

The art in data science

Matt Monihan, Product @ RJMetrics
@mattmonihan






Officially the first person to have ever worn a VR headset in an isolation tank. [#virtualreality](#)
[@HalcyonFloats](#)



Reply to Halcyon Floats



“The formula for Ancient Aliens is always, fact, loose fact, fabrication, in that order. And it’s always amazing.”

Dan Clifton Episode #7

MERCENARY

Data science is useful for two things

1. Creating an input to an automated process
2. Generating ideas that inspire change

Communicating an idea is a PR
challenge as much as it is a design
and technical challenge.

This is real life...



The Onion 
@TheOnion

 **Follow**

Company's Employees Spend Entire Day Touching Base
<http://onion.com/etvVY9>

12:00 PM - 21 Mar 2011



311



43

Why?

- 1. They lack a single source of truth**
- 2. They don't trust the numbers**
3. They don't understand how the information is meaningful

Solution

- Consolidate the data to one location
- Designate an owner
- Defend the truth with documentation

Our Stack



What does it mean to communicate with data “at scale?”

4 Aspects

1. Truth is owned/maintained
2. The analysis is transparent
3. The analysis is self-serve, but the conclusion is directed
4. If the analysis is recurring, it needs to be reliable.

Why?

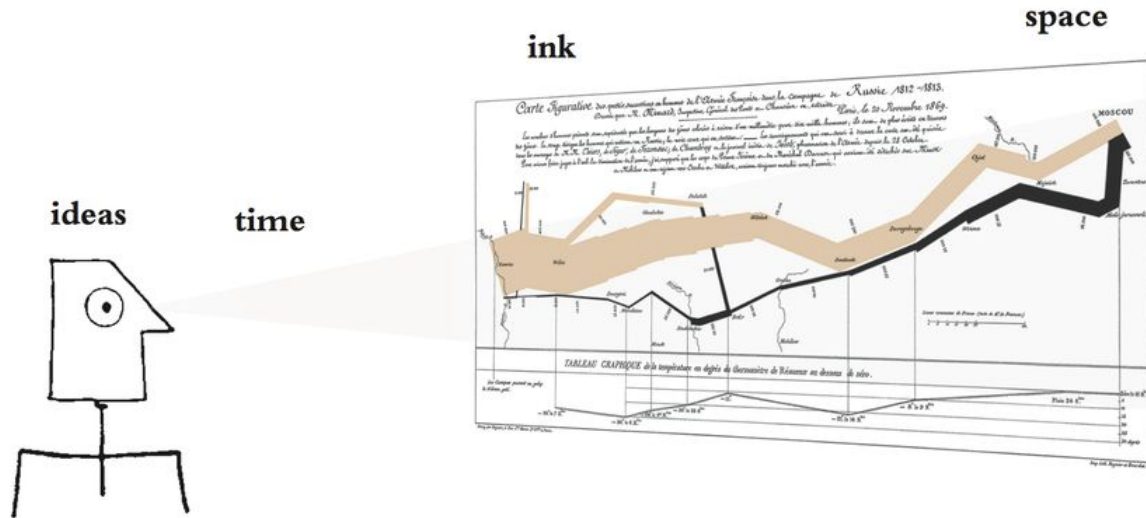
1. They lack a single source of truth
2. They don't trust the numbers
- 3. They don't understand how the information is meaningful**

Principles of Graphical Excellence

Graphical excellence is the well-designed presentation of interesting data—a matter of *substance*, of *statistics*, and of *design*.

Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency.

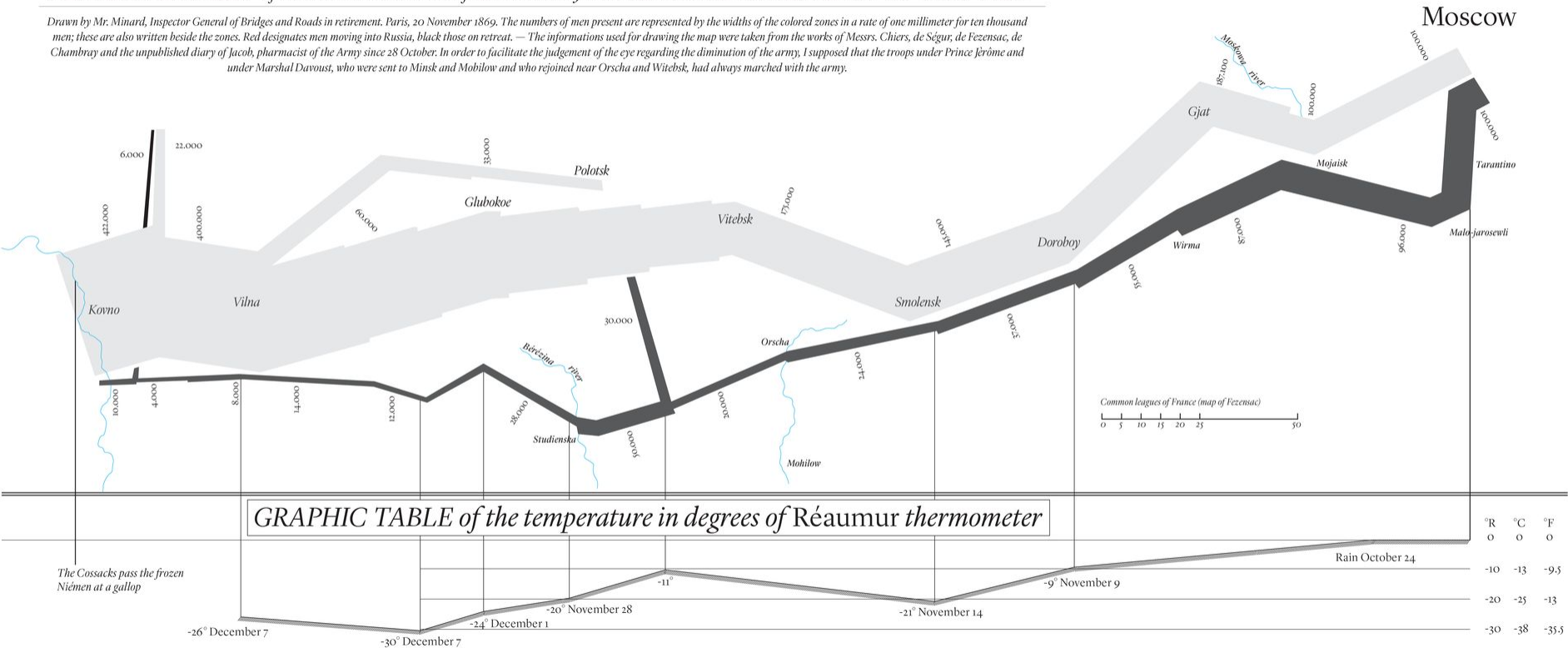
Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.

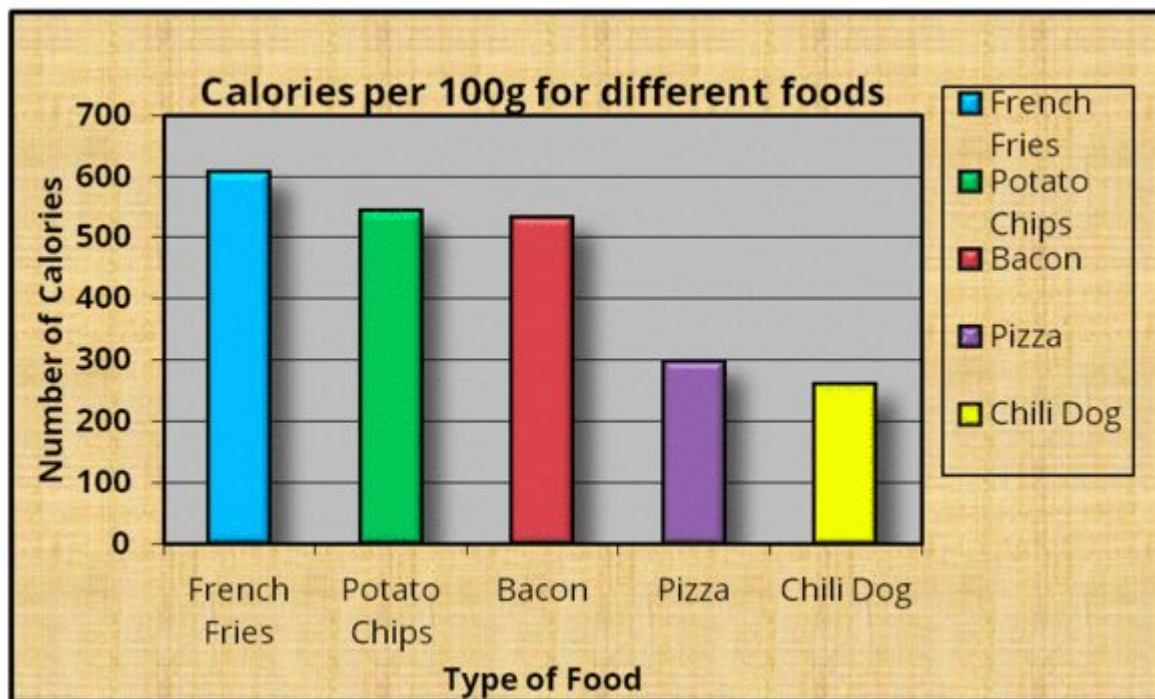


Increase your Data/Ink ratio

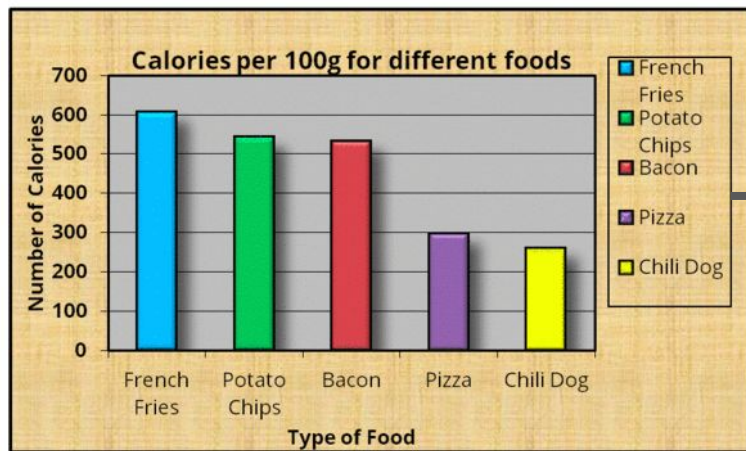
FIGURATIVE MAP of the successive losses in men of the French Army in the RUSSIAN CAMPAIGN OF 1812-1813

Drawn by Mr. Minard, Inspector General of Bridges and Roads in retirement, Paris, 20 November 1869. The numbers of men present are represented by the widths of the colored zones in a rate of one millimeter for ten thousand men; these are also written beside the zones. Red designates men moving into Russia, black those on retreat. — The informations used for drawing the map were taken from the works of Messrs. Chiers, de Ségur, de Fenezac, de Chambray and the unpublished diary of Jacob, pharmacist of the Army since 28 October. In order to facilitate the judgement of the eye regarding the diminution of the army, I supposed that the troops under Prince Jérôme and under Marshal Davoust, who were sent to Minsk and Mobilow and who rejoined near Orscha and Witebsk, had always marched with the army.

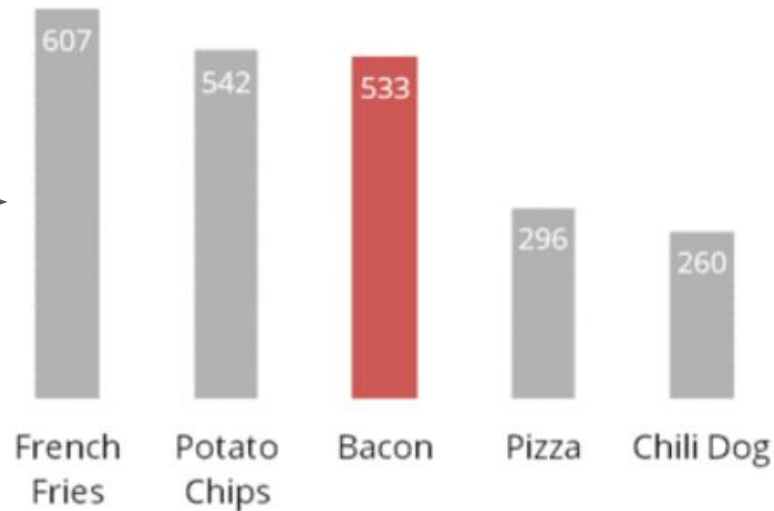




Remove
to improve
(the **data-ink** ratio)



Calories per 100g



Role	Name	Year of the...	Debut	Number of Fans	Takedown Rate
Face (The Hero)	The Ultimate Warrior	Tiger	May-2011	97320.00	86.2
Face (The Hero)	Hulk Hogan	Oxen	Jan-2008	988551.00	61.978
Face (The Hero)	Macho Man Randy Savage	Monkey	Feb-2008	157618.00	59.29
Face (The Hero)	Hacksaw Jim Duggan	Pig	Mar-2008	30300.00	53.4332
Face (The Hero)	Superfly Jimmy Snuka	Dragon	Mar-2008	12341.00	52.7
Heel (The Bad Guy)	Rowdy Roddy Piper	Rooster	Jun-1968	71645.00	45.4
Heel (The Bad Guy)	The Million Dollar Man Ted DiBiase	Rat	Apr-1975	449342.00	43.7689
Heel (The Bad Guy)	Mr. Perfect Curt Henning	Rat	May-1980	13773.00	38
Heel (The Bad Guy)	Jake the Snake Roberts	Snake	Jul-1975	5609.00	37.99
Jobber (The Unknown)	Brad Smith	Sheep	Aug-2008	1103.00	36.316
Jobber (The Unknown)	Ted Duncan	Sheep	Aug-2008	200.00	33.61
Jobber (The Unknown)	Joey the Uber Nerd Cherdarchuk	Snake	Aug-2008	5.00	21.0196

Remove
to improve
the **data tables** edition

Role	Name	Year of the...	Debut	Thousands of Fans	Takedown Rate
Face (The Hero)	The Ultimate Warrior	Tiger	May-2011	97.3	86.2
	Hulk Hogan	Oxen	Jan-2008	988.6	62.0
	Macho Man Randy Savage	Monkey	Feb-2008	157.6	59.3
	Hacksaw Jim Duggan	Pig	Mar-2008	30.3	53.4
	Superfly Jimmy Snuka	Dragon	Mar-2008	12.3	52.7
Heel (The Bad Guy)	Rowdy Roddy Piper	Rooster	Jun-1968	71.6	45.4
	The Million Dollar Man Ted DiBiase	Rat	Apr-1975	449.3	43.8
	Mr. Perfect Curt Henning	Rat	May-1980	13.8	38.0
	Jake the Snake Roberts	Snake	Jul-1975	5.6	38.0
Jobber (The Unknown)	Brad Smith	Sheep	Aug-2008	1.1	36.3
	Ted Duncan	Sheep	Aug-2008	0.2	33.6
	Joey the Uber Nerd Cherdarchuk	Snake	Aug-2008	0.0	21.0

New York, NY ★ 🏠

⌂ midtown | Report | Change Station ▼

Forecast History Calendar Rain / Snow Health **NEW!**

Elev 80 ft 40.75 °N, 73.99 °W | Updated 11 sec ago



Overcast

50.5 °F

Feels Like 50.5 °F



Wind Variable
Gusts 2.5 mph

Today is forecast to be **NEARLY THE SAME** temperature as yesterday.

Today

High 48 | Low 37 °F

☁ 0% Chance of Precip.

Yesterday

High 51.6 | Low 43.3 °F

Precip. 0.2 in [Radar Loop]

Pressure	30.22 in
Visibility	10.0 miles
Clouds	Overcast 3200 ft
Dew Point	37 °F
Humidity	60%
Rainfall	0.01 in
Snow Depth	Not available.

Sun & Moon



7:19 am



4:37 pm



Waning Gibbous,

64% visible

METAR KNYC 311751Z AUTO VRB05KT 10SM
OVC032 08/01 A3006 RMK AO2 SLP172
T00830011 10083 20078 58018 \$

10-Day Weather Forecast

Graph

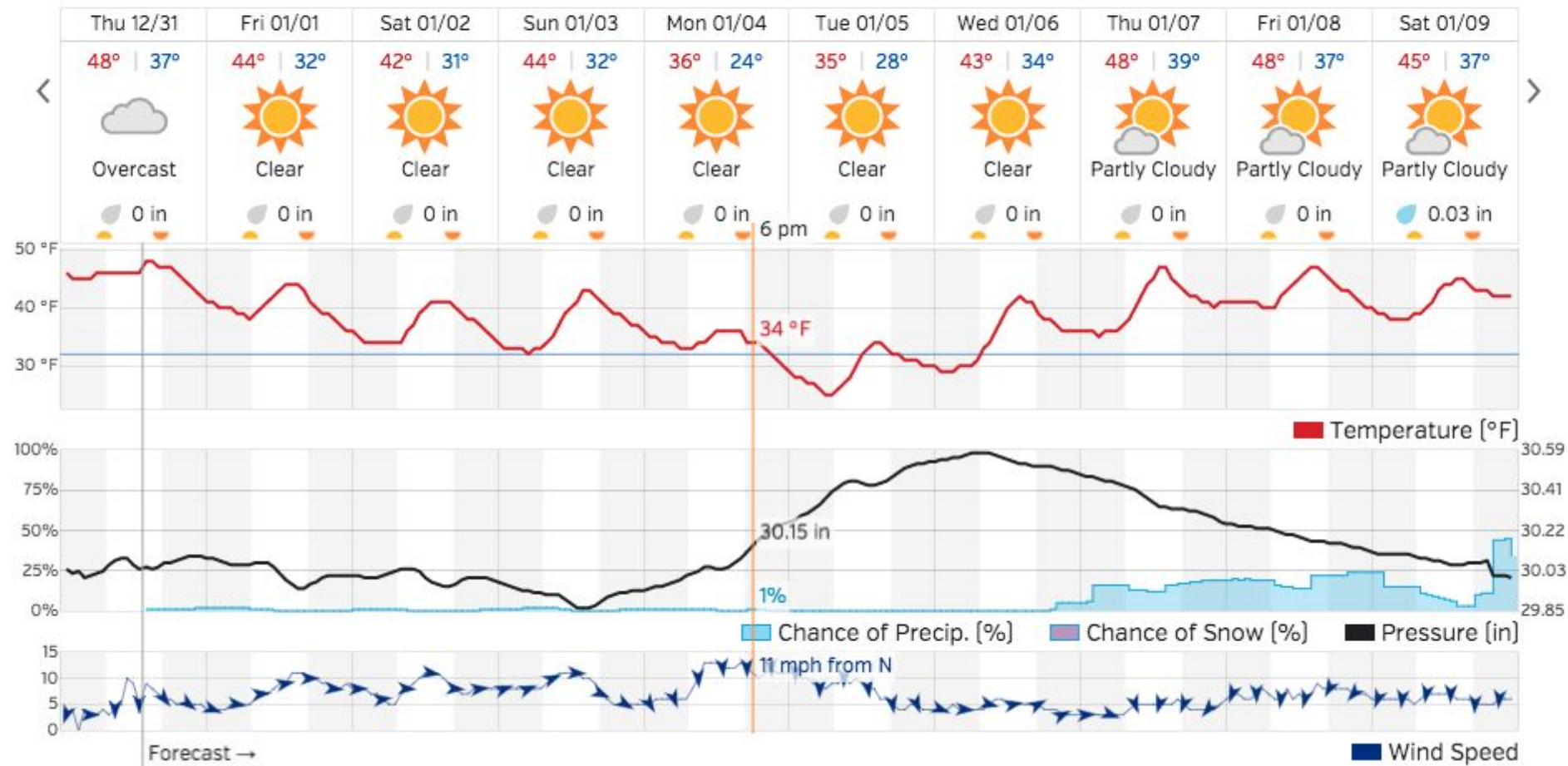
Table

Descriptive

Daily

Hourly

Customize



% of students with the following math assessment scores

Assessment Score Range	This class	My other class	School	District
<=50%	10%	9%	10%	12%
51-60%	4%	10%	11%	12%
61-70%	17%	23%	23%	26%
71-80%	20%	25%	26%	24%
81-90%	33%	23%	24%	20%
91-100%	17%	11%	10%	9%

Why was this a good design?

Attribute	% Weight
Comprehensive information	9
Important information is highlighted	9
Use of graphics whenever appropriate	9
Good choice and design of graphics	9
Aesthetically pleasing visual design	8
Sufficient information to decide if action is necessary	8
Good hierarchy of importance by salience	7
Good support for comparisons	7
Legibility	7
Organization is clear	6
Good hierarchy of importance by position	6
Everything is visible without scrolling or paging	5
Clear meanings	4
Good use of space	3
Scalable design	3
Overall	100

May 1, 2012
Tuesday

Grade 10 Algebra Course

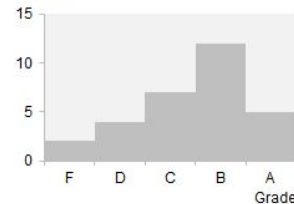
Note: All scores are expressed as percentage of points earned out of the total points possible.

HELP

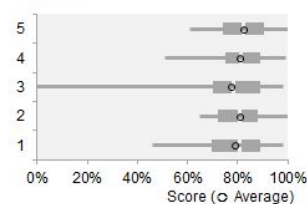
80% Term Complete		Course Grades					Class Discipline					Assignment Scores				
		Current					Tardy					Centered on Average				
		Previous Course					Referrals					Below Average				
		Last Roll					Late Assign.					Assignments 1 to 5				
		Current Grade					Last Assign.					Last Assign.				
		F	D	C	B	A	0	5	10	15	20	2	3	4	5	68%
Frederick Chandler																
Bae Kim																
Fiona Reeves																
Brian Francis																
Anthony Harper																
Christopher Murphy																
Kirsten Holmes																
Roshawn Dawson																
Nikolas Mikhailovich																
James Martin																
Blaine Harper																
George Smith																
Regan Potrero																
Britta Jones																
Scott Ortiz																
Xu Mei																
Jaime Goss																
Samuel Miller																
Maria Garcia																
Jose Domingo																
Lawrence Parker																
Fariah Jackson																
Sarah Jameson																
David Chenowith																
Alison Perry																
Amala Singh																
Hannah Li																
James Snow																
Donald Chase																
Holly Norton																

Grade and Assignment Score Distribution

Students

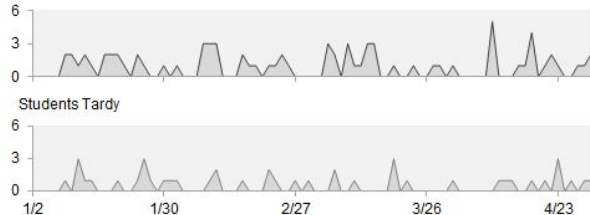


Assignment

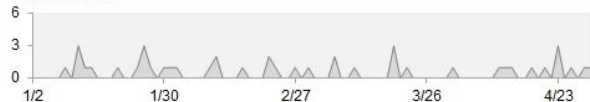


Attendance (excluding weekends)

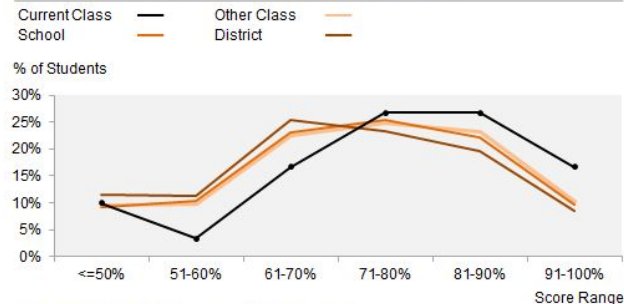
Students Absent



Students Tardy



Standardized Math Assessment Score Distribution



Standardized Math Assessment Median Score

Current Class	Other Class	School	District
79.0%	77.4%	74.2%	71.9%

The potential
risks of spending
too much time
ONLINE

Is the Internet **Bad** for Your Health?

Almost
everyone
uses the
Internet

In America:



Worldwide:



It's predicted that by 2016...

There will be
3 billion
Internet users
(almost half the world's
population)

The Internet
economy will reach
\$4.2 trillion
in the G-20
economies

It will rank as one of
the world's
top five
economies
(ahead of Germany)

How can using the Internet
affect our brain?

Addiction

Brain damage

Use small multiples

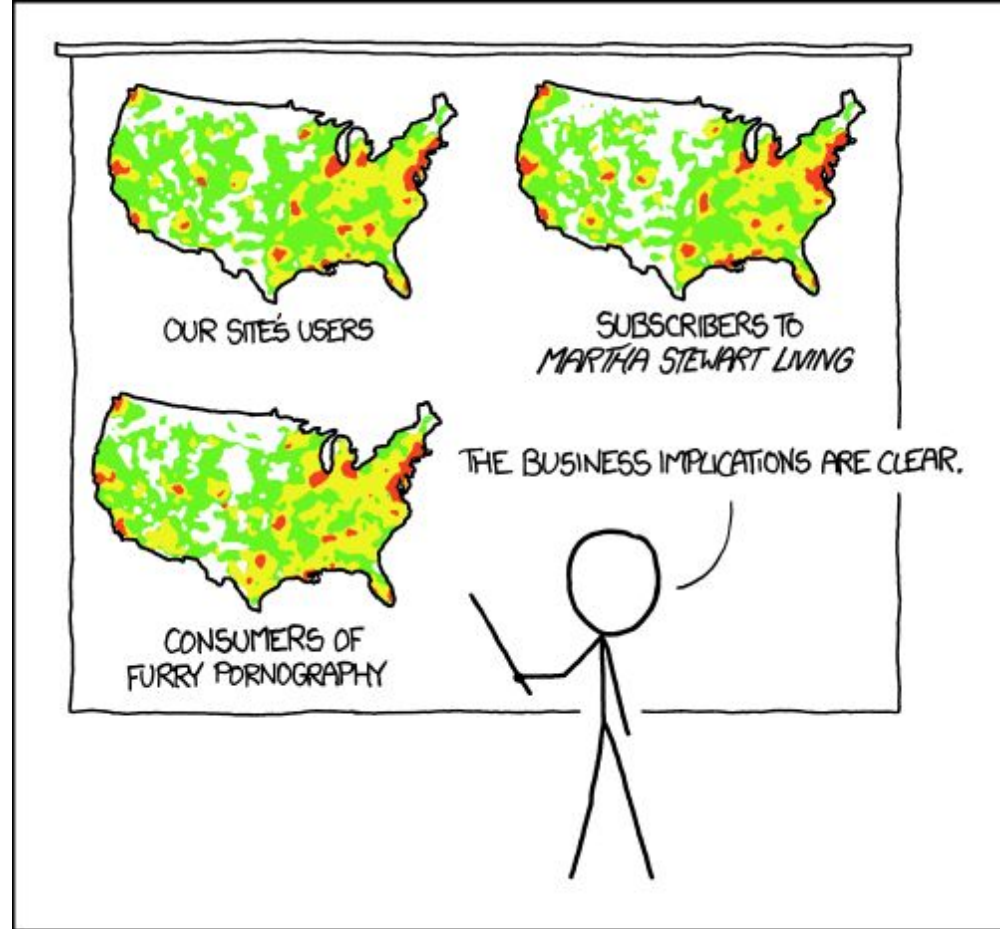
US Drought Index, 1895-2014



US Drought Index, 1895-2014



When dealing with geospatial data...

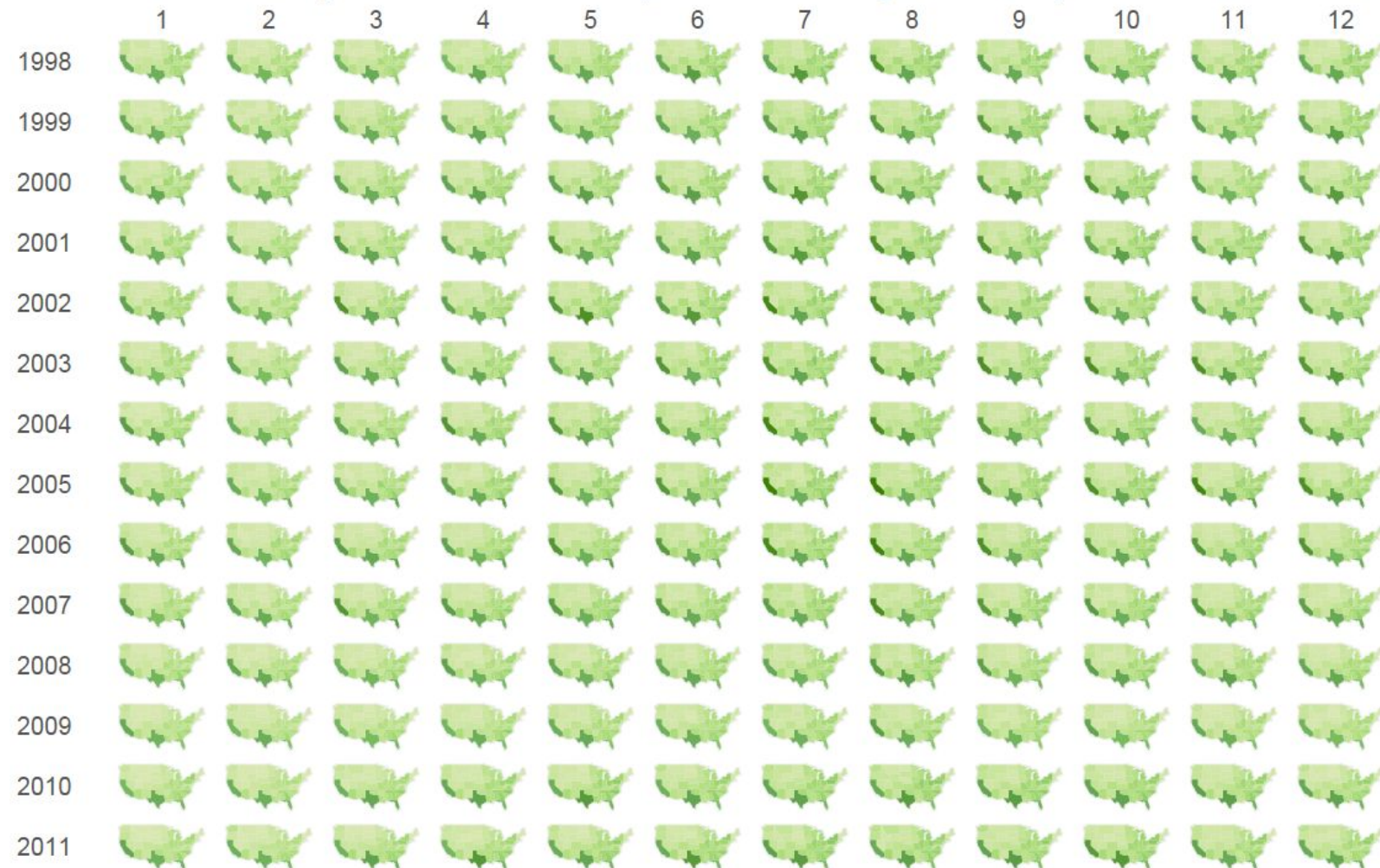


PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE
BASICALLY JUST POPULATION MAPS

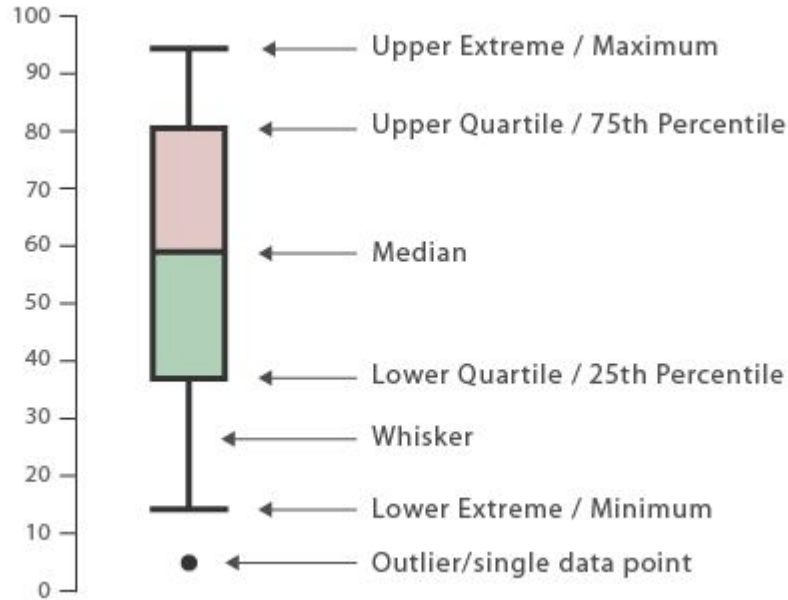
US Road Fatalities, 1998-2011

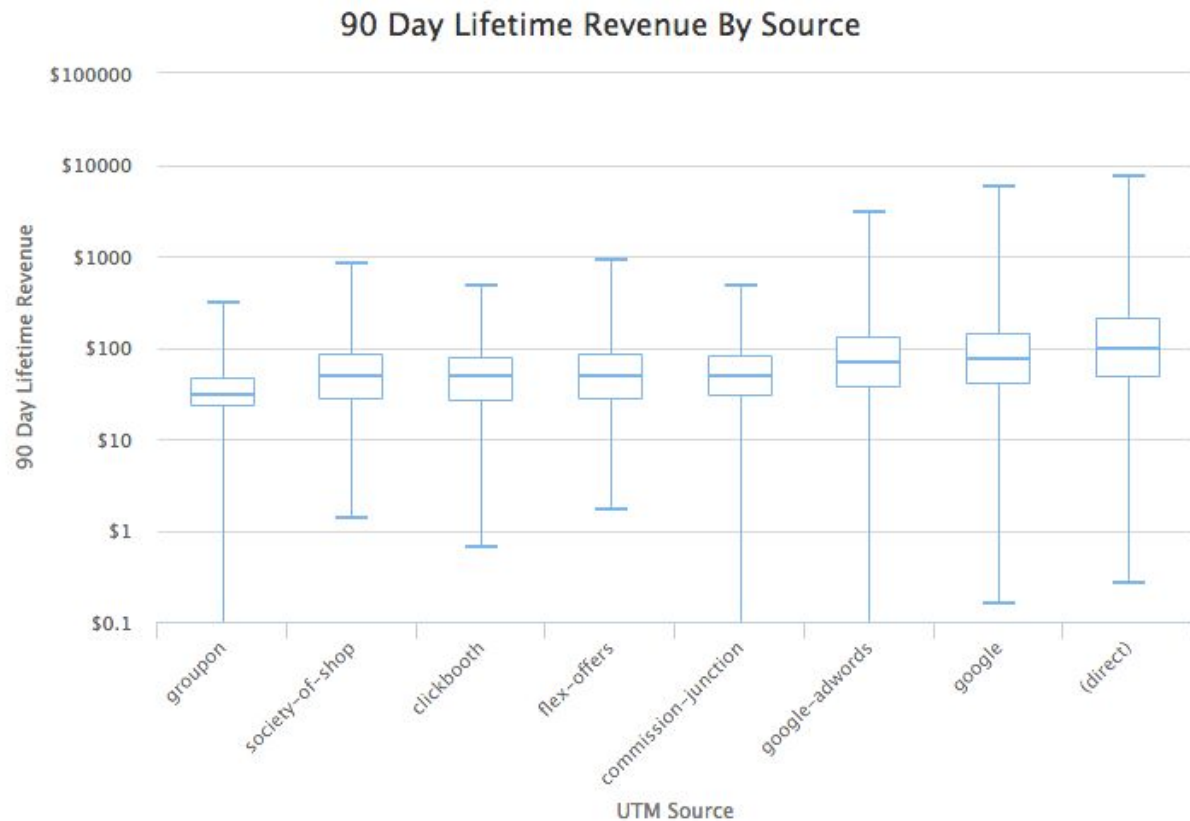
source: <http://www.fars.nhtsa.dot.gov/Main/index.aspx>

Fatalities 1 363



Use box and whisker plots in lieu of averages





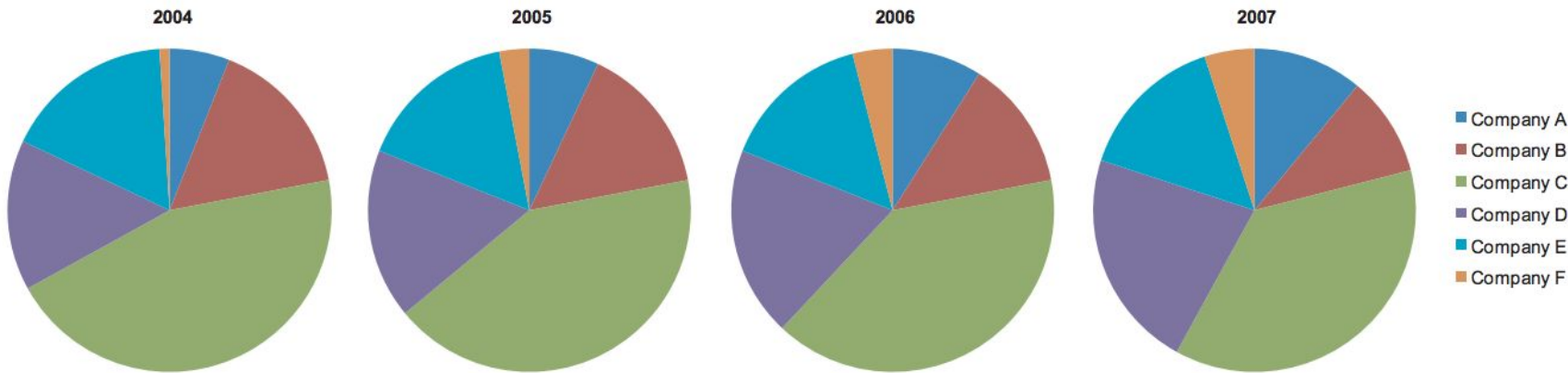
Customers whose first order source is a direct visit have a higher 90 day lifetime revenue than other sources.

With such a large range, an average wouldn't provide an accurate picture of reality.

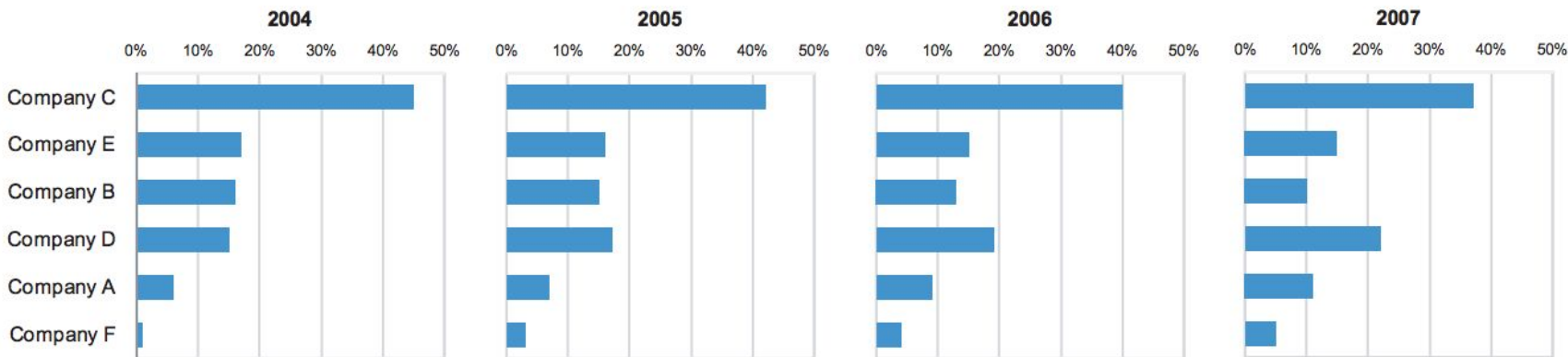
[Explore data](#)

“Save the Pies for dessert”

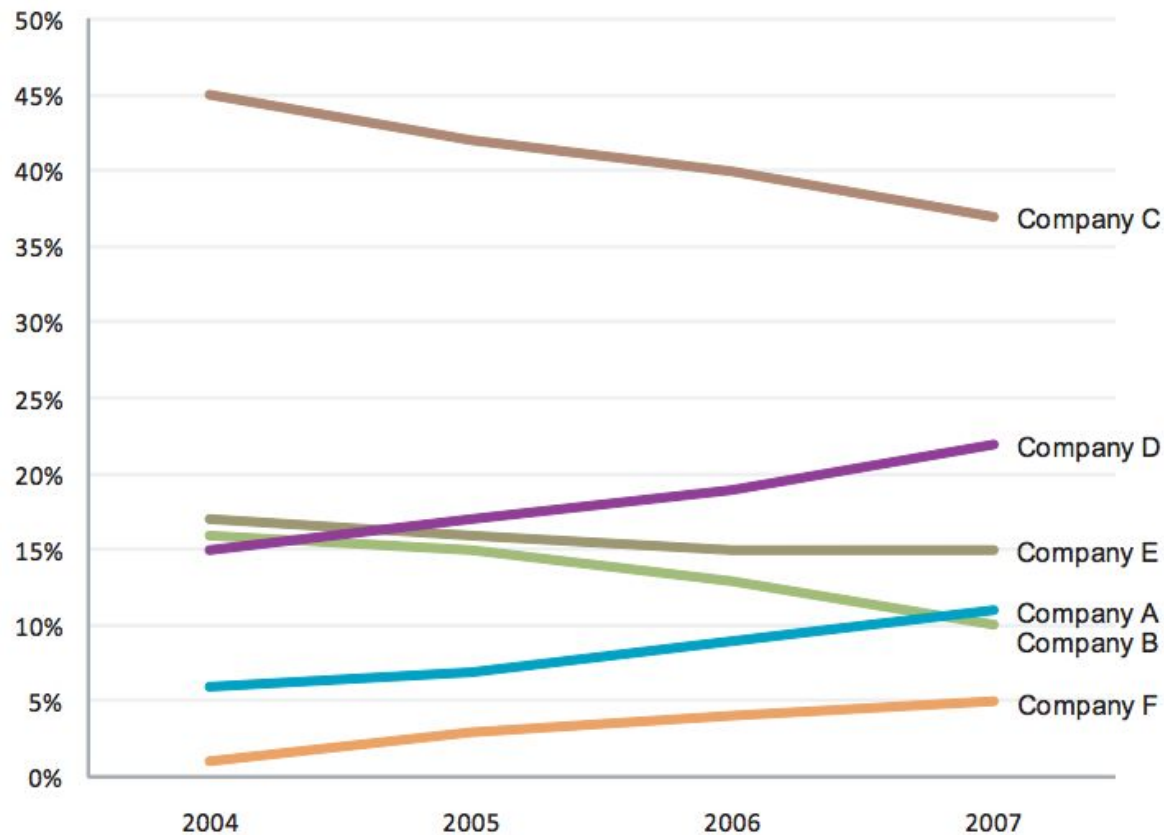
“The only worse design than a pie chart is several of them, for then the viewer is asked to compare quantities located in spatial disarray both within and between pies” - Ed Tufte



Try to follow the changes of these various companies and how they compare to one another through time. It is nearly impossible. Notice how easily you can do it, however, using the following display:

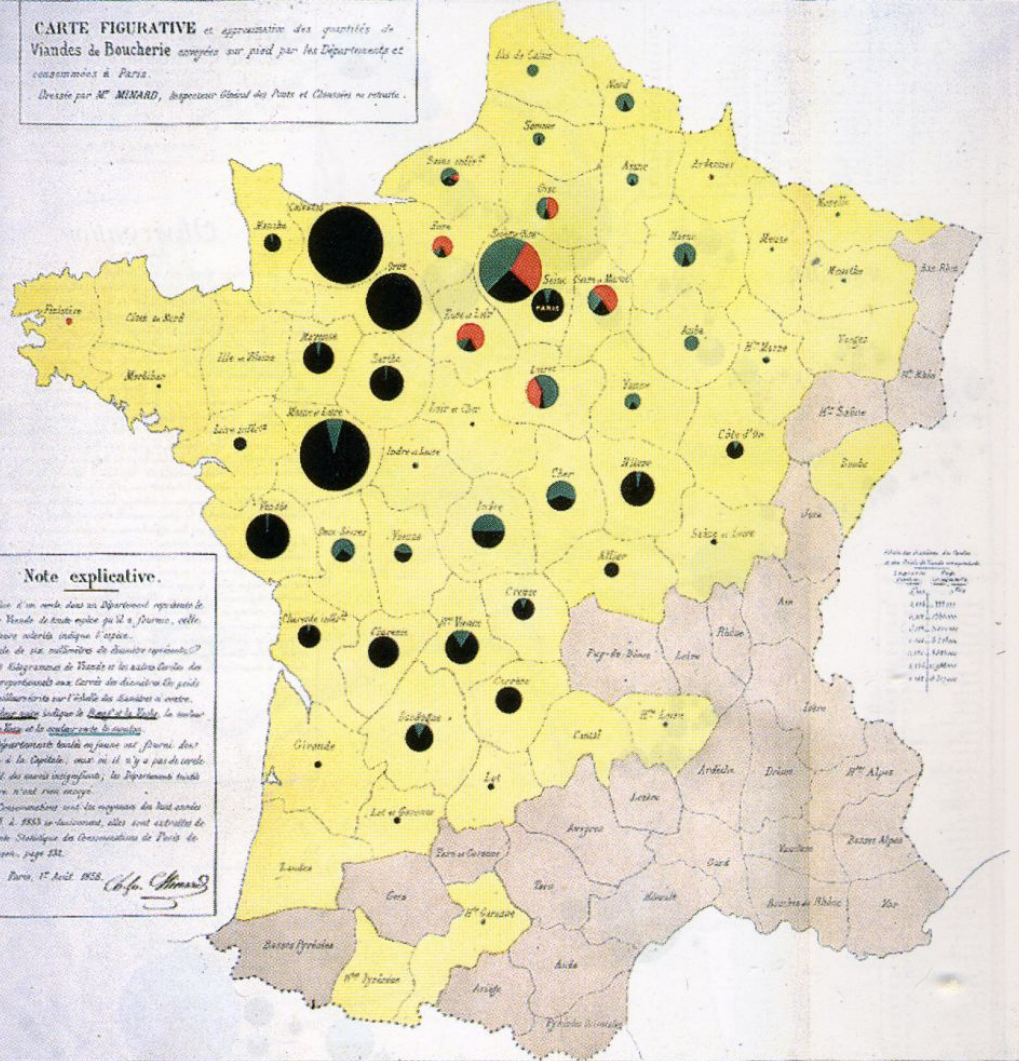


Company Percentages of Market Share by Year



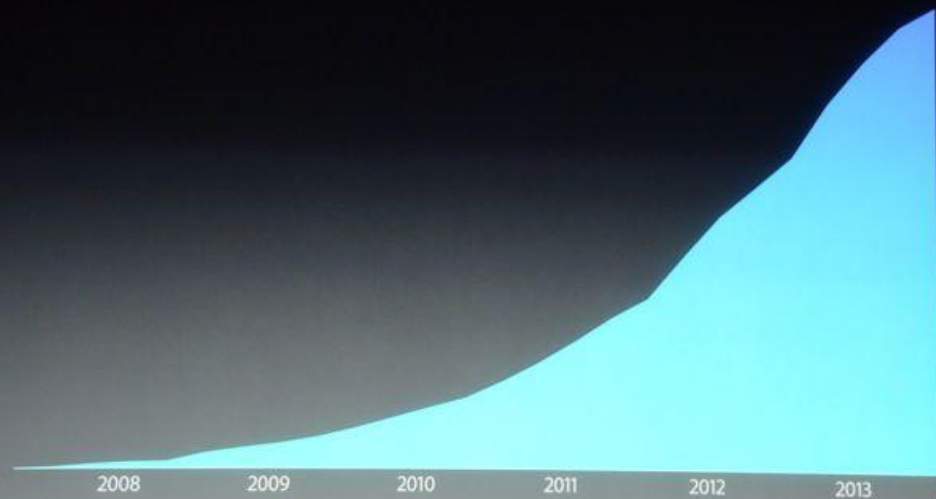
CARTE FIGURATIVE et approximation des quantités de
Viandes de Boucherie consommées sur pied par les Départements et
consommées à Paris.

Dressée par M. MINARD, Inspecteur Général des Pêches et Chasseurs de France.

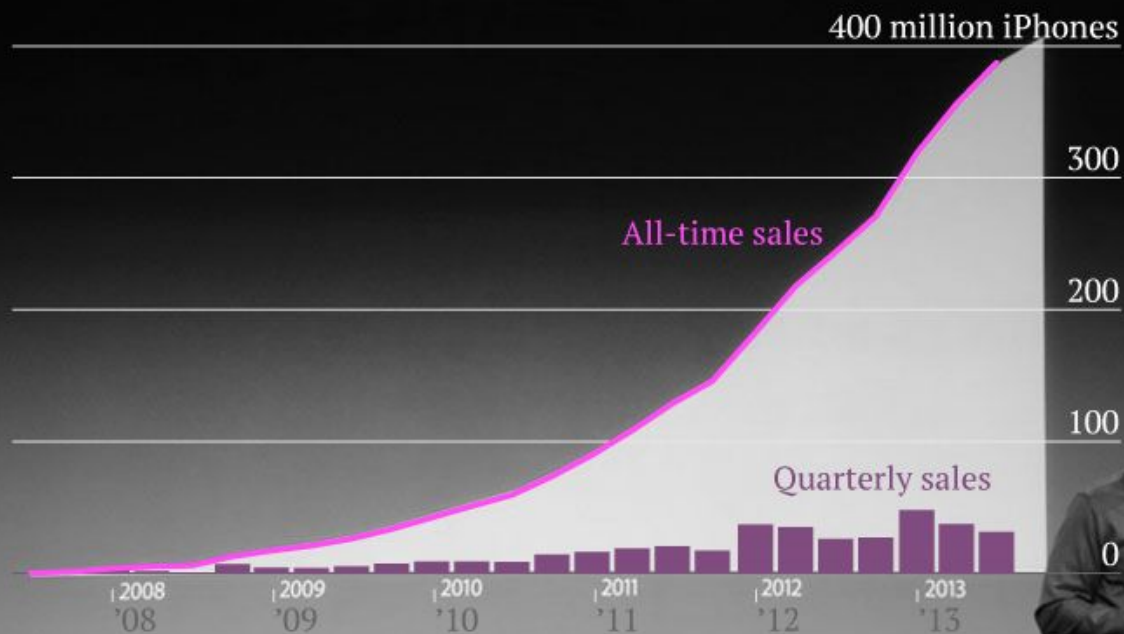


Cattle sent from around France for consumption in Paris. Charles Minard 1858.

Cumulative iPhone sales

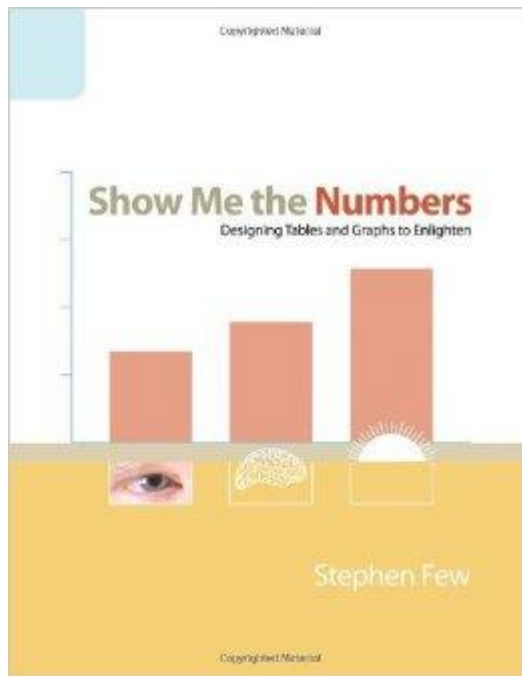


Cumulative iPhone sales

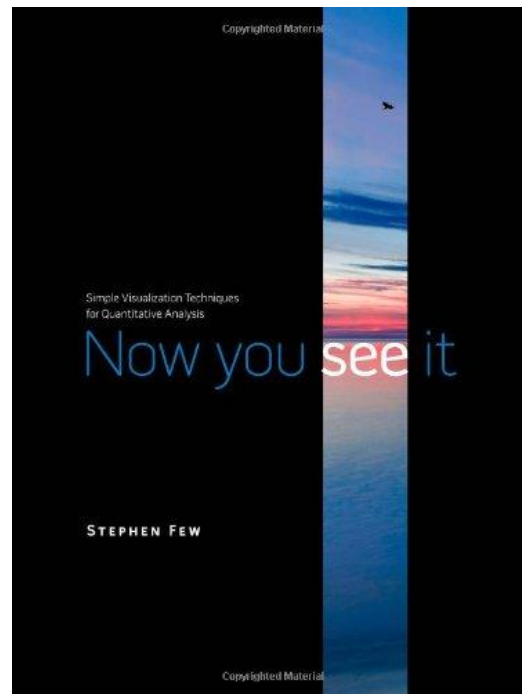


Check out these books by Stephen Few

Show me the numbers

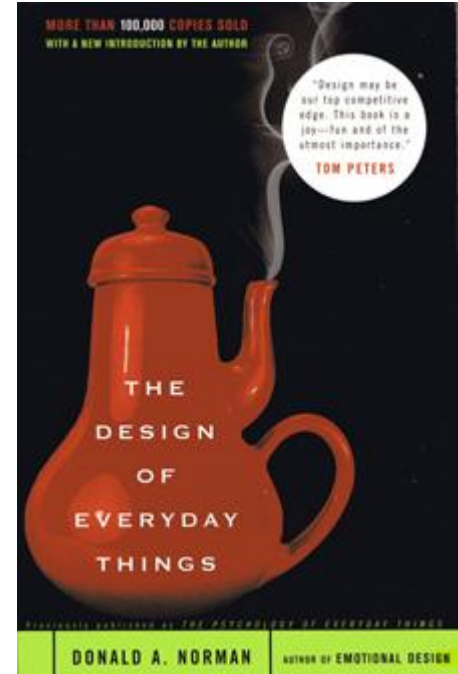


Now you see it



Knowledge in the world vs Knowledge in the head.

Put as much knowledge in the world as possible, to
reduce cognitive load, and the need to context switch.



\$66k

Data doesn't speak
for itself

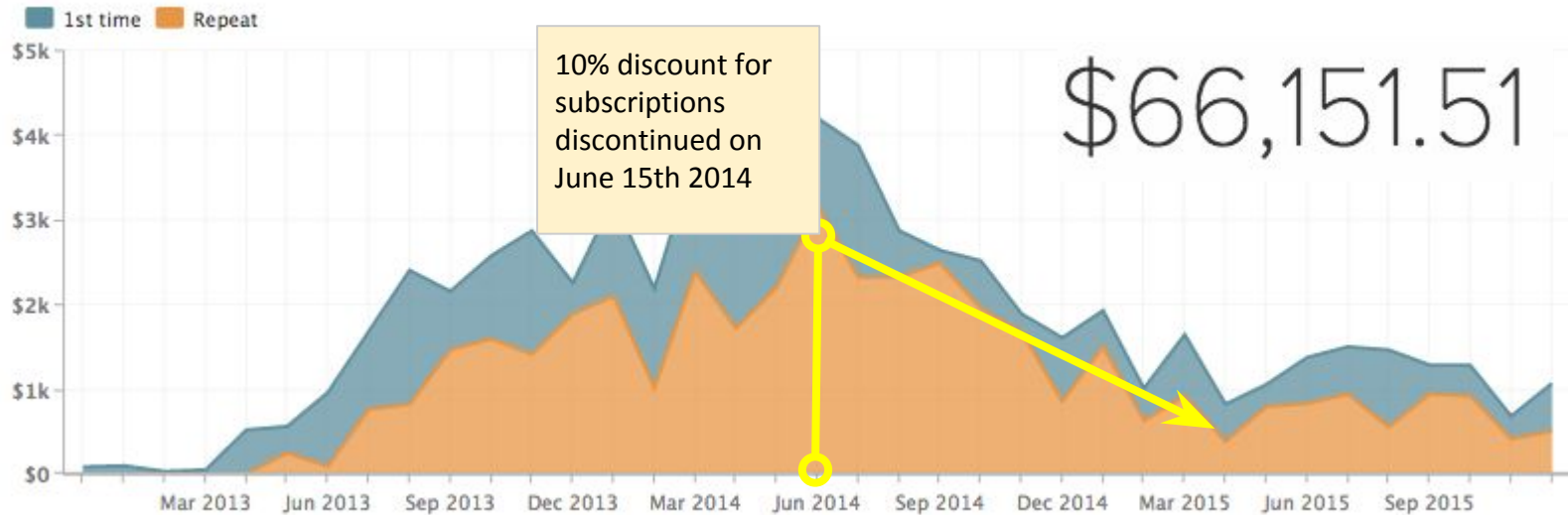
Reduce cognitive load by providing context.



Revenue is defined as all purchases. It does not account for returns or merchant fees.

[See the SQL Query](#): `Select SUM(revenue)`
from orders...

Are we bringing in fewer new customers or losing the ones we already had?



Recommend: Reinstating the 10% discount for subscription purchases.

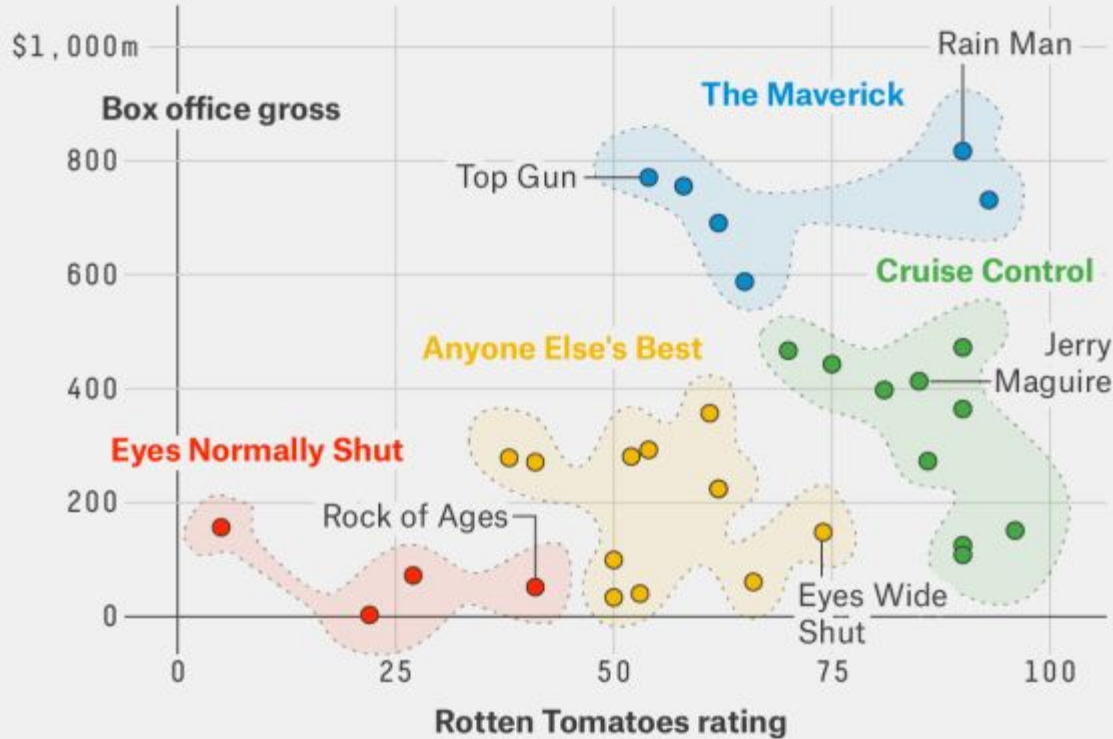
The Four Types Of Tom Cruise Movies

By WALT HICKEY



Tom Cruise Is Impossible

Box office gross in 2014 dollars vs. Rotten Tomatoes rating

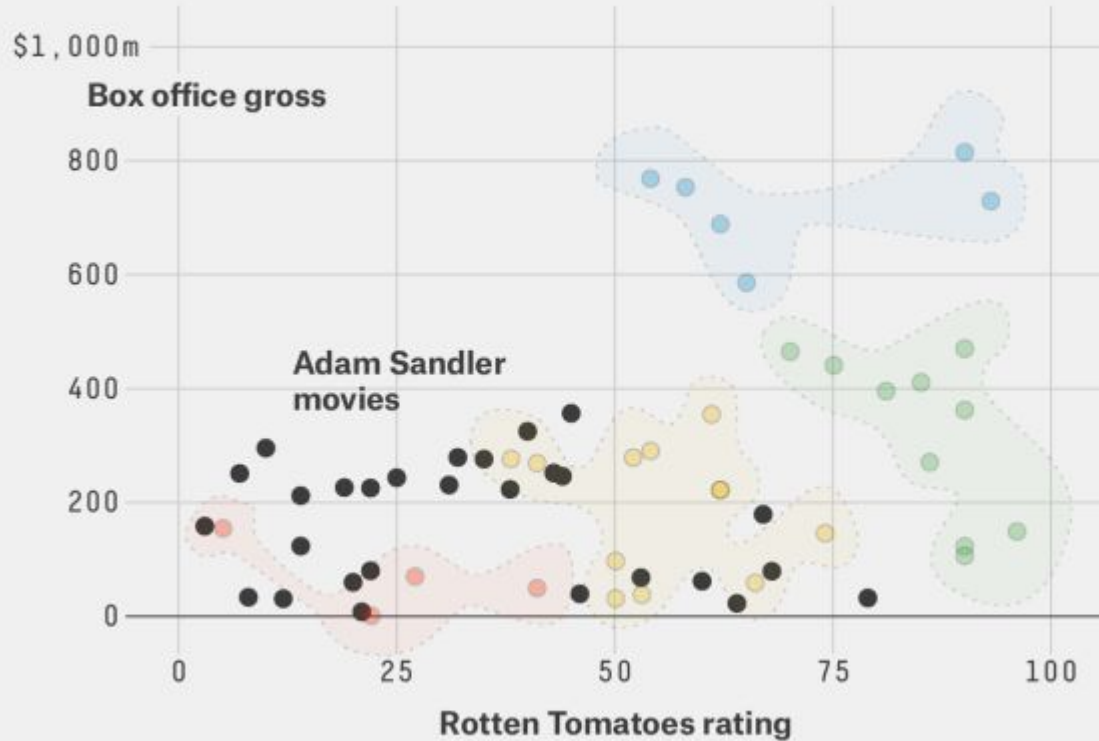


“His movies are better-reviewed (81 percent have ratings of 50 or higher on Rotten Tomatoes) and make more money than anyone else’s.”

Walt Hickey. fivethirtyeight

Mediocre Tom Cruise Is Better Than Most

Box office gross in 2014 dollars vs. Rotten Tomatoes rating



Good Adam Sandler is still not that good.

Questions?

Matt Monihan, Product @ RJMetrics
@mattmonihan