

# Coding the Twitter Sphere:

Humans and Machines Learning Together

Dr. Stuart Shulman

@stuartwshulman

stu@texifter.com

# Acknowledgements

The National Science Foundation

Mark J. Hoy

# Conflict of Interest Disclosure

I am the sole manager of Texifter

We sell Discover Text licenses

We sell Gnip data licenses

# A Master Metaphor: Sifter



# An Open Source Kernel



Like 120

register for a free account  
forgot password?

Maintained by texifter in a [FISMA-compliant](#) environment

[Home](#) | [About QDAP](#) | [About CAT](#) | [PCAT](#) | [Terms of Service](#) | [Privacy Statement](#) | [CAT Help Wiki](#) | [Contact Us](#)

## Welcome to the Coding Analysis Toolkit (CAT)

CAT is a free service of the Qualitative Data Analysis Program (QDAP), and hosted by the University Center for Social and Urban Research, at the University of Pittsburgh, and QDAP-UMass, in the College of Social and Behavioral Sciences, at the University of Massachusetts Amherst. CAT was the 2008 winner of the "Best Research Software" award from the organized section on Information Technology & Politics in the American Political Science Association.

For the CAT Quick Start Guide, you can view the PDF file [CAT Quickstart Guide](#) or watch the [CAT Tutorial - February 23, 2009](#)

May 5, 2010 - CAT is now an open source project! You can host your own version of CAT from the project source code at:  
<http://sourceforge.net/projects/catoolkit/>

Read about the latest innovations in text analytics on the [Texifter blog](#) or via a specially curated [@scoopit page](#)

### CAT Statistics

There are currently **5,936** primary CAT accounts and **1,014** sub-accounts. CAT users have uploaded **5,793** coded datasets and **8,425** raw datasets. They have coded a total of **1,482,741** items and adjudicators have made **164,177** validation choices in CAT.



[What CAT Does](#)

[CAT Features](#)

[Praise for CAT](#)

### What can you do in CAT?

- Efficiently code raw text data sets
- Annotate coding with shared memos
- Manage team coding permissions via the Web
- Create unlimited collaborator sub-accounts
- Assign multiple coders to specific tasks
- Easily measure inter-rater reliability
- Adjudicate valid & invalid coder decisions
- Report validity by dataset, code or coder
- Export coding in RTF, CSV or XML format
- Archive or share completed projects

### What file types can CAT import?

- Plain text
- HTML
- CAT XML
- Merged ATLAS.ti coding

### CAT Resources

[Raw Data Preparation Guide](#)  
[ATLAS.ti Upload Preparation](#)

Have you tried [DiscoverText?](#)  
*Featuring the Facebook Graph & Twitter APIs*



# Three Primary Tasks in CAT

## What can you do in CAT?

---

Efficiently code raw text data sets

Annotate coding with shared memos

Manage team coding permissions via the Web

Create unlimited collaborator sub-accounts

Assign multiple coders to specific tasks

Easily measure inter-rater reliability

Adjudicate valid & invalid coder decisions

Report validity by dataset, code or coder

Export coding in RTF, CSV or XML format

Archive or share completed projects

# Classification of Text

A 2500 year-old problem

Plato argued it would be frustrating

It still is...

# Grimmer & Stewart "Text as Data"

## Political Analysis (2013)

Volume is a problem for scholars

Coders are expensive

Groups struggle to accurately label text at scale

Validation of both humans and machines is "essential"

Some models are easier to validate than others

All models are wrong

Automated models enhance/amplify, but don't replace humans

There is no one right way to do this

"Validate, validate, validate"

"What should be avoided then, is the blind use of any method without a validation step."



# coderrank (Patent Pending)

Coder:

**# Datasets:** 31  
**# Annotations:** 15480  
**# Validations:** 313  
**% Correctness:** 96.49 %

Coder:

**# Datasets:** 23  
**# Annotations:** 7676  
**# Validations:** 351  
**% Correctness:** 82.62 %

Coder:

**# Datasets:** 73  
**# Annotations:** 18300  
**# Validations:** 2790  
**% Correctness:** 93.26 %

Coder:

**# Datasets:** 32  
**# Annotations:** 15283  
**# Validations:** 276  
**% Correctness:** 42.75 %

Coder:

**# Datasets:** 109  
**# Annotations:** 84793  
**# Validations:** 7202  
**% Correctness:** 91.53 %

# Three Important Books

"Google is not just a company, it is an entirely new way of thinking about understanding who we are and what we want. Jarvis has done something really important: extend that approach to business and culture, revealing just how revolutionary it is." —Chris Anderson, author of *The Long Tail*

## What Would Google Do?

Written and Read by  
Jeff Jarvis

U N A B R I D G E D

NATIONAL BESTSELLER

JAMES GLEICK

BESTSELLING AUTHOR OF CHAOS

THE INFORMATION

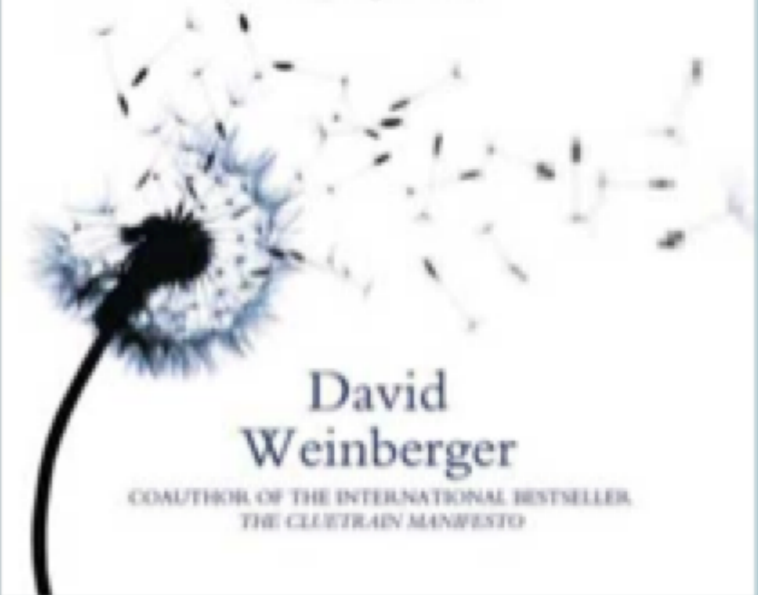
A HISTORY, A THEORY, A FLOOD

"BRILLIANT, ILLUMINATING AND SEXY THEORETICAL."  
—THE NEW YORK TIMES

## Everything Is Miscellaneous

THE POWER OF THE  
NEW DIGITAL DISORDER

"Perfectly placed to tell us what's really new  
about the second-generation Web."  
—Los Angeles Times

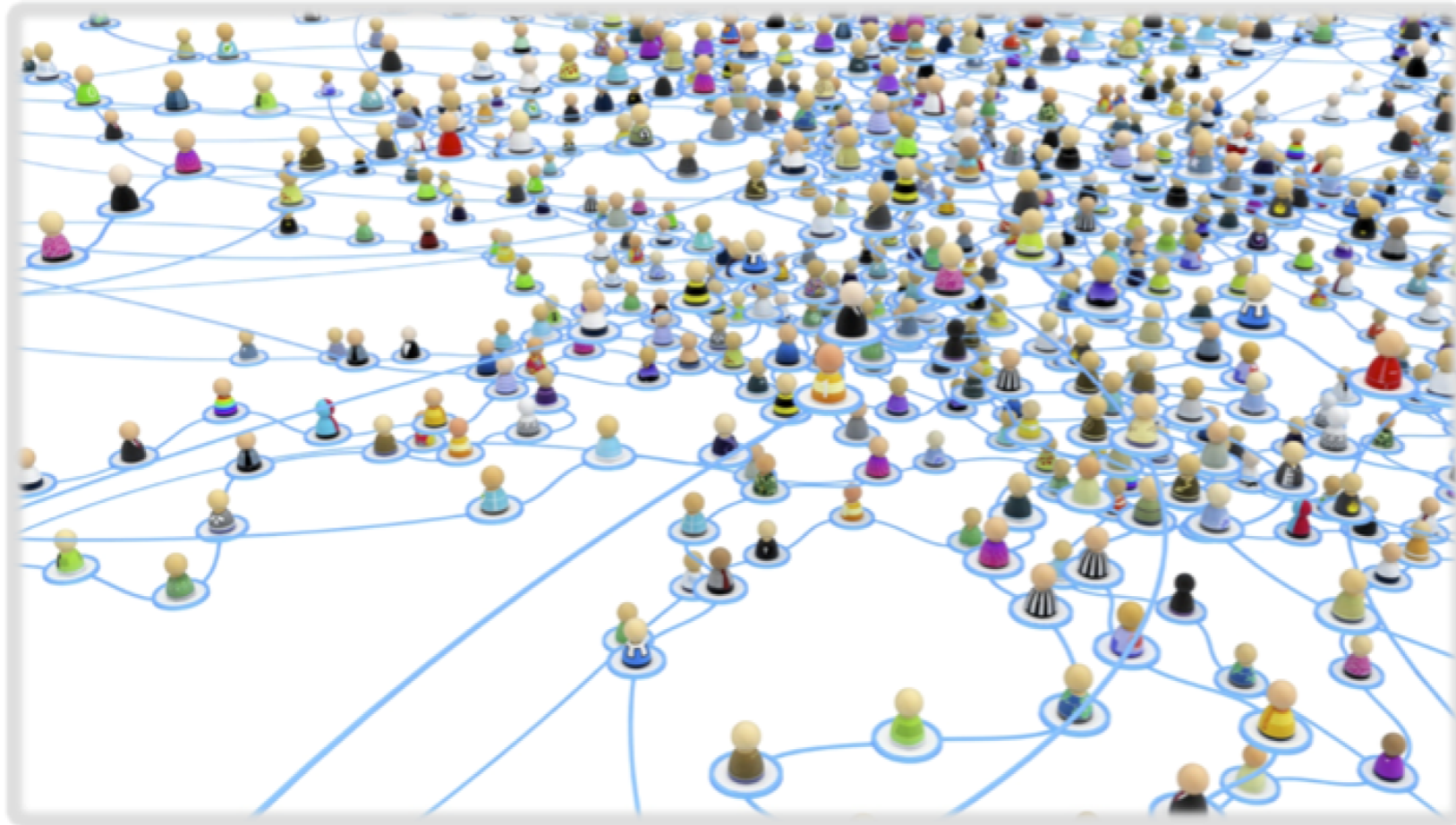


David  
Weinberger

COAUTHOR OF THE INTERNATIONAL BESTSELLER  
THE CLEUTRAIN MANIFESTO

# One Particularly Important Idea

**Crowdsourcing brings widely distributed wisdom to process of text analysis**



**“This is really the biggest paradigm shift in innovation since the Industrial Revolution”**

- MIT professor Eric von Hippel, specialist in innovation management

# Five Pillars of Text Analytics

Search

Filter

Code

Cluster

Classify

You can execute all five using DT



# Pillar #1: Search

- ... that shutting down traffic was his "favorite thing about being Governor" - <http://t>
- ...er affected E.Coast traffic. Very bad. Obama HHS & IRS abuse of power affects millions natio
- ...not called over GWB traffic - <http://t.co/0FMHGnrBwL> <http://t.c>
- ...erage of Christie's Traffic Scandal Than in Last Six Months on IRS <http://t.co/FtvB62n7G>
- ...er affected E.Coast traffic. Very bad. Obama HHS & IRS abuse of power affects millions natio
- ...ee them languish in traffic, and hear the lamentations of their
- ...hristie ordered the traffic jam. How do u feel?
- ...KE ON GOV CHRISTIE? TRAFFIC JAM APOLOGY? CHRISTIE IS TOO OOOOOO FAT!! FRACK THE FAT' CABOT OIL& GA
- ...logan: Think of the traffic studies I can do with the presidential motorcade. #bridgegat
- ...t knowing about the traffic is like Obama not knowing about Benghazi is he giving O a pa
- ...ave been broken' in traffic scandal: <http://t.co/KZmG9F8pp>
- ...ave been broken' in traffic scandal: <http://t.co/c3kMAWLyg>
- ...ave been broken' in traffic scandal: <http://t.co/SuSBcnRJ3>
- ...ave been broken' in traffic scandal: <http://t.co/gfoT5WPc1>
- ...not called over GWB traffic <http://t.co/wSHmVwymZS> #bridgegat
- ...of blocking so much traffic it's Christie. (Pause) He's a very large man
- ...not called over GWB traffic <http://t.co/wSHmVwymZS> #bridgegat

# Search for Negative Cases



## Agatha Christie

Writer

Dame Agatha Mary Clarissa Christie, DBE was an English crime writer of novels, short stories, and plays. [Wikipedia](#)

**Born:** September 15, 1890, [Torquay, United Kingdom](#)

**Died:** January 12, 1976, [Wallingford, Oxfordshire, United Kingdom](#)











# Defined Search (Multi-term)

## Search and Browse Archive

Christie - Twitter API (Archive)

Showing 1 to 100 of 4,376 total

Advanced filters

-  Add to Bucket: [New](#) | [Existing](#) | [Selected](#)  1 of 44
-  ...ign money with gov: **Republican** Steve Lonegan, running for Congress, promotes... <http://t.co/bVa2jom6Y>
  -  ...e's "leadership" in **bridge** scandal: <http://t.co/E8TmAELhu>
  -  Numbers Monday: **Bridge**-Gate Helps Christie on Twitter, but Cruz Wins the Week <http://t.co/17Q8fH8x8>
  -  ...n 2 lanes. National **GOP** shuts down the whole govt. When it comes to petty, some GOPers think bigger <http://t.co/...>
  -  ...ers: Embattled N.J. **governor** faces probe over Sandy funds: Congressman <http://t.co/uWLBZ3YGm>
  -  Chris Christie: A **Bridge** too Far? BRIDGEGATE: WHO WILL GO TO JAIL?????? <http://t.co/UsH9Qbeyvw> via @YahooFinanc
  -  ...stie Unaware He Was **Governor** <http://t.co/EllusQCvT>
  -  ...nefit of doubt amid **bridge** sca... <http://t.co/mntmB0hzd>
  -  ... Christie blocked a **bridge** but Cory Bookers was the one that had an imaginary friend dealing drugs under it
  -  ... bring back that GW **Bridge** character at Marve
  -  ...ign money with gov: **Republican** Steve Lonegan, running for Congress, promot... <http://t.co/MP4iFmmfQ>
  -  ...s striking that the **GOP** golden boys of '09-Christie and McDonnell-were hit with major scandals
  -  ... Christie blocked a **bridge** but Cory Bookers was the one that had an imaginary friend dealing drugs under it
  -  ...ign money with gov: **Republican** Steve Lonegan, running for Congress, promot... <http://t.co/Gus6ZsGiL>
  -  ...ign money with gov: **Republican** Steve Lonegan, running for Congress, promot... <http://t.co/ntZ7Ohskp>
  -  ..."can't imagine" the **governor** isn't telling the truth. <http://t.co/...>
  -  RT @johnlair58: .@**GOP** chairman: 'We're a young, fresh party'; he left out #liberal (he cites liberal Chris Christie)




Showing 1 to 100 of 4,376 total

# Pillar #2: Filters

## Search and Browse Dataset

Relevant Tweets? ([Dataset](#))

Showing 1 to 62 of 62 total **Filters applied**

   Add to Bucket: [New](#) | [Existing](#) | [Selected](#)

### Advanced filters

Search query:

Defined search:

Filter by date:  to

Filter by meta: 

Filter by annotations:

Filter by coding:

Filter by classification:

Set Classification Bounds: [set classification boundary filter](#)

Selected filters:

<input type="text" value="No, not Christie"/>	<input type="text" value="Winning"/>	<a href="#">Remove</a>
<input type="text" value="(items not coded)"/>		<a href="#">Remove</a>

[Add filter](#)

[Add filter](#)

[Add filter](#)

[Add filter](#)

[close](#)


# Another Common Filter

**Advanced filters**

Search query:

Defined search:

Filter by date:  to

Filter by meta:   =  [Add filter](#)

Filter by annotations:  [Add filter](#)

Filter by coding:  [Add filter](#)

Selected filters:

from_user:	Contains	christie	<a href="#">Remove</a>
(items not coded)			<a href="#">Remove</a>

[close](#)



TopMeta Discovery



Top values for: favourites\_count:

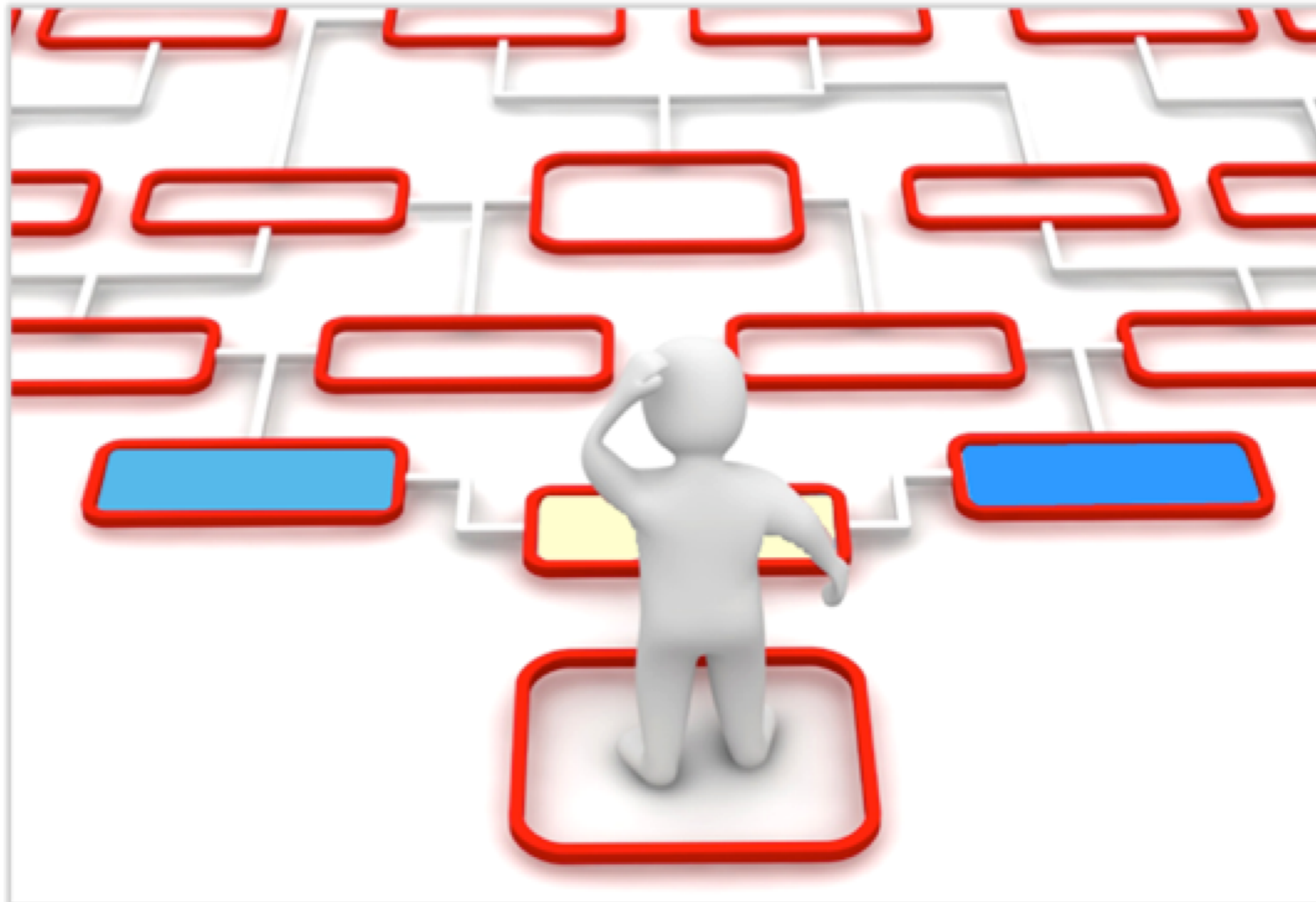
1 of 46

Meta Value ▾	Total	Filter
146363	1	
45411	1	
44173	2	
43412	1	
43411	2	
41698	1	
39859	2	
39195	1	
36395	1	
36394	1	

Showing 1 to 10 of 458 total

# Pillar#3: Human Coding

**Annotation enhances your analysis by  
*applying human interpretation to machine results***



# Keystroke Coding is Fast


**DiscoverText** Relevant Tweets v3 (Dataset) [return to Dashboard](#)

Sort: Key Code # columns: 6 594/1407 : 430

(1) Yes, Chris Christie (3) No, not Christie

### Document Metadata

created\_at: 2014-01-12 23:03:07  
followers\_count: 3290  
friends\_count: 2829  
from\_user: SydneyNewsdaily  
screen\_name: SydneyNewsdaily  
source: dlvr.it  
url\_mention: http://t.co/f95whEAXQv  
url\_mention\_expanded: http://dlvr.it/4hKzNQ  
user\_created\_at: Thu Jun 28 00:43:45 +0000 2012  
user\_description: Your best source of Sydney News on Twitter  
user\_location: USA, Maine  
username: Sydney Newsdaily

 **Sydney Newsdaily** [Follow](#)  
@SydneyNewsdaily

NSW prosecutors consider murder charge after Daniel Christie dies in Sydney ... - ABC Online [dlvr.it/4hKzNQ](http://dlvr.it/4hKzNQ)

12:03 AM - 13 Jan 2014

← ↻ ★



# Coding Off a List is Faster

**Search and Browse Dataset**  
Relevant Tweets v2 (Dataset)  
Showing 1 to 100 of 793 total

Advanced filters:

1 of 8

**Code Items by List** ✕

(1) Yes, Chris Christie (3) No, not Christie

Cancel Code

- ...RT @warriorwoman9  
<http://t.co/0uwid9G4u> codable
- ...ts New Theory About  
<http://t.co/loBpbUQSk> codable
- ...T @lordxmen2k: The "Chris Christie" Code: He Used Sandy Relief Funds To Make Himself Look Good & The Bridge Lane Closures To Exact Revenge codable
- Rove: Bridge scandal shows Chris Christie is 'what we want' in a president <http://t.co/KKompActRI> lol #p2 #tcot #uniteblue #pjnet #lntyhb codable
- ...ts New Theory About Chris Christie's Bridge GionnyScandalChris Christie, Mark Sokolich, Chris... <http://t.co/UscODPigF> codable
- ...s expected for Gov. Chris Christie aides over bridge scandal: <http://t.co/mMhQEMiaLJ> codable
- The "Chris Christie" Code: He Used Sandy Relief Funds To Make Himself Look Good & The Bridge Lane Closures To Exact Revenge On Democrats codable
- ...O to Rightbloggers, Chris Christie Finally Earns Respect With His Bridge to Benghazi <http://t.co/sh4N0HnaE5> via @villagevoic codable
- Karl Rove: Chris Christie's Bridge Scandal Response Gives 'Street Cred' with Tea Party <http://t.co/xWkqdPNorh> via @BreitbartNews BS KAR codable
- Say What?: Chris Christie Bridge Scandal - KGAN-TV CBS 2 Iowa - Top Stories: <http://t.co/WNfxKDiyl> codable
- ...s expected for Gov. Chris Christie aides over bridge scandal: <http://t.co/NRG9kTAQv> codable

Showing 1 to 100 of 793 total

# Data Cleaning is Fundamental



This unit is coded with: *No, off topic*

Edit Coding



**A. Z.**

@AldanaZaza

Follow

Cuando Agatha Christie se enteró q su secretaria y su marido eran amantes creó una escena de crimen y se desapareció unas semanas.

5:58 PM - 12 Jan 2014

1 RETWEET



# Pillar #4: Clustering

Archive name: **Christie - Twitter API**










[Return to archive details](#)

Near-duplicates (1839 clusters)



[Create bucket from clusters](#)

[Create dataset from clusters](#)

Similar Clusters	Cluster Options
 (214)	<a href="#">create bucket</a>   <a href="#">create dataset</a>   <a href="#">rename cluster</a>
 (157)	<a href="#">create bucket</a>   <a href="#">create dataset</a>   <a href="#">rename cluster</a>
 (130)	<a href="#">create bucket</a>   <a href="#">create dataset</a>   <a href="#">rename cluster</a>
 (94)	<a href="#">create bucket</a>   <a href="#">create dataset</a>   <a href="#">rename cluster</a>
 (78)	<a href="#">create bucket</a>   <a href="#">create dataset</a>   <a href="#">rename cluster</a>
 (77)	<a href="#">create bucket</a>   <a href="#">create dataset</a>   <a href="#">rename cluster</a>
 (66)	<a href="#">create bucket</a>   <a href="#">create dataset</a>   <a href="#">rename cluster</a>
 (65)	<a href="#">create bucket</a>   <a href="#">create dataset</a>   <a href="#">rename cluster</a>
 (60)	<a href="#">create bucket</a>   <a href="#">create dataset</a>   <a href="#">rename cluster</a>

## Search and Browse Archive Cluster

Christie - Twitter API / (Archive Cluster)

[Advanced filters](#)

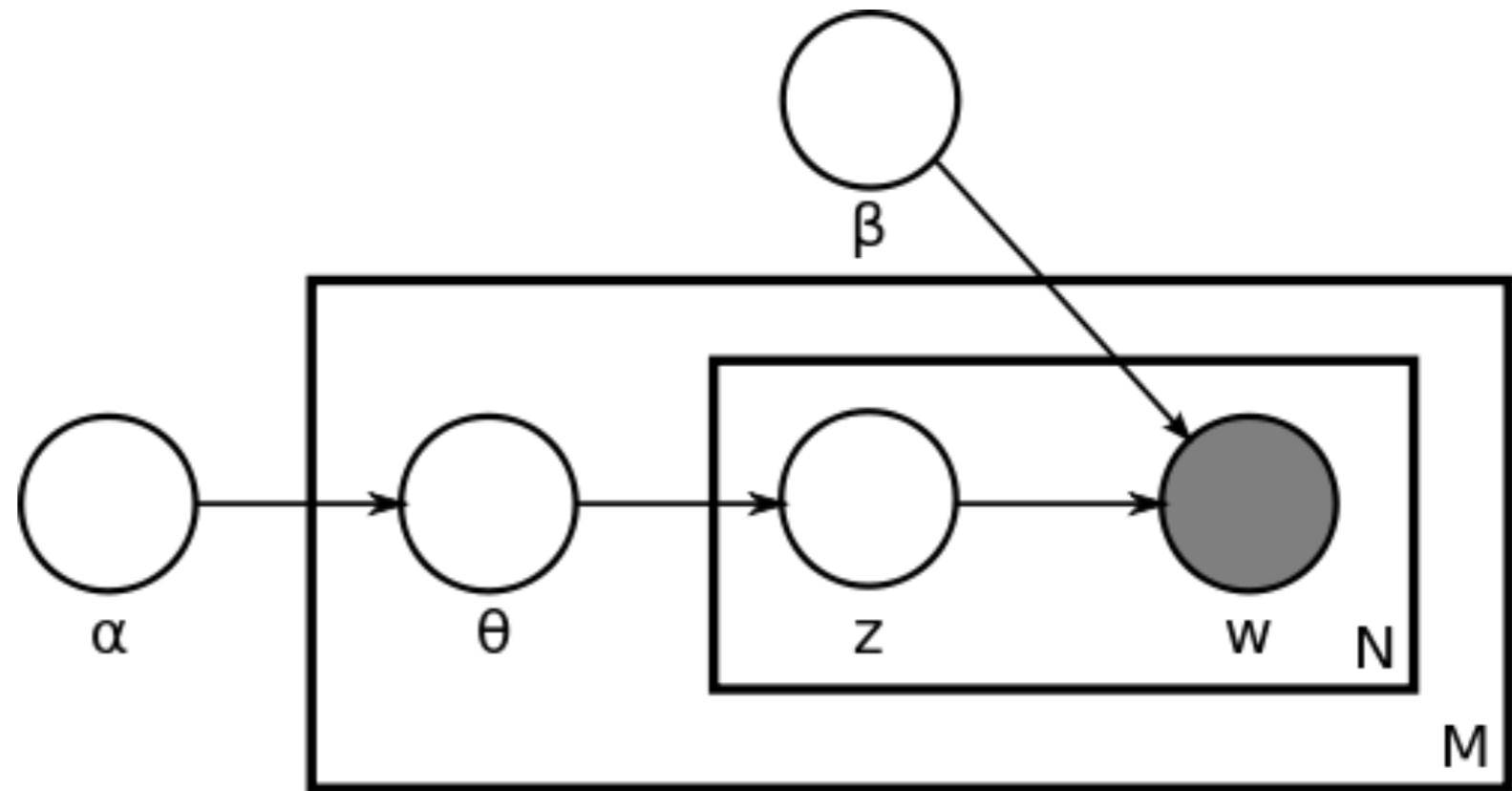
Showing 1 to 100 of 214 total

  Add to Bucket: [New](#) | [Existing](#) | [Selected](#)



-  Making federal case against Christie on lane closures tough: expert - New York Daily News: New  
<http://t.co/JxCyCH7Md6>
-  Making federal case against Christie on lane closures tough: expert - New York Daily News: New  
<http://t.co/dJUB5m4tXT>
-  Making federal case against Christie on lane closures tough: expert - New York Daily News: New  
<http://t.co/FVsGYQ1Ev5>
-  Making federal case against Christie on lane closures tough: expert - New York Daily News: New  
<http://t.co/mCMVRriPwa>
-  Making federal case against Christie on lane closures tough: expert - New York Daily News: New  
<http://t.co/T2PoA4f6xN>
-  Making federal case against Christie on lane closures tough: expert - New York Daily News: New  
<http://t.co/gcpSVKYOWQ>
-  Making federal case against Christie on lane closures tough: expert - New York Daily News: New  
<http://t.co/9Uv8D6pLX7>
-  Making federal case against Christie on lane closures tough: expert - New York Daily News: New  
<http://t.co/aYi6SWUFzT>

# Latent Dirichlet Allocation (LDA) Topic Models



# LDA on the Christie Data

Topic 1 : christie, sandy, christies, funds, relief, feds, investigating, daily, gov, feminized

Topic 2 : with, daniel, didnt, after, murder, time, agatha, death, former, mayor

Topic 3 : bridge, about, traffic, more, scandal, chris, nj, some, just, says

Topic 4 : like, gop, bridgegate, what, 2016, know, now, will, bully, dont

Topic 5 : obama, benghazi, impeachment, dem, have, probe, lawmaker, floats, possibility, gwob

Topic 6 : jersey, over, stages, still, aides, grief, bogus, hes, news, subpoenas

Topic 7 : rove, closures, karl, york, while, federal, party, tea, governor, president

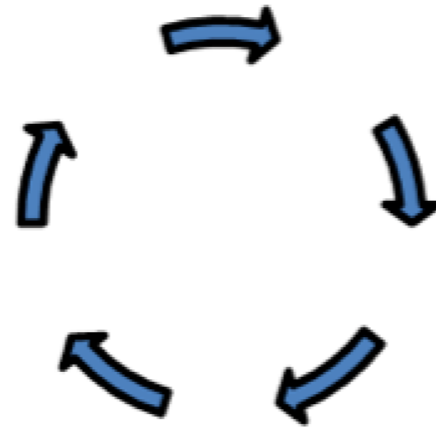
Topic 8 : irs, political, been, show, republicans, media, get, laws, word, scandals



# Pillar #5: Machine-Learning



What Humans Do Best



What Computers Do Best

**Humans and machines learning together**



# Create a Dataset to Code

Any archive or bucket

Use the random sampling tool

Standard: All coders get all items

Triage: Coders get next uncoded item

# Select from Three Coding Styles

Default: Mutually Exclusive Codes

Option 1: Non-Mutually Exclusive Codes

Option 2: User-Defined Codes  
(Grounded Theory)

# Assign Peers to Code a Dataset

How many coders?

How many items need to be coded?

How many test or training sets?

There are no cookbook answers

# Look at Inter-Rater Reliability

Highly reliable coding (easy tasks)

Unreliable coding (interesting tasks)

If humans can't, neither can machines

Some tasks better suited for machines

# Adjudication: The Secret Sauce

Expert review or consensus process

Invalidate false positives

Identify strong and weak coders

Exclude false positives from training sets



# Adjudicate Dataset

Dataset Details > Validate Dataset

Dataset: Fear Arousal >95 v1



Code: **Fear Rejection**

(1) Valid

(2) Skip to next

(3) Not valid

Validations remaining: 90  
[\[+\] Change Code Filter](#)

40.00% of users coded this as Fear Rejection

- Soooo Tired of Dat Commercial of Tha Woman Name Terry w| Tha Hole n Her Throat & Da Blonde Wig #NoOffense

## Coder choices

Coder	Codes
Shulman, Stu	Fear Rejection
Pruitt, Katie	Fear Arousal
Apostoleris, Lucas	Fear Arousal
Liew, Jasy	Fear Neutral
Eanes, Ryan	Fear Rejection

# Adjudicate Dataset

[Dataset Details](#) > [Validate Dataset](#)

Dataset: Fear Arousal >95 v1



Code: **Fear Arousal**

Adjudicated as **Valid**

[Clear / Edit Adjudication](#)

100.00% of users coded this as Fear Arousal

Validations remaining: **89**

[\[+\] Change Code Filter](#)

Holy shit that commercial of that lady Terri who used to be a smoker is freaky

## Coder choices

Coder	Codes
Shulman, Stu	Fear Arousal
Pruitt, Katie	Fear Arousal
Apostoleris, Lucas	Fear Arousal
Liew, Jasy	Fear Arousal
Eanes, Ryan	Fear Arousal

# Use Classification Scores as Filters

Iteration plays a critical role

Train, classify, filter

Repeat until the model is trusted

Each round weeds out false positives

# Classifier Histograms: More Filtering

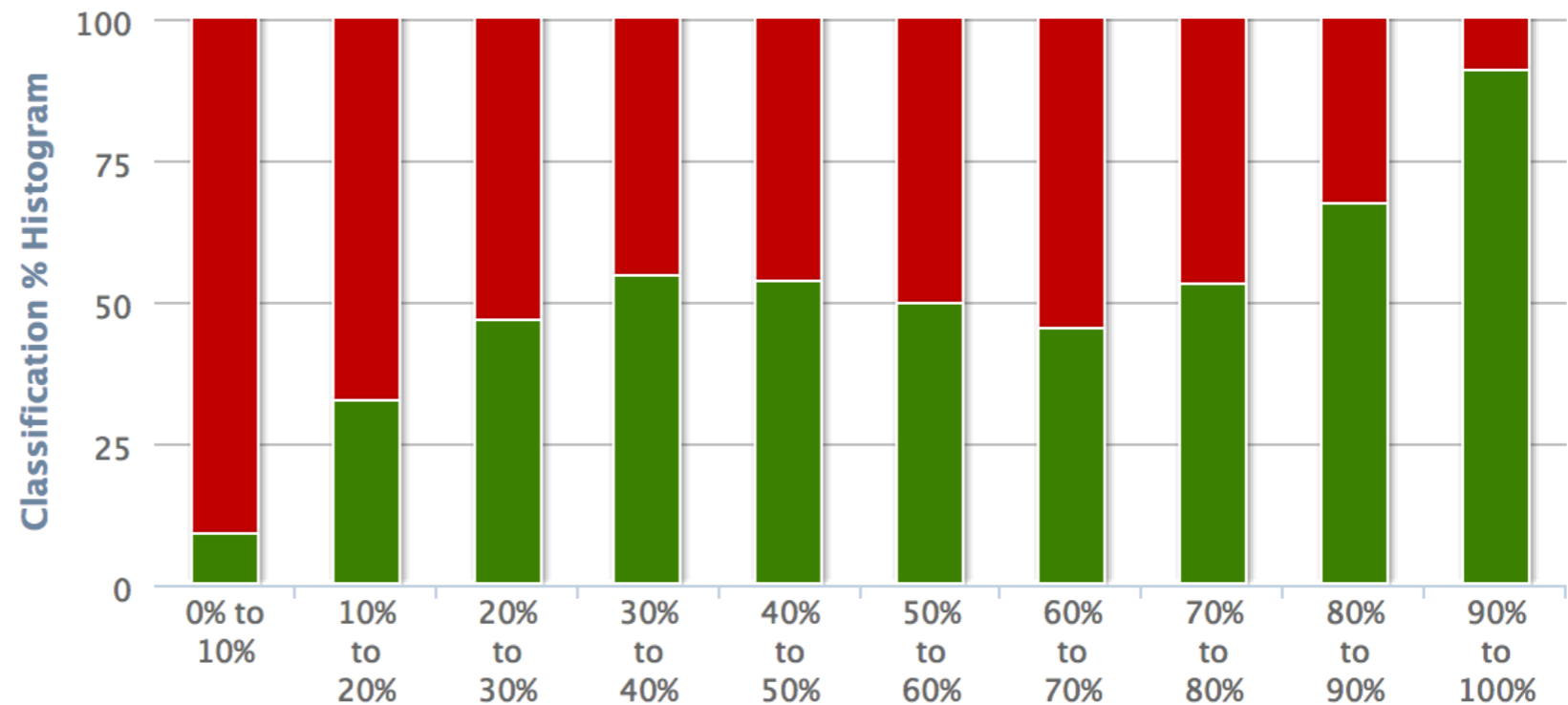
Classification Boundary Filter



Classification Histogram



- No, not Christie
- Yes, Chris Christie



Histogram Range: 0% to 100%



select all

clear all

apply filter



# sifter

Search and retrieve data from the complete history of Twitter

<http://sifter.texifter.com>

# Rules & Filtering

Before the PowerTrack API will deliver activities to you through the streaming connection, it must be configured with a set of rules. These rules are applied across the firehose of the data source, and with those activities that match being sent through your API connection.

Put simply, PowerTrack rules allow you to build a list of requirements for the data that will be sent through your stream. For example:

*I want activities where the text mentions the keyword “gnip”*

Or, for a more realistic example:

*I want activities which meet the following requirements*

- contain any of these keywords: **gnip, data, social**
- and contain any of these keywords: **love, hate, like, dislike**
- and which contain a link
- but which don't contain any of these keywords: **ping, gnop**

The above requirements can be translated into a PowerTrack rule:

```
(gnip OR data OR social) (love OR hate OR like OR dislike) has:links -(ping OR gnop)
```

## Operators

In the example above, the keywords and 'has:links' are Operators. Operators tell PowerTrack what types of activities to deliver. For example, the keyword operator 'gnip' above tells PowerTrack that an activity must contain a tokenized keyword in the text, while the has:links operator adds a requirement that the activity contain a URL in its body.

Operators may be either positive or negative.



# Thanks for Listening

Dr. Stuart Shulman

@stuartwshulman

stu@texifter.com

[discovertext.com](http://discovertext.com)

[sifter.texifter.com](http://sifter.texifter.com)