Geometry of Diversity and Determinantal Point Processes: Representation, Inference and Learning

Ben Taskar

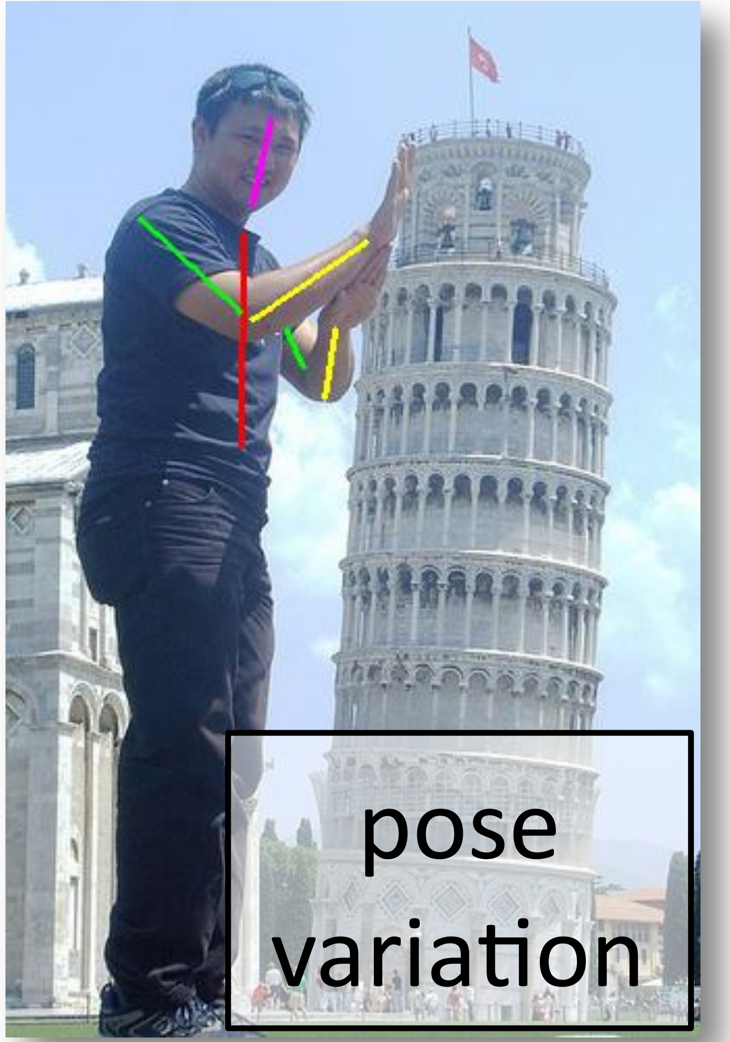
Joint work with Jennifer Gillenwater and Alex Kulesza

University of Pennsylvania


CARLOW UNIVERSITY
SAFER DATA MINING
Map Reduce
Map Reuse
Map Recycle
Green Data Processing
THE FREE AIN'T FREE
FREE VARIABLES!
BAYeSIANS AGAINST
SUPPORT VECTOR MACHINES

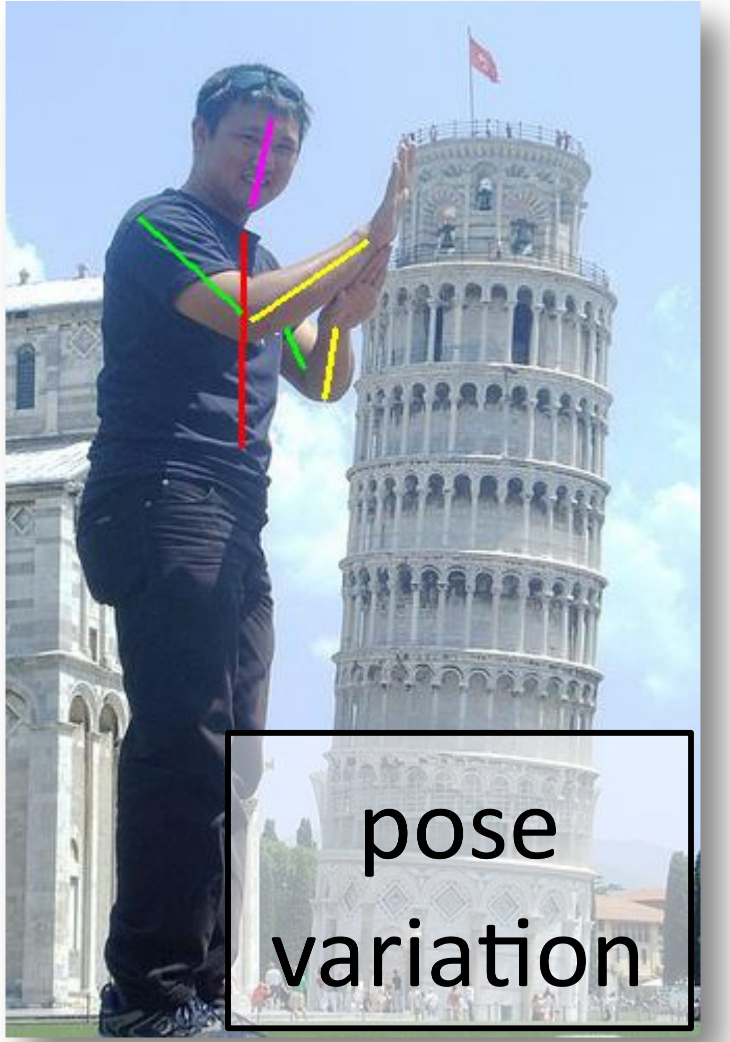

Eichner & Ferrari BMVC09

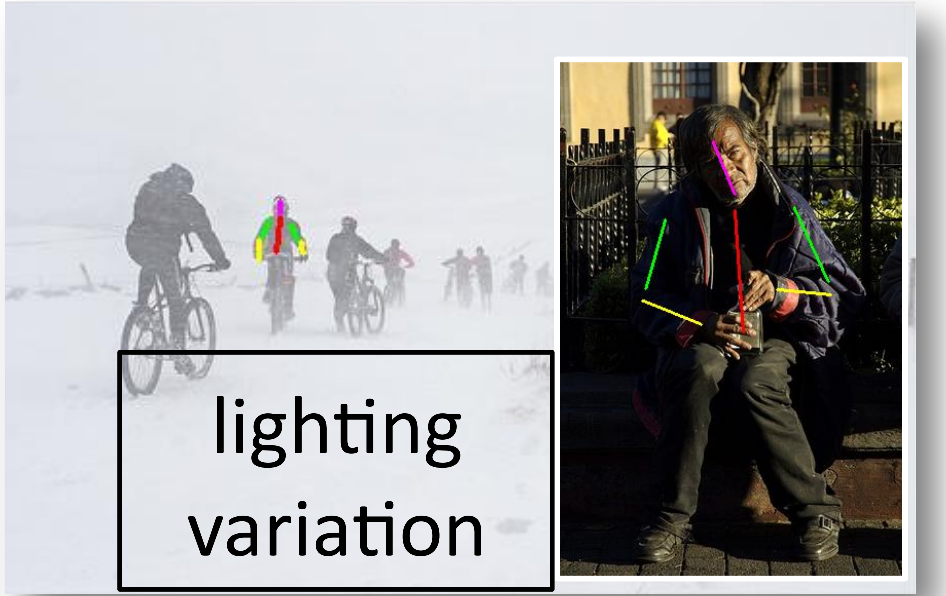
Human pose estimation: what's so hard?

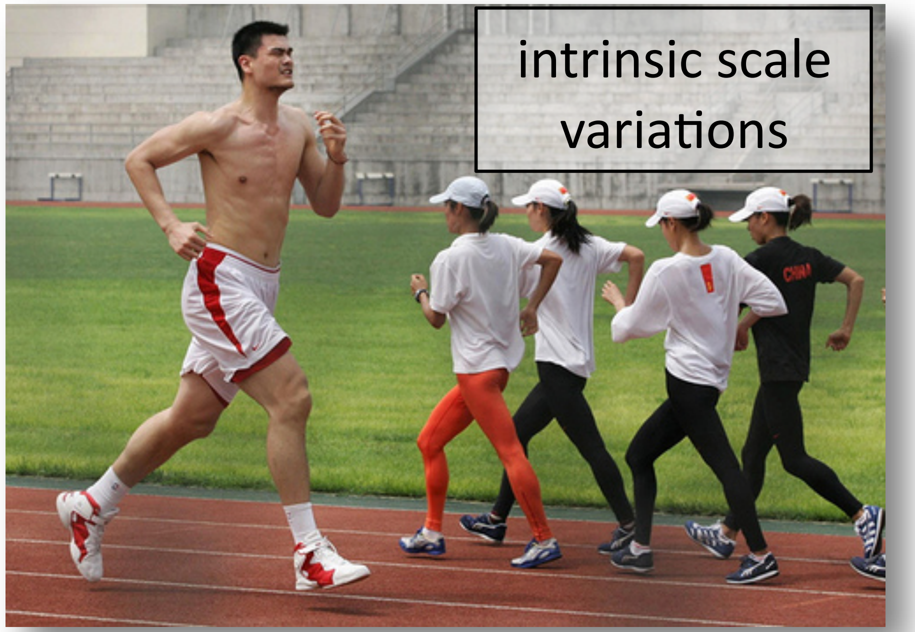


Human pose estimation: what's so hard?

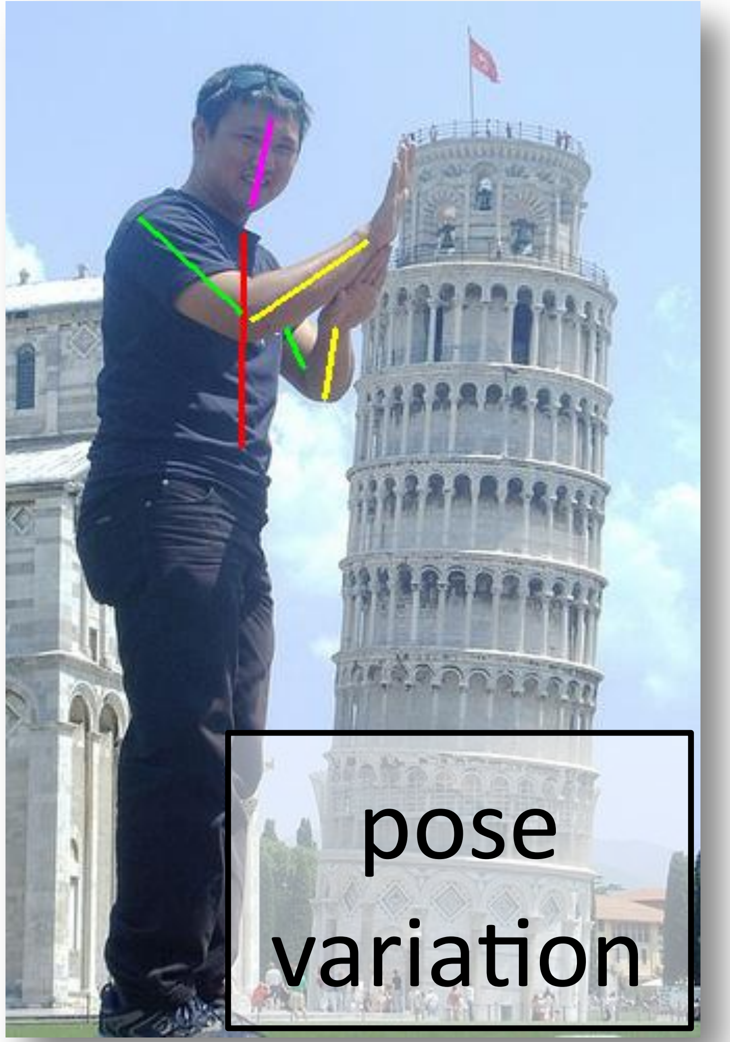
Human pose estimation: what's so hard?

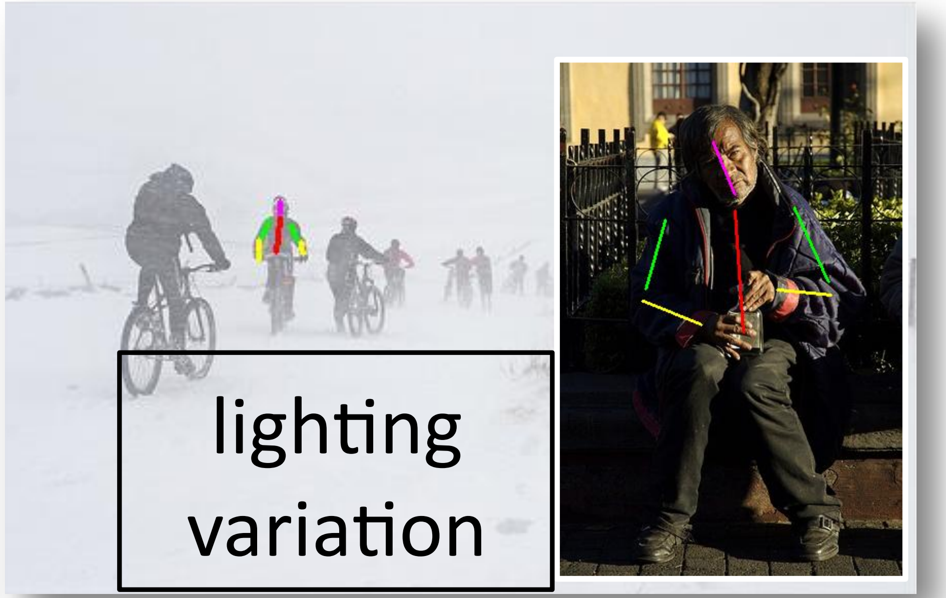


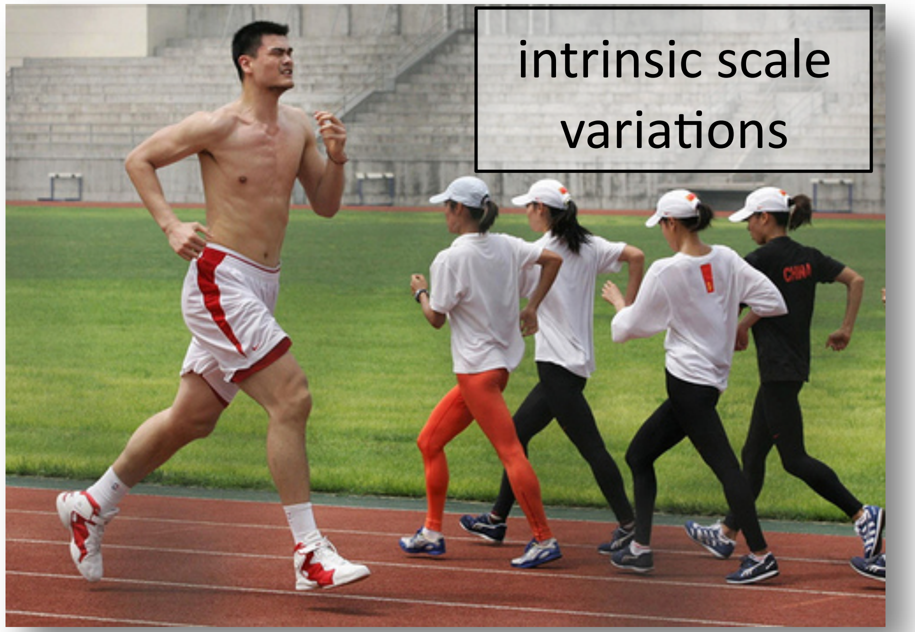




Human pose estimation: what's so hard?



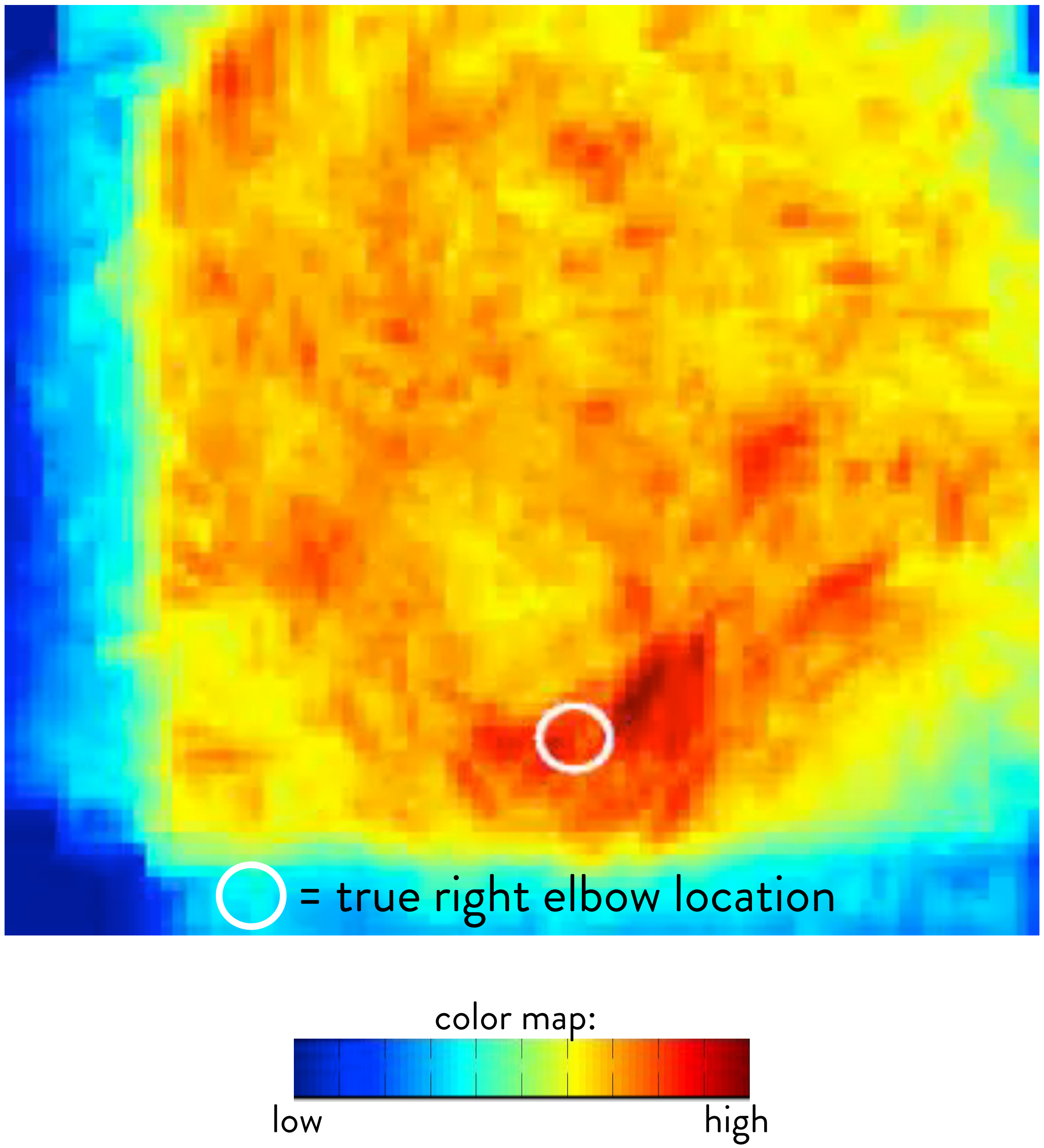




Human pose estimation: what's so hard?

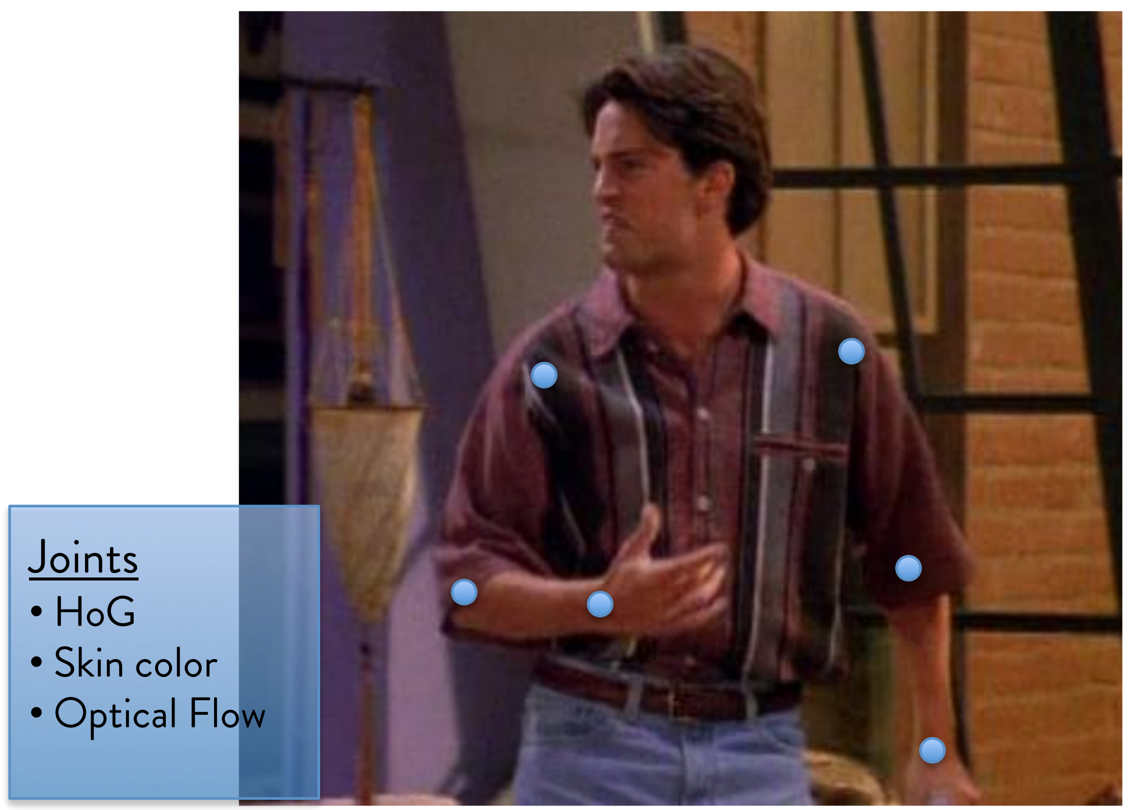
Local signal is weak

State-of-the-art right elbow detector
[HoG+SVM+etc]



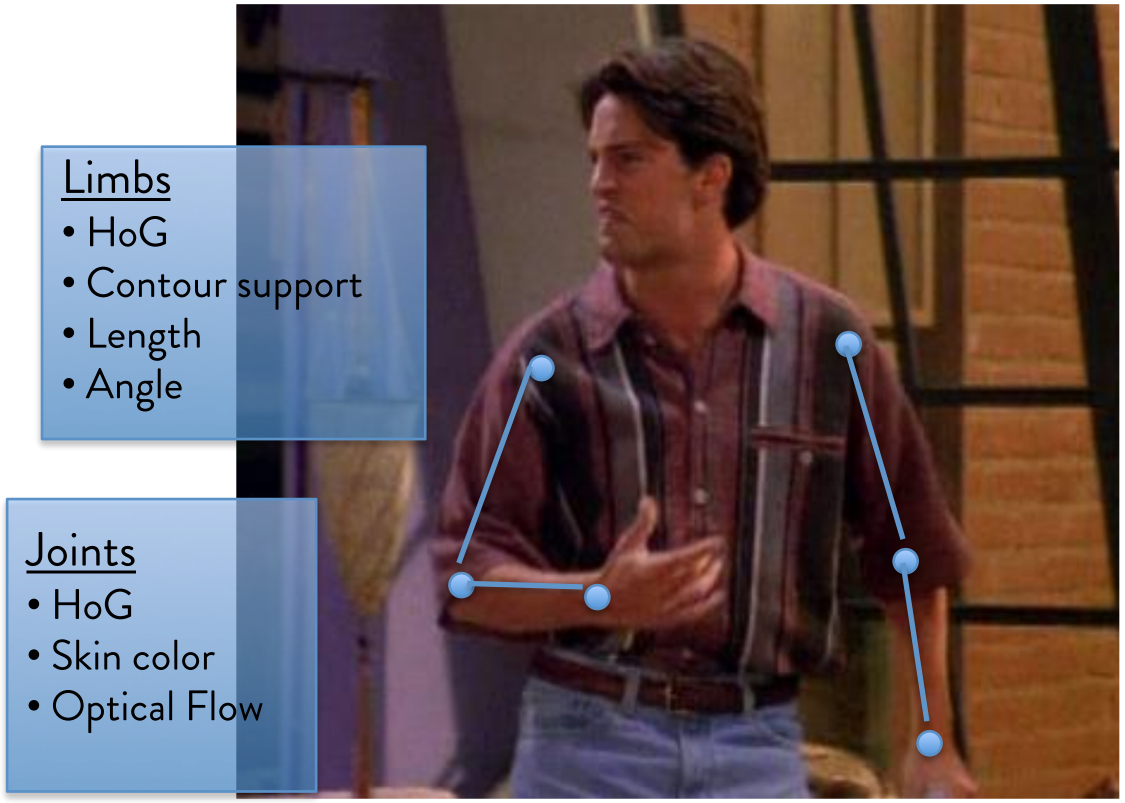
Graphical Models vs. Tyranny of Small Decisions

- Detecting joints and parts in isolation is hard
- Need to capture relationships *between* joints



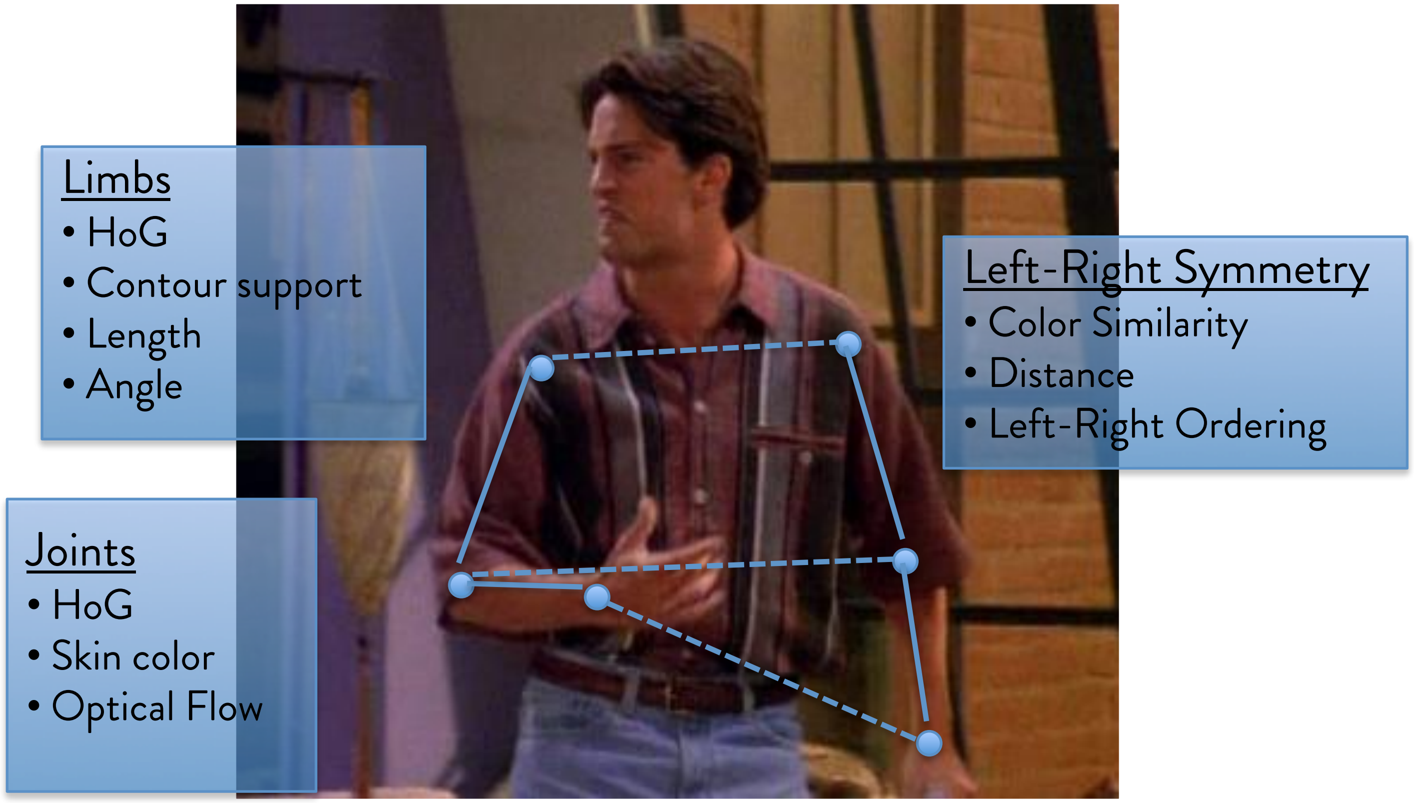
Graphical Models vs. Tyranny of Small Decisions

- Detecting joints and parts in isolation is hard
- Need to capture relationships *between* joints



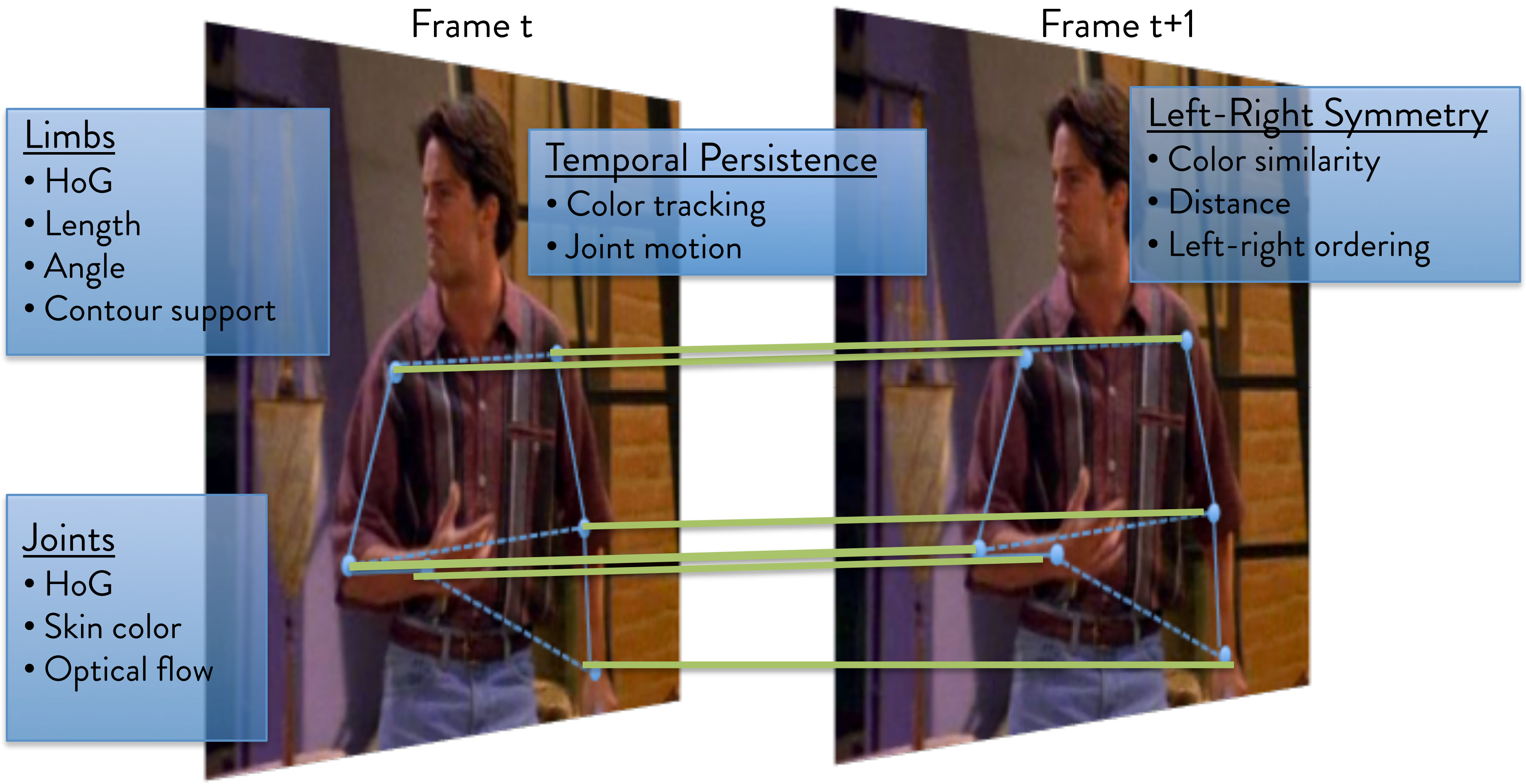
Graphical Models vs. Tyranny of Small Decisions

- Detecting joints and parts in isolation is hard
- Need to capture relationships *between* joints



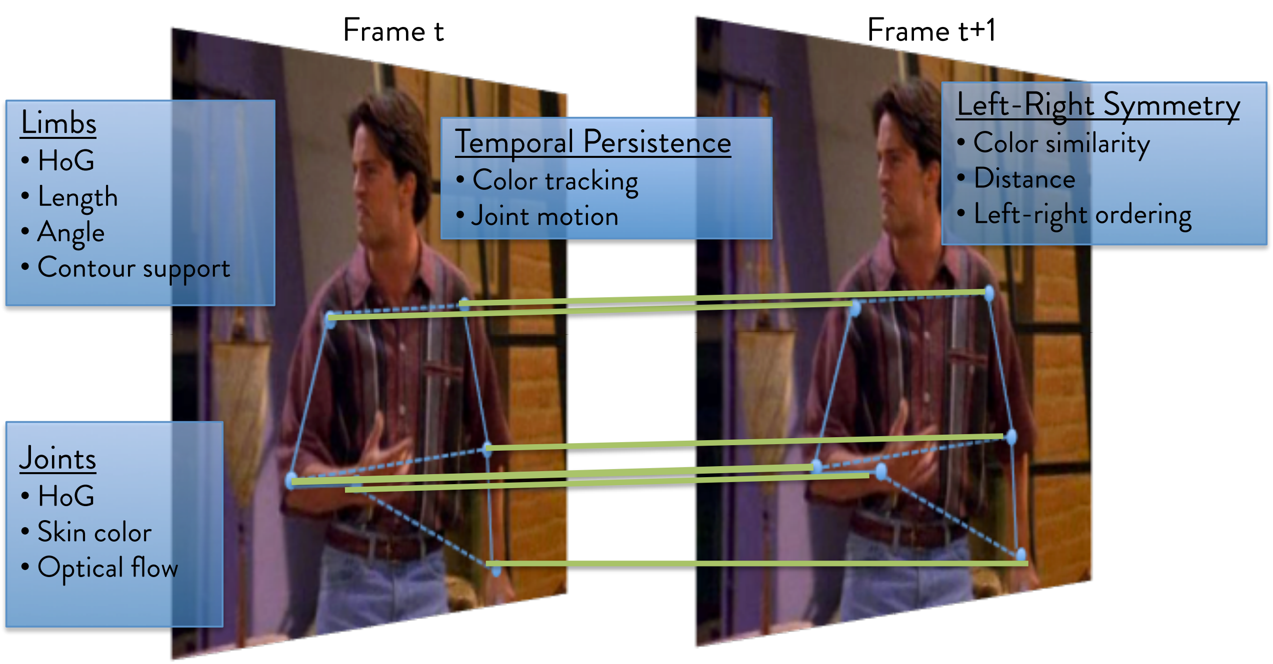
Graphical Models vs. Tyranny of Small Decisions

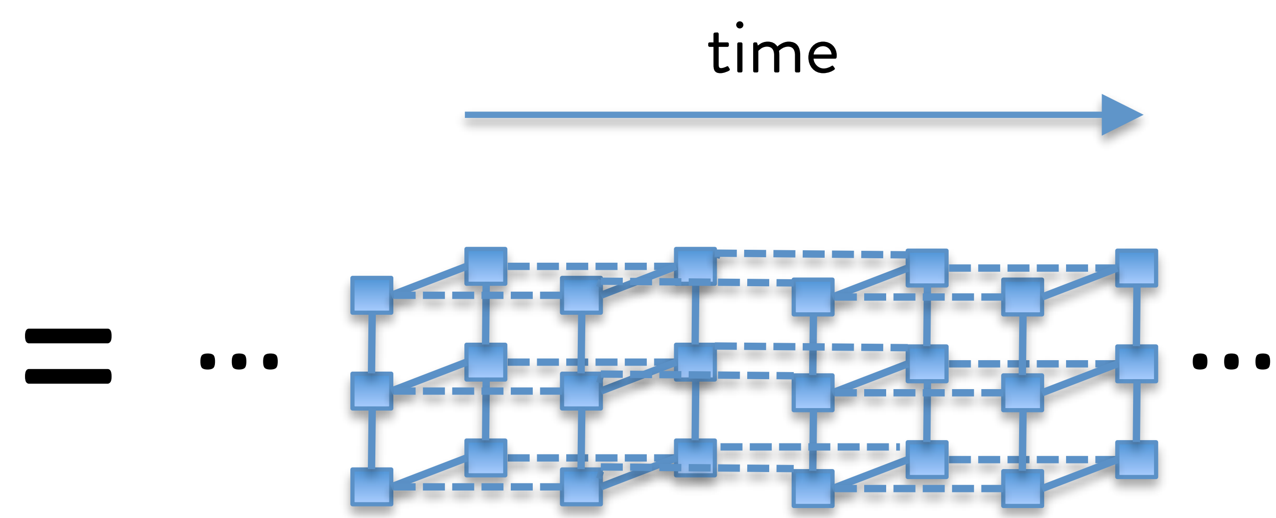
- Detecting joints and parts in isolation is hard
- Need to capture relationships *between* joints



Graphical Models vs. Tyranny of Small Decisions

- Detecting joints and parts in isolation is hard
- Need to capture relationships *between* joints



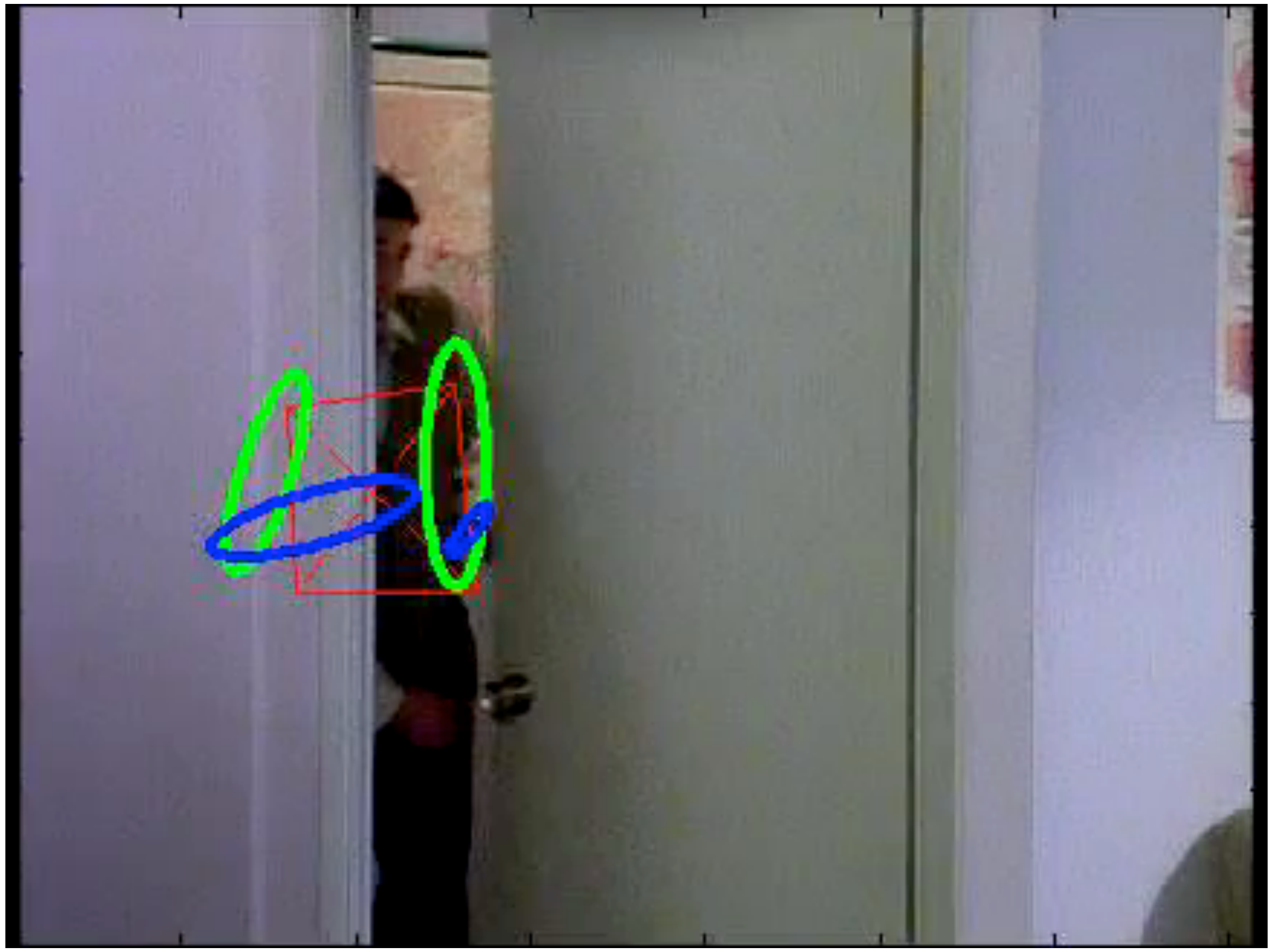


$$P(\mathbf{y}) \propto \prod_{i \in \mathcal{V}} \phi_i(y_i) \prod_{ij \in \mathcal{E}} \phi_{ij}(y_i, y_j)$$

Graphical Models vs. Tyranny of Small Decisions

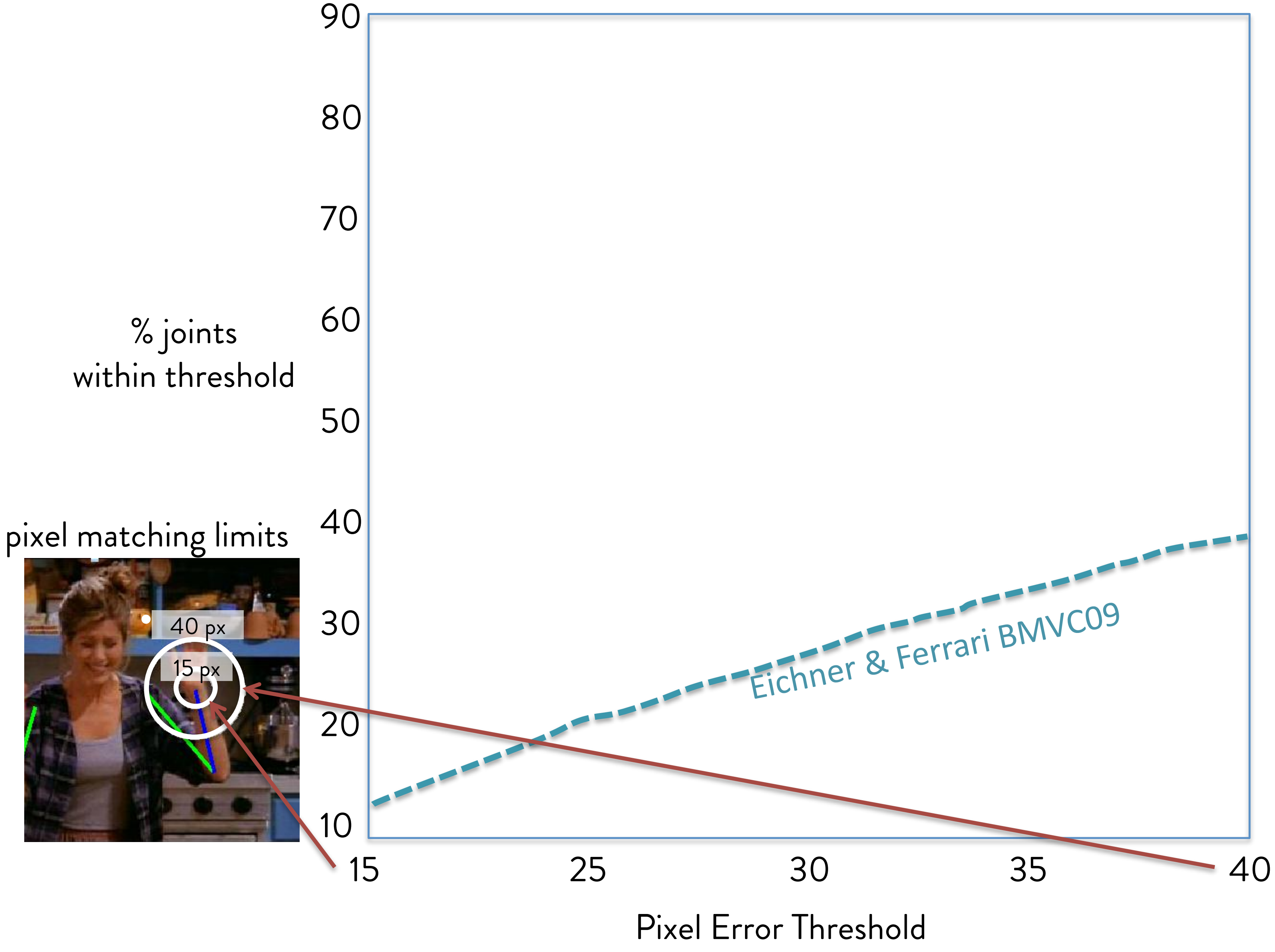
- Detecting joints and parts in isolation is hard
- Need to capture relationships *between* joints

Eichner & Ferrari BMVC09



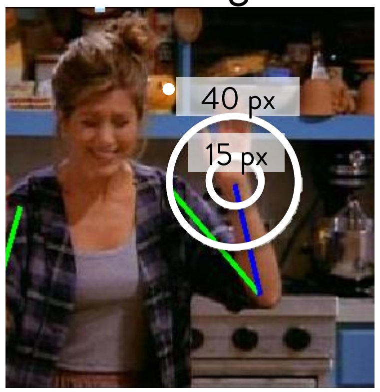
Sapp, Weiss & Taskar CVPR11

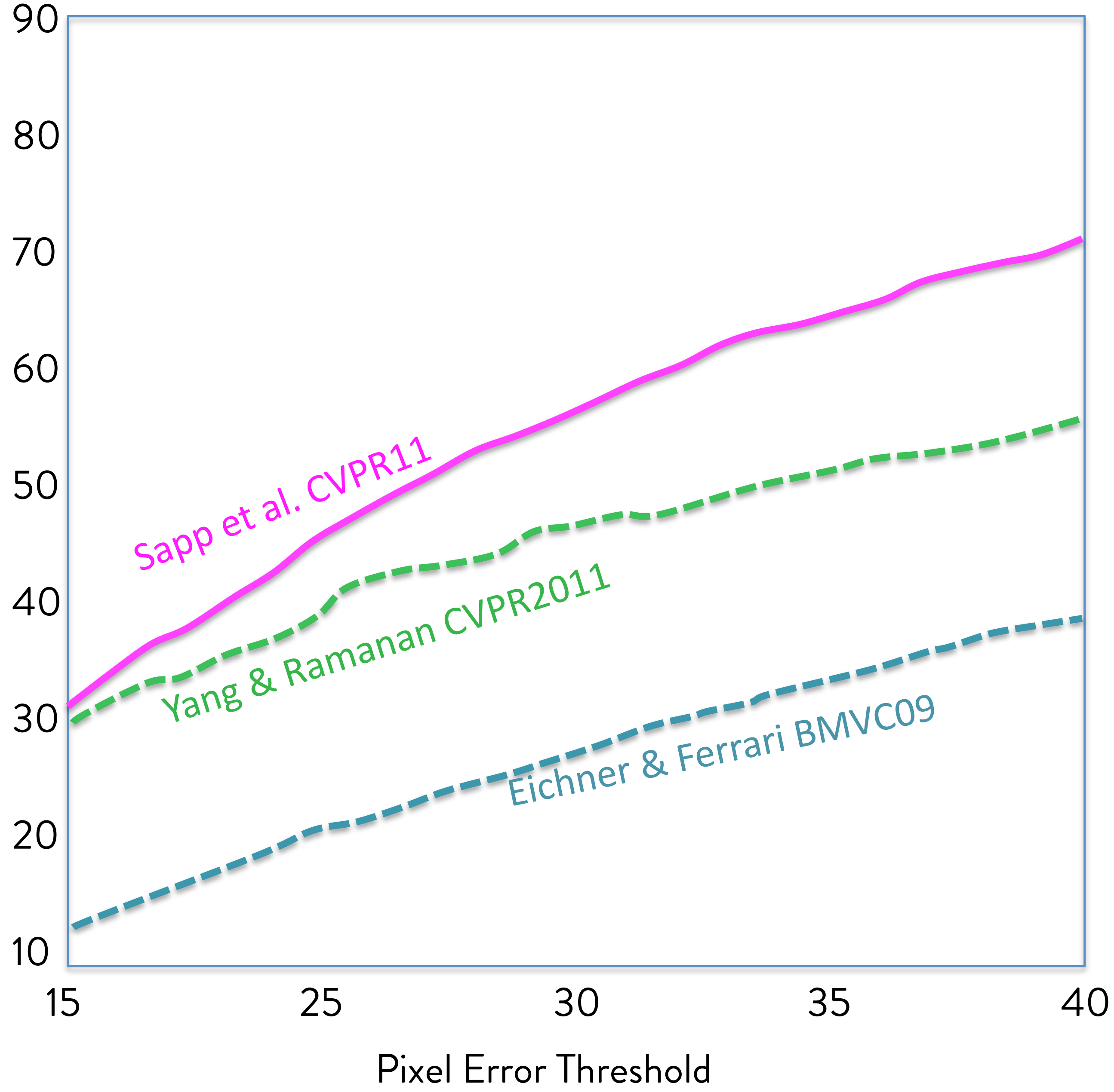
Accuracy of Wrist Localization



Accuracy of Wrist Localization

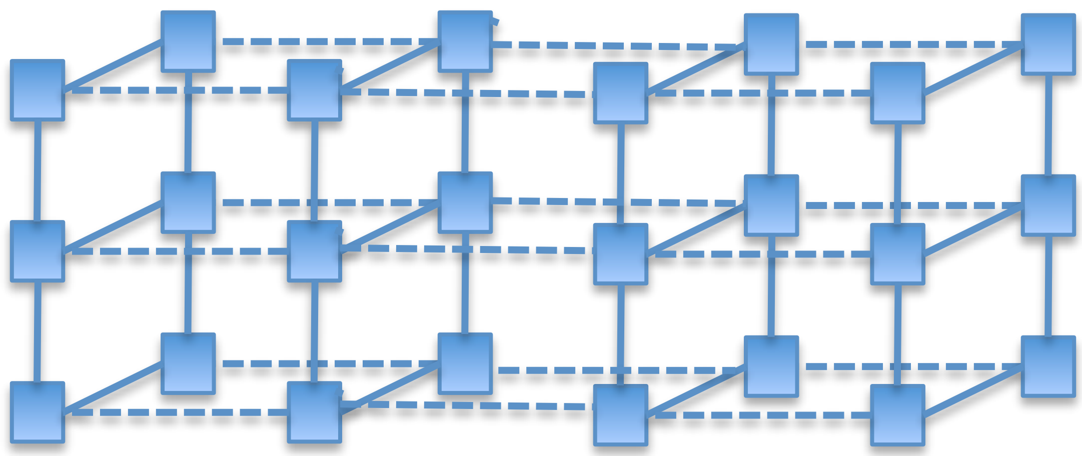
pixel matching limits





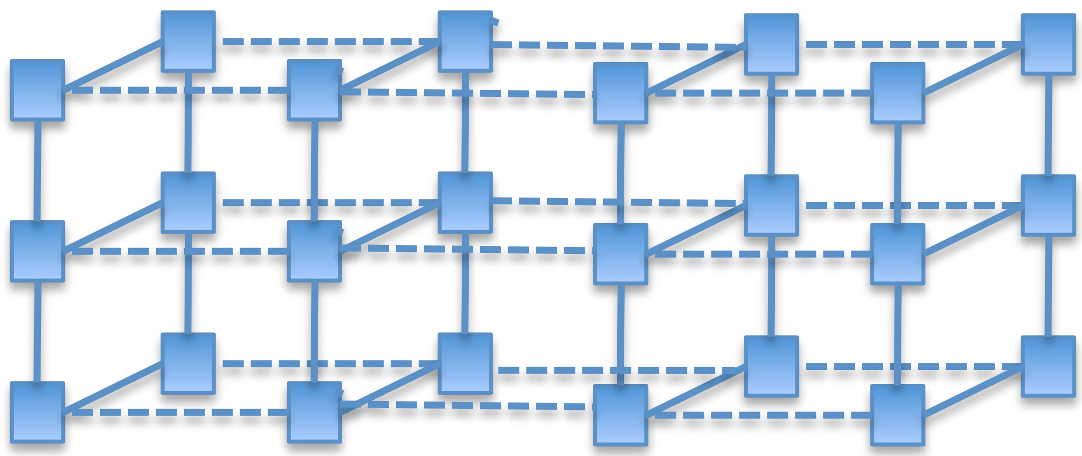
The Catch

- Problem: inference exponential in # of joints (10^{24})



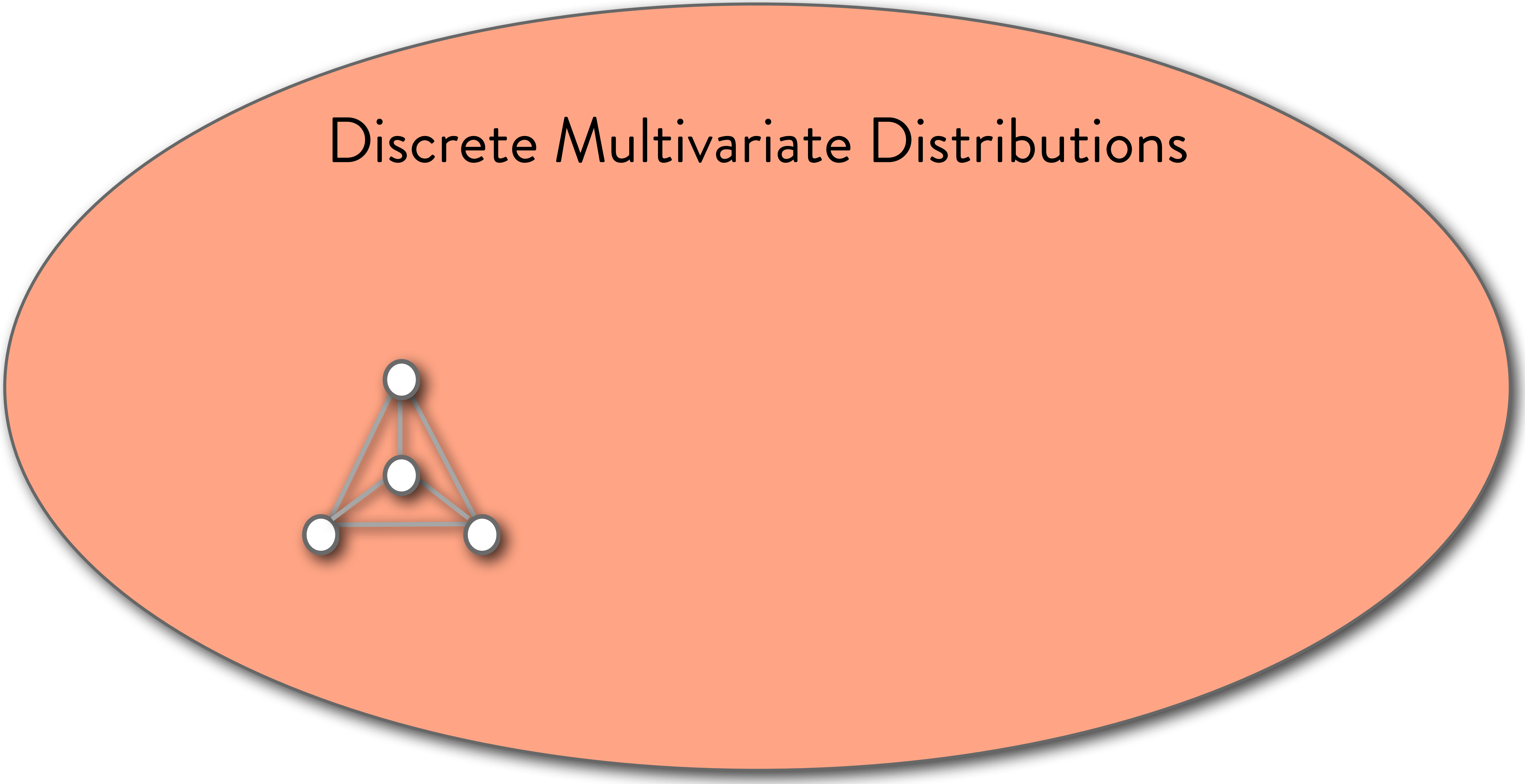
The Catch

- Problem: inference exponential in # of joints (10^{24})

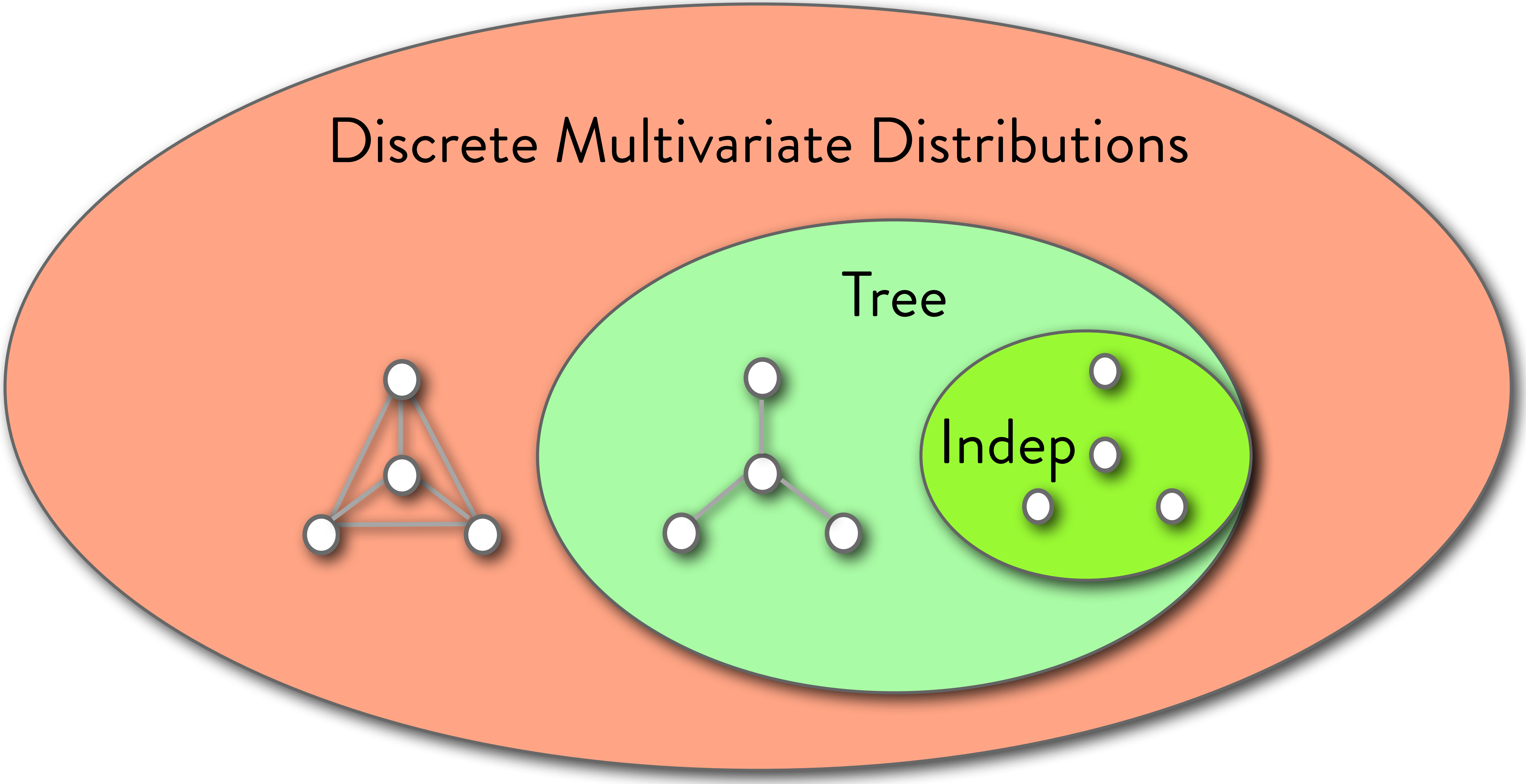


- Structured prediction cascades [Weiss & Taskar, 10]
  - Efficient, accurate inference & learning (with high-probability)
  - Using a coarse-to-fine cascade of graphical models

Graphical Models vs. Tyranny of Small Decisions

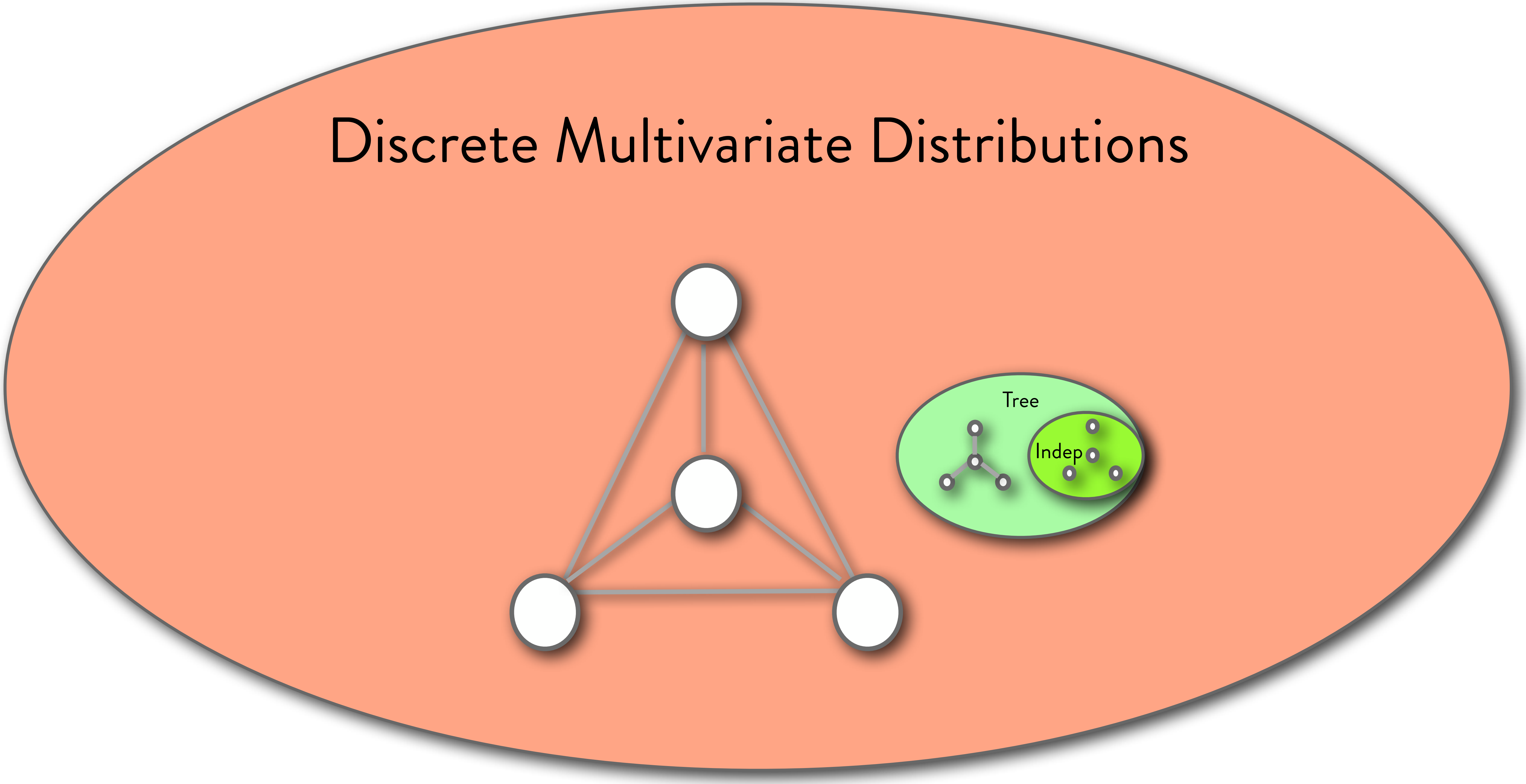


Graphical Models vs. Tyranny of Small Decisions



*Not shown: Dalvi & Suciu 07, Poon & Domingos 11, planar and log-supermodular models

Graphical Models vs. Tyranny of Small Decisions



*Not shown: Dalvi & Suciu 07, Poon & Domingos 11, planar and log-supermodular models

A distribution over sets of poses?

A distribution over sets of poses?

- Uncertainty over number
- Spatial repulsion
- Tractable?

Image search: “jaguar”

Relevance only:

...

Image search: “jaguar”

Relevance only:

...

Relevance + diversity:

...

Summarization

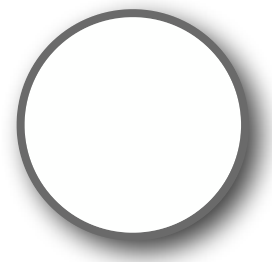
Summarization

Frequency only:

- Romney expected to claim nomination
- Romney wins three primaries
- Romney tightens grip in GOP race
- Romney is unpopular, likely nominee

Graphical models?

Graphical models?

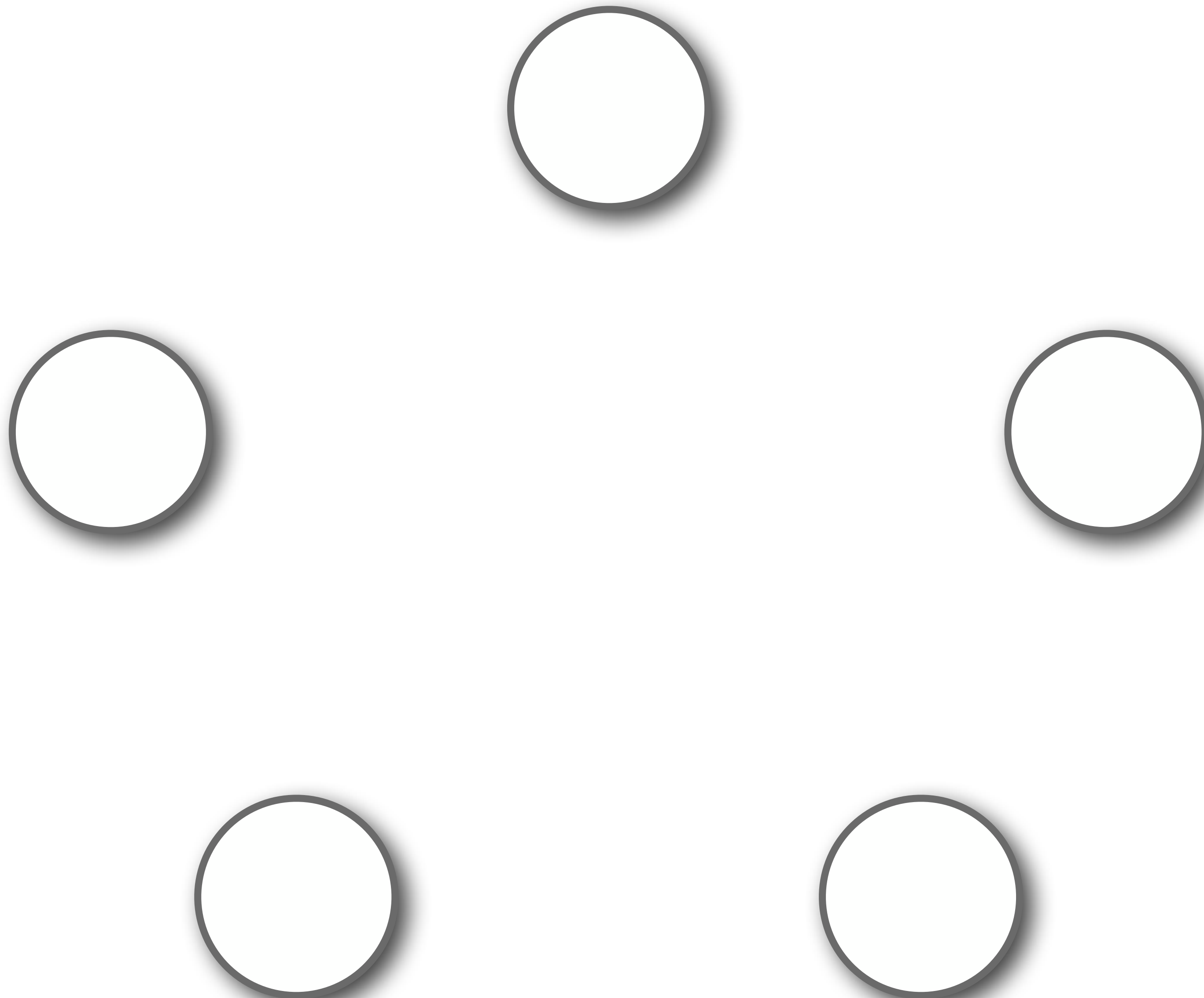


sentence i

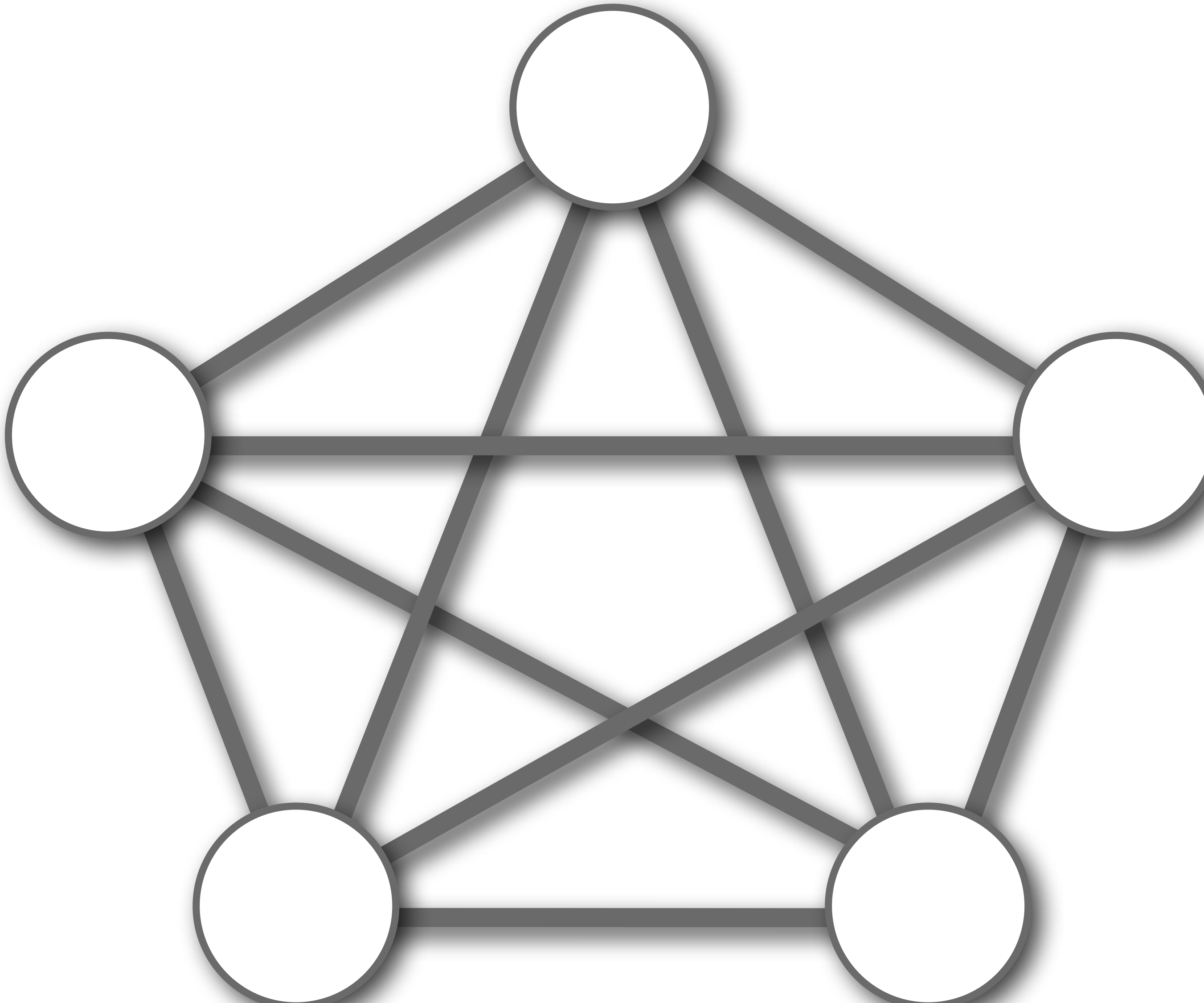
Graphical models?

sentence *i*

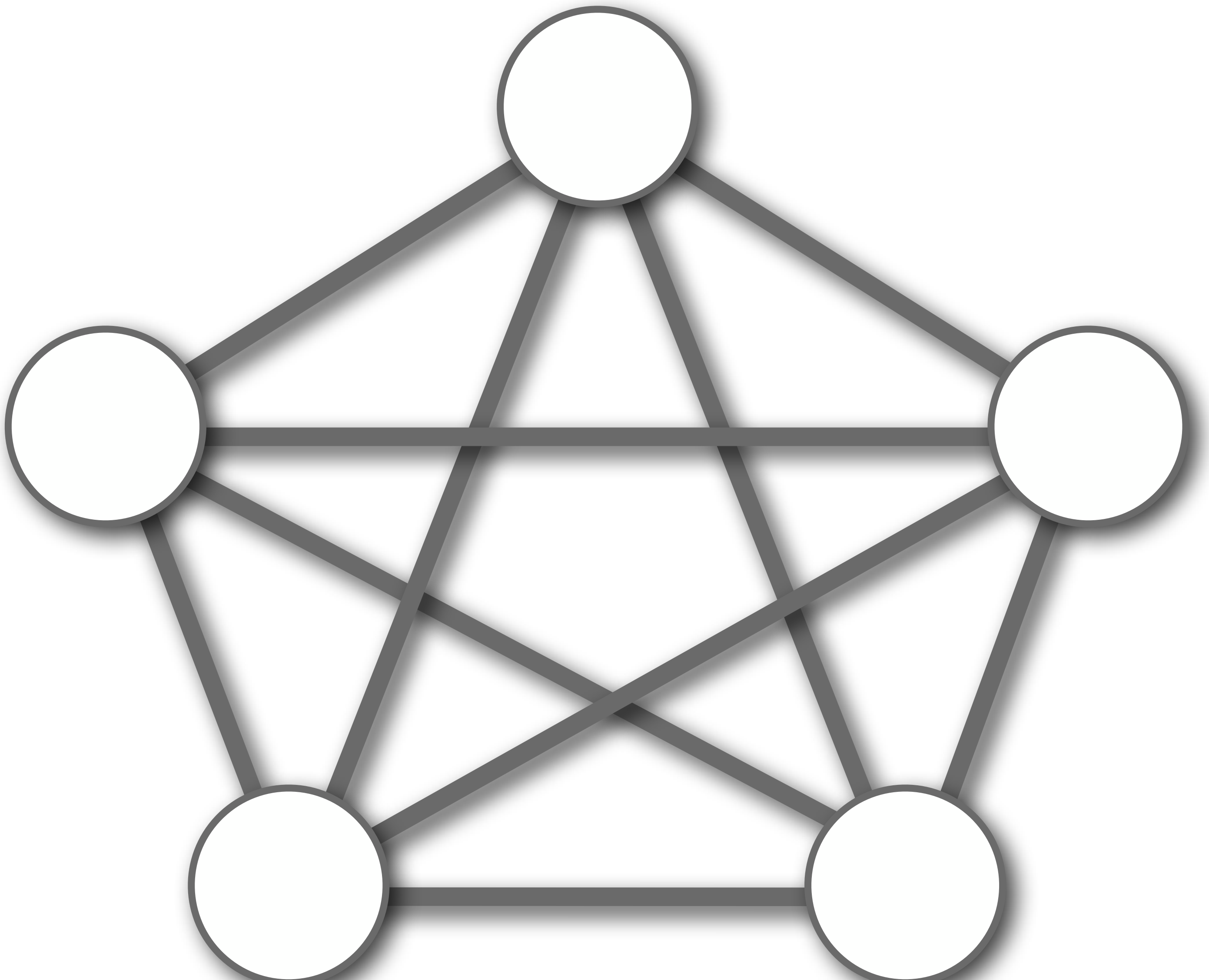
Graphical models?



Graphical models?

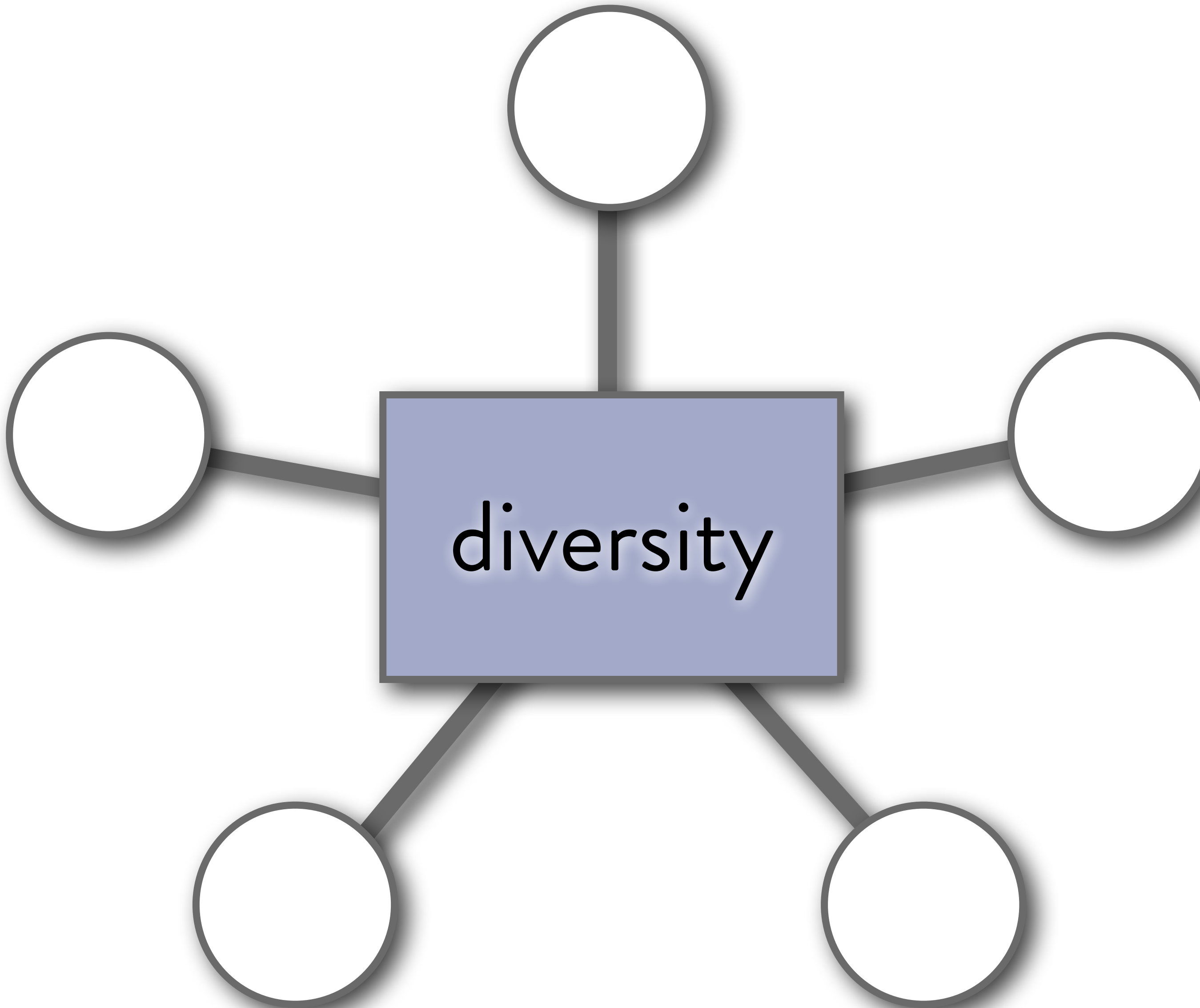


Graphical models?

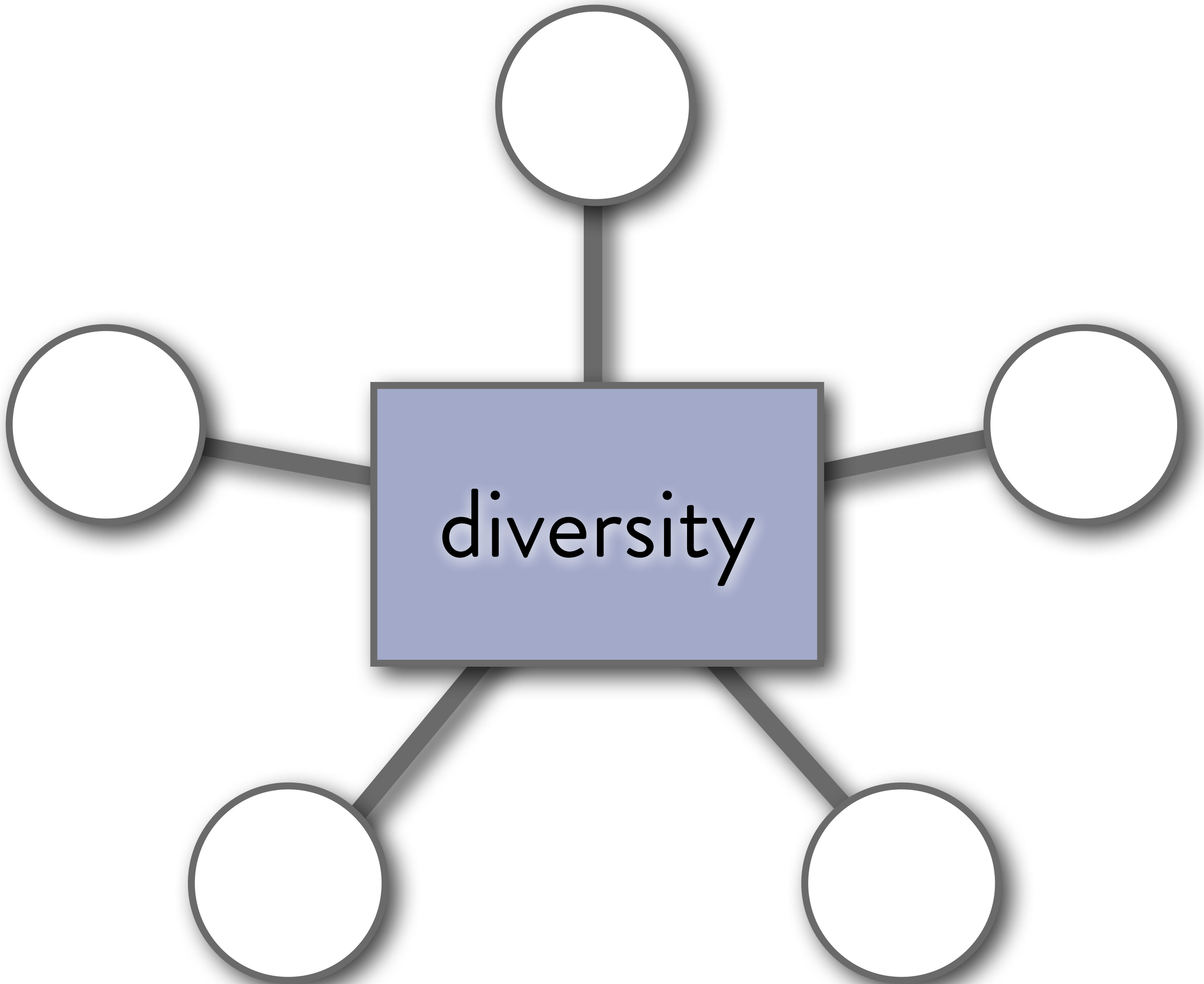


Local negative interactions + **many cycles** = hard

Determinantal point processes (DPPs)



Determinantal point processes (DPPs)



Global, **negative** interactions are easy

Determinantal point processes

Quality, diversity, and learning

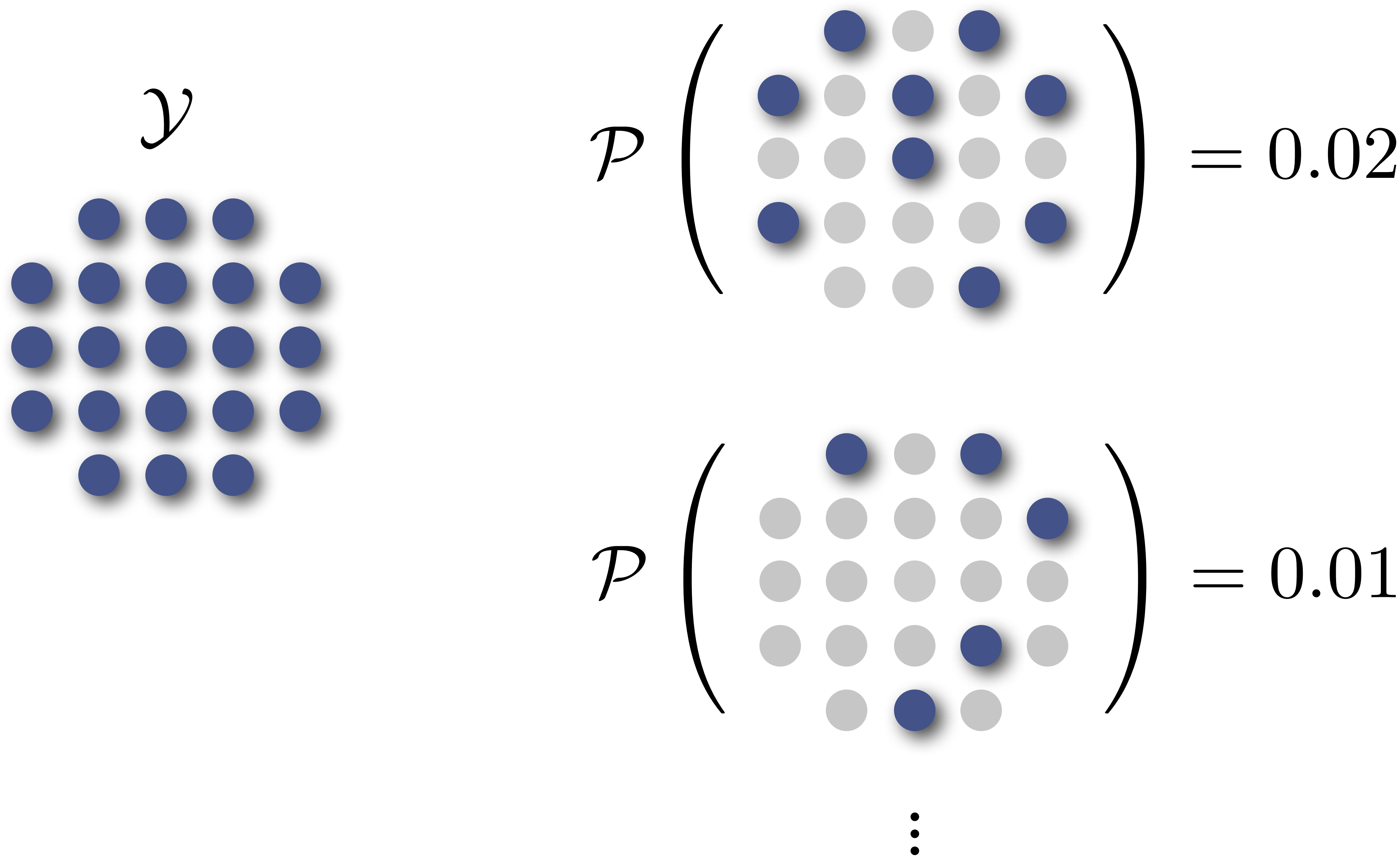
Sampling

k-DPPs (fixed cardinality)

Structured DPPs

News threading

Discrete point process



Discrete point process

- 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

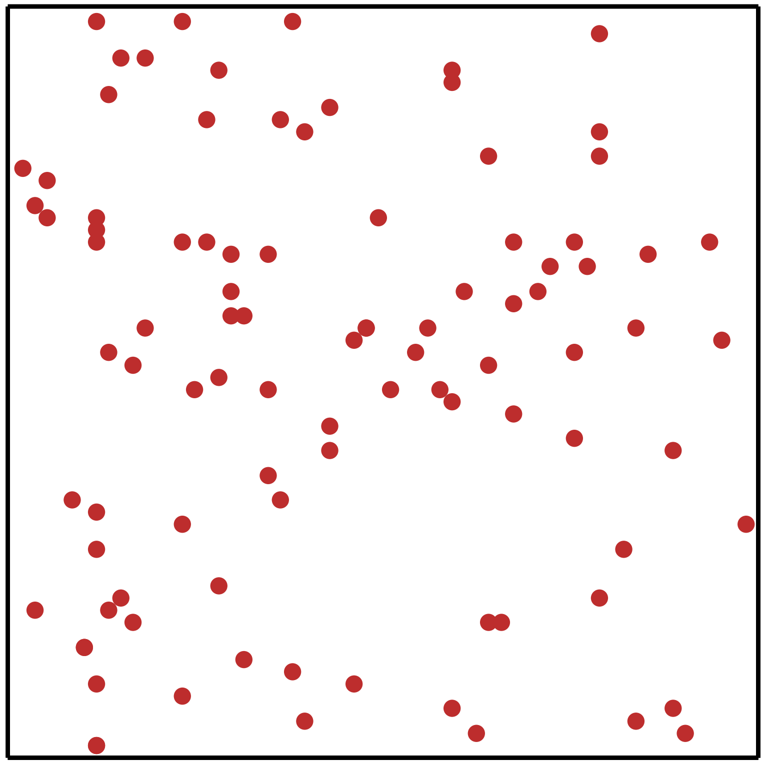
- Each element *i* included with probability p_i:

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

Independent point process

- Each element *i* included with probability p_i:

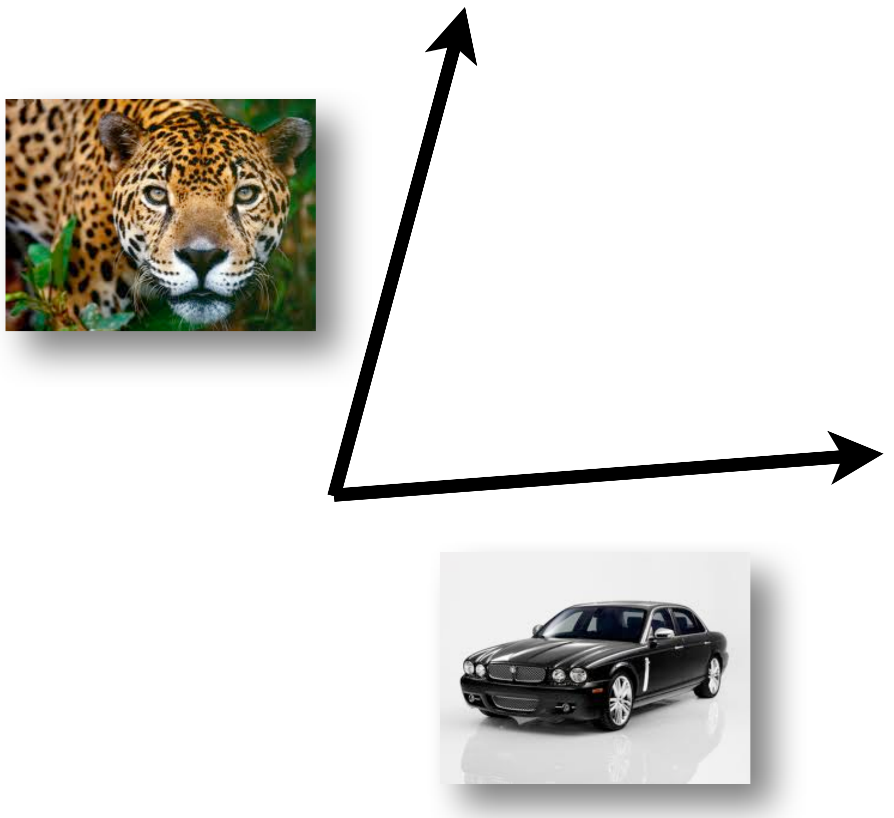
$$\mathcal{P}(\boldsymbol{Y} = Y) = \prod_{i \in Y} p_i \prod_{i \notin Y} (1 - p_i)$$



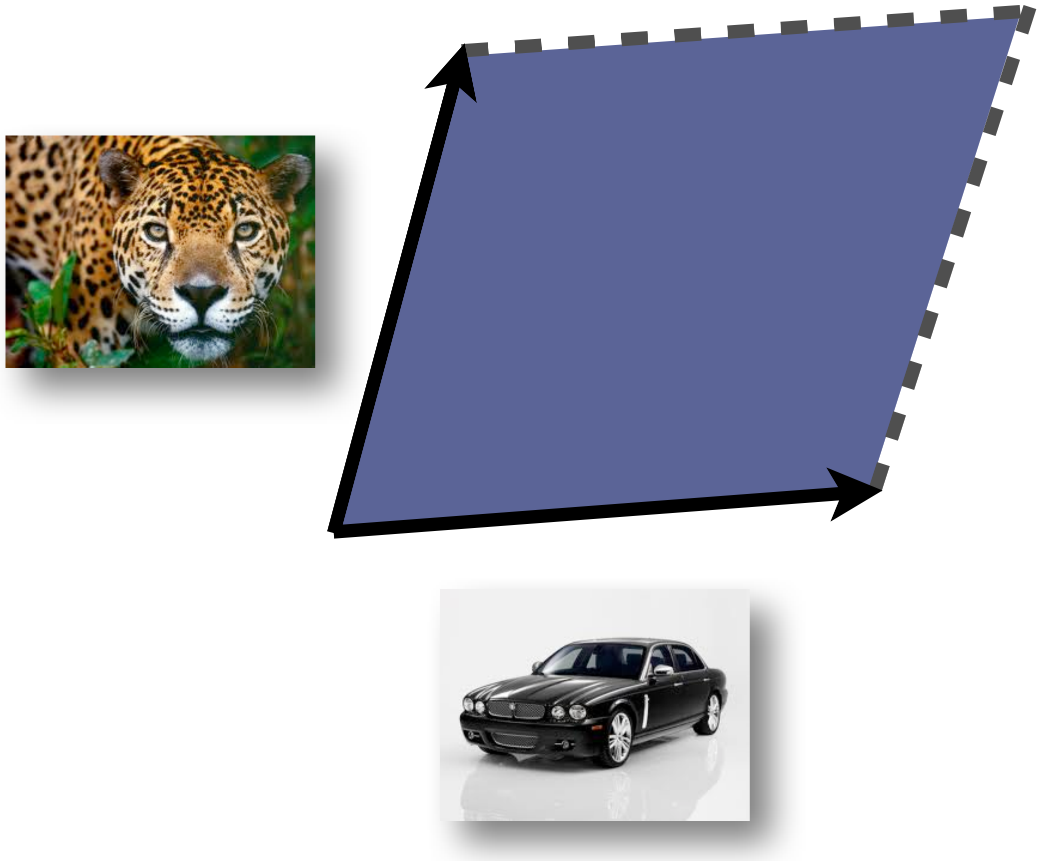
Feature function **g** on items in $\mathcal{Y}$

$\boldsymbol{g}\Big($  $\Big)$

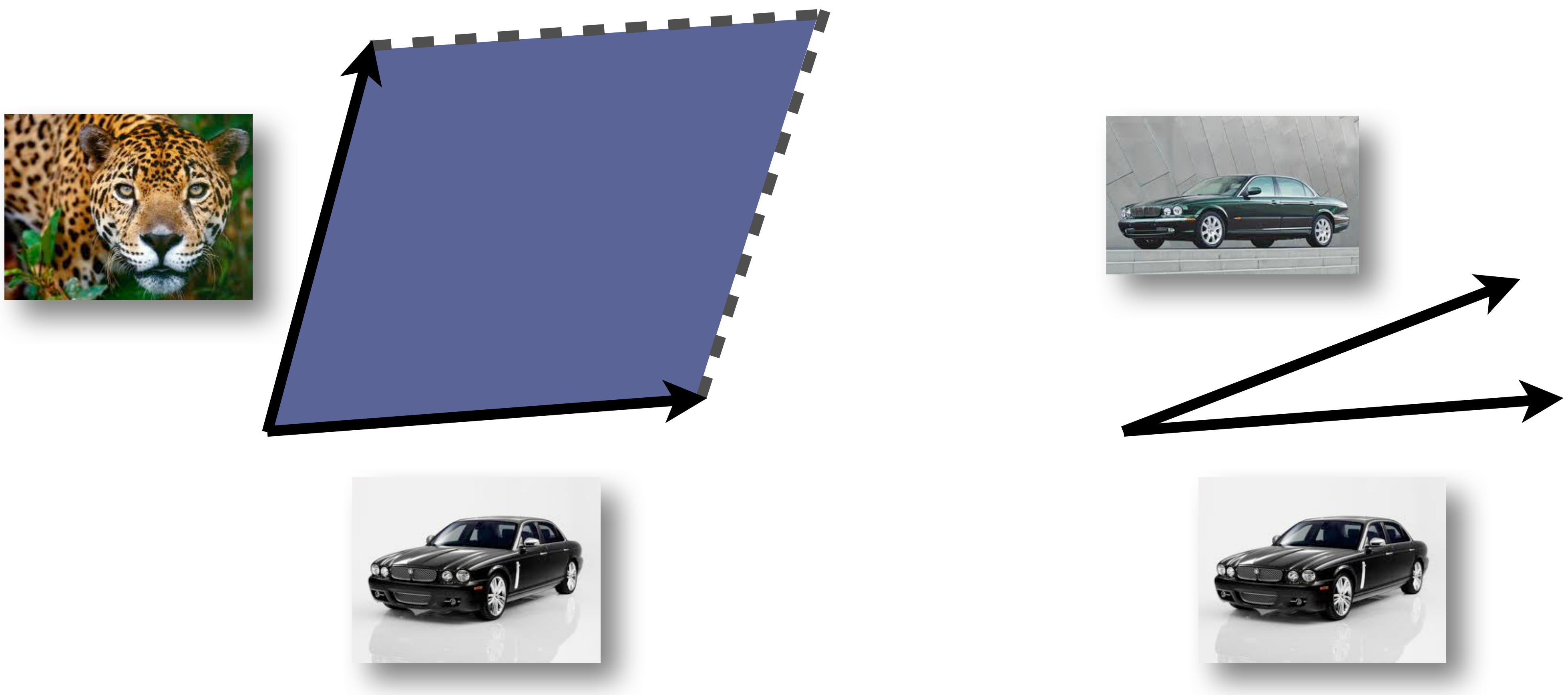
Feature function **g** on items in $\mathcal{Y}$



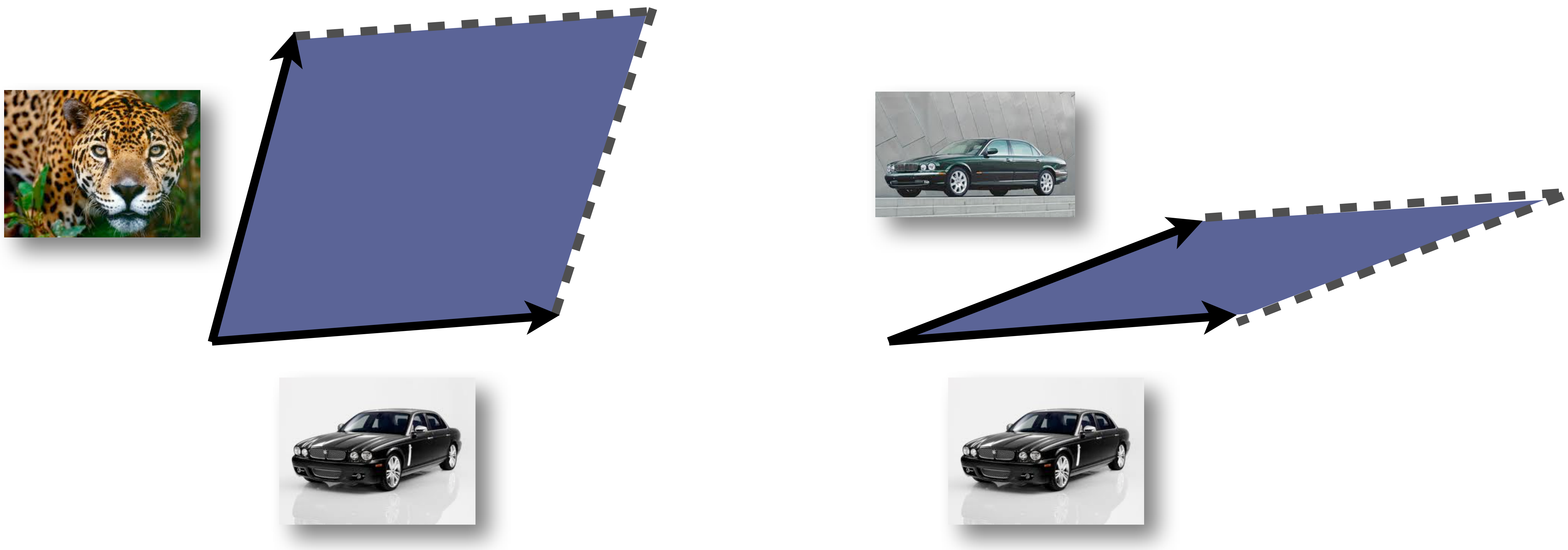
Feature function $\mathbf{g}$ on items in $\mathcal{Y}$

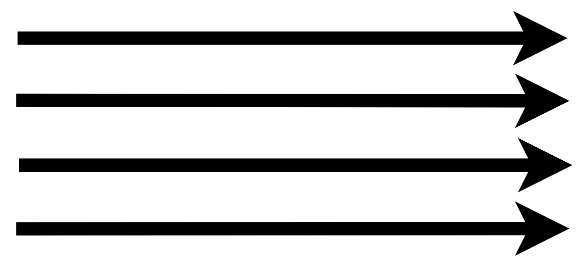


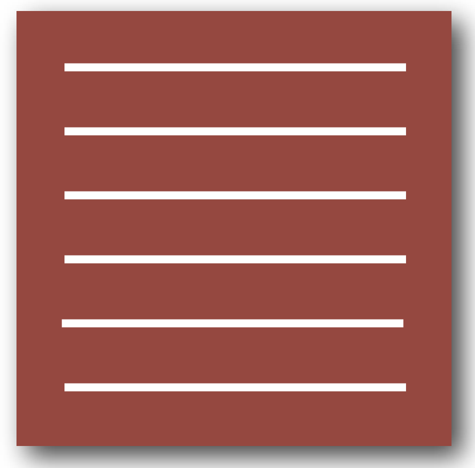
Feature function $\mathbf{g}$ on items in $\mathcal{Y}$



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







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

Determinantal point process

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

[Macchi, 1975]

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

- Given an $N \times N$ symmetric p.s.d. matrix L

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

[Macchi, 1975]

Determinantal point process

- Given an $N \times N$ symmetric p.s.d. matrix L

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

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

[Macchi, 1975]

Determinantal point process

- Given an $N \times N$ symmetric p.s.d. matrix L

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

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

[Macchi, 1975]

Determinantal point process

- Given an $N \times N$ symmetric p.s.d. matrix L

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

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

[Macchi, 1975]

Determinantal point process

- Given an $N \times N$ symmetric p.s.d. matrix L

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

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

[Macchi, 1975]

DPP inference

- Normalization:

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

- Marginals, conditioning (N^3 or faster)
- Exact sampling (N^3 or faster)
- MAP / mode is NP-hard, but log-submodular

DPP inference

- Normalization:

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

- Marginals, conditioning (N^3 or faster)
- Exact sampling (N^3 or faster)
- MAP / mode is NP-hard, but log-submodular

DPP inference

- Marginals:

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

DPP inference

- Marginals:

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

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

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

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

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

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

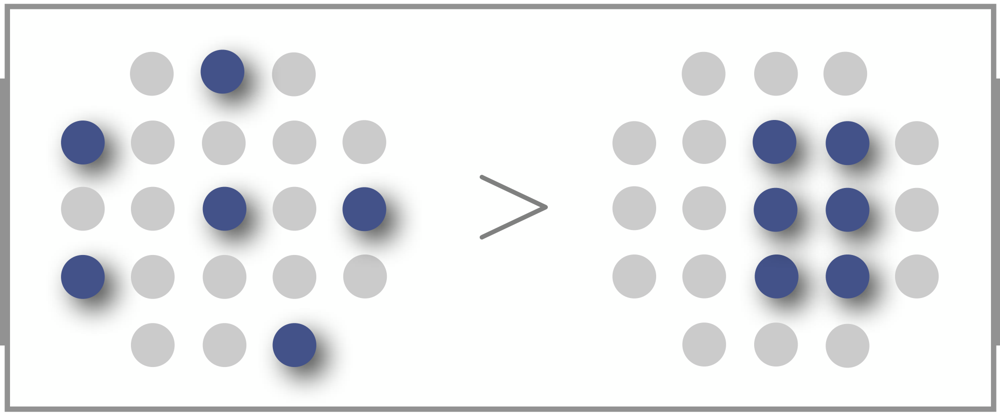
$$\mathcal{P}(i \in \boldsymbol{Y}) = \det(K_{ii}) = K_{ii}$$

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

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

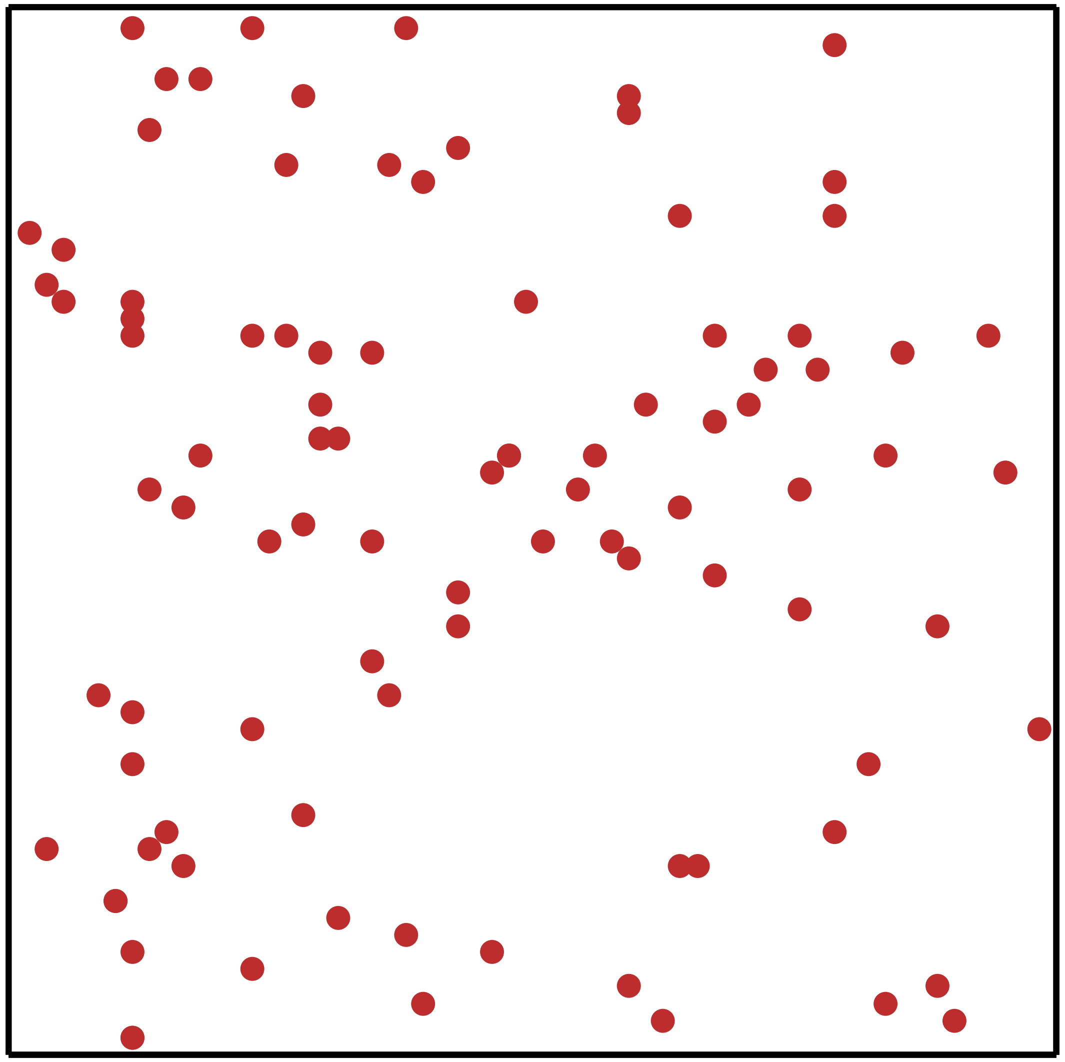
$$\mathcal{P}(i \in \boldsymbol{Y}) = \det(K_{ii}) = K_{ii}$$

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

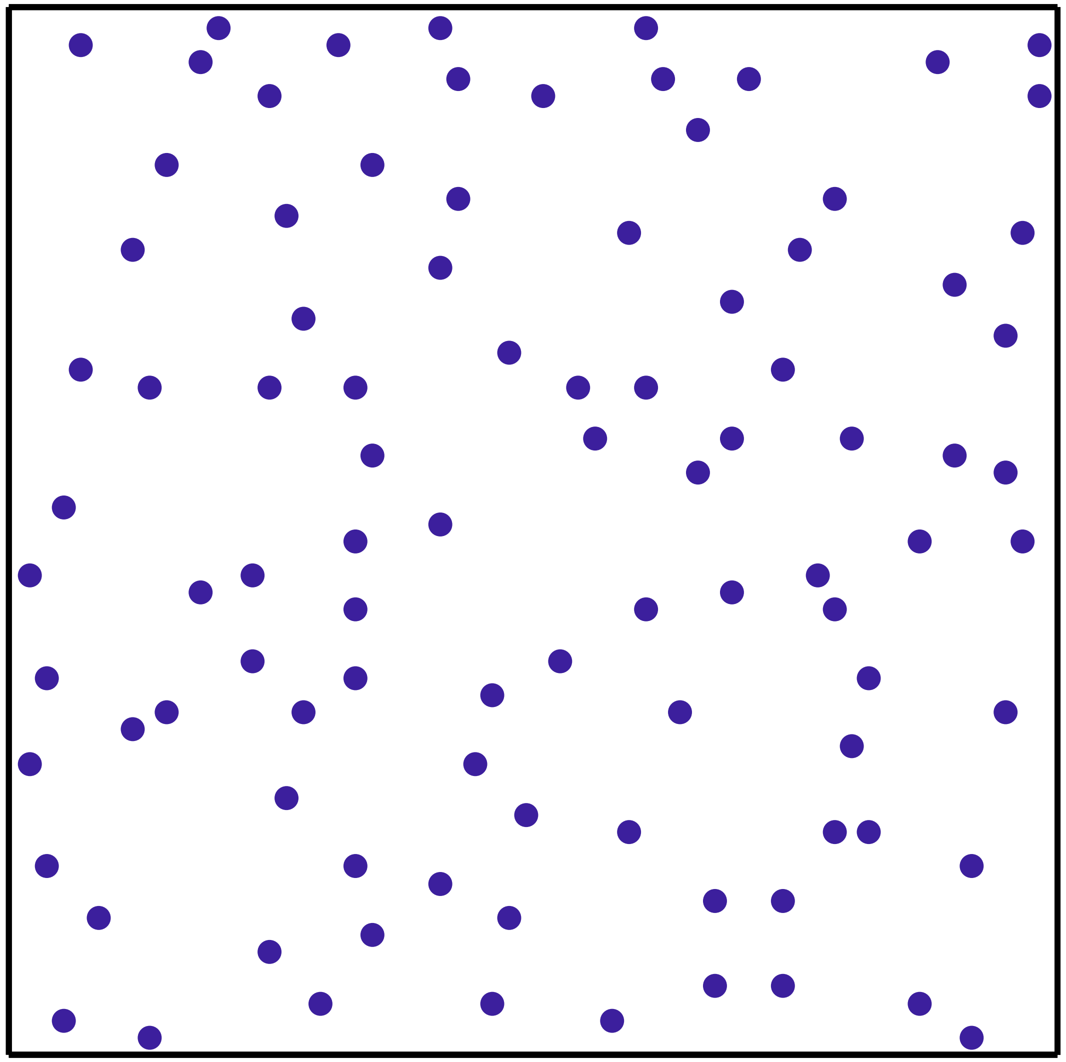


Diversity

Point process samples



Independent



DPP

Determinantal point processes

Quality, diversity, and learning

Sampling

k-DPPs (fixed cardinality)

Structured DPPs

News threading


L
=

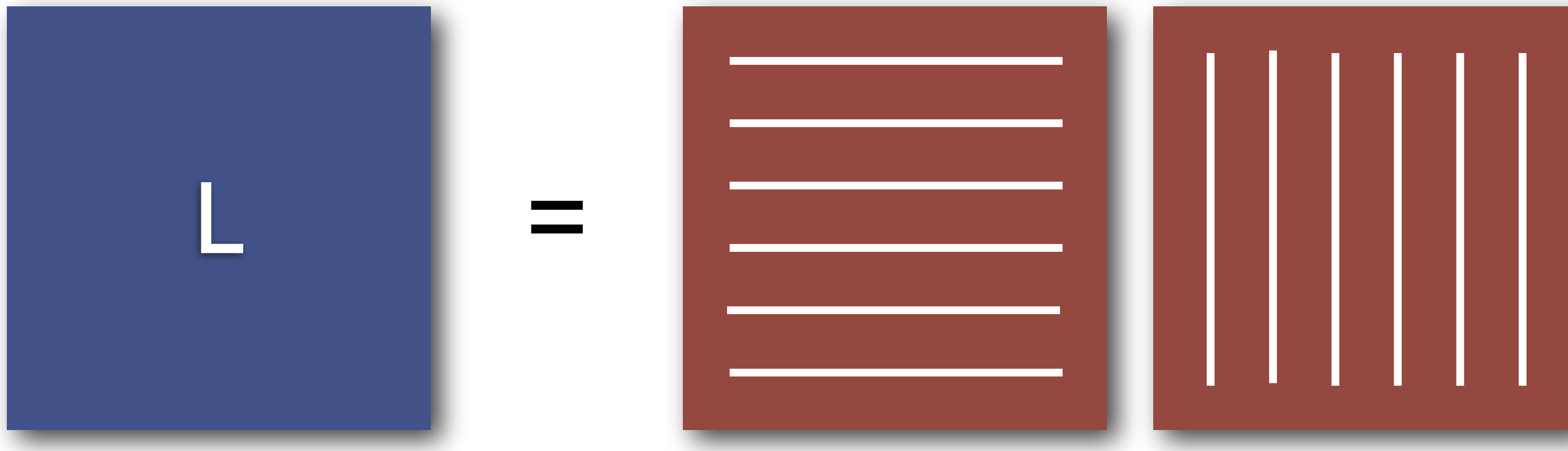
$$L_{ij} = \boldsymbol{g}(i)^\top \boldsymbol{g}(j)$$

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

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

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

Quality score



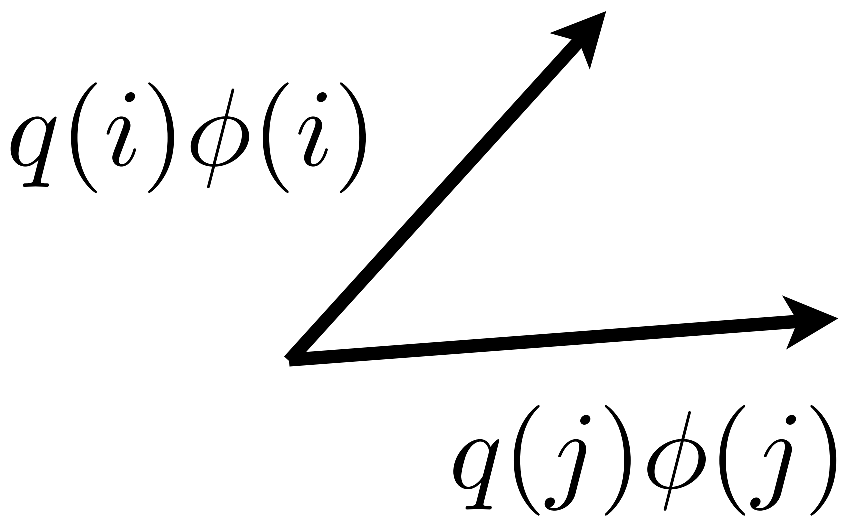
$$L_{ij} = q(i)\phi(i)^\top \phi(j) q(j)$$

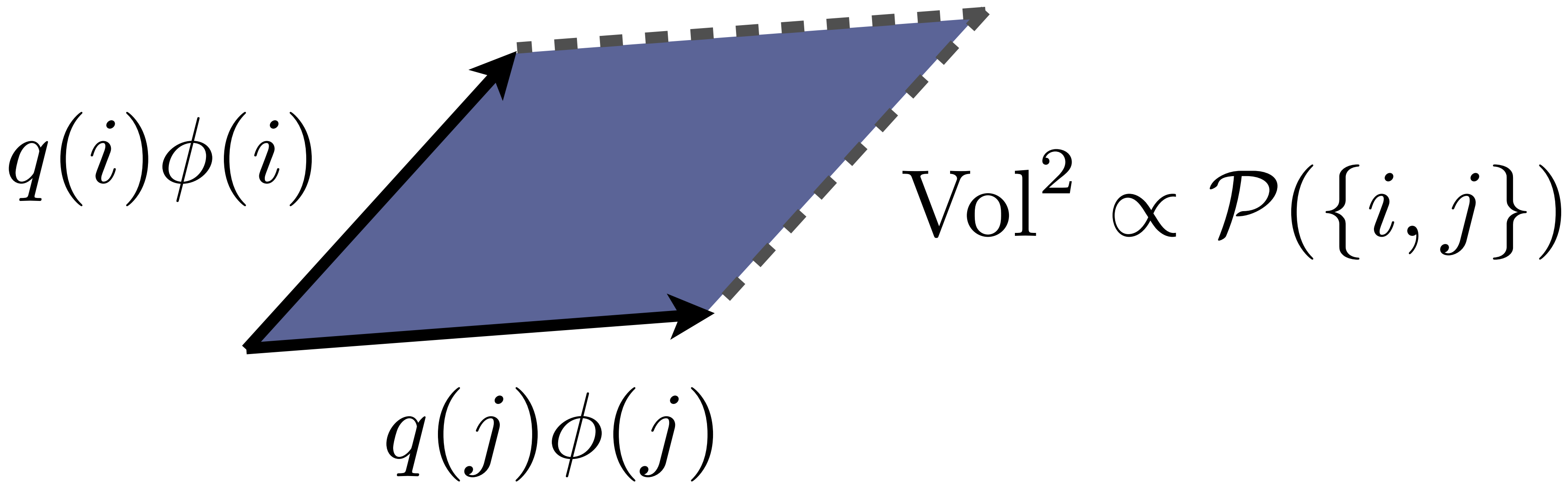
$$q(i) \in \mathbb{R}_+$$

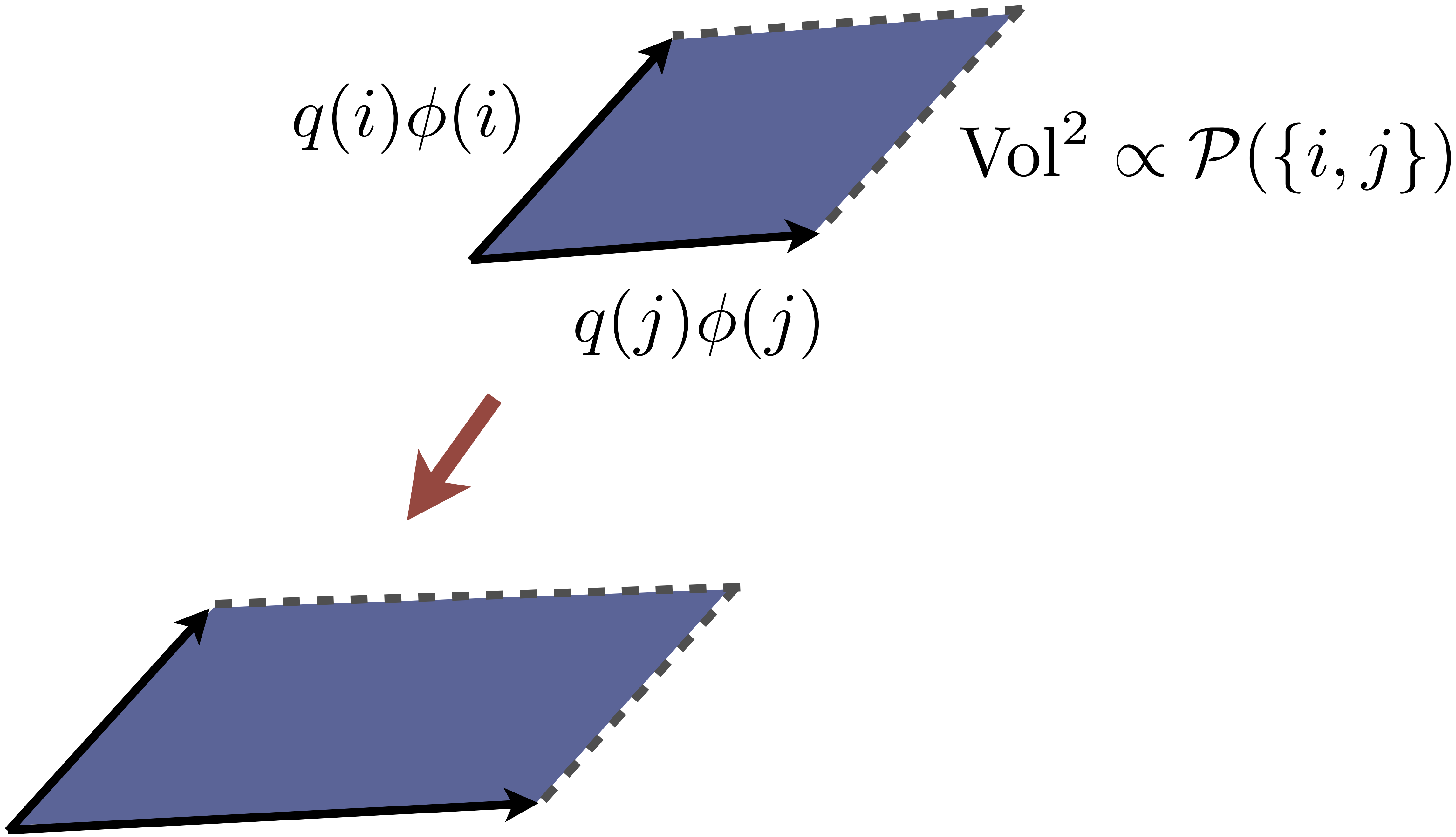
Quality score

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

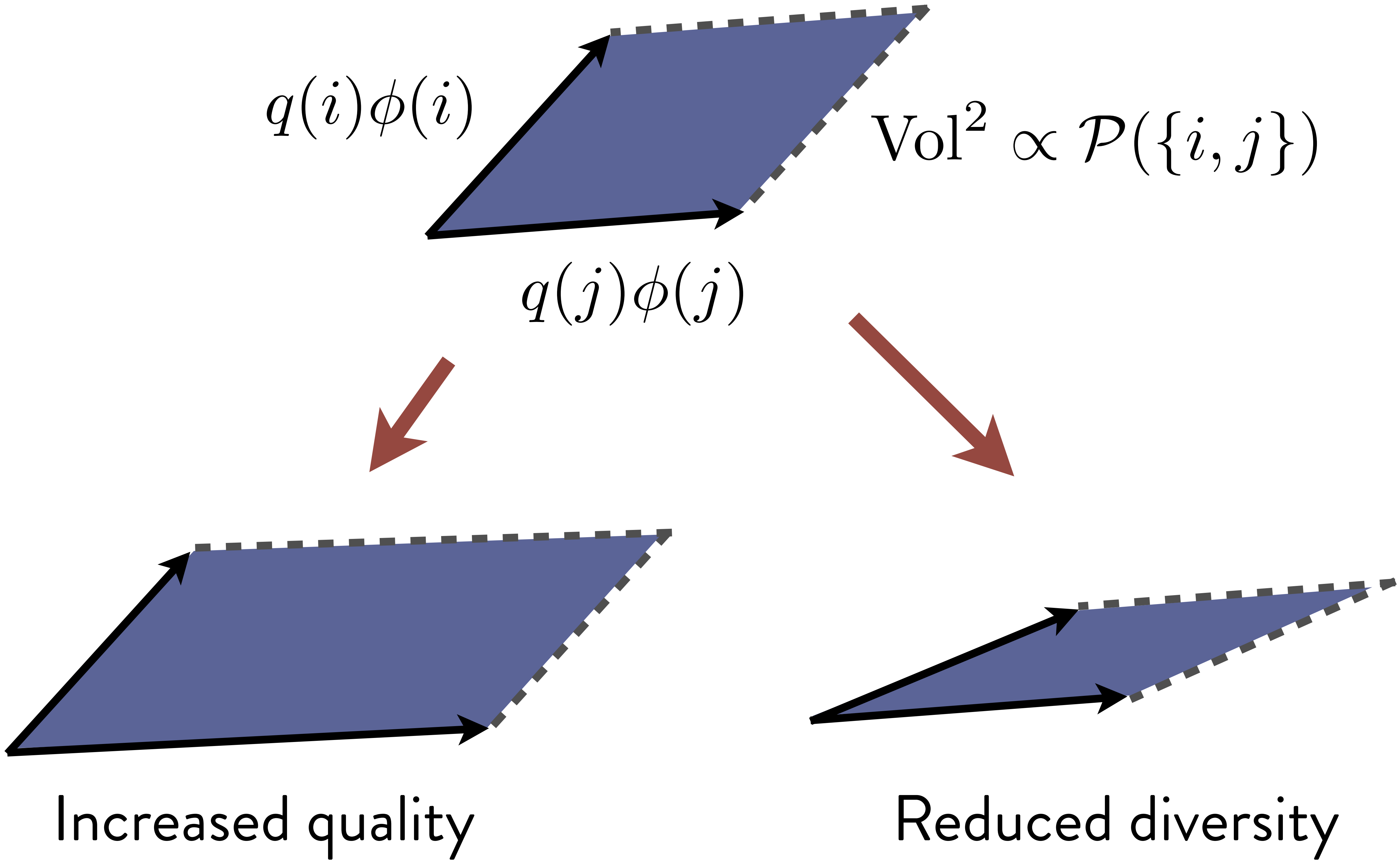
Diversity features


$q(i)\phi(i)$
$q(j)\phi(j)$


$q(i)\phi(i)$
$\mathrm{Vol}^2 \propto \mathcal{P}(\{i, j\})$
$q(j)\phi(j)$



Increased quality


$q(i)\phi(i)$
$\mathrm{Vol}^2 \propto \mathcal{P}(\{i,j\})$
$q(j)\phi(j)$
Increased quality
Reduced diversity

Quality vs. 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

Quality vs. **diversity**

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

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

  - Optimize θ by maximum likelihood
  - Can find global optimum in $O(N^3)$

Quality vs. **diversity**

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

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

  - Optimize θ by maximum likelihood
  - Can find global optimum in $O(N^3)$
- Don't yet know how to learn **diversity** efficiently **(a natural parametrization is NP-hard)**

Determinantal point processes

Quality, diversity, and learning

Sampling

k-DPPs (fixed cardinality)

Structured DPPs

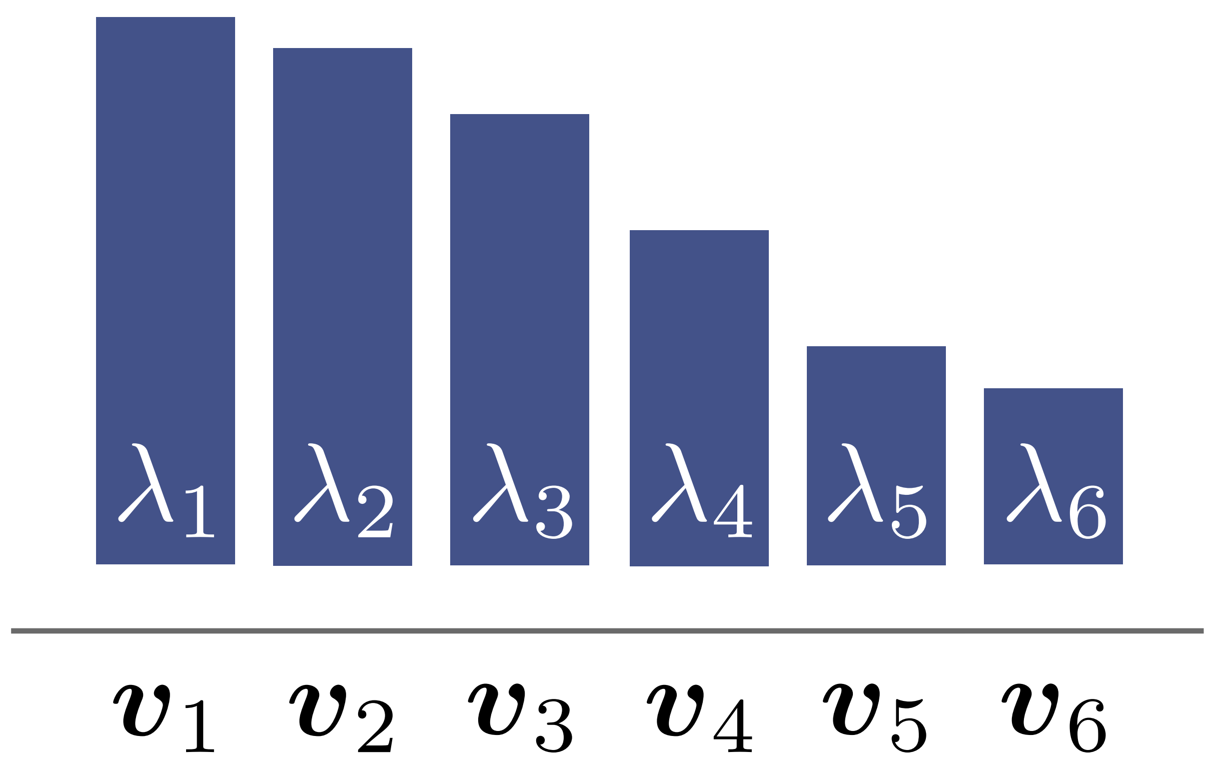
News threading

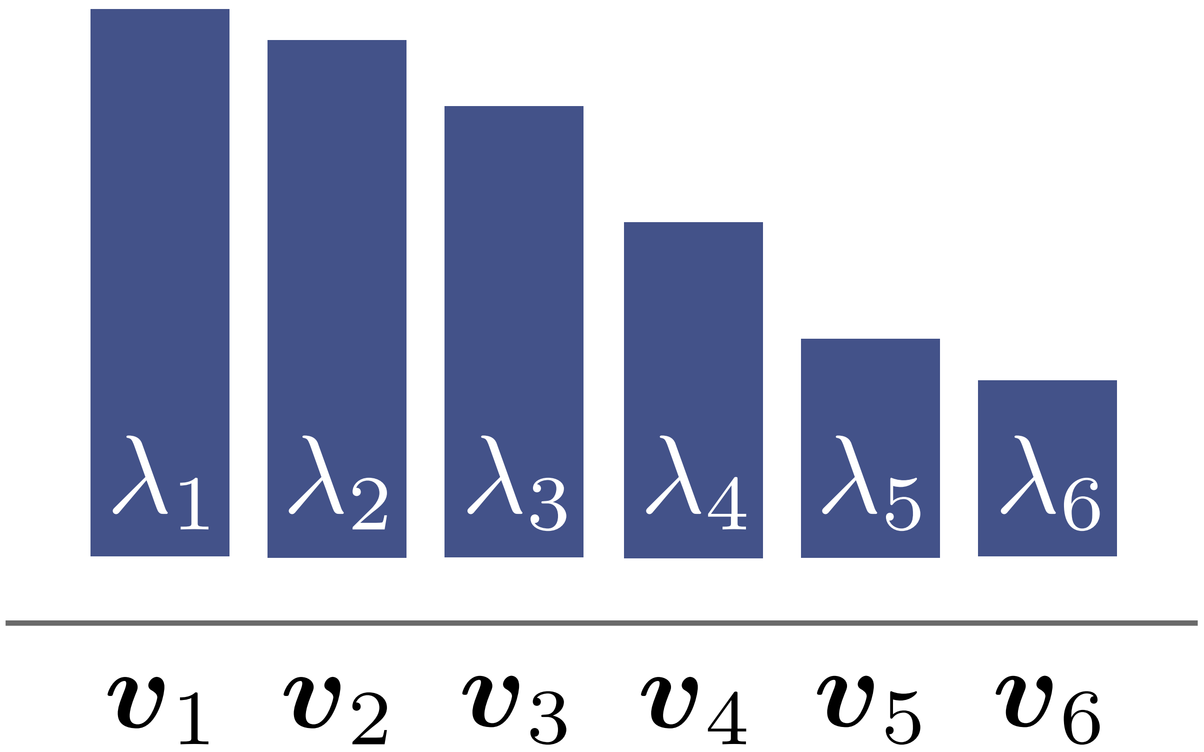
Eigendecomposition

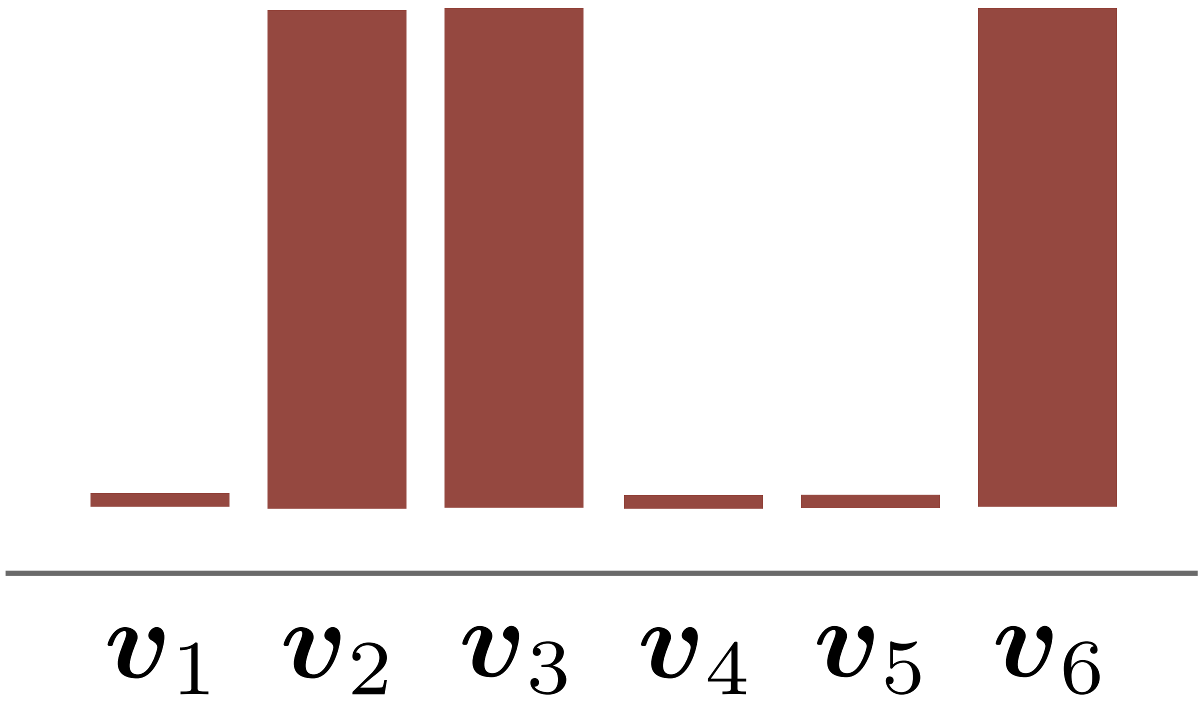
$$L = \sum_{n=1}^{N} \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^\top$$

Eigendecomposition

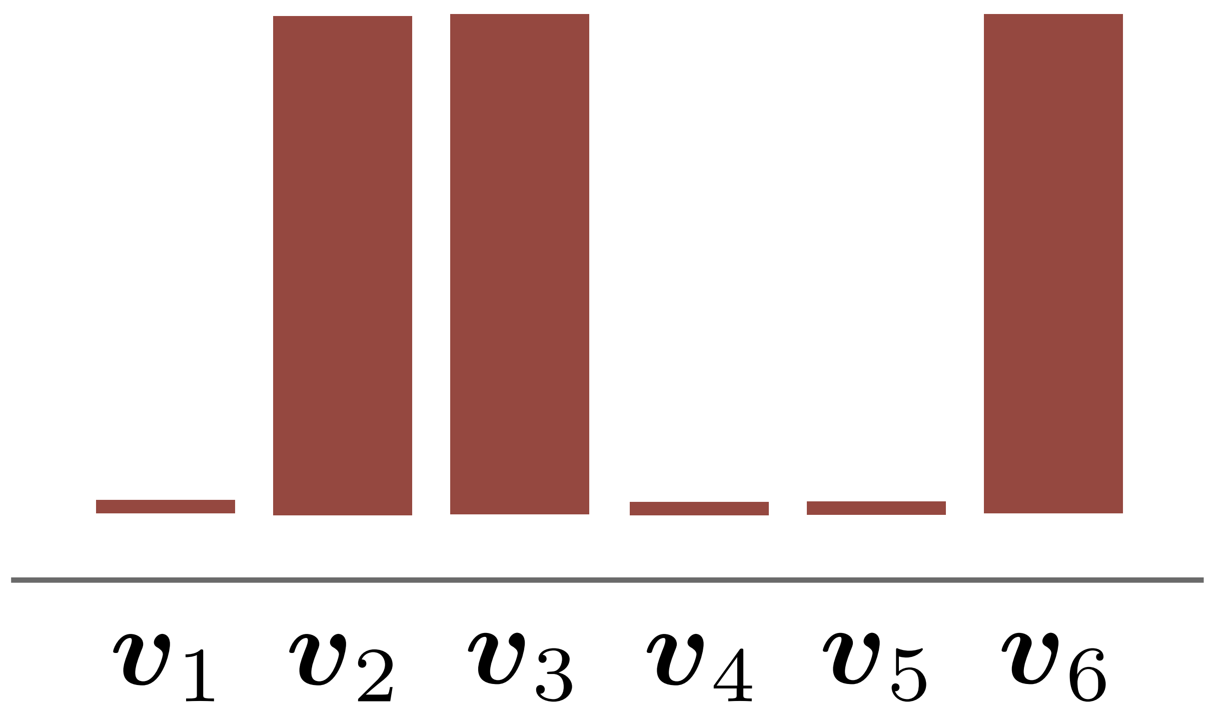
$$L = \sum_{n=1}^{N} \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^\top$$



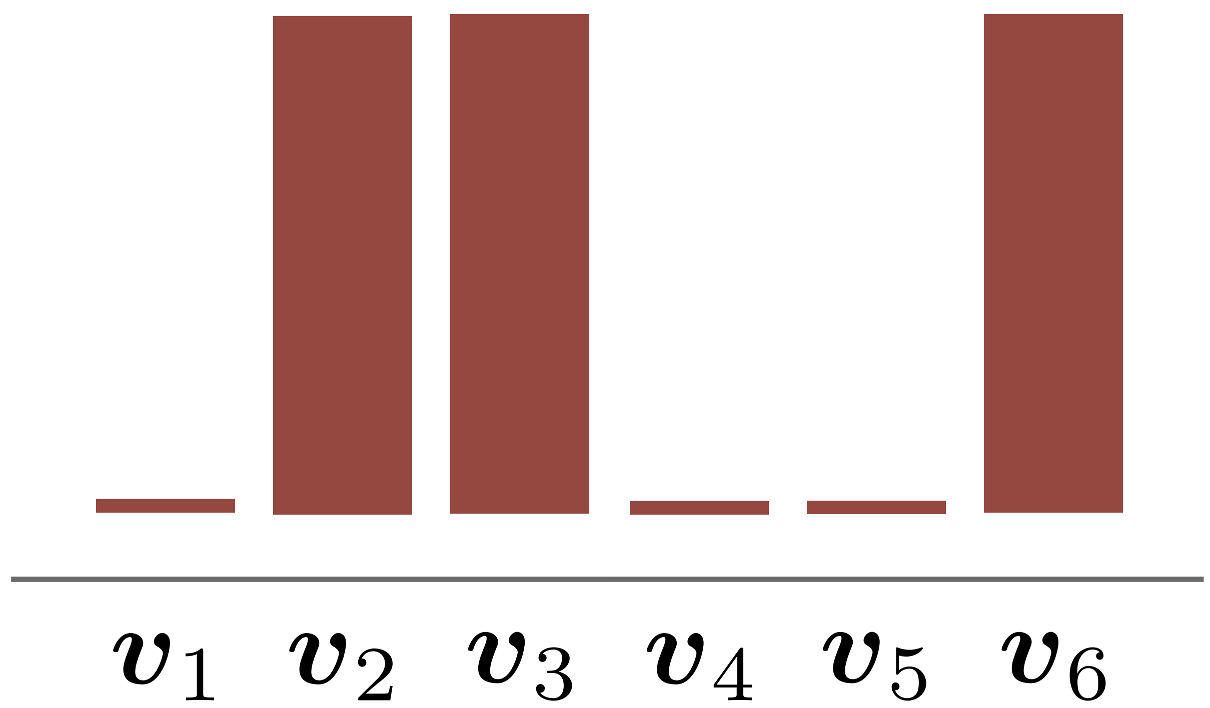

λ_1 λ_2 λ_3 λ_4 λ_5 λ_6
$\boldsymbol{v}_1$ $\boldsymbol{v}_2$ $\boldsymbol{v}_3$ $\boldsymbol{v}_4$ $\boldsymbol{v}_5$ $\boldsymbol{v}_6$


$\boldsymbol{v}_1$ $\boldsymbol{v}_2$ $\boldsymbol{v}_3$ $\boldsymbol{v}_4$ $\boldsymbol{v}_5$ $\boldsymbol{v}_6$

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



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



- Easy to sample in polynomial time
- $\mathcal{P}^J$ only supported on sets of size $|J|$

Key insight

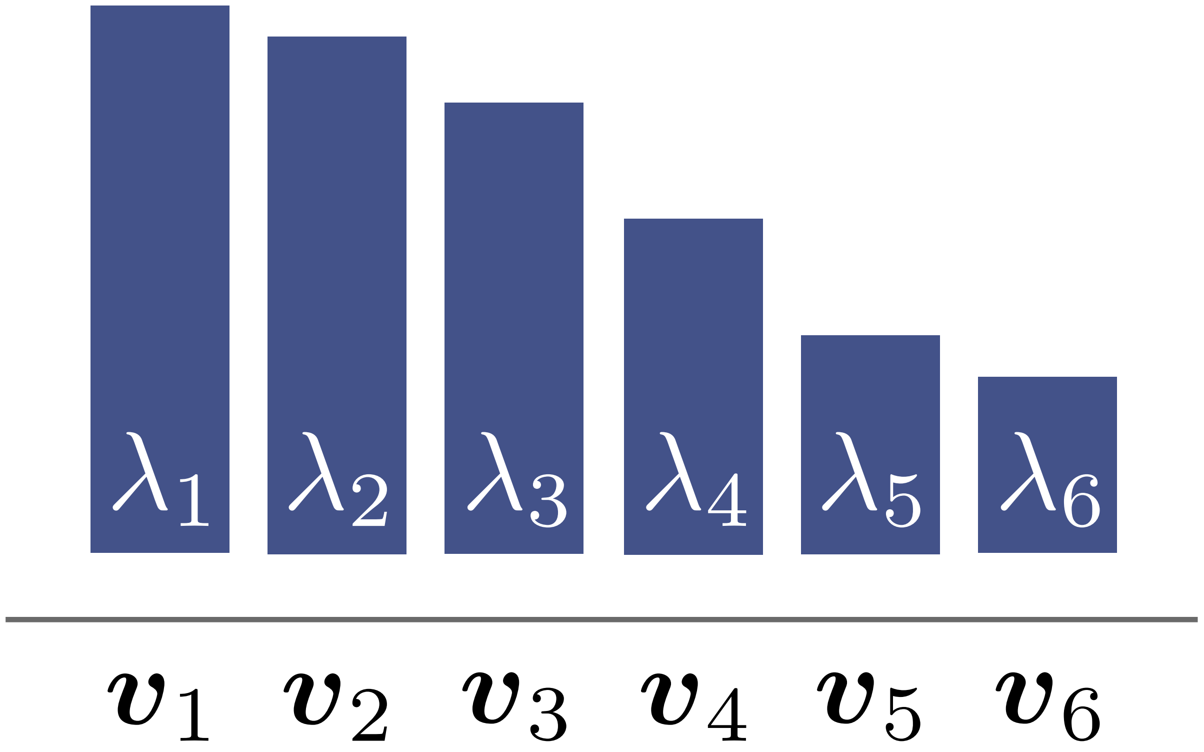
Every DPP is a "factored" mixture of its elementary DPPs:

$$\mathcal{P} \propto \sum_{J \subseteq \{1,\dots,N\}} \mathcal{P}^J \underbrace{\prod_{n \in J} \lambda_n}_{\text{mixture weight}}$$

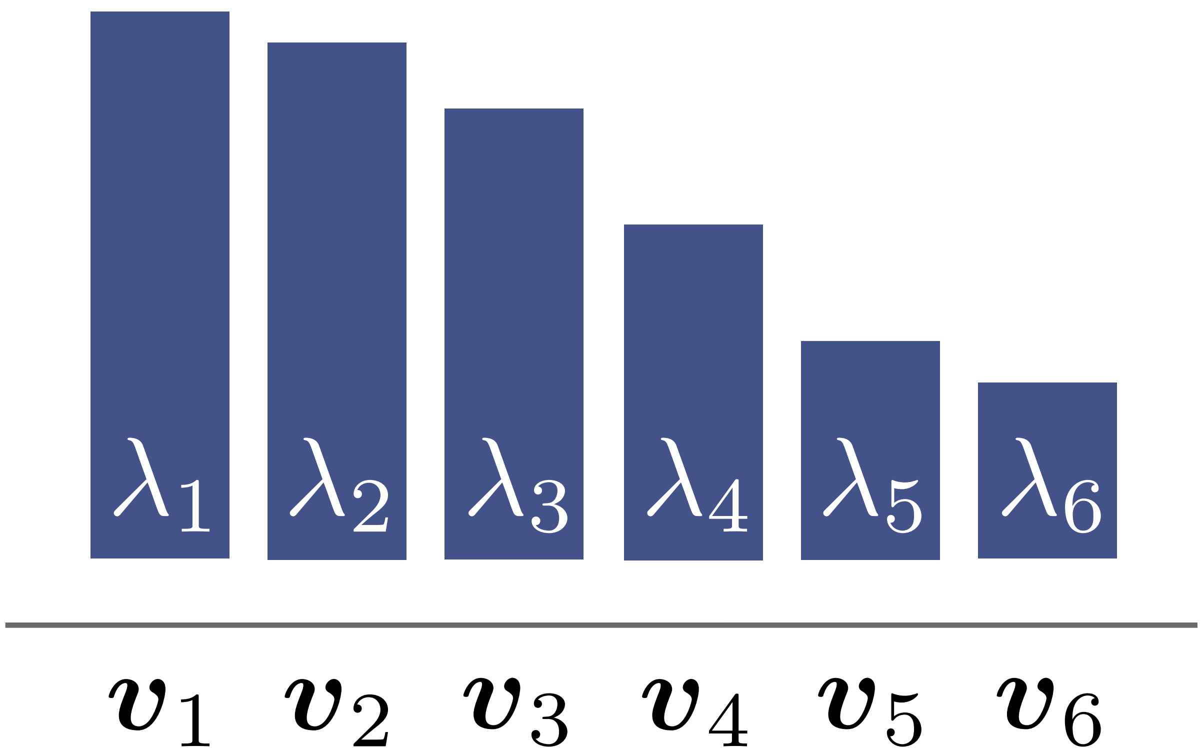
[Hough et al, 2006]

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

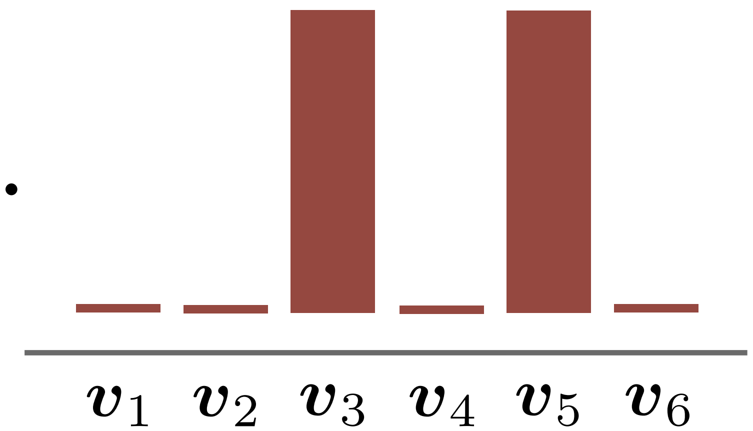
mixture weight



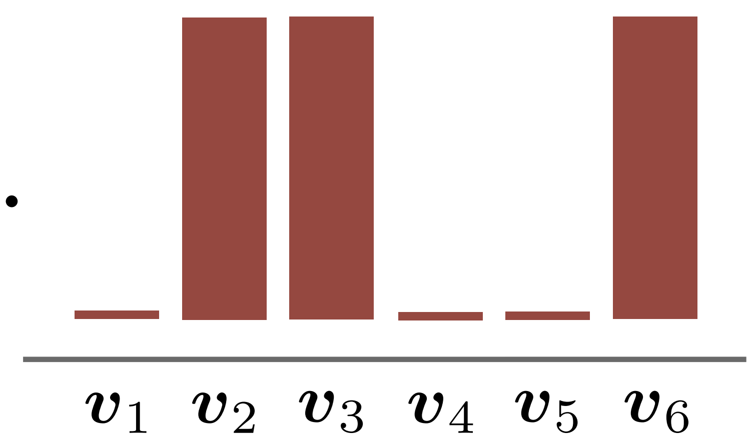
$$\mathcal{P} \propto \sum_{J \subseteq \{1,\ldots,N\}} \mathcal{P}^J \underbrace{\prod_{n \in J} \lambda_n}_{\text{mixture weight}}$$



$\lambda_3 \lambda_5 \cdot$



$+ \lambda_2 \lambda_3 \lambda_6 \cdot$



$+ \cdots$

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

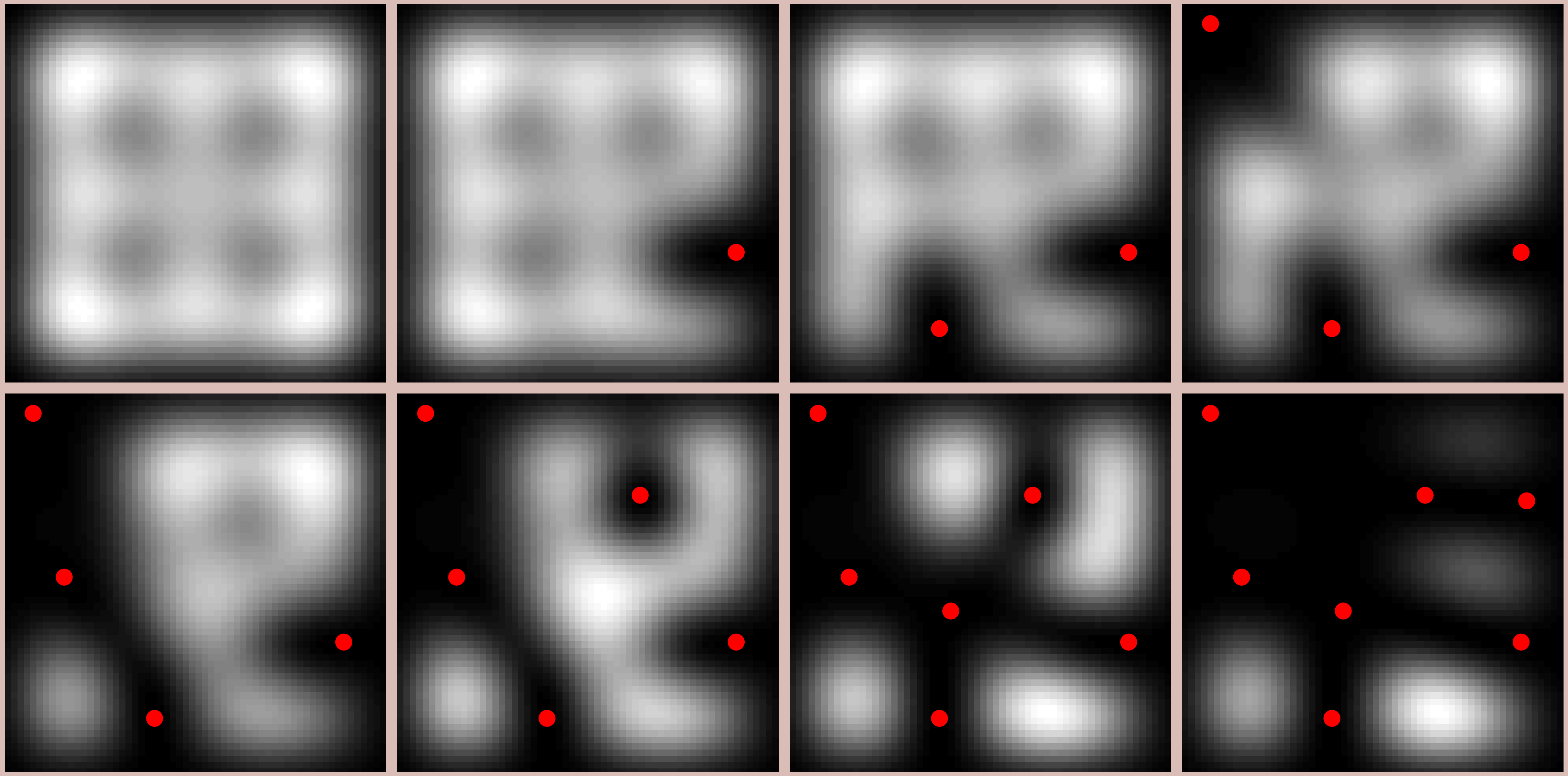
Choose elementary DPP $\mathcal{P}^J$ by mixture weight:

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

- 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 $\mathcal{P}^J$



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 $\mathcal{P}^J$

Consequences

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

Consequences

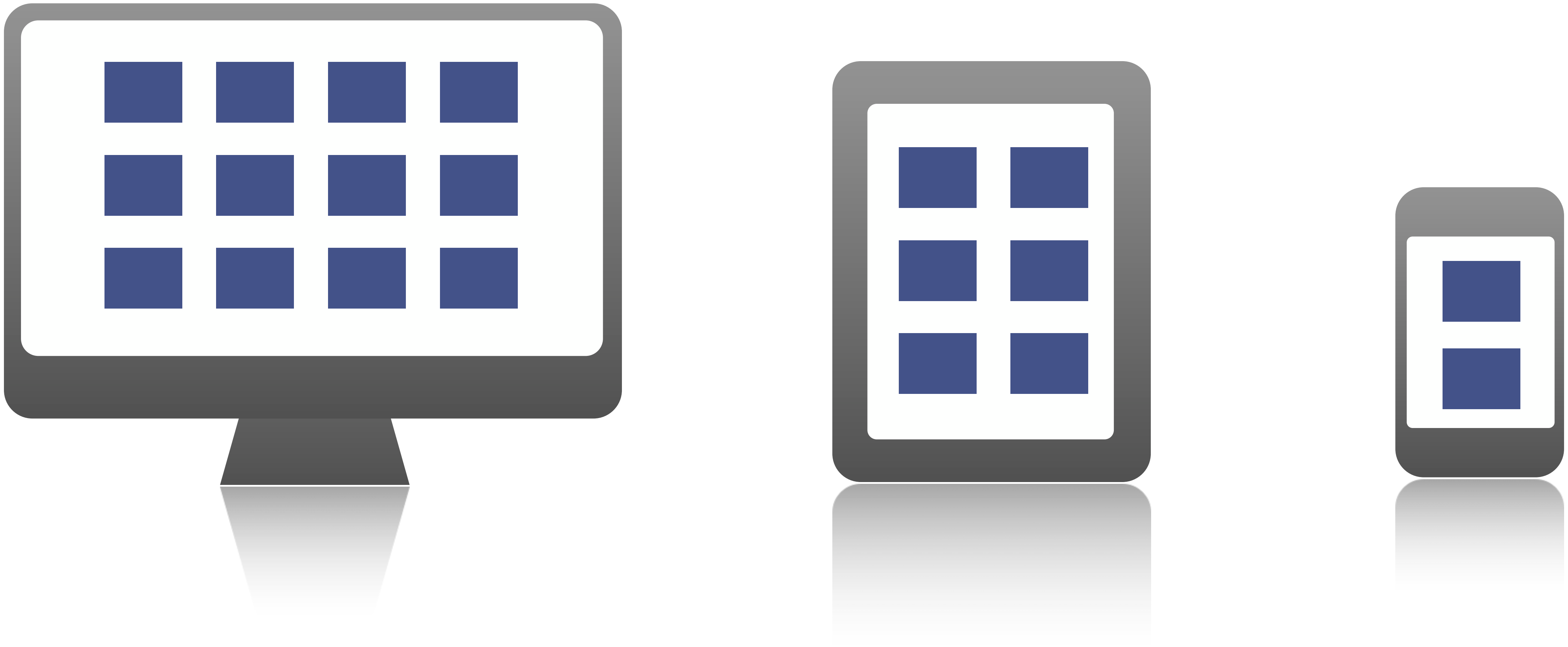
- Phase one determines:
  - **Size** of sample ($|J|$)
  - Likely **content** of sample (eigenvectors)

➡ **Size** and **content** are tied

➡ **Size** is sum of Bernoulli variables

What if we need exactly *k* diverse items?

What if we need exactly *k* diverse items?



Determinantal point processes

Quality, diversity, and learning

Sampling

k-DPPs (fixed cardinality)

Structured DPPs

News threading

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!

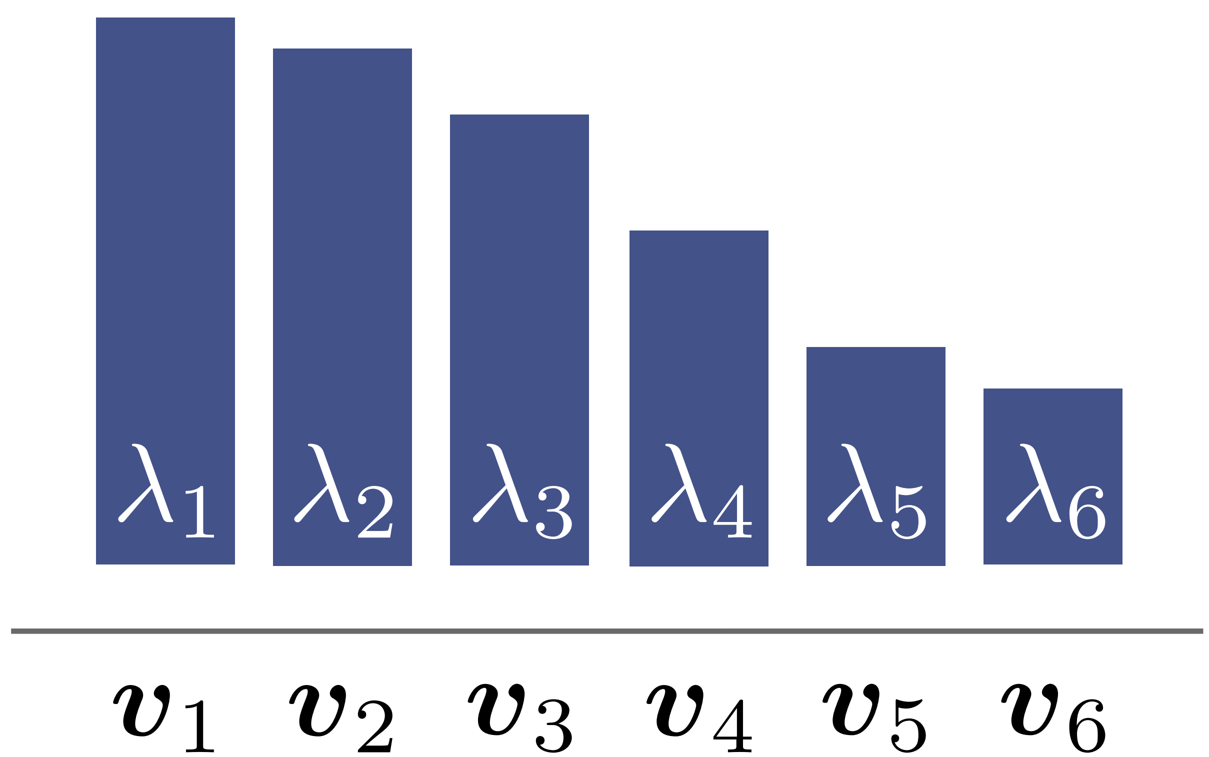
DPP

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

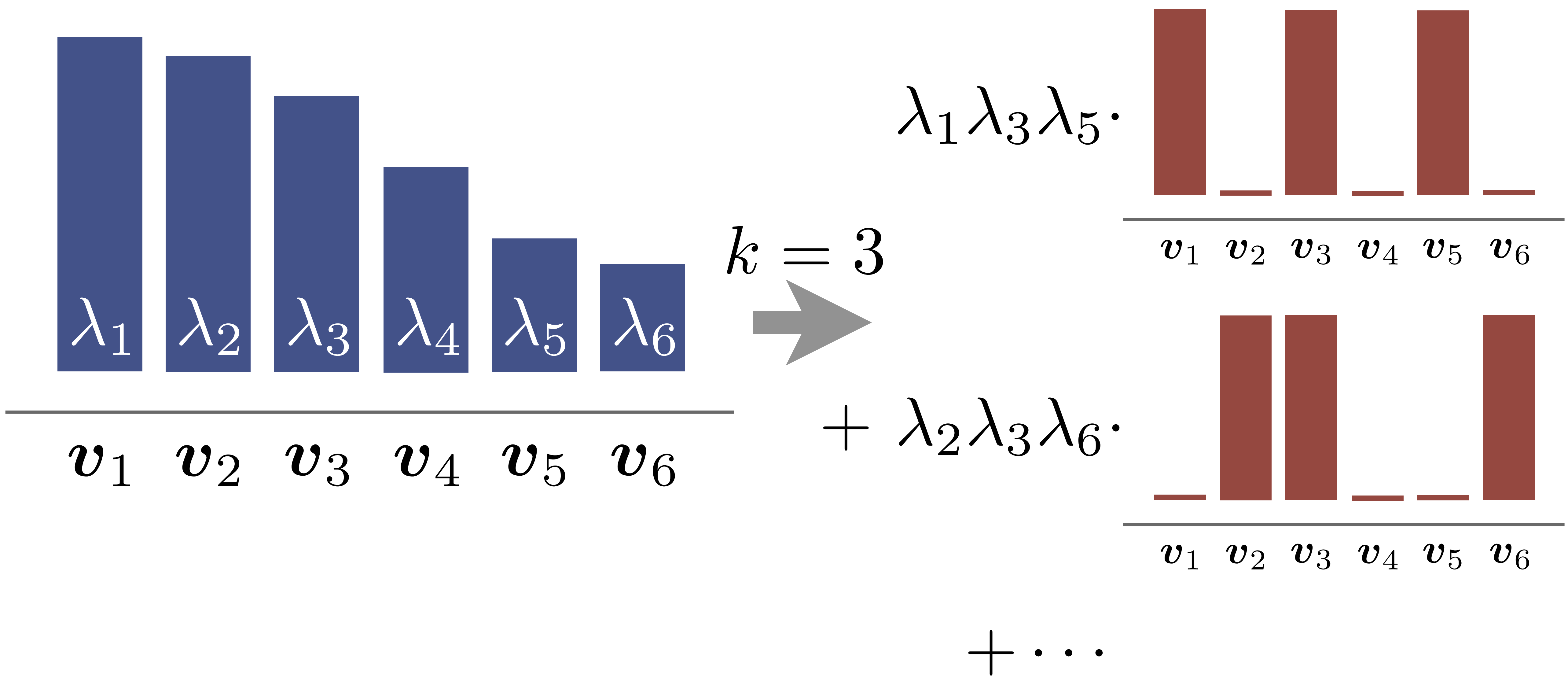
k-DPP

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

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



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



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:

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

- Runtime of new PHASE ONE is $O(Nk)$
- PHASE TWO is unchanged

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

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

Length

Quality features

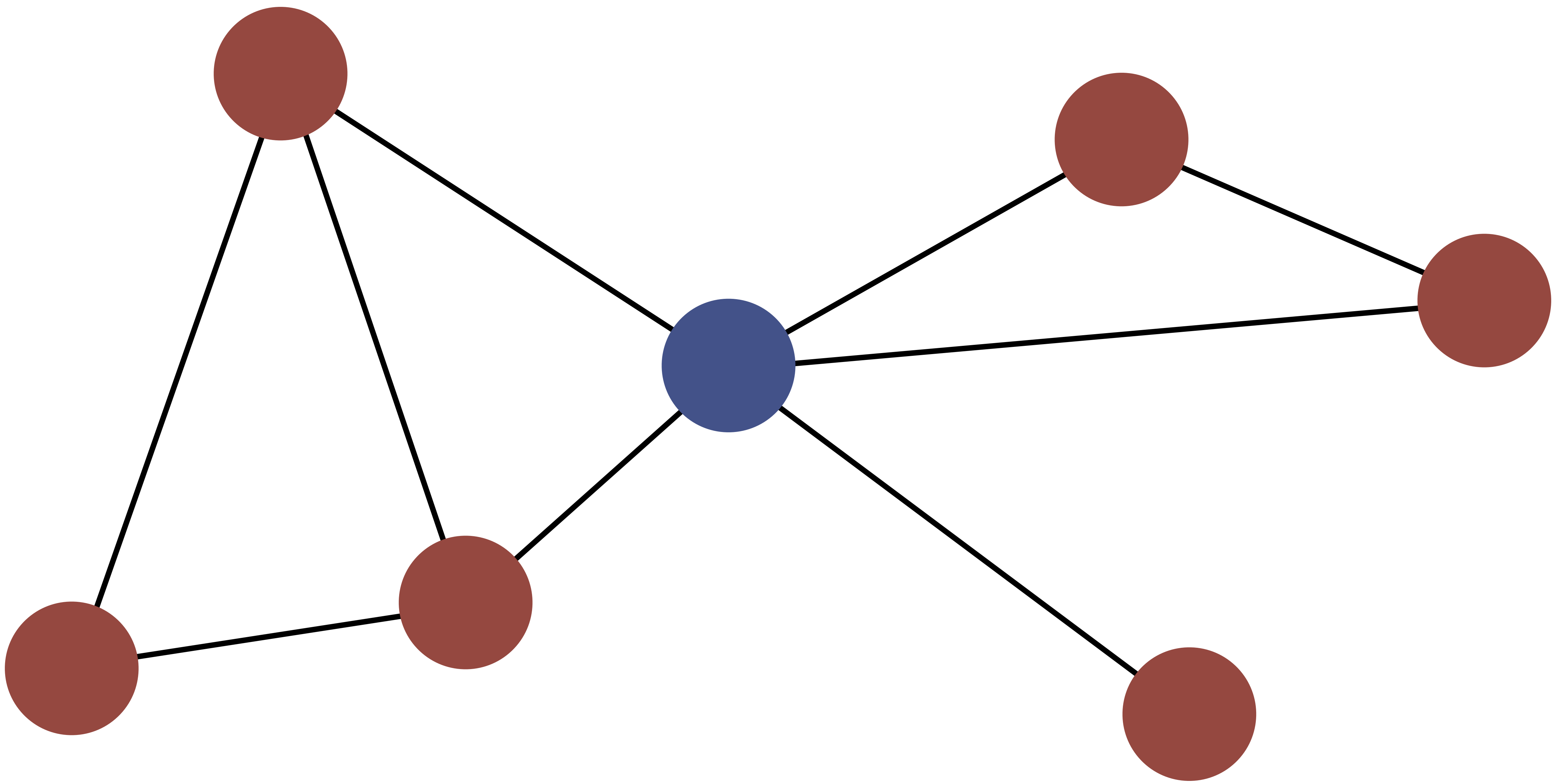
2. Pizza Hut has released the limited edition dish after the success of its cheese and BBQ crusts.

Position in article

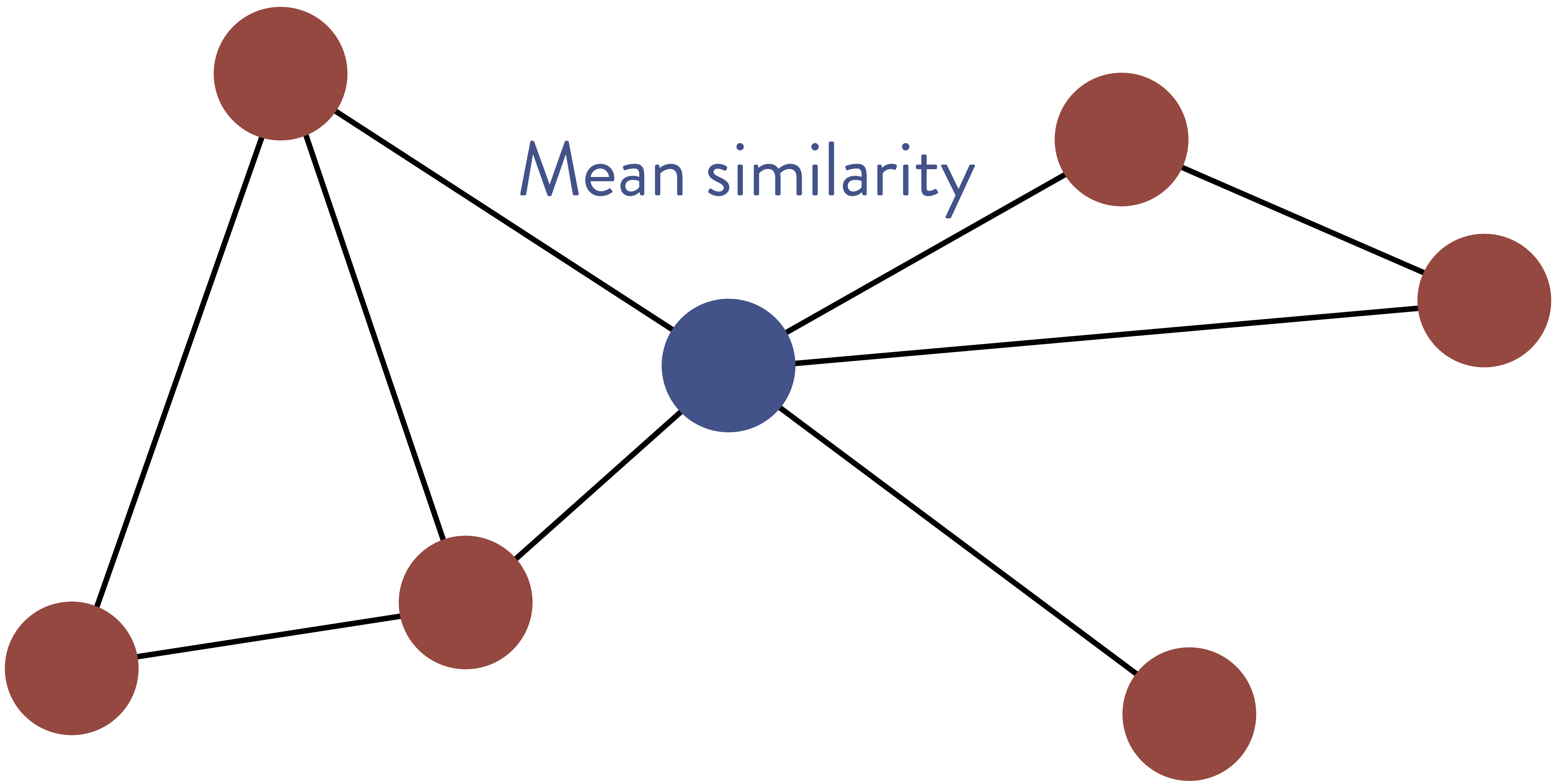
3. Dubbed the "pizza dog", the 14-inch feast is only available for delivery and costs up to £19.49.

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

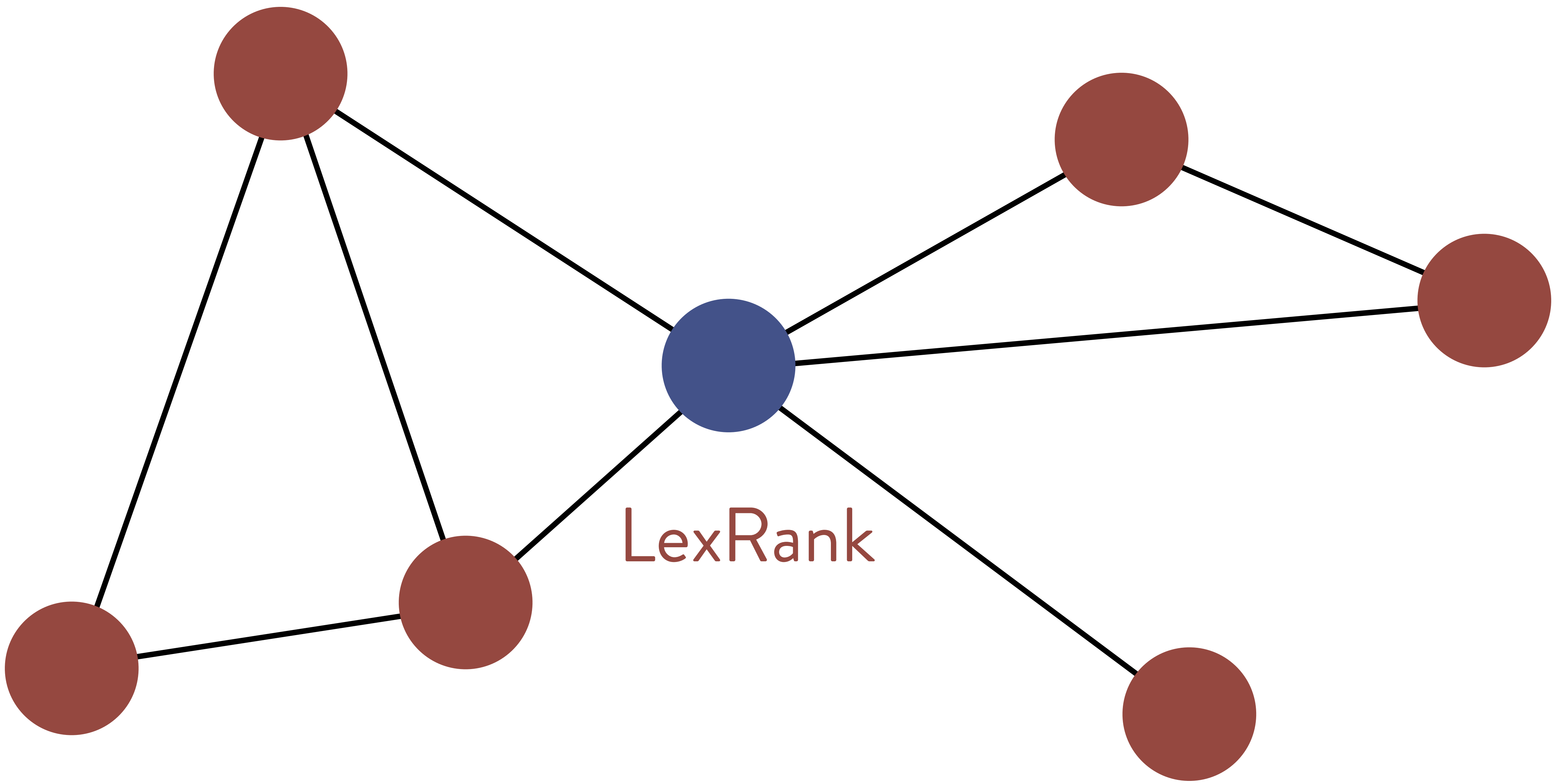
Quality features



Quality features



Quality features



Diversity features

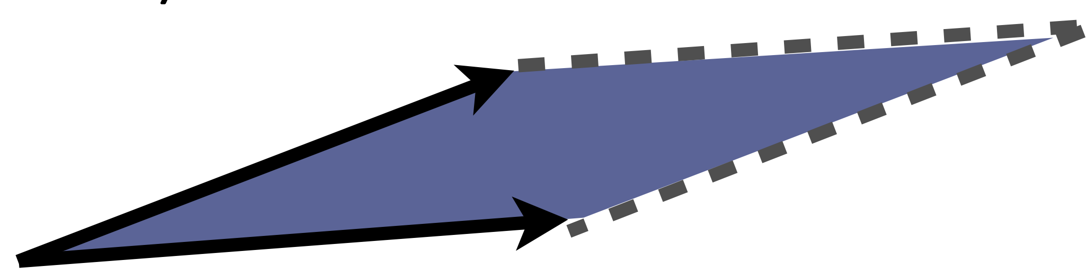
- ϕ fixed to tf-idf vectors: cosine similarity

$$\phi\left(\begin{array}{l}\text{Dubbed the “pizza dog”, the 14-inch feast is only}\\ \text{available for delivery and costs up to £19.49.}\end{array}\right)$$

Diversity features

- ϕ fixed to tf-idf vectors: cosine similarity

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

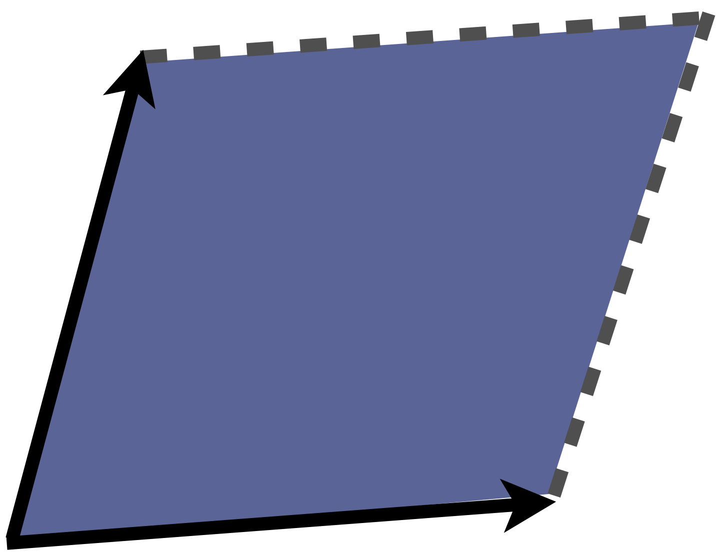


Dubbed the "pizza dog", the 14-inch feast is only available for delivery and costs up to £19.49.

Diversity features

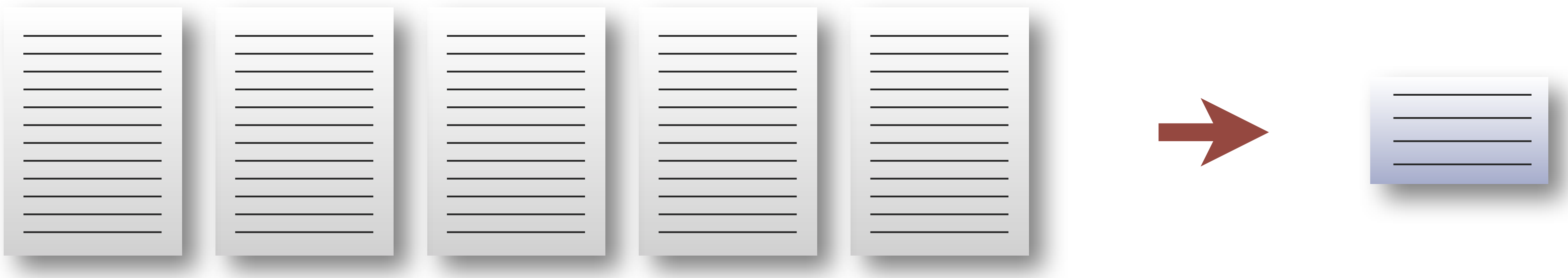
- ϕ fixed to tf-idf vectors: cosine similarity

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.

News summarization



- **Input**: 10 news articles, ~250 sentences
- **Output**: 665 character summary
- **Eval**: ROUGE metric (four human summaries)
- Learn on DUC 03, test on DUC 04 data

System	ROUGE-1F	ROUGE-1R	R-SU4F
Begin	32.08	32.69	10.37
MMR*	37.58	38.05	13.06
Best in 2004	37.87	38.20	13.19
SubMod**	38.90	39.35	-

[*Carbonell and Goldstein, 1998] [**Lin and Bilmes, 2012]

System	ROUGE-1F	ROUGE-1R	R-SU4F
Begin	32.08	32.69	10.37
MMR*	37.58	38.05	13.06
Best in 2004	37.87	38.20	13.19
SubMod**	38.90	39.35	-
DPP MAP	38.96	39.15	13.83

[*Carbonell and Goldstein, 1998] [**Lin and Bilmes, 2012]

System	ROUGE-1F	ROUGE-1R	R-SU4F
Begin	32.08	32.69	10.37
MMR*	37.58	38.05	13.06
Best in 2004	37.87	38.20	13.19
SubMod**	38.90	39.35	-
DPP MAP	38.96	39.15	13.83
DPP MinRisk	**40.33**	**41.31**	**14.13**

[*Carbonell and Goldstein, 1998] [**Lin and Bilmes, 2012]

Determinantal point processes

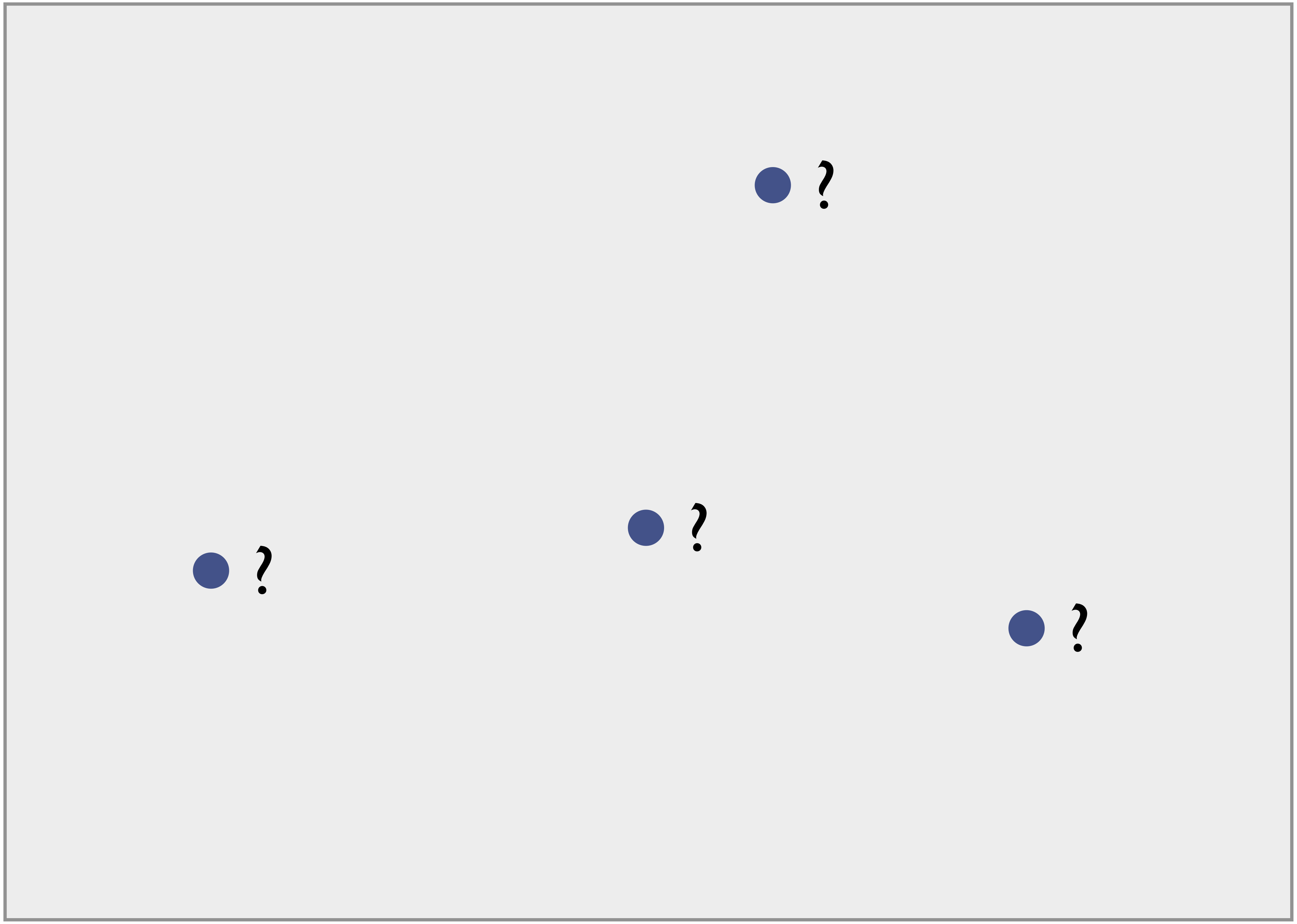
Quality, diversity, and learning

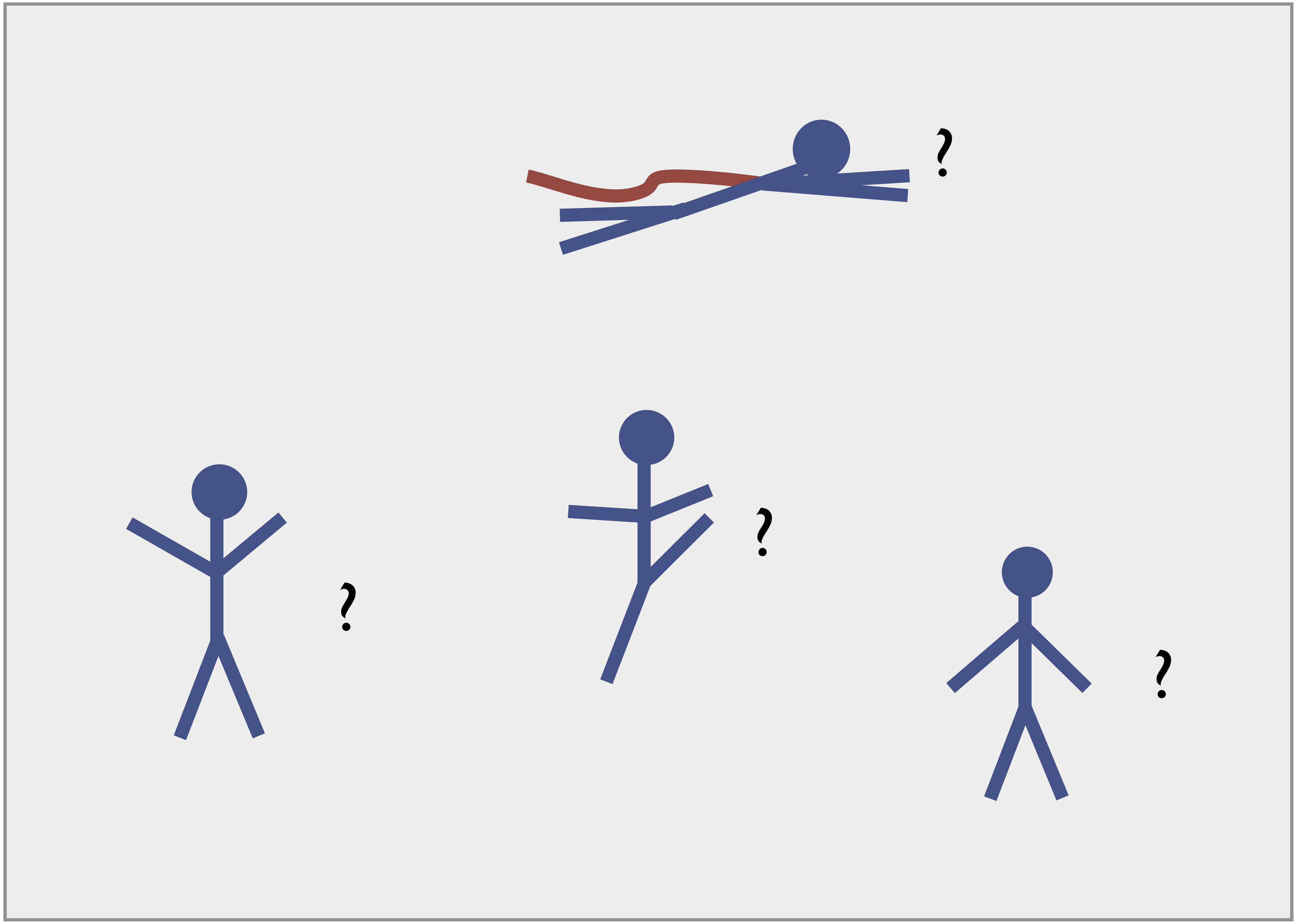
Sampling

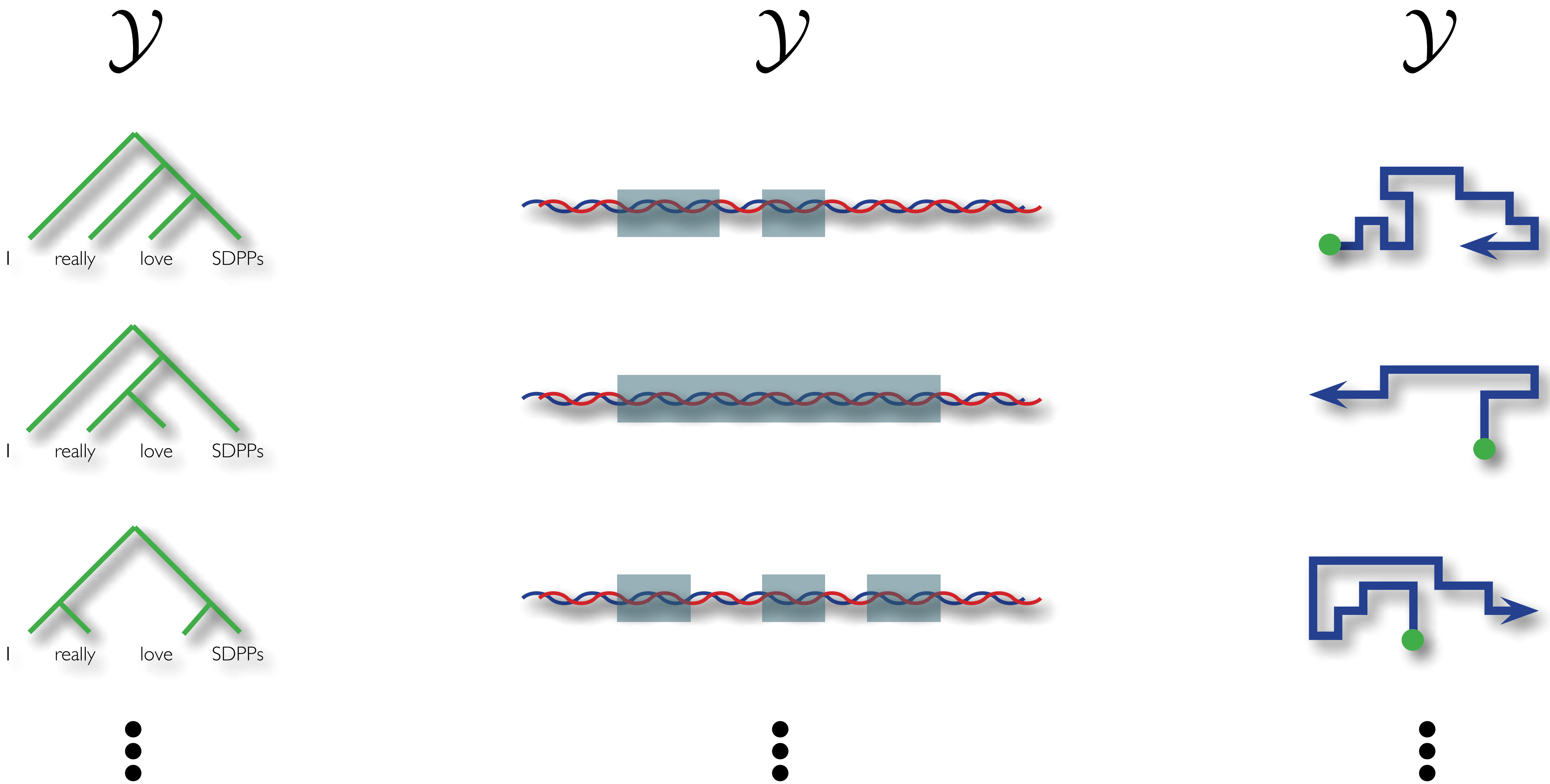
k-DPPs

Structured DPPs

News threading






$\mathcal{Y}$
$\mathcal{Y}$
$\mathcal{Y}$
I really love SDPPs
I really love SDPPs
I really love SDPPs

Structured DPPs

- Exponentially many complex “items”
- Can’t even write down N x N kernel
- But can still compute marginals and sample!

Structured DPPs

- Exponentially many complex "items"
- Can't even write down N x N kernel
- But can still compute marginals and sample!

1. Factorized model
2. Dual representation of L
3. Second order message-passing

1. Factorization

- Quality scores factor multiplicatively:

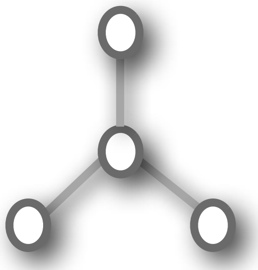
$$q(i) = \prod_{v \in \mathcal{V}} q_v(i_v) \prod_{vu \in \mathcal{E}} q_{vu}(i_v, i_u)$$

- Diversity features factor additively:

1. Factorization

- Quality scores factor multiplicatively:

$$q(i) = \prod_{v \in \mathcal{V}} q_v(i_v) \prod_{vu \in \mathcal{E}} q_{vu}(i_v, i_u)$$

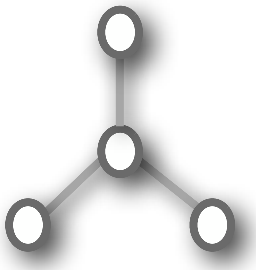
e.g., tree 

- Diversity features factor additively:

1. Factorization

- Quality scores factor multiplicatively:

$$q(i) = \prod_{v \in \mathcal{V}} q_v(i_v) \prod_{vu \in \mathcal{E}} q_{vu}(i_v, i_u)$$

e.g., tree 

- Diversity features factor additively:

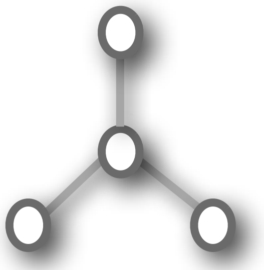
$$\phi(i) = \sum_{v \in \mathcal{V}} \phi_v(i_v) + \sum_{vu \in \mathcal{E}} \phi_{vu}(i_v, i_u)$$

1. Factorization

- Quality scores factor multiplicatively:

$$q(i) = \prod_{v \in \mathcal{V}} q_v(i_v) \prod_{vu \in \mathcal{E}} q_{vu}(i_v, i_u)$$

e.g., tree



- Diversity features factor additively:

$$\phi(i) = \sum_{v \in \mathcal{V}} \phi_v(i_v) + \sum_{vu \in \mathcal{E}} \phi_{vu}(i_v, i_u)$$

e.g., $\phi(i)^\top \phi(j)$

spatial overlap

Quality

Quality

Quality

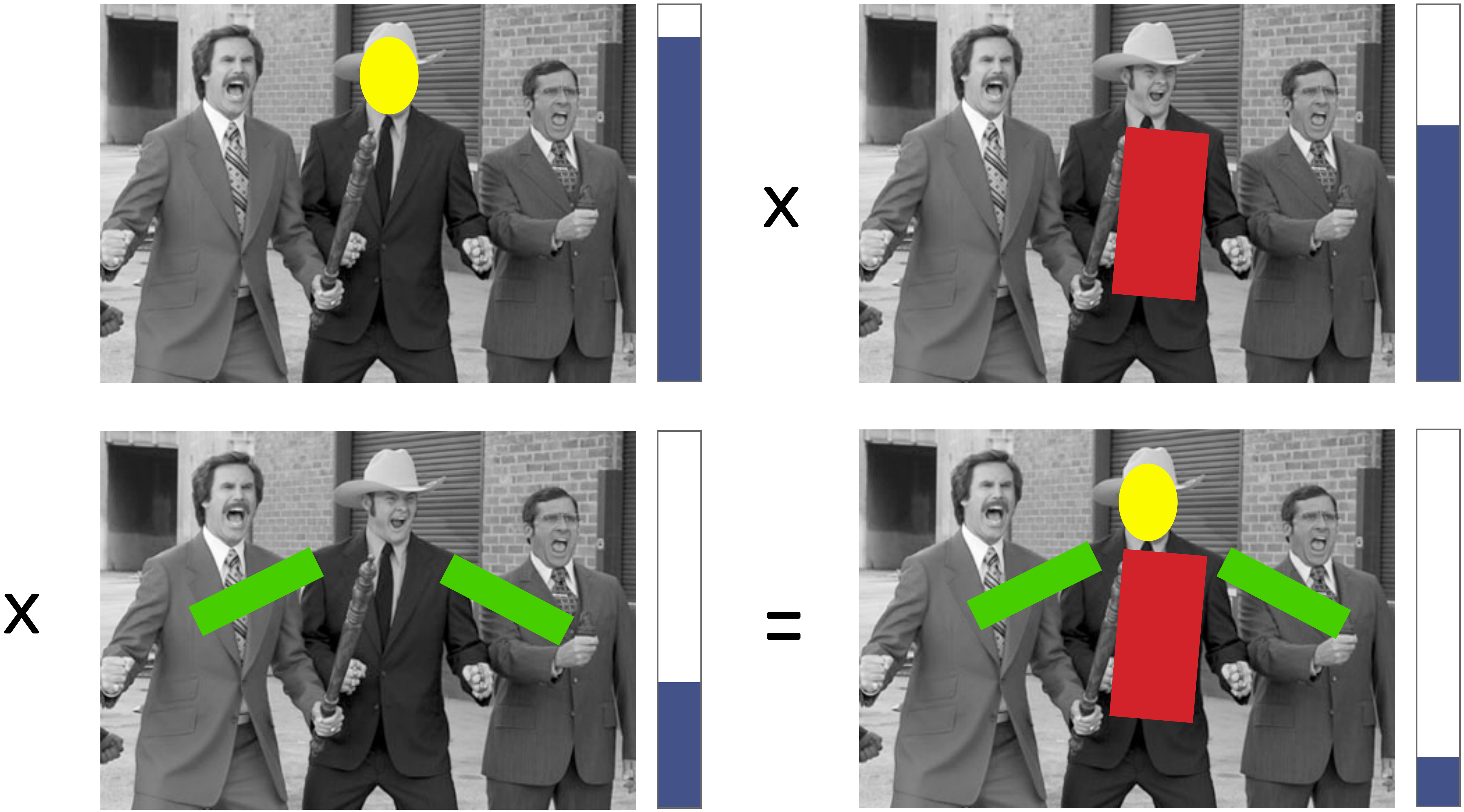
x

Quality

x

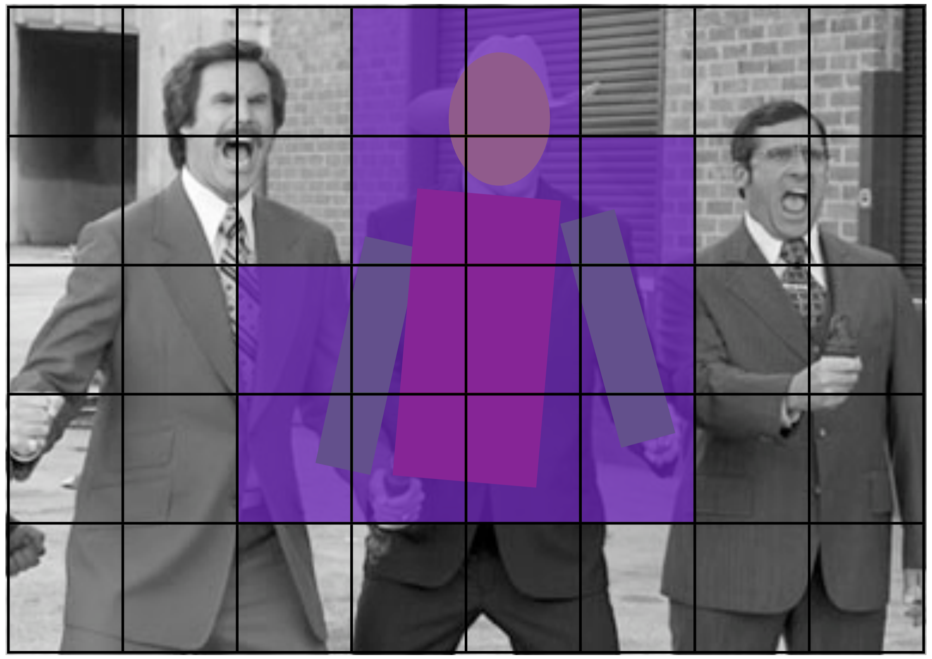
x

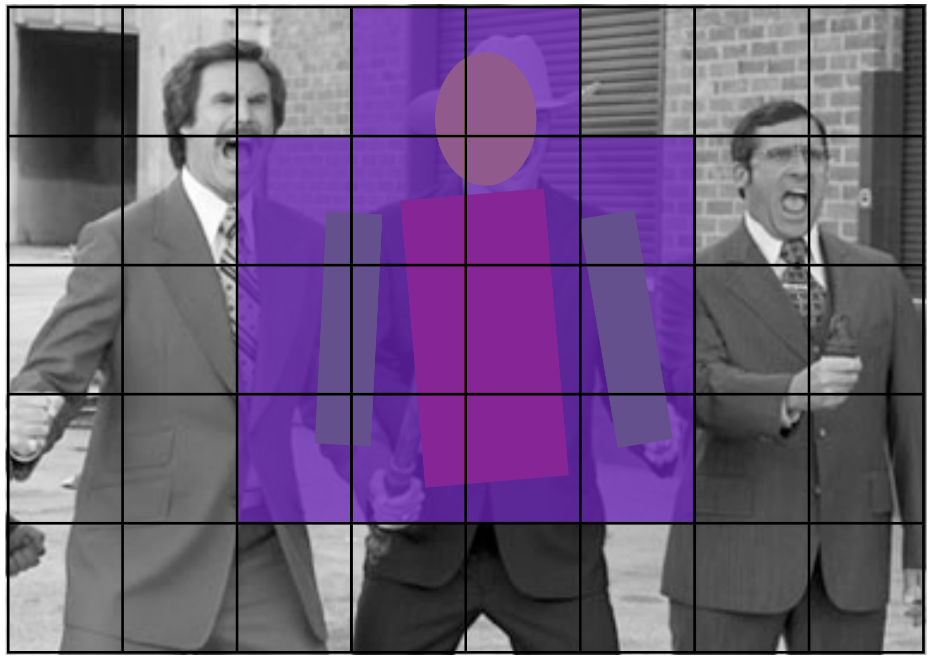
Quality



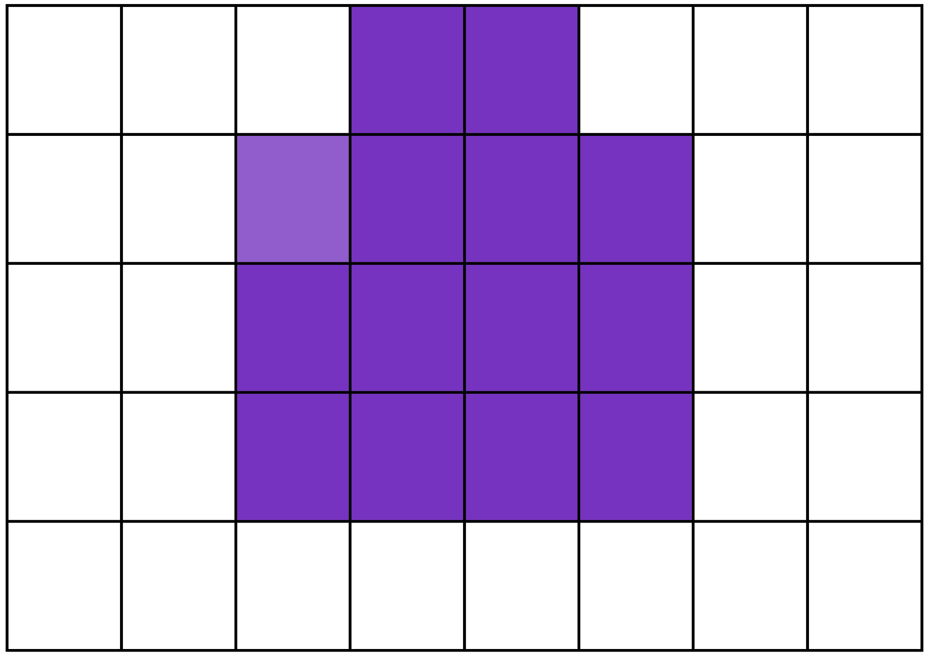
Diversity

Diversity



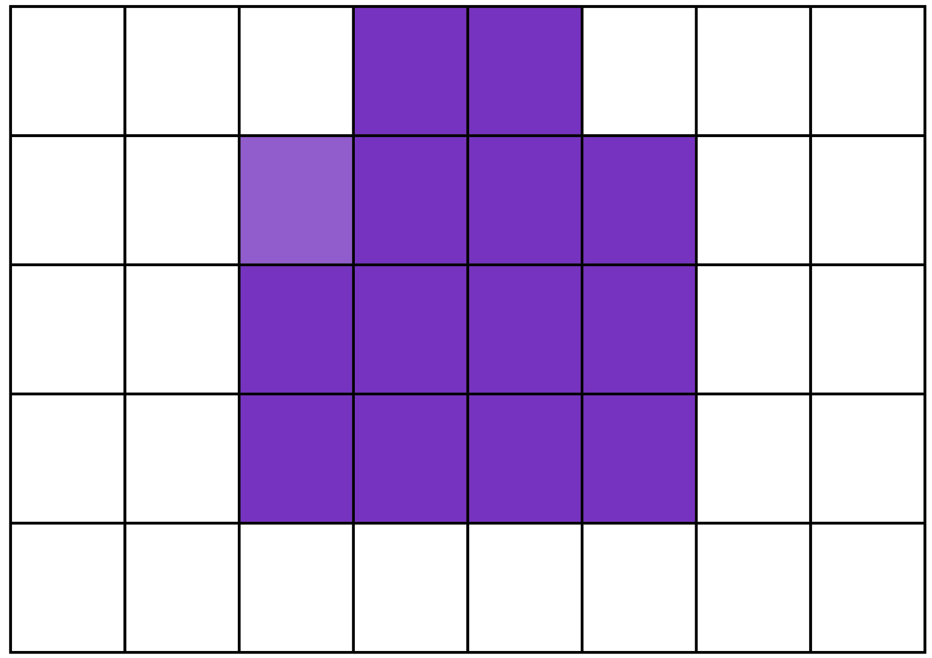


Diversity

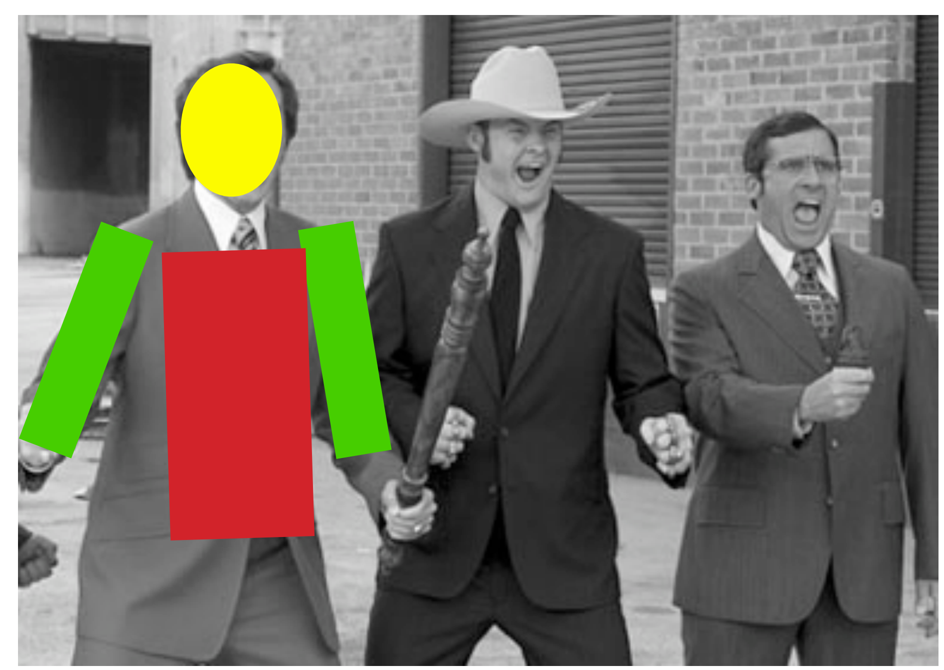


Low diversity

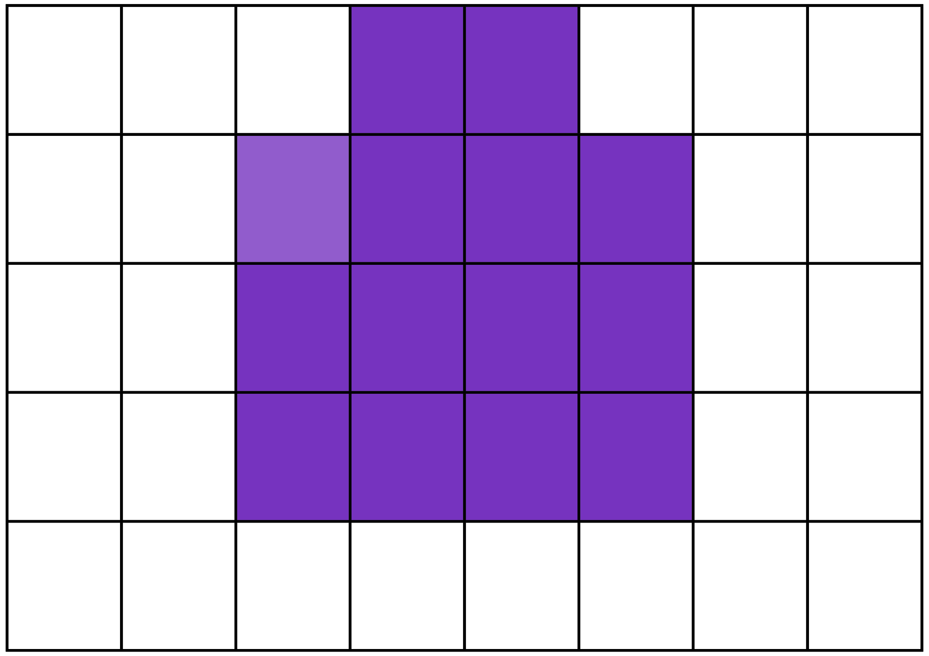
Diversity



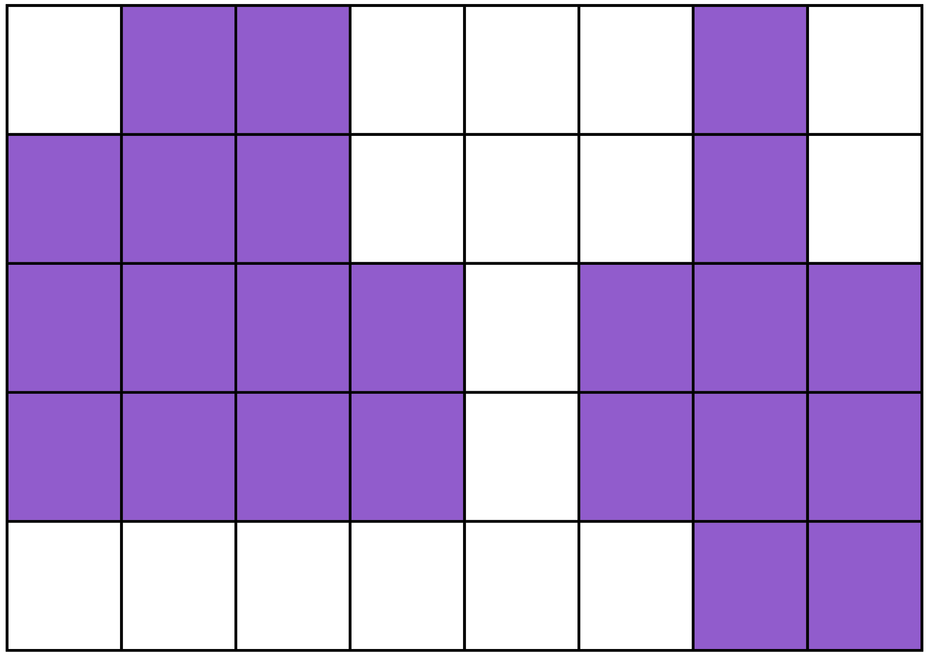
Low diversity



Diversity

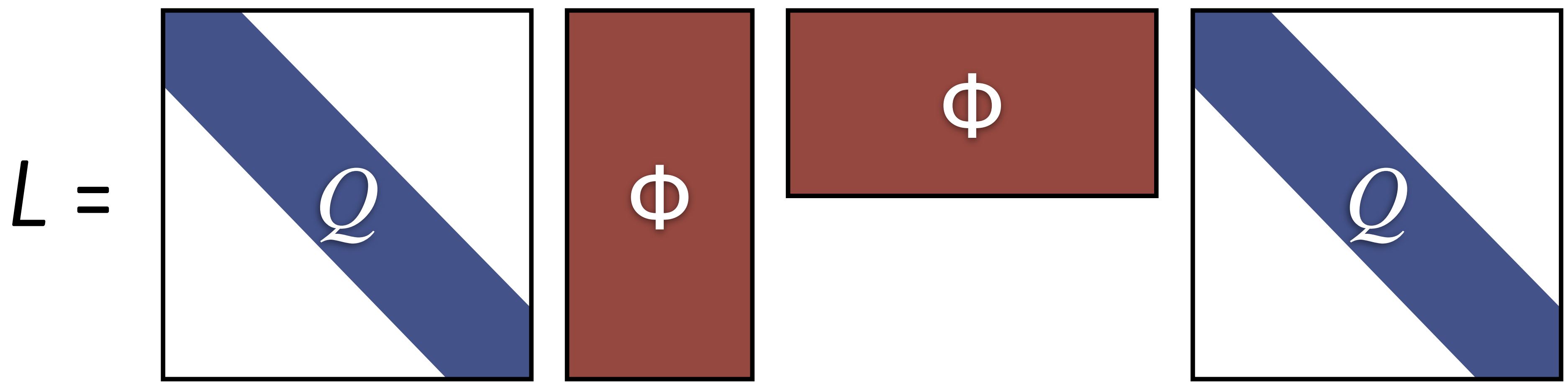


Low diversity



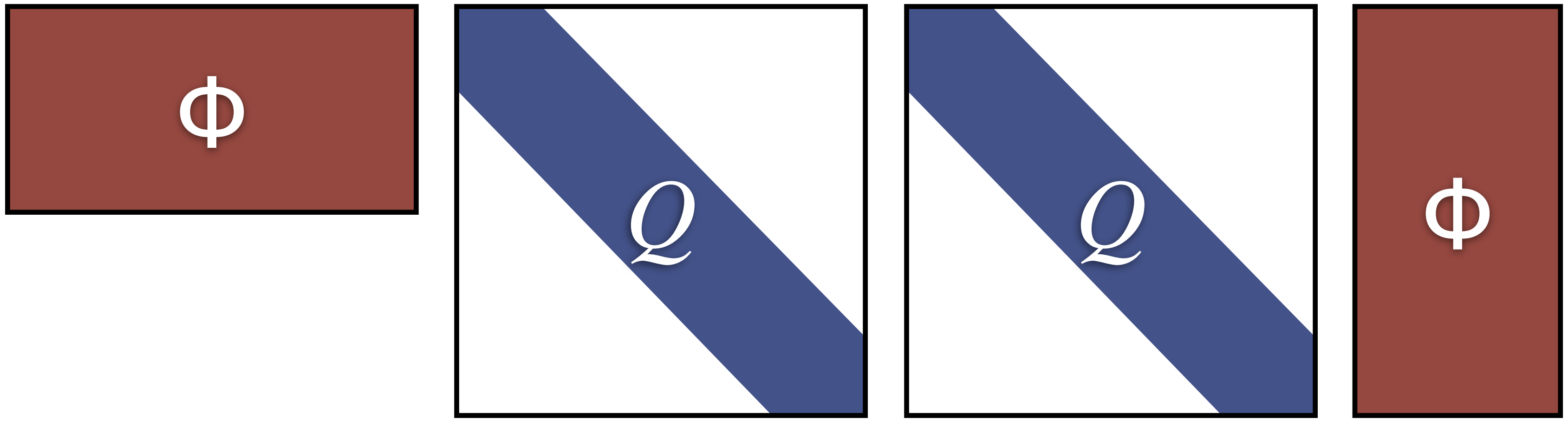
High diversity

2. Dual representation

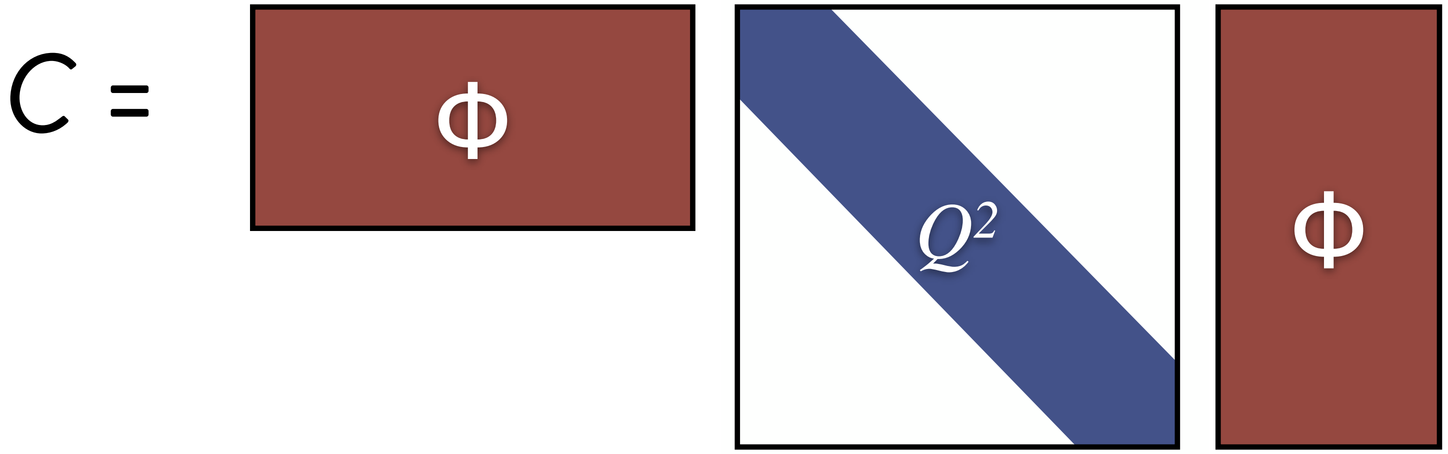


$$L_{ij} = q(i)\phi(i)^\top \phi(j)q(j)$$

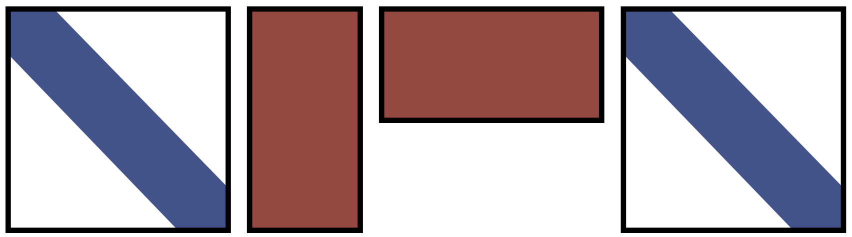
2. Dual representation



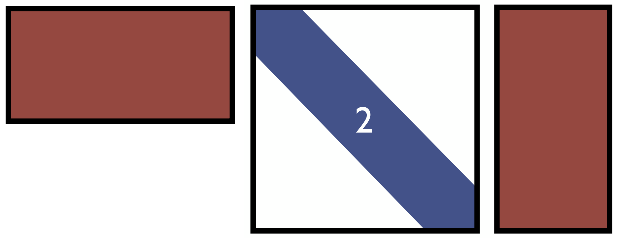
2. Dual representation



2. Dual representation

$L =$ 

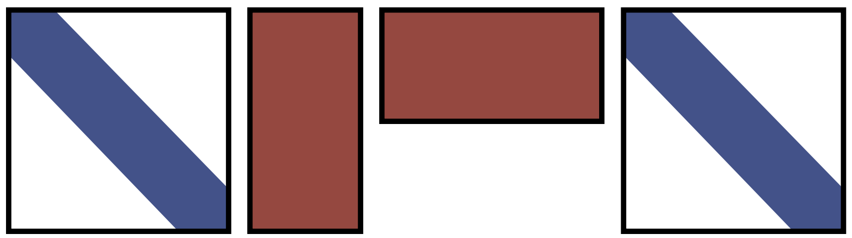
$N \times N$

$C =$ 

$D \times D$

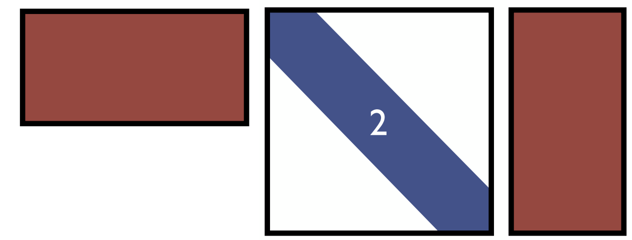
2. Dual representation

L =



N x *N*

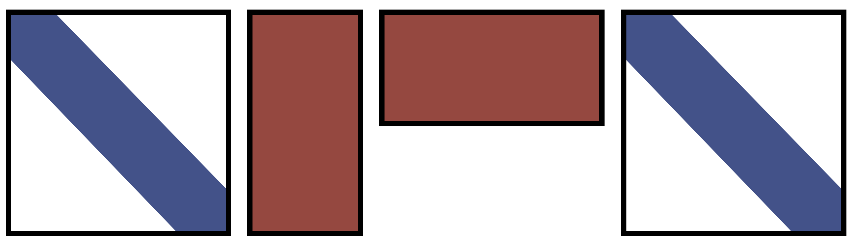
C =



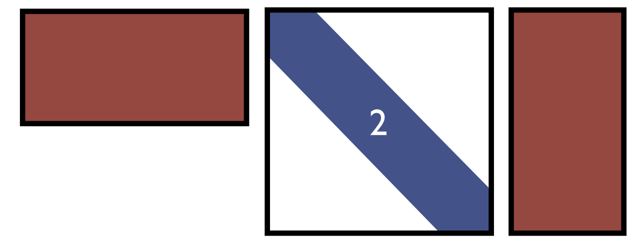
D x *D*

- C and *L* have the same non-zero eigenvalues, and related eigenvectors
- Use C for sampling and other inference!

2. Dual representation

$L =$ 

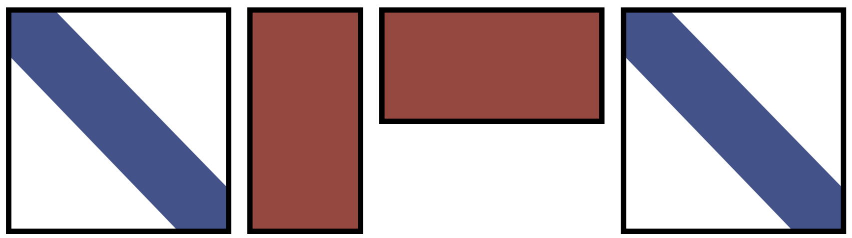
$N \times N$

$C =$ 

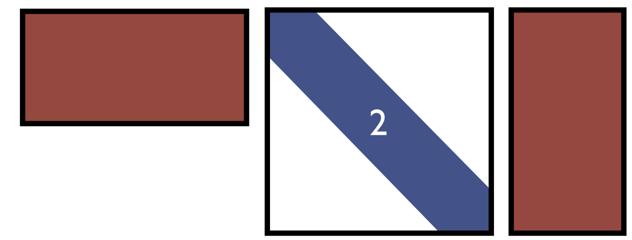
$D \times D$

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

2. Dual representation

$L =$ 

$N \times N$

$C =$ 

$D \times D$

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

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

3. Second-order message passing

3. Second-order message passing

- Can compute feature covariance using message passing **if** q is a tree

3. Second-order message passing

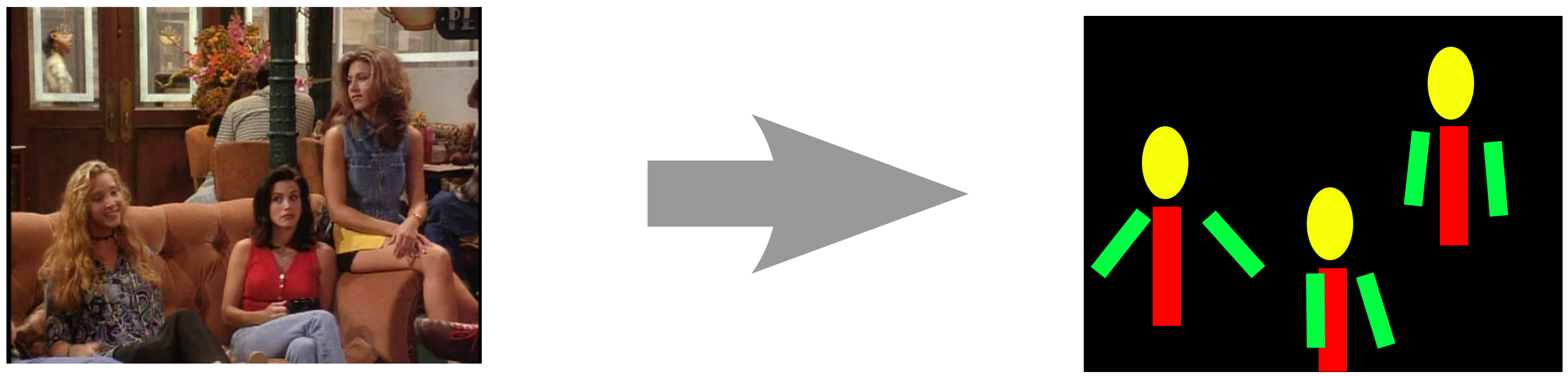
- Can compute feature covariance using message passing **if** q is a tree
- Use special semiring sum-product [Li & Eisner,09]

3. Second-order message passing

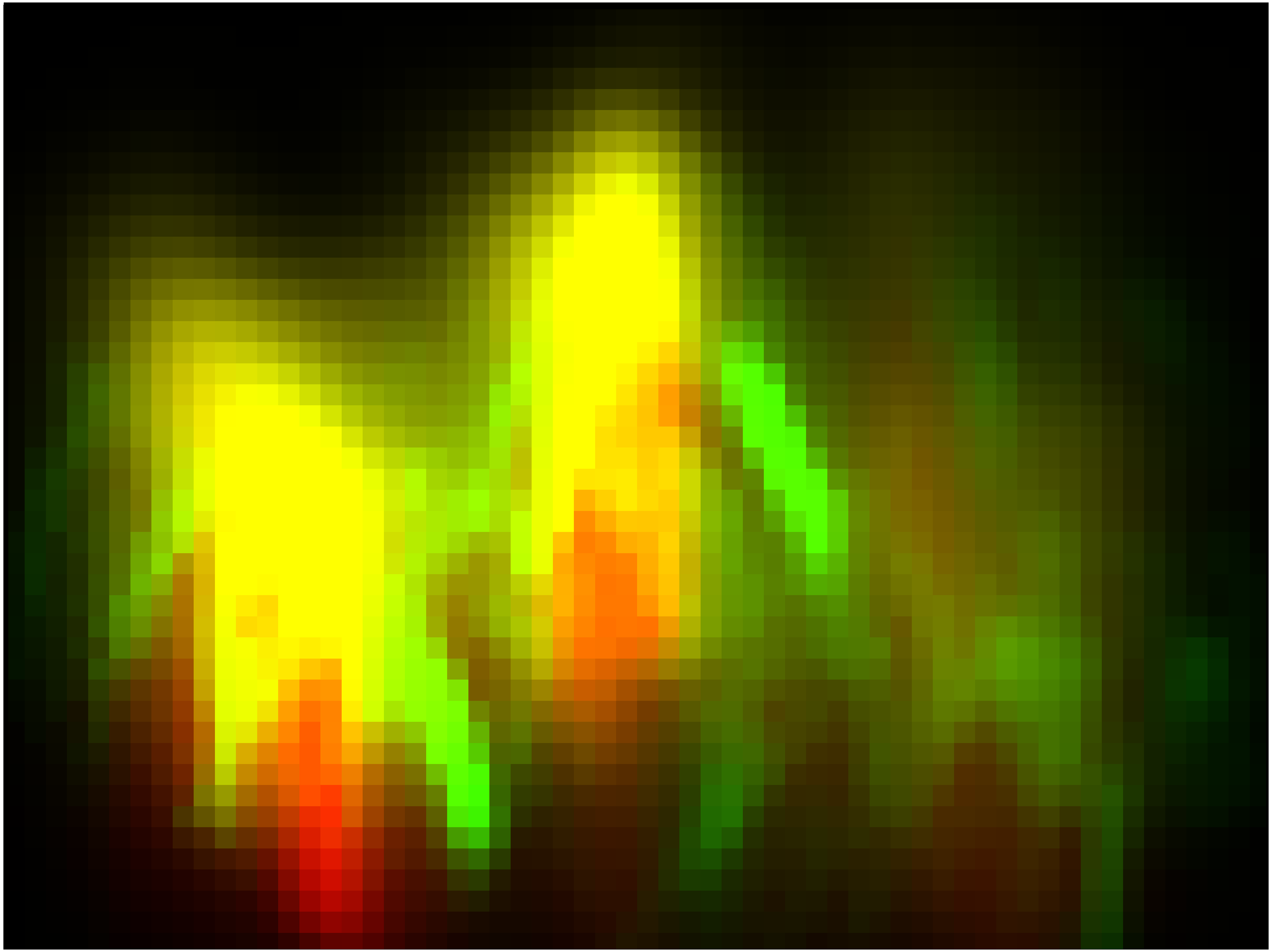
- Can compute feature covariance using message passing **if** q is a tree
- Use special semiring sum-product [Li & Eisner,09]
- Linear in number of nodes
- Quadratic in number of diversity features D

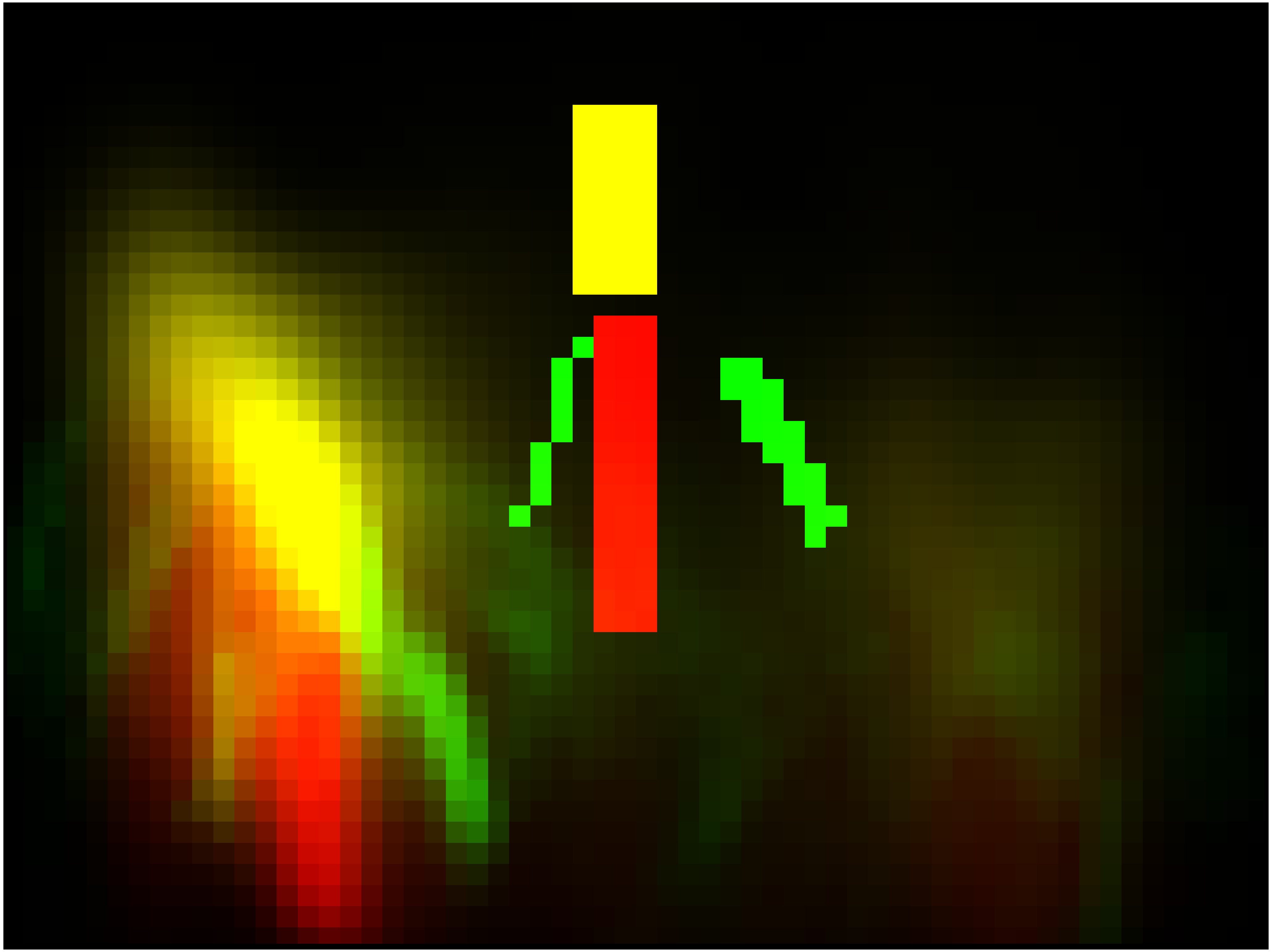
$$O(D^2 \log N)$$

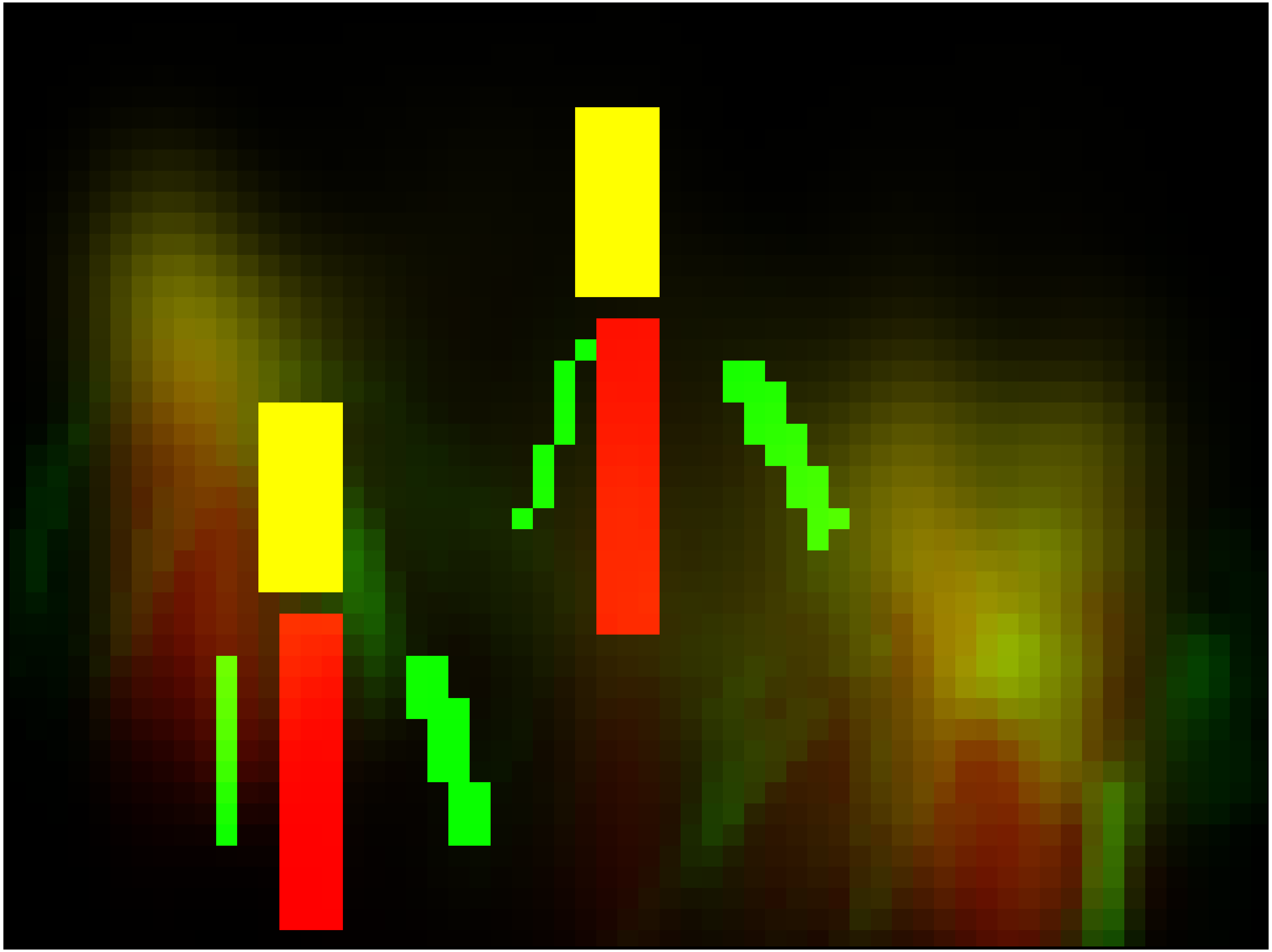
Multiple-pose estimation

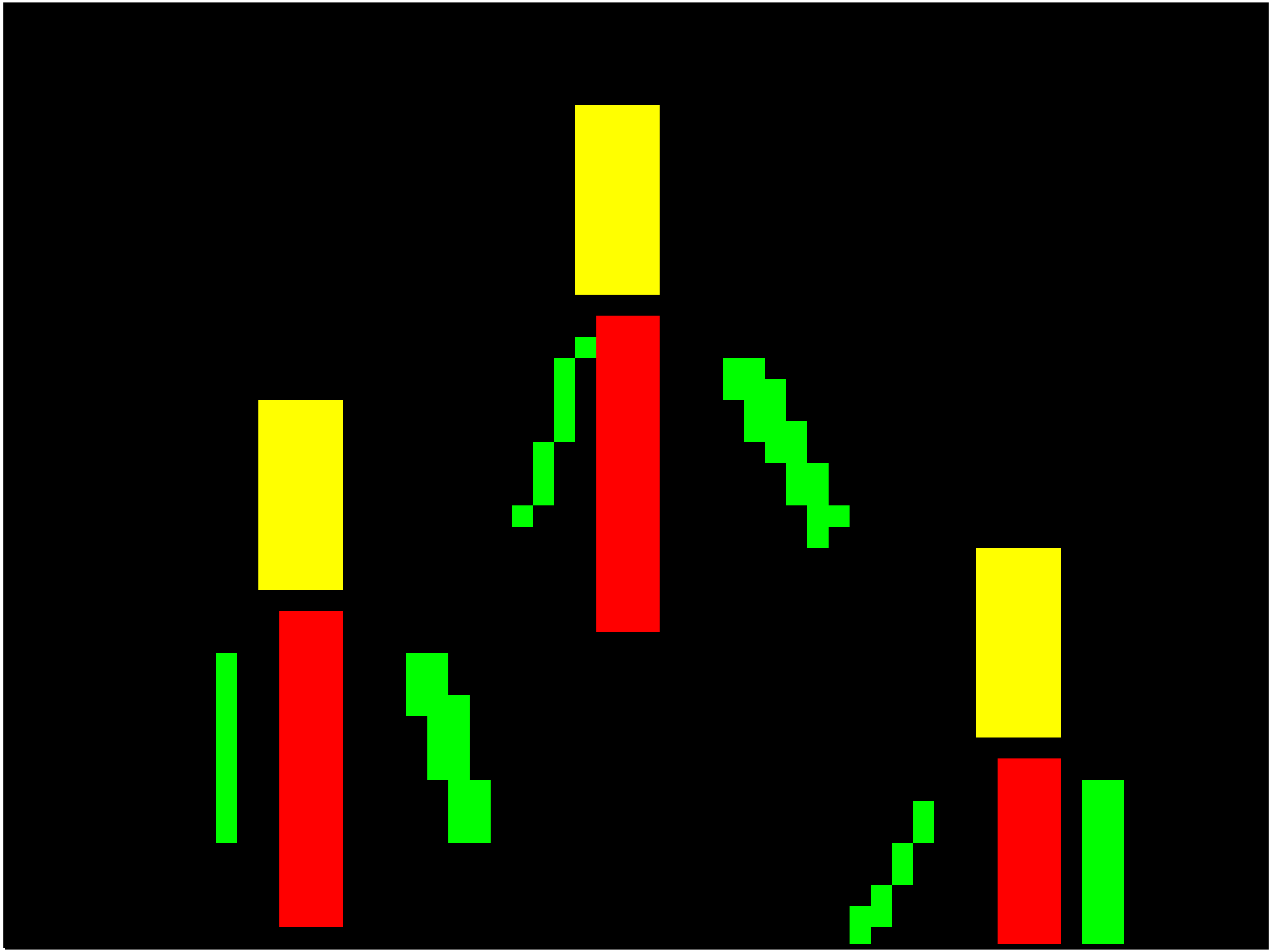


- Images from TV shows
  - 3+ people/image, similar scale, hand labeled
- Trained quality model, spatial diversity model

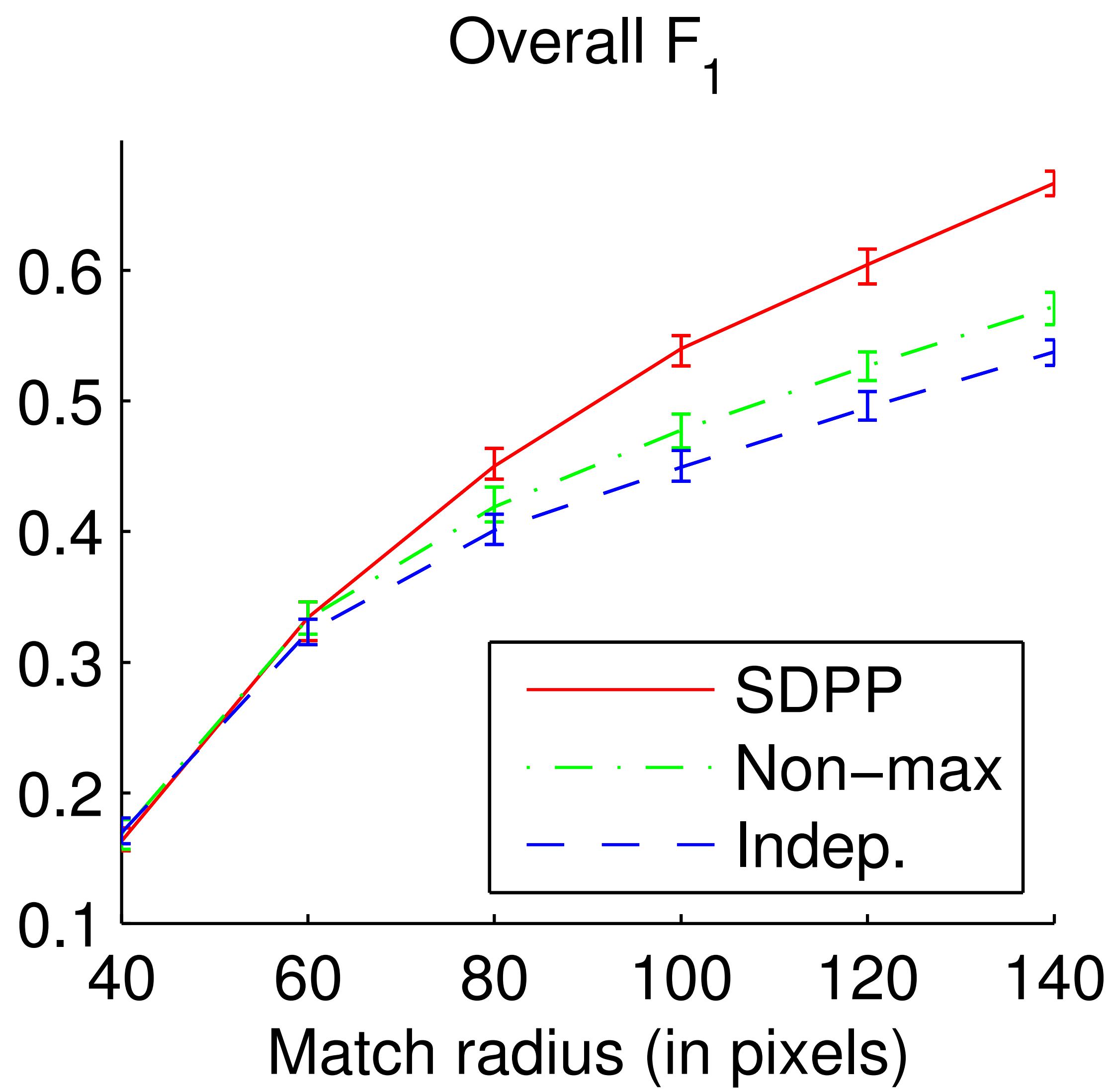




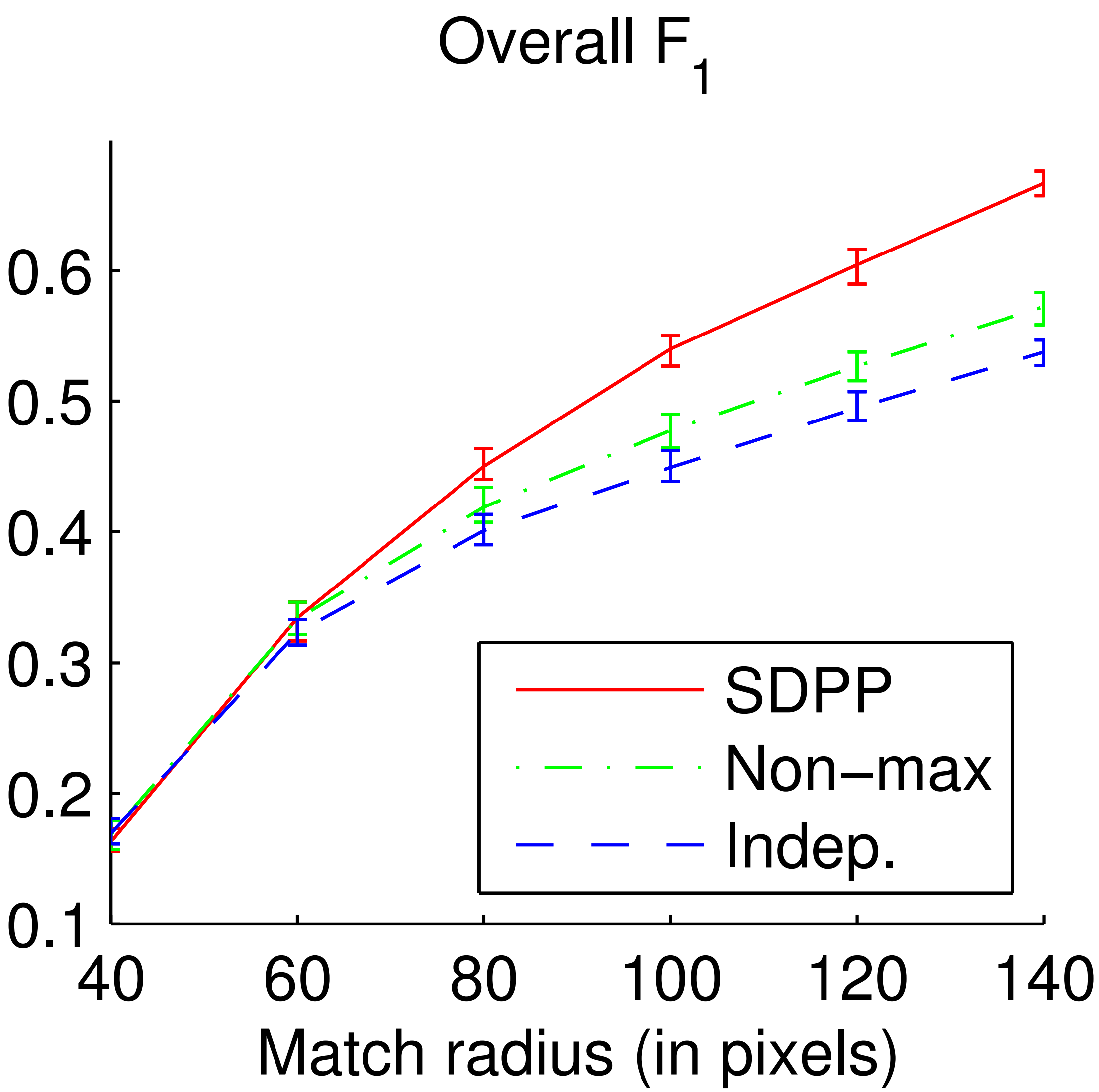


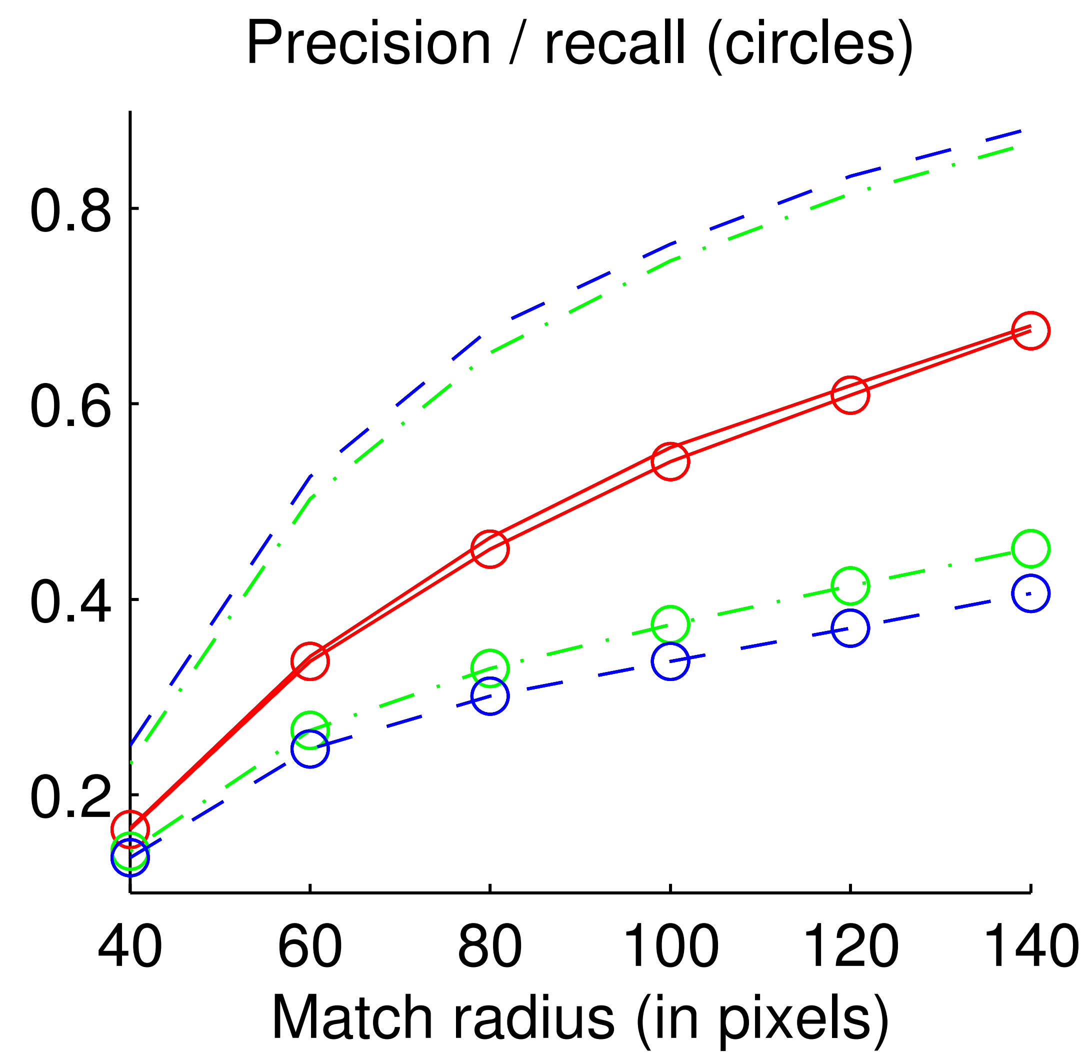


Pose accuracy



Pose accuracy





Determinantal point processes

Quality, diversity, and learning

Sampling

k-DPPs

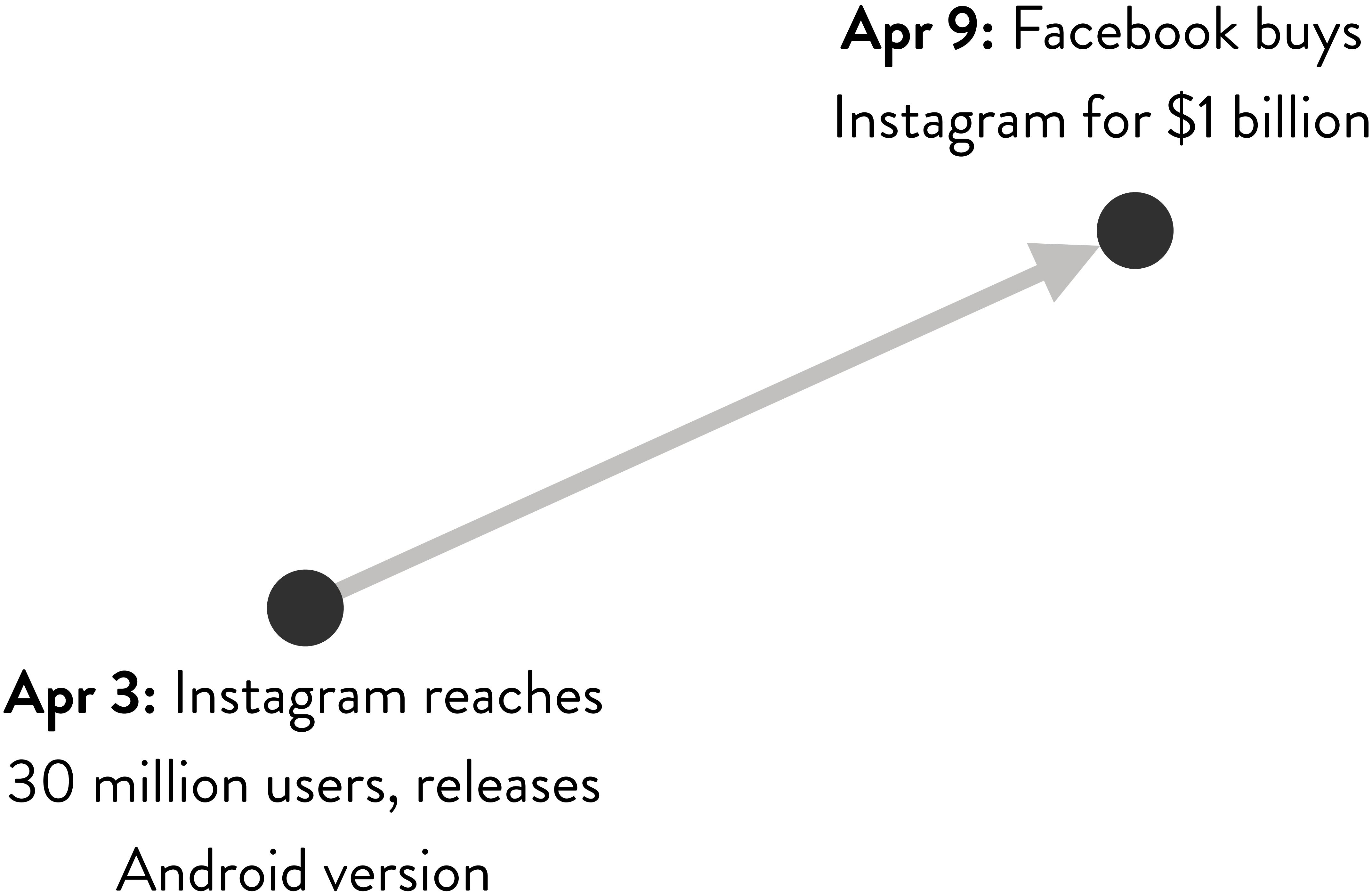
Structured DPPs

News threading

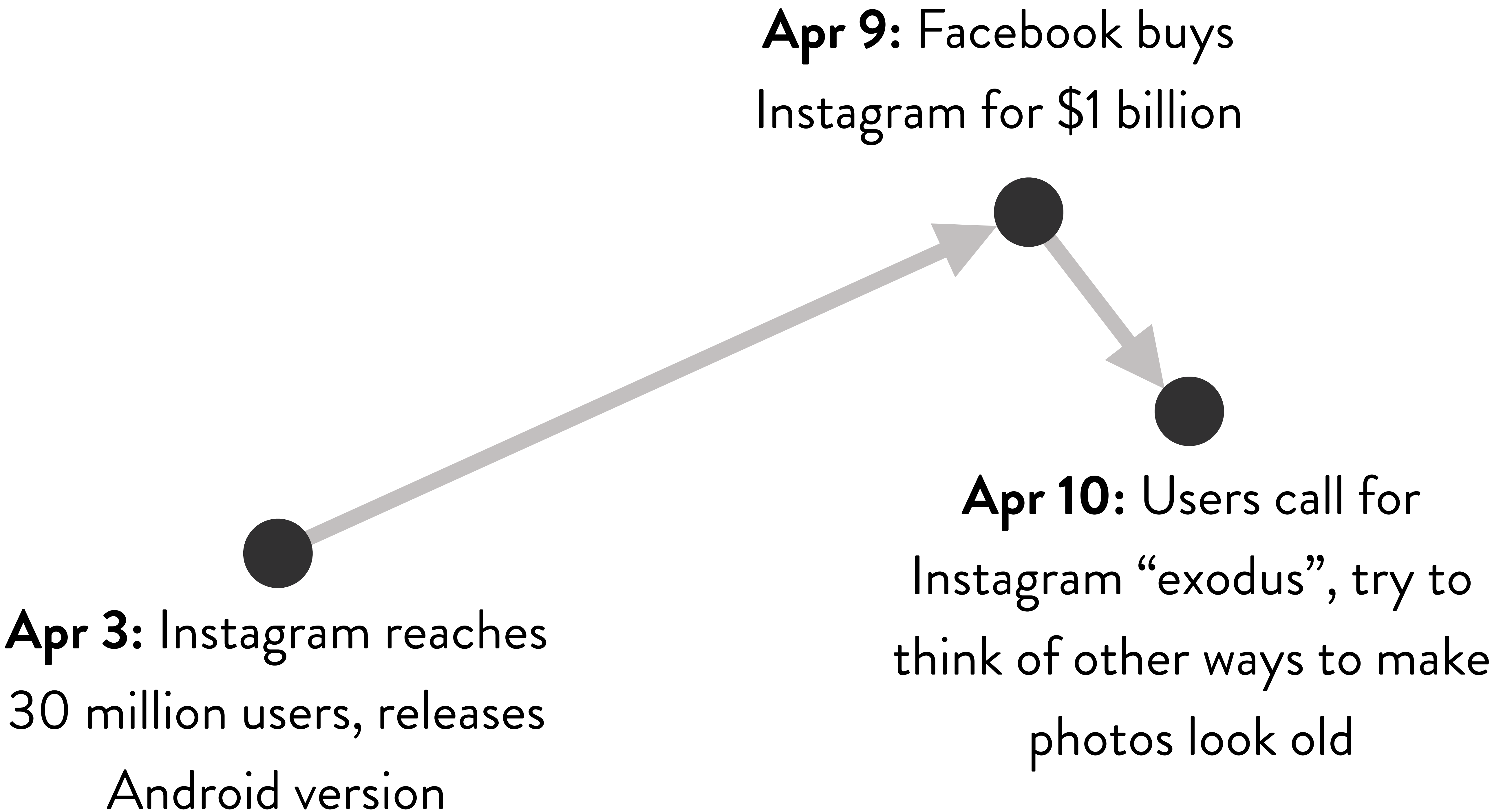
News happens

Apr 3: Instagram reaches 30 million users, releases Android version

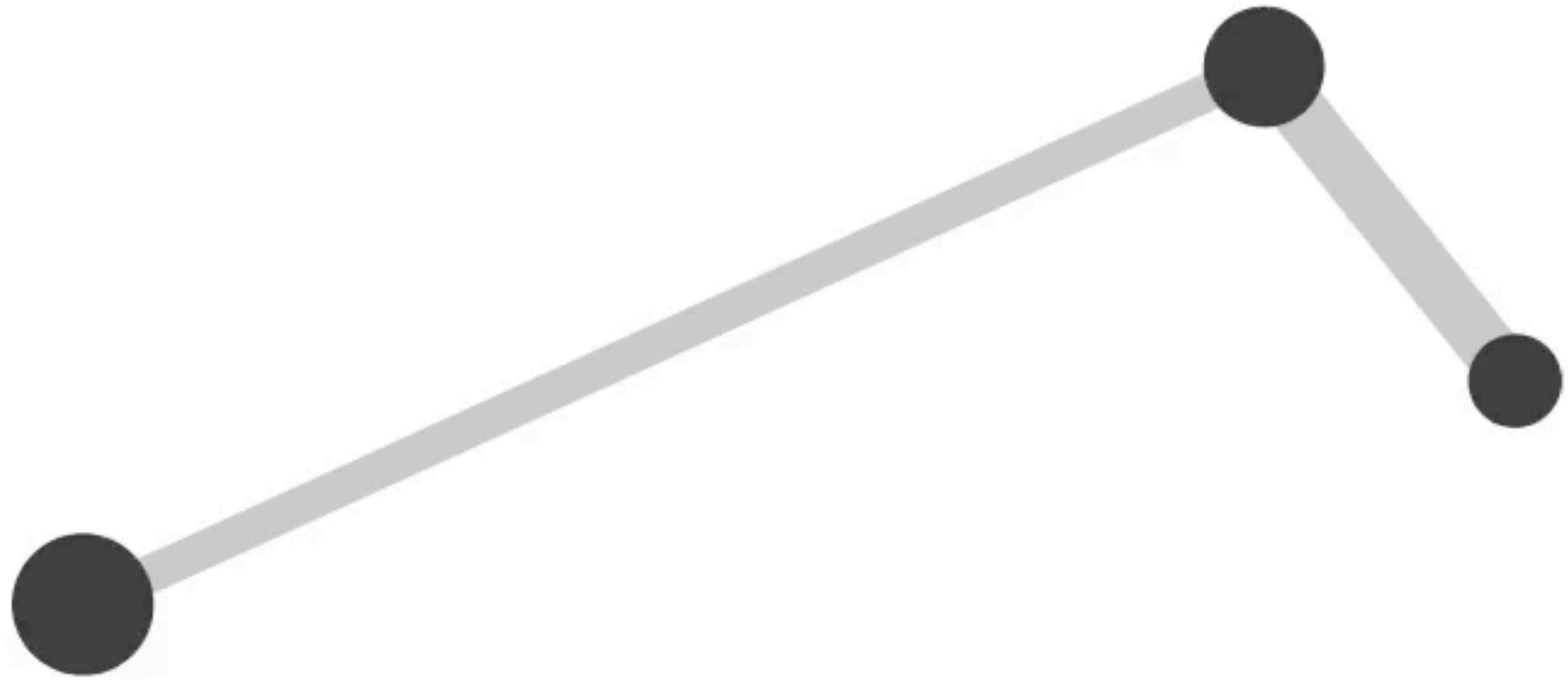
News happens



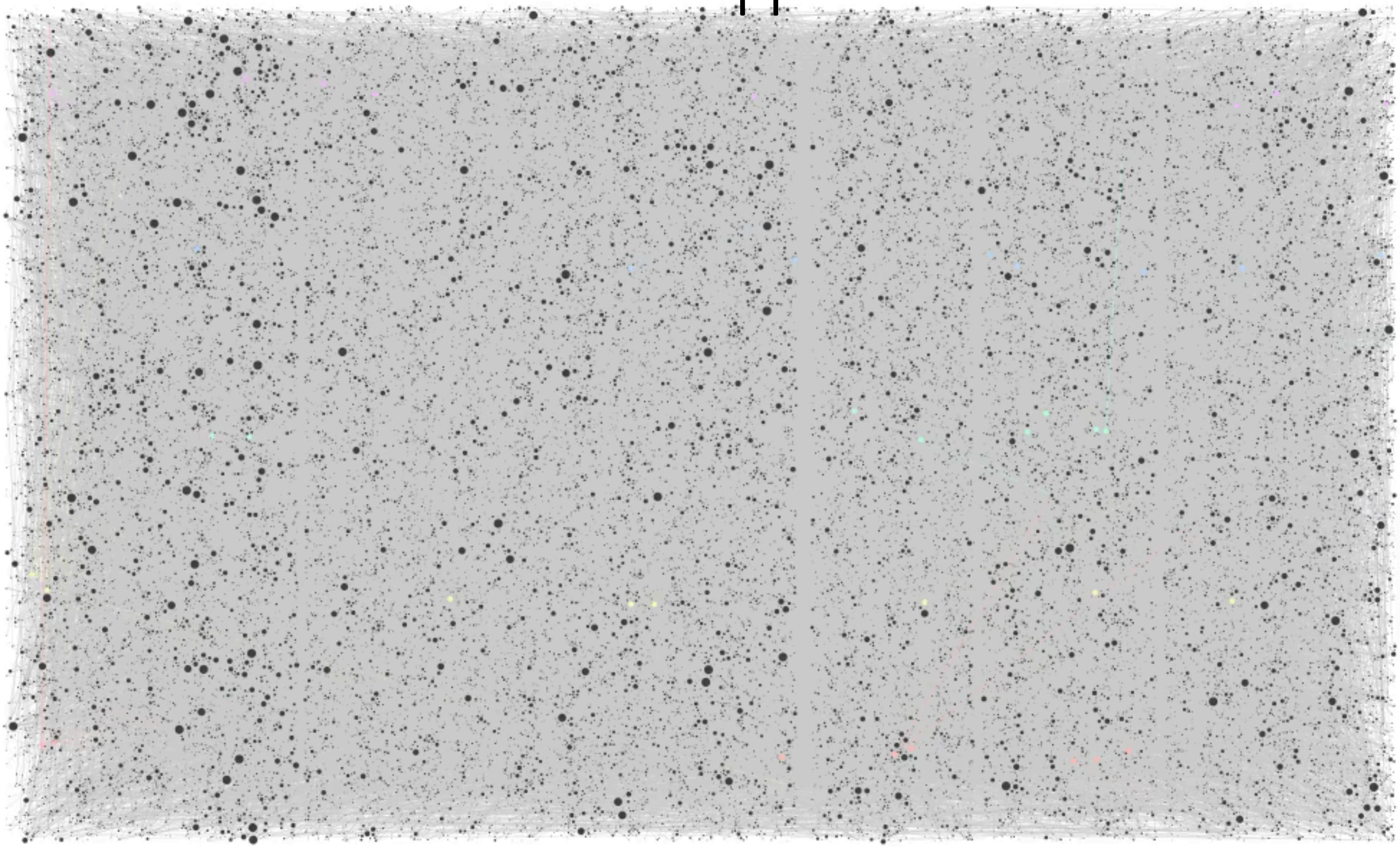
News happens

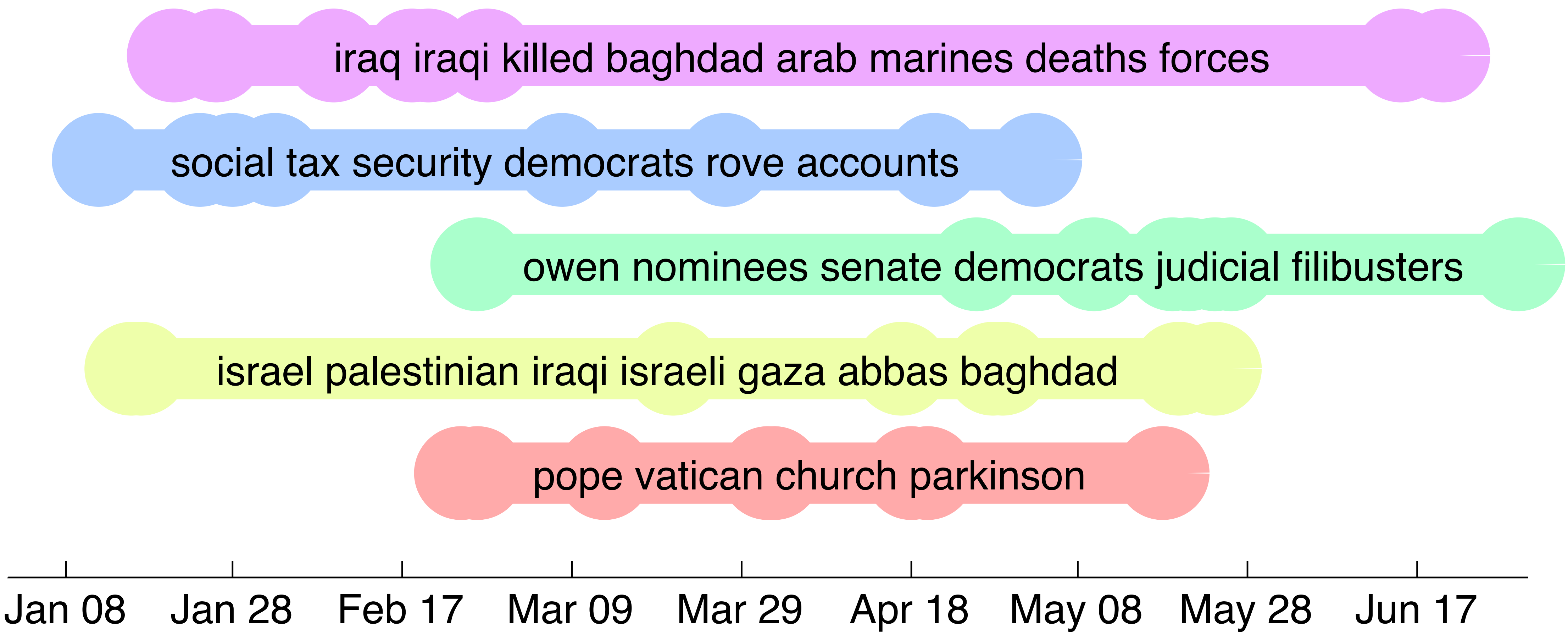


News happens

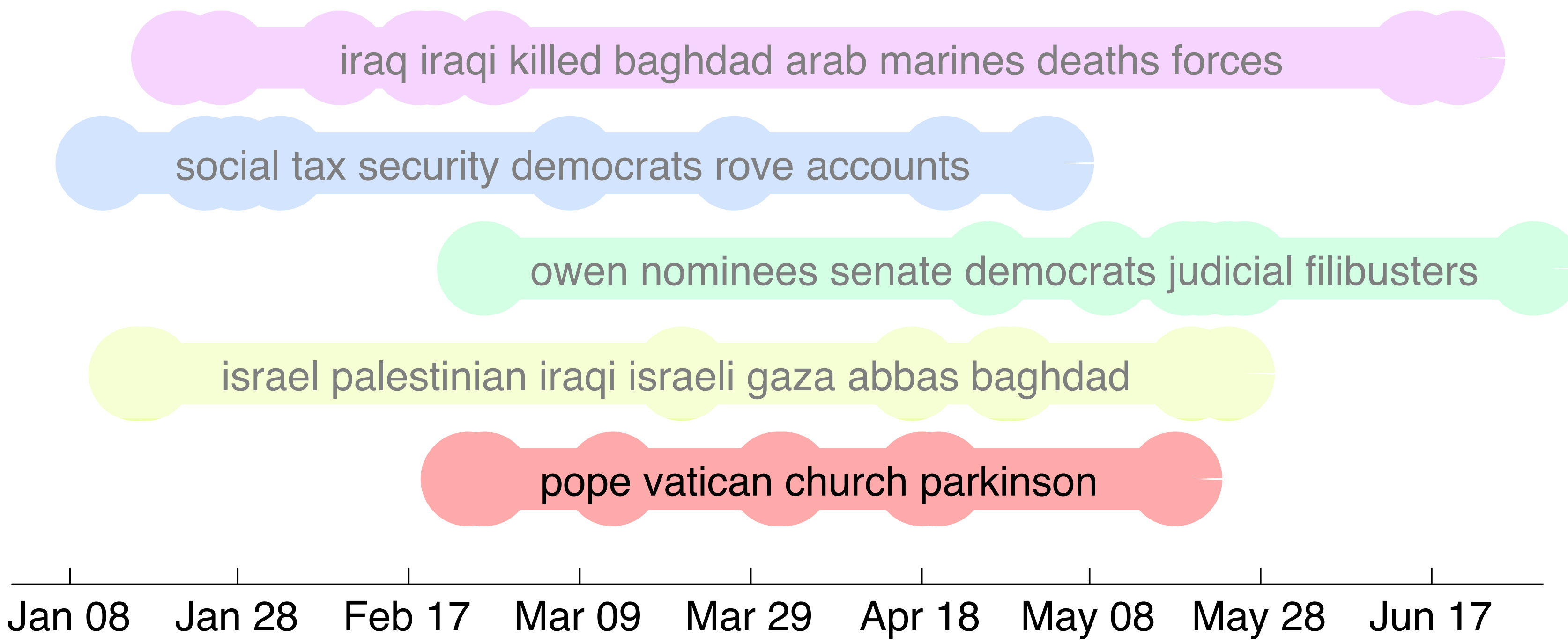


News happens



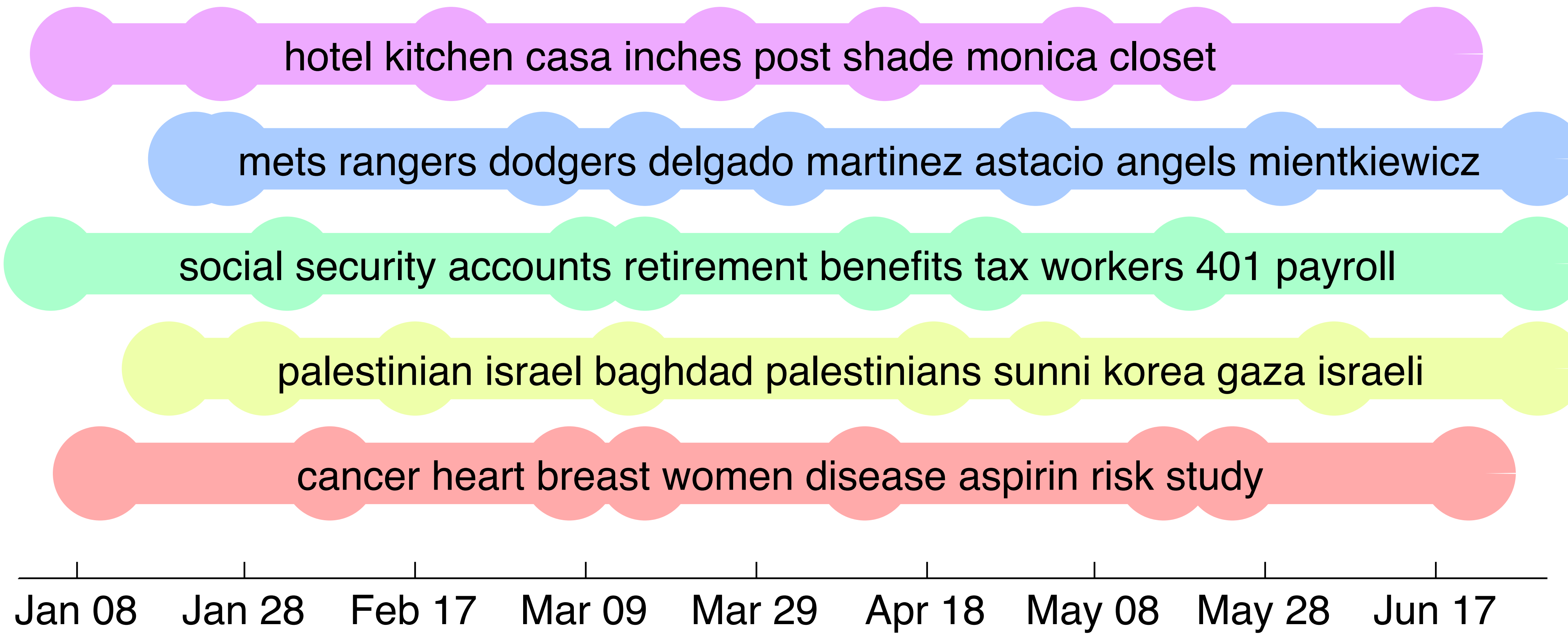


iraq iraqi killed baghdad arab marines deaths forces
social tax security democrats rove accounts
owen nominees senate democrats judicial filibusters
israel palestinian iraqi israeli gaza abbas baghdad
pope vatican church parkinson
Jan 08
Jan 28
Feb 17
Mar 09
Mar 29
Apr 18
May 08
May 28
Jun 17

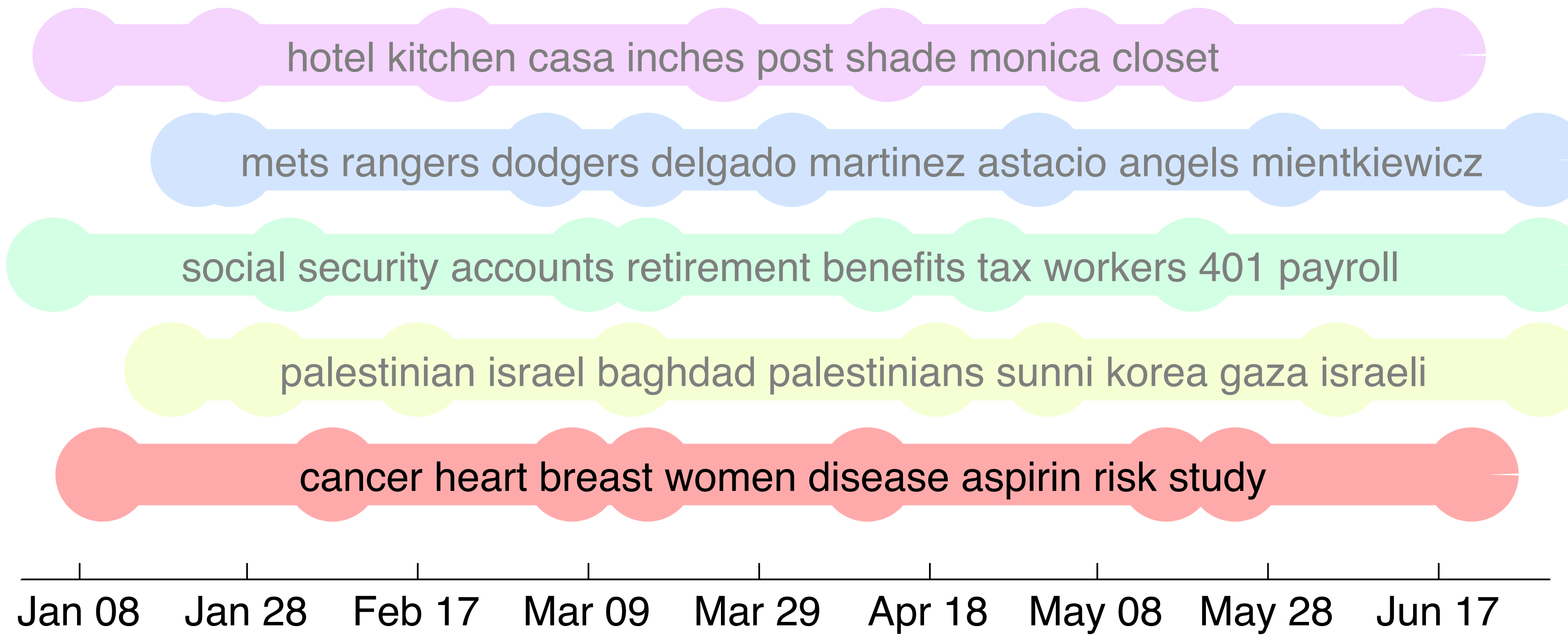


Feb 24: Parkinson's Disease Increases Risks to Pope
Feb 26: Pope's Health Raises Questions About His Ability to Lead
Mar 13: Pope Returns Home After 18 Days at Hospital
Apr 01: Pope's Condition Worsens as World Prepares for End of Papacy
Apr 02: Pope, Though Gravely Ill, Utters Thanks for Prayers
Apr 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

Dynamic topic model



[Blei & Lafferty, 2006]



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

[Blei & Lafferty, 2006]

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, and volume-preserving random projections to scale

Scale

- ~35,000 articles per six month time period

Scale

- ~35,000 articles per six month time period
- About 10^{360} possible sets of threads
- D = 36,356-dimensional diversity features

Scale

- ~35,000 articles per six month time period
- About 10^{360} possible sets of threads
- D = 36,356-dimensional diversity features
- Naively, each second-order message is 200 TB
- Using random projections to approximate volumes
  We show need only log(# articles) projections

Results: Human summaries & Turk ratings

System	*k*-means
Rouge1	16.5
Rouge2	0.69
Rouge-SU4	3.76
Coherence	2.73

Results: Human summaries & Turk ratings

System	*k*-means	DTM
Rouge1	16.5	14.7
Rouge2	0.69	0.75
Rouge-SU4	3.76	3.44
Coherence	2.73	3.19

Results: Human summaries & Turk ratings

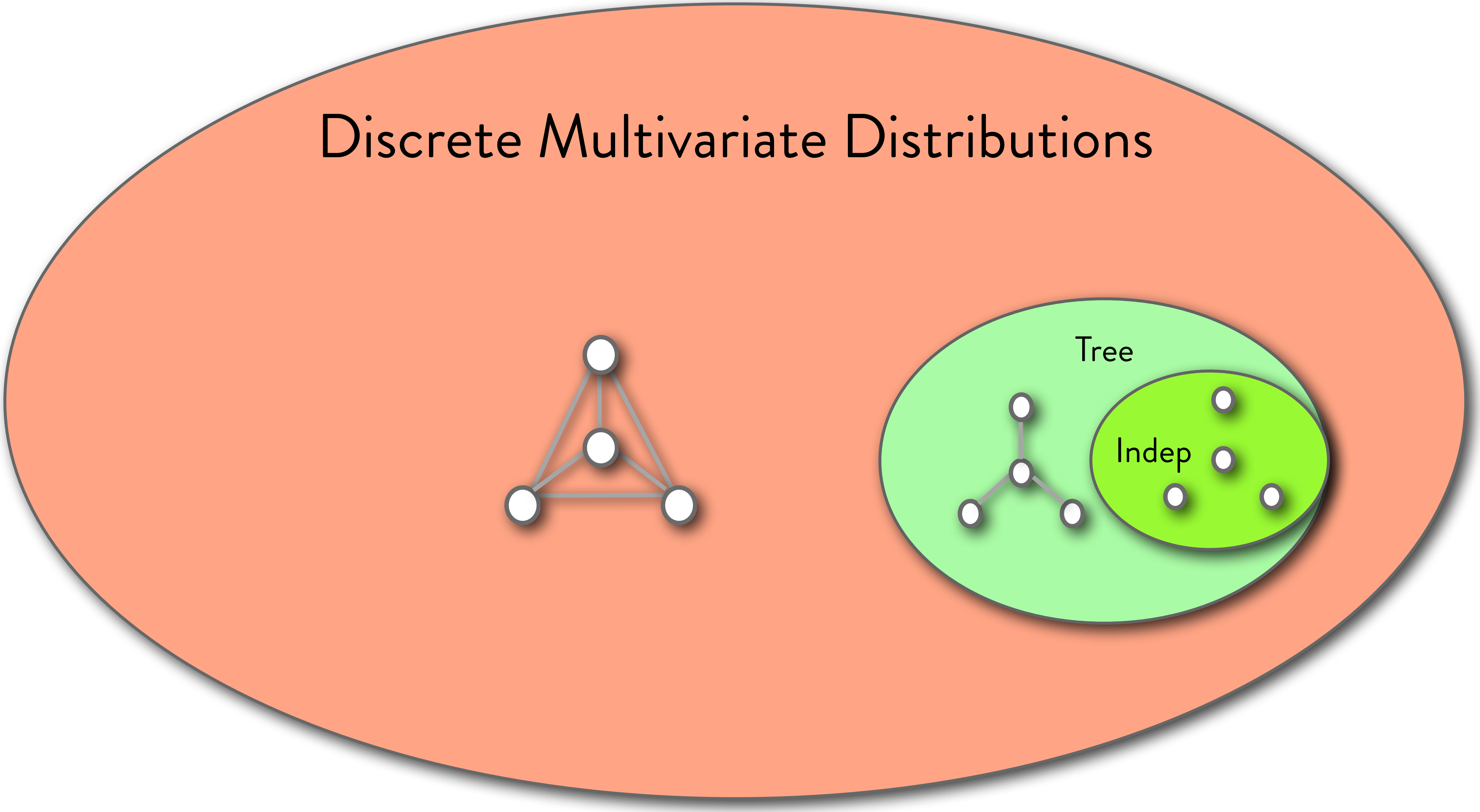
System	*k*-means	DTM	*k*-SDPP
Rouge1	16.5	14.7	**17.2**
Rouge2	0.69	0.75	**0.89**
Rouge-SU4	3.76	3.44	**3.98**
Coherence	2.73	3.19	**3.31**

Results: Human summaries & Turk ratings

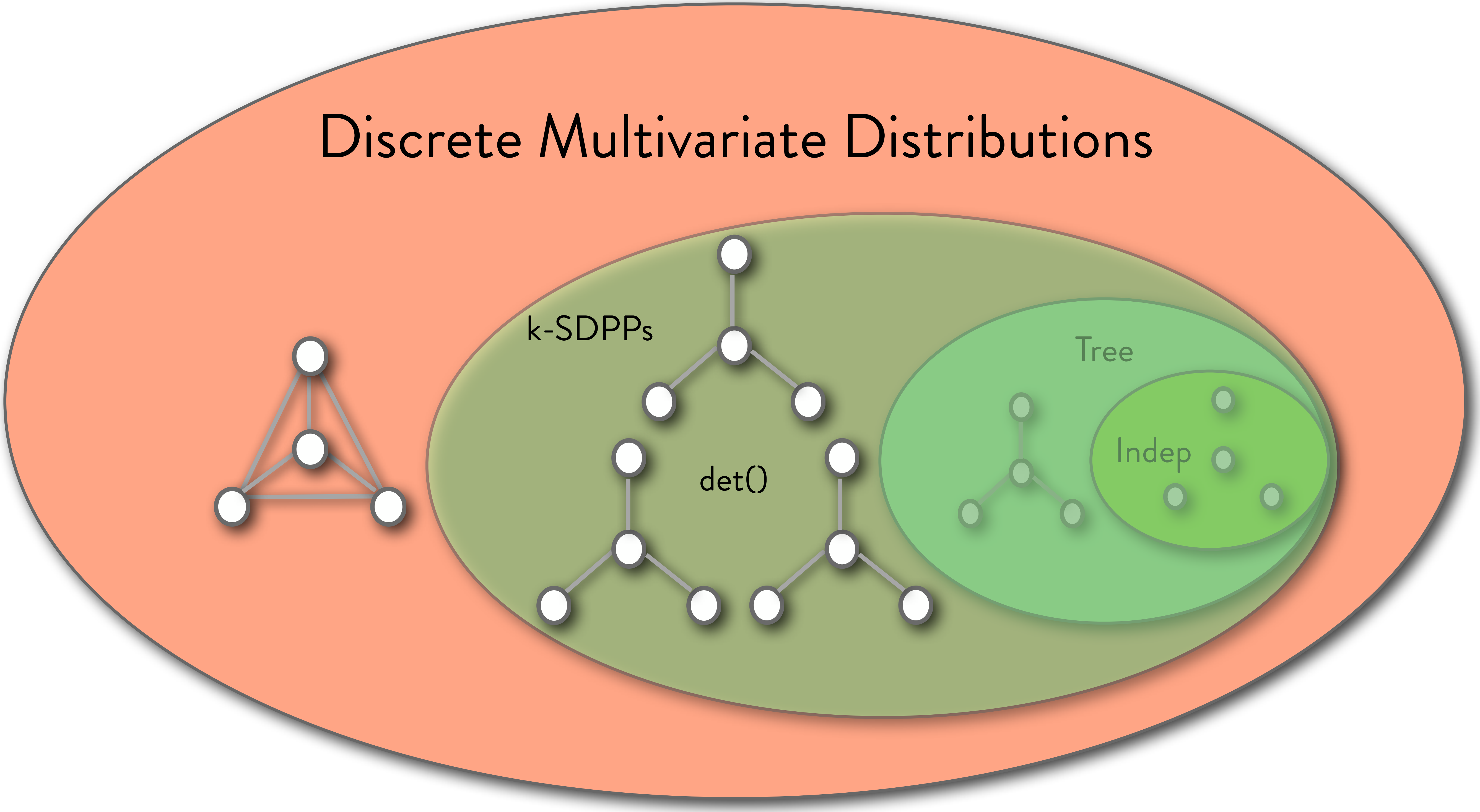
System	*k*-means	DTM	*k*-SDPP
Rouge1	16.5	14.7	**17.2**
Rouge2	0.69	0.75	**0.89**
Rouge-SU4	3.76	3.44	**3.98**
Coherence	2.73	3.19	**3.31**
Runtime (s)	626	19,434	**252**

Conclusion

k-SDPPs vs. Tyranny of Small Decisions



k-SDPPs vs. Tyranny of Small Decisions



- DPPs capture **global**, **negative** correlations
- Efficient normalization, marginals, sampling
- Our contributions:
  - **representation**
  - **learning**
  - **inference**
  - **structure**

make DPPs useful for modeling real-world data.

Papers, Tutorial, Code

- Relevant Papers: see my webpage (NIPS10, UAI11, ICML11, EMNLP12,NIPS12)
- Tutorial:

  http://arxiv.org/abs/1207.6083 (117 pages)
- Matlab Code:

  http://www.cis.upenn.edu/~kulesza/code/dpp.tgz