6.883 Learning with Combinatorial Structure

Stefanie Jegelka

Organization

- participate: questions, discussion, … highly welcome! ☺

- class website: http://people.csail.mit.edu/stefje/fall15/
- Piazza for Q&A: piazza.com/mit/fall2015/6883 please sign up!
- Listeners: register to access class materials

- Grade: 45% homework, 45% project, 10% scribe
- Homework: ok to discuss in groups, but each person must hand in a solution & acknowledge collaborations

- TA: Zi Wang, office hours will be posted
- If you email me: put 6.883 in subject

Organization

- textbook & class material: no single one;
  material will be pointed out as we go
- we will discuss foundations & (very) recent research papers

Homework 0:

- fill out the survey
- sign up

What is this class about?

What is this class about?

- Recall: regression / classification

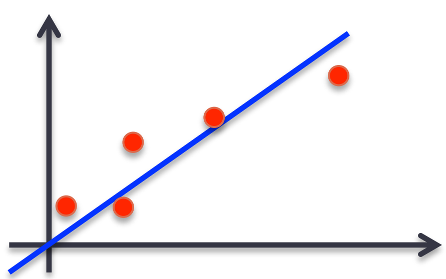
Observe samples $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$.

$x_i \in \mathbb{R}^d$. regression: $y_i \in \mathbb{R}$, classification: $y_i \in \{0, 1\}$

Problem: find a function $f \in \mathcal{F}$ that predicts y well: $\hat{y} = f(x)$

e.g. linear function $f(x) = w^\top x$

minimize $\mathbb{E}[\text{loss}(f(x), y)]$



1. we can do this (with enough samples)
2. … if loss is convex

Example 1: Structured prediction

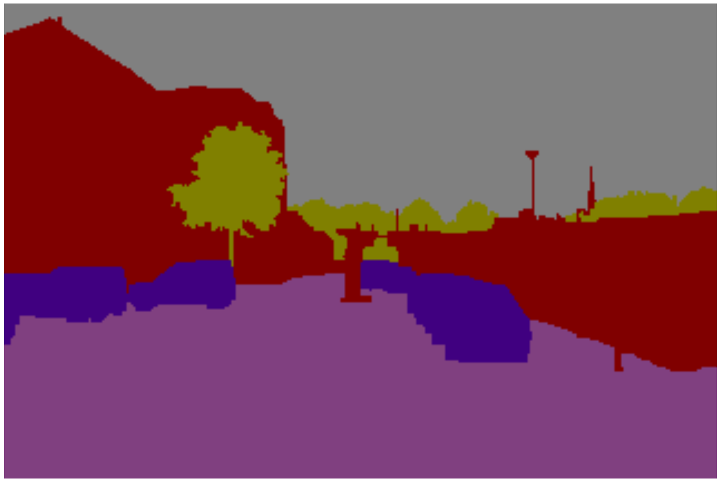
Observe samples $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$.

find a function that predicts y from x: $\hat{y} = f(x)$

What if y is not a scalar?

x

y



Structured prediction

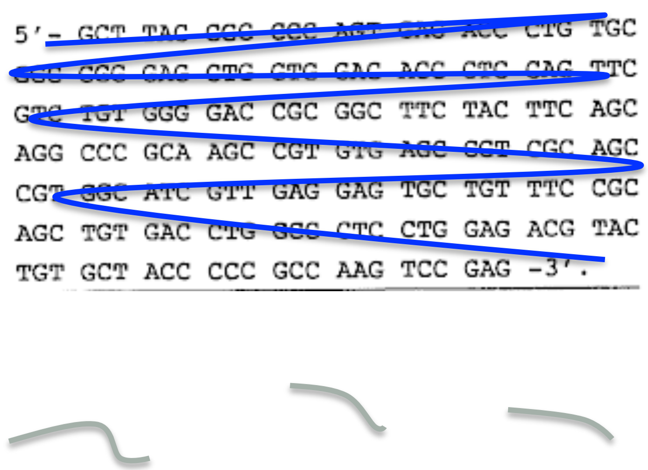
Observe samples $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$.

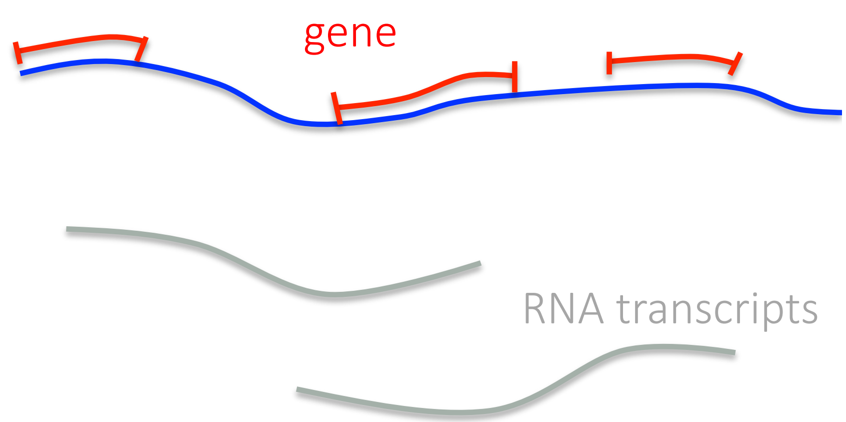
find a function that predicts y from x: $\hat{y} = f(x)$

What if y is not a scalar?

x

y





Structured prediction

Observe samples $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$.

find a function that predicts y from x: $\hat{y} = f(x)$

$$y = (y^1, \ldots, y^d), \quad y^i \in \mathcal{Y}^i$$

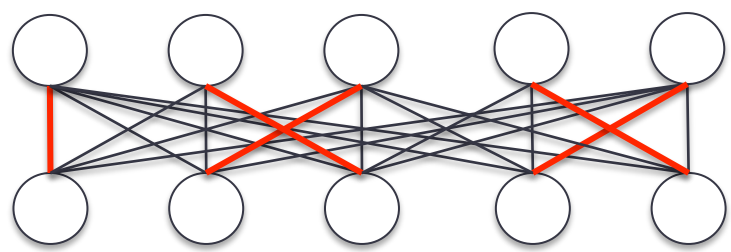
components y^i interdependent and often discrete

Example formulation:

$$f(x) = \arg\max_{y \in \mathcal{Y}} \; g(x, y)$$

e.g. $g(x, y) = p(y|x, \theta)$

$$g(x, y) = w^\top \phi(x, y)$$



$\mathcal{Y} =$ all matchings

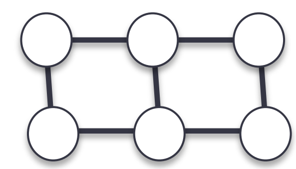
$$y = (1, 0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

Structured prediction

Observe samples $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$.

find a function that predicts y from x: $\hat{y} = f(x)$

$y = (y^1, \ldots, y^d), \quad y^i \in \mathcal{Y}^i$



components y^i interdependent and often discrete

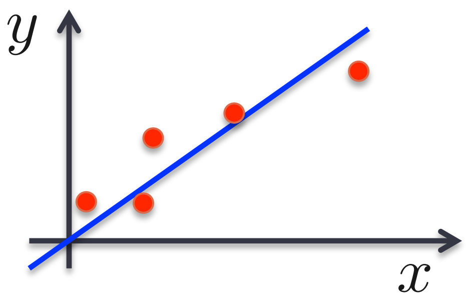
Example formulation:

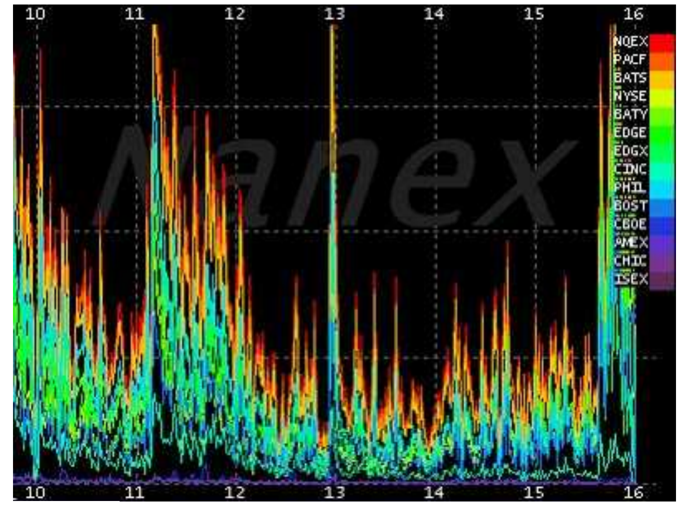
$$\underbrace{f(x) = \underbrace{\arg\max_{y \in \mathcal{Y}} \; g(x, y)}_{\substack{\text{combinatorial optimization problem} \\ \text{-- can we solve this?}}}}_{\text{can we learn such a function?}}$$

e.g. $g(x, y) = p(y | x, \theta)$

$g(x, y) = w^\top \phi(x, y)$

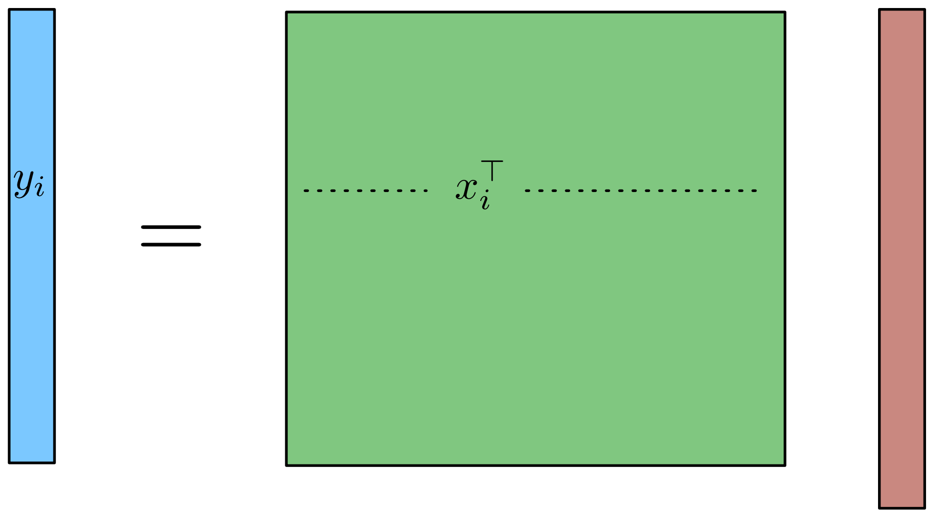
Another example





Example 2: High-dimensional data

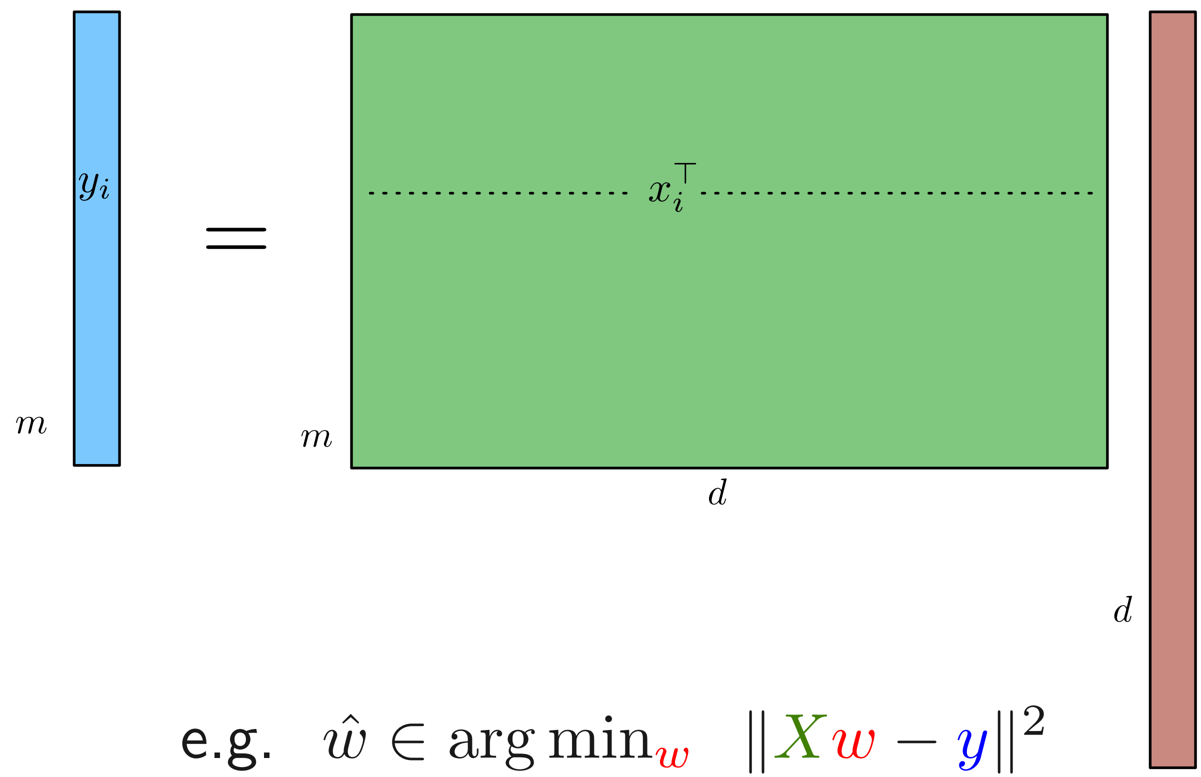
$$y \; = \; Xw \; + \; \epsilon$$



e.g. $\hat{w} \in \arg\min_{w} \quad \|Xw - y\|^2$

High-dimensional data

$$y = Xw + \epsilon$$

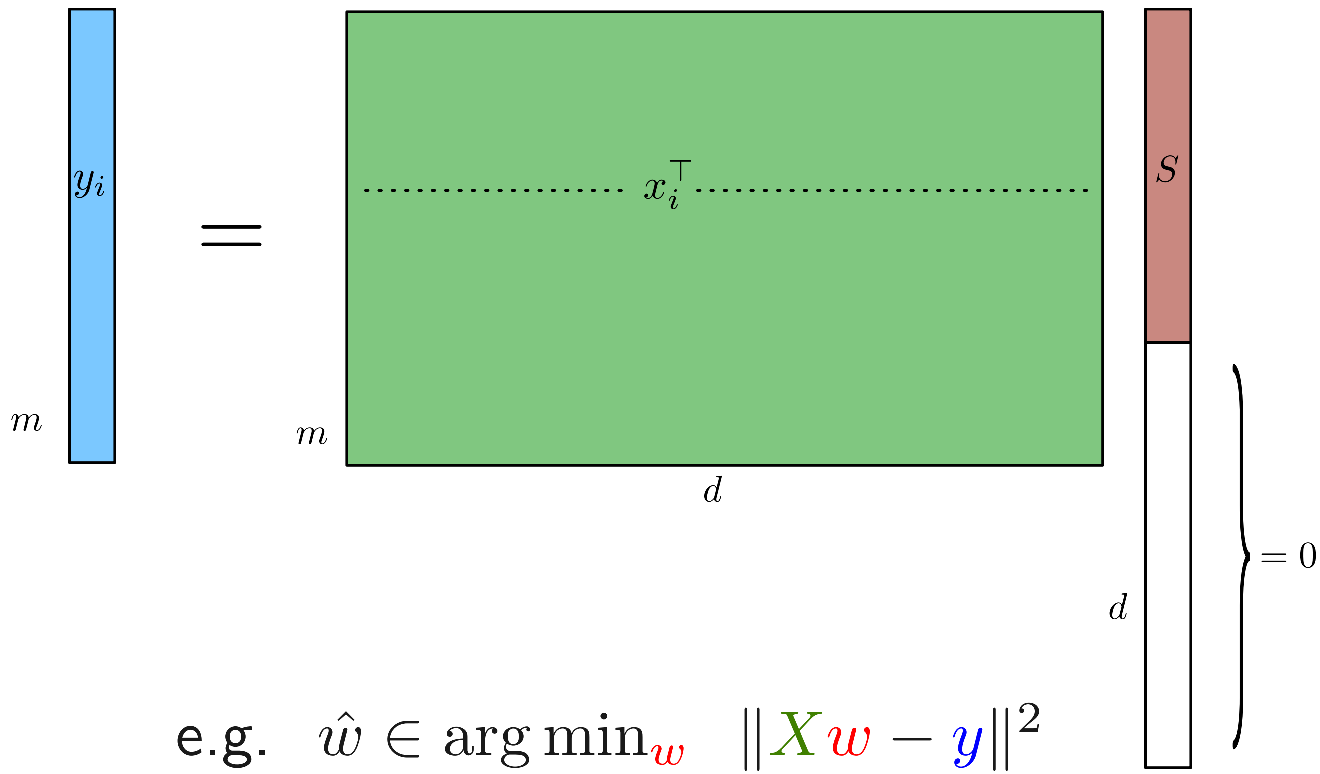


e.g. $\hat{w} \in \arg\min_{w} \quad \|Xw - y\|^2$

without noise: infinitely many solutions! (a whole subspace)

High-dimensional data

$$y \;=\; Xw \;+\; \epsilon$$



e.g. $\hat{w} \in \arg\min_{w} \quad \|Xw - y\|^2$

without noise: infinitely many solutions! (a whole subspace)

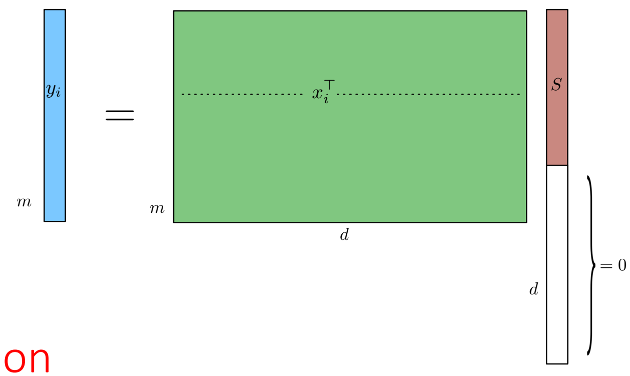
key insight: w is not arbitrary – only *k* nonzero entries

High-dimensional data

$$y \;=\; Xw \;+\; \epsilon$$

e.g. $\hat{w} \in \arg\min_w \quad \|Xw - y\|^2$

s.t. $\|w\|_0 \leq k$

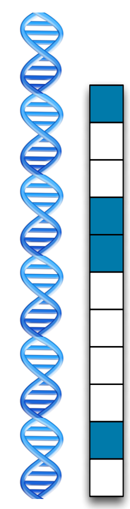


combinatorial optimization problem!

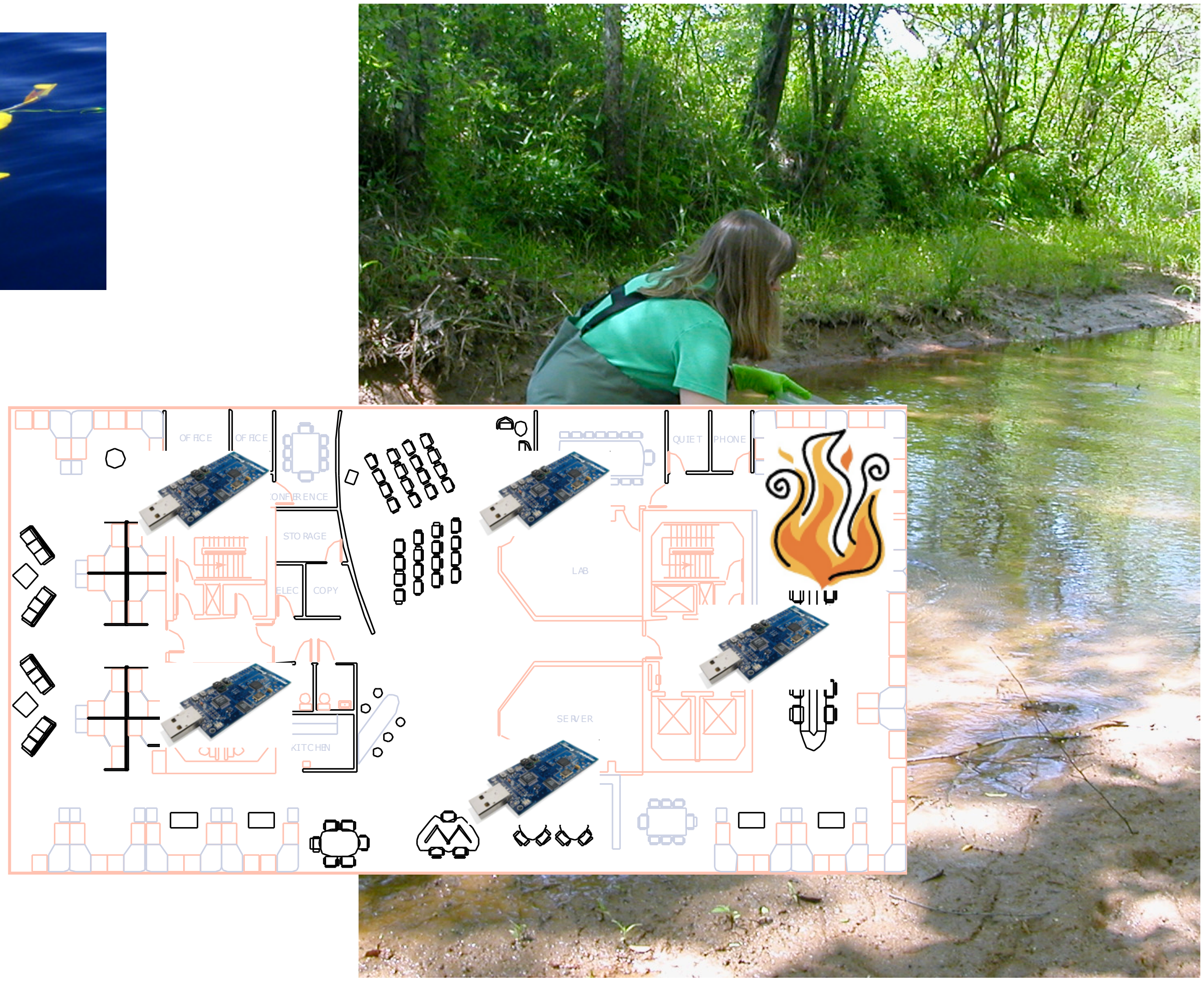
More generally:

- signal is a sparse combination of "atoms"
  low-rank matrix, sparse graph, …
- signal has certain sparsity pattern

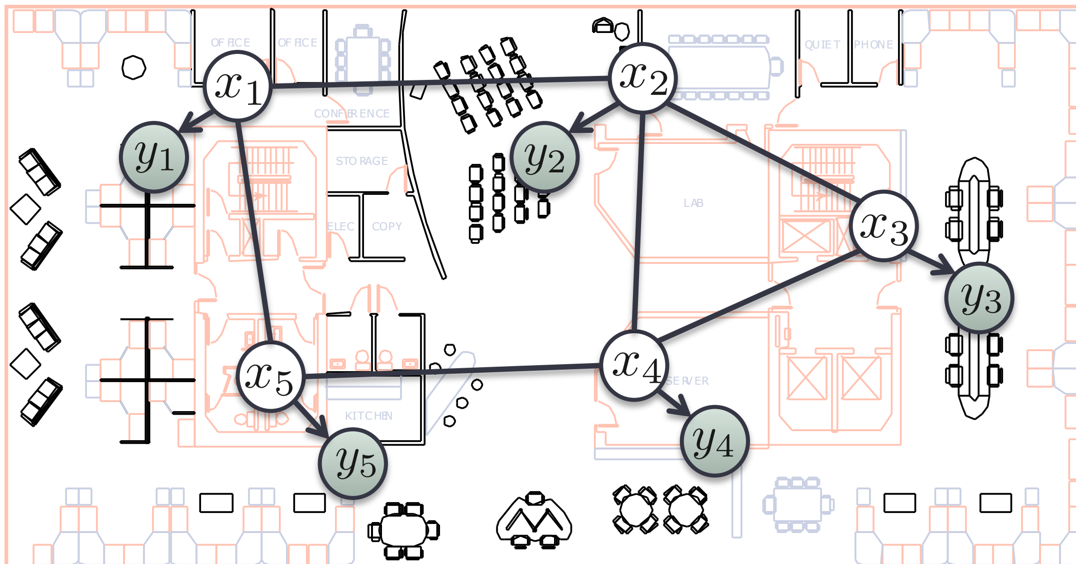
How phrase this?



Example 3: Sensing & monitoring



Sensing & monitoring



$$y = x + \epsilon$$

$$p(x, y) = p(y_1, \ldots, y_n)\, p(x_1, \ldots, x_n | y_1, \ldots, y_n)$$

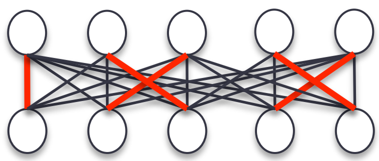
pick set A of locations to maximize

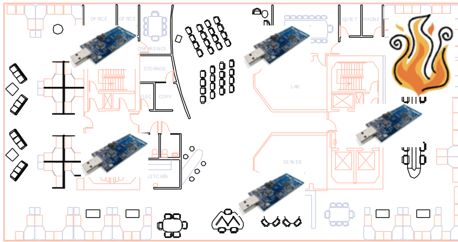
$$F(A) = I(\mathsf{x}; \mathsf{y}_A) = H(\mathsf{x}) - H(\mathsf{x} | \mathsf{y}_A)$$

uncertainty before observing

uncertainty after observing

Recap ...





ground set of items $\mathcal{V}$

subsets $S \subseteq \mathcal{V}$

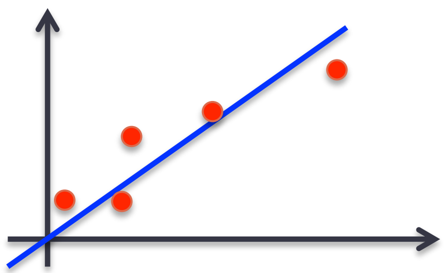
constraints: $|S| \leq k, \quad S$ a matching, $\ldots$

set function: $F(S) \qquad F : 2^{\mathcal{V}} \to \mathbb{R}_+$

- Mathematical models?
- Algorithms?
- Analysis?

this may be hard ... ☹

What is "nice"?



convex loss function
over convex set

$$\min_{w} \sum_{i=1}^{n} (w^\top x_i - y_i)^2$$

Can we still use convex optimization? How?

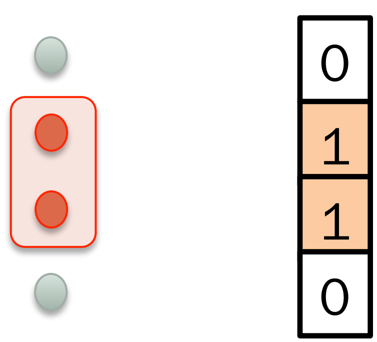
We'll learn about appropriate algorithms & their analysis

What makes our setting easier?

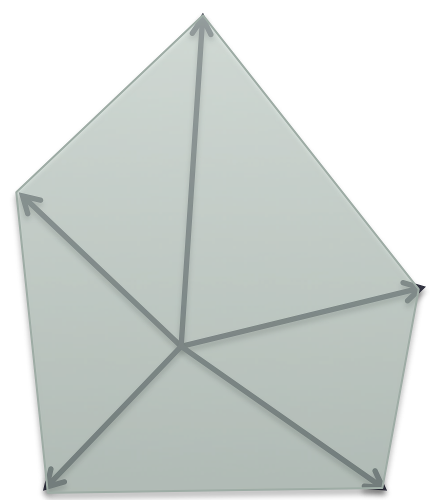
Mathematical structure.

- Some combinatorial problems are nice. Which?

- Relaxations

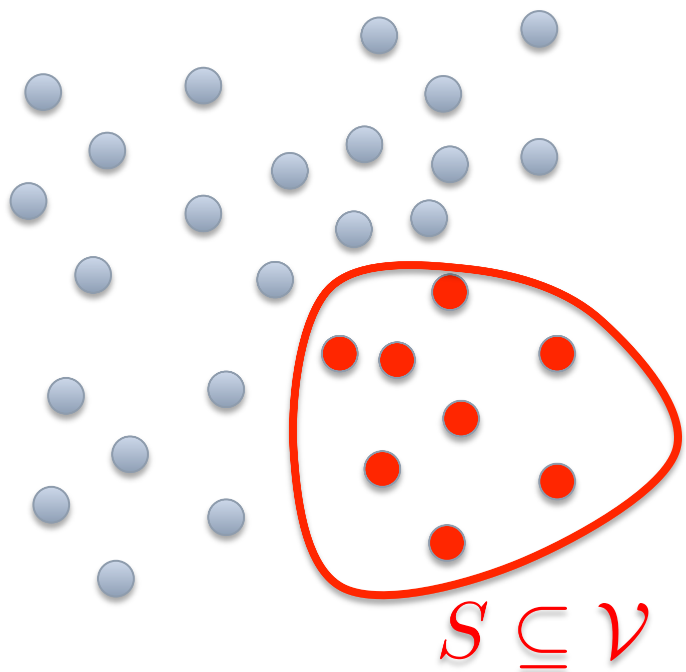


$S \subseteq \mathcal{V}$ $\quad$ $x = 1_S$



Is this easier?
It depends.
Geometry important
for statistics &
computation

Set functions



$$F(S)$$

"discrete analog of convexity"?

Convex functions *(Lovász, 1983)*

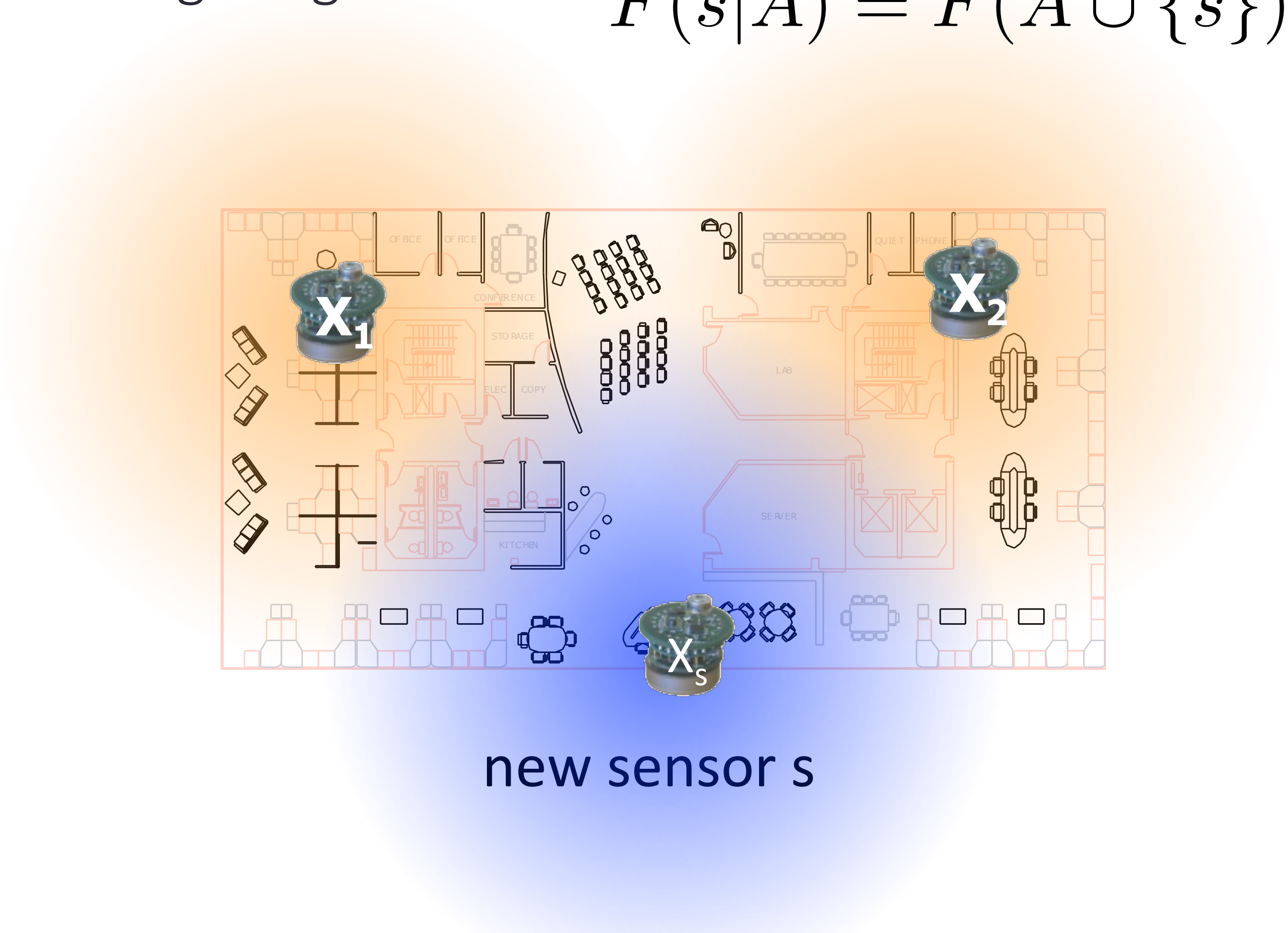
- "**occur in many models** in economy, engineering and other sciences", "often the only nontrivial property that can be stated in general"
- **preserved** under many operations and transformations: larger effective range of results
- sufficient structure for a "mathematically beautiful and practically useful **theory**"
- efficient **minimization**

> "It is less apparent, but we claim and hope to prove to a certain extent, that a similar role is played in discrete optimization by ***submodular set-functions***" [...]
> they share the above four properties.

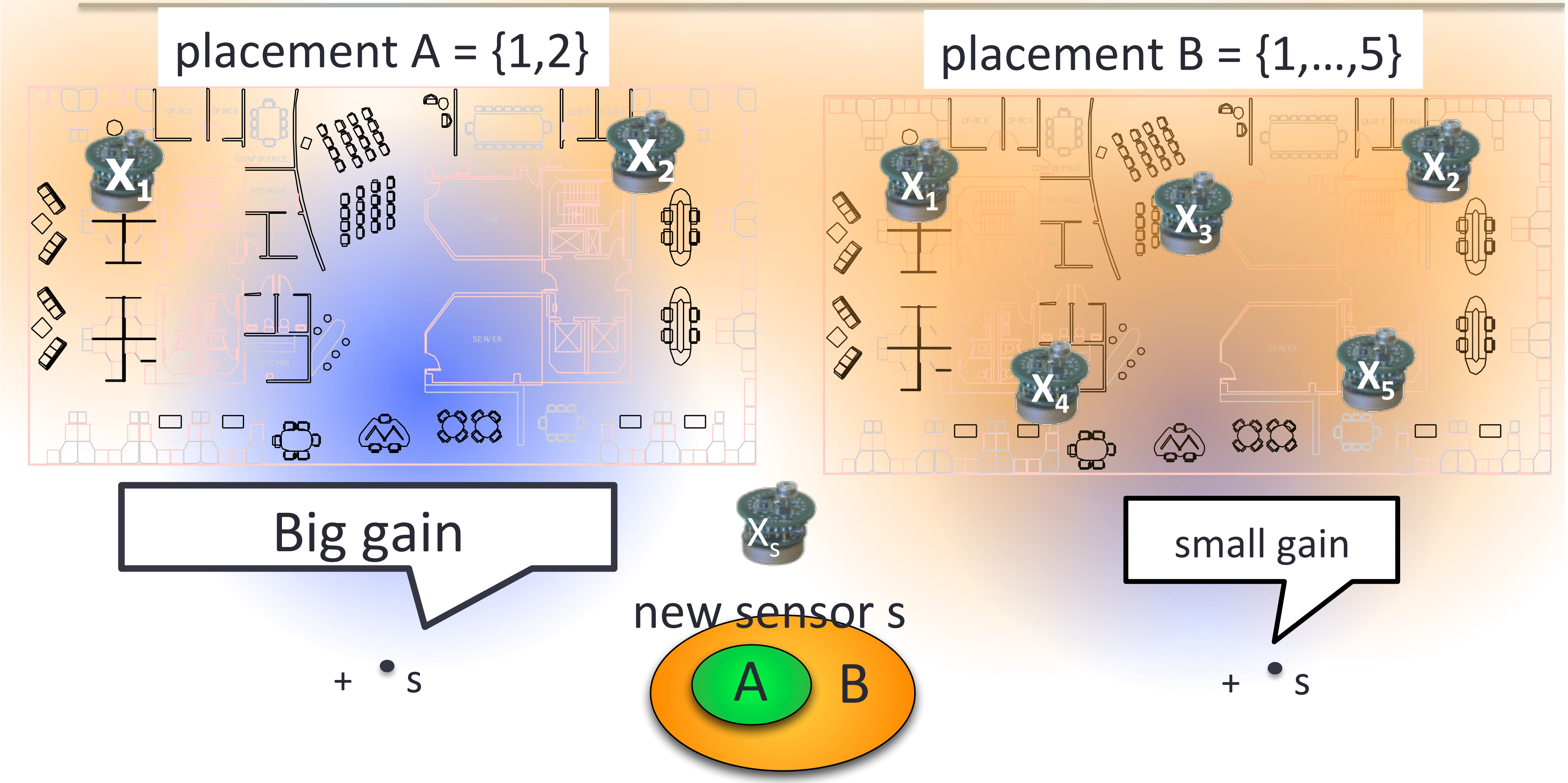
Marginal gain

- Given set function $F : 2^V \rightarrow \mathbb{R}$

- Marginal gain: $F(s|A) = F(A \cup \{s\}) - F(A)$



Diminishing marginal gains

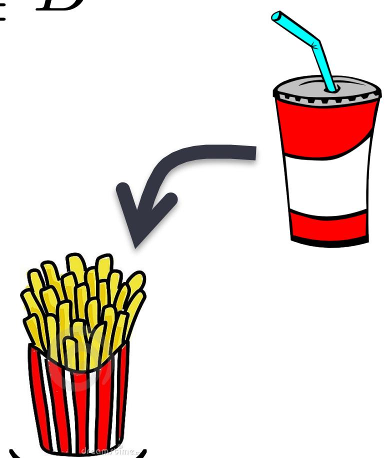


$$A \subseteq B$$

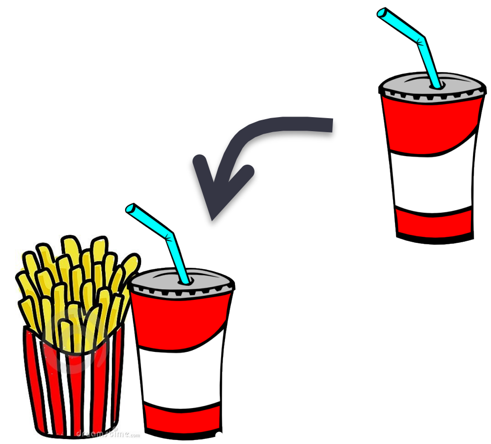
$$F(A \cup s) - F(A) \quad \geq \quad F(B \cup s) - F(B)$$

Submodularity

$A \subseteq B$



A



B

$$F(A \cup s) - F(A) \quad \geq \quad F(B \cup s) - F(B)$$

extra cost:
one drink

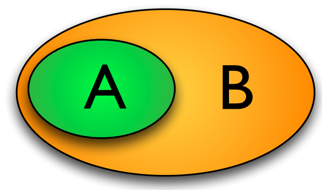
extra cost:
free refill ☺

diminishing marginal costs

Supermodular set functions

- Submodularity: diminishing marginal gains

$$F(A \cup e) - F(A) \quad \geq \quad F(B \cup e) - F(B)$$



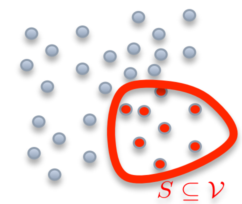
- Supermodularity: increasing marginal gains

$$F(A \cup e) - F(A) \quad \leq \quad F(B \cup e) - F(B)$$

$\emptyset$

Why is submodularity useful?

- occurs in many learning problems:
  rank, independence, diversity, cohesion, graphs, ...

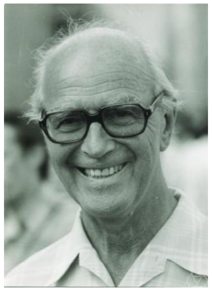


- associated with very “nice” polyhedra

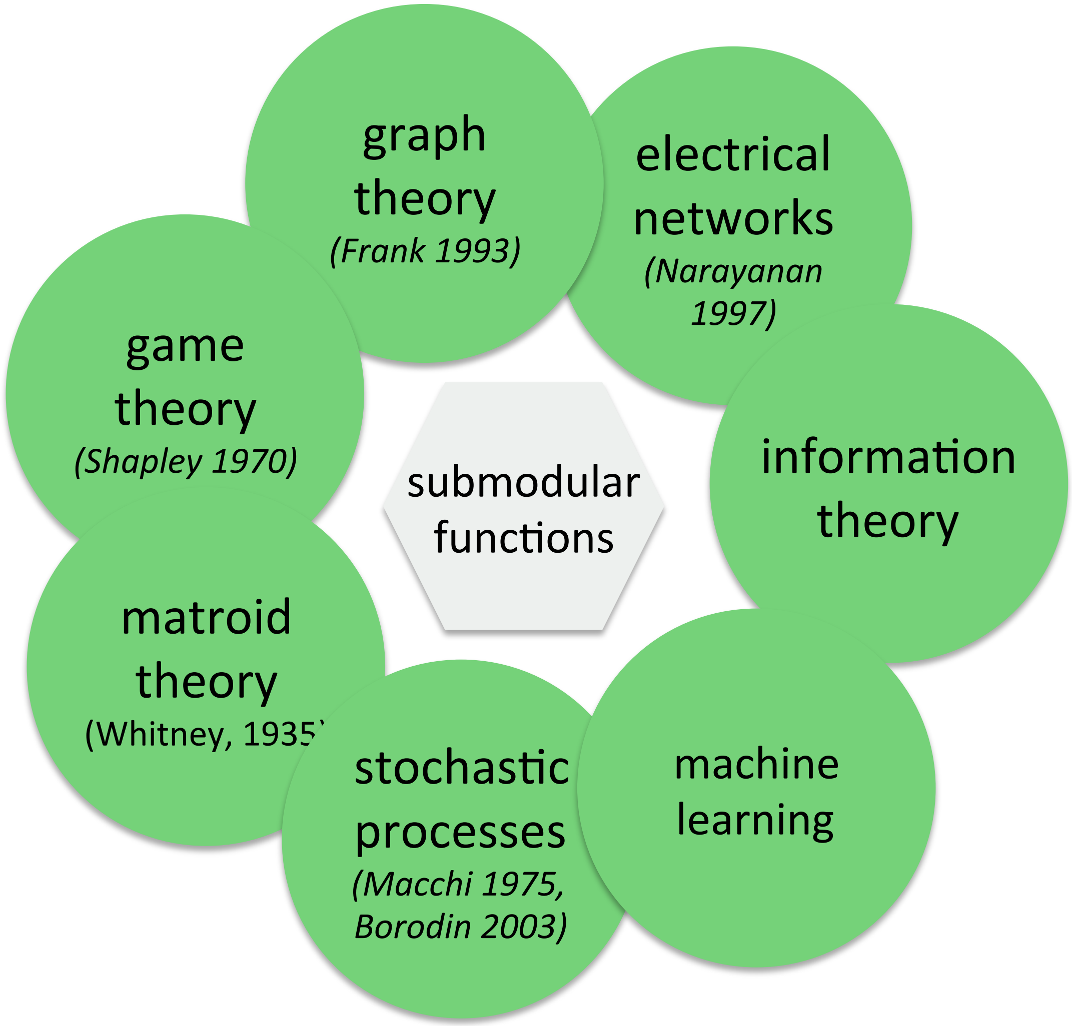
- close connections to convexity

- optimization: convex optimization, greedy algorithms, ...

The big picture


G. Choquet

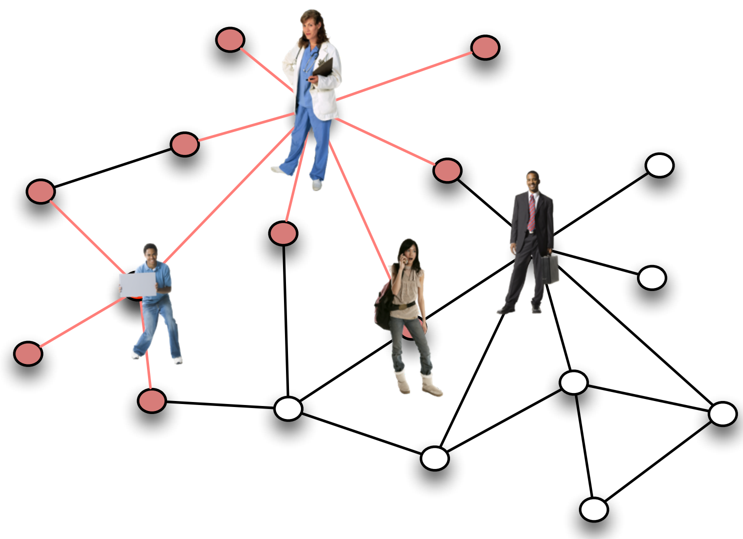

J. Edmonds



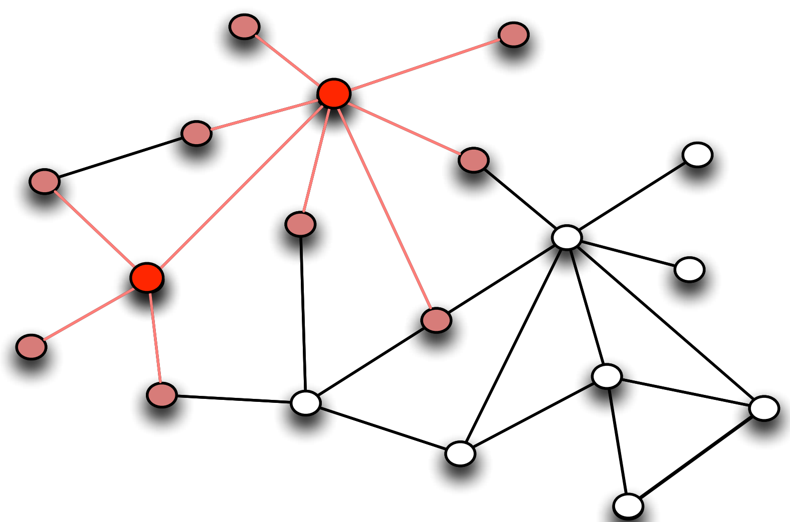

L.S. Shapley


L. Lovász

Diffusion processes on graphs



Diffusion processes on graphs



- information propagates
- node v becomes active if random threshold exceeded:

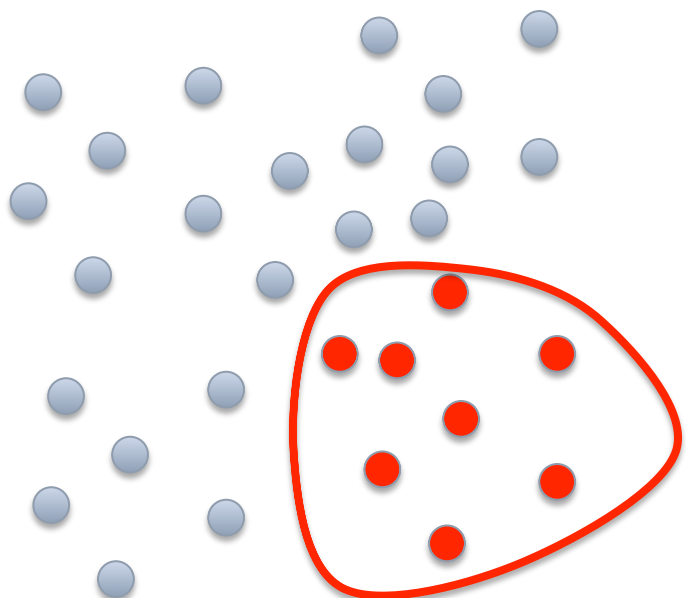
$$a_v(N_v) \geq \theta$$

activation function

set of active neighbors

- # active nodes after t steps?
- Which set of nodes is most influential?

Set functions … and point processes



$S \subseteq \mathcal{V}$

so far:

set function $F : 2^{\mathcal{V}} \to \mathbb{R}_+$

$\min \mid \max \quad F(S)$

Point process:

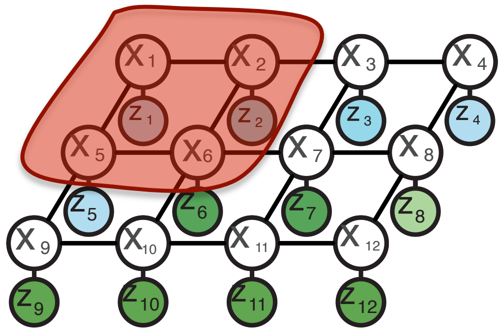
distribution over subsets:

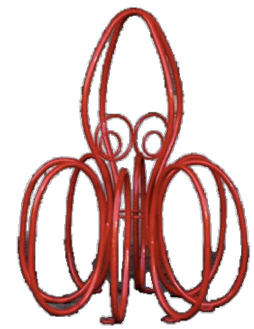
$P(S)$

Questions:

- mode?
- marginal probabilities?
- sampling?
- learning? …

Point processes -- examples





$$P(x|z) \propto P(z|x)\, P(x)$$

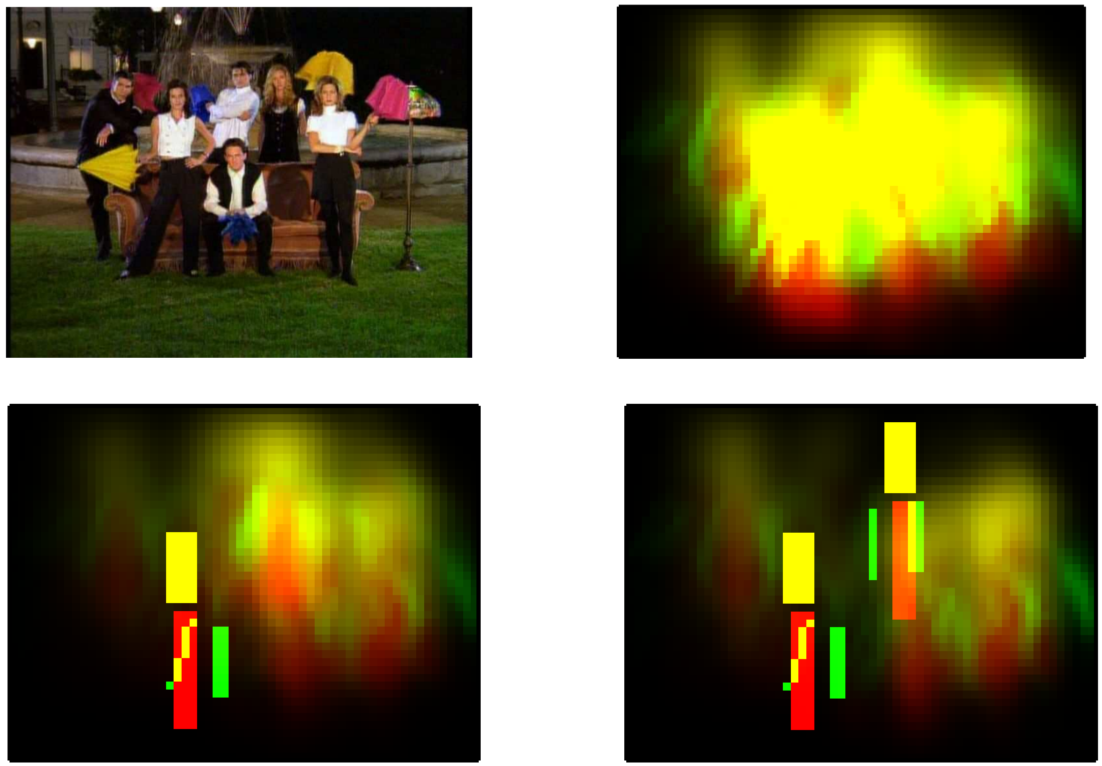
labels

pixel values

$$x \in \{0,1\}^n$$

would like: nearby points are both selected or not selected
spatial coherence, “attractive” --- positive correlations

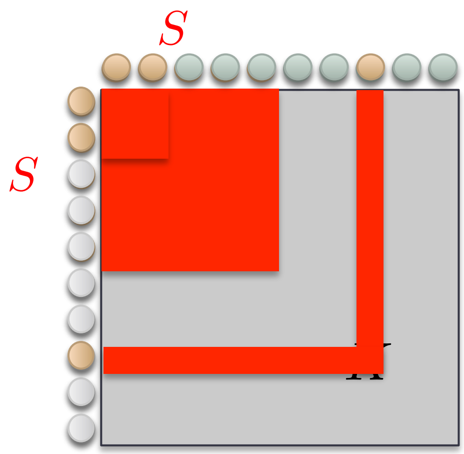
Diversity



$$P(S \mid \mathrm{data}) \;\propto\; P(S)\; P(\mathrm{data} \mid S)$$

would like: “spread out”, repulsion, diversity

Determinantal point processes



- normalized similarity matrix K
- sample Y:

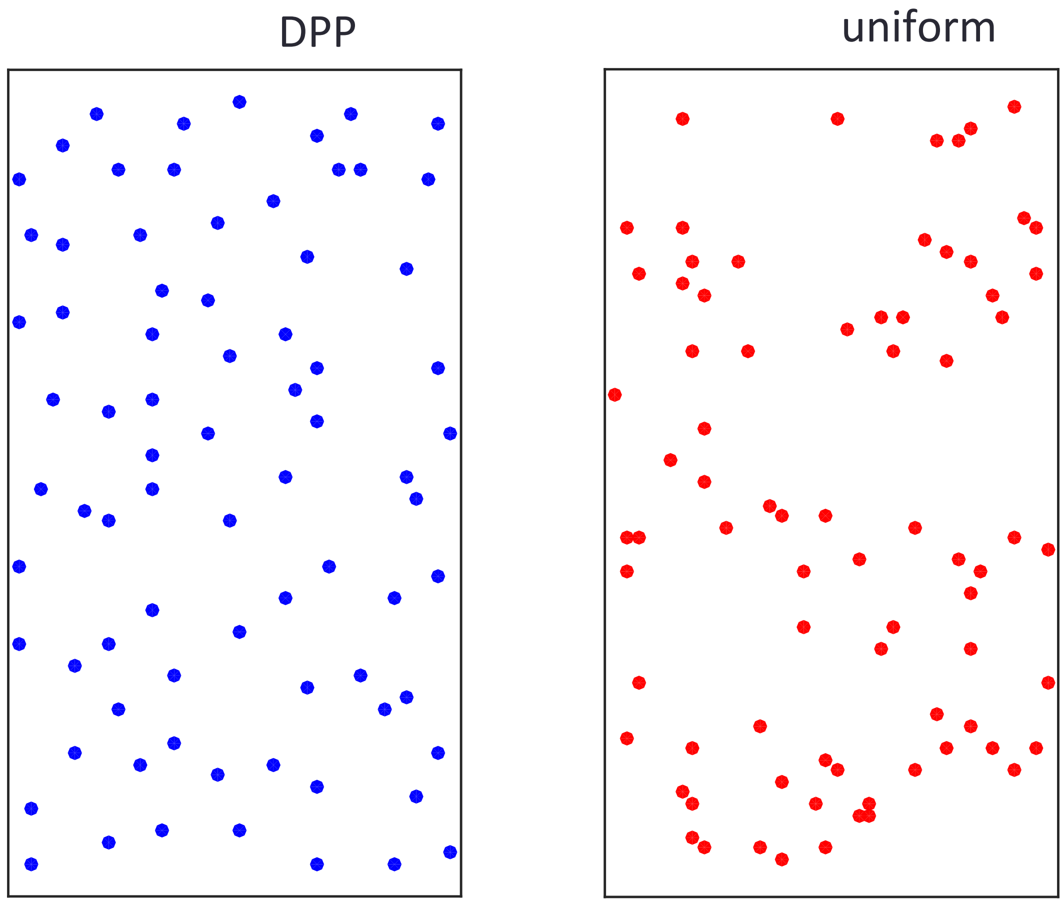
$$P(S \subseteq Y) = \det(K_S)$$

$$P(e_i \in Y) = K_{ii}$$

$$\begin{aligned} P(e_i, e_j \in Y) &= K_{ii}K_{jj} - K_{ij}^2 \\ &= P(e_i \in Y)P(e_j \in Y) - K_{ij}^2 \end{aligned}$$

repulsion

DPP sample



similarities:

$$s_{ij} = \exp(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2)$$

$$\sigma^2 = 35$$

Why is this useful?

- representation makes many things closed form / tractable: linear algebra.

Representations…

- set functions
- graphs
- convex functions
- polyhedra
- determinants
- polynomials

Common questions

- combinatorial predictions
- combinatorial regularization
- selecting informative subsets
- processes defined by combinatorial objects
- point processes …

- Mathematical models? how phrase as inference/learning/optimization problem?
- Algorithms? Convex, combinatorial optimization.
- Analysis? Is this tractable? Can we do learning? Can we give any guarantees? How much time will this take? …

Coarse syllabus

Properties & Algorithms

- basic convex analysis
  convex sets & functions, norms, subdifferentials, optimality conditions, duality
- convex optimization
  non-smooth optimization, conditional gradient method, proximal gradient, splitting & dual decomposition & others
- submodularity & convexity
- submodular maximization
- scalability
- determinantal point processes
- online learning

Formulations & Applications

- structured prediction
- combinatorial norms & regularization
- spread of influence, diversity, information gain, point processes, …

Lots of connections

Example: the same algorithm for:

- Learning to predict structures (structured prediction)
- Finding the minimum of a submodular set function
- Generate “pseudo-samples” to approximate moments
- Learning with many sparsity- and low-rank inducing regularizers
- Learning with combinatorial norms
- Finding the mode (MAP) for certain graphical models
- Approximating partition functions
- …

… after suitable formulation ☺

Goals of the class

- understand formulations of combinatorial learning problems
  be able to formulate problems mathematically
- understand underlying mathematical principles
  (these are often shared among many problems – surprisingly many connections!)
  be able to recognize mathematical structure to exploit
- understand algorithmic techniques & their connections: what applies? why do they work?
  be able to select, derive & analyze appropriate algorithms
- Have fun with some beautiful math!
- ➔ basis to explore and play with it on your own!! ☺

Upcoming seminars

- MIT-MSR Machine Learning Seminar
  Andreas Krause (ETH): Inference and Learning with Probabilistic Submodular Models
  Thursday Sep 10, 4pm, 32-G449

- Stochastics & Statistics Seminar
  Robert Freund (MIT): An extended Frank-Wolfe Method with Application to Low-Rank Matrix Completion
  Friday, Sep 11, 11am, 32-124

- Algorithms & Complexity Seminar
  Morteza Zadimoghaddam: Randomized Composable Core-sets for Distributed Submodular and Diversity Maximization
  Friday, Sep 11, 4pm, 32-G575