Week 9: Big Data

Slava Mikhaylov

POLS3003 Data Science and Big Data Analytics

Week 9 Outline

Big Data

A/B Testing

- Potential outcomes framework
- Regression analysis of experiments

Technologies

- Business intelligence
- Architecture

Visualization

Big Data

From Wiki

- Volume: big data doesn't sample; it just observes and tracks what happens
- Velocity: big data is often available in real-time
- Variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion
- Machine Learning: big data often doesn't ask why and simply detects patterns
- Digital footprint: big data is often a cost-free byproduct of digital interaction

Components of Big Data

McKinsey report:

- Techniques for analyzing data (A/B testing, machine learning, natural language processing)
- Big Data technologies, like business intelligence, cloud computing and databases
- Visualization.

A/B Testing

A/B Testing

- This is simply experimental and quasi-experimental research design to test an intervention (or several interventions) against a control group.
- The key interest is in identifying the causal effect of an intervention.

Do hospitals make people healthier?

The National Health Interview Survey (NHIS): "During the past 12 months, was the respondent a patient in a hospital overnight?", "Would you say your health in general is excellent, very good, good, fair, poor?"

- Simple analysis of data suggests that hospitals make people sicker. But people who go to the hospital are probably less healthy to begin with.

Do hospitals make people healthier? Formalized

- Hospital treatment is described by a binary random variable $D_i = 0, 1$
- Health status (outcome) is Y_i
- Is Y_i affected by D_i?

Potential outcomes framework

- In idealized world we imagine what might have happened to a person who went to the hospital if he/she had not gone, and vice versa.
- Potential outcomes model of causality: Jerzy Neyman (1923) and Donald Rubin (1978).

Rubin (1974):

Intuitively, the causal effect of one treatment, E, over another, C, for a particular unit and an interval of time from t_1 to t_2 is the difference between what would have happened at time t_2 if the unit had been exposed to E initiated at t_1 and what would have happened at t_2 if the unit had been exposed to C initiated at t_1: 'If an hour ago I had taken two aspirins instead of just a glass of water, my headache would now be gone,' or 'because an hour ago I took two aspirins instead of just a glass of water, my headache is now gone.' Our definition of the causal effect of the E versus C treatment will reflect this intuitive meaning.

Things to note:

- Potential outcomes and covariates are fixed for each i. Treatments and response indicators are stochastic.
- Effects are defined by letting only treatments vary, not units.
- Thus, causal effects are defined only for units that can conceivably receive different treatment values.
- The test for the above is "manipulation" (Holland,1986).

- Holland (1986) : "For causal inference, it is critical that each unit be potentially exposable to any one of the causes."
- Angrist and Krueger (1999): "The problem of ambiguous counterfactuals is typically resolved by focusing on hypothetic manipulations in the world as is."

Formalizing potential outcomes framework

For any individual there are two potential health outcomes:

$$\text{Potential outcome} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

The difference between two potential outcomes Y_{1i} and Y_{0i} is the causal effect of going to the hospital for individual i.

Observed outcome as a combination of potential outcomes

$$Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} = Y_{0i} + (Y_{1i} - Y_{0i})D_i \qquad (1)$$

- where $(Y_{1i} - Y_{0i})$ is the causal effect of hospitalization for an individual.
- Since we never observe both potential outcomes for any one person, we must compare the average health effects for two groups (hospitalized and not).

Fundamental problem of causal inference (Holland, 1986):

For each i potential outcomes for all treatments exist, but we only observe the potential outcome for the treatment value that i receives.

- "Scientific solution": Use theory to determine when units are interchangeable.
- "Statistical solution": Study averages.

The comparison of average health conditional on hospitalization status:

$$\underbrace{E[Y_i|D_i=1]-E[Y_i|D_i=0]}_{\text{Observed difference in average health}} = \underbrace{E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=1]}_{\text{Average treatment effect on the treated (ATT)}} \underbrace{+E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]}_{\text{Selection bias}}.$$

where $E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=1]=E[Y_{1i}-Y_{0i}|D_i=1]$ is the average causal effect of hospitalization on those who were hospitalized.

Selection bias and random assignment

Selection problem is solved via random assignment because it makes D_i independent of potential outcomes.

$$\begin{aligned} E[Y_i|D_i=1] - E[Y_i|D_i=0] &= E[Y_{1i}|D_i=1] - E[Y_{0i}|D_i=0] \\ &= E[Y_{1i}|D_i=1] - E[Y_{0i}|D_i=1] \\ &= E[Y_{1i} - Y_{0i}|D_i=1] \\ &= E[Y_{1i} - Y_{0i}]. \end{aligned}$$

by virtue of independence of Y_{0i} and D_i we swap $E[Y_{0i}|D_i=1]$ for $E[Y_{0i}|D_i=0]$ in the second line.

Solving the most important problem of empirical research

Random assignment of D_i eliminates selection bias

Regression analysis of experiments

Assume that the treatment effect is constant (the same for all individuals), $y_{1i} - y_{0i} = \rho$. We can rewrite (1) as

$$Y_i = \underbrace{\alpha}_{E[Y_{0i}]} + \underbrace{\rho}_{(Y_{1i}-Y_{0i})} D_i + \underbrace{\eta_i}_{Y_{0i}-E(Y_{0i})}, \quad (2)$$

where η_i is the random part of Y_{0i}.

Evaluating conditional expectation under $D_i = 0, 1$

$$E[Y_i|D_i = 1] = \alpha + \rho + E[\eta_i|D_i = 1]$$
$$E[Y_i|D_i = 0] = \alpha + E[\eta_i|D_i = 0],$$

so that

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \underbrace{\rho}_{\text{Treatment effect}} + \underbrace{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]}_{\text{Selection bias}}.$$

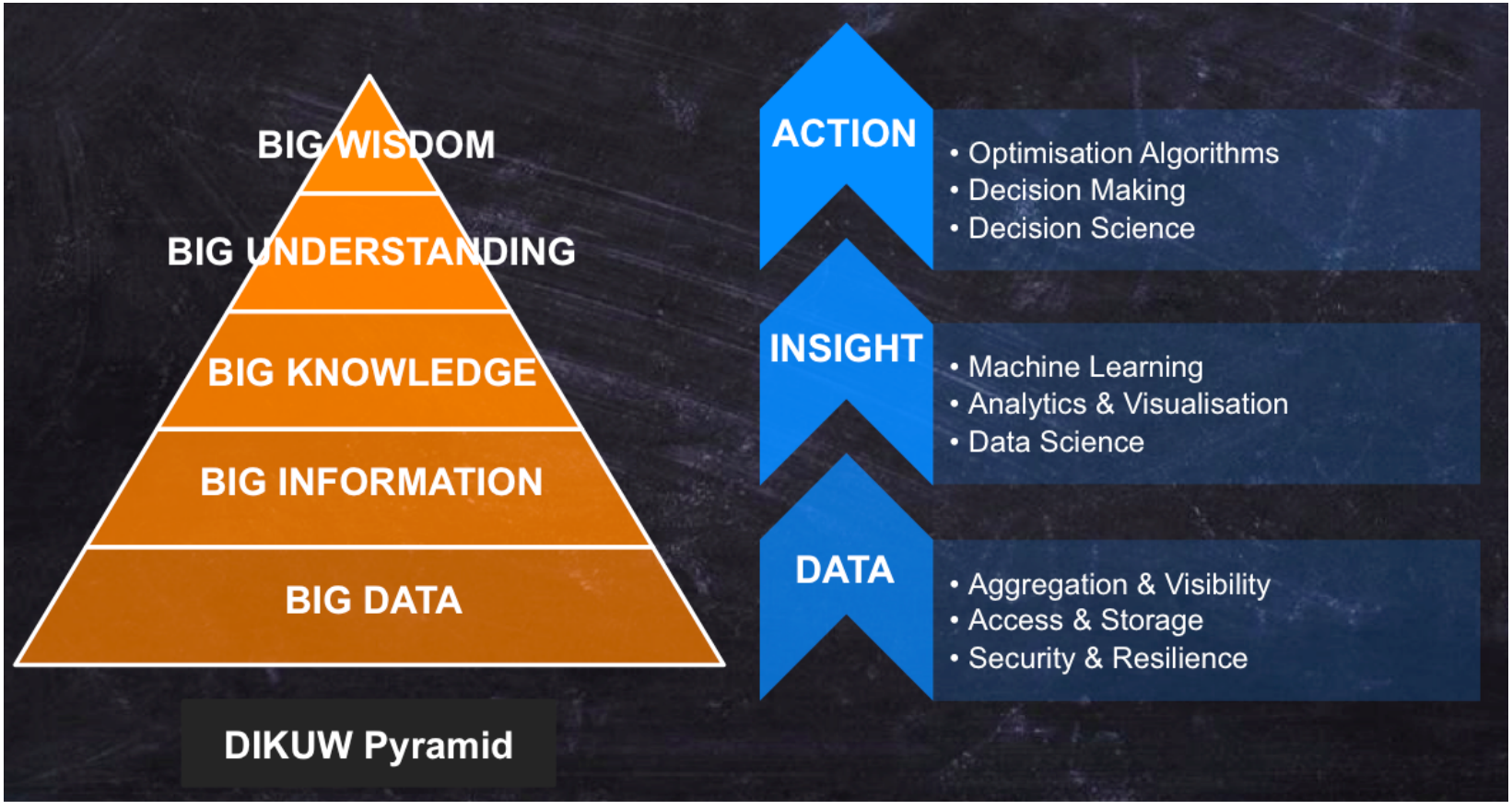
- Selection bias is the correlation between the regression error, η_i, and the regressor, D_i.
- Since,

$$E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0] = E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0],$$

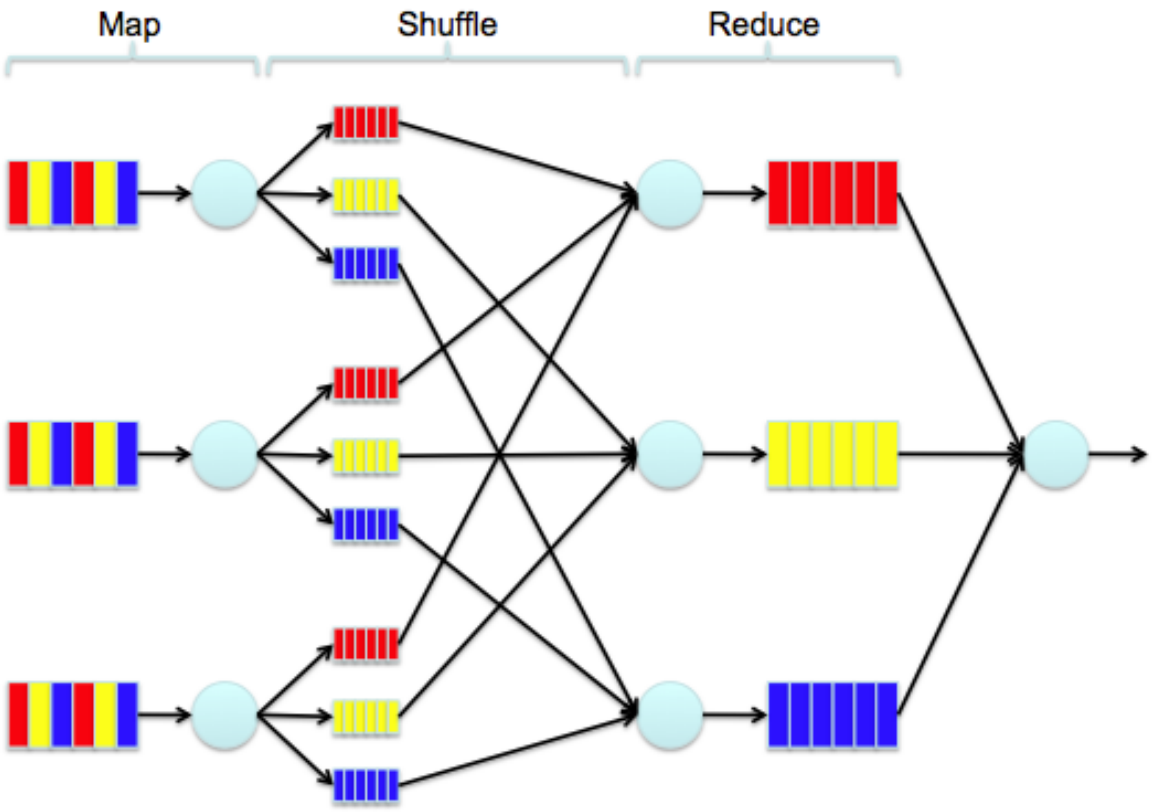
this correlation reflects the difference in (no-treatment) potential outcomes between those who get treated and those who don't.

Technologies

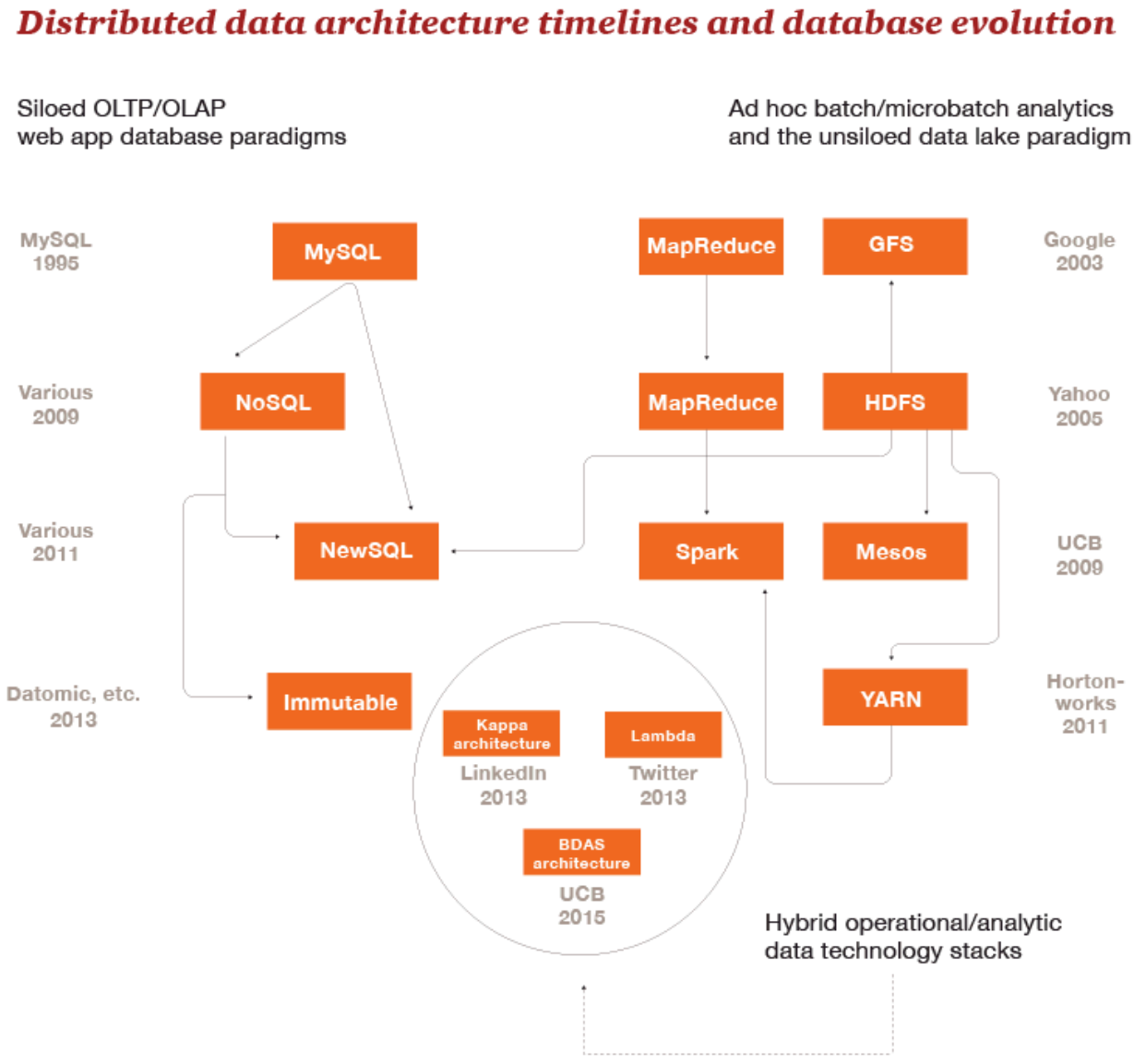
Business intelligence



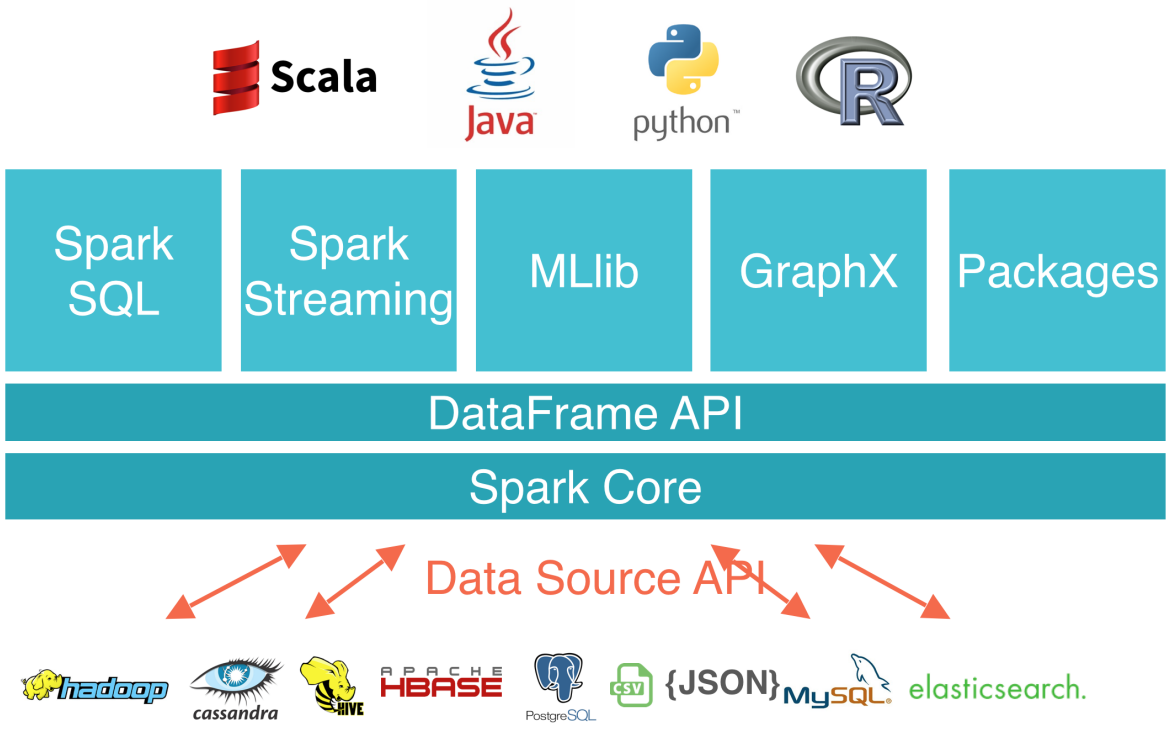
MapReduce



Architecture evolution



Typical data analytic stack



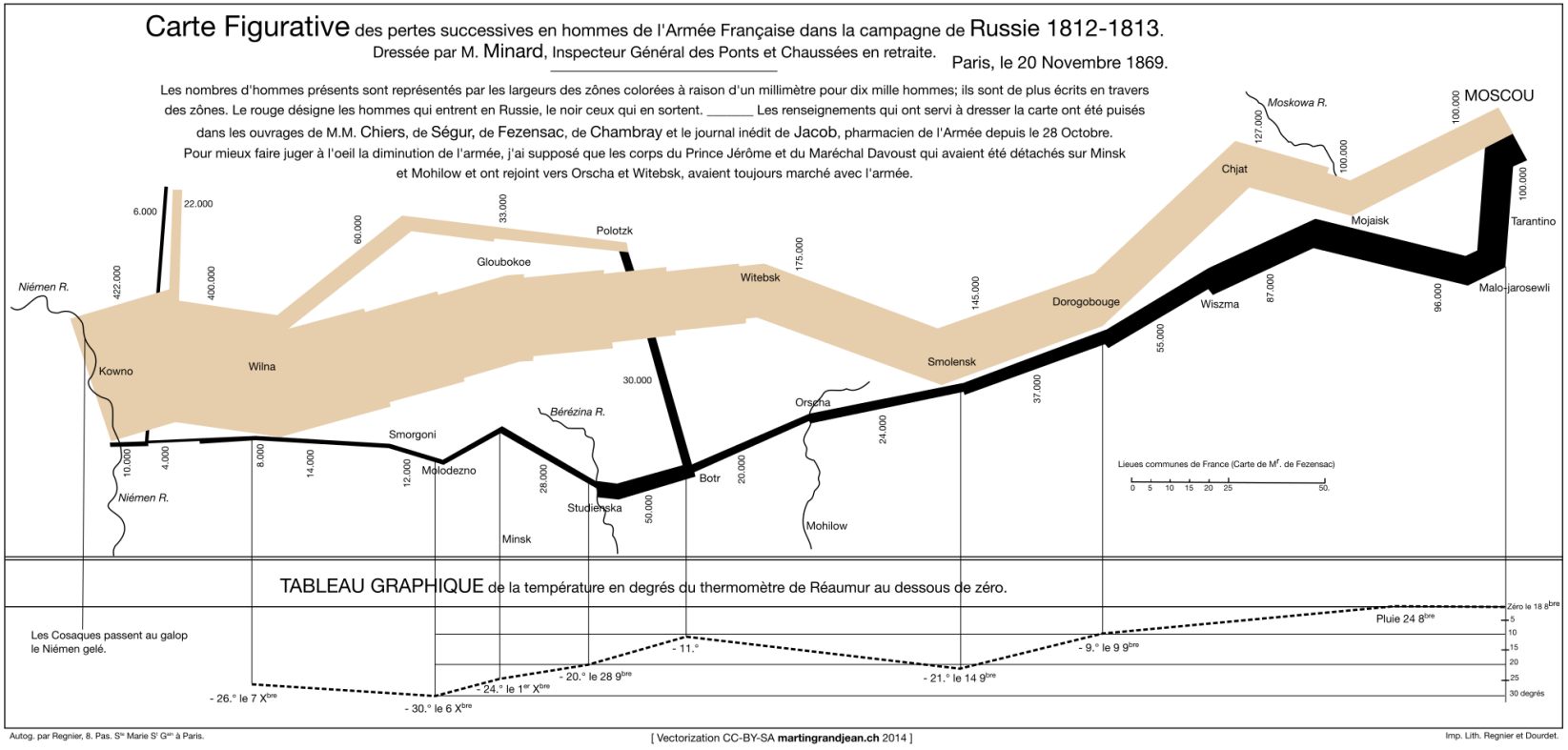
Visualization

Data visualization

Edward Tufte's "The Visual Display of Quantitative Information":

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
- avoid distorting what the data has to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

Charles Joseph Minard's Napoleon's March diagram



Autog. par Regnier, 8. Pas. Ste Marie S' Gen à Paris.

Imp. Lith. Regnier et Dourdet.