Determinantal Point Processes

Joint work with Jennifer Gillenwater and Alex Kulesza

Image search: “jaguar”

Relevance only:

...

Image search: "jaguar"

Relevance only:

...

Relevance + diversity:

...

Summarization

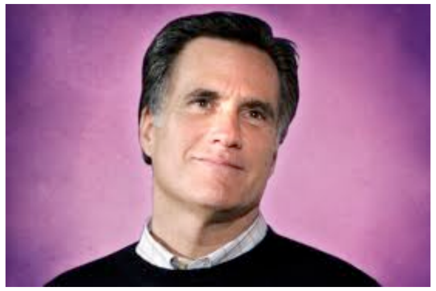
Summarization

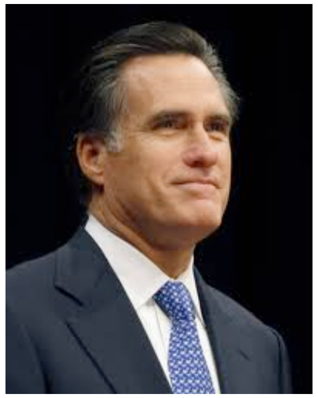
Summarization

Salience only:

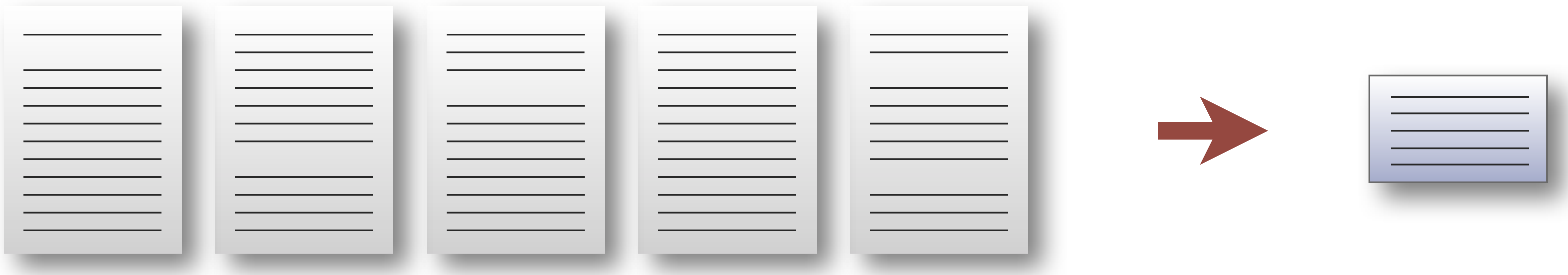
- Romney expected to claim nomination
- Romney wins three primaries
- Romney tightens grip in GOP race
- Romney is unpopular, likely nominee







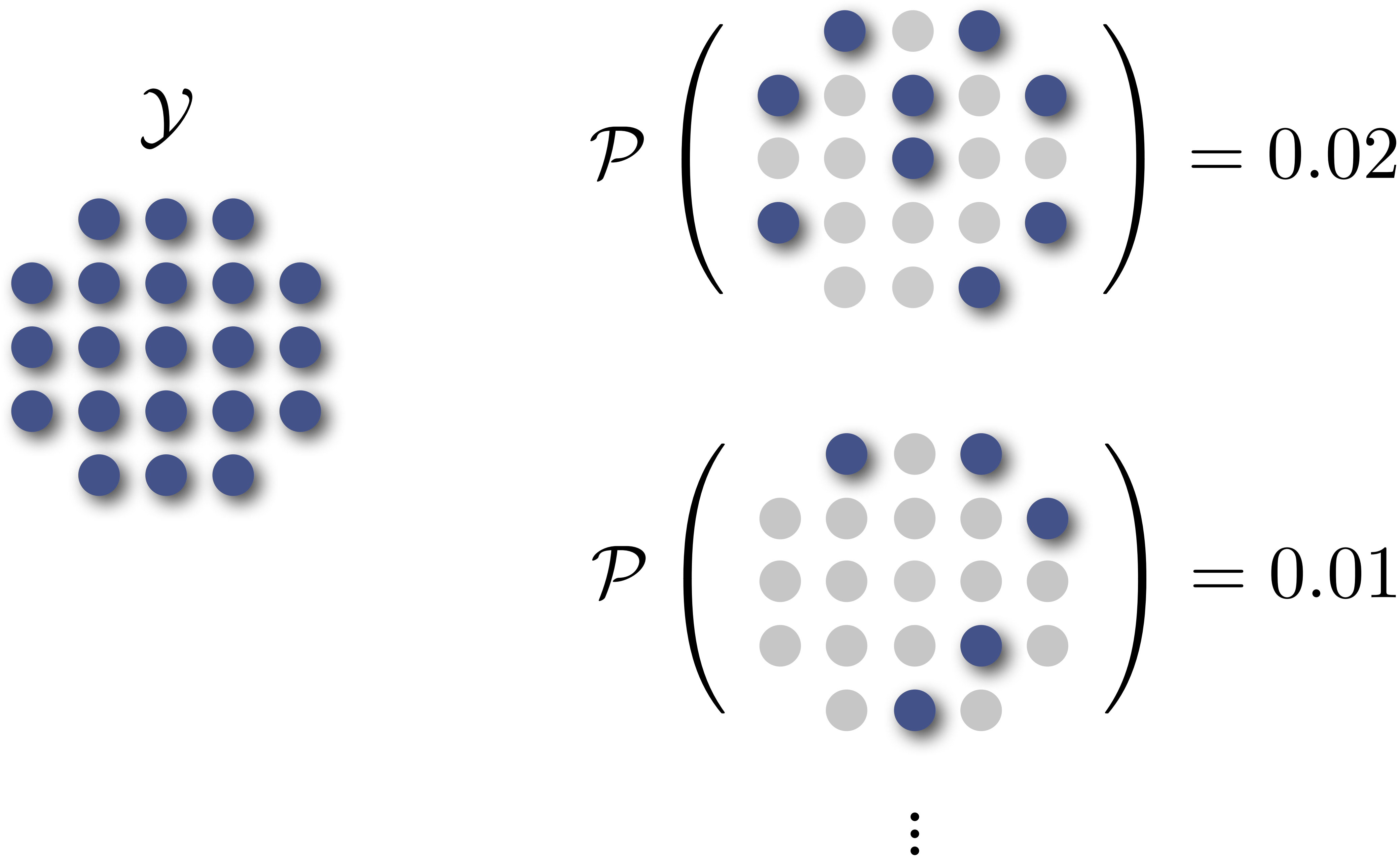

Summarization



Salience + coverage:

- Perry surging ahead of GOP pack
- Bachmann jumps into primary lead
- Herman Cain now leading in polls
- Gingrich leads Romney in national poll
- Santorum takes slight lead in GOP race
- Romney the inevitable nominee

Point processes



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:

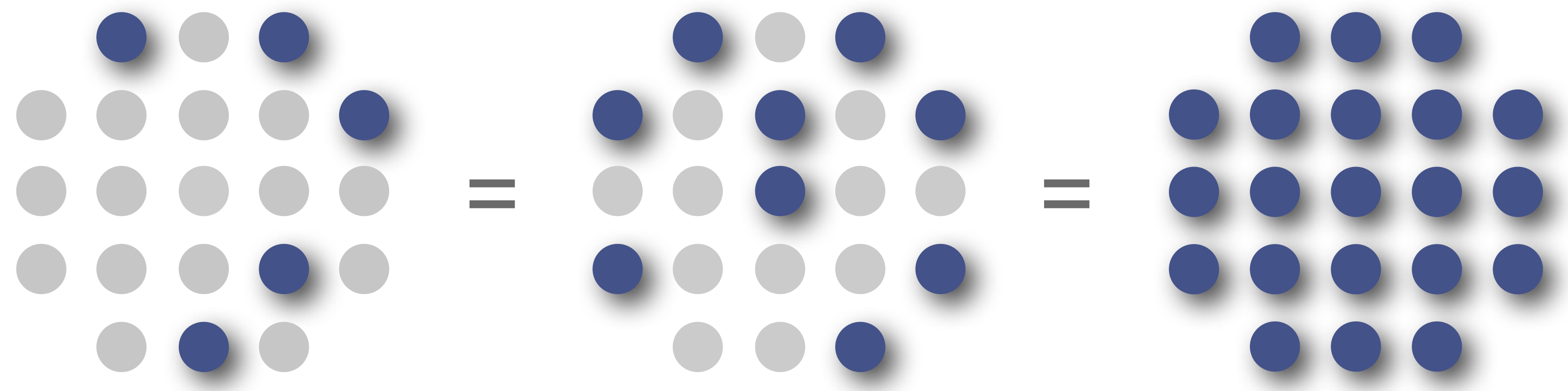
$$\mathcal{P}(Y) = \prod_{i \in Y} p_i \prod_{i \notin Y} (1 - p_i)$$

Independent point process

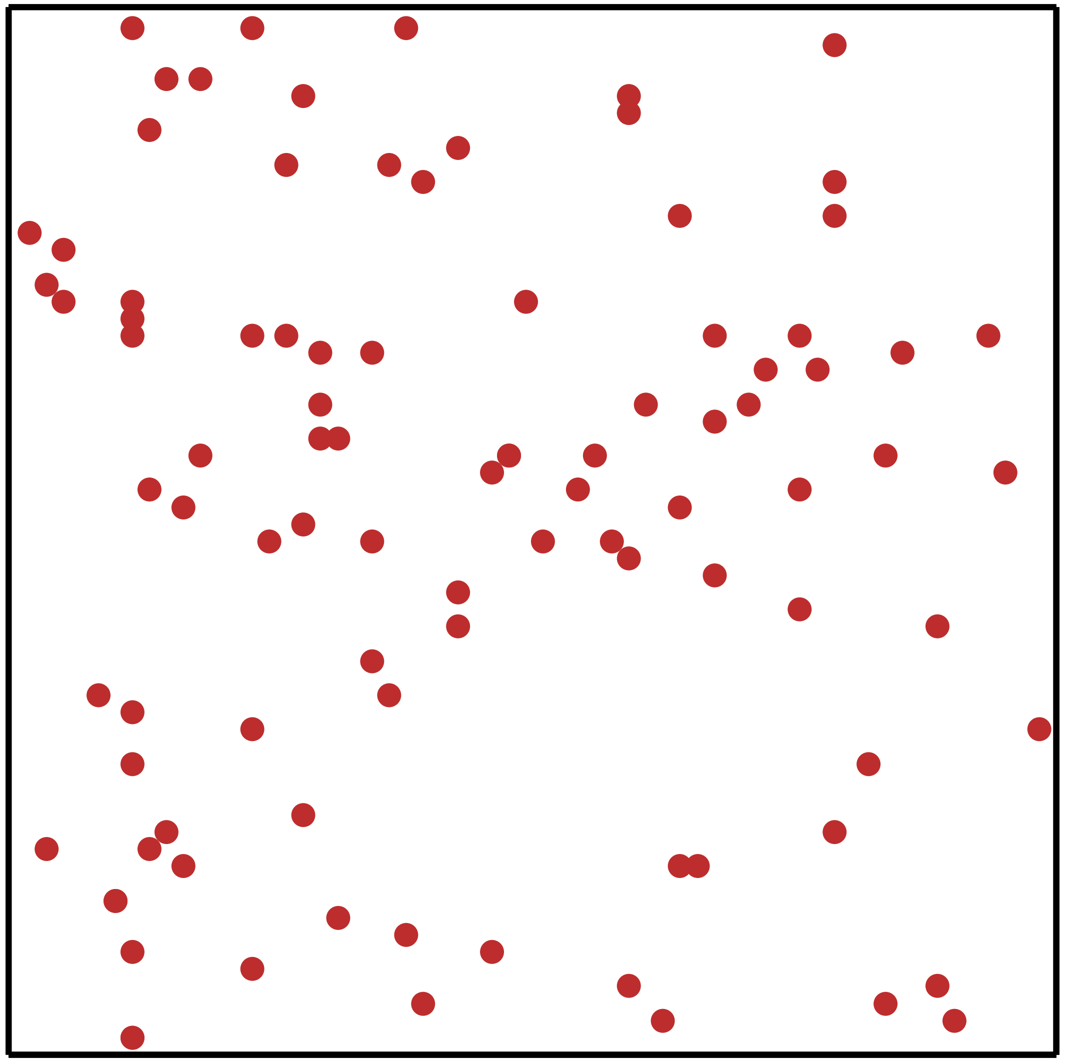
- Each element *i* included with probability p_i:

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

- For example, uniform:

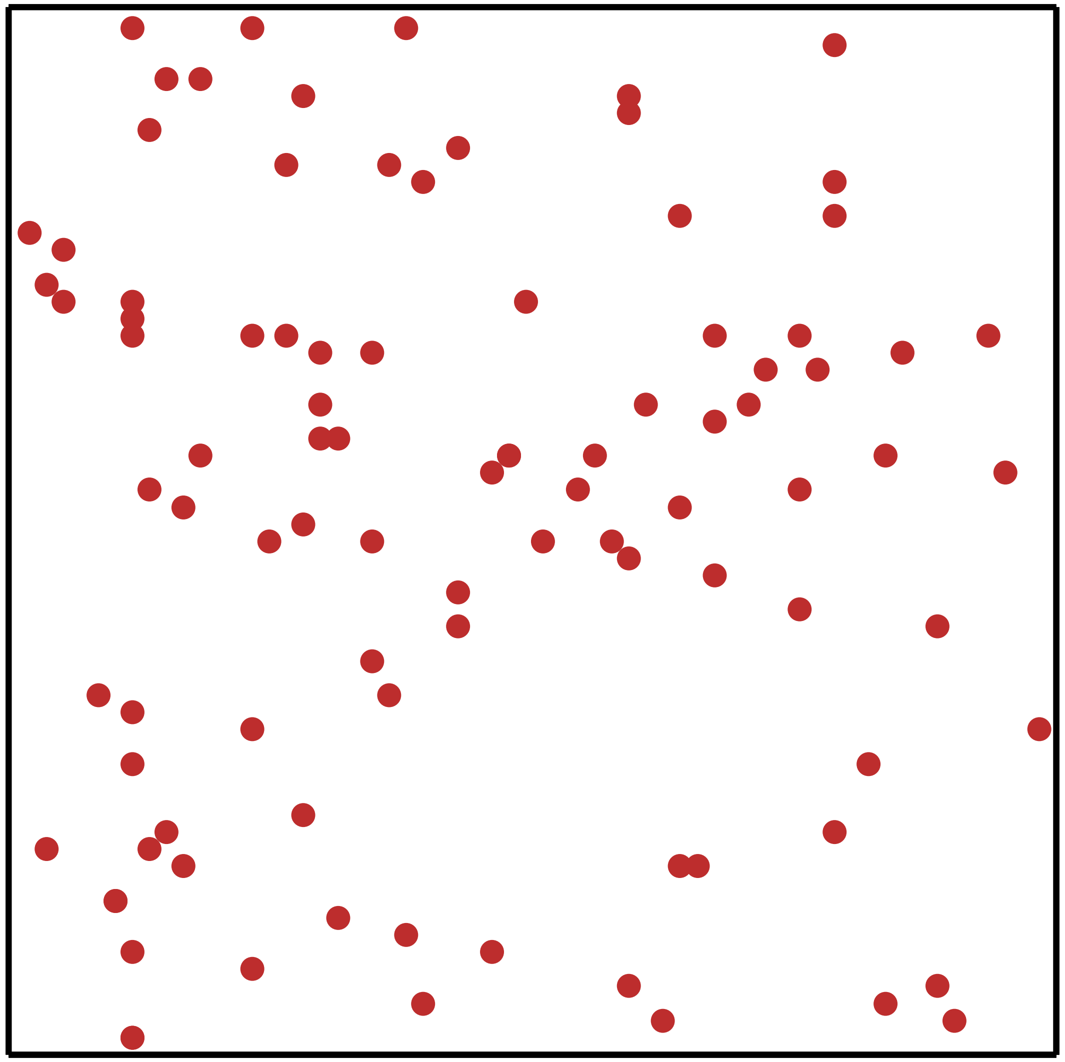


Point process samples

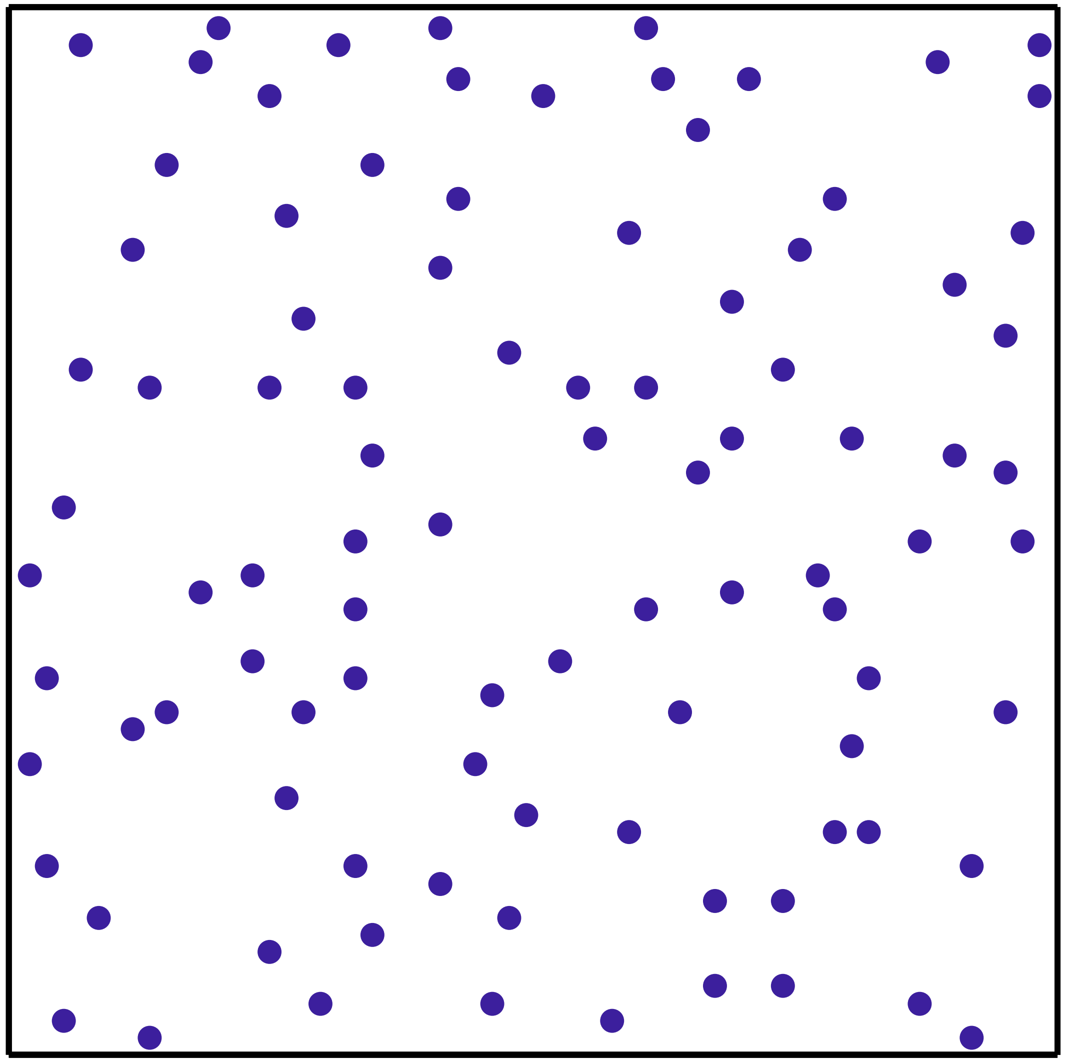


Independent

Point process samples



Independent

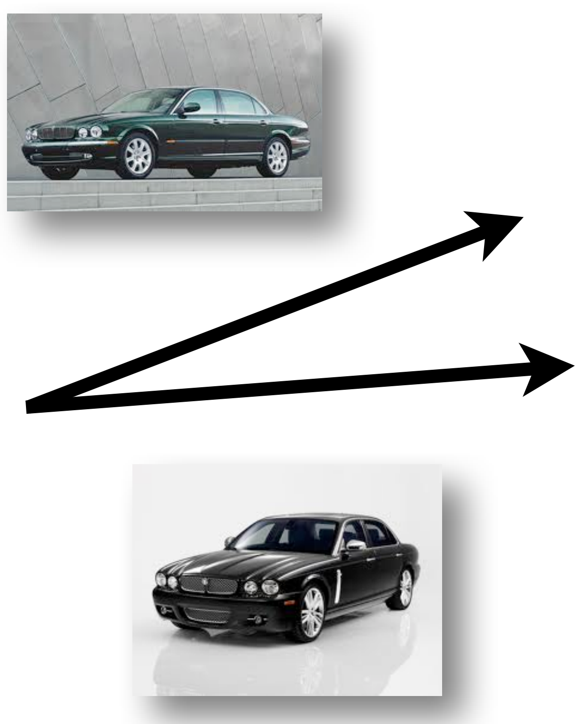


DPP

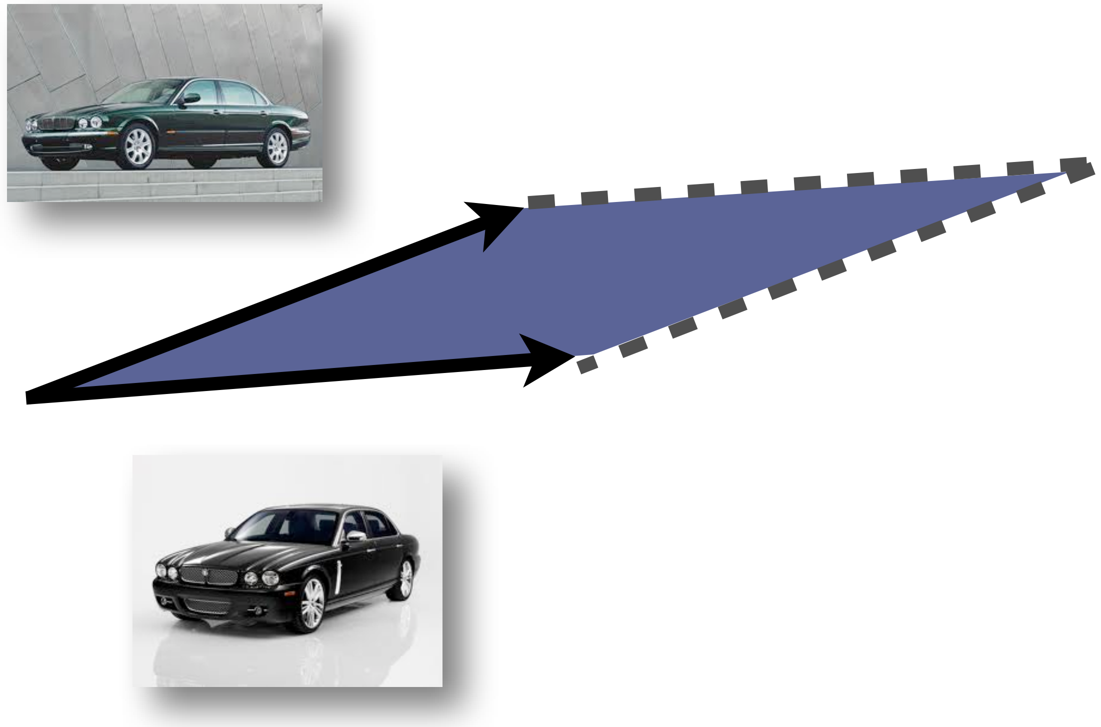
Feature function **g** on items in $\mathcal{Y}$

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

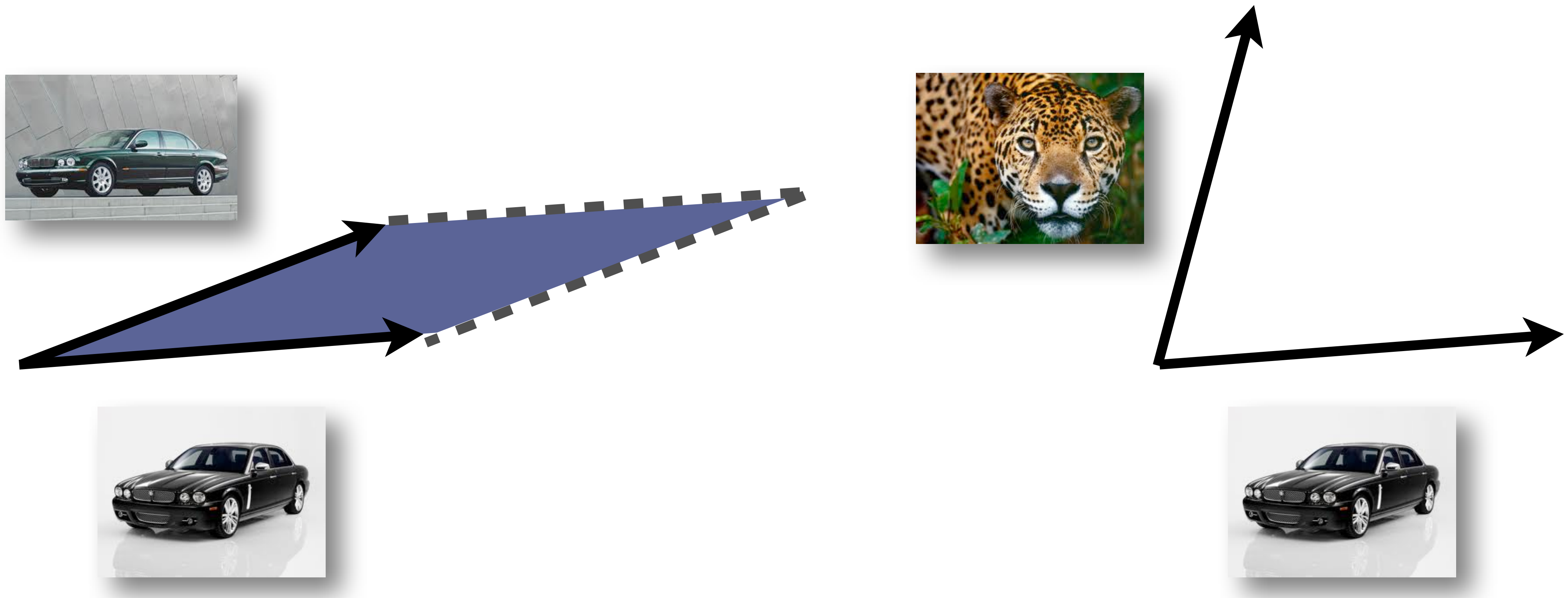
Feature function $\mathbf{g}$ on items in $\mathcal{Y}$



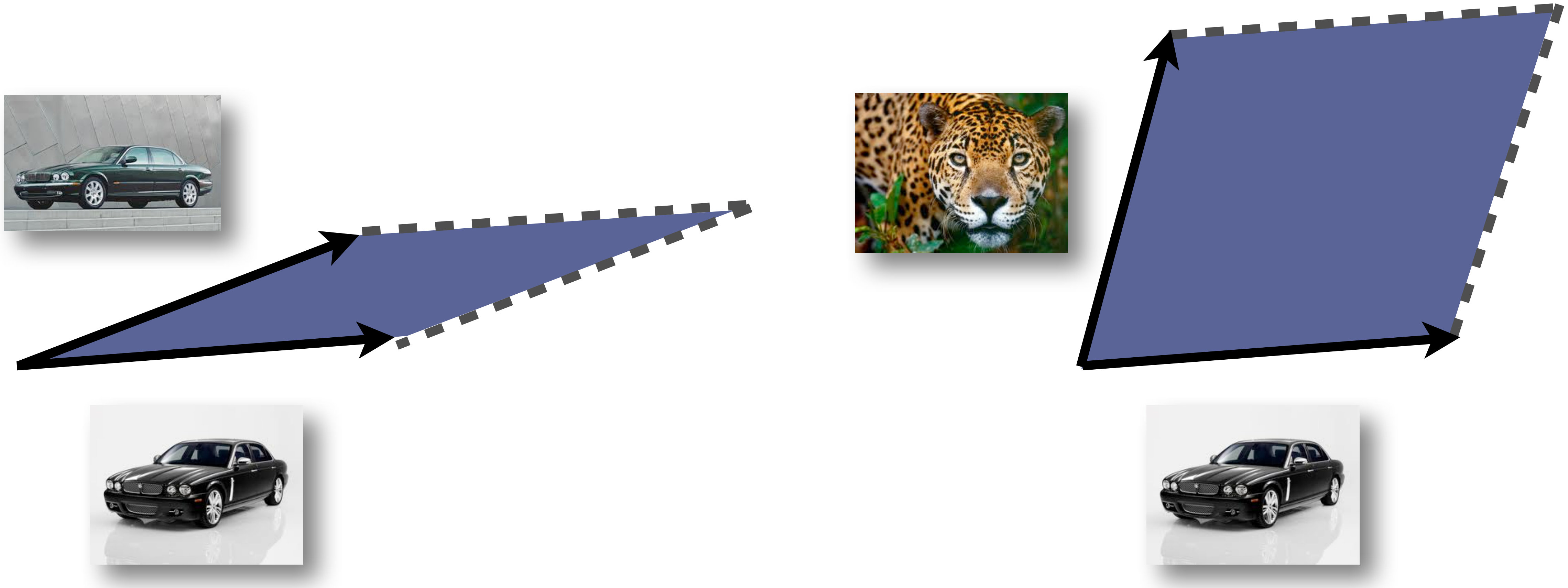
Feature function $\mathbf{g}$ on items in $\mathcal{Y}$

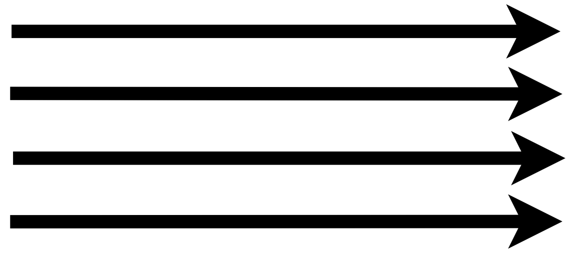


Feature function $\mathbf{g}$ on items in $\mathcal{Y}$



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

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

[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]

Determinantal point process

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

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

[Macchi, 1975]

Inference: normalization

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

Inference: normalization

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

Inference: marginals

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

Inference: marginals

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

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

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

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

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

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

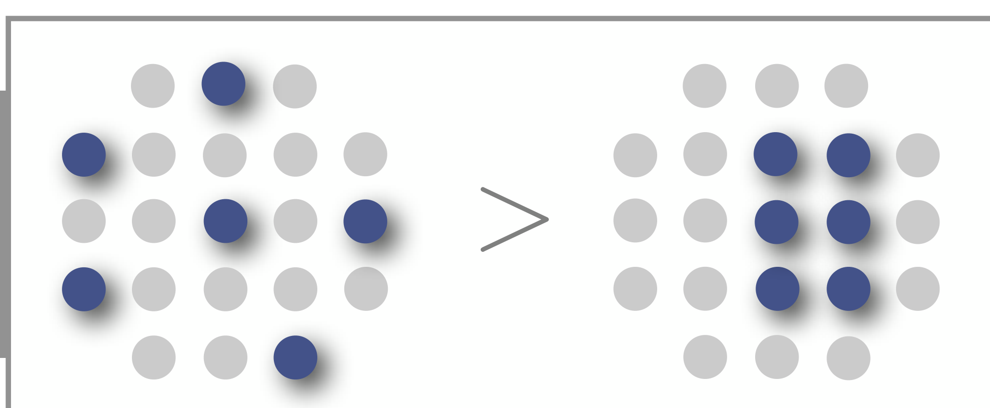
$$\mathcal{P}(i \in \boldsymbol{Y}) = \det(K_{ii}) = K_{ii}$$

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

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

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

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



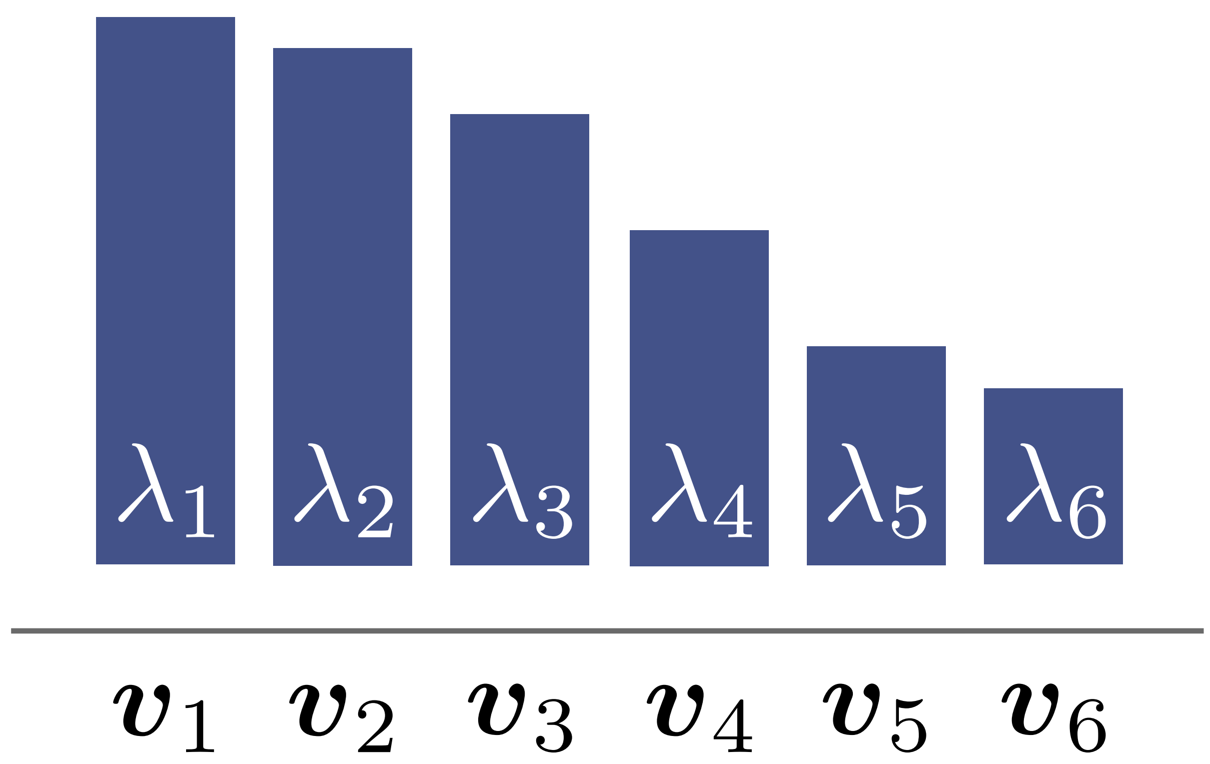
Diversity

Sampling: requires eigendecomposition

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

Sampling: requires eigendecomposition

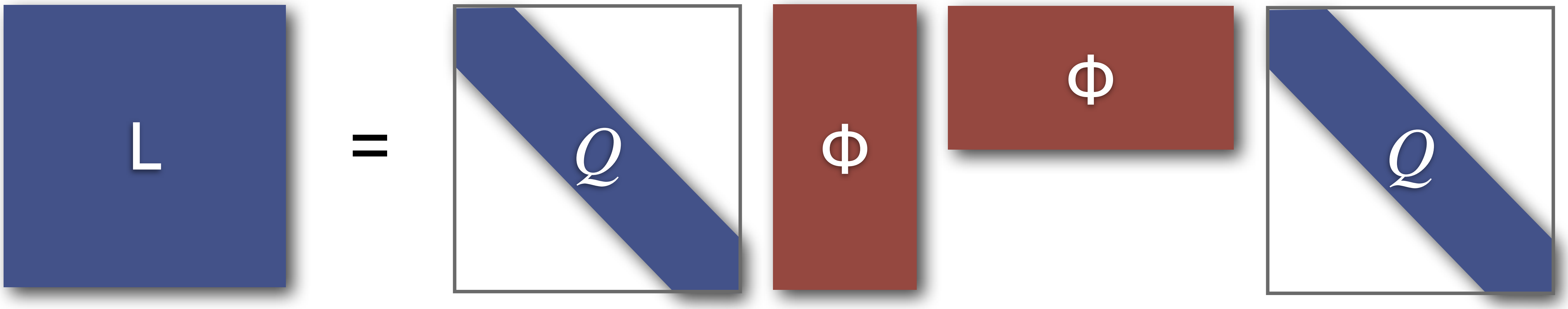
$$K = \sum_{n=1}^{N} \lambda_n \boldsymbol{v}_n \boldsymbol{v}_n^\top$$



Quality vs. diversity

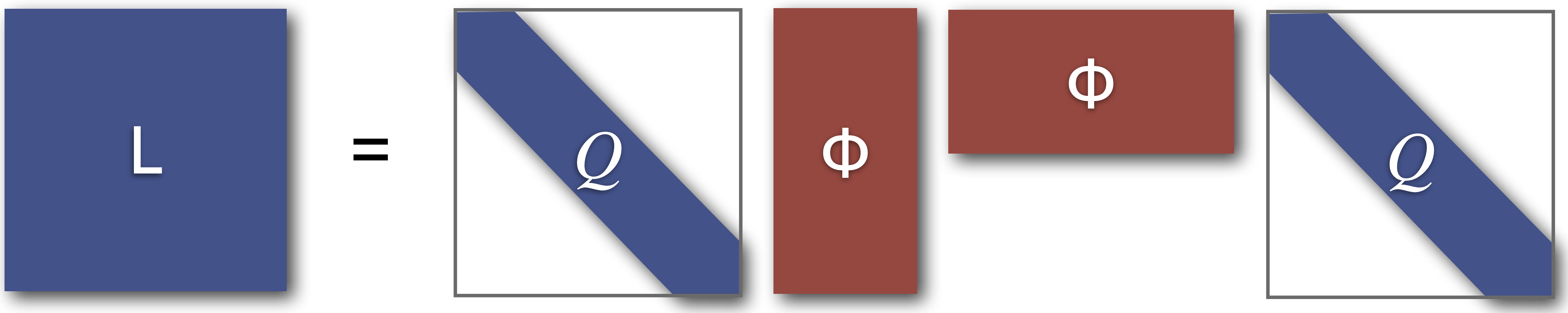
$$L_{ij} = \boldsymbol{g}(i)^\top \boldsymbol{g}(j)$$

Quality vs. diversity



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

Quality vs. diversity

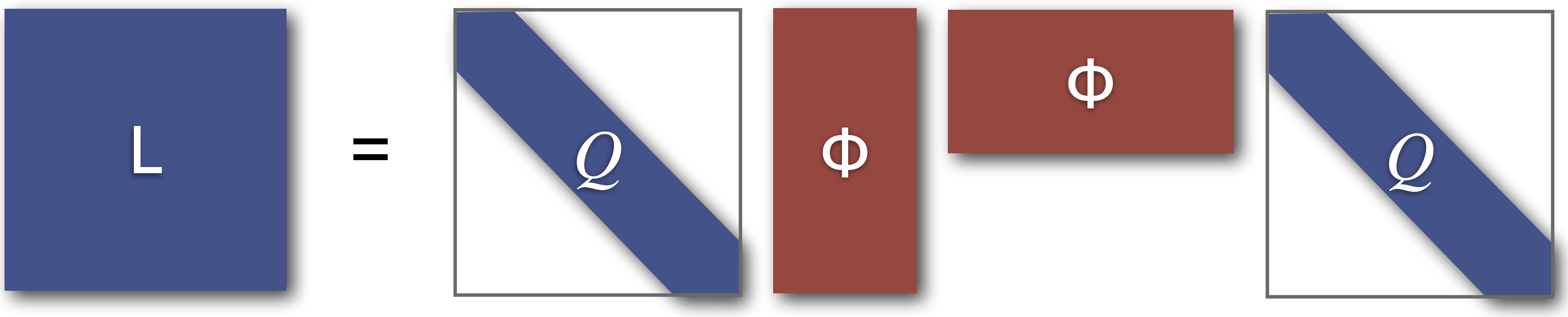


$$L_{ij} = q(i)\phi(i)^\top \phi(j)q(j)$$

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

Quality score

Quality vs. diversity



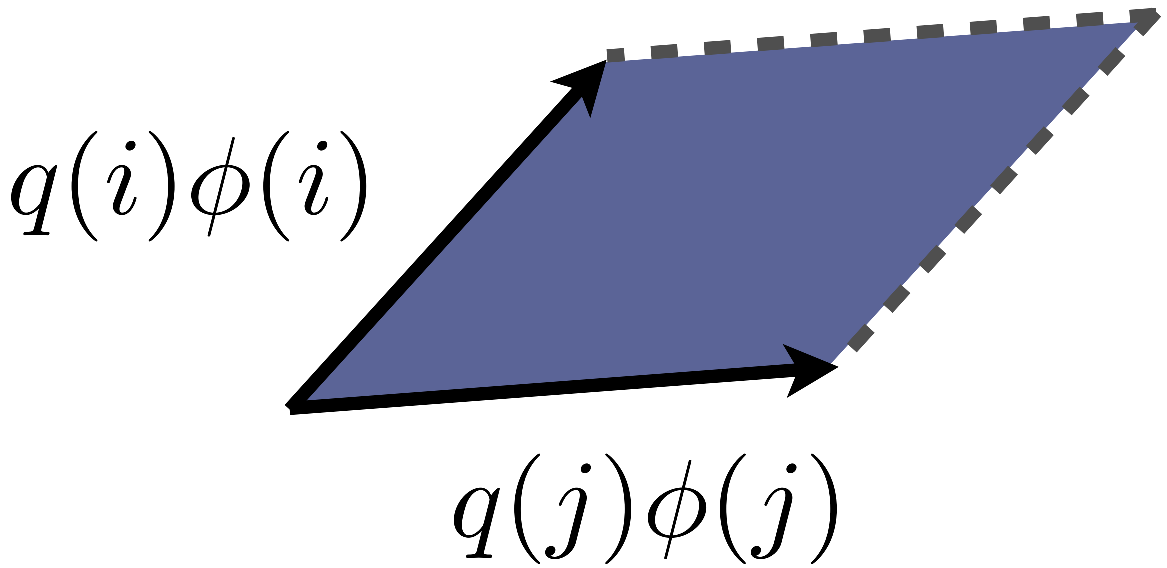
$$L_{ij} = q(i)\phi(i)^\top \phi(j)q(j)$$

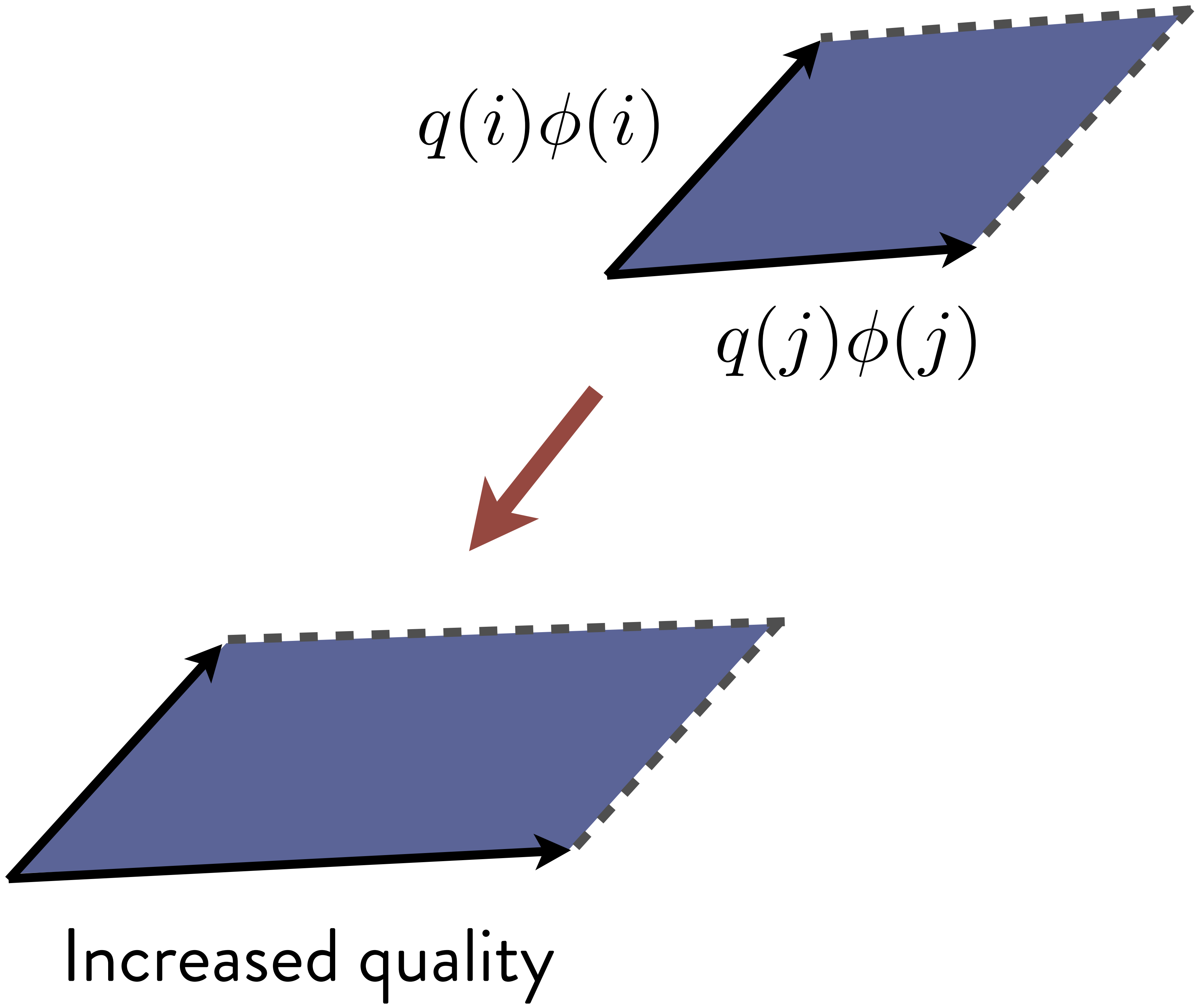
$$q(i) \in \mathbb{R}_+$$

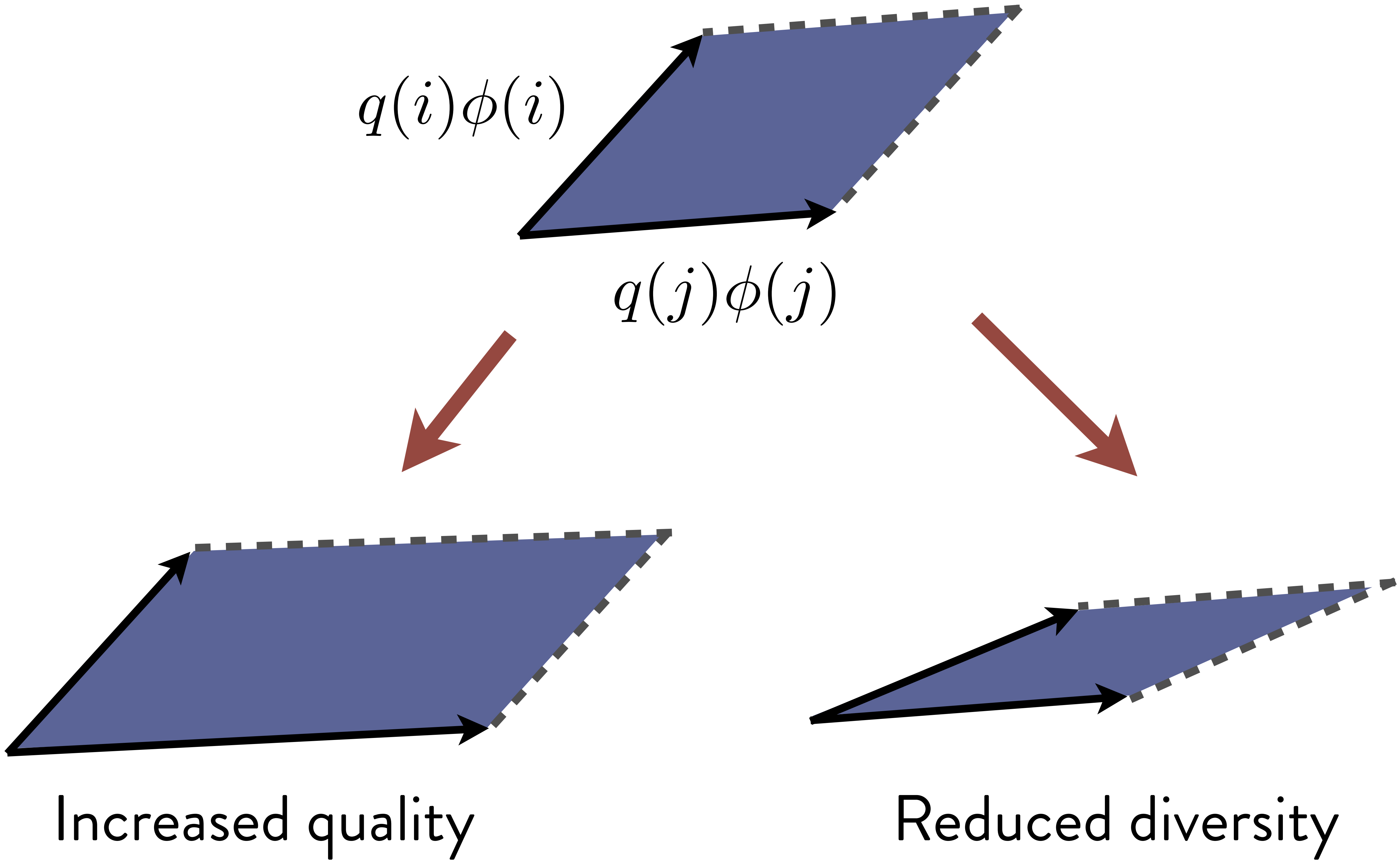
Quality score

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

Diversity features


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


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


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

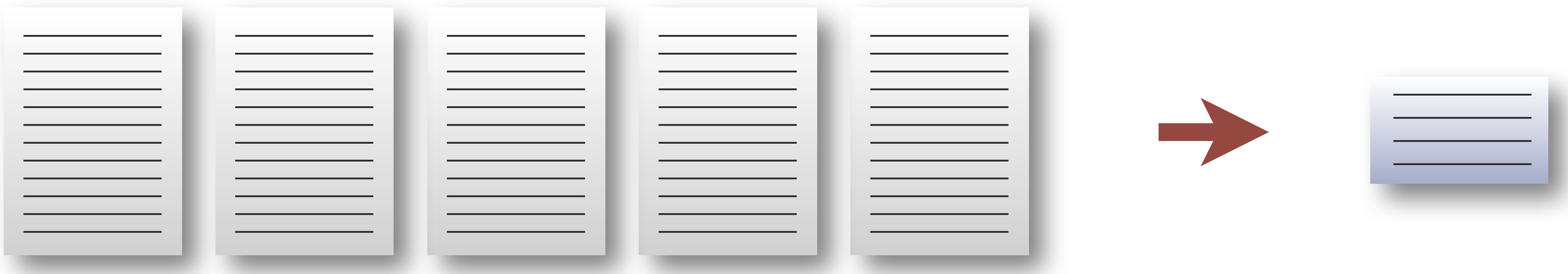
Quality vs. **diversity**

- 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**

News summarization



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

System	ROUGE-1F	ROUGE-1R	R-SU4F

System	ROUGE-1F	ROUGE-1R	R-SU4F
MMR	37.58	38.05	13.06

System	ROUGE-1F	ROUGE-1R	R-SU4F
MMR	37.58	38.05	13.06
Peer 65	37.87	38.20	13.19

System	ROUGE-1F	ROUGE-1R	R-SU4F
MMR	37.58	38.05	13.06
Peer 65	37.87	38.20	13.19
SubMod*	39.78	40.43	-

System	ROUGE-1F	ROUGE-1R	R-SU4F
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

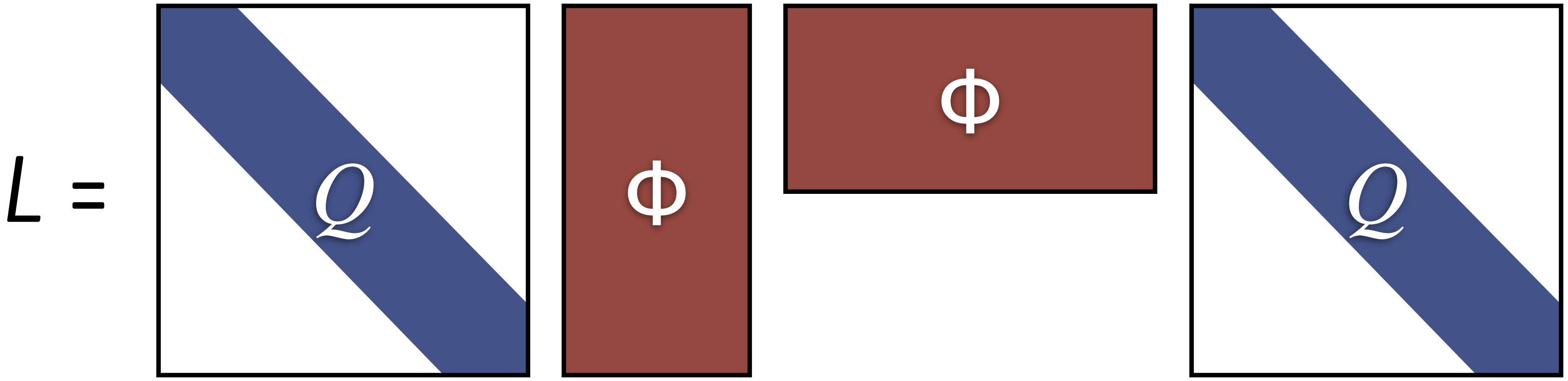
[*Lin and Bilmes, 2012]

System	ROUGE-1F	ROUGE-1R	R-SU4F
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**

[*Lin and Bilmes, 2012]

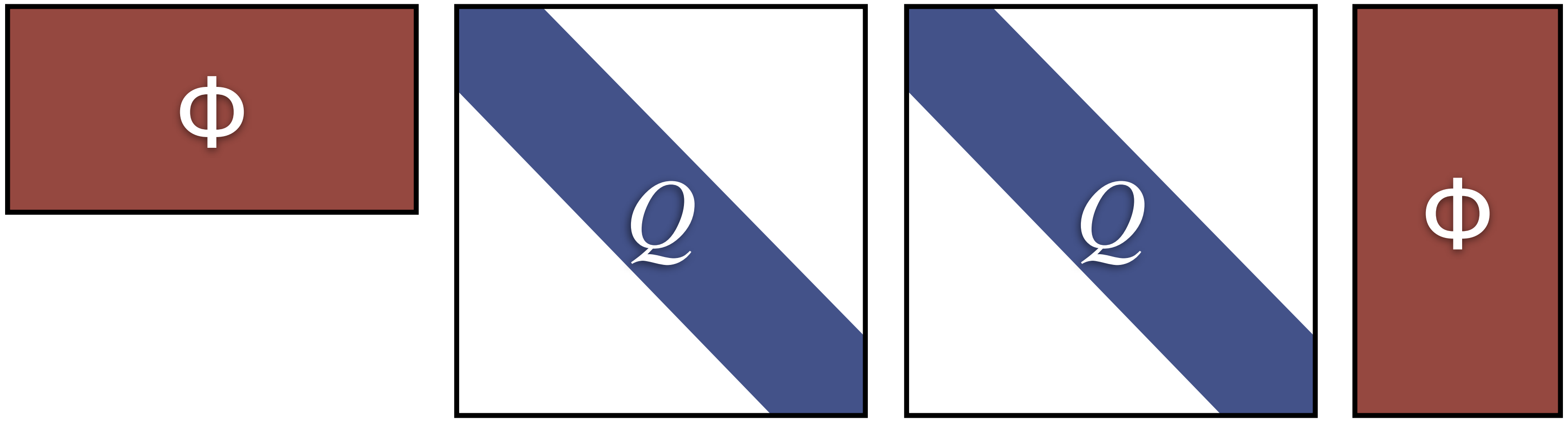
Large N?

Dual representation

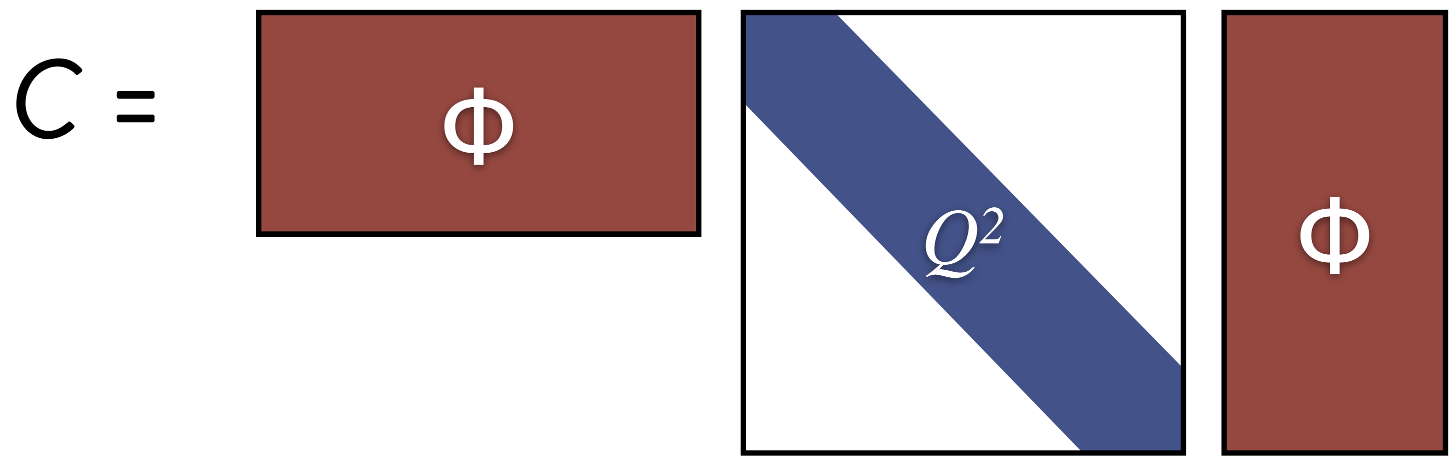


$$L_{ij} = q(i)\phi(i)^\top \phi(j) q(j)$$

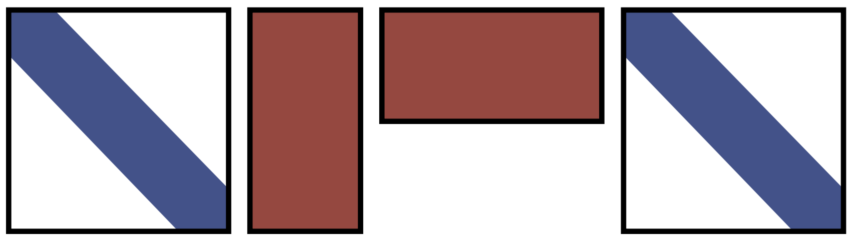
Dual representation



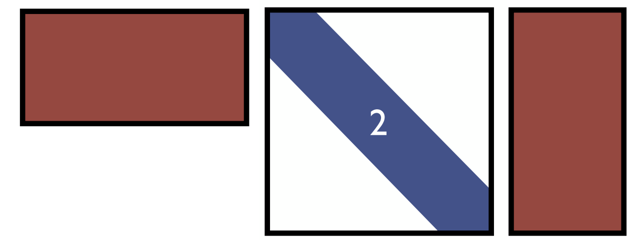
Dual representation



Dual representation

$L =$ 

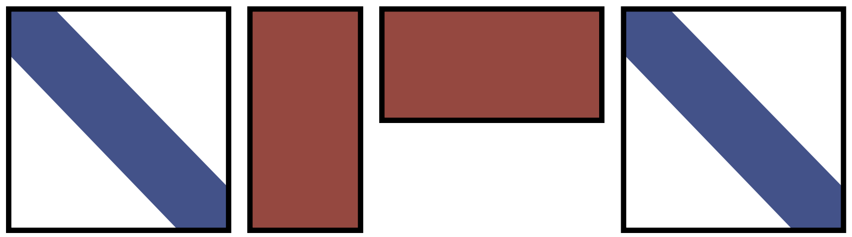
N x N

$C =$ 

D x D

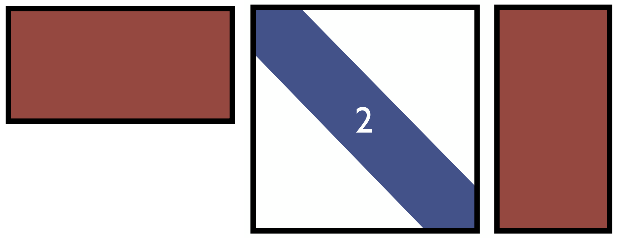
Dual representation

$L =$



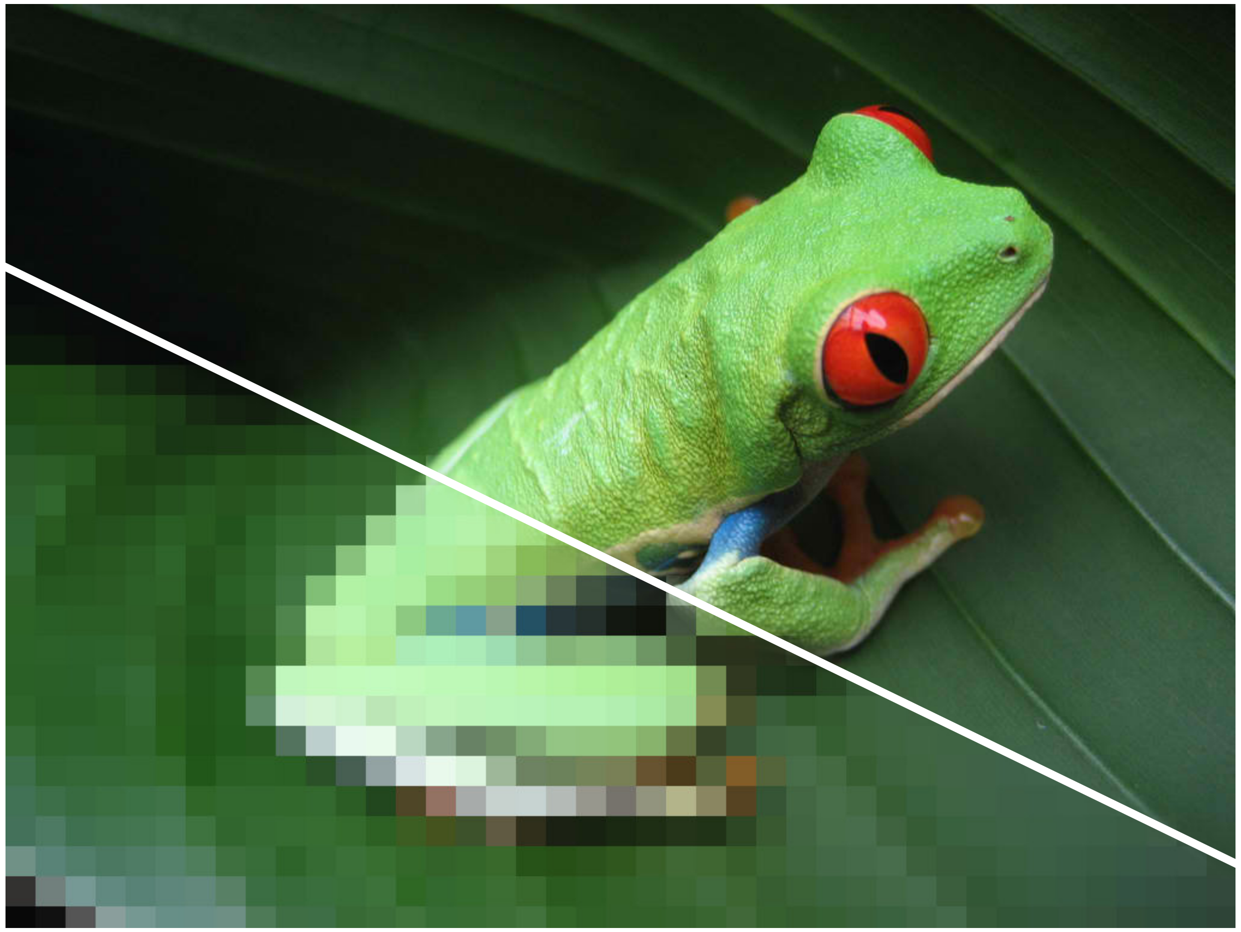
$N \times N$

$C =$

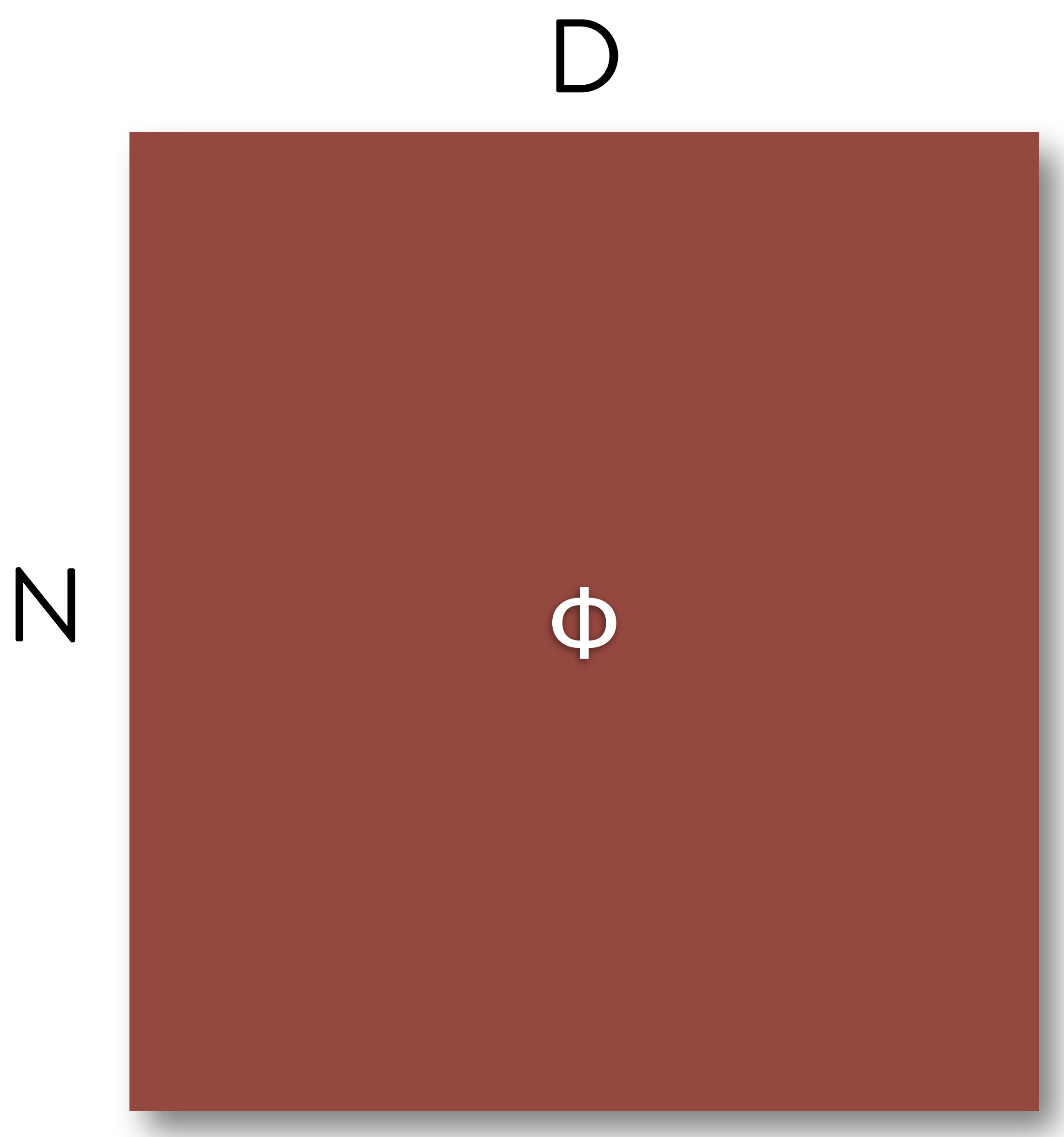


$D \times D$

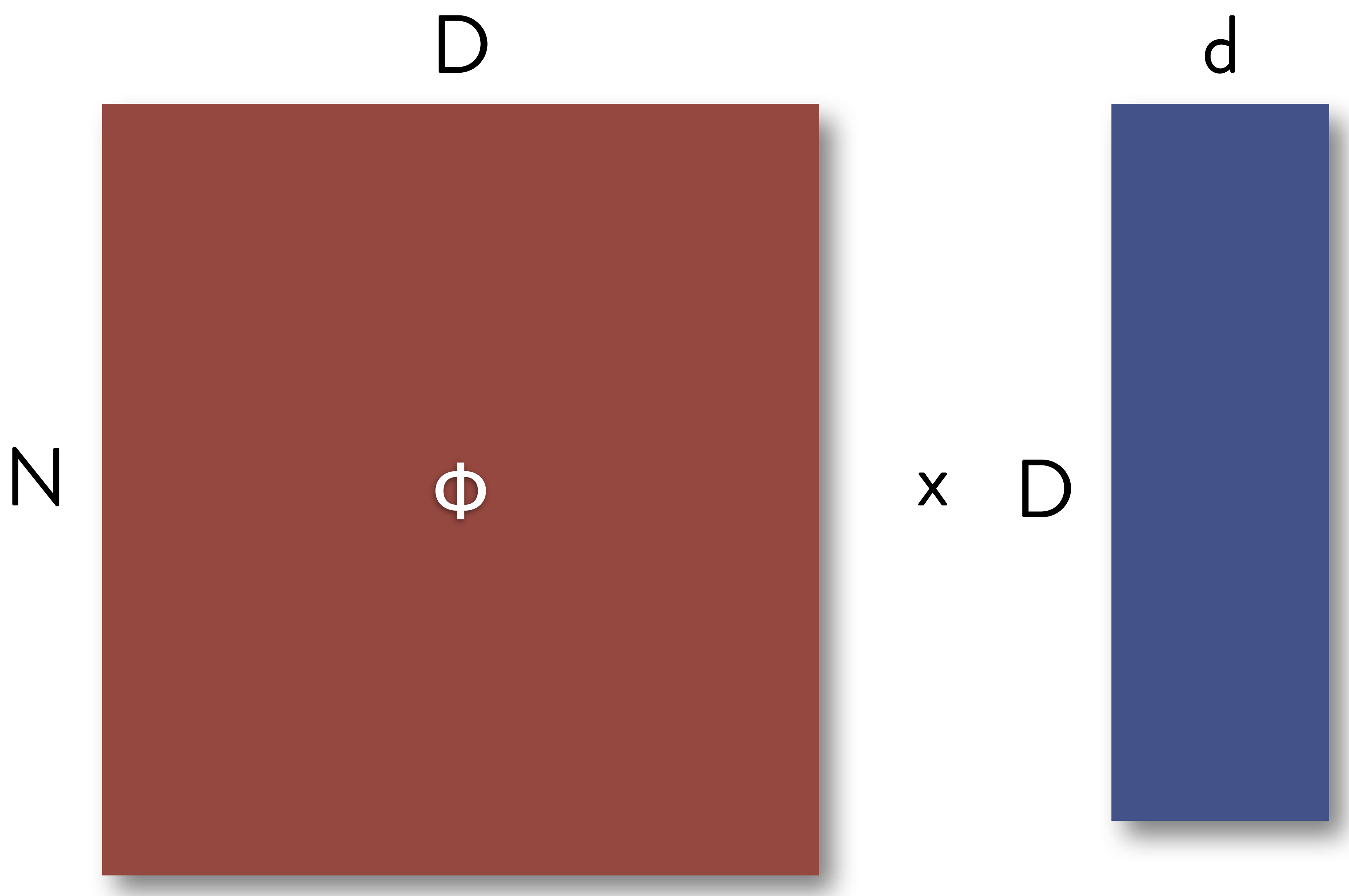
- C and L have same (non-zero) eigenvalues
- Eigenvectors are related
- Use C for sampling and other inference



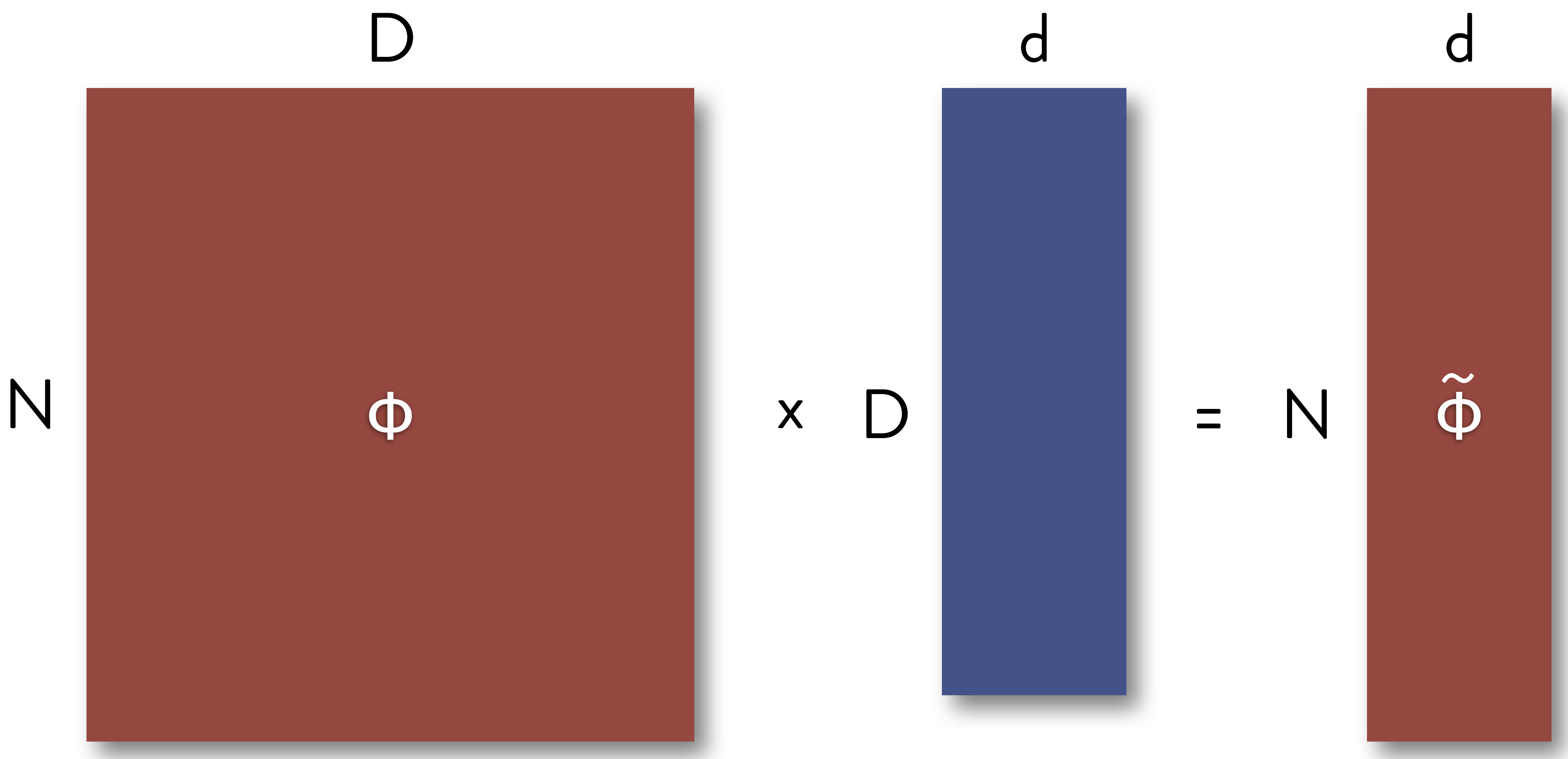

Projection



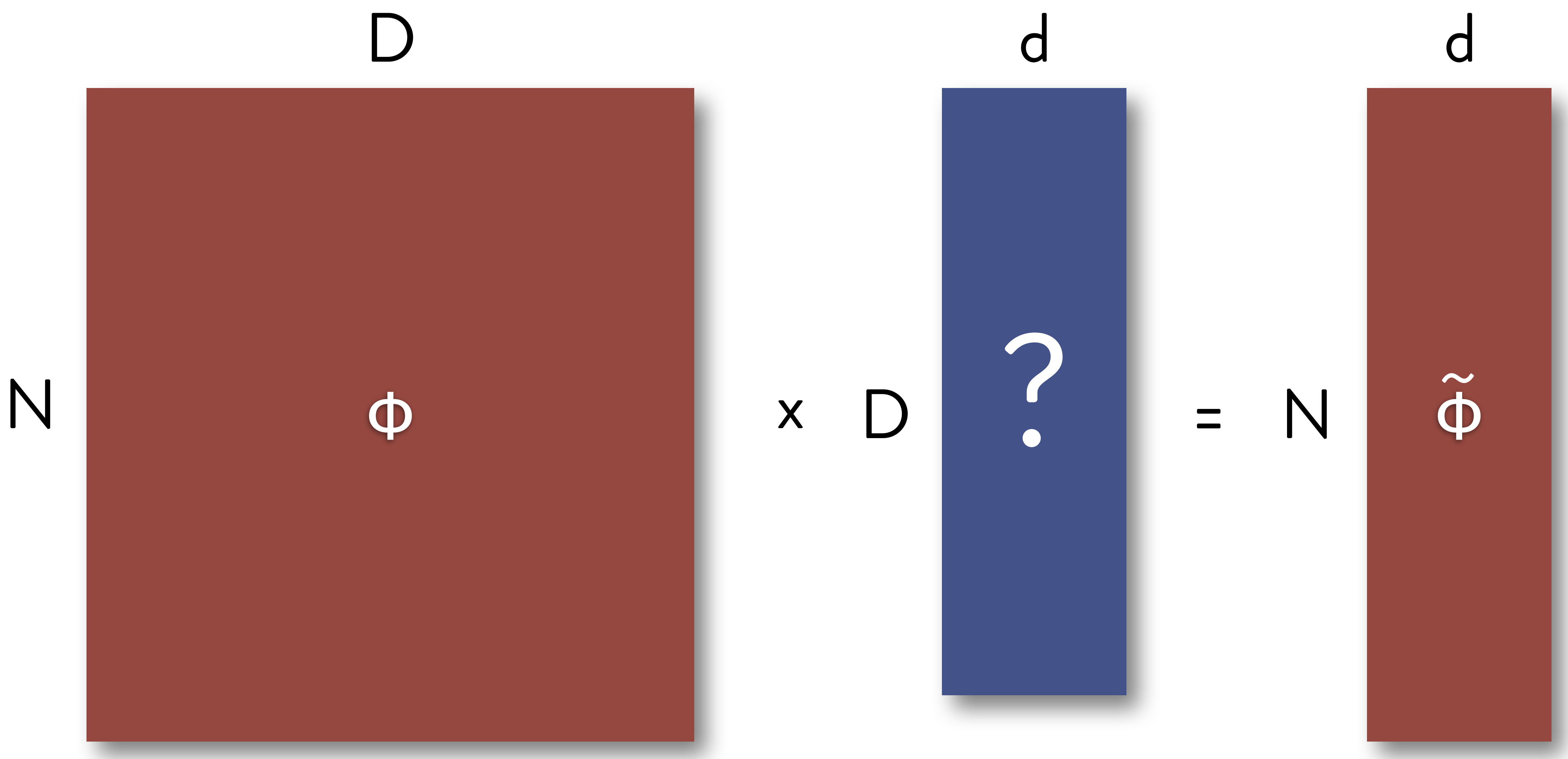
Projection



Projection

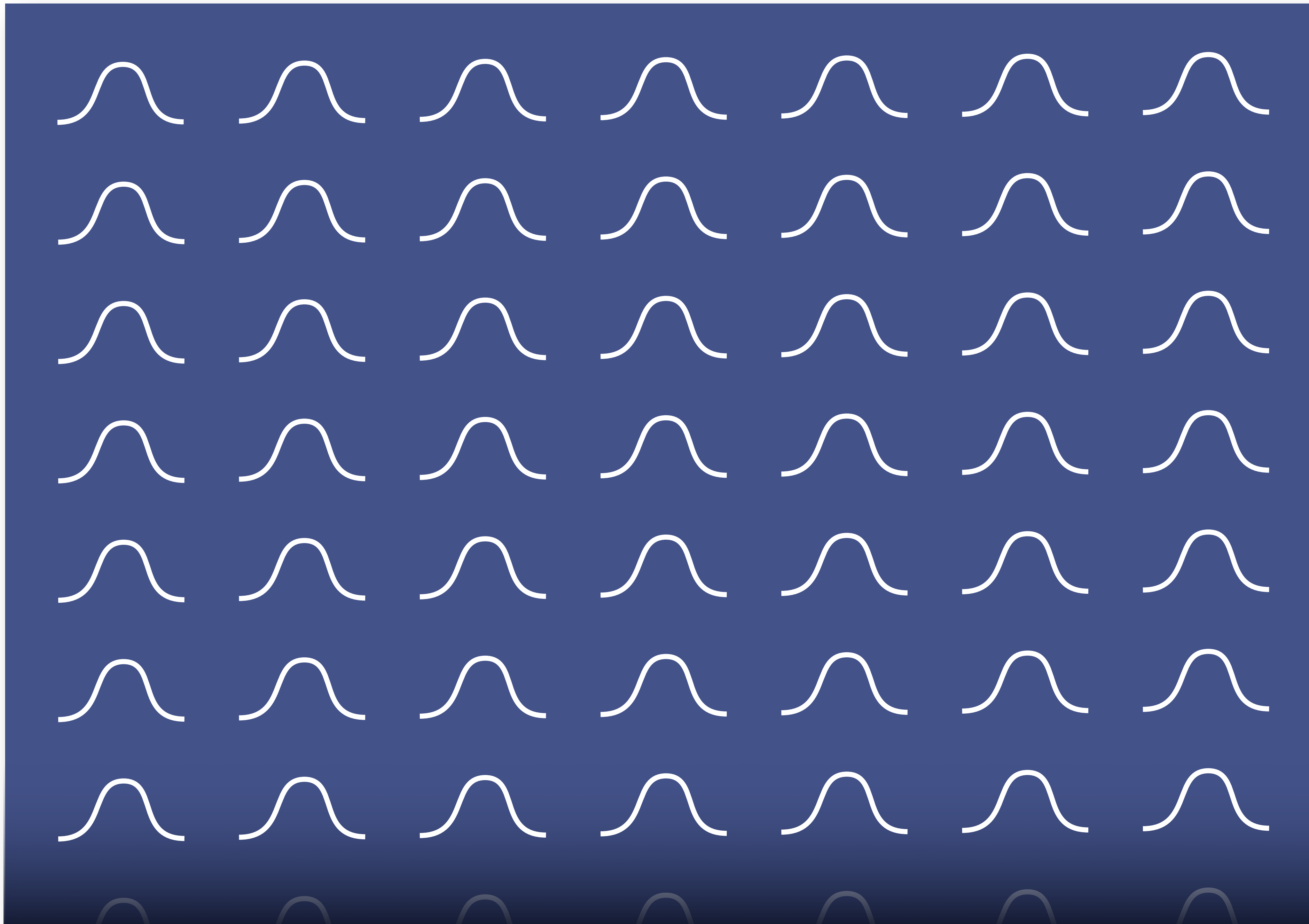


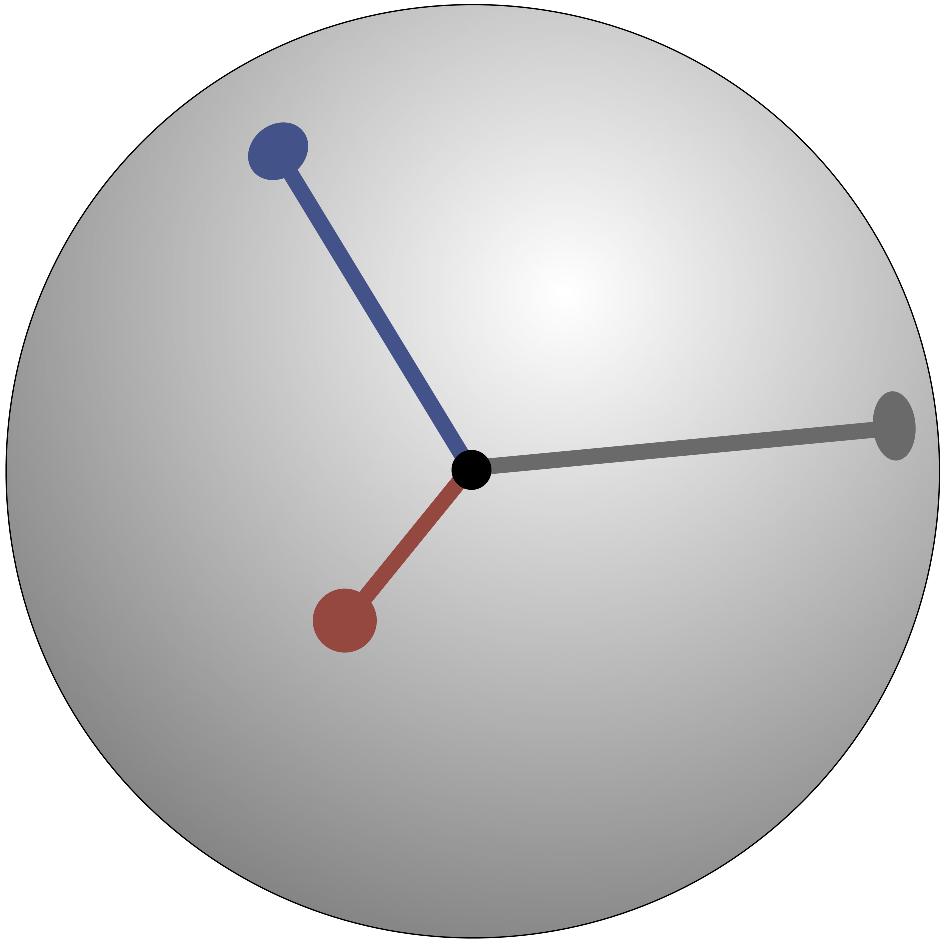
Projection

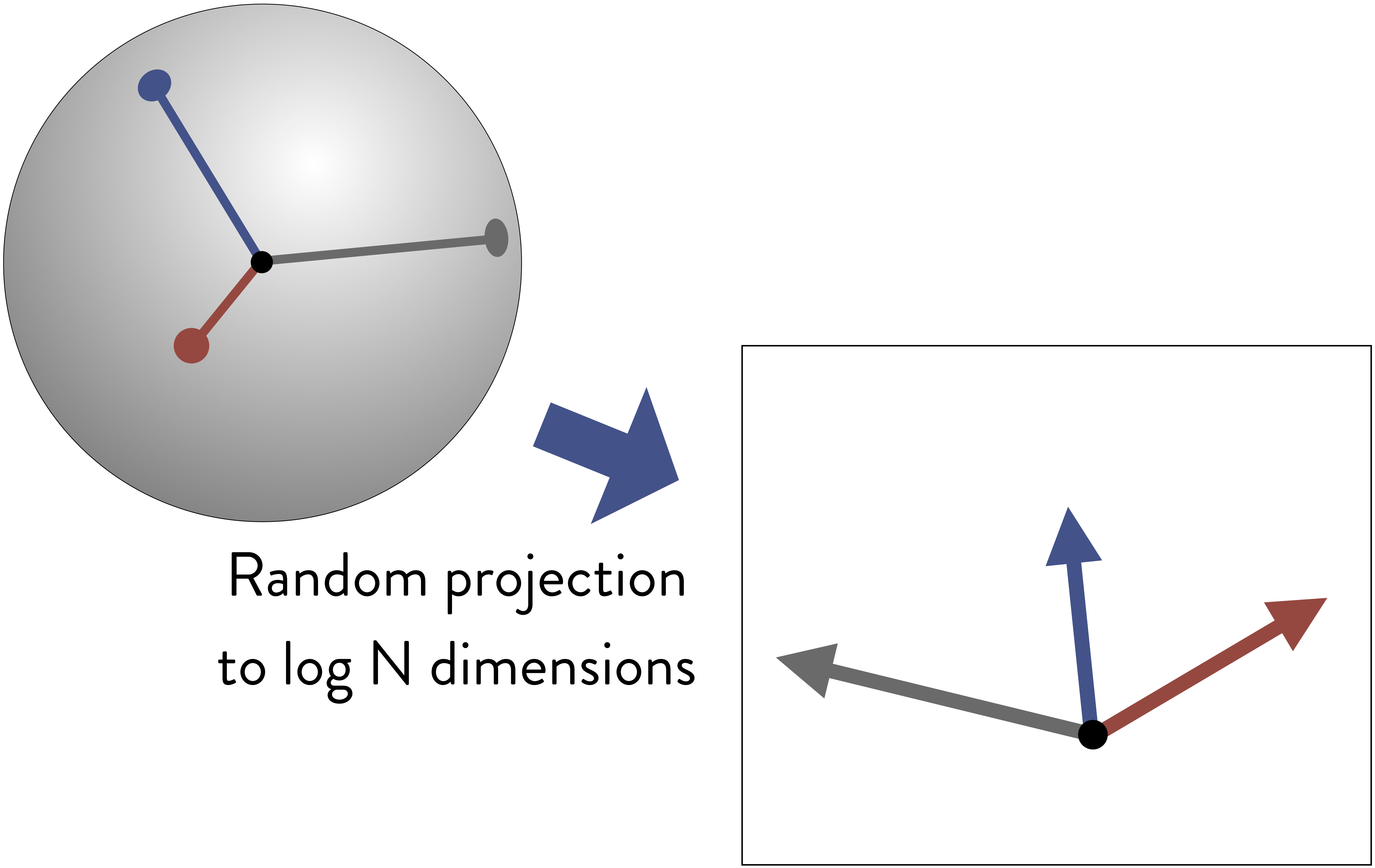


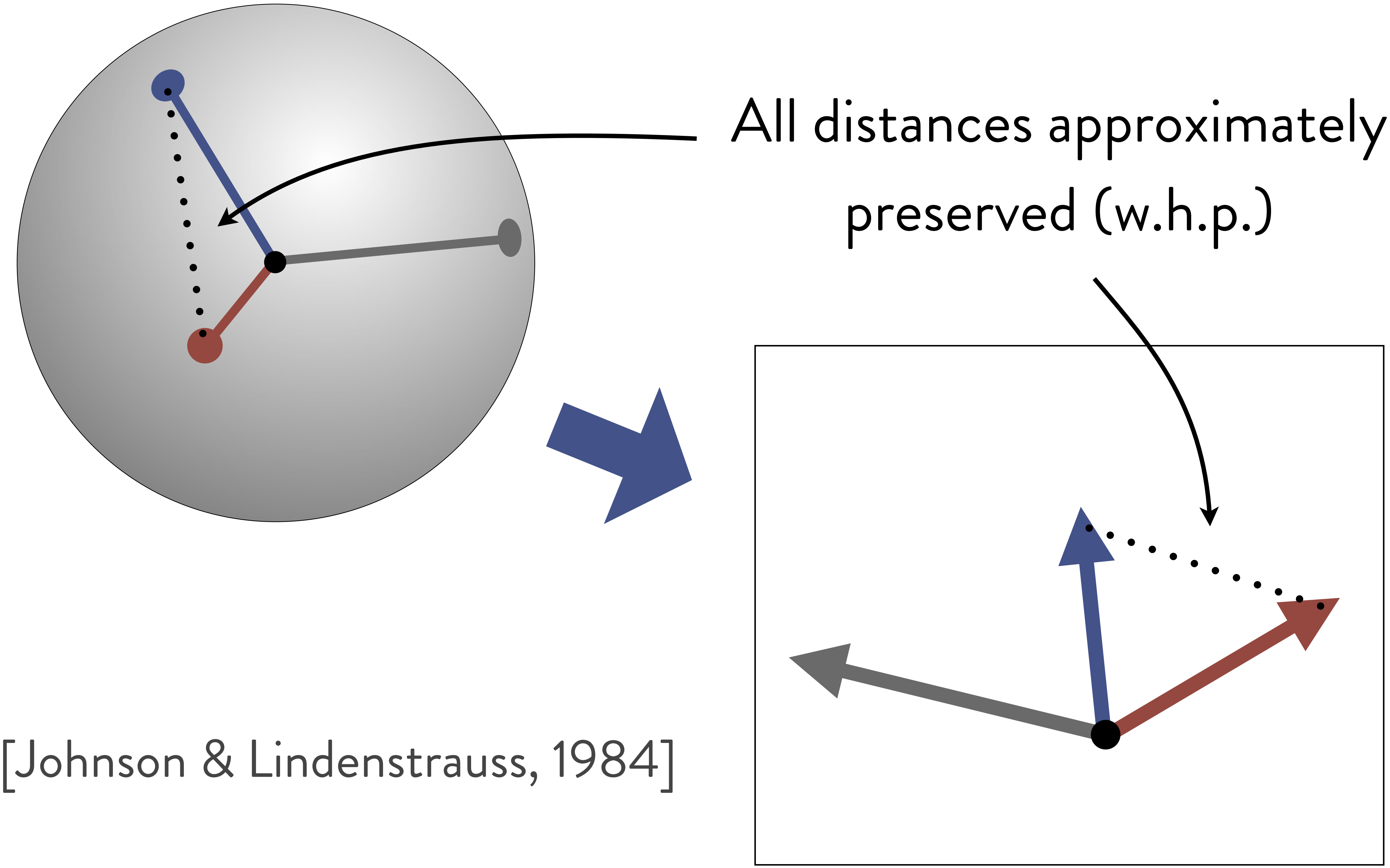
Random projection

Random projection

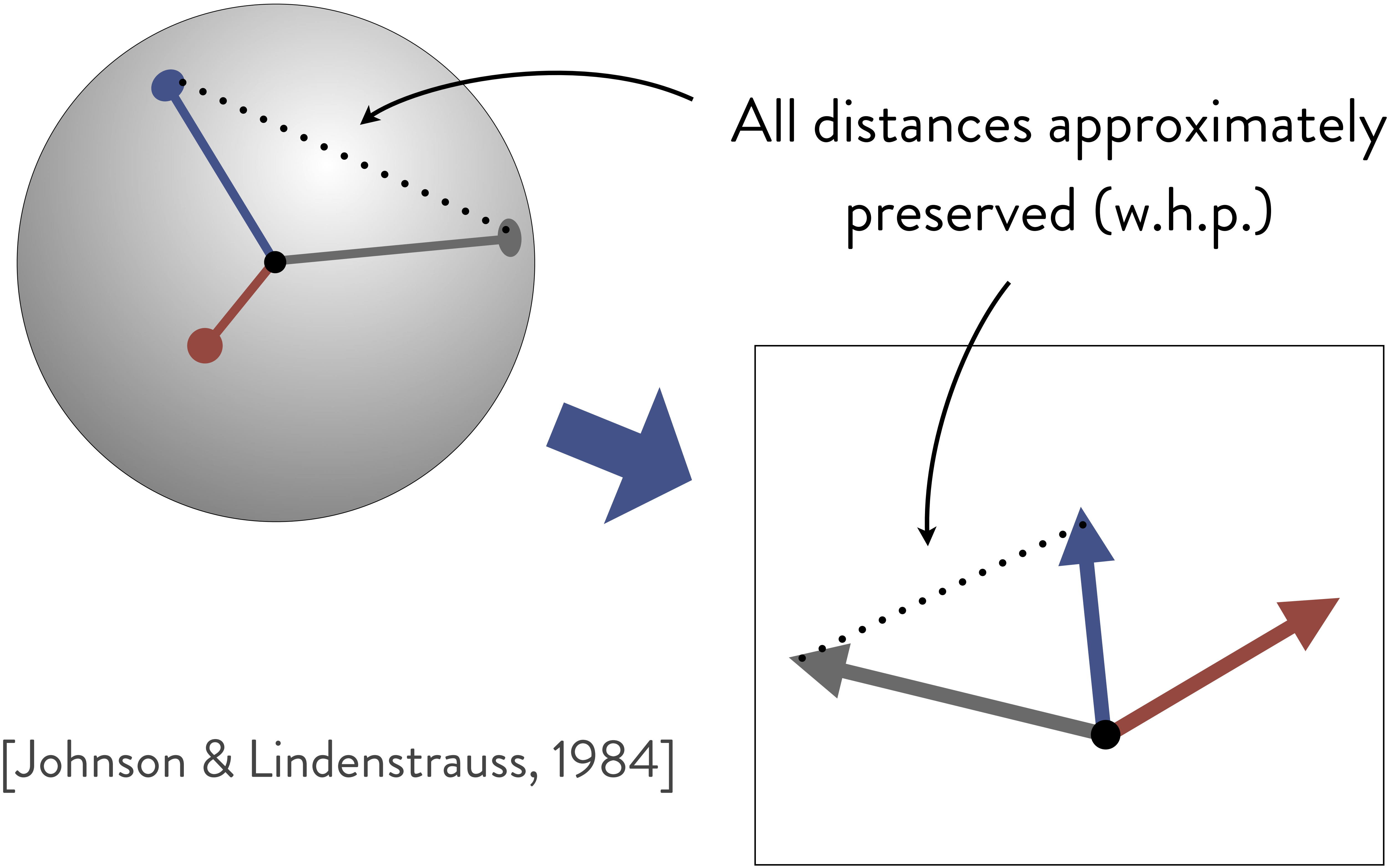




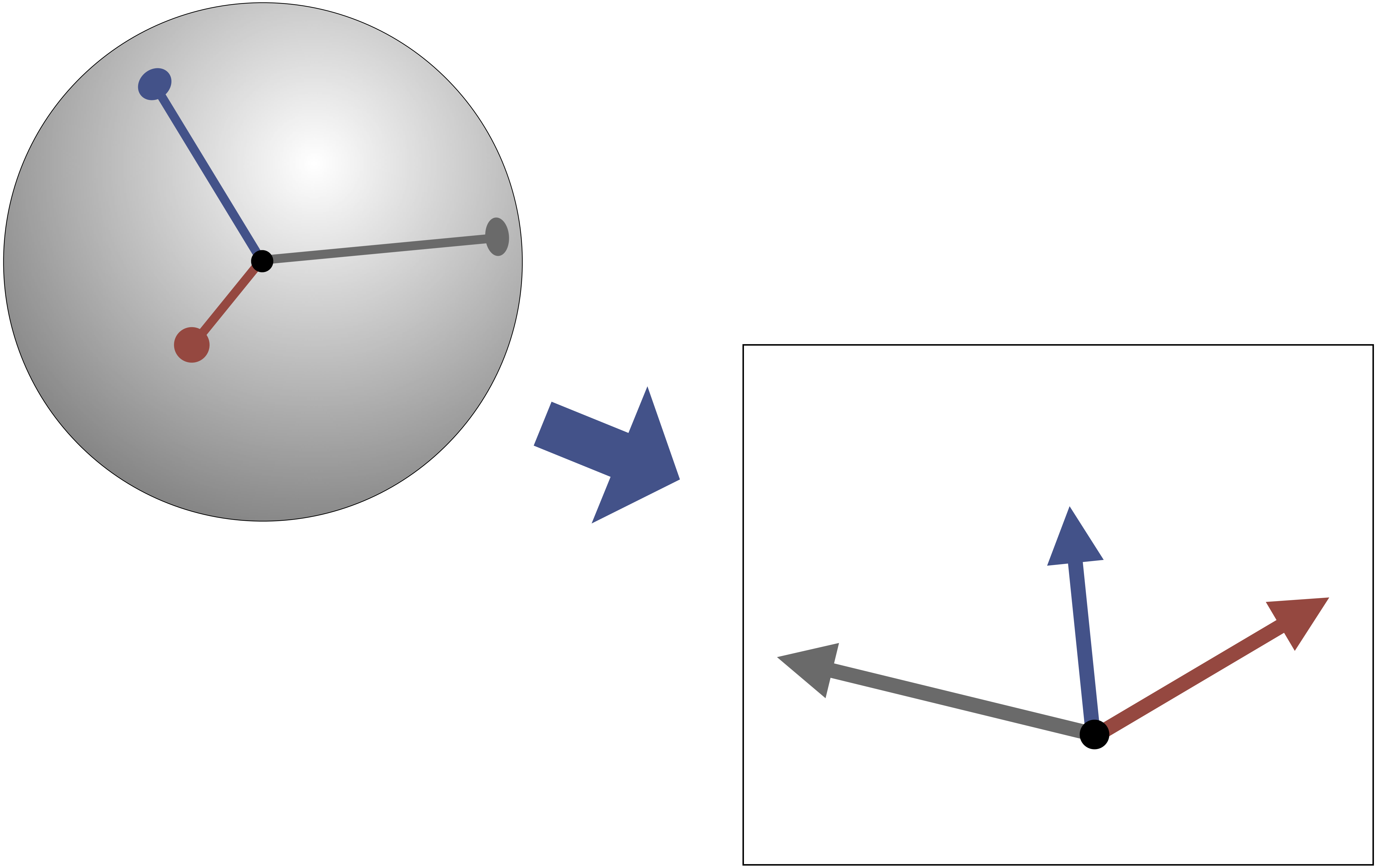

Random projection
to log N dimensions

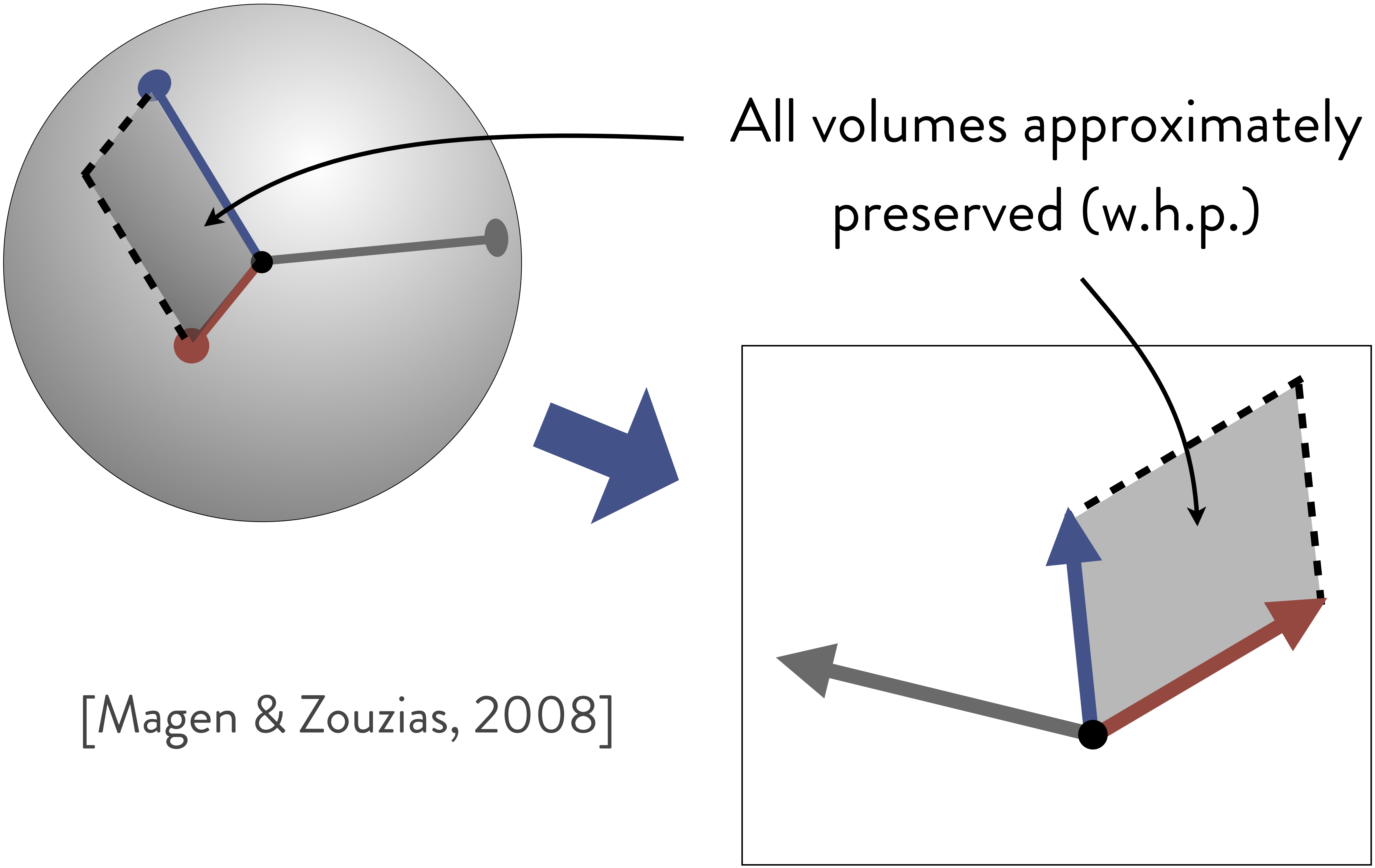


[Johnson & Lindenstrauss, 1984]

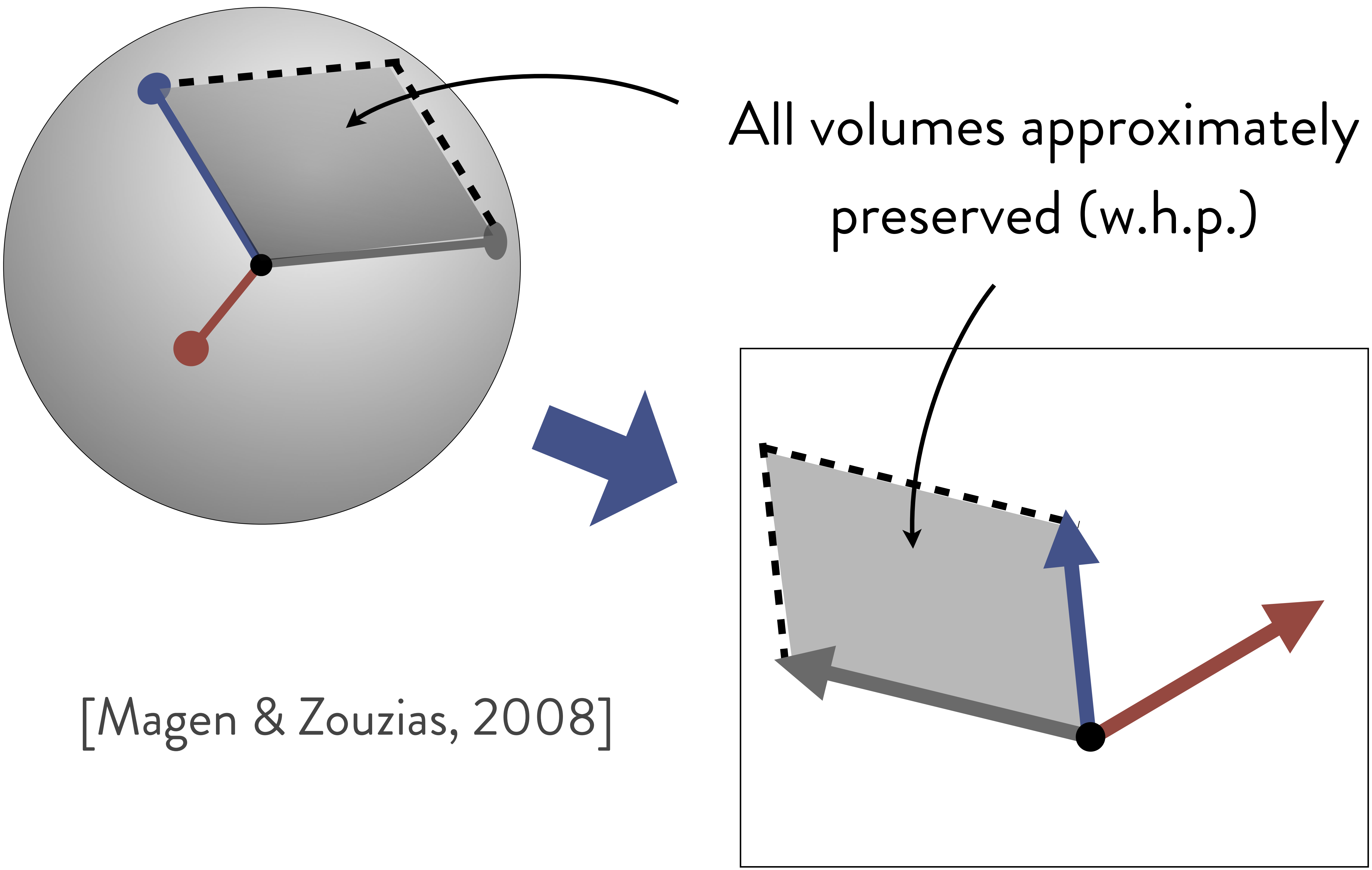


[Johnson & Lindenstrauss, 1984]





[Magen & Zouzias, 2008]


All volumes approximately preserved (w.h.p.)
[Magen & Zouzias, 2008]

Random projection for DPPs

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

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

Random projection for DPPs

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

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

- Logarithmic in N, no dependence on D

Random projection for DPPs

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

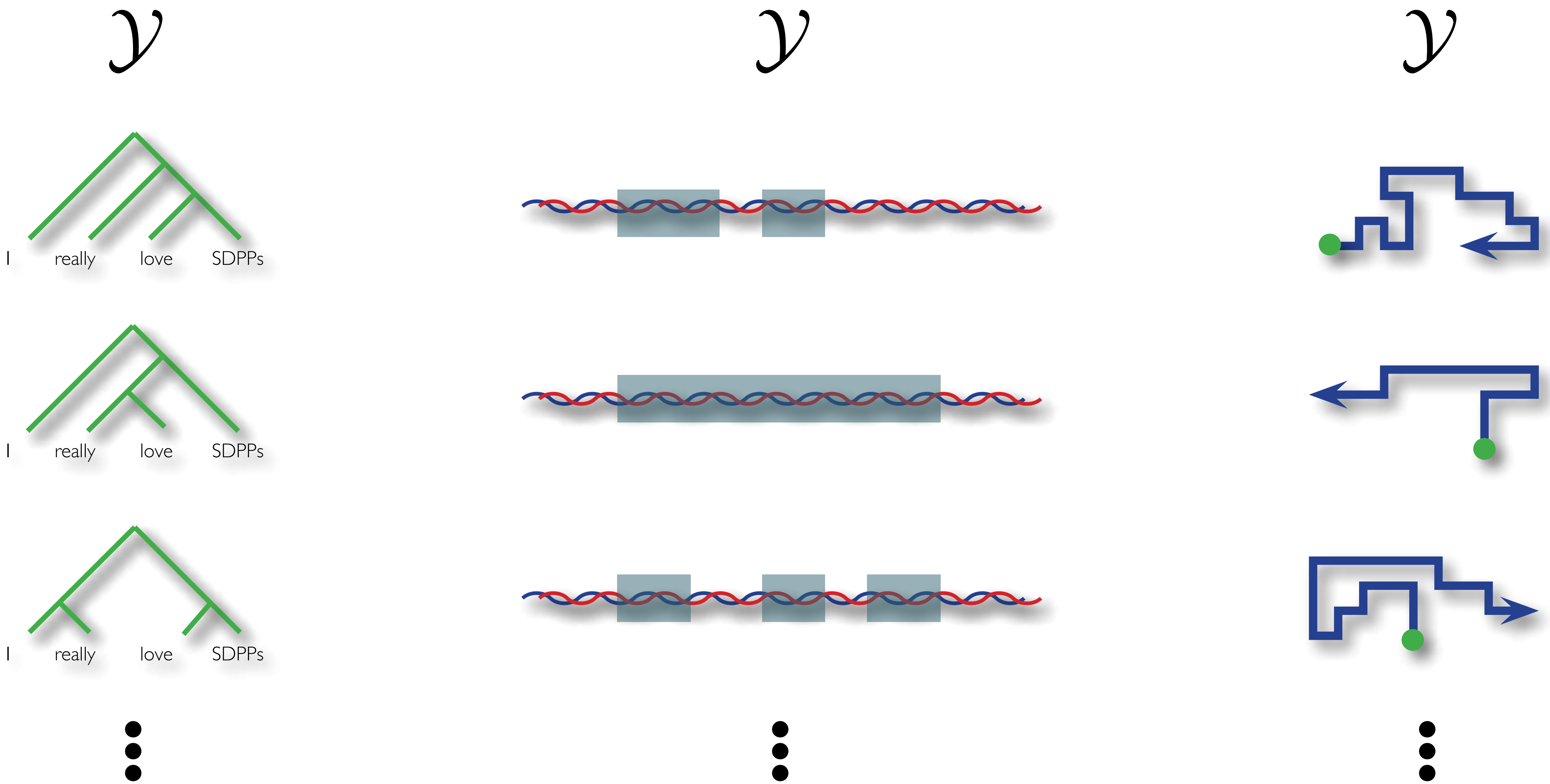
$$\|\mathcal{P} - \tilde{\mathcal{P}}\|_1 \leq O(\epsilon) \ .$$

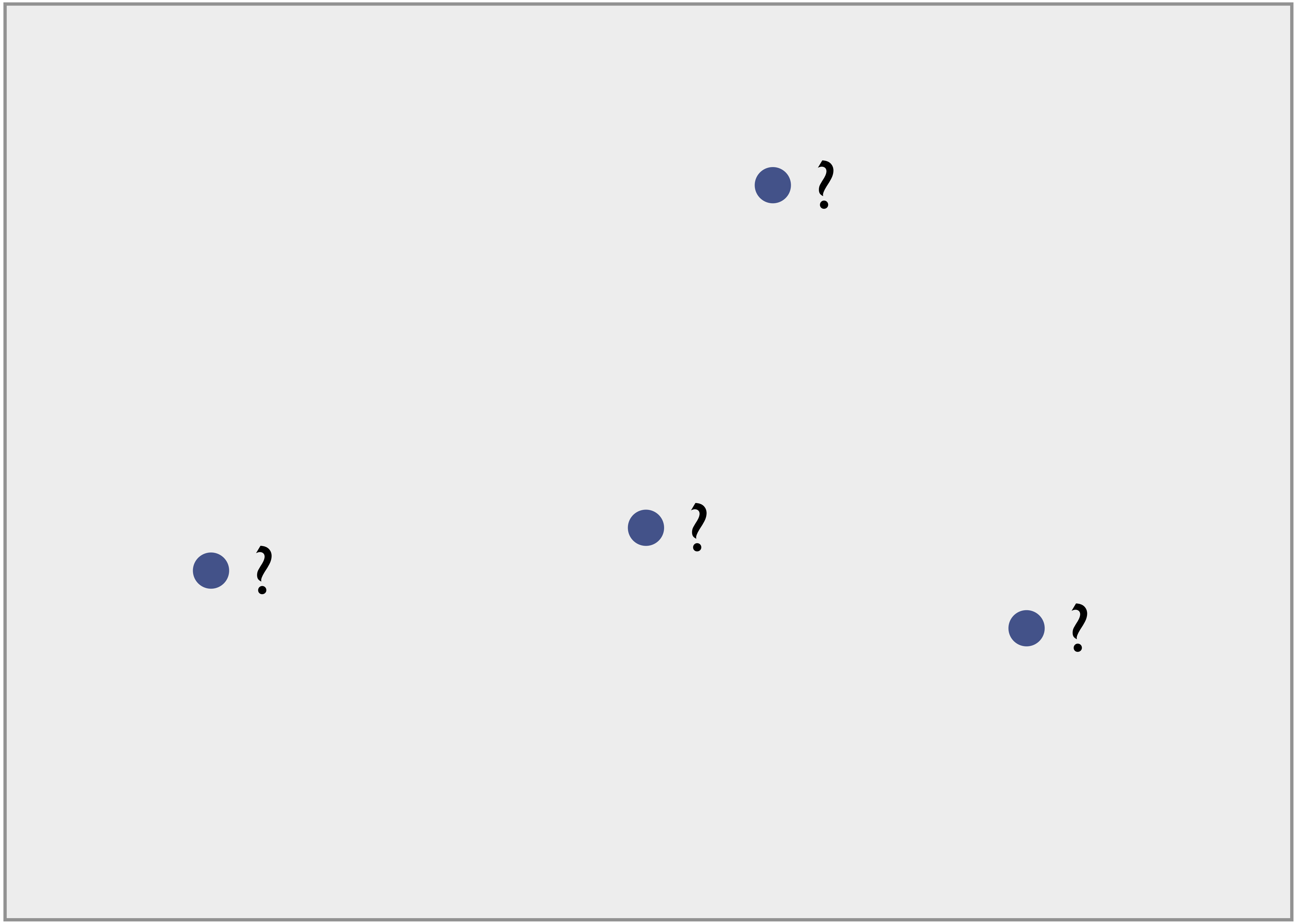
- Logarithmic in N, no dependence on D
- Small, d x d dual representation

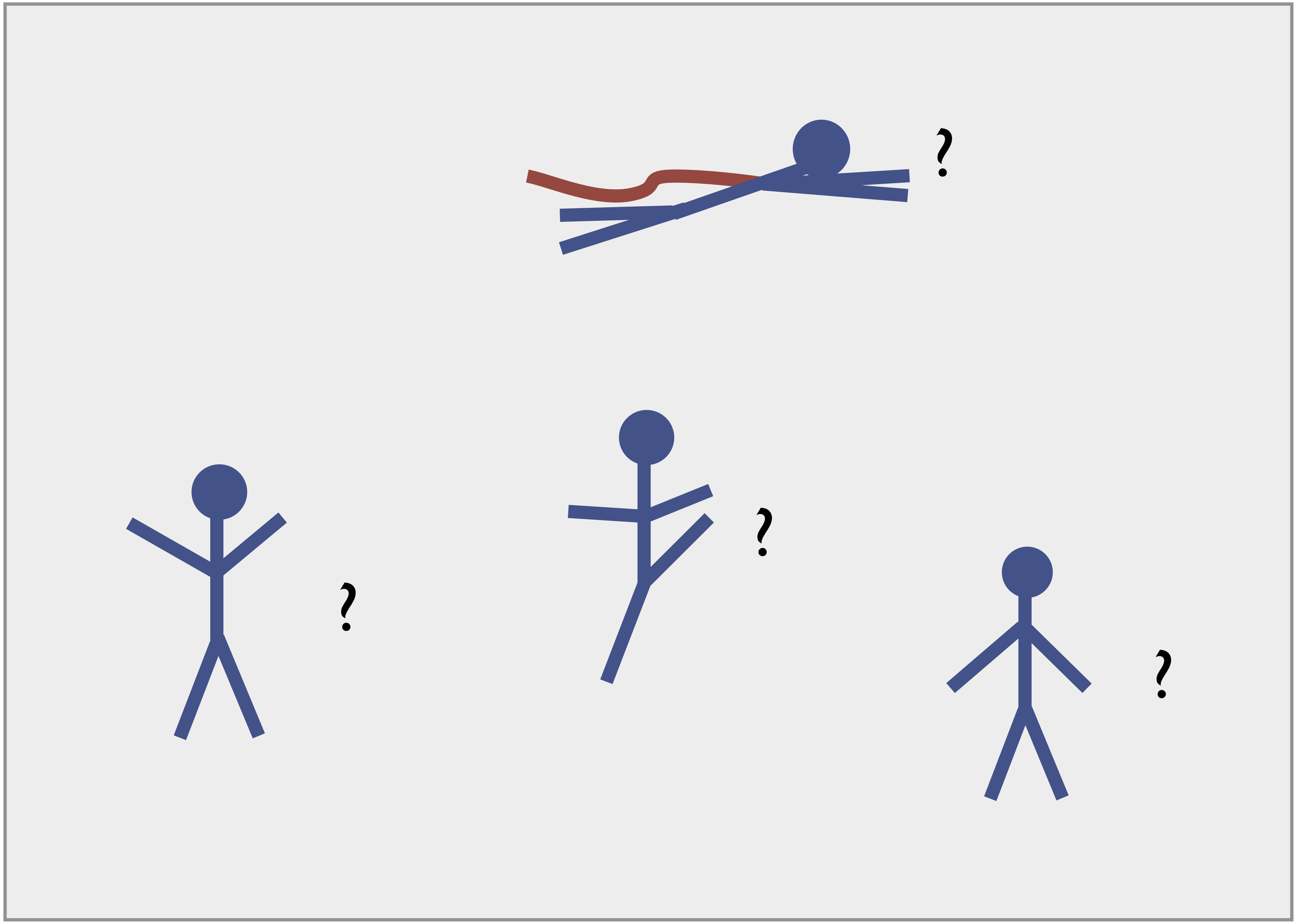
DPPs at scale

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

Exponential N?







Structured DPPs

- Exponentially many complex "items"
- Can't even handle O(*N)*
- But can still compute marginals and sample!

Structured DPPs

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

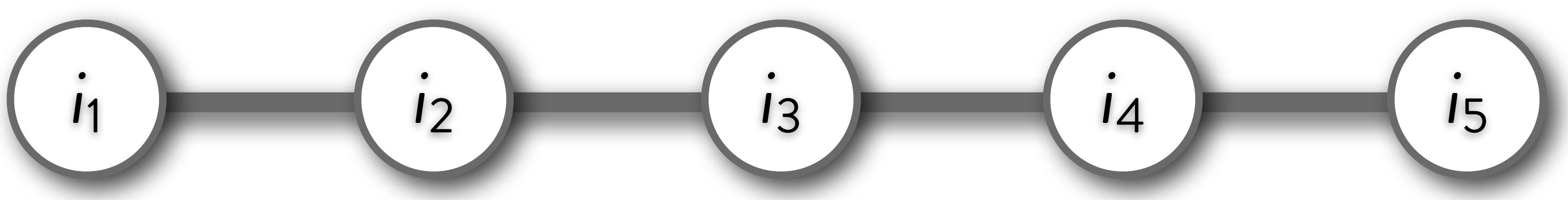
1. Factorized model
2. Dual DPPs
3. Second order message-passing

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:

$$q(\boldsymbol{i}) = \prod_{\alpha} q(i_\alpha)$$

- Diversity features factor additively:

1. Factorization

- Quality scores factor multiplicatively:

$$q(\boldsymbol{i}) = \prod_{\alpha} q(i_\alpha)$$

e.g., MRF

- Diversity features factor additively:

1. Factorization

- Quality scores factor multiplicatively:

$$q(\boldsymbol{i}) = \prod_{\alpha} q(i_\alpha)$$

e.g., MRF

- Diversity features factor additively:

$$\phi(\boldsymbol{i}) = \sum_{\alpha} \phi(i_\alpha)$$

1. Factorization

- Quality scores factor multiplicatively:

$$q(\boldsymbol{i}) = \prod_{\alpha} q(i_\alpha)$$

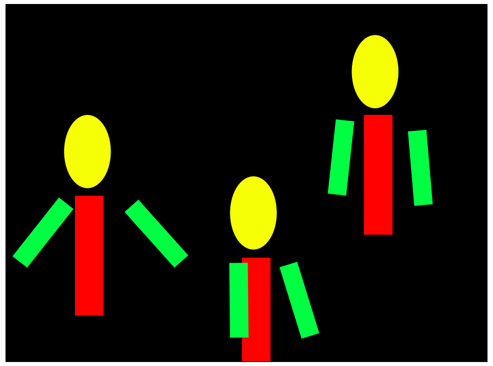
e.g., MRF

- Diversity features factor additively:

$$\phi(\boldsymbol{i}) = \sum_{\alpha} \phi(i_\alpha)$$

e.g., Hamming

Multiple-pose estimation



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

Quality

Quality

Quality

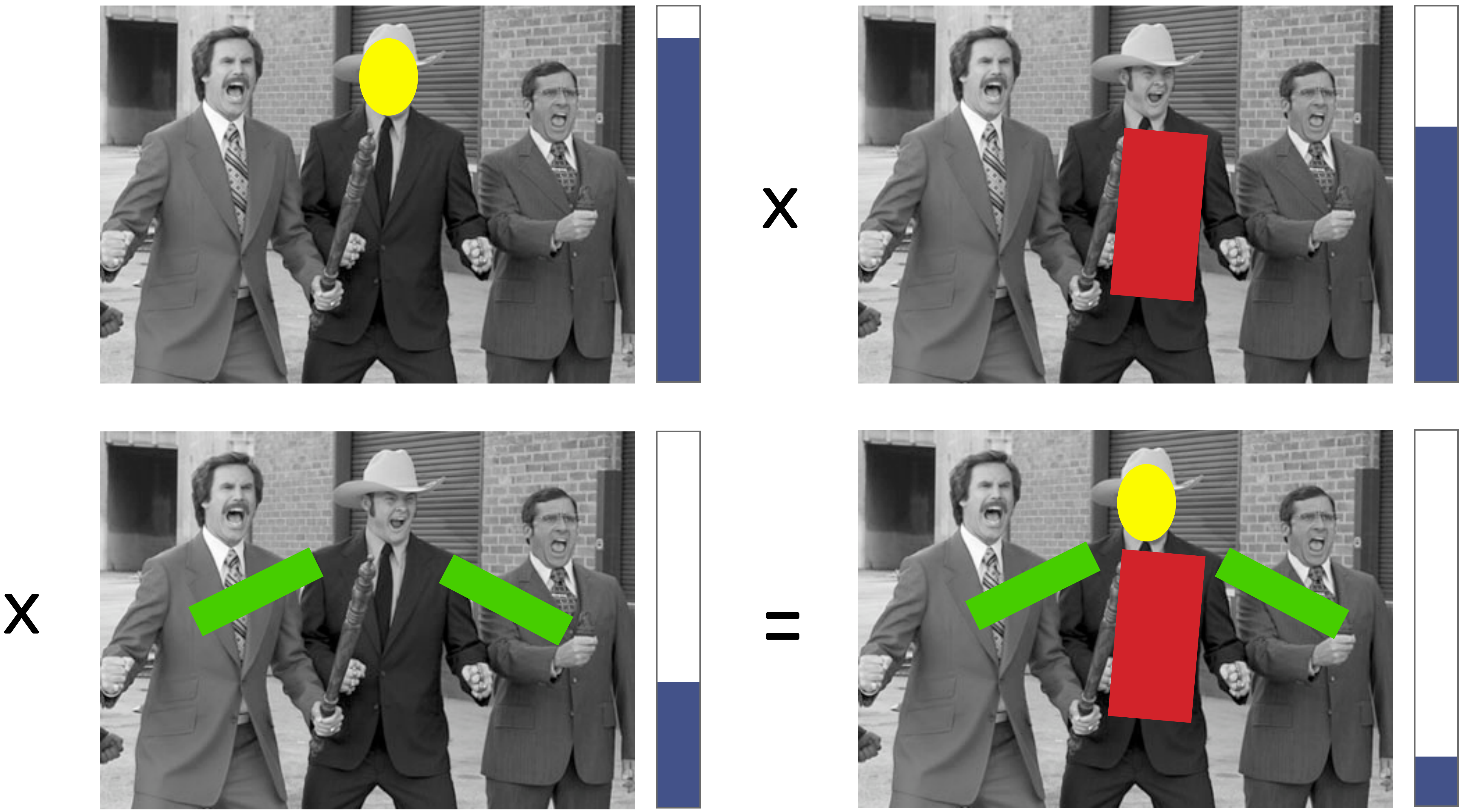
x

Quality

x

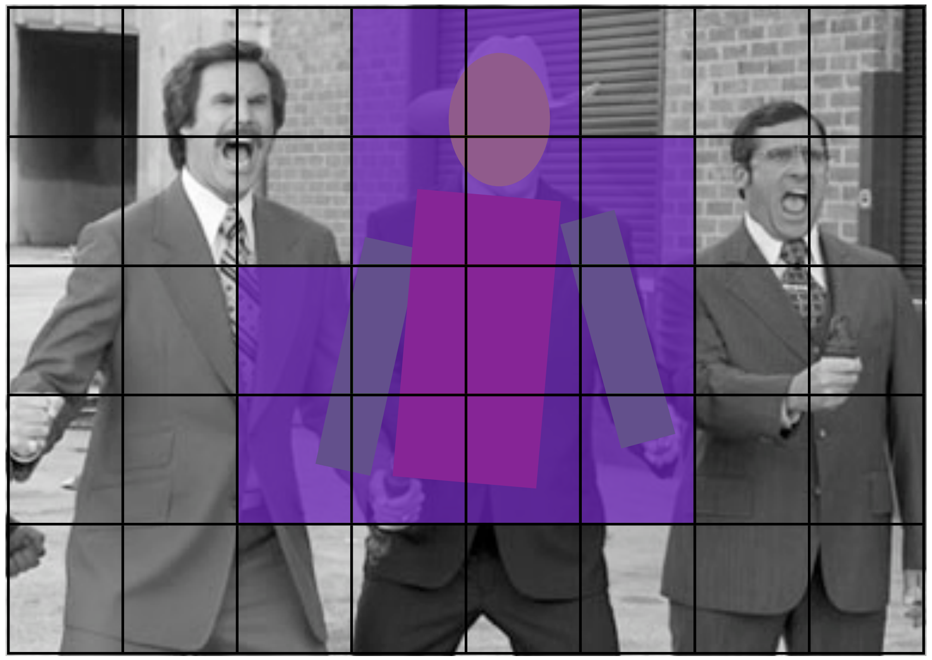
x

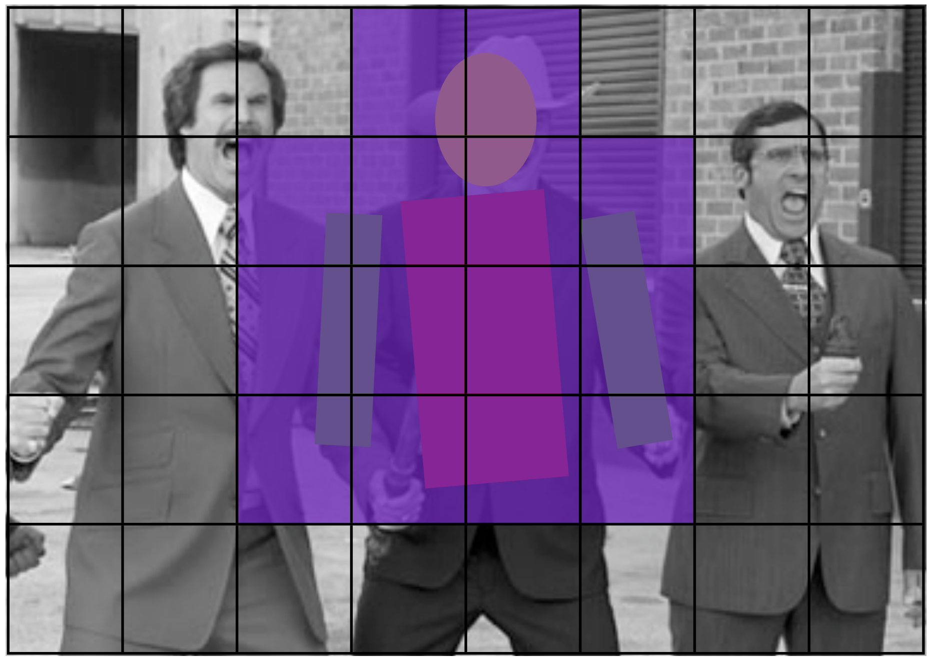
Quality



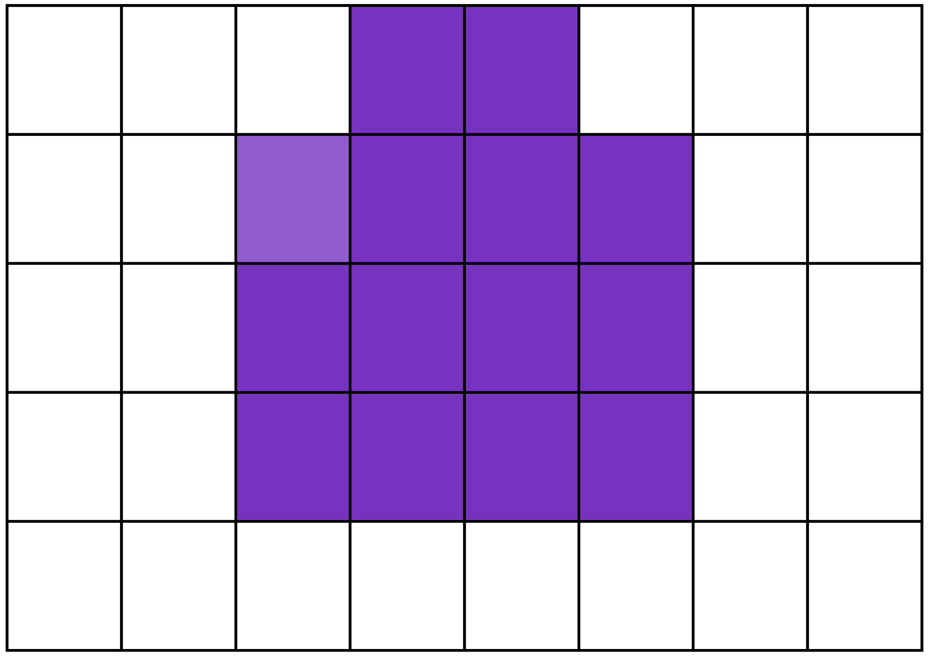
Diversity

Diversity



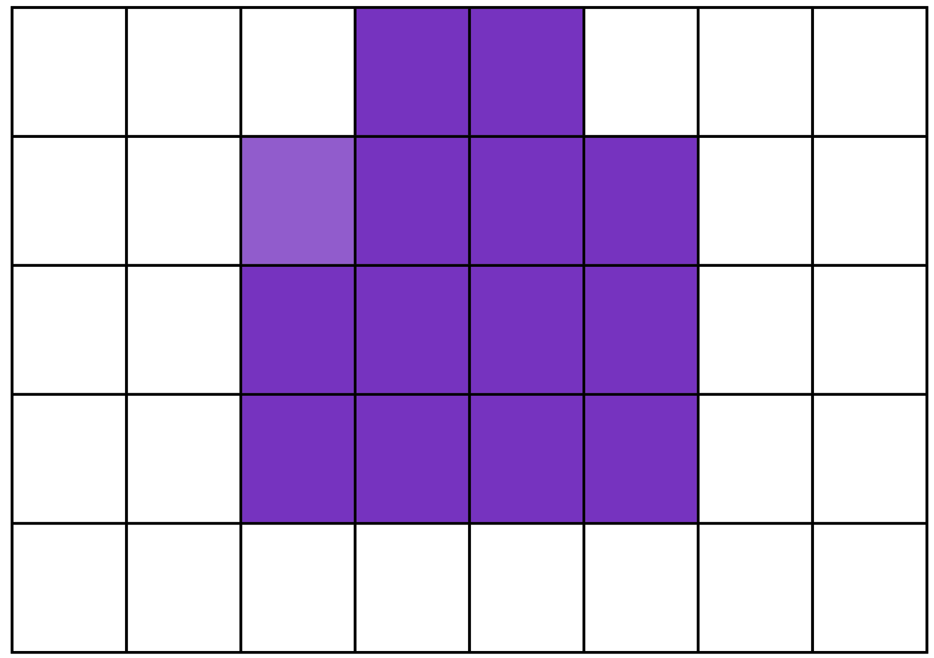


Diversity

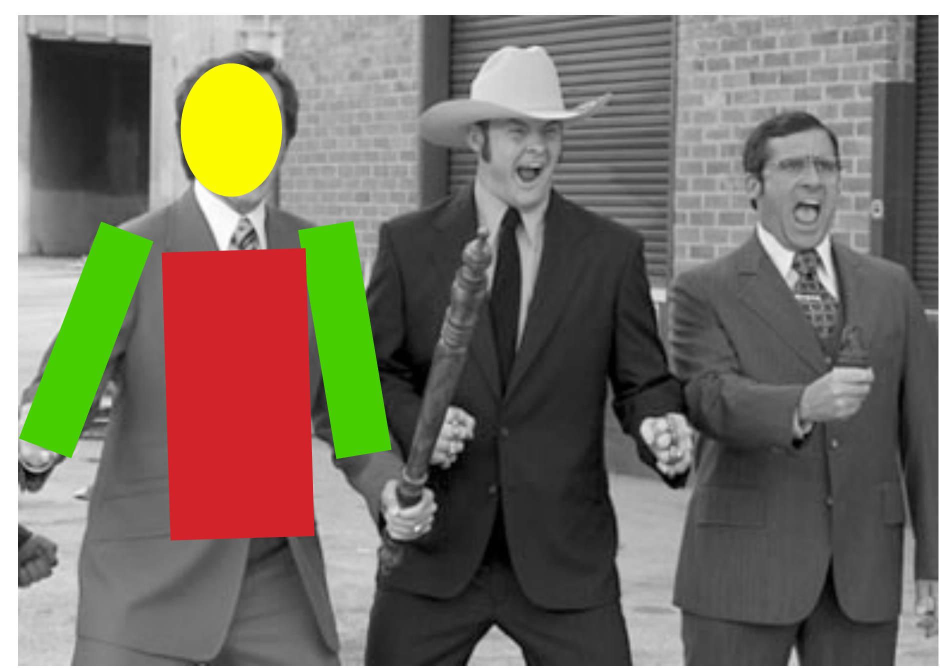


Low diversity

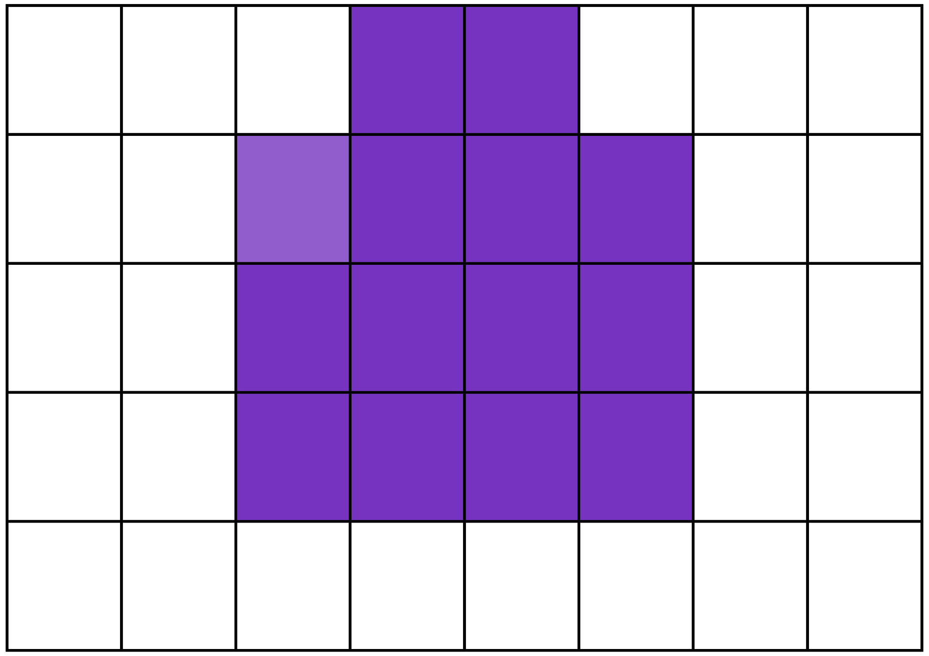
Diversity



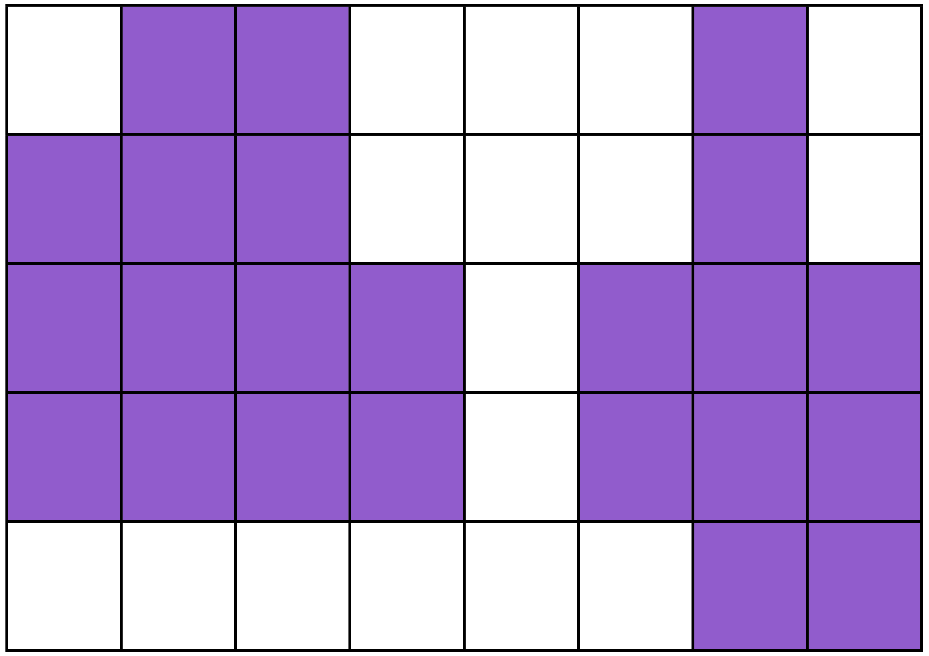
Low diversity



Diversity

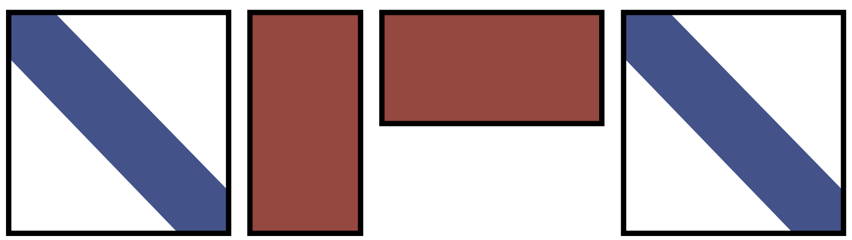


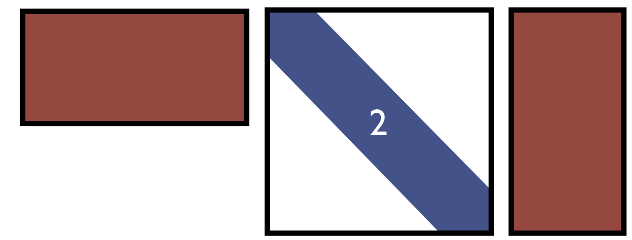
Low diversity



High diversity

2. Dual representation

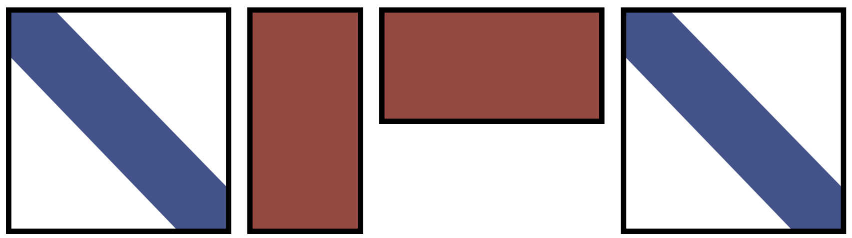
$L =$ 

$C =$ 

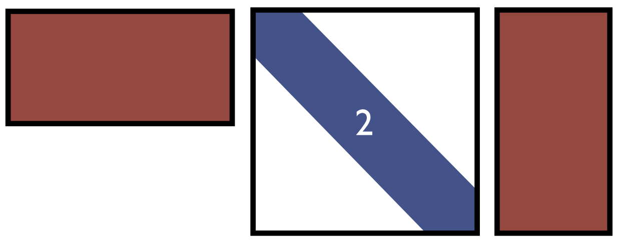
$N \times N$

$D \times D$

2. Dual representation

$L =$ 

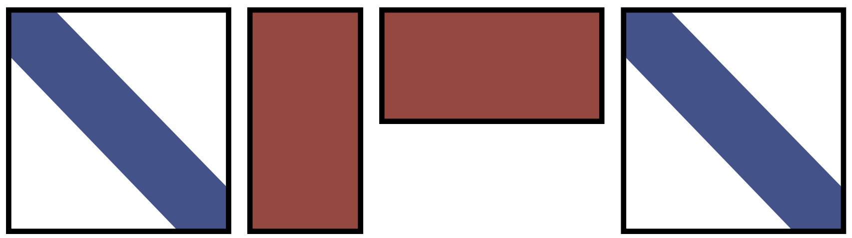
$N \times N$

$C =$ 

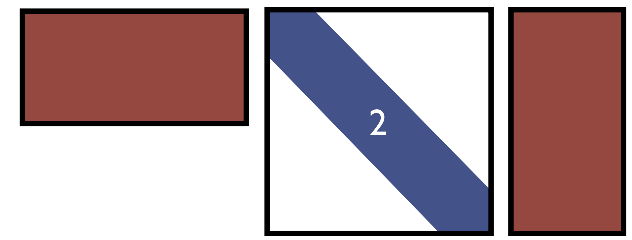
$D \times D$

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

2. Dual representation

L = 

N x N

C = 

D x D

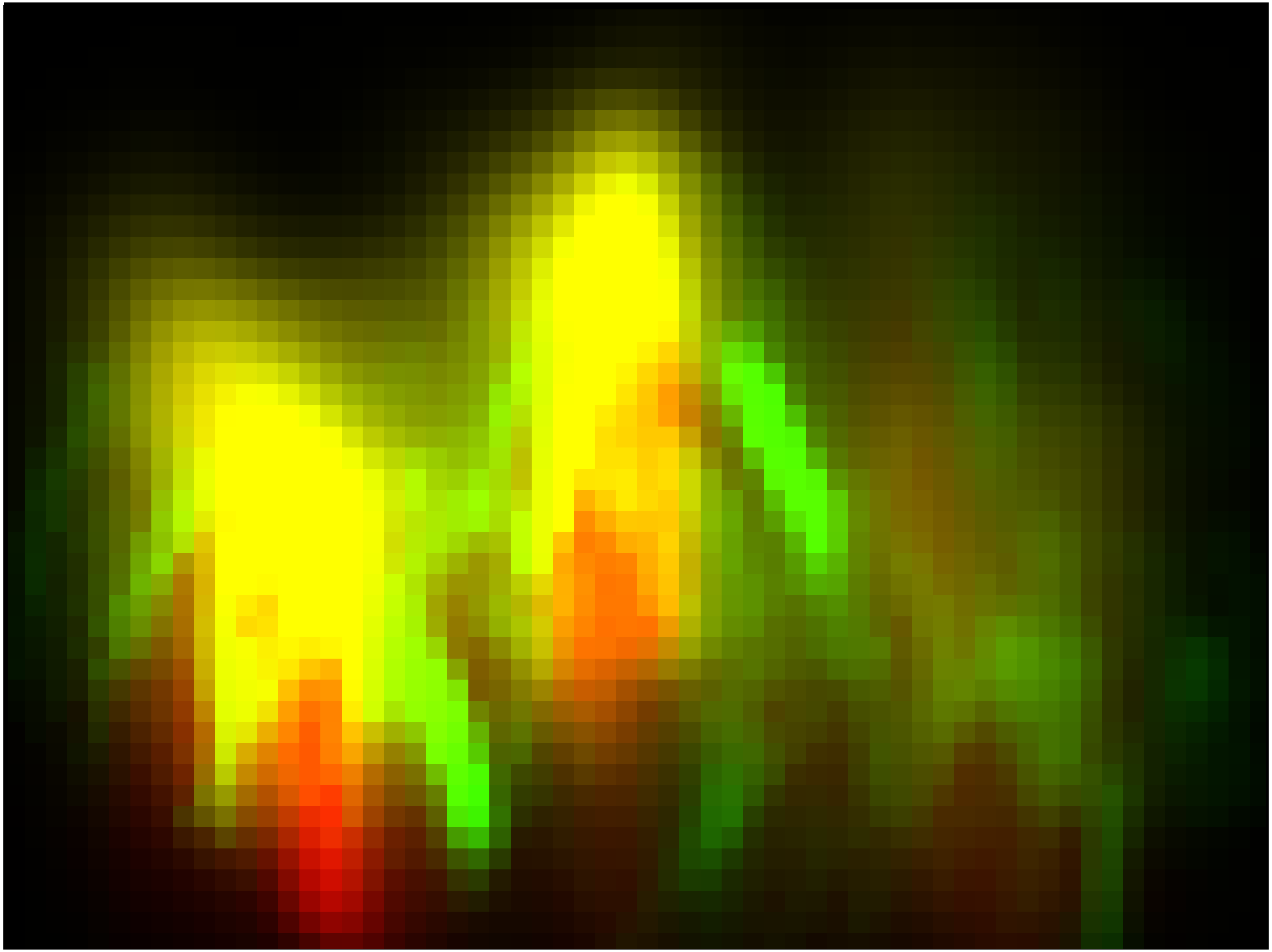
$$C_{rl} = \sum_{\boldsymbol{i}} q^2(\boldsymbol{i})\phi_r(\boldsymbol{i})\phi_l(\boldsymbol{i})$$

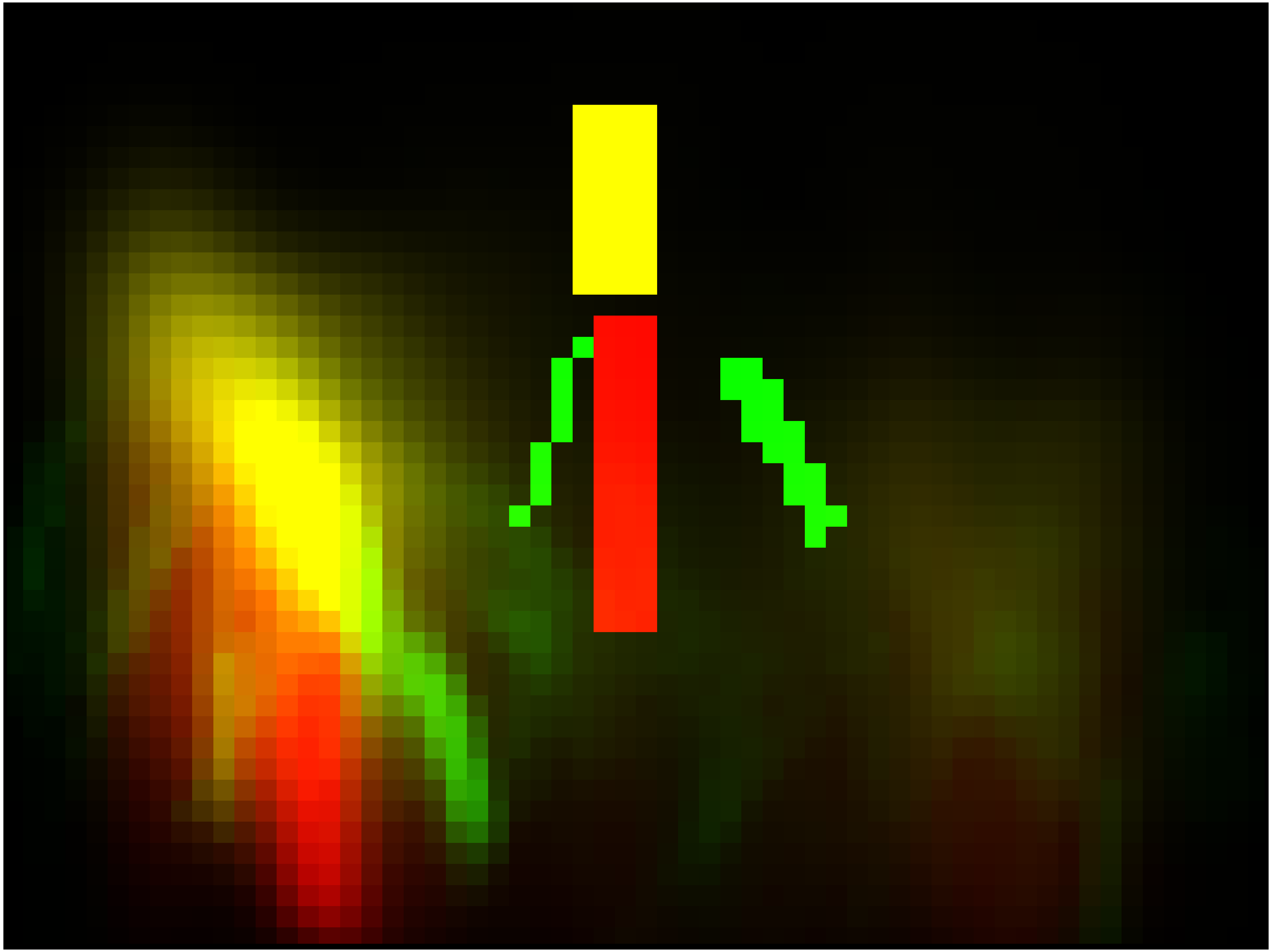
C is covariance of ϕ under $\Pr(\boldsymbol{i}) \propto q^2(\boldsymbol{i})$

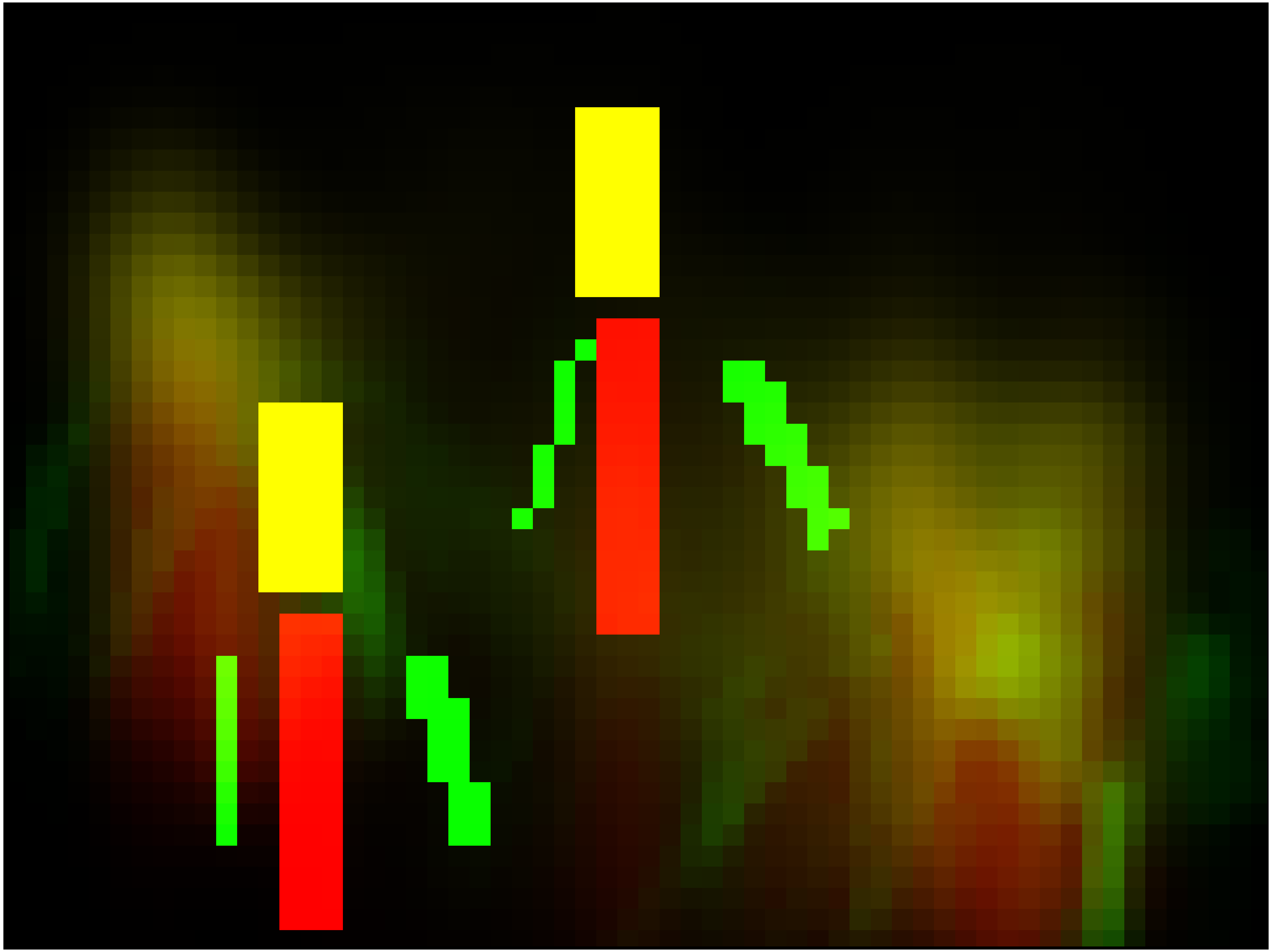
3. Second-order message passing

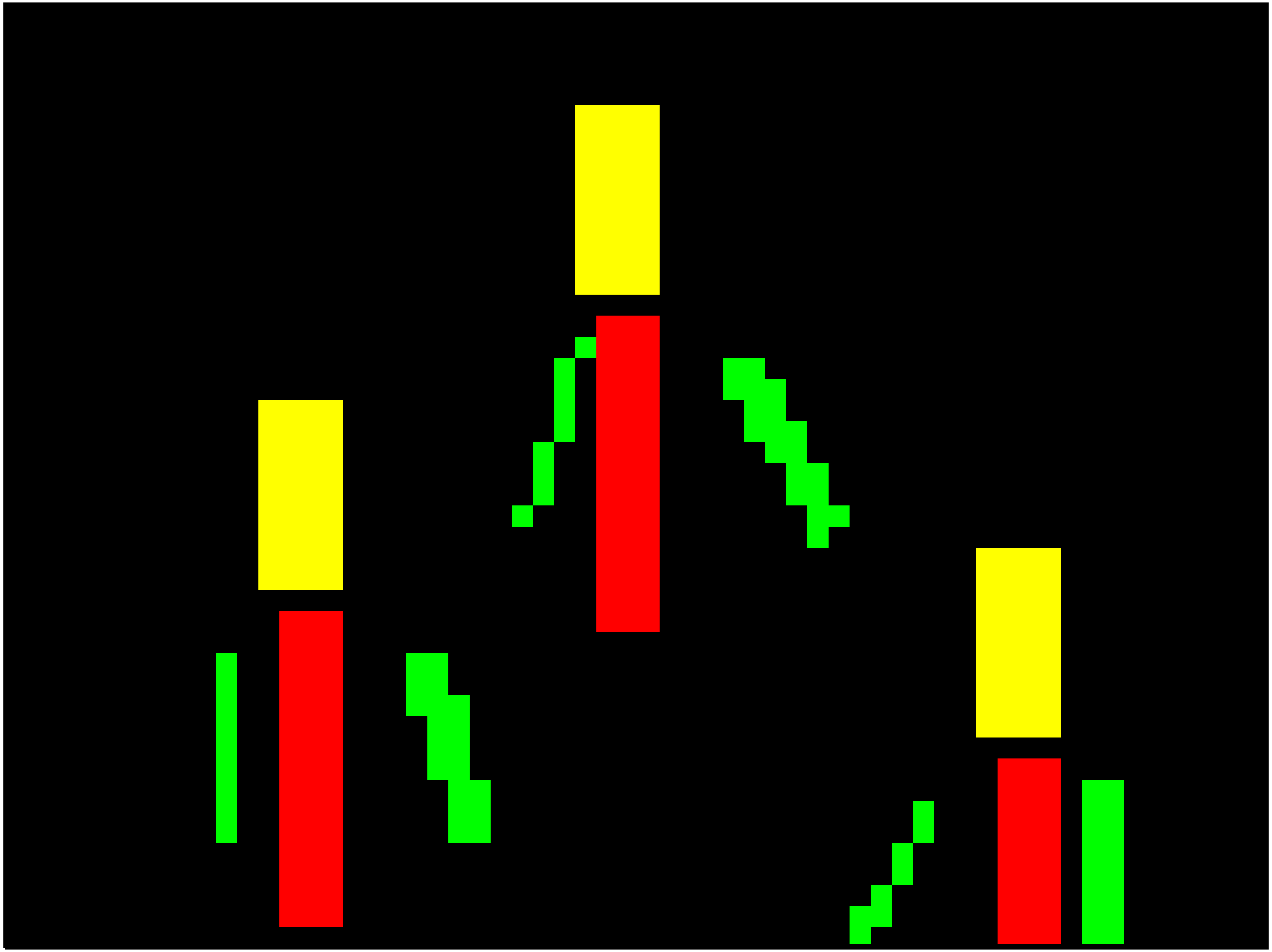
- Can compute feature covariance using message passing when graph is a tree
- Use special semiring in place of sum-product
- Linear in number of nodes
- Quadratic in dimension of diversity features ϕ

[Li + Eisner, 2009]

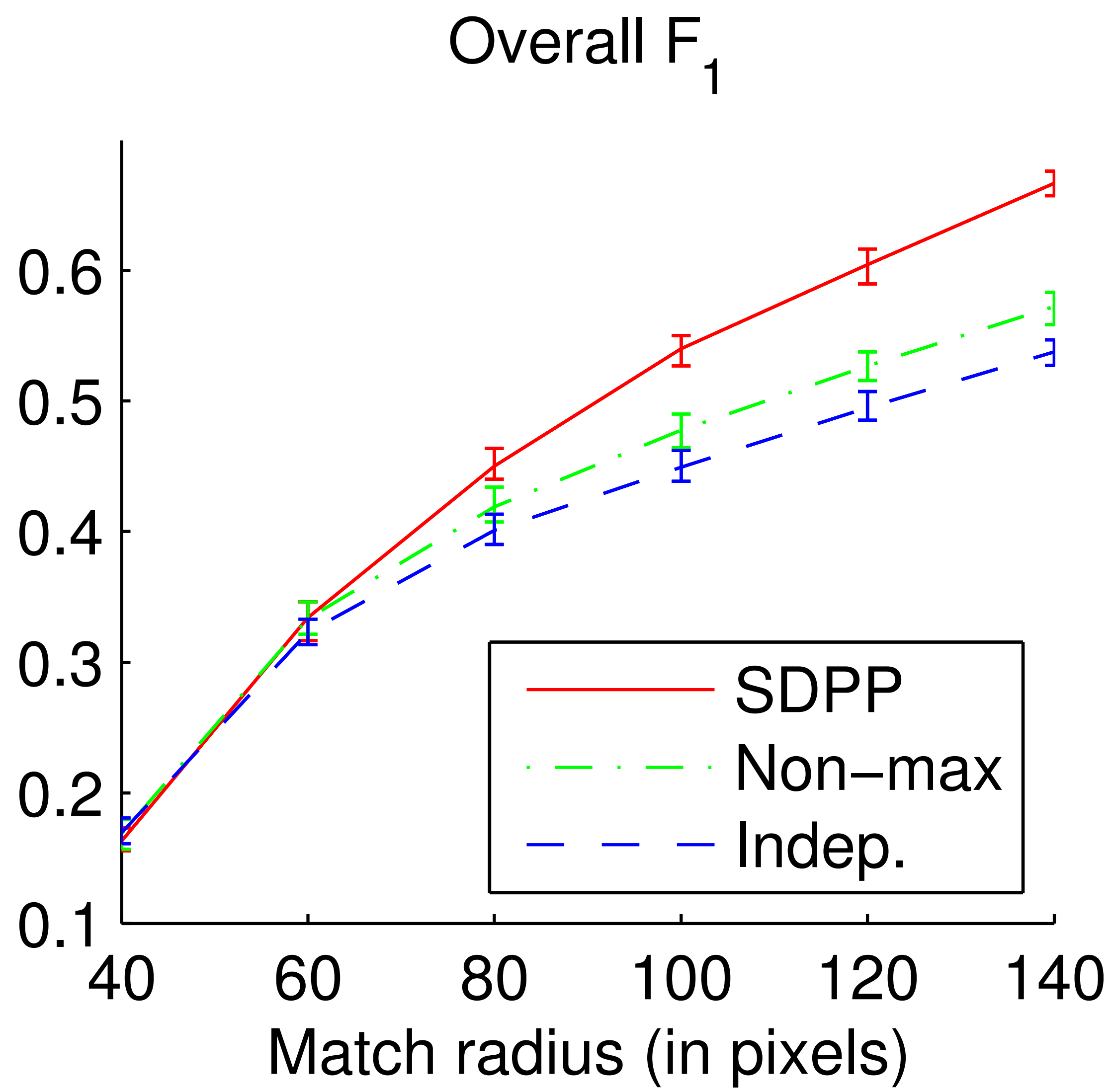




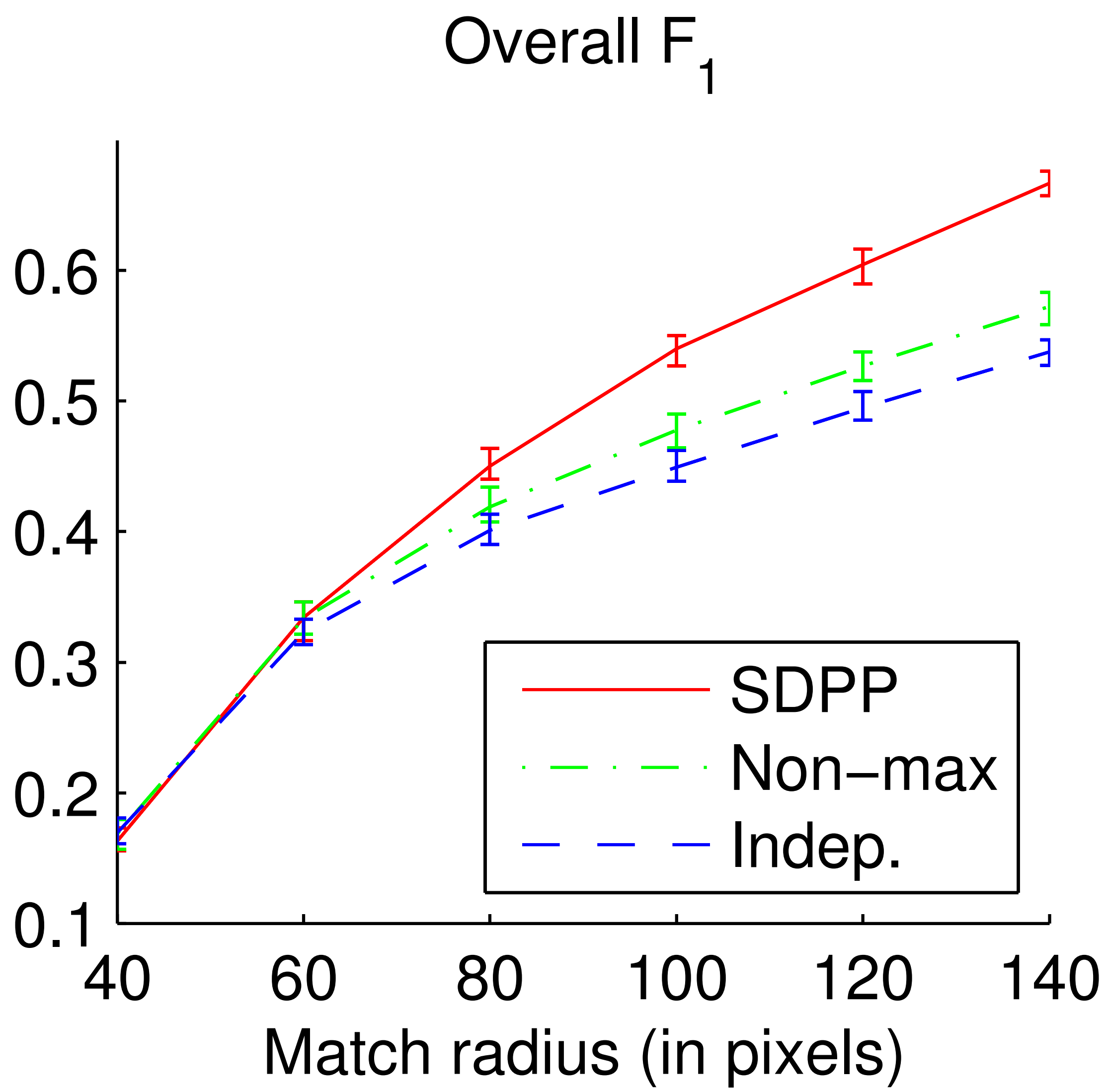


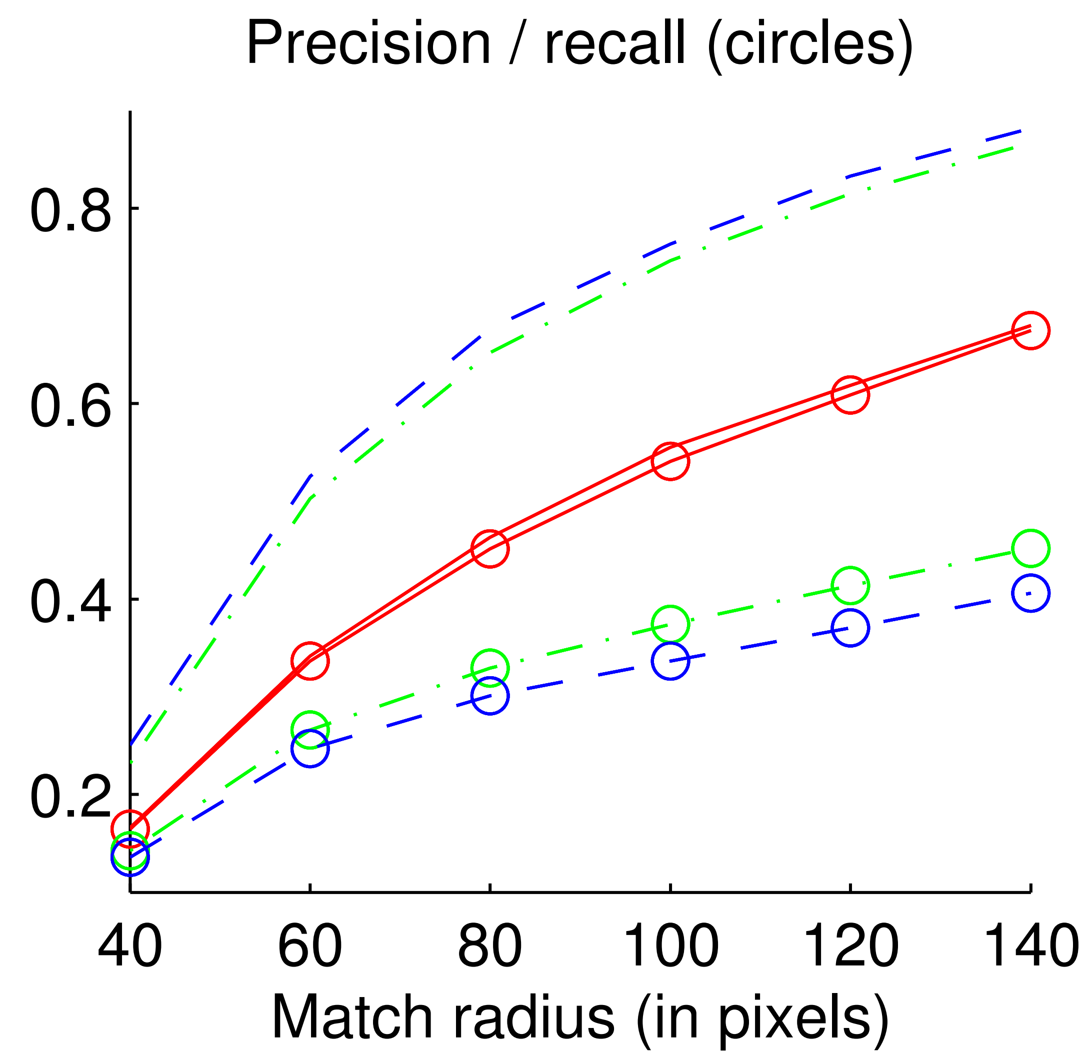


Pose accuracy



Pose accuracy



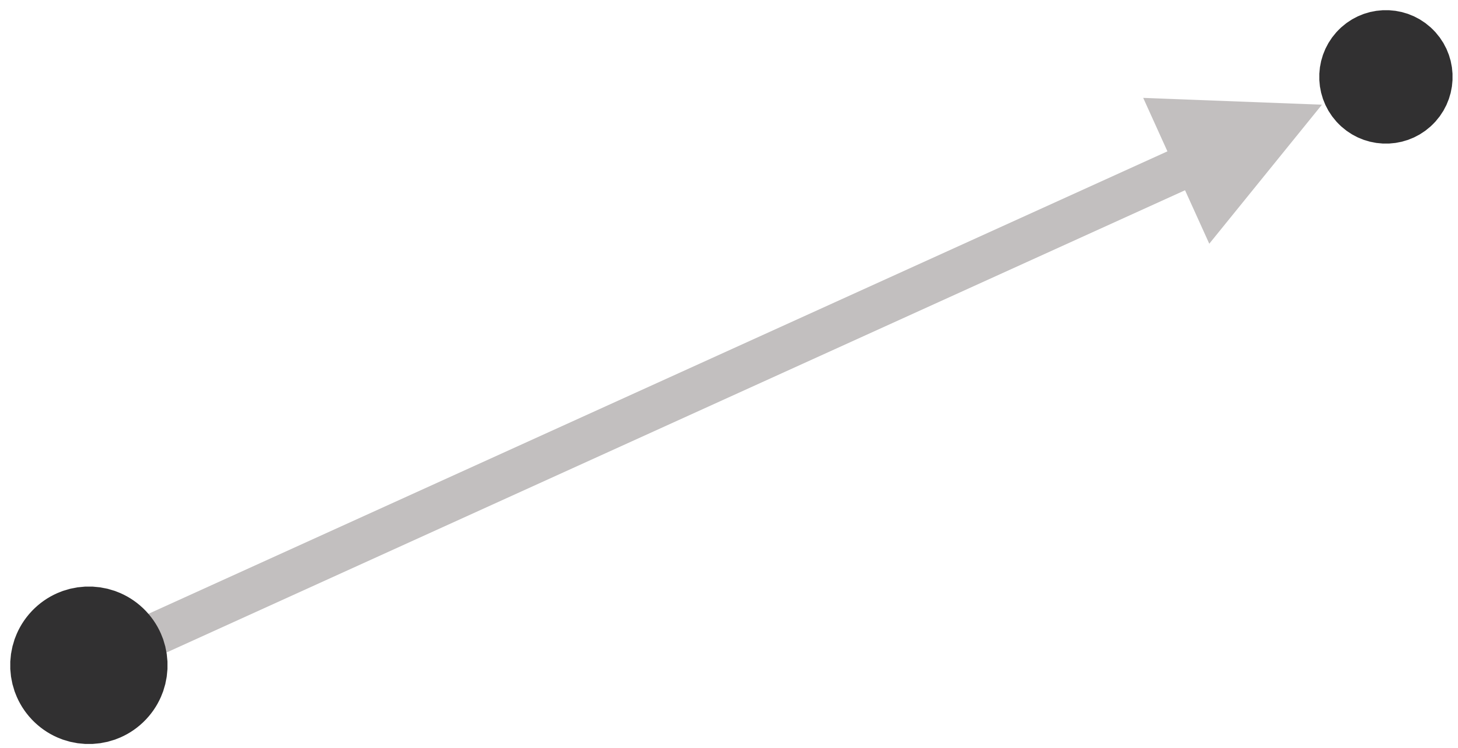


News threading

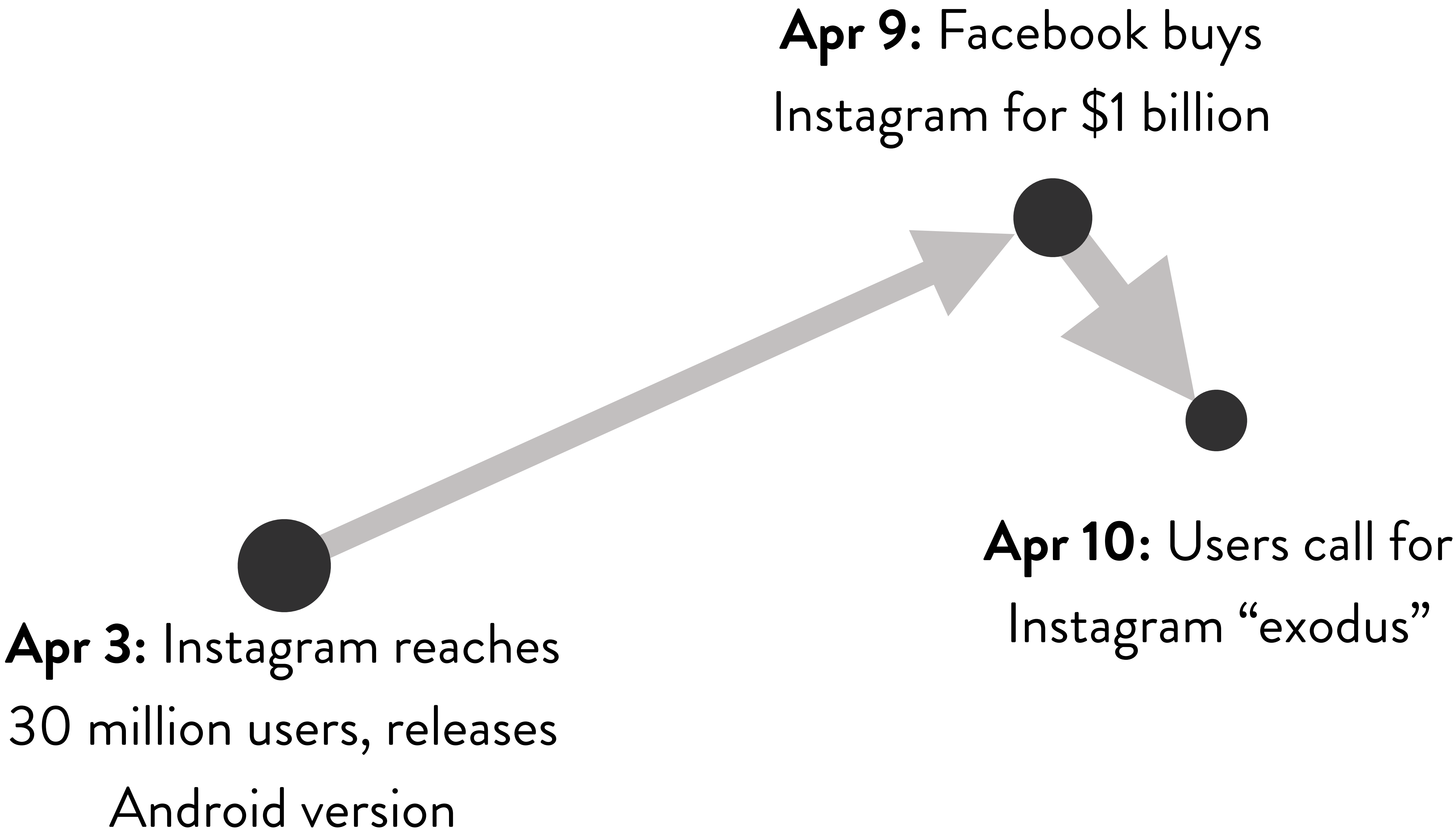
- **Input**: large news corpus
- **Output**: threads of articles
  - Each thread narrates a major story
  - Threads are diverse to cover many stories
- Combine *k*-DPPs, structured DPPs, dual DPPs, and random projection

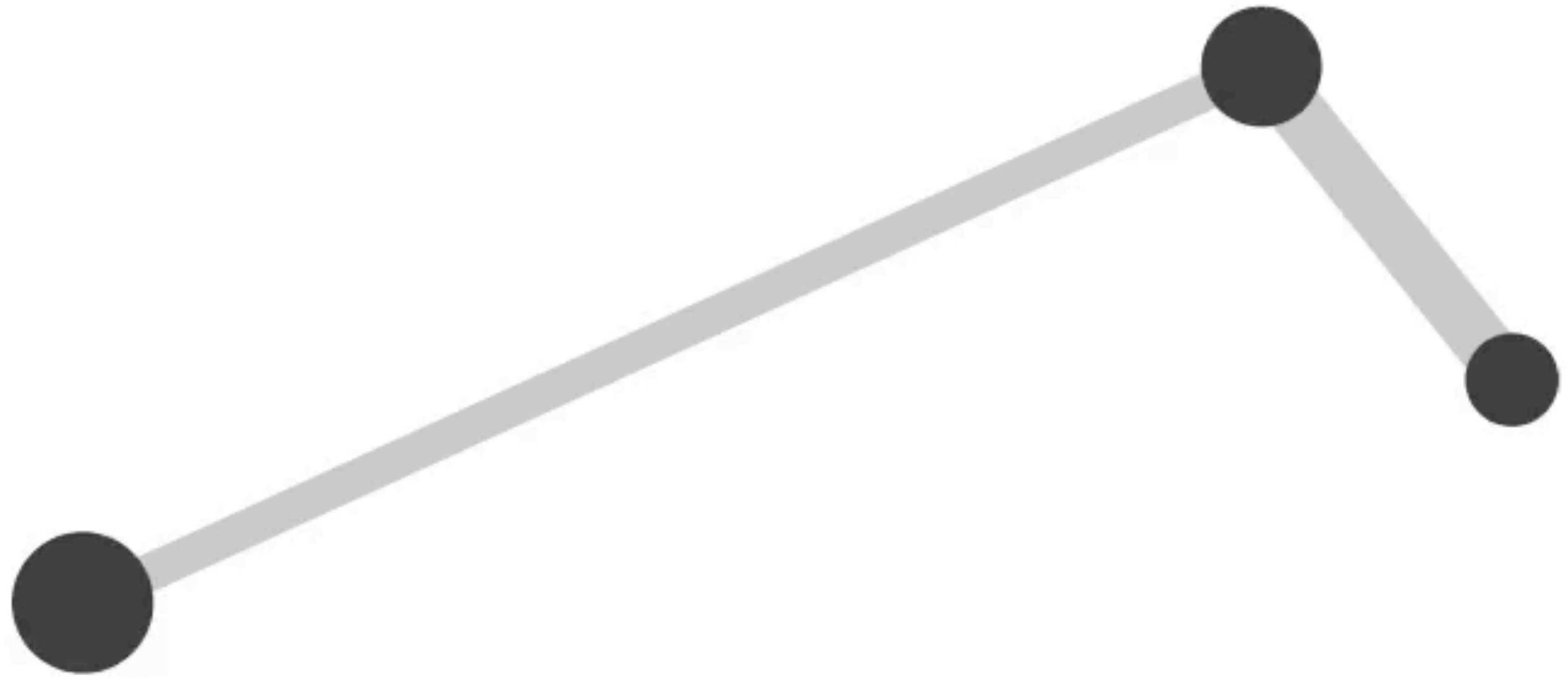
Apr 3: Instagram reaches 30 million users, releases Android version

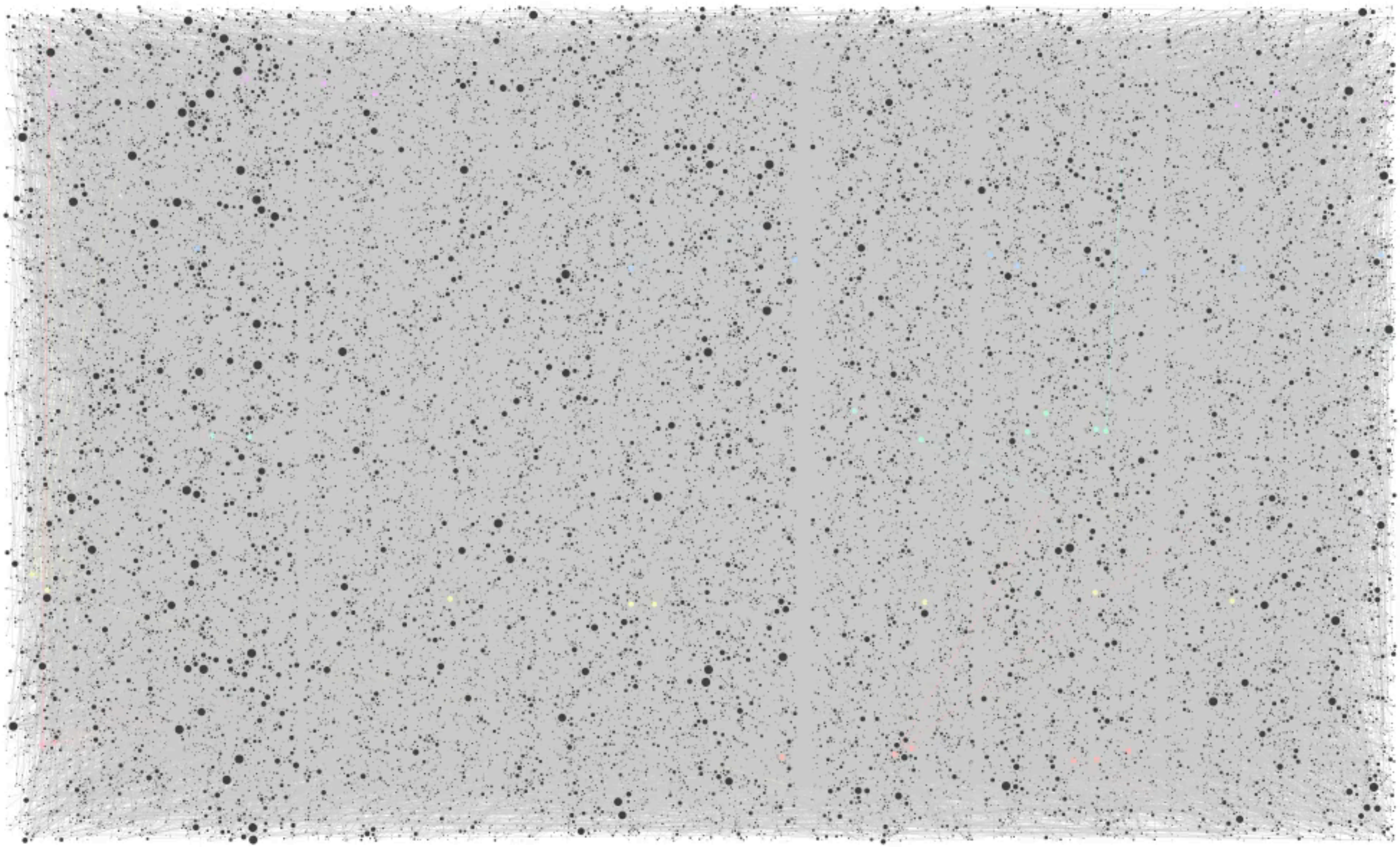
Apr 9: Facebook buys Instagram for $1 billion



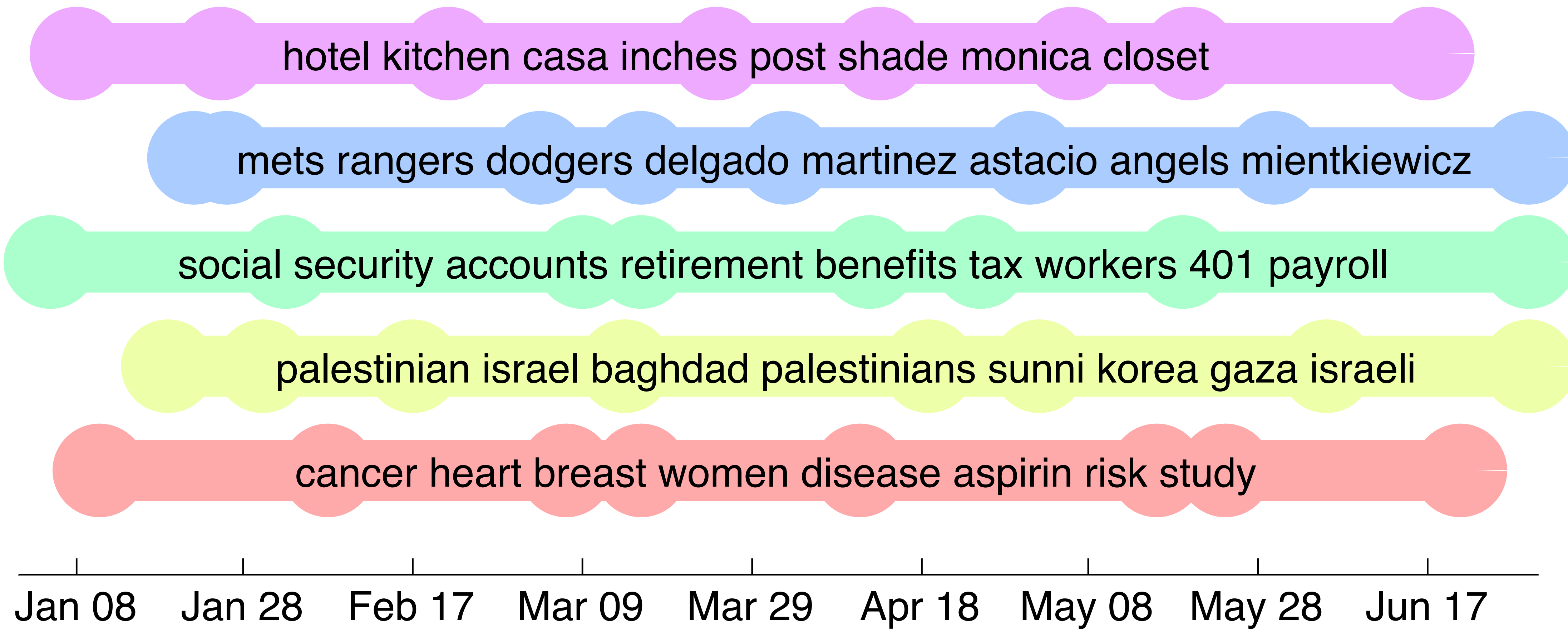
Apr 3: Instagram reaches 30 million users, releases Android version

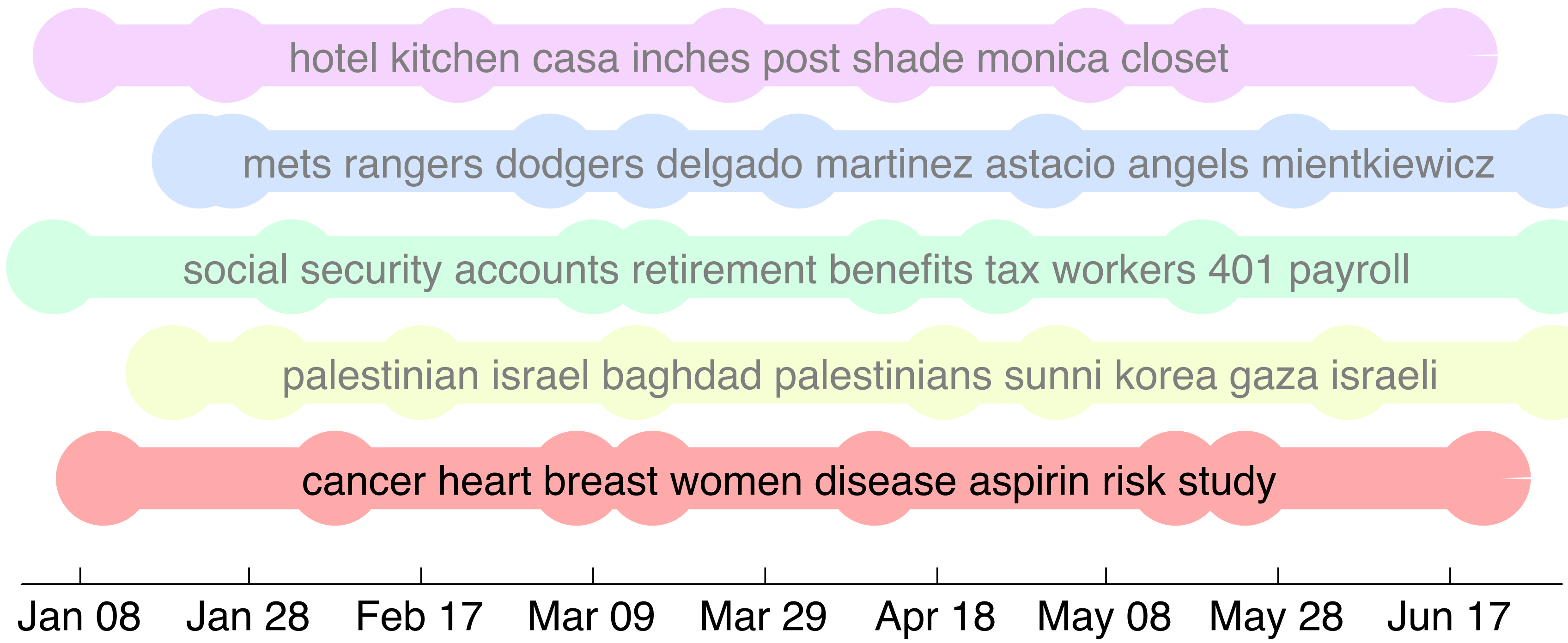

Apr 9: Facebook buys Instagram for $1 billion
Apr 10: Users call for Instagram "exodus"
Apr 3: Instagram reaches 30 million users, releases Android version





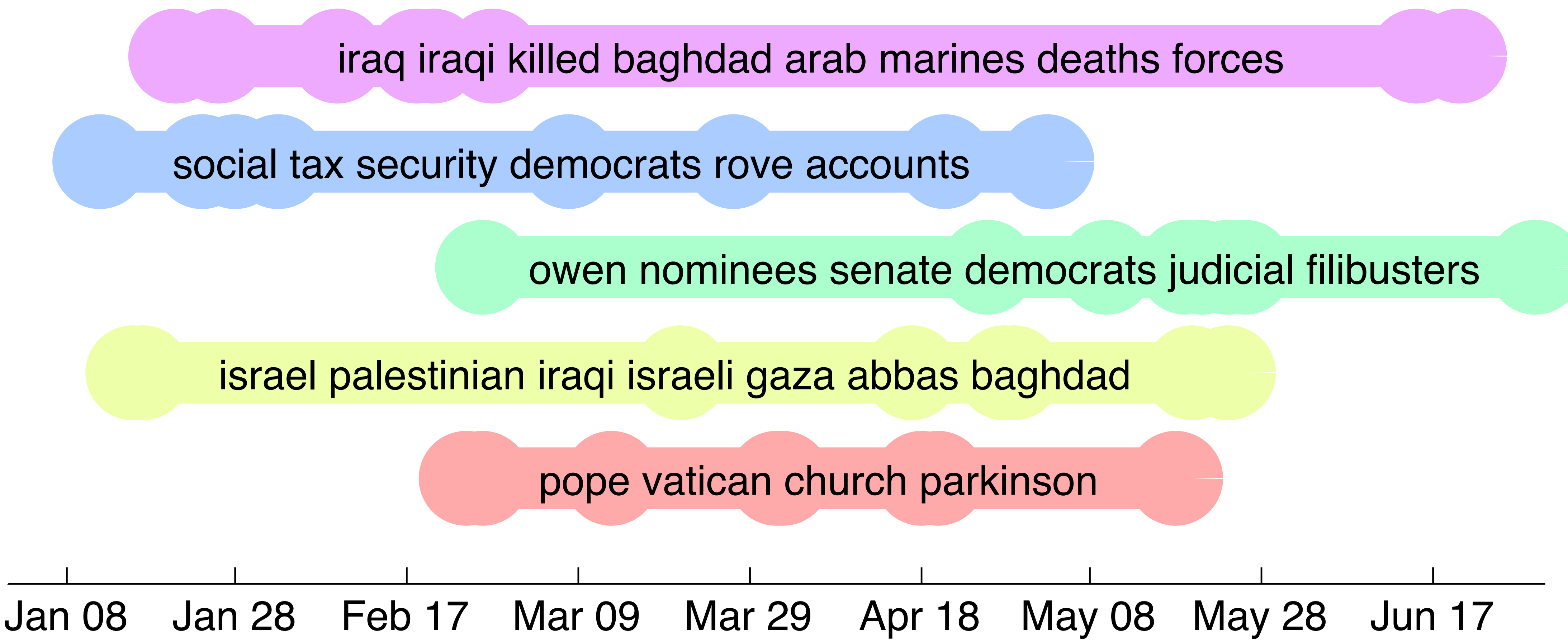
Dynamic topic model

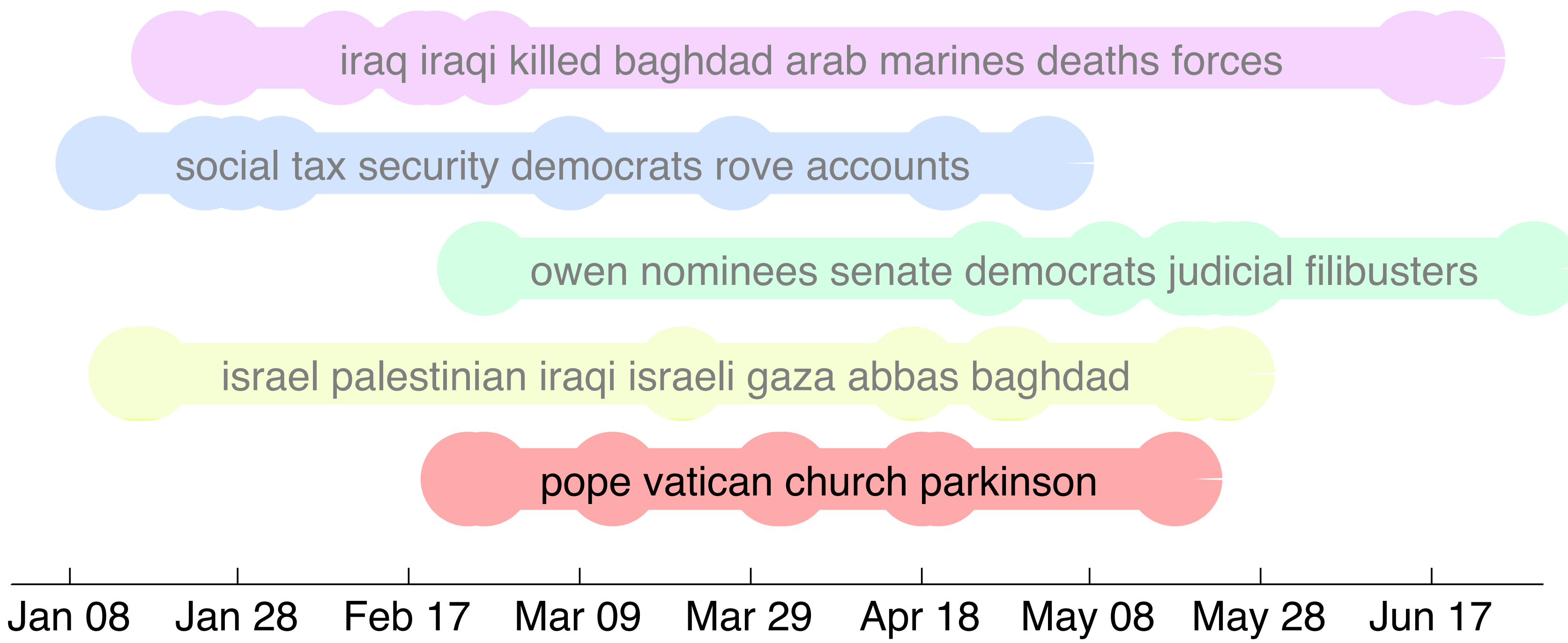




Jan 11: Study Backs Meat, Colon Tumor Link
Feb 07: Patients Still Don't Know How Often Women Get Heart Disease
Mar 07: Aspirin Therapy Benefits Women, but Not the Way It Aids Men
Mar 16: Radiation Therapy Doesn't Increase Heart Disease Risk
Apr 11: Personal Health: Women Struggle for Parity of the Heart
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**
  - amazon mechanical turk to evaluate thread **quality**

Results: Human summaries & ratings

System	
ROUGE-1F	
R-SU4F	
Coherence	
Interlopers	

Results: Human summaries & ratings

System	k-means
ROUGE-1F	16.5
R-SU4F	3.76
Coherence	2.73
Interlopers	0.71

Results: Human summaries & ratings

System	*k*-means	DTM	
ROUGE-1F	16.5	14.7	
R-SU4F	3.76	3.44	
Coherence	2.73	3.19	
Interlopers	0.71	1.10	

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

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**

- 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

Supporting Materials

- ML Foundations & Trends Survey

  http://arxiv.org/abs/1207.6083 (Pre-print, 120 pages)

- Matlab Code:

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