

# Interactive Data Visualization

Danne Woo and Peter Darche

danne@dannewoo.com  
@dannewoo  
www.dannewoo.com

pdarche@gmail.com  
@pdarche  
www.peterdarche.com

# Danne Woo

Design Technologist

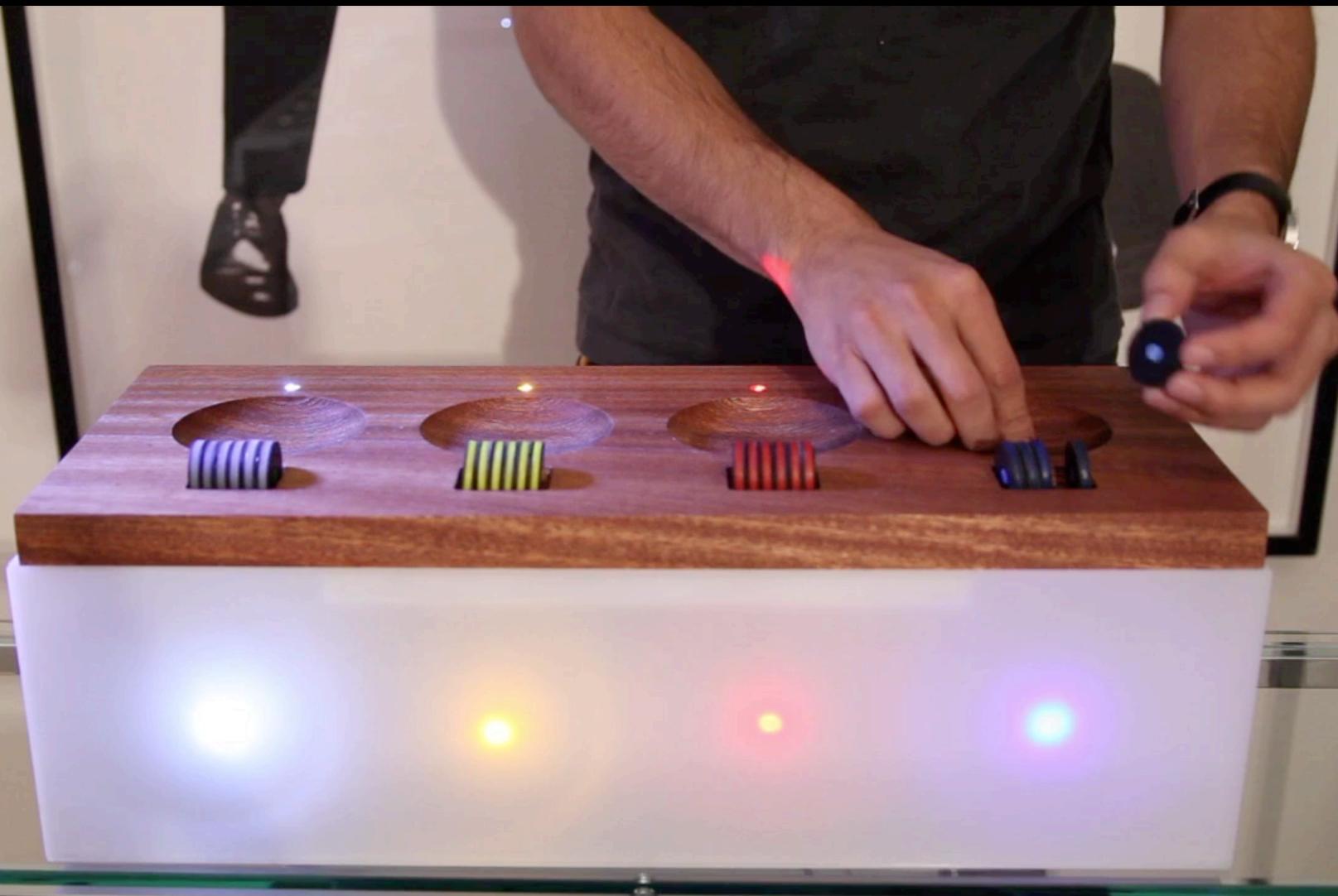
Interaction Design Professor

Software Designer/Developer

# SPLAT

Join team two at [splat.in/2](http://splat.in/2)







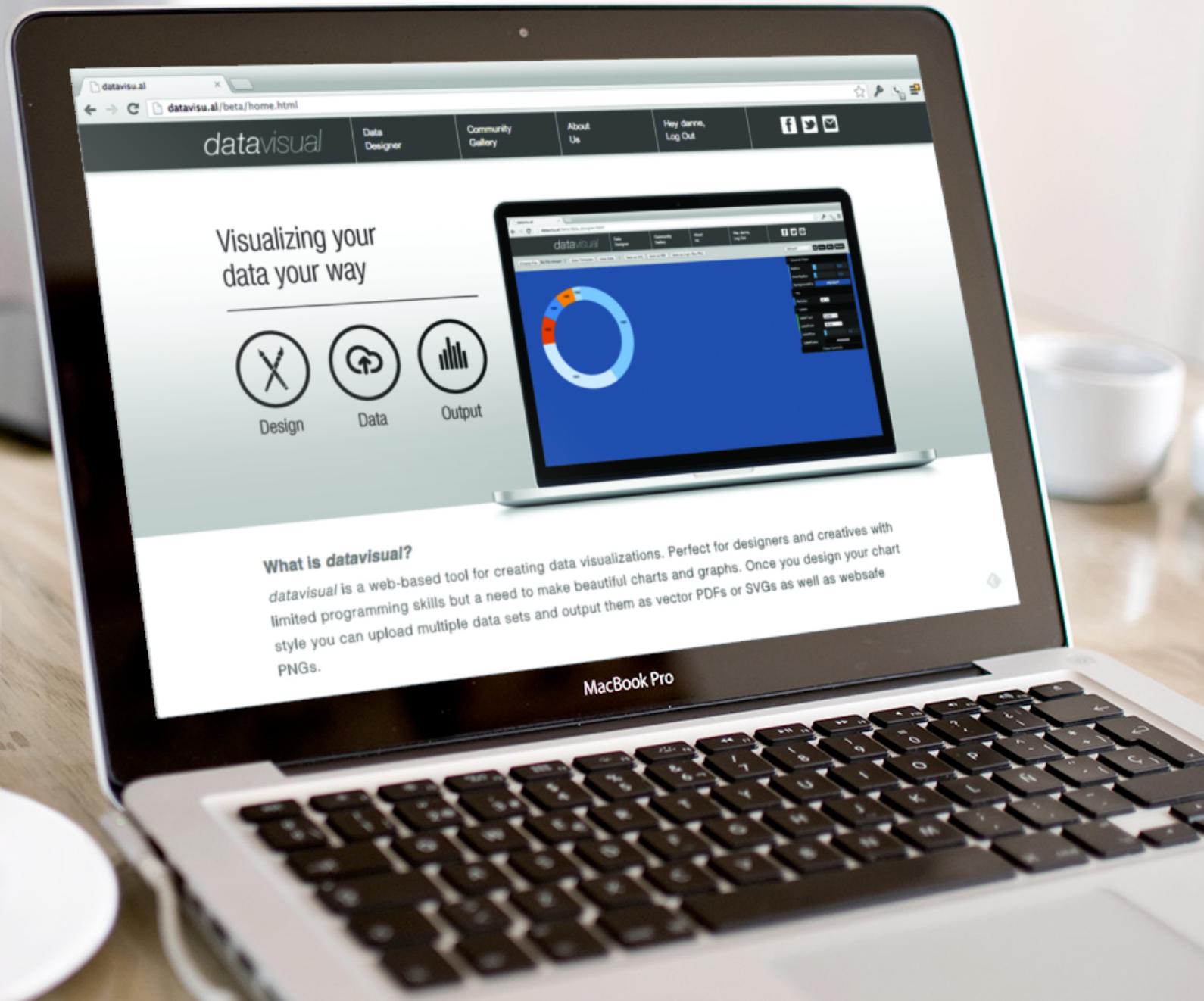
# Peter Darche

Data Scientist

Programmer

Software Designer/Developer









*datavisual*

**Your turn.**

Who are you?

Where are you from?

Why did this class interest you?

What type of data interests you most?

# Course Structure

**Day 1** Data Sourcing, Scrubbing and Analysis

**Day 2** Designing Data

**Day 3 + 4** Data, d3.js and the Web

**Day 5** Live Data and APIs

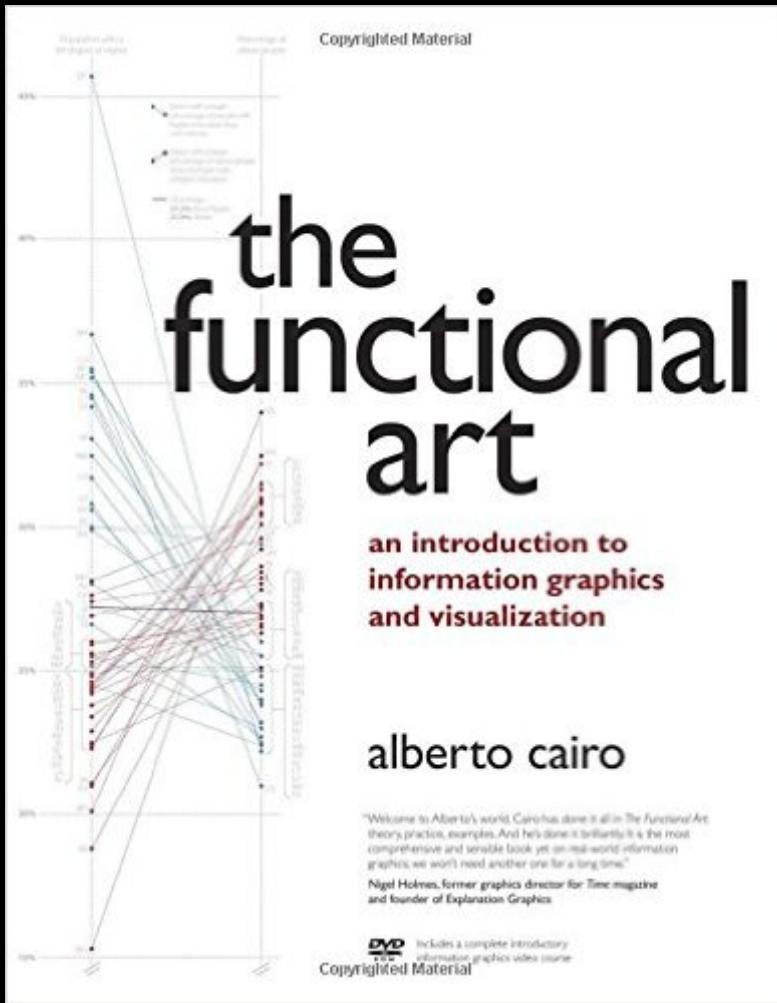
# **Project Brief**

Design and build an interactive data visualization or information graphic that visually clarifies our reliance on electricity, how we generate power, how we use it and/or its effects on our planet. Formulate a question around this topic, use this question as your basis for your research, collect data from reputable sources to support your answer to this question and create an interactive visualization using this data.

# Suggested Readings

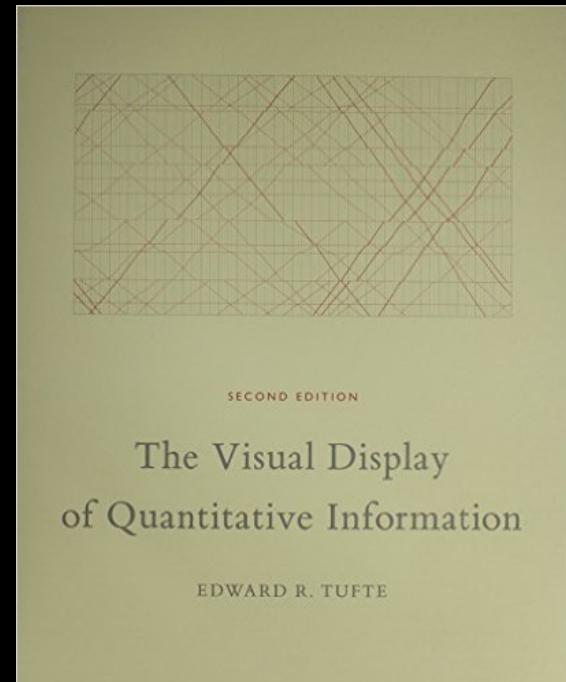
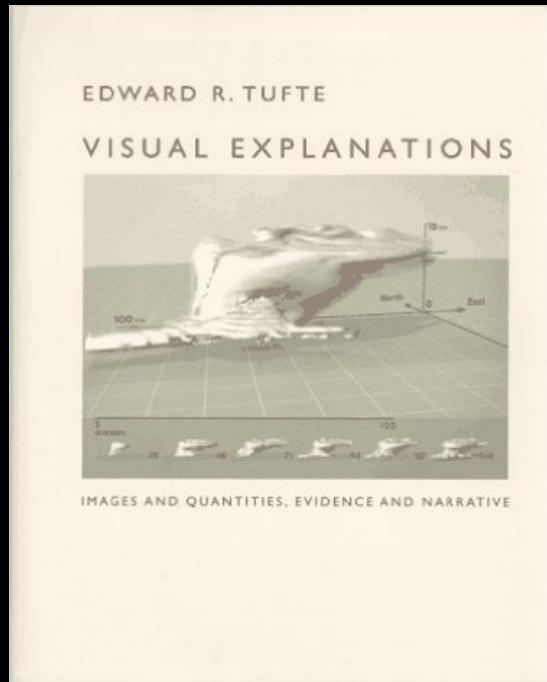
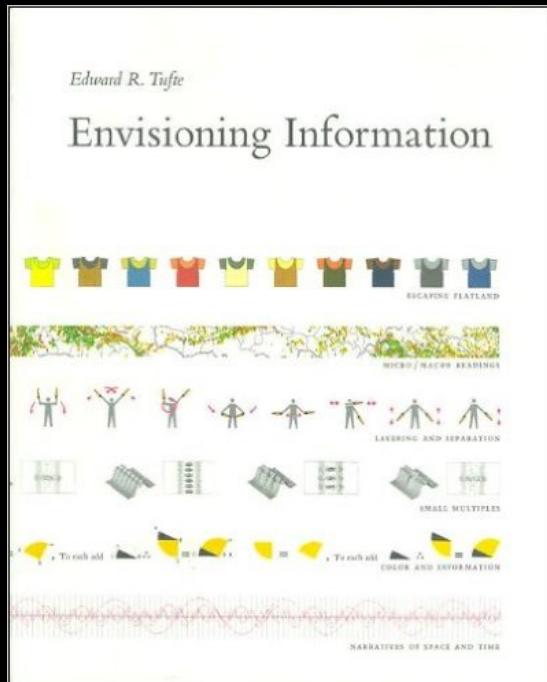
The Functional Art  
By Alberto Cairo

[www.thefunctionalart.com](http://www.thefunctionalart.com)

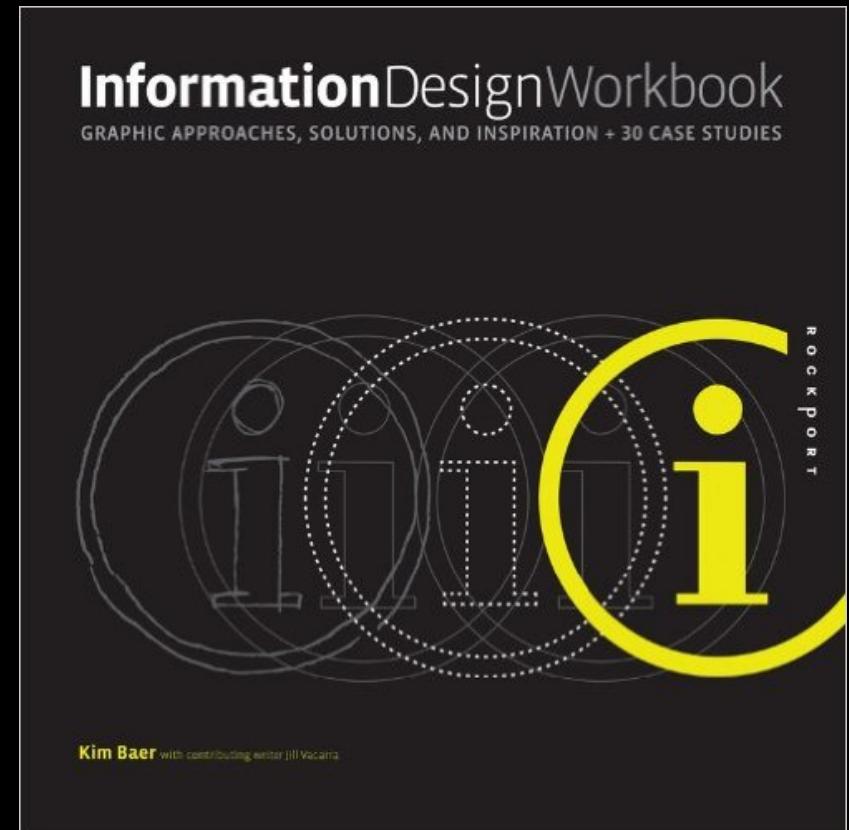
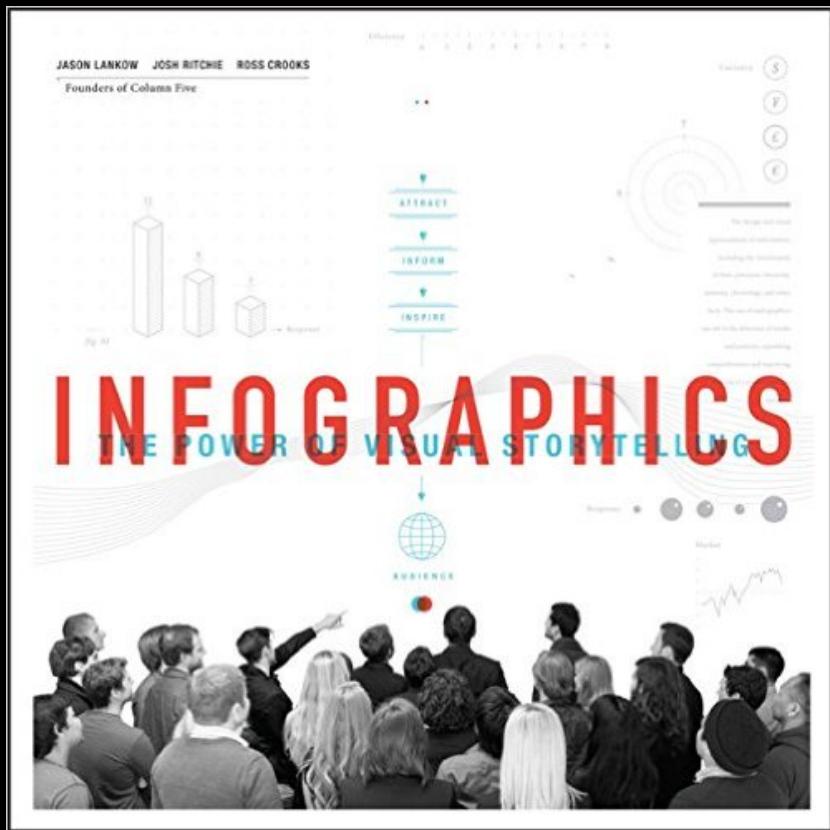


# Suggested Readings

## Edward Tufte Books

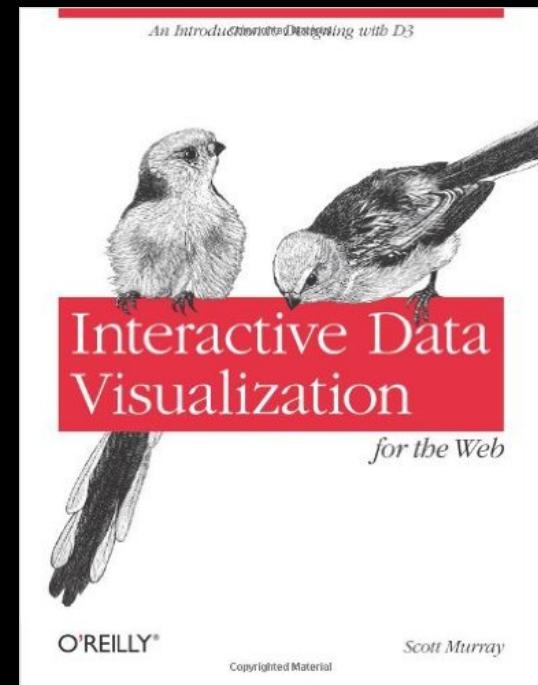
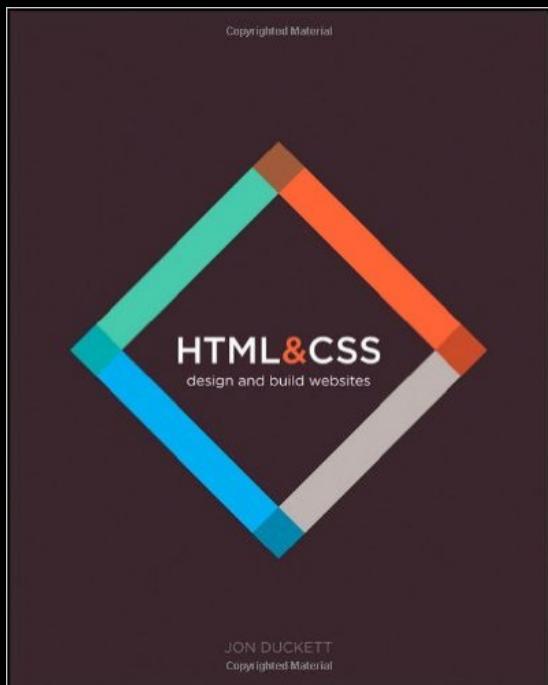


# Suggested Readings



# Suggested Readings

## Web Development



# Git Repo

github.com/dannewoo/ciid-data-viz

The screenshot shows a GitHub repository page for 'ciid-data-viz' owned by 'dannewoo'. The repository has 7 commits, 1 branch, 0 releases, and 1 contributor. The latest commit was made 2 months ago. The README.md file contains the following text:

## CIID Data Viz workshop (April 23rd - 27th, 2018)

### Day 1 - Data Pipeline

- The Data Pipeline
  - Data Collection/Sourcing

# 14-Week Class

dataviz.dannewoo.com

The screenshot shows a web browser window with a dark theme. The title bar reads "Data Visualization" and the address bar shows the URL "dataviz.dannewoo.com". The main content area displays the course information.

**Data Visualization**

**MENU**

- Introduction

**SYLLABUS**

- Week 1: Introduction to Data: Source, Scrub, Analyze and Visualization (Data Pipeline)
- Week 2: Sourcing Data, Collecting Data and Big Data
- Week 3: Scrubbing and Analysis
- Week 4: Data Visualization (Illustrator vs Code)
- Week 5: Infographics
- Week 6: Data Storytelling and

**Introduction**

Queens College  
ARTS 370 – 01  
Data Visualization

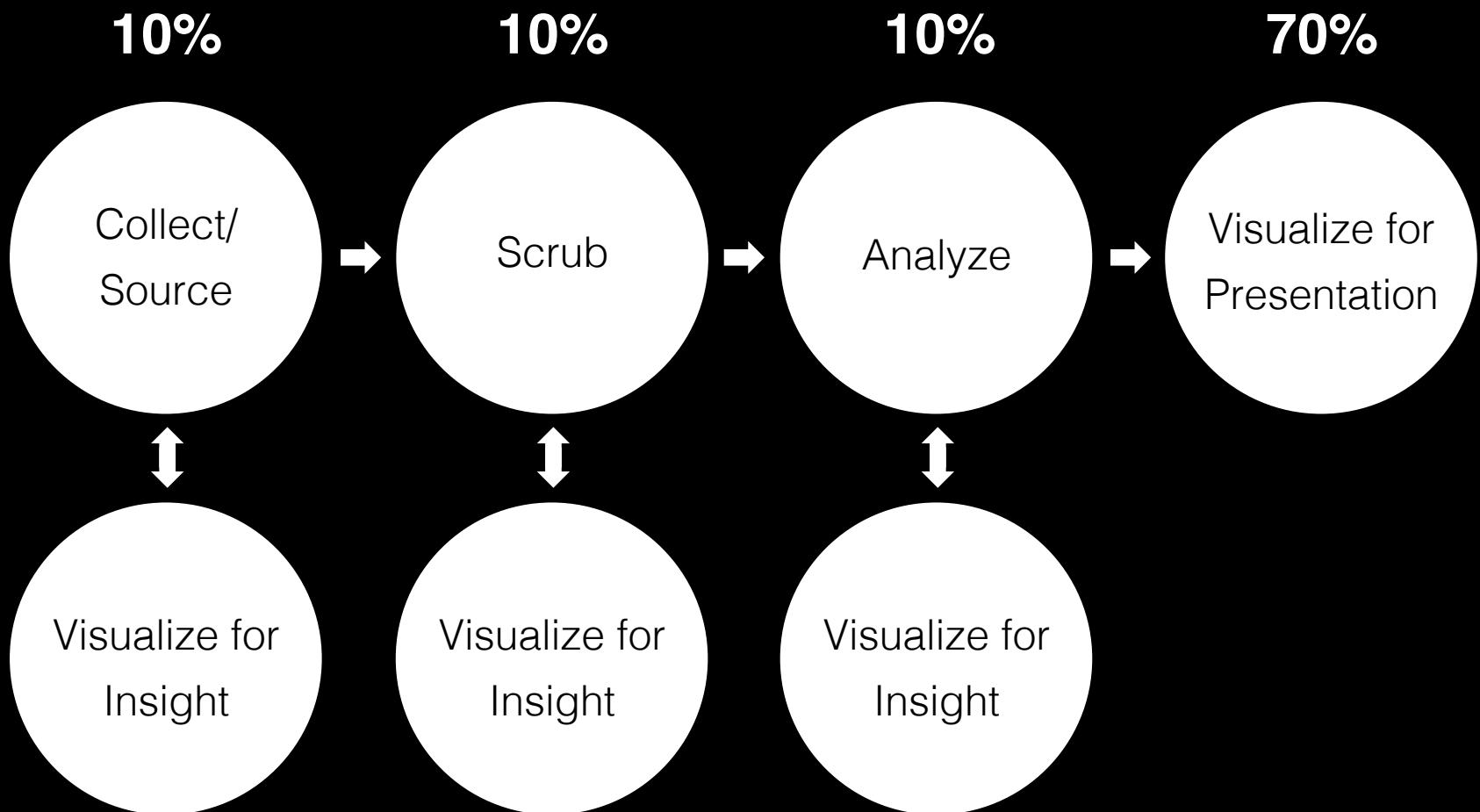
Fall 2015  
Wednesday 1:40PM – 5:30PM  
I-Building 203

**Course Description**

The massive amounts of data that we produce as a culture is steadily rising year after year. This evergrowing sea of information needs to be understood. Since we are all naturally visual people, the best way to understand this data is to graphically interpret it as data visualizations.

Over the course of this semester we will cover this entire process.

# Data Pipeline



# What is Data?

A collection of facts.

Numbers

Observations

Words

Pictures

Measurements

Videos

Descriptions

Audio

# Data

## Data Types – Structured vs Unstructured

Rank This Week	Song	Artist
1	Despacito	Luis Fonsi & Daddy Yankee Featuring Justin Bieber
2	Wild Thoughts	DJ Khaled Featuring Rihanna & Bryson Tiller
3	Bodak Yellow (Money Moves)	Cardi B
4	Believer	Imagine Dragons
5	Attention	Charlie Puth
6	Unforgettable	French Montana Featuring Swae Lee
7	There's Nothing Holdin' Me Back	Shawn Mendes
8	That's What I Like	Bruno Mars
9	Shape Of You	Ed Sheeran
10	Rake It Up	Yo Gotti Featuring Nicki Minaj
11	Strip That Down	Liam Payne Featuring Quavo
12	Bank Account	21 Savage
13	Body Like A Back Road	Sam Hunt
14	Slow Hands	Niall Horan
15	Congratulations	Post Malone Featuring Quavo
16	I'm The One	DJ Khaled Featuring Justin Bieber, Quavo, Chance The Rapper & Lil Wayne
17	Redbone	Childish Gambino
18	Sorry Not Sorry	Demi Lovato
19	Humble.	Kendrick Lamar
20	Friends	Justin Bieber + BloodPop
21	Something Just Like This	The Chainsmokers & Coldplay
22	Feels	Calvin Harris Featuring Pharrell Williams, Katy Perry & Big Sean
23	Feel It Still	Portugal. The Man

Comin' over in my direction

So thankful for that, it's such a blessin', yeah

Turn every situation into heaven, yeah

Oh-oh, you are

My sunrise on the darkest day

Got me feelin' some kind of way

# Data

Structured or Unstructured?

Cincinnati Reds vs. San Diego Padres

espn.go.com/mlb/boxscore?gameId=350811125

**GAME HQ**  
Scores for Aug 11, 2015

**Final** Series: Game 2 of 3

**Reds 6** (49-62, 21-36 away) vs **Padres 11** (54-60, 26-28 home)

Watch Highlights

10:10 PM ET, August 11, 2015  
Petco Park, San Diego, California

1	2	3	4	5	6	7	8	9	R	H	E
CIN	0	0	0	1	2	0	0	0	6	11	2
SD	3	5	3	0	0	0	0	-	11	11	1

W: C. Rea (1-0)  
L: M. Lorenzen (3-8)

**Recap** **Box Score** **Play-By-Play** **Photos 40** **Conversation 24**

**C Cincinnati Reds**

Hitters	AB	R	H	RBI	BB	SO	#P	Avg	OBP	SLG
Phillips 2B	4	0	2	0	1	0	15	.286	.325	.378
Schumaker LF	3	0	0	1	1	1	19	.213	.296	.287
Votto 1B	4	1	1	0	0	2	15	.299	.430	.515
Frazier 3B	4	1	3	2	0	0	17	.260	.313	.530
Bruce RF	4	0	0	0	0	1	18	.246	.324	.468
De Jesús SS	4	0	0	0	0	1	15	.276	.339	.438
Barnhart C	4	2	3	0	0	0	9	.259	.338	.353
Lorenzen P	0	0	0	0	0	0	0	.276	.276	.345
Axelrod P	2	0	0	0	0	1	6	.000	.000	.000
a-Byrd PH	1	0	0	0	0	1	6	.243	.296	.468
Badenhop P	0	0	0	0	0	0	0	.000	.000	.000
Parra P	0	0	0	0	0	0	0	.000	.000	.000

**San Diego Padres**

Hitters	AB	R	H	RBI	BB	SO	#P	Avg	OBP	SLG
Solarte 3B	4	3	2	0	1	0	21	.270	.324	.414
B. Norris P	0	0	0	0	0	0	0	.000	.000	.000
Garcés P	0	0	0	0	0	0	0	.000	.000	.000
Alonso 1B	5	2	2	2	0	0	13	.272	.353	.366
Kemp RF	3	2	1	0	1	1	17	.256	.304	.392
M. Upton Jr. CF	1	0	0	0	0	0	5	.222	.294	.361
J. Upton LF	4	2	1	1	1	0	18	.252	.332	.447
Gyorko 2B	4	1	2	4	1	0	15	.235	.296	.356
Venable CF-RF	3	0	0	0	1	1	18	.255	.317	.378
Hedges C	4	0	2	1	0	1	14	.200	.244	.293
Amarista SS	4	0	0	0	0	0	18	.204	.261	.304
Rea P	3	1	1	0	0	2	13	.333	.333	.333

Ford es la Marca Favorita en América.<sup>1</sup>  
Presiona un vehículo para un oferta especial.

DISEÑA Y COTIZA Go Further Visita a tu Concesionario Ford Local

# Data

Structured or Unstructured?

```
19:03:46, 12ZH1bPJZDGHyUTHjQgYvUqXq87LqrGRKp, 9294716,
1B2YvBPW23S8v5ouPdKm4kozSzYNzmGb43, 240370,
8b7dfc4c42328512196333f2c23e1923b003e154edffc2486d071eba2578b5ba
19:03:45, 1LwtxmX13HZTDDcUEzmoW2Av2GxDBHeAMD, 399980000,
70066209f0bb4a2b0196ff3a0adff952d77cacf96890766c0d333371a7af0d70
19:03:44, 12BApnQD51QcqsjrBR8DCjcLdijHoDAs7P, 81342236,
fd53151661778c8162e4db747d38bdb9a9a0dafce4a375e5b2824291de67a782
19:03:44, 1DbCAqvSHPxKrGUc5K7E7UxBhsAk97Dya, 6827706,
8147b14900c0587b5f5ac1e57fb1bf17cb62fdf3896bde7021eb05f17462cb8
19:03:42, 1Dhh8W17PkRFrBYK7CZRZGHdUrGLGYfRK8, 999000000,
438dae53f595c20dc51e972ebfc6cf96046272608e4fa1d6ef2f593a20c82bb
19:03:42, 1BAyaPneSPVheVYvX2p4Qrm9BygtWyGVx2, 306651770,
```

# Data

Structured or Unstructured?

KATY PERRY on Twitter: "NZ Tickets & VIP packages for both #WITNESSTHETOUR Auckland shows are on sale now bit.ly/KatyPerry\_NZ"

The screenshot shows a tweet from Katy Perry (@katyperry) on the Twitter mobile interface. The tweet includes a profile picture of Katy Perry, a bio mentioning New Zealand tour dates, and a link. Below the tweet is a large, dark, sequined photograph of Katy Perry sitting on a staircase. The interface shows a navigation bar at the top with various icons and links, and a sidebar on the left with user information.

# Data

## Structured or Unstructured?

Ken: I have "heard" from two sources in this office that Enron will announce my successor today. I have not been advised of the selection and am concerned that this announcement will not treat my retirement in an appropriate and deserving manner...

Frankly, I think this matter has not been handled in a caring way and I have been accused of giving information on the selection to Tom DeLay, which is untrue. My office mates are all more aware of the candidates and their status than I have been.

I regret having to contact you on this matter but feel that despite my total cooperation, allegiance and dedication to Enron, I am not being treated with the courtesy and sensitivity that this matter deserves.

# Data

## Structured or Unstructured?

```
Delivered-To: kenneth.lay@enron.com
Received: by 10.112.161.231 with SMTP id xv7csp2952801bb; Tue, 25 Sep 2000 01:01:10 -0800 (PST)
Return-Path: <bounce+a5d1e6.0977-kenneth.lay=enron.com@enron.com>
Received: from mail-s83.mailgun.info (mail-s83.mailgun.info. [184.173.153.211]) by mx.enron.com with ESMTP id
z8si1471337yhb.134.2013.11.19.01.01.08 for <kenneth.lay@enron.com>; Tue, 25 Sep 2000 01:01:08 -0800 (PST)
Received-SPF: pass (enron.com: domain of bounce+a5d1e6.0977-kenneth.lay=enron.
com@enron.com designates 184.173.153.211 as permitted sender) client-ip=184.173.153.211;
Date: Tue, 25 Sep 2000 09:00:03 +0000
From: joe.hillings@enron.com
To: kenneth.lay@enron.com
Subject: Selection of My Successor – My Concern
Message-Id: <528b2893.J26eP57xAov4mdbI%joe.hillings@enron.com>
User-Agent: Mozilla/4.0 (compatible; Lotus-Notes/5.0; Windows-NT)
Mime-Version: 1.0
Content-Type: text/plain; charset="us-ascii"
X-Mailgun-Sid: WyI2M2M0NyIsICJhbWFkZXVzQGRlY29kZWQuY28iLCAiMDk3NyJd
Sender: joe.hillings@enron.com
Content-Transfer-Encoding: 7bit
```

# Data

Structured or Unstructured?



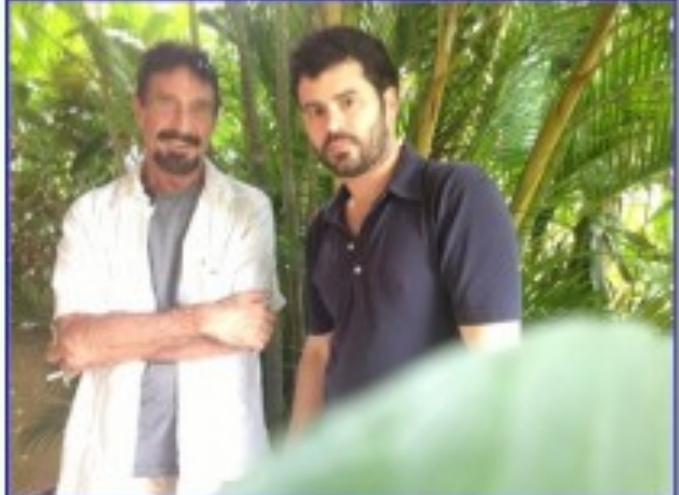
# Data

## Structured or Unstructured?

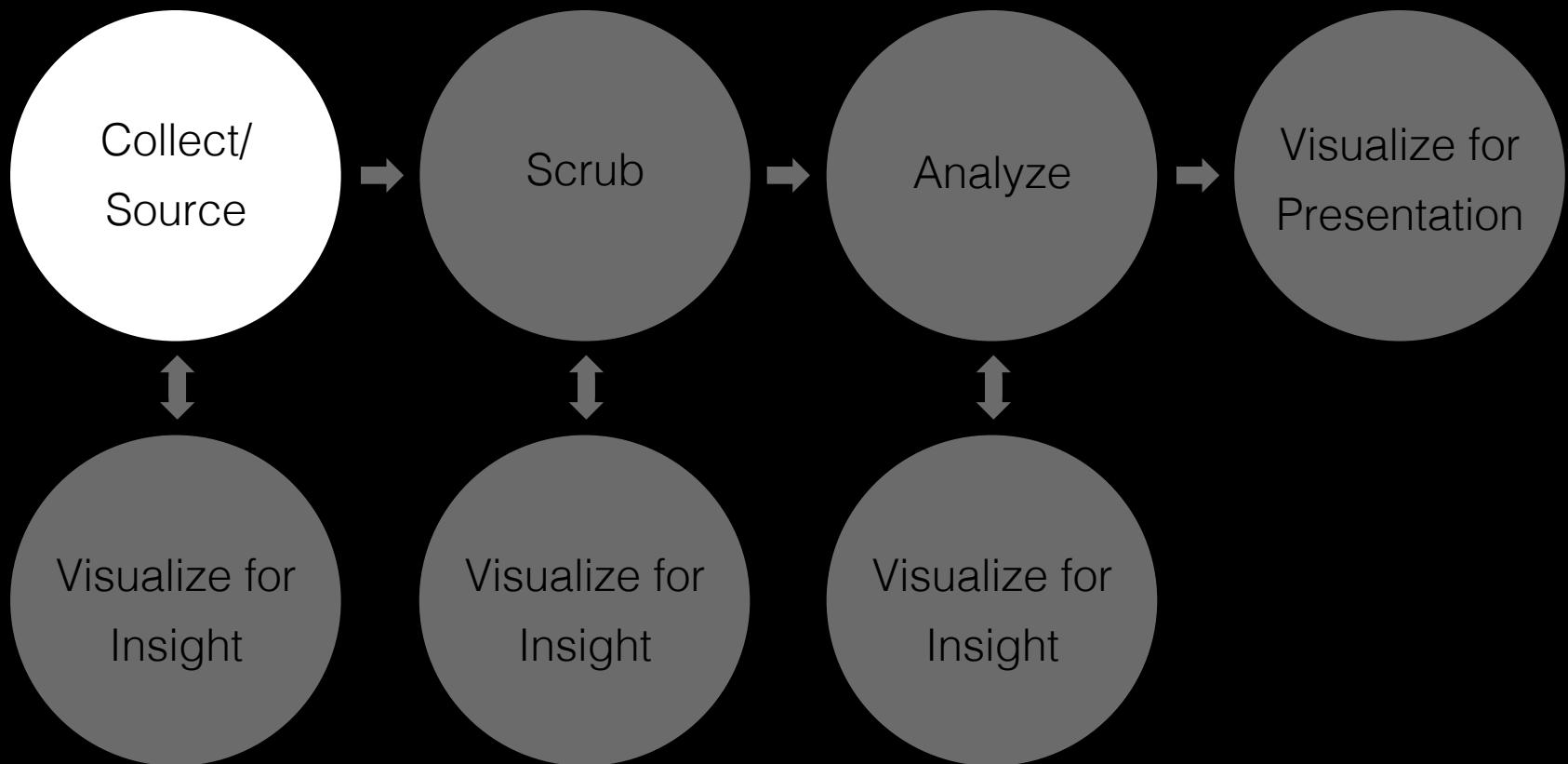
### Basic Image Information

Camera:	Apple iPhone 4S
Lens:	4.3 mm
Exposure:	Auto exposure, Program AE, 1/20 sec, f/2.4, ISO 125
Flash:	Off, Did not fire
Date:	December 3, 2012 12:26:08PM (timezone not specified) (2 hours, 44 minutes, 59 seconds ago, assuming image timezone of 6 hours behind GMT)
Location:	Latitude/longitude: 15° 39' 29.4" North, 88° 59' 31.8" West (15.660167, -88.992167)  Photos on Jeffrey's blog that are near this location.  Map via embedded coordinates at: <a href="#">Google</a> , <a href="#">Yahoo</a> , <a href="#">Wikimapia</a> , <a href="#">OpenStreetMap</a> , <a href="#">Bing</a> (also see the <a href="#">Google Maps</a> pane below) Altitude: 7,152,159.68 m Timezone guess from <a href="#">earthtools.org</a> : 6 hours behind GMT
File:	480 × 640 JPEG 132,481 bytes (0.13 megabytes) Image compression: 88% 4% crop of the 3,264 × 2,448 (8.0 megapixel) original
Color Encoding:	<b>WARNING:</b> Color space tagged as sRGB, without an embedded color profile. Windows and Mac browsers and apps treat the colors randomly.  Images for the web are most widely viewable when in the sRGB color space and with an embedded color profile. See my <a href="#">Introduction to Digital-Image Color Spaces</a> for more information.
Image URL:	<a href="http://assets.vice.com/content-images/contentimage/noslug/libcid75e0012f775d7dd621ef549fb673.jpg">http://assets.vice.com/content-images/contentimage/noslug/libcid75e0012f775d7dd621ef549fb673.jpg</a>  Apply other tools to this image via <a href="#">ImgOpt.com</a> .

© P P Main image displayed here at 70% width (49% the area of the original)



# Data Pipeline



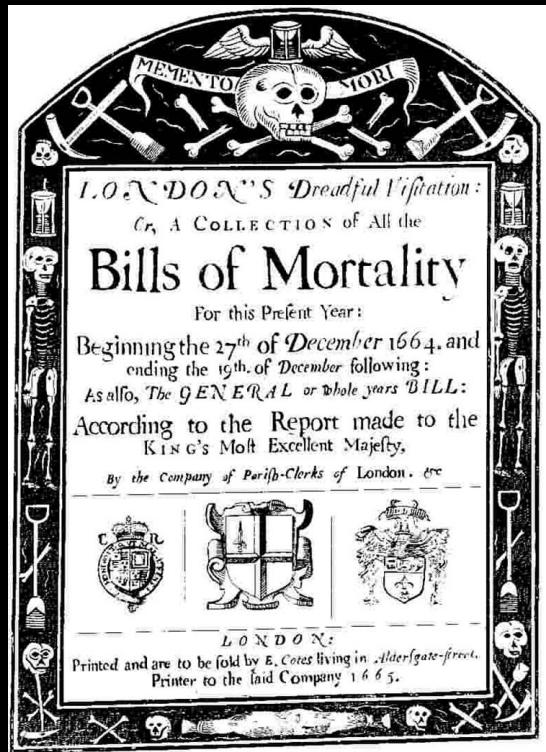
# Collecting Data

Blombos Ocher Plaque – 75000BC



# Collecting Data

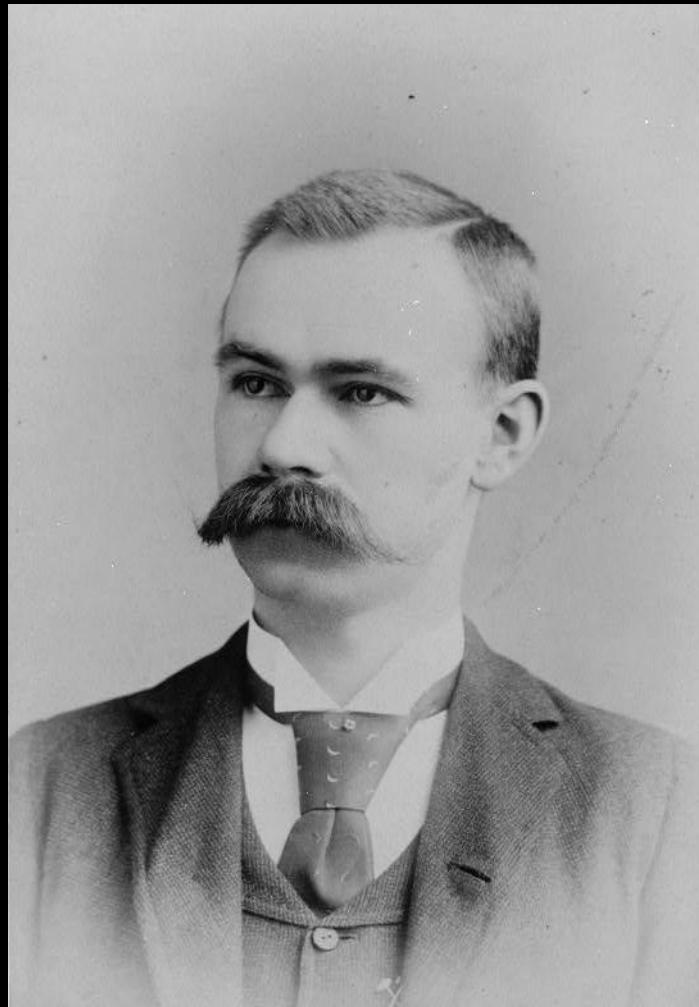
John Graunt – 1662



	1647	1648	1649	1650	1651	1652	1653	1654	1655	1656	1657	1658	1659	1660	1629	1630	1631	1632	1633	1634	1635	1636	1637	1638	In 20 Years,							
Abortive, and stillborn	335	329	327	351	380	381	384	433	483	419	463	467	421	544	499	439	410	445	500	475	507	523	1793	2005	1542	1587	1832					
Aged	916	835	889	696	780	834	864	974	743	892	869	1176	9091	1095	579	712	661	671	704	623	794	714	2475	2814	3330	4542	3080					
Ague, and Fever:	1260	884	751	970	1038	1212	1382	1371	689	875	999	1800	2303	2148	956	1021	1115	1108	953	1279	1022	2300	4118	6235	3865	4903	4163					
Apoplex, and fadainly	68	74	64	74	106	111	118	86	92	102	113	138	91	67	22	36	17	24	35	20	75	85	280	421	445	177	1306					
Bleach					1	3	7	2		1																						
Blasted	4	1				6	6		4		5	5	3	8	10	13	6	4		4	54	14	5	12	14	16	99					
Bleeding	3	2	5	1	3	4	3	2	7	3	5	4	7	2	5	2	5	4	4	3	16	7	14	12	19	17	65					
Bloody Flux, Scouring, and Flux	155	176	802	289	833	762	200	386	168	368	362	233	346	251	449	438	352	348	278	512	346	330	1587	1406	1422	2181	1161	1597	7818			
Burnt, and Scalded	3	6	10	5	11	8	5	7	10	5	7	4	6	6	3	10	7	5	1	3	12	3	25	19	24	31	26	19	125			
Calenture	1					1		2	1																							
Cancer, Gangrene, and Fistula	26	29	31	19	31	53	36	37	73	31	24	35	63	52	20	14	23	28	27	30	24	30	85	112	105	157	150	114	669			
Wolf						8																							8			
Canker, Sore-mouth, and Thrush	66	28	54	42	68	57	53	72	44	81	19	27	73	68	6	4	4	1	7	74	15	79	190	244	161	133	689					
Childbed	161	106	114	117	206	213	158	192	177	201	236	225	220	194	150	157	112	171	132	143	163	230	590	668	498	709	839	490	3364			
Chromes, and Infants	1369	1254	1065	990	1237	1280	1050	1343	1089	1393	1162	1144	858	1123	2590	2378	2035	2208	2130	2315	2113	1895	2277	8453	4078	4910	4788	4510	32106			
Colic, and Wind	103	71	85	82	76	102	81	101	85	120	113	179	116	107	48	57					37	50	105	87	341	359	497	247	1389			
Cold, and Cough										41	36	21	58	30	31	33	24	10	58	51	55	45	54	50	57	174	207	00	77	140	43	598
Convulsion, and Cough	2423	2200	2388	1988	2350	2410	2286	2386	2606	3184	2757	3610	2982	3414	1827	1910	1711	1757	1754	1955	2080	2477	5157	8260	8999	9914	2157	7197	44887			
Convulsion	684	491	530	493	569	653	666	818	702	1027	807	841	742	1031	52	87	18	21	221	336	418	709	408	1734	2198	2065	3377	1324	9073			
Cramp						1														0	0	0	0	0	0	0	0					
Cut of the Stone	2	1	3		1	1	2	4	1	3	5	40	48							5	1	5	2	5	10	0	4	13	47	38		
Drophy, and Tympany	185	434	421	508	444	550	617	704	660	706	631	931	646	872	235	252	279	280	266	250	329	389	1048	1734	1538	2121	2982	1303	9623			
Drowned	47	40	30	27	49	50	30	30	43	49	63	57	48	43	33	29	34	37	32	32	45	139	147	144	182	215	130	827				
Excessive drinking						2																						2				
Executed	8	17	29	43	24	12	19	21	19	22	20	18	7	18	19	13	12	18	13	13	62	52	97	76	79	55	384					
Fainted in a Bath						1																										
Falling-Sicknes	3	2	2	3		3	4	1	4	3	1	4	5	3	10	7	7	2	5	0	8	27	21	10	8	8	0	74				
Flox, and small pox	139	400	1150	184	525	1272	119	812	1294	823	835	409	1523	354	72	40	58	531	72	1354	293	127	701	1840	1913	2755	3361	2785	10576			
Found dead in the Streets	6	6	9	8	7	9	14	4	3	4	9	11	2	6	13	33	26	6	13	8	24	24	83	69	23	34	27	29	243			
French-Pox	18	29	15	18	21	20	20	29	23	25	53	51	31	17	12	12	7	17	12	22	53	48	80	81	130	83	392					
Frighted	4	4	1	3		3	2		1																							
Gout	9	5	13	9	7	7	5	6	8	7	13	14	2	2	5	3	4	4	5	7	8	14	24	35	25	36	28	134				
Grief	12	13	16	7	17	14	11	17	10	13	12	13	4	18	20	22	11	17	15	20	71	50	48	59	43	47	279					
Hanged, and made away themselves	11	10	13	14	9	14	15	9	14	16	24	18	18	30	8	6	15	3	8	7	37	18	48	47	72	32	222					
Jaundice	57	35	39	49	41	43	57	71	61	41	46	77	102	70	47	59	35	43	33	45	54	63	184	197	180	212	225	188	998			
Jaw-faln-	1	1			3				2	2		3	1			10	16	13	8	10	4	11	47	35	02	5	6	10	95			
Impostume	75	61	65	59	80	105	79	90	92	122	80	134	105	96	58	76	73	74	50	62	73	130	283	315	260	354	428	228	1639			
Itch					1															10	00	10	01						11			
Killed by several Accidents	27	57	39	94	47	45	57	58	52	43	52	47	55	47	54	55	47	46	49	41	51	60	202	201	217	207	194	148	1021			
King's Evil	27	26	22	19	22	20	26	26	27	24	23	28	28	54	16	25	18	38	35	20	26	69	97	150	94	94	103	66	537			
Lethargy	3	4	3	4	4	3	10	9	4	6	2	6	4	2	2	2	2	2	2	5	7	13	21	21	9	67						
Leproy					1															2	2	2	1	1	3	3	06					
Liver-town Splicen, and Rakers	53	66	60	50	66	73	67	65	52	50	38	51	8	15	94	112	99	87	82	77	90	99	393	356	213	360	191	158	1421			

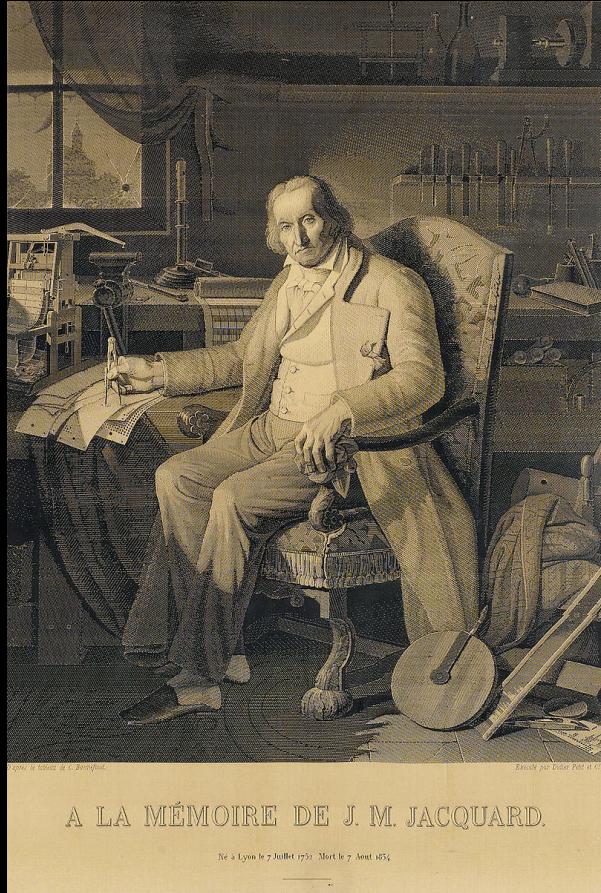
# Collecting Data

Herman Hollerith – 1890



# Collecting Data

## Jacquard Loom



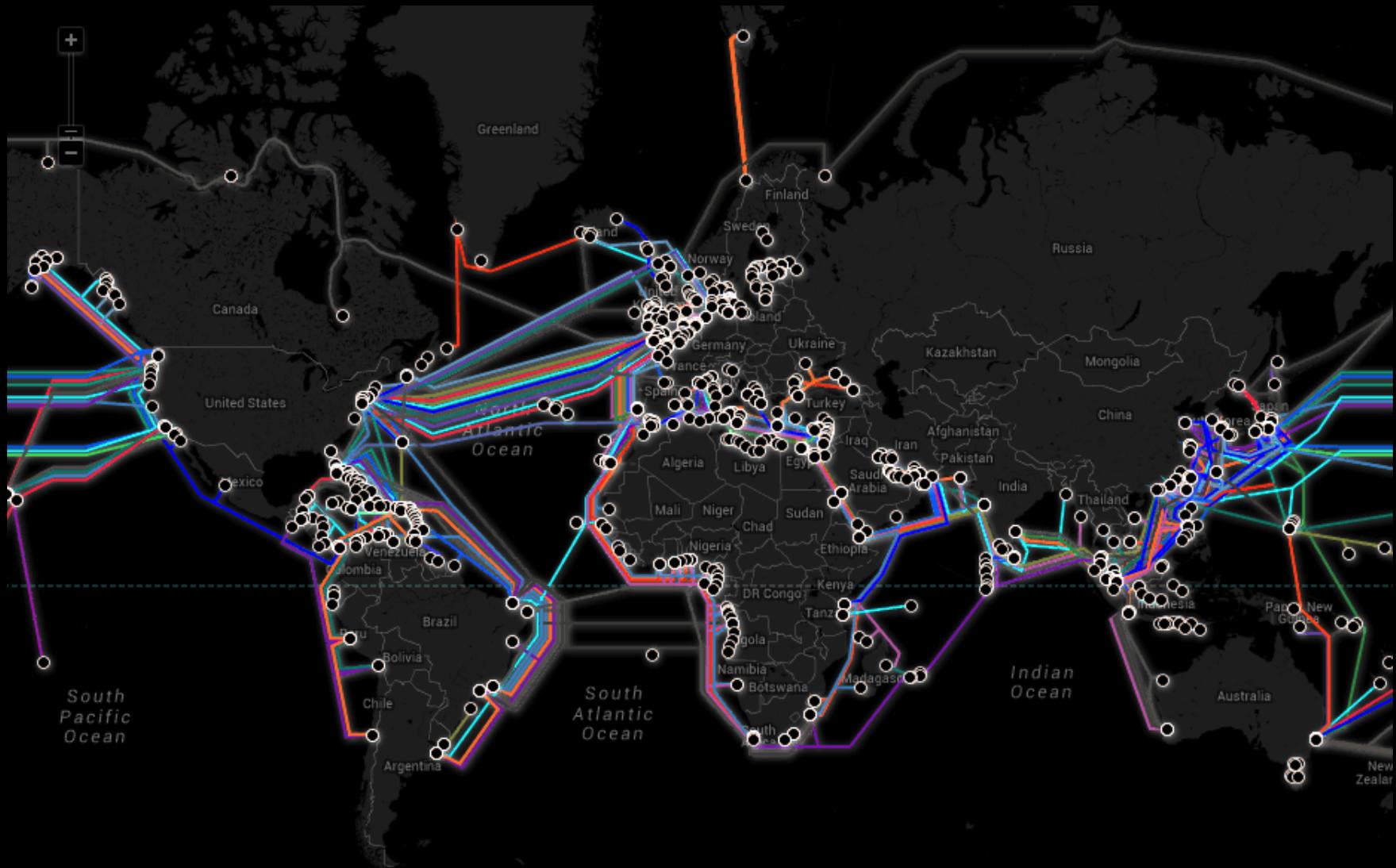
# Collecting Data

Herman Hollerith – 1890



# Collecting Data

The Internet – 1982



# Collecting Data

World Wide Web – 1989



# Collecting Data

How the Web Works



← http:// →



# Collecting Data

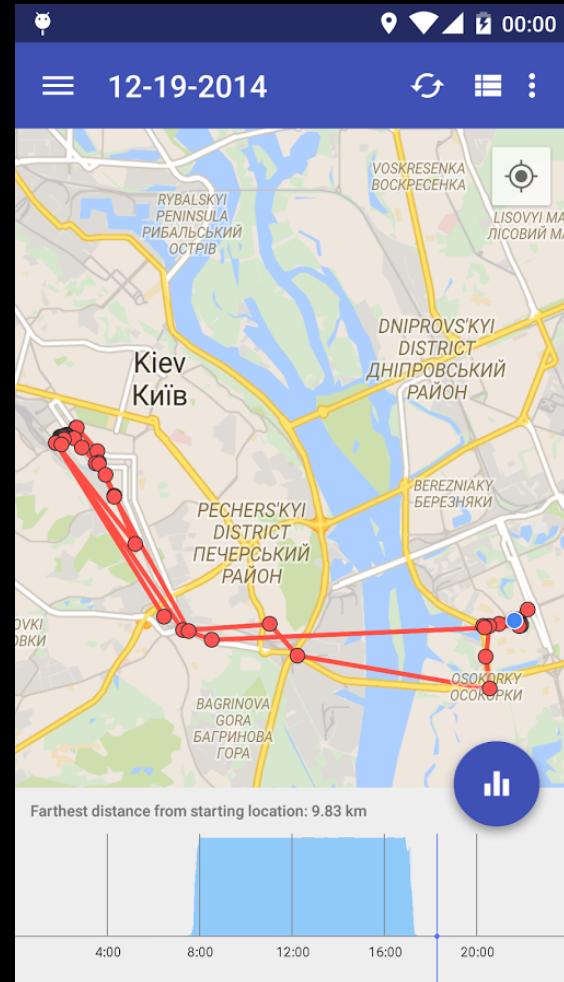
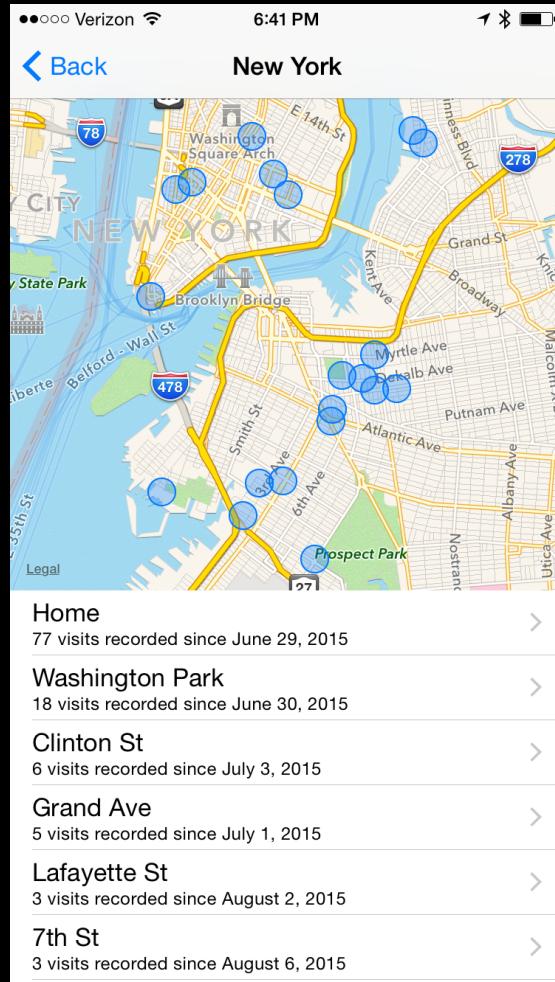
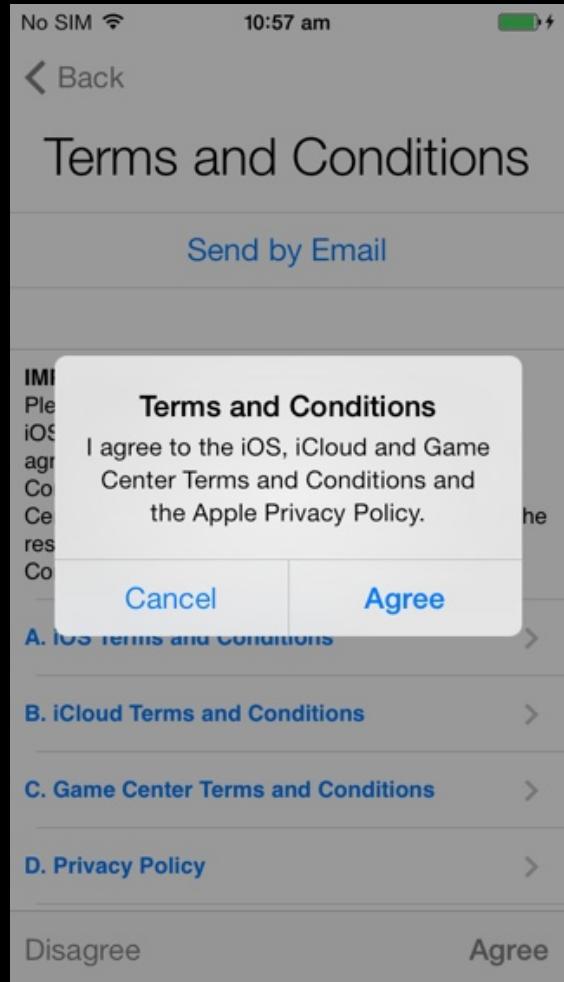
Every Company is a Data Company



**NETFLIX**

# Collecting Data

Every Company is a Data Company



# Collecting Data

## Take Back Your Data

A bite of Me  
by Federico Zannier

Home Updates 0 Backers 111 Comments 1

New York, NY Conceptual Art



Federico Zannier

111 backers  
\$1,271 pledged of \$500 goal  
22 days to go

Back This Project  
\$1 minimum pledge

This project will be funded on Wednesday Jun 5, 10:44am EDT.

Project by Federico Zannier Brooklyn, NY Contact me

263 people like this. Sign Up to see what your friends like.

I've data mined myself. I've violated my own privacy. Now I am selling it all. But how much am I worth?

Launched: May 6, 2013  
Funding ends: Jun 5, 2013

Remind me

First created 0 backed  
Federico Zannier 704 trends  
Website: myprivacy.info

# Collecting Data

Nicholas Felton



# Collecting Data

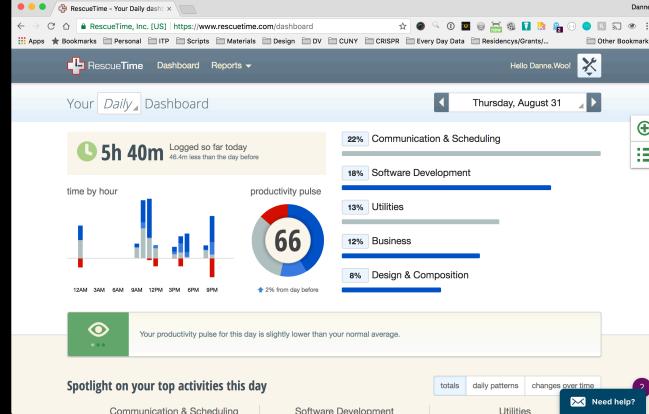
## Personal Data – Software

This screenshot shows the OpenPaths website interface. At the top, there's a header with links for Home, About, FAQ, Tools, Projects, My Data, Account, and Logout. A message at the top asks if Google Chrome should save the password. Below the header, the title "Data for dannewoo" is displayed, followed by a subtext "This is your personal data." On the left, there's a map with several red location markers and a "View my map" button. To the right of the map are download options for CSV, JSON, and KML. A section titled "OpenPaths API" contains fields for Access key (QBENBjKMCKYURWFKAJUNK) and Secret key (SHOW). Below this, a note explains that the dataset can be accessed programmatically through the API, which uses OAuth 2.0 authentication. It also states that anyone with these tokens will be able to access your data. There's a link to reset the secret key. At the bottom, there are "Requests" and "Connect and Import" buttons.

Open Paths



Feltron Reporter App



Rescue Time



Apple iPhone Health Sensors

# Collecting Data

## Personal Data – Hardware



Fitbit



Jawbone



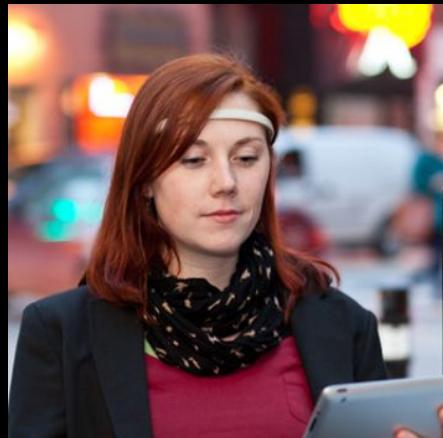
Nike Fuelband



Apple Watch

# Collecting Data

## Personal Data – Hardware



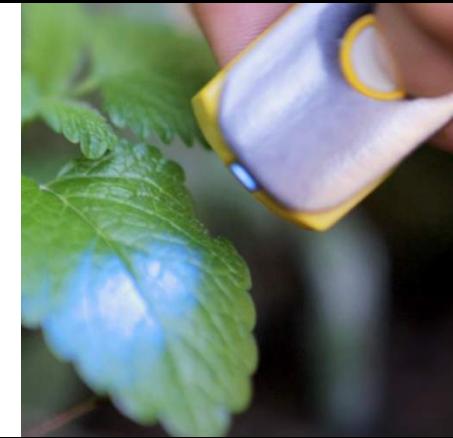
Muse Headband



Nest



Vessyl



Scio

# Collecting Data

The Connected Us



vocativ

Thursdays at 11pm 

# Collecting Data

Rodrigo Narciso – CH4



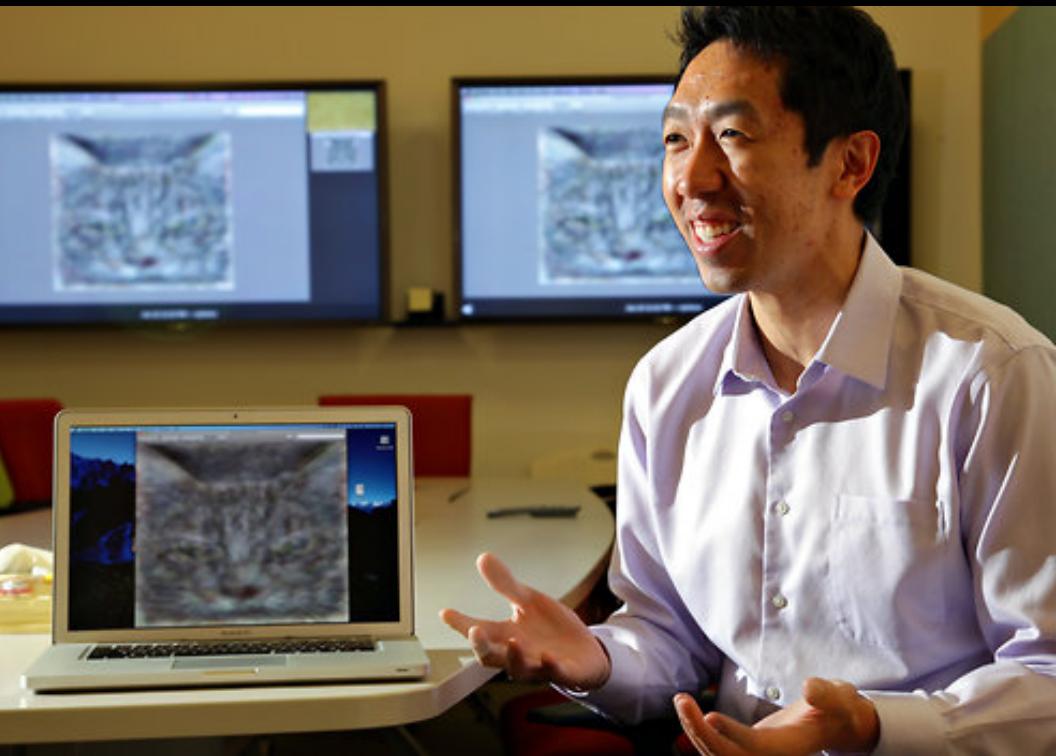
# Sourcing Data

The Internet (Not Just Cat Memes)



# Sourcing Data

But mostly cat memes



# Sourcing Data

## Open Datasets

The screenshot shows the homepage of the Guardian's Data blog. The main headline is "How socioeconomic disadvantage varies area-by-area in Australia". Below it is a map of Australia color-coded by socioeconomic disadvantage. To the right is a photo of a coach and the text "A new dawn: how a new coach can impact an NRL club". Below the map are three smaller images: a landscape, a rock, and a political donation network diagram. The footer includes links for "All today's stories" and "+ More DataBlog".

[theguardian.com/data](https://theguardian.com/data)

The screenshot shows the homepage of the US Census Bureau. The main feature is a map of the United States with a color key for median age. Below the map are sections for "POPULATION CLOCK", "QUICKFACTS", and "Did You Know?". The "QUICKFACTS" section highlights that 12.4% of people in Sioux City, Iowa are 65 years or older. The "Did You Know?" section notes that the nation's population has a distinctly older age profile than it did 15 years ago. Other sections include "U.S. Census Bureau Economic Indicators" and "Stat of the Day".

[census.gov](https://census.gov)

The screenshot shows the homepage of the FBI UCR. It features the FBI seal and the title "Uniform Crime Reporting". The page explains the history of the program, noting its origins in 1929 and its purpose to provide reliable uniform crime statistics for the nation. It also mentions the four annual publications: Crime in the United States, National Incident-Based Reporting System, Law Enforcement Officers Killed and Assaulted (LEOKA), and Hate Crime Statistics. The page includes a "Explore Our Services" section with links to publications like "Law Enforcement Officers Killed and Assaulted (LEOKA)" and "National Incident-Based Reporting System (NIBRS)". A "Latest Releases" section shows recent reports like "Preliminary Semimonthly UCR Report - February through June 2016".

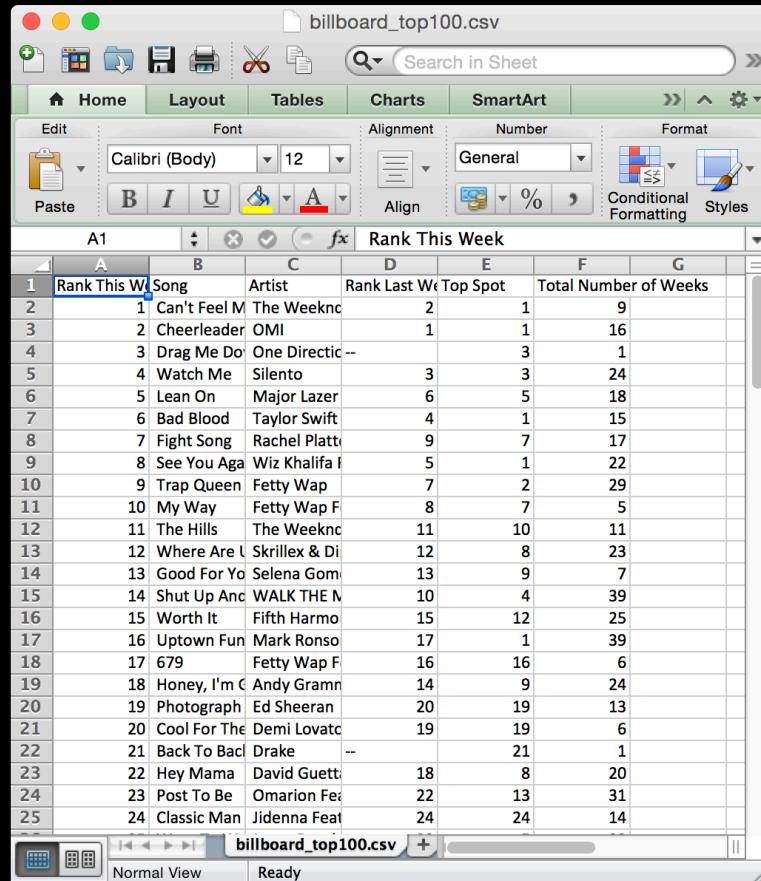
[ucr.fbi.gov](https://ucr.fbi.gov)

The screenshot shows the homepage of the World Bank Open Data. It features the World Bank logo and the title "World Bank Open Data". The page emphasizes "Free and open access to global development data". It includes a search bar, a browse menu, and a "Most Recent" section featuring news items like "New Partnership for Capacity Development in Household Surveys for Welfare Analysis". On the right, there is a "WHAT YOU CAN LEARN WITH OPEN DATA" section and a "Atlas of Sustainable Development Goals 2017" graphic.

[data.worldbank.org](https://data.worldbank.org)

# Sourcing Data

## Data Formats - Tabular



The screenshot shows a Microsoft Excel spreadsheet titled "billboard\_top100.csv". The data consists of 25 rows of song information, with columns labeled A through G. The columns represent: Rank This Week, Song, Artist, Rank Last Week, Top Spot, and Total Number of Weeks. The data includes songs like "Can't Feel Me" by M. The Weeknd at rank 1, "Cheerleader" by OMI at rank 2, and "Bad Blood" by Taylor Swift at rank 6.

	A	B	C	D	E	F	G
1	Rank This Wk	Song	Artist	Rank Last Wk	Top Spot	Total Number of Weeks	
2	1	Can't Feel Me	M. The Weeknd	2	1	9	
3	2	Cheerleader	OMI	1	1	16	
4	3	Drag Me Down	One Direction	--	3	1	
5	4	Watch Me	Silento	3	3	24	
6	5	Lean On	Major Lazer	6	5	18	
7	6	Bad Blood	Taylor Swift	4	1	15	
8	7	Fight Song	Rachel Platten	9	7	17	
9	8	See You Again	Wiz Khalifa ft. Charlie Puth	5	1	22	
10	9	Trap Queen	Fetty Wap	7	2	29	
11	10	My Way	Fetty Wap ft. Cardi B	8	7	5	
12	11	The Hills	The Weeknd	11	10	11	
13	12	Where Are U Now	Skrillex & Diplo ft. Justin Bieber	12	8	23	
14	13	Good For You	Selena Gomez	13	9	7	
15	14	Shut Up And Dance	Walk The Moon	10	4	39	
16	15	Worth It	Fifth Harmony	15	12	25	
17	16	Uptown Funk	Mark Ronson ft. Bruno Mars	17	1	39	
18	17	679	Fetty Wap ft. 21 Savage	16	16	6	
19	18	Honey, I'm Good	Andy Grammer	14	9	24	
20	19	Photograph	Ed Sheeran	20	19	13	
21	20	Cool For The Weekend	Demi Lovato	19	19	6	
22	21	Back To Back	Drake	--	21	1	
23	22	Hey Mama	David Guetta ft. Nicki Minaj	18	8	20	
24	23	Post To Be	Omarion ft. Chris Brown	22	13	31	
25	24	Classic Man	Jidenna ft. Burna Boy	24	24	14	

**CSV – Comma Separated Value**

**TSV – Tab Separated Value**

**XLS – Excel**

# Sourcing Data

## HTML

### Tables

Easily Copy and Paste

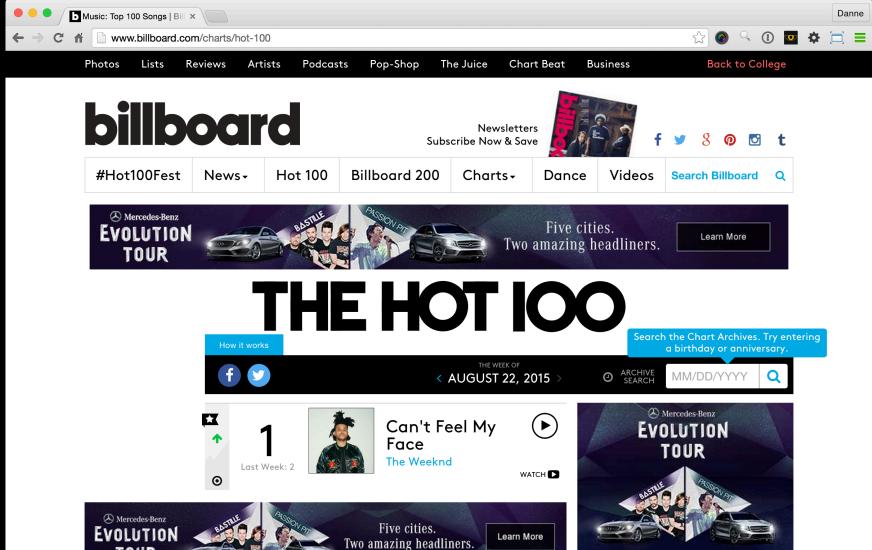
This screenshot shows a web browser displaying the 'Football Statistics' page from whoscored.com. The main content area is titled 'Team Statistics' and includes tabs for 'Summary', 'Defensive', 'Offensive', and 'Detailed'. Below these tabs is a 'View' dropdown with options for 'Overall', 'Home', and 'Away'. The main table lists 19 football teams along with their tournament, goals, shots per game, discipline, possession percentage, pass percentage, aerials won, and rating. To the right of the table are two charts: 'Best Form (Last 6 Matches)' showing winning streaks for Benfica, 1860 Muenchen, Bayern Munich, Barca, Tottenham, and Stockport; and 'Winning Streak (Longest Winning Pattern)' showing streaks for Benfica, Ludogorets Razgrad, Fiorentina, Nimes, and Sporting Gijon.

R Team	Tournament	Goals	Shots pg	Discipline	Possession%	Pass%	Aerials Won	Rating
1 Paris Saint Germain	French Ligue 1	96	16.9	65 6	62.8	89.6	9.4	7.21
2 Barcelona	La Liga	76	15.6	59 0	60.7	87.2	9.5	7.18
3 Bayern Munich	Bundesliga	76	18	59 2	61.9	87.2	16.8	7.16
4 Manchester City	Premier League	90	17.6	52 2	66.4	88.9	13.6	7.15
5 Real Madrid	La Liga	76	18.3	56 4	57.6	87.8	10.8	7.11
6 Juventus	Serie A	74	14.8	48 1	56.0	87.0	12	7.08
7 Manchester United	Premier League	63	13.4	57 1	53.1	82.8	16.9	7.01
8 Atletico Madrid	La Liga	50	10.9	72 3	48.4	80.3	14.9	7.01
9 Liverpool	Premier League	75	16.9	39 1	57.2	83.3	15.7	7.00
10 Tottenham	Premier League	64	17	42 2	58.8	83.9	16.4	6.98
11 Lyon	French Ligue 1	64	14.5	62 2	54.2	84.4	15.1	6.98
12 Monaco	French Ligue 1	77	12.6	59 1	51.0	81.2	16.7	6.98
13 Marseille	French Ligue 1	64	16	72 2	55.2	83.7	16	6.98
14 Lazio	Serie A	73	14.6	54 3	51.0	83.5	12.2	6.97
15 Napoli	Serie A	64	17.3	42 2	60.3	87.7	10.1	6.97
16 Roma	Serie A	50	18	39 2	56.4	83.5	17.2	6.96
17 Chelsea	Premier League	53	16.1	34 4	54.2	84.4	14.6	6.94
18 Inter	Serie A	50	15	55 0	54.4	85.7	14.4	6.94
19 Arsenal	Premier League	58	16.1	49 0	58.6	84.2	17.8	6.92

[whoscored.com/statistics](http://whoscored.com/statistics)

### Complex Layout

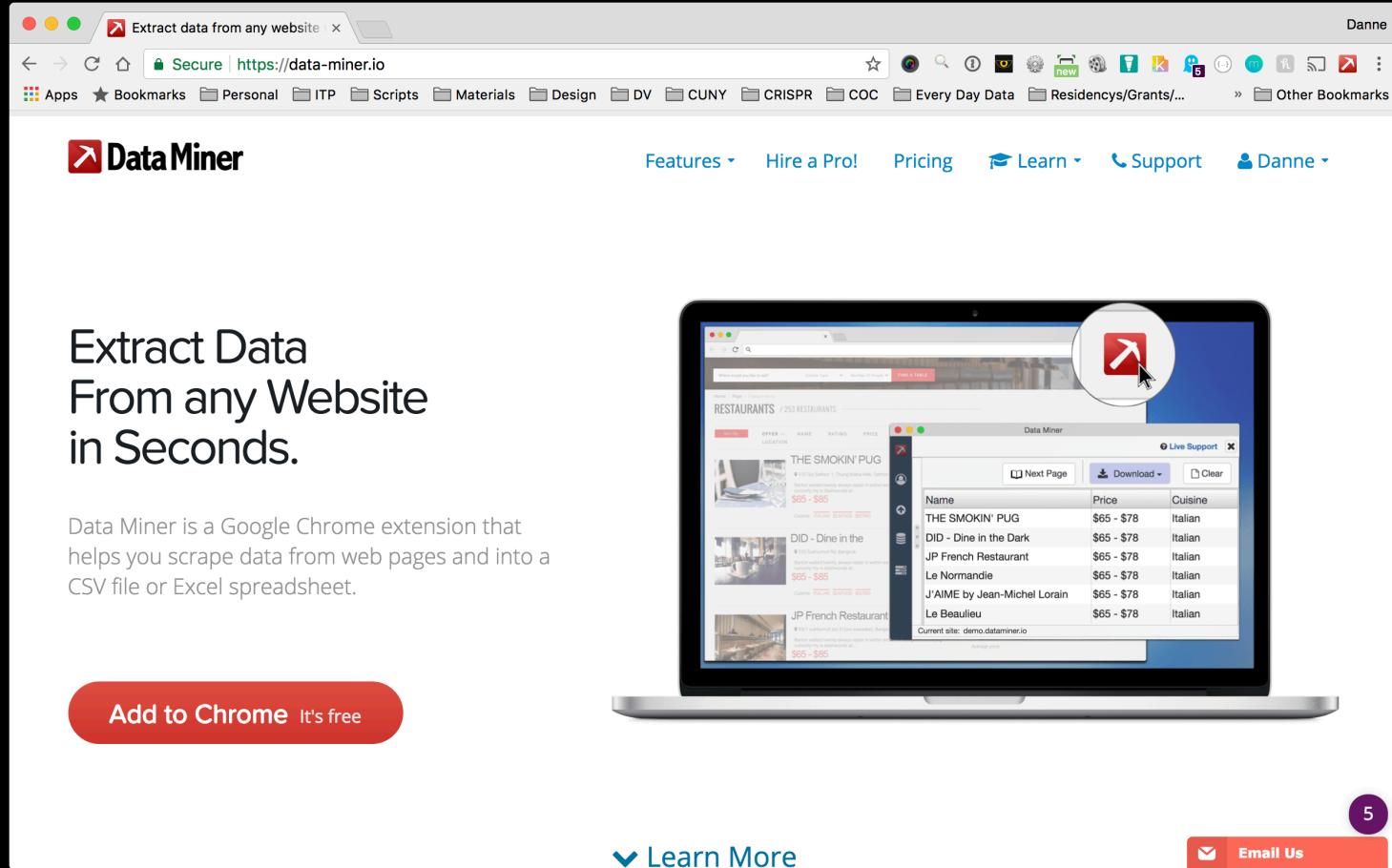
HTML Web Scrapping – Data Miner or Python



[billboard.com/charts/hot-100](http://billboard.com/charts/hot-100)

# Sourcing Data

## Web Scraping



The screenshot shows a Mac OS X desktop with a browser window open to <https://data-miner.io>. The browser's address bar shows "Extract data from any website". The Data Miner logo is at the top left. The main content area features a large heading "Extract Data From any Website in Seconds." and a description of the extension's purpose. Below this is a red "Add to Chrome" button. To the right, there's a visual representation of a laptop displaying a restaurant search page and the Data Miner extension interface, which shows a table of scraped data. A circular callout highlights the extension icon in the browser toolbar. At the bottom, there are "Learn More" and "Email Us" buttons, and a purple notification bubble with the number "5".

Extract data from any website

Secure | https://data-miner.io

Apps Bookmarks Personal ITP Scripts Materials Design DV CUNY CRISPR COC Every Day Data Residencys/Grants/... Other Bookmarks

Data Miner

Features Hire a Pro! Pricing Learn Support Danne

Extract Data From any Website in Seconds.

Data Miner is a Google Chrome extension that helps you scrape data from web pages and into a CSV file or Excel spreadsheet.

Add to Chrome It's free

Learn More Email Us

5

# Sourcing Data

## HTML based Data Tables (Tabula Free Software)



PDF  
Resilience saving lives today, investing for tomorrow

232

2006-2015.pdf

Home Tools 2006-2015.pdf 1 / 12 82.5% Danne

TABLE 1 Total number of reported disasters,<sup>1</sup> by continent and by year (2006-2015)

	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	Total <sup>3</sup>
Africa	202	184	173	156	135	165	124	115	101	116	1,471
Americas	105	133	143	115	146	131	114	106	116	124	1,233
Asia	308	262	240	233	253	235	210	228	228	240	2,437
Europe	98	104	58	75	99	49	91	69	80	70	793
Oceania	18	11	13	19	18	15	14	12	12	24	156
Very high human development <sup>2</sup>	124	118	104	99	113	94	119	123	112	113	1,119
High human development	202	224	193	183	213	159	165	155	192	183	1,869
Medium human development	199	158	170	154	161	165	126	130	126	146	1,535
Low human development	206	194	160	162	164	177	143	122	107	132	1,567
Total	731	694	627	598	651	595	553	530	537	574	6,090

Source: EM-DAT, CRED, University of Louvain, Belgium

<sup>1</sup> In Tables 1-13, 'disasters' refer to those with a natural and/or technological trigger only, and do not include wars, conflict-related famines, diseases or epidemics.

<sup>2</sup> See note on UNDP's Human Development Index country status in the disaster definitions section in the introduction to this annex.

<sup>3</sup> Since slow-onset disasters can affect the same country for a number of years, it is best to use figures on total numbers to calculate annual averages over a decade rather than as absolute totals (see the data definitions and methodology section in the introduction to this annex).

Note: Some totals in the table may not correspond due to rounding.

With 574 disasters reported, 2015 is the year of the decade with the fourth lowest number of disasters, very far below the peak of 2006.

Among continents, the number of disasters was the third lowest of the decade in Africa and the fourth lowest in Europe. In contrast, this number was the fourth highest of the decade in the Americas and in Asia; and Oceania suffered its highest number of disasters of the decade.

In 2015, the number of disasters was at its third lowest level in countries with low human development and at its fourth lowest in medium and high human development countries.

In countries with very high human development, the number of disasters was, in 2014, the fifth lowest in the decade.

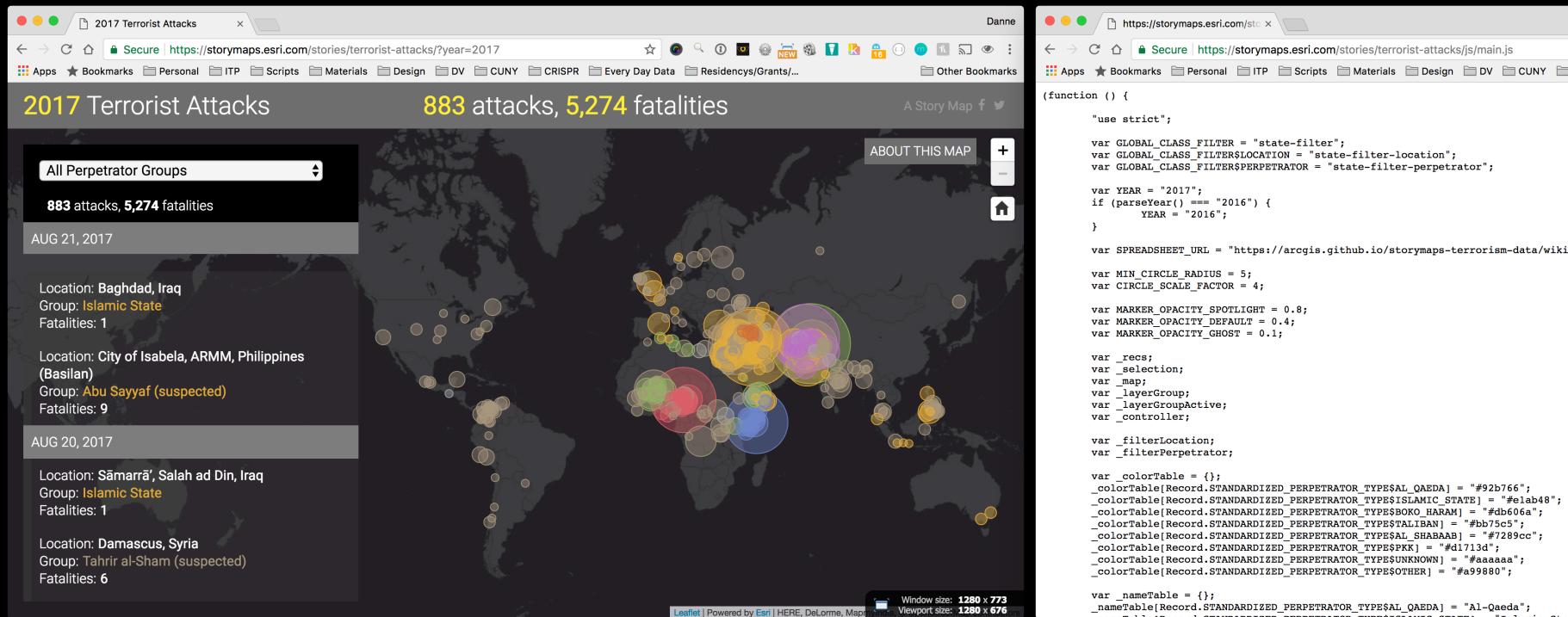
The distribution of disasters among continents is, in 2015, very similar to the annual average for the decade.

World Disasters Report 2016

ANNEEX Disaster data

# Sourcing Data

Interactive Data Visualizations (Look at the code, view source)



# Sourcing Data

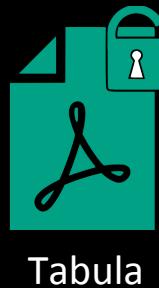
## Open Dataset Formats



Download file  
usually as a  
XLS, CSV, TSV,  
JSON or XML  
format



Select, copy  
and paste  
entire table  
into Excel or  
Google Sheets



Tabula



NO <table>



View the  
JavaScript  
source code  
and find any  
reference to a  
data file (CSV,  
JSON, XML)



Data Miner



Web Scrapping

# Sourcing Data

## APIs (Application Programming Interfaces)

The screenshot shows the New York Times Developers API documentation page. The main navigation bar includes links for Overview, Available APIs, Events, APIs (which is highlighted in green), Beta, Open Source, Blog, and Careers. The 'Available APIs' section lists several categories:

- THE ARTICLE SEARCH API**: Search Times articles from 1851 to today, retrieving headlines, abstracts and links to associated multimedia.
- THE BEST SELLERS API**: Get data from all New York Times best-seller lists, including rank history for specific best sellers.
- THE CAMPAIGN FINANCE API**: Get presidential campaign contribution and expenditure data based on United States Federal Election Commission filings.
- THE COMMUNITY API**: Get comments by NYTimes.com users.
- THE CONGRESS API**: Get U.S. Congressional vote data, including information about specific House and Senate members.
- THE DISTRICTS API**: Get political districts based on a pair of coordinates. Currently, the Districts API is limited to New York City.
- THE EVENT LISTINGS API**: Get information about hand-picked events in New York City and the surrounding area.
- THE GEOGRAPHIC API**: Use linked data to enhance location concepts used in The New York Times' controlled vocabulary.

developer.nytimes.com

The screenshot shows the Instagram Developer Documentation page. The left sidebar has a navigation menu with links like Overview, Authentication, Restrict API Requests, Real-time, Mobile Sharing, API Console, Endpoints, Limits, Embedding, Libraries, Support, and Platform Developers. The main content area features a large image of an iPhone with the Instagram logo, and the text "Hello Developers." Below it, a paragraph describes the purpose of the API. A call-to-action button says "Register Your Application then dive into the documentation". At the bottom, there's a "Getting Started" section with three numbered steps: Register, Authenticate, and Start making requests!

instagram.com/developer

The screenshot shows the OpenWeatherMap API documentation page. It features a header with the OpenWeatherMap logo and links for Home, Weather, Maps, API, Price, Technology, Stations, and News. Below the header, a section titled "Free weather API for developers" is shown. It includes sections for "Current weather data", "5 and 16-days forecast", and "Historical data". Each section contains brief descriptions and links to more details. The footer includes a "more" link for each section.

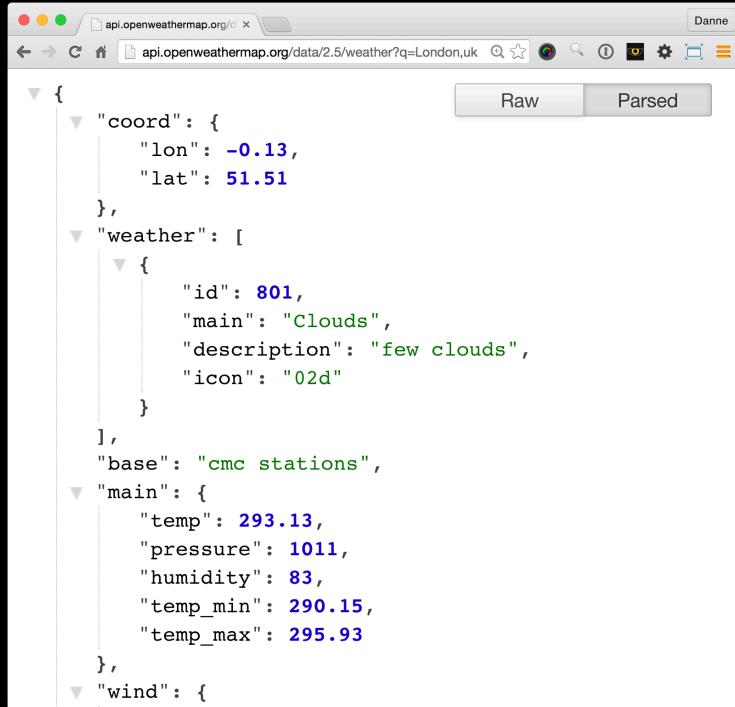
openweathermap.org/api

The screenshot shows the Twitter Developers flight conference landing page. The background features the Golden Gate Bridge. The main heading is "flight" with the subtitle "Twitter's First Annual Conference for mobile Developers". It specifies the date as "San Francisco | October 22nd". Below the main title, there are three calls to action: "Build Better Apps", "Monetize Your Apps", and "Tap Into Twitter Data". A "GNIP" logo is also present. In the bottom right corner, there's a "Tweets" box with a message about the OAuth / Auth issue.

dev.twitter.com

# Sourcing Data

## Data Formats - Nested



A screenshot of a browser window showing the JSON response from the OpenWeatherMap API. The URL in the address bar is `api.openweathermap.org/data/2.5/weather?q=London,uk`. The JSON data is displayed in a tree-view format with collapsible sections. The main object contains fields like "coord", "weather", "base", "main", and "wind". The "weather" field is expanded, showing an array of objects with "id", "main", "description", and "icon" properties. The "main" field is also expanded, showing "temp", "pressure", "humidity", "temp\_min", and "temp\_max" values.

```
{  
  "coord": {  
    "lon": -0.13,  
    "lat": 51.51  
  },  
  "weather": [  
    {  
      "id": 801,  
      "main": "Clouds",  
      "description": "few clouds",  
      "icon": "02d"  
    }  
  ],  
  "base": "cmc stations",  
  "main": {  
    "temp": 293.13,  
    "pressure": 1011,  
    "humidity": 83,  
    "temp_min": 290.15,  
    "temp_max": 295.93  
  },  
  "wind": {  
    "speed": 4.6,  
    "deg": 180,  
    "gusts": null  
  }  
}
```

**JSON – JavaScript Object Notation**  
**XML – Extensible Markup Language**

# Sourcing Data

APIs (Application Programming Interfaces)

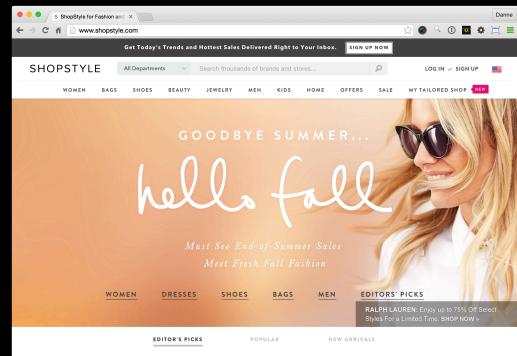
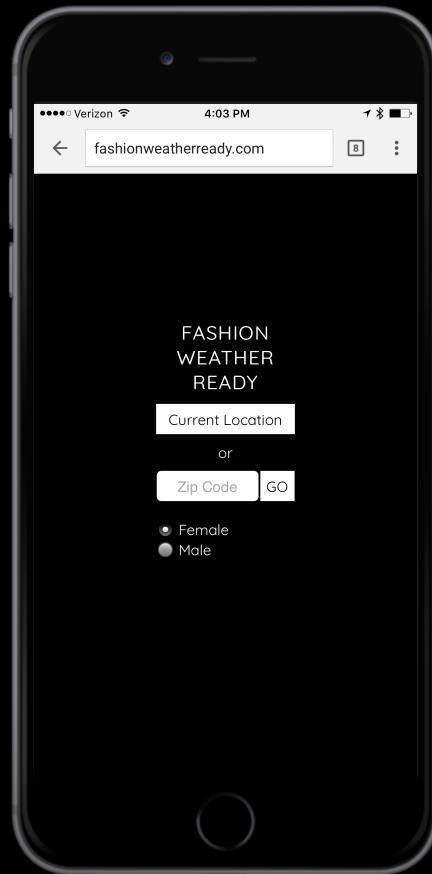


10,000+ Apps  
Using Foursquares  
Location Data API



# Sourcing Data

## APIs (Application Programming Interfaces)



fashionweatherready.com

# Big Data

John Mashey – 1998

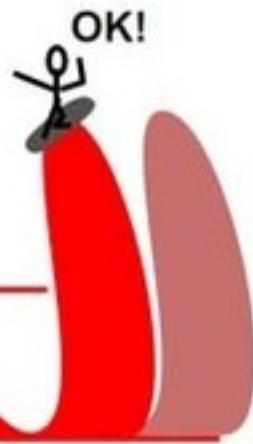
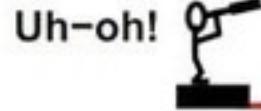
The SGI logo, consisting of the lowercase letters "sgi" in a bold, black, sans-serif font.

## *Big Data ... and the Next Wave of **InfraStress***

**John R. Mashey  
Chief Scientist, SGI**

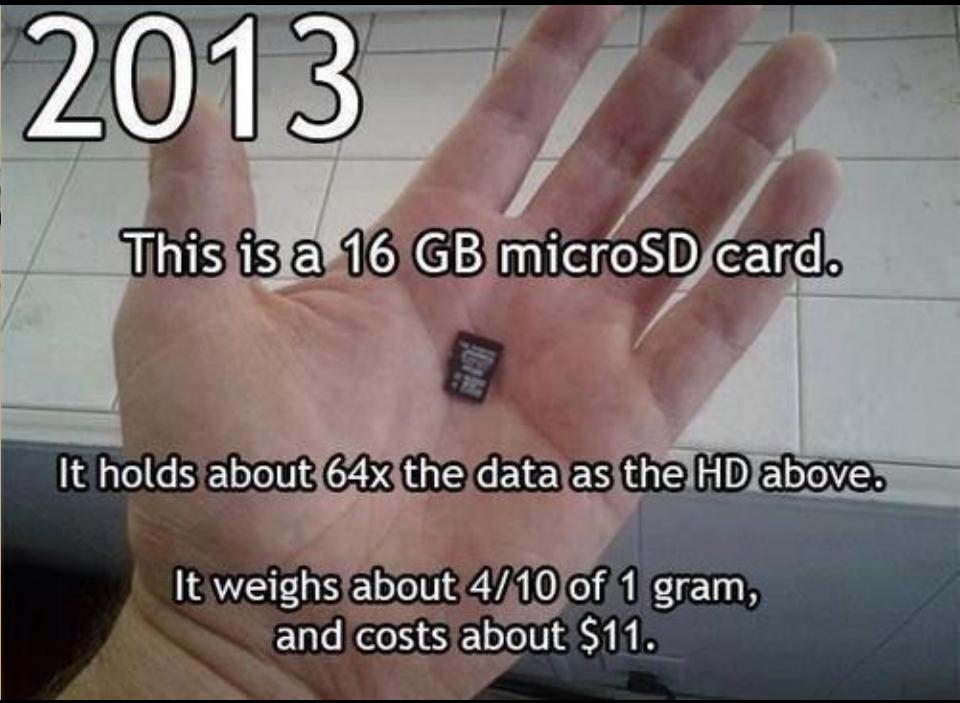
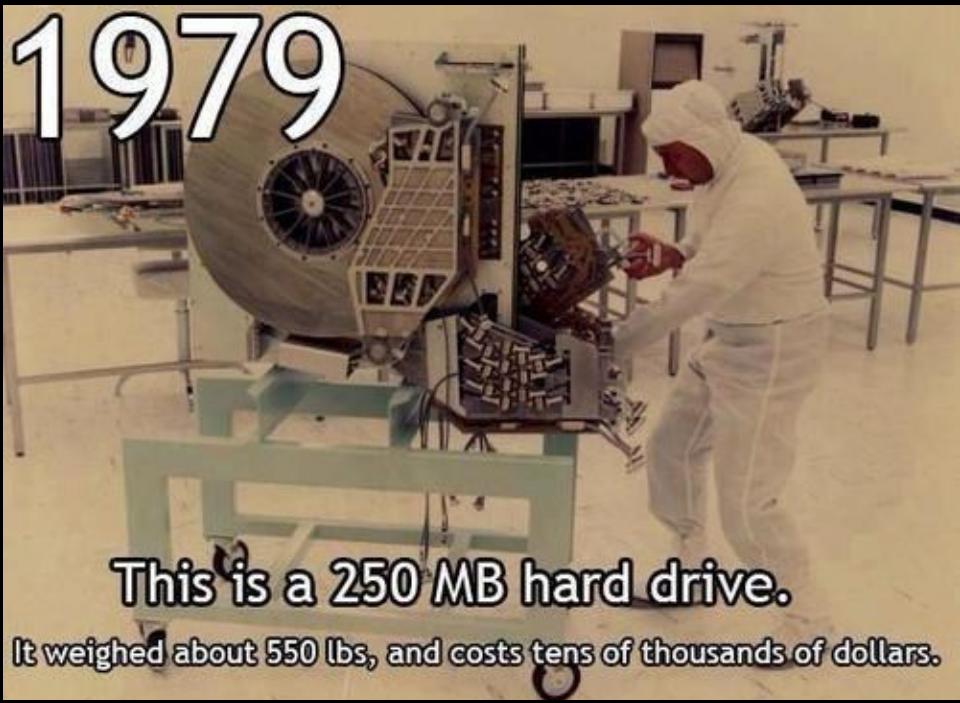


*Technology Waves:  
NOT technology for technology's sake  
**IT'S WHAT YOU DO WITH IT**  
But if you don't understand the trends  
**IT'S WHAT IT WILL DO TO YOU***



# Big Data

More Storage Capabilities

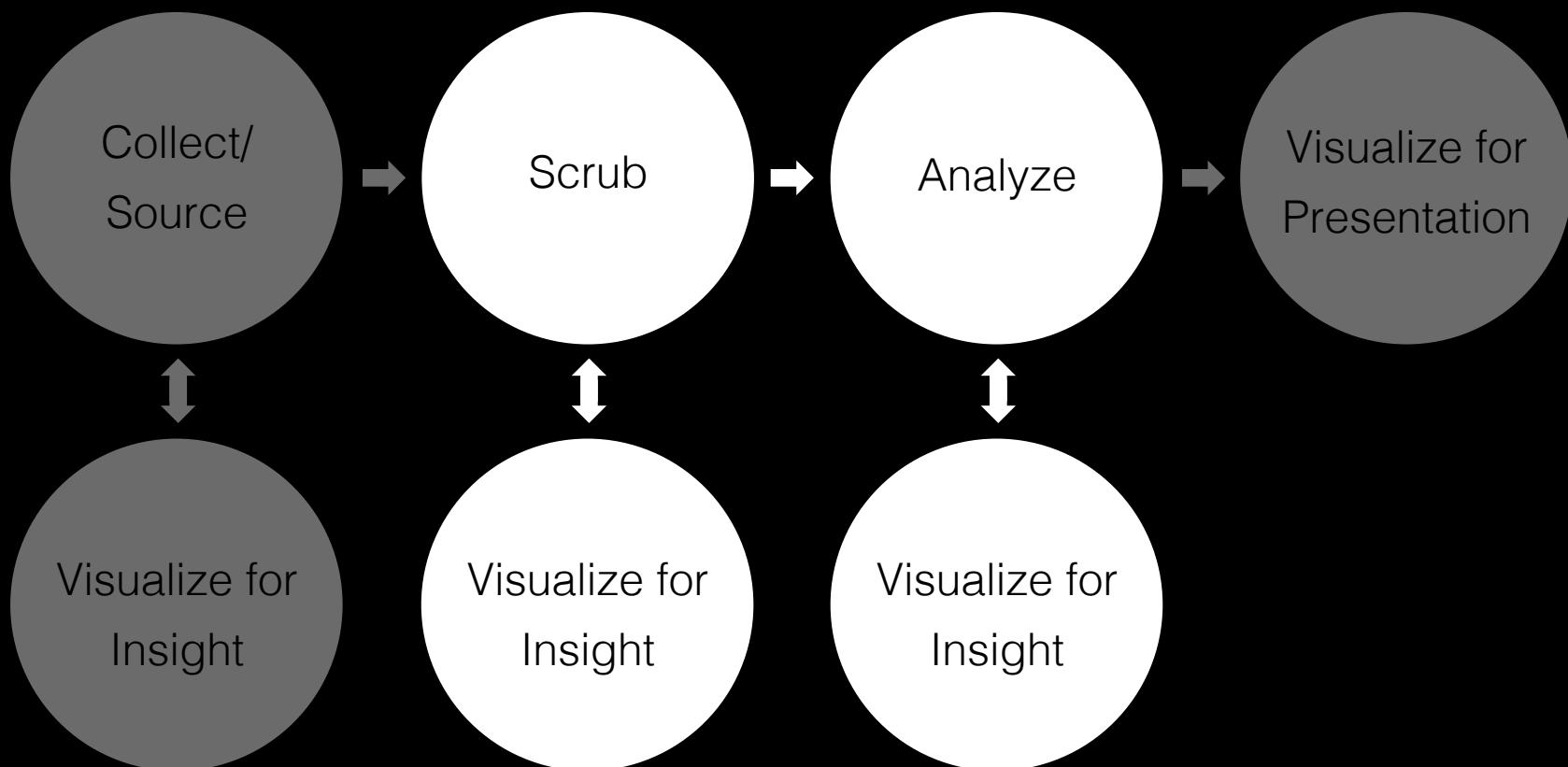


# Big Data

Needs more than one computer



# Data Pipeline



# Data Scrubbing

Date	Country	Address	City	Min Value	Max Value
4/23/2015	United States	3245 First Street	Los Angeles	4	9
4/24/2015		28 Orange Ave	Los Angeles	2	7
3/18/2014		763 Main St	Los Angeles	5	9
2/6/2015		72 Second Ave	Los Angeles	6	9
5/27/2014		87 Thurston Street	Los Angles	3	7
6/30/2013		652 Greene Street	New York	4	6
5/7/2014		887 Clinton Ave	Nuw York	3	8
3/3/2015		554 Washington Street	New York	3	9
12/23/2015		92 Waverly Ave	New York	2	5
2015/2/24	Germany	Am Falkpl. 5	Berlin	1	7
2014/3/26		Mehringdamm 32	Berlin	4	6
2014/6/8		Röntgenstraße 7	Berlin	1	8
2014/11/12		Rudi-Dutschke-Straße 26	Berlin	2	6

# Data Scrubbing

80% Time Spent

Combining Data Files

Data Formatting

Multiple Date Formats

Missing Data

Spaces Replaced with Strange Characters

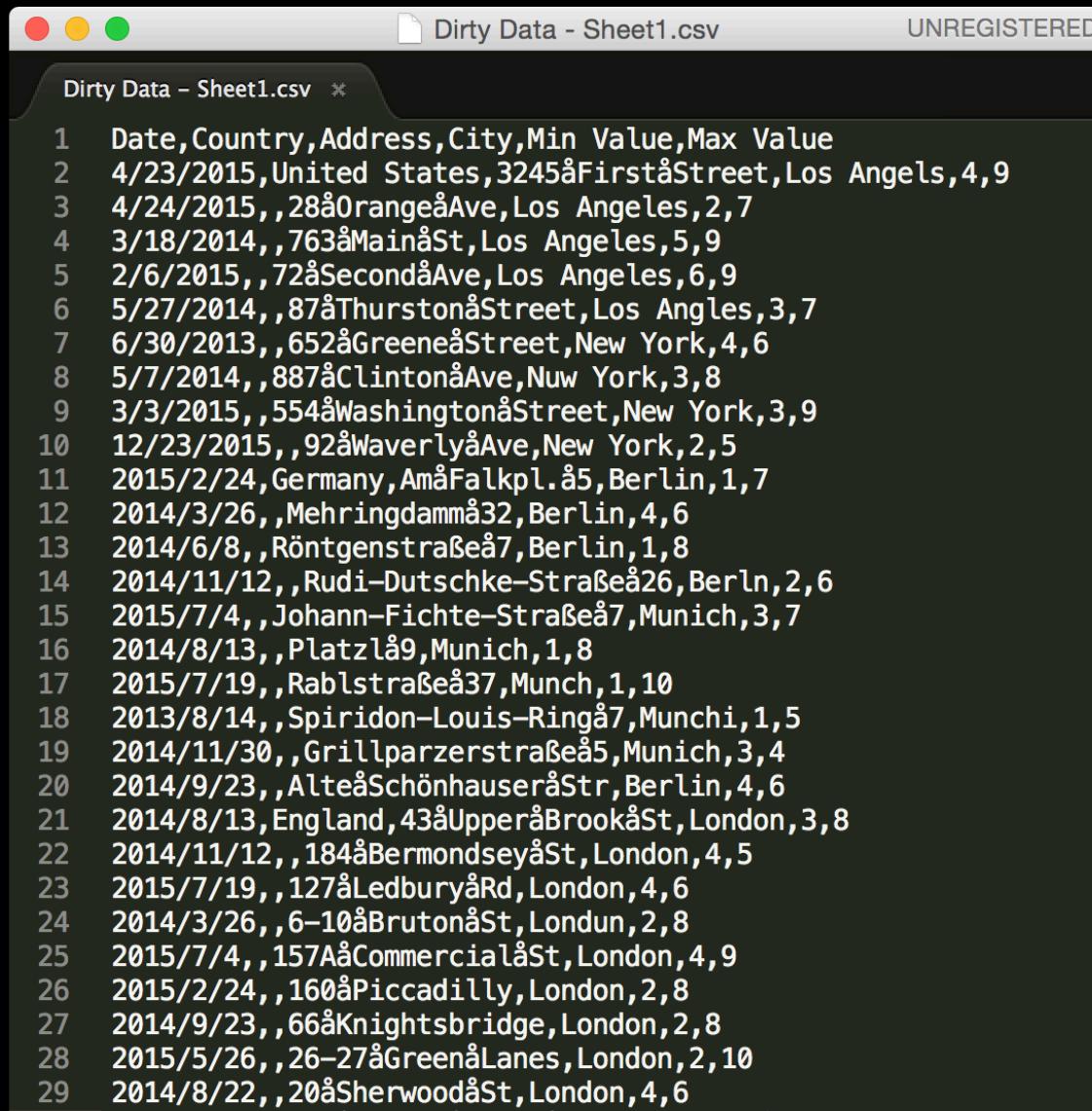
Random HTML Code

Spelling Errors

Etc.

# Data Scrubbing

Text Editor (Sublime Text)



The screenshot shows a Sublime Text window titled "Dirty Data - Sheet1.csv". The status bar indicates the file is "UNREGISTERED". The content of the file is a CSV dataset with 29 rows. The columns are labeled: Date, Country, Address, City, Min Value, and Max Value. The data contains numerous errors, such as missing commas, extra characters like 'å', and misspellings like "Los Angels" instead of "Los Angeles".

	Date	Country	Address	City	Min Value	Max Value
1	4/23/2015	United States	3245 First Street	Los Angels	4	9
2	4/24/2015	,	28 Orange Ave	Los Angeles	2	7
3	3/18/2014	,	763 Main St	Los Angeles	5	9
4	2/6/2015	,	72 Second Ave	Los Angeles	6	9
5	5/27/2014	,	87 Thurston Street	Los Angles	3	7
6	6/30/2013	,	652 Greene Street	New York	4	6
7	5/7/2014	,	887 Clinton Ave	Nuw York	3	8
8	3/3/2015	,	554 Washington Street	New York	3	9
9	12/23/2015	,	92 Waverly Ave	New York	2	5
10	2015/2/24	Germany	Am Falkpl. 5	Berlin	1	7
11	2014/3/26	,	Mehringdamm 32	Berlin	4	6
12	2014/6/8	,	Röntgenstraße 7	Berlin	1	8
13	2014/11/12	,	Rudi-Dutschke-Straße 26	Berln	2	6
14	2015/7/4	,	Johann-Fichte-Straße 7	Munich	3	7
15	2014/8/13	,	Platzl 9	Munich	1	8
16	2015/7/19	,	Rablstraße 37	Munch	1	10
17	2013/8/14	,	Spiridon-Louis-Ring 7	Munchi	1	5
18	2014/11/30	,	Grillparzerstraße 5	Munich	3	4
19	2014/9/23	,	Alte Schönhauser Str	Berlin	4	6
20	2014/8/13	England	43 Upper Brook St	London	3	8
21	2014/11/12	,	184 Bermondsey St	London	4	5
22	2015/7/19	,	127 Ledbury Rd	London	4	6
23	2014/3/26	,	6-10 Bruton St	Londun	2	8
24	2015/7/4	,	157A Commercial St	London	4	9
25	2015/2/24	,	160 Piccadilly	London	2	8
26	2014/9/23	,	66 Knightsbridge	London	2	8
27	2015/5/26	,	26-27 Green Lanes	London	2	10
28	2014/8/22	,	20 Sherwood St	London	4	6
29						

# Data Scrubbing

Google Sheets or Excel

The screenshot shows a Google Sheets document titled "Dirty Data". The spreadsheet contains 24 rows of data with columns A through G. Column A lists dates from 2013 to 2015. Column B lists countries (United States, Germany, England). Column C lists addresses with various encoding errors. Column D lists cities. Columns E and F contain numerical values (Min Value and Max Value). Column G is empty. The browser address bar shows the URL of the Google Sheets document.

	A	B	C	D	E	F	G
1	Date	Country	Address	City	Min Value	Max Value	
2	4/23/2015	United States	3245åFirståStreet	Los Angels	4	9	
3	4/24/2015		28åOrangeåAve	Los Angeles	2	7	
4	3/18/2014		763åMainåSt	Los Angeles	5	9	
5	2/6/2015		72åSecondåAve	Los Angeles	6	9	
6	5/27/2014		87åThurstonåStr	Los Angles	3	7	
7	6/30/2013		652åGreeneåStr	New York	4	6	
8	5/7/2014		887åClintonåAve	Nuw York	3	8	
9	3/3/2015		554åWashington	New York	3	9	
10	12/23/2015		92åWaverlyåAve	New York	2	5	
11	2015/2/24	Germany	AmåFalkpl.å5	Berlin	1	7	
12	2014/3/26		Mehringdammå3	Berlin	4	6	
13	2014/6/8		Röntgenstraßeå	Berlin	1	8	
14	2014/11/12		Rudi-Dutschke-S	Berlin	2	6	
15	2015/7/4		Johann-Fichte-S	Munich	3	7	
16	2014/8/13		Platzlå9	Munich	1	8	
17	2015/7/19		Rablstraßeå37	Munch	1	10	
18	2013/8/14		Spiridon-Louis-R	Munchi	1	5	
19	2014/11/30		Grillparzerstraße	Munich	3	4	
20	2014/9/23		AlteåSchönhaus	Berlin	4	6	
21	2014/8/13	England	43åUpperåBrook	London	3	8	
22	2014/11/12		184åBermondse	London	4	5	
23	2015/7/19		127åLedburyåR	London	4	6	
24	2014/3/26		6-10åBrutonåSt	Londun	2	8	

# Data Scrubbing

## Open Refine

Dirty Data Sheet1 csv - Open

localhost:3333/project?project=2265808547941

Refine OPEN Dirty Data Sheet1 csv Permalink

Facet / Filter Undo / Redo 0

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?  
[Watch these screencasts](#)

29 rows

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 29 next > last »

All	Date	Country	Address	City	Min Value	Max Value
1.	4/23/2015	United States	3245 First Street	Los Angeles	4	9
2.	4/24/2015		28 Orange Ave	Los Angeles	2	7
3.	3/18/2014		763 Main St	Los Angeles	5	9
4.	2/6/2015		72 Second Ave	Los Angeles	6	9
5.	5/27/2014		87 Thurston Street	Los Angles	3	7
6.	6/30/2013		652 Greene Street	New York	4	6
7.	5/7/2014		887 Clinton Ave	Nuw York	3	8
8.	3/3/2015		554 Washington Street	New York	3	9
9.	12/23/2015		92 Waverly Ave	New York	2	5
10.	2015/2/24	Germany	Am Falkpl. 5	Berlin	1	7
11.	2014/3/26		Mehringdamm 32	Berlin	4	6
12.	2014/6/8		Rontgenstra e 7	Berlin	1	8
13.	2014/11/12		Rudi-Dutschke-Stra e 26	Berlin	2	6
14.	2015/7/4		Johann-Fichte-Stra e 7	Munich	3	7
15.	2014/8/13		Platz 9	Munich	1	8
16.	2015/7/19		Rablstra e 37	Munch	1	10
17.	2013/8/14		Spiridon-Louis-Ring 7	Munchi	1	5
18.	2014/11/30		Grillparzerstra e 5	Munich	3	4
19.	2014/9/23		Alte Sch nhauser Str	Berlin	4	6
20.	2014/8/13	England	43 Upper Brook St	London	3	8
21.	2014/11/12		184 Bermondsey St	London	4	5
22.	2015/7/19		127 Ledbury Rd	London	4	6
23.	2014/3/26		6-10 Bruton St	Londun	2	8
24.	2015/7/4		157A Commercial St	London	4	9
25.	2015/2/24		160 Piccadilly	London	2	8
26.	2014/9/23		66 Knightsbridge	London	2	8

# Data Scrubbing

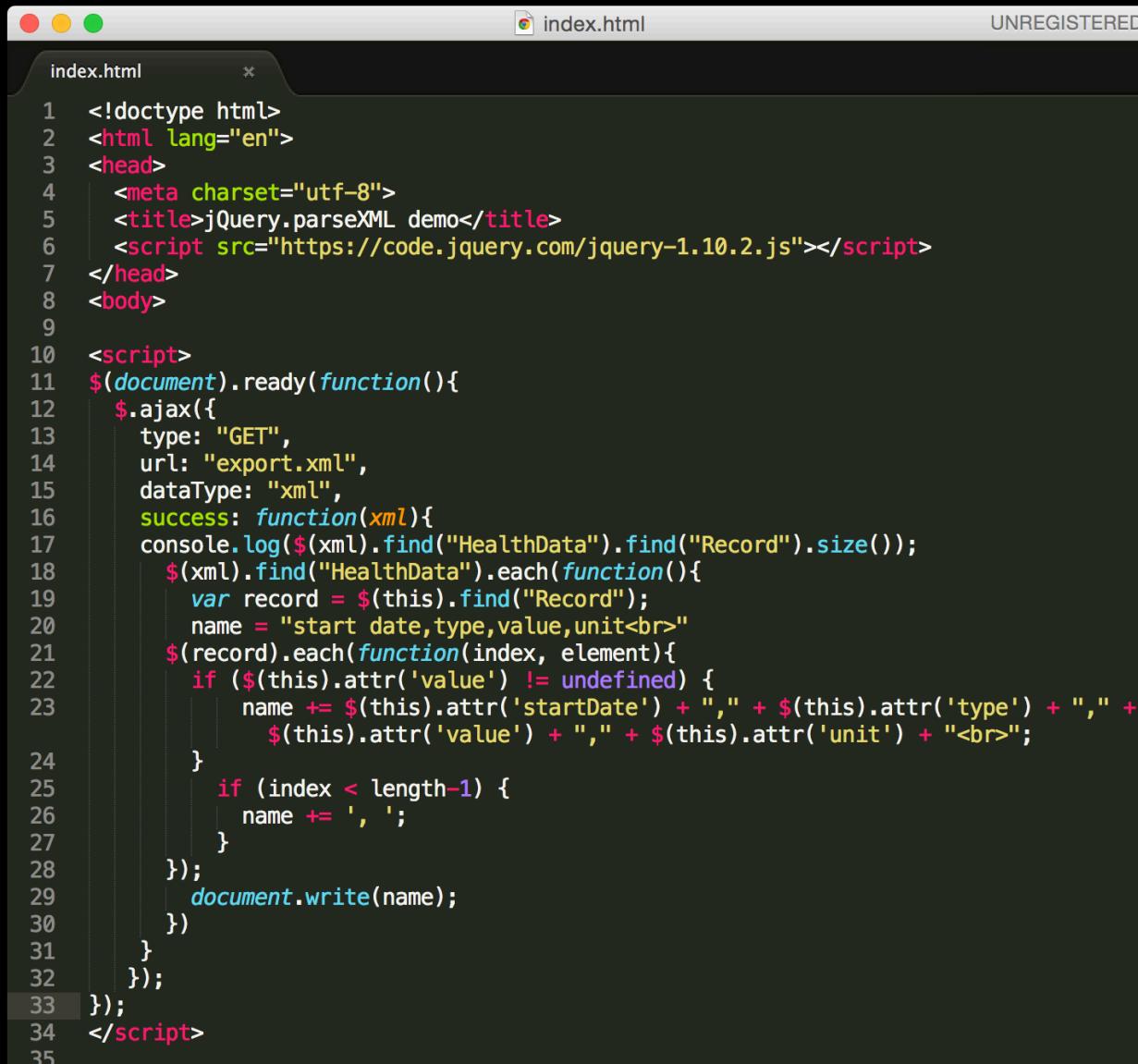
Converting XML > CSV



```
export.xml UNREGISTERED
1  <?xml version="1.0" encoding="UTF-8"?>
2  <HealthData locale="en_US">
3  <ExportDate value="20150814122840-0400"/>
4  <Me HKCharacteristicTypeIdentifierDateOfBirth="19820630" HKCharacteristicTypeIdentifierBiologicalSex="2"
HKCharacteristicTypeIdentifierBloodType="0"/>
5  <Record type="HKQuantityTypeIdentifierHeight" source="Danne's Apple Watch" unit="m" creationDate="20150630182606-0400" startDate="20150630182606-0400" endDate="20150630182606-0400" value="1.778"/>
6  <Record type="HKQuantityTypeIdentifierBodyMass" source="Danne's Apple Watch" unit="kg" creationDate="20150630182606-0400" startDate="20150630182606-0400" endDate="20150630182606-0400" value="65.7709"/>
7  <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150630122800-0400" endDate="20150701122800-0400" min="0.95" max="1.85" average="1.17278" recordCount="30"/>
8  <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150701122800-0400" endDate="20150702122800-0400" min="1.01667" max="1.55" average="1.17302" recordCount="21"/>
9  <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150702122800-0400" endDate="20150703122800-0400" min="0.9" max="1.55" average="1.17742" recordCount="31"/>
10 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150703122800-0400" endDate="20150704122800-0400" min="0.766667" max="1.61667" average="1.15045" recordCount="37"/>
11 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150704122800-0400" endDate="20150705122800-0400" min="0.9" max="1.88333" average="1.19608" recordCount="51"/>
12 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150705122800-0400" endDate="20150706122800-0400" min="0.816667" max="1.4" average="1.13947" recordCount="19"/>
13 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150706122800-0400" endDate="20150707122800-0400" min="1.01667" max="1.61667" average="1.16818" recordCount="22"/>
14 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150707122800-0400" endDate="20150708122800-0400" min="1.01667" max="1.53333" average="1.21759" recordCount="36"/>
15 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150708122800-0400" endDate="20150709122800-0400" min="0.9" max="1.66667" average="1.14167" recordCount="28"/>
16 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150709122800-0400" endDate="20150710122800-0400" min="1.01667" max="1.61667" average="1.25873" recordCount="21"/>
17 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150710122800-0400" endDate="20150711122800-0400" min="0.766667" max="1.78333" average="1.22593" recordCount="36"/>
18 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150711122800-0400" endDate="20150712122800-0400" min="1.01667" max="1.23333" average="1.11" recordCount="5"/>
19 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150712122800-0400" endDate="20150713122800-0400" min="1.2" max="1.53333" average="1.34583" recordCount="8"/>
20 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150713122800-0400" endDate="20150714122800-0400" min="1.01667" max="1.66667" average="1.22467" recordCount="25"/>
21 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150714122800-0400" endDate="20150715122800-0400" min="0.68333" max="1.58333" average="1.25" recordCount="15"/>
22 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150715122800-0400" endDate="20150716122800-0400" min="1.03333" max="1.78333" average="1.3375" recordCount="8"/>
23 <Record type="HKQuantityTypeIdentifierHeartRate" source="Danne's Apple Watch" unit="count/s" startDate="20150716122800-0400" endDate="20150717122800-0400" min="1.03333" max="1.78333" average="1.3375" recordCount="8"/>
```

# Data Scrubbing

Python or JavaScript



The screenshot shows a browser window with the title "index.html" and status bar "UNREGISTERED". The content area displays the source code of an HTML file. The code includes an XML parsing script using jQuery. It performs an AJAX GET request to "export.xml", parses the XML, and then iterates through the "Record" elements to extract "start date", "type", "value", and "unit" attributes, concatenating them into a single string separated by commas and line breaks, and finally writing this string to the document.

```
index.html
1  <!doctype html>
2  <html lang="en">
3  <head>
4      <meta charset="utf-8">
5      <title>jQuery.parseXML demo</title>
6      <script src="https://code.jquery.com/jquery-1.10.2.js"></script>
7  </head>
8  <body>
9
10 <script>
11 $(document).ready(function(){
12     $.ajax({
13         type: "GET",
14         url: "export.xml",
15         dataType: "xml",
16         success: function(xml){
17             console.log($(xml).find("HealthData").find("Record").size());
18             $(xml).find("HealthData").each(function(){
19                 var record = $(this).find("Record");
20                 name = "start date,type,value,unit<br>"
21                 $(record).each(function(index, element){
22                     if ($(this).attr('value') != undefined) {
23                         name += $(this).attr('startDate') + "," + $(this).attr('type') + "," +
24                             $(this).attr('value') + "," + $(this).attr('unit') + "<br>";
25                     }
26                     if (index < length-1) {
27                         name += ', ';
28                     }
29                 });
30                 document.write(name);
31             })
32         });
33     });
34 </script>
35 
```

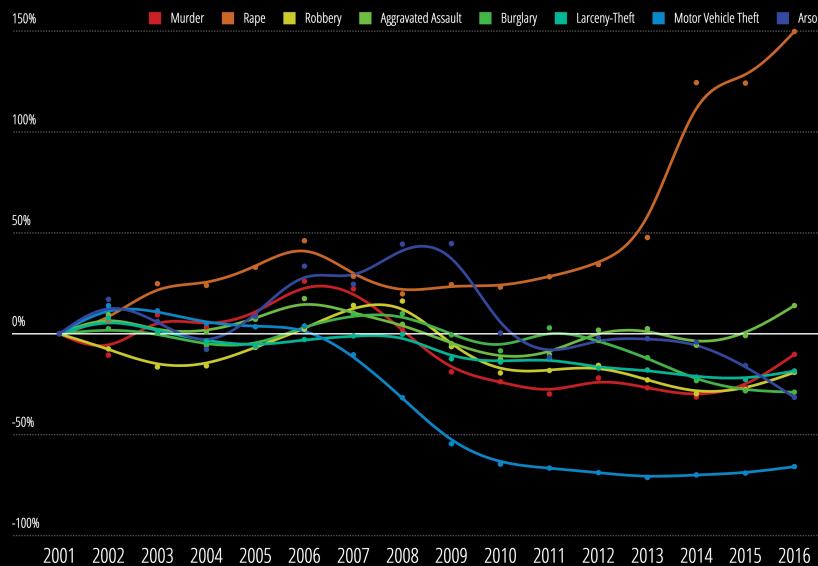
# Data Scrubbing

## Combining Datasets

### Sheriff Joe's Crime Record

Last Friday Trump issued his first pardon for Joe Arpaio, the controversial former sheriff of Maricopa County, Arizona. Sheriff Joe first gained national recognition for his participation in the debunked Obama birther conspiracy and has since been criticized for his policing methods considered by many to be racist. After Joe refused to follow a court order to stop his racial profiling practices, he was found guilty of contempt. One aspect that makes this pardon unusual is that it happened before his sentencing.

The chart below shows the yearly change in Sheriff Joe's violent crime record since 2001. While most violent crimes either went down or stayed the same, rape cases skyrocketed.



Source: [azps.gov/about/reports/crime](http://azps.gov/about/reports/crime) and [city-data.com/crime/crime-Arizona.html](http://city-data.com/crime/crime-Arizona.html)



2013  
CRIME  
IN  
ARIZONA  
REPORT

2012  
CRIME  
IN  
ARIZONA  
REPORT

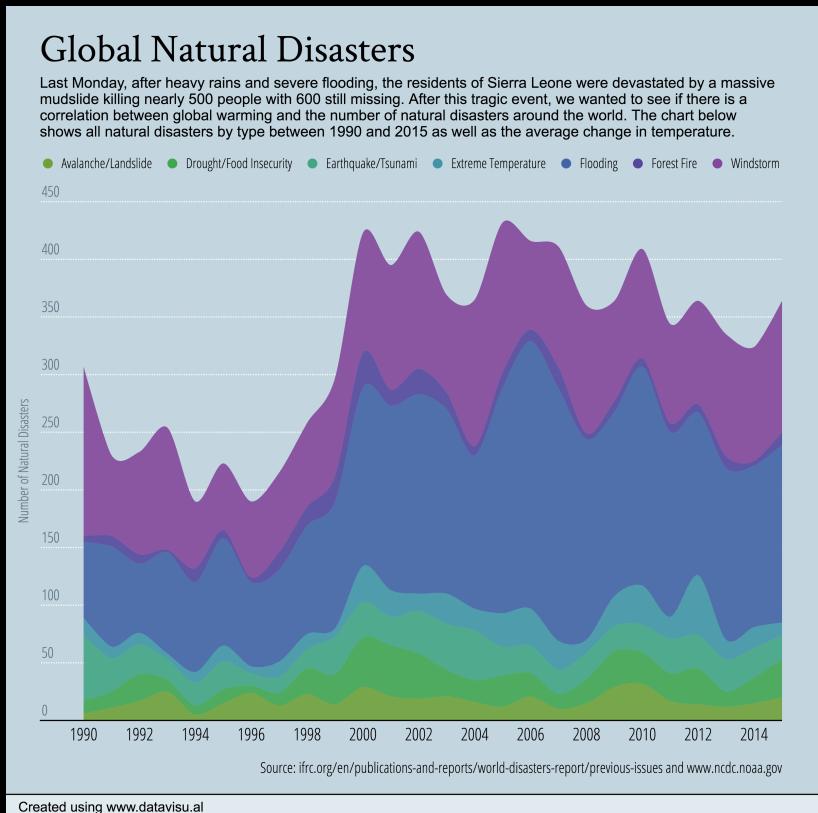
2011  
CRIME  
IN  
ARIZONA  
REPORT

2011  
CRIME  
IN  
ARIZONA  
REPORT

2010  
CRIME  
IN  
ARIZONA  
REPORT

# Data Scrubbing

## Reformatting Datasets



natural-disasters-type-2.xlsx

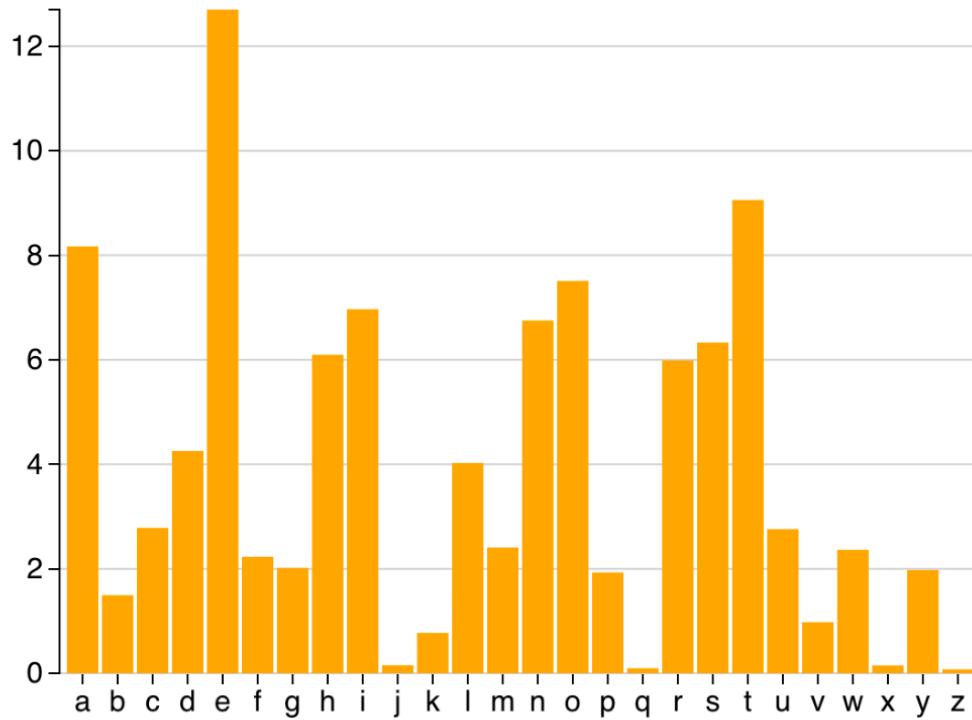
The screenshot shows a Microsoft Excel spreadsheet titled "natural-disasters-type-2.xlsx". The ribbon menu includes Home, Layout, Tables, Charts, SmartArt, Formulas, Data, Review, and a gear icon. The Home tab is selected. The toolbar includes icons for Paste, Bold, Italic, Underline, Insert, Cut, Copy, Paste, Font Size (12), Alignment, Number (General), Format, Conditional Formatting, Styles, Actions, and Themes.

The table has columns labeled "Years", "Avalanche/L Drought/Foo Earthquake/ Extreme Tem Flooding", "Forest Fire", and "Windstorm". The data spans from row 1 to 27, with the first row serving as the header. The "Years" column contains values from 1990 to 2015. The other columns contain numerical values representing the frequency or intensity of different types of natural disasters.

	Years	Avalanche/L Drought/Foo Earthquake/ Extreme Tem Flooding	Forest Fire	Windstorm
1	1990	6	11	57
2	1991	11	14	29
3	1992	17	23	26
4	1993	25	10	18
5	1994	5	8	20
6	1995	15	12	24
7	1996	24	6	11
8	1997	13	11	14
9	1998	23	22	17
10	1999	14	26	33
11	2000	29	43	31
12	2001	21	44	25
13	2002	19	39	37
14	2003	21	23	40
15	2004	16	19	43
16	2005	12	27	25
17	2006	21	20	24
18	2007	10	13	21
19	2008	15	21	23
20	2009	29	31	22
21	2010	32	27	24
22	2011	17	24	30
23	2012	14	31	29
24	2013	12	13	28
25	2014	15	22	26
26	2015	20	33	21
27				
28				
29				
30				

# Data Analysis

Al-Kindi, 850



# Data Analysis

John Tukey, 1977



# Data Analysis

John Tukey, 1977

I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Mean of x

9

Variance of x

11

Mean of y

7.50

Variance of y

4.122

Correlation between x and y

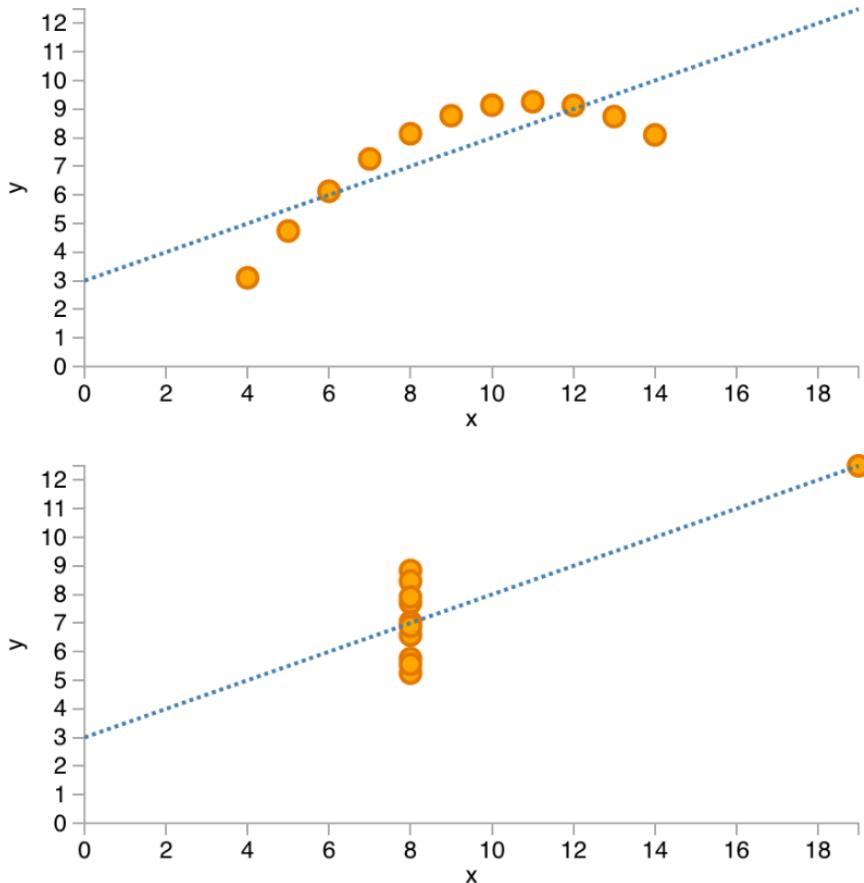
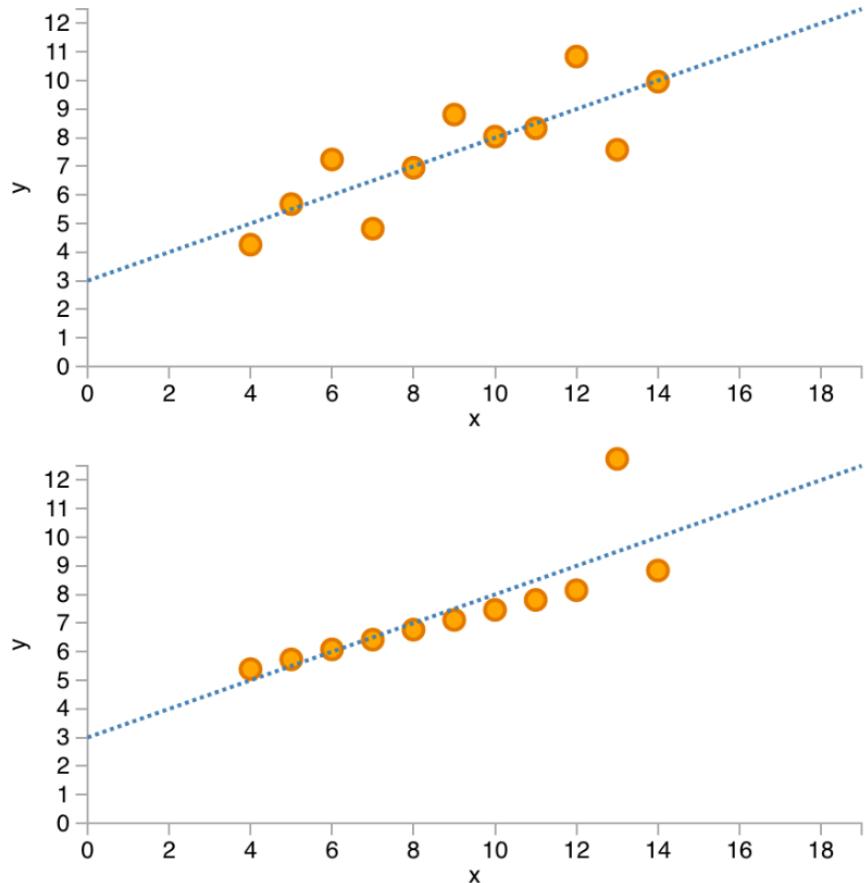
0.816

Linear regression line

$y = 3.00 + 0.50x$

# Data Analysis

John Tukey, 1977



# Data Analysis

## Google Sheets and Excel

The screenshot shows a Google Sheets spreadsheet titled "Dirty Data". The spreadsheet has a green header row with columns labeled A through H. The data starts at row 1, with columns A, C, D, E, F, G, and H having dropdown menus. The first few rows of data are as follows:

	Date	Country	Address	City	Min Value	Max Value	Average	
1	12/23/2015	United States	92 Waverly Ave	New York	2	5	3.5	
2	4/24/2015	United States	28 Orange Ave	Los Angeles	2	7	4.5	
3	4/23/2015	United States	3245 First Street	Los Angeles	4	9	6.5	
4	3/3/2015	United States	554 Washington	New York	3	9	6	
5	2/6/2015	United States	72 Second Ave	Los Angeles	6	9	7.5	
6	5/27/2014	United States	87 Thurston Stre	Los Angeles	3	7	5	
7	5/7/2014	United States	887 Clinton Ave	New York	3	8	5.5	
8	3/18/2014	United States	763 Main St	Los Angeles	5	9	7	
9	6/30/2013	United States	652 Greene Stre	New York	4	6	5	
10	7/19/2015	Germany	Rablstraße 37	Munich	1	10	5.5	
11	7/4/2015	Germany	Johann-Fichte-S	Munich	3	7	5	
12	2/24/2015	Germany	Am Falkpl. 5	Berlin	1	7	4	
13	11/30/2014	Germany	Grillparzerstraße	Munich	3	4	3.5	
14	11/12/2014	Germany	Rudi-Dutschke-S	Berlin	2	6	4	
15	9/23/2014	Germany	Alte Schänhouse	Berlin	4	6	5	
16	8/13/2014	Germany	Platzl 9	Munich	1	8	4.5	
17	6/8/2014	Germany	Röntgenstraße 7	Berlin	1	8	4.5	
18	3/26/2014	Germany	Mehringdamm 3:	Berlin	4	6	5	
19	8/14/2013	Germany	Spiridon-Louis-R	Munich	1	5	3	
20	7/19/2015	England	127 Ledbury Rd	London	4	6	5	
21	7/4/2015	England	157A Commercial	London	4	9	6.5	
22	5/26/2015	England	26-27 Green Lanes	London	2	10	6	
23	2/24/2015	England	160 Piccadilly	London	2	8	5	

The spreadsheet includes a toolbar with various icons and a status bar at the bottom.

# Data Analysis

## Pivot Tables

The screenshot shows a Google Sheets document titled "Dirty Data". The main spreadsheet contains a table with dates in column A and cities in row 1 (Berlin, London, Los Angeles, Munich, New York). The "Report Editor" sidebar is open, showing settings for Rows, Columns, Values, and Filters.

**Rows - Add field**

Group by: Date  
Order: Ascending ▾  
Sort by: Date ▾  
 Show totals

**Columns - Add field**

Group by: City  
Order: Ascending ▾  
Sort by: City ▾  
 Show totals

**Values - Add field**

Display: Average  
Summarize by: AVERAGE ▾

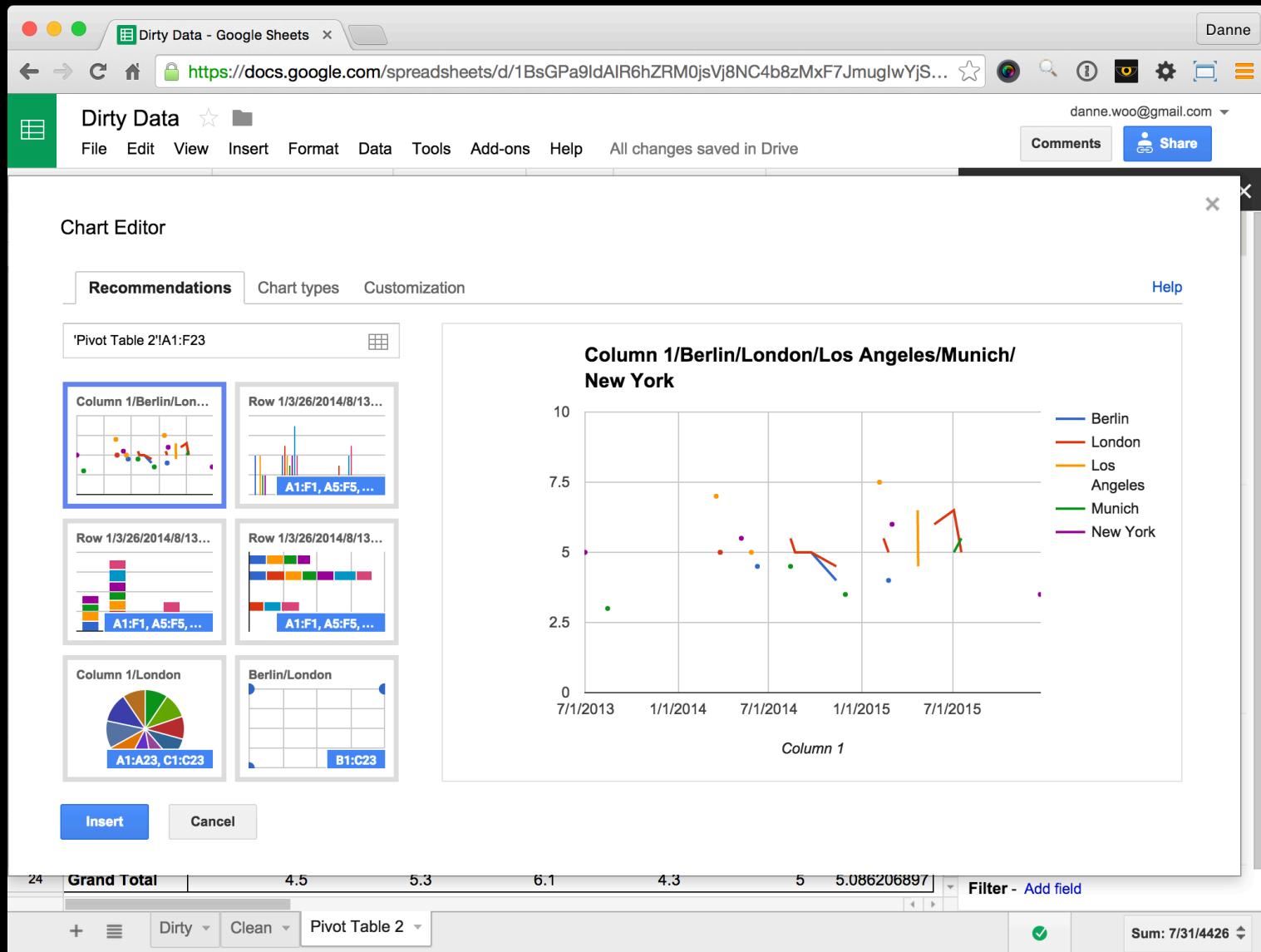
**Filter - Add field**

The main table data is as follows:

	A	B	C	D	E	F
1		Berlin	London	Los Angeles	Munich	New York
2	6/30/2013					
3	8/14/2013					3
4	3/18/2014			7		
5	3/26/2014	5	5			
6	5/7/2014					
7	5/27/2014			5		
8	6/8/2014	4.5				
9	8/13/2014		5.5		4.5	
10	8/22/2014		5			
11	9/23/2014	5	5			
12	11/12/2014	4	4.5			
13	11/30/2014				3.5	
14	2/6/2015			7.5		
15	2/14/2015		5.5			
16	2/24/2015	4	5			
17	3/3/2015					
18	4/23/2015			6.5		
19	4/24/2015			4.5		
20	5/26/2015		6			
21	7/4/2015		6.5		5	
22	7/19/2015		5		5.5	
23	12/23/2015					
24	Grand Total	4.5	5.3	6.1	4.3	

# Data Analysis

## Visualize for Insight



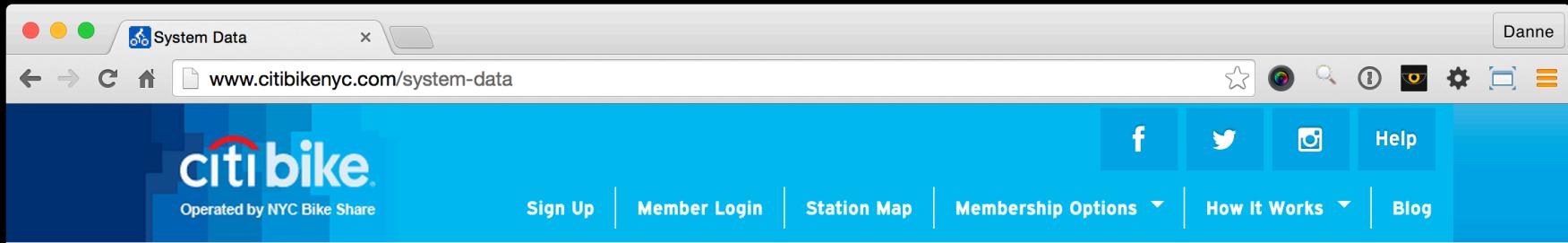
# Data Analysis

CitiBike Data / [bit.ly/qc-data-citibike](http://bit.ly/qc-data-citibike)



# Data Analysis

CitiBike Data - [citibikenyc.com/system-data](http://www.citibikenyc.com/system-data)



## System Data

Where do Citi Bikers ride? When do they ride? How far do they go? Which stations are most popular? What days of the week are most rides taken on? We've heard all of these questions and more from you and now we are happy to provide the data sets to help you discover the answers to these questions and more. We invite developers, engineers, statisticians, artists, academics and other members of the interested public to use the data we provide for analysis, development, visualization and whatever else moves you.

This data is provided according to the [NYCBS Data Use Policy](#).

### Citi Bike Trip Histories

Below are links to downloadable files of Citi Bike trip data. The data includes:

- Trip Duration (seconds)
- Start Time and Date
- Stop Time and Date
- Start Station Name
- End Station Name
- Station ID
- Station Lat/Long
- Bike ID
- User Type (Customer = 24-hour pass or 7-day pass user; Subscriber = Annual Member)
- Gender (Zero=unknown; 1=male; 2=female)
- Year of Birth

# Data Analysis

CitiBike Data – [bit.ly/qc-data-citibike](https://bit.ly/qc-data-citibike)

CitiBike Data 2014 - Google Sheets

<https://docs.google.com/spreadsheets/d/1Xy8lAVRlhGAWIGFmURIV74LBOaXPeucHHoer3bOhPYA/edit#gid=0>

danne.woo@gmail.com

Comments Share

	A	B	C	D	E	F	G	H	I	J	K
1	Date	Trips over the past day	Cumulative trips	Miles traveled today	Miles traveled to date	Total Annual Members	Annual Member ID	24-Hour Passes	7-Day Passes	Purchased (midnight to 11:59 pm)	
2	1/1/2014	6559	6323722	9.254	11,243,581	95971	29	268	10		
3	1/2/2014	9334	6333056	11.736	11,255,317	96000	42	66	3		
4	1/3/2014	1288	6334344	1.855	11,257,172	96042	23	1	1		
5	1/4/2014	2494	6336838	3.706	11,260,878	96065	19	14	1		
6	1/5/2014	2937	6339775	3.72	11,264,598	96084	18	17	4		
7	1/6/2014	10481	6350256	13.016	11,277,613	96102	22	30	7		
8	1/7/2014	7144	6357400	8.2	11,285,814	96124	28	5	4		
9	1/8/2014	10162	6367562	11.768	11,297,582	96152	19	28	4		
10	1/9/2014	14571	6382133	18.401	11,315,983	96171	18	74	10		
11	1/10/2014	10685	6392818	13.433	11,329,416	96189	36	49	5		
12	1/11/2014	8333	6401151	10.735	11,340,151	96225	22	101	16		
13	1/12/2014	13549	6414700	19.141	11,359,291	96247	29	350	22		
14	1/13/2014	22377	6437077	30.829	11,390,121	96276	72	224	31		
15	1/14/2014	10849	6447926	13.271	11,403,392	96348	43	45	4		
16	1/15/2014	23500	6471426	32.303	11,435,695	96391	62	241	30		
17	1/16/2014	21528	6492954	28.807	11,464,502	96453	56	208	37		
18	1/17/2014	21840	6514794	30.773	11,495,275	96509	67	282	34		
19	1/18/2014	10222	6525016	14.596	11,509,870	96576	40	176	15		
20	1/19/2014	9768	6534784	13.591	11,523,461	96616	38	194	13		
21	1/20/2014	14581	6549365	20.103	11,543,564	96654	47	133	11		
22	1/21/2014	4171	6553536	5.429	11,548,993	96701	32	5	1		
23	1/22/2014	2762	6556298	4.061	11,553,054	96733	8	4	3		
24	1/23/2014	5743	6562041	8.336	11,561,390	96741	16	9	1		

# Data Analysis

CitiBike Usage based on temperature?

New York Weather - AccuWeather

www.accuweather.com/en/us/new-york-ny/10007/weather-forecast/349727

World > North America > United States > New York > New York

AccuWeather.com in partnership with abc7 for New York, NY

Follow us on Facebook Twitter YouTube 8+ English (US), °F Login

United States WEATHER New York, NY LOCAL WEATHER AIR QUALITY ALERT

Now 4:31 pm EDT Weekend Extended Month Radar MinuteCast® WATCH VIDEOS

Cloudy 86° Hi 95° Lo 76° RealFeel 106°

Today Aug 17 Partly sunny

Tonight Aug 17 Partly cloudy, warm and humid

Tomorrow Aug 18 Mostly sunny, hot and humid

RealFeel® 89°

MinuteCast® Hourly

Cloudy 86°

VIDEO: Utah Hikers Take Cover During Rainstorm as Floodwaters Pour Over Cliff

WATCH: Best Way to Beat the Intense Heat in NYC

VIDEO: Terrified Woman on Quad Bike Films Koala Chasing After Her

Propertyshark propertyshark.com/promotionalcode 15% off all subscriptions. Limited time offer. Subscribe today

USAA® Medicare Solutions usaa.com/MedicareSolutions We Are Here To Help You Make Important Decisions. Learn More.

Epson® Printers epson.com/Printers Print, Copy, Scan from Home. Epson Quality for Your Family.

Cloudy 86°

DT & HUMID Monday

RealFeel® Temps 55°-105°

Cooler beaches

Video Weather Forecast

New York Radar

BUFFALO NEW YORK PITTSBURGH

Hartford Connecticut Bridgeport Stamford Stamford R.I.

Paterson Allentown Elizabeth New York

Harrisburg Philadelphia

Danne

# Data Analysis

R

Screenshot of the R environment showing various windows and plots.

**R Console:**

```
rgl.sr> ylen <- ylim[2] - ylim[1] + 1
rgl.sr> colorlut <- terrain.colors(ylen)
rgl.sr> col <- colorlut[y - ylim[1] + 1]
rgl.sr> rgl.clear()
rgl.sr> rgl.surface(x, z, y, color = col)
```

**R Data Editor:**

height	weight
58	115
59	117
60	120
61	123
62	126
63	129
64	132
65	135
66	139
67	142
68	146
69	150
70	154
71	159
72	164

**Quartz (2) – Active:**

**R Workspace Browser:**

Object	Type	Structure
dati	data.frame	dim: 20 4
g	factor	levels: 10
l	numeric	length: 12
n	numeric	length: 1
opar	list	length: 2
pie.sales	numeric	length: 6
pin	numeric	length: 2
scale	numeric	length: 1
usr	numeric	length: 4
women	data.frame	dim: 15 2
height	numeric	length: 15
weight	numeric	length: 15
x	numeric	length: 87

**R Package Manager:**

status	Package	Description
<input checked="" type="checkbox"/>	graphics	The R Graphics Package
<input type="checkbox"/>	grid	The Grid Graphics Package
<input type="checkbox"/>	lattice	Lattice Graphics
<input checked="" type="checkbox"/>	methods	Formal Methods and Classes
<input type="checkbox"/>	mgcv	GAMs with GCV smoothness estimation

**RGL device 1 (active):**

# Data Analysis

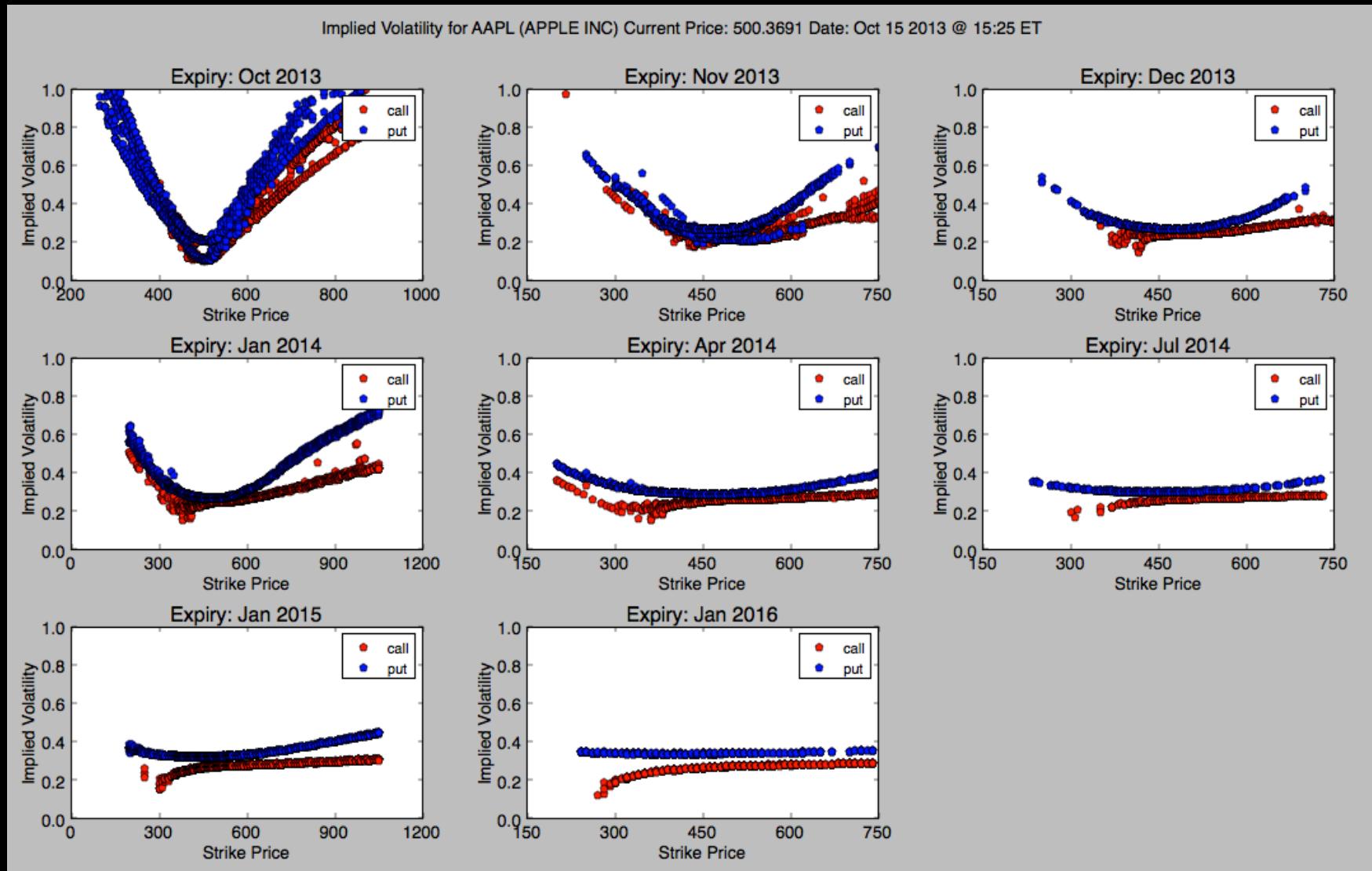
## R Studio

The screenshot shows the RStudio interface with the following components:

- Code Editor:** The left pane displays the R code for creating an interactive plot. The code uses the `manipulate` function from the `manipulate` package to create a plot of stopping distance versus car speed. It includes controls for color, type, pch, axes, and labels.
- Console:** The bottom-left pane shows the R console output, which is identical to the code in the editor.
- Manipulator:** A modal window titled "Manipulate" is open, allowing users to change parameters: color (blue), type (b), pch (19), and two checkboxes for "Draw Axes" and "Draw Labels".
- Help:** The right pane shows the help documentation for the `manipulate` function, including its description and usage examples.
- Plot:** The main pane displays the resulting plot titled "Speed and Stopping Distances of Cars". The x-axis is "Car Speed (mph)" ranging from 5 to 25, and the y-axis is "Stopping Distance (feet)" ranging from 0 to 100. The plot shows a scatter of points with blue lines connecting them, representing individual car trajectories.

# Data Analysis

## Python and Pandas



# Data Analysis

Companies Predicting the Future

Google

NETFLIX



MAJOR LEAGUE BASEBALL



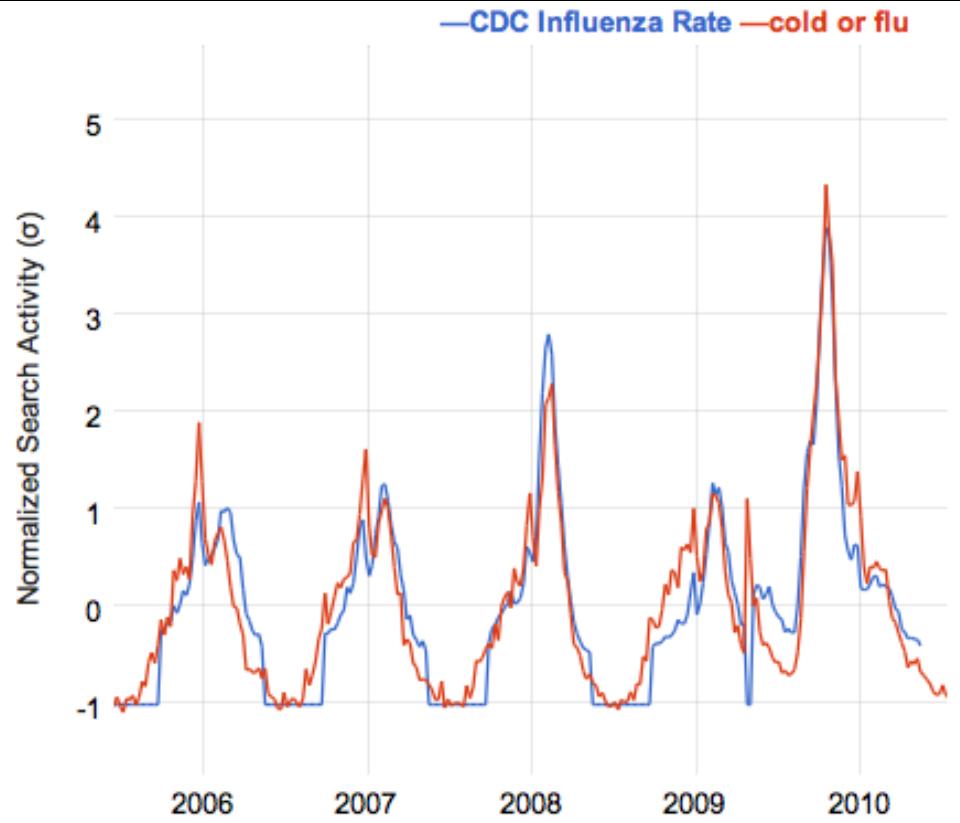
Spotify®

# Data Analysis

Companies Predicting the Future

# Google

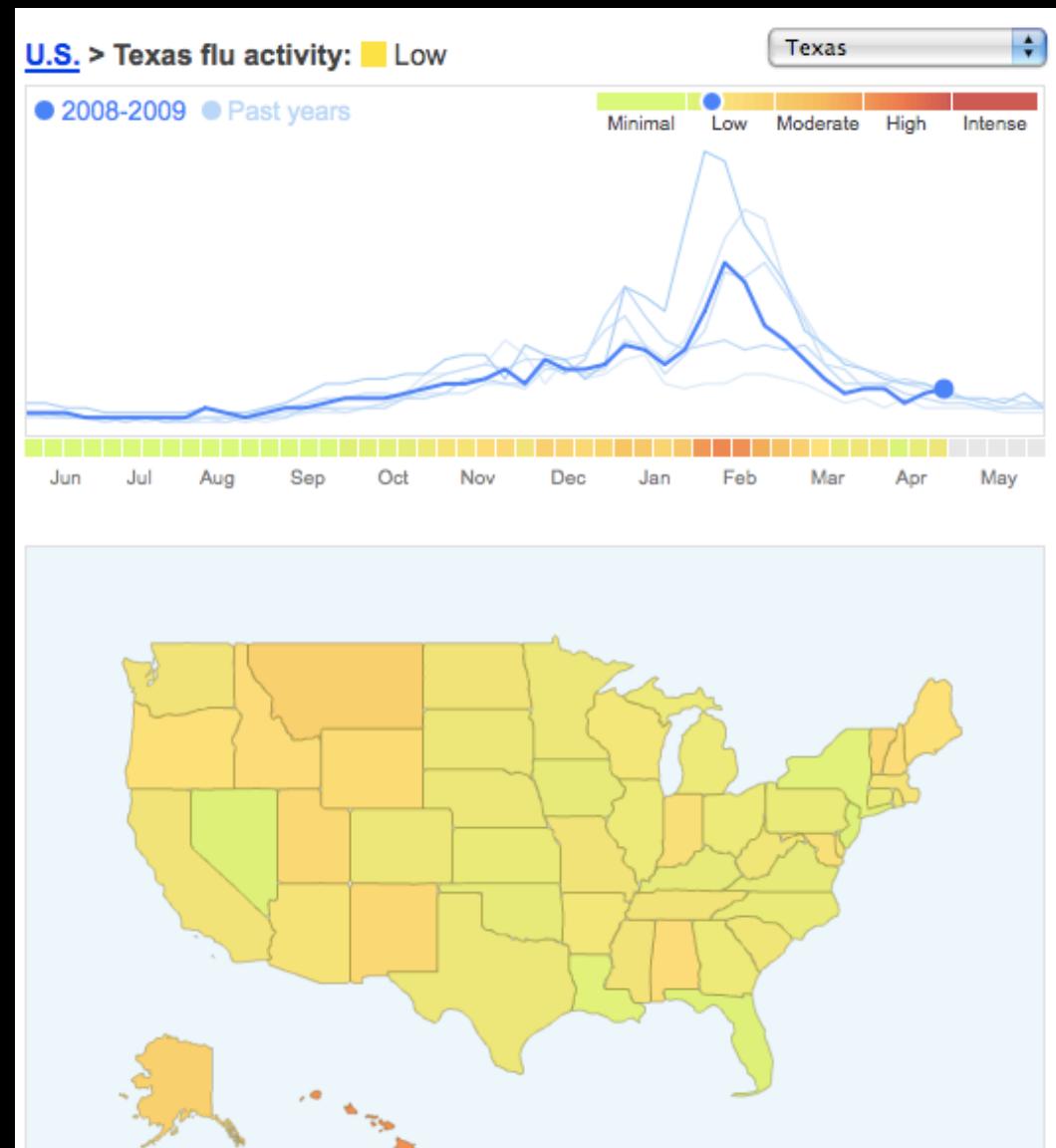
Correlated with CDC Influenza Rate
0.9261 cold or flu
0.9018 how to treat the flu
0.8983 treat the flu
0.8925 cure the flu
0.8918 treatment for flu
0.8910 what to do when you have the flu
0.8907 cold and flu symptoms
0.8889 remedies for the flu
0.8870 is it the flu
0.8860 flu home remedies



# Data Analysis

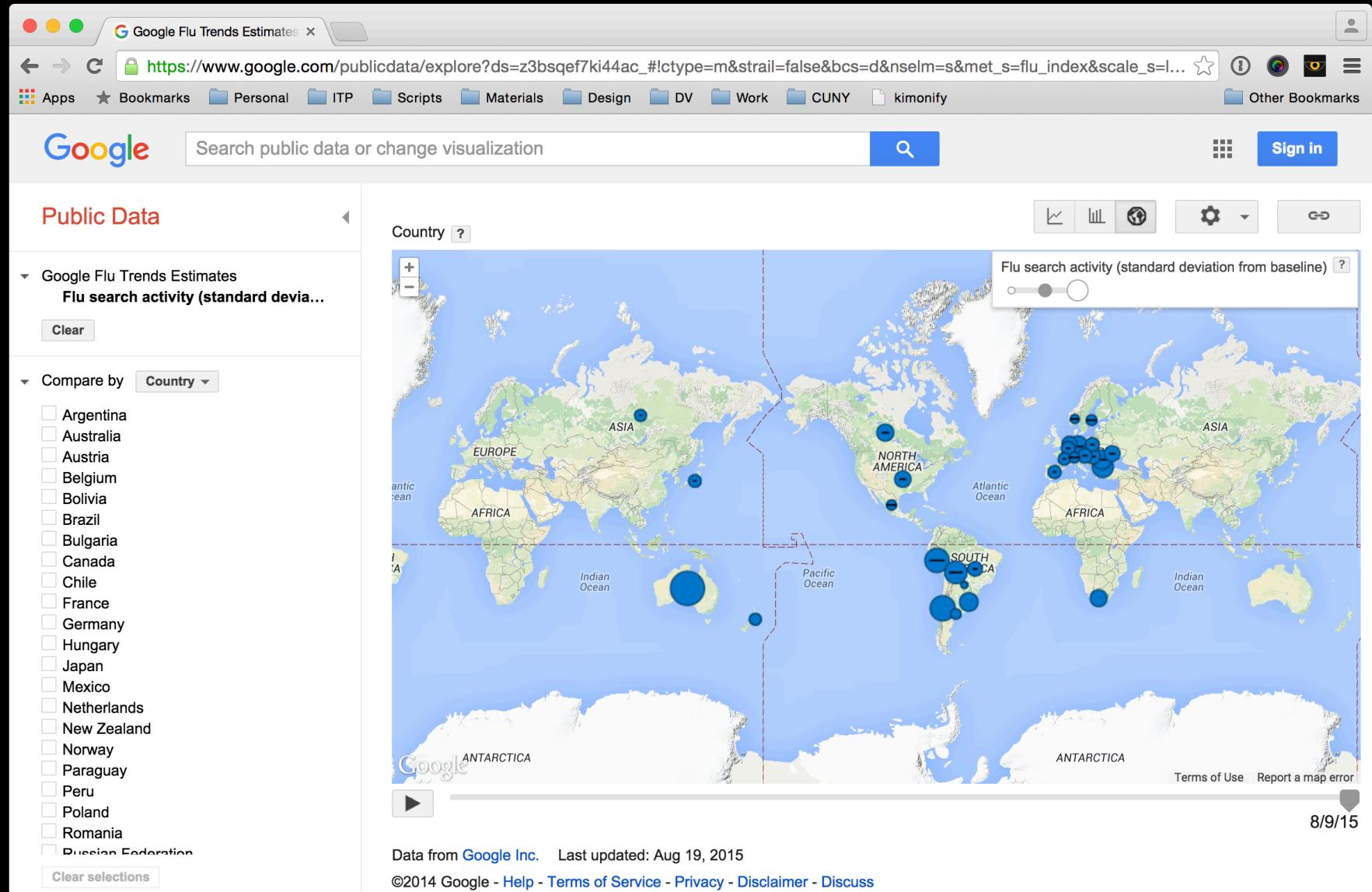
Companies Predicting the Future

Google



# Data Analysis

## Companies Predicting the Future



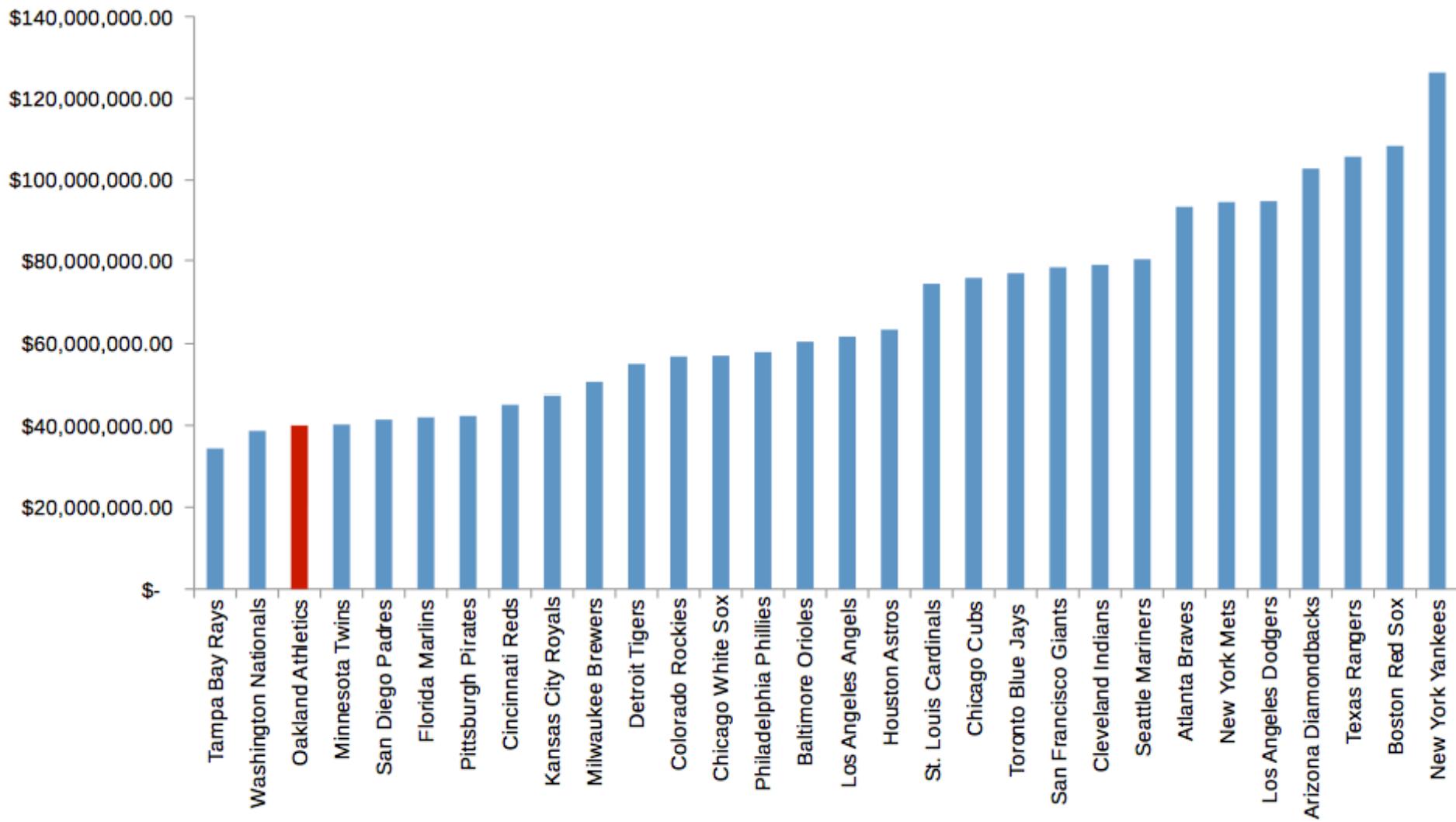
# Data Analysis

Companies Predicting the Future



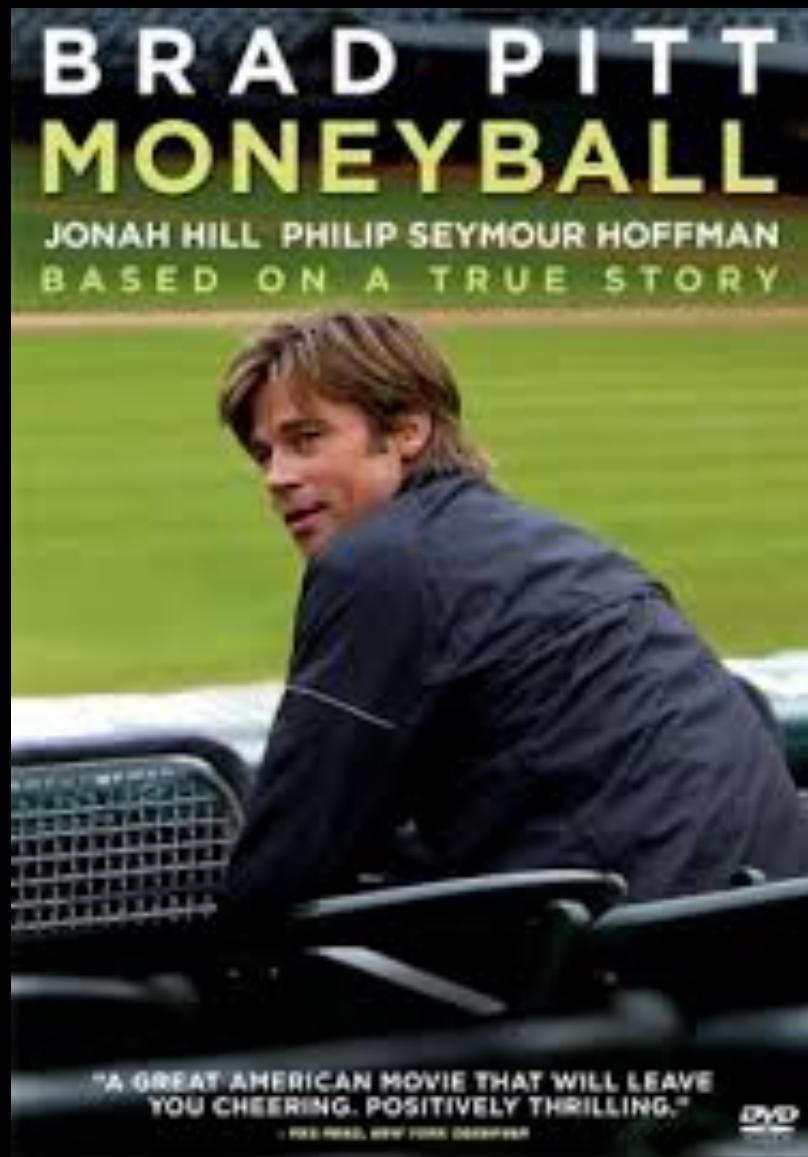
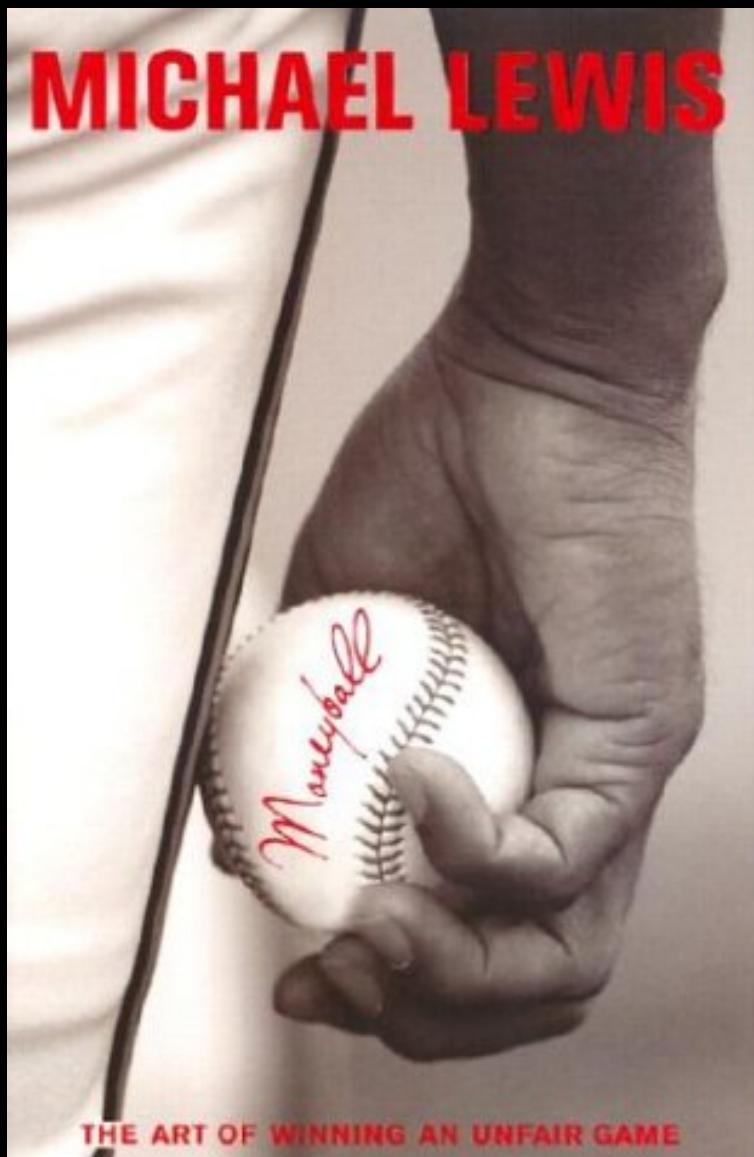
# Data Analysis

## Companies Predicting the Future



# Data Analysis

Companies Predicting the Future



# Data Analysis

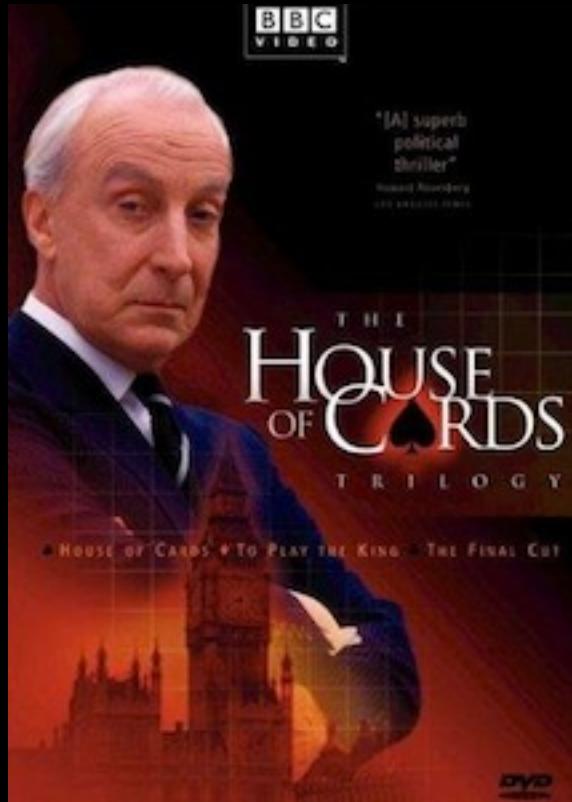
## Companies Predicting the Future



- More than 25 million users
- About 30 million plays per day (and it tracks every time you rewind, fast forward and pause a movie)
- More than 2 billion hours of streaming video watched during the last three months of 2011 alone
- About 4 million ratings per day
- About 3 million searches per day
- Geo-location data
- Device information
- Time of day and week (it now can verify that users watch more TV shows during the week and more movies during the weekend)
- Metadata from third parties such as Nielsen
- Social media data from Facebook and Twitter

# Data Analysis

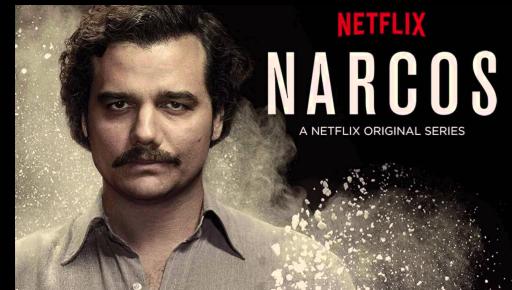
## Companies Predicting the Future



- Data of their subscriber viewing preferences showed they would enjoy a remake of the 1990 BBC miniseries, House of Cards.
- Same users who loved this miniseries also loved movies starring Kevin Spacey or directed by David Fincher.
- Because of this they financed a remake of the BBC drama with Spacey and Fincher.
- Targeted advertising before and after launch.

# Data Analysis

Companies Predicting the Future



# Data Analysis

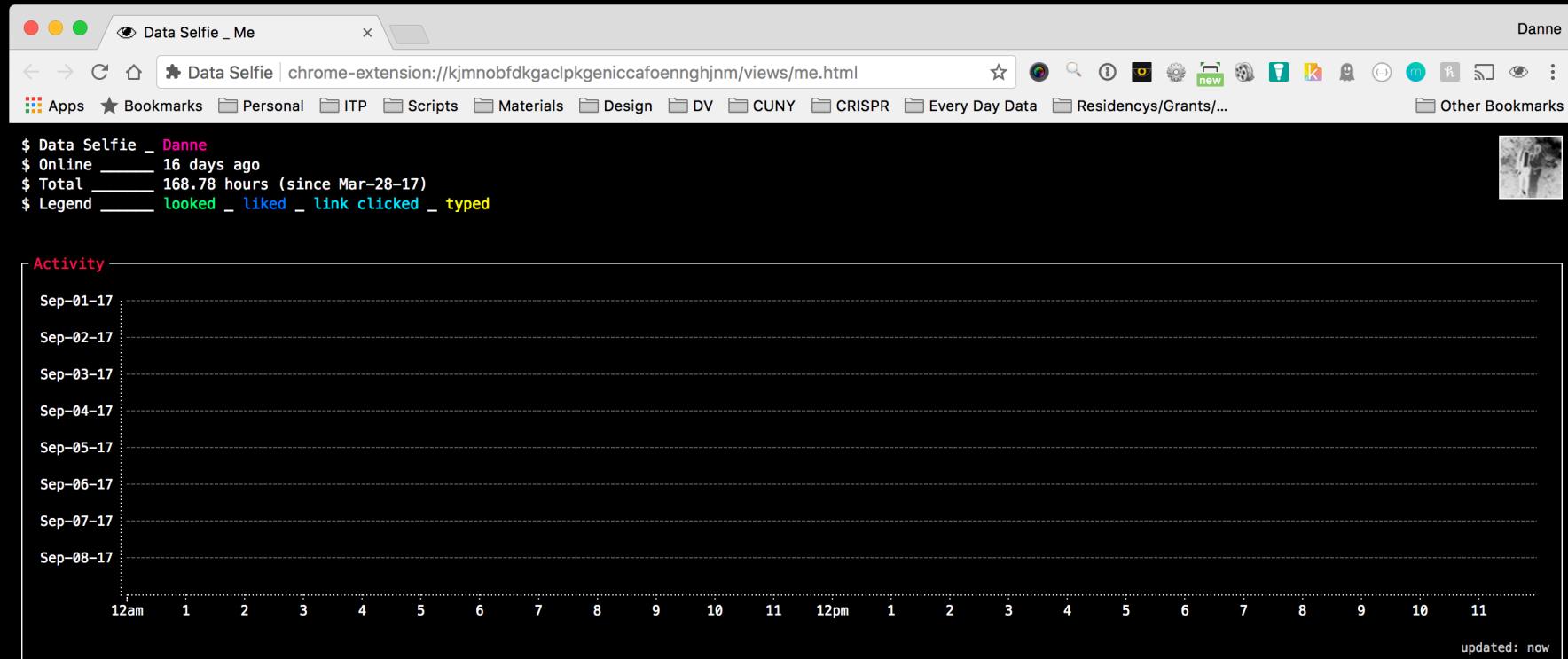
Know Who is Tracking You / Ghostery

The screenshot shows a web browser window with the following details:

- Title Bar:** GHOSTERY Download Add-on
- Address Bar:** https://www.ghostery.com/en/try-us/download-add-on/
- Toolbar:** Apps, Bookmarks, Personal, ITP, Scripts, Materials, Design, DV, Work, CUNY, kimonify, Other Bookmarks
- Content Area:**
  - Test your site now!
  - Sign up for our privacy newsletter!
  - Social media links: Facebook, Twitter, LinkedIn, Google+
  - Login link
  - Navigation links: WHY GHOSTERY, OUR SOLUTIONS, TRY US, INTELLIGENCE, SUPPORT, ABOUT US, SEARCH
- Breadcrumbs:** Home > Try Us / Download Add-On
- Main Headline:** TRY US  
DOWNLOAD GHOSTERY ADD-ON
- Desktop Browser Options:**
  - Opera 5.4.8: ADD TO OPERA
  - Firefox 5.4.8: ADD TO FIREFOX
  - Chrome 5.4.8: ADD TO CHROME
  - Safari 5.4.8: ADD TO SAFARI

# Data Analysis

## Know Who is Tracking You / Data Selfie



Top friends (10 of 402)	
Time spent (in sec) on friends' posts	
5015	Damian Monzillo
3698	Matthew Oomhawfur
3579	Victor De La Cruz
3418	Summer DuBois
3110	Kristin Burtoft

Top pages (10 of 494)	
Time spent (in sec) on pages' posts	
2855	Brooklyn Research
1957	Datavisual
1935	G00D
1562	Adrián González
1304	Science Friday

Top Likes (10 of 39)	
Likes for posts, photos or videos	
6	Surya Mattu
5	Matthew Oomhawfur
3	Andy Rementer
3	Nancy Rose
3	Michael Venutolo-Mantovani