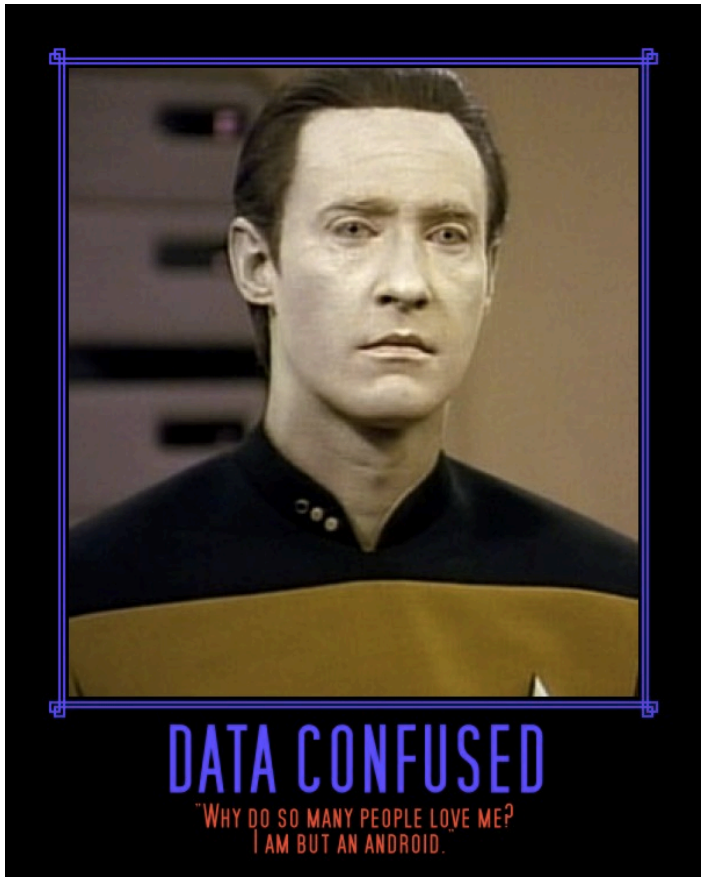
Introduction to Data Science

UNDERSTANDING YOUR DATA

BRIAN D'ALESSANDRO

WHAT IS DATA ANYWAYS?

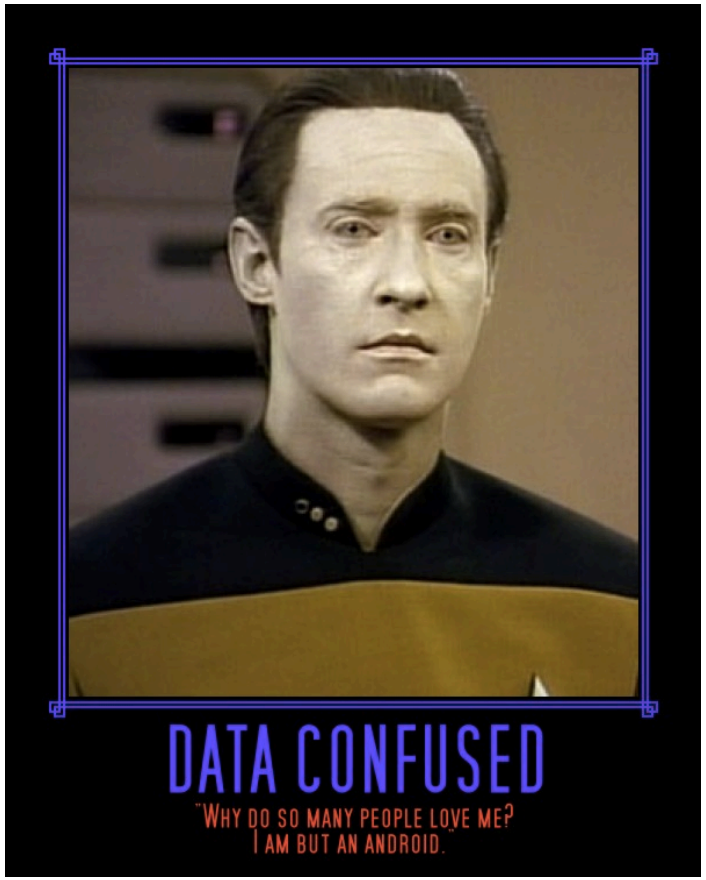


Files
(i.e. .txt, .binary, .csv)?

Programming Objects
(i.e. tuples, arrays, hashMaps, dicts)?

Variables and their distribution
(i.e. X, P(X))?

WHAT IS DATA ANYWAYS?



For now we're going to focus on this meaning of data.

Variables and their distribution (i.e. X, P(X))?

SOMETHING YOU SHOULD KNOW

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

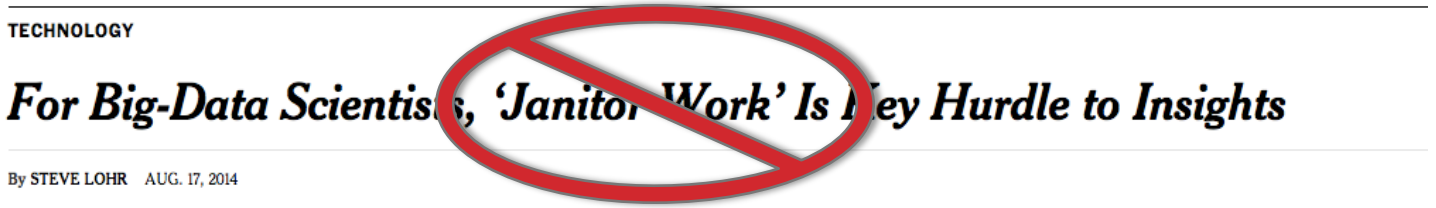
By STEVE LOHR AUG. 17, 2014

> Yet far too much handcrafted work —
> what data scientists call "data wrangling," "data munging" and "data janitor work" — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets.

Source:http://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html?_r=0

DON'T FRET ABOUT GARBAGE HEADLINES

Don't think of data preparation as a nuisance or pejorative

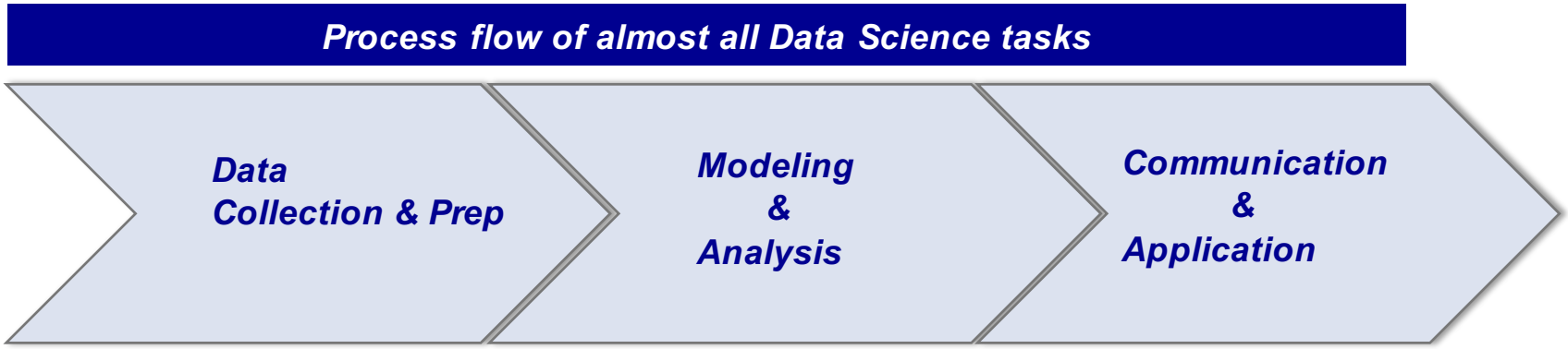

TECHNOLOGY

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

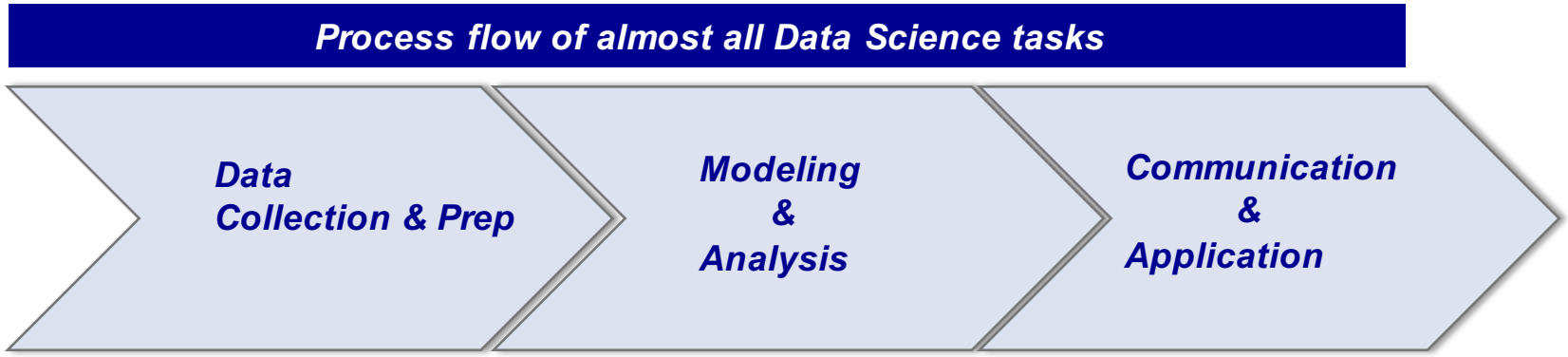
Data preparation and understanding is integral to great analysis. It is well worth spending 80% of your time getting it right!

DATA PREP => SUCCESS



This is the "*sexy*" part,
getting all the research and attention from competitions
…
and it is indeed fun!

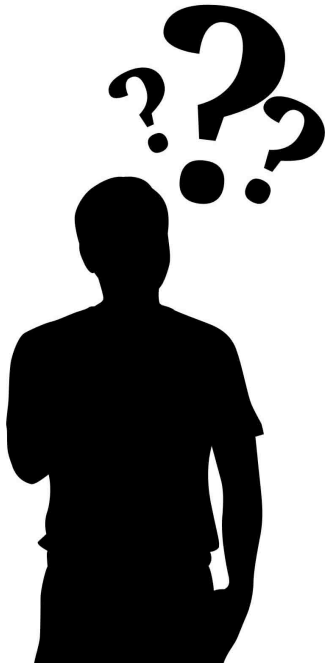
DATA PREP => SUCCESS



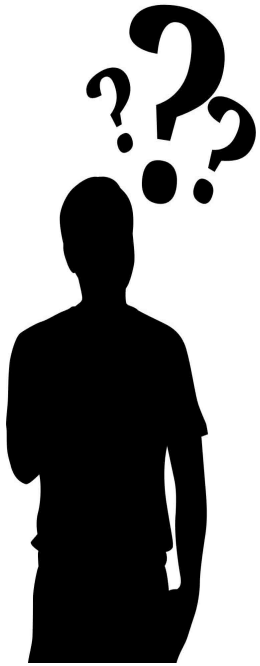
These are what make you a good and effective data scientist however

2 KEY QUESTIONS ABOUT DATA

1. Where did my data come from?
2. What does it look like?



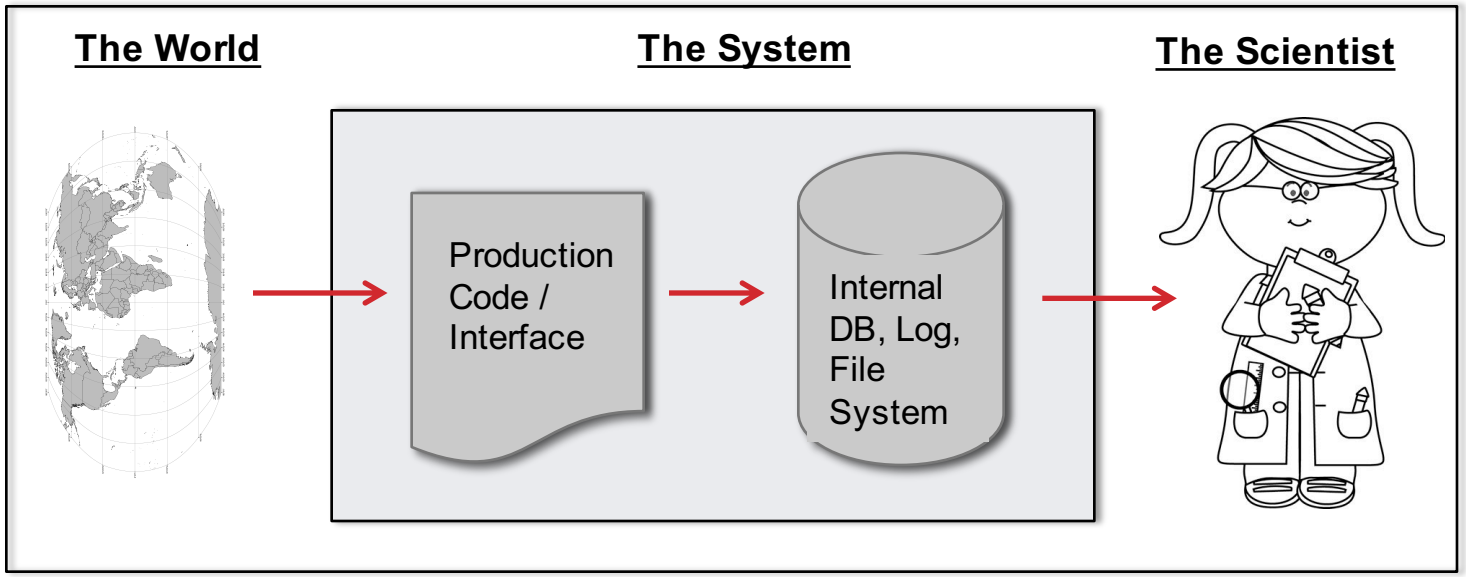
ALWAYS KNOW THE SOURCE



1. Where did my data come from?
   - Internal ETL processes
   - Production logging/sampling
   - Web scraping/ API
   - Survey/Panel

Rule of thumb: If you did not pull your own data, always be skeptical (that it is indeed what you need), and spend extra time validating it.

COMMON DATA FLOW



Challenges we often face:

- Often the data scientist doesn't write/own the code that produces the data. (Selection Bias, Unknown Unknowns)
- The system intervening on the world changes the nature of the data we collect from it. (Selection Bias, Negative Feedback Loop)
- We have no control often of data that streams into the system. (Concept Drift & Selection Bias)