Active Learning and Optimized Information Gathering

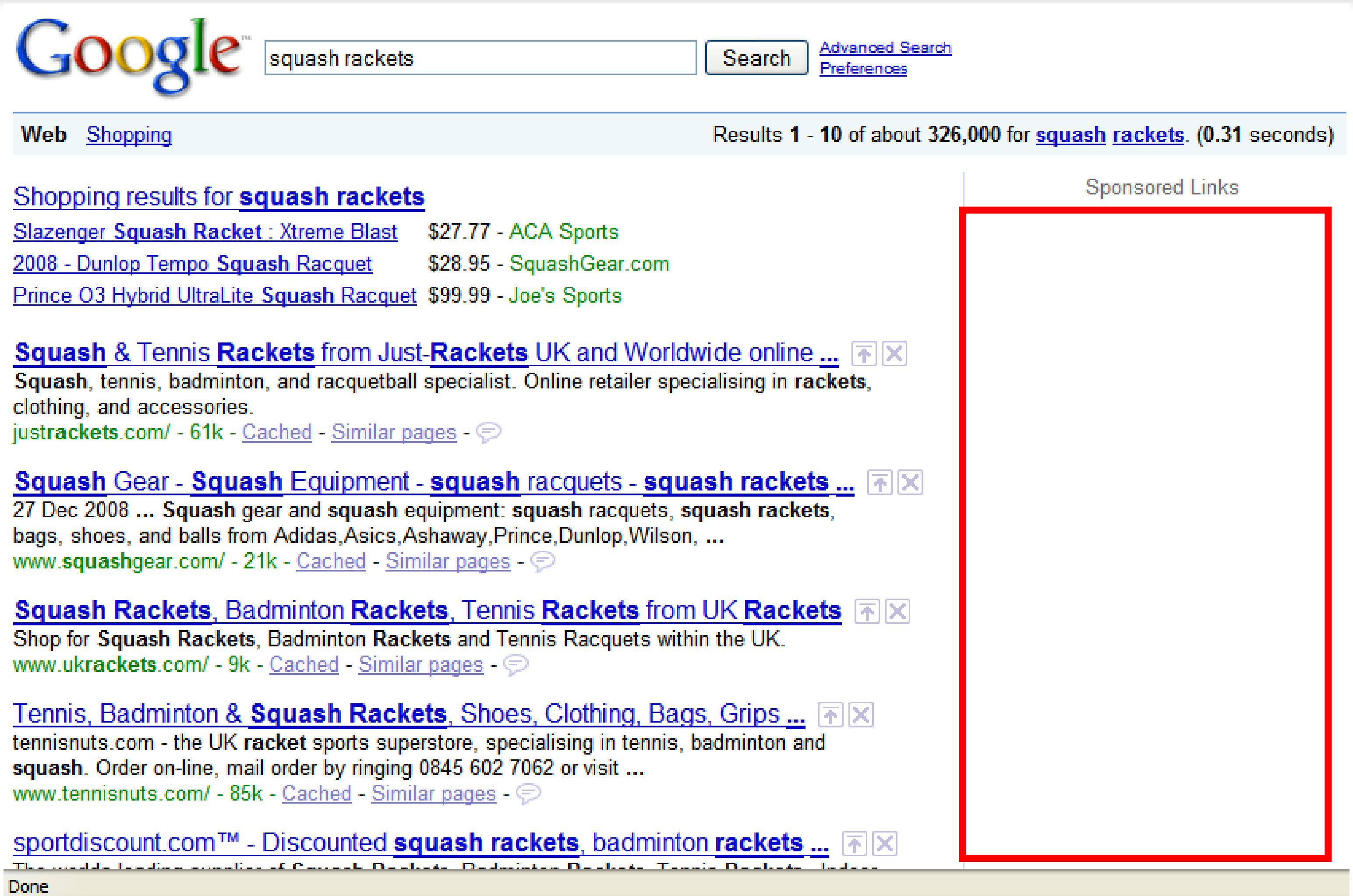
Lecture 19 – Summary

CS 101.2

Andreas Krause

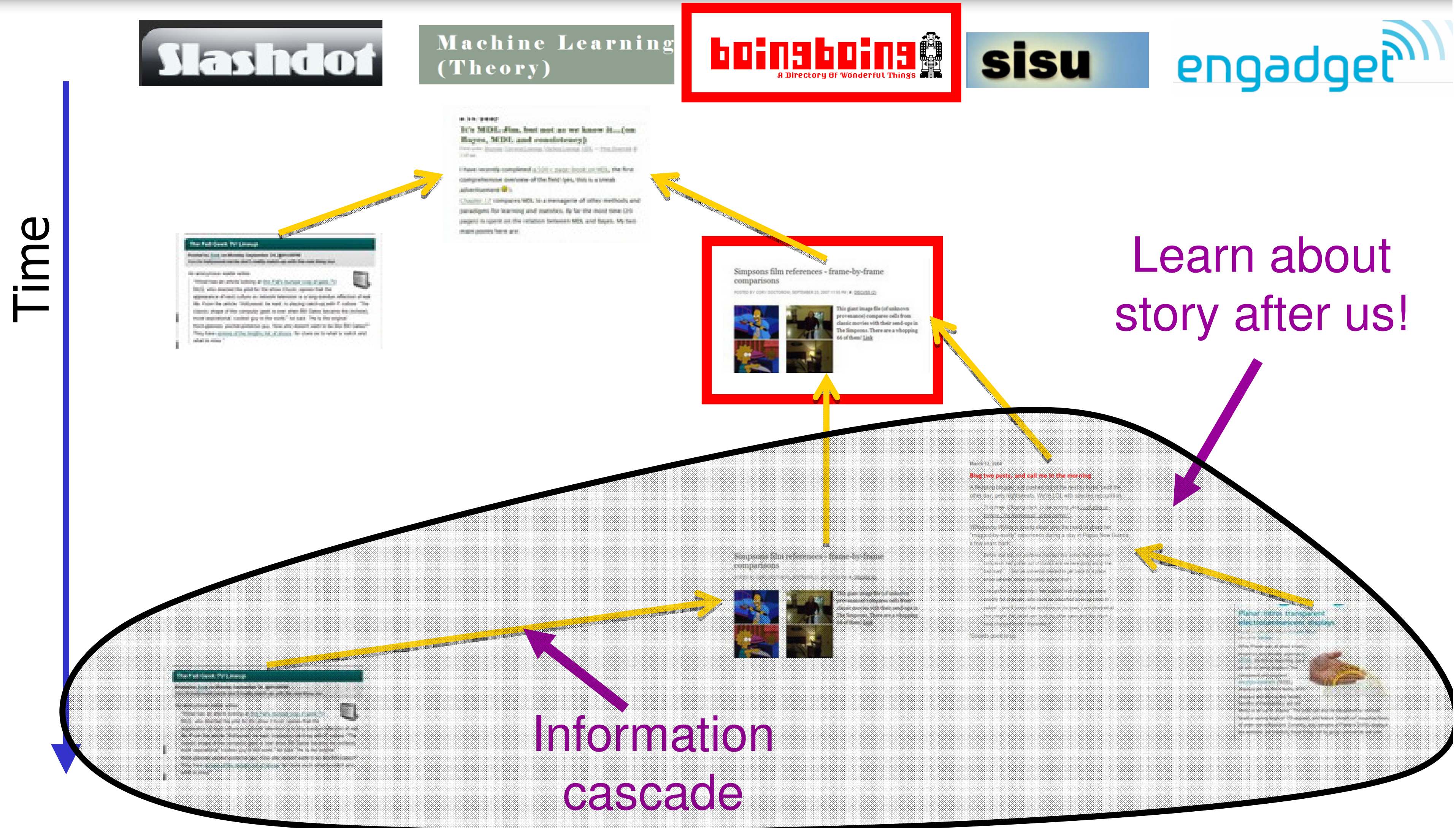
How can we get **most useful** information
at **minimum cost?**

Sponsored search



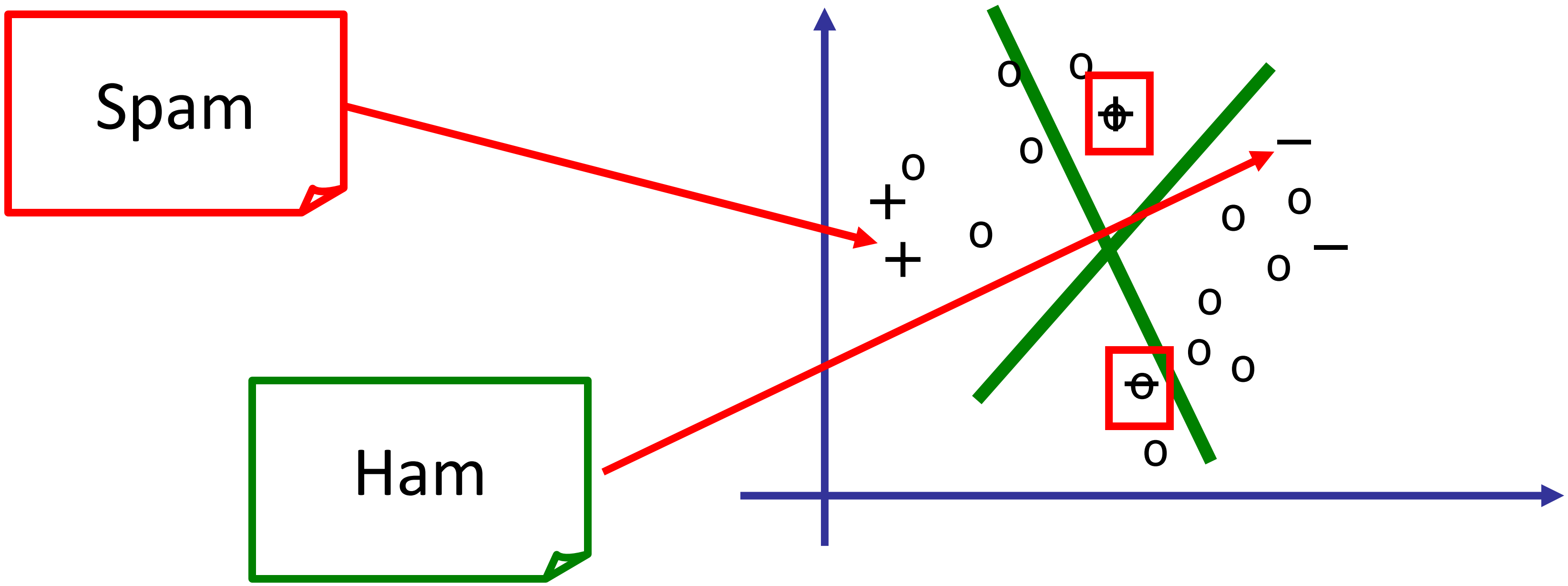
Which ads should be displayed to maximize revenue?

Which blogs to read



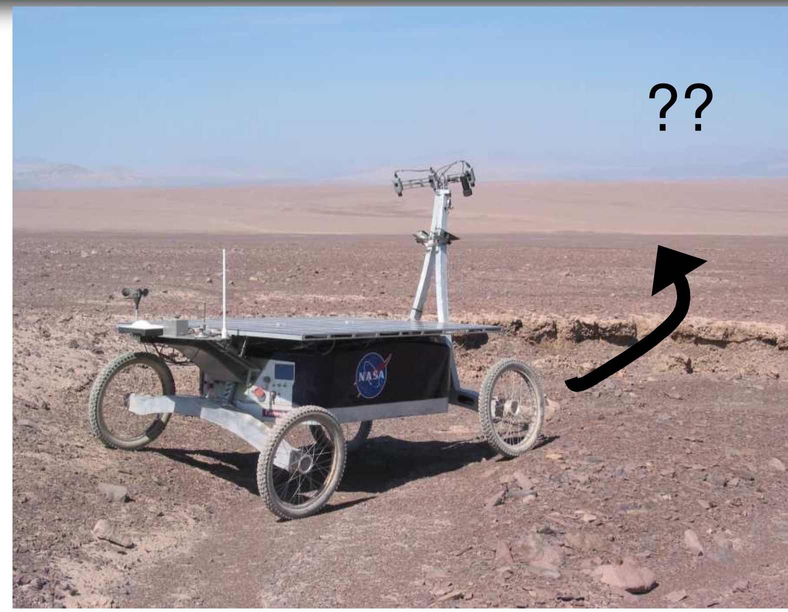
Which blogs should we read to learn about big cascades early?

Spam or Ham?

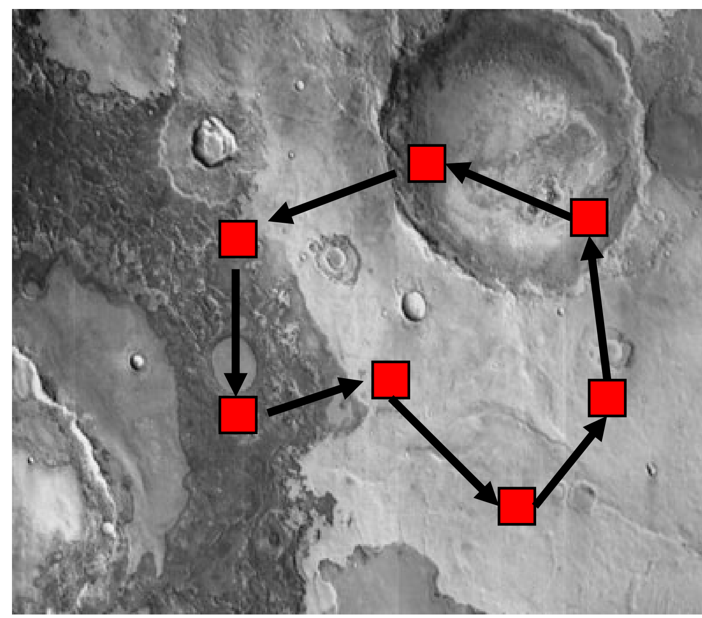


- Labels are expensive (need to ask expert)
- Which labels should we obtain to maximize classification accuracy?

Automated environmental monitoring



- Robots collect measurements
- Limited capacity requires selection



Key intellectual questions

- How can a machine choose experiments that allow it to maximize its performance in an unfamiliar environment?
- How can a machine tell “interesting and useful” data from noise?
- How can we develop tools that allow us to cope with the overload of information?
- How can we automate curiosity?

What you've learned in this class

- Bandit problems, Exploration / Exploitation tradeoffs
- Online algorithms, regret minimization
- Reinforcement learning and MDPs
- Learning theory (PAC learning, VC dimension,...)
- Active learning (pool-based, label complexity..)
- Uncertainty sampling
- Kernel methods (Gaussian processes, SVMs, …)
- Value of information
- Bayesian modeling
- Bayesian experimental design
- Submodular function optimization
- Sparsity (Sparse PCA, Compressed sensing, …)
- Applications (Human learning, robotics, sensor networks, neuroscience, …)

Big picture

- Three types of approaches
  1. Online decision making
  2. Statistical active learning
  3. Combinatorial approaches

- All approaches specify
  - a **goal** of the information gathering task
  - a class of **queries** that can be posed
- This allows to develop algorithms for selecting most useful information

Overview of approaches

Approach	**Goal**	**Queries**
Online optimization (bandits, experts,…)	**Maximize** a noisy function	Function values at selected inputs
Active learning for classification	Learn a hypothesis (identify function **level sets**)	Labels for selected inputs
Active learning for regression	**Estimate** a function everywhere	Function values at selected inputs
Bayesian experimental design	Allow **inferences** in prob. model	Subset of variables

Approaches vary in

- Assumptions made about the world
  - **Bayesian** (prior distribution over states of the world)
  - **Frequentist** (no prior, but iid noise)
  - **Adversarial** (oblivious, adaptive, …)
- Adaptivity
  - **A priori** approaches select all observations before measurements are made
  - **Sequential** approaches choose observations based on prior observations
  - Multi-stage
- Guarantees about solutions
  - Regret guarantees
  - Improvement in sample complexity
  - Approximation guarantees for fixed sample size

Summary online prediction

- Natural formalism for studying exploration / exploitation tradeoffs
- Often, algorithms are very robust:
  - Can deal with adversarial noise
- Many extensions to practical settings
  - Exploit structure in pay-off function
  - Exploit context dependency
- (Often) lead to practical algorithms
- Can only be used for noisy function optimization

Summary statistical active learning

- Only select most useful samples to quickly learn complex hypotheses
- Can get exponential improvement in sample complexity!!
  - Threshold functions
  - Homogeneous linear separators
- Can suffer from sampling bias
  - Pool based active learning is a principled way around this
- Positive results often make strong assumptions
- For noisy data, often only fallback guarantees

Summary combinatorial approaches

- Select informative variables to facilitate decision making
  - Value of information, Bayesian experimental design
- Strongest theoretical results for a priori selection problems
- Can accommodate complex constraints
  - Varying cost functions
  - Informative path planning
- Lead to very practical and efficient algorithms
- Have to make fairly strong assumptions (Bayesian prior)

Final project

- Writeup due March 17 (next Tuesday), 11:59pm
- 8 pages NIPS format
- Clearly discuss
  - Problem statement
  - Formal model used to address problem
  - Approach used to solve the problem
  - Experimental results / proofs

Project Poster Session

- Tuesday March 17 1pm-2:30pm
- Second floor Powell-Booth (CACR Atrium)
- Easels and poster boards will be made available
  - Can pick up poster boards (32” by 40”) on Monday in my office
- Tell other people to come (will have cookies ☺)
- Will have a **best project award** (public vote)!!

Course feedback

- Your feedback is important!!
  - What was good, what should be improved?
  - Design of new, machine learning / AI related courses
- PLEASE fill out
  - Online survey (TQFR)
  - Written form (distributed in class)