Active Search with Complex Models: Graphs, Sparsity, and Pattern Finding

Yifei Ma

PhD Student, advised by Jeff Schneider
Machine Learning Department
School of Computer Science
Carnegie Mellon University

Active Search

Like Beer Tasting

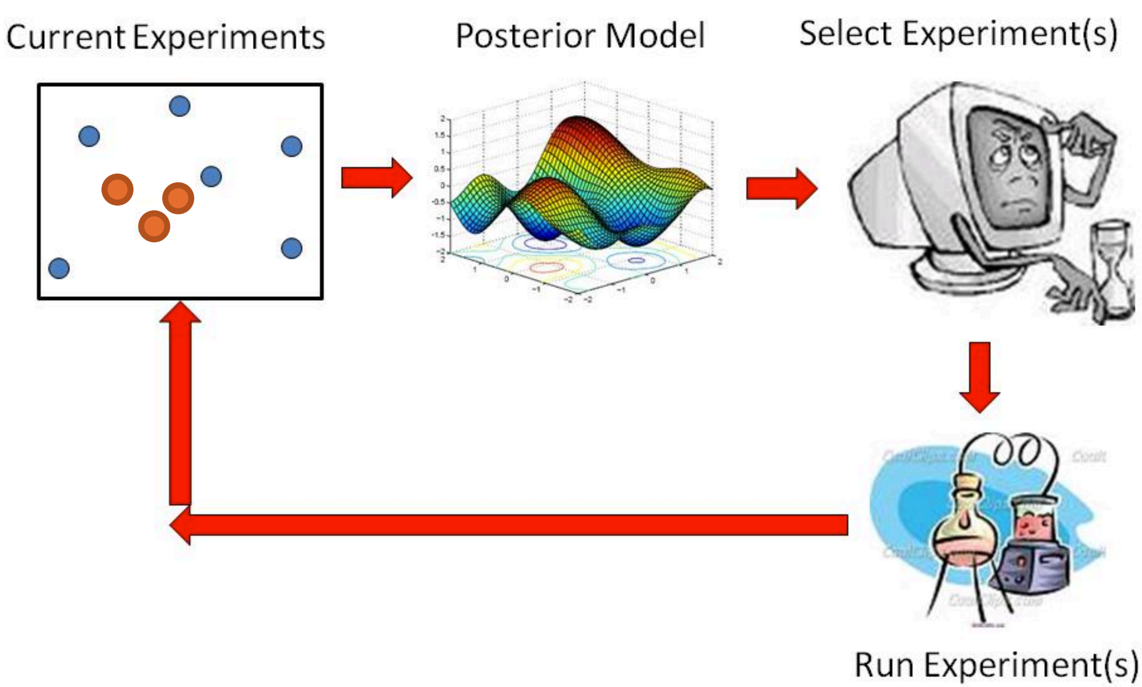
Pictures in [Audibert and Munos 2011]

Active Search Common Paradigm

Assume: A pool of unlabeled data

Goal: find all positive instances quickly

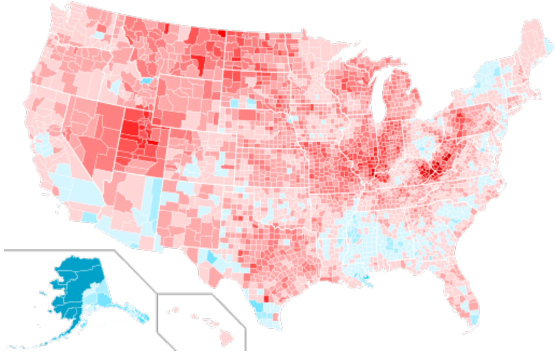
Action: present instances and get labels



Artwork by [Javad Azimi]

Applications

Application	Active Search Allows
Product Recommendation	New Users w/ Little Purchase History
Information Retrieval	Relevant but Underspecified Results
Environmental Monitoring	All Polluted Areas
Opinion Polling	All Winning/Swing States
Hazard/Survivor Search	Localize All Signal Hotspots



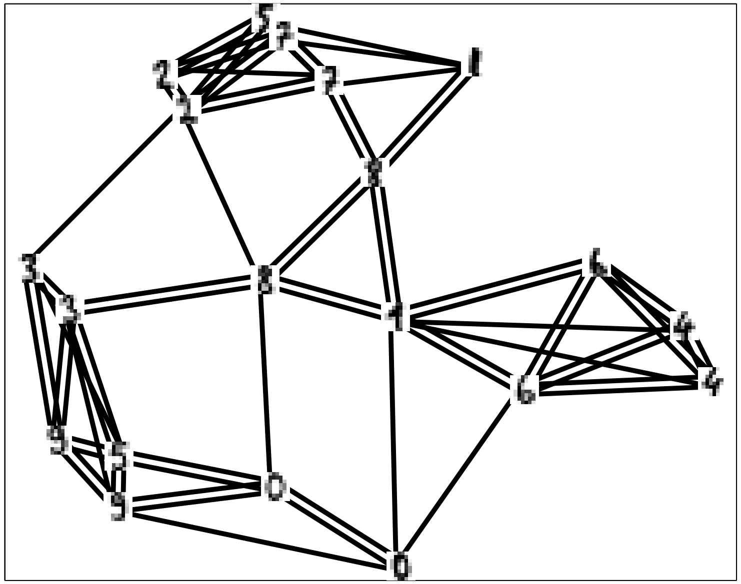
Outline

App	Challenge	Previous state-of-the-art	Contribution	Papers
Rec / Retrieval	Similarity features	Linear models	Graphs	NIPS 2013; UAI 2015
Monitoring / Polling	Reward defined by a group of points	Point rewards	Group rewards	AISTATS 2014; 2015
Surveillance	Sparse signal	Point measurements	Aggregate measurements	AAAI 2017

Idea 1: Active Search on Graphs

Graphs can represent complex information

- High-dim sparse features, links, hierarchical structures.



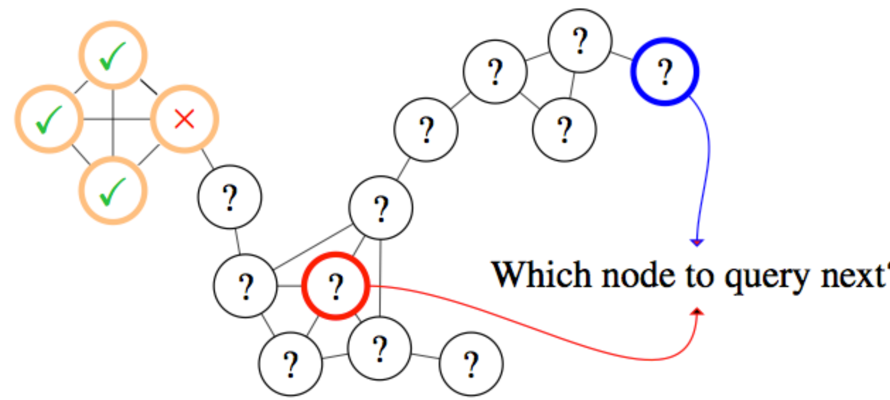
Nearest-neighbor graphs using raw image pixels

Problem Definition

Assume: known graph; unknown labels

Task: find all  nodes using the fewest label queries

Question: which nodes to query?



Task breakdown:

Exploration (learning): reduce model uncertainty [NIPS 2013]

Exploitation (search): find all positives [UAI 2015]

Good Exploration Similar to Experimental Designs

Optimal Design [Gergonne, J. D. 1815]

Design experiments to optimize some criterion (e.g. variance, entropy)

Blind of actual observations

Eg. regression

$y_i = x_i^T\beta$

design x_i, observe y_i, learn β?

D-optimality

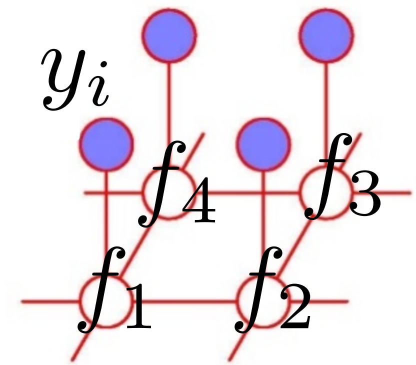
V-optimality

Σ-optimality – Our contribution #1

Kernel regression/Gaussian process

[By NOAA (http://www.photolib.noaa.gov/htmls/theb1982.htm) [Public domain], via Wikimedia Commons]

Gaussian Random Fields [Zhu 2004]



A Bayesian model for label propagation

Prior $$E(\mathbf{f}) = \frac{1}{2}\sum_{i\sim j}(f_i - f_j)^2 = \frac{1}{2}\mathbf{f}^\top L\mathbf{f}$$

$$p(\mathbf{f}) \sim \mathcal{N}(0, L^{-1})$$

$\mathbf{f}=(f_1,\ldots,f_n)^T$: node values.

L=D-A : graph Laplacian.

$$L = \begin{bmatrix} 4 & -1 & & -1 \\ -1 & 4 & -1 & \\ & -1 & 4 & -1 \\ -1 & & -1 & 4 \end{bmatrix}$$

Observe $f_s=y_s$, posterior is Gaussian with

Mean values $$\hat{\mathbf{f}}_{u|s} = D_{uu}^{-1}(A_{us}\mathbf{y}_s + A_{uu}\hat{\mathbf{f}}_{u|s})$$

and covariance matrix

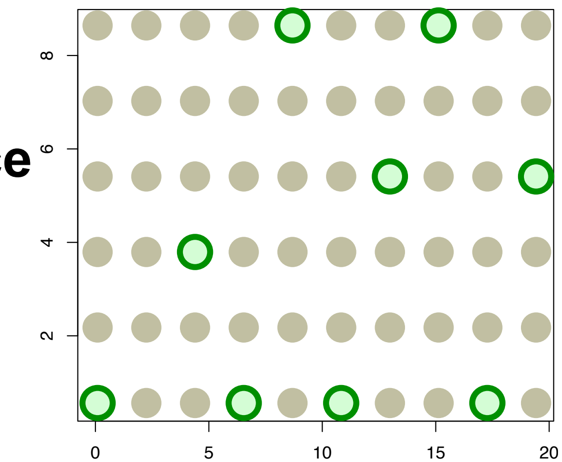
$$C_{(s)} = \begin{pmatrix} 0 & & \\ & (L_{uu})^{-1} & \\ & & 0 \end{pmatrix}$$

Baseline 1: D-Optimality

Minimize posterior differential entropy

$$\min_s \det(C_{(s)}) = \det\left((L_u)^{-1}\right)$$

Greedy application maximizes marginal variance



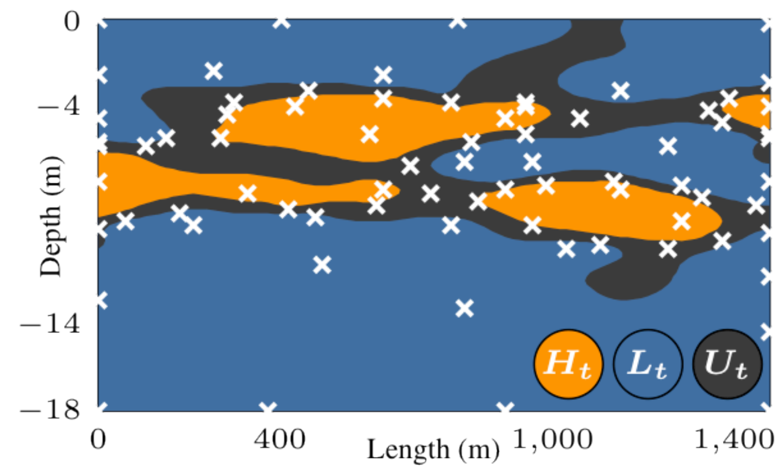
$$I_{(s)}(f;y_i) \simeq \log(1 + C_{(s)}(i,i)/\sigma_n^2)$$

Near-optimal sensor placement [Krause 2008]

GP-Bandit [Srinivas 2010]

Level set estimation [Gotovos 2013]

Bandits on graphs [Valko 2014]



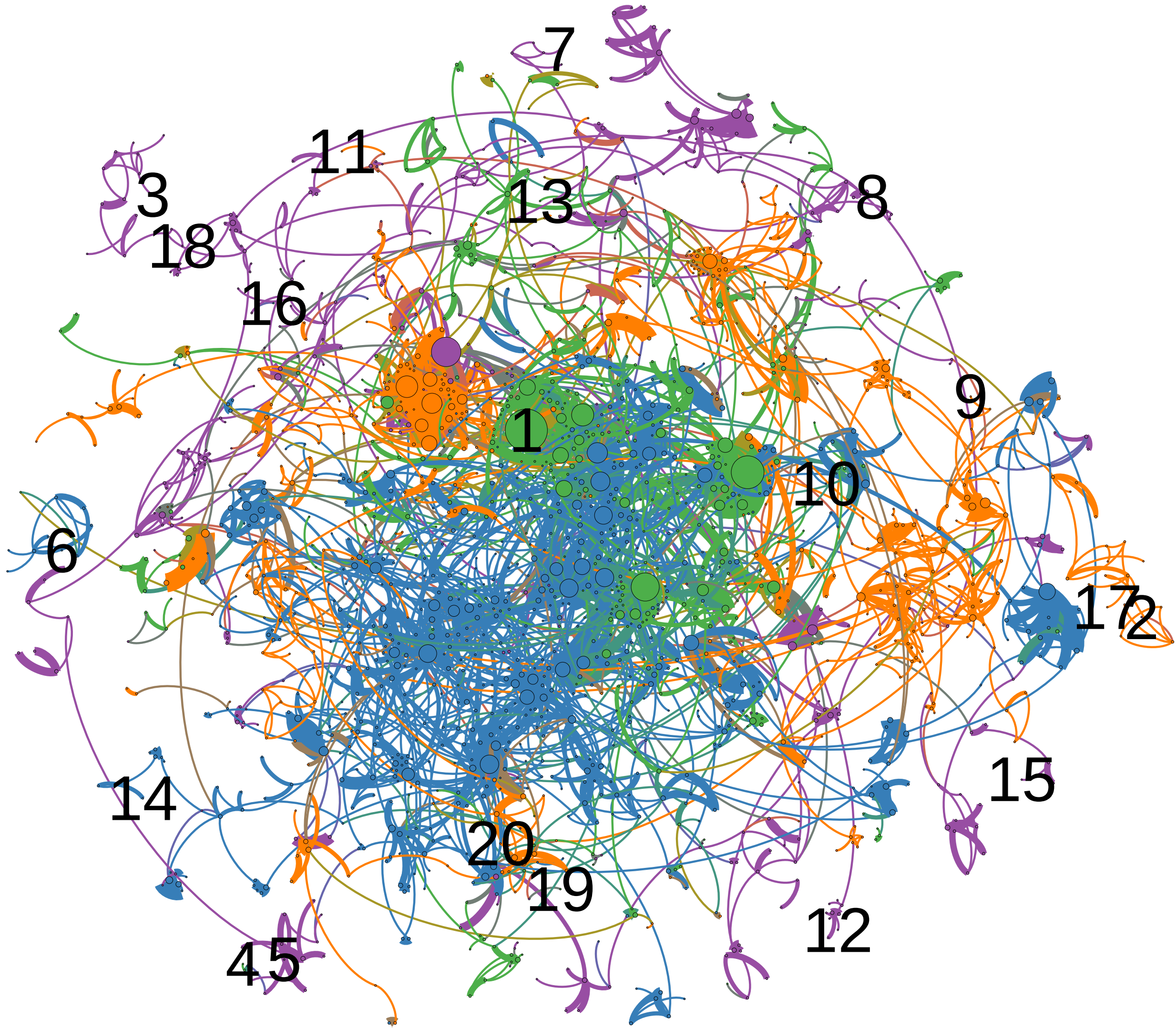
Waste samples at boundaries

Baseline 1: D-Optimality Picks Outliers

Choose the periphery

DBLP Coauthorship graph
1711 nodes, 2898 edges.
Labels (author area):

- Machine learning
- Data mining
- Information retrieval
- Database

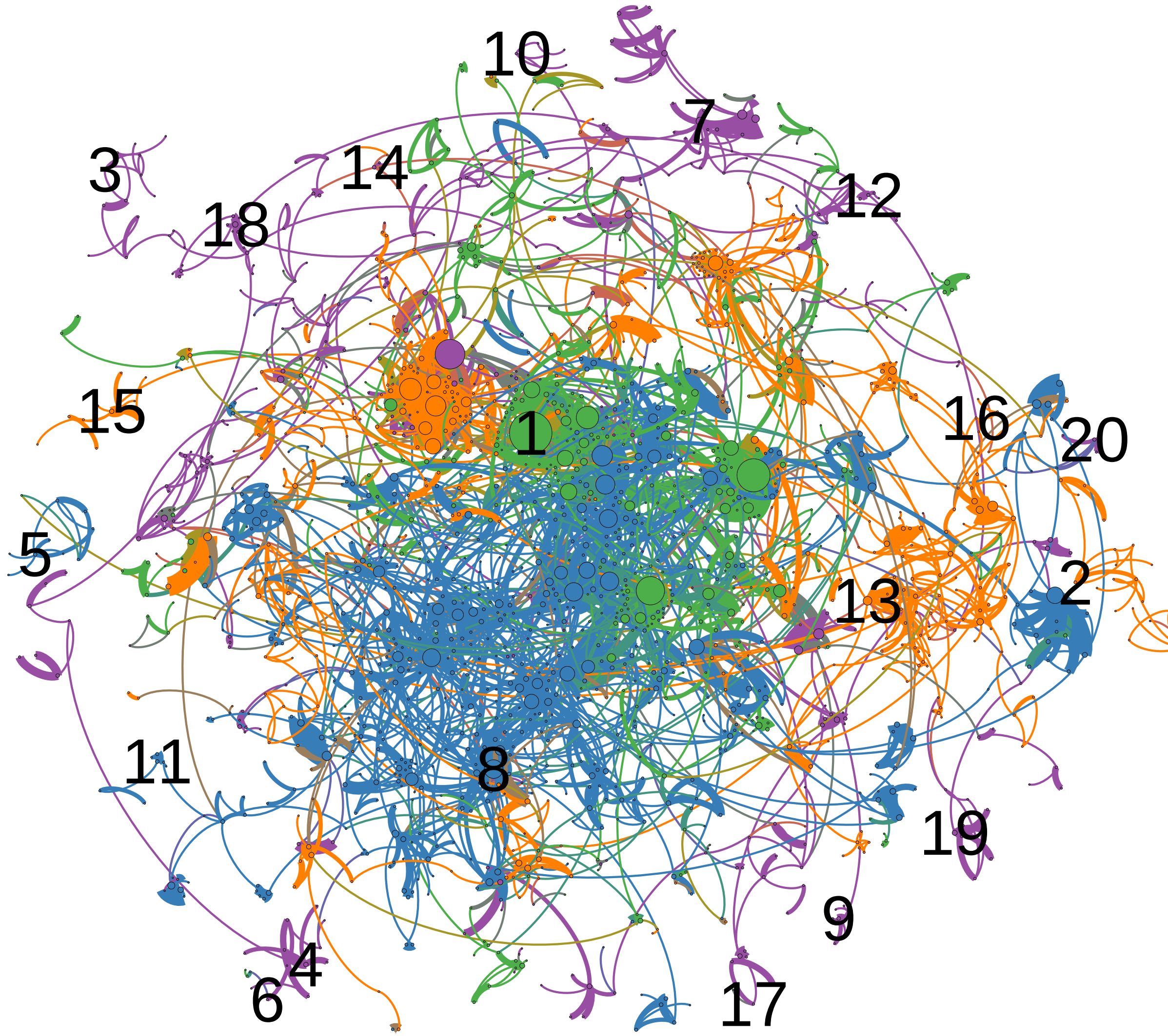


Baseline 2: V-Optimality

Minimize trace of posterior variance [Ji 2012]

$$\min_{s} \mathrm{tr}(C_{(s)}) = \mathrm{tr}\left((L_u)^{-1}\right)$$

Improves but not ideal



Our Approach: Σ-Optimality and Active Surveying

Bayesian optimal active search and survey [Garnett 2012]

Aims to predict the average of node values

$$\frac{f \cdot \mathbf{1}}{n}\Big| y_s \sim \mathcal{N}\Big(\frac{\hat{f}_{(s)} \cdot \mathbf{1}}{n}, \frac{\mathbf{1}^\top C_{(s)} \mathbf{1}}{n^2}\Big)$$

Bayesian risk minimization

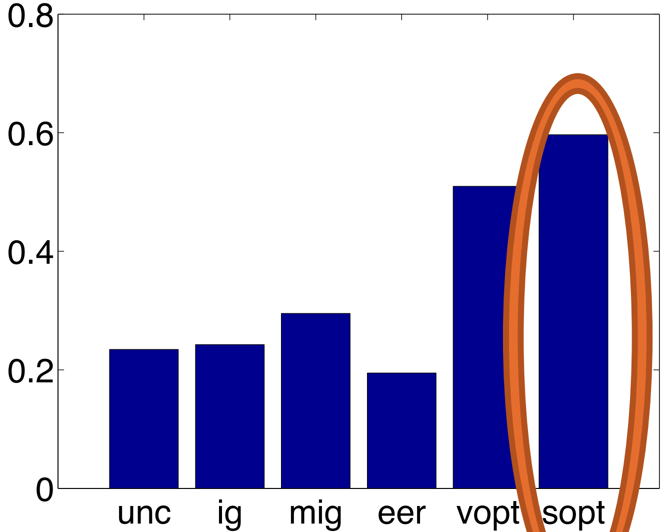
$$\min_s \mathbf{1}^\top C_{(s)} \mathbf{1}$$

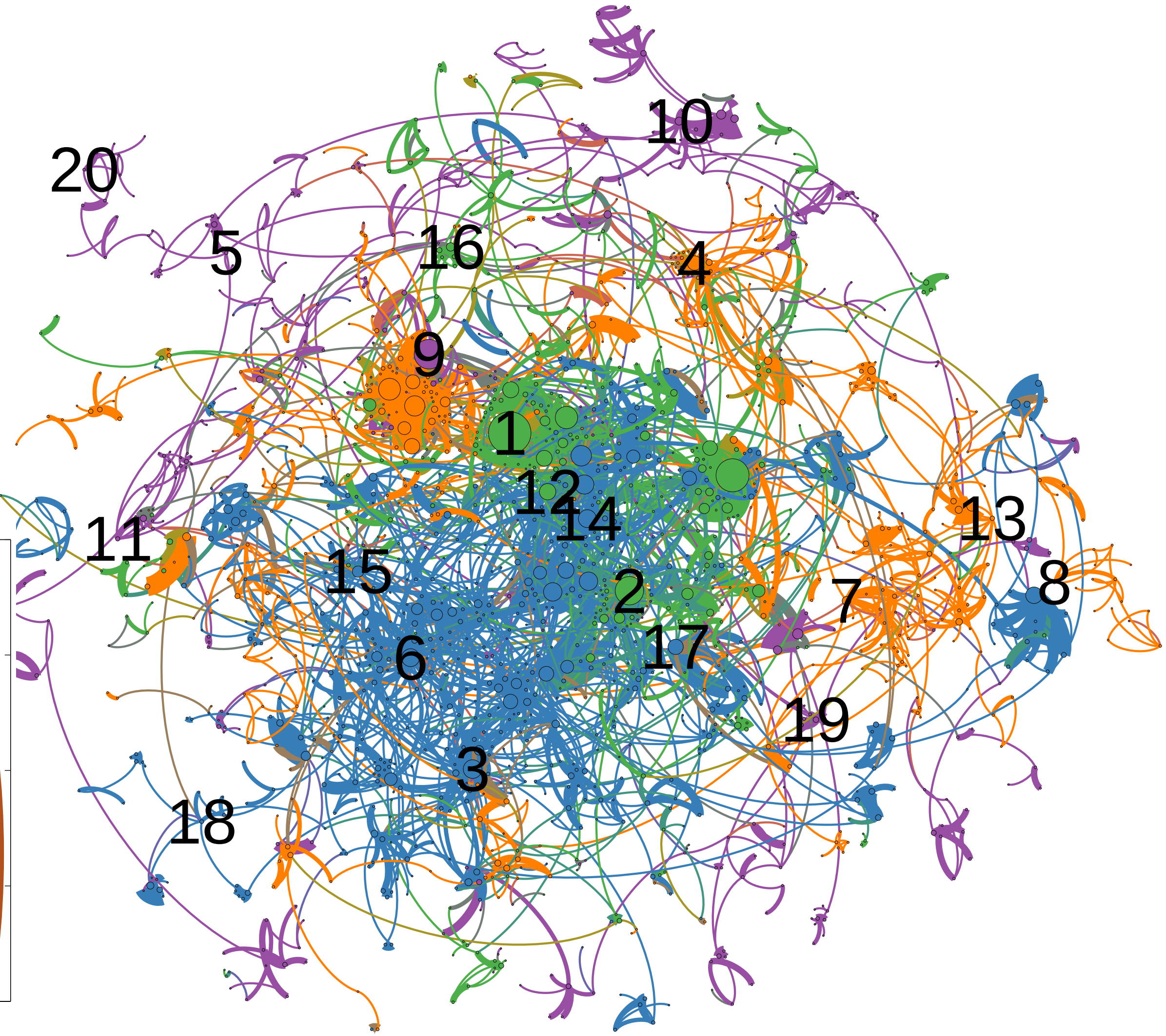
Σ-Optimality on Graphs

$$\min_s \mathbf{1}^\top C_{(s)} \mathbf{1}$$

Cluster centers!

Better active learning accuracy





Insights? Break It Down to Greedy Application

Write current covariance matrix

$$C_{(s)} = \left(\rho_{ij}\sigma_i\sigma_j\right)_{ij}$$

Apply Woodbury matrix inversion formula

$$\text{[D-Opt Krause 2008]}\ v^{t+1} = \arg\max_i \sigma_i^2$$

$$\text{[V-Opt Ji 2012]}\ v^{t+1} = \arg\max_i \sum_j (\rho_{ij}\sigma_j)^2$$

$$[\ \Sigma\text{-Opt Ours}\ ]\ v^{t+1} = \arg\max_i \sum_j \rho_{ij}\sigma_j$$

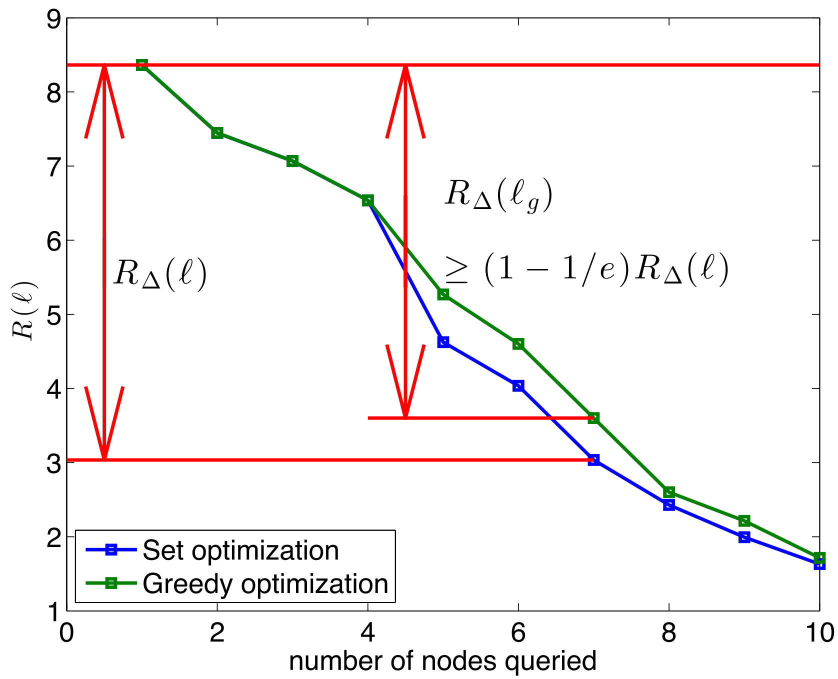
The Idea: ***L-1 more robust than L-2***

Theoretical Guarantees for Greedy Update

Monotone decreasing risk

Diminishing returns (submodularity)

Both V-opt[1] & Σ-opt[2]



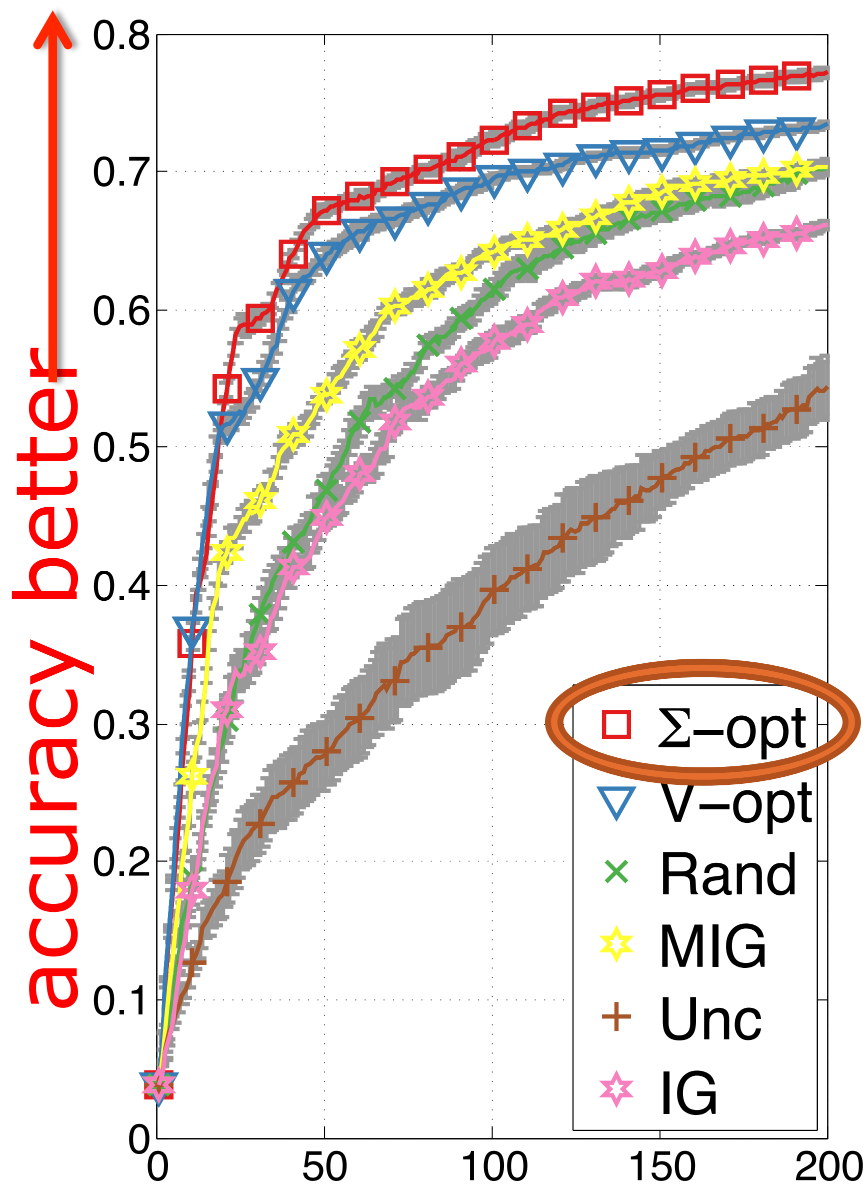
$$\begin{cases} \min_{v'} \left(\mathbf{1}^\top (L_{u^k \setminus \{v'\}})^{-1}\mathbf{1}\right) \\ \ell^{k+1} = \ell^k \cup \{v'_*\} \end{cases}$$

vs

$$\min_{\ell^{k+1}} R_\Sigma(\ell^{k+1}) = \left(\mathbf{1}^\top (L_{u^{k+1}})^{-1}\mathbf{1}\right)$$

[1]Friedlan & Gaubert, 2011; Ma et. al. NIPS workshop 2012. [2]Ma et. al. NIPS 2013

Isolet 1+2+3+4



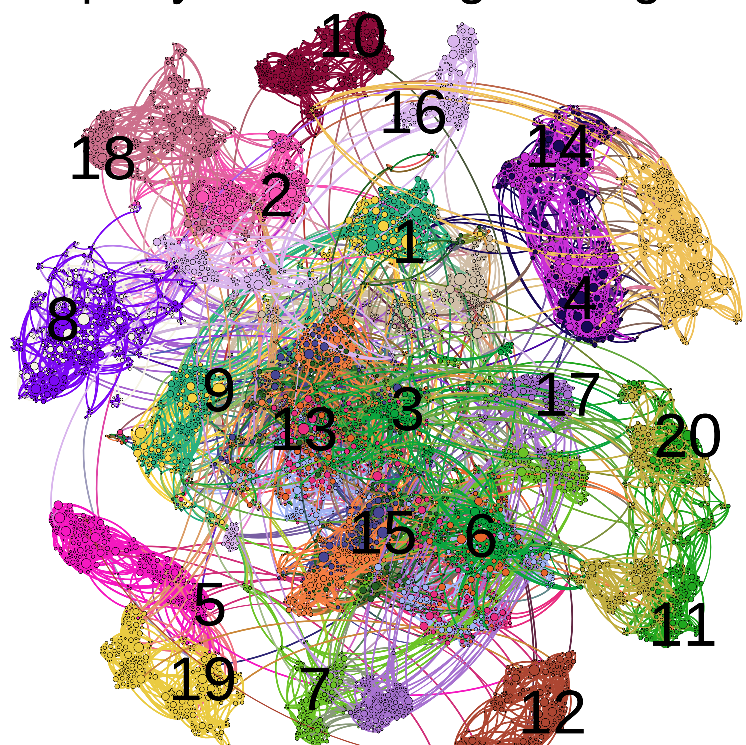
6238 Spoken letter recordings

617 dimensional frequency feature

5-nearest neighbor graph from raw input

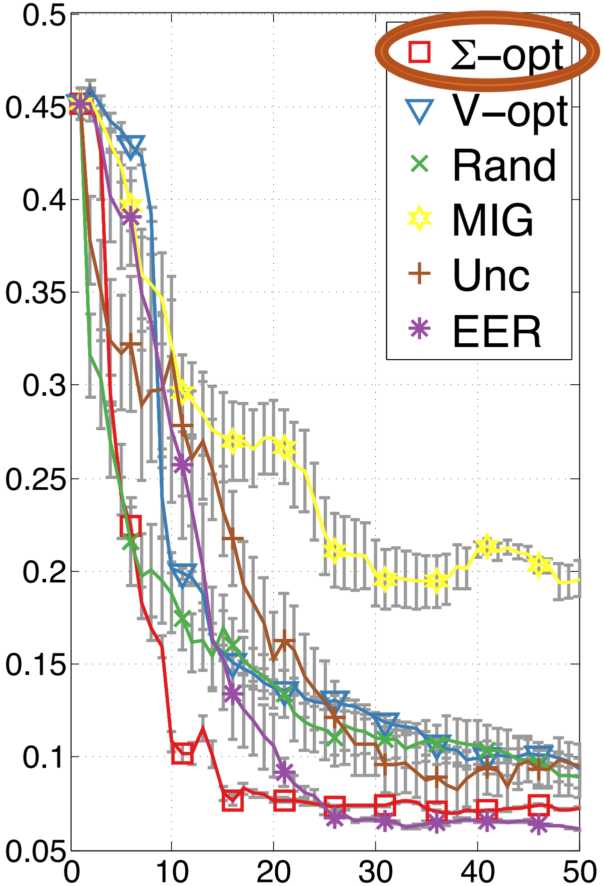
Random subsample 70% instances

First query fixed at largest degree

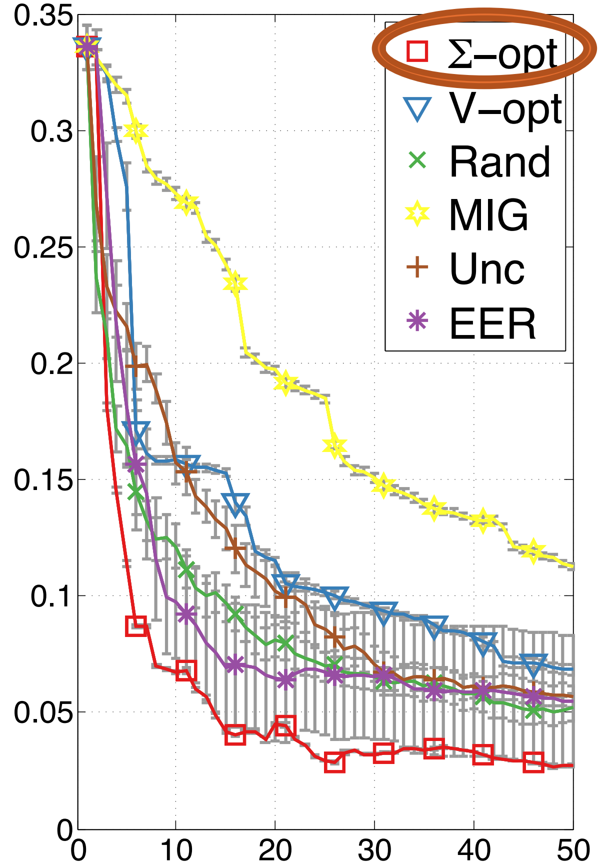


Active Surveying

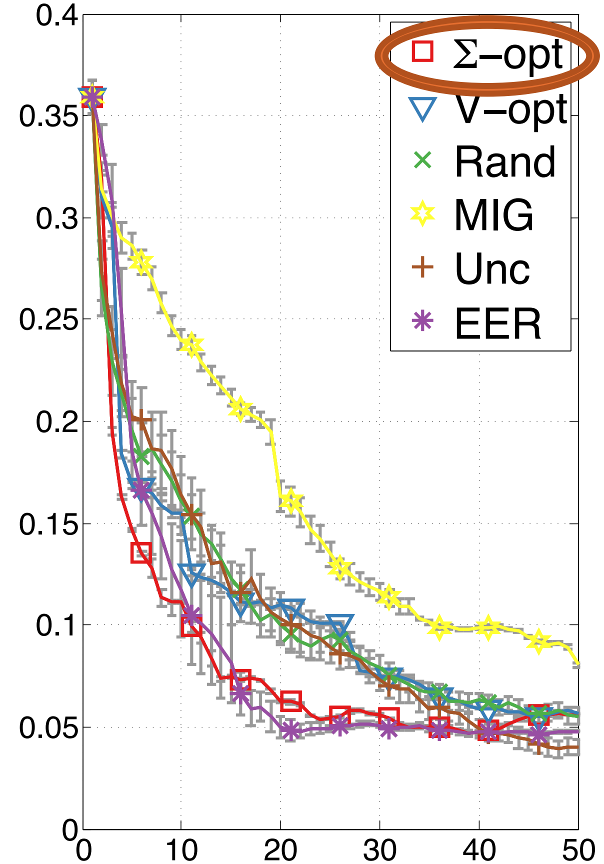
better



DBLP Coauthorship



Cora Citation



Citeseer Citation

Active Search on Graphs [Ma 2015]

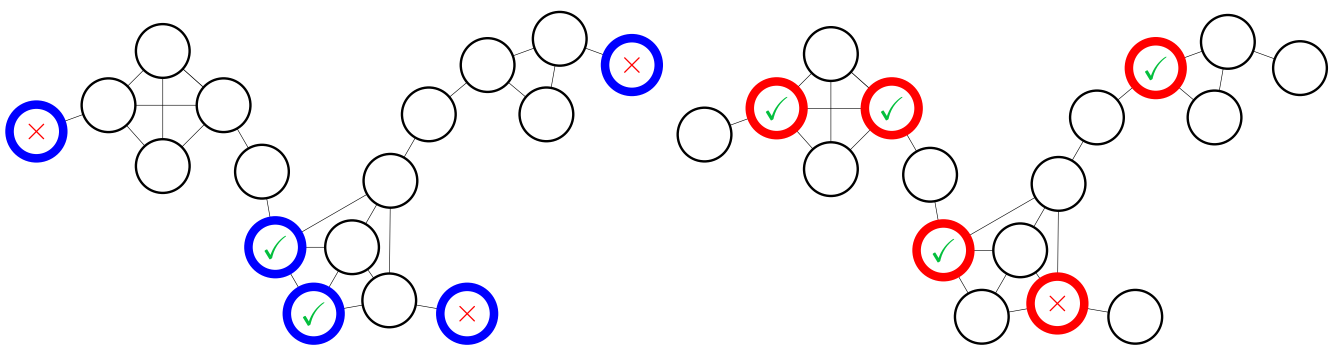
Auton Lab

Goal: find all positive nodes

Pick nodes by

$$\arg\max_i \mu_t(i) + \alpha_t \cdot s_t(i)$$

$$\text{where,} \quad s_t(i) = \sum_j \rho_{ij}\sigma_j$$



Choices in previous work   Choices by our algorithm

Active Search on Graphs [Ma 2015]

Select observations based on

$$\arg\max_i \mu_t(i) + \alpha_t \cdot s_t(i)$$

$$\text{where,} \quad s_t(i) = \sum_j \rho_{ij}\sigma_j$$

Experiment

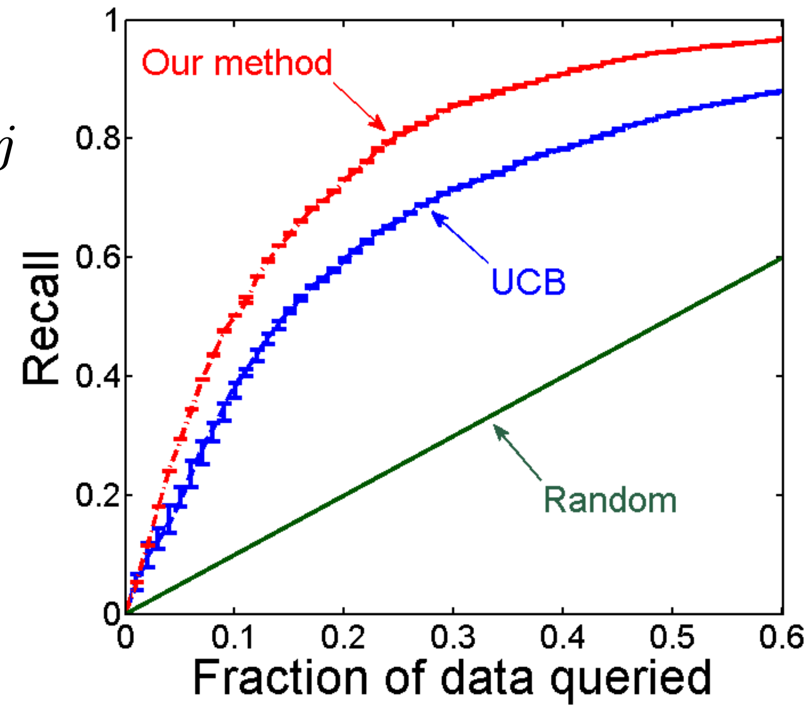
Nodes: 5000 populated places

Edges: wikipedia links

Search: 725 capitals

among countries, cities,

towns and villages



Regret Analysis

Define Regret	$R_T = \max_{v_t^*, \text{non-repeat}} \sum_{t=1}^{T} f(v_t^*) - f(v_t)$
Define Information	$\gamma_T = \max_{\lvert S \rvert \le T} \mathcal{I}(\mathbf{y}_S; f)$
Assume	$\sqrt{\mathbf{f}^\top \tilde{\mathcal{L}} \mathbf{f}} \le B, \quad \text{proper } \alpha_t$ $\gamma_T \le d_T^* \log\left(1 + \frac{T}{\sigma_n^2 \omega_0}\right),$
GP-SOPT.TT/TOPK	$\tilde{O}\left(k\sqrt{T}\left(B\sqrt{d_T^*} + d_T^*\right)\right)$, any T.
Compare With	$\tilde{O}\left(\sqrt{T}\left(B\sqrt{d_T^*} + d_T^*\right)\right)$, [ref 5]

Summary: Active Search on Graphs

Graphs can represent complex information

- Links, sparse features, hierarchical structures.

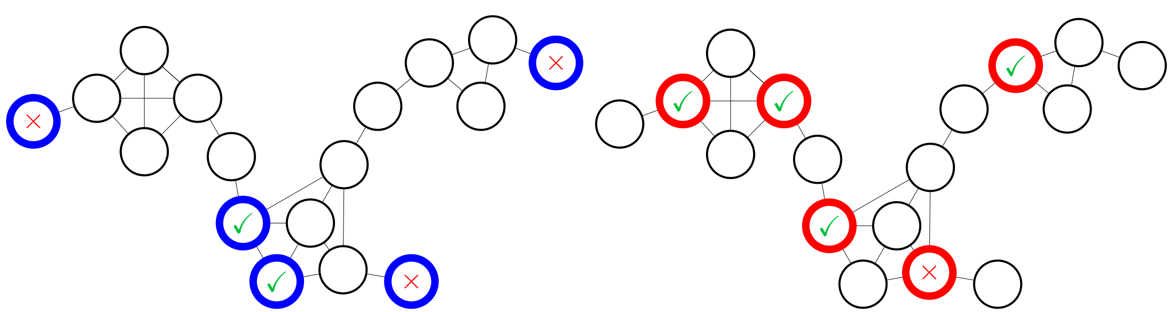
Better exploration:

- Σ-Optimality, UCB

Submodularity for global optimality

$$\arg\max_i \mu_t(i) + \alpha_t \cdot s_t(i)$$

$$\text{where,} \quad s_t(i) = \sum_j \rho_{ij}\sigma_j$$



Choices in previous work   Choices by our algorithm

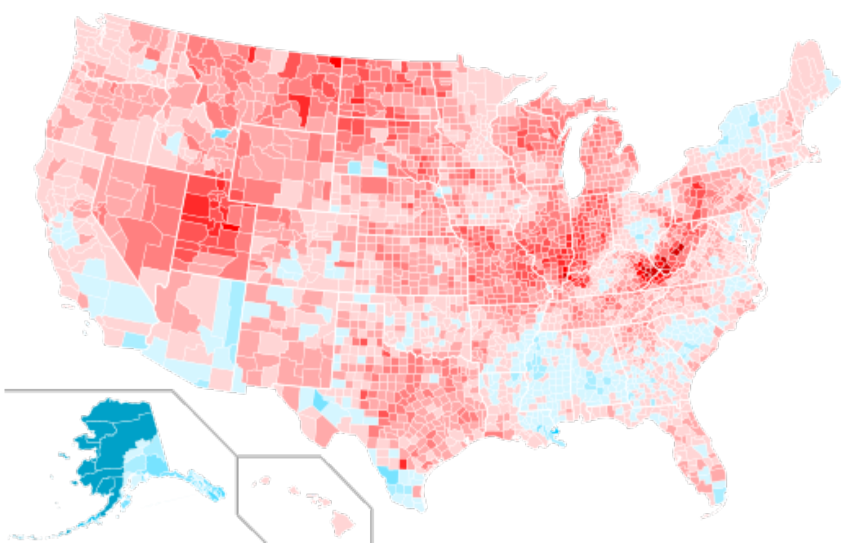
Outline

App	Challenge	Previous state-of-the-art	Contribution	Papers
Rec / Retrieval	Similarity features	Linear models	Graphs	NIPS 2013; UAI 2015
Monitoring / Polling	Reward defined by a group of points	Point rewards	Group rewards	AISTATS 2014; 2015
Surveillance	Sparse signal	Point measurements	Aggregate measurements	AAAI 2017

Idea 2: Patterns Defined by a Group of Points [Ma 2014]

Environmental monitoring

Public opinion search



Problem Definition

Point actions

On the upper level

Pay to observe

Assume GP connection

Region rewards

On the lower level

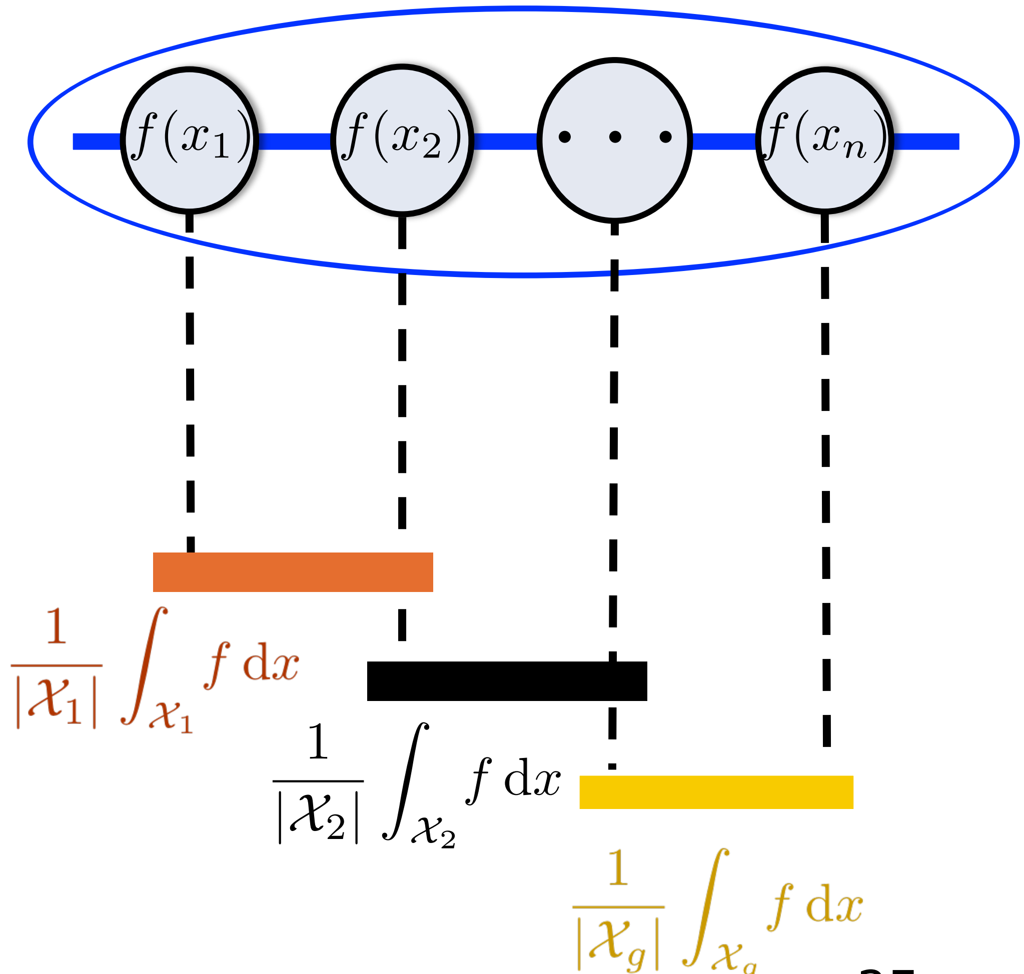
Region integral > threshold

Input

GP prior

Region definitions

Threshold



Algorithm

Maximize 1-step look-ahead expected reward

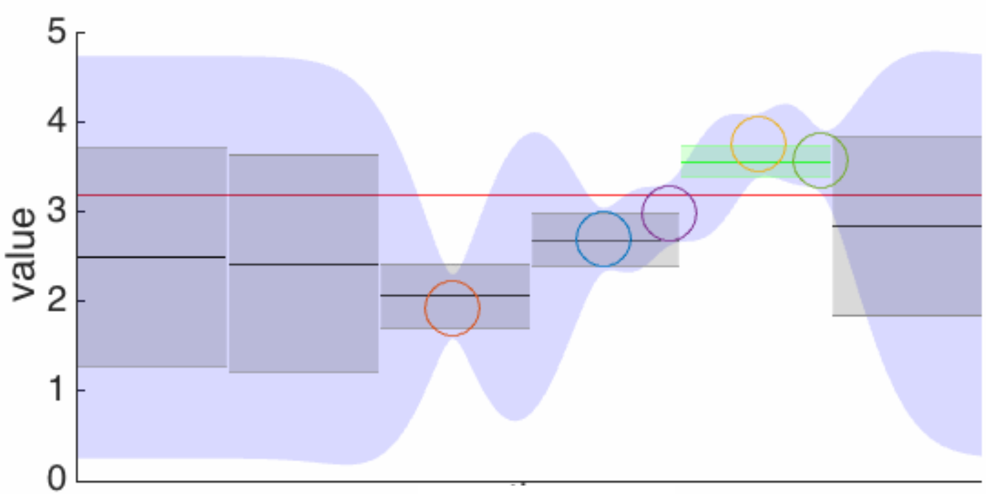
$$\max_{x_{t+1}} \int p_t(y_{t+1}|x_{t+1}) \cdot \sum_{g \in \mathcal{G}_t} \mathbf{1}(\mathrm{reward}_g \mid x_{1:t+1}, y_{1:t+1})\, \mathrm{d}y_{t+1}$$

Analytical solutions when reward on region integral

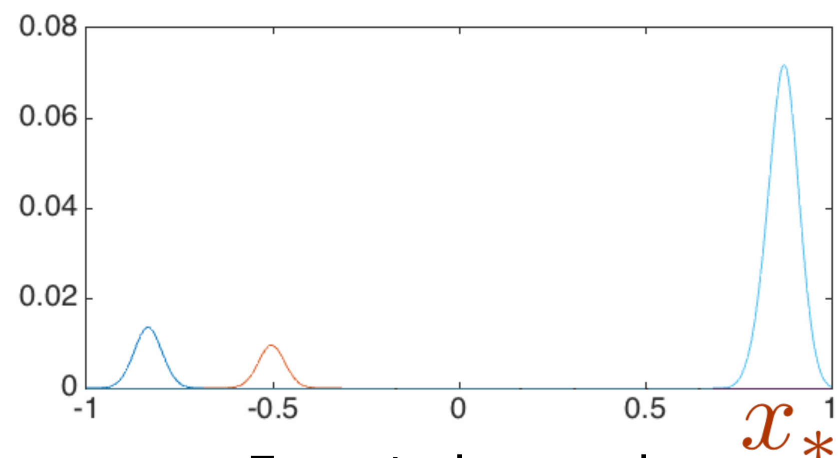
Algorithm

Maximize 1-step look-ahead expected reward

$$\max_{x_{t+1}} \int p_t(y_{t+1}|x_{t+1}) \cdot \sum_{g \in \mathcal{G}_t} \mathbf{1}(\text{reward}_g \mid x_{1:t+1}, y_{1:t+1}) \, \mathrm{d}y_{t+1}$$



Posterior



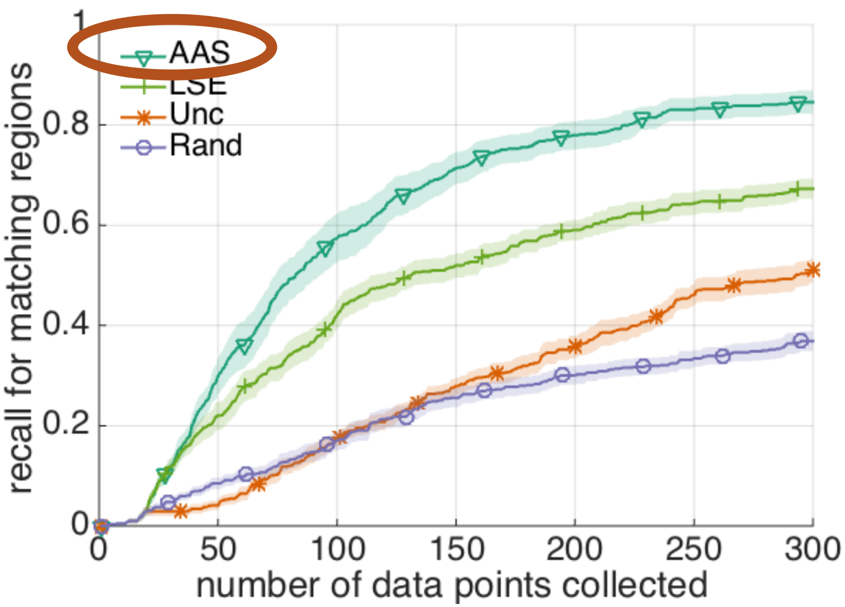
Expected reward

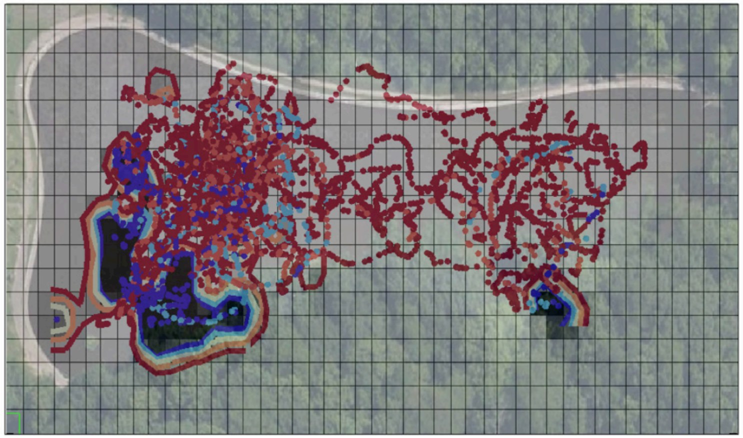
Circles: collected; blue: GP posterior; gray/green: post. of region integrals.

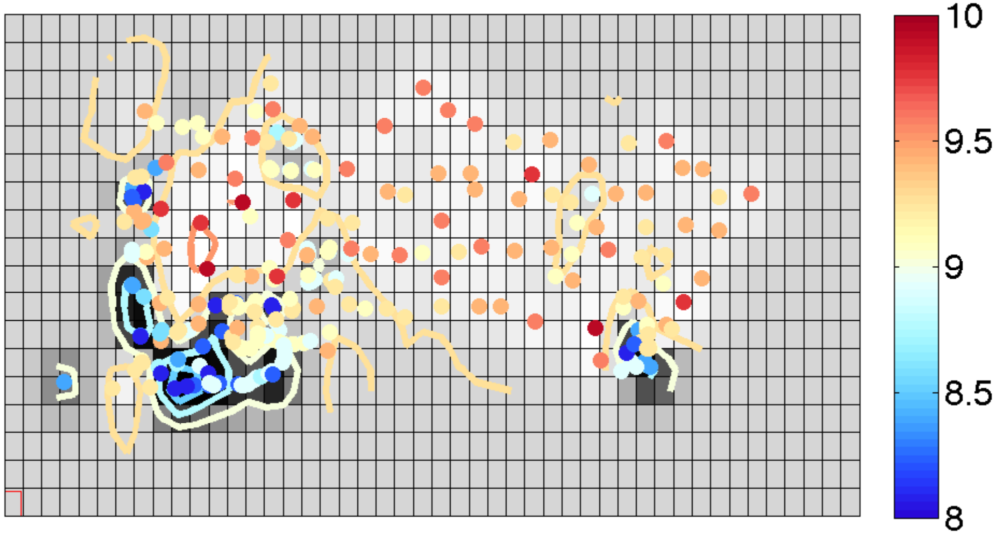
Water Quality (Dissolved Oxygen)

Recall of target regions

Re-picked measurements

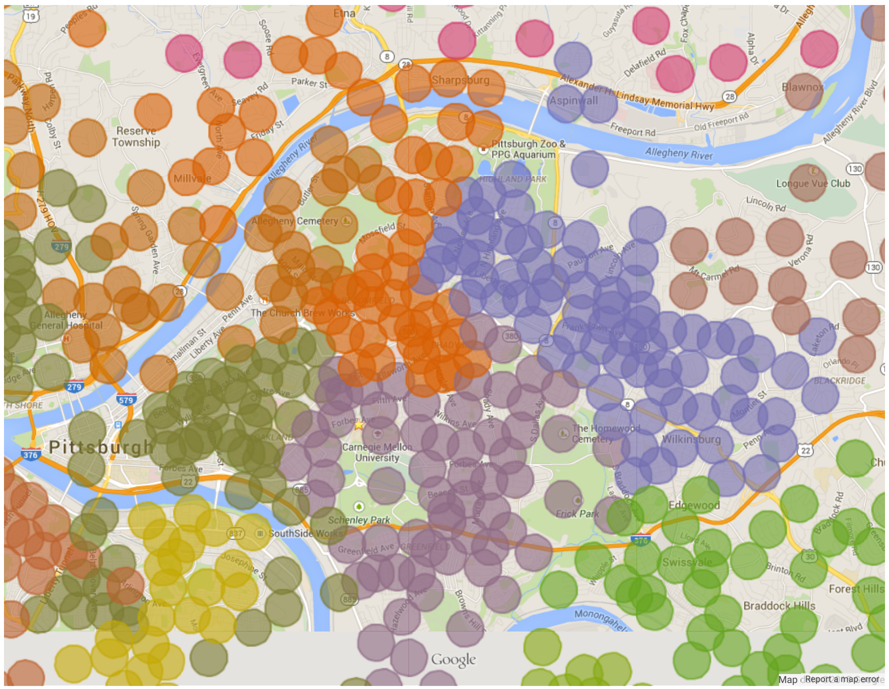




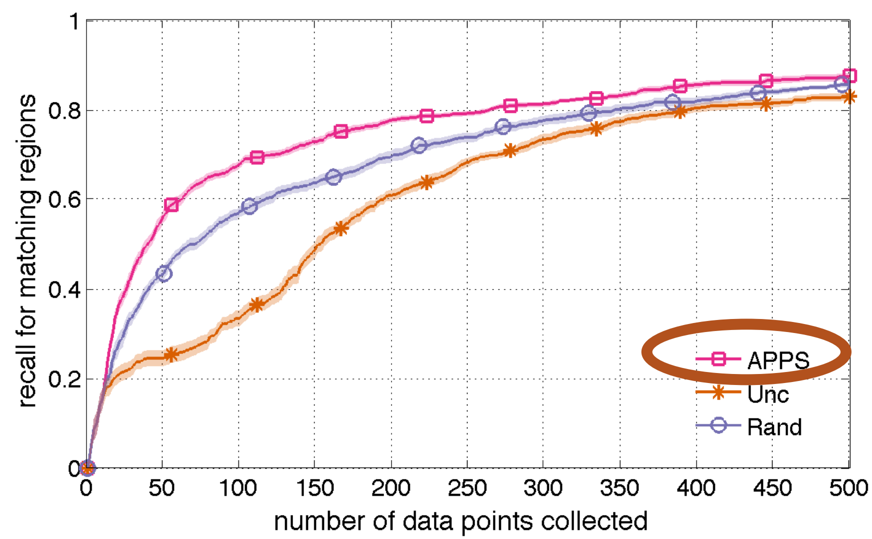


PA Election (Races vs. Precincts)

Search for positive electoral races with precinct queries.

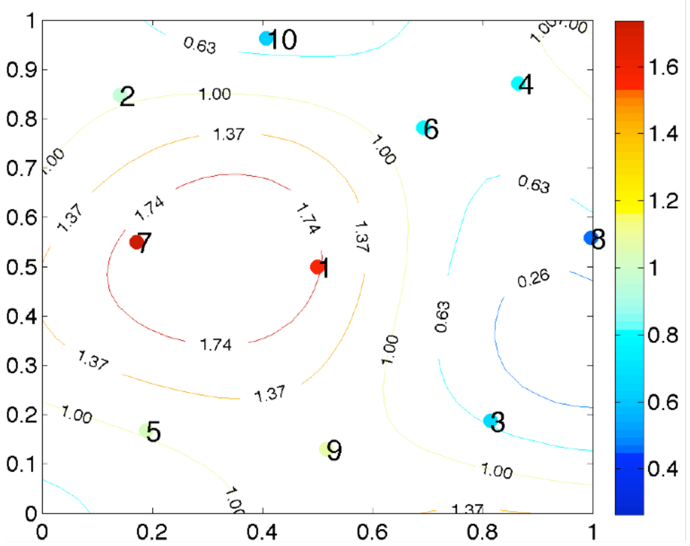


Dots: precinct centers, same color: races;
Build kernel on precincts by demographic info.



Alternative Intuition

Assuming regions are independent



Select points in a region

Variance reduction of the integral

Bayesian quadrature [Minka 2000]

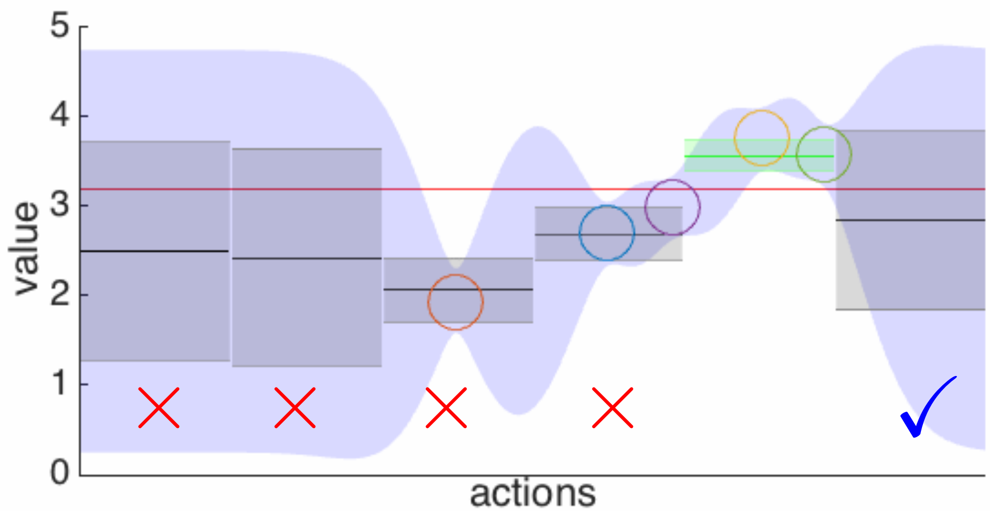
Σ-optimality

Select a region

High posterior mean

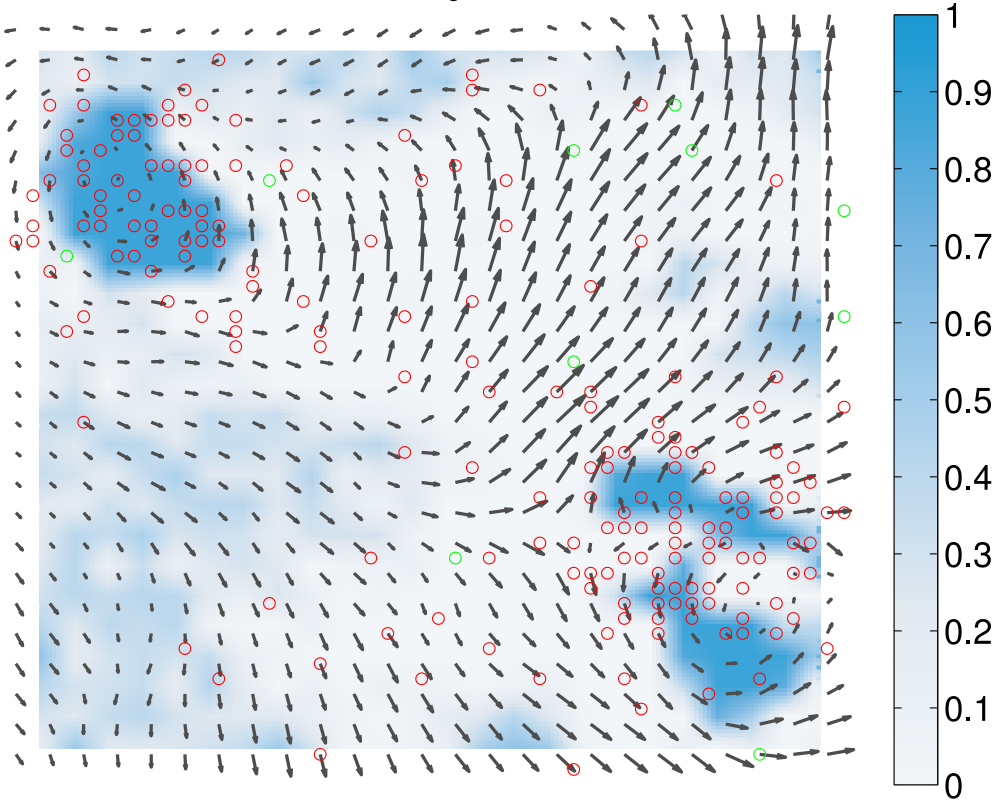
and

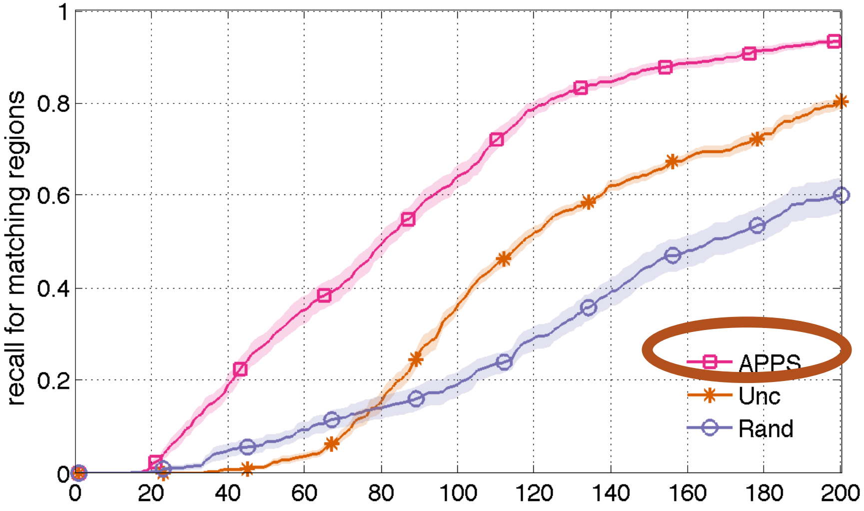
High variance reduction

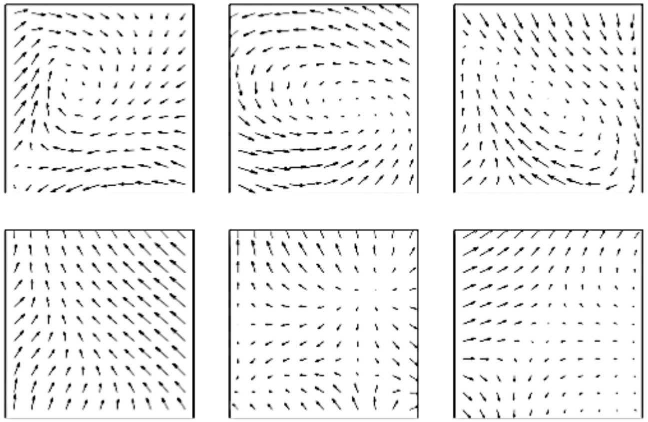


Identify Fluid Flow Vortices via Point Observations [Ma 2015]

Observe point vectors
Objective overlapping windows of 11x11 that contain a vortex
Classifier 2-layer neural net







Summary: Active Search for Region Rewards

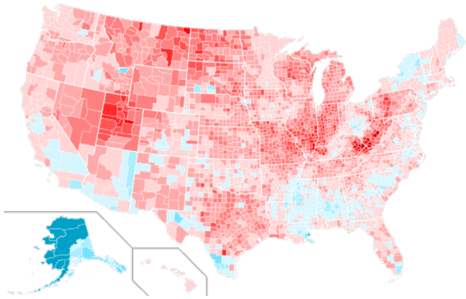
Bayesian optimization for integral rewards

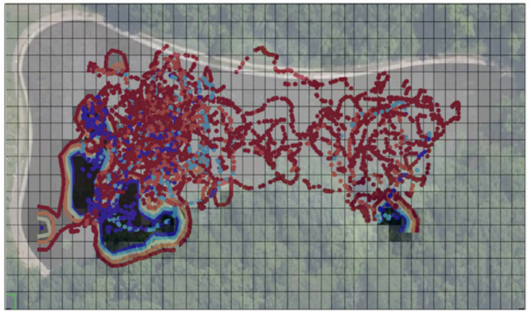
Connection to Bayesian quadrature and Σ-optimality

Expected reward balances exploration / exploitation

Empirical results on applications

Connections to multi-task BO





Outline

App	Challenge	Previous state-of-the-art	Contribution	Papers
Rec / Retrieval	Similarity features	Linear models	Graphs	NIPS 2013; UAI 2015
Monitoring / Polling	Reward defined by a group of points	Point rewards	Group rewards	AISTATS 2014; 2015
Surveillance	Sparse signal	Point measurements	Aggregate measurements	AAAI 2017

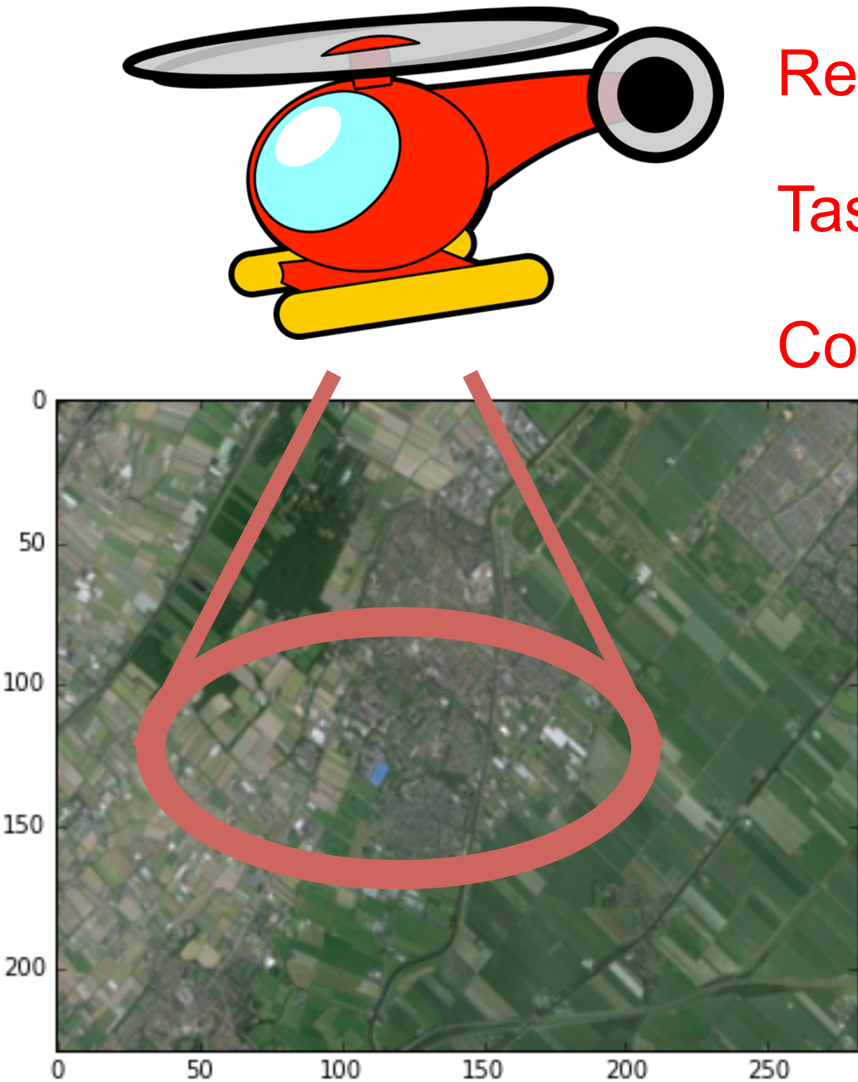
Sparse Rewards and Region Sensing

Region sensing (aggregate value)

Task: localize the sources

Control: both altitude and position

- Radiation
- Gas leaks
- Survivors



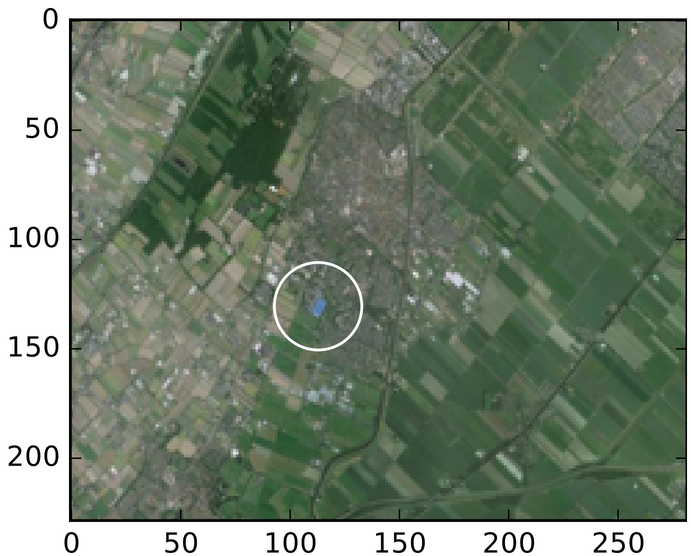
Auton Lab
Carnegie Mellon
Washington University in St. Louis

Demo Active Search

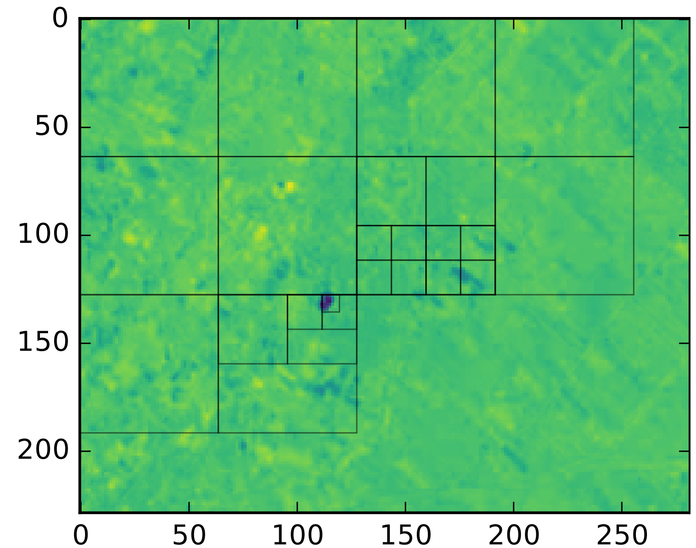
Find blue colors on a real satellite image

Simulate search and rescue in open areas

Used a blue filter on the RGB values, yielding scalar outcomes



(a) True point values

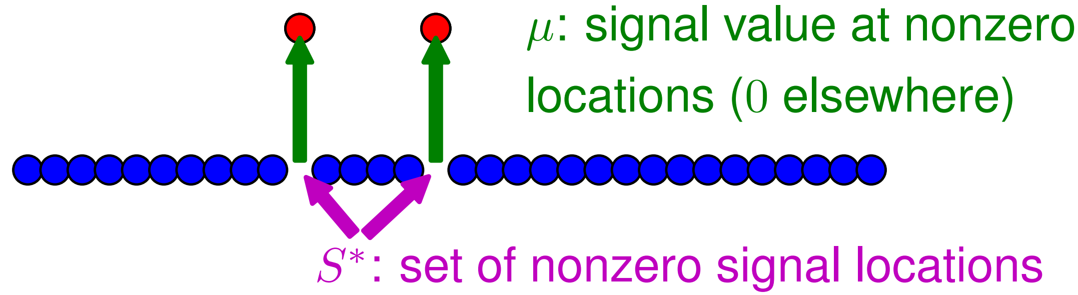


(b) search sequence

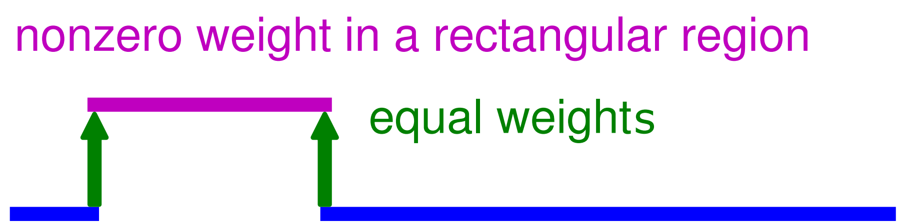
Problem Formulation

Sensing model $$y_t = \mathbf{x}_t^\top \boldsymbol{\beta}^* + \varepsilon_t,\ \varepsilon_t \sim \mathcal{N}(0,1)$$

$\boldsymbol{\beta}^* \in \mathbb{R}_+^n$
k-sparse signal



$\mathbf{x}_t \in \mathbb{R}_+^n,\ \|\mathbf{x}_t\|_2 = 1$
aggregate measurement



Objective: design x_t to recovery the support of β^*

Algorithm

Assume uniform prior $\boldsymbol{\beta} \in \{\mu\mathbf{e}_1, \mu\mathbf{e}_2, \ldots, \mu\mathbf{e}_n\}$

Repeat

Pick $\arg\max_{\mathbf{x}_t} I_{t-1}\big(\boldsymbol{\beta}; y(\mathbf{x}_t)\big)$

Observe y_t

Update $\pi_t(\boldsymbol{\beta}) \propto \pi_{t-1}(\boldsymbol{\beta})\phi(y_t - \mathbf{x}_t^\top \boldsymbol{\beta})$

For k-sparse, repeat the above to find each signal

Information Gain

Equivalent to marginal entropy,

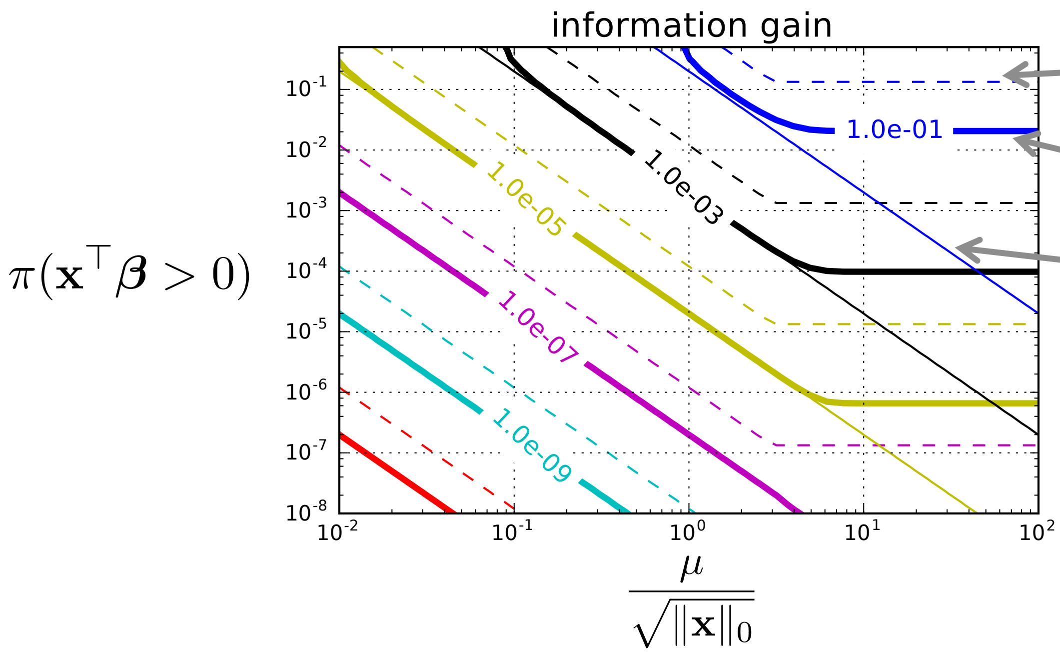
$$I(\boldsymbol{\beta}; y(\mathbf{x})) \simeq H(y(\mathbf{x}))$$

e.g., for binary search on the prior,

$$y(\mathbf{x}) \sim \begin{cases} \mathcal{N}\left(\frac{\mu}{\sqrt{\|\mathbf{x}\|_0}}, 1\right) & \text{w.p. } \pi(\mathbf{x}^\top \boldsymbol{\beta} > 0) = \frac{1}{2}; \\ \mathcal{N}(0, 1) & \text{otherwise.} \end{cases}$$

For noiseless, the entropy is log(2).

Information Gain Can Be Bounded



- ✓ Pinsker's inequality
- ✓ True information
- Jensen's inequality

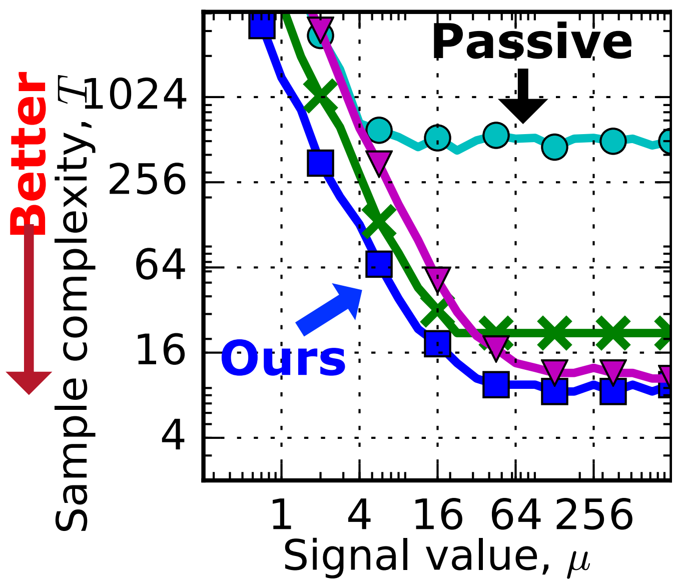
On the prior:

$$\min\left\{\frac{\mu^2}{12n}, \frac{1}{8}\right\} \leq I_0(\boldsymbol{\beta}, y(\mathbf{x})) \leq \frac{\mu^2}{2n}$$

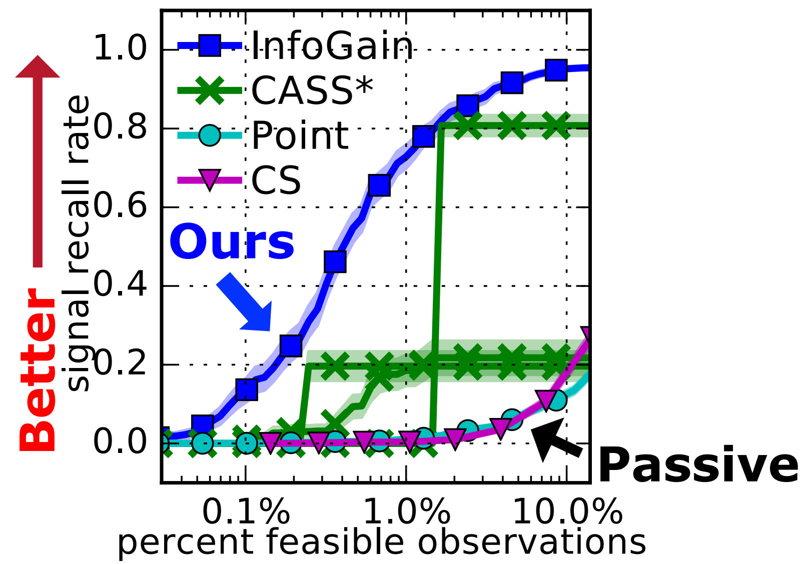
Theoretical and Empirical Results

Theoretically optimal: uses $\tilde{O}\left(\frac{n}{\mu^2} + k^2\right)$ measurements

Significantly better than passive sensing under region constraints



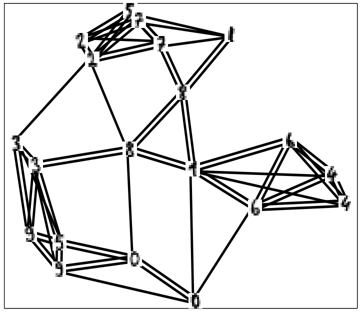
Number of measurements in simulated data

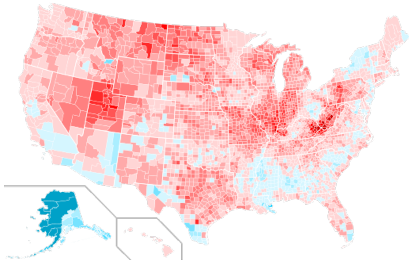


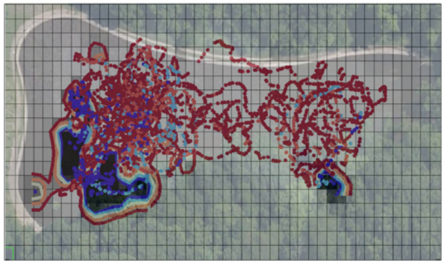
Average search progression on satellite images

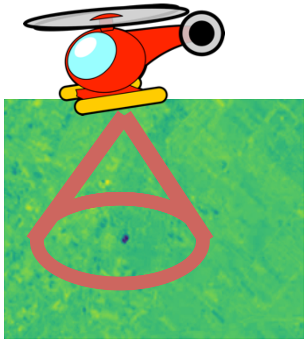
Summary

App	Challenge	Previous state-of-the-art	Contribution	Papers
Rec / Retrieval	Similarity features	Linear models	Graphs	NIPS 2013; UAI 2015
Monitoring / Polling	Reward defined by a group of points	Point rewards	Group rewards	AISTATS 2014; 2015
Surveillance	Sparse signal	Point measurements	Aggregate measurements	AAAI 2017







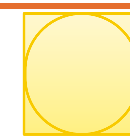


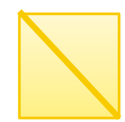
The Goal: Compare Sequential Active Learning Algos

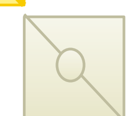
Sequentially Select *s* for
$$\begin{cases} P(y_u|y_s) \propto \mathcal{N}(y_u; \hat{y}_u, L_u^{-1}) \\ L = \begin{pmatrix} L_u & L_{us} \\ L_{su} & L_s \end{pmatrix}, \quad \hat{y}_u = -L_u^{-1} L_{us} y_s \end{cases}$$

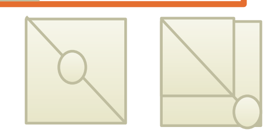
***s*: labeled, *u*: unlabeled. (*u*,*s*): complementary**

Possible strategies: (at step k with u^k unlabeled)

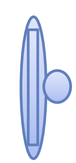
Σ-Optimality[1] $$\min_{v'} \left(\mathbf{1}^\top (L_{u^k \setminus \{v'\}})^{-1} \mathbf{1} \right)$$


V-Optimality[2] $$\min_{v'} \operatorname{tr}\left((L_{u^k \setminus \{v'\}})^{-1} \right)$$


Info Gain (IG)[3] $$\max_{v'} \left(L_{u^k}^{-1} \right)_{v',v'}$$


Mutual (MIG)[3] $$\max_{v'} \left(L_{u^k}^{-1} \right)_{v',v'} \Big/ \left((L_{\ell^k \cup \{v'\}})^{-1} \right)_{v',v'}$$


Uncertainty[4] $$\min_{v'} \left| \hat{y}_{v'} \right|$$


E Error (EER)[4] $$\max_{v'} \mathbb{E}_{y_{v'}} \left[\left(\sum_{u_i \in u} \left| \hat{y}_{u_i} \right| \middle| y_{v'} \right) \middle| y_{\ell^k} \right]$$


[1]Ma et. al. 2013. [2]Zhu et. al. 2003; Ji & Han, 2012. [3]Krause et. al. 2008. [4]Settles 2012.