Summarizing Large Data Sets

Jeffrey A. Bilmes

Professor
Departments of Electrical Engineering
& Computer Science and Engineering
University of Washington, Seattle
http://melodi.ee.washington.edu/~bilmes

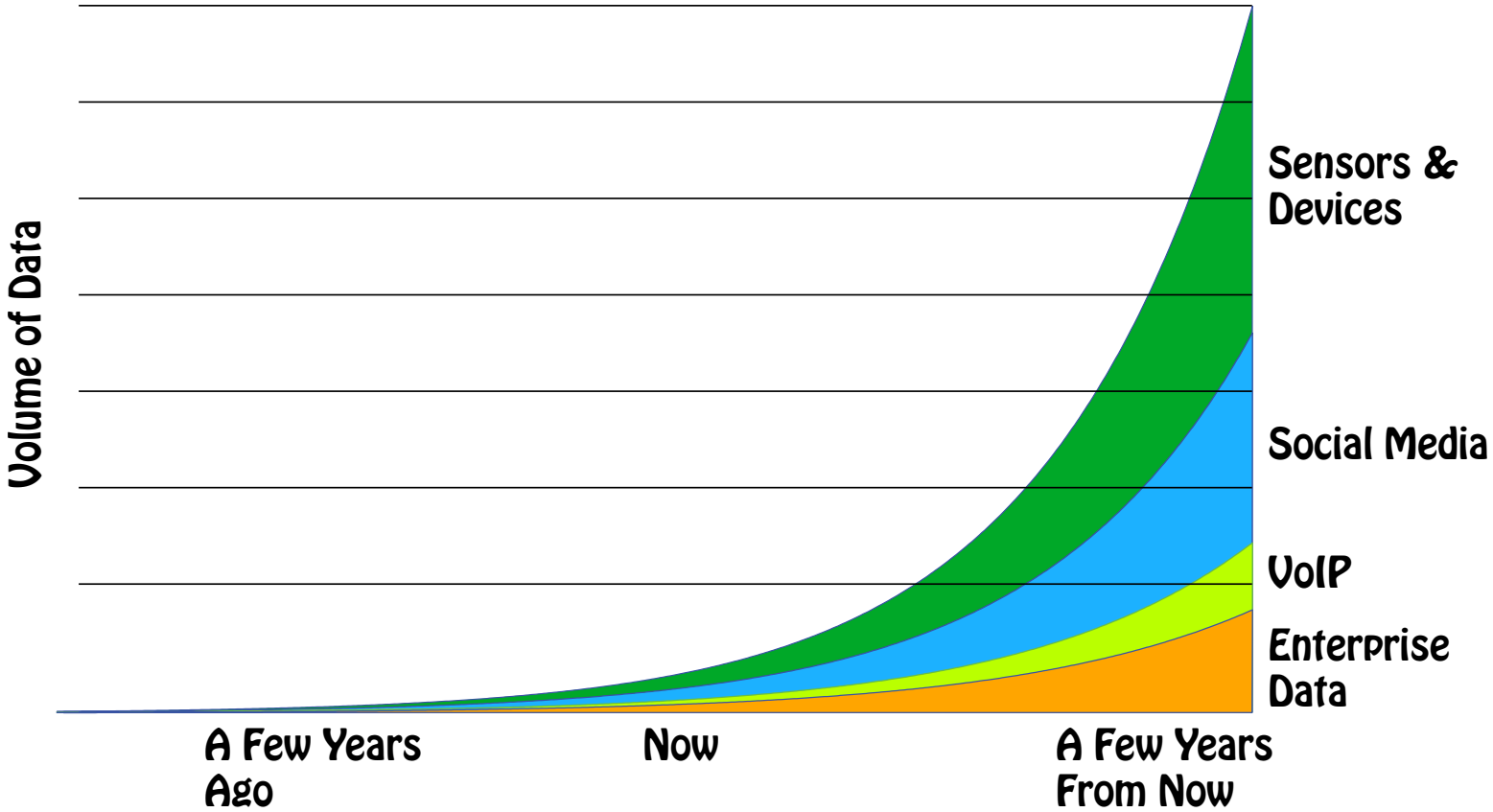
Friday, March 27th, 2015

Outline

Outline

Modern Times

- Big Data is Really Big and Getting Even Bigger.



- Every day, we create 2.5 quintillion bytes (2.5 billion gigabytes) of data (source: IBM).
- 90% of the world's data has been created in the last two years.

Big Data in Machine Learning

Statistics and Machine Learning

- "There's no data like more data", more samples reduces sampling error, higher statistical significance, and better p values.

Big Data in Machine Learning

Statistics and Machine Learning

- "There's no data like more data", more samples reduces sampling error, higher statistical significance, and better p values.

Computational Consequences:

- More expensive, computational resource demands (compute and storage), distributed implementations, more complicated

Big Data in Machine Learning

Statistics and Machine Learning

- "There's no data like more data", more samples reduces sampling error, higher statistical significance, and better p values.

Computational Consequences:

- More expensive, computational resource demands (compute and storage), distributed implementations, more complicated
- Research opportunities to address new computational challenges

1. systems programming, parallel and distributed computing, network topologies, efficient databases.
2. Examples: map reduce, Hadoop, GraphLab, HaLoop, Greenplum, Asterix, Spark, SystemML, MLBase, Myria, etc.

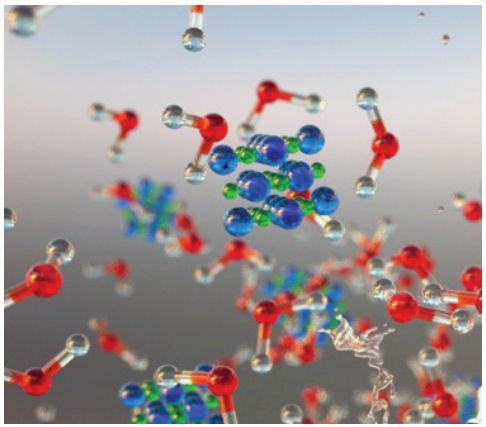
Bigger is Different

H2O molecules

Bigger is Different

H2O molecules

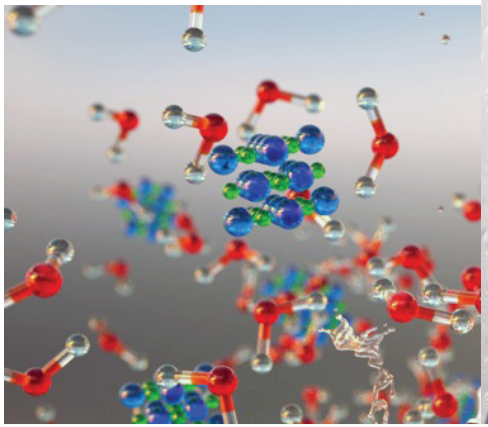
- small (n-body)



Bigger is Different

H2O molecules

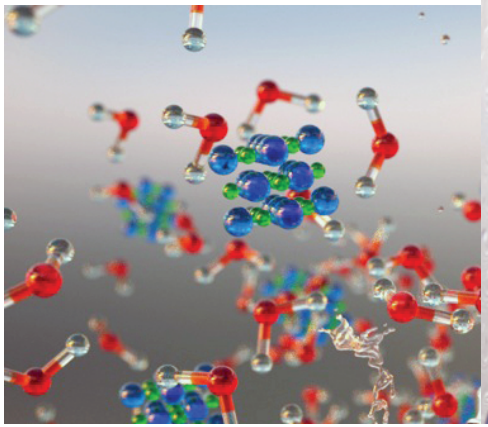
- small (n-body)
- medium (fluid dynamics, viscosity, compresibility),

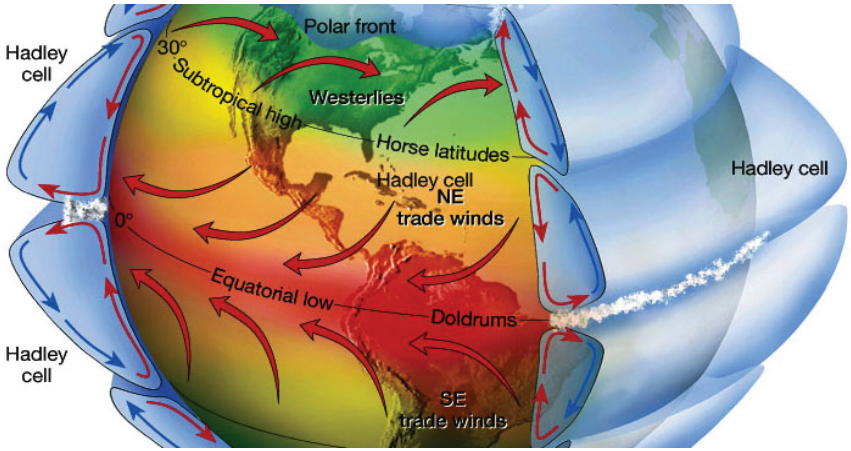


Bigger is Different

H2O molecules

- small (n-body)
- medium (fluid dynamics, viscosity, compresibility),
- large (global weather systems, meteorology).





Same underlying molecular collision events!

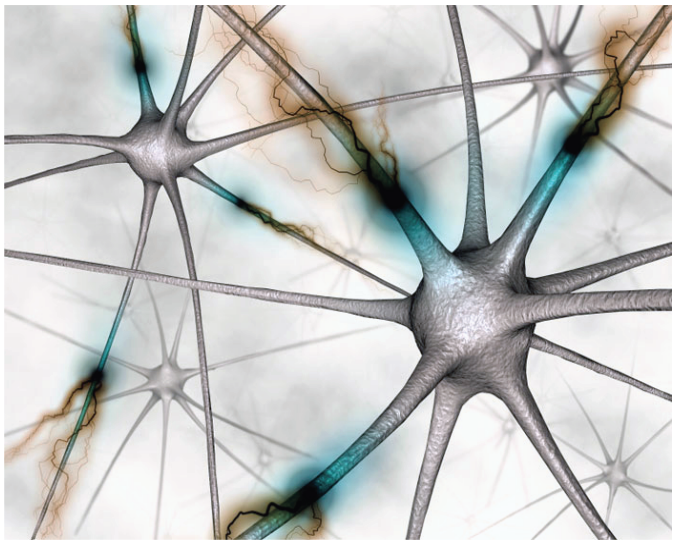
Bigger is Different

Neurons

Bigger is Different

Neurons

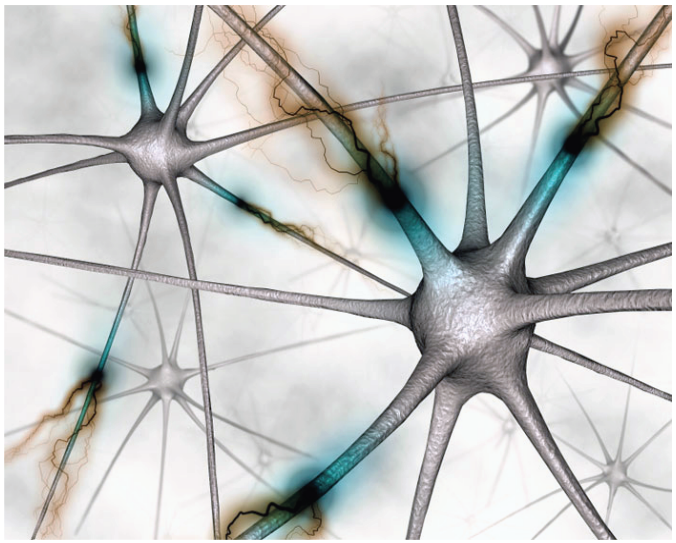
- small (neural spike trains, population coding)



Bigger is Different

Neurons

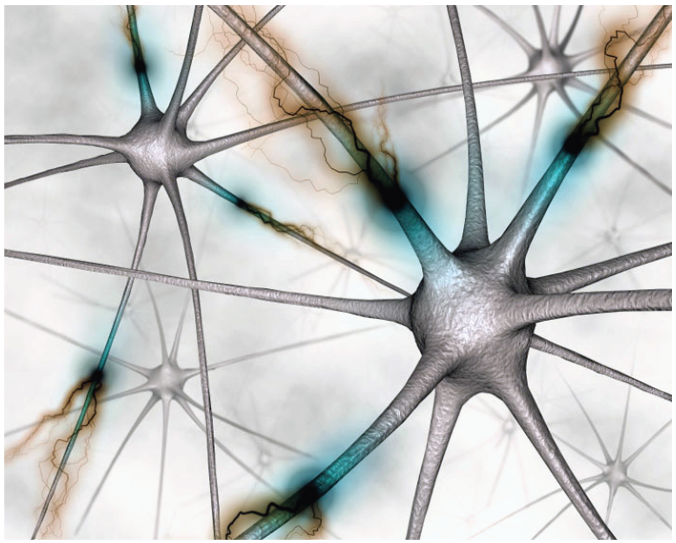
- small (neural spike trains, population coding)
- medium (intelligence, consciousness, psychology)



Bigger is Different

Neurons

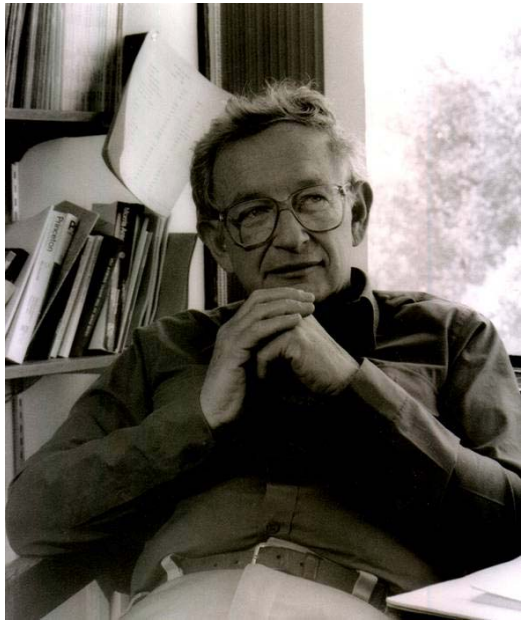
- small (neural spike trains, population coding)
- medium (intelligence, consciousness, psychology)
- large (society, social choice, wisdom of the crowd)



Same underlying electrical and chemical impulses.

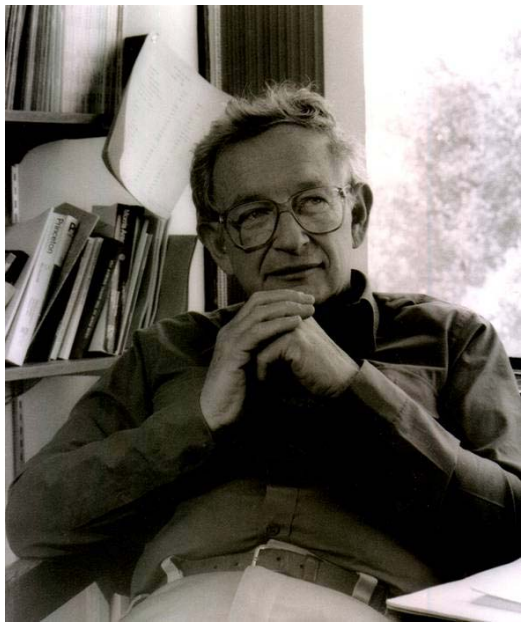
Bigger is Different

- "More is Different", P.W. Anderson, 1972 (Nobel laureate). "The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe."



Bigger is Different

- "More is Different", P.W. Anderson, 1972 (Nobel laureate). "The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe."
- "...alterations of being ...are not only the transition of one magnitude into another, but a transition from quantity into quality," Hegel, The Science of Logic, 1816



Big Data is Different Data: A Proposition

- Hypothesis: extremely large data sets offer qualitatively different capabilities than small data sets.

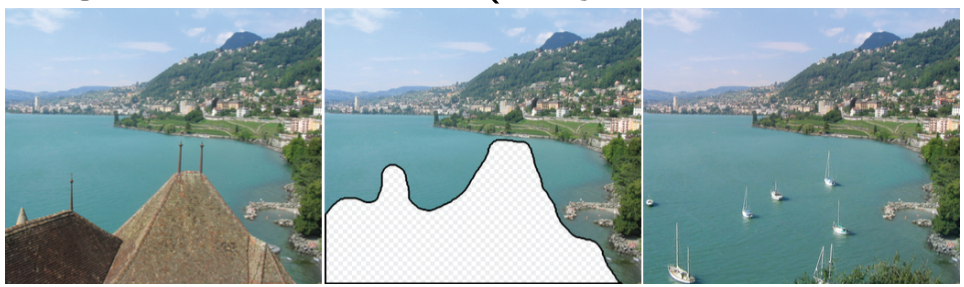
Big Data is Different Data: A Proposition

- Hypothesis: extremely large data sets offer qualitatively different capabilities than small data sets.
- Some Evidence: Image Completion (Hays & Efros, 2007)

"our initial experiments . . . on a dataset of ten thousand images were very discouraging. However, increasing the image collection to two million yielded a qualitative leap in performance"

Big Data is Different Data: A Proposition

- Hypothesis: extremely large data sets offer qualitatively different capabilities than small data sets.
- Some Evidence: Image Completion (Hays & Efros, 2007)

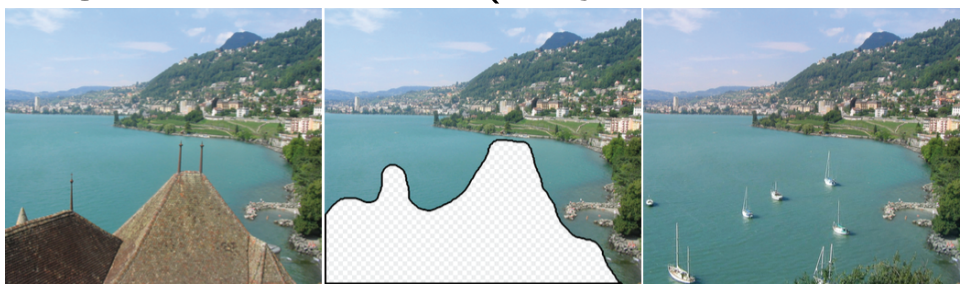


"our initial experiments . . . on a dataset of ten thousand images were very discouraging. However, increasing the image collection to two million yielded a qualitative leap in performance"

- Problem: Big data sets are big, unwieldy, computationally challenging, and highly redundant.

Big Data is Different Data: A Proposition

- Hypothesis: extremely large data sets offer qualitatively different capabilities than small data sets.
- Some Evidence: Image Completion (Hays & Efros, 2007)



"our initial experiments . . . on a dataset of ten thousand images were very discouraging. However, increasing the image collection to two million yielded a qualitative leap in performance"

- Problem: Big data sets are big, unwieldy, computationally challenging, and highly redundant.
- Research Quest: Can statistical predictions and actions be made cost effectively using the right small data?

Outline

Sets and set functions

We are given a finite "ground" set of objects:

$V =$ 

Also given a set function $f : 2^V \to \mathbb{R}$ that valuates subsets $A \subseteq V$.

Ex: $f(V) = 6$

Sets and set functions

Subset $A \subseteq V$ of objects:

$$A = \left\{ \right.$$

$$\left. \right\}$$

Also given a set function $f : 2^V \to \mathbb{R}$ that valuates subsets $A \subseteq V$.
Ex: $f(A) = 1$

Sets and set functions

Subset $B \subseteq V$ of objects:

$B = \{$      $\}$

Also given a set function $f : 2^V \to \mathbb{R}$ that valuates subsets $A \subseteq V$.
Ex: $f(B) = 6$

Two Equivalent Submodular Definitions

Definition (submodular)

A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A, B \subseteq V$, we have that:

$$f(A) + f(B) \geq f(A \cup B) + f(A \cap B) \tag{1}$$

Definition (submodular (diminishing returns))

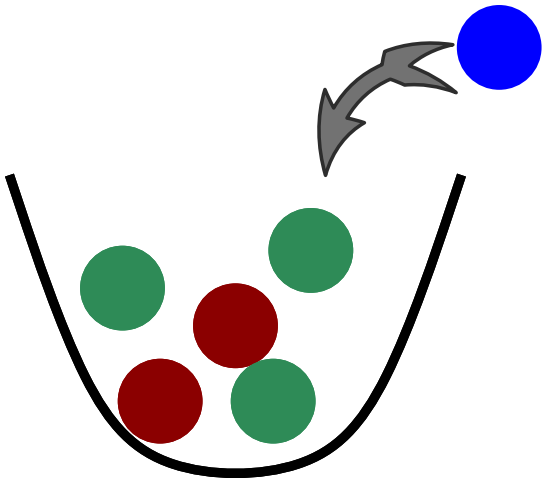
A function $f : 2^V \to \mathbb{R}$ is submodular if for any $A \subseteq B \subset V$, and $v \in V \setminus B$, we have that:

$$f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B) \tag{2}$$

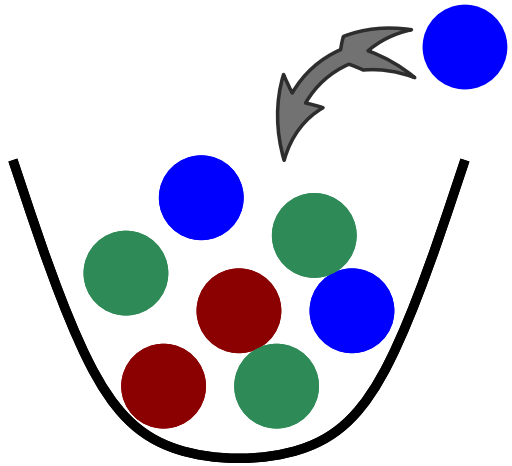
Incremental "value", "gain", or "cost" of v decreases (diminishes) as the context in which v is considered grows from A to B.

Example: Number of Colors of Balls in Urns

- Consider an urn containing colored balls. Given a set S of balls, $f(S)$ counts the number of distinct colors.



Initial value: 2 (colors in urn).
New value with added blue ball: 3



Initial value: 3 (colors in urn).
New value with added blue ball: 3

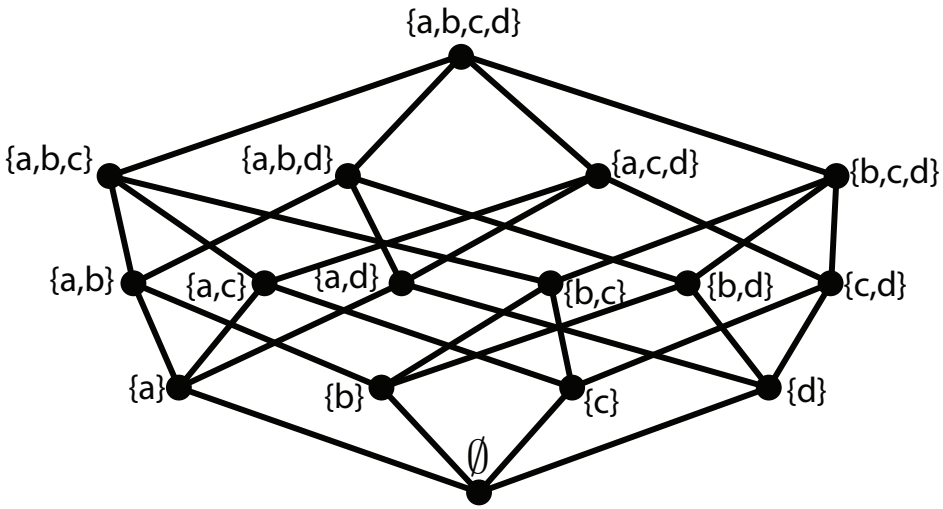
- Submodularity: Incremental Value of Object Diminishes in a Larger Context (diminishing returns). Thus, f is submodular.

Discrete Optimization

- Unconstrained minimization and maximization:

$$\min_{X \subseteq V} f(X) \quad (3) \qquad \max_{X \subseteq V} f(X) \quad (4)$$

- Knowing nothing about f, need 2^n queries for any quality assurance on candidate solution. Otherwise, solution can be unboundedly poor!!

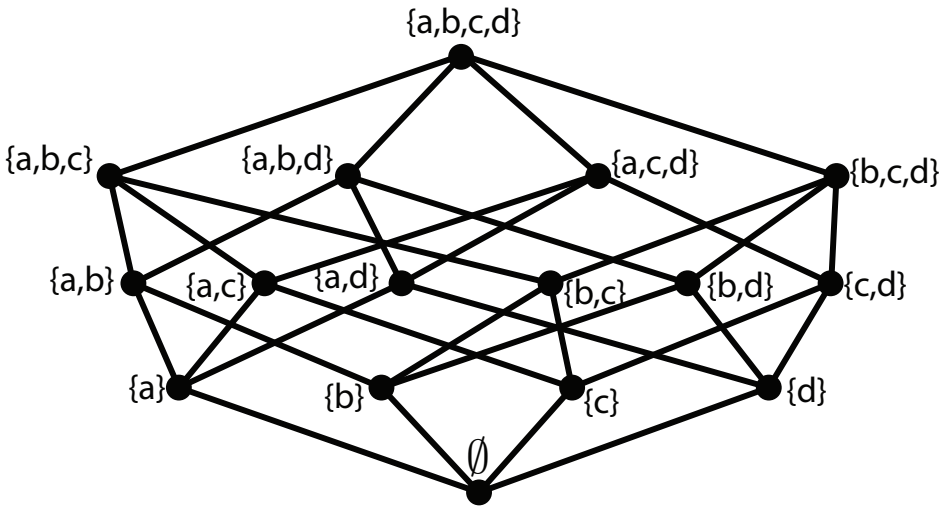


Discrete Optimization

- Unconstrained minimization and maximization:

$$\min_{X \subseteq V} f(X) \quad (3) \qquad \max_{X \subseteq V} f(X) \quad (4)$$

- Knowing nothing about f, need 2^n queries for any quality assurance on candidate solution. Otherwise, solution can be unboundedly poor!!



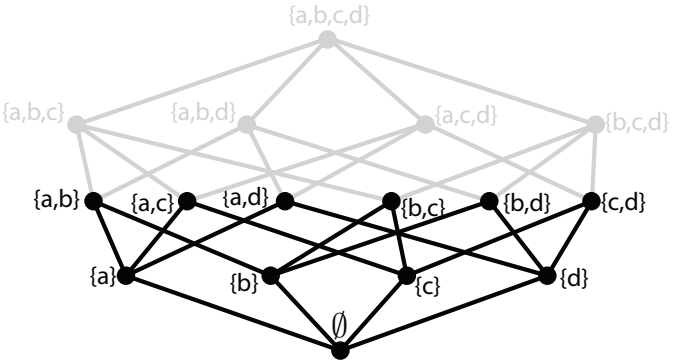
- When f is submodular, however, Eq. (3) is polytime, and Eq. (4) is constant-factor approximable.

Constrained Discrete Optimization

- Constrained case: interested only in a subset of subsets $\mathcal{S} \subseteq 2^V$.

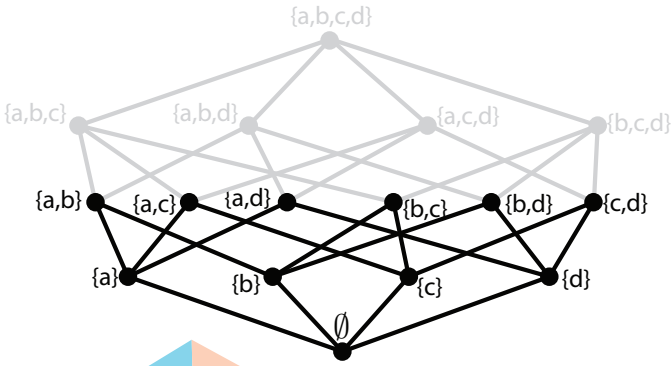
Constrained Discrete Optimization

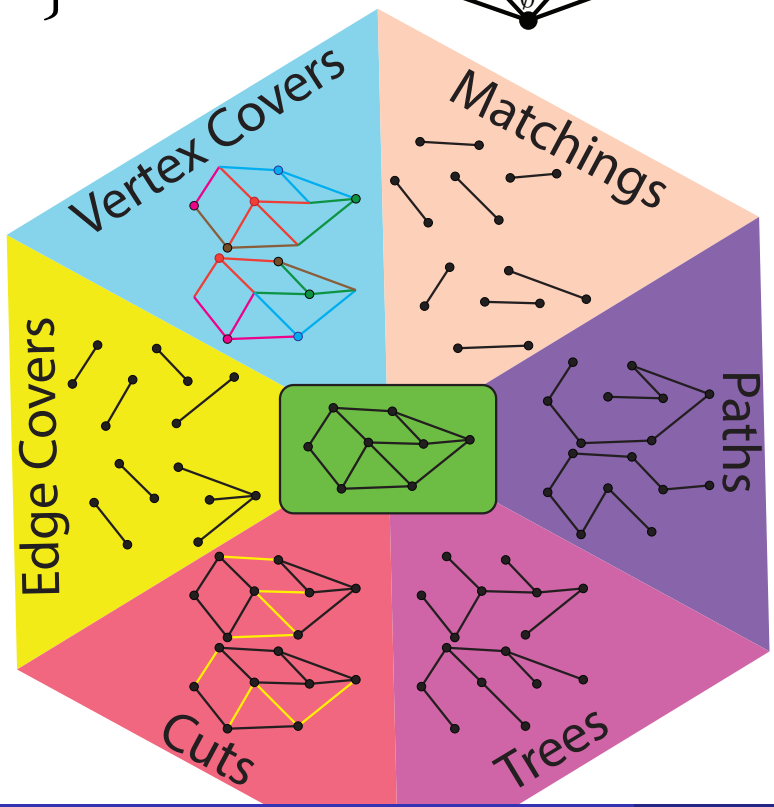
- Constrained case: interested only in a subset of subsets $\mathcal{S} \subseteq 2^V$.
- Ex: Bounded size $\mathcal{S} = \{S \subseteq V : |S| \leq k\}$, or given cost vector w and budget, bounded cost $\left\{S \subseteq V : \sum_{s \in S} w(s) \leq b\right\}$.



Constrained Discrete Optimization

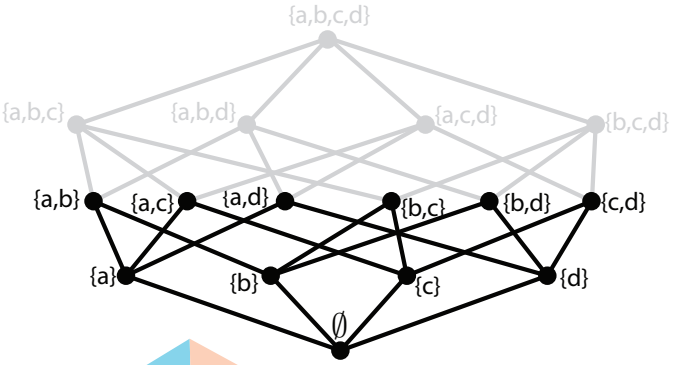
- Constrained case: interested only in a subset of subsets $\mathcal{S} \subseteq 2^V$.
- Ex: Bounded size $\mathcal{S} = \{S \subseteq V : |S| \leq k\}$, or given cost vector w and budget, bounded cost $\left\{S \subseteq V : \sum_{s \in S} w(s) \leq b\right\}$.
- Ex: feasible sets $\mathcal{S}$ as combinatorial objects.



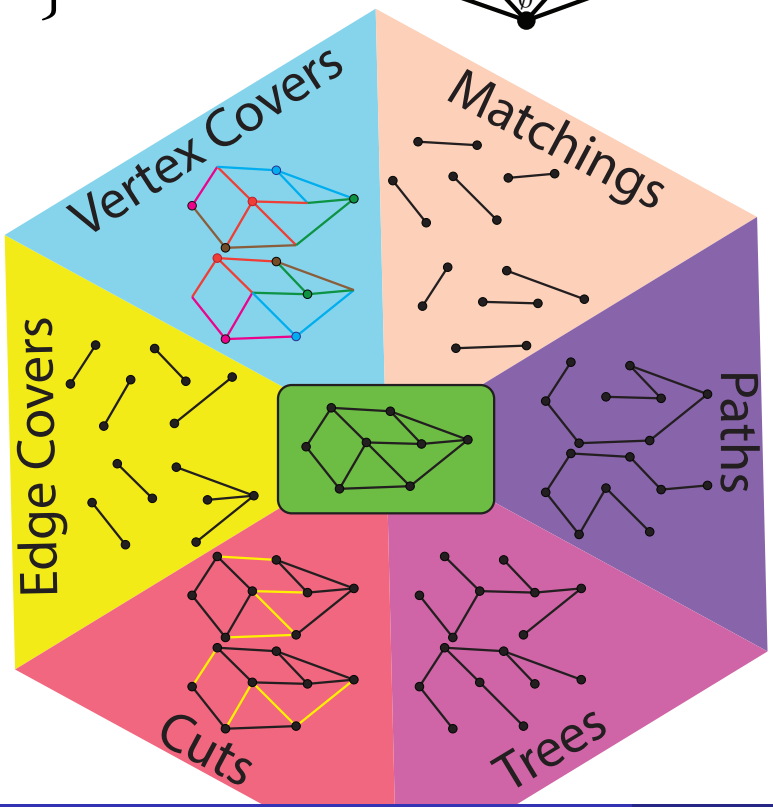


Constrained Discrete Optimization

- Constrained case: interested only in a subset of subsets $\mathcal{S} \subseteq 2^V$.
- Ex: Bounded size $\mathcal{S} = \{S \subseteq V : |S| \leq k\}$, or given cost vector w and budget, bounded cost $\left\{S \subseteq V : \sum_{s \in S} w(s) \leq b\right\}$.



- Ex: feasible sets $\mathcal{S}$ as combinatorial objects.



- Ex: feasible sets $\mathcal{S}$ as sub-level sets of g, $\mathcal{S} = \{S \subseteq V : g(S) \leq \alpha\}$, sup-level sets $\mathcal{S} = \{S \subseteq V : g(S) \geq \alpha\}$

Constrained Discrete Optimization

- Constrained discrete optimization problems:

$$\begin{array}{ll} \text{maximize} & f(S) \\ \text{subject to} & S \in \mathcal{S} \end{array} \quad (5) \qquad \begin{array}{ll} \text{minimize} & f(S) \\ \text{subject to} & S \in \mathcal{S} \end{array} \quad (6)$$

  where $\mathcal{S} \subseteq 2^V$ is the feasible set of sets.

Constrained Discrete Optimization

- Constrained discrete optimization problems:

$$\begin{array}{ll} \text{maximize} & f(S) \\ \text{subject to} & S \in \mathcal{S} \end{array} \quad (5) \qquad \begin{array}{ll} \text{minimize} & f(S) \\ \text{subject to} & S \in \mathcal{S} \end{array} \quad (6)$$

  where $\mathcal{S} \subseteq 2^V$ is the feasible set of sets.
- Fortunately, when f (and g) are submodular, these problems can often be solved with guarantees, often very efficiently! *(Feige, Mirrokni & Vondrák 20XX; Goel, Karande, Tripathi & Wang; Svitkina & Fleischer 2010; Jegelka & Bilmes 2011, Iyer, Jegelka, & Bilmes 2013, Iyer & Bilmes 2014, and many others).*

The Entropy Function

- Given a joint distribution $p(x_V)$ over $|V|$ random variables:

$$f(A) \triangleq H(X_A) = -\sum_{x_A} p(x_A) \log p(x_A) \tag{7}$$

with $p(x_V)$ joint probability distribution over X_V.

- Shannon's incredible 1948 paper stated that entropy is subadditive $H(X_A) + H(X_B) \leq H(A \cup B)$. McGill 1954 stated that further conditioning reduces entropy, $H(X_A|X_B) \geq H(X_A|X_B, X_C)$.
- This condition is identical to submodularity!

A Salmagundi of Submodularity

- Many diverse functions are submodular. E.g., set cover, vertex cover, edge cover, graph cut, bipartite neighborhoods, facility location, sums of weighted concave composed with additive functions, matrix rank, matroid rank, entropy, KL-divergence functions, quantum entropy, log determinant, spectral functions applied to Hermitian matrix, etc.

A Salmagundi of Submodularity

- Many diverse functions are submodular. E.g., set cover, vertex cover, edge cover, graph cut, bipartite neighborhoods, facility location, sums of weighted concave composed with additive functions, matrix rank, matroid rank, entropy, KL-divergence functions, quantum entropy, log determinant, spectral functions applied to Hermitian matrix, etc.
- All submodular functions express a form of "abstract combinatorial independence" or "generalized complexity"
- Given submodular f, $\exists$ a notion of "independence" , i.e., $A \perp\!\!\!\perp B$:

$$f(A \cup B) = f(A) + f(B), \tag{8}$$

- and a notion of "conditional independence" , i.e., $A \perp\!\!\!\perp B | C$:

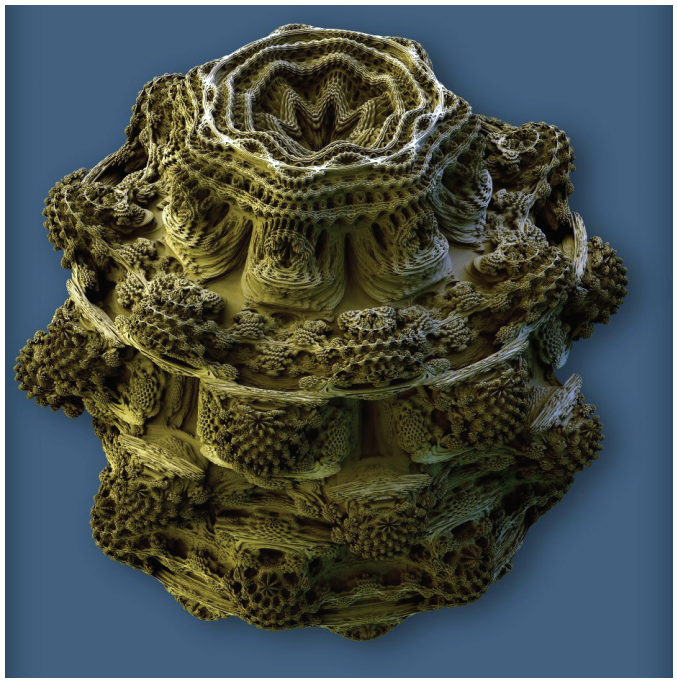
$$f(A \cup B \cup C) + f(C) = f(A \cup C) + f(B \cup C) \tag{9}$$

Submodularity as a Model for Real-World Data

- Submodularity: useful model to valuate set of data objects.
- Meaning of valuations typically depend on if function is to be maximized or minimized.

Maximization
diversity, coverage, span, irredundancy, and information.

Minimization
cooperative costs, complexity, roughness, and irregularity.



Big Data Summarization: Submodular Approaches

- We are given a set indexed by V
- Approach: 1) find a good function $f : 2^V \rightarrow \mathbb{R}_+$ that represents information in V. 2) Then optimize f to obtain a subset.

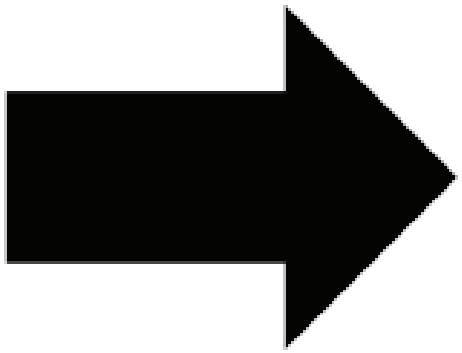
1) Heuristic: design f by hand, hoping that f is a good proxy for the information within V. Acknowledge that f is a surrogate objective, guarantees are only in terms of f.
2) Alternatively: attempt to learn f, or some aspects of a good f, in some fashion based on training data.

- We report on both kinds of results for summarizing large data sets below.

Outline

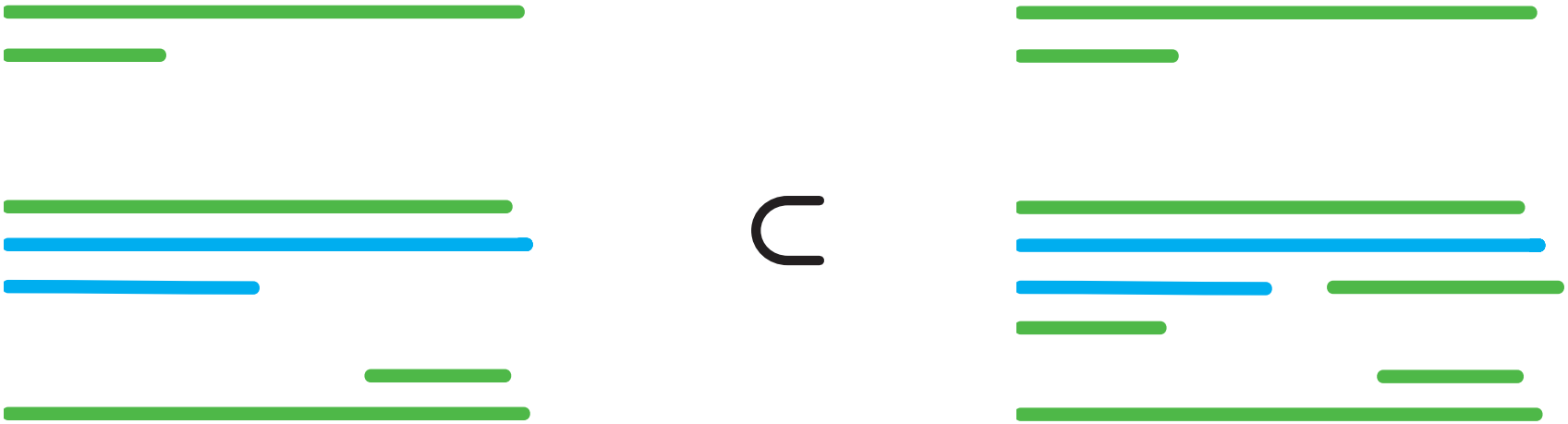
Document Summarization

Given a large set of documents, summarize it down to the essential set of sentences.



Extractive Document Summarization

- We extract sentences (green) as a summary of the full document



- The summary on the left is a subset of the summary on the right.
- Consider adding a new (blue) sentence to each of the two summaries.
- Marginal benefit of adding the new (blue) sentence to the smaller (left) summary is no less than the marginal benefit of adding blue sentence to the larger (right) summary.

Submodularity for document summarization?

Exists many unintentional uses of submodularity in NLP community.

- E.g., maximum marginal relevance (MMR) (Carbonell & Goldstein, 1998) has a diminishing returns property.
- Modified MMR - (McDonald, 2007)
- Concept-based approaches (Filatova & Hatzivassiloglou 2004; Takamura & Okumura, 2009; Riedhammer et al., 2010; Qazvinian et al., 2010).
- Automatic evaluation of candidate summarizes are submodular: ROUGE-N (Lin 2004) and Pyramid (Nenkova & Passonneau, 2004).
- Both ROUGE-N and Pyramid are parameterized by good quality summarizes produced by humans, used only for evaluation.

Structured Learning of Submodularity

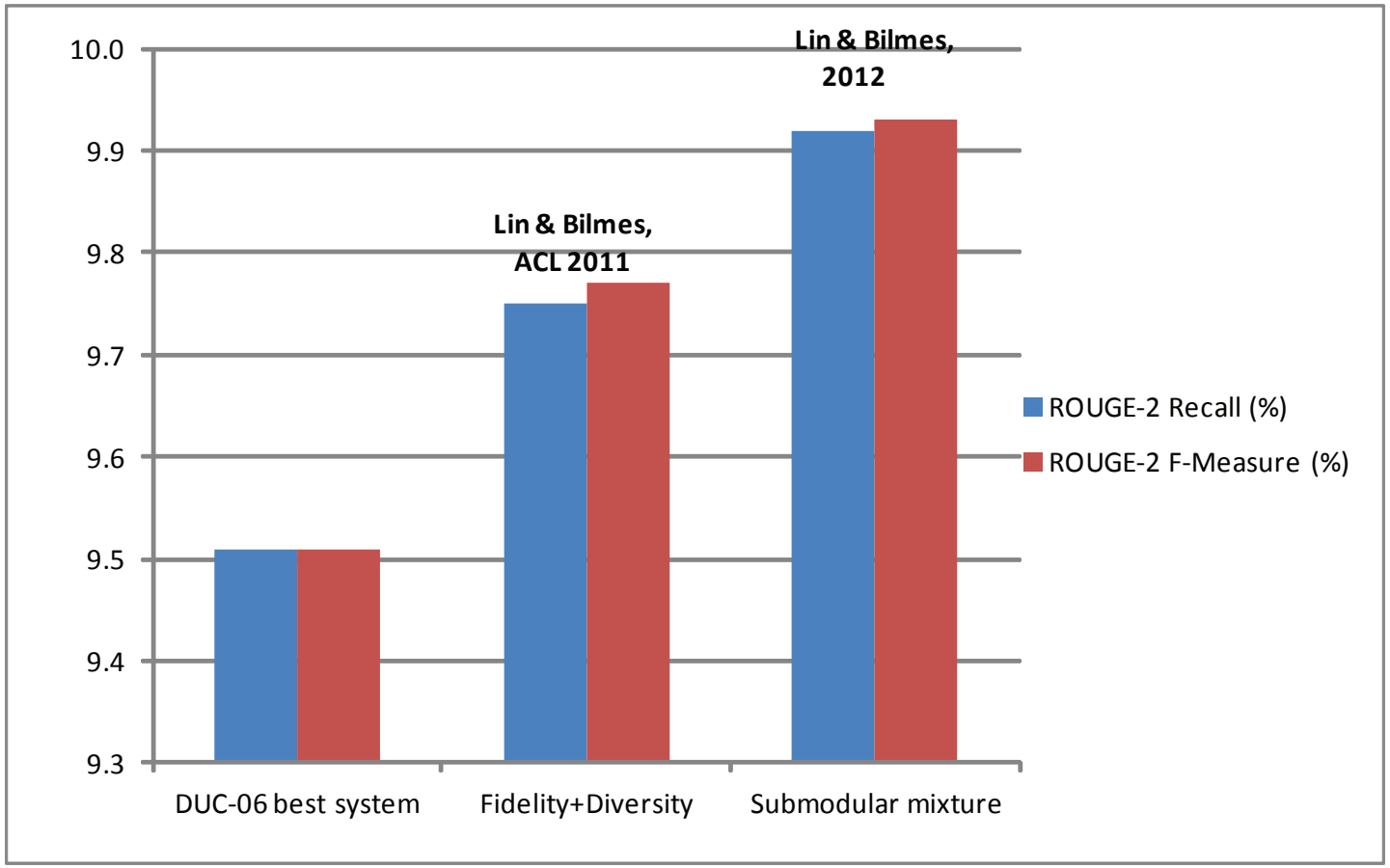
- $\mathbf{w}^\top \mathbf{f}_t(\mathbf{y})$ is a convex combination of submodular functions.
- Mixture weights can be learnt via structured max-margin.

$$\begin{aligned} \underset{\mathbf{w}\geq 0, \xi_t}{\text{minimize}} \quad & \frac{1}{T}\sum_t \xi_t + \frac{\lambda}{2}\|\mathbf{w}\|^2 && (10) \\ \text{subject to} \quad & \mathbf{w}^\top \mathbf{f}_t(\mathbf{y}^{(t)}) \geq \max_{\mathbf{y}\in\mathcal{Y}_t} \left(\mathbf{w}^\top \mathbf{f}_t(\mathbf{y}) + \ell_t(\mathbf{y})\right) - \xi_t, \forall t \\ & \xi_t \geq 0, \forall t. && (11) \end{aligned}$$

- Exponential set of constraints reduced to an embedded optimization problem, "inference."

Typical Results (DUC-06)

Rouge-2: higher is better



Outline

As the data set size grow . . .

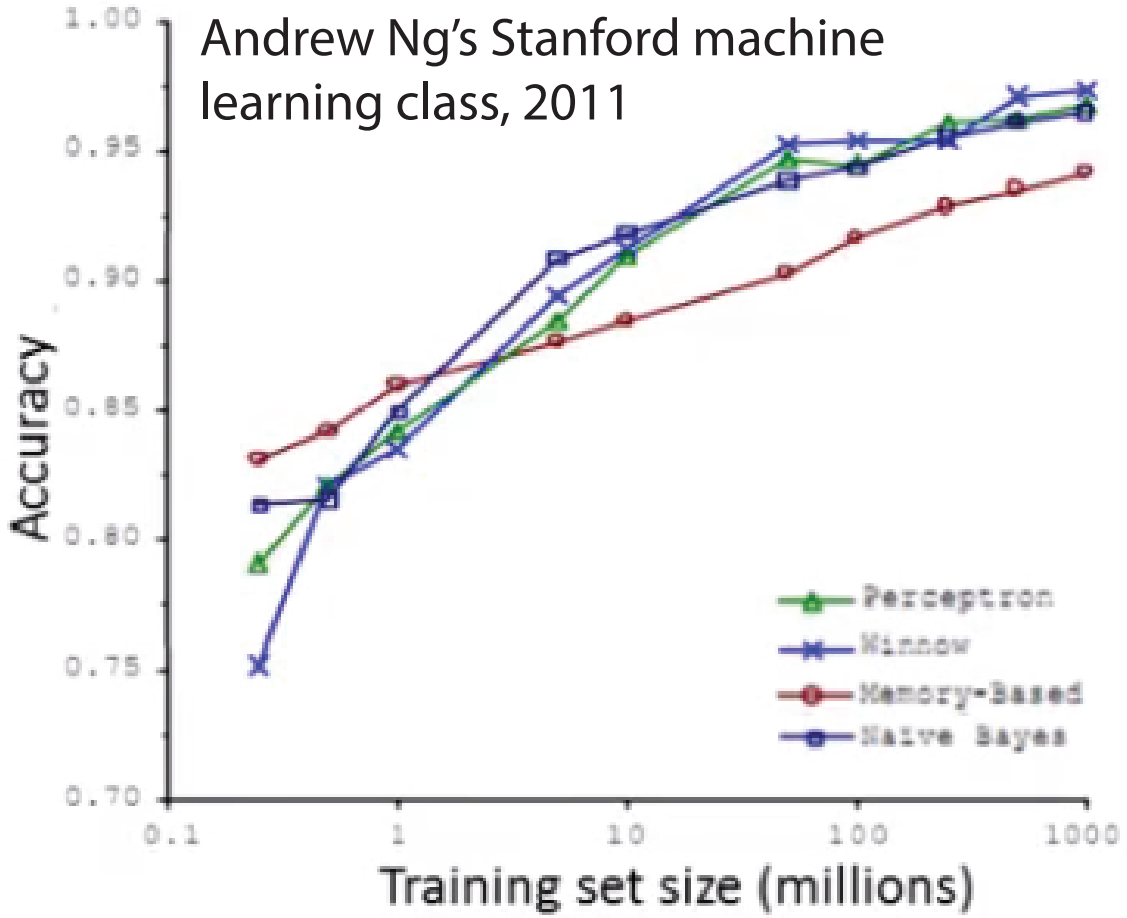
There is no data like more data

As the data set size grow . . .

~~There is no data like more data~~ $\Rightarrow$ more data is like no more data.

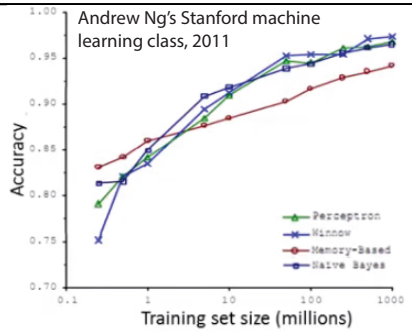
As the data set size grow . . .

~~There is no data like more data~~ $\Rightarrow$ more data is like no more data.

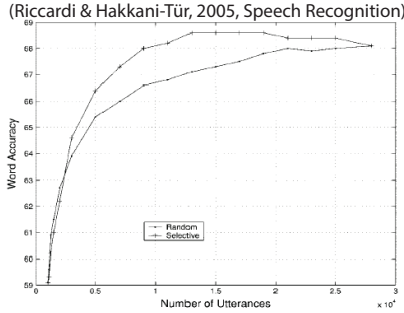


As the data set sizes grow . . .

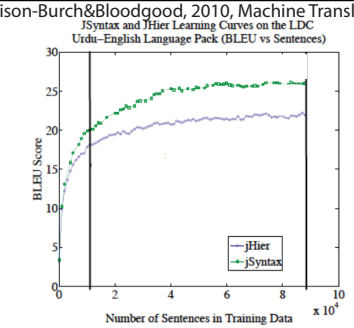
Andrew Ng's Stanford machine learning class, 2011



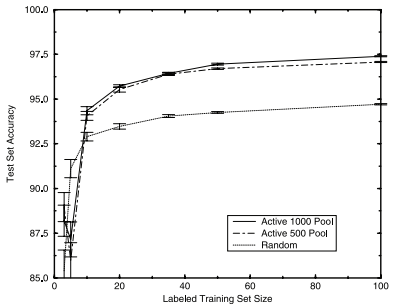
(Riccardi & Hakkani-Tür, 2005, Speech Recognition)



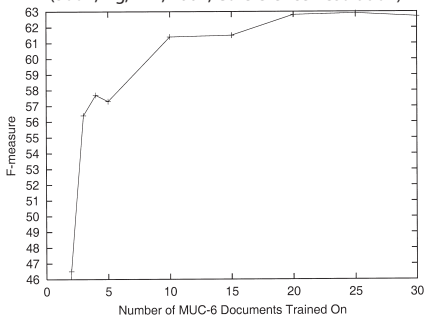
(Callison-Burch&Bloodgood, 2010, Machine Translation)



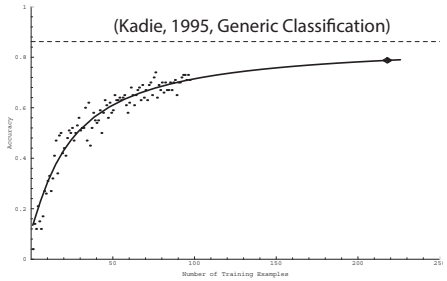
Tong & Koller, 2001



(Soon, Ng, Lim, 2001, Coreference Resolution)

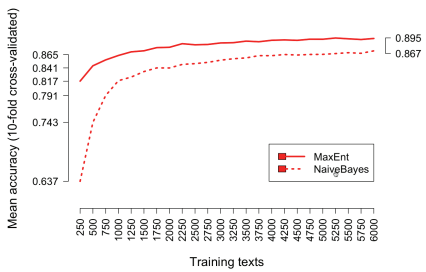


(Kadie, 1995, Generic Classification)

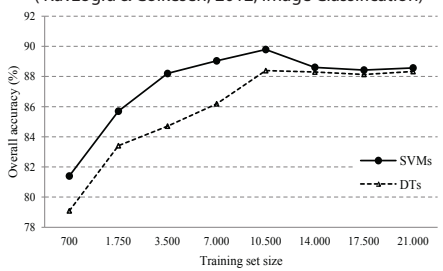


Sentiment Tutorial, Chris Potts, Stanford Ling., 2011

Sentiment-aware + neg. marking; OpenTable; 6000 test reviews

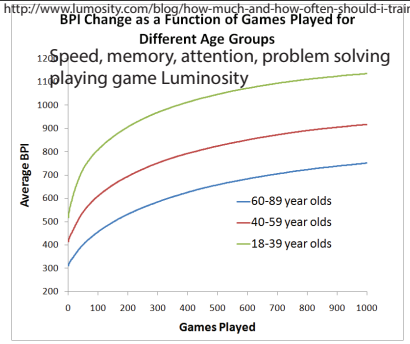


(Kavzoglu & Colkesen, 2012, Image Classification)



http://www.lumosity.com/blog/how-much-and-how-often-should-i-train/

Speed, memory, attention, problem solving playing game Luminosity



Practical Goals: Submodular Proxies

- Research question: Can statistical predictions be cost effective using small data?
- Using a submodular model that is easy to evaluate, can we produce useful and easy to obtain training data subsets that still perform well enough?

Outline

Speech Subset Selection

- Corpus Summarization: Given a large set of speech utterances (training corpus) $V = \{v_1, v_2, \ldots, v_n\}$, choose a small subset $A \subseteq V$ that is representative of V.
- Goal: training on summary should yield highest accuracy possible.
- Focus on drastic reductions in training set (one to two orders magnitude) to reduce model design time.

Submodular Switchboard: GMM and DNN

	1%	5%	10%	20%	all
Rand	52.1± 1.5	38.2±0.2	35.1±0.3	34.4±0.2	
HE (words)	49.6	36.5	34.8	N/A	
HE (3-phones)	47.5	37.6	34.2	N/A	31.0
SM (3-phones)	**47.5**	**35.7**	**33.3**	**32.6**	

Table : Word error rates, random (Rand), histogram-entropy (HE), the submodular (SM) system. Histogram-entropy results saturate after 10%.

	1%	5%	10%	20%	all
Rand	43.7±0.5	34.3±0.9	31.5±0.5	29.8±0.2	
HE (3-phones)	42.8	33.9	31.3	N/A	26.0
SM (3-phones)	**41.1**	**31.8**	**29.3**	**28.2**	

Table : Word error rates for DNN system.

Outline

Data subset selection for machine translation

- Statistical Machine Translation (SMT): automatically translate from one human language to another.
- Common problems in SMT: 1) test data is from a target domain while training data is mixed-domain; 2) phrase translation table, when based on all training data, can be massive.
- Solution: choose and then train using only a (domain-specific) subset of training data.
- Many previous approaches (e.g., n-gram overlap (Ittycheriah & Roukos, 2007), coverage of unseen n-grams (Eck et al. 2005), feature decay approach (Biçici & Yuret, 2011-2013)) are inadvertently submodular.
- Some (e.g., Moore & Lewis, 2010) are only modular.
- We approach directly using submodular functions.

Feature based submodular functions

- V is set of sentences, U is a set of features (n-grams in our work).

Feature based submodular functions

- V is set of sentences, U is a set of features (n-grams in our work).
- Feature-based submodular functions:

$$f(X) = \sum_{u \in U} w_u \phi_u(m_u(X)) \tag{12}$$

where $w_u > 0$ is a feature weight, $m_u(X) = \sum_{x \in X} m_u(x)$ is a non-negative modular function specific to feature u, $m_u(x)$ is a relevance score, a non-negative scalar score indicating the relevance of feature u in object x, and ϕ_u is a u-specific non-negative non-decreasing concave function.

Results on Two Standard Translation Corpora

Method	**Data Subset Sizes**				
	10%	20%	30%	40%	100%
	NIST				
Random	0.3991	0.4142	0.4205	0.4220	0.4257
Cross-entropy	0.4235	0.4292	0.4290	0.4292	
Submodular-5	0.4285	**0.4356**	0.4333	**0.4324**	
Submodular-6	**0.4302***	**0.4334**	**0.4371***	**0.4349**	
	EUROPARL				0.2651
Random	0.2590	0.2652	0.2677	0.2697	
Cross-entropy	0.2639	0.2687	0.2704	0.2723	
Submodular-5	0.2653	**0.2727**	0.2697	0.2720	
Submodular-6	**0.2697***	**0.2700**	**0.2740***	0.2723	

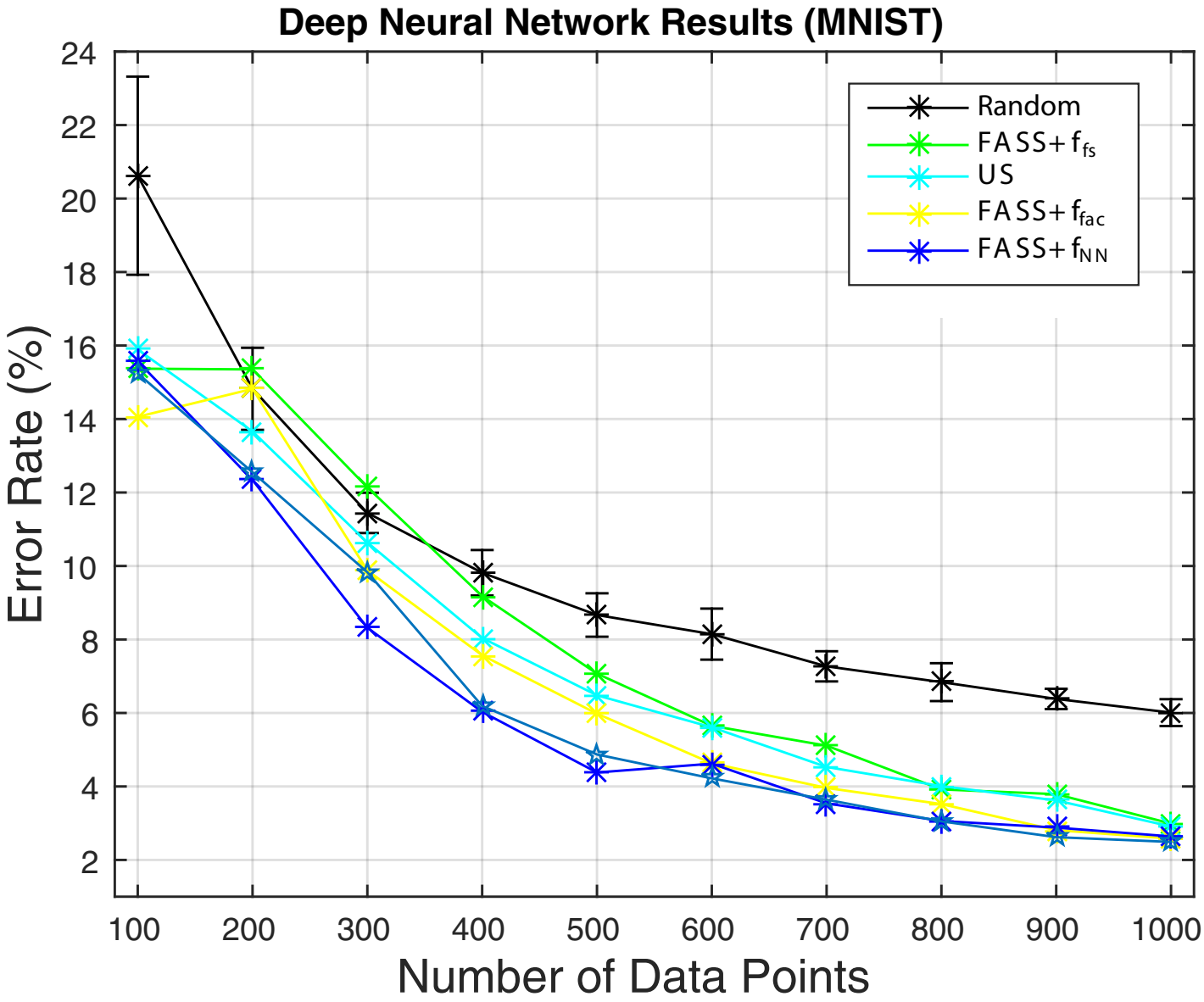
BLEU test-set scores (higher is better) for random, cross-entropy (standard baseline), and various submodular methods. Bold = significant over best Xent system.

Outline

Handwritten Image Recognition

- Task: Train a machine learning system to classify hand-written digits correctly.
- A standard data set MNIST (Lecun,'98).
- Deep Neural Networks are state-of-the art on this kind of data, but they are very slow to train and greatly benefit from GPU hardware.

Image Recognition, Submodular Selection



FASS = Filtered Active Submodular Selection

US = Uncertainty Sampling

f_{fs}, f_{fac}, and f_{NN} are three submodular functions.

Outline

Modern Image collections

Many images, also that have a higher level gestalt than just a few.

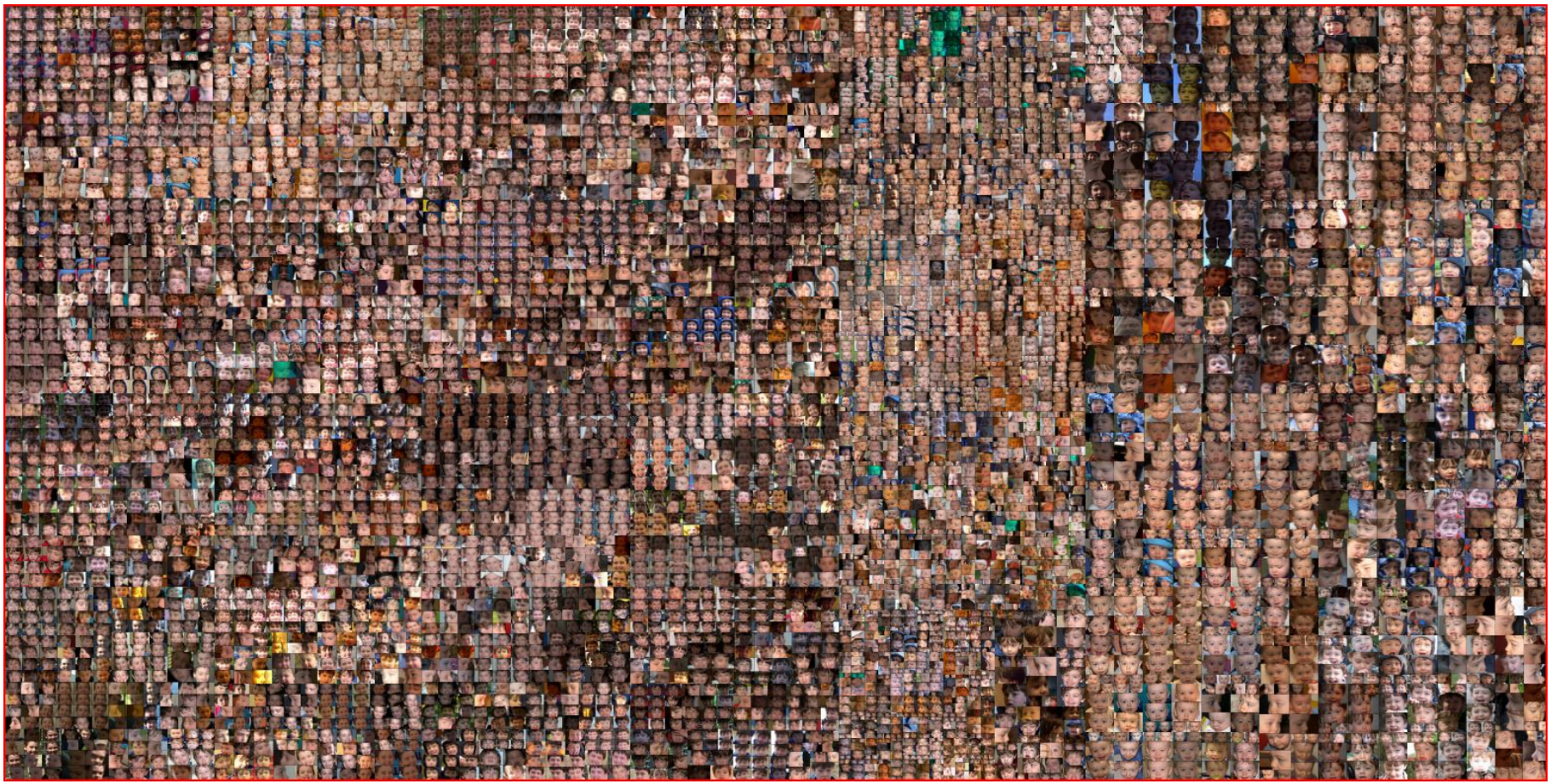


Image Summarization

Task: Summarize collection of images by representative subset of the images

Applications:

- Summarizing your holiday pictures.
- Summarizing image search results
- Efficient browsing of image collections
- Video frame summarization

Image Summarization - Data Collection

Data Statistics

- 14 image collections with 100 pictures each
- $\sim$ 400 human summaries for every image collection, via Amazon Turk, about 5500 summaries total!

Example collections:

Image Summarization

Super-Pixel Based V-Rouge

Whole collection:

3 best summaries:

3 medium summaries:

3 worst summaries:

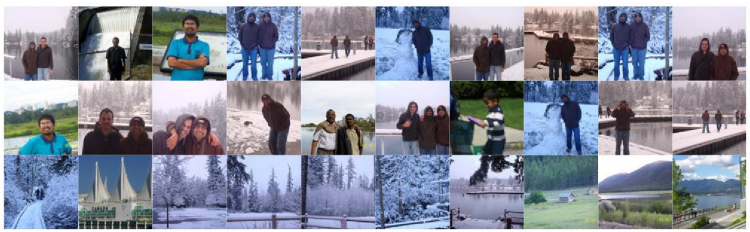
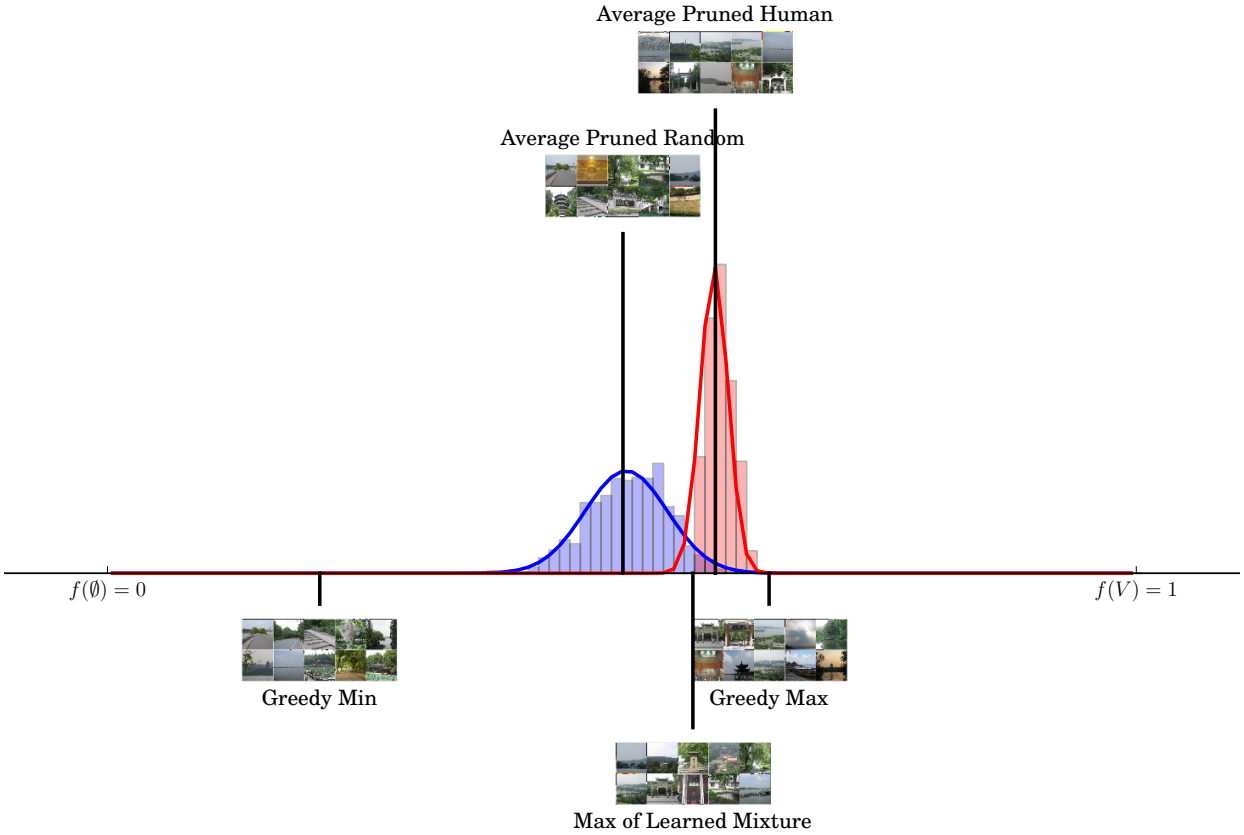


Image Summarization

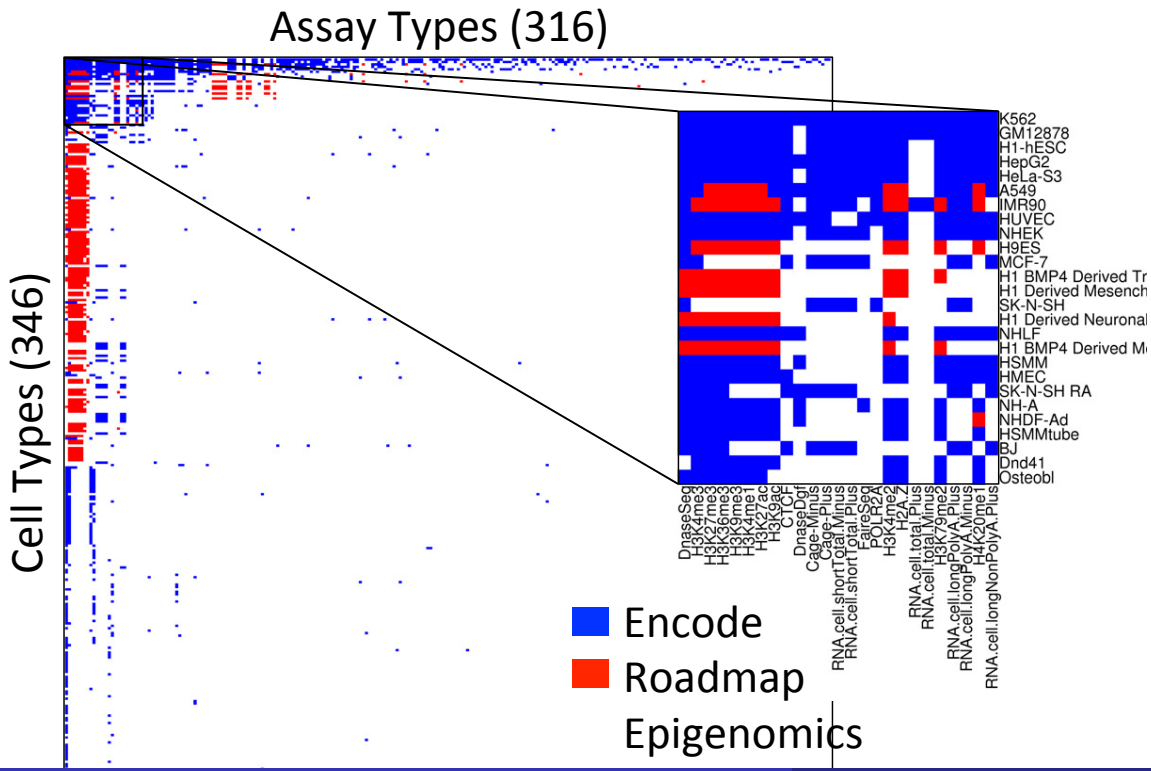
Typical Results - Learnt mixture using Max-Margin



Outline

Assay Selection

- Functional genomics: improve understanding of gene function (e.g., transcription, translation, and forms of interactions).
- Genomic assays measure a DNA activity in cell-type specific nucleus. E.g., transcription, transcription factor binding, DNA accessibility, covalent modification of the histone proteins, etc.
- Cost prohibitive: The cost of genomics assays limits their application. Ideally, to fully characterize a cell type, one desires every possible assay type, but too costly.



Assay Selection: Typical Results

- Fortunately, assay types are often redundant.
- Goal: choose a subset of assay types such that remaining assay types can be deduced.
- Below are some preliminary results on summarizing genomic assays, quality of summaries measured in a variety of ways.

metric	whole data set	averaged random	facility location	saturated coverage
annotation prediction	0.434	0.270±0.039	0.319	**0.330**
assay prediction (SVR)	0.690	0.977±0.028	**0.898**	0.943
assay prediction (LR)	0.728	1.032±0.025	**0.969**	0.975
function prediction (F-measure)	0.0489	0.0029±0.0025	0.0065	**0.0081**
function prediction (AUC-PR)	0.076	0.027±0.009	0.038	**0.042**

Outline

Acknowledgments

Thanks to the following amazing people (former & current students, and current colleagues):

Mukund Narasimhan, Hui Lin, Andrew Guillory, Stefanie Jegelka, Sebastian Tschiatschek, Kai Wei, Yuzong Liu, Rishabh Iyer, Jennifer Gillenwater, Yoshinobu Kawahara, Katrin Kirchhoff, Carlos Guestrin, & Bill Noble.

The End

The End: Thank you!

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