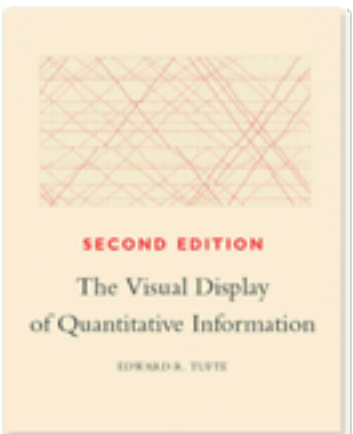
GENERAL ASSEMBLY

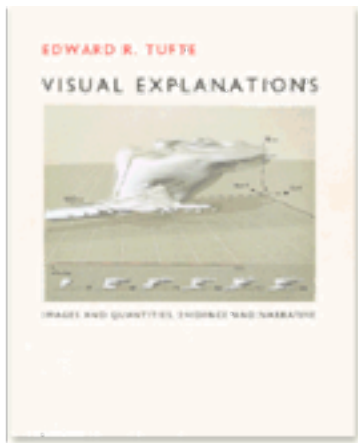
INTRO TO DATA SCIENCE
LECTURE 3:
EXPLORATORY DATA ANALYSIS

KNOWLEDGE DISCOVERY & STORY TELLING

PIONEER OF DATA VISUALIZATION - EDWARD TUFTE (edwardtufte.com)

Professor Emeritus of Political Science, Statistics, and Computer Science at Yale
Wrote, designed, self-published 4 classic books on data visualization





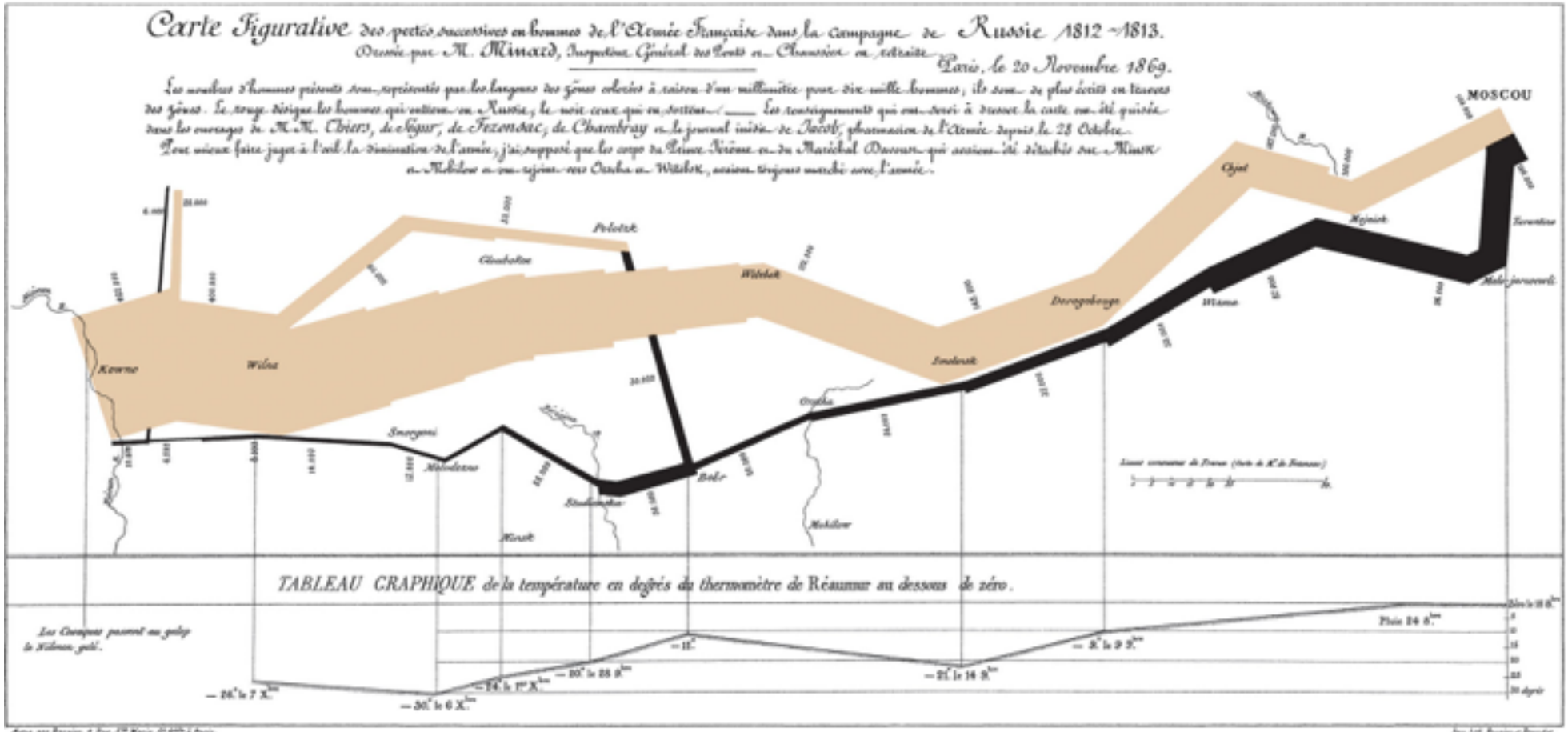
- In 1975 Tufte taught Statistics to a group of journalists who were visiting the school to study economics
- He developed a set of lectures on statistical graphics, which became joint seminars with John Tukey.
- John Tukey is a pioneer in the field of information design.
- The material was the foundation for “The Visual Display of Quantitative Information”
- Tufte coined the following phrases:
    - “chartjunk” = useless, non-informative or information-obscuring elements in quantitative displays
    - “data-ink ratio” = excessive decoration of visual displays

Tufte's believed in:

- Use data-rich illustrations that present *all* available data
- Close examination: every data point has a value
- General examination: only trends and patterns can be observed

- Folks who did it well:
  - Charles Joseph Minard - Napolean's March
  - Dr John Snow - London Dot Map

Napleon's Russian Campaign of 1812



number of troops (1mm = 10000), distance travelled, temperature, latitude and longitude, direction, location relative to specific dates

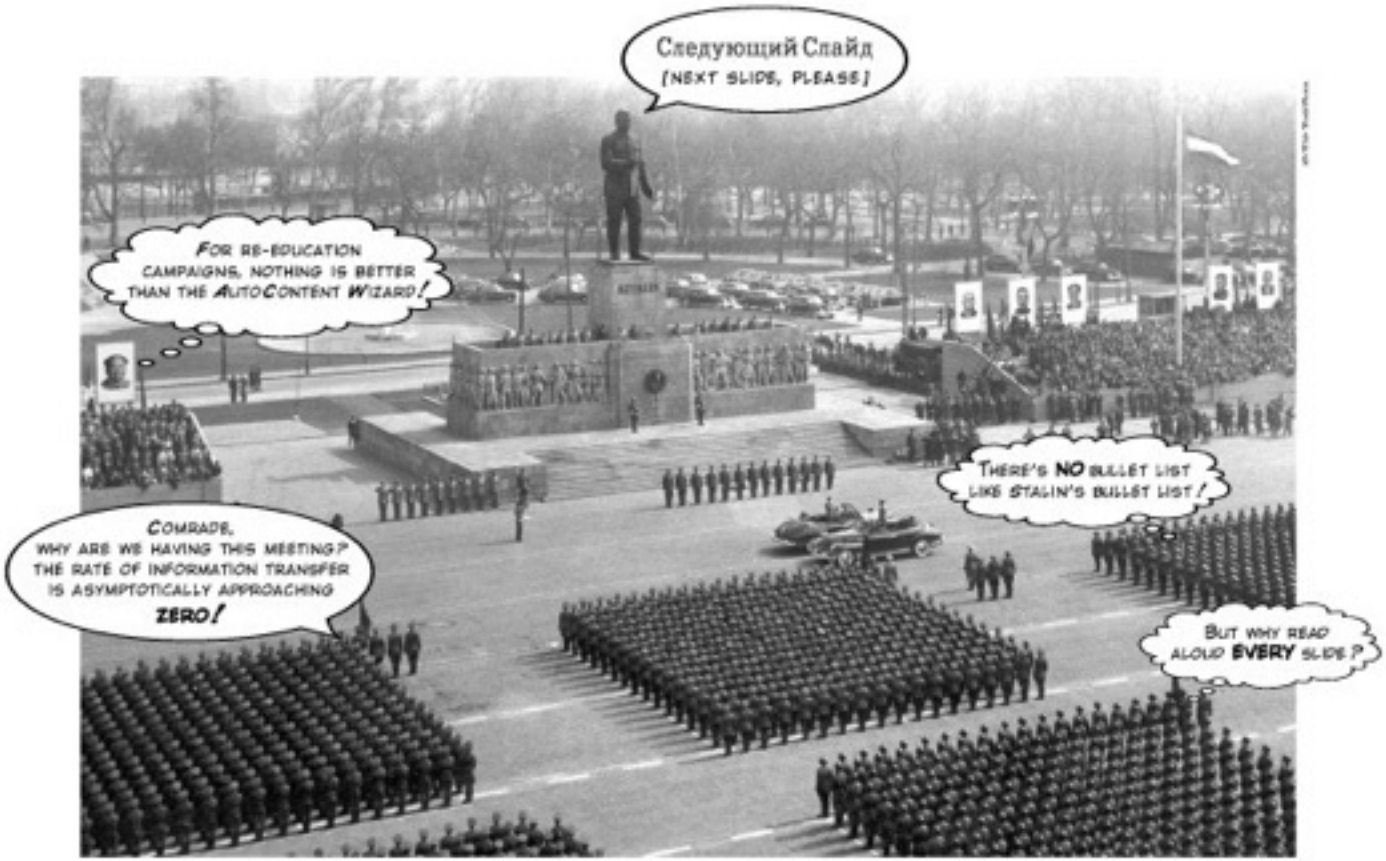
Tufte makes serious criticism of Powerpoint, in an essay entitled “The Cognitive Style of PowerPoint”

- guide and reassure the presenter rather than enlighten the audience

- unhelpfully simplistic charts

- poor typography

- simplistic thinking



Edward Tufte, *The Cognitive Style of PowerPoint*

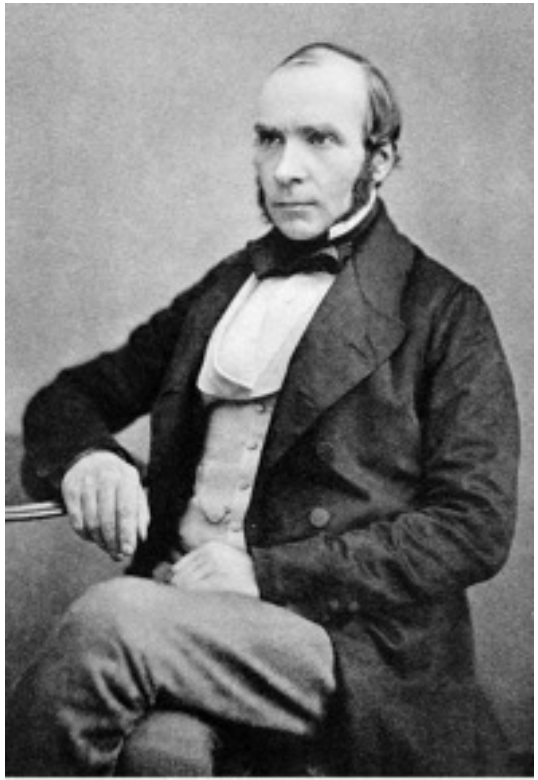
Tufte examined the way NASA engineers used Powerpoint in the events that lead to the Space Shuttle Challenger Disaster

- style designed to persuade, rather than inform

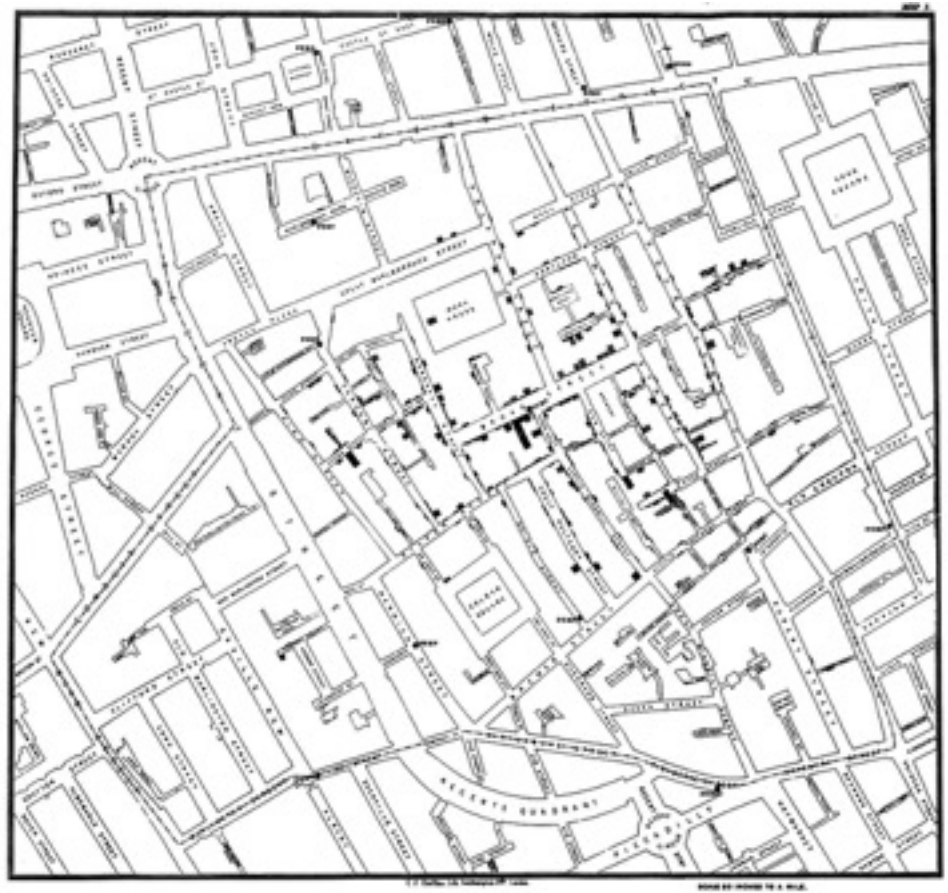
- his analysis was included in the official report

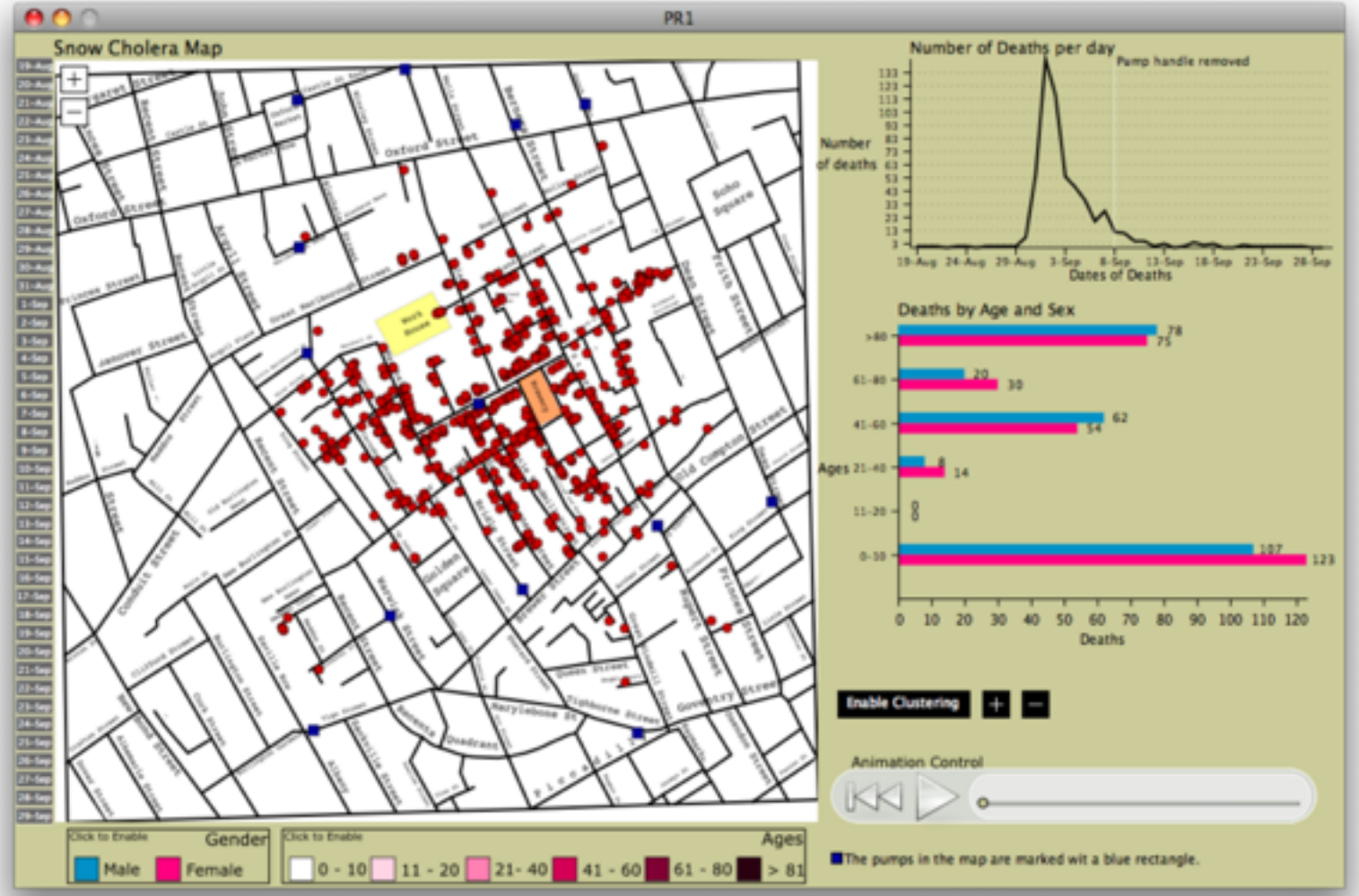
- of specific note: an engineering detail buried in small type on a crowded slide with 6 bullet points

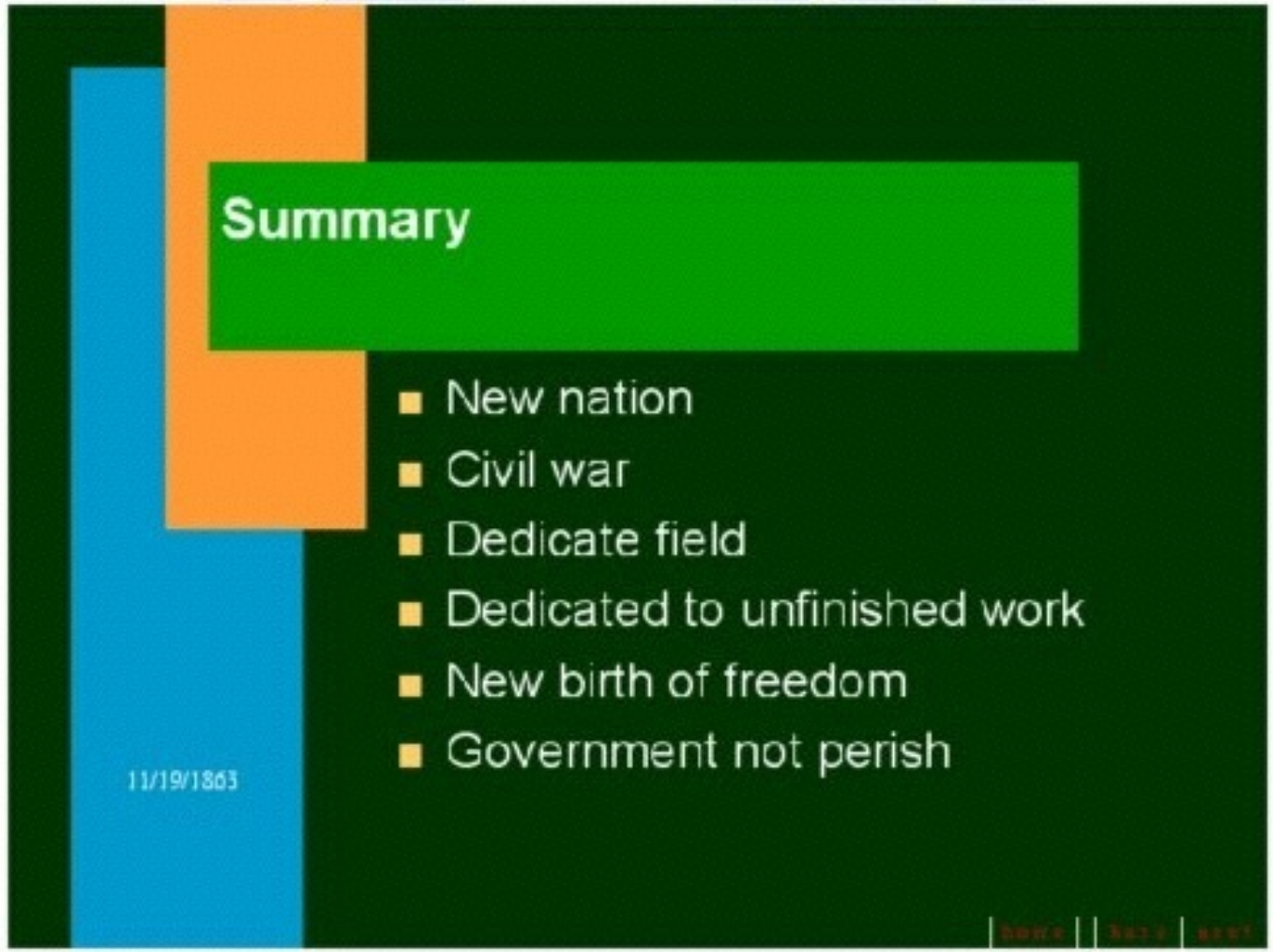
- such a detail presented in a regular engineering white paper



John Snow




PR1
Snow Cholera Map
Number of Deaths per day
Pump handle removed
Number of deaths
Dates of Deaths
Deaths by Age and Sex
>80
61-80
41-60
Ages 21-40
11-20
0-10
78
75
20
30
62
14
8
14
8
8
107
123
0 10 20 30 40 50 60 70 80 90 100 110 120
Deaths
Enable Clustering
Animation Control
Click to Enable
Gender
Male
Female
Click to Enable
Ages
0 - 10
11 - 20
21 - 40
41 - 60
61 - 80
> 81
The pumps in the map are marked wit a blue rectangle.


Summary
New nation
Civil war
Dedicate field
Dedicated to unfinished work
New birth of freedom
Government not perish
11/19/1863

EXPLORATORY DATA ANALYSIS

“Exploratory Data Analysis” is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there.“

- John Tukey

John Tukey was Professor Emeritus of Political Science, Statistics, and Computer Science at Yale

- coined the term ‘bit’
- the boxplot
- FFT

- Gain intuition about the data
- Inspect and compare distributions (data transformation)
- Sanity checking
- Handling categorical variables
- Identifying missing data, and subsequently handling them
- Identifying outliers, and subsequently handling them
- Identifying out-of-range values
- Identifying impossible data combinations
- Summarize the data/summary statistics

Knowledge Discovery

The Training Set

- ***Mean***
- ***Variance***
- ***Correlation***

Francis Anscombe

- constructed 4 datasets in 1973
- to demonstrate:
  - the importance of graphing data before analyzing it
  - the effect of outliers

1. Convenience (e.g. percentages vs. original data, radians vs degrees)

2. Reducing skewness
   a. take roots or logarithms or reciprocals (common)
   b. take squares, cubes
   c. http://en.wikipedia.org/wiki/Skewness

3. Equalizing “spread”
   a. Each data set or subset having about the same spread or variability is ***homoscedasticity***: the opposite is called ***heteroscedasticity***
   b. http://en.wikipedia.org/wiki/Heteroscedasticity

4. Scaling/Normalization

Most Common:

1. Reciprocal
2. Logarithm
3. Square/Cube root
4. Power