Picking an analysis tool: how to recognize the limits of your tools

James (JD) Long

Cory House 🏠
@housecor

A conference speaker's primary impact isn't teaching...It's getting you excited enough to learn more.

Conf speaking is sales, for ideas.

7:00 AM - Jul 30, 2017

♡ 1,443 💬 502 people are talking about this

My Only Points

- If you can manipulate data, you're more likely to be useful
- You may need to switch tools to remain useful
- Useful people get through life better

So Who Am I?

Morgan Stanley

KOREAN AIR

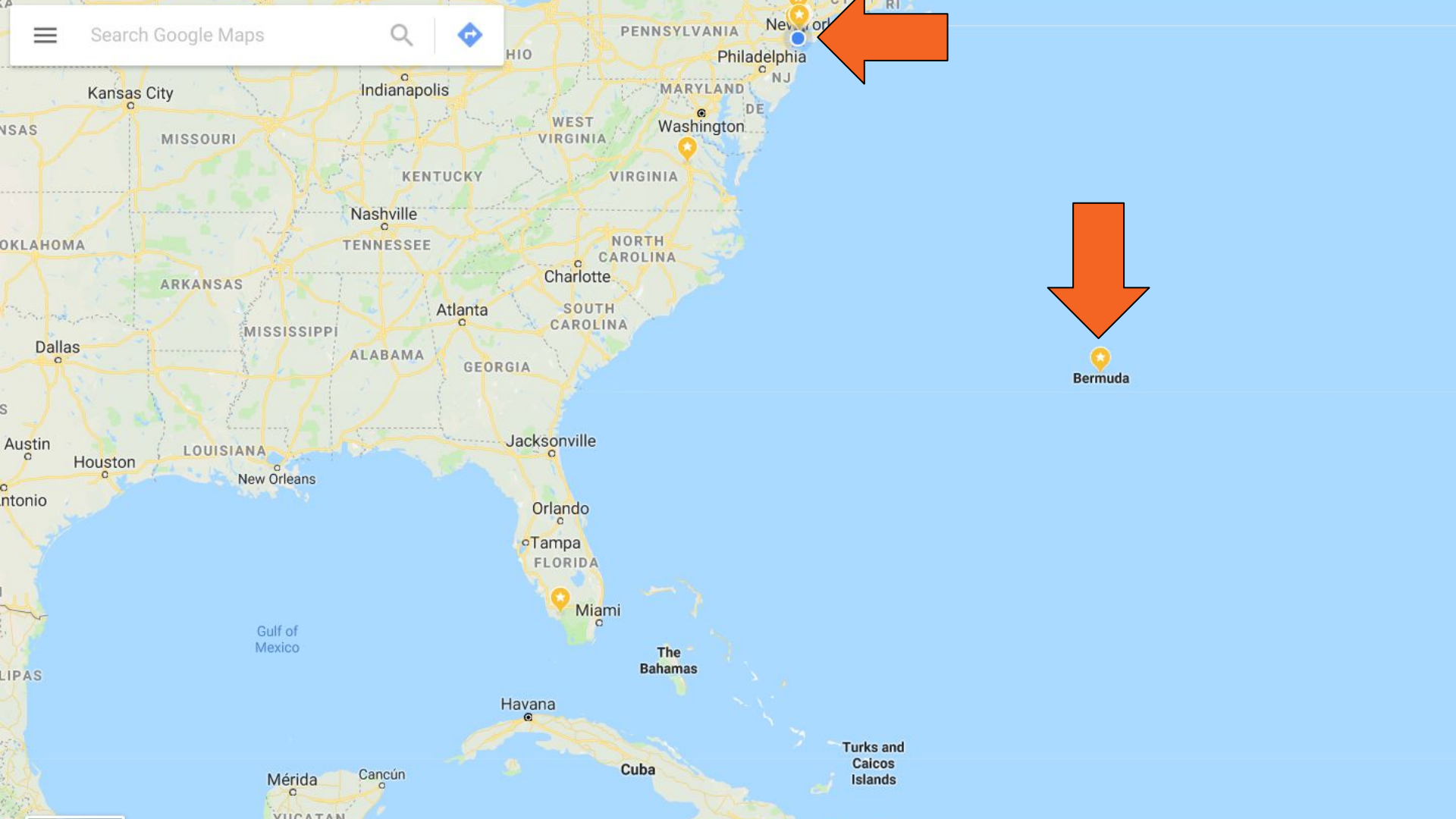
Tyson Foods, Inc.

MIRANT

PSEG

Swiss Re


St George's Channel
Ferry Point Park
St David's Island
Great Head Park
Coney Island Park
Castle Harbour
Cooper's Island Nature Reserve
Castle Harbour
Harrington Sound
Tucker's Town
Ireland Island
Flatt's Village
Ocean View Golf Course
Spittal Pond Nature Reserve
Somerset Village
Grand Bermuda
Hamilton
Great Sound
Ely's Harbour
Paget Marsh
Elbow Beach
Little Sound
Marley Beach
Port Royal Golf Course
Barnes Corner Park


Search Google Maps
PENNSYLVANIA
Philadelphia
NJ
MARYLAND
DE
Washington
Kansas City
Indianapolis
MISSOURI
WEST VIRGINIA
KENTUCKY
VIRGINIA
Nashville
TENNESSEE
NORTH CAROLINA
Charlotte
ARKANSAS
Atlanta
SOUTH CAROLINA
MISSISSIPPI
ALABAMA
GEORGIA
Dallas
Austin
Houston
LOUISIANA
New Orleans
Jacksonville
Orlando
Tampa
FLORIDA
Miami
Gulf of Mexico
The Bahamas
Havana
Cuba
Turks and Caicos Islands
Mérida
Cancún
Bermuda

What’s the most used data analysis tool in the world?


Excel

Spreadsheets? Sure!

- Easy to know where to start... just add data
- Simple data structure
- Visual formula creation: click on the numbers you want
- See the numbers change!
- Easily accessible to almost everyone

Excel is Programming

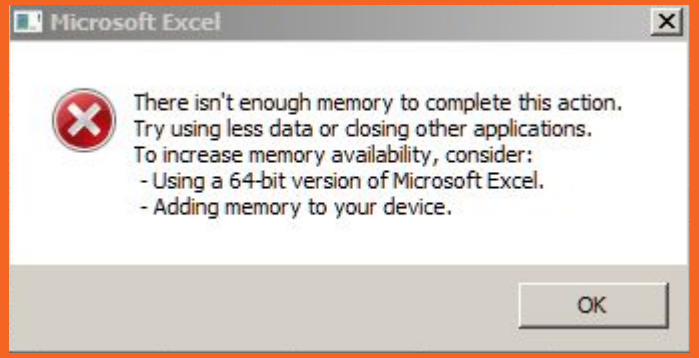
“The purpose of programming is to find a sequence of instructions that will automate performing a specific task or solving a given problem”

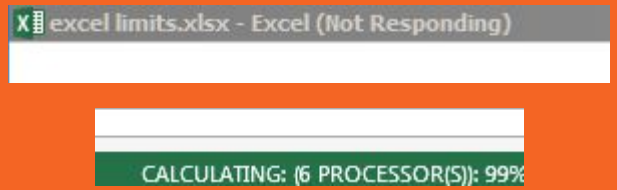
Don’t let anyone tell you it’s not real programming

So What's the Problem?


USING THE RIGHT TOOL
SAVES
LOTS OF PAIN

What Does it Feel Like When You Outgrow Excel?





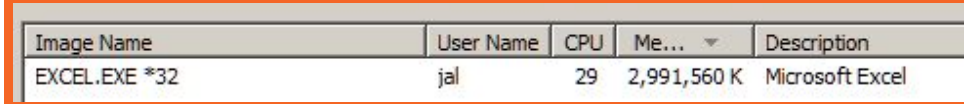
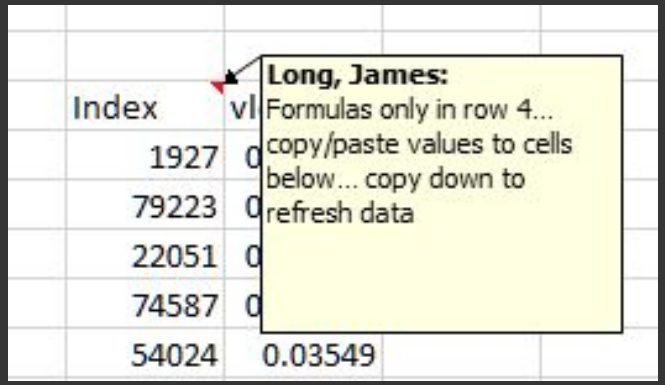
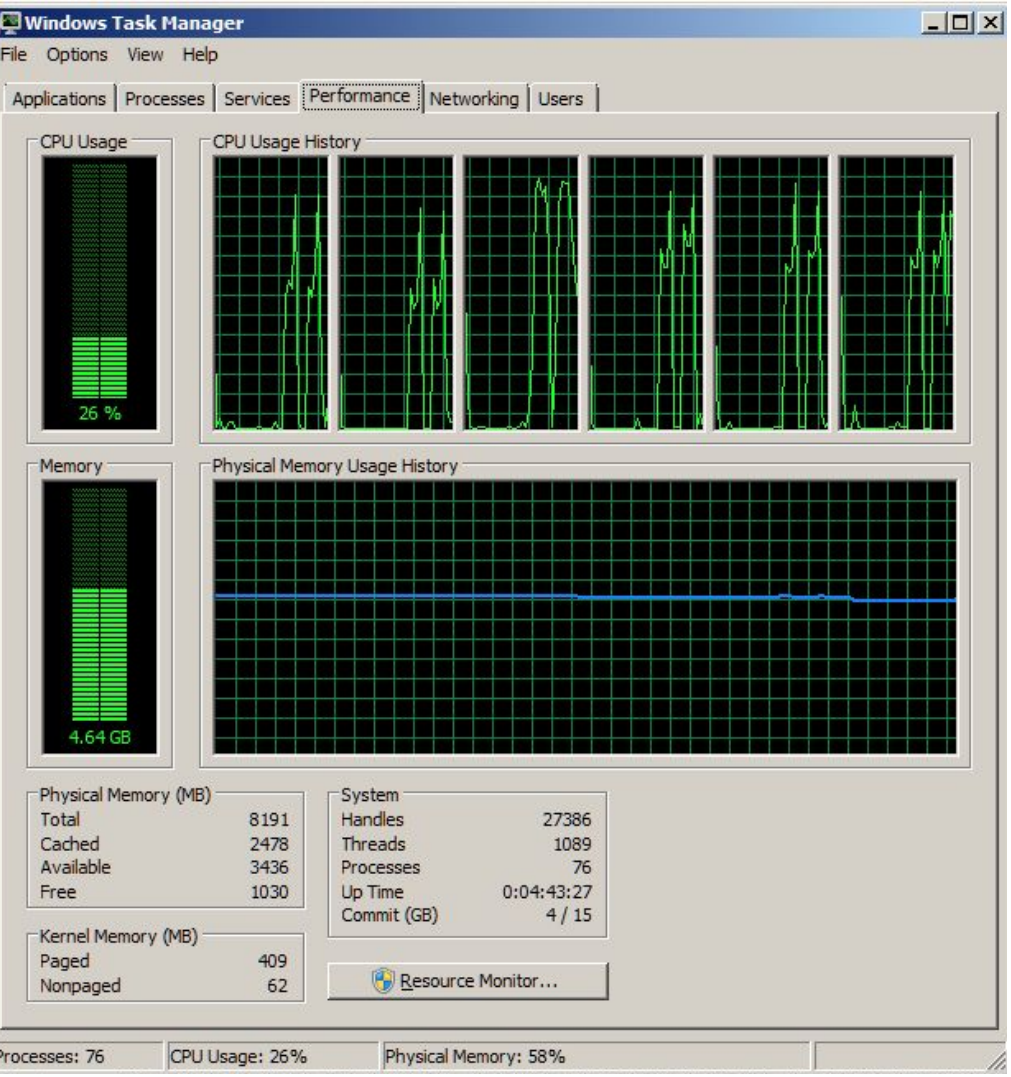


Image Name	User Name	CPU	Me...	Description
EXCEL.EXE *32	jal	29	2,991,560 K	Microsoft Excel

800MB File



This is how you punish future you… or your colleagues



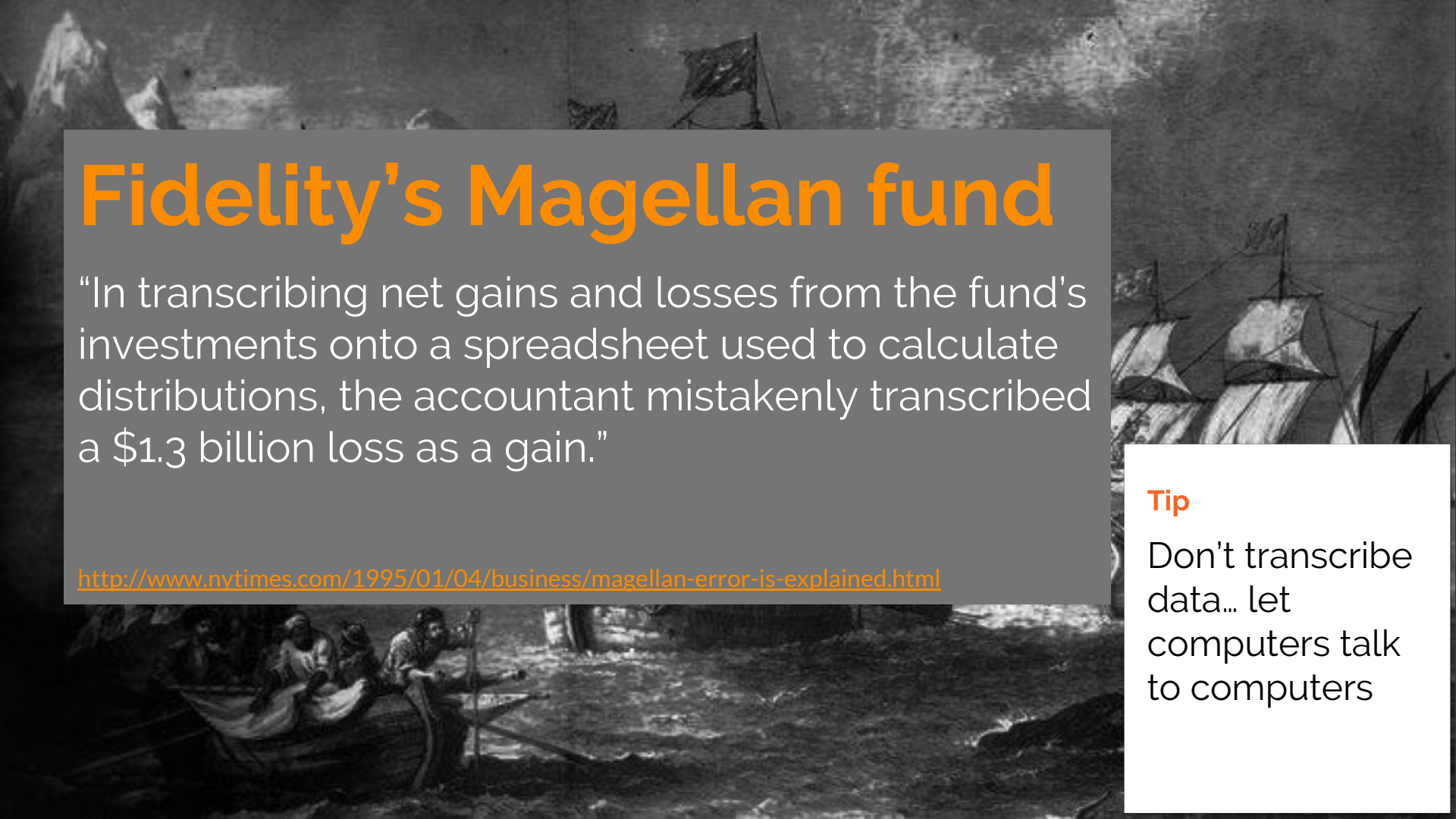
But not just computing performance...

THE WALL STREET JOURNAL.

BUSINESS | CFO JOURNAL

Stop Using Excel, Finance Chiefs Tell Staffs

Ubiquitous spreadsheet software that revolutionized accounting hasn't kept up, CFOs say



Fidelity's Magellan fund

"In transcribing net gains and losses from the fund's investments onto a spreadsheet used to calculate distributions, the accountant mistakenly transcribed a $1.3 billion loss as a gain."

http://www.nytimes.com/1995/01/04/business/magellan-error-is-explained.html

Tip

Don't transcribe data… let computers talk to computers

2010 Scientific paper redacted

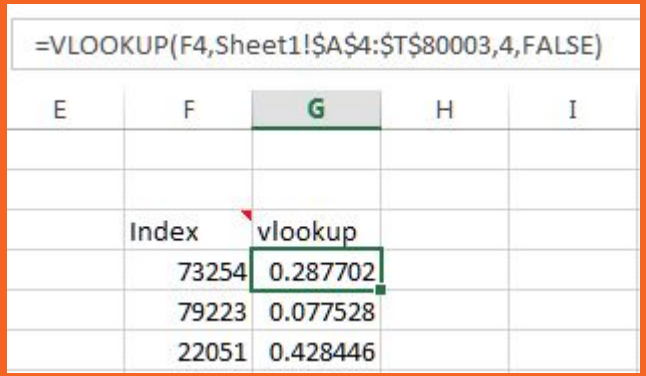
Formula did not include full range...

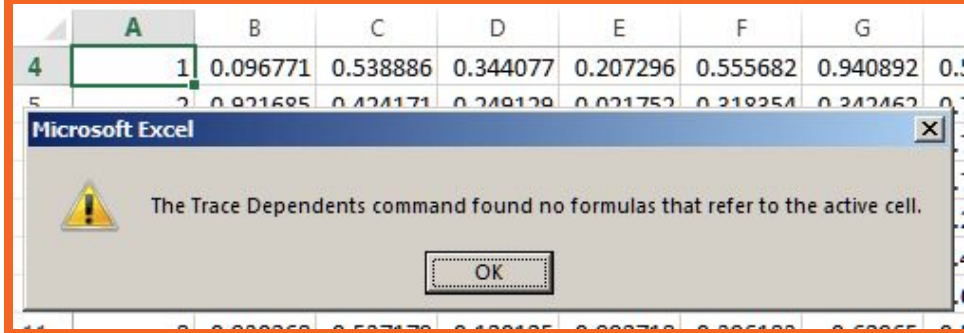
	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

http://www.businessinsider.com/editor-talks-reinhart-rogoff-excel-error-2013-4

Spreadsheets cannot be “read” linearly...

Each cell must be inspected to see where it’s used... and that doesn’t always work





Other Excel Pain Points

Creating multiple models within a spreadsheet can be a maintenance nightmare

Graphs may not refresh when a new row of data is added to a table

Pivot Tables may gain or lose rows breaking formulas

These are actual things that have caused me pain...

Alternatives?

Learning a little bit of coding will go a long way...

It's not just for comp sci nerds...

Example of well commented R code

This is the same as what I did in my unstable Excel toy example

```
# create a 250 col x 80000 row matrix of random uniform (0,1) numbers
random_numbers <- matrix( runif(250 * 80000), ncol=250)

#square root of the random numbers
sqrt_random_numbers <- sqrt(random_numbers)

#mean of each column
column_means <- colMeans(sqrt_random_numbers)

# create a random index of 80000 values between 1 and 80000 then get the data from sqrt_random_numbers
# from the 4th column and put that all in a vector
resample_index <- sample(1:80000, 80000, replace=T)

## pulls the item from the 4th column of the squared matrix that correspond to the resampled index
resorted_4th_column <- sqrt_random_numbers[ resample_index ,4 ]
```

Run time: 1.006 seconds

Run it yourself: https://gist.github.com/CerebralMastication/489e0c728dcdca398cebd043b9866669


But JD, I hear that big data is all the hotness right now?

3 Buckets for Data Size

Small Data: Easy to use on your PC

16GB RAM ~
4 GB Data

Medium Data: Fits on a server

My server:
504GB ~ 126GB

EC2:
1952GB ~ 488GB

Big Data: Needs to be spread across multiple servers

500GB+ Data

Assuming we need 4x data size of RAM to work with the data

“Big Data” bucket

90% - can be reduced to “small data” problems through sampling, subsetting, or summarizing

9% - can be reduced to multiple small data problems through chunking and iterating

1% - irreducibly big

If requires many jobs, then stay in R or Python but use a parallel backend: Hadoop or Spark for example.

If summary or reduced data is needed, possibly a column data store like Amazon Redshift easily queryable from R or Python

For irreducibly big data problems special engineering will be needed… these are rare.

Source: Hadley Wickham at RStudio: https://www.rstudio.com/resources/webinars/pipelines-for-data-analysis-in-r/

James (JD) Long
jdlong@gmail.com
Twitter:@cmastication
GitHub:

https://github.com/CerebralMastication/WestPointPresentation