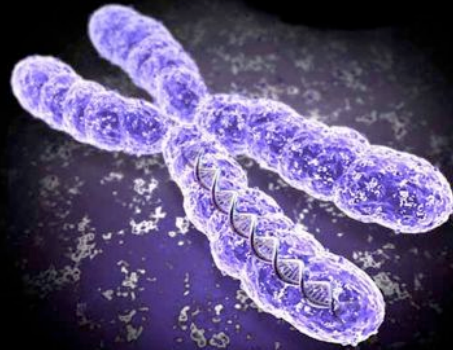


Estimating the Unseen (getting more from your data)

Gregory Valiant

UC Berkeley -> Microsoft -> Stanford

Data, Data, Everywhere



Data, Data, Everywhere

New sorts/scales of datasets

New sorts of *algorithmic* challenges

Central Challenge: How to use data *efficiently*

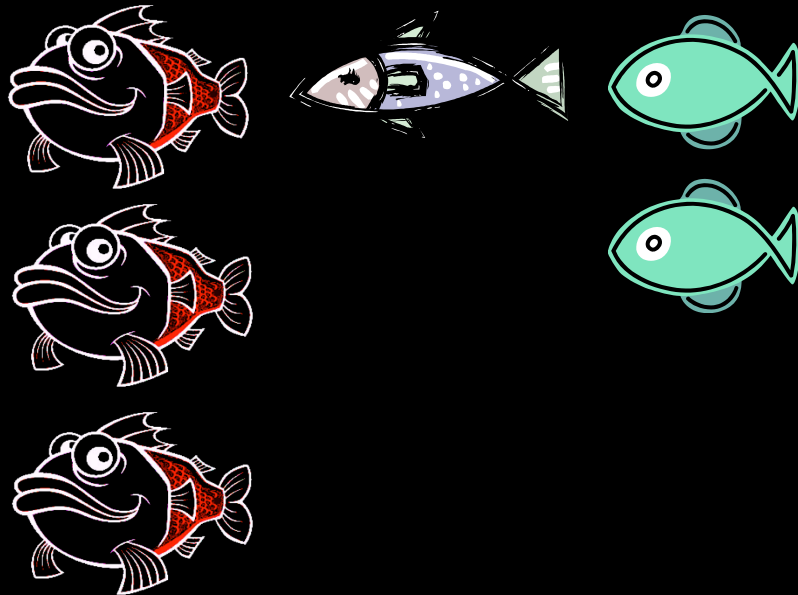
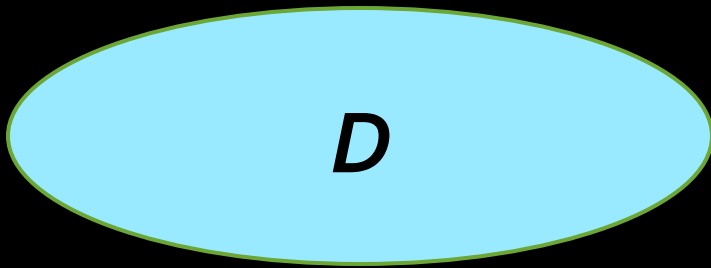
3 reasons Data *efficiency* is **increasingly** important
~~despite~~ so much data
because

- In some domains, datasets can't grow much more
- Large datasets, but even more complex objects
- In general: the more resources, the more potential....

Many fundamental questions still unanswered!!

A Basic Question

Given independent samples from a distribution (of discrete support):



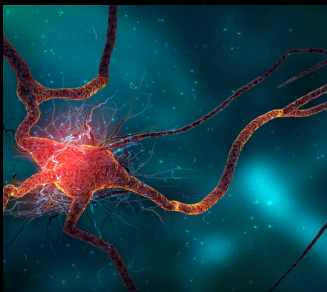
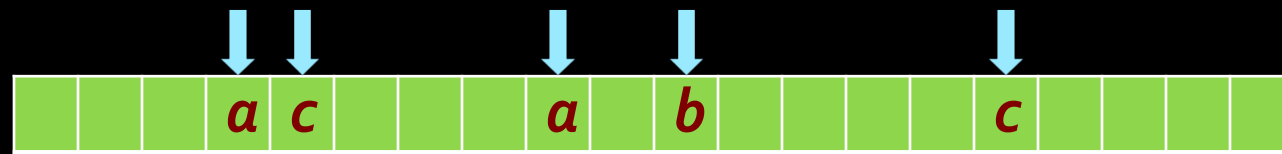
Empirical distribution \leftrightarrow optimally approximates *seen* portion of distribution

- What can we infer about the *unseen* portion?
- How can inferences about the unseen portion yield better estimates of distribution properties?

Some concrete problems



Q1: Given a length n vector, how many indices must we look at to estimate **# distinct** elements, to $\pm \epsilon n$ (w.h.p)? [**distinct elements problem**]



Q2: Given indep. samples from D supported on $\{1, \dots, n\}$, how many samples required to estimate **entropy(D)** to within $\pm \epsilon$ (w.h.p)?

myspace

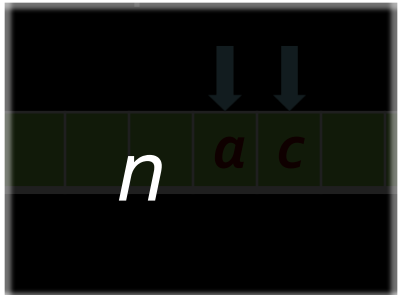
vs

myspace

Q3: Given samples from D_1 and D_2 supported on $\{1, 2, \dots, n\}$, how many samples required to estimate **Dist(D_1, D_2)** to within $\pm \epsilon$ (w.h.p)?

...

Some concrete problems

	Trivial	Previous	Answer
Distinct Elements	 <p>n</p>	$O(n)$ [Bar Yossef et al. '01] [P. Valiant, '08] [Raskhodnikova et al. '09]	
Entropy		$O(n)$ [Batu et al. '01, '02] [Paninski, '03, '04] [Dasgupta et al, '05]	$\Theta\left(\frac{n}{\log n}\right)$
Distance	$O(n \log n)$	$O(n)$ [Goldreich et al. '96] [Batu et al. '00, '01]	
...	

R.A. Fisher's Butterflies

How many new species if I observe for another period?

$$h_1 - h_2 + h_3 - h_4 + h_5 - \dots$$



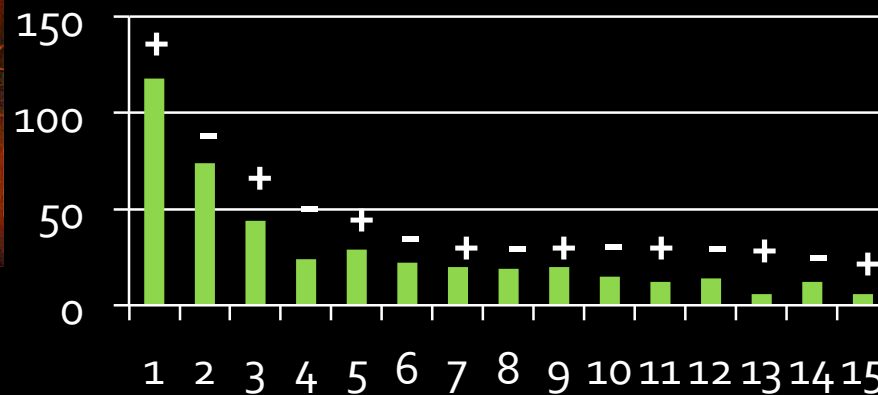
Turing's Enigma Codewords

Probability mass of unseen codewords



$$h_1 / (\text{number of samples})$$

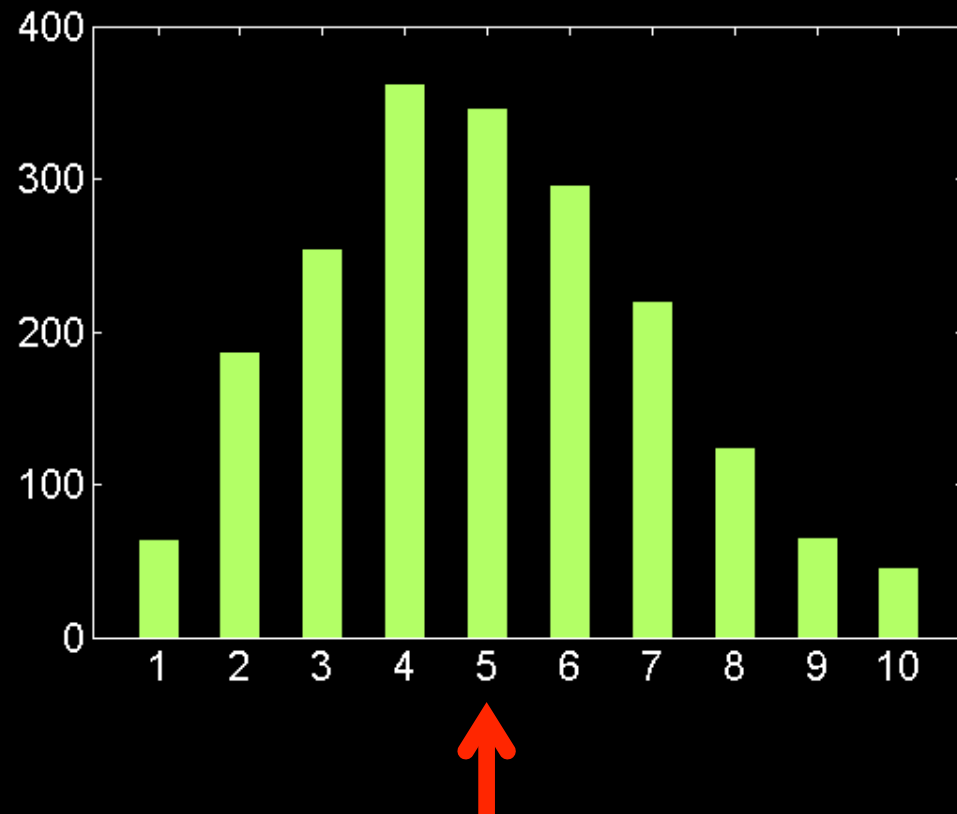
Corbet's Butterfly Data



("Histogram" of the samples)

Reasoning Beyond the Empirical Distribution

Histogram based on 10000 samples:

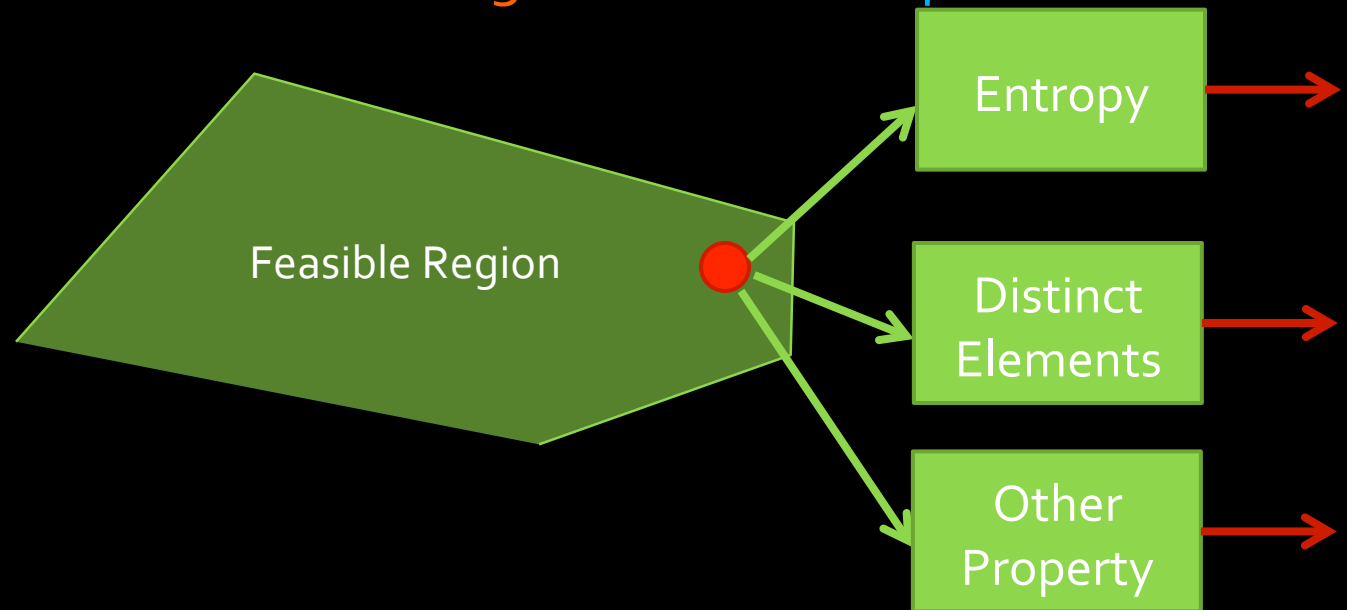


Maximum Likelihood Interpretation

What is the distribution that
maximizes the likelihood of yielding
the observed **histogram**
(among distributions of support n) ?

Linear Programming

“Find distributions whose **expected histogram** are close to the **observed histogram of the samples**”

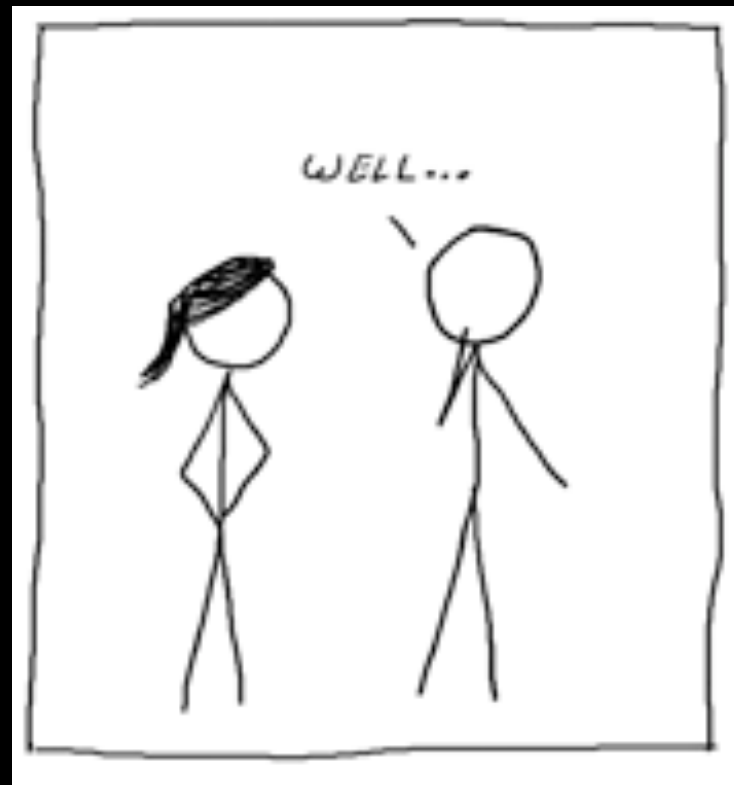


Technical challenge: show “diameter” of feasible region is small

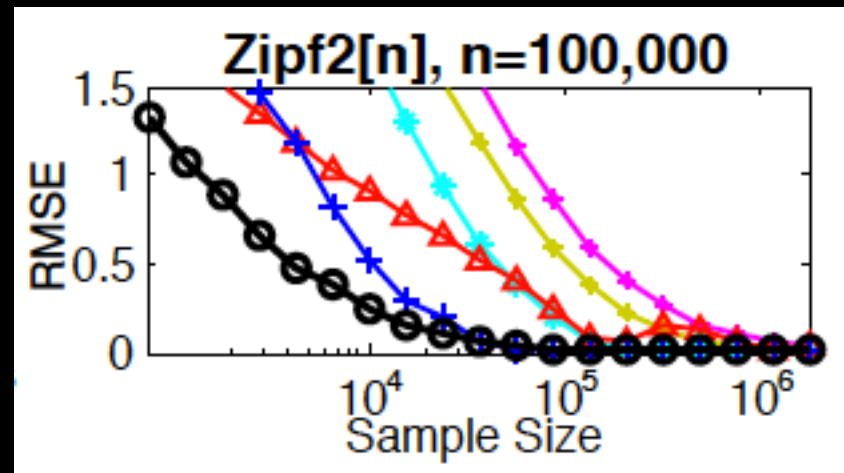
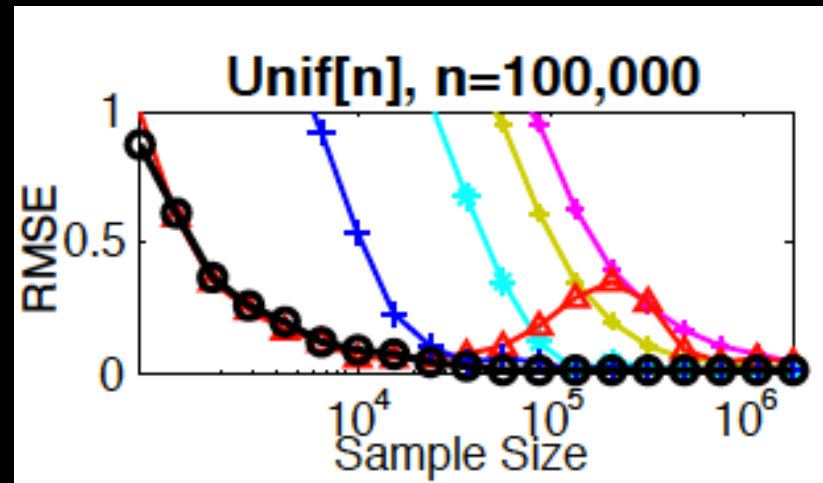
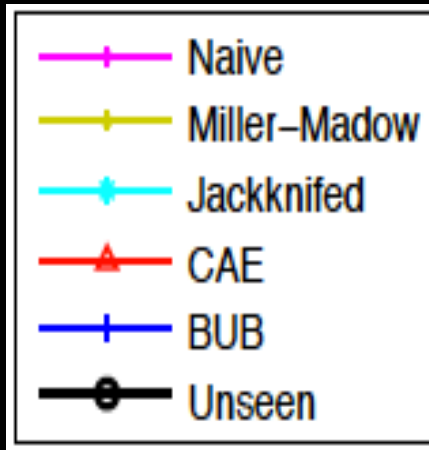
$\Theta(n/\log n)$ samples, and **OPTIMAL**

So...does this actually work in practice?

YES!!

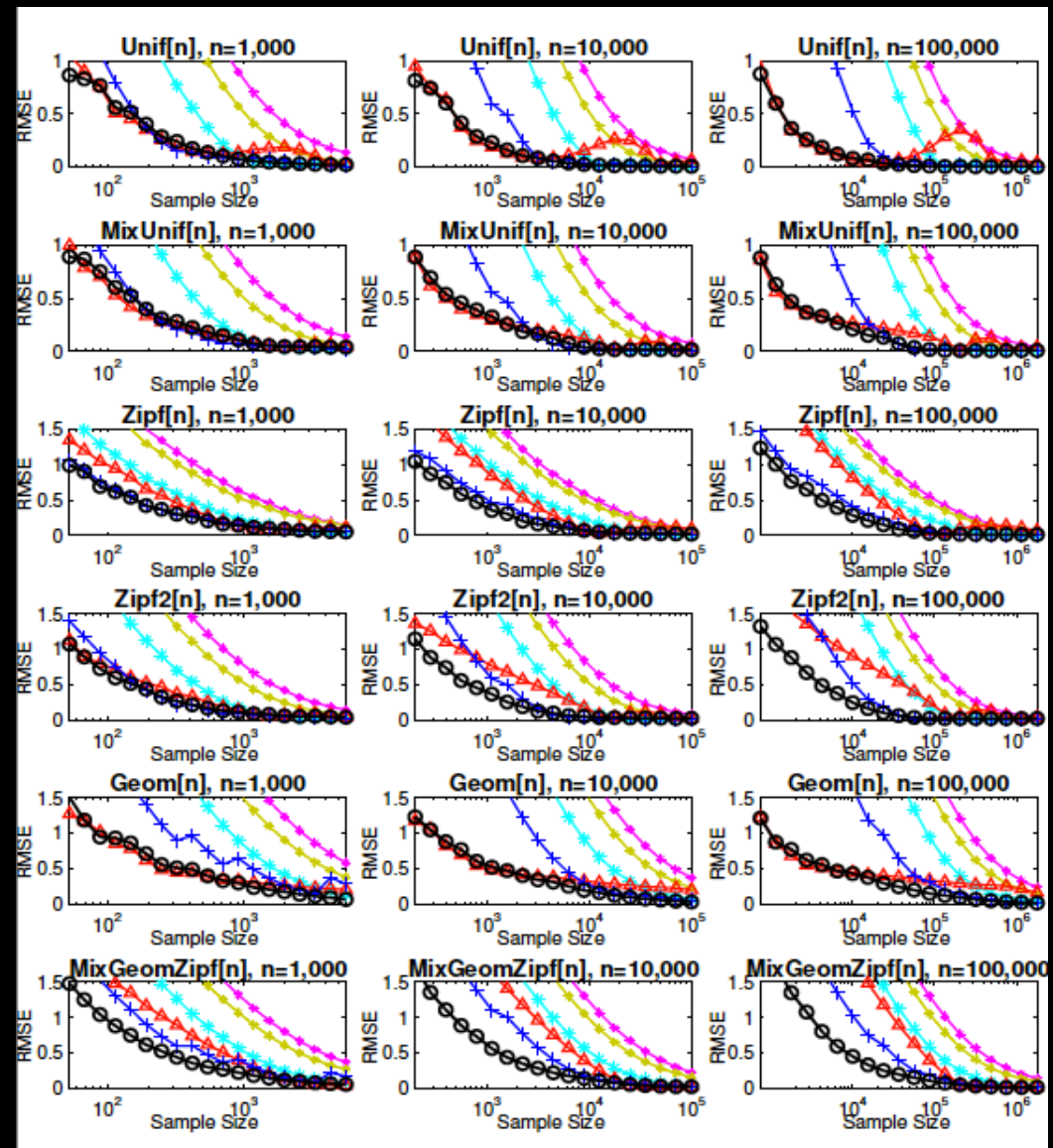
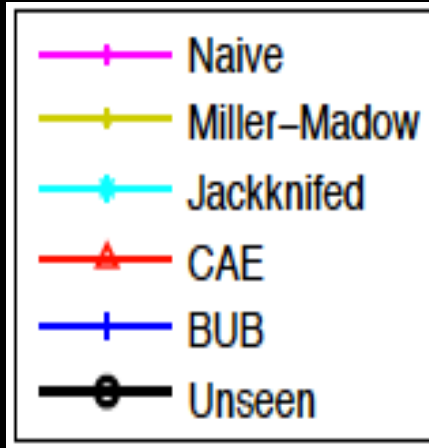


Performance in Practice (entropy)



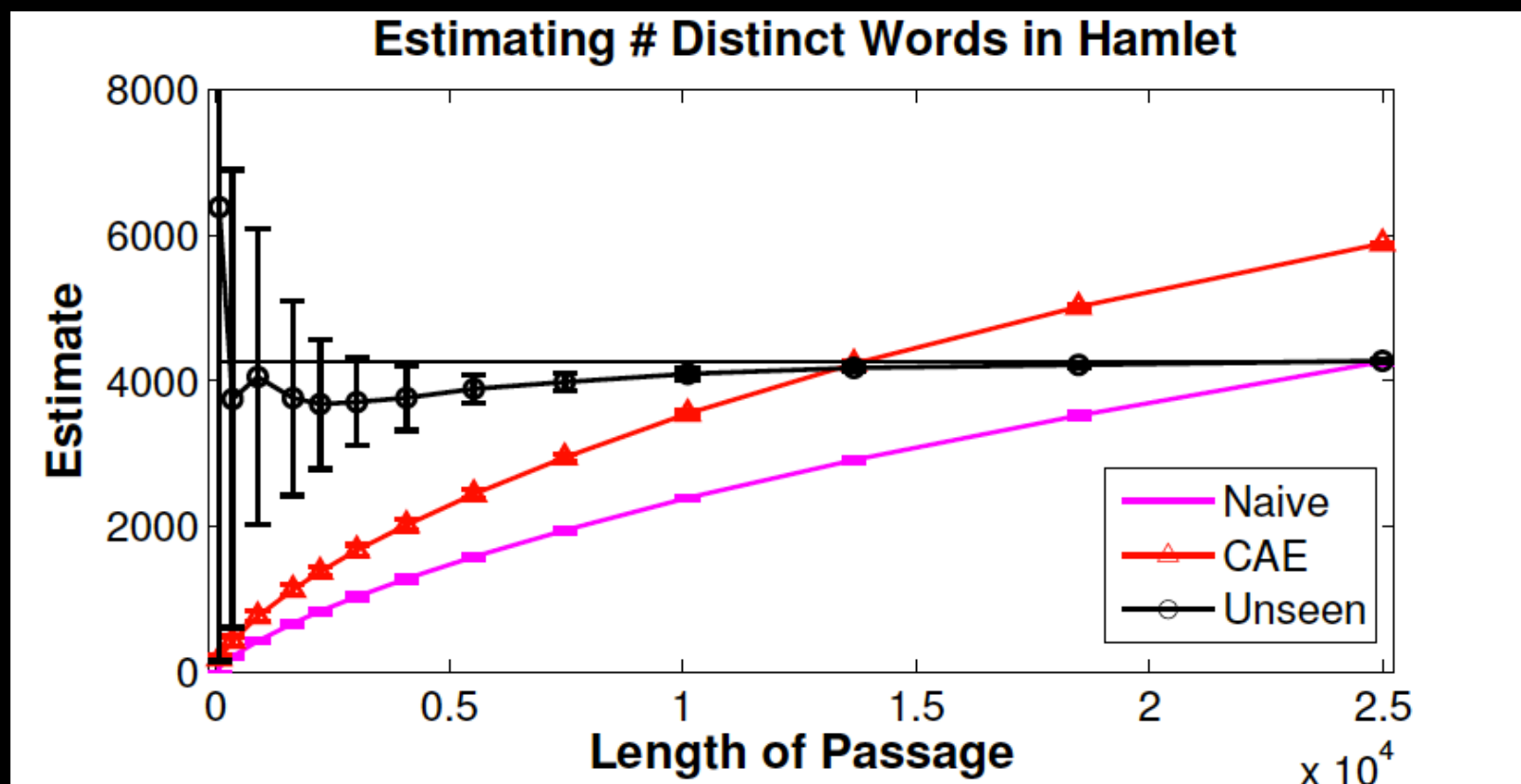
Zipf: power law distr.
 $p_j \propto 1/j$ (or $1/j^c$)

Performance in Practice (entropy)



Performance in Practice (support size)

Task: Pick a (short) passage from *Hamlet*, then estimate # distinct words in *Hamlet*



The Big Picture

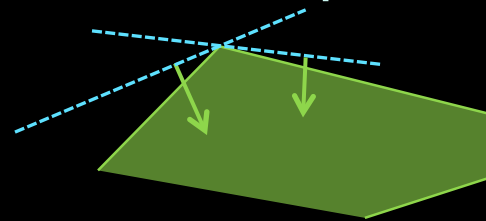
Estimating Statistical Properties

100+ years of statistics

“Linear estimators”

$$c_1 \cdot h_1 + c_2 \cdot h_2 + c_3 \cdot h_3 + \dots$$

[Our Approach]
Linear Programming
Substantial improvements!



“What richness of algorithmic machinery is necessary to effectively solve these problems?”

Thm [VV]:

Exist near-optimal “Linear Estimators”, but...