COS10022 – DATA SCIENCE PRINCIPLES

Dr Pei-Wei Tsai (Lecturer, Unit Convenor)
ptsai@swin.edu.au, EN508d

SWINBURNE UNIVERSITY OF TECHNOLOGY

Week 11
Communication & Model Deployment

COS10022 - Data science Principles

Key Questions

- Which aspects of modeling results need to be communicated?
- How do we visualize results effectively?

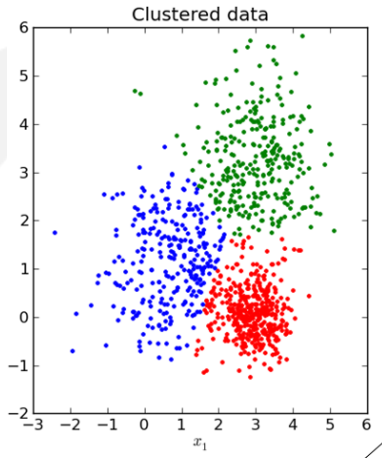
- What and how to communicate project deliverables (results) to different groups of stakeholder?
- What are some general model deployment best practices?
- A case in point.
  Deployment practices of K-Means Clustering in customer segmentation applications.

Learning Outcomes

This lecture supports the achievement of the following learning outcomes:

3. **Describe the processes within the Data Analytics Lifecycle.**
4. **Analyse business and organisational problems and formulate them into data science tasks.**
5. **Evaluate suitable techniques and tools for specific data science tasks.**
6. **Develop analytics plan for a given business case study.**

Phase 5 – Communicate Results



Model
k-means **clustering**
Naïve Bayes **classification**
Simple linear **regression**

Suppose that Alice and Bob uses the following words with probabilities as show below. now, can you guess who is the sender for the content: "Wonderful Love"?

Alice	Bob
Love [0.1]	Wonderful [0.5]
Wonderful [0.1]	Love [0.3]
Great [0.8]	Deal [0.3]

Phase 5 – Communicate Results

In the 5th phase of the Data Analytics Lifecycle, the data science team, in collaboration with major stakeholders, determines if the results of the project are a **success** or a **failure** based on the criteria developed in Phase 1.

The team should identify **key findings**, quantify the **business value**, and develop a narrative to **summarize and convey findings** to stakeholders.

Phase 5 – Communicate Results

Key activities:

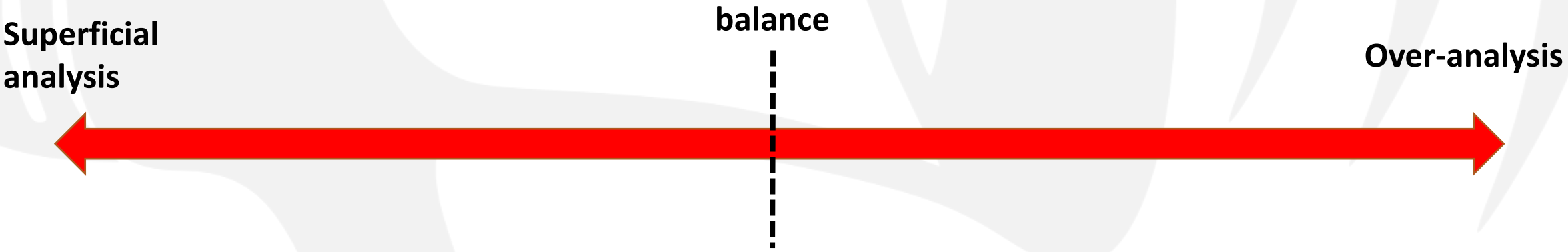
1. **Compare the outcomes of the modeling to the criteria established for success and failure.**
2. **Determine if the data science project succeeded or failed in its objectives.**
3. **Communicate model(s) that addresses the analytical challenge in the most appropriate way.**
4. **Reflect on the project, the encountered obstacles and what can be improved in the future.**

The data science team can move to the next phase if the team has documented and reported the *key findings*, and major *insights* have been derived from the analysis.

Success or failure criteria

The success or failure criteria are determined by whether the data and the chosen analytics models are able to accept or reject the initial hypotheses formulated in Phase 1.

Rejecting a hypothesis does **not** always equate a failure. Instead, a failure usually refers to the inability to strike the balance between two possible analytics extremes.



Success or failure criteria

When assessing whether the data and the model accept or reject a hypothesis:

- Determine if the results are **significant** and **valid**.
  - Reject the null hypothesis ($p < 0.05$)
  - Accept the null hypothesis ($p \geq 0.05$)
  - Internal validity:
    - Experimental design meets the requirements of scientific method employed
  - External validity :
    - Validity of employing the conclusions outside of the context of a study (about generalizing)

- Identify components of the results which are in line with the **initial hypotheses**, as well as those which are **surprising**.

The significance of results

- **Statistical Significance**
  The **paired sample t-test**, sometimes called the dependent sample t-test, is a statistical procedure used to determine whether the **mean difference** between two sets of observations is zero (**null hypothesis**).

  Statistical significance is determined by looking at the **p-value**. The p-value gives the probability of observing the test results under the null hypothesis. A statistically significant result should reject the null hypothesis (p-value < 0.05).

  For example, if your evaluation results suggest that model A has a higher accuracy than model B and the p-value < 0.05, then the performance difference is said to be statistically significant. Otherwise, there is no real difference between the performance of both models.

  Null hypothesis (H_0) : There is no difference between the performance of Model A and Model B.
  Alternative hypothesis (H_A) : Model A has a higher accuracy than Model B.

The significance of results

- **Practical Significance**
  Statistical significance itself doesn't necessarily imply that your results have practical significance.

  When the data is large enough, even very small differences between the performance of two analytics models may be deemed as statistically significant by the paired sample t-test. It is possible, however, that from a practical such a small difference is not useful.

  Practical significance depends on the subject matter.

The significance of results

Both statistical significance and practical significance need to be met in order to draw a meaningful conclusion from your analysis.

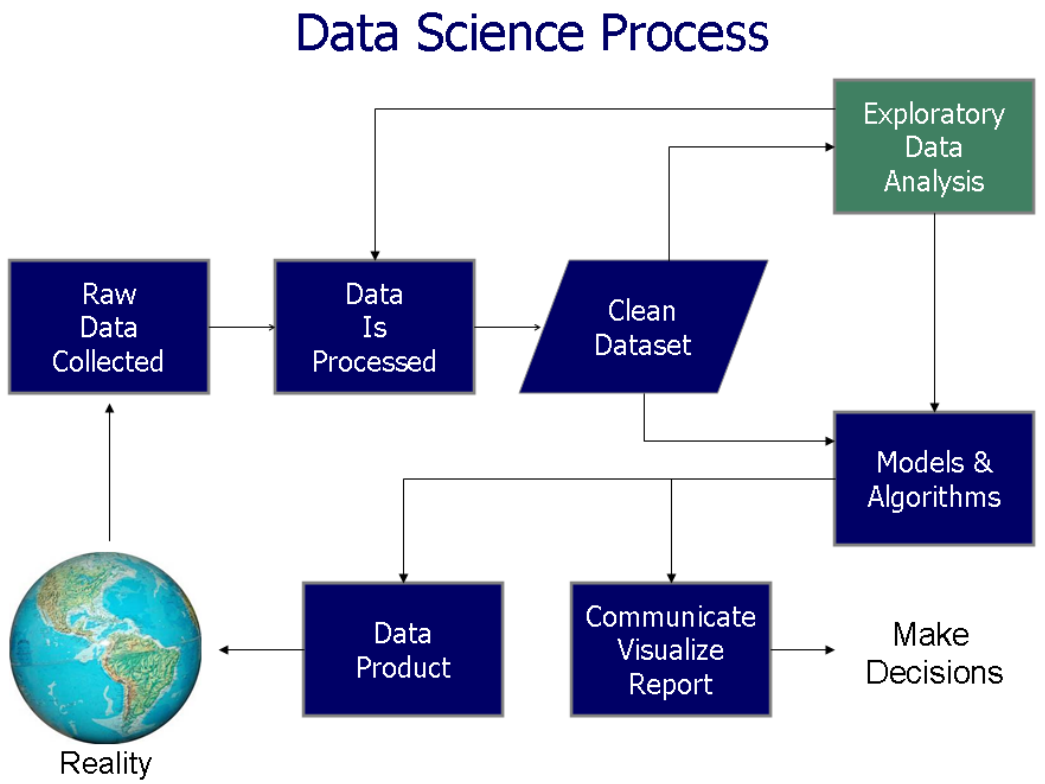
Further discussion of significance testing is beyond the scope of this unit. More information can be found here:

- http://www.datasciencecentral.com/profiles/blogs/statistical-significance-and-its-part-in-science-downfalls
- http://www.statisticssolutions.com/manova-analysis-paired-sample-t-test/
- http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/introductory-concepts/p-value-and-significance-level/what-is-pvalue/

Data visualisation basics

As the volume and complexity of data has grown, data scientists have become more reliant on using crisp visuals to illustrate key ideas and portray rich data in a simple way.

Data visualisation involves the creation and study of the visual representation of data, where the primary goal is to communicate information clearly and efficiently.



Adapted from: Schutt, R. and O'Neil, C., 2013.
Doing data science: Straight talk from the frontline. O'Reilly Media, Inc.

Data visualisation basics

There are many tools that support the creation of effective data visualisations.

Open Source Tools	Commercial Tools
R (`lattice`, `ggplot2`)	Tableau
GGobi / Rggobi	Spotfire
Gnuplot	QlikView
PyPlot, Matplotlib (Python)	Adobe Illustrator
OpenLayers (JavaScript)	
D3.js (JavaScript)	
Inkscape	

Data visualisation basics

Lozovsky, V. 2008. "Table vs Graph", *Information Builders.* Link: http://www.informationbuilders.com/new/newsletter/9-2/05_lozovsky

It is more difficult to observe key insights when data is presented in **tables** instead of in **charts**. What follows are two different approaches to visualising just the same data.

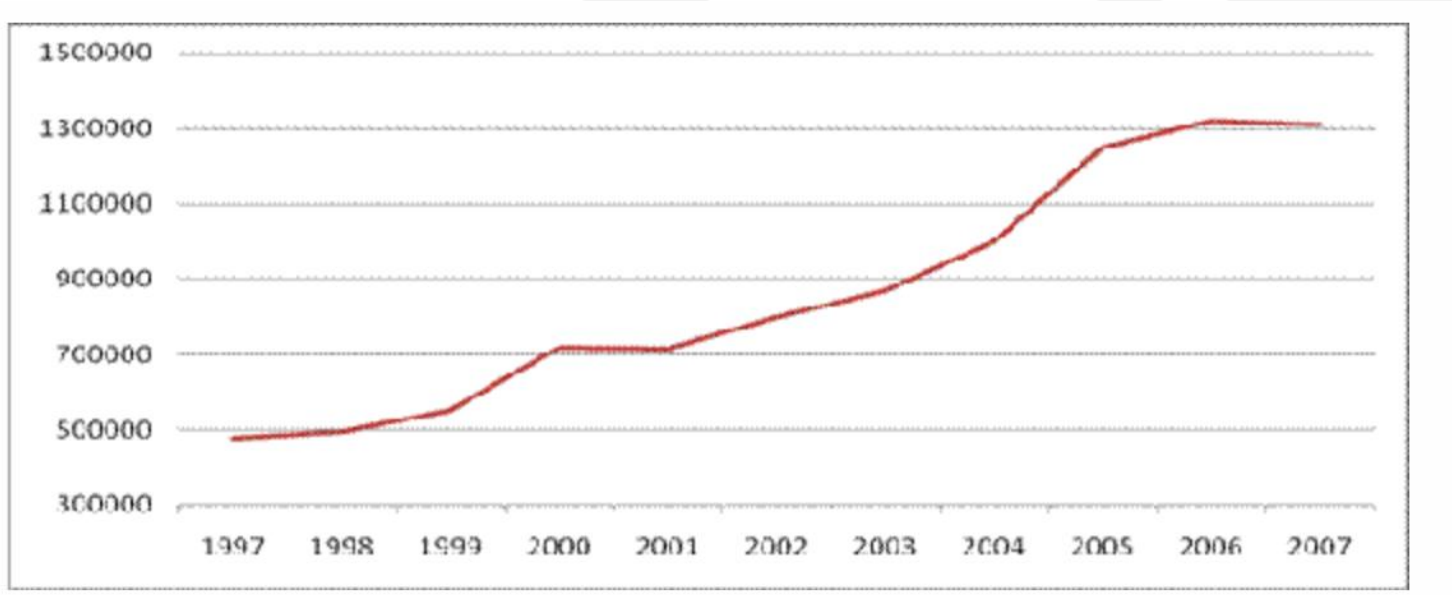
As an example, here is a quarterly survey of Manhattan real estate sales:

Manhattan Market Matrix	Current Qtr	%Chg	Prior Qtr	%Chg	Prior Year Qtr
Average Sales Price	$1,439,909	5.1%	$1,369,486	17.6%	$1,224,840
Average Price per Square Foot	$1,180	3.1%	$1,144	18.2%	$998
Median Sales Price	$850,000	-1.7%	$864,397	6.4%	$799,000
Number of Sales	2,518	-28%	3,499	3.2%	2,441
Days on Market	131	6.9%	123	-12.4%	149
Listing Discount	2.7%		2.0%		2.8%
Listing Inventory	5,133	-1.4%	5,204	-13.5%	5,934

TABLE

Data visualisation basics

Lozovsky, V. 2008. "Table vs Graph", *Information Builders.* Link: http://www.informationbuilders.com/new/newsletter/9-2/05_lozovsky



CHART

Charts generally work best when the visualisation needs to communicate a message that is contained in the **shape** of the data and the **relationships** between variables.

Data Visualisation Basics

- Similar things can be found in our daily life. The speed meter is one of the typical example.

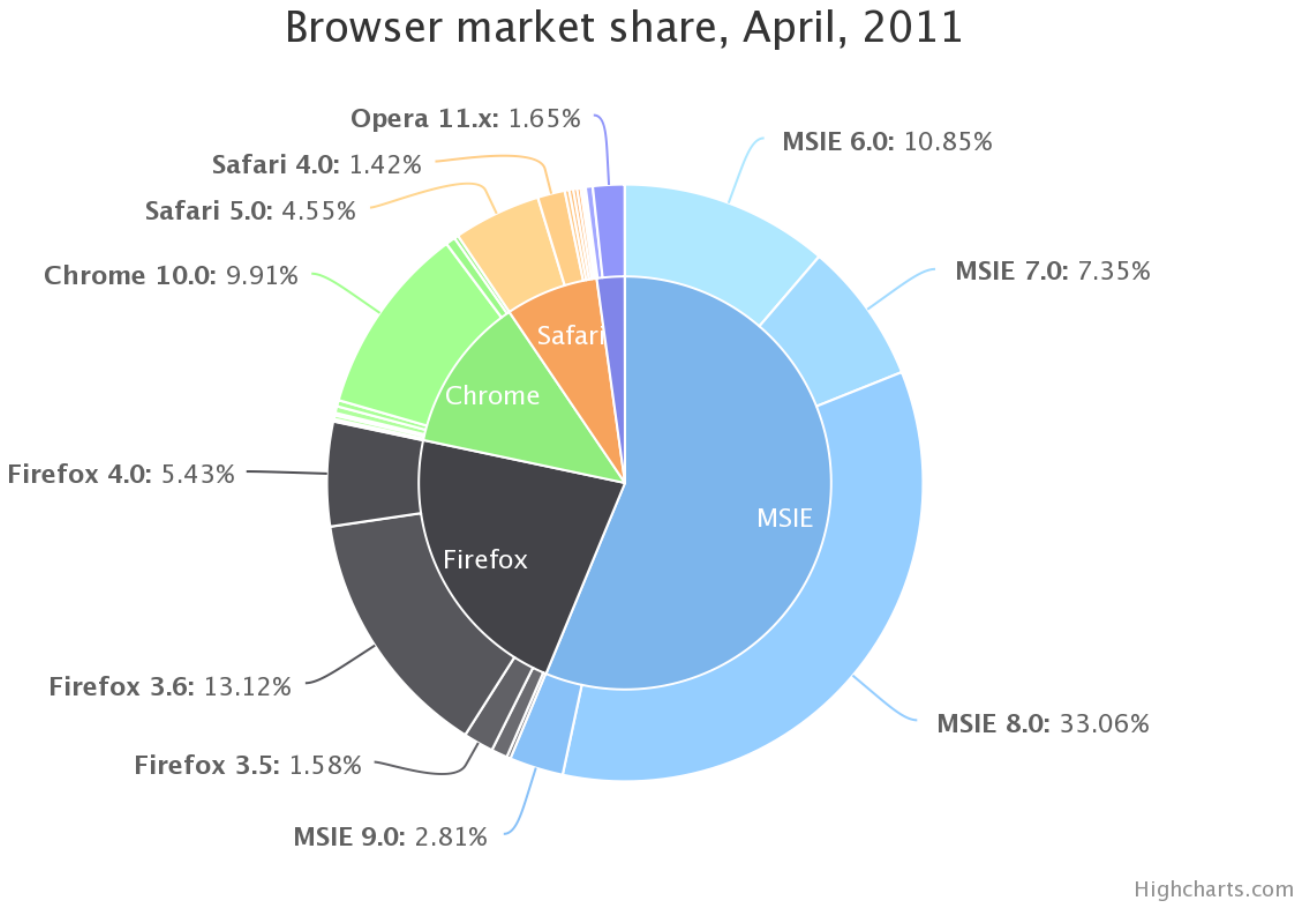
Data visualisation basics

It is important to know when to use **a particular type of chart** or **graph** to express a given kind of data. The objective is to find the best chart for expressing the data **clearly** so the visual does not impede the message, but supports the audience in taking away the intended message.

Data to be visualised	Type of chart
Components (parts of whole)	Pie chart
Item	Bar chart
Time series	Line chart
Frequency	Line chart of histogram
Correlation	Scatterplot, side-by-side bar charts

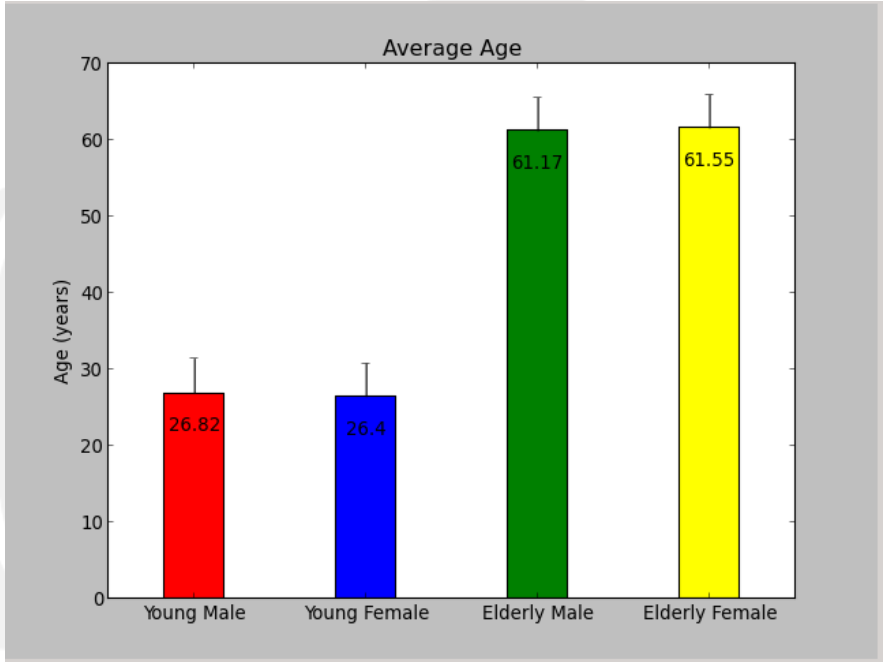
A pie chart generated with R

(source: http://stackoverflow.com/questions/26748069/ggplot2-pie-and-donut-chart-on-same-plot)

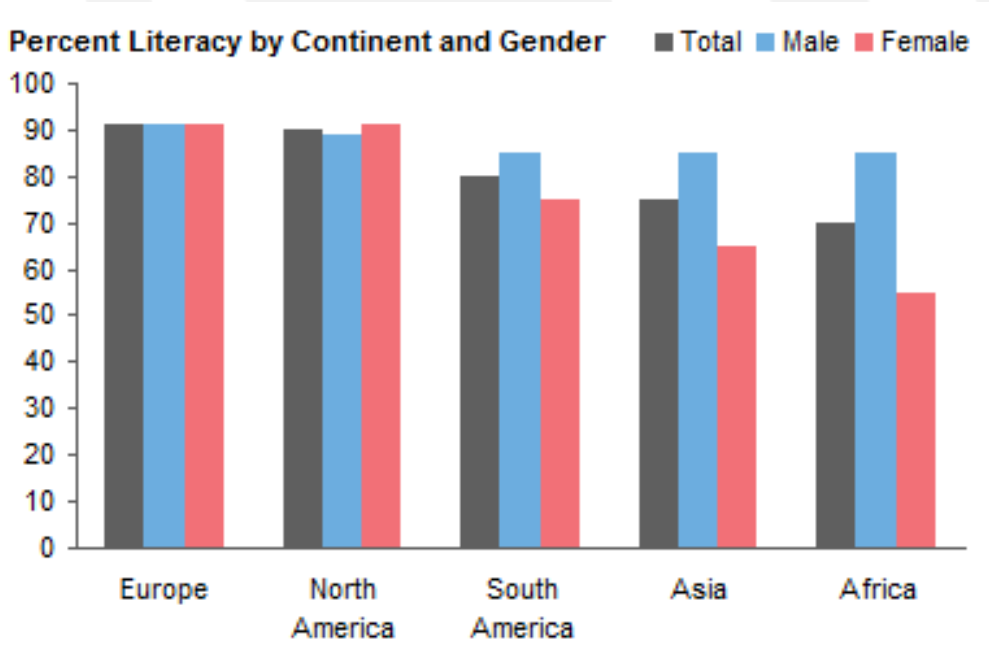


A bar chart plotted with PyPlot

(source: http://stackoverflow.com/questions/13312820/how-do-i-plot-just-the-positive-error-bar-with-pyplot-bar)



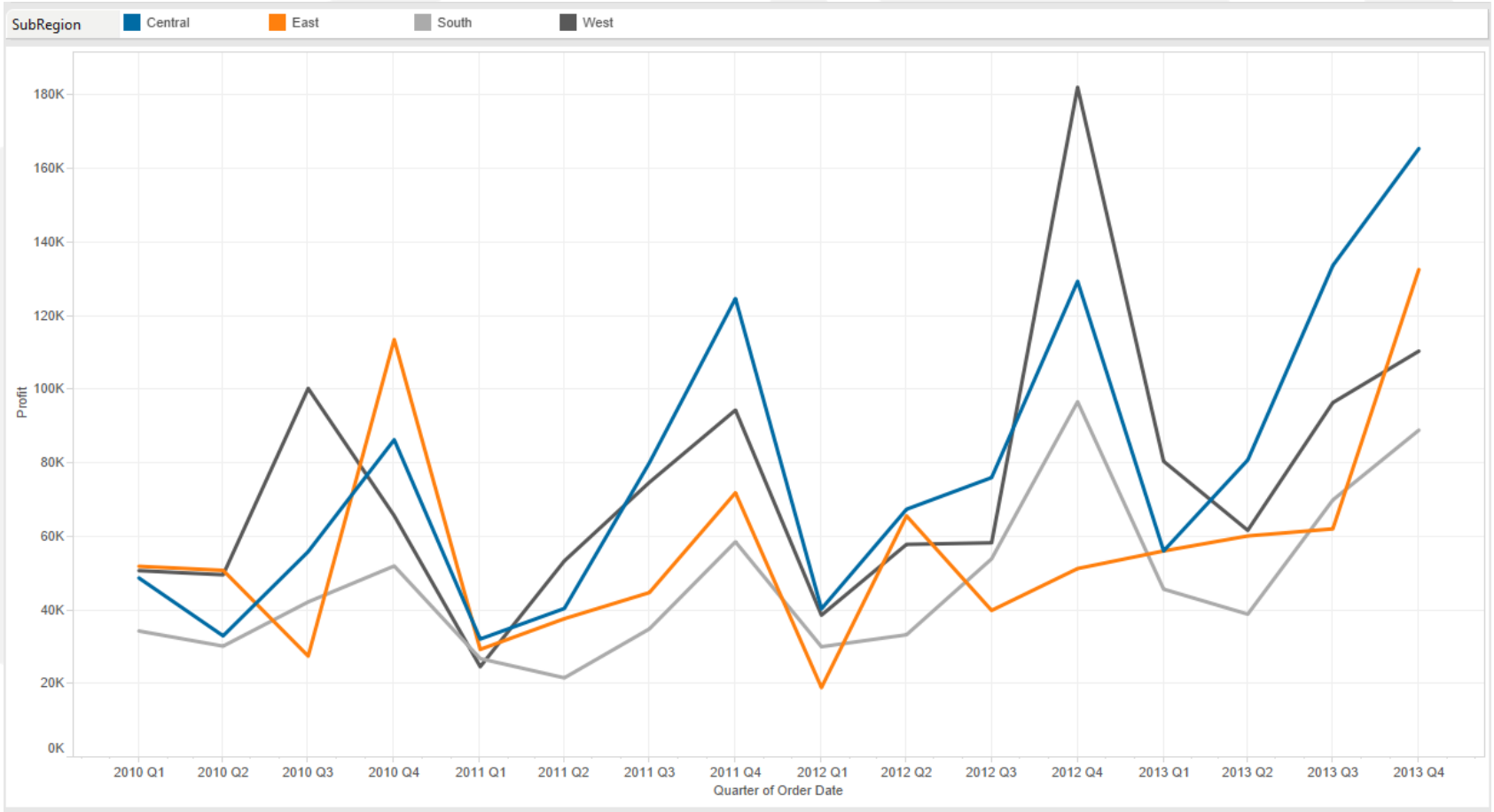
Simple bar chart



Side-by-side bar chart

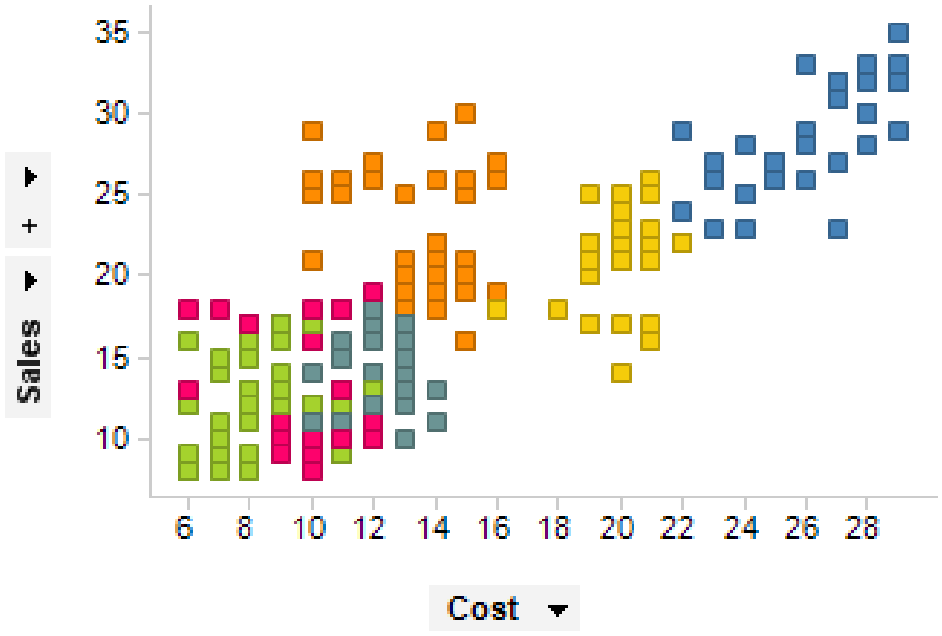
A line chart plotted with Tableau

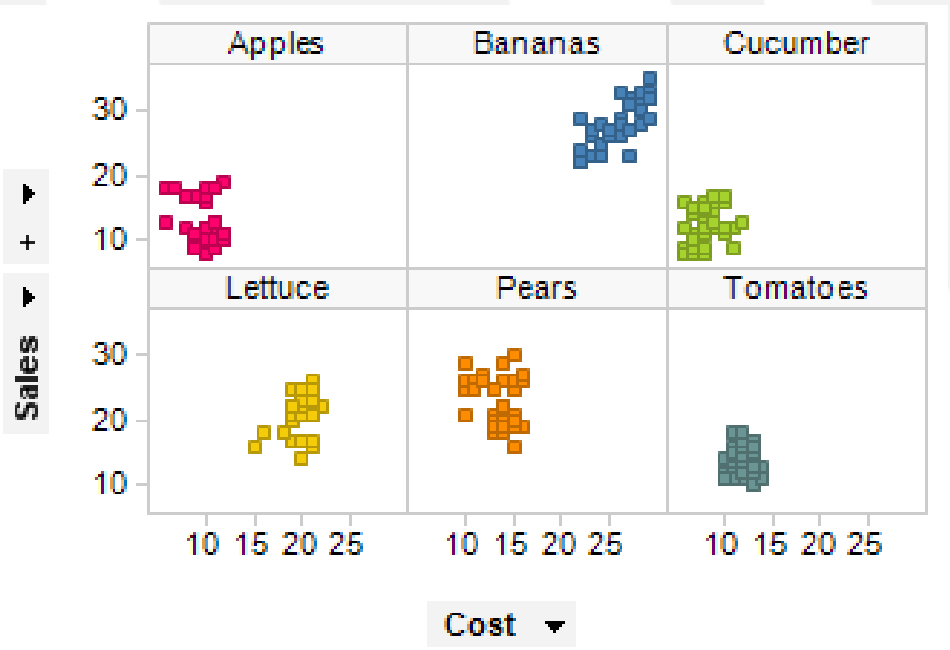
(source: https://www.interworks.com/blog/ccapitula/2014/10/28/tableau-essentials-chart-types-line-charts-continuous-discrete)



Scatterplots created with Spotfire

(source: https://docs.tibco.com/pub/spotfire/6.5.2/doc/html/scat/scat_what_is_a_scatter_plot.htm)





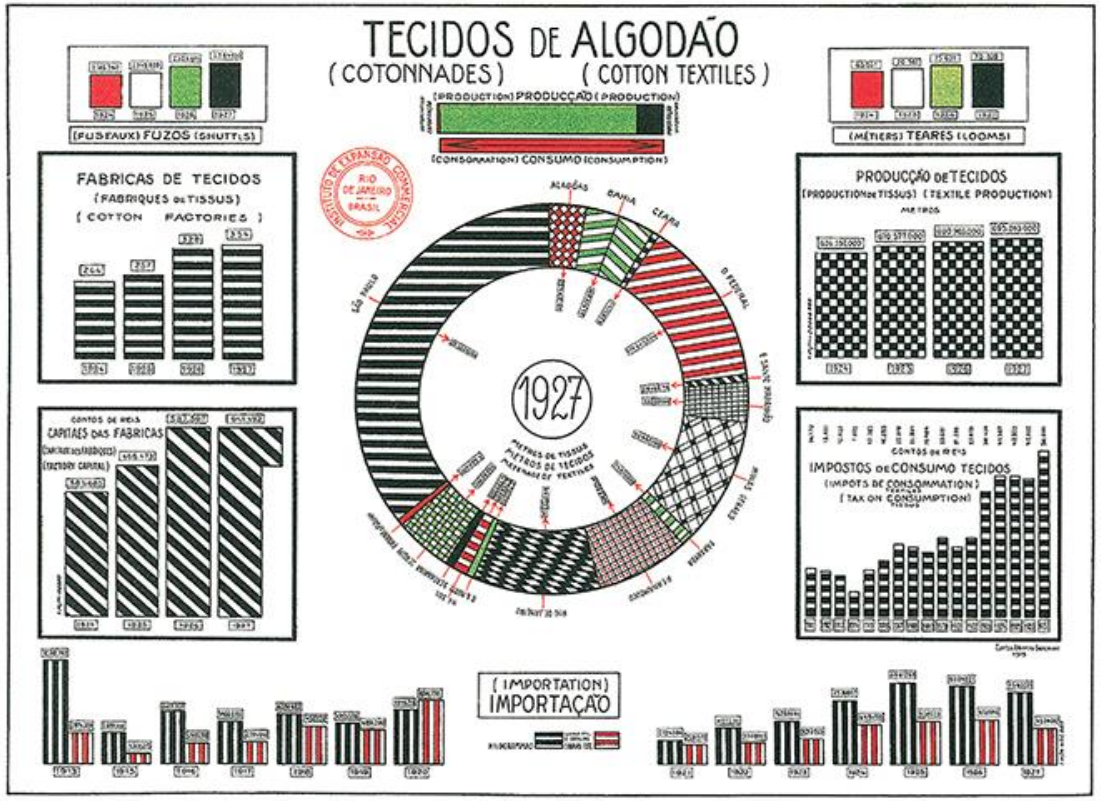
Data visualisation best practices

Tufte, E. 1983. *The Visual Display of Quantitative Information.* Link: https://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=00040Z

Avoid chart junks.

Software packages may add extraneous visualisation elements that could convolute the presented information. Instead, strive for simplicity.

On the right is an example of highly 'noisy' data visualisation (Tufte 1983).



Engage #1: Chart dos and don’ts

European Environment Agency 2013. Link: http://www.eea.europa.eu/data-and-maps/daviz/learn-more/chart-dos-and-donts

Which principle(s) of chart visualisation do you often violate?

Data visualisation best practices

Be aware of "Data-Ink Ratio".

Data-Ink refers to the actual portion of a graphic that portrays the data, while **non-Data Ink** refers to labels, edges, colours, and other decoration.

Hence, **Data-Ink Ratio** could be thought of as:

$$\frac{\text{Data - ink}}{\text{Total ink used to print the graphic}}$$

Also equals to:
1.0 – proportion of a graphic that can be erased without loss of data-information

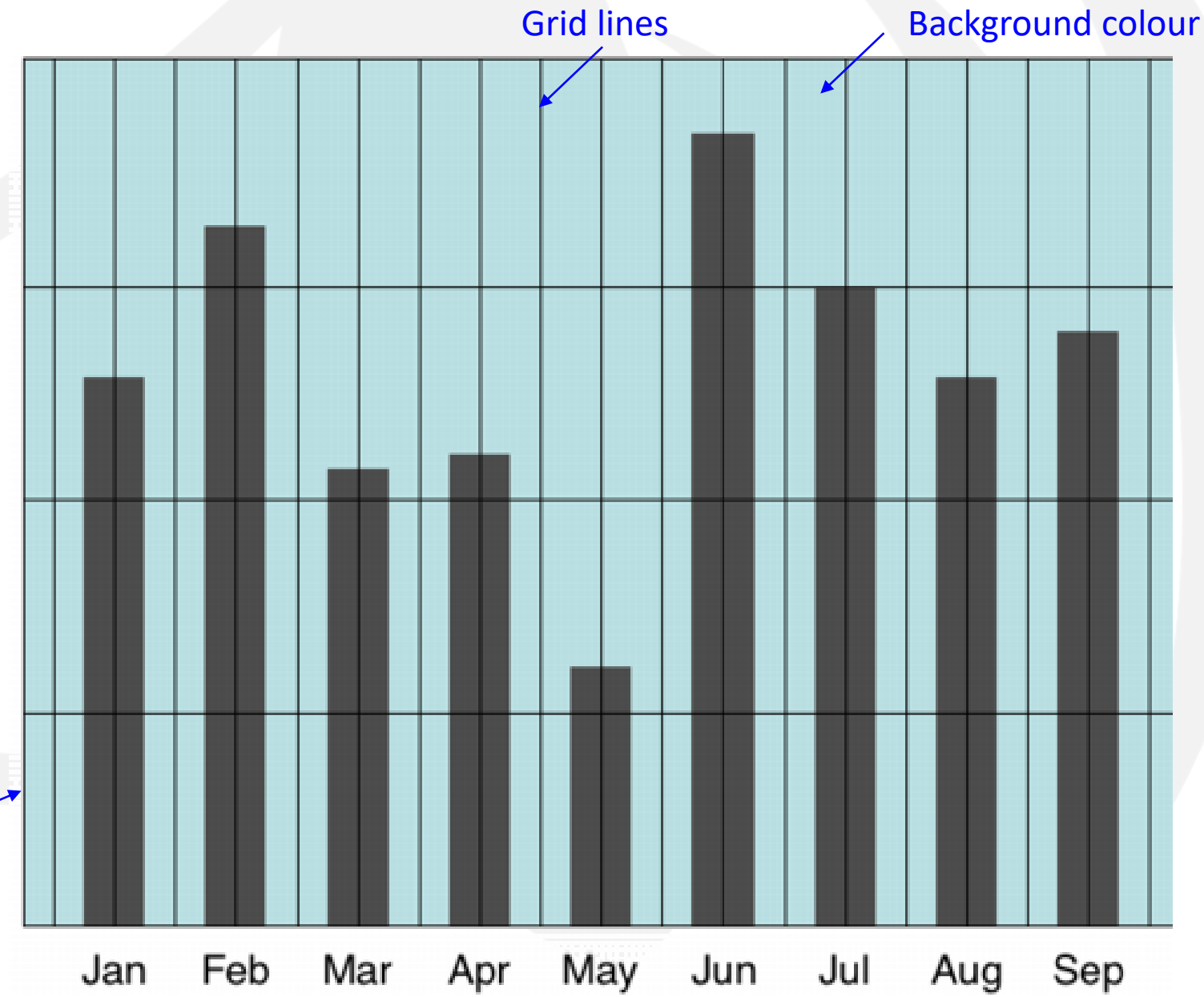
The greater the ratio, the more data rich it is and the fewer distractions it has. According to Edward Tufte who pioneered the concept, the goal of data visualisation is to design display with the highest possible data-ink ratio (that is, as close to the total of 1.0), without eliminating something that is necessary for effective communication.

Data Visualisation Best Practices

Source: http://www.infovis-wiki.net/index.php/Data-Ink_Ratio

Example of a graph with a low Data-Ink Ratio

The border around the graph, the background colour and the grid lines are all unnecessary data ink.



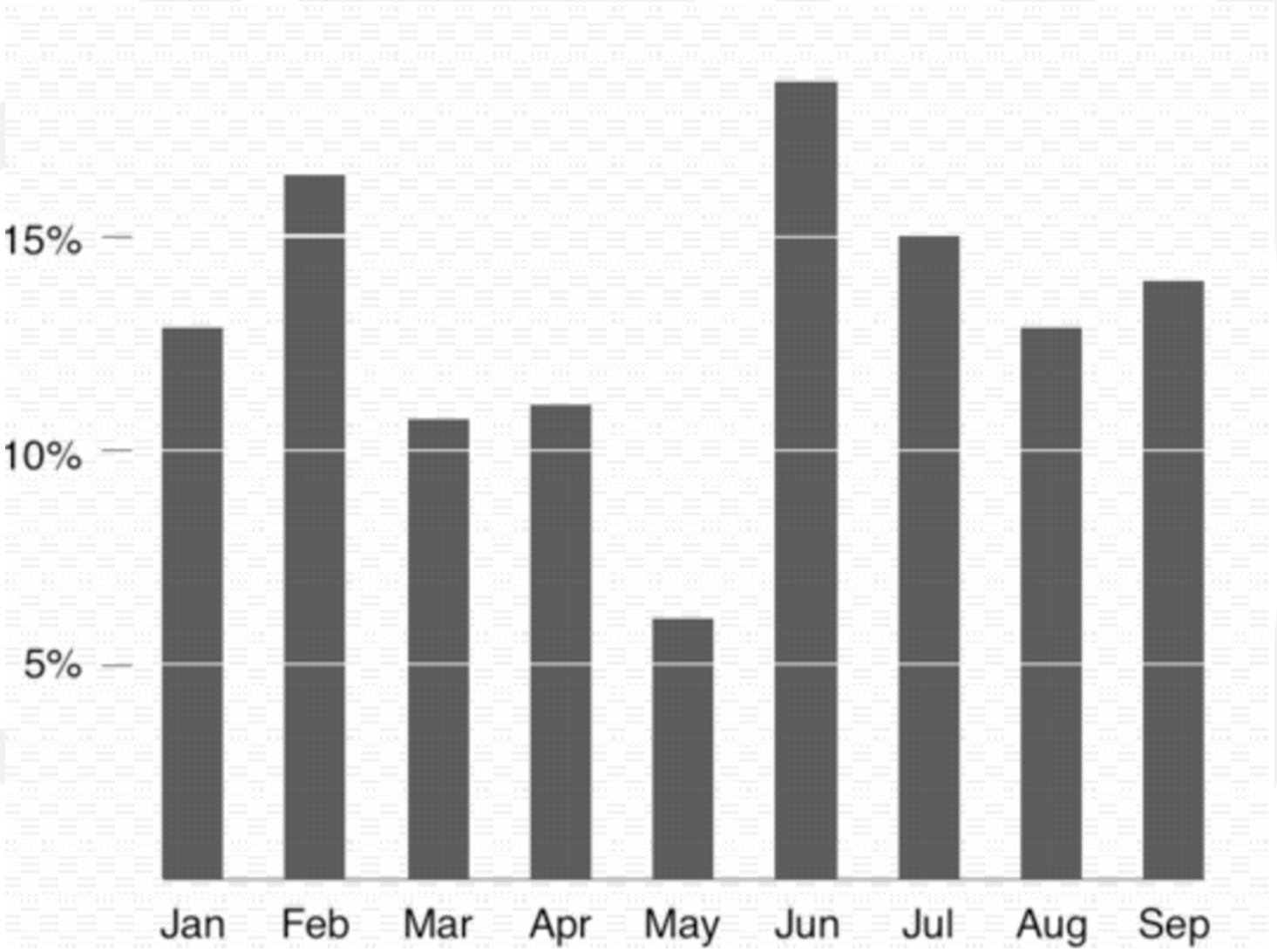
Data visualisation best practices

Source: http://www.infovis-wiki.net/index.php/Data-Ink_Ratio

Example of a graph with a high Data-Ink Ratio

The border around the graph, the background colour and the grid lines have been deleted, and have thus drawn the viewer's attention to horizontal scales that are data-ink.

There is nothing else to distract and the key features of the data stand out clearly.

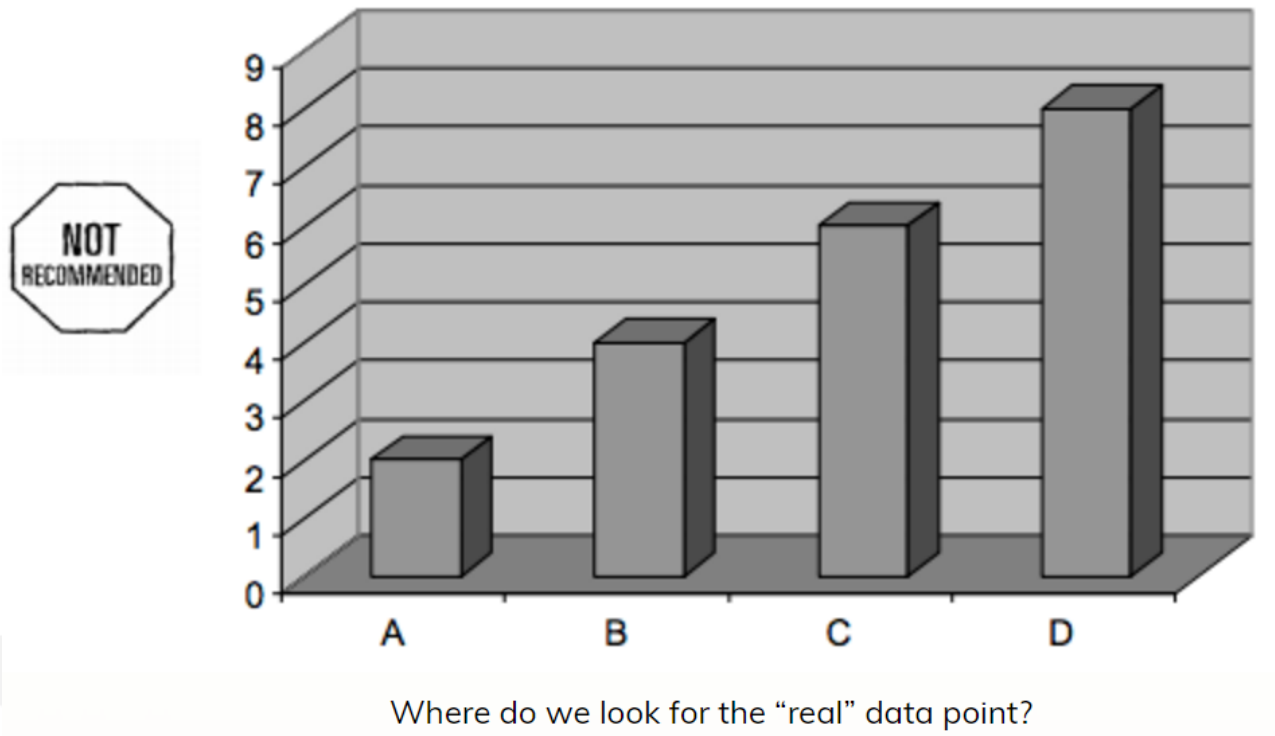


Data visualisation best practices

Avoid using 3D visualisation in most graphics.

Usually, 3D visualisations add unnecessary shading, depth, or dimensions to graphics. 3D charts often distort scales and axes, thus impeding viewer recognition.

The 3D bar chart on the right is not recommended as we are not sure if the front of the bar or the back of the bar is **where the data really is** and this could all be based on which software program you used to generate it.



Where do we look for the "real" data point?

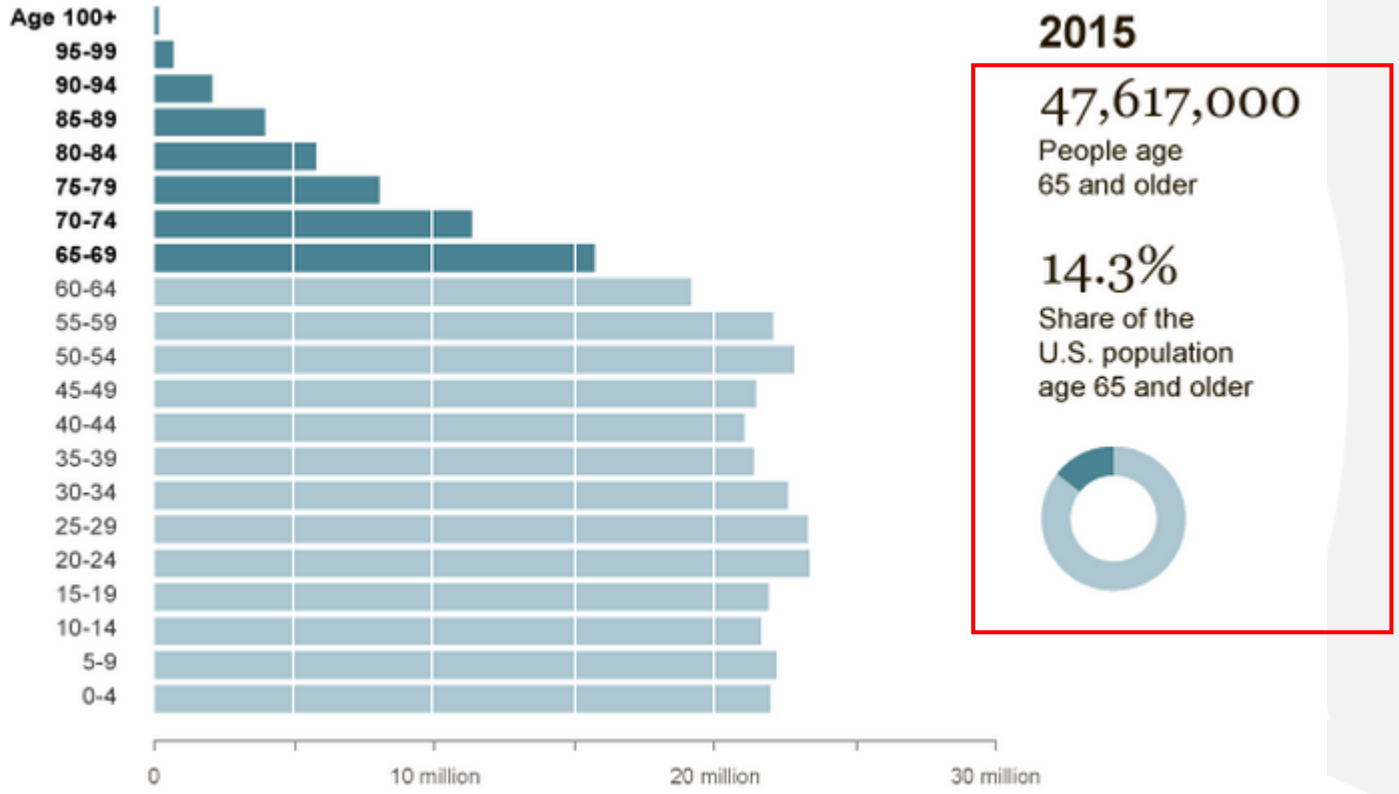
Source: http://www.dasheroo.com/blog/qwickbyte-why-you-should-avoid-three-dimensional-bar-charts/

Data visualisation best practices

- Most data presentations and graphs can be improved by simply removing visual distractions (removing ‘chart junk’).
- Communicate clearly and simply.
- Use colour in a deliberate way.
- Take time to provide the context of a visualisations.
- Avoid unnecessary embellishment and focus on trying to find the best, simplest method for transmitting your message.

Data visualisation best practices

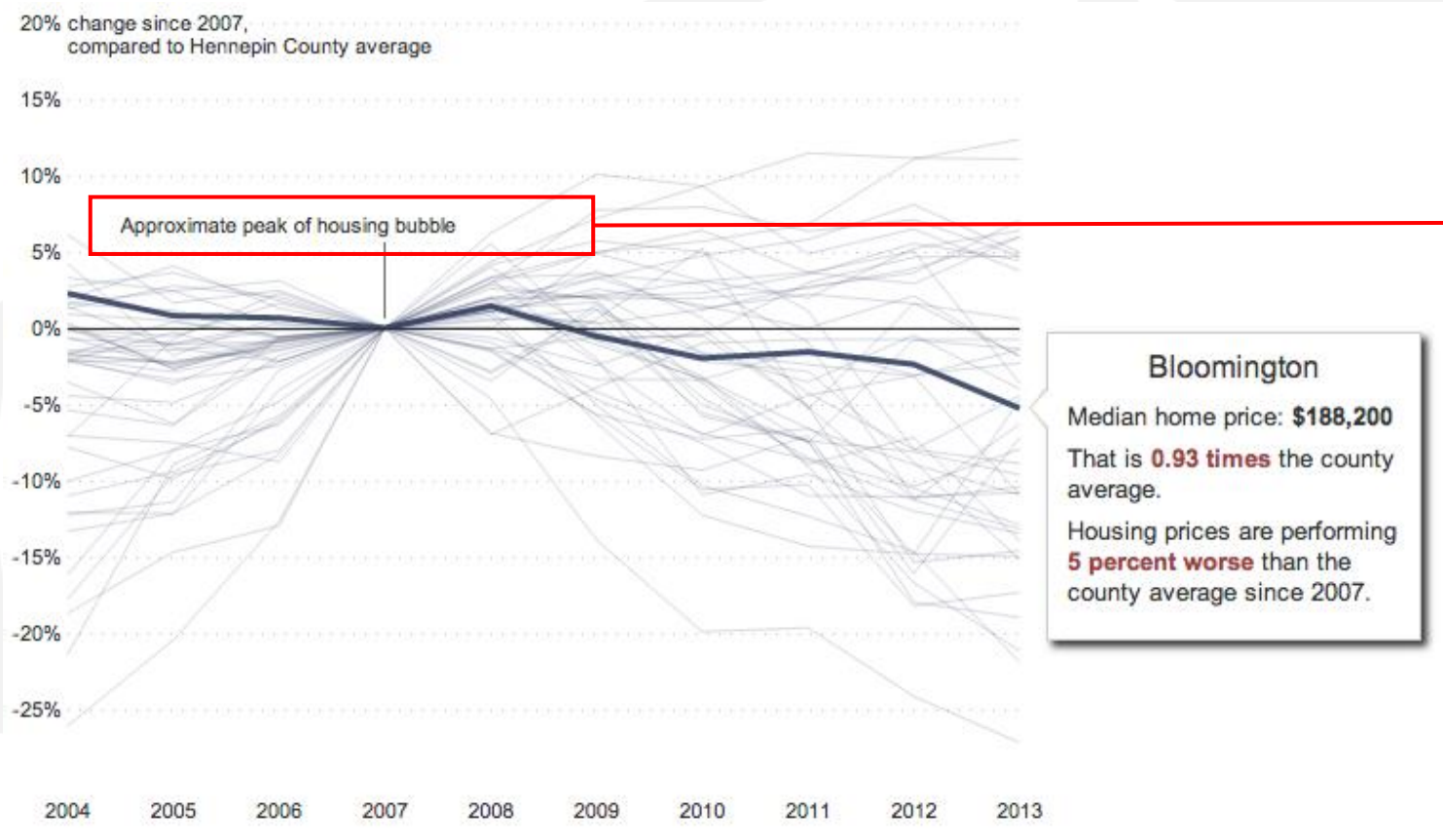
Diakopoulos, N 2013. “Storytelling with data visualization: Context is king”, *Tow Center for Digital Journalism*. Link: http://towcenter.org/storytelling-with-data-visualization-context-is-king/



This “**observational**” annotation provides the **context** of the visualisation by supporting reflection on a data value or group of values that are already depicted in the visualisation.

Data visualisation best practices

Diakopoulos, N 2013. "Storytelling with data visualization: Context is king", *Tow Center for Digital Journalism*. Link: http://towcenter.org/storytelling-with-data-visualization-context-is-king/



This "**additive**" annotation provides the **context** of the visualisation by including external information that is not clearly depicted in the visualisation yet highly relevant to understanding the information.

Phase 6 - Operationalise

Goals

The data science team delivers final reports, briefings, code, and technical documents.

The team may run a pilot project to implement the models in a production environment.

Developing deliverables for multiple audiences

- There are at least two (2) categories of audience: **executive** or **technical** audiences.
- Different stakeholders have different roles, interests and stakes in the analytic project. As such, the project output needs to be tailored to a specific audience.

Developing deliverables for multiple audiences

Analytic project stakeholders.

+ **Business User** | Someone who understands the domain areas and usually benefits from the result.

+ **Project Sponsor** | Responsible for the birth of the project.

+ **Project Manager** | Ensures that key milestones and objectives are met on time and at the expected quality.

+ **Business Intelligence Analyst** | Provides business domain expertise based on a deep understanding of the data, KPIs, key metrics, and business intelligence from a reporting perspective.

+ **Database Administrator** | Provisions and configures the database environment to support the analytics needs of the working team.

+ **Data Engineer** | Leverage deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox.

+ **Data Scientist** | Provides subject matter expertise for analytical techniques, data modeling, and applying valid analytical techniques to given business problems.

Developing deliverables for multiple audiences

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Project goals	List top 3 – 5 agree-upon goals.	
Main findings	Emphasize key messages	
Approach	High-level methodology	• High-level methodology • Relevant details on modelling techniques and technology
Model description	Overview of the modelling techniques	
Key points supported with data	Support key points with simple charts and graphics (e.g. bar chart)	• Show details to support the key points • Analyst-oriented charts and graphs, such as ROC curves and histograms • Visuals of key variables and significance of each

Developing deliverables for multiple audiences

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Model details	Omit, or discuss only at a high level	• Show the code or main logic of the model, and include model type, variables, and technology used to execute the model and score data • Identify key variables and impact of each • Describe expected model performance and any caveat • Details description of the modelling technique • Discuss variables, scope, and predictive power

Developing deliverables for multiple audiences

Presentation Component	Project Sponsor Presentation	Analyst Presentation
Recommendations	Focus on business impact, including risks and ROI. Give the sponsor highly relevant and actionable points to help her champion the proposed analytic solution within the organisation.	Supplement recommendations with implications for the modelling or for deploying in a production environment.

ROI (Return On Investment): A ratio between net profit and cost of investment.

Model deployment best practices

Brownlee, J 2016, 'Deploy your predictive model to production', *Machine Learning Mastery*
Link: http://machinelearningmastery.com/deploy-machine-learning-model-to-production/

Assume that you have built and trained an analytic model (either K-Means clusters, Naïve Bayes, or Linear Regression model).

To deploy the model means to:

- create a standalone program that make ad-hoc predictions using the model, **or**
- incorporate the model into your existing software or enterprise systems to support regular business decisions.

Either way, there are some **best practices** to follow.

Model deployment best practices

Brownlee, J 2016, 'Deploy your predictive model to production', *Machine Learning Mastery*
Link: http://machinelearningmastery.com/deploy-machine-learning-model-to-production/

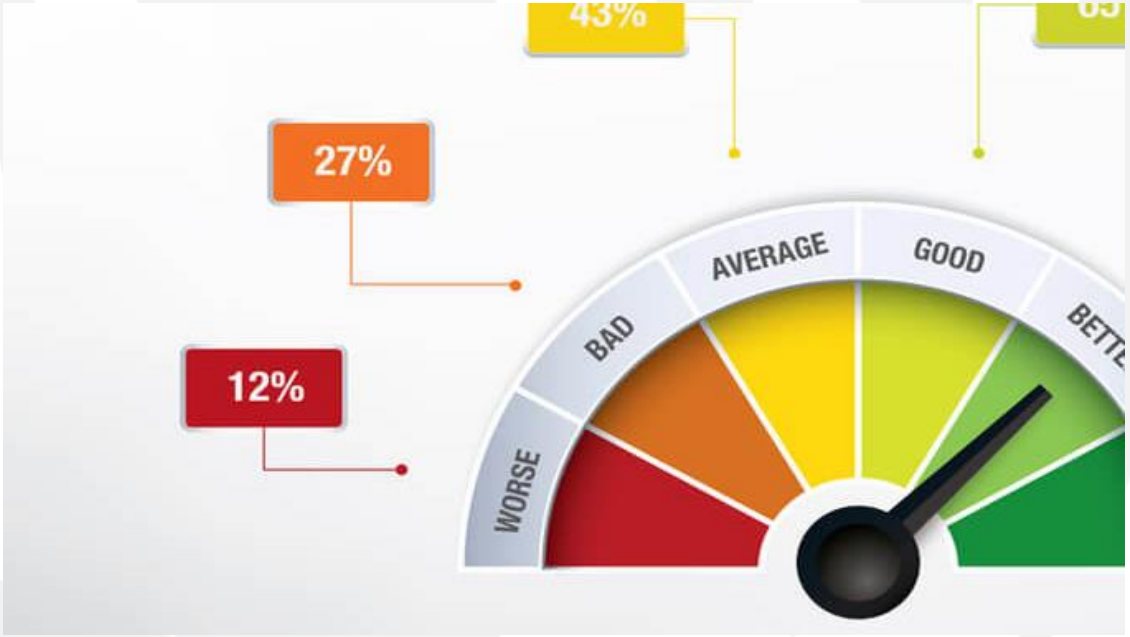
Best Practice #1

Specify performance requirements

Decide what constitutes good and bad performance of the model. Express these in terms of the evaluation metrics scores (accuracy, TPR, precision, etc.).

Specifying the performance requirements is useful as they allow you to:

- assess if the model performs as expected in the deployment setting
- continuously improve the model in the future



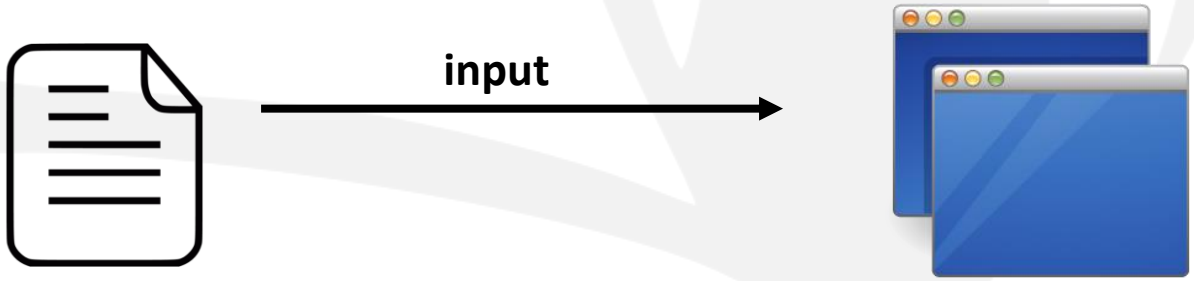
Model deployment best practices

Brownlee, J 2016, 'Deploy your predictive model to production', *Machine Learning Mastery*
Link: http://machinelearningmastery.com/deploy-machine-learning-model-to-production/

Best Practice #2

Separate model coefficients from the computer program that will make the prediction.

Treat the learned model coefficients as external inputs to your software (similar to software configuration settings). DO NOT "hardcode" the model coefficients in your program as it will complicate the process of updating your model next time.



model_coefficients.txt **a computer program that implements the model**

Model deployment best practices

Brownlee, J 2016, 'Deploy your predictive model to production', *Machine Learning Mastery*
Link: http://machinelearningmastery.com/deploy-machine-learning-model-to-production/

Best Practice #3

Develop automated tests for your model

The goal is to quickly check if the model continues to work at the expected level of performance. To build automated tests:

1. Collect a small sample of data on which to make predictions.
2. Use the program that implements the model to make prediction on this small sample data.
3. Confirm that the program produces your expected results.

The idea is that if the program performs worse than the expected level in this automated test, the model is broken and should not be immediately deployed in real business applications.

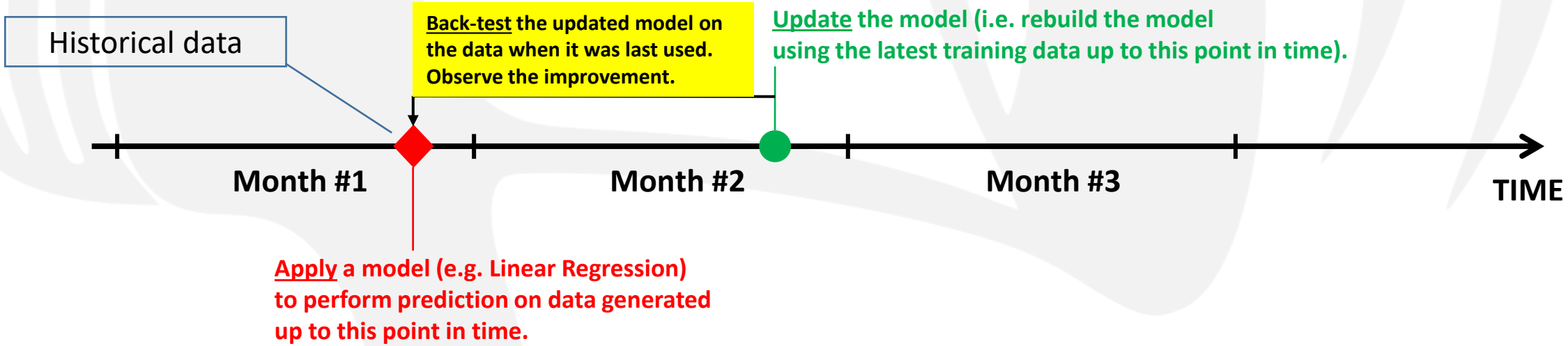
Model deployment best practices

Brownlee, J 2016, 'Deploy your predictive model to production', *Machine Learning Mastery*
Link: http://machinelearningmastery.com/deploy-machine-learning-model-to-production/

Best Practice #4

Develop Back-Testing and Now-Testing infrastructure

- Data on which the predictions are made will change from time to time as new data are being generated. A **Back-Testing** infrastructure tests any new update to a model on historical data to [determine if the update has truly made improvement to the older model or not].



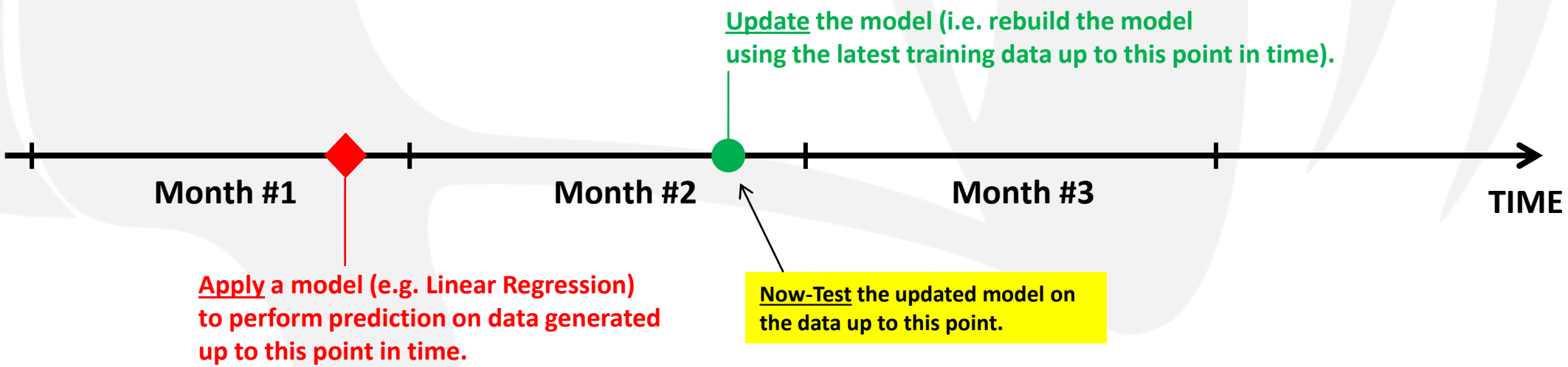
Model deployment best practices

Brownlee, J 2016, 'Deploy your predictive model to production', *Machine Learning Mastery*
Link: http://machinelearningmastery.com/deploy-machine-learning-model-to-production/

Best Practice #4

Develop Back-Testing and Now-Testing infrastructure

- Assuming that the updated model has outperformed a previous model in Back-Testing, subsequently perform Now-Testing. This is a test of the new model on the latest data. The goal is to [check if the new model is ready for deployment].



Model deployment best practices

Brownlee, J 2016, 'Deploy your predictive model to production', *Machine Learning Mastery*
Link: http://machinelearningmastery.com/deploy-machine-learning-model-to-production/

Best Practice #5

Strictly evaluate each model update

A model needs to be updated from time to time.

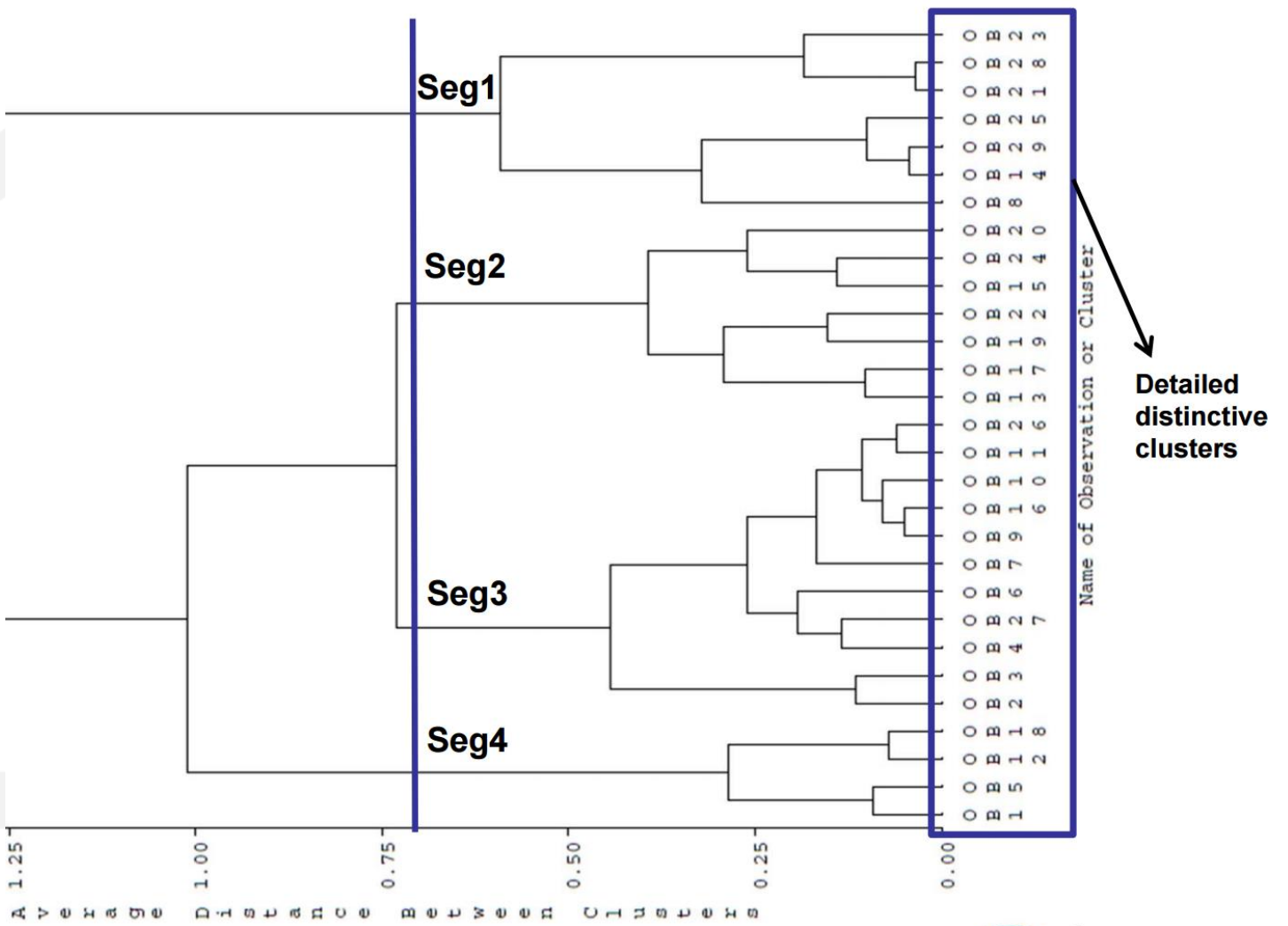
- Test each new update and be highly critical. Give the model every chance to fail.
- Apply Best Practices #1 – #4 above on each new model update.
- Roll out the new model within a limited production environment or in a 'beta release' to allow feedback. This mitigates risks.
- Accept the new model once it meets the set minimum performance requirements.
- Consider incrementally updating the minimum performance requirements as model performance improves.

Example:
K-Means Clustering deployment practices in customer segmentation applications

A popular application of K-Means Clustering algorithm is to segment customers into different groups of distinct characteristics.

A practical challenge with using K-Means is choosing the right number of clusters to generate.

The **Bottom-up approach** initially generates a flexible number of clusters (i.e. segments) and then aggregates similar clusters to form larger clusters that maximize the average distance between clusters.



Source: SAS, https://goo.gl/2hYGdQ

a LoyaltyOne program

K-Means Clustering deployment practices in customer segmentation applications

Chen, J 2014, 'Retail customer segmentation using SAS', Calgary SAS Users Group meeting
Link: https://goo.gl/2hYGdQ

After the clusters are generated ….

Segments profiling

- Using cluster building variables to profile the segments (e.g. RFM framework – *Recency, Frequency, Monetary)*.
- Use additional data source to profile them.
- Paint a clear picture of the segments.

K-Means Clustering deployment practices in customer segmentation applications

Chen, J 2014, 'Retail customer segmentation using SAS', Calgary SAS Users Group meeting
Link: https://goo.gl/2hYGdQ

Validate segmentation effectiveness

- Test campaigns with segmenting strategy.
- Measure campaign results by segments.
- Validate segments with market research survey.

Example: https://towardsdatascience.com/clustering-algorithms-for-customer-segmentation-af637c6830ac

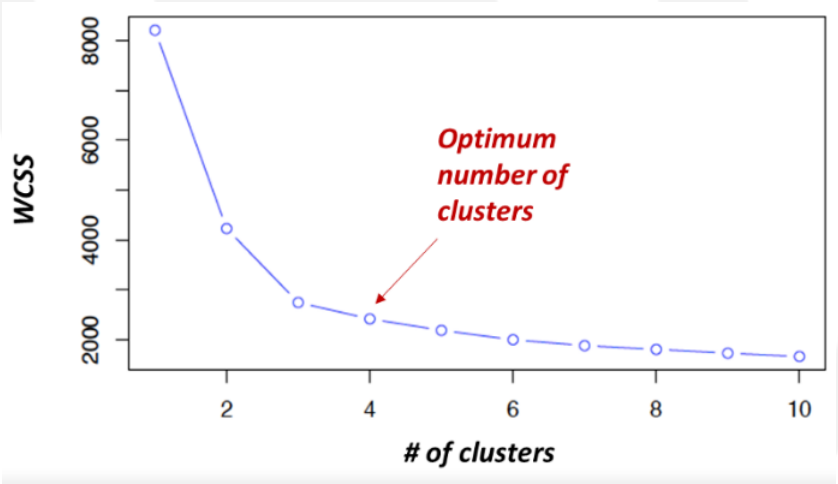
Example: k-means clustering for customer segmentation

Context

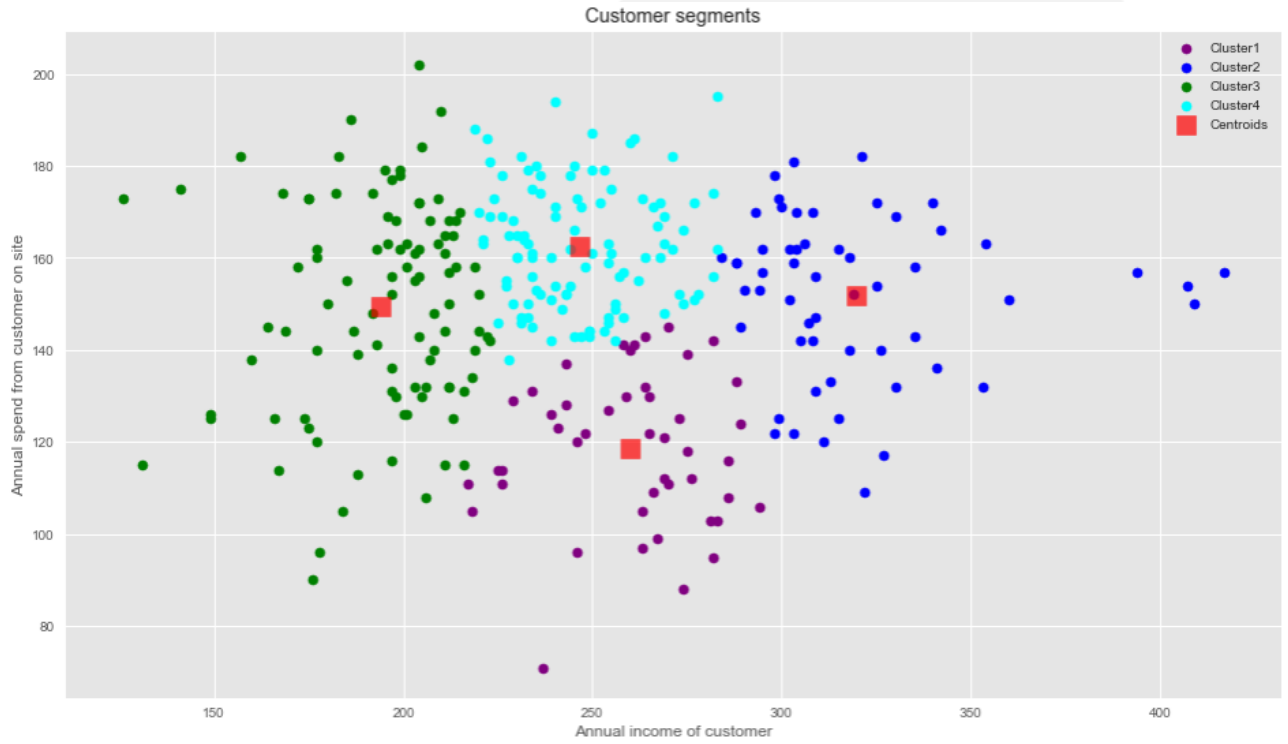
In today's competitive world, it is crucial to understand customer behavior and categorize customers based on their demography and buying behavior. This is a critical aspect of customer segmentation that allows marketers to better tailor their marketing efforts to various audience subsets in terms of promotional, marketing and product development strategies.

About the data set

The dataset consists of Annual income (in $000) of 303 customers and their total spend (in $000) on an e-commerce site for a period of one year. Let us explore the data using *numpy* and *pandas* libraries in python.

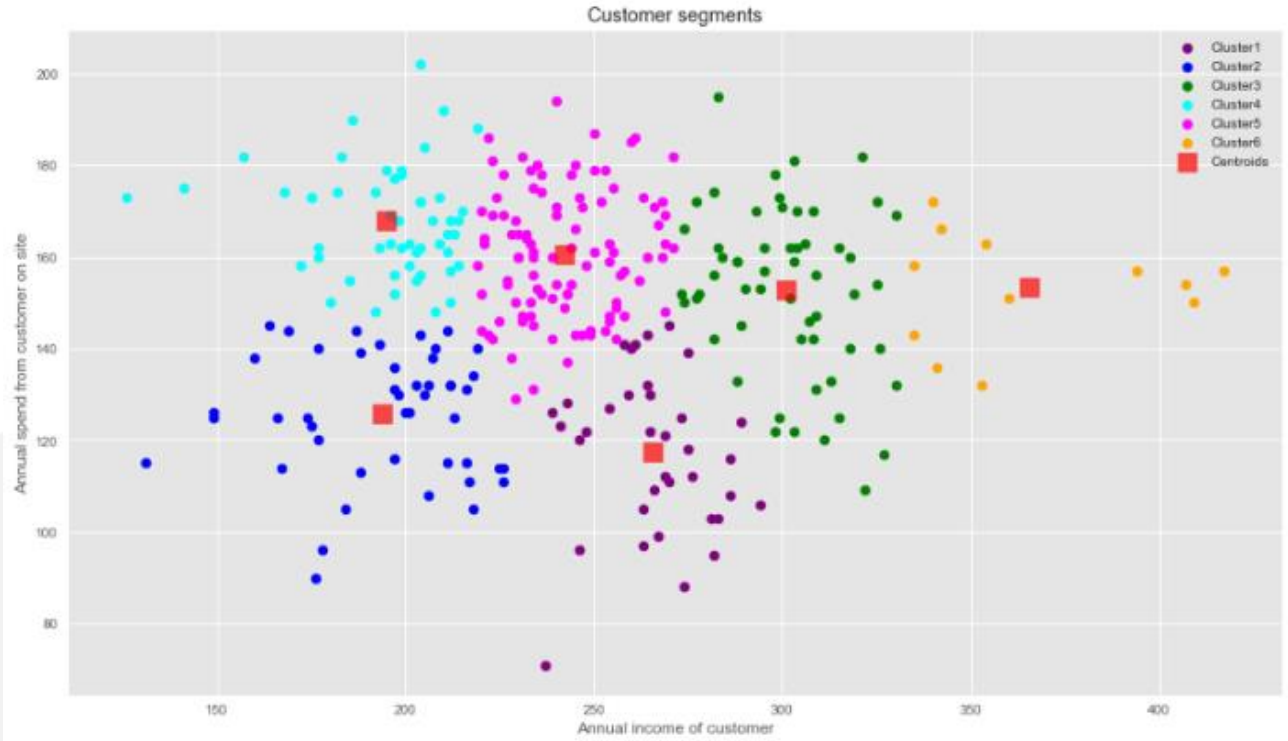


WCSS = Within Cluster Sum of Square



Segments profiling
Using cluster building variables to profile the segments

1. Cluster 1: Customers with medium annual income and low annual spend
2. Cluster 2: Customers with high annual income and medium to high annual spend
3. Cluster 3: Customers with low annual income
4. Cluster 4: Customers with medium annual income but high annual spend



Cluster plot : k=6

Segments profiling:
Setting the number of clusters to 6 provide more meaningful customer segmentation

1. Cluster 1: Medium income, low annual spend
2. Cluster 2: Low income, low annual spend
3. Cluster 3: High income, high annual spend
4. Cluster 4: Low income, high annual spend
5. Cluster 5: Medium income, low annual spend
6. Cluster 6: Very high income, high annual spend

1. Cluster 1: Medium income, low annual spend
2. Cluster 2: Low income, low annual spend
3. Cluster 3: High income, high annual spend
4. Cluster 4: Low income, high annual spend
5. Cluster 5: Medium income, low annual spend
6. Cluster 6: Very high income, high annual spend

Marketing strategies for the customer segments

Based on the 6 clusters, we could formulate marketing strategies relevant to each cluster:

- A typical strategy would focus certain promotional efforts for the high value customers of Cluster 6 & Cluster 3.
- Cluster 4 is a unique customer segment, where in spite of their relatively lower annual income, these customers tend to spend more on the site, indicating their loyalty. There could be some discounted pricing based promotional campaigns for this group so as to retain them.
- For Cluster 2 where both the income and annual spend are low, further analysis could be needed to find the reasons for the lower spend and price-sensitive strategies could be introduced to increase the spend from this segment.
- Customers in clusters 1 and 5 are not spending enough on the site in spite of a good annual income—further analysis of these segments could lead to insights on the satisfaction / dissatisfaction of these customers or lesser visibility of the e-commerce site to these customers. Strategies could be evolved accordingly.

K-Means Clustering deployment practices in customer segmentation applications

Chen, J 2014, 'Retail customer segmentation using SAS', Calgary SAS Users Group meeting
Link: https://goo.gl/2hYGdQ

Review, iterate, deploy again

- Update segments regularly.
- Monitor the migration of segments.
- Gather feedback and campaign response results.
- Maintain and improve the process and model.

K-Means Clustering deployment practices in customer segmentation applications

Another good practical example of deploying K-Means Clustering for customer segmentation:

http://www.kimberlycoffey.com/blog/2016/8/k-means-clustering-for-customer-segmentation

The article contains some R code snippets. Don’t let the codes intimidate you; importantly, read the author’s discussions.

Texts and Resources

Unless stated otherwise, the materials presented in this lecture are taken from:

- Dietrich, D. ed., 2015. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*. EMC Education Services.