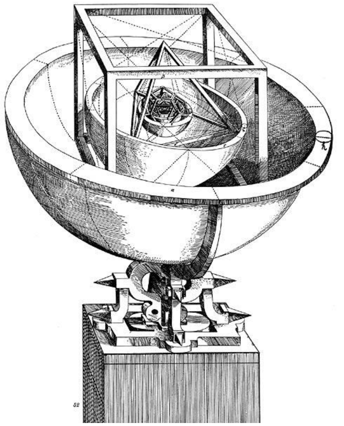
Two Decades of Property Testing

Madhu Sudan
Microsoft Research

Kepler's Big Data Problem



Tycho Brahe (~1550-1600):

- Wished to measure planetary motion accurately.
- To confirm sun revolved around earth … (+ other planets around sun)
- Spent 10% of Danish GNP

Johannes Kepler (~1575-1625s):

- Believed Copernicus's picture: planets in circular orbits.
- Addendum: Ratio of orbits based on Löwner-John ratios of platonic solids.
- "Stole" Brahe's data (1601). Source: Michael Fowler, "Galileo & Einstein", U. Virginia
- Worked on it for nine years.
- Disproved Addendum; Confirmed Copernicus (circle -> ellipse); discovered laws of planetary motion.

- **Nine Years?**

The challenge of analyzing big data

- Standard method:
  - Propose concept class.
  - LEARN (parameters of) best fitting concept in class to data in hand.
  - TEST to see if this is a good enough fit.
- Bottleneck
  - LEARNing is expensive; wasted if TEST rejects.
  - Can we TEST before we LEARN?

Don’t be
Ridiculous!

- Yes: This is PROPERTY TESTING!!

Property Testing

- Sublinear time algorithms:
  - Algorithms running in time o(input), o(output).
    - Probabilistic.
    - Correct on (approximation to) input.
    - Random access to input, output implicit.
- Property testing:
  - Restriction of sublinear time algorithms to decision problems (output = YES/NO).
  - What decision problem?
    - $\exists$ concept within class that fits data?

      $\Leftrightarrow$ Does data have Property?
- Amazing fact: Many non-trivial algorithms exist!

Example 1: Polling

- Is the majority of the population Red/Blue
  - $C = \cup_{\alpha > .5} C_\alpha$; $C_\alpha = \{$populations with α fraction Red$\}$
  - Can Test for $\alpha \geq .5$ by random sampling.
    - Accept w.h.p. if $\alpha \geq .5$
    - Reject w.h.p. if $\alpha < .5 - \epsilon$
  - Sample size $\propto$ $\Theta\left(\frac{1}{\epsilon^2}\right)$
    - Independent of size of population
- Other similar examples: (basic statistical parameters; averages, quantiles, variance …)

Example 2: Linearity

- Can test for homomorphisms:
  - Given: $f: G \to H$ (G, H finite groups), is f essentially a homomorphism?
  - Test:
    - Pick $x, y \in G$ uniformly, ind. at random;
    - Verify $f(x) \cdot f(y) = f(x \cdot y)$
  - Completeness: accepts homomorphisms w.p. 1
    - (Obvious)
  - Soundness: Rejects f w.p prob. Proportional to its "distance" (margin) from homomorphisms.
    - (Not obvious, [BlumLubyRubinfeld'90])

Linearity Analysis

- Given $f: G \to H$ that usually passes test, "pretend" it is close to a homomorphism $g: G \to H$.
  - Locally decode g
    - $\forall x,\ g(x) \triangleq f(x.r) \cdot f(r)^{-1}$ for random $r \in G$
    - Prove:
      1. g is close to f. (Easy)
      2. g is a homomorphism. (Challenging)
         - Why should $f(x.r) \cdot f(r)^{-1} = f(x.s) \cdot f(s)^{-1}$?
           - [Requires some algebraic reasoning.]

- Note: New elements in analysis!

A subtle change

- Compare:
  - f usually satisfies $f(x.y) = f(x) \cdot f(y)$.
  - Population has close to 50% Reds.

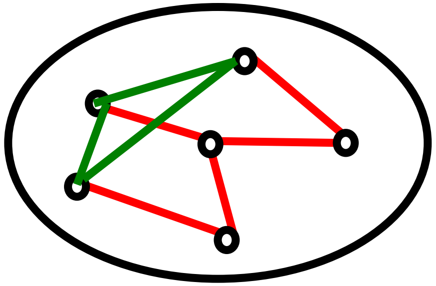
  Vs.

  - f is close to g that always satisfies $g(x.y) = f(x) \cdot g(y)$
  - Population is close to one with exactly 50% Reds.
- Notions same for Polling; not Homomorphisms.
- Latter notion is generalizable to any property!
- Notion of choice in Property Testing

History (slightly abbreviated)

- [Blum,Luby,Rubinfeld – S'90]
  - Linearity + application to program testing
- [Babai,Fortnow,Lund – F'90]
  - Multilinearity + application to PCPs (MIP).
- [Rubinfeld+S.]
  - Low-degree testing + Definition
- [Goldreich,Goldwasser,Ron]
  - Graph property testing + systematic study
- Since then … many developments
  - More graph properties, statistical properties, matrix properties, properties of Boolean functions …
  - More algebraic properties

Example 3: Δ-free-ness



- Given graph G, is it free of triangles?

- Test:
  - Pick vertices u, v, w at random.
  - Accept if u, v, w don't form a triangle

- Analysis: [Alon-Shapira]
  - Use Szemeredi's regularity lemma.
    - Can partition any graph into $O_\epsilon(1)$ parts.
    - Between each part edges "random".
    - If some three well-connected partitions form triangle; then many triangles, else close to triangle-free

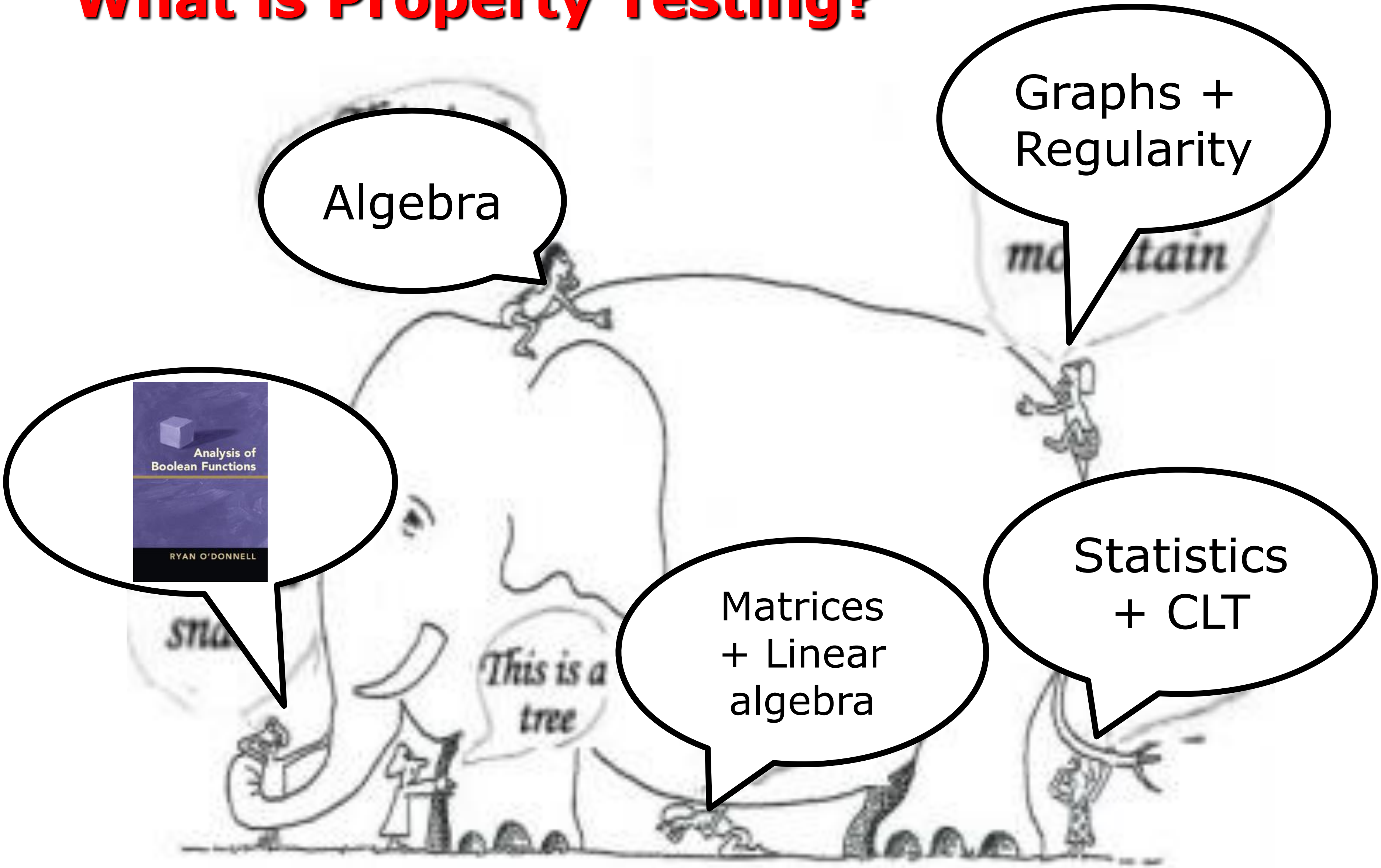
Example 4: Long code/Junta testing

- Given $f: \{0,1\}^n \rightarrow \{0,1\}$ does it depend on few coordinates. [BGS, Håstad, FKRSS… Blais]
  - Motivation: data = genome; f represents some disease;
  - Junta-testing: Disease caused by few features?
  - Testing before learning?
- Fuzzy Test: [KKMO, MOO]
  - Pick $x \sim U(\{0,1\}^n)$ and y ϵ-noisy copy of x.
  - Accept iff $f(x) = f(y)$; Repeat
- Analysis:
  - If f function of very few variables ⇒ Accept w.h.p.
  - If f depends on many variables ⇒ Reject w.p. $\frac{1}{2}$.
  - Techniques: Fourier analysis, Influence of variables, hypercontractivity …

Example 5: Distribution Testing

- Given samples from unknown distribution P on $[n]$
- Determine if $H(P) \geq k$
- [Batu et al.,Valiant,Valiant²]:
  - #samples needed $= \Theta(\frac{n}{\log n})$!
  - Techniques:
    - Multivariate Central Limit Theorem
    - Stein's method
    - Hermite polynomials …

What is Property Testing?



(Dense) Graph Property Testing

- Theorem [AlonFischerNewmanShapira]:
  Graph property P is $O(1)$-query testable
  $\Leftrightarrow$ P is "determined by regularity".
  - Suggested by [Goldreich,Goldwasser,Ron]
  - In particular implies all hereditary properties are testable [Alon Shapira]

- Nice characterization of testability?
  - Uniform test for all graph properties.
  - Single unifying analysis e.g. Δ-freeness & 3-colorability

Contrast with Low-degree testing

- Given $f: \mathbb{F}_q^n \to \mathbb{F}_q$; Is $\deg(f) \le d$?
- Roughly, BLR deals with $d = 1$;
- $d \le q/$
  - Tes
  - Mal
  - Ana

Why no unification?

- $d \ge q$ ... Ron'03]
  - Test: $\deg(f|_A) \le d$ for subspace A; $\dim(A) = d + 1$?
  - Analysis a la BLR, RS; many changes
- d, q arbitrary: [KaufmanRon'04] Analysis a la …

Aside: Importance of Low-degree Testing

- Central element in PCPs.
  - Till [Dinur'06] – no proof without (robust) low-degree testing.
  - Since: Best proofs (smallest, tightest parameters etc.) rely (in/)directly on improvements to low-degree tests.
- Connected to Gowers Norms:
  - [Viola-Wigderson'07]: [AKKLR]⇒Hardness Amplification
- Yield Locally Testable Codes
  - Best in high-rate regime.
  - [BarakGopalanHåstadMekaRaghavendraSteurer'12]: [BKSSZ'11]⇒ Small-set expanders.

Some (introspective) questions

- What is qualitatively novel about linearity testing relative to classical statistics?
- Why are the mathematical underpinnings of different themes so different?
- Why is there no analog of "graph property testing" (broad class of properties, totally classified wrt testability) in algebraic world?
  - What is the context for low-degree testing?
- Answer to all: Invariance!

Invariance?

- Property $P \subseteq \{f : D \to R\}$

- Property P invariant under 1-1 $\pi: D \to D$, if

$$f \in P \;\Rightarrow f \circ \pi \in P$$

- Property P invariant under group G if

$$\forall \pi \in G \Rightarrow P \text{ is invariant under } \pi.$$

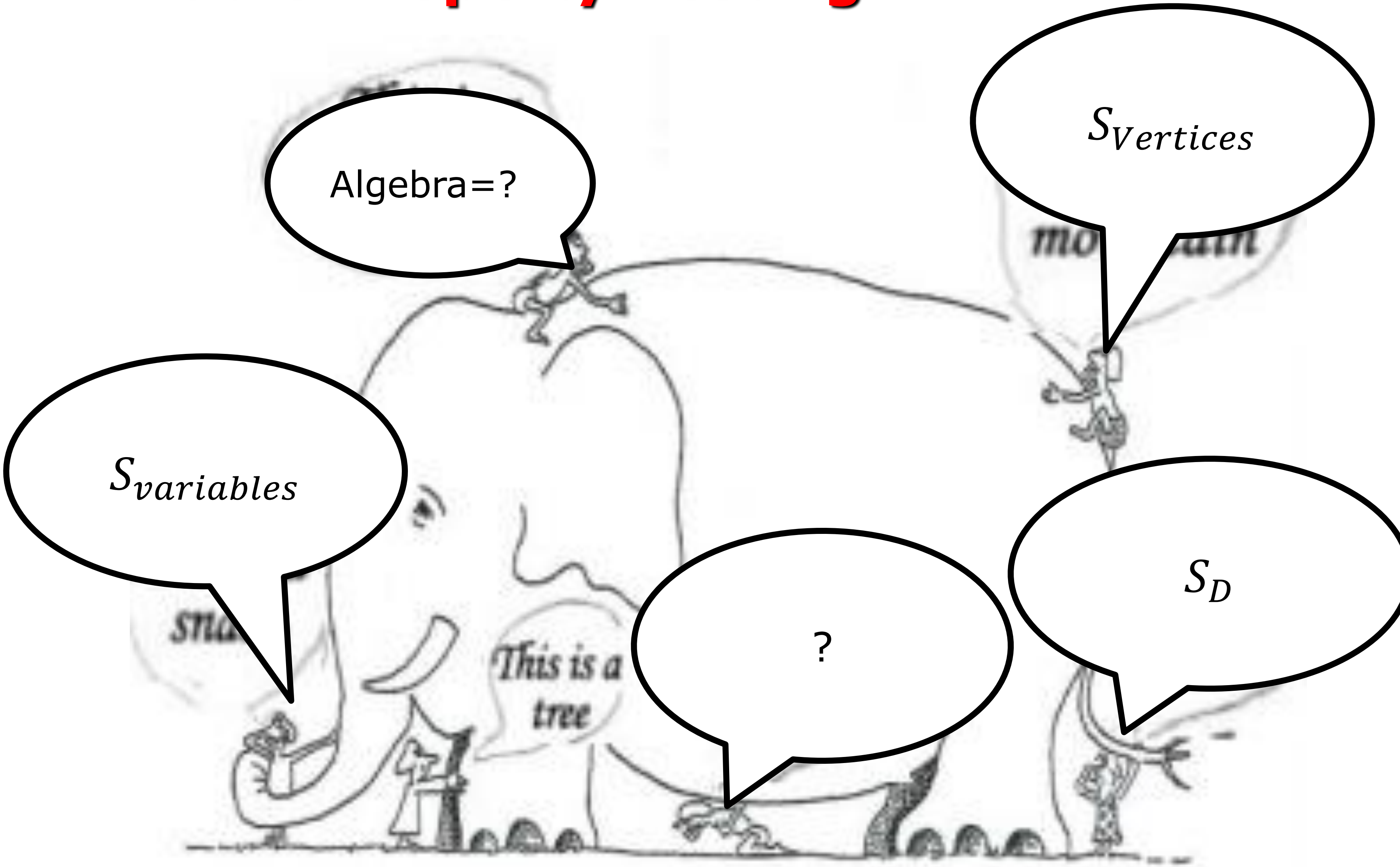
  - (Maximal) G is invariance class of P.

- Main Observation: Different property tests unified/separated by invariance class.

Invariances (contd.)

- Some examples:
  - Classical statistics: Invariant under all permutations.
  - Graph properties: Invariant under vertex renaming.
  - Boolean properties: Invariant under variable renaming.
  - Matrix properties: Invariant under mult. by invertible matrix.
  - Algebraic Properties = ?
- Answers to (introspective) questions.
  - Classical statistics only dealt with S_D
  - Different invariances $\Rightarrow$ different techniques.
  - Context for algebra?

What is Property Testing?



Abstracting algebraic properties

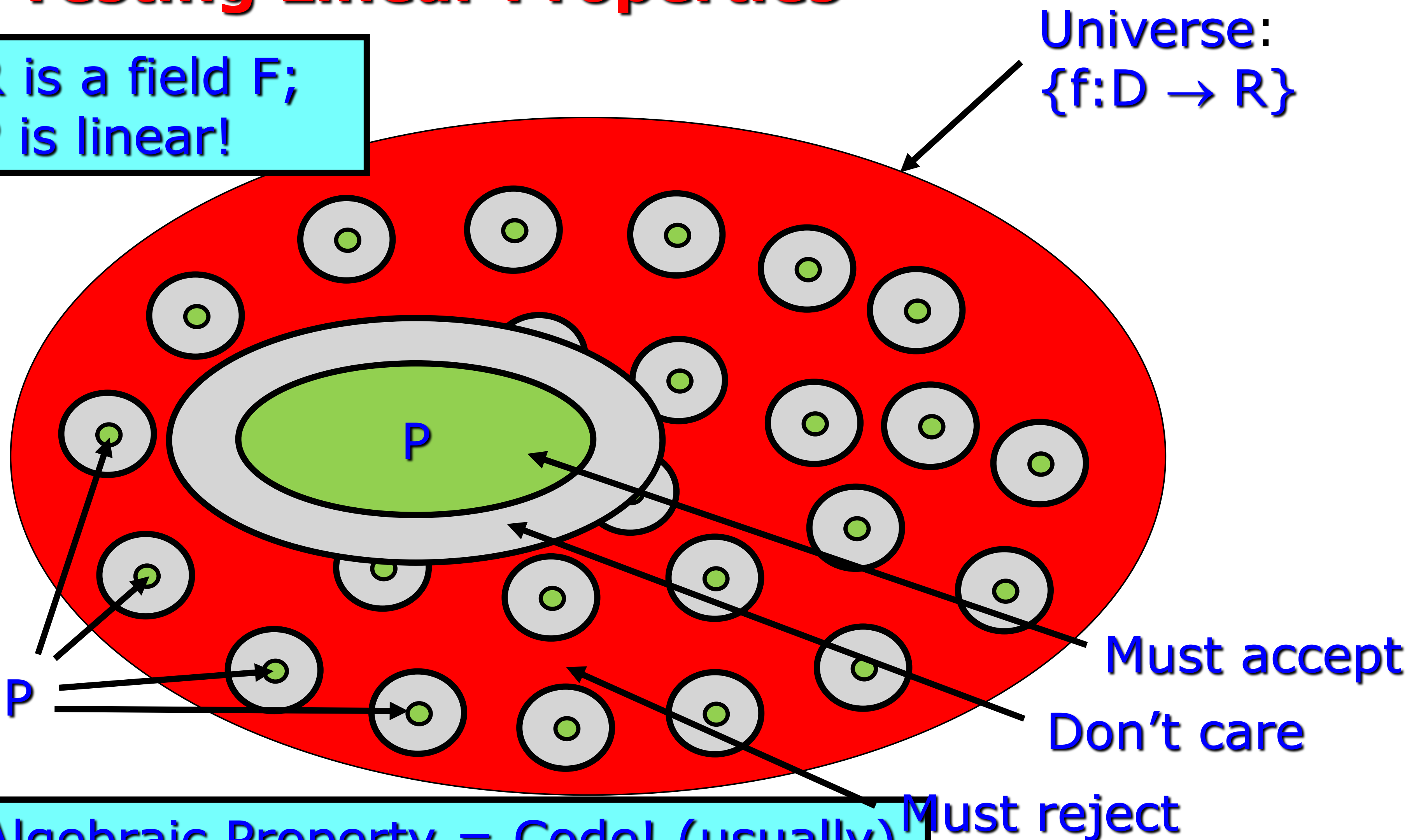
- [Kaufman+S.'08]
- Affine Invariance:
  - Domain = Big field ($\mathbb{F}_{q^n}$)

    or vector space over small field ($\mathbb{F}_q^n$).
  - Property invariant under affine transformations of domain ($x \mapsto A.x + b$)

- Linearity of Properties:
  - Range = small field ($\mathbb{F}_q$)
  - Property = vector space over range.

Testing Linear Properties

R is a field F;
P is linear!



Algebraic Property = Code! (usually)

Why study affine-invariance?

- Common abstraction of properties studied in [BLR], [RS], [ALMSS], [AKKLR], [KR], [KL], [JPRZ].
  - (Variations on low-degree polynomials)

- Hopes
  - Unify existing proofs
  - Classify/characterize testability
  - Find new testable codes (w. novel parameters)

- Rest of the talk: Brief summary of findings

Results 1: AKKLR Conjecture

- P k-locally testable ⇒ P satisfies k-local constraint
- AKKLR Conjecture: k-local constraint + symmetry (2-transitive invariance)⇒ P k'-locally testable.
- Theorem [Kaufman+S.'08]: $P \subseteq \{f: \mathbb{F}_Q^n \to \mathbb{F}_q\}$ has k-local constraint ⇒ $k'(k, Q)$-locally testable.
  - Notion of "single-orbit" ⇒ Unification!
  - Structure of affine-invariant properties.
- Theorem [Grigorescu,Kaufman,S.08]:
  $\exists P \subseteq \{\mathbb{F}_{2^n} \to \mathbb{F}_2\}$ with 8-local constraint
  that is not $\log\log\log n$-LDPC.
- Thm[BMSS'11]: ∃ $O(1)$-LDPC that is not $O(1)$-LTC.

Results 2: Accidental +ve

- [Bhattacharyya,Kopparty,Schoenebeck,S.,Zuckerman'10]:
  - Goal: Test low-degree polynomials over $\mathbb{F}_2$.
  - Hope: Use known better tests from the 90s.
  - Result: New technique + stronger result:
    - [AKKLR] natural test rejects $\Omega(1)$-far f'ns w.p. $\Omega(2^{-d})$.
    - Ours: same test rejects $\Omega(2^{-d})$-far w.p. $\Omega(1)$.
- [Ron-Zewi,S'12]: Better query complexity for low-degree testing, when $d > \frac{q}{2}; q = 2^t$.
  - When $d < q/2$; q-queries suffice.
  - When $\frac{q}{2} < d < q$; known tests made q^2-queries.
  - Our result: $O(q)$-queries suffice.
  - Techniques: single-orbit, structure of affine-invariance…
- Non-linear affine-invariant properties …

Results 3: Lifting

- An annoying way to construct locally constrained properties:
  - Define base property $B \subseteq \{f: \mathbb{F}_q \to \mathbb{F}_q\}$.
  - n-Lifted property $= \{f: \mathbb{F}_q^n \to \mathbb{F}_q \mid f|_{line} \in B\ \forall line\}$

Bad News + Bad News = Good News!

  - But ... result after complicated usage and analysis.
- [Friedl,S'95]:If $\frac{q}{2} < d < q, \exists f: \mathbb{F}_q^n \to \mathbb{F}_q; \deg(f) > d$; such that on every $line$, $\deg(f|_{line}) \leq d$.
  - (Reason for "accidental result 2" on last slide.)

Result 3: Lifting (contd.)

- [Guo,Kopparty,S.'13] Take any base property and lift it:
  - Inherits rel. distance of base property.
  - Testable by [Kaufman+S.'08].
  - Rate = ?; Needs adhoc analysis.
- Base property = deg. d poly with $\frac{q}{2} < d < q$:
  - Code has much higher rate
    - Rate $\rightarrow 1$ for constant dimensional lifts, as $\frac{d}{q} \rightarrow 1$.
    - Gives only known codes of rate $\rightarrow 1$ that are simultaneously sub-linearly locally testable and decodable.

Conclusions

- Returning to bigger picture:
  - Invariance explains the diversity in property testing.
  - Different invariance classes ⇒ different techniques.
  - Same invariance class ⇒ same techniques?
- Need to investigate:
  - Properties of real-valued functions!
  - Properties invariant (only) under variable renaming (a la junta-testing).
  - Invariances of "inference problems"?
  - Queries vs. samples?

Thank You