BIBA: Business Intelligence and Big Data

2018-10-24

Jens Ulrik Hansen

Storytelling with data

Today’s program

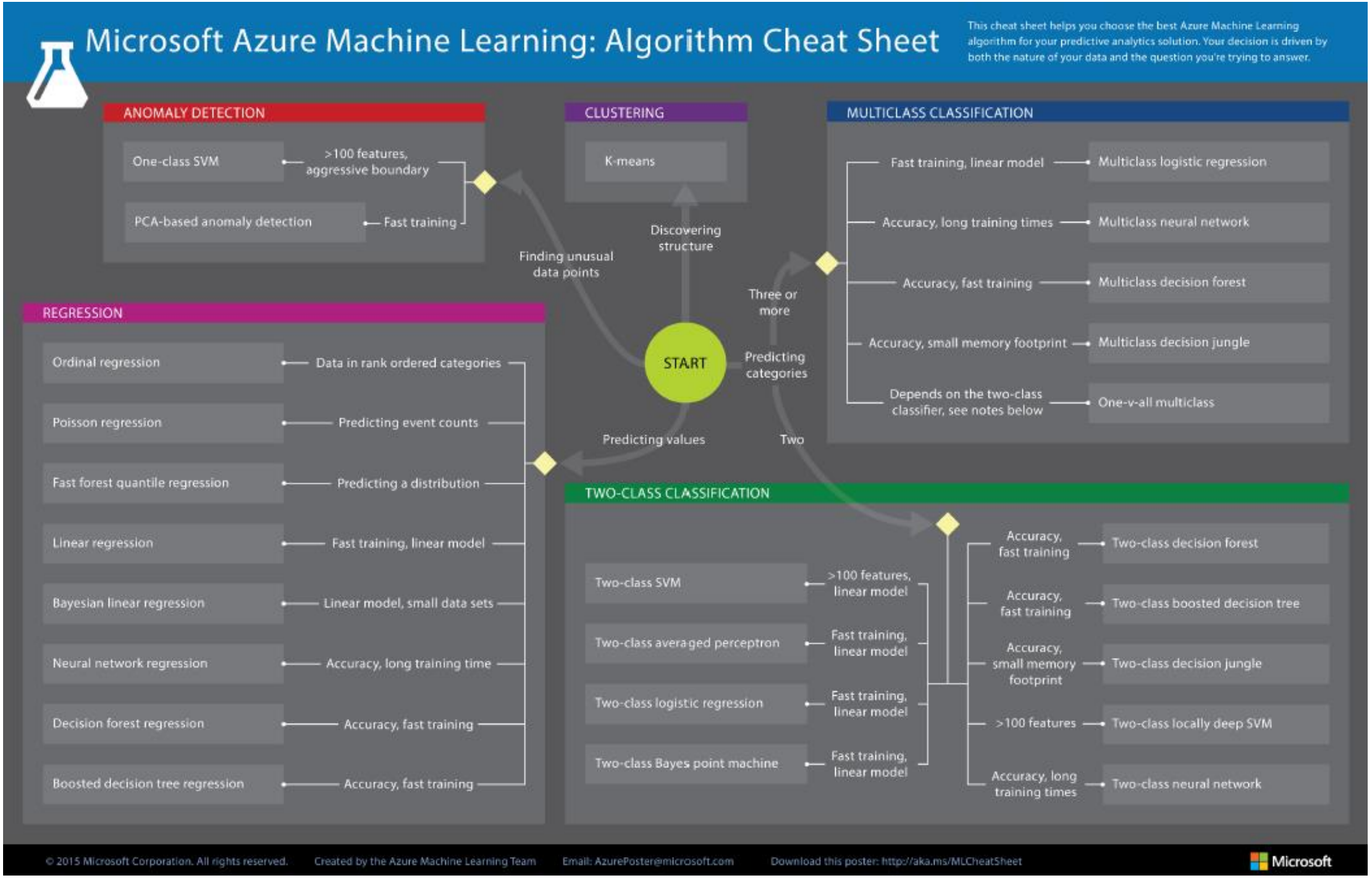
- General feedback on the modeling synopsis hand-in
- Recap on the data science process
- Storytelling with data
- “Storytelling with data” – a few examples in R
- Reporting and Dashboard
- Other BI tools

General feedback on the modeling synopsis hand-in

General feedback on the modeling synopsis hand-in

- Only 6 hand-ins this time
- All of them are on the right track, but some have progressed more on the modeling part than others
- I can provide some individual comments at the end or in the breaks

https://docs.microsoft.com/en-us/azure/machine-learning/studio/algorithm-cheat-sheet



Recap on the data science process

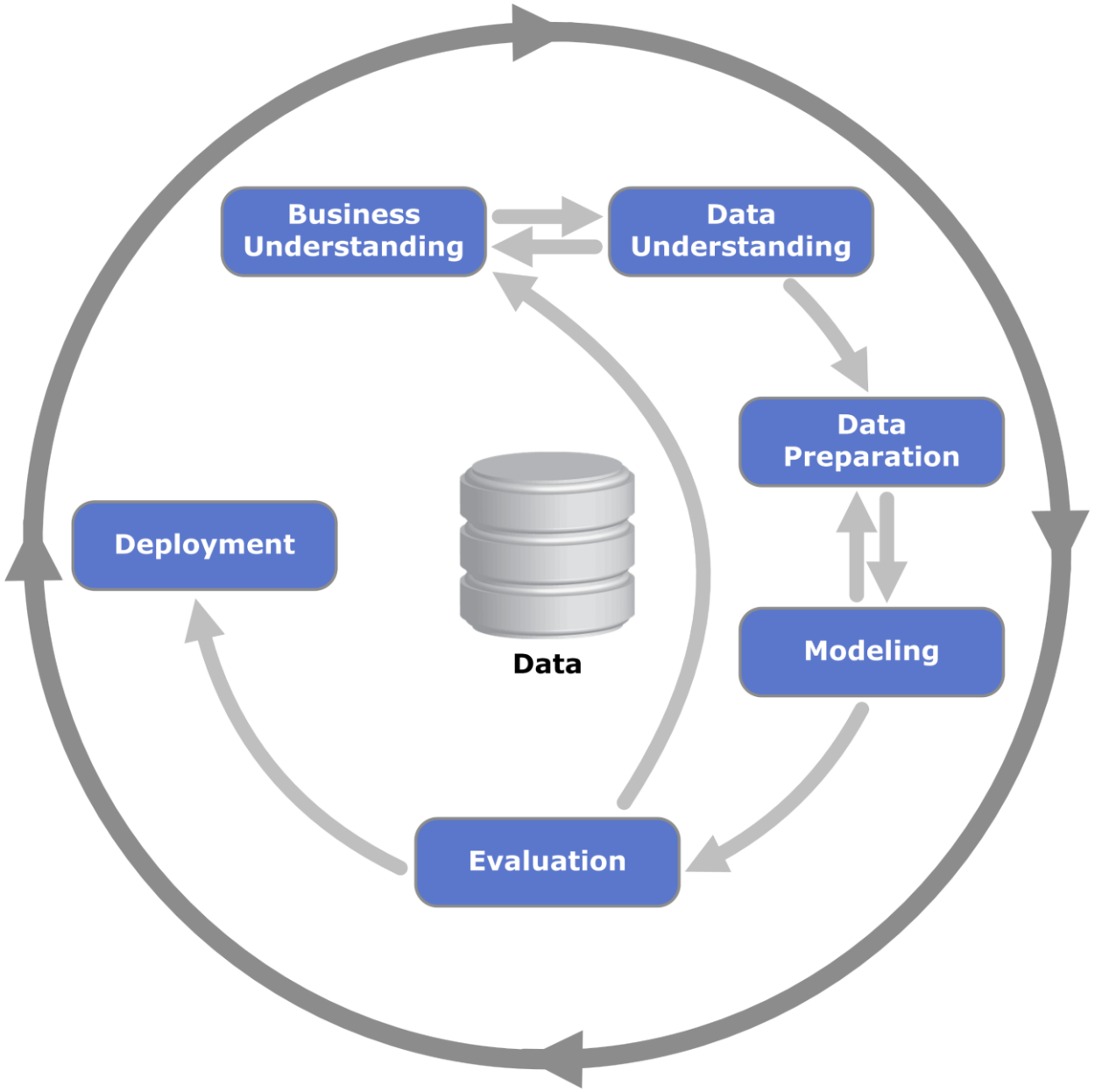
The data science/BI process recap

- Using data (and data analysis) to solve business problems
  1. identify business problem
  2. collect data
  3. prepare data
  4. analyze data
  5. conclude and communicate

- Other models of the Data Science Process: CRISP-DM, KDD, SEMMA, TDSP, …
- The Data Science Process is different from the Software Development Process

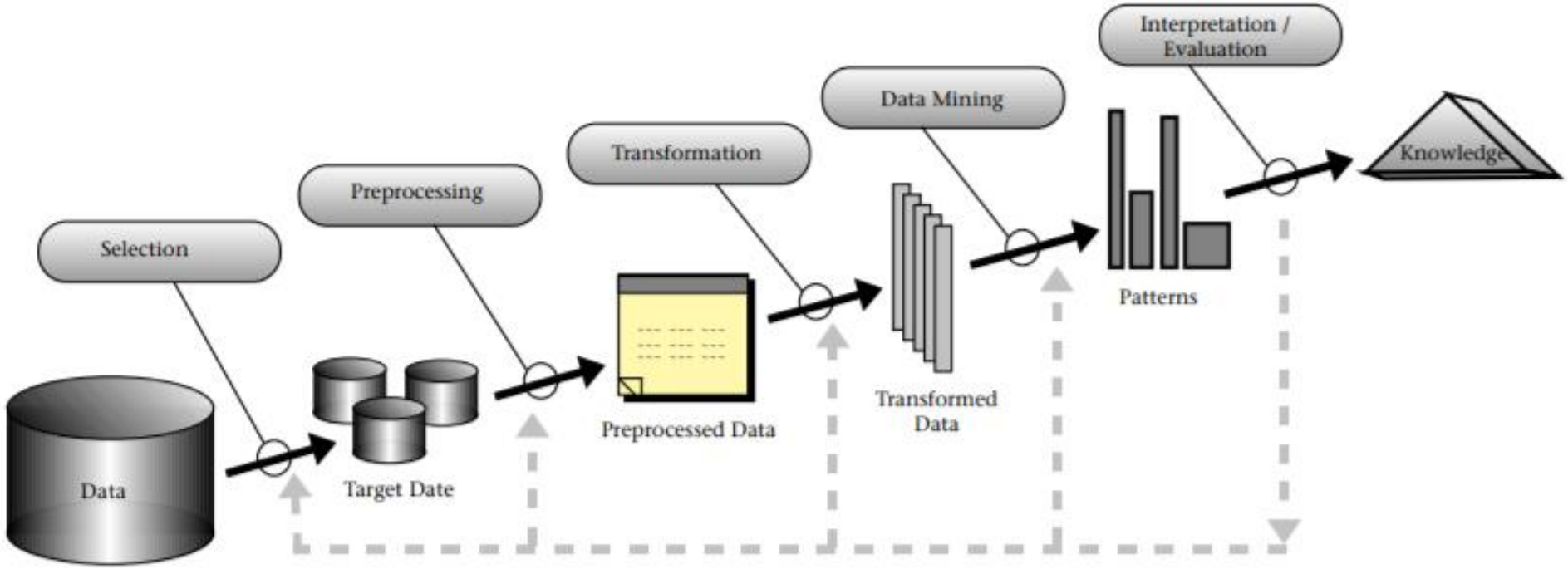
CRISP-DM

- ***Cross-industry standard process for data mining***
  - https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining



KDD

- ***Knowledge Discovery in Databases***
  - https://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

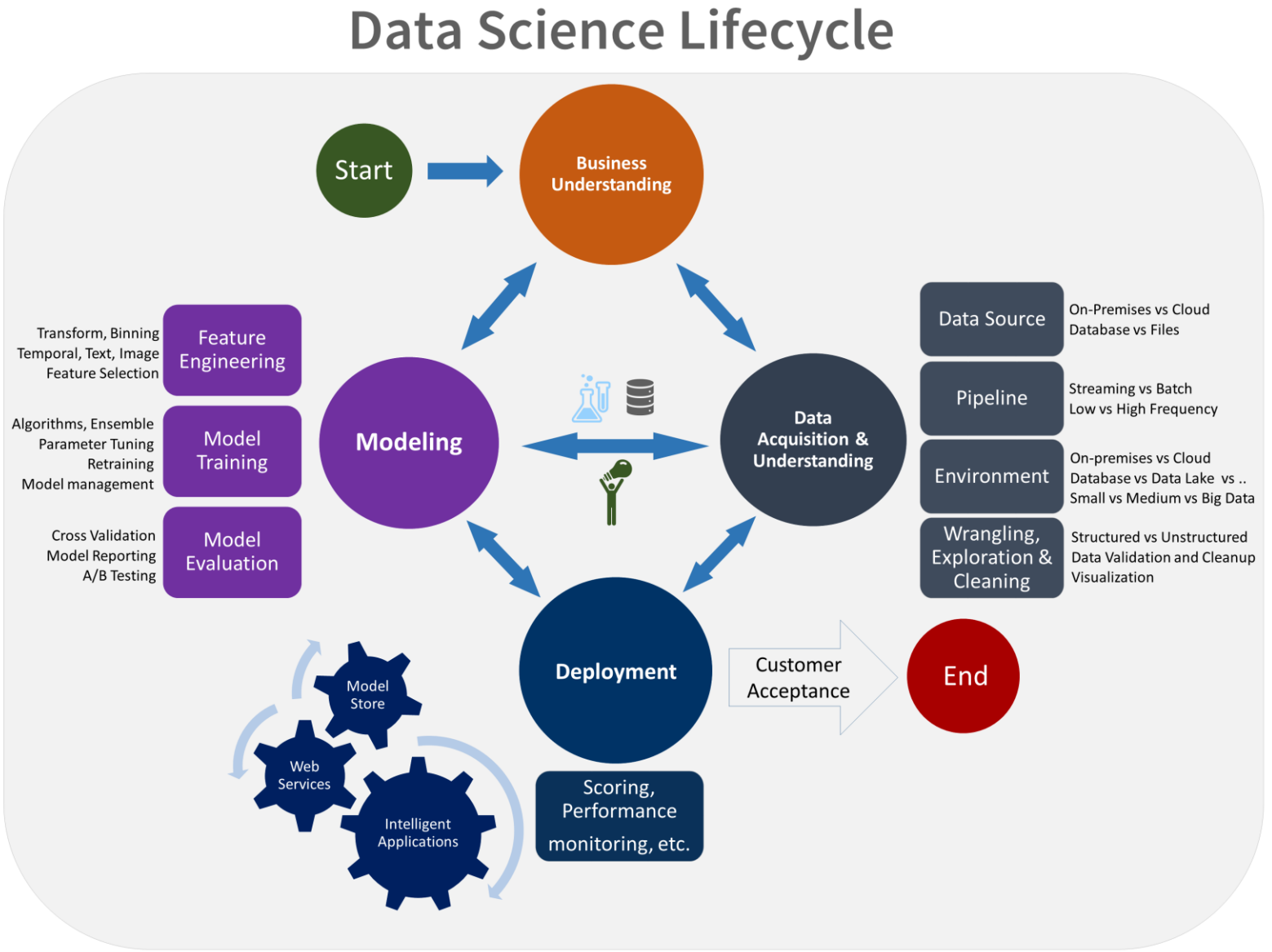


SEMMA

- ***Sample, Explore, Modify, Model, and Assess***
  - SAS Institute
  - https://en.wikipedia.org/wiki/SEMMA

TDSP

- ***Team Data Science Process***
  - Microsoft
  - https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview



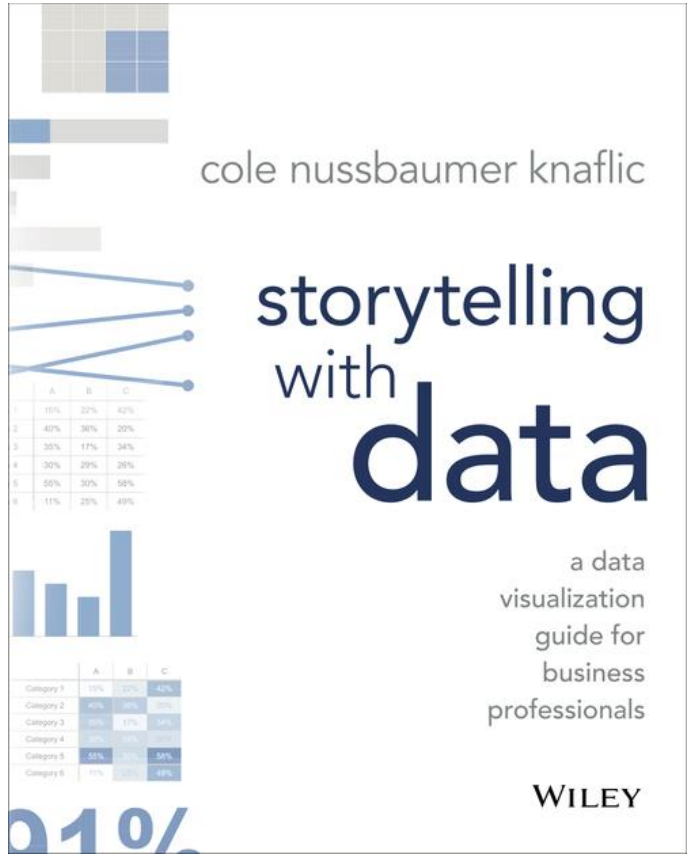
Storytelling with data

Storytelling with data

- Why do we need storytelling when visualizing data?
  - If we want actions to happen, proper storytelling is effective
- Do not chose a particular visualization based on the type of data you have, but chose it based on the story you want to tell
- Storytelling with data vs visualization
  - Visualization is a fantastic tool for storytelling (with data)
  - But storytelling with data is more than just pretty visualizations
  - (And data visualizations can be used for more than just storytelling)
- The visualizations/presentation is the only aspect of your analysis your audience see
  - No one ever sees (or care about) the super nice code you wrote to clean the data or the clever trick to did to reduce the model error by 50%
- Avoid to heavy cognitive load of your audience

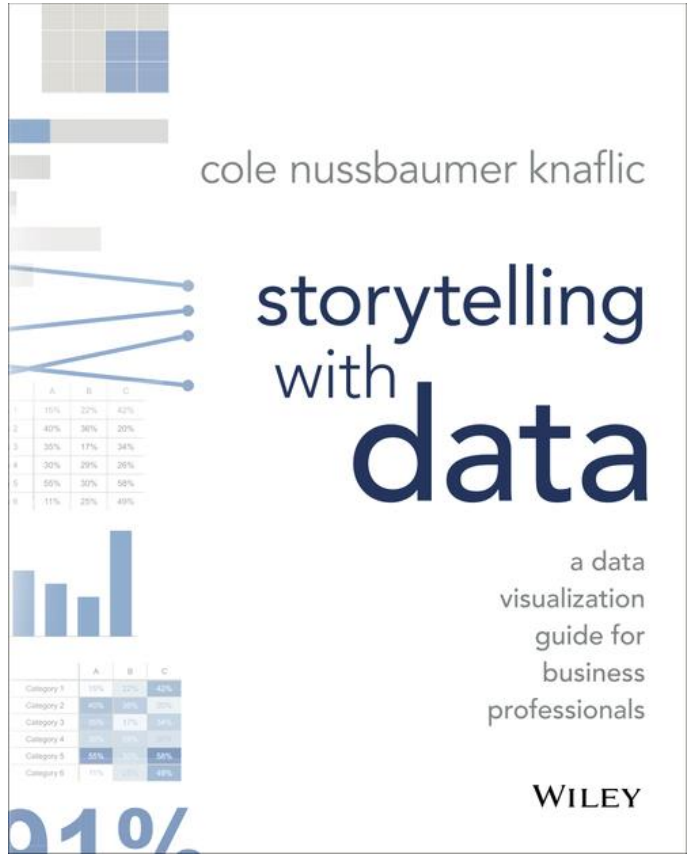
Storytelling with data

- Cole Nussbaumer Knaflic's book: "Storytelling with data"
- Examples of great visualizations and storytelling with data:
  - Ted Ed Talk by David McCandless: https://www.youtube.com/watch?v=5Zg-C8AAIGg
  - Ted Talk by Hans Rosling: https://www.ted.com/talks/hans_rosling_at_state
  - All Hans Rosling's other Ted Talks



Storytelling with data – 6 key lessons

1. Understand the context
2. Choose an appropriate visual display
3. Eliminate clutter
4. Focus attention where you want it
5. Think like a designer
6. Tell a story



1. Understand the context

- There is a difference between *Exploratory* data analysis and *Explanatory* data analysis
  - In Explanatory data analysis you have to convince other than yourself
- Before you begin to visualize data, you need to understand the context (for the need to communicate).
- Ask the following questions:
  - Who is your audience?
  - What do you need them to know or do?
  - What data is available that will help make my point?

  -> The Who, What, How

2. Choose an appropriate visual display

- More on this in a moment…

3. Eliminate clutter

- Every single thing you add to a blank screen will increase the cognitive load of your audience
  - Thus, eliminate everything that is not necessary

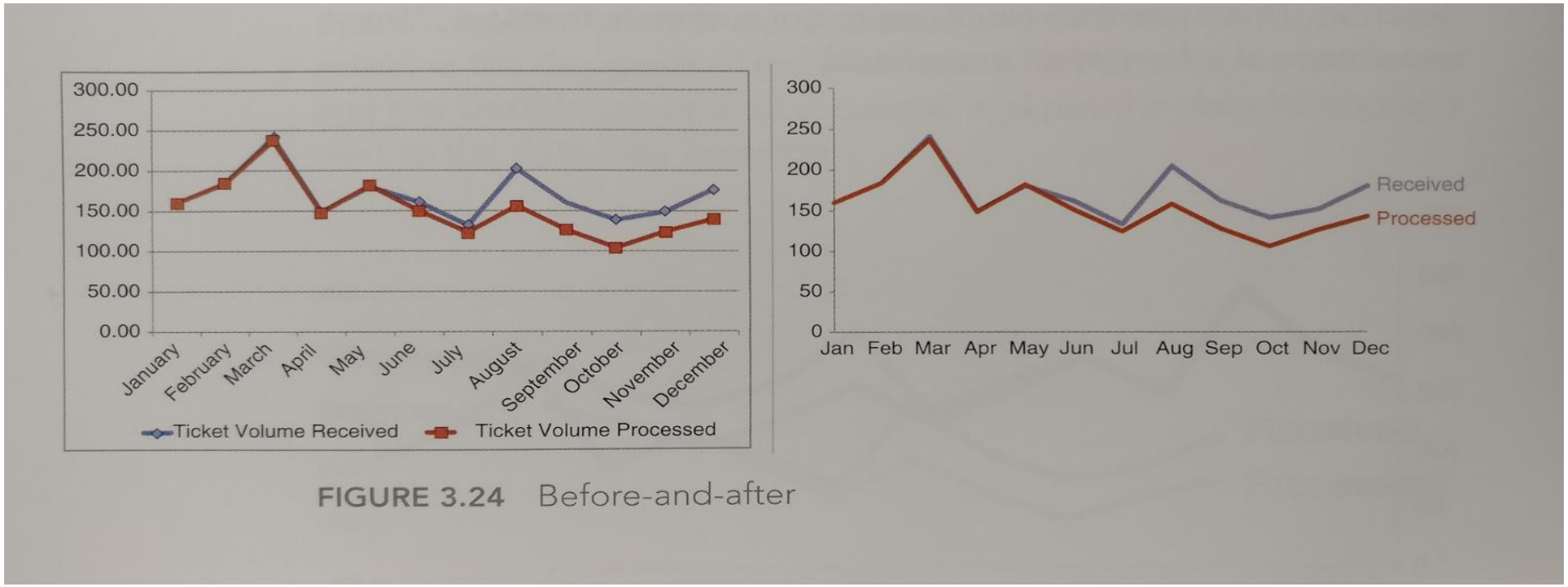


FIGURE 3.24 Before-and-after

4. Focus attention where you want it

- Visual ques can be used to focus our audience's attention to where we want it
- Our iconic memory is drawn to preattentive attributes such as color, size, and position on page
  - We immediately focus our attention at the color that stands out, for instance
- Proximity and similarity in color between the visual data and the corresponding text
- We usually start in the upper left corner and do zigzagging across to the lower right corner without any other visual ques

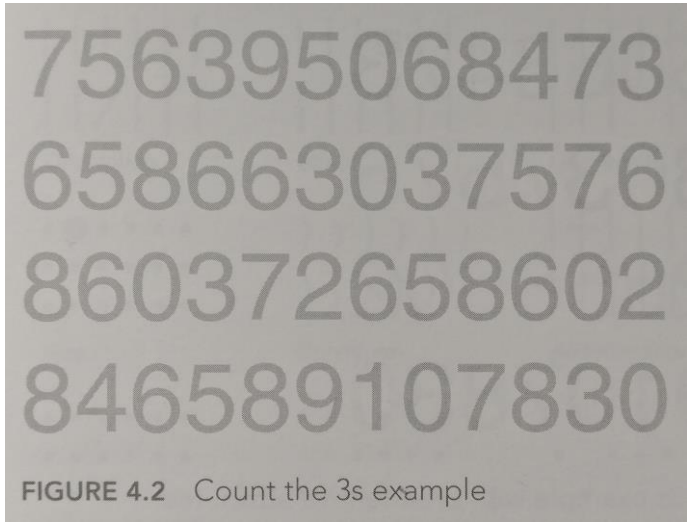


FIGURE 4.2 Count the 3s example

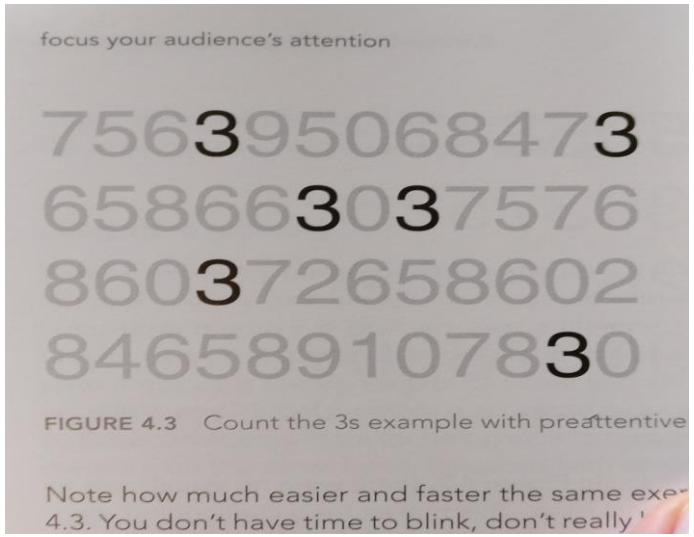


FIGURE 4.3 Count the 3s example with preattentive

Note how much easier and faster the same exe
4.3. You don't have time to blink, don't really

5. Think like a designer

- General principles of design applies...

6. Tell a story

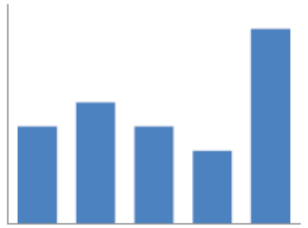
- A story we remember in ways we cannot remember data
- A story can help bring data/facts to life
  - the story needs to be clearly visible in the graph as well as the call to action
- A story has
  - A plot: What context is essential to understand?
  - Twists: What is interesting about the data and what it shows?
  - Ending: What do you want your audience to do?
    - A particular call to action or just start a conversation

2. Choose an appropriate visual display

- What is the best way to show the data you want to communicate?
- See http://www.datavizproject.com/ for a ”complete” list of possible graphs

Simple text


Scatterplot

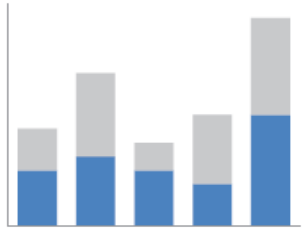

Vertical bar

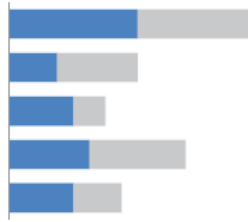

Horizontal bar

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Table

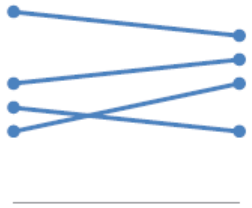

Line

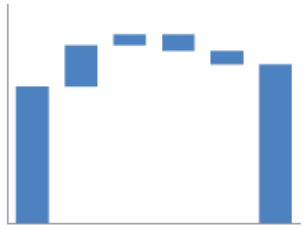

Stacked vertical bar

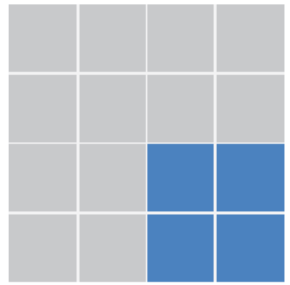

Stacked horizontal bar

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

Heatmap

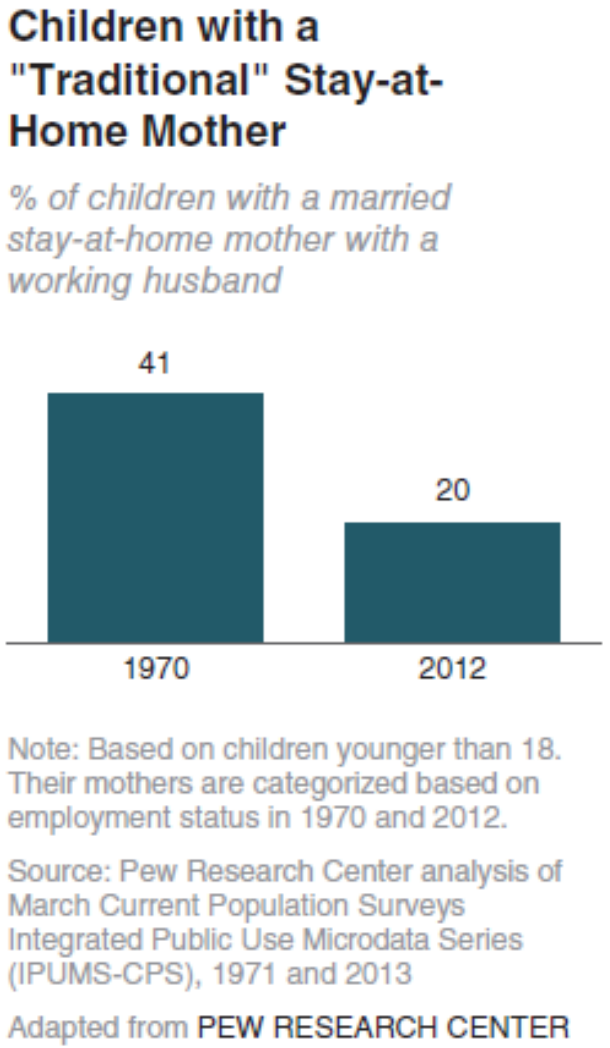

Slopegraph


Waterfall


Square area

2. Choose an appropriate visual display

- Simple text
  - When you have a number or two to share
  - Comparing two numbers can be easier



2. Choose an appropriate visual display

- Tables
  - Interact with our verbal system – we *read* them
  - Reading along columns or rows, compare numbers
  - Also good when there are numbers in different units
  - Not so good in a live presentation
  - “Don’t let heavy borders or shading compete for attention” – the data should stand out

Heavy borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Light borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

Minimal borders

Group	Metric A	Metric B	Metric C
Group 1	$X.X	Y%	Z,ZZZ
Group 2	$X.X	Y%	Z,ZZZ
Group 3	$X.X	Y%	Z,ZZZ
Group 4	$X.X	Y%	Z,ZZZ
Group 5	$X.X	Y%	Z,ZZZ

2. Choose an appropriate visual display

- Heatmaps
  - A kind of table
  - Visualize data in a tabular form
  - Provide visuals cues to point of interest

Table

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

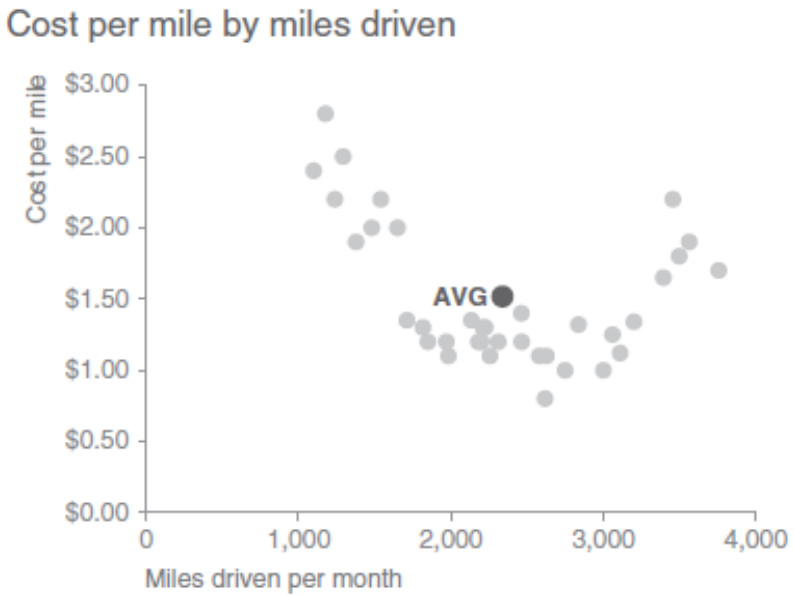
Heatmap

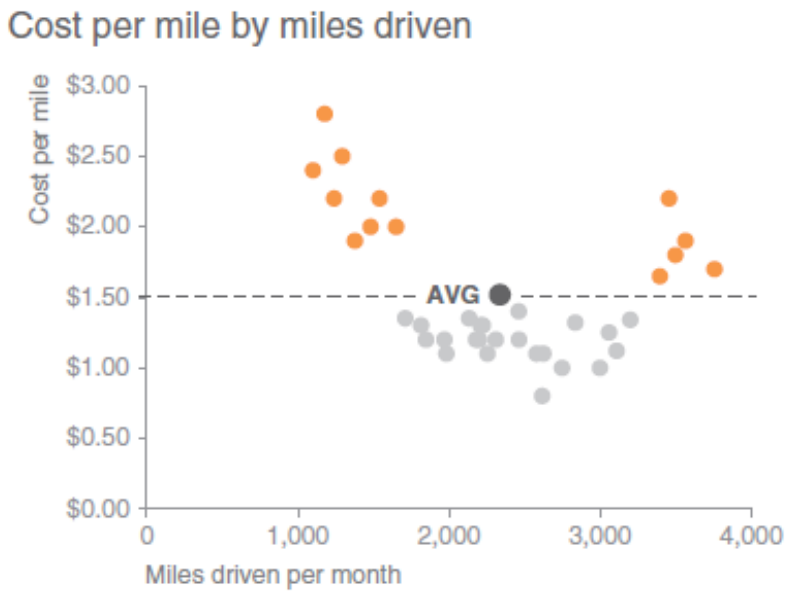
LOW-HIGH

	A	B	C
Category 1	15%	22%	42%
Category 2	40%	36%	20%
Category 3	35%	17%	34%
Category 4	30%	29%	26%
Category 5	55%	30%	58%
Category 6	11%	25%	49%

2. Choose an appropriate visual display

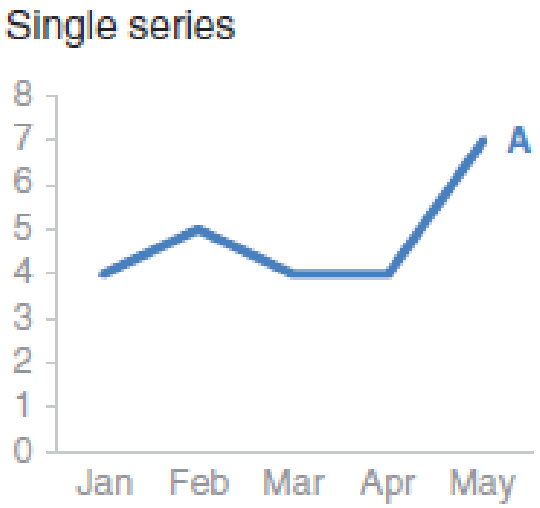
- Graphs
  - Speaks to our visual system which is faster at processing information
- Scatterplot
  - Useful for showing relationship between two numerical variables

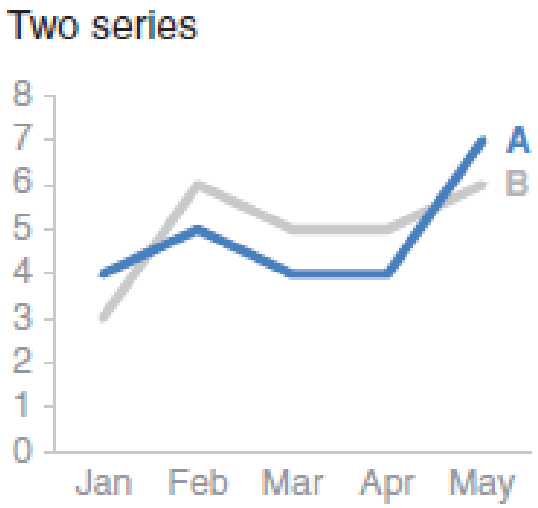


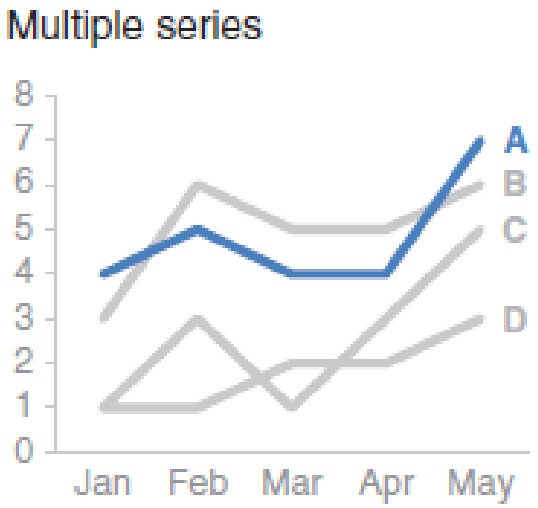


2. Choose an appropriate visual display

- Line graph
  - Useful with numerical data and time series data
  - Date/time on the x-axis – use consistent intervals
  - A line graph can also be used to visualize (confidence) intervals

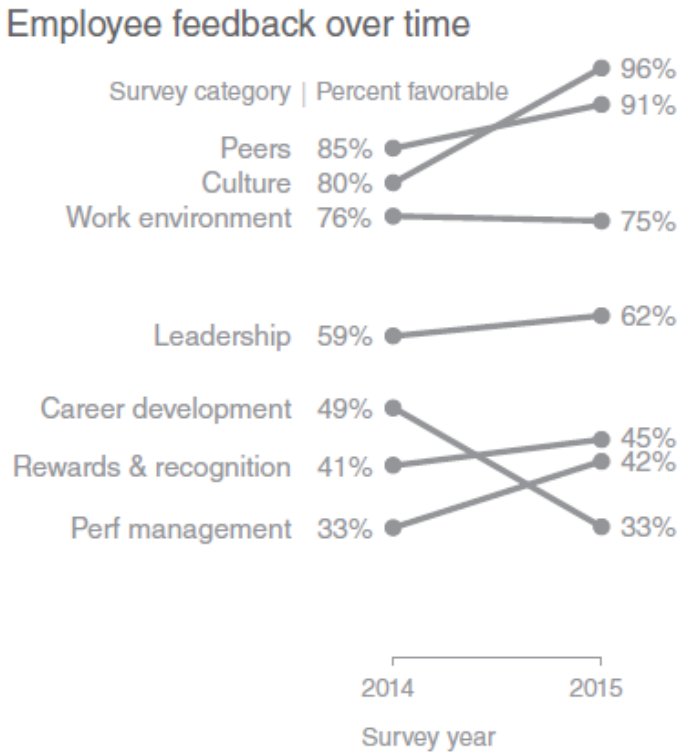


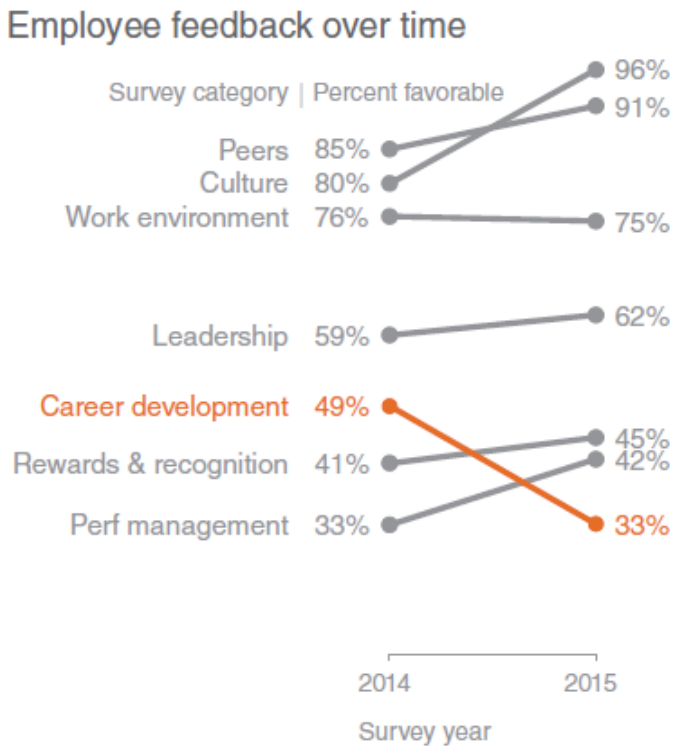




2. Choose an appropriate visual display

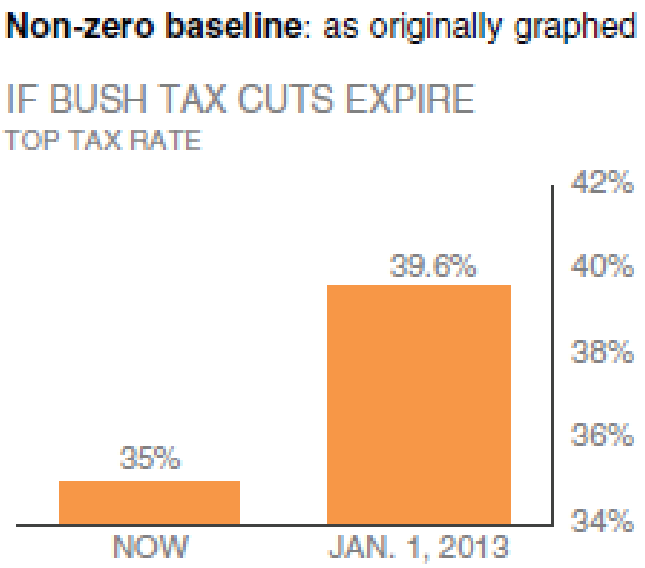
- Slope graph
  - When you have two time points or data points for comparison and want to show relative increase and decrease
  - Visualize decrease or increase
  - The decrease or increase is intuitively visible without further explanation
  - Not part of standard plotting packages
  - Can be problematic with many overlapping lines

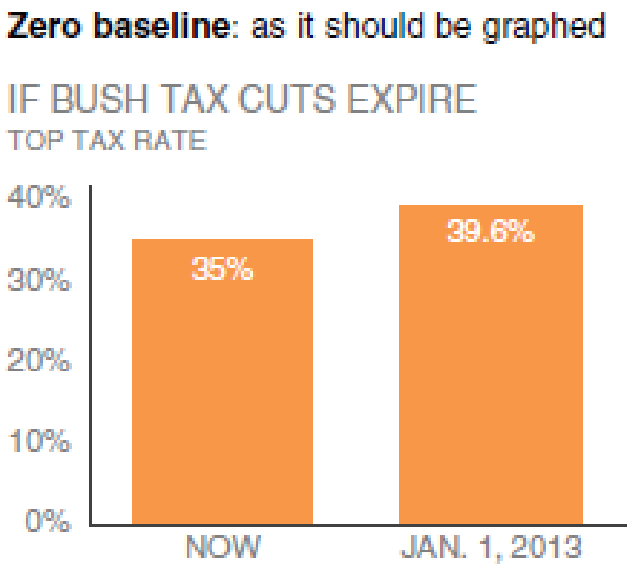




2. Choose an appropriate visual display

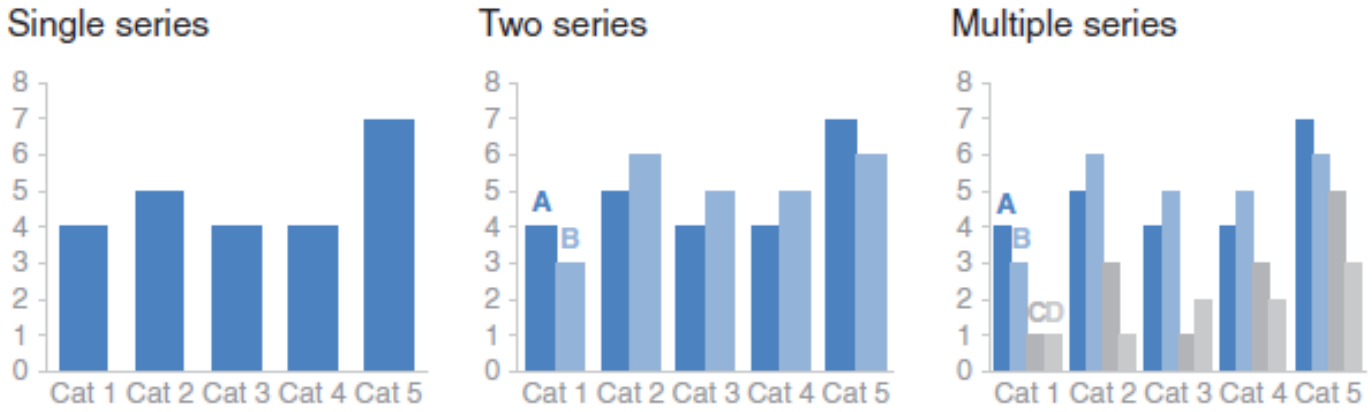
- Bar charts
  - Useful for categorical data
  - Easy for our eyes to read – it is easy to spot the highest and smallest bar
  - A zero baseline is important to not leave false impressions in the viewers mind
  - Width of the bars

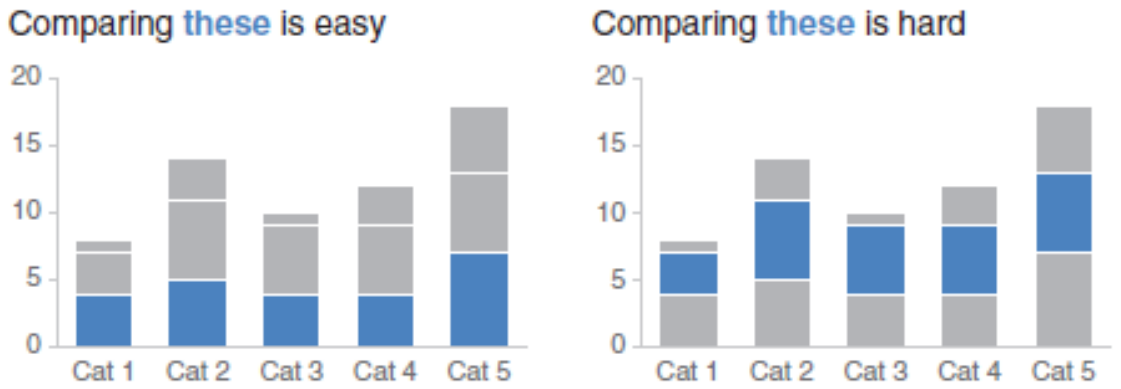




2. Choose an appropriate visual display

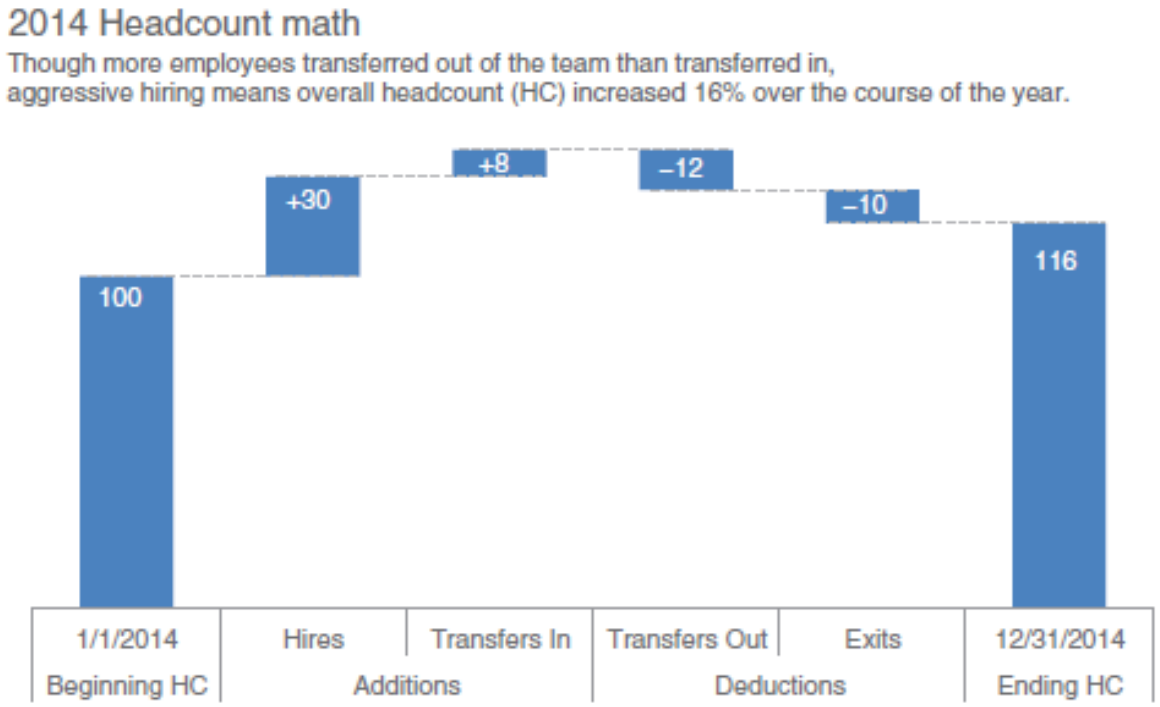
- Bar charts
  - Vertical bar chart
    - The standard
  - Stacked vertical bar chart
    - Can easily become overwhelming
    - Hard to compare subcomponents across categories





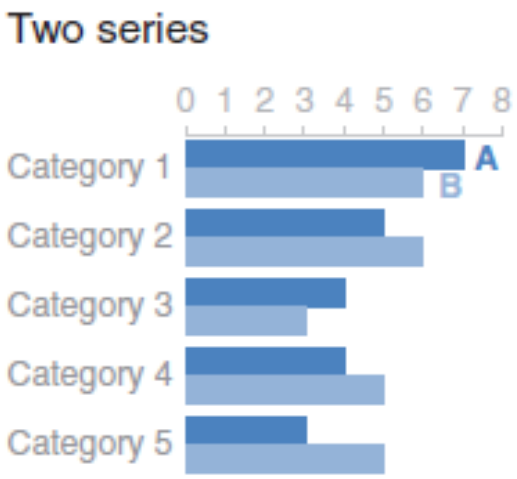
2. Choose an appropriate visual display

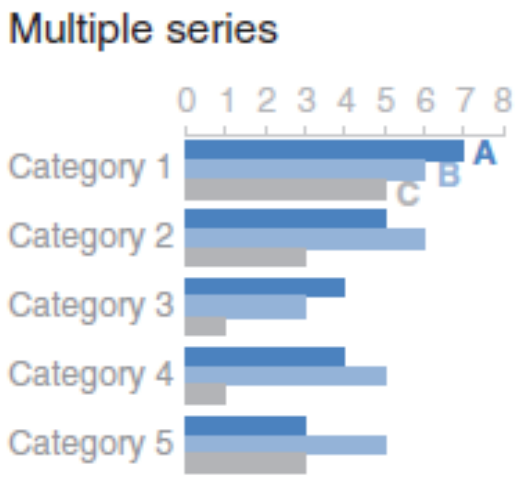
- Waterfall chart
  - Pull apart pieces (with both positive and negative
  - A decomposition of sales in the marketing mix models we saw earlier



2. Choose an appropriate visual display

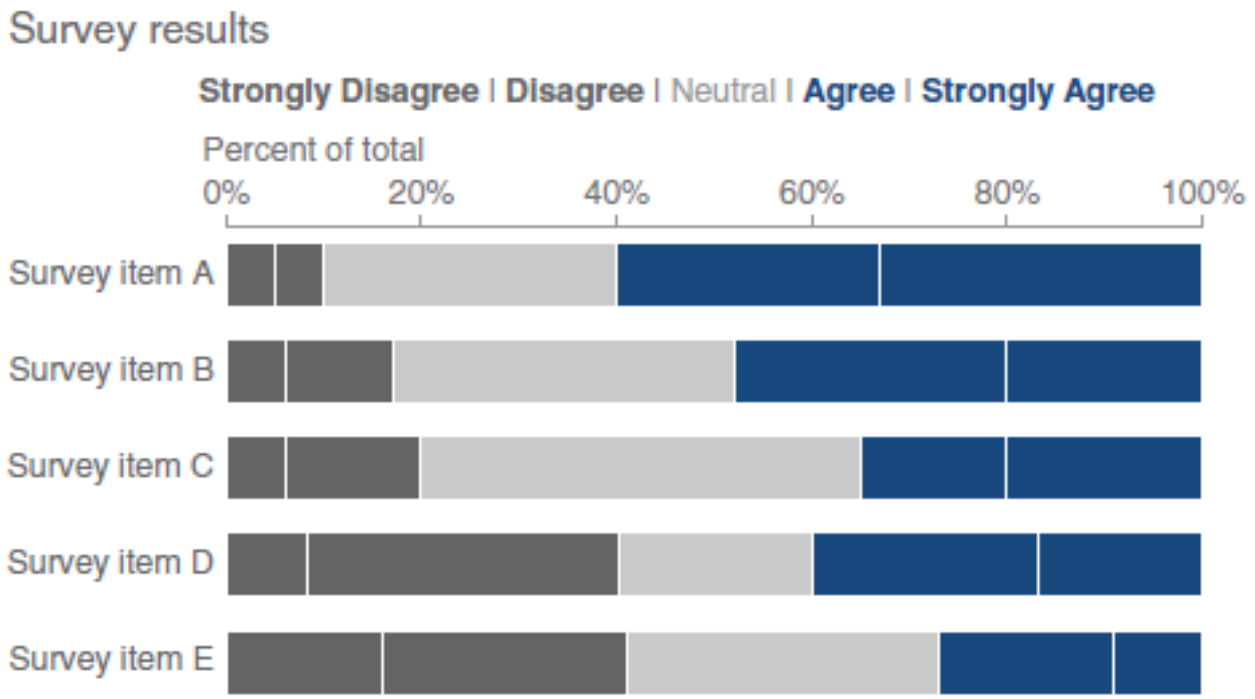
- Horizontal bar chart
  - Extremely easy to read
  - Useful with long category names
  - Our eyes hit the category names before the data
  - Ordering of categories – use natural order if there is one, otherwise think strategically and maybe order be increasing or decreasing size





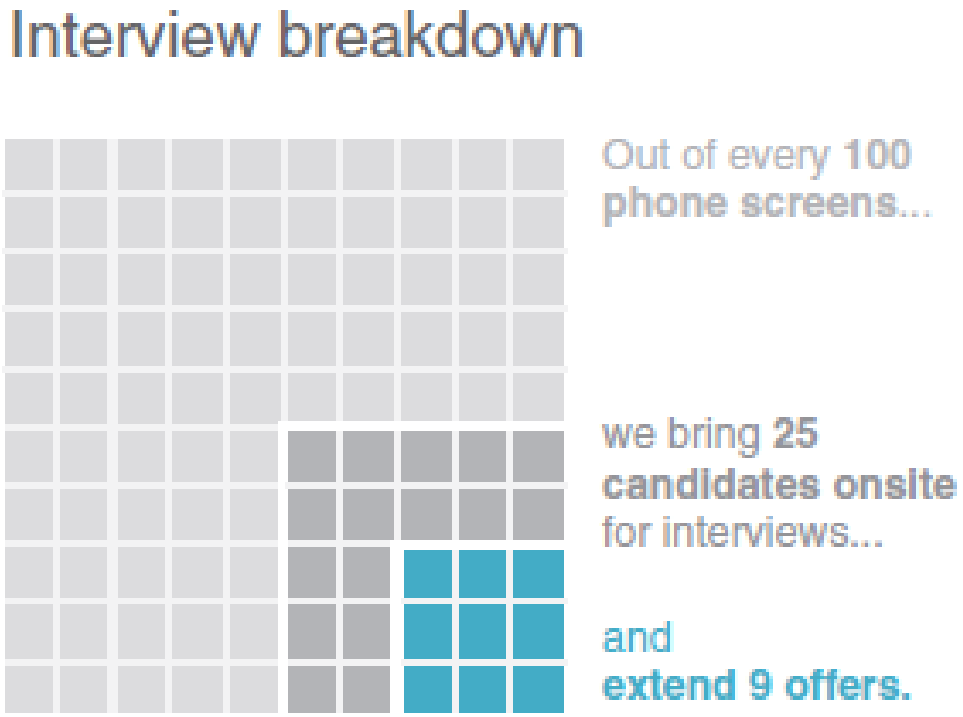
2. Choose an appropriate visual display

- Stacked horizontal bar chart
  - Like stacked vertical bar chart
  - Easy comparison of the left-most pieces and the right-most pieces



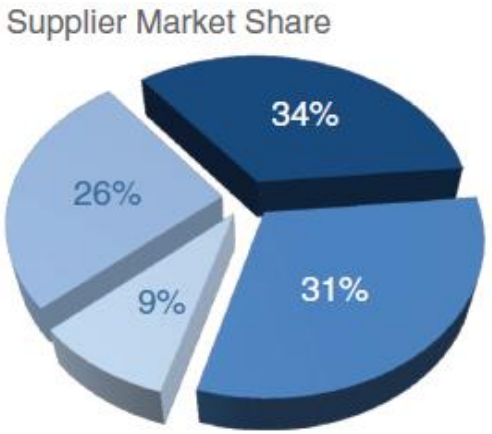
2. Choose an appropriate visual display

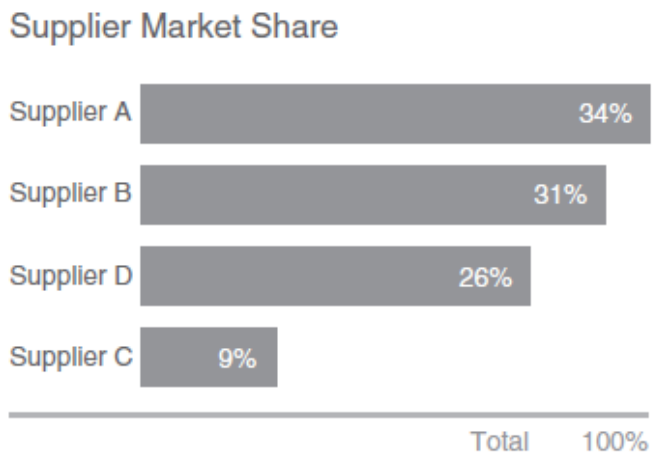
- Area graphs
  - Harder to read
  - Good when the difference in numbers are beyond linear

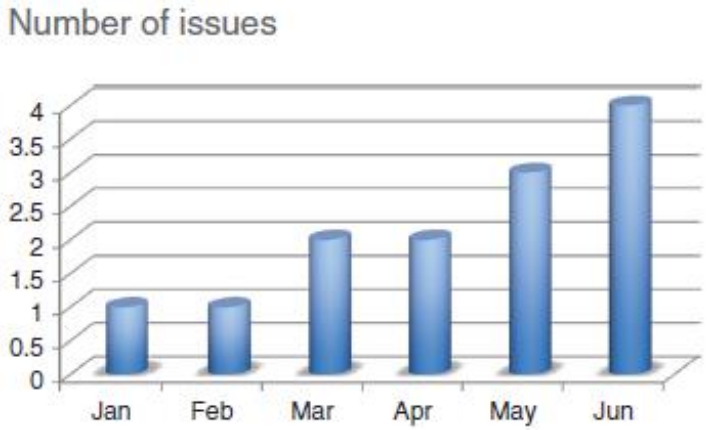


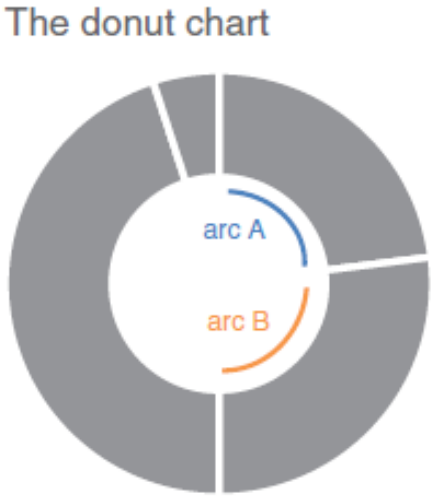
2. Choose an appropriate visual display

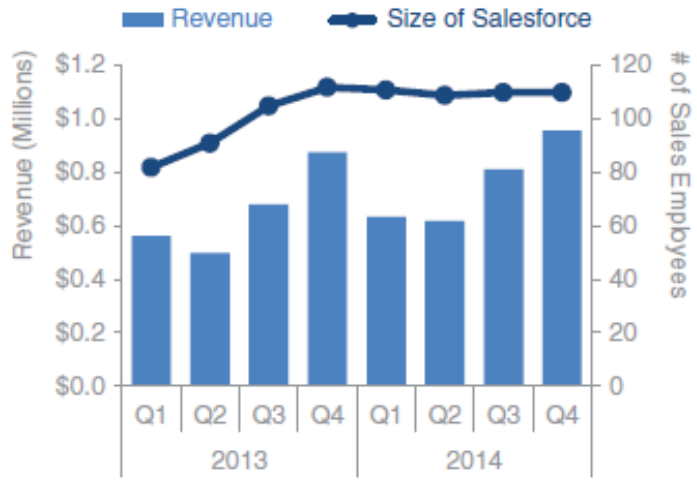
- To be avoided
  - Pie and donut charts
    - Hard to read and compare
  - 3D charts
    - Hard to read exact values
  - Secondary y-axis
    - Heavy load on reading











Storytelling with data in R

Storytelling with data in R

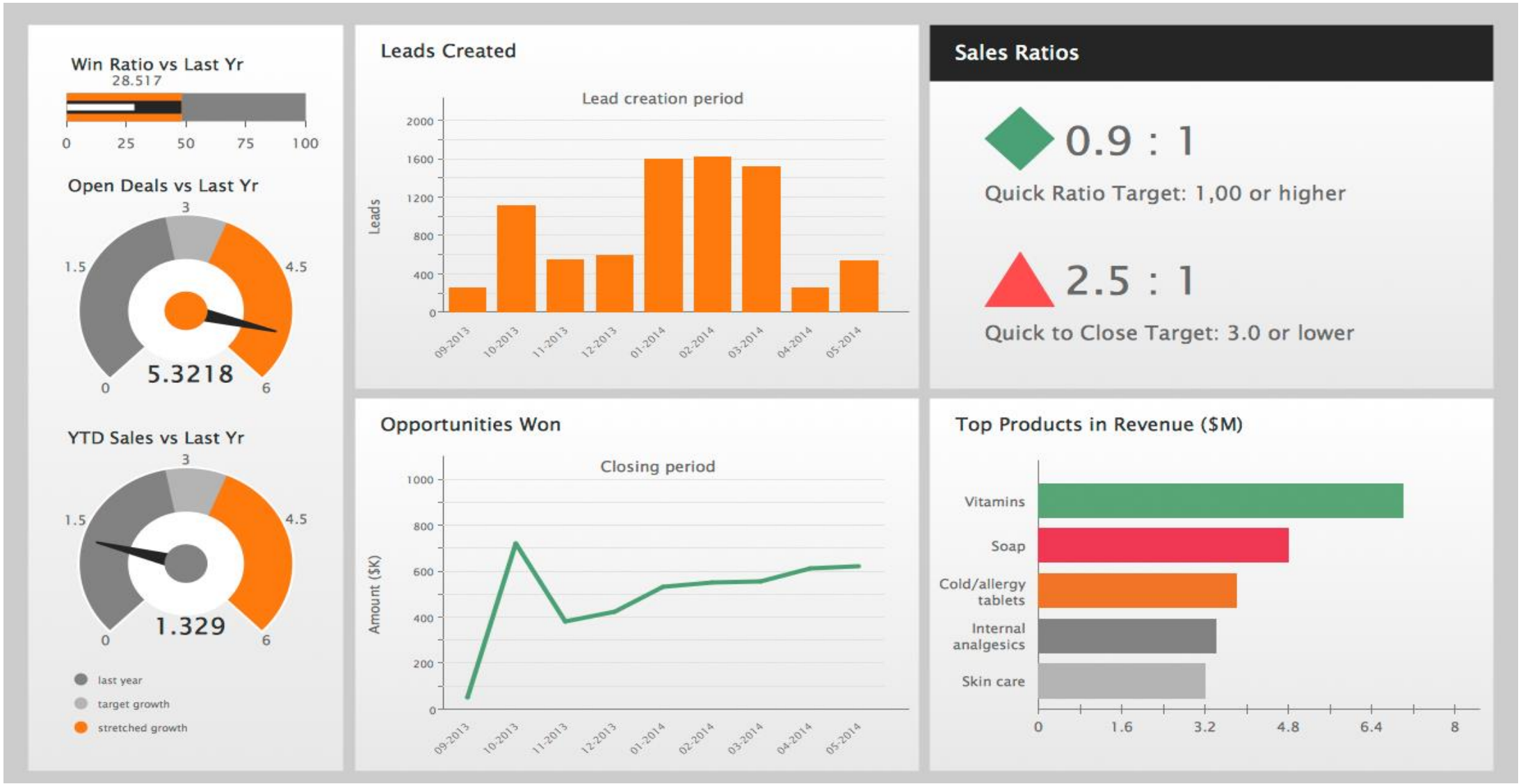
- See the Jupyter notebook “7.1 Storytelling with data in R.ipynb”

Reporting and Dashboard

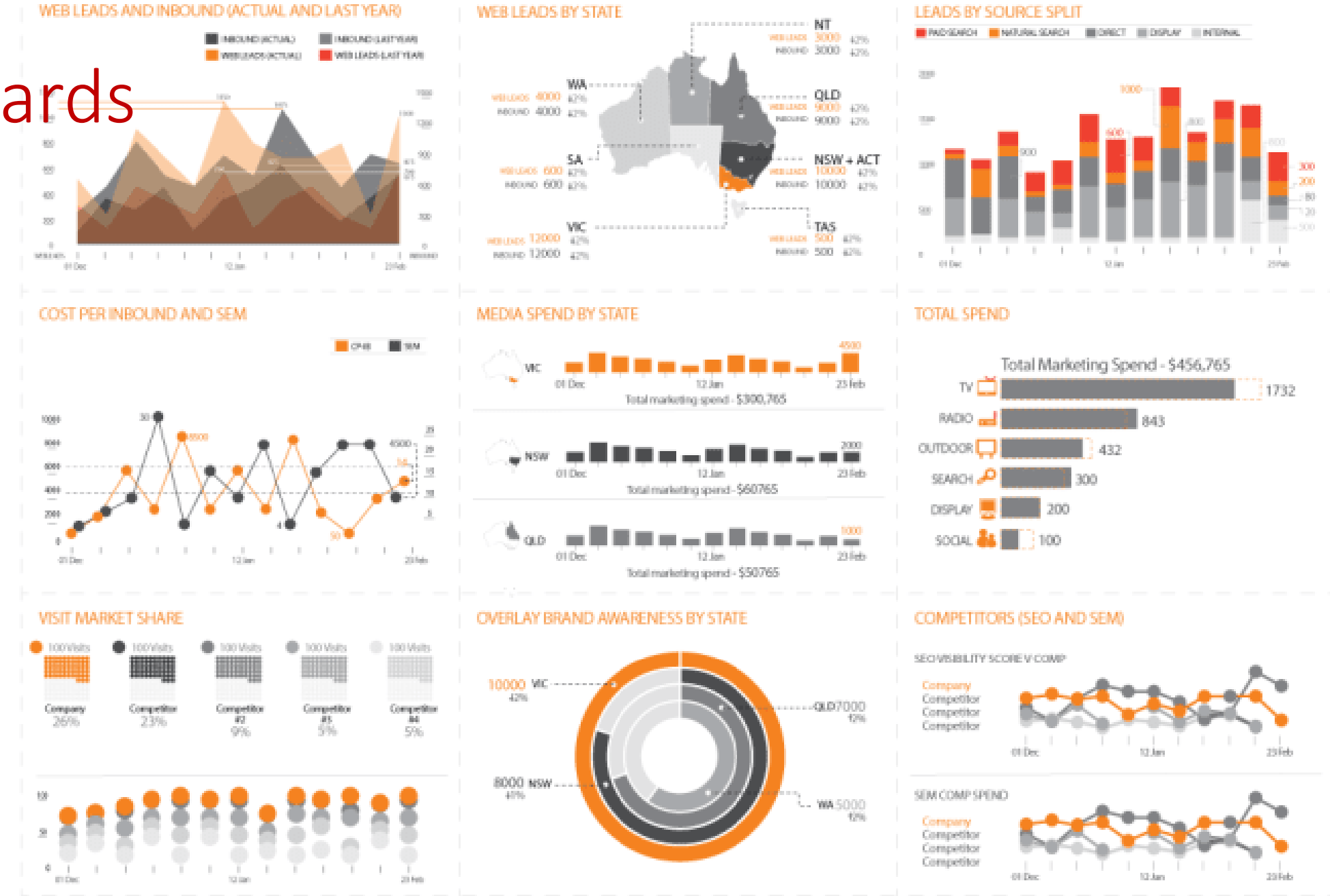
Reporting in general

- There are different ways of reporting Business Intelligence or Data Science findings
  - One time written report/slides
  - One time presentation
  - Dashboards
  - Interactive reports/dashboards
  - Excel sheets
  - Self-service BI (the user finds the data and the graphs themselves)
- The lessons of storytelling with data and visualization applies
- Reporting is were Data Science, IT, and Business meet
  - Whoever is involved in the creating the report needs to keep the final end-user/audience in mind
  - Different skilled people might be involved at different levels
  - There are several technical solutions depending on the type of report and the skill level of the involved stakeholders

Dashboards


Win Ratio vs Last Yr
28.517
0
25
50
75
100
Open Deals vs Last Yr
3
1.5
4.5
0
5.3218
6
YTD Sales vs Last Yr
3
1.5
4.5
0
1.329
6
last year
target growth
stretched growth
Leads Created
Lead creation period
Leads
2000
1600
1200
800
400
0
09-2013
10-2013
11-2013
12-2013
01-2014
02-2014
03-2014
04-2014
05-2014
Sales Ratios
0.9 : 1
Quick Ratio Target: 1,00 or higher
2.5 : 1
Quick to Close Target: 3.0 or lower
Opportunities Won
Closing period
Amount ($K)
1000
800
600
400
200
0
09-2013
10-2013
11-2013
12-2013
01-2014
02-2014
03-2014
04-2014
05-2014
Top Products in Revenue ($M)
Vitamins
Soap
Cold/allergy tablets
Internal analgesics
Skin care
0
1.6
3.2
4.8
6.4
8

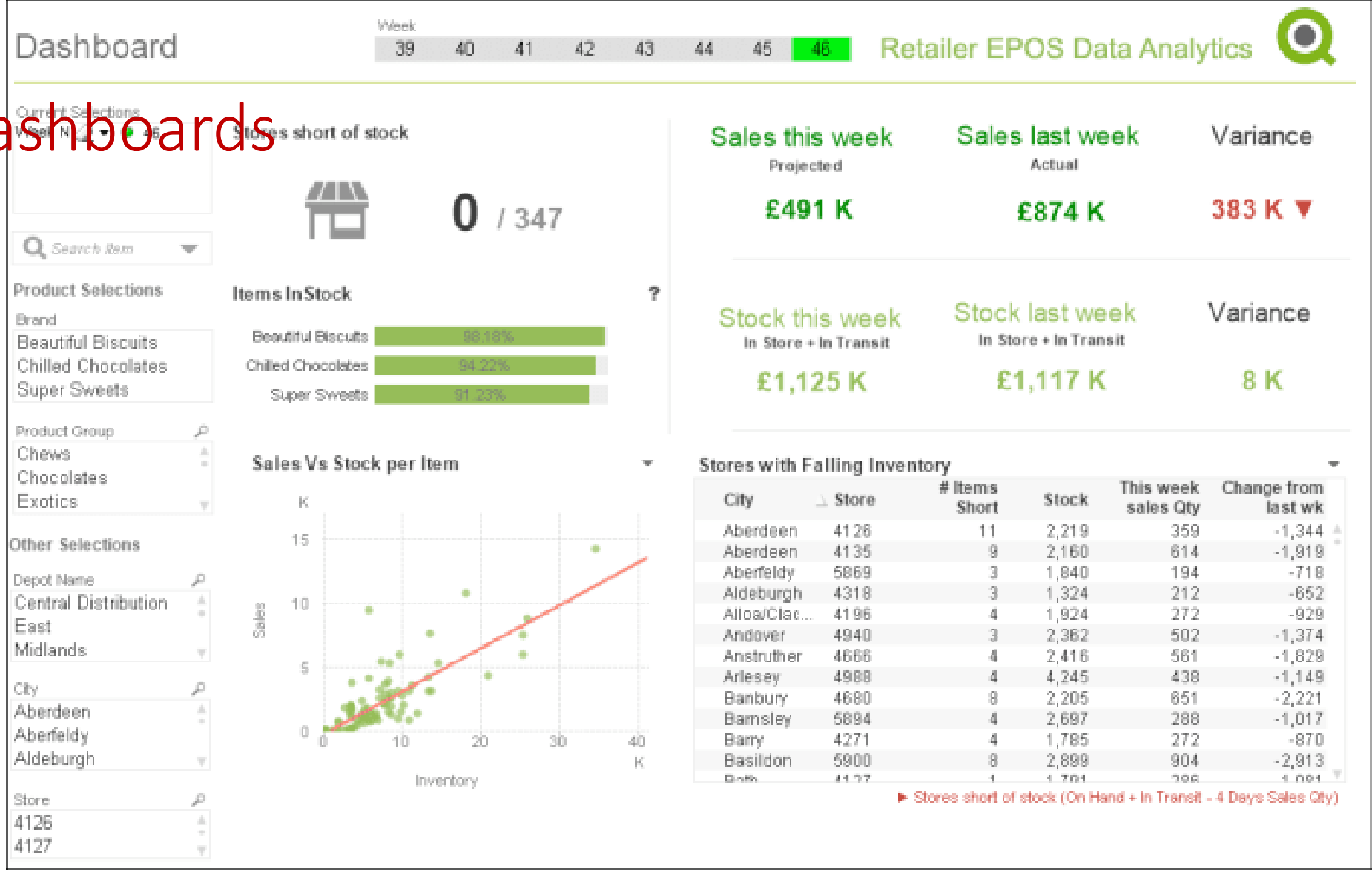
Dashboards


WEB LEADS AND INBOUND (ACTUAL AND LAST YEAR)
WEB LEADS BY STATE
NT
WA
QLD
SA
NSW + ACT
VIC
TAS
LEADS BY SOURCE SPLIT
COST PER INBOUND AND SEM
MEDIA SPEND BY STATE
VIC
NSW
QLD
Total marketing spend - $300,765
Total marketing spend - $60765
Total marketing spend - $50765
TOTAL SPEND
Total Marketing Spend - $456,765
TV 1732
RADIO 843
OUTDOOR 432
SEARCH 300
DISPLAY 200
SOCIAL 100
VISIT MARKET SHARE
Company 26%
Competitor 23%
Competitor #2 9%
Competitor #3 5%
Competitor #4 5%
OVERLAY BRAND AWARENESS BY STATE
10000 VIC 42%
8000 NSW 45%
QLD 7000 12%
WA 5000 12%
COMPETITORS (SEO AND SEM)
SEO VISIBILITY SCORE V COMP
SEM COMP SPEND
Company
Competitor
Competitor
Competitor

Dashboards


Power BI
Executive Metrics Dashboard
Share Dashboard
$85.22M
$2.22M
$4.07M
Total Invoice
Discount Savings
Marketing Site Traffic
Feedback Ratings
AccountCount
Customer Feedback Trend
Variance to Budget
Sales Pipeline
Customer Size vs Revenue
Total Invoice, Budget Amount
AccountCount
Number of Visits, Expenditures ($Mil)
Marketing Campaign Stats
Marketing Campaign Metrics
NORTH AMERICA
EURO

Dashboards


Dashboard
Week
39 40 41 42 43 44 45 46
Retailer EPOS Data Analytics
Current Selections
Stores short of stock
0 / 347
Search Item
Product Selections
Brand
Beautiful Biscuits
Chilled Chocolates
Super Sweets
Product Group
Chews
Chocolates
Exotics
Other Selections
Depot Name
Central Distribution
East
Midlands
City
Aberdeen
Aberfeldy
Aldeburgh
Store
4126
4127
Items In Stock
Beautiful Biscuits 98.18%
Chilled Chocolates 94.22%
Super Sweets 91.23%
Sales Vs Stock per Item
K
15
10
5
0
Sales
0
10
20
30
40
K
Inventory
Sales this week
Projected
£491 K
Sales last week
Actual
£874 K
Variance
383 K ▼
Stock this week
In Store + In Transit
£1,125 K
Stock last week
In Store + In Transit
£1,117 K
Variance
8 K
Stores with Falling Inventory
City
Store
Items Short
Stock
This week sales Qty
Change from last wk
Aberdeen 4126 11 2,219 359 -1,344
Aberdeen 4135 9 2,160 614 -1,919
Aberfeldy 5869 3 1,840 194 -718
Aldeburgh 4318 3 1,324 212 -652
Alloa/Clac... 4196 4 1,924 272 -929
Andover 4940 3 2,362 502 -1,374
Anstruther 4666 4 2,416 561 -1,829
Arlesey 4988 4 4,245 438 -1,149
Banbury 4680 8 2,205 651 -2,221
Barnsley 5894 4 2,697 288 -1,017
Barry 4271 4 1,785 272 -870
Basildon 5900 8 2,899 904 -2,913
► Stores short of stock (On Hand + In Transit - 4 Days Sales Qty)

Dashboards

- Analogy to the dashboards of a car, airplane or super tanker
  - Provide all the relevant information to navigate safely by the one in charge
  - A picture is worth more than a 1000 words: Simple graphs and statistics can save for numerous reports and meetings
- Purpose of dashboards
  - Provide managers with a fast and useful data overview
  - To track KPIs and measure performance in general
  - To inform the user with timely, accurate, important and actionable insights
- Different dashboards for different parts of the company
  - The top manager needs to track total revenue and other high-level KPIs
  - A software development team might need to track the number of users or bug reports in the part of the software they are responsible for
- Dashboards can provide information tailored to long-term strategic decisions as well as real-time information for making immediate decisions

Dashboards

- Characteristics of good dashboards:
  - Tailored to specific needs/business objectives
  - The shown measures/KPIs have been chosen carefully to match business objectives (– first business objectives then KPIs)
  - It is kept simple – don’t add more KPIs just because you can!
  - Information (and additional information) is easily accessible to the user
  - Use of a variety of charts, graphs, tables, and speedometers
  - It is interactive
  - It could include external content
  - Subscribe to simple user interface best practices
  - Remember what we talked about when discussing storytelling

Reporting in R

- Jupyter notebooks – you know these already
- R scripts – use scripts to generate graphs etc.
- RMarkdown
- Rnotebook
- Flexdashboards
- Shiny

RMarkdown

- RMarkdown allows you to do the communication part of Data Science/Business Intelligence in an easy way using R
- It provide a unifying way of writing reports in various formats, slides, dashboards etc. all using R code and Markdown language
- An RMarkdown file is a plain text file with the extension ".Rmd"
- Such a file contains three types of content:
  - ***An (optional) YAML header*** specifying the type of document
  - ***Chucks of R code*** where you can show code examples and output of data analysis and modeling as well as plots
  - ***Text in a simple Markdown format*** that explain your analysis of what have you
- Good examples are the slides from last time (and today)

RMarkdown

- To create an RMarkdown document in RStudio go through the menus “File → New file → RMarkdown...”
- Choose what kind of RMarkdown document you would like to create
- Note, if you want to produce pdf output, you need a TeX distribution installed. (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux)
- The output field in the YAML header specifies what type of document and output you get.
- To “compile”/knit your RMarkdown document, click the Knit button or press “Cmd/Ctrl + Shift + K”
- RPubs (https://rpubs.com/) allows for an easy way to publish your RMarkdown documents online (-It can be done by a click of a button from RStudio!)
- Let us look at some examples (slides from today’s lecture)

RNotebook

- A notebook format not unlike Jupyter notebooks
- Used in through RStudio

Flexdashboards

- Flexdashboards is an easy way to make dashboards in RMarkdown
- To use it, install the “flexdashboard” R package
- To create a flexdashboard, when creating a new RMarkdown document, chose “From Template”, “Flex Dashboard”.
- Dashboards can be made interactively using “Shiny”
- Examples:
  - https://beta.rstudioconnect.com/jjallaire/htmlwidgets-ggplotly-geoms/
  - https://jjallaire.shinyapps.io/shiny-crandash/
  - See documentation and more examples at: http://rmarkdown.rstudio.com/flexdashboard/index.html

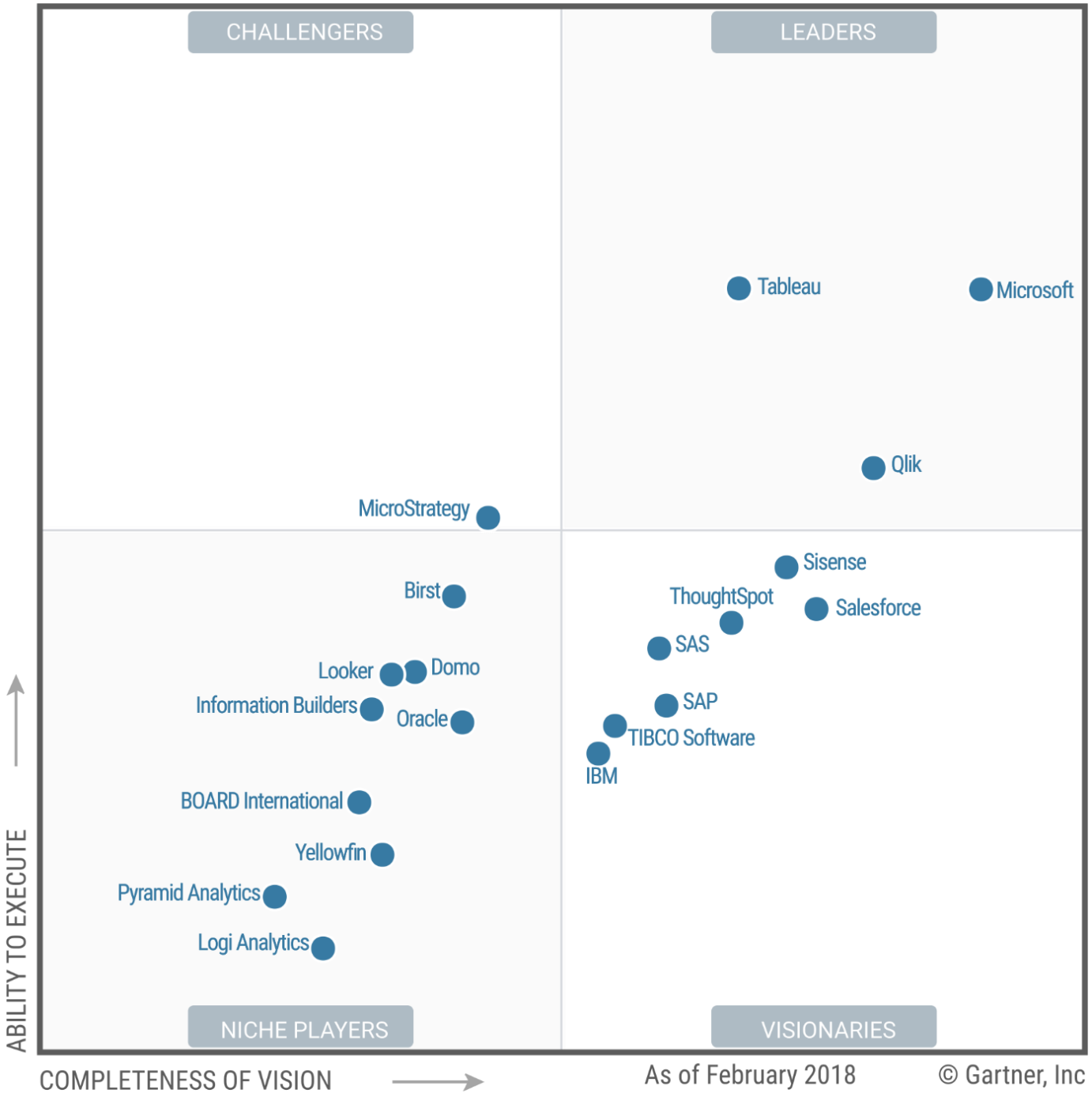
Shiny

- A web application framework for R that can turn your analyses into interactive web applications
- No knowledge of HTML, CSS, or JavaScript is required
- A Shiny App consists of:
- A ui.R script that specifies input and output
- A server.R script that runs R code to create the output (potentially based on the input)
- Examples
  - http://shiny.rstudio.com/gallery/kmeans-example.html
  - http://shiny.rstudio.com/gallery/word-cloud.html
- A Shiny server is needed. You can install your own, or you can also publish your apps publicly on Rstudio's Shiny server at: http://www.shinyapps.io/
- Shiny-like code can be used in flexdashboards to make them interactively too
- See documentation at: http://shiny.rstudio.com/
- See Chapter 8 of the book "Introduction to R for Business Intelligence" by Jay Gendron

Other BI tools

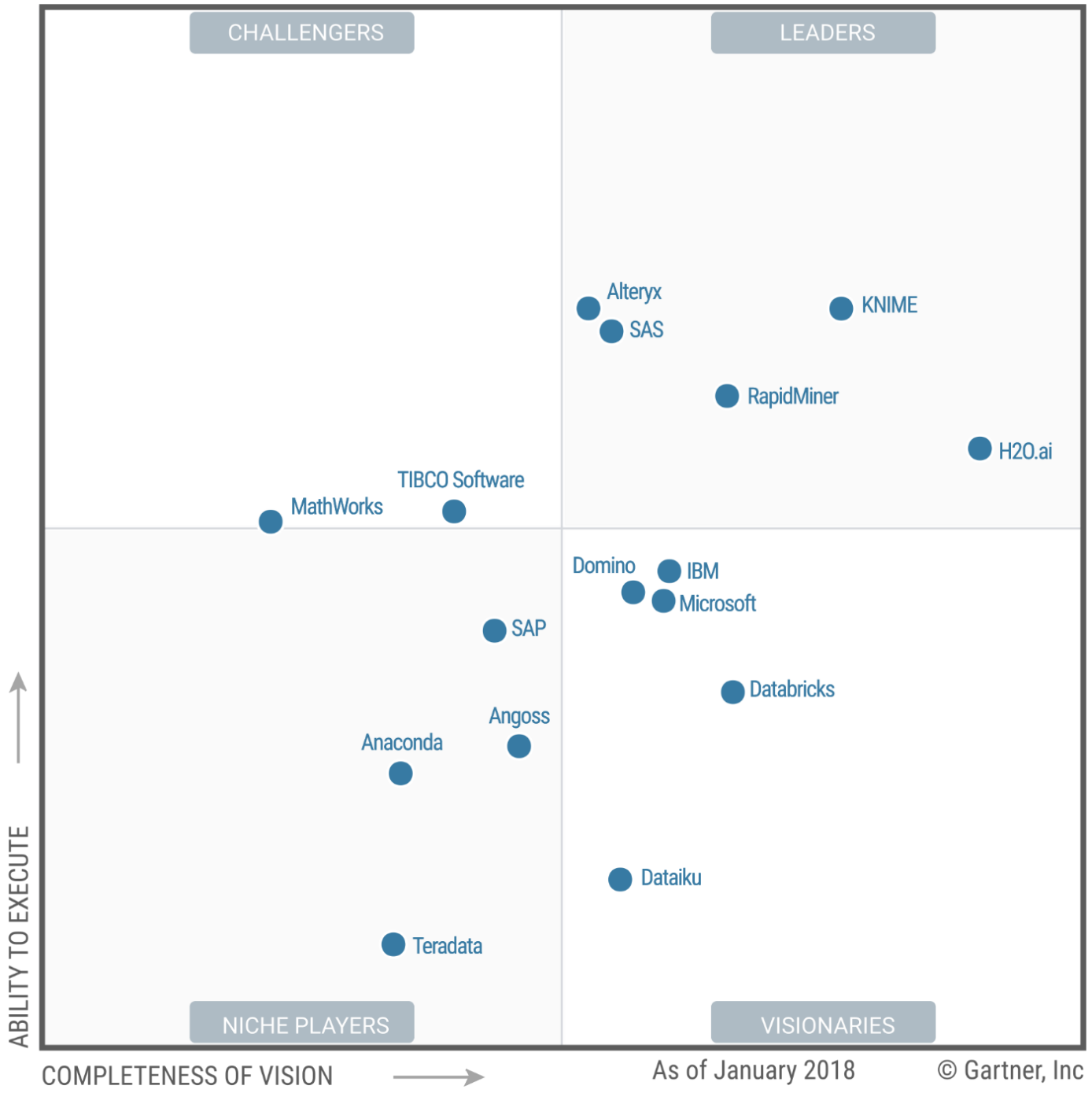
Other BI tools

- **Business Intelligence platforms**
- *Gartner's Magic Quadrant for Analytics and Business Intelligence Platforms*
  - https://www.gartner.com/doc/reprints?id=1-3TXXSLV&ct=170221&st=sb



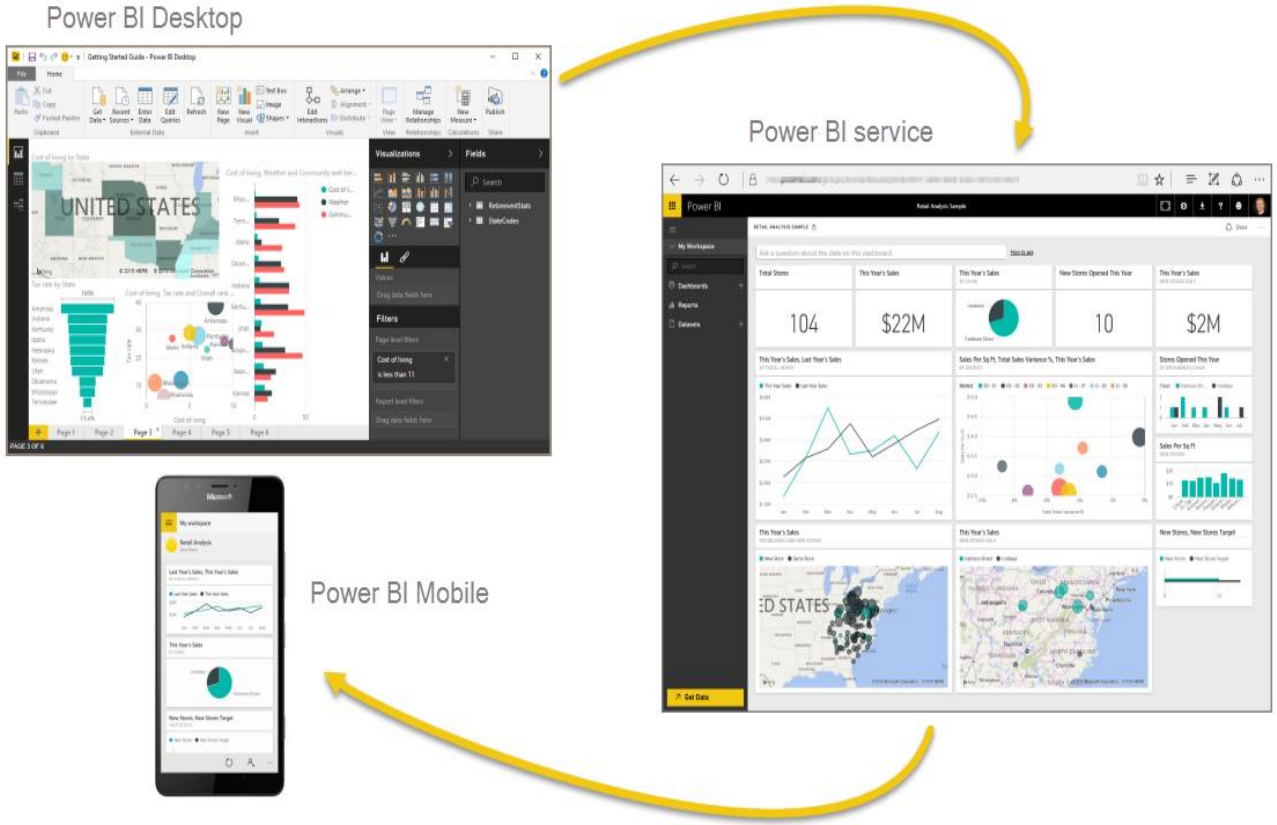
Other BI tools

- **Data Science and Machine-Learning Platforms**
- *Magic Quadrant for Data Science and Machine-Learning Platforms*
  - https://www.gartner.com/doc/reprints?id=1-4RMUF0K&ct=180222&st=sb



Microsoft Power BI

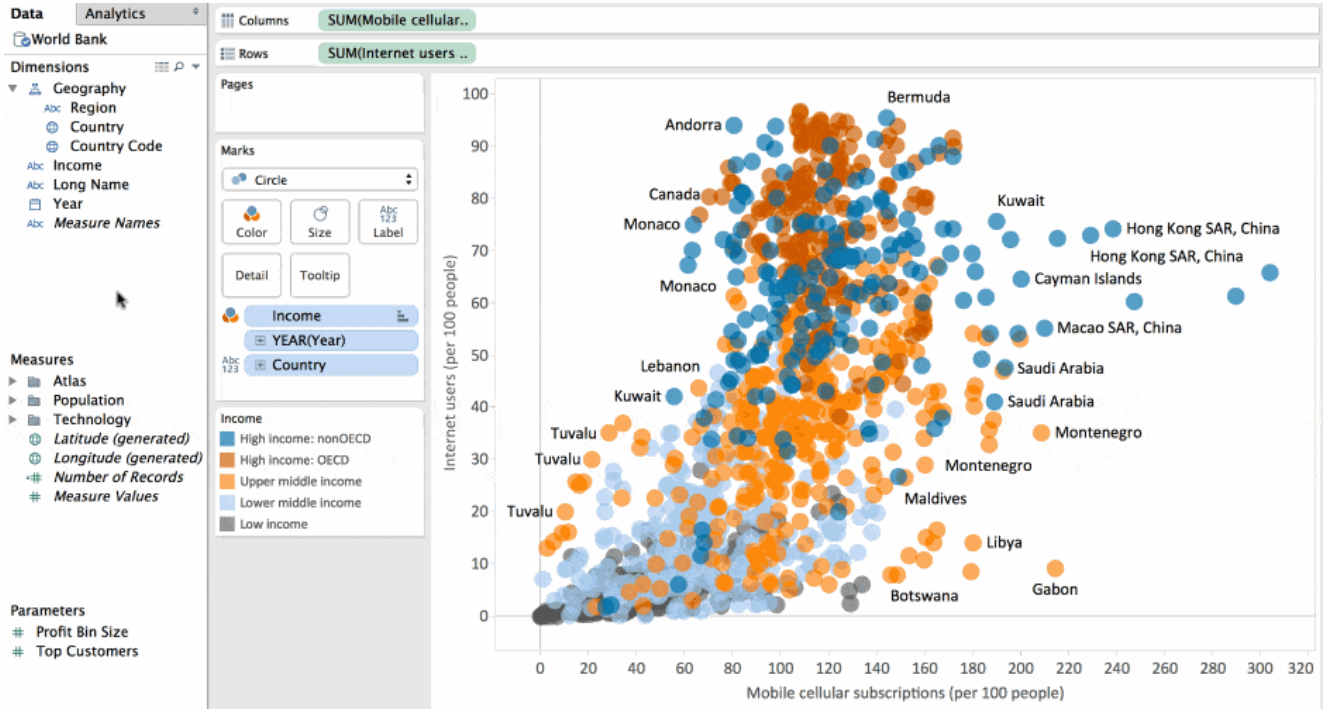
- A collection of software services to connect data for various sources and visualize it in graphs, dashboards and reports that are shareable
- Website: https://powerbi.microsoft.com
- Three components:
  - Power BI Desktop, Service and Mobile
- Power BI Desktop also comes in a free version



Microsoft Power BI

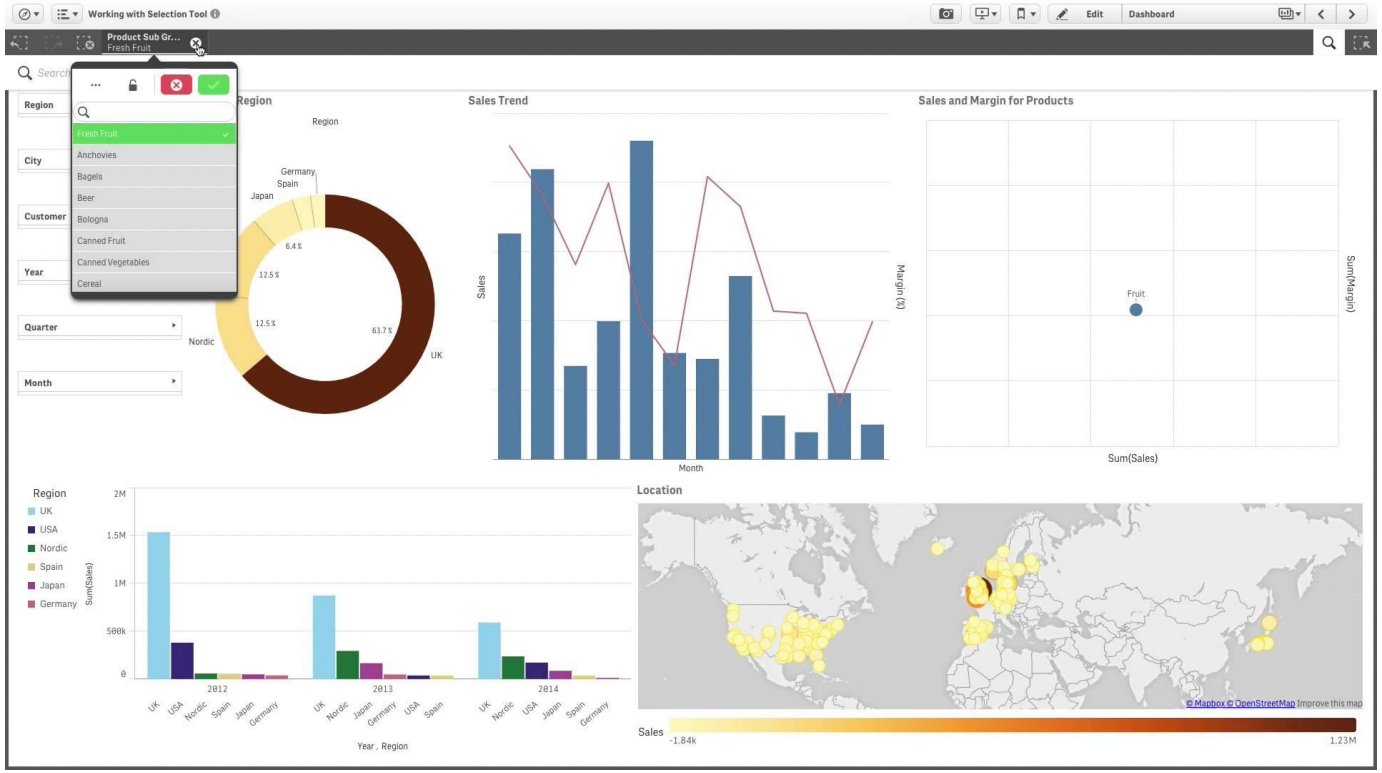
- Let’s have a look...
  - https://www.youtube.com/watch?v=yKTSLffVGbk&feature=youtu.be

Tableau



- Mission statement:
  - "We help people see and understand their data"
- Popular tool for visualizing data and make dashboards
- One of the most widely used BI tool
- Comes in a desktop, server and cloud version
- Free 14 days trial or 1 year for students available from: https://www.tableau.com/
- Videos
  - Tableau intro: https://www.youtube.com/watch?v=Qo6W-oBO9XM
  - 2019 BI trends: https://www.youtube.com/watch?v=vyWbDCn_ncA&list=PL_qx68DwhYA8g_GffE7fOQ0VPg9A3oSfZ
  - Data storytelling is the new language of corporations: https://www.youtube.com/watch?v=3J8T-QNygts&index=8&list=PL_qx68DwhYA8g_GffE7fOQ0VPg9A3oSfZ

Qlik



- Website: https://www.qlik.com
- Qlik Sense Product Tour: https://www.youtube.com/watch?v=85QHuNNeaCg
- Examples
  - Quooker controls their entire supply chain with Qlik: https://www.youtube.com/watch?v=fal-9dl3Z5Q&index=17&list=PLW1uf5CQ_gSpRDqce6rmMzvik1nxw6NWC
  - Ottawa Paramedics: https://www.youtube.com/watch?v=KA6vQVjA1ak&index=10&list=PLW1uf5CQ_gSpRDqce6rmMzvik1nxw6NWC