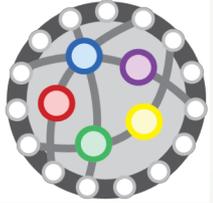


ELAC 2019 INTRODUCTION TO DATA SCIENCE

Day 3

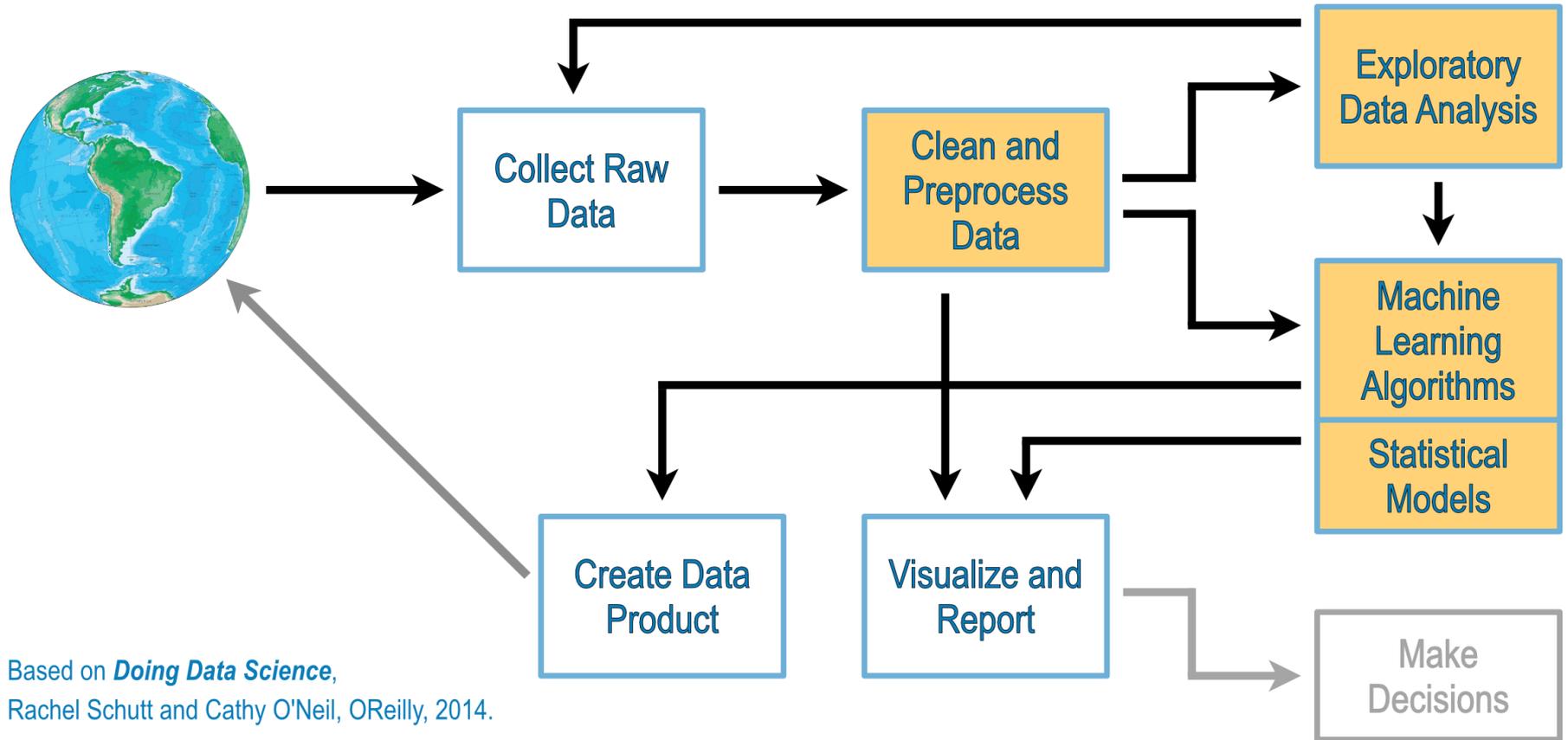
Rafael Santos - rafael.santos@inpe.br
www.lac.inpe.br/~rafael.santos/talks.html

Introduction to Data Science

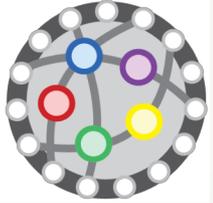


About this Lecture

Where are we?



Introduction to Data Science



Exploratory Data Analysis

Exploratory Data Analysis

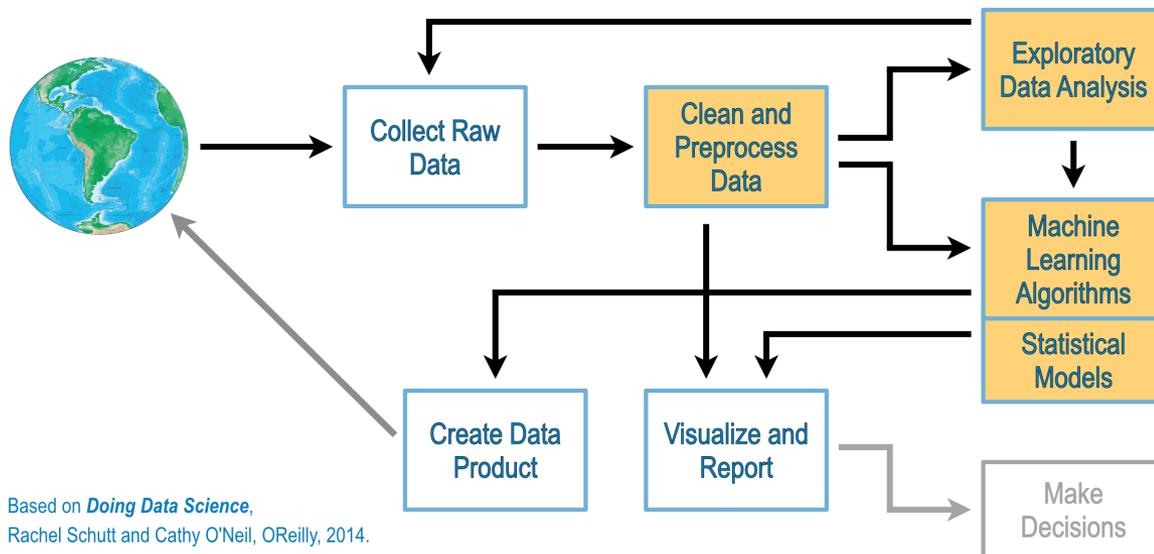
- “*Exploratory data analysis*” is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there.

— John Tukey

- Contrast it with *Confirmatory Data Analysis*, in which we have a hypothesis or model and try to confirm or deny it.

Exploratory Data Analysis

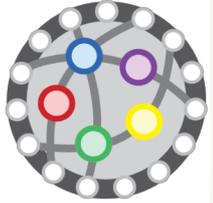
- Basic tools: graphs, plots, basic statistics.
 - ▣ Explore and describe data and relations.
 - ▣ Gain intuition about the data.
 - ▣ Change, add, transform variables.
 - ▣ Eventually *go back*.



Exploratory Data Analysis: Steps

- Load the data. Make sure it is **tidy**.
- Get basic statistics about the variables.
- Create new variables (segmentation, discretization, comparison).
- Combine existing variables (ratio).
- Explore relations between variables.
- Plot the data.
- Document what you've found (even what you *think* you've found).

Introduction to Data Science



EDA in R

“R Programming”

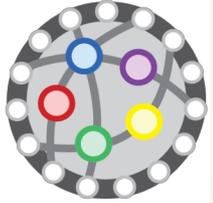


EDA in R (Part II)

- Let's switch to a browser:

<http://www.lac.inpe.br/~rafael.santos/r.html>

Introduction to Data Science



Visualization in R

Visualization in R

- Way too much ~~complex~~ flexible!
- Three approaches to graphics:
 - ▣ Base R graphics
 - ▣ **lattice** package
 - ▣ **ggplot2** package

Base R Graphics

□ Pros:

- Good documentation and examples.
- Customizable.

□ Cons:

- Customization may be hard work.

Lattice package

□ Pros:

- Multi-panel plotting made easy!
- Easy to add data summaries.

□ Cons:

- Monolithic style.
- Extreme customization is hard.

ggplot2 package

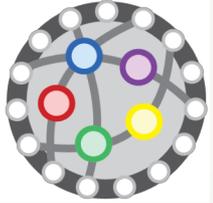
□ Pros:

- Grammar of Graphs!
- Prefers data frames.
- Multi-panel plotting made easy (facets)!
- Use aesthetics to represent variables.
- Combinations of layers.

□ Cons:

- Grammar of Graphs?
- Monolithic style.
- Extreme customization is hard.

Introduction to Data Science

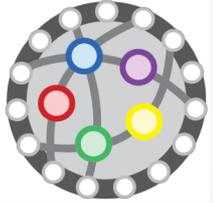


References

References

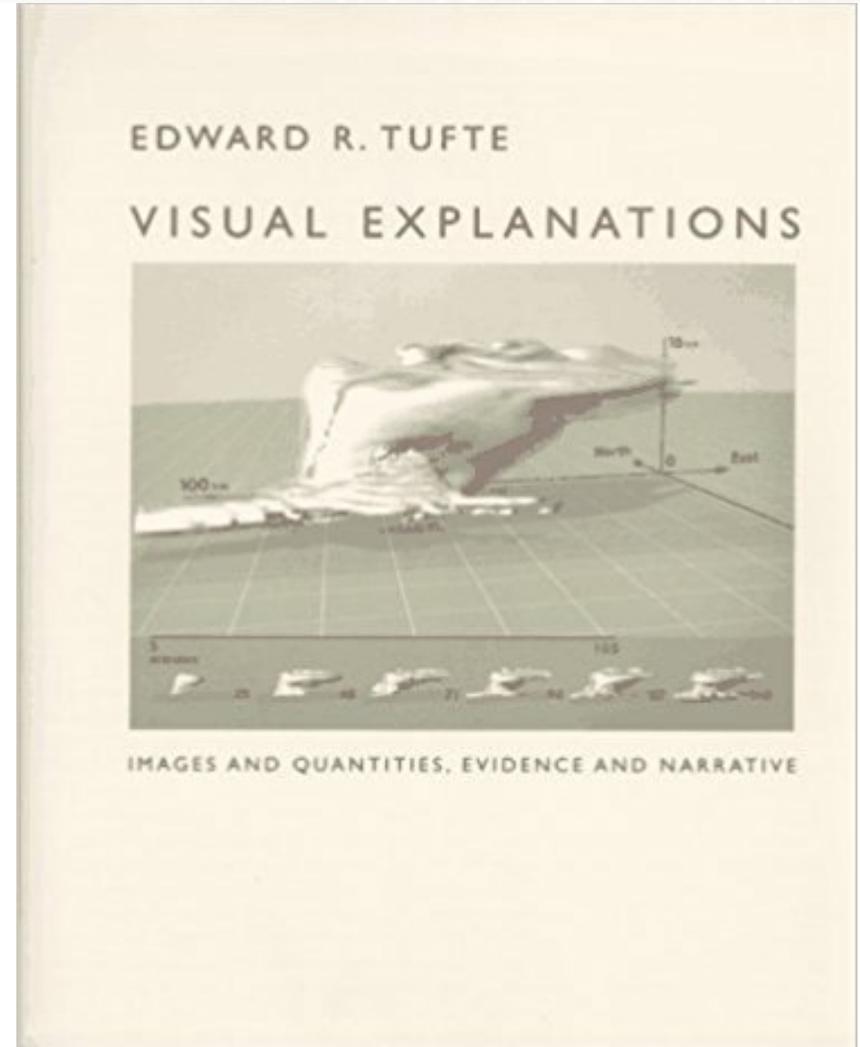
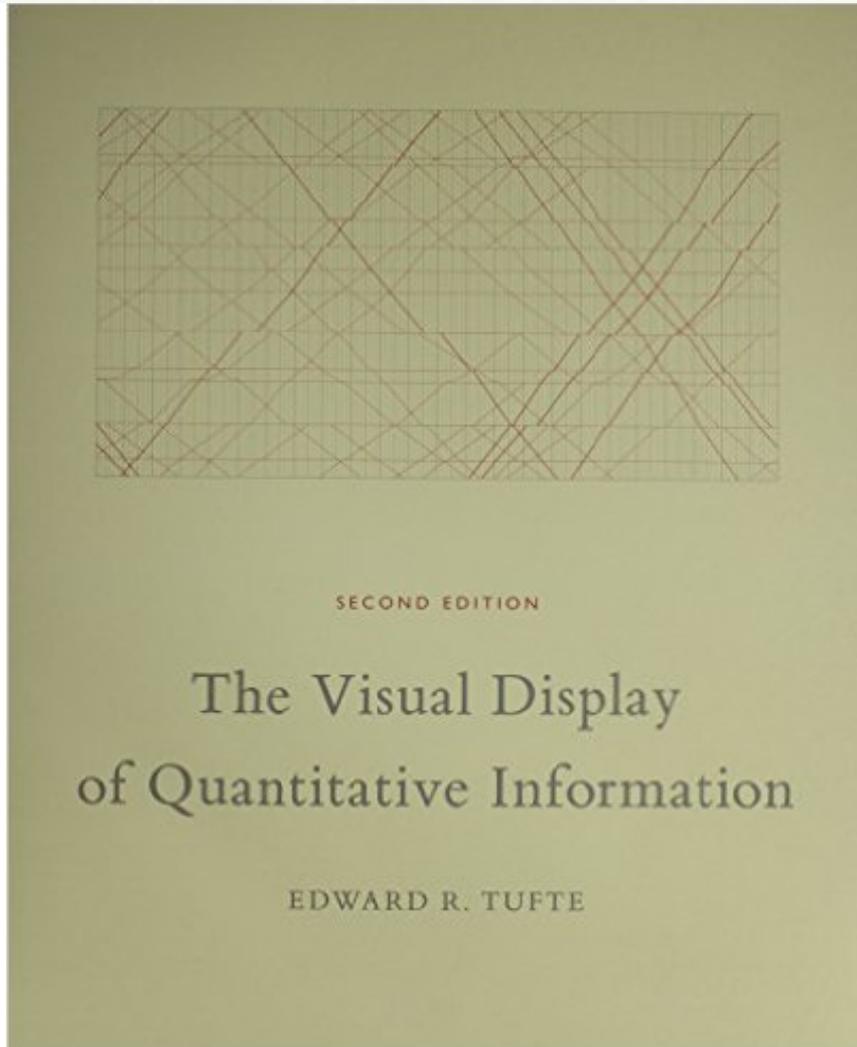
- Motivation: *Doing Data Science*, Rachel Schutt and Cathy O'Neil, O'Reilly, 2014
- Lots of code from StackOverflow!

Introduction to Data Science

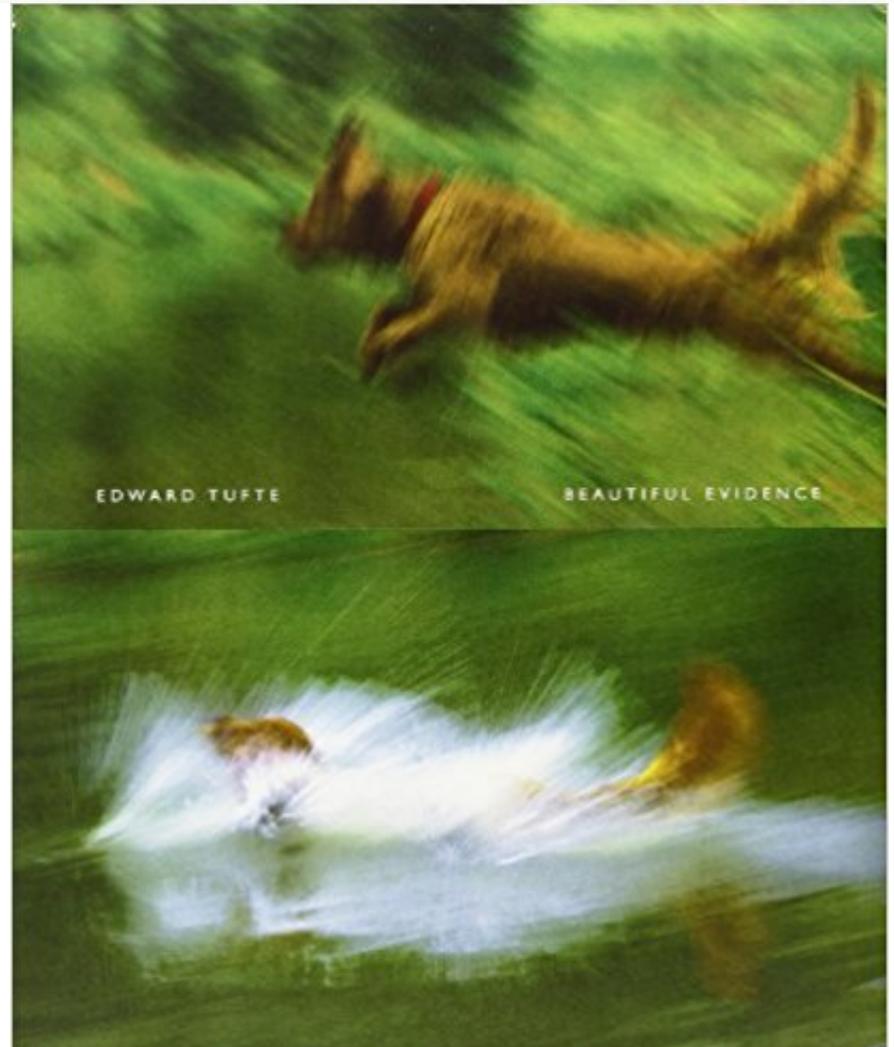
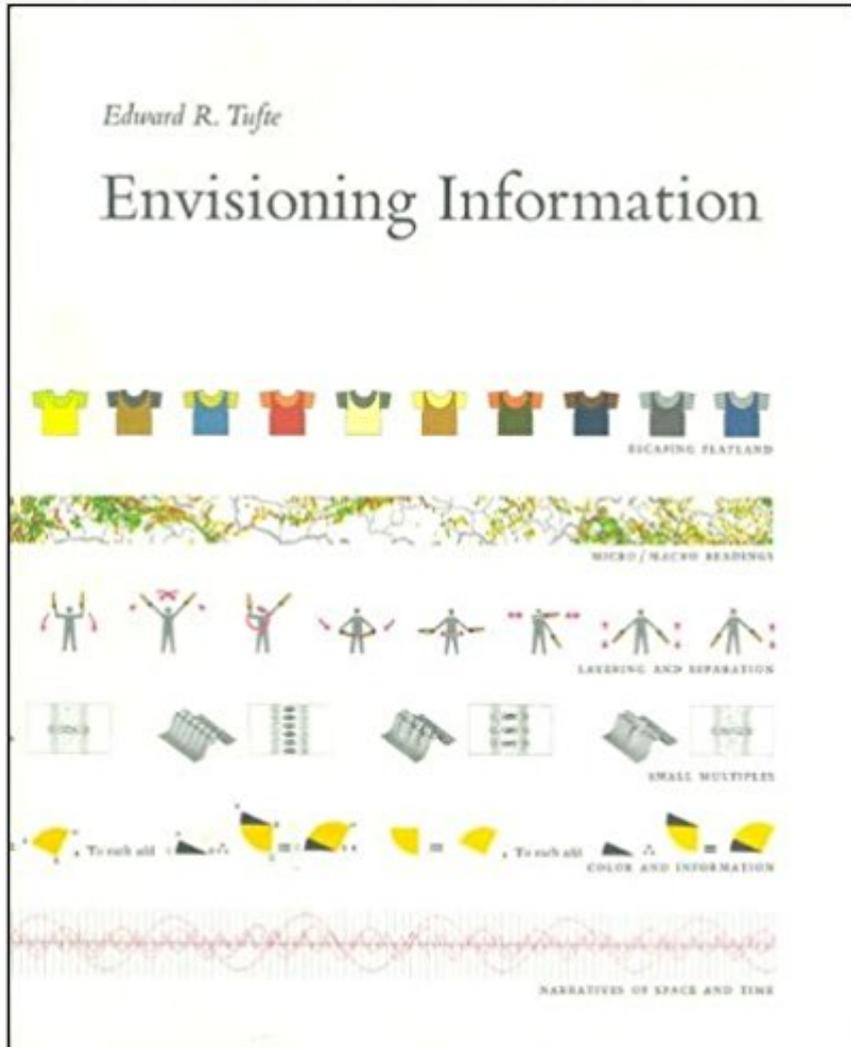


Digression: Edward Tufte and Visualization

References



References



Edward Tufte and Super Graphics

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 ont été en Russie, le noir ceux qui en sont restés. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Légras, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 23 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 Davout, qui avaient été détachés sur Minsk et Mohilow et qui rejoignirent Orscha et Wilna, avaient toujours marché avec l'armée.

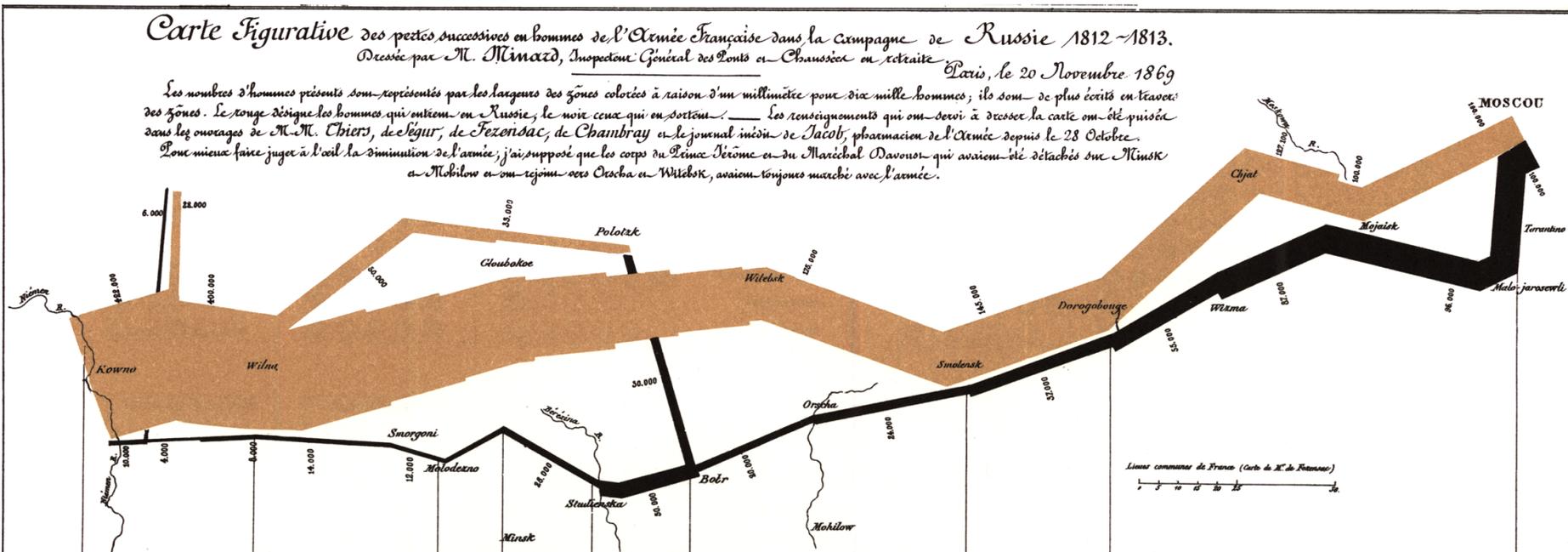
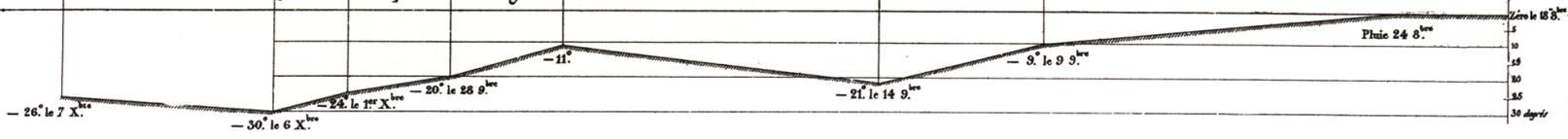


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

Les Cosaques passent au galop le Niémen gelé.



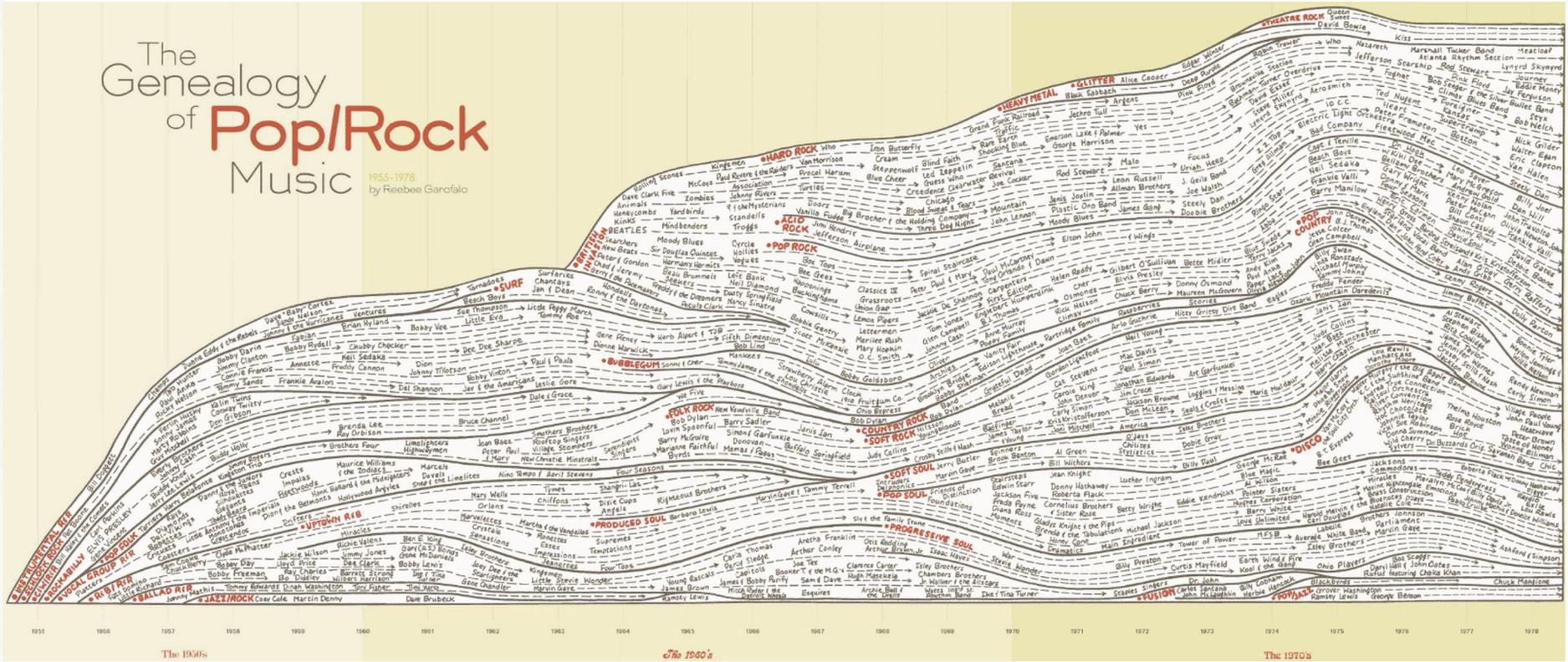
Atting, par Regnier, à Par. 5^{me} Marie 52 6^{me} à Paris.

Imp. Lith. Regnier et Deroyat.

Edward Tufte and Super Graphics

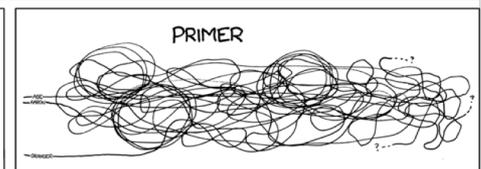
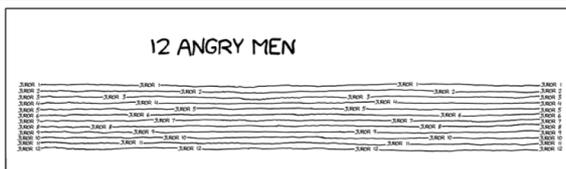
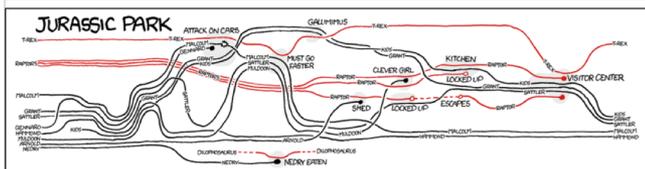
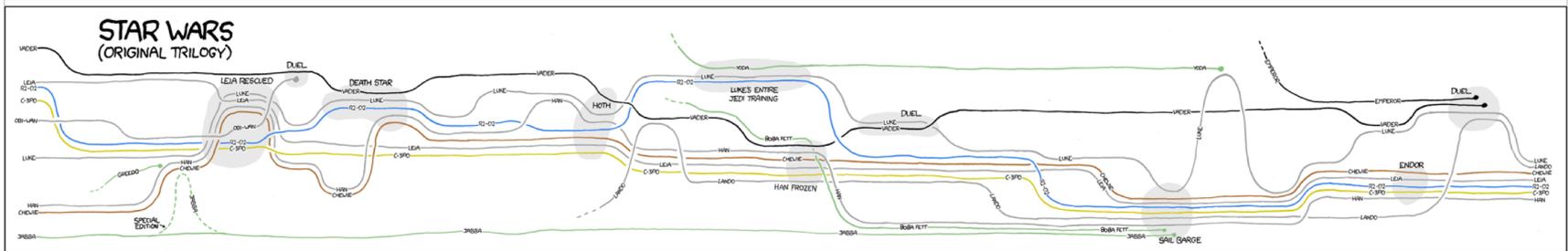
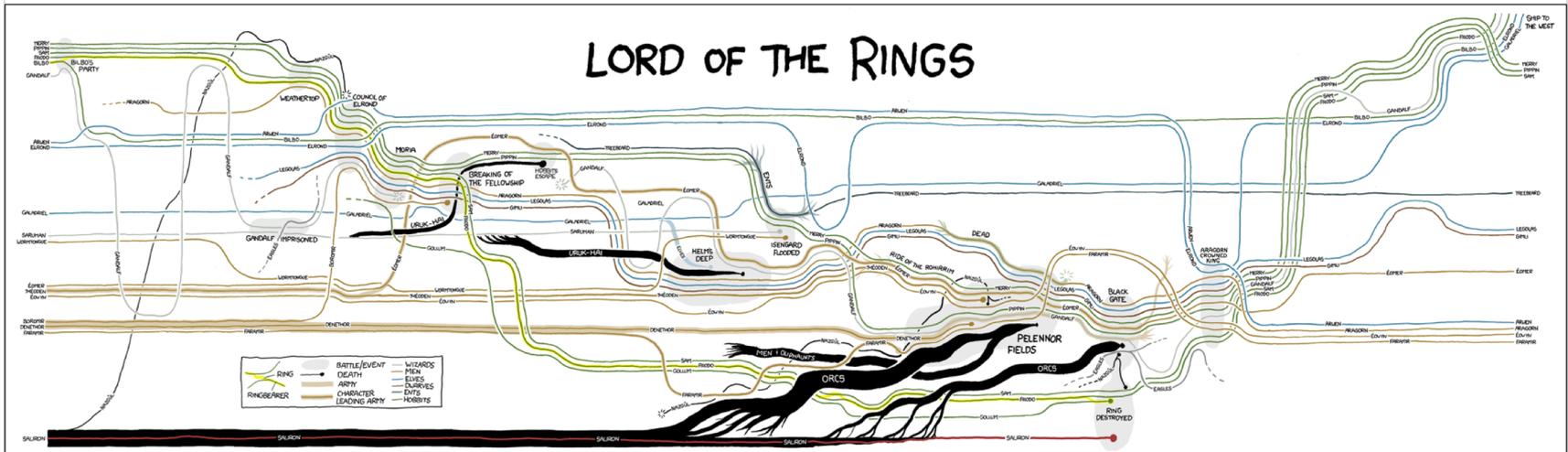
The Genealogy of Pop/Rock Music

1955-1978
by Reebec Garofalo



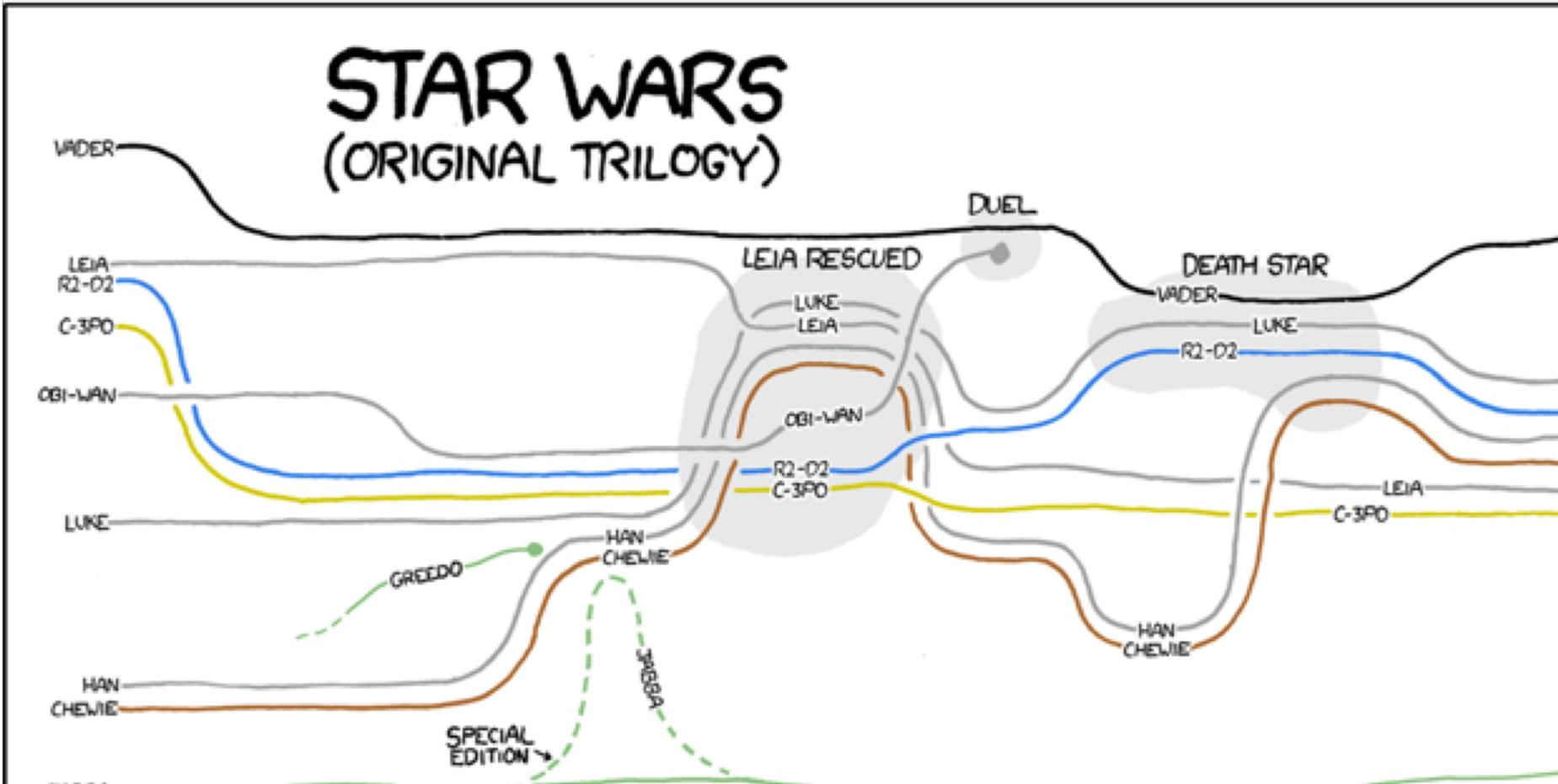
XKCD and Super Graphics

THESE CHARTS SHOW MOVIE CHARACTER INTERACTIONS. THE HORIZONTAL AXIS IS TIME. THE VERTICAL GROUPING OF THE LINES INDICATES WHICH CHARACTERS ARE TOGETHER AT A GIVEN TIME.



XKCD and *Super Graphics*

STAR WARS (ORIGINAL TRILOGY)



Shameless Advertising

- Applied Computing Graduate Program at INPE:
 - http://www.inpe.br/pos_graduacao/cursos/cap/
- CAP's Annual Workshop (September 2019):
 - <http://www.inpe.br/worcap/>
- Grants!
 - <http://www.inpe.br/bolsas> (IC)
 - <http://www.inpe.br/pci>

rafael.santos@inpe.br