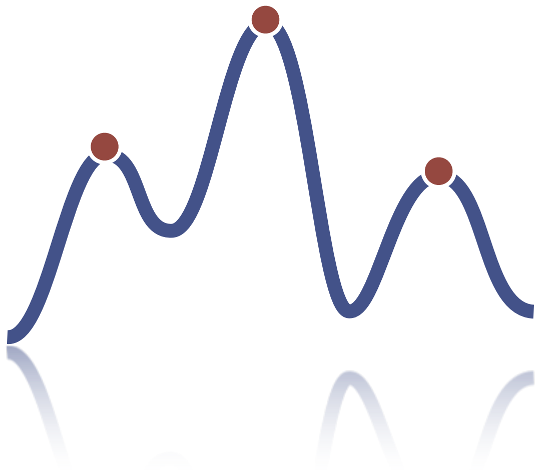
Determinantal Point Processes

Alex Kulesza

with Ben Taskar and Jennifer Gillenwater

Previously...

- M-best MAP
- Diverse M-best MAP
- Sampling



Use **single-output** model **multiple times**

Use **single-output** model **multiple times**

$$\mathcal{P}(\boldsymbol{y}_1) \qquad \mathcal{P}(\boldsymbol{y}_2) \qquad \mathcal{P}(\boldsymbol{y}_3)$$

A unified approach:

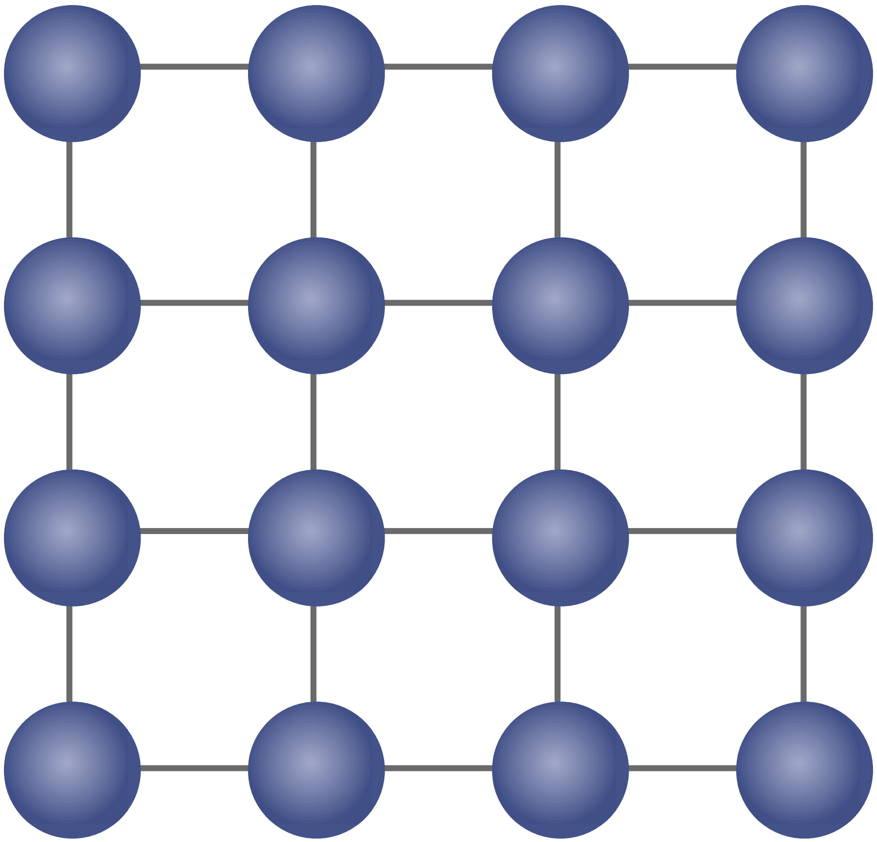
Explicitly model **sets** of **multiple** outputs

$$\mathcal{P}(\{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3\})$$

Explicitly model **sets** of **multiple** outputs

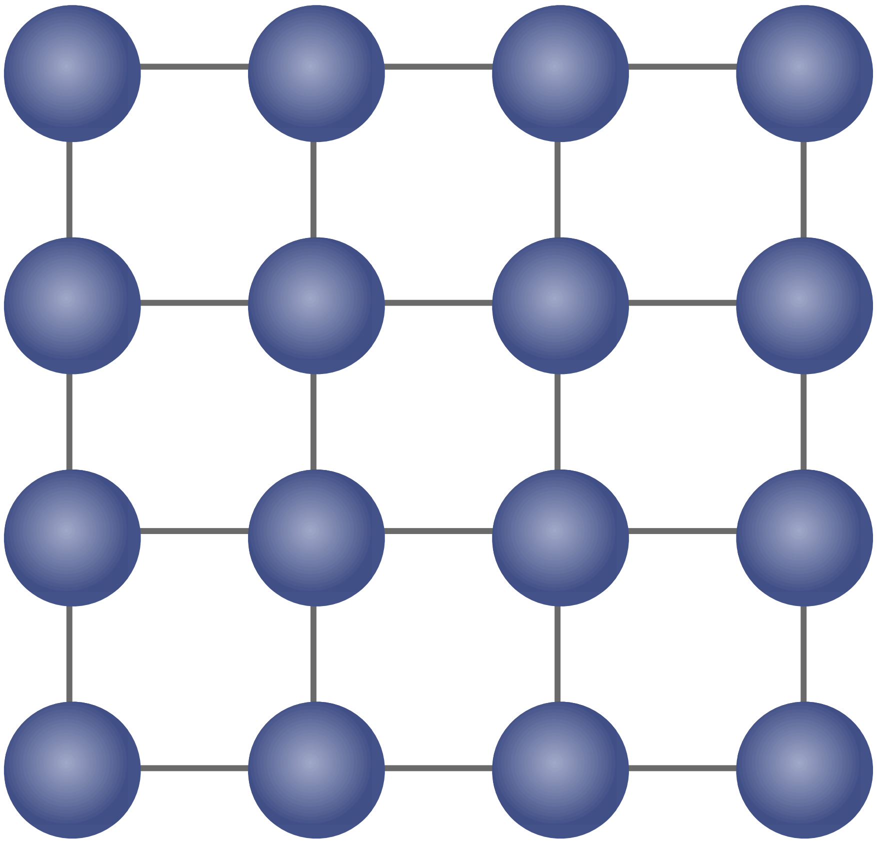
$$\mathcal{P}(\{\boldsymbol{y}_1, \boldsymbol{y}_2, \boldsymbol{y}_3\})$$

- Sample entire sets of multiple predictions
- Marginal and conditional probabilities
- How can this be efficient?



10,000

pixels

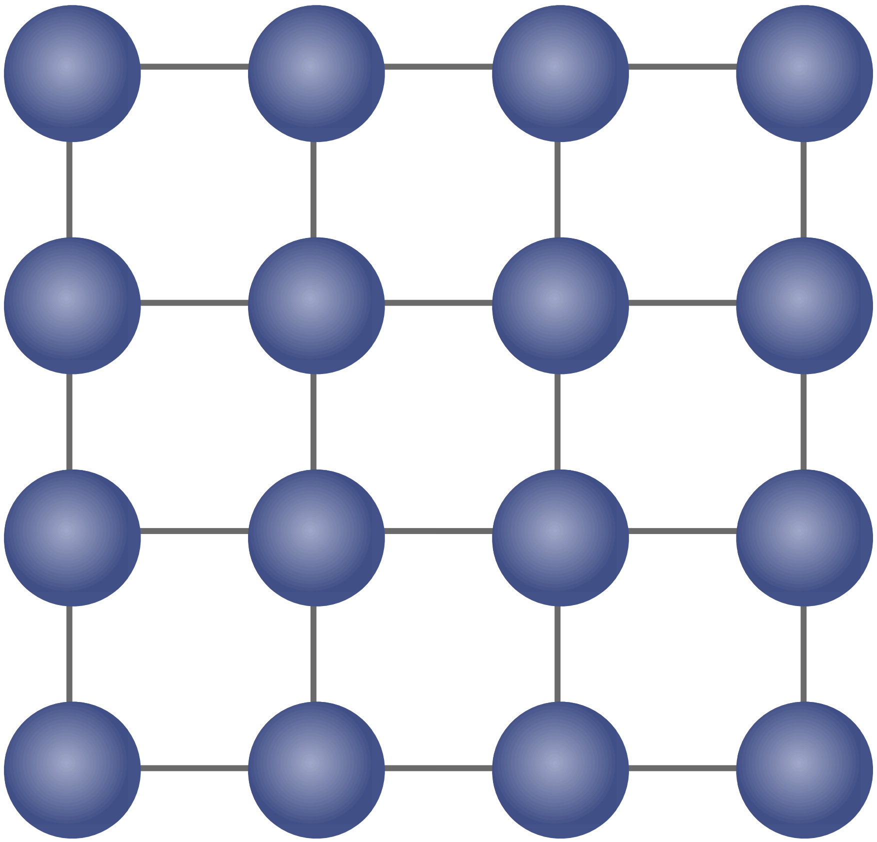


10,000

pixels

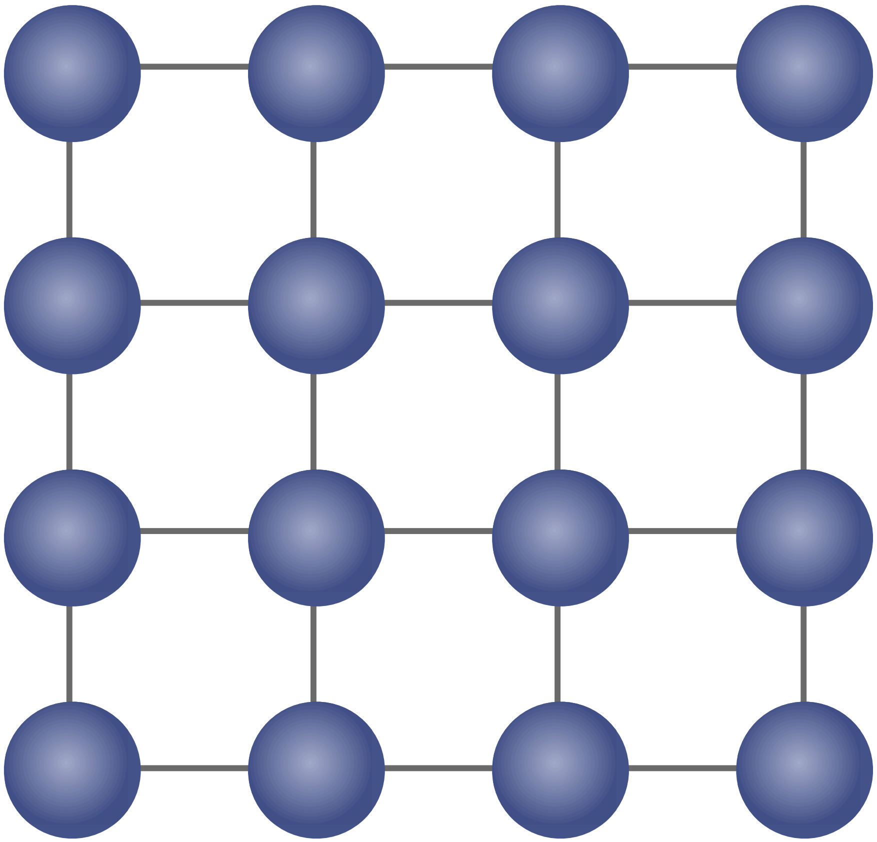
10

labels



$10{,}000^{10}$

structures

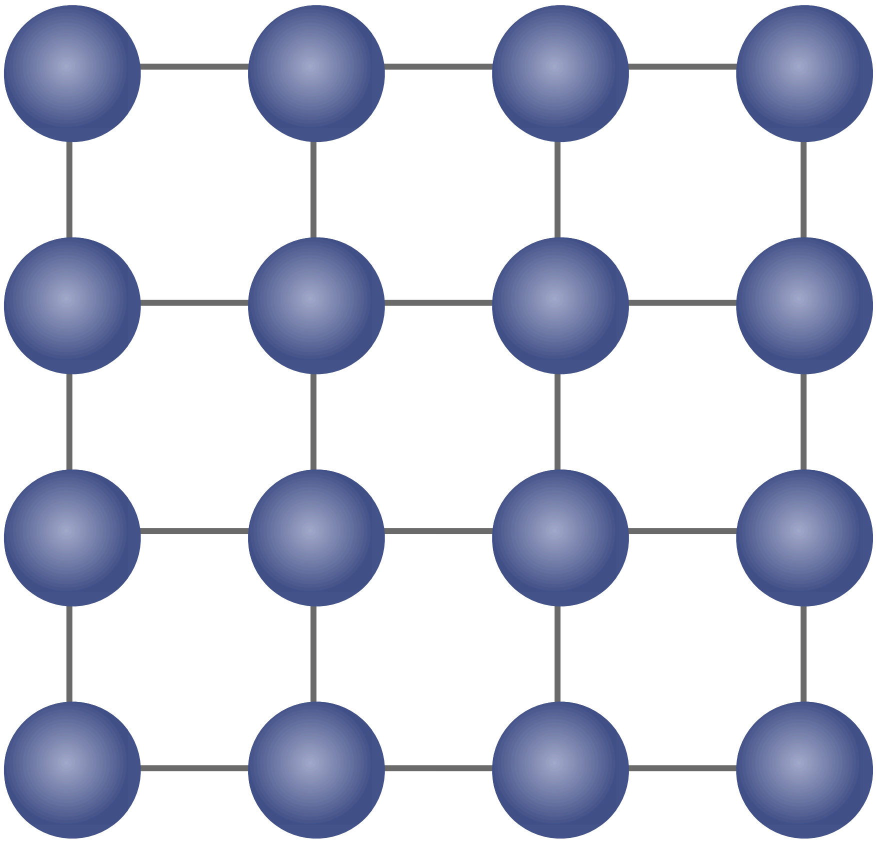


$10{,}000^{10}$

structures

10

predictions



$$\left(10{,}000^{10}\right)^{10}$$

sets of structures

Determinantal Point Processes

- Encode **diversity** using kernel matrix
- Linear algebra makes inference easy (and fun)
- Probabilistic models of diverse sets of objects
- We will extend to **structured** objects
- But let's start at the beginning...

Image search: “jaguar”

Relevance only:

…

Relevance + diversity:

…

Summarization

Importance only:

- NSA collecting customers' phone records
- NSA, Verizon surveillance program revealed
- NSA's phone snooping a different kind of creepy

Summarization

Importance + coverage:

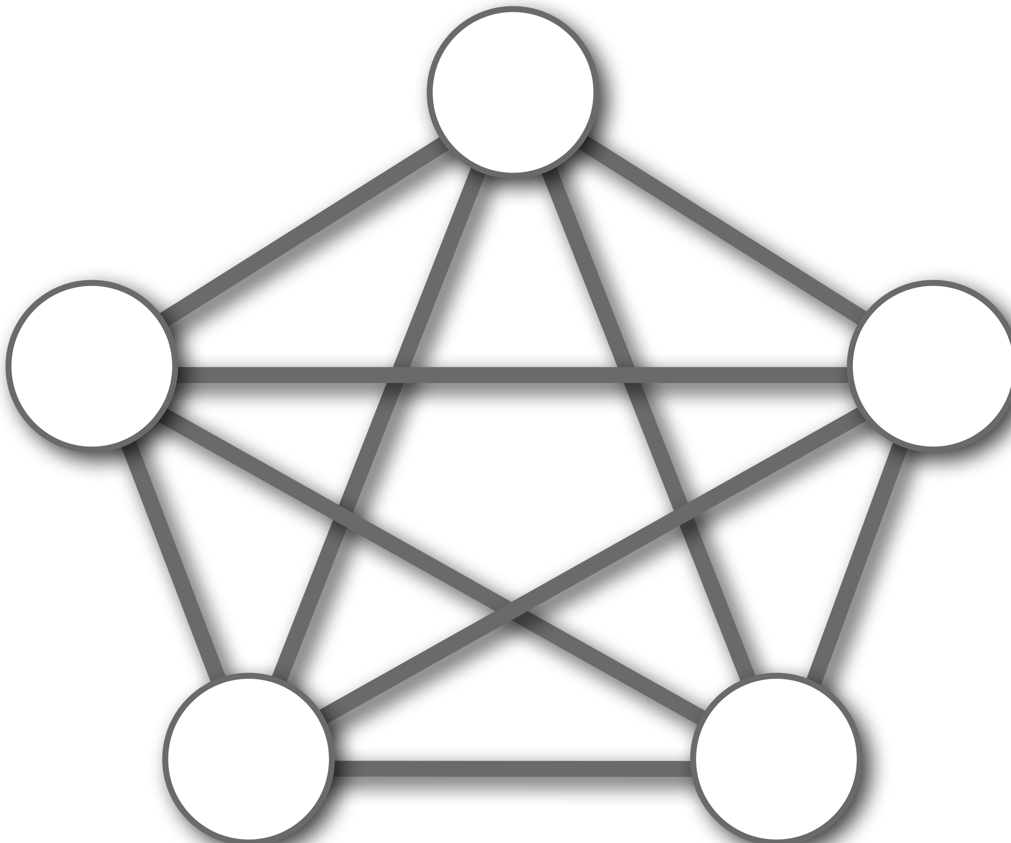
- NSA collecting phone records
- PRISM: How the NSA wiretapped the Internet
- GCHQ taps fibre-optic cables
- Google, Apple, Facebook deny PRISM involvement

Google Microsoft verizon facebook

Graphical models?

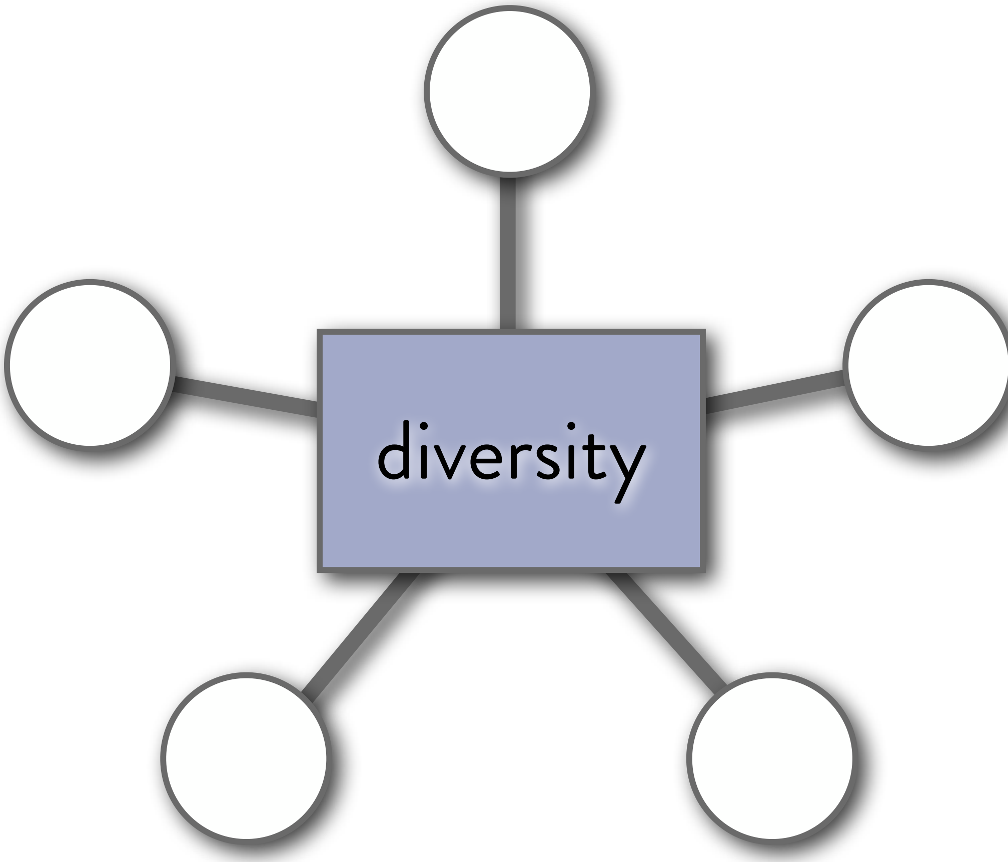
item *i*

Graphical models?



Loopy, **negative** interactions are hard

Determinantal point processes (DPPs)



Global, **negative** interactions are easy

Supporting Materials

- **Tech report:** http://arxiv.org/abs/1207.6083 (120 pages, with all the proofs!)
- **Matlab Code:** http://www.eecs.umich.edu/~kulesza/code/dpp.tgz

Outline

Part I Representation, inference, comparison to other models, learning

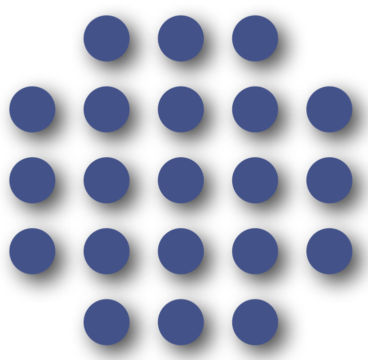
Part II Large-scale inference, extensions, sets of structures, applications

Part I

- Representation
- Inference: Marginals, Conditionals
- Inference: Sampling
- DPPs vs MRFs
- Learning

Discrete point processes

$\mathcal{Y}$



$\mathcal{P}$  $= 0.02$

$\mathcal{P}$  $= 0.01$

$\vdots$

Discrete point processes

- N items (e.g., images or sentences):

$$\mathcal{Y} = \{1, 2, ..., N\}$$

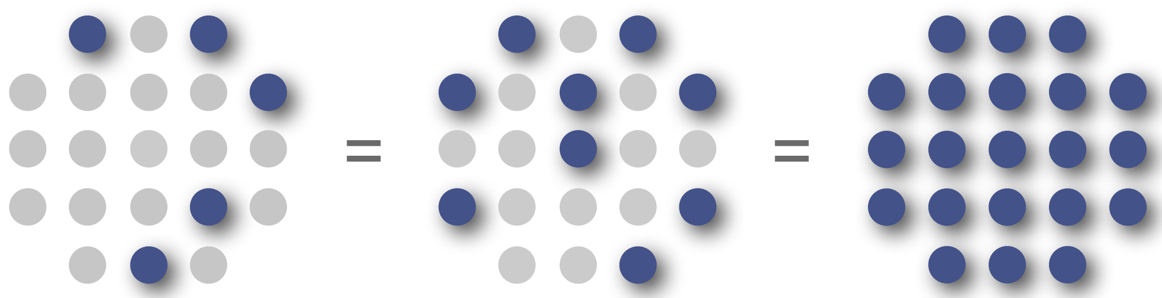
- 2^N possible subsets
- Probability measure $\mathcal{P}$ over subsets $Y \subseteq \mathcal{Y}$

Independent point process

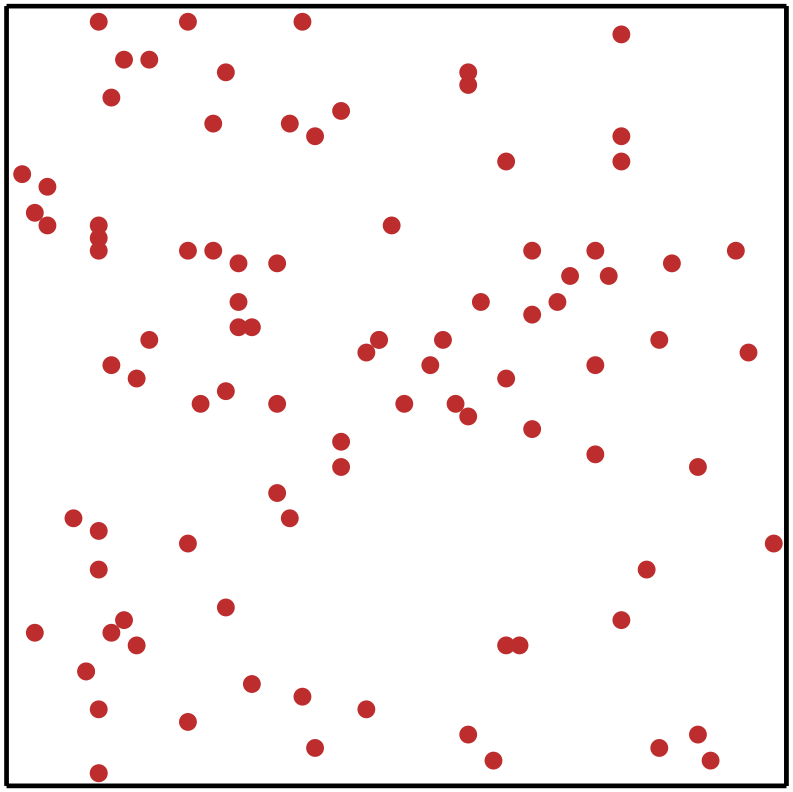
- Each element *i* included with probability p_i:

$$\mathcal{P}(Y) = \prod_{i \in Y} p_i \prod_{i \notin Y} (1 - p_i)$$

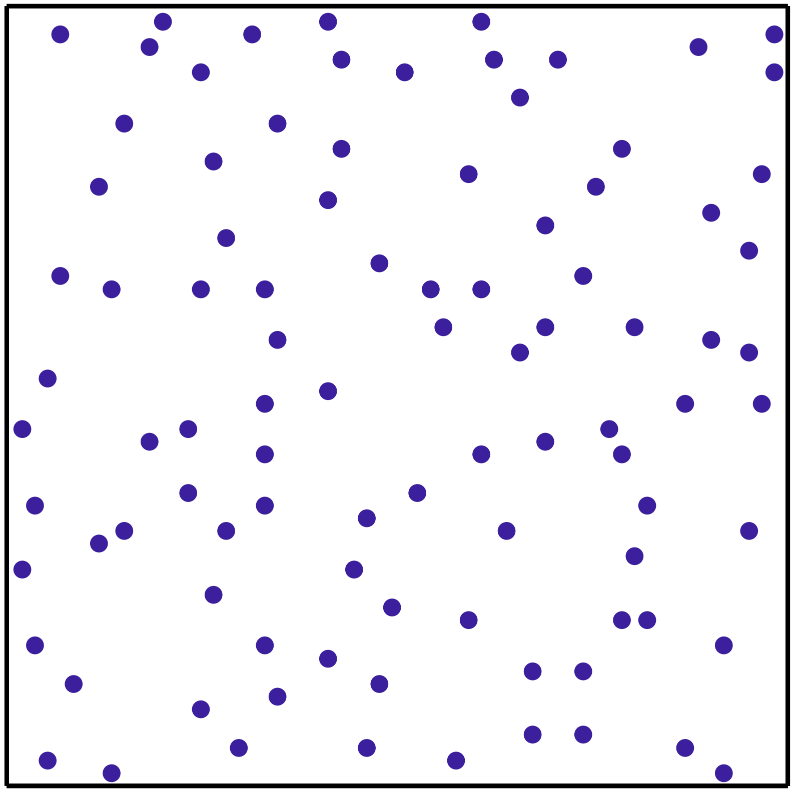
- For example, uniform:



Point process samples



Independent

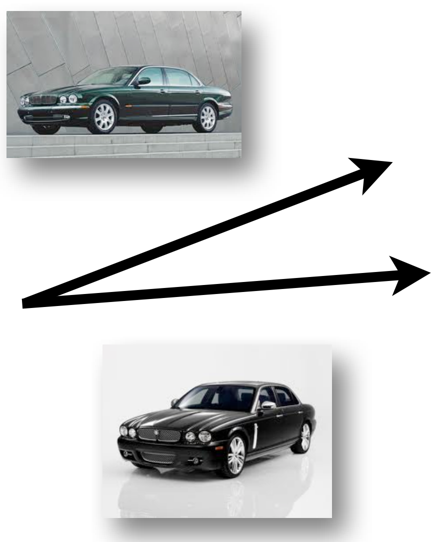


DPP

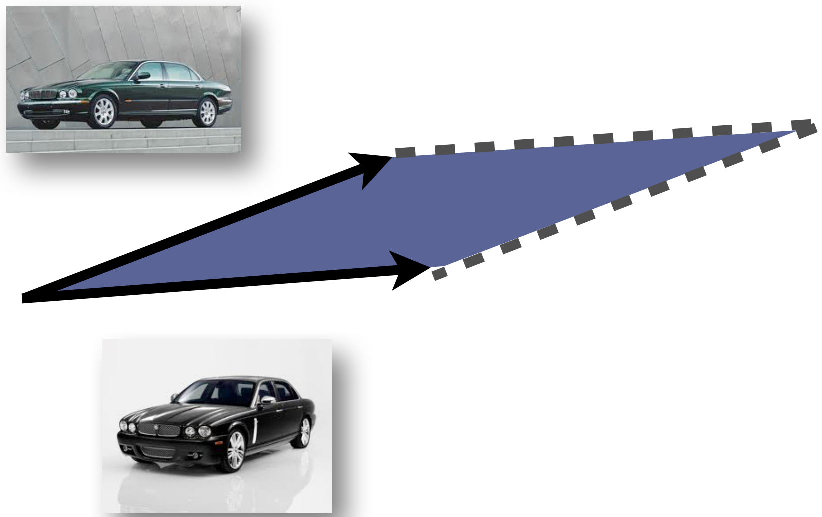
Feature function **g** on items in $\mathcal{Y}$

$\boldsymbol{g}\left(\right.$  $\left.\right)$

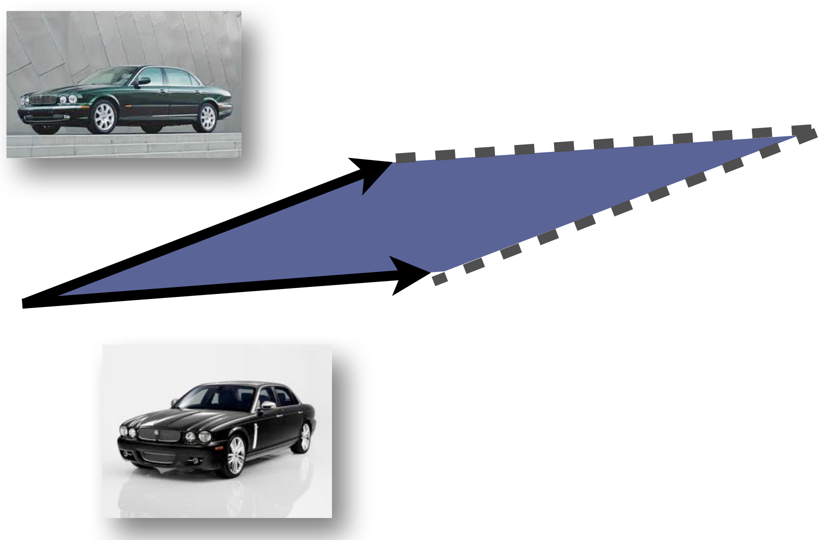
Feature function **g** on items in $\mathcal{Y}$

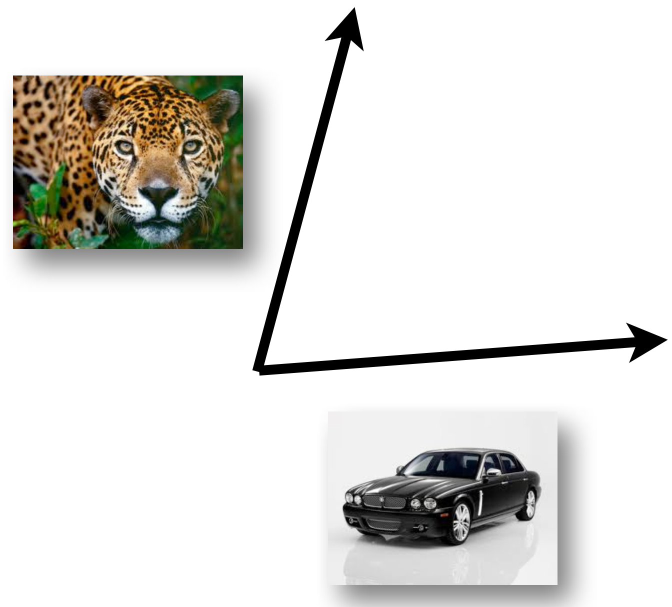


Feature function **g** on items in $\mathcal{Y}$

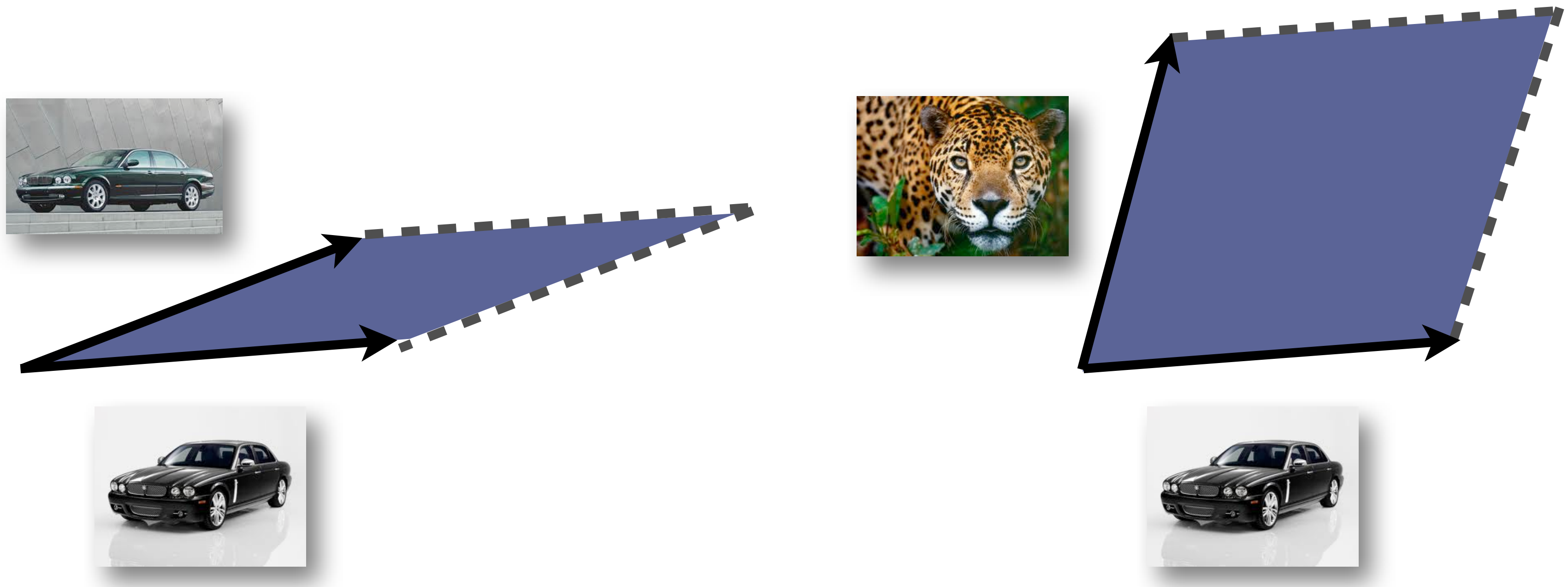


Feature function $\mathbf{g}$ on items in $\mathcal{Y}$





Feature function **g** on items in $\mathcal{Y}$



$$L_{ij} = \boldsymbol{g}(i)^\top \boldsymbol{g}(j)$$

Determinantal point process

$$\mathcal{P}(Y) \propto \det(L_Y)$$

= squared volume spanned by $\boldsymbol{g}(i),\ i \in Y$

[Macchi, 1975]

Determinantal point process

$$\mathcal{P}(Y) \propto \det(L_Y)$$

$$L = \begin{pmatrix} L_{11} & L_{12} & L_{13} & L_{14} \\ L_{21} & L_{22} & L_{23} & L_{24} \\ L_{31} & L_{32} & L_{33} & L_{34} \\ L_{41} & L_{42} & L_{43} & L_{44} \end{pmatrix}$$

[Macchi, 1975]

Determinantal point process

$$\mathcal{P}(Y) \propto \det(L_Y)$$

$$\mathcal{P}(\{2,4\}) \quad \begin{matrix} L_{11} & L_{12} & L_{13} & L_{14} \\ L_{21} & L_{22} & L_{23} & L_{24} \\ L_{31} & L_{32} & L_{33} & L_{34} \\ L_{41} & L_{42} & L_{43} & L_{44} \end{matrix}$$

[Macchi, 1975]

Determinantal point process

$$\mathcal{P}(Y) \propto \det(L_Y)$$

$$\mathcal{P}(\{2,4\}) \quad \begin{matrix} L_{11} & L_{12} & L_{13} & L_{14} \\ L_{21} & L_{22} & L_{23} & L_{24} \\ L_{31} & L_{32} & L_{33} & L_{34} \\ L_{41} & L_{42} & L_{43} & L_{44} \end{matrix}$$

[Macchi, 1975]

Determinantal point process

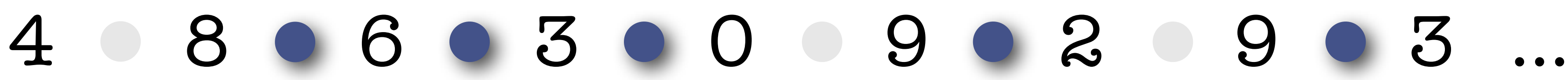
$$\mathcal{P}(Y) \propto \det(L_Y)$$

$$\mathcal{P}(\{2,4\}) \propto \begin{vmatrix} L_{22} & L_{24} \\ L_{42} & L_{44} \end{vmatrix}$$

[Macchi, 1975]

4 8 6 3 0 9 2 9 3 ...

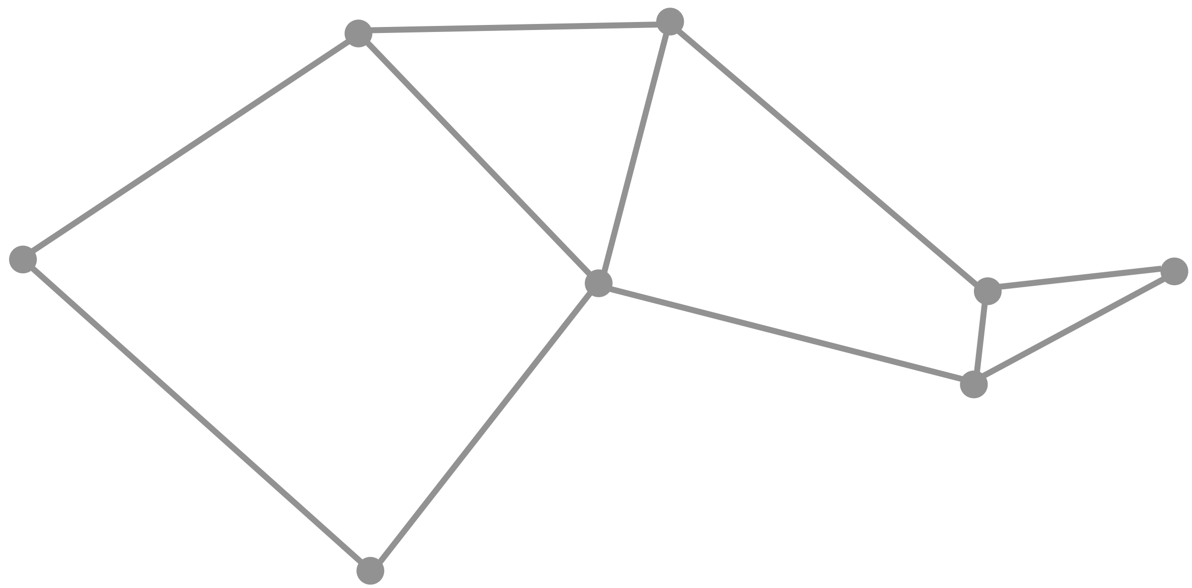
[Borodin et al, 2010]



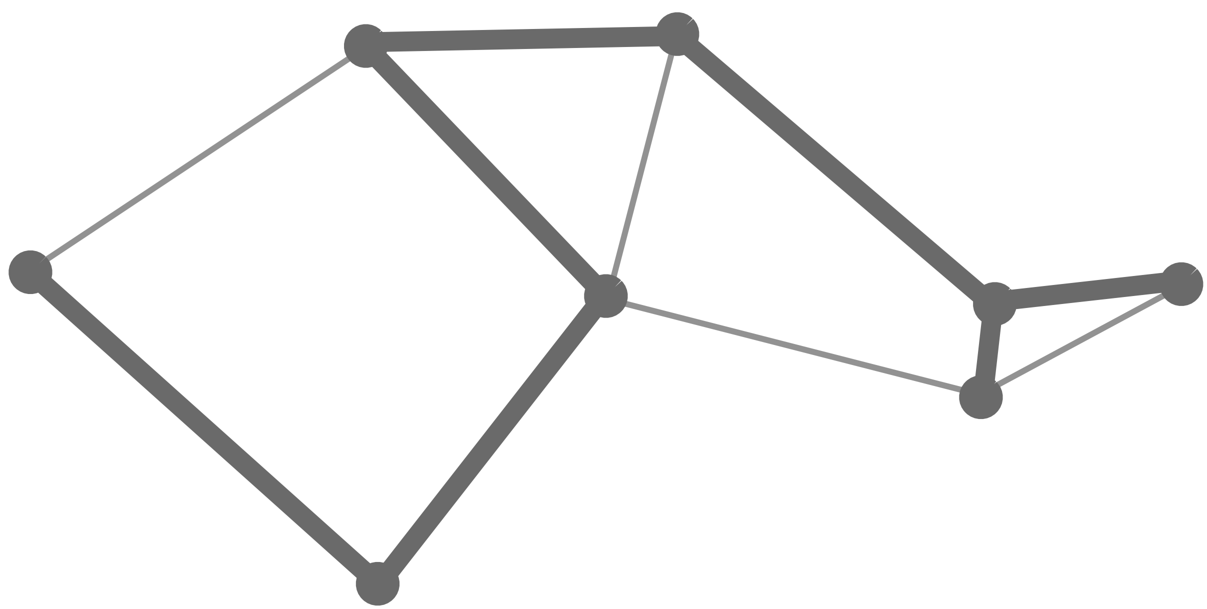
[Borodin et al, 2010]

4 8 6 3 0 9 2 9 3 …

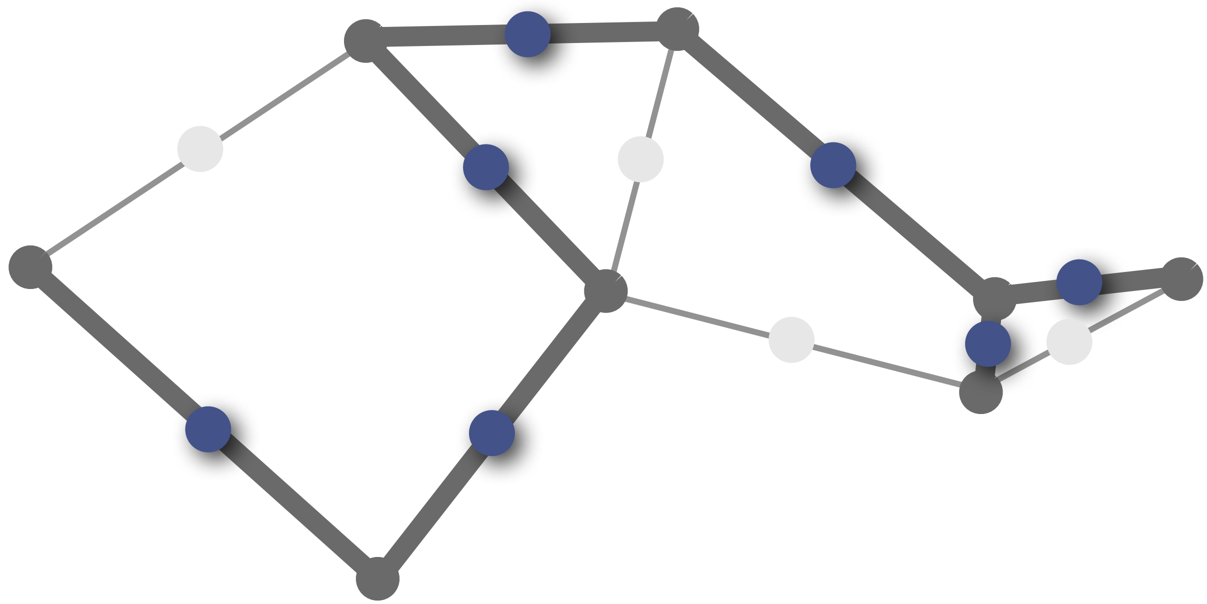
[Borodin et al, 2010]



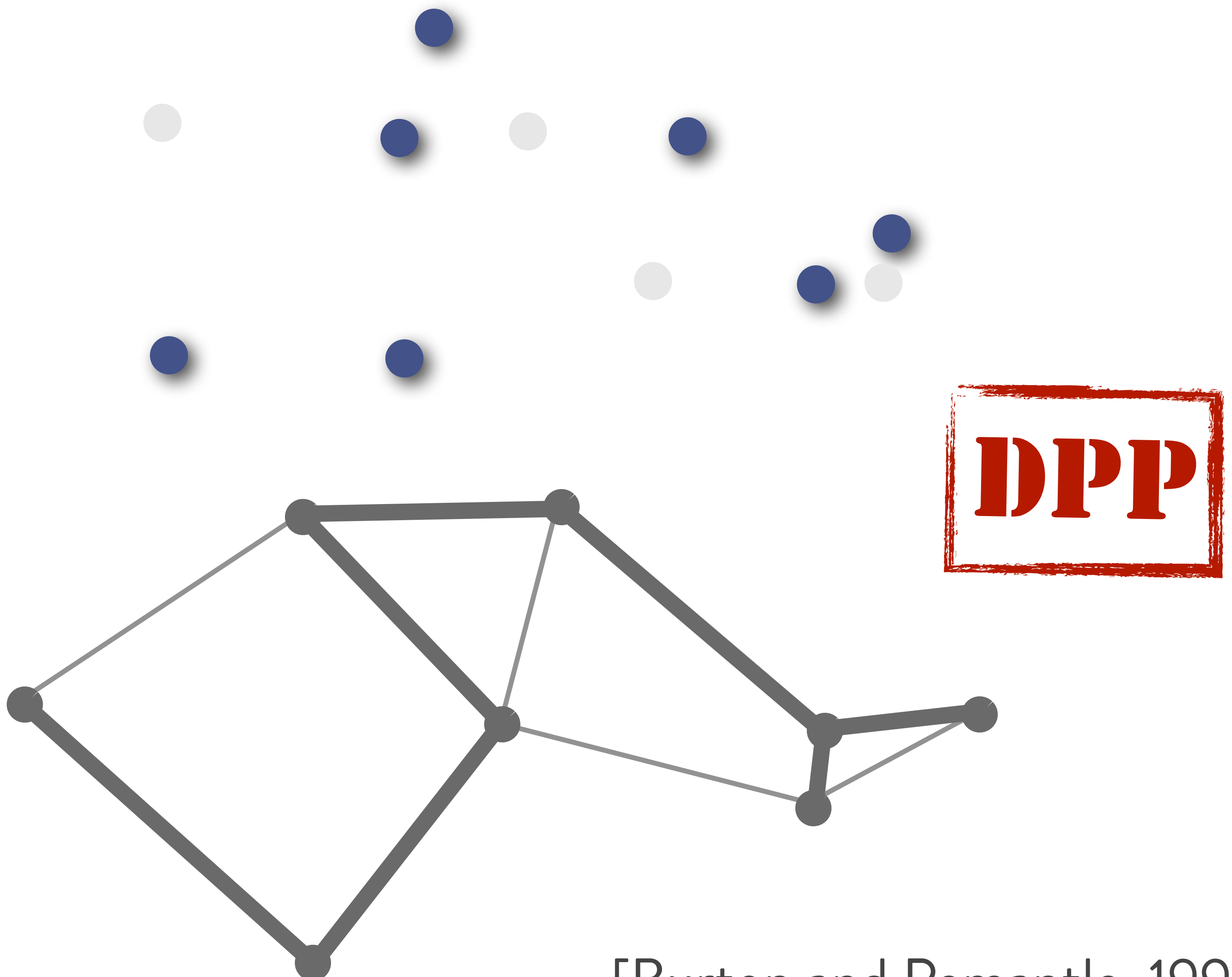
[Burton and Pemantle, 1993]



[Burton and Pemantle, 1993]

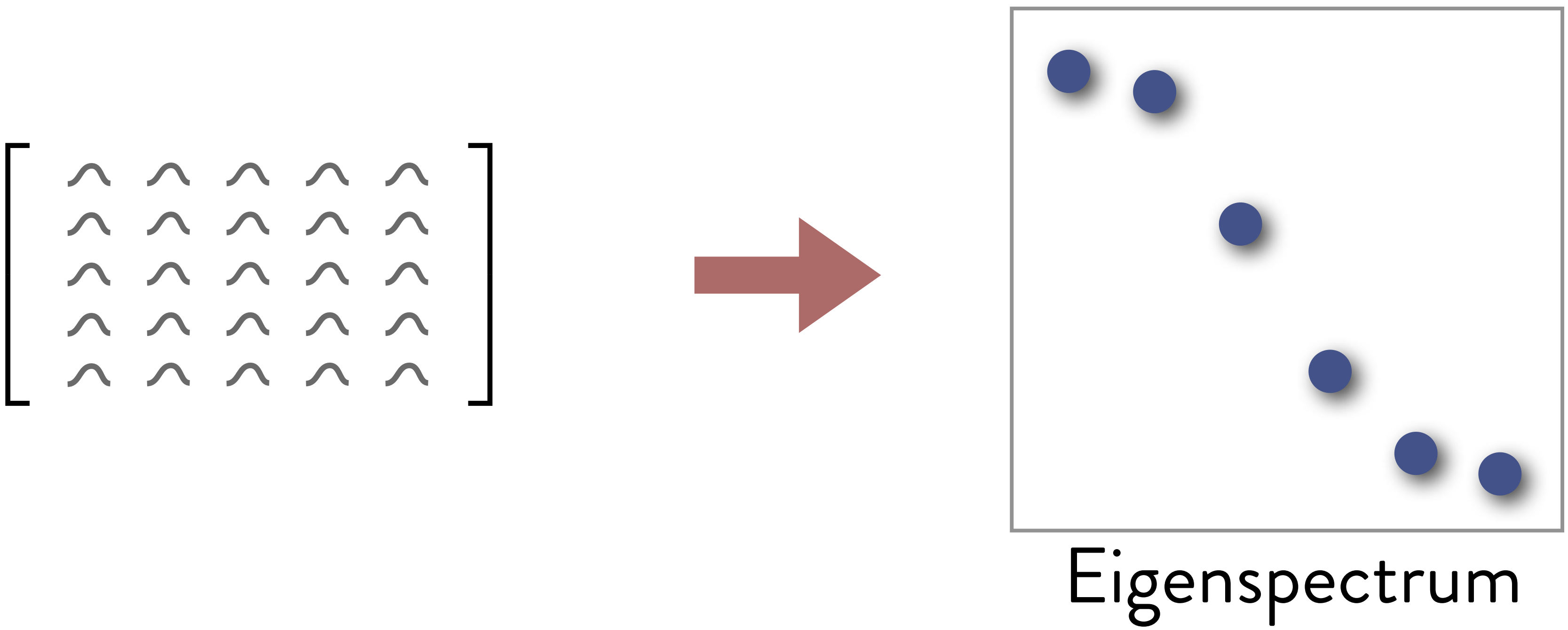


[Burton and Pemantle, 1993]



[Burton and Pemantle, 1993]

[Dyson, 1970]



[Dyson, 1970]

[Dyson, 1970]

Part I

Representation

Inference: Marginals, Conditionals

Inference: Sampling

DPPs vs MRFs

Learning

Inference: normalization

?

$$\mathcal{P}(Y) \propto \det(L_Y)$$

Inference: normalization

$$\mathcal{P}(Y) = \frac{\det(L_Y)}{\det(L + I)}$$

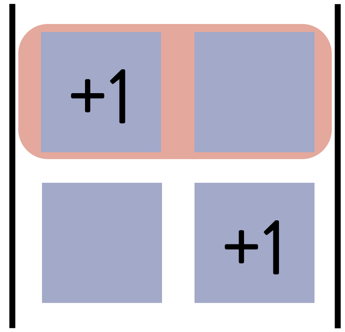
Multilinearity of determinants

$$\begin{vmatrix} — & \alpha R_1 & — \\ — & R_2 & — \\ — & R_3 & — \\ & \vdots & \end{vmatrix} = \alpha \begin{vmatrix} — & R_1 & — \\ — & R_2 & — \\ — & R_3 & — \\ & \vdots & \end{vmatrix}$$

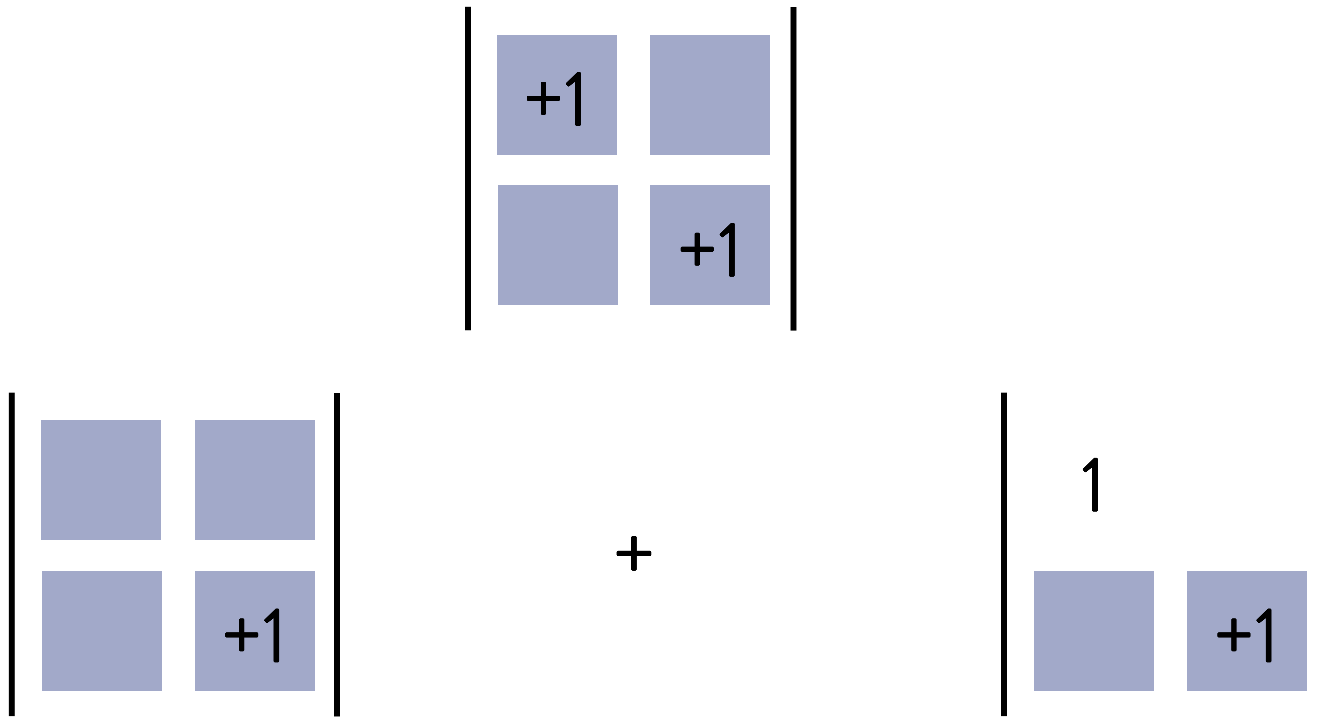
$$\begin{vmatrix} — & R_1 + R_1' & — \\ — & R_2 & — \\ — & R_3 & — \\ & \vdots & \end{vmatrix} = \begin{vmatrix} — & R_1 & — \\ — & R_2 & — \\ — & R_3 & — \\ & \vdots & \end{vmatrix} + \begin{vmatrix} — & R_1' & — \\ — & R_2 & — \\ — & R_3 & — \\ & \vdots & \end{vmatrix}$$

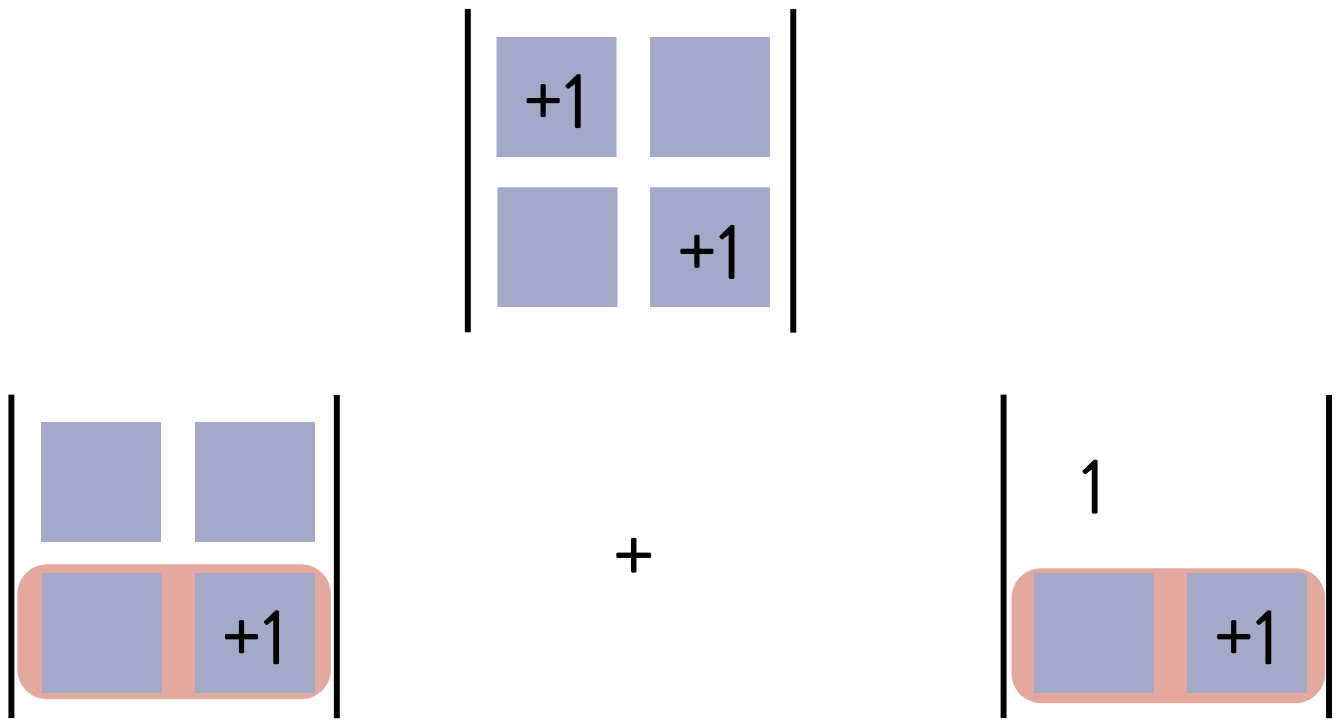
L+I

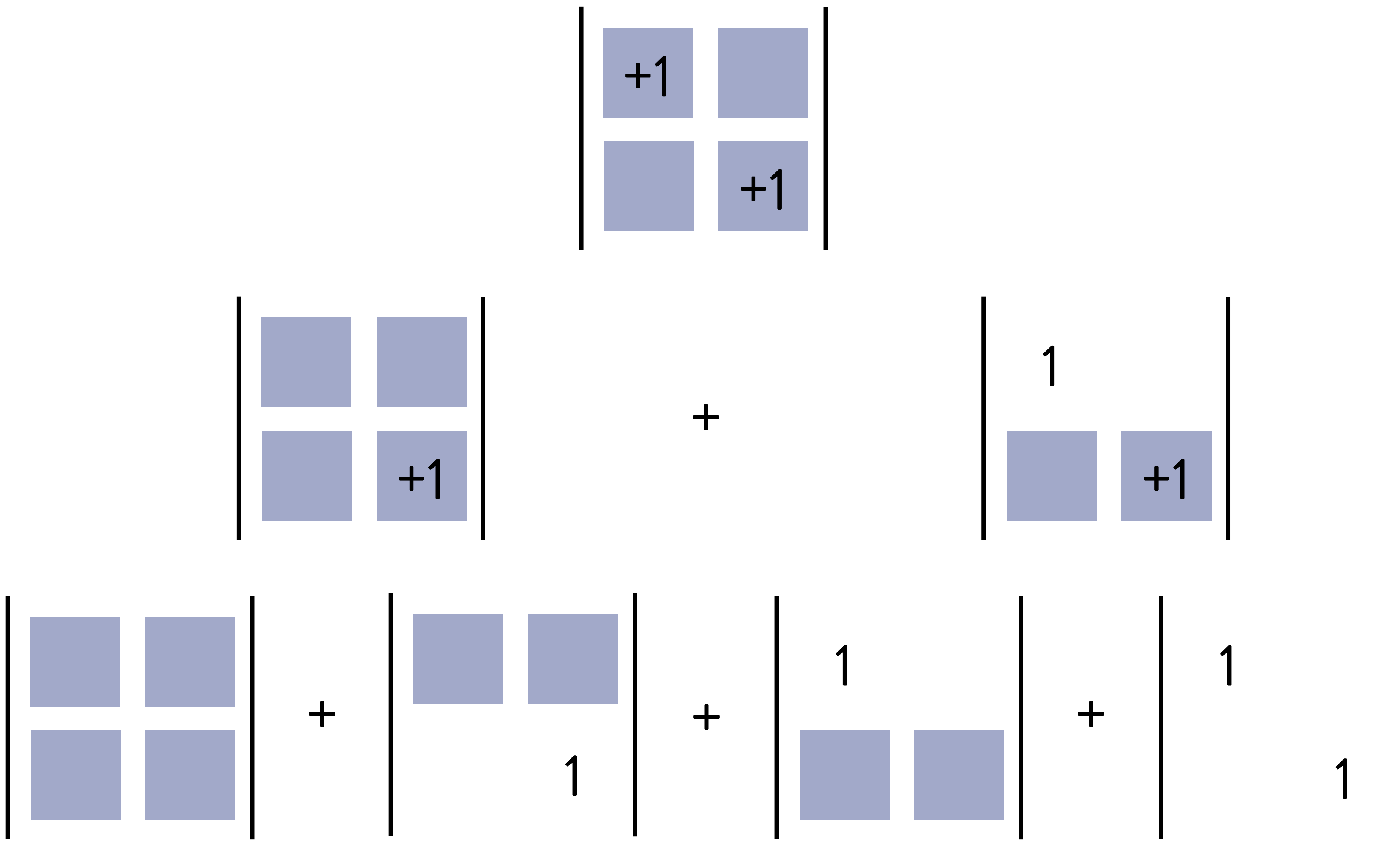
$$\begin{vmatrix} +1 & \\ & +1 \end{vmatrix}$$

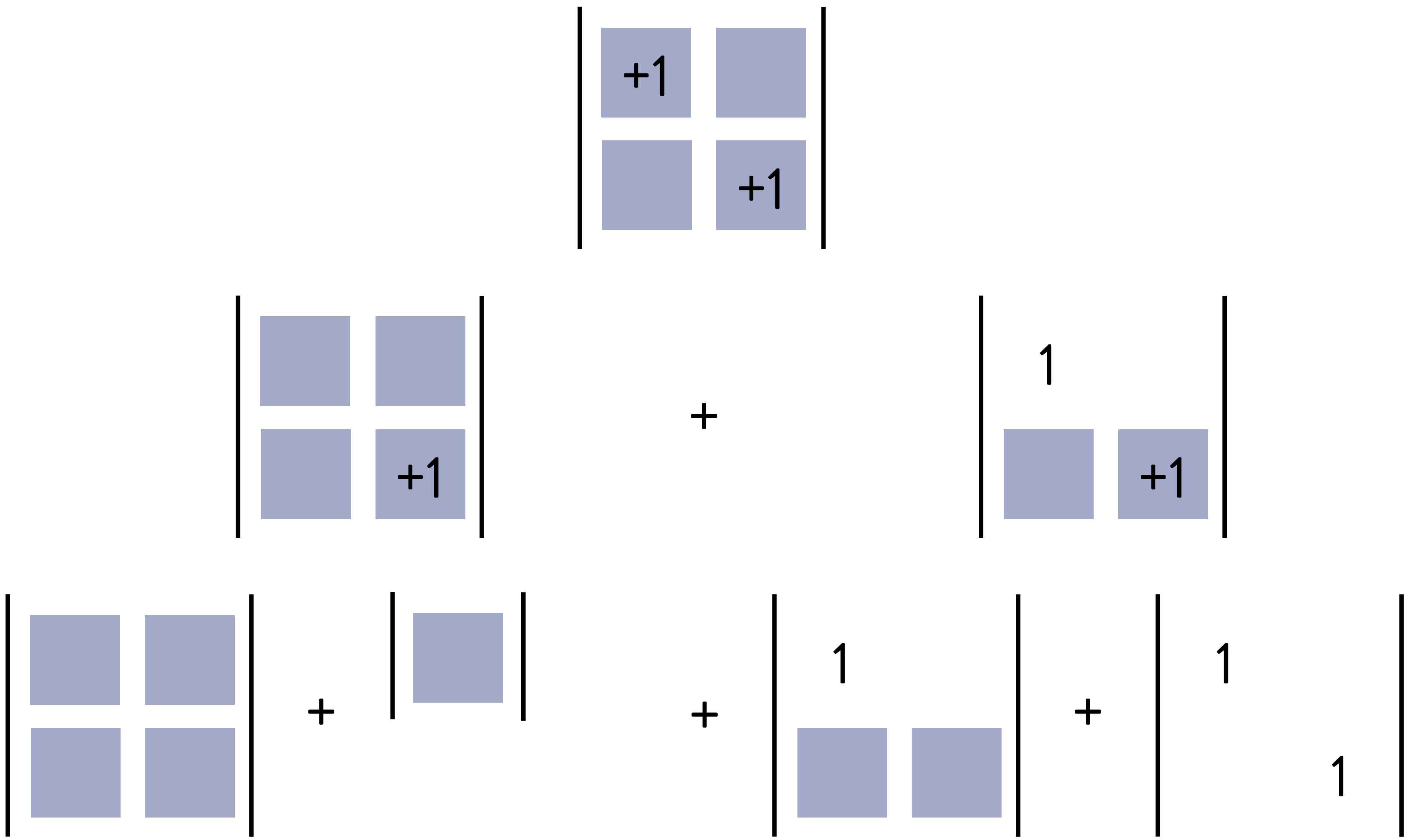


+1
+1

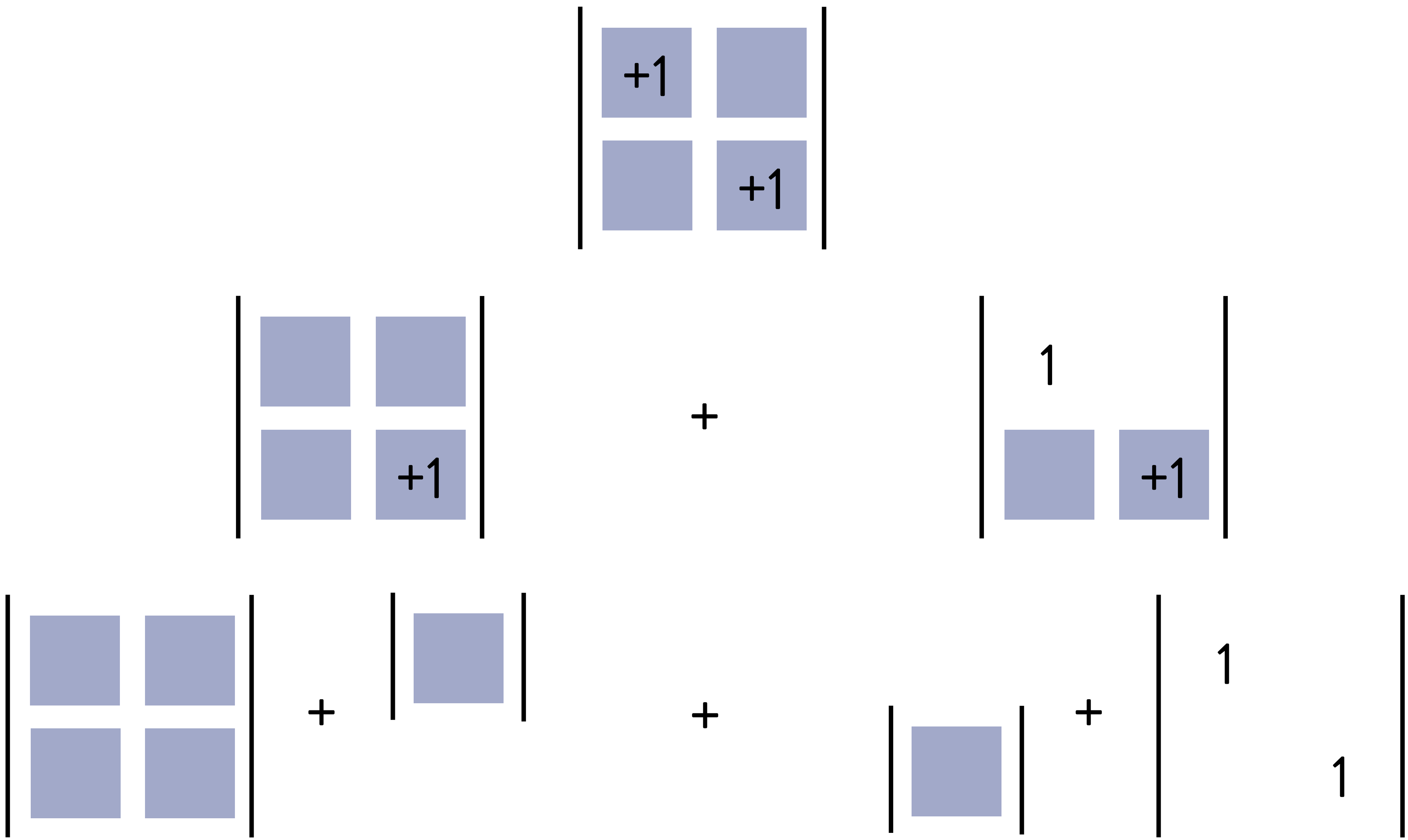

+1
+1
+1
+
1
+1

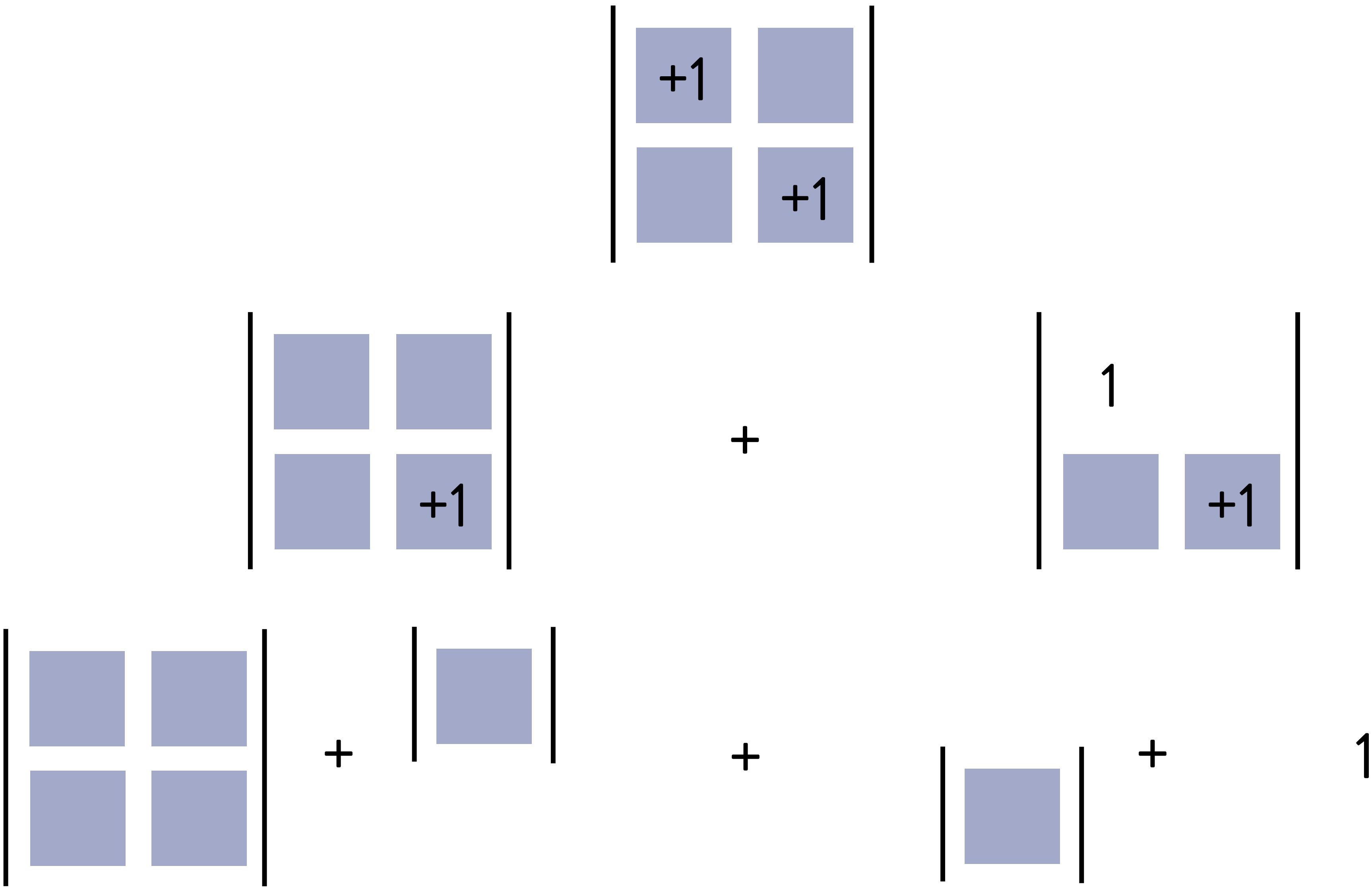

+1
+1
+1
+
1
+1

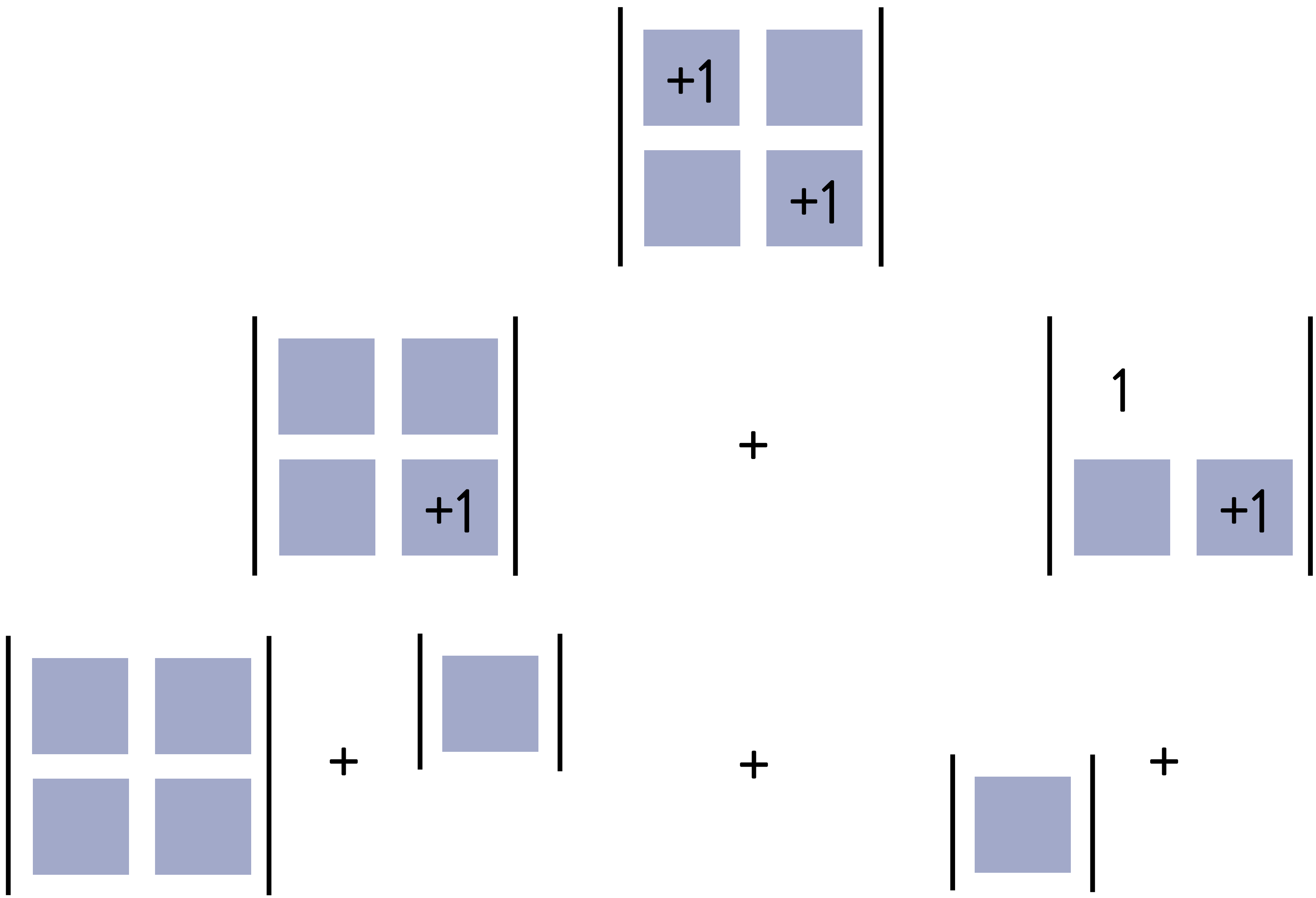

+1
+1
+1
+
1
+1
+
1
+
1
+
1
1



+1
+1
+1
+
1
+1
+
+
1
+
1
1

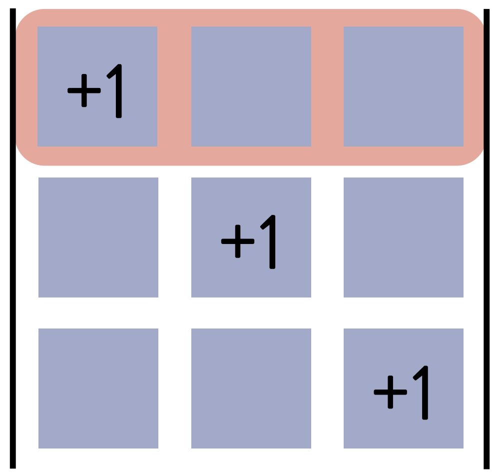

+1
+1
+1
+
1
+1
+
+
+
1
1


+1
+1
+1
+
1
+1
+
+
+
1



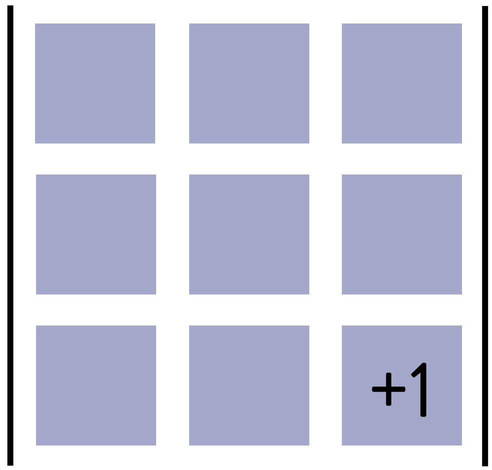
$\propto \mathcal{P}(\{1,2\}) \quad \mathcal{P}(\{1\}) \quad \mathcal{P}(\{2\}) \quad \mathcal{P}(\varnothing)$

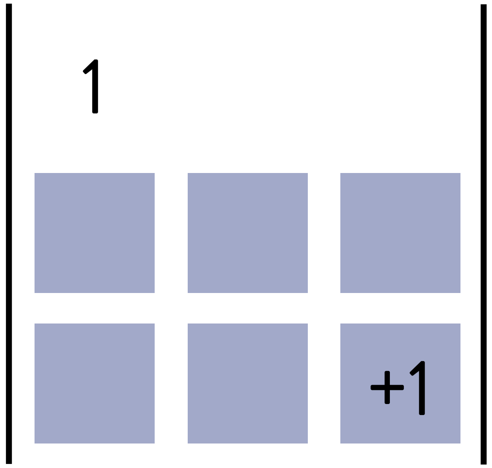
$$\begin{vmatrix} +1 & & \\ & +1 & \\ & & +1 \end{vmatrix}$$

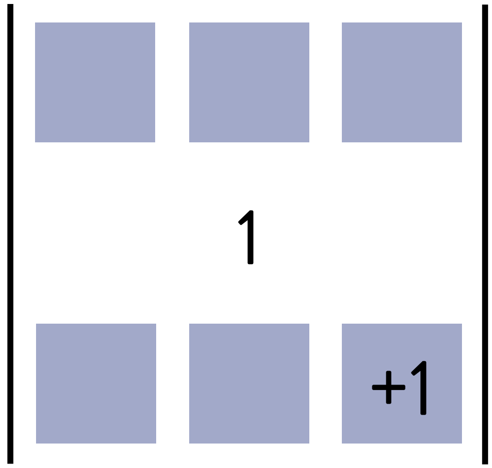

+1
+1
+1

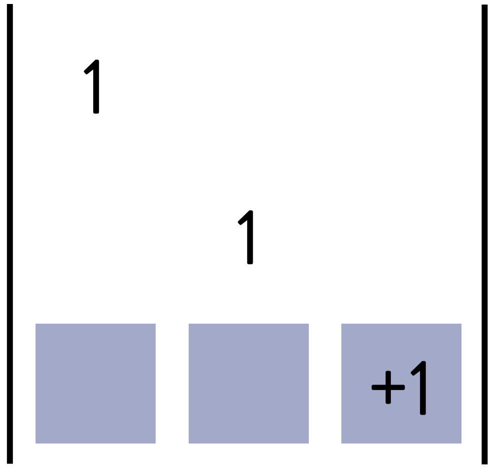

+1
+1
1
+1
+1

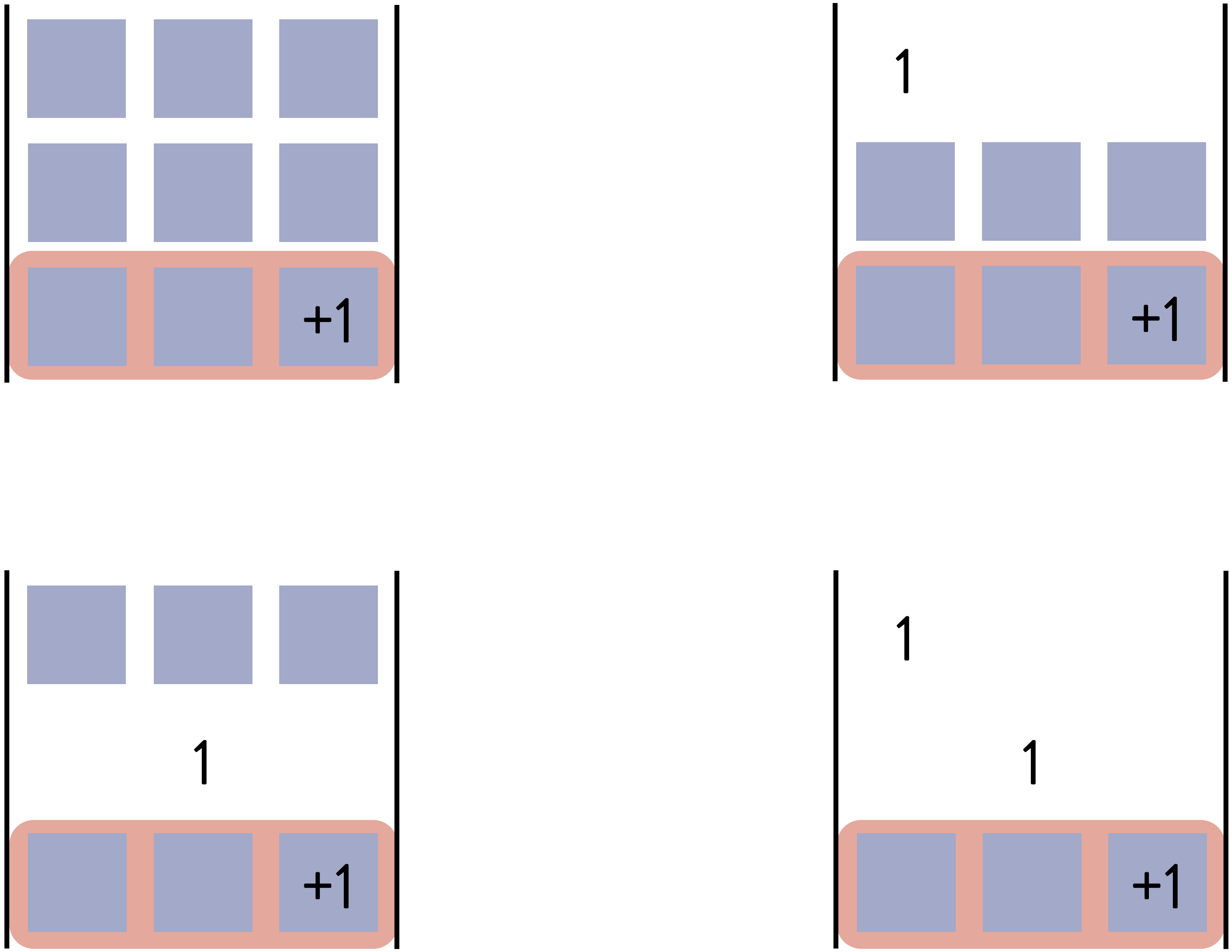

+1
+1
1
+1
+1

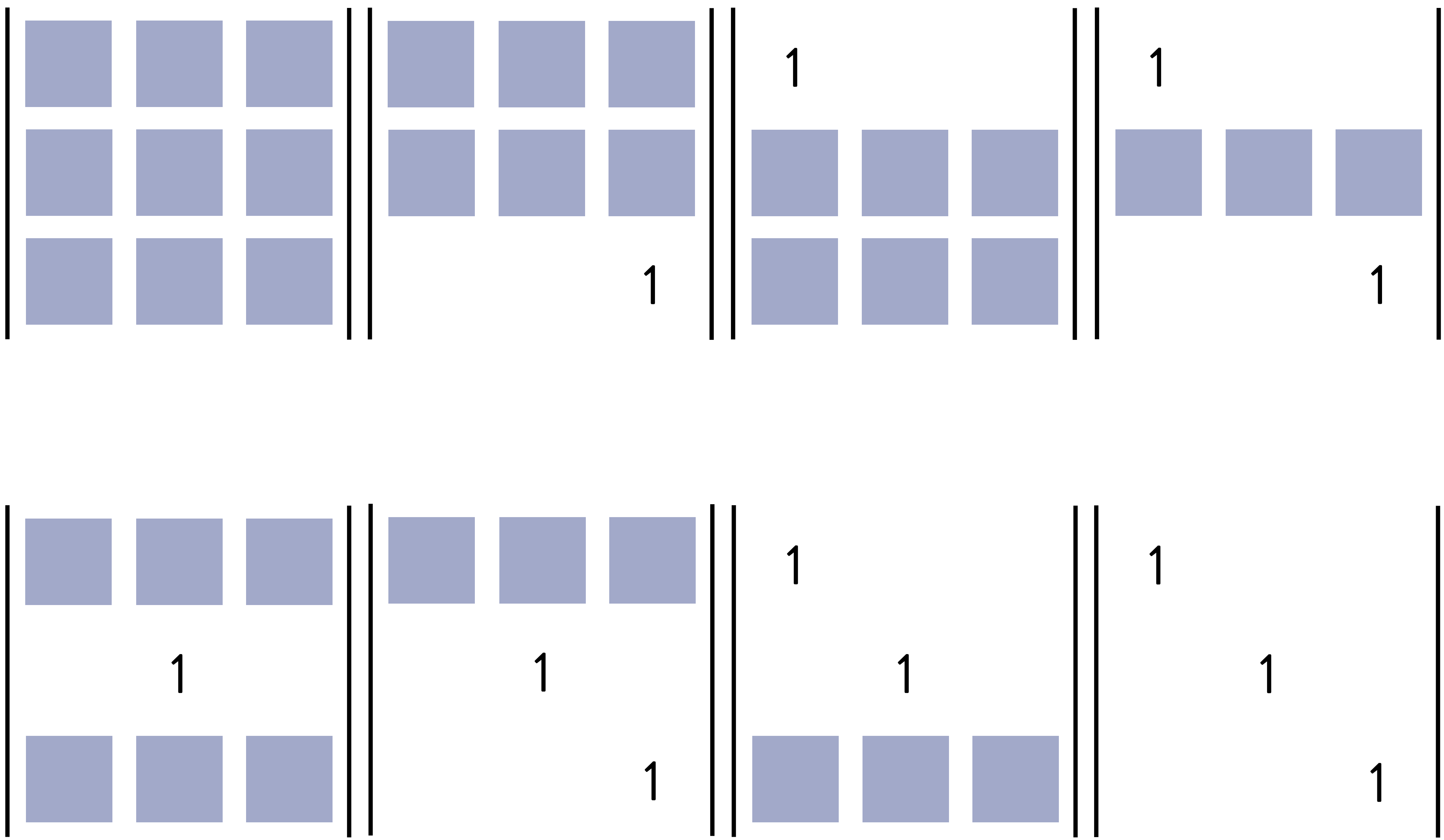

+1

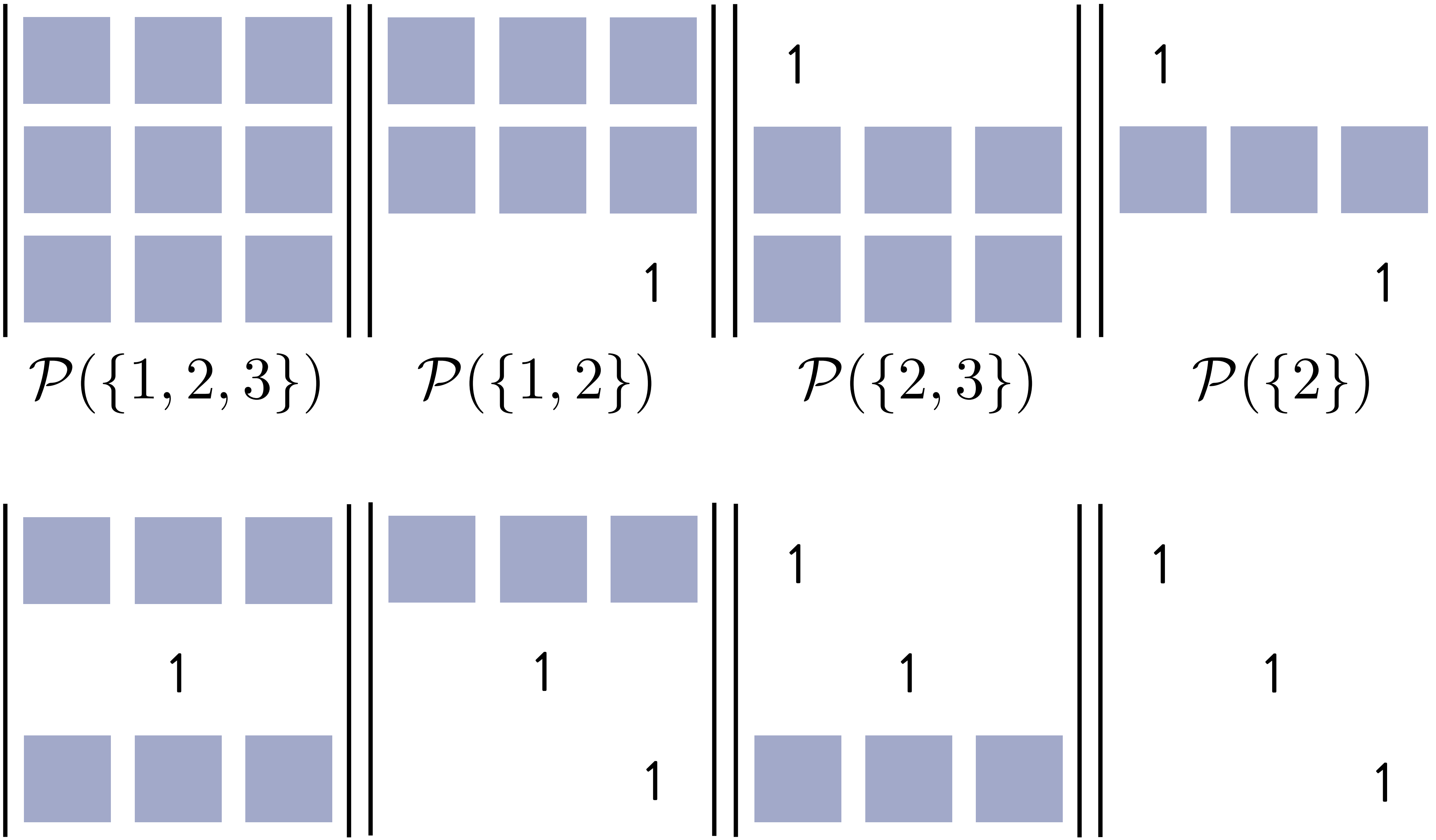

1
+1


1
+1


1
1
+1


+1
1
+1
1
+1
1
1
+1





$\mathcal{P}(\{1,2,3\})$ $\mathcal{P}(\{1,2\})$ $\mathcal{P}(\{2,3\})$ $\mathcal{P}(\{2\})$

$\mathcal{P}(\{1,3\})$ $\mathcal{P}(\{1\})$ $\mathcal{P}(\{3\})$ $\mathcal{P}(\varnothing)$

Inference: marginals

$$\mathcal{P}(A \subseteq \boldsymbol{Y}) = \det(K_A)$$

$$K = L(L + I)^{-1}$$

$$\mathcal{P}(A \subseteq \boldsymbol{Y}) = \det(K_A)$$

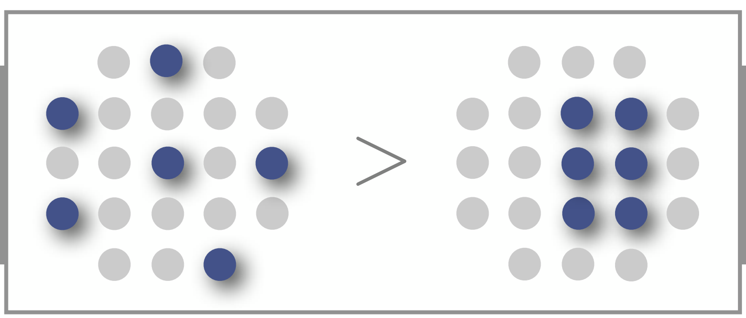
$$\mathcal{P}(i \in \boldsymbol{Y}) = \det(K_{ii}) = K_{ii}$$

$$\mathbb{E}[|\boldsymbol{Y}|] = \sum_i \mathcal{P}(i \in \boldsymbol{Y}) = \text{trace}(K)$$

$$\mathcal{P}(A \subseteq \boldsymbol{Y}) = \det(K_A)$$

$$\mathcal{P}(i \in \boldsymbol{Y}) = \det(K_{ii}) = K_{ii}$$

$$\begin{aligned}\mathcal{P}(i, j \in \boldsymbol{Y}) &= \det \begin{pmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{pmatrix} \\ &= K_{ii}K_{jj} - K_{ij}K_{ji} \\ &= \mathcal{P}(i \in \boldsymbol{Y})\mathcal{P}(j \in \boldsymbol{Y}) - K_{ij}^2\end{aligned}$$



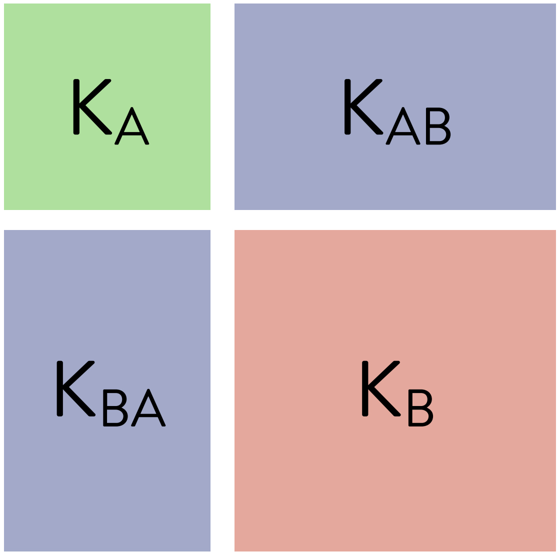
Diversity

Inference: conditioning

$$\mathcal{P}(B \subseteq \boldsymbol{Y} | A \subseteq \boldsymbol{Y}) = ?$$

Inference: conditioning

$$K_{A \cup B} =$$



Schur complement:

$$\det(K_{A \cup B}) = \det(K_A)\det(K_B - K_{BA}K_A^{-1}K_{AB})$$

Inference: conditioning

$$\det(K_{A\cup B}) = \det(K_A)\det(K_B - K_{BA}K_A^{-1}K_{AB})$$

$$\mathcal{P}(B \subseteq \boldsymbol{Y} | A \subseteq \boldsymbol{Y}) = \frac{\mathcal{P}(A \cup B \subseteq \boldsymbol{Y})}{\mathcal{P}(A \subseteq \boldsymbol{Y})}$$

$$= \frac{\det(K_{A\cup B})}{\det(K_A)}$$

$$= \det(K_B - K_{BA}K_A^{-1}K_{AB})$$

Inference: conditioning

$$\mathcal{P}(B \subseteq \boldsymbol{Y} | A \subseteq \boldsymbol{Y}) = \det(K_B - K_{BA} K_A^{-1} K_{AB})$$

$$= \det(\left[K - K_{*A} K_A^{-1} K_{A*}\right]_B)$$

DPPs closed under conditioning

Part I

Representation

Inference: Marginals, Conditionals

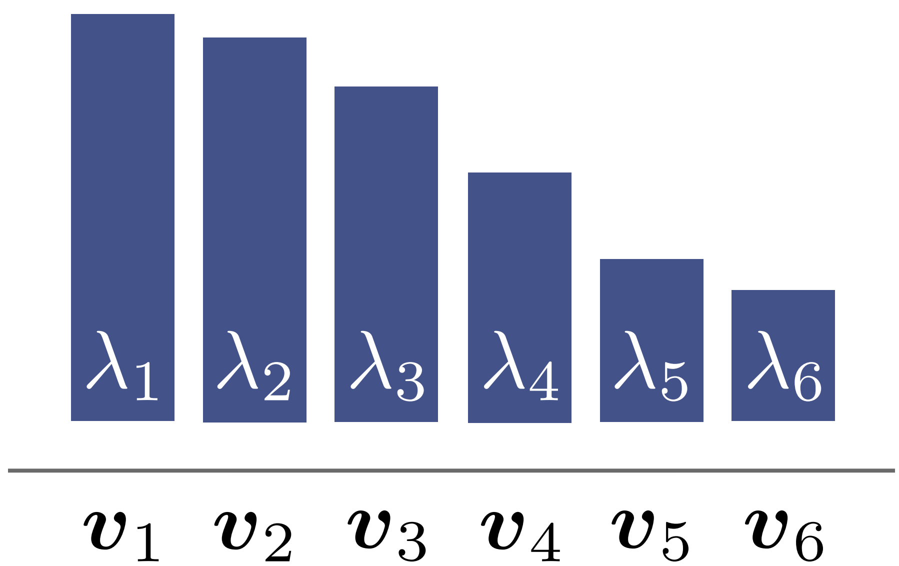
Inference: Sampling

DPPs vs MRFs

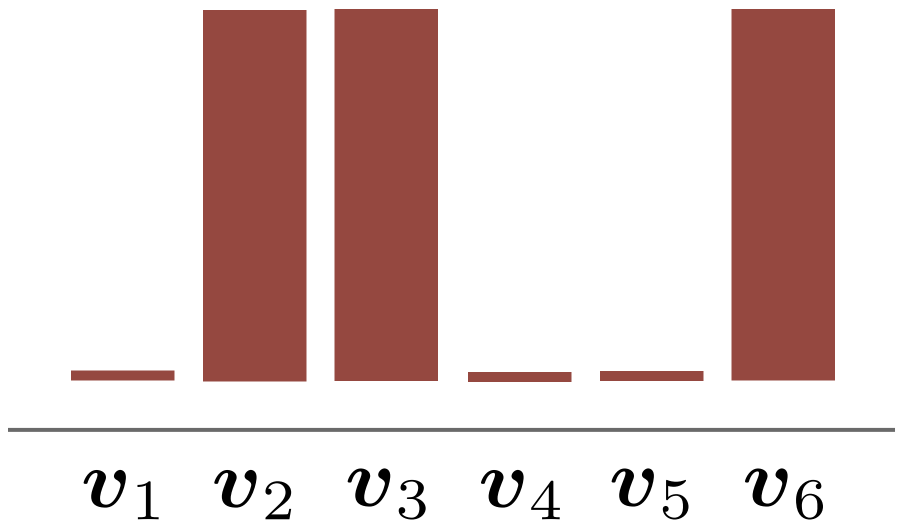
Learning

Eigendecomposition

$$K = \sum_{n=1}^{N} \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^\top$$



Elementary DPP $\mathcal{P}^{\{2,3,6\}}$



- $\mathcal{P}^J$ only supported on sets of size $|J|$
- Exact sampling in $O(|J|^2 N)$

Elementary DPPs

- The marginal kernel of P^J is $K^J = \sum_{n \in J} \boldsymbol{v}_n \boldsymbol{v}_n^\top$
- Expected size $\mathbb{E}[|\boldsymbol{Y}|] = trace(K^J) = \sum_{n \in J} ||\boldsymbol{v}_j||^2 = |J|$
- Since $rank(K^J) = |J|$, $Pr(|\boldsymbol{Y}| > |J|) = 0$
- Hence $Pr(|\boldsymbol{Y}| = |J|) = 1$

Key insight

Every DPP is a "factored" mixture of its elementary DPPs:

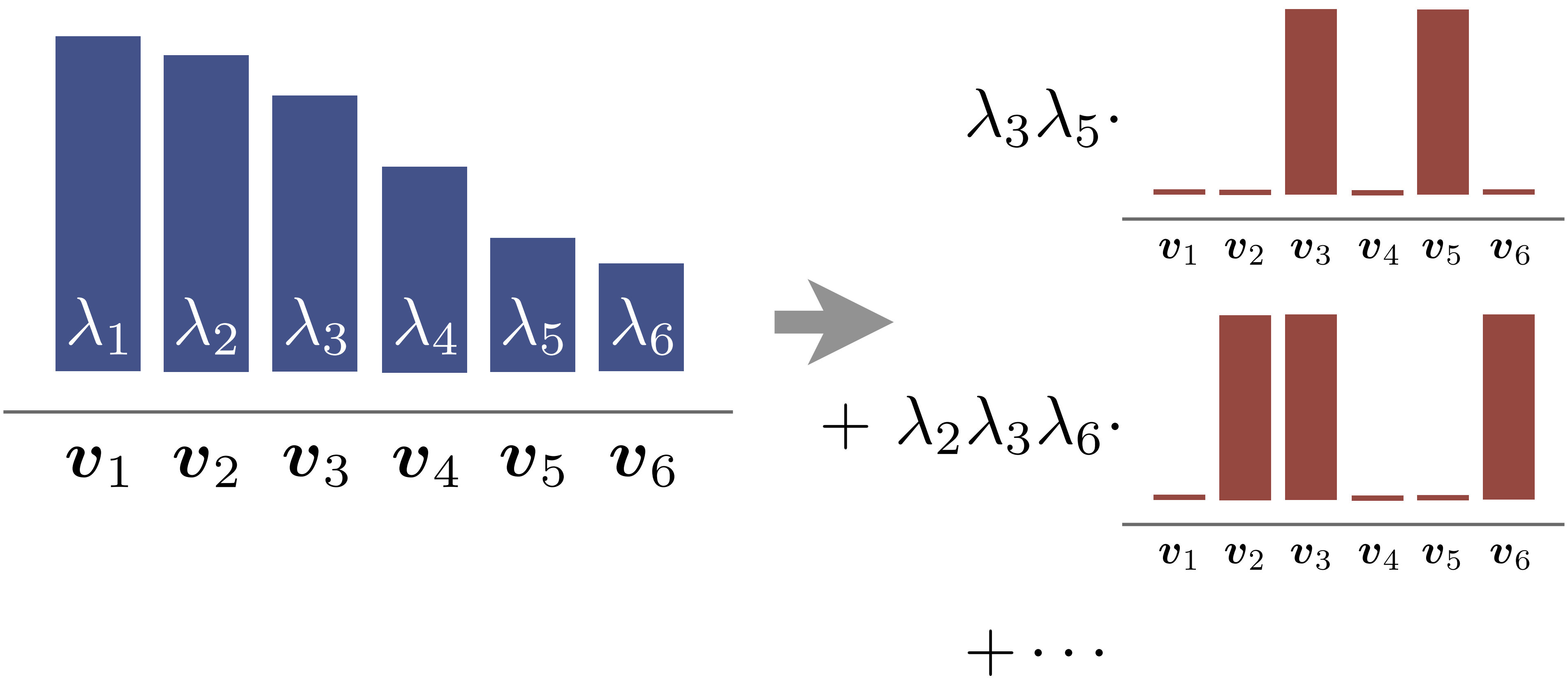
$$\mathcal{P} \propto \sum_{J \subseteq \{1,\ldots,N\}} \mathcal{P}^J \prod_{n \in J} \lambda_n$$

mixture weight

[Hough et al, 2006]

$$\mathcal{P} \propto \sum_{J \subseteq \{1,\ldots,N\}} \mathcal{P}^J \prod_{n \in J} \lambda_n$$

mixture weight



Sampling algorithm

PHASE ONE

Choose elementary DPP $\mathcal{P}^J$ by mixture weight:

$$\Pr(J) \propto \prod_{n \in J} \lambda_n$$

PHASE TWO

Draw sample from $\mathcal{P}^J$

PHASE ONE

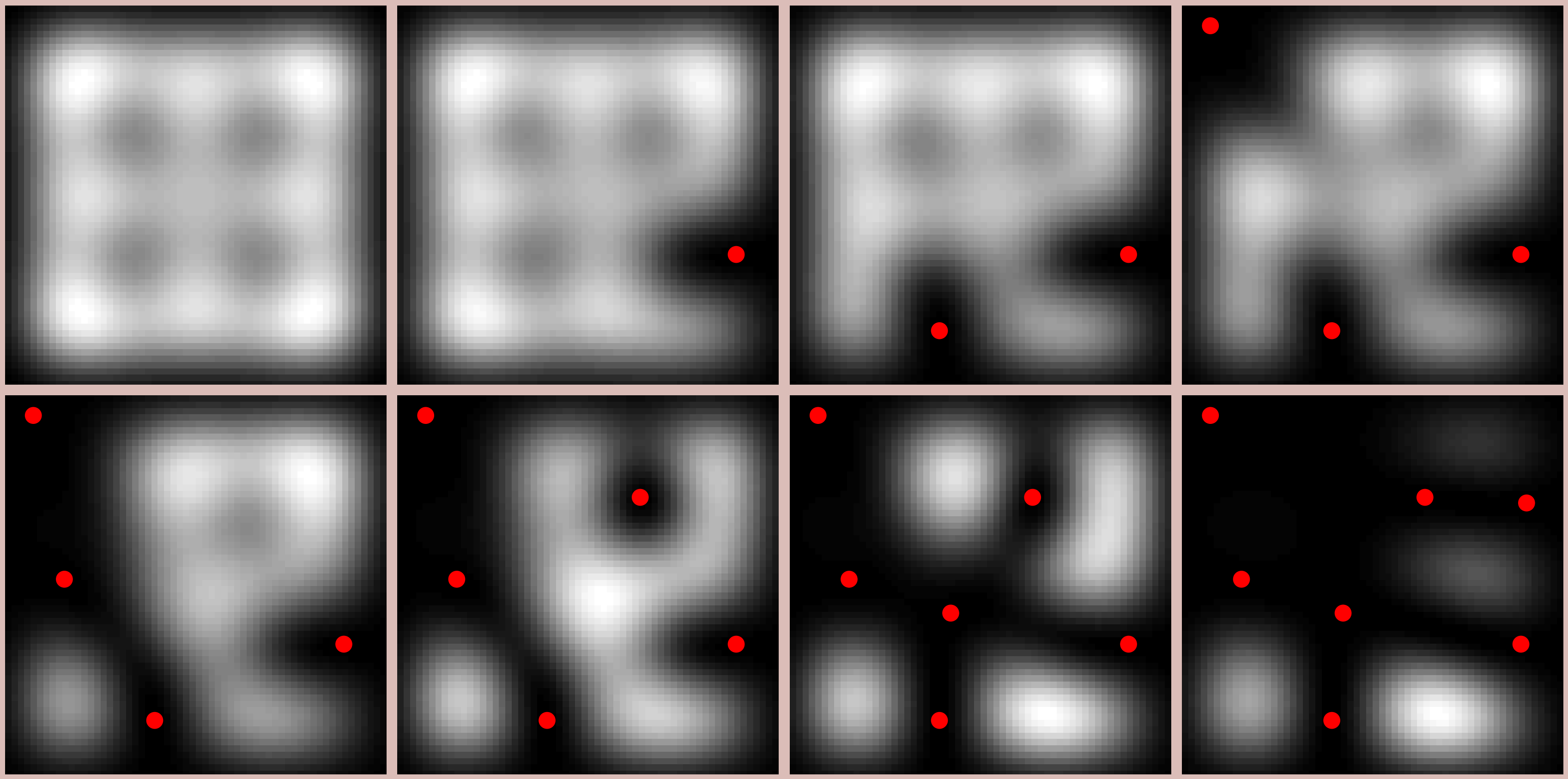
Choose elementary DPP $\mathcal{P}^J$ by mixture weight:

$$\Pr(J) \propto \prod_{n \in J} \lambda_n$$

- Let $J = \varnothing$
- For $n = 1, 2, \ldots, N$
  - $J \leftarrow J \cup \{n\}$ with probability $\frac{\lambda_n}{\lambda_n + 1}$

Phase two

Draw sample from $\mathcal{P}^J$



Phase two

Draw sample from $\mathcal{P}^J$

- Let $Y = \varnothing$, K is the kernel of $\mathcal{P}^J$
- For 1 to $|J|$
  - Choose i with probability $\propto K_{ii}$
  - $Y \leftarrow Y \cup \{i\}$
  - Update K to condition on event $i \in \boldsymbol{Y}$

PHASE TWO

Draw sample from $\mathcal{P}^J$

- Let $Y = \varnothing$, K is the kernel of $\mathcal{P}^J$
- For 1 to $|J|$
  - Choose i with
  - $Y \leftarrow Y \cup \{i\}$
  - Update K to condition on event $i \in \boldsymbol{Y}$

Could be expensive!
But with lazy eval, $O(|J|^2 N)$.

Consequences

- Phase one determines:
  - **Size** of sample ($|J|$)
  - Likely **content** of sample (eigenvectors)

➡ **Size** and **content** are tied

➡ **Size** is sum of Bernoulli variables

Part I

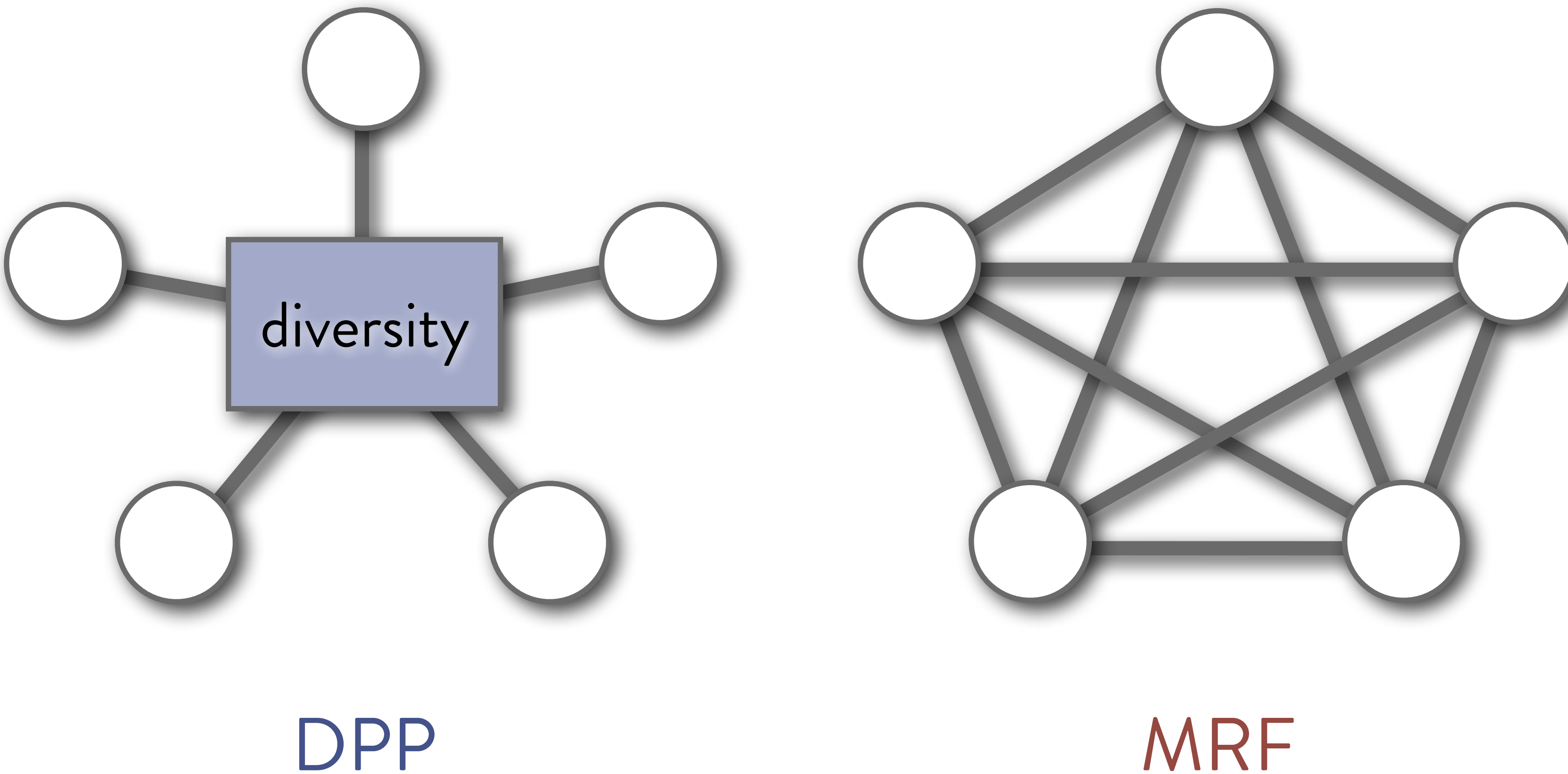
Representation

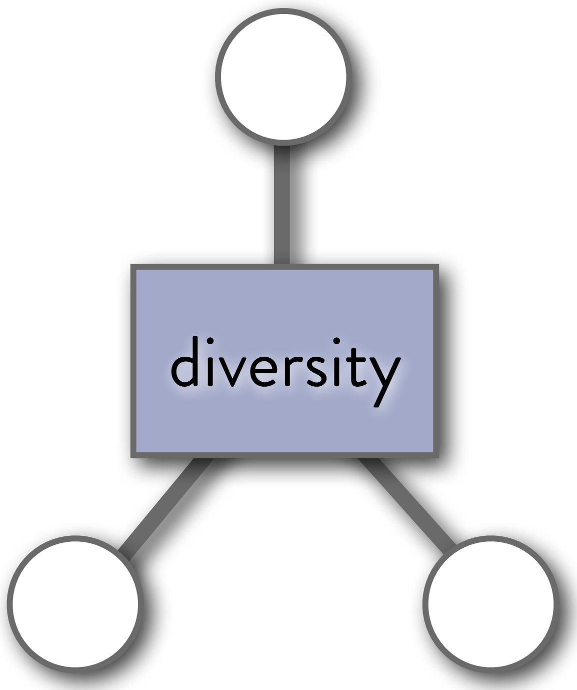
Inference: Marginals, Conditionals

Inference: Sampling

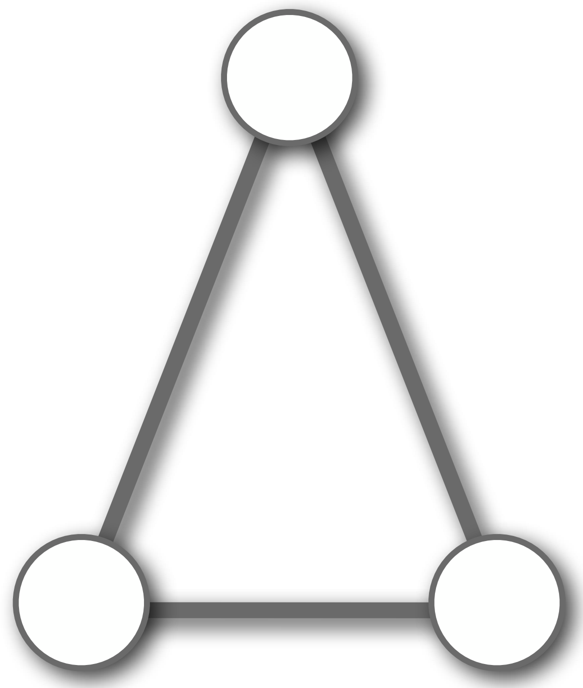
DPPs vs MRFs

Learning

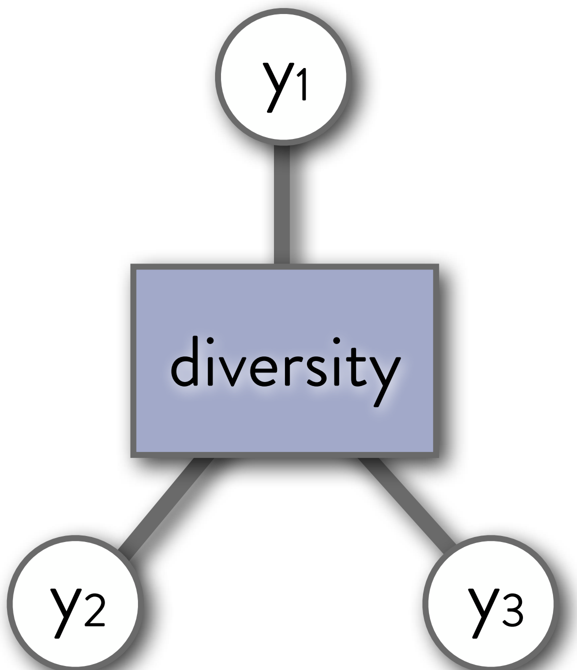

diversity
DPP
MRF



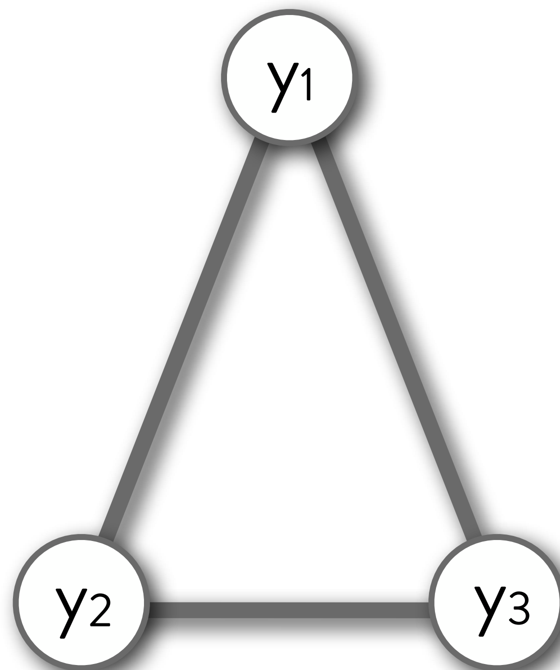
DPP



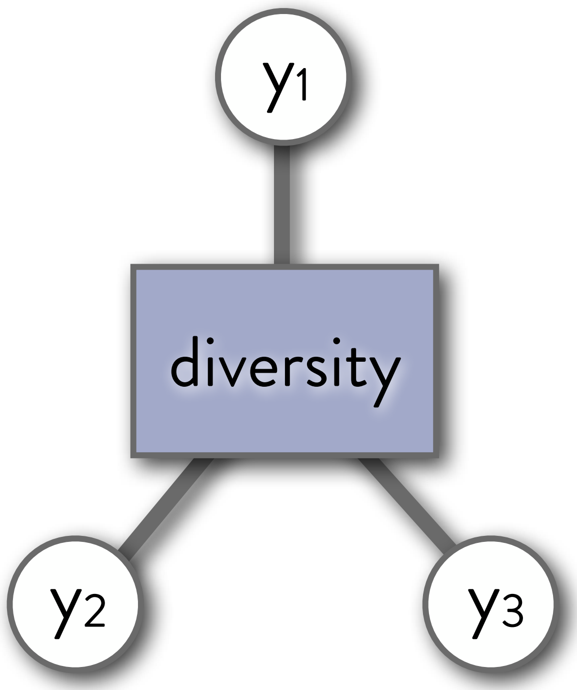
MRF



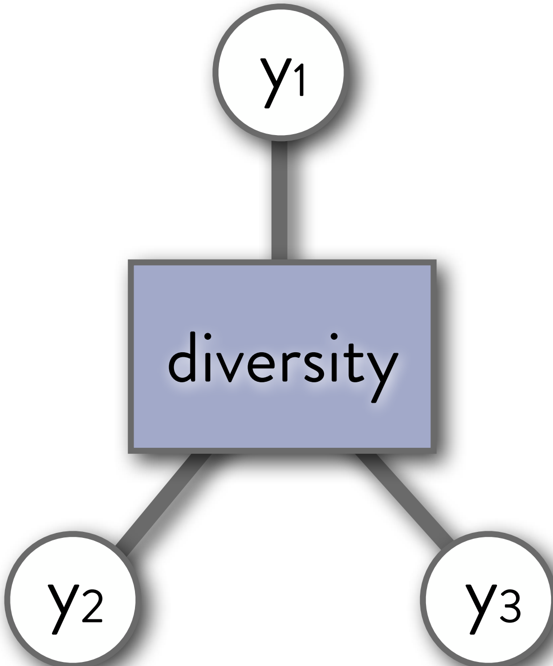
DPP



MRF



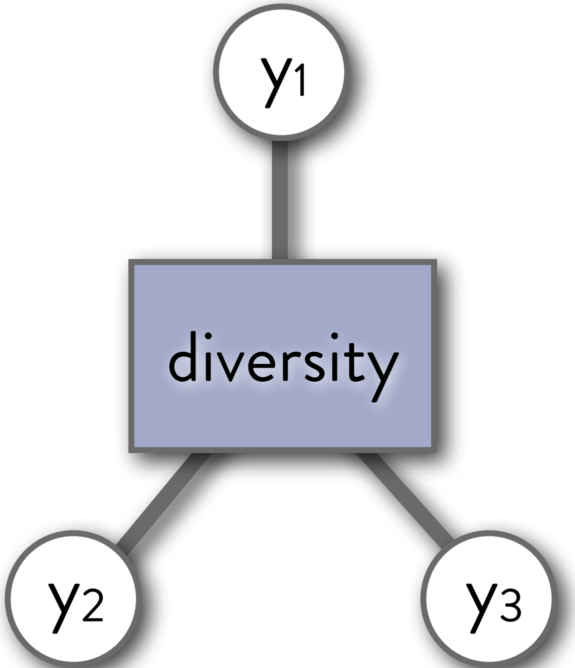
DPP



DPP

$$\mathcal{P}(Y) \propto \det(L_Y)$$

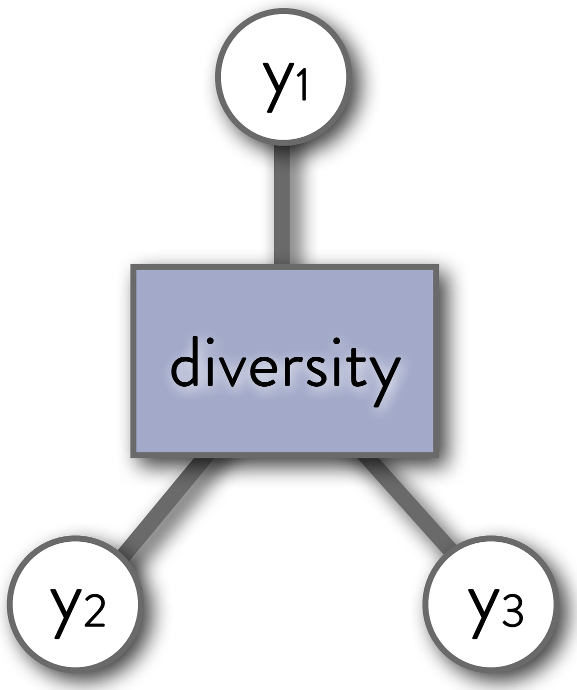
$$\begin{matrix} L_{11} & L_{12} & L_{13} \\ L_{21} & L_{22} & L_{23} \\ L_{31} & L_{32} & L_{33} \end{matrix}$$



DPP

$$\mathcal{P}(Y) \propto \det(L_Y)$$

$$\begin{matrix} L_{11} & L_{12} & L_{13} \\ & L_{22} & L_{23} \\ & & L_{33} \end{matrix}$$

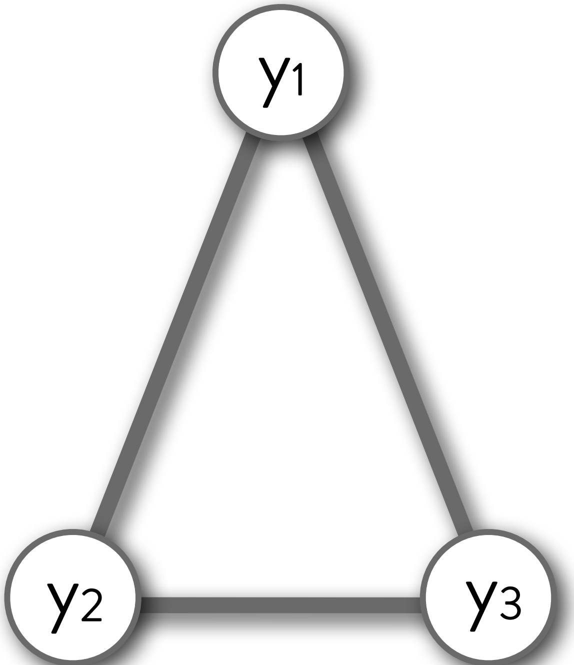


DPP

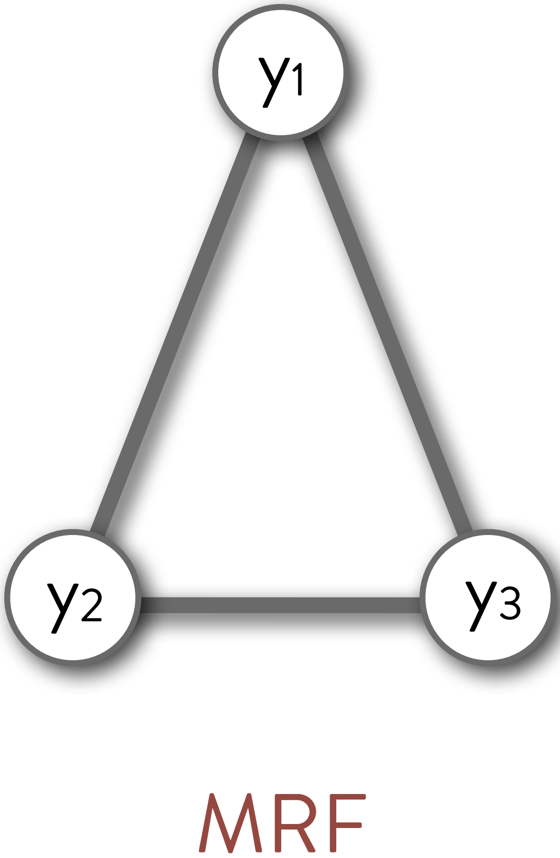
$$\mathcal{P}(Y) \propto \det(L_Y)$$

$$\begin{matrix} L_{11} & L_{12} & L_{13} \\ & L_{22} & L_{23} \\ & & L_{33} \end{matrix}$$

$$L \succeq 0$$

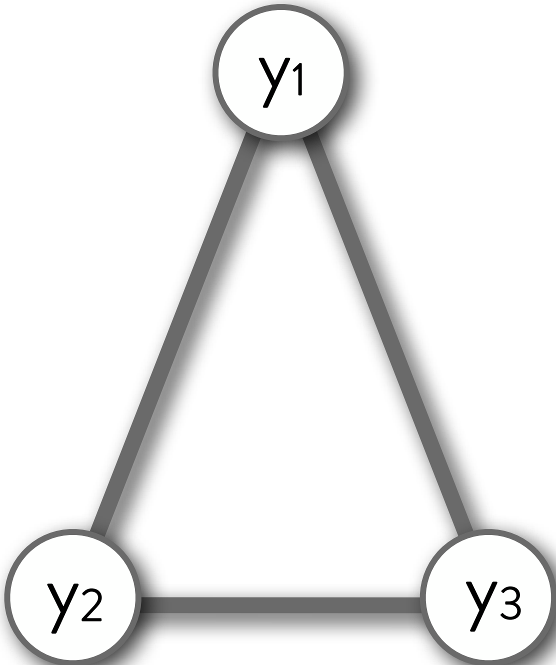


MRF



$$\mathcal{P}(Y) \propto \exp\left(\sum_i w_i y_i + \sum_{i<j} w_{ij} y_i y_j\right)$$

w_1	w_2	w_3
w_{12}	w_{13}	w_{23}



MRF

$$\mathcal{P}(Y) \propto$$
$$\exp\left(\sum_i w_i y_i + \sum_{i<j} w_{ij} y_i y_j\right)$$

w_1	w_2	w_3
w_{12}	w_{13}	w_{23}

$$w_{ij} \leq 0$$

DPP

L_{11}	L_{12}	L_{13}
	L_{22}	L_{23}
		L_{33}

MRF

w_1	w_2	w_3
w_{12}	w_{13}	w_{23}

y_1	y_2	y_3
0	0	0
1	0	0
0	1	0
0	0	1
1	1	0
1	0	1
0	1	1
1	1	1

DPP

$$\begin{matrix} L_{11} & L_{12} & L_{13} \\ & L_{22} & L_{23} \\ & & L_{33} \end{matrix}$$

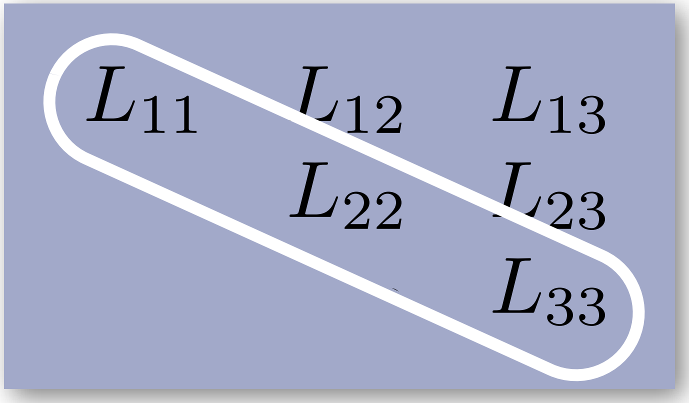
MRF

$$\begin{matrix} w_1 & w_2 & w_3 \\ w_{12} & w_{13} & w_{23} \end{matrix}$$

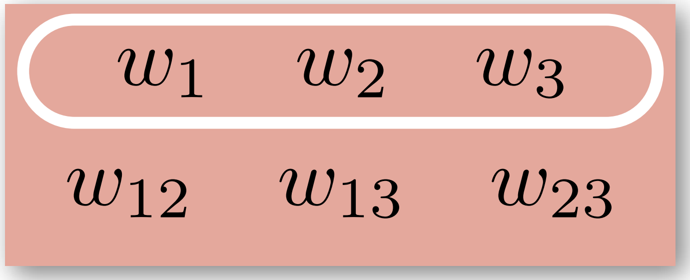
y_1	y_2	y_3
0	0	0
1	0	0
0	1	0
0	0	1
1	1	0
1	0	1
0	1	1
1	1	1

Arbitrary

DPP



MRF



y_1	y_2	y_3
0	0	0
1	0	0
0	1	0
0	0	1
1	1	0
1	0	1
0	1	1
1	1	1

Arbitrary

DPP

L_{12} L_{13}
L_{23}

MRF

w_{12} w_{13} w_{23}

y_1	y_2	y_3	
0	0	0	
1	0	0	Arbitrary
0	1	0	
0	0	1	
1	1	0	
1	0	1	
0	1	1	
1	1	1	Fix this

DPP

	L_{12}	L_{13}
		L_{23}

MRF

w_{12} w_{13} w_{23}

y_1	y_2	y_3	
0	0	0	
1	0	0	Arbitrary
0	1	0	
0	0	1	
1	1	0	
1	0	1	Plot these
0	1	1	
1	1	1	Fix this

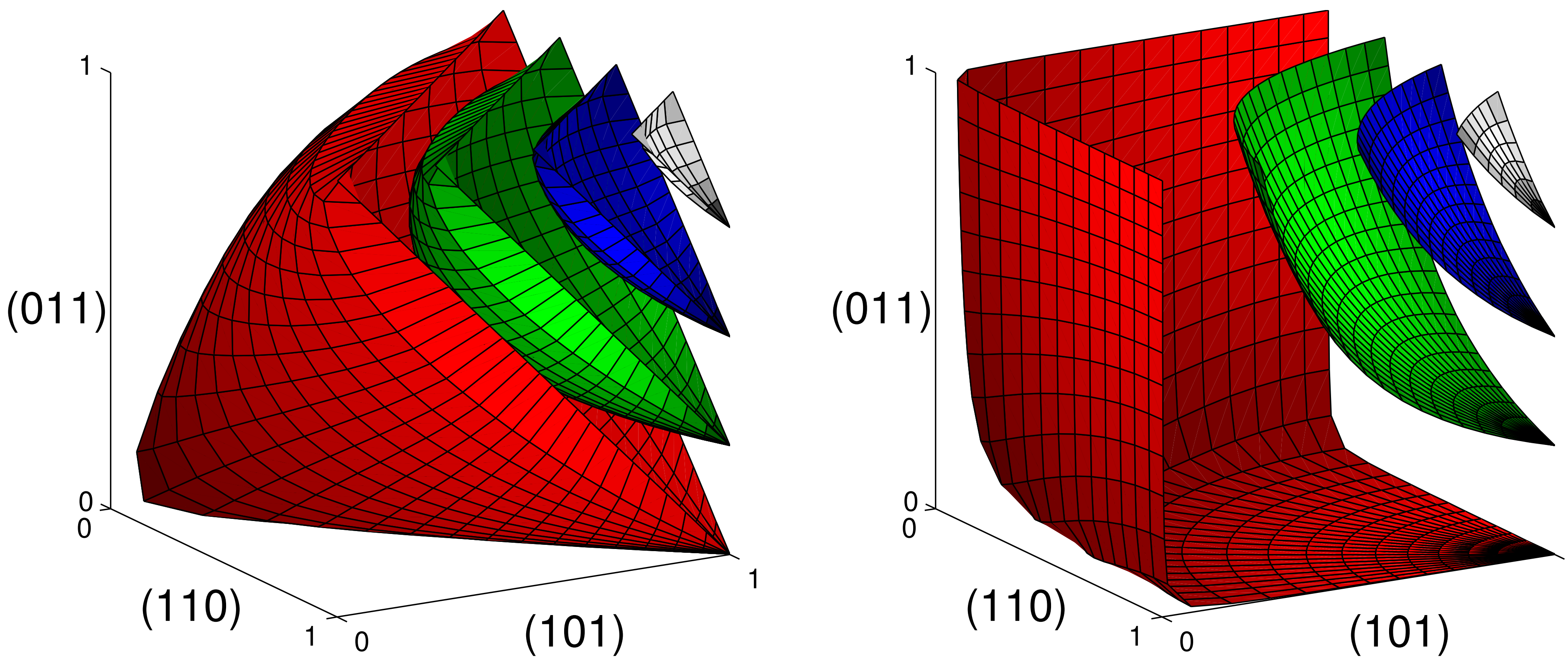
DPP

	L_{12}	L_{13}
		L_{23}

MRF

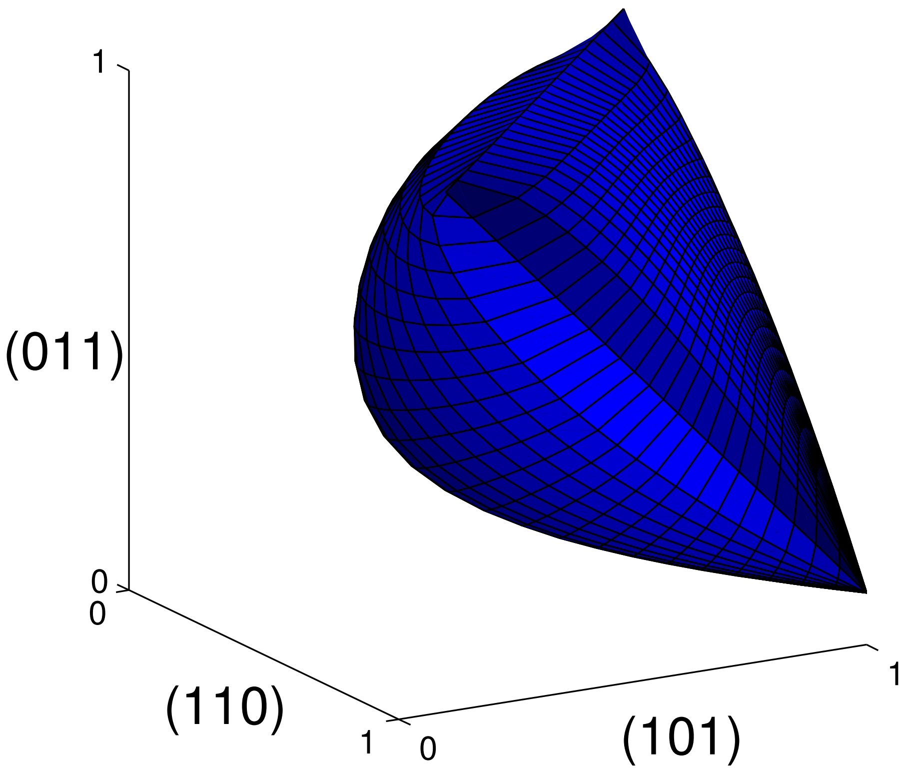
w_{12} w_{13} w_{23}

(111) : 0.001 0.25 0.5 0.75

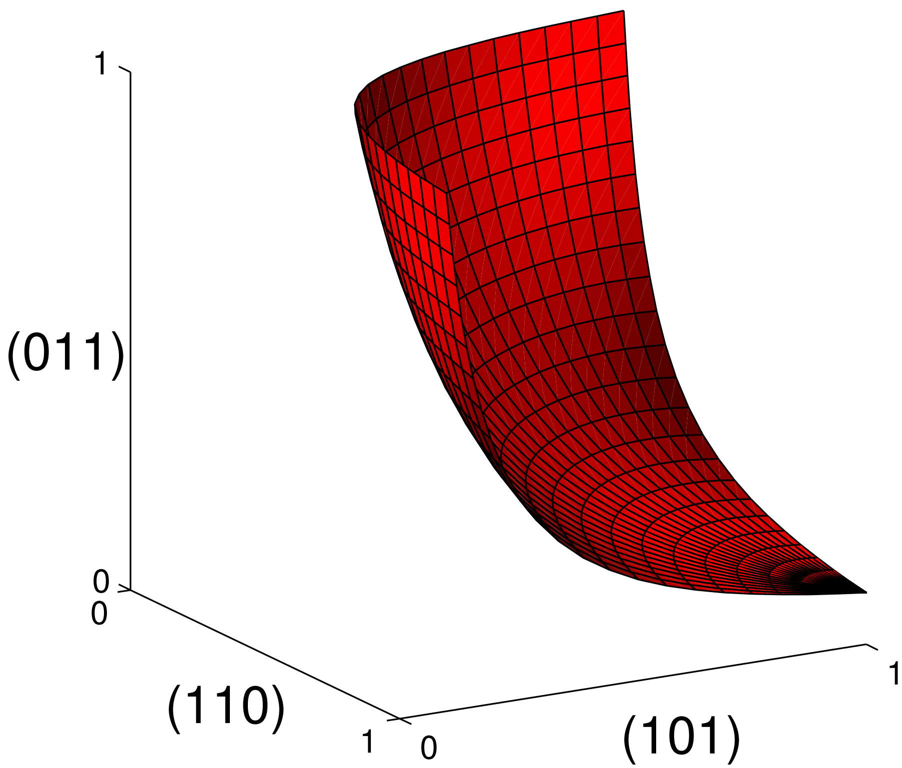


DPP

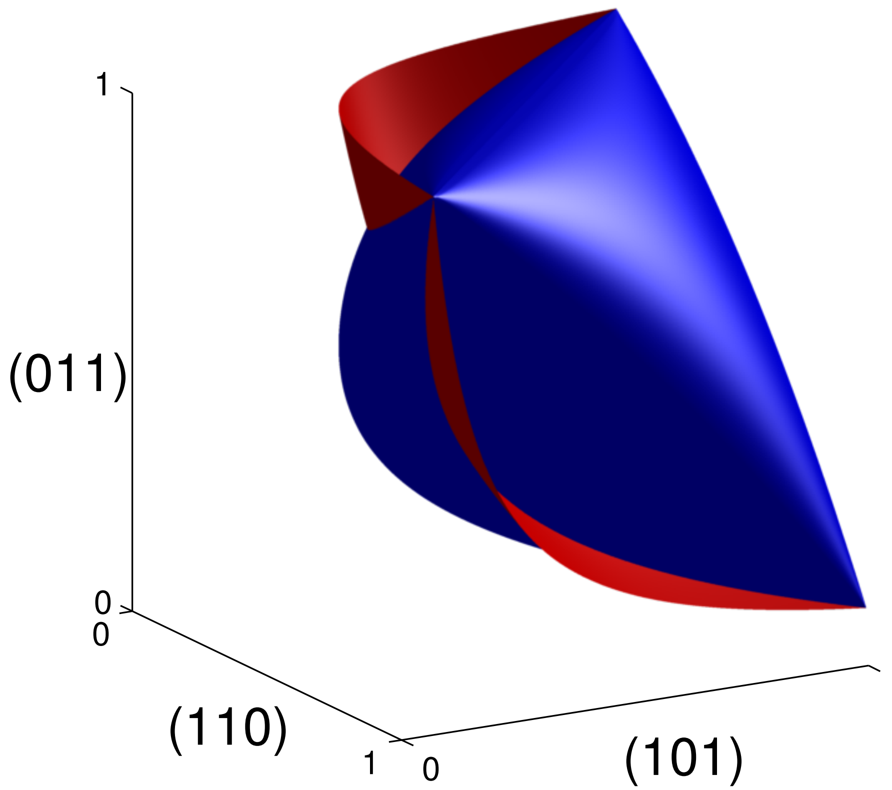
MRF



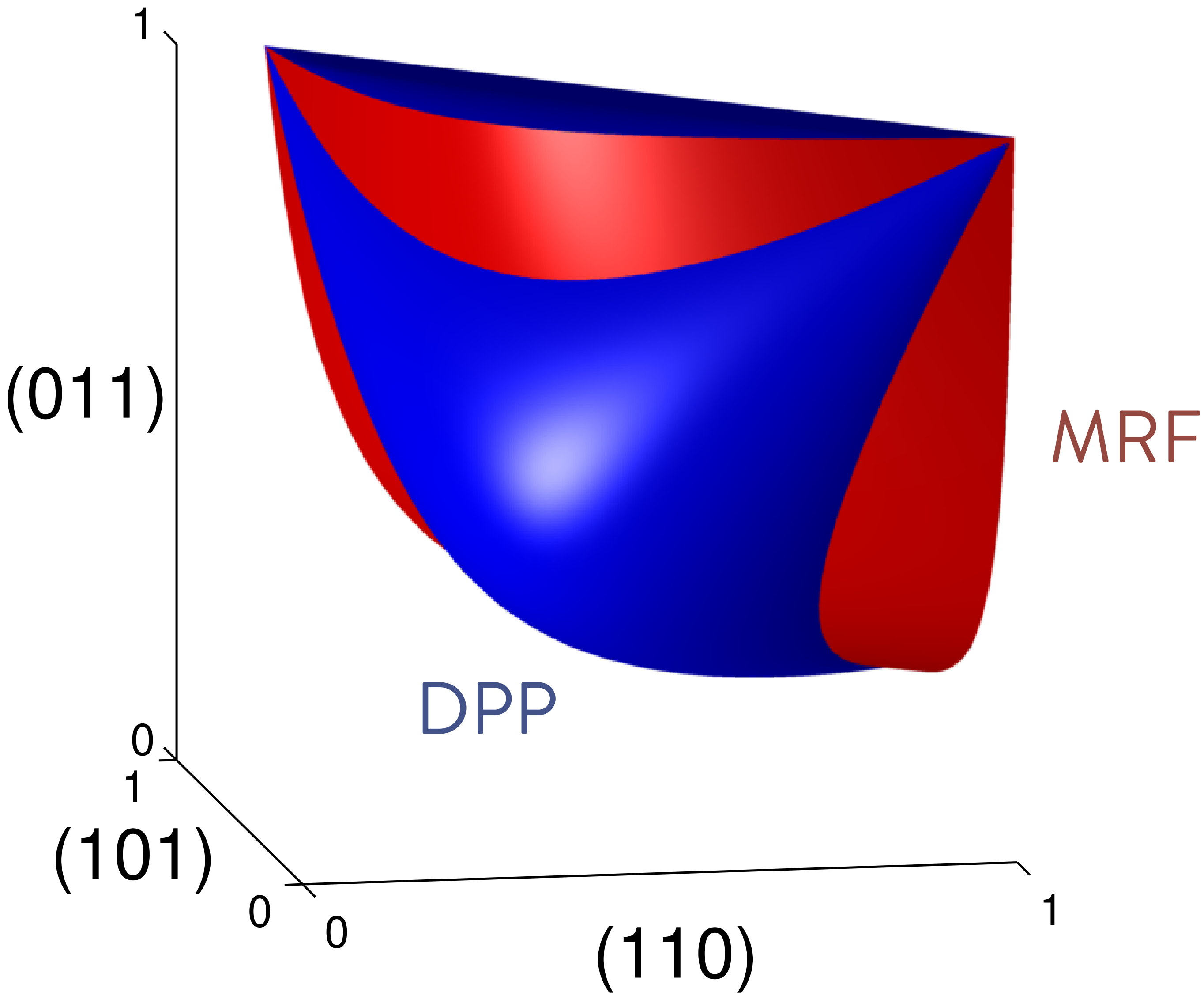
DPP



MRF



1
(011)
0
0
(110)
1
0
(101)
1


1
(011)
0
1
(101)
0
0
(110)
1
MRF
DPP

	Gaussian	DPP
Parameters	$O(N^2)$	$O(N^2)$
Closure	marginalization, conditioning	marginalization, conditioning
Independence	given by zeros of Σ^{-1}	given by zeros of K^{-1} (context specific)
Sufficient Statistics	1^{st} + 2^{nd} moments	1^{st} + 2^{nd} + 3^{rd} moments

Term 'determinant' first introduced by Gauss in *Disquisitiones arithmeticae* (1801)

Part I

Representation

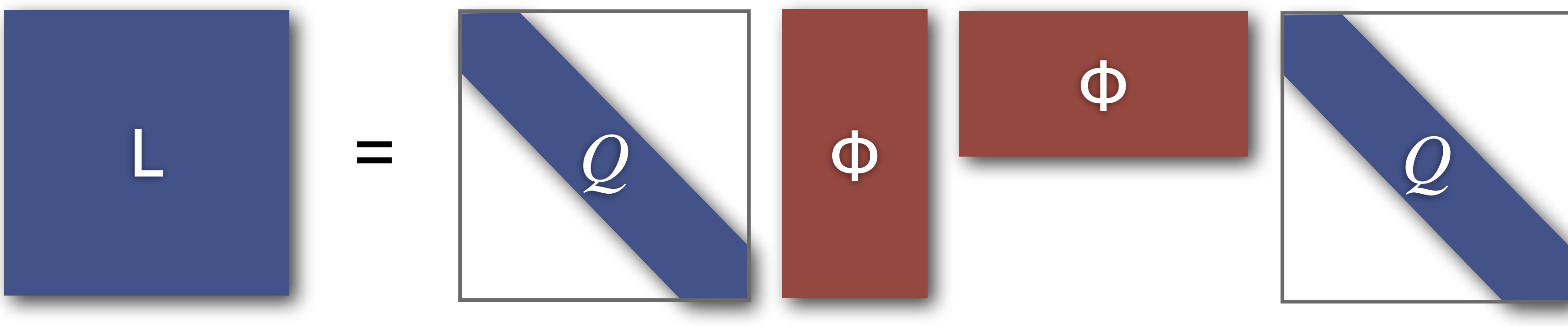
Inference: Marginals, Conditionals

Inference: Sampling

DPPs vs MRFs

Learning

$$L_{ij} = \boldsymbol{g}(i)^\top \boldsymbol{g}(j)$$



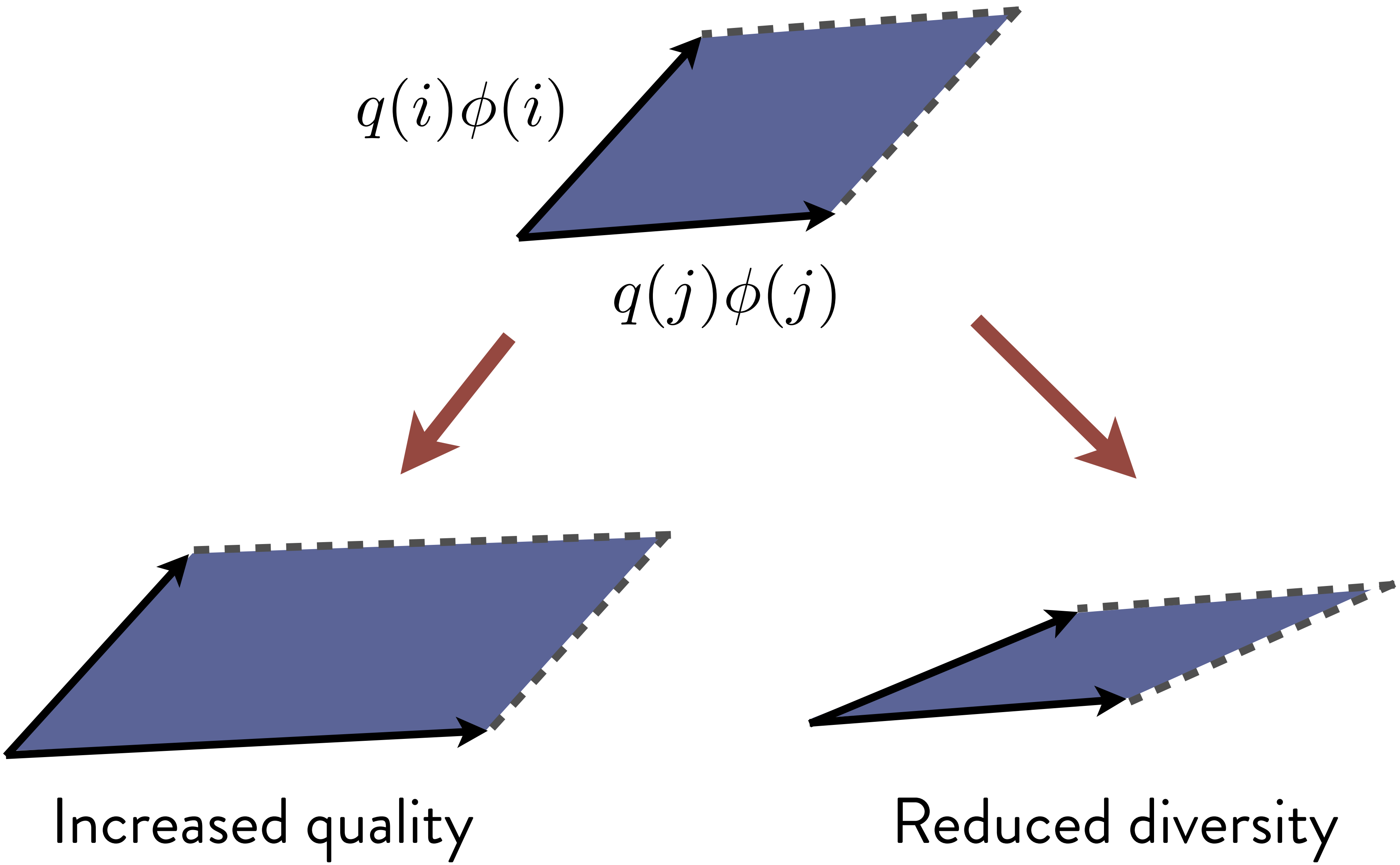
$$L_{ij} = q(i)\phi(i)^\top \phi(j)q(j)$$

$q(i) \in \mathbb{R}_+$

Quality score

$\phi(i) \in \mathbb{R}^D, \; \|\phi(i)\|^2 = 1$

Diversity features


$q(i)\phi(i)$
$q(j)\phi(j)$
Increased quality
Reduced diversity

$$\mathcal{P}(\boldsymbol{Y} = Y) \propto \det(L_Y)$$

$$= \det(\{q(i)\phi(i)^\top \phi(j)q(j)\}_{i,j\in Y})$$

$$= \det\left(\phi(Y)^\top \phi(Y)\right) \prod_{i\in Y} q^2(i)$$

$$\mathcal{P}(\boldsymbol{Y} = Y) \propto \det(L_Y)$$

$$= \det(\{q(i)\phi(i)^\top \phi(j)q(j)\}_{i,j \in Y})$$

$$= \det\left(\phi(Y)^\top \phi(Y)\right) \prod_{i \in Y} q^2(i)$$

Balance quality and diversity

$$\mathcal{P}(\boldsymbol{Y} = Y) \propto \det(L_Y)$$

$$= \det(\{q(i)\phi(i)^\top \phi(j)q(j)\}_{i,j\in Y})$$

$$= \det\left(\phi(Y)^\top \phi(Y)\right) \prod_{i\in Y} q^2(i)$$

Balance quality and diversity

Quality vs. **diversity**

- Intuitive and natural tradeoff
- Log-linear **quality** model:

$$q(i) = \exp(\theta^\top \boldsymbol{f}(i))$$

  - Optimize θ by maximum likelihood
- Open question: how to learn **diversity**

- Log-likelihood of training example Y:

Quality **Diversity** **Normalization**

$$\theta^\top \sum_{i \in Y} \boldsymbol{f}(i) + \log\det(\phi(Y)^\top \phi(Y)) - \log(Z)$$

- Concave in θ; gradient is:

$$\sum_{i \in Y} \boldsymbol{f}(i) - \sum_{Y'} \mathcal{P}(Y') \sum_{j \in Y'} \boldsymbol{f}(j)$$

Gradient of log-likelihood:

$$\sum_{i \in Y} \boldsymbol{f}(i) - \sum_{Y'} \mathcal{P}(Y') \sum_{j \in Y'} \boldsymbol{f}(j)$$

Gradient of log-likelihood:

$$\sum_{i \in Y} \boldsymbol{f}(i) - \sum_{Y'} \mathcal{P}(Y') \sum_{j \in Y'} \boldsymbol{f}(j)$$

$$= \sum_{i \in Y} \boldsymbol{f}(i) - \sum_{j} \boldsymbol{f}(j) \sum_{Y' \ni j} \mathcal{P}(Y')$$

Gradient of log-likelihood:

$$\sum_{i \in Y} \boldsymbol{f}(i) - \sum_{Y'} \mathcal{P}(Y') \sum_{j \in Y'} \boldsymbol{f}(j)$$

$$= \sum_{i \in Y} \boldsymbol{f}(i) - \sum_{j} \boldsymbol{f}(j) \sum_{Y' \ni j} \mathcal{P}(Y')$$

marginal of *j*

Gradient of log-likelihood:

$$\sum_{i \in Y} \boldsymbol{f}(i) - \sum_{Y'} \mathcal{P}(Y') \sum_{j \in Y'} \boldsymbol{f}(j)$$

$$= \sum_{i \in Y} \boldsymbol{f}(i) - \sum_{j} \boldsymbol{f}(j) \; K_{jj}$$

Compute gradient efficiently

News summarization

- **Input**: 10 news articles, ~250 sentences
- **Output**: 665 character summary
- **Eval**: ROUGE metric (four human summaries)

Hot dog in pizza is the stuff of dreams

- A gut-busting pizza has been launched – with a hot dog sausage stuffed in the crust.
- Pizza Hut has released the limited edition dish after the success of its cheese and BBQ crusts.
- Dubbed the "pizza dog", the 14-inch feast is only available for delivery and costs up to £19.49.

[The Sun, 4/12/12]

Quality features

- Dubbed the “pizza dog”, the 14-inch feast is only available for delivery and costs up to £19.49.

Quality features

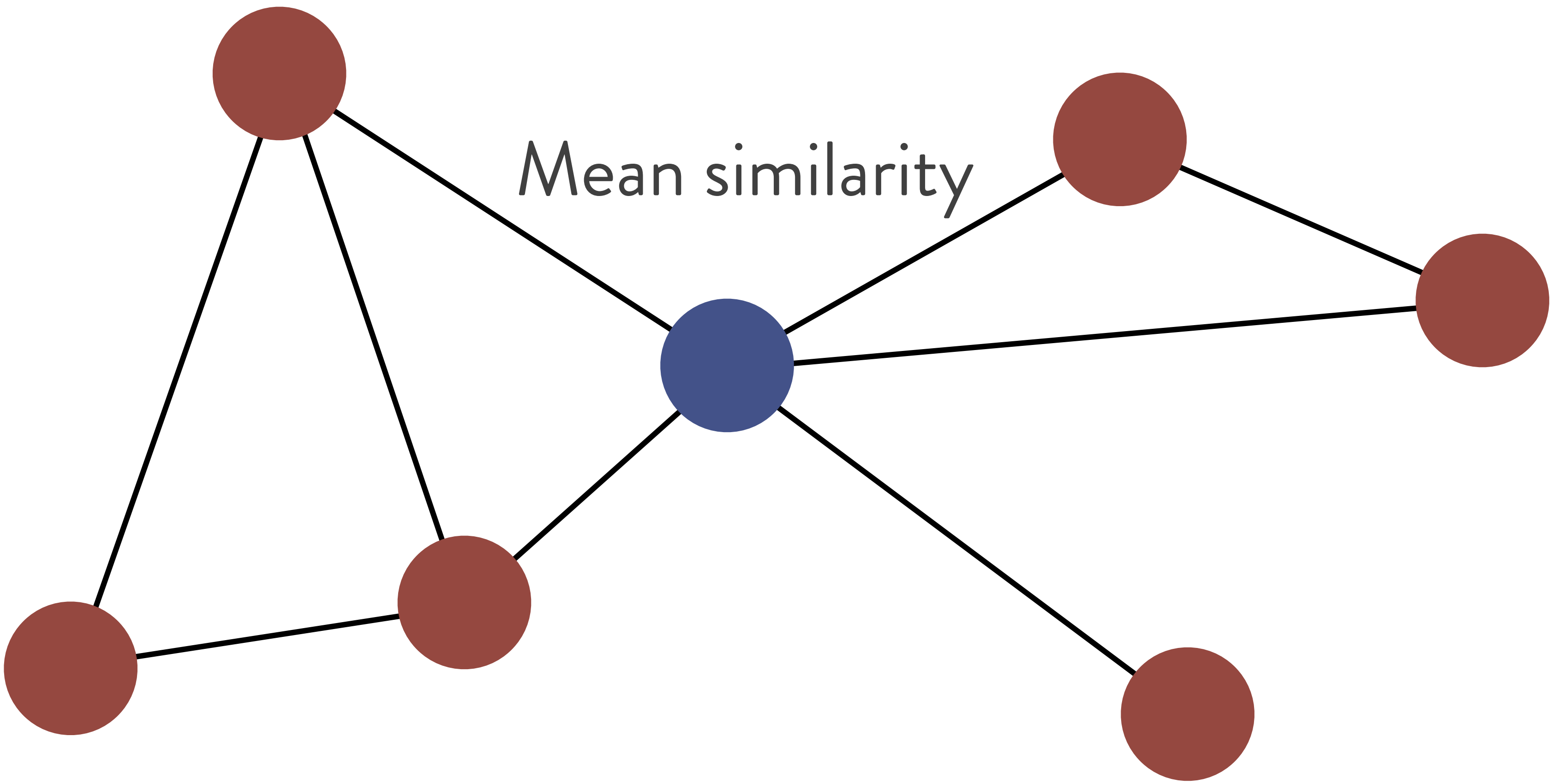
Position in article

2. Pizza Hut has released the limited edition dish after the success of its cheese and BBQ crusts.

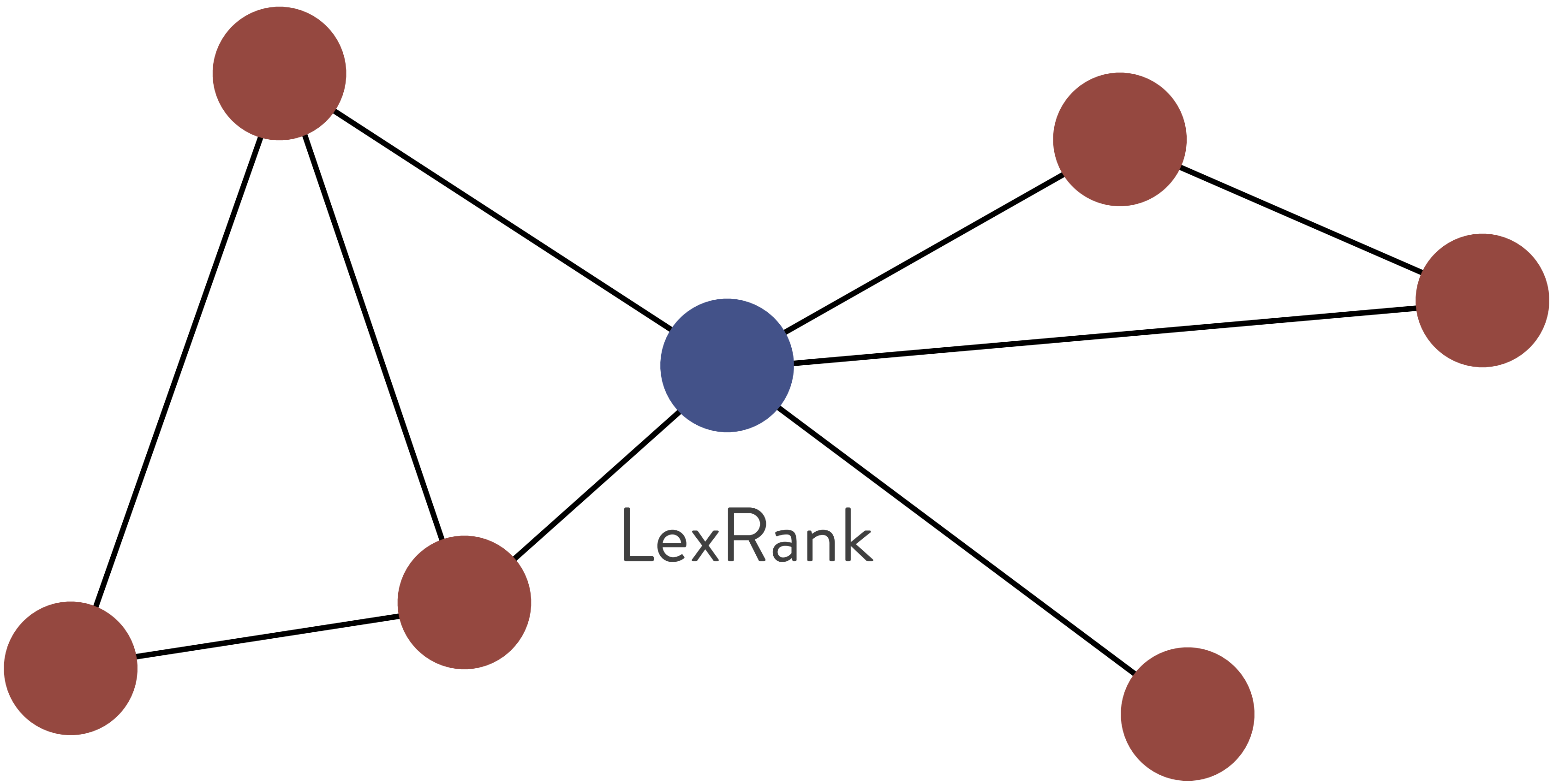
3. Dubbed the "pizza dog", the 14-inch feast is only available for delivery and costs up to £19.49.

4. The firm was the first to stuff its crusts and has been selling the hot dog variety in Thailand and Japan since 2007.

Quality features



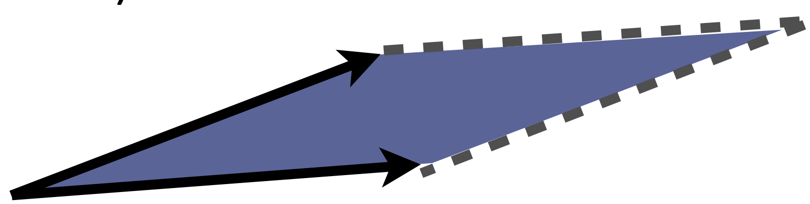
Quality features



Diversity features

- ϕ are tf-idf vectors: cosine similarity

The 14-inch "pizza dog" is available for delivery.

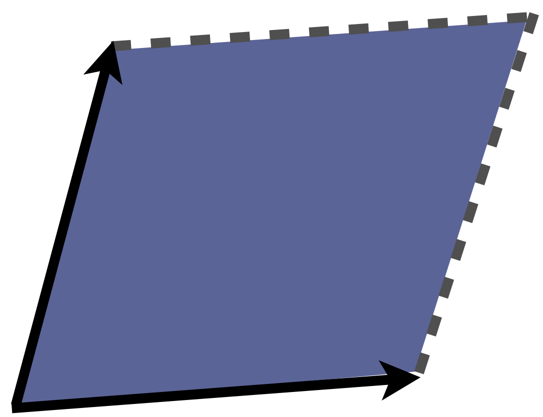


Dubbed the "pizza dog", the 14-inch feast is only available for delivery and costs up to £19.49.

Diversity features

- ϕ are tf-idf vectors: cosine similarity

Sadly, this caloric coma is not available in the U.S. yet.



Dubbed the “pizza dog”, the 14-inch feast is only available for delivery and costs up to £19.49.

Greedy MAP decoding

- Initialize summary Y to empty
- Add sentence i maximizing:

$$\frac{\log \mathcal{P}(Y \cup \{i\}|X) - \log \mathcal{P}(Y|X)}{\text{length}(i)}$$

Until budget full

✓ Simple, fast, good results

– Inexact, ignores loss

[Lin and Bilmes, 2010]

Minimum Bayes risk decoding

- Choose Y to maximize:

$$\mathbb{E}_{Y^*}\left[\textsc{rouge-1f}(Y, Y^*)\right]$$

[Goel and Byrne, 2000]

Minimum Bayes risk decoding

- Choose Y to maximize:

$$\mathbb{E}_{Y^*}\left[\text{ROUGE-1F}(Y, Y^*)\right]$$

[Goel and Byrne, 2000]

Minimum Bayes risk decoding

- Draw samples: $Y^1, Y^2, \ldots, Y^R$
- Choose Y to maximize:

$$\mathbb{E}_{Y^*} \left[\text{ROUGE-1F}(Y, Y^*) \right]$$

[Goel and Byrne, 2000]

Minimum Bayes risk decoding

- Draw samples: $Y^1, Y^2, \ldots, Y^R$
- Choose Y to maximize:

$$\frac{1}{R}\sum_{r=1}^{R} \text{ROUGE-1F}(Y, Y^r)$$

[Goel and Byrne, 2000]

Minimum Bayes risk decoding

- Draw samples: $Y^1, Y^2, \ldots, Y^R$
- Choose Y^s to maximize:

$$\frac{1}{R}\sum_{r=1}^{R} \text{ROUGE-1F}(Y^s, Y^r)$$

[Goel and Byrne, 2000]

Minimum Bayes risk decoding

- Draw samples: $Y^1, Y^2, \ldots, Y^R$
- Choose Y^s to maximize:

$$\frac{1}{R}\sum_{r=1}^{R} \text{ROUGE-1F}(Y^s, Y^r)$$

✓ Loss-sensitive, improves results

– Slower

[Goel and Byrne, 2000]

System	ROUGE-1F	ROUGE-1R	R-SU4F
Begin	32.08	32.69	10.37
MMR	37.58	38.05	13.06
Peer 65	37.87	38.20	13.19
SubMod*	39.78	40.43	-
DPP greedy	38.96	39.15	13.83
DPP MBR	**40.33**	**41.31**	**14.13**
LR+DPP	37.96	38.31	13.13

[*Lin and Bilmes, 2012]

Part I Representation, inference, comparison to other models, learning

Break

Part II Large-scale inference, extensions, sets of structures, applications

Part II

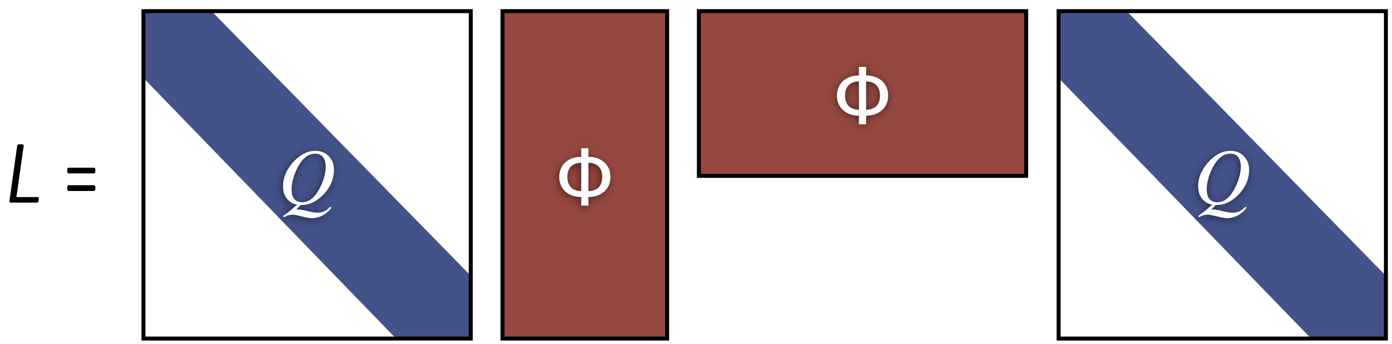
Large-scale DPPs

k-DPPs

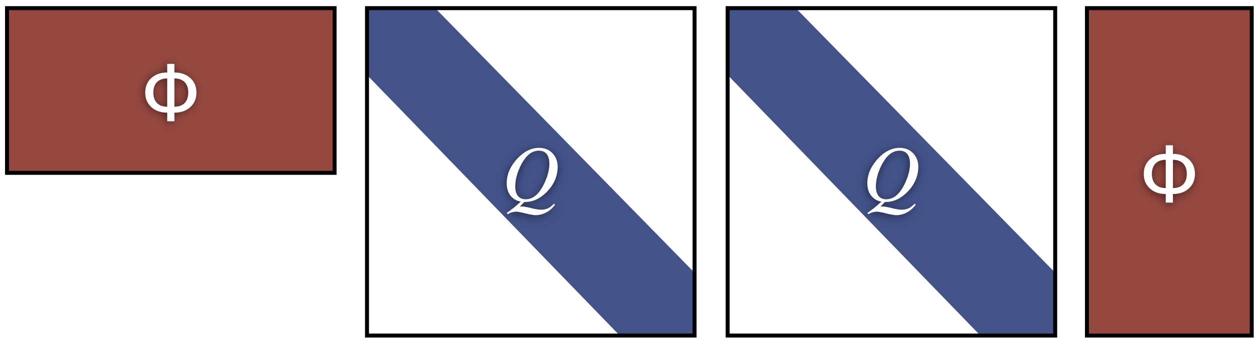
Structured DPPs

News threading

Conclusion

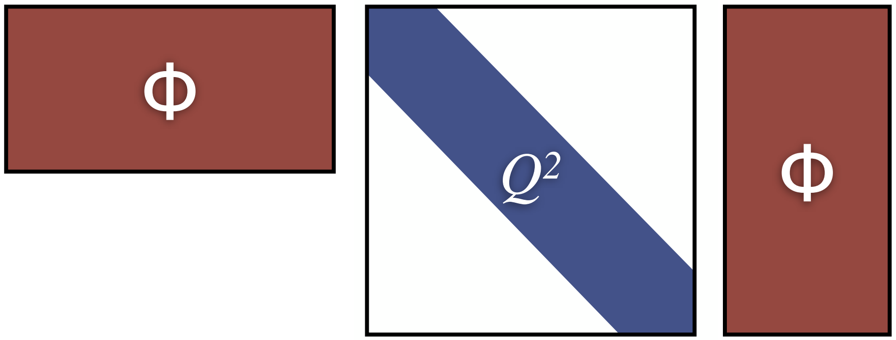


$$L_{ij} = q(i)\phi(i)^\top \phi(j) q(j)$$

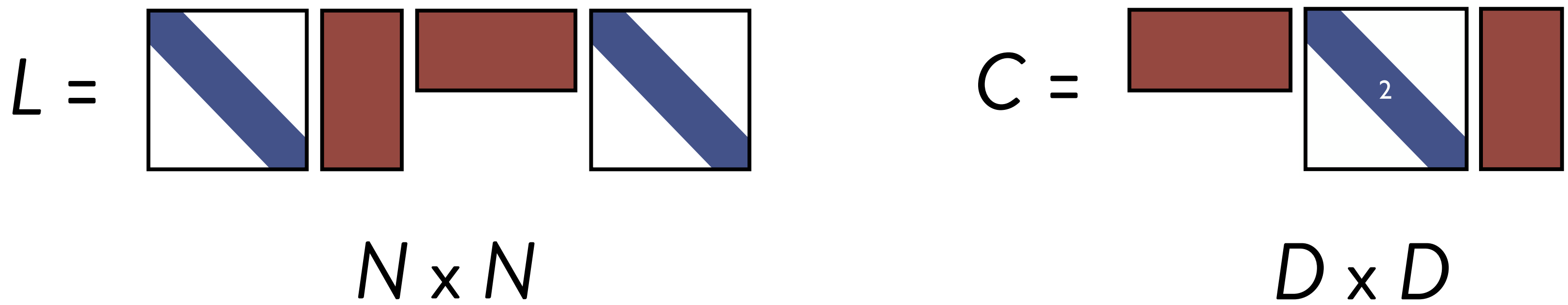


Φ
Q
Q
Φ

$C =$



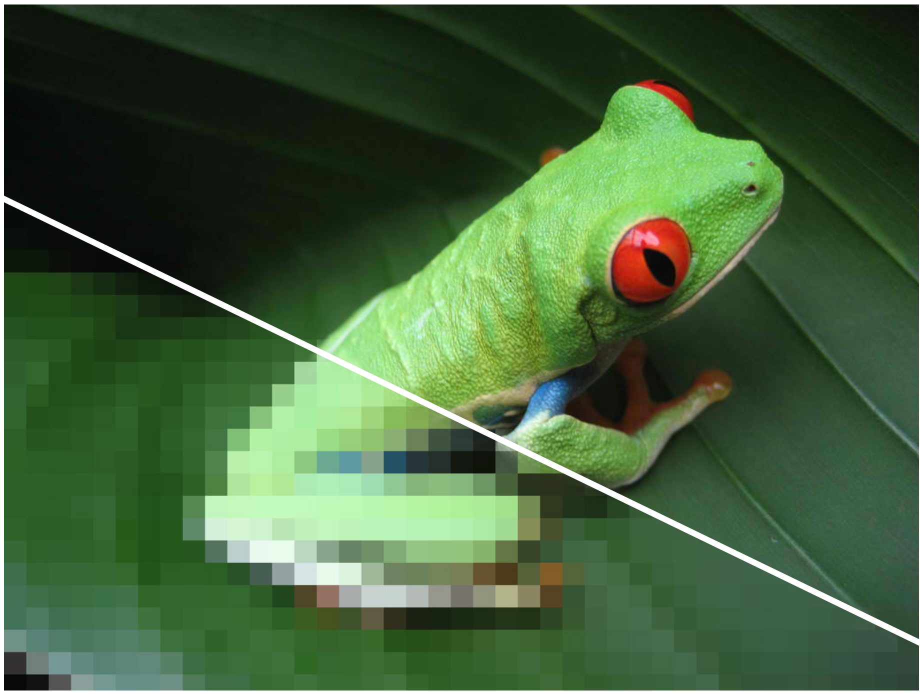
Dual representation



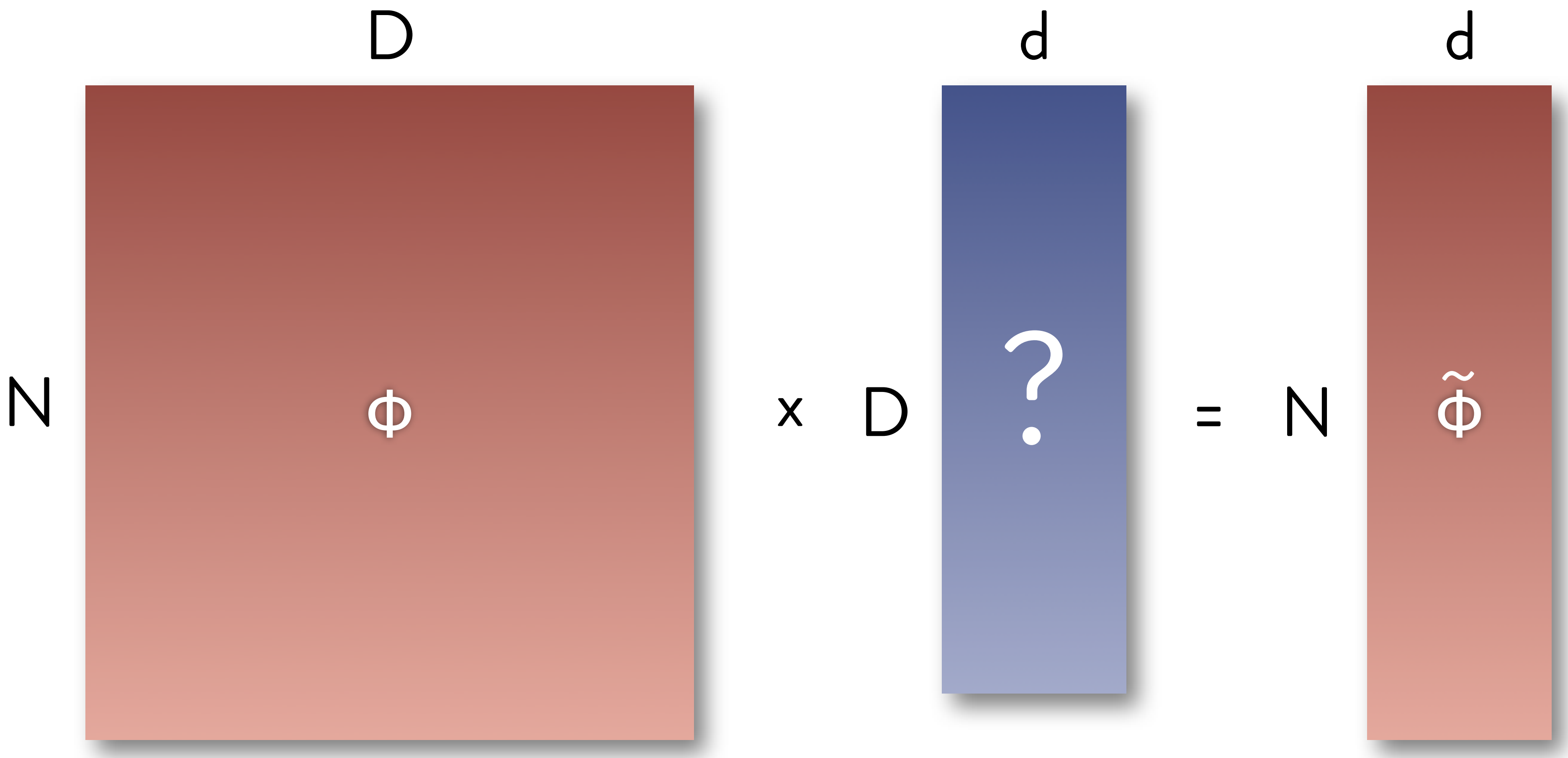
- C and L have same (non-zero) eigenvalues
- Eigenvectors are related
- Use C for sampling and other inference

DPPs at scale

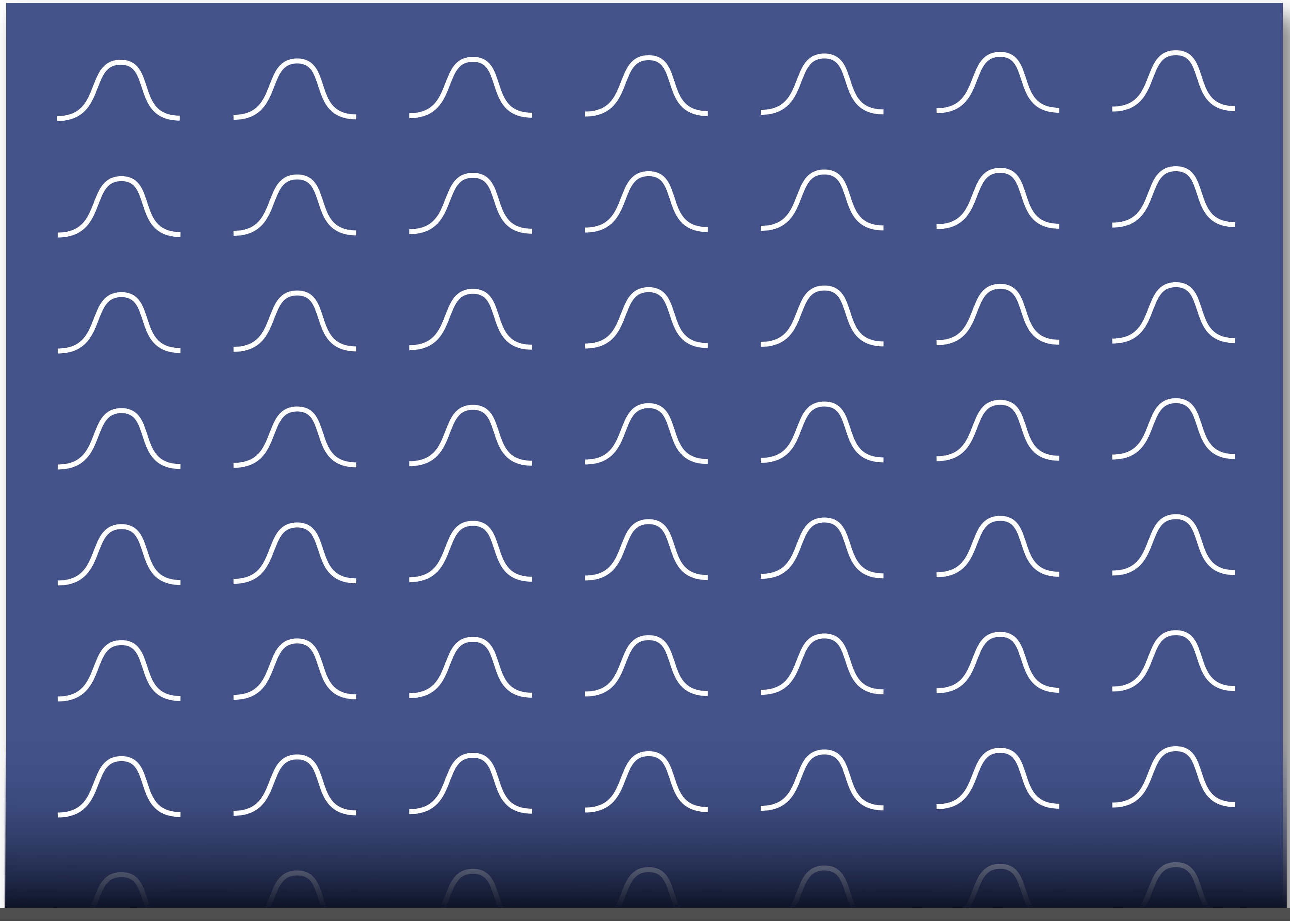
	Small N	Large N
Small D	Standard DPP or dual DPP	Dual DPP
Large D	Standard DPP	?

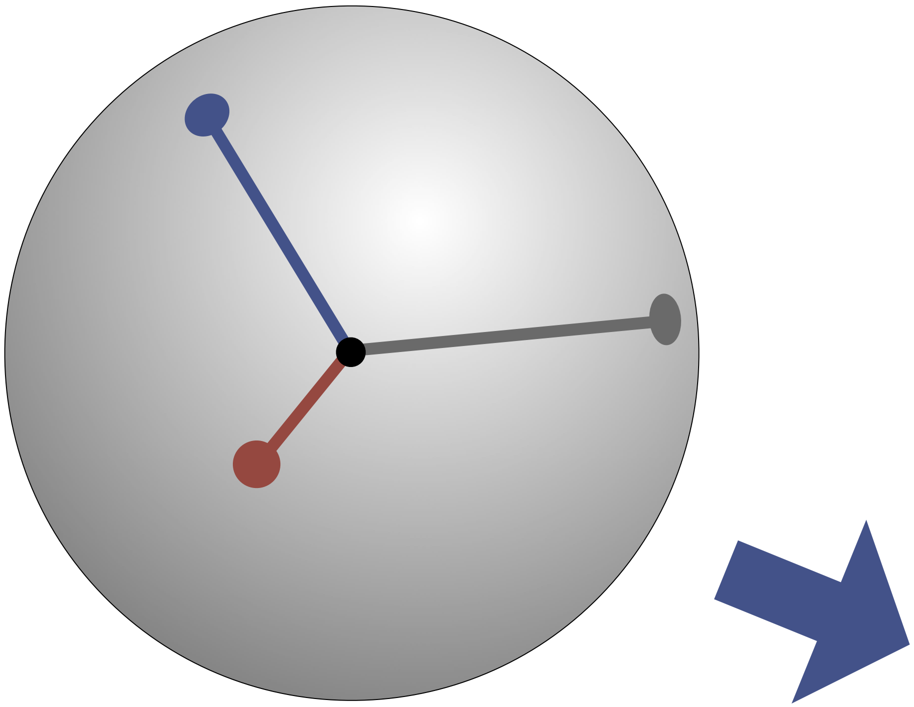


Projection

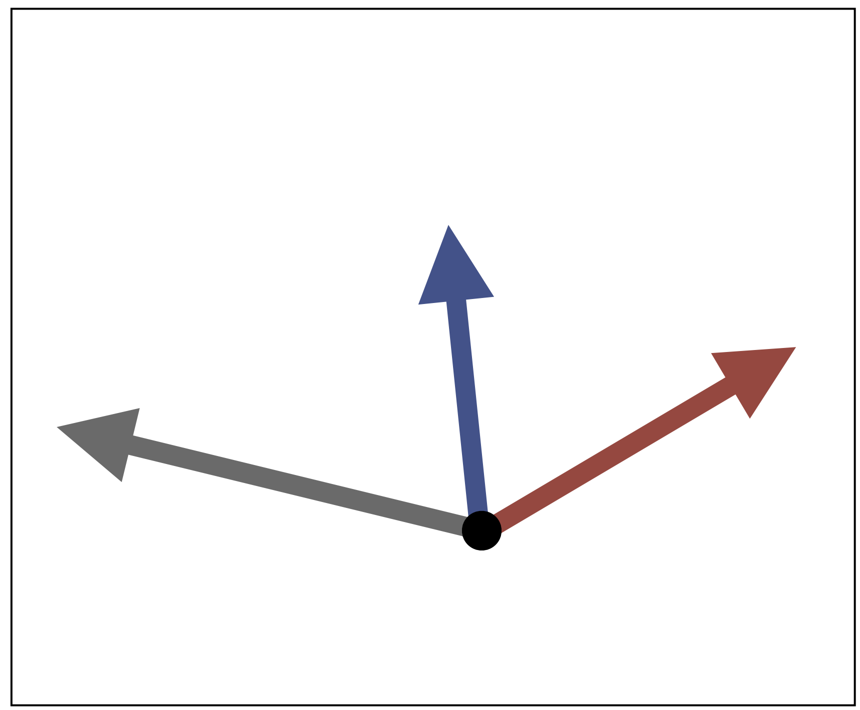


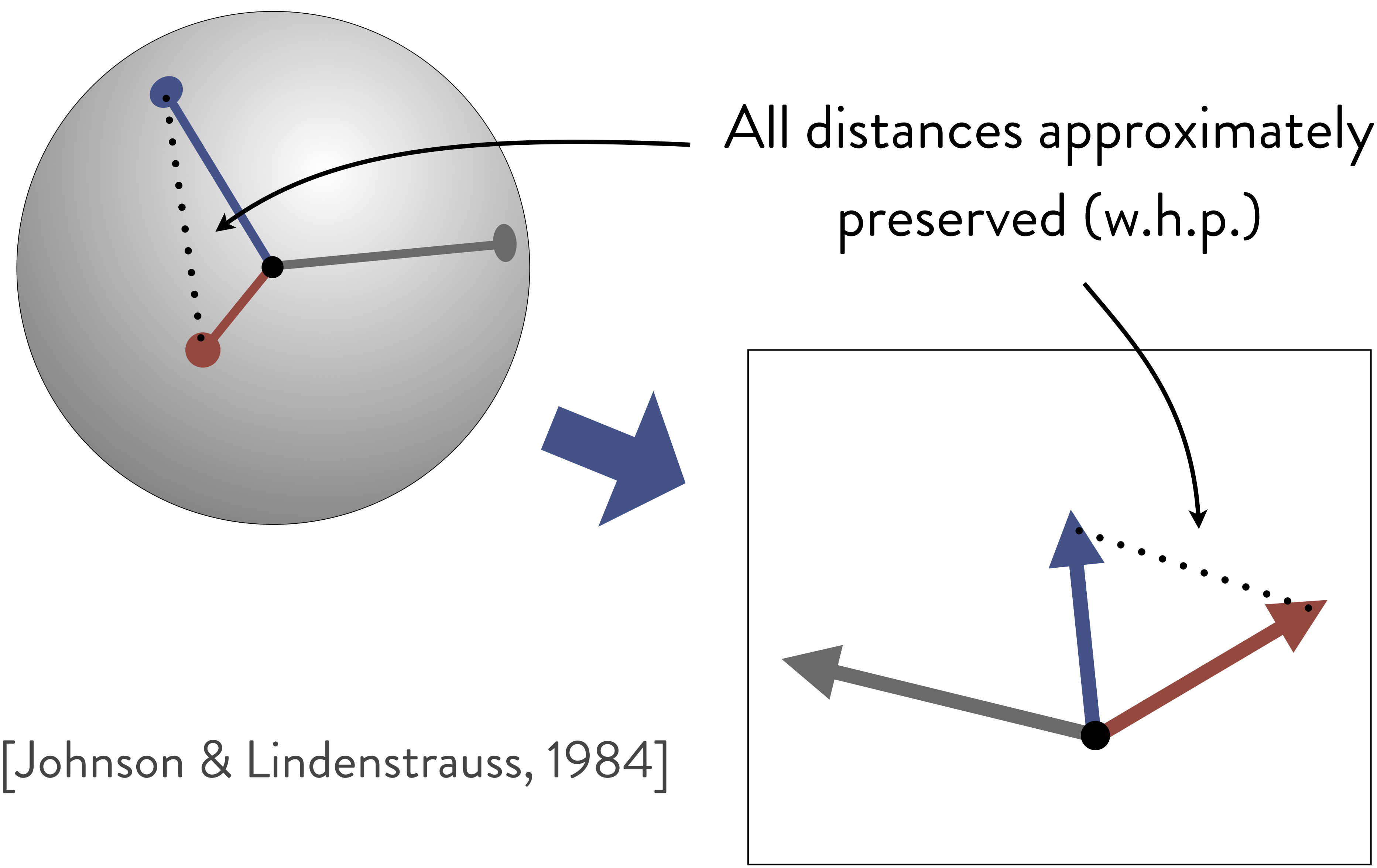
Random projection



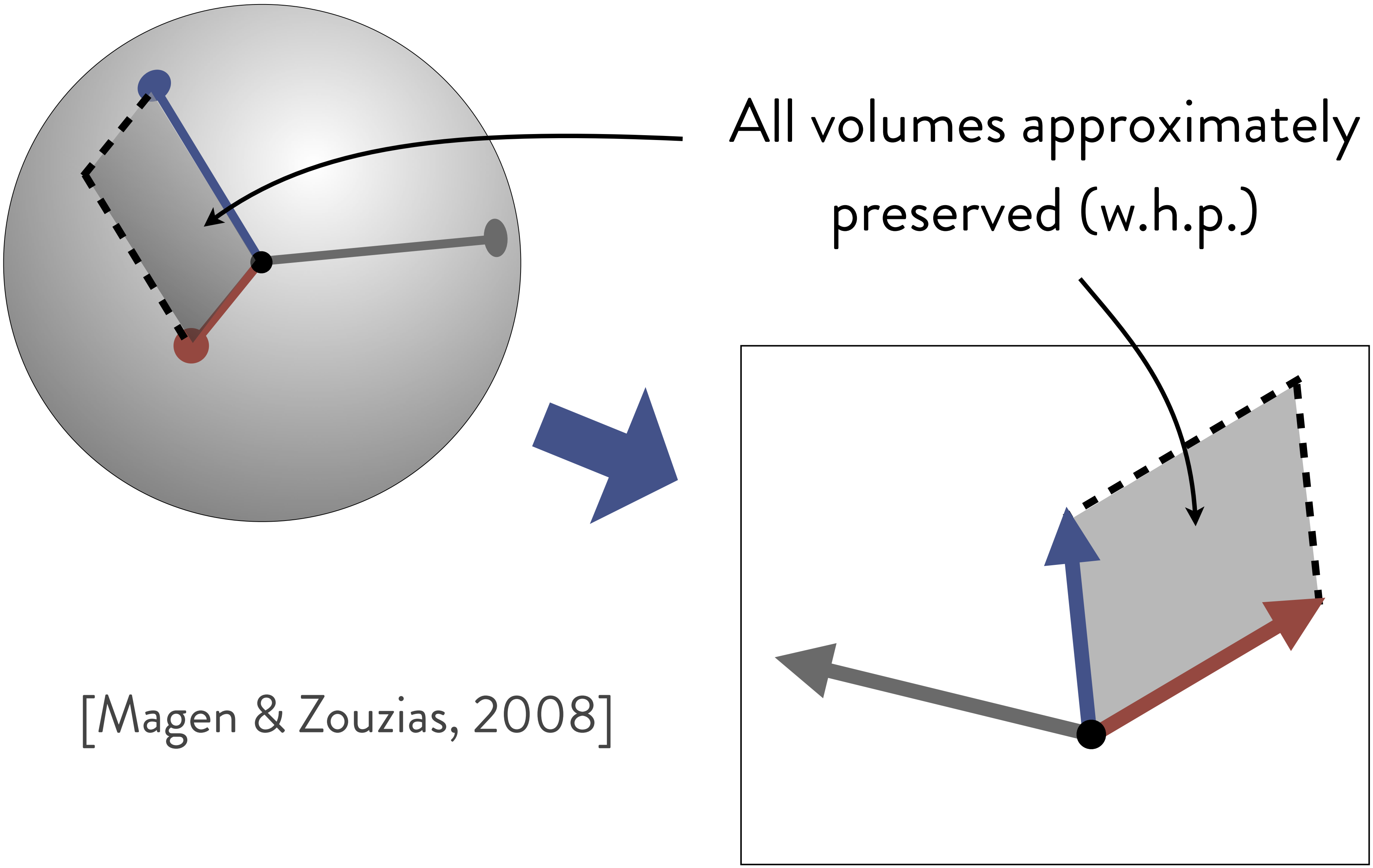


Random projection
to log N dimensions





[Johnson & Lindenstrauss, 1984]



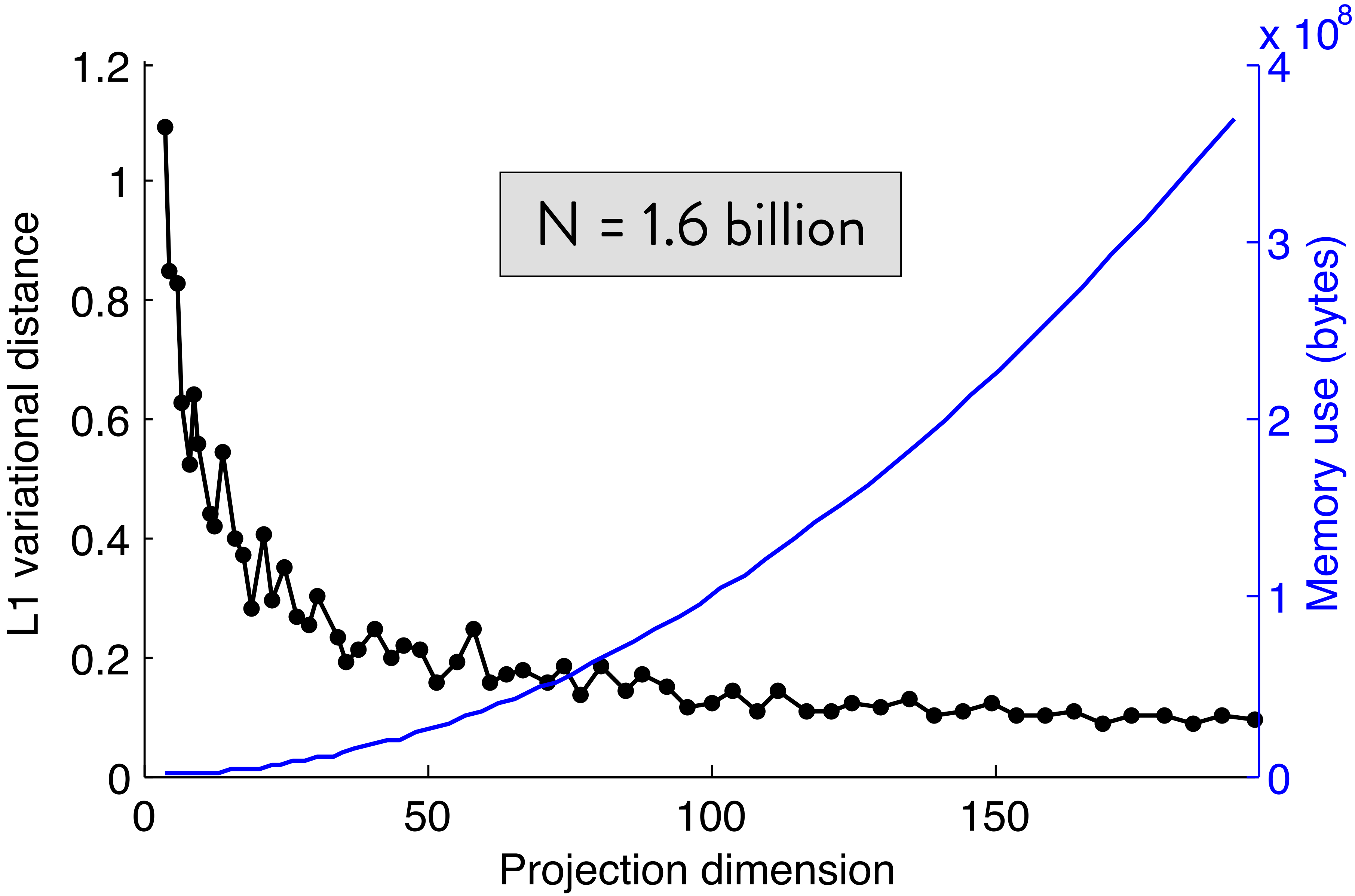
[Magen & Zouzias, 2008]

Random projection for DPPs

- **Theorem**: For $d = O\left(\frac{\log N}{\epsilon^2}\right)$ random projections, with high probability we have

$$\|\mathcal{P} - \tilde{\mathcal{P}}\|_1 \leq O(\epsilon) \ .$$

- Logarithmic in N, no dependence on D
- Small, d x d dual representation


N = 1.6 billion
L1 variational distance
Memory use (bytes)
x 10^8
Projection dimension
0
0.2
0.4
0.6
0.8
1
1.2
0
1
2
3
4
0
50
100
150

DPPs at scale

	Small N	Large N
Small D	Standard DPP or dual DPP	Dual DPP
Large D	Standard DPP	Random projection dual DPP

Part II

Large-scale DPPs

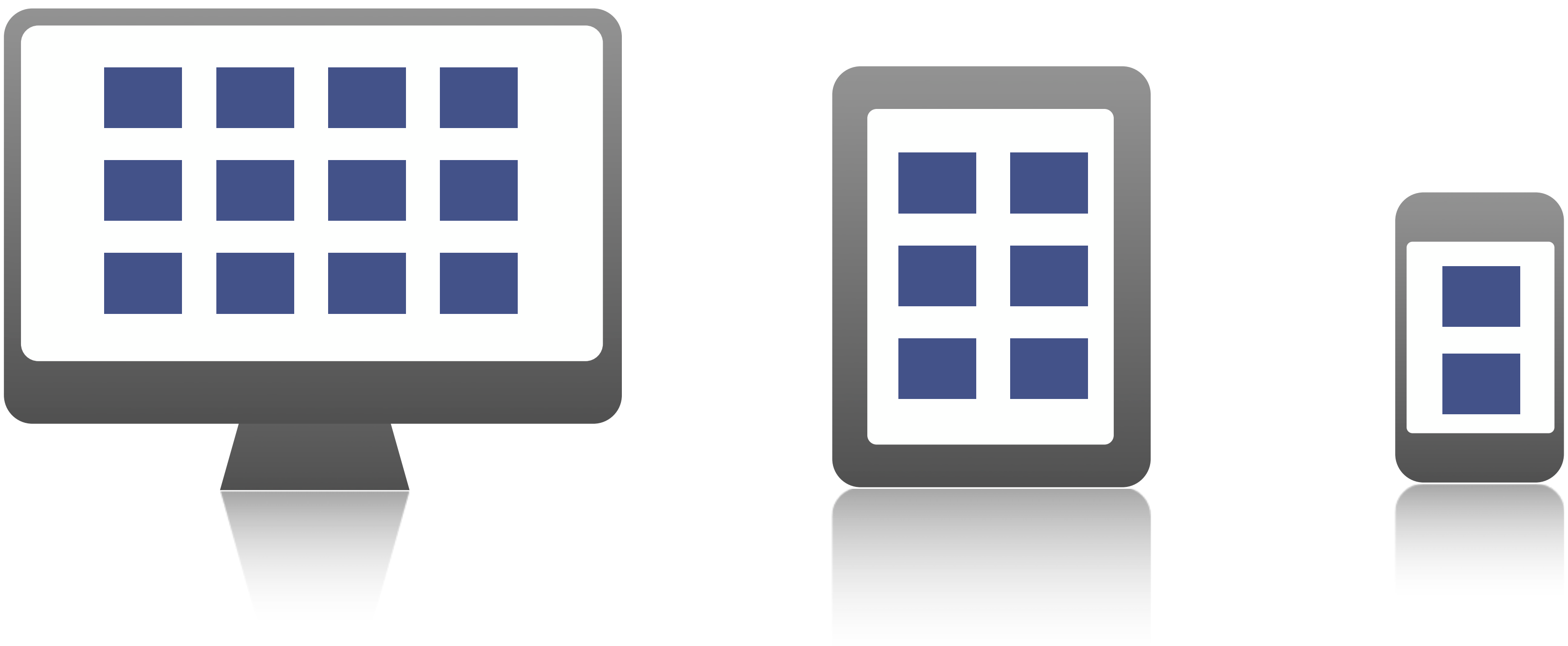
k-DPPs

Structured DPPs

News threading

Conclusion

What if we need exactly *k* diverse items?



k-DPPs

- Simple idea: condition DPP on target size k

$$\mathcal{P}^k(Y) = \frac{\det(L_Y)}{\sum_{|Y'|=k} \det(L_{Y'})}$$

- Can choose k at test time
- But inference (naively) looks exponential!

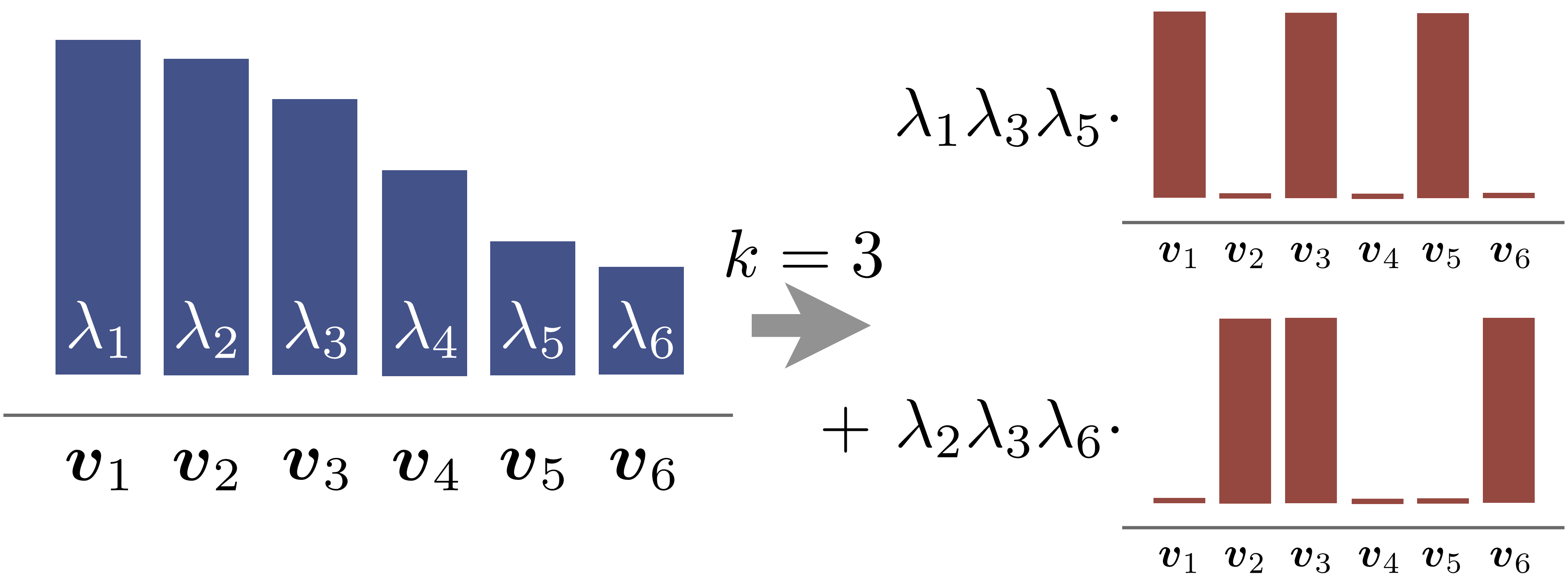
DPP

$$\mathcal{P} \propto \sum_{J \subseteq \{1,\ldots,N\}} \mathcal{P}^J \prod_{n \in J} \lambda_n$$

k-DPP

$$\mathcal{P} \propto \sum_{\substack{J \subseteq \{1,\ldots,N\} \\ |J| = k}} \mathcal{P}^J \prod_{n \in J} \lambda_n$$

$$\mathcal{P} \propto \sum_{\substack{J \subseteq \{1,\dots,N\} \\ |J| = k}} \mathcal{P}^J \prod_{n \in J} \lambda_n$$



$+ \cdots$

k-DPP sampling

- Need new PHASE ONE to pick $|J| = k$
- No longer independent:
  - Once we pick one, can only pick k-1 more

k-DPP sampling

- Solution: recursion on elementary symmetric polynomials:

$$e_k^N = \sum_{\substack{J \in \{1,\ldots,N\} \\ |J|=k}} \prod_{n \in J} \lambda_n$$

- Using dynamic prog. PHASE ONE is $O(Nk)$
- PHASE TWO is unchanged

Image search



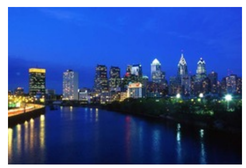
- 2,016 images from Google Image Search
  - 3 categories: cars, cities, dog breeds
- Diversity judgments: Amazon Mechanical Turk





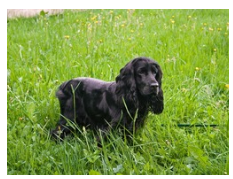



LOVE


Philadelphia







Learning

- Learn mixture of 55 “expert” k-DPPs:
  - SIFT
  - Color histograms
  - GIST
  - Center only / all pairs

k=2

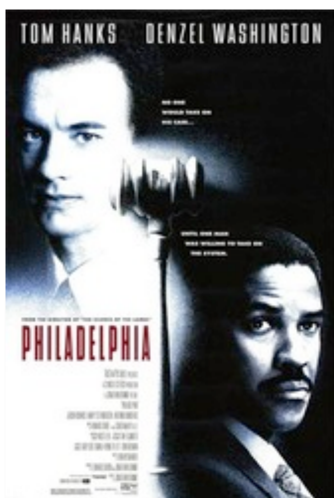
“porsche”

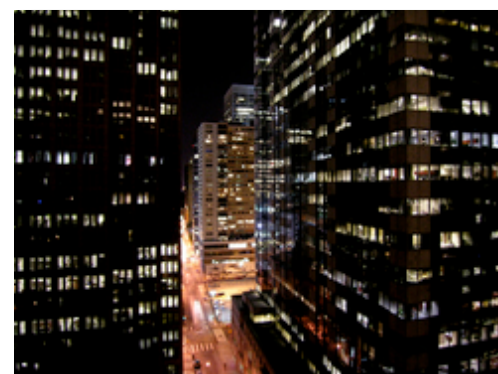
k=4

"philadelphia"

k=2

k=4

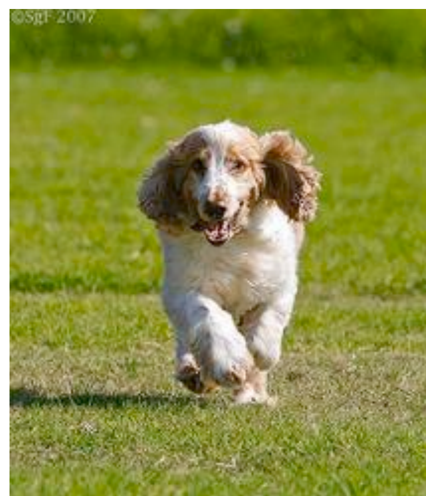




“cocker spaniel”

k=2

k=4



Labeling accuracy

System	Cars	Cities	Dogs
Single MMR*	55.95	56.48	56.23
Mixture MMR*	59.59	60.99	57.39
Mixture *k*-DPP	**64.58**	61.29	**59.84**

*[Carbonell and Goldstein, 1998]

Part II

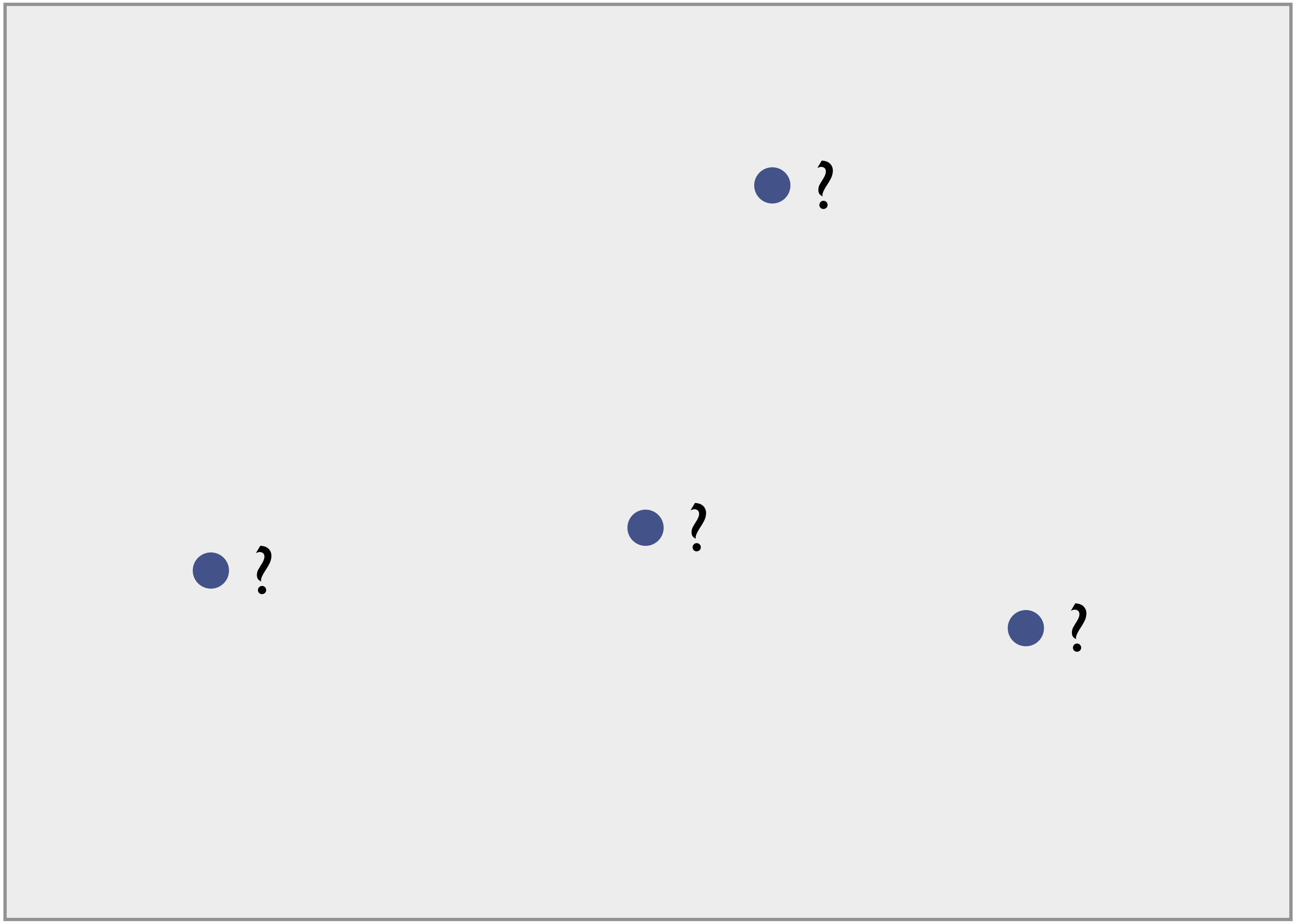
Large-scale DPPs

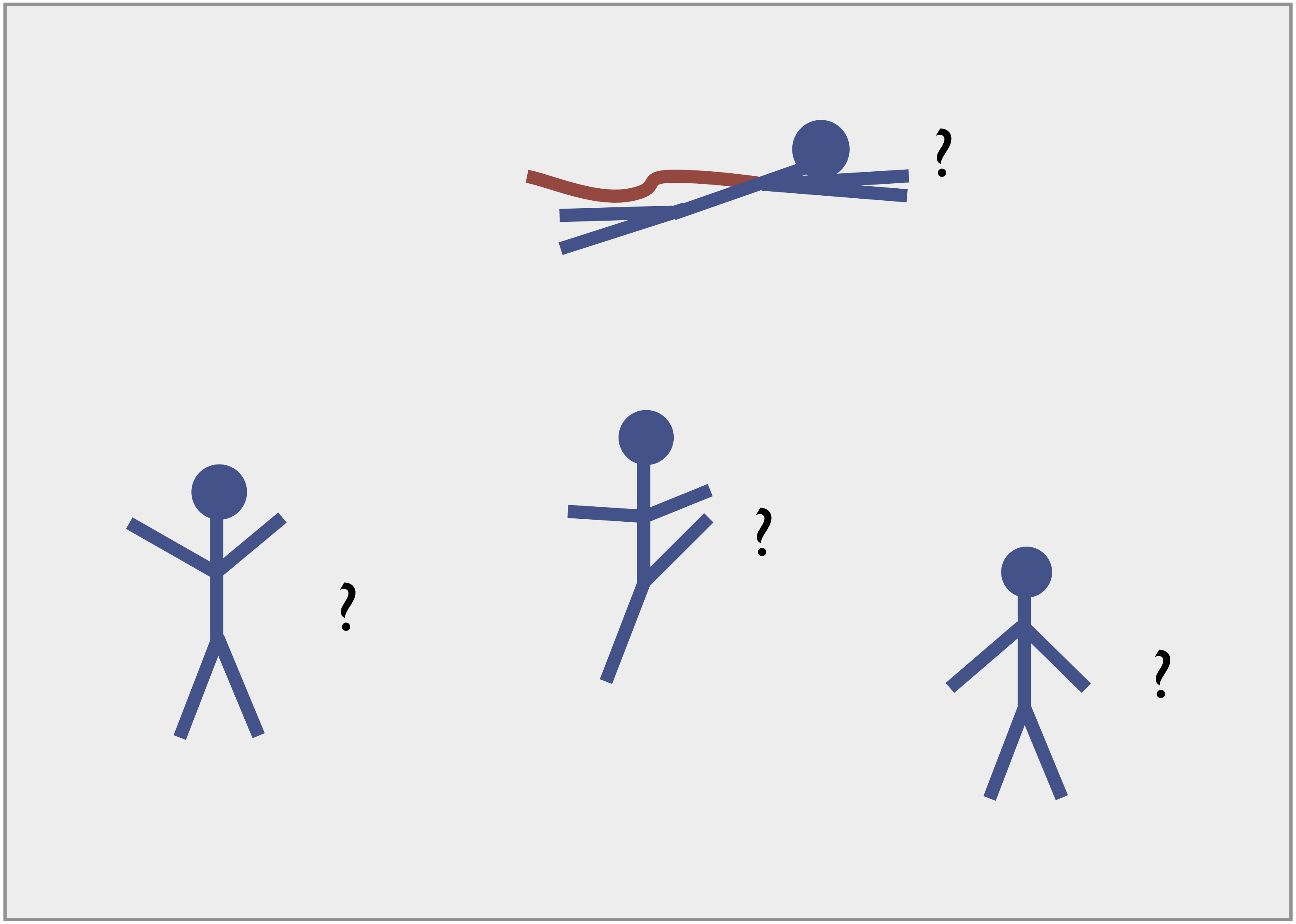
k-DPPs

Structured DPPs

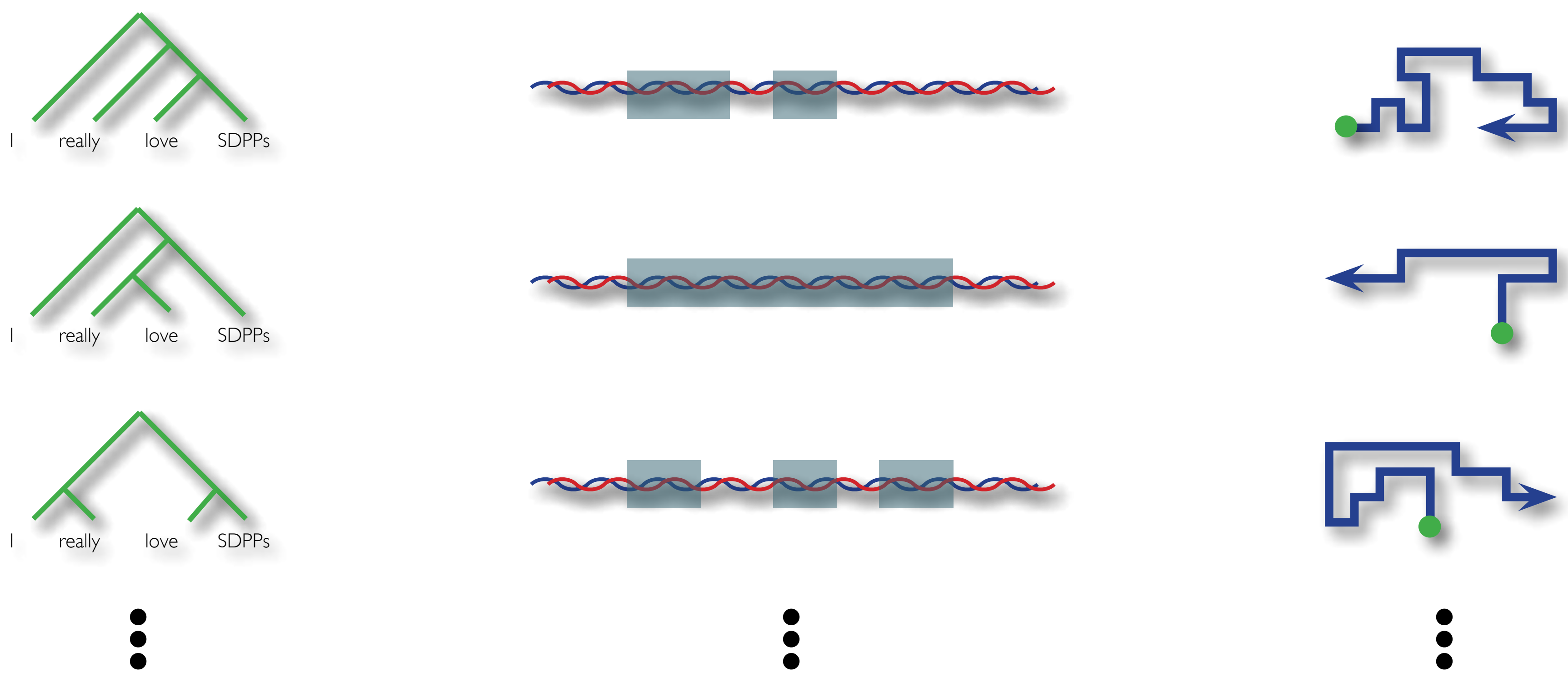
News threading

Conclusion





$\mathcal{Y}$ $\mathcal{Y}$ $\mathcal{Y}$



Structured DPPs

- Exponentially many complex “items”
- Can’t even handle O(N)
- But can still compute marginals and sample!

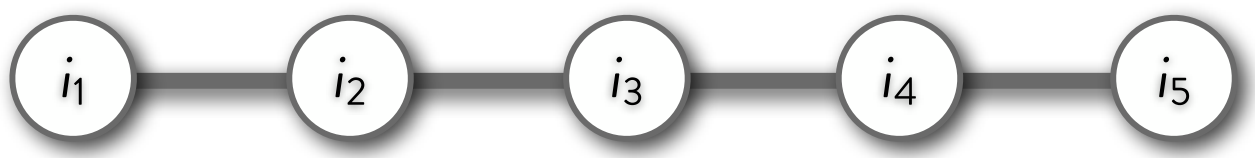
1. Factorized model
2. Dual DPPs
3. Second order message-passing

Structure

- Each item $\boldsymbol{i} \in \mathcal{Y}$ is a structure with factors α:

$$\boldsymbol{i} = \{i_\alpha\}$$

- For instance, standard sequence model:



1. Factorization

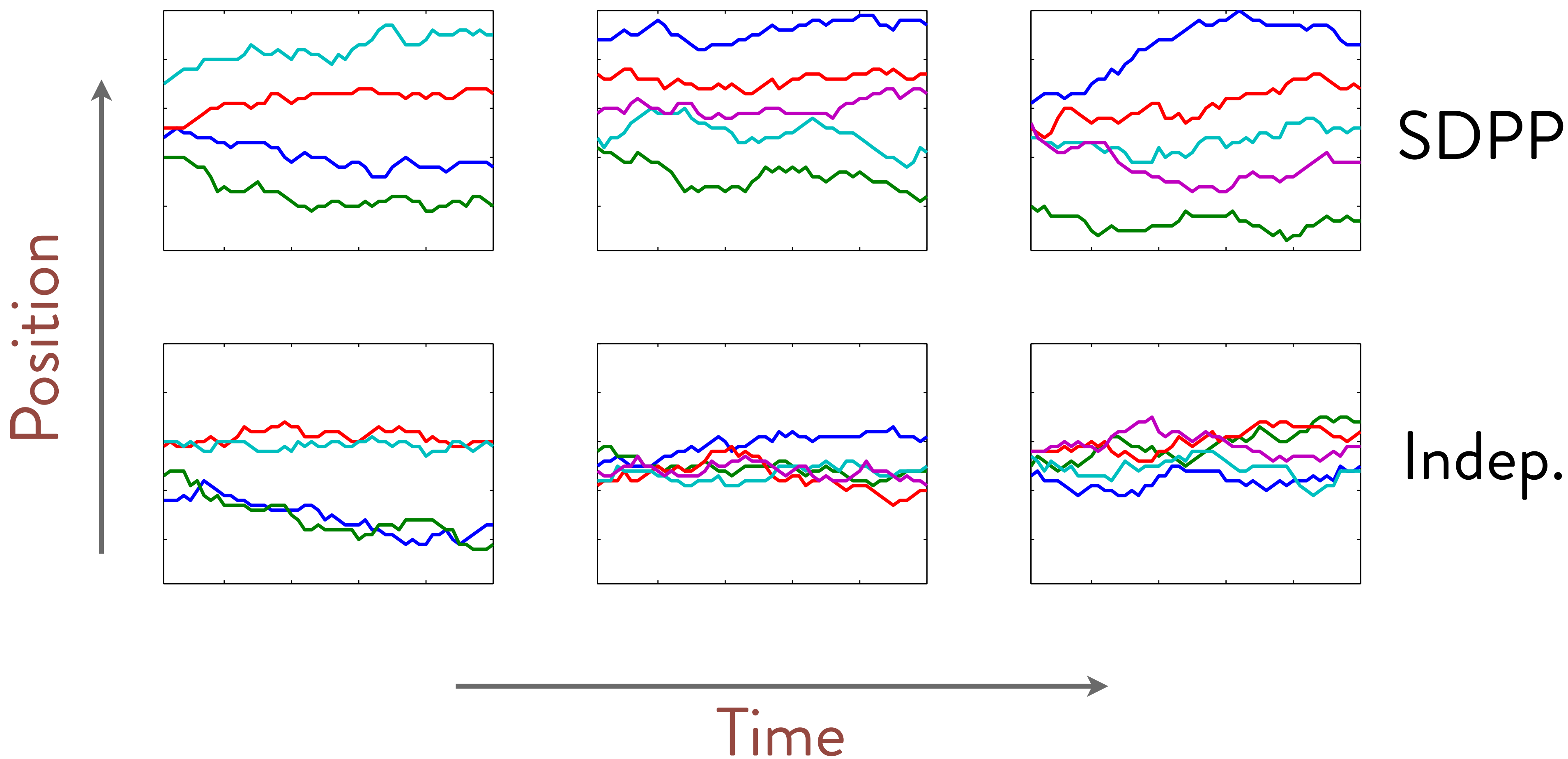
- Quality scores factor multiplicatively:

$$q(\boldsymbol{i}) = \prod_{\alpha} q(i_\alpha) \qquad \textbf{e.g., MRF}$$

- Diversity features factor additively:

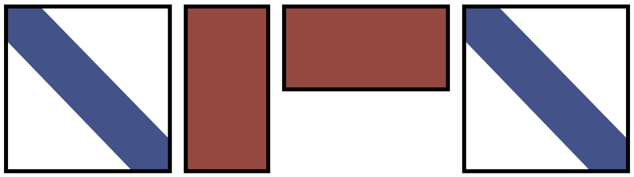
$$\phi(\boldsymbol{i}) = \sum_{\alpha} \phi(i_\alpha) \qquad \textbf{e.g., Hamming}$$

Synthetic particle tracking



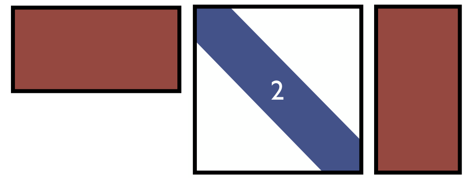
2. Dual representation

L =



N x N

C =



D x D

$$C_{rl} = \sum_{\boldsymbol{i}} q^2(\boldsymbol{i}) \phi_r(\boldsymbol{i}) \phi_l(\boldsymbol{i})$$

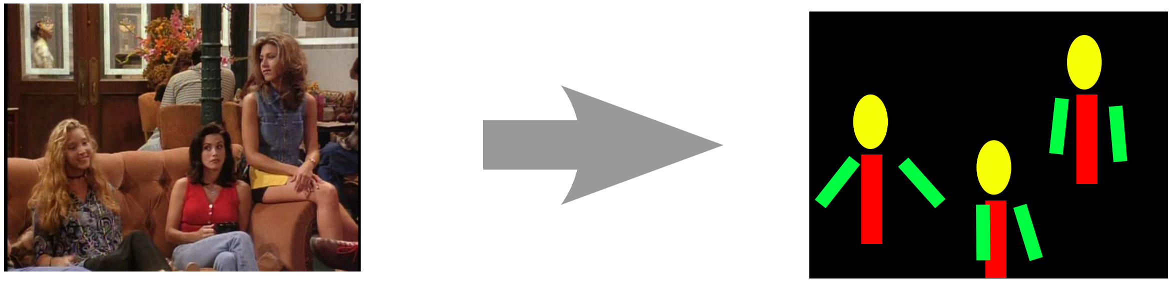
C is covariance of ϕ under $\Pr(\boldsymbol{i}) \propto q^2(\boldsymbol{i})$

3. Second-order message passing

- Can compute feature covariance using message passing when graph is a tree
- Use special semiring in place of sum-product
- Linear in number of nodes
- Quadratic in dimension of diversity features ϕ

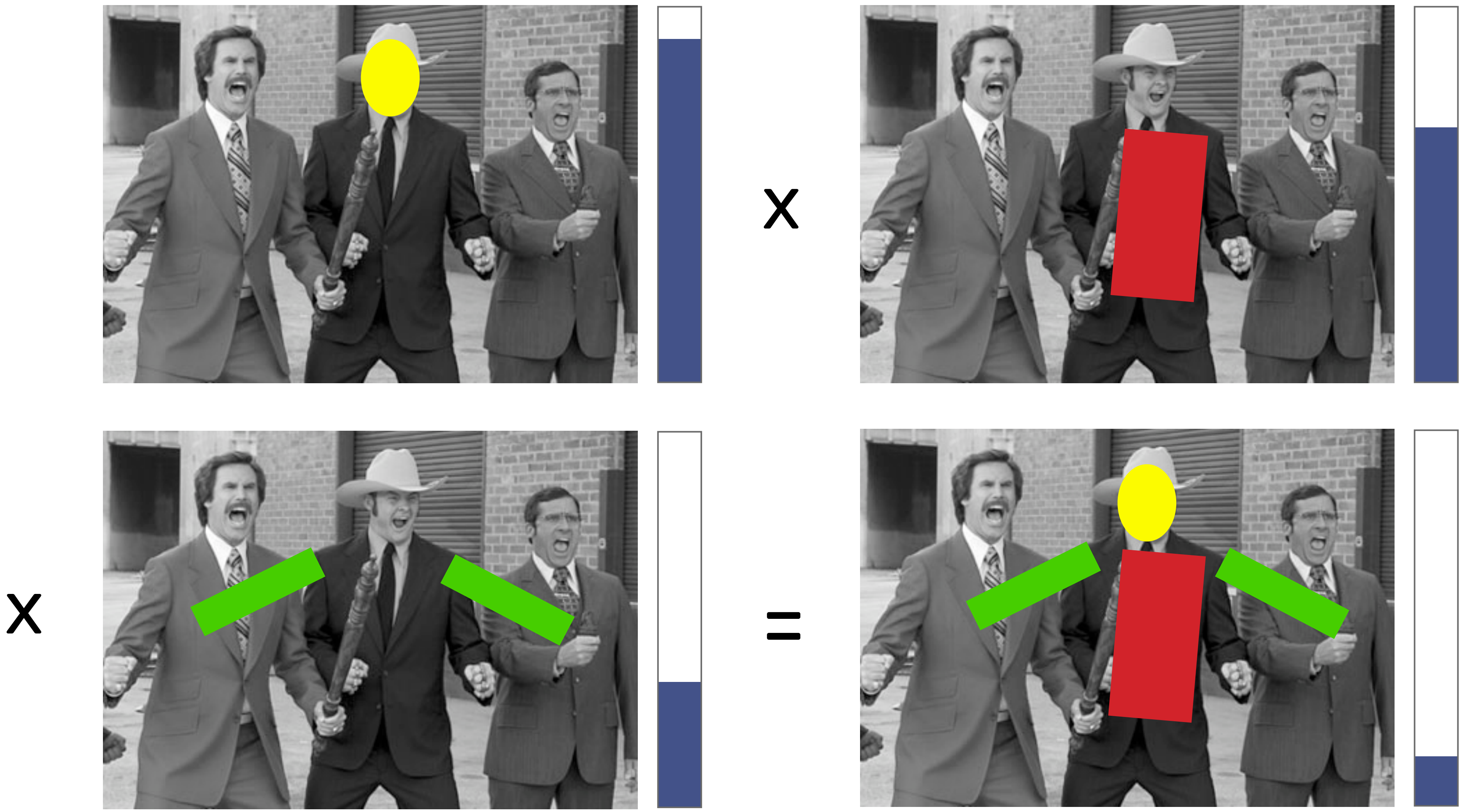
[Li + Eisner, 2009]

Multiple-pose estimation



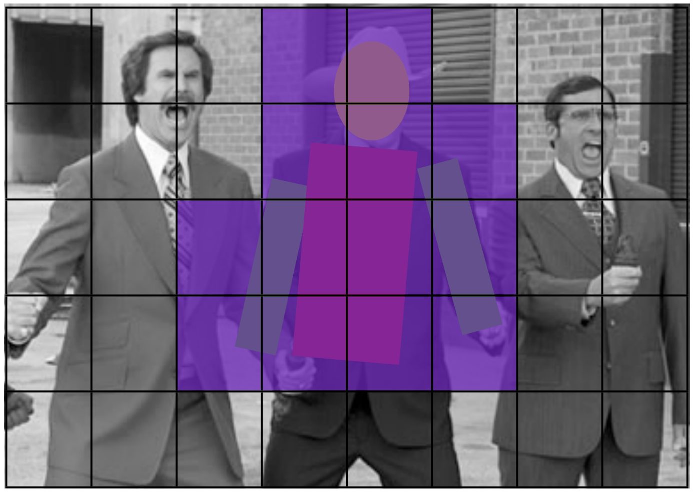
- Images from TV shows
  - 3+ people/image, similar scale, hand labeled
- Trained quality model, spatial diversity model

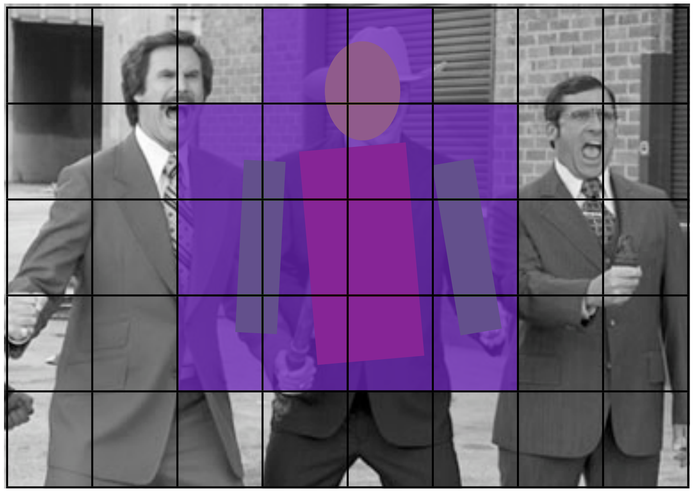
Quality



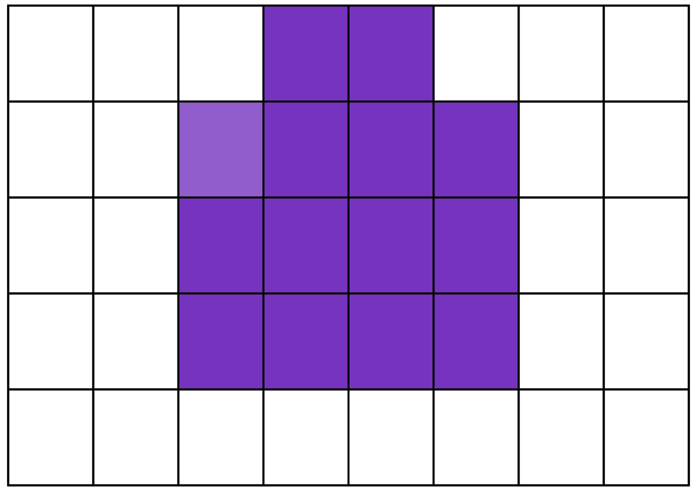
Diversity

Diversity



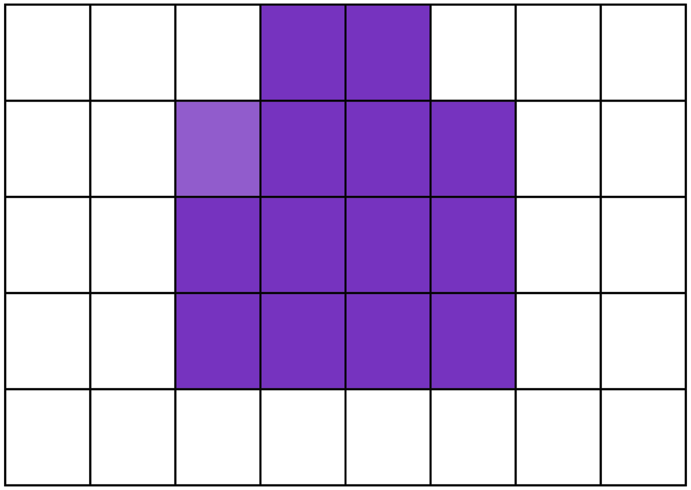


Diversity



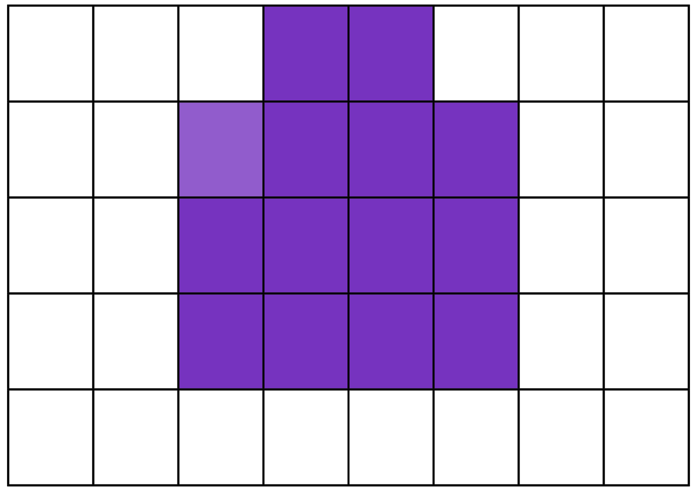
Low diversity

Diversity

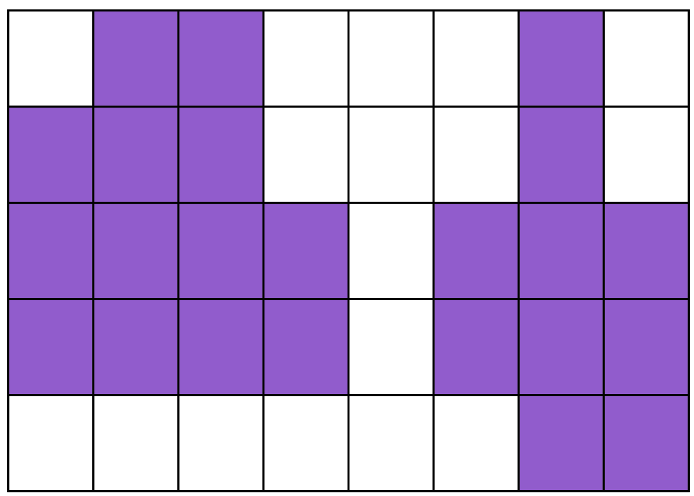


Low diversity

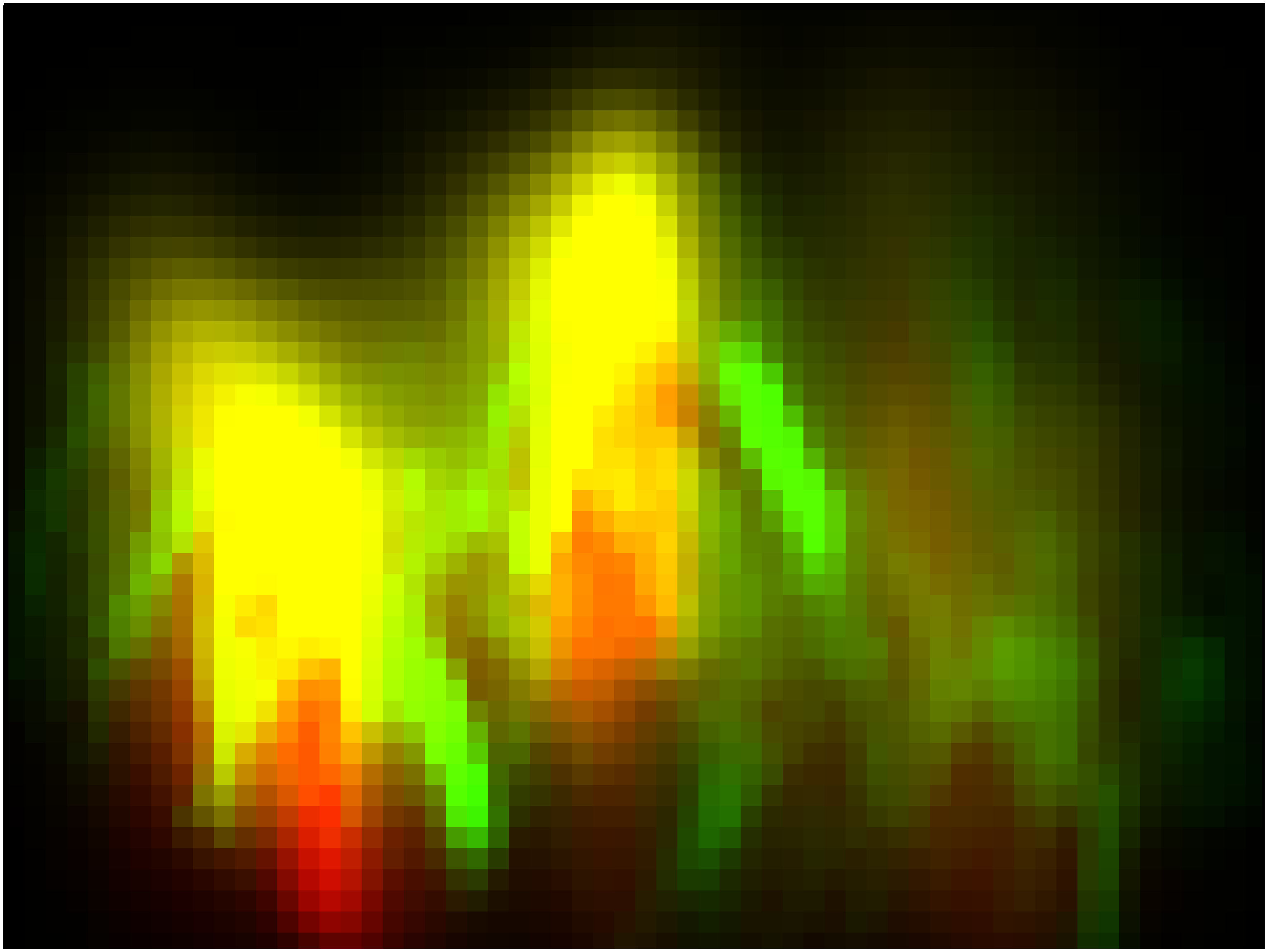
Diversity

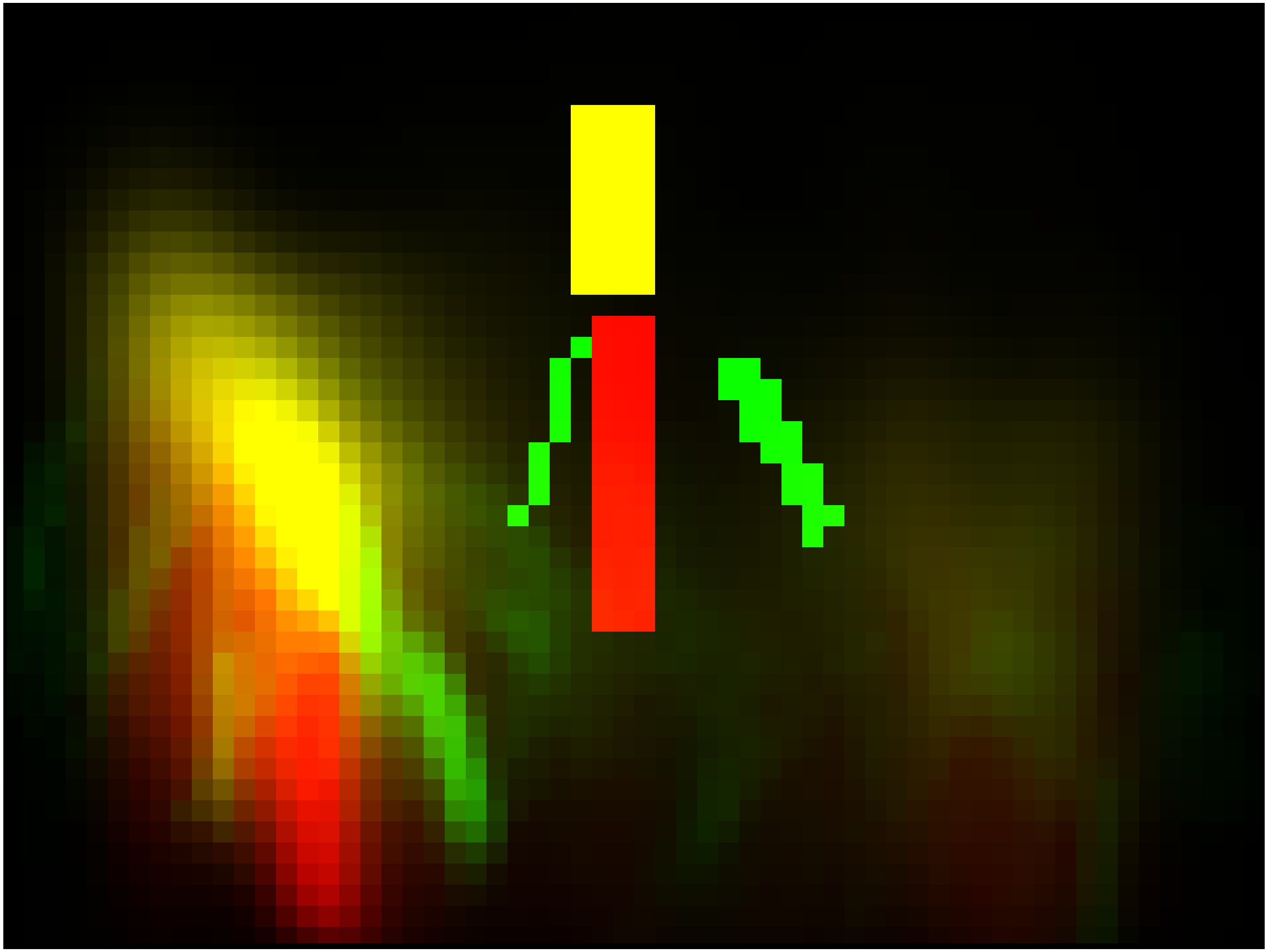


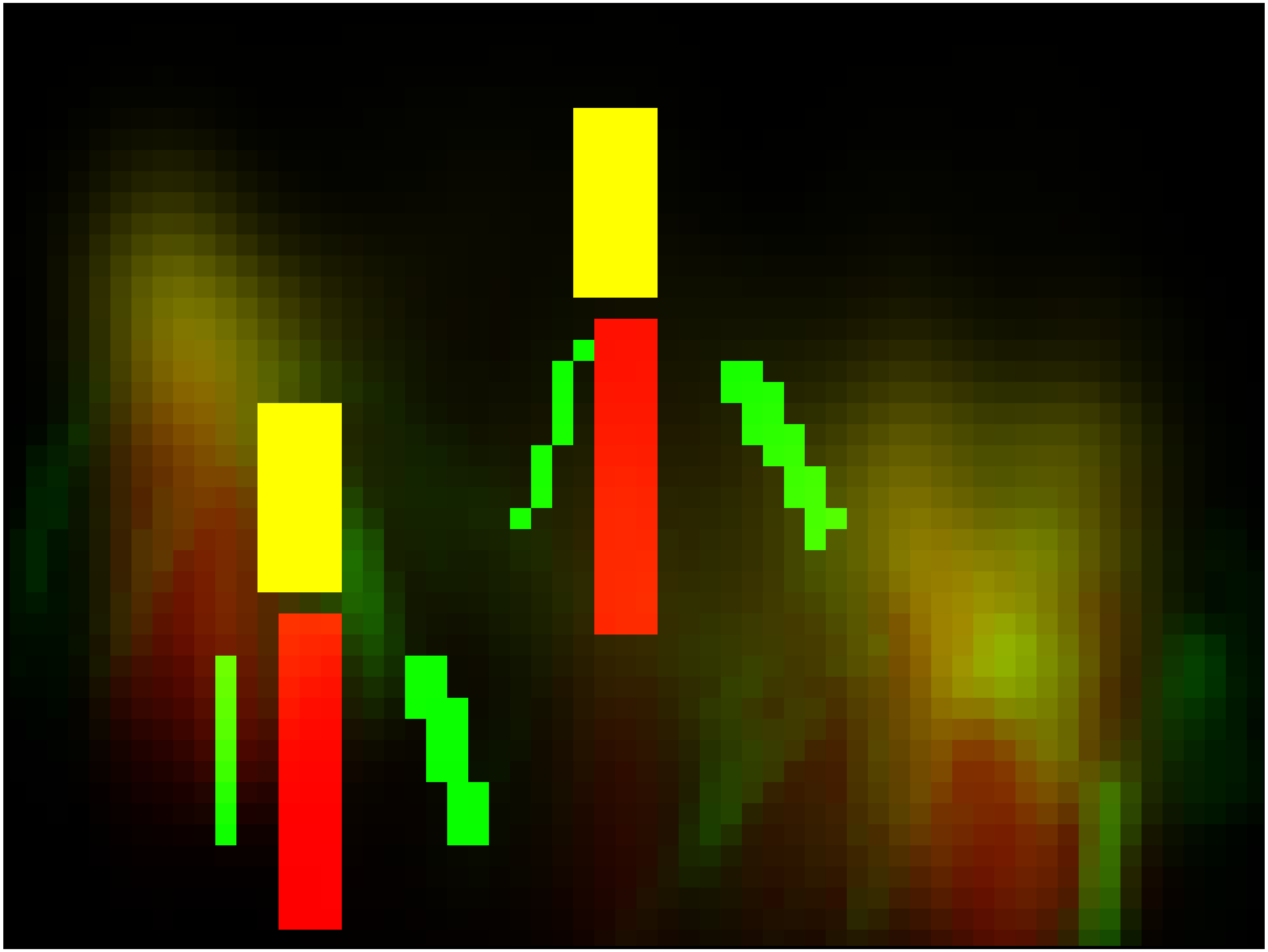
Low diversity

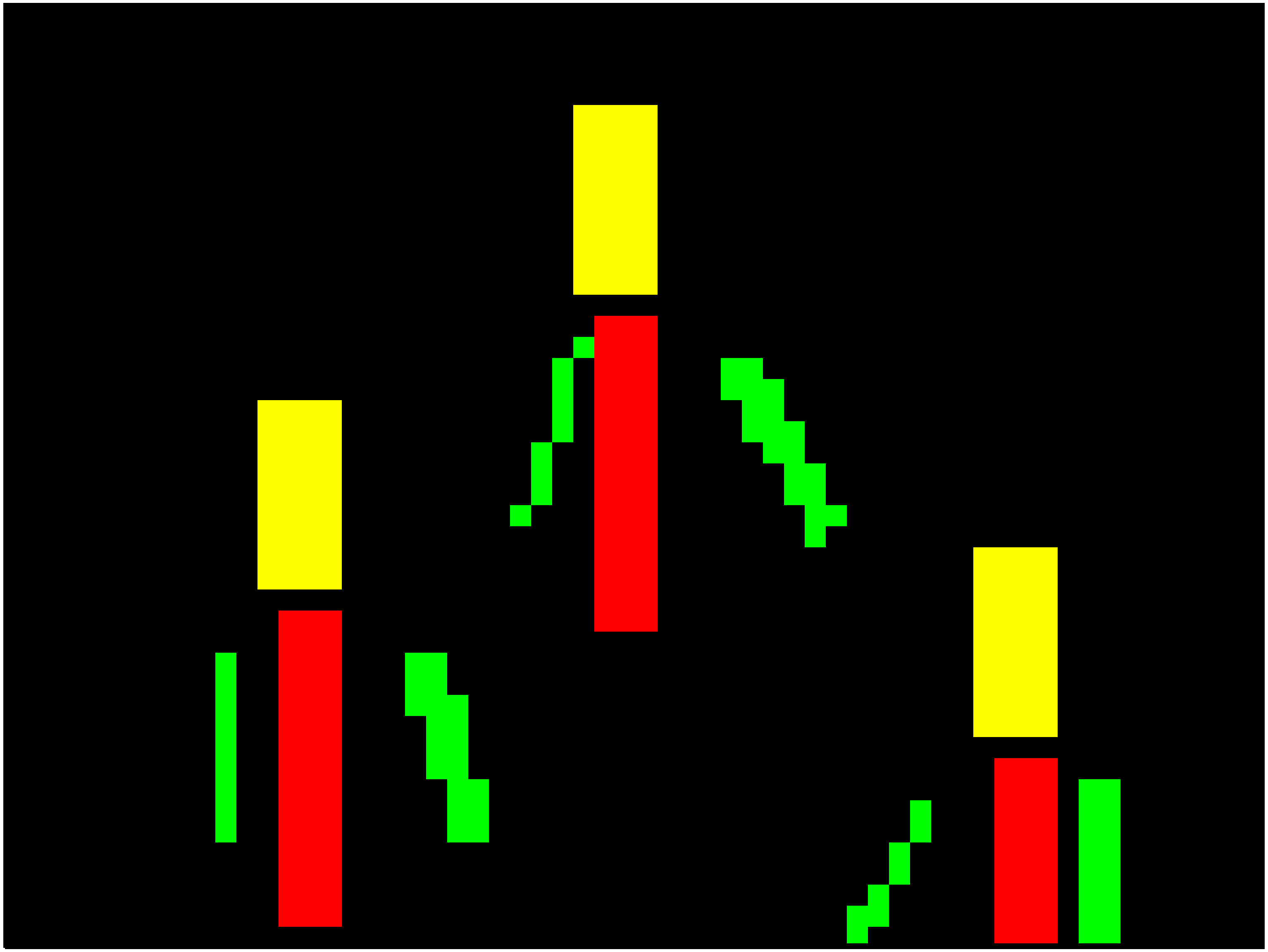


High diversity

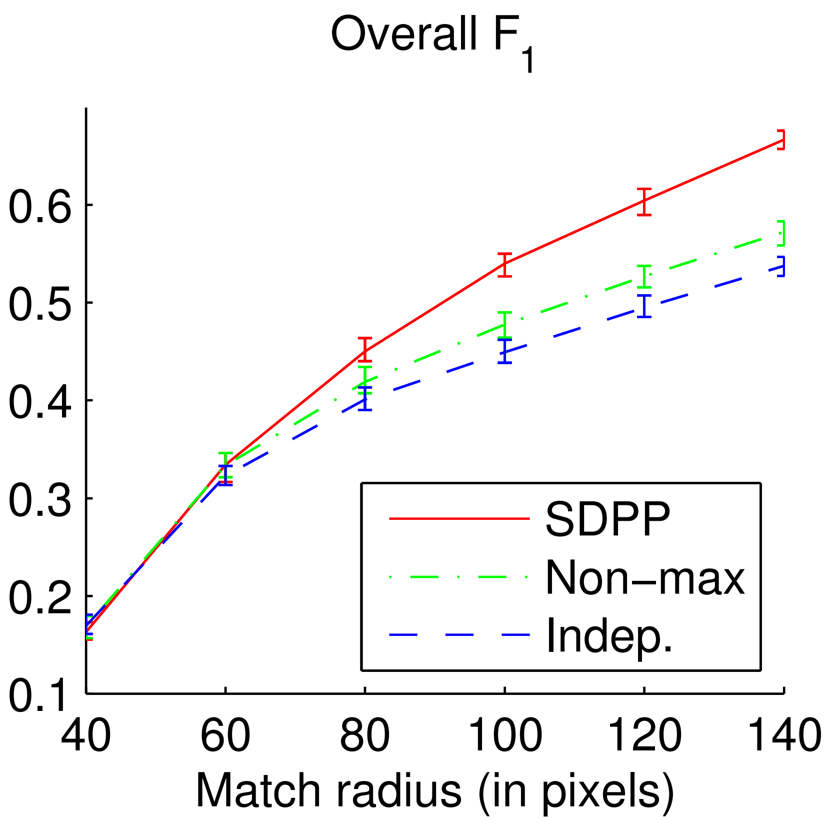


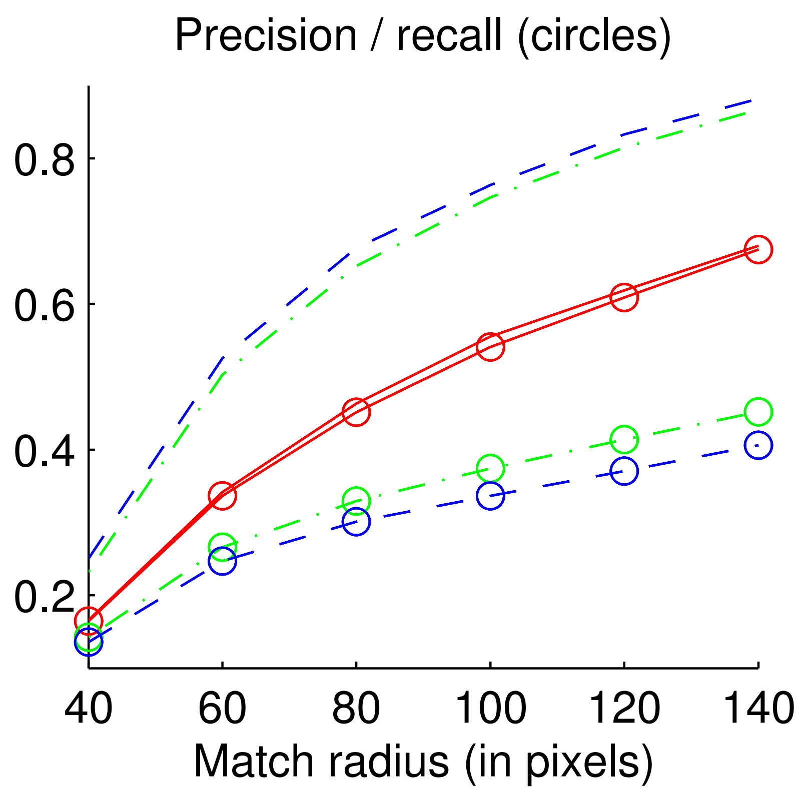






Pose accuracy





Part II

Large-scale DPPs

k-DPPs

Structured DPPs

News threading

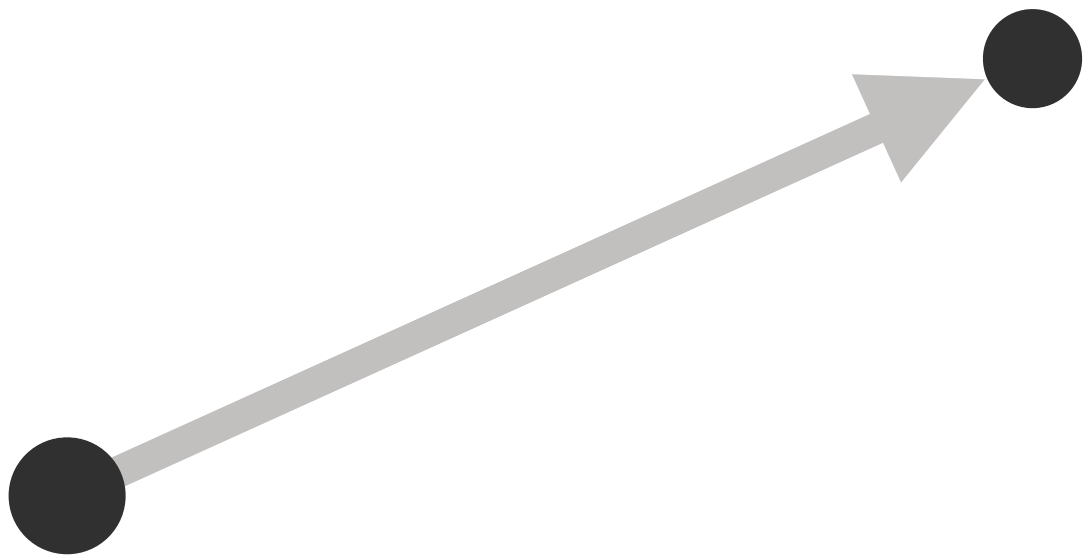
Conclusion

News threading

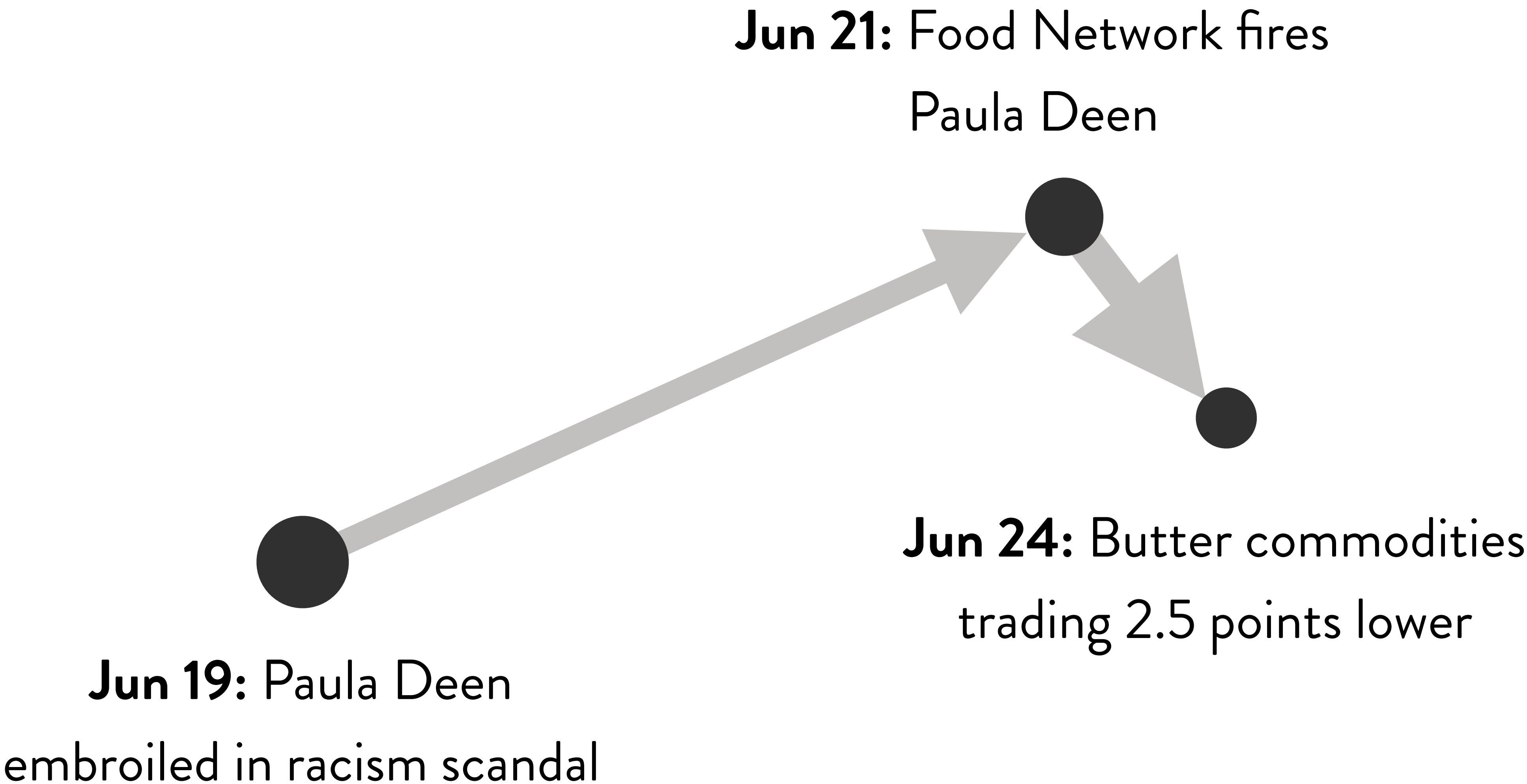
- **Input**: large news corpus
- **Output**: threads of articles
  - Each thread narrates a major story
  - Threads are diverse to cover many stories
- Combine *k*-DPPs, structured DPPs, dual DPPs, and random projection

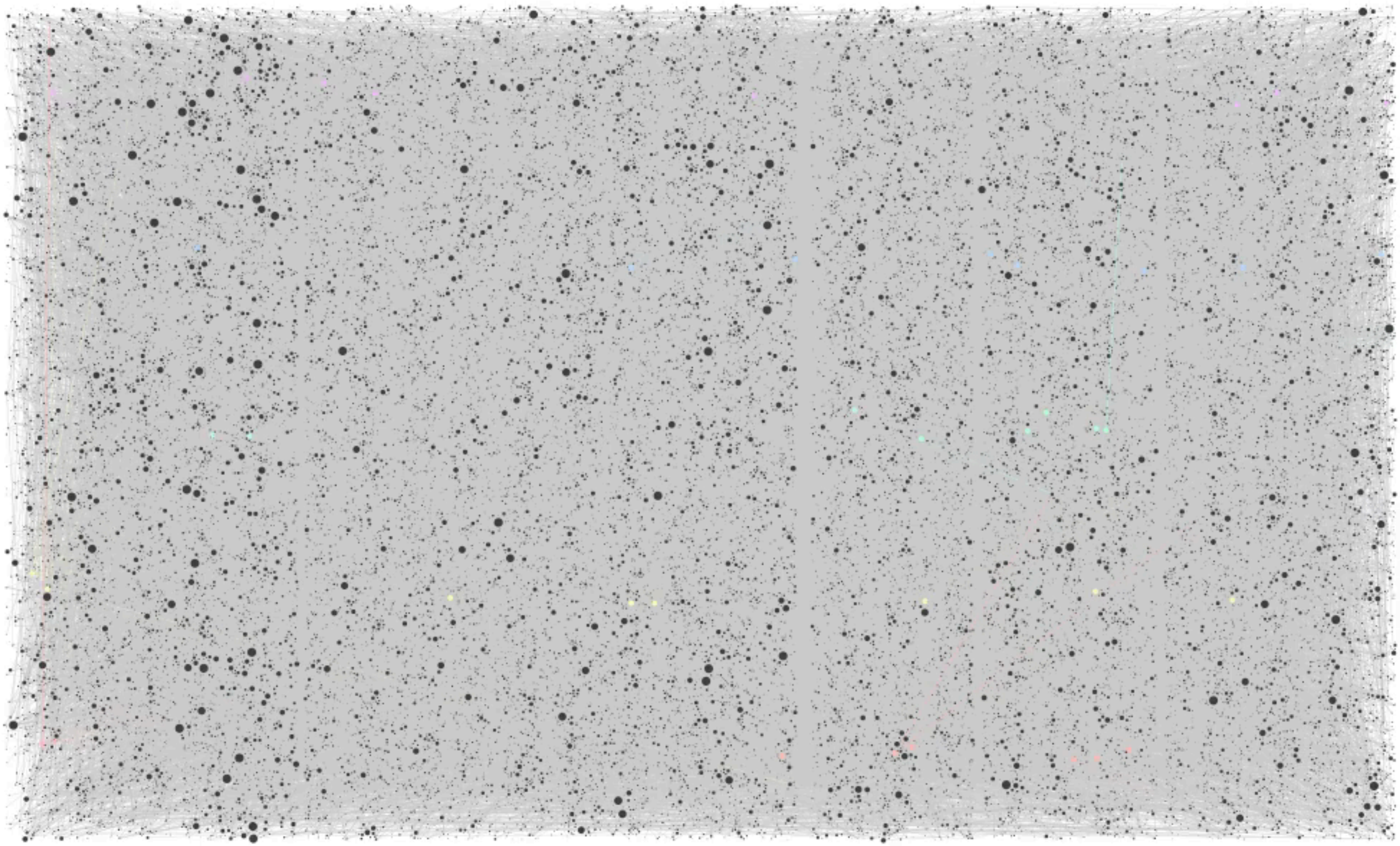
Jun 19: Paula Deen embroiled in racism scandal

Jun 21: Food Network fires Paula Deen

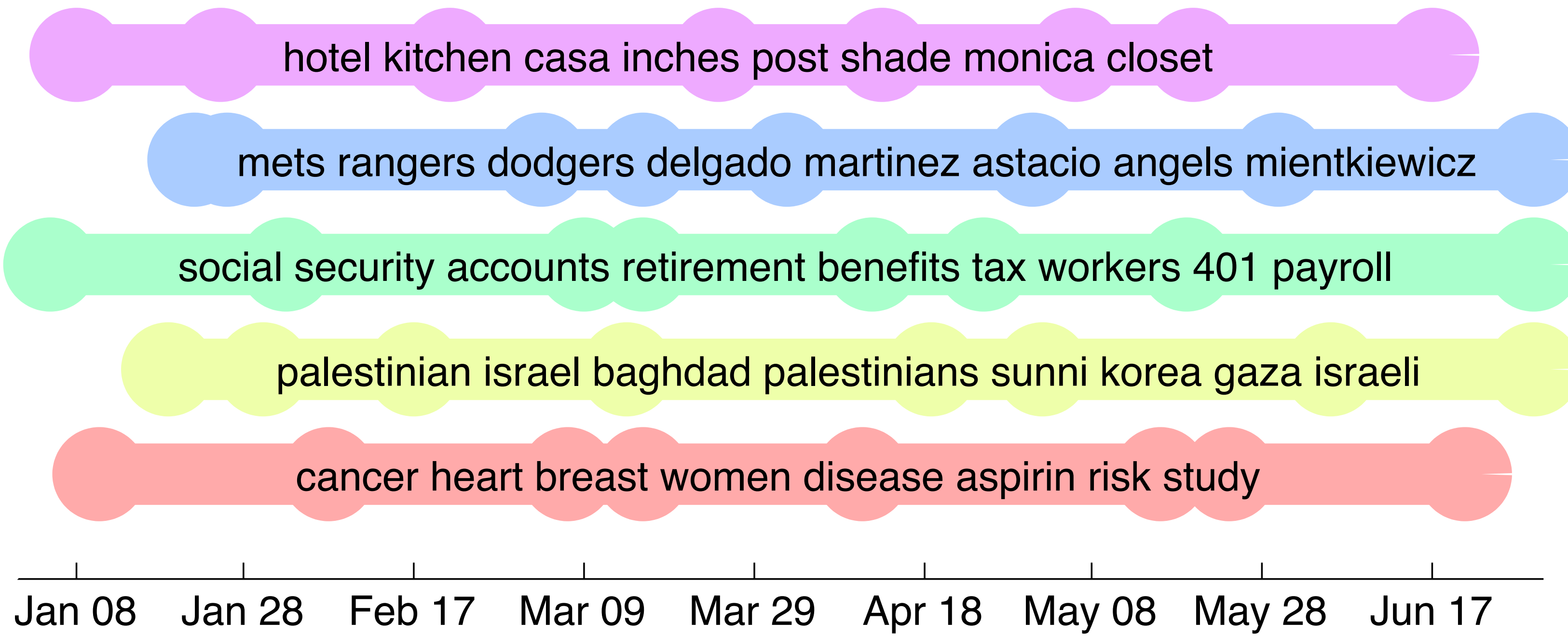


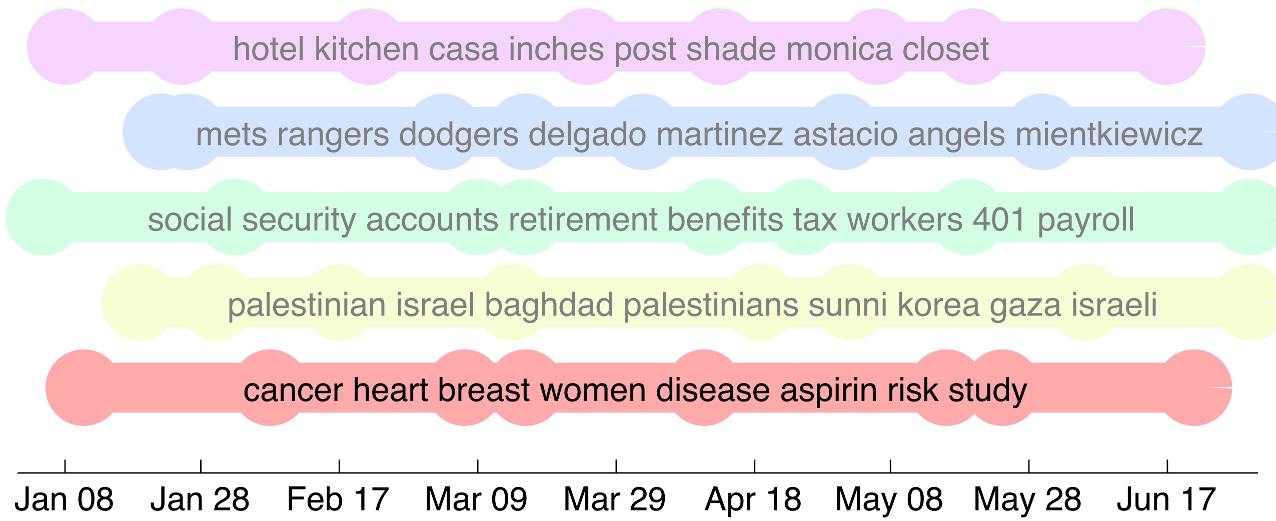
Jun 19: Paula Deen embroiled in racism scandal


Jun 21: Food Network fires Paula Deen
Jun 24: Butter commodities trading 2.5 points lower
Jun 19: Paula Deen embroiled in racism scandal



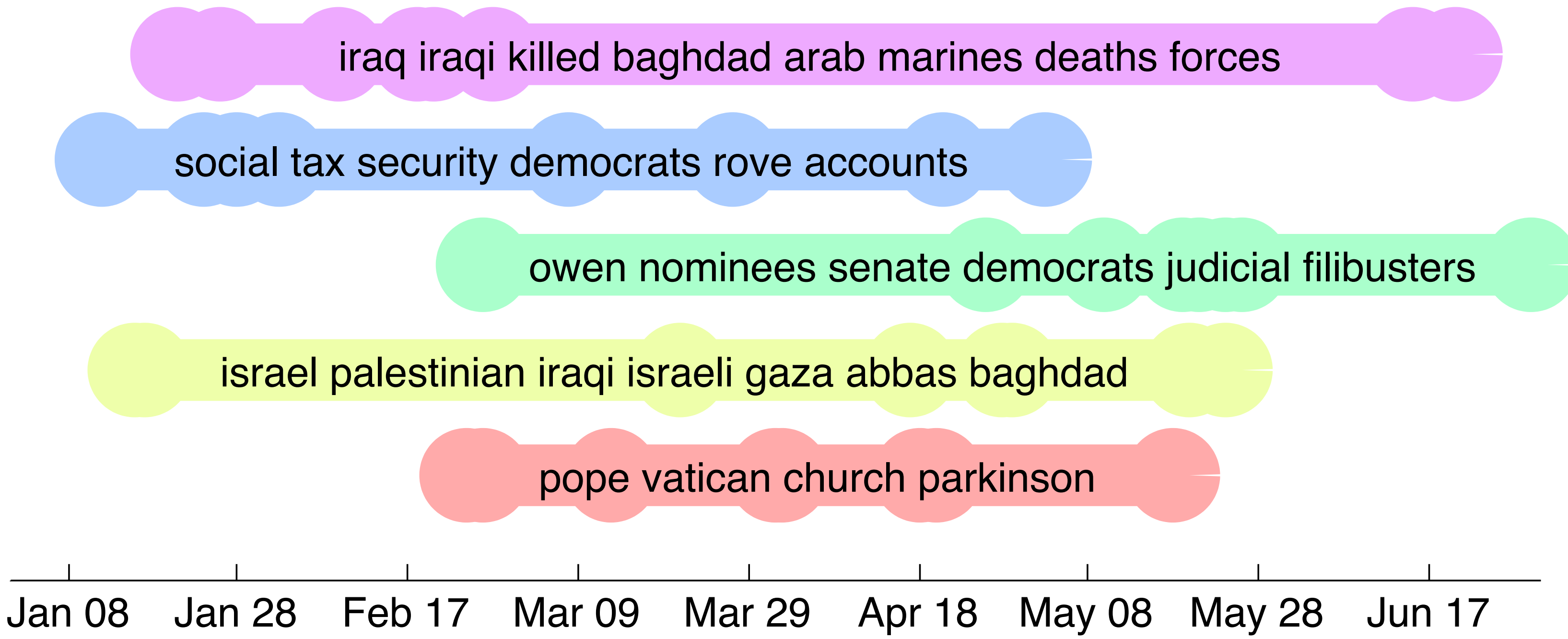
Dynamic topic model

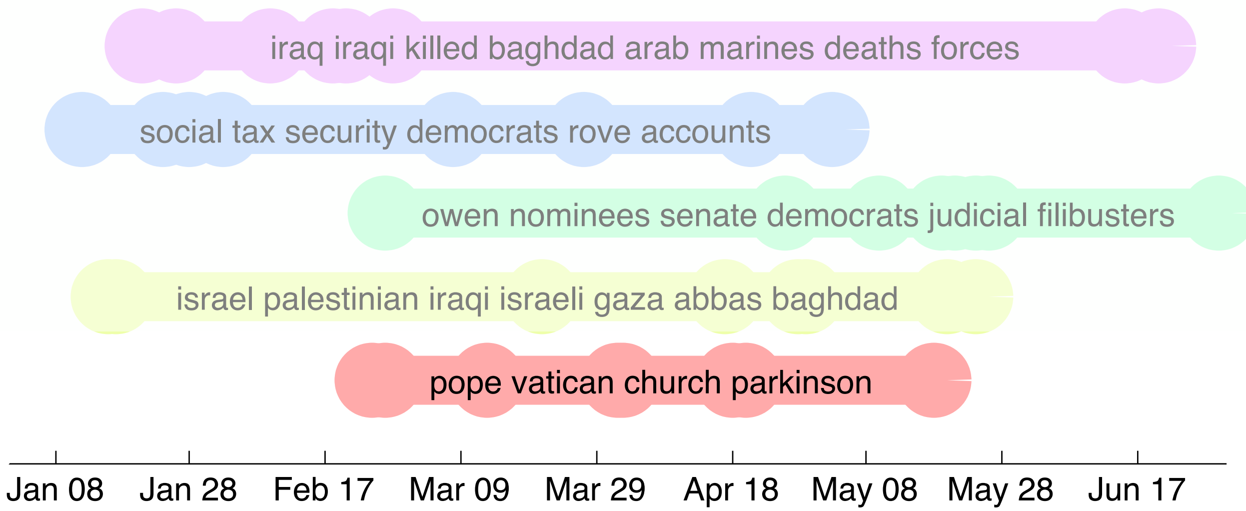




Jan 11: Study Backs Meat, Colon Tumor Link
Feb 07: Patients Still Don't Know How Often Women Get Heart Disease
Mar 07: Aspirin Therapy Benefits Women, but Not the Way It Aids Men
Mar 16: Radiation Therapy Doesn't Increase Heart Disease Risk
Apr 11: Personal Health: Women Struggle for Parity of the Heart
May 16: Black Women More Likely to Die from Breast Cancer
May 24: Studies Bolster Diet, Exercise for Breast Cancer Patients
Jun 21: Another Reason Fish is Good for You

DPP threads


iraq iraqi killed baghdad arab marines deaths forces
social tax security democrats rove accounts
owen nominees senate democrats judicial filibusters
israel palestinian iraqi israeli gaza abbas baghdad
pope vatican church parkinson
Jan 08
Jan 28
Feb 17
Mar 09
Mar 29
Apr 18
May 08
May 28
Jun 17



Feb 24: Parkinson's Disease Increases Risks to Pope
Feb 26: Pope's Health Raises Questions About His Ability to Lead
Mar 13: Pope Returns Home After 18 Days at Hospital
Apr 01: Pope's Condition Worsens as World Prepares for End of Papacy
Apr 02: Pope, Though Gravely Ill, Utters Thanks for Prayers
Apr 18: Europeans Fast Falling Away from Church
Apr 20: In Developing World, Choice [of Pope] Met with Skepticism
May 18: Pope Sends Message with Choice of Name

Scale

- ~35,000 articles per six month time period
- About 10^{360} possible sets of threads
- *D* = 36,356-dimensional diversity features
- Naively, requires 1600 TB of memory
- Use random projection to make it efficient

Evaluation

- Gold timelines too expensive
  - Human news summaries to evaluate **content**
  - amazonmechanical turk (beta, Artificial Artificial Intelligence) to evaluate thread **quality**

Results: Human summaries & ratings

System	*k*-means	DTM	*k*-SDPP
ROUGE-1F	16.5	14.7	**17.2**
R-SU4F	3.76	3.44	**3.98**
Coherence	2.73	3.19	**3.31**
Interlopers	0.71	1.10	1.15
Runtime (s)	626	19,434	**252**

Part II

Large-scale DPPs

k-DPPs

Structured DPPs

News threading

Conclusion

- DPPs model **global**, **negative** correlations
- Efficient inference:
  - normalization
  - marginals
  - conditioning
  - sampling
- Extensions make DPPs useful for modeling and learning from large-scale real-world data

Food Processing

Dirichlet Process, aka
Chinese Restaurant Process

Beta-Bernouli Process, aka
Indian Buffet Process

Determinantal Process, aka
Antisocial Coffeeshop Process

Supporting Materials

- Tech report (120 pages, with all the proofs!) http://arxiv.org/abs/1207.6083
- Matlab Code: http://www.eecs.umich.edu/~kulesza/code/dpp.tgz