UNIVERSITÄT DES SAARLANDES

max planck institut informatik

Sublinear Algorithms

Lecture 04: Measurements I

European Research Council

Established by the European Commission

Last 3 lectures:

Streaming Algorithms

Main requirement: Maintain *sublinear* space

Often of interest: fast query and update time

Next 3 lectures:

Measurements

Access an object using queries/measurements,

e.g. linear combinations of its components

Main requirement:

Achieve *sublinear* measurement complexity

Sparse Recovery

Recovery of sparse structures using queries from a restricted class

Principle: Often 1% of an object carries 99% of the information

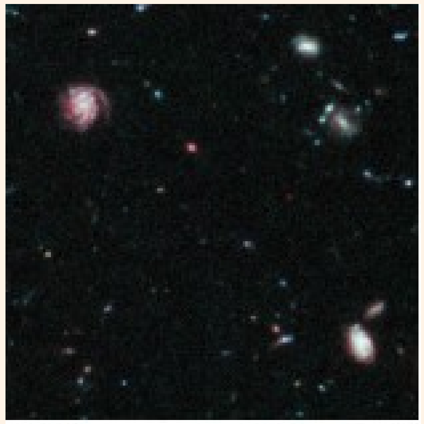
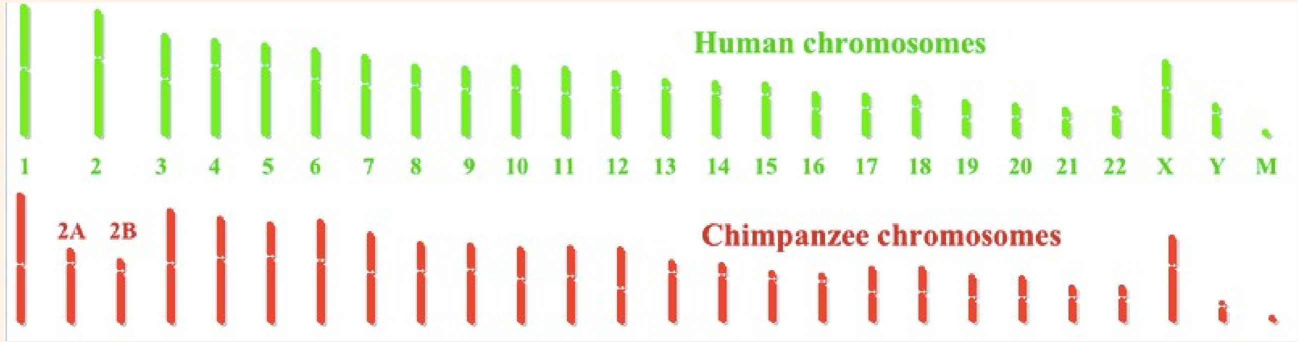


Image from Hubble telescope

Dense Image

Sparse Transformation of a Dense Image

Audio signals
are sparse

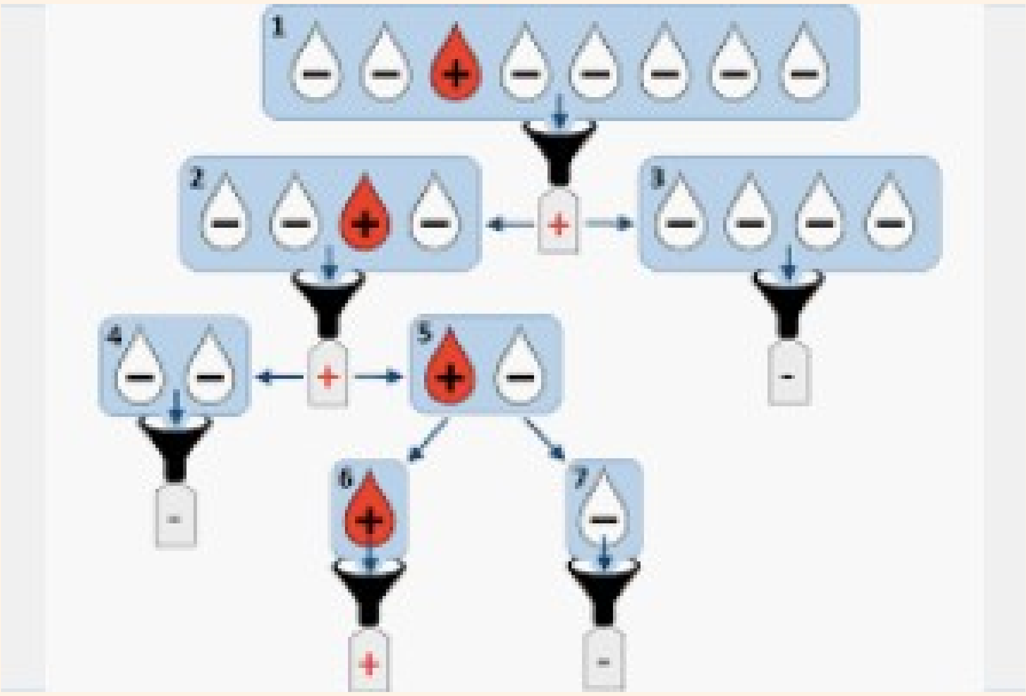


Difference between human and chimpanzee DNA
is a sparse vector

Combinatorial Group Testing
The syphilis problem

Dorfman (1942):
At most *k* of my *n* soldiers are infected with syphilis.
I need to find them using less than n tests!

Pool together patients.
If at least one is infected, then test outcome is ‘’Positive”
If none of them is infected, then test outcome is ‘’Negative”



Re-gained interest during COVID-19 pandemic

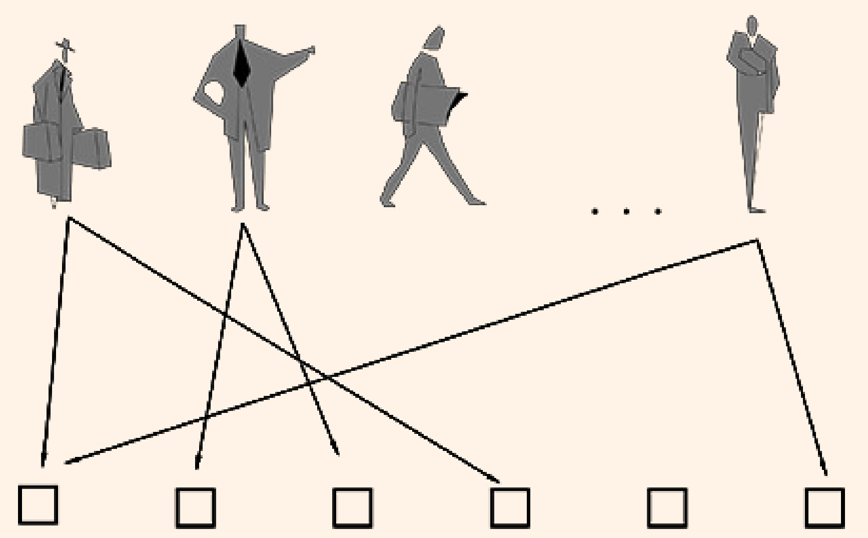
Combinatorial Group Testing
The basic set-up

Recover k-sparse $x \in \{\text{FALSE}, \text{RIGHT}\}^n$ using queries of the form $\bigvee_{i \in S} x_i$

Design measurements $S_1, S_2, \ldots, S_m$

such that, given for all $j \in [m] : y_j = \bigvee_{i \in S_j} x_i$

we can recover $x \in \{\text{FALSE}, \text{RIGHT}\}^n$ if it is *k*-sparse



In matrix form (Boole Algebra): $y = Mx$ where $M_{j,i} = (i \in S_j)$

(Looks like linear sketching, right?)

Main Goal: Minimize number of measurements (tests)

Combinatorial Group Testing
Guarantees

Non-uniform Group Testing: Recover a *fixed* k-sparse vector x with some target probability δ

Uniform Group Testing: Design measurements which allow recovery of *every* k-sparse vector x

$$\Sigma_{k,n} := \text{set of } x \in \{\textsc{False}, \textsc{True}\}^n \text{ with at most } k \text{ of the } x_i \text{ equal to } \textsc{True}$$

Non-uniform Group Testing
(with δ failure probability)

$$\forall x \in \Sigma_{k,n} : \Pr\{\textsc{Find}(y) \neq x\} \leq \delta$$

Uniform Group Testing
(with δ failure probability)

$$\Pr\{\exists x \in \Sigma_{k,n} : \textsc{Find}(y) \neq x\} \leq \delta$$

Uniform is NOT *Deterministic*

Non-uniform Group Testing

$O(k \log n)$ measurements suffice

In every measurement S_j include *i* with probability 1/k

No false negatives: If *i* is infected (xi = True), then all the tests it participates in will be positive.

False positive: A non-infected *i* (xi = False) such that all the tests it participates in turn out to be positive.

Consider test *j* in which *i* participates in

$$\Pr\{y_j = 1\} = 1 - \left(1 - \frac{1}{k}\right)^k \leq 1 - \frac{1}{e}$$

Why?

$$\Pr\{i \text{ false positive}\} = (1 - \frac{1}{e})^{c \log n} \leq \frac{1}{\text{poly}(n)}$$

+ *union bound*

Find() : Keep every *i* which participates only in positive tests.

Recovery Time $O(n \log n)$

Failure Probability $\frac{1}{\text{poly}(n)}$

Uniform Group Testing, or
One Matrix to rule them all

Find(): Keep every i which participates only in positive tests.

Disjunct matrices
A set of measurements with the following combinatorial property

$$\forall T \subseteq [n], |T| = k+1 :, \exists j \in [m] \text{ and } i \in T :$$
$$T \cap S_j = \{i\}$$

Disjunct matrices allow uniform group testing

Proof: Assume there exists a false positive i.
Let T be the set of infected individuals along with i.
Then T violates the disjunctness property.

Some bounds on measurements:

$$\Omega(k^2 \log_k n)$$
$$O\left(k^2 \min\left\{\log n, (\log_k n)^2\right\}\right)$$

Gap since the 60s:
connected to a major question in
Coding Theory

Group Testing and the CountMin sketch

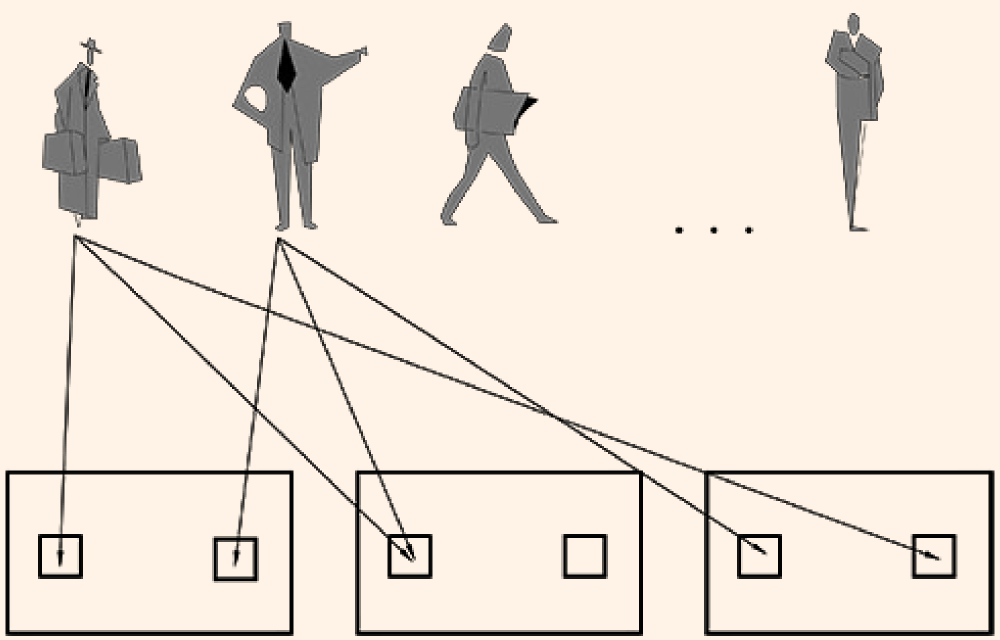
Pick (random) hash functions $h_r : [n] \to [2k]$ for $r \in [R]$

Pick measurements $h_r^{-1}(b) = \{i \in [n] : h_r(i) = b\}$

#measurements $2kR$

Uniform

$R = \Theta(k \log n)$



Non-uniform

$R = \Theta(\log n)$

In CountMin sketch $R = \log(1/\delta)$ drives down the failure probability

You may think of $k = 1/\varepsilon$

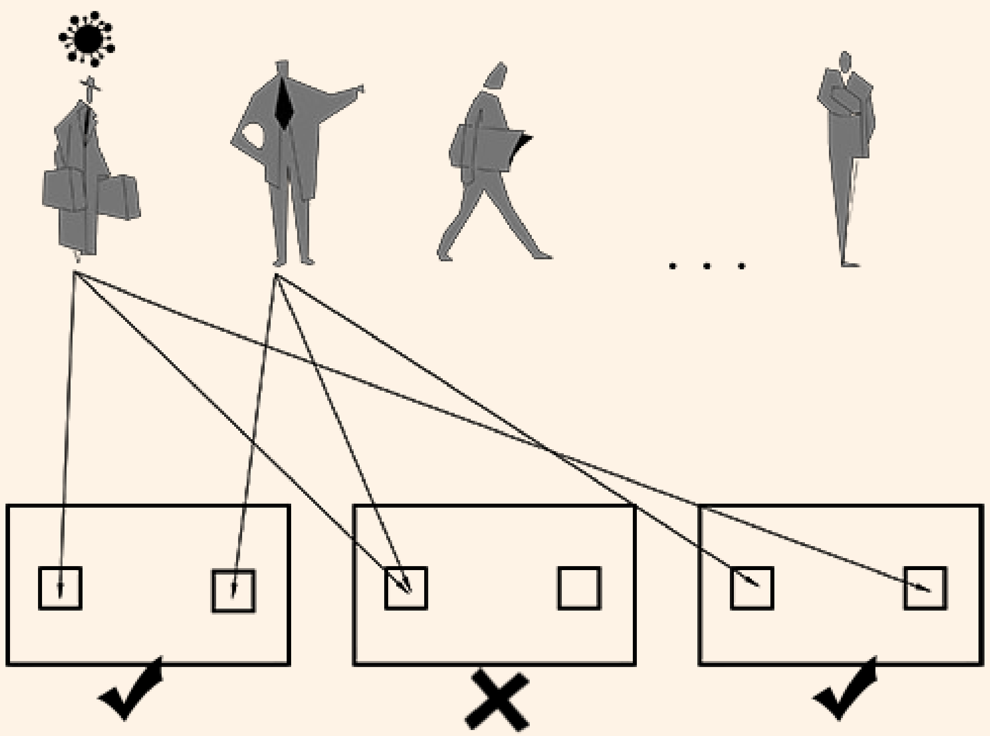
Analysis of disjunctness

Find(): Keep every *i* which participates only in positive tests.

Fix *(T,i), T* a *k*-element subset of [n] and *i* an element outside of *T*

$$\Pr\{h_r(i) \in h_r(T)\} \leq \frac{k}{2k} = \frac{1}{2}$$

$$\Pr\{h_r(i) \in h_r(T), \forall r \in [R]\} = \left(\frac{1}{2}\right)^R$$



$$\binom{n}{k} \approx e^{k \log(n/k)}$$

$$\Pr\{\text{disjunctness fails}\} \leq n \cdot \binom{n}{k} \cdot \left(\frac{1}{2}\right)^R \leq \frac{1}{3}$$ since $R = \Theta(k \log n)$

#measurements: $O(k^2 \log n)$

Recovery Time: $O(nk \log n)$

Two-stage group testing

We are allowed *two* rounds of adaptivity

Narrow down the set of possible infected individuals to *2k*, then perform a test on each one.

The correct combinatorial construct
List-disjunct matrices

Written in terms of hash functions

$$\forall S, T \subseteq [n], S \cap T = \emptyset, |S| = k, |T| = k+1,$$
$$\exists j \in T, r \in [R] : h_r(j) \notin h_r(S)$$

Find(): Keep every *i* which participates only in positive tests.

↓

At most k false negatives

$O(k \log(n/k))$ measurements suffice (and is optimal)

Analysis of List-Disjunctness

Fix *(S,T)*.

Fix *i* in *T*.

$$\Pr\{h_r(i) \in h_r(S)\} \leq \frac{k}{2k} = \frac{1}{2}$$

Probability that all *i* in *T* appear as false positive in one repetition:

$$\Pr\{h_r(i) \in h_r(S), \forall i \in T\} \leq \left(\frac{1}{2}\right)^{k+1}$$

Probability that there exists a false negative

$$\Pr\{h_r(i) \in h_r(S), \forall i \in T, r \in [R]\} \leq \left(\frac{1}{2}\right)^{(k+1)R}$$

And now a union-bound over all pairs *(S,T)*

$$\binom{n}{k} \cdot \binom{n-k}{k+1} \cdot \left(\frac{1}{2}\right)^{(k+1)R} \qquad \binom{n}{k} \approx e^{k \log(n/k)}$$

…smaller than *1/3* by choosing $R = \Theta(\log(n/k))$

Compact Representations

All the above constructions can be analyzed using
O(k)-wise independent hash functions

Can store measurements in *O(k R log n)* bits of space

For disjunct matrices and non-uniform group testing,
2-wise independence suffices (in problem set)

↓

Polynomial time derandomization for disjunct matrices

Practical Considerations

-Constraint 1: Cannot pick arbitrarily large groups

Additional Parameter: *Maximum group size*

↓

Sparse Group Testing

-Constraint 2: Some tests might fail

↓

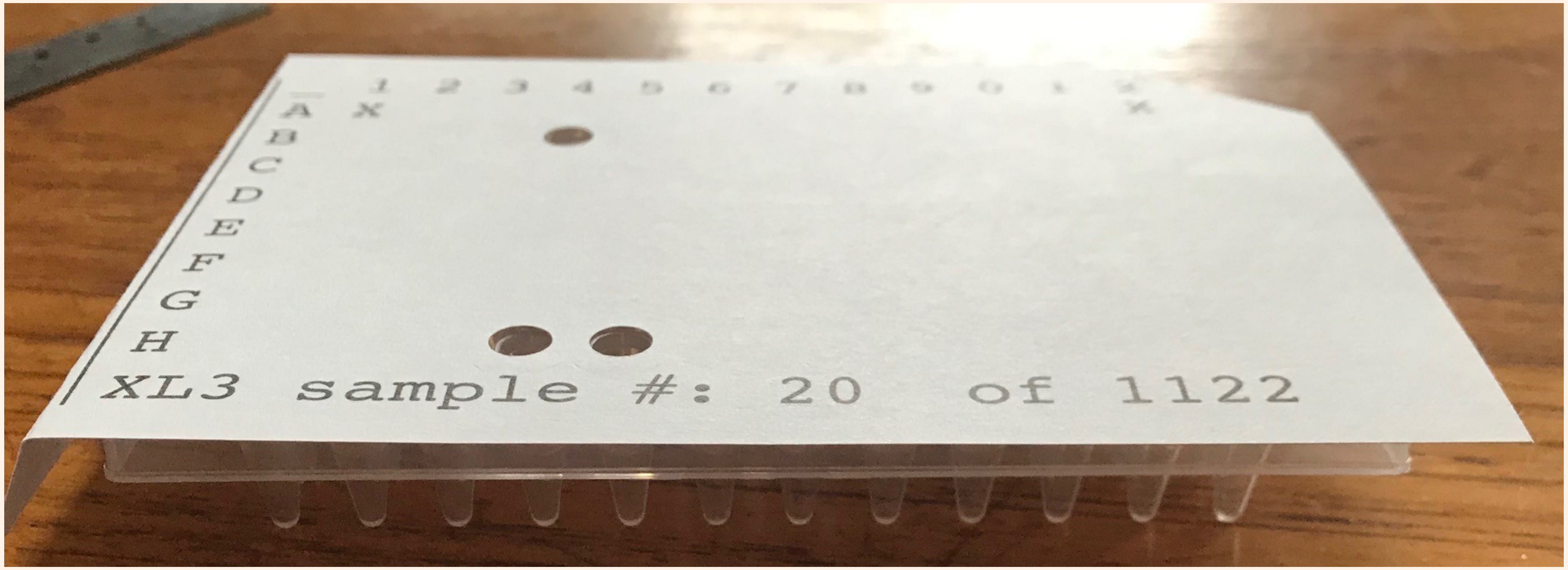
Error-Correcting Group Testing

-Constraint 3: Some times measurements must belong to a constrained ensemble

↓

Application-dependent group testing

Bonus



Origami Assay paper template for group testing design,
designed during COVID-19 pandemic

Provides paper templates to guide the technician
on how to allocate patient samples across the test wells

Disclaimer: I do not understand how it works

Outline

Non-uniform Group Testing

Uniform Group Testing:

Disjunct Matrices and List-Disjunct Matrices

Group Testing from CountMin sketches

Thank you