SIBB512
Exploratory Data Analysis
Nov 28, 2019

Exploratory Data Analysis

Bhoom Suktitipat, MD, PhD
[bhoom.suk@mahidol.edu]
Graduate Program in Medical Bioinformatics
Department of Biochemistry
Faculty of Medicine Siriraj Hospital
&
Integrative Computational BioScience Center (ICBS)
Mahidol University

Objectives

- Introduction to Statistics
- Exploratory Data Analysis
- Graphical EDA
  - Exploring Relationship
- Non-graphical EDA
  - Measuring Central Tendency (mean/median)
  - Spreads, Precision & Accuracy (SD
- Envisioning Information

Odds update every 5 minutes | *Last updated: 12:52AM EDT on Aug 26, 2016*

Election Betting Odds

By Maxim Lott and John Stossel

Why This Beats Polls | Odds from Betfair and PredictIt | How People Bet

President | Congress | Third Party in Debate | Charts

Chances of winning...

Senate Control

72.5%

▲ 12.0% since 8pm 5/22

DEM

27.5%

▼-12.0%

REP

House of Representatives Control

94.5%

▲ 10.0% since 8pm 5/22

REP

5.5%

▼-10.0%

DEM

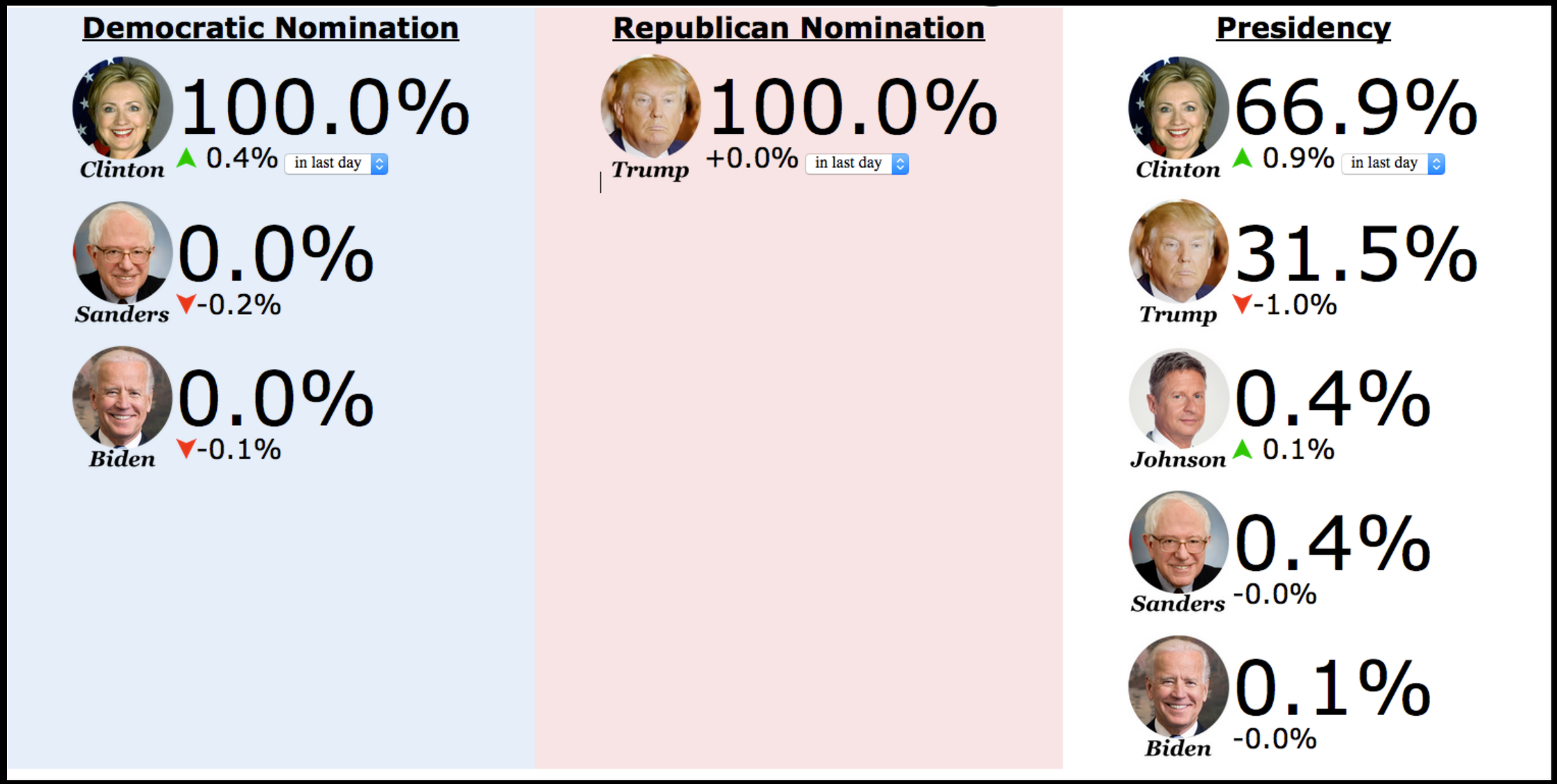
This convention page currently reports bets from PredictIt. Betfair to be averaged in if/when it has a liquid market for Congress

About these odds and FAQ | By Maxim Lott and John Stossel | Odds update every 5 minutes

| Tweet

https://electionbettingodds.com/ 2016-08-26

Statistics is a tool that helps quantify uncertainty.



https://electionbettingodds.com/ 2016-07-28

Statistics is a tool that helps quantify uncertainty.

Statistics

SCIENCE

TABLE OF CONTENTS

WRITTEN BY:

Dennis J. Sweeney
David R. Anderson
Thomas A. Williams

Statistics, the science of collecting, analyzing, presenting, and interpreting data. Governmental needs for census data as well as information about a variety of economic activities provided much of the early impetus for the field of statistics. Currently the need to turn the large amounts of data available in many applied fields into useful information has stimulated both theoretical and practical developments in statistics.

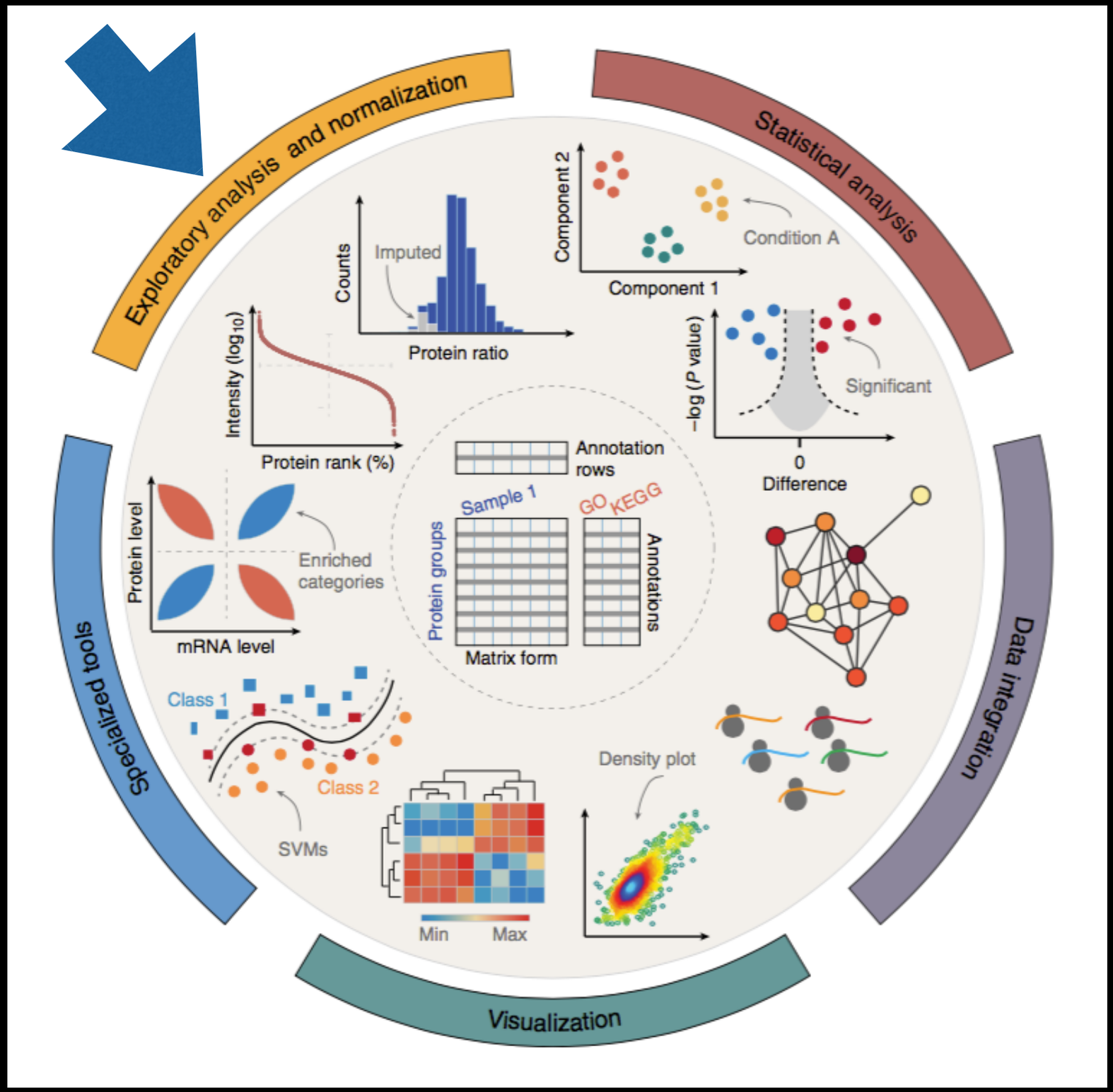
Data are the facts and figures that are collected, analyzed, and summarized for presentation and interpretation. Data may be classified as either quantitative or qualitative. Quantitative data measure either how much or how many of something, and ... (100 of 12,460 words)

ENCYCLOPÆDIA BRITANNICA

Statistics

- **Descriptive statistics**
  - **Describing the basic features of the data**
- **Inferential statistics**
  - Making propositions about a population/phenomenon.

Typical Data Life Cycle



Perseus Data Analysis Platform. Nature Method (2016) doi:10.1038/nMeth.3901

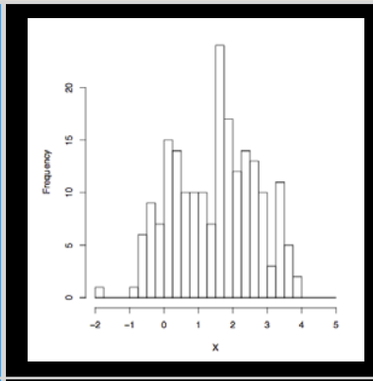
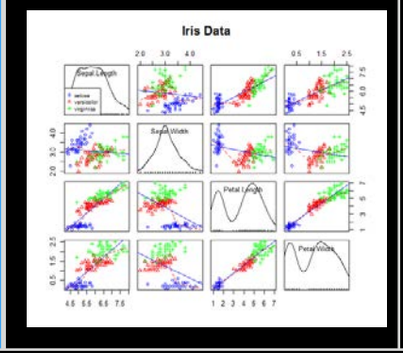
Exploratory Data Analysis

- an approach to analyzing data sets to summarize their main characteristics, ***often with visual methods.***
- Seeing what the data can tell us.
- No formal hypothesis testing
- No model

Image: Wikipedia

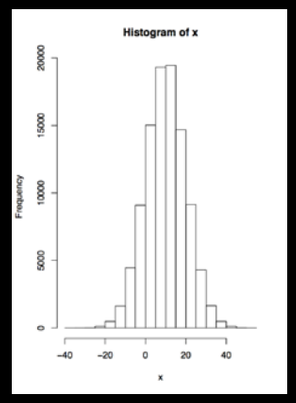
John Wilder Tukey

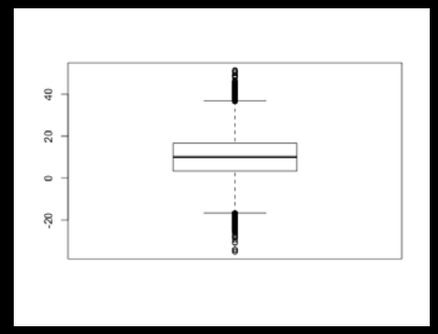
Exploratory Data Analysis

	Graphical	Non-graphical
Univariate		Central Tendency, Spread
Multivariate		Contingency Table (2x2)

EDA Graphical Technique

- **Box plot**
- **Histogram**
- **Scatter plot**
- Stem-and-leaf plot

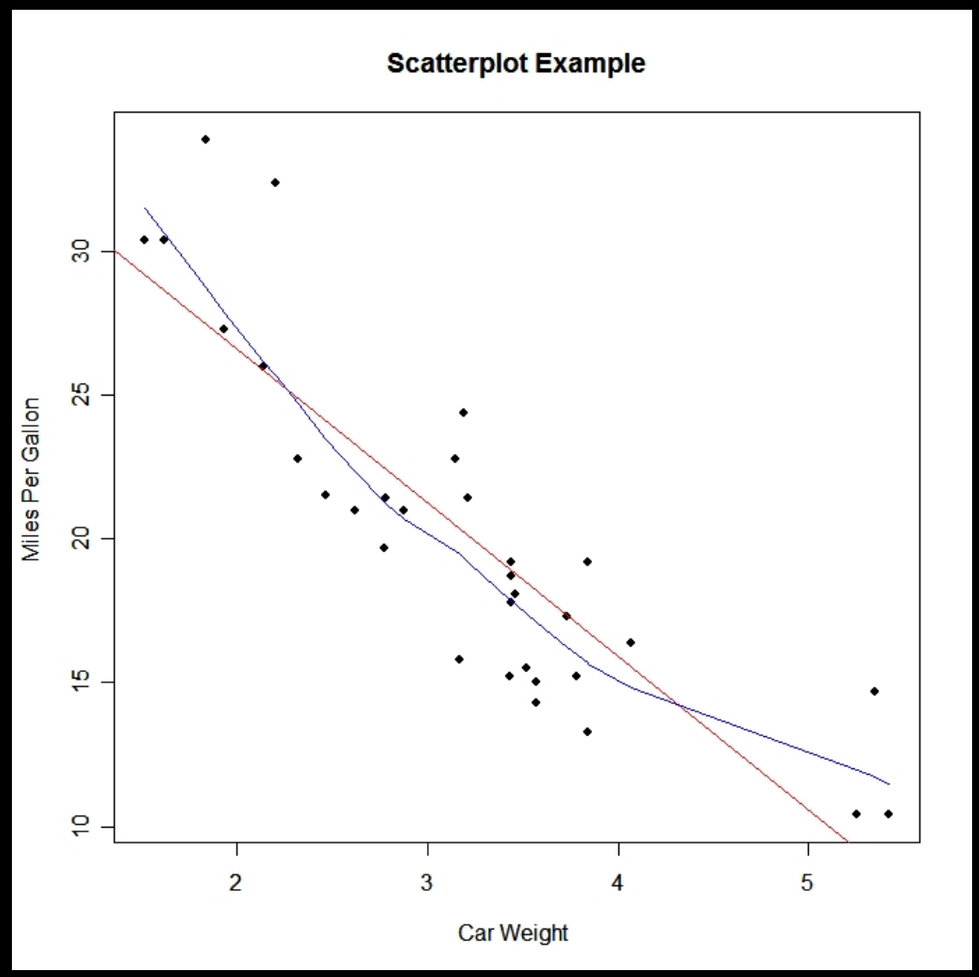




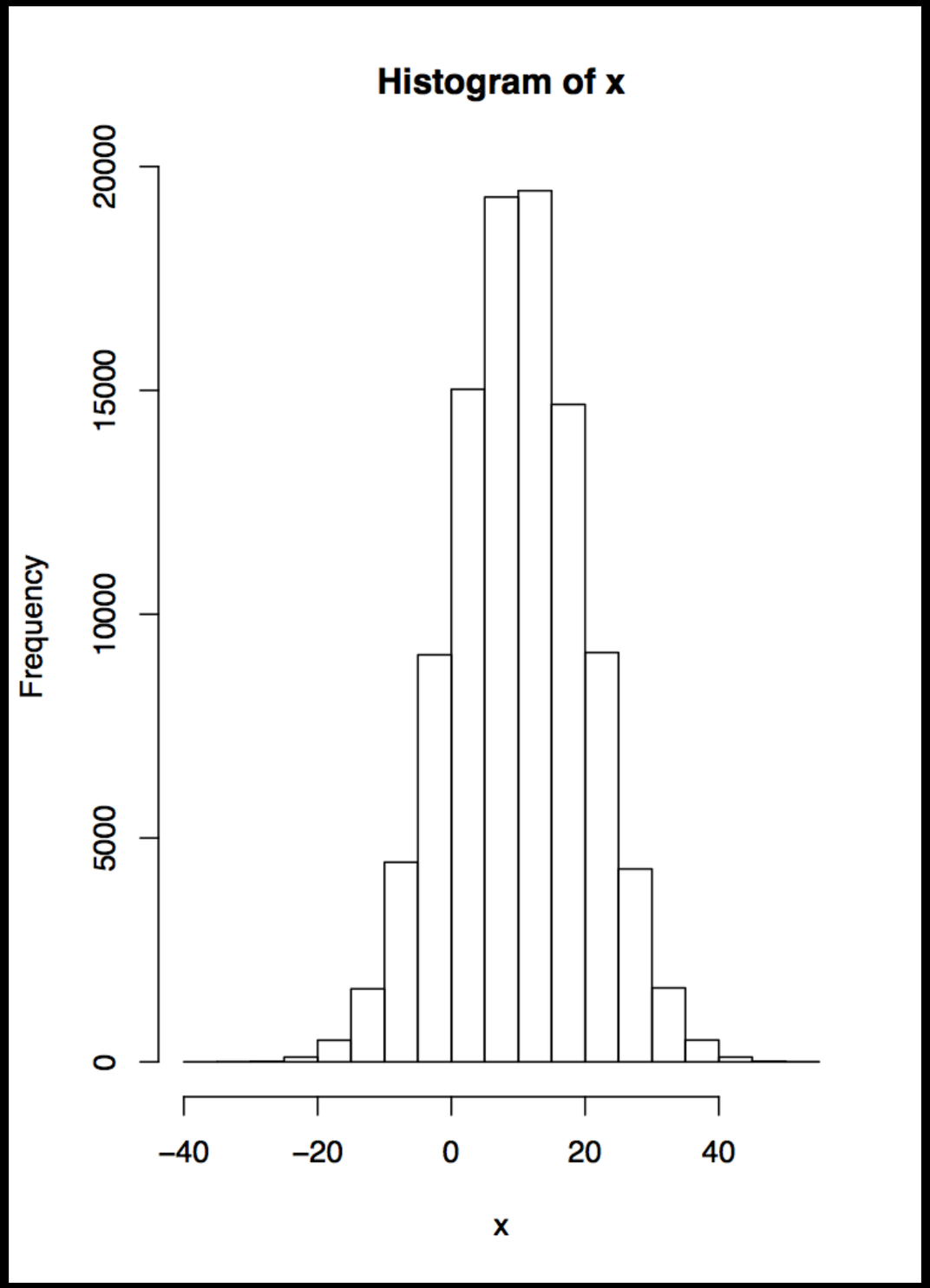

```
> stem(c(15,16,21,23,23,26,26,30,32,41))

  The decimal point is 1 digit(s) to the right of the |

  1 | 56
  2 | 13366
  3 | 02
  4 | 1
```



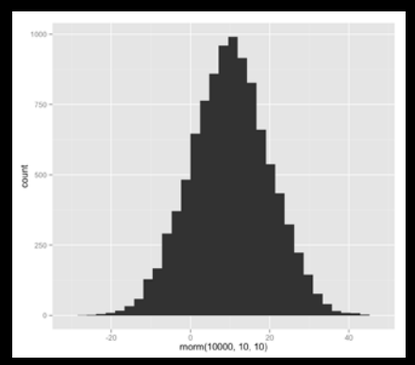
Histogram

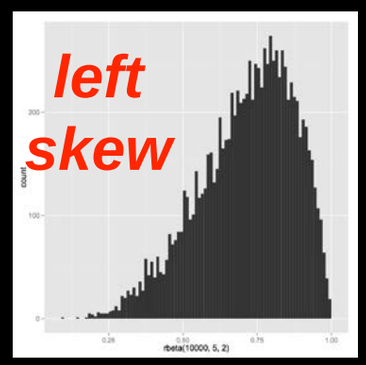

Histogram of x
Frequency
0
5000
10000
15000
20000
−40
−20
0
20
40
x

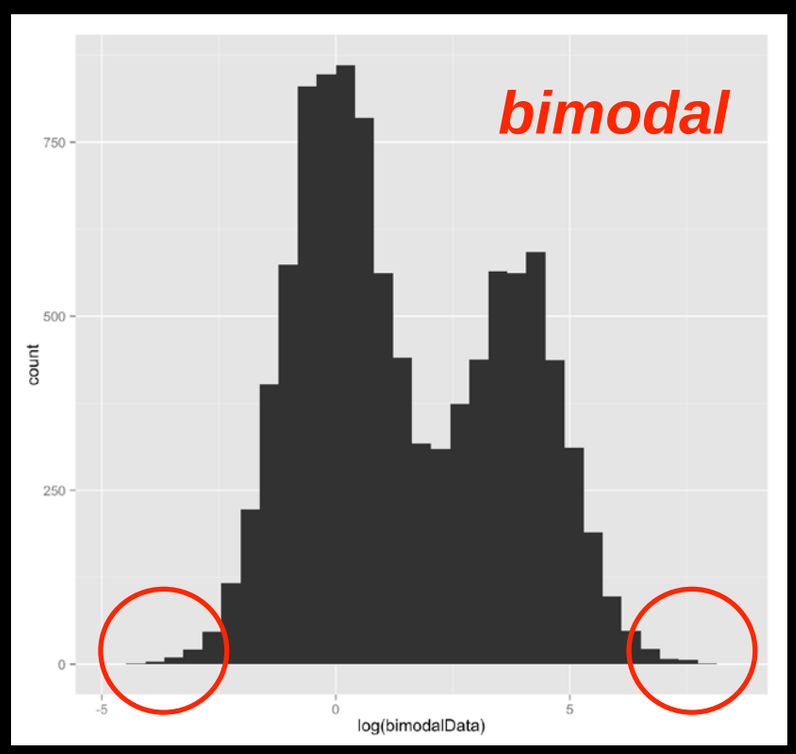
Histogram

- Shape of distribution
  - Symmetry, Skewness, modality, outliers

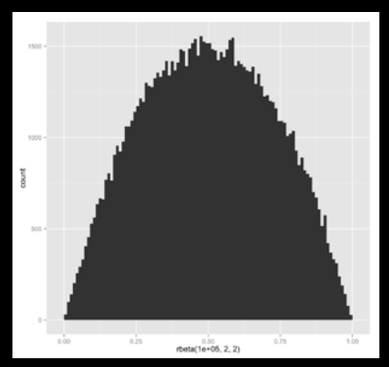
normal/Gaussian distribution

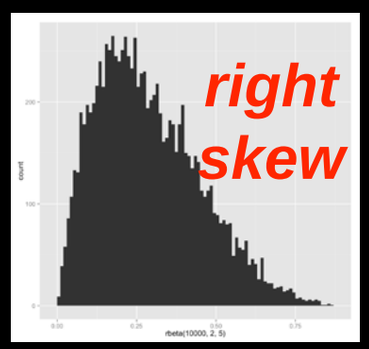




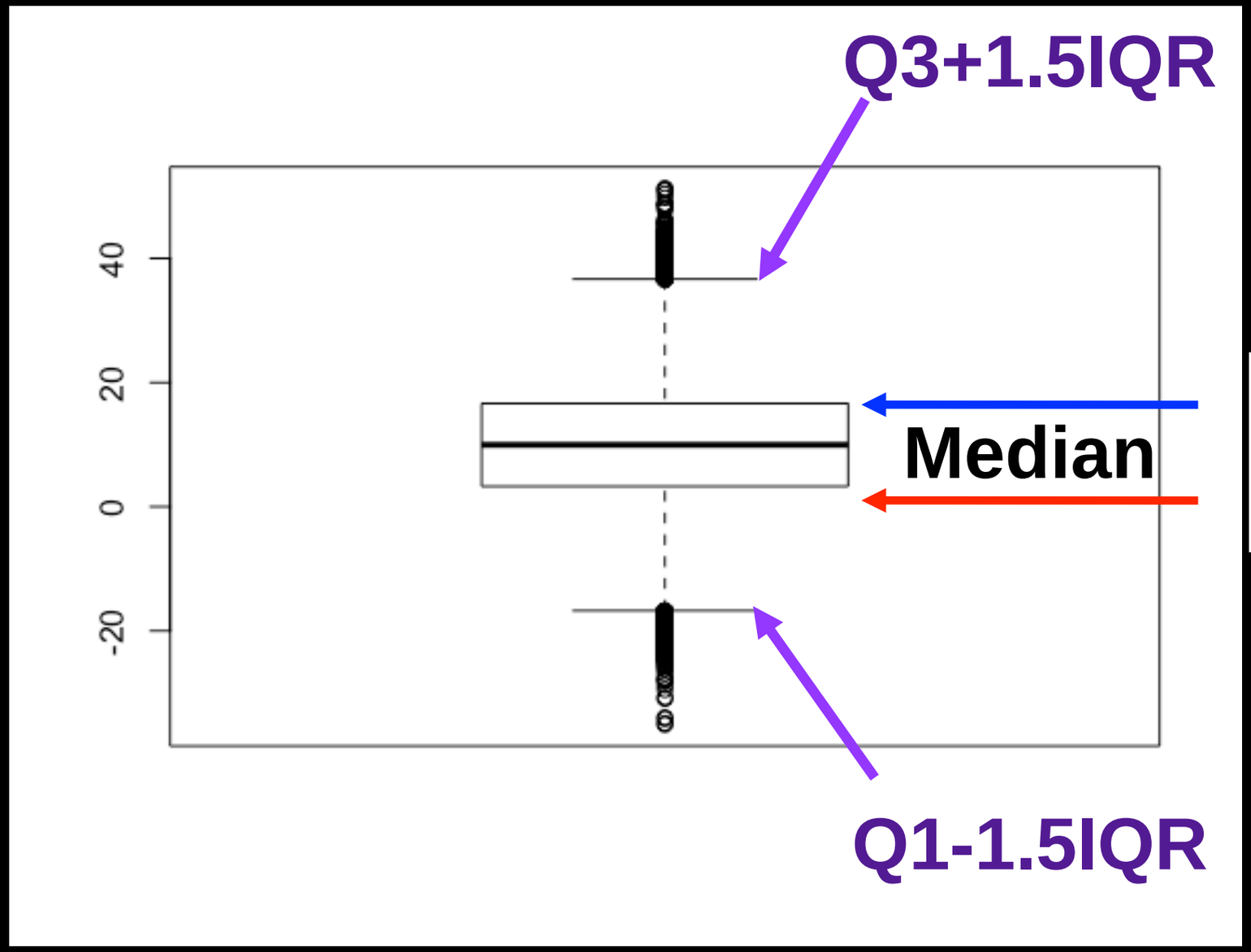


t-distribution



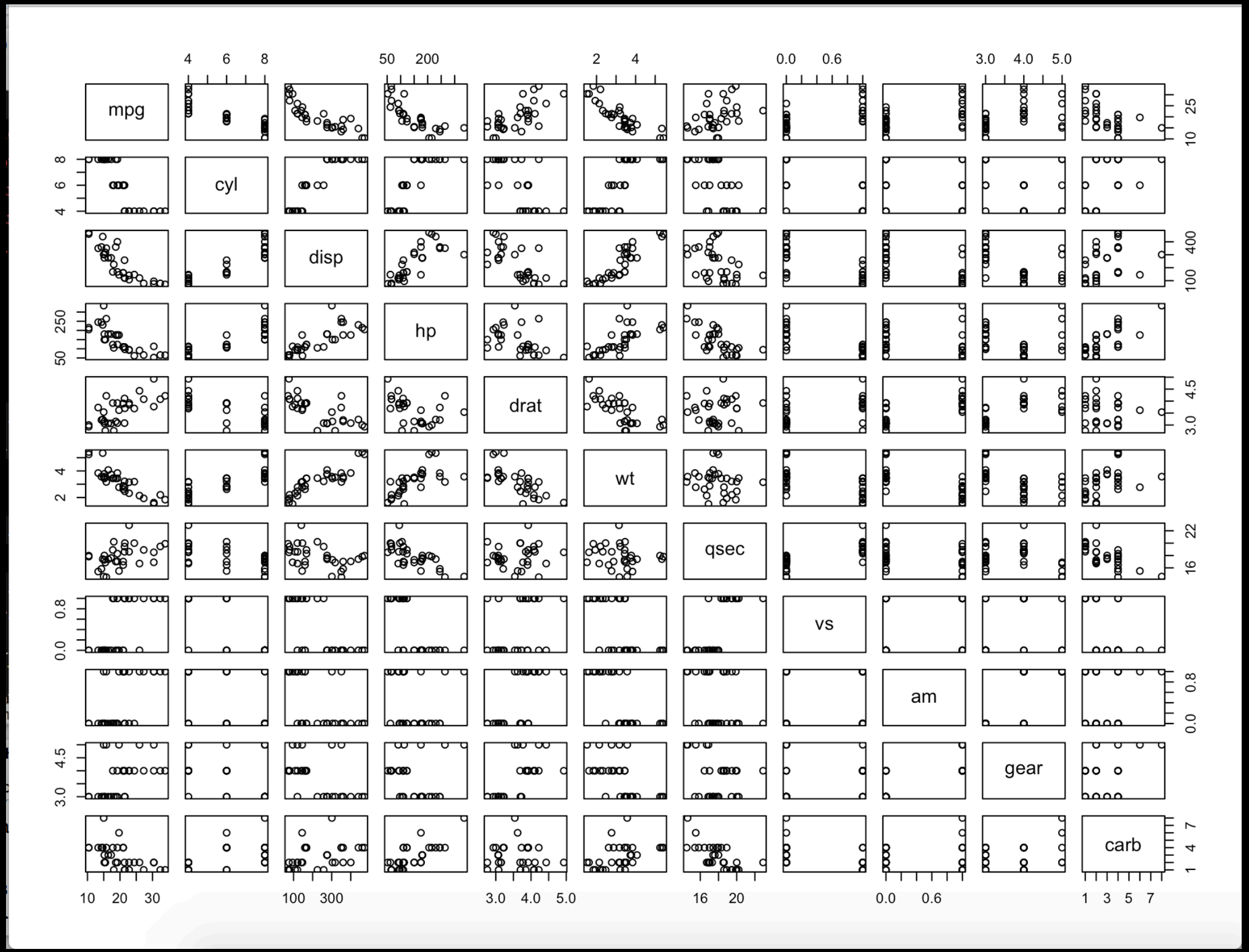


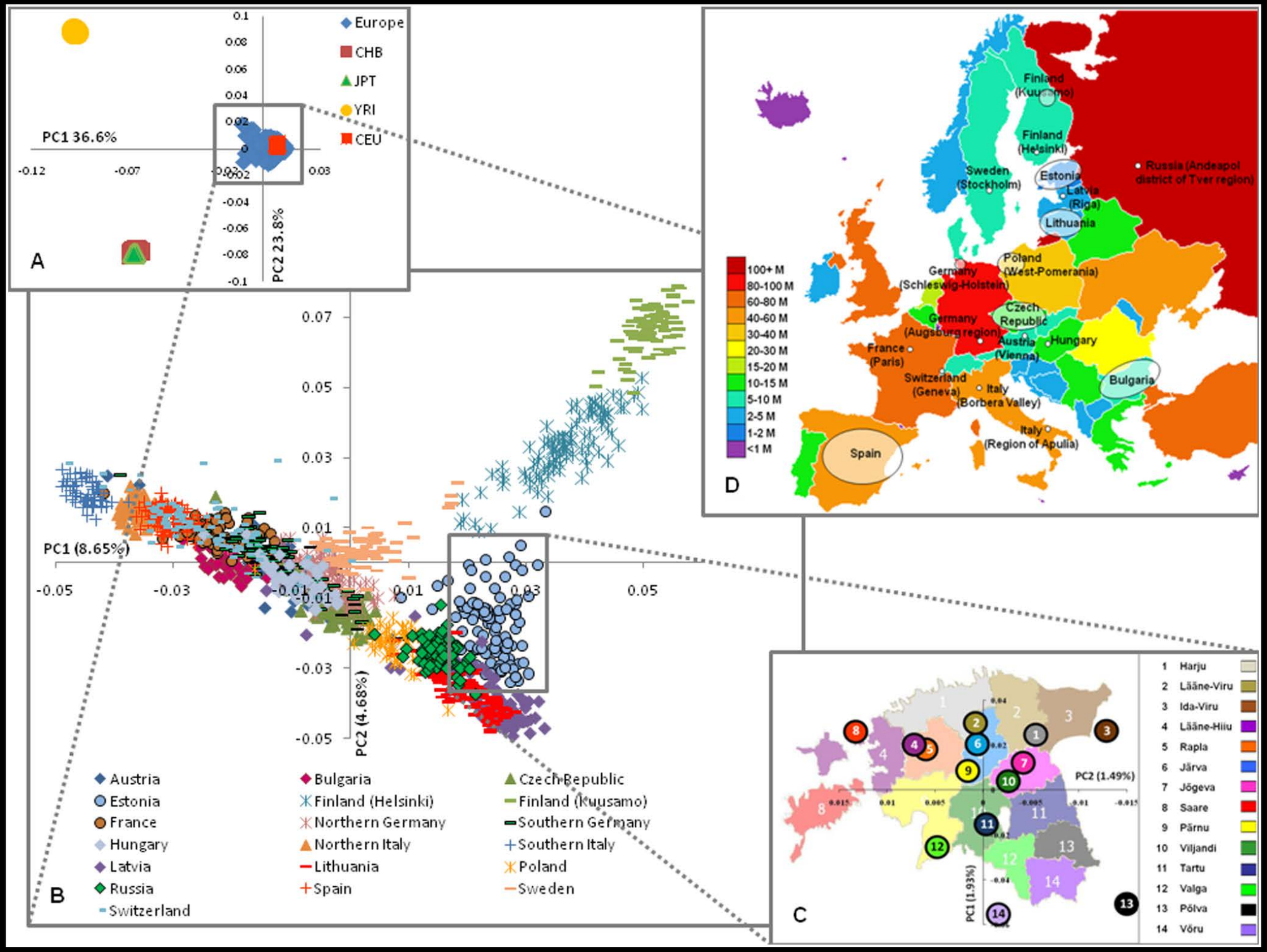
Boxplot



Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-35.090	3.277	9.961	9.960	16.640	51.300

Exploring Relationship

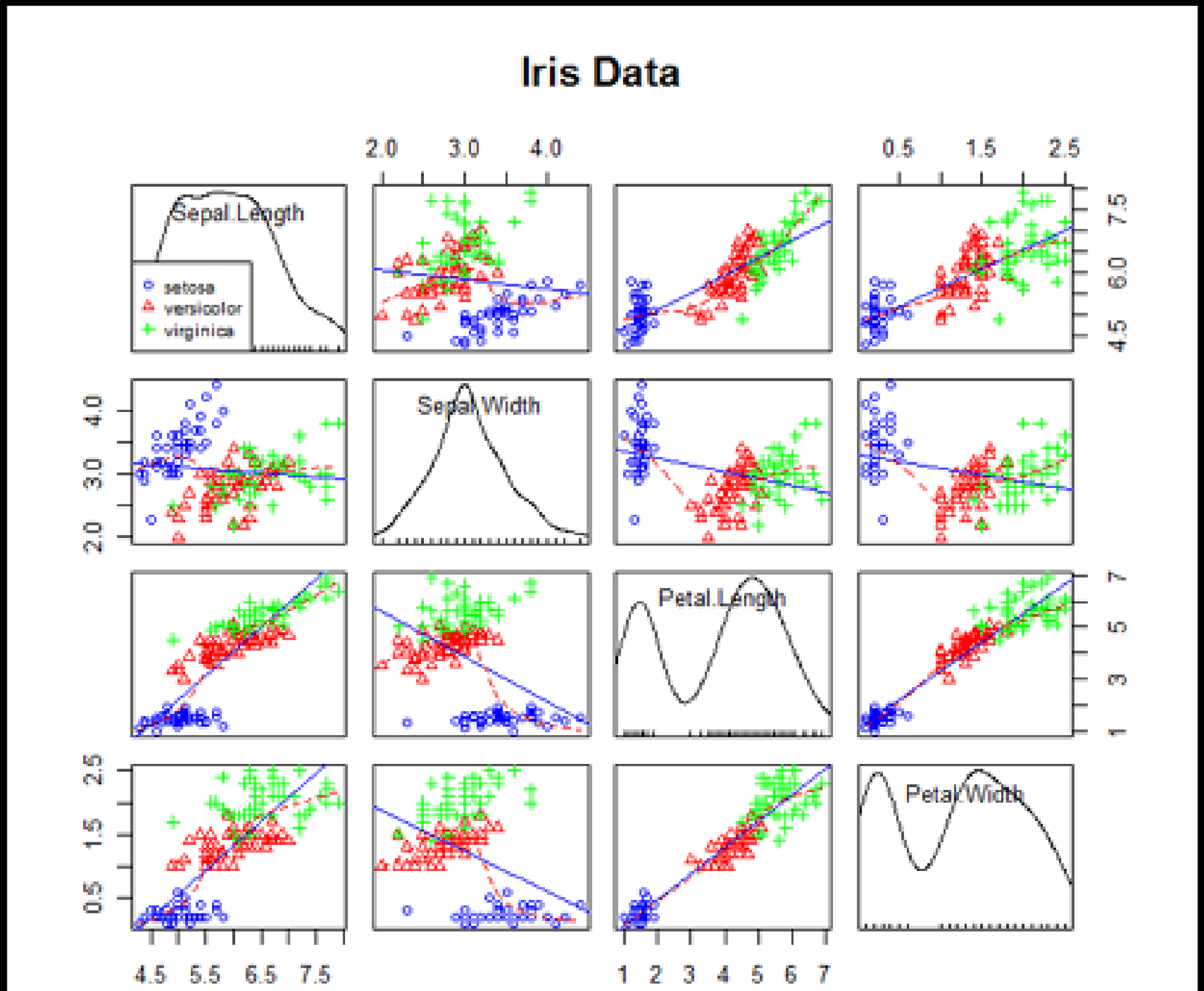




Principal Component Analysis from genetic data

Further Readings

www.statmethods.net/advgraphs/index.html



Non-graphical EDA

- Measures of central tendency
  - Mean, median, mode
- Measures of spread
  - Variability, variance, SD
- Shape of the distribution
- Outliers

Central Tendency

Mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Average Arithmetic mean

Central Tendency

Median vs Mean

3, 4, 6, 7, 8, 10, 15

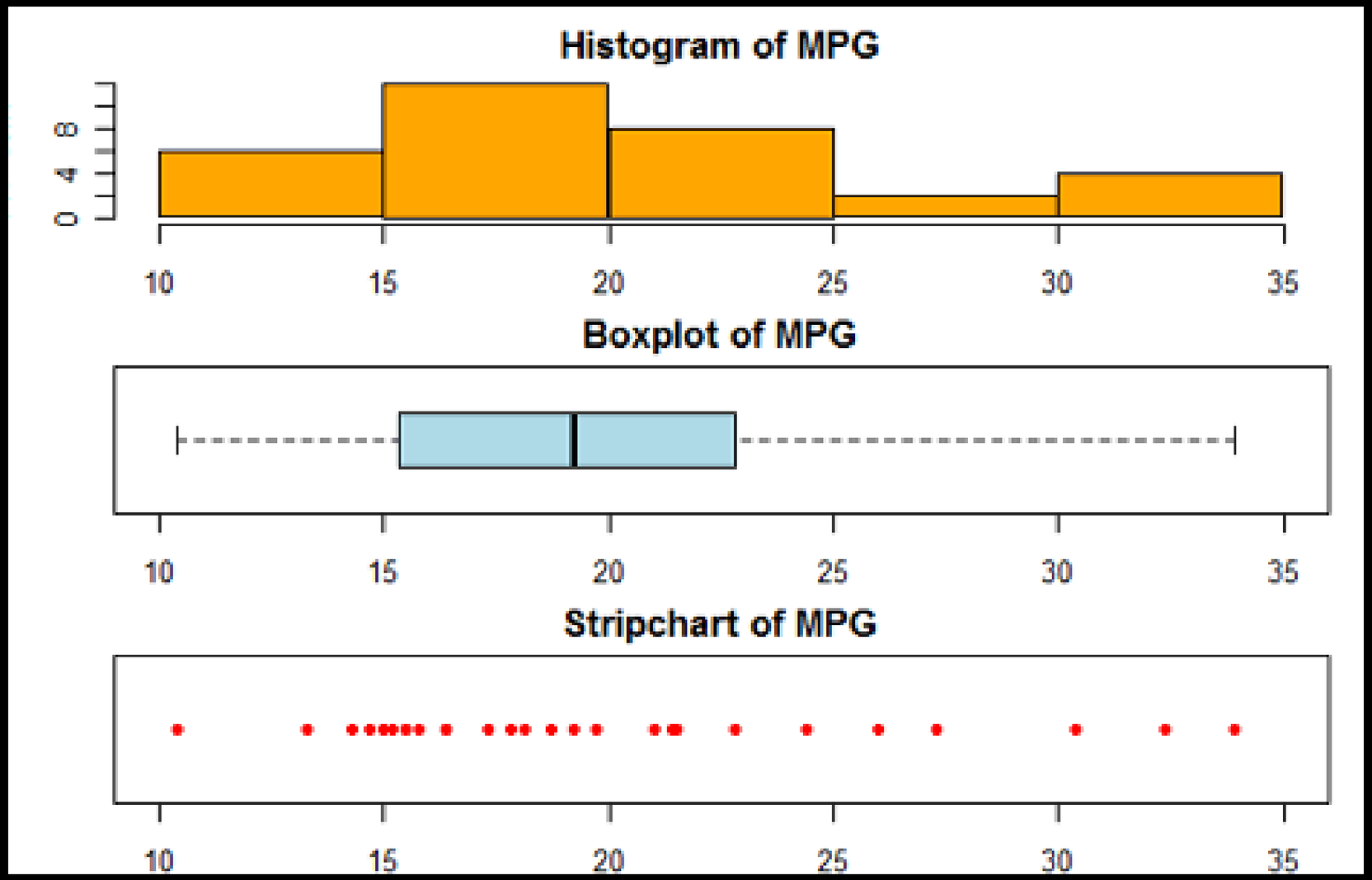
$$\bar{X} = 7.57$$

3, 4, 6, 7, 8, 10, 150

$$\bar{X} = 26.86$$

Measures of Spread

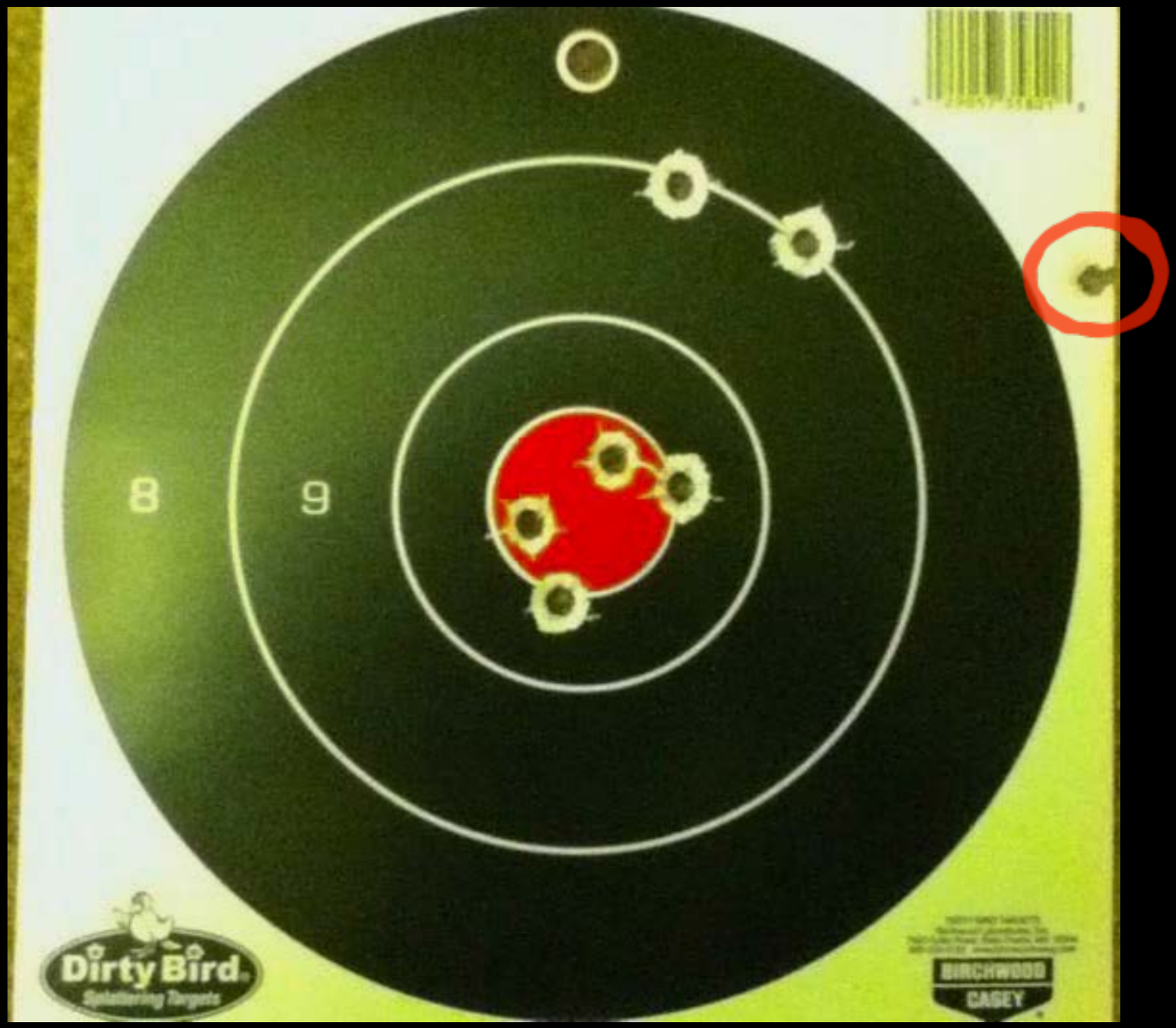
Dispersion, Variability

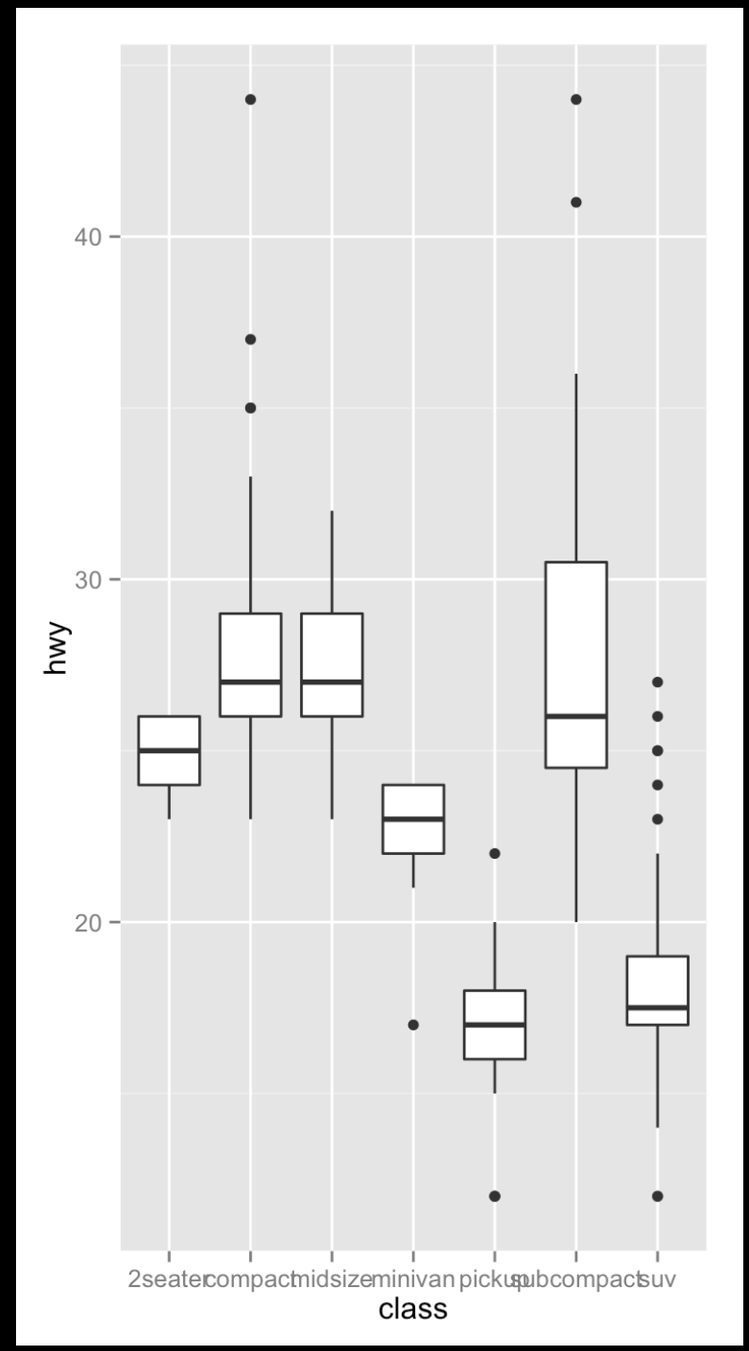


http://www.statmethods.net/graphs/index.html

Measures of Spread

Dispersion, Variability

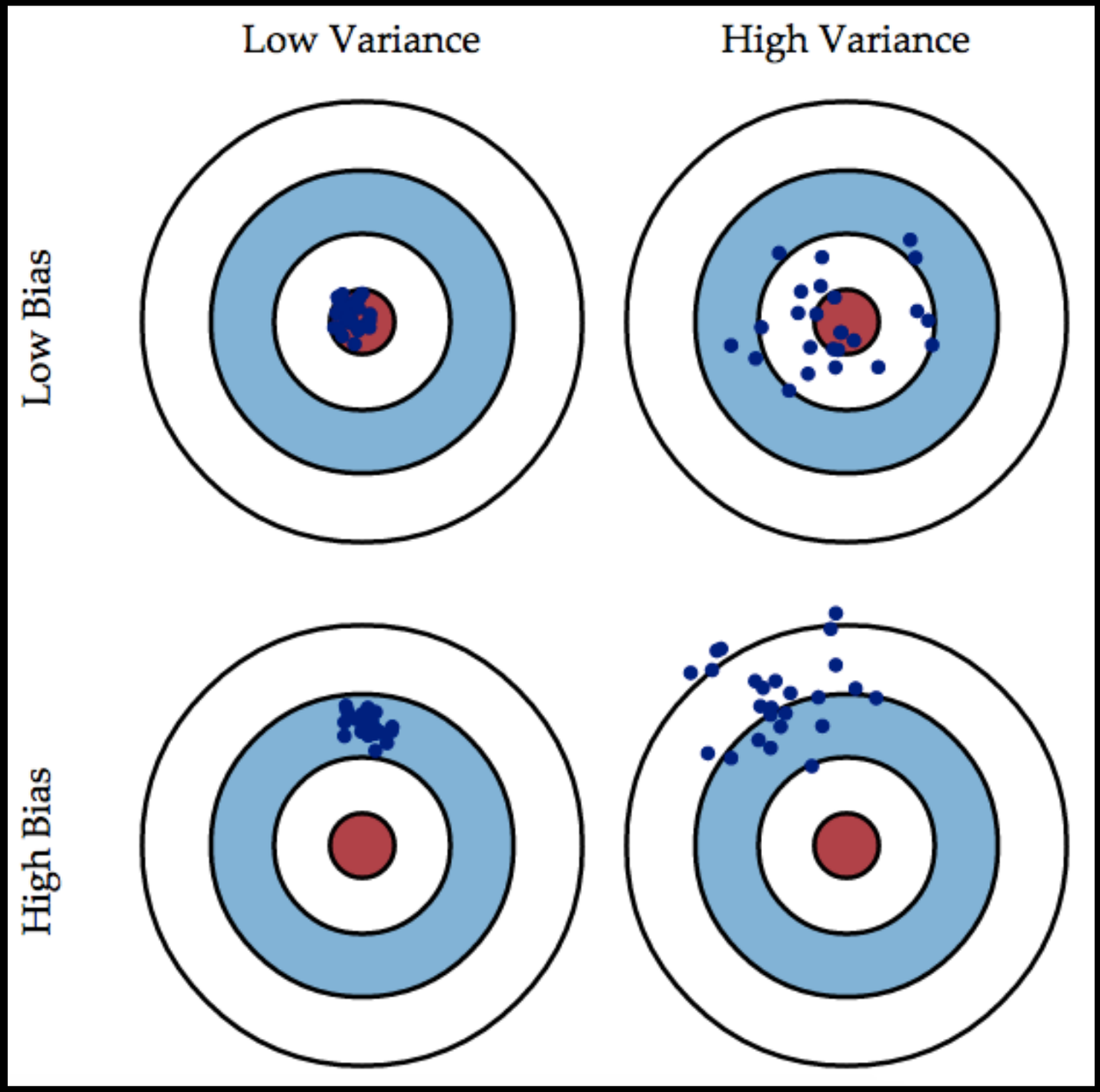




Measures of Spread

Precision = Low Variance

Accuracy
=
Unbias/Valid



Deviation

Deviation from the central tendency

$$X_i - \bar{X}$$

$$s^2 = \text{Variance} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n} \text{or } \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$s = \text{sqrt (Variance)} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Standard Deviation **(S.D.) measures variability**

Sample statistics vs Population statistics

s vs σ

an estimator from a sample vs a parameter of a population

$\widehat{\theta}$ is an estimator for θ

a hat or a caret denotes an estimator

Standard Error (S.E.)

- **a measure of the statistical accuracy of an estimate,** equal to the standard deviation of the theoretical distribution of a large population of such estimates.

- Not to confuse with S.D.
  BMJ. 2005 Oct 15; 331(7521): 903.

Standard Error of Mean (SEM)

$$SE_{\bar{x}} = \frac{SD}{\sqrt{n}}$$

a measure of precision of the sample mean

Can we use SE to replace SD?

So, if we want to say how widely scattered some measurements are, we use the standard deviation.

If we want to indicate the uncertainty around the estimate of the mean measurement, we quote the standard error of the mean.

For a large sample, a 95% confidence interval is obtained as the values 1.96×SE either side of the mean.

95% Confidence Interval (95%CI)

Current male smokers with an average daily dose of >30 cigarettes had

ORs of **103.5** **(95% CI 74.8-143.2)** for SqCC,

Sample Statistics

95%CI
Measure of Precision

Int J Cancer. 2012 Sep 1; 131(5): 1210–1219.

Other Measure of Spread

Range: min - max

Inter-Quartile Range: P_{25} - P_{75}

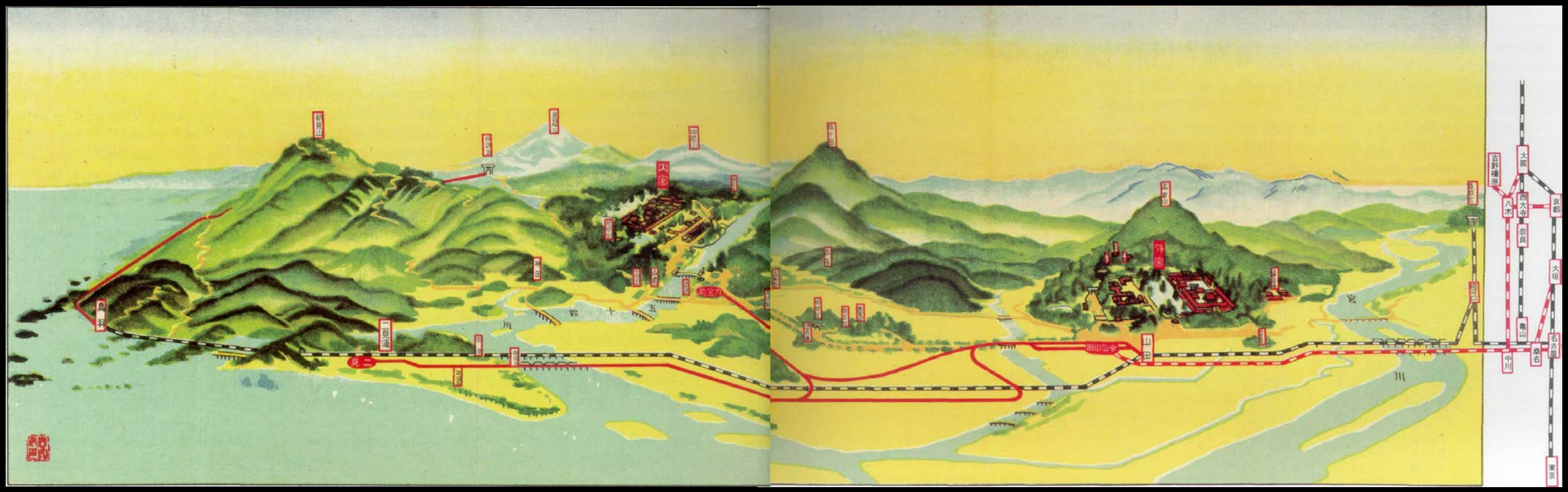
Edward Tufte

THE RENOWNED THEORIST of analytical design, Edward Tufte, was described by the ***New York Times*** as “the Leonardo da Vinci of data” for his pioneering work in the display and analysis of visual evidence.

He develops the fundamental theory of analytical design and proposes methods for display for nearly every type of evidence (time series, images, causal arrows, data tables, statistical graphics, public presentations).

http://www.skeptic.com/eskeptic/04-11-30/

Escaping the Flatland



Escaping this flatland is the essential task of envisioning information — for all the interesting worlds (physical, biological, imaginary, human) that we seek to understand are inevitably and happily multivariate in nature. Not flatlands.

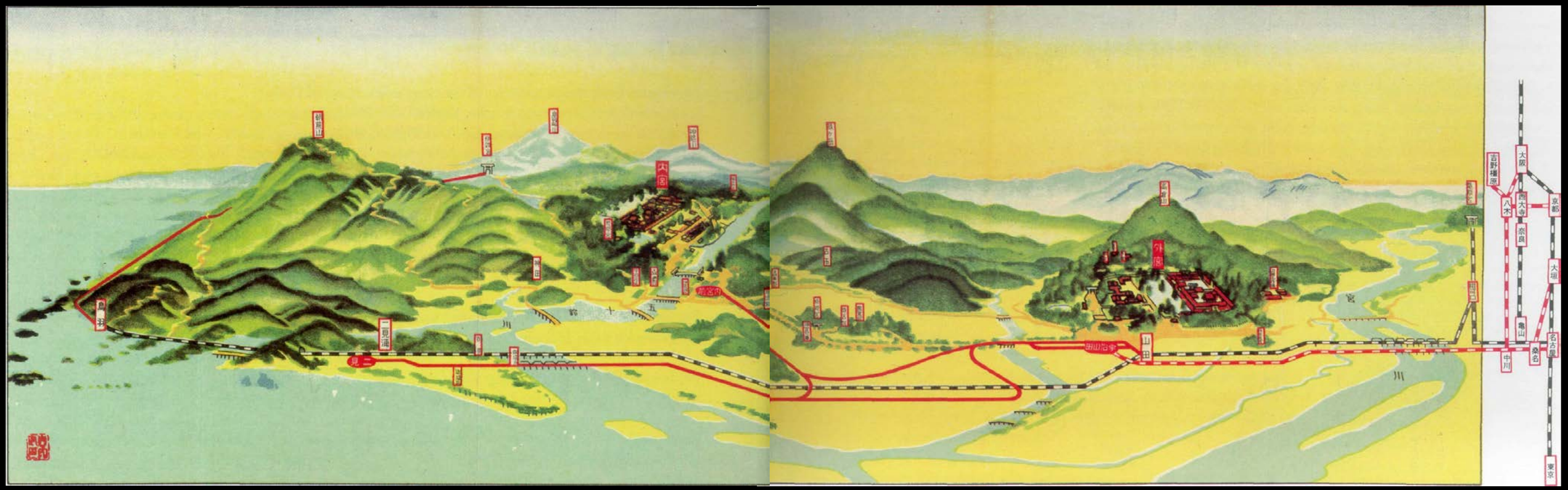
Guide for Visitors to Ise Shrine

(Ise, Japan; no date; published between October 1948 and April 1954, according to The Library, Ise Shrine, Mie Prefecture).

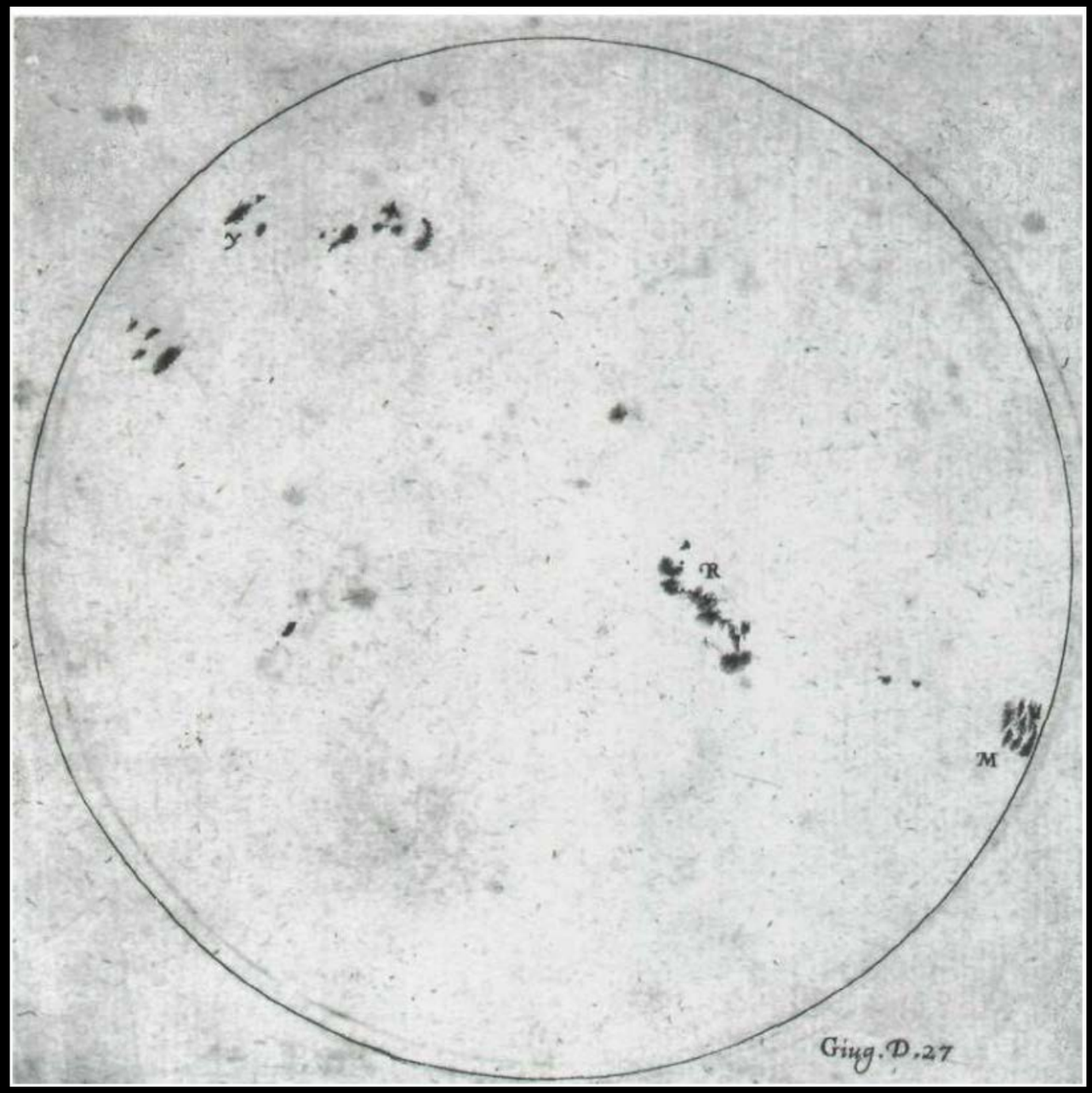
Escaping the Flatland

1.Increase the number of dimensions

2.Increase the data density



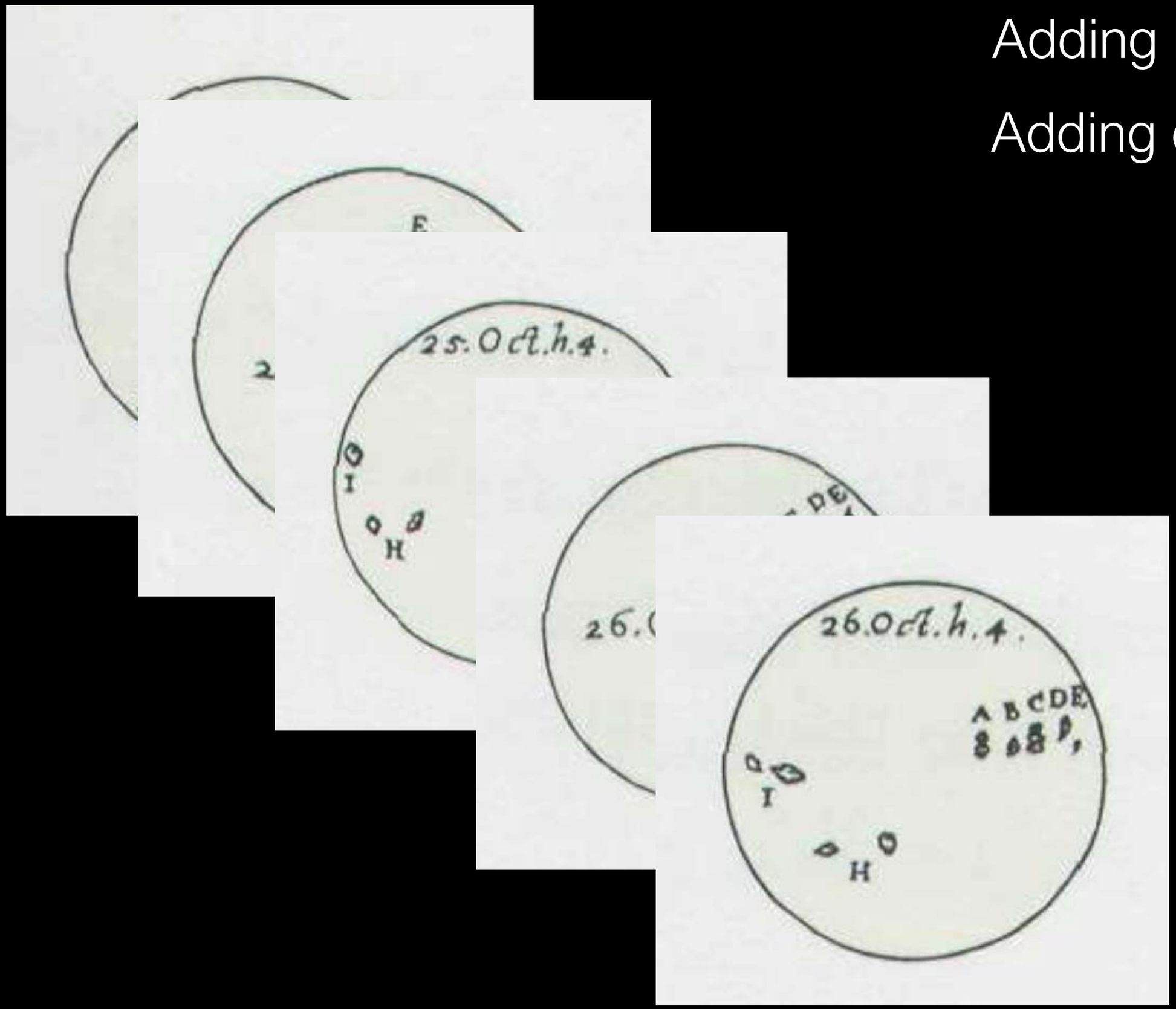
The Sunspots

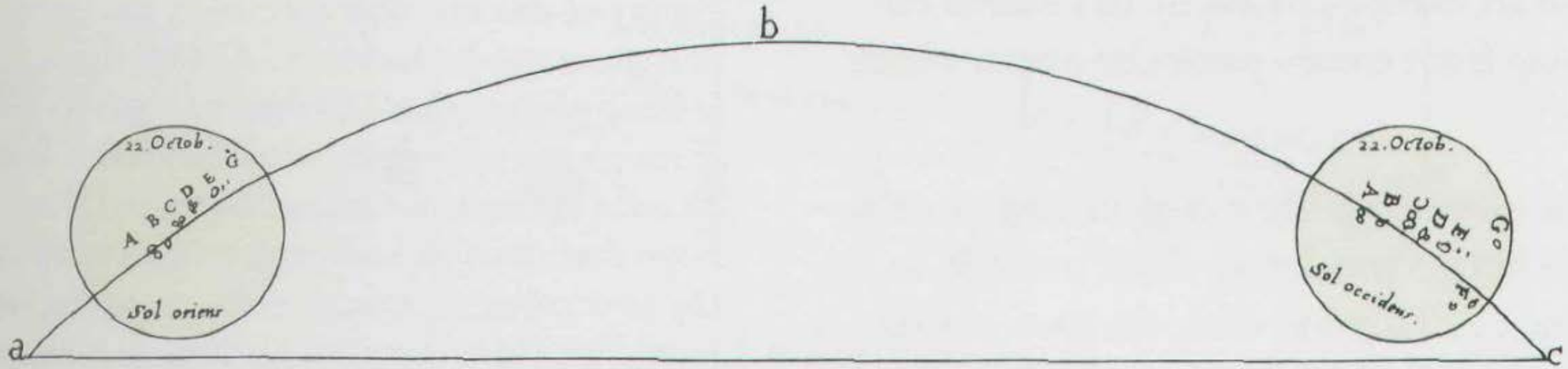


George Sarton, "Early Observations of the Sunspots," Isis, 37 (May 1947), 69-71

Adding Labels

Adding date & time





a c, horizon, a b c, arcus ſolis diurnus. Soloriens ex parte a, maculas exhibet quas vides, occidens verò c, easdem ratione primi motus, nonnihil inuertit. Et hanc matutinam vespertinamq; mutationem, omnes maculæ quotidie ſubeunt. Quod ſemel exhibuiſſe et monuiſſe, ſufficiat

Showing the sunspots location in 2D

- time
- labels
- orientation to the location in the sky

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 zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. 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'oeil 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.

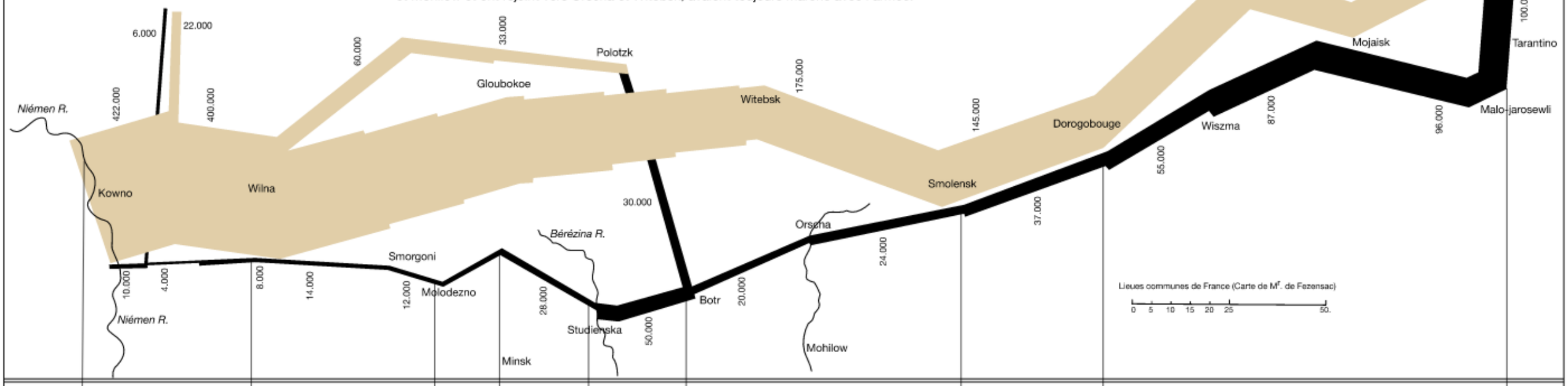
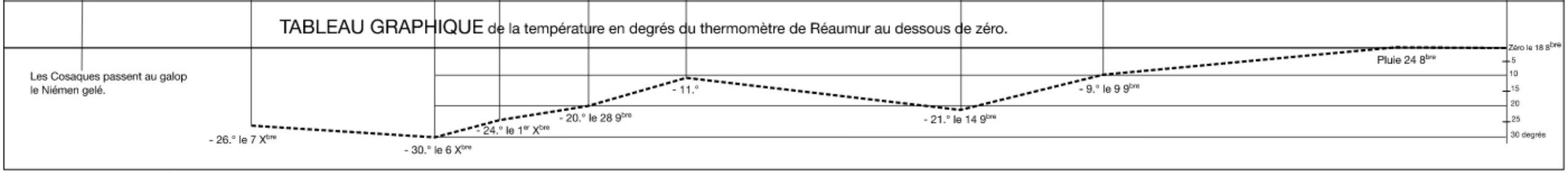


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.



Autog. par Regnier, 8. Pas. Sr Marie S' Gr à Paris.

[Vectorization CC-BY-SA **martingrandjean.ch** 2014]

Imp. Lith. Regnier et Dourdet.

Charles Joseph Minard's vectorized map (1869) displaying the movements and the number of Napoleonic troops during the Russian campaign (1812-1813), as well as the temperature on the return path.

Suggested Reading

- Tufte E. Envisioning Information. Cheshire, CT, USA: Graphics Press; 1990. (Chapter 1)

- Wickham H. ggplot2: Elegant Graphics for Data Analysis. 1st ed. 2009. Corr. 3rd printing 2010 edition. Springer; 2010. 213 p.