The joys of Clean Data!

Silicon Valley Big Data Science Meetup

23 Sep 2015

Matt Dowle

Overview

- For beginners
- Examples from my background
- Tools along the way
- Live demo of “80% munging”
- How H2O fits in
- Q & A

1996 – Lehman Brothers

- Just graduated – Applied Maths & Computing
- Dividend claims
- Cleaning at source; e.g. data entry typos
- Estimate cash flows, alerts etc
- Nothing fancy
- Tools: VB & Sybase
- How I accidentally created messy data

1999 - Salomon Brothers

- Equity risk model
  - Multiple time series regression (10 year)
  - DEM proxy for EUR prior to 1 Jan 1999
  - IPOs get their sector's median; e.g. France Telecom
  - Abbey National X0004445**5**
  - 90% of the lines of code was not the regression

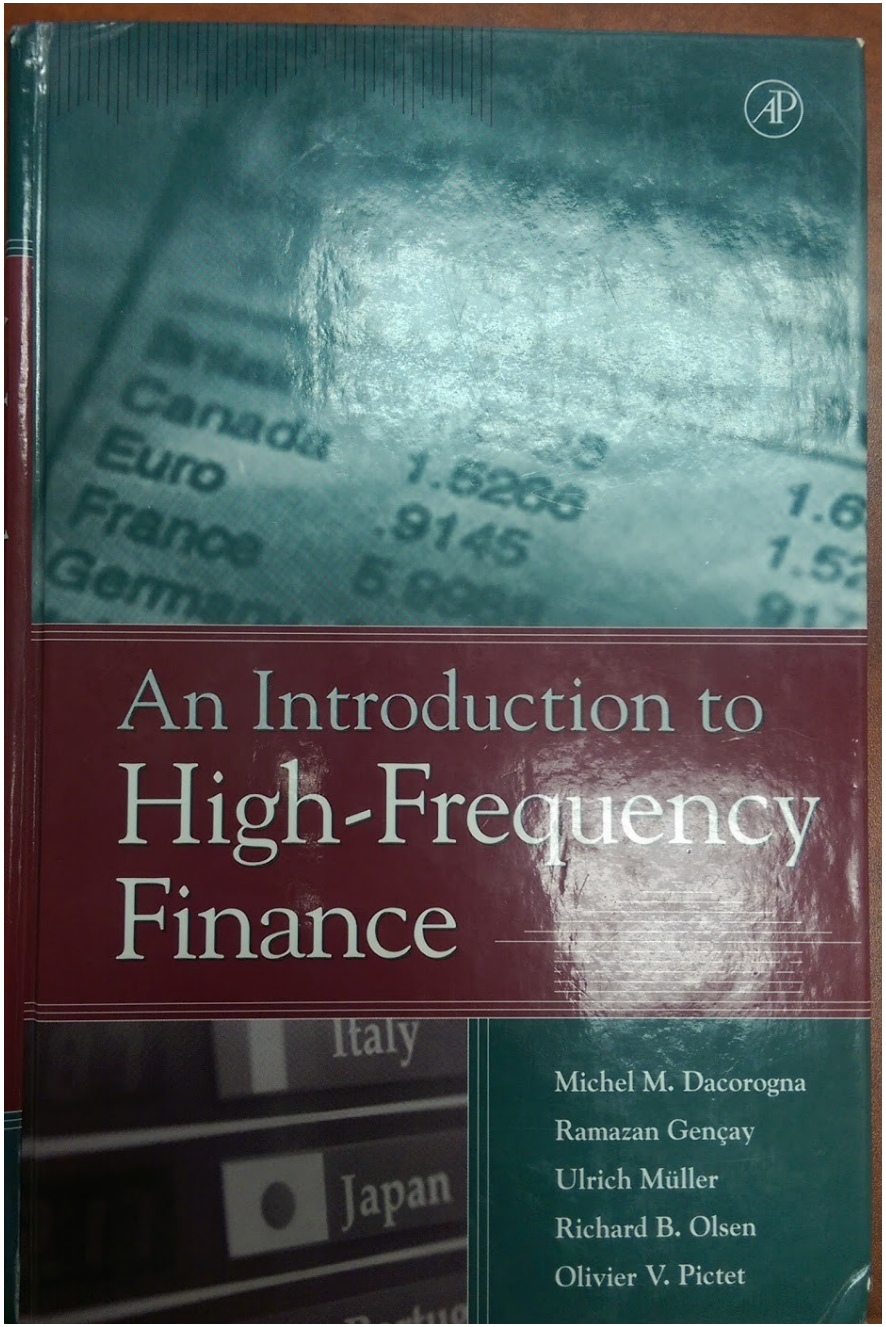
2002 - Citigroup

- Pairs Trading
- 200 most liquid stocks
- 200 x 199 / 2 = 19,900 pairs
- Stock splits, id changes
- Dickey Fuller test for stationarity
- Bollinger bands => buy/sell signal
- Excel spreadsheet to clients with embedded S-PLUS plot, daily, 50 custom variants
- Rebalance => orphan & surrogate pairs

2004
moved to fund management

Bigger data

e.g. 25TB



4

ADAPTIVE DATA CLEANING

Over cleaning

1. I queried for intra-day auctions

   `select from quote where bid>ask`

   No results; i.e. all bid<ask.

   Asked data provider

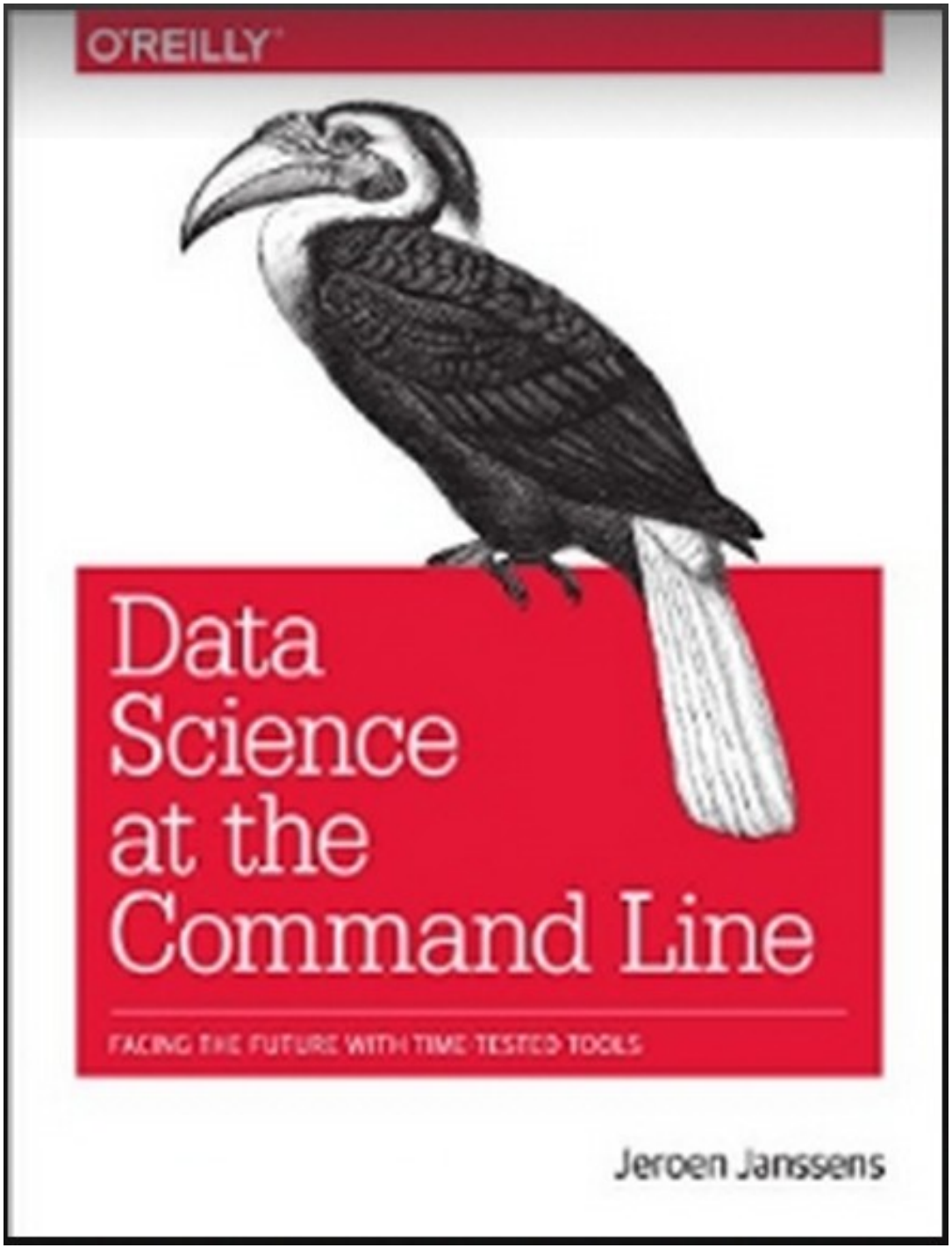
   Grrrrr

2. Negative prices can be correct

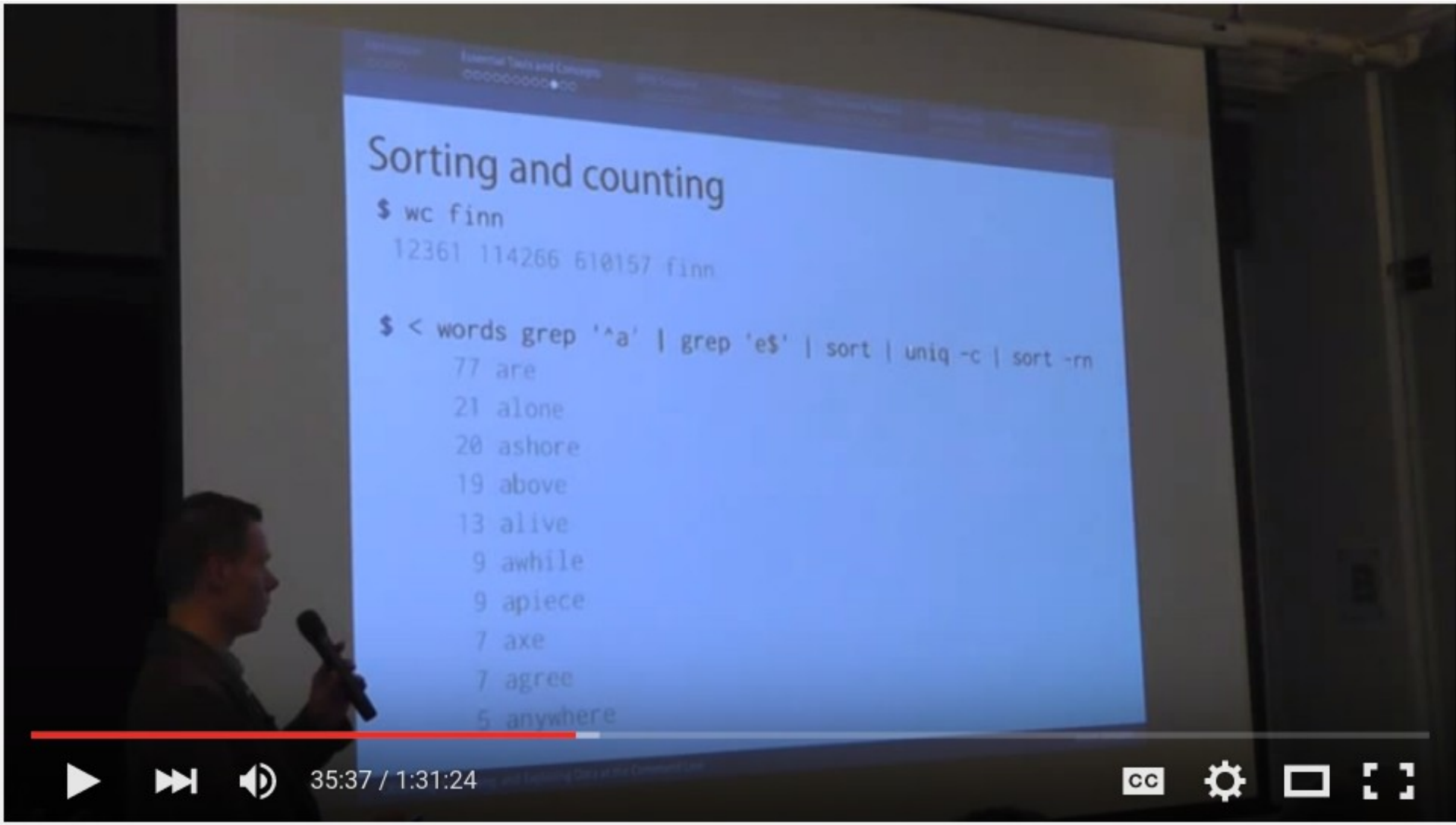
Tools

KDB http://kx.com/
@kxsystems

OneTick https://www.onetick.com/
@OneMarketData


O'REILLY
Data Science at the Command Line
FACING THE FUTURE WITH TIME-TESTED TOOLS
Jeroen Janssens

https://www.youtube.com/watch?v=QxpOKbv-KQU



alias	**csvsql**	**json2csv**	**shuf**
awk	**csvstack**	**less**	**sort**
aws	**csvstat**	**parallel**	**split**
bc	**curl**	**paste**	**sql2csv**
bigmler	**cut**	**pbc**	**tail**
body	**dseq**	**python, R and r**	**tapkee**
cat	**find**	**Rio**	**tee**
cols	**for**	**Rio-scatter**	**tr**
csvcut	**grep**	**run_experiment**	**tree**
csvgrep	**head**	**sample**	**uniq**
csvjoin	**header**	**scrape**	**wc**
csvlook	**in2csv**	**sed**	**weka**
csvsort	**jq**	**seq**	**xml2json**

- Can be faster than loading the whole file into R or Python
- Can be faster workflow
- Pre-processing before loading into R or Python

tidyr by Hadley Wickham

https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html

- Untidy data defined as :
  - Column headers are values, not variable names.
  - Multiple variables are stored in one column.
  - Variables are stored in both rows and columns.
  - Multiple types of observational units stored in the same table.
  - A single observational unit is stored in multiple tables.
- Solves by: gathering, separating and spreading
- That's the *shape* of the data. Yes, good, but not the kind of messy data I'm talking about in this presentation.

To illustrate

- In June 2013, RStudio made available download logs from their CRAN mirror
  http://blog.rstudio.org/2013/06/10/rstudio-cran-mirror/
- R-Bloggers search “CRAN download stats”
  154 results; e.g.

  http://www.r-bloggers.com/finally-tracking-cran-packages-downloads/

  https://github.com/metacran/cranlogs

  http://www.r-bloggers.com/working-with-the-rstudio-cran-logs/

  http://www.r-bloggers.com/cran-download-statistics-of-any-packages-rstats/

  http://www.r-bloggers.com/my-r-packages-worldmap-of-downloads/

March 2015

vozag Rankings, Reviews & Data Of The World

Home About Us

Top 100 R Packages by Downloads

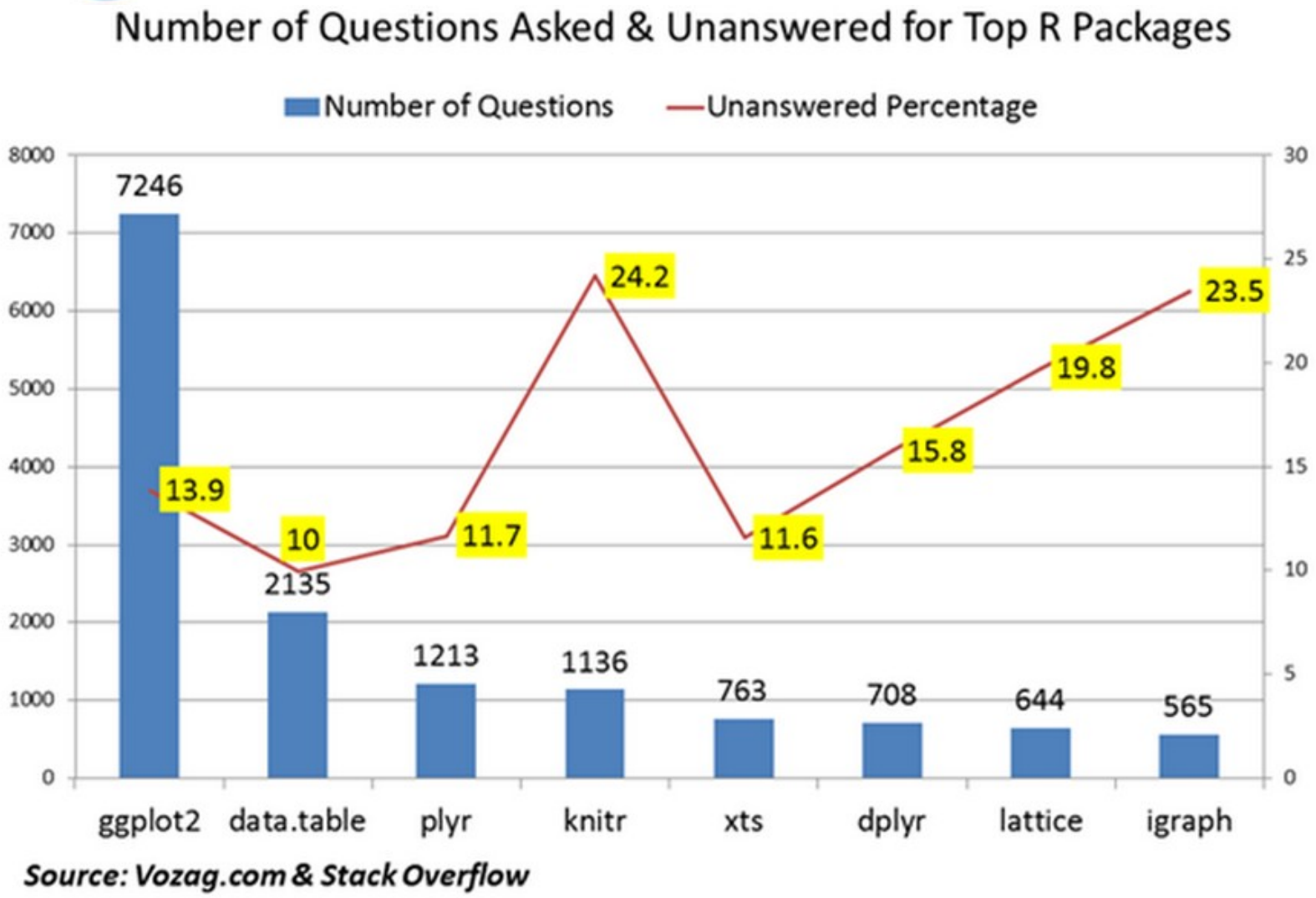
Enter email Get Research By Email

We analyzed data from Cran daily download data to understand the top R packages that were downloaded. Here is the list based on download data for a single day: Feb 28th 2015.

Rank	Package	No. of Downloads
1	Rcpp	1960
2	ggplot2	1785
3	digest	1709
4	reshape2	1651
5	plyr	1634
6	rJava	1577
7	stringr	1549
8	RColorBrewer	1497
9	colorspace	1372
10	manipulate	1363

Data Science Central

April 2015



- Comparing the top downloaded packages with the most discussed packages shows little correlations between them. For Instance, ggplot2 has the most questions asked & is the second highest downloaded package **but data.table package (the second highest ranked R package for questions asked) is not even in the top 100 packages downloaded.** Knitr is another example which is in the top 5 questions asked, but is 27th ranked in downloaded packages.
- **So- does the R community need to focus on packages that have the highest questions to resolve their issues rather than the ones with the most downloads?**

Let's look at the data!

Live demo of munging

Observations and comments on meetup video recording

https://livestream.com/h2oai/events/4369265

(~ 40 mins in)

“Big data”

1. **Data > 240GB**

   needle-in-haystack e.g. fraud

2. **Data < 240GB**

   compute intensive, parallel 100's cores

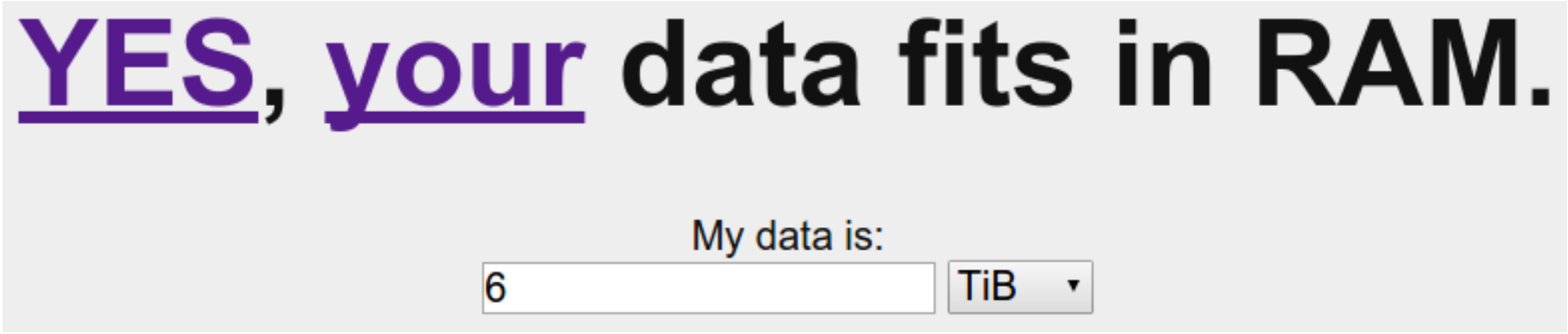
3. **Data < 240GB**

   feature engineering > 240GB

Speed for i) production and ii) interaction

NB: 240GB is currently largest available on EC2

http://yourdatafitsinram.com/



Dell PowerEdge R920 60 core

(4 * Intel® Xeon® E7-8880L 2.2GHz, 37.5M Cache, 15 Core)

with 1.5TB RAM $60k (96 * 16GB)

with 6TB $150k-$200k? (96 * 64GB)

But, still “only” 60 cores

In the office here we already have 2.5TB RAM and 320 cores on 10 machines.

So do many businesses.

Thank you.

Q & A