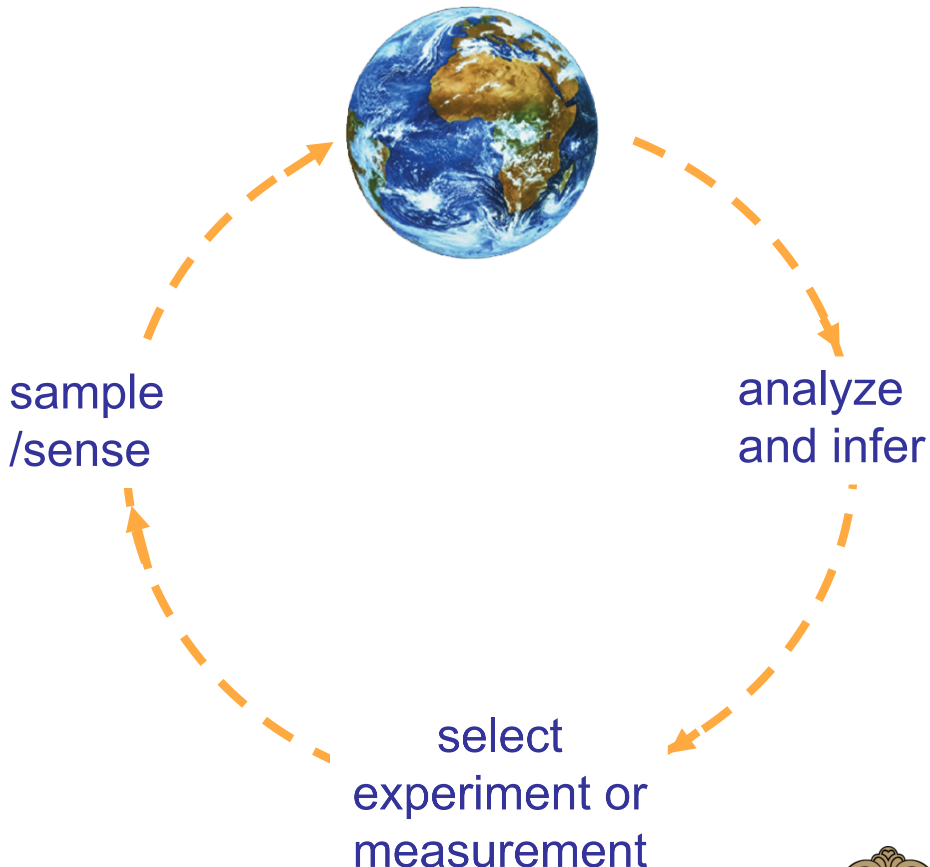
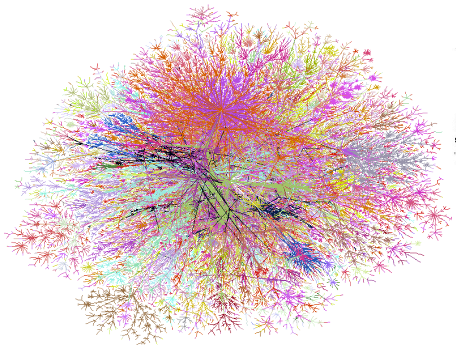
Feedback in Information and Decision Systems

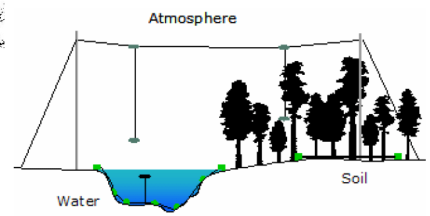


(please see notes below each slide for more information)

THE UNIVERSITY of WISCONSIN MADISON
Rob Nowak
www.ece.wisc.edu/~nowak

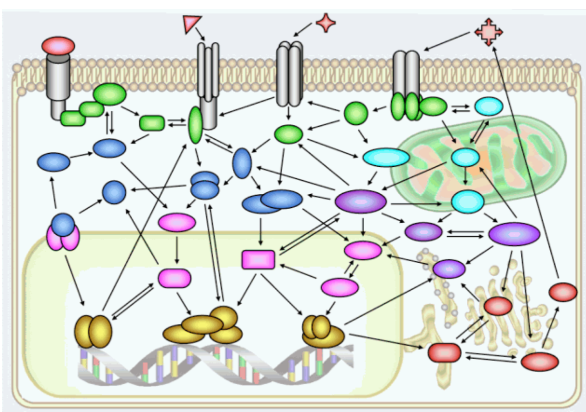
Grand Challenge:
Understanding Large Networked Systems



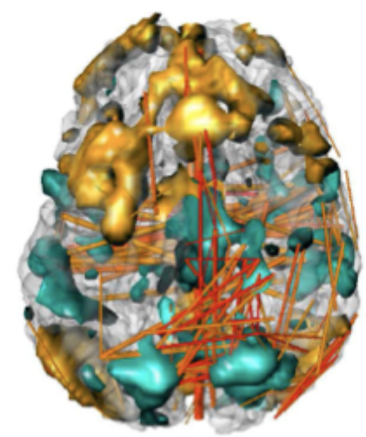


Technological Networks
(Internet Mapping Project, US power grid, UCLA CENS)

Social Networks



Biological Networks
(JMDBase)



Brain Networks
(Worsley et al, 2005)

Challenges:

- Inferring structure & function of the network
- Optimized design & resource allocation
- Pattern analysis & anomaly detection

Challenges for Information and Decision Systems

Measurement

impossible to measure/observe large networks everywhere and all the time

☞ incomplete/missing and indirect data is the norm

Uncertainty

experiments/measurements are noisy, corrupted or otherwise unreliable

☞ info-processing and decision-making must be robust to uncertainty

Diversity

data from disparate sources

☞ integration of info from sensors, experiments, databases, human intell, etc.

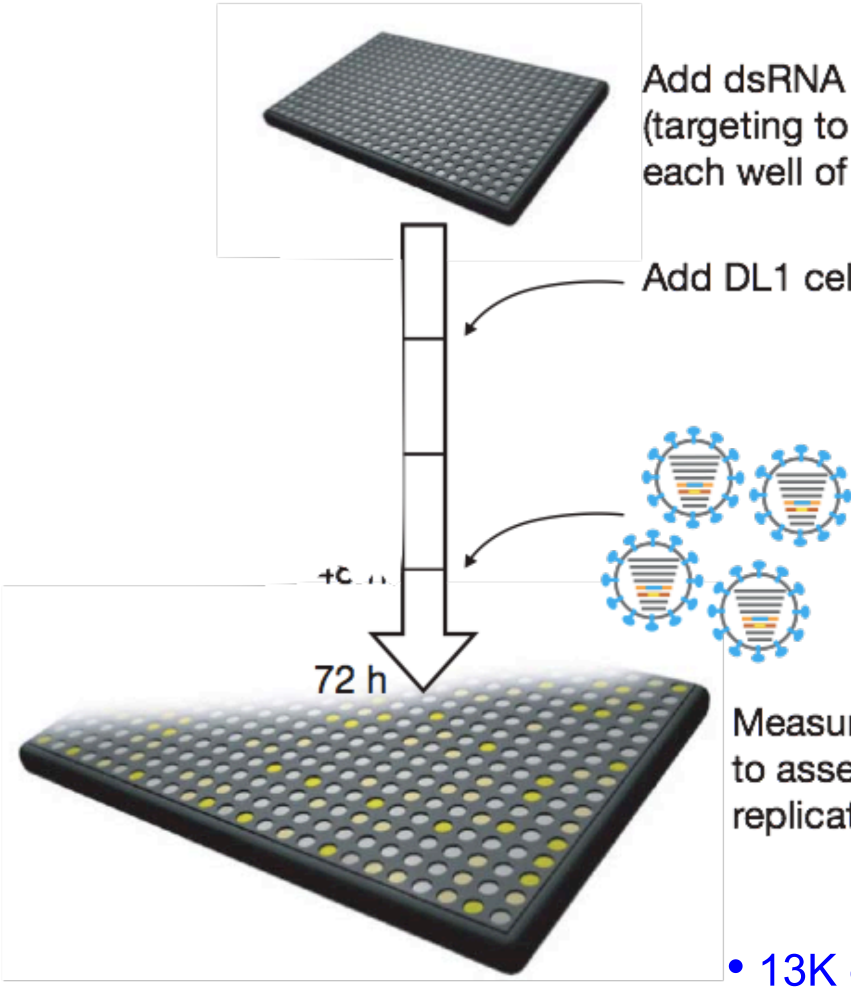
Complexity

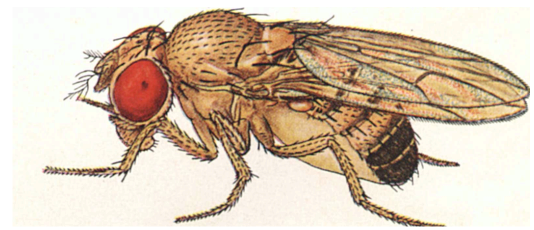
networked systems can be ultra-high dimensional

☞ modeling/approximation is formidable, mathematically & computationally

Motivation: Systems Biology
(Ahlquist Lab, UW-Madison)

Drosophila RNAi screen identifies host genes important for influenza virus replication. Hao et al, Nature 2008.





- 13K genes, only a few involved in virus replication
- very noisy data
- determine virus replication network/pathways

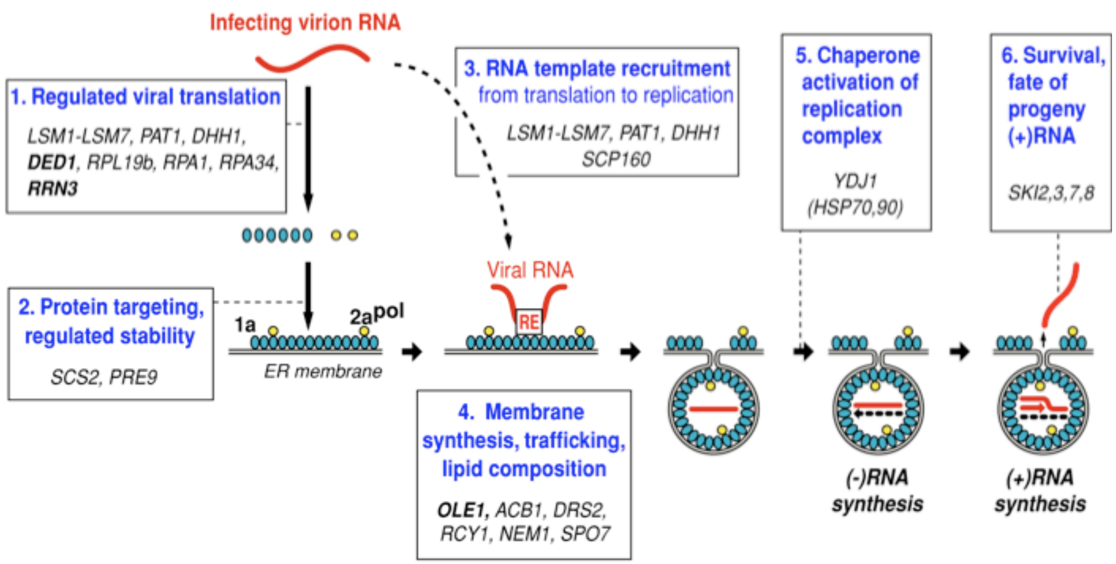
Motivation: Systems Biology

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

Multistage Adaptive Experiments:

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

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



Next step: test m-fold gene deletion strains... $\binom{13000}{2} \approx 85,000,000$!

Feedback in Info-Processing and Decision-Making

Given: A collection of models or hypotheses $\mathcal{F}$.

Goal: Identify correct/best model $f^* \in \mathcal{F}$ from experimental data.

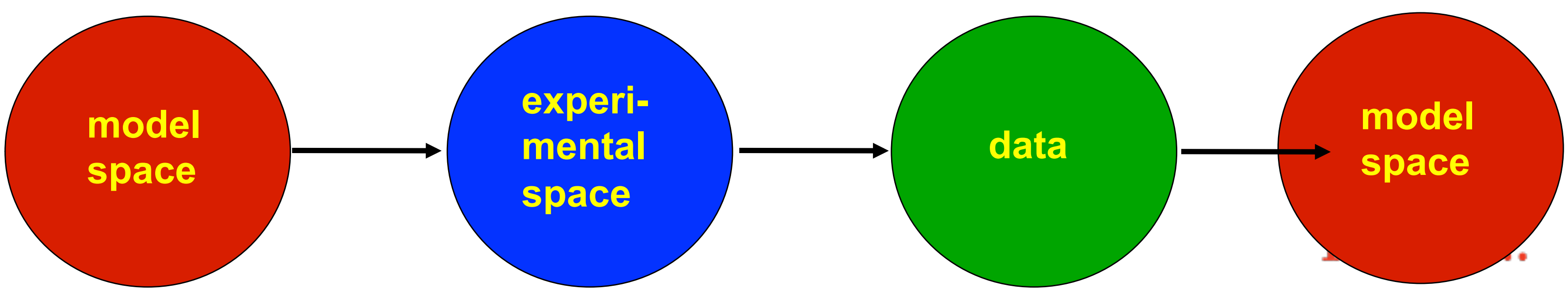
Optimize: Minimize number of experiments/measurements.

Let $N := N(\epsilon, \mathcal{F}, P)$ be the ϵ-covering number of $\mathcal{F}$ wrt measure P.
Suppose each measurement/experiment yields K bits.

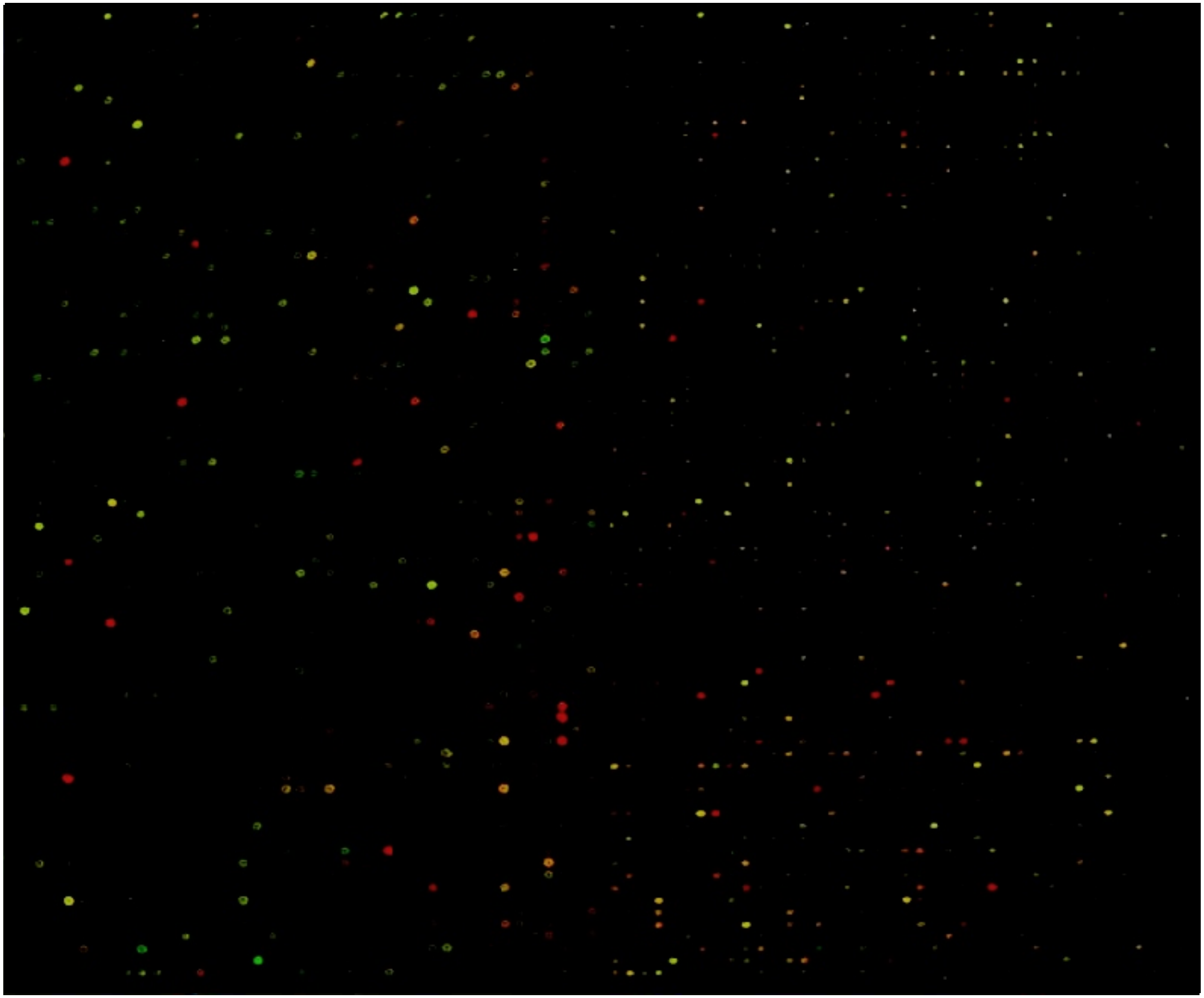
Is it possible to learn an ϵ-good approximation to f^* from $O\left(K^{-1} \log N\right)$ experiments/measurements?

"Active Learning Using Arbitrary Binary Valued Queries," Kulkarni, Mitter, Tsitsiklis, Machine Learning, 11, 23-35 (1993)

In general, many more experiments are required, because some bits in each experiment are redundant. But sometimes, sequentially adaptive experiments or measurements can identify a near-optimal f in $O(K^{-1} \log N)$ steps.



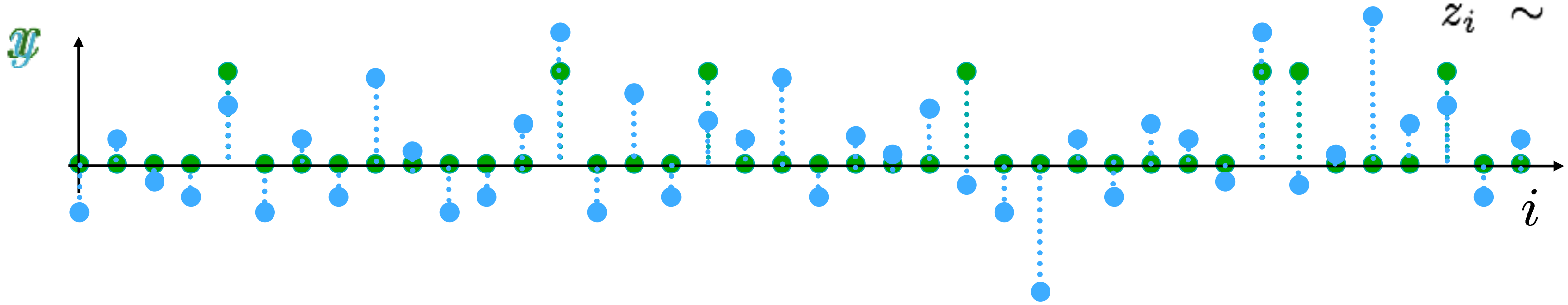
Sequential Adaptive Experimentation



Quantifying Gains from Feedback

$$y_i = x_i + z_i\,,\ i = 1, \ldots, N$$

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



Let $\mathcal{S} := \{i = 1, \ldots, N \ :\ x_i \neq 0\}$. Assume sparsity: $|\mathcal{S}| \ll N$.

$\widehat{\mathcal{S}}(y)$ denotes an estimator of $\mathcal{S}$.

$$\text{FDP}(\widehat{\mathcal{S}}) := \frac{|\widehat{\mathcal{S}} \backslash \mathcal{S}|}{|\widehat{\mathcal{S}}|} = \frac{\#\text{ falsely discovered components}}{\text{total }\#\text{ discovered components}}$$

$$\text{NDP}(\widehat{\mathcal{S}}) := \frac{|\mathcal{S} \backslash \widehat{\mathcal{S}}|}{|\mathcal{S}|} = \frac{\#\text{ missed components}}{\#\text{ true non-zero components}}$$

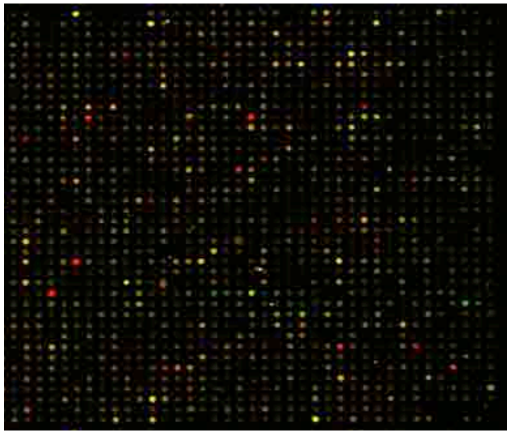
To guarantee $\text{FDP}(\widehat{\mathcal{S}}) \xrightarrow{P} 0\,,\ \text{NDP}(\widehat{\mathcal{S}}) \xrightarrow{P} 0$ as $N \to \infty$, we require

$$\text{SNR} \sim \log N\ ,\ \text{using non-adaptive measurements}$$

$$\text{SNR} \sim \log \ldots \log N\ ,\ \text{using adaptive measurements}$$

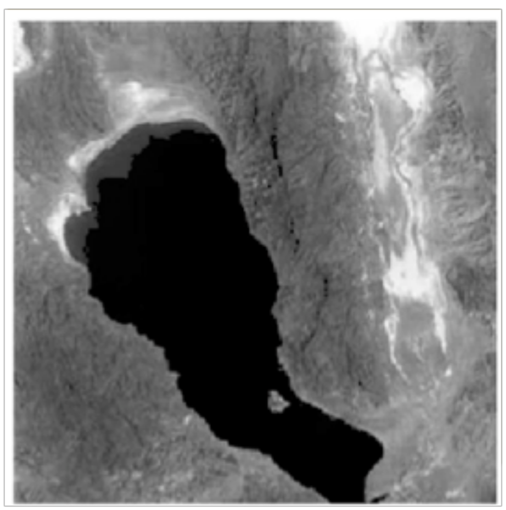
J. Haupt, R. Castro and RN, “Distilled Sensing: Selective Sampling for Sparse Signal Recovery,” AISTATS 2009

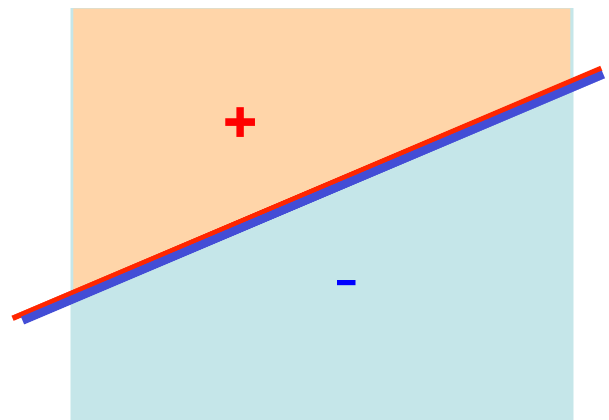
The Power of Feedback in Sensing and Data Selection



Sparse Signal Recovery:
non-adaptive sensing $\Rightarrow$ SNR $\quad O(\log n)$
adaptive sensing $\Rightarrow$ SNR $\sim O(\log \cdots \log n)$

Image Reconstruction:
uniform sampling n points $\Rightarrow$ error $\sim O(n^{-1/2})$
adaptive sampling n points $\Rightarrow$ error $\sim O(n^{-2})$





Classification:
passive learning $\Rightarrow$ sample complexity $n \sim 1/\mathrm{err}$
active learning $\Rightarrow$ sample complexity $n \sim \log(1/\mathrm{err})$

Scientific and Engineering Discovery is a Closed-Loop Process

Do we have the right theory and methods for it ?

Paths forward:

- Closing the loop between data acquisition and analysis
- Do ‘more with less’ or ‘less with more’ data
  (sublinear complexity algorithms)
- Integrating disparate information sources (including humans)
- Man-machine systems

more information: www.ece.wisc.edu/~nowak