# Taming probability distributions over big domains

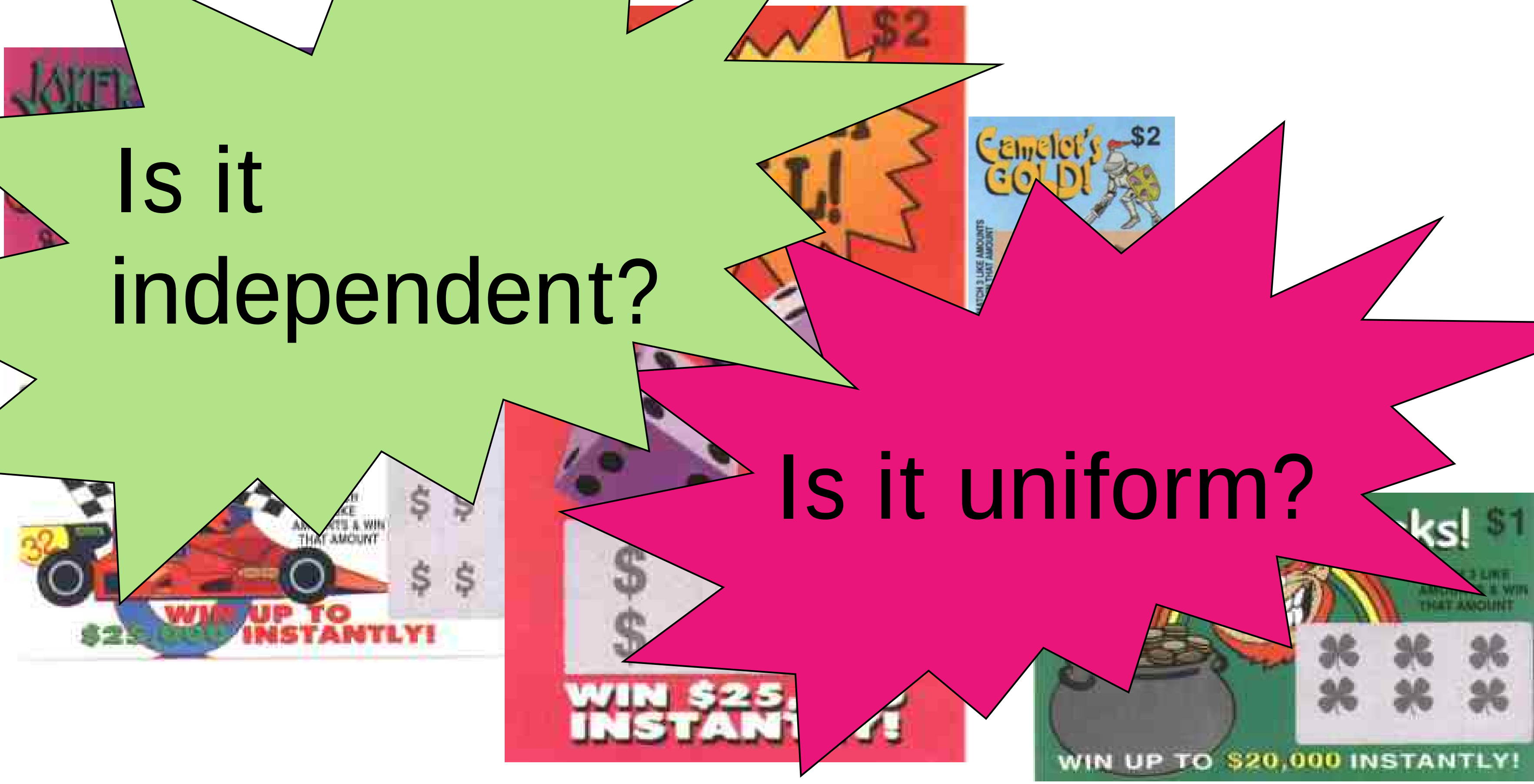
Ronitt Rubinfeld
MIT and Tel Aviv University

# What properties do your big distributions have?

# Play the lottery?

Is it independent?

Is it uniform?



# Is the lottery unfair?

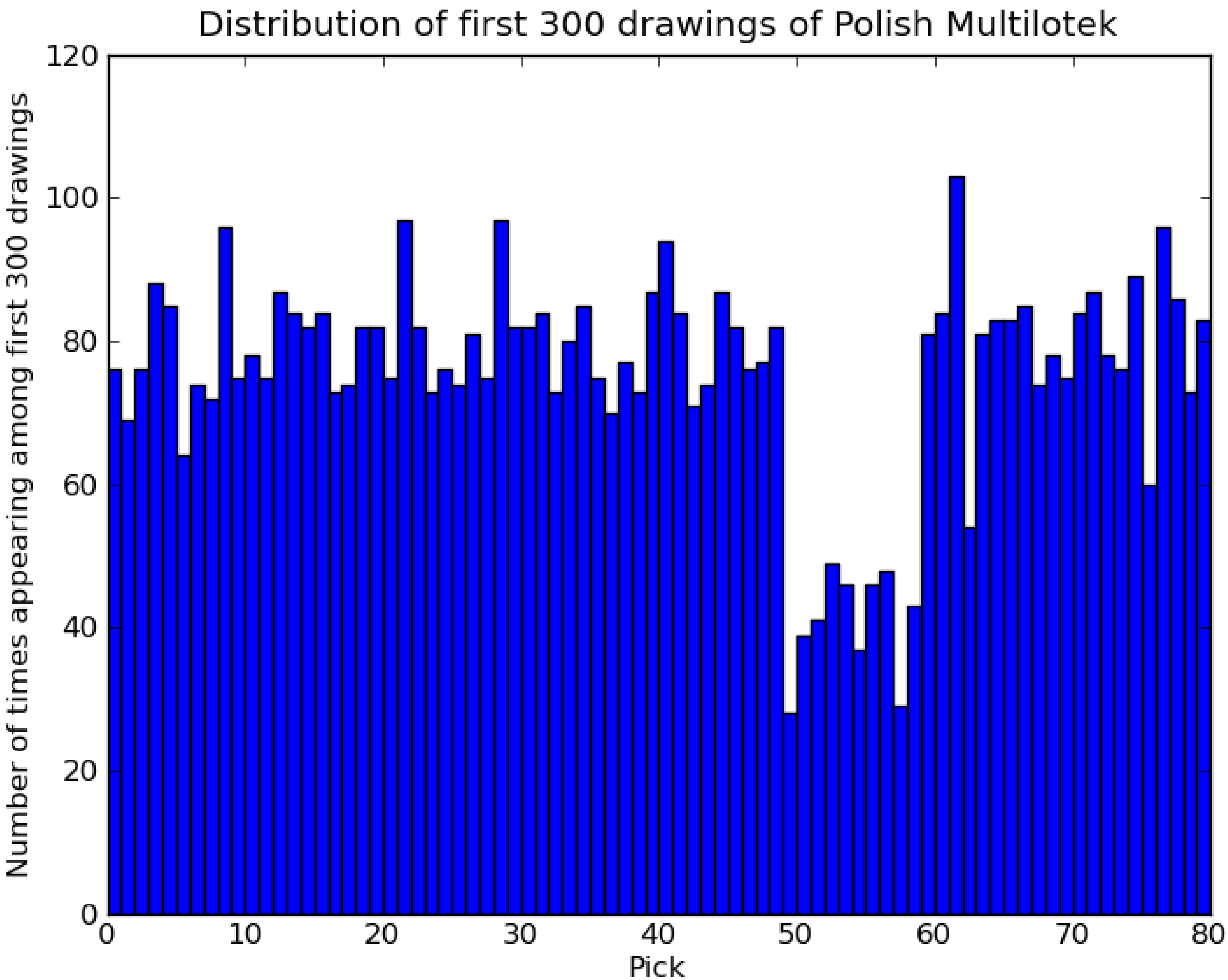
- From Hitlotto.com:  Lottery experts agree, past number histories can be the key to predicting future winners.

# True Story!

- Polish lottery Multilotek
  - Choose “uniformly” at random distinct 20 numbers out of 1 to 80.
  - Initial machine biased
    - e.g., probability of 50-59 too small

- Past results:

http://serwis.lotto.pl:8080/archiwum/wyniki_wszystkie.php?id_gra=2




Thanks to Krzysztof Onak (pointer) and Eric Price (graph)

# New Jersey Pick 3,4 Lottery

- New Jersey Pick *k ( =3,4)* Lottery.
  - Pick *k* digits in order.
  - $10^k$ possible values.
  - Assume lottery draws iid
- Data:
  - Pick 3 - 8522 results from 5/22/75 to 10/15/00
    - $\chi^2$–test gives 42% confidence
  - Pick 4 - 6544 results from 9/1/77 to 10/15/00.
    - fewer results than possible values
    - $\chi^2$–test gives no confidence

# Distributions on BIG domains

- Given samples of a distribution, need to know, e.g.,
  - entropy
  - number of distinct elements
  - “shape” (monotone, bimodal,…)
  - closeness to uniform, Gaussian, Zipfian…
  - Ability to generate the distribution?

- No assumptions on shape of distribution
  - i.e., smoothness, monotonicity, normal distribution,…

- Considered in statistics, information theory, machine learning, databases, algorithms, physics, biology,…

# Key Question

- How many samples do you need in terms of domain size?
  - Do you need to estimate the probabilities of each domain item?
  - Can sample complexity be *sublinear* in size of the domain?

Rules out standard statistical techniques

# Our Aim:

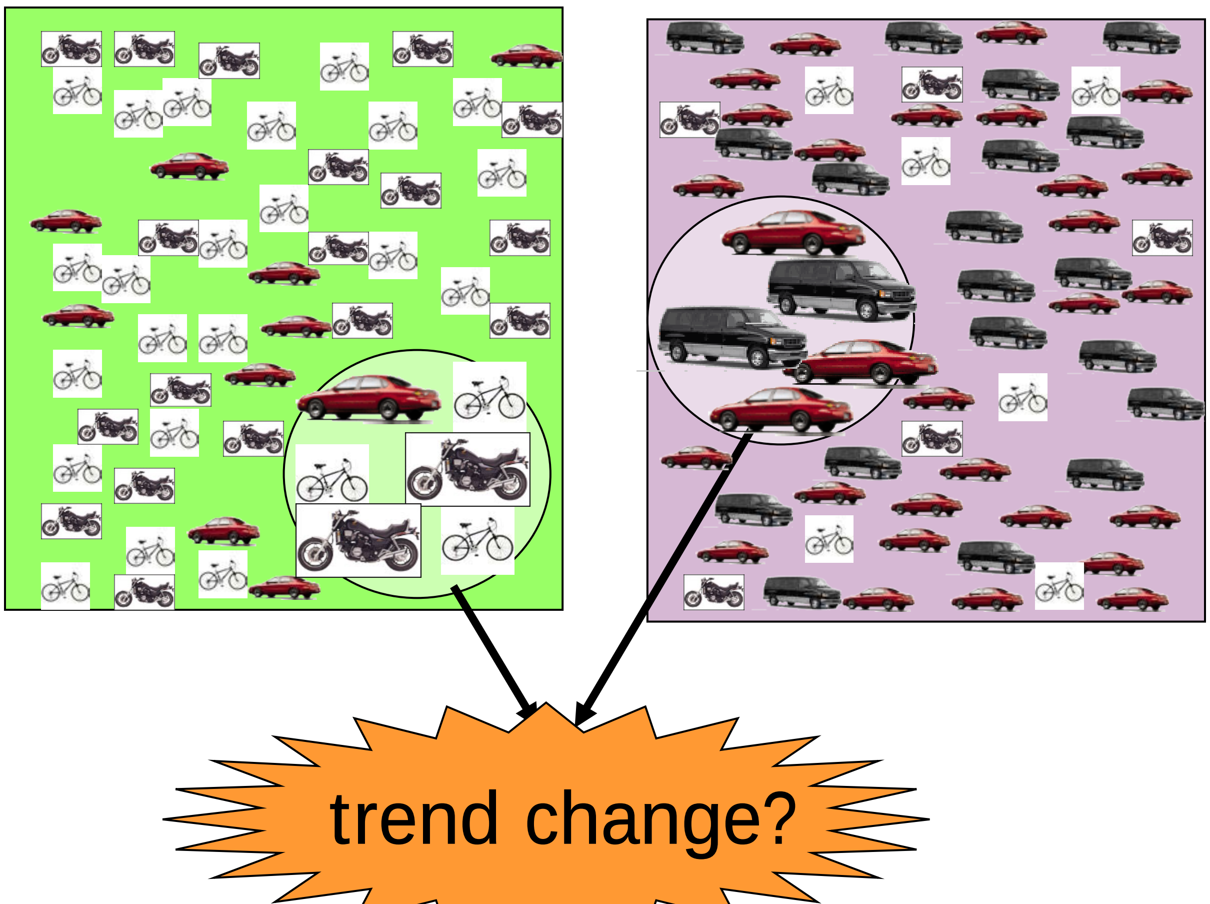
Algorithms with sublinear sample complexity

# Some other interesting properties…

# Testing closeness of two distributions:

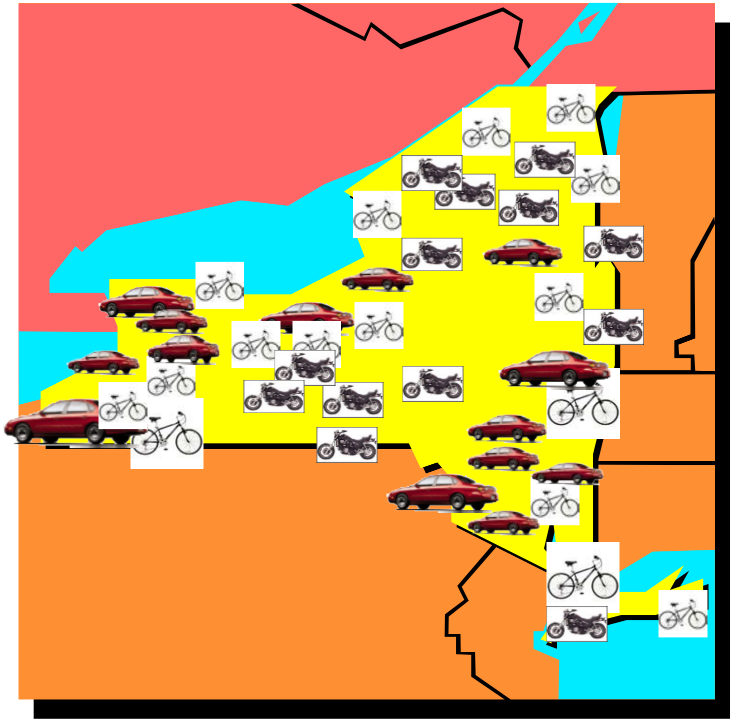
Transactions of 20-30 yr olds

Transactions of 30-40 yr olds



# Testing Independence:

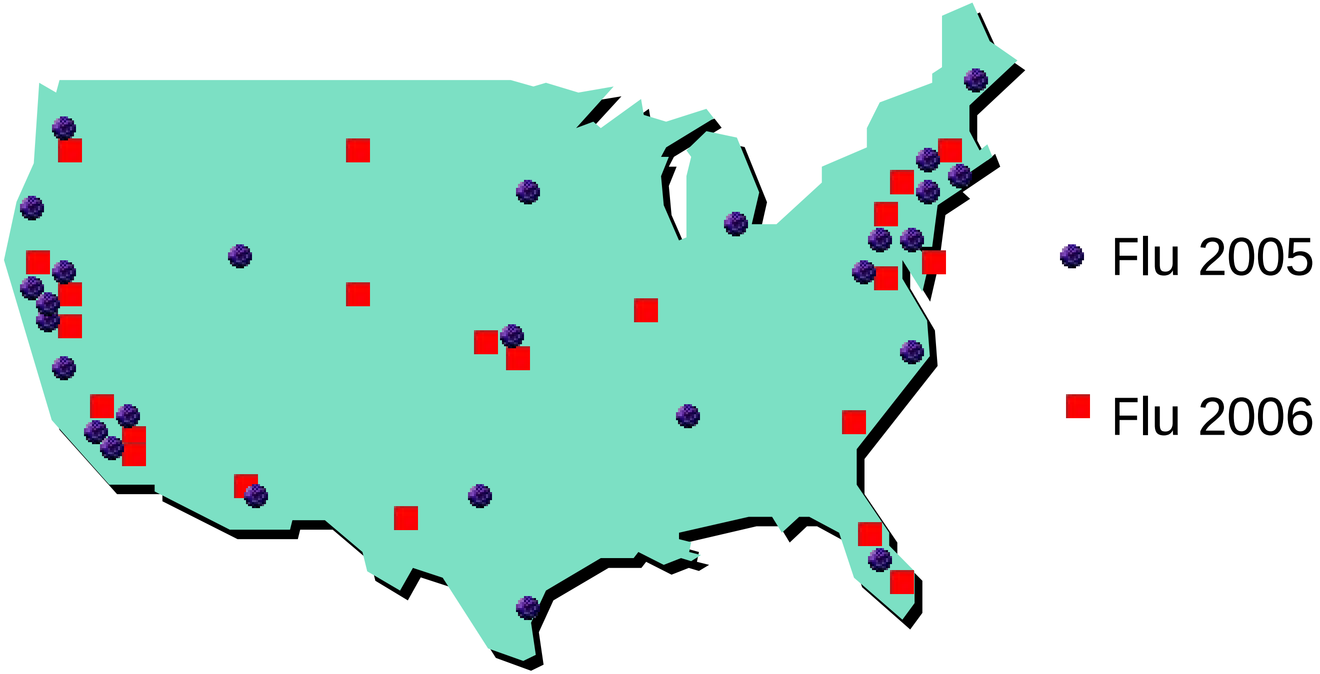
Shopping patterns:



Independent of zip code?

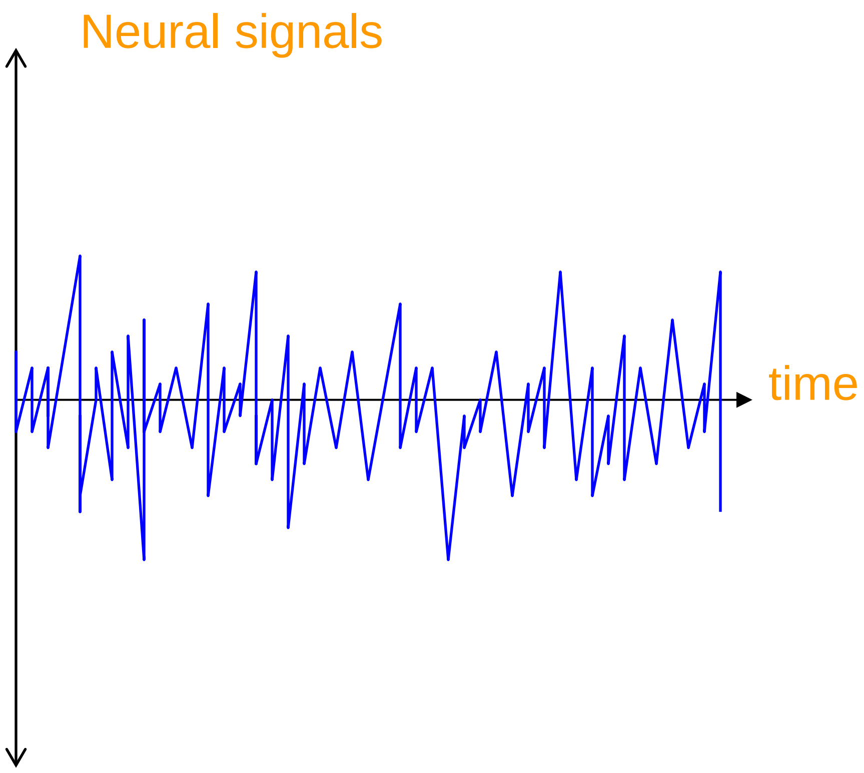
# Outbreak of diseases

- Similar patterns?
- Correlated with income level?
- More prevalent near large airports?



# Information in neural spike trails

[Strong, Koberle, de Ruyter van Steveninck, Bialek  ’98]




- Each application of stimuli gives sample of signal (spike trail)

- Entropy of (discretized) signal indicates which neurons respond to stimuli

# Compressibility of data

# Distribution property testing in algorithm design

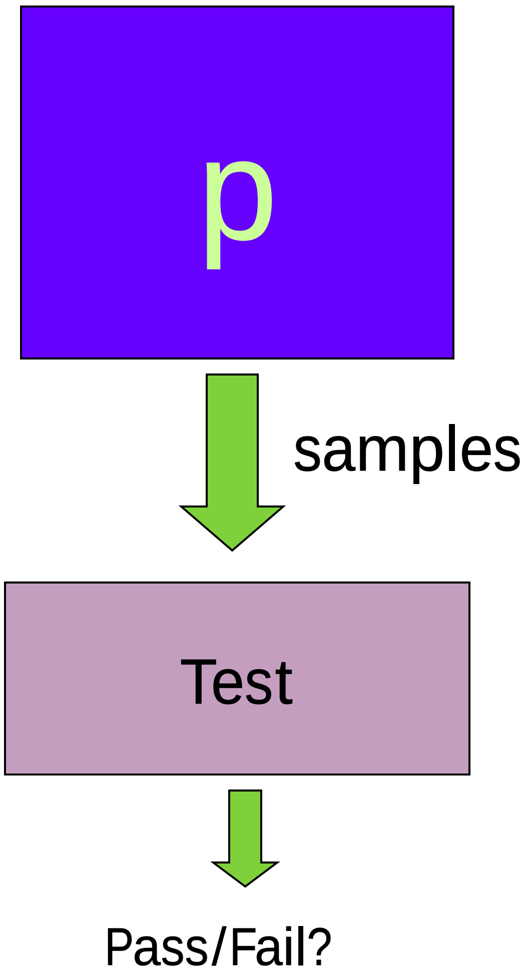
- Testing expansion, rapid mixing and cluster structure

  [Goldreich Ron] [Batu Fortnow Rubinfeld Smith White] [Czumaj Sohler] [Kale Seshadri] [Nachmias Shapira][Czumaj Peng Sohler]

- Testing graph isomorphism

  [Fisher Matsliah] [Onak Sun]

# Our usual model:

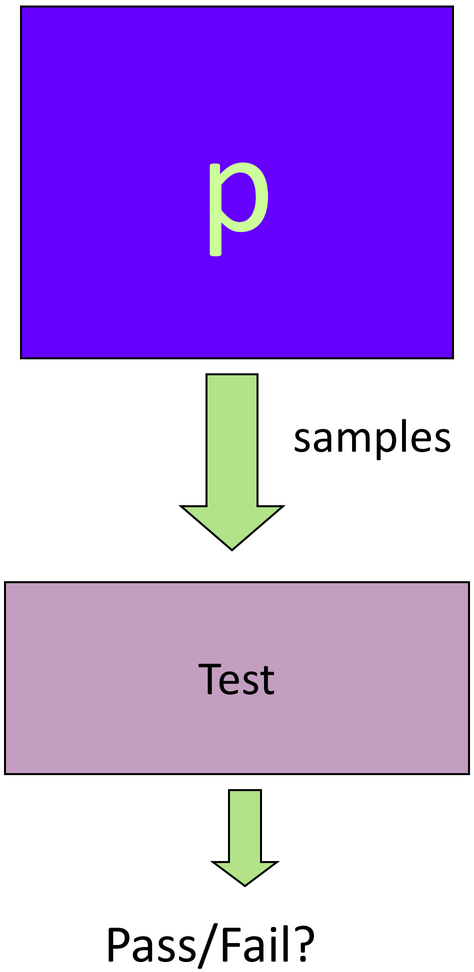



- *p* is arbitrary black-box distribution over *[n]*, generates iid samples.
- $p_i$ = Prob[ *p* outputs *i* ]
- Sample complexity in terms of *n*?

# Similarities of distributions

- Are $p$ and $q$ close or far?
  - $q$ is known to the tester
    - $q$ is uniform
  - $q$ is given via samples

# Is *p* uniform?




- Theorem: ([Goldreich Ron] [Batu Fortnow R. Smith White] [Paninski]) Sample complexity of distinguishing

$$p = U$$

from $||p - U||_1 > \varepsilon$ is $\theta(n^{1/2})$

$$||p - U||_1 = \Sigma \left| p_i - \frac{1}{n} \right|$$

# Upper bound for $L_2$ distance [Goldreich Ron]

- $L_2$ distance: $\left|\left|p - q\right|\right|_2^2 = \sum(p_i - q_i)^2$

- $||p\text{-}U||_2^2 = \Sigma(p_i - 1/n)^2$
  $= \Sigma p_i^2 - 2\Sigma p_i/n + \Sigma 1/n^2$
  $= \Sigma p_i^2 - 1/n$

- Estimate collision probability to estimate $L_2$ distance from uniform

# Testing uniformity [GR][BFRSW]

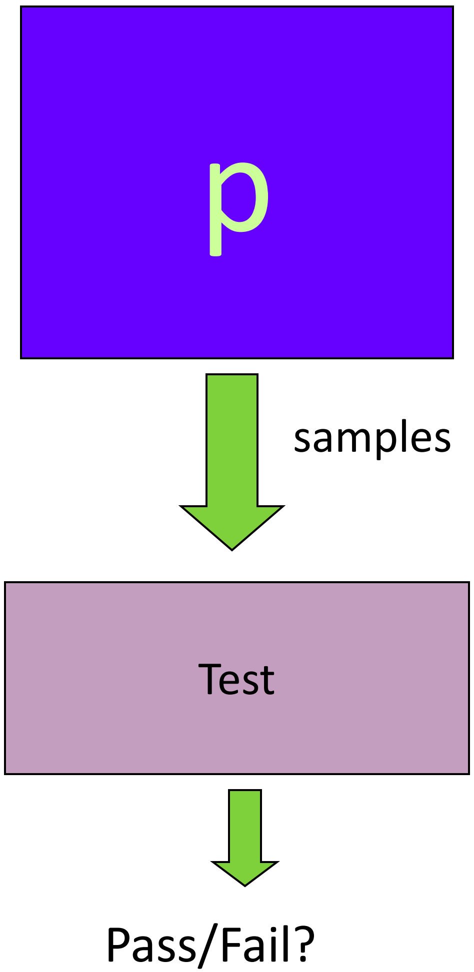
- Upper bound: Estimate collision probability + bound $L_\infty$ norm
  - Issues:
    - Collision probability of uniform is 1/n
    - Pairs not independent
    - Relation between $L_1$ and $L_2$ norms
  - Comment: [P] uses different estimator

- Easy lower bound: $\Omega(n^{1/2})$
  - Can get $\Omega(n^{1/2}/\varepsilon^2)$ [P]

# Back to the lottery…

plenty of samples!

# Is *p* uniform?




- Theorem: ([Goldreich Ron][Batu Fortnow R. Smith White] [Paninski]) Sample complexity of distinguishing

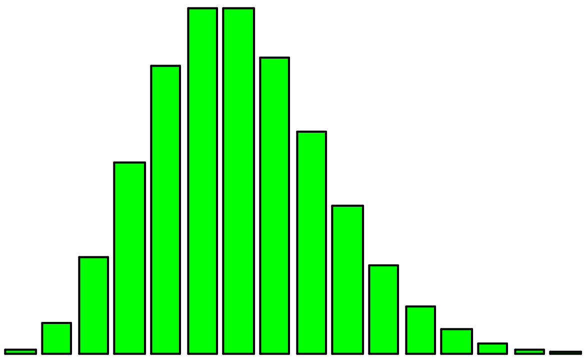
  $p=U$

  from $|p-U|_1>\varepsilon$ is $\theta(n^{1/2})$

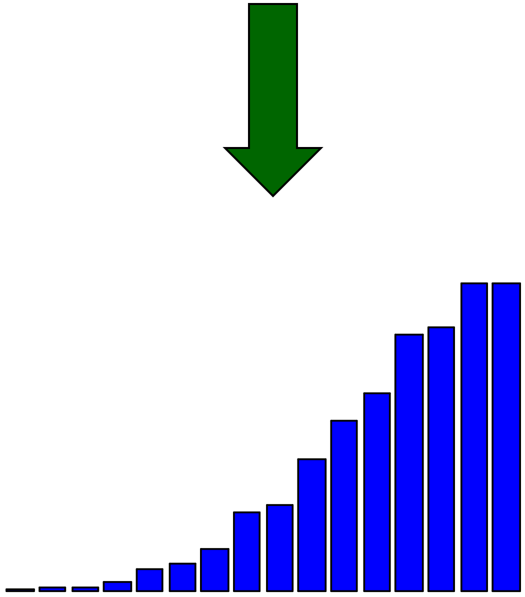
- Nearly same complexity to test if p is any *known* distribution [Batu Fischer Fortnow Kumar R. White][Onak]: "Testing identity"

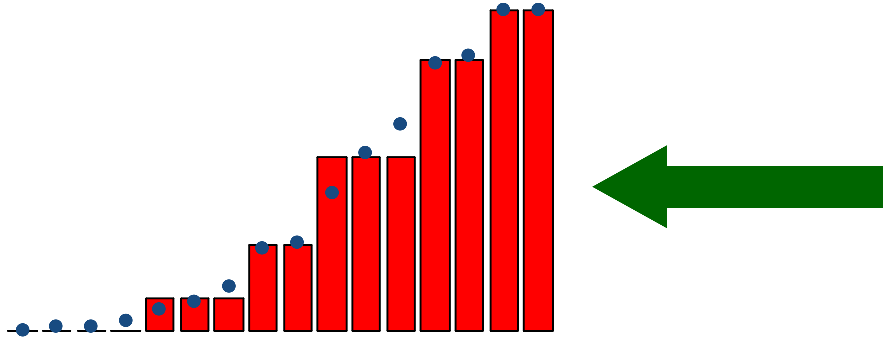
# Testing identity via testing uniformity on subdomains:

q (known)



- *(Relabel domain so that q monotone)*
- Partition domain into *O(log n)* groups, so that each partition almost "flat" --
  - differ by <(1+ε) multiplicative factor
  - *q* close to uniform over each partition
- Test:
  – Test that *p* close to uniform over each partition
  – Test that *p* assigns approximately correct total weights to each partition

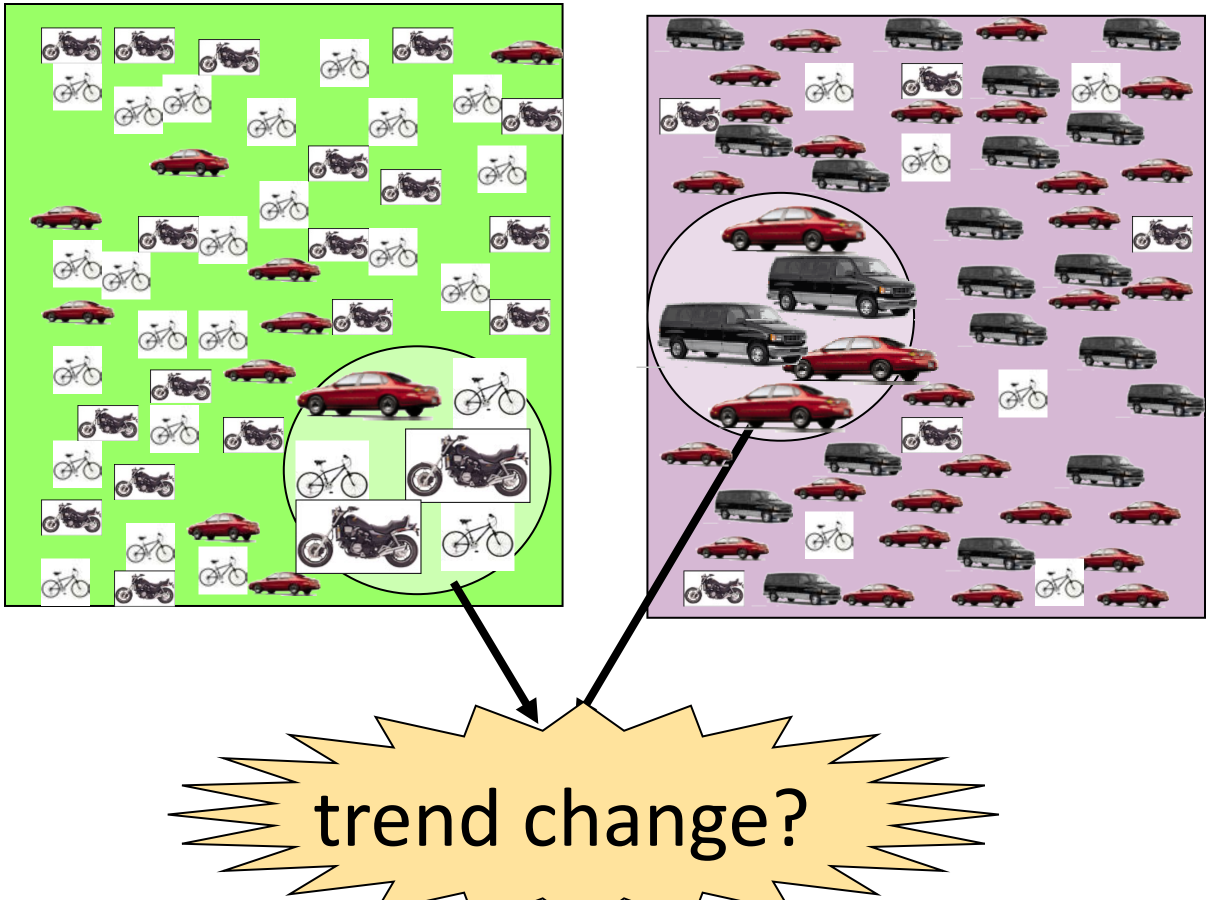




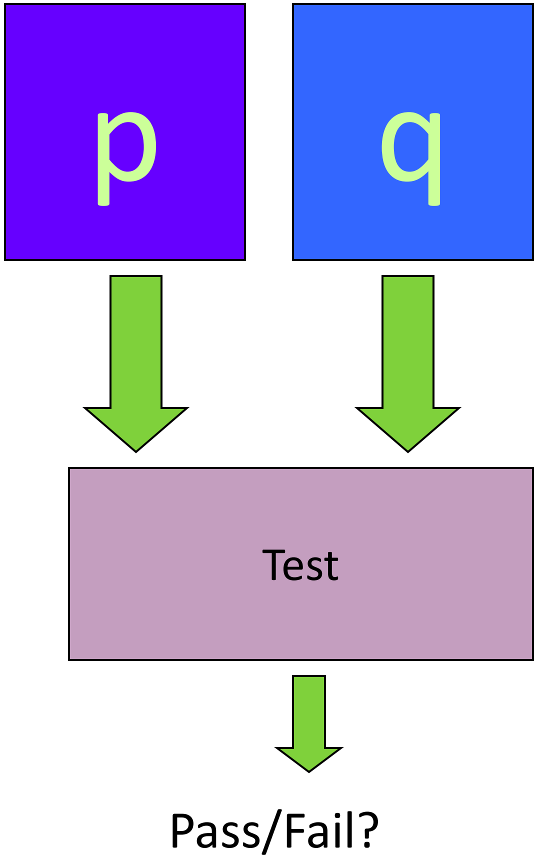
# Testing closeness of two distributions:

Transactions of 20-30 yr olds

Transactions of 30-40 yr olds



# Testing closeness




Theorem: ([BFRSW] [P. Valiant] [Chan Diakonikolas Valiant Valiant])

Sample complexity of distinguishing

$p=q$

from $||p-q||_1 > \varepsilon$

is $\theta(n^{2/3})$

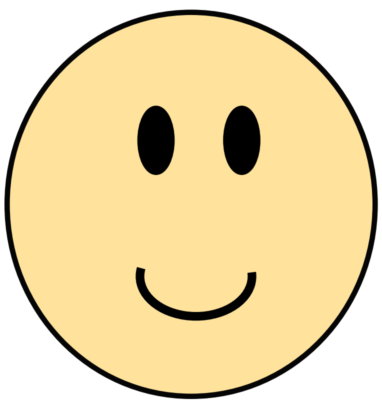
# Why so different?

- Collision statistics are all that matter
- Collisions on “heavy” elements can hide collision statistics of rest of the domain
- Construct pairs of distributions where heavy elements are identical, but “light” elements are either identical or very different

# Approximating the distance between two distributions?

Distinguishing whether $|p\text{-}q|_1 < \varepsilon$ or $|p\text{ -}q|_1$ is $\Theta(1)$ requires $\theta(\frac{n}{\log n})$ samples

[V08, G. Valiant P. Valiant 11]



or

?

# Collisions tell all

- Algorithms:
  - Use collisions to determine “wrong” behavior
- Lower bounds:
  - For symmetric properties, collision statistics are only relevant information
  - Need new analytical tools since not independent

# What about joint properties of many distributions?

# Some questions (and answers):

- Are they all equal?
- Can they be clustered into k groups of similar distributions?
- Do they all have the same mean?

See [Levi Ron R. 2011, Levi Ron R. 2012]

# More properties:

- Independence and limited Independence: [Batu Fischer Fortnow Kumar R. White] [Levi Ron R.][Alon Andoni Kaufman Matulef R. Xie] [Haviv Langberg]
- Entropy, support size and other information theoretic quantities [Guha McGregor Venkatasubramanian]
- Monotonicity over general posets [Batu Kumar R.] [Bhattacharyya Fischer R. P. Valiant]
- *K*-histogram distributions [Levi Indyk R.]
- *K*-modal distributions [Daskalakis Diakonikolas Servedio]
- Poisson Binomial Distributions [Daskalakis Diakonikolas Servedio]

And lots more!

# Many other properties to consider!

- Higher dimensional flat distributions
- Mixtures of *k* Gaussians
- “Junta”-distributions
- Generated by a small Markovian process
- …

# Dependence on n

- o(n)
- But usually $n^{\alpha}$ for some $0 < \alpha < 1$

Is this good or bad?

nontrivial

but still daunting!

# Getting past the lower bounds

- Restricted classes of distributions
  - Structured distributions
  - Competitive closeness testing -- compare to best symmetric
- Other distance measures
- More powerful query models

# Special distributions (spoiler alert)

- Can we take advantage of special structure of the distribution?
  - E.g., monotone/*k*-modal distributions, Poisson Binomials, Sums of independent integer random variables, … BEAUTIFUL STRUCTURAL THEOREMS!
  - A general way to compete with the optimal?
  - See later talks TODAY!

# Other distance measures:

- L2 distance
- Information theoretic distances [Guha McGregor Venkatasubramanian]
- Earth Mover Distance [Doba Nguyen$^2$ R.]

# More power to the tester!

# What kind of queries?

- Samples of distribution
- Queries to probability density function (pdf-queries): “What is p(i)?”
- Queries to cumulative distribution function (cdf-queries): “What is p([1..x])?” [Canonne R.]
- Samples of conditional distribution [Chakraborty Fischer Goldhirsh Matsliah] [Canonne Ron Servedio]
  - Which conditioning predicates?
    - Arbitrary subsets, ranges, pairs of domain elements...

# Example 1:

Distribution comes from a file that has already been sorted

1,1,1,1,2,4,4,10,11,13,13,13,13,13,15,99,99,253,666,666,…

- Samples in $O(1)$ time
- pdf queries in $O(\log n)$ time
- cdf queries in $O(\log n)$ time

# Example 2:

Google *n*-gram data

- Frequencies (Pdf) for each sequence of n words
- Samples of sequences

# Example 3:

Database provides extra support

- E.g. Needletail [Kim Madden Parameswaran]
  - Samples
  - Conditional samples for simple predicates
    - i.e. random entry $x$ s.t. $x_i = r$

# Can it help to have pdf queries (rather than samples)?

# YES!

$$\frac{2}{n}, 0, 0, 0, 0, \frac{2}{n}, \frac{2}{n}, \frac{2}{n}, 0, 0, \frac{2}{n}, 0, \frac{2}{n}, 0, \frac{2}{n}, \frac{2}{n}, 0, 0, 0, \frac{2}{n}, \frac{2}{n}, 0, 0, 0, \frac{2}{n}, \frac{2}{n}, \frac{2}{n}, \frac{2}{n}, 0$$

Testing uniformity?

Samples only: need $\sqrt{n}$

Given pdf queries: O($1/\epsilon$)

# Are probability distribution function (pdf) queries better than samples?

# No!

0 0 0 0 0 0 0 0 0 0 0 0 ¼ 0 0 ¼ 0 0 0 0 0 0 ¼ 0 0 ¼ 0 0 0 0 0 0

What is entropy?

Given samples:  estimate quickly

Given pdf queries:  look for needles in haystack

# Can we multiplicatively approximate entropy from samples?

- In general, no!
  - $\approx 0$ entropy distributions are hard to distinguish with any number of samples
- entropy big enough:
  - $\gamma$-multiplicatively approximate the entropy with $\theta(n^{1/\gamma^2})$ samples (if entropy > $\Omega(\gamma)$) [Batu Dasgupta R. Kumar] [Valiant]
  - better bounds in terms of support size [Brautbar Samorodnitsky]

# Can we multiplicatively approximate entropy from other queries?

- From pdf queries (only):

  $\Omega(n)$ for any approximation
- From pdf queries + samples:

  $\theta(\log n)$

[BDKR][Guha McGregor Venkatasubramanian]

# What about additive estimates of entropy?

- Samples only: $\theta(n/\log n)$ [Valiant Valiant]
- Samples + cdf, Samples+ pdf: $polylog(n)$ [Canonne R]
  - Sample to estimate $E[\log(\frac{1}{p(x)})]$

# Closeness of distributions

$O(\frac{1}{\epsilon})$ samples suffice for testing closeness

# Relative power of different oracles?

# Samples + pdf vs. cdf queries [Canonne R.]

- Cdf is pretty powerful:
  - Given samples + cdf, can simulate samples + pdf in 2 queries
  - Given cdf, can simulate samples in O(log n) queries
- What about other direction?
  - Some evidence that cdf queries are more powerful…

# Samples and pdf/cdf vs. conditional samples?

More efficient closeness testing/distance approximation algorithms in samples+pdf/cdf models

But no separation results!

# $(1/\log n)$- close to monotone distributions

- Two classes of close-to-monotone distributions:
  - One “heavy” element – probability 1-1/log n
  - Entropy difference comes from VERY FEW or JUST PLAIN FEW other “light” elements

- Estimating entropy difference requires $\Omega(\log n)$ sample+pdf queries
- Needs only $O(\log^2 \log n)$ sample+cdf queries!

# Questions for the oracles

- Comparison of powers of different oracle models?
- Approximate queries?
- Improvements to other learning/testing problems in these models?
- What queries should we convince DB systems to implement?

# More open directions

Other properties?

Non-iid samples?

# Conclusion:

- Distribution testing problems are everywhere
- For many problems, we need a lot fewer samples than one might think!
- Many COOL ideas and techniques have been developed
- Lots more to do!

# Thank you