DATA VISUALIZATION

Architecture of a BI system

Data selection, extraction, transformation and load

Data visualization and analytics: descriptive, predictive prescriptive

Data storage, in relational or multi-dimensional tables

OLAP: on-line analytical processing
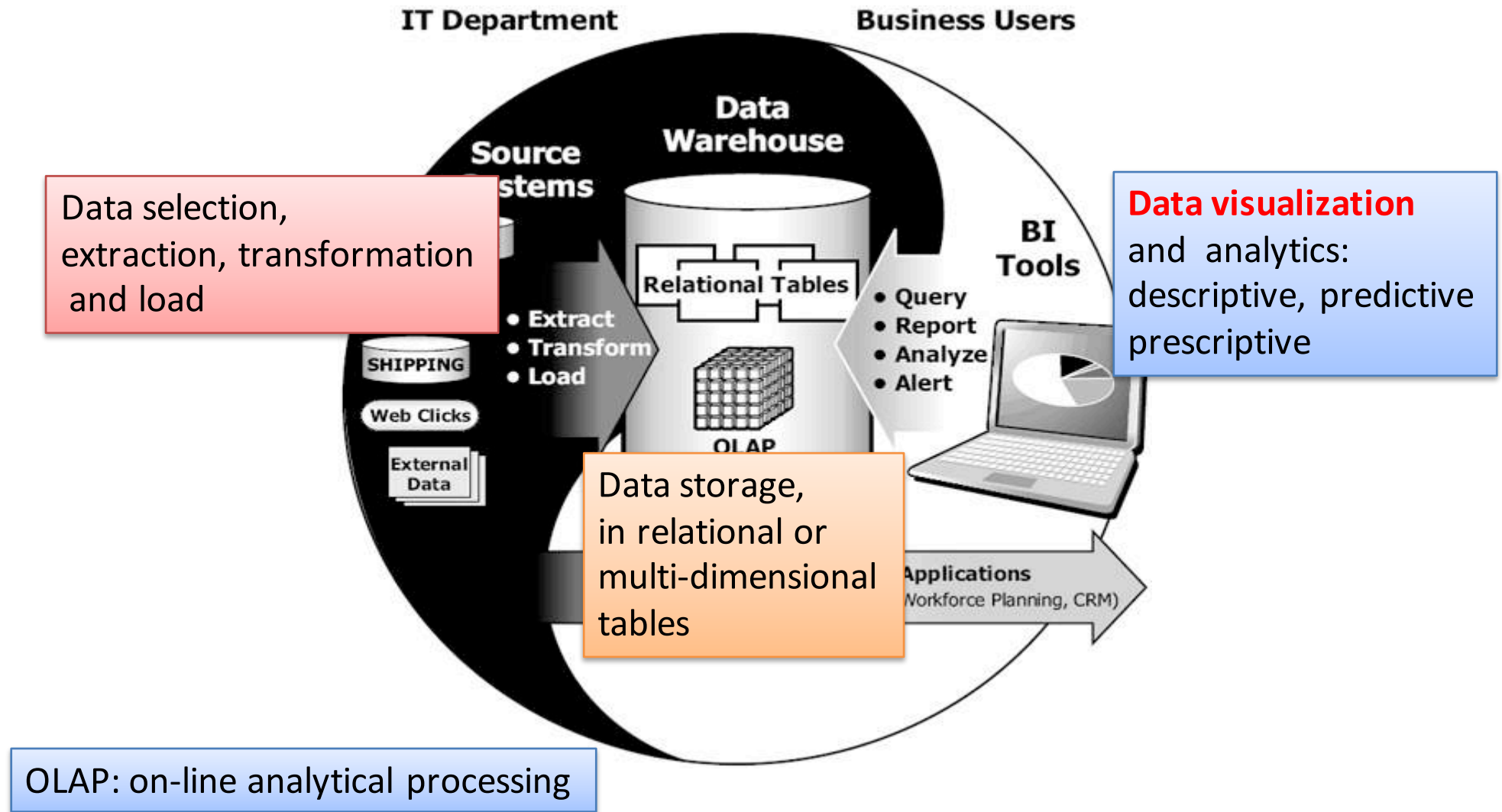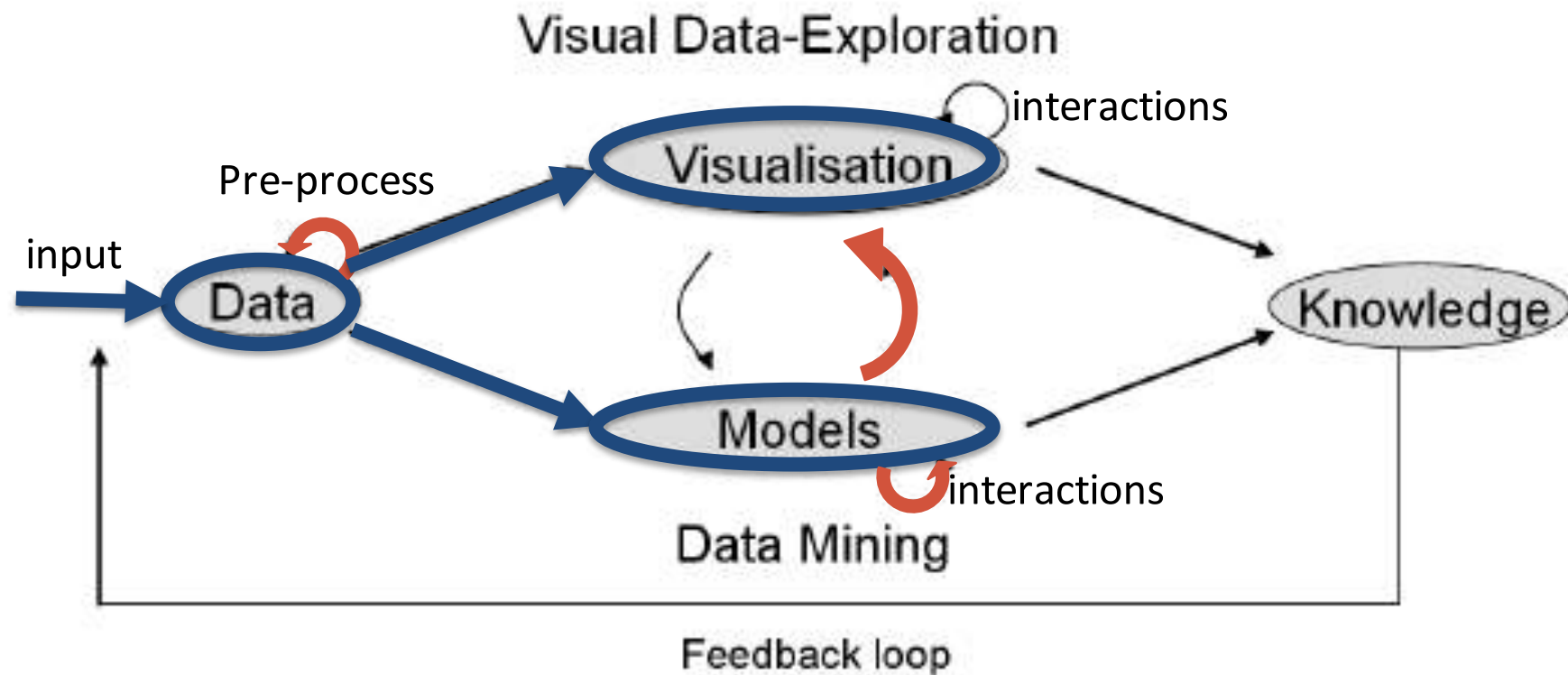
Outline

- **What is Visualization?**
- **Why Visualize?**
- **Data → Visualization**
- **How to tell if a visualization is "appropriate"**
- **Guidelines**

Keim's Visual Analytics Model

Image source: Keim, Daniel, et al. *Visual analytics: Definition, process, and challenges*. Springer Berlin Heidelberg, 2008.

Visualization: what is?

Visual **representations** of data that reinforce human **cognition**

Perhaps a more helpful question:

What are some ways
a "visualization" can be **useful**?

But first off: Value of Visualization

- Provably **much better than written reports**, since:

- Reduce Memory Load
  - Working memory is limited
  - Offload storage/organization to the diagram

- Reduce Search Time
  - Pre-attentive (constant-time) search
  - Spatially-indexed patterns store the "facts"

- Enable Perceptual Inference
  - Map inference to pattern finding

Some very "old" examples: cholera outbreak in 1854

- In 1854, cholera broke out in London
  - 127 peo
    days
  - 616 peo
- Initial exp
        "Mia
- Dr. John Sr
  outbreak
- How did h
  - Talked t
  - Identifie
  - Used ma
  - Convinc

More info here: http://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

Another old example: 1869

"Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812-1813."
Drawn by Mr. Charles Minard, Inspector General of Bridges and Roads in retirement.
Paris, 20 November 1869.

Minard's multi-dimensional map

- Minard was a pioneer of the use of graphics in engineering and statistics.
- He is most well known for his cartographic depiction of numerical data on a map of Napoleon's disastrous losses suffered during the Russian campaign of 1812
- The illustration depicts Napoleon's army departing the Polish-Russian border. A thick band illustrates the size of his army at specific geographic points during their advance and retreat.
- It displays **six types of data in two dimensions**:
  - the number of Napoleon's troops;
  - the distance traveled;
  - temperature;
  - latitude and longitude;
  - direction of travel;
  - and location relative to specific dates without making mention of Napoleon;

So when vizualizations are useful?

Example: Traditional excel table vrs "fancy" visualization

Tasks	Start	Duration (Days)	% Complete	02 - Jan - 12	09 - Jan - 12	16 - Jan - 12	23 - Jan - 12	30 - Jan - 12	06 - Feb - 12	13 - Feb - 12	20 - Feb - 12	27 - Feb - 12	05 - Mar - 12
Task 1	1/2/12	35	26%										
Sub Task 1.1	1/2/12	10	40%										
Sub Task 1.2	1/12/12	25	20%										
Task 2	1/2/12	66	88%										
Sub Task 2.1	1/2/12	30	100%										
Sub Task 2.2	2/1/12	25	80%										
Sub Task 2.3	2/27/12	10	70%										
Task 3	1/2/12	10	40%										
Sub Task 3.1	1/2/12	10	40%										

Useful=help us understanding the data

- What types of data?

- What kind of "understanding" we want to convey?

- How to connect data with (good) visualizations?

Types of data

- Data can be classified in three groups

- Qualitative (Attributes)
    1. Nominal
    2. Ordinal

- Quantitative (Metrics)
    3. Numeric

Qualitative:Nominal data

- Data that be counted, but **not** ordered or aggregated.
- Examples:
  - Products – Books, Movies, Music
  - Gender – Male, Female
  - State – Virginia, Nevada, California

Qualitative: Ordinal data

- Data that can be counted and **ordered**, but not aggregated.
- Examples:
  - Date – 1/1/2014, 1/2/2014…
  - Grades – A, B, C…
  - Ranks – Like, Neutral, Dislike

Metrics

- Quantitative data that can be **counted, ordered, and aggregated**.
- Examples:
  - Revenue, Cost, Profit
  - Number of Customers
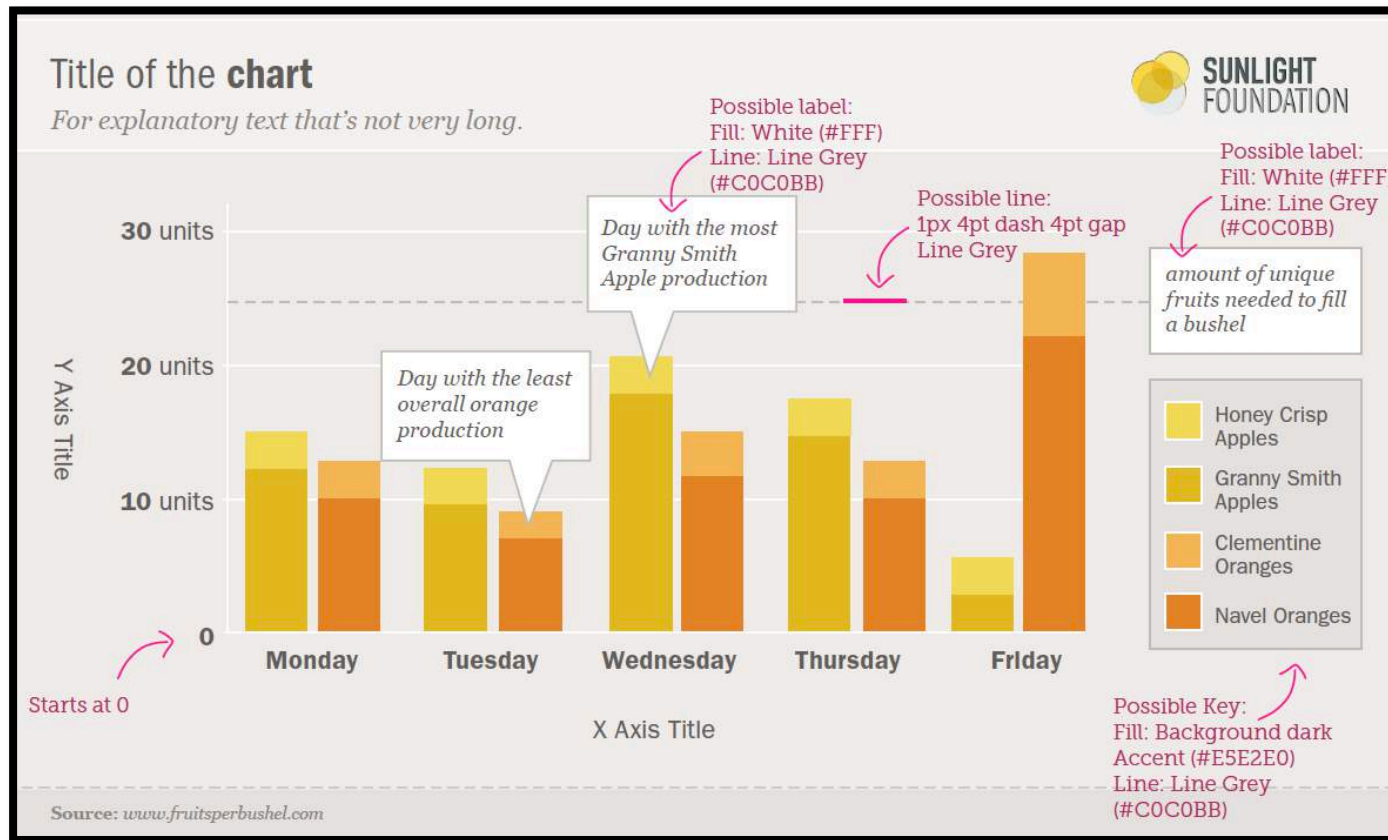 - Temperature
  - Time

Ordinal Attributes and Metrics

- Some data can be used as either ordinal or metrics. Their classification is dependent on usage.
- Examples:
  - Age
  - Scores

Types of visualizations

- Bar charts (histograms)
- Line charts
- Scatterplots
- Maps
- Pie Charts
- Network (graphs)
- ...many others (will see creative examples)

Bar Charts (histograms)

Line charts
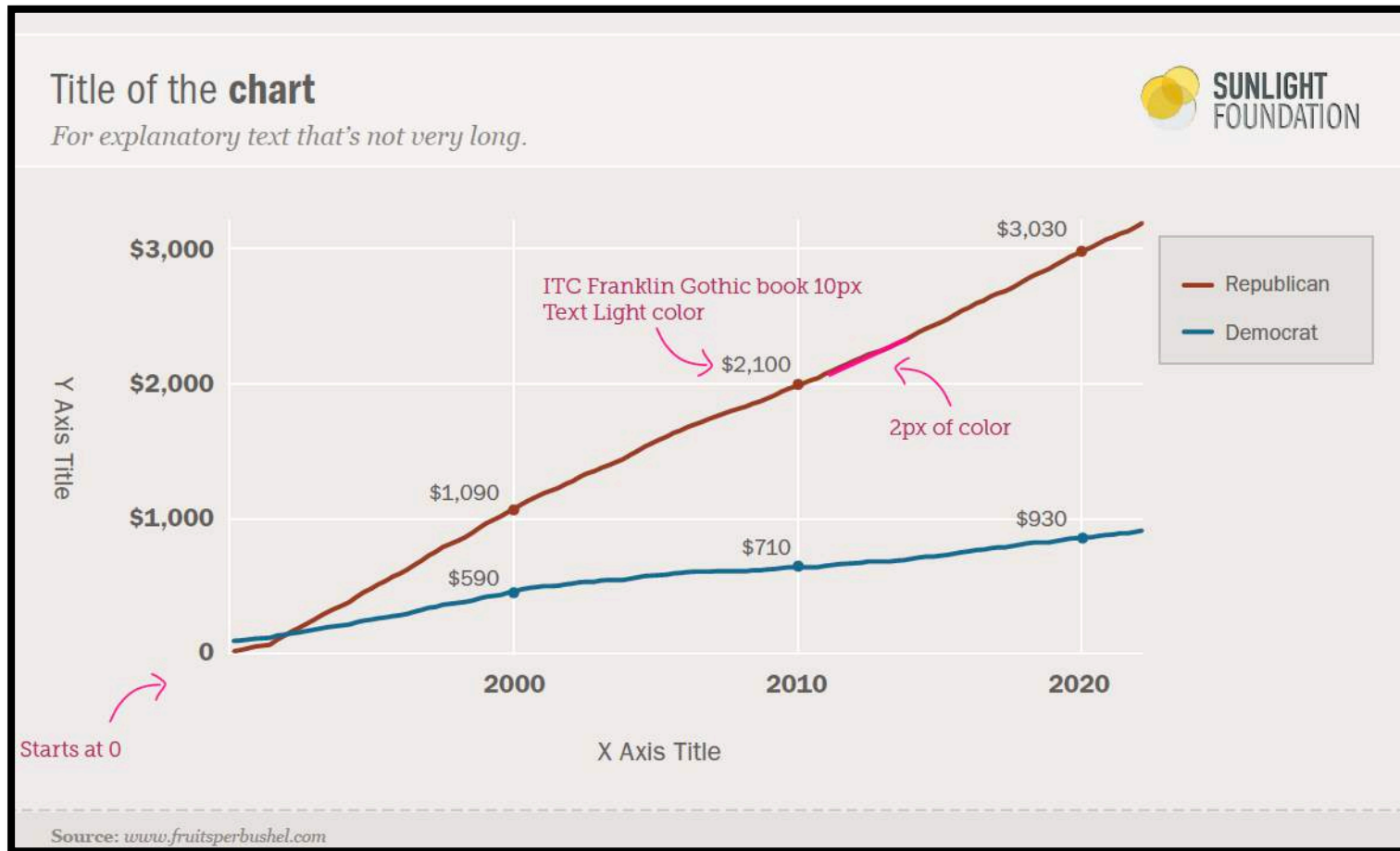
Scatterplots

- Shows the relationship between two **continuous variables**
- Each point in the plot represents an observation
- You can change color or symbol to **highlight groups**
- Sometimes it is nice to show a trend line (regression)
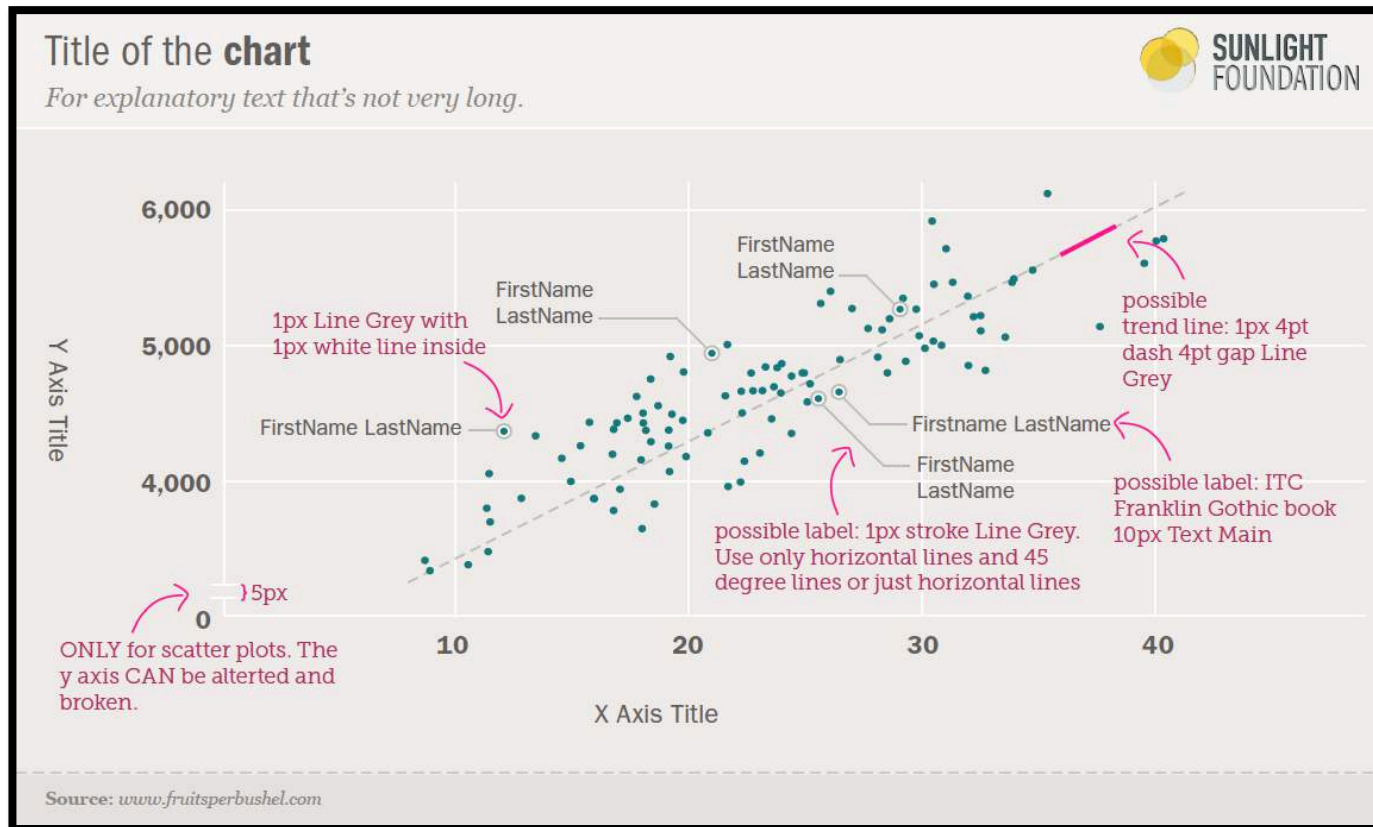
Maps

- Useful for analysis with a strong geographic component
- Remember: color scale comparisons are harder for humans than size comparisons. Keep this in mind as you choose between a map or another layout.
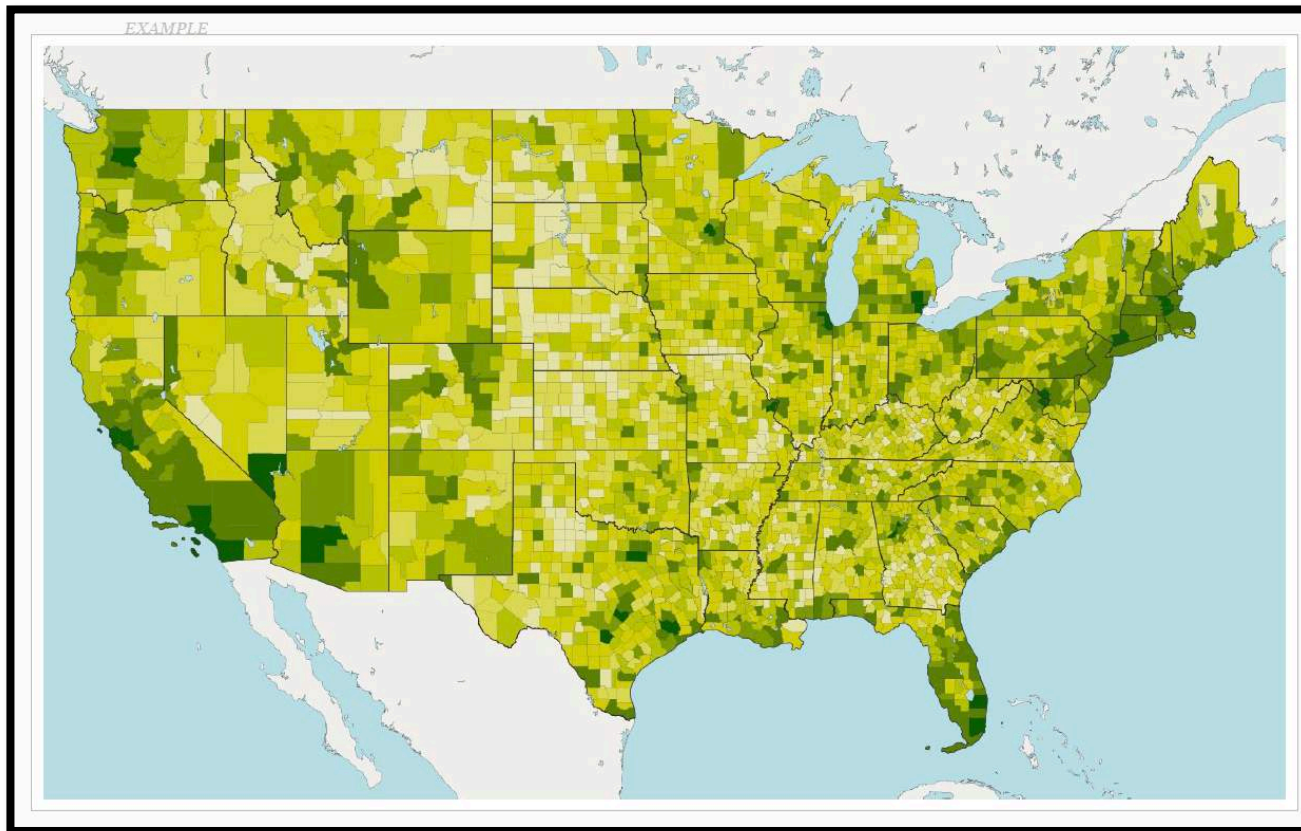
Pie Charts

- Almost never the right choice (angular comparison is hard)
- Use only if the following 2 conditions are met:
    - You want to show the relative relationship between 2-3 things
    - They add up to 100%
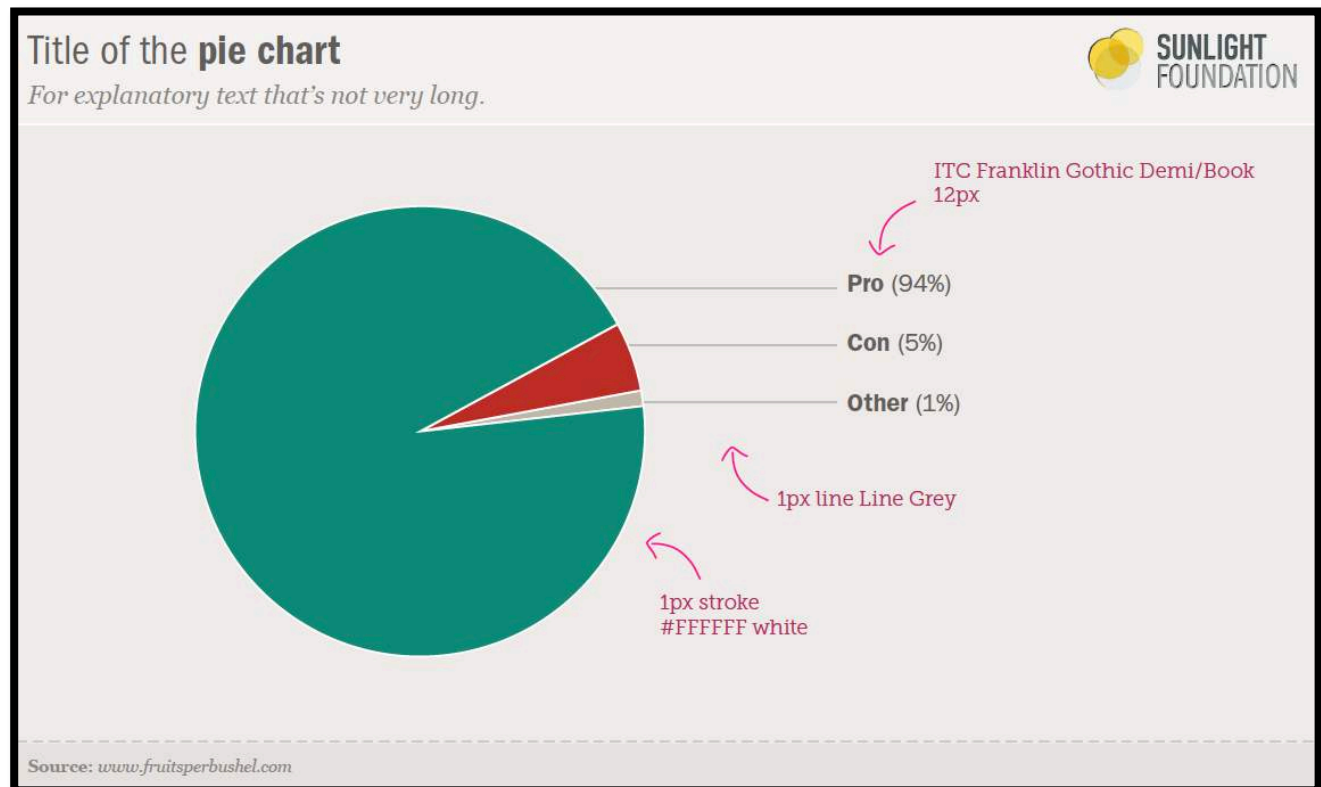
Networks (graphs)

- Useful for showing the relationships between entities
- Can use color, size, etc. to encode additional information about nodes/edges
- Caveat: network diagrams quickly become hairballs for large, dense data.
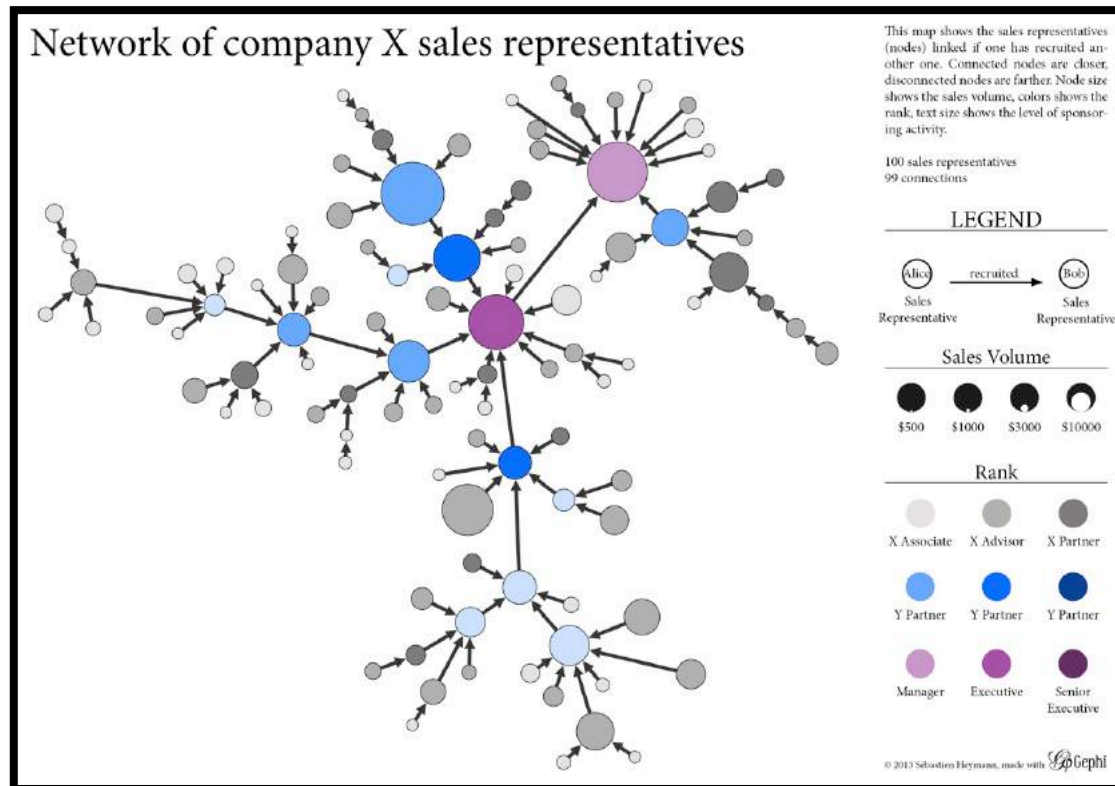
Connecting Data To Visualization

- Data have types
- Visualizations have types
- Reports have communication objectives

- **How do we map one onto the other?**
  1. **Depending on the information we want to convey**
  2. **Depending on the type of data we want to visualize**

SELECTING VISUALIZATIONS

Depending on the information we want to convey

Do you want to compare values?

- Charts are perfect for comparing one or many value sets, and they can easily show the low and high values in the data sets.
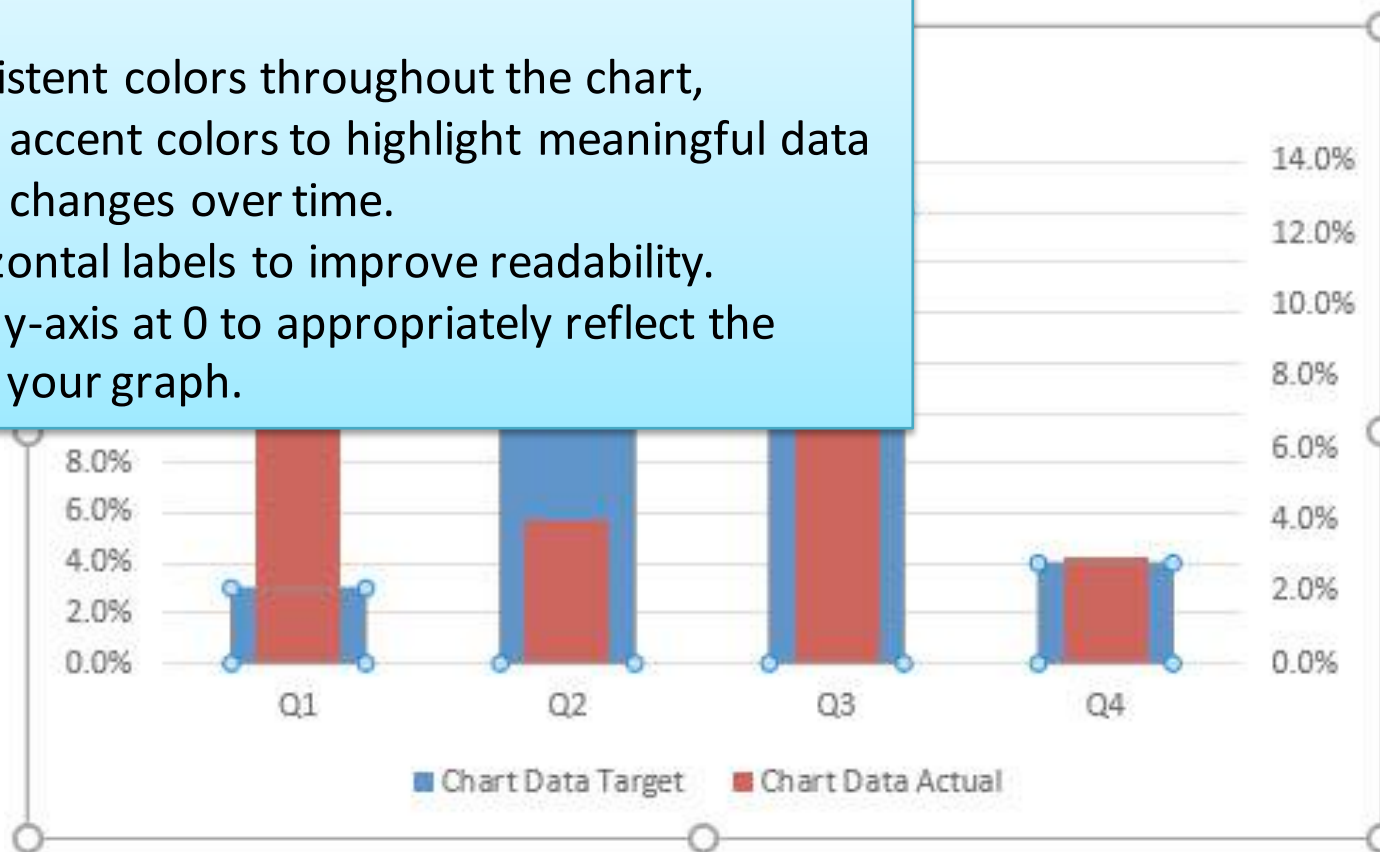- To create a **comparison chart**, use these types of graphs:
  - Column
  - Bar
  - Circular Area
  - Line
  - Scatter Plot
  - Bullet

A **column chart** is used to show a **comparison** among different items, or it can show a comparison of items over time.

Design Best Practices for Column Charts:

- Use consistent colors throughout the chart, selecting accent colors to highlight meaningful data points or changes over time.
- Use horizontal labels to improve readability.
- Start the y-axis at 0 to appropriately reflect the values in your graph.

14.0%

12.0%

10.0%

8.0%

8.0%   6.0%

6.0%   4.0%

4.0%

2.0%   2.0%

0.0%   0.0%

          Q1   Q2   Q3   Q4

          ■ Chart Data Target   ■ Chart Data Actual

A **bar chart**, basically a horizontal column chart, should be used to avoid clutter when one data label is long (e.g., "Individual contributors" or if you have more than 10 items to compare.

Excuses for being late to class

| Excuse	
I had no clean pants to wear	
I thought it was Saturday	
I forgot to set my alarm	
It was still too dark, I thought it was still nighttime	
I got stuck in traffic	

A **line ch**art reveals trends or progress over time and can be used to show many different categories of data. You should use it when you chart a continuous data set.

Stock Monitoring, 2010

O Portfolio 3 O Portfolio 1 O Portfolio 2

A **bullet graph** reveals progress toward a goal, compares this to another measure, and provides context in the form of a rating or performance.


Sunkost: Sales Target of 2010
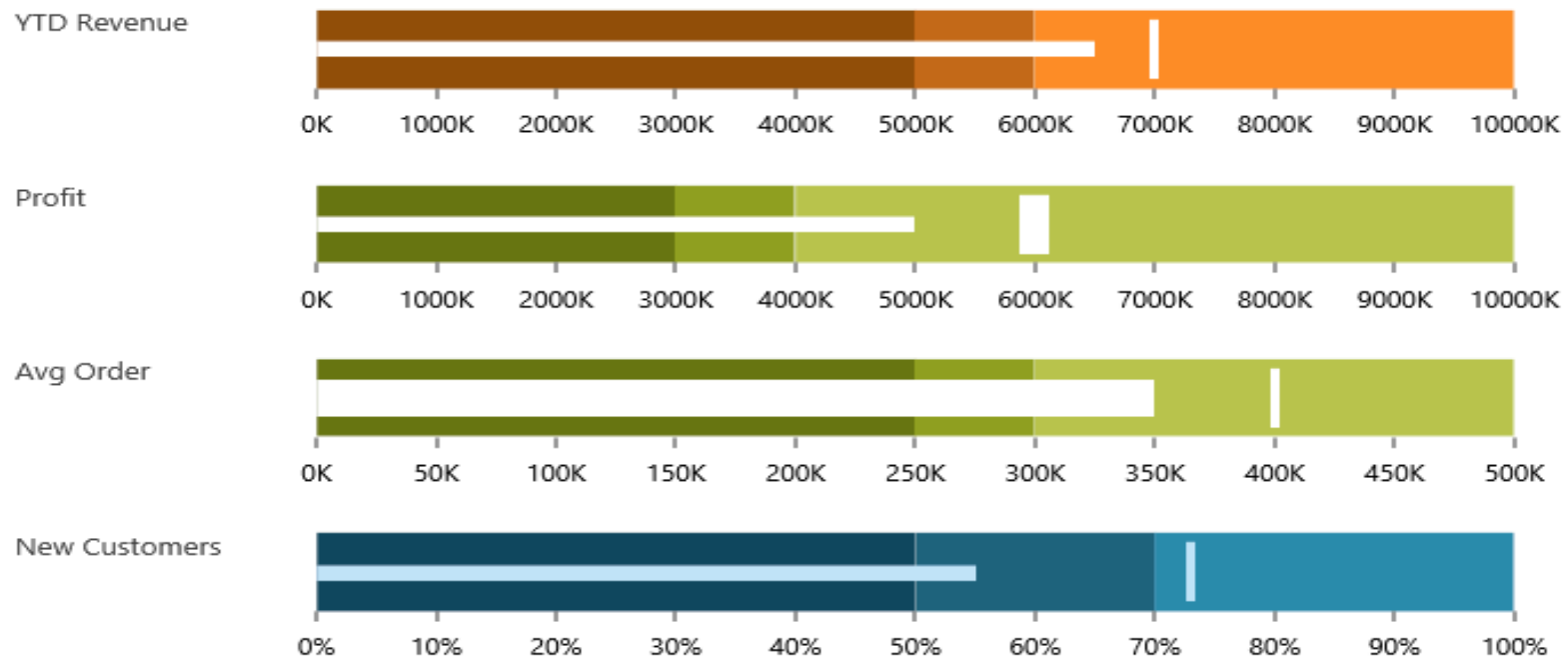
The sales totals to date for 2013 (white horizontal line), clearly exceed the total sales for all of 2012 (the begginning of the middle range). The 2013 sales numbers suggest that our new marketing campaign is successful, resulting in increased product penetration and a significant sales boost, working our way up to the targeted goals for the whole year (vertical white lines).

In this example, vertical with bar is the target, horizontal white line is what has been achieved

Do you want to show the composition of something?

- Use this type of chart to show how individual parts make up the whole of something, such as the device type used for mobile visitors to your website or total sales broken down by sales rep. Better suited to compare categories.
- To show composition, use these charts:
  - Pie
  - Stacked Bar
 - Stacked Column
  - Area
  - Waterfall

A **stacked** (bar or column) **chart** is used to break down and compare parts of a whole. Each bar in the **chart** represents a whole, and segments in the bar represent different parts or categories of that whole.
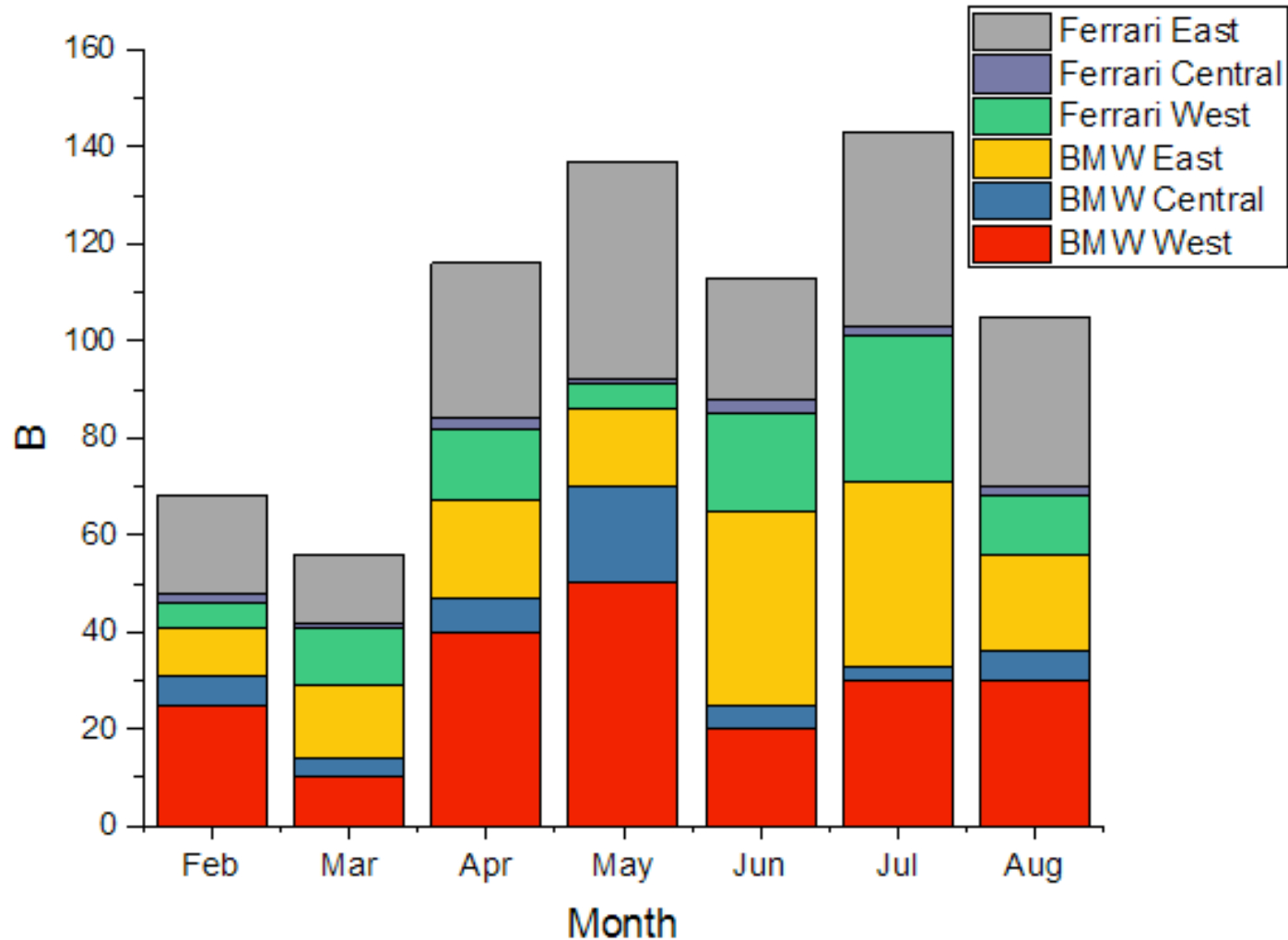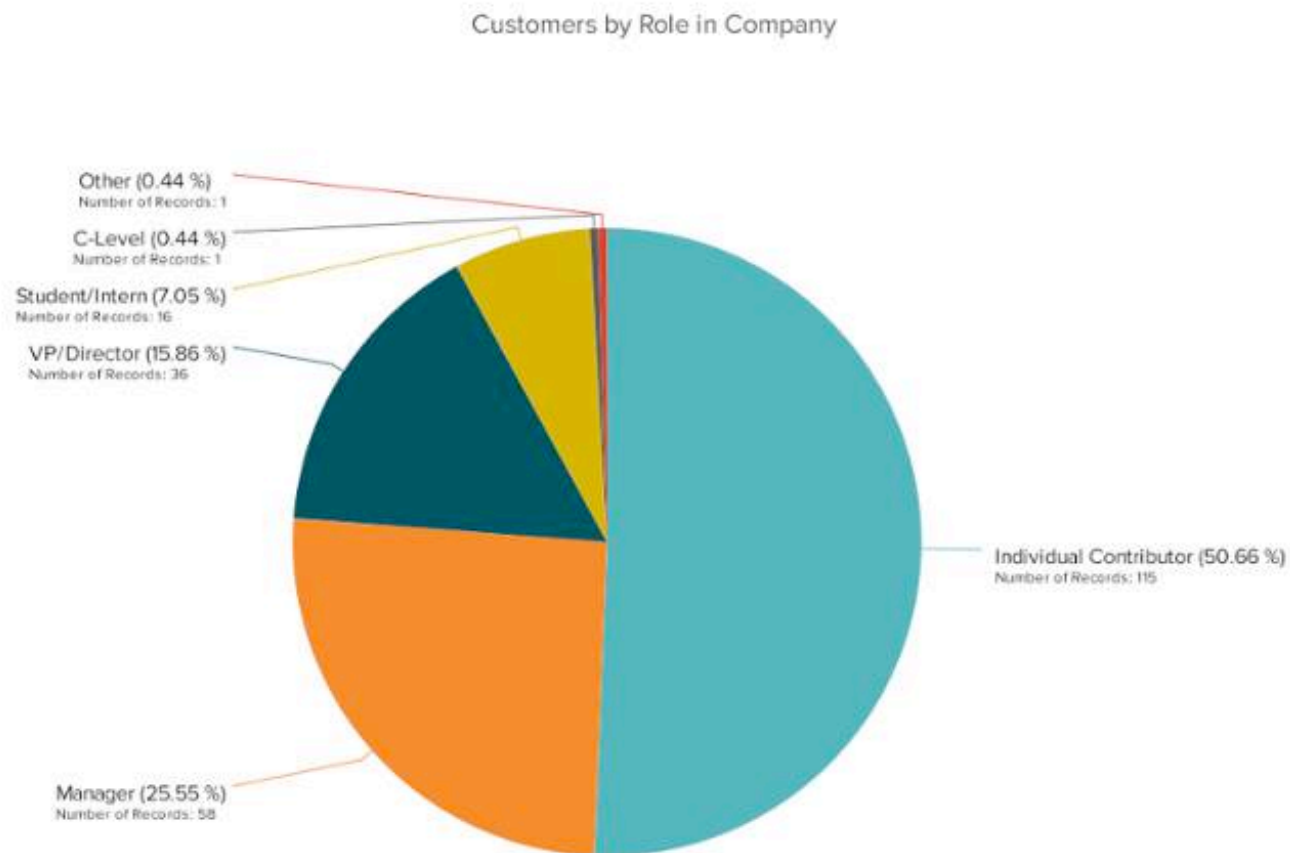
An **area chart** is basically a line chart, but the space between the x-axis and the line is filled with a color or pattern. It is useful for showing **part-to-whole relations**, such as showing individual sales reps' contribution to total sales for a year. It helps you analyze both overall and individual trend information.

Profit made by a chain of grocery stores

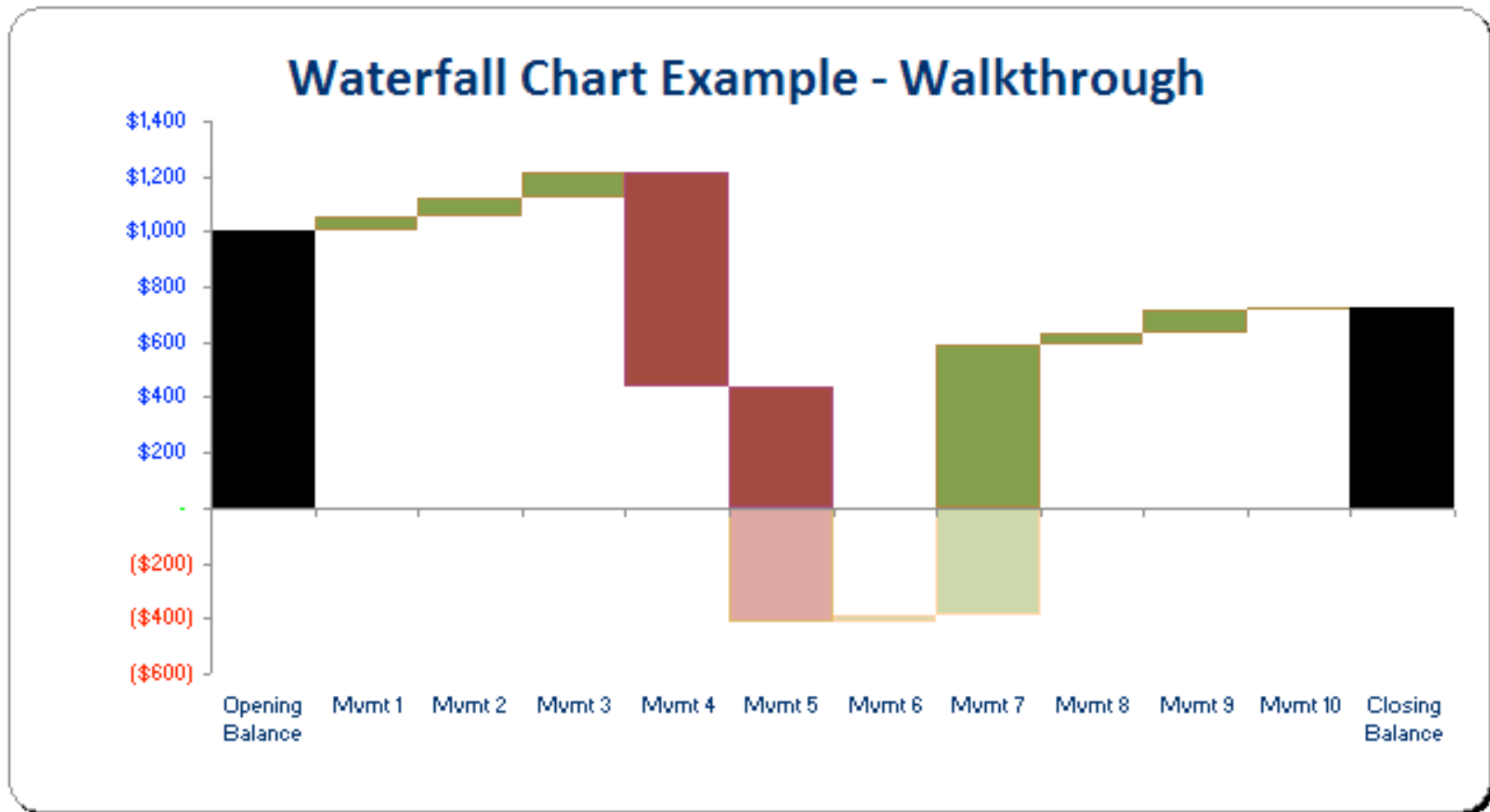Legend: Store 1, Store 2, Store 3, Store 4

Stream charts (a.k.o. area charts)
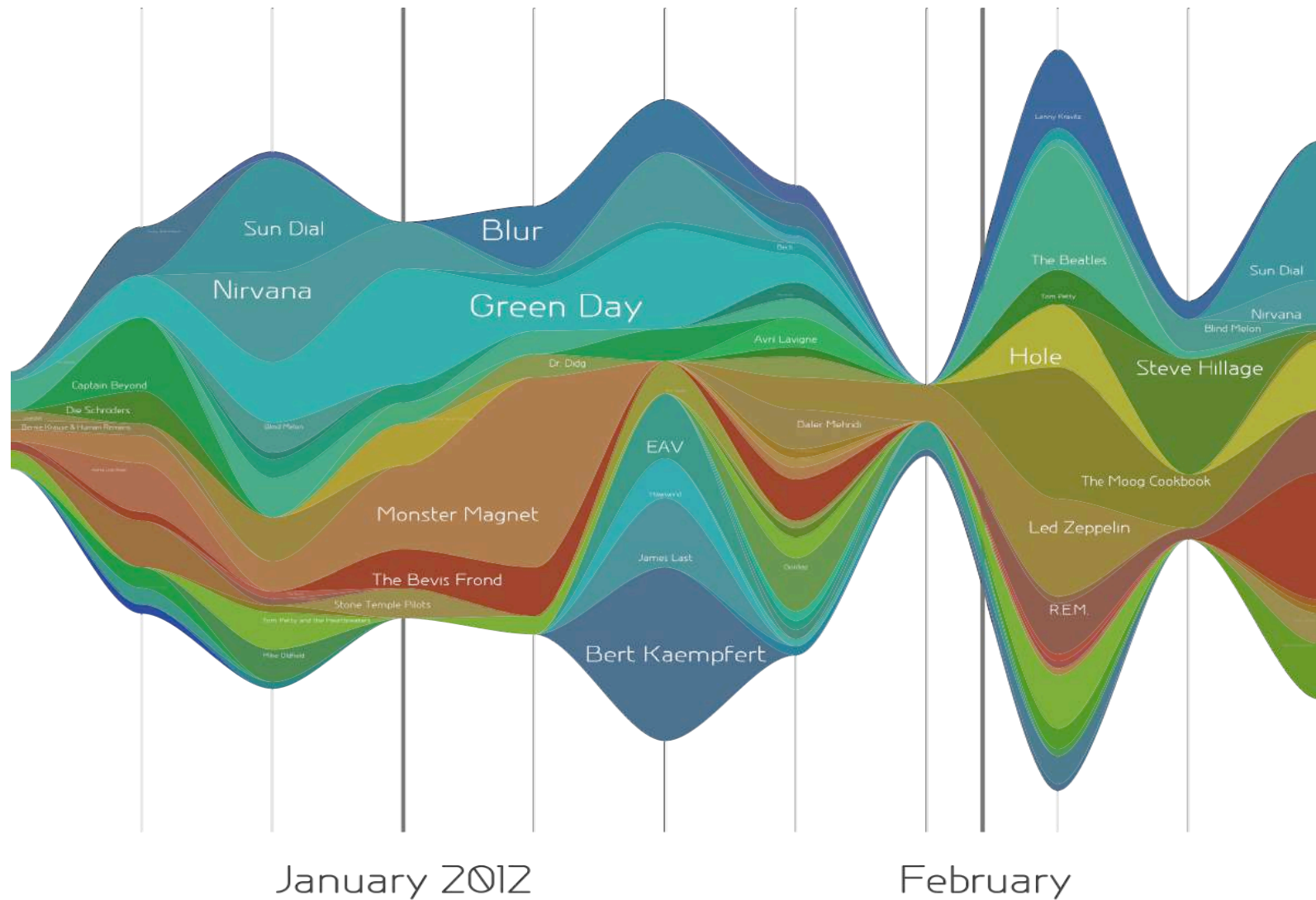
A **pie chart** shows a static number and how **categories** represent part of a whole -- the composition of something. A pie chart represents numbers in percentages, and the total sum of all segments needs to equal 100%.

Customers by Role in Company

Other (0.44 %)
Number of Records: 1

C-Level (0.44 %)
Number of Records: 1

Student/Intern (7.05 %)
Number of Records: 16

VP/Director (15.86 %)
Number of Records: 36

Individual Contributor (50.66 %)
Number of Records: 115

Manager (25.55 %)
Number of Records: 58

DataHero

A **waterfall chart** should be used to show how an initial value is affected by intermediate values -- either positive or negative -- and resulted in a final value. This should be used to reveal the composition of a number. An example of this would be to showcase how subsequent movements affected the total balance (increases in value are coloured green and decreases are coloured red).

Waterfall Chart Example - Walkthrough

$1,400	
$1,200	
$1,000	
$800	
$600	
$400	
$200	
-	
($200)	
($400)	
($600)	

Opening Balance Mvmt 1 Mvmt 2 Mvmt 3 Mvmt 4 Mvmt 5 Mvmt 6 Mvmt 7 Mvmt 8 Mvmt 9 Mvmt 10 Closing Balance

Do you want to understand the distribution of your data?

- Distribution charts help you to understand outliers, the normal tendency, and the range of information in your values.

- Use these charts to show distribution:
 - Scatter Plot
  - Line
  - Column
  - Bar

A **scatter chart** will show the relationship between two different variables or it can reveal the distribution trends. It should be used when there are many different data points, and you want to highlight similarities in the data set. This is useful when looking for outliers or for understanding the distribution of your data.

Customer Happines by Response Time
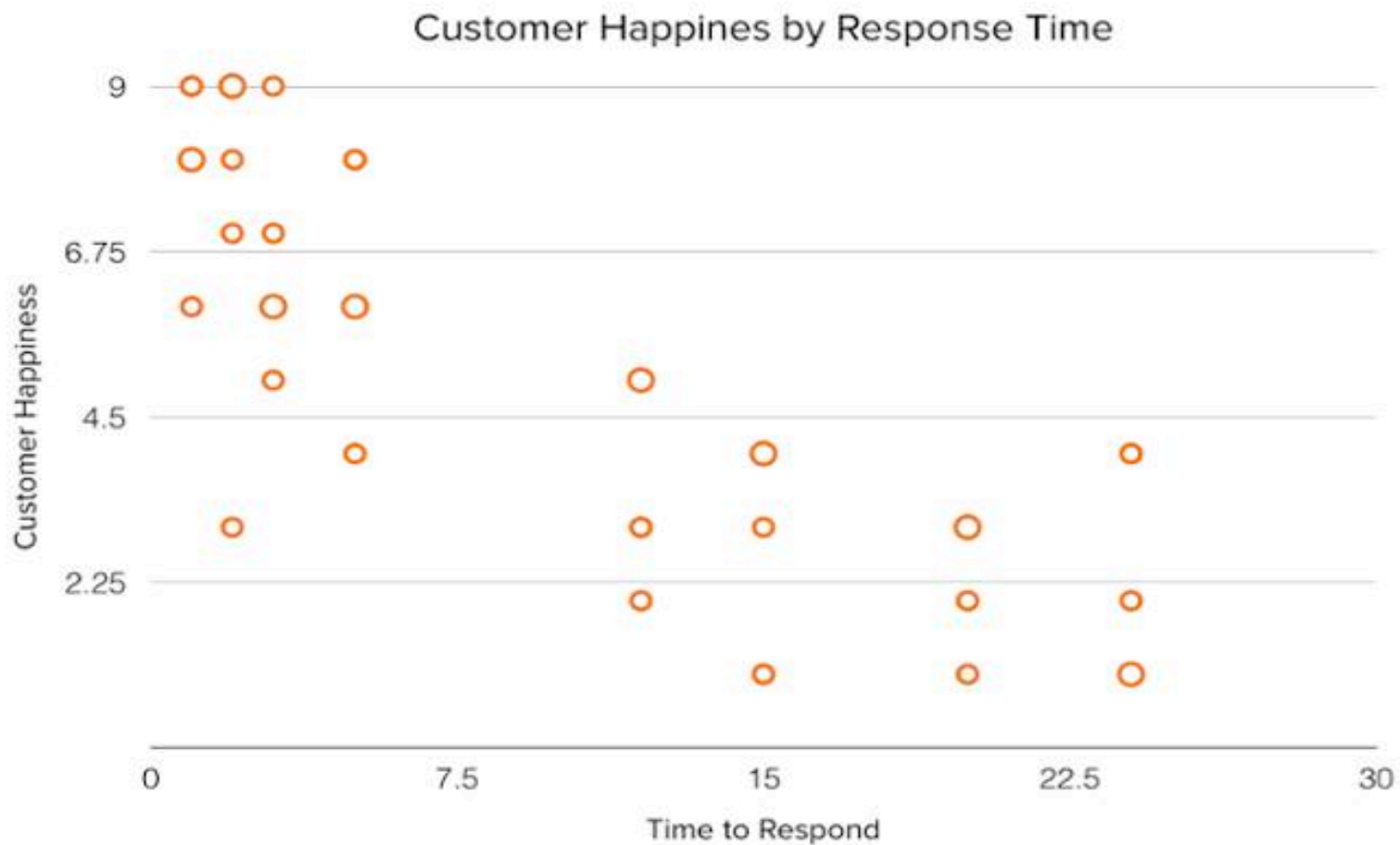
Are you interested in analyzing trends in your data set?

- If you want to know more information about how a data set performed during a specific time period, there are specific chart types that do extremely well.

- You should choose a:
  - Line
  - Dual-Axis Line
  - Funnels
  - Column

A **dual axis** chart allows you to plot data using **two y-axes and a shared x-axis.** It's used with three data sets, one of which is based on a **continuous** set of data and another which is better suited to being grouped by **category**. This should be used **to visualize a correlation or the lack thereof** between these three data sets.

Revenues and Profit by Month for 2014
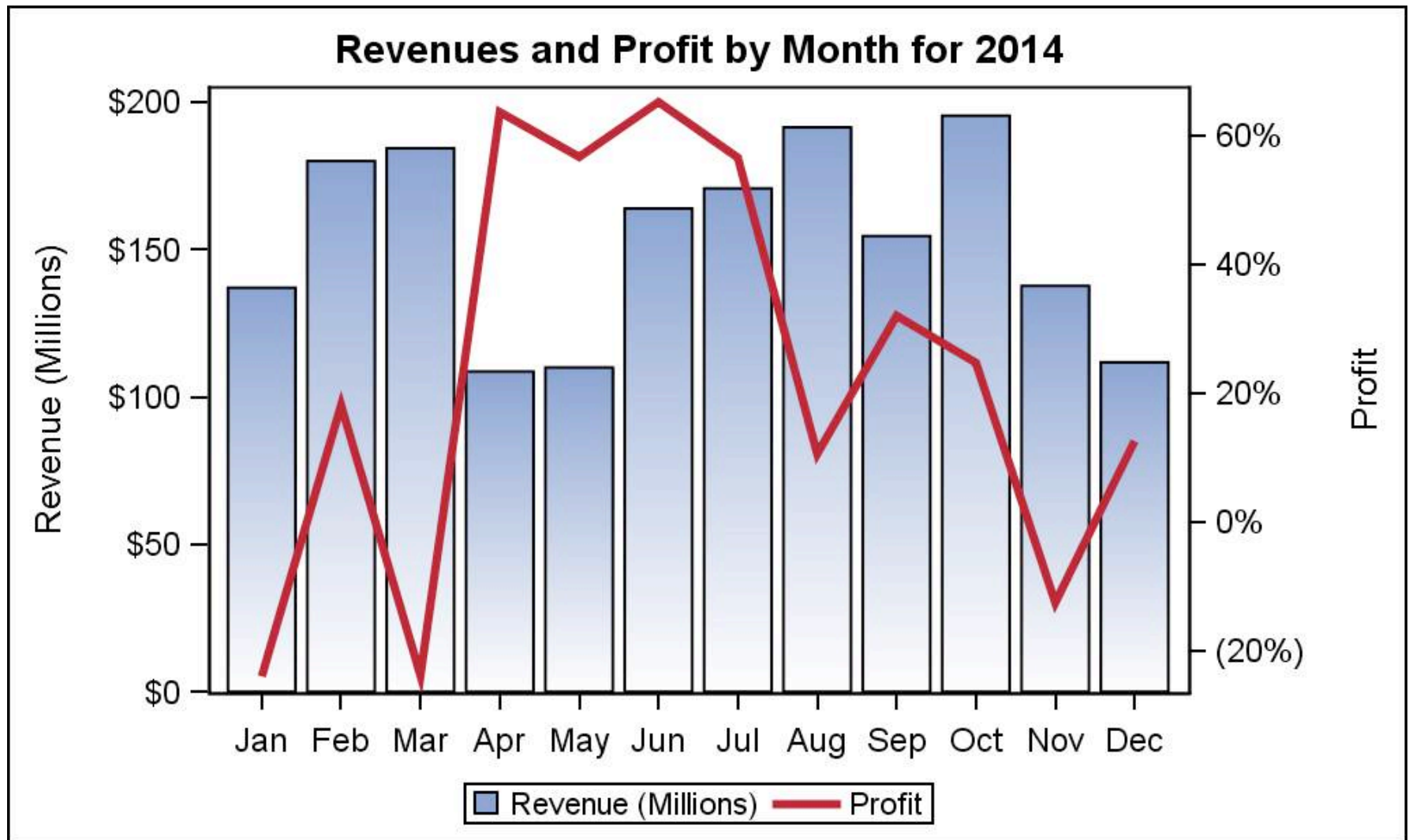
A **funnel chart** shows a series of steps and the completion rate for each step. This can be used to track the sales process or the conversion rate across a series of pages or steps (e.g. from contacts to contracts).

100%
IDENTIFIED 250

LOST 20
UNQUALIFIED 9

60%
CONTACTED 150

LOST 15

40%
DISCUSSION 100

LOST 35

6%
WON

15

Do you want to better understand the relationship between value sets?

- Relationship charts are suited to showing how one variable relates to one or numerous different variables. You could use this to show how something positively effects, has no effect, or negatively effects another variable.
- When trying to establish the relationship between things, use these charts:
  - Scatter Plot
  - Heat maps
  - Bubble
  - Line
  - Networks
  - Spirals

A **bubble chart** is similar to a scatter plot in that it can show distribution or relationship. There is a third dimension, which is indicated by the size of the bubble or circle.

Hours Spent Online by Age and Gender
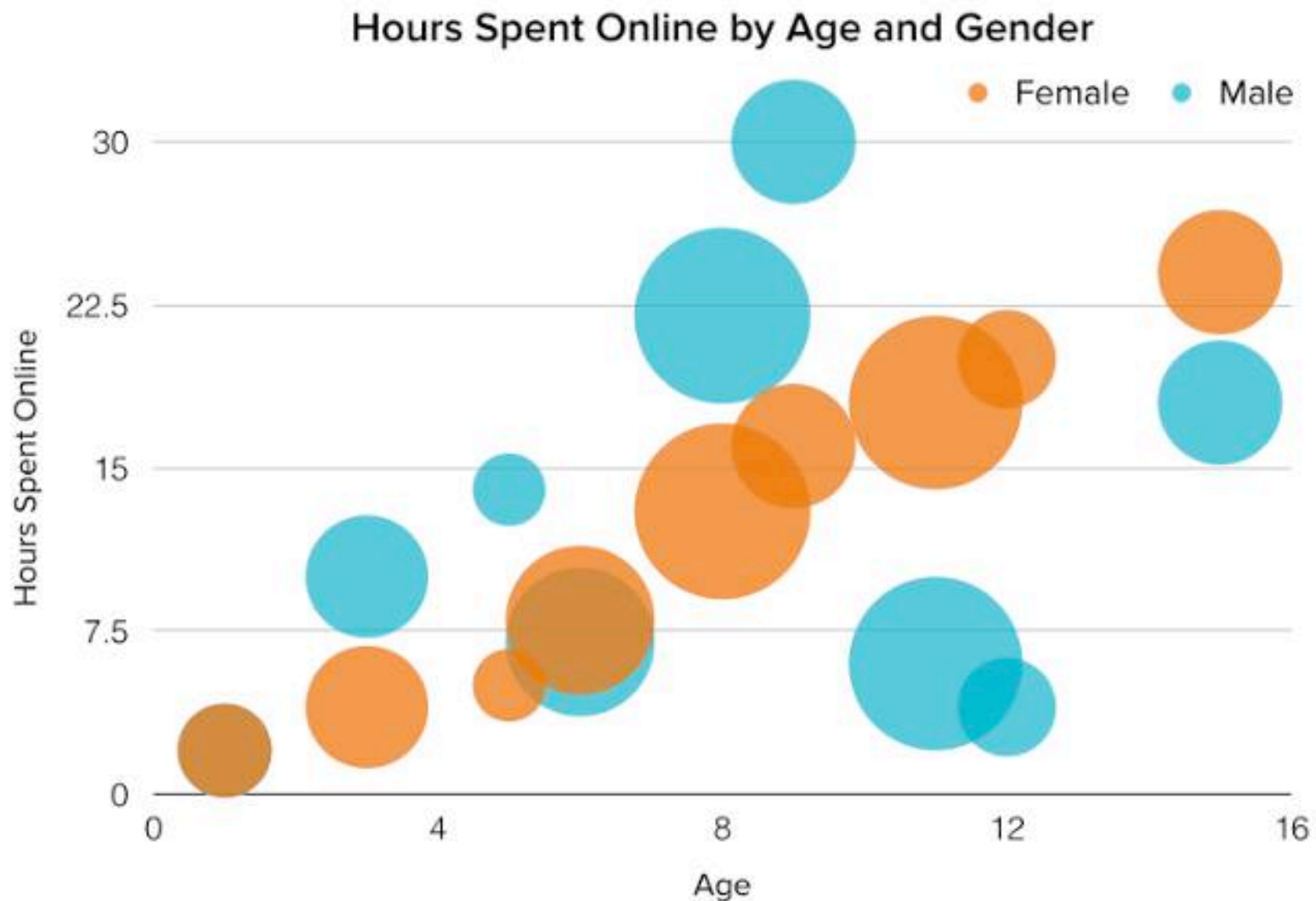
A **heat map** shows the relationship between two items and provides rating information, such as high to low or poor to excellent. The rating information is displayed using varying colors or saturation.

Highest Degree vs. Class Identification

HIGHEST DEGREE	1 - lower class	2 - working class	3 - middle class	4 - upper class
0 - Less than HS	67	144	78	8
1 - High School	91	552	337	17
2 - Junior College	9	87	72	4
3 - Bachelor	6	90	230	23
4 - Graduate	2	32	136	21

SUBJECTIVE CLASS IDENTIFICATION

0   200   400

DataHero

Heat maps can also be geographical maps
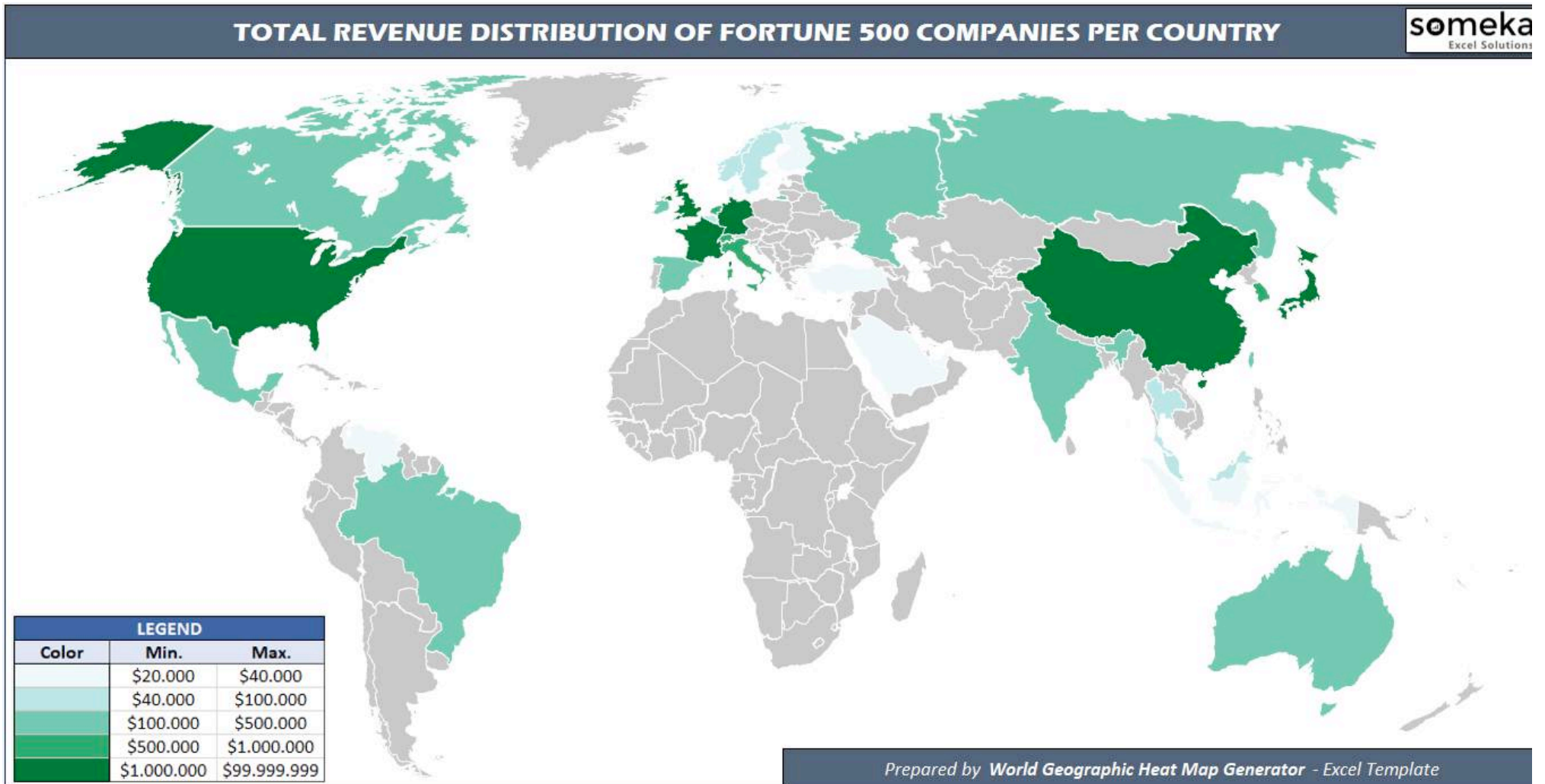
Network graphs are useful to show complex, non-numerical relations between entities

THE COMMERCIAL GRAPH
An example of visualizing complex business ecosystems through data.

COMPANY B
★ ★ ★ ⯪ ☆

COMPANY C
★ ★ ★ ★ ☆

Buyer

Supplier

COMPANY A
★ ★ ★ ☆ ☆

COMPANY G
★ ★ ★ ☆ ☆

Supplier

Supplier

COMPANY D
★ ★ ★ ★ ☆

Buyer

Buyer

COMPANY F
★ ★ ★ ⯪ ☆

COMPANY E
★ ★ ★ ☆ ☆

Although they quickly become unreadable
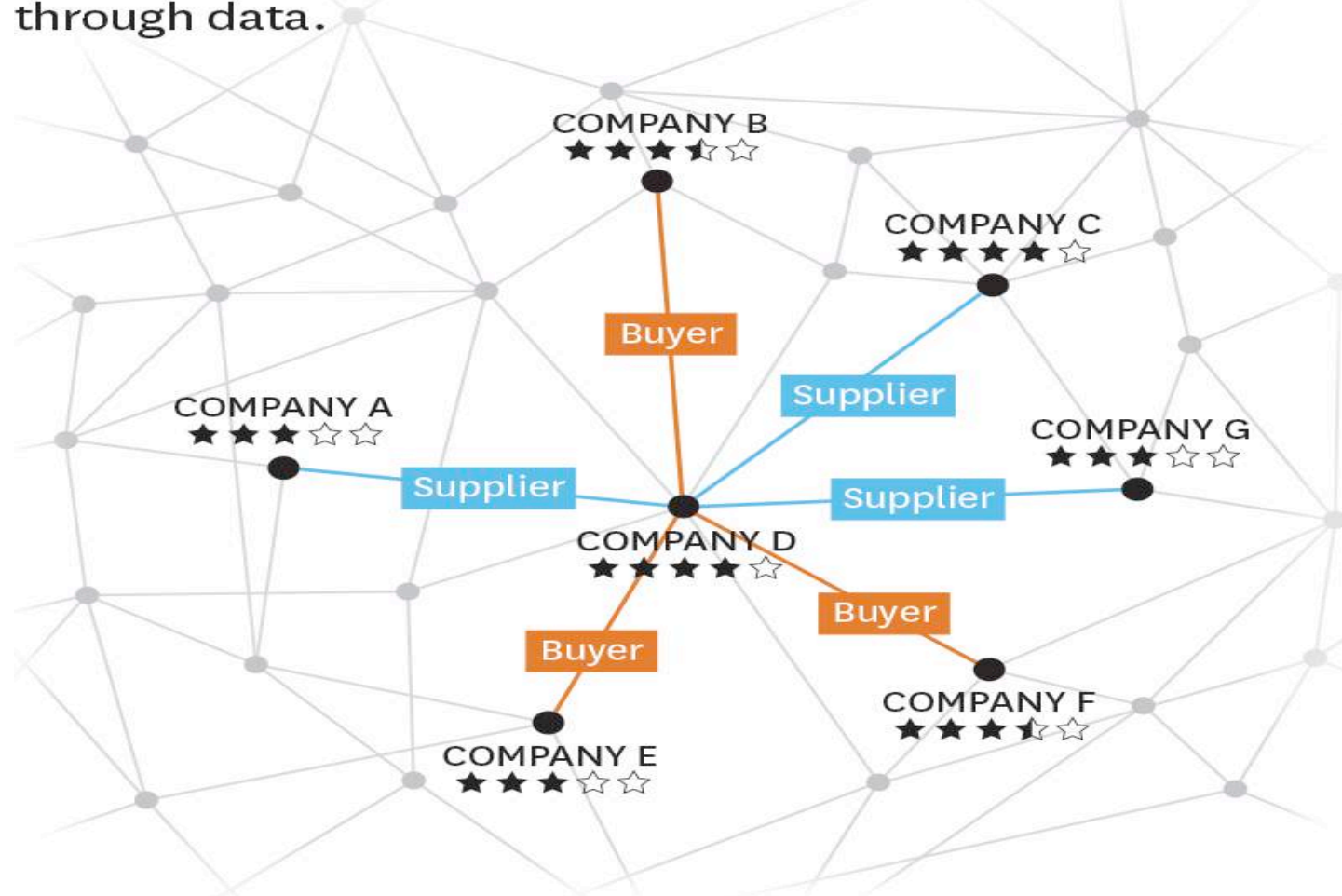
Spirals

- Spiral Plots are ideal for showing large data sets, usually to show trends over a large time period. This makes Spiral Plots great for displaying periodic patterns. Colour can be assigned to each period to break them up and to allow some comparison between each period.

- So for example, if we were to show data over a year, we could assign a colour for each month on the graph.
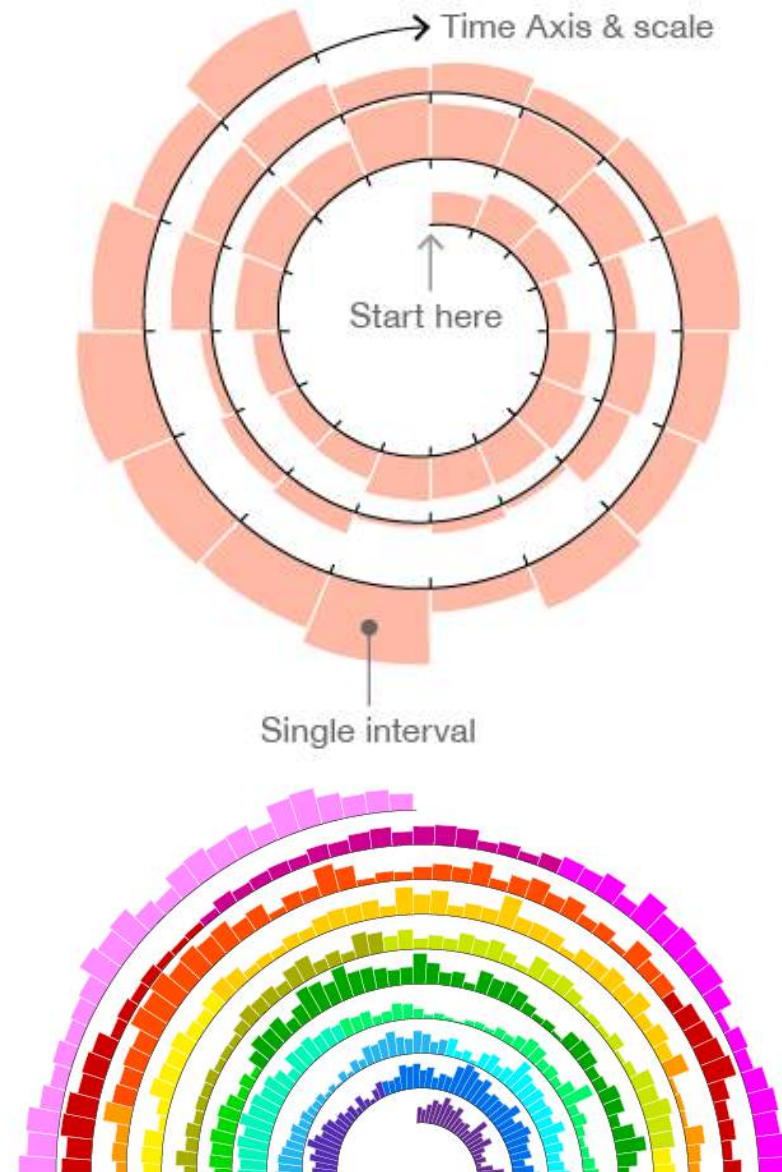
Spirals uses

Here each trip around the circle represent a day, colors represent queue lenghts in streets
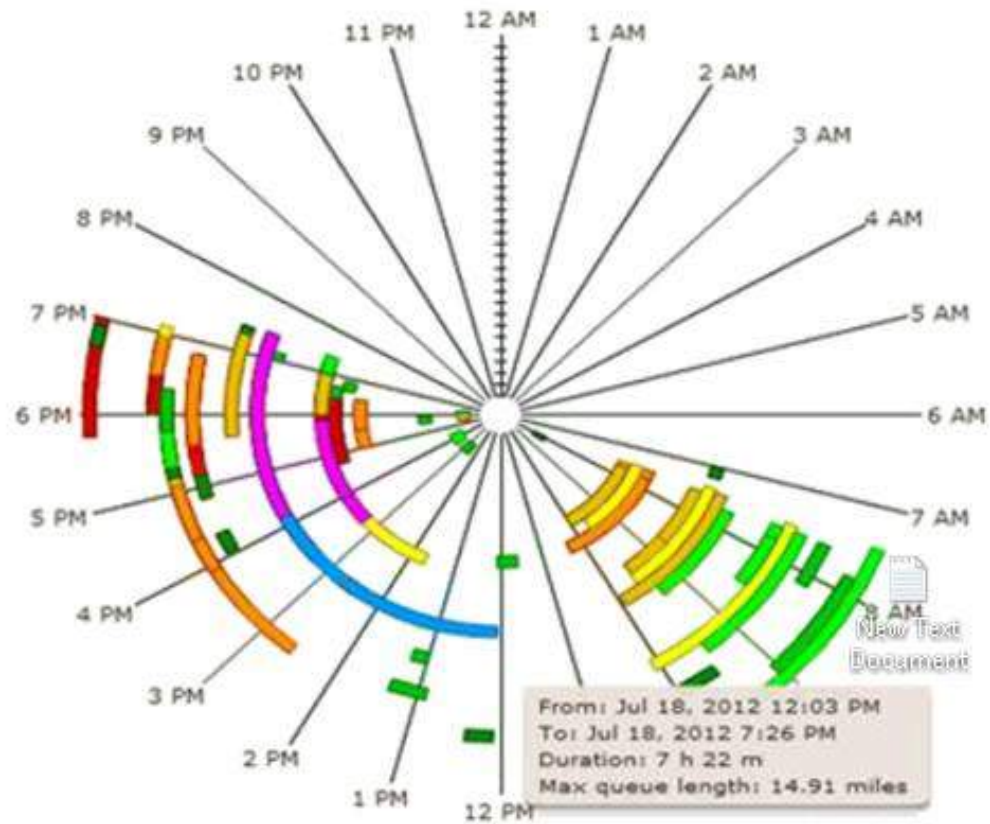
Spiral uses

Here the spiral center is a "target" attribute, and dots represent attributes (single or sets of attributes) that influence the target. Strenght of influence decrease as we move far from the centre along the spiral line
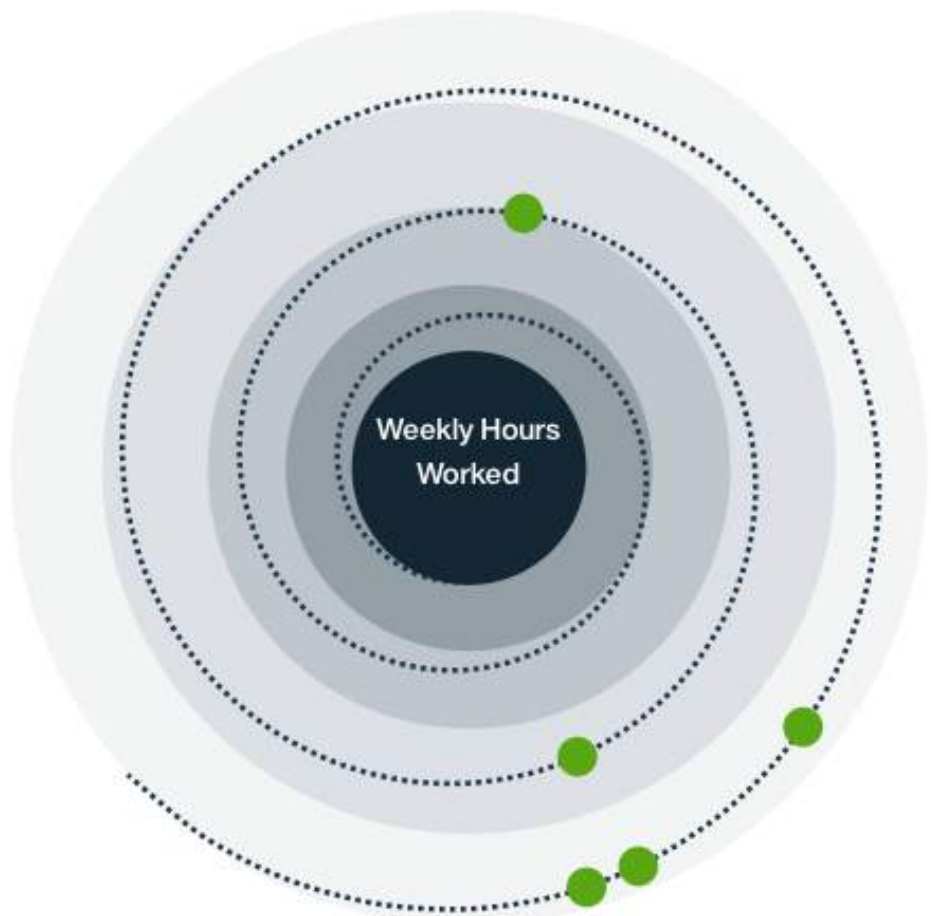
Drivers	Strength
Employment Status	67%
Weekly Earnings	56%
Age Range	24%
Age	19%
Weekly Earnings Bucket	18%

SELECTING VISUALIZATIONS

Depending on the type of data we want to visualize

Certain visualizations are better for specific types of data

	Metric	Attribute (Nominal)	Attribute (Ordinal)
Attribute (Nominal)	Bar Heatmap	Network	Line w/ Break-By Bar w/ Break-By
Attribute (Ordinal)	Column Line		Scatter Grid
Metric	Scatter/Bubble		

Attribute (Nominal) and Metric

Attributes (nominal) and metrics

Comparative Analysis

Bar Chart - Sorted
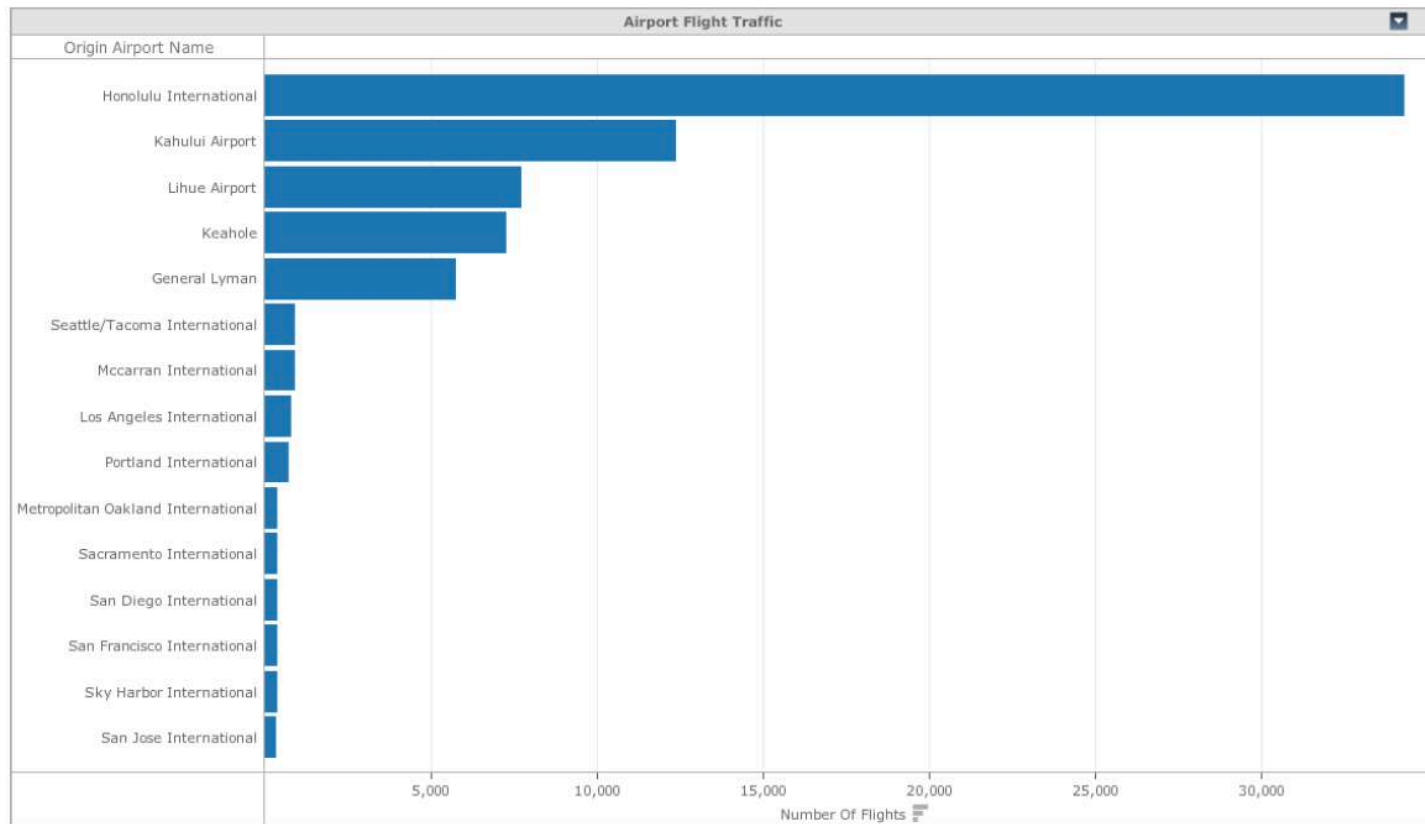
Comparative Analysis

Avoid: Line Chart – Implies continuity between points
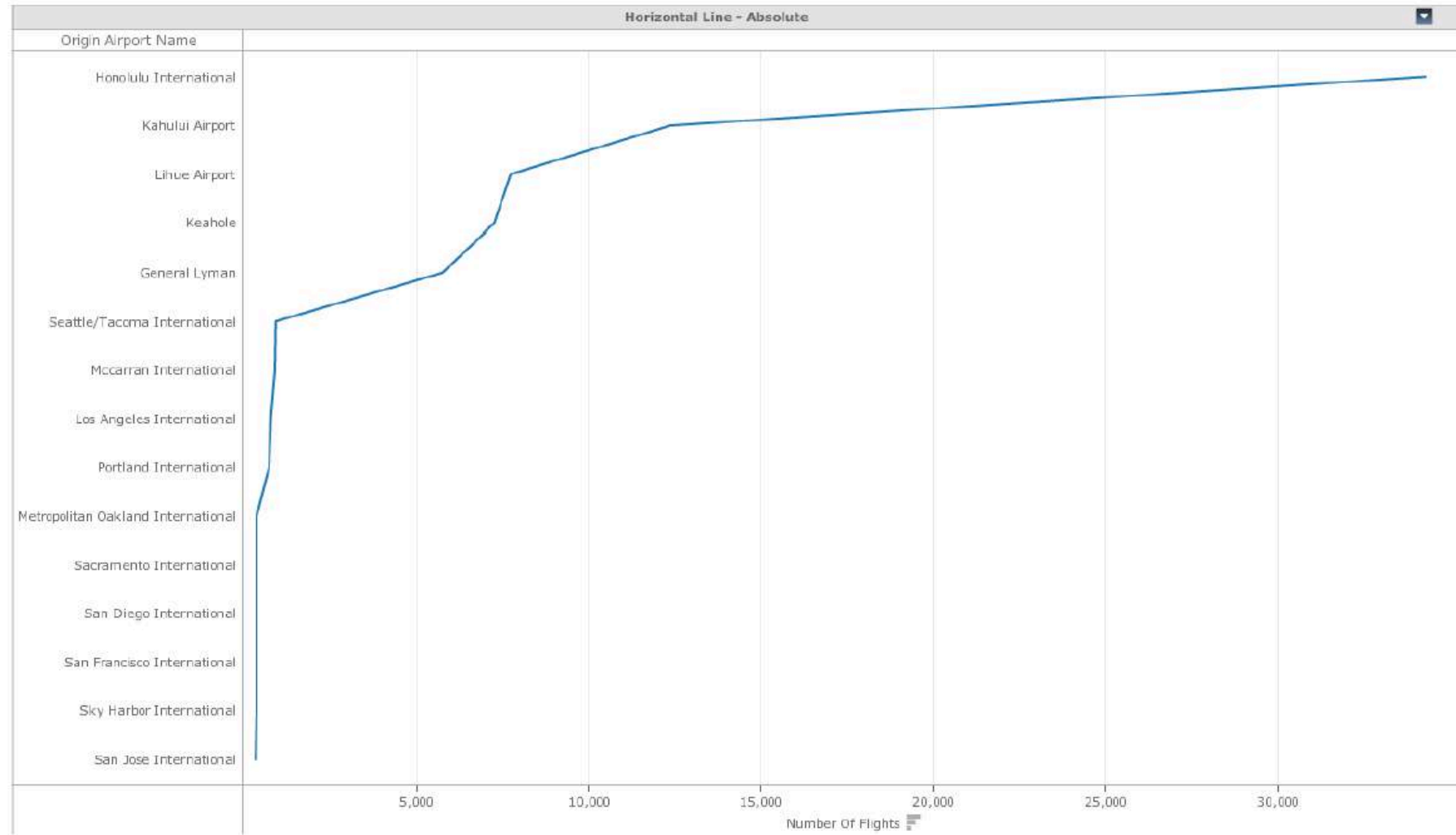
Attribute (Nominal) and Metric

Contribution Analysis – Few Elements

Pie Chart
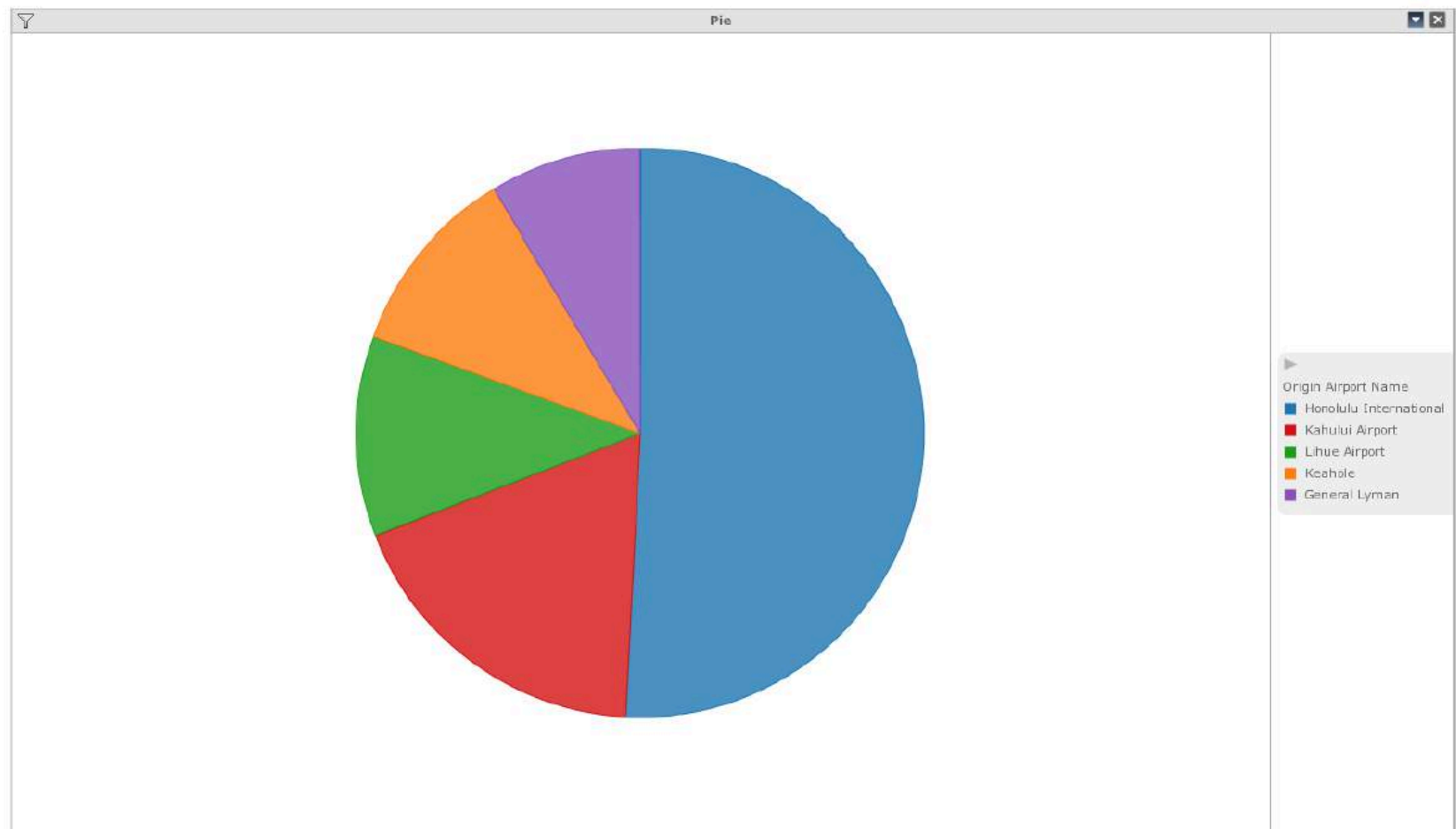
Attribute (nominal) and Metric

Many elements: heat map

	2/10/2014	2/17/2014	2/24/2014	3/3/2014	3/10/2014	3/17/2014	3/24/2014	3/31/2014
P-11857-N3		$56000	$40000	$72000	$96000	$56000	$40000	$72000
P-13498-X5	$367000	$122000	$615000	$613000	$492000	$123000	$976000	$123000
P-13498-X7	$118900	$237800	$240900	$240900	$244000	$118900	$122000	$118900
P-13498-X9		$350000	$280000	$210000	$350000	$280000	$210000	$350000
P-27842-S1		$59800	$70200	$39000	$46800	$54600	$59800	$70200
P-47112-P1					$16000	$16000		
P-47124-P2							$16000	
P-N3T54-200	$12250	$14000	$7000	$12250	$15750	$7000	$7000	$7000
P-N3T57-100	$8000	$28000	$28000	$8000	$8000	$10000	$8000	$16000

Attribute (Ordinal) and Metric

Time-Series Analysis – Few Elements

Column Chart
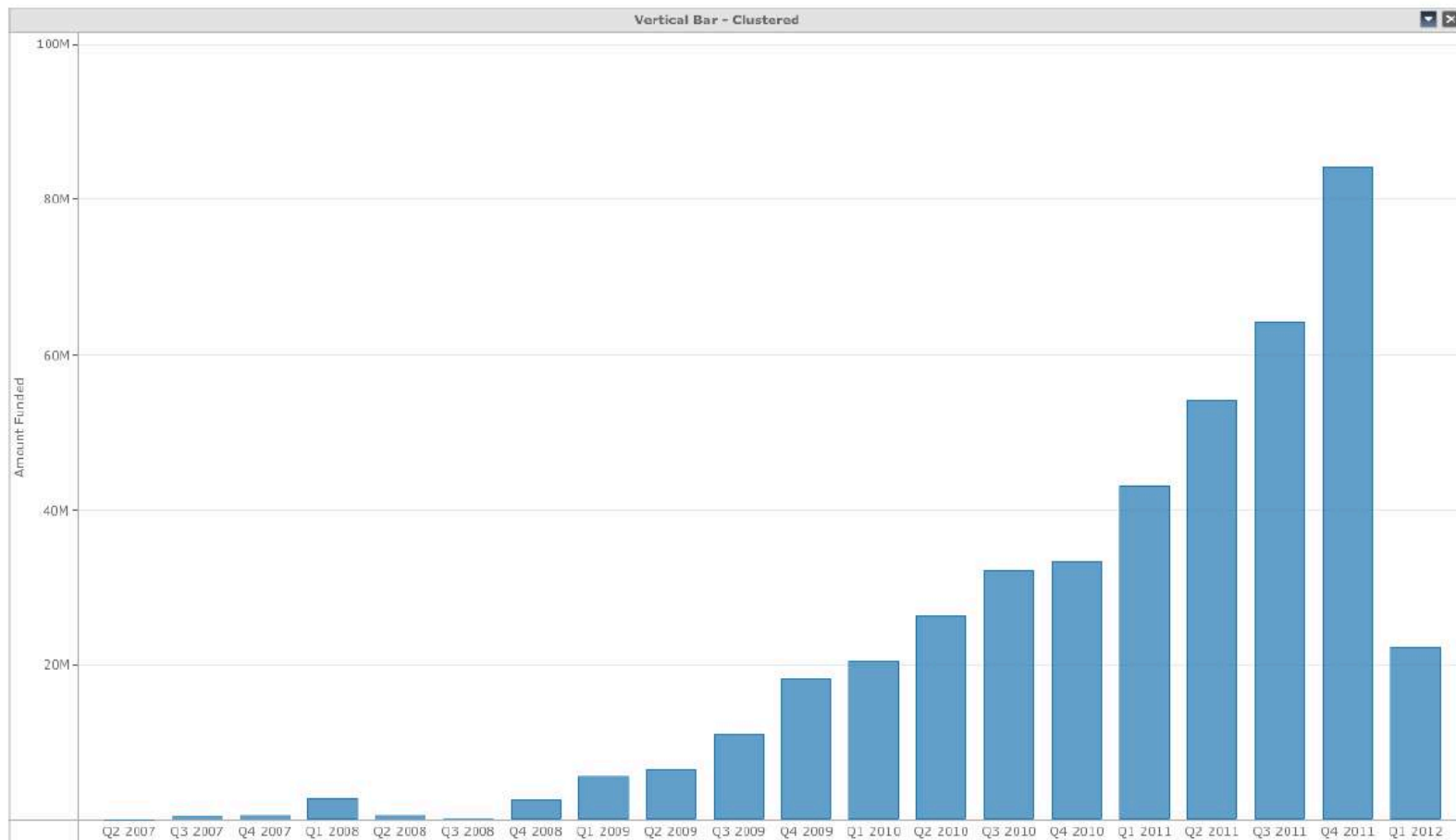
Attribute (Ordinal) and Metric

Time-Series Analysis – Many Elements

Line Chart
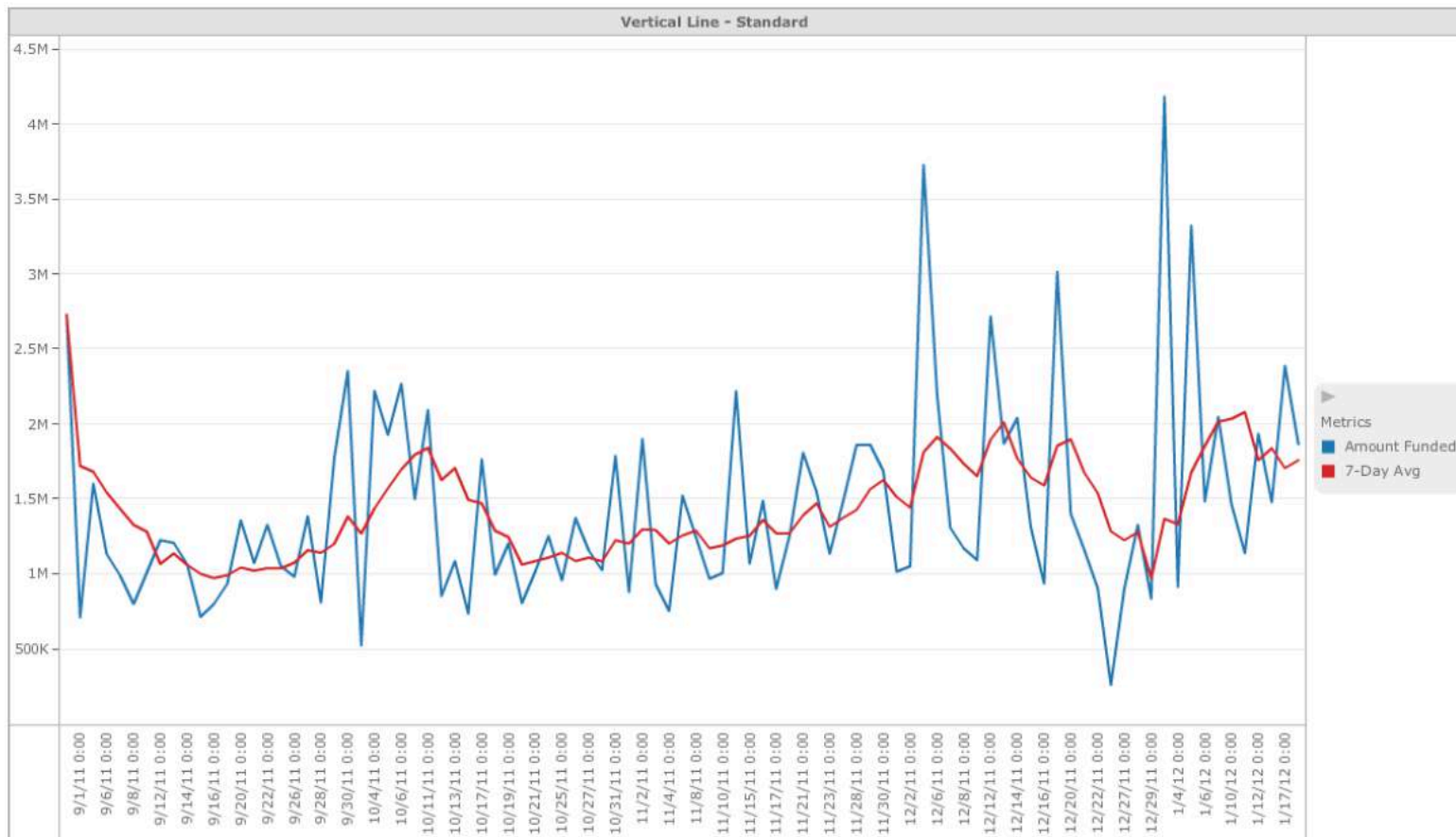
Time-Series Analysis

Avoid: Pie Chart – Removes ordinality
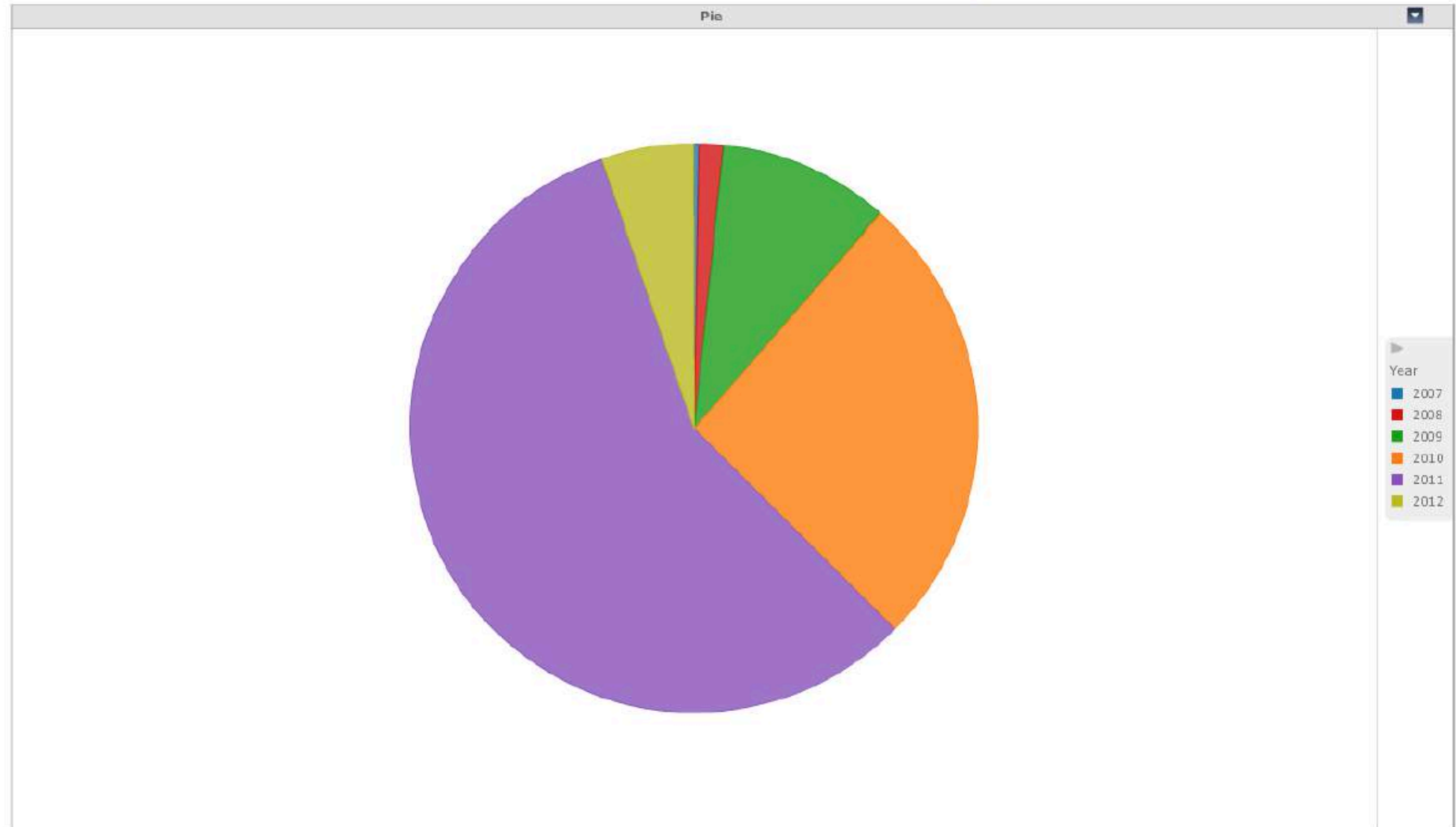
Metric and Metric

Correlation Analysis

Scatter Plot
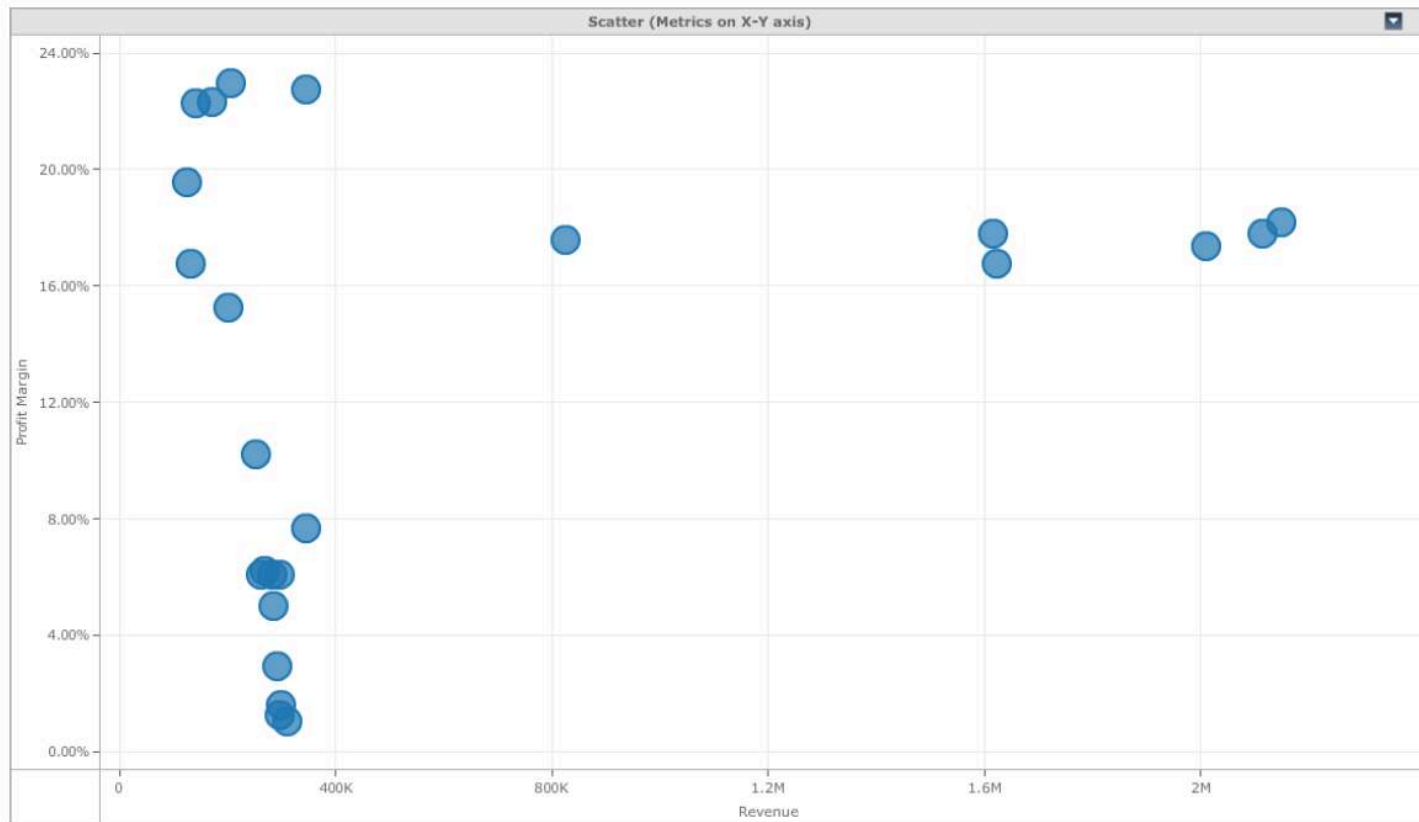
Attribute (Nominal) and Attribute (Nominal)

Market Basket or Network Analysis

Network Visualization

Market Basket or Network Analysis

Avoid: Scatter Grid – Implies ordinality

Attribute (Nominal) and Attribute (Ordinal)
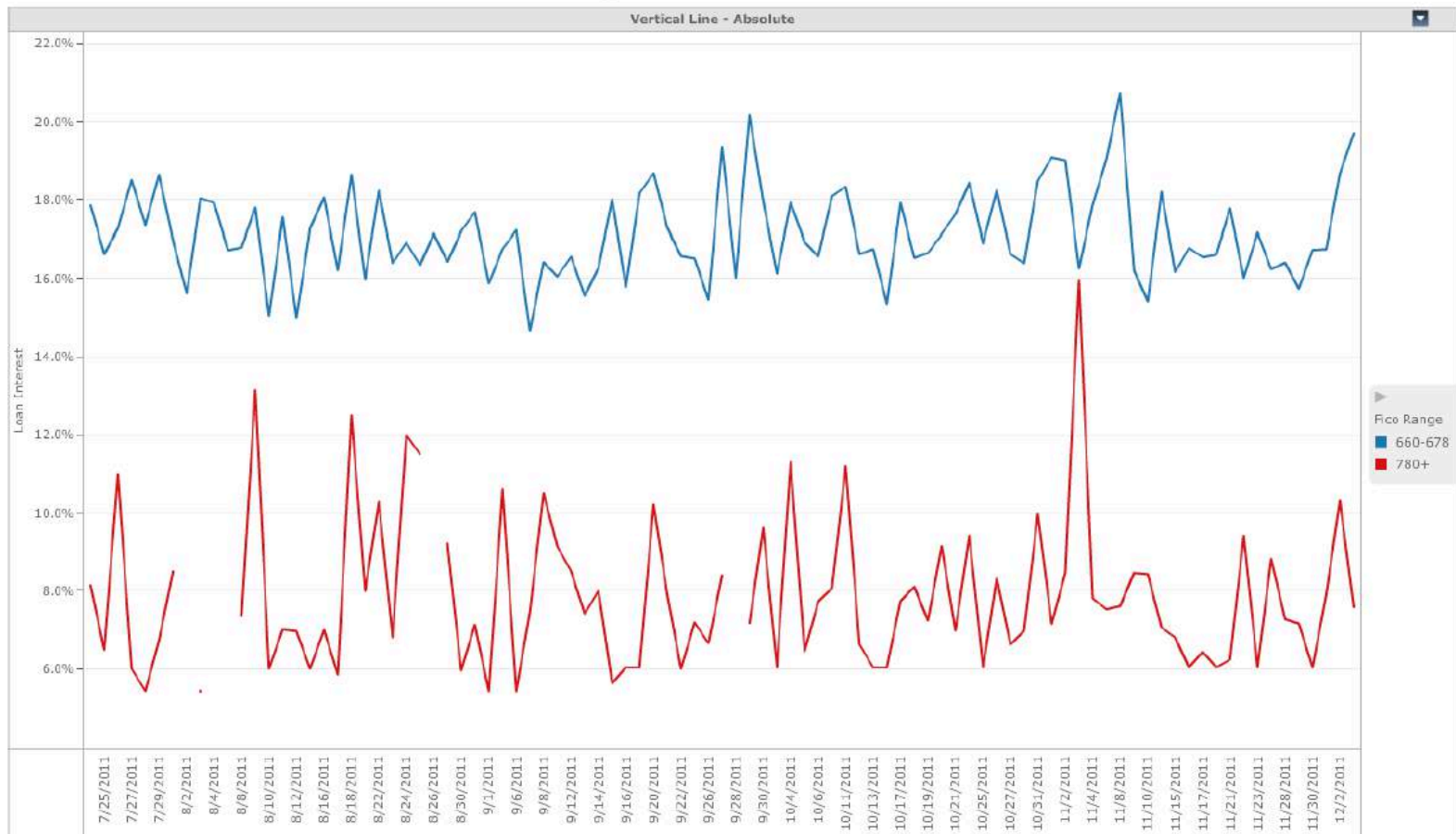
Attribute (Ordinal) and Attribute (Ordinal)

How to improve a visualization

- Colors
- Saturations
- Size
- Interpretability
- Performance
- Layout
- Interactivity

Enhancing Visualizations for Additional Insights

Appropriate Visual Enhancements

	Attribute (Nominal)	Attribute (Ordinal)	Metric
Color Hue	x	x	x
Color Saturation		x	x
Size		x	x

Color (Hue) to Identify Groups of Attributes

Color (Saturation) to Highlight Metric Patterns
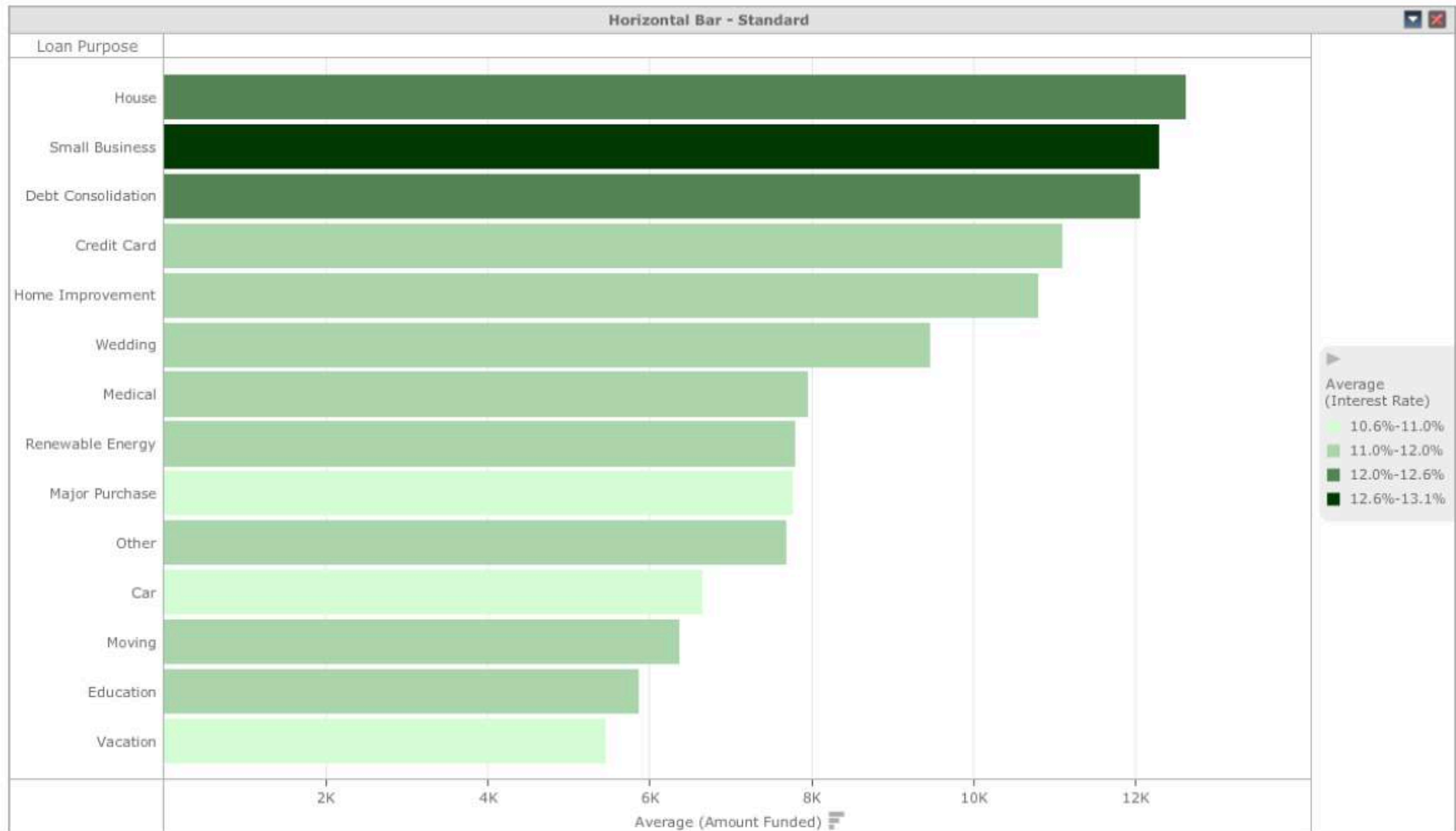
Adding Size to Emphasize Metric Trends
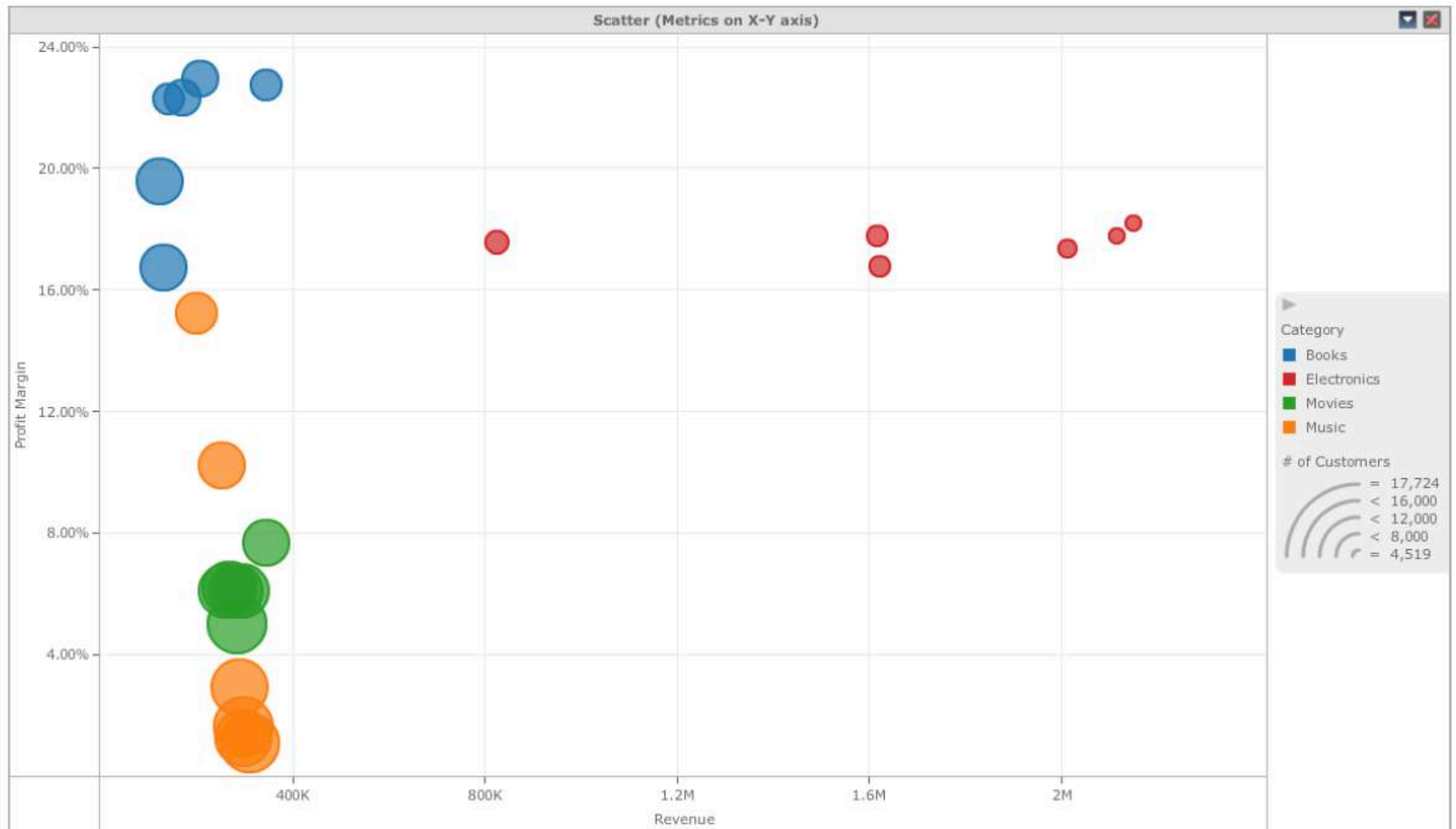
More hints on colors

Colors Should Enhance Data Comprehension, Not Distract

Use Fewer Than 6 Colors
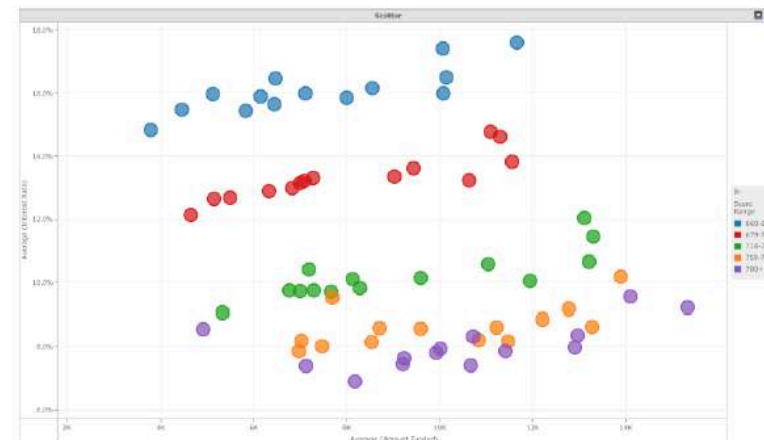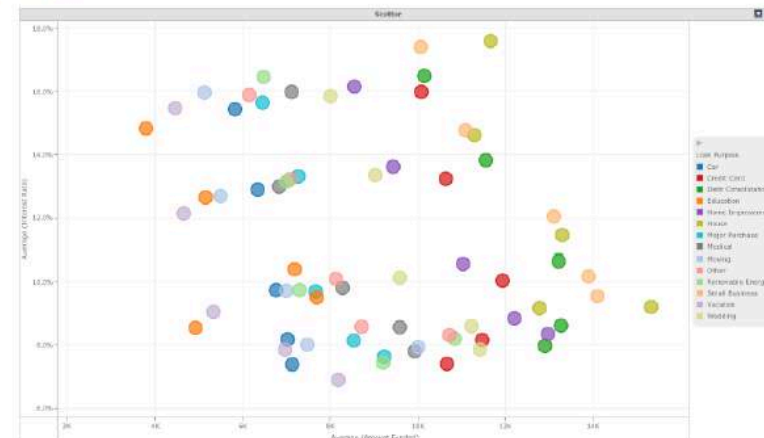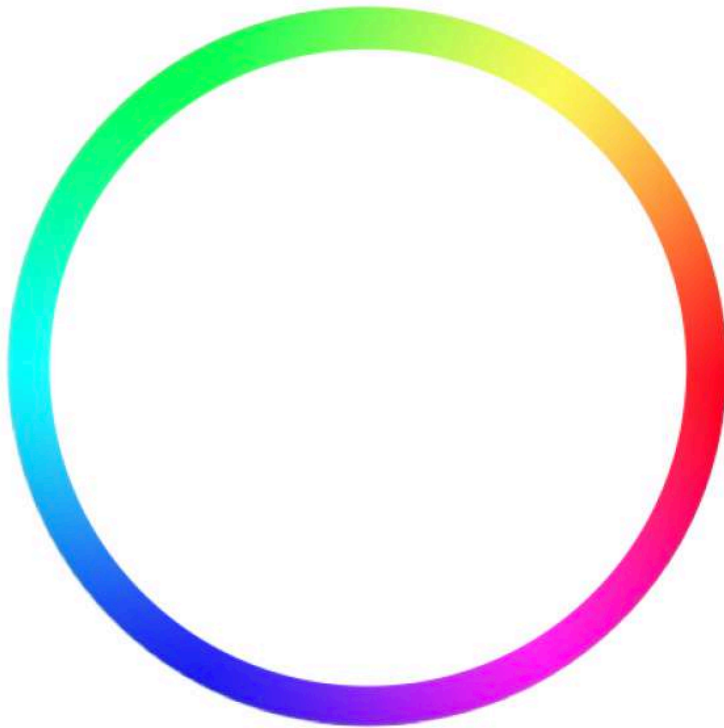
More hints on colors

Use Colors to Emphasize Comparisons
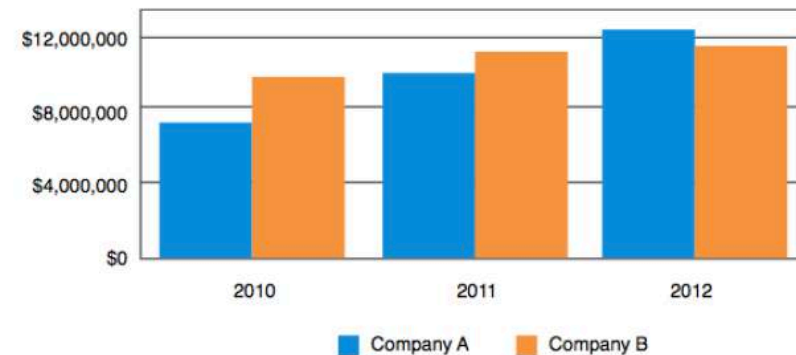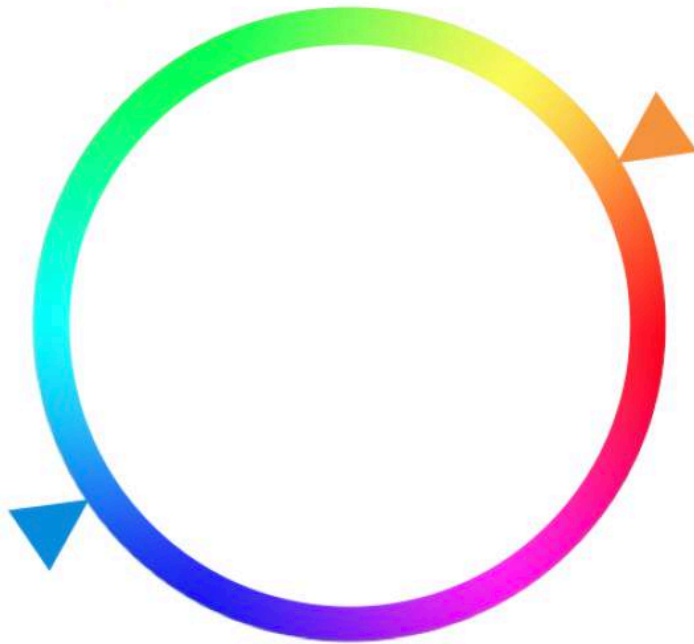
More hints on colors

Avoid color confusion!

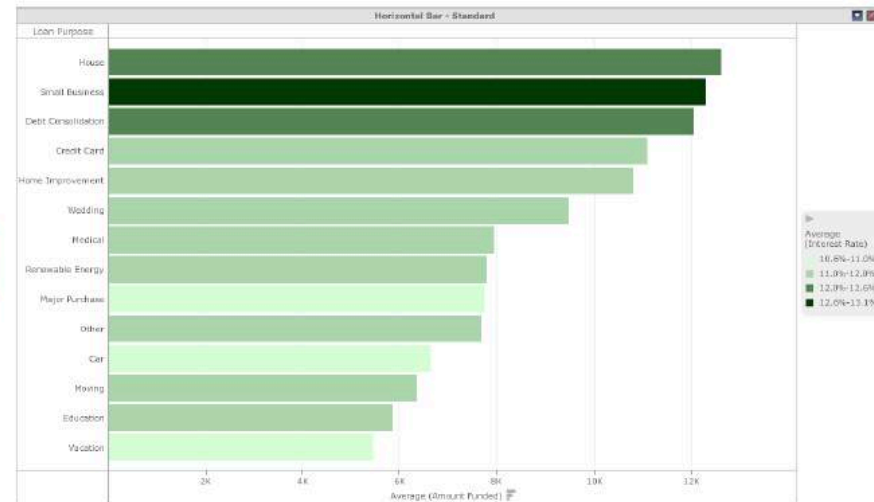Why should we both change from our home kits?
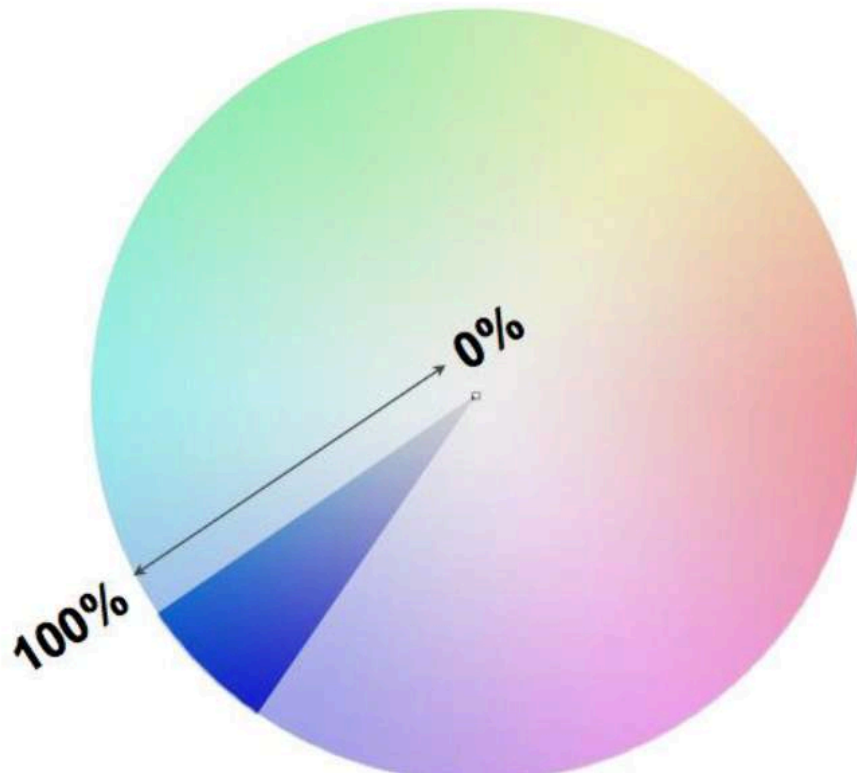
Home Jerseys

Wales

Portugal
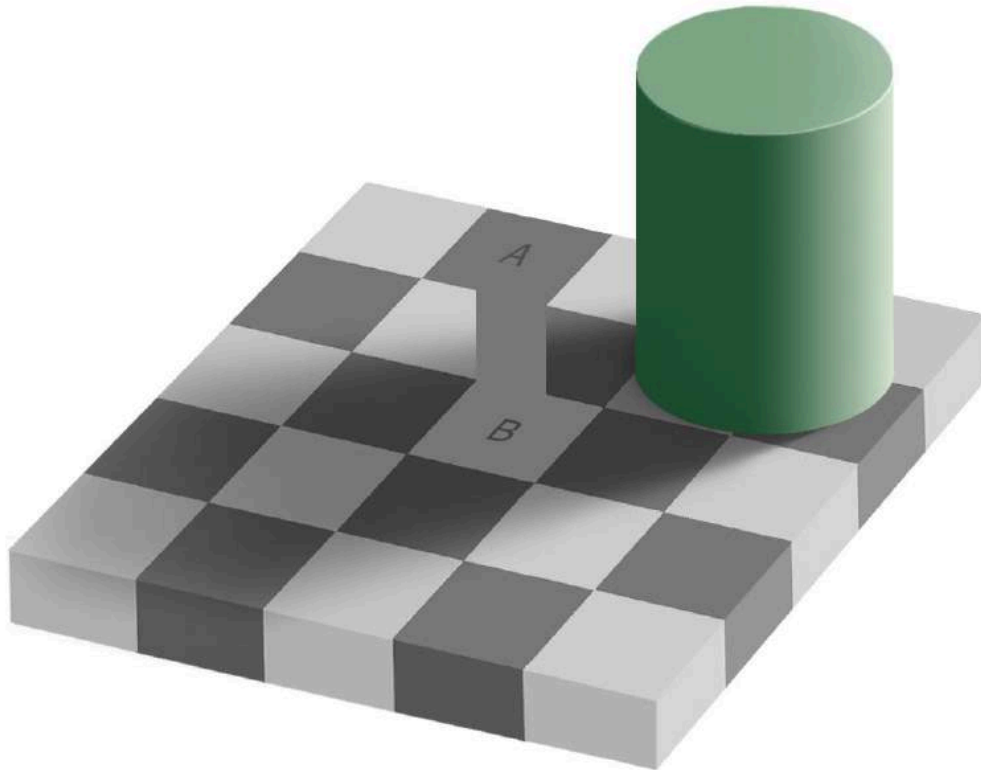
More hints on colors

Use Color Saturation Correctly

Less Saturation: Smaller Values
More Saturation: Greater Values

More hints on colors

Color Constancy Can Confound Data Comprehension
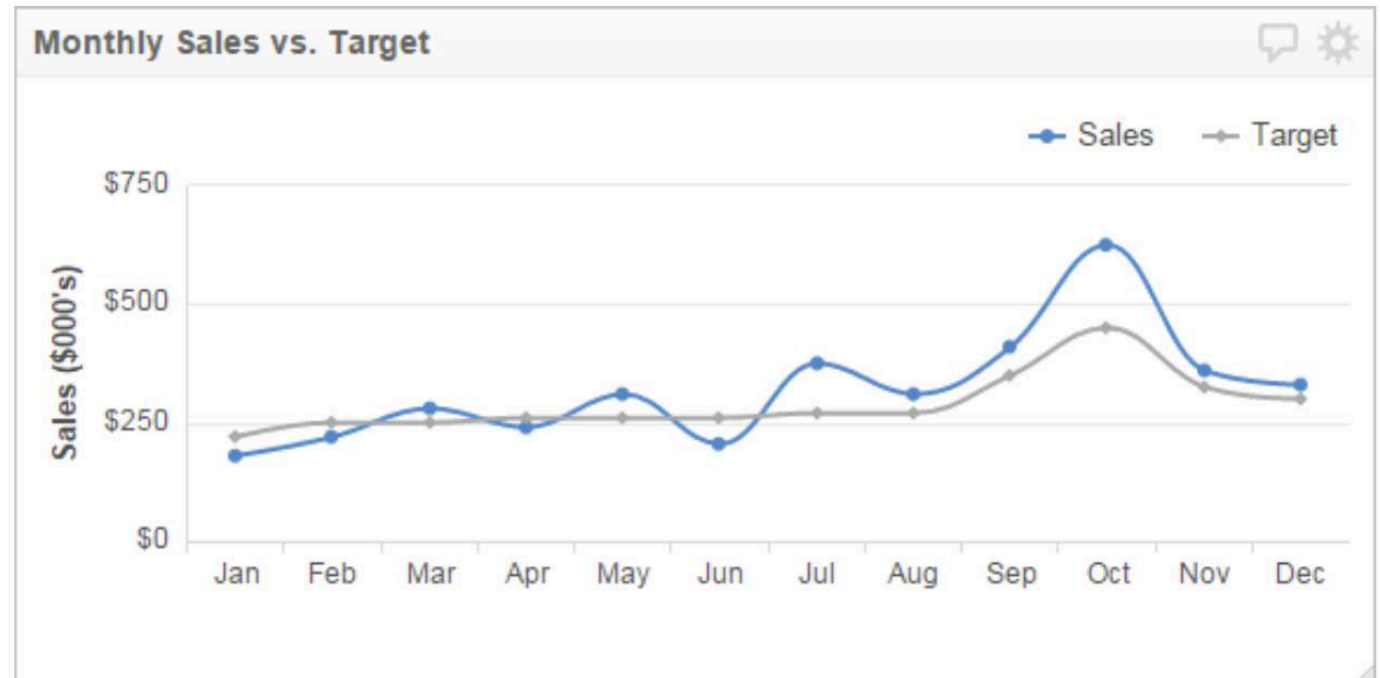
Avoid Color Gradients for Backgrounds

Interpretability: avoid users do the math

Looks typical right?

- Which months were below target?
- Which months were above target?
- And by how much?

Monthly Sales vs. Target

Sales (in $000's): Jan ~$175, Feb ~$215, Mar ~$275, Apr ~$240, May ~$310, Jun ~$200, Jul ~$375, Aug ~$300, Sep ~$410, Oct ~$625, Nov ~$360, Dec ~$330
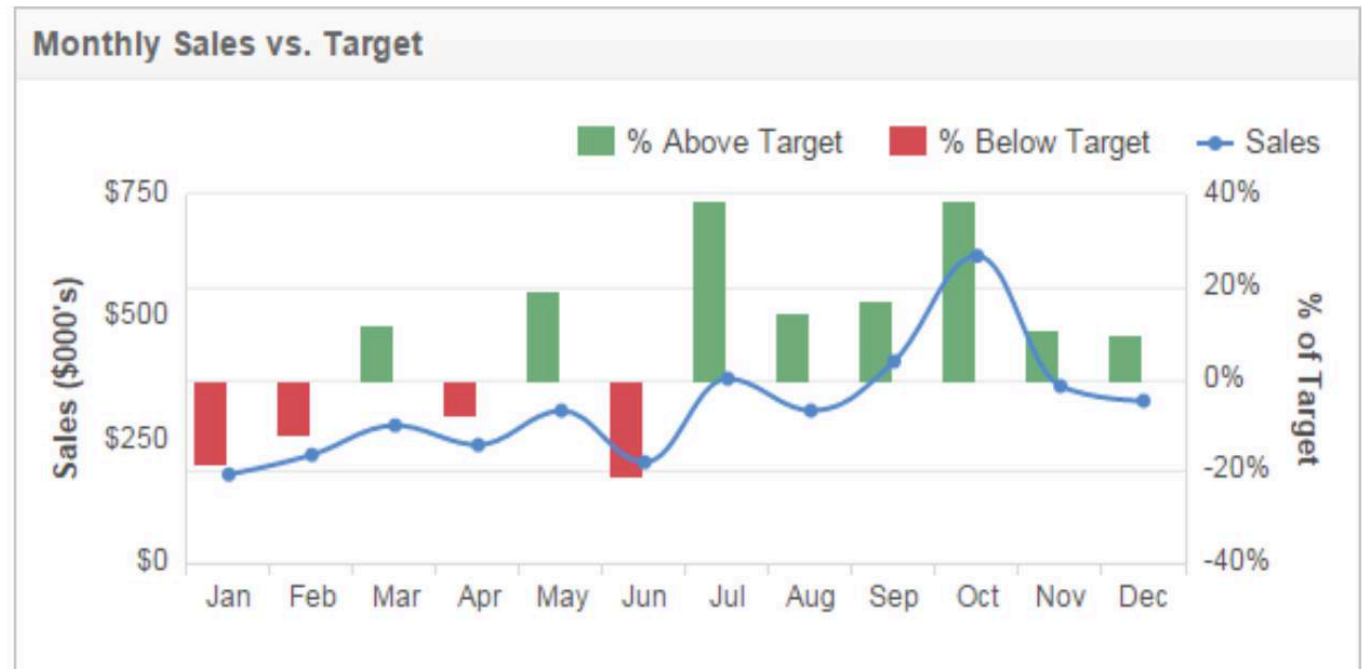
Target values track near $250–$450

Interpretability: avoid users do the math

Go the extra mile to make it great

- Shows Actuals
- Shows Good and Bad months
- Quantifies good and bad
- Uses and Overlay and dual Y Axis.

Interpretability: sloppy labelling

Make it idiot proof, I mean "self explanatory"!

- What is the unit of measure on the Y axis?
- Are you using a log scale? (If so, mention it)
- Are the numbers shown in K's, M's or B's?
- If Currency, which currency?
- Net Revenue or Gross Revenue?
- How long is the rolling window being shown?
- How long does each bar represent?

🛑 Real Time Revenue

✅ Real Time Gross Revenue ($) last 1 hour in 4 minute buckets

Real-time Revenue

Apr 14, 2016 12:04 AM
Projected_Rev: 345.21
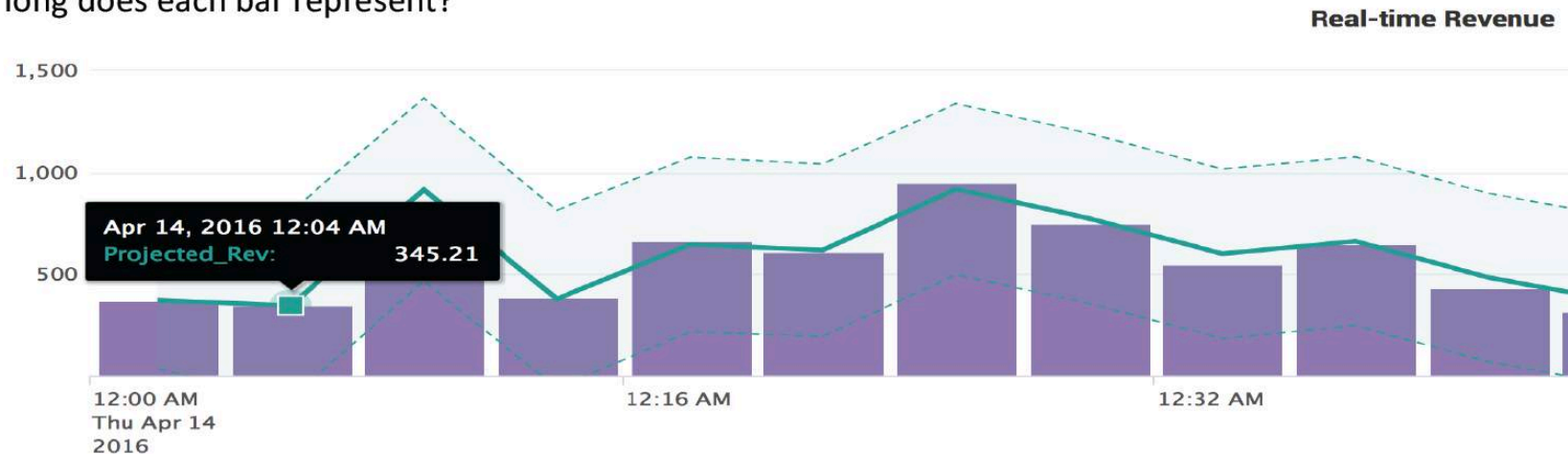
Performance: reduce load times

- Limit the number of objects on a single screen
- Limit real-time searches
- Specify filters to reduce the data
- Look for long-running searches and schedule them in advance (e.g. do not load twitter data at at query time
- Use **summary indexes** to reduce the search load of the dashboard
- Look for commonality between searches. When found, combine several searches into one
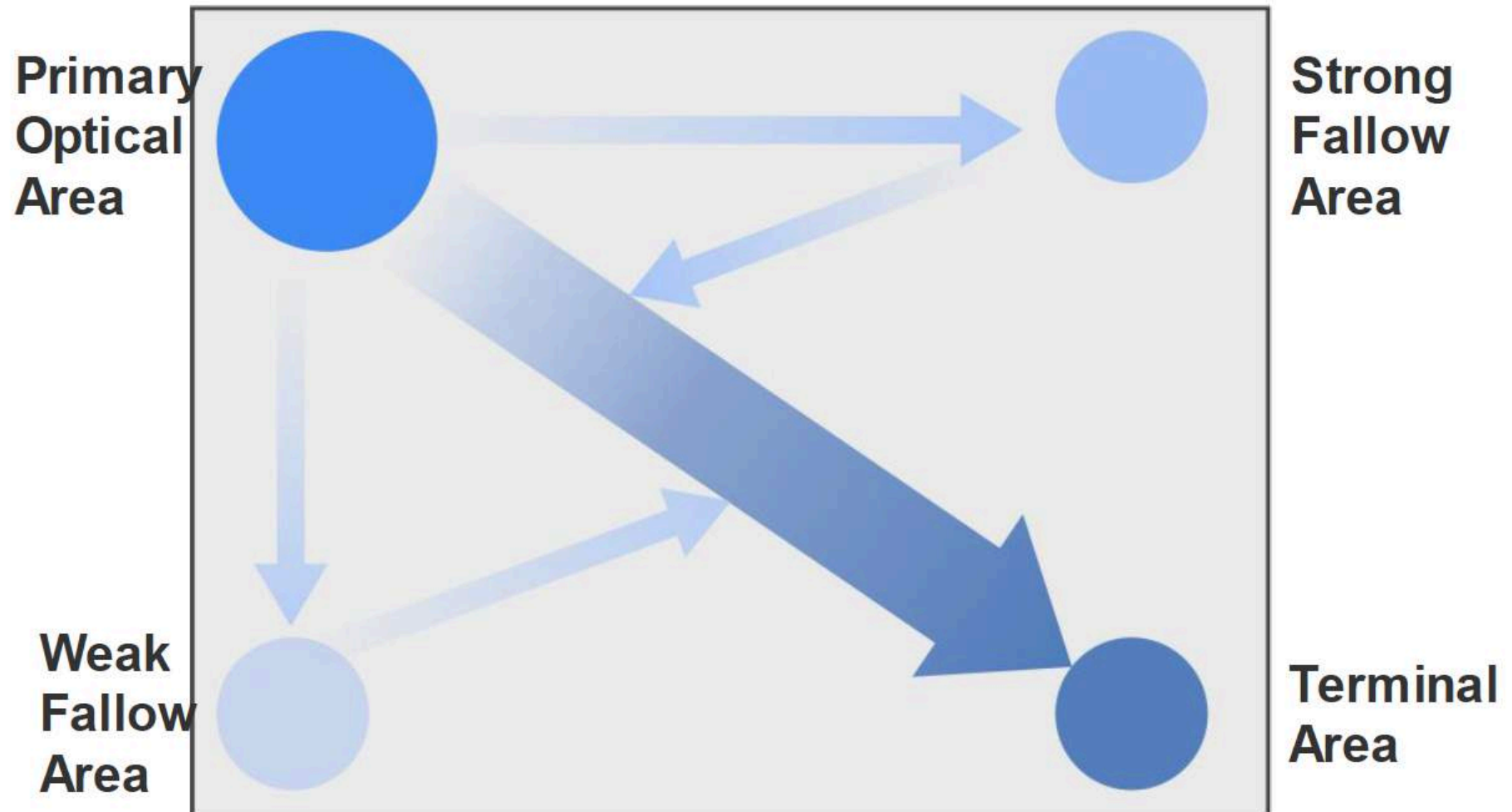
Wait Time
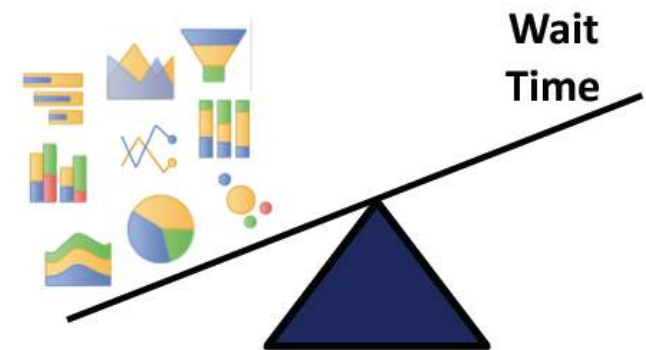
Layout: size to the right width

Orientation

- Landscape mode on a mobile will require scrolling to see content
- Are your key metrics at the top?
- Sizes to the width of the app
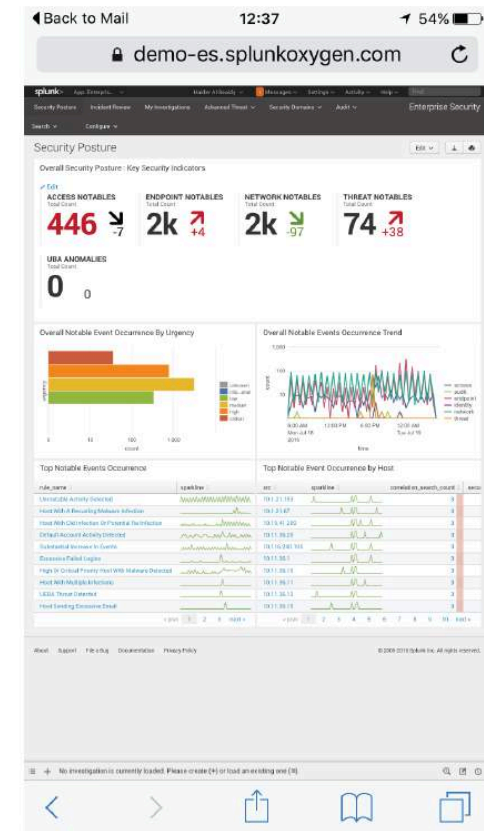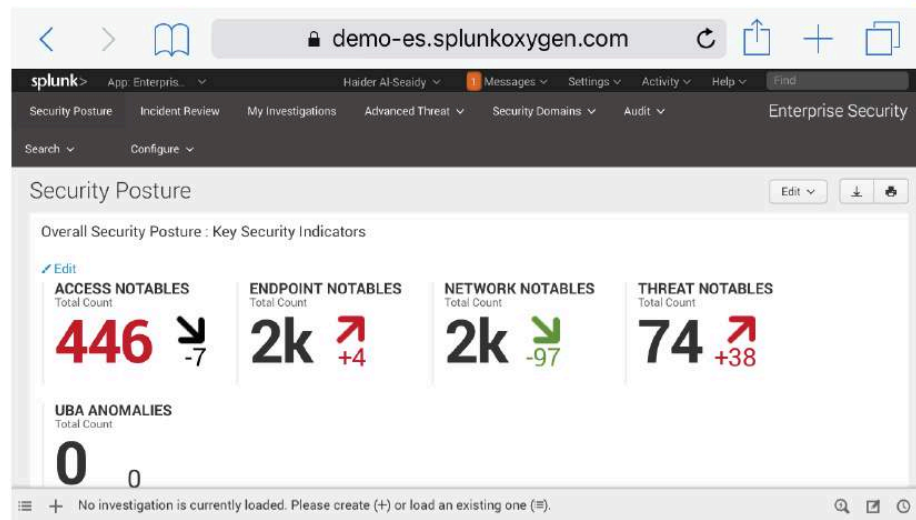
Layout:
People Have a Bias in How They Read and Scan Content

Layout: place most relevant content in primary optical area

Layout: present content hierarchically
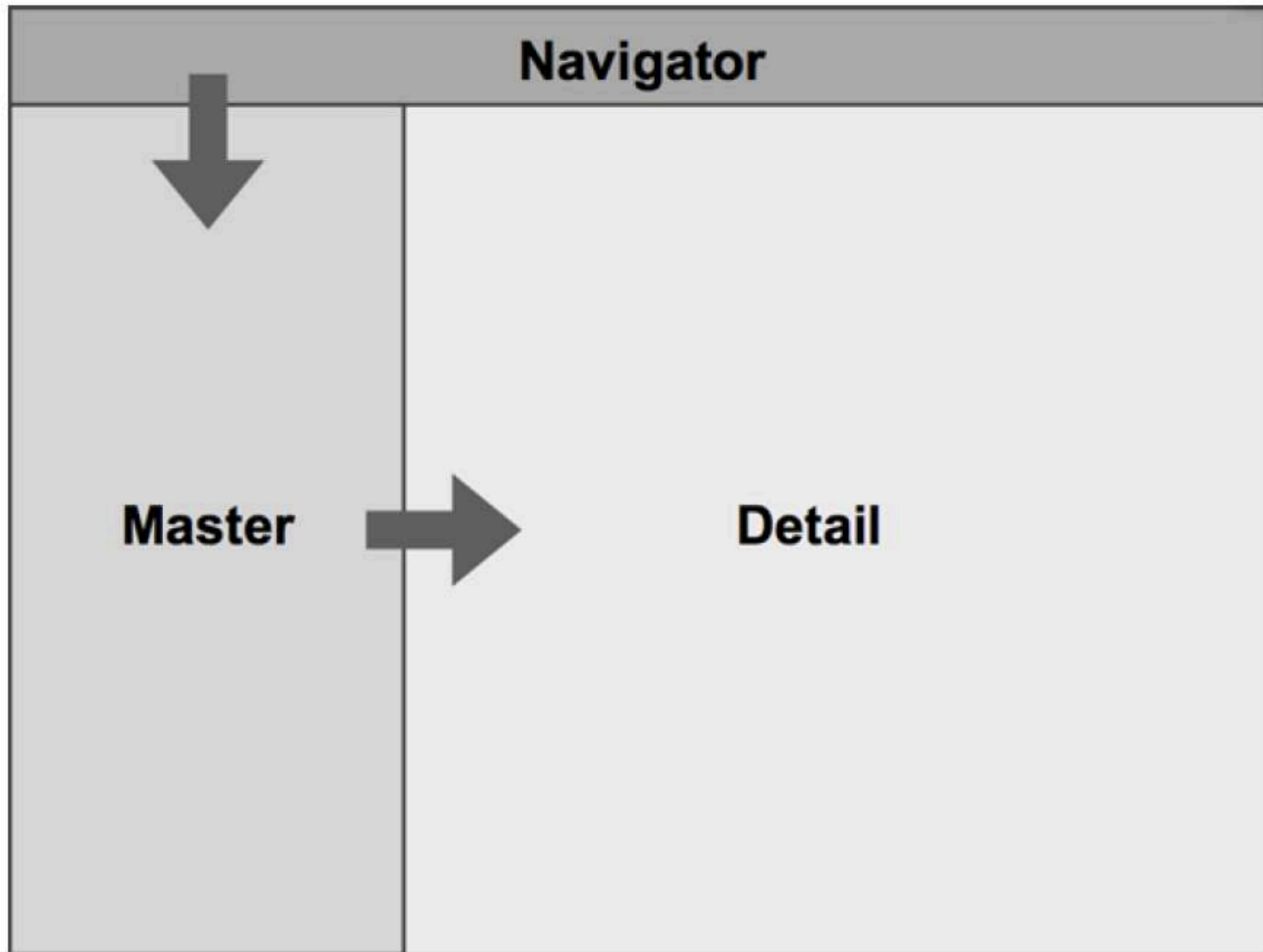
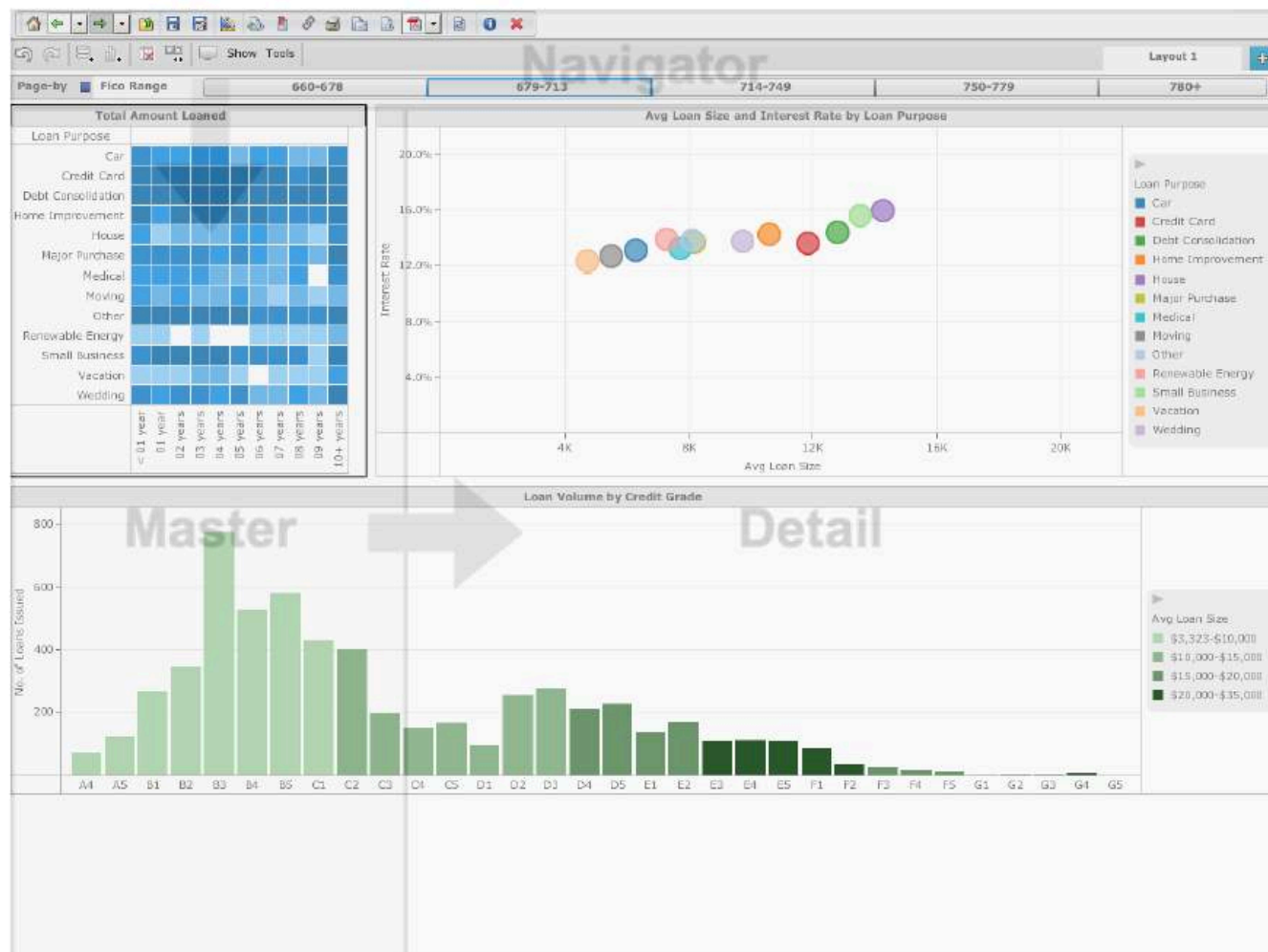# Layout: present content hierarchically

Interactivity & animation

- Interactivity and animation are the latest and coolest features for presenting information
- User can adapt a visualization to his/her own needs and curiosity, interacting with the map
- Interactive maps – when well designed – greatly improve the efficacy of an interaction
- Some very cool visualizations: (see more onhttp://blog.visme.co/best-information-graphics-2016/

The New York Times' project on where people born in a state move to

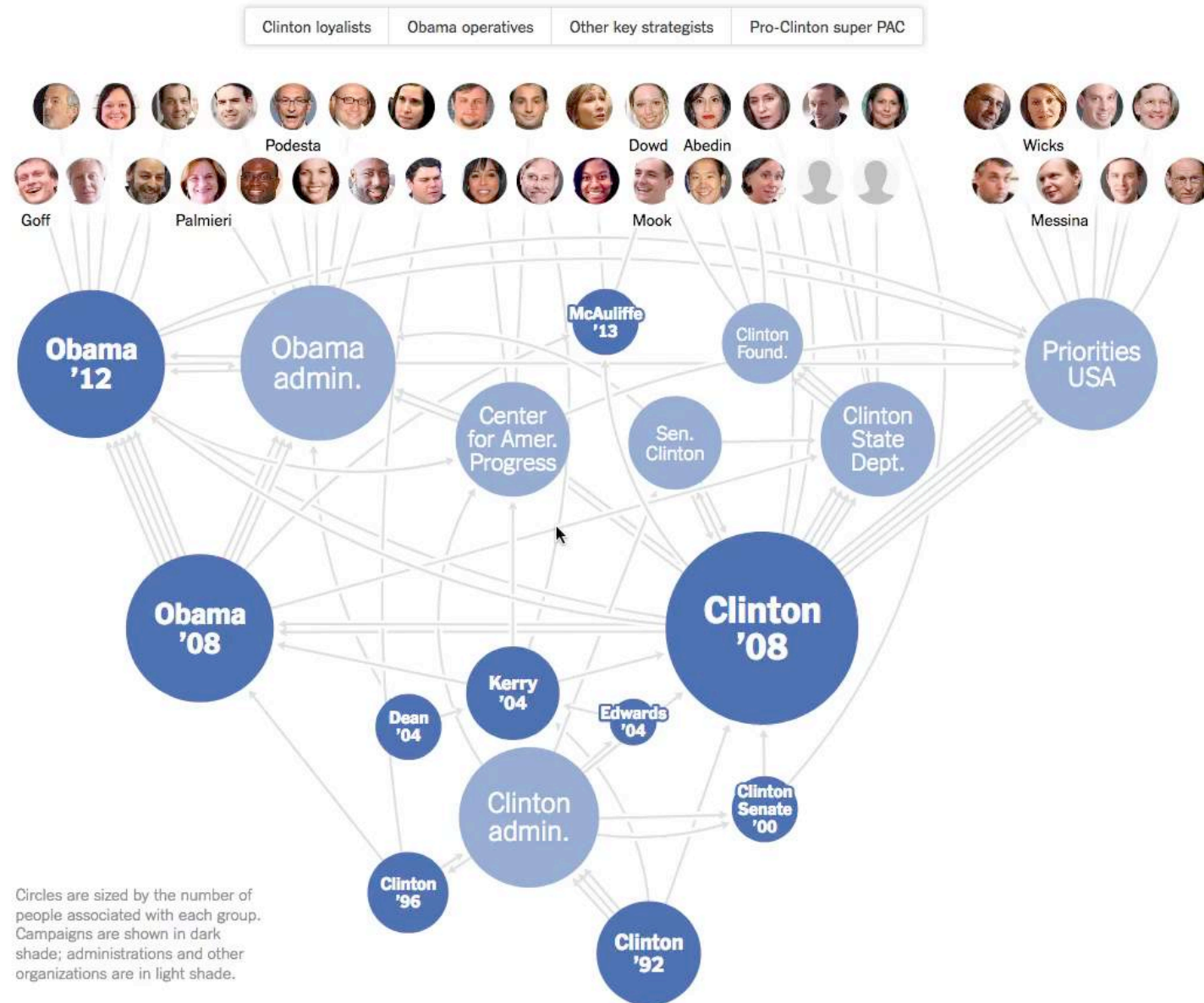Where people born in Washington, D.C. **have moved to:**

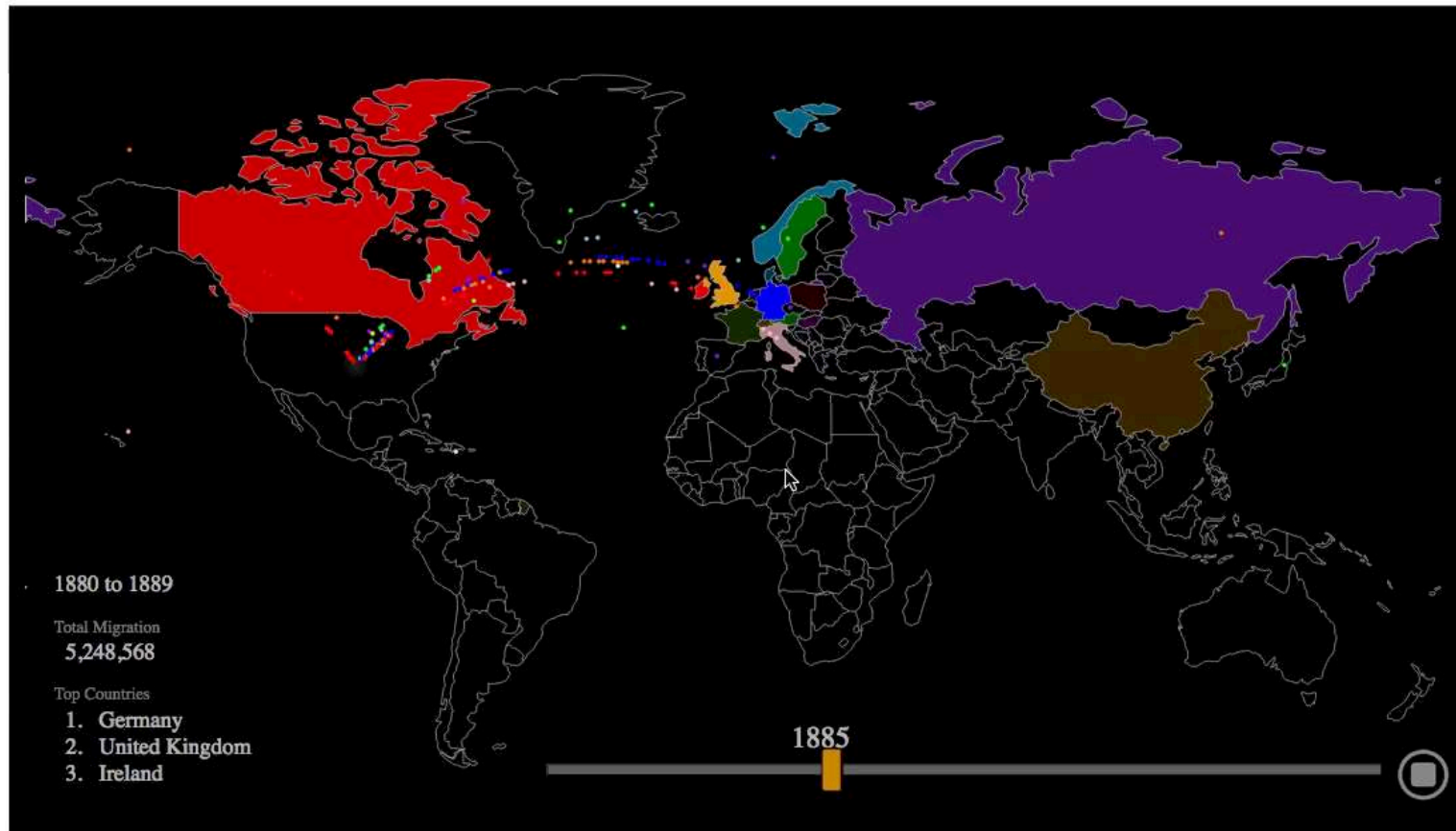New! ○ Switch to Migration into Washington, D.C.
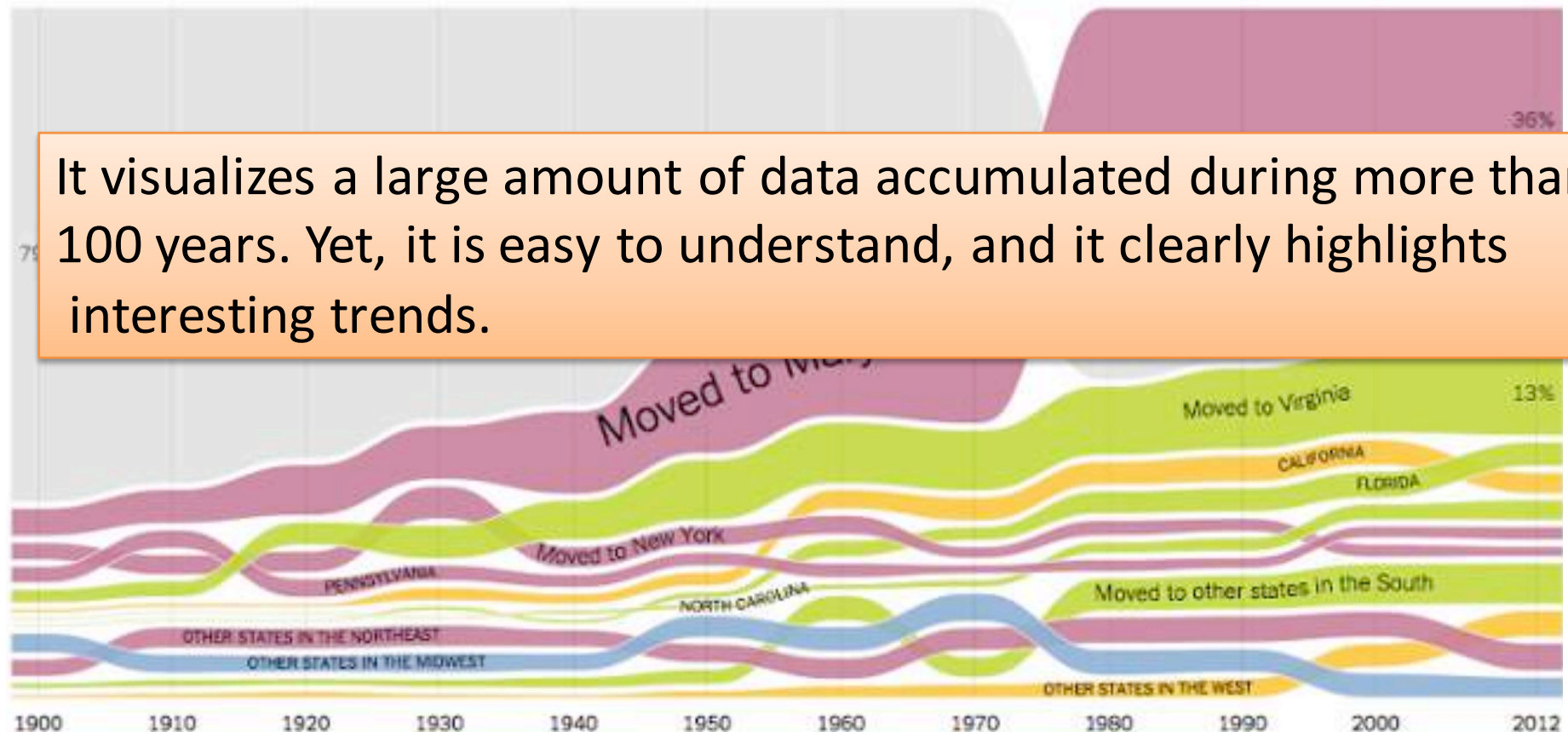
It visualizes a large amount of data accumulated during more than 100 years. Yet, it is easy to understand, and it clearly highlights interesting trends.

36%

Moved to Maryland

Moved to Virginia   13%

CALIFORNIA

FLORIDA

PENNSYLVANIA

Moved to New York

NORTH CAROLINA

Moved to other states in the South

OTHER STATES IN THE NORTHEAST

OTHER STATES IN THE MIDWEST

OTHER STATES IN THE WEST

1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2012

Politicians & Political campaigns

Immigration

Two Centuries of U.S. Immigration (1 dot = 10,000 people)

1880 to 1889

Total Migration
5,248,568

Top Countries
1. Germany
2. United Kingdom
3. Ireland

1885

Full screen interactive map / HD video

Lissen to Wikipedia
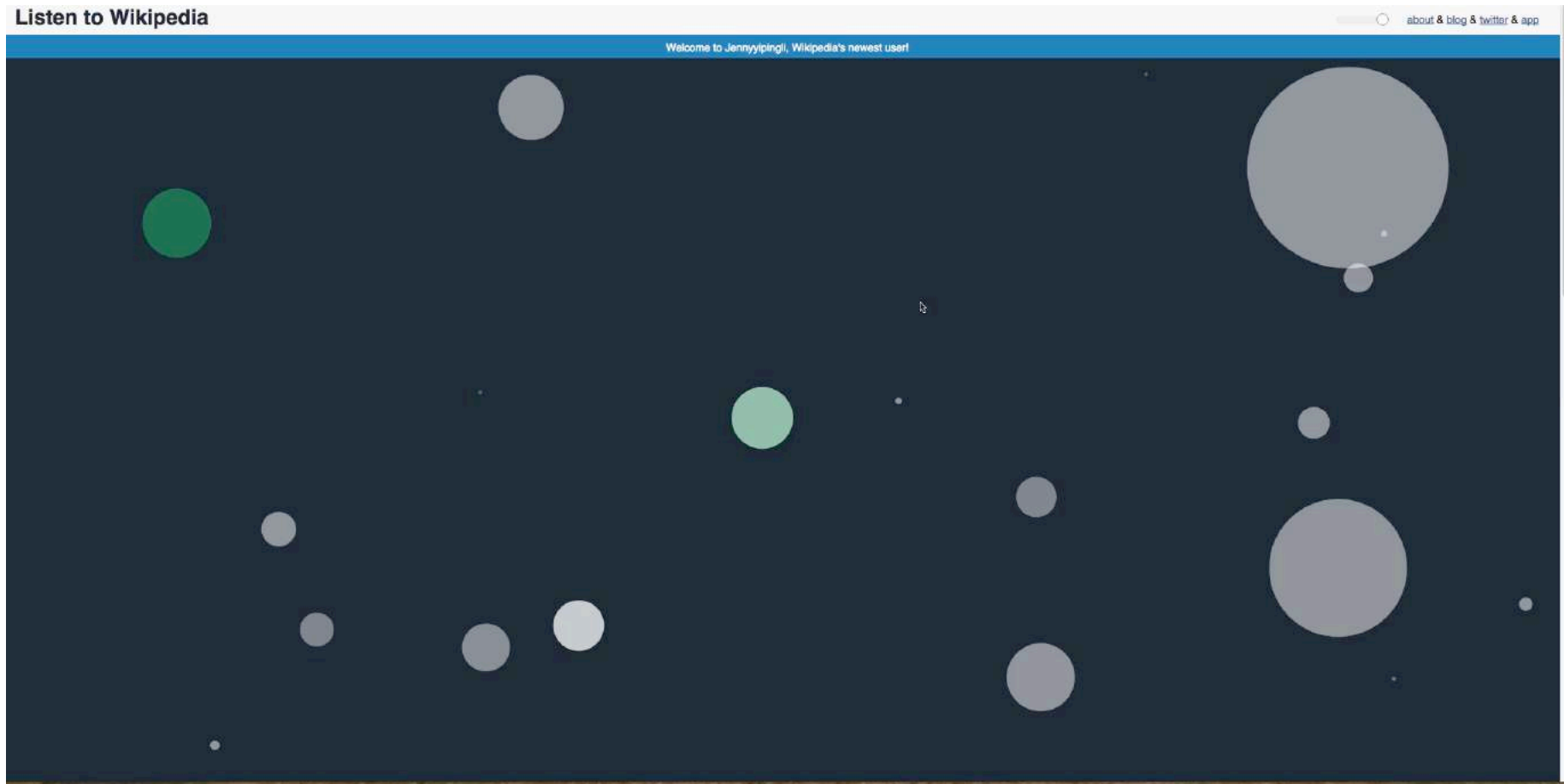
International Trades

Interactive: Mapping the Flow of International Trade

Homework

- Search on the web for fancy and new visualizations

- Can you imagine how a given visualization that you found can be useful for some business application?

- In your final project, when designing your report with Watson, pay attention to indications provided in this presentation when designing the report layout and visualizations.