

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

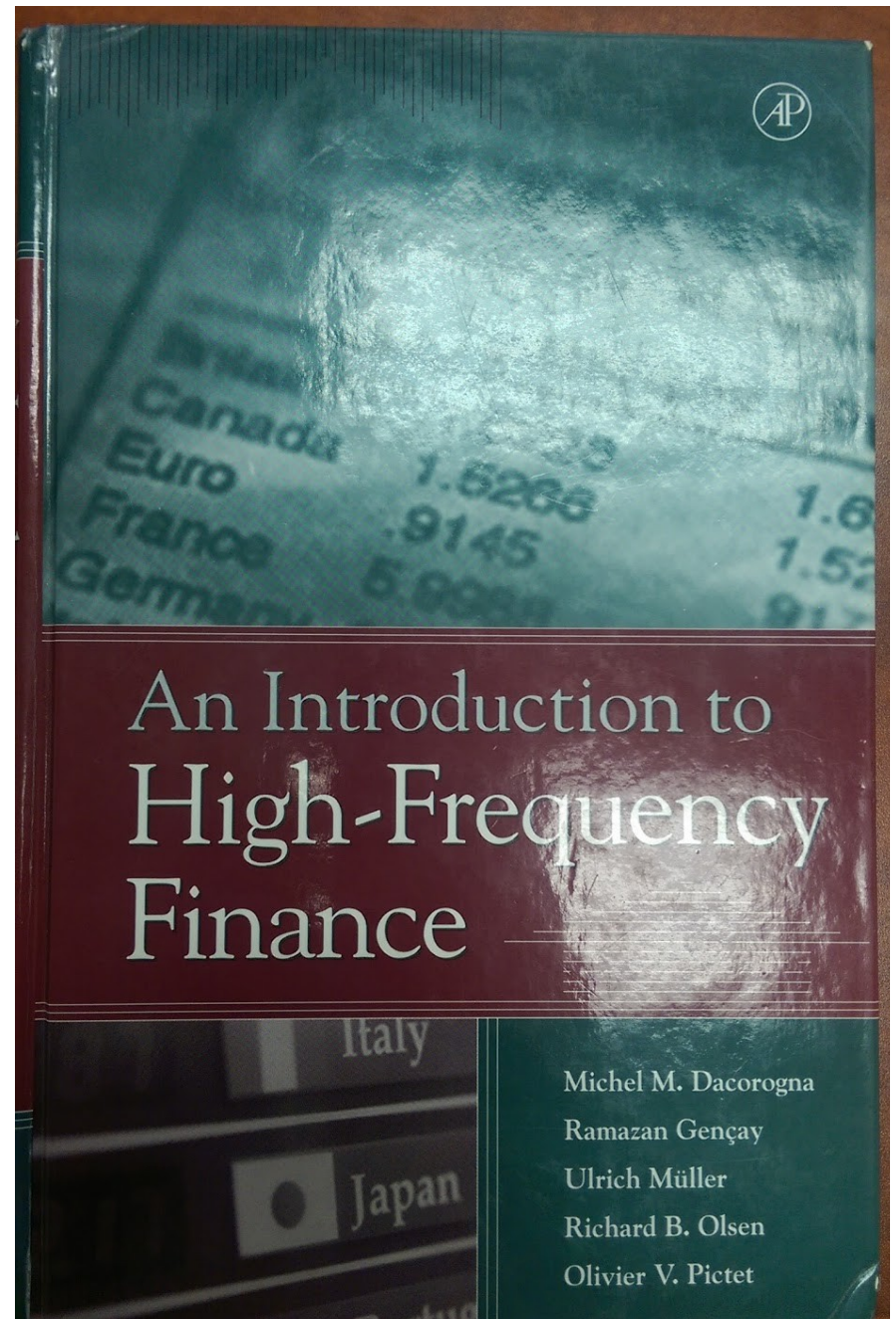
- Equity risk model
 - Multiple time series regression (10 year)
 - DEM proxy for EUR prior to 1 Jan 1999
 - IPOs get their sector's median; e.g. France Telecom
 - Abbey National X000445**5**
 - 90% of the lines of code was not the regression

2002 - Citigroup

- Pairs Trading
- 200 most liquid stocks
- $200 \times 199 / 2 = 19,900$ 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

4.1	Introduction: Using a Filter to Clean the Data	82
4.2	Data and Data Errors	84
4.2.1	Time Series of Ticks	84
4.2.2	Data Error Types	85
4.3	General Overview of the Filter	86
4.3.1	The Functionality of the Filter	86
4.3.2	Overview of the Filtering Algorithm and Its Structure	88
4.4	Basic Filtering Elements and Operations	88
4.4.1	Credibility and Trust Capital	89
4.4.2	Filtering of Single Scalar Quotes: The Level Filter	91
4.4.3	Pair Filtering: The Credibility of Returns	93
4.4.4	Computing the Expected Volatility	96
4.4.5	Pair Filtering: Comparing Quote Origins	98
4.4.6	A Time Scale for Filtering	100
4.5	The Scalar Filtering Window	103
4.5.1	Entering a New Quote in the Scalar Filtering Window	104
4.5.2	The Trust Capital of a New Scalar Quote	104
4.5.3	Updating the Scalar Window	106
4.5.4	Dismissing Quotes from the Scalar Window	107
4.5.5	Updating the Statistics with Credible Scalar Quotes	108
4.5.6	A Second Scalar Window for Old Valid Quotes	108
4.6	The Full-Quote Filtering Window	109
4.6.1	Quote Splitting Depending on the Instrument Type	110
4.6.2	The Basic Validity Test	110
4.6.3	Transforming the Filtered Variable	112
4.7	Univariate Filtering	113

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>

Sorting and counting

```
$ wc finn
```

```
12361 114266 610157 finn
```

```
$ < words grep '^a' | grep 'e$' | sort | uniq -c | sort -rn
```

```
77 are
```

```
21 alone
```

```
20 ashore
```

```
19 above
```

```
13 alive
```

```
9 awhile
```

```
9 apiece
```

```
7 axe
```

```
7 agree
```

```
5 anywhere
```



35:37 / 1:31:24



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/>

Top 100 R Packages by Downloads



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

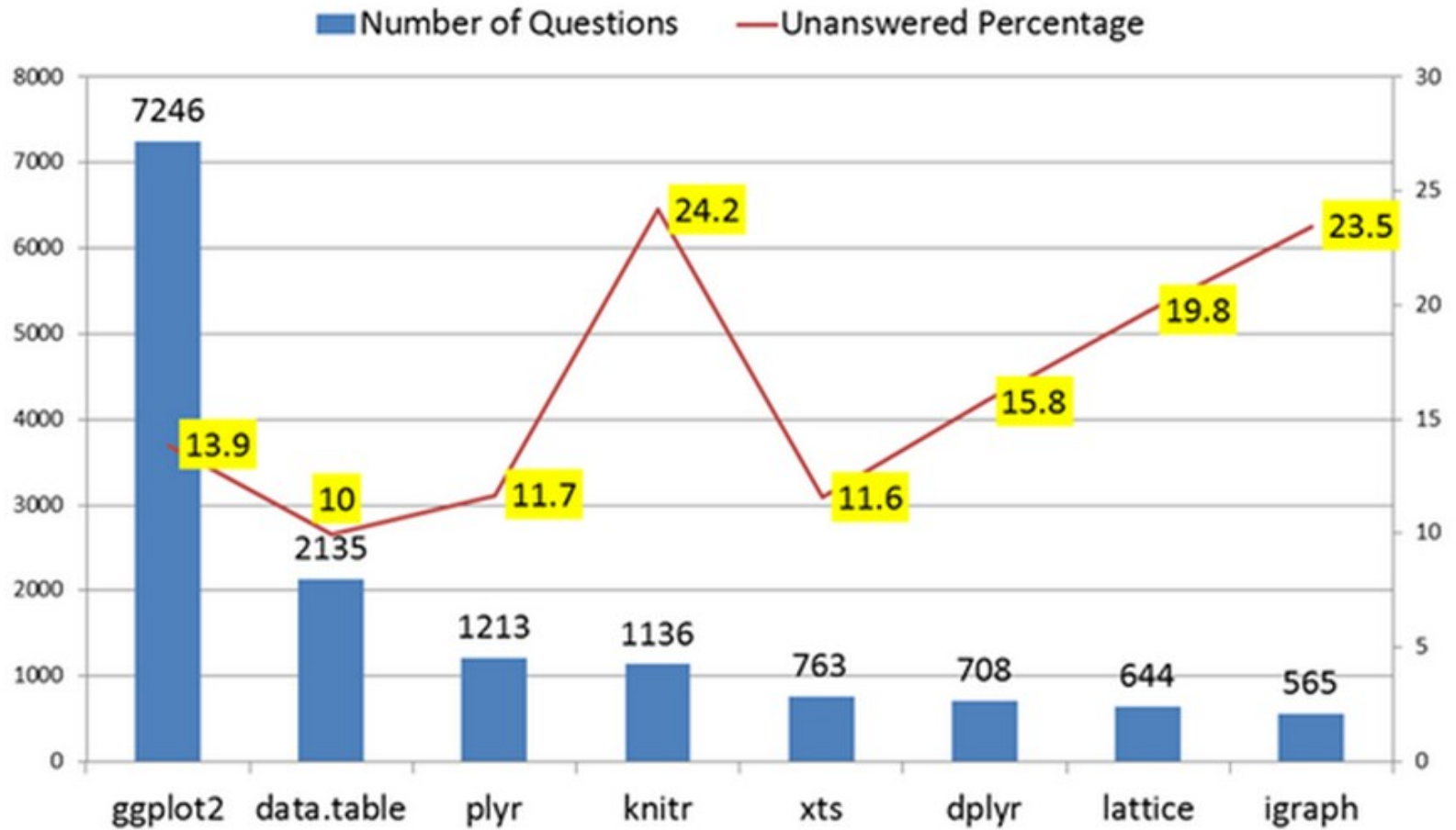
March
2015



Data Science Central

Number of Questions Asked & Unanswered for Top R Packages

April
2015



Source: *Vozag.com & Stack Overflow*



Data Science Central

- 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/>

YES, your data fits in RAM.

My data is:

TiB ▼

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