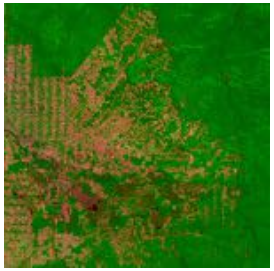
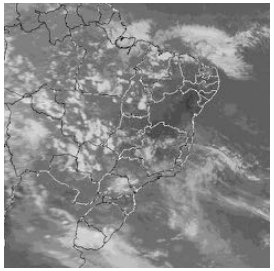




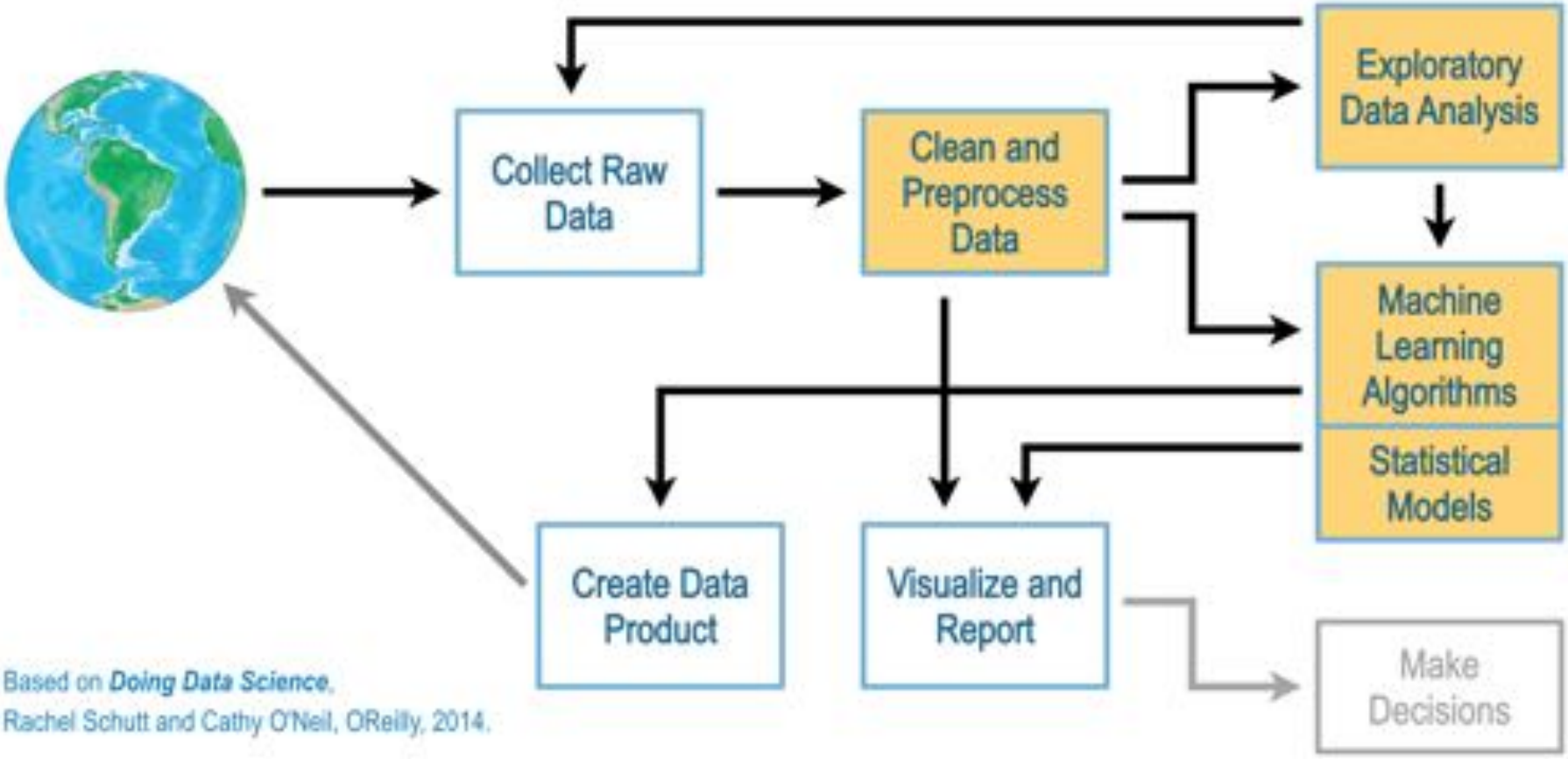
CAP-394
INTRODUCTION TO DATA SCIENCE

Rafael Santos – rafael.santos@inpe.br
Gilberto Ribeiro – gilberto.queiroz@inpe.br
www.lac.inpe.br/~rafael.santos/cap394.html

Updated in 2019

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.*

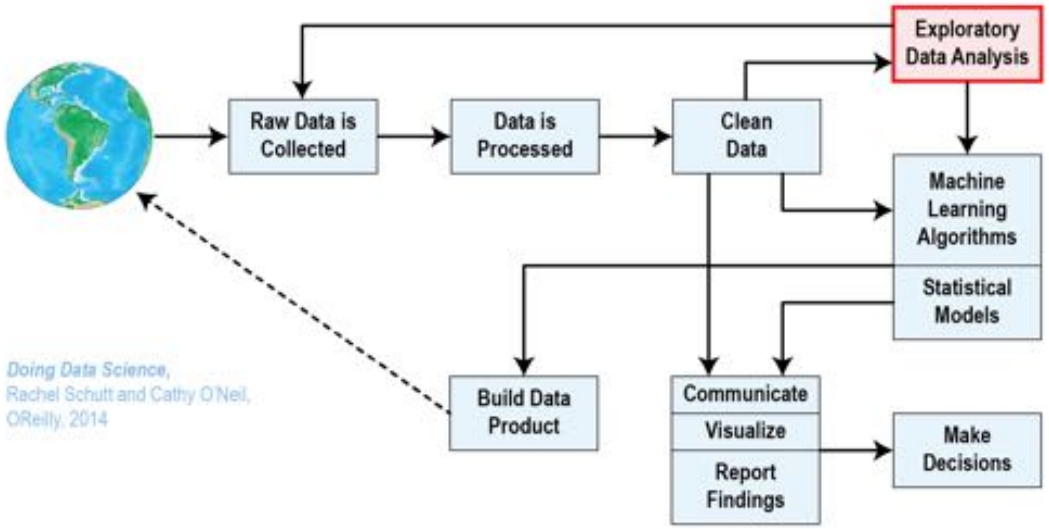
  — John Tukey

- 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

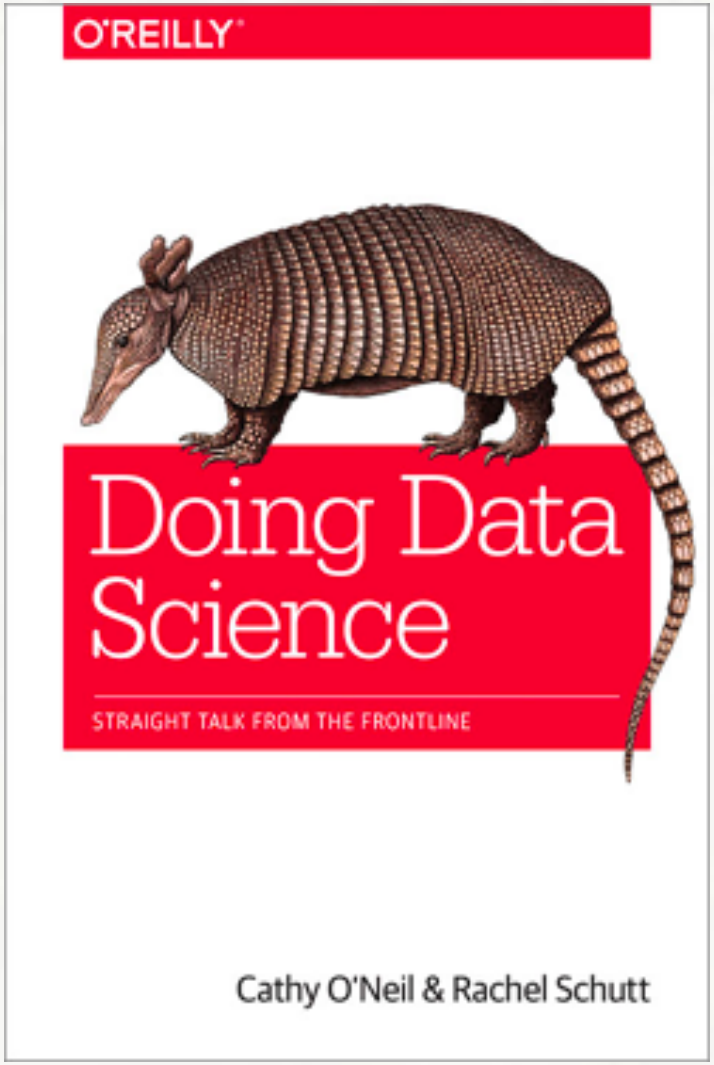
“R Programming”

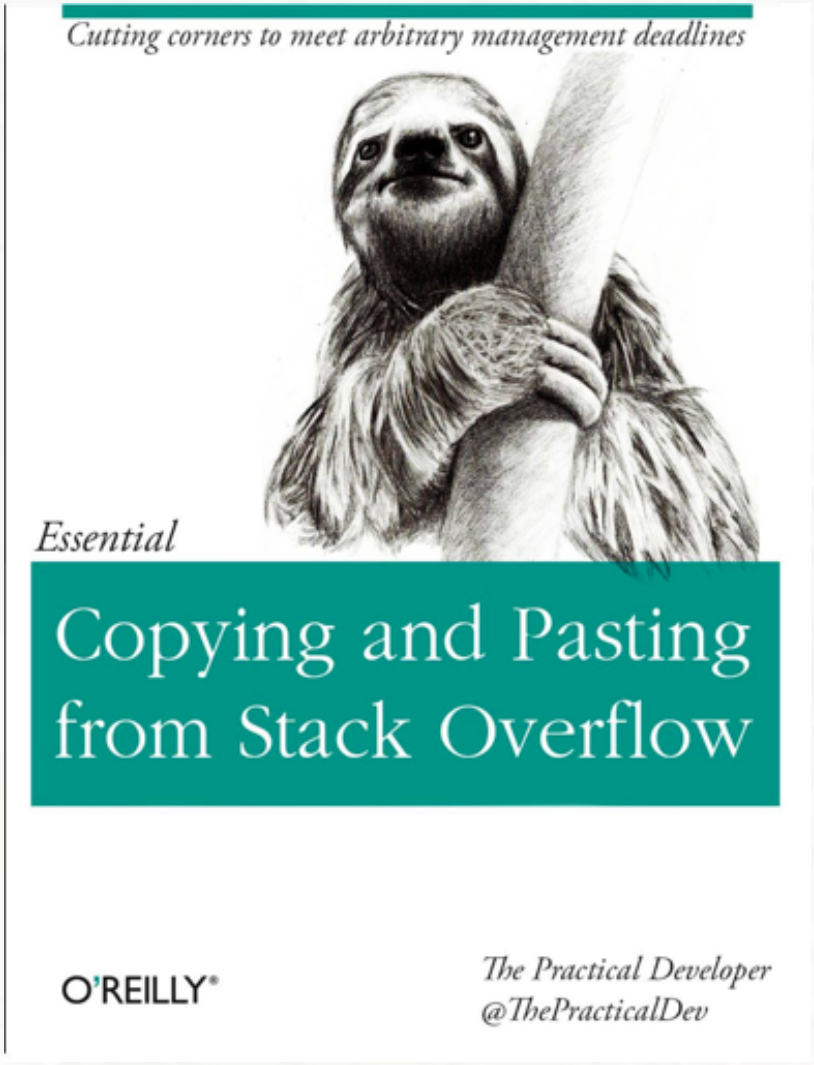
EDA in R

- Let's switch to a browser: http://www.lac.inpe.br/~rafael.santos/r.html

References

References


O'REILLY
Doing Data Science
STRAIGHT TALK FROM THE FRONTLINE
Cathy O'Neil & Rachel Schutt


Cutting corners to meet arbitrary management deadlines
Essential
Copying and Pasting from Stack Overflow
O'REILLY
The Practical Developer
@ThePracticalDev

Your Project

Research in Data Science

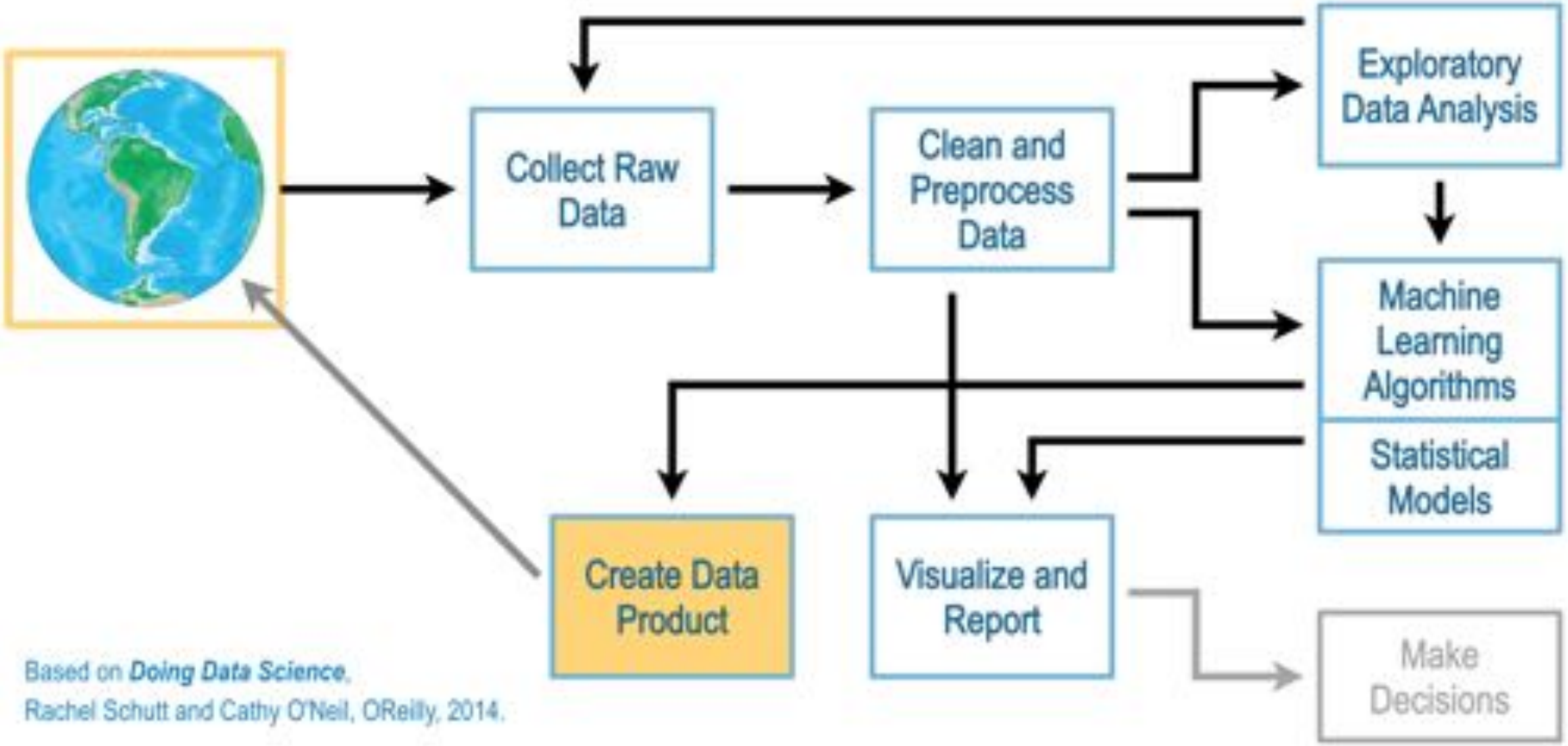
- Basic
  - New algorithms and variations.
  - Reference implementations.
  - Support tools (e.g. databases, data access, abstraction, automation).
- Applied
- Get your data, start doing:
  - Cleaning, munging.
  - EDA, visualization.
  - Basic model creation.
- …with real data, in a reproducible way.

Are we going the Basic way?

- There is ***nothing*** basic to it!

1. Develop a new (or derived) algorithm.
   - Justify why!
   - Package and document it extensively!
   - Add reproducible test cases!
   - Publish your code!
2. Develop a data access or analysis tool.
   - Better if it is thematic.
   - Document, add test cases, etc.

Are we going the Applied way?



A Comic Book guide to the Applied Data Science way

1. Think about your data.
   - Where is it? How can you access it?
   - How is it stored, formatted?

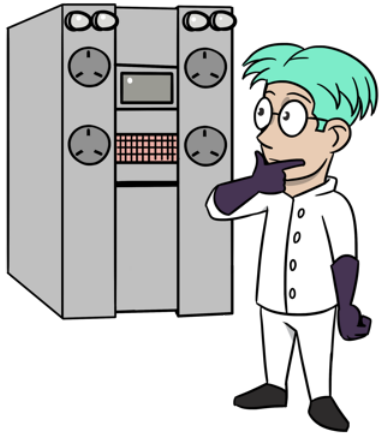
2. Which are interesting questions about it?
   - Can we answer those questions with the data or do we need more data?
   - Can we get more data?

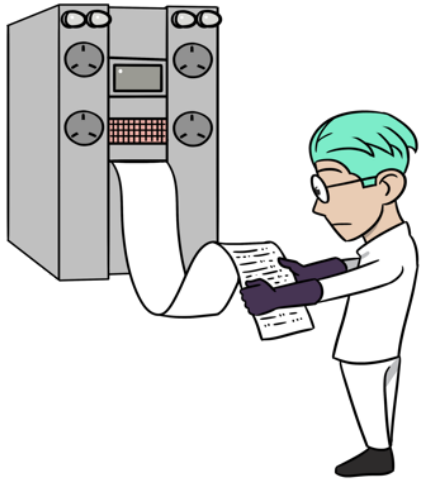
A Comic Book guide to the Applied Data Science way

3. Create code to explore it.
   - Explore its structure, completeness, features.
   - Do some basic statistics, EDA, visualization.
   - Don’t worry about failures in the code:
     worry about failures in the data!

A Comic Book guide to the Applied Data Science way

4. Consider hypotheses about your data.
   - Make sure it makes (at lease some) sense!
   - Check the data again!
   - Learn how to create models.
   - Apply and evaluate models on your data.

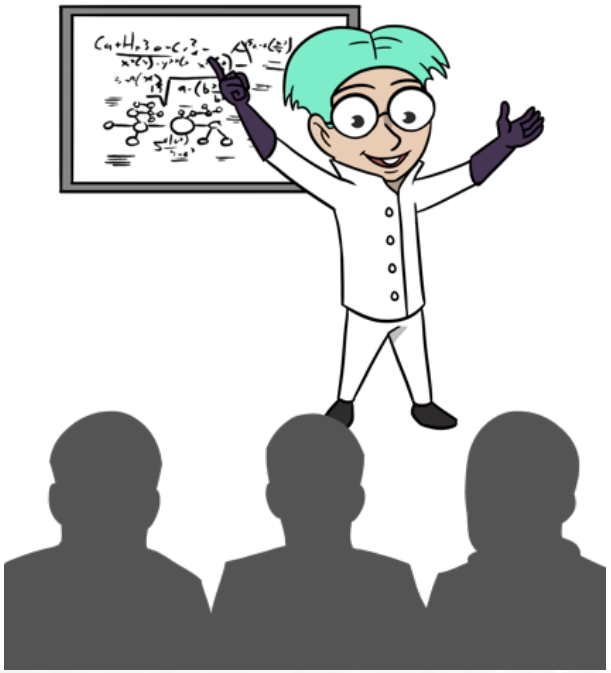




A Comic Book guide to the Applied Data Science way

5. Communicate and document your results.

- Intermediate results if they help to tell a story about the data.
- Even bad results if they can teach us something!
- Have you been using notebooks?
- Can you create a new data product?

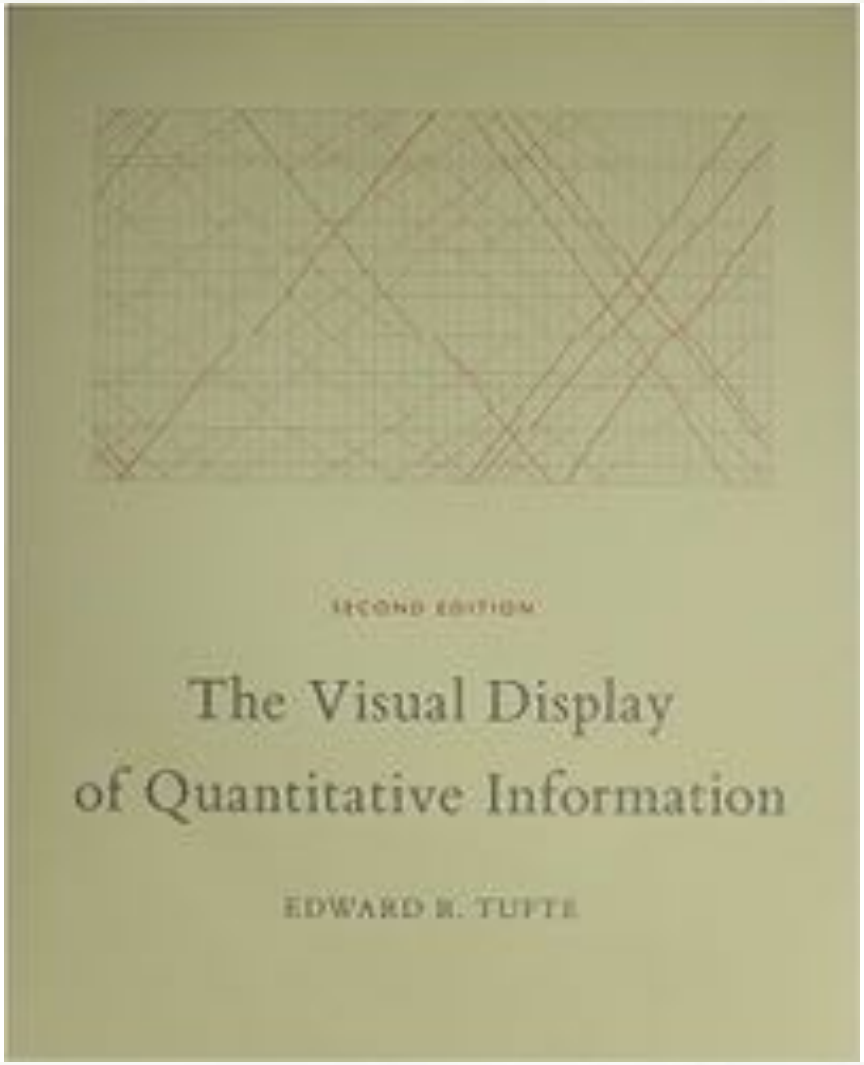


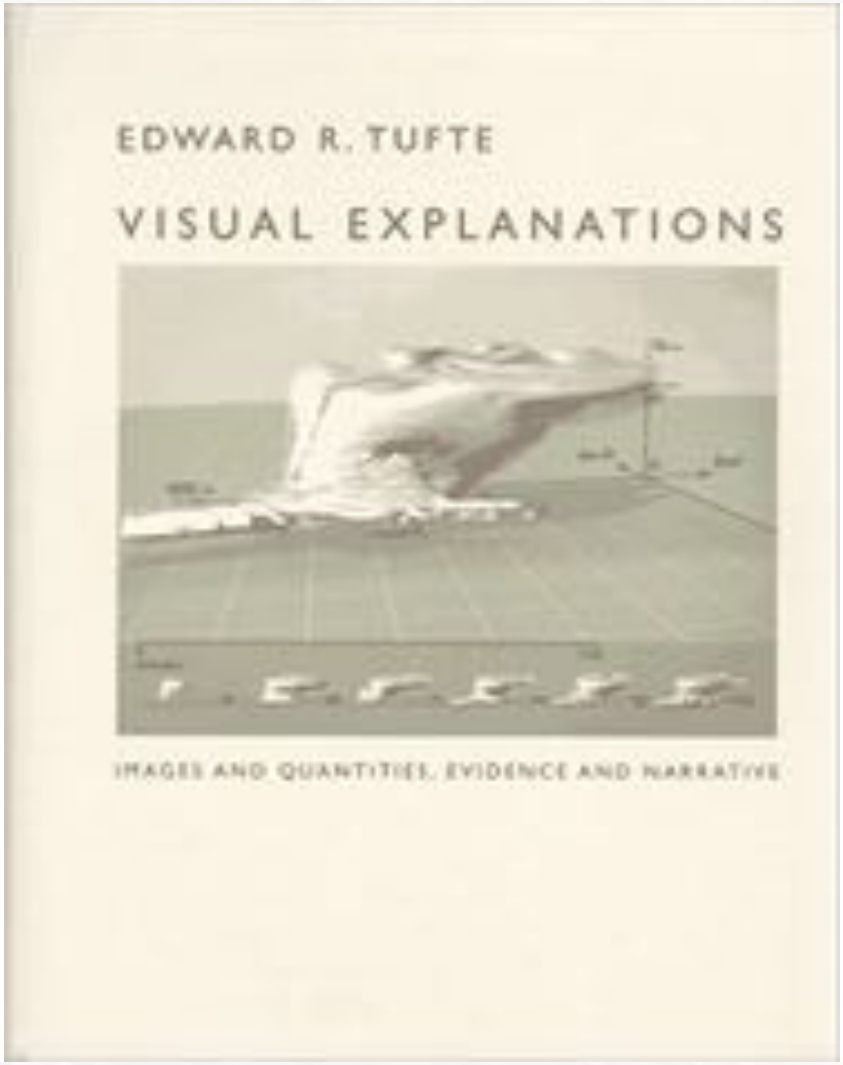
A Comic Book guide to the Applied Data Science way

- A *very common* alternative approach:

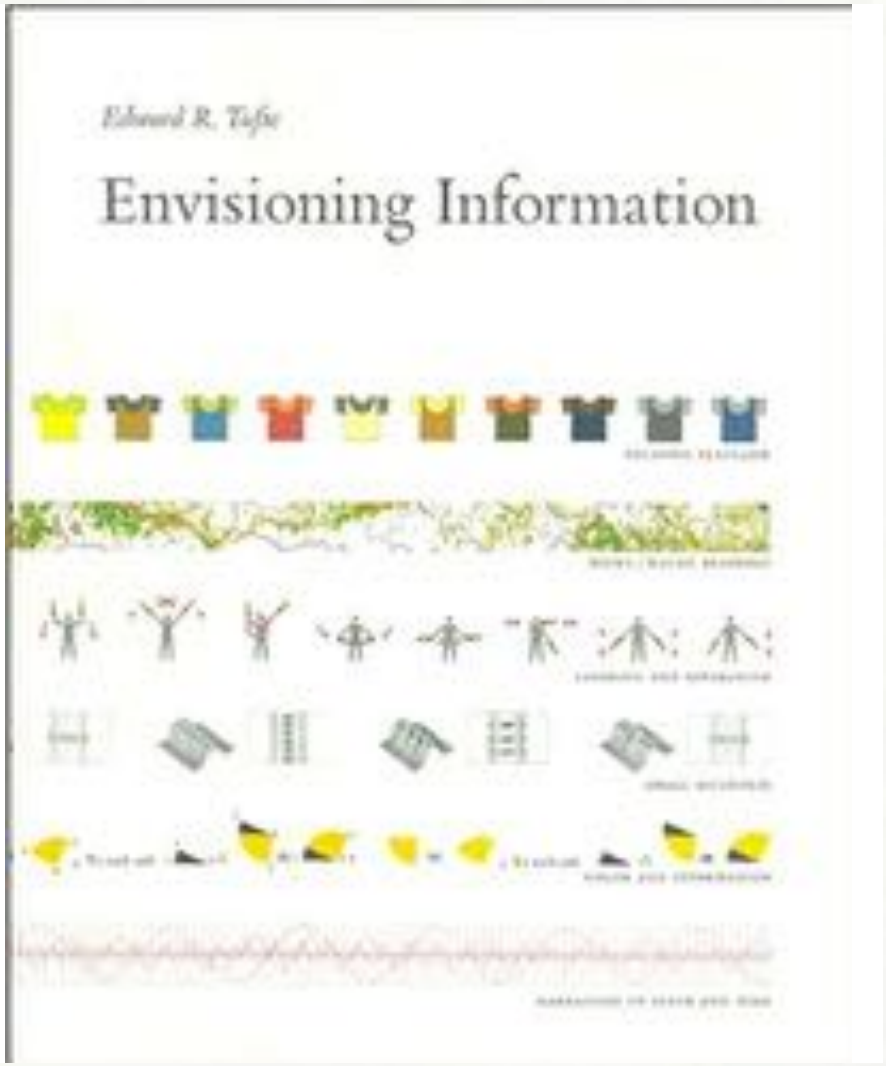
Digression: Edward Tufte and Visualization

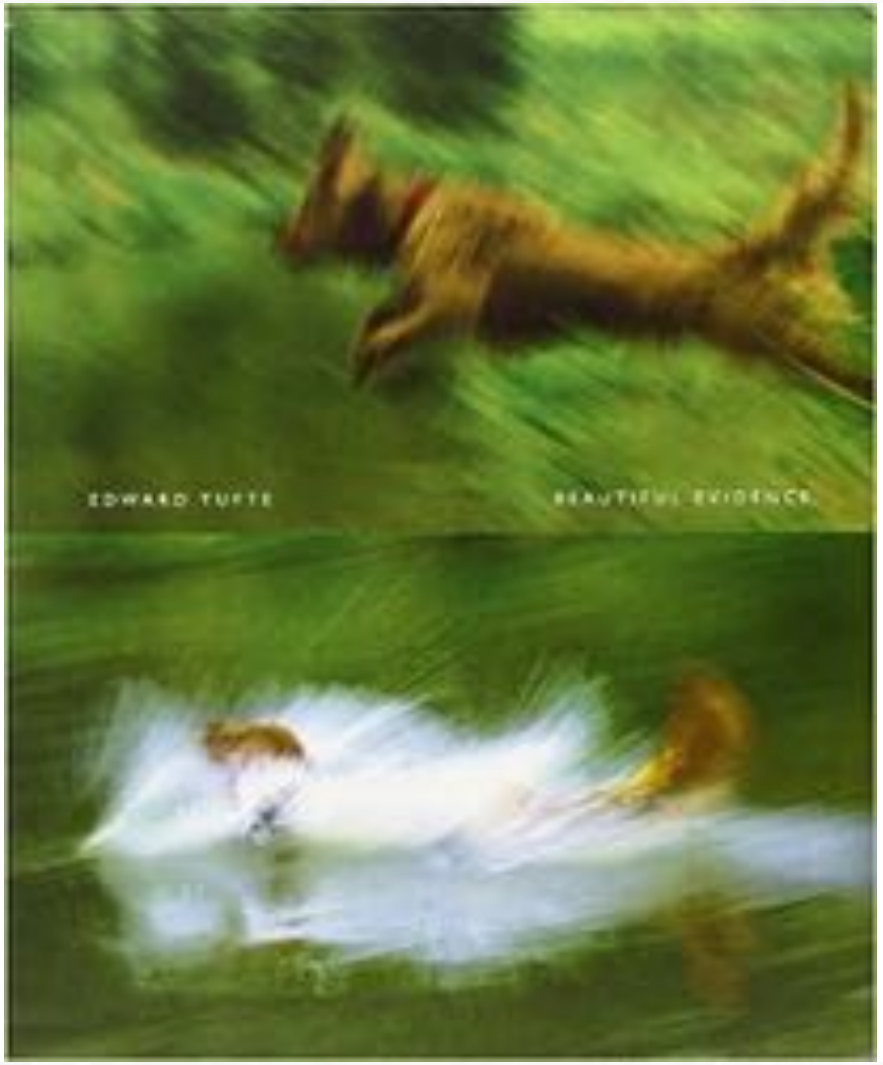
References


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

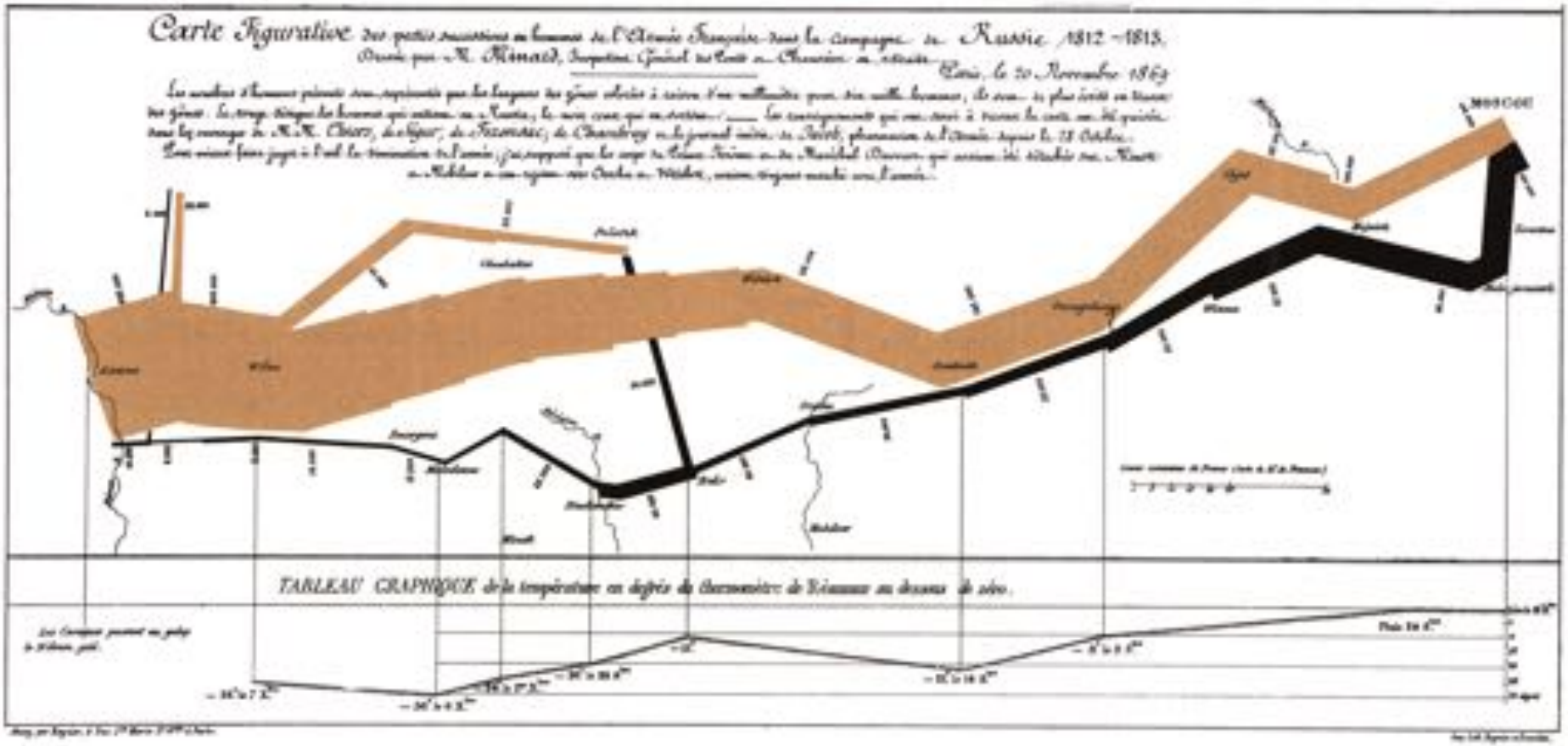

EDWARD R. TUFTE
VISUAL EXPLANATIONS
IMAGES AND QUANTITIES, EVIDENCE AND NARRATIVE

References

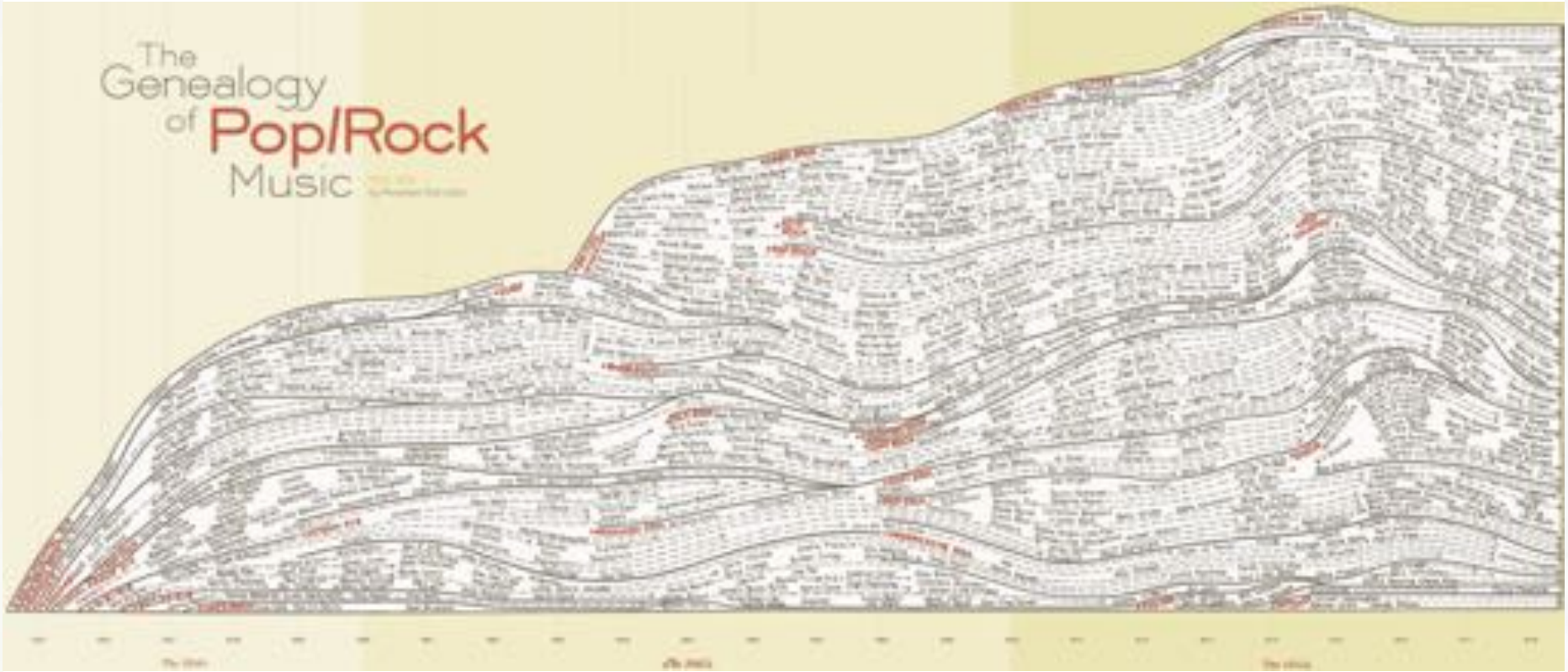




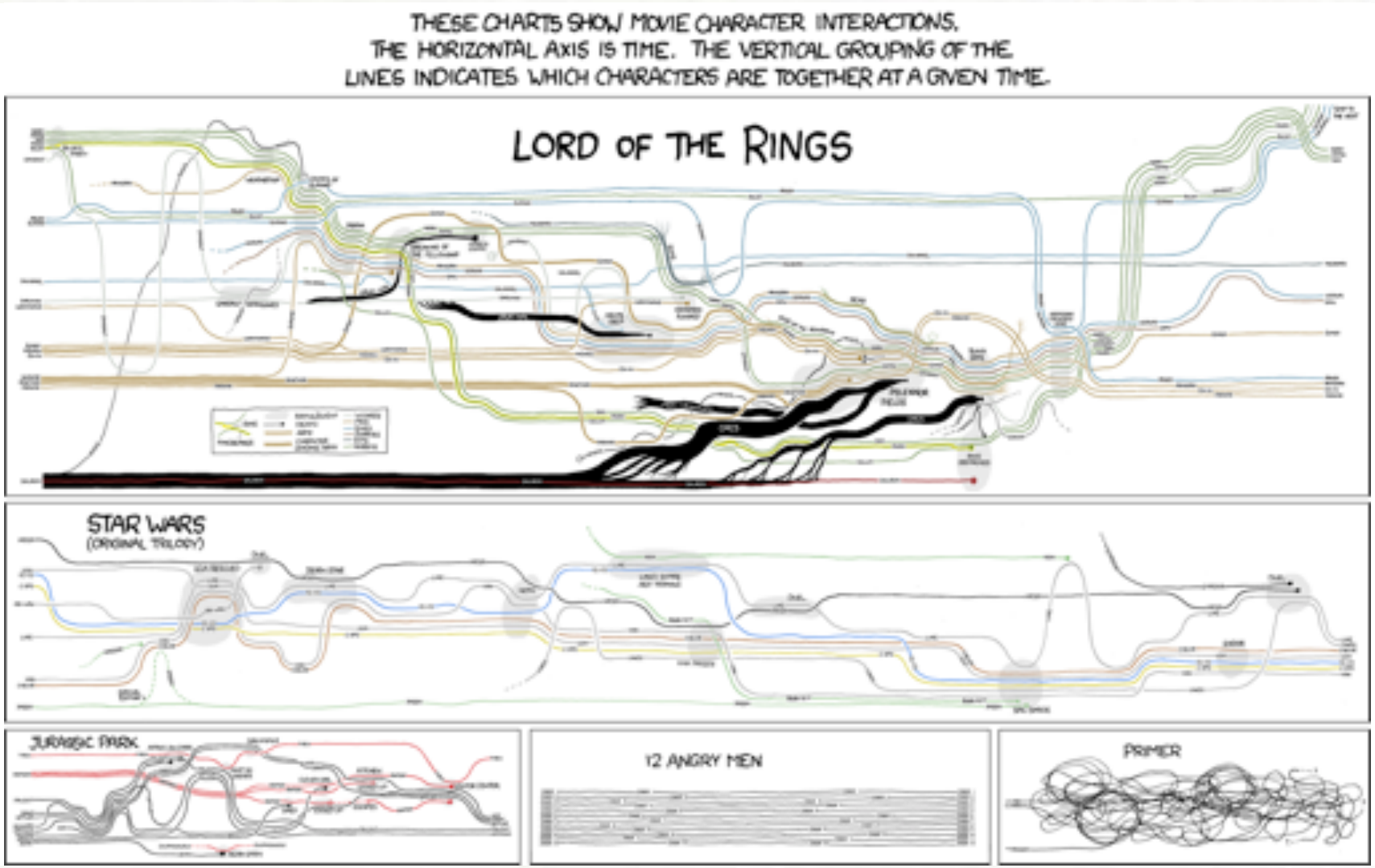
Super Graphics



Super Graphics



Super Graphics (XKCD)



XKCD and *Super Graphics*