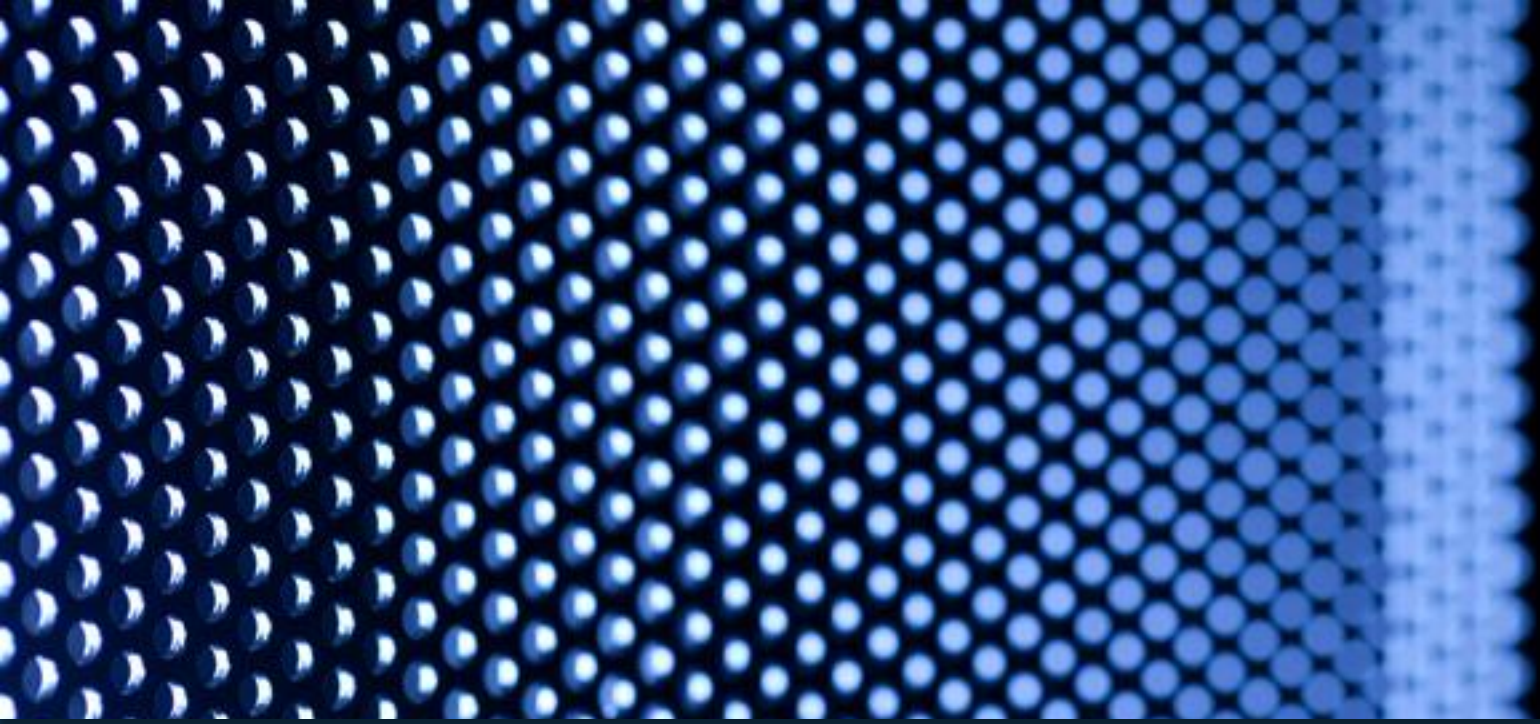


VERISIGN

Analyzing the Keystroke Dynamics of Web Identifiers

Andrew G. West

Verisign Labs

Presented at: ACM Web Science 2017 – Troy, NY, USA

BIG IDEA + TAKEAWAYS + OUTLINE

IDEA: Analyze **keystroke dynamics** of **web identifiers**
(typing speed, accuracy, latencies) (domains, usernames, hashtags, etc.)

METHOD: Crowdsource identifier typings via MTurk

Broad measurements

- Typing time *grows linearly* with identifier length (≈290 ms/char.)
- *Familiarity* increases type-ability for identifiers and extensions

Features/models predicting type-ability

- Keyboard topology is a poor predictor of type-ability
- Ease of *identifier tokenization* and *bi-gram composition* are better
- Models tepidly effective; one σ is 2.8 secs. for 12-char. identifier

Tokenization clues in typing latency

- *Maximum intra-character latency happens on word boundaries*
  3× to 4× more often than by random chance

MOTIVATIONS + APPLICATIONS

Shared props. of web identifiers/namespaces:

- Uniqueness (availability) requirements
- Desire for short memorable strings
- Absence of delimiters (e.g., spaces)

FairWinds Partners: "… of the top 250 most highly trafficked websites ...typosquatting costs … those sites **$364 million and 448 million impressions per year**" (2010)

YOU'VE GOT OPTIONS

10 tips for choosing the perfect domain name

By Andrea Rowland

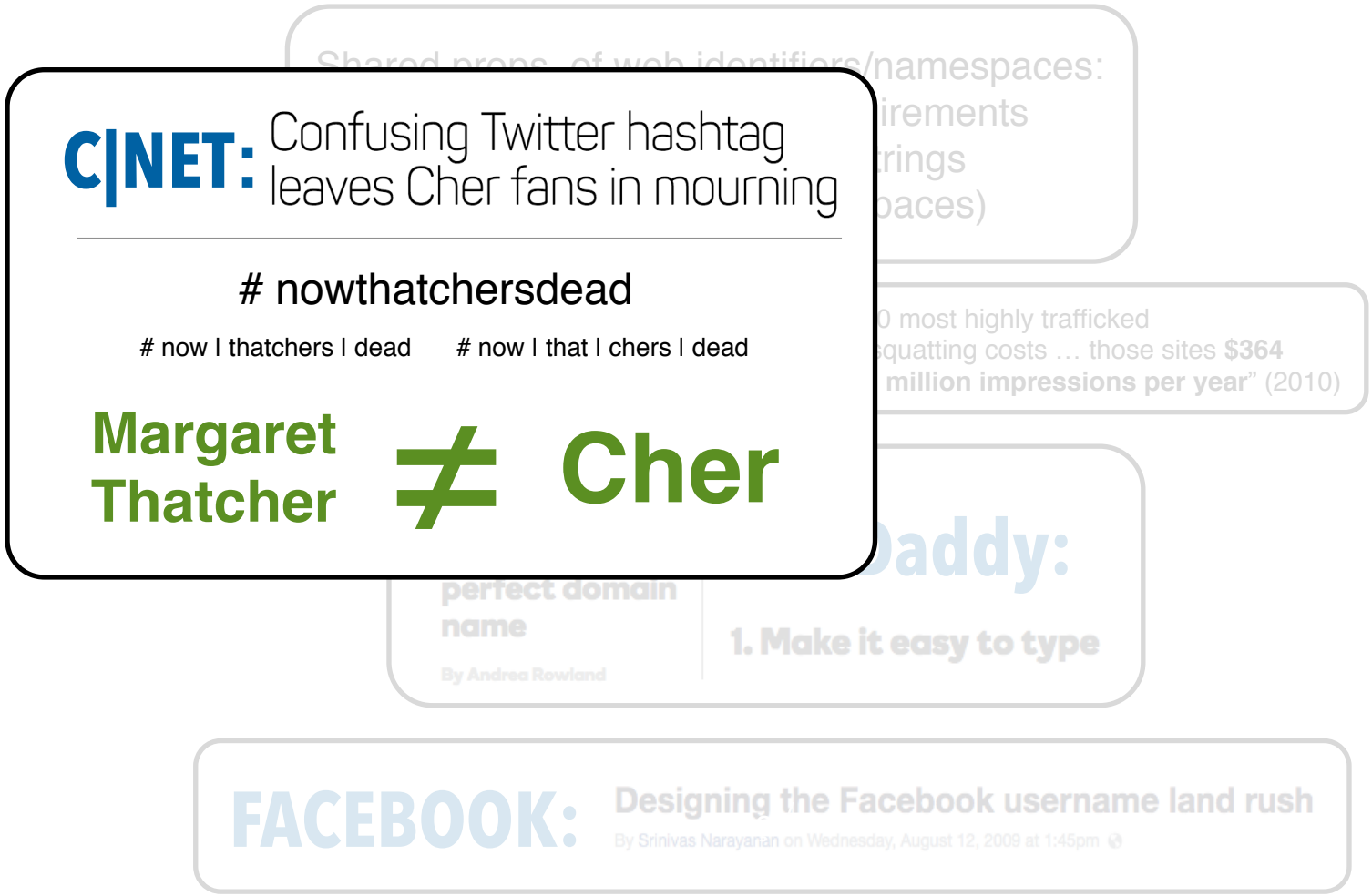
GoDaddy:

1. Make it easy to type

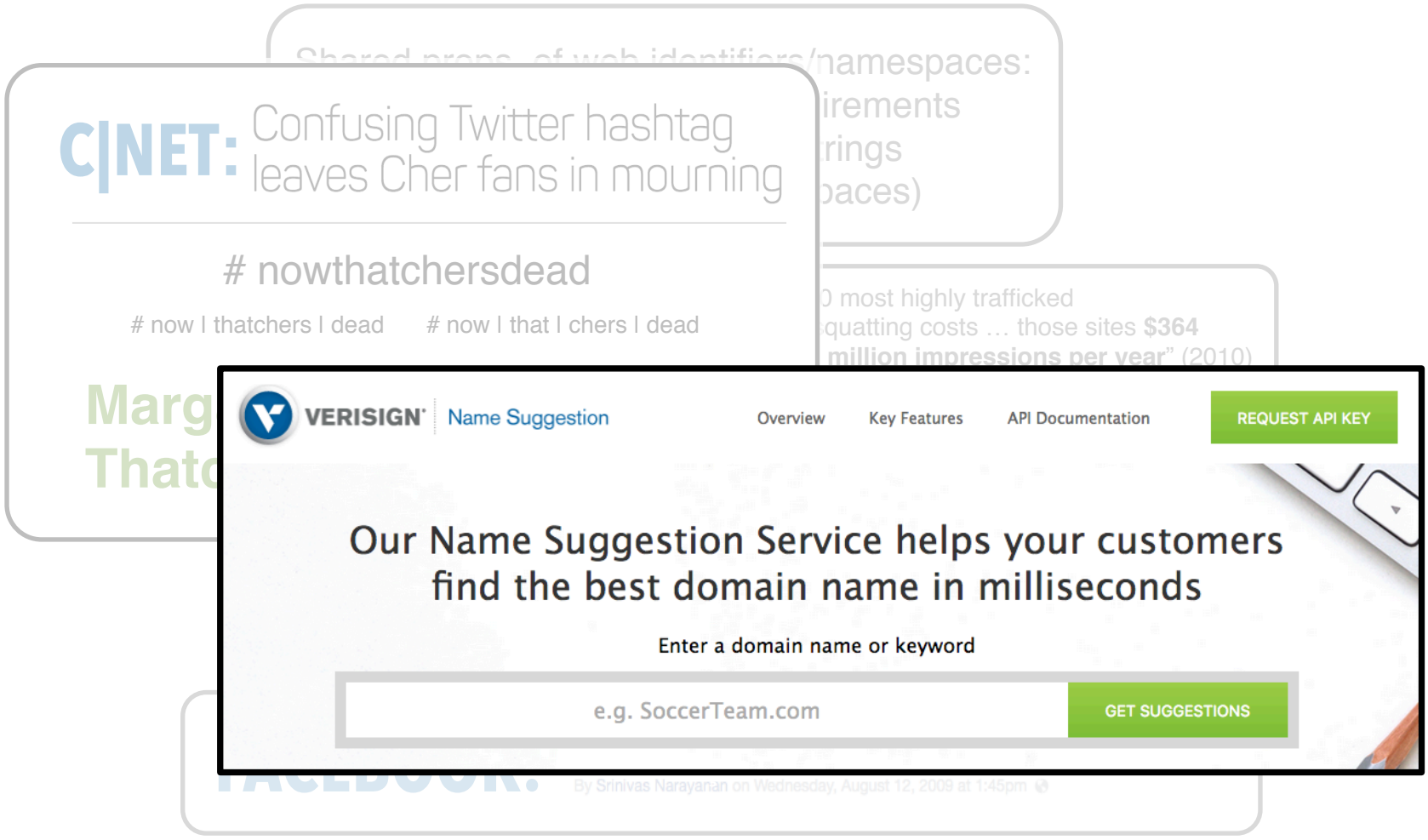
FACEBOOK: **Designing the Facebook username land rush**

By Srinivas Narayanan on Wednesday, August 12, 2009 at 1:45pm

MOTIVATIONS + APPLICATIONS



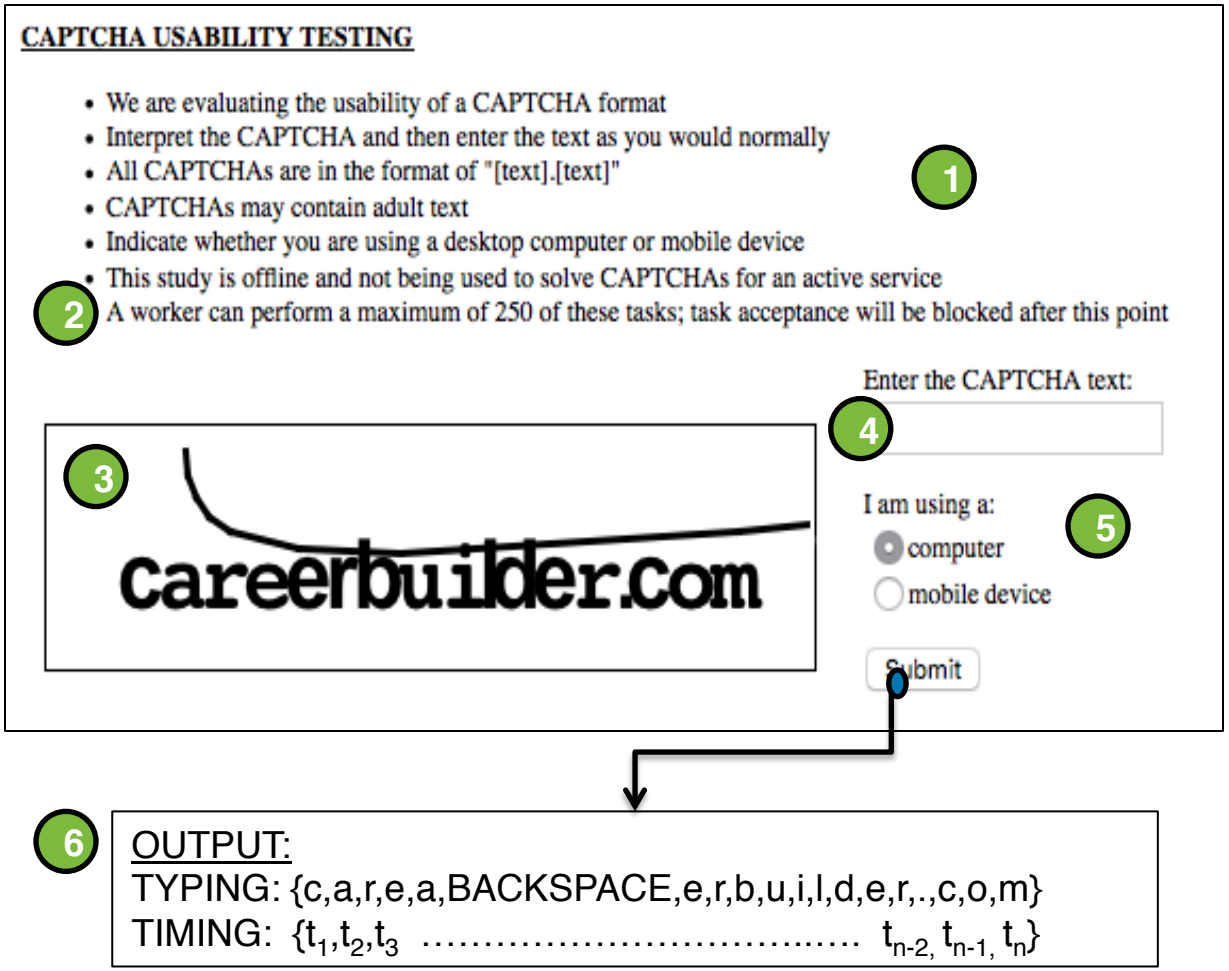
MOTIVATIONS + APPLICATIONS



Data collection

Crowdsourcing identifier typings via Mechanical Turk

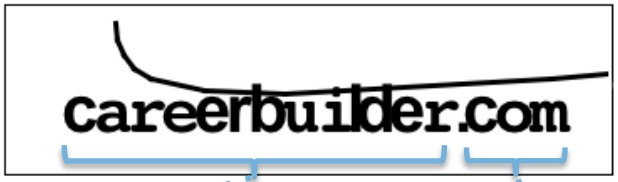
EXPERIMENTAL DESIGN: CAPTCHAS



1. Instruction set
2. Participation limit
3. CAPTCHA image
4. Solution form
5. Device type
6. Output

IDENTIFIERS TO BE TYPED

Why domain names?



Second-level domain (SLD)

Top-level domain (TLD)

SEGMENTATION CORPUS:
21,896 domains that multiple humans visited and *consensus* tokenized: [adomain.com] -> {a | domain}

RANDOM
14,650 doms.

COM/NET where:
- Not foreign script
- Not parked
- Not redirected
- Has content

POPULAR
7,250 doms.

Domains from "Alexa Top 10k"

These are domains we'd expect to be type-in traffic!

TLDs[1] are randomly assigned to corpus SLDs with following probabilities:

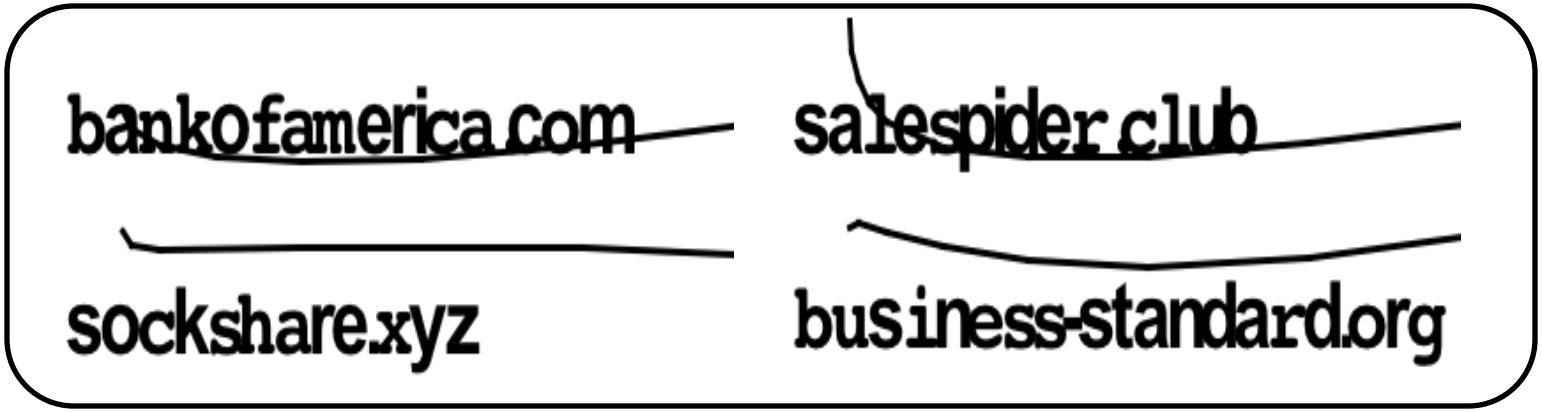
20% - COM	Legacy (40%)
10% - NET	
10% - ORG	
10% - DE	ccTLD (30%)
10% - CN	
10% - UK	
10% - XYZ	New gTLD (30%)
10% - CLUB	
10% - ONLINE	

[1] These TLDs were selected to compare the most popular extension of each type per industry statistics (www.ntldstats.com; Verisign 07-2016 DNIB)

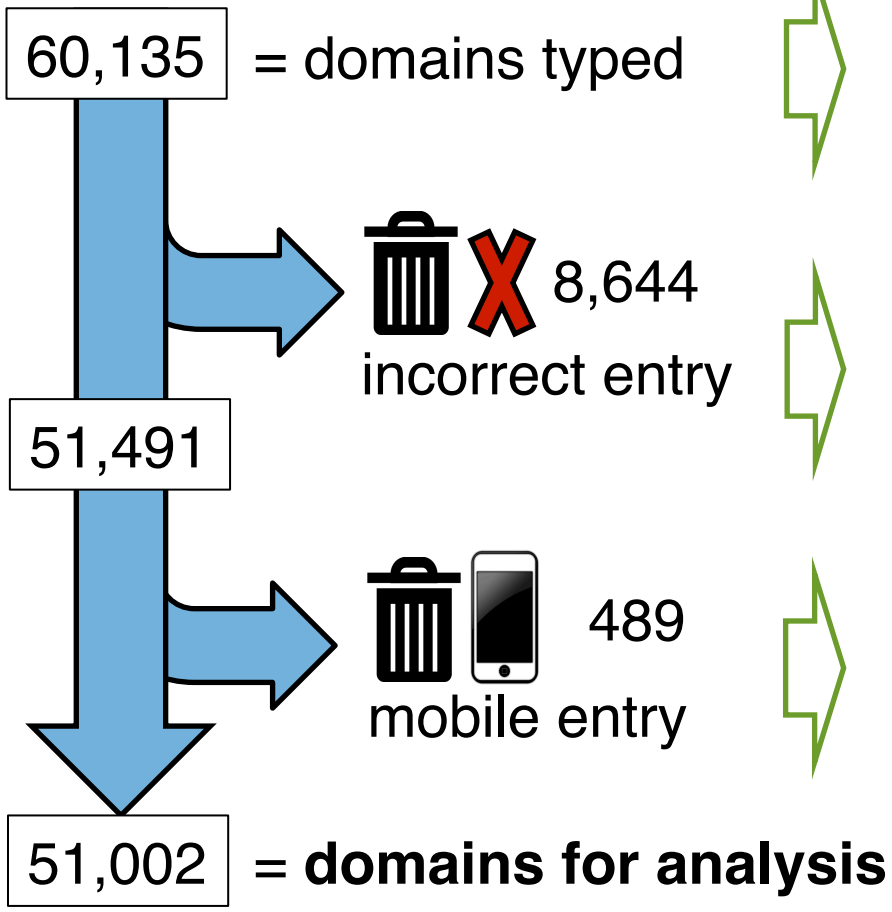
EXAMPLE STRINGS + CAPTCHAS

Experiment randomly chooses from 22k CAPTCHA to display

umclimbing	norwegianblues	monicaandbryce	2300fannin
oaklandbakery	makehomenow	montserrattreviews	cozycatboarding
philsmithroofing	lincolncitybuilder	enkipools	nationalcarparks
letsbuytickets	goldmedalplumbing	stlouisvideoeditor	greenbuildingguide
stadiumflooring	killnoise	javahouserealty	ecoartforchange
davidreedagency	littleavenuecakes	bestmetrocarpetcln	yogurtville



DATA FILTERING



Mechanical Turk
- $200 per 10k typings
- $1250 cost all-in

Final text != CAPTCHA
- Typos
- “Dot” exclusions
- Technical issues
- Backspaces OK

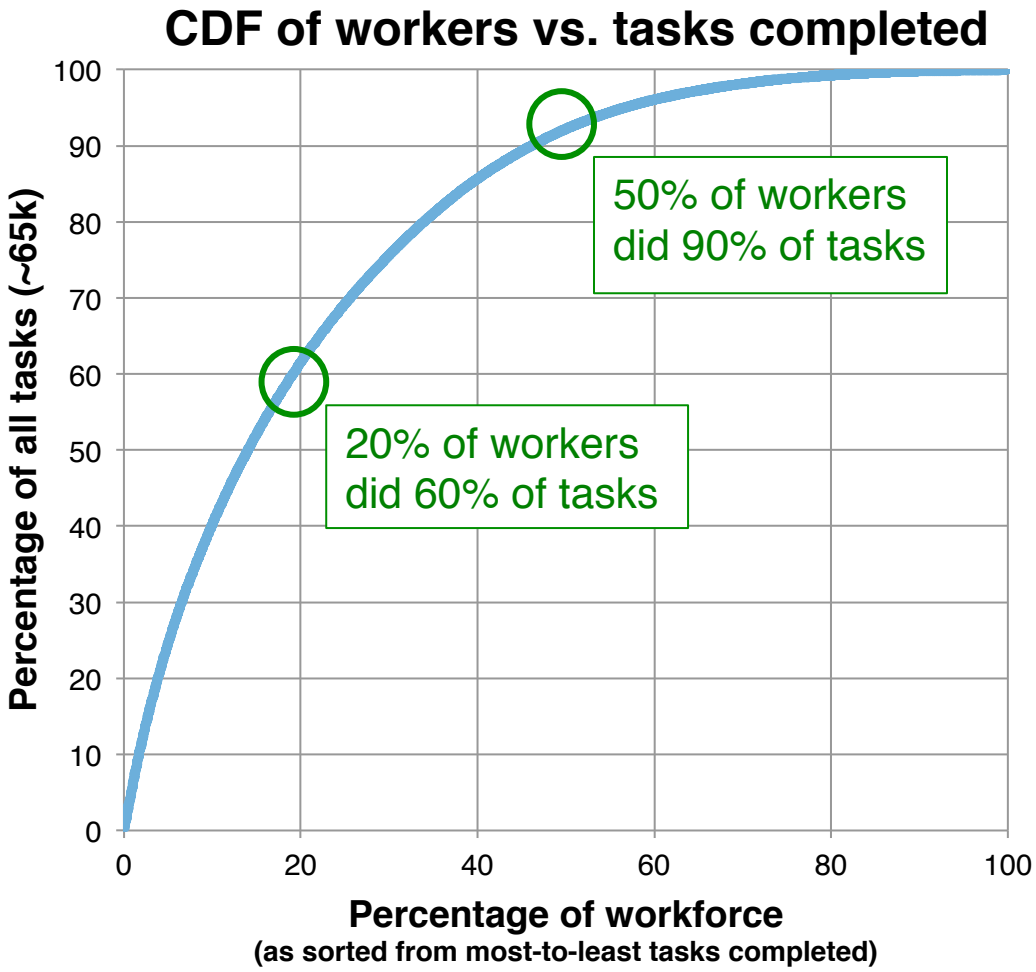
Mobile input devices
- Self-reported
- UserAgent string

WORKER POPULATION + BIASES (1)

1,442 unique workers

- Median = 21 tasks
- Average = 41.7
- 20 workers did max. 250 tasks
- MTurk generally representative of Internet pop.[2,3]

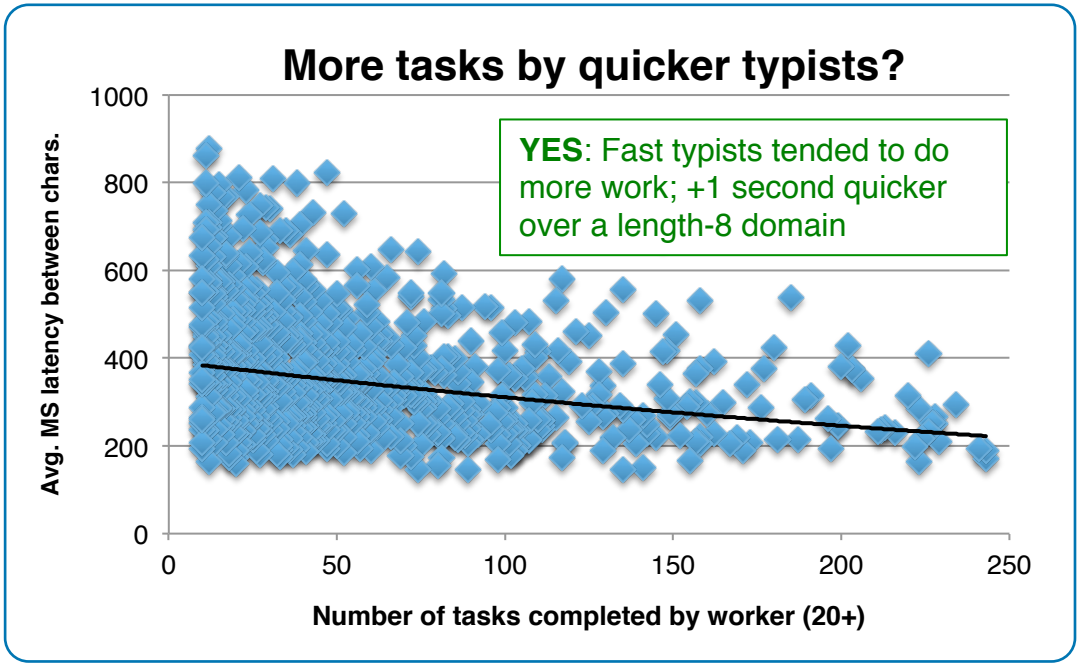
[2] http://demographics.mturk-tracker.com/
[3] http://blog.turkprime.com/2015/03/the-new-new-demographics-on-mechanical.html



WORKER POPULATION + BIASES (2)

Economics of MTurk:

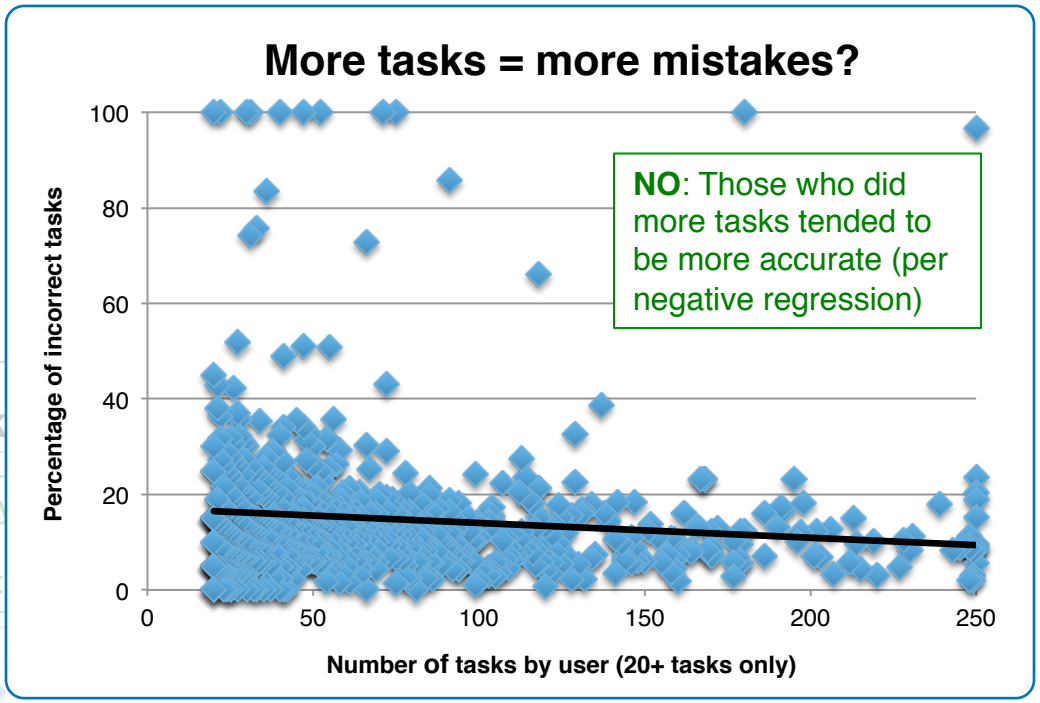
- Task self-selection
- Finite work pools
- Optimizing payout
- Worker reputation

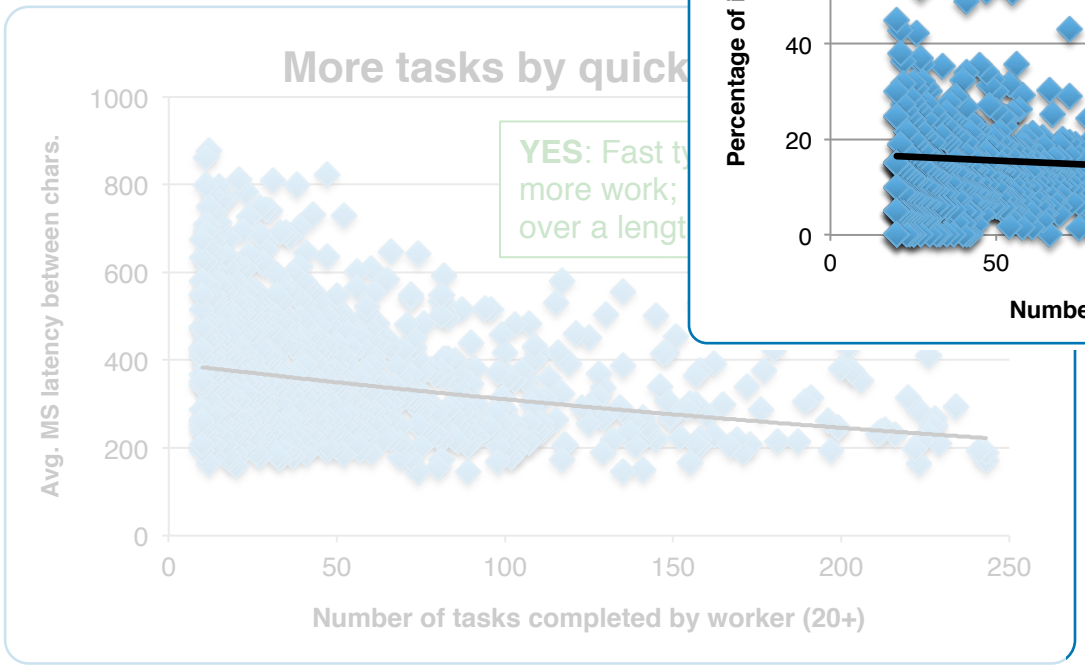


WORKER POPULATION + BIASES (2)

Economics of MTurk:

- Task self-selection
- Finite work pools
- Optimizing payout
- Worker reputation



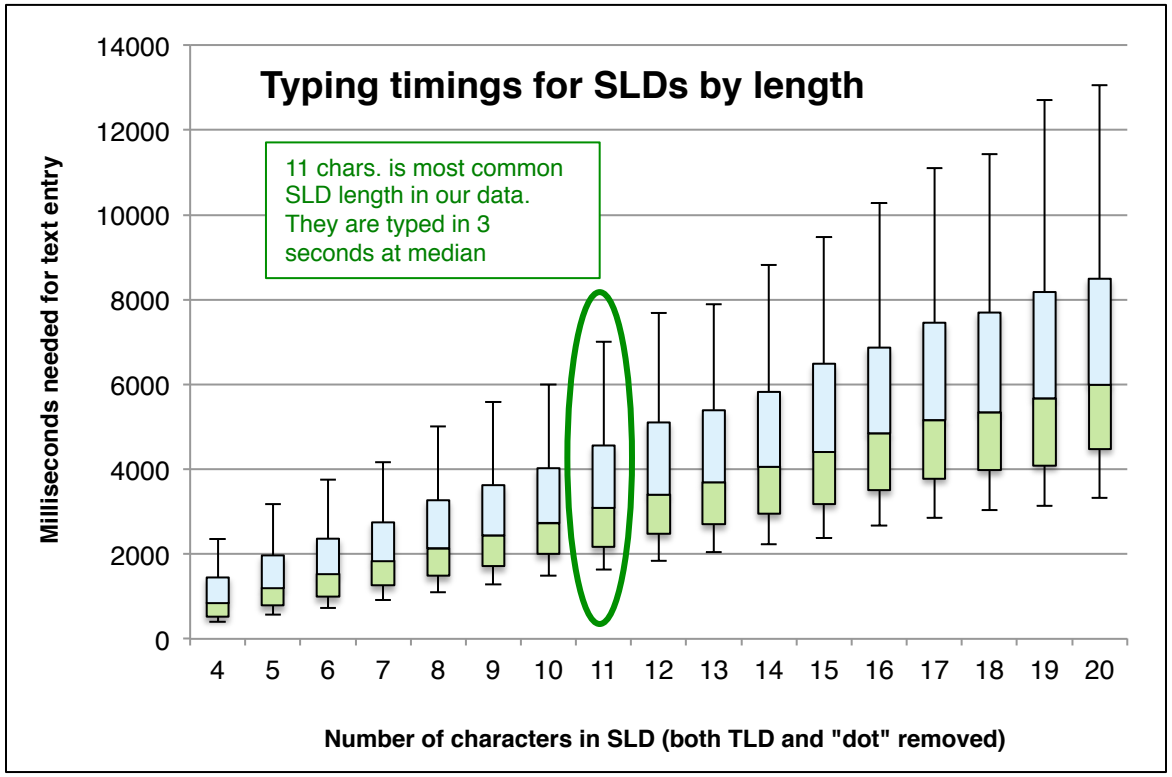


Analysis: Broad measurements

Focusing on domain SLD and TLD type-ability

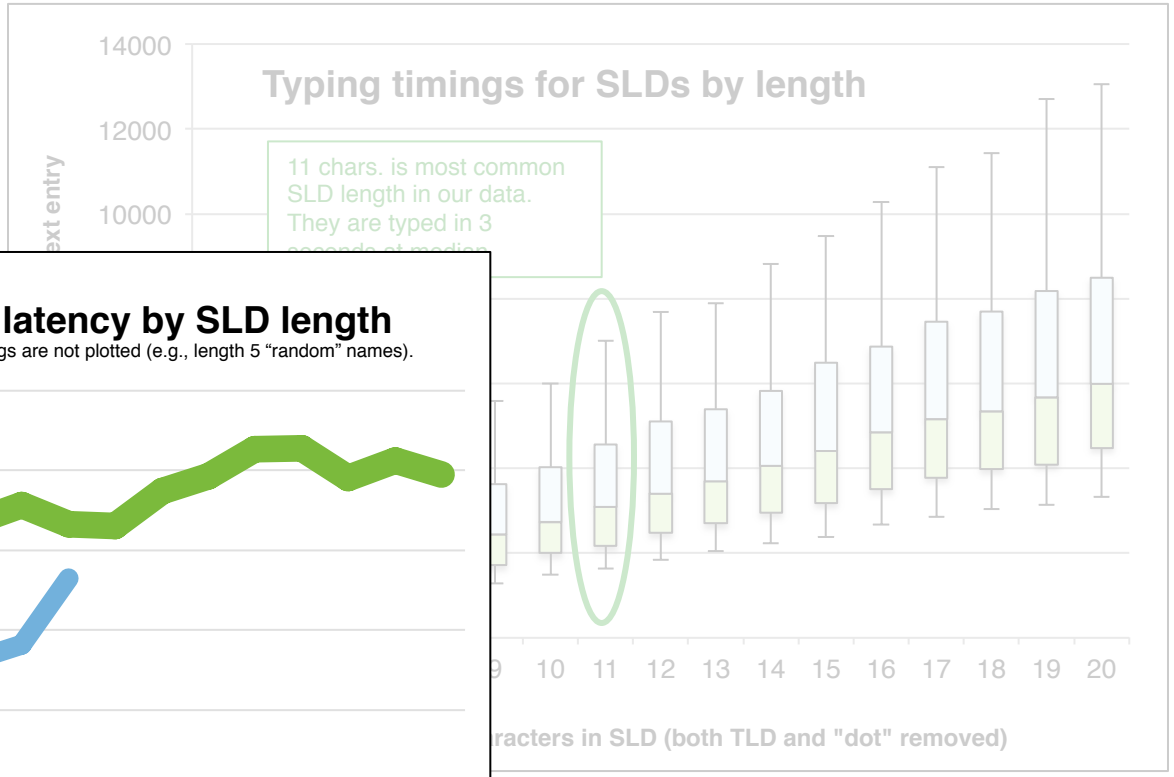
SLD (LEFT-OF-DOT) TYPING TIMES

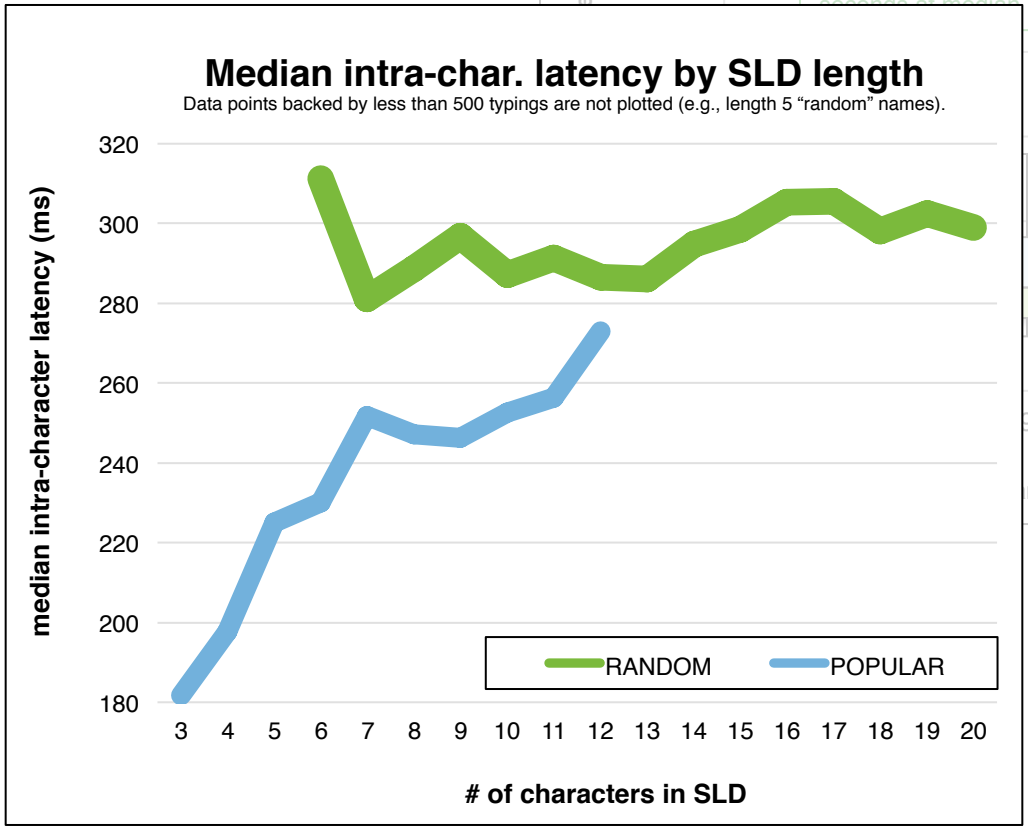
TAKEAWAY: Sanity check; longer identifiers take longer to type



SLD (LEFT-OF-DOT) TYPING TIMES

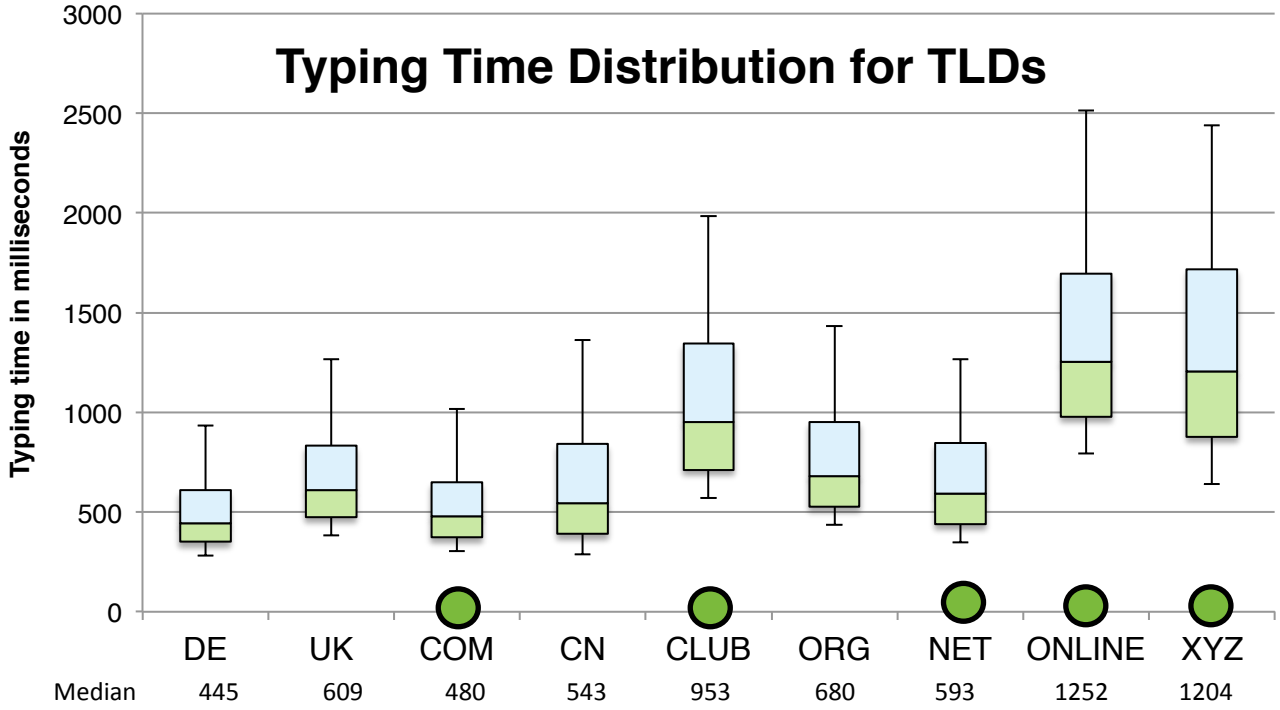
TAKEAWAY: Sanity check; longer identifiers take longer to type

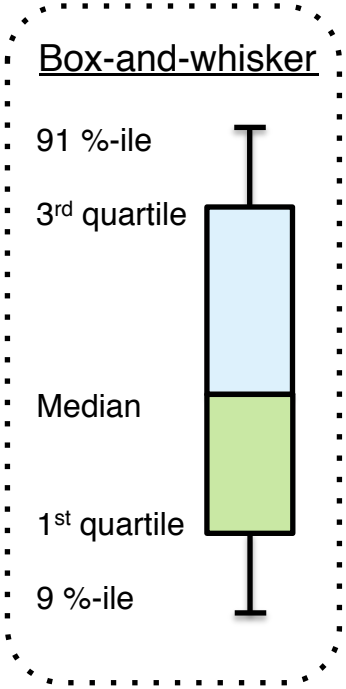




TAKEAWAY: *Linear* growth in "random" typing times reveal no hidden costs in seeking out longer names in mature namespaces

TLD (EXTENSION) TYPING TIMES





TAKEAWAY : Pervasive legacy TLDs (COM; NET) are typed at speeds comparable to 2-char. ccTLDs; muscle memory?

TAKEAWAY : New gTLDs tend to be slowest (even when length-normalized); perhaps reflecting lack of awareness[4]

[4] ICANN. Global Registrant Survey, https://newgtlds.icann.org/en/reviews/cct/global-registrant-survey-15sep16-en.pdf

SLD + TLD BACKSPACE (BS) STATISTICS

PROPERTY	ALL
# SLDs	51,002
SLDs w/backspace	8,229
Backspace (BS) presses	19,284
Avg. SLD length	12.0
% SLDs w/BS	16.1%
Actual/min keypresses	106.1%
BS/actual presses	3.0%

TAKEAWAY: Errors are not uncommon with 16% of SLD typings using backspace; hinting at need for typosquatting protection

TLD	Chars./BS (normalized)	TLDs per backspace
COM	105.88	35.29
NET	61.98	20.66
DE	48.50	24.25
ORG	44.52	14.84
UK	35.54	17.77
ONLINE	34.41	6.88
CLUB	33.99	8.50
XYZ	19.08	6.36
CN	16.19	8.10

TAKEAWAY: Pervasive legacy TLDs are most accurately typed [e.g., a user will only need backspace once per 35 typings of ".COM" (106 chars.)]

Analysis: Features/models predicting typeability

Can we predict the typing time of a string?

PREDICTING TYPING TIMES

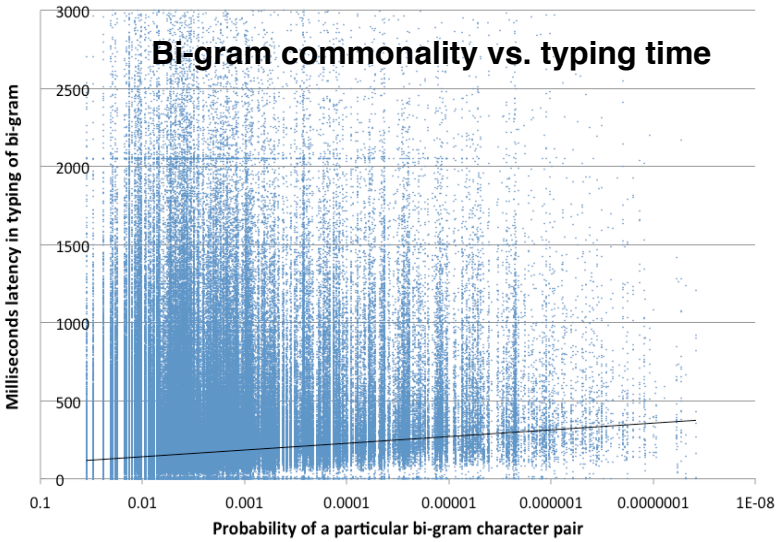
Predicted value is ratio on (0, ∞) that captures how fast a domain was typed *relative to specific typist's speed* on other domains. Value = 1.0 implies a median degree of typing difficulty.

Features for typing time prediction; RRIF = Regression ReliefF for feature efficacy

	FEATURE	RRIF	DESCRIPTION	
	distance	0.04	Finger travel distance (meters)	Keyboard topology
	same_hand_%	0.10	Percentage key transitions on same hand	
	same_fing_%	0.04	Percentage key transitions on same finger	
★	row1_%	0.16	Percentage key transitions to/from numeric keys	
	repeat_%	0.04	Percentage key transitions back to same character	
	pinky_%	0.10	Percentage keys typed with pinky finger	
	dom_length	0.05	Length of the SLD in characters	String properties
	seg_words	0.06	Number of words in SLD; per corpus	
★★	seg_diff	0.49	Ease of tokenizing SLD; per algorithm	
★	2gram_prob	0.16	Log-sum of bigram probabilities in SLD	

MOST INDICATIVE FEATURES

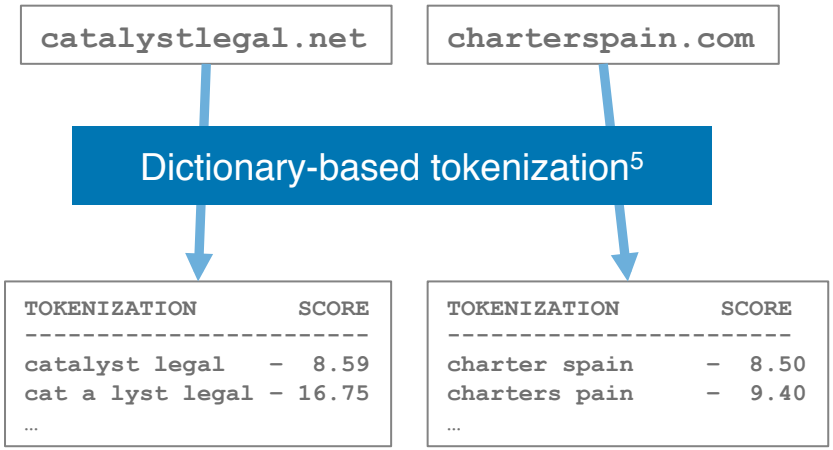
TAKEAWAY: Typeability can be improved by choosing longer keyword-rich strings rather than introducing numbers, unfamiliar abbreviations, atypical letter patterns, or ambiguity to produce available and/or shorter identifiers.



English bi-gram commonality (per Google n-gram corpus) is predictive of intra-character typing latency

TH	HE	IN	ER	…	WQ	QZ
3.5%	3.1%	2.4%	2.0%	[snip 670]	0.0%	0.0%

Ambiguity in performing tokenization

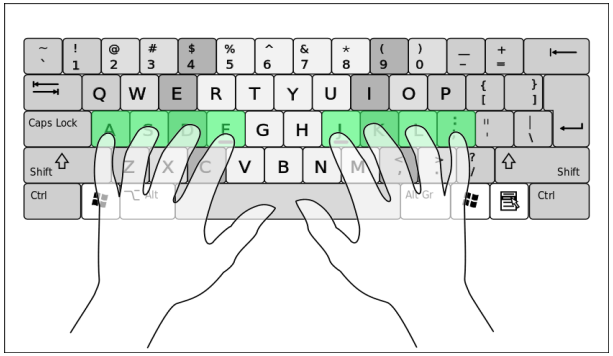


Ratio of scores between most probable tokenization and 2nd most probable tokenization captures tokenization ambiguity

[5] Jeremy Kun. "Word Segmentation, of Makingsenseofthis". https://jeremykun.com/2012/01/15/word-segmentation/

TYPING MODEL PEFORMANCE

- Built model from all features using Support Vector Regression (SVR)
- Evaluated using 10-fold cross-validation
- **TAKEAWAY**: Middling performance
  - RMSE = 0.3132
  - 12-character domain typically typed in 4 seconds. If model predicts a median typing time (output=1.0), then 4.0 ± 1.4 = 2.6 to 5.4 seconds would be one σ
  - Admittedly, not the strongest result
  - Useful in tagging outliers
- Can backspaces/typos be predicted?
  - Re-using features to predict if backspace will occur in a string barely outperformed random chance
  - **TAKEAWAY**: Predicting *where* an error will occur and *what* character mis-strike will occur appears very challenging
  - Might this be typist specific?




```
+0.7254
+0.6537 * (norm) row1_%
+0.5336 * (norm) 2gram_prob
-0.3262 * (norm) seg_diff
+0.1214 * (norm) distance
+0.0825 * (norm) dom_length
+0.0781 * (norm) pinky_%
+0.0752 * (norm) same_fing_%
+0.0222 * (norm) same_hand_%
```

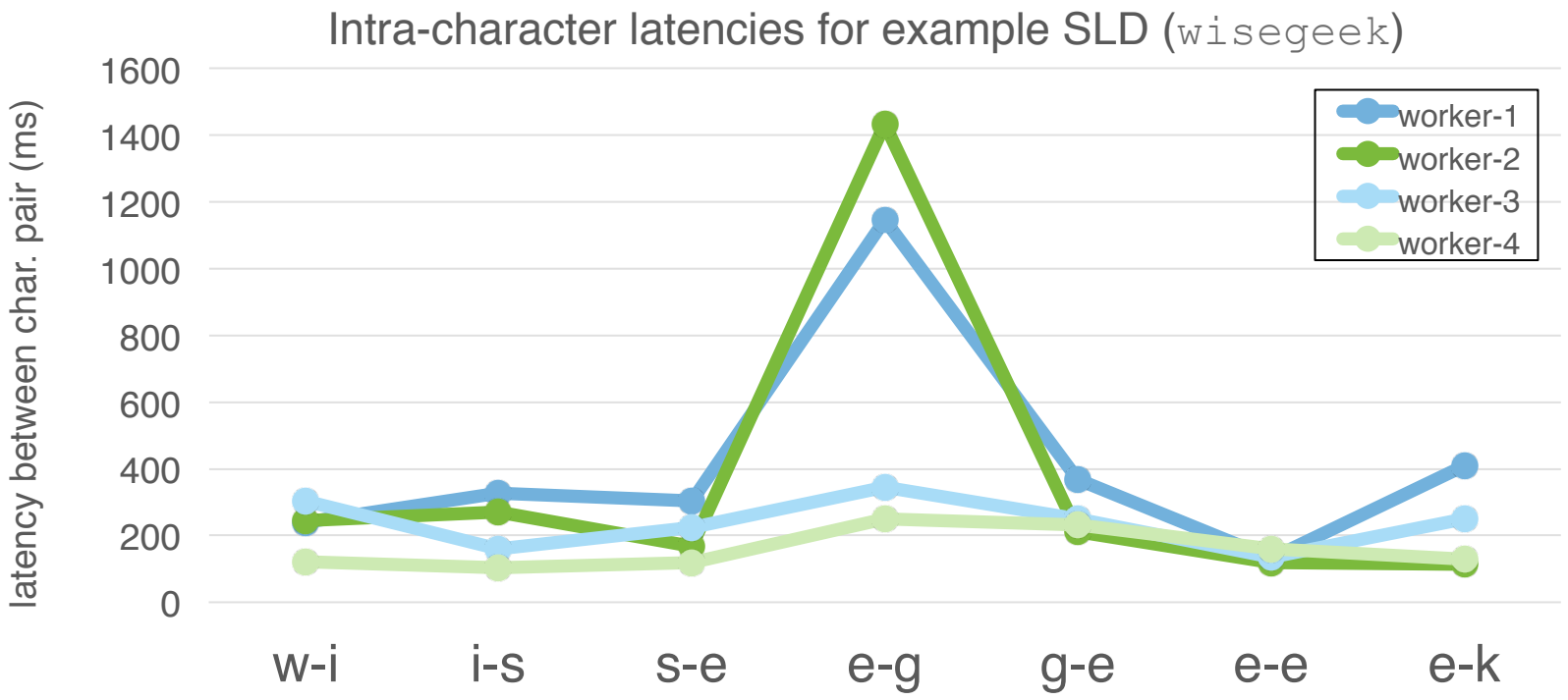
Linear regression model of typing time prediction

Analysis: Tokenization clues in typing latency

Are subconscious pauses indicative of word boundaries?

WORD BOUNDARY PREDICTION VIA LATENCY

- **HYPOTHESIS**: Users subconsciously insert brief (millisecond to second granularity) pauses in typing cadence when they knowingly encounter a word boundary in an identifier.



TESTING THE LATENCY HYPOTHESIS

TAKEAWAY: Maximum intra-character latency occurs on a word boundary 3× to 4× more often than random chance would suggest

PROOF-OF-CONCEPT

- Take all SLDs like “wisegeek”
  - Two words; humans and algorithms agree on tokenization; algorithmic tokenization is confident (> median) ; no numbers or hyphens; where human typed w/o BS
- Does the largest intra-char. latency match the word boundary?
  - Yes, 41.6% of the time
  - Lift of 4× over random chance (9.9%; calculated from SLD lengths)

TOUGHER SETUP

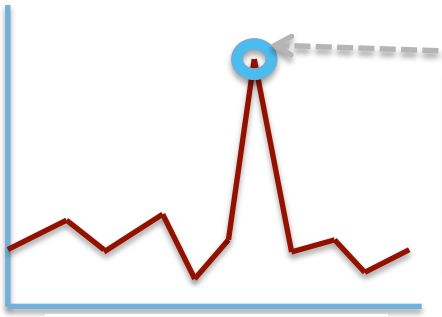
- Computational tokenization confidence is in lowest quartile
  - Latency accurate 31.5% of time
- Authoritative human corpus disagrees with CPU tokenization
  - Latency accurate 34.0% of time

ONLY PoC; challenges remain

LATENCY HYPOTHESIS APPLIED

Instrumented form

Inter-keystroke latency



characters typed

Statistical tokenizer (dictionary-based)

SPLIT – LIKELIHOOD

---

choose I spain – 50%
chooses I pain – 50%

Resolve

Process graph: Identify local maxima

Large temporal pause between characters “e” and “s”

choose I spain

NameSuggestion

Spinning on, e.g.: “Barcelona”, “espana”, “espanol”, etc.

chooses I pain

NameSuggestion

Spinning on, e.g.: “Tylenol”, “massa[illegible]”, “painkiller”, etc.

Future work + conclusions

FUTURE WORK + CONCLUSIONS

- Future work
  - Text entry on mobile devices
  - Role of type-in traffic / search engines / auto-correct
- 51,000 crowd-sourced web identifier typings revealed:
  - Typing times grow linearly with identifier length
  - Familiar domain extensions are quickest and most accurately typed
  - Domains that are easily tokenized offer typeability benefits
  - Models can predict typing times with moderate accuracy
  - Typing latencies offer clues about identifier tokenization
- Applications
  - Takeaways for name consumers + marketing departments
  - Algorithmic foundations for those who distribute/suggest identifiers

VERISIGN®