WWW.DAUPHINE.PSL.EU

# Portfolio sampling with K-means


Fabrice Riva

fabrice.riva@dauphine.psl.eu


Dauphine | PSL
UNIVERSITÉ PARIS

# Motivation

- Investors should hold a diversified portfolio
- Financial theory strongly supports diversification:
  - Diversification allows to eliminate idiosyncratic risk (more on this in next slides)
    - 15 stocks are enough on average to eliminate idiosyncratic risk
    - Number of stocks depends on country stock exchange characteristics and time periods
  - CAPM: no compensation in expected returns for idiosyncratic risk
  - Two-fund separation theorem (Tobin, 1958): for an investor, the optimal portfolio is a mix of the risk-free asset and the market portfolio
- The rise of ETFs (Exchange-Traded Funds) supports the view that investors seek to diversify their portfolios:
  - First ETF (SPY) launched in the United States in 1993. As of October 2022, 10,127 ETFs available worldwide (source : etfgi.com)
  - ETF assets under management (AuM) around $9,847 bn in October 2023

# Issues and proposed solution

- Well-diversified portfolios $\Rightarrow$ investors should hold a large number of stocks
- Issue: a well-diversified portfolio entails investors to manage a large number of assets. Management costs increase with the number of assets, which decreases the benefits of diversification
- Questions :
  - Is it possible to replicate the returns of a large portfolio with a small subset of its constituents?
  - How to select stocks within a portfolio in order to achieve best replication of that portfolio?
  - How many stocks should be selected?
- Proposed solution: *K*-means
  - Identification of clusters within the initial portfolio, i.e. groups of stocks whose return time-series are "similar"
  - For each cluster, identification of the stock that is the most representative of the whole cluster, i.e. identification of each cluster's centroid
  - Investment in the portfolio built from the various centroid stocks

# Data - I

- CRSP database
- Monthly data over June 2007 – December 2020: 168 months / 14 years
- 300 randomly chosen stocks that match the following criteria:
  - Data available from beginning to end of period
  - No change in activity sector of a firm over the period: used only to test if clusters match firm activity sectors
- Objective: replicate as closely as possible, with only a few stocks, the return time series of the 300-stock equally-weighted portfolio
- Train set: January 2007 – February 2018, i.e. 80% of the whole period. 134 months including 2008 global financial crisis
- Test set: March 2018 – December 2020, i.e. 20% of the whole period. 34 months including Covid 19 period

# Data - II

- Partial view of the data (first 5 observations)

| PERMNO | date | TICKER | RET | vwretd | sector_1dgt |
|---:|---:|---:|---:|---:|---:|
| 10104 | 2007-01-31 | ORCL | 0.001167 | 0.019387 | 7 |
| 10104 | 2007-02-28 | ORCL | -0.042541 | -0.014006 | 7 |
| 10104 | 2007-03-30 | ORCL | 0.103469 | 0.012954 | 7 |
| 10104 | 2007-04-30 | ORCL | 0.036955 | 0.039834 | 7 |
| 10104 | 2007-05-31 | ORCL | 0.030851 | 0.038953 | 7 |

- 300 stocks × 168 months = 50,400 observations in total

# Implementation

- Python + pandas
- matplotlib library for plots
- Scikit-learn API for $K$-means implementation (in reality $K$-means ++)
  - Computation of clusters
  - For each cluster, identification of cluster's centroid

# Agenda

- Stock, portfolio and risk basics
- K-means baseline algorithm
- K-means ++
- How to determine the number of clusters
- Applications
  - Iris dataset
  - Stock clustering: portfolio sampling with small number of stocks

# Stock and portfolio basics

# Stocks

- $P_{i,t}$: stock $i$ price at date $t$
- $R_{i,t}$: stock $i$ discrete return from date $t-1$ to date $t$. Ignoring dividends:

$$R_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}$$

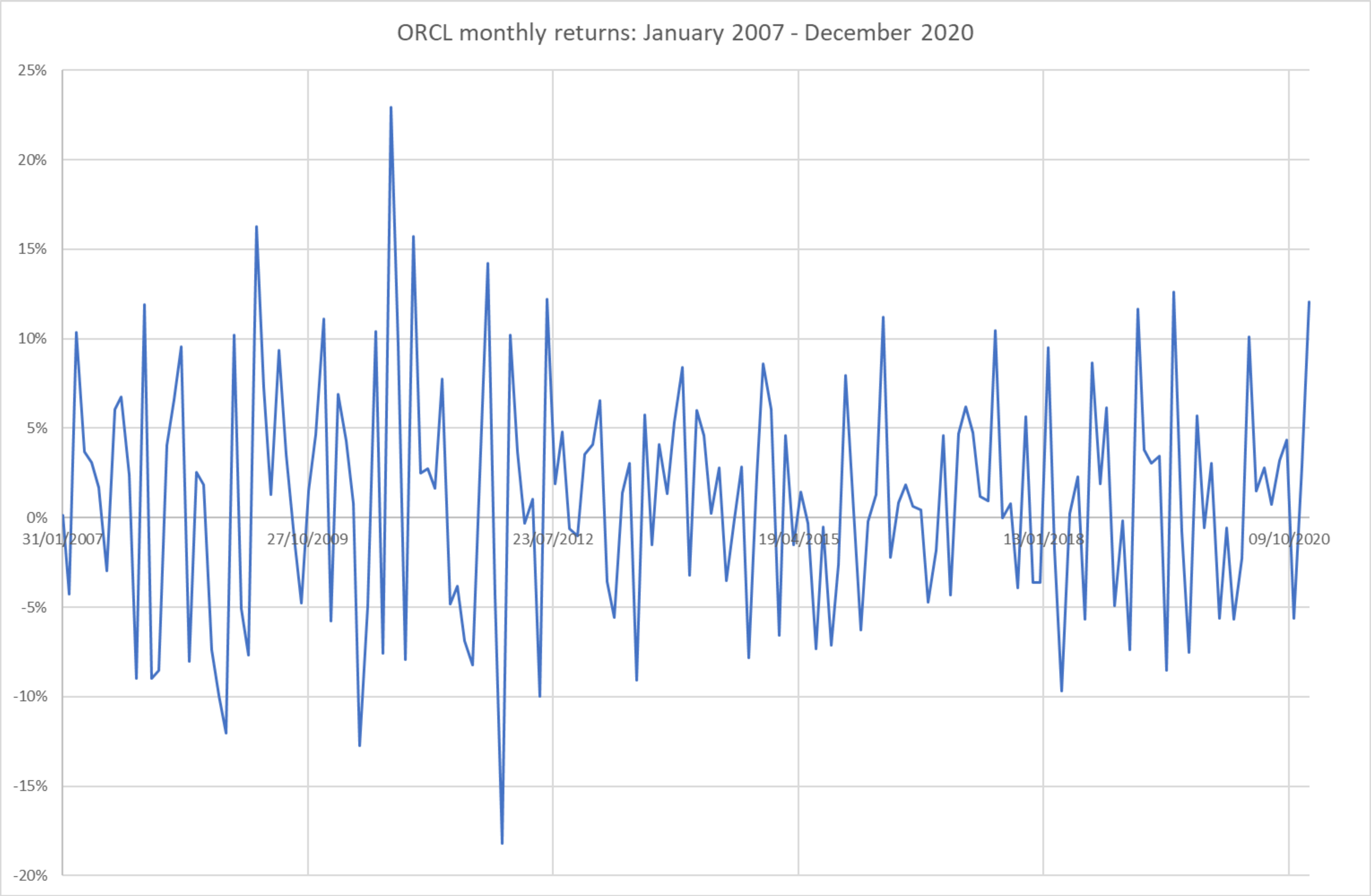
- Also, continuous return:

$$r_{i,t} = \log(1 + R_{i,t})$$

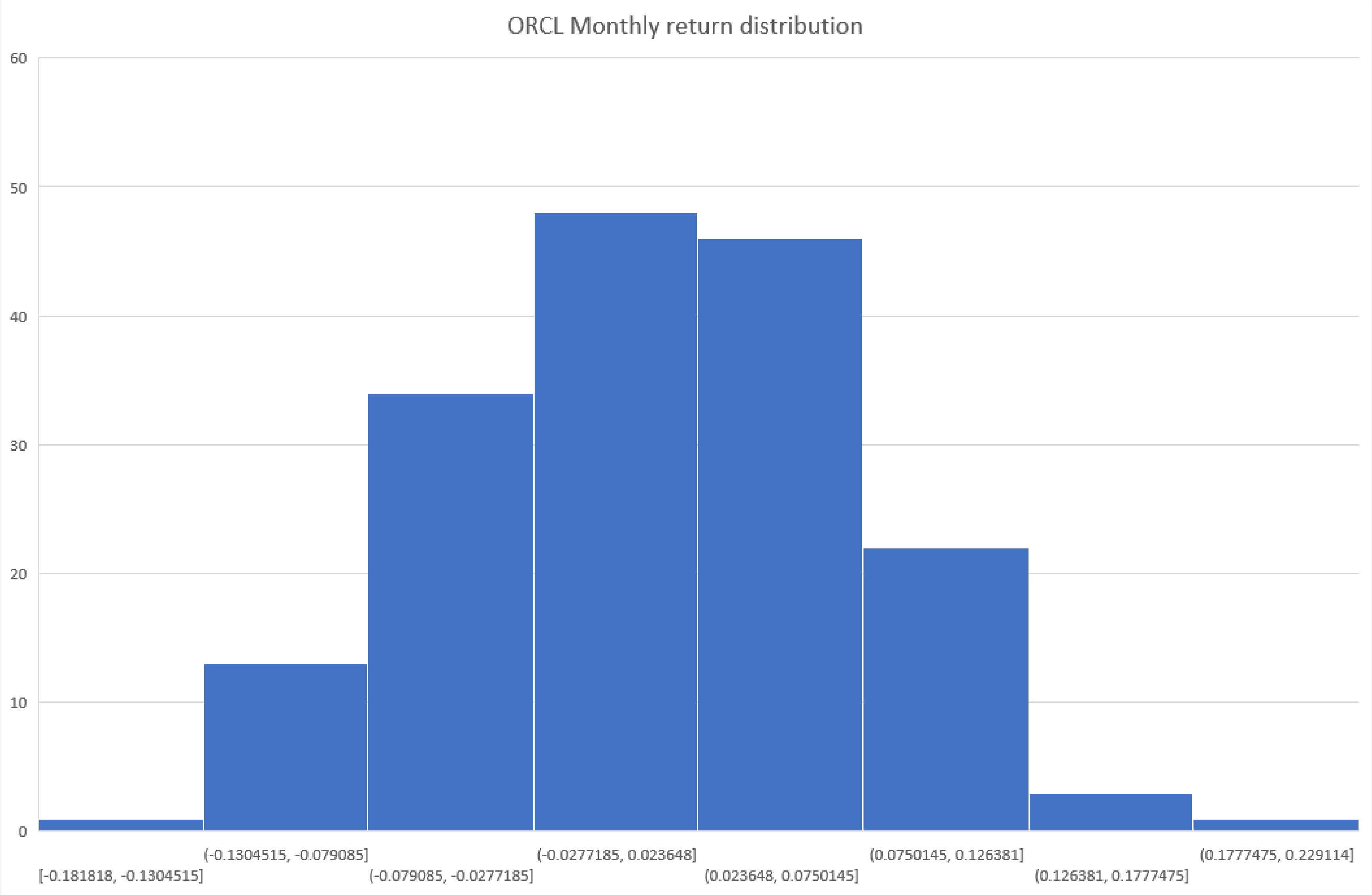
- Stock prices, and therefore returns, are highly unpredictable $\Rightarrow R_{i,t}$ is modeled as the outcome of a random variable
- For tractability reasons, returns are often modeled as the outcome of a stationary normal distribution:

$$R_{i,t} \sim \mathcal{N}(\mu_i, \sigma_i)$$

# Random behavior of stock returns: ORCL



# Random behavior of stock returns: ORCL



# Risk and portfolios

- One way to measure stock $i$ risk is $\sigma_i$. In finance, $\sigma_i$ (expressed on an annual basis) is called the volatility of stock $i$
- Investors can reduce the risk they bear by holding a portfolio of stocks instead of a single individual stock
- A portfolio is a combination of assets. In finance, portfolios are represented by a vector of weights $x = [x_1 \ x_2 \cdots x_n]^T$, where $x_j$ represents the weight of asset $j$ in portfolio $x$:
  - Weight $x_j$ corresponds to the dollar amount invested in stock $j$ divided by the total dollar amount invested in portfolio $x$
  - A portfolio must be feasible, which implies $\sum_{j=1}^{n} x_j = 1$
  - Weights can be **negative** if short selling is allowed: in that case, the investor borrows the stock to a lender (broker) and sells it to a trader. The investor will later purchase the asset in order to return it to the lender. If the price has fallen (increased) in the meantine, the investor will make a profit (loss) equal to the price drop (increase)

# Portfolio risk - I

- Denoting $\mu = [\mu_1 \ \mu_2 \ \cdots \ \mu_n]^T$ the vector of individual stocks mean returns, portfolio $x$ mean (or expected) return is computed as:

$$\mu_x = x^T \mu$$

- What about the risk of portfolio $x$?
- It is assumed that individual stock returns are correlated and that their cross-sectional dependance can be modeled using variance-covariance matrix $\Sigma$:

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \cdots & \sigma_{n,n} \end{bmatrix}$$

where $\sigma_{i,j}$ denotes the covariance of asset $i$ and asset $j$ returns

# Portfolio risk - II

- Using $\Sigma$, the variance of the returns of portfolio $x$ is given by the following expression:

$$\sigma_x^2 = x^T \Sigma x$$

- The above expression is equivalent to:

$$\sigma_x^2 = \underbrace{\sum_{i=1}^{n} x_i^2 \sigma_{i,i}}_{(a)} + \underbrace{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} x_i x_j \sigma_{i,j}}_{(b)} \tag{1}$$

- The $(a)$ term is the sum of the individual variances while the $(b)$ term is (twice) the sum of cross covariances

# Variance of an equally-weighted portfolio

- Equally-weighted portfolio means $x_1 = x_2 = \cdots = x_n = \frac{1}{n}$
- Assuming that none of the stocks has an infinite variance, $\lim_{n \to \infty} (a) = 0$ in $(1)$
  $\Rightarrow$ individual variances vanish in a well-diversified portfolio
- Let $\bar{\sigma}$ be the average covariance of the $n$ stocks. Clearly:

$$\bar{\sigma} = \frac{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \sigma_{i,j}}{n(n-1)}$$

- $(b)$ in $(1)$ can be rewritten as $\frac{n(n-1)\bar{\sigma}}{n^2}$ , so that $\lim_{n \to +\infty} (b) = \bar{\sigma}$
- Conclusion: the variance of portfolio $x$ tends to approach the average covariance of stocks in $x$
- **Portfolio risk** of an asset: covariances matter more than variances

# Market model - I

- The market model (Sharpe, 1963) aims at decomposing an asset's returns
- According to the model, changes in an asset prices, hence in its returns, are the outcome of:
  - Marketwide information (e.g. macroeconomic events, business climate, etc.) . Marketwide information has a **systematic** impact, hence affects stocks as a whole. However, each stock reacts differently based on its own **sensitivity** to this type of information
  - Company specific (idiosyncratic, non-systematic) information. These are stock specific movements which are **not correlated with the rest of the market**
- From the above, the return on stock $i$ at date $t$ can be decomposed as:

Average return on the stock if $R_{m,t} = 0$

Stock sensitivity to market

$$R_{i,t} = \alpha_i + \beta_i R_{m,t} + \epsilon_{i,t}$$

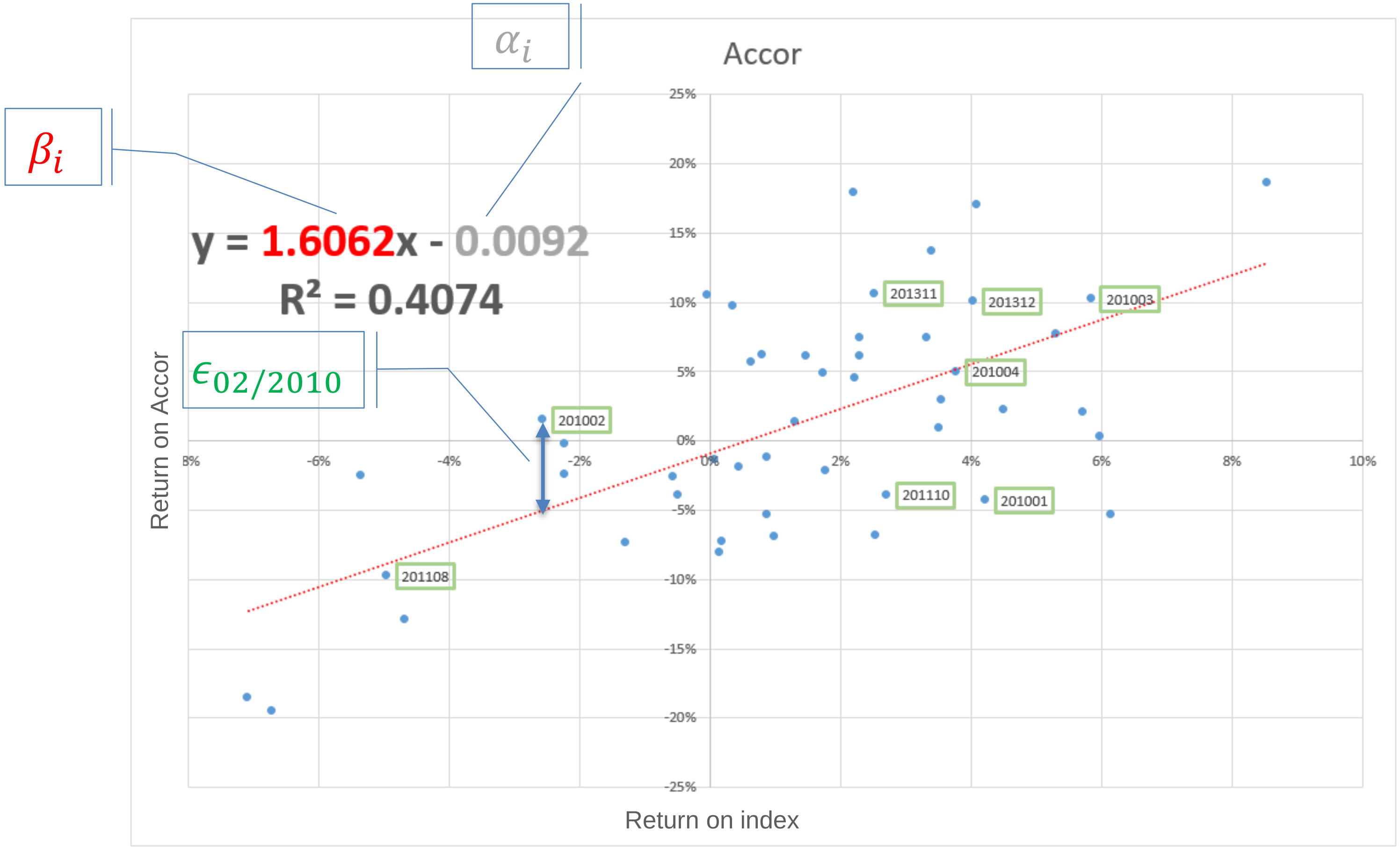
Stock return

Market return

Stock specific return component

# Market model - II

- Estimating parameters $\alpha_i$ and $\beta_i$ requires isolating the two components of the stock's returns, i.e. one component correlated with the market and the other independent / orthogonal to the market
- This is achieved by using OLS, where the rate of return on the stock is regressed on the market rate of return (practically a general stock market index)

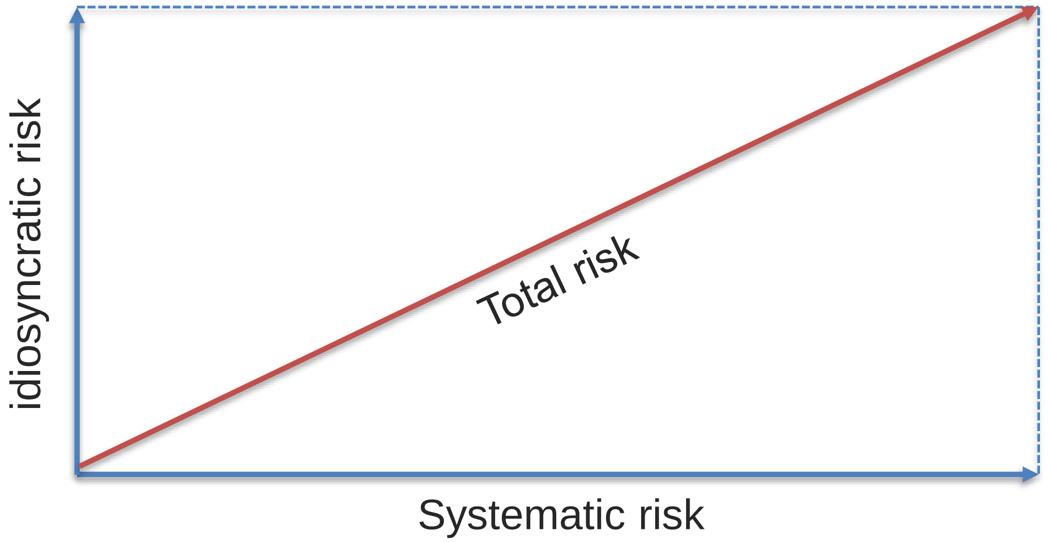


# Risk decomposition from market model

- The variance of the returns of a stock based on market model is given by:

$$\begin{aligned} \sigma_i^2 &= \sigma^2(\alpha_i + \beta_i R_m + \epsilon_i) \\ &= \beta_i^2 \sigma_m^2 + \sigma_{\epsilon_i}^2 \end{aligned}$$

- The $\beta_i^2 \sigma_m^2$ term corresponds to (square of) systematic risk
- The $\sigma_{\epsilon_i}^2$ term corresponds to the (square of) idiosyncratic or diversifiable risk



# Elimination of idiosyncratic risk - I

- The $\beta$ of a $n$-stock portfolio is given by:

$$\begin{aligned}\beta_x &= \frac{cov(R_x, R_m)}{\sigma_m^2} = \frac{cov(\sum_{i=1}^n x_i R_i, R_m)}{\sigma_m^2} = \sum_{i=1}^n x_i \frac{cov(R_i, R_m)}{\sigma_m^2} \\ &= \sum_{i=1}^n x_i \beta_i\end{aligned}$$

- Therefore, portfolio $x$'s $\beta$ is the **weighted sum** of the $\beta$s of portfolio $x$ constituents
- From market model, the variance of the returns of portfolio $x$ is given by:

$$\begin{aligned}\sigma_x^2 &= \sigma^2\left[\sum_{i=1}^n x_i R_i\right] \\ &= \sigma^2\left[\sum_{i=1}^n x_i(\alpha_i + \beta_i R_m + \epsilon_i)\right] \\ &= \sigma^2[\sum_{i=1}^n x_i \beta_i R_m] + \sigma^2[\sum_{i=1}^n x_i \epsilon_i]\end{aligned}$$

# Elimination of idiosyncratic risk - II

- $\lim_{n \to +\infty} \sum_{i=1}^{n} x_i \beta_i = \beta_m = 1 \Rightarrow \lim_{n \to +\infty} \sigma^2 [\sum_{i=1}^{n} x_i \beta_i R_m] = \sigma_m^2$
- Assuming that the various $\epsilon_i$s are not cross-correlated, that each of them has a small variance, and that the $x_i$ weights are not too different, i.e. $x_i \approx \frac{1}{n} \forall i$:

$$\lim_{n \to +\infty} \sigma^2 \left[ \sum_{i=1}^{n} x_i \epsilon_i \right] \approx \lim_{n \to +\infty} \frac{1}{n^2} \sum_{i=1}^{n} \sigma_{\epsilon_i}^2 \approx 0$$

- Conclusion: for a well-diversified portfolio, the stock (company) idiosyncratic risk vanishes (hence its name diversifiable risk) and the variance (or volatility) of the portfolio converges to the variance (volatility) of the market portfolio

# CAPM

- CAPM (Capital Asset Pricing Model) is one of the most fundamental relationships in finance
- Discovered independently by Sharpe (Nobel Prize), Lintner and Mossin in the 60's
- CAPM establishes the relationship that connects the expected (future) return of an asset to its risk:

$$\mathbb{E}(R_i) = R_f + \beta_i \mathbb{E}(R_m - R_f)$$

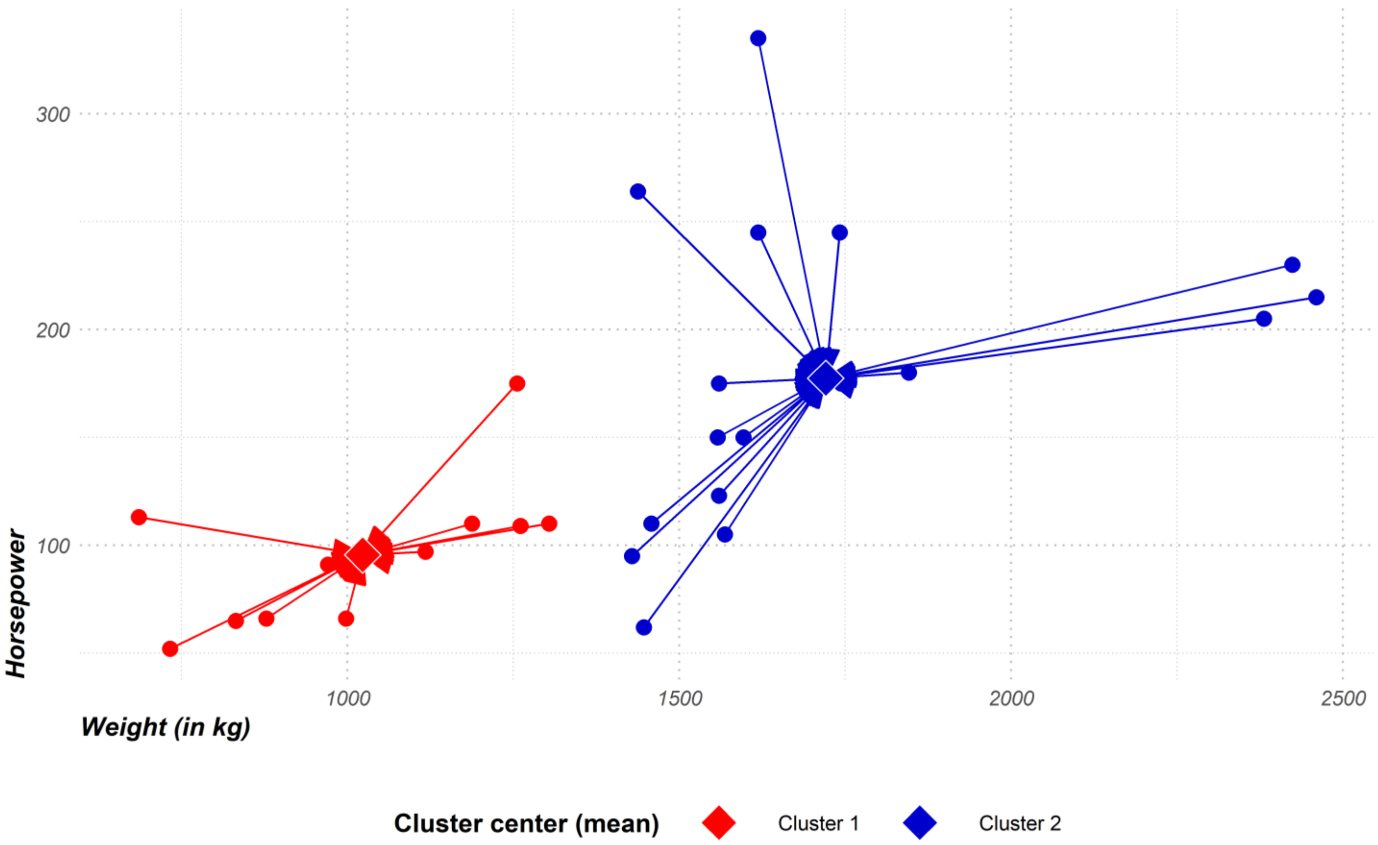
  where $R_f$ is the return of a risk-free asset (typically a long-term Government bond) and $\mathbb{E}(R_m)$ is the expected return of the market
- Takeaway for this course: Idiosyncratic risk does not show up in CAPM since $\beta_i$ matters only $\Rightarrow$ useless risk since it is not compensated by extra return

# $K$-means

# Baseline algorithm

# What is $K$-means?

$K$-means clustering is an unsupervised method that aims at partitioning $n$ observations into $K$ clusters. Each observation belongs to the cluster with the nearest mean, serving as a prototype to the cluster



# Definitions and notations

- $x_i \in \mathbb{R}^p, i \in \{1, \ldots, n\}$ are the observations we want to partition
- $\mu_k \in \mathbb{R}^p, k \in \{1, \ldots, K\}$ are the means, where $\mu_k$ is the center (or centroid) of cluster $k$. We will denote $\boldsymbol{M}$ the associated matrix
- $z_i^k$ are indicator variables associated to $x_i$ such that $z_i^k = 1$ if $x_i$ belongs to cluster $k$ and $z_i^k = 0$ otherwise. We will denote $\boldsymbol{Z}$ the matrix whose components are equal to $z_i^k$
- Finally we define the distorsion $J(\boldsymbol{M}, \boldsymbol{Z})$ by:

$$J(\mathbf{M}, \mathbf{Z}) = \sum_{k=1}^{K} \sum_{i=1}^{n} z_i^k ||x_i - \mu_k||^2$$

where $\|x_i - \mu_k\|^2 = \sqrt{\sum_{j=1}^{p} \left(x_{i,p} - \mu_{k,p}\right)^2}$

- Identifying the clusters means finding the $\boldsymbol{M}$ and $\boldsymbol{Z}$ matrices that minimize $J(\boldsymbol{M}, \boldsymbol{Z})$

# Example - I

- We make the following assumptions:
  - $x_1 = (5.1, 3.5, 1.4, 0.2)$, $x_2 = (4.9, 3.0, 1.4, 0.2)$, $x_3 = (7.0, 3.2, 4.7, 1.4)$
  - At some iteration $j$:
    - $\mu_1 = (5.7\ 3.8, 4.7, 1.2)$, $\mu_2 = (5.0, 3.4, 1.5, 0.2)$
    - $x_1$ and $x_3$ are assigned to cluster 1, and $x_2$ to cluster 2

- Expression of $x$:

$$\mathbf{X} = \begin{pmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3.0 & 1.4 & 0.2 \\ 7.0 & 3.2 & 4.7 & 1.4 \end{pmatrix}$$

- Expression of $\boldsymbol{M}$:

$$\mathbf{M} = \begin{pmatrix} 5.7 & 3.8 & 4.7 & 1.2 \\ 5.0 & 3.4 & 1.5 & 0.2 \end{pmatrix}$$

- Expression of $z$:

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

# Example - II

- Computing initial distorsion

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **X** | 5.1 | 3.5 | 1.4 | 0.2 | | **Z** | 1 | 0 | 1 | | | |
| 2 | | 4.9 | 3.0 | 1.4 | 0.2 | | | 0 | 1 | 0 | | | |
| 3 | | 7.0 | 3.2 | 4.7 | 1.4 | | | | | | | | |
| 4 | | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | |
| 6 | **M** | 5.7 | 3.8 | 4.7 | 1.2 | | | | | | | | |
| 7 | | 5.0 | 3.4 | 1.5 | 0.2 | | | | | | | | |
| 8 | | | | | | | | | | | | | |
| 9 | | | | | | **Distorsion** | | | | | | | $z_i^k$ |
| 10 | Cluster 1 | 0.36 | 0.09 | 10.89 | 1.00 | =(E1-E$6)^2 | | 3.51 | =SQRT(SUM(B10:E10)) | | | | 1 |
| 11 | | 0.64 | 0.64 | 10.89 | 1.00 | =(E2-E$6)^2 | | 3.63 | =SQRT(SUM(B11:E11)) | | | | 0 |
| 12 | | 1.69 | 0.36 | 0.00 | 0.04 | =(E3-E$6)^2 | | 1.45 | =SQRT(SUM(B12:E12)) | | | | 1 |
| 13 | Cluster 2 | 0.01 | 0.01 | 0.01 | 0.00 | =(E1-E$7)^2 | | 0.17 | =SQRT(SUM(B13:E13)) | | | | 0 |
| 14 | | 0.01 | 0.16 | 0.01 | 0.00 | =(E2-E$7)^2 | | 0.42 | =SQRT(SUM(B14:E14)) | | | | 1 |
| 15 | | 4.00 | 0.04 | 10.24 | 1.44 | =(E3-E$7)^2 | | 3.96 | =SQRT(SUM(B15:E15)) | | | | 0 |
| 16 | | | | | | | | | | | | | |
| 17 | | **Total distorsion** | | | | | | **5.3828** | =SUMPRODUCT(H10:H15;M10:M15) | | | | |

# The algorithm

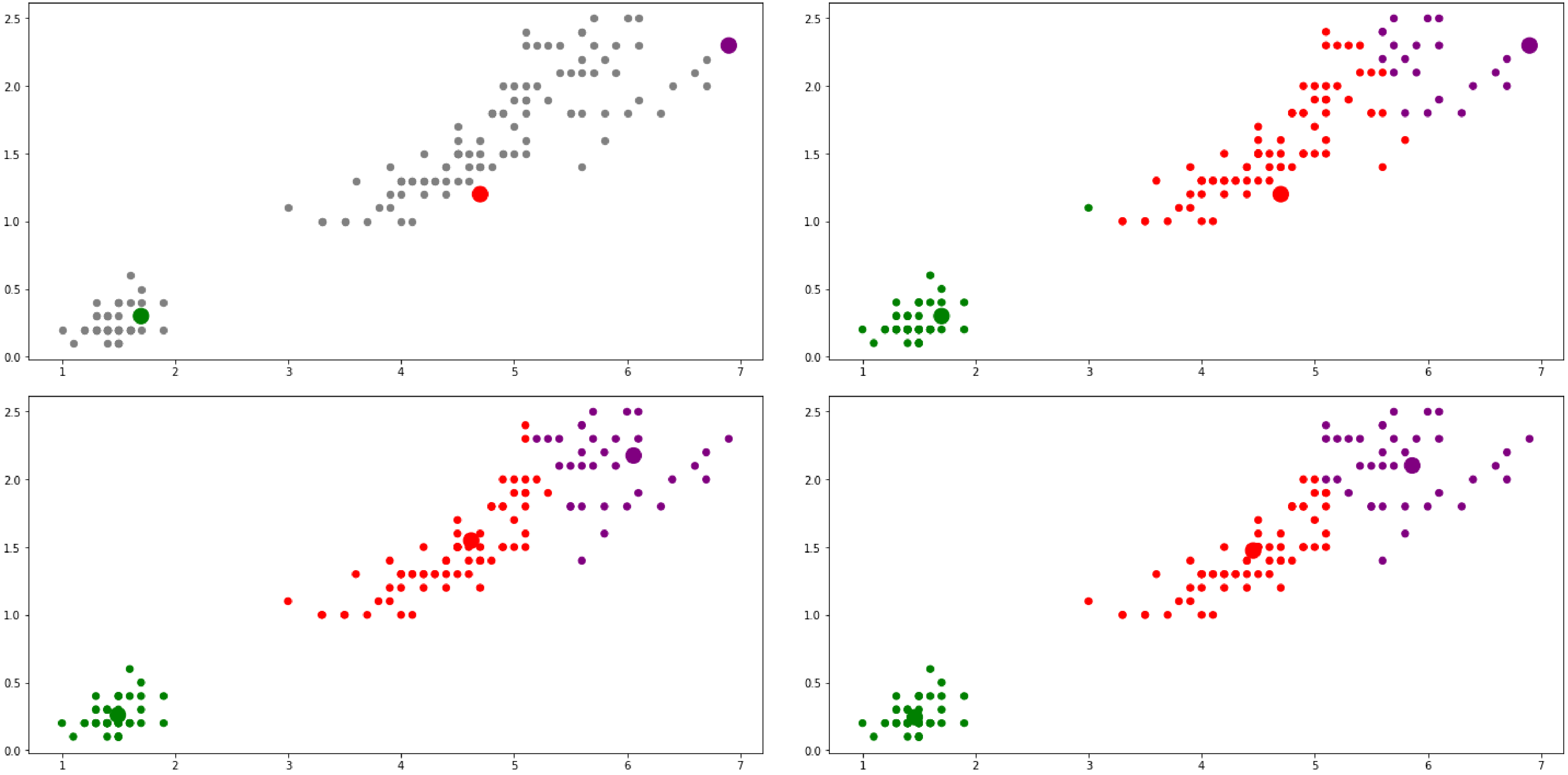
- It is possible to show that minimizing $J(\boldsymbol{M}, \boldsymbol{Z})$ can be achieved using the following algorithm (see for example Baukhage, 2015: https://arxiv.org/pdf/1512.07548.pdf):
  - Step 0: randomly initialize $\boldsymbol{M}$ by selecting randomly $K$ datapoints in the sample
  - Step 1: minimize $J$ with respect to $\boldsymbol{Z}$: $z_i^k = 1$ if $||x_i - \mu_k||^2 = \min_s ||x_i - \mu_s||^2$. In other words, $x_i$ must be associated to the nearest center $\mu_k$
  - Step 2: minimize $J$ with respect to $\mu$: $\mu_k = \frac{\sum_i z_i^k x_i}{\sum_i z_i^k}$
  - Step 3: come back to step 1 until convergence

# Example - III

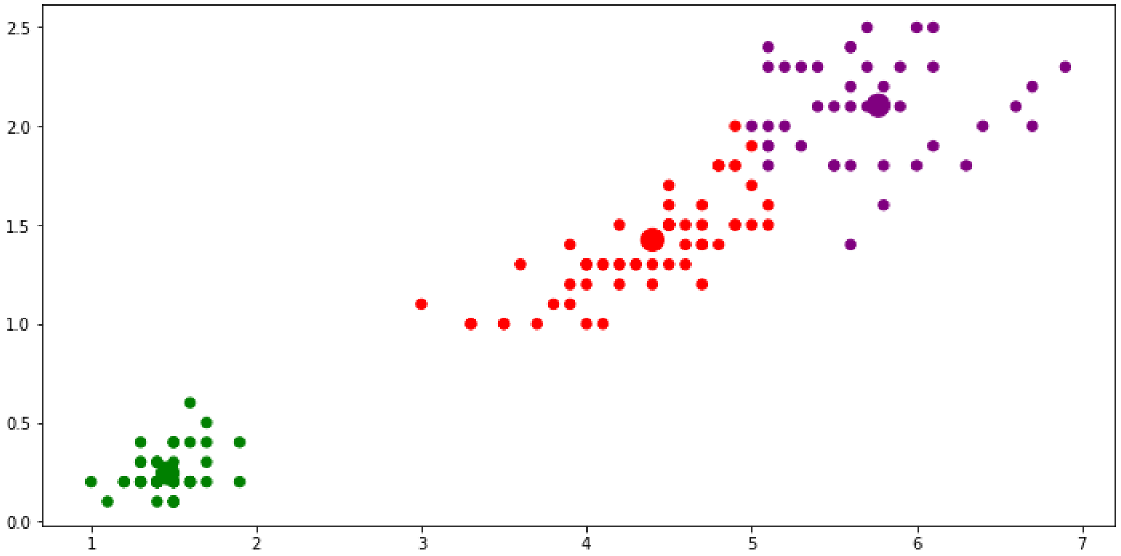
- Assignment to appropriate cluster based on previous distorsions and computation of new distorsion

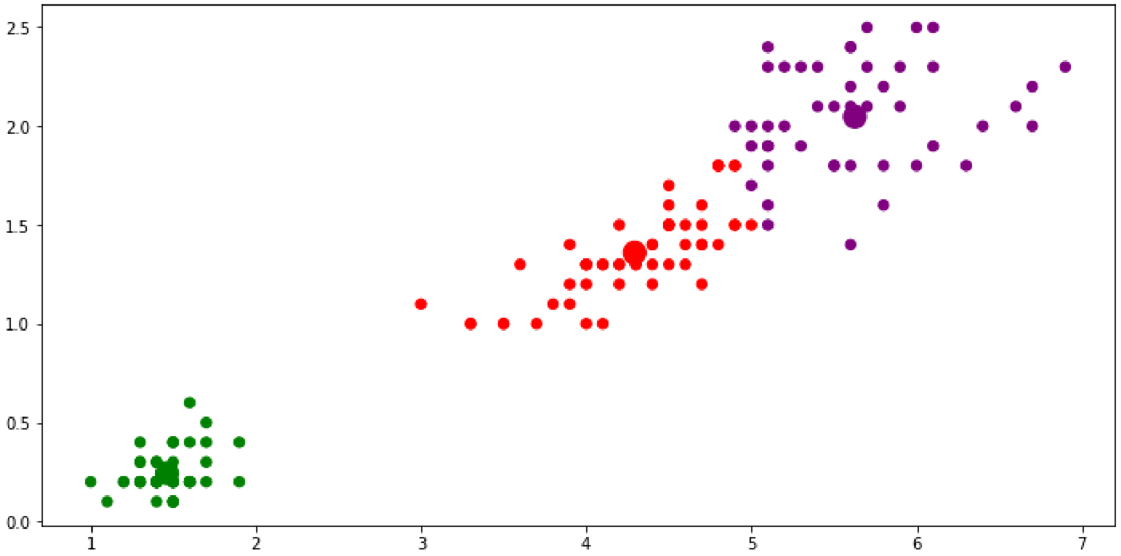
| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 19 | Assignment to nearest cluster -> new Z | | | | | | | | | | | | | |
| 20 | | | | | | | | | | | | | | |
| 21 | | **Z** | 0 | 0 | 1 | cluster 1 | | | | | | | | |
| 22 | | | 1 | 1 | 0 | cluster 2 | | | | | | | | |
| 23 | | | | | | | | | | | | | | |
| 24 | $z_i^k$ | | | | | **New centroid values** | | | | | | | | |
| 25 | 0 | | Cluster 1 | | 7 | 3.2 | 4.7 | 1.4 | =SUMPRODUCT(E1:E3;$A$25:$A$27)/SUM($A$25:$A$27) | | | | | |
| 26 | 0 | | Cluster 2 | | 5 | 3.25 | 1.4 | 0.2 | =SUMPRODUCT(E1:E3;$A$28:$A$30)/SUM($A$28:$A$30) | | | | | |
| 27 | 1 | | | | | | | | | | | | | |
| 28 | 1 | | | | | | | | | | | | | |
| 29 | 1 | | | | | | | | | | | | | |
| 30 | 0 | | | | | | | | | | | | | |
| 31 | | | | | | | | | | | | | | |
| 32 | | | | | | | **Distorsion** | | | | | | | |
| 33 | Cluster 1 | 3.61 | 0.09 | 10.89 | 1.44 | =(B1-E$25)^2 | | 4.00 | =SQRT(SUM(B33:E33)) | | | | | |
| 34 | | 4.41 | 0.04 | 10.89 | 1.44 | =(B2-E$25)^2 | | 4.10 | =SQRT(SUM(B34:E34)) | | | | | |
| 35 | | 0.00 | 0.00 | 0.00 | 0.00 | =(B3-E$25)^2 | | 0.00 | =SQRT(SUM(B35:E35)) | | | | | |
| 36 | Cluster 2 | 0.01 | 0.06 | 0.00 | 0.00 | =(B1-E$26)^2 | | 0.27 | =SQRT(SUM(B36:E36)) | | | | | |
| 37 | | 0.01 | 0.06 | 0.00 | 0.00 | =(B2-E$26)^2 | | 0.27 | =SQRT(SUM(B37:E37)) | | | | | |
| 38 | | 4.00 | 0.00 | 10.89 | 1.44 | =(B3-E$26)^2 | | 4.04 | =SQRT(SUM(B38:E38)) | | | | | |
| 39 | | | | | | | | | | | | | | |
| 40 | | **Total distorsion** | | | | | | **0.5385** | =SUMPRODUCT(H33:H38;A25:A30) | | | | | |

# $K$-means in action: Iris dataset - I



# $K$-means in action: Iris dataset - II

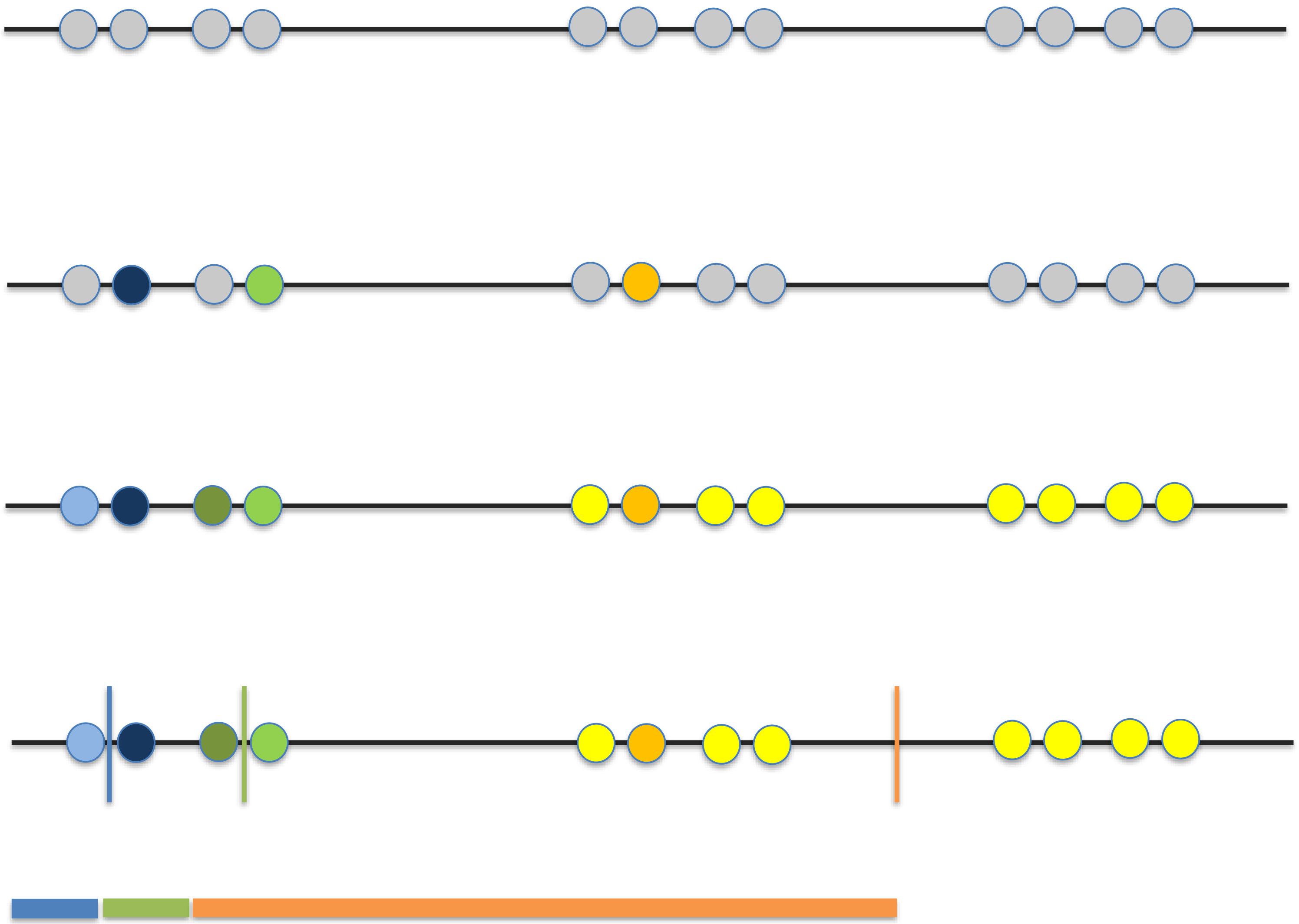


# $K$-means++

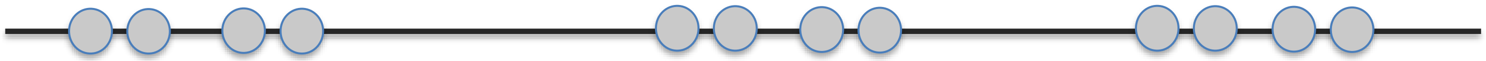
# Allocation to wrong cluster - I

- Hypothetical dataset:
- Assume we randomly choose the following 3 centroids at step 0:
- Assignment of examples to clusters based on distances
- New centroid values
- Total variation within clusters:

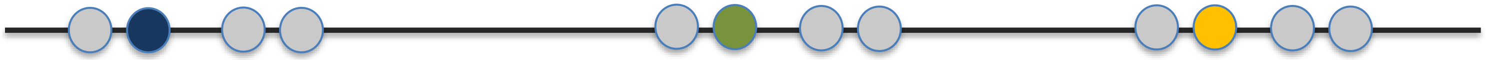


# Allocation to wrong cluster - II

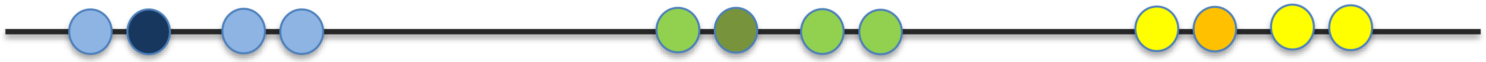
- Same dataset as before



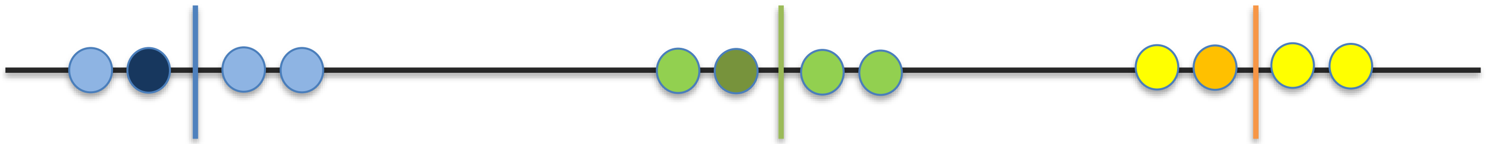
- Assume the randomly chosen centroids at step 0 are as follows



- Assignment of examples to clusters based on distances



- New centroid values:



- Total variation within clusters:

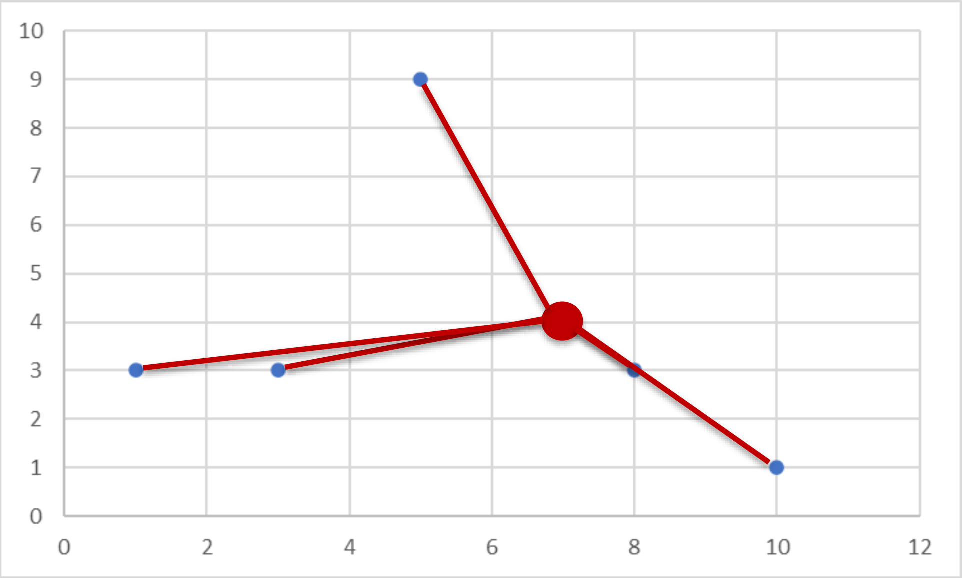
# $K$-means++

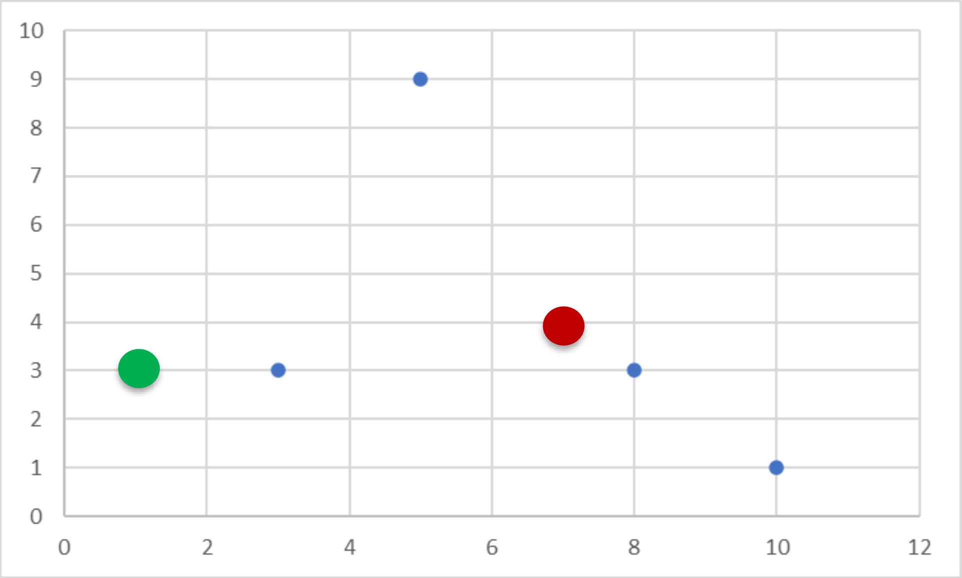
- One way to avoid the previous issue is to run the $K$-means algorithm several times and to keep the clustering that achieves the lowest variation within clusters
- Yet, a nice solution to overcome bad initialization of the centroids is to use $K$-means++
- The $K$-means++ algorithm is as follows:
  - 1.a: Choose an initial centroid $c_1$ at random (with uniform distribution) from $\boldsymbol{X}$
  - 1.b: Choose the next centroid $c_i$, selecting $c_i = x'$ from $\boldsymbol{X}$ with probability $\frac{D(x')^2}{\sum_{x \epsilon X} D(x')^2}$, where $D(x)$ denotes the shortest distance from a data point $x$ to the closest centroid already chosen
  - 1.c: repeat step 1.b until the $K$ centroids are chosen
  - 2. Proceed as with the standard $K$-means algorithm

# Example - IV

- In this example, the objective is to assign the instances of a small dataset to 3 clusters

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | =(A7-C7)^2+(B7-D7)^2 | | |
| 2 | | | | | | | | | |
| 3 | | | | | | | | | =E7/$E$13 |
| 4 | | | | | | | | | |
| 5 | x | | $c_1$ | | $D(x,c_1)^2$ | prob | cum prob | | |
| 6 | 7 | 4 | 7 | 4 | - | - | | | |
| 7 | 8 | 3 | 7 | 4 | 2 | 0.02 | 0.02 | | |
| 8 | 5 | 9 | 7 | 4 | 29 | 0.28 | 0.30 | =G7+F8 | |
| 9 | 3 | 3 | 7 | 4 | 17 | 0.17 | 0.47 | | |
| 10 | 1 | 3 | 7 | 4 | 37 | 0.36 | 0.83 | $c_2$ | |
| 11 | 10 | 1 | 7 | 4 | 18 | 0.17 | 1 | | |
| 12 | | | | | | | | | |
| 13 | | | | Total | 103 | =SUM(E7:E11) | | | |
| 14 | | | | | | | | | |
| 15 | Uniformly distributed random number | | | | | 0.78 | =RAND() | | |

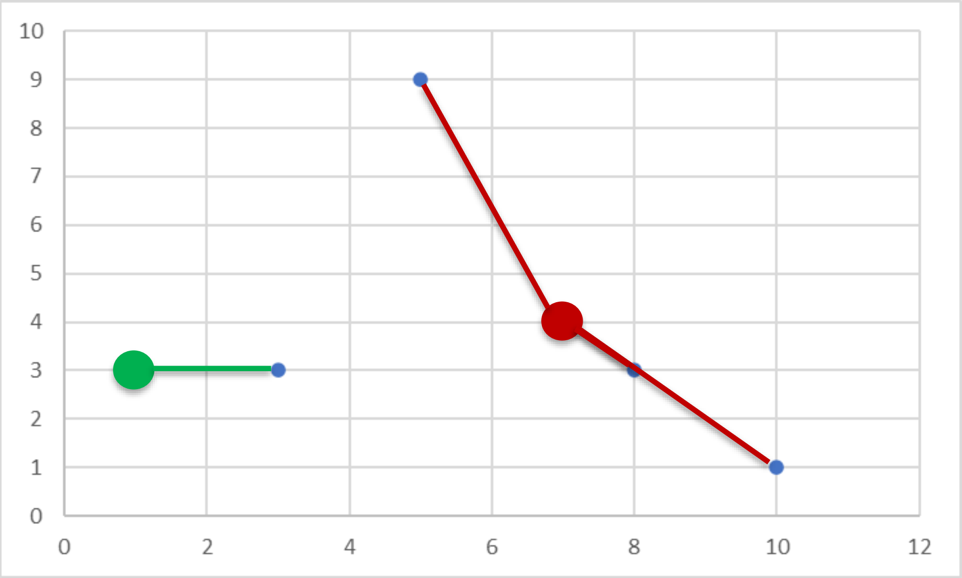


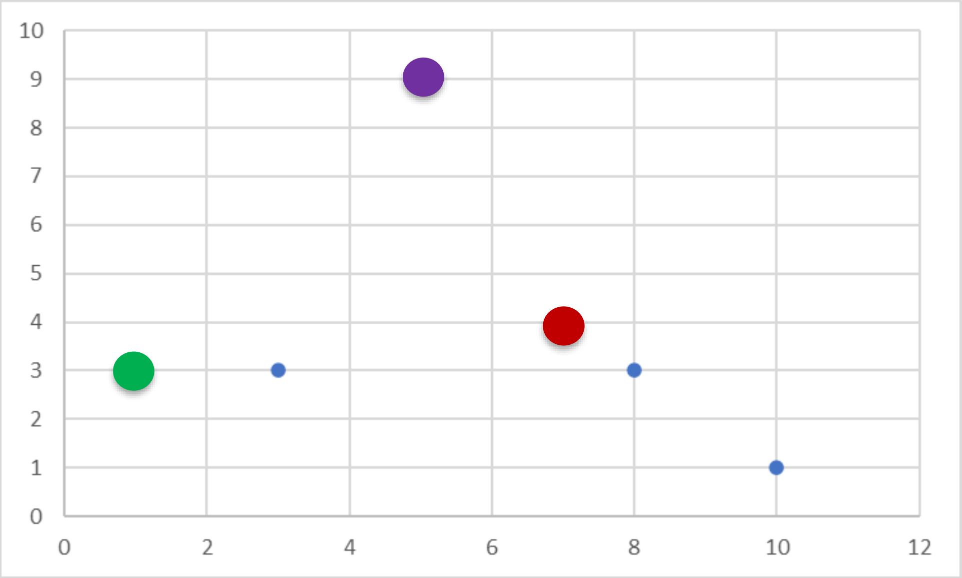



# Example - V

| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | x | | $c_1$ | | $c_2$ | | $D(x,c_1)^2$ | $D(x,c_2)^2$ | $\min(D(x,c_i)^2)$ | prob | cum prod | |
| 18 | 7 | 4 | 7 | 4 | 1 | 3 | - | - | | | | |
| 19 | 8 | 3 | 7 | 4 | 1 | 3 | 2 | 49 | 2 | 0.04 | 0.04 | |
| 20 | 5 | 9 | 7 | 4 | 1 | 3 | 29 | 52 | 29 | 0.55 | 0.58 | $c_3$ |
| 21 | 3 | 3 | 7 | 4 | 1 | 3 | 17 | 4 | 4 | 0.08 | 0.66 | |
| 22 | 1 | 3 | 7 | 4 | 1 | 3 | - | - | | | 0.66 | |
| 23 | 10 | 1 | 7 | 4 | 1 | 3 | 18 | 85 | 18 | 0.34 | 1.00 | |
| 24 | | | | | | | | | | | | |
| 25 | | | | | | | | Total | 53 | | | |
| 26 | | | | | | | | | | | | |
| 27 | Uniformly distributed random number | | | | | 0.47 | =RAND() | | | | | |

- Once the 3 clusters are initialized, the standard $K$-means algorithm is ran with centroids $(7,4)$, $(1,3)$ and $(5,9)$



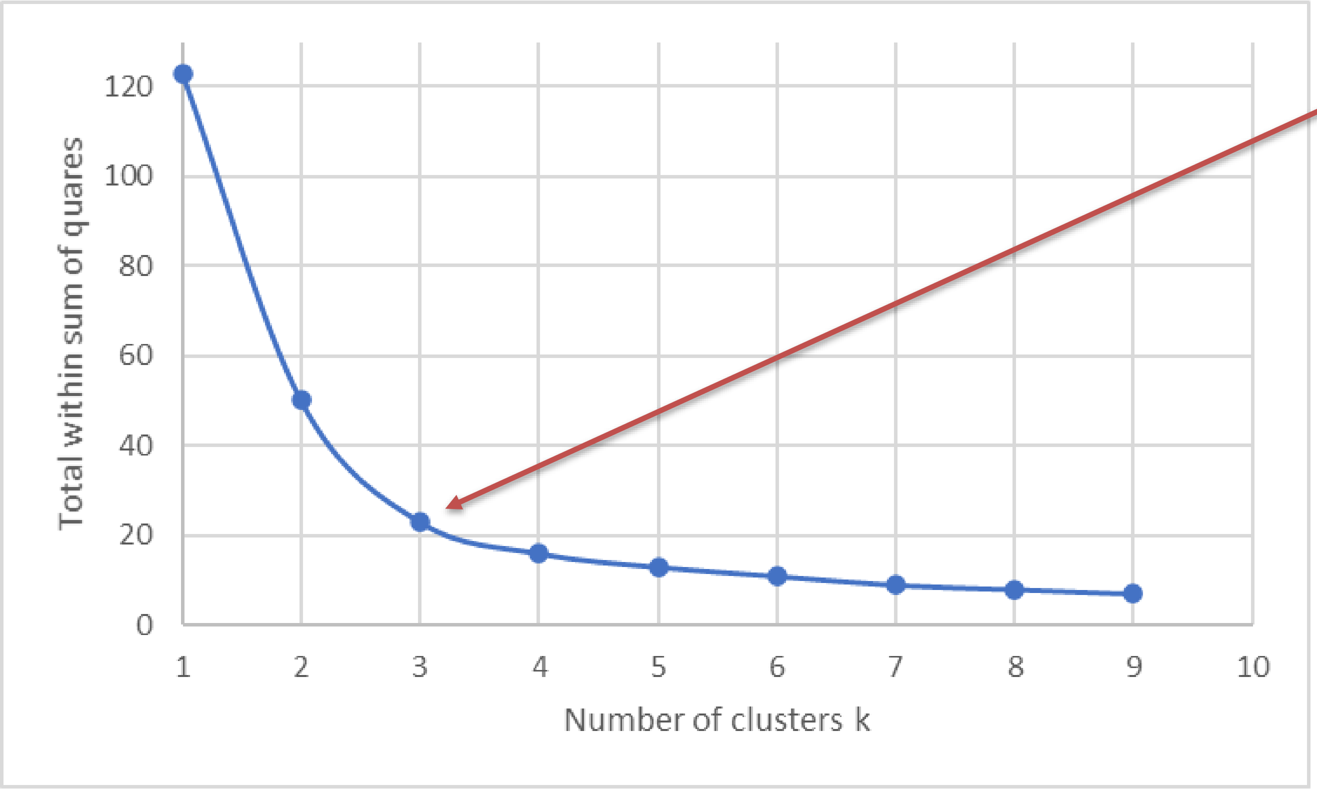
# How many clusters?

# Metrics

- As $K$-means is an unsupervised method, the actual number of clusters is unknown
- However, a variety of measures (currently around 30) have been proposed for evaluating clustering results and determine the appropiate number of clusters
- We will cover two of them only
  - The "Elbow" method
  - The silhouette statistic (Rousseeuw, JCAM 1987)

# The "Elbow" method

- The "Elbow" method: plot the within clusters sum of squared distances against the number of clusters. A rule of thumb if to set the optimal $K$ as the one such that the slope of the graph goes from steep to shalow (elbow)




Optimal $K$

# The silhouette coefficient - I

- Assume the examples have been clustered via any technique (e.g. $K$-means) into $K$ clusters
- For sample $i \in C_i$, let define $a(i)$ such that:

$$a(i) = \frac{1}{n_i - 1} \sum_{j \in C_i, i \neq j} D(i, j)$$

where $n_k$ is the number of samples in cluster $i$ and $D(i, j)$ is the distance between samples $i$ and $j$. So, $a(i)$ is the mean distance between $i$ and all other samples within the same cluster.

- Next, let define $b(i)$ such that:

$$b(i) = \min_{k \neq i} \frac{1}{n_k} \sum_{j \in C_k} D(i, j)$$

- $b(i)$ is the mean distance of $i$ to the cluster with the smallest mean distance (neighboring cluster)

# The silhouette coefficient - II

- The silhouette coefficient $s(i)$ of sample $i$ is given by:

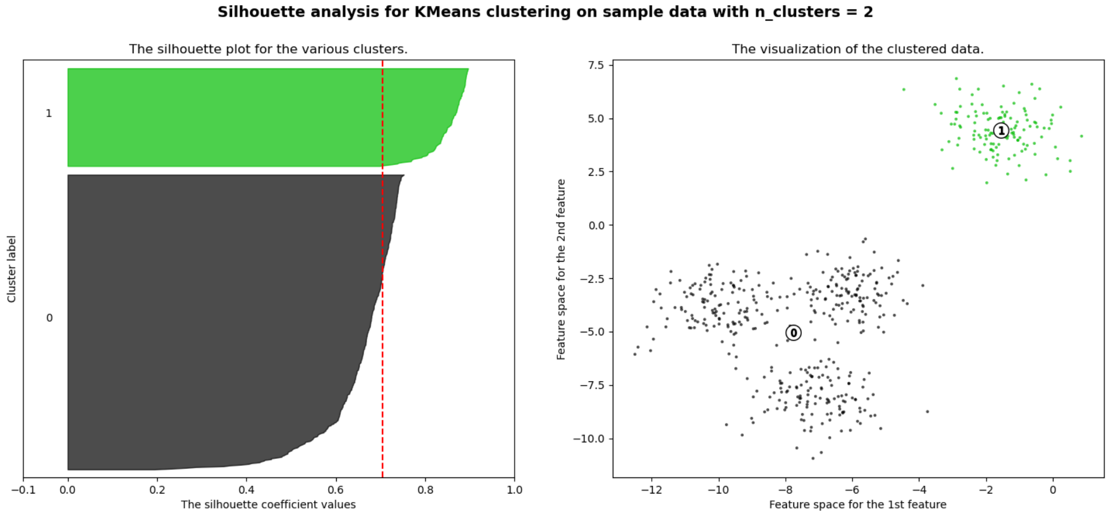
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

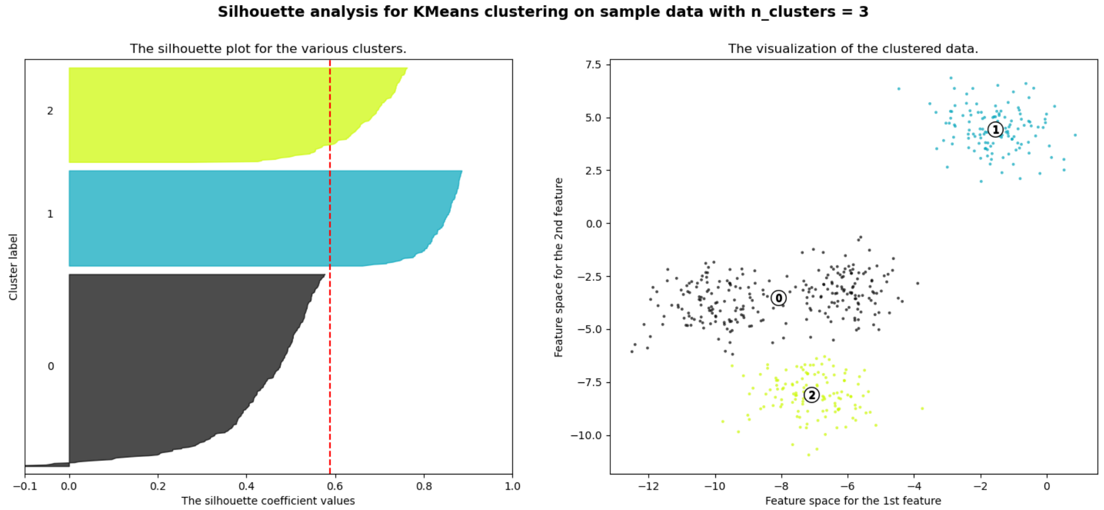
- From the definition of $s(i)$, it is clear that $-1 \leq s(i) \leq 1$.
  - $s(i)$ close to 1 requires that $a(i) \ll b(i)$
    - A low $a(i)$ means that $i$ is well matched to its own cluster
    - A large $b(i)$ means that it is badly matched with its neighboring cluster
    - So $s(i)$ close to 1 means that sample $i$ is appropriately clustered.
  - $s(i)$ close to -1 requires that $a(i) \gg b(i)$
    - A high $a(i)$ means that sample $i$ is not well matched to its own cluster
    - A small $b(i)$ means that it matches well with its neighboring cluster
    - Therefore, $i$ should be assigned to its neighboring cluster
  - Finally, $s(i) = 0$ means that the sample is on the border of two clusters

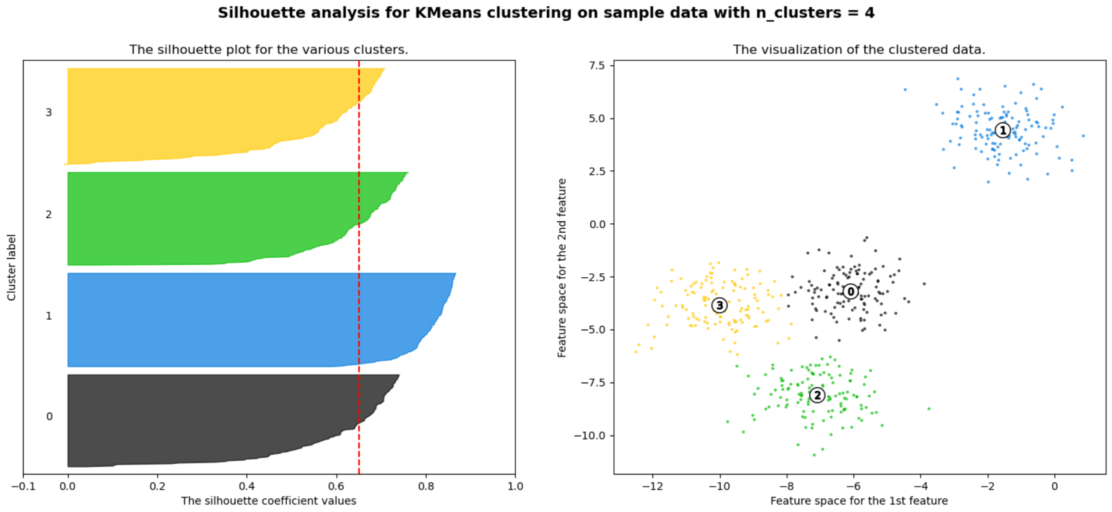
# The silhouette coefficient - III

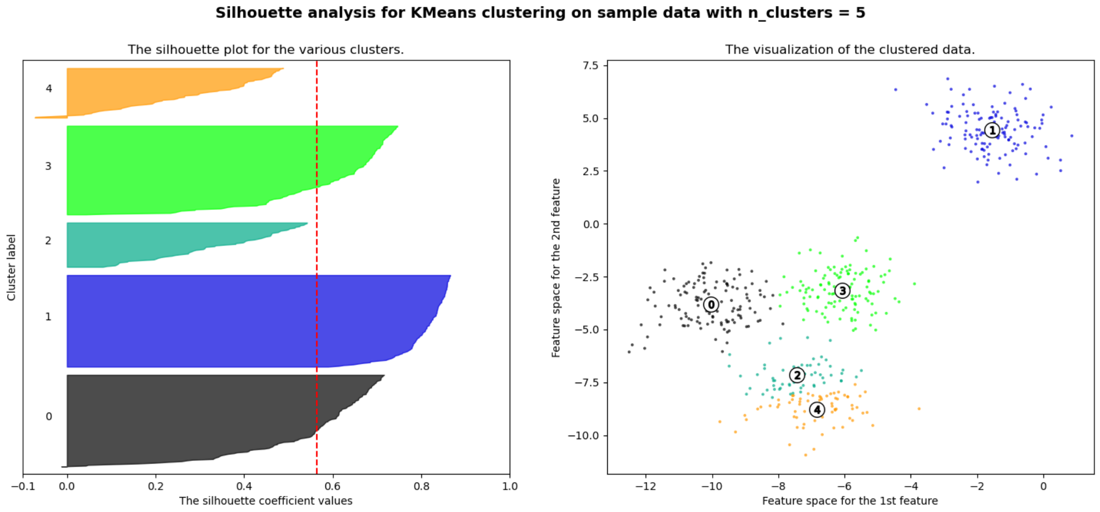
- The mean $s(i)$ over all samples **within a cluster** measures how tightly grouped all the points in the cluster are
- The mean $s(i)$ over all samples **in the dataset** is a measure of how appropriately the data have been clustered

# Silhouette plot examples




Source: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html

# Use cases

# Jupyter notebooks and data

- Iris example:
  - Data: iris_data.txt
    - Features: Sepal Length, Sepal Width, Petal Length, Petal Width
    - Labels: Iris-setosa, Iris-versicolor, Iris-virginica
  - Jupyter notebook: iris-solution.ipynb
- Portfolio sampling:
  - Data: stocks.csv
  - Jupyter notebook: Stock sampling using k-means - tutorial version.ipynb