# *Analiza danych*

prof. UAM dr hab. Tomasz Górecki

tomasz.gorecki@amu.edu.pl
http://drizzt.home.amu.edu.pl

Zakład Statystyki Matematycznej i Analizy Danych
Wydział Matematyki i Informatyki
Uniwersytet im. Adama Mickiewicza w Poznaniu

# Generowanie dowolnych danych

## *Funkcja `sample()`*

Funkcja sample(), domyślnie generuje dane bez powtórzeń.

## *Predefiniowane wektory*

W R znajdują się pewne predefiniowane wektory, z których możemy losować elementy za pomocą funkcji sample. Są to: letters (małe litery), LETTERS (wielkie litery), month.abb (skróty trzyliterowe angielskich nazw miesięcy), month.name (angielskie nazwy miesięcy).

## *Funkcja `gl()`*

Funkcja gl() służy do generowania zmiennych czynnikowych.

gl(n = l. poziomów, k = l. powtórzeń, length = długość, labels = poziomy)

# Funkcje dotyczące rozkładów prawdopodobieństwa

| Rozkład | Dystrybuanta | Gęstość | Kwantyl | Generator |
|---|---|---|---|---|
| dwumianowy | pbinom | dbinom | qbinom | rbinom |
| Poissona | ppois | dpois | qpois | rpois |
| ujemny dwumianowy | pnbinom | dnbinom | qnbinom | rnbinom |
| geometryczny | pgeom | dgeom | qgeom | rgeom |
| hipergeometryczny | phyper | dhyper | qhyper | rhyper |
| wielomianowy | | dmultinom | | rmultinom |
| jednostajny | punif | dunif | qunif | runif |
| beta | pbeta | dbeta | qbeta | rbeta |
| wykładniczy | pexp | dexp | qexp | rexp |
| gamma | pgamma | dgamma | qgamma | rgamma |
| normalny | pnorm | dnorm | qnorm | rnorm |
| logarytmiczno-normalny | plnorm | dlnorm | qlnorm | rlnorm |
| Weibulla | pweibull | dweibull | qweibull | rweibull |
| chi-kwadrat | pchisq | dchisq | qchisq | rchisq |
| $t$ | pt | dt | qt | rt |
| Cauchy'ego | pcauchy | dcauchy | qcauchy | rcauchy |
| $F$ | pf | df | qf | rf |

# Rodzaje błędów

## *Błąd pomiarowy*

Rachunek błędów jest to zespół zagadnień na pograniczu metrologii, statystyki i matematyki. Obejmuje zasady opracowywania i prezentacji wyników doświadczalnych. Wszelkie wyniki pomiarów pozbawione dyskusji błędów, a zwłaszcza określenia błędu pomiarowego (różnica pomiędzy wynikiem pomiaru, a prawdziwą wartością), są w istocie wyłącznie wskazaniami. Błąd pomiarowy nie powstaje jedynie w wyniku pomyłki, jest on nieodłącznym czynnikiem procesu pomiarowego.

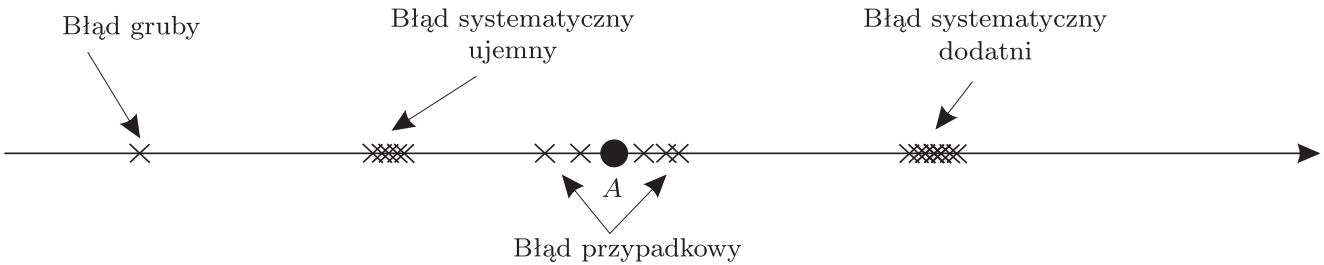
# Rodzaje błędów

## *Elementy błędu pomiarowego*

Błąd pomiarowy składa się z dwóch elementów:

1. Błąd statystyczny, przypadkowy – błąd wynikający z ogółu wpływów środowiska, których często nie można zidentyfikować czy wyeliminować. Charakteryzuje się niewielką wartością oraz losowym znakiem i wartością. Związany jest z pomiarem i nie można go całkowicie wyeliminować. Redukuje się go poprzez powtarzanie pomiarów i uśrednianie.
2. Błąd systematyczny – błąd wynikający z zastosowanej metody pomiaru lub innych przyczyn (np. niedających się wykluczyć, ale znanych zjawisk mających wpływ na pomiar). Charakteryzuje się stałym znakiem, tj. zawsze powoduje zawyżenie lub zaniżenie wartości wyniku pomiaru, wpływa jednakowo na wszystkie pomiary. Błąd systematyczny o znanej wartości nazywamy poprawką.

# Rodzaje błędów pomiarowych



# Obserwacja odstające

## *Błąd gruby*

Wyróżnia się również błąd gruby (pomyłkę), który jest pewną odmianą błędu przypadkowego, w sensie jego przypadkowego pojawiania się. Ma miejsce, gdy któryś z wyników pomiaru odbiega znacznie od pozostałych. Możemy wówczas podejrzewać, że pewne zdarzenie wypaczyło wynik eksperymentu. Błędy grube mogą wynikać np. ze złego odczytania skali przyrządu pomiarowego, pomyłki przy zapisie miejsca przecinka, pomiaru błędnego obiektu. Wyniki takie powinny zostać odrzucone podczas analizy statystycznej.

## *Obserwacja odstająca*

Obserwacja odstająca (*ang. outlier*) to obiekt, który tak bardzo różni się od innych obserwacji, iż powstaje podejrzenie, że został on wygenerowany przez inny mechanizm.

# Obserwacja odstające

## *Wartość odstająca vs anomalia*

W analizie danych wyróżnią się jeszcze pojęcia anomalii (*anomaly*), czyli obserwacji niezgodnych z oczekiwanym wzorcem innych elementów w zbiorze danych. Różnica między anomalią i obserwacją odstającą jest bardzo subtelna i zazwyczaj nie ma znaczenia. Oba pojęcia odnoszą się do punktów danych mających wyjątkowo niskie prawdopodobieństwo wystąpienia. Subtelna różnica polega na tym, że nazwanie czegoś anomalią sugeruje, że mamy hipotezę wygenerowania jej przez inny proces niż ten generujący normalne dane. Nazywanie tego wartością odstającą jest bardziej opisowe i nie wyklucza, że jest to statystyczny przypadek. Można również powiedzieć, że anomalia odbiega od oczekiwań modelu, podczas gdy wartość odstająca odbiega od większości danych.

# Wykrywanie obserwacji odstających

## *Wykrywanie obserwacji odstających (błędów grubych)*

1. Kryterium Chauveneta – obliczamy średnią oraz wariancję z całej próby, następnie dla podejrzanej obserwacji liczymy

$$t = \frac{|x - \bar{x}|}{s}.$$

Jeśli możemy założyć normalność pomiarów, to znajdujemy prawdopodobieństwo, że zmienna losowa będzie oddalona od średniej o nie mniej niż $ts$, czyli

$$p = P(|X - \bar{x}| \geq ts) = 1 - P(|X - \bar{x}| < ts) = 2 - 2 \cdot \Phi(t).$$

Dysponując próbą o liczebności $n$, spodziewamy się, że poza tym przedziałem znajdzie się $np$ obserwacji. Odrzucamy obserwację jeśli $np < 1/2$.

2. Kryterium Peirce'a – pozwala wykryć więcej niż jedną obserwację odstającą przy założeniu normalności danych.

# Wykrywanie obserwacji odstających

## *Eliminacja błędów grubych (obserwacji odstających)*

3. Filtr HAMPELA – odrzucamy obserwacje spoza przedziału

$$[\widetilde{X} - 4.5 \cdot \text{MAD}, \widetilde{X} + 4.5 \cdot \text{MAD}],$$

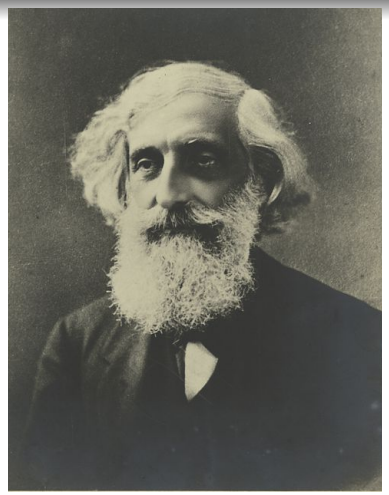
gdzie $\widetilde{X}$ jest medianą zbioru $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$ oraz

$$\text{MAD}(X) = \text{Mediana}|\boldsymbol{X} - \widetilde{X}|$$

jest bezwzględnym odchyleniem medianowym (*ang. median absolute deviation*).

# Wykrywanie obserwacji odstających

Benjamin Peirce
(1809-1880)



William Chauvenet
(1820-1870)

Frank Hampel
(1941-2018)

## *Bibliografia*

Chauvenet, W. (1863). *A manual of spherical and practical astronomy*. American Journal of Science 2(36):378–384.

Hampel, F.R. (1971). *A general qualitative definition of robustness*. Annals of Mathematics Statistics 42(6):1887–1896.

Peirce, B. (1852). *Criterion for the rejection of doubtful observations*. Astronomical Journal 2(45):161–163.

# Wykrywanie obserwacji odstających

## *Eliminacja błędów grubych (obserwacji odstających)*

3. Test statystyczny:
   - Test Dixona (test Q) – stosujemy go dla próbek o liczności do 30, dodatkowo zakładamy normalność danych.
   - Test Grubbsa – najpopularniejszy, zakładamy normalność danych.
   - Test Rosnera – do wykrywania wielu (do 10) obserwacji odstających równocześnie, zakładamy normalność danych (próba o liczności co najmniej 25).

# Wykrywanie obserwacji odstających

Bernard A. Rosner
(????-????)


Wilfrid J. Dixon
(1915-2008)

Frank E. Grubbs
(1913-2000)

## *Bibliografia*

Dixon, W.J. (1950). *Analysis of extreme values*. Annals of Mathematical Statistics 21(4):488–506.

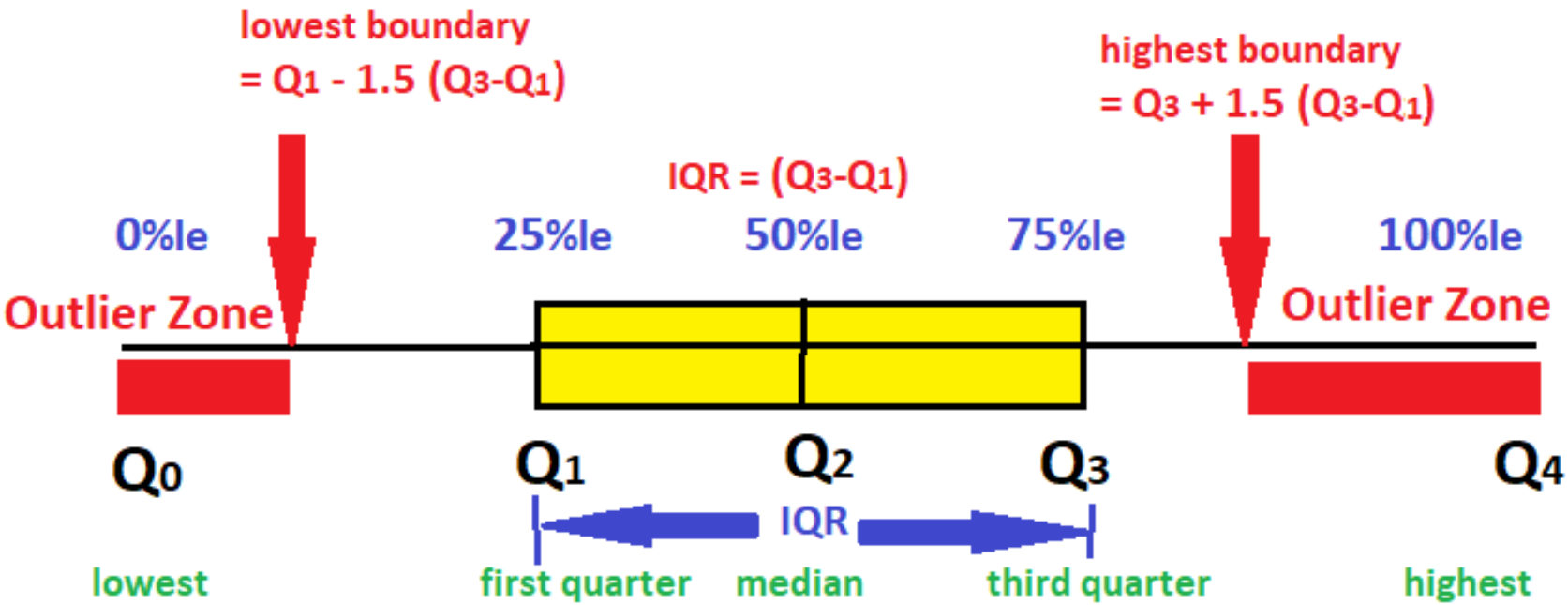
Grubbs, F.E. (1950). *Sample criteria for testing outlying observations*. Annals of Mathematical Statistics 21(1):27–58.

Rosner, B.A. (1975). On the Detection of Many Outliers. Technometrics 17(2):221–227.

# Wykrywanie obserwacji odstających

## *Eliminacja błędów grubych (obserwacji odstających)*

5. Wykres pudełkowy.

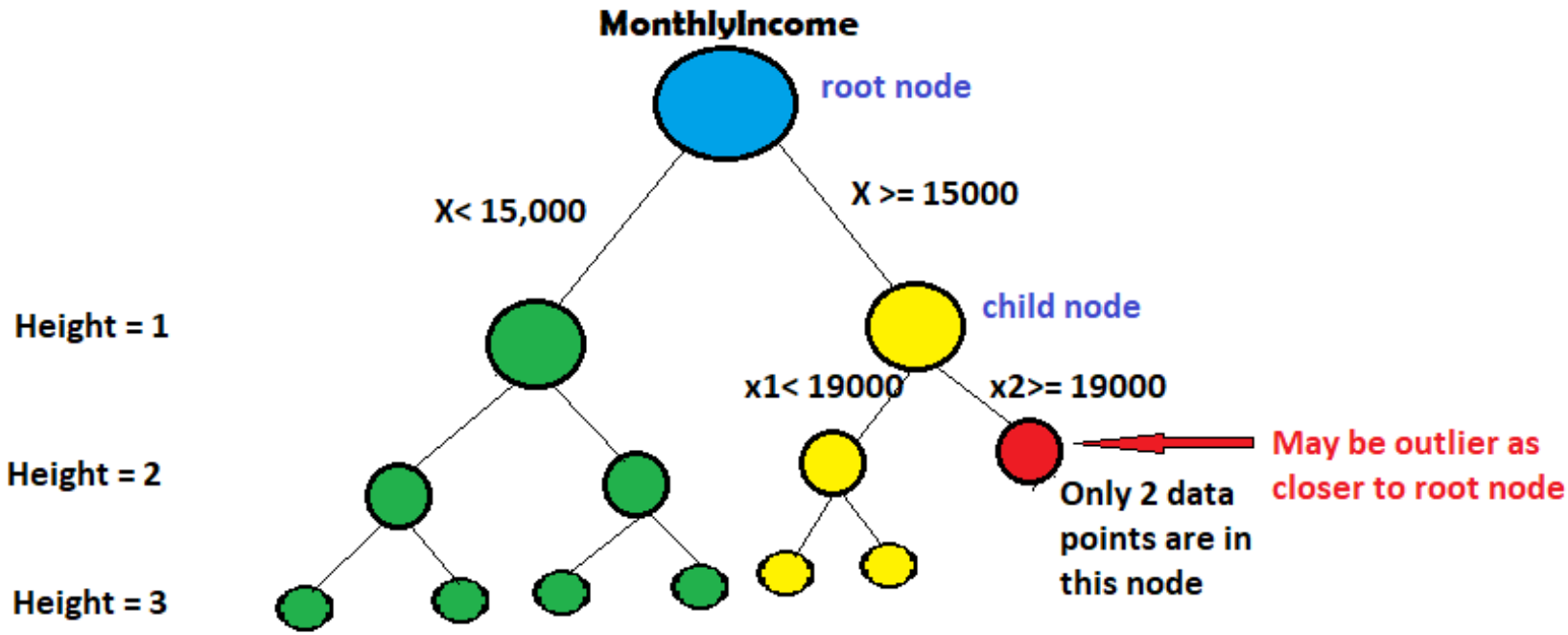


# Wykrywanie obserwacji odstających

## *Eliminacja błędów grubych (obserwacji odstających)*

6. Las izolacyjny (*ang. isolation forest*) – używany zwłaszcza dla danych wielowymiarowych. Tworzymy drzewa decyzyjne stosując losowe podziały. Obserwacje, które stają się szybko małymi liśćmi są bardziej prawdopodobne jako obserwacje odstające, ponieważ anomalie są bardziej podatne na izolację przy losowym podziale.

# Wykrywanie obserwacji odstających

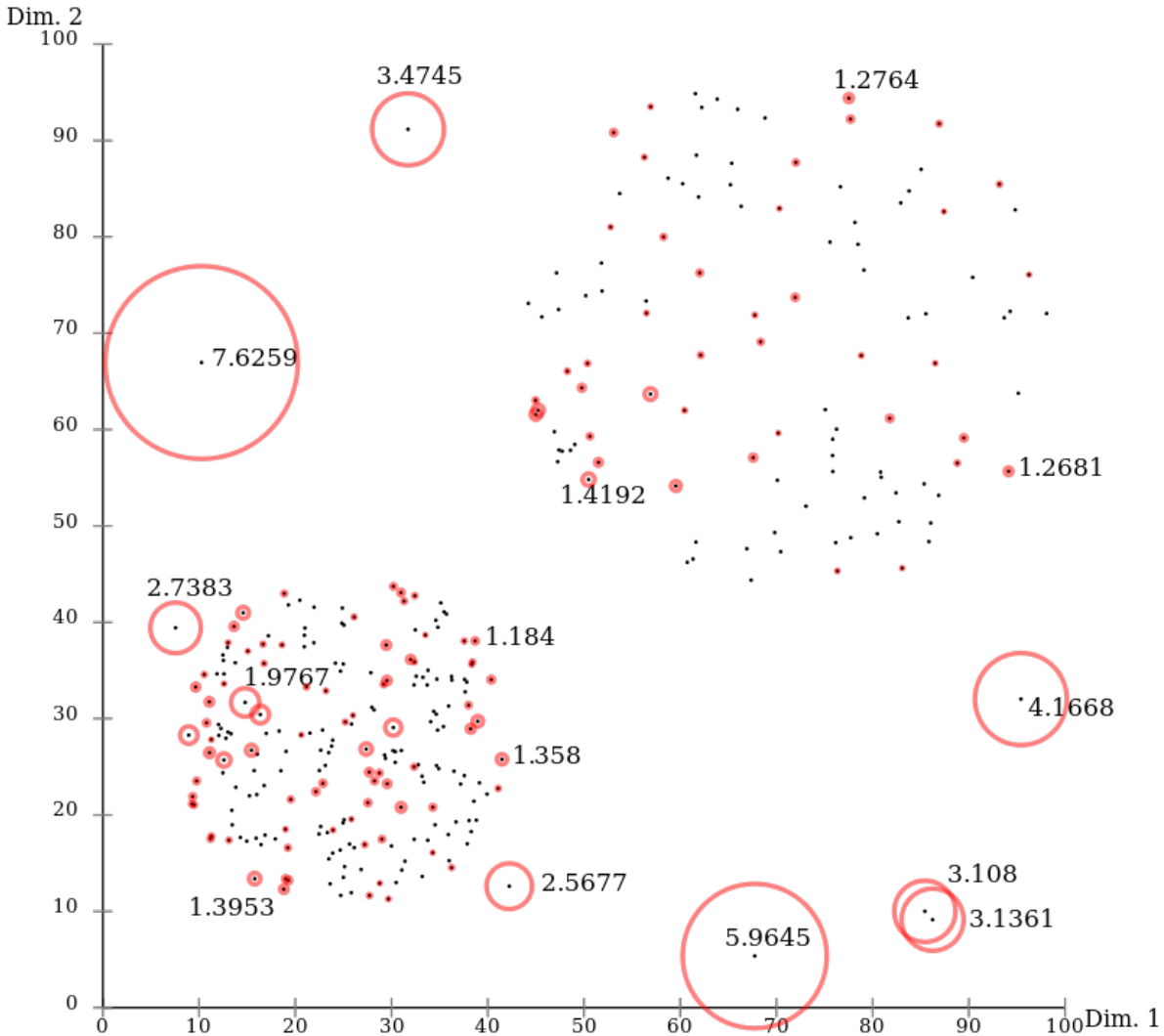


# Wykrywanie obserwacji odstających

## *Eliminacja błędów grubych (obserwacji odstających)*

7. Współczynnik odstawania lokalnego (*ang. local outlier factor (LOF)*) – bazuje na lokalnych gęstościach (podobnie jak algorytm DBSCAN w analizie skupień). Mała gęstość wokół punktu wskazuje na niego jako na obserwację odstającą.

# Wykrywanie obserwacji odstających



# Wykrywanie obserwacji odstających

## *Bibliografia*


Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J. (2000). *LOF: Identifying Density-based Local Outliers*. Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD. pp. 93–104.


Liu, F.T, Ting, K.M., Zhou, Z.-H. (2008). *Isolation-based Anomaly Detection*. ACM Transactions on Knowledge Discovery from Data 6:1–39.

# Wykrywanie obserwacji odstających

## *Eliminacja błędów grubych (obserwacji odstających) – przykład*

Wyniki oznaczeń zawartości jonów miedzi ($Cu^{2+}$) w próbce ścieków [$mg/dm^3$] wyglądają następująco: 0,875, 0,863, 0,876, 0,868, 0,771, 0,881, 0,878, 0,869, 0,866. Czy jakaś z obserwacji może zostać uznana za odstającą, jeśli zakładamy, że zawartość jonów ma rozkład normalny?

Korzystając z kryterium Chauveneta otrzymujemy

$$t = \frac{|0.771 - 0.8608|}{0.0341} = 2.63.$$

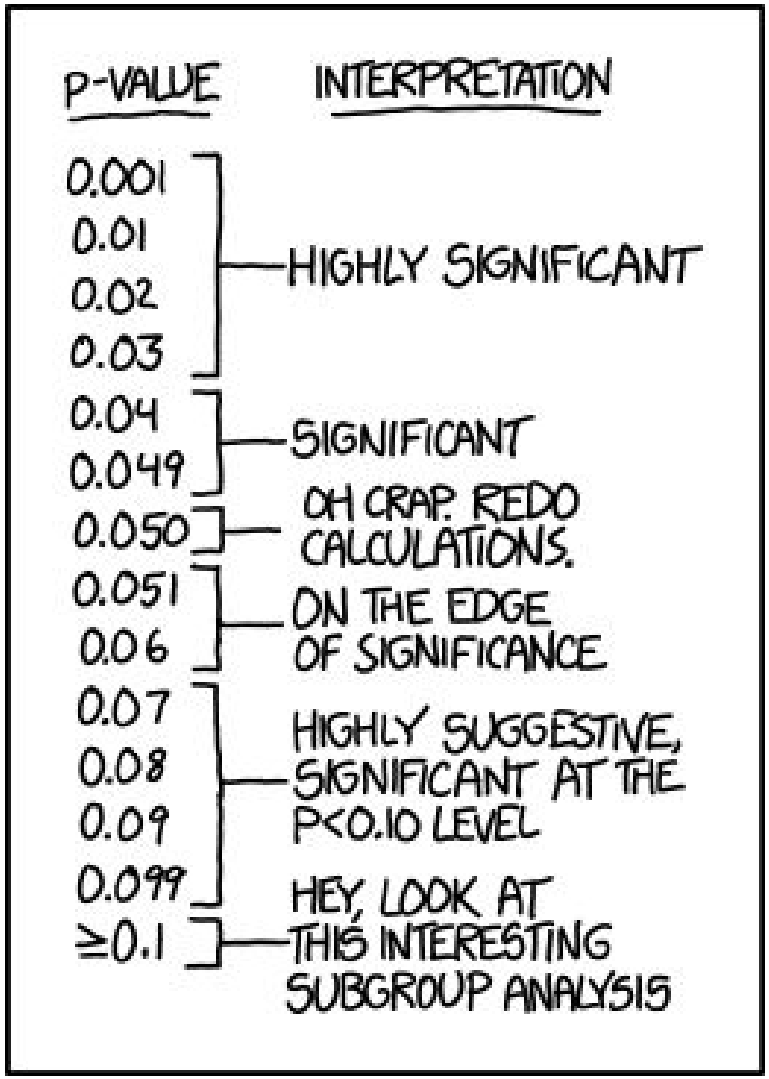
$$p = 2 - 2 \cdot \Phi(2.63) = 0.0085$$

$$np = 9 \cdot 0.0085 = 0.077 < 1/2.$$

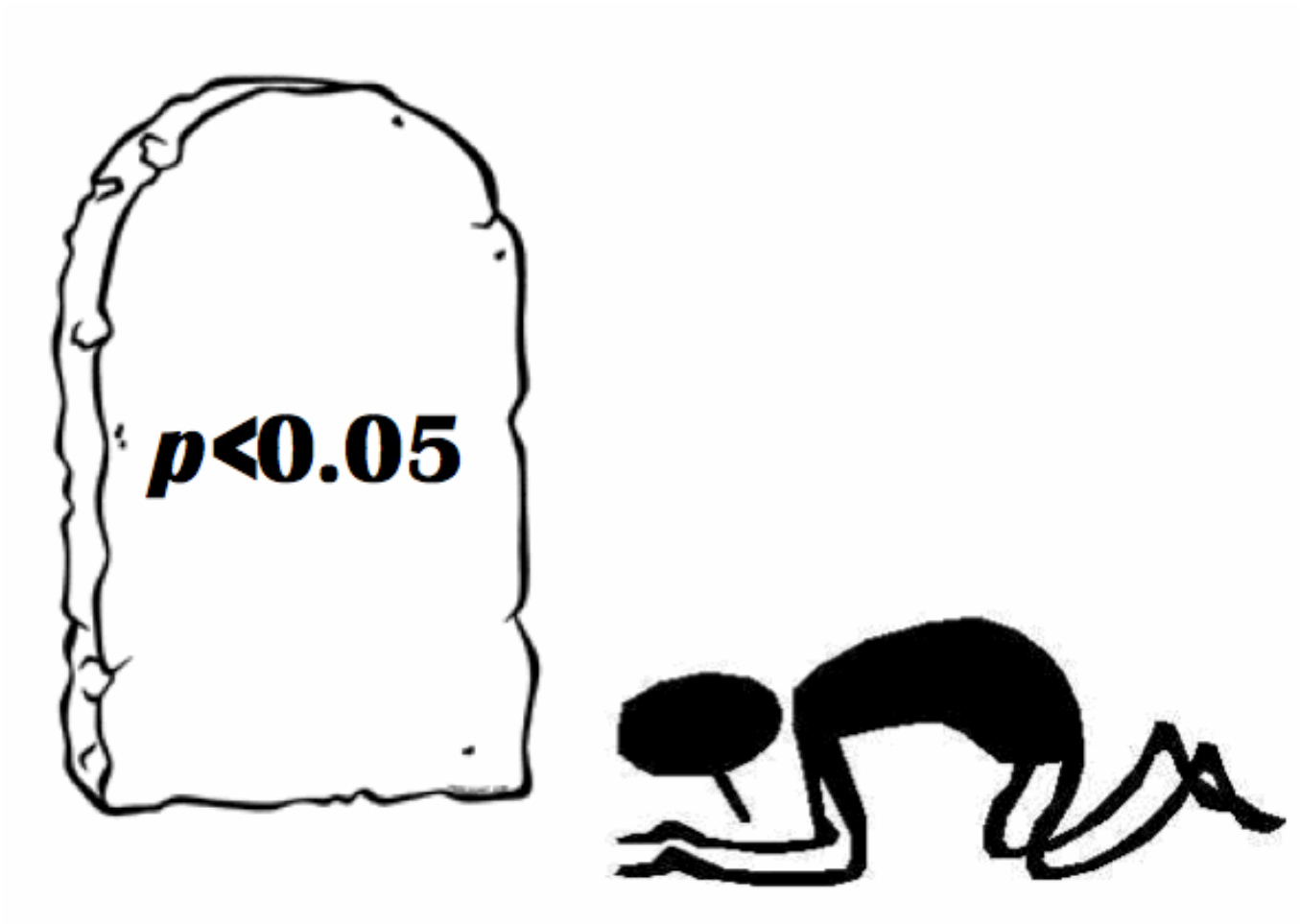
Zatem, najmniejsza obserwacja jest obserwacją odstającą.

# *p*-wartość

*p*-wartość (ang. *p*-value) to prawdopodobieństwo (przy prawdziwości $H_0$) otrzymania wartości równej lub bardziej ekstremalnej niż zaobserwowana. P-wartość pozwala bezpośrednio ocenić wiarygodność hipotezy. Im p-wartość jest większa, tym bardziej brak nam podstaw, aby w nią wątpić. Mała p-wartość świadczy przeciwko hipotezie zerowej.



# p-wartość



# Typy braków w danych

Typy brakujących danych (*ang. missing data*) reprezentują zależności statystyczne pomiędzy wartościami zaobserwowanymi cech, a rozkładem prawdopodobieństwa wartości brakujących. Poprawne rozpoznanie typu brakujących danych jest kluczowe w tworzeniu prawidłowych modeli imputacji danych.

- MCAR (*ang. Missing Completely At Random*) – braki nie zależą ani od wartości zaobserwowanych, ani od brakujących.
- MAR (*ang. Missing At Random*) – braki zależą tylko od zaobserwowanych obserwacji, a nie od innych braków.
- MNAR (*ang. Missing Not At Random*) – braki zależą od wartości brakujących.

# Typy braków w danych

## Types of missingness

**Missing Completely at Random**

(MCAR)

*No systematic relationship between missing data and other values*

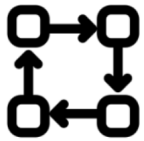
*Data entry errors when inputting data*

**Missing at Random**

(MAR)

*Systematic relationship between missing data and other observed values*

*Missing ozone data for high temperatures*



**Missing Not at Random**

(MNAR)

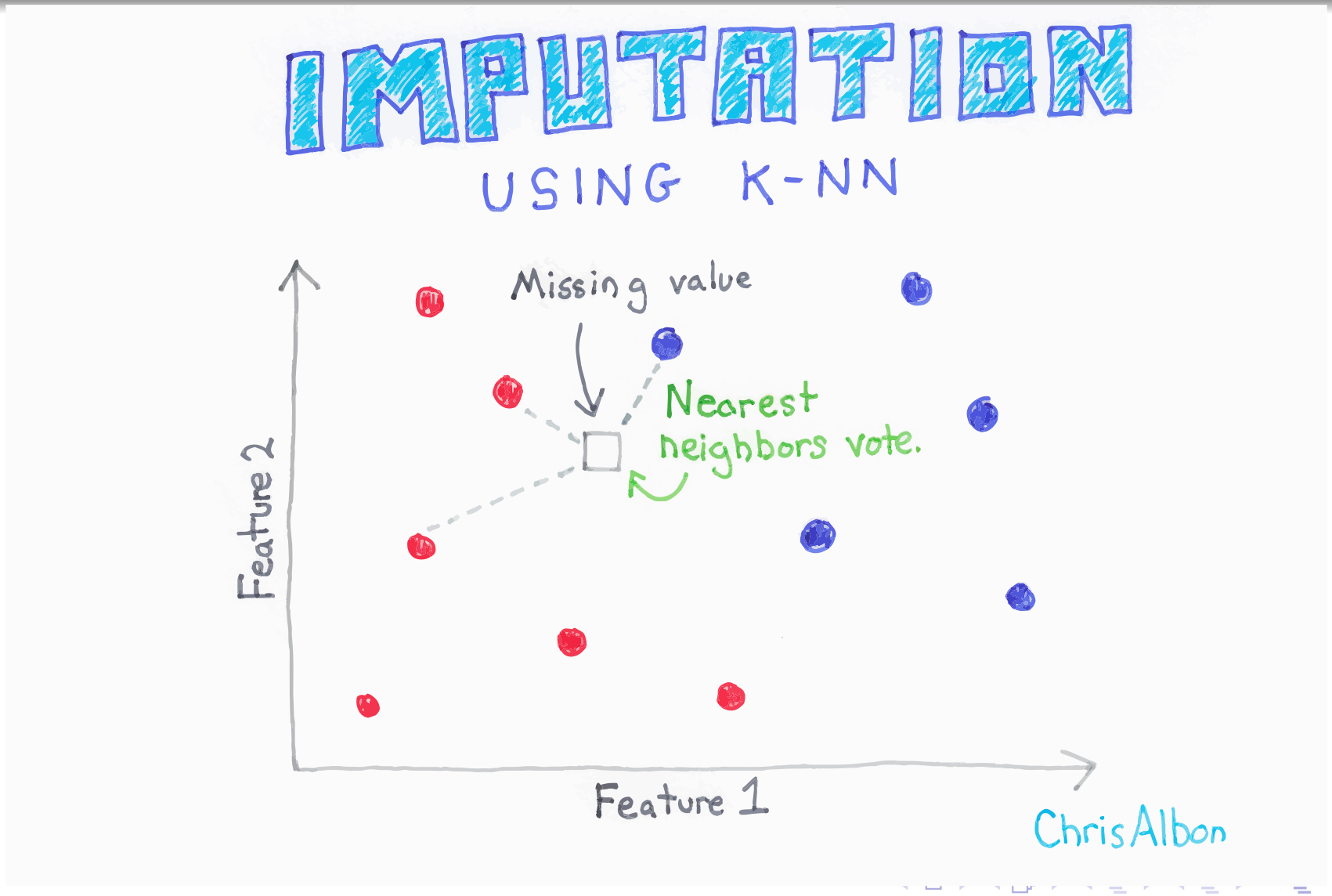
*Systematic relationship between missing data and unobserved values*

*Missing temperature values for high temperatures*

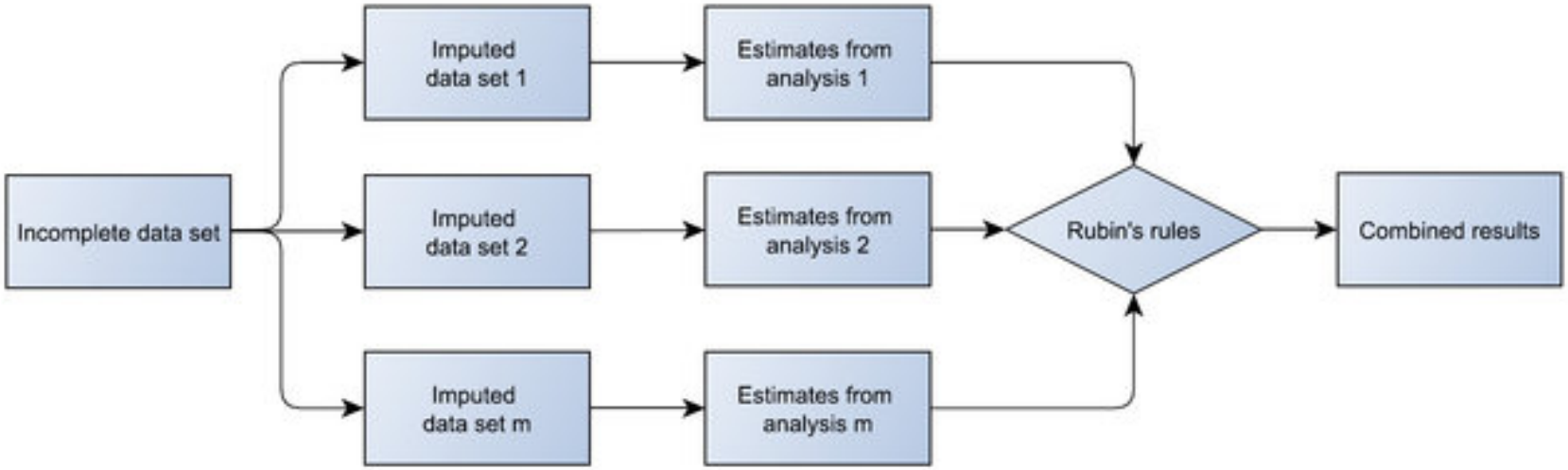
# Radzenie sobie z brakami danych

1. Usunięcie wierszy/kolumn z brakami w danych,
2. Uzupełnienie danych (imputacja):
   1. Stała wartość (np. ostatnia zaobserwowana),
   2. Średnia, mediana, dominanta ze znanych wartości,
   3. Estymacja za pomocą modelu predykcyjnego,
   4. Wielokrotna imputacja (MICE, Amelia II) – wykonywane są wielokrotne imputacja na bazie modelu probabilistycznego i ostateczna imputacja jest średnią z wielu imputacji pośrednich.

# Radzenie sobie z brakami danych



# Radzenie sobie z brakami danych




## *Bibliografia*

van Buuren, S., Groothuis-Oudshoorn, K. (2011). *mice: Multivariate Imputation by Chained Equations in R*. Journal of Statistical Software 45(3):1–67.

King, G., Honaker, J., Joseph, A., Scheve, K. (2001). *Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation*. American Political Science Review 95(1):49-69.

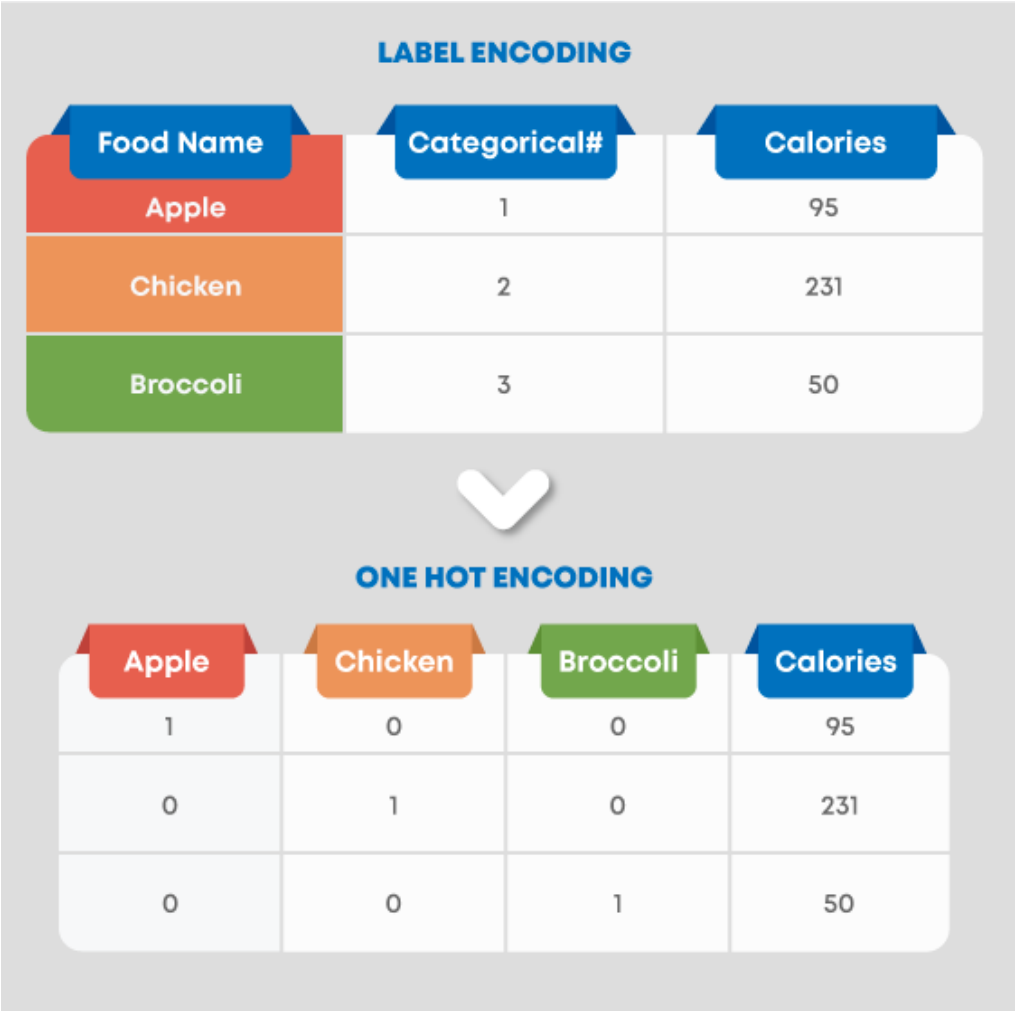
# Kodowanie zmiennych

Większość algorytmów uczenia maszynowego/sztucznej inteligencji działa lepiej z liczbowymi danymi wejściowymi. Dlatego jednym z głównych wyzwań, przed którym staje analityk, jest konwersja danych tekstowych/kategorycznych na dane liczbowe. Istnieje wiele sposobów konwersji wartości kategorycznych na wartości liczbowe.

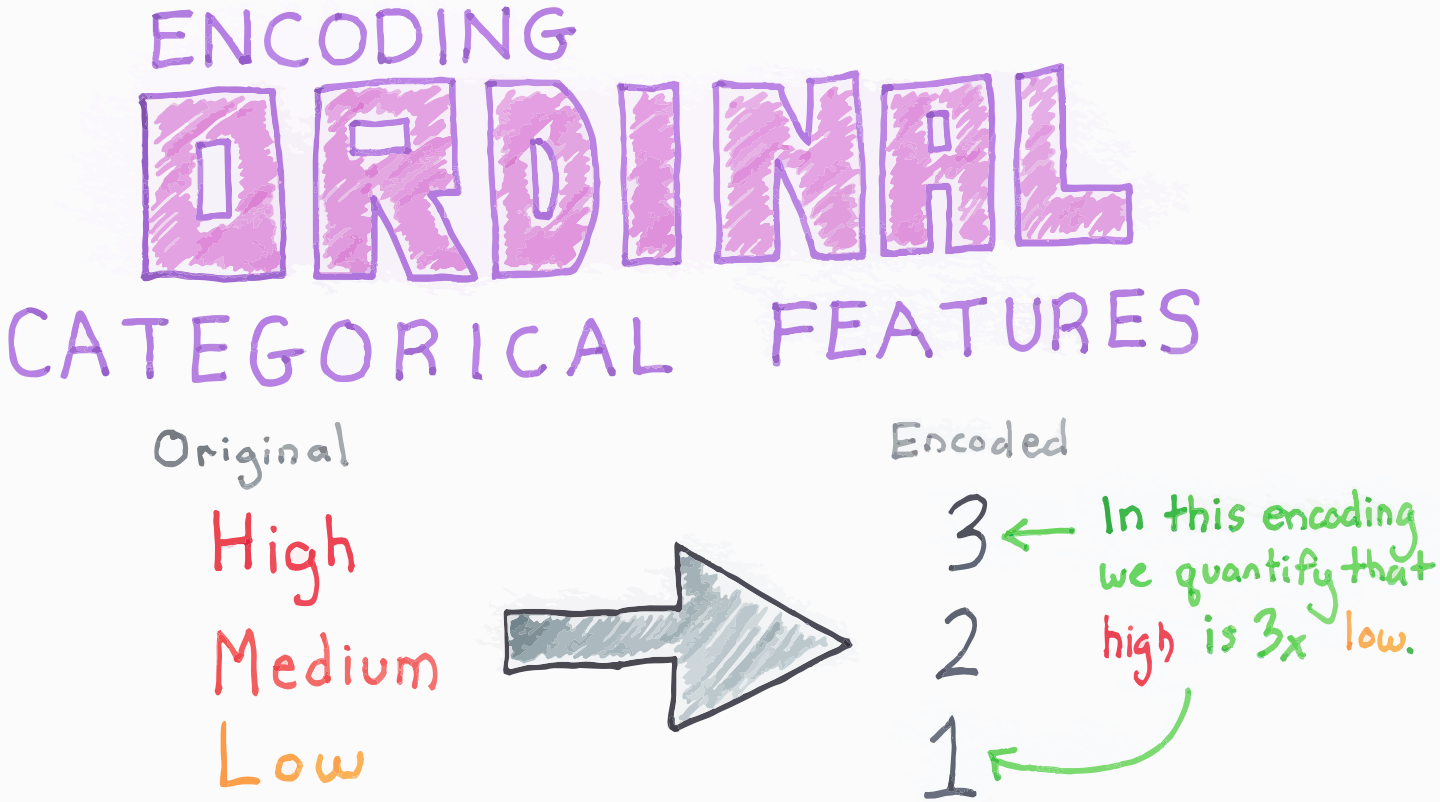
# Kodowanie etykiet

Kodowanie etykiet (*ang. label encoding, ordinal encoding*). Jeśli istnieje jakaś znana relacja (lub kolejność) pomiędzy poziomami lub etykietami w zmiennej jakościowej, wtedy najlepiej jest użyć techniki kodowania etykiet – konwertowanie etykiet na postać numeryczną.

# Kodowanie etykiet



# Kodowanie etykiet

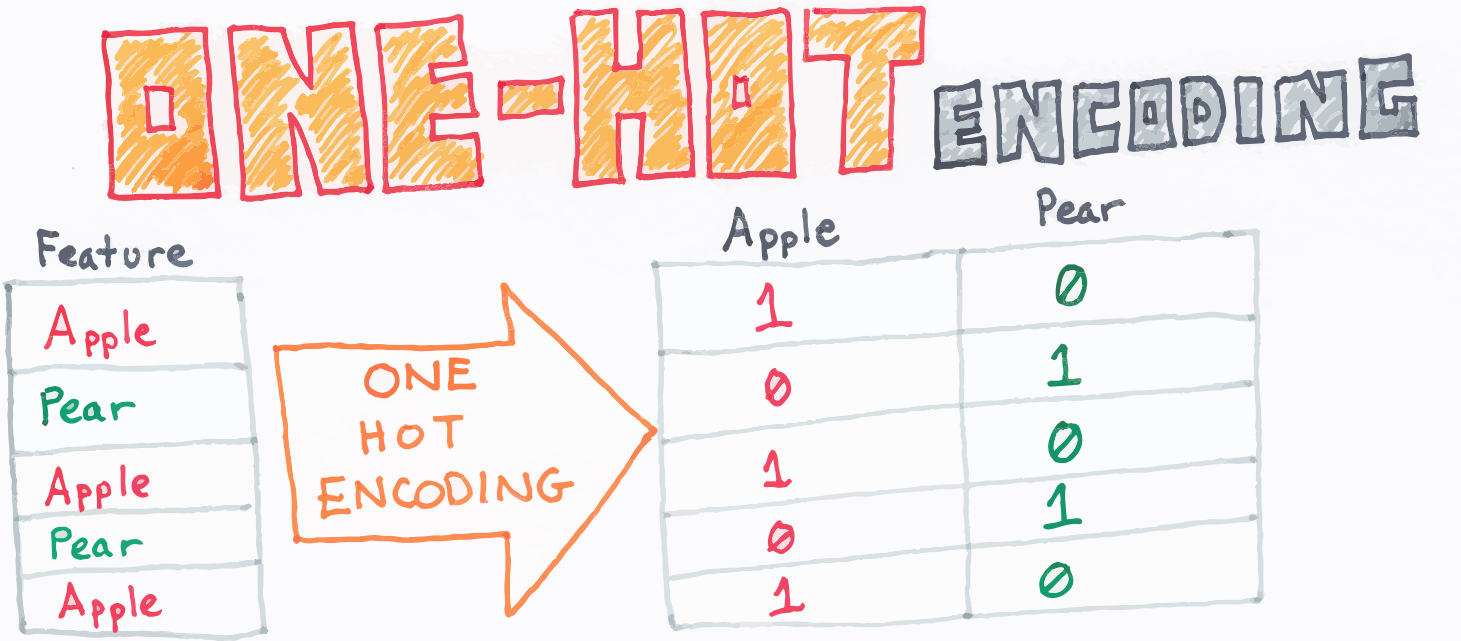



Many machine learning algorithms require numerical feature values. This might seem simple but it is important to notice that we are encoding the interval between categories.

## Gorące kodowanie

Gorące kodowanie (*ang. one hot encoding*). Chociaż kodowanie etykiet jest proste, ma ono tę wadę, że wartości numeryczne mogą być błędnie interpretowane przez algorytmy jako mające w sobie jakąś hierarchię/porządek. Ten problem porządku jest rozwiązywany w innym powszechnym podejściu alternatywnym, zwanym kodowaniem gorącym. W tej strategii każda wartość kategorii jest przekształcana w nową kolumnę i przypisywana jest do niej wartość 1 lub 0 (notacja dla prawda/fałsz).

# Gorące kodowanie



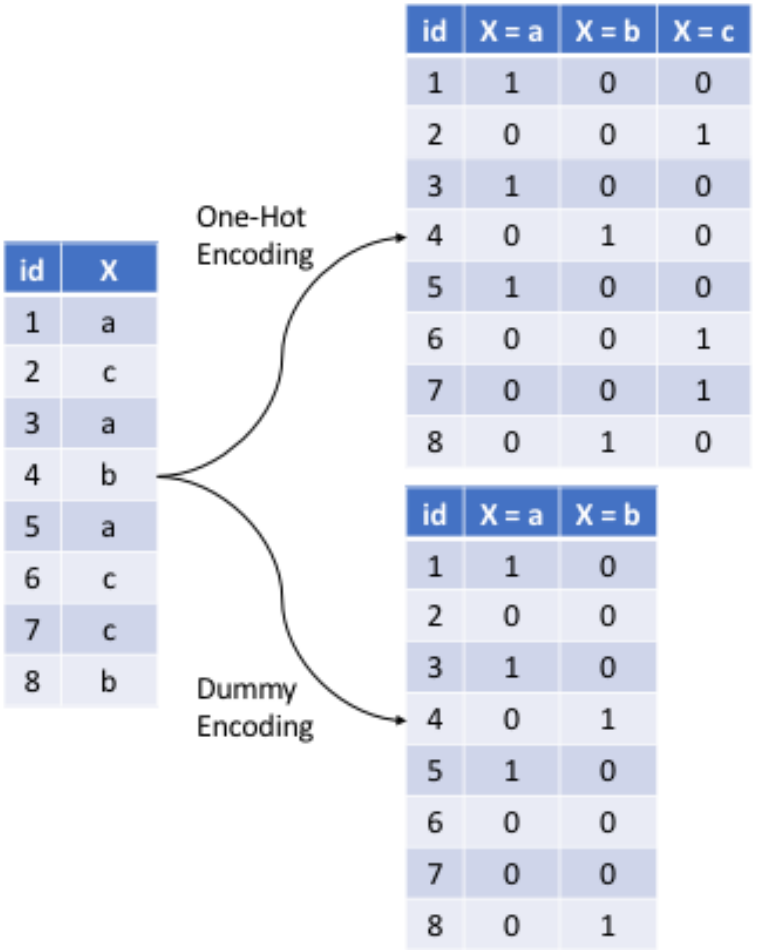

One-hot encoding allows us to turn nominal categorical data into features with numerical values, while not mathematically imply any ordinal relationship between the classes.

ChrisAlbon

## Gorące kodowanie

Chociaż to podejście eliminuje problemy związane z hierarchią/porządkiem, ma jednak wadę polegającą na dodaniu większej liczby kolumn do zbioru danych. Może to spowodować, że liczba kolumn znacznie się rozszerzy, jeśli mamy wiele unikalnych wartości w kolumnie kategorii. Zwróćmy uwagę, że w tej technice ostatnia kolumna jest nadmiarowa (jej zawartość wynika z zawartości poprzednich kolumn). Taka wersja kodowania gorącego nazywana jest kodowaniem naiwnym (*ang. dummy encoding*). Kodowanie naiwne jest niewielkim ulepszeniem w stosunku do kodowania gorącego, mianowicie wykorzystuje $N-1$ cech (kolumn) do reprezentowania $N$ etykiet/kategorii.

# Gorące kodowanie




| id | X |
|---|---|
| 1 | a |
| 2 | c |
| 3 | a |
| 4 | b |
| 5 | a |
| 6 | c |
| 7 | c |
| 8 | b |

| id | X = a | X = b | X = c |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 0 | 1 |
| 3 | 1 | 0 | 0 |
| 4 | 0 | 1 | 0 |
| 5 | 1 | 0 | 0 |
| 6 | 0 | 0 | 1 |
| 7 | 0 | 0 | 1 |
| 8 | 0 | 1 | 0 |

| id | X = a | X = b |
|---|---|---|
| 1 | 1 | 0 |
| 2 | 0 | 0 |
| 3 | 1 | 0 |
| 4 | 0 | 1 |
| 5 | 1 | 0 |
| 6 | 0 | 0 |
| 7 | 0 | 0 |
| 8 | 0 | 1 |

## Kodowanie binarne

Kodowanie binarne (*ang. binary encoding*). Podczas gdy kodowanie gorące rozwiązuje problem nierównych wag nadawanych kategoriom w ramach cechy, nie jest ono zbyt użyteczne, gdy istnieje wiele kategorii, ponieważ spowoduje to utworzenie wielu nowych kolumn, co może skutkować przekleństwem wymiarowości (*ang. curse of dimensionality*). Koncepcja „klątwy wymiarowości" mówi o tym, że w przestrzeniach wielowymiarowych pewne algorytmy przestają działać poprawnie. W takiej sytuacji można spróbować zastosować kodowanie binarne. W tej technice, najpierw kategorie są kodowane jako porządkowe, następnie te liczby całkowite są konwertowane na kod binarny, a ostatecznie cyfry z tego ciągu binarnego są dzielone na osobne kolumny. Dzięki temu dane są kodowane w mniejszej ilości wymiarów niż w przypadku kodowania gorącego.

# Kodowanie binarne

| Temperature |
|---|
| Hot |
| Cold |
| Very Hot |
| Warm |
| Hot |
| Warm |
| Warm |
| Hot |
| Hot |
| Cold |

| Order |
|---|
| 1 |
| 2 |
| 3 |
| 4 |
| 1 |
| 4 |
| 4 |
| 1 |
| 1 |
| 2 |

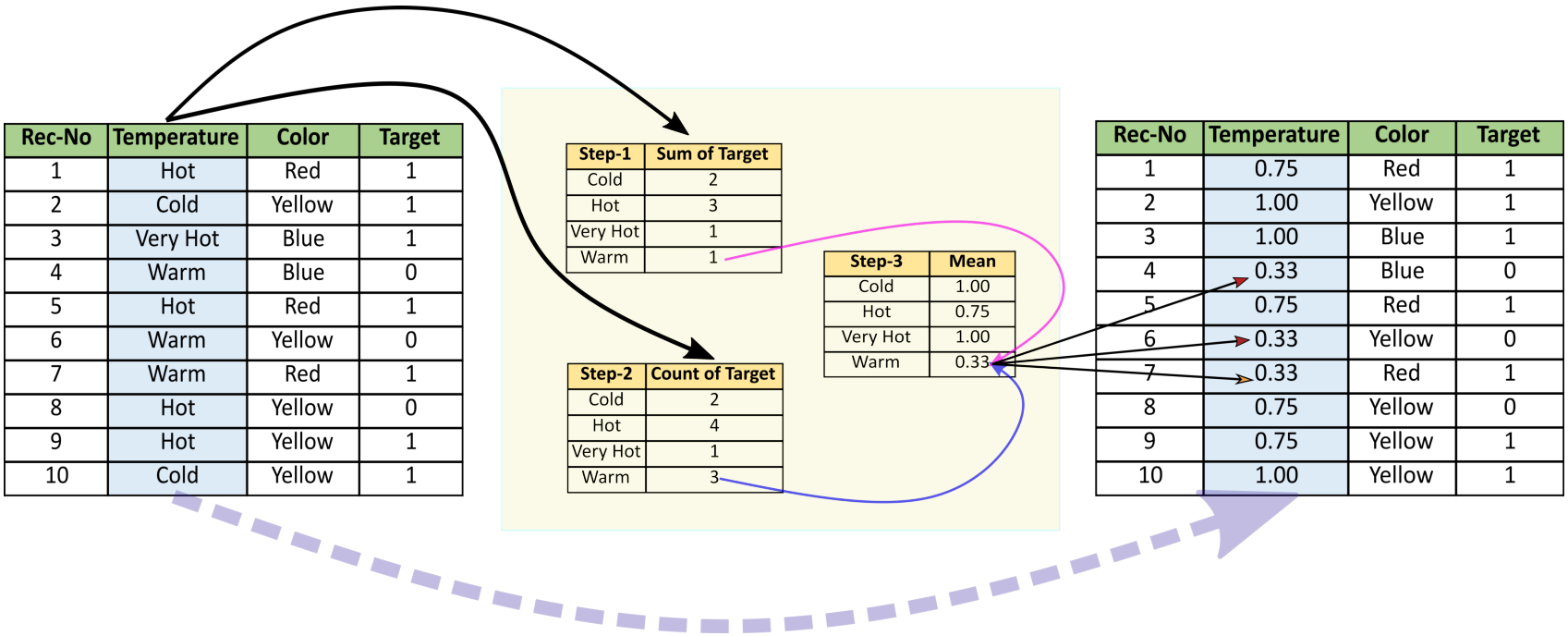
| Binary |
|---|
| 001 |
| 010 |
| 011 |
| 100 |
| 001 |
| 100 |
| 100 |
| 001 |
| 001 |
| 010 |

| Temperature_0 | Temperature_1 | Temperature_2 |
|---|---|---|
| 0 | 0 | 1 |
| 0 | 1 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 0 |
| 0 | 0 | 1 |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 1 | 0 |

# Kodowanie średnią

Kodowanie średnią (*ang. mean encoding, target encoding*). Kodowanie średnią jest podobne do kodowania etykiet, z tą różnicą, że tutaj etykiety są skorelowane bezpośrednio z odpowiedzią (zmienną objaśnianą). W kodowaniu średnią dla każdej kategorii etykieta jest ustalana jako średnia wartość zmiennej docelowej na danych uczących dla tej kategorii (w klasyfikacji jest to prawdopodobieństwo przynależności do klas, dodajemy zatem: `liczba kategorii` $\times$ `liczb klas` nowych cech). Obliczamy statystyki na danych uczących i używamy ich na zbiorze testowym. Jeśli kategoria ze zbioru testowego nie występuje w zbiorze uczącym, używamy statystyki obliczonej na pełnym zbiorze uczącym (bez podziału na kategorie).

# Kodowanie średnią



| Rec-No | Temperature | Color | Target |
|---|---|---|---|
| 1 | Hot | Red | 1 |
| 2 | Cold | Yellow | 1 |
| 3 | Very Hot | Blue | 1 |
| 4 | Warm | Blue | 0 |
| 5 | Hot | Red | 1 |
| 6 | Warm | Yellow | 0 |
| 7 | Warm | Red | 1 |
| 8 | Hot | Yellow | 0 |
| 9 | Hot | Yellow | 1 |
| 10 | Cold | Yellow | 1 |

| Step-1 | Sum of Target |
|---|---|
| Cold | 2 |
| Hot | 3 |
| Very Hot | 1 |
| Warm | 1 |

| Step-2 | Count of Target |
|---|---|
| Cold | 2 |
| Hot | 4 |
| Very Hot | 1 |
| Warm | 3 |

| Step-3 | Mean |
|---|---|
| Cold | 1.00 |
| Hot | 0.75 |
| Very Hot | 1.00 |
| Warm | 0.33 |

| Rec-No | Temperature | Color | Target |
|---|---|---|---|
| 1 | 0.75 | Red | 1 |
| 2 | 1.00 | Yellow | 1 |
| 3 | 1.00 | Blue | 1 |
| 4 | 0.33 | Blue | 0 |
| 5 | 0.75 | Red | 1 |
| 6 | 0.33 | Yellow | 0 |
| 7 | 0.33 | Red | 1 |
| 8 | 0.75 | Yellow | 0 |
| 9 | 0.75 | Yellow | 1 |
| 10 | 1.00 | Yellow | 1 |

## Kodowanie średnią

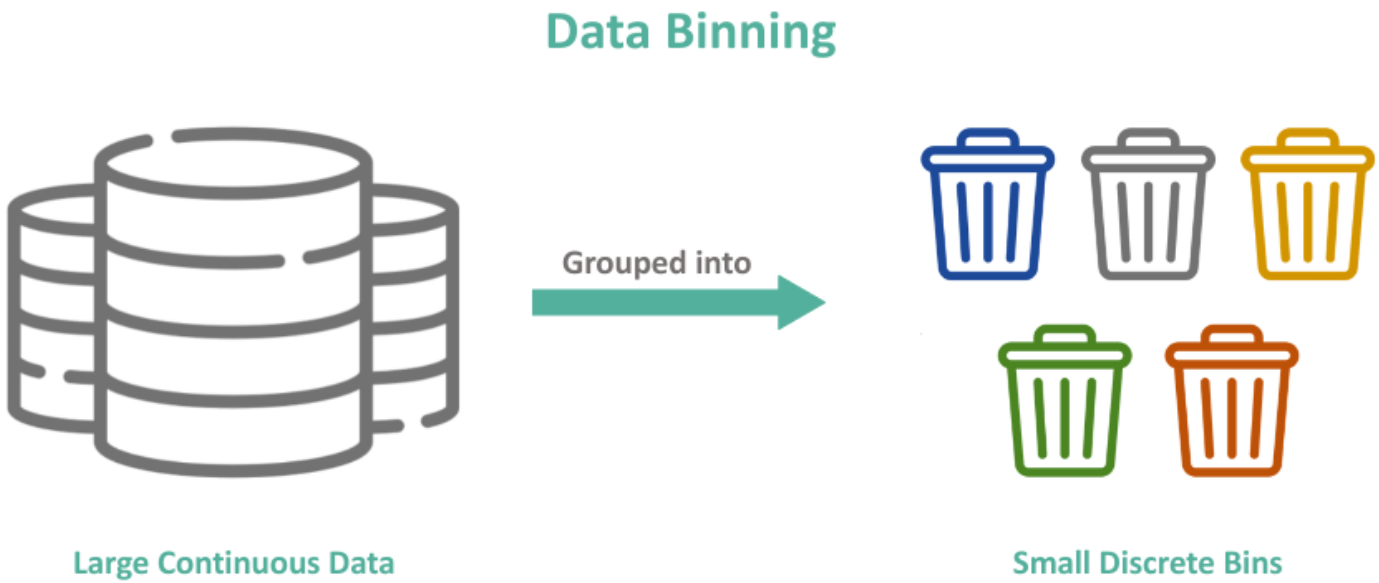
Taka metoda kodowania ma również pewien problem: nadmierne dopasowanie. Poleganie na średniej wartości nie zawsze jest dobrym pomysłem, gdy liczba wartości użytych do obliczenia średniej jest mała. Chodzi o to, że średniej nie można ufać, ponieważ jest zbyt mało wartości. Sztuczka polega na „wygładzeniu" średniej przez uwzględnienie średniej oceny ze wszystkich obserwacji. Innymi słowy, jeśli nie ma wielu obserwacji, powinniśmy polegać na globalnej średniej, natomiast jeśli jest wystarczająca liczba obserwacji, możemy bezpiecznie polegać na średniej lokalnej. Czyli:

$$\bar{x}_i = \frac{n_i \cdot \bar{x}_{\text{local}} + m \cdot \bar{x}_{\text{global}}}{n + m},$$

gdzie $n_i$ jest liczebnością konkretnej kategorii $i$, a $m$ jest parametrem wygładzania. Dla $m = 0$ bierzemy zwykłe średnie. Parametr $m$ jest dość intuicyjny: wymagamy co najmniej $m$ obserwacji w grupie, aby średnia kategorii przewyższyła średnią globalną.

## Kubełkowanie

Kubełkowanie danych (*ang. binning, bucketing*), jest procesem używanym do minimalizacji efektów błędów obserwacji. Jest to proces przekształcania zmiennych numerycznych na ich odpowiedniki kategorialne lub przedziałowe. Zatem, kubełkowanie przekształca cechę z wartościami ciągłymi w cechę kategorialną lub przedziałową.



## Kubełkowanie

Główne techniki:

- Kubełkowanie o stałej szerokości (*ang. equal width (distance) binning*) – dzieli zakres danych na przedziały o ustalonej, równej szerokości. Nie należy stosować do rozkładów skośnych.
- Kubełkowanie o stałej częstotliwości (*ang. equal frequency (depth) binning, quantile binning*) – dane są rozdzielane na kubełki w taki sposób, aby każdy kubełek zawierał mniej więcej taką samą liczbę danych. Ta metoda może skutecznie radzić sobie z wartościami odstającymi i skośnymi danymi.

# Po co nam normalizacja?

Celem normalizacji (*ang. normalization*) jest doprowadzenie zmiennych do porównywalności. Uzyskuje się to poprzez pozbawienie mian wyników pomiaru oraz ujednolicenie ich rzędów wielkości. Pierwszy cel normalizacji jest jednoznaczny. Cel drugi nie jest jednoznaczny, a zatem dopuszcza w tym zakresie różne rozwiązania. Ogólnie rzecz biorąc ujednolicenie rzędów wielkości uzyskuje się przez wprowadzenie jednolicie określonej wartości zerowej dla wszystkich zmiennych, a następnie przeskalowanie wartości zmiennych. Wszystkie wzory normalizacyjne są uważane za przekształcenia liniowe i powinny być zaimplementowane do zmiennych mierzonych na skalach mocnych (przedziałowa i ilorazowa).

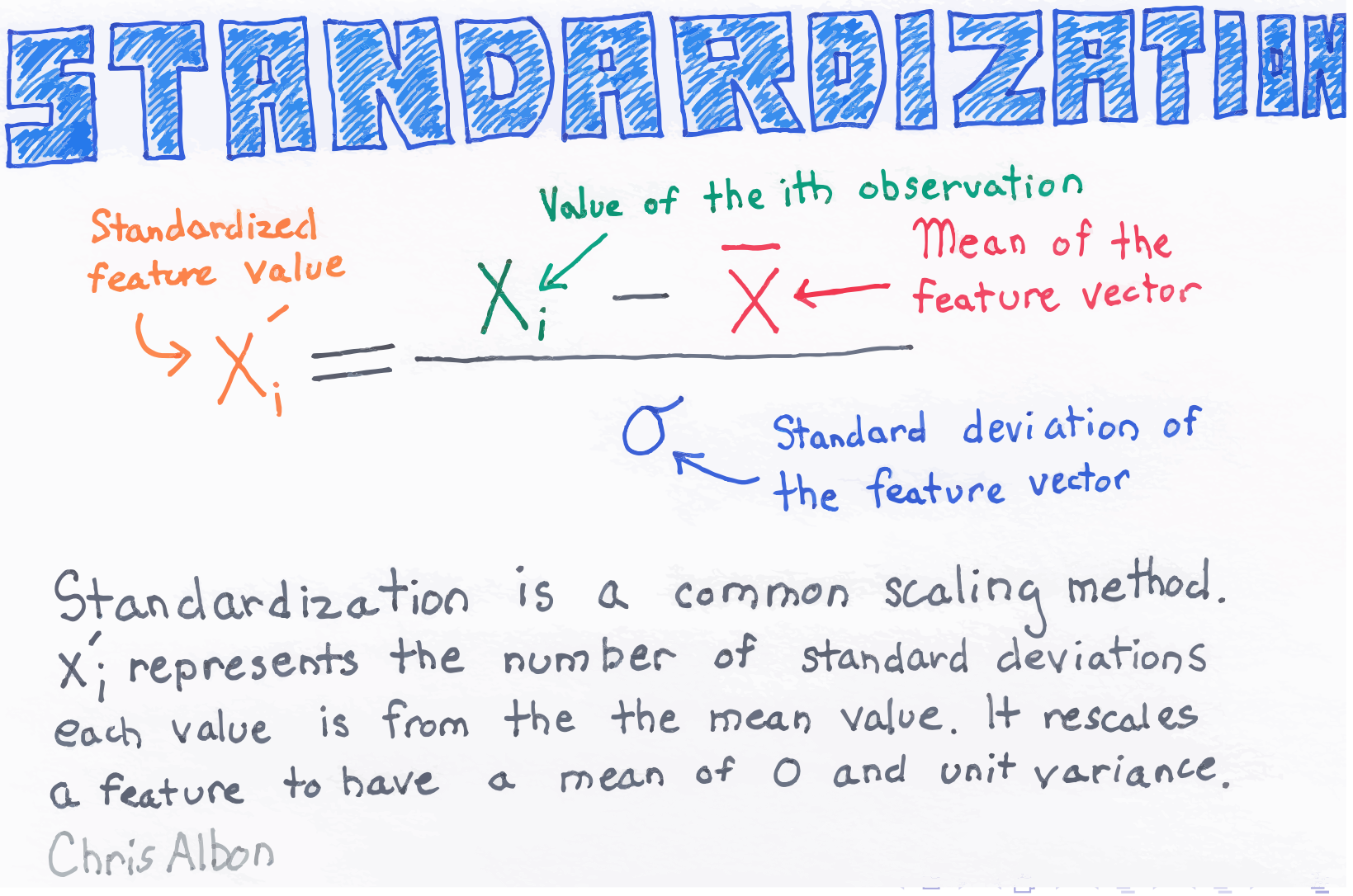
## Standaryzacja

Najczęściej spotykanym sposobem normalizacji jest standaryzacja (*ang. standardization, z-score normalization*). Celem standaryzacji zmiennej jest modyfikacja rozkładu tak aby miał wartość oczekiwaną 0 i odchylenie standardowe 1:

$$z_i = \frac{x_i - \bar{x}}{s},$$

gdzie $\bar{x}$ oznacza średnią, a $s$ odchylenie standardowe. Przekształcenie takie jest liniowe i monotoniczne oraz nie zmienia kształtu rozkładu. Po standaryzacji otrzymujemy wskaźnik, który mówi nam o odległości od średniej (0 w tym przypadku) wyrażonej w odchyleniach standardowych (1 w tym przypadku). Zatem np. $z = 2.5$ oznacza, że jesteśmy w odległości 2.5 odchylenia standardowego od 0.

# Standaryzacja



## Unitaryzacja

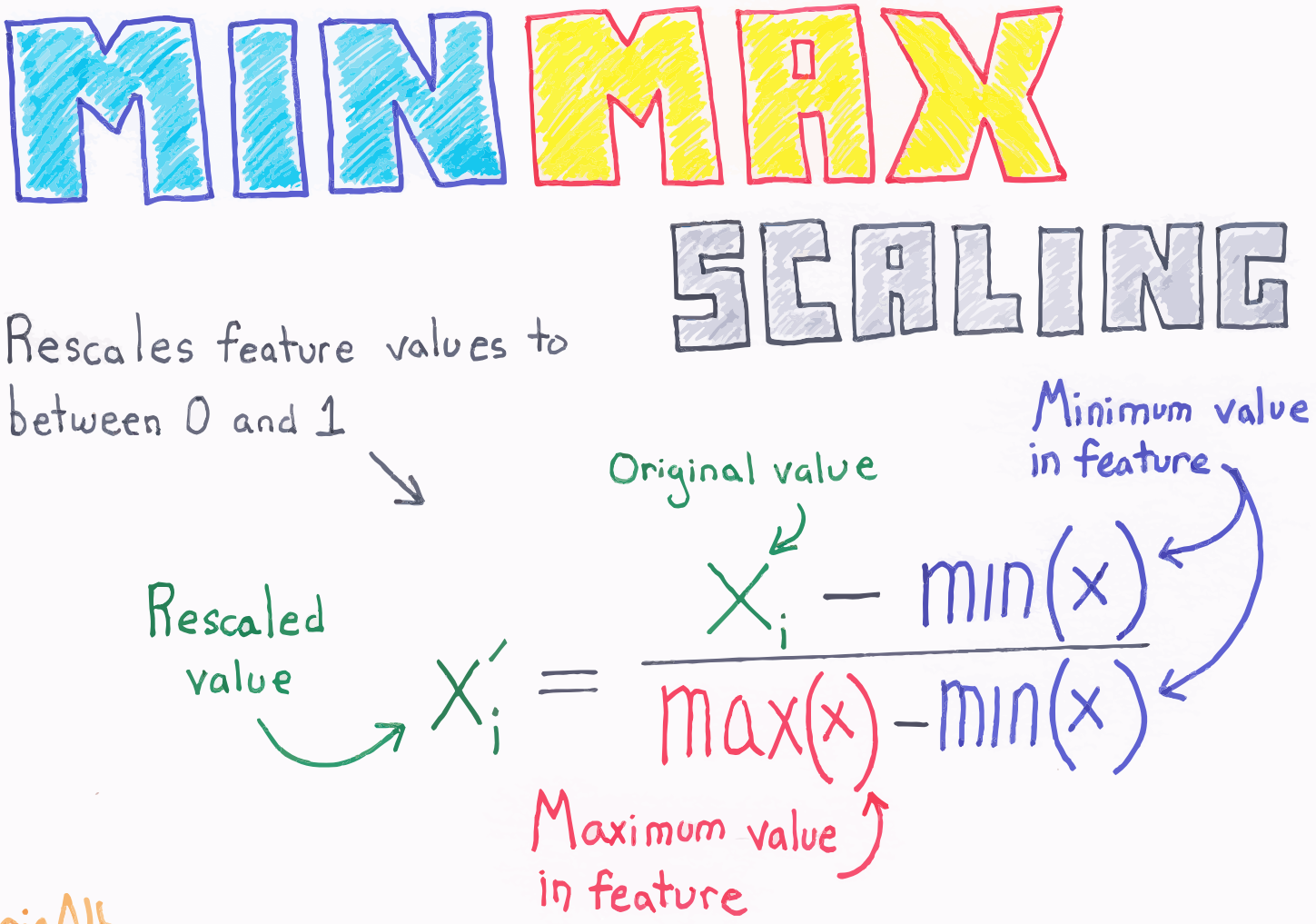
Celem unitaryzacji (*ang. unitarization, unity-based normalization, min-max feature scaling*) jest uzyskanie zmiennych o ujednoliconym zakresie zmienności, definiowanym przez różnicę pomiędzy ich wartościami maksymalnymi i minimalnymi

$$z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}.$$

W wyniku zastosowania powyższej formuły otrzymujemy zmienne o wartościach należących do przedziału $[0, 1]$. Powyższą formułę można przekształcić tak, aby transformować dane do dowolnego odcinka $[a, b]$:

$$z_i = a + \frac{(x_i - x_{\min})(b - a)}{x_{\max} - x_{\min}}.$$

# Unitaryzacja



# Standaryzacja vs unitaryzacja

- Jeśli dane nie mają rozkładu normalnego używana jest zazwyczaj unitaryzacja. Dla rozkładów normalnych preferowana jest standaryzacja.
- Wartości zmiennej po unitaryzacji są zazwyczaj w przedziale $[0, 1]$. Po standaryzacji nie są ograniczone.
- Unitaryzacja jest silnie wrażliwa na obserwacje odstające, podczas gdy standaryzacja jedynie w nieznacznym stopniu.
- Unitaryzacja jest używana gdy algorytm nie ma założeń, co do rozkładu. W przeciwnym razie używana jest standaryzacja.

# Standaryzacja vs unitaryzacja