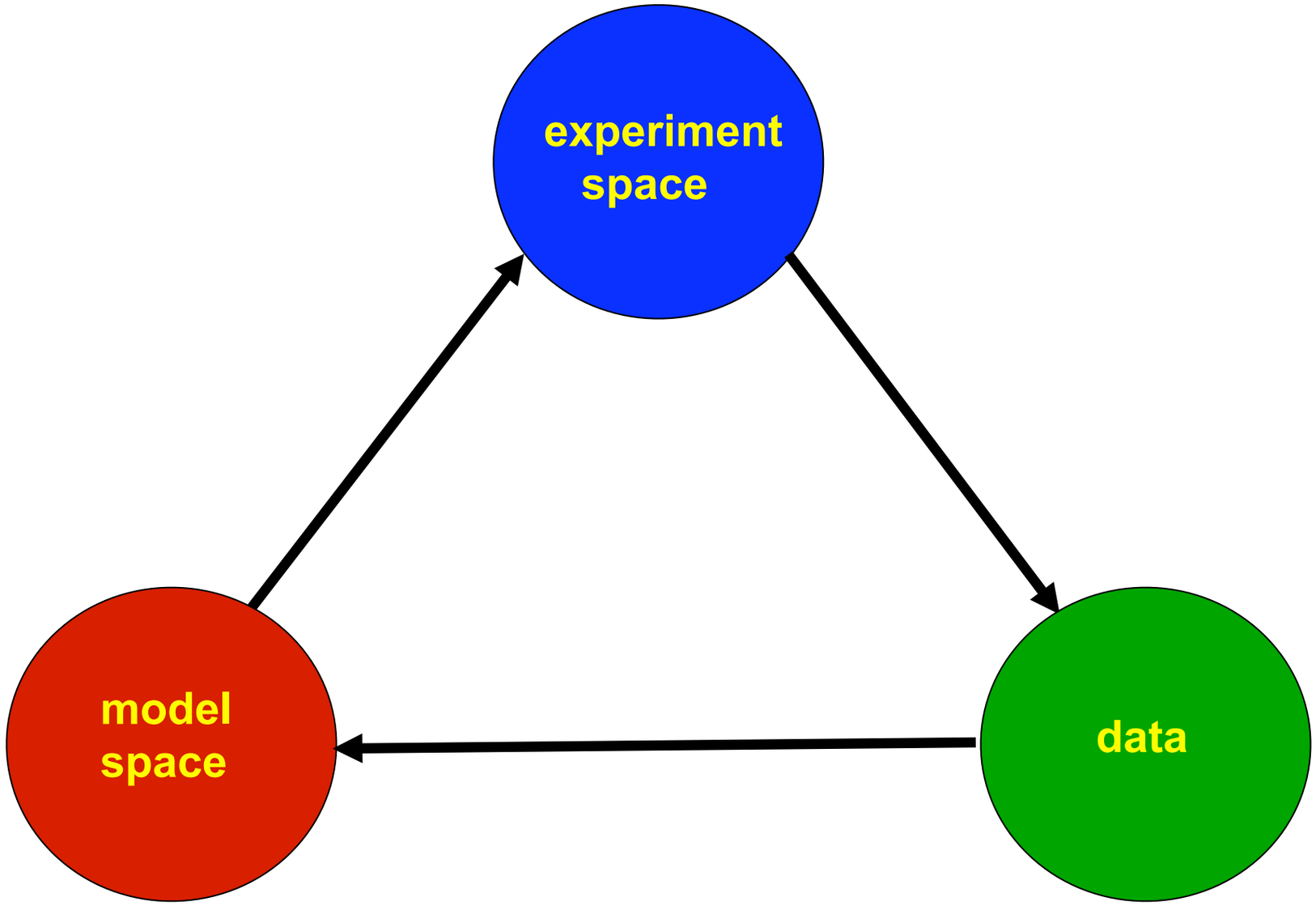
# Adaptive Information and Optimization




Rob Nowak
www.ece.wisc.edu/~nowak

WIDDOW 31 January 2011

Joint work with
R. Castro, J. Haupt, M. Malloy


WIDOW
"Echoes of White Sister, Giuffria, Icon, Surgin and Early Bon Jovi"
- Dave Ling, Classic Rock Magazine
The Critically Acclaimed Self Titled Debut Album
From Australia's Melodic Hard Rock Sensation
Available Now on AOR Heaven Records
THE HOME OF MELODIC ROCK
AOR HEAVEN
WWW.AORHEAVEN.COM

# Adaptive Information

**Goal:** Estimate an unknown object $x \in \mathcal{X}$ from scalar samples

**Information**: samples of the form $y_1(x), \ldots, y_n(x)$,
the values of certain functionals of $x$

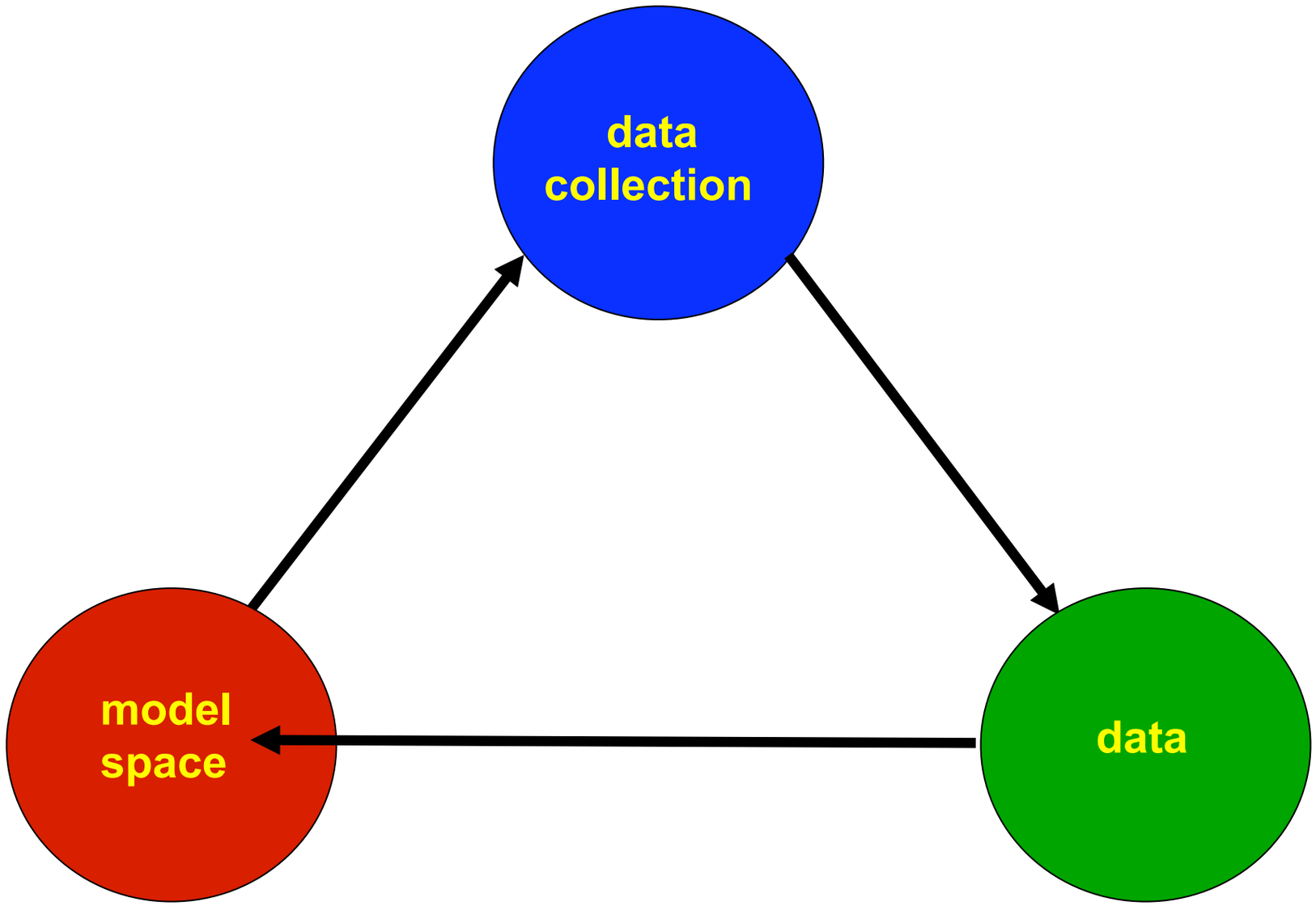
**Non-Adaptive Information**: $y_1, y_2, \cdots \in \mathcal{Y}$ non-adaptively chosen (deterministically or randomly) independent of $x$

**Adaptive Information**: $y_1, y_2, \cdots \in \mathcal{Y}$ are selected sequentially and $y_i$ can depend on previously gathered information, i.e., $y_1(x), \ldots, y_{i-1}(x)$

**Does adaptivity help?**

# Feedback from Data Analysis to Data Collection

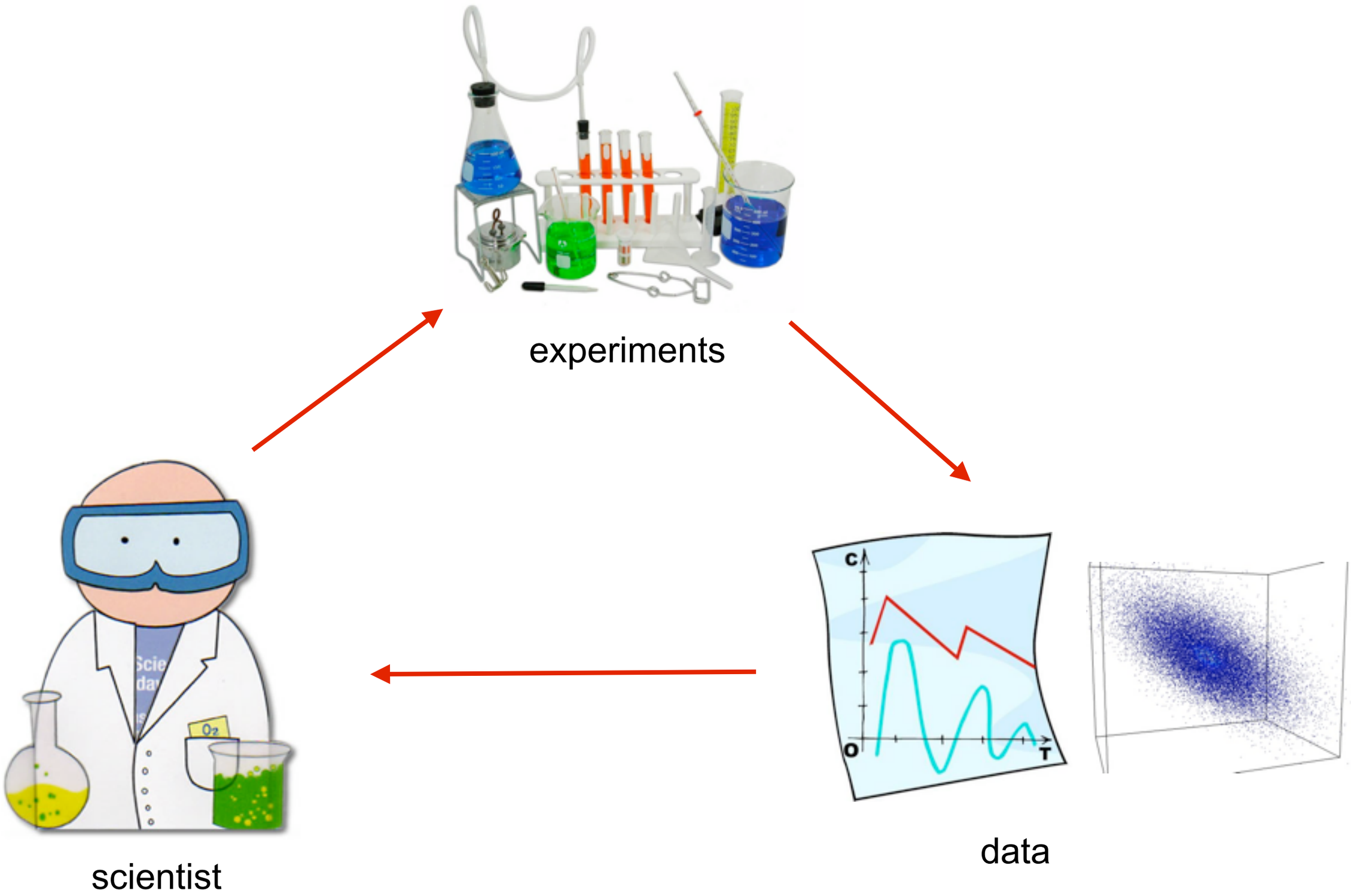
$\mathcal{Y}$: possible measurements/experiments



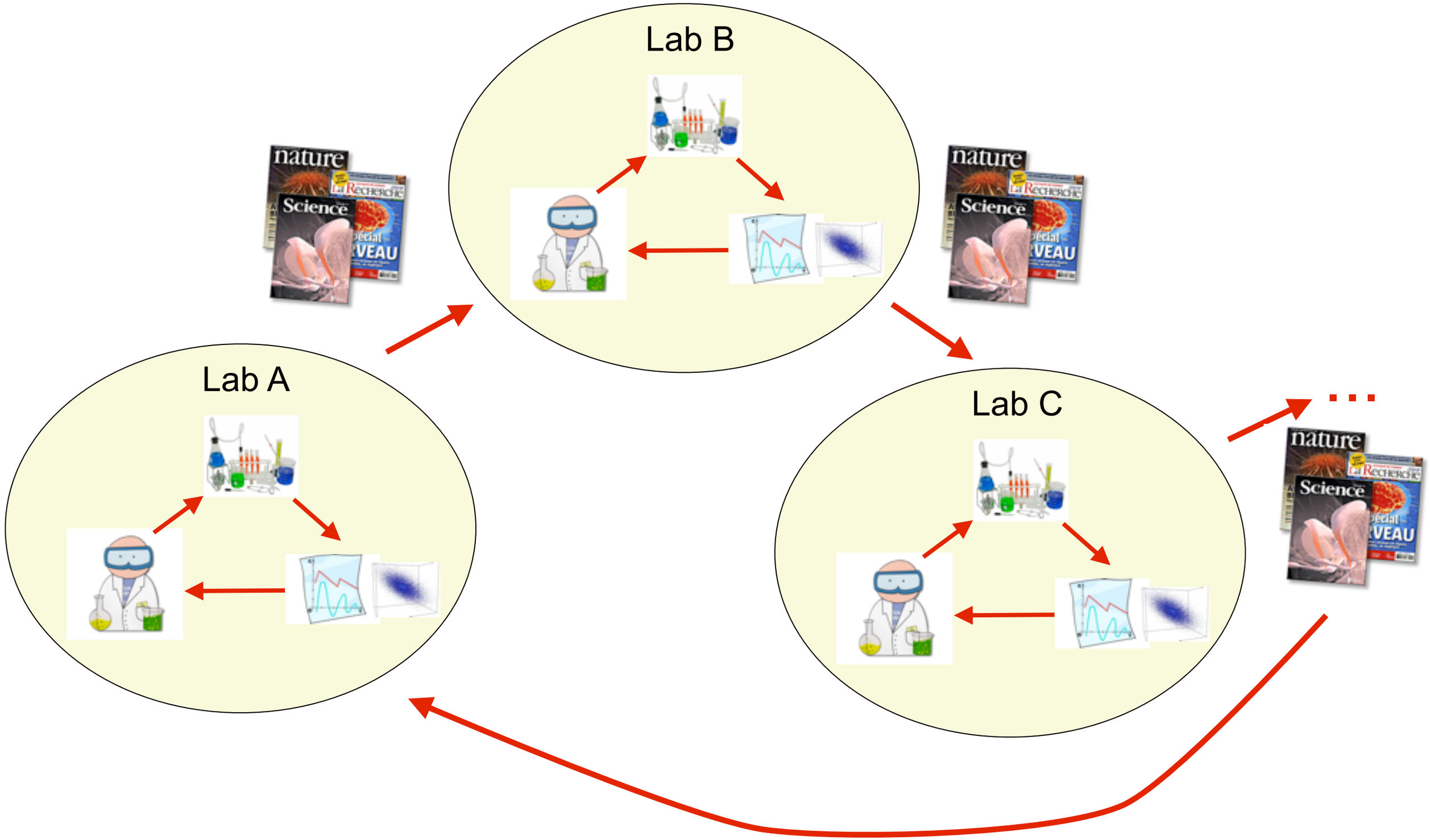

$\mathcal{X}$: models/hypotheses under consideration

$y_1(x), y_2(x), \ldots$: information/data

# The Scientific Process in a Laboratory

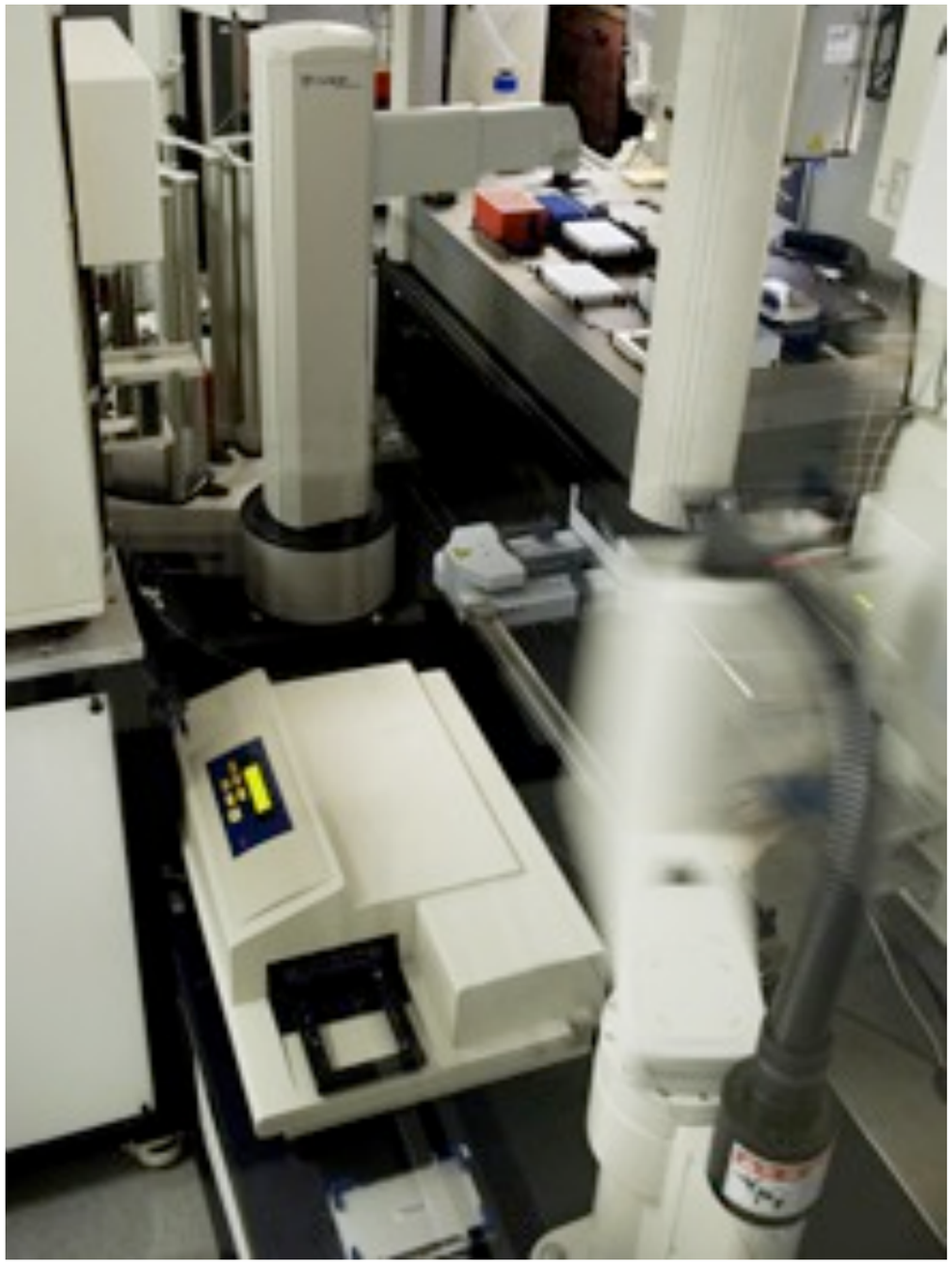


# The Scientific Process at Large



# Robot Scientist

www.aber.ac.uk/compsci/Research/bio/robotsci/



Wired Magazine, April 2009:

For the first time, a robotic system has made a novel scientific discovery with virtually no human intellectual input.

Scientists designed "Adam" to carry out the entire scientific process on its own: formulating hypotheses, designing and running experiments, analyzing data, and deciding which experiments to run next. "It's a major advance," says David Waltz of the Center for Computational Learning Systems at Columbia University. "Science is being done here in a way that incorporates artificial intelligence. It's automating a part of the scientific process that hasn't been automated in the past."

*Adam is the first automated system to complete the cycle from hypothesis, to experiment, to reformulated hypothesis without human intervention.*

# Adaptive vs. Non-Adaptive: Three Situations

The "bare minimum" number of measurements depends on intrinsic complexity of $\mathcal{X}$. In practice, the minimum number depends on jointly on $\mathcal{X}$ and $\mathcal{Y}$.

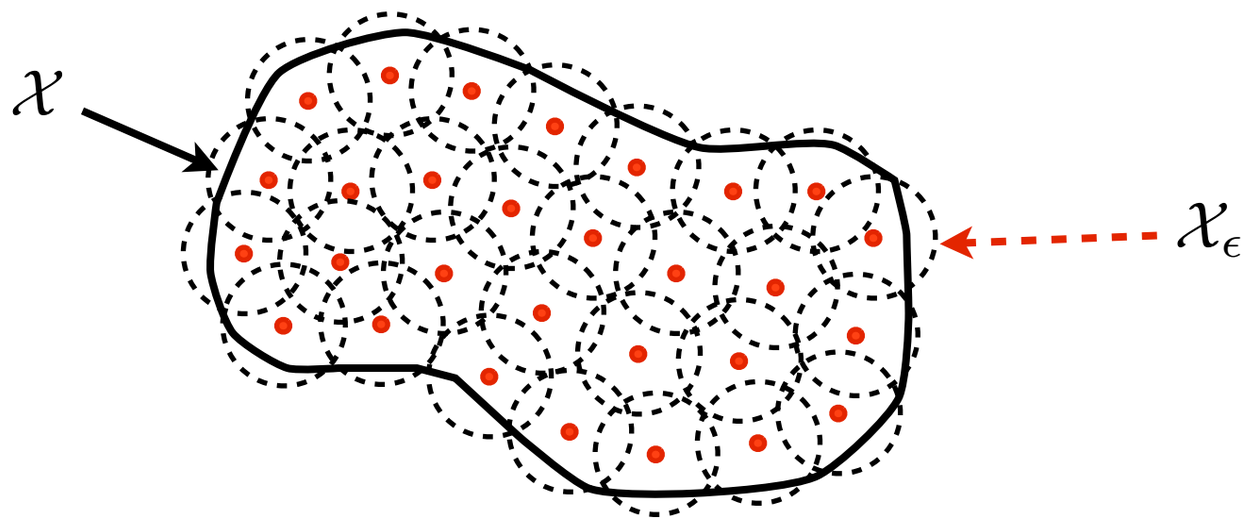
**Equal and Good**:
adaptive and non-adaptive equally informative and require about the bare minimum of measurements

**Equal and Bad**:
adaptive and non-adaptive equally (non)-informative and require many more measurements then the bare minimum

**Good and Bad**:
adaptive requires bare minimum number of measurments, non-adaptive requires many more

# The Bare Minimum

Assume $\mathcal{X}$ is equipped with metric $d$ and is compact.




Let $\mathcal{X}_\epsilon \subset \mathcal{X}$ be a finite subset of size $N_\epsilon$ having the property that any element of $\mathcal{X}$ is within distance $\epsilon$ of an element in $\mathcal{X}_\epsilon$

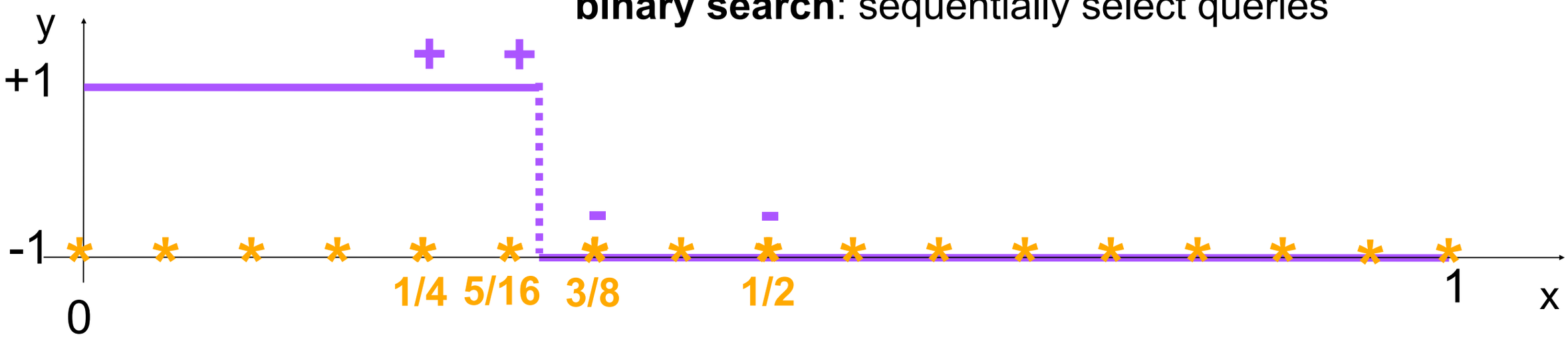
**Metric Entropy:** Need at least $\log N_\epsilon$ bits of information to approximately determine any $x \in \mathcal{X}$

Ex. suppose $\mathcal{X} = [0,1]^d$. we can take a uniform grid of points spaced $\epsilon$ apart as our cover. Then $N_\epsilon = (\frac{1}{\epsilon})^d$ and $\log N_\epsilon = d \log(1/\epsilon)$.

# Binary Search

$$\mathcal{X} = \{\text{subsets } [0, \tfrac{1}{N}], [0, \tfrac{2}{N}], \ldots, [0, 1]\}$$
$$\mathcal{Y} = \text{“membership queries”}$$

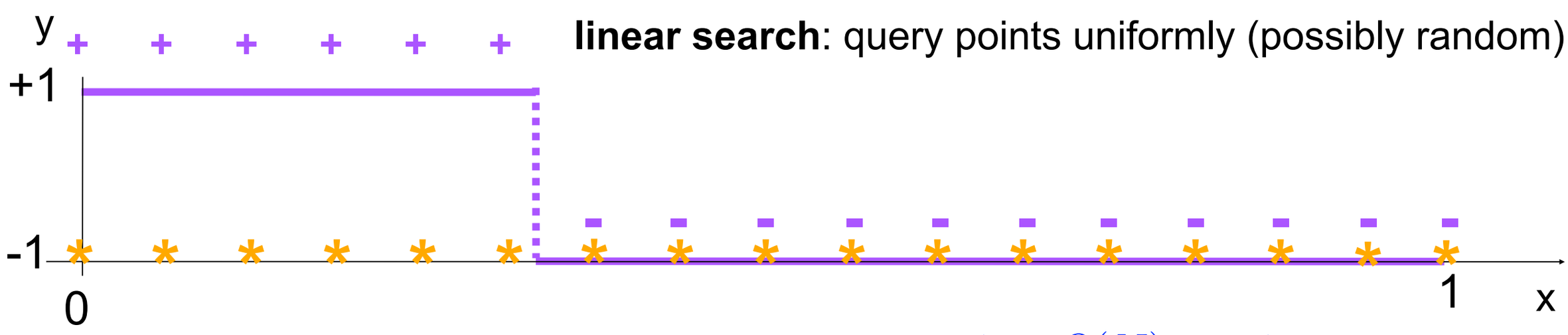
**binary search**: sequentially select queries




**1/3 = 0101…**

requires $\log_2 N$ queries

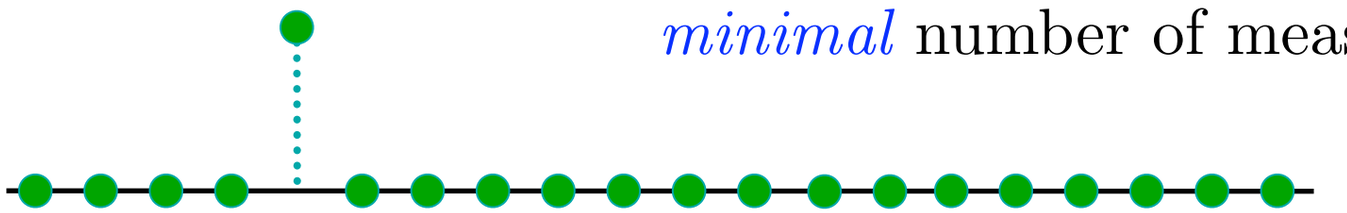
**linear search**: query points uniformly (possibly random)




requires $O(N)$ queries

# Does Adaptivity Help ?

identify a sparse signal $x \in \mathbb{R}^n$ from a *minimal* number of measurements



**Point measurments**: $y = \langle x\,,\ \delta_k \rangle = x_k$

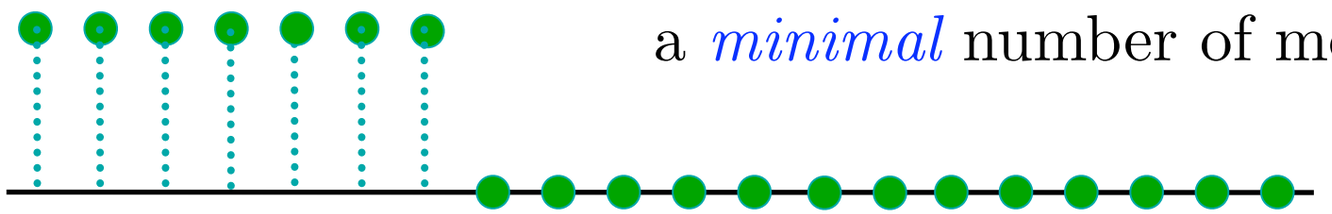
$O(n)$ measurements (random or adaptive) are needed to recover $x$

**Compressed Sensing**: $y \ = \ \langle x\,,\, \phi \rangle$ where $\phi \in \{-1, 1\}^n$

$O(\log n)$ measurements (random or adaptive) are needed to recover $x$

**Adaptivity doesn't help**

# Does Adaptivity Help ?



identify a threshold signal $x \in \mathbb{R}^n$ from a *minimal* number of measurements

**Point measurments**: $y = \langle x\,,\ \delta_k \rangle = x_k$

$O(n)$ random measurements are needed to recover $x$

$O(\log n)$ adaptive measurements are needed to recover $x$ (binary search)

**Compressed Sensing**: $y \ = \ \langle x\,,\, \phi \rangle$ where $\phi \in \{-1, 1\}^n$

$O(\log n)$ random measurements are needed to recover $x$

**Adaptivity may help, depending on nature of signal and measurements**

# Optimizing Information Collection

**Goal:** Estimate an unknown object $x \in \mathcal{X}$ from scalar samples

**Information**: samples of the form $y_1(x), \ldots, y_n(x)$,
the values of certain functionals of $x$

**Adaptive Information**: $y_1, y_2, \cdots \in \mathcal{Y}$ are selected sequentially and $y_i$ can depend on previously gathered information, i.e., $y_1(x), \ldots, y_{i-1}(x)$

**Dynamic Programming:** $K > 0$ measurement/experiment steps

$$\min_{\widehat{x}\,,\, y_1,\ldots,y_K} \; \max_{x \in \mathcal{X}} \; d(x\,,\, \widehat{x}(y_1, \ldots, y_K))$$

computationally prohibitive in all but very low-dimensional, simple problems

# Greedy Strategies

**Ex. Binary Information:** for each $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,

$$y(x) = \begin{cases} +1 & , \quad \text{if } x \text{ predicts a positive outcome on } y \\ -1 & , \quad \text{if } x \text{ predicts a negative outcome on } y \end{cases}$$

optimal procedure is a search tree; construction is NP-complete (Hyafil & Rivest '76)

**Splitting Algorithm**

initialize: $n = 0$, $\mathcal{X}_0 = \mathcal{X}$

while $|\mathcal{X}_n| > 1$

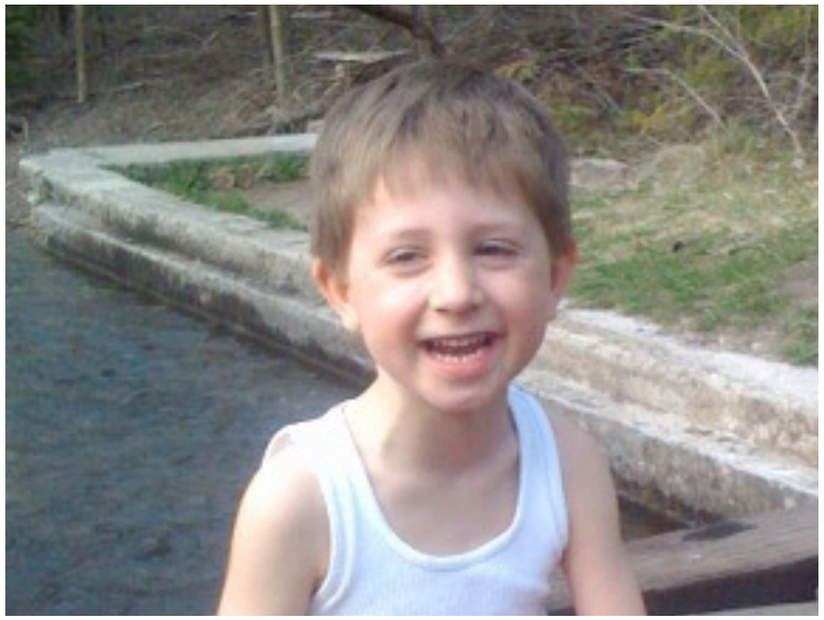
1) Select $y_n = \arg\min_{y \in \mathcal{Y}} |\sum_{x \in \mathcal{X}_n} y(x)|$

2) Perform $y_n$ to obtain information $y_n(x^*)$

3) Set $\mathcal{X}_{n+1} = \{x \in \mathcal{X}_n : y_n(x) = y_n(x^*)\}$, $n = n + 1$

Splitting Algorithm is near-optimal (average depth is within $\log |\mathcal{X}|$ factor of optimal)

**depth of optimal tree depends on nature of $\mathcal{X}$ and $\mathcal{Y}$**

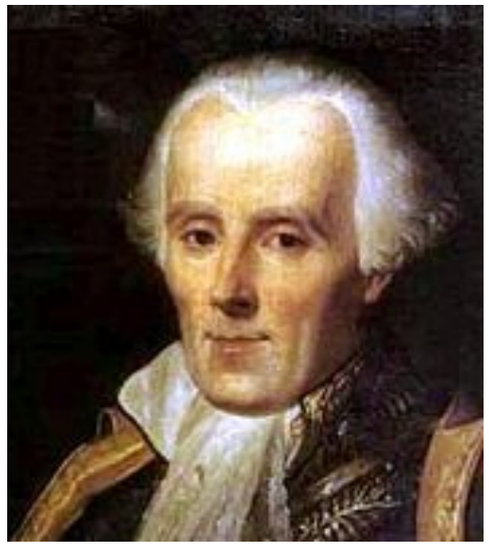



“Is the person wearing a hat ?”

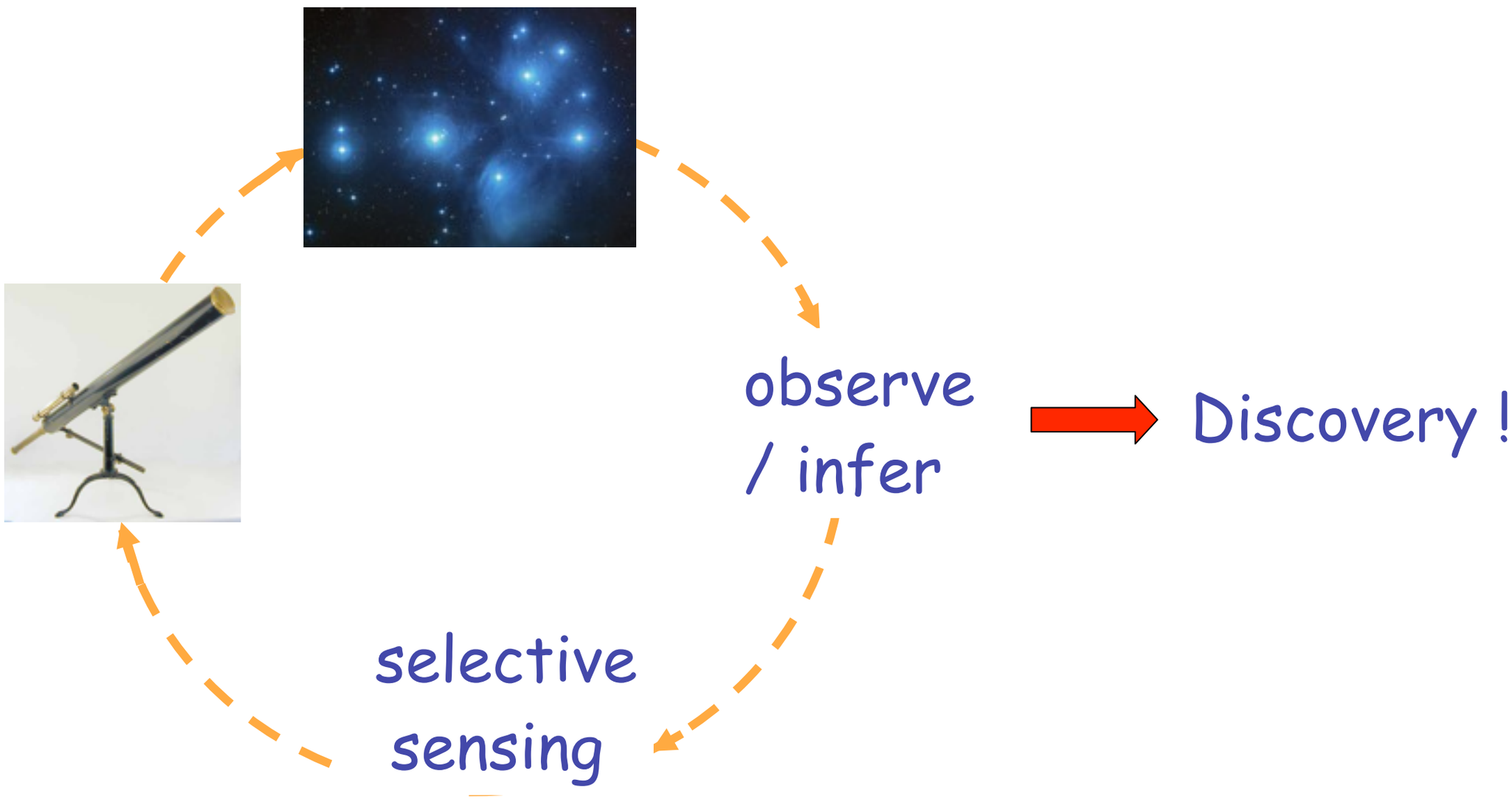
“Does the person have blue eyes ?”

splitting algorithm is quite effective if responses are reliable

# Laplace



Decided to make new astronomical measurements when “the discrepancy between prediction and observation [was] large enough to give a high probability that there is something new to be found.”  Jaynes (1986)



# Probabilistic Splitting Algorithm

**Probabilistic Splitting Algorithm**

initialize: $p_0(x) =$ uniform

for $n = 0, 1, \ldots, k-1$

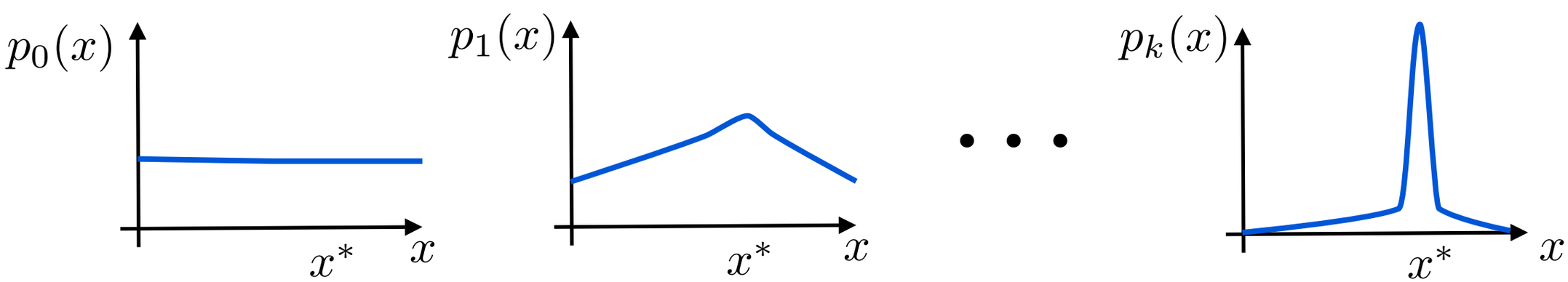
1) Select $y_{n+1}$ to maximize $\mathbb{E}_y \left[ \int p_n(x|y) \log \frac{p_n(x|y)}{p_n(x)} \, dx \right]$

2) Perform $y_{n+1}$ to obtain information $y_{n+1}(x^*)$

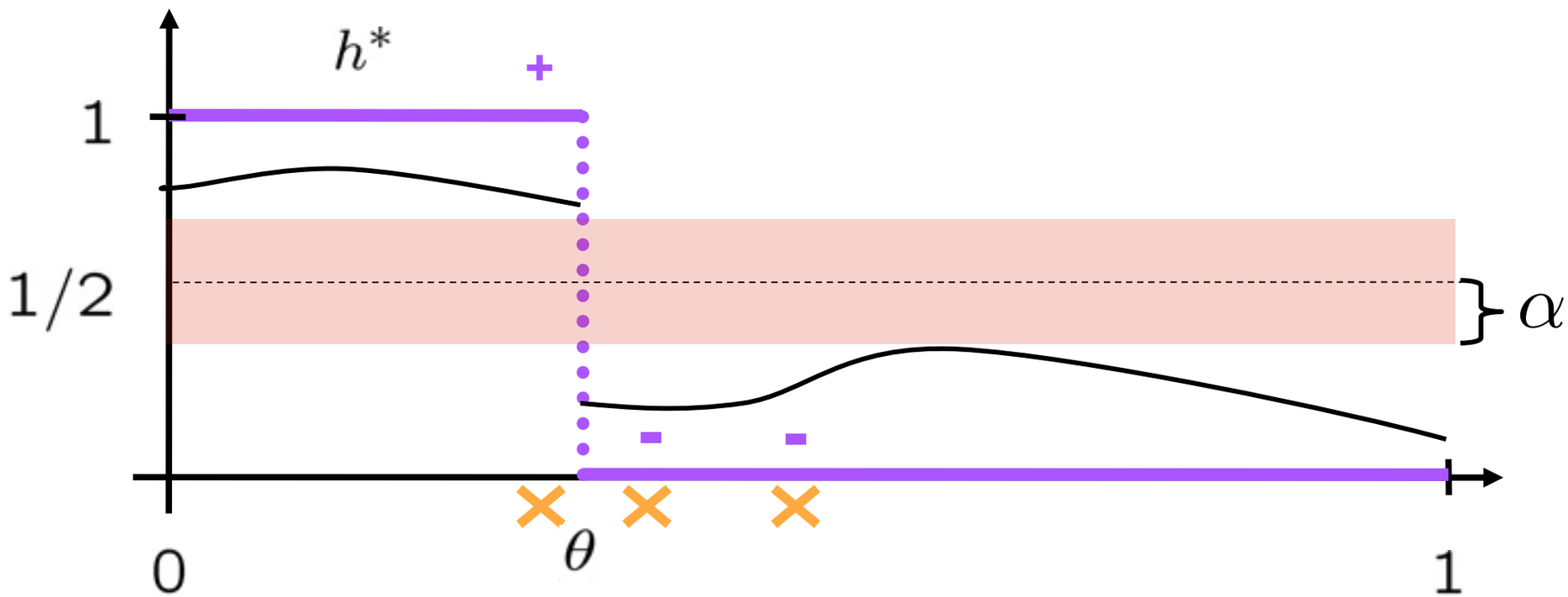
3) $y_{n+1}(x^*)$ + Bayes rule: $p_n(x) \to p_{n+1}(x)$

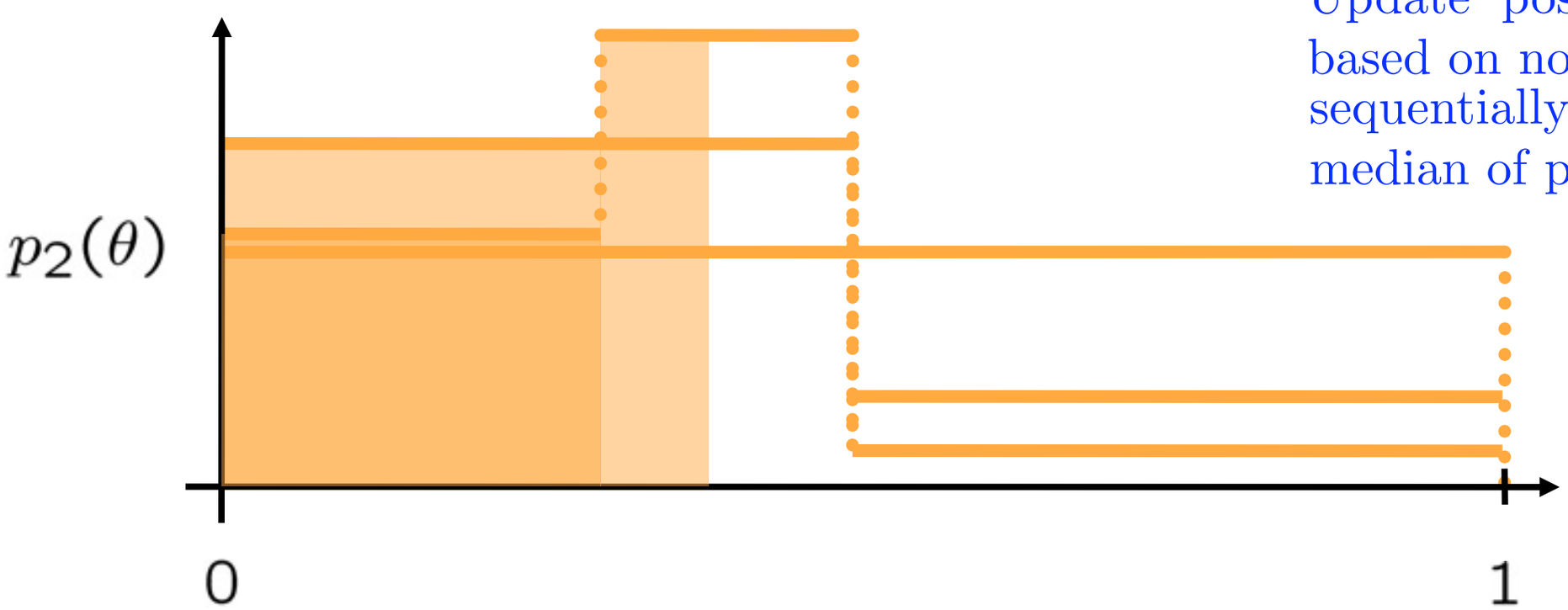
output: $\arg\max_x p_k(k)$

**"Information-Gain"**
(Shannon '48, Lindley '56)



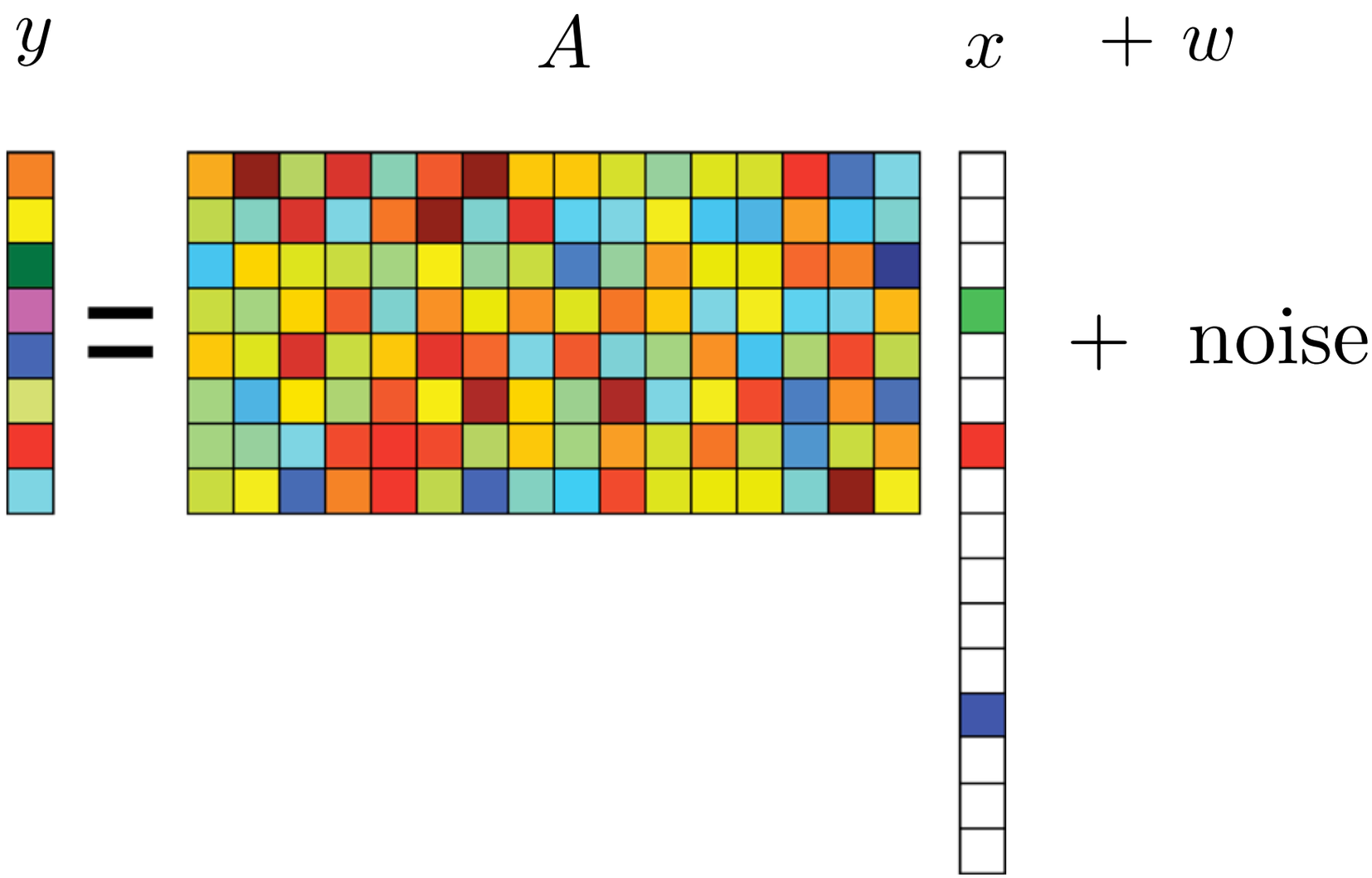
# Ex. Noisy Binary Search (Burnashev & Zigangirov '74)




Update 'posterior' density based on noise bound $\alpha$ sequentially take samples at median of posterior density

# Sparsity and High-Dimensional Models




$y \;=\; A\,x \;+\; w$ , with $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ (but sparse), $w \sim \mathcal{N}(0, I)$

# Signal Processing Gear

back to shop

## y = phi x (Dark T-Shirt)

$18.99

**Fit:** Standard

Not too tight, not too loose.

**Fabric Thickness:**

Thin — Thick

1. Color: (Charcoal)
2. Size: Large — Size Chart
3. Qty: 1

**ADD TO CART**

AVAILABILITY: In Stock.
Product Number: 030-469487567

Like — Sign Up to see what your friends like.

Share |

A visual representation of the math behind compressive sensing

Look cool without breaking the bank. Our durable, high-quality, pre-shrunk 100% cotton t-shirt is what to wear when you want to go comfortably casual. Preshrunk, durable and guaranteed.

- 5.6 oz. 100% cotton
- Standard fit

### Other items by Signal Processing Gear:

y = phi x (Mug)

y = phi x (Large Mug)

y = phi x (Light T-Shirt)

# Experimental Design

$$y \;=\; Ax \;+\; w$$

experimental design: how to design $A$ ?

**Constraints:**

- sample budget: $A$ is $m \times n$ with $m \leq k < n$
- precision budget: $\|A\|_F^2 \leq$ Constant

**Sequential Design:** how to chose $A_1, \ldots, A_k$ to max prob of identifying $x$?

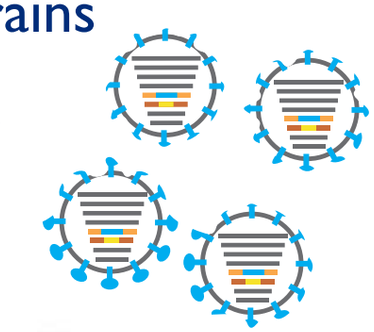
$$\begin{aligned} y_1 &= A_1 x + w_1 \\ y_2 &= A_2 x + w_2 \\ &\vdots \\ y_k &= A_k x + w_k \end{aligned}$$

# Application: Inferring Biological Pathways

virus

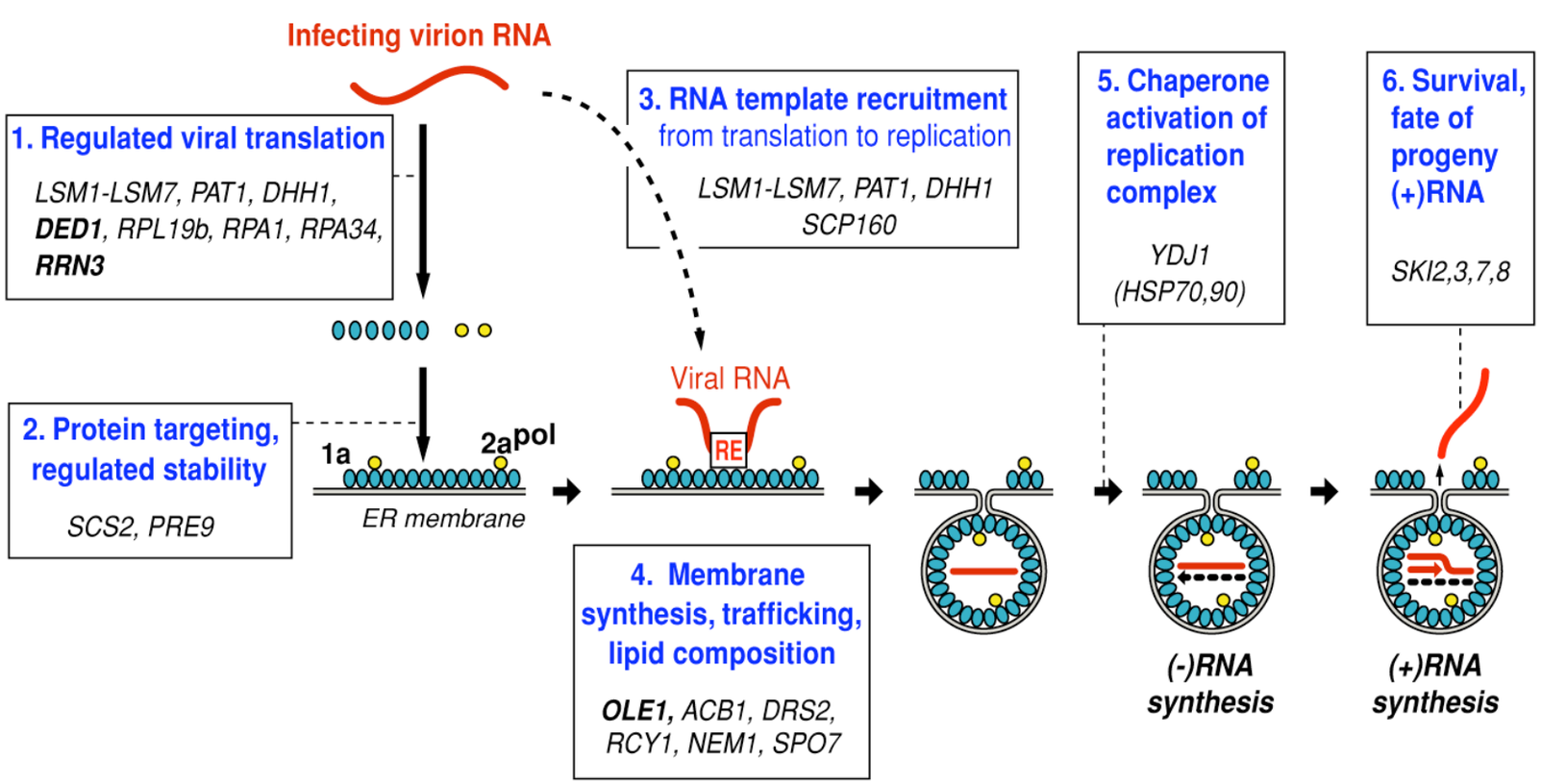
fruit fly

13,071 single-gene knock-down cell strains



infect each strain with fluorescing virus

microwell array




approximately 100 significant genes/protiens discovered by Ahlquist and Kawaoka Labs at UW-Madison

# Challenge: High-Dimensionality and Low SNR

nature Vol 454|14 August 2008|doi:10.1038/nature07151

## *Drosophila* RNAi screen identifies host genes important for influenza virus replication

Linhui Hao[1,2]*, Akira Sakurai[3]*†, Tokiko Watanabe[3], Ericka Sorensen[1], Chairul A. Nidom[5,6], Michael A. Newton[4], Paul Ahlquist[1,2] & Yoshihiro Kawaoka[3,7,8,9]

How do they confidently determine the ~100 out of 13K genes hijacked for virus replication from extremely noisy data?

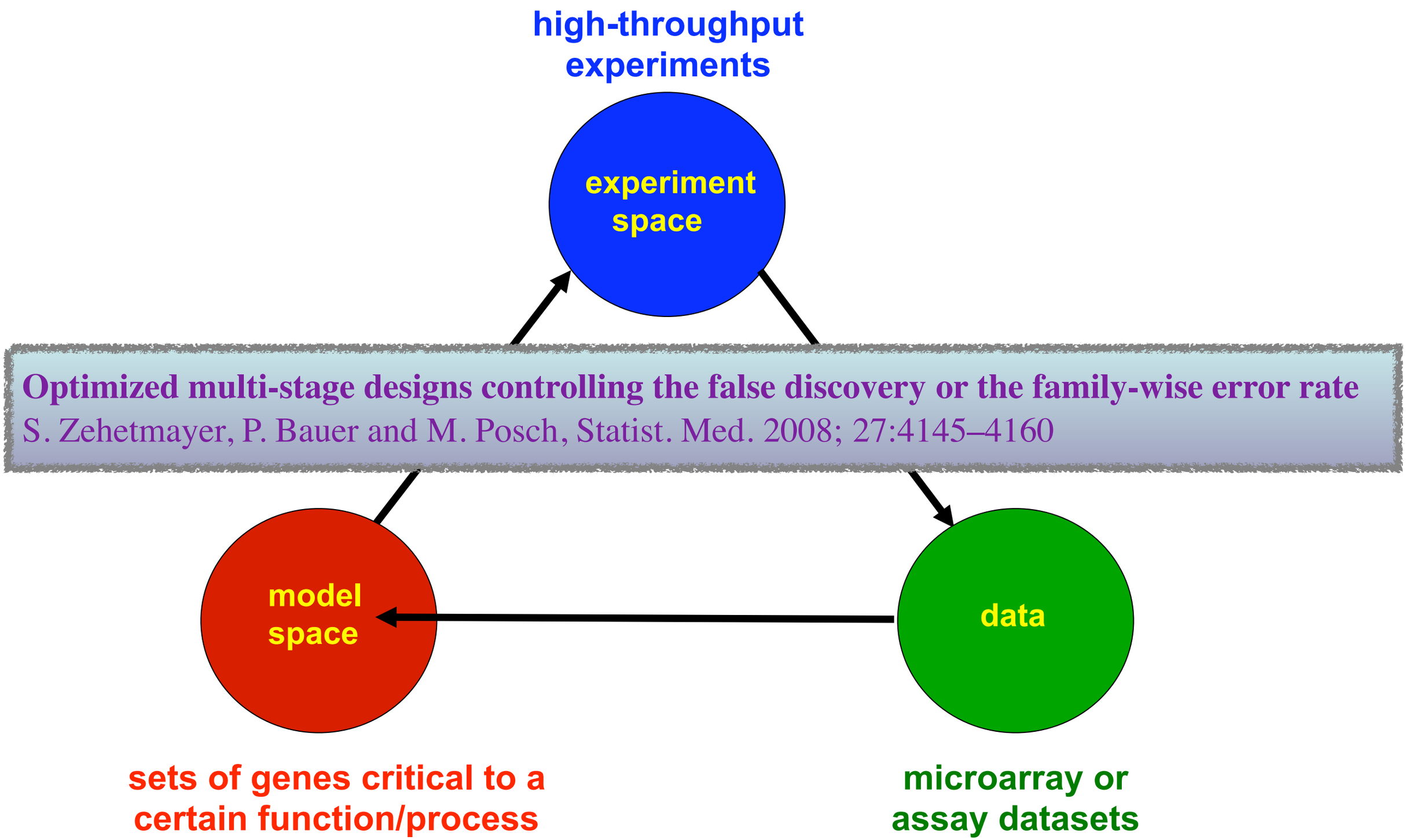
**Sequential Experimental Design:**

**Stage 1**: assay all 13K strains, twice; keep all with significant fluorescence in one or both assays for 2nd stage (13K $\rightarrow$ 1K)

**Stage 2**: assay remaining 1K strains, 6-12 times; retain only those with statistically significant fluorescence (1K $\rightarrow$ 100)

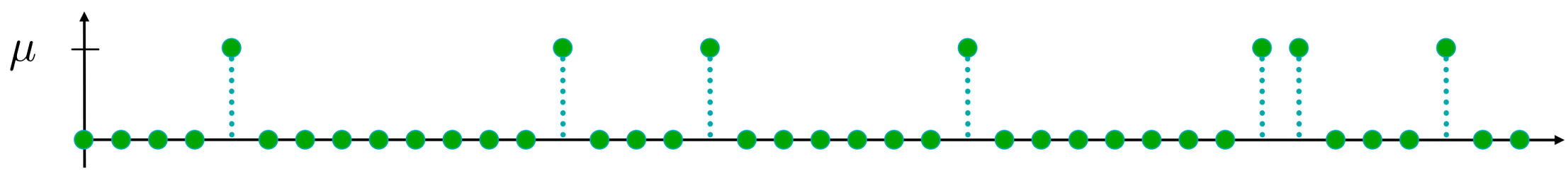
vastly more efficient that replicating all 13K experiments many times

# Feedback from Data Analysis to Data Collection



# Sparse Signal Model

Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be an unknown sparse vector;
most (or all) of its components $x_i$ are equal to zero.




$$x_i = \begin{cases} \mu > 0\,, & i \in \mathcal{S} \\ 0\,, & i \notin \mathcal{S} \end{cases}, \text{ where } |\mathcal{S}| \ll n$$

signal **support set**

deterministic
but unknown

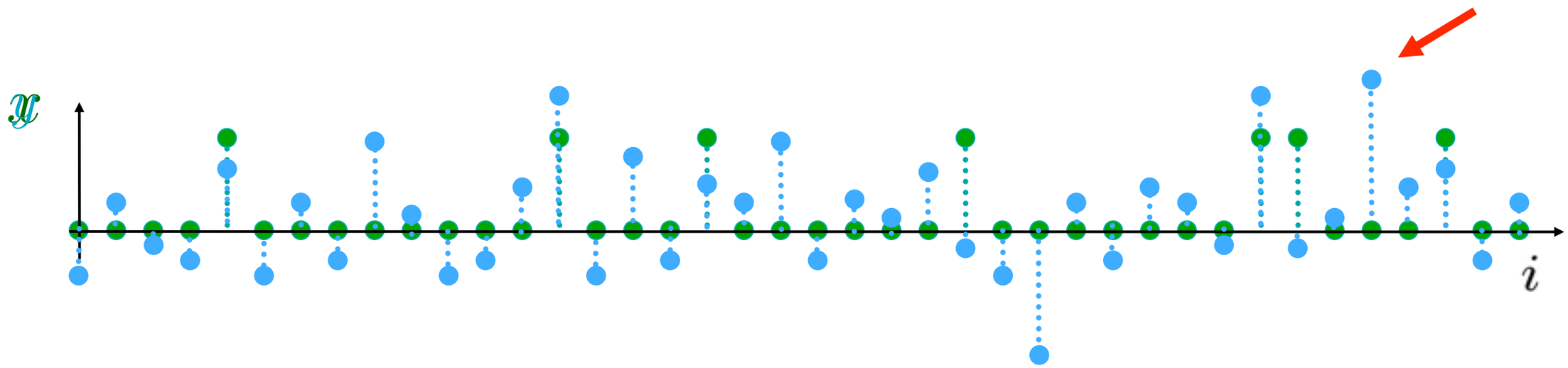
**Assume sublinear sparsity level:** $|\mathcal{S}| \ll n$

number of signal
components

# Noisy Observation Model

$$y_i \;=\; x_i \;+\; z_i\,,\;\; i = 1, \ldots, n$$

$$z_i \;\overset{\text{iid}}{\sim}\; \mathcal{N}(0,1)$$



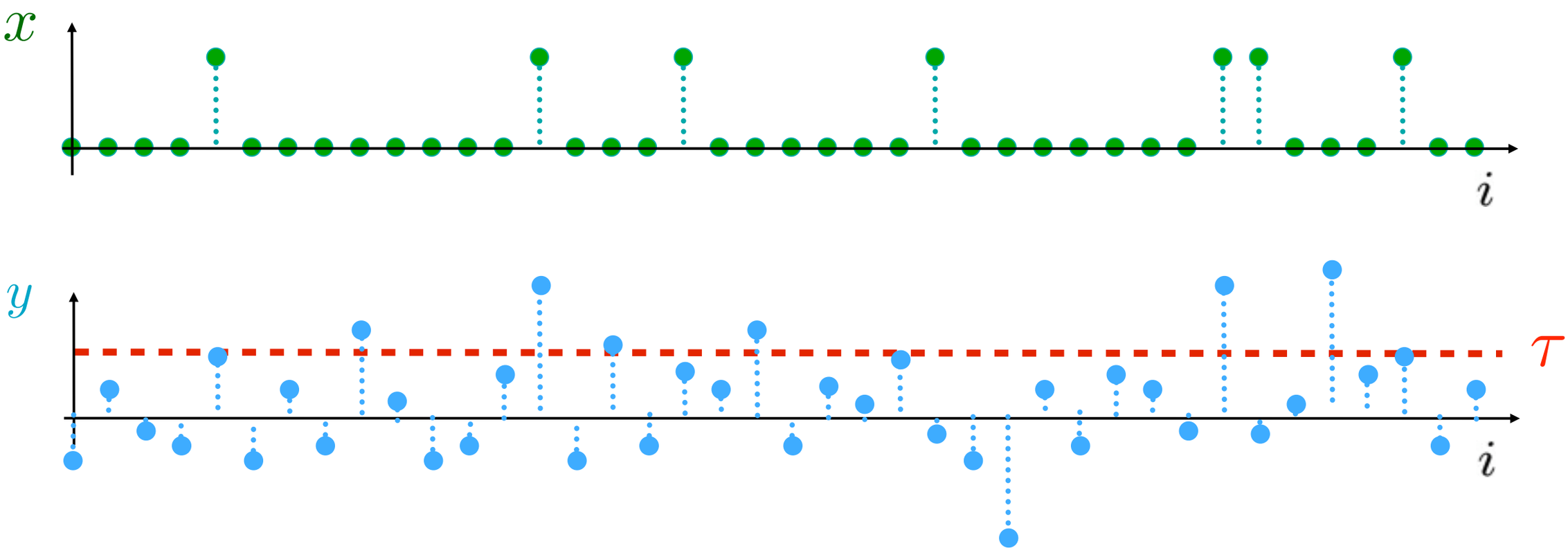

Suppose we want to locate just **one** signal component: $\widehat{i} \;=\; \arg\max_i y_i$

Even if no signal is present, $\max_i y_i \;\sim\; \sqrt{2\log n}$

**It is *impossible* to reliably detect signal components weaker than $O(\sqrt{\log n})$**

# Threshold Tests

Our goal is to estimate the set of non-zero components: $\mathcal{S} := \{i \,:\, x_i \neq 0\}$




**Definition 1** *A* threshold test *is an estimator of the form:*

$$\widehat{\mathcal{S}}_\tau(y) \;:=\; \{i \in \{1, \ldots, n\} : y_i \geq \tau > 0\}$$

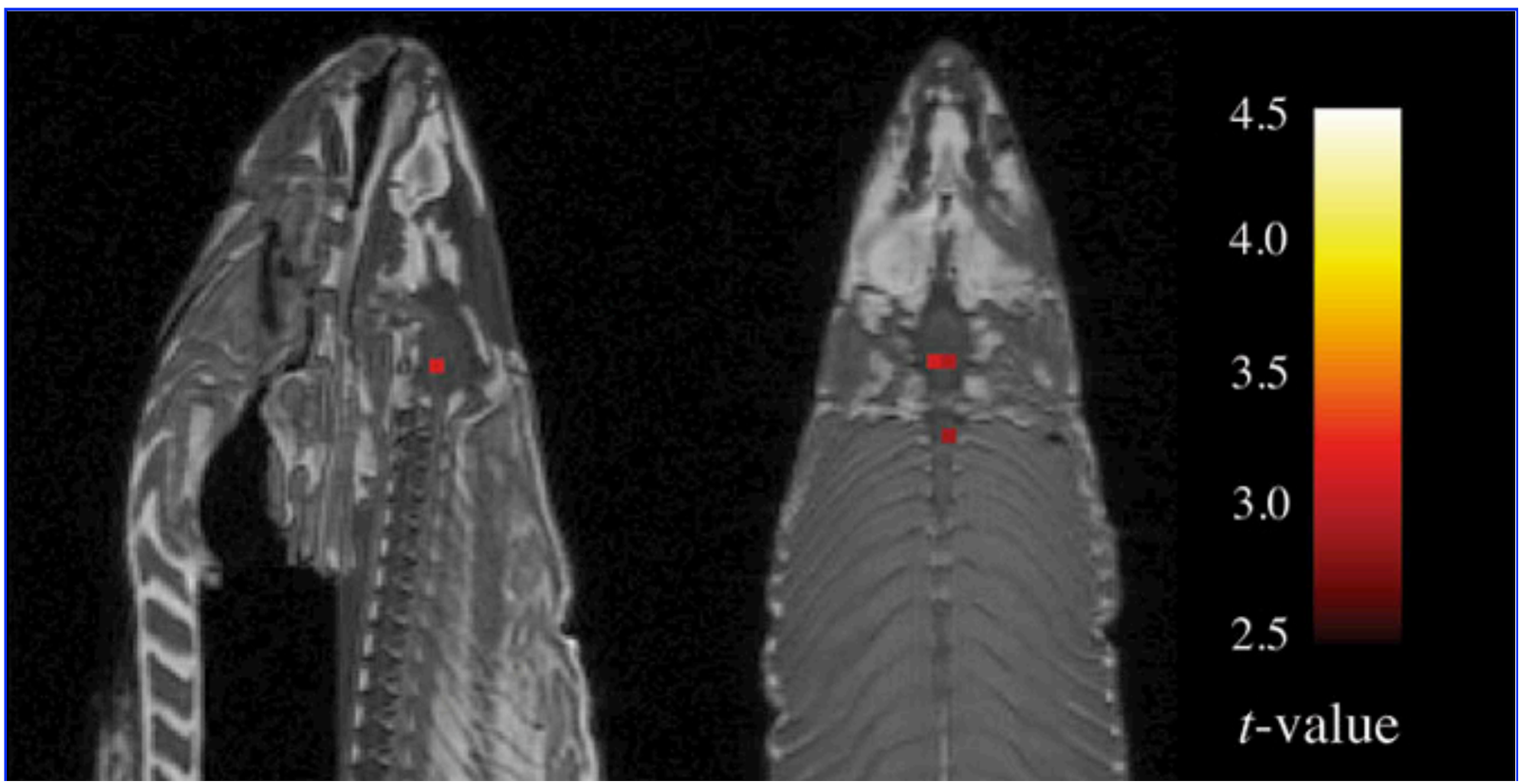
**Bonferroni Correction**: To keep the error level small (e.g., less than 5%) the threshold must be on the order of $\sqrt{\log n}$.

# Is there really a problem ?

Wired Science
News for Your Neurons
Previous post
Next post

## Scanning Dead Salmon in fMRI Machine Highlights Risk of Red Herrings

By Alexis Madrigal September 18, 2009 | 5:37 pm | Categories: Brains and Behavior



# An Alternative: Sequential Experimental Design

Instead of the usual non-adaptive observation model

$$y_i \;=\; x_i \;+\; z_i\,,\;\; i = 1, \ldots, n$$

suppose we are able to sequentially collect several independent measurements of each component of $x$, according to

$$y_{i,j} \;=\; x_i \;+\; \gamma_{i,j}^{-1/2}\, z_{i,j}, \quad i = 1, \ldots, n, \;\; j = 1, \ldots, k$$

where

$j$ indexes the measurement steps

$k$ denotes the total number of steps

$z_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$

$\gamma_{i,j} \geq 0$ controls the precision of each measurement

**Total precision budget is constrained, but the choice of $\gamma_{i,j}$ can depend on past observations $\{y_{i,\ell}\}_{\ell<j}$.**

# Experimental (Precision) Budget

sequential measurement model

$$y_{i,j} \;=\; x_i \;+\; \gamma_{i,j}^{-1/2}\, z_{i,j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k$$

The precision parameters $\{\gamma_{i,j}\}$ are required to satisfy

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \gamma_{i,j} \;\leq\; n$$

For example, the usual non-adaptive, single measurement model corresponds to taking $k = 1$, and $\gamma_{i,1} = 1$, $i = 1, \ldots, n$. This baseline can be compared with adaptive procedures by allowing $k > 1$ and variable $\{\gamma_{i,j}\}$ satisfying budget.

**Precision parameters control the SNR per component.**

SNR is increased/decreased by

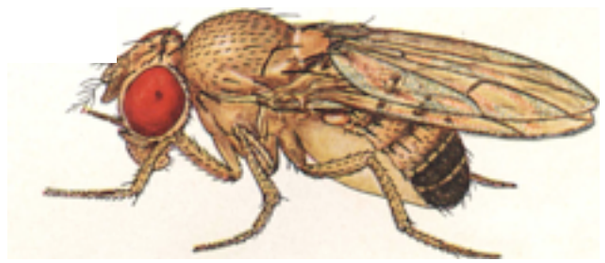
— more/fewer repeated samples or

— longer/shorter observation times

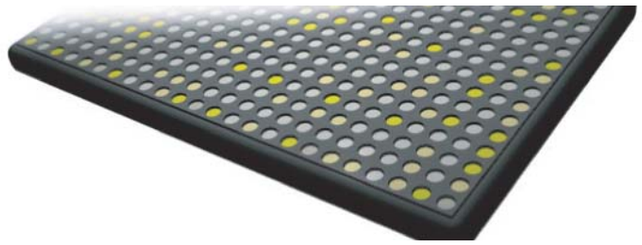
# Fruit Fly Example

virus

fruit fly



assay



How to find genes involved in virus replication ?

**Sequential Design Idea**

Budget: **k assays**, n tests/assay

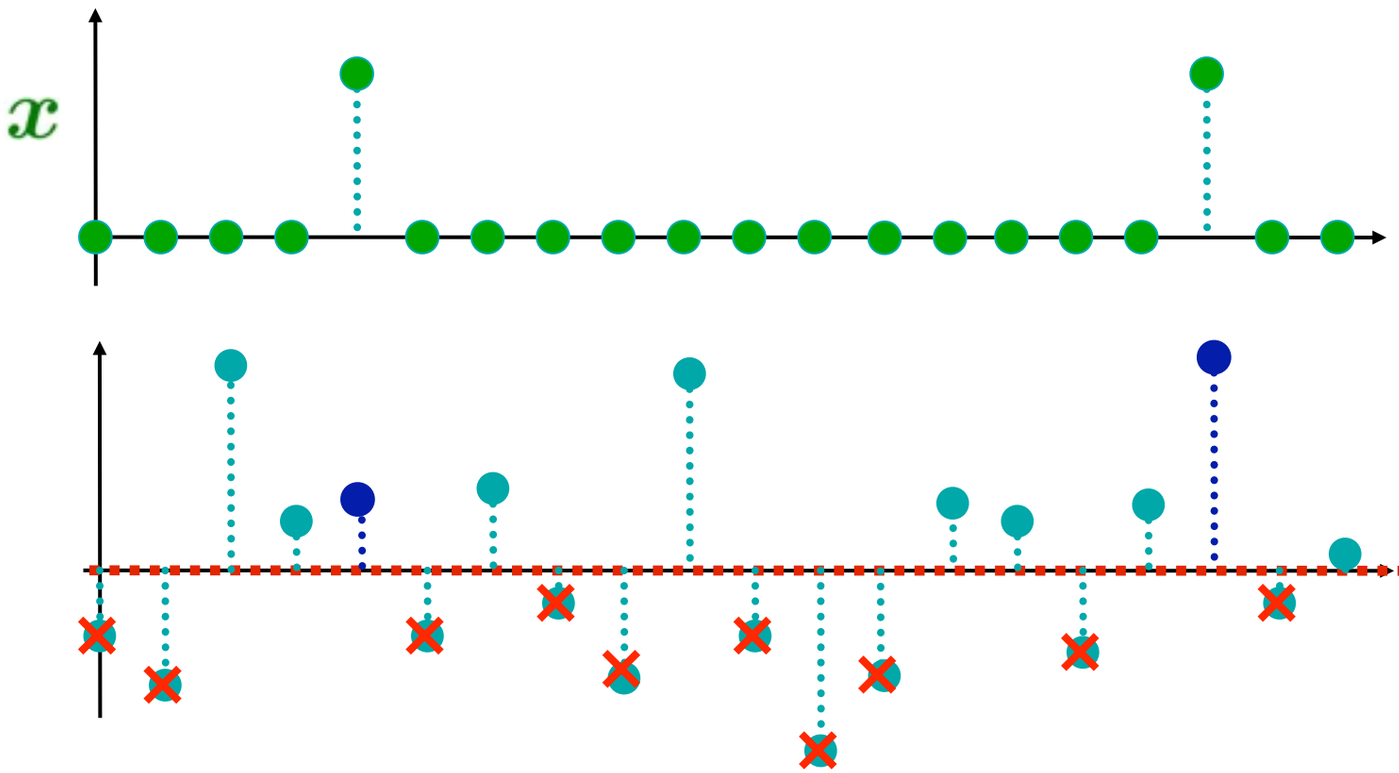
**Assay 1**: measure fluorescence of all n genes; discard n/2 genes with weakest fluorescence.

**Assay 2**: measure fluorescence for remaining n/2 genes, each tested twice (double SNR); discard n/4 genes with weakest fluorescence.

**Assay 3**: measure fluorescence for remaining n/4 genes, each tested four times (quadruple SNR); discard n/8 genes with weakest fluorescence.

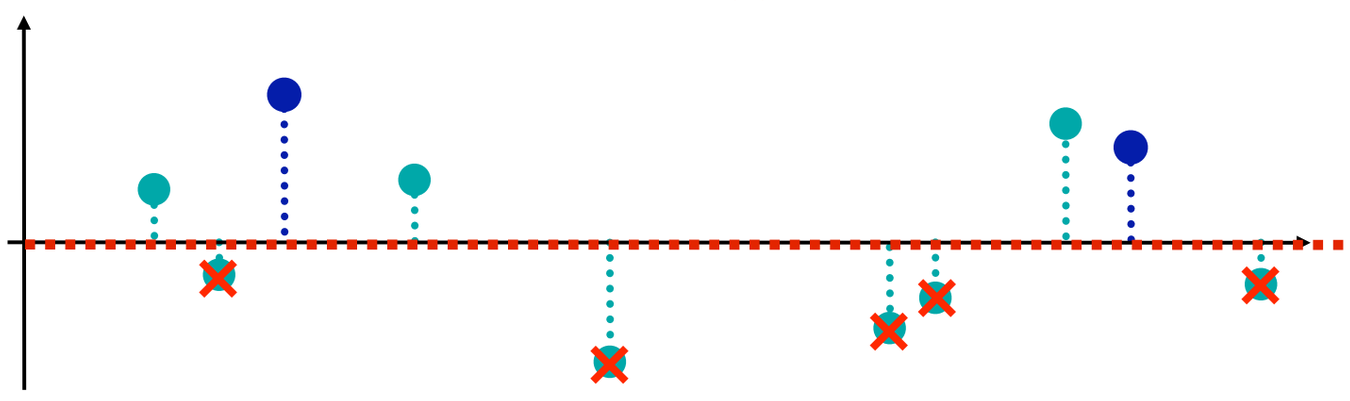
continue "distilling"....

# Idealized Example




Take $k = 3$ steps and split precision budget uniformly ($n/3$ per step)

$$y_{i,1} = x_i + \mathcal{N}(0, 3)$$



$$y_{i,2} = x_i + \mathcal{N}\left(0, \frac{3}{2}\right)$$

$$y_{i,3} = x_i + \mathcal{N}\left(0, \frac{3}{4}\right)$$

# Distilled Sensing

***Simple* Distilled Sensing**

initialize: $\mathcal{S}_0 = \{1, \ldots, n\}$, $\gamma_{i,j}^{-1} = 2 + \epsilon$, $\epsilon > 0$

for $j = 1, \ldots, k$

1) measure: $y_{i,j} \sim \mathcal{N}(x_i, 2 + \epsilon)$ , $i \in \mathcal{S}_{j-1}$

2) threshold: $\mathcal{S}_j = \{i : y_{i,j} \geq 0\}$

end

output: $\mathcal{S}_k = \{i : y_{i,k} > 0\}$

total precision budget: $\mathbb{E}\left[\sum_{i,j} \gamma_{i,j}\right]$

$$
\begin{aligned}
&= \frac{1}{2+\epsilon} \sum_{j=1}^{k} \mathbb{E}|\mathcal{S}_{j-1}| \\
&\leq \frac{1}{2+\epsilon} \sum_{j=1}^{k} \left( \frac{n - |\mathcal{S}|}{2^{j-1}} + |\mathcal{S}| \right) \\
&\leq \frac{2(n - |\mathcal{S}|)}{2+\epsilon} + k|\mathcal{S}| \leq n
\end{aligned}
$$

(for $n$ large)

probability of error:

$$
\begin{aligned}
\mathbb{P}(\mathcal{S}_k \neq \mathcal{S}) &= \mathbb{P}\left(\{\mathcal{S}^c \cap \mathcal{S}_k \neq \emptyset\} \cup \{\mathcal{S} \cap \mathcal{S}_k^c \neq \emptyset\}\right) \\
&\leq \mathbb{P}\left(\mathcal{S}^c \cap \mathcal{S}_k \neq \emptyset\right) + \mathbb{P}\left(\mathcal{S} \cap \mathcal{S}_k^c \neq \emptyset\right)
\end{aligned}
$$

# False Positives

$$\mathbb{P}(\mathcal{S}_k \neq \mathcal{S}) \quad \leq \quad \mathbb{P}\left(\mathcal{S}^c \cap \mathcal{S}_k \neq \emptyset\right) \; + \; \mathbb{P}\left(\mathcal{S} \cap \mathcal{S}_k^c \neq \emptyset\right)$$

$$\begin{aligned}
\mathbb{P}\left(\mathcal{S}^c \cap \mathcal{S}_k \neq \emptyset\right) \quad &= \quad \mathbb{P}\left(\bigcup_{i \notin \mathcal{S}} \bigcap_{j=1}^{k} y_{i,j} > 0\right) \\
&\leq \quad \sum_{i \notin \mathcal{S}} \mathbb{P}\left(\bigcap_{j=1}^{k} y_{i,j} > 0\right) \\
&= \quad \sum_{i \notin \mathcal{S}} 2^{-k} \;=\; \frac{n-s}{2^k}
\end{aligned}$$

# False Negatives

$$\mathbb{P}(\mathcal{S}_k \neq \mathcal{S}) \quad \leq \quad \mathbb{P}\left(\mathcal{S}^c \cap \mathcal{S}_k \neq \emptyset\right) \; + \; \mathbb{P}\left(\mathcal{S} \cap \mathcal{S}_k^c \neq \emptyset\right)$$

$$\begin{aligned} \mathbb{P}\left(\mathcal{S} \cap \mathcal{S}_k^c \neq \emptyset\right) \quad &= \quad \mathbb{P}\left( \bigcup_{j=1}^{k} \bigcup_{i \in \mathcal{S}} y_{i,j} < 0 \right) \\ &\leq \quad \frac{k|\mathcal{S}|}{2} \, \exp\left( -\frac{\mu^2}{2(2+\epsilon)} \right) \end{aligned}$$

# Probability of Error Bound

$$
\begin{aligned}
\mathbb{P}(\mathcal{S}_k \neq \mathcal{S}) &\leq \mathbb{P}\left(\mathcal{S}^c \cap \mathcal{S}_k \neq \emptyset\right) + \mathbb{P}\left(\mathcal{S} \cap \mathcal{S}_k^c \neq \emptyset\right) \\
&\leq \frac{n-s}{2^k} + \frac{k|\mathcal{S}|}{2} \exp\left(-\frac{\mu^2}{2(2+\epsilon)}\right) \\
&= \frac{n-s}{2^k} + \frac{1}{2} \exp\left(-\frac{(\mu^2 - 2(2+\epsilon)\log(k|\mathcal{S}|))}{2(2+\epsilon)}\right)
\end{aligned}
$$

Consider high-dimensional limit as $n \to \infty$ and take $k = \log_2 n^{1+\epsilon}$

$$
\mathbb{P}(\mathcal{S}_k \neq \mathcal{S}) \leq \underbrace{\frac{n-s}{2^k}}_{\to 0} + \frac{1}{2} \exp\left(-\frac{(\mu^2 - 2(2+\epsilon)\log(|\mathcal{S}|(1+\epsilon)\log_2 n))}{2(2+\epsilon)}\right)
$$

Second term tends to zero if

$$
\mu \geq \sqrt{2(2+\epsilon)\log(|\mathcal{S}|(1+\epsilon)\log_2 n)}
$$

# Gains of Sequential Design

non-adaptive threshold:

$$\mu \;\geq\; \sqrt{2\log n}$$

DS threshold:

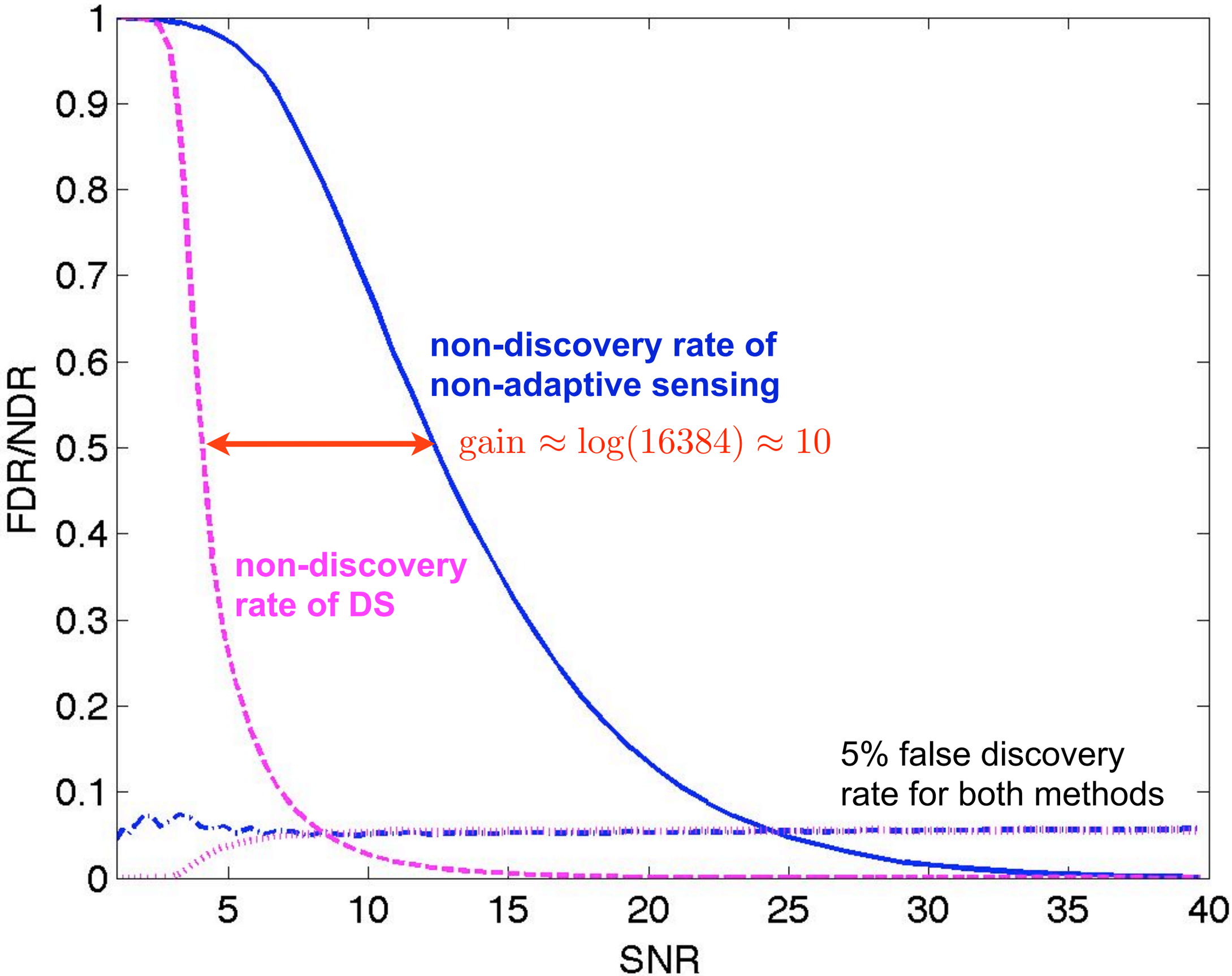
$$\begin{aligned}\mu &\geq \sqrt{2(2+\epsilon)\log(|\mathcal{S}|(1+\epsilon)\log_2 n)} \\ &\approx \sqrt{4\log|\mathcal{S}|}\end{aligned}$$

**We get a gain whenever $|\mathcal{S}| \preceq n^{1/2}$**

**Punchline:** In ultra-sparse setting, say $|\mathcal{S}| = C\log n$, DS drives error to zero if $\mu \geq \sqrt{(8+\epsilon)\log\log n}$, compared to the non-adaptive requirement $\mu \geq \sqrt{2\log n}$.

# Example

$n = 2^{14}, \ \|x\|_0 = \sqrt{n} = 128$



# Conclusions

## Sequential Experimental Designs for High-Dimensional Models

thresholds for recovery in high-dimensional limit:

| | |
|---|---|
| **non-adaptive designs** | $\text{SNR} \sim \log n$ |
| **sequential designs** | $\text{SNR} \sim$ arbitrarily slowly growing function of $n$ |

Distilled Sensing: Adaptive Sampling for Sparse Detection and Estimation
J. Haupt, R. Castro, and RN, **arXiv:1001.5311v2**

## Geometry of Sequential Inference

number of membership queries required to learn a set to $\epsilon$ accuracy:

| | |
|---|---|
| **non-adaptive** | $\#$ queries $\sim 1/\epsilon$ |
| **adaptive** | $\#$ queries $\sim \log(1/\epsilon)$ |

The Geometry of Generalized Binary Search, RN, **arXiv:0910.4397**