Massachusetts
Institute of
Technology

Submodularity

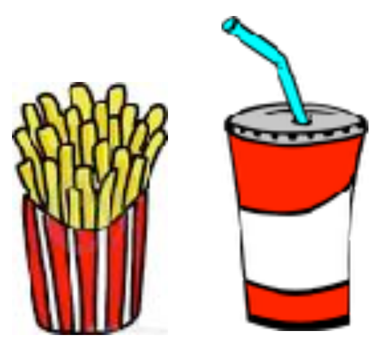
Stefanie Jegelka, MIT
MLSS London, 2019

Set functions

ground set

$\mathcal{V} = \{$  $\}$

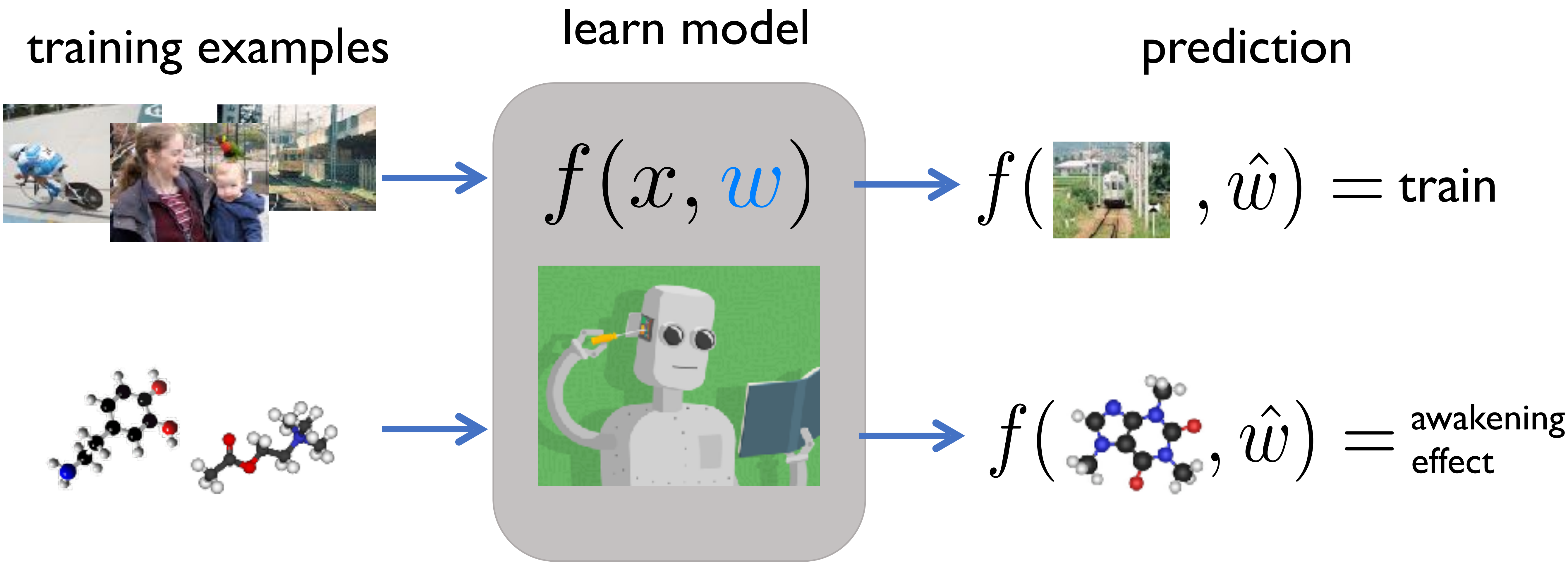
$$F : 2^{\mathcal{V}} \to \mathbb{R}$$

$F($  $)$ = cost of buying items together, or

utility, or

probability, …

Subset selection in Machine Learning



Subset selection in Machine Learning

training examples

learn model

$f(x, w)$

prediction

$f($ $, \hat{w}) =$ train

Informative Subsets

- Compression
- Summarization

- Placing sensors
- Designing experiments

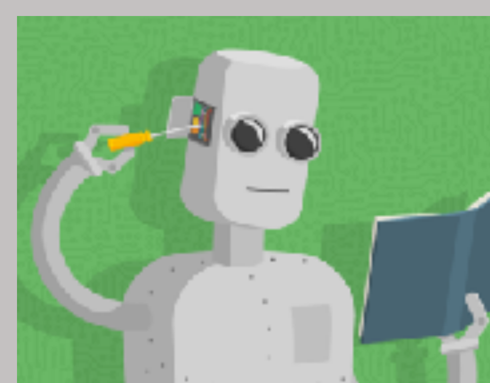
$$F(S) = \text{“information”}$$

Subset selection in Machine Learning

training examples

learn model

$$f(x, w)$$



prediction

$f($  $, \hat{w}) =$ train

Variable (Coordinate) Selection

Only use few coordinates of x in $f(x, w)$

$$f(x, w) = \sum_{i=1}^{d} w_i x_i$$

$$F(S) = \text{“coherence”}$$

Subset selection in Machine Learning

training examples

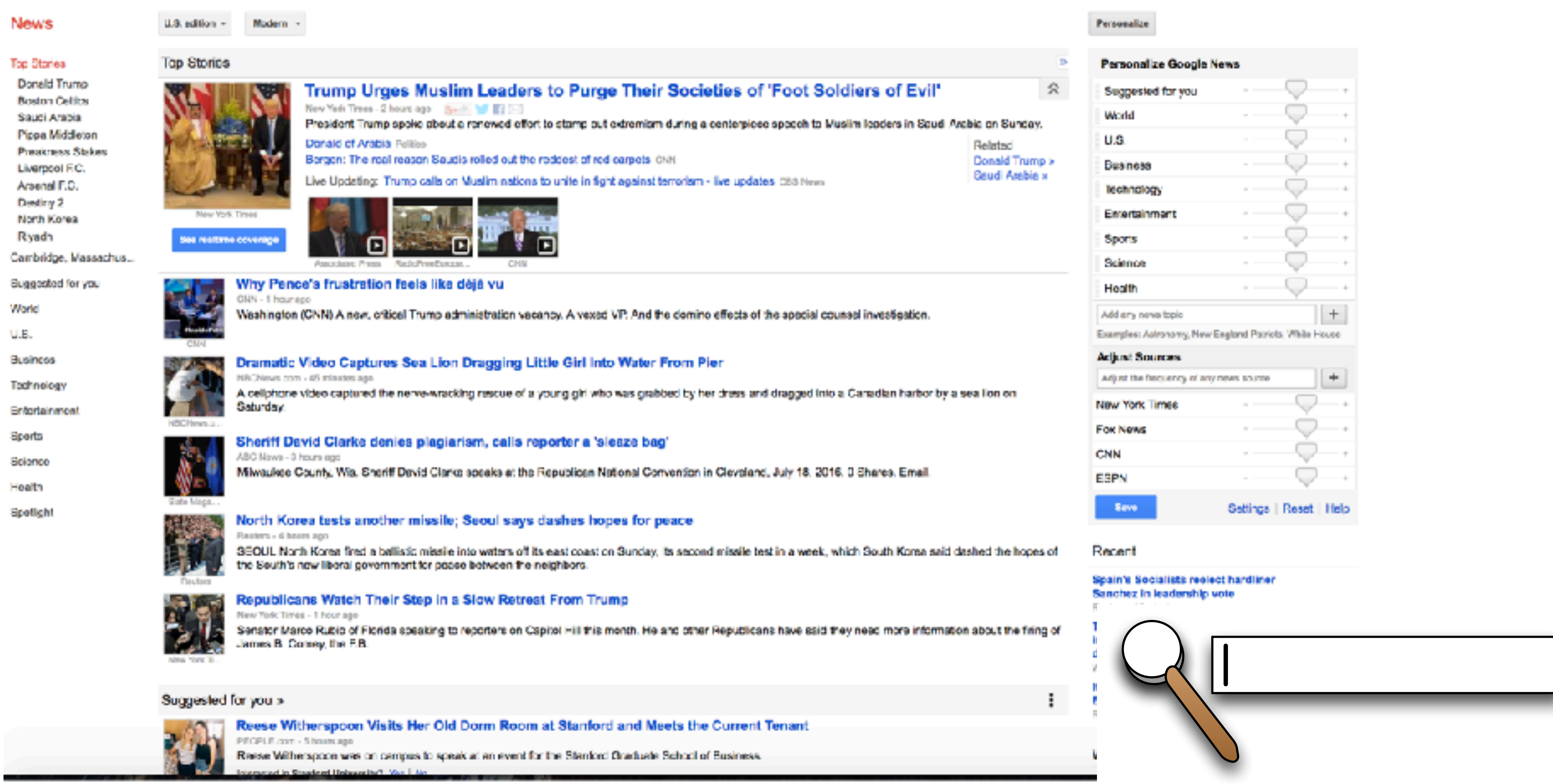
learn model

$f(x, w)$

prediction

$f($  $, \hat{w}) =$ train

Summarization & Recommendation



$$F(S) = \text{relevance} + \text{diversity or coverage}$$

Common formalization

Find a subset S that
maximizes / minimizes a set function $F(S)$

Can be difficult in general: exponentially many subsets!

Rescue: mathematical structure — **submodularity**

Why is submodularity interesting?

Importance of convex functions *(Lovász, 1983):*

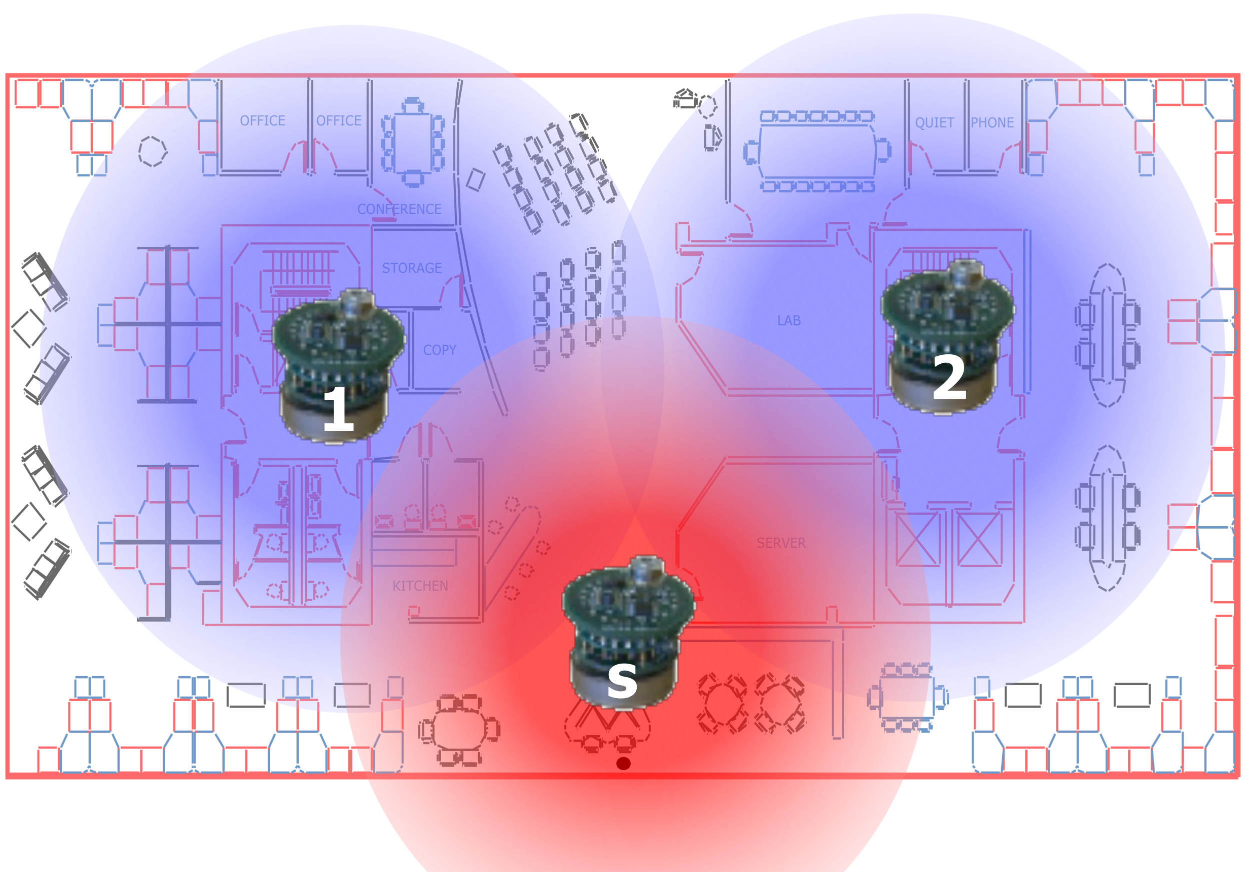
- *"occur in many models in economy, engineering and other sciences", "often the only nontrivial property that can be stated in general"*
- *preserved under many operations and transformations: larger effective range of results*
- *sufficient structure for a "mathematically beautiful and practically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

Marginal gain

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

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



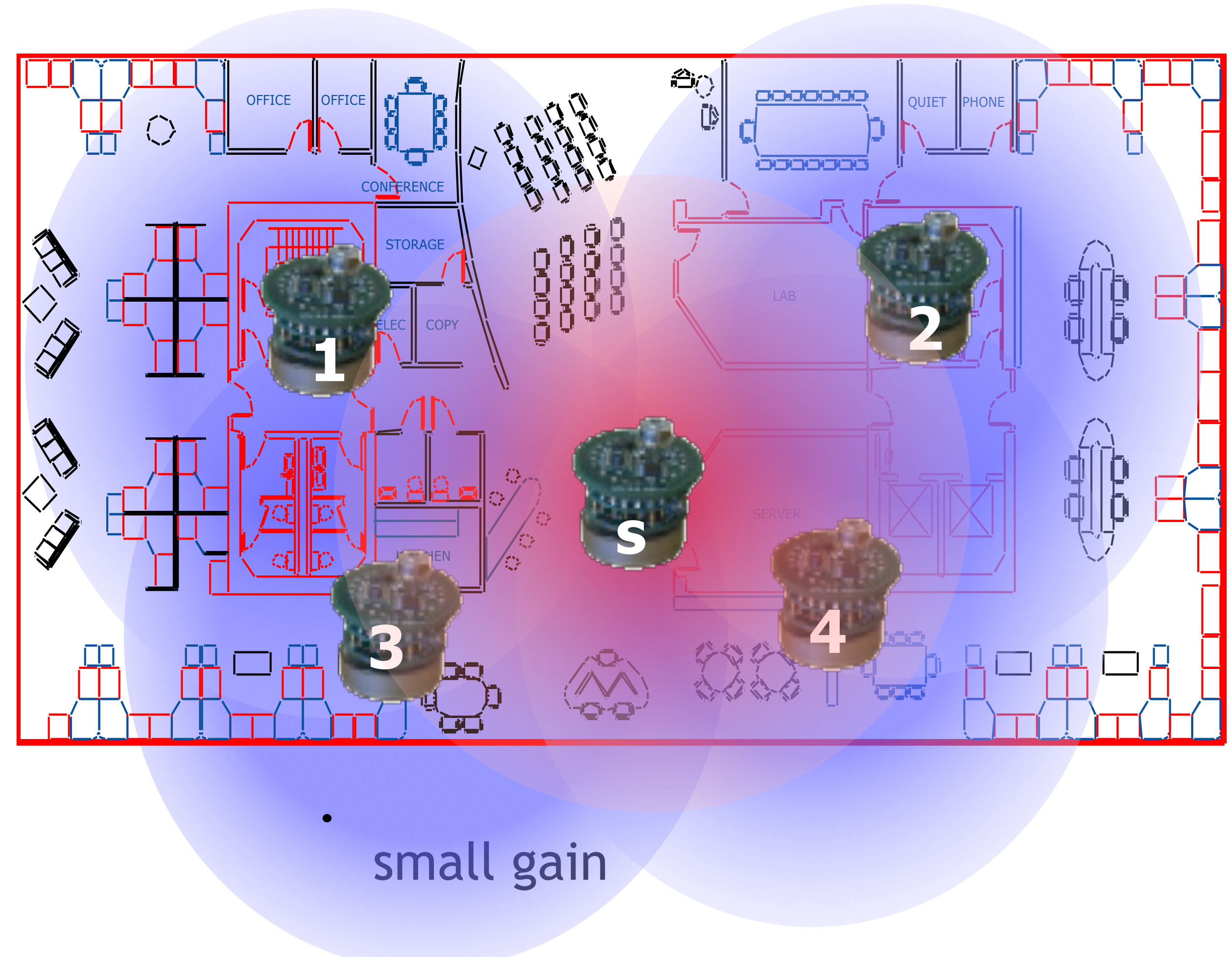
new sensor s

Diminishing gains

placement A = {1,2}

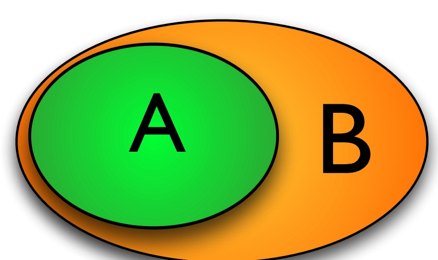
Big gain

placement B = {1,2,3,4}



small gain

for all $A \subseteq B$
and s not in B



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

Submodularity

set function: $F(S)$

- diminishing returns: $\forall S \subseteq T,\ a \notin T$

$$\boxed{F(S \cup \{a\})} - \boxed{F(S)} \geq \boxed{F(T \cup \{a\})} - \boxed{F(T)}$$

- equivalent general definition: $\forall\ A, B \subseteq V$

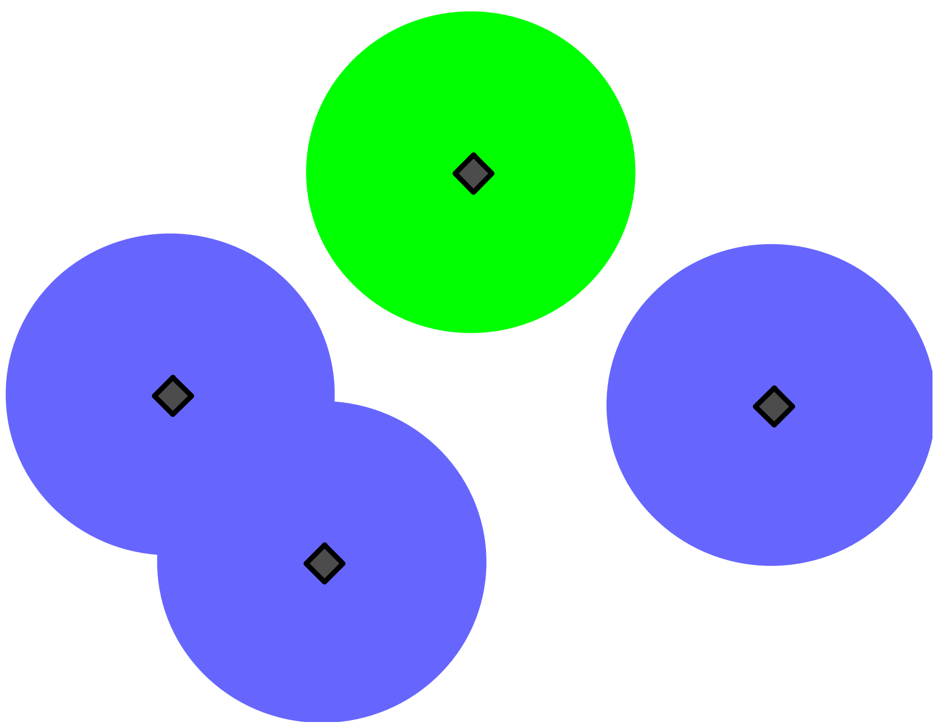
$$\boxed{F(A)} + \boxed{F(B)} \geq \boxed{F(A \cup B)} + \boxed{F(A \cap B)}$$

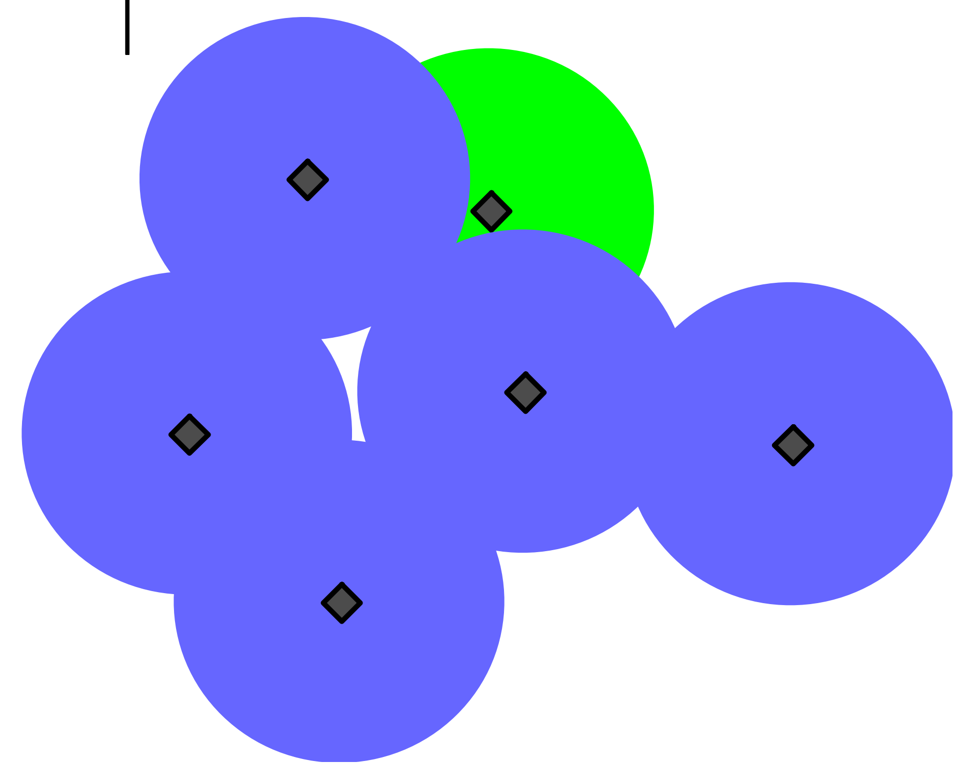
Roadmap

- Examples of submodular functions
- Maximizing submodular functions
- Minimizing submodular functions
- Further connections & perspectives

Example: cover

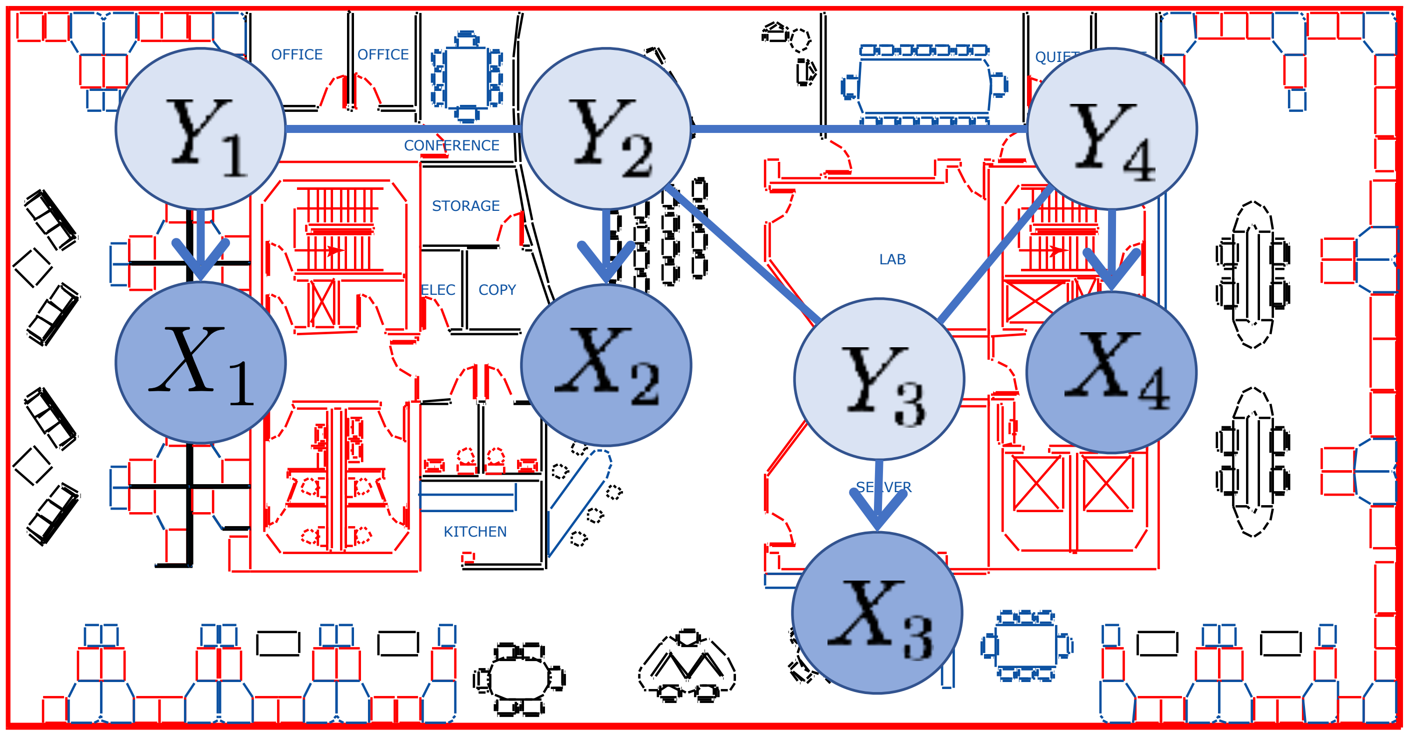
$$F(S) = \left| \bigcup_{v \in S} \mathrm{area}(v) \right|$$





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

Example: sensing



- $\mathcal{V}$ = random variables we can possibly observe
- Utility to have sensors in locations *A*:

$$F(A) = \underbrace{H(\mathbf{Y})}_{\text{uncertainty about temperature before sensing}} - \underbrace{H(\mathbf{Y} \mid \mathbf{X}_A)}_{\text{uncertainty about temperature after sensing}} = I(\mathbf{Y}; \mathbf{X}_A)$$

***uncertainty** about temperature*
***before** sensing*

***uncertainty** about temperature*
***after** sensing*

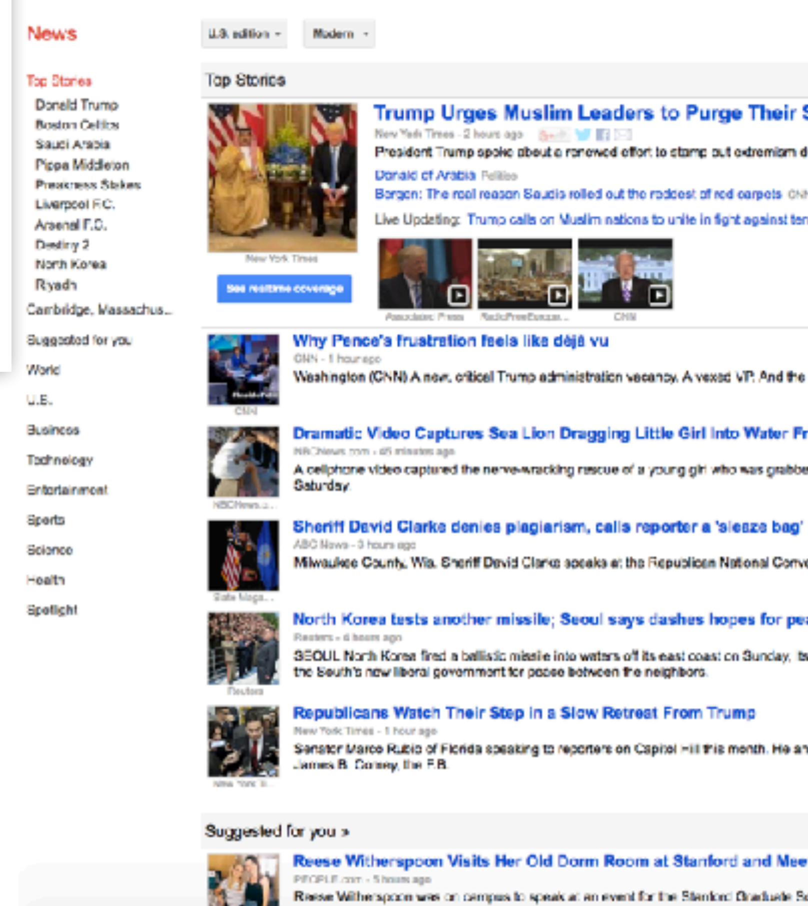
Mutual information

Recommendation & Summarization

We want:
relevance & coverage
diversity
personalization

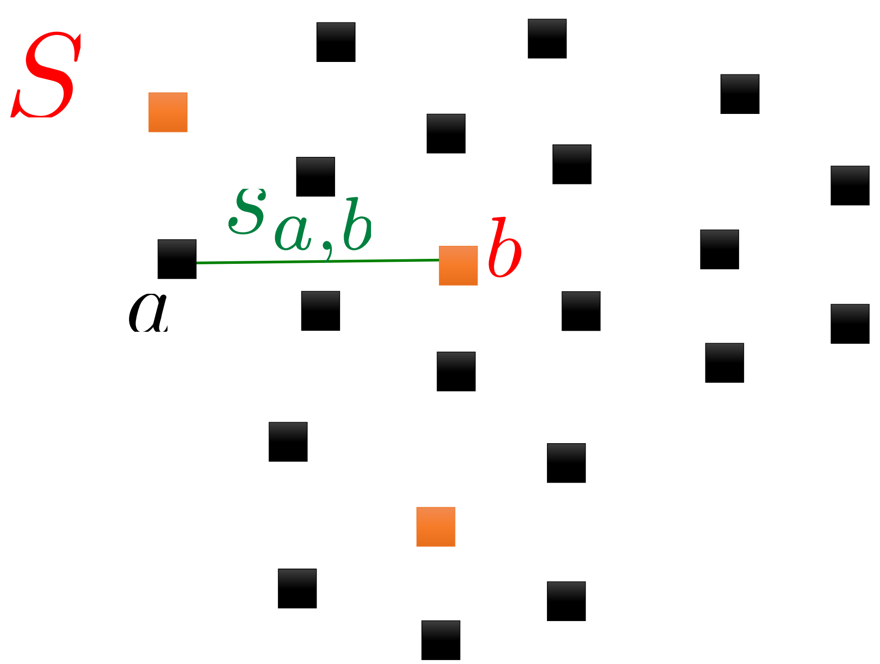
If you bought this,
you may want
to add …

??

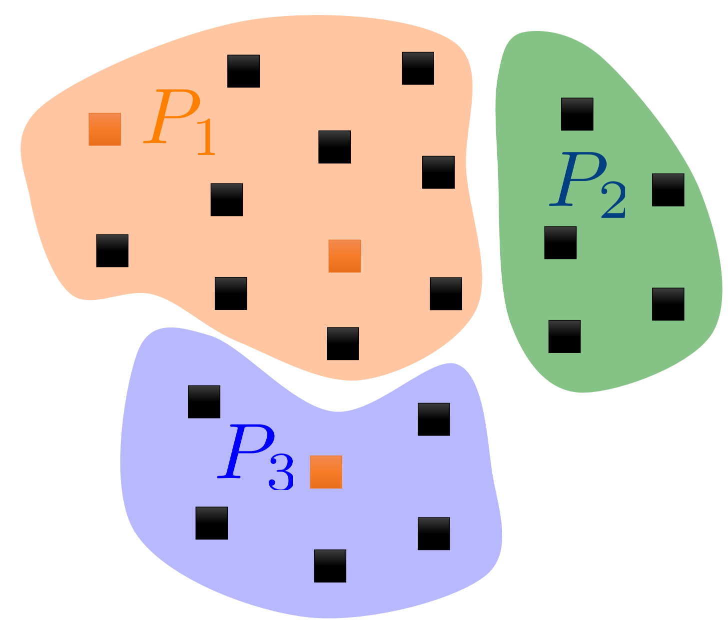


What could F(S) be?

$$F(S) = \sum_{a \in \mathcal{V}} \max_{b \in S} s_{a,b}$$

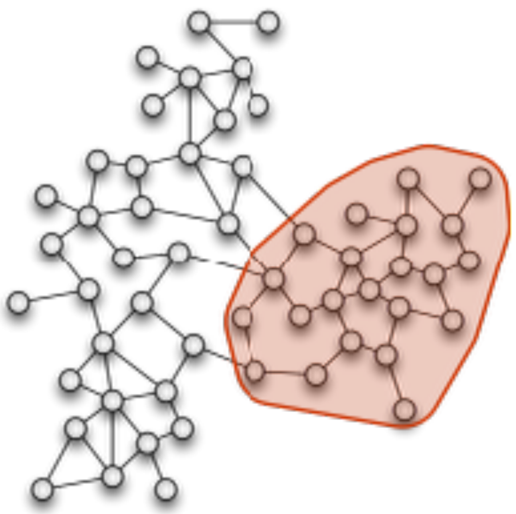


$$F(S) = \sum_{j} \sqrt{|S \cap P_j|}$$



More in the practical!

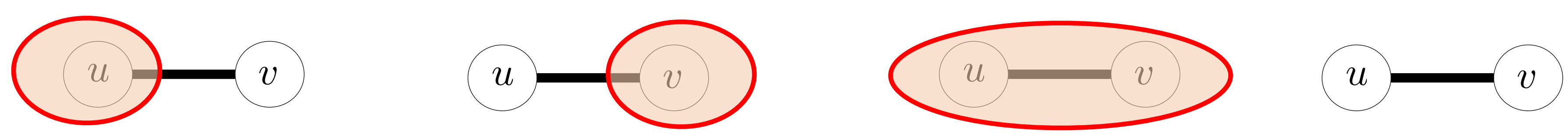
Example: graph cuts



cut for one edge:

$$F(S) = \sum_{u \in S, v \notin S} w_{uv}$$

$$\overset{F(S)}{F(\{u\})} + \overset{F(T)}{F(\{v\})} \geq \overset{F(S \cup T) + F(S \cap T)}{F(\{u, v\}) + F(\emptyset)}$$



w_{uv} w_{uv} 0 0

- cut of one edge is submodular!
- large graph: sum of edges

sum of submodular functions is submodular

Examples of submodular functions

- Entropy of discrete random variables
- Mutual information

- Matrix rank (as a function of columns)

- Coverage
- Spread in social networks

- Graph cuts
- … many others!

Submodular functions (almost) everywhere!

graph theory *(Frank 1993)*

electrical networks *(Narayanan 1997)*

game theory *(Shapley 1970)*

submodular functions

information theory *(Fujishige 1978)*

matroid theory (Whitney, 1935)

stochastic processes *(Macchi 1975, Borodin 2003)*

machine learning

Roadmap

- Examples of submodular functions
- Maximizing submodular functions
- Minimizing submodular functions
- Further connections & perspectives

Roadmap

- Maximizing submodular functions
- Minimizing submodular functions

Key Question:
Submodularity = Discrete Convexity or Discrete Concavity?
(Lovász, Fujishige, Murota, …)

Roadmap

- Maximizing submodular functions: *discrete concavity*
- Minimizing submodular functions: *discrete convexity*

Key Question:
Submodularity = Discrete Convexity or Discrete Concavity?
(Lovász, Fujishige, Murota, …)

Maximizing a submodular function?

$$\boxed{\max_S \; F(S) \;\; \text{s.t.} \;\; |S| \leq k}$$

NP-hard ☹️

Maximizing a submodular function?

$$\max_{S} \; F(S) \;\; \text{s.t.} \;\; |S| \leq k$$

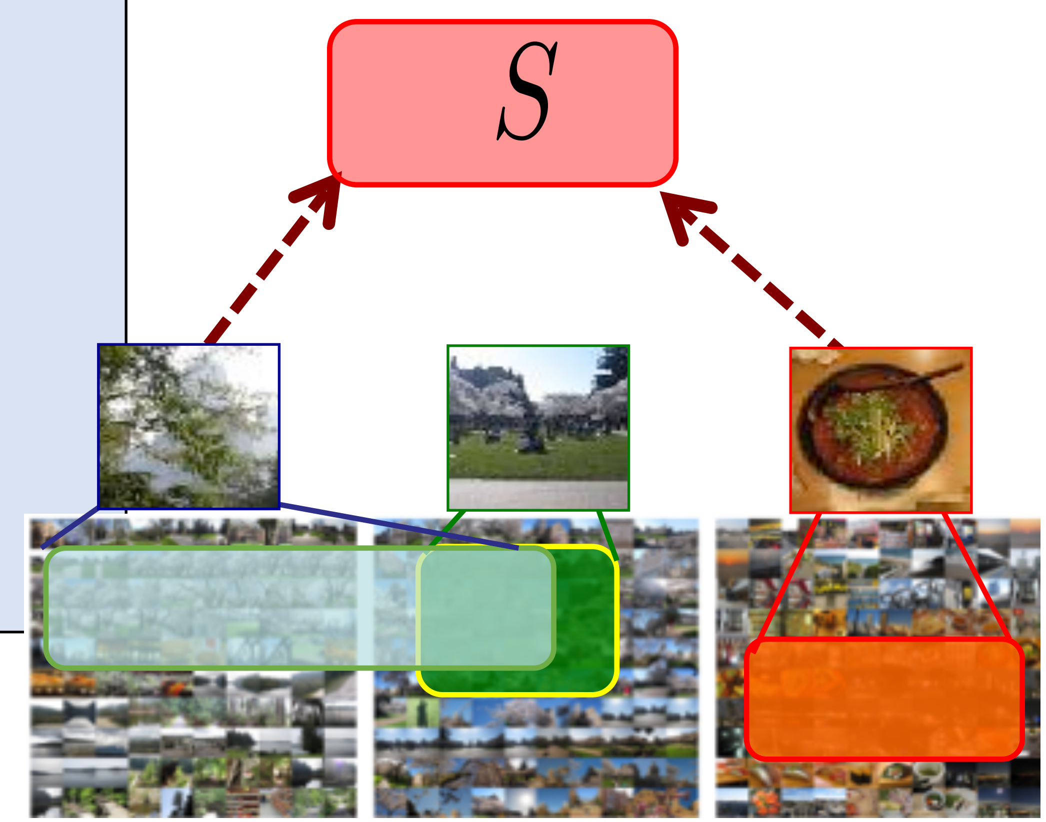
greedy algorithm:

$S_0 = \emptyset$

for *i* = *0*, …, *k-1*

$$e^* = \arg \max_{e \in \mathcal{V} \setminus S_i} F(S_i \cup \{e\})$$

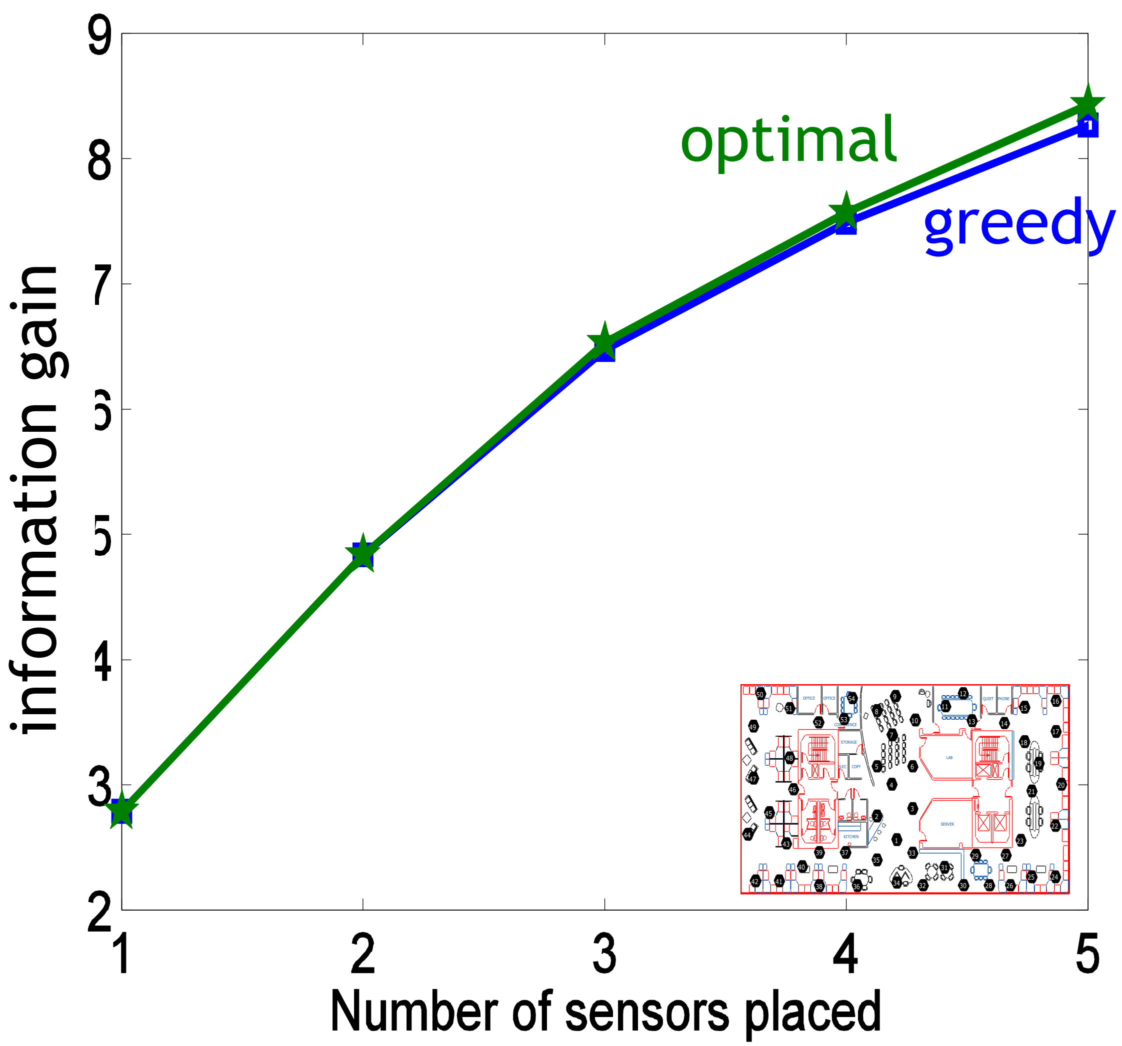
$$S_{i+1} = S_i \cup \{e^*\}$$



How “good” is S_k ?

How good is greedy?

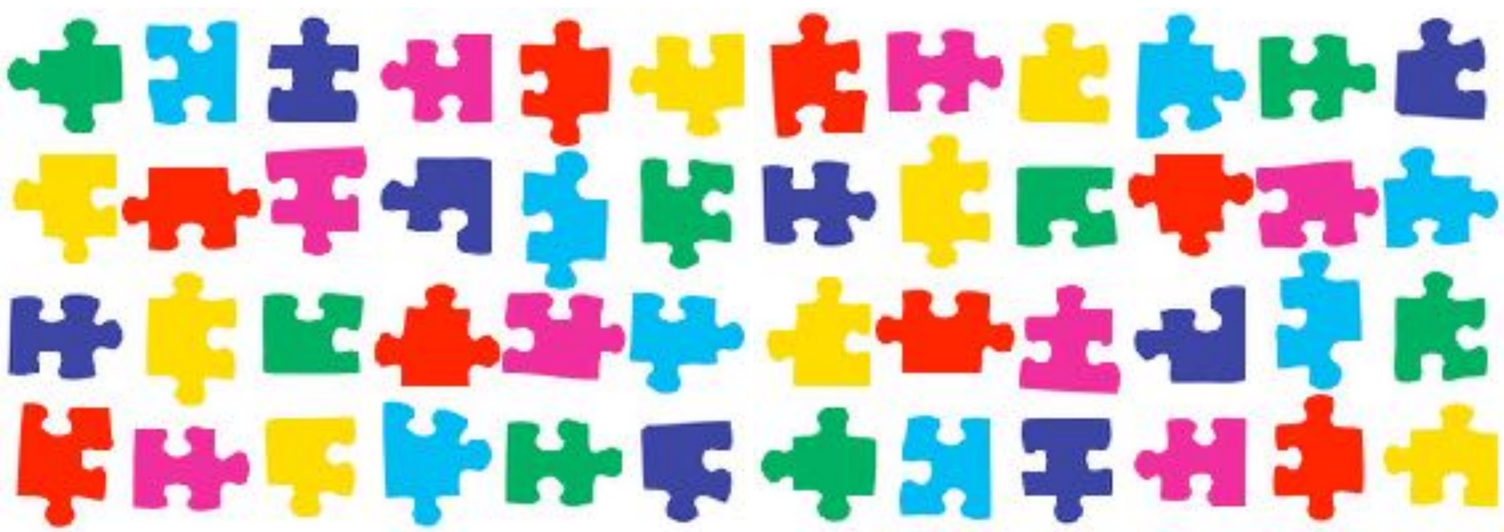
empirically:



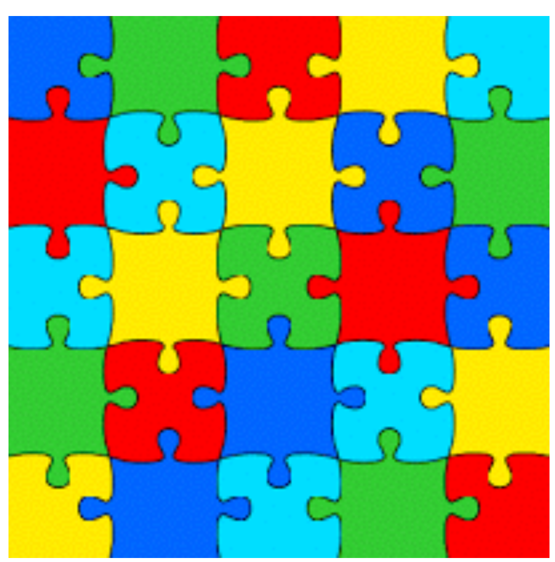
Theorem *(Nemhauser, Wolsey, Fisher 1978)*:
If F is monotone submodular, then Greedy is guaranteed to achieve at least 63% of optimum:

$$F(S_k) \geq \left(1 - \frac{1}{e}\right) F(S^*)$$

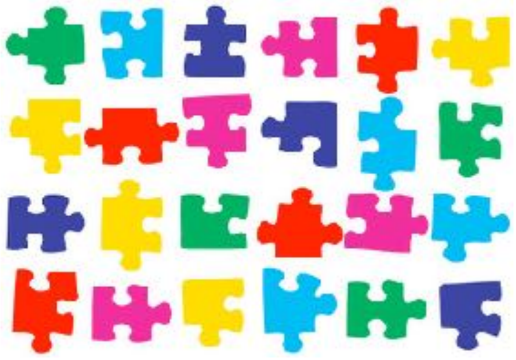
Greedy can fail … without submodularity

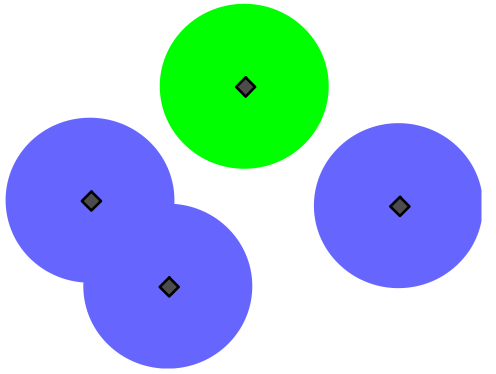


But: this never happens with diminishing returns! ☺️

If $S =$  then $F(S) = 100.$
Otherwise, $F(S) = 0$

Recap: why does greedy work?



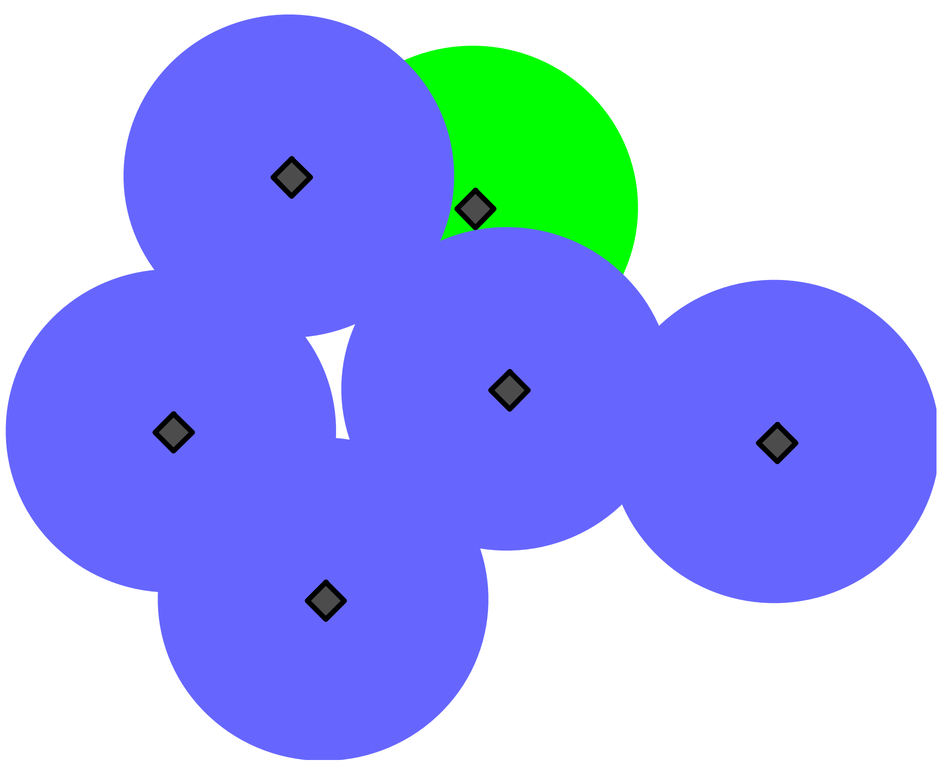


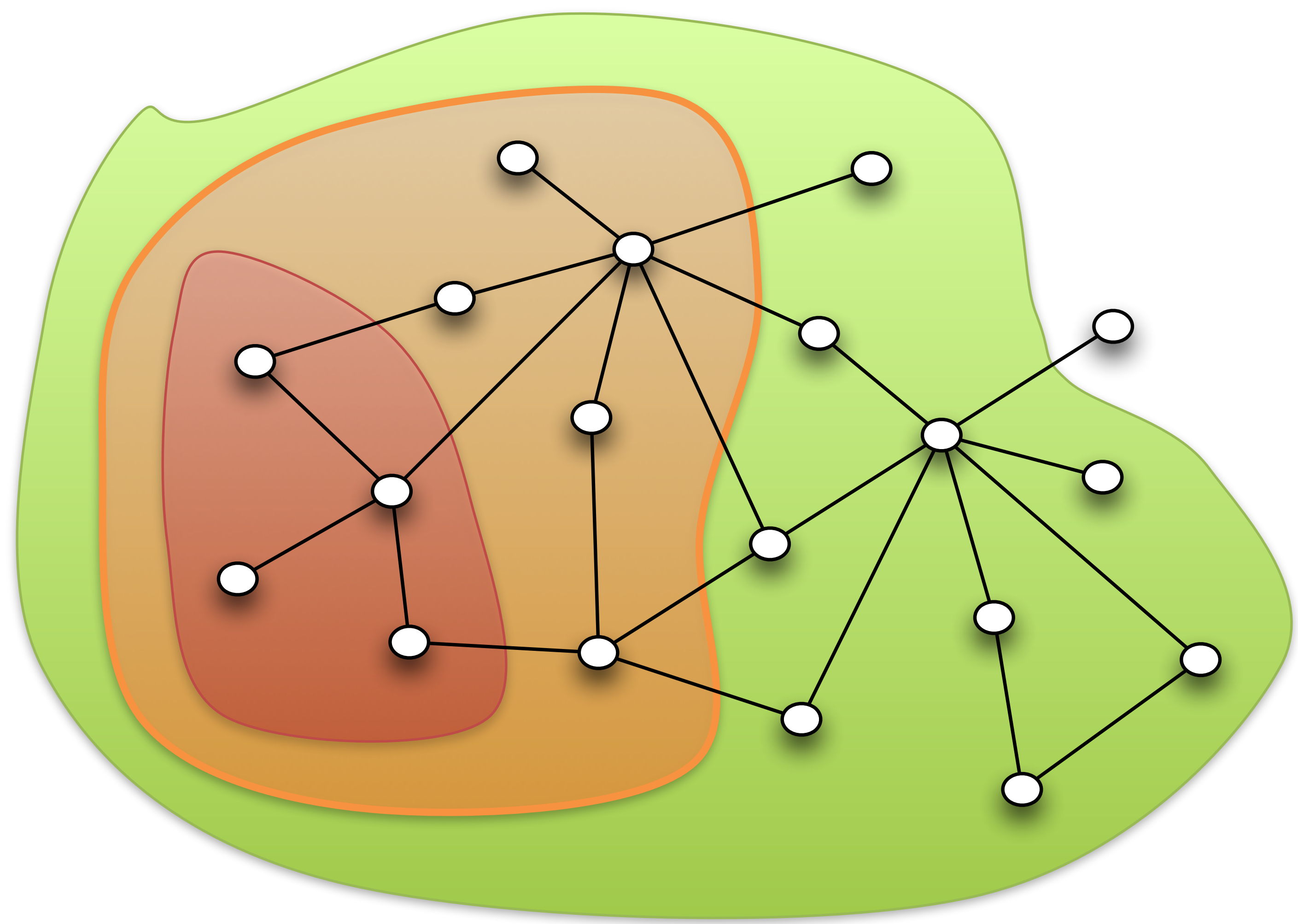
1. **Submodularity**: global information from local information
   Marginal gain of single item gives information about global value

2. **Monotonicity**: items can never harm (= reduce *F*)

Monotonicity

$$\text{if} \quad S \subseteq T \quad \text{then} \quad F(S) \le F(T)$$





3 5 1

Beyond greedy?

- Non-monotone functions?
- Large-scale greedy?
- Other constraints?

Greedy++

Non-monotone functions

$$F(A) = \left| \bigcup_{a \in A} \mathrm{area}(a) \right| - \sum_{a \in A} c(a)$$

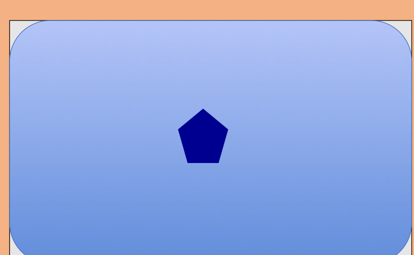
greedy solution:

$$F(A) = 40$$

optimal solution

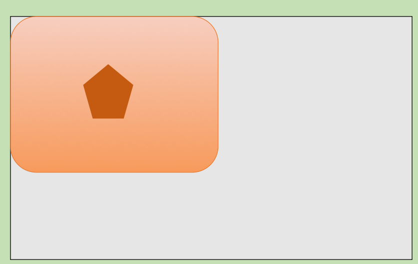
$$F(A) = 95$$

sensor 1



coverage: 100
cost: -60
value 40

sensor 2

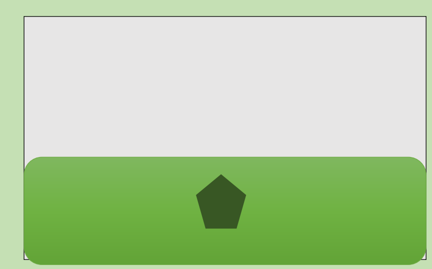


coverage: 30
cost: - 1
value 29

sensor 3

coverage: 30
cost: - 1
value 29

sensor 4



coverage: 40
cost: - 3
value 37

Random Sampling

- Flip a coin for each item. Select with probability 1/2.

- Approximation Guarantee:

$$\mathbb{E}[F(S)] \geq \frac{1}{4}F(S^*)$$

Can we do better?

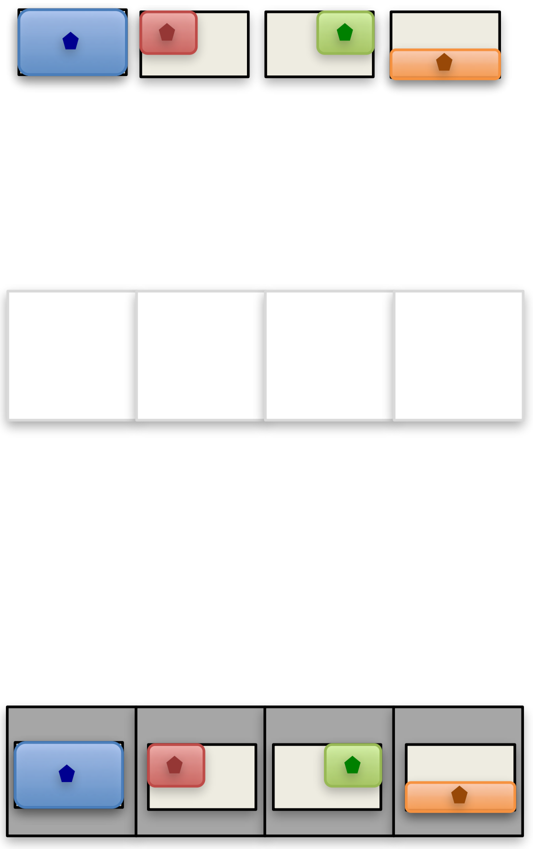
Double (bidirectional) greedy

Start: $A = \emptyset, \; B = \mathcal{V}$

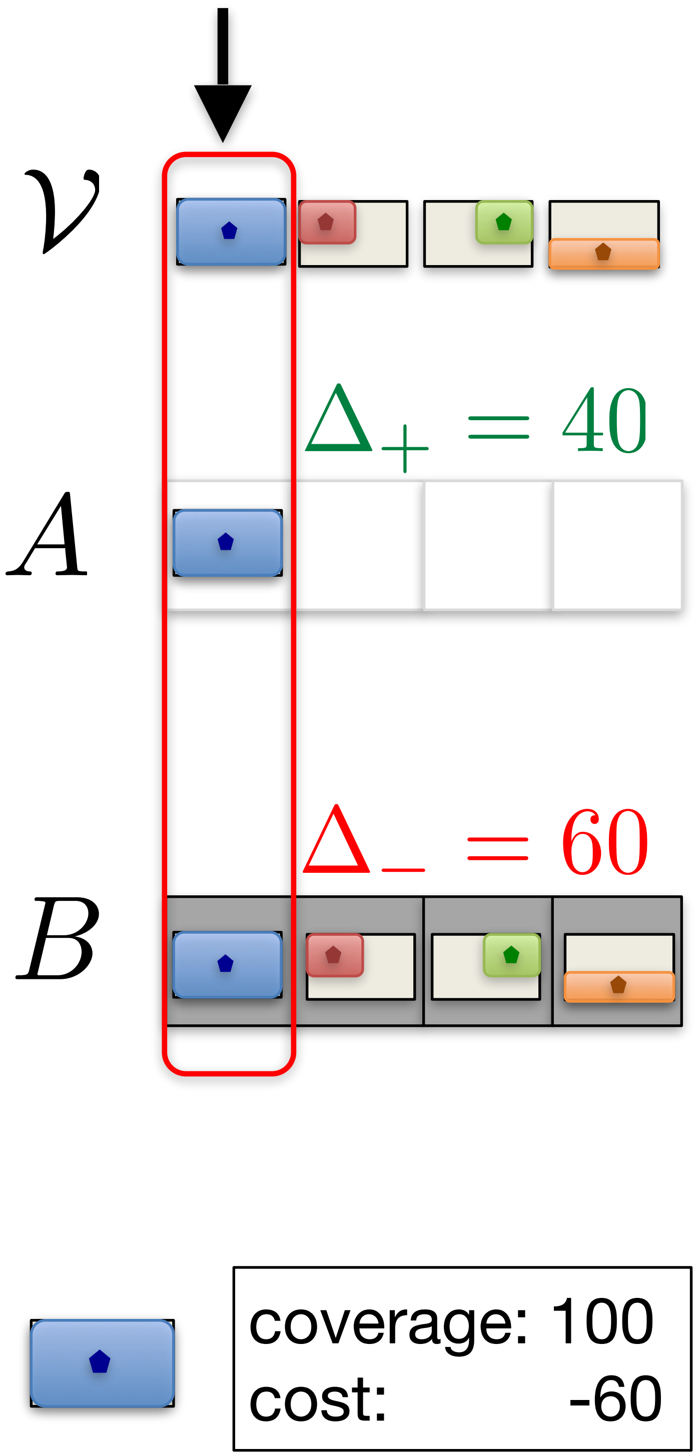
$\mathcal{V}$

A **Definitely keep**

B **Not (yet) decided to remove**



Double (bidirectional) greedy



Start: $A = \emptyset,\ B = \mathcal{V}$

for *i=1, ..., n* *//add or remove?*

- gain of adding (to A):

$$\Delta_+ = [\,F(A \cup a_i) - F(A)\,]_+$$

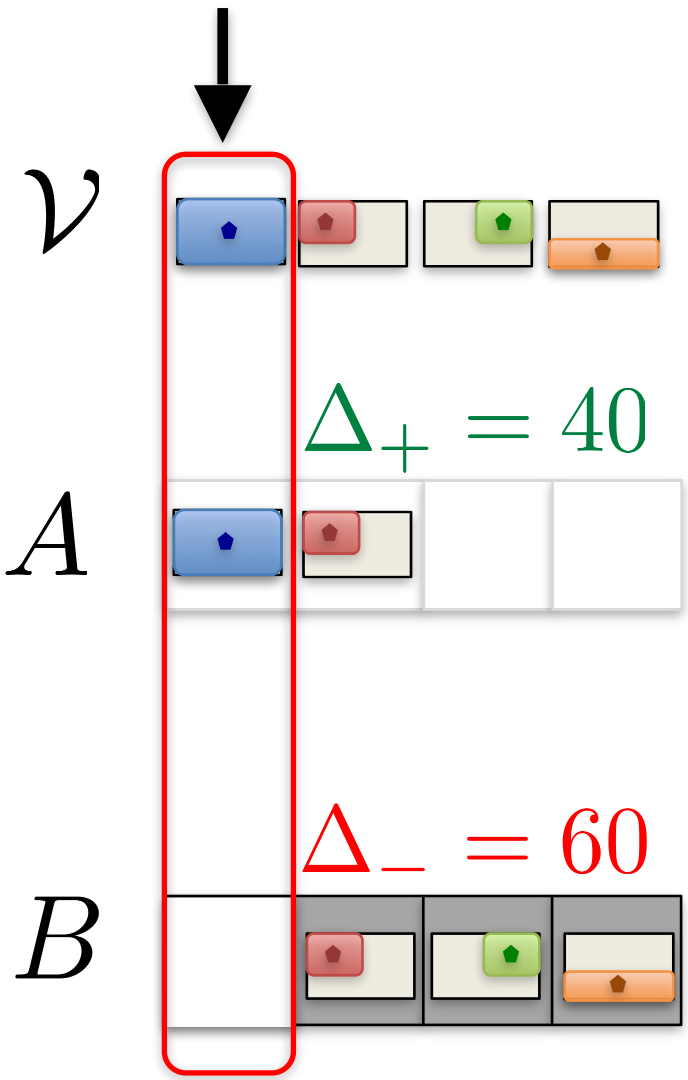
- gain of removing (from B):

$$\Delta_- = [\,F(B \setminus a) - F(B)\,]_+$$

add with probability

$$\mathbb{P}(\text{add}) = \frac{\Delta_+}{\Delta_+ + \Delta_-}$$ = 40%

Double (bidirectional) greedy



Start: $A = \emptyset,\ B = \mathcal{V}$

for *i=1, …, n* *//add or remove?*

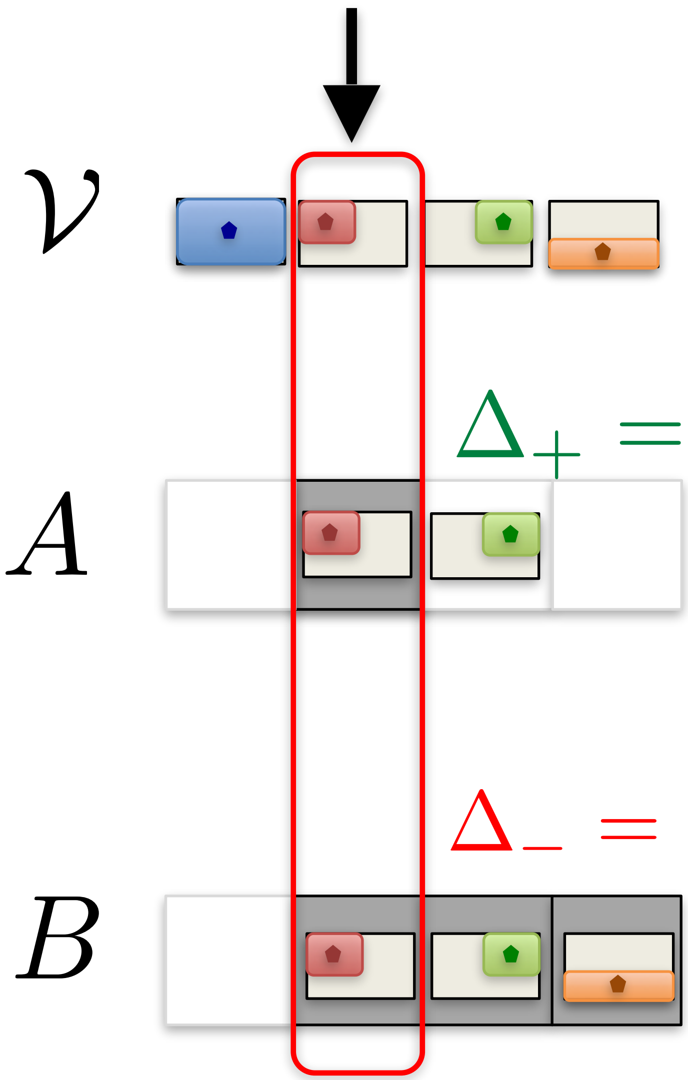
add with probability

$$\mathbb{P}(\text{add}) = \frac{\Delta_+}{\Delta_+ + \Delta_-}$$

add to A or $\boxed{\text{remove from B}}$

coverage: 100
cost: -60

Double (bidirectional) greedy



$\Delta_+ = 29$

$\Delta_- = [-29]_+ = 0$

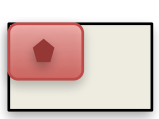
Start: $A = \emptyset, \; B = \mathcal{V}$

for *i=1, …, n* *//add or remove?*

add with probability

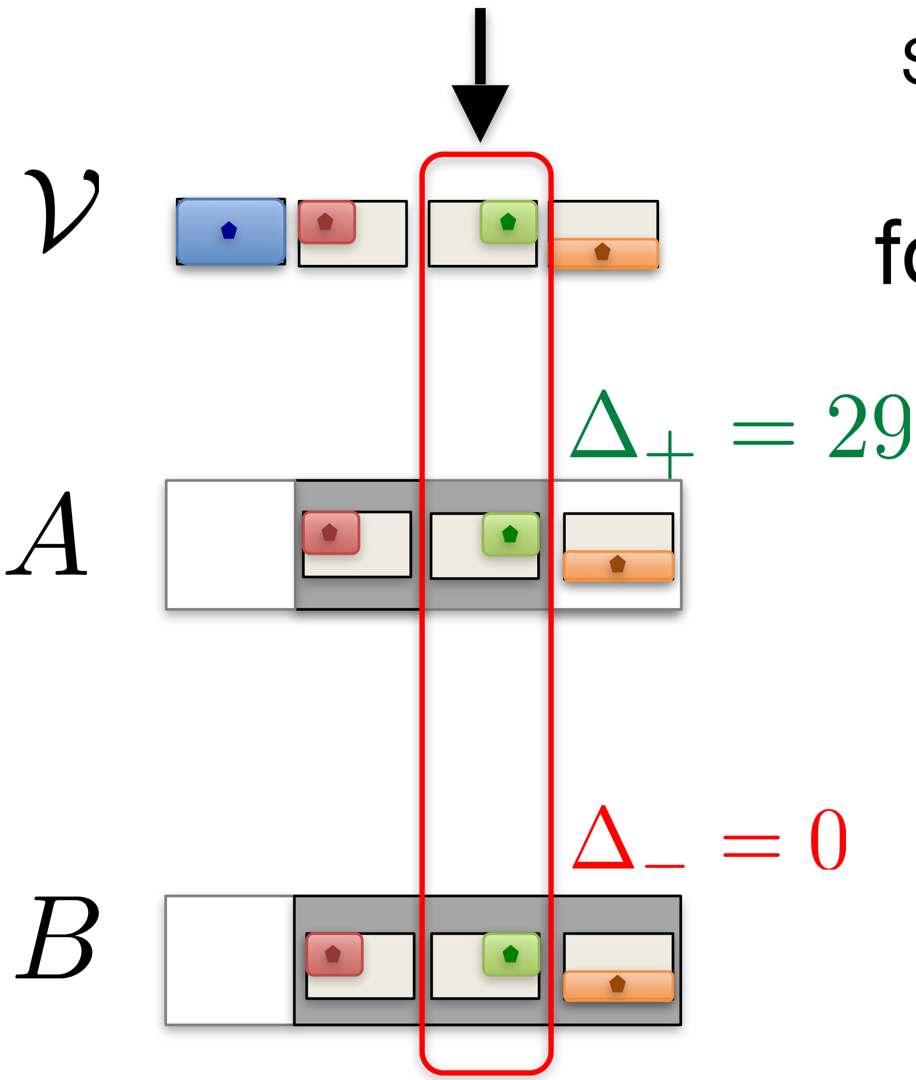
$$\mathbb{P}(\text{add}) = \frac{\Delta_+}{\Delta_+ + \Delta_-} = \frac{29}{29}$$

add to A or remove from B



coverage:	30
cost:	- 1

Double (bidirectional) greedy



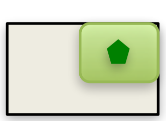
Start: $A = \emptyset,\ B = \mathcal{V}$

for *i*=1, …, *n* *//add or remove?*

add with probability

$$\mathbb{P}(\text{add}) = \frac{\Delta_+}{\Delta_+ + \Delta_-} = \frac{29}{29}$$

add to A or remove from B



coverage: 30
cost: - 1

Double greedy

$$\max_{S \subseteq \mathcal{V}} \; F(S)$$

Theorem *(Buchbinder, Feldman, Naor, Schwartz '12)*

F submodular, S_g solution of double greedy. Then

$$\mathbb{E}[F(S_g)] \geq \frac{1}{2} F(S^*)$$

optimal solution

Beyond greedy?

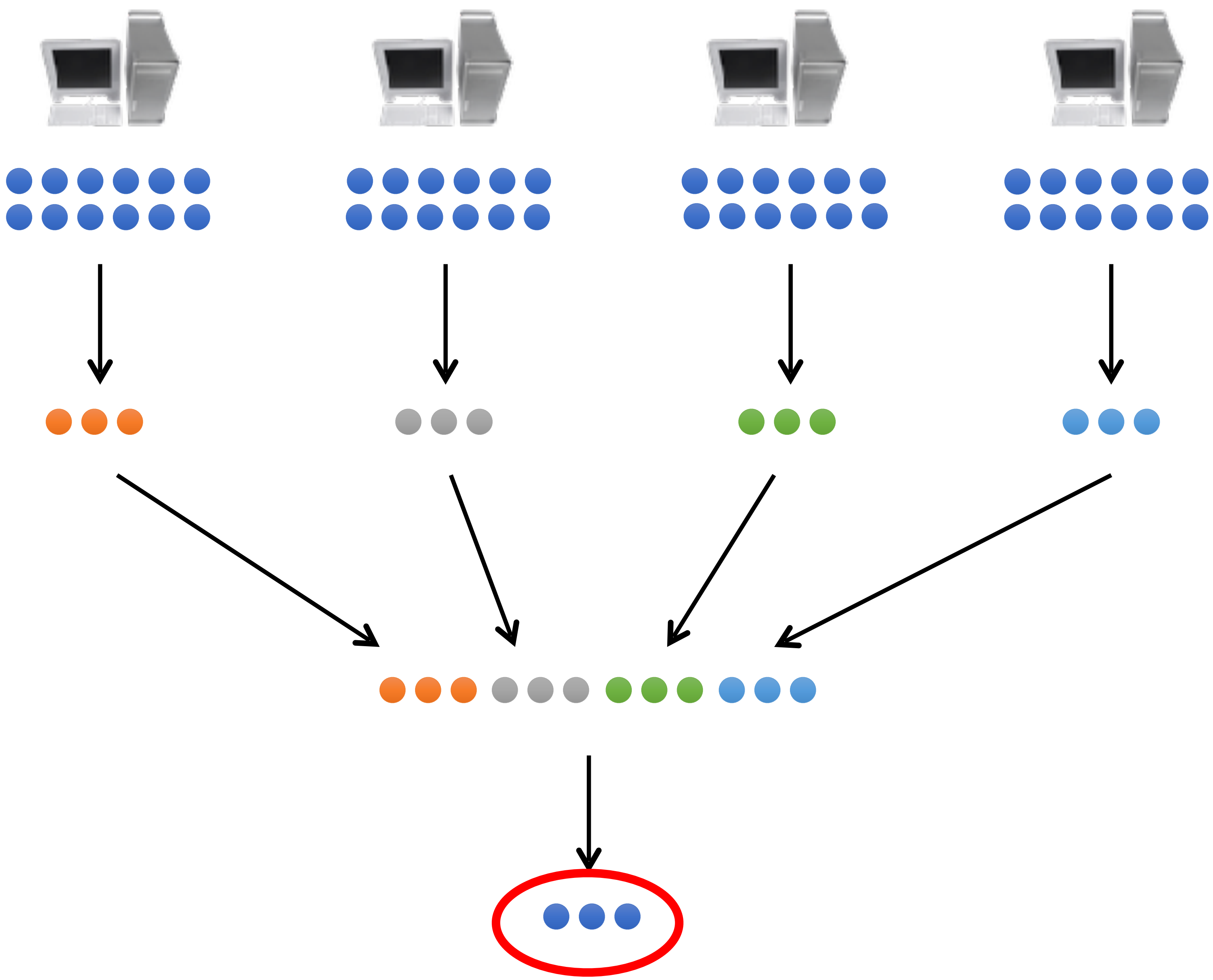
- Non-monotone functions?
  Double greedy (or repeated greedy)

- Large-scale greedy?

- Other constraints?

Greedy++

Distributed greedy algorithms



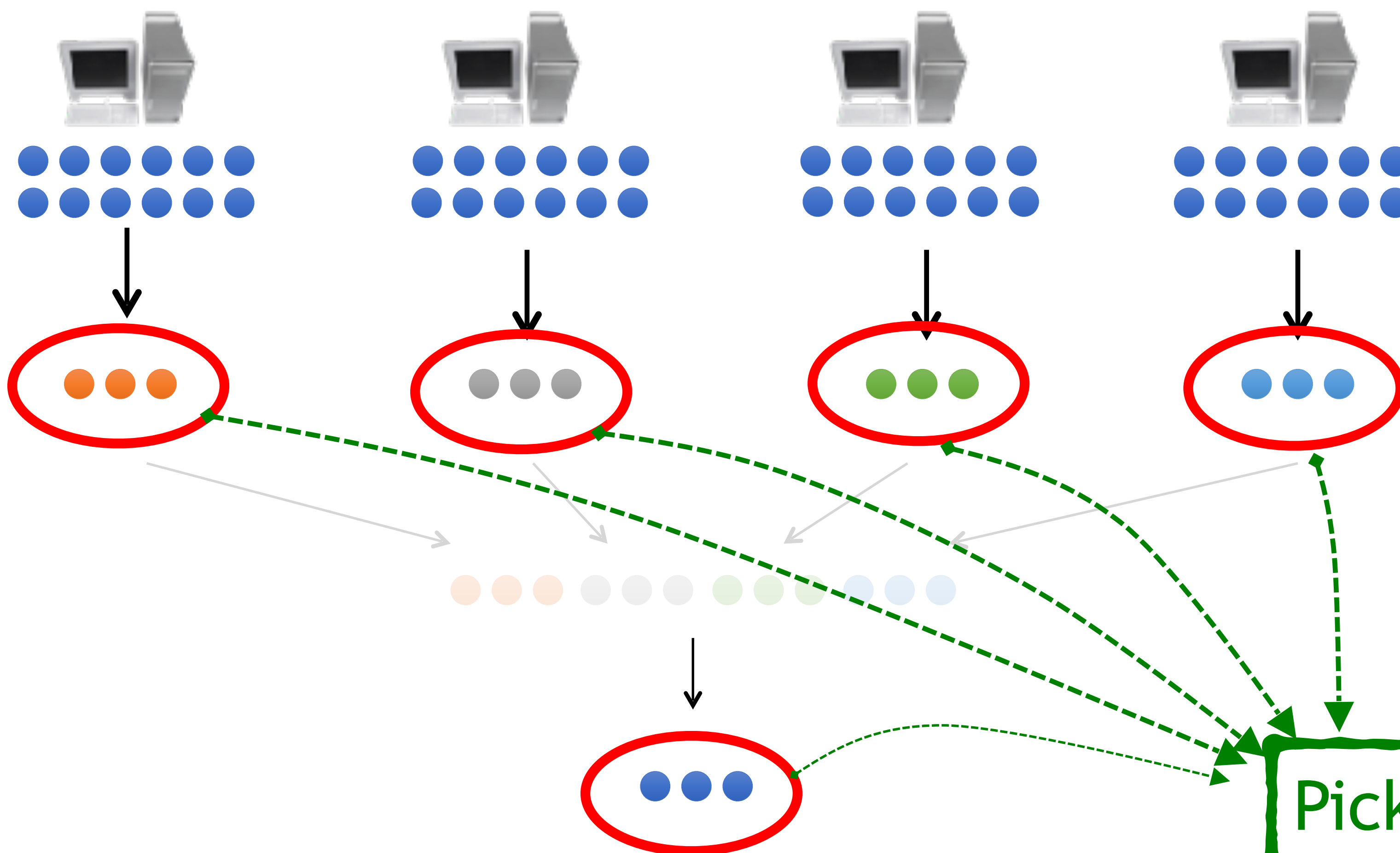
greedy is sequential.
pick in parallel??

pick *k* elements
on each machine.

combine and run
greedy again.

Is this useful?

Distributed greedy algorithms



pick in parallel
from *m* machines

Is this useful?

Pick the best of m+1 solutions

Random partition:

$$\frac{1}{2}\left(1 - \frac{1}{e}\right)$$

Even better with geometric structure

(Mirzasoleiman-Karbasi-Sarkar-Krause'13, da Ponte Barbosa-Ene-Nguyen-Ward'15)

More scalability

- Streaming algorithms
- Stochastic greedy

- **Adaptivity** *(Balkanski-Singer 2017, Balkansi-Rubinstein-Singer 2018, …)*
    - # rounds of parallel queries
    - needs adaptive sampling: e.g.
      up-sample growing set or down-sample increasing set, distribution evolves over rounds

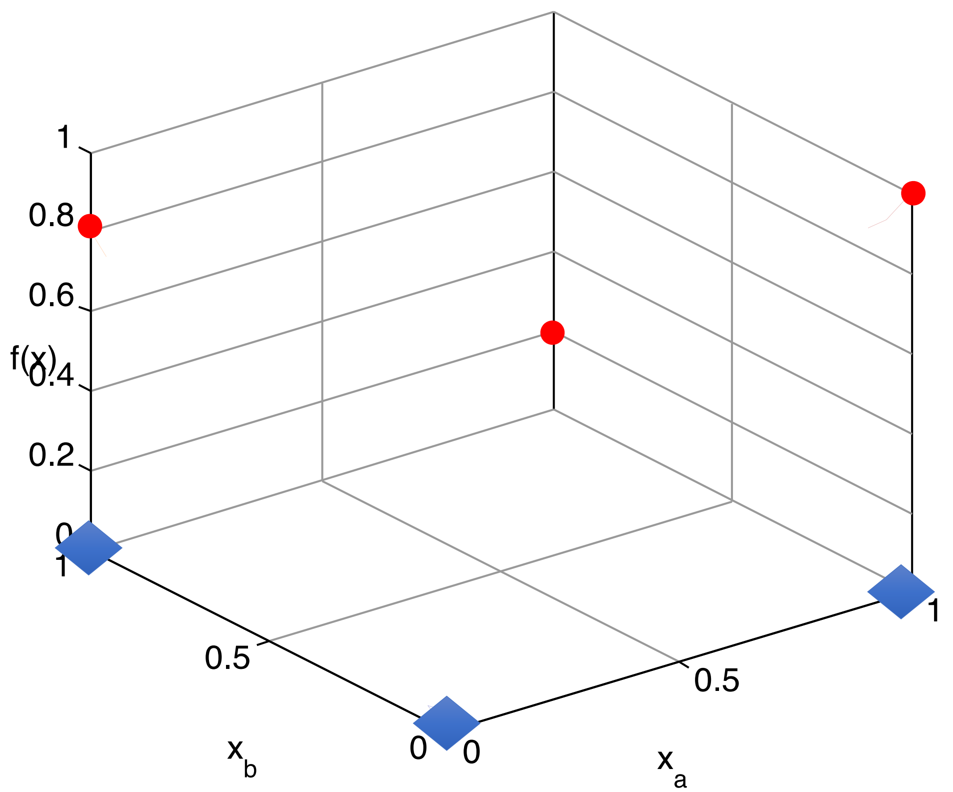
Beyond greedy?

- Non-monotone functions?
  Double greedy (or repeated greedy)
- Large-scale greedy?
  Distributed, streaming, stochastic,…
- Other constraints?

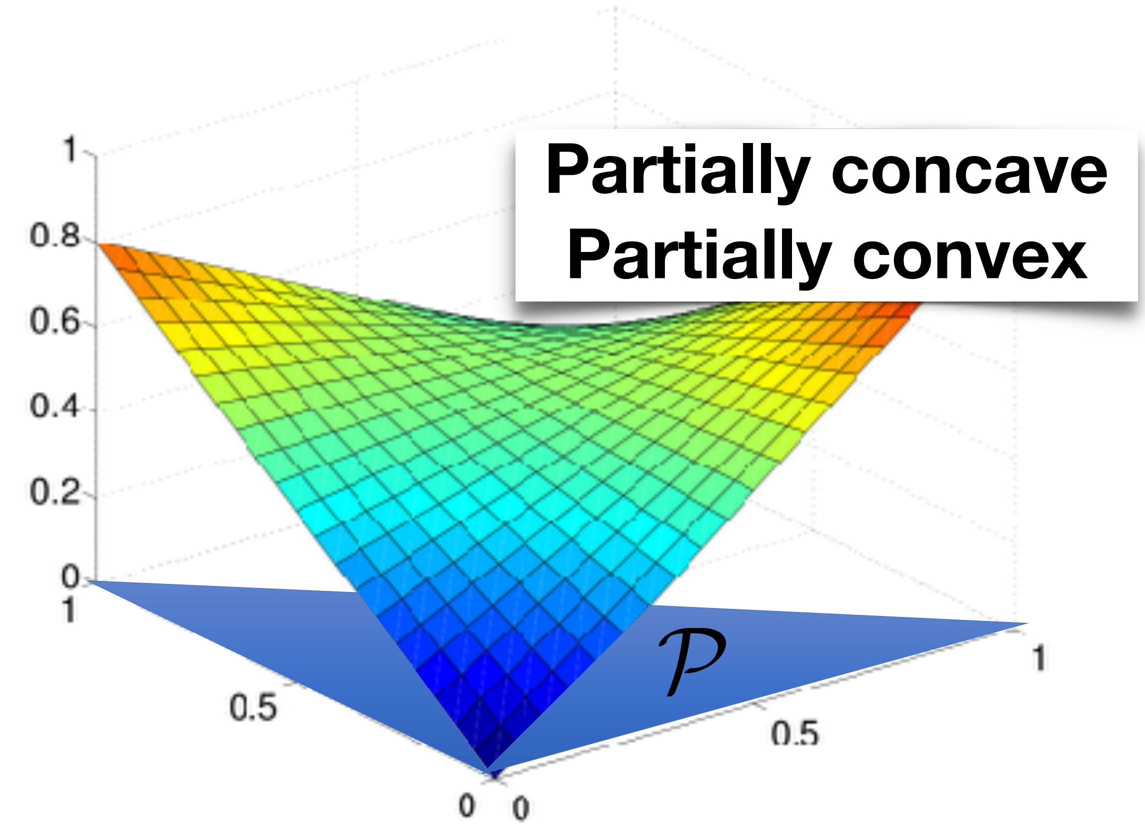
Continuous Optimization!

Relax: Discrete to continuous

$\max F(S)$



$\max \quad f_M(x)$



Algorithm: “continuous greedy”

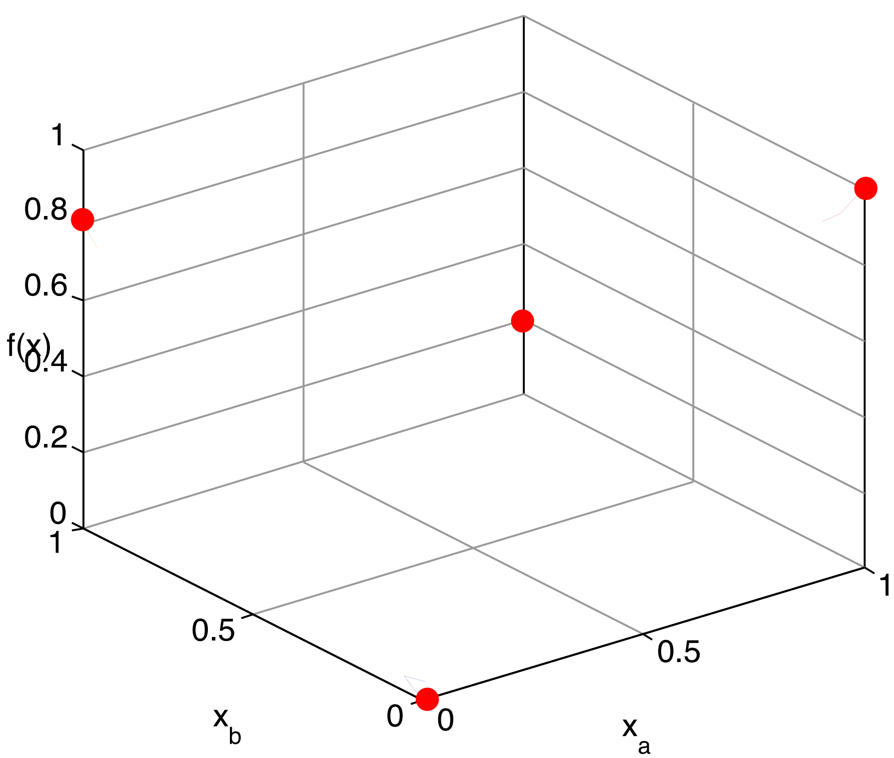
1. approximately maximize f_M over $\mathcal{P} = \mathrm{conv}(\mathcal{I})$
2. round to discrete set

(Vondrák ’08; Calinescu-Chekuri-Pal-Vondrák ‘11; Kulik-Shachnai-Tamir’11)

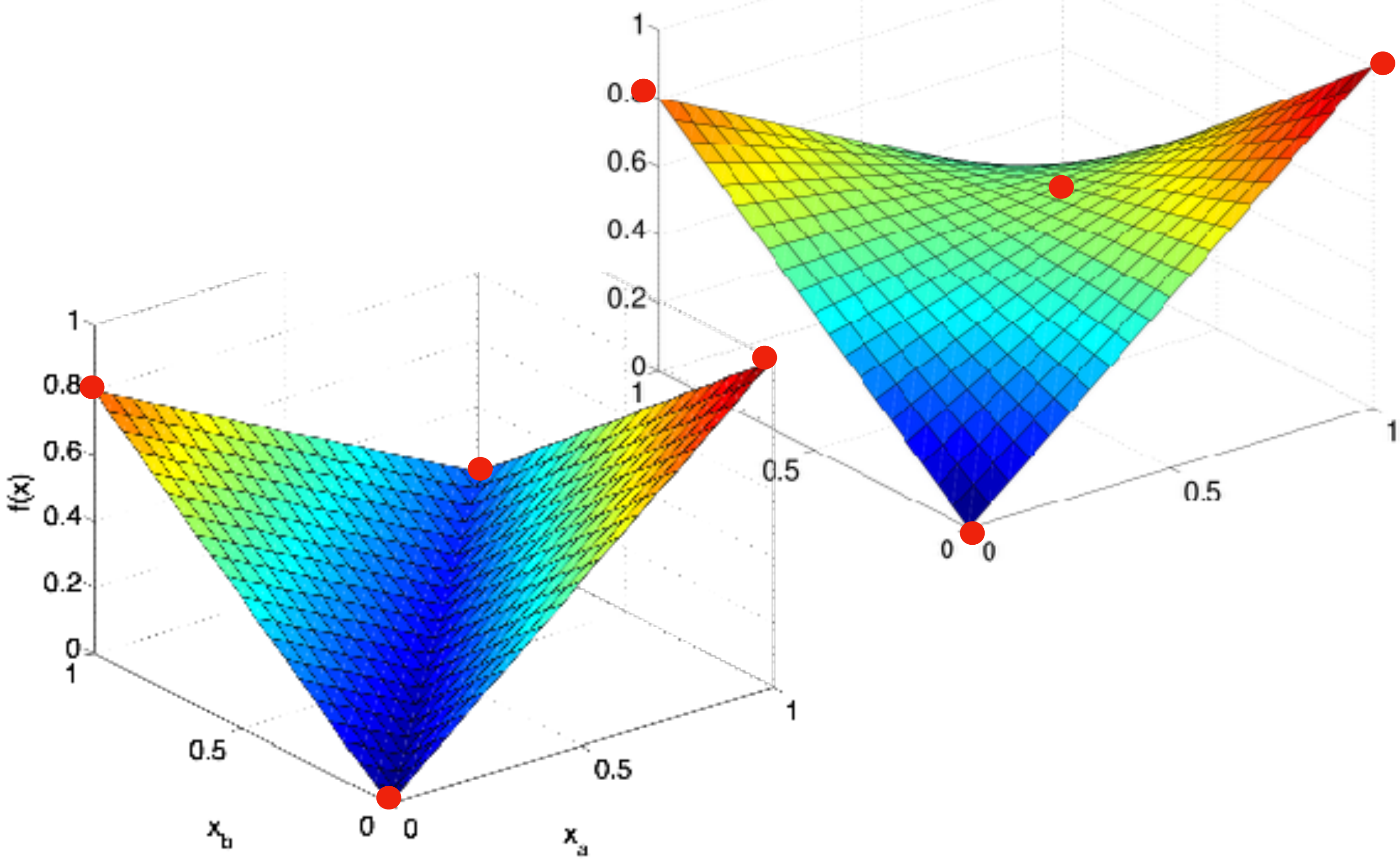
Between discrete & continuous optimization

nonlinear extension/optimization

$$F : \{0,1\}^n \to \mathbb{R}$$



$$f : [0,1]^n \to \mathbb{R}$$



$$\min_{x \in \mathcal{C} \subseteq \{0,1\}^n} F(x)$$

$$\min_{z \in \mathrm{conv}(\mathcal{C}) \subseteq [0,1]^n} f(z)$$

Generic construction

$F : \{0,1\}^n \to \mathbb{R} \longrightarrow f : [0,1]^n \to \mathbb{R}$

discrete set:
T = {*a*,*d*}

1	*a*	*a*	.5
0	*b*	*b*	.5
0	*c*	*c*	0
1	*d*	*d*	.8

continuous *z*

- Define probability measure over *subsets* (joint over coordinates) such that marginals agree with z:

$$\mathbb{P}(i \in S) = z_i$$

- Extension: $\boxed{f(z) = \mathbb{E}[F(S)]}$

- for discrete z: $f(z) = F(z)$

Independent coordinates

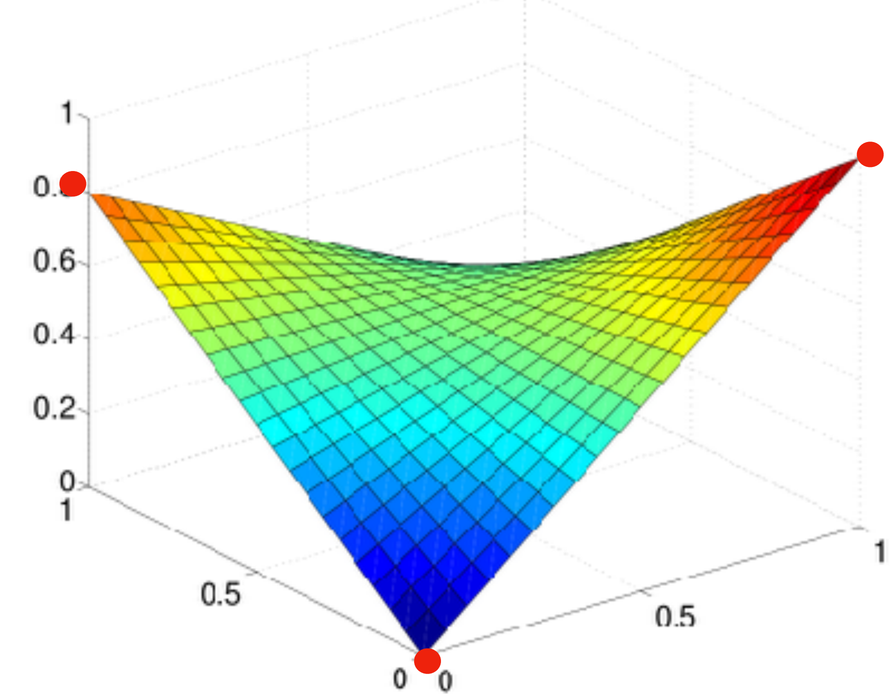
$$\boxed{f(z) = \mathbb{E}[F(S)]}$$

a	.5
b	.5
c	0
d	.8

$$P(S) = \prod_{i \in S} z_i \cdot \prod_{j \notin S} (1 - z_j)$$

- $f(z)$ is a multilinear polynomial: ***multilinear extension***

- neither convex nor concave…
  but enough for optimization



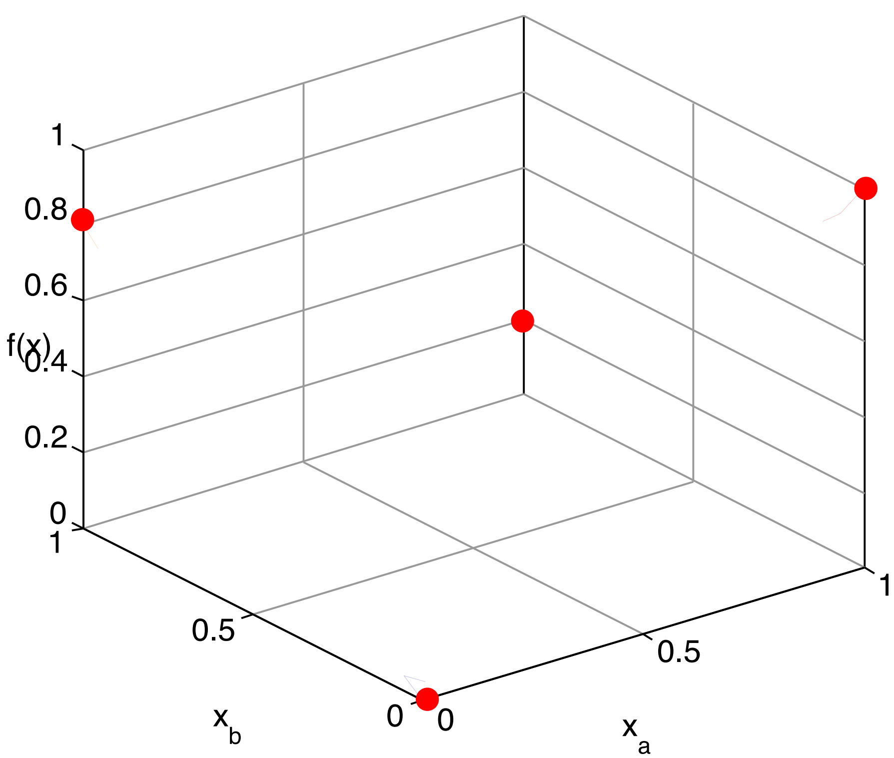
Roadmap

- Examples of submodular functions
- Maximizing submodular functions
  - greedy
  - double greedy
  - continuous greedy
- Minimizing submodular functions
- Further connections & perspectives

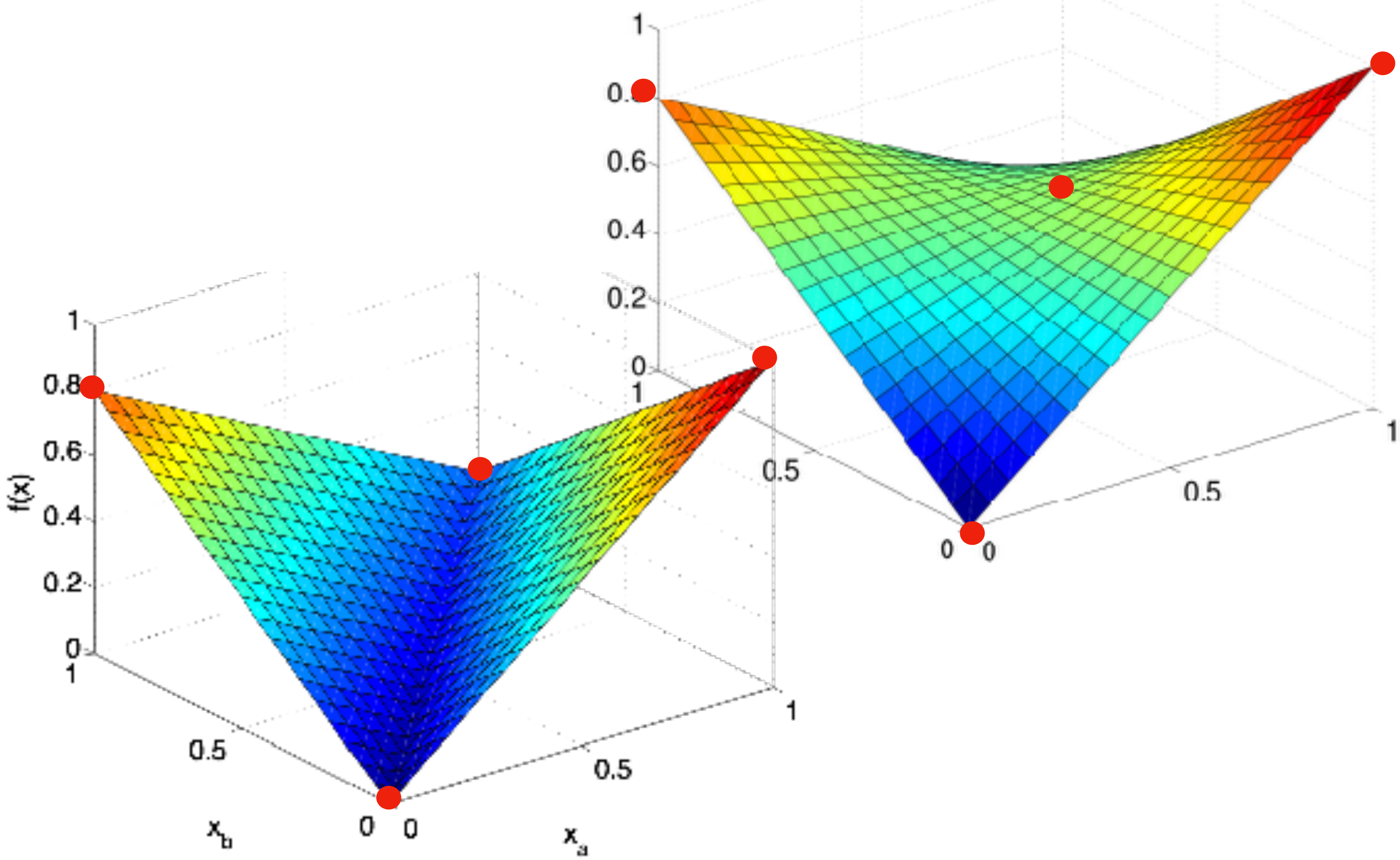
Between discrete & continuous optimization

nonlinear extension/optimization

$$F : \{0,1\}^n \to \mathbb{R}$$



$$f : [0,1]^n \to \mathbb{R}$$



$$\min_{x \in \mathcal{C} \subseteq \{0,1\}^n} F(x)$$

$$\min_{z \in \mathrm{conv}(\mathcal{C}) \subseteq [0,1]^n} f(z)$$

Lovász extension

$$\boxed{f(z) = \mathbb{E}[F(S)]}$$

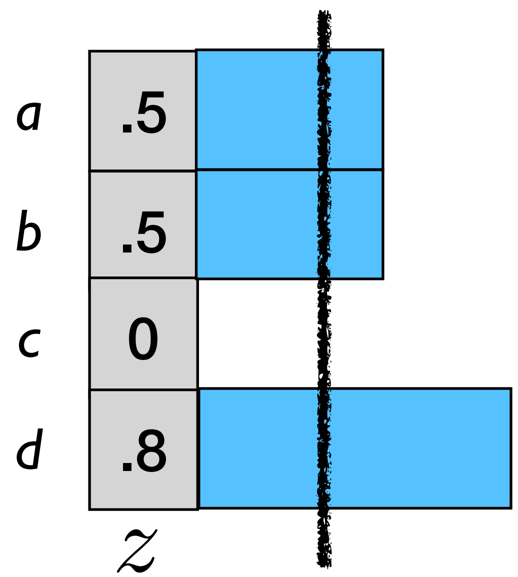
$$\mathbb{P}(i \in S) = z_i$$

Lovász extension

$$\boxed{f(z) = \mathbb{E}[F(S)]}$$

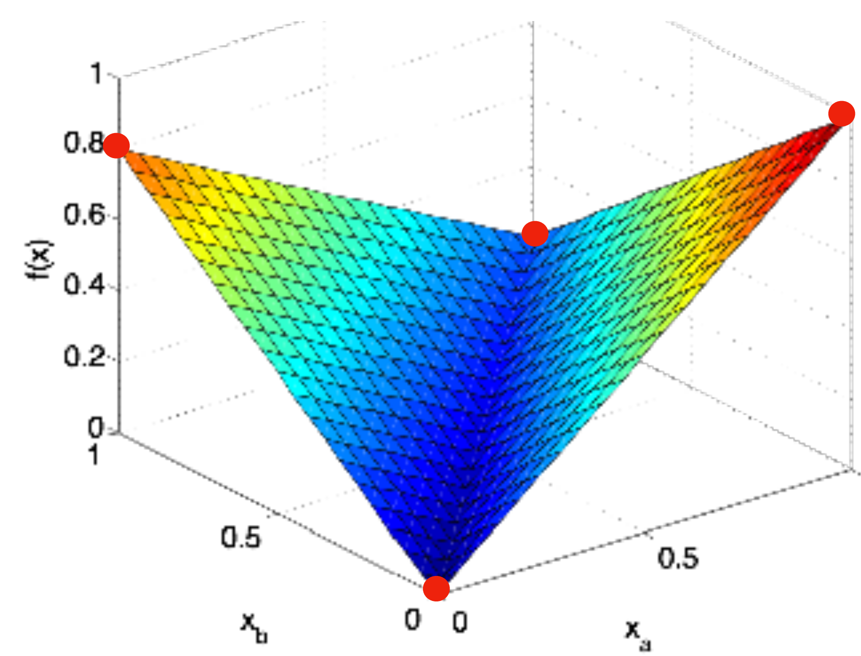
$$\mathbb{P}(i \in S) = z_i$$

- "coupled" distribution defined by level sets



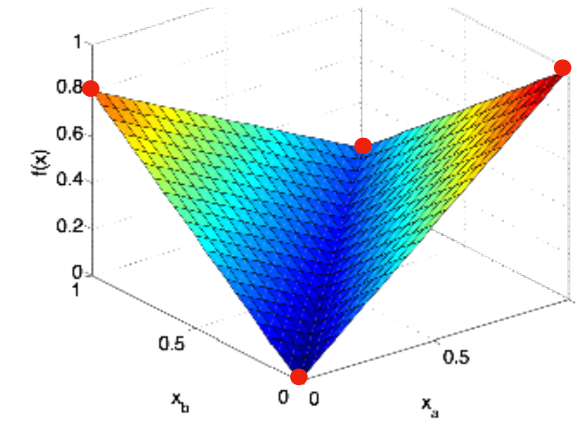
$$S_0 = \{\},\ S_1 = \{d\},\ S_2 = \{a, b, d\},$$
$$S_3 = \{a, b, c, d\}$$

Theorem (Lovász 1983)
$f(z)$ is convex iff $F(S)$ is submodular.



Implications of the Lovász extension

$$f(z) = \mathbb{E}[F(S)]$$



- piecewise linear & convex
- can compute subgradient of *f(z) in O(n log n)*

➡ *convex optimization*

- rounding: use one of the level sets of z^*

	z
a	.5
b	.5
c	0
d	.8

exact convex relaxation! $$\min_{S \subseteq V} F(S) = \min_{z \in [0,1]^n} f(z)$$

Submodular minimization: a brief overview

$$\boxed{\min_{z \in [0,1]^n} f(z)}$$

convex optimization

- **ellipsoid method** *(Grötschel-Lovász-Schrijver 81)*
- **subgradient method** *(improved: Chakrabarty-Lee-Sidford-Wong 16)*

combinatorial optimization

- **network flow based** *(Schrijver 00, Iwata-Fleischer-Fujishige-01)*
  $O(n^4 T + n^5 \log M)$ *(Iwata 03)*, $O(n^6 + n^5 T)$ *(Orlin 09)*

convex + combinatorial

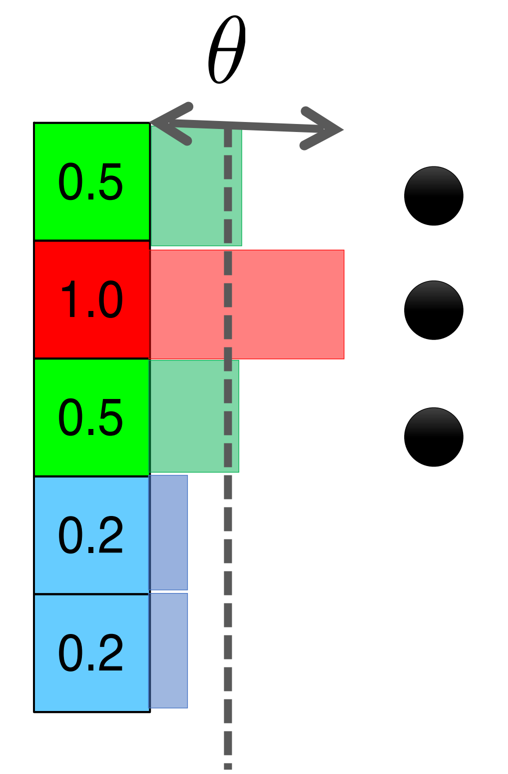
- **cutting planes** *(Lee-Sidford-Wong 15)*

$$O(n^2 T \log nM + n^3 \log^c nM) \qquad O(n^3 T \log^2 n + n^4 \log^c n)$$

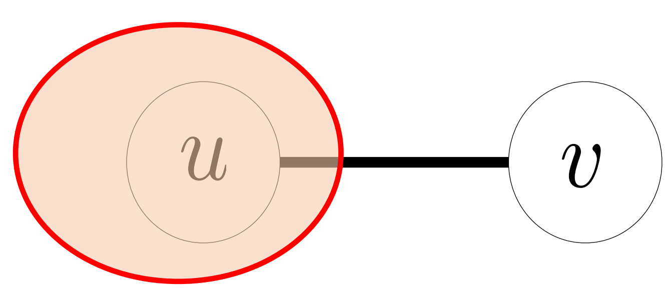
Examples of Lovasz extensions

1. $$F(S) = \min\{|S|, 1\}$$

$$f(x) = \max_i x_i$$



2. Cut function: 2 items (nodes)

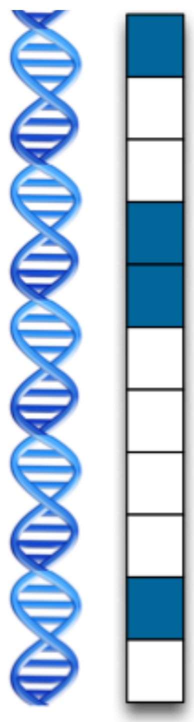


$$F(S) = \begin{cases} 1 & \text{if } |S| = 1 \\ 0 & \text{otherwise.} \end{cases}$$

$$f(x) = |x_u - x_v|$$

Applications

- structured sparsity *(Bach 10)*



Only use few coordinates of x in $f(x, w)$

Convex regularizer from submodular function!

- decomposition & parallel algorithms
  (Komodakis-Paragios-Tziritas 11, Stobbe-Krause 10, Jegelka-Bach-Sra 13, Nishihara-Jegelka-Jordan 14, Ene-Nguyen 15)

- variational inference *(Djolonga-Krause 14)*

How far does relaxation go?

- **strongly convex** version:

$$\min_{z \in [0,1]^n} f(z) \quad \dashrightarrow \quad \min_{z \in \mathbb{R}^n} f(z) + \frac{1}{2}\|z\|^2$$

$$\updownarrow$$

dual: $$\min_{s \in \mathcal{B}_F} \frac{1}{2}\|s\|^2$$

- Fujishige-Wolfe / minimum-norm point algorithm
- actually solves *parametric submodular minimization*

- **But:** no relaxation is tight for constrained minimization
typically hard to approximate

Roadmap

- Examples of submodular functions
- Maximizing submodular functions: *discrete concavity*
- Minimizing submodular functions: *discrete convexity*
- Further connections & perspectives
  - nonconvex optimization
  - discrete probability

Submodularity beyond sets

- sets: for all subsets $A, B \subseteq V$

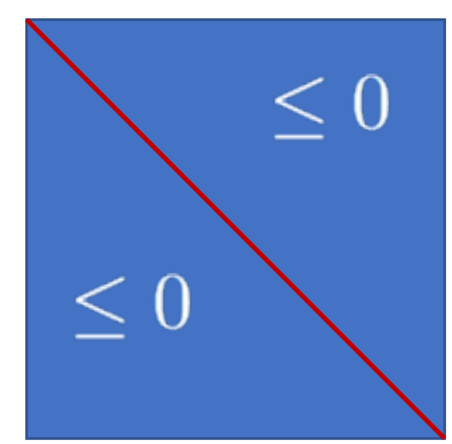
$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$$

- replace sets by vectors:

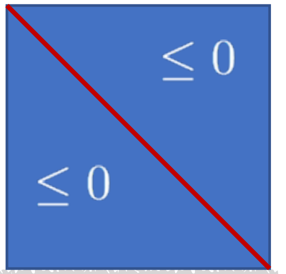
$$F(x) + F(y) \geq F(x \vee y) + F(x \wedge y)$$

- or: Hessian has all off-diagonals <= 0. *(Topkis 1978)*

$$\frac{\partial^2 F}{\partial x_i \partial x_j} \leq 0 \qquad \forall i \neq j$$

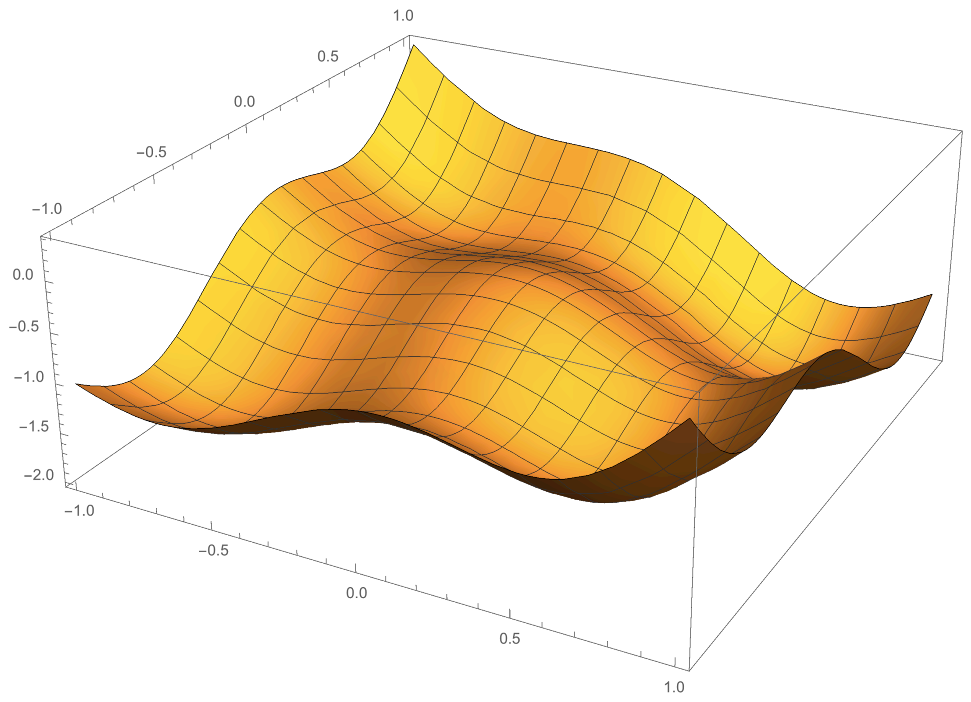


Examples



$$F(x) + F(y) \geq F(x \vee y) + F(x \wedge y)$$

$$\frac{\partial^2 F}{\partial x_i \partial x_j} \leq 0 \quad \forall i \neq j$$



submodular function can be convex, concave or neither!

- any separable function $F(x) = \sum_{i=1}^{n} F_i(x_i)$
- $F(x) = g(x_i - x_j)$ for concave g
- $F(x) = h\left(\sum_i x_i\right)$ for convex h

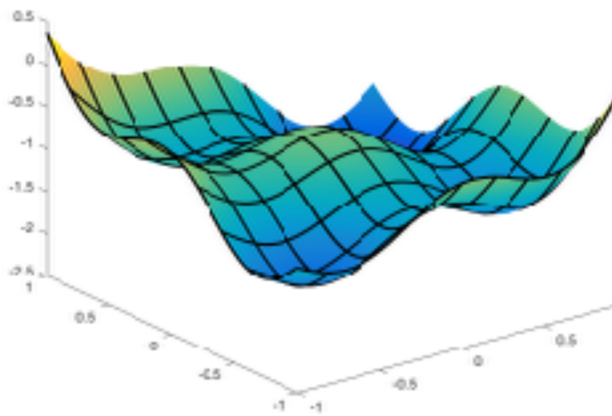
Optimization

Maximization

- "diminishing returns" stronger than submodularity
- with DR, many results generalize (incl. "continuous greedy")
  (Kapralov-Post-Vondrák 2010, Soma et al 2014-15, Ene & Nguyen 2016, Bian et al 2016, Gottschalk & Peis 2016)

Minimization

- discretize continuous functions
- then same approaches as for set functions
  (Schrijver 2000, Orlin 2007, Bach 2015)



➡ *nonconvex optimization via discrete optimization!*

Discrete Probability

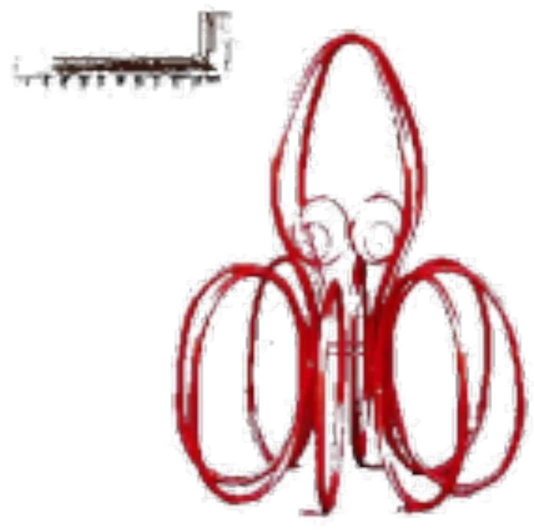
- Distribution over subsets
- Log-submodularity / log-supermodularity

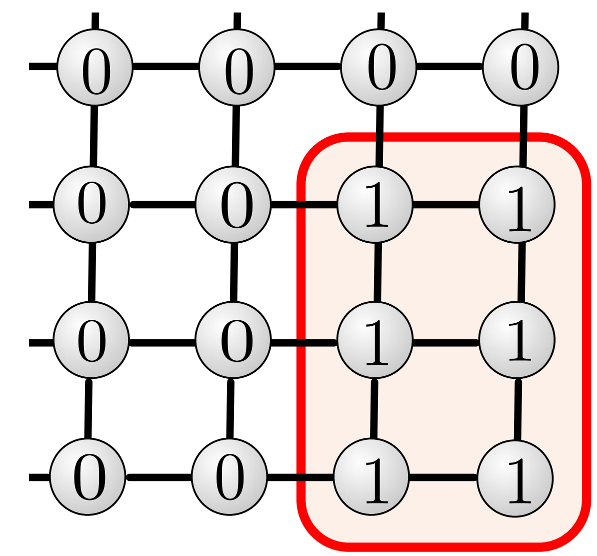
$$P(S) \propto \exp(F(S))$$

- equivalently: n binary random variables

Log-supermodularity

E.g. ferromagnetic Ising model / Conditional Random Field





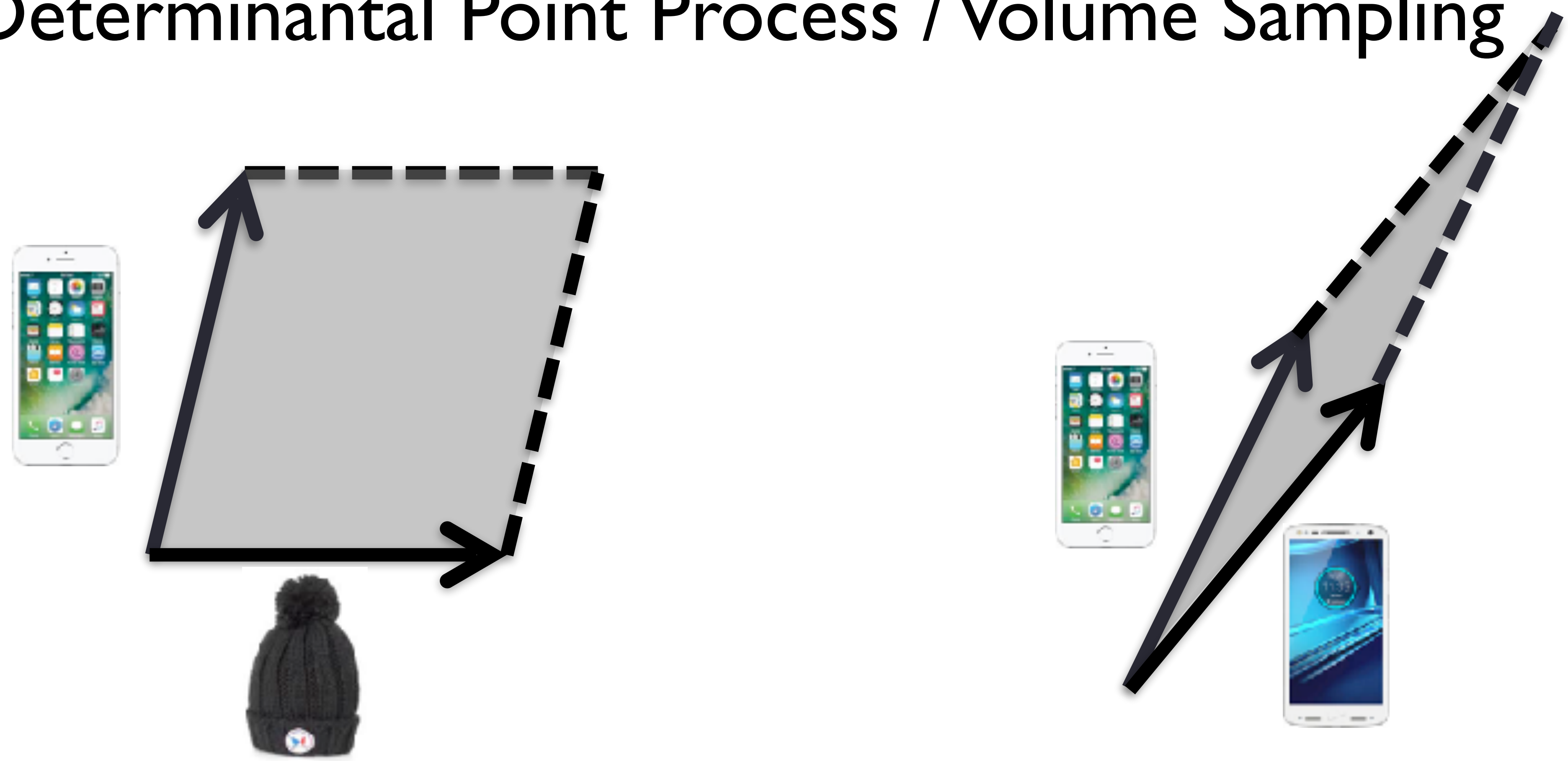
strong positive correlation:
"multivariate totally positive of order 2", "positive association"

Benefits:
finding the mode = minimizing a submodular function
approximating partition function & marginals …

(Fortuin-Kasteleyn-Ginibre 71, Kolmogorov-Zabih 04, Djolonga-Krause 14, …)

Log-Submodularity

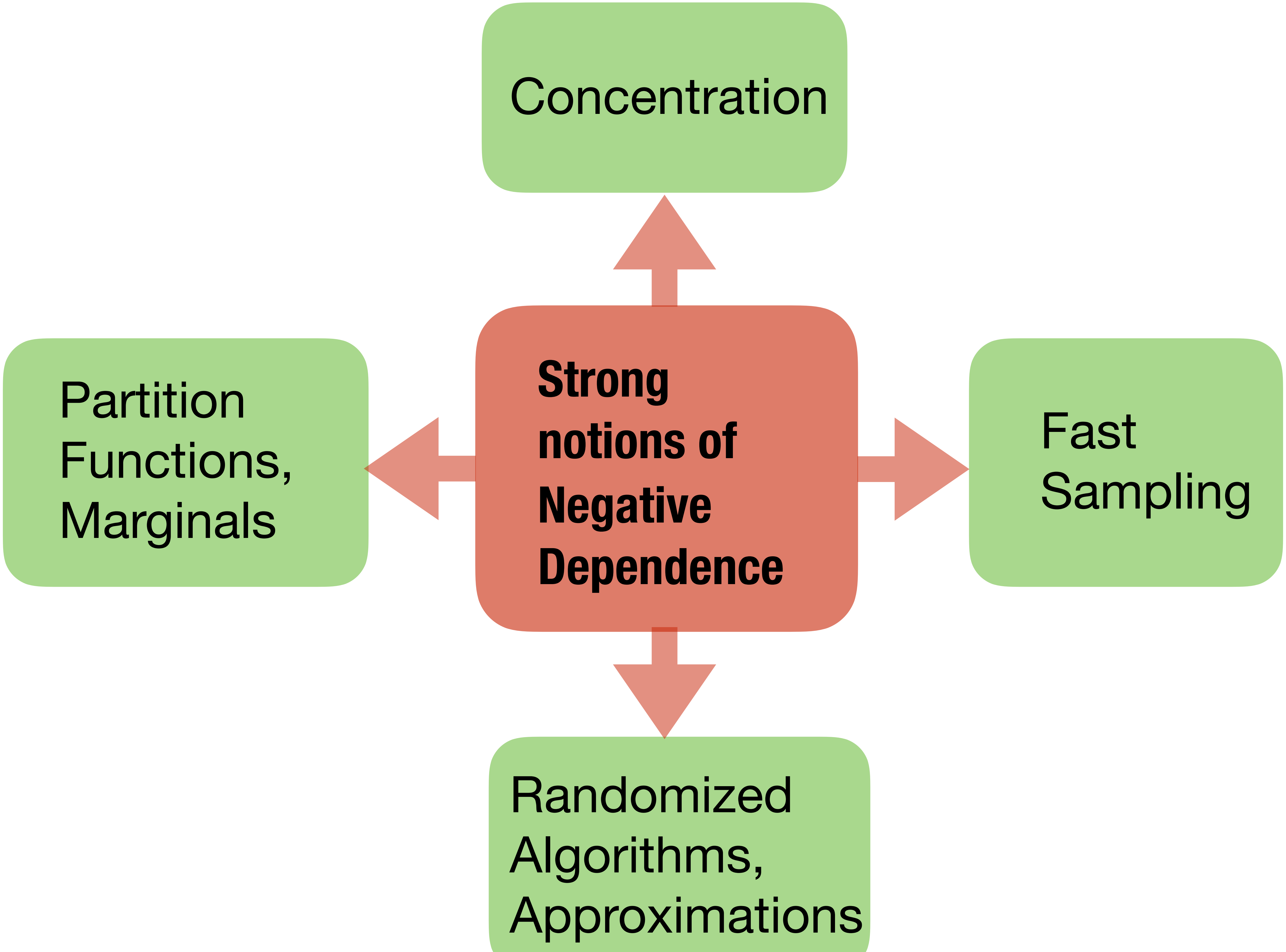
E.g. Determinantal Point Process / Volume Sampling



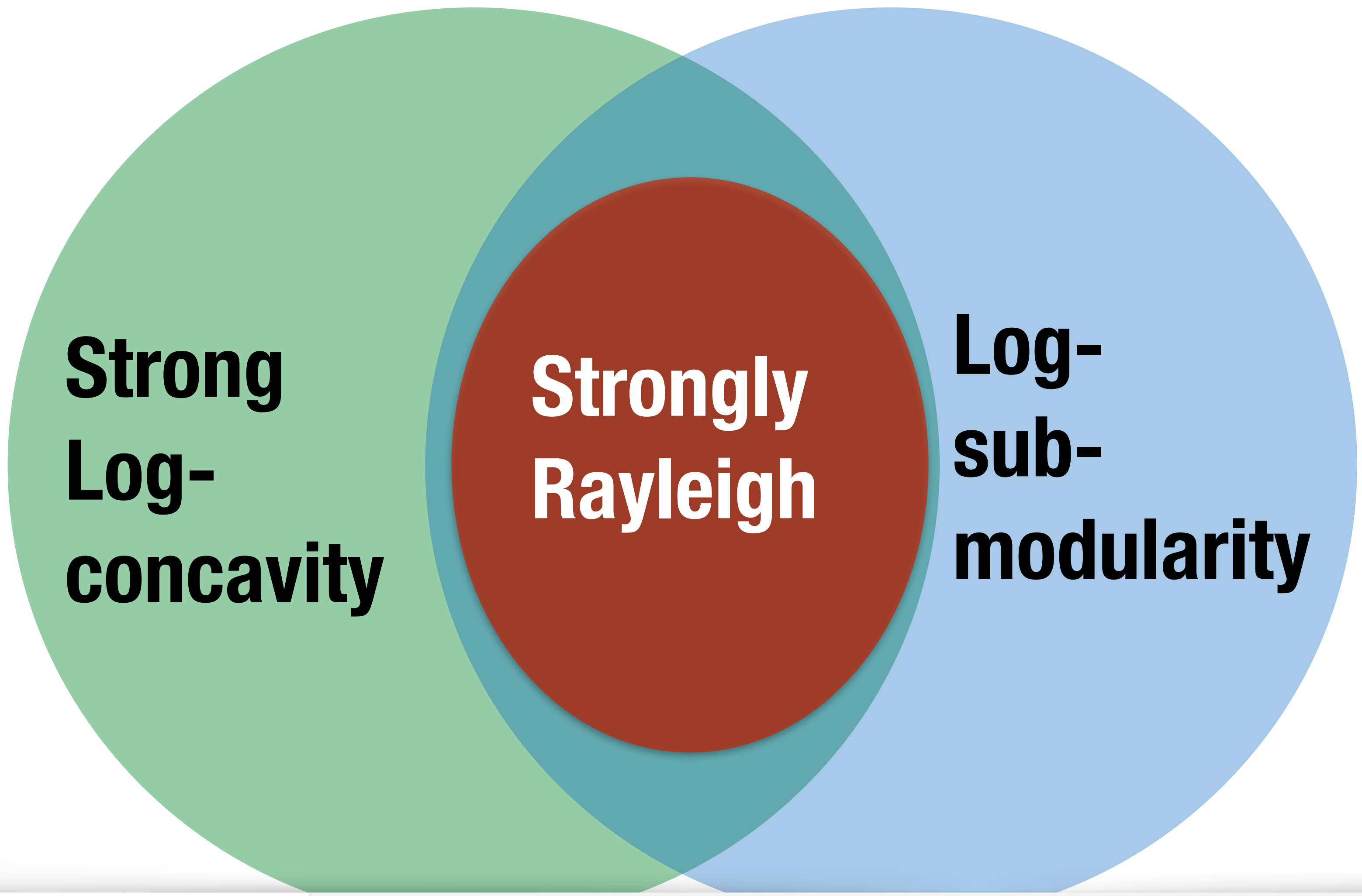
Negative Correlation / Repulsion

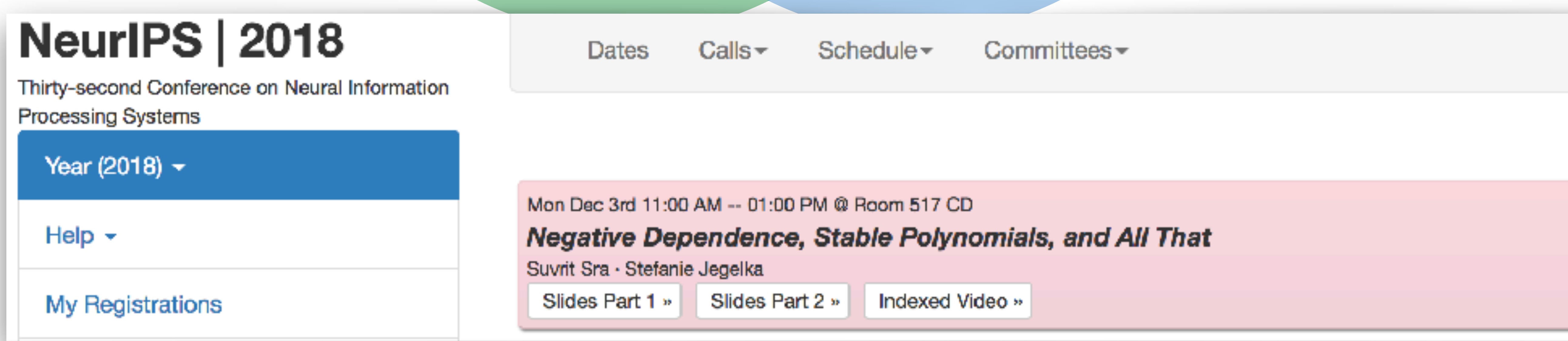
Sub-Family has strong negative dependence

(Macchi 75, Feder-Mihail 82, Borodin 02, Deshpande-Rademacher-Vempala-Wang 06, Borcea-Bränden 09, Borcea-Bränden-Liggett 09, Kulesza-Taskar 12, Anari-Oveis Gharan-Rezaei 16, Li-Jegelka-Sra 16, …)


Concentration
Partition Functions, Marginals
Strong notions of Negative Dependence
Fast Sampling
Randomized Algorithms, Approximations

Notions of Negative Dependence





Other directions

- **Approximate Submodularity**
  Optimization results hold with additional constants
- **Stochastic Optimization**
  Submodular function is an expectation
- **Robust Optimization**
  Adversarial Perturbations of Inputs
- **Sequential Decisions / Planning**

- **Learning Submodular Functions**

nonconvex optimization
lattice / continuous submodularity
many optimization results generalize

probability measures
log-supermodular
log-submodular
sampling, mode, approx. partition function

submodular set functions

convexity:
minimization
maximize coherence

dim. returns (concavity):
maximization
maximize diversity

many examples:

- linear/modular functions
- entropy
- mutual information
- rank functions
- coverage
- diffusion in networks
- volume
- graph cut …