

“...a wealth of information creates a poverty of attention.”

- Herbert Simon, 1971



CoderRank: Creating Gold Standards

Dr. Stuart W. Shulman
Founder & CEO, Texifter
@stuartwshulman

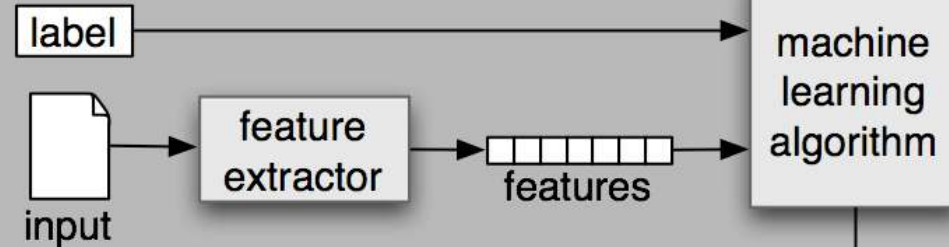
Text Classification

A 2500 year-old problem

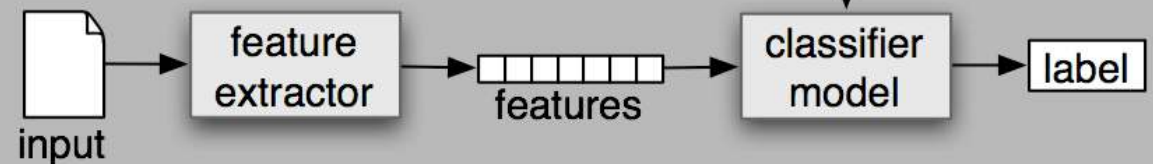
Plato argued it would be frustrating. It still is.



(a) Training



(b) Prediction



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.”

Free, Open-Source, Web-based Text Analytics Toolkit

**CAT**
Coding Analysis Toolkit ✓ Like 25

register for a free account
forgot password?

usernamepasswordlogin

Home | About CAT | DiscoverText | Terms of Service | Privacy Statement | CAT Help Wiki | Contact Us

Welcome to the Coding Analysis Toolkit (CAT)

CAT is a free service hosted by Texifter. Load, code, and annotate text data in teams. Measure inter-rater reliability and adjudicate differences between coders. Report on the accuracy of codes and coders over time. Train better coders through systematic iterations. 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 here:
[CAT Quickstart Guide](#)

To view a tutorial on using CAT, click here:
[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/>

CAT Statistics

There are currently **11,865** primary CAT accounts and **1,509** sub-accounts. CAT users have uploaded **8,839** coded datasets and **15,035** raw datasets. They have coded a total of **2,227,092** items and adjudicators have made **198,220** validation choices in CAT.

discovertext

If you like CAT, you'll love [DiscoverText](#). DiscoverText is a cloud-based, collaborative text analytics solution. Generate valuable insights about customers, products, employees, news, citizens, and more. [Sign up for a 30 day free trial](#).

What CAT DoesCAT FeaturesPraise 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](#)
- [Merging HUs in ATLAS.ti](#)

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

Original Software Kernel: Tools for Measurement

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

Avoid Tennis Elbow



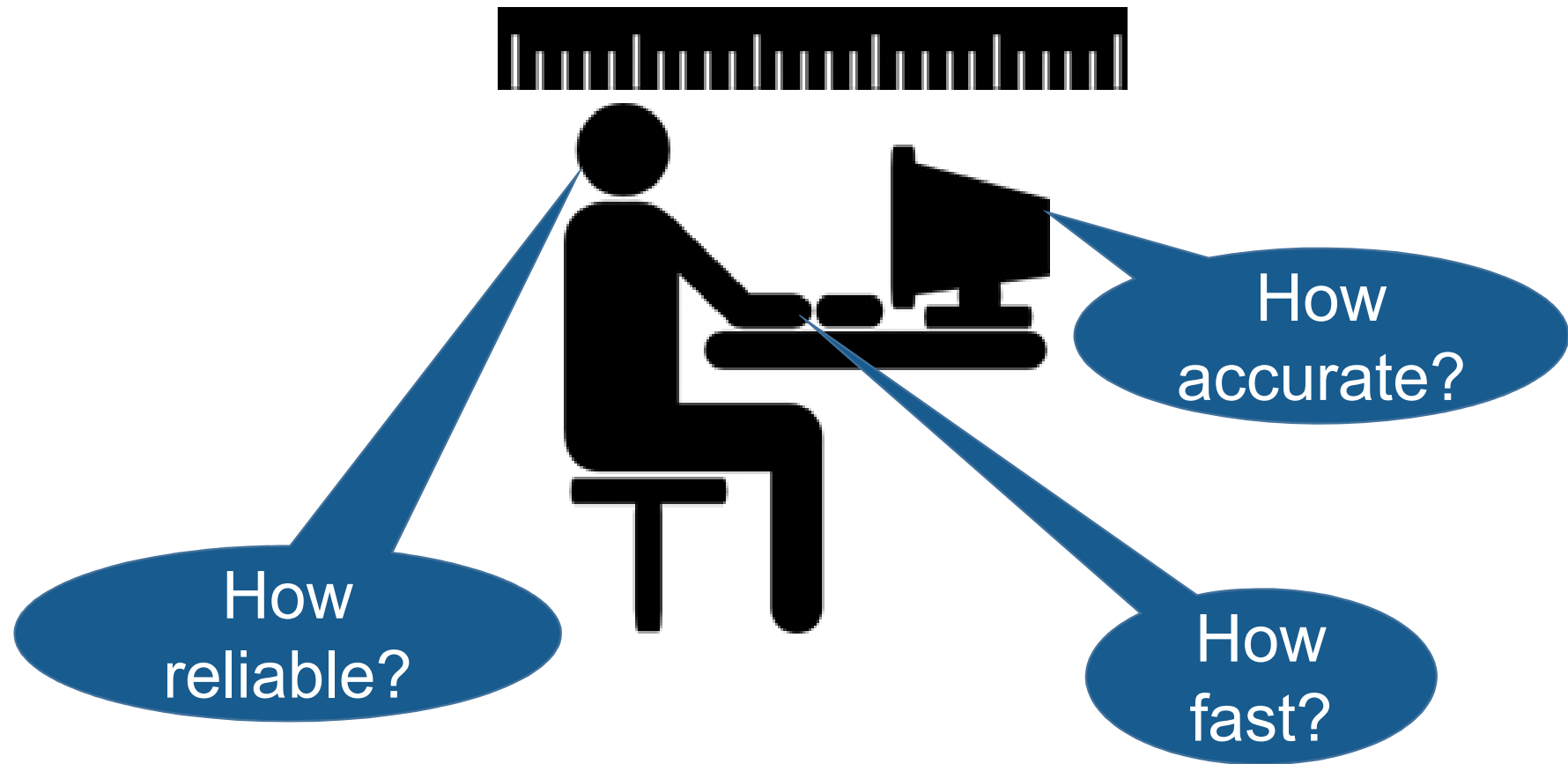
Items load to the screen and the coder hits the keystroke

Keystroke Human Coding

The screenshot displays the Keystroke Human Coding interface. At the top, there is a toolbar with various icons for file operations and editing. Below the toolbar, a header bar contains a 'Sort: Key Code' dropdown, a '# columns: 6' dropdown, a 'STOP' button, and a progress indicator '0/179 : 3'. The main area is divided into four tabs: '(1) News or Current Events', '(2) Sports', '(3) Food or Restaurants', and '(5) Other'. A tweet from 'Europafoot' is shown, with the text 'Mercato - Lyon : Un grand d'Angleterre rêve de Lacazette - europafoot.com/mercato-lyon-u... dans #Football'. Two blue callout bubbles are present: one labeled 'Codes' pointing to the tab bar, and another labeled 'Data' pointing to the tweet text.

Human coding can be distributed to individuals, groups & crowds

Computer Science & National Science Foundation: Measure Everything



Annotator Speed

Coder Stats

Total Coding Time: 02:04:30

Avg. Coding Time: 6s

Coder	Units Coded	Avg. Coding Time	Total Coding Time
Redacted	200 (100.00%)	6.52s	00:21:44
	200 (100.00%)	4.65s	00:15:30
	200 (100.00%)	10.32s	00:34:24
	200 (100.00%)	4.30s	00:14:19
	200 (100.00%)	4.30s	00:14:19
	200 (100.00%)	7.27s	00:24:14

Interrater Reliability: A Critical Measurement

Standard Comparisons

[Dataset details](#) > Standard Comparisons

Dataset: v3 Malaysia Non-Exacts

[Perform another comparison](#)

Current view: Comparison Table

Code	Coder 5981	Coder 9953	Coder 9955	Coder 2668	Coder 2279	Coder 4341	Exact Match	Partial Match	Kappa
MH-17	11	13	27	20	8	11	7	12	0.69
Not MH-17	189	187	173	180	192	189	172	17	0.96
Totals	200	200	200	200	200	200	179	29	0.94

Adjudication

Dataset: MH17 Test 2

Code: **Not MH-17**

(1) Valid

(2) Skip to next

(3) Not valid

Validations remaining: 200
[\[+\] Change Code Filter](#)

20.00% of users coded this as Not MH-17

Document Metadata

author user id: christy18
extkey: 415379051995629
inserted: 7/22/2014 9:35:32 AM
is segmented: true
itemid: 45094031055
language: Chinese - Simplified
messagetype: post
original text: MH17事后第一天，美国就宣布其一套雷达系统看到了一枚地空导弹从乌克兰境内飞向马航，还有另一套看到了马航被击中的热信号，如此精确和反应迅速的雷达系统.....在几个月前，为啥却看不到马航370.....一架在空中飞了8个小时.....比导弹大那么多倍的777.....去哪了？
postsinthread: 1
published: 7/22/2014 9:34:37 AM
threadid: 415379051995629
url: http://t.qq.com/p/t/415379051995629

Coder choices

Coder	Codes
Coder 5611	MH-17
Coder 0043	MH-17
Coder 5981	MH-17
Coder 9783	MH-17
Coder 2279	Not MH-17

MH 17 事后第一天，美国就宣布其一套雷达系统看到了一枚地空导弹从乌克兰境内飞向马航，还有另一套看到了马航被击中的热信号，如此精确和反应迅速的雷达系统.....在几个月前，为啥却看不到马航370.....一架在空中飞了8个小时.....比导弹大那么多倍的777.....去哪了？

Dataset: MH17 Test 1

[View statistics on other datasets](#)

Total Valid Answers: 1010 / 1209 (83.54%)

Validations by Coder:

Coder	Valid Answers
Coder 3510	4 / 4 (100.00%)
Coder 9817	197 / 200 (98.50%)
Coder 5611	196 / 200 (98.00%)
Coder 0043	195 / 200 (97.50%)
Coder 9783	169 / 200 (84.50%)
Coder 5981	124 / 200 (62.00%)
Coder 2279	122 / 200 (61.00%)
Coder 6549	3 / 5 (60.00%)

Validations by Code:

MH-17	672 / 681 (98.68%)
Not MH-17	338 / 528 (64.02%)

Dataset: MH17 Test 2

[View statistics on other datasets](#)

Total Valid Answers: 734 / 800 (91.75%)

Validations by Coder:

Coder	Valid Answers
Coder 5611	197 / 200 (98.50%)
Coder 9783	184 / 200 (92.00%)
Coder 0043	179 / 200 (89.50%)
Coder 5981	174 / 200 (87.00%)

Validations by Code:

MH-17	564 / 567 (99.47%)
Not MH-17	170 / 233 (72.96%)

 **coderrank**

Patent issued March 1, 2016

CoderRank for enhanced machine-learning is our key innovation

 **discovertext**



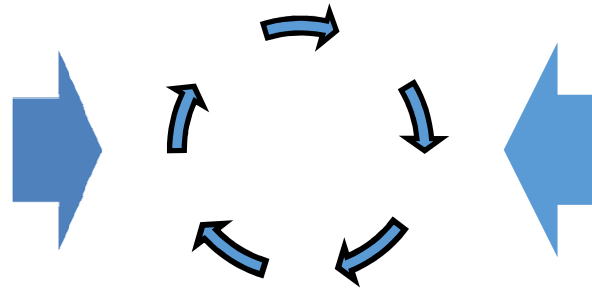
CoderRank for Enhanced Machine-learning

CoderRank is to text analytics what PageRank was to search. Just as Google said not all web pages are created equal, Texifter argues that not all humans are created equal. When training machines, it is best to rely most on the humans most likely to create a valid observation. We proposed **a unique way to rank humans on trust and knowledge vectors.**

Active Learning engines and human coding tools combine...



what humans do best...



with what computers do best.

Humans and machines learning together

It is always good to keep humans “in-the-loop”

Word sense disambiguation (relevance)



Word sense disambiguation (relevance)



Word sense disambiguation (relevance)



Word sense disambiguation (relevance)



Yes



No

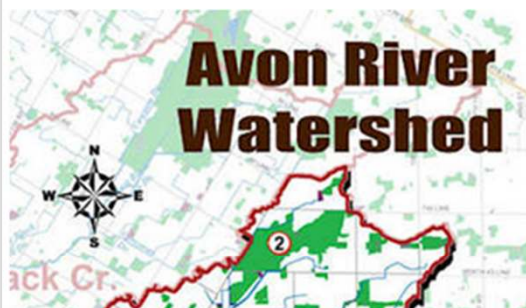


No

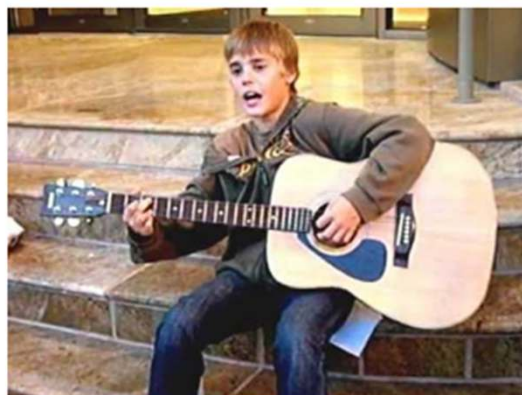
AVON

the company for women

Yes

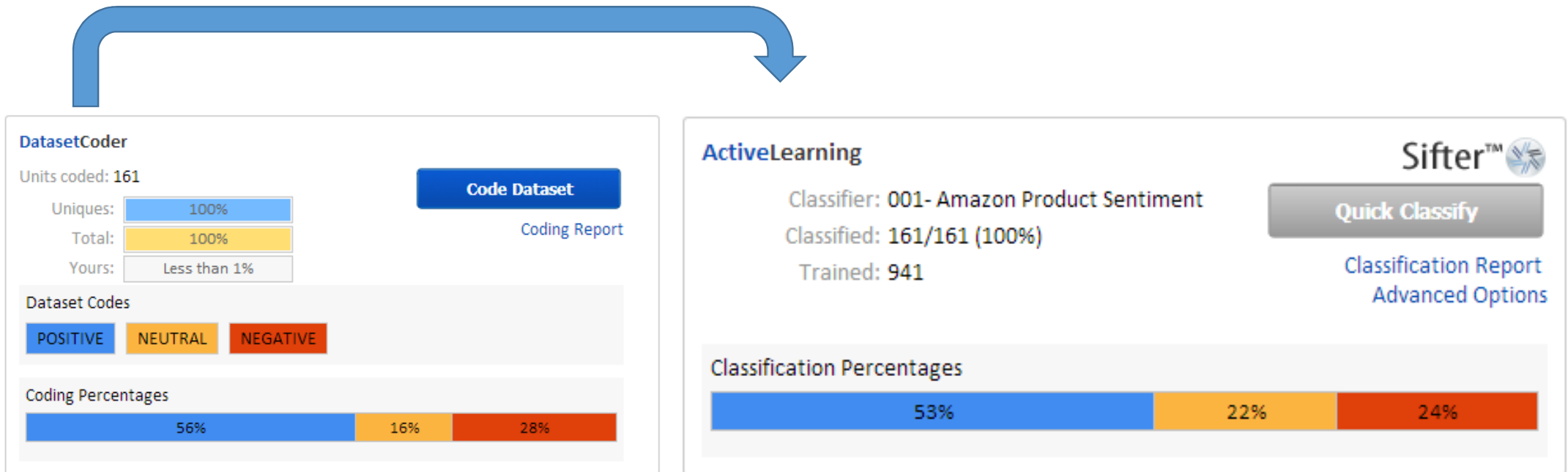


No



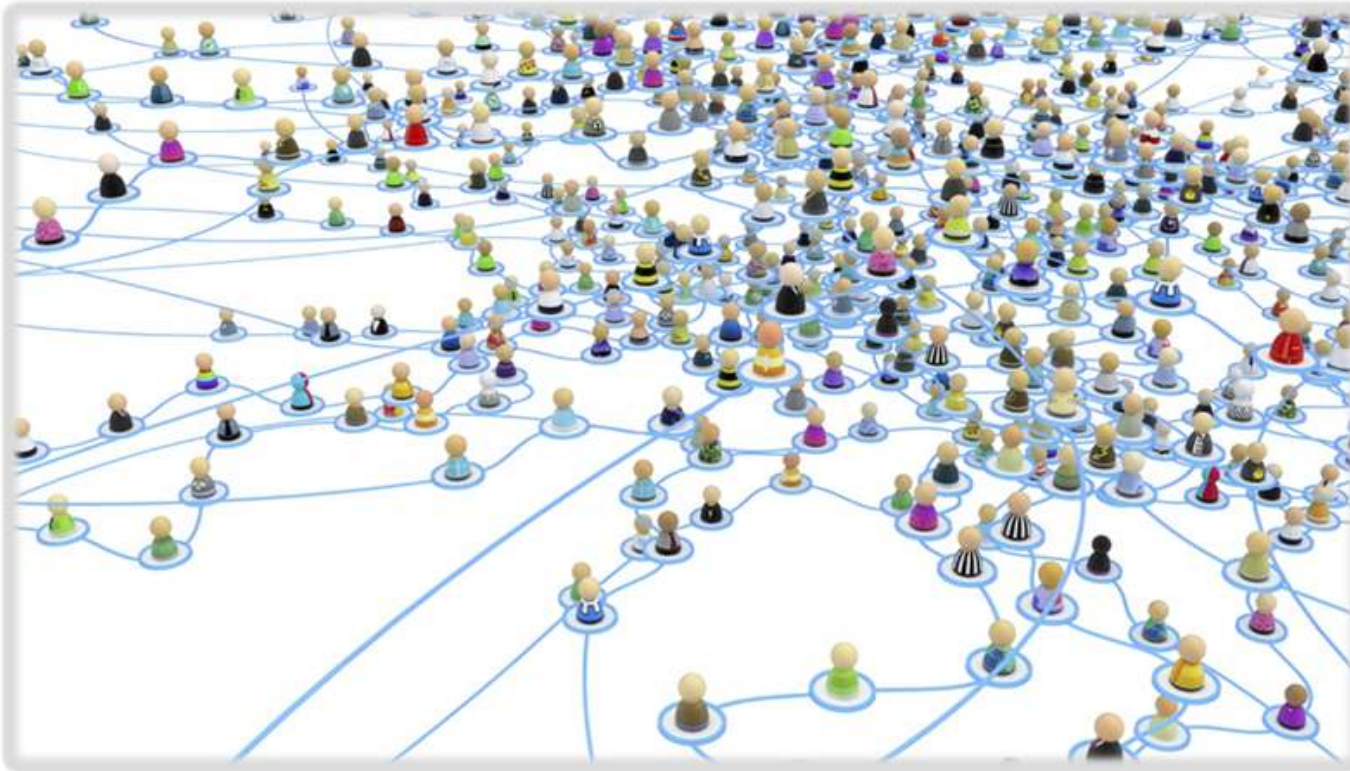
No

Human coding converts into machine classifiers



Accumulated human coding becomes training data via machine-learning

Distribute coding for synchronous & asynchronous collaboration



Crowdsourcing accelerates the insight generation

Thank-you for listening!

@stuartwshulman
stu@texifter.com