spreadsheets

@JennyBryan

@jennybc

STAT 545

@STAT545

http://stat545.com

relevant links, credits, and slides:

https://github.com/jennybc/2016-06_spreadsheets

Rich FitzJohn

Research Software Engineer
University College London

@rgfitzjohn

@richfitz

spreadsheets:
a dystopian moonscape
of unrecorded
user actions

— Gordon Shotwell

some of my best friends use spreadsheets

I supported myself for ~4 years doing spreadsheets

~1 billion use Microsoft Office

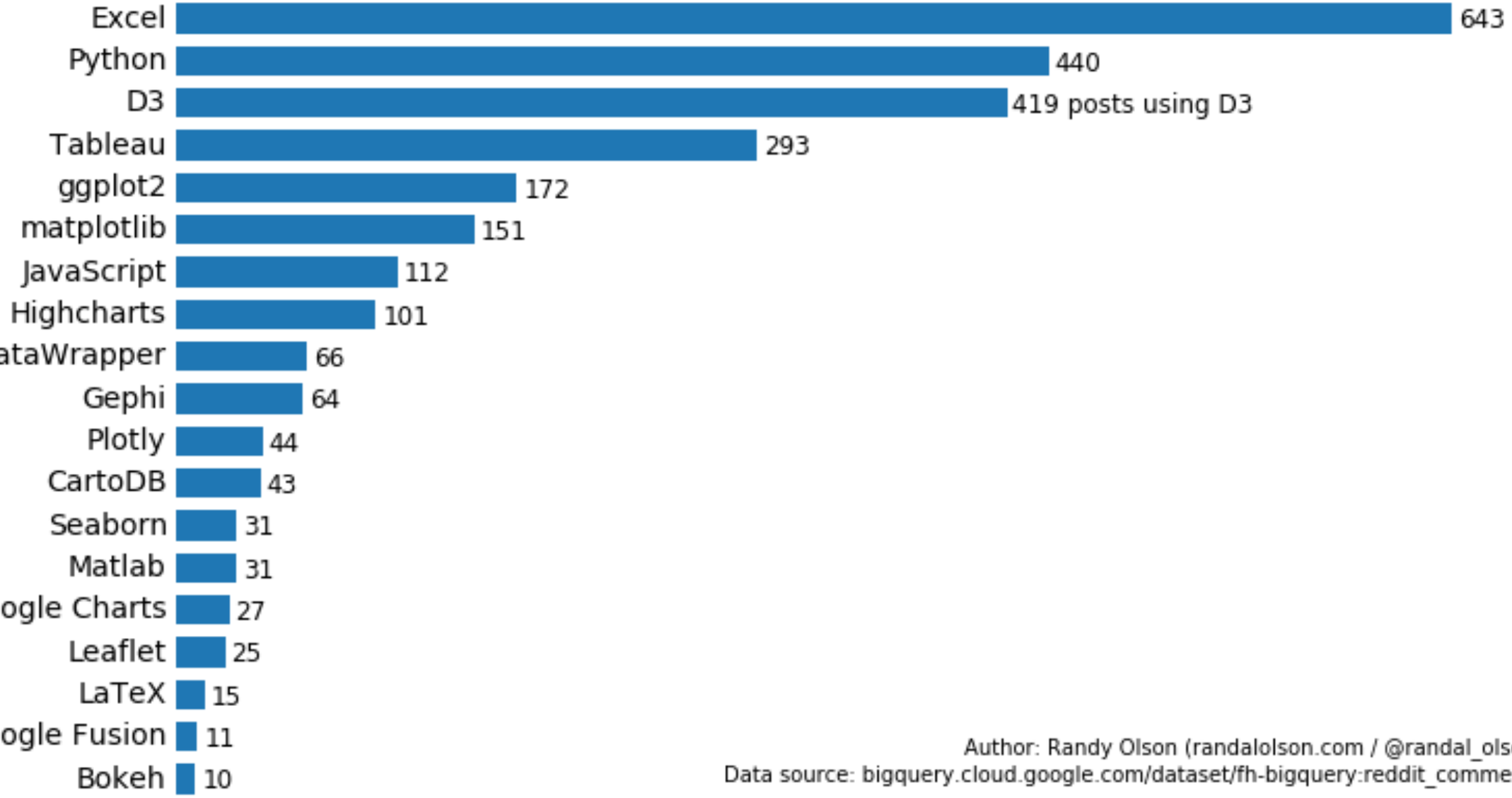
~650 million use spreadsheets

>50% use formulas

1 - 5 million people use Python

250K - 1 million people use R

Tools used in /r/DataIsBeautiful OC posts, 2014-2016



Author: Randy Olson (randalolson.com / @randal_olson)
Data source: bigquery.cloud.google.com/dataset/fh-bigquery:reddit_comments

you go into data analysis with the tools you know, not the tools you need

spreadsheets combine:

- data
- logic
- figures
- formatted tables

+ reactivity

spreadsheets users
use workbooks
like I would use
a data analytic git repo

a data analytic project:

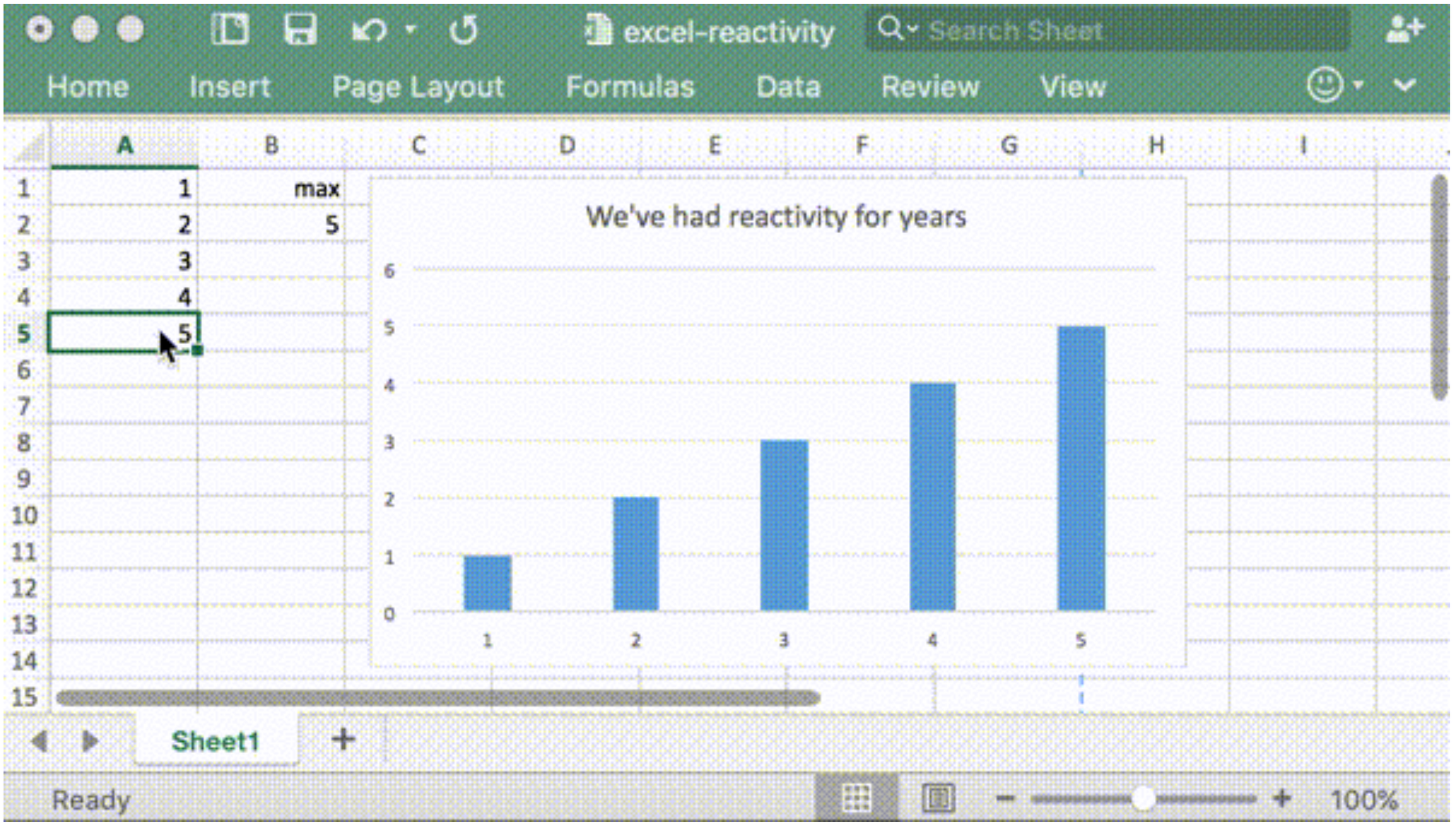
data

.R, .Rmd

.png, .svg

.md, .html, .pdf, Shiny app

+ build and deploy


excel-reactivity
Search Sheet
Home
Insert
Page Layout
Formulas
Data
Review
View
A
B
C
D
E
F
G
H
I
max
5
We've had reactivity for years
6
5
4
3
2
1
0
1
2
3
4
5
Sheet1
Ready
100%

syntax bullshittery

spreadsheets are not going away

deal with it

Jenny Bryan @JennyBryan · Apr 20

I'm seeking TRUE, crazy spreadsheet stories. Happy to get the actual sheet or just a description of the crazy. Also: I can keep a secret.

Based on surveys made by members of the Staff of Indiana University and supported by the National Research Council with Rockefeller Foundation funds

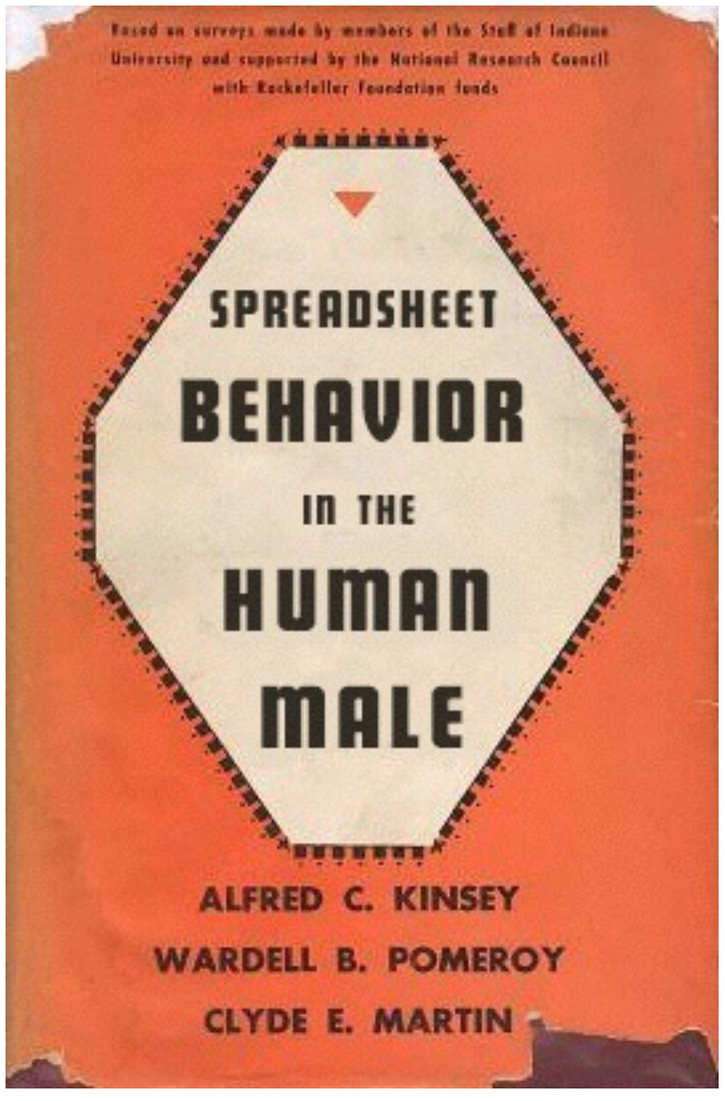
SEXUAL BEHAVIOR IN THE HUMAN MALE

ALFRED C. KINSEY
WARDELL B. POMEROY
CLYDE E. MARTIN

Based on surveys made by members of the Staff of Indiana University and supported by the National Research Council with Rockefeller Foundation funds

SPREADSHEET BEHAVIOR IN THE HUMAN MALE

ALFRED C. KINSEY
WARDELL B. POMEROY
CLYDE E. MARTIN



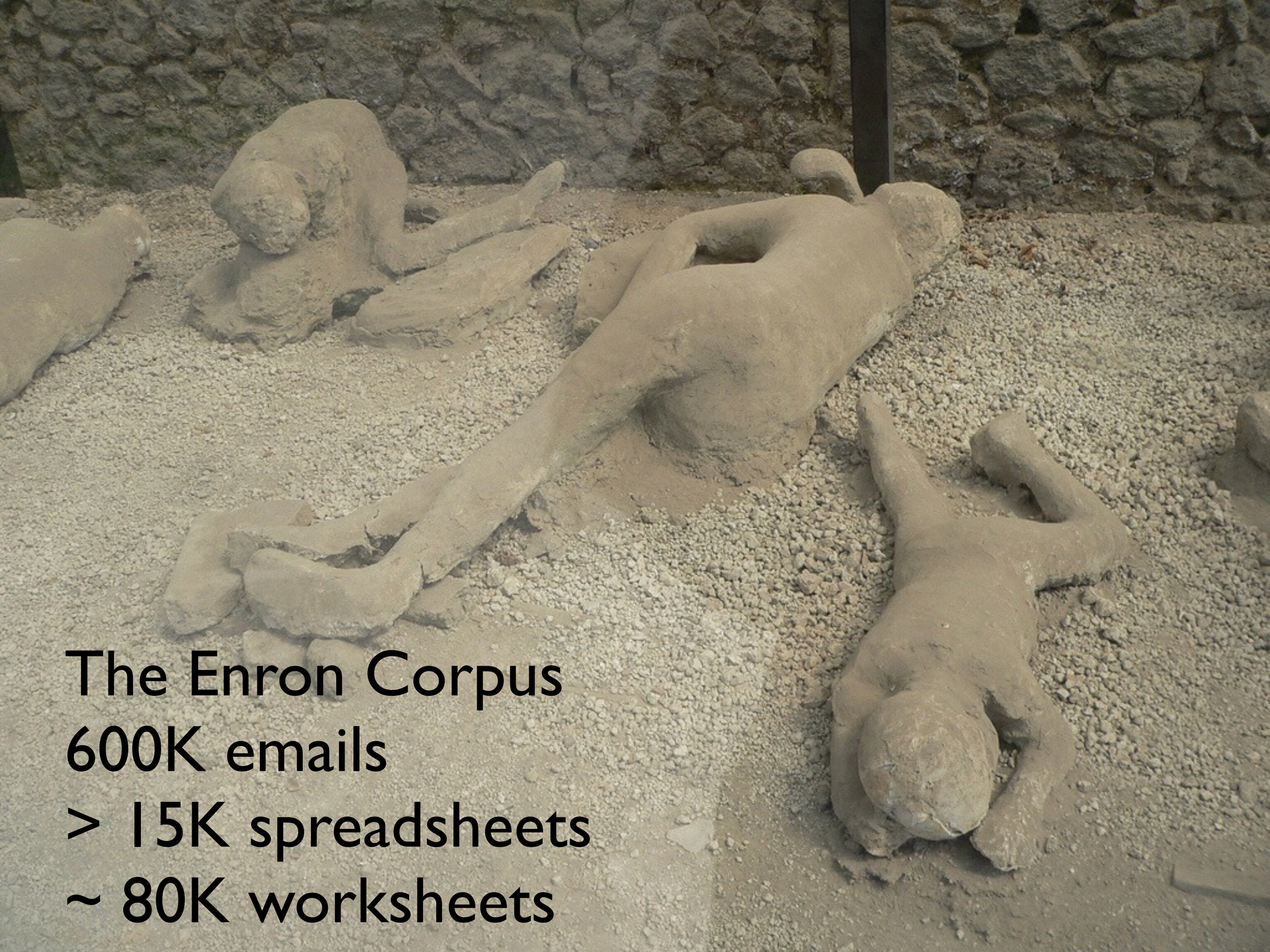
what you THINK people are doing

!=

what you think people SHOULD be doing

!=

what people ARE ACTUALLY doing



The Enron Corpus
600K emails
> 15K spreadsheets
~ 80K worksheets

Enron North America - West Gas

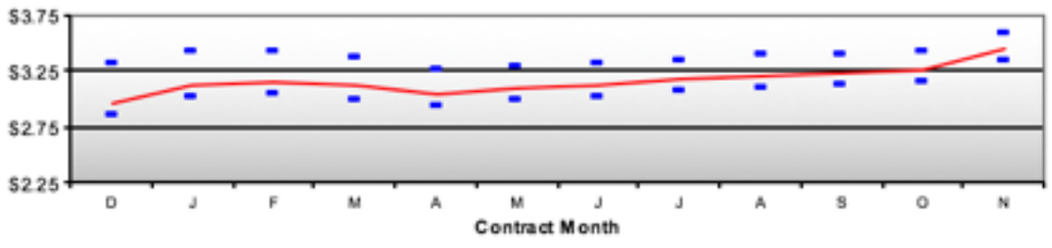
November 9, 2001

ENA - West Gas Contacts

Houston Office		Regional Offices		
Barry Tycholiz	(713) 853-1587	Mark Whitt	(303) 575-6473	Denver
Kim Ward	(713) 853-0685	Paul Lucci	(303) 575-6474	Denver
Stephanie Miller	(713) 853-1688	Tyrell Harrison	(303) 575-6478	Denver
Philip Polsky	(713) 853-5181	Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

	NYMEX	
	SETTLE	Δ
Cash		
ROM		
Dec-01	2.960	0.090
Dec-01 to Mar-02	3.088	0.083
Apr-02 to Oct-02	3.166	0.084
Nov-02 to Mar-03	3.651	0.090
One Year Strip*	3.165	0.084



IF NWPL Rocky Mountains			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
1.890	1.910		
2.060	2.080		
2.395	2.415	(0.565)	(0.545)
2.594	2.614	(0.494)	(0.474)
2.581	2.601	(0.585)	(0.565)
3.356	3.376	(0.295)	(0.275)
2.634	2.654	(0.530)	(0.510)

IF CIG Rocky Mountains				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	1.940	1.960		
ROM	1.960	1.980		
Dec-01	2.345	2.365	(0.615)	(0.595)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)
One Year Strip*	2.551	2.571	(0.614)	(0.594)

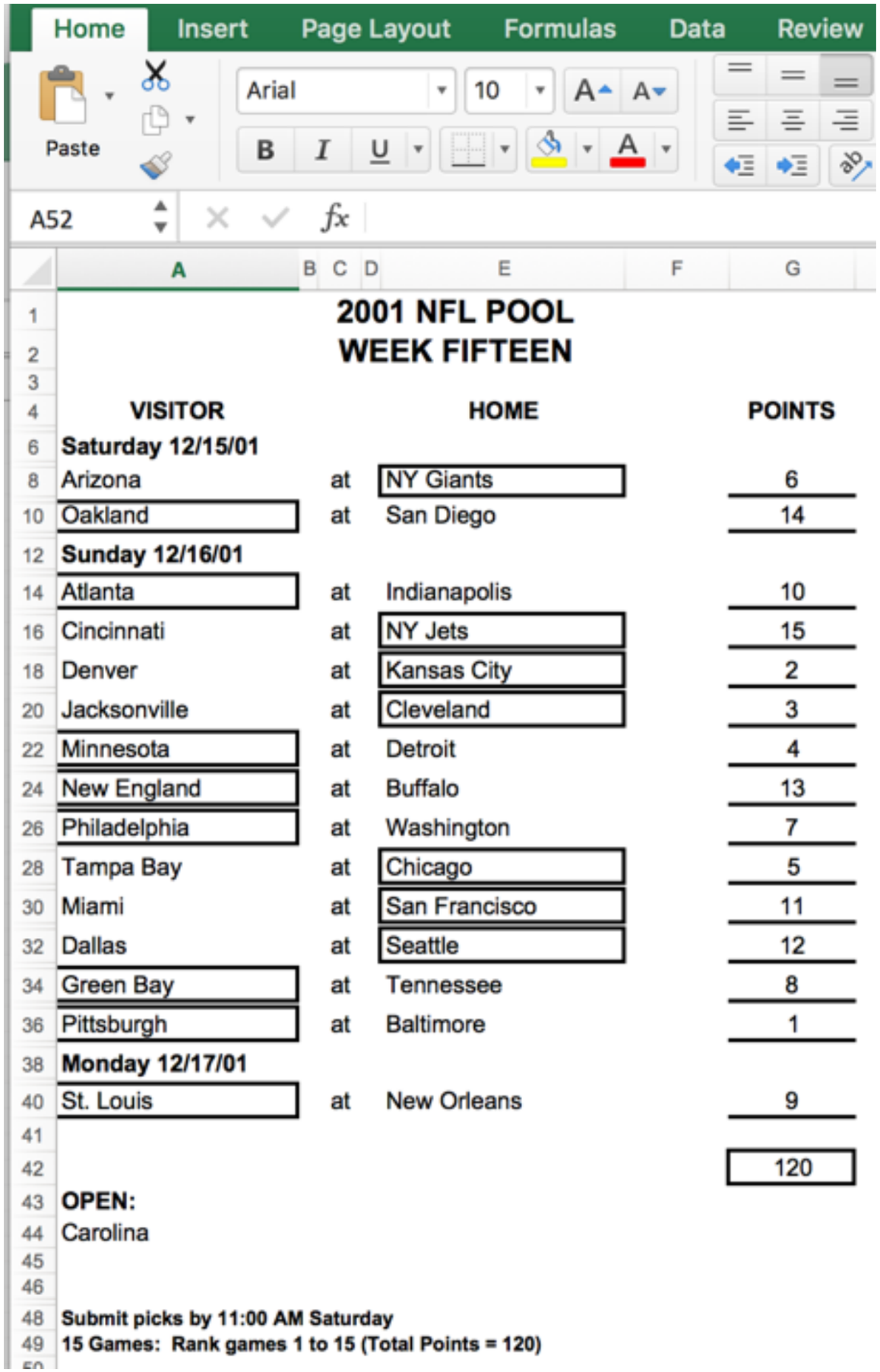
IF EL Paso Permian				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.375	2.395		
ROM	2.420	2.440		
Dec-01	2.700	2.720	(0.260)	(0.240)
Dec-01 to Mar-02	2.855	2.875	(0.233)	(0.213)
Apr-02 to Oct-02	3.009	3.029	(0.158)	(0.138)
Nov-02 to Mar-03	3.499	3.519	(0.153)	(0.133)
One Year Strip*	2.982	3.002	(0.182)	(0.162)

IF EL Paso San Juan				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.450	2.470		
ROM	2.350	2.370		
Dec-01	2.560	2.580	(0.400)	(0.380)
Dec-01 to Mar-02	2.743	2.763	(0.345)	(0.325)
Apr-02 to Oct-02	2.801	2.821	(0.365)	(0.345)
Nov-02 to Mar-03	3.421	3.441	(0.230)	(0.210)
One Year Strip*	2.817	2.837	(0.347)	(0.327)

AECO / NIT				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.376	2.396		
ROM	2.398	2.418		
Dec-01	2.552	2.572	(0.408)	(0.388)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)
One Year Strip*	2.676	2.696	(0.488)	(0.468)

IF NWPL Canadian Border (Sumas)				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.480	2.500		
ROM	2.460	2.480		
Dec-01	2.800	2.820	(0.160)	(0.140)
Dec-01 to Mar-02	2.892	2.912	(0.196)	(0.176)
Apr-02 to Oct-02	2.796	2.816	(0.370)	(0.350)
Nov-02 to Mar-03	3.706	3.726	0.055	0.075
One Year Strip*	2.880	2.900	(0.285)	(0.265)

IF PEPL TX-OK				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.530	2.550		
ROM	2.530	2.550		
Dec-01	2.828	2.848	(0.133)	(0.113)
Dec-01 to Mar-02	2.958	2.978	(0.130)	(0.110)
Apr-02 to Oct-02	3.046	3.066	(0.120)	(0.100)
Nov-02 to Mar-03	3.531	3.551	(0.120)	(0.100)
One Year Strip*	3.041	3.061	(0.123)	(0.103)


Home
Insert
Page Layout
Formulas
Data
Review
Paste
Arial
10
A52
2001 NFL POOL
WEEK FIFTEEN
VISITOR
HOME
POINTS
Saturday 12/15/01
Arizona at NY Giants 6
Oakland at San Diego 14
Sunday 12/16/01
Atlanta at Indianapolis 10
Cincinnati at NY Jets 15
Denver at Kansas City 2
Jacksonville at Cleveland 3
Minnesota at Detroit 4
New England at Buffalo 13
Philadelphia at Washington 7
Tampa Bay at Chicago 5
Miami at San Francisco 11
Dallas at Seattle 12
Green Bay at Tennessee 8
Pittsburgh at Baltimore 1
Monday 12/17/01
St. Louis at New Orleans 9
120
OPEN:
Carolina
Submit picks by 11:00 AM Saturday
15 Games: Rank games 1 to 15 (Total Points = 120)

	EUSES	Enron
Number of spreadsheets analyzed	4,447	15,770
Number of spreadsheets with formulas	1,961	9,120
Number of worksheets	16,853	79,983
Maximum number of worksheets	106	175
Number of non-empty cells	8,209,095	97,636,511
Average number of non-empty cells per spreadsheet	1,846	6,191
Number of formulas	730,186	20,277,835
Average of formulas per spreadsheet with formulas	372	2,223
Number of unique formulas	65,143	913,472
Number of unique formulas per spreadsheet with formulas	33	100

from Hermans, Murphy-Hill

data in formatting

Plot: 2			
Date collecte	Species	Sex	Weight
1/8/14	NA		
1/8/14	DM	M	44
1/8/14	DM	M	38
1/8/14	OL		
1/8/14	PE	M	22
1/8/14	DM	M	38
1/8/14	DM	M	48
1/8/14	DM	M	43
1/8/14	DM	F	35
1/8/14	DM	M	43
1/8/14	DM	F	37
1/8/14	PF	F	7
1/8/14	DM	M	45
1/8/14	OT		
1/8/14	DS	M	157
1/8/14	OX		
2/18/14	NA	M	218
2/18/14	PF	F	7
2/18/14	DM	M	52
	measurement device not calibrated		

	A	B	C
1	species	tail length	
2	Allactaga balikunica	**177.32**	
3	Allactaga bullata	165.2	
4	Allocricetulus eversmanni	**18.64**	
5	Apodemus uralensis	84.89	
6	Arvicola amphibius	105.14	
7	Brachytarsomys albicauda	230.02	
8	Brachyuromys betsileoensis	83.32	
9	Cardiocranius paradoxus	**74.36**	
10	Castor fiber	379.89	
11	Cricetulus barabensis	24.72	
12			
13	**bold = needs checking**		
14	yellow = from different source		
15			

small multiples

EXTRAPOLATED FROM TRIALS

Mean Peak Force (k | Mean Dev. Force (k | Mean Time to P.F. | Mean time to relax (ms)

Time (mi	Control	AAPH	Control	AAPH	Control	AAPH	Control	AAPH
0	11.69	N/A	11.32	N/A	412.2	N/A	441	N/A
5	10.53	7.53	8.92	6.25	373.8	390.7	360.7	435.3
8	9.809	8.165	8.218	6.534	379.2	397.8	350.2	440.5
12	9.255	9.113	6.723	6.129	391.8	390.7	421.3	461.7
18	7.54	8.823	5.341	5.757	395.5	393.2	516.8	543
24	6.311	8.351	3.059	5.379	384.9	377	501.4	550.3
40	2.723	8.353	-2.184	6.008	379.1	407.2	587.6	502.3
50	0.48	8.343	-5.46	6.052	375.5	424.2	641.4	546.8

Mean Net Work (mJ/g | Mean Shortening Work (mJ/g)

Time (mi	Control	AAPH	Control	AAPH
10	0.182	0.14	0.769	0.668
15	0.15	0.139	0.713	0.75
20	0.13	0.126	0.626	0.711
30	0.072	0.131	0.489	0.727
45	-0.01	0.141	0.275	0.634

Peak Force SE | Dev. Force SE | Time to P.F. SE | Time to relax SE

Time (mi	Control	AAPH	Control	AAPH	Control	AAPH	Control	AAPH
0	2.407	N/A	2.326	N/A	29.59	N/A	56.64	N/A
5	1.87	1.333	1.683	1.601	28.5	31.98	45.74	52.37
8	1.842	2.162	1.531	1.789	29.24	31.45	35.59	56.77
12	1.549	2.133	1.138	1.474	28.57	25.19	53.11	48.51
18	1.341	2.034	1.038	1.402	26.35	25.75	67.42	55.03
24	1.202	2.095	0.866	1.451	39.16	19.83	73.33	54.63
40	1.545	2.223	1.663	1.558	53.48	22.41	105.7	45.23
50	2.088	2.072	2.422	1.759	63.41	22.14	127.1	54.85

Net Work SE | Shortening Work SE

Time (mi	Control	AAPH	Control	AAPH
10	0.029	0.031	0.13	0.167
15	0.02	0.025	0.09	0.147
20	0.019	0.023	0.077	0.126
30	0.02	0.025	0.068	0.135
45	0.033	0.025	0.144	0.125

data in formulas

Excel File Edit View Insert Format Tools Data Window Help

snaildata

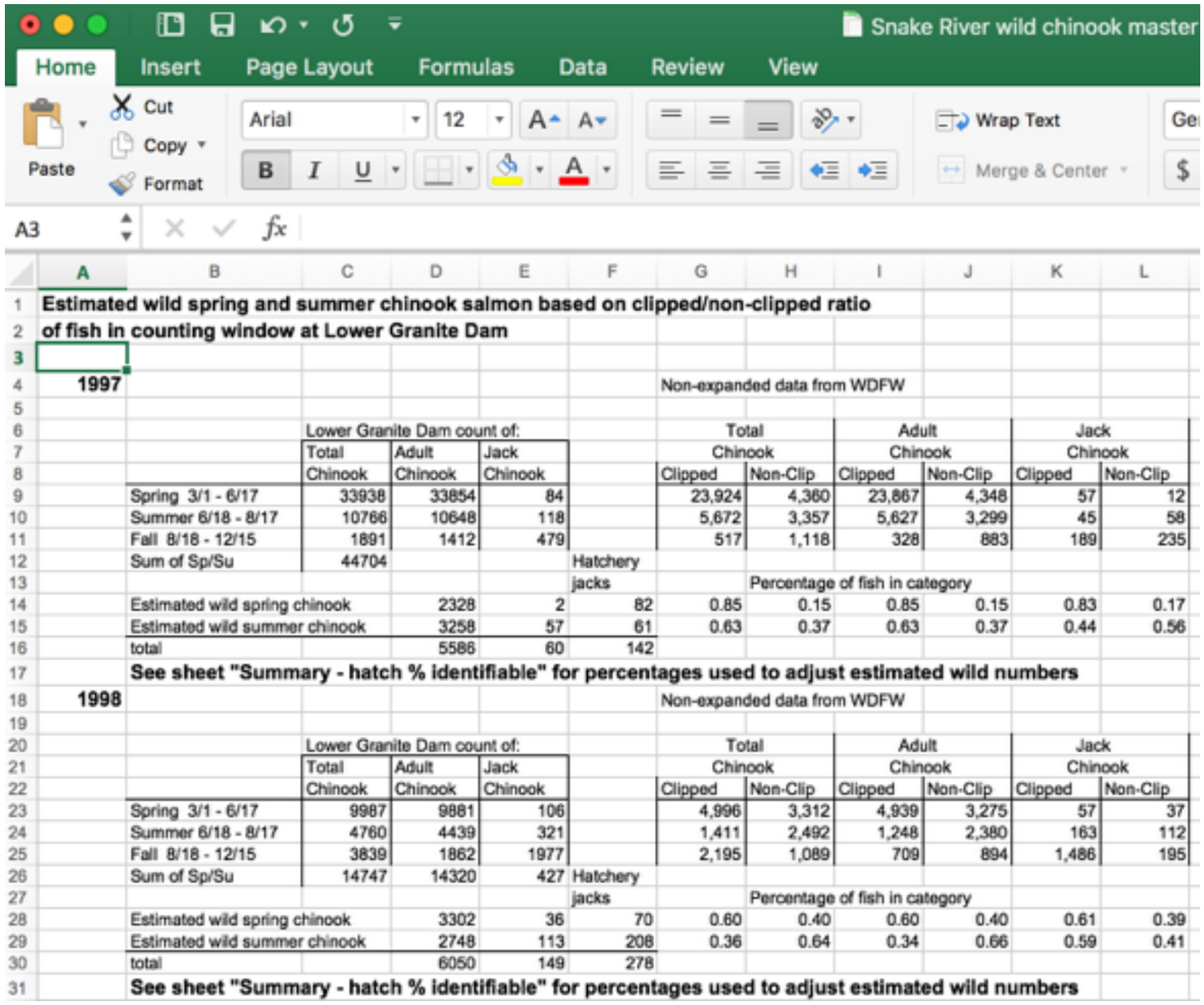
Home Insert Page Layout Formulas Data Review View

C5 =(3.6946*10^-6)/'Old snails'!J26

	A	B	C	D	E	F	G	H	I
1	oldsnail27_	peak force (kN/m^2)		developed force (kN/n		time to pf (ms)		time to relax (ms)	
2	Time (min)	Control	AAPH	Control	AAPH	Control	AAPH	Control	AAPH
3	0	15.50528	N/A	15.34497	N/A	362	N/A	251	N/A
4	5	15.79448	11.6517	12.52871	10.25405	346	335	244	278
5	8	15.15288	13.33965	12.02323	11.77049	343	349	239	278
6	12	14.46832	14.76727	10.72342	11.30112	358	353	280	298
7	18	12.05681	15.77029	8.95424	11.33722	337	371	313	317
8	24	11.5714	16.32993	6.679	11.69828	334.4244	369	324.122	317
9	40	8.429	16.33932	1.239	12.02323	319.074	413	385.13	336
10	50	6.465	15.78004	-2.161	13.97295	309.48	429	423.26	349

=(3.6946*10^-6)/'Old snails'!J26

data in (merged) column headers



Snake River wild chinook master

Home | Insert | Page Layout | Formulas | Data | Review | View

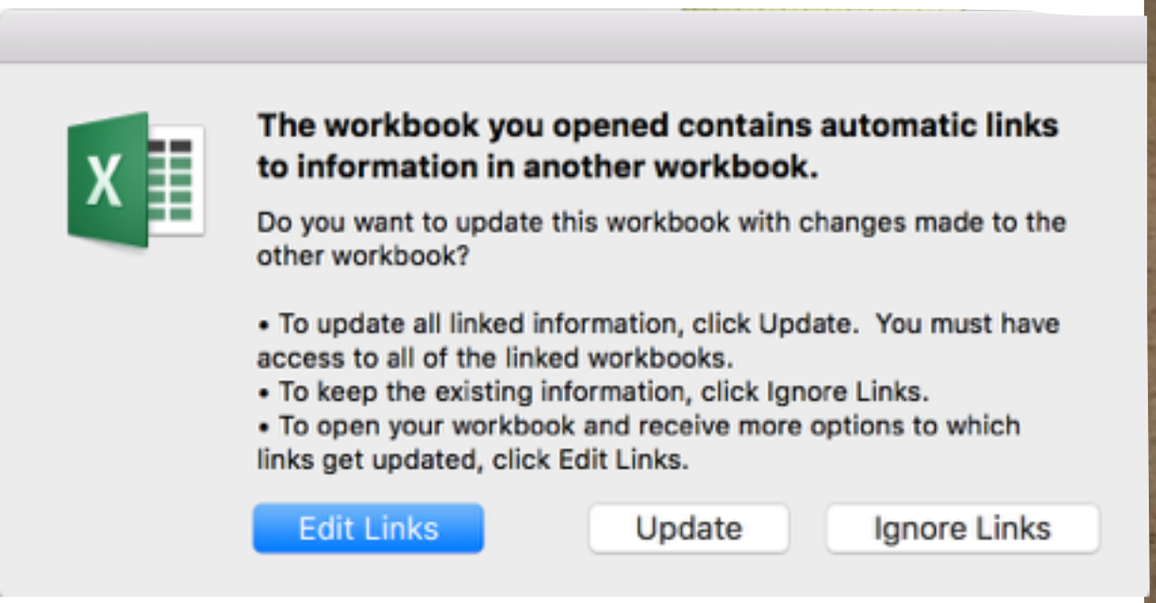
Paste | Cut | Copy | Format | Arial | 12 | Wrap Text | Merge & Center | $

A3

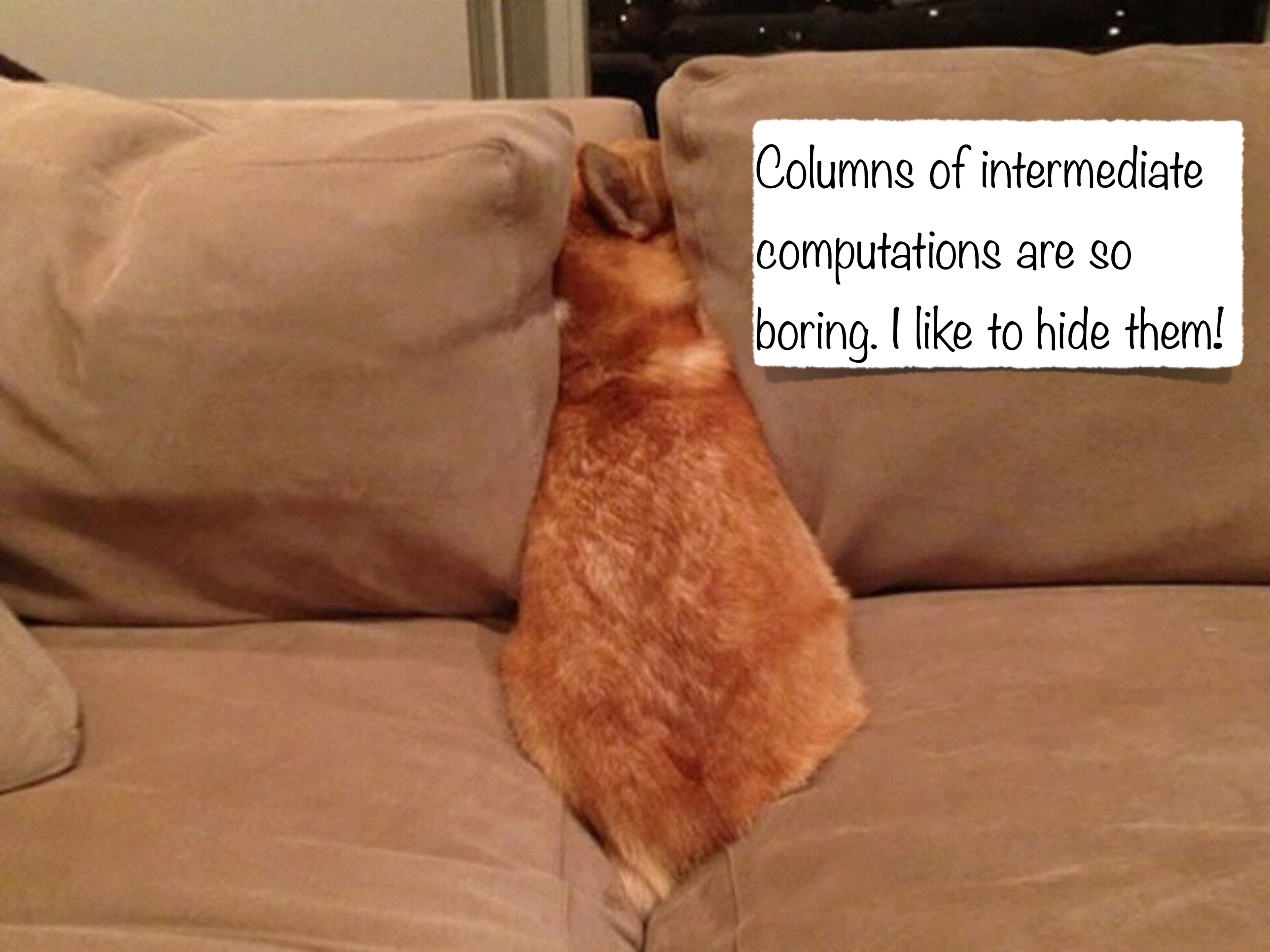
	A	B	C	D	E	F	G	H	I	J	K	L
1	**Estimated wild spring and summer chinook salmon based on clipped/non-clipped ratio**											
2	**of fish in counting window at Lower Granite Dam**											
3												
4	**1997**						Non-expanded data from WDFW					
5												
6			Lower Granite Dam count of:				Total Chinook		Adult Chinook		Jack Chinook	
7			Total Chinook	Adult Chinook	Jack Chinook		Clipped	Non-Clip	Clipped	Non-Clip	Clipped	Non-Clip
8		Spring 3/1 - 6/17	33938	33854	84		23,924	4,360	23,867	4,348	57	12
9		Summer 6/18 - 8/17	10766	10648	118		5,672	3,357	5,627	3,299	45	58
10		Fall 8/18 - 12/15	1891	1412	479		517	1,118	328	883	189	235
11		Sum of Sp/Su	44704			Hatchery						
12						jacks		Percentage of fish in category				
13		Estimated wild spring chinook		2328	2	82	0.85	0.15	0.85	0.15	0.83	0.17
14		Estimated wild summer chinook		3258	57	61	0.63	0.37	0.63	0.37	0.44	0.56
15		total		5586	60	142						
16		**See sheet "Summary - hatch % identifiable" for percentages used to adjust estimated wild numbers**										
17	**1998**						Non-expanded data from WDFW					
18												
19			Lower Granite Dam count of:				Total Chinook		Adult Chinook		Jack Chinook	
20			Total Chinook	Adult Chinook	Jack Chinook		Clipped	Non-Clip	Clipped	Non-Clip	Clipped	Non-Clip
21		Spring 3/1 - 6/17	9987	9881	106		4,996	3,312	4,939	3,275	57	37
22		Summer 6/18 - 8/17	4760	4439	321		1,411	2,492	1,248	2,380	163	112
23		Fall 8/18 - 12/15	3839	1862	1977		2,195	1,089	709	894	1,486	195
24		Sum of Sp/Su	14747	14320	427	Hatchery						
25						jacks		Percentage of fish in category				
26		Estimated wild spring chinook		3302	36	70	0.60	0.40	0.60	0.40	0.61	0.39
27		Estimated wild summer chinook		2748	113	208	0.36	0.64	0.34	0.66	0.59	0.41
28		total		6050	149	278						
29		**See sheet "Summary - hatch % identifiable" for percentages used to adjust estimated wild numbers**										

	A	B	C
1	name	mass	threat status
2	Rodentia		
3	rat	56	LC
4	mouse	90	LC
5	squirrel	24	CR
6	Primates		
7	gibbon	7000	LC
8	bushbaby	678	EN
9	Chiroptera		
10	myotis	45	NT
11	smaller myotis	48	NT
12	fishing bat	89	VU
13	bulldog bat	33	DD

My workbook has 17 sheets, but I transposed the data matrix in sheet 4, at random, for absolutely no reason.



We have formulas that refer to cells in the other. But you will only ever get one of us.


Columns of intermediate computations are so boring. I like to hide them!

ALGORITHMS
BY COMPLEXITY

MORE COMPLEX →

LEFTPAD

QUICKSORT

GIT MERGE

SELF-DRIVING CAR

GOOGLE SEARCH BACKEND

SPRAWLING EXCEL SPREADSHEET BUILT UP OVER 20 YEARS BY A CHURCH GROUP IN NEBRASKA TO COORDINATE THEIR SCHEDULING

machine readable

&

human readable

code
can be
machine & human readable

data
can be
machine & human readable

jennybc / lotr

branch: master | lotr / lotr_clean.tsv

file | 684 lines (683 sloc) | 42.64 kb | Open | Edit | Raw | Blame | History | Delete

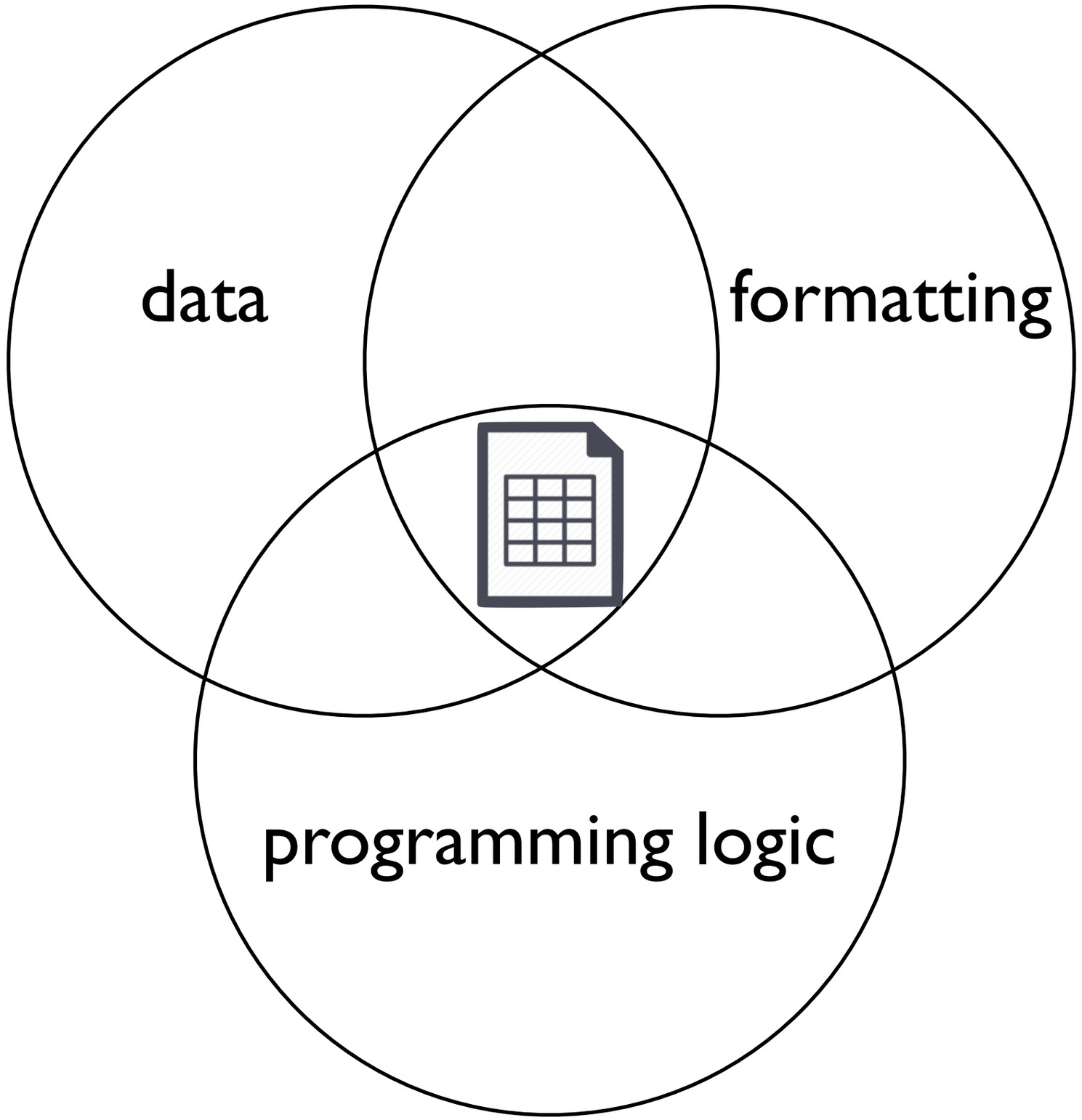
Search this file...

	Film	Chapter	Character	Race	Words
2	The Fellowship Of The Ring	01: Prologue	Bilbo	Hobbit	4
3	The Fellowship Of The Ring	01: Prologue	Elrond	Elf	5
4	The Fellowship Of The Ring	01: Prologue	Galadriel	Elf	460
5	The Fellowship Of The Ring	02: Concerning Hobbits	Bilbo	Hobbit	214
6	The Fellowship Of The Ring	03: The Shire	Bilbo	Hobbit	70
7	The Fellowship Of The Ring	03: The Shire	Frodo	Hobbit	128
8	The Fellowship Of The Ring	03: The Shire	Gandalf	Wizard	197
9	The Fellowship Of The Ring	03: The Shire	Hobbit Kids	Hobbit	10

a spreadsheet

is often neither

machine nor human readable


data
formatting
programming logic

what are the problems?

which ones can we solve?

via training

via tooling

be realistic,
be fair,
be precise

let 1,000 flowers bloom!

Two angles on the Spreadsheet Problem:

Create new spreadsheet implementations
that use, e.g., R for computation
and visualization.

Accept spreadsheets as they are.
Create tools to get goodies out and into, e.g., R.
Maybe write back into sheets?

Stencila

Sheets

Spreadsheets are the probably the most widely used environment for data analysis and end user programming. Their ubiquity comes from the intuitive simplicity of a live, reactive, see-your-data-and-do-stuff with it interface.

Stencila sheets combine the benefits of the spreadsheet interface with power of languages like R and Python.

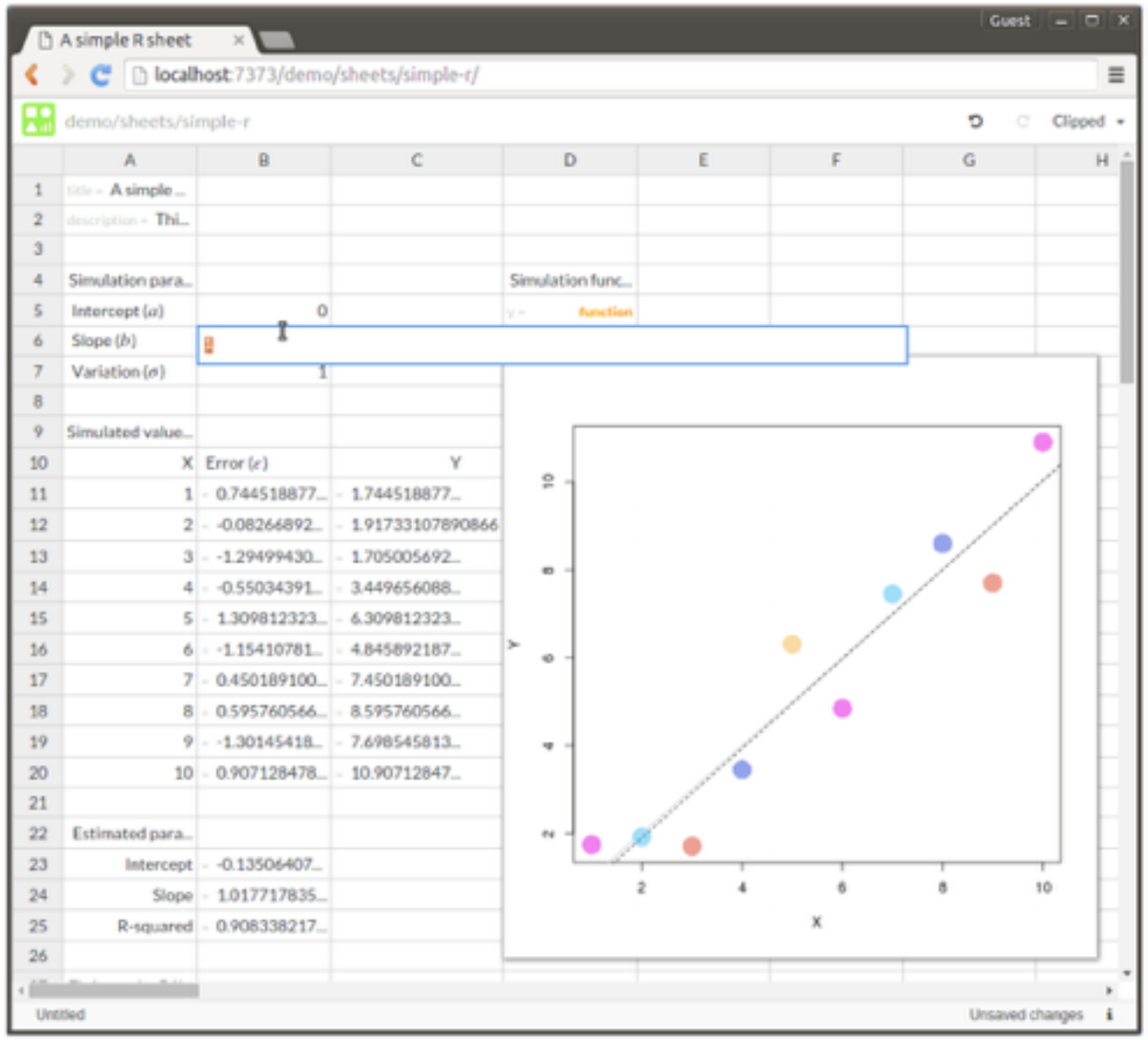
Built from the ground up for transparency, testability and version control, we think Stencila sheets are the next step in the evolution of the spreadsheet.

More:

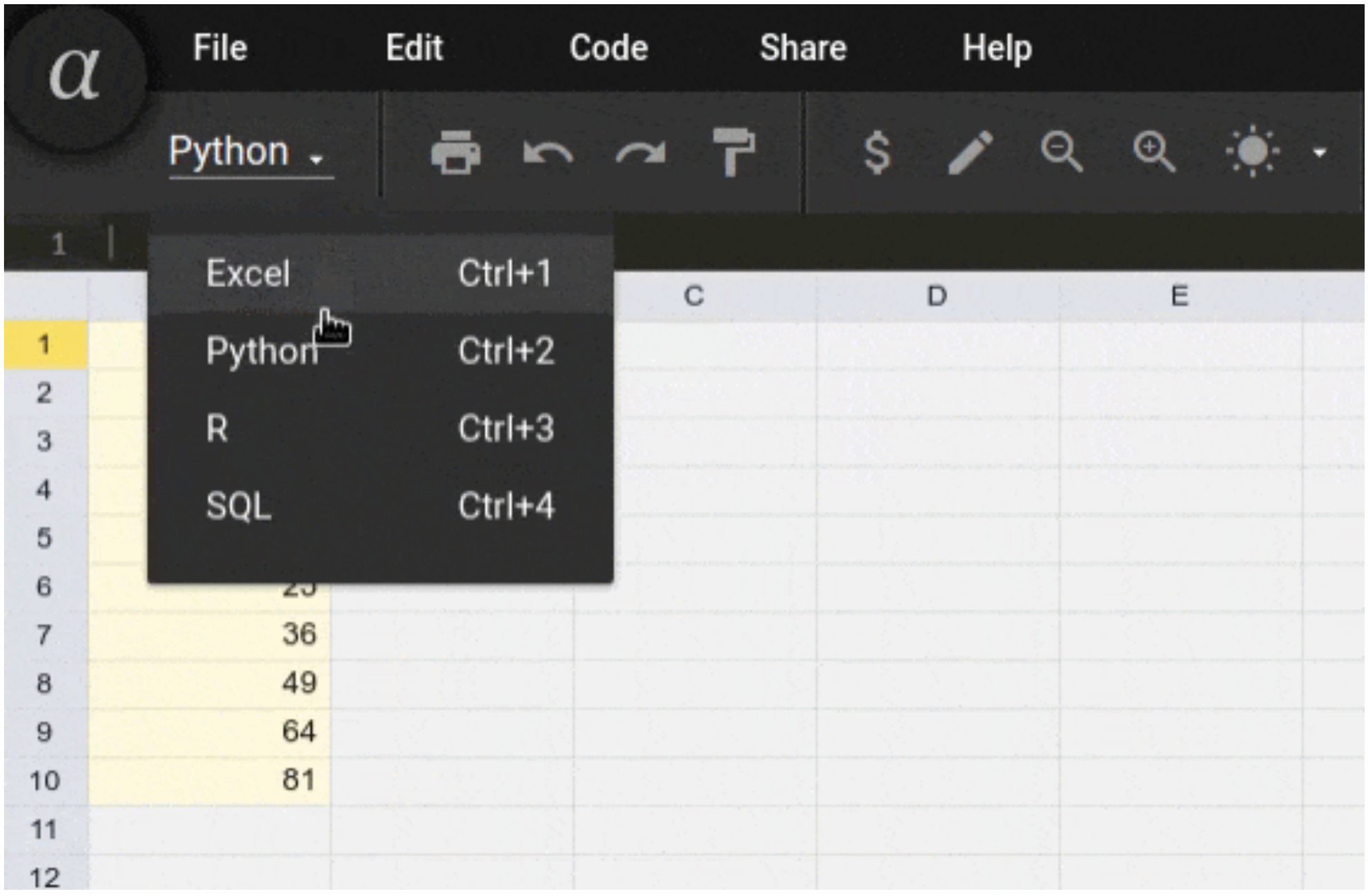
Spreadsheets are dead, long live reactive programming environments!

Getting under Stencila sheets

A spreadsheet file format for humans

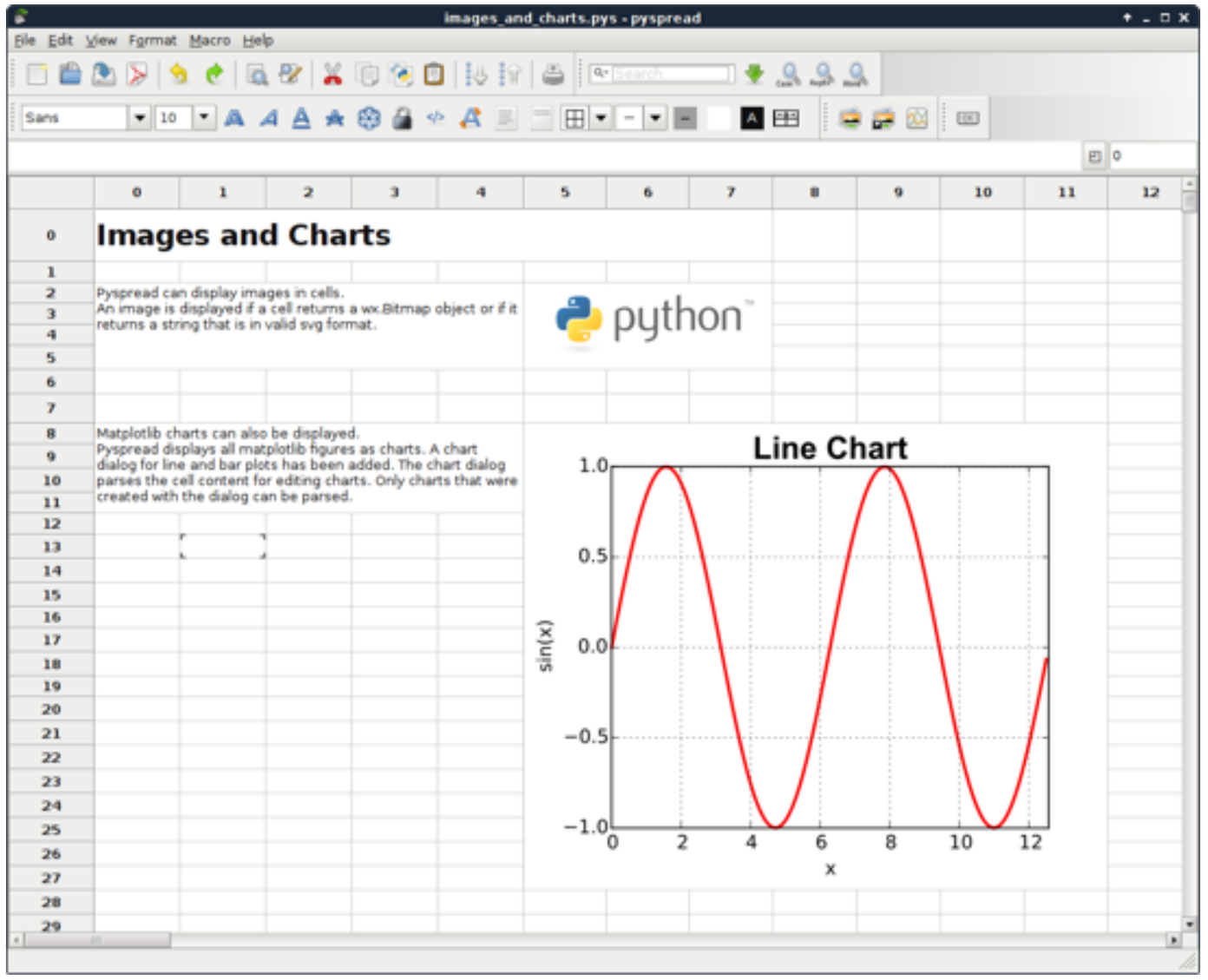


AlphaSheets
“collaborative, programmable spreadsheets”



pyspread

Python power for your tables



Pyspread is a non-traditional spreadsheet application that is based on and written in the programming language **Python**.

The goal of pyspread is to be the most pythonic spreadsheet.

Pyspread expects Python expressions in its grid cells, which makes a spreadsheet specific language obsolete. Each cell returns a Python object that can be accessed from other cells. These objects can represent anything including lists or matrices.

Pyspread is free software. It is released under the **GPL v3**. You can find the sorce code at **github**.

Jean Adams
@JeanVAdams

If your collaborator asks, “In what form would you like the data?” you should respond, “In its current form.” via @kwbroman

Enron North America - West Gas

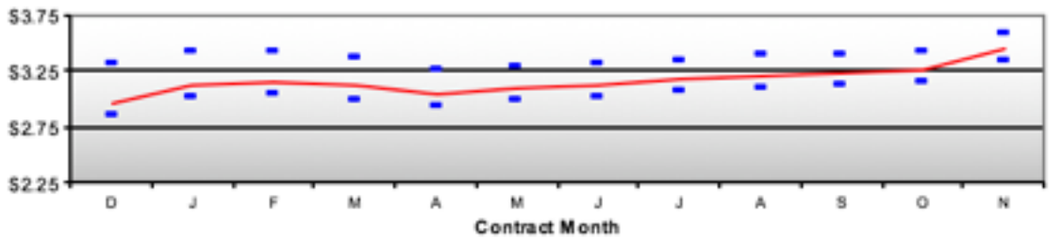
November 9, 2001

ENA - West Gas Contacts

Houston Office		Regional Offices		
Barry Tycholiz	(713) 853-1587	Mark Whitt	(303) 575-6473	Denver
Kim Ward	(713) 853-0685	Paul Lucci	(303) 575-6474	Denver
Stephanie Miller	(713) 853-1688	Tyrell Harrison	(303) 575-6478	Denver
Philip Polsky	(713) 853-5181	Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

	NYMEX	
	SETTLE	Δ
Cash		
ROM		
Dec-01	2.960	0.090
Dec-01 to Mar-02	3.088	0.083
Apr-02 to Oct-02	3.166	0.084
Nov-02 to Mar-03	3.651	0.090
One Year Strip*	3.165	0.084



IF NWPL Rocky Mountains			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
1.890	1.910		
2.060	2.080		
2.395	2.415	(0.565)	(0.545)
2.594	2.614	(0.494)	(0.474)
2.581	2.601	(0.585)	(0.565)
3.356	3.376	(0.295)	(0.275)
2.634	2.654	(0.530)	(0.510)

IF CIG Rocky Mountains				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	1.940	1.960		
ROM	1.960	1.980		
Dec-01	2.345	2.365	(0.615)	(0.595)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)
One Year Strip*	2.551	2.571	(0.614)	(0.594)

IF EL Paso Permian				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.375	2.395		
ROM	2.420	2.440		
Dec-01	2.700	2.720	(0.260)	(0.240)
Dec-01 to Mar-02	2.855	2.875	(0.233)	(0.213)
Apr-02 to Oct-02	3.009	3.029	(0.158)	(0.138)
Nov-02 to Mar-03	3.499	3.519	(0.153)	(0.133)
One Year Strip*	2.982	3.002	(0.182)	(0.162)

IF EL Paso San Juan				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.450	2.470		
ROM	2.350	2.370		
Dec-01	2.560	2.580	(0.400)	(0.380)
Dec-01 to Mar-02	2.743	2.763	(0.345)	(0.325)
Apr-02 to Oct-02	2.801	2.821	(0.365)	(0.345)
Nov-02 to Mar-03	3.421	3.441	(0.230)	(0.210)
One Year Strip*	2.817	2.837	(0.347)	(0.327)

AECO / NIT				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.376	2.396		
ROM	2.398	2.418		
Dec-01	2.552	2.572	(0.408)	(0.388)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)
One Year Strip*	2.676	2.696	(0.488)	(0.468)

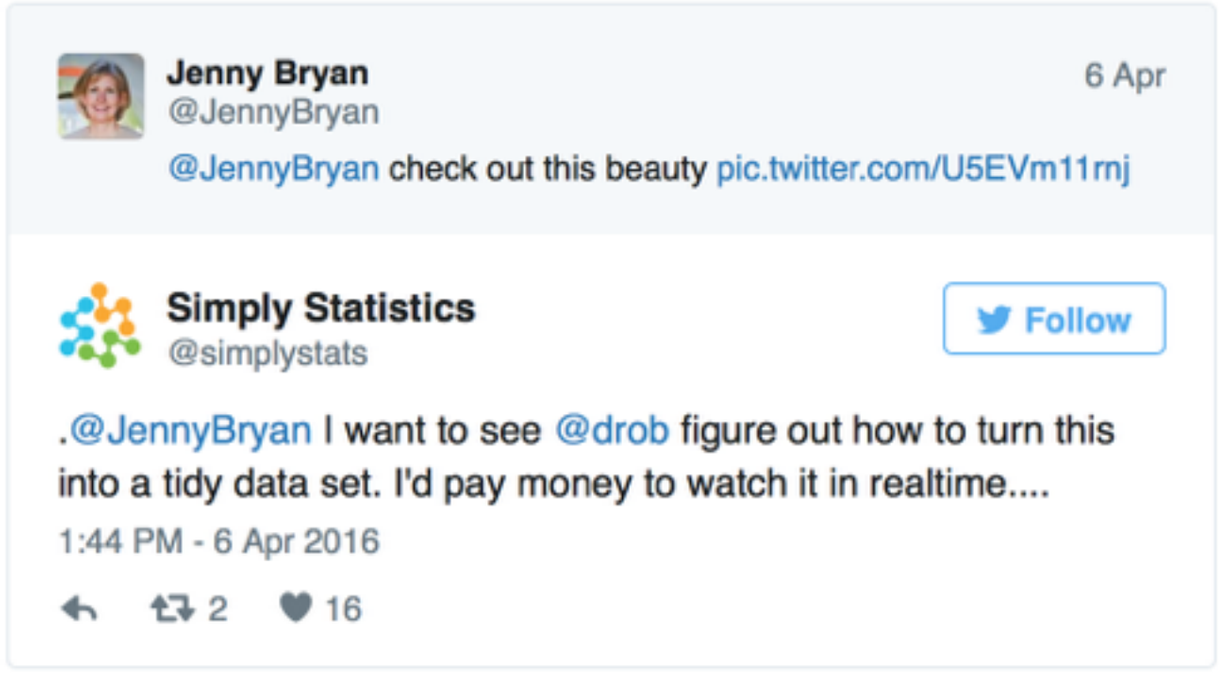
IF NWPL Canadian Border (Sumas)				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.480	2.500		
ROM	2.460	2.480		
Dec-01	2.800	2.820	(0.160)	(0.140)
Dec-01 to Mar-02	2.892	2.912	(0.196)	(0.176)
Apr-02 to Oct-02	2.796	2.816	(0.370)	(0.350)
Nov-02 to Mar-03	3.706	3.726	0.055	0.075
One Year Strip*	2.880	2.900	(0.285)	(0.265)

IF PEPL TX-OK				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.530	2.550		
ROM	2.530	2.550		
Dec-01	2.828	2.848	(0.133)	(0.113)
Dec-01 to Mar-02	2.958	2.978	(0.130)	(0.110)
Apr-02 to Oct-02	3.046	3.066	(0.120)	(0.100)
Nov-02 to Mar-03	3.531	3.551	(0.120)	(0.100)
One Year Strip*	3.041	3.061	(0.123)	(0.103)

Tidying an untidyable dataset

David Robinson

April 7, 2016



Jenny Bryan 6 Apr
@JennyBryan

@JennyBryan check out this beauty pic.twitter.com/U5EVm11rnj

Simply Statistics Follow
@simplystats

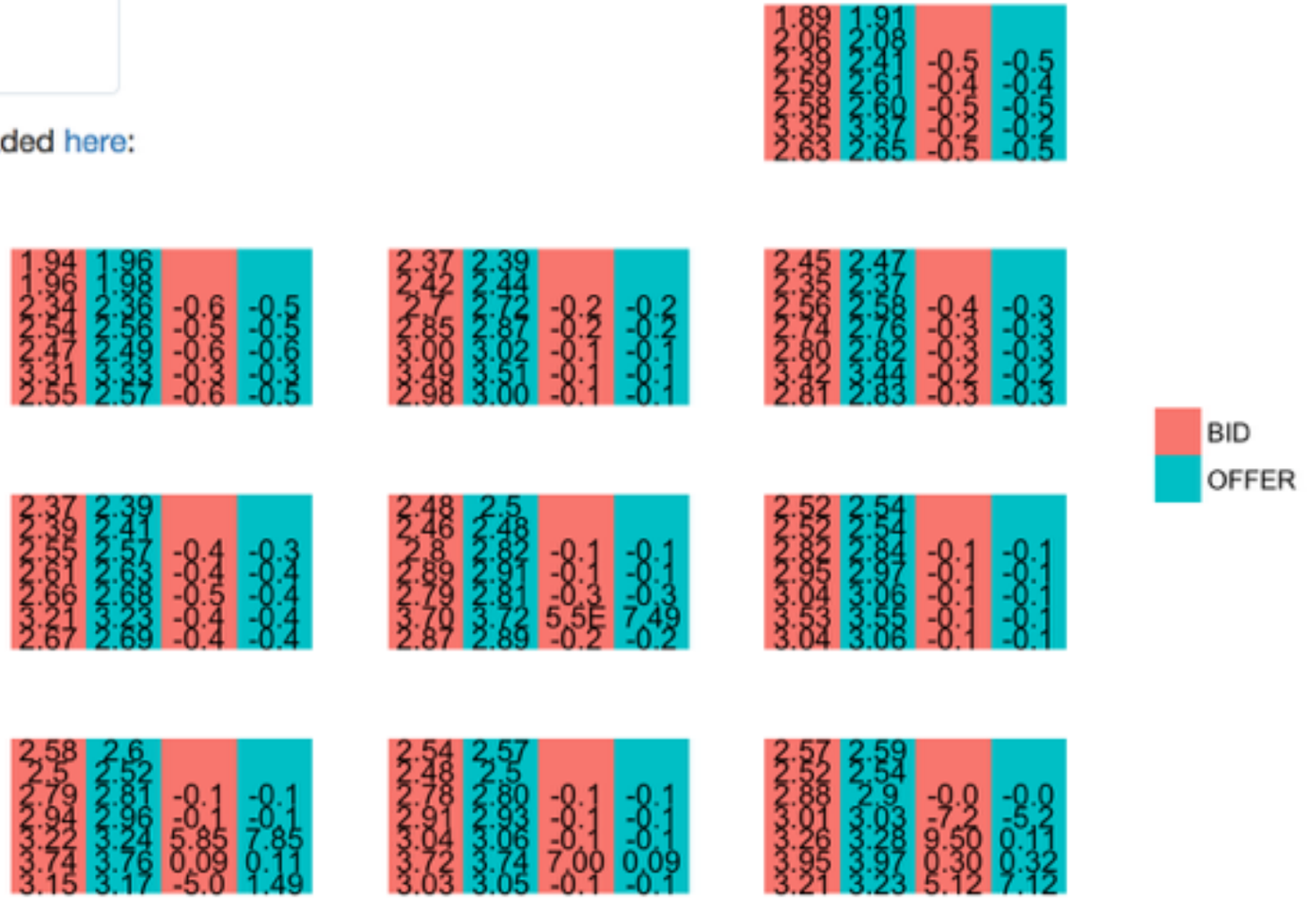
.@JennyBryan I want to see @drob figure out how to turn this into a tidy data set. I'd pay money to watch it in realtime....

1:44 PM - 6 Apr 2016

2 16

Challenge accepted. Here's the data Jenny sent, which can be downloaded here:

~150 lines of code later …



readxl: CRAN, GitHub
openxlsx: CRAN, GitHub

XLConnect: CRAN, GitHub
xlsx: CRAN, GitHub

gdata: CRAN, R-Forge
… and more

What do we want?

no tricky dependency … no Java
agnostic re: Excel, Google Sheet, ill-formed csv

expose
- (unformatted) data
- (unevaluated) formulas
- formatting

detect / propose views
handle merged cells, weird headers

How are we doing it?

define the **linen** object = spreadsheet receptacle
- document meta-data
- worksheet meta-data
- cell data, broadly defined

rexcel & googlesheets create linen objects
- simple? return a data frame!
- not? expose linen object for more processing …

Sheets API v3
(XML)

Google Apps
Script / Sheets
API v4 (JSON)

(XML)

linen

workbook
worksheets
cells

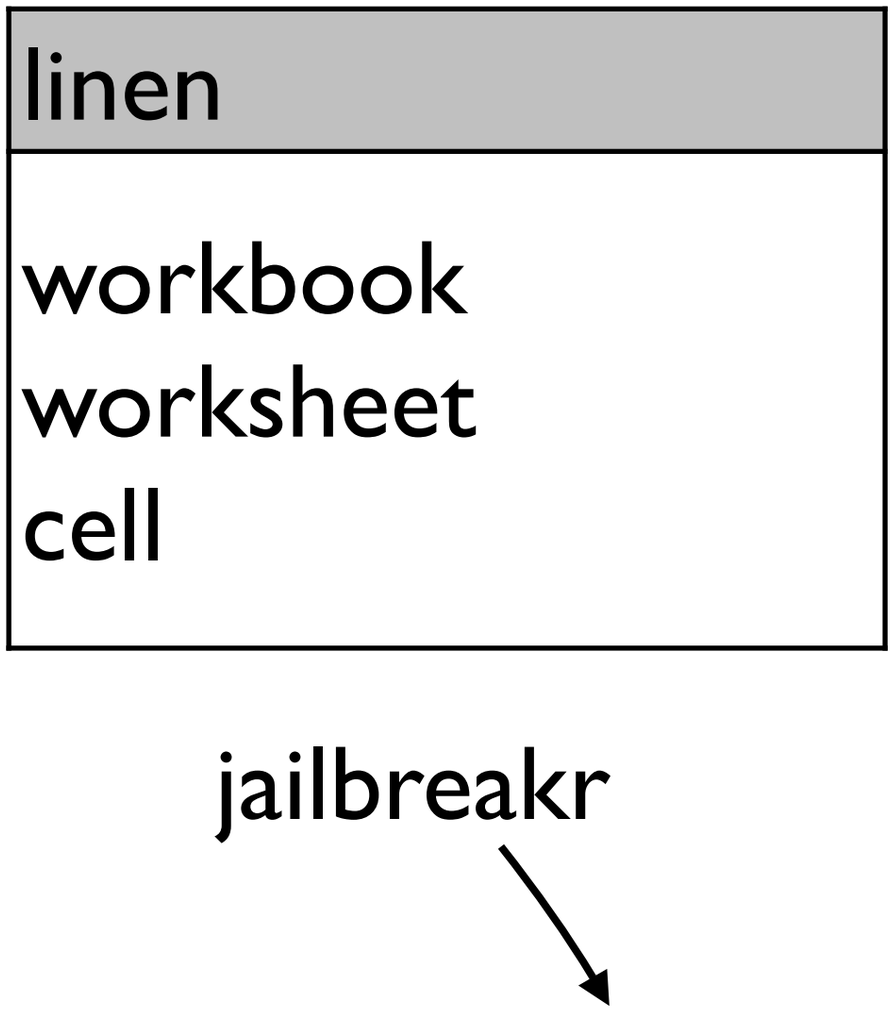
googlesheets

rexcel

data frame

data frame

googlesheets



rexcel

jailbreakr

multiple views, data frames
unformatted data, formatting
unevaluated formulas
figures?

Enron North America - West Gas

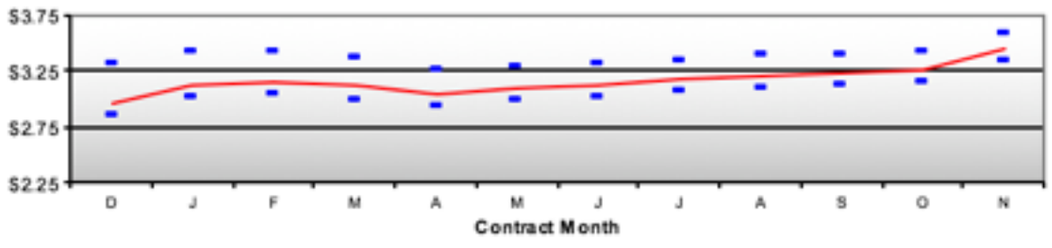
November 9, 2001

ENA - West Gas Contacts

Houston Office		Regional Offices		
Barry Tycholiz	(713) 853-1587	Mark Whitt	(303) 575-6473	Denver
Kim Ward	(713) 853-0685	Paul Lucci	(303) 575-6474	Denver
Stephanie Miller	(713) 853-1688	Tyrell Harrison	(303) 575-6478	Denver
Philip Polsky	(713) 853-5181	Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

	NYMEX	
	SETTLE	Δ
Cash		
ROM		
Dec-01	2.960	0.090
Dec-01 to Mar-02	3.088	0.083
Apr-02 to Oct-02	3.166	0.084
Nov-02 to Mar-03	3.651	0.090
One Year Strip*	3.165	0.084



IF NWPL Rocky Mountains			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
1.890	1.910		
2.060	2.080		
2.395	2.415	(0.565)	(0.545)
2.594	2.614	(0.494)	(0.474)
2.581	2.601	(0.585)	(0.565)
3.356	3.376	(0.295)	(0.275)
2.634	2.654	(0.530)	(0.510)

IF CIG Rocky Mountains				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	1.940	1.960		
ROM	1.960	1.980		
Dec-01	2.345	2.365	(0.615)	(0.595)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)
One Year Strip*	2.551	2.571	(0.614)	(0.594)

IF EL Paso Permian				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.375	2.395		
ROM	2.420	2.440		
Dec-01	2.700	2.720	(0.260)	(0.240)
Dec-01 to Mar-02	2.855	2.875	(0.233)	(0.213)
Apr-02 to Oct-02	3.009	3.029	(0.158)	(0.138)
Nov-02 to Mar-03	3.499	3.519	(0.153)	(0.133)
One Year Strip*	2.982	3.002	(0.182)	(0.162)

IF EL Paso San Juan				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.450	2.470		
ROM	2.350	2.370		
Dec-01	2.560	2.580	(0.400)	(0.380)
Dec-01 to Mar-02	2.743	2.763	(0.345)	(0.325)
Apr-02 to Oct-02	2.801	2.821	(0.365)	(0.345)
Nov-02 to Mar-03	3.421	3.441	(0.230)	(0.210)
One Year Strip*	2.817	2.837	(0.347)	(0.327)

AECO / NIT				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.376	2.396		
ROM	2.398	2.418		
Dec-01	2.552	2.572	(0.408)	(0.388)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)
One Year Strip*	2.676	2.696	(0.488)	(0.468)

IF NWPL Canadian Border (Sumas)				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.480	2.500		
ROM	2.460	2.480		
Dec-01	2.800	2.820	(0.160)	(0.140)
Dec-01 to Mar-02	2.892	2.912	(0.196)	(0.176)
Apr-02 to Oct-02	2.796	2.816	(0.370)	(0.350)
Nov-02 to Mar-03	3.706	3.726	0.055	0.075
One Year Strip*	2.880	2.900	(0.285)	(0.265)

IF PEPL TX-OK				
	Fixed Price		Basis	
	BID	OFFER	BID	OFFER
Cash	2.530	2.550		
ROM	2.530	2.550		
Dec-01	2.828	2.848	(0.133)	(0.113)
Dec-01 to Mar-02	2.958	2.978	(0.130)	(0.110)
Apr-02 to Oct-02	3.046	3.066	(0.120)	(0.100)
Nov-02 to Mar-03	3.531	3.551	(0.120)	(0.100)
One Year Strip*	3.041	3.061	(0.123)	(0.103)

Enron North America - West Gas

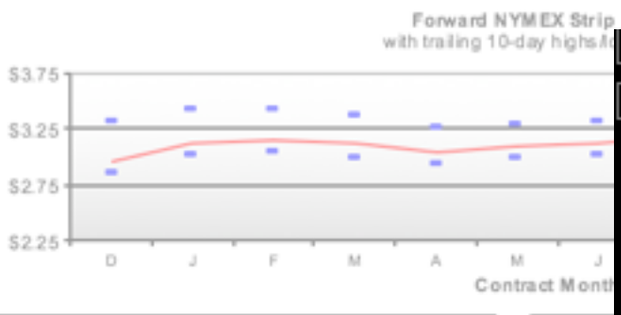
November 9, 2001

ENA - West Gas Contacts

Houston Office		Regional Offices		
Barry Tycholiz	(713) 853-1587	Mark Whitt	(303) 575-6473	Denver
Kim Ward	(713) 853-0685	Paul Lucci	(303) 575-6474	Denver
Stephanie Miller	(713) 853-1688	Tyrell Harrison	(303) 575-6478	Denver
Philip Polsky	(713) 853-5181	Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

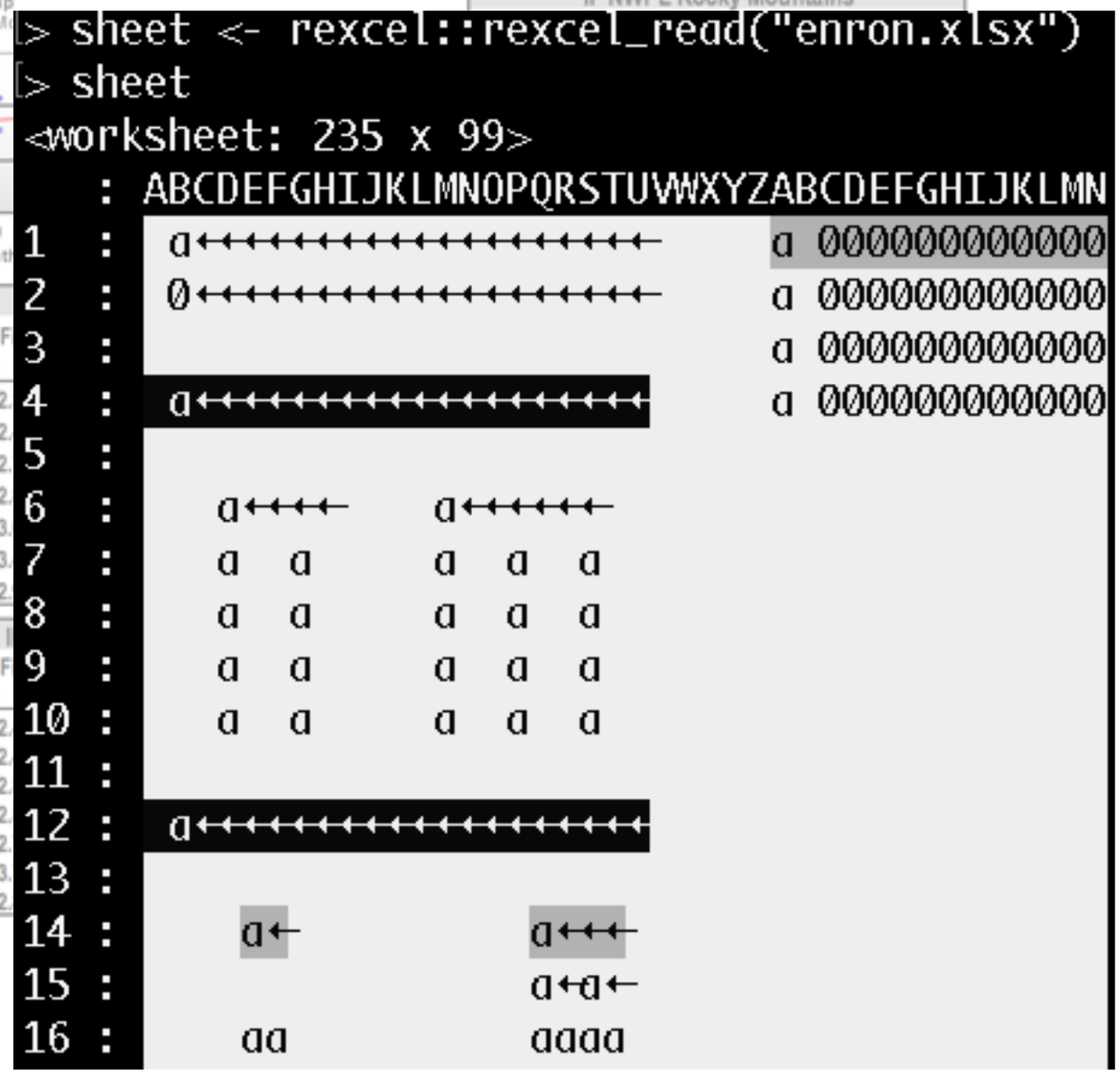
	NYMEX SETTLE	NYMEX Δ
Cash		
ROM		
Dec-01	2.960	0.090
Dec-01 to Mar-02	3.088	0.083
Apr-02 to Oct-02	3.166	0.084
Nov-02 to Mar-03	3.651	0.090
One Year Strip*	3.165	0.084



IF NWPL Rocky Mountains

	IF CIG Rocky Mountains Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	1.940	1.960		
ROM	1.960	1.980		
Dec-01	2.345	2.365	(0.615)	(0.595)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)
One Year Strip*	2.551	2.571	(0.614)	(0.594)

	AECO / NIT Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	2.376	2.396		
ROM	2.398	2.418		
Dec-01	2.552	2.572	(0.408)	(0.388)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)
One Year Strip*	2.676	2.696	(0.488)	(0.468)



	AC	AD	AE	AF	AG
115					
116			**Congratulations to Dan Foulston**		
117			**of Canadian Western Natural Gas!!!**		
118			**He is the winner of our weekend getaway**		
119			**to Lake Louise. Dan's guess of $1.375 C/GJ**		
120			**was the closest to the average Aeco**		
121			**day gas price for the first**		
122			**15 days of December.**		

```
113:
114:
115:
116:                                    a
117:                                    a
118:                                    a
119:                                    a
120:                                    a
121:                                    a
122:                                    a
```

Enron North America - West Gas

November 9, 2001

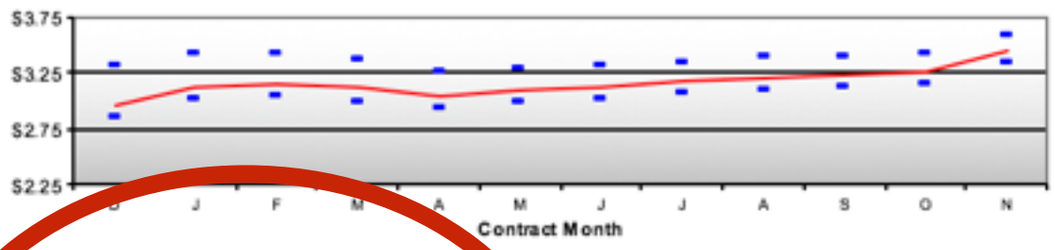
ENA - West Gas Contacts

Houston Office		Regional Offices		
Barry Tycholiz	(713) 853-1587	Mark Whitt	(303) 575-6473	Denver
Kim Ward	(713) 853-0685	Paul Lucci	(303) 575-6474	Denver
Stephanie Miller	(713) 853-1688	Tyrell Harrison	(303) 575-6478	Denver
Philip Polsky	(713) 853-5181	Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

	NYMEX	
	SETTLE	Δ
Cash		
ROM		
Dec-01	2.960	0.090
Dec-01 to Mar-02	3.088	0.083
Apr-02 to Oct-02	3.166	0.084
Nov-02 to Mar-03	3.651	0.090
One Year Strip*	3.165	0.084

Forward NYMEX Strip
with trailing 10-day highs/lows



IF NWPL Rocky Mountains			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
1.890	1.910		
2.060	2.080		
2.395	2.415	(0.565)	(0.545)
2.594	2.614	(0.494)	(0.474)
2.581	2.601	(0.585)	(0.565)
3.356	3.376	(0.295)	(0.275)
2.634	2.654	(0.530)	(0.510)

IF CIG Rocky Mountains			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
1.940	1.960		
1.960	1.980		
2.345	2.365	(0.615)	(0.595)
2.548	2.568	(0.540)	(0.520)
2.471	2.491	(0.695)	(0.675)
3.311	3.331	(0.340)	(0.320)
2.551	2.571	(0.614)	(0.59

IF EL Paso Permian			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
2.375	2.395		
2.420	2.440		
2.700	2.720	(0.260)	(0.240)
2.855	2.875	(0.233)	(0.213)
3.009	3.029	(0.158)	(0.138)
3.499	3.519	(0.153)	(0.133)
2.982	3.002	(0.182)	(0.162)

IF EL Paso San Juan			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
2.450	2.470		
2.350	2.370		
2.560	2.580	(0.400)	(0.380)
2.743	2.763	(0.345)	(0.325)
2.801	2.821	(0.365)	(0.345)
3.421	3.441	(0.230)	(0.210)
2.817	2.837	(0.347)	(0.327)

AECO / NIT			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
2.376	2.396		
2.398	2.418		
2.552	2.572	(0.408)	(0.388)
2.616	2.636	(0.472)	(0.452)
2.661	2.681	(0.505)	(0.485)
3.216	3.236	(0.435)	(0.415)
2.676	2.696	(0.488)	(0.468)

IF NWPL Canadian Border (Sumas)			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
2.480	2.500		
2.460	2.480		
2.800	2.820	(0.160)	(0.140)
2.892	2.912	(0.196)	(0.176)
2.796	2.816	(0.370)	(0.350)
3.706	3.726	0.055	0.075
2.880	2.900	(0.285)	(0.265)

IF PEPL TX-OK			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
2.530	2.550		
2.530	2.550		
2.828	2.848	(0.133)	(0.113)
2.958	2.978	(0.130)	(0.110)
3.046	3.066	(0.120)	(0.100)
3.531	3.551	(0.120)	(0.100)
3.041	3.061	(0.123)	(0.103)

Enron North America - West Gas

November 9, 2001

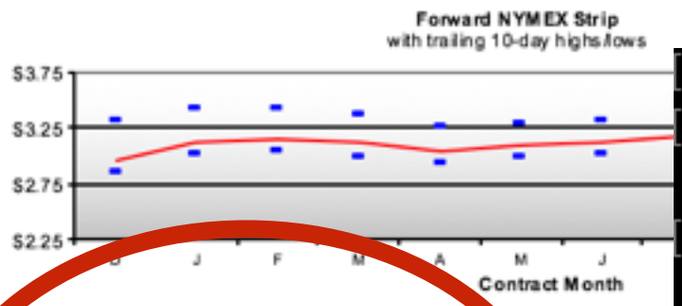
ENA - West Gas Contacts

Houston Office	
Barry Tycholiz	(713) 853-1587
Kim Ward	(713) 853-0685
Stephanie Miller	(713) 853-1688
Philip Polsky	(713) 853-5181

Regional Offices		
Mark Whitt	(303) 575-6473	Denver
Paul Lucci	(303) 575-6474	Denver
Tyrell Harrison	(303) 575-6478	Denver
Dave Fuller	(503) 464-3732	Portland

Forward Prices (US$/MMBtu)

	NYMEX SETTLE	NYMEX Δ
Cash		
ROM		
Dec-01	2.960	0.090
Dec-01 to Mar-02	3.088	0.083
Apr-02 to Oct-02	3.166	0.084
Nov-02 to Mar-03	3.651	0.090
One Year Strip*	3.165	0.084



IF CIG Rocky Mountains	Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	1.940	1.960		
ROM	1.960	1.980		
Dec-01	2.345	2.365	(0.615)	(0.595)
Dec-01 to Mar-02	2.548	2.568	(0.540)	(0.520)
Apr-02 to Oct-02	2.471	2.491	(0.695)	(0.675)
Nov-02 to Mar-03	3.311	3.331	(0.340)	(0.320)
One Year Strip*	2.551	2.571	(0.614)	(0.59

AECO / NIT	Fixed Price BID	Fixed Price OFFER	Basis BID	Basis OFFER
Cash	2.376	2.396		
ROM	2.398	2.418		
Dec-01	2.552	2.572	(0.408)	(0.388)
Dec-01 to Mar-02	2.616	2.636	(0.472)	(0.452)
Apr-02 to Oct-02	2.661	2.681	(0.505)	(0.485)
Nov-02 to Mar-03	3.216	3.236	(0.435)	(0.415)
One Year Strip*	2.676	2.696	(0.488)	(0.468)

IF NWPL Rocky Mountains	Fixed Price BID	Basis
Cash	2.375	
ROM	2.420	
Dec-01	2.700	
Dec-01 to Mar-02	2.855	
Apr-02 to Oct-02	3.009	
Nov-02 to Mar-03	3.499	
One Year Strip*	2.982	

IF NW	Fixed P BID
Cash	2.480
ROM	2.460
Dec-01	2.800
Dec-01 to Mar-02	2.892
Apr-02 to Oct-02	2.796
Nov-02 to Mar-03	3.706
One Year Strip*	2.880

```
> views <- jailbreakr::split_sheet(sheet)
> length(views)
[1] 40
> head(views, 3)
[[1]]
<worksheet_view: 2 x 20 (of 235 x 99)>
  : BCDEFGHIJKLMNOPQRSTU
1: a←←←←←←←←←←←←←←←←←←←
2: 0←←←←←←←←←←←←←←←←←←←

[[2]]
<worksheet_view: 1 x 20 (of 235 x 99)>
  : BCDEFGHIJKLMNOPQRSTU
4: a←←←←←←←←←←←←←←←←←←←

[[3]]
<worksheet_view: 1 x 20 (of 235 x 99)>
  : BCDEFGHIJKLMNOPQRSTU
12: a←←←←←←←←←←←←←←←←←←←
```

IF CIG Rocky Mountains			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
1.940	1.960		
1.960	1.980		
2.345	2.365	(0.615)	(0.595)
2.548	2.568	(0.540)	(0.520)
2.471	2.491	(0.695)	(0.675)
3.311	3.331	(0.340)	(0.320)
2.551	2.571	(0.614)	(0.594)

```
> x2 <- jailbreakr::split_metadata(x1, FALSE, min_data_block=1)
> x3 <- jailbreakr::split_headers_apply(x2, 2)
> x3
<worksheet_view: 7 x 4 (of 235 x 99)>
(with headers)
(with data: metadata)
  : GHIJ
28: 00
29: 00
30: 0000
31: 0000
32: 0000
33: 0000
34: 0000
```

IF CIG Rocky Mountains			
Fixed Price		Basis	
BID	OFFER	BID	OFFER
1.940	1.960		
1.960	1.980		
2.345	2.365	(0.615)	(0.595)
2.548	2.568	(0.540)	(0.520)
2.471	2.491	(0.695)	(0.675)
3.311	3.331	(0.340)	(0.320)
2.551	2.571	(0.614)	(0.594)

```
> x3$values()
     Fixed Price:BID Fixed Price:OFFER Basis:BID Basis:OFFER
[1,] 1.94            1.96              NA        NA
[2,] 1.96            1.98              NA        NA
[3,] 2.345           2.365             -0.615    -0.595
[4,] 2.54775         2.56775           -0.54     -0.52
[5,] 2.471286        2.491286          -0.695    -0.675
[6,] 3.311           3.331             -0.34     -0.32
[7,] 2.55075         2.57075           -0.61375  -0.59375
```

rsheets https://github.com/rsheets

Repositories | People 2 | Teams 0 | Settings

Filters ▾ | Find a repository… | New repository

rexcel

R ★6 1

Extracts spreadsheet data fromExcel workbooks and puts into linen format

Updated 8 days ago

enron_corpus_linen

R ★0 0

Code to process the enron corpus

Updated 9 days ago

cellranger

R ★12 3

Helper functions to work with spreadsheets and the "A1:D10" style of cell range specification

Updated on May 25

linen

R ★5 0

General representation of spreadsheet data, plus some limited low-level operations on that data

Updated on May 13

README

★0 0

Updated on May 3

jailbreakr

R ★3 0

Get out of Excel free.

Updated on May 3

People 2 >

jennybc
Jennifer (Jenny) Bryan

richfitz
Rich FitzJohn

Invite someone

googlesheets — linen / cellranger — rexcel

raw object — data frame

jailbreakr

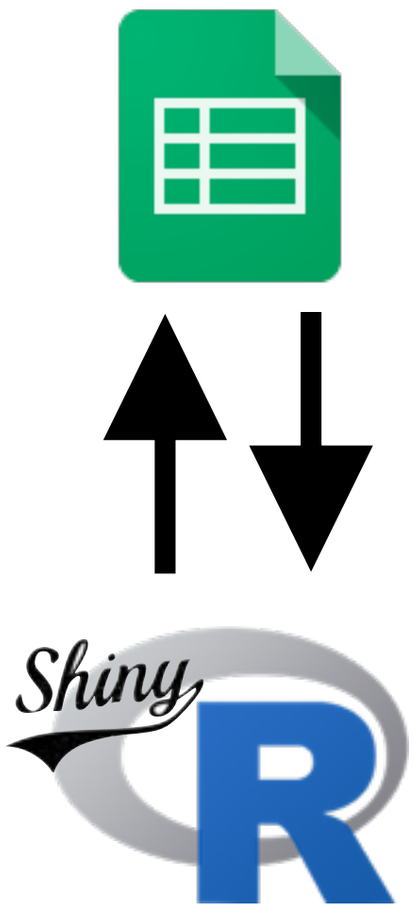
raw object — data frame

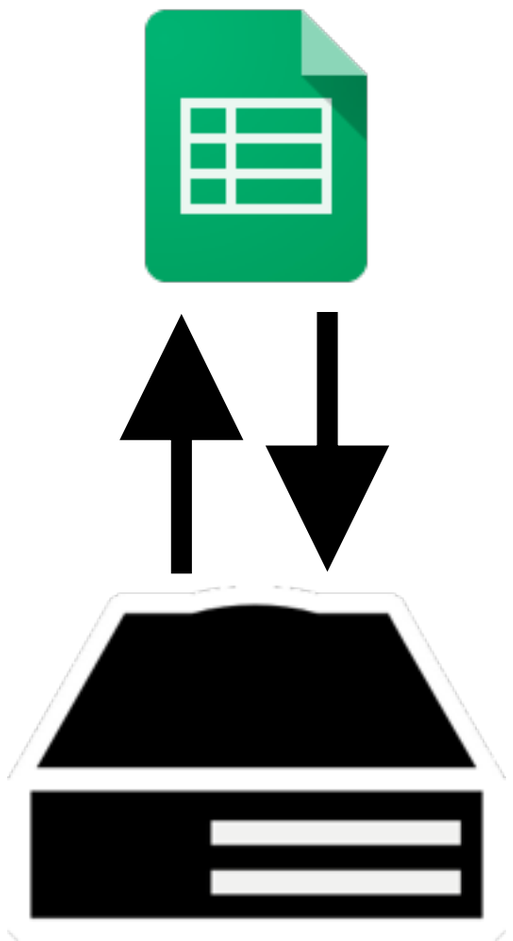
multiple data frames

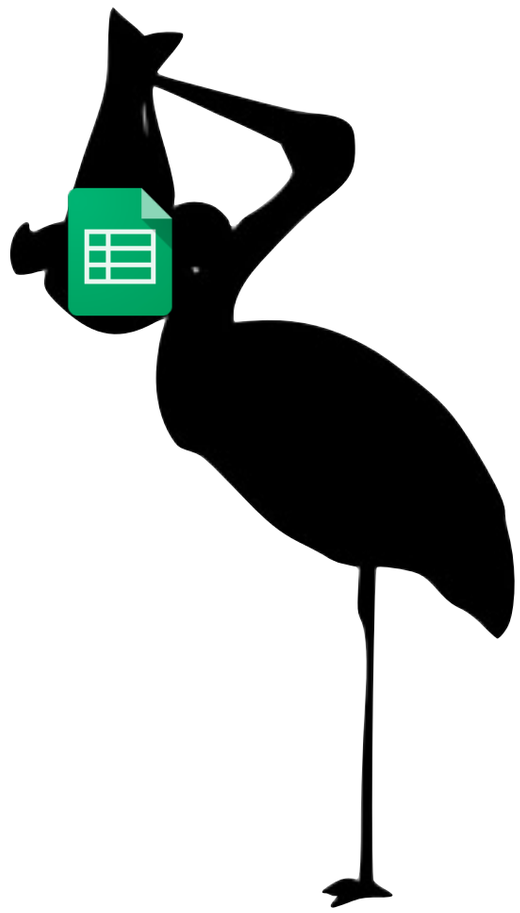
formulas, formatting, figures?

bonus content

googlesheets







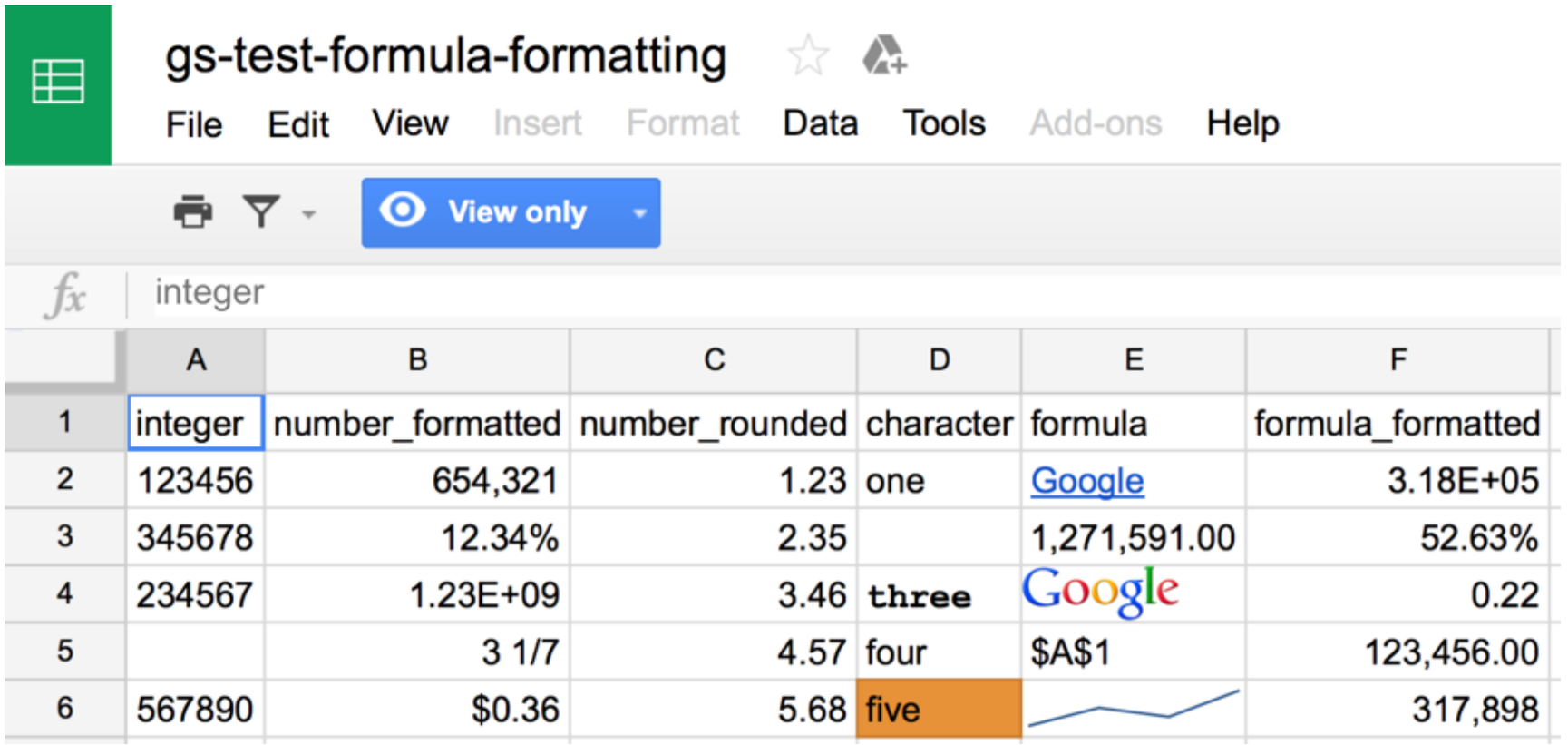
gs-test-formula-formatting

File Edit View Insert Format Data Tools Add-ons Help

View only

integer

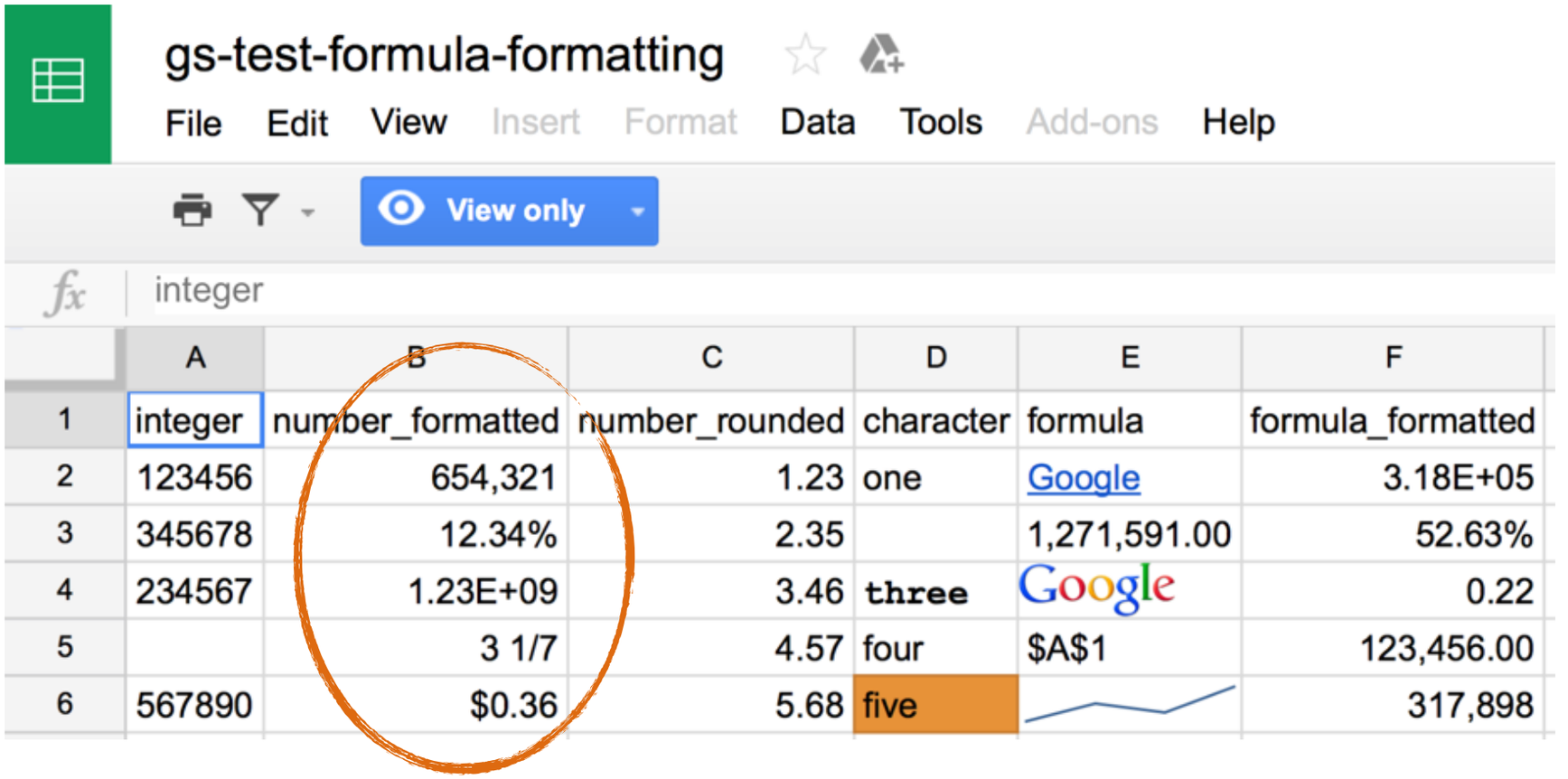
	A	B	C	D	E	F
1	integer	number_formatted	number_rounded	character	formula	formula_formatted
2	123456	654,321	1.23	one	Google	3.18E+05
3	345678	12.34%	2.35		1,271,591.00	52.63%
4	234567	1.23E+09	3.46	three	Google	0.22
5		3 1/7	4.57	four	A1	123,456.00
6	567890	$0.36	5.68	five		317,898



	A	B	C	D	E	F
1	integer	number_formatted	number_rounded	character	formula	formula_formatted
2	123456	654,321	1.23	one	Google	3.18E+05
3	345678	12.34%	2.35		1,271,591.00	52.63%
4	234567	1.23E+09	3.46	three	Google	0.22
5		3 1/7	4.57	four	A1	123,456.00
6	567890	$0.36	5.68	five		317,898

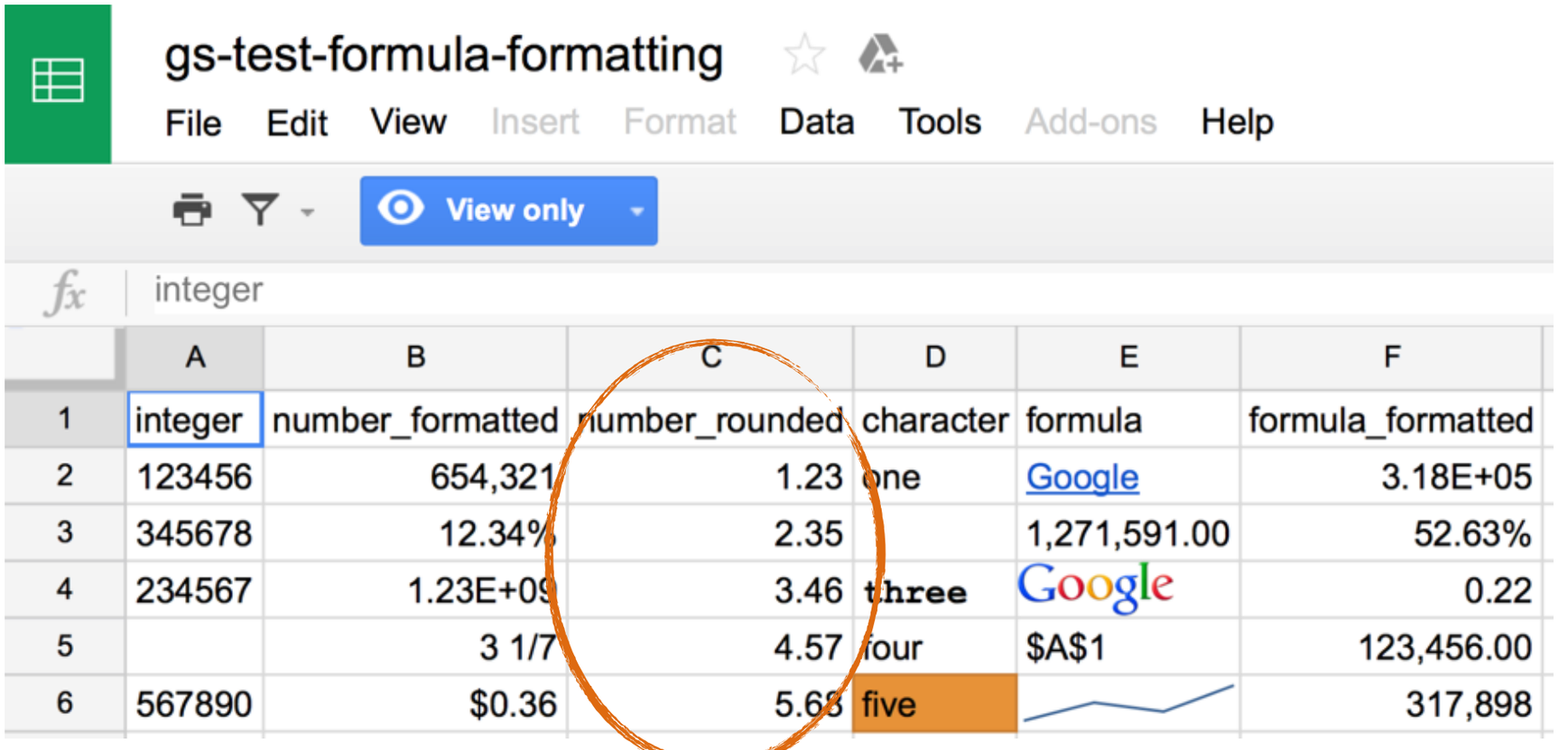
default read does not necessarily give you what you want with numeric formatting and formulas

```
gs_ff() %>%
  gs_read() %>%
  select(-integer)
#> Accessing worksheet titled 'Sheet1'.
#> No encoding supplied: defaulting to UTF-8.
#> Source: local data frame [5 x 5]
#>
#>   number_formatted number_rounded character     formula formula_formatted
#>              (chr)          (dbl)     (chr)       (chr)             (chr)
#> 1          654,321           1.23       one      Google          3.18E+05
#> 2           12.34%           2.35        NA 1,271,591.00           52.63%
#> 3         1.23E+09           3.46     three          NA              0.22
#> 4            3 1/7           4.57      four       $A$1        123,456.00
#> 5           $0.36           5.68      five          NA           317,898
```



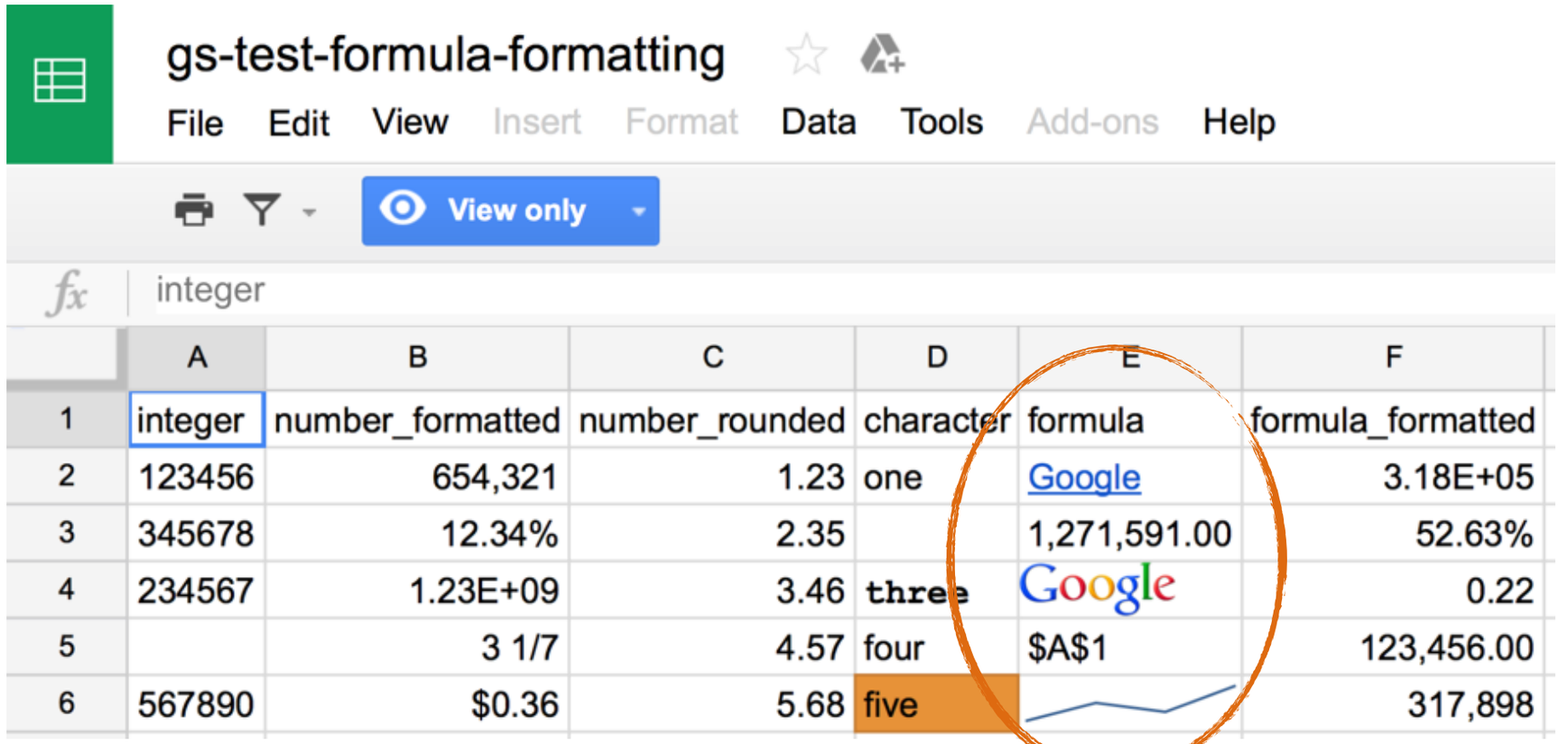
	A	B	C	D	E	F
1	integer	number_formatted	number_rounded	character	formula	formula_formatted
2	123456	654,321	1.23	one	Google	3.18E+05
3	345678	12.34%	2.35		1,271,591.00	52.63%
4	234567	1.23E+09	3.46	three	Google	0.22
5		3 1/7	4.57	four	A1	123,456.00
6	567890	$0.36	5.68	five		317,898

```
cf <- gs_read_cellfeed(gs_ff())
cf %>%
  filter(row > 1, col == 2) %>%
  select(value, input_value, numeric_value) %>%
  readr::type_convert()
#> <tibble [5 x 3]>
#>      value  input_value numeric_value
#>      <chr>        <dbl>         <dbl>
#> 1  654,321 6.543210e+05  6.543210e+05
#> 2   12.34% 1.234000e+01  1.234000e-01
#> 3 1.23E+09 1.234568e+09  1.234568e+09
#> 4    3 1/7 3.141593e+00  3.141593e+00
#> 5    $0.36 3.600000e-01  3.600000e-01
```



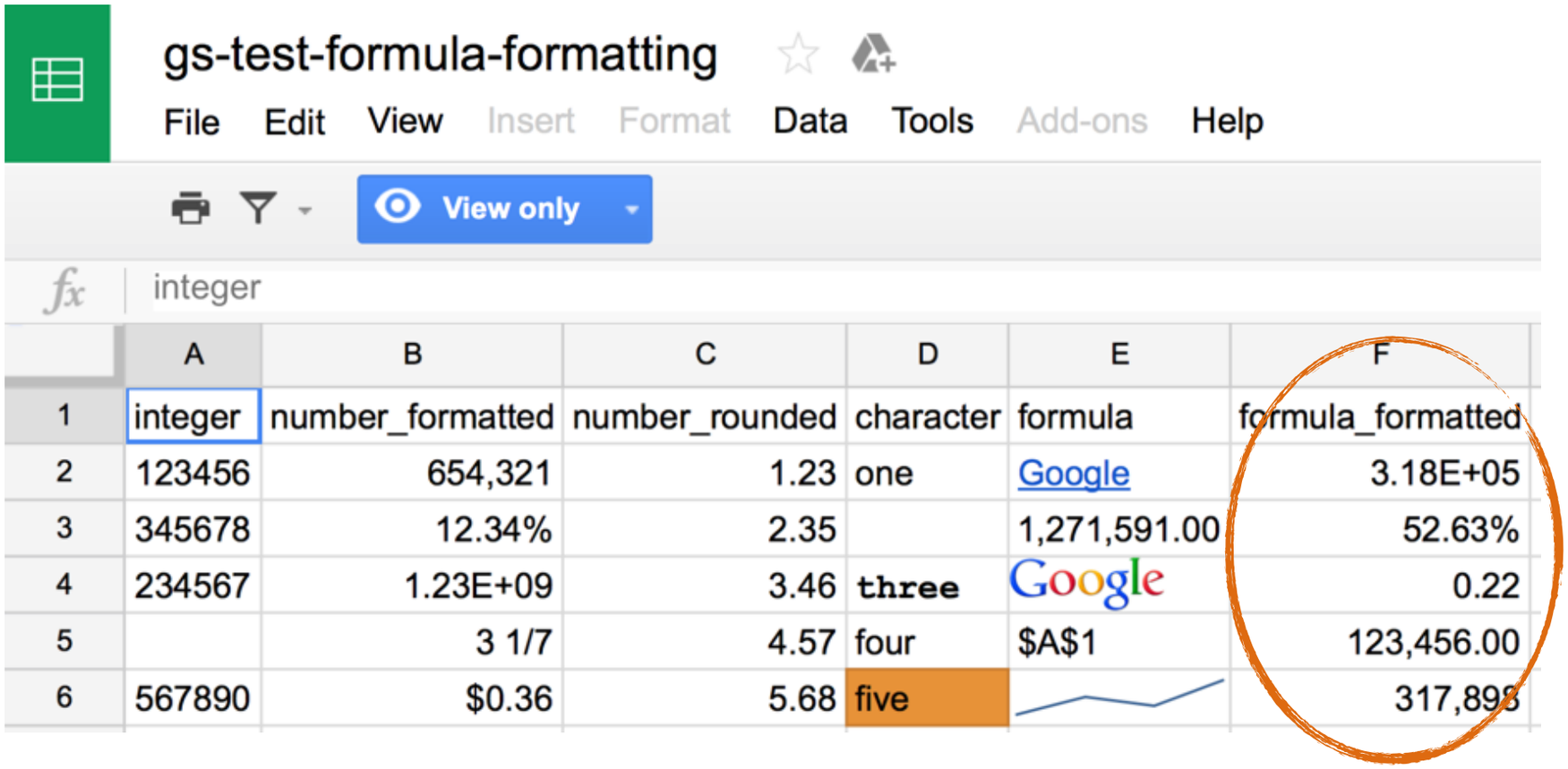
	A	B	C	D	E	F
1	integer	number_formatted	number_rounded	character	formula	formula_formatted
2	123456	654,321	1.23	one	Google	3.18E+05
3	345678	12.34%	2.35		1,271,591.00	52.63%
4	234567	1.23E+09	3.46	three	Google	0.22
5		3 1/7	4.57	four	A1	123,456.00
6	567890	$0.36	5.68	five		317,898

```
cf <- gs_read_cellfeed(gs_ff())
cf %>%
  filter(row > 1, col == 3) %>%
  select(value, input_value, numeric_value) %>%
  readr::type_convert()
#> <tibble [5 x 3]>
#>   value input_value numeric_value
#>   <dbl>       <dbl>         <dbl>
#> 1  1.23      1.2345        1.2345
#> 2  2.35      2.3456        2.3456
#> 3  3.46      3.4567        3.4567
#> 4  4.57      4.5678        4.5678
#> 5  5.68      5.6789        5.6789
```



	A	B	C	D	E	F
1	integer	number_formatted	number_rounded	character	formula	formula_formatted
2	123456	654,321	1.23	one	Google	3.18E+05
3	345678	12.34%	2.35		1,271,591.00	52.63%
4	234567	1.23E+09	3.46	three	Google	0.22
5		3 1/7	4.57	four	A1	123,456.00
6	567890	$0.36	5.68	five		317,898

```
cf <- gs_read_cellfeed(gs_ff())
cf %>%
  filter(row > 1, col == 5) %>%
  select(value, input_value, numeric_value) %>%
  mutate(input_value = substr(input_value, 1, 43)) %>%
  readr::type_convert()
#> <tibble [5 x 3]>
#>          value                                    input_value numeric_value
#>          <chr>                                          <chr>         <dbl>
#> 1       Google =HYPERLINK("http://www.google.com/","Google            NA
#> 2 1,271,591.00                    =sum(R[-1]C[-4]:R[3]C[-4])       1271591
#> 3          <NA> =IMAGE("https://www.google.com/images/srpr/            NA
#> 4          $A$1                                 =ADDRESS(1,1)            NA
#> 5          <NA>           =SPARKLINE(R[-4]C[-4]:R[0]C[-4])            NA
```



	A	B	C	D	E	F
1	integer	number_formatted	number_rounded	character	formula	formula_formatted
2	123456	654,321	1.23	one	Google	3.18E+05
3	345678	12.34%	2.35		1,271,591.00	52.63%
4	234567	1.23E+09	3.46	three	Google	0.22
5		3 1/7	4.57	four	A1	123,456.00
6	567890	$0.36	5.68	five		317,898

```
cf <- gs_read_cellfeed(gs_ff())
cf %>%
  filter(row > 1, col == 6) %>%
  select(value, input_value, numeric_value) %>%
  readr::type_convert()
#> <tibble [5 x 3]>
#>        value                    input_value numeric_value
#>        <chr>                          <chr>         <dbl>
#> 1   3.18E+05 =average(R[0]C[-5]:R[4]C[-5])  3.178978e+05
#> 2     52.63%         =R[-1]C[-5]/R[1]C[-5]  5.263144e-01
#> 3       0.22         =R[-2]C[-5]/R[2]C[-5]  2.173942e-01
#> 4 123,456.00    =min(R[-3]C[-5]:R[1]C[-5])  1.234560e+05
#> 5    317,898           =average(R2C1:R6C1)  3.178978e+05
```