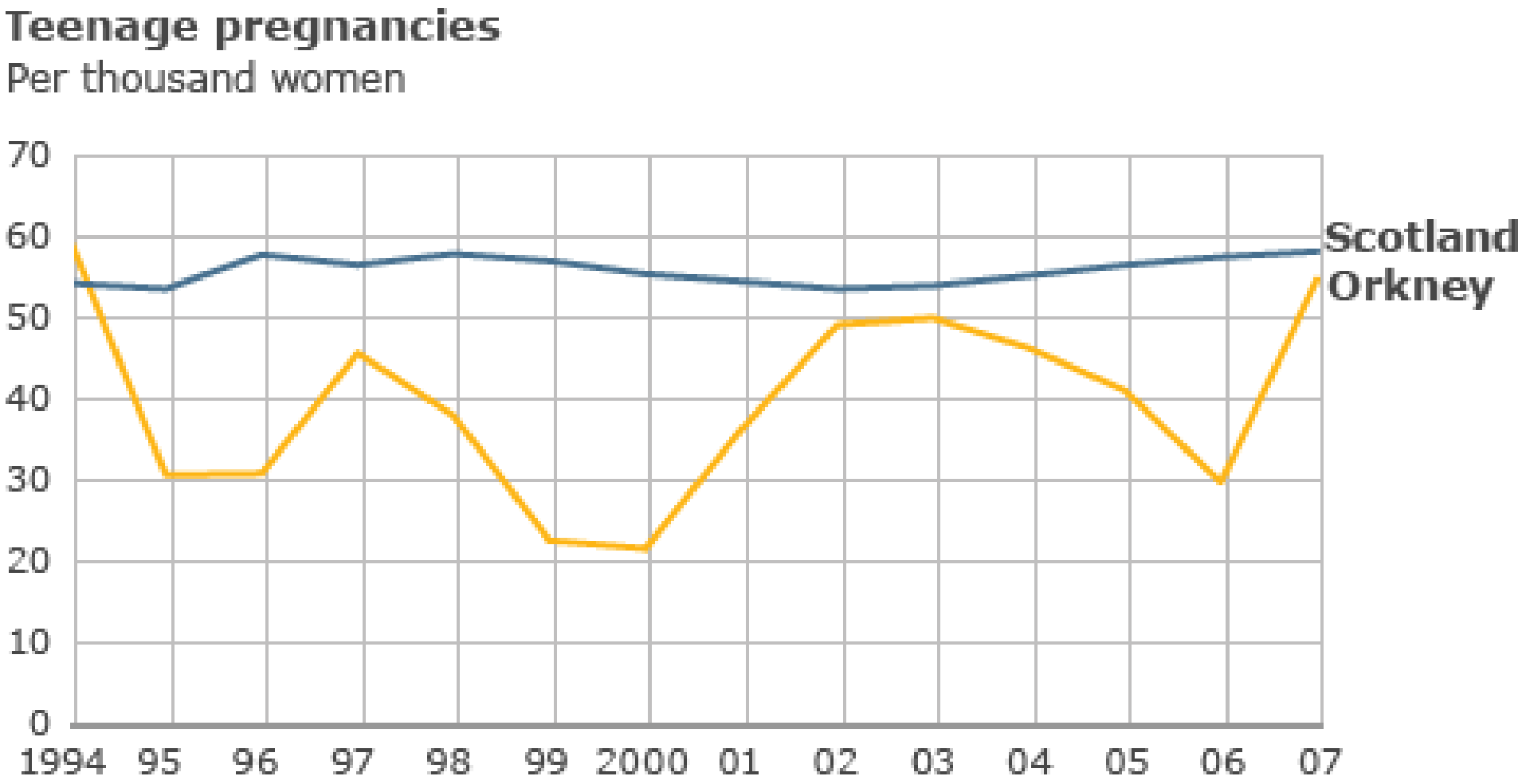
Tweetbook
Say goodbye to worktime boredom. Follow us on Facebook or Twitter

Magazine Monitor
Paper Monitor, Your Letters, Quote of the Day, Caption Competition and more

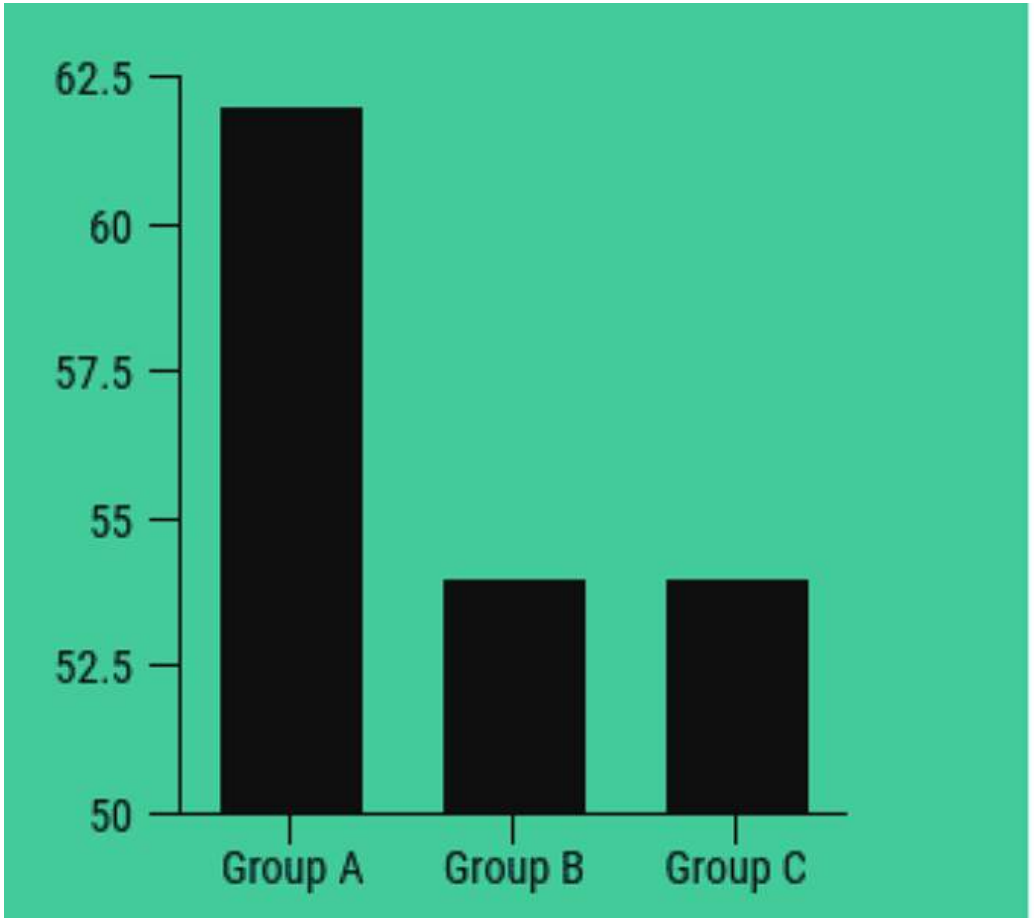
2) Myląca wizualizacja rezultatów



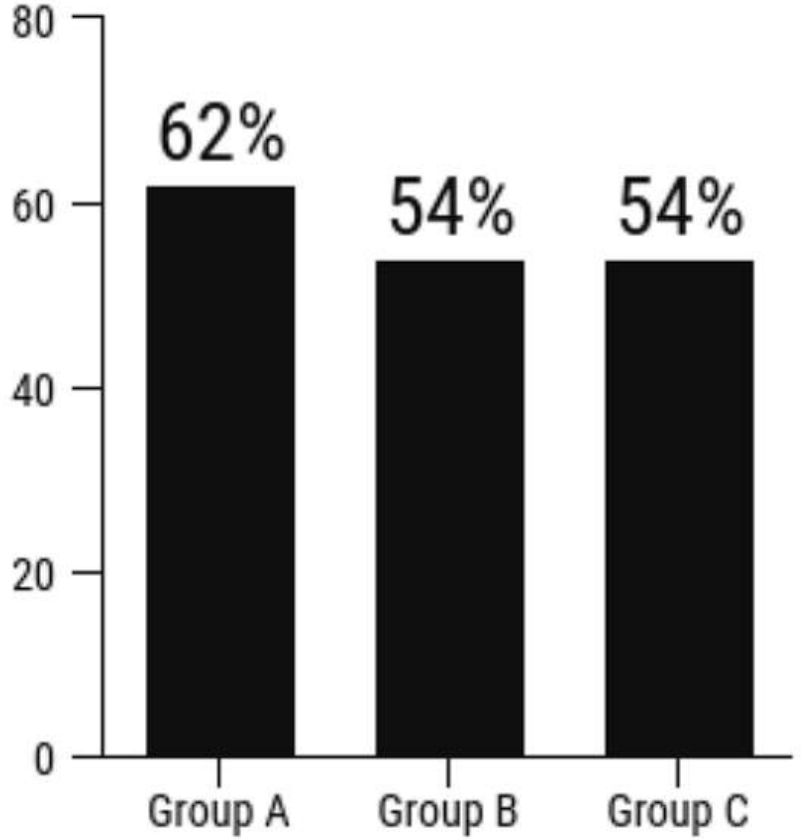
2) Myląca wizualizacja rezultatów



2) Myląca wizualizacja rezultatów

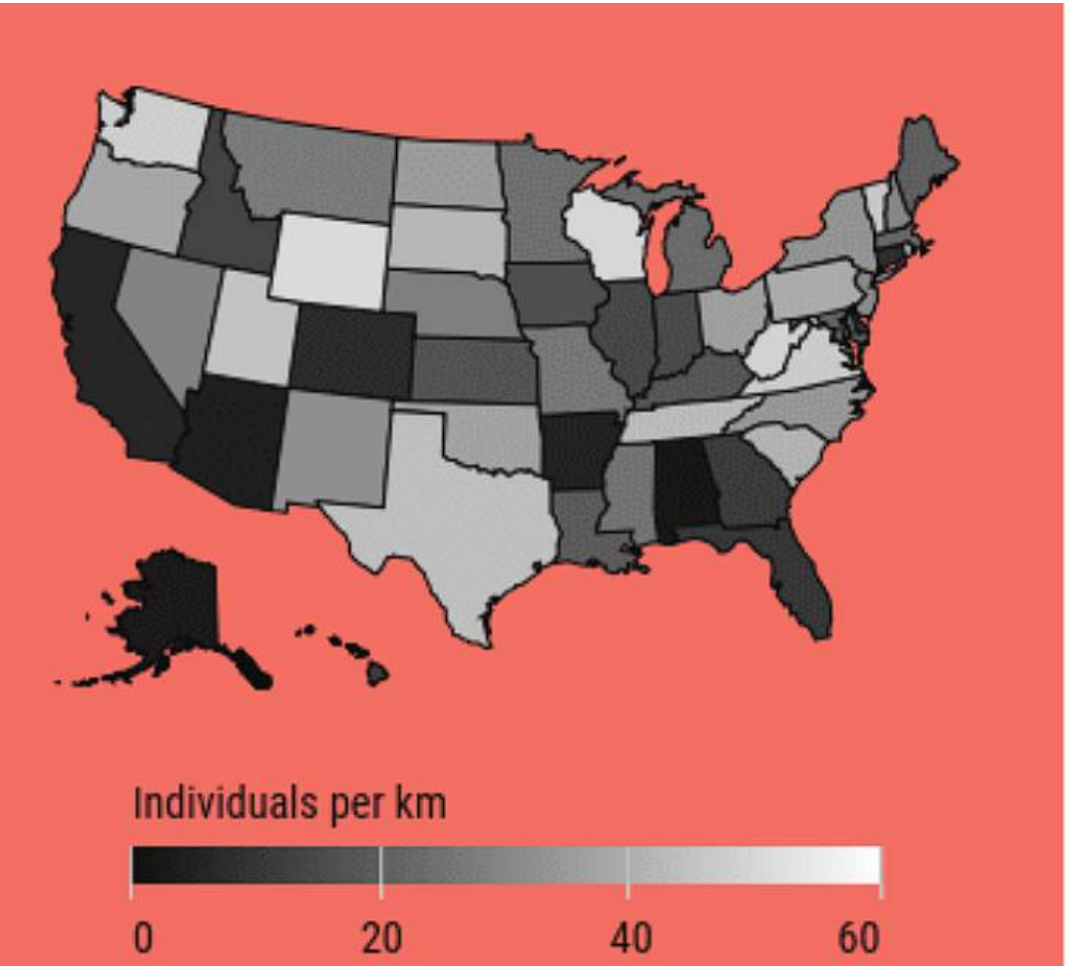


Źle

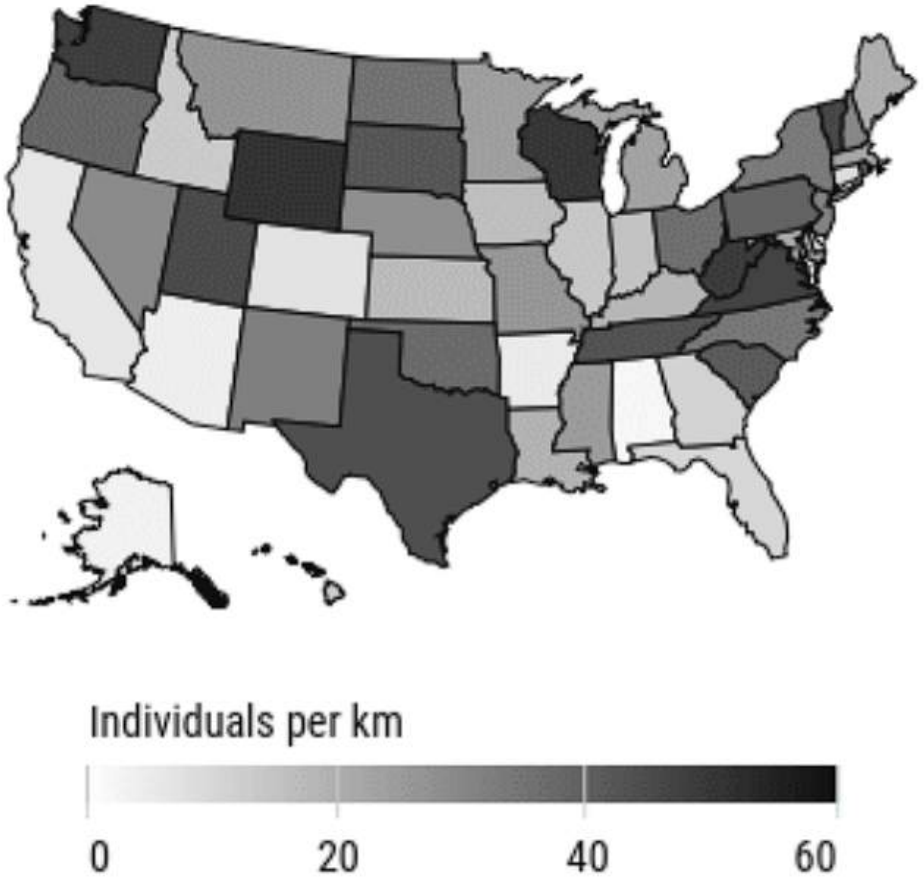


Dobrze

2) Myląca wizualizacja rezultatów



Źle



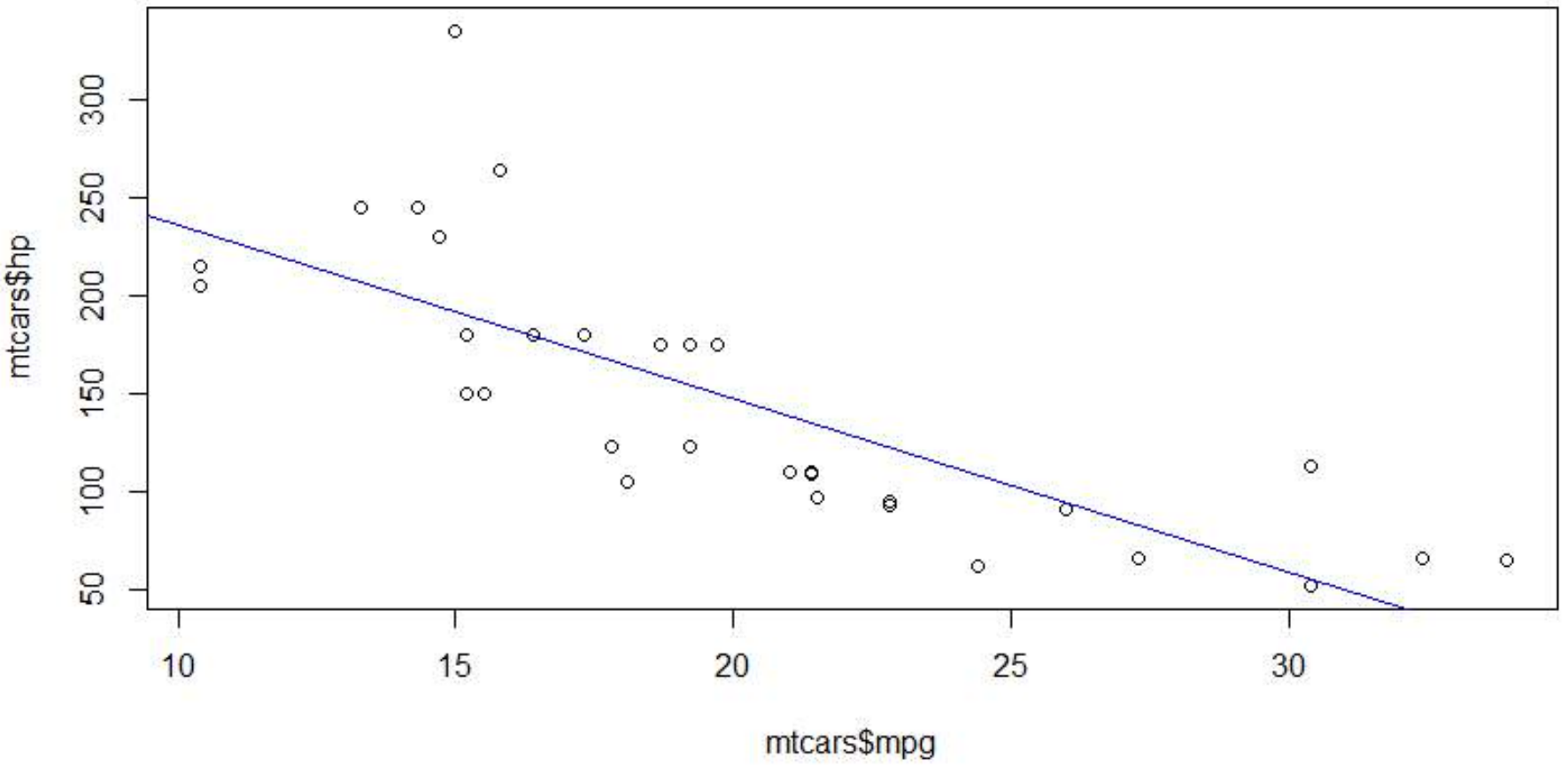
Dobrze

3) Nieadekwatne wnioskowanie

- Korelacja *versus* wywoływanie
- Niezrozumienie losowości i prawdopodobieństwa zajścia zdarzeń
- Błąd atomistyczny
- Błędne wnioskowanie z wizualizacji
- Wcześniej wspomniane błędy związane z danymi

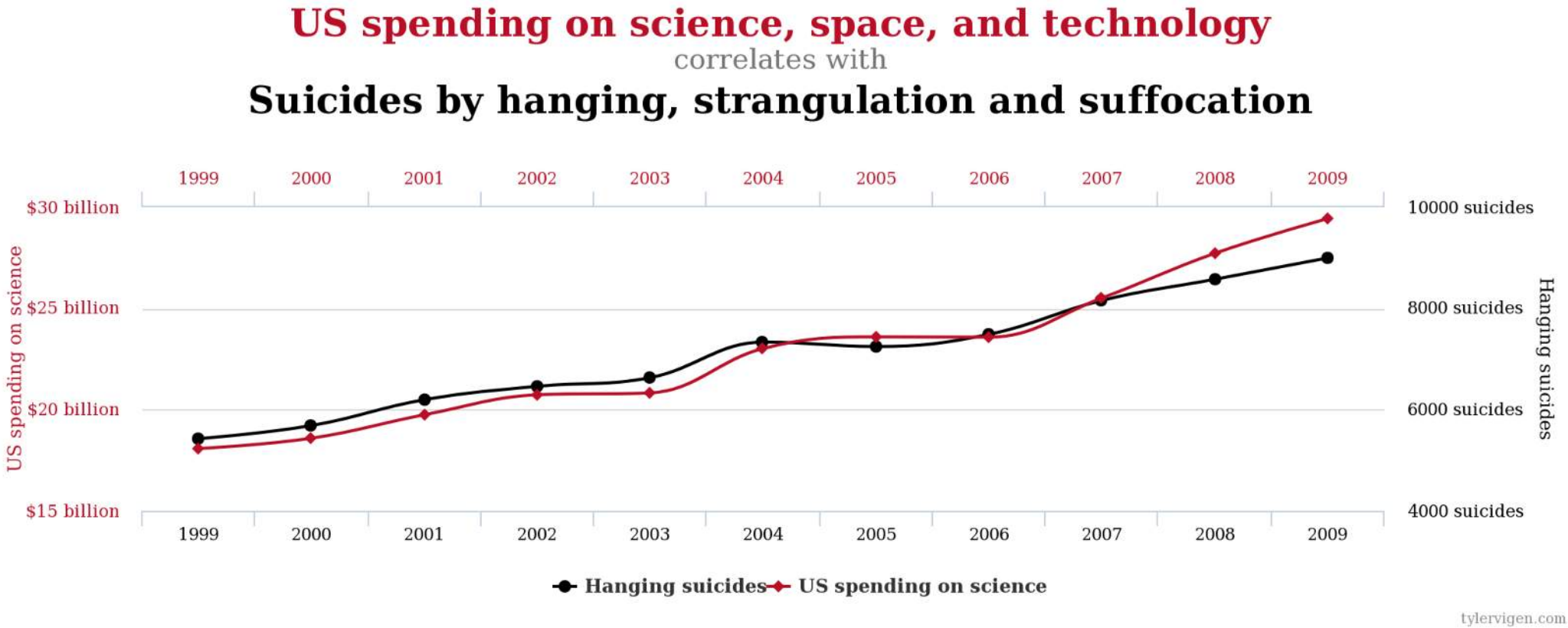
3) Nieadekwatne wnioskowanie

Korelacja *versus* wywoływanie



3) Nieadekwatne wnioskowanie

Korelacja *versus* wywoływanie



4) Celowe fałszowanie danych

10/26/2009, 00.00 | SOUTH KOREA

Send to a friend

Sentenced for fraud Hwang Woo-suk, pioneer of (false) "human cloning"

56 year old scientist also charged with embezzlement and breach of laws on bioethics. He falsified stem cell research, claiming to have cloned cells from healthy patients. The prosecutor asks for four years in prison; sentence due in the next few hours.

Seoul (AsiaNews / Agencies) - A court in Seoul sentenced for fraud the controversial South Korean scientist Hwang Woo-suk, famous for experiments on stem cells and human cloning. He is also charged with fraud, embezzlement and violation of laws on bioethics. He was celebrated and revered as a national hero for having led South Korea at the forefront of scientific research; revelations about the falsification of his experiments shocked the entire nation.

Pomiar (zebranie) danych

Pomiar (zebranie) danych

Typy zmiennych:

- **stymulanty** - zmienne, których wysokie wartości są pożądane
- **destymulanty** - zmienne, których wysokie wartości są niepożądane
- **nominanty** - zmienne, których odchylenia od poziomu najkorzystniejszego (optymalnego poziomu nasycenia) są niepożądane

Pomiar (zebranie) danych

- **Populacja celowa** – to czym się będziemy zajmować jako całość
- **Populacja badana** (operat losowania) – formalny zbiór jednostek, który potencjelnie możemy zbadać
- **Próba badawcza** – grupa jednostek wylosowanych do badania

Losowe schematy doboru prób badawczych

- Opierają się na losowości i wykorzystaniu rachunku prawdopodobieństwa, aby zmniejszyć ryzyko błędu
- Wykorzystywane są generatory liczb (pseudo)losowych dostępne w pakietach do obliczeń statystycznych
- Najczęściej wykorzystywane w badaniach ilościowych

Losowe schematy doboru prób badawczych

- Dobór losowy prosty
- Dobór losowy warstwowy
- Dobór losowy systematyczny
- Dobór zespołowy

Nielosowe schematy doboru prób badawczych

- Nie wykorzystuje się losowania i rachunku prawdopodobieństwa
- Badacz sam dokonuje wyboru konkretnych jednostek do badania
- Najczęściej wykorzystywane w badaniach jakościowych

Nielosowe schematy doboru prób badawczych

- Dobór celowy
- Dobór kwotowy
- Dobór oparty na dostępności respondentów
- Dobór metodą kuli śnieżnej

Wielkość próby badawczej

- Zarówno zbyt mała, jak i zbyt duża próba badawcza niesie ze sobą określone problemy
- Przyjmuje się zazwyczaj założenie o 95% **poziomie ufności** i przedziale **błędu losowego** +/- 3%
- Wielkość próby badawczej zależy od np. konkretnego problemu, rozproszenia danych, dostępnego czasu, *response rate*

Eksploracja danych

- Czyszczenie, transformacja i skalowanie zmiennych:
  - Transformacja logarytmiczna
  - Normalizacja
  - Standaryzacja
- Metody grupowania (klasteryzacji):
  - metody niehierachiczne (*k-means*)
  - metody hierarchiczne
  - metody oparte na gęstości (*DBscan*)
- Uczenie maszynowe (*ML*)

Eksploracja danych

'Water, water everywhere, nor any drop to drink'

The Rime of the Ancient Mariner

Eksploracja danych

'Water, water everywhere, nor any drop to drink'

The Rime of the Ancient Mariner

- *Data mining; digging into data*
- (Kolejny) etap pracy z danymi mający na celu wykorzystanie mocy obliczeniowej komputera do odkrycia wzorców, prawidłowości, trendów, schematów w zbiorze danych

Eksploracja danych

'Water, water everywhere, nor any drop to drink'

The Rime of the Ancient Mariner

- *Data mining; digging into data*
- (Kolejny) etap pracy z danymi mający na celu wykorzystanie mocy obliczeniowej komputera do odkrycia wzorców, prawidłowości, trendów, schematów w zbiorze danych
- Ukierunkowuje dalsze badania lub jest ich efektem finalnym
- Najczęściej opiera się na konkretnych algorytmach implementowanych w pakietach statystycznych

Eksploracja danych - problemy

- w dużych bazach danych mogą zostać odkryte tysiące reguł/wzorców
- różni użytkownicy są zainteresowani różnymi typami prawidłowości

Eksploracja danych - problemy

- w dużych bazach danych mogą zostać odkryte tysiące reguł/wzorców
- różni użytkownicy są zainteresowani różnymi typami prawidłowości
- trudności w zrozumieniu tych wielowarstwowych powiązań
- algorytmy eksploracji danych są często intensywne obliczeniowo → *Amazon Web Services; Microsoft Azure; Nvidia HPC, Google Cloud*

Eksploracja danych

- Czyszczenie danych,
- Transformacja i skalowanie zmiennych:
  - Transformacja logarytmiczna
  - Normalizacja
  - Standaryzacja

Czyszczenie danych

1) Inspekcja: Wykrycie nieoczekiwanych, nieprawidłowych i niespójnych danych
(obserwacje odstające, duplikaty, brakujące wartości)

2) Czyszczenie: Naprawa lub usuwanie wykrytych anomali

3) Weryfikacja: Po oczyszczeniu dane są ponowanie kontrolowane w celu sprawdzenia poprawności.

Transformacja i skalowanie zmiennych

- Poprawa interpretowalności danych
- Ujednolicenie skal
- Uporządkowanie prezentacji graficznej
- Głębszy wgląd w dane
- Spełnienie założeń metod eksploracji danych/ wnioskowania statystycznego

Transformacja logarytmiczna

- Dane często są mocno skrzywione, lub skoncentrowane wokół jednej wartości
- Wnioskowanie z takich danych jest utrudnione/niemożliwe
- Transformacja logarytmiczna ma na celu upodobnienie takiego zbioru do rozkładu normalnego
- $x \rightarrow log(x)$

Transformacja logarytmiczna

- Przykład

Normalizacja zmiennych

- Ma na celu przekształcenie zmiennych tak, aby zawierały się one w przedziale 0 – 1 (porównywalność)
- Zmniejsza złożoność danych
- Redukuje anomalie w zbiorze
- Normalizacja *min – max*:

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

Normalizacja zmiennych

- Przykład

Standaryzacja zmiennych

- Przekształcenie zmiennych tak, aby miały średnią = 0 i odchylenie standardowe = 1
- Podobnie jak normalizacja *min-max,* standaryzacja ma na celu ułatwienie porównywania zmiennych o różnych skalach numerycznych

$$X_{stand} = \frac{X - \bar{X}}{\sigma}$$

Standaryzacja zmiennych

- Przykład

Eksploracja danych

- Czyszczenie, transformacja i skalowanie zmiennych:
  - Transformacja logarytmiczna
  - Normalizacja
  - Standaryzacja
- Metody klasteryzacji (grupowania):
  - metody niehierachiczne (*k-means*)
  - metody hierarchiczne
  - metody oparte na gęstości (*DBSCAN*)
- Uczenie maszynowe (ML)

Metody klasteryzacji (grupowania)

- Wykorzystywane w celu identyfikacji grup skupiających podobne do siebie jednostki
- Składa się na nie wiele algorytmów różniących się zarówno sposobem wykrywania grup jak i różnicami w ich definicji

Metody klasteryzacji (grupowania)

- Wykorzystywane w celu identyfikacji grup skupiających podobne do siebie jednostki
- Składa się na nie wiele algorytmów różniących się zarówno sposobem wykrywania grup jak i różnicami w ich definicji
- Algorytmy grupowania zmiennych dzielimy na:
  - niehierarchiczne (oparte na centroidach, np. *kmeans*),
  - hierarchiczne (np. *drzewo klasyfikacyjne*),
  - oparte na gęstości (łączą obszary o wysokiej gęstości, np. *DBscan*)

Algorytm k-średnich (*k-means*)

- Prace rozpoczęły się w latach 50-tych XX w. i obejmowały różnych badaczy i dyscypliny: Steinhous (1956), Lloyd (1957), Jancey (1966)
- Klasyczny algorytm *k-means* został po raz pierwszy opisany przez MacQeena (1967).

NOKIA BELL LABS

SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS

J. MacQUEEN
UNIVERSITY OF CALIFORNIA, LOS ANGELES

1. Introduction

The main purpose of this paper is to describe a process for partitioning an N-dimensional population into k sets on the basis of a sample. The process, which is called 'k-means,' appears to give partitions which are reasonably efficient in the sense of within-class variance. That is, if p is the probability mass

Algorytm k-średnich (k-means)

- Mając ustaloną liczbę skupień (*k*), przyporządkowuje obserwacje do klastrów tak, aby średnie w klastrach były jak najbardziej różne od siebie.
- Różnice między obserwacjami są mierzone w kategoriach jednej z kilku miar odległości (np. *euklidesową, Chebysheva, Manhattan*)

Algorytm k-średnich (k-means)

1) Określenie liczby klastrów (*k*) do utworzenia

2) Wybierz losowo k obiektów z zestawu danych jako początkowe centra klastrów

3) Przypisz każdą obserwację do najbliższego centroida, w oparciu o odległość euklidesową

4) Dla każdego z *k* klastrów aktualizuj centroid poprzez obliczenie nowych wartości średnich dla punktów w klastrze.

5) Iteruj kroki 3 i 4 do momentu, gdy przypisania klastrów przestaną się zmieniać

Algorytm k-średnich

https://www.youtube.com/watch?v=5I3Ei69I40s

Algorytm k-średnich (k-means)

Problemy:

- Wymaga wybrania z góry odpowiedniej liczby klastrów
- Uzyskane wyniki końcowe są wrażliwe na początkowy losowy wybór centrów klastrów

Algorytm k-średnich (k-means)

Problemy:

- Wymaga wybrania z góry odpowiedniej liczby klastrów
- Uzyskane wyniki końcowe są wrażliwe na początkowy losowy wybór centrów klastrów
- Jest wrażliwy na wartości odstające.
- Działa dobrze dla wyraźnie odseparowanych klastrów
- Zmiana kolejności danych może prowadzić do innych wyników klasteryzacji

Algorytm k-średnich (k-means)

- Przykład

Metody hierarchiczne (drzewo klasyfikacyjne)

- Sibson (1973), Defays (1977) i Rohlf (1982) jako „prehistoria” rodziny metod
- Efektem działania są obiekty pogrupowane w klastry według ich hierarchii.

Volume 16, Issue 1
1973

Article Contents

JOURNAL ARTICLE

SLINK: An optimally efficient algorithm for the single-link cluster method

R. Sibson

The Computer Journal, Volume 16, Issue 1, 1973, Pages 30–34,
https://doi.org/10.1093/comjnl/16.1.30
Published: 01 January 1973

SLINK : An optimally efficient algorithm for the single-link cluster method

R. Sibson

King's College Research Centre, King's College, Cambridge, and Cambridge University Statistical Laboratory

The SLINK algorithm carries out single-link (nearest-neighbour) cluster analysis on an arbitrary dissimilarity coefficient and provides a representation of the resultant dendrogram which can readily be converted into the usual tree-diagram. The algorithm achieves the theoretical order-of-magnitude bounds for both compactness of storage and speed of operation, and makes the application of the single-link method feasible for a number of OTU's well into the range 10^3 to 10^4. The algorithm is easily programmable in a variety of languages including FORTRAN.

(Received January 1972)

Metody hierarchiczne (drzewo klasyfikacyjne)

- Sibson (1973), Defays (1977) i Rohlf (1982) jako „prehistoria” rodziny metod
- Efektem działania są obiekty pogrupowane w klastry według ich hierarchii.
- Nie wymaga wcześniejszego określenia liczby skupień; wymaga jednak wskazania metody obliczania podobieństwa pomiędzy obserwacjami

Metody hierarchiczne (drzewo klasyfikacyjne)

- Sibson (1973), Defays (1977) i Rohlf (1982) jako „prehistoria” rodziny metod
- Efektem działania są obiekty pogrupowane w klastry według ich hierarchii.
- Nie wymaga wcześniejszego określenia liczby skupień; wymaga jednak wskazania metody obliczania podobieństwa pomiędzy obserwacjami
- Algorytm tworzy drzewopodobny obiekt graficzny o nazwie ***dendrogram***
- Poprzez odcinanie gałęzi dendrogramu użytkownik formuje porządaną liczbę grup

Drzewo klasyfikacyjne - algorytm

1) Umieść każdy punkt danych w jego własnym klastrze.

2) Zidentyfikuj najbliższe (najpodobniejsze) dwa klastry i połącz je w jeden klaster.

3) Powtarzaj krok 2, aż wszystkie punkty danych znajdą się w jednym klastrze.

Drzewo klasyfikacyjne

Problemy:

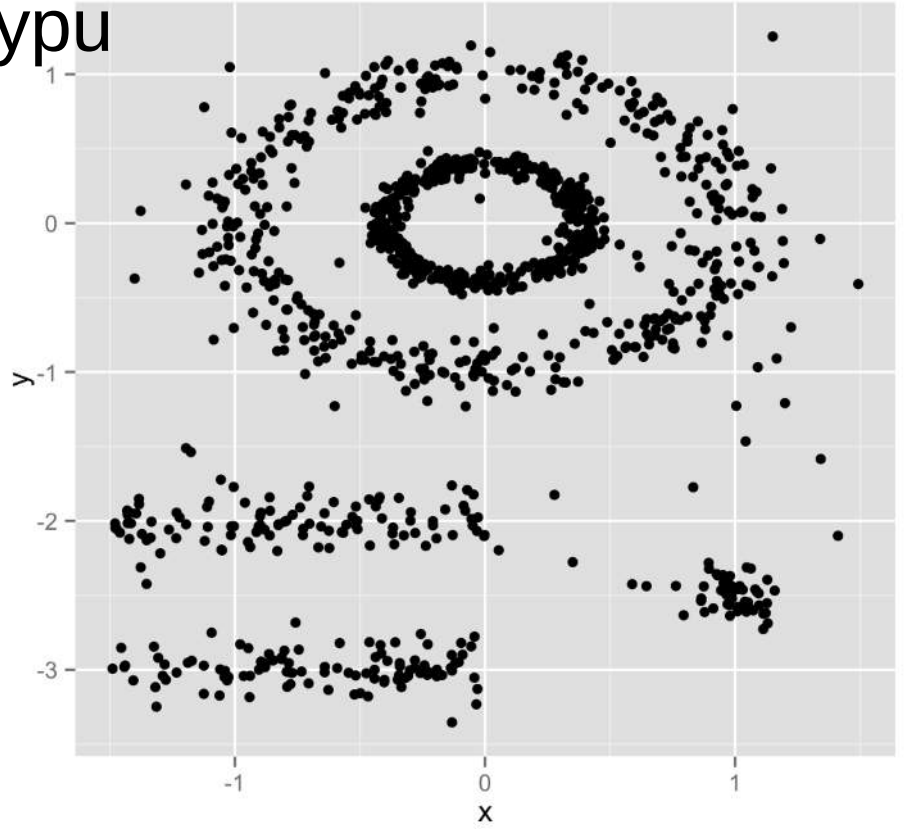
- Określenie miary podobieństwa/ niepodobieństwa pomiędzy obserwacjami
- Określenie tzw. „miejsc odcięcia” dendrogramu
- Wrażliwy na obserwacje odstające
- Działa dobrze dla wyraźnie odesparowanych klastrów

Drzewo klasyfikacyjne - algorytm

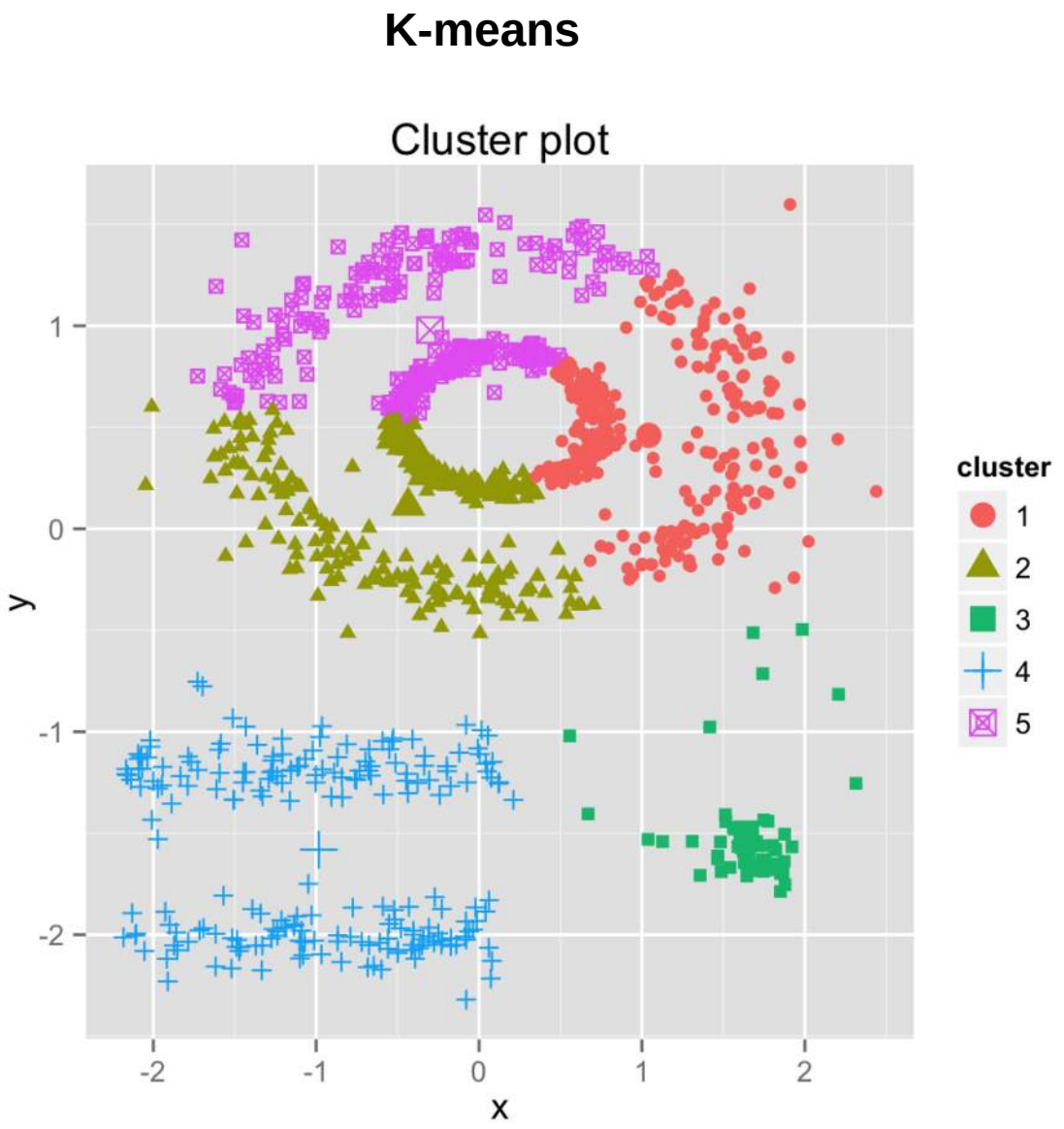
Przykład

Klasyfikcja oparta na gęstości - dbscan

- W praktyce dane (zwłaszcza przestrzenne) znajdują się często w z góry określonych grupach i zawierają tzw. szum oraz wartości odstające
- Klasyczne algorytmy miałyby problem w klasyfikacji danych tego typu



Klasyfikcja oparta na gęstości - dbscan



Klasyfikcja oparta na gęstości - dbscan

- ***D****ensity-****B****ased* ***S****patial* ***C****lustering and* ***A****pplication with* ***N****oise*
- DBSCAN (Ester et al. 1996) jest odporny na wskazane problemy
- Nie wymaga wskazania liczby skupień

A Density-Based Algorithm for Discovering Clusters

A Density-Based Algorithm for Discovering Clusters

in Large Spatial Databases with Noise

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu

Institute for Computer Science, University of Munich
Oettingenstr. 67, D-80538 München, Germany
{ester | kriegel | sander | xwxu}@informatik.uni-muenchen.de

Abstract

Clustering algorithms are attractive for the task of class identification in spatial databases. However, the application to large spatial databases rises the following requirements for clustering algorithms: minimal requirements of domain knowledge to determine the input parameters, discovery of clusters with arbitrary shape and good efficiency on large databases. The well-known clustering algorithms offer no solution to the combination of these requirements. In this paper, we present the new clustering algorithm DBSCAN relying on a density-based notion of clusters which is designed to discover clusters of arbitrary shape. DBSCAN requires only one input parameter and supports the user in determining an appropriate value for it. We performed an experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data and real data of the SEQUOIA 2000 benchmark. The results of our experiments demonstrate that (1) DBSCAN is significantly more effective in discovering clusters of arbitrary shape than the well-known algorithm CLARANS, and that (2) DBSCAN outperforms CLARANS by a factor of more than 100 in terms of efficiency.

Keywords: Clustering Algorithms, Arbitrary Shape of Clusters, Efficiency on Large Spatial Databases, Handling Nlj4-275oise.

1. Introduction

are often not known in advance when dealing with large databases.

(2) Discovery of clusters with arbitrary shape, because the shape of clusters in spatial databases may be spherical, drawn-out, linear, elongated etc.

(3) Good efficiency on large databases, i.e. on databases of significantly more than just a few thousand objects.

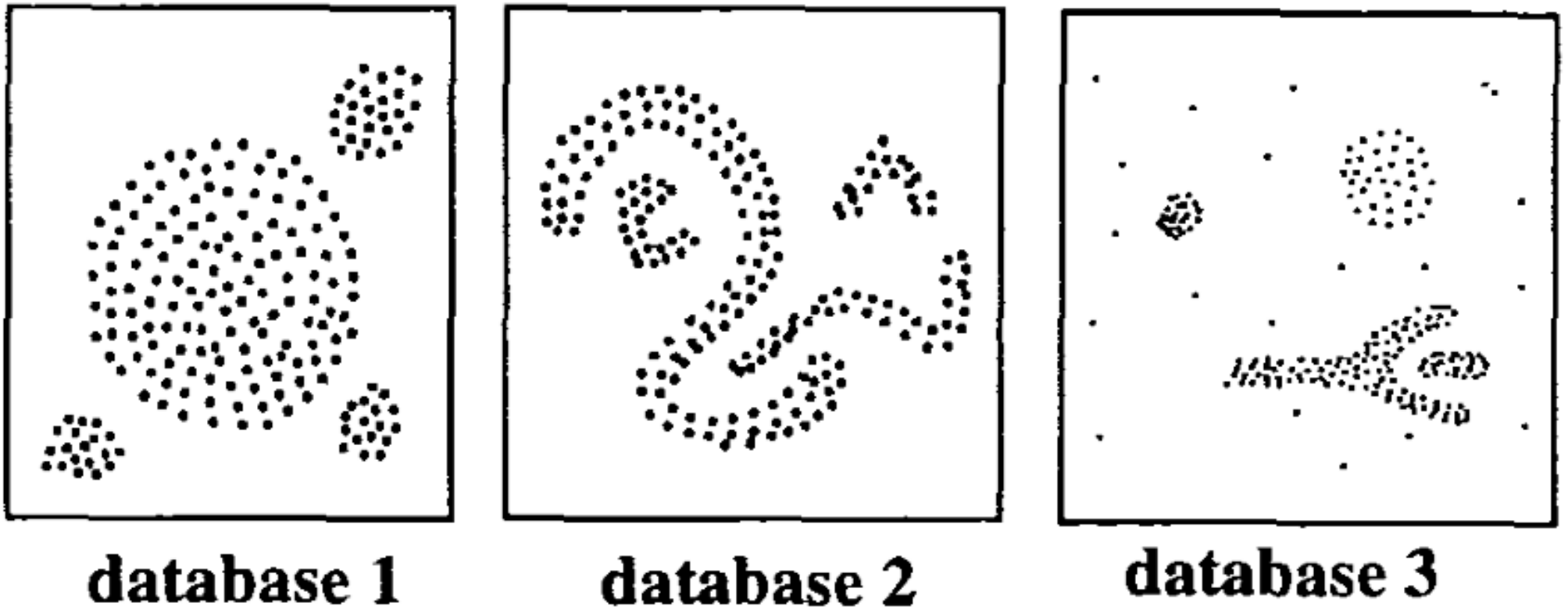
The well-known clustering algorithms offer no solution to the combination of these requirements. In this paper, we present the new clustering algorithm DBSCAN. It requires only one input parameter and supports the user in determining an appropriate value for it. It discovers clusters of arbitrary shape. Finally, DBSCAN is efficient even for large spatial databases. The rest of the paper is organized as follows. We discuss clustering algorithms in section 2 evaluating them according to the above requirements. In section 3, we present our notion of clusters which is based on the concept of density in the database. Section 4 introduces the algorithm DBSCAN which discovers such clusters in a spatial database. In section 5, we performed an experimental evaluation of the effectiveness and efficiency of DBSCAN using synthetic data and data of the SEQUOIA 2000 benchmark. Section 6 concludes with a summary and some directions for future research.

Klasyfikcja oparta na gęstości - dbscan

- *Density-Based Spatial Clustering and Application with Noise*
- DBSCAN (Ester et al. 1996) jest odporny na wskazane problemy
- Nie wymaga wskazania liczby skupień
- Wykrywa klastry o dowolnym kształcie
- Podstawowa idea wywodzi się z intuicyjnej dla człowieka metody klastrowania
- Zastosowanie głównie do danych przestrzennych (GIS)

Klasyfikcja oparta na gęstości - dbscan

Klastry to obszary o zwiększonej gęstości, oddzielone obszarami o małej gęstości



Ester et al, 1996

Klasyfikcja oparta na gęstości - dbscan

- Algorytm wymaga wskazania dwóch parametrów: *epsilon i minimum points*

  1) Dla każdego punktu x oblicz odległość między x a innymi punktami.

  2) Znajdź wszystkie punkty sąsiednie w odległości *epsilon* od punktu początkowego.

Klasyfikcja oparta na gęstości - dbscan

- Algorytm wymaga wskazania dwóch parametrów: *epsilon i minimum points*

  1) Dla każdego punktu x oblicz odległość między x a innymi punktami.

  2) Znajdź wszystkie punkty sąsiednie w odległości *epsilon* od punktu początkowego.

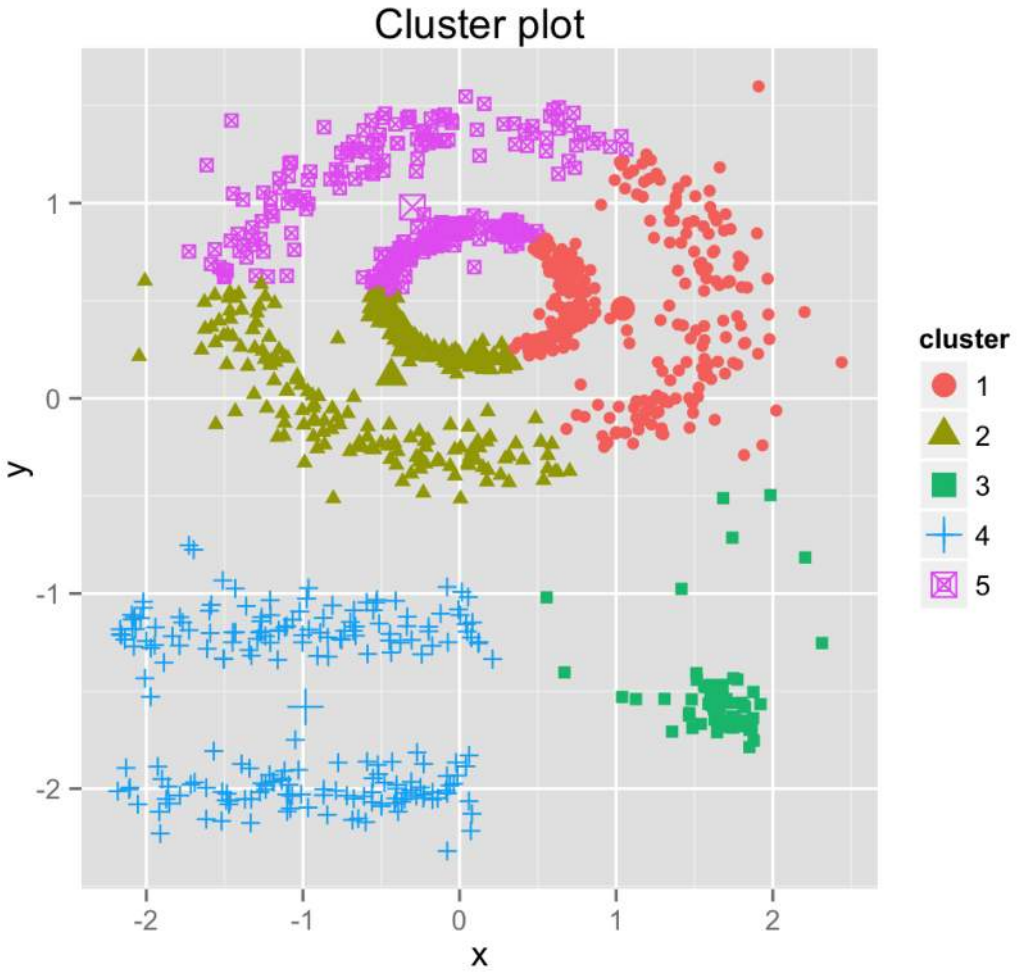
  3) Dla każdego punktu, jeśli nie jest on jeszcze przypisany do klastra, utwórz nowy klaster (jeżeli liczba sąsiadów >= *minimum points*).

  4) Znajdź wszystkie jego gęsto połączone punkty (*epsilon*) i przypisz je do tego samego klastra co punkt główny.

  5) Iteruj przez pozostałe nieodwiedzone punkty w zbiorze danych.

Klasyfikcja oparta na gęstości - dbscan

- Przykład



Eksploracja danych – bardziej zaawansowane metody

- Transformacja i skalowanie zmiennych:
  - Transformacja logarytmiczna
  - Normalizacja
  - Standaryzacja
- Metody grupowania (klasteryzacji):
  - metody niehierachiczne (k-means)
  - metody hierarchiczne
  - metody oparte na gęstości (dbscan)
- **Uczenie maszynowe (ML)**

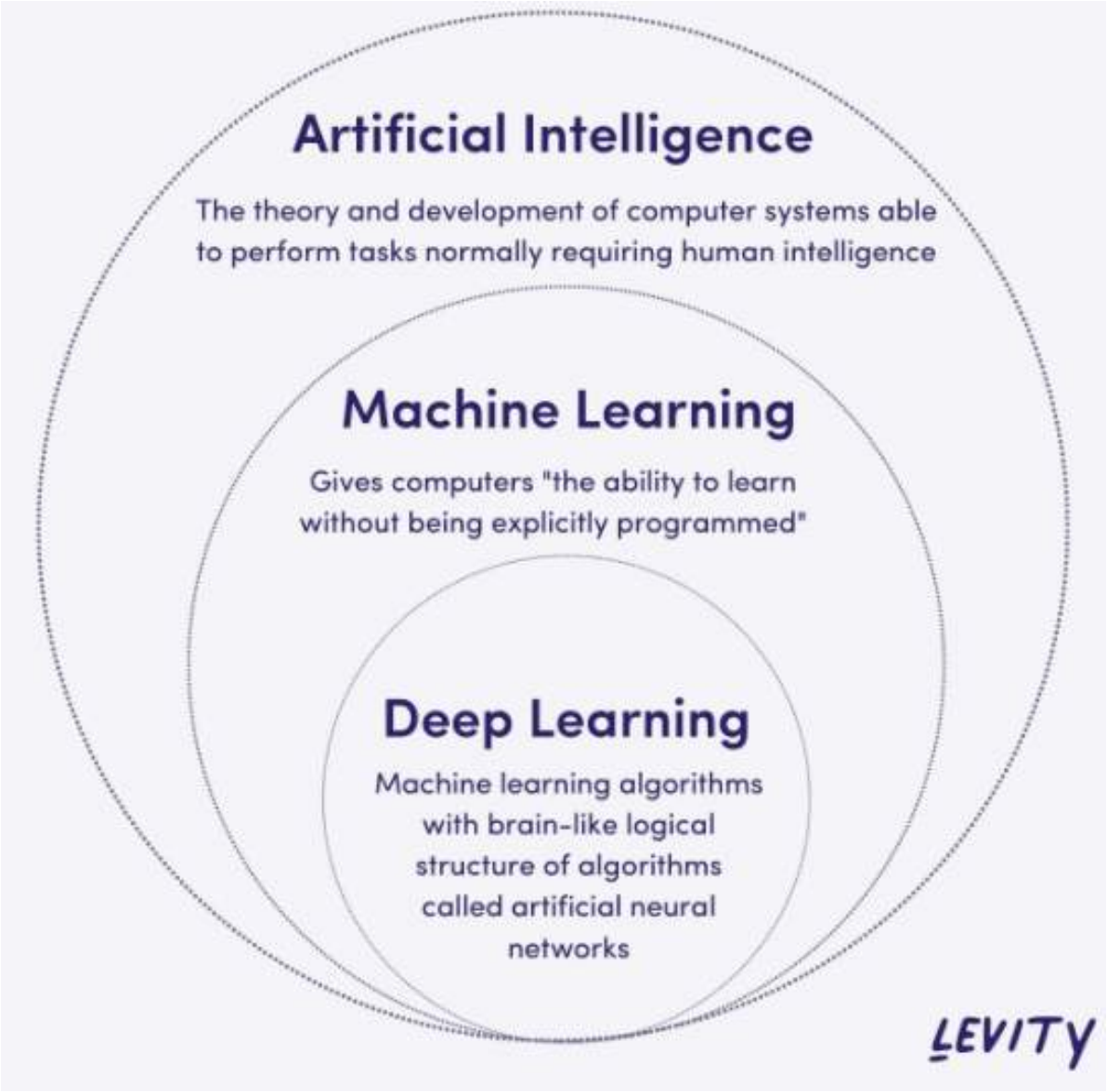
Uczenie maszynowe jako część AI

- Badania rozpoczęły się w Darthmouth College (USA) w 1956 r.
- Pierwsze implementacje obejmowały strategie gry w szachy, rozwiązywanie problemów matematycznych
- W tym czasie naukowcy wierzyli w szybki postęp i opracowanie uogólnionej sztucznej inteligencji (AGI)

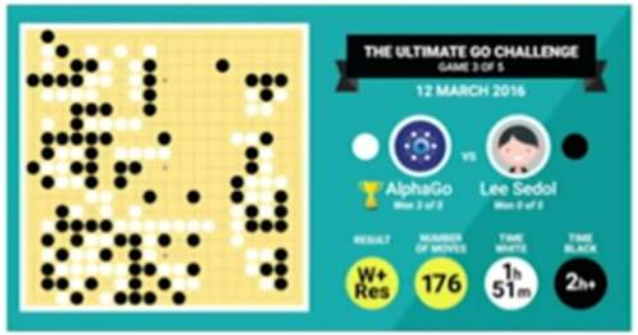
Uczenie maszynowe jako część AI

- Badania rozpoczęły się w Darthmouth College (USA) w 1956 r.
- Pierwsze implementacje obejmowały strategie gry w szachy, rozwiązywanie problemów matematycznych
- W tym czasie naukowcy wierzyli w szybki postęp i opracowanie uogólnionej sztucznej inteligencji (AGI)
- Okres zastoju od lat 70 do końca lat 90 XX w.
- Powolny rozwój nastąpił pod koniec lat 90-tych i na początku XXI w.
- Skokowy wzrost aplikacji od roku 2015
- Zmiana paradygmatu AI (do czasu...)

Sztuczna inteligencja i uczenie maszynowe

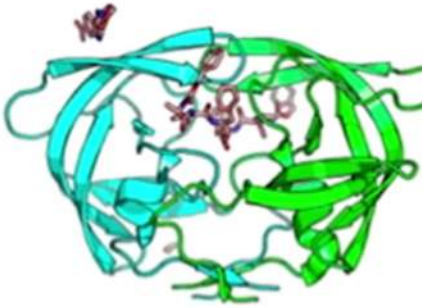


Uczenie maszynowe jest wszędzie

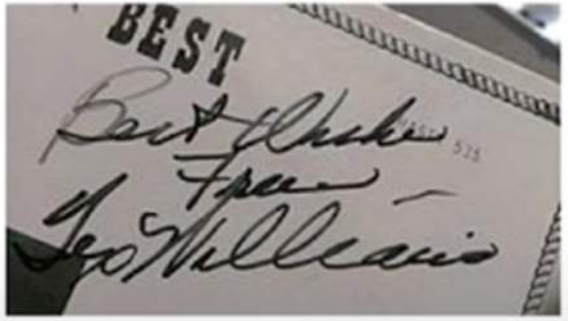


AlphaGo

Recommendation systems



Drug discovery



Character recognition

2σ TWO SIGMA

Hedge fund stock predictions

Voice assistants

Assisted driving

Face detection/recognition

Cancer diagnosis

Uczenie maszynowe

- Dziedzina badań, która skupia się na systemach komputerowych, które mogą uczyć się na podstawie danych.

Uczenie maszynowe

- Dziedzina badań, która skupia się na systemach komputerowych, które mogą uczyć się na podstawie danych.
- Systemy ML (modele) potrafią uczyć się konkretnych zadań na podstawie analizy dużej liczby przykładów, np. model ML może nauczyć się jak rozpoznać samochód na podstawie obserwacji dużej liczb aut.

Uczenie maszynowe

- Brak programowania reguł wprost (przez programistę) – model może nauczyć się rozwiązywać konkretny problem bez predefiniowanych konkretnych reguł
- Model uczy się sam, jakie charakterystyki są istotne, aby rozpoznać dany obiekt

Uczenie maszynowe

- Brak programowania reguł wprost (przez programistę) – model może nauczyć się rozwiązywać konkretny problem bez predefiniowanych konkretnych reguł
- Model uczy się sam, jakie charakterystyki są istotne, aby rozpoznać dany obiekt
- Istotna jest ilość i jakość danych
- Modele ML potrafią wykrywać wzorce, schematy w danych
- ML wspiera podejmowanie decyzji w oparciu o dane (data-driven decisions)

Uczenie maszynowe

- Nienadzorowane (*unsupervised learning*)
- Nadzorowane (*supervised learning*)
- Posiłkowane (*reinforcement learning*)

Uczenie maszynowe

Etapy budowy modelu ML:

1) Zgromadzenie/pozyskanie danych

2) Przygotowanie danych do dalszej analizy (porządkowanie, usuwanie obserwacji odstających)

3) Wybór metody i modelu

4) Trenowanie modelu (w przypadku metod nadzorowanych)

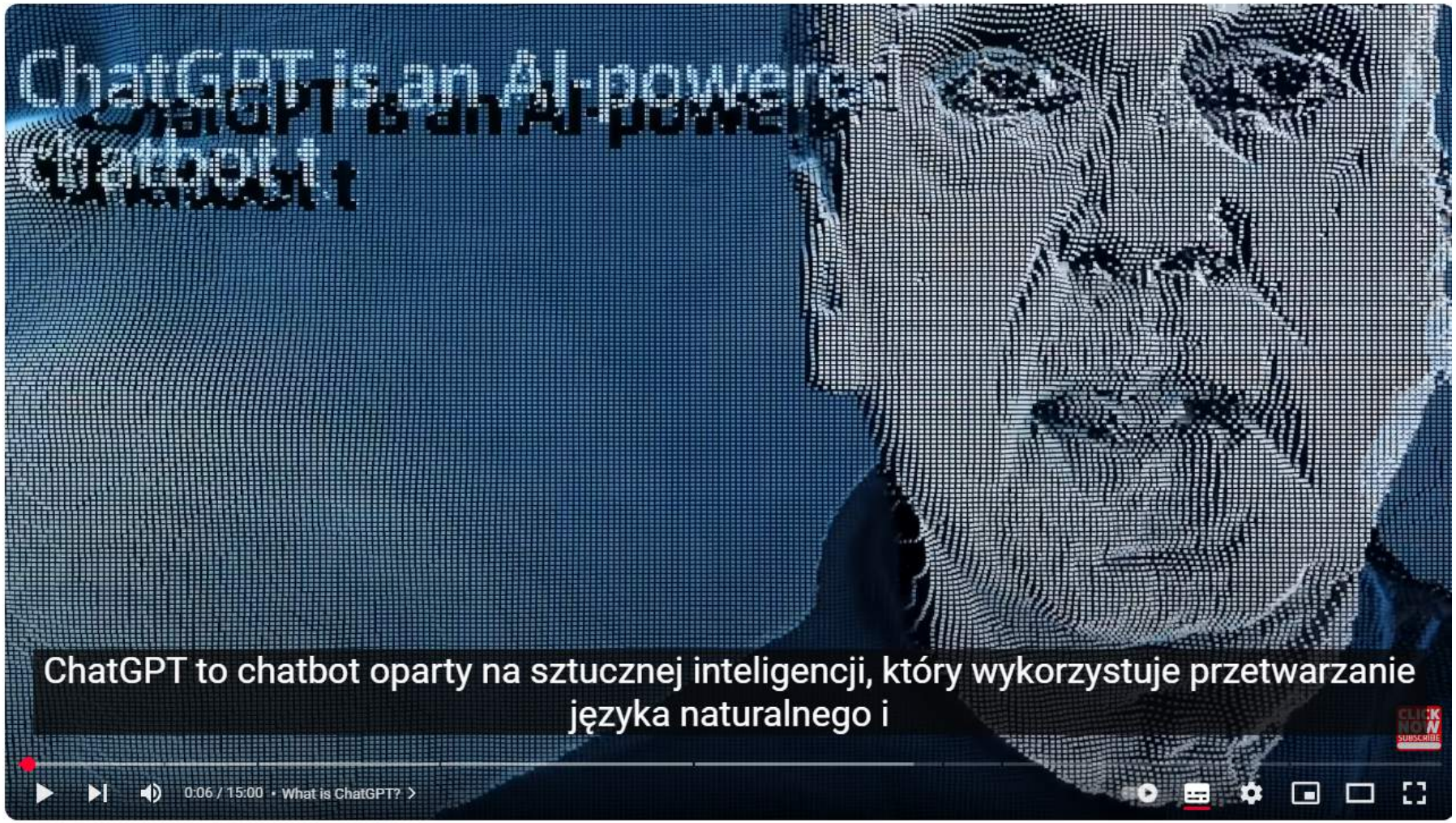
5) Ewaluacja, określenie i pomiar błędów

6) Dopasowanie parametrów

Uczenie maszynowe

https://www.youtube.com/watch?v=nKW8Ndu7Mjw&ab_channel=GoogleCloudTech

How ChatGPT works?



https://www.youtube.com/watch?v=WAiqNav2cRE

Uczenie maszynowe

- Przykład – uczenie nadzorowane:

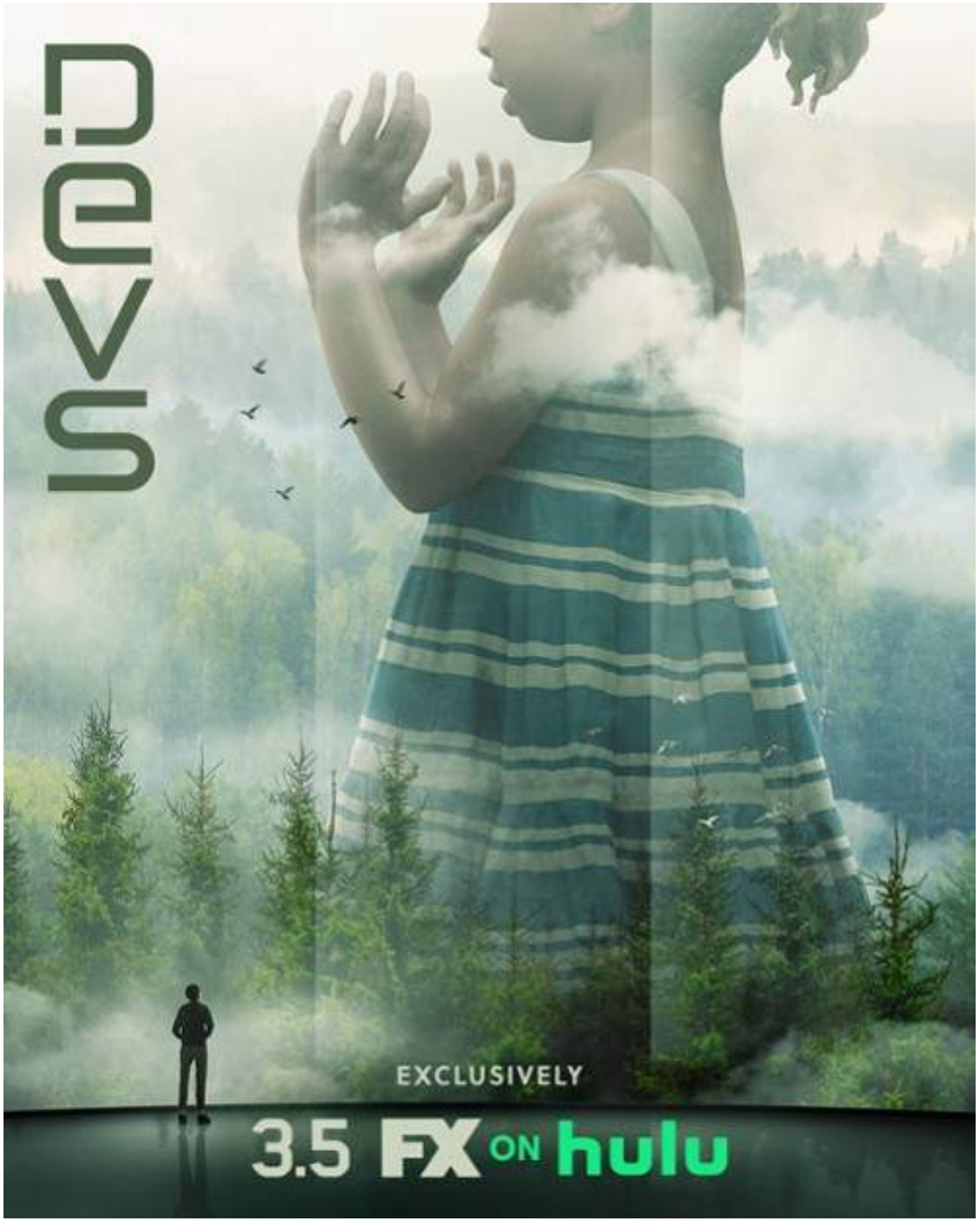
  1) wykorzystamy zdjęcia satelitarne powiatu śremskiego, a także informacje o klasach pokryciach terenu dostępne na www.s2glc.cbk.waw.pl

  2) zdjęcia i dane muszą zostać przygotowane do pracy pod kątem modelu ML (np. przycinanie, przekształcanie, wartości odstające)

  3) losowo wybierzemy część danych do nauki wykrywania klas zagospodarowania terenu przez model ML

  4) użyjemy wytrenowanego modelu do klasyfikacji zagospodarowania terenu powiatu śremskiego (zdjęcia satelitarne)

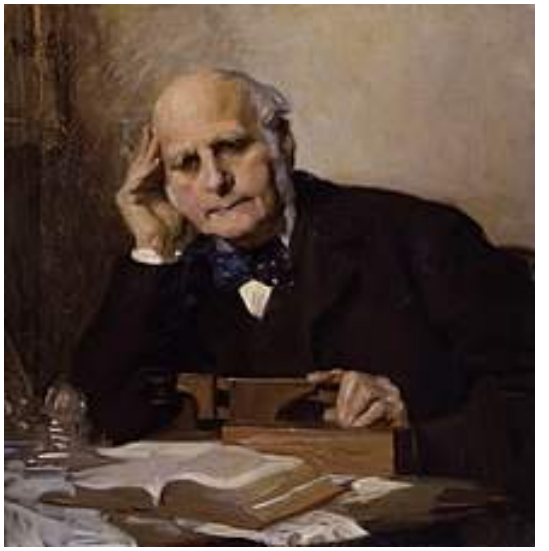
  5) ocenimy jakość klasyfikacji przeprowadzonej przez model



DEVS (2020) – HBO MAX

Regresja statystyczna

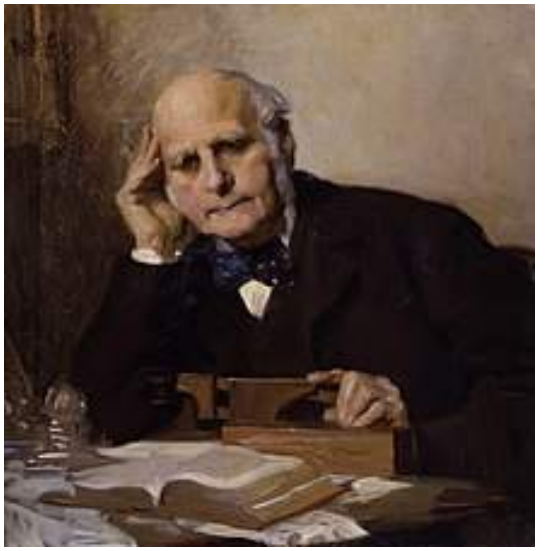
Regresja



1822-1911

- Ojcem był sir Francis Galton (1875)
- Stworzył fundament teoretyczny i empiryczny regresji analizując wzrost ludzi
- Koncepcja ta została potem rowinięta przez Karla Pearsona i Udny Yule (1896, 1930)

Regresja



1822-1911

- Ojcem był sir Francis Galton (1875)
- Stworzył fundament teoretyczny i empiryczny regresji analizując wzrost ludzi
- Koncepcja ta została potem rowinięta przez Karla Pearsona i Udny Yule (1896, 1930)
- Dwa rodzaje modeli: regresja prosta i regresja wieloraka
- Jest szeroko stosowanym narzędziem statystycznym służącym do ustalenia zależności pomiędzy zmiennymi.

Regresja statystyczna

- Celem jest określenie wpływu zestawu zmiennych niezależnych na zmienną zależną.
- W modelu regresji wielkość zmiany zmiennej zależnej ilustruje się jako wielokrotność zmiany zmiennych niezależnych.

Regresja statystyczna

- Celem jest określenie wpływu zestawu zmiennych niezależnych na zmienną zależną.
- W modelu regresji wielkość zmiany zmiennej zależnej ilustruje się jako wielokrotność zmiany zmiennych niezależnych.
- R. Fisher i założenia modelu parametrycznego (niezależność, normalność rozkładu, linowa zależność)

Regresja – założenie o niezależności zmiennych

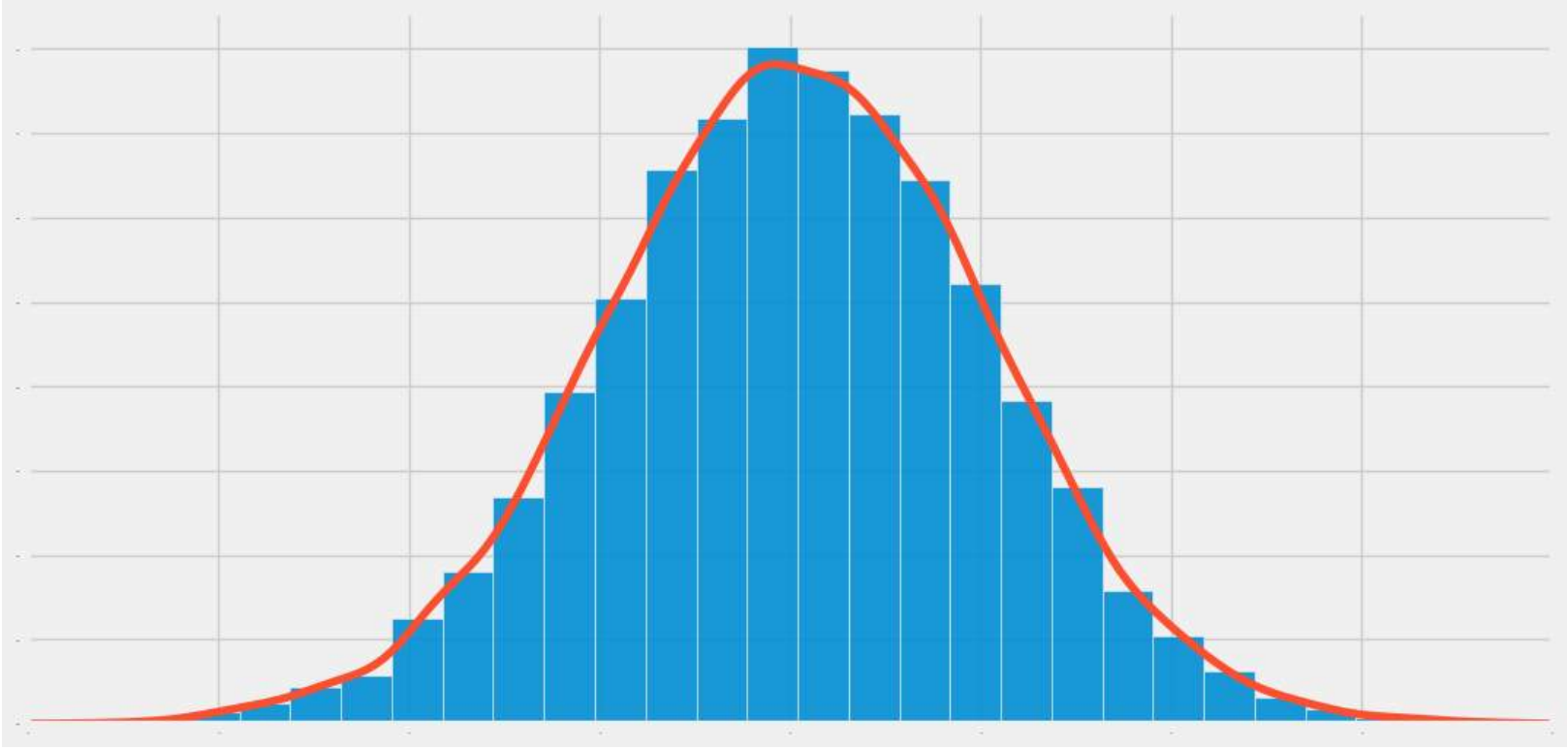
- Zależność oznacza istnienie (jakiegoś) połączenia pomiędzy zmiennymi
- Założenie o niezależności oznacza, że dane nie są powiązane ze sobą

Regresja – założenie o niezależności zmiennych

- Zależność oznacza istnienie (jakiegoś) połączenia pomiędzy zmiennymi
- Założenie o niezależności oznacza, że dane nie są powiązane ze sobą
- Brak związku pomiędzy uczestnikami badania
- Kluczowy więc jest etap gromadzenia/wyboru danych

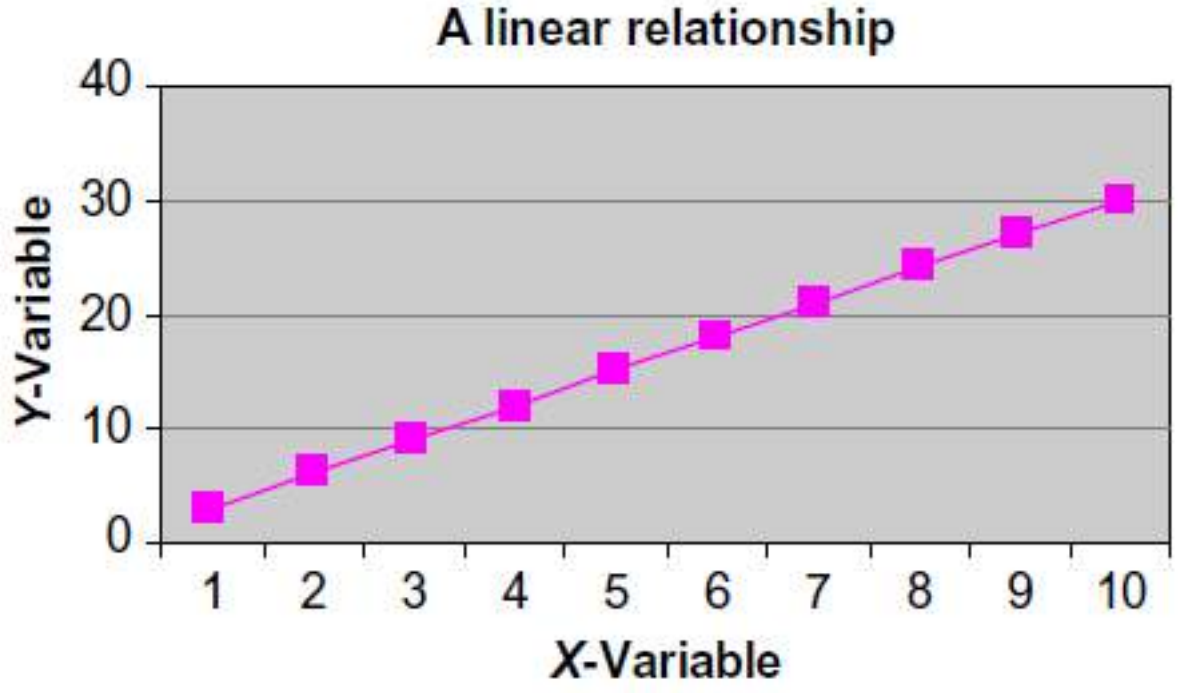
Regresja – założenie o normalności rozkładu

Rozkład wartości każdej zmiennej w zbiorze danych jest rozkładem normalnym skupionym wokół wartości średniej



Regresja – założenie o linowej zależności pomiędzy zmiennymi

- Zmienna niezależna wywiera liniowy efekt na zmienną zależną
- Efekt ten może być zilustrowany linią prostą



Metody regresji

- Regresja prosta
- Regresja wieloraka

...

regresja logistyczna, nieliniowa, krokowa

Regresja prosta

- Służy do przewidywania wartości zmiennej y na podstawie jednej zmiennej przewidującej x
- Celem jest zbudowanie modelu matematycznego (wzoru, formuły), który określa y jako funkcję zmiennej x.

Regresja prosta

- Służy do przewidywania wartości zmiennej y na podstawie jednej zmiennej przewidującej x
- Celem jest zbudowanie modelu matematycznego (wzoru, formuły), który określa y jako funkcję zmiennej x.
- Po zbudowaniu statystycznie istotnego modelu, można go wykorzystać do przewidywania przyszłych wyników zmiennej y na podstawie nowych wartości x

Regresja prosta

- Ogólny model regresji prostej to wzór na linię (y=ax+b):

$$Y = \mathrm{B}_0 + \mathrm{B}_1 X + e$$

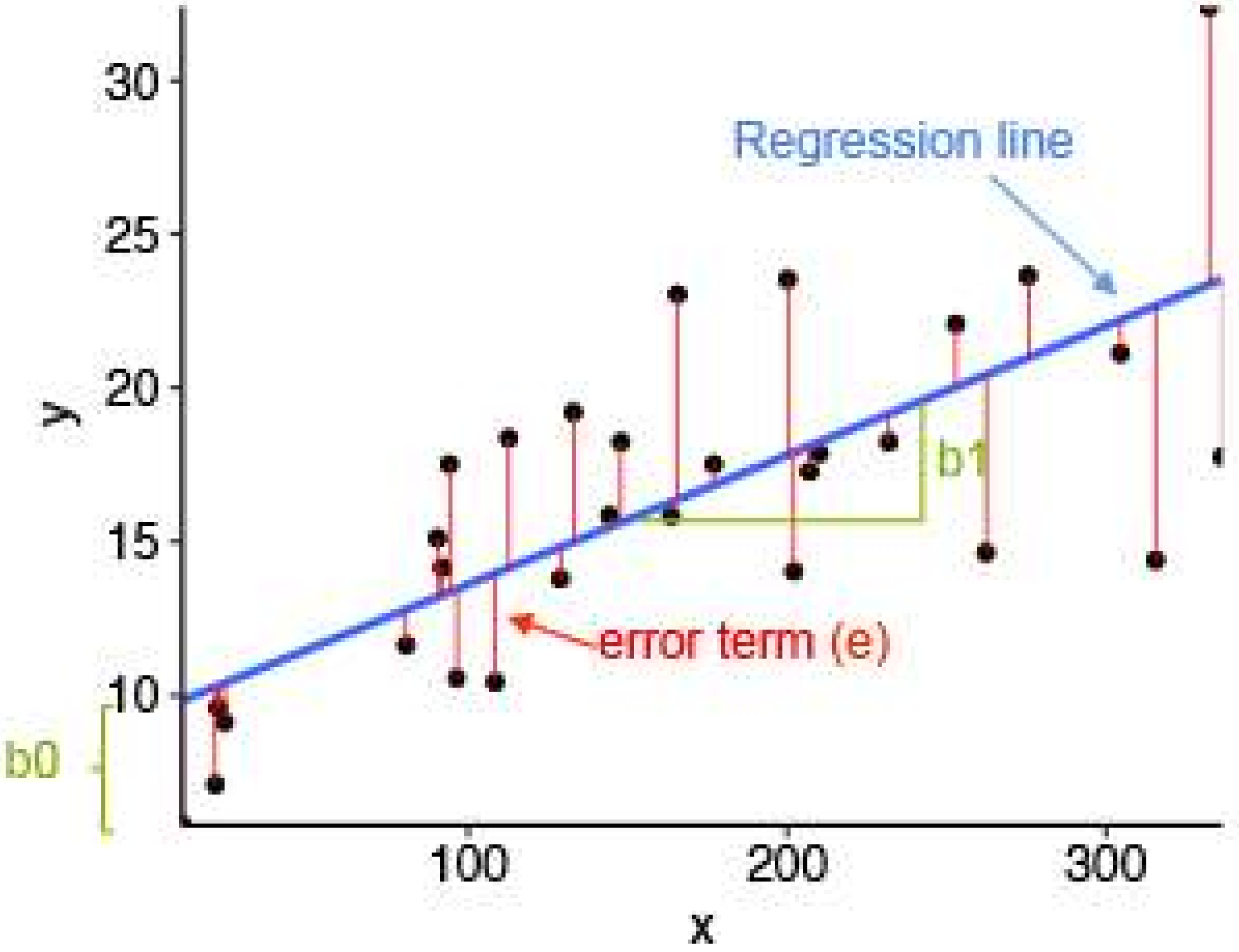
Gdzie:

Y - zmienna zależna
X – zmienna niezależna
B_0 – wyraz wolny
B_1 – współczynnik modelu regresji
e – składnik losowy modelu (błąd)

Regresja prosta

- Najprostszą metodą identyfikacji parametrów modelu jest metoda najmniejszych kwadratów (*ordinary least squares* (OLS))
- Bazuje ona na minimalizacji odległości punktów (obserwacji) od linii regresji

Regresja prosta



Regresja prosta

Przykład: Wpływ budżetu reklamowego youtube na wielkosć sprzedaży

Regresja wieloraka

- To rozszerzenie regresji prostej na przypadki z wieloma zmiennymi niezależnymi
- Wykorzystywana do predykcji wartości zmiennej Y podstawie wartości (wielu) zmiennych objaśniających X

Regresja wieloraka

- Ogólny model regresji wielorakiej to wzór:

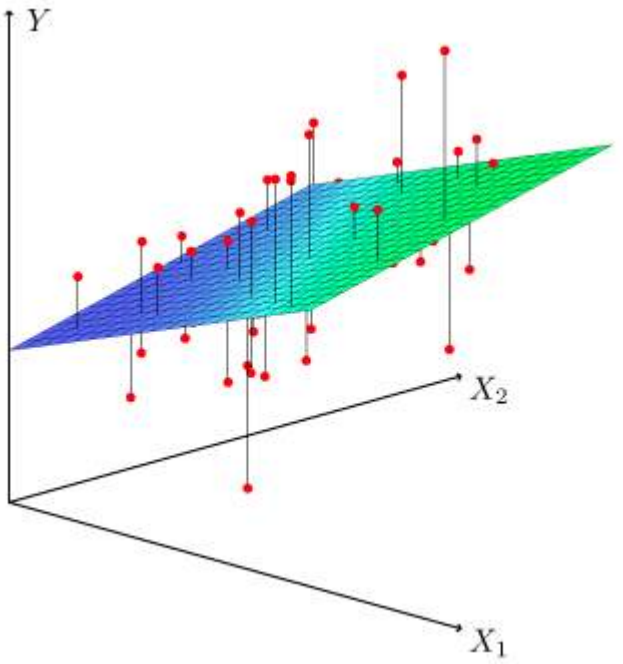
$$Y = \mathrm{B}_0 + \mathrm{B}_1 X_1 + \mathrm{B}_2 X_2 + \ldots + \mathrm{B}_n X_n + e$$

Gdzie:

Y - zmienna zależna
X_1 X_2 X_n – zmienne niezależne
B_0 – wyraz wolny
B_1 , B_2, B_n współczynniki modelu regresji
e – składnik losowy modelu (błąd)

Regresja wieloraka

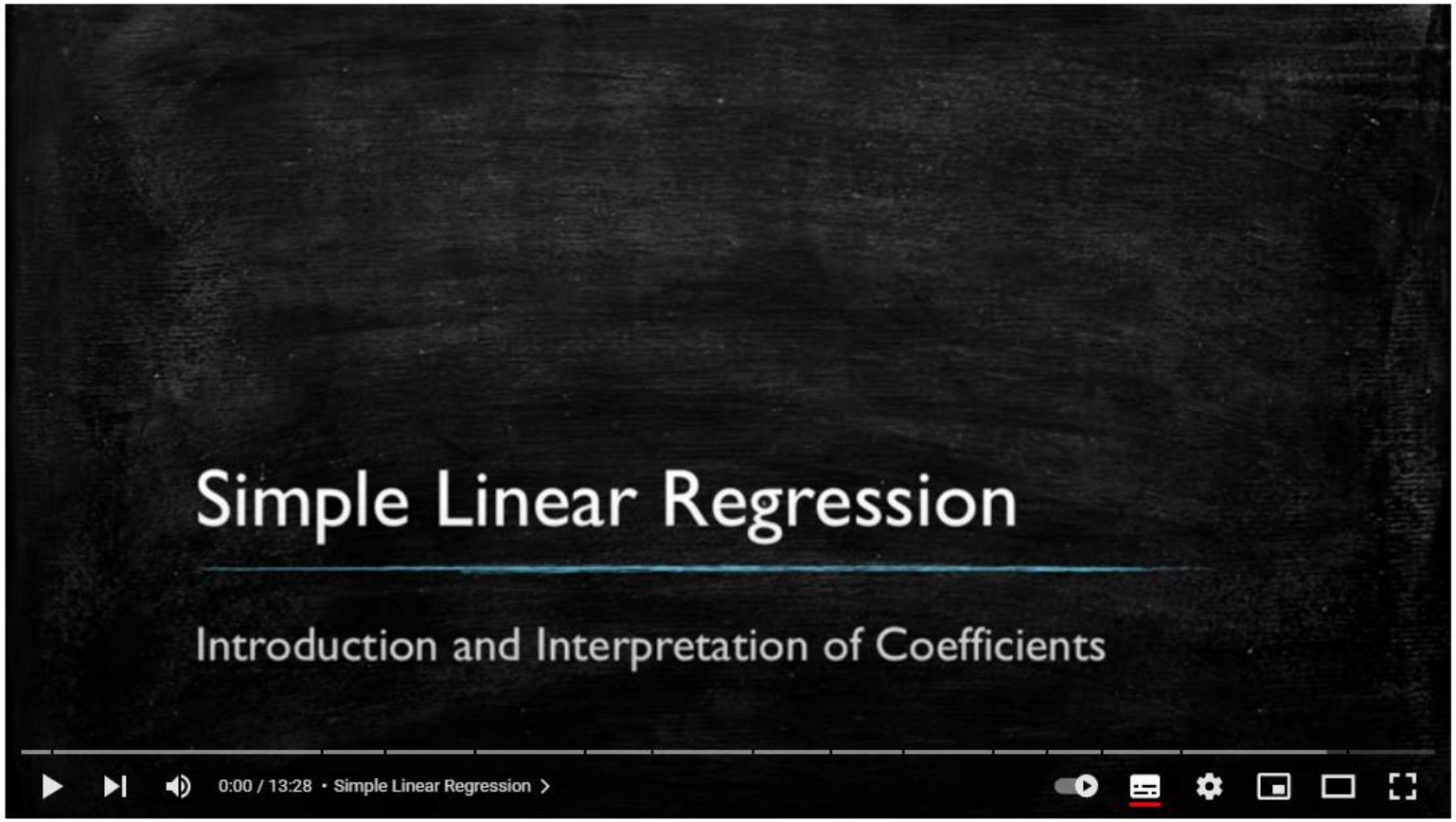
- Szacowanie wartości parametrów odbywa się tak samo jak w przypadku regresji prostej (OLS), jednak dla wielu wymiarów
- Chodzi o znalezienie n-wymiarowej płaszczyzny, która przechodzi najbliżej punktów współrzędnych (danych)



Regresja wieloraka

- Przykład: Wpływ budżetu reklamowego youtube, facebook i gazeta na wielkosć sprzedaży

Regresja - podsumowanie



https://www.youtube.com/watch?v=owI7zxCqNY0&ab_channel=dataminingincae