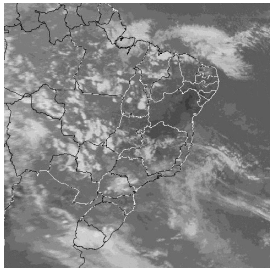
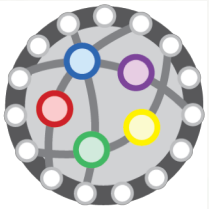


ELAC 2019
INTRODUCTION TO DATA SCIENCE

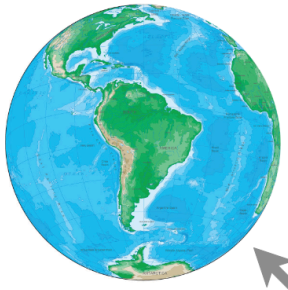
Day 3

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



About this Lecture

Where are we?



Collect Raw Data

Clean and Preprocess Data

Exploratory Data Analysis

Machine Learning Algorithms

Statistical Models

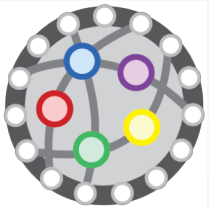
Create Data Product

Visualize and Report

Make Decisions

Based on ***Doing Data Science***,
Rachel Schutt and Cathy O'Neil, OReilly, 2014.

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.*

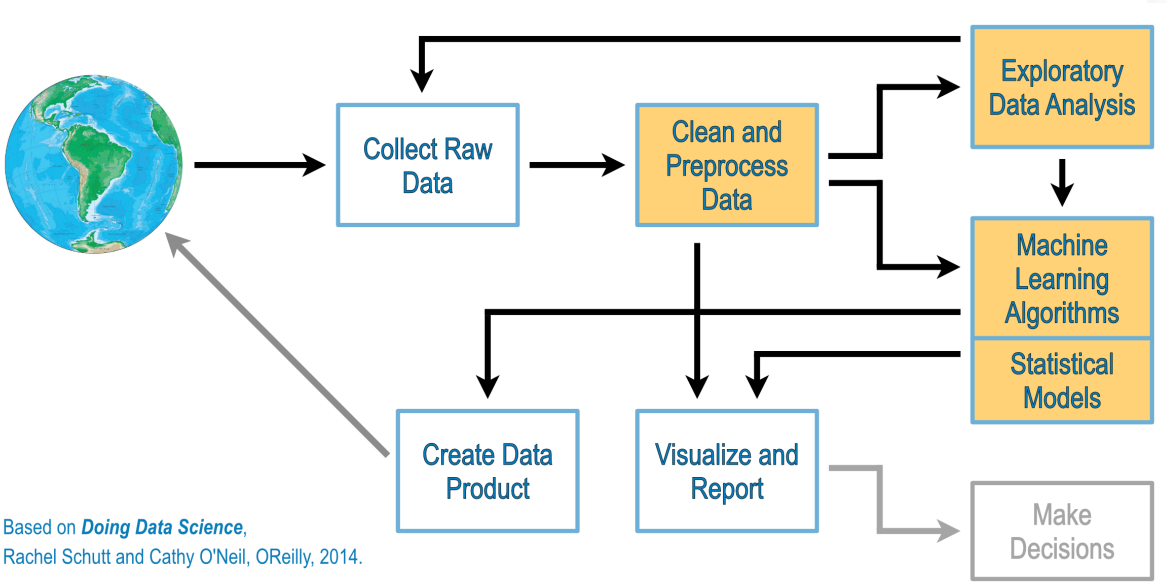
  — John Tukey

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

Doing Data Science, Rachel Schutt and Cathy O'Neil, OReilly, 2014

Exploratory Data Analysis

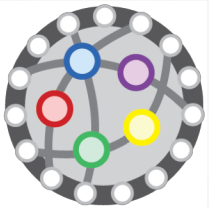
- Basic tools: graphs, plots, basic statistics.
  - Explore and describe data and relations.
  - Gain intuition about the data.
  - Change, add, transform variables.
  - Eventually *go back*.



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



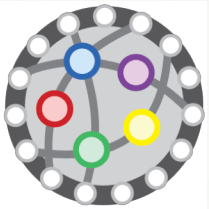
EDA in R

“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

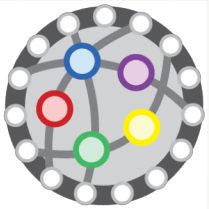
- Pros:
  - Good documentation and examples.
  - Customizable.
- Cons:
  - Customization may be hard work.

Lattice package

- Pros:
  - Multi-panel plotting made easy!
  - Easy to add data summaries.
- Cons:
  - Monolithic style.
  - Extreme customization is hard.

ggplot2 package

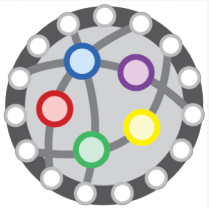
- Pros:
  - Grammar of Graphs!
  - Prefers data frames.
  - Multi-panel plotting made easy (facets)!
  - Use aesthetics to represent variables.
  - Combinations of layers.
- Cons:
  - Grammar of Graphs?
  - Monolithic style.
  - Extreme customization is hard.



References

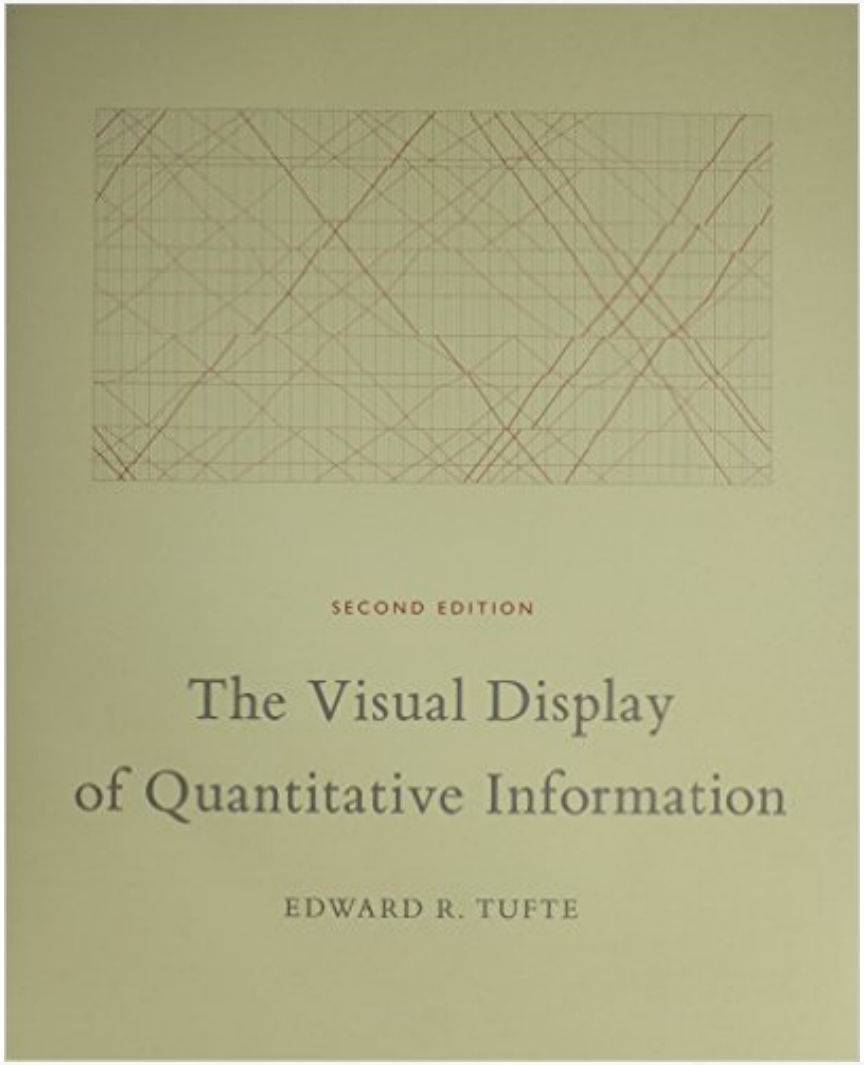
References

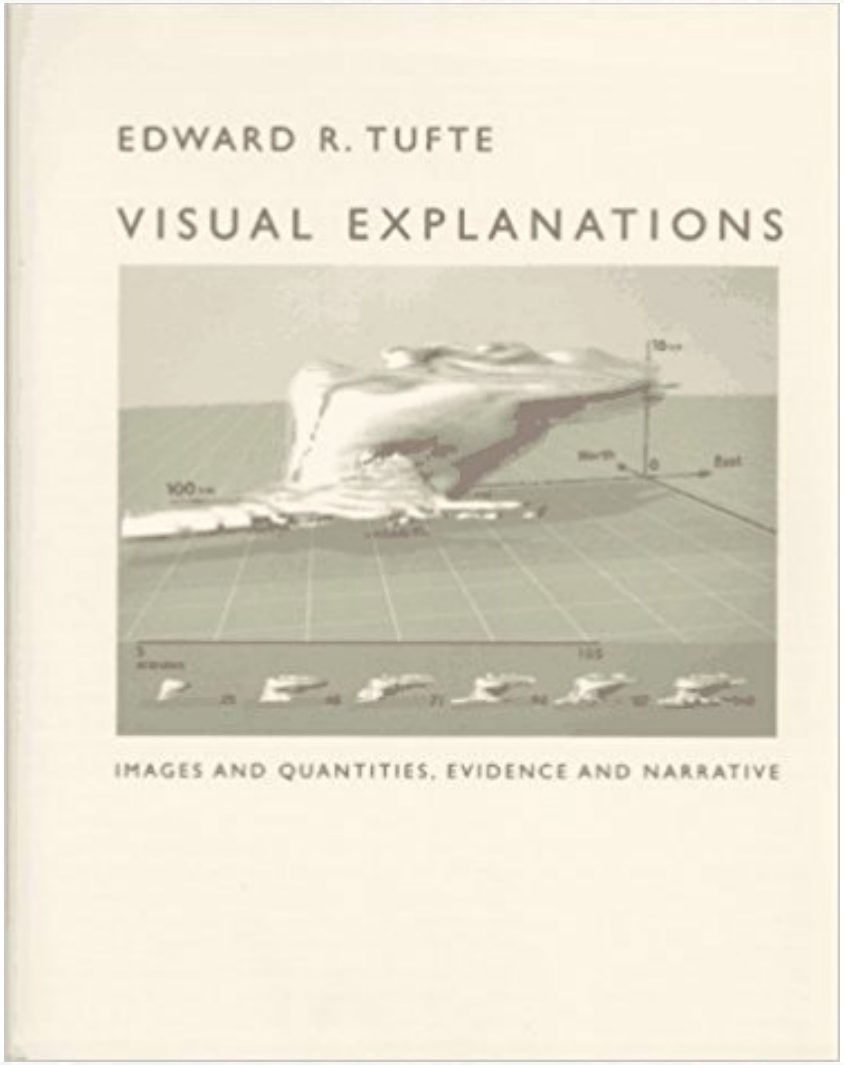
- Motivation: *Doing Data Science, Rachel Schutt and Cathy O'Neil, OReilly, 2014*
- Lots of code from StackOverflow!



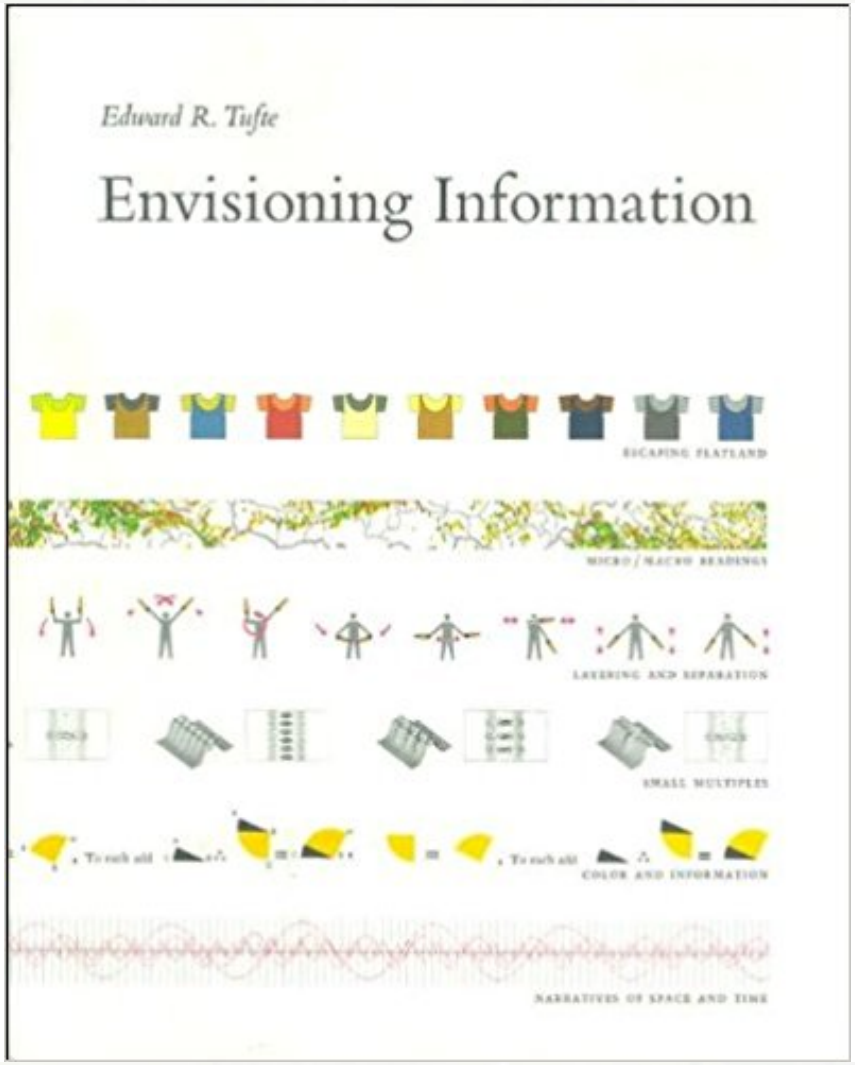
Digression: Edward Tufte and Visualization

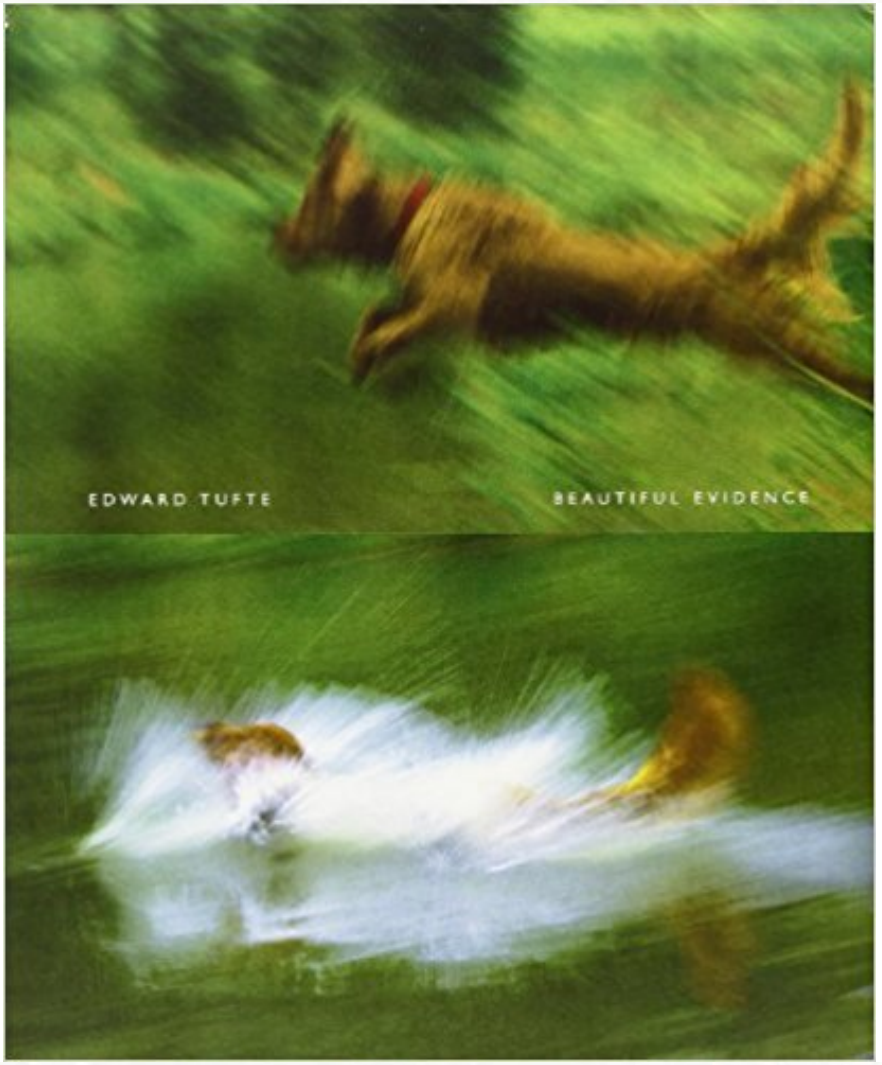
References


SECOND EDITION
The Visual Display
of Quantitative Information
EDWARD R. TUFTE

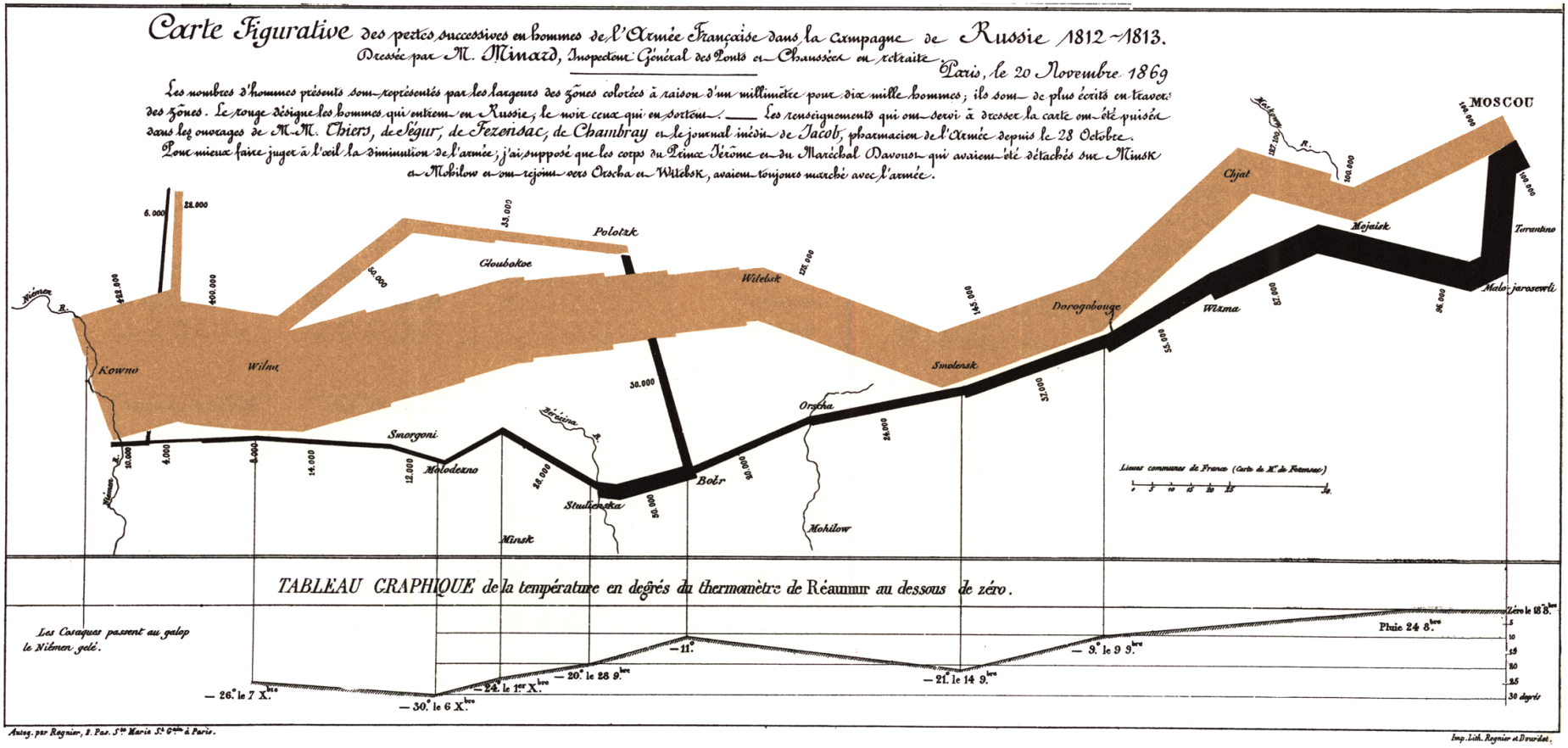

EDWARD R. TUFTE
VISUAL EXPLANATIONS
IMAGES AND QUANTITIES, EVIDENCE AND NARRATIVE

References

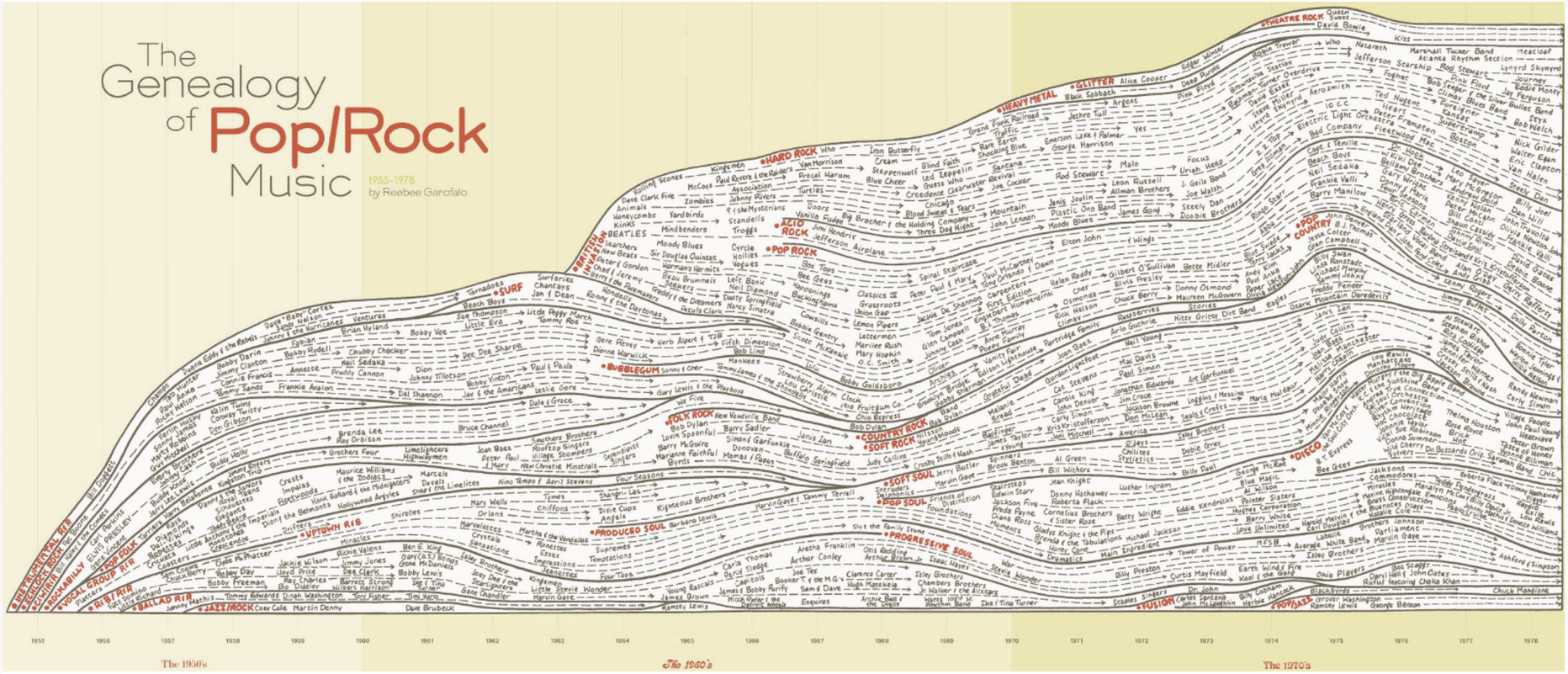

Edward R. Tufte
Envisioning Information


EDWARD TUFTE
BEAUTIFUL EVIDENCE

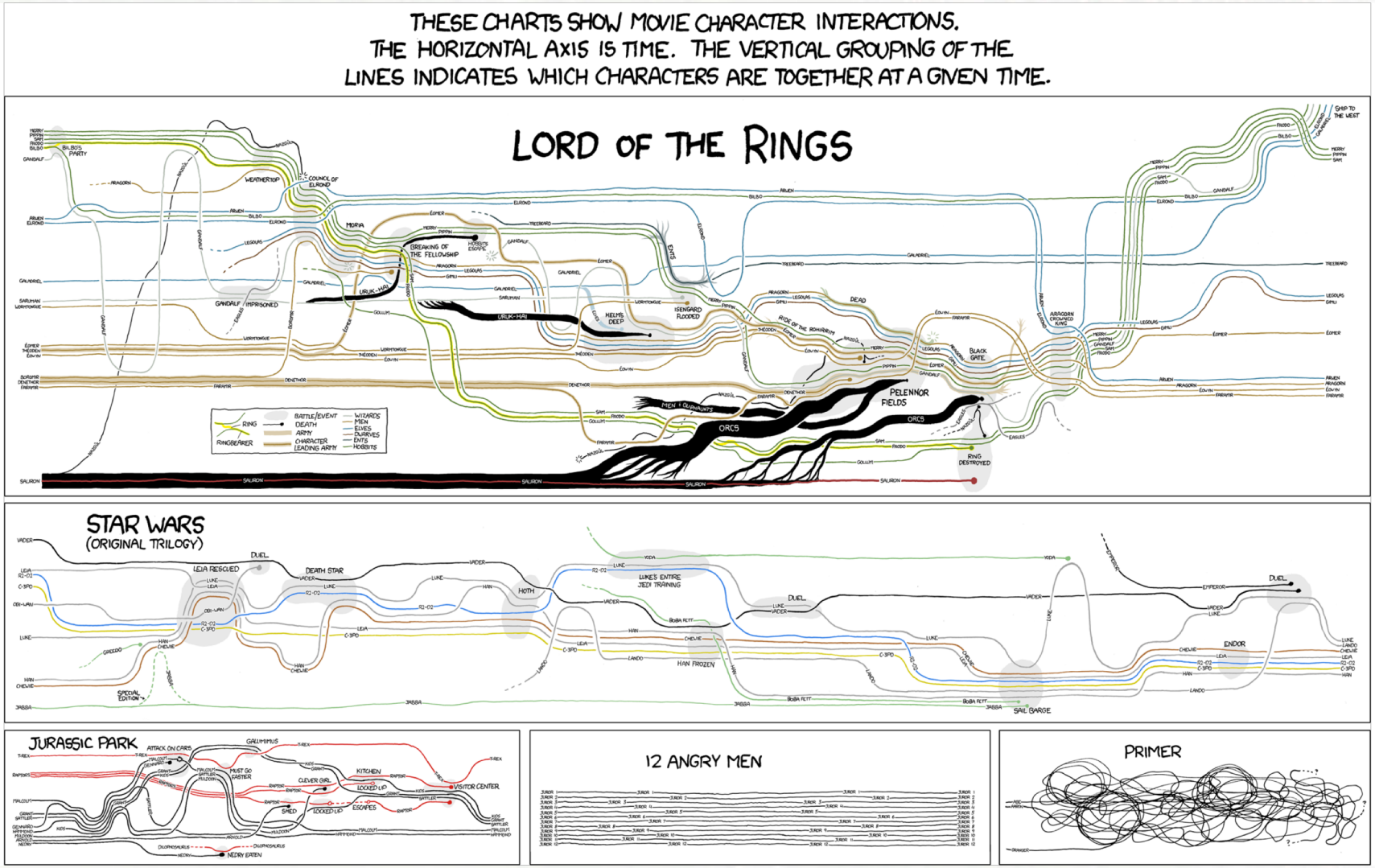
Edward Tufte and *Super Graphics*



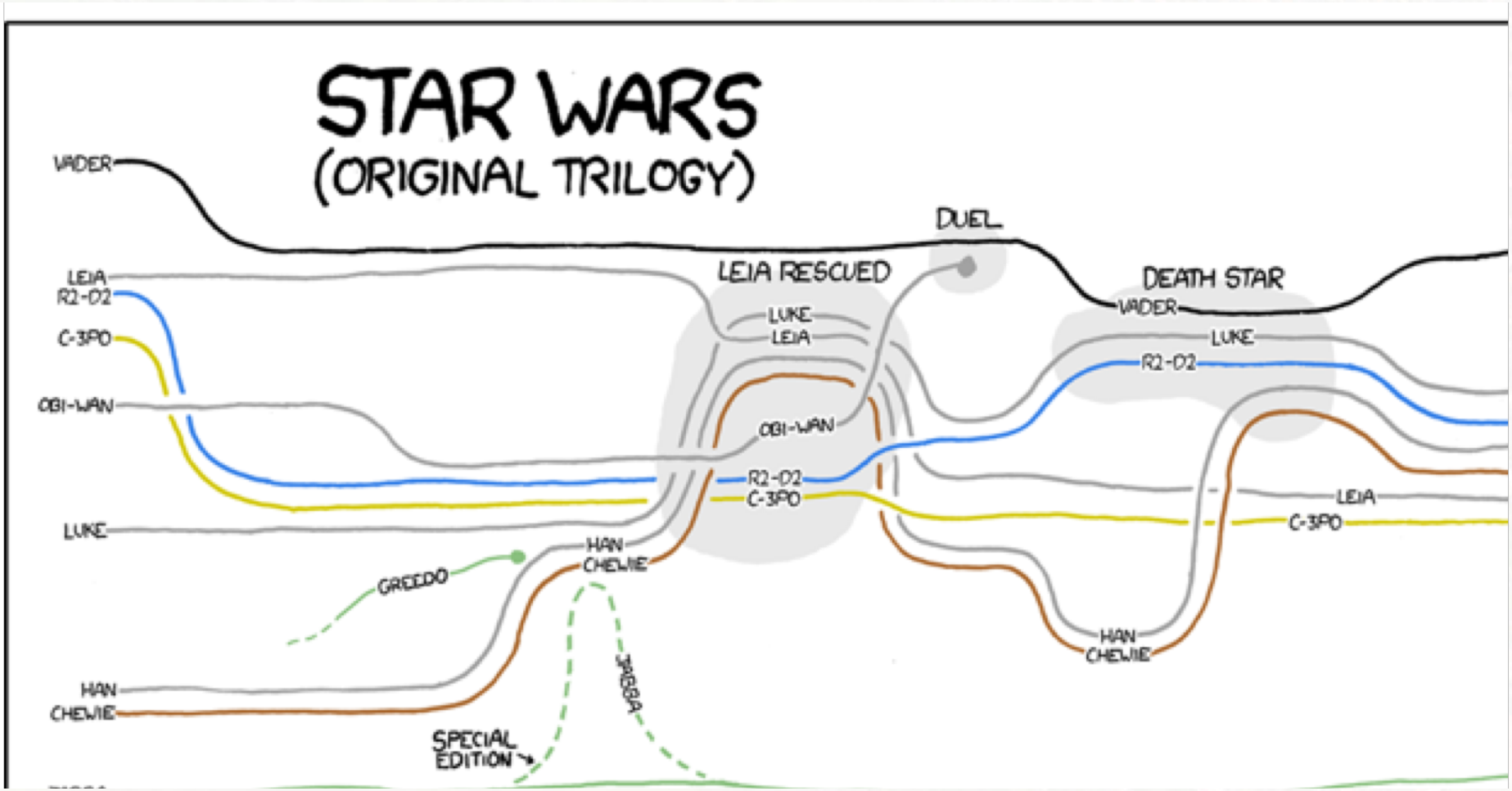
Edward Tufte and *Super Graphics*



XKCD and *Super Graphics*



XKCD and *Super Graphics*



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