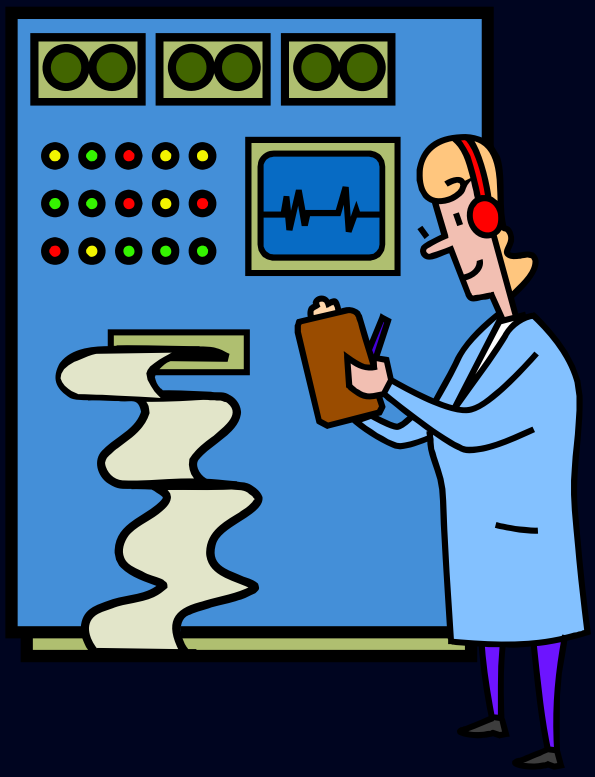
Algebraic Property Testing

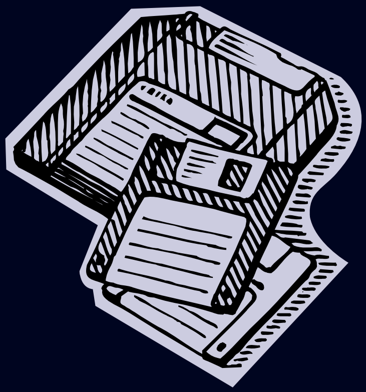
Madhu Sudan
MIT CSAIL

Joint work with Tali Kaufman (IAS).

Classical Data Processing



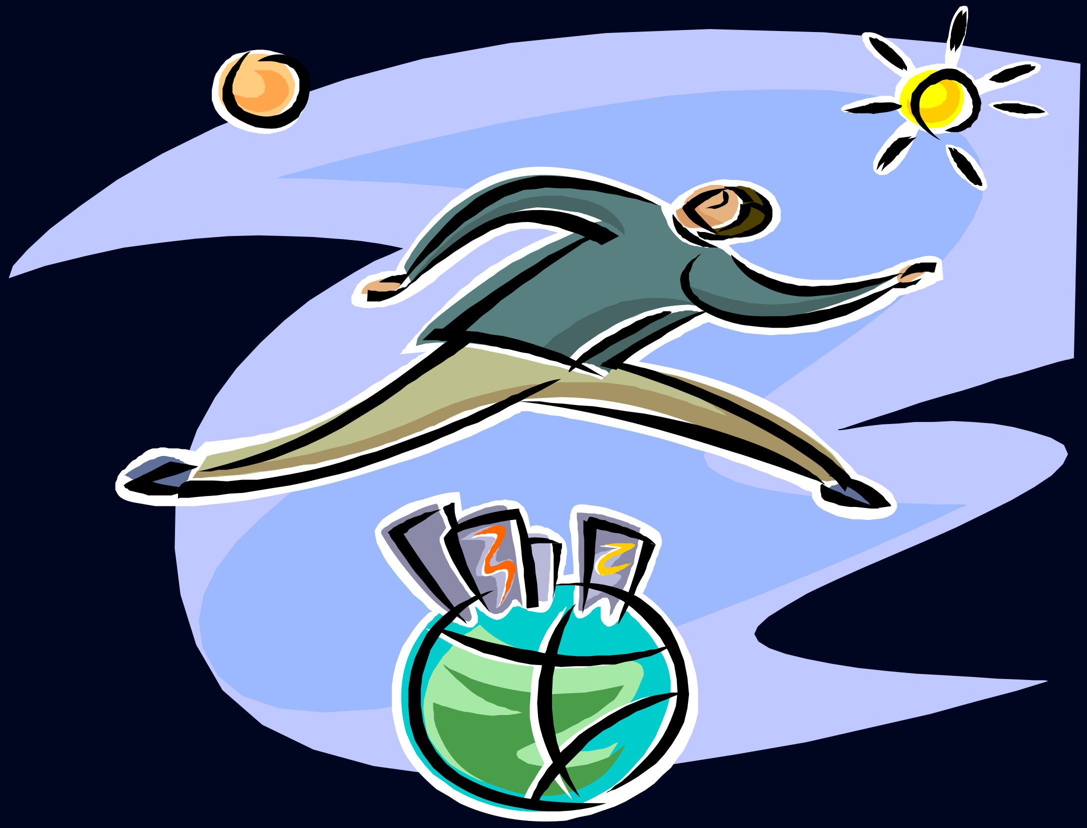
Big computers



Small Data

Modern Data Processing

Small computers



Enormous Data

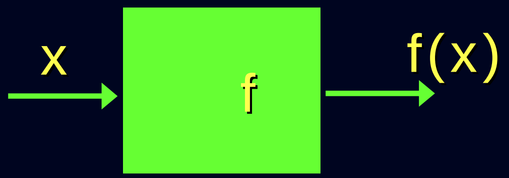
Needs new algorithmic paradigm

- Imbalance not a question of technology :
  - I.e., not because computing speeds are growing less fast than memory capacity.
- Imbalance is a function of expectations:
  - E.g., Users expect to be able to “analyze” the WWW, using a laptop. But WWW includes millions other such laptops.
- Need: Sublinear time algorithms
  - That “estimate” rather than “compute” some given function.

Property Testing

Data: $f : D \to R$

f given by a sampling box



Property: $\mathcal{F} \subseteq \{f : D \to R\}$

q-query Test: Samples f-box q times.

Accepts if $f \in \mathcal{F}$

Hope: ~~Rejects if $f \notin \mathcal{F}$~~ Impossible with $q \ll |D|$

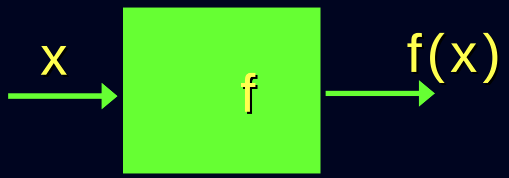
Rejects if f is δ-far from $\mathcal{F}$

f δ-close to $\mathcal{F}$ if $\exists\, g \in \mathcal{F}$ s.t. $\Pr_{x \in D}[f(x) \neq g(x)] \leq \delta$

Property Testing

Data: $f : D \to R$

f given by a sampling box



Property: $\mathcal{F} \subseteq \{f : D \to R\}$

q-query Test: Samples f-box q times.

~~Accepts if $f \in \mathcal{F}$~~

Hope: ~~Rejects if $f \notin \mathcal{F}$~~

Impossible with $q \ll |D|$

Rejects if f is δ-far from $\mathcal{F}$

f δ-close to $\mathcal{F}$ if $\exists\, g \in \mathcal{F}$ s.t. $\Pr_{x \in D}[f(x) \neq g(x)] \leq \delta$

Example: Linearity Testing

- [Blum, Luby, Rubinfeld '90]

$$D = \mathbb{F}_2^n ; R = \mathbb{F}_2$$

- Property = Linearity

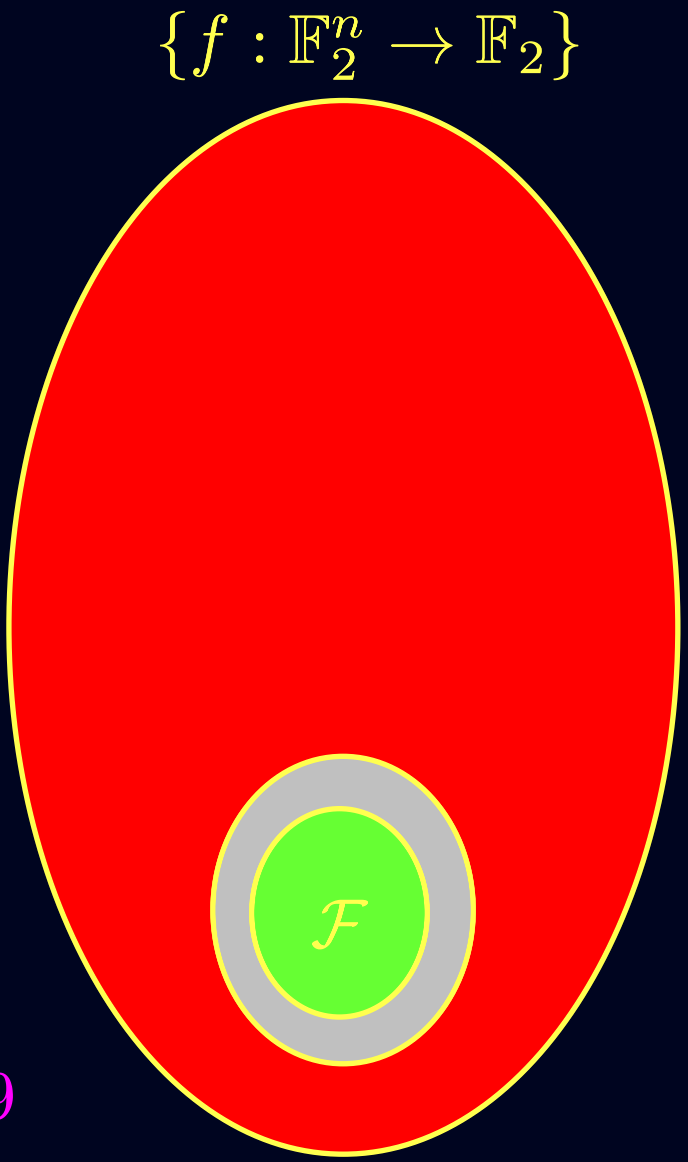
$$\mathcal{F} = \{f | \forall x, y f(x) + f(y) = f(x+y)\}$$

- Test: Pick random x, y

  Accept if $f(x) + f(y) = f(x+y)$

- Non-trivial analysis:

  f δ-far from $\mathcal{F}$ $\Rightarrow$ reject w.p. $2\delta/9$



Example: Linearity Testing

- [Blum, Luby, Rubinfeld '90]

$$D = \mathbb{F}_2^n; R = \mathbb{F}_2$$

- Property = Linearity

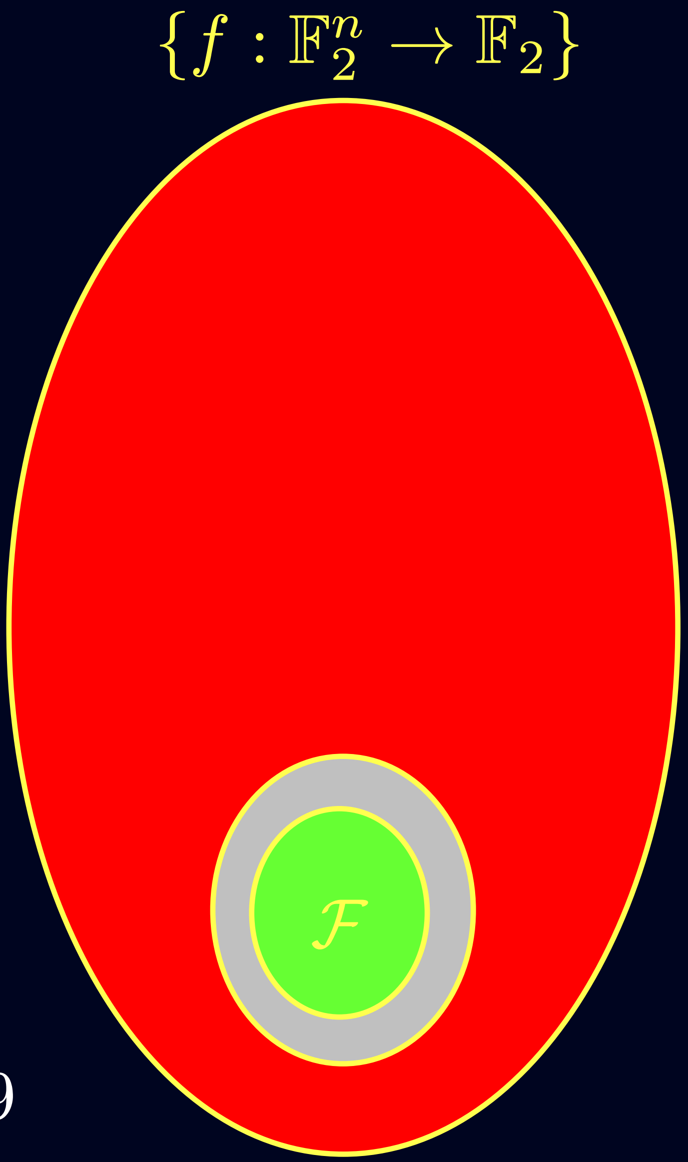
$$\mathcal{F} = \{f | \forall x, y f(x) + f(y) = f(x+y)\}$$

- Test: Pick random x, y

  Accept if $f(x) + f(y) = f(x+y)$

- Non-trivial analysis:

  f δ-far from $\mathcal{F}$ $\Rightarrow$ reject w.p. $2\delta/9$



Property Testing: Abbreviated History

- Prehistoric: Statistical sampling
  - E.g., “Is mean/median at least X”.
- Linearity Testing [BLR’90], Multilinearity Testing [Babai, Fortnow, Lund ’91].
- Graph/Combinatorial Property Testing [Goldreich, Goldwasser, Ron ’94].
  - E.g., Is a graph “close” to being 3-colorable.
- Algebraic Testing [GLRSW,RS,FS,AKKLR,KR,JPSZ]
  - Is multivariate function a polynomial (of bounded degree).
- Graph Testing [Alon-Shapira, AFNS, Borgs et al.]
  - Characterizes graph properties that are testable.

This Talk

- Abstracting Algebraic Property Testing

- Generic Theorem:

  If $\mathcal{F} \subseteq \mathbb{F}^n \to \mathbb{F}$ is closed under addition,

  and under affine tranformations of the coordinates,

  and is locally characterized

  then it is testable.

- Motivations:
  - Generalizes, unifies previous algebraic works
  - Initiates systematic study of testability for algebraic properties
  - Sheds light on testing and invariances of properties.

Property Testing vs. “Statistics”

- Classical Statistics (Mean, Median, Quantiles):
  - Also run in sublinear time.
  - So what’s special about “linearity testing”?
- Classical statistics work on “symmetric” properties:

  $\mathcal{F}$ closed under arbitrary permutation on D.
- Linear functions closed under much smaller group of permutations:

  $\mathcal{F}$ closed under linear maps $L : \mathbb{F}_2^n \to \mathbb{F}_2^n$

  $$|D|^{\log |D|} \text{ such maps vs. } |D|!.$$
- Simlarly, graph properties have nice invariances

Is testing a corollary of invariance?

- Example 1: $D = R; \mathcal{F}_1 = \{Id | Id(x) = x\}$
  - Invariant group trivial. Testing easy.
- Example 2: $D = \{1, \ldots, n\}$; $R = \{1, \ldots, n^9\}$;
  $\mathcal{F}_2 = \{f | x < y \Rightarrow f(x) < f(y)\}$
  - Invariant group still trivial. No O(1) local tests.

- Conclusion: Testing not necessarily a consequence of invariance.
  - If we believe this to be the case for the linearity test, must prove it!!

Is testing a corollary of invariance?

- Example 1: $D = R; \mathcal{F}_1 = \{Id | Id(x) = x\}$
  - Invariant group trivial. Testing easy.
- Example 2: $D = \{1, \ldots, n\}$; $R = \{1, \ldots, n^9\}$;
  $\mathcal{F}_2 = \{f | x < y \Rightarrow f(x) < f(y)\}$
  - Invariant group still trivial. No O(1) local tests.

- Conclusion: Testing not necessarily a consequence of invariance.
  - If we believe this to be the case for the linearity test, must prove it!!

Part II: Formal Definitions & Results

Linear Invariance

$\mathcal{F}$ is Linear Invariant if

- $\mathcal{F}$ is a linear subspace (of $\mathbb{F}^{|\mathbb{F}|^n}$)
- $f \in \mathcal{F}$ and $L : \mathbb{F}^n \to \mathbb{F}^n$ linear $\Rightarrow f \circ L \in \mathcal{F}$

(Affine Invariance defined similarly)

- Examples:
  - Linear functions,
  - n-variate polynomials of degree $\leq d$,
  - homogenous polynomialsof degree d,
  - $\mathcal{F}_1 + \mathcal{F}_2$

Testing, constraints, characterizations

- Suppose $\mathcal{F}$ has a k-query test.
- Then members of $\mathcal{F}$ satisfy a k-local constraint.

Constraint: $C = (x_1, \ldots, x_k \in \mathbb{F}^n$; subspace $V \subseteq \mathbb{F}^k)$

$\forall\, f \in \mathcal{F}, \quad f$ satisfies C i.e., $\langle f(x_1), \ldots, f(x_k) \rangle \in V$

- E.g., in the linear case: $f(\alpha) + f(\beta) = f(\alpha + \beta)$

$$C = (\alpha, \beta, \alpha + \beta; V = \{000, 011, 101, 110\})$$

Characterization: $\mathcal{C} = \{C_1, \ldots, C_m\}$

$f \in \mathcal{F} \quad \Leftrightarrow f$ satisfies $\mathrm{C}_1, \ldots, C_m$.

(Linear-Invariant) Algebraic Characterizations

- Characterizations require many constraints!
- Linear (affine) invariance turns one constraint into many.

$C = (x_1, \ldots, x_k; V)$ constraint and L linear (affine)
$\Rightarrow C \circ L = (L(x_1), \ldots, L(x_k); V)$ is also a constraint.

- (Linearity) Example:

$(L(\alpha), L(\beta), L(\alpha + \beta); V)$ constraint for linearity

- Algebraic Characterization: Single constraint C s.t.

$\{C \circ L | L$ linear (affine) $\}$ characterize $\mathcal{F}$.

Main Theorems

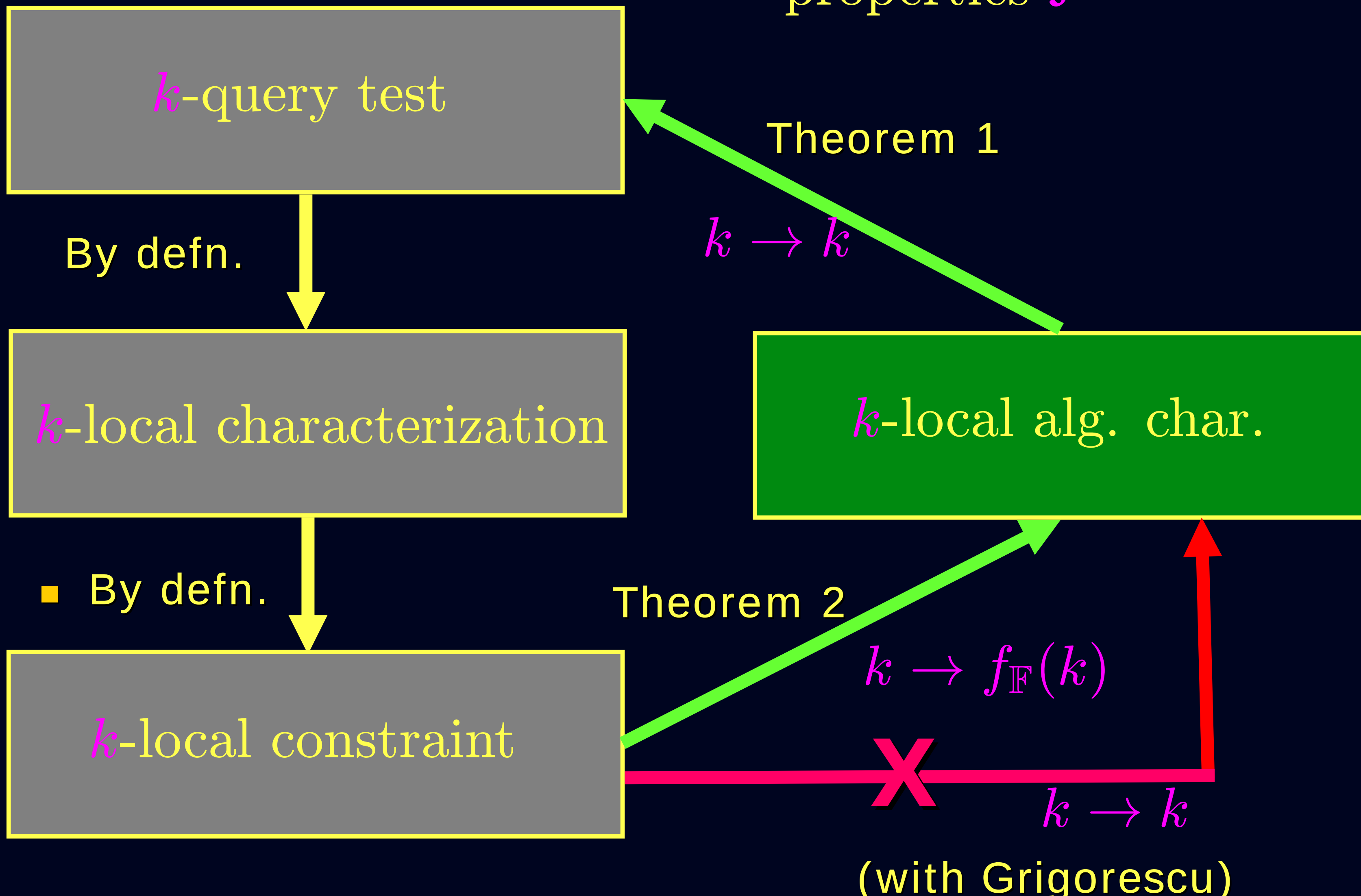
Theorem: $\mathcal{F}$ affine-invariant
and has k-local algebraic characterization,
implies it has a k-query property test.

- Unifies, simplifies, and extends previous algebraic tests

Theorem: $\mathcal{F}$ affine-invariant
and has k-local constraint
$\Rightarrow$ $\mathcal{F}$ has $f_{\mathbb{F}}(k)$-local algebraic characterization.

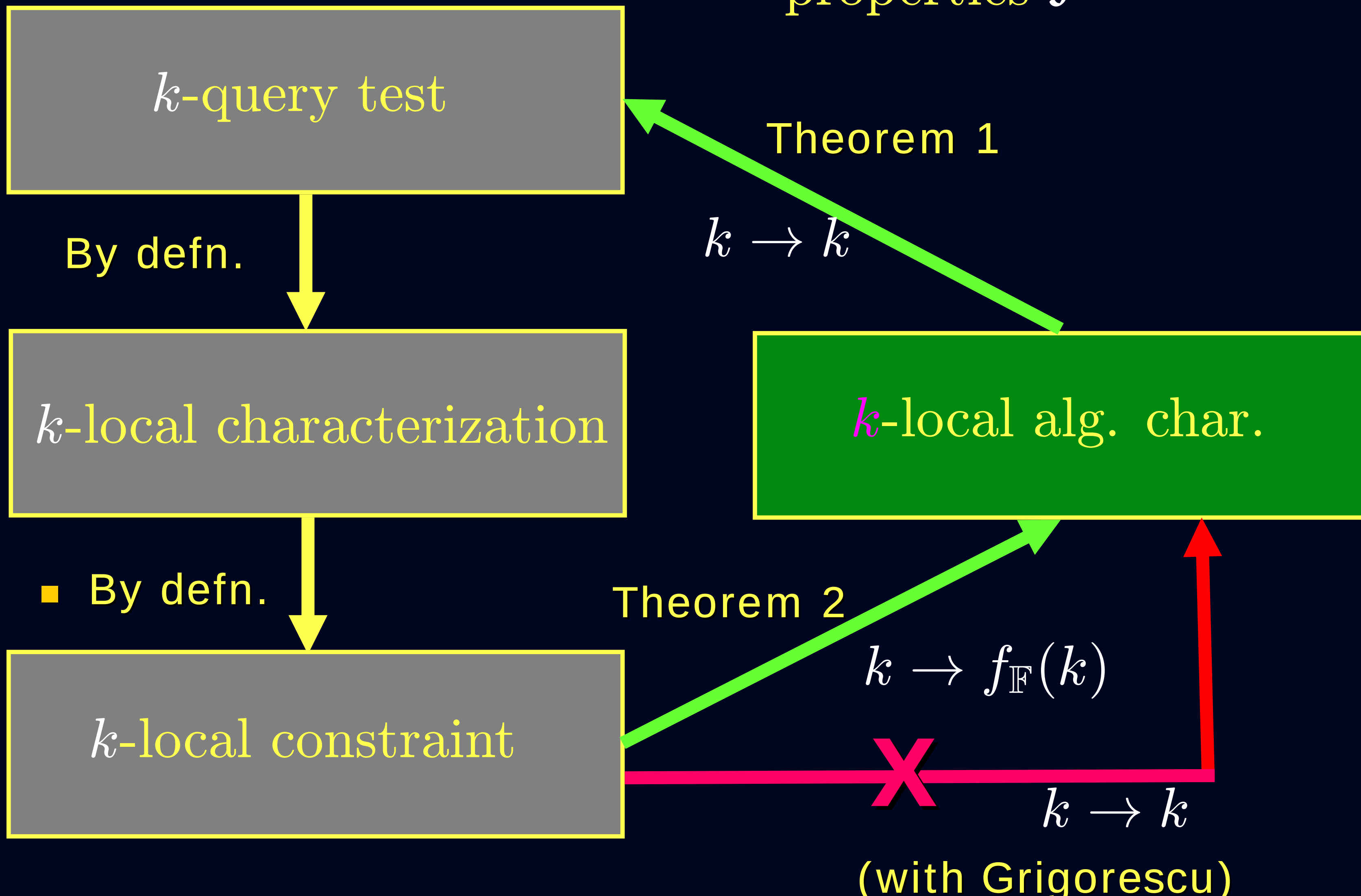
Pictorially

For affine-invariant properties $\mathcal{F}$



Pictorially

For affine-invariant properties $\mathcal{F}$



Part III: BLR (and our) analysis

BLR Analysis: Outline

- Have f s.t. $\Pr_{x,y}[f(x) + f(y) \neq f(x+y)] = \delta < 2/9$.
  Want to show f close to some $g \in \mathcal{F}$.
- Define $g(x) =$ most likely $_y\{f(x+y) - f(y)\}$.
- If f close to $\mathcal{F}$ then g will be in $\mathcal{F}$ and close to f.
- But if f not close? g may not even be uniquely defined!
- Steps:
  - Step 0: Prove f close to g
  - Step 1: Prove "most likely" is overwhelming majority.
  - Step 2: Prove that g is in $\mathcal{F}$.

BLR Analysis: Step 0

- Define $g(x) =$ most likely $_y\{f(x+y) - f(y)\}$.

Claim: $\Pr_x[f(x) \neq g(x)] \leq 2\delta$

– Let $B = \{x | \Pr_y[f(x) \neq f(x+y)f(y)] \geq \frac{1}{2}\}$

– $\Pr_{x,y}[$linearity test rejects $|x \in B] \geq \frac{1}{2}$

$\Rightarrow \Pr_x[x \in B] \leq 2\delta$

– If $x \notin B$ then $f(x) = g(x)$

$\mathrm{Vote}_x(y)$

BLR Analysis: Step 1

- Define $g(x) =$ most likely $_y\{f(x+y) - f(y)\}$.
- Suppose for some x, $\exists$ two equally likely values.
  Presumably, only one leads to linear x, so which one?
- If we wish to show g linear,
  then need to rule out this case.

Lemma: $\forall\, x$, $\Pr_{y,z}[\mathrm{Vote}_x(y) \neq \mathrm{Vote}_x(z))] \leq 4\delta$

$\text{Vote}_x(y)$

BLR Analysis: Step 1

- Define $g(x) =$ most likely $_y\{f(x+y) - f(y)\}$.
- Suppose for some x, $\exists$ two equally likely values.
  Presumably, only one leads to linear x, so which one?
- If we wish to show g linear,
  then need to rule out this case.

Lemma: $\forall\, x$, $\Pr_{y,z}[\text{Vote}_x(y) \neq \text{Vote}_x(z))] \leq 4\delta$

Vote$_x(y)$

BLR Analysis: Step 1

- Define $g(x) =$ most likely $_y\{f(x+y) - f(y)\}$.

Lemma: $\forall\, x,\ \Pr_{y,z}[\mathrm{Vote}_x(y) \neq \mathrm{Vote}_x(z))] \leq 4\delta$

?	$f(y)$	$-f(x+y)$
$f(z)$	$f(y+z)$	$-f(y+2z)$
$-f(x+z)$	$-f(2y+z)$	$f(x+2y+2z)$

Prob. Row/column sum non-zero $\leq \delta$.

$\mathrm{Vote}_x(y)$

BLR Analysis: Step 1

- Define $g(x) =$ most likely $_y\{f(x+y) - f(y)\}$.

Lemma: $\forall\, x$, $\Pr_{y,z}[\mathrm{Vote}_x(y) \neq \mathrm{Vote}_x(z))] \leq 4\delta$

?	$f(y)$	$-f(x+y)$
$f(z)$	$f(y+z)$	$-f(y+2z)$
$-f(x+z)$	$-f(2y+z)$	$f(x+2y+2z)$

Prob. Row/column sum non-zero $\leq \delta$.

BLR Analysis: Step 2 (Similar)

Lemma: If $\delta < \frac{1}{20}$, then $\forall\, x, y,\ g(x) + g(y) = g(x+y)$

$g(x)$	$g(y)$	$-g(x+y)$
$f(z)$	$f(y+z)$	$-f(y+2z)$
$-f(x+z)$	$-f(2y+z)$	$f(x+2y+2z)$

Prob. Row/column sum non-zero $\leq 4\delta$.

Our Analysis: Outline

- f s.t. $\Pr_L[\langle f(L(x_1),\ldots,f(L(x_k))\rangle \in V] = \delta \ll 1$.
- Define $g(x) = \alpha$ that maximizes
  $\Pr_{\{L|L(x_1)=x\}}[\langle \alpha, f(L(x_2)),\ldots,f(L(x_k))\rangle \in V]$

- Steps:
  - Step 0: Prove f close to g
  - Step 1: Prove "most likely" is overwhelming majority.
  - Step 2: Prove that g is in $\mathcal{F}$.

Our Analysis: Outline

- f s.t. $\Pr_L[\langle f(L(x_1),\ldots,f(L(x_k))\rangle \in V] = \delta \ll 1$.
- Define $g(x) = \alpha$ that maximizes $\Pr_{\{L \mid L(x_1)=x\}}[\langle \alpha, f(L(x_2)),\ldots,f(L(x_k))\rangle \in V]$

Same as before

- Steps:
  - Step 0: Prove f close to g
  - Step 1: Prove "most likely" is overwhelming majority.
  - Step 2: Prove that g is in $\mathcal{F}$.

$\mathrm{Vote}_x(L)$

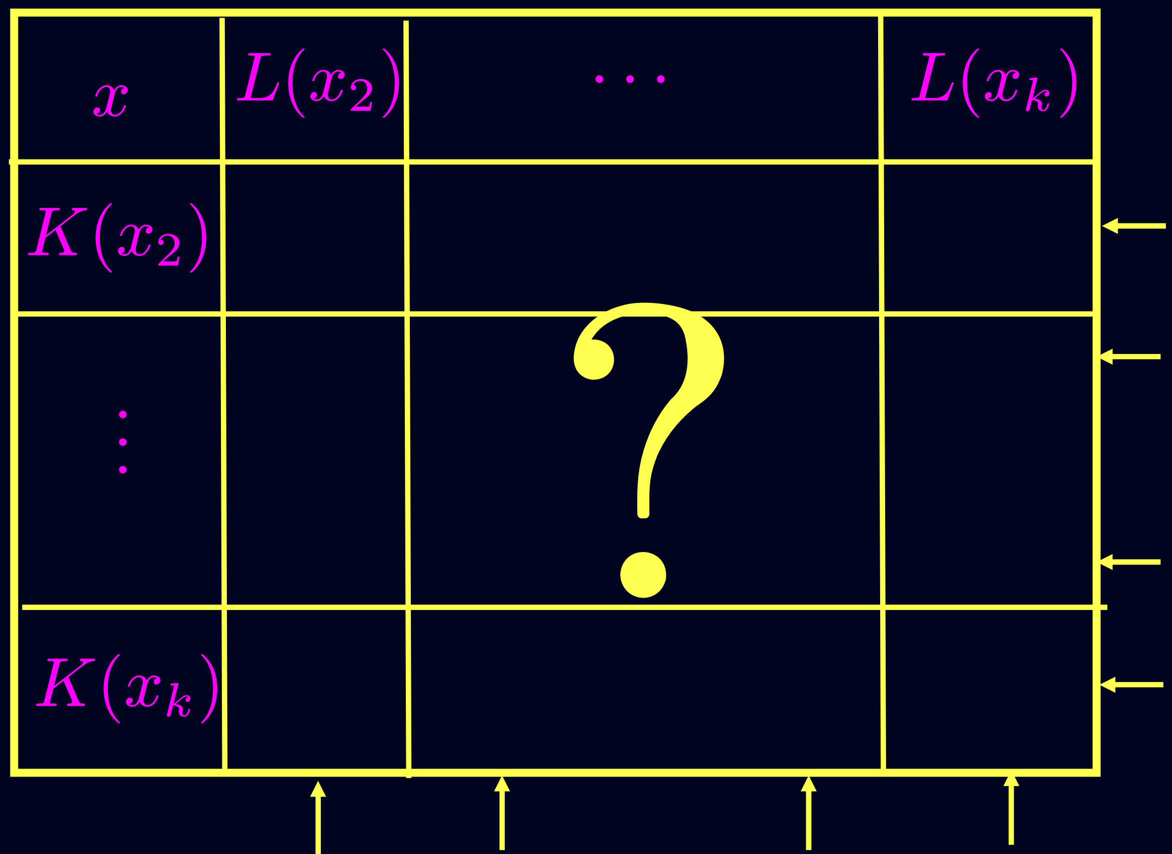
Matrix Magic?

- Define $g(x) = \alpha$ that maximizes
  $\Pr_{\{L \mid L(x_1)=x\}}[\langle \alpha, f(L(x_2)), \ldots, f(L(x_k))\rangle \in V]$

Lemma: $\forall\, x$, $\Pr_{L,K}[\mathrm{Vote}_x(L) \neq \mathrm{Vote}_x(K))] \leq 2(k-1)\delta$

x	$L(x_2)$	$\cdots$	$L(x_k)$
$K(x_2)$			
$\vdots$		?	
$K(x_k)$			

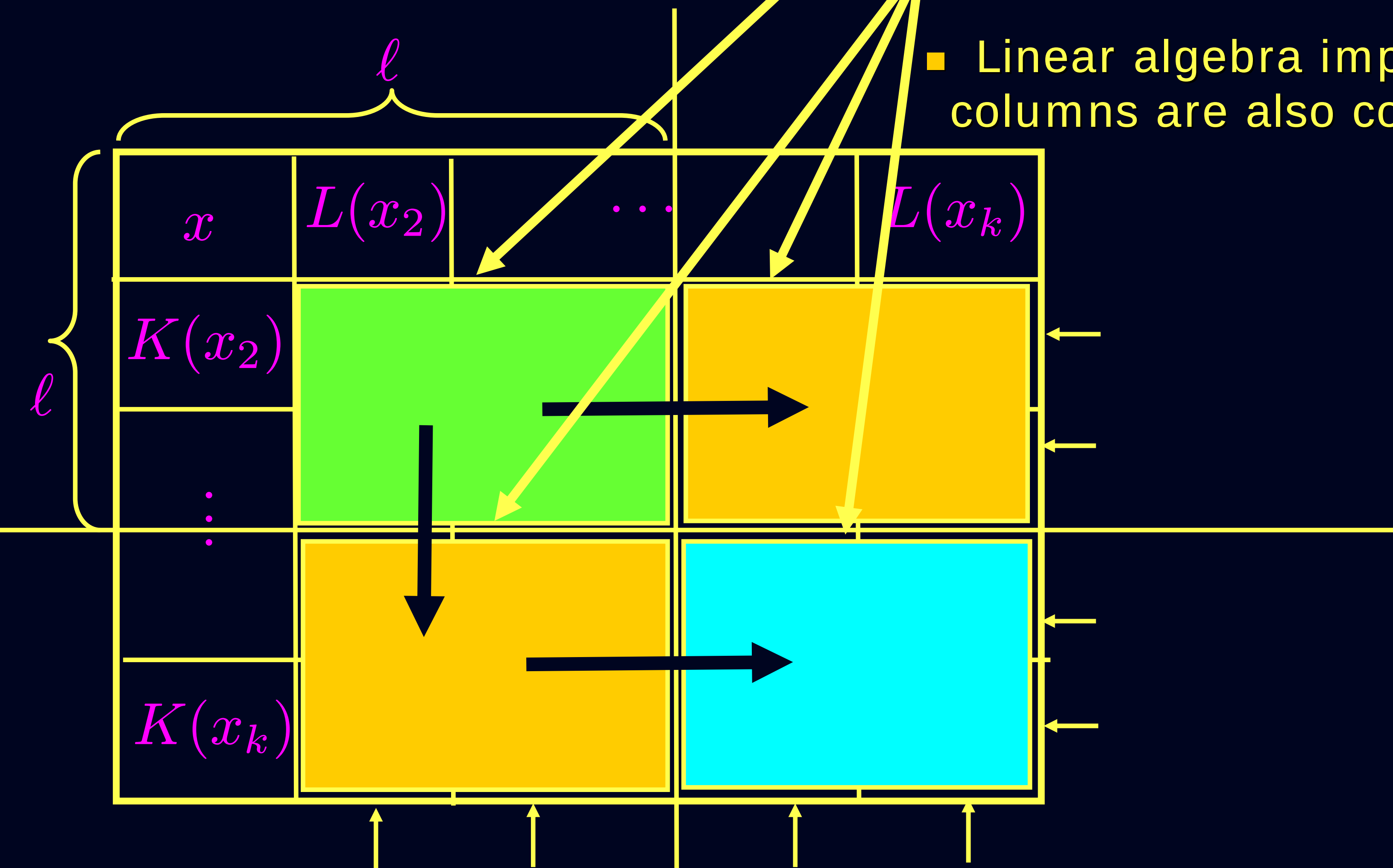
Matrix Magic?



- Want marked rows to be random constraints.
- Suppose $x_1, \ldots, x_\ell$ linearly independent; and rest dependent on them.

Matrix Magic?

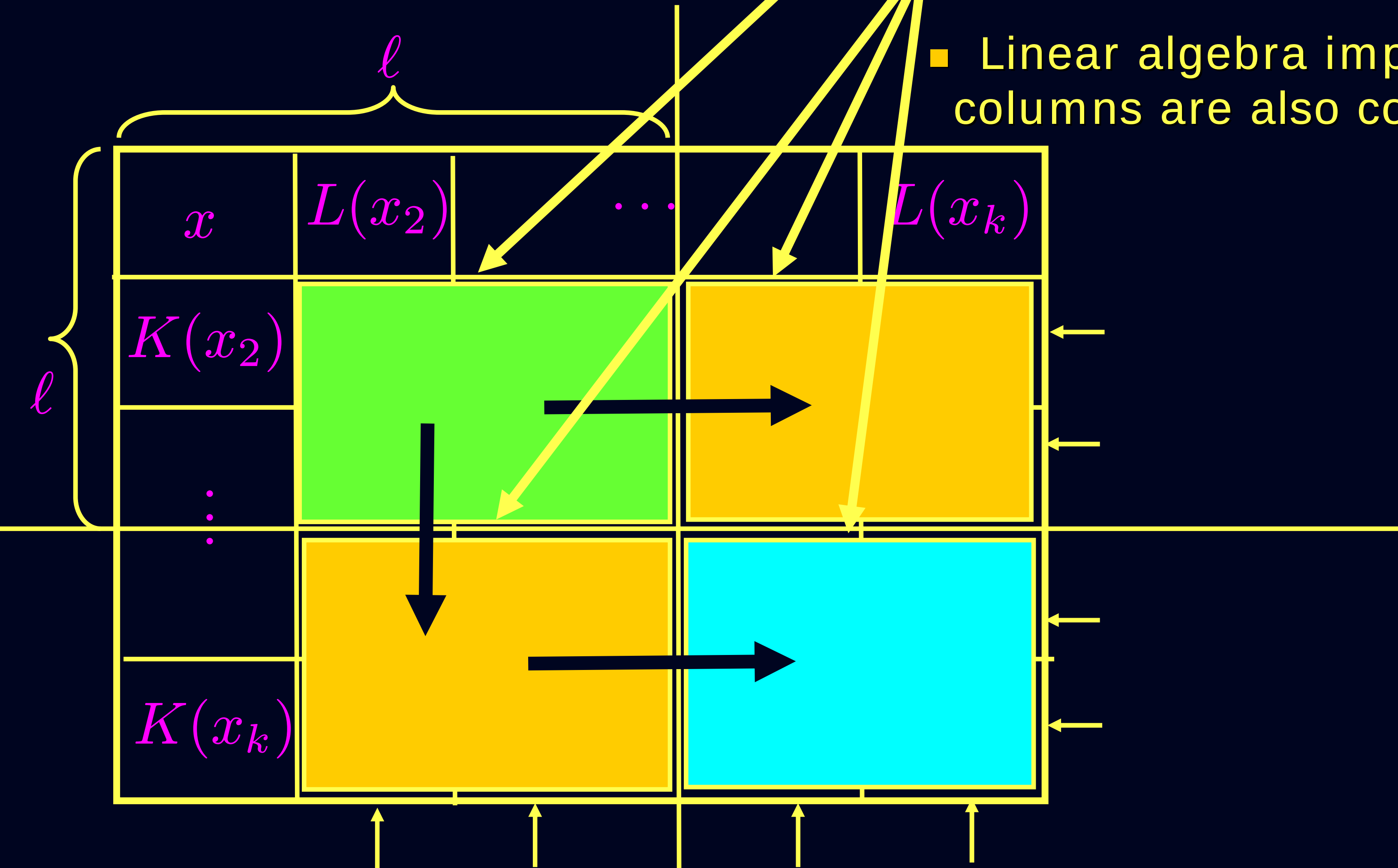
- Fill with random entries
- Fill so as to make constraints
- Linear algebra implies final columns are also constraints.



- Suppose $x_1, \ldots, x_\ell$ linearly independent; and rest dependent on them.

Matrix Magic?

- Fill with random entries
- Fill so as to make constraints
- Linear algebra implies final columns are also constraints.



- Suppose $x_1, \ldots, x_\ell$ linearly independent; and rest dependent on them.

Conclusions

- Linear/Affine-invariant properties testable if they have local constraints.
- Gives clean generalization of linearity and low-degree tests.
- Future work: What kind of invariances lead to testability (from characterizations)?