

# INTRODUCTION

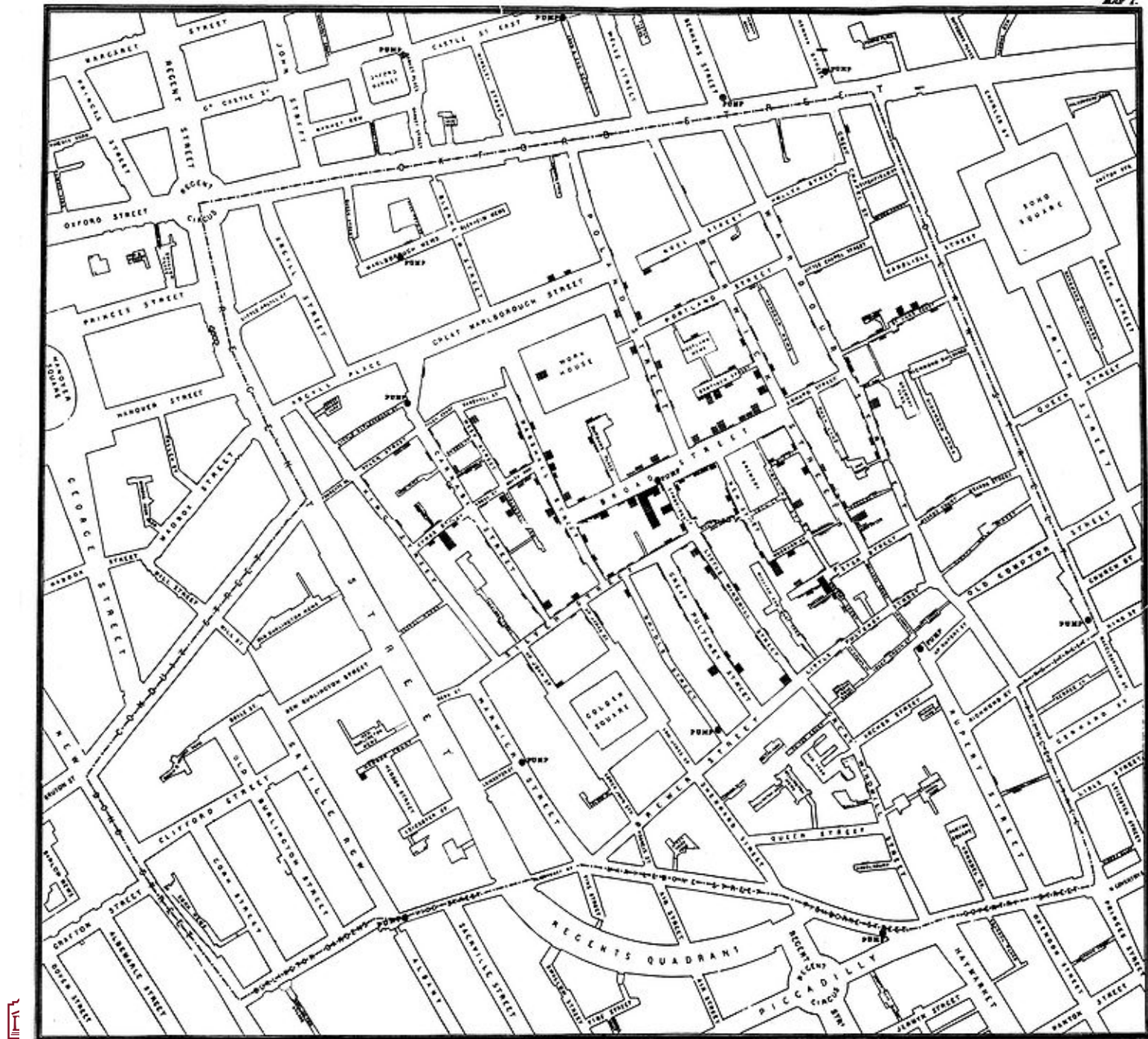
---

“Discovery is no longer limited by the collection and processing of data, but rather management, analysis, and visualization.”

@DamianMingle

## London's Cholera Outbreak of 1854

Physician John Snow links the outbreak to a contaminated well by plotting number of cases on a map, jump-starting the science of epidemiology.



*Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.*  
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite. Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. — Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Chiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre. Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

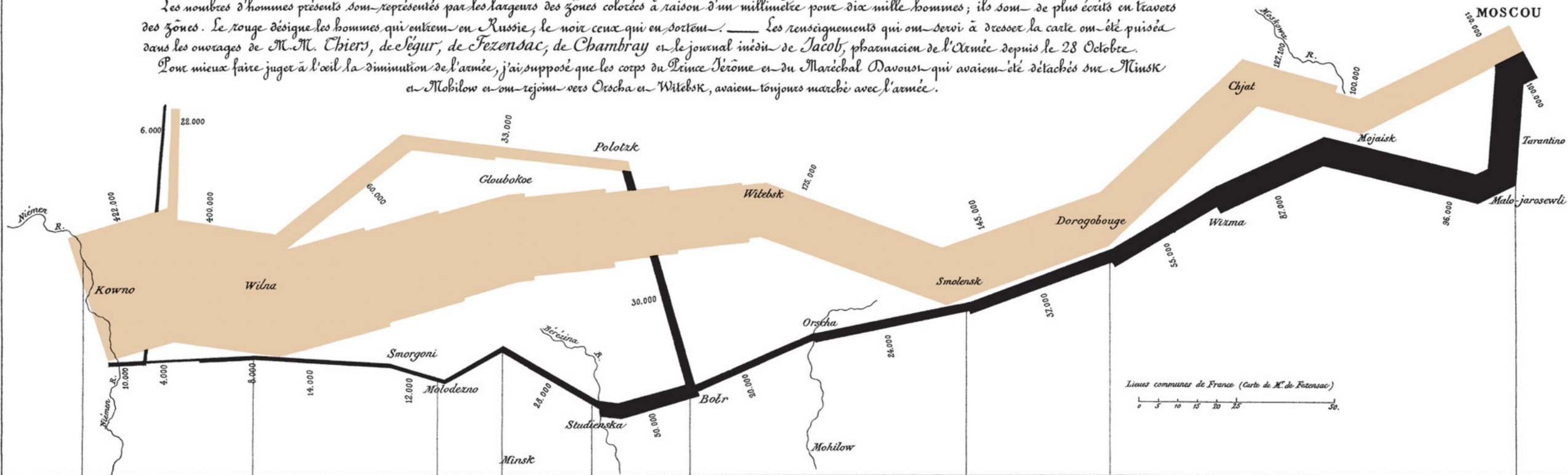
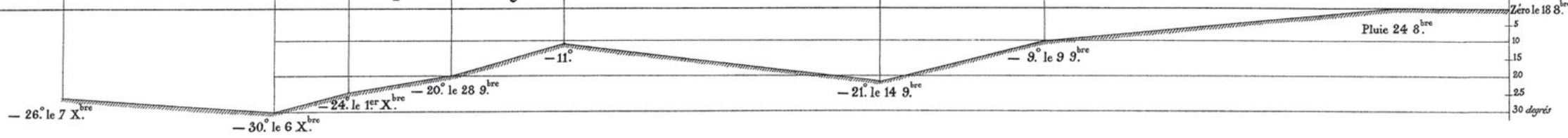


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niemen, gelé.



Minard's March to Moscow

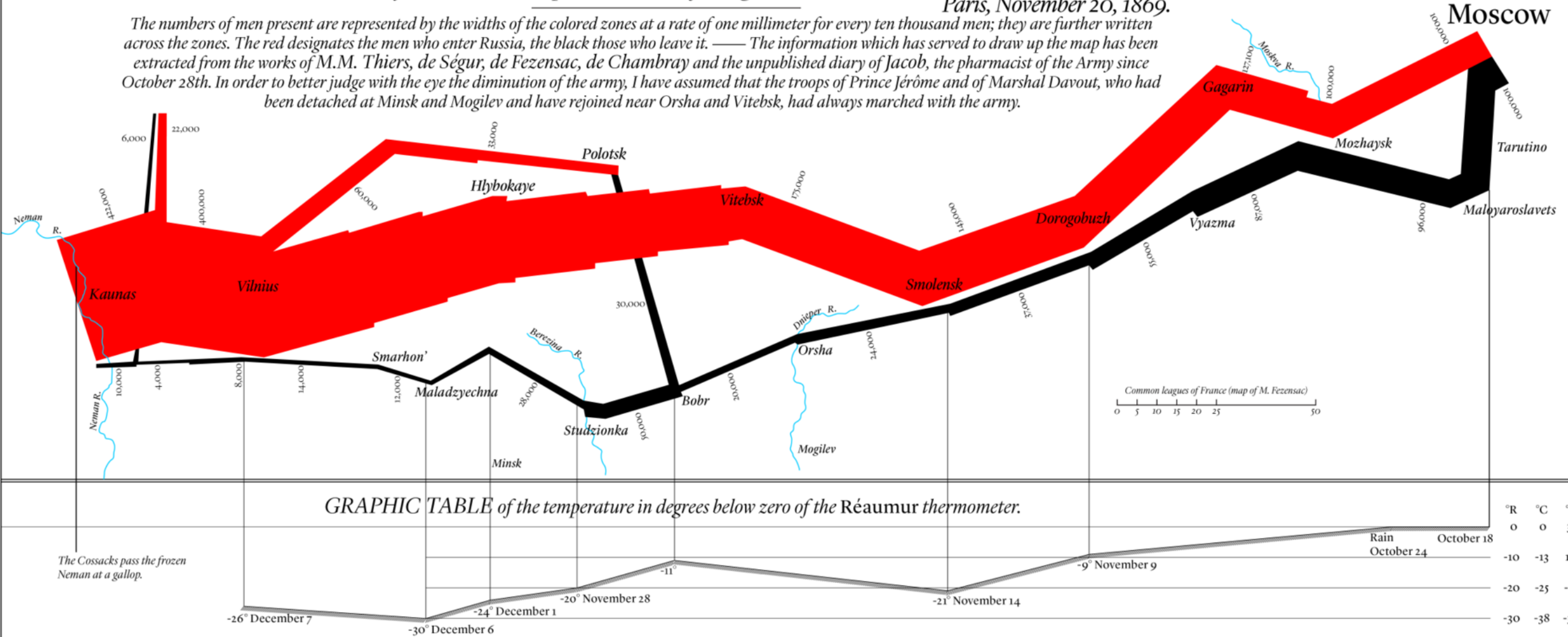


# Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812 ~ 1813

Drawn by M. Minard, Inspector General of Bridges and Roads (retired).

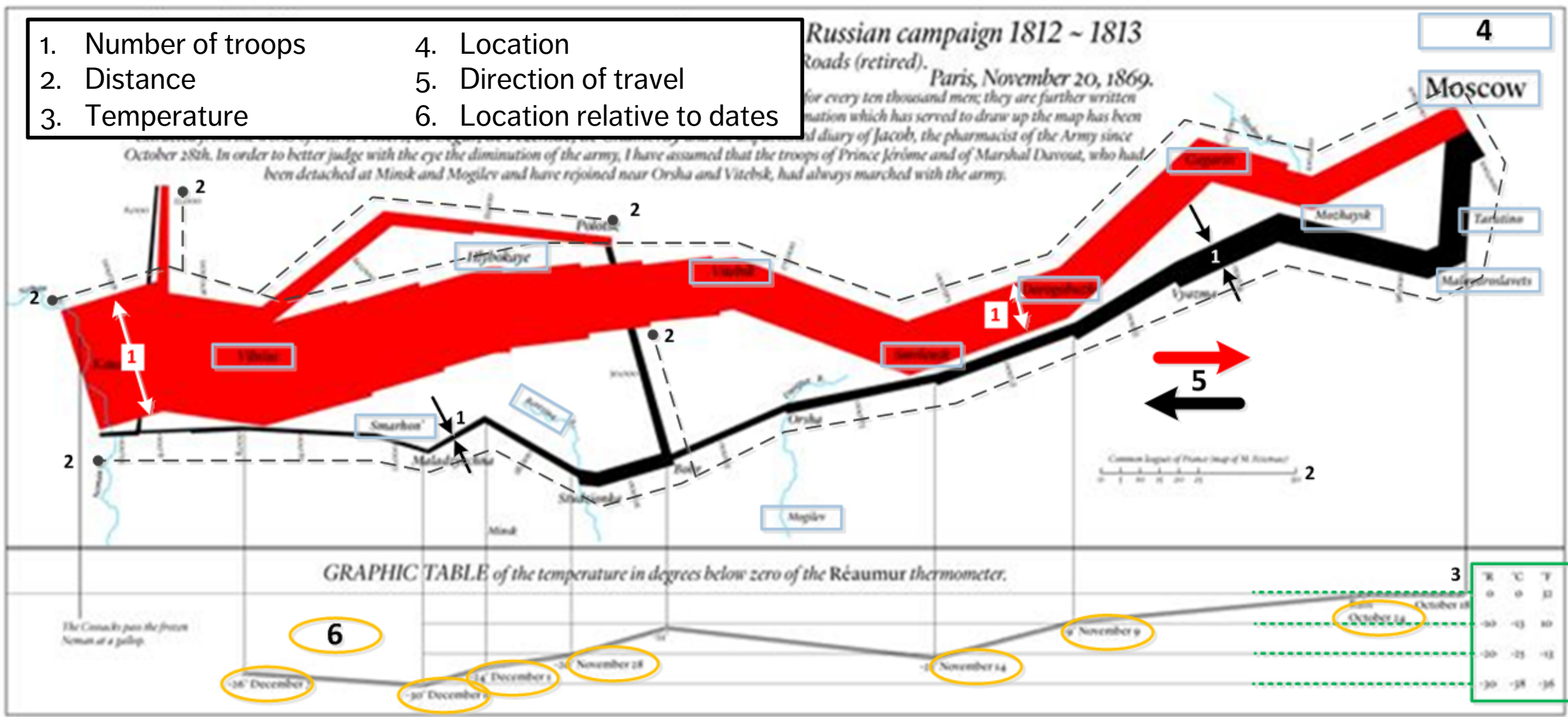
Paris, November 20, 1869.

The numbers of men present are represented by the widths of the colored zones at a rate of one millimeter for every ten thousand men; they are further written across the zones. The red designates the men who enter Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M.M. Thiers, de Ségur, de Fezensac, de Chambray and the unpublished diary of Jacob, the pharmacist of the Army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of Prince Jérôme and of Marshal Davout, who had been detached at Minsk and Mogilev and have rejoined near Orsha and Vitebsk, had always marched with the army.



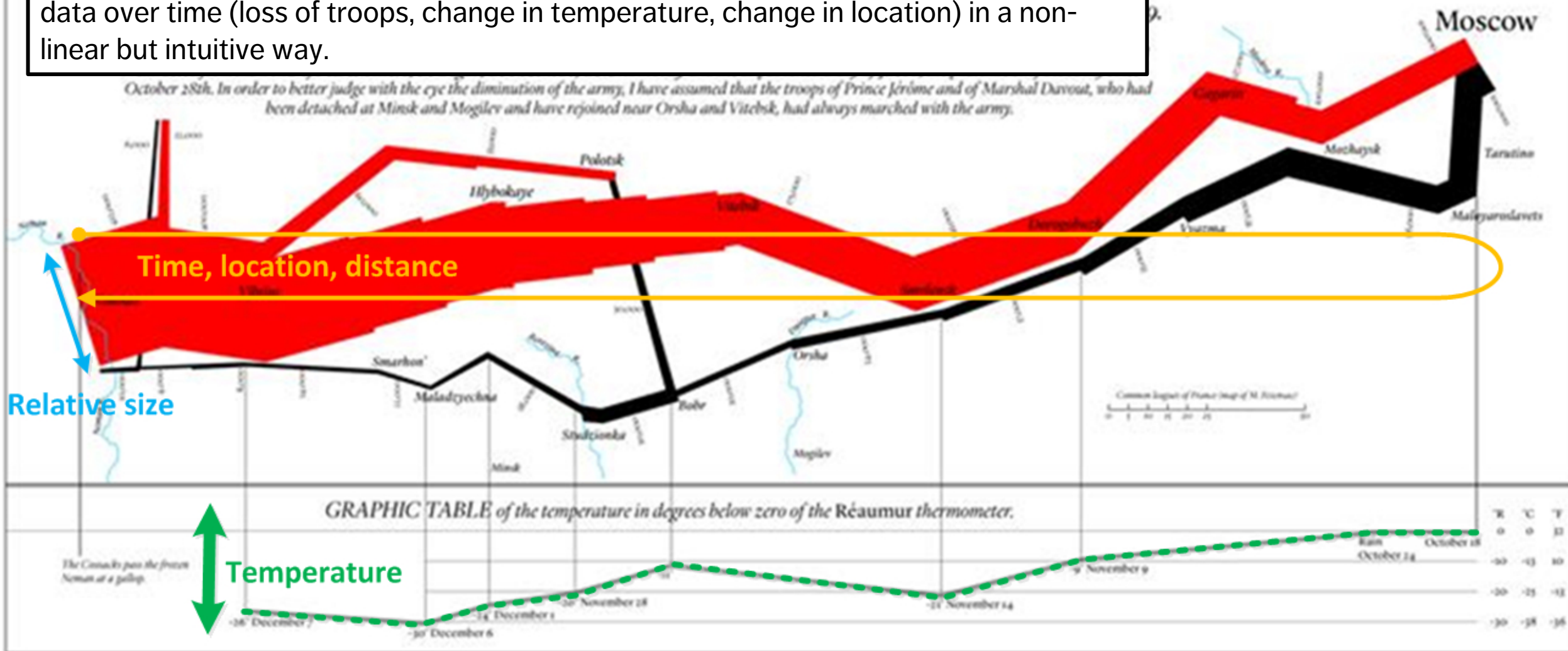
## Minard's March to Moscow

1. Number of troops
2. Distance
3. Temperature
4. Location
5. Direction of travel
6. Location relative to dates



Minard's March to Moscow

**Change over time without a linear time axis** – the ability to represent the *change* in data over time (loss of troops, change in temperature, change in location) in a non-linear but intuitive way.



Minard's March to Moscow



# INFOGRAPHICS

Created for **story-telling** purposes (**subjective**)

Intended for a **specific** audience

**Self-contained** and discrete

Graphic design aspect is key

Cannot usually be re-used with other data

Can incorporate **unquantifiable** information





# INFOGRAPHICS

Created for **story-telling** purposes (**subjective**)

Intended for a **specific** audience

**Self-contained** and discrete

Graphic design aspect is key

Cannot usually be re-used with other data

Can incorporate **unquantifiable** information



# DATA VISUALIZATION

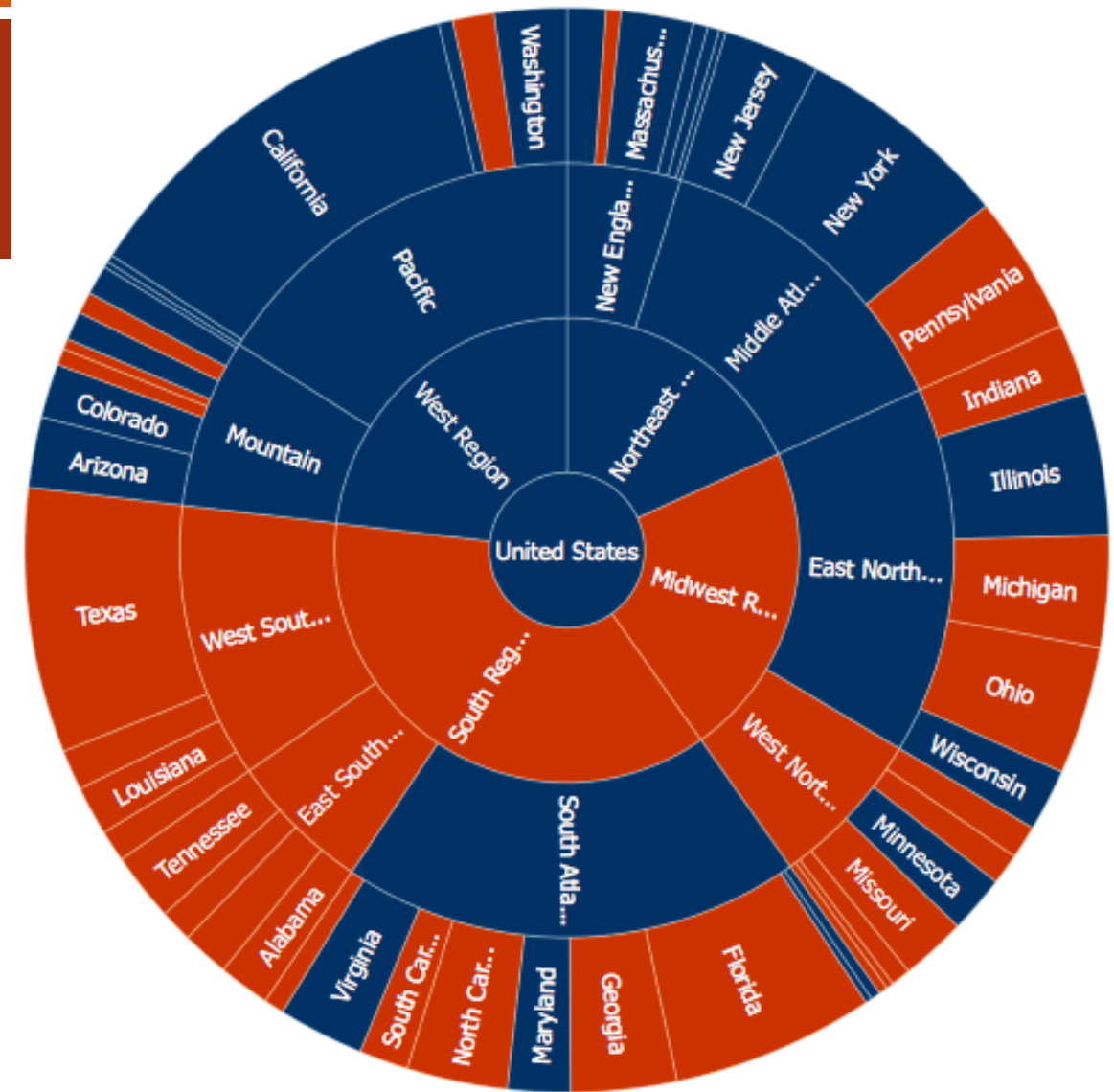
A **method**, as well as an item (**objective**)

Typically focuses on the **quantifiable**

Used to make sense of the data or to make it **accessible** (datasets can be massive and unwieldy)

May be generated automatically

The look and feel are less important than the **insights conveyed** by the data



Size Population Color Median Household Income

Low Income High Income

# DATA VISUALIZATION

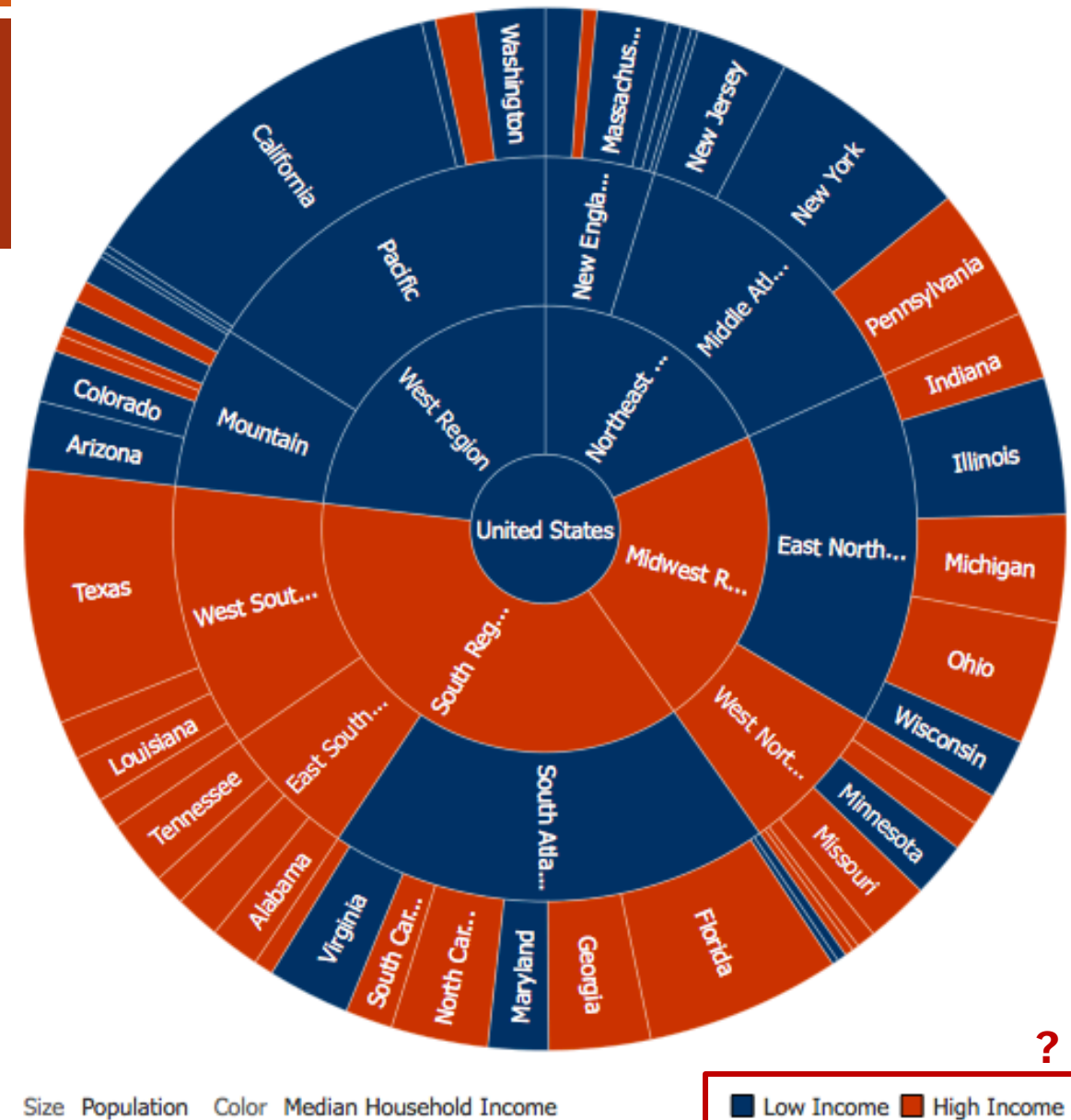
A **method**, as well as an item (**objective**)

Typically focuses on the **quantifiable**

Used to make sense of the data or to make it **accessible** (datasets can be massive and unwieldy)

May be generated automatically

The look and feel are less important than the **insights conveyed** by the data



# DATA UP TO THE 20<sup>TH</sup> CENTURY

In the 20<sup>th</sup> century, data problems were mostly related to

- **engineering** (design of machines)
- **sciences** (formulation of theories)

Problems were solved **empirically**, **theoretically**, or through **computation**.



# DATA UP TO THE 20<sup>TH</sup> CENTURY

Engineers equipped machines with sensors  $\Rightarrow$  used data to assess if the machines behaved as expected & to improve designs.

Scientists set up experiments  $\Rightarrow$  used data to test the validity of theories.

- Experiments are expensive; relatively few data points are generated.

Data contained additional information which is often ignored.

- Example: Mendel's experimental data, analyzed by Fisher, found to be too good to be true.

# DATA IN THE 21<sup>ST</sup> CENTURY

In the 21<sup>st</sup> century, there is:

- there is **more data**
- it's mostly **digital**
- it's mostly **observed** (rather than generated by designed experiment)

Problems are solved **empirically, theoretically**, through **computation** and/or **data exploration/visualization**.

# DATA IN THE 21<sup>ST</sup> CENTURY

**Empirically:** observe and describe what happens

**Theoretically:** generalize and build models and generalizations to understand what happens

**Computationally:** design computer simulations to better understand what happens

**Data Exploration/Visualization:** the new approach to understanding

## EXERCISE

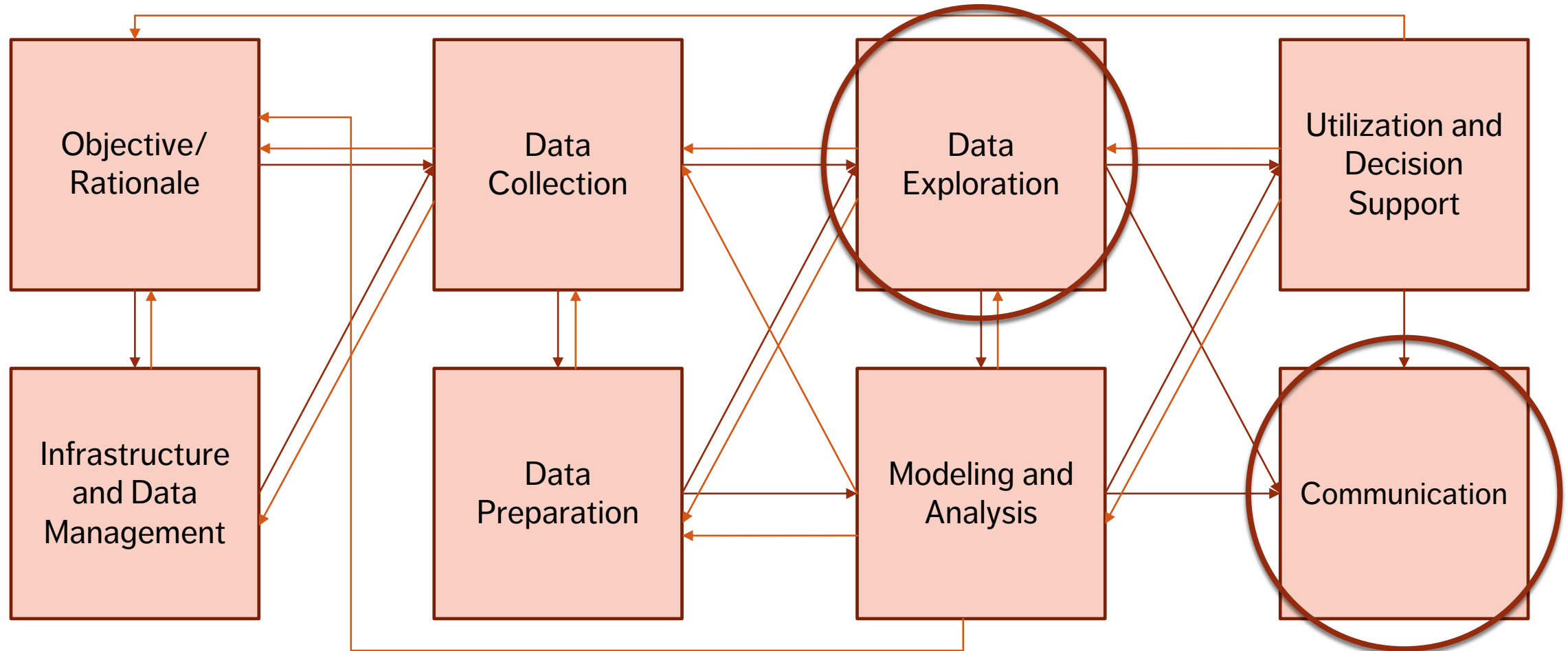
In teams or individually, identify a few data visualizations that appeal to you (professionally, esthetically, or both).

What is the story being told by the visualization?

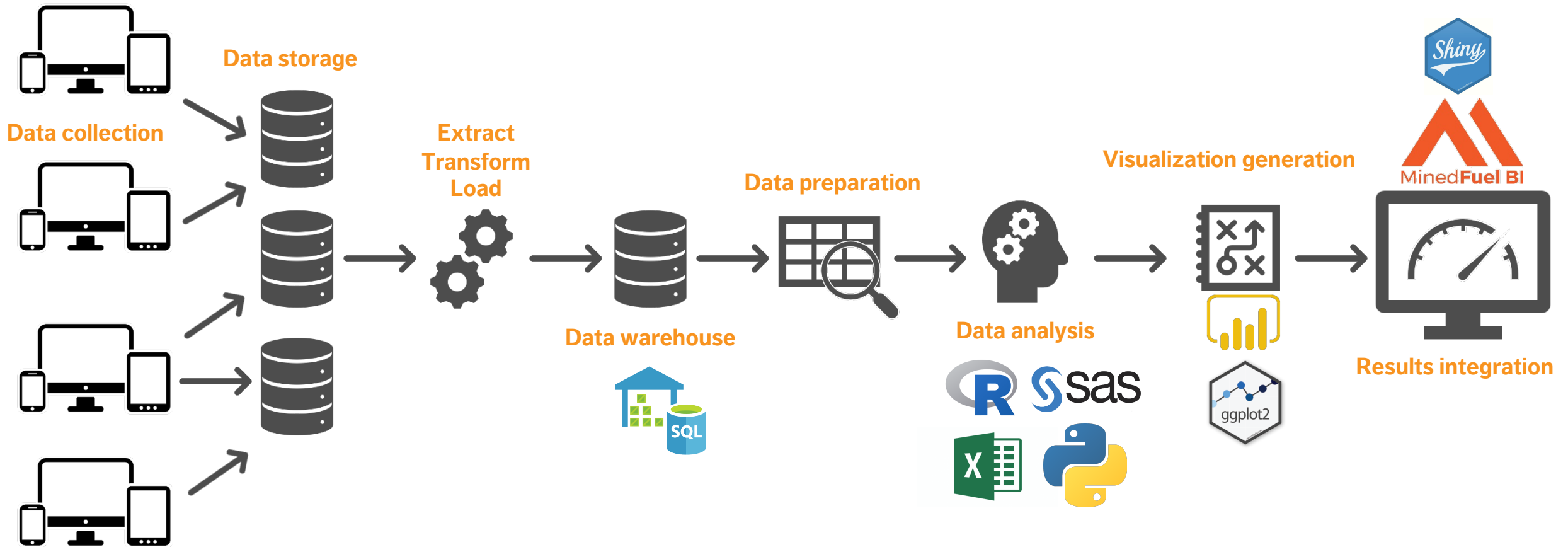
What kind of data is needed to build these visualizations?



# THE (MESSY) ANALYSIS PROCESS



# DATA ENVIRONMENT



Visualizations account for only ~10% of the process.

## EXERCISE

In teams or individually, identify work scenarios for which data visualization could prove useful.

What insight could be drawn from such visualizations?

Would such visualizations get a buy-in from your supervisors/employers?

How much work would be required to get from design to completion? Are the obstacles mostly of a technical nature? Related to data procurement?

# VISUALIZATION AND DATA EXPLORATION



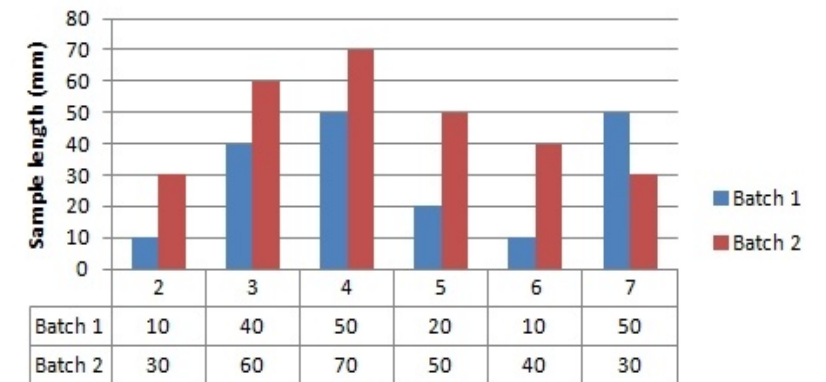
# OVERVIEW

## The past is **data-driven**:

- mostly Excel (or reporting tools like Cognos)
- mostly numbers, tables and non-interactive graphs
- distributed on desktop computers, by email, in PowerPoint presentation
- static, mostly backwards looking (lagging indicators)
- KPIs and dashboards were somewhat contrived

North Region Unit Sales by City July 2006

Region	Jan-06	Feb-06	Mar-06	Apr-06	May-06	Jun-06	Jul-06
<b>Actuals</b>							
Seattle	111	653	1,598	3,411	3,972	5,092	5,291
Boise	26,779	27,867	29,153	30,557	33,402	35,400	35,451
Portland	33,078	34,401	37,535	39,916	41,357	45,306	46,671
Spokane	25,417	26,669	28,092	29,020	29,674	30,501	30,831
North Region	199,841	211,053	226,789	242,957	256,605	273,640	277,771
<b>Plan</b>							
Seattle	693	468	790	1,383	2,205	3,180	4,211
Boise	29,525	26,062	27,088	28,269	29,536	30,821	32,161
Portland	32,276	34,708	36,737	38,857	41,066	43,364	45,751
Spokane	30,500	26,644	27,987	29,430	30,994	32,594	34,231
North Region	191,783	203,916	216,524	230,474	246,390	263,378	281,221
<b>Variance</b>							
Seattle	-582	185	808	2,029	1,767	1,912	1,071
Boise	-2,746	1,805	2,064	2,288	3,866	4,578	3,281
Portland	802	-307	798	1,059	291	1,942	921
Spokane	-5,082	25	105	-410	-1,320	-2,093	-3,391
North Region	-8,057	7,137	10,365	12,483	10,215	10,261	-3,411



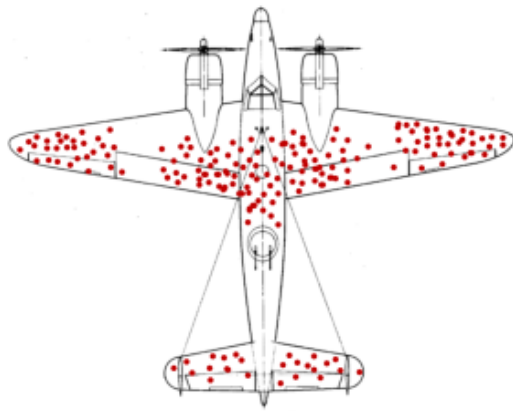
# OVERVIEW

The future is **story-driven**:

- new tools: Power BI, R, Qlickview etc.
- mostly visualizations, occasional numbers and tables
- distributed on the web (internal and external)
- dynamic and both backwards and forwards looking (leading and lagging indicators)
- data for everyone



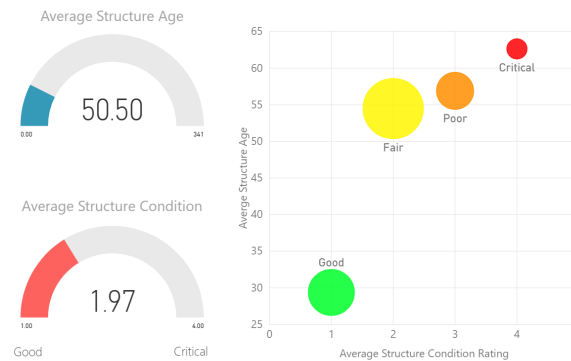
# DEFINING CONTEXT



Seconds

## Directory of Federal Real Property (DFRP) Dashboard

You have selected 20,186 properties that contain 35,148 structures



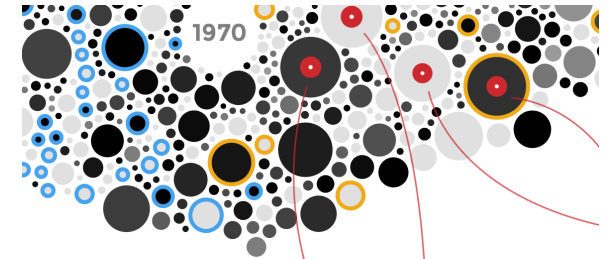
Minutes

## Access to Information and Privacy (ATIP) search

You have currently selected 28,711 requests totaling 6,597,612 pages of information



Fraction of Hour



## The Beatles

No other artist or band has more songs in the Top 2000 as the Beatles. With 38 songs they are responsible for 14% of all titles before 1970. Nonetheless, only 5 years ago they still had 50 songs in the list.

- 4 Piano Man  
Billy Joel | 1974
- 5 Child in Time  
Deep Purple | 1972

Hours

← Infographics/Data Viz →

← Dashboards →

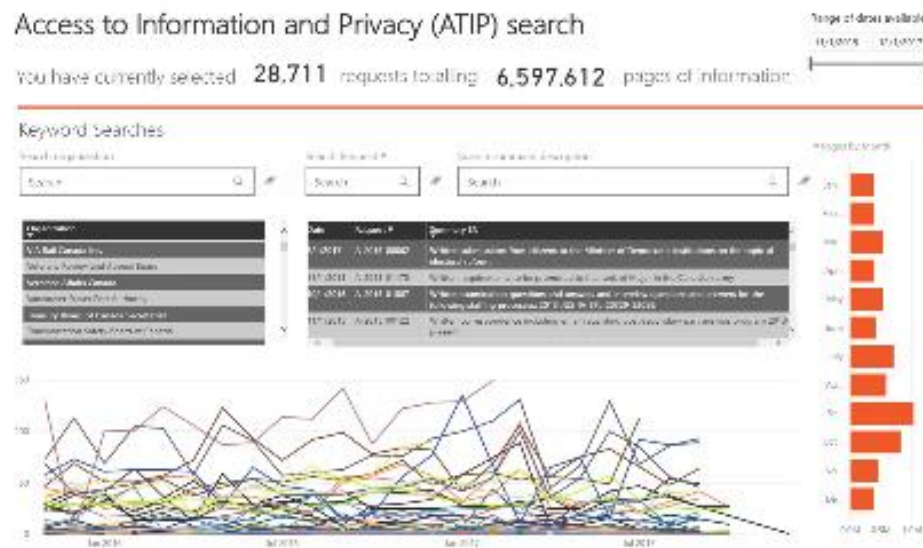
← Reports →

← Data Art →

# EXPLORATORY VS. EXPLANATORY ANALYSIS

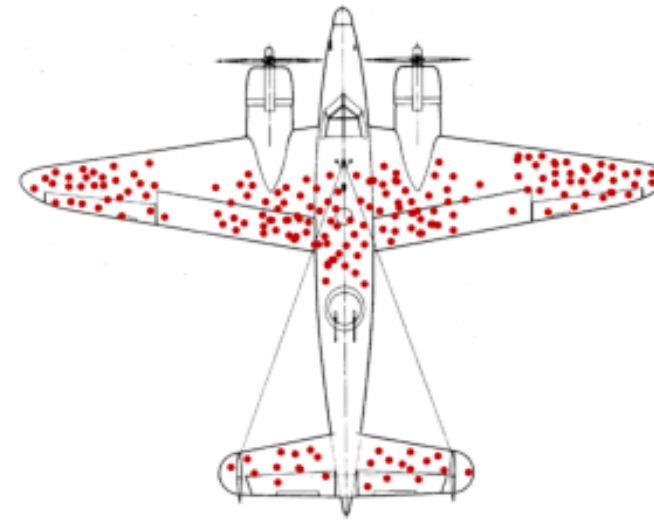
**Exploratory:** understanding the **DATA** (associated with reports)

**Explanatory:** communicating a **STORY** (associated with dashboards and data viz)



Exploratory

VS.



Explanatory

## SOME BASIC QUESTIONS

What system does your data represent – objects, attributes, relationships?

**How** does it represent this system – i.e. the data model?

Who made this dataset? When? For what purpose?

Assuming a flat file – what do the rows represent? What do the columns represent?

Do you even have enough information (e.g. **metadata**) to answer these questions?

Where can you find more information?

# NON-VISUALIZATION BASED SUMMARIES OF YOUR DATASET

Cl	N03	NH4
Min. : 0.222	Min. : 0.000	Min. : 5.00
1st Qu.: 10.994	1st Qu.: 1.147	1st Qu.: 37.86
Median : 32.470	Median : 2.356	Median : 107.36
Mean : 42.517	Mean : 3.121	Mean : 471.73
3rd Qu.: 57.750	3rd Qu.: 4.147	3rd Qu.: 244.90
Max. : 391.500	Max. : 45.650	Max. : 24064.00
NA's : 16	NA's : 2	NA's : 2

season  
Length: 340  
Class : character  
Mode : character

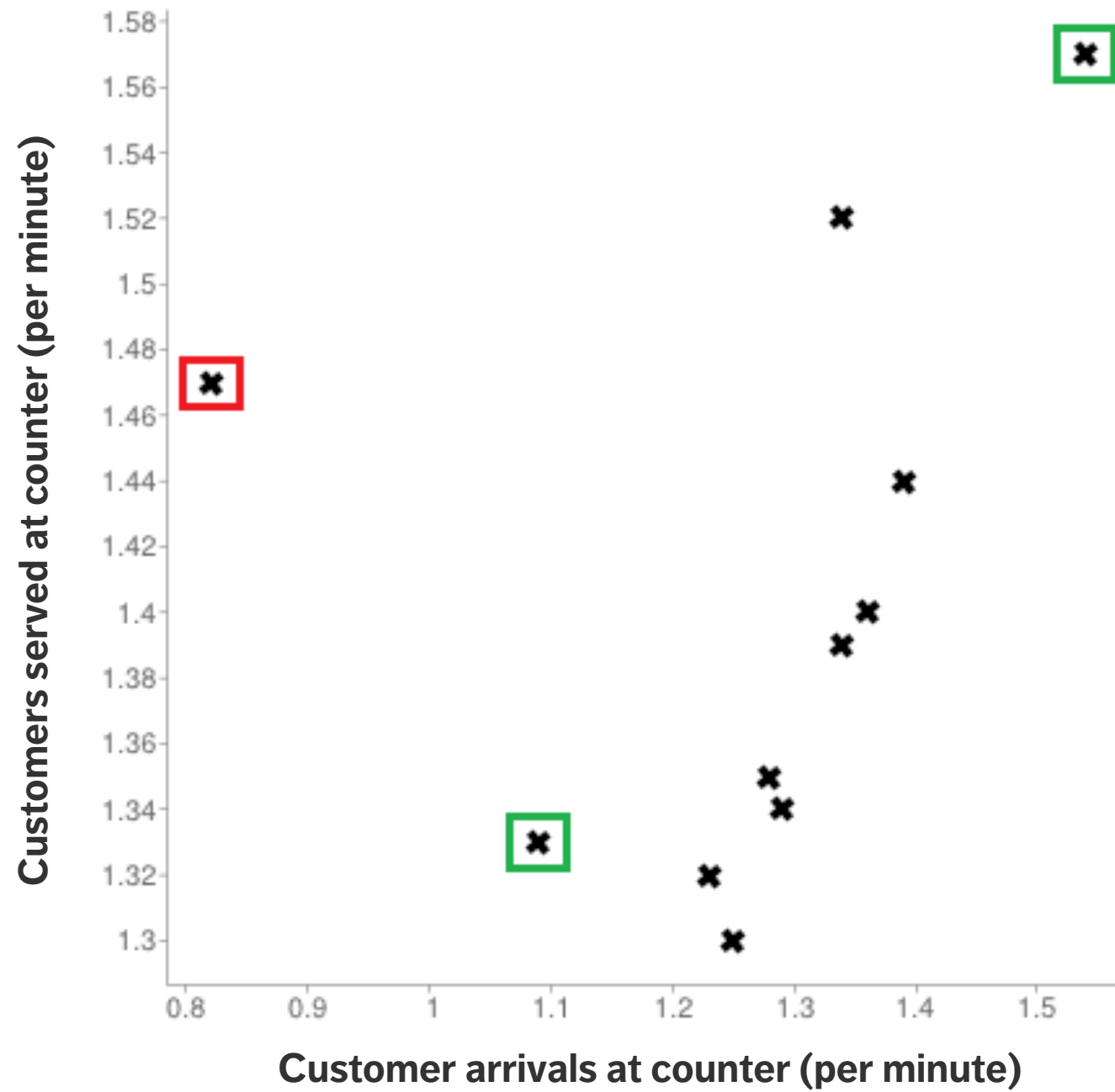
autumn spring summer winter  
80 84 86 90

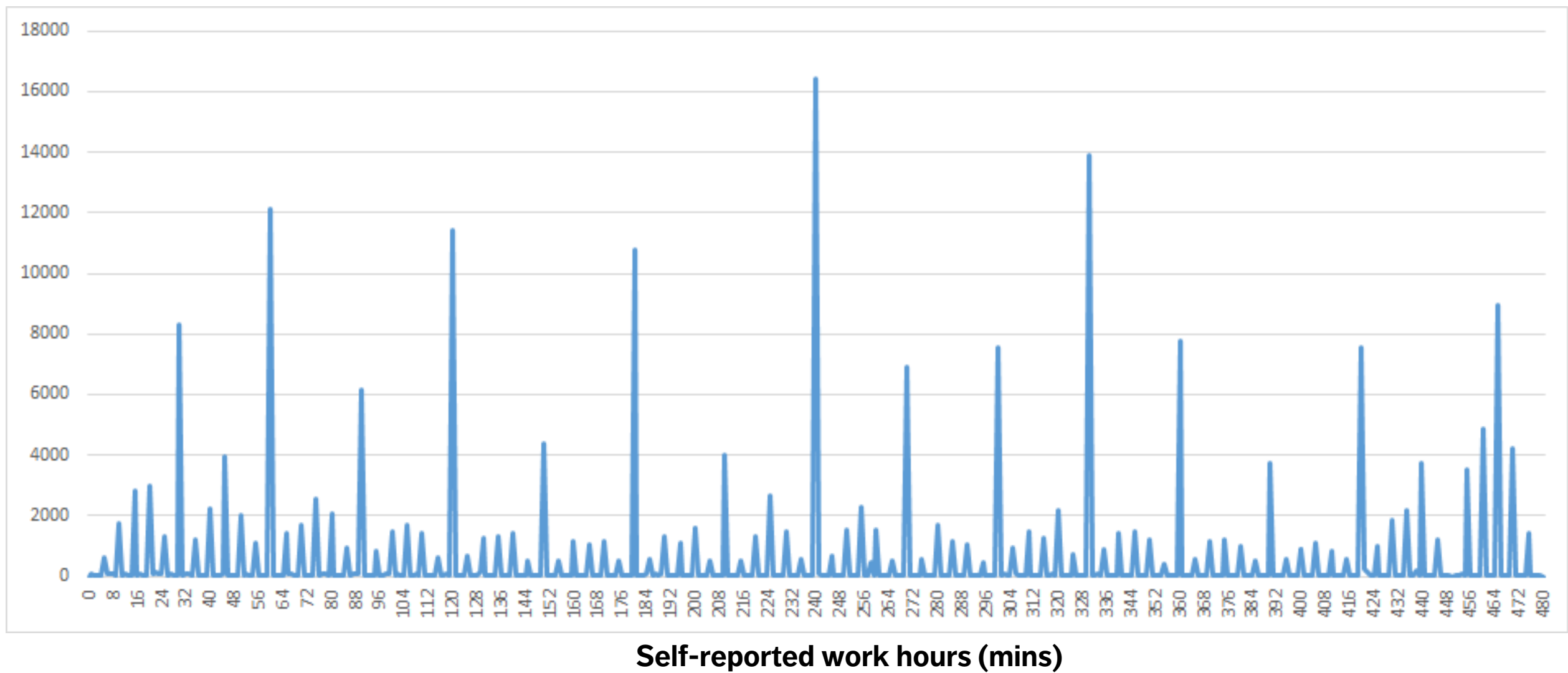


# PRE-ANALYSIS USE

Data visualization can be used to set the stage for analysis:

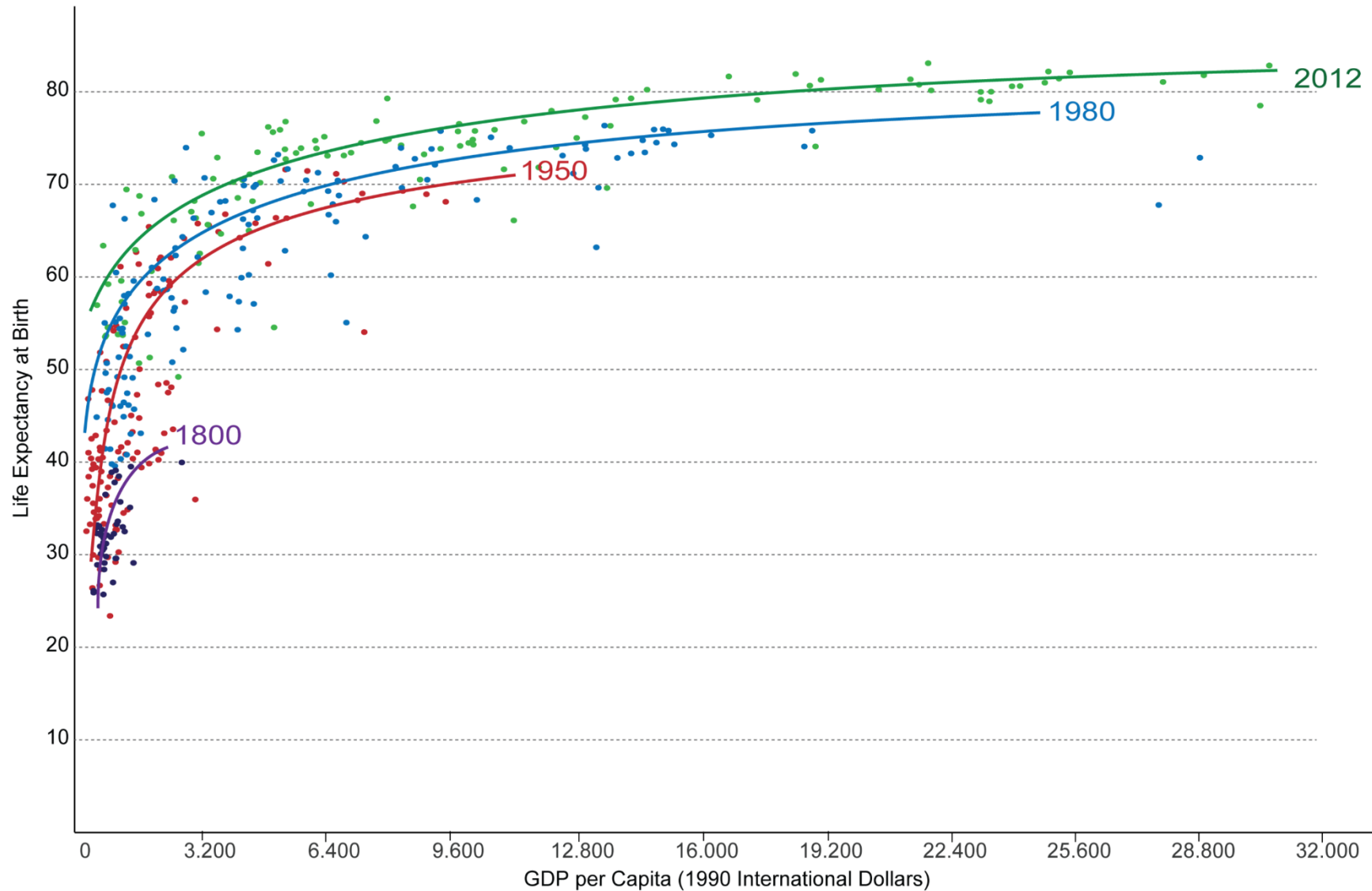
- **detecting anomalous entries**  
invalid entries, missing values, outliers
- **shaping the data transformations**  
binning, standardization, Box-Cox transformations, PCA-like transformations
- **getting a sense for the data**  
data analysis as an art form, exploratory analysis
- **identifying hidden data structure**  
clustering, associations, patterns informing the next stage of analysis





## Life Expectancy vs. GDP per Capita from 1800 to 2012 – by Max Roser

GDP per capita is measured in International Dollars. This is a currency that would buy a comparable amount of goods and services a U.S. dollar would buy in the United States in 1990. Therefore incomes are comparable across countries and across time.



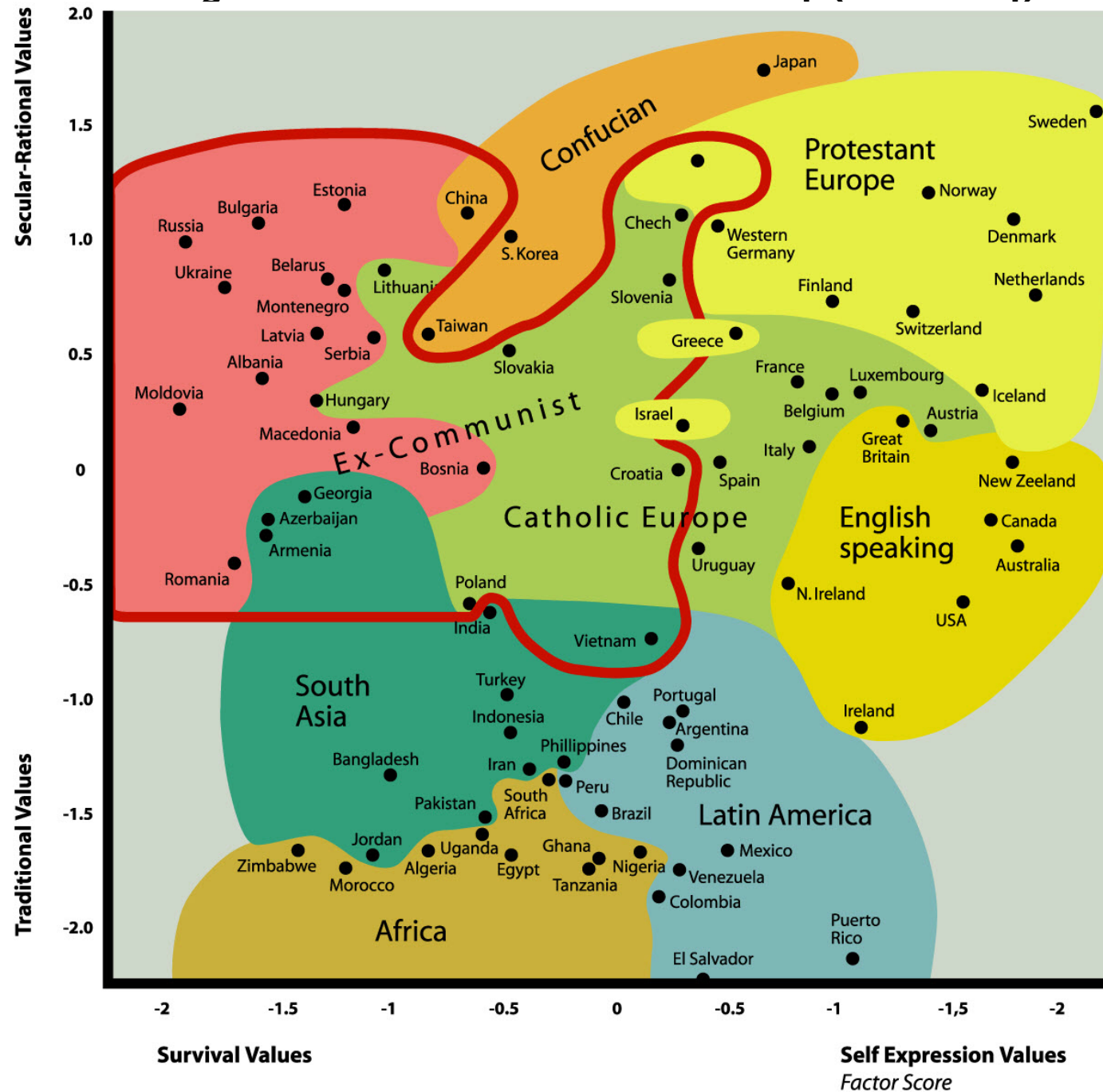
This graph displays the correlation between life expectancy and GDP per capita.

Countries with higher GDP have a higher life expectancy, in general.

The relationship seems to follow a logarithmic trend: the unit increase in life expectancy per unit increase in GDP decreases as GDP per capita increases.

# Inglehart-Welzel's Global Cultural Map (2010-2014)

[[https://en.wikipedia.org/wiki/World\\_Values\\_Survey](https://en.wikipedia.org/wiki/World_Values_Survey)]



## Traditional values

importance of religion, parent-child ties, deference to authority and traditional family values.

## Secular-rational values

less emphasis on religion, traditional family values and authority.

## Survival values

emphasis on economic and physical security.

## Self-expression values

high priority to environmental protection, growing tolerance of foreigners, gays and lesbians and gender equality

# DISCUSSION

Which of the pre-analysis uses of visualization is most relevant to your work?

# BASICS OF DASHBOARDING



# REPORTING AND DEPLOYMENT

An analysis can only be as good as how it is **communicated** and/or **deployed**.

## Crucial Questions:

- Who is in receipt of the report(s)?
- How are the workflows deployed into production?
- Can data insights be turned into useful policies?

Automatic reporting should be audited and validated **regularly**.

# REPORTING AND DEPLOYMENT

**Communication** should occur at various stages of the project, not solely upon completion:

- keep sponsors / clients aware of broad lines
- technical details may be avoided, but documented nonetheless

**Ideal scenario:** analysis software is also reporting software

- minimizes human error related to cut-and-paste
- removes the need for keeping analysis and reporting separate
- makes sharing the work with other project member easier

Simplify the process further by deploying directly to the Web.

# DISCUSSION

What are your favourite reporting tools?

How much should you test a product before deployment?

What's the cost of deploying a faulty product?

# DASHBOARDS

A **dashboard** is any visual display of data used to monitor conditions and/or facilitate understanding.

## Examples:

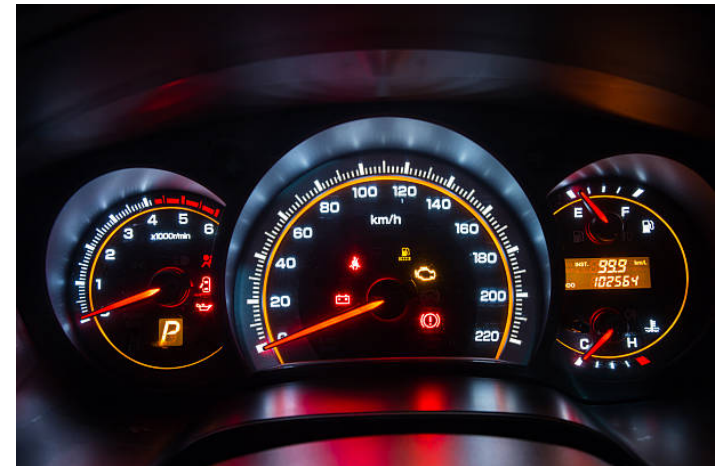
- interactive display that allows people to explore motor insurance claims by city, province, driver age, etc.
- PDF showing key audit metrics that gets e-mailed to a Department's DG on a weekly basis.
- wall-mounted screen that shows call centre statistics in real-time.
- mobile app that allow hospital administrators to review wait times on an hourly- and daily-basis for the current year and the previous year.

## SOME QUESTIONS TO CONSIDER

In a car's dashboard, a small number of **key indicators** (speed, gasoline level, lights, etc.) need to be understood **at a glance**. A dashboard design that does not take these two characteristics under consideration can have catastrophic consequences.

The following questions need to be answered prior to the dashboard being designed:

- Who is the dashboard's **consumer**?
- What **story** does the dashboard tell?
- What data (categories) will be used?
- What will **appear** on the dashboard?
- How can the dashboard **help** the consumer?



# DASHBOARD DESIGN GUIDELINES

Nick Smith suggests the following 6 Golden Rules:

- **Consider the audience** (who are you trying to inform? does the DG really need to know that the servers are operating at 88% capacity?)
- **Select the right type of dashboard** (operational, strategic/executive, analytical)
- **Group data logically, use space wisely** (split functional areas: product, sales/marketing, finance, people, etc.)
- **Make the data relevant to the audience** (scope and reach of data, different dashboards for different departments, etc.)
- **Avoid cluttering the dashboard** (present the most important metrics only)
- **Refresh your data at the right frequency** (real-time, daily, weekly, monthly, etc. )



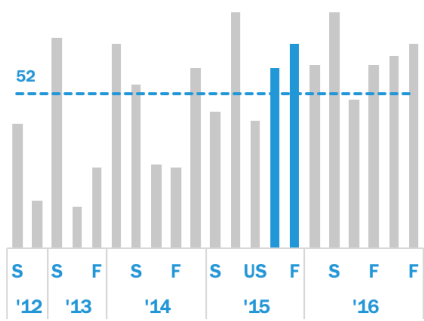


✔ Meets or Exceeds Target    ➕ Near Target    ✖ Needs Improvement    ⚙ Measuring    📊 Collecting Data

# Course Metrics

[<https://bigbookofdashboards.com/dashboards.html>]

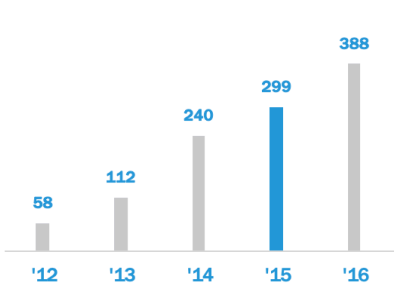
## Students



1097

Total Students in five years

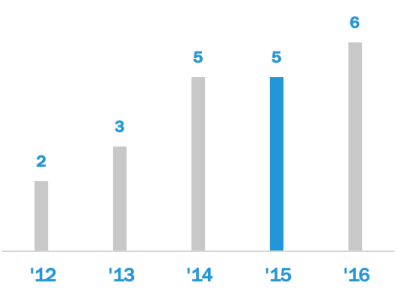
## Enrollments



687

Total Students in 2015-2016

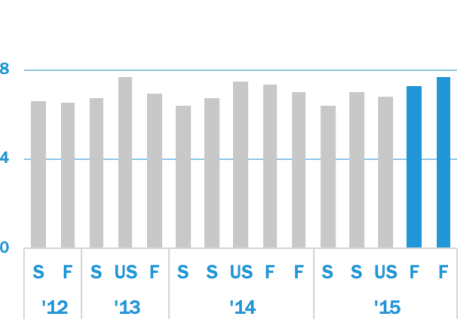
## Classes



21

Total Classes in five years

## Ratings



7.7 of 8

Most recent instructor rating (out of 8.0)

## Semesters

### 2015 Fall Semester 001

## Questions

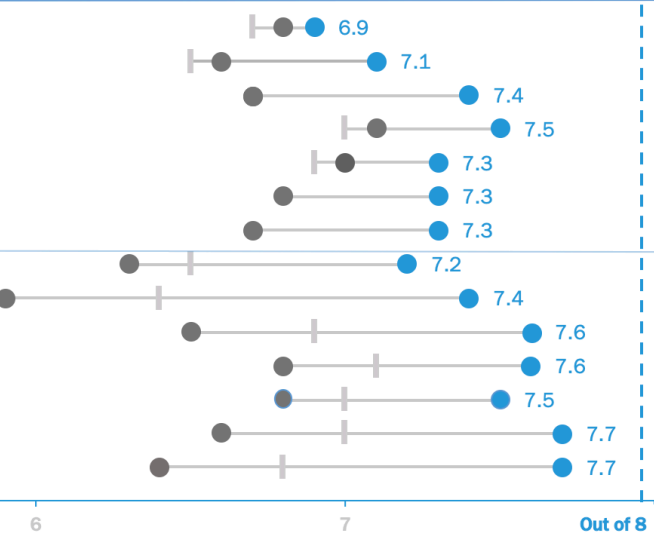
- I developed specific skills and competencies
- Overall, this was an excellent course
- The instructor communicated clearly
- The Instructor graded fairly
- The instructor was well organized
- The instructor interacted well with students
- Overall, this instructor was excellent

### 2015 Fall Semester 002

- I developed specific skills and competencies
- Overall, this was an excellent course
- The instructor communicated clearly
- The Instructor graded fairly
- The instructor was well organized
- The instructor interacted well with students
- Overall, this instructor was excellent

● BANA ■ College ● Shaffer

## Ratings



## COURSE METRICS DASHBOARD – STRENGTHS

Easy-to-see key metrics

Simple color scheme

Potential to be static or interactive

Both overview and details are clear

# DISCUSSION

There are no perfect dashboards – no collection of charts will ever suit everyone who encounters it.

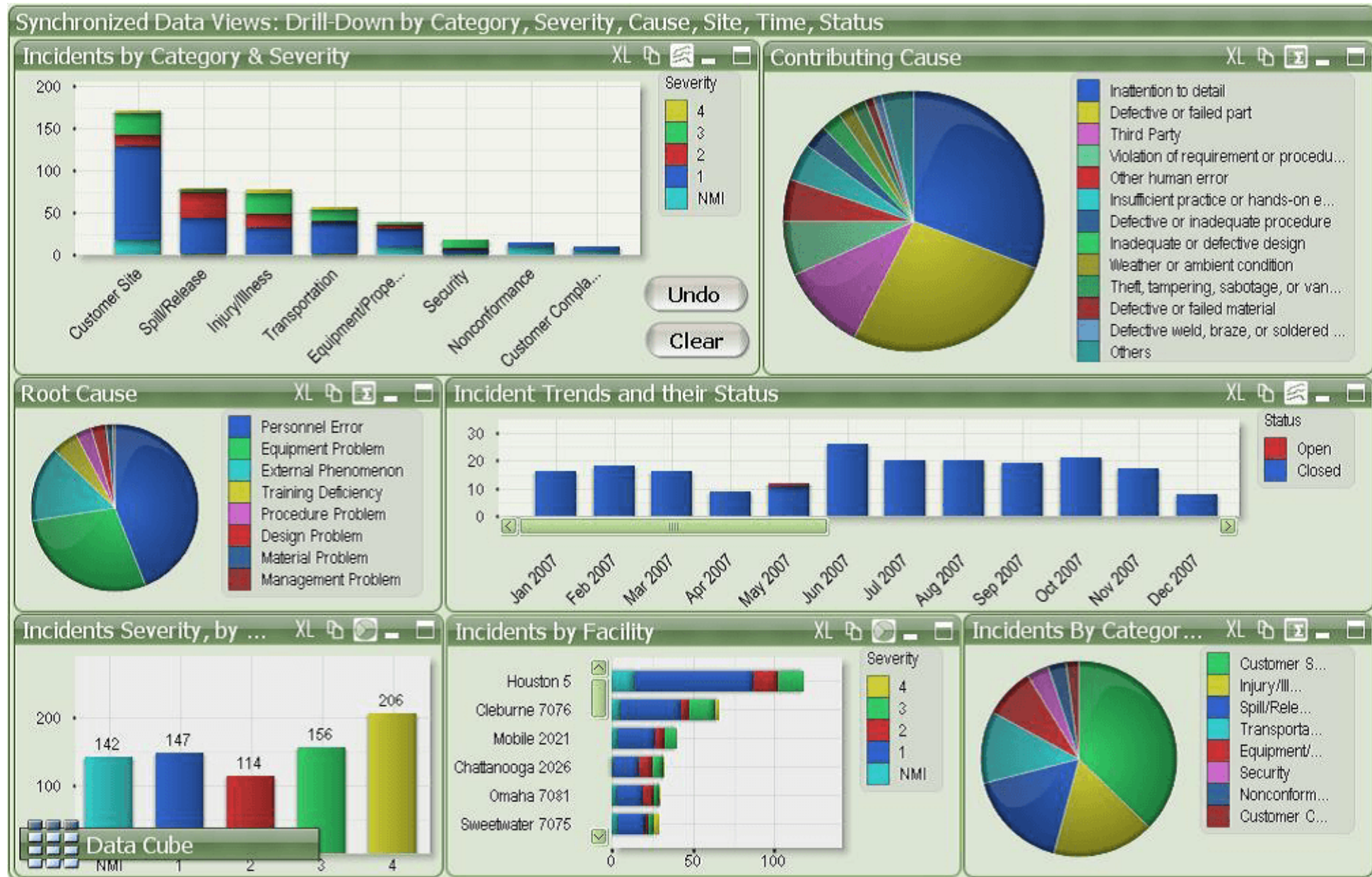
All dashboards should be **truthful** and **functional**, but dashboards that are also **elegant** (delightful, enjoyable) will take you further.

All dashboards are **incomplete**. Good dashboards will still lead to dead ends, but they should allow users to ask: “Why? What is the root cause of a problem?”

**Tools:** Excel, Power BI, Tableau, R + Shiny, Geckoboard, Matillion, etc.

## EXERCISE

Consider the following dashboards. Can you figure out, at a glance, who their audience is? What are their strengths? What are their limitations? How could you improve them?



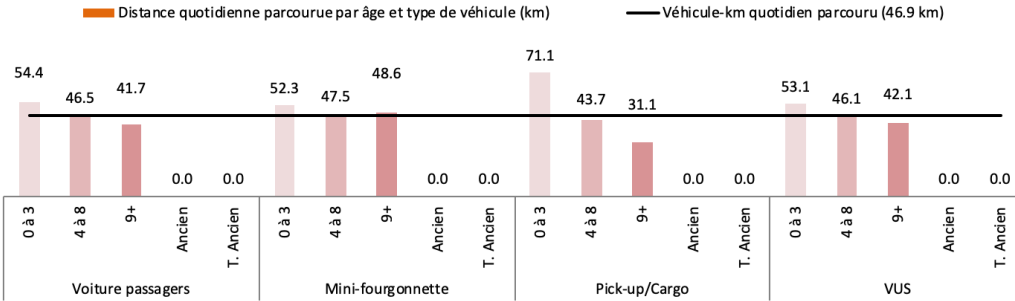




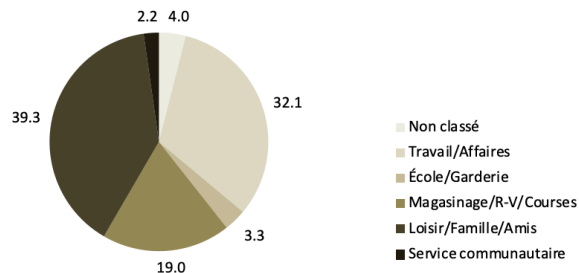
# Ontario – 1er trimestre 2012

## Caractéristiques des déplacements

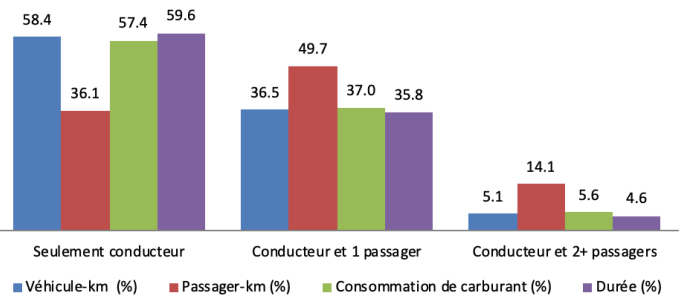
Véhicule-km quotidien parcouru par âge et type de véhicule



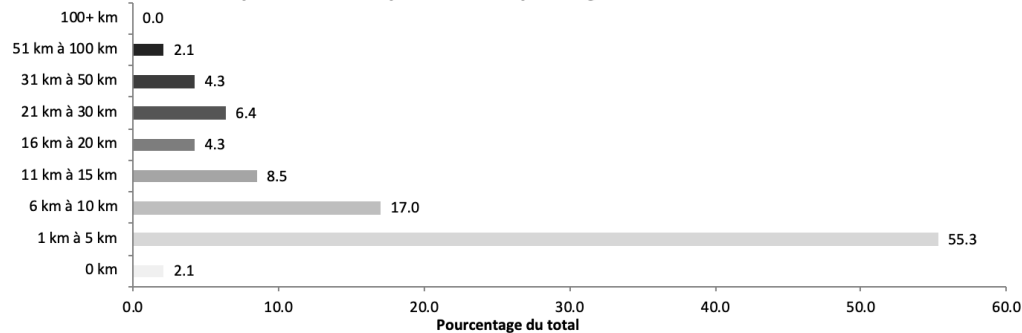
Passager-km quotidien parcouru par but des déplacements (%)



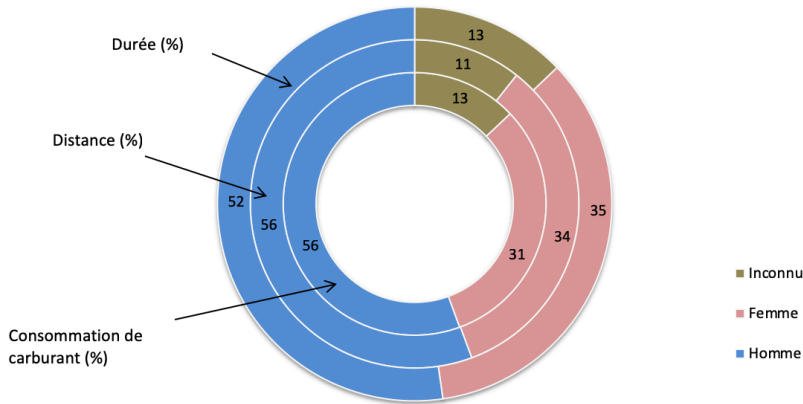
Distance, passager-km parcouru, durée et consommation de carburant par occupation



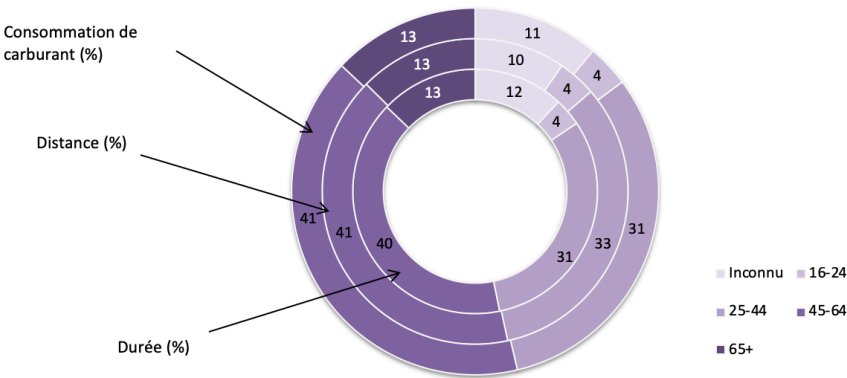
Proportion de déplacements par segments de distance



Durée, distance et consommation de carburant par sexe

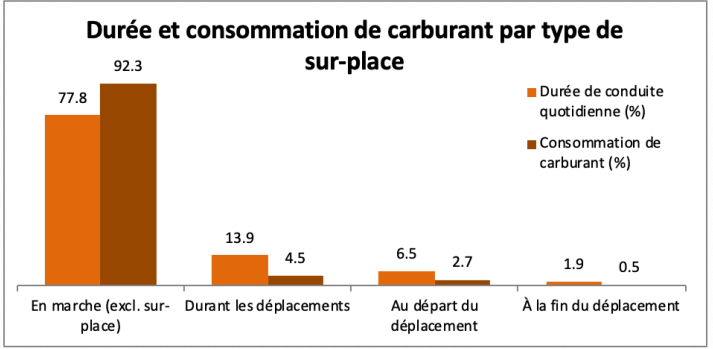
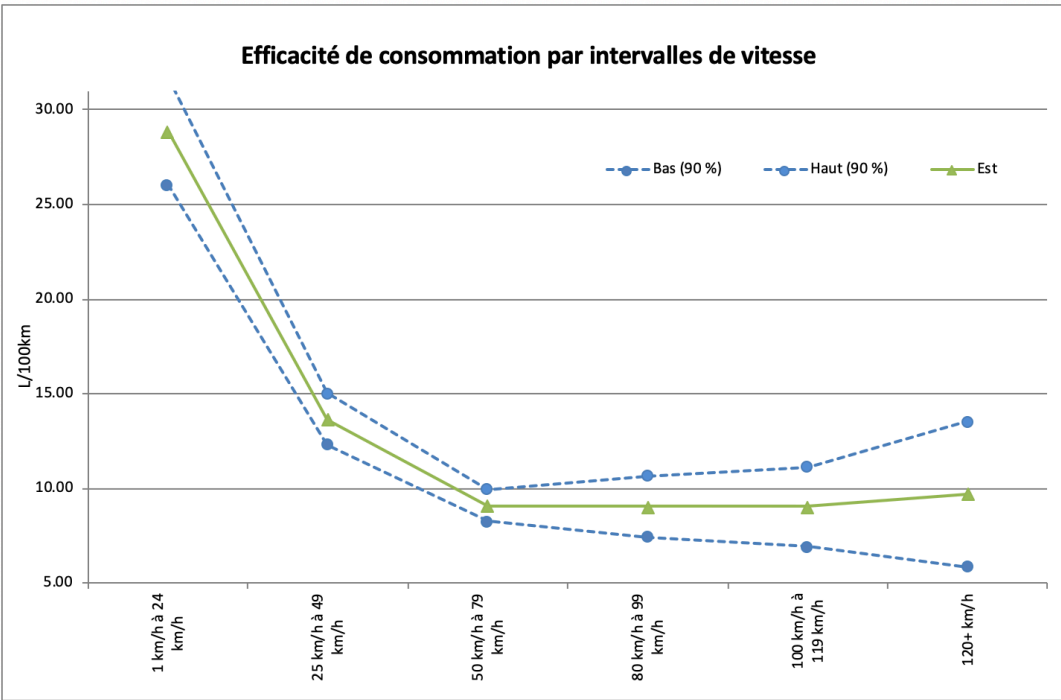
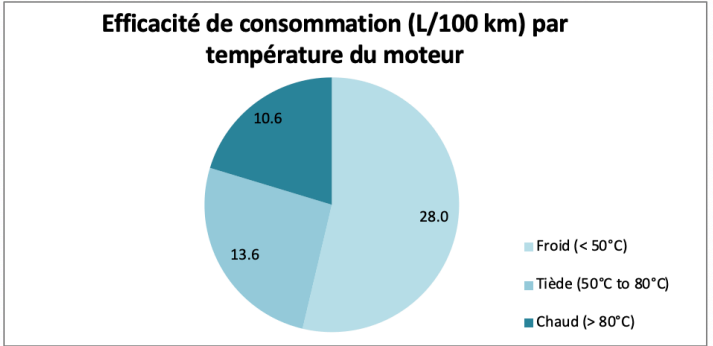
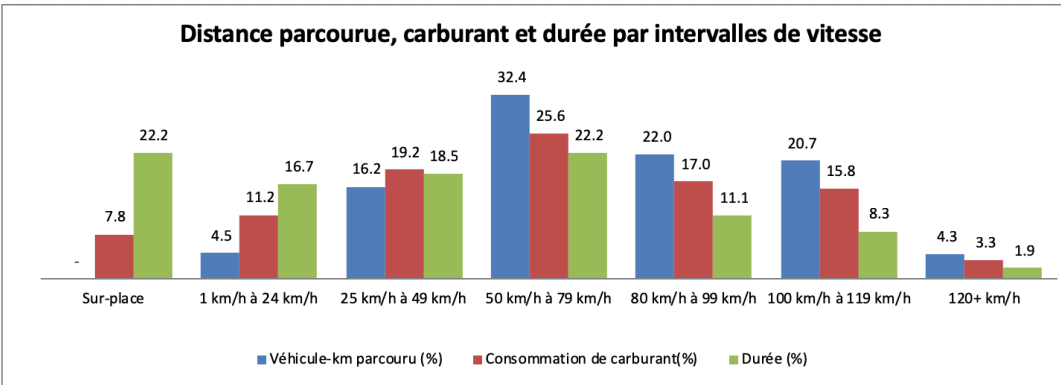
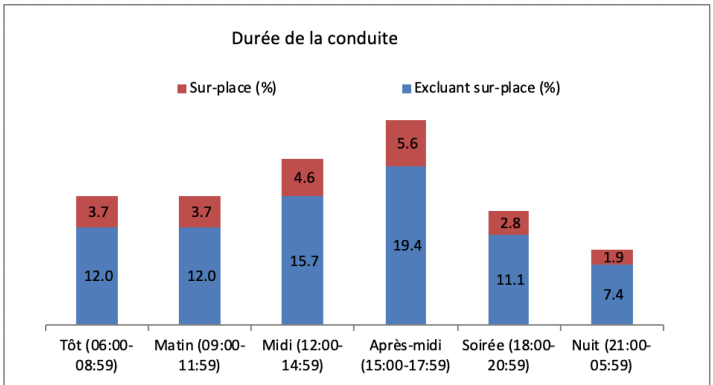


Consommation de carburant, distance et durée par âge des conducteurs



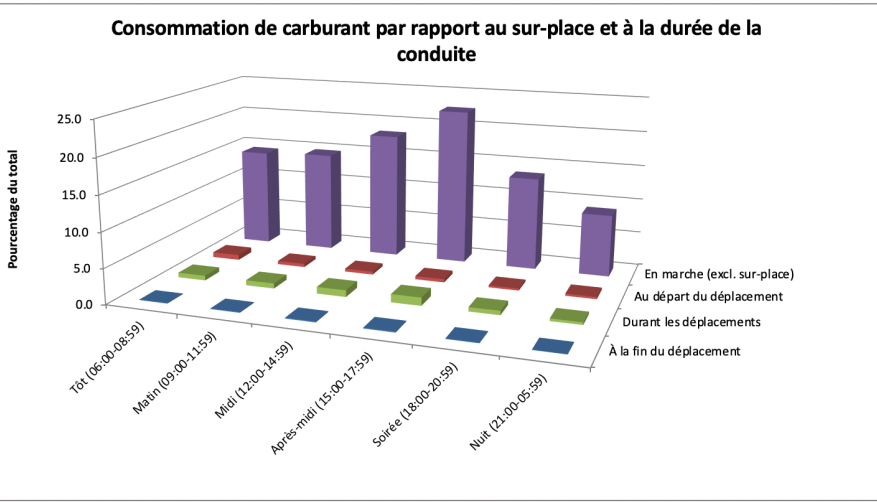
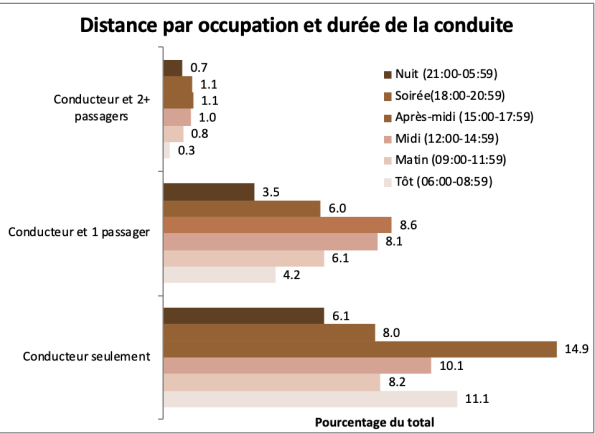
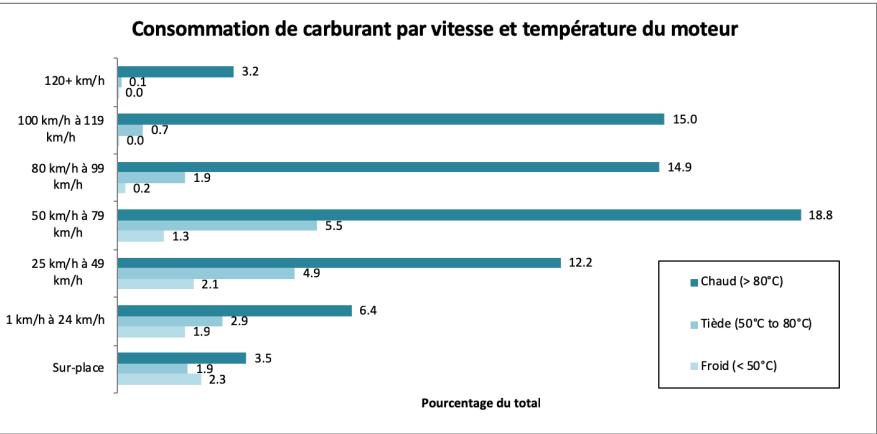
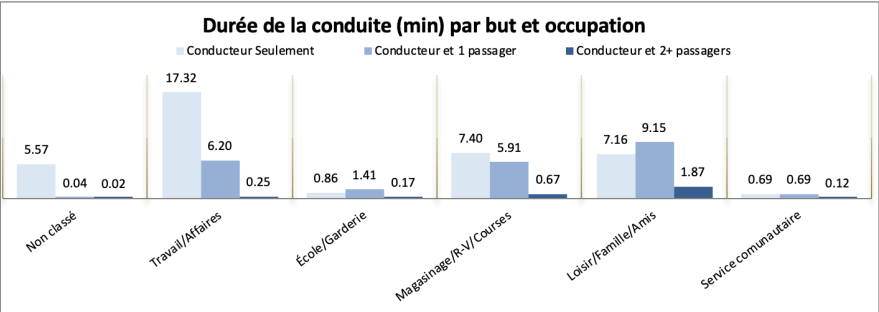
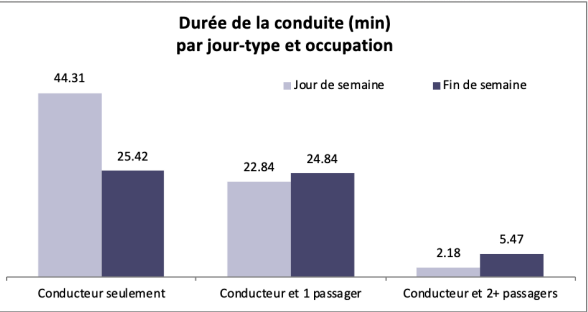
# Ontario – 1er trimestre 2012

## Sous-caractéristiques des déplacements



# Ontario – 1er trimestre 2012

## Caractéristiques mixtes sur les déplacements



# What-If Analysis: Impact of Minimum Wage

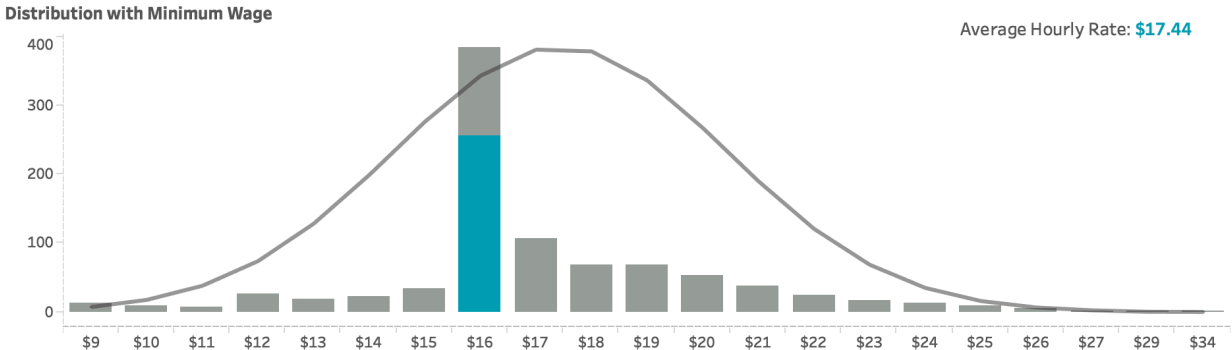
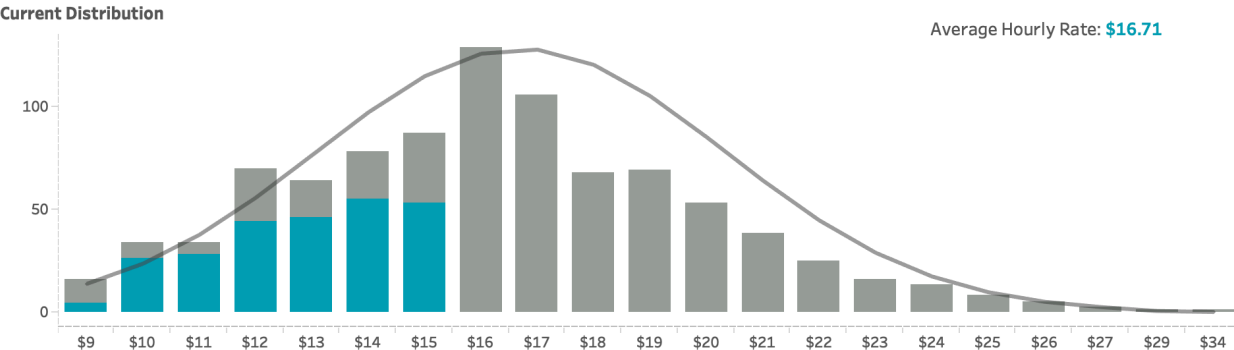
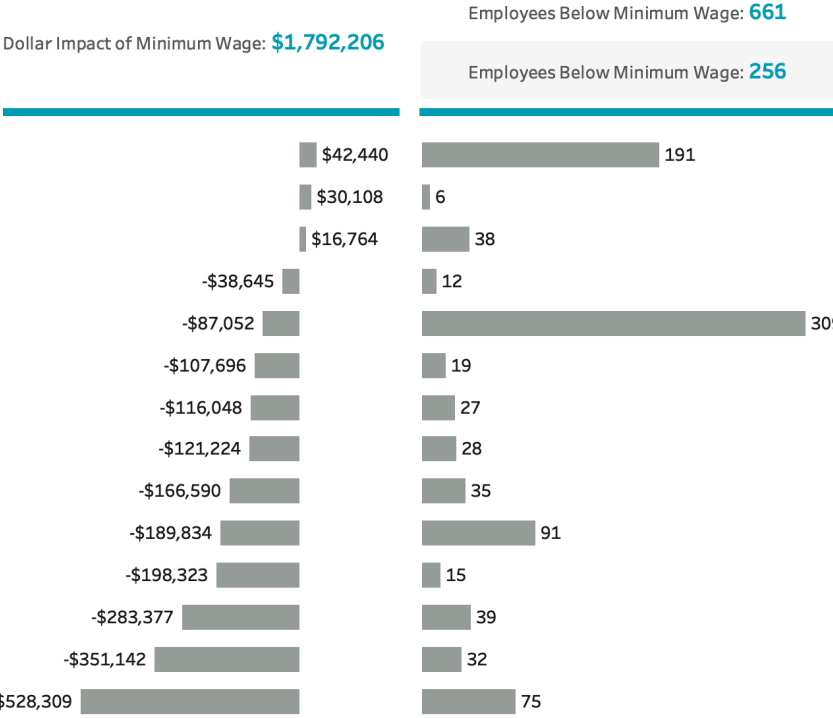
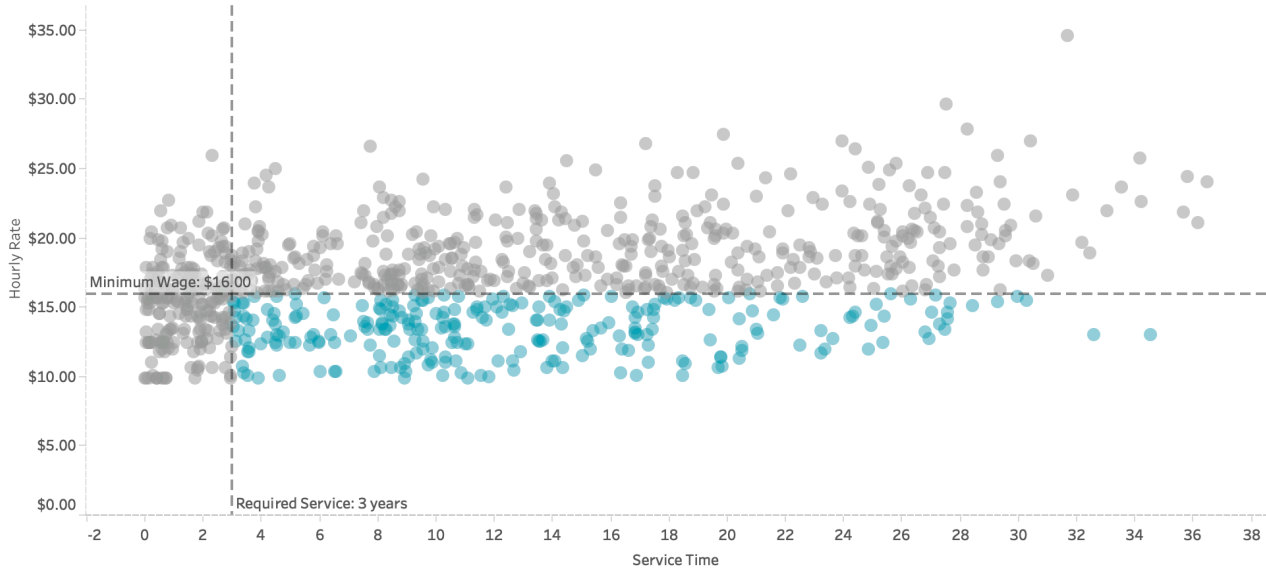
[<https://bigbookofdashboards.com/dashboards.html>]



Proposed Minimum Wage

Required Service

Developed by Matt Chambers  
<http://sirvialot.blogspot.com/>



# LIMITATIONS

Dashboard #1: not glanceable, overuse of colour, pie charts??

Dashboard #2: 3D visualizations, distracting borders and background, lack of filtered data, insufficient labels and context

Dashboards #3 – 4: ...

## EXERCISE

In teams or individually, identify a scenario for which a dashboard could prove useful.

Determine specific questions that the dashboard could help answer or insights that it could provide.

Identify data sources and data elements that could be fed into your dashboard.

Design a display (with pen and paper) with mock charts.

What are the strengths and limitations of your dashboard? Is it functional? Elegant?