

An introduction to Statistical Learning

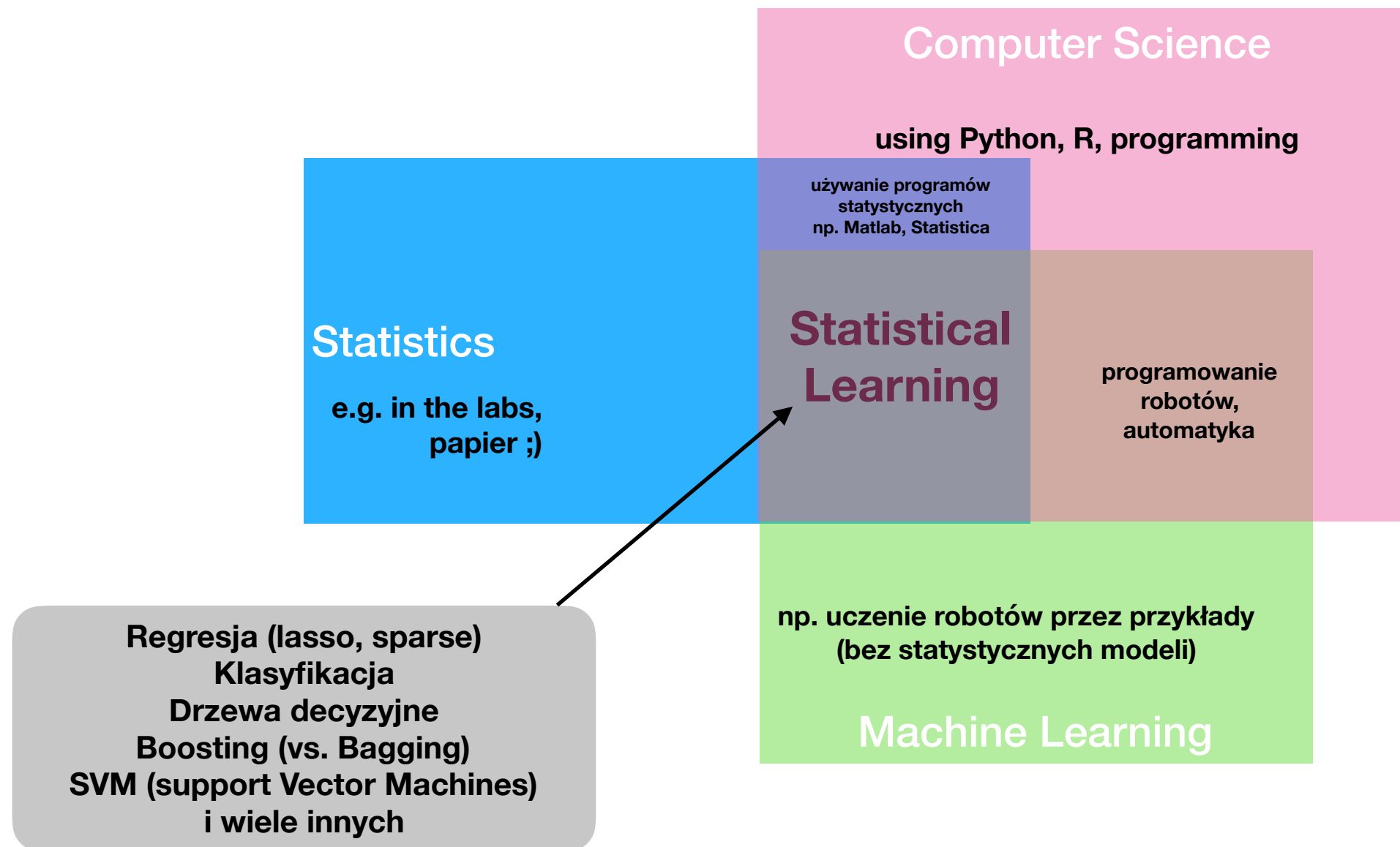
Lila Gmerek / Week 1 / 29.04.2019

Links

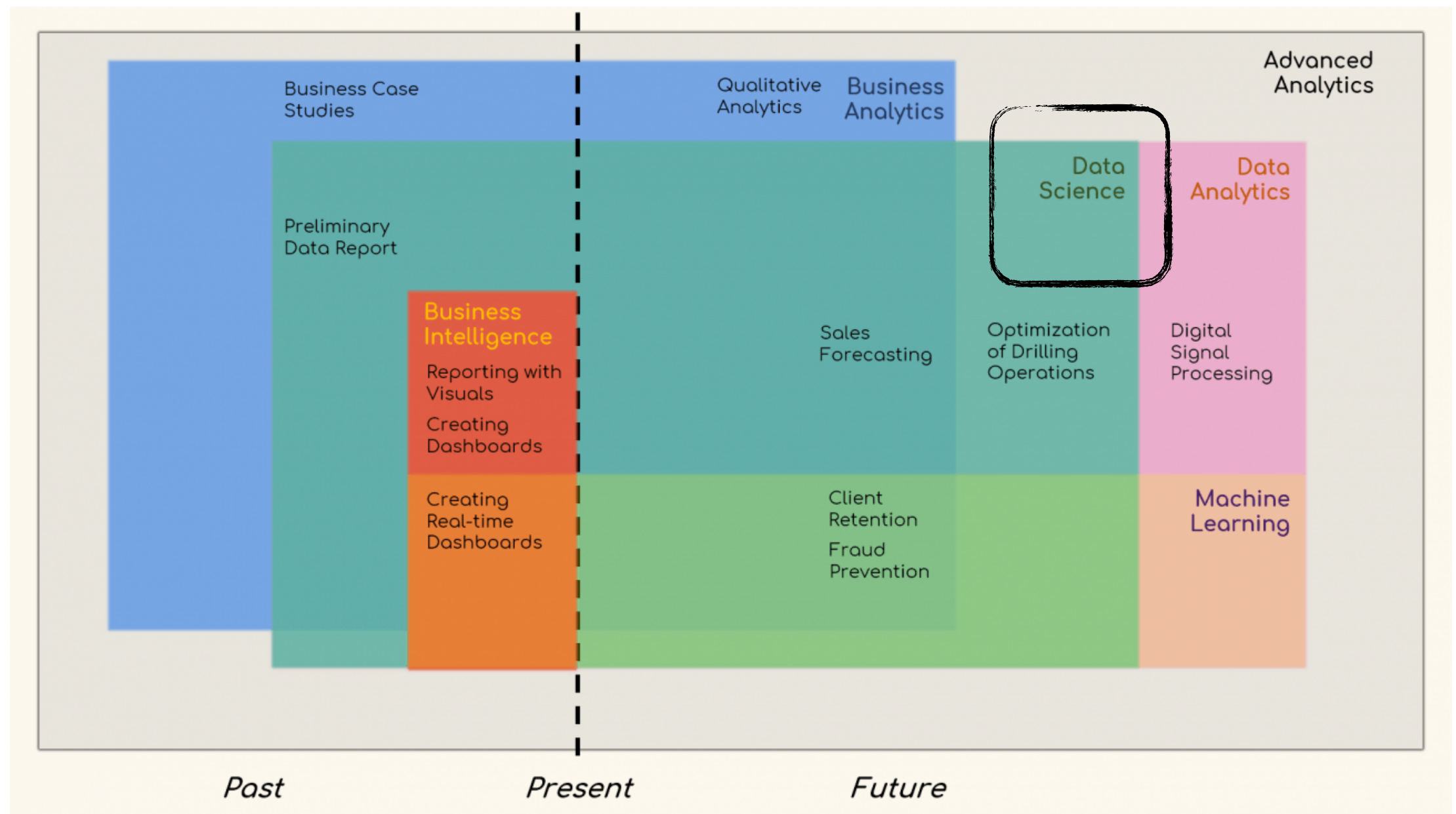
- Book <http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf>
- Github: <https://github.com/erg0-0/ML-study-group-pl-fc>

Introduction. An Overview of Statistical Learning

Statistical learning refers to a set of tools for modeling and understanding complex datasets. It is a recently developed area in statistics and blends with parallel developments in computer science and, in particular, machine learning. The field encompasses many methods such as the lasso and sparse regression, classification and regression trees, and boosting and support vector machines.



Trochę inne ujęcie tych definicji - źródło to 365Data.com



<https://sintelix.com/blog/machine-learning-v-data-science-v-analytics/>

Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as supervised or unsupervised. Broadly speaking, supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs. Problems of this nature occur in fields as diverse as business, medicine, astrophysics, and public policy. With unsupervised statistical learning, there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data.

Narzędzia:

- uczenie nadzorowane
- uczenie nienadzorowane (uczenie zależności, struktur)

CEL:

- stworzenie modelu predykcyjnego/szacującego
- rezultat (*output*) wynika z danych wejściowych (1+ *input*)

Dyskusja:

- Jaka jest różnica między Machine Learningiem a Data Science?
- Czy Machine Learning to Deep learning czy Deep learning to machine learning?
- Co ma wspólnego big data (engineering) z machine learningiem?

- Deep learning Wikipedia: https://en.wikipedia.org/wiki/Deep_learning
- Machine Learning / Uczenie maszynowe: https://pl.wikipedia.org/wiki/Uczenie_maszynowe
- Data Science: https://en.wikipedia.org/wiki/Data_science
- Big Data: https://en.wikipedia.org/wiki/Big_data

Przykłady zastosowania Machine Learning, które będą pojawiać się w podręczniku

1. WAGE DATA - analiza danych na temat zarobków - REGRESJA

Cel analizy: ZROZUMIEĆ DANE i przewidywać przyszłość :) analiza_czynnikow, ktore wplywaja na wysokosc zarobków (wage) grup mężczyzn z regionu atlantyckiego w US.

Featery/cechy analizowane / input : wiek (age), edukacja (education), rok(year) → wplyw na zarobki (wage)

Natura problemu: regresja

Oczekiwany output: (procentowy) wplyw danego czynnika na zarobki / prawdopodobieństwo etc.

UWAGA! mamy już jakieś dane dla konkretnej grupy z PRZESZŁOŚCI! Chcemy to przewidywać w PRZYSZŁOŚCI może też dla innych grup!

Wnioski na bazie wykresów (następny slajd, wykresy):

Czy lepiej analizować te czynniki oddzielnie, czy razem?

1. Wage jako funkcja wieku. Zarobki rosną wraz z wiekiem, ale spadają po 60ym roku zycia (fig. 1, niebieska linia)
2. Wage jako funkcja roku. Wzrost o ok. 10k\$ miedzy 2003-2009
3. Wage jako funkcja edukacji. Zarobki rosna wraz z poziomem edukacji.

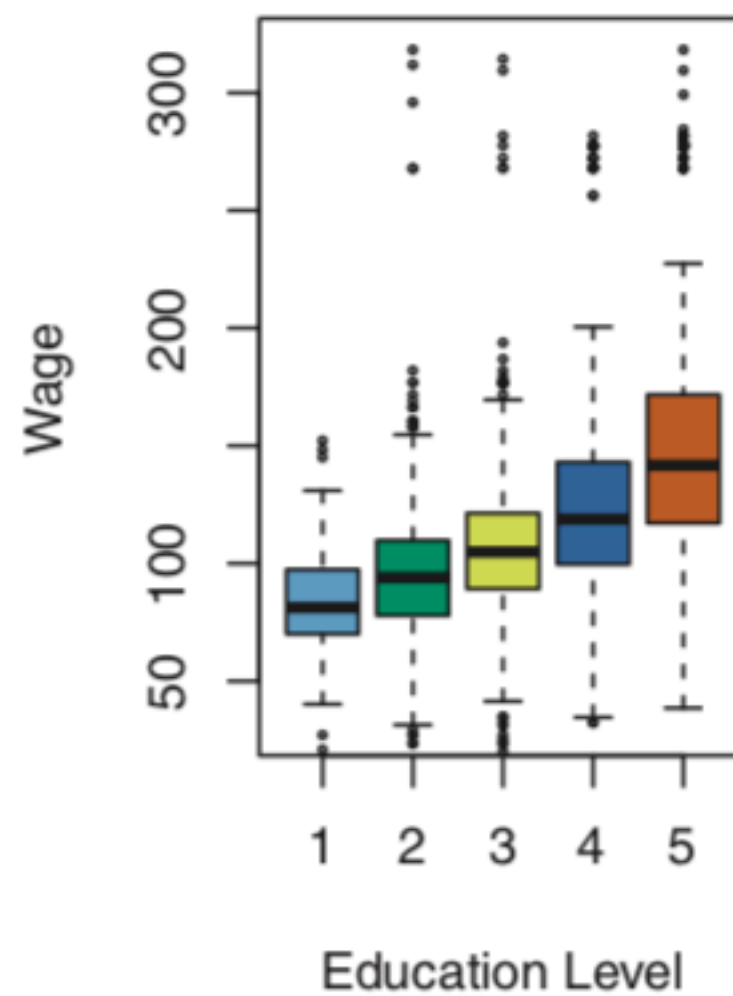
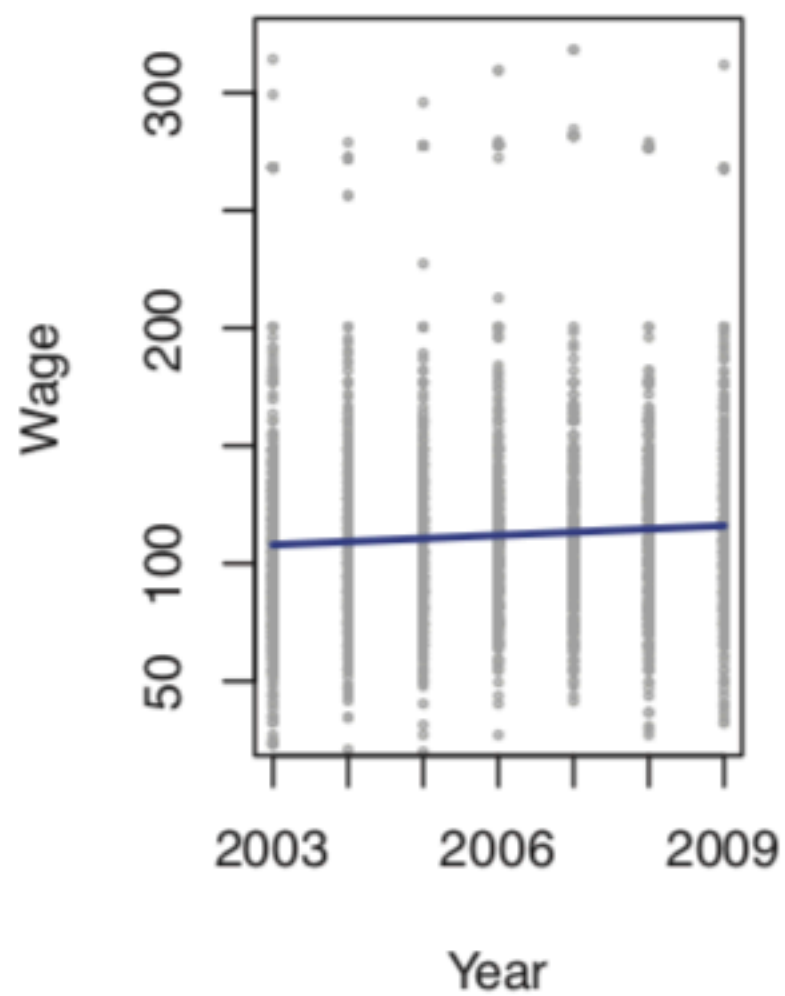
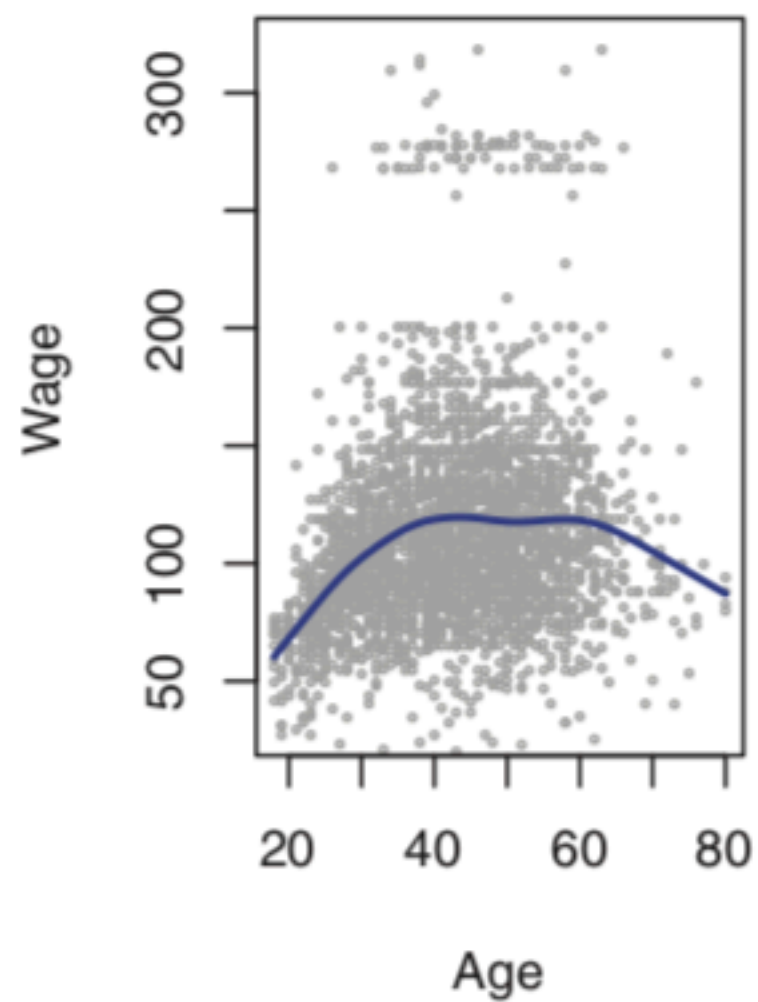
Odpowiedz: Najlepszą predykcję da kombinacja tych 3 czynników: age, education, year. Analizowanie ich oddzielnie może zaburzyć predykcję.

Prosta zależność liniowa jest widoczna tylko na wykresie wage vs. year. Potrzebny jest model nieco bardziej złożony.

Jakiego modelu predykcyjnego użyć?

Supervised learning - uczenie maszynowe nadzorowane

1. Rozdział 1 - regresja liniowa (widac zwlaszcza na wykresie year/wage)
2. Rozdział 7 - modelowanie nie-liniowe, regresja wielomianowa , GAMs itp.



Kiedy problem zarobków jest natury regresyjnej?

The Wage data involves predicting a continuous or quantitative output value.

Wartości predykowane są liczbami ciągłymi, a zasadniczo po prostu mają wartość numeryczną, ciągłą (tzw. float).

- jaka jest prawdopodobieństwo, że polski klient kupi nowe płatki naszej firmy ,WasabiFlakes' ?
- jaka jutro będzie temperatura powietrza? jaka będzie cena nieruchomości w Bostonie?
- o ile wzrośnie/spadnie sprzedaż płatków Wasabi Flakes w następstwie celowej kampanii marketingowej?

Kiedy ten problem jest natury klasyfikacyjnej?

Jeśli chcesz przewidywać wartości nie-numeryczne - kategoryczne albo jakościowe (categorical or qualitative)

- czy klient kupi WasabiFlakes?(tak/nie) , czy pacjent będzie miał raka? (tak/nie)
- czy jutro będzie padać? (tak/nie); czy na zdjęciu jest kwiat irys setosa, versicolor czy virginica? (grupowanie 1,2 czy 3?)
- do których klientów zaadresować kampanię płatków WasabiFlakes? [klastryzacja, grupowanie]

2. Stock Market Data (Smarket Data)

Cel predykcji: *The goal is to predict whether the index will increase or decrease on a given day using the past 5 days' percentage changes in the index.* (nie przewiduje się wartości numerycznej, ale kategorię)

Natura problemu: klasyfikacja

Featury/cechy analizowane/ input: indeks giełdowy, data (dzień)

Oczekiwany output: czy market performance(obserwacja) idzie w górę <Up bucket> czy w dół <Down bucket>

Uwaga! Mamy dane historyczne! Chcemy to przewidzieć w przyszłości!

Wnioski na bazie wykresów:

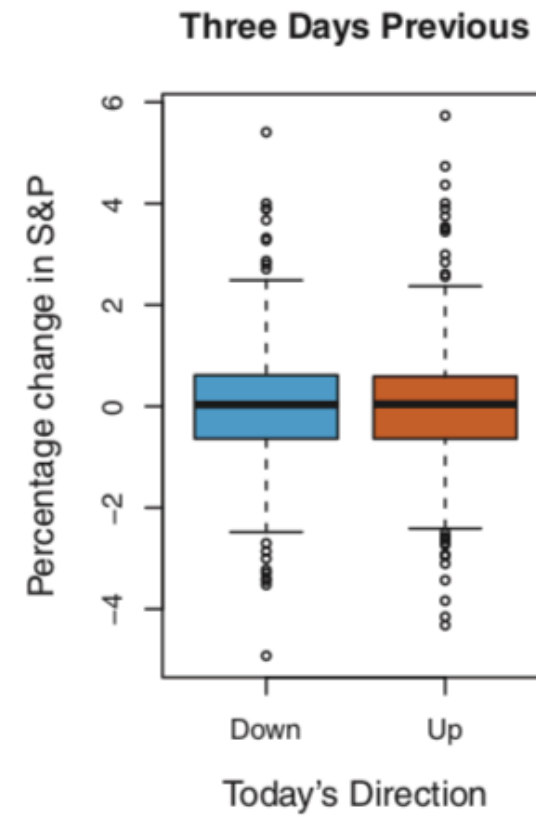
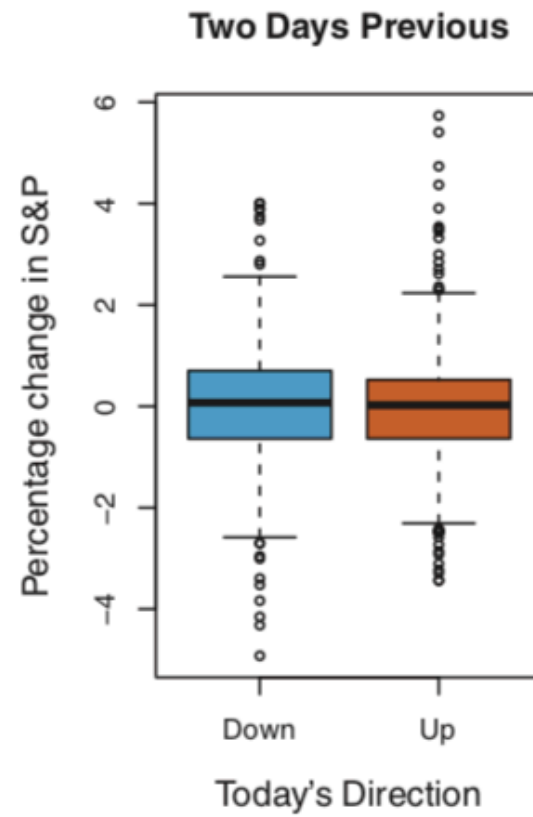
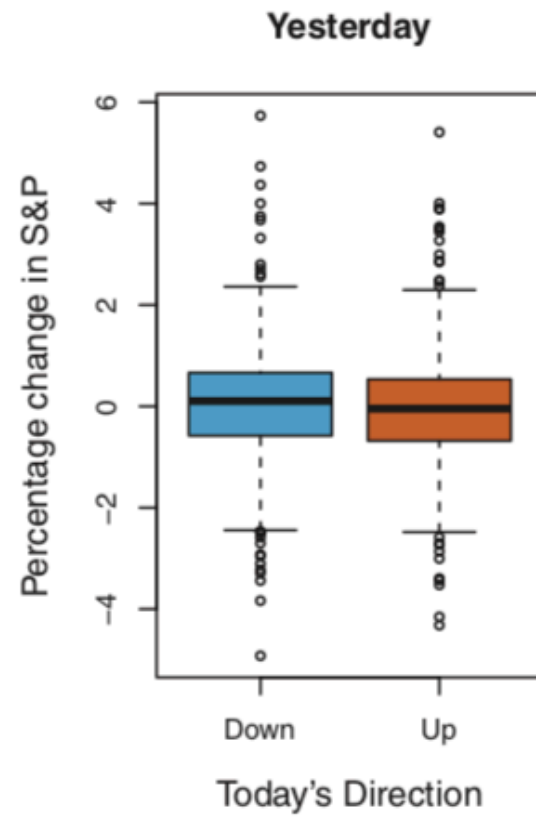
1. Nie ma prostej zależności między poprzednim dniem(dniami) a wysokością indeksu.

Gdyby była, gra na giełdzie nie byłaby taka trudna, a wszyscy byliby bogaci ;-).

2. Są modele, które oceniają z 60% dokładnością trend indeksu w cyklu 5-letnim.

Jakiego modelu predykcyjnego użyć?

Rozdział 4 - metody statystyczne klasyfikacyjne - Supervised Learning



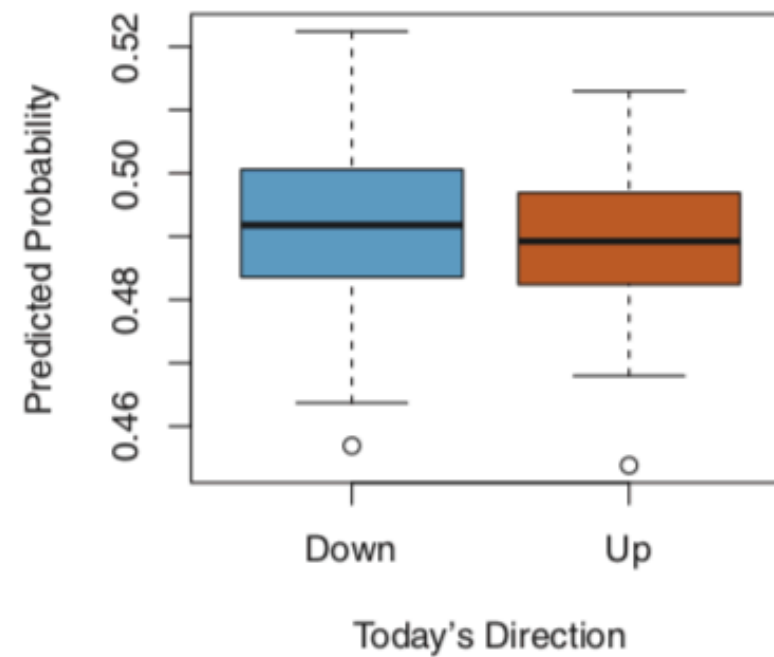


FIGURE 1.3. We fit a quadratic discriminant analysis model to the subset of the **Smarket** data corresponding to the 2001–2004 time period, and predicted the probability of a stock market decrease using the 2005 data. On average, the predicted probability of decrease is higher for the days in which the market does decrease. Based on these results, we are able to correctly predict the direction of movement in the market 60% of the time.

3. Gene Expression Data - NCI60

Cel predykcji:

In a marketing setting, we might have demographic information for a number of current or potential customers. We may wish to understand which types of customers are similar to each other by grouping individuals according to their observed characteristics.

Dataset NCI60 - we are interested in determining whether there are groups, or clusters, among the cell lines based on their gene expression measurements

Natura problemu: (klasyfikacja) **clustering - NIE predykcja, analiza teraźniejszości, NIE przyszłości**

Featery/cechy analizowane/ input: dane o klientach (X cech m.in. historia zakupów itd.)

Dataset NCI60 - 6.830 obserwacji 64 linii komórek rakowych(? cancer cell lines)

Oczekiwany output: algorytm ma pogrupować klientów na podstawie najlepszych znalezionych zależności

Algorytm ma pogrupować cell lines na podstawie ich pomiarów ekspresji genetycznej. Nie wiemy jak - gdybyśmy wiedzieli - nie trzeba by tworzyć modelu. Danych jest zbyt dużo!

Uwaga! NIE MA wartości oczekiwanej! Nie wiemy jak pogrupować klientów - algorytm sam ma na to „wpaść”!

Wnioski na bazie wykresów NCI60:

1. Aby przedstawić dane na wykresie zastosowano redukcję wymiarów (dimentions) do dwóch (fig. 1.4), co umożliwiło podział na 4 główne grupy. Każdy punkt to jeden z 64 cell lines
2. Drugi wykres to pierwszy + uwzględnienie 14 rodzajów raka za pomocą koloru+kształtu. Widać pewne zależności np. konkretny rodzaj raka pojawia się blisko siebie w danym klastrze linii komórkowych.
3. Utrata informacji jest nieunikniona.

Jakiego modelu użyć?

Rozdział 10 - metody statystyczne klasyfikacyjne - Uczenie Maszynowe nienadzorowane (Unsupervised Learning)

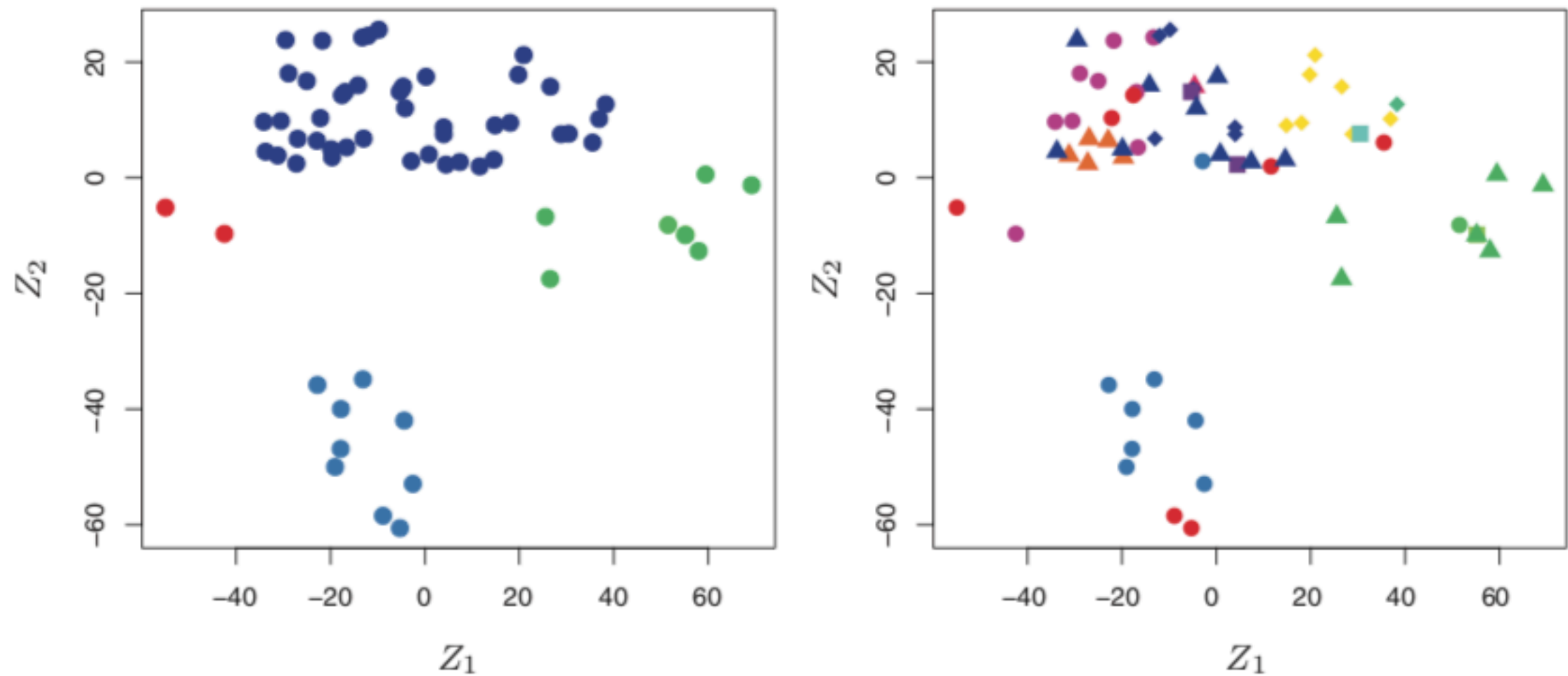
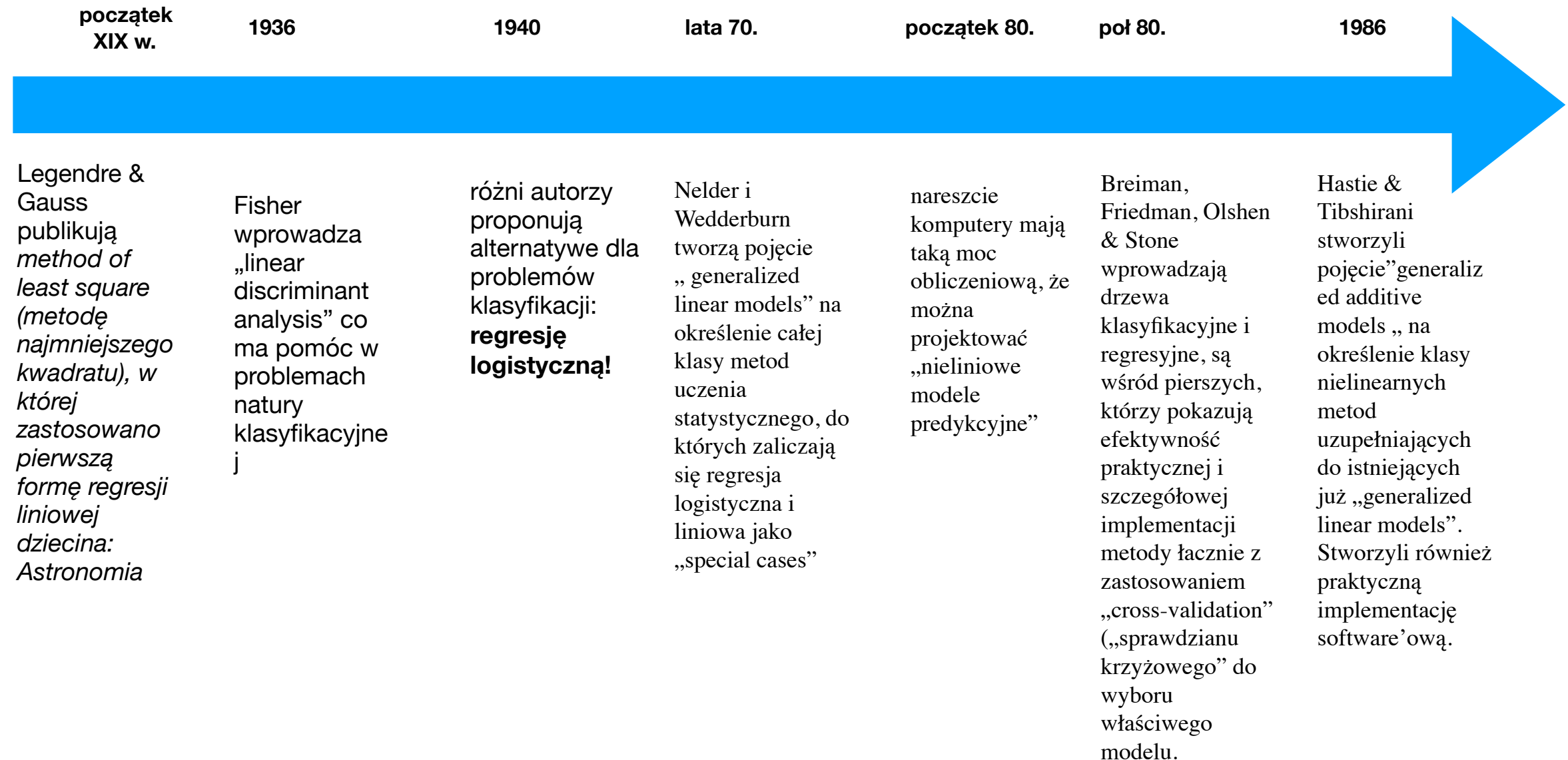


FIGURE 1.4. Left: Representation of the **NCI60** gene expression data set in a two-dimensional space, Z_1 and Z_2 . Each point corresponds to one of the 64 cell lines. There appear to be four groups of cell lines, which we have represented using different colors. Right: Same as left panel except that we have represented each of the 14 different types of cancer using a different colored symbol. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.

Brief History of Statistical Learning



Od tego czasu Statistical Learning, a potem i bardziej modne Data Science przeżywa rozkwit. Dostępność software’u oraz mocy obliczeniowych, a także przyjazność dla użytkownika nie stanowią już takiej przeszkody dla szerszej społeczności.