Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms

Saranga Komanduri

Patrick Gage Kelley, Michelle L. Mazurek, Richard Shay, Tim Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Julio López

Carnegie Mellon

CyLab Usable Privacy and Security Laboratory
http://cups.cs.cmu.edu/

Recent Data Breaches

	Affected users
Gawker	1,300,000
Sony	25,000,000
Battlefield Heroes	550,000
Sega	1,300,000
Booz Allen Hamilton	90,000
Bloggtoppen	90,000
Valve	700,000

“The passwords are stored encrypted, but with enough effort and depending on the quality of the password, they can be cracked. This, I'm afraid, is a serious threat; it means that anyone who uses the same email/password on other systems is now vulnerable to a malicious attacker using that information to access their account.”

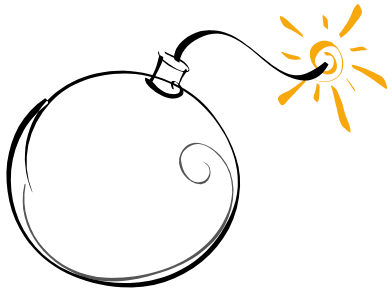
Jeremy White, CEO of Codeweavers
October 2011

Threat Model

Offline Attack

- Attacker has password file

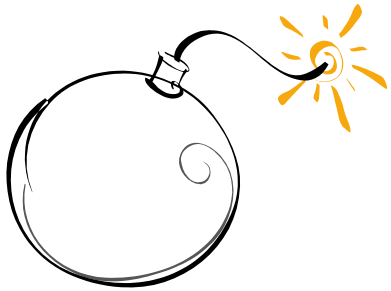
- Needs to guess passwords to crack them



Threat Model

Offline Attack

- Attacker has password file
- Needs to guess passwords to crack them
- Attacker can make many guesses
- Smart guessing strategy



Guessing Strategy

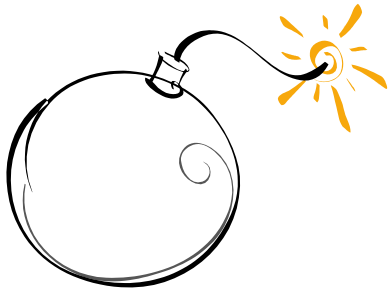
Dumb attacker	Smart attacker
aaaaaaaa	123456789
aaaaaaab	password
aaaaaaac	iloveyou
aaaaaaad	princess
aaaaaaae	12345678
…	…

Smart attacker uses data to crack passwords more quickly

Threat Model

Offline Attack

- Attacker has password file
- Needs to guess passwords to crack them
- Attacker can make many guesses
- Smart guessing strategy



Password-composition Policies

- Intended to make passwords harder to guess

Password-composition Policies

WIKIPEDIA

Log in / create account

From Wikipedia, the free encyclopedia

Login error
Passwords must be at least 1 character.

Password Requirements

Adhere to the following password requirements, when selecting your Andrew account pas

Must Contain
• At least 8-characters. • At least one uppercase alphabetic character (e.g., A-Z). • At least one lowercase alphabetic character (e.g., a-z). • At least one number (e.g., 0-9). • At least one special character (e.g., ~!@#$%^&*()_-+=).
Cannot Contain
• Known information (i.e., first name, last name, Andrew userID, date of birth, 9-digit Carnegie Mellon ID number, SSN, job title). • Four or more occurrences of the same character (e.g., aaaa, 2222, a123a345a678a).* • A word that is found in a standard dictionary.* **Note**: Verify that the letters within your password do not spell a word after you remove any non-alphabetical or special characters. The system checks all of the letters of the password together. Details... ***This requirement does not apply to Andrew account passwords that are more than 19 characters in length (e.g., passphrase).***
Additional Policies
• Last five passwords cannot be used. • Cannot be changed more than four times in a day.

Existing Guidance

NIST Special Publication 800-63
Version 1.0.2

National Institute of Standards and Technology
Technology Administration
U.S. Department of Commerce

Electronic Authentication Guideline

Recommendations of the National Institute of Standards and Technology

William E. Burr
Donna F. Dodson
W. Timothy Polk

INFORMATION SECURITY

Computer Security Division
Information Technology Laboratory
National Institute of Standards and Technology

Existing Guidance

National Institute of Standards and Technology
Technology Administration
U.S. Department of Commerce

- NIST guide not based on empirical evidence
- No empirical data on user behavior

Password-composition Policies

- Users can struggle to create and remember complex passwords [Zviran & Haga 1999, Procter et al. 2002, Yan et al. 2004, Vu et al. 2007, and many others…]

- Security can suffer if usability is poor [Sasse et al. 2001, and many others…]

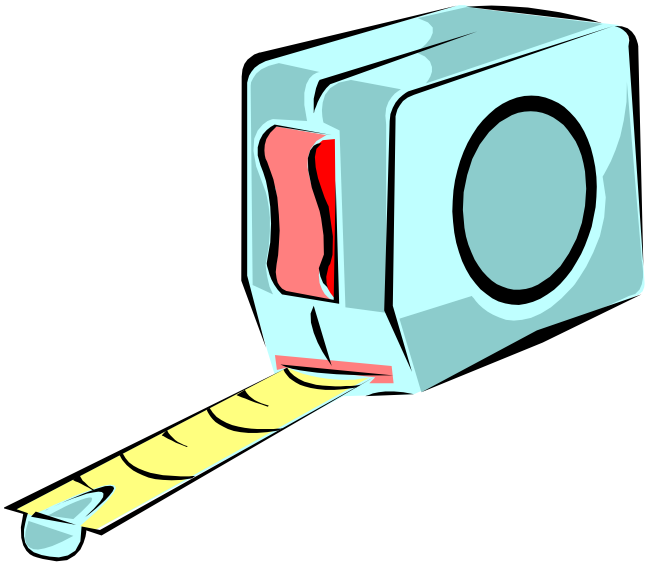
Contributions

- Measured guessability across seven password-composition policies
  - Threat model: offline attack
- Studied the impact of tuning and data selection on policy evaluation
- Compare security metrics across policies
  - Correlate security with usability

Policy Metrics

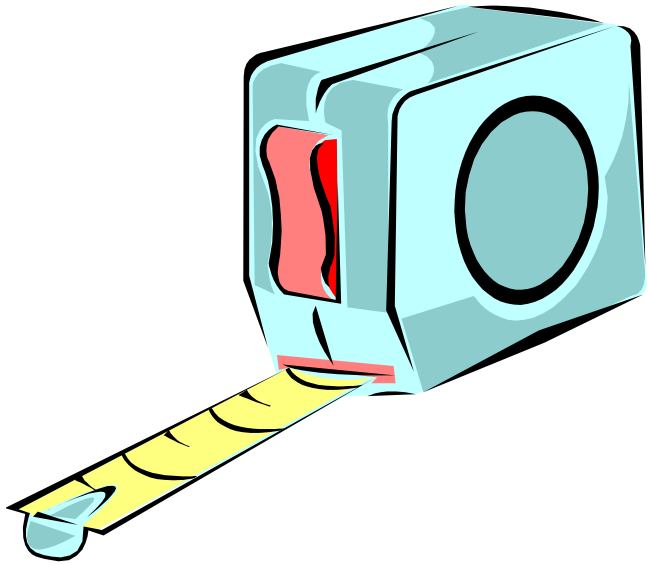
- Guessability
  - Measure of how easy it is to guess passwords

- Estimated entropy [Our previous work 2010]



Policy Metrics

- Guessability
  - Measure of how easy it is to guess passwords
- Estimated entropy [Our previous work 2010]
- NIST entropy [NIST SP 800-63]
- Usability [Our previous work 2011]
  - Login failures
  - Reported sentiment
  - Writing down



Guessability

- Measure of password strength

  Stronger = less guessable

- Guess number: The number of attempts needed to guess a password

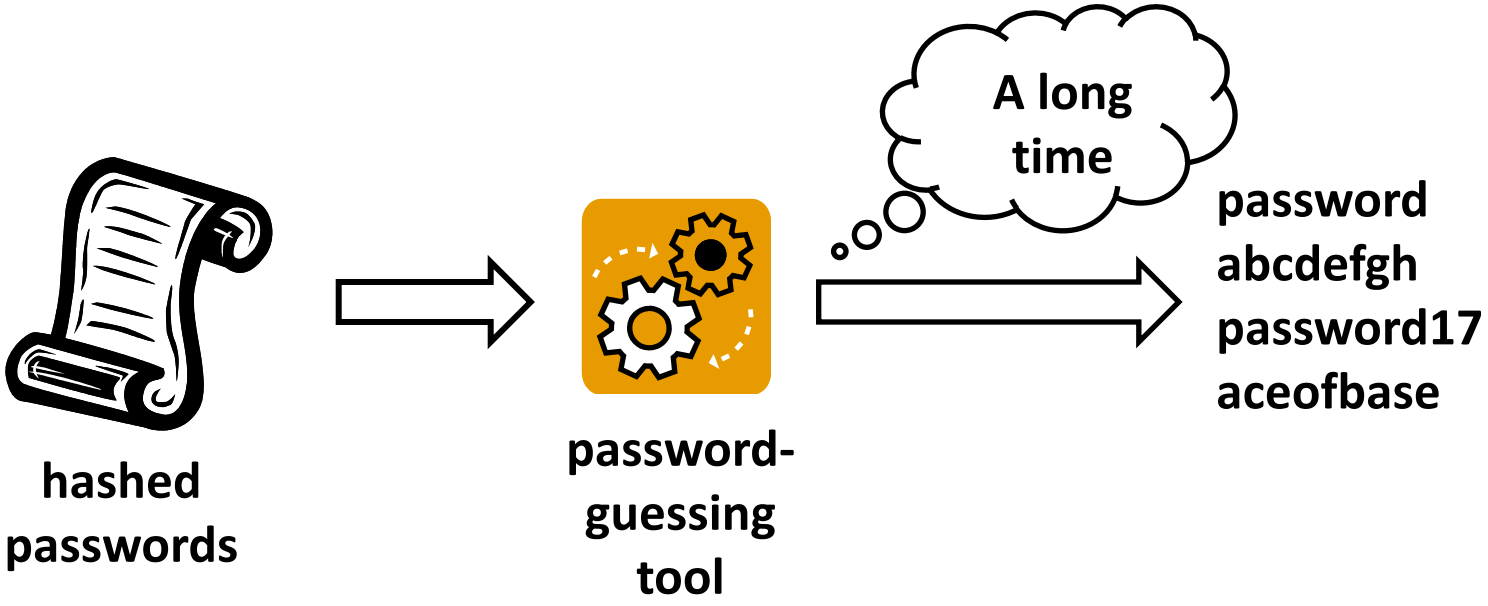
Guessability

Bob's password	Attacker's guesses
`iloveyou`	1 `123456789`
	2 `password`
	3 `iloveyou`
	4 `princess`
	…

Guessability

Bob's password	Attacker's guesses	
`iloveyou`	1 123456789	
	2 password	Guess number
	3 iloveyou	3
	4 princess	
	…	

Measuring Guessability



Traditional approach: Run cracking tool

Offline Attack Speed

Single-core CPU

1,500	guesses/s	sha512
130,000,000	guesses/day	sha512
2,200,000,000	guesses/day	md5

Mid-level GPU

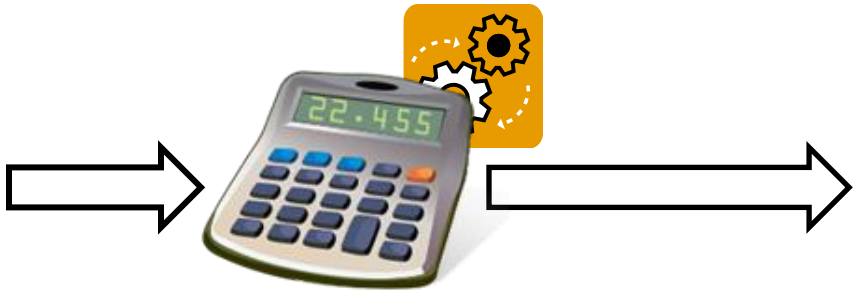
34,000,000,000	guesses/day	md5

Source: John the Ripper Test Mode and Wiki (openwall.info)

Measuring Guessability

password
abcdefgh
password17
aceofbase
jnfksl834df

plaintext passwords



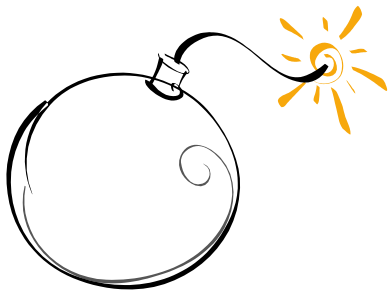
password-guessing *calculator*

password: 2
abcdefgh: 19546
password17: 1.4×10^6
aceofbase: 3×10^4
jnfksl834df: never

Our approach: Calculate guess numbers directly

Threat Model

- Offline attacker that can make a huge number of guesses
  - This paper: 50 trillion (5×10^{13}) guesses on each password
    - 25,000 CPU days with MD5 hashes



Selecting an Attacker

- John the Ripper
- Markov model [Narayanan and Shmatikov 2005]
- Weir's probabilistic context-free grammar [Weir et al. 2009]

Selecting an Attacker

- John the Ripper
- Markov model [Narayanan and Shmatikov 2005]
- Weir's probabilistic context-free grammar
  - Performed best
  - Previous work found similar result [Weir et al. 2010, Zhang et al. 2010]

Training Data

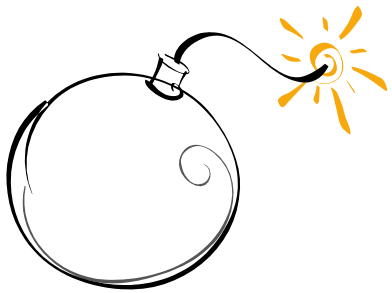
- Leaked datasets
  - RockYou (32M passwords)
  - MySpace (47K passwords)

Training Data

- Leaked datasets
  - RockYou (32M passwords)
  - MySpace (47K passwords)
- Dictionaries
  - Openwall (40M passwords)
  - Unix dictionary (235K words)
  - Inflection list (162K words)
- Collected passwords (12K total passwords)

Threat Model

- Offline attacker that can make up to 50 trillion guesses
- Order of guesses based on Weir's algorithm
  - Attacker learns from training data
    - Leaked data plus collected passwords
    - Attacker has limited knowledge of the target policy



Data Collection

- Mechanical Turk used for anonymous recruitment and payment
  - Enabled study of many participants
    - 1,000+ per condition
  - Well-designed studies can produce high-quality data [Burhmester et al. 2011]
  - Workers prevented from participating multiple times
  - Payment: 55¢ + 70¢

Study Design

- Hypothetical email scenario for password creation

Steps:

1. Create a password under a randomly assigned condition
2. Take a survey
3. Recall password
4. Return in two days

Condition: Basic8

password

NIST estimate: 18 bits

Condition: Dictionary8

sapsword

NIST estimate: 24 bits

Condition: Comprehensive8

Sapsword1!

NIST estimate: 30 bits

Condition: Basic16

passwordpassword

NIST estimate: 30 bits

Condition: Blacklist x 3

- Blacklists:
  - Easy: 235K Unix dictionary
  - Medium: 40M entry cracking wordlist
  - Hard: 5B guesses from Weir

- Only requirement is that candidate password is not on a blacklist

NIST estimate: 24 bits

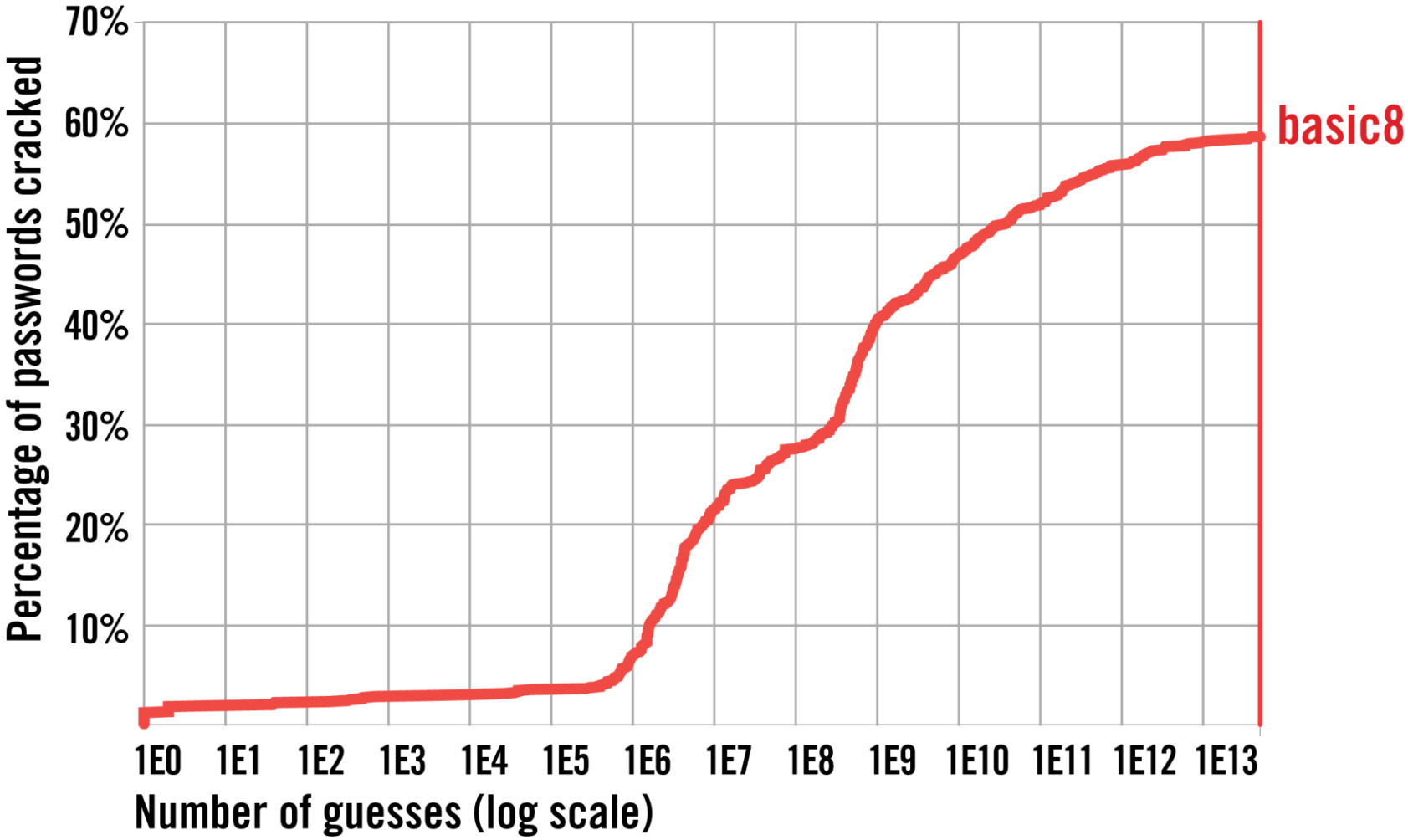
Contributions

- Measured guessability across seven password-composition policies
  - Threat model: offline attack

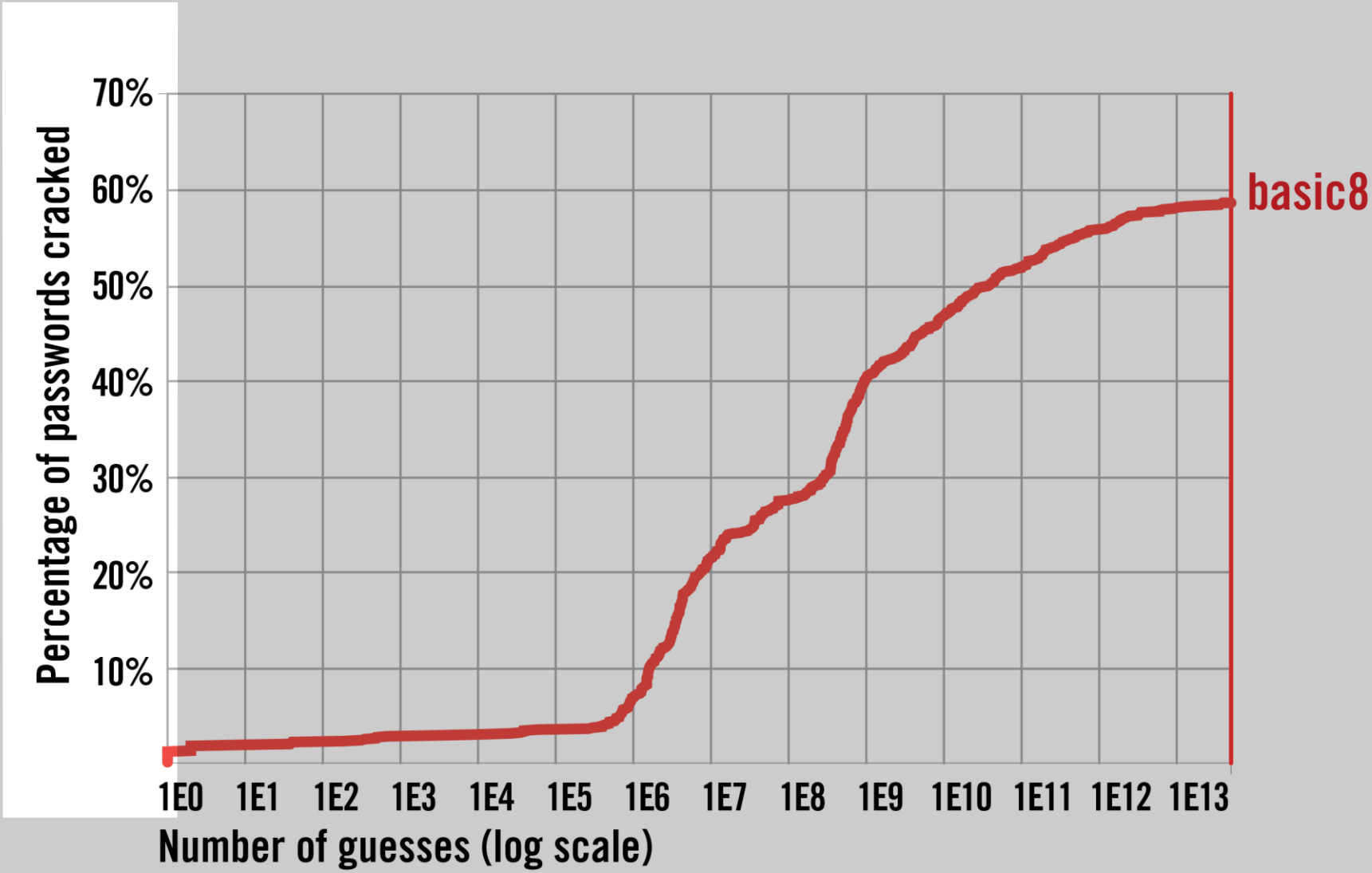
- Studied the impact of tuning and test-set selection on policy evaluation

- Compare security metrics across policies
  - Correlate security with usability

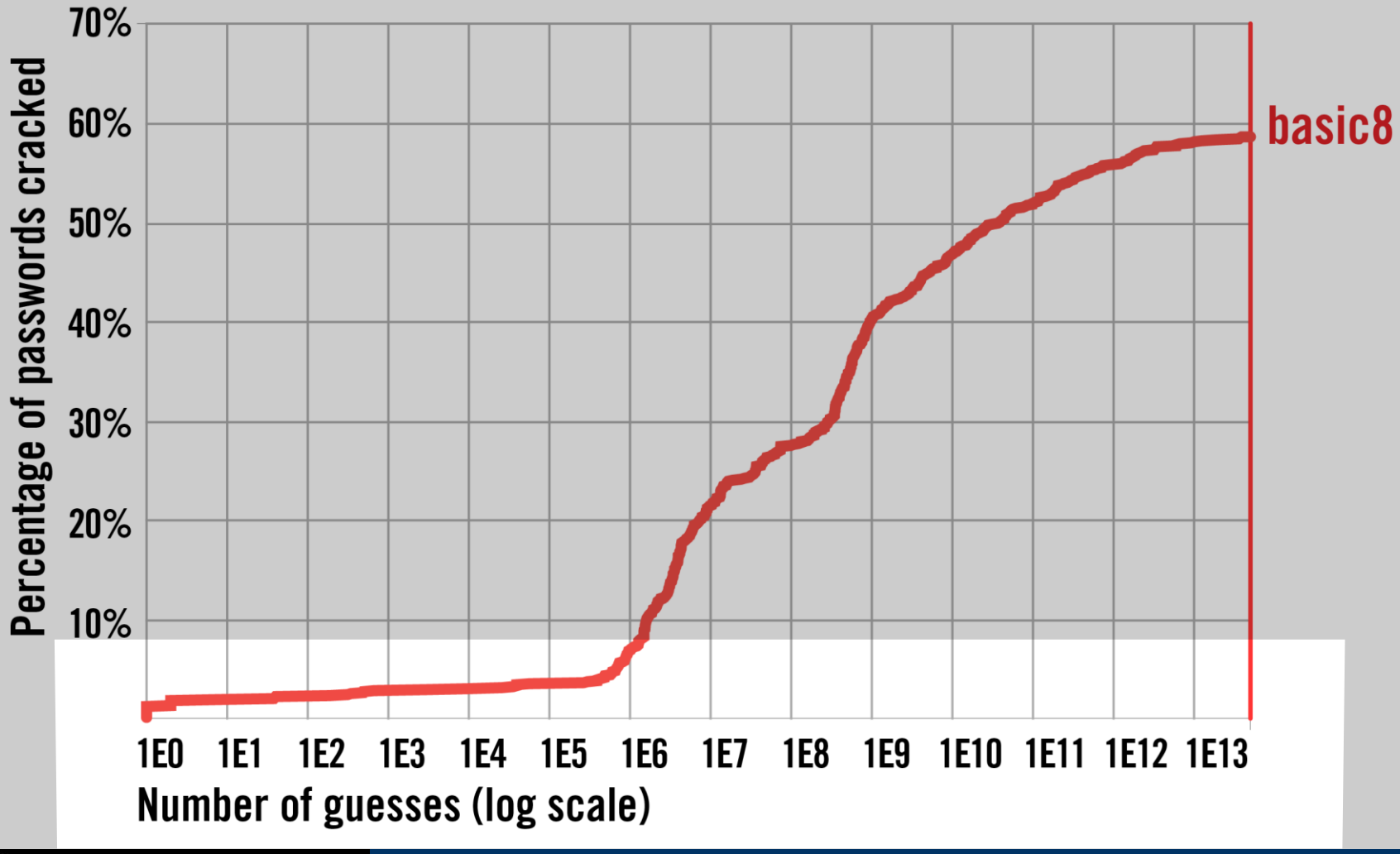
Guessability Results – Basic8



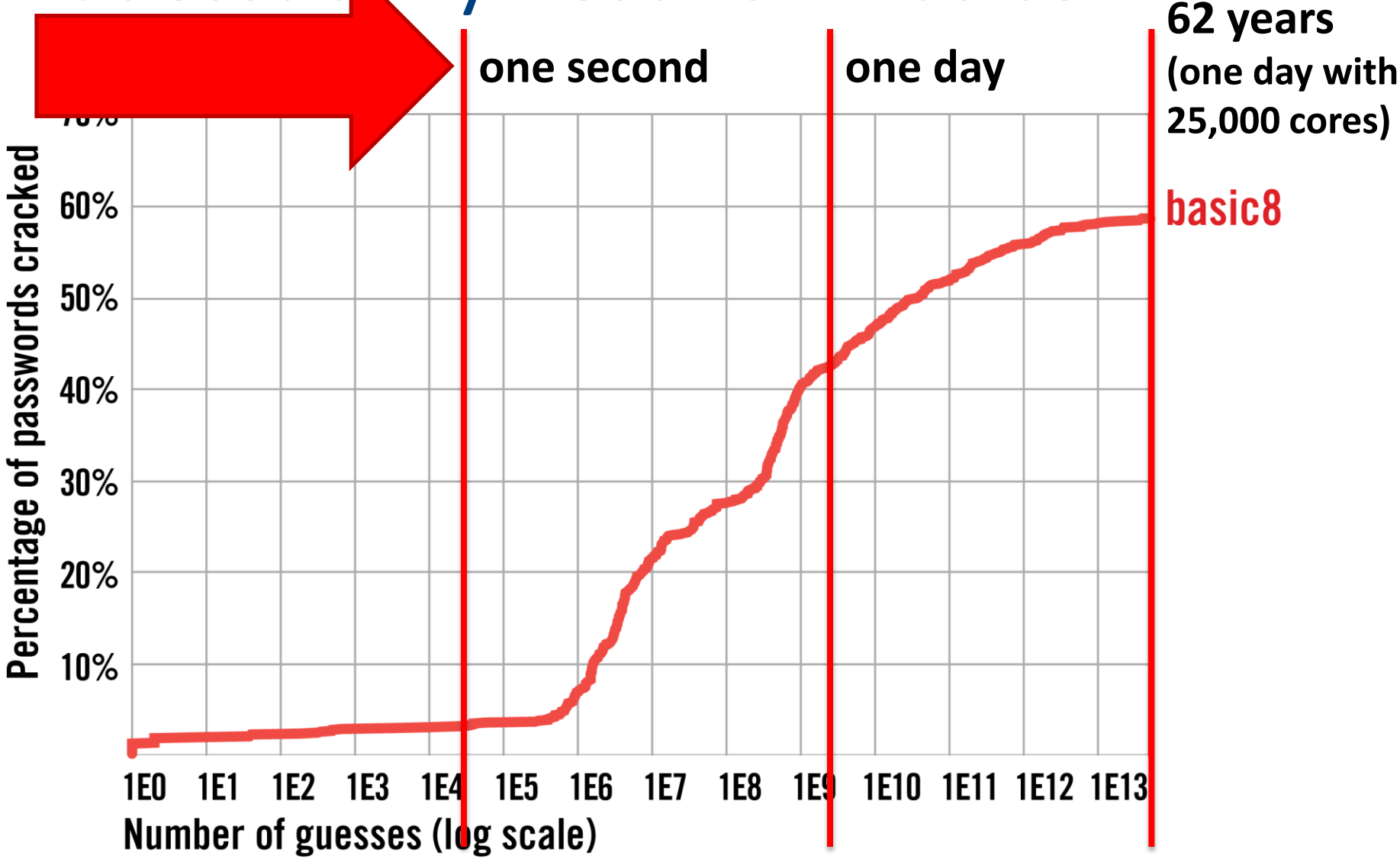
Guessability Results – Basic8



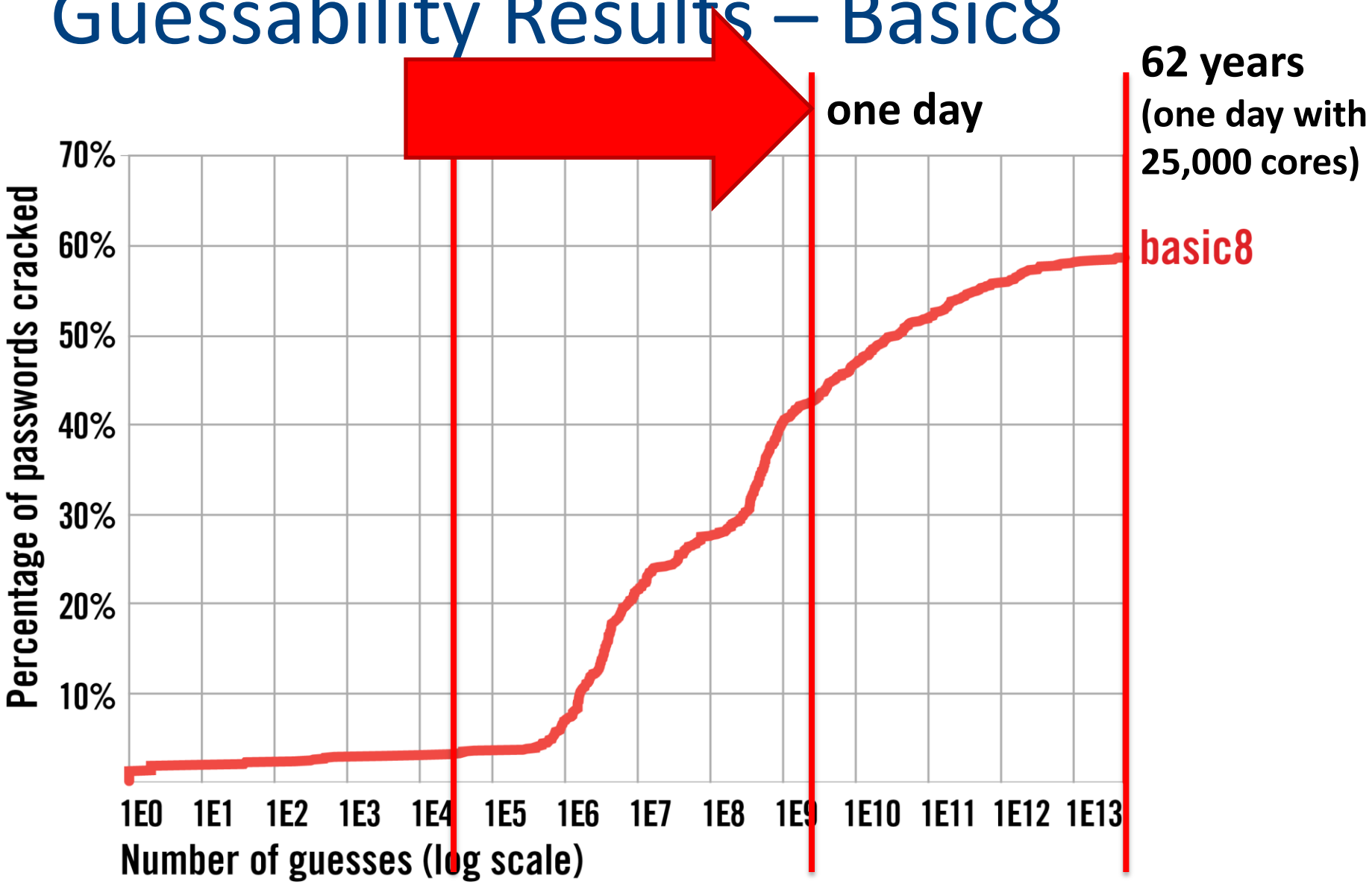
Guessability Results – Basic8



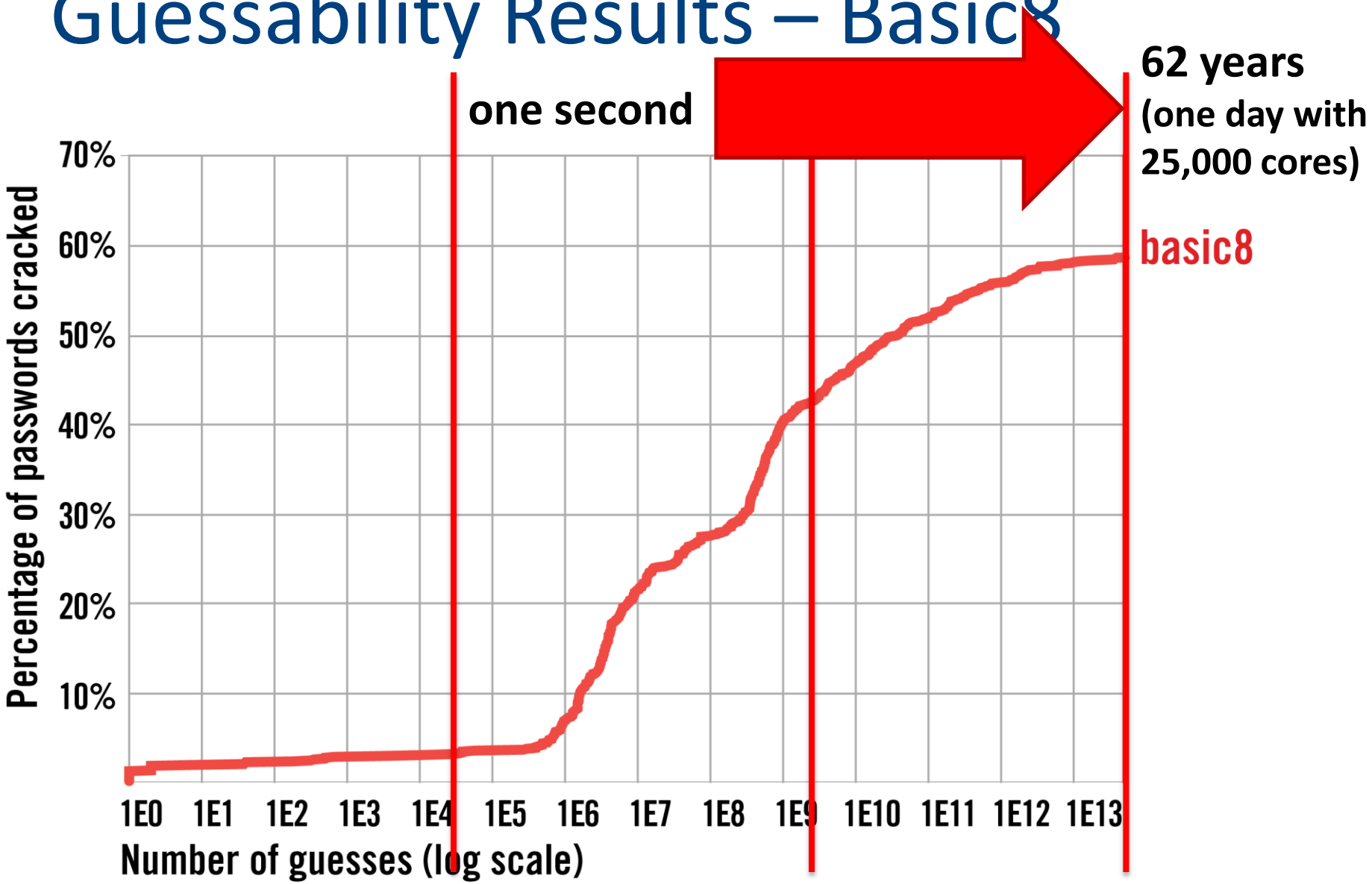
Guessability Results – Basic8



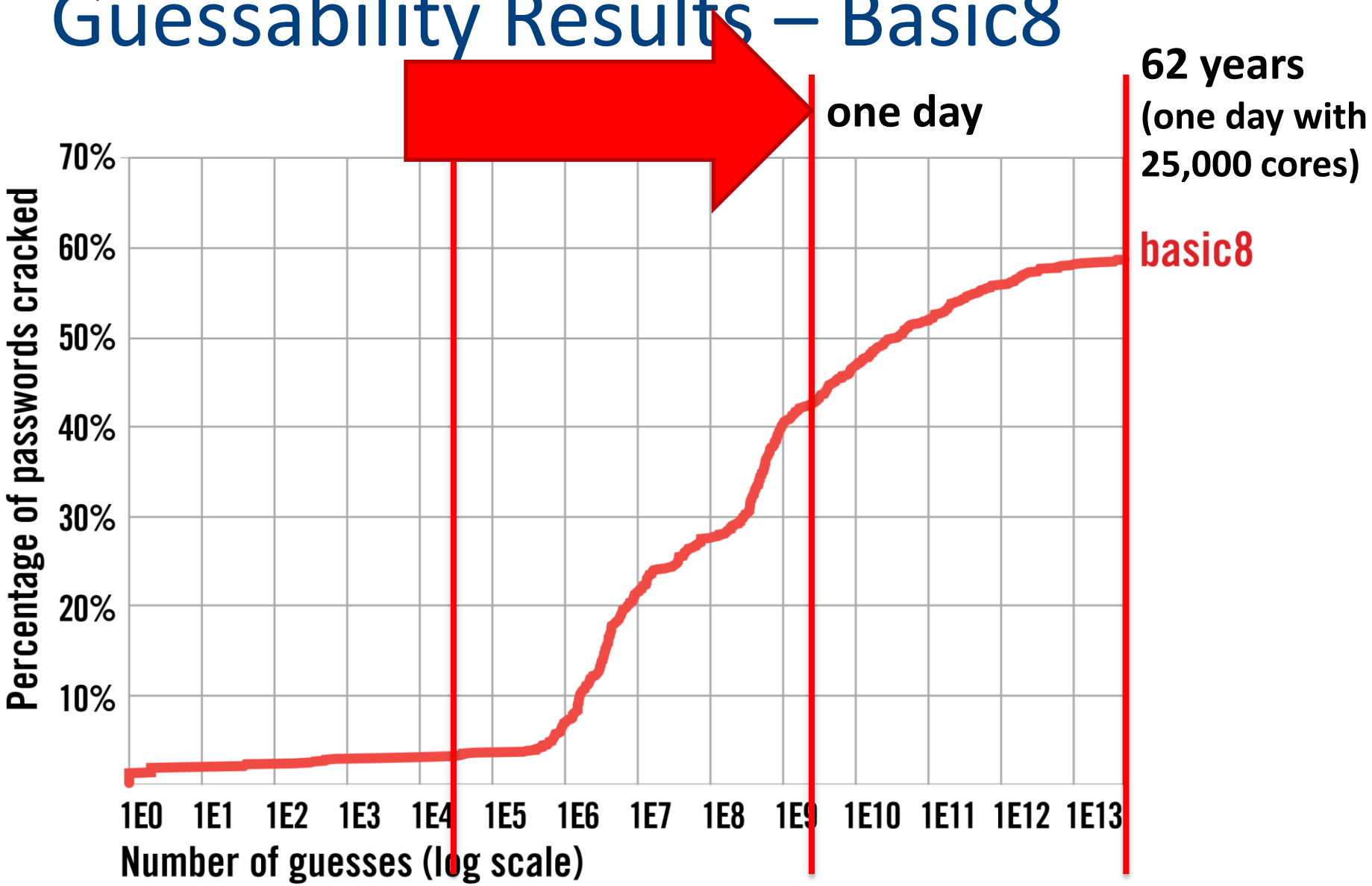
Guessability Results – Basic8



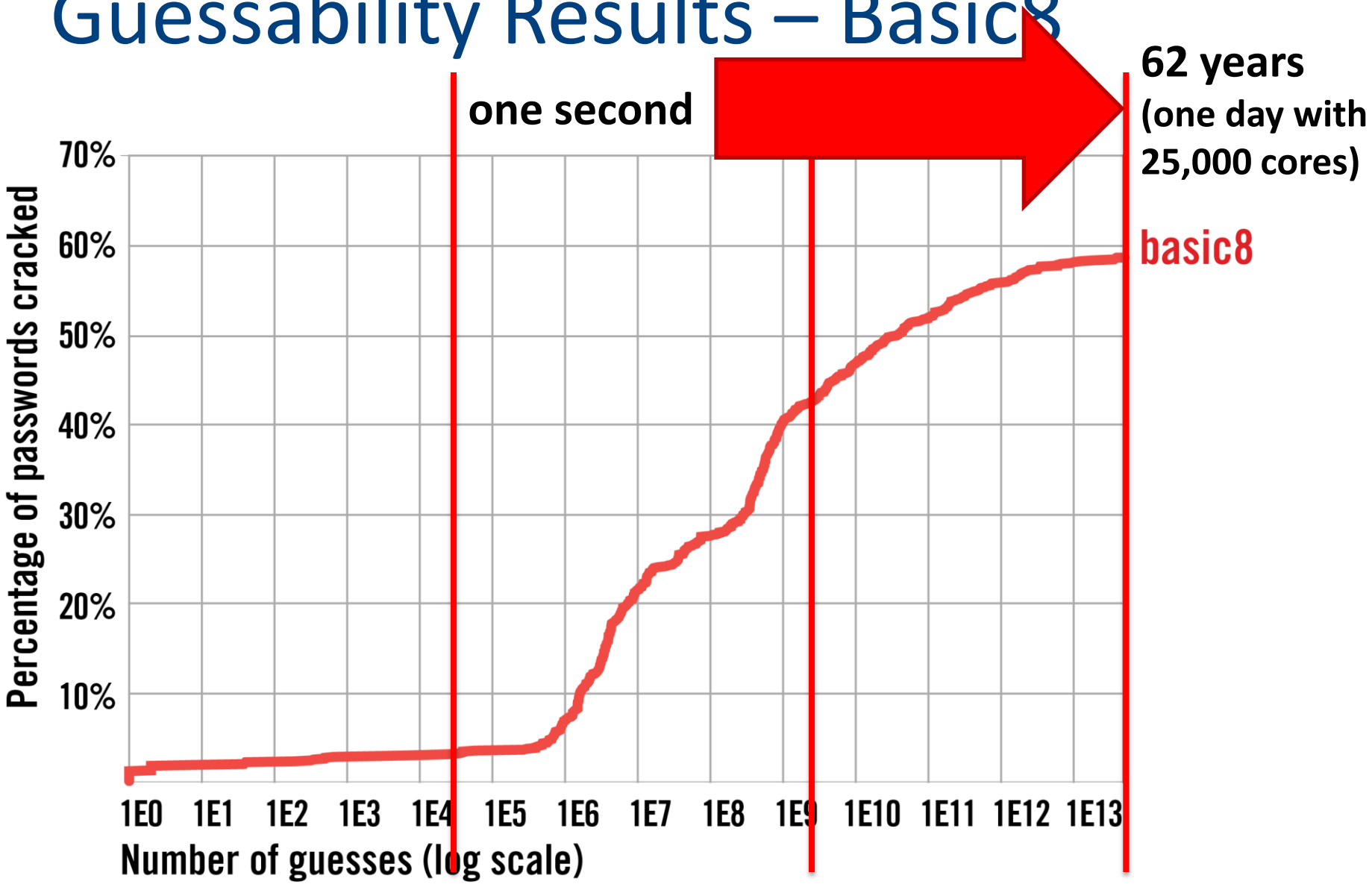
Guessability Results – Basic8



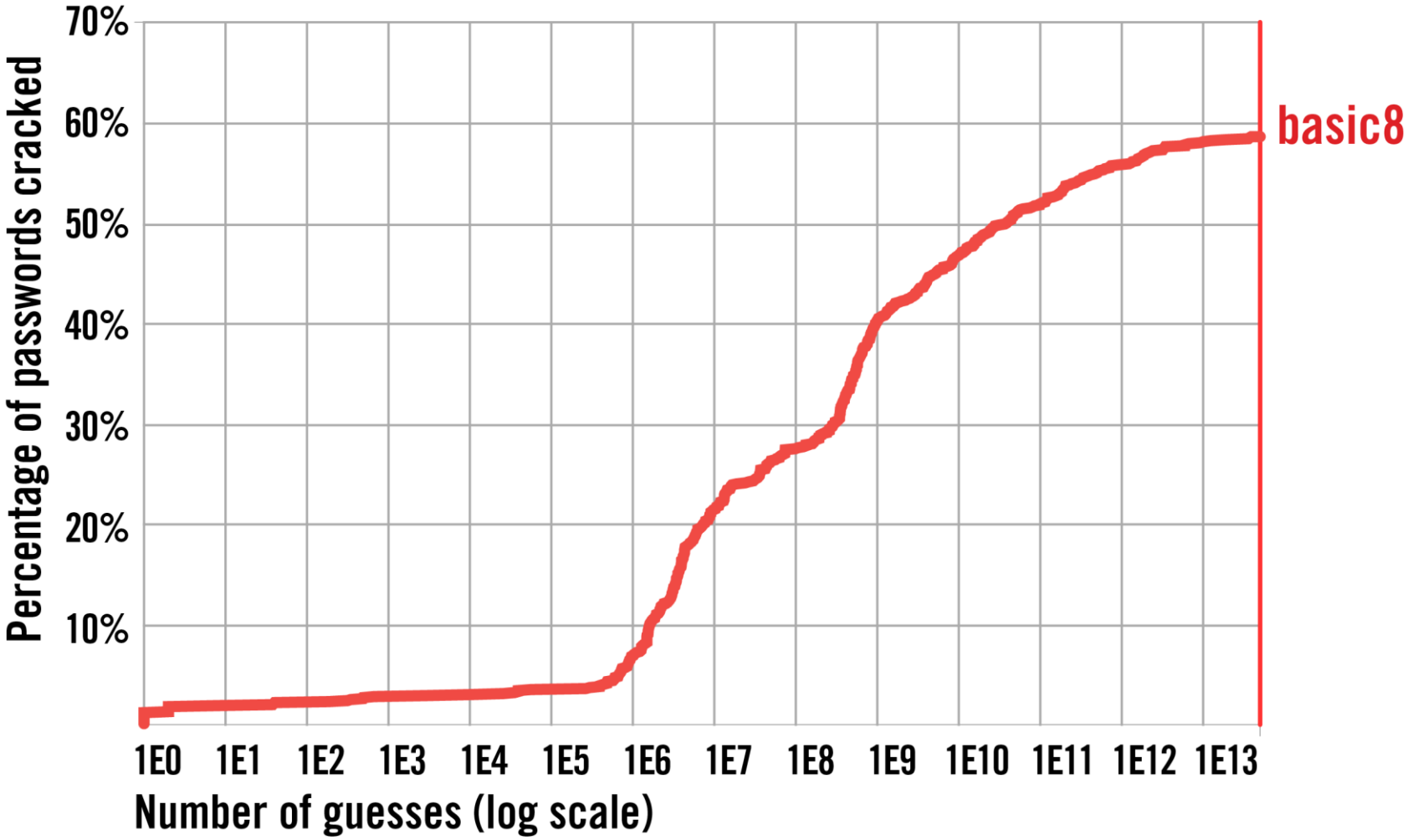
Guessability Results – Basic8



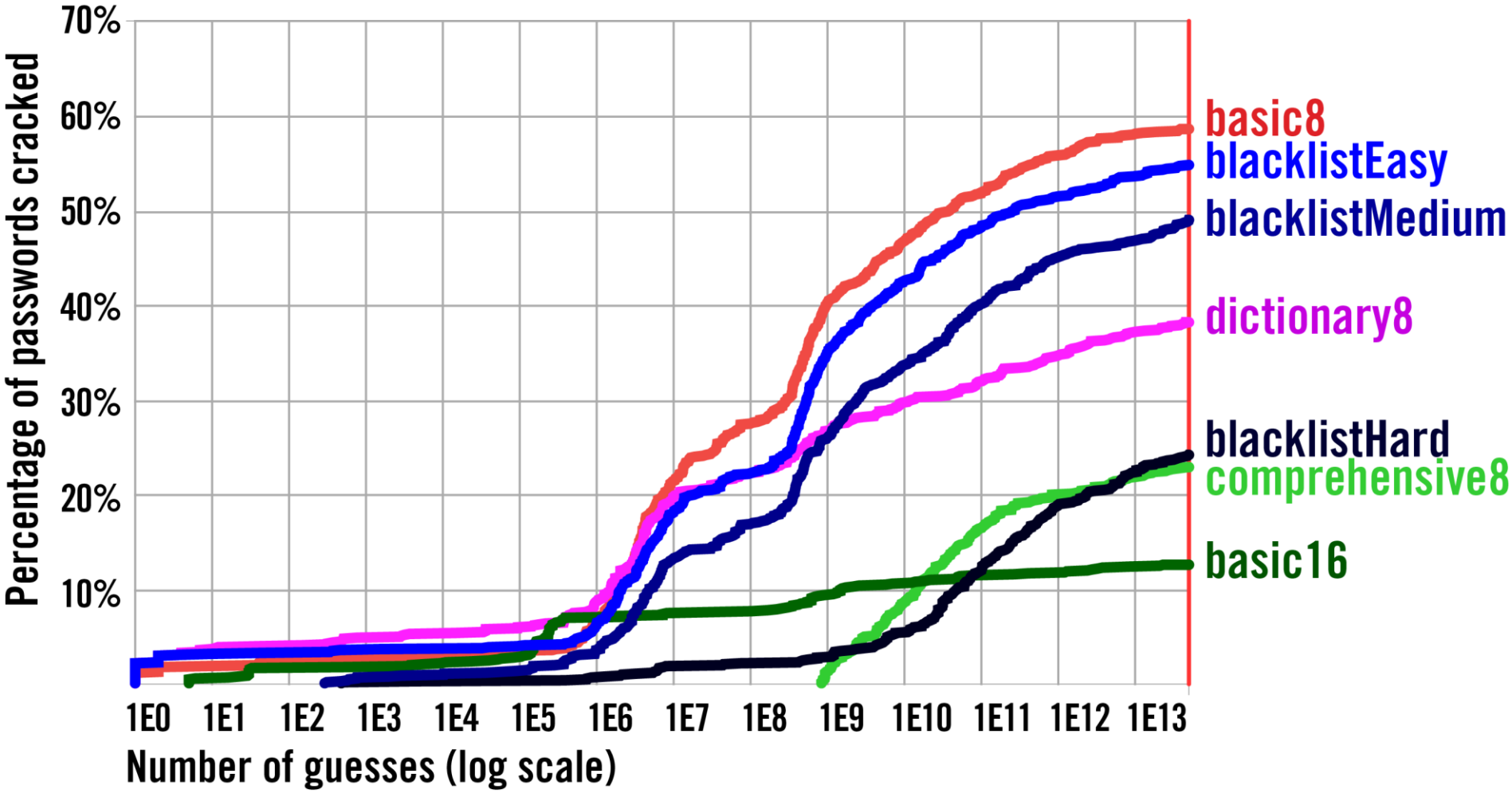
Guessability Results – Basic8



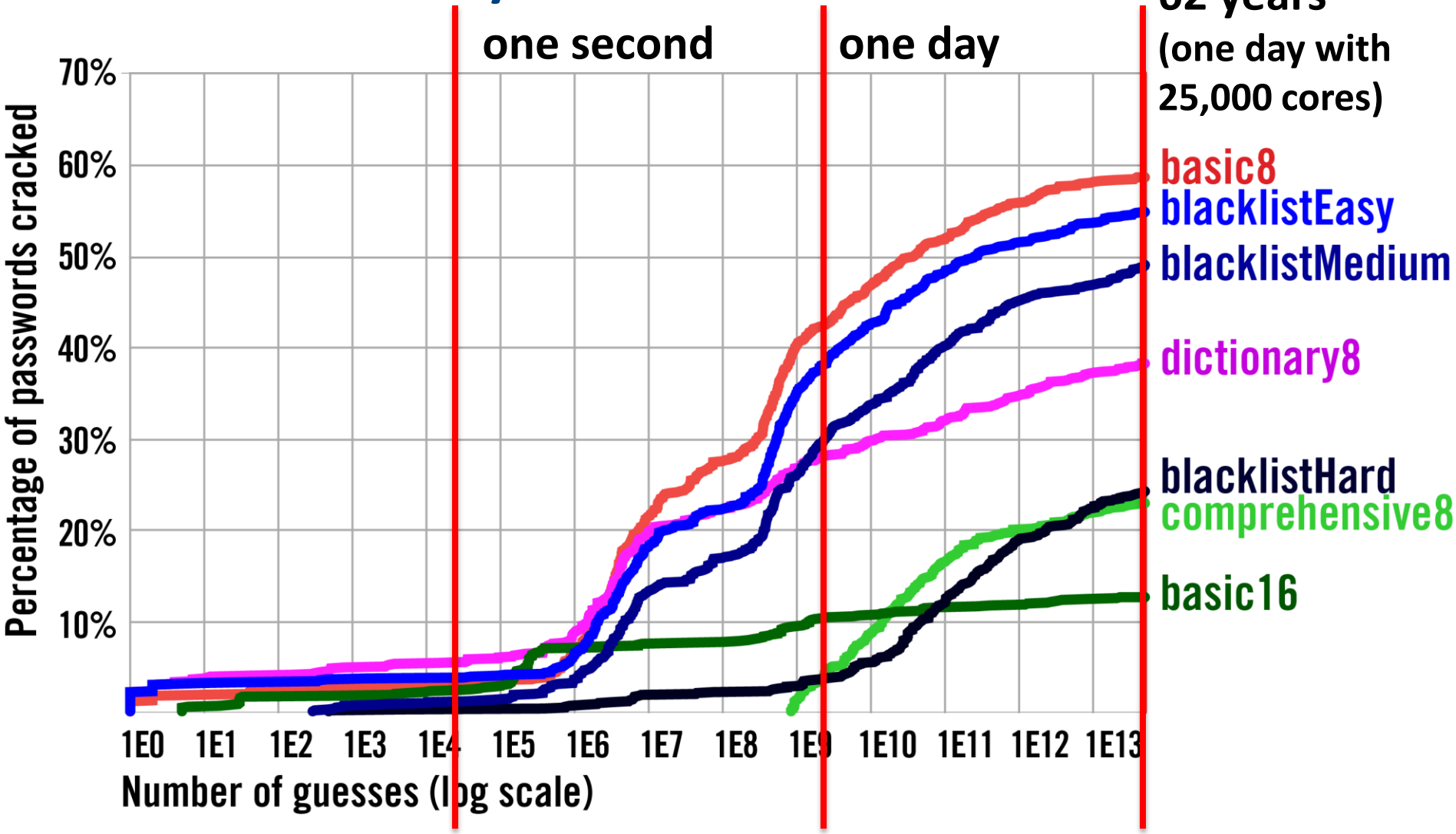
Guessability Results – Basic8



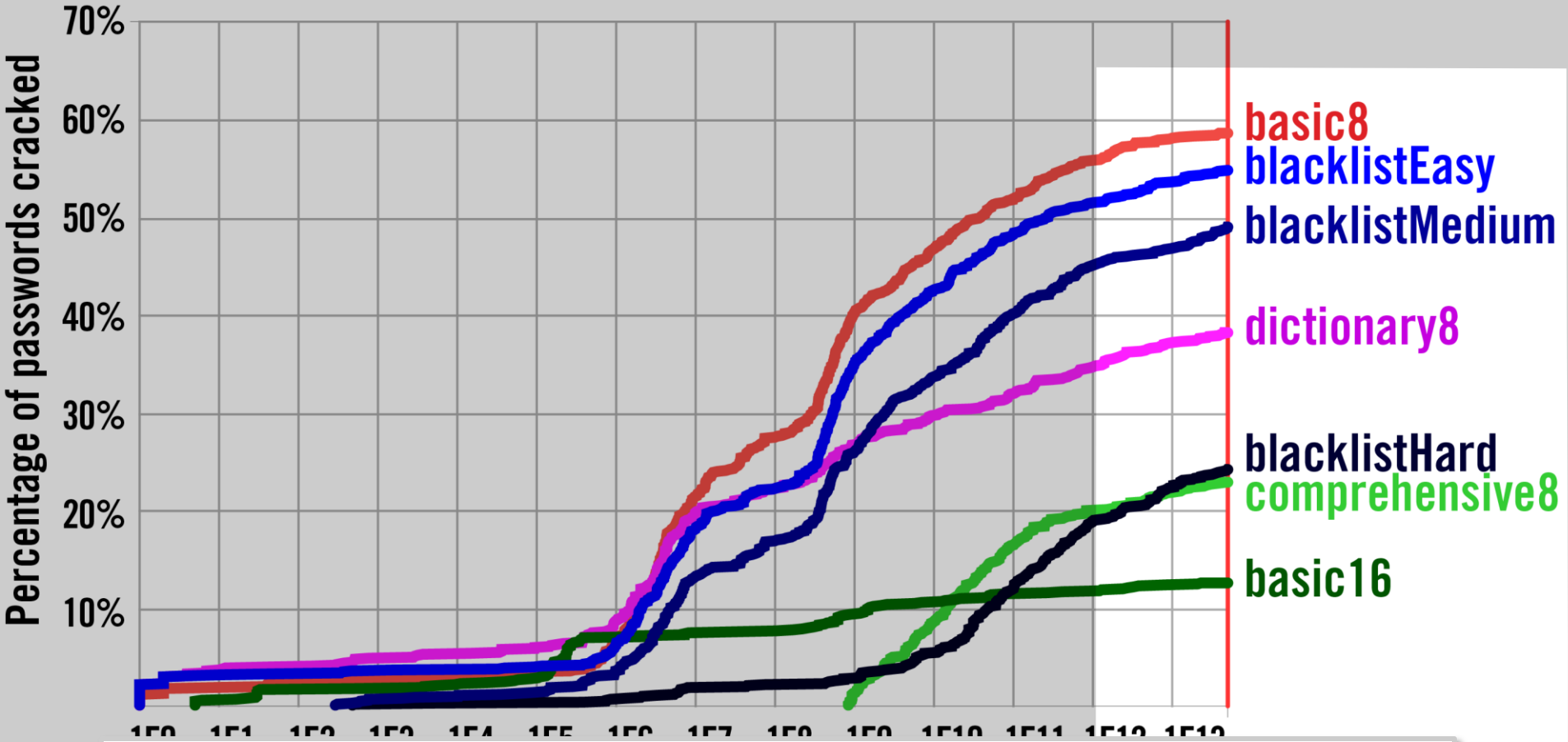
Guessability Results



Guessability Results

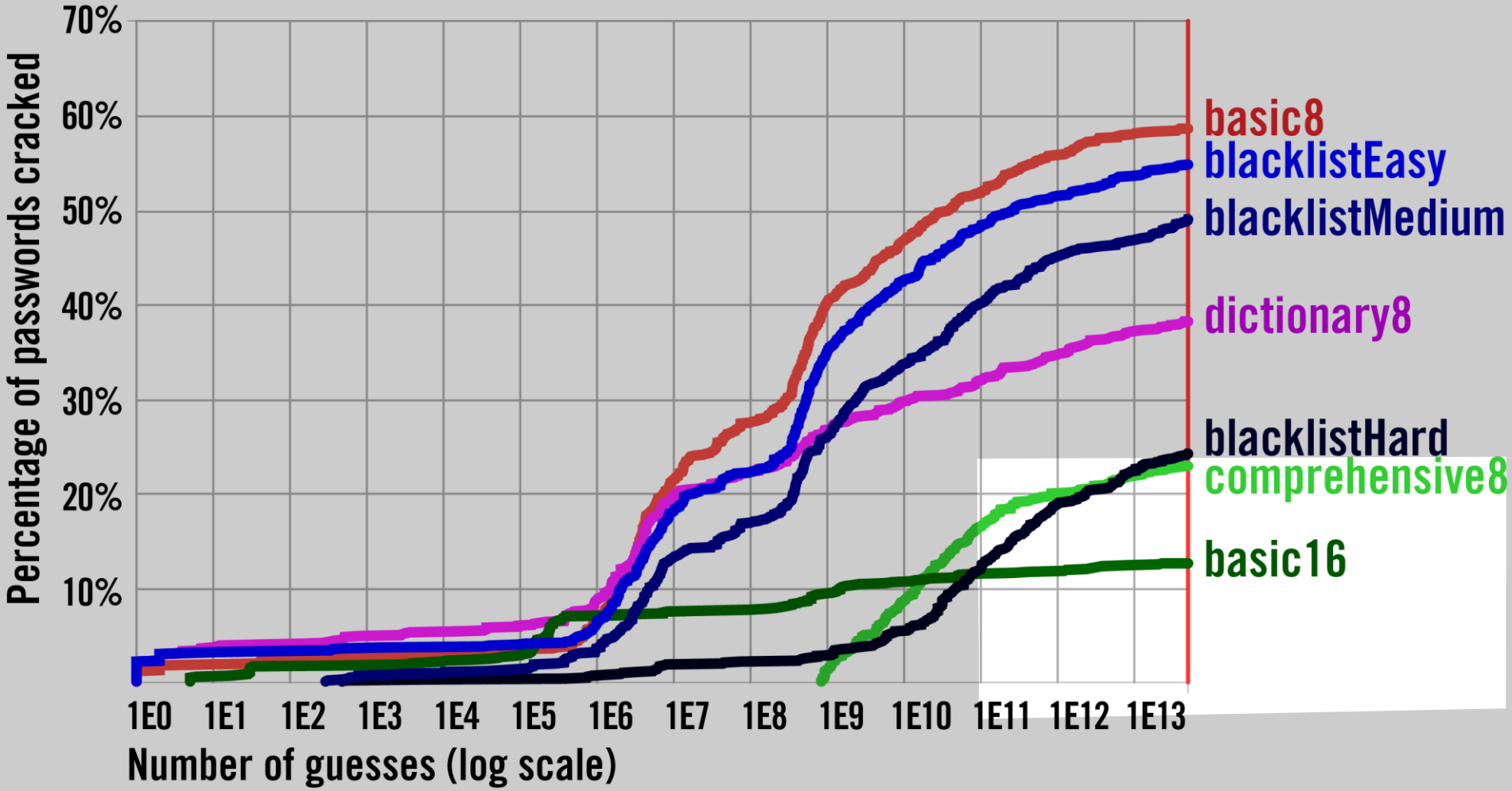


Guessability Results

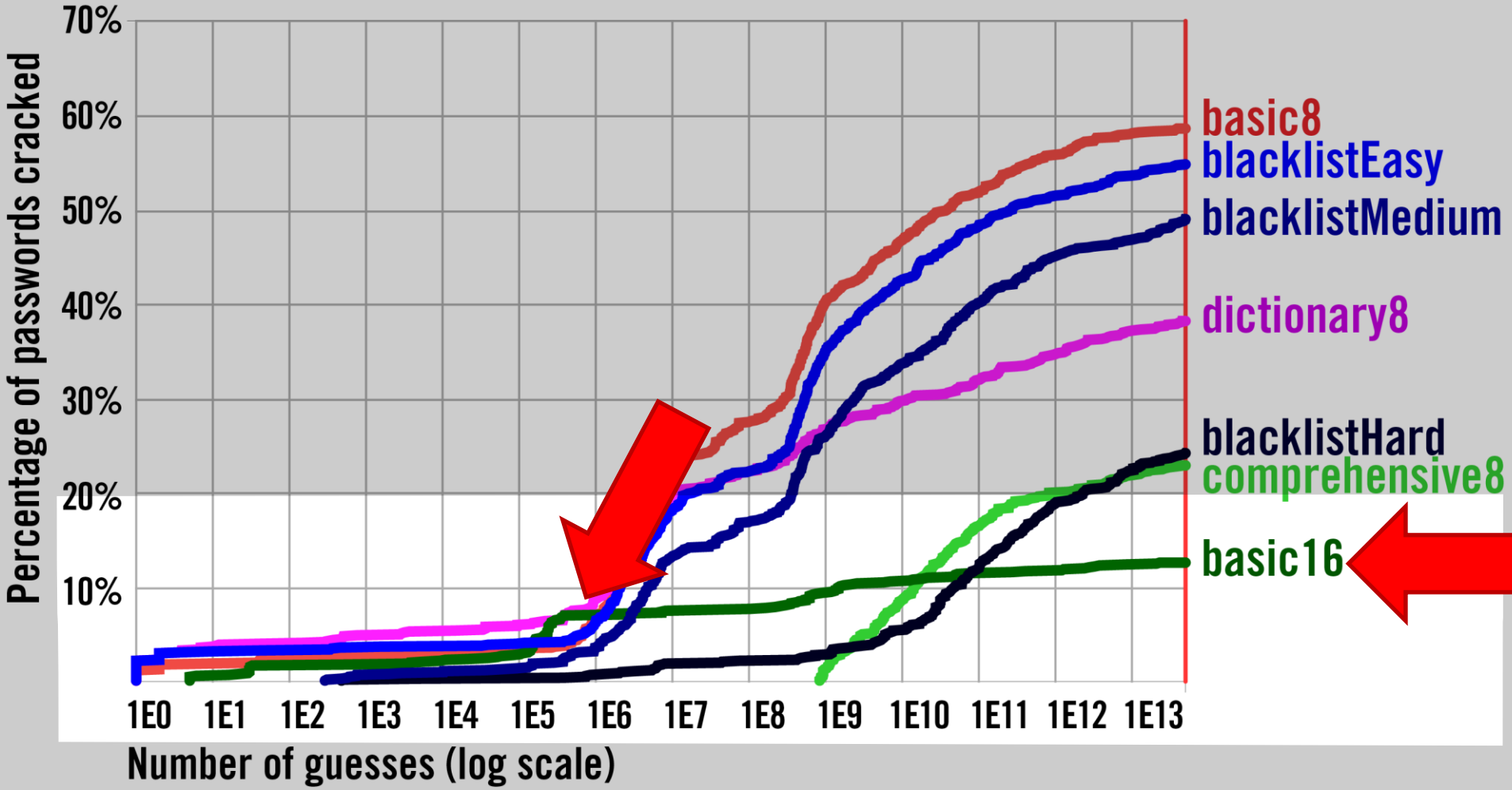


Basic16 performs best (13%), basic8 is worst (60%)

Guessability Results



Guessability Results



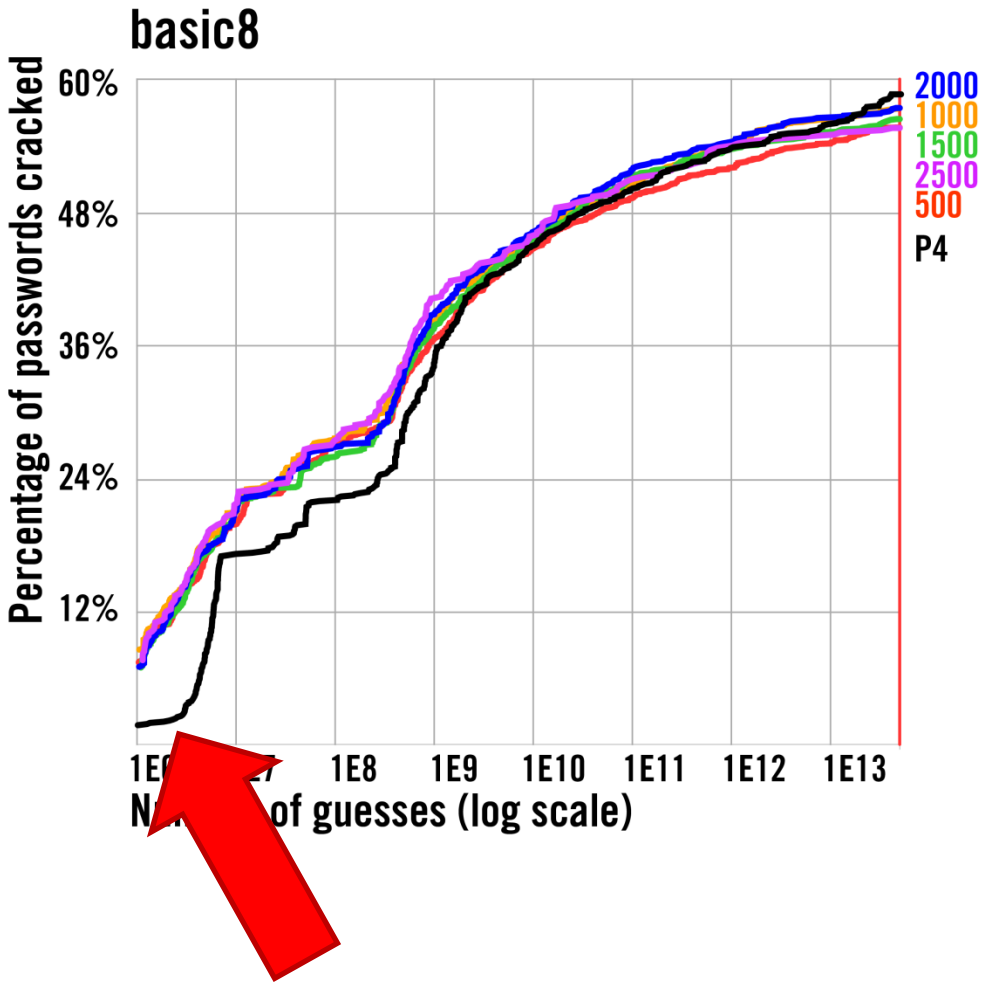
Contributions

- Measured guessability across seven password-composition policies
  - Threat model: offline attack

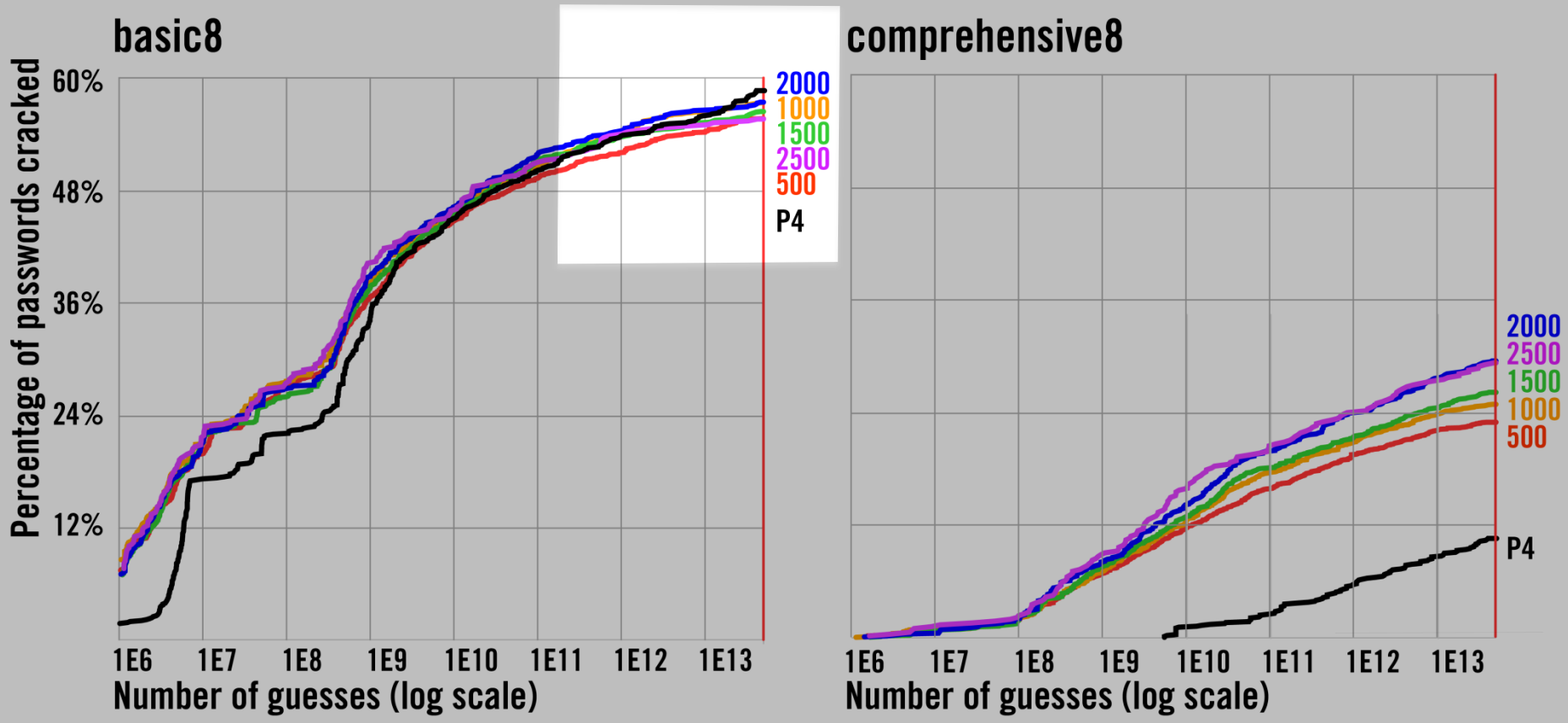
- Studied the impact of tuning and test-set selection on policy evaluation

- Compare security metrics across policies
  - Correlate security with usability

Increasing Training Data

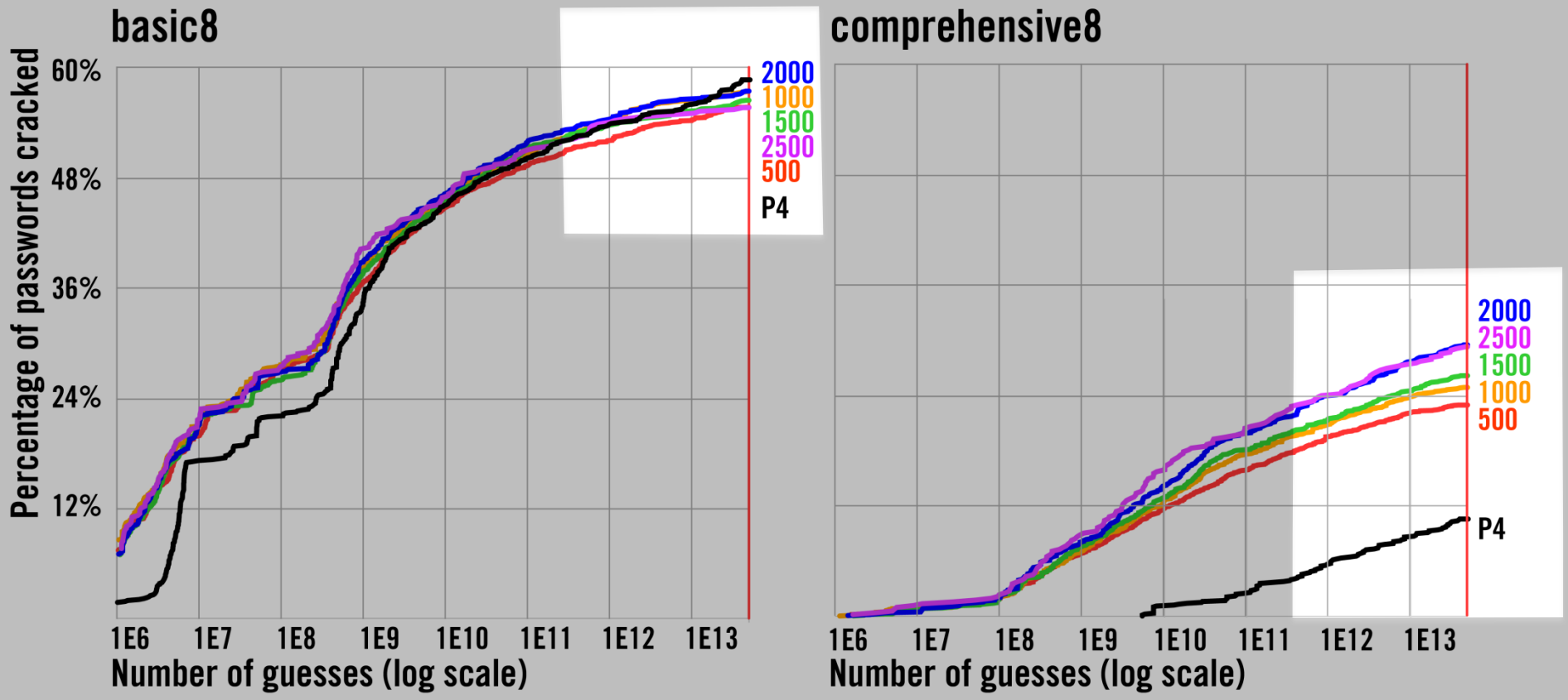


Increasing Training Data



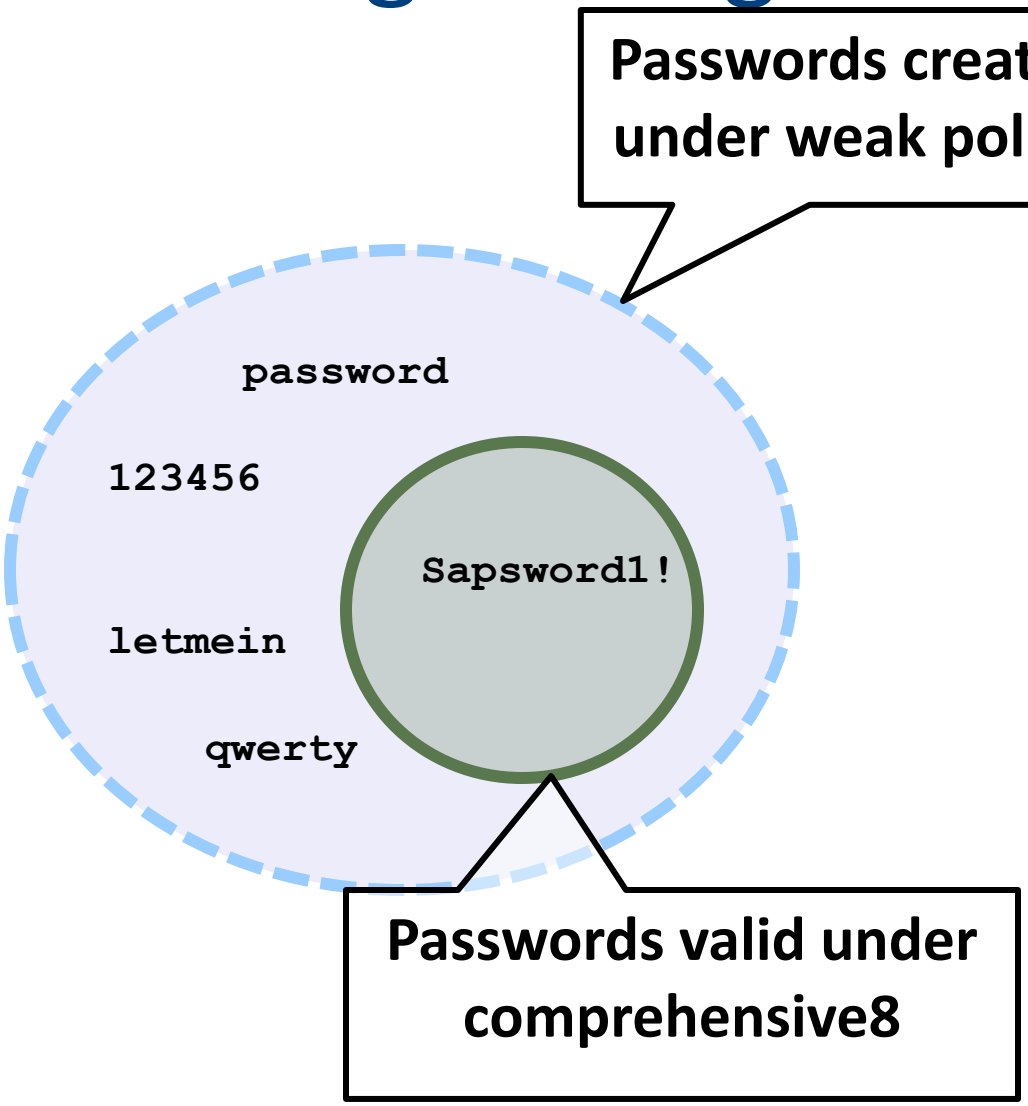
Basic8 does not benefit from additional data

Increasing Training Data

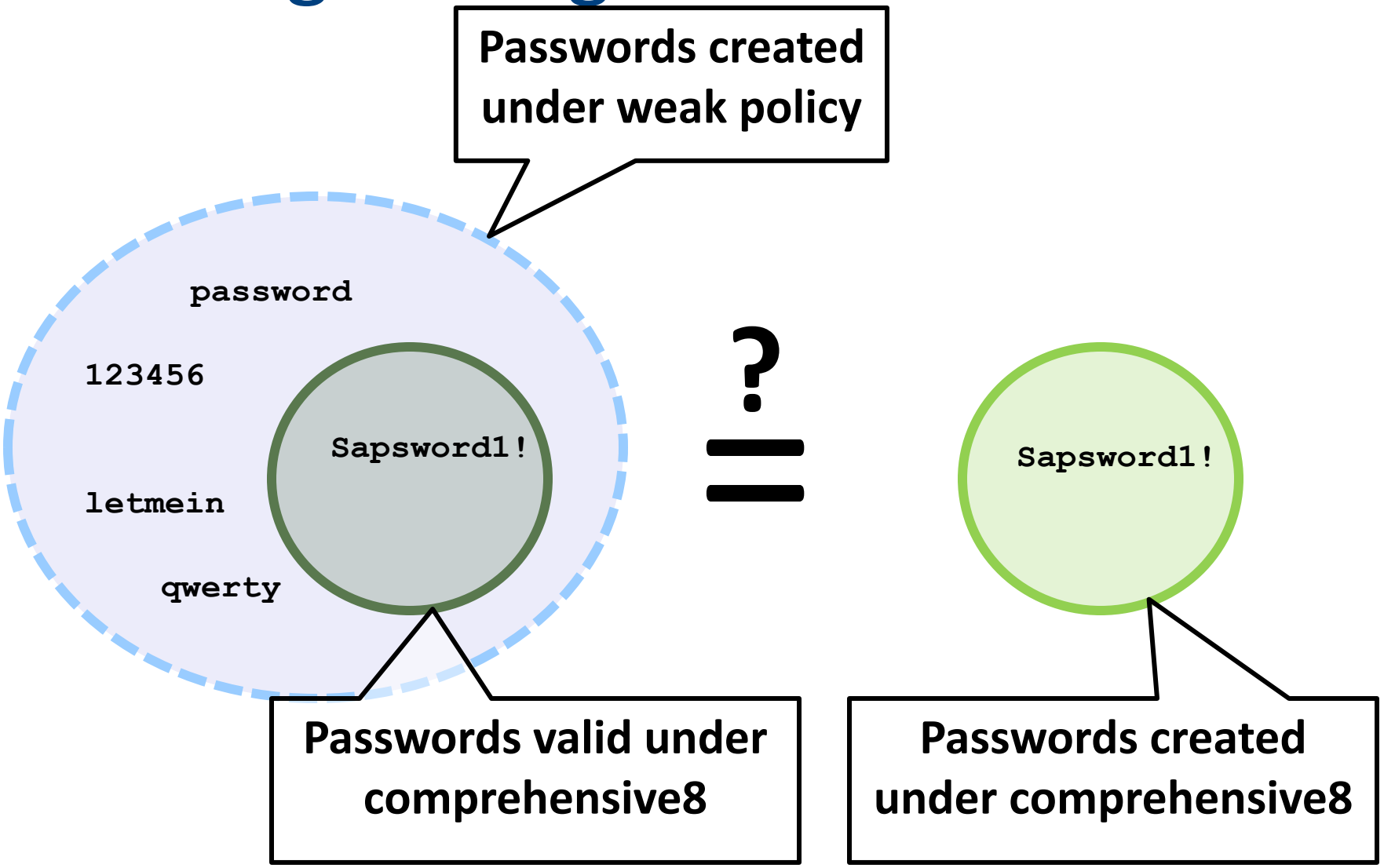


Target-policy passwords needed for complex policies

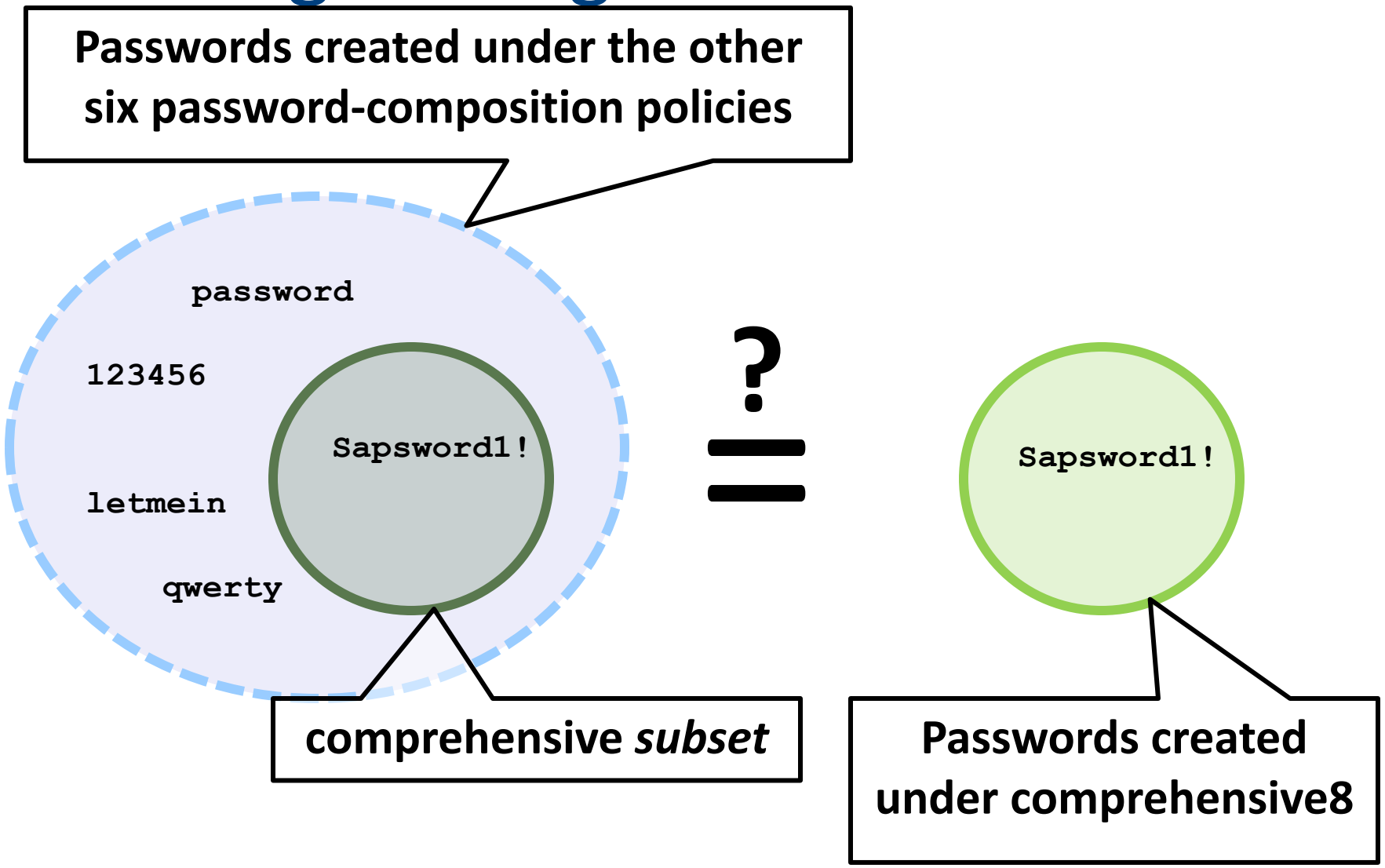
Choosing the Right Test Data



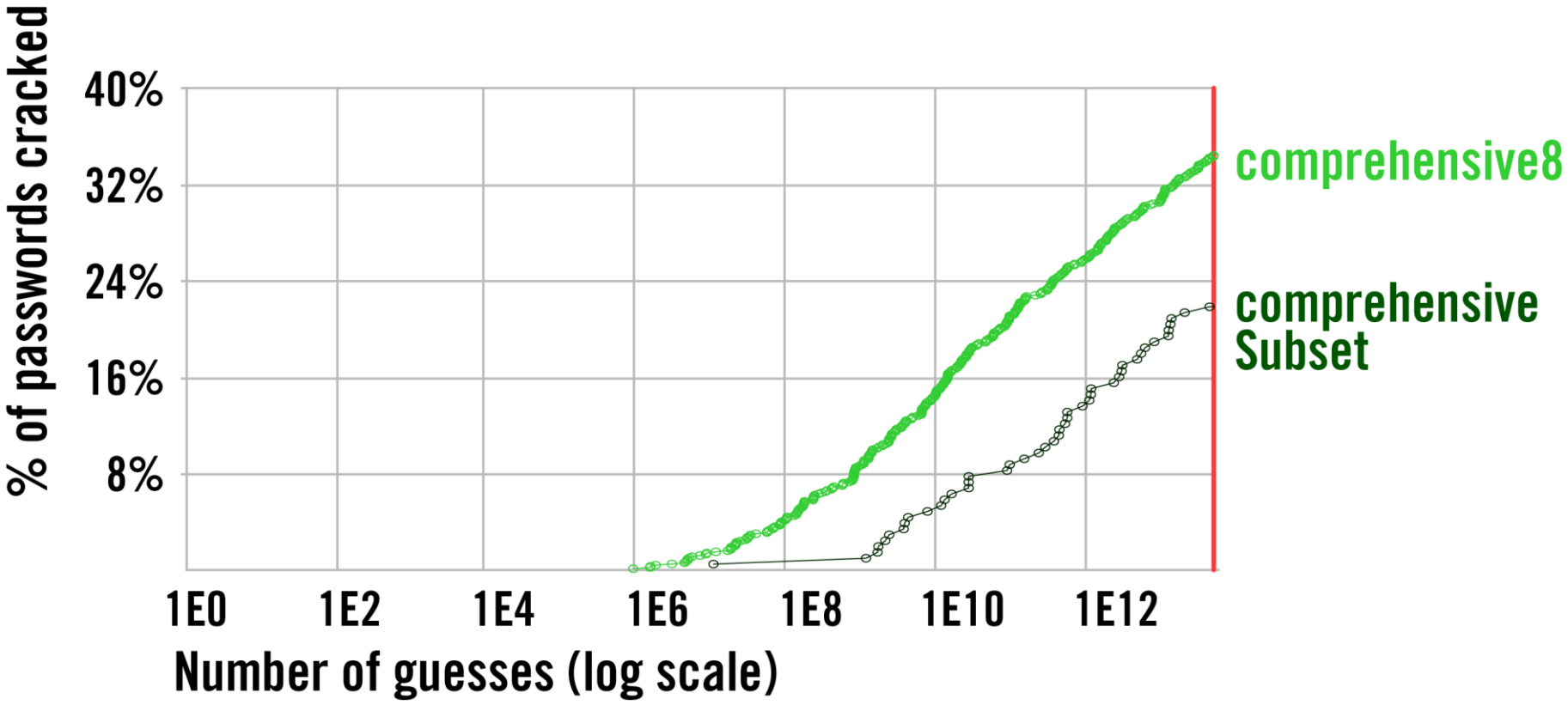
Choosing the Right Test Data



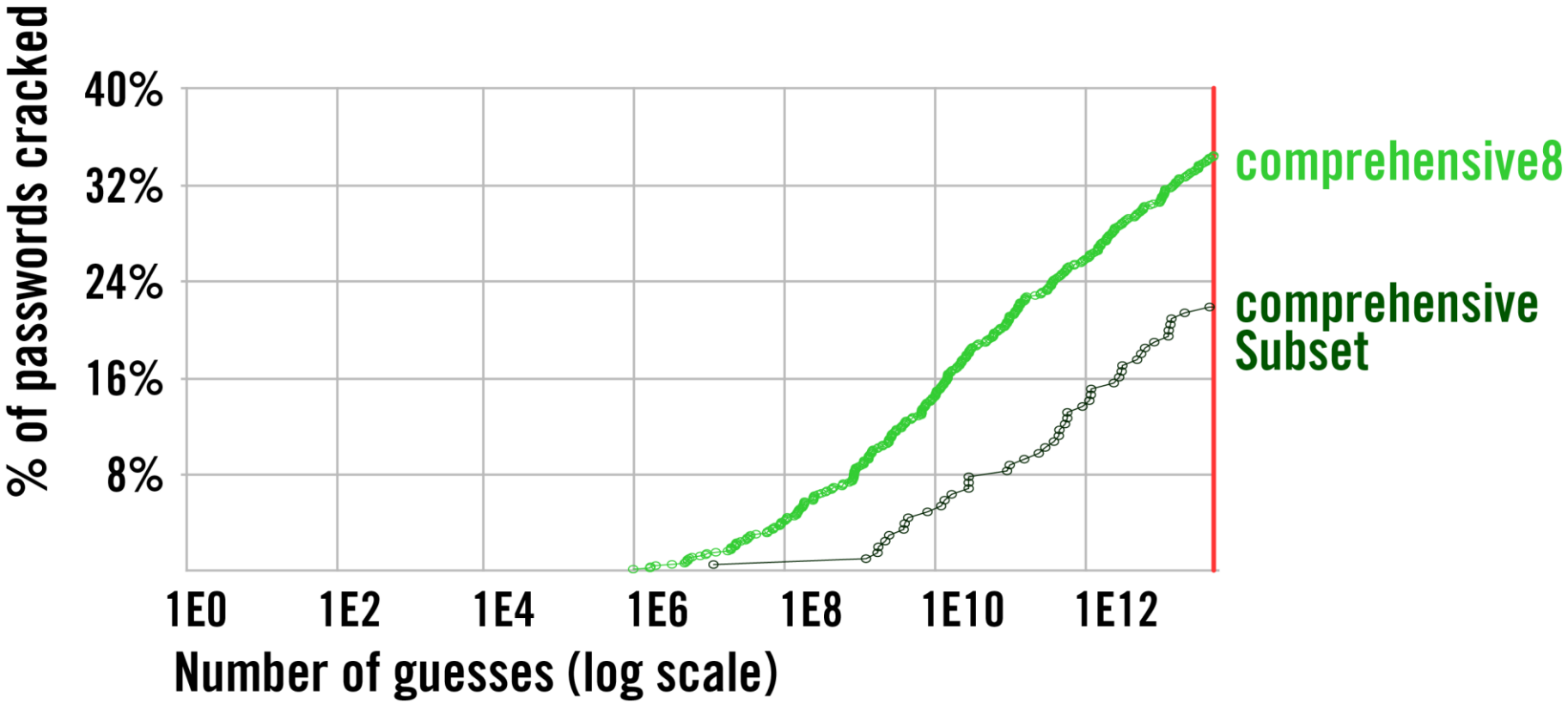
Choosing the Right Test Data



Choosing the Right Test Data



Choosing the Right Test Data



Carefully choosing test passwords is critical when evaluating policies

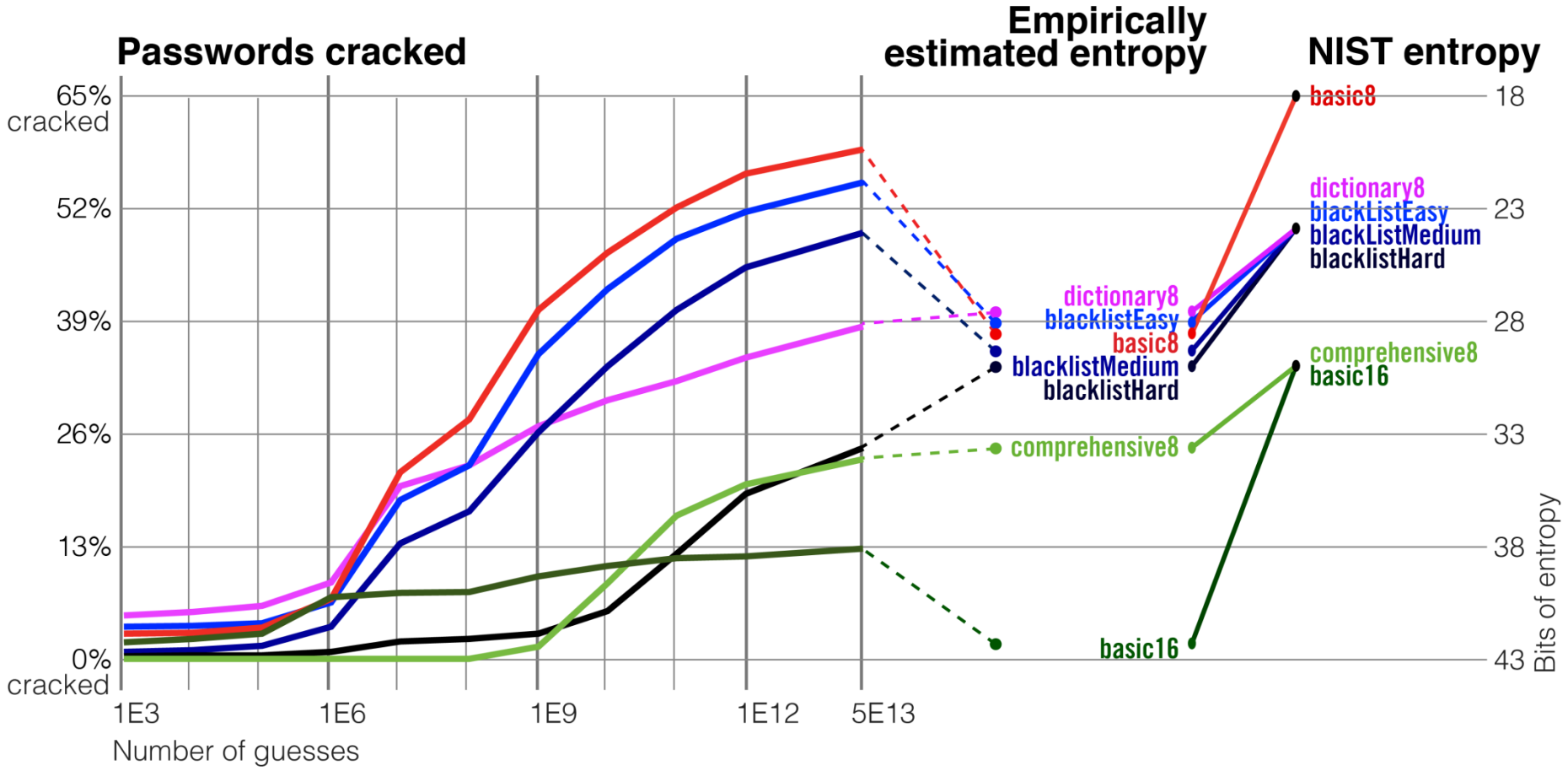
Contributions

- Measured guessability across seven password-composition policies
  - Threat model: offline attack

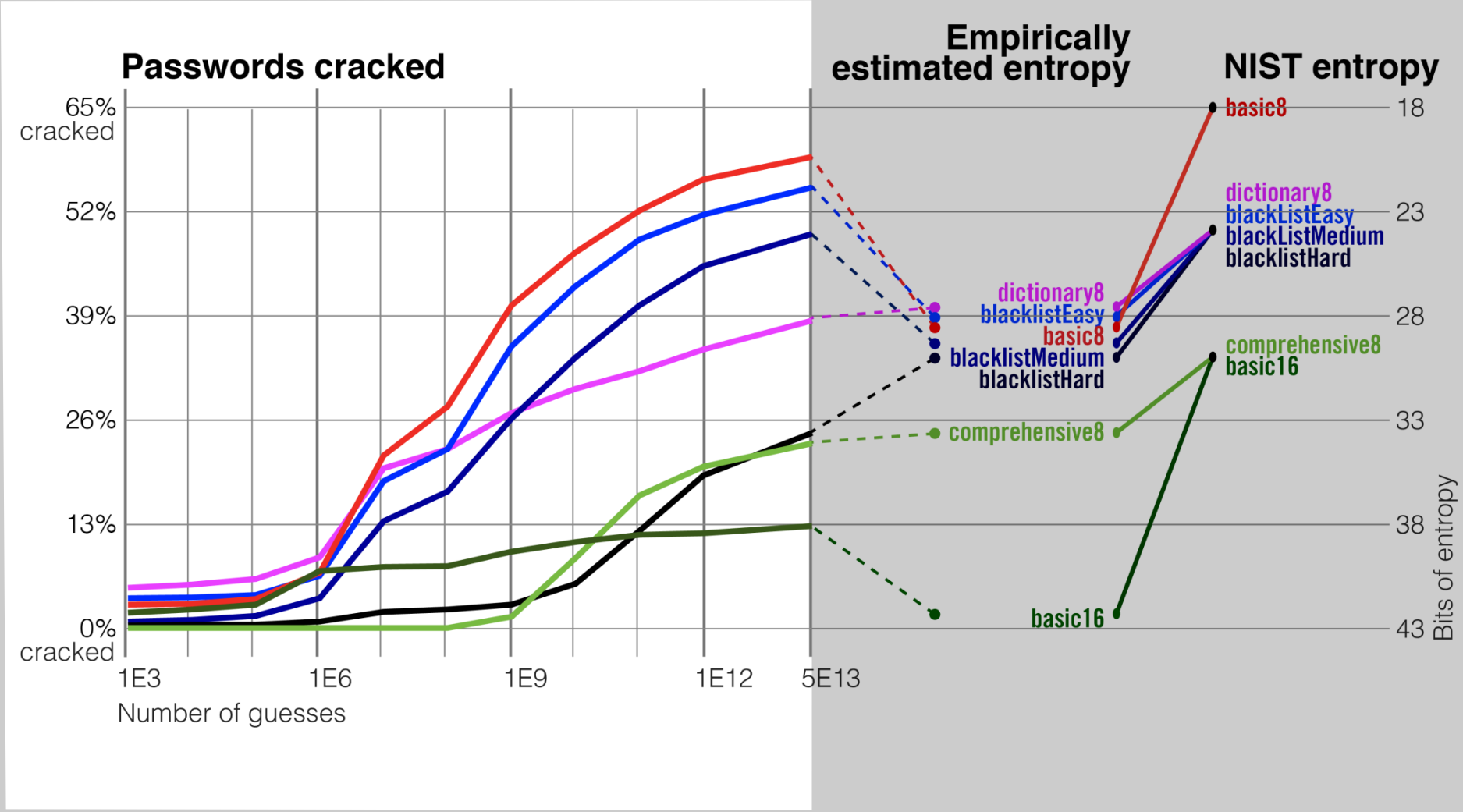
- Studied the impact of tuning and test-set selection on policy evaluation

➢ Compare security metrics across policies
  - Correlate security with usability

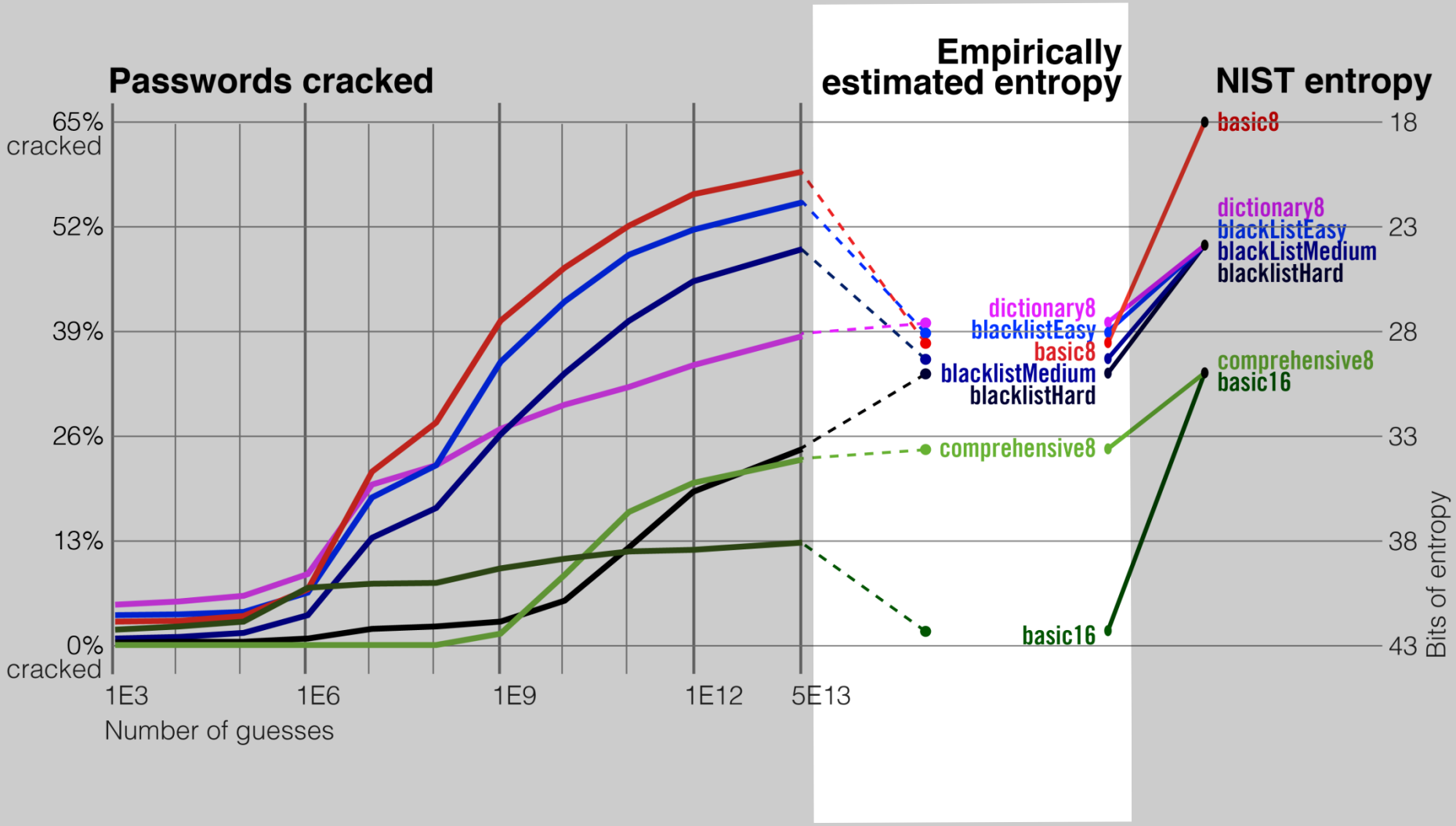
Comparing Metrics



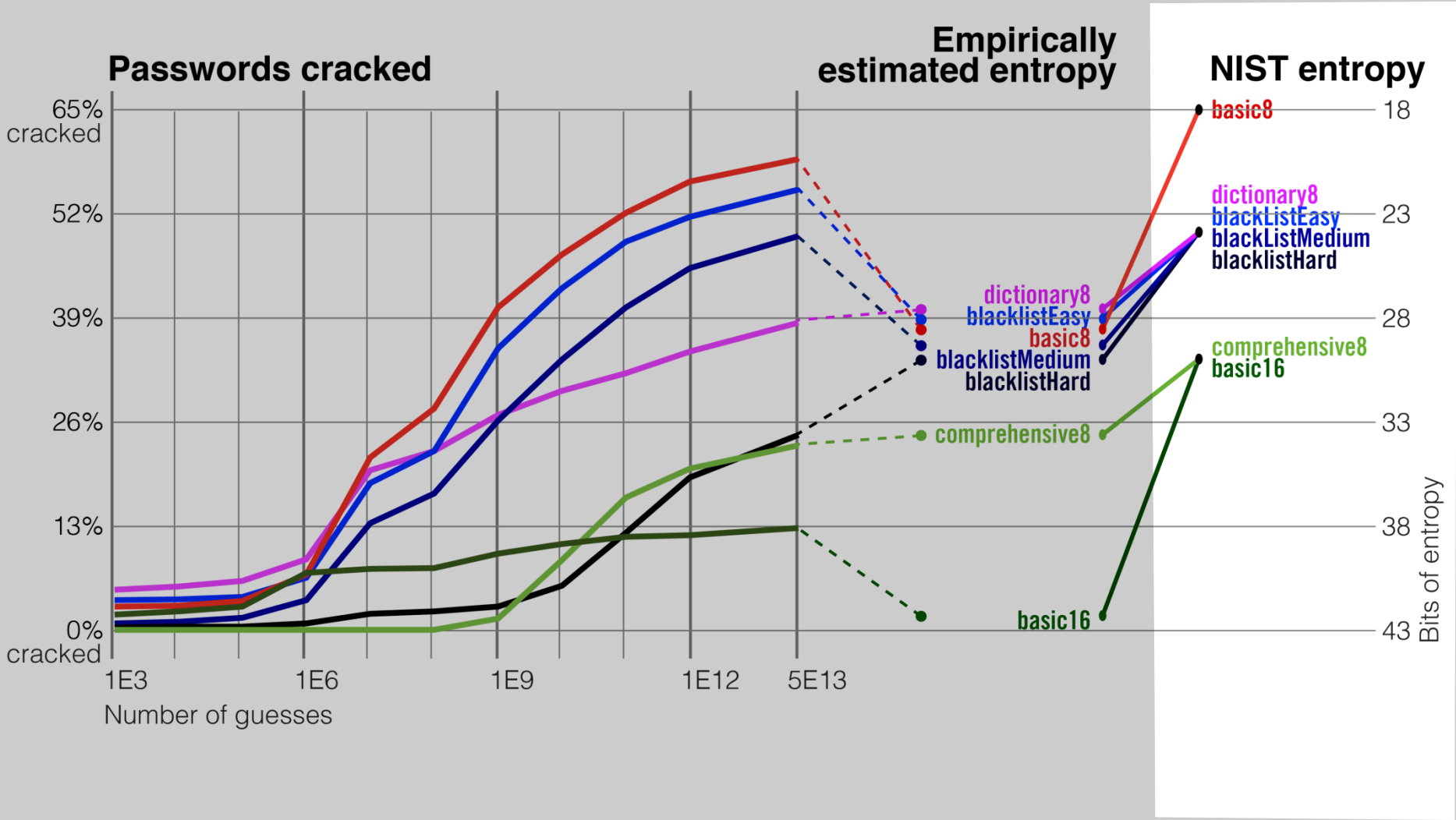
Comparing Metrics



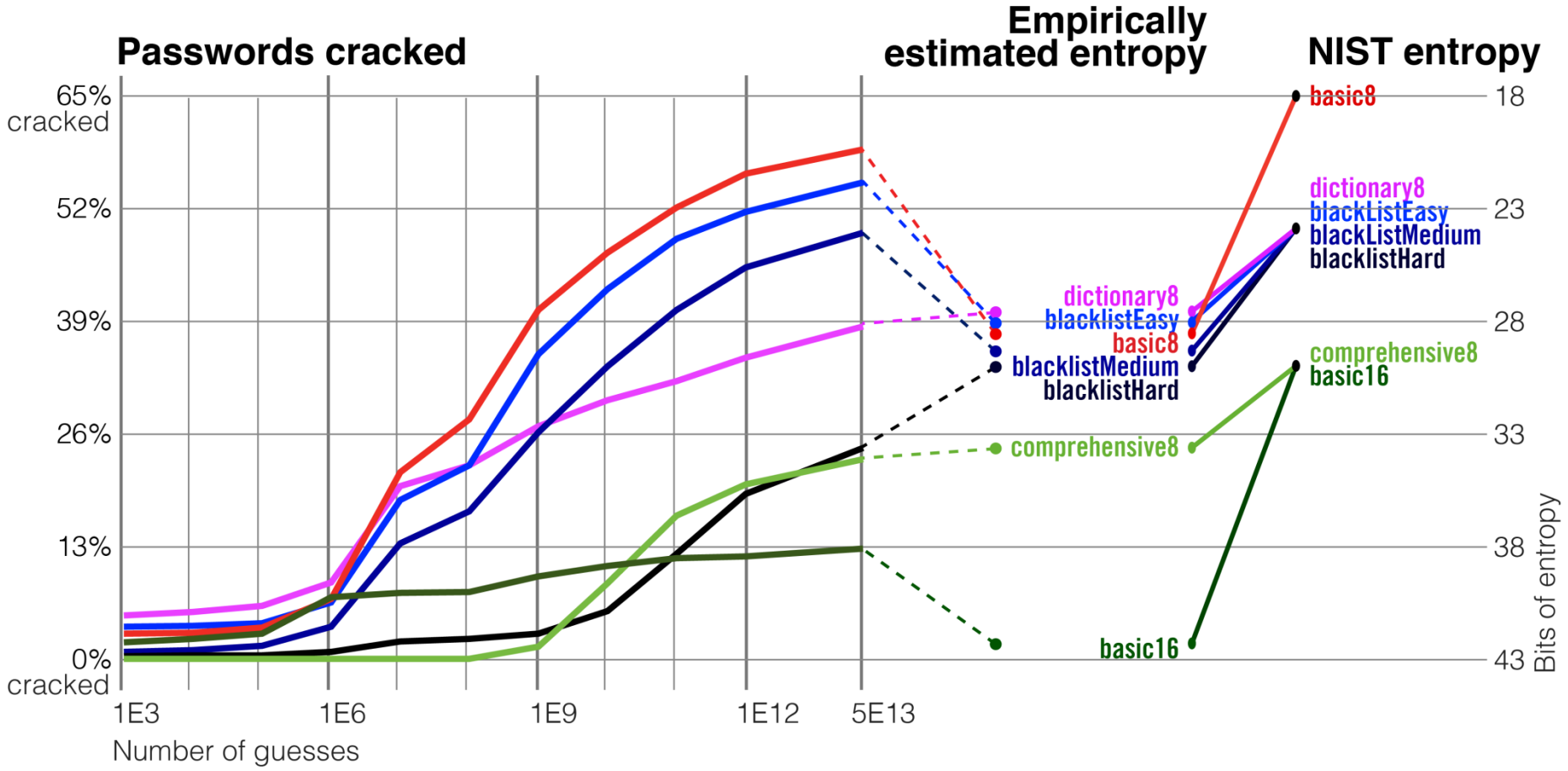
Comparing Metrics



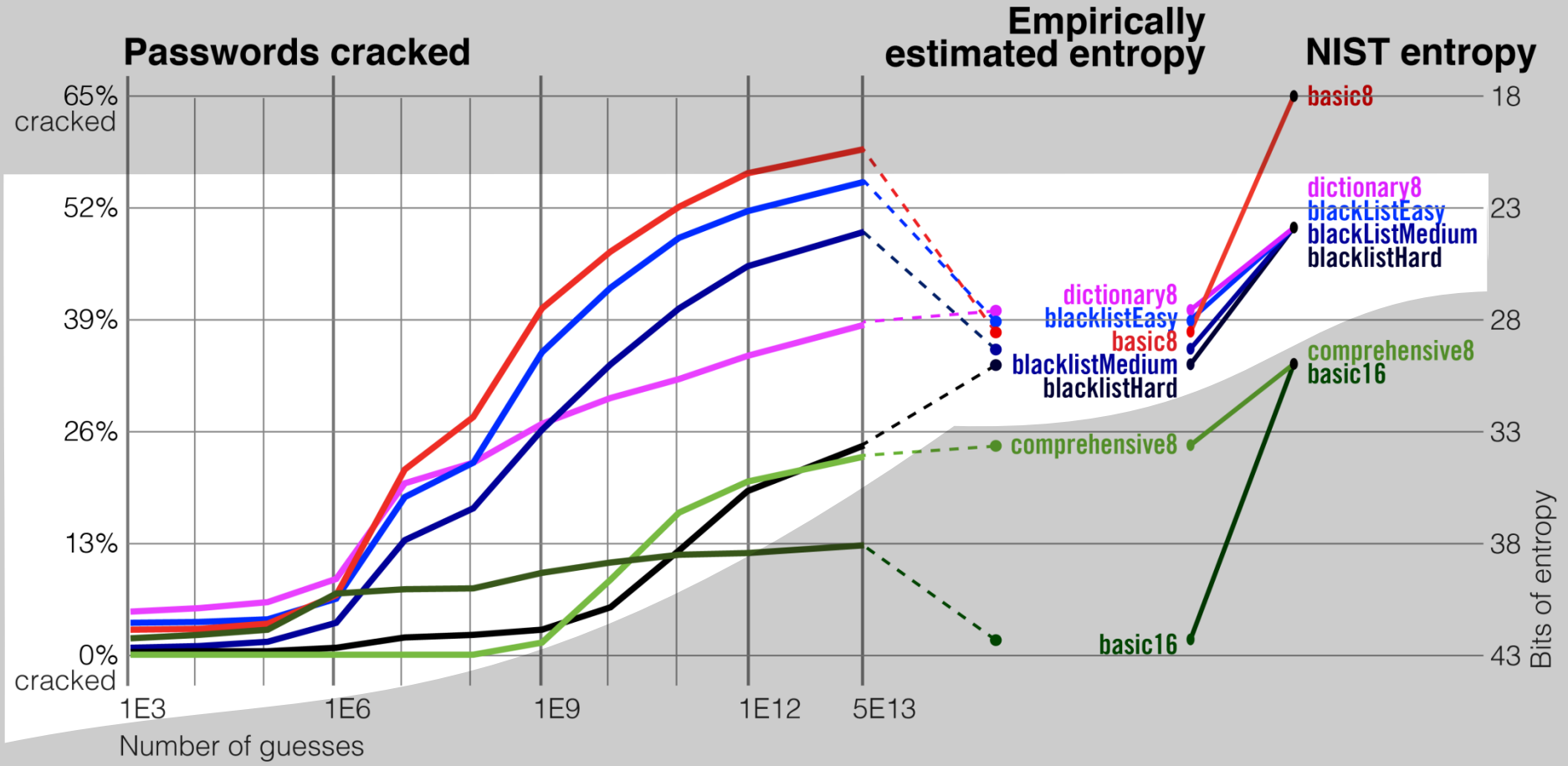
Comparing Metrics



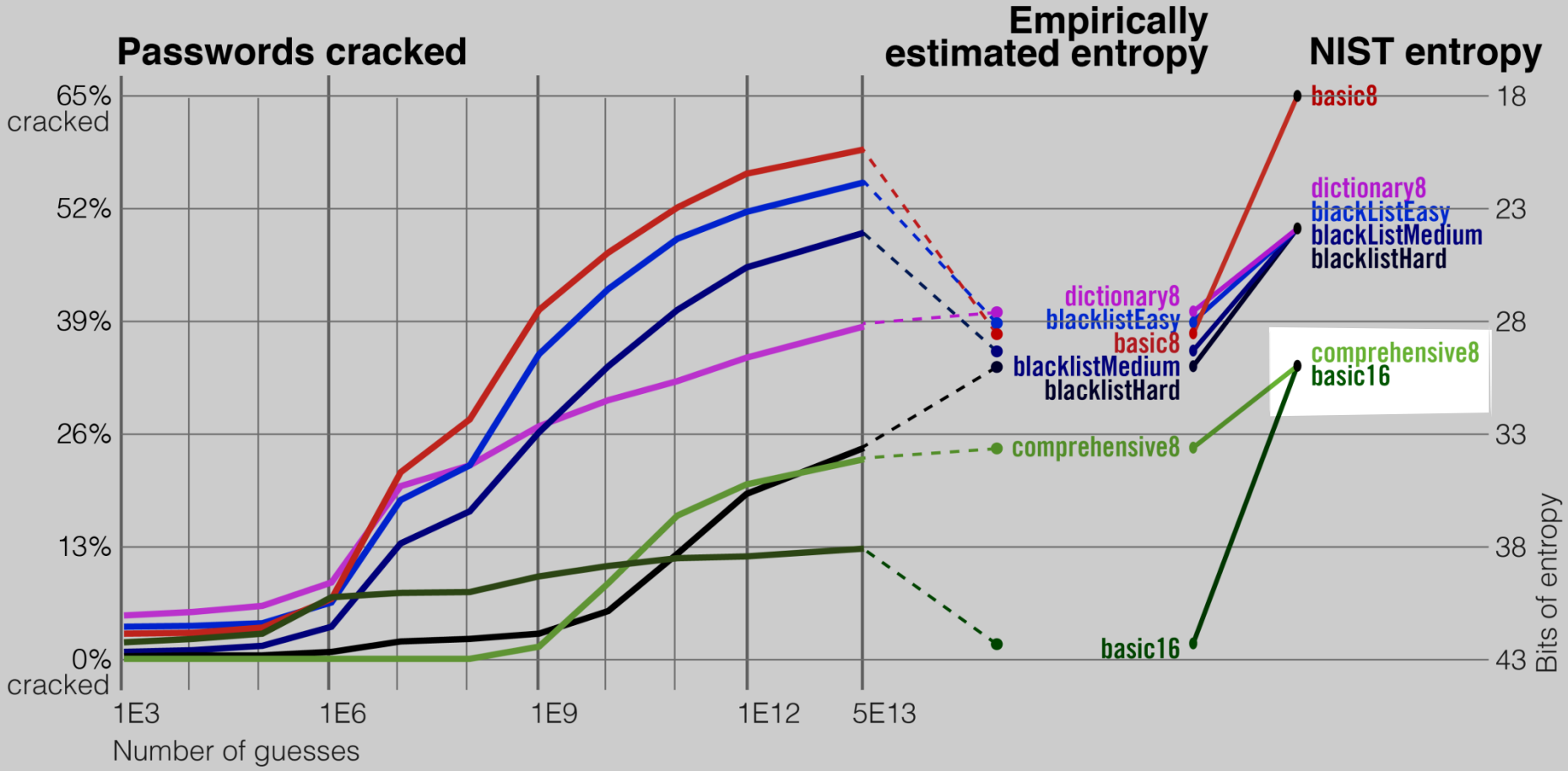
Comparing Metrics



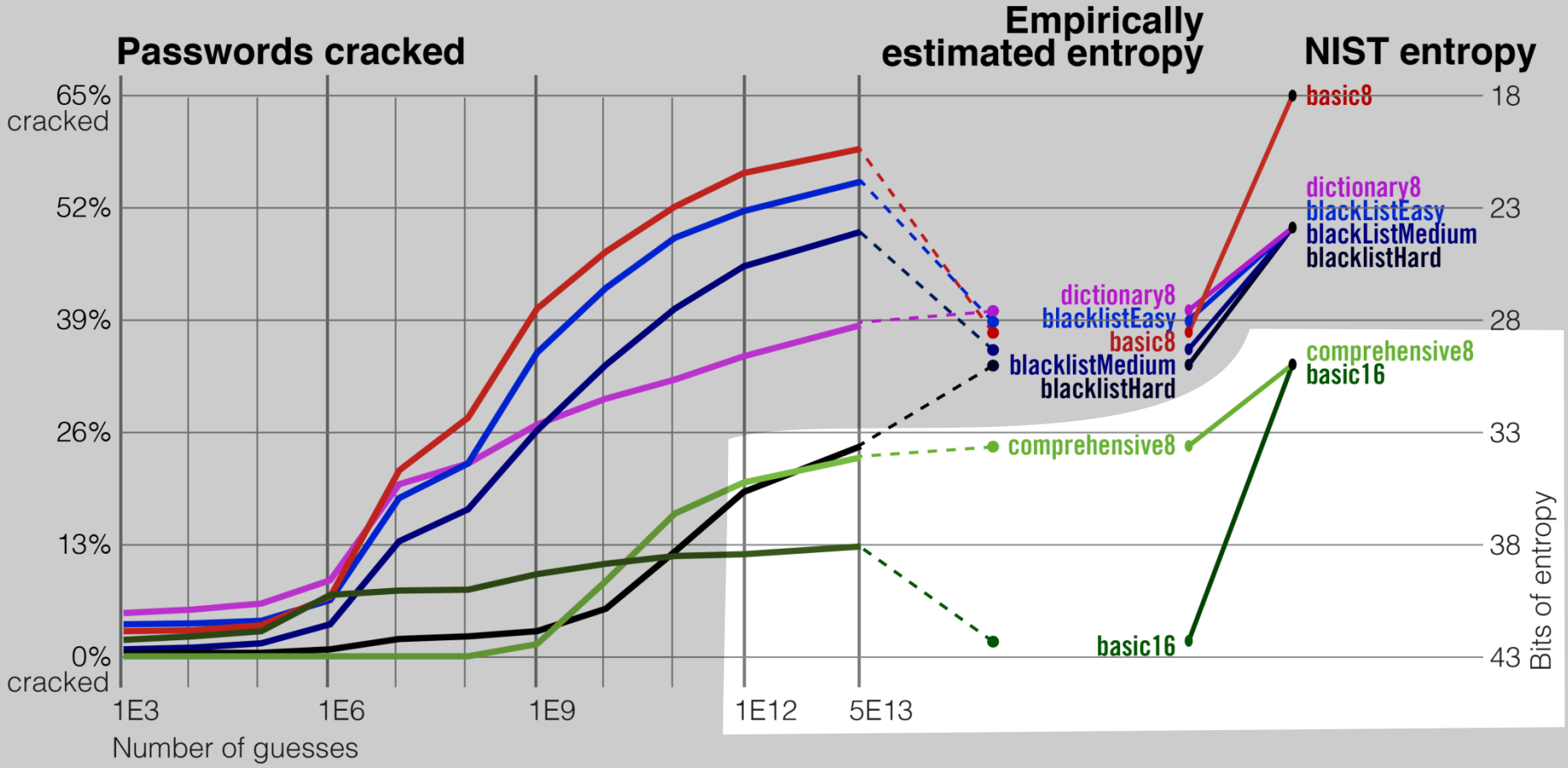
Comparing Metrics



Comparing Metrics



Comparing Metrics



Usability - Basic16 & Comprehensive8

- Basic16 is more usable [Our previous work 2011]
  - Fewer participants wrote down password (50% vs. 33%)
  - Self-reported difficulty and annoyance was lower

Basic16 appears to be more secure and more usable than comprehensive8

Conclusions

- In some cases, more secure ≠ less usable

Conclusions

- In some cases, more secure ≠ less usable

- Complex policies are tricky to analyze
  - Need high-quality training data
  - Important to choose test data carefully

Conclusions

- In some cases, more secure ≠ less usable
- Complex policies are tricky to analyze
  - Need high-quality training data
  - Important to choose test data carefully
- Existing guidance is not very helpful

Cylab Usable Privacy and Security Laboratory

http://cups.cs.cmu.edu/

Carnegie Mellon

Questions?

Existing Guidance

- NIST guide not based on empirical evidence
  - Provides a means of "scoring" password policies

NIST would like to obtain more data on the passwords users actually choose, but, where they have the data, system administrators are understandably reluctant to reveal password data to others. – [Burr 2006]

NIST
National Institute of Standards and Technology
Technology Administration
U.S. Department of Commerce

Weir's Algorithm

Presented at Oakland in 2009

- Learns probabilities from training data

- Generates new guesses based on likelihood

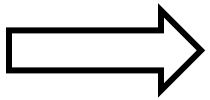
Weir's Algorithm [Weir et al. (Oakland) 2009]

Learned Elements

Training data

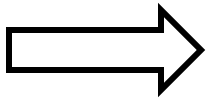
pass#word

Best!123



strings		symbols		digits	
pass	⅓	#	½	123	1
word	⅓	!	½		
best	⅓				

structures



$L_4S_1L_4$ ½

$(UL_3)S_1D_3$ ½

Weir's Algorithm [Weir et al. (Oakland) 2009]

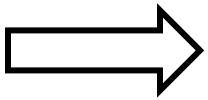
Learned Elements

strings		symbols		digits	
pass	⅓	#	½	123	1
word	⅓	!	½		
best	⅓				

structures

$L_4S_1L_4$ ½

$(UL_3)S_1D_3$ ½



Guesses

Pass#123	1/12
Pass!123	1/12
Word#123	1/12
Word!123	1/12
Best#123	1/12
Best!123	1/12
pass#pass	1/36
pass#word	1/36
pass#best	1/36
pass!pass	1/36
pass!word	1/36
…	

Weir's Algorithm [Weir et al. (Oakland) 2009]

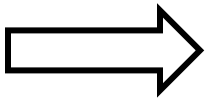
Learned Elements

strings		symbols		digits	
pass	⅓	#	½	123	1
word	⅓	!	½		
best	⅓				

structures

$L_4S_1L_4$ ½

$(UL_3)S_1D_3$ ½



Lookup Table

Pass#123	1/12	1
pass#pass	1/36	7
Total guesses:		24

Basic8 frequencies

12345678	1.3%
Password	0.7%
123456789	0.6%

Five appeared twice

Rest were unique

N = 1000

Demographics

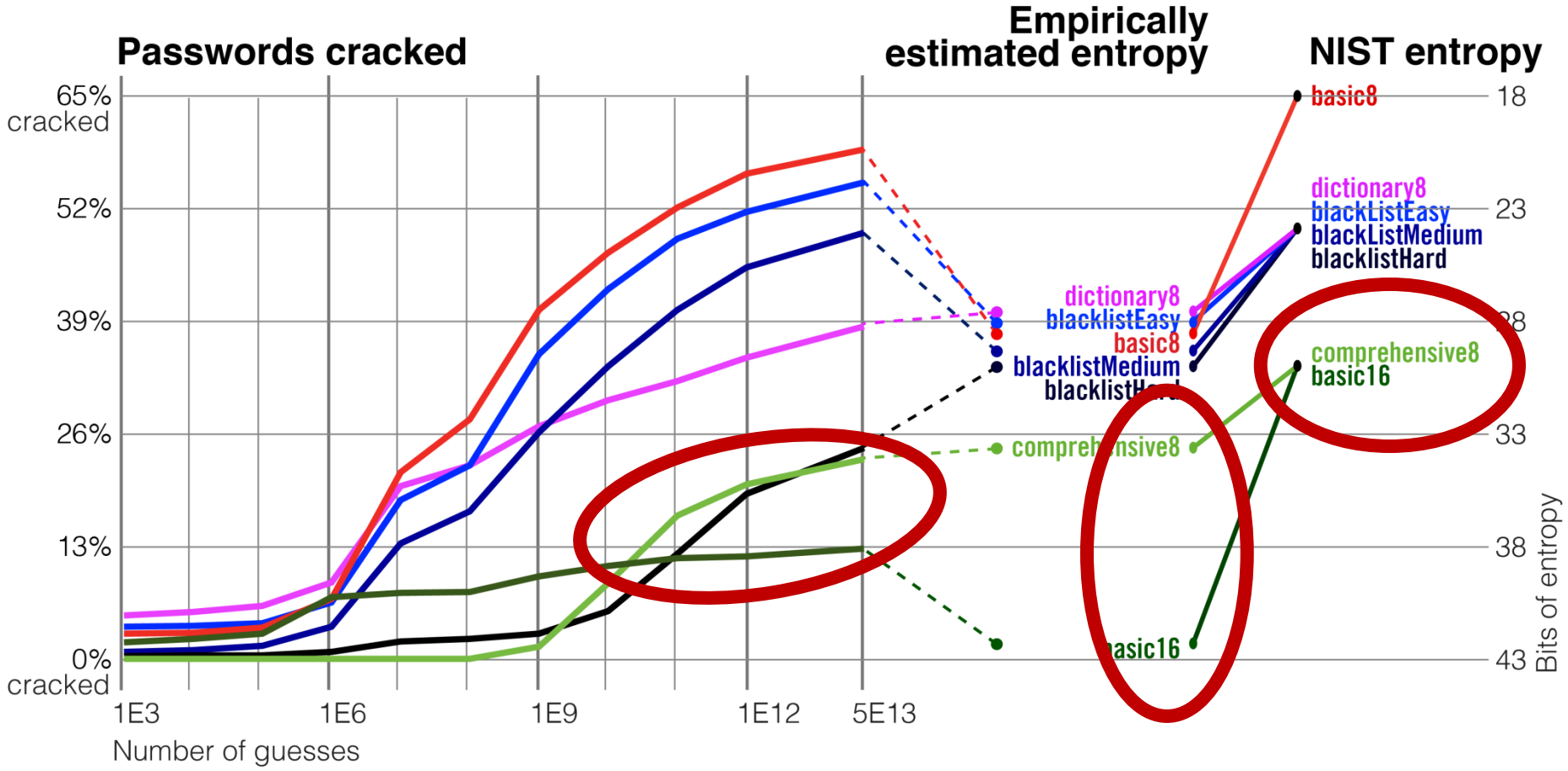
- 1,000 participants per condition
- 51% male, 47% female
- Mean age: 29.8 years
- No significant difference across conditions
- 2,889 returned within three days of follow-up email

Hypothetical Email Scenario

Imagine that your main email service provider has been attacked, and your account became compromised. You need to create a new password for your email account, since your old password may be known by the attackers. Because of the attack, your email service provider is also changing its password rules.

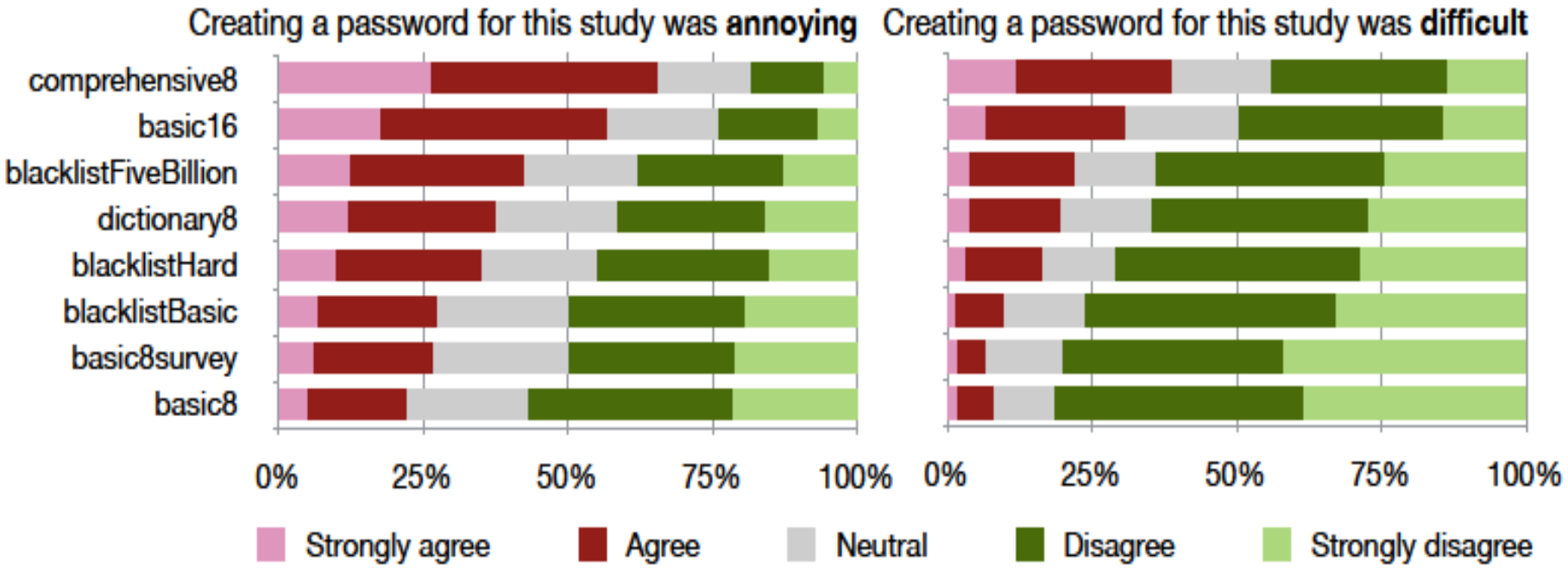
Please follow the instructions below to create a new password for your email account. We will ask you to use this password in a few days to log in again so it is important that you remember your new password. Please take the steps you would normally take to remember your email password and protect this password as you normally would protect the password for your email account. Please behave as you would if this were your real password!

Comparing Metrics



Basic16 vs Comprehensive8

- Basic16 requires significantly fewer attempts in password creation
  - 53% vs 18% success on first attempt, $p < 0.001$
  - 1.66 vs 3.35 attempts total, $p < 0.001$
- Comprehensive8 participants had significantly higher dropout rates
  - 19% vs 25%, $p < 0.001$


Creating a password for this study was annoying
Creating a password for this study was difficult
comprehensive8
basic16
blacklistFiveBillion
dictionary8
blacklistHard
blacklistBasic
basic8survey
basic8
0%
25%
50%
75%
100%
Strongly agree
Agree
Neutral
Disagree
Strongly disagree