



# Submodular Functions: Definitions, Examples, and Information Measures

## Lecture 7

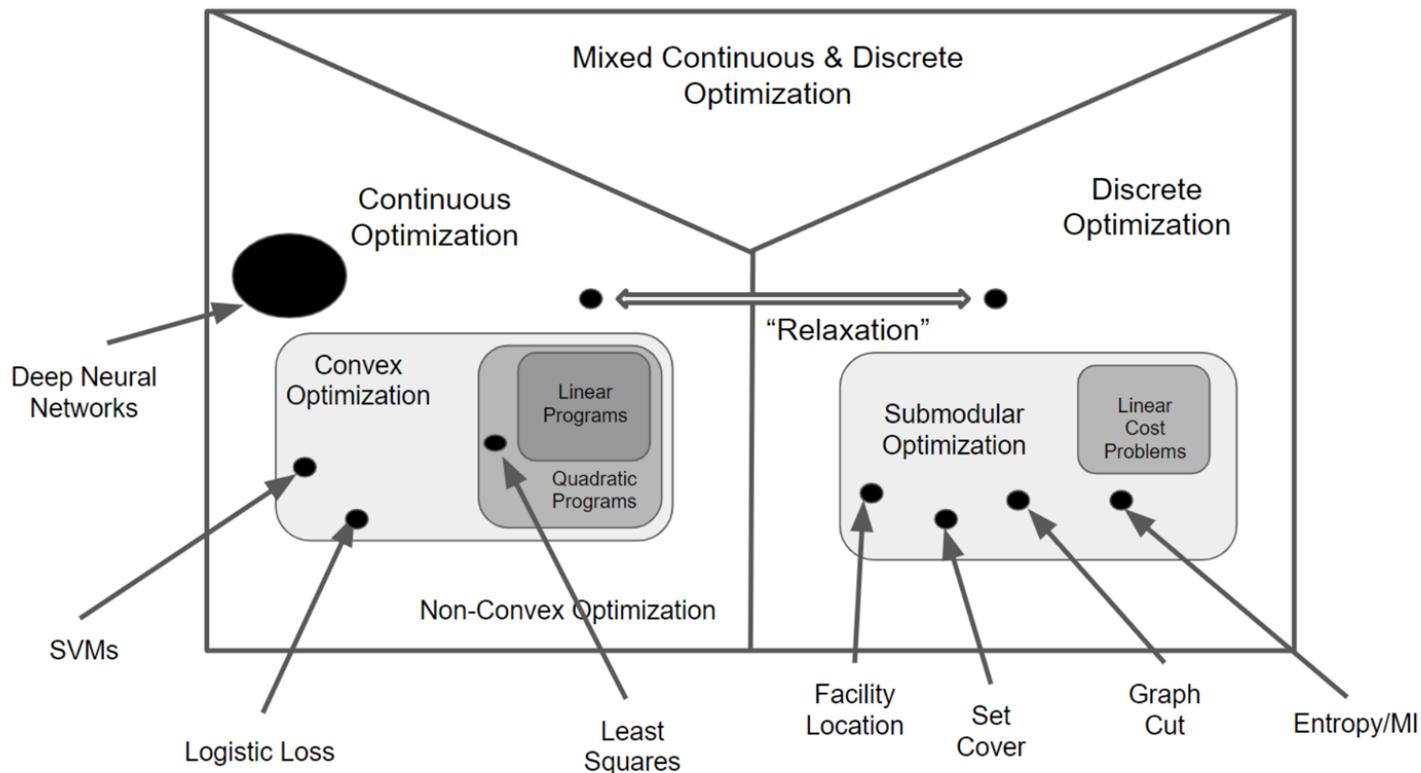
---

Advanced Topics in Optimization For Machine Learning  
(CS 7301)

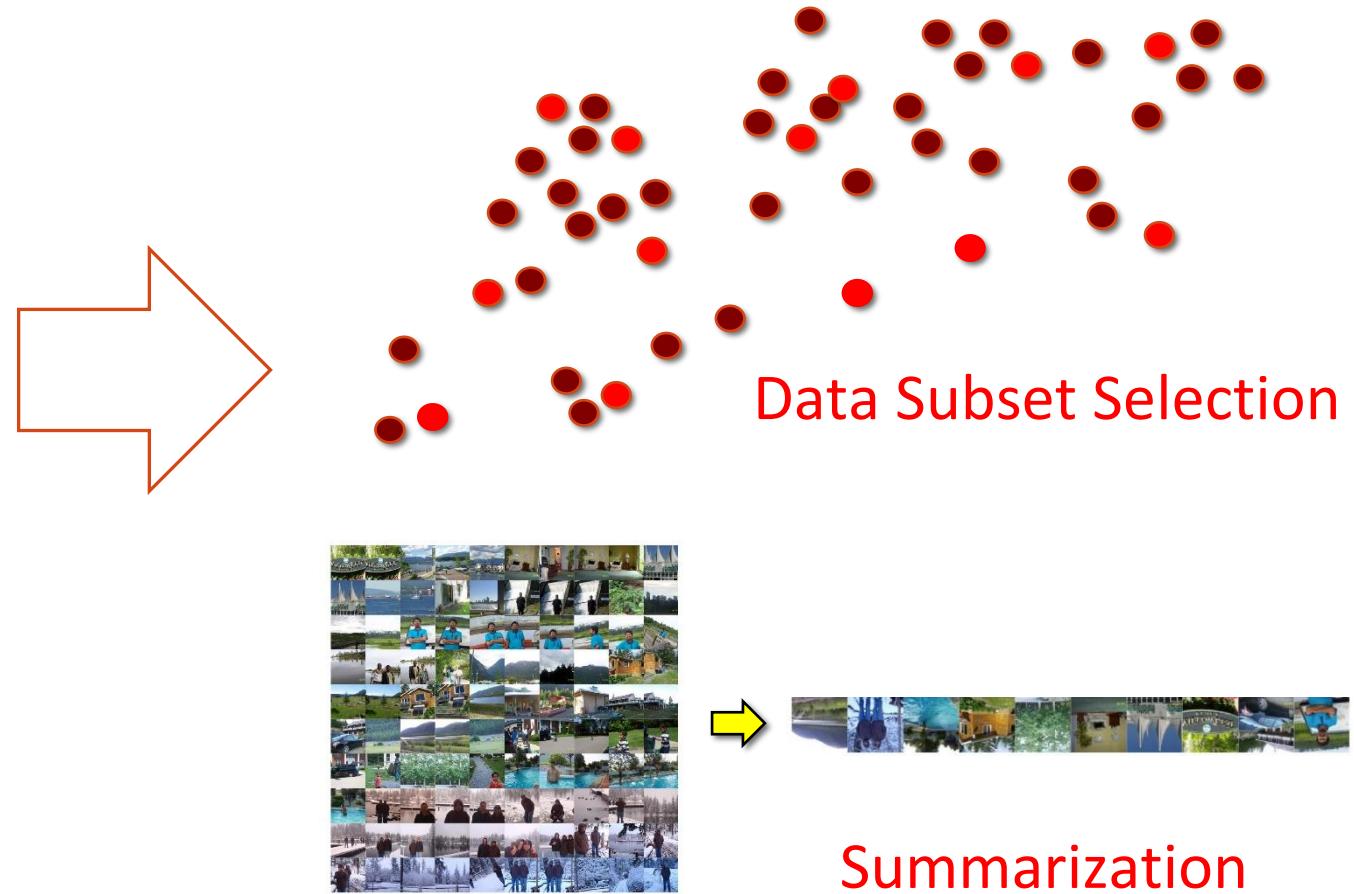
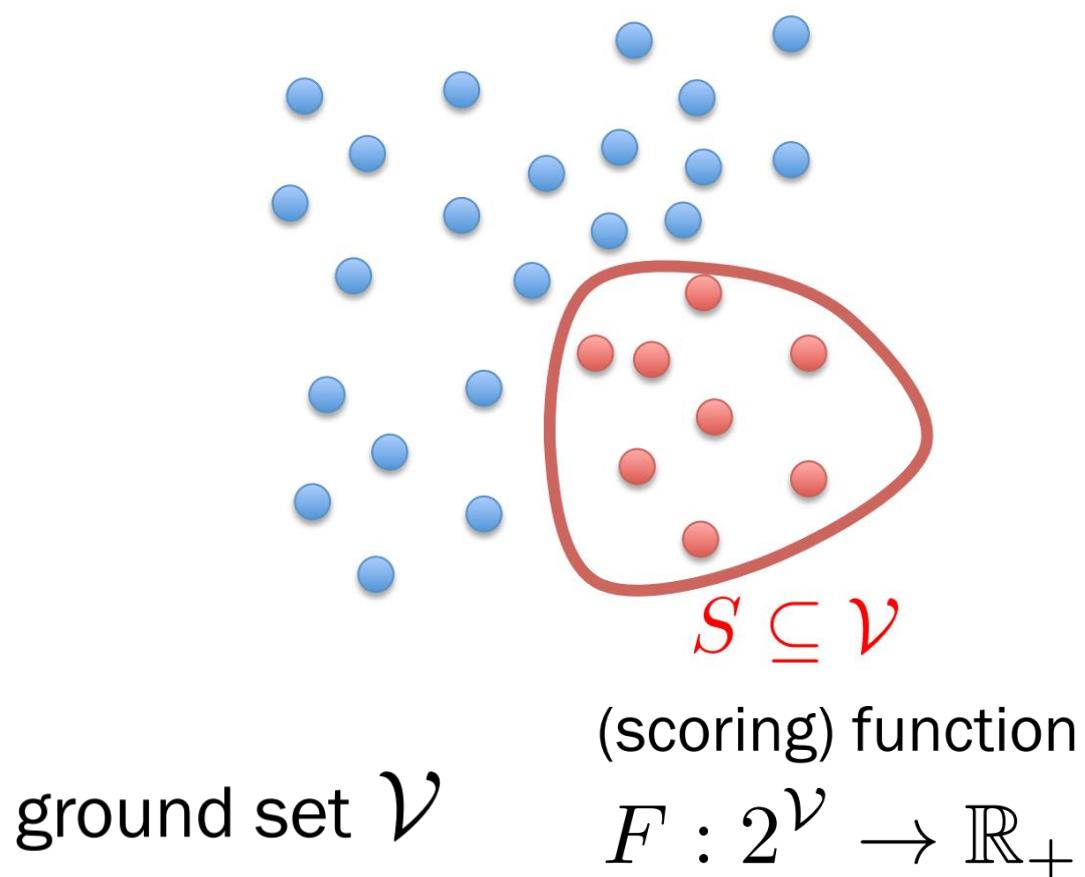
Instructor: Rishabh Iyer

# Big Picture: Continuous and Discrete Optimization

---



# Discrete Optimization



# Outline

---

- ❑ Discrete Optimization in Machine Learning
- ❑ **Lecture 7.1: Submodular Functions: Properties and Examples**
  - ❑ Definition and Intuition of Submodularity
  - ❑ Modeling Power of Submodular/Set Functions
  - ❑ Examples of Submodular Functions
  - ❑ Properties of Submodular Functions
- ❑ **Lecture 7.2: Submodular Information Measures (SIMs)**
  - ❑ Definitions of SIMs: Conditional Gain, Submodular Mutual Information, Submodular Conditional Mutual Information
  - ❑ Properties of SIMs
  - ❑ Examples of SIMs
  - ❑ Applications of SIMs

# Acknowledgements

---

Slides borrowed from several sources:

1. Submodular Optimization course at UW from Jeff Bilmes
2. Tutorial on Submodular Optimization by Stefanie Jegelka, Andreas Krause and Jeff Bilmes at ICML and NIPS
3. Some of my own tutorials at WACV, IJCAI, ECAI, SPCOM etc.

# Useful Material

---

- Fujishige, “Submodular Functions and Optimization”, 2005
- Narayanan, “Submodular Functions and Electrical Networks”, 1997
- Welsh, “Matroid Theory”, 1975
- Oxley, “Matroid Theory”, 1992 (and 2011).
- Lawler, “Combinatorial Optimization: Networks and Matroids”, 1976.
- Schrijver, “Combinatorial Optimization”, 2003
- Gruenbaum, “Convex Polytopes, 2nd Ed”, 2003.
- Additional readings that will be announced here.

# Useful material

---

- Jeff's Class: [https://people.ece.uw.edu/bilmes/classes/ee563\\_spring\\_2018/](https://people.ece.uw.edu/bilmes/classes/ee563_spring_2018/)
- Stefanie Jegelka & Andreas Krause's 2013 ICML tutorial:  
<http://techtalks.tv/talks/submodularity-in-machine-learning-new-directions-part-i/58125/>
- Jeff's NIPS, 2013 tutorial on submodularity:  
<http://melodi.ee.washington.edu/~bilmes/pgs/b2hd-bilmes2013-nips-tutorial.html> and <http://youtu.be/c4rBof38nKQ>
- Andreas Krause's web page: <http://submodularity.org>
- Francis Bach's updated 2013 text: [http://hal.archives-ouvertes.fr/docs/00/87/06/09/PDF/submodular\\_fot\\_revisd\\_hal.pdf](http://hal.archives-ouvertes.fr/docs/00/87/06/09/PDF/submodular_fot_revisd_hal.pdf)
- Tom McCormick's overview paper on submodular minimization:  
<http://people.commerce.ubc.ca/faculty/mccormick/sfmchap8a.pdf>
- My recent papers on Submodular Information Measures  
(<http://proceedings.mlr.press/v132/iyer21a/iyer21a.pdf>, <https://arxiv.org/pdf/2103.00128.pdf>)

# Motivation: Big Data in Machine Learning

---

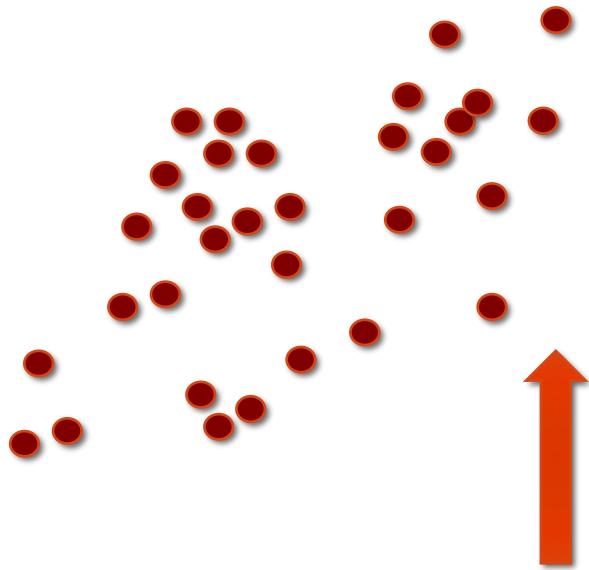


Small Data

Each Data Point represents  
Visual features in n-dimensional space  
(for example objects or scenes in an  
image)

# Motivation: Big Data in Machine Learning

---

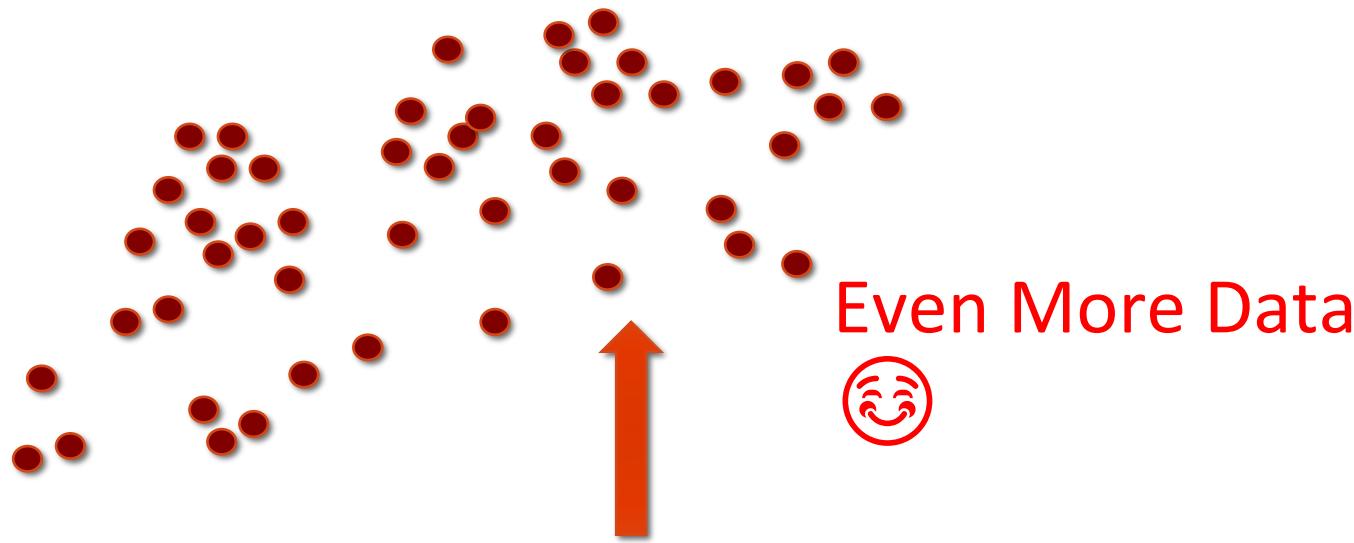


Little More Data

Each Data Point represents  
Visual features in n-dimensional space  
(for example objects or scenes in an  
image)

# Motivation: Big Data in Machine Learning

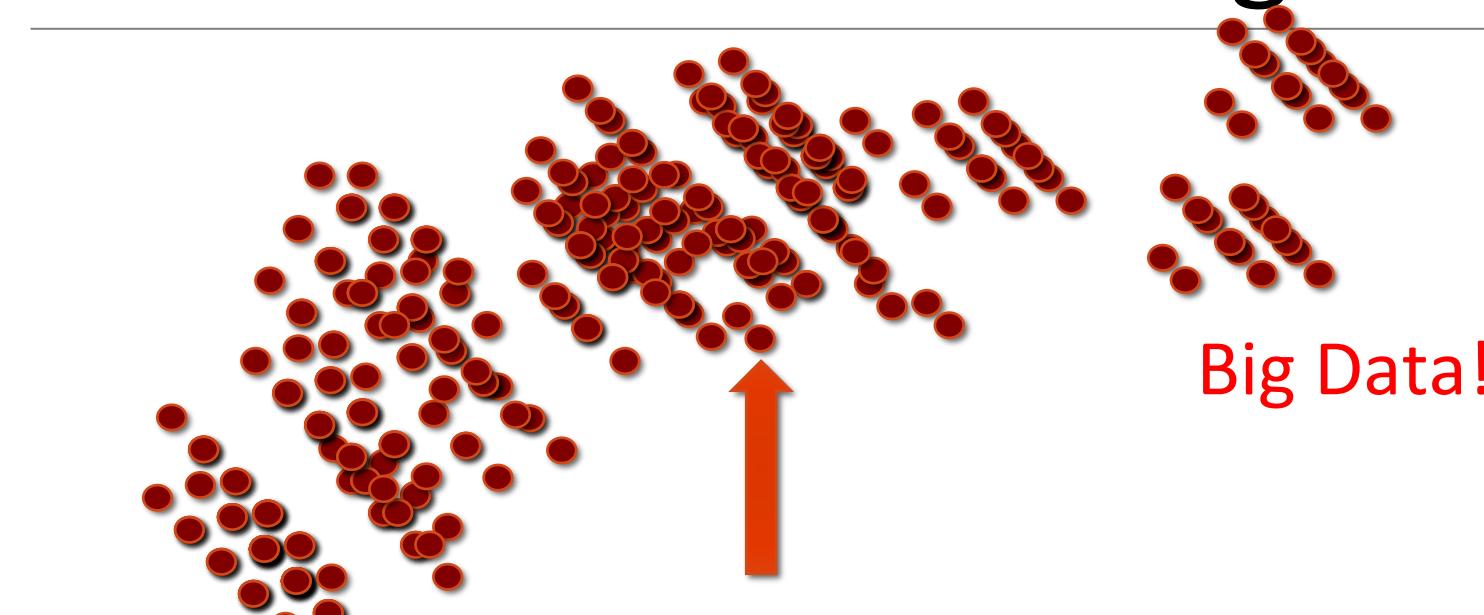
---



The more Data we get, the better Is our understanding of the space, And correspondingly, the better will be the performance of Statistical algorithms

# Motivation: Big Data in Machine Learning

---



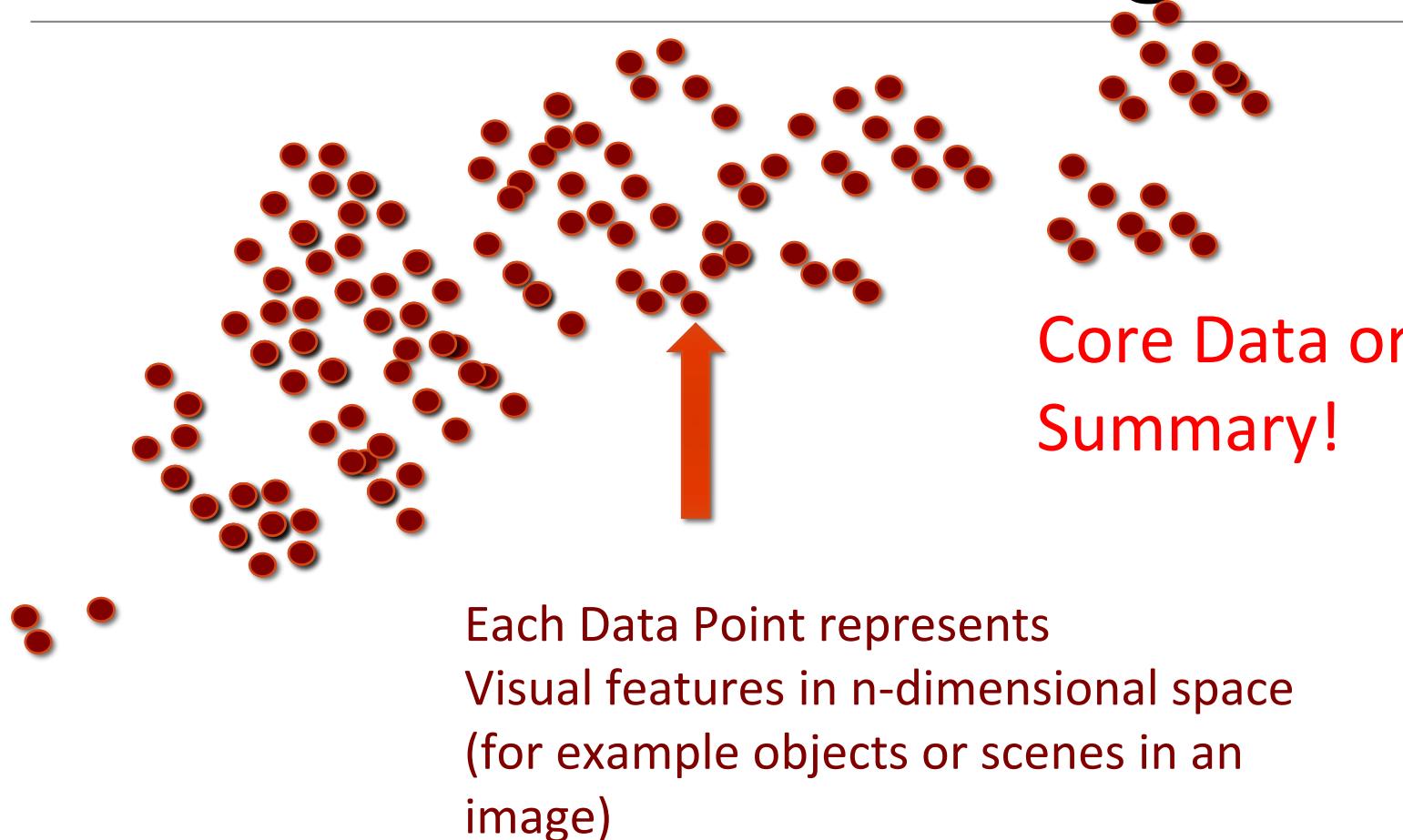
Each Data Point represents  
Visual features in n-dimensional space  
(for example objects or scenes in an  
image)

Has the complete picture  
of the distribution of data.

However, Clearly there is a  
lot of redundancy,  
And wasteful utilization  
of the resources!

# Motivation: Big Data in Machine Learning

---

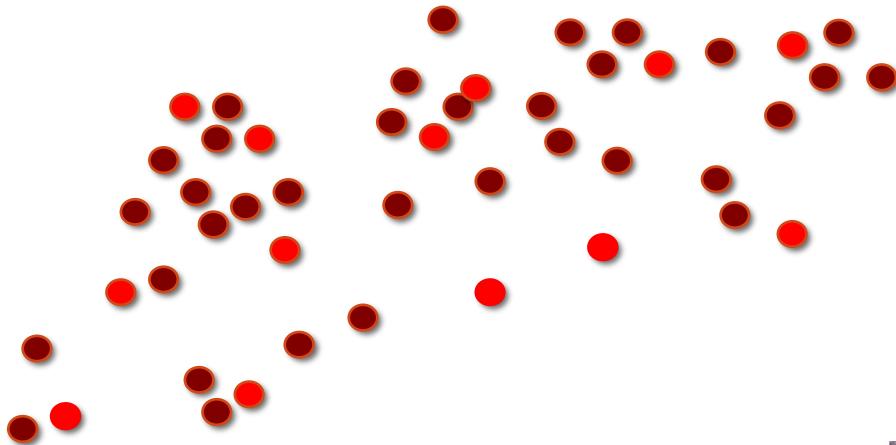


Maintains the Representation of the Data, while requiring Much lesser number Of samples

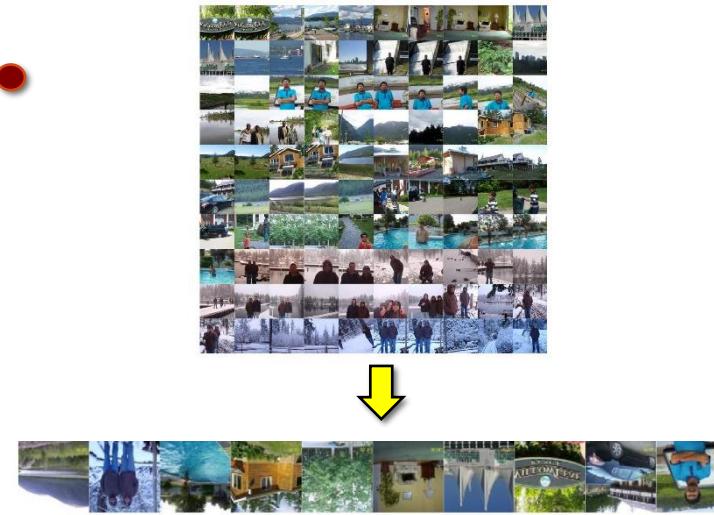
⇒ Lesser Training and Labeling Costs!

# Combinatorial Subset Selection Problems

---



Data Subset Selection



Summarization



Image Segmentation

Subset Selection Problems everywhere!

# Discrete Optimization in Machine Learning

---

- MAP inference in Probabilistic Models: Ising Models, DPPs
- Feature Subset Selection
- Data Partitioning
- Data Subset Selection
- Data Summarization: Text, Images, Video Summarization
- Social networks, Influence Maximization
- Natural Language Processing: words, phrases, n-grams, syntax trees, semantic structures
- Computer Vision: Image Segmentation, Image Correspondence
- Genomics and Computational Biology: cell types or assay selection, selecting peptides and proteins

# Outline

---

- ❑ Discrete Optimization in Machine Learning
- ❑ **Lecture 19 & 20**
  - ❑ Definition and Intuition of Submodularity
  - ❑ Modeling Power of Submodular/Set Functions
  - ❑ Examples of Submodular Functions
  - ❑ Examples of Submodular Optimization
- ❑ Lectures 21 & 22
  - ❑ Optimization Algorithms for Different Function Classes and Constraints
  - ❑ Optimization Algorithms for Different Settings (Streaming/Distributed)
  - ❑ Practical Implementation tricks

# Combinatorial Subset Selection Problems

$$V = \{ \text{Banana, Milk, Apple, } \\ \text{Strawberry, Car, Laptop, } \\ \text{T-shirt, Book, Coffee} \}$$

$$f : 2^V \rightarrow \mathbb{R}$$

$$A = \{ \text{Banana, } \\ \text{Strawberry, } \\ \text{Book, } \}$$

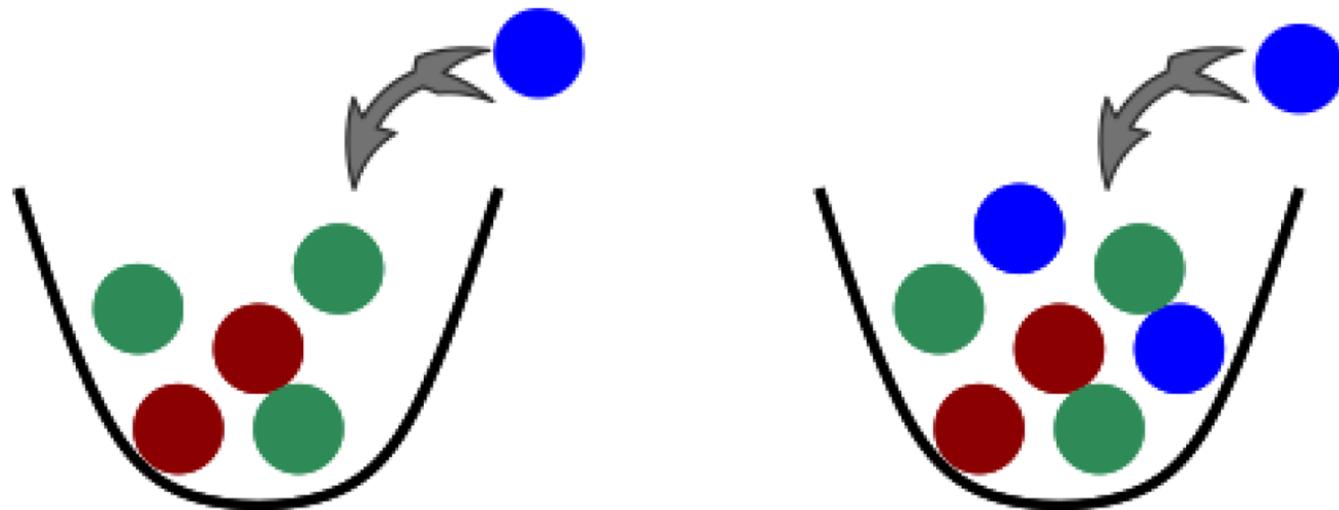
Choose Subset  $A \subseteq V$   
 $f(A) = 22$

General Set function Optimization: very hard!

What if there is some special structure?

# Submodular Functions

$$f(A \cup v) - f(A) \geq f(B \cup v) - f(B), \text{ if } A \subseteq B$$



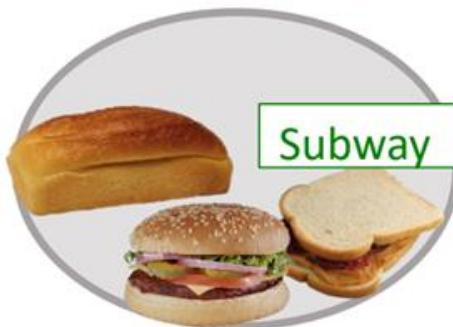
Negative of a  
Submodular  
Function is a  
Super-modular  
Function!

$f = \#$  of distinct colors of balls in the urn.

# Submodular Functions

$$f(\text{French Fries} \cup \text{McDonald's Soda}) - f(\text{French Fries}) \geq f(\text{Hamburger} \cup \text{French Fries} \cup \text{McDonald's Soda}) - f(\text{Hamburger})$$

Diminishing Returns!



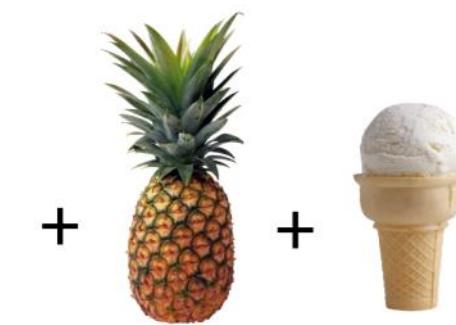
The more items  
you buy,  
the more the  
discount!

# Modular Functions

---

- each element  $e$  has a weight  $w(e)$

$$F(S) = \sum_{e \in S} w(e)$$



$$A \subset B$$

$$F(A \cup e) - F(A) = w(e) \quad = \quad F(B \cup e) - F(B) = w(e)$$

Modular Functions are both submodular and super-modular!

# Monotone Submodular Functions

---

- A set function is called **monotonic** if

$$A \subseteq B \subseteq V \Rightarrow F(A) \leq F(B)$$

- Examples:

- **Influence** in social networks [Kempe et al KDD '03]

- For discrete RVs, **entropy**  $F(A) = H(X_A)$  is monotonic:  
Suppose  $B = A \cup C$ . Then

$$F(B) = H(X_A, X_C) = H(X_A) + H(X_C | X_A) \geq H(X_A) = F(A)$$

- **Information gain**:  $F(A) = H(Y) - H(Y | X_A)$

# Submodularity (almost) everywhere



**THEORY OF CAPACITIES<sup>(1)</sup>**  
by Gustave CHOQUET<sup>(2)(3)</sup>.

## INTRODUCTION

This work originated from the following significance had been emphasized by M. Brelot  
Is the interior Newtonian capacity of an arbitrary subset  $X$  of the space  $R^3$  equal to the exterior capacity of  $X$ ?



## Cores of Convex Games<sup>1)</sup>

By LLOYD S. SHAPLEY<sup>2)</sup>

*ct:* The core of an  $n$ -person game is the set of feasible outcomes that are stable under the coalition of players. A convex game is defined as one that is balanced. In this paper it is shown that the core of a convex game is not empty. The structure of the core is further analyzed. It is further shown that certain other cooperative solutions are also stable. The value of a convex game is the center of gravity of the core.

## Submodular Functions, Matroids, and Certain Polyhedra\*

Jack Edmonds

National Bureau of Standards, Washington, D.C., U.S.A.

I

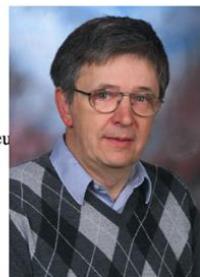


The viewpoint of the subject of matroids, and related areas of lattice theory has always been, in one way or another, abstraction of algebraic dependence or equivalently, abstraction of the incidence relations in geometric representation of algebra. Often one of the main derived facts is that all bases have the same cardinality. (See Van der Waerden, Section 33.)

## Submodular functions and convexity

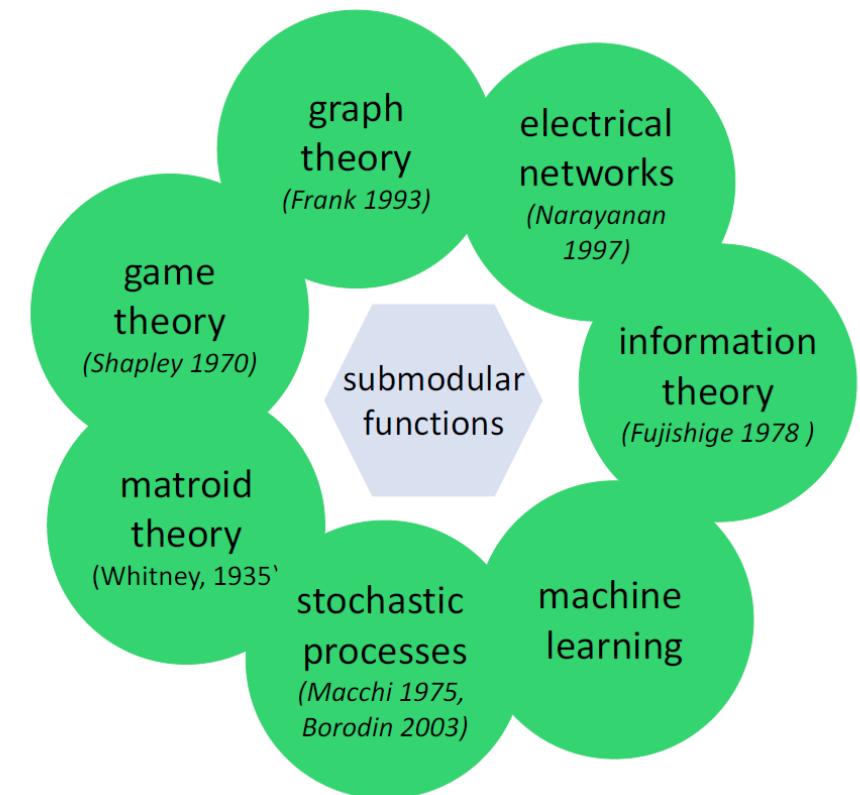
L. Lovász

Eötvös Loránd University, Department of Analysis I, Múzeum krt. 6, Budapest, Hungary



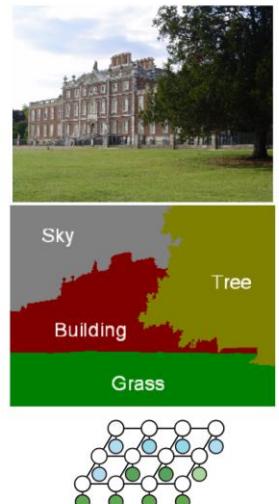
## 0. Introduction

In “continuous” optimization convex functions play a central role. Besides elementary tools like differentiation, various methods for finding the minimum of a convex function constitute the main body of nonlinear optimization.



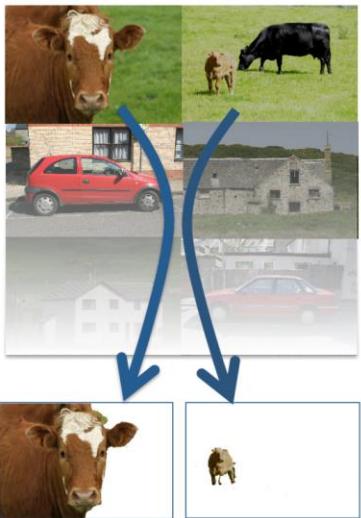
# Submodular Optimization in Machine Learning

---



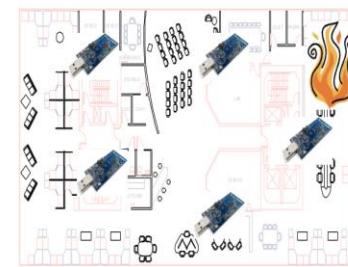
$F(S)$  = coherence + likelihood

Discrete Labeling



$F(S)$  = relevance + diversity or coverage

Summarization



- where put sensors?
- which experiments?
- summarization

$F(S)$  = “information”

Sensor Placement

# Submodular Functions for Summarization

---



Subset



C



Superset

# Submodular Functions for Summarization



Subset



C



Superset

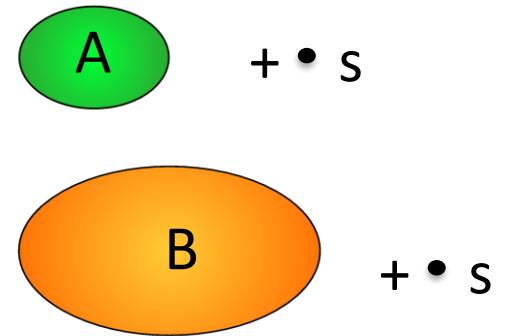
Information gain reduces with larger sets!

# Two Equivalent Definitions of Submodularity

---

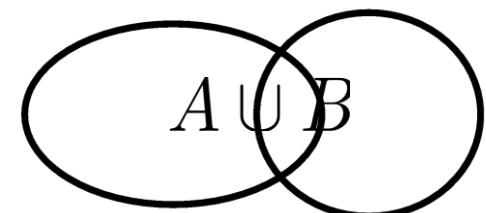
- Diminishing gains: for all  $A, B \subseteq V$

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



- 
- Union-Intersection: for all  $A, B \subseteq V$

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



# Instantiations of Set Functions

---

## Representation Functions

- Facility Location Function (k-mediods clustering)
- Graph Cut Family, Saturated Coverage

## Diversity Functions

- Dispersion Functions (Min, Sum, Min-Sum)
- Determinantal Point Processes

## Coverage Functions

- Set Cover Function
- Probabilistic Set Cover Function
- Feature Based Functions

## Importance Functions

- Modular Functions

## Information Functions

- Mutual Information
- Entropy

## Discounted Cost Functions

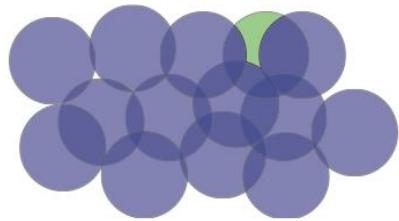
- Clustered Concave over Modular Functions
- Cooperative Costs and Saturations

## Complexity Functions

- Bipartite Neighborhood Functions

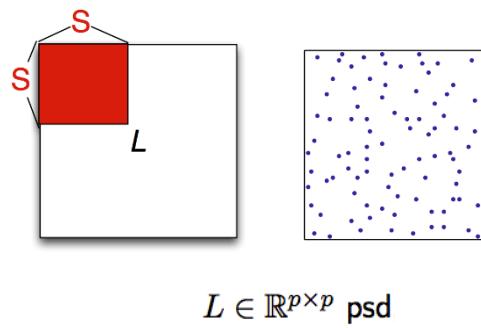
# Facets of Submodularity: Maximization

---



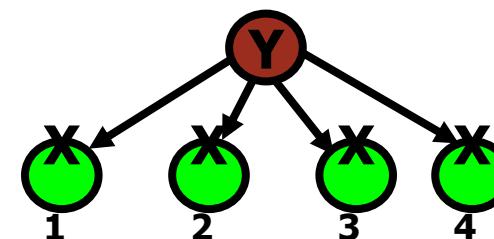
$$F(A) = \cup_{s \in A} \text{area}(s)$$

**Coverage**  
e



$$F(A) = \log \det(L_A)$$

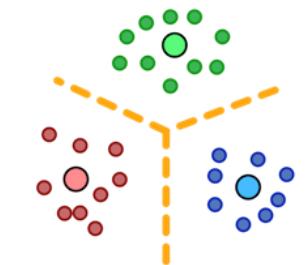
**Diversity**



$$F(A) = H(X_A)$$

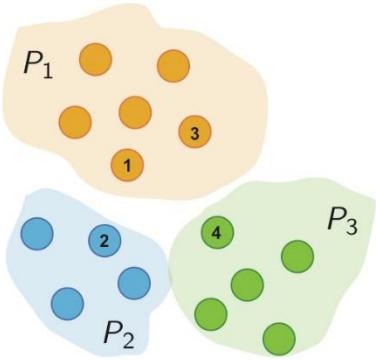
**Information**

$$F(A) = \sum_{i \in V} \max_{j \in A} s_{ij}$$



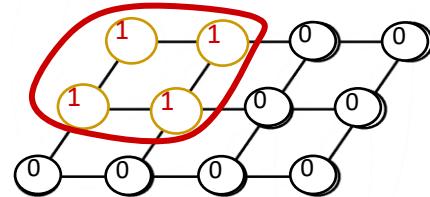
**Representation**

# Facets of Submodularity: Minimization



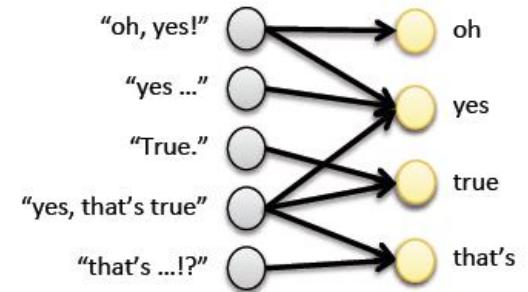
$$F(A) = \sum_{i=1}^3 \sqrt{\sum_{j \in A \cap P_i} r_i}$$

**Cooperative Costs**



$$E(\mathbf{x}; \mathbf{z}) = \sum_i E_i(x_i) + \sum_{ij} E_{ij}(x_i, x_j)$$

**Attractive Potentials**



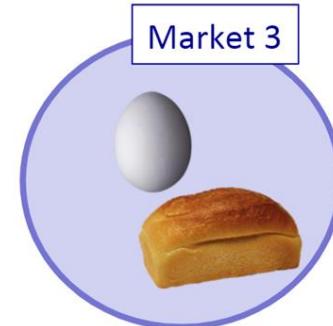
$$F(A) = \gamma(A)$$

**Complexity**

# Co-operative Costs



cost:  
time to shop  
+ price of items

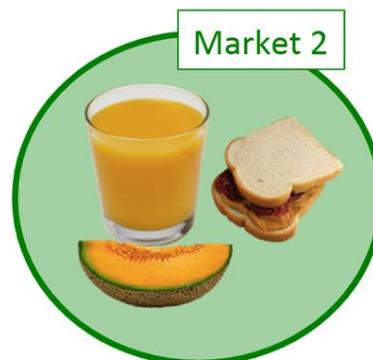


$$\begin{aligned} F(\text{coffee}, \text{orange slice}, \text{sandwich}) &= \text{cost}(\text{coffee}) + \text{cost}(\text{orange slice}, \text{sandwich}) \\ &= t_1 + 1 + t_2 + 2 \end{aligned}$$

= #shops + #items

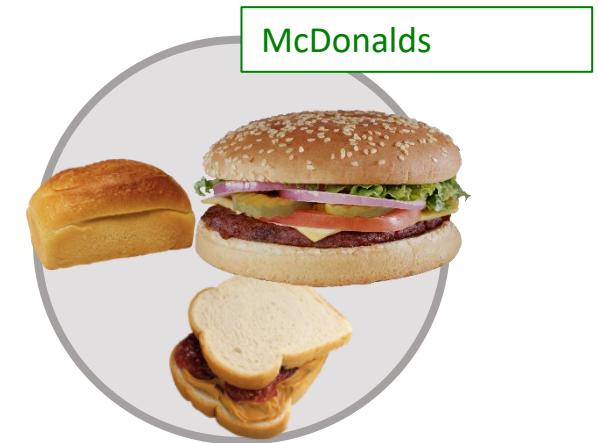
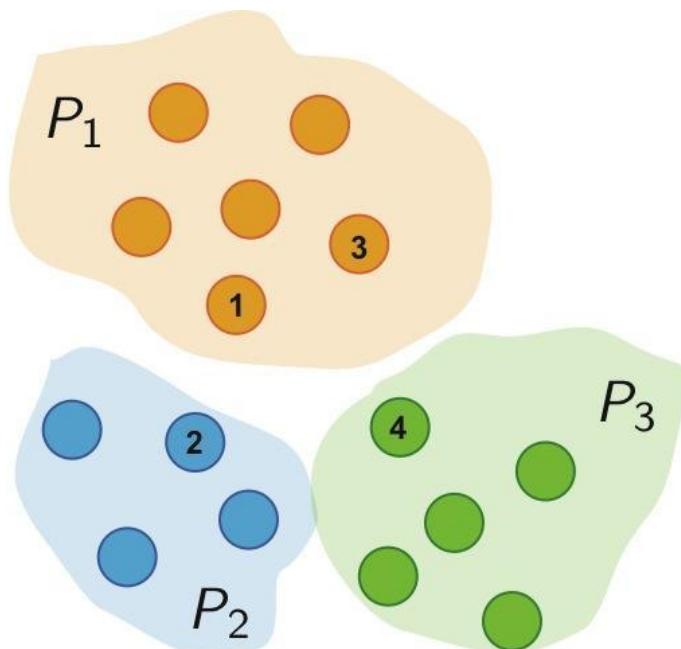


submodular?



# Cooperative Costs

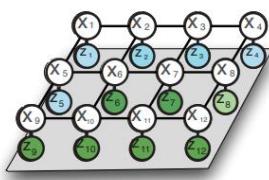
---



$$F(A) = \sum_{i=1}^3 \sqrt{\sum_{j \in A \cap P_i} r_i}$$

Iyer-Bilmes 2013, Jegelka-Bilmes 2011, ...

# Attractive Potentials

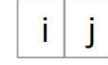


$$P(\mathbf{x} \mid \mathbf{z}) \propto \exp(-E(\mathbf{x}; \mathbf{z}))$$

$$E(\mathbf{x}; \mathbf{z}) = \sum_i E_i(x_i) + \sum_{ij} E_{ij}(x_i, x_j)$$

spatial coherence:

$$E_{ij}(1, 0) + E_{ij}(0, 1) \geq E_{ij}(0, 0) + E_{ij}(1, 1)$$



$$S = \{i\}$$

$$T = \{j\}$$

$$S \cap T = \emptyset$$

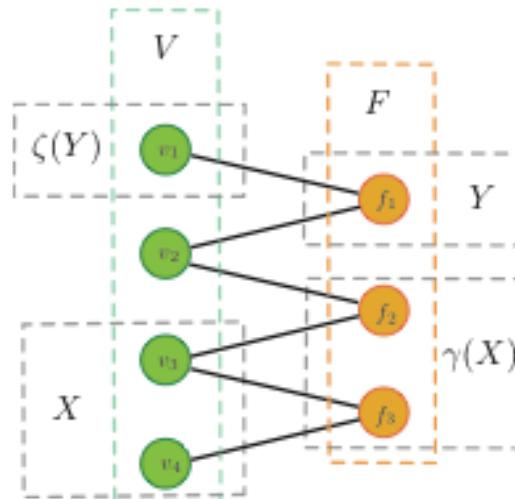
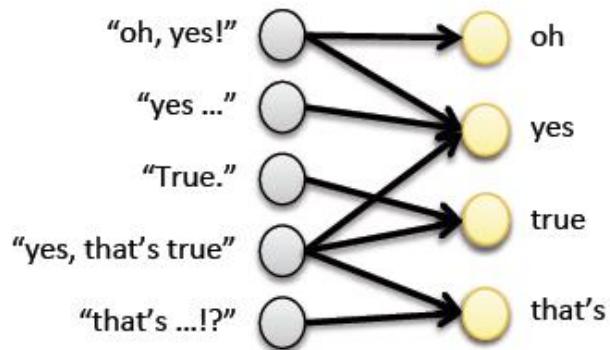
$$S \cup T$$

$$F(S) + F(T) \geq F(S \cup T) + F(S \cap T)$$



Boykov-Jolly 2001, ...

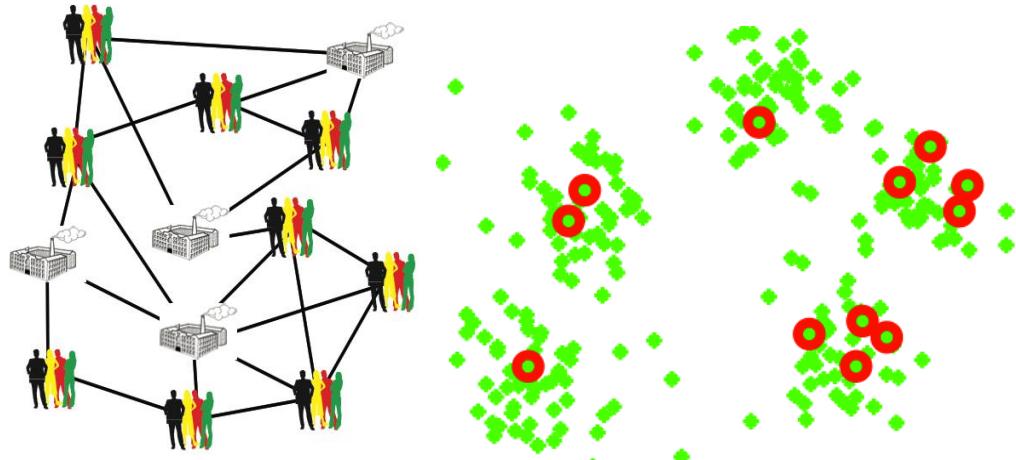
# Complexity Functions



Example: Selecting a limited complexity subset for quick Experimental turn around time!

$$F(A) = \gamma(A)$$

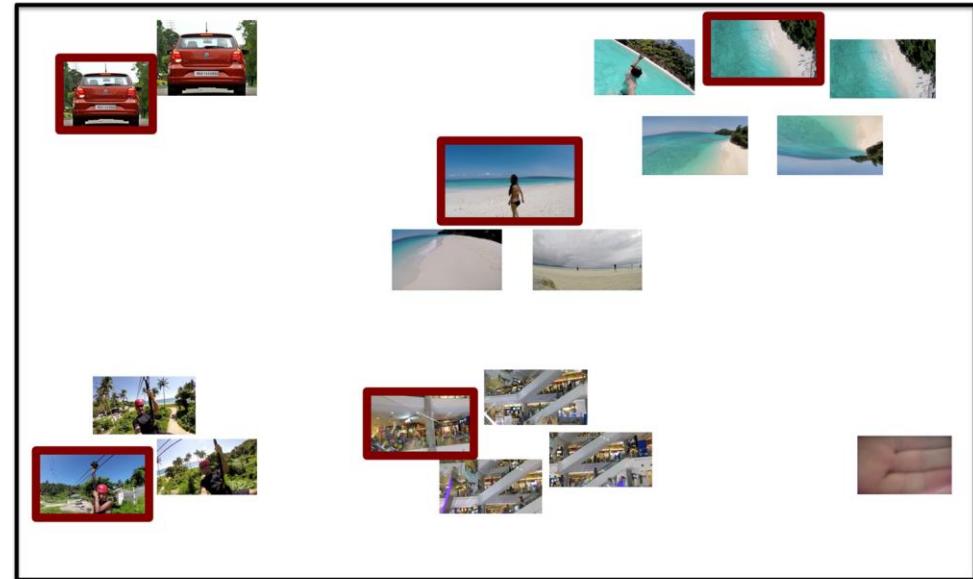
# Representation Functions



Facility Location	$\sum_{i \in V} \max_{k \in X} s_{ik}$
Saturated Coverage	$\sum_{i \in V} \min\{\sum_{j \in X} s_{ij}, \alpha_i\}$
Graph Cut	$\lambda \sum_{i \in V} \sum_{j \in X} s_{ij} - \sum_{i, j \in X} s_{ij}$

↑

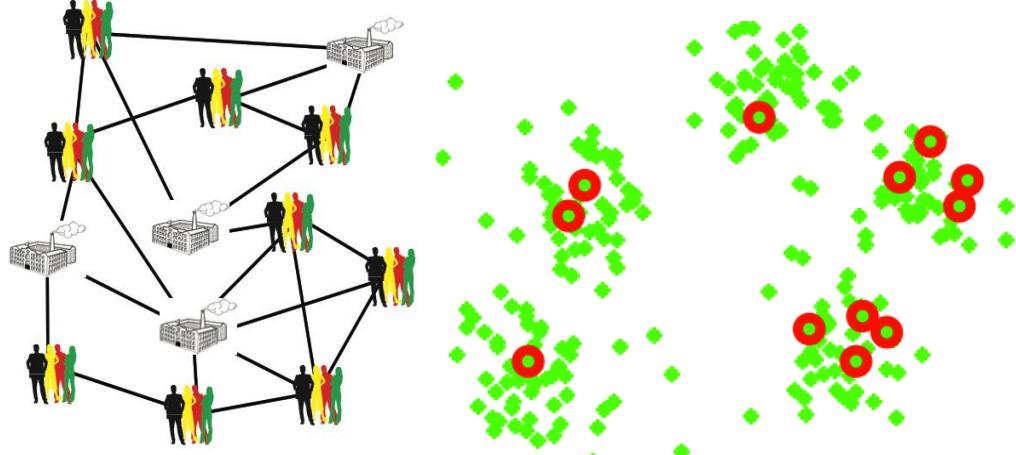
Similarity Kernel



Representation Functions  
Picks Centroids

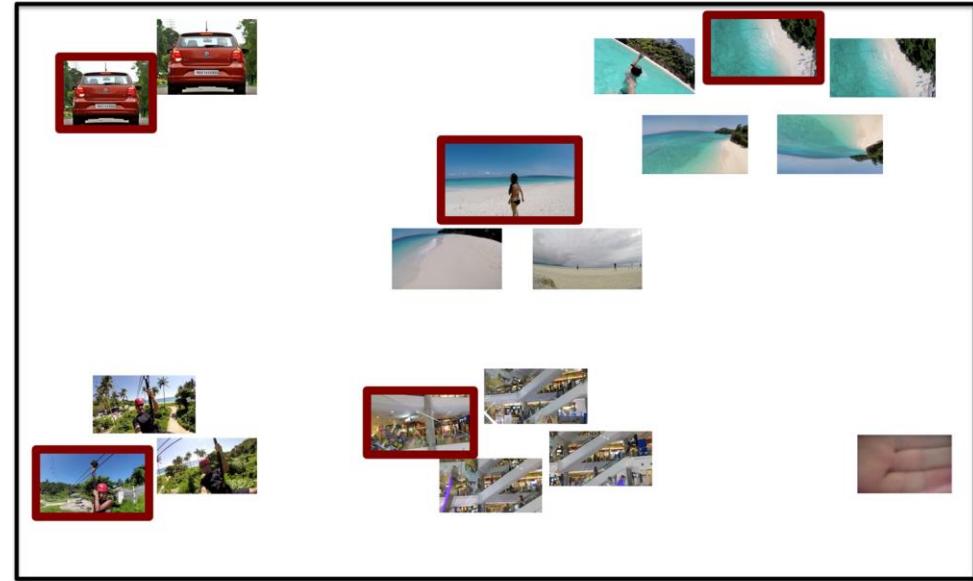
Iyer 2015, Kaushal et al 2019, Tschiatchek et al 2014, ...

# Representation Functions



Facility Location	$\sum_{i \in V} \max_{k \in X} s_{ik}$
Saturated Coverage	$\sum_{i \in V} \min\{\sum_{j \in X} s_{ij}, \alpha_i\}$
Graph Cut	$\lambda \sum_{i \in V} \sum_{j \in X} s_{ij} - \sum_{i, j \in X} s_{ij}$

**Graph Cut is not monotone submodular  
when  $\lambda < 2$**

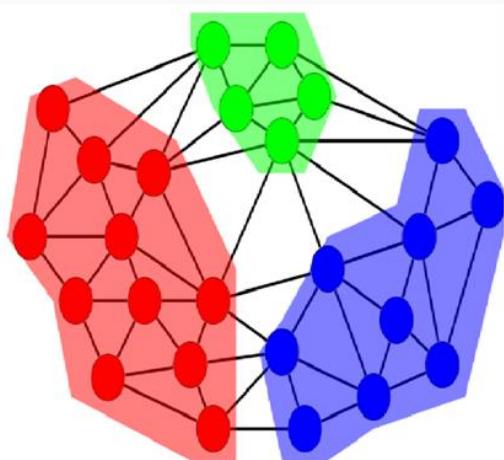


Representation Functions  
Picks Centroids

Iyer 2015, Kaushal et al 2019, Tschiatchek et al 2014, ...

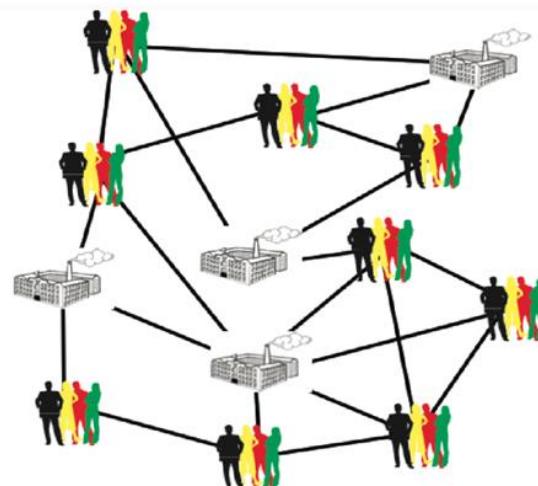
# Representation Functions

Characterized by **similarity kernel**  $s_{ij}$  between elements  $i$  and  $j$



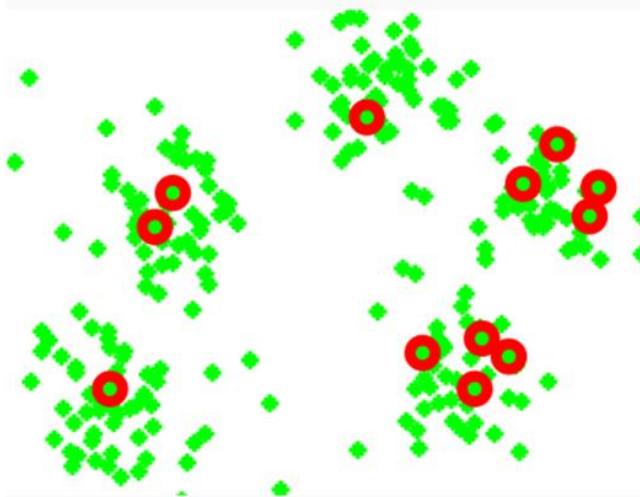
**Graph Cut:** Not monotone  
submodular when  $\lambda < 2$

$$\lambda \sum_{i \in V} \sum_{j \in X} s_{ij} - \sum_{i,j \in X} s_{ij}$$



**Facility Location:**

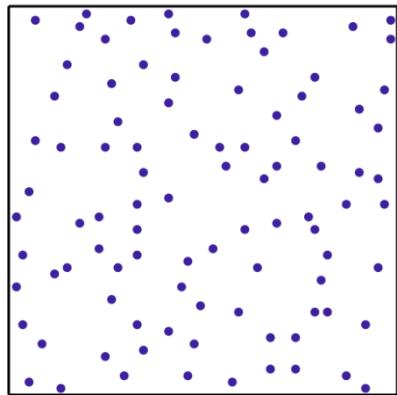
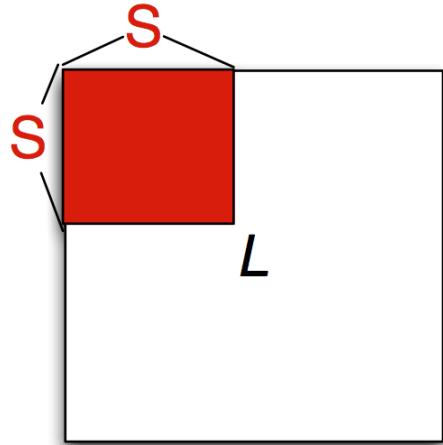
$$\sum_{i \in V} \max_{k \in X} s_{ik}$$



**Saturated Coverage:**

$$\sum_{i \in V} \min \left\{ \sum_{j \in X} s_{ij}, \alpha_i \right\}$$

# Diversity Functions: DPPs

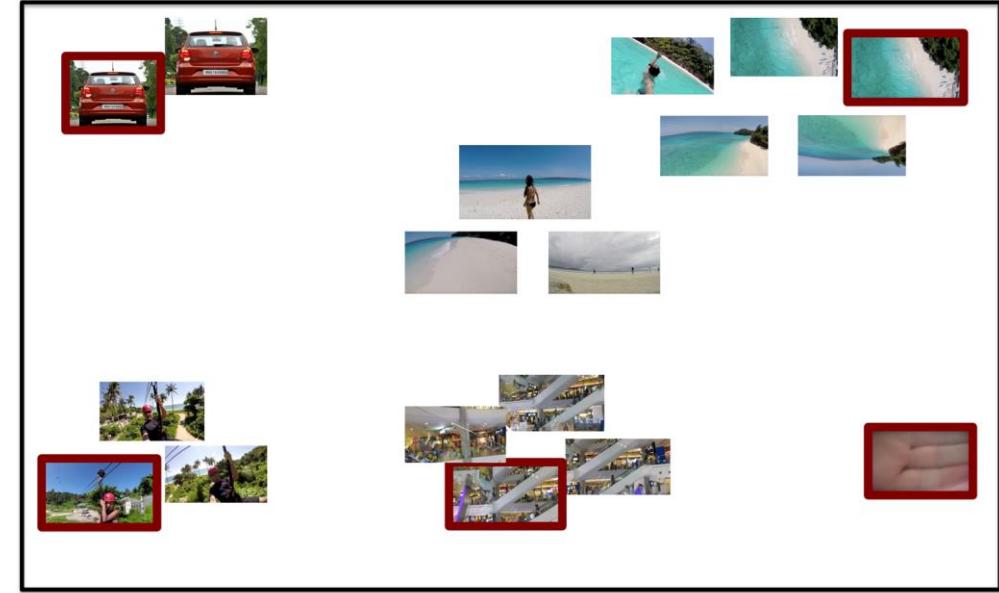


Determinantal Point Processes

$$F(S) = \log \det(L_S)$$



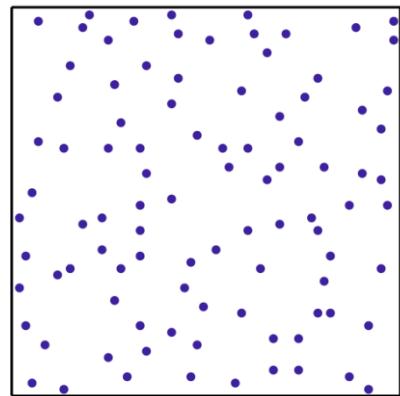
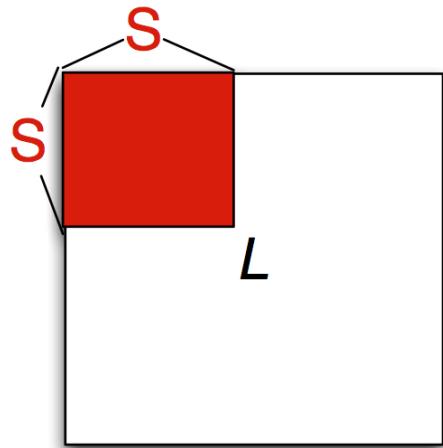
Similarity Kernel



Diversity Functions  
Picks items as different as possible!

Kulesza-Taskar 2012, ...

# Diversity Functions: DPPs

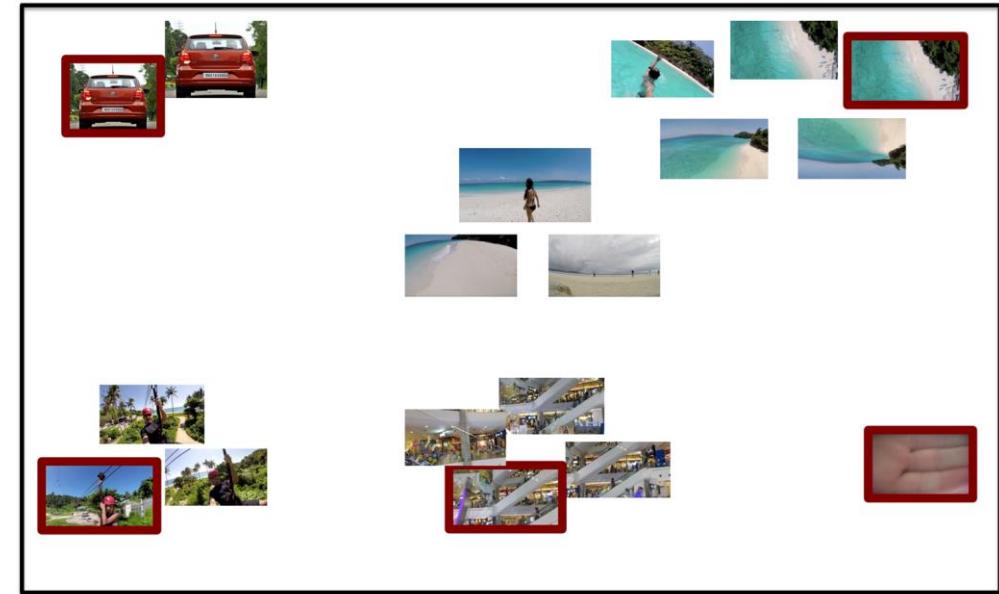


Determinantal Point Processes

$$F(S) = \log \det(L_S)$$



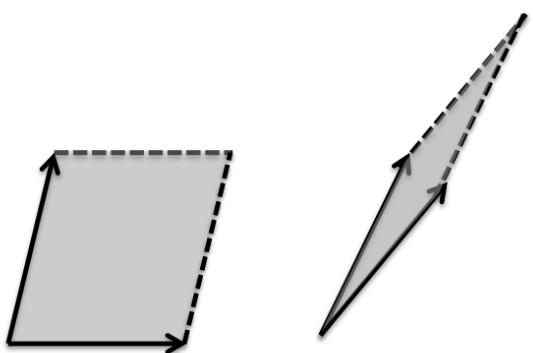
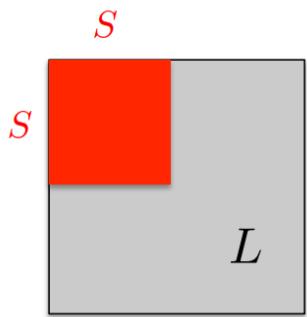
Log-Det Function is Non-Monotone Submodular!



Diversity Functions  
Picks items as different as possible!

Kulesza-Taskar 2012, ...

# Determinantal Point Processes



- similarity matrix  $L$

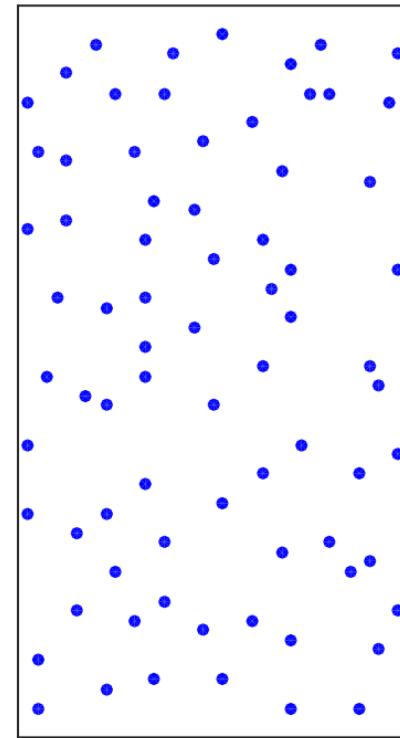
$$L_{ij} = x_i^\top x_j$$

- sample set  $Y$ :

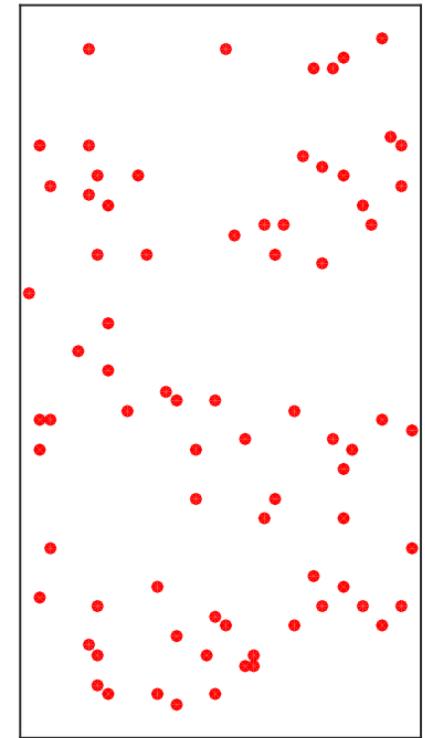
$$\begin{aligned} P(Y = S) &\propto \det(L_S) \\ &= \text{Vol}(\{x_i\}_{i \in S})^2 \end{aligned}$$

$F(S) = \log \det(K_S)$   
is submodular!

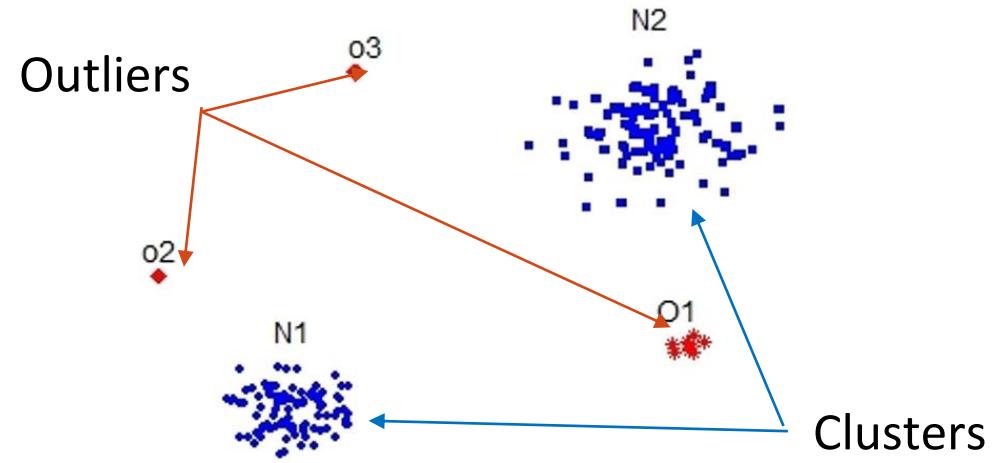
DPP



uniform

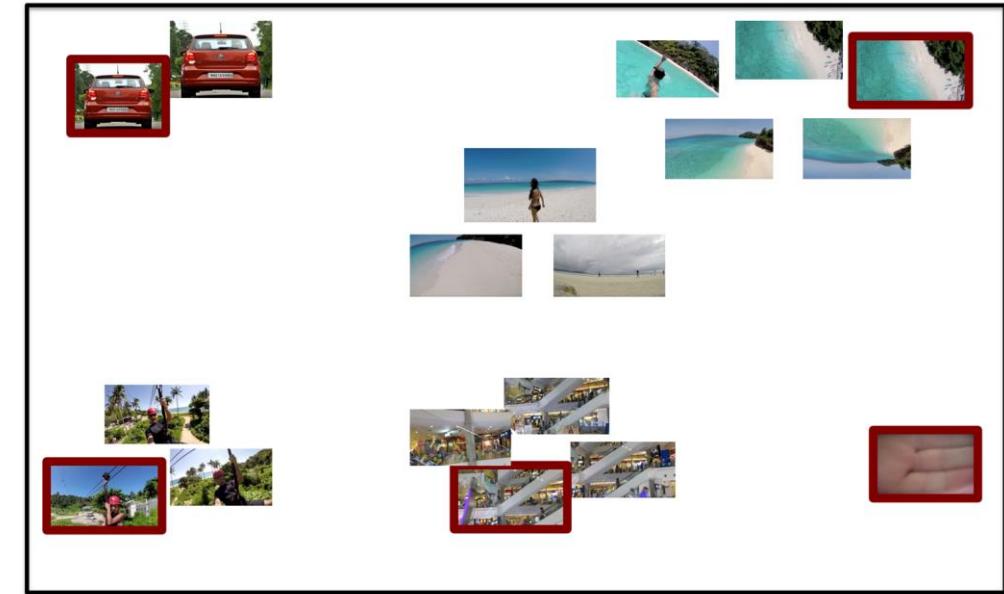


# Diversity Functions: Dispersion



Dispersion Min	$\min_{k,l \in X, k \neq l} d_{kl}$
Dispersion Sum	$\sum_{k,l \in X} d_{kl}$
Dispersion Min-Sum	$\sum_{k \in X} \min_{l \in X} d_{kl}$

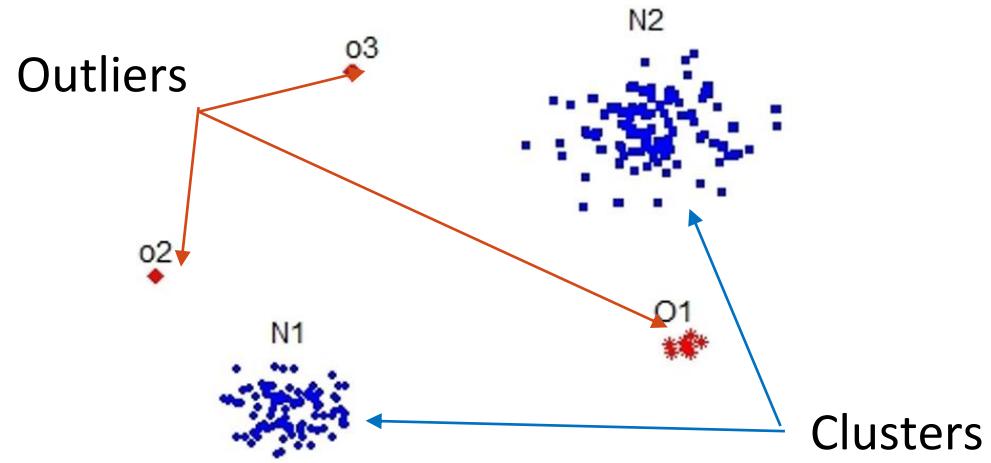
Distance Measure



Diversity Functions  
Picks items as different as possible!

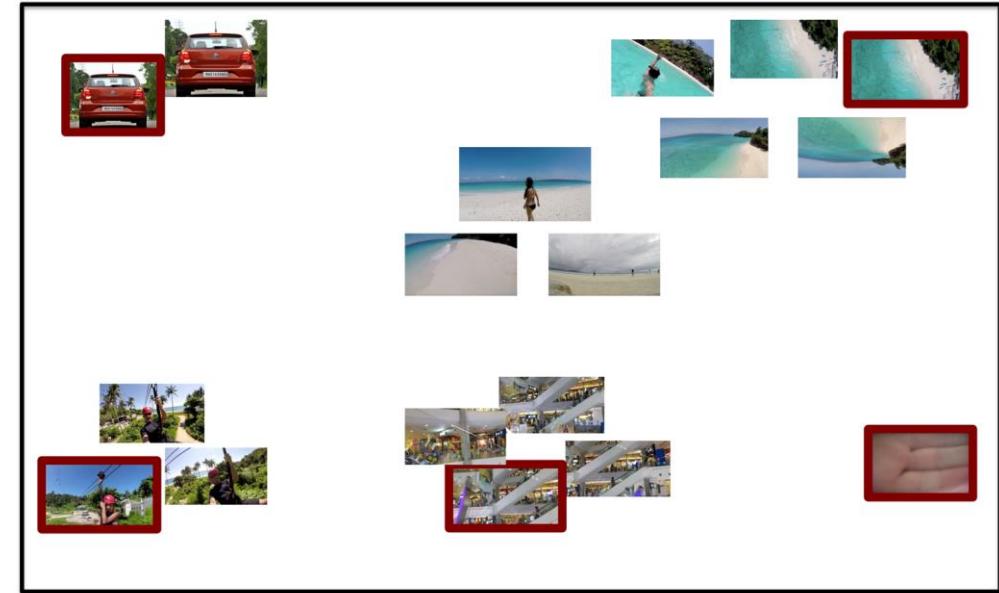
Dasgupta et al 2013, Chakraborty et al 2015

# Diversity Functions: Dispersion



Dispersion Min	$\min_{k,l \in X, k \neq l} d_{kl}$
Dispersion Sum	$\sum_{k,l \in X} d_{kl}$
Dispersion Min-Sum	$\sum_{k \in X} \min_{l \in X} d_{kl}$

Dispersion Sum and Dispersion Min Not Submodular!

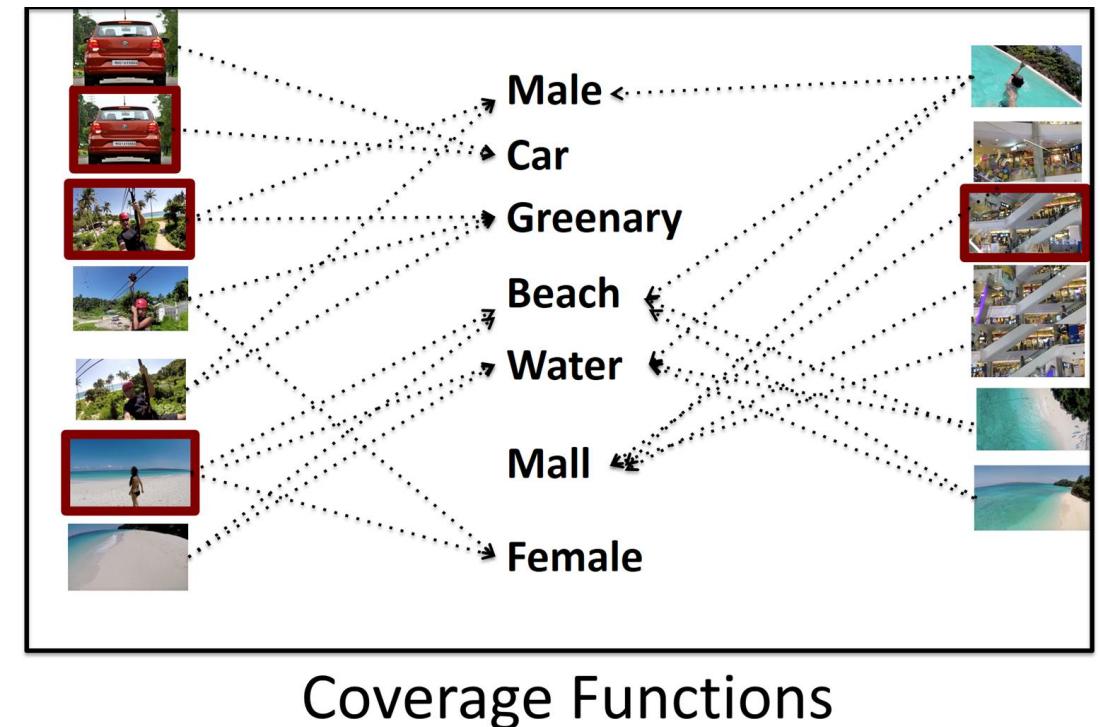
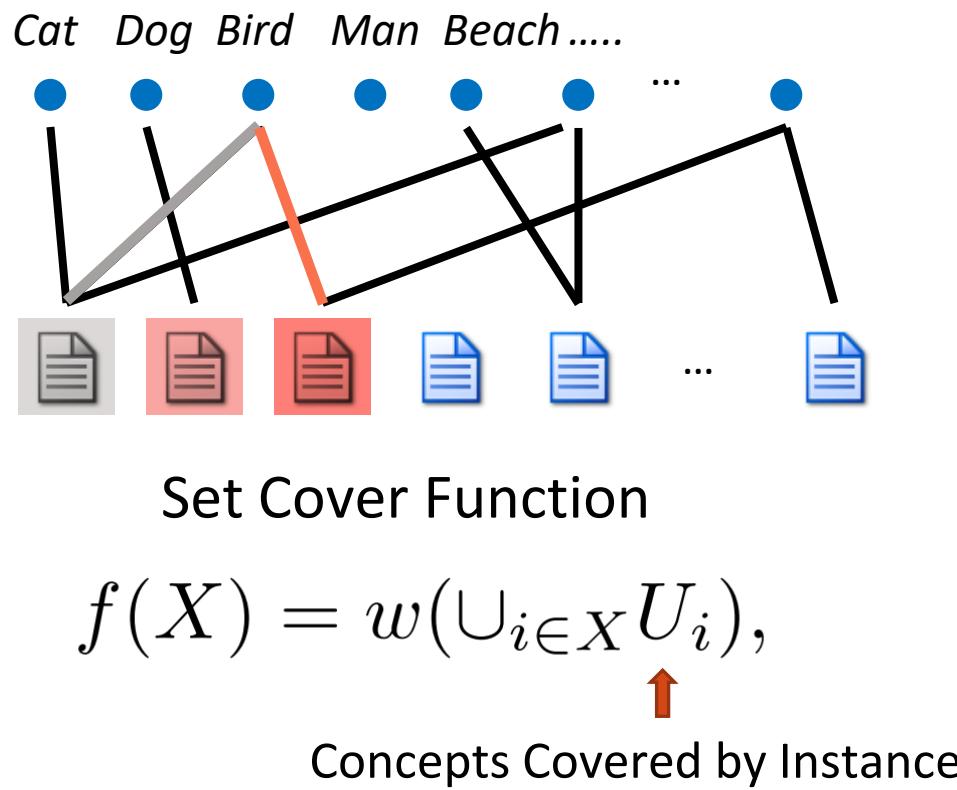


Diversity Functions

Picks items as different as possible!

Dasgupta et al 2013, Chakraborty et al 2015

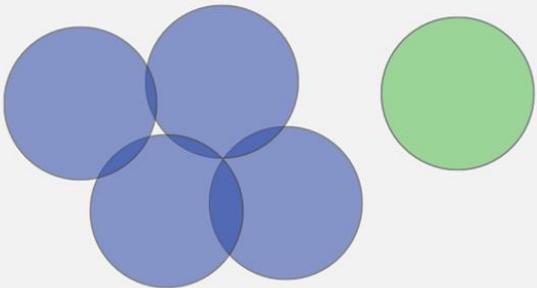
# Coverage Functions



Select instances which “cover” all concepts

Wolsey et al 1982, ...

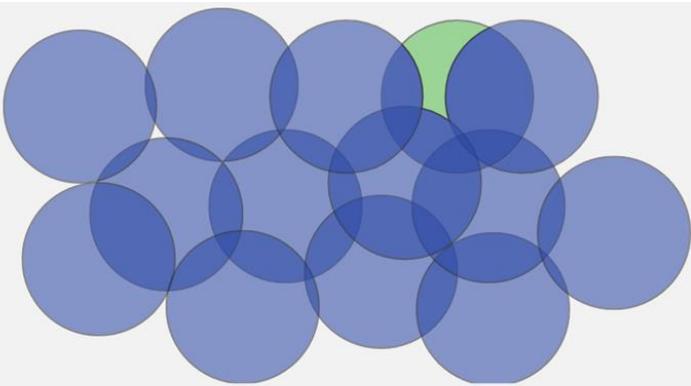
# Why is Set Cover Submodular?



Gain (value) of  $v$  in context of  $A$ :

$$f(A \cup \{v\}) - f(A) = f(\{v\})$$

We get full value  $f(\{v\})$  in this case since the area of  $v$  has no overlap with that of  $A$ .

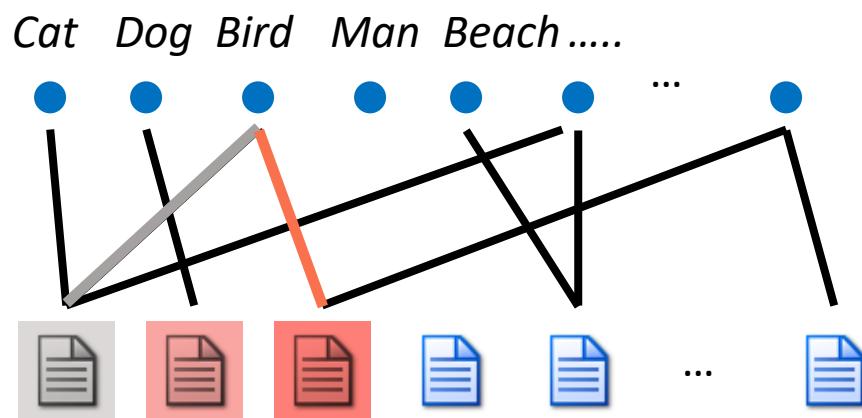


Incremental value of  $v$  in the context of  $B \supset A$ :

$$\begin{aligned} f(B \cup \{v\}) - f(B) &< f(\{v\}) \\ &= f(A \cup \{v\}) - f(A) \end{aligned}$$

So benefit of  $v$  in the context of  $A$  is greater than the benefit of  $v$  in the context of  $B \supseteq A$ .

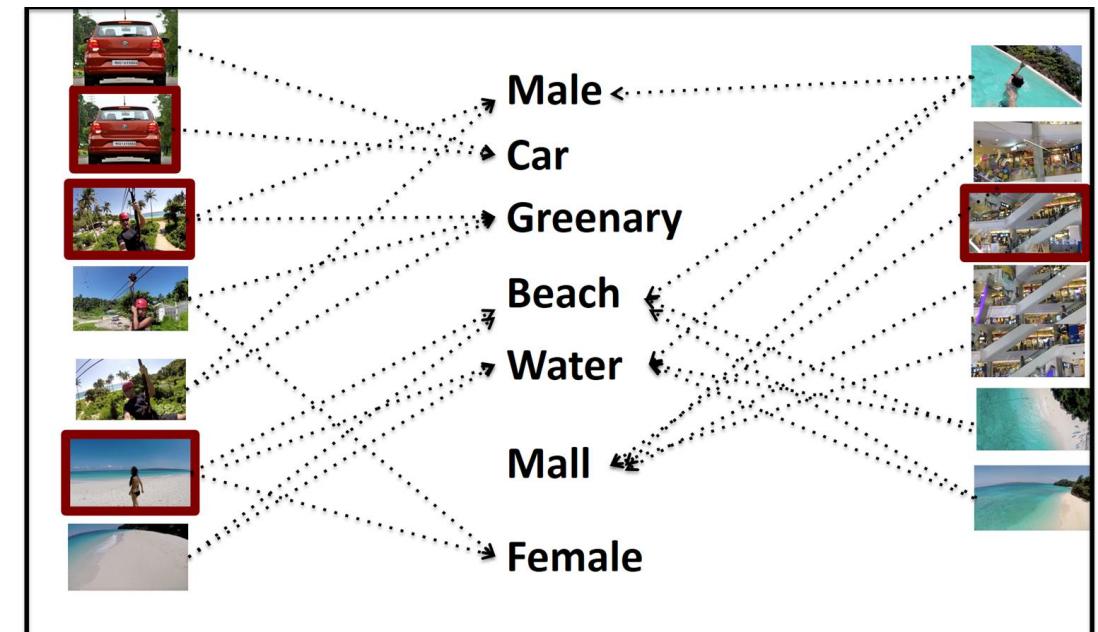
# Coverage Functions



Probabilistic Set Cover Function

$$f(X) = \sum_{i \in \mathcal{U}} w_i [1 - \prod_{j \in X} (1 - p_{ij})].$$

Probability that Image i covers concept j

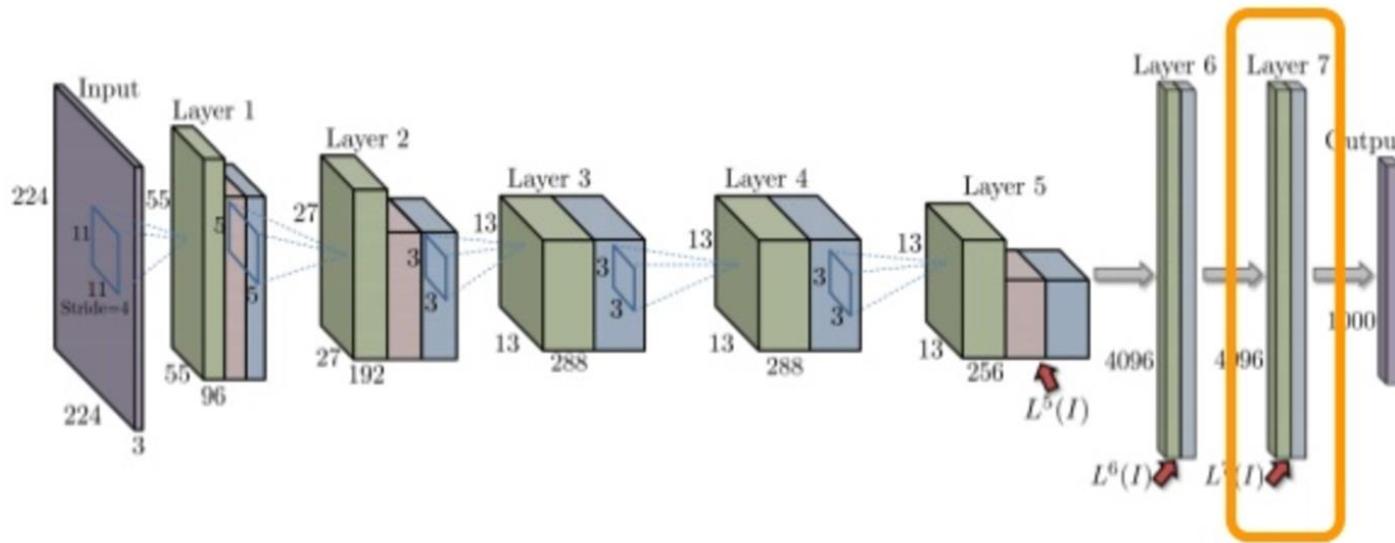


Coverage Functions

Allow for Probability of covering concepts

El-Arini & Guestrin 2013, ...

# Feature Based Functions



Feature Based Functions

$$f_{\text{fea}}(S) = \sum_{u \in \mathcal{U}} g(m_u(S)).$$



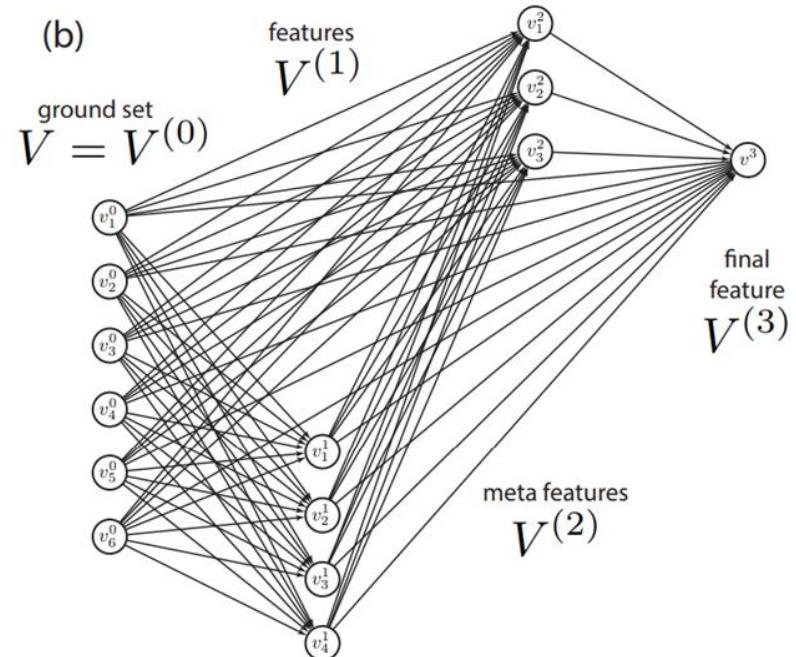
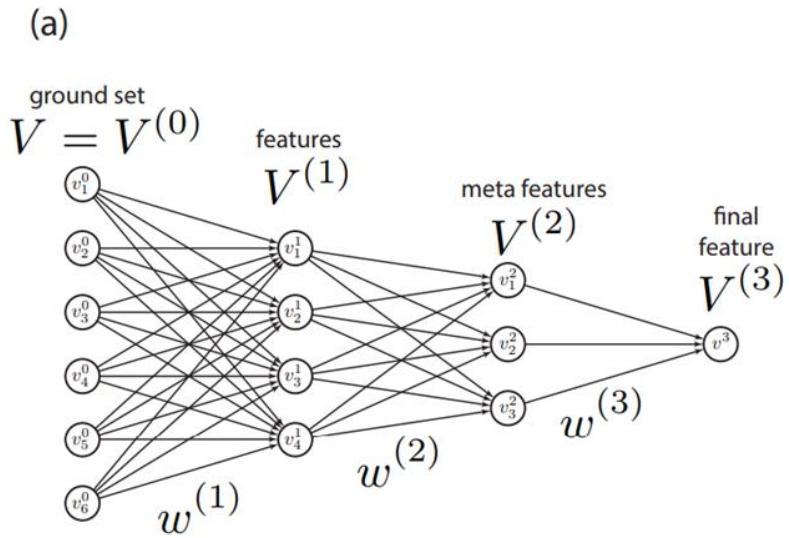
Total Contribution of Feature  $u$  in the Set of Images  $S$

Achieve  
Uniformity in  
Feature  
Coverage

Wei-Iyer et al 2014 ...

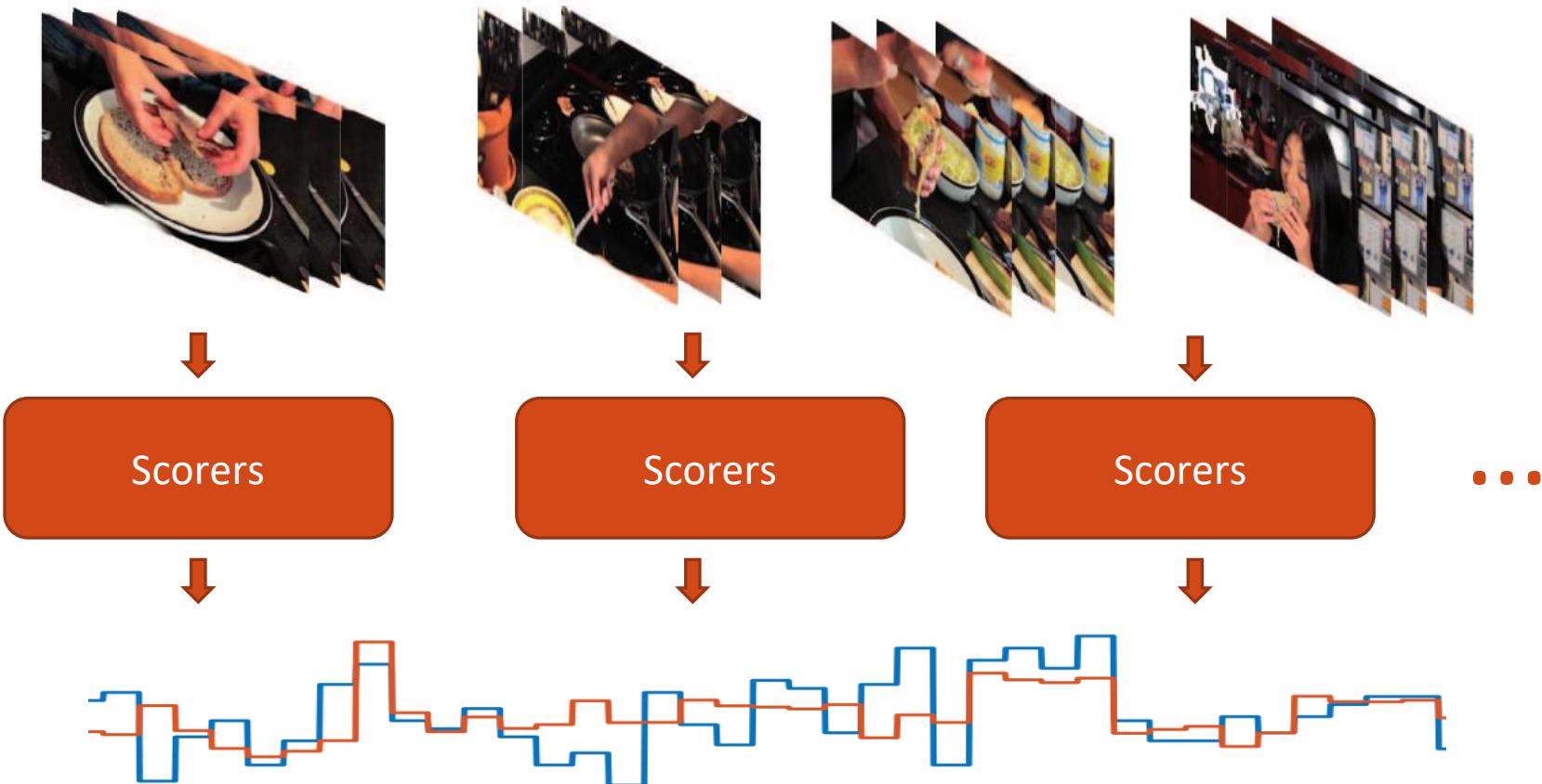
# Nested Feature Based Functions

---



$$\bar{f}(A) = \phi_{v^K} \left( \sum_{v^{K-1} \in V^{(K-1)}} w_{v^K}^{(K)}(v^{K-1}) \phi_{v^{K-1}} \left( \dots \sum_{v^2 \in V^{(2)}} w_{v^3}^{(3)}(v^2) \phi_{v^2} \left( \sum_{v^1 \in V^{(1)}} w_{v^2}^{(2)}(v^1) \phi_{v^1} \left( \sum_{a \in A} w_{v^1}^{(1)}(a) \right) \right) \right) \right)$$

# Importance Functions



# Information Functions

$X_1, \dots, X_n$  discrete random variables:  $X_e \in \{1, \dots, m\}$

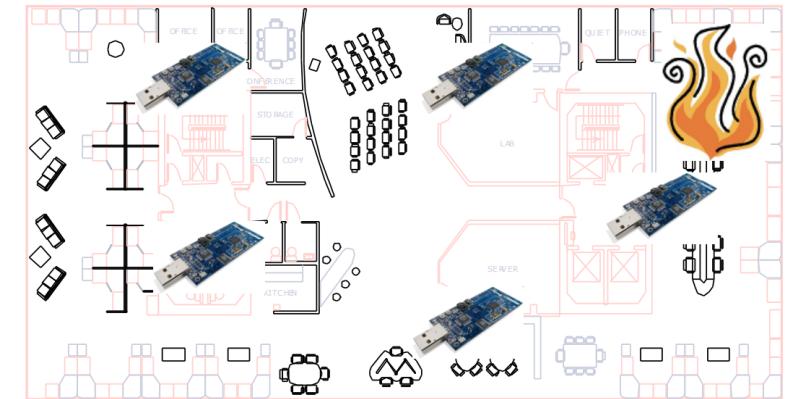
$F(S) = H(X_S)$  = joint entropy of variables indexed by  $S$

$$H(X_e) = \sum_{x \in \{1, \dots, m\}} P(X_e = x) \log P(X_e = x)$$

$$A \subset B, e \notin B \quad F(A \cup e) - F(A) \geq F(B \cup e) - F(B)??$$

$$\begin{aligned} H(X_{A \cup e}) - H(X_A) &= H(X_e | X_A) \\ &\leq H(X_e | X_B) \quad \text{"information never hurts"} \\ &= H(X_{B \cup e}) - H(X_B) \end{aligned}$$

discrete entropy is submodular!



Entropy  
Mutual Information  
Information Gain

...

Krause et al 2008, ...

# Information Gain as a Submodular Function

---

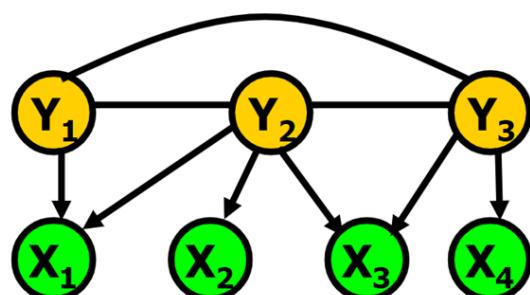
$Y_1, \dots, Y_m, X_1, \dots, X_n$  discrete RVs

$$F(A) = I(Y; X_A) = H(Y) - H(Y | X_A)$$

- $F(A)$  is NOT always submodular

If  $X_i$  are all conditionally independent given  $Y$ ,  
then  $F(A)$  is submodular!

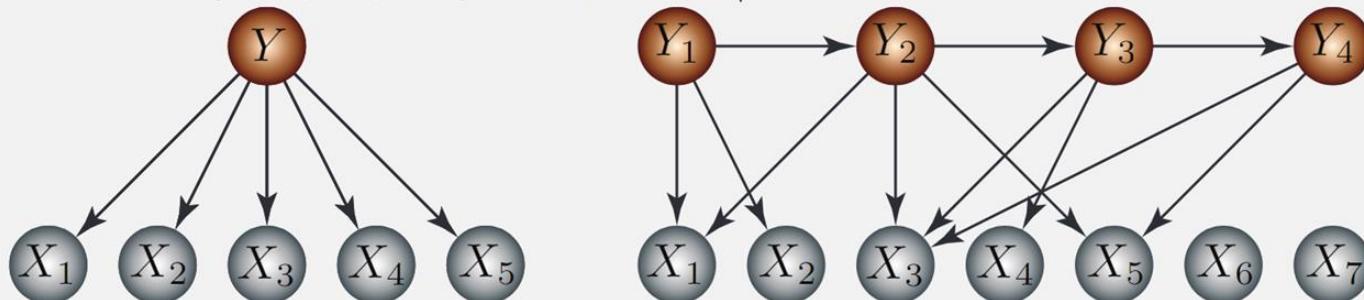
[Krause & Guestrin '05]



Proof:  
“information never hurts”

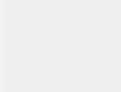
# Information Gain: Feature Selection

- Naïve Bayes property:  $X_A \perp\!\!\!\perp X_B | Y$  for all  $A, B$ .



- When  $X_A \perp\!\!\!\perp X_B | Y$  for all  $A, B$  (the Naïve Bayes assumption holds), then

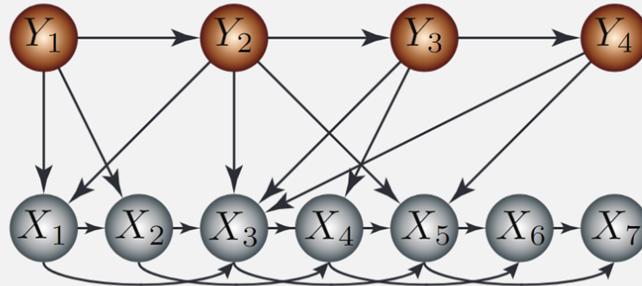
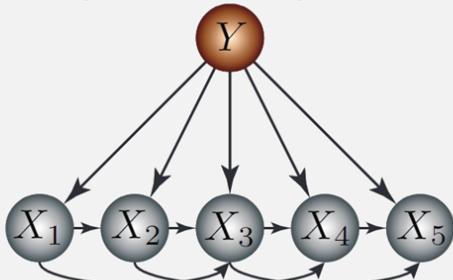
$$f(A) = I(Y; X_A) = H(X_A) - H(X_A|Y) = H(X_A) - \sum_{a \in A} H(X_a|Y)$$



is submodular (submodular minus modular).

# Feature Selection

- Naïve Bayes property fails:



- $f(A)$  naturally expressed as a difference of two submodular functions

$$f(A) = I(Y; X_A) = H(X_A) - H(X_A|Y),$$

which is a DS (difference of submodular) function.

- Alternatively, when Naïve Bayes assumption is false, we can make a submodular approximation (Peng-2005). E.g., functions of the form:

$$f(A) = \sum_{a \in A} I(X_a; Y) - \lambda \sum_{a, a' \in A} I(X_a; X_{a'}|Y)$$

where  $\lambda \geq 0$  is a tradeoff constant.

# Summary of Function Classes

---

## **Monotone Submodular:**

- 1) Facility Location
- 2) Saturated Coverage
- 3) Feature Based Functions
- 4) Deep Submodular Functions
- 5) Set Cover
- 6) Probabilistic Set Cover
- 7) Complexity Function
- 8) Entropy

## **Non Monotone Submodular**

- 1) Log Determinant (DPP)
- 2) Graph Cut ( $\lambda < 2$ )
- 3) Mutual Information (NB)

## **Non Submodular Functions**

- 1) Disparity Min
- 2) Disparity Sum

# Summary of Function Classes

General Set Functions

Supermodular Functions

Modular Functions

Submodular Functions

Monotone  
Submodular  
Functions

Non-  
Monotone  
Submodular  
Functions

Dispersion  
Functions

# Properties of Submodular Functions

---

- ❑ Convex Combinations of Submodular Functions are Submodular
- ❑ Intersections with Fixed Sets is Submodular (Restrictions)
- ❑ Unions with Fixed Sets is Submodular (Conditioning)
- ❑ Complement Functions are Submodular (Reflection)
- ❑ Minimum and Maximum of Submodular Functions
- ❑ Submodularity and Convexity
- ❑ Submodularity and Concavity

# Convex Combinations of Submodular Functions

---

$F_1, \dots, F_m$  submodular functions on  $V$  and  $\lambda_1, \dots, \lambda_m > 0$

Then:  $F(A) = \sum_i \lambda_i F_i(A)$  is submodular

Submodularity closed under nonnegative linear combinations!

Extremely useful fact:

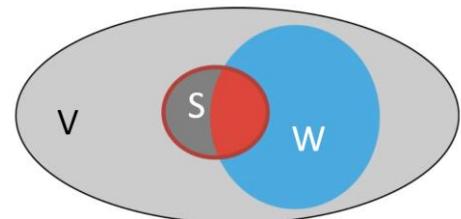
- $F_\theta(A)$  submodular  $\rightarrow \sum_\theta P(\theta) F_\theta(A)$  submodular!
- Multicriterion optimization
- A basic proof technique! ☺

# More Properties

---

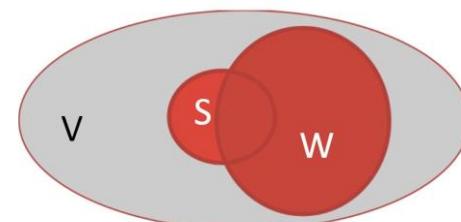
- **Restriction:**  $F(S)$  submodular on  $V$ ,  $W$  subset of  $V$

Then  $F'(S) = F(S \cap W)$  is submodular



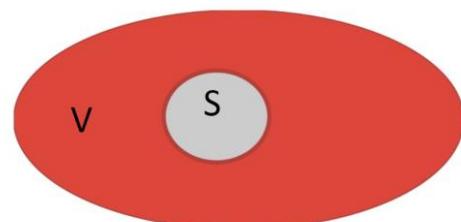
- **Conditioning:**  $F(S)$  submodular on  $V$ ,  $W$  subset of  $V$

Then  $F'(S) = F(S \cup W)$  is submodular



- **Reflection:**  $F(S)$  submodular on  $V$

Then  $F'(S) = F(V \setminus S)$  is submodular



# Concave over Submodular

---

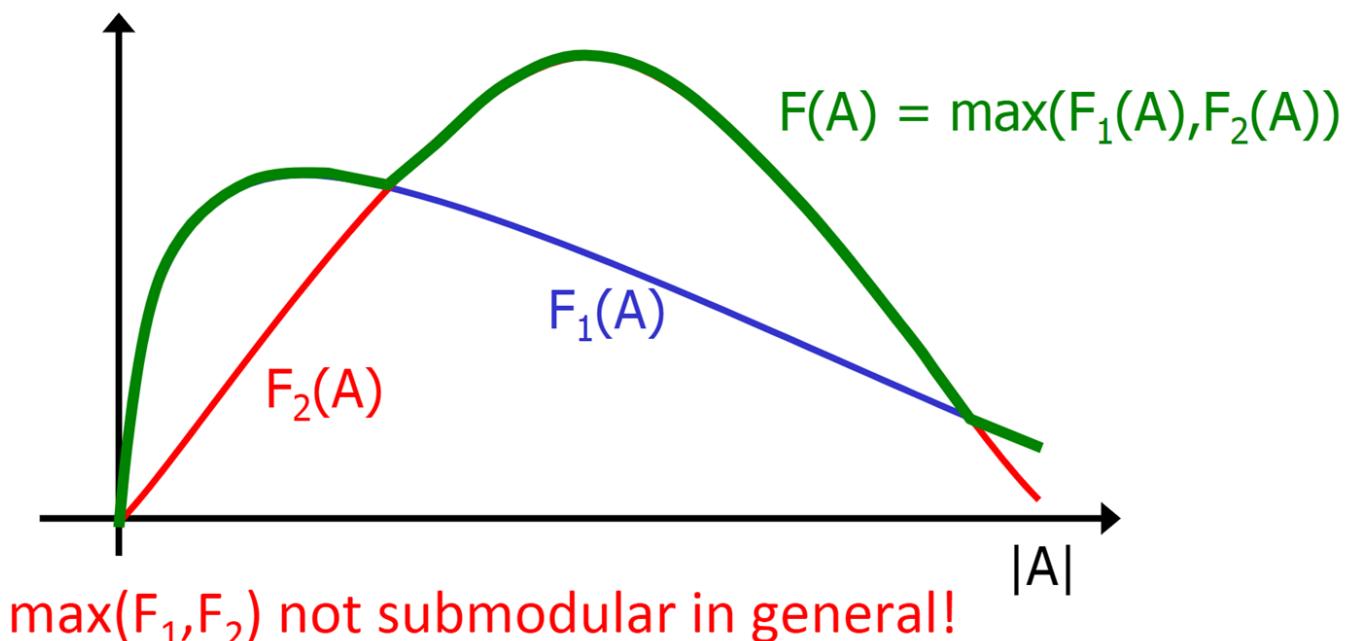
- Concave over Monotone Submodular functions are Submodular:
  - Given a concave function  $g$ , and monotone submodular function  $f$ ,  $g(f(S))$  is submodular!
  - Hence given a positive modular function  $m$ ,  $g(m(S))$  is submodular and hence Feature based functions are submodular
  - This does not hold if  $m$  is positive and negative (i.e.  $f$  is non-monotone)

# Maximum of Submodular Functions

---

- $F_1(A), F_2(A)$  submodular. What about

$$F(A) = \max\{ F_1(A), F_2(A) \} \quad ?$$



# Minimum of Submodular Functions

---

Well, maybe  $F(A) = \min(F_1(A), F_2(A))$  instead?

	$F_1(A)$	$F_2(A)$
{}	0	0
{a}	1	0
{b}	0	1
{a,b}	1	1

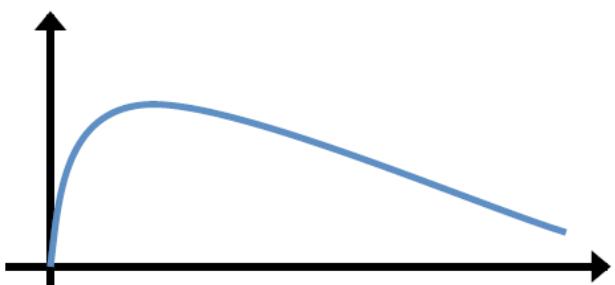
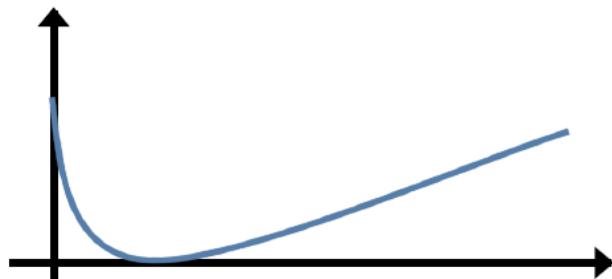
$$\begin{aligned}F(\{b\}) - F(\{\}) &= 0 \\ &< \\ F(\{a,b\}) - F(\{a\}) &= 1\end{aligned}$$

$\min(F_1, F_2)$  not submodular in general!

# Is Submodularity like Convexity or Concavity?

---

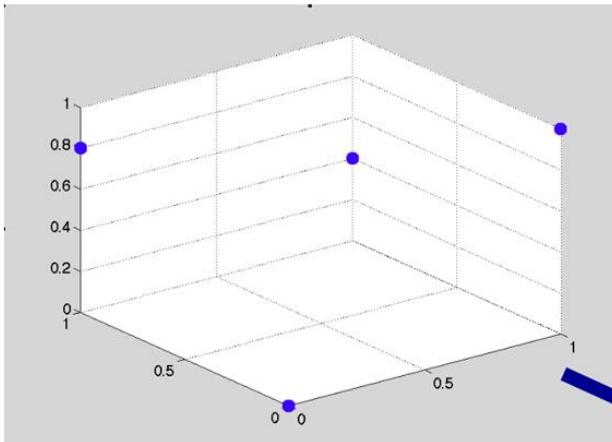
discrete convexity ....



... or concavity?

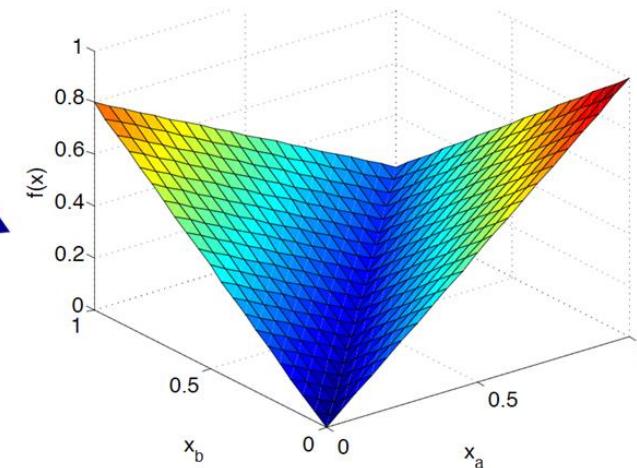
# Convex Aspects of Submodular Functions

---



- convex extension
- duality
- efficient minimization

But this is only  
half of the story...



# Concave Aspects of Submodular Functions

---

- submodularity:

$A \subseteq B, s \notin B :$

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

A

+ $\bullet$  s

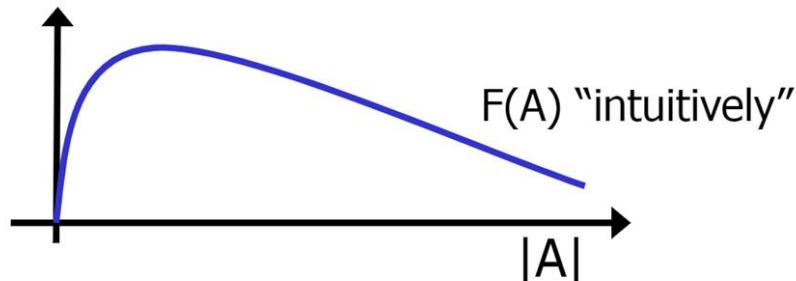
B

+ $\bullet$  s

- concavity:

$a \leq b, s > 0 :$

$$f(a + s) - f(a) \geq f(b + s) - f(b)$$

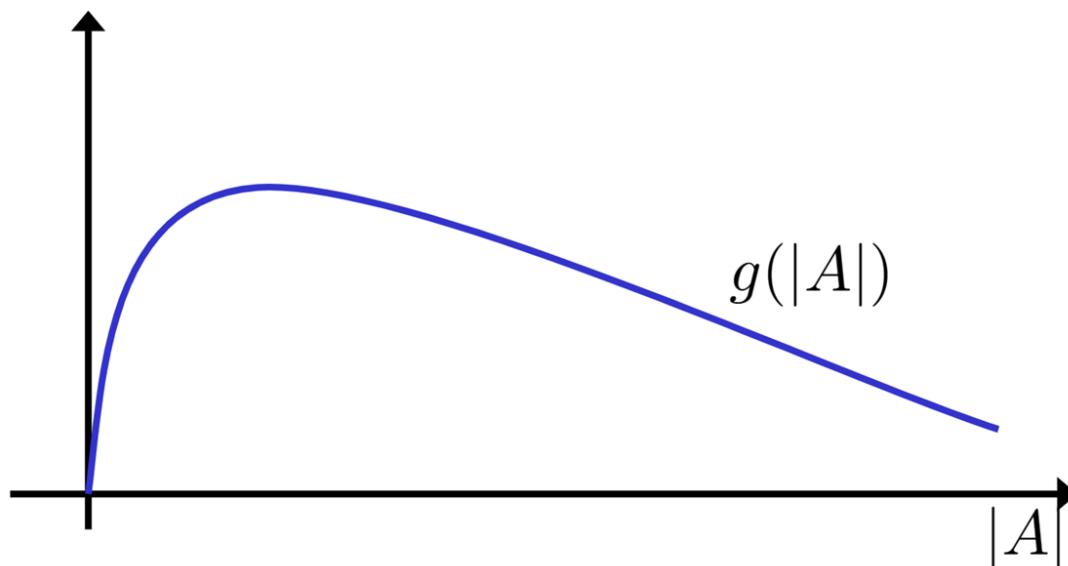


# Concave Aspects of Submodular Functions

---

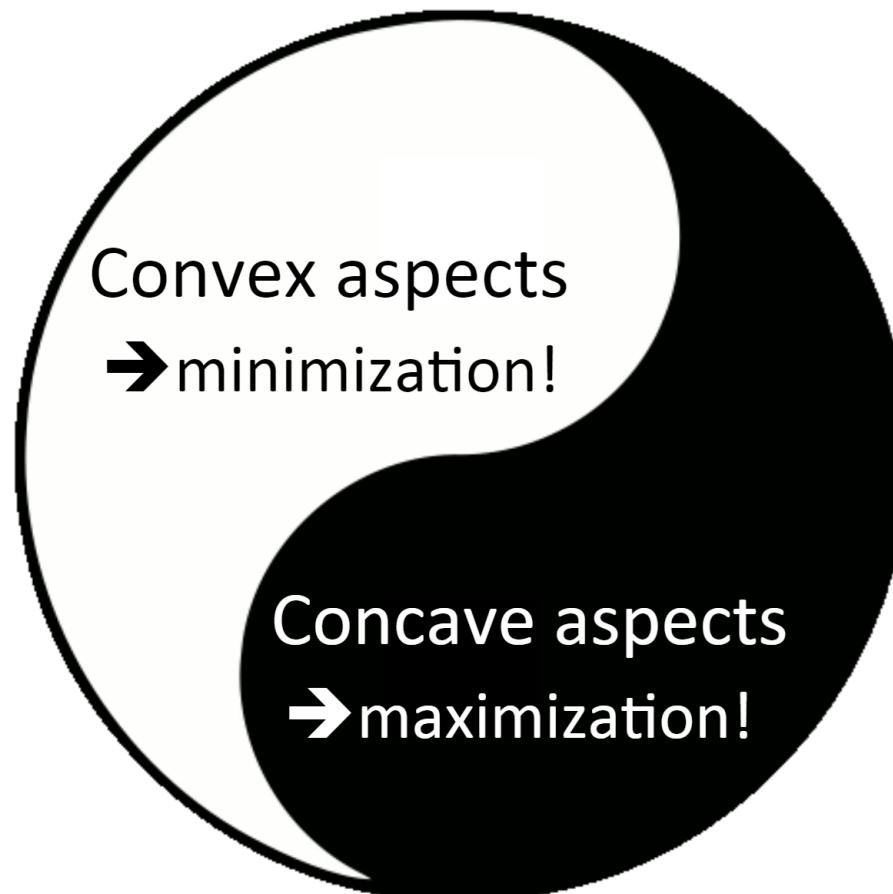
- suppose  $g : \mathbb{N} \rightarrow \mathbb{R}$  and  $F(A) = g(|A|)$

$F(A)$  submodular if and only if ...  $g$  is concave



# Two Faces of Submodularity

---



# Outline

---

- ❑ Discrete Optimization in Machine Learning
- ❑ Lecture 7.1: Submodular Functions: Properties and Examples
  - ❑ Definition and Intuition of Submodularity
  - ❑ Modeling Power of Submodular/Set Functions
  - ❑ Examples of Submodular Functions
  - ❑ Properties of Submodular Functions
- ❑ **Lecture 7.2: Submodular Information Measures (SIMs)**
  - ❑ Definitions of SIMs: Conditional Gain, Submodular Mutual Information, Submodular Conditional Mutual Information
  - ❑ Properties of SIMs
  - ❑ Examples of SIMs
  - ❑ Applications of SIMs

# Submodular Information Functions

---



Subset



$\subset$



Superset

# Submodular Information Functions



Subset



$\subset$



Superset

Information gain reduces with larger sets!

# Submodular Information Functions

---

$X_1, \dots, X_n$  discrete random variables:  $X_e \in \{1, \dots, m\}$

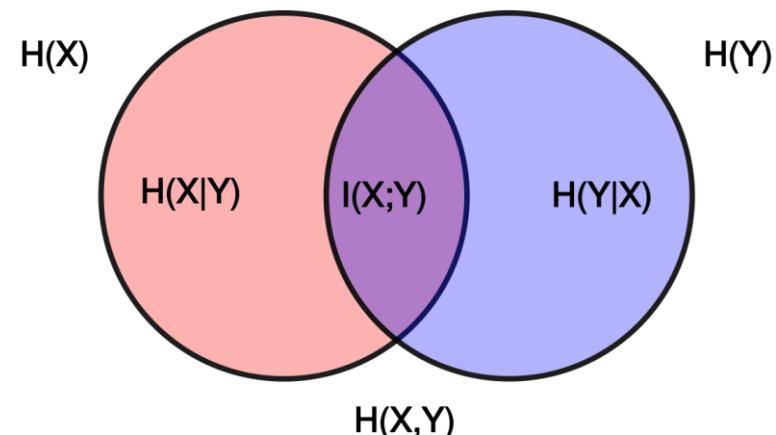
$F(S) = H(X_S)$  joint entropy of variables indexed by  $S$

$$H(X_e) = \sum_{x \in \{1, \dots, m\}} P(X_e = x) \log P(X_e = x)$$

$$A \subset B, e \notin B \quad F(A \cup e) - F(A) \geq F(B \cup e) - F(B) ??$$

$$\begin{aligned} H(X_{A \cup e}) - H(X_A) &= H(X_e | X_A) \\ &\leq H(X_e | X_B) \quad \text{"information never hurts"} \\ &= H(X_{B \cup e}) - H(X_B) \end{aligned}$$

discrete entropy is submodular!



# Recall: Information Theoretic Concepts

---

- **Entropy:** Given a set of random variables,  $X_1, \dots, X_n$ , the Entropy of a **subset** of random variables:  $H(X_A) = -\sum_{X_A} P(X_A) \log P(X_A)$
- **Mutual Information:** Given a set of random variables,  $X_1, \dots, X_n$ , and sets  $A, B \subseteq V$ , the Mutual Information  $I(X_A; X_B) = H(X_A) + H(X_B) - H(X_{A \cup B})$
- **Conditional Entropy:** Given a set of random variables,  $X_1, \dots, X_n$ , and sets  $A, B \subseteq V$ , the Conditional Entropy  $H(X_A|X_B) = H(X_{A \cup B}) - H(X_B)$
- **Conditional Mutual Information:** Given a set of random variables,  $X_1, \dots, X_n$ , and sets  $A, B, C \subseteq V$ , the Conditional Mutual Information  $I(X_A; X_B|X_C) = H(X_A|X_C) + H(X_B|X_C) - H(X_{A \cup B}|X_C)$

# Shannon Inequalities and Entropy

---

The Entropy Function  $F$  over sets  $A, B \subseteq V$  or Random variables satisfies the following three properties:

- ① Normalized:  $F(\emptyset) = 0$
- ② Monotone:  $F(A) \geq F(B)$  if  $A \supseteq B$
- ③ Two Alternative (Submodularity):  
$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$$

Observations:

- Property 3 implies non-negativity of mutual information and conditional mutual information!
- Properties 1 and 2 imply non-negativity of entropy and conditional entropy!

# Shannon Inequalities & Poly-Matroids

---

**Definition:** A Poly-Matroid Function  $f : 2^V \Rightarrow \mathbb{R}$  is a set function which satisfies:

- ① Normalized:  $F(\emptyset) = 0$
- ② Monotone:  $F(A) \geq F(B)$  if  $A \supseteq B$
- ③ Two Alternative (Submodularity):  
$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$$

Observations:

- Poly-Matroid Functions generalize Entropy!

# Submodular Information Measures

- **Submodular Information Functions:** Given a set of data-points  $V = \{1, \dots, n\}$ , the Information of a **set** of points is  $F(A)$  where  $F$  is a polymatroid function.
- **Submodular Mutual Information:** Given a set of data-points  $V = \{1, \dots, n\}$ , and sets  $A, B \subseteq V$ , the Submodular Mutual Information  $I_F(A; B) = F(A) + F(B) - F(A \cup B)$
- **Conditional Gain:** Given a set of data-points  $V = \{1, \dots, n\}$ , and sets  $A, B \subseteq V$ , the Conditional Gain  $F(A|B) = F(A \cup B) - F(B)$
- **Conditional Submodular Mutual Information:** Given a set of data points,  $1, \dots, n$ , and sets  $A, B, C \subseteq V$ , the Conditional Submodular Mutual Information  
$$I_F(A; B|C) = F(A|C) + F(B|C) - F(A \cup B|C)$$

# Submodular Information Measures

---

## Intuition of the Information Theoretic Quantities

1. **Submodular Information:** Clear from name
2. **Submodular Conditional Gain:** Gain in Information in Adding set A to set B.
3. **Submodular Mutual Information:** Joint Information between sets A and B based on F
4. **Independence:** Maximally different sets A and B in terms of F

# Submodular Information $\supset$ Entropic Information

- Polymatroid Functions strictly generalize Entropy (Zhang & Yeung 1997-98).
- There exist certain inequalities which Entropic measures satisfy but they cannot be shown by Shannon Inequalities!
- One such Inequality which holds for all convex combinations of Entropic functions: Given four random variables  $X_1, X_2, X_3, X_4$ :

$$I(X_3; X_4) - I(X_3; X_4|X_1) - I(X_3; X_1|X_2) \leq \frac{1}{2}I(X_1; X_2) + \frac{1}{4}I(X_1; X_3, X_4) + \frac{1}{4}I(X_2; X_3, X_4)$$

- In general, submodular information measures will not satisfy this.
- In fact, there exist Set Cover, Facility Location and Matroid Rank Functions which do not satisfy the inequality above.

# Independence between Sets

- Given two sets  $A, B$ , we say that  $A \perp_f B$  iff

$$I_F(A; B) = 0$$

- Given sets  $A, B, C$ ,  $A \perp_f B|C$  iff

$$I_F(A; B|C) = 0$$

- The above notions of independence (joint independence) imply pairwise independence  $A \perp B$  iff  $a \perp_f b, \forall a \in A, b \in B$ .
- The above generalize (conditional) independence between sets of random variables.

# Multi-Set Submodular Information

- Given sets  $A_1, \dots, A_k \subseteq V$ , we can define two kinds of Mutual Information measures:
  - Submodular Total Correlation:

$$C_F(A_1; \dots; A_k) = \sum_{i=1}^k F(A_i) - F(\cup_i A_i)$$

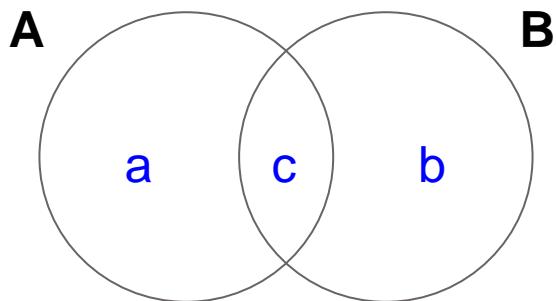
- Submodular Multi-Set Mutual Information:

$$I_f(A_1; A_2; \dots; A_k) \triangleq - \sum_{T \subseteq [k]} (-1)^{|T|} f(\cup_{i \in T} A_i)$$

- Conditional variants can similarly be defined!

# Submodular Information for k = 2 and 3

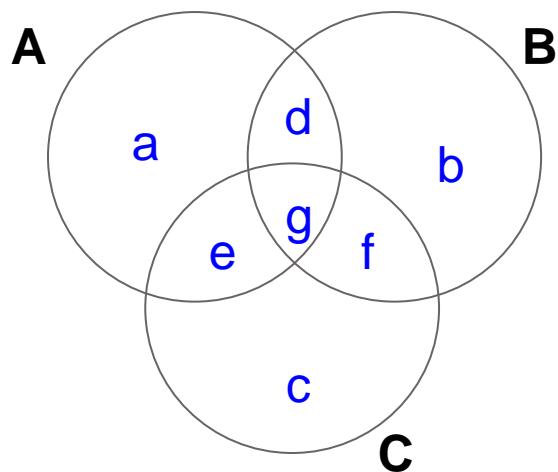
---



a:  $F(A|B)$

b:  $F(B|A)$

c:  $I_F(A; B)$



a:  $I_F(A|B \cup C)$

b:  $I_F(B|A \cup C)$

c:  $I_F(C|A \cup B)$

d:  $I_F(A; B|C)$

e:  $I_F(A; C|B)$

f:  $I_F(B; C|A)$

g:  $I_F(A; B; C)$

# Modular Functions

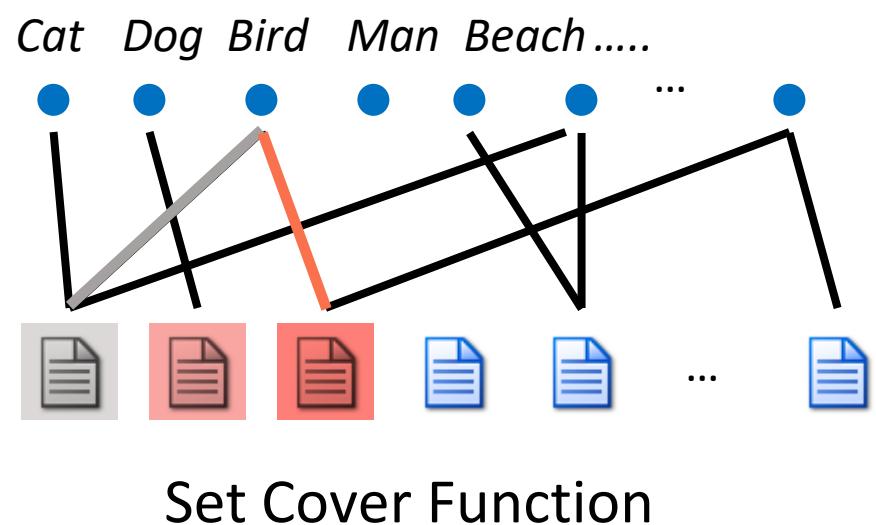
---



- $F(A) = w(A)$
- $F(A|B) = w(A \setminus B)$
- $I_F(A; B) = w(A \cap B)$
- $I_F(A_1; \dots; A_k) = w(\cap_{i=1}^k A_i)$
- Independence:  $A \cap B = \emptyset$

# Weighted Set Cover

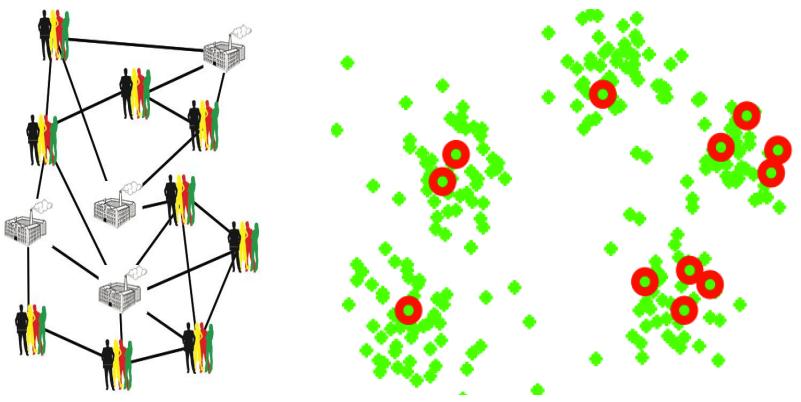
---



- Define  $\gamma(A) = \cup_{i \in A} U_i$ .
- $F(A) = w(\gamma(A))$
- $F(A|B) = w(\gamma(A) \setminus \gamma(B))$
- $I_F(A; B) = w(\gamma(A) \cap \gamma(B))$
- $I_F(A_1; \dots; A_k) = w(\cap_{i=1}^k \gamma(A_i))$
- Independence:  $\gamma(A) \cap \gamma(B) = \emptyset$

# Facility Location

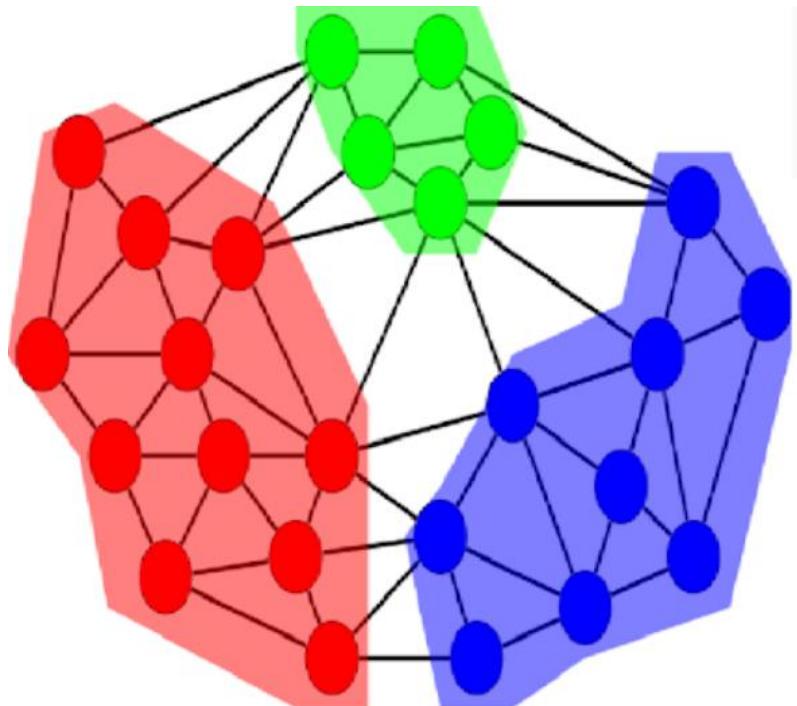
---



- $F(A) = \sum_{i \in V} \max_{a \in A} s_{ia}$
- $F(A|B) = \sum_{i \in V} \max(0, \max_{a \in A} s_{ia} - \max_{b \in B} s_{ib})$
- $I_F(A; B) = \sum_{i \in V} \min(\max_{a \in A} s_{ia}, \max_{b \in B} s_{ib})$
- $I_F(A_1; \dots; A_k) = \sum_{i \in V} \min(\max_{a_1 \in A_1} s_{ia_1}, \dots, \max_{a_k \in A_k} s_{ia_k})$

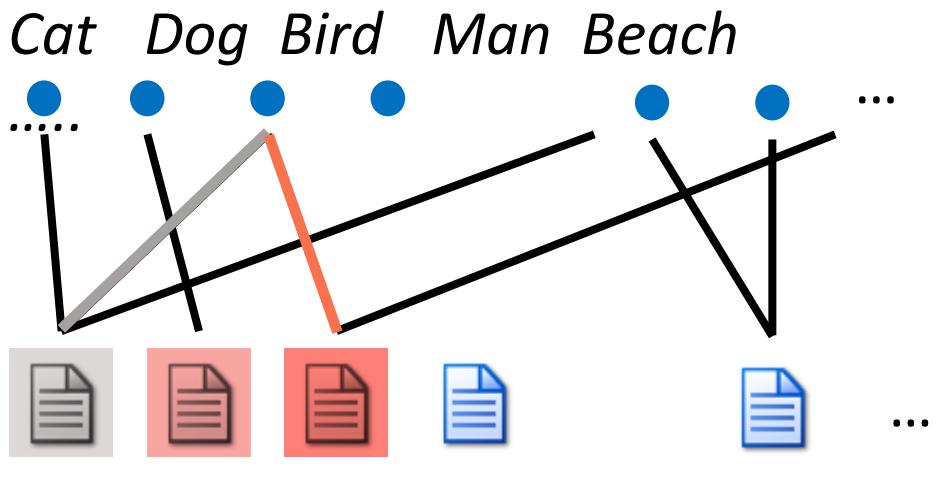
# Penalized Sum Coverage

---



- $F(A) = \sum_{i \in V} \sum_{j \in A} s_{ij} - \frac{1}{2} \sum_{i,j \in A} s_{ij}$
- $I_F(A; B) = \sum_{i \in A, j \in B} s_{ij}$
- $I_F(A; V \setminus A)$  is exactly the graph cut function!
- Independence:  $A \perp_f B$  iff  $s_{ij} = 0$  for  $i \in A, j \in B$ .

# Probabilistic Coverage



# Probabilistic Set Cover

$$f(X) = \sum_{i \in \mathcal{U}} w_i [1 - \prod_{j \in X} (1 - p_{ij})].$$


# Probability that Image i covers concept j

- Define  $P_i(A) = \prod_{a \in A} (1 - p_{ia})$ .
  - $F(A) = \sum_{i \in U} w_i (1 - \prod_{a \in A} (1 - p_{ia}))$
  - $F(A|B) = \sum_{i \in U} w_i P_i(B) (1 - P_i(A \setminus B))$
  - $I_F(A; B) = \sum_{i \in U} w_i (1 - P_i(A))(1 - P_i(B))$
  - $I_F(A_1; \dots; A_k) = \sum_{i \in U} w_i \prod_j j = 1^k (1 - P_i(A_j))$

# Non-Negativity, Monotonicity and Upper/Lower Bounds

- If  $F$  is a polymatroid function...
- Non-Negativity:
  - $F(A)$ ,  $F(A|B)$ ,  $I_F(A; B)$  and  $I_F(A; B|C)$  are all non-negative
- Monotonicity:
  - $F(A)$  and  $F(A|B)$  are monotone in  $A$  for a fixed  $B$
  - $I_F(A; B)$  is monotone in  $A$  for a fixed  $B$  and vice-versa.
  - $I_F(A; B|C)$  is monotone in  $A$  for fixed  $B, C$  and monotone in  $B$  for fixed  $A, C$ .
- Upper/Lower Bounds:
  - $0 \leq F(A|B) \leq F(A)$
  - $0 \leq F(A \cap B) \leq I_F(A; B) \leq \min(F(A), F(B))$
  - $0 \leq F(A \cap B|C) \leq I_F(A; B|C) \leq \min(F(A|C), F(B|C))$

# Submodularity

A Slight Detour...

- Monotonicity = First order partial derivatives non-negative:  
 $F(i|A) = F(A \cup i) - F(A) \geq 0$
- Submodularity = Second order partial derivatives non-positive:  
 $F(i; j|A) = F(i|A \cup j) - F(i|A) \leq 0$
- A **subclass** of submodular functions can be characterized by additionally requiring third order partial derivatives to be non-negative:  $F(i; j; k|A) = F(i; j|A \cup k) - F(i; j|A) \geq 0$
- Many subclasses of functions like Set Cover, Facility Location and Concave over Modular with Power/Log functions satisfy this!
- However, even the simple uniform matroid rank function:  
 $f(A) = \min(|A|, k)$  does not satisfy this..

# Submodularity Cont..

- Submodularity of  $F(A)$  and  $F(A|B)$ :
  - $F(A)$  and  $F(A|B)$  are both submodular in  $A$  for a given  $B$  by definition.
  - $F(A|B)$  is however, neither submodular nor supermodular in  $B$  for a given  $A$ .
- Submodularity of  $I_F(A; B)$  and  $I_F(A; B|C)$ 
  - If  $F$  is such that **its third order partial derivatives are non-negative**, then  $I_F(A; B)$  and  $I_F(A; B|C)$  are both submodular in  $A$  for a fixed  $B$  and  $C$ .
  - In general, however, they are a difference of submodular functions.