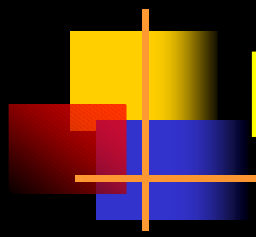
Some Rigorous Approaches to Data Mining

Christos H. Papadimitriou

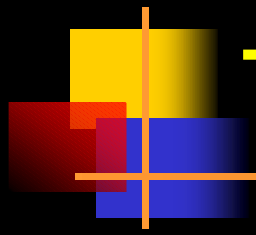
www.cs.berkeley.edu/~christos



Data mining:

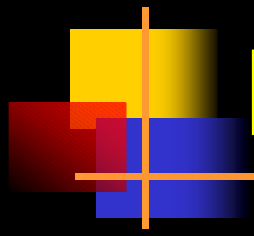
- the killer app of the new era
- important research issues
- *important social issues*
- fundamentally *economic* activity

  "aggregation is a necessary evil, since we do not have data, and, even if we had them, they would be hard to analyze"
- source of neat theoretical problems



This talk:

- clustering and segmentation (joint work with Jon Kleinberg and Prabhakar Raghavan)
- mining the web graph (Bar-Yossef, Berg, Chien, Fakcharoenphol, Weitz, VLDB 2000)
- thoughts on privacy

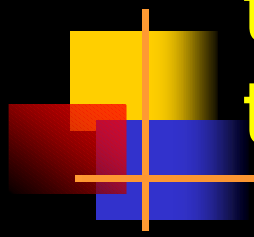


Data mining:

discovering interesting patterns in data

- patterns: clusters, associations, ...

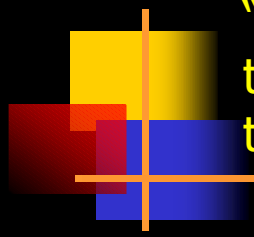
- interesting: ?



"a pattern is interesting to the extent to which it advances the objectives of the enterprise"

a *microeconomic* view of data mining

"patterns you can take to the bank"



"a pattern is interesting
to the extent to which it advances
the objectives of the enterprise"

Every enterprise is facing an optimization problem like this:

$$\max_{x \in D} f(x)$$

objective

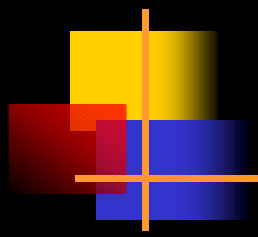
decision

domain of all possible decisions

Or, more accurately:

$$\max_{x \in D} \sum_{i \in C} f_i(x)$$

set of all *customers*



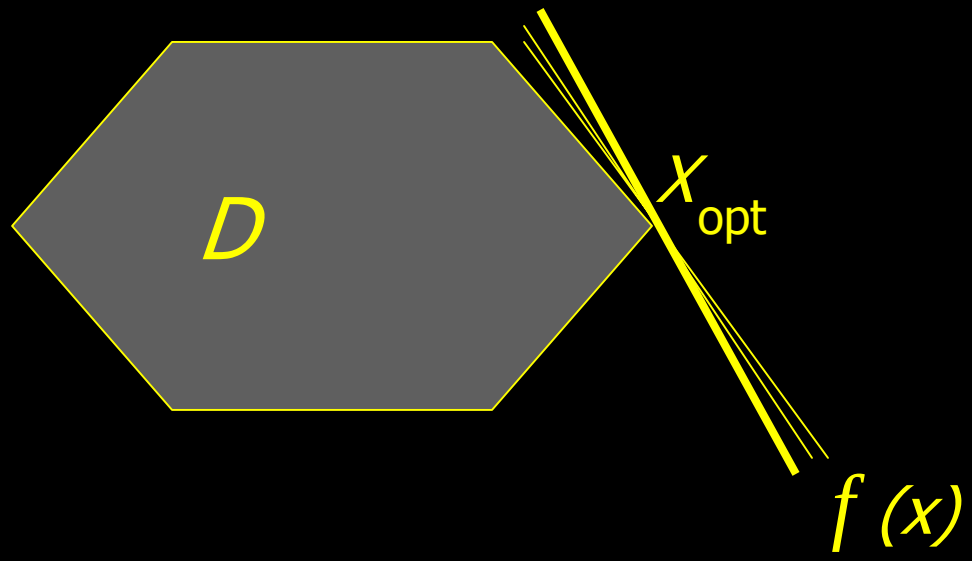
$$\max_{D} \sum_{i \in C} f_i(x)$$

Data mining can help in:

1. increasing the accuracy in the parameters of $f = \Sigma f_i$ that affect most critically the decision x (*data-driven sensitivity analysis*)

2. *segmenting* the population C of customers (*microeconomically-driven clustering*)

Data-driven sensitivity analysis



Associations which are likely to affect the values of parameters of *f* in a way that may change the optimum decision are "interesting."

Microeconomically-driven clustering

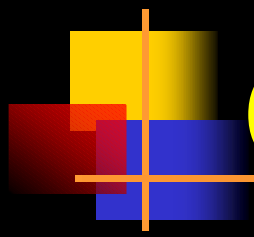
Recall: $\max_{D} \sum_{i \in C} f_i(x)$

$$\max_{x,\, y \in D} \sum_{i \in C} \max\{ f_i(x), f_i(y) \}$$

equal!

$$\max_{S \subseteq C} \left[\max_{x \in D} \sum_{i \in S} f_i(x) + \max_{x \in D} \sum_{i \notin S} f_i(x) \right]$$

cluster → $S \subseteq C$

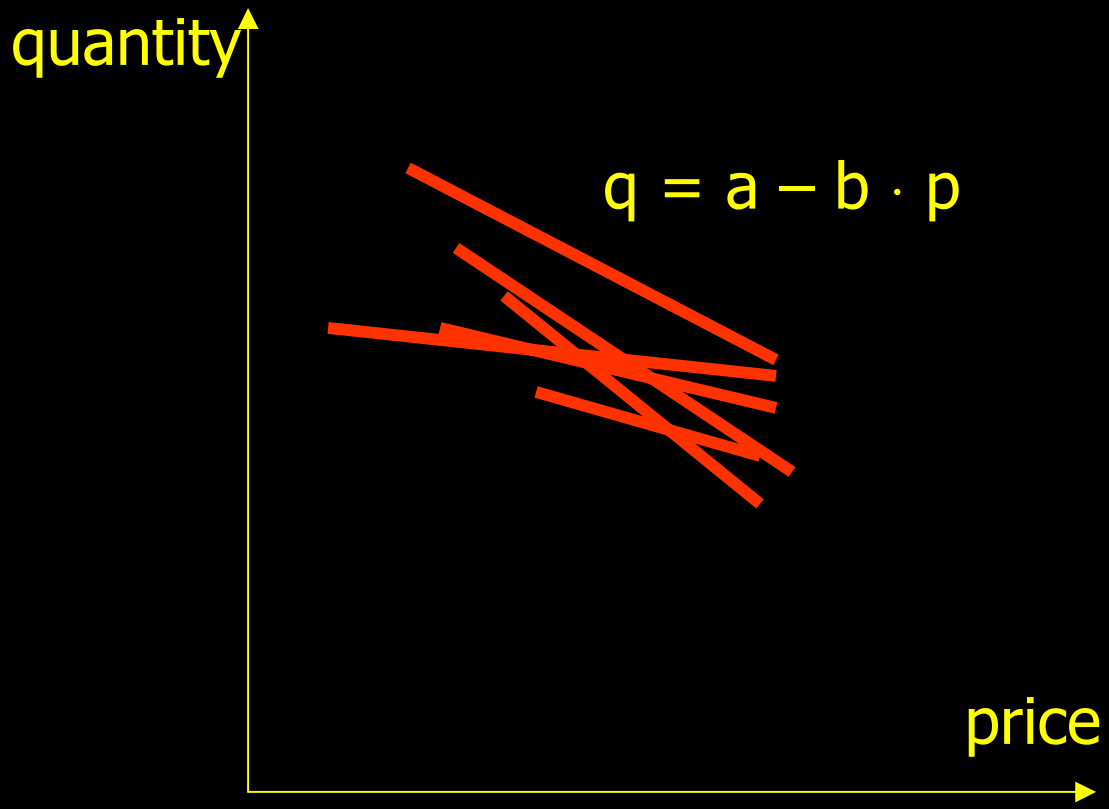


(A parenthesis: Clustering)

- before WW2, optimization was considered a single important problem, about to be solved
- simplex was criticized as too specialized
- since then, we realized that optimization is *many* problems: linear, convex, quadratic, multi-objective, unconstrained optimization, combinatorial and network optimization, NP-completeness… We solved the easy ones, and learned to live with the others.

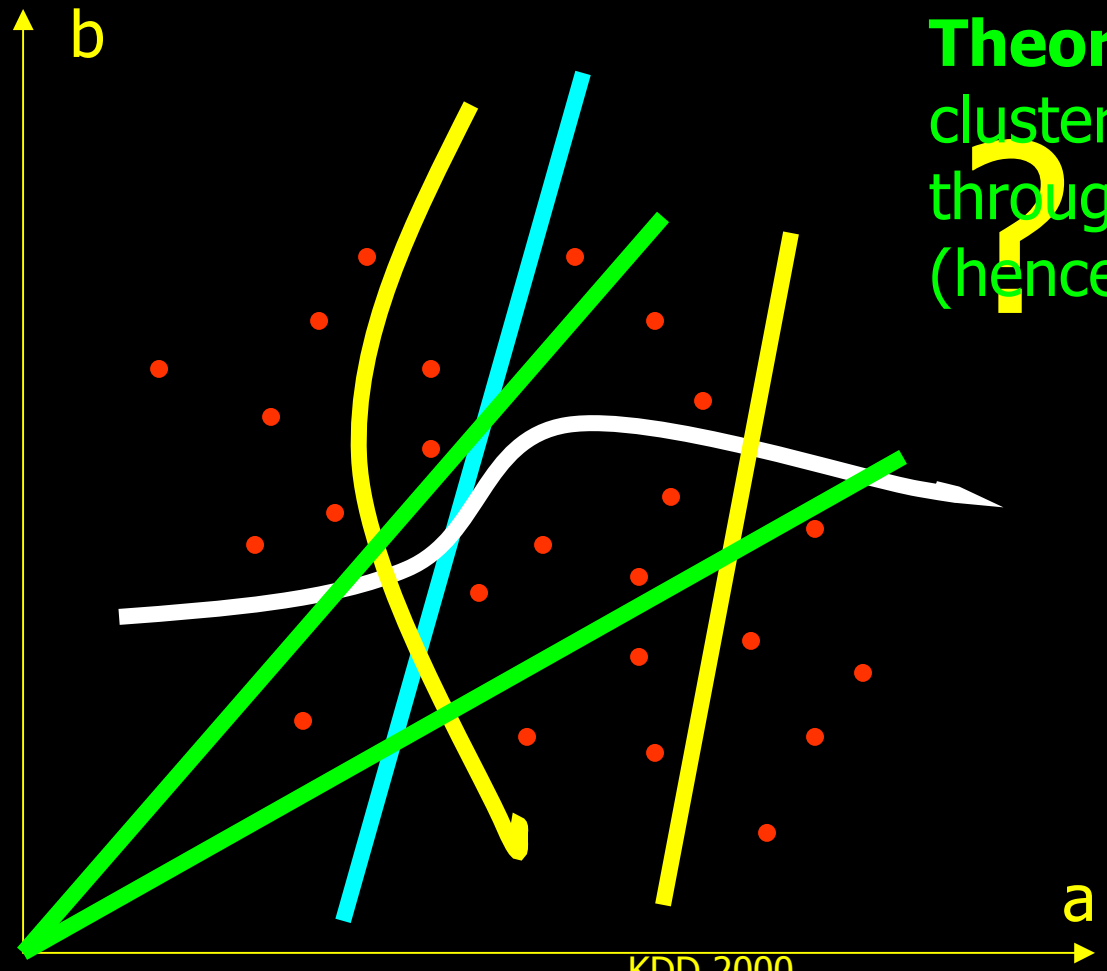
- is this a useful lesson of history for clustering research?

Example: market segmentation



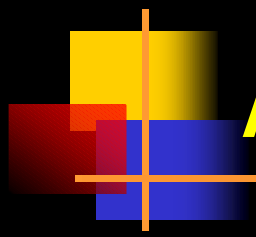
Goal: maximize revenue

or, in the a – b plane:



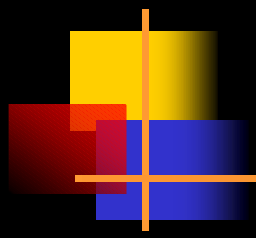
Theorem: Optimum clustering is by lines through the origin (hence: $O(n^2)$ DP)

?



Also:

- more than two segments
- variable no. of segments, each with an incremental cost
- if uncertainty above a threshold, segmentation beats customization
- mass customization
- a new genre of optimization problems
- rigorous use of sampling: approximation



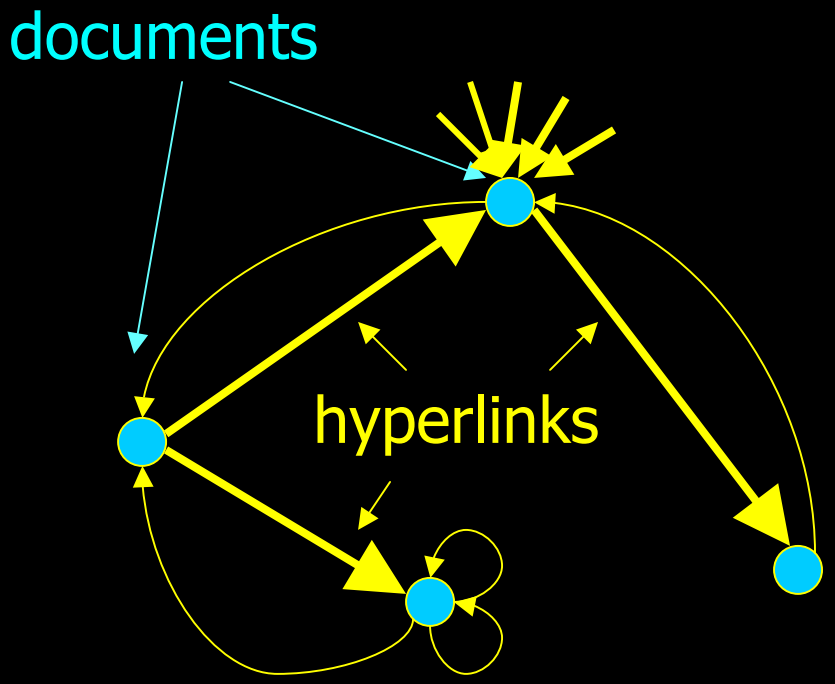
Talking of sampling…

- how do you sample the web?

  [Bar-Yossef, Berg, Chien, Fakcharoenphol, Weitz, VLDB 2000]

- e.g.: 42% of web documents are in html. *How do you find that?*
- *What is a "random" web document?*

The web as a graph
(cf. [Kleinberg 1998, IBM/CLEVER 1998-]



Idea: random walk

Problems:

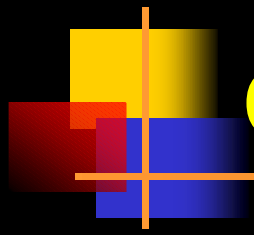
1. asymmetric ✓
2. uneven degree ✓
3. 2nd eigenvalue?

$\lambda = 0.99999$

the web walker: results

- [Jerrum-Sinclair 1989]:
  mixing time is ~log $N/(1-\lambda)$
- WW mixing time: 3,000,000
- actual WW mixing time: 100

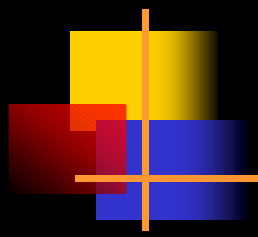
- .com 49%, .jp 9%,
  .edu 7%, .ch 0.8%



on privacy

- arguably the most crucial and far-reaching challenge and mission of the community
- least understood (e.g., is it rational?)
- www.sims.berkeley.edu/~hal, ~/pam, [Stanford Law Review, ca. 2000]

thoughts on privacy



- also an economic problem
- surrendering private information is either good or bad for you
- selling mailing lists vs. selling aggregate information: false dilemma
- the revelation principle: customized privacy
- proposed principle: *take into account the objective of the individual in data-mining*