Subspaces alignment and graph convolutional networks performance

Paul Expert
Centre for the Mathematics of Precision Healthcare
Global Digital Health Unit

Yifan Qian, Tom Rieu, Pietro Panzarasa, Mauricio Barahona

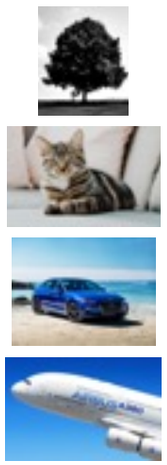
WCNTDA, UKM, 25th September 2019

Imperial College London

EPSRC
Engineering and Physical Sciences Research Council

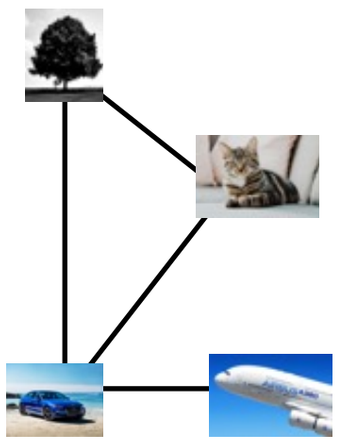
Queen Mary University of London

Outline



- Many types of artificial neural networks architecture: Multi-Layer Perceptrons, Recurrent Neural Networks, Autoencoders, Convolutional Neural Networks, Graph Convolutional Neural Networks.

- They are extremely good at image and speech recognition and classification in general with the increase of computational power and the availability of large dataset (data is the new oil, « Big Data », Alpha XX, etc …).

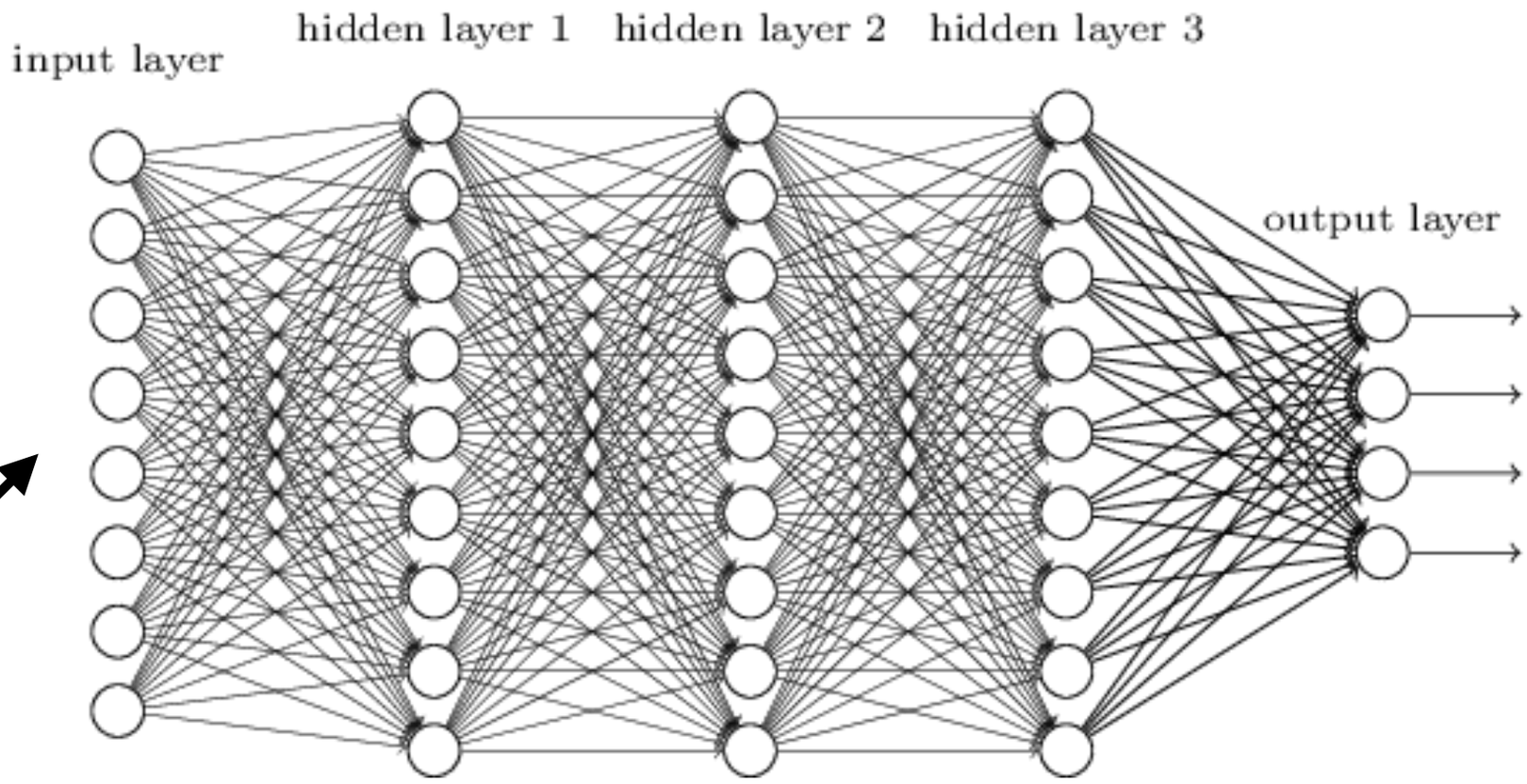


- Is using relational information between objects to be classified always advantageous?
  And can we predict whether is advantageous?

Prototypical goal: classification

Deep neural network

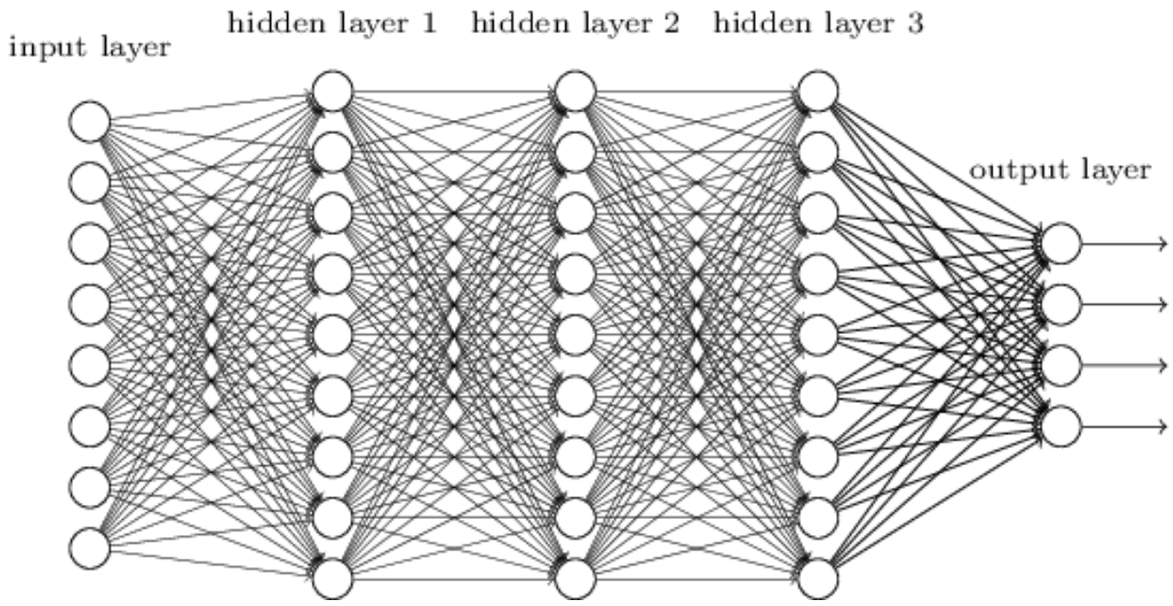
Features



It's a tree!

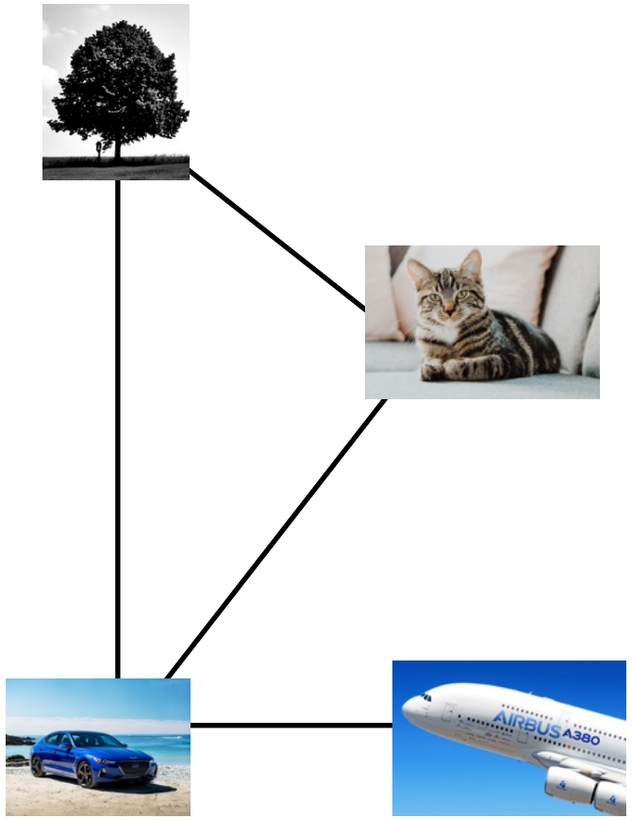
http://playground.tensorflow.org/

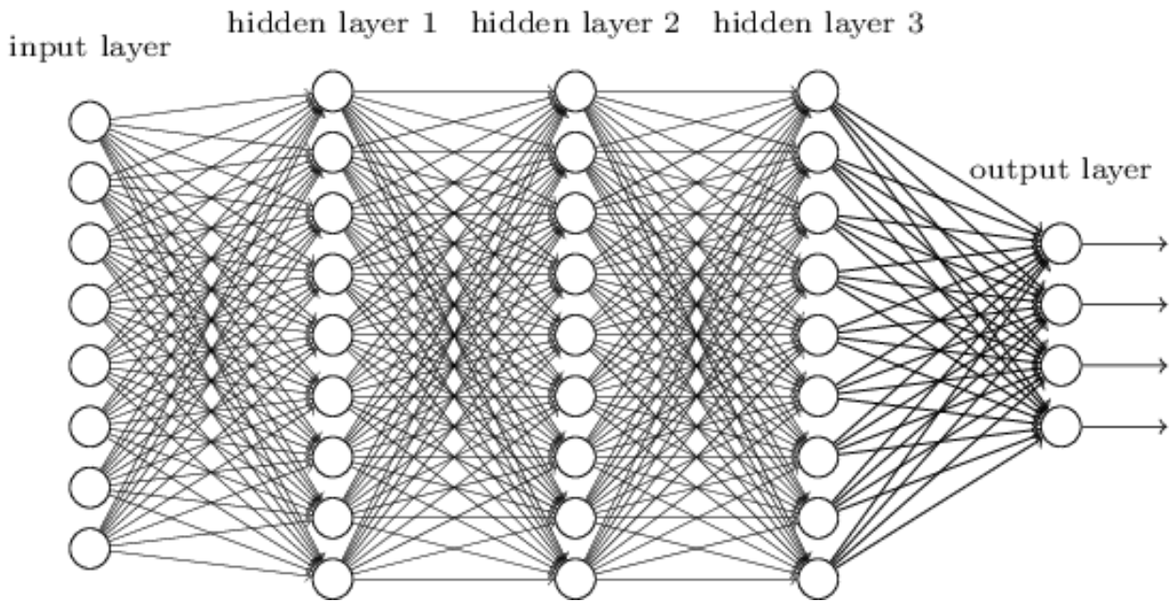
Independent versus interdependent data



Tree
Cat
Car
Plane

Independent versus <u>interdependent</u> data

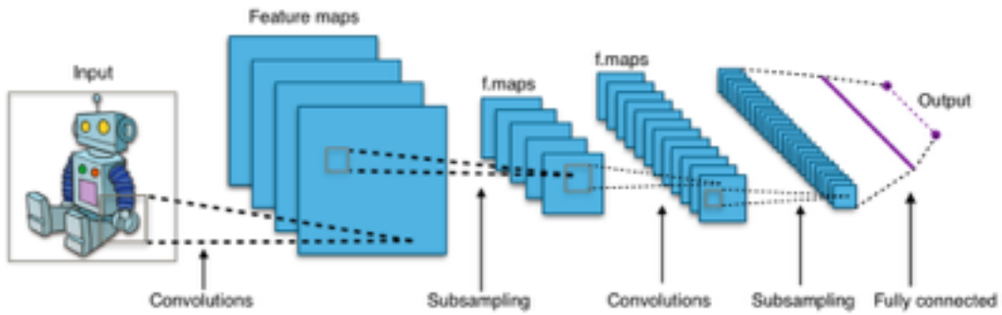




Tree
Cat
Car
Plane

(Very) Short intro about Convolutional Neural Networks

Slide filters across an image.



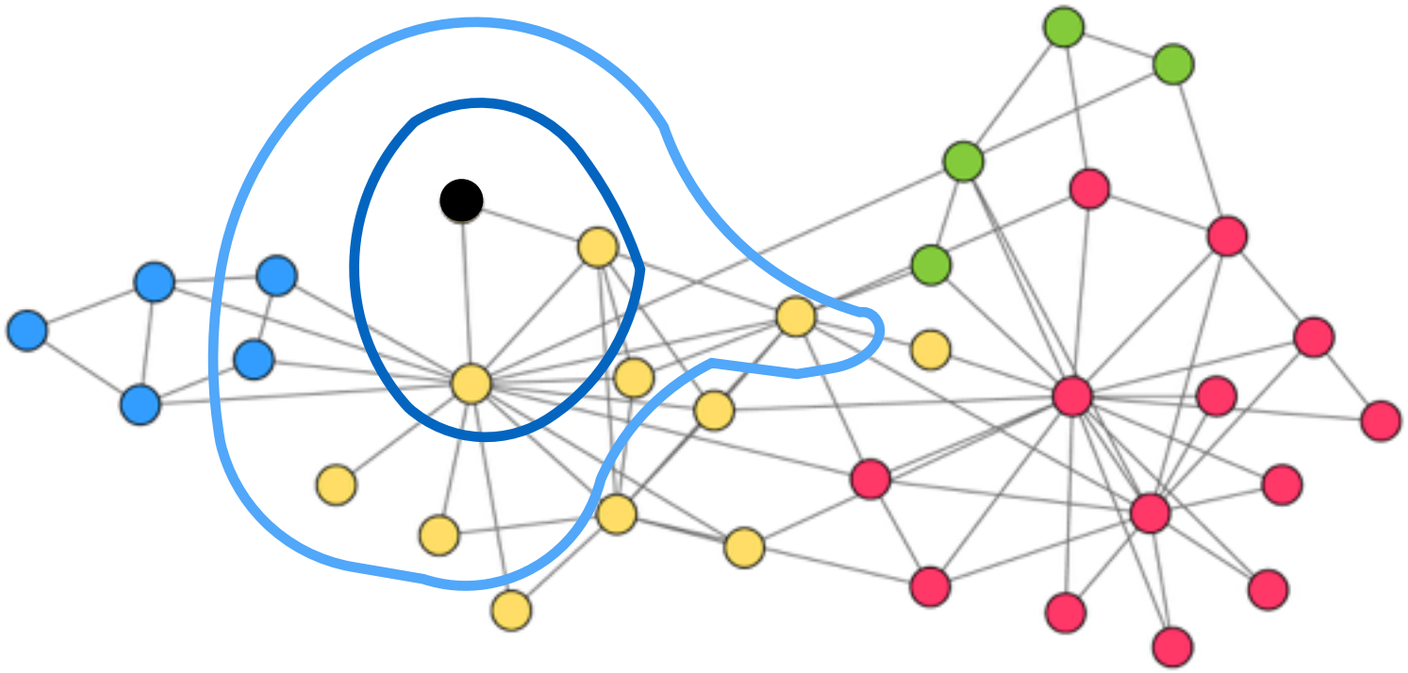
Good for « Euclidean » data, but how to generalise to « non-Euclidean » i.e. graph structured data?

Y. LeCun, Y. Bengio, and G. Hinton, Nature 521, 436 (2015).

https://ujjwalkarn.me/2016/08/11/intuitive-explanation-convnets/

(Very) Short intro about Graph Convolutional Networks

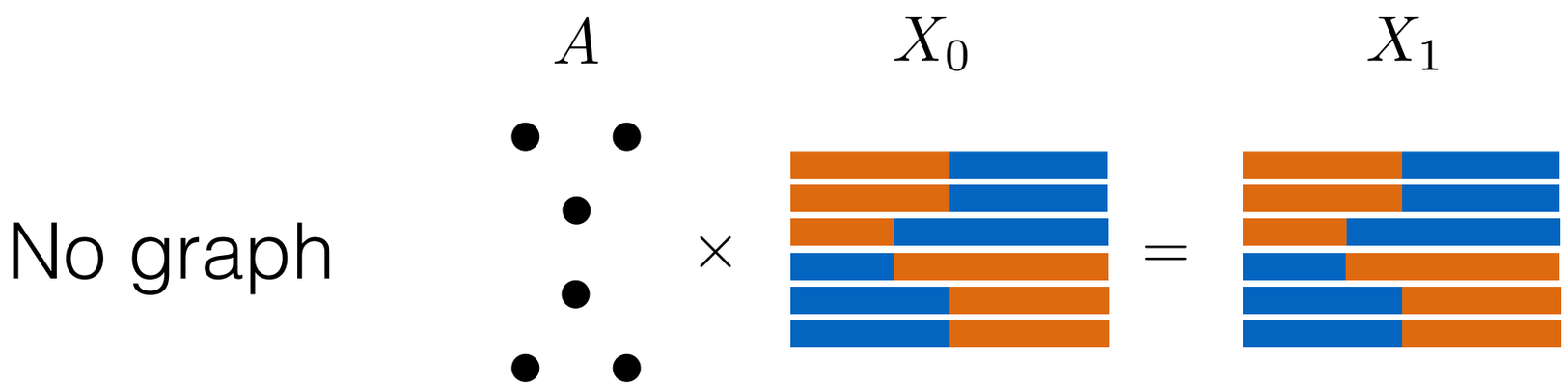
How to average information/find similar nodes in a principled way in a relational graph?



Define a convolution operation in the spectral domain of the graph.

D. I. Shuman *et al.*, *IEEE Sig Proc Mag,* 2013
Bronstein *et al.,* Geometric deep learning: going beyond Euclidean data, *IEEE Sig Proc Mag, 2017*
Wu *et al.*, A Comprehensive Survey on Graph Neural Networks, *arxiv 2019*

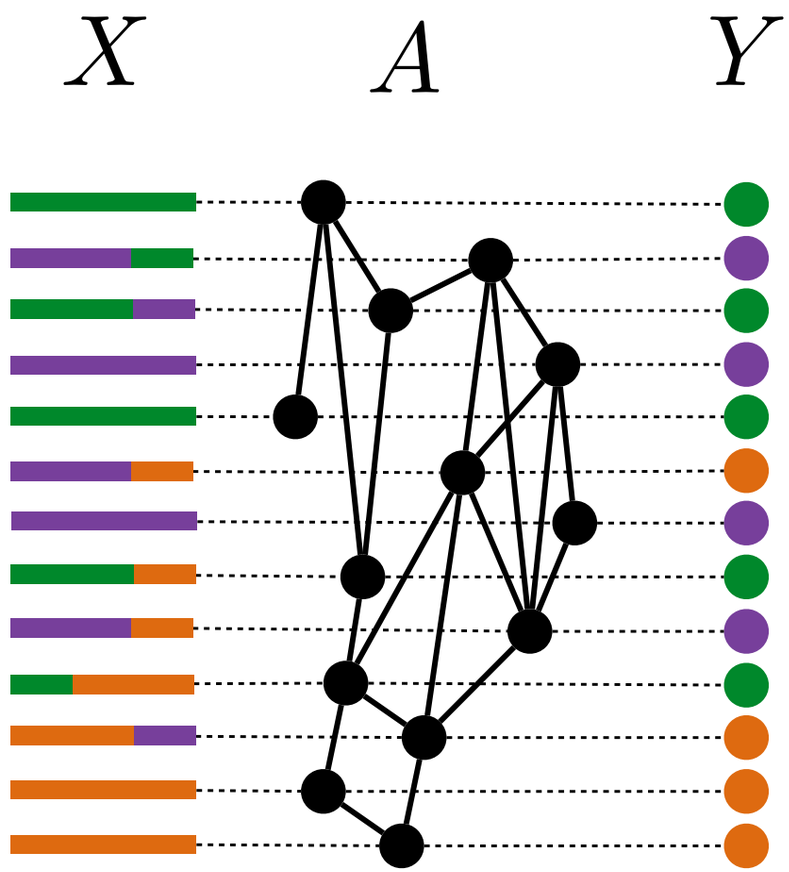
What happens in a « graph convolution »?



Each node is an object with its associated features

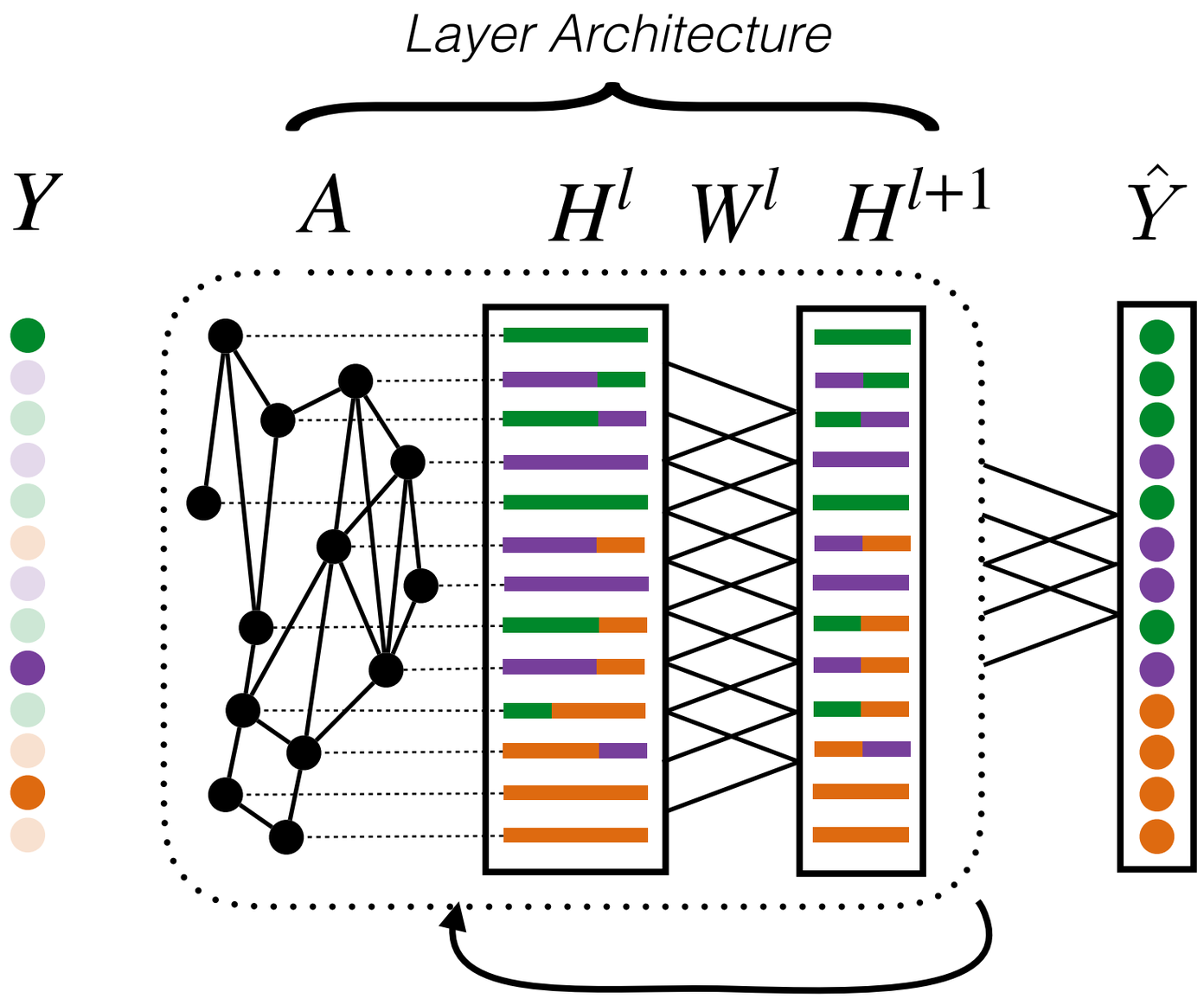
Graph Convolutional Neural Networks

Task: node classification



X: features associated to the nodes, *F*
A: relational graph between nodes, *N*
Y: ground truth, *C*

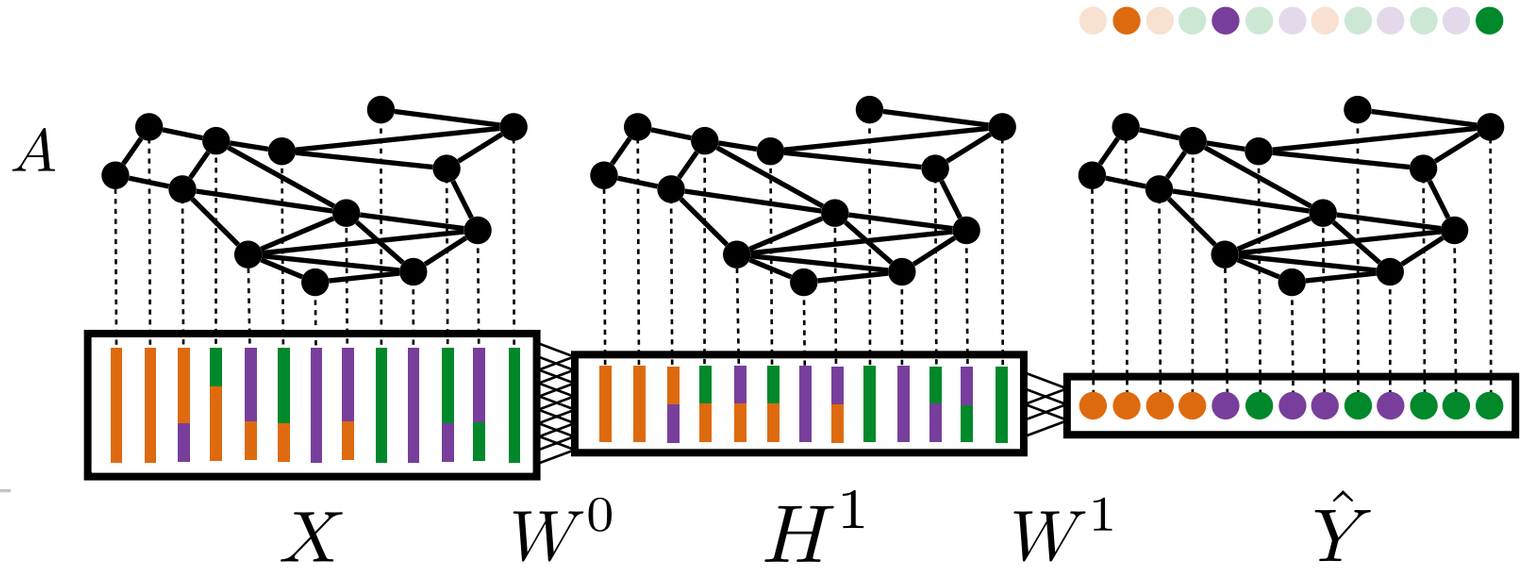
Typical GCN architecture

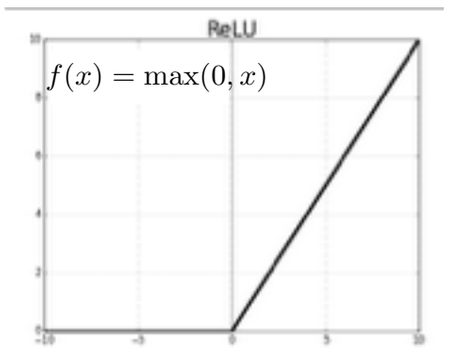


Transductive semi-supervised learning

Kipf & Welling, « Semi-supervised classification with graph convolutional networks » *ICLR 2017*

Simplest model





ReLU

softmax

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{j=1}^{C} \exp(x_j)}$$

$$\begin{aligned}\hat{Y} &= f(X, A) \\ &= \text{softmax}(\hat{A} H^1 W^1) \\ &= \text{softmax}(\hat{A}\ \text{ReLU}(\hat{A} X W^0)\ W^1)\end{aligned}$$

Convolution

Kipf & Welling, « Semi-supervised classification with graph convolutional networks » *ICLR 2017*

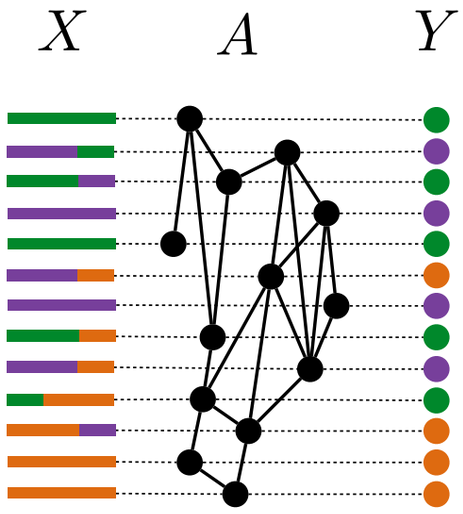
Seems awesome !!

Method	**Citeseer**	**Cora**	**Pubmed**	**NELL**
ManiReg [3]	60.1	59.5	70.7	21.8
SemiEmb [28]	59.6	59.0	71.1	26.7
LP [32]	45.3	68.0	63.0	26.5
DeepWalk [22]	43.2	67.2	65.3	58.1
ICA [18]	69.1	75.1	73.9	23.1
Planetoid* [29]	64.7 (26s)	75.7 (13s)	77.2 (25s)	61.9 (185s)
GCN (this paper)	**70.3** (7s)	**81.5** (4s)	**79.0** (38s)	**66.0** (48s)
GCN (rand. splits)	67.9 ± 0.5	80.1 ± 0.5	78.9 ± 0.7	58.4 ± 1.7

Accuracy: number of correct predictions out of all predictions

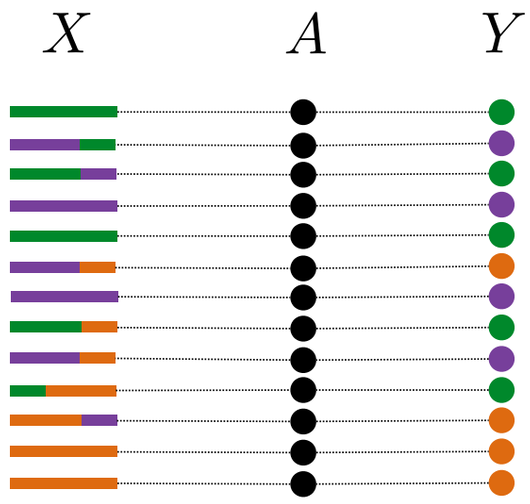
Kipf & Welling, « Semi-supervised classification with graph convolutional networks » *ICLR 2017*

But is it really?



GCN

VS



MLP

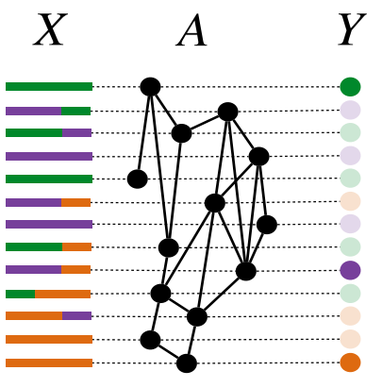
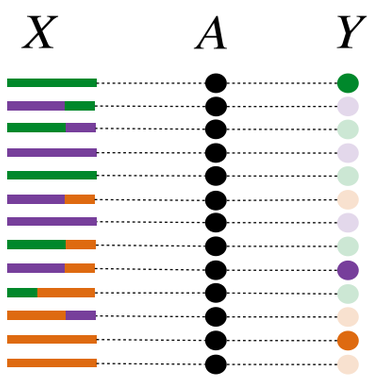
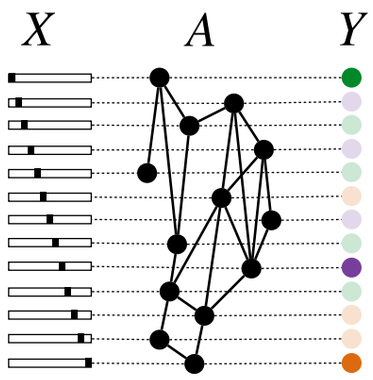
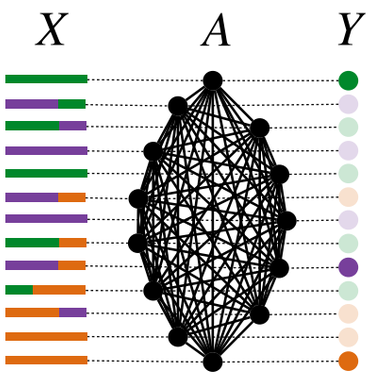
Data sets	GCN	MLP
Constructive example	**0.932** ± **0.006**	0.416 ± 0.010
CORA	**0.811** ± **0.005**	0.548 ± 0.014
AMiner	**0.748** ± **0.005**	0.547 ± 0.013
Wikipedia	0.392 ± 0.010	**0.450** ± **0.007**

Datasets

- Toy model with planted communities (SBM)
- CORA: benchmark citation network, from time immemorial
- Aminer: computer science citation network
- Wikipedia: home made, 12 classes

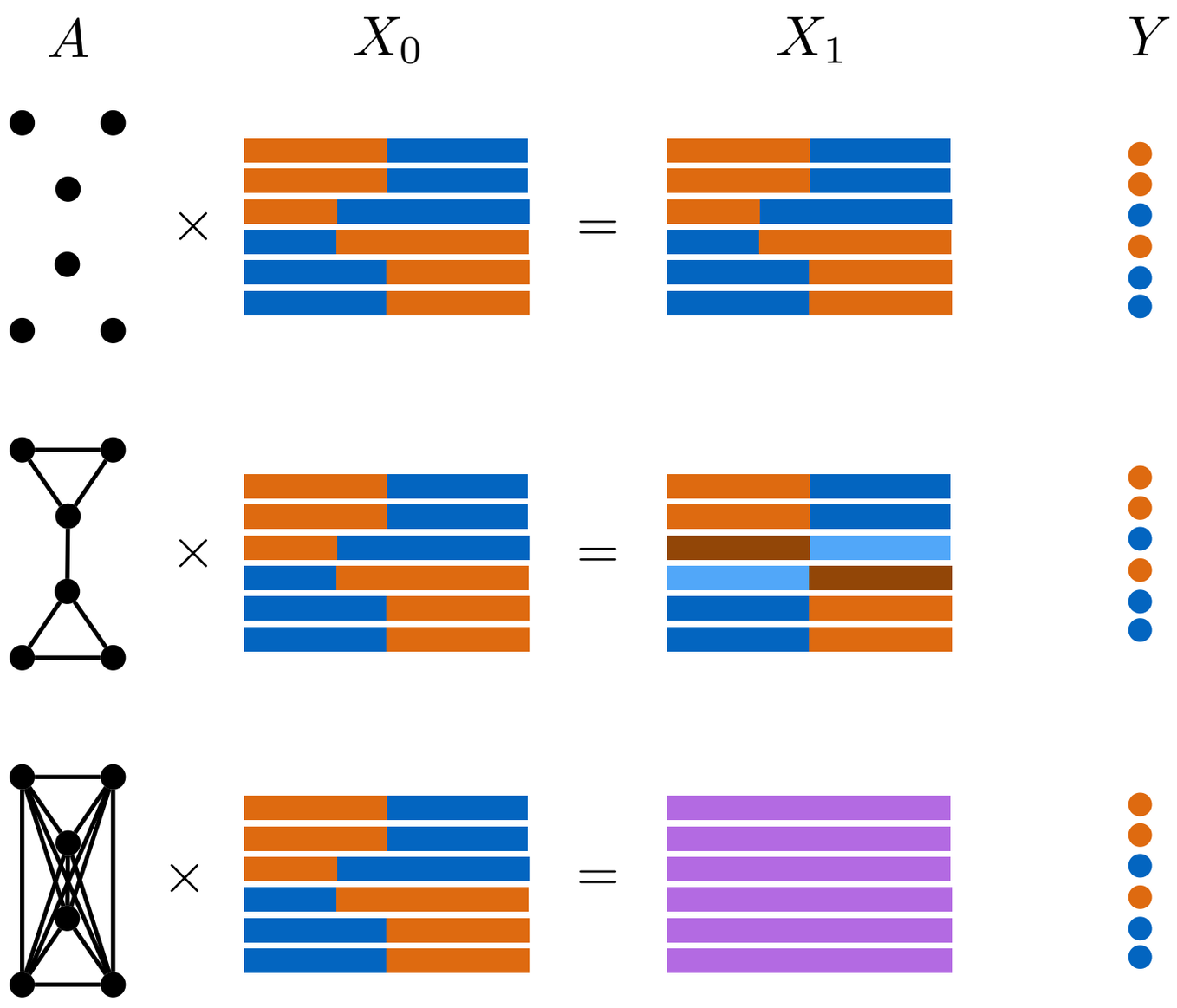
Data sets	Nodes	Edges	Features	Classes
Constructive example	$1,000$	$6,541$	500	10
CORA	$2,485$	$5,069$	$1,433$	7
AMiner	$2,072$	$4,299$	500	7
Wikipedia	$20,525$	$215,056$	100	12
Wikipedia I	$2,414$	$8,163$	100	5
Wikipedia II	$1,858$	$8,444$	100	5

MLP is just one limit case

	GCN	Limit cases		
		MLP (no graph)	No features	Complete graph
				
Data sets				
Constructive example	**0.932 ± 0.006**	0.416 ± 0.010	0.764 ± 0.009	0.100 ± 0.003
CORA	**0.811 ± 0.005**	0.548 ± 0.014	0.691 ± 0.006	0.121 ± 0.066
AMiner	**0.748 ± 0.005**	0.547 ± 0.013	0.591 ± 0.006	0.123 ± 0.045
Wikipedia	0.392 ± 0.010	**0.450 ± 0.007**	0.254 ± 0.037	O.O.M.
Wikipedia I	**0.861 ± 0.006**	0.796 ± 0.005	0.824 ± 0.003	0.163 ± 0.135
Wikipedia II	0.566 ± 0.021	**0.659 ± 0.011**	0.347 ± 0.012	0.155 ± 0.176

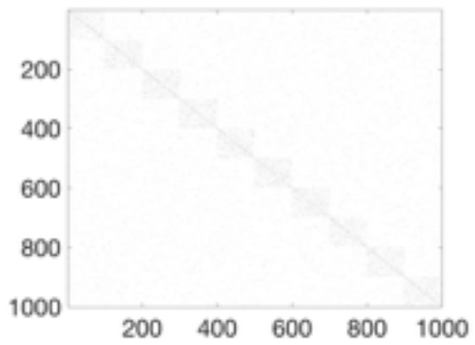
Could the performance be related to the congruence of information in the ingredients?

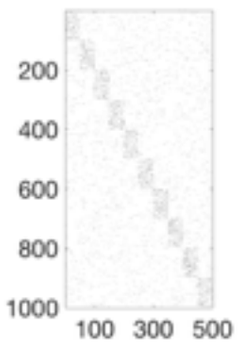
It gets more convoluted with the ground truth.

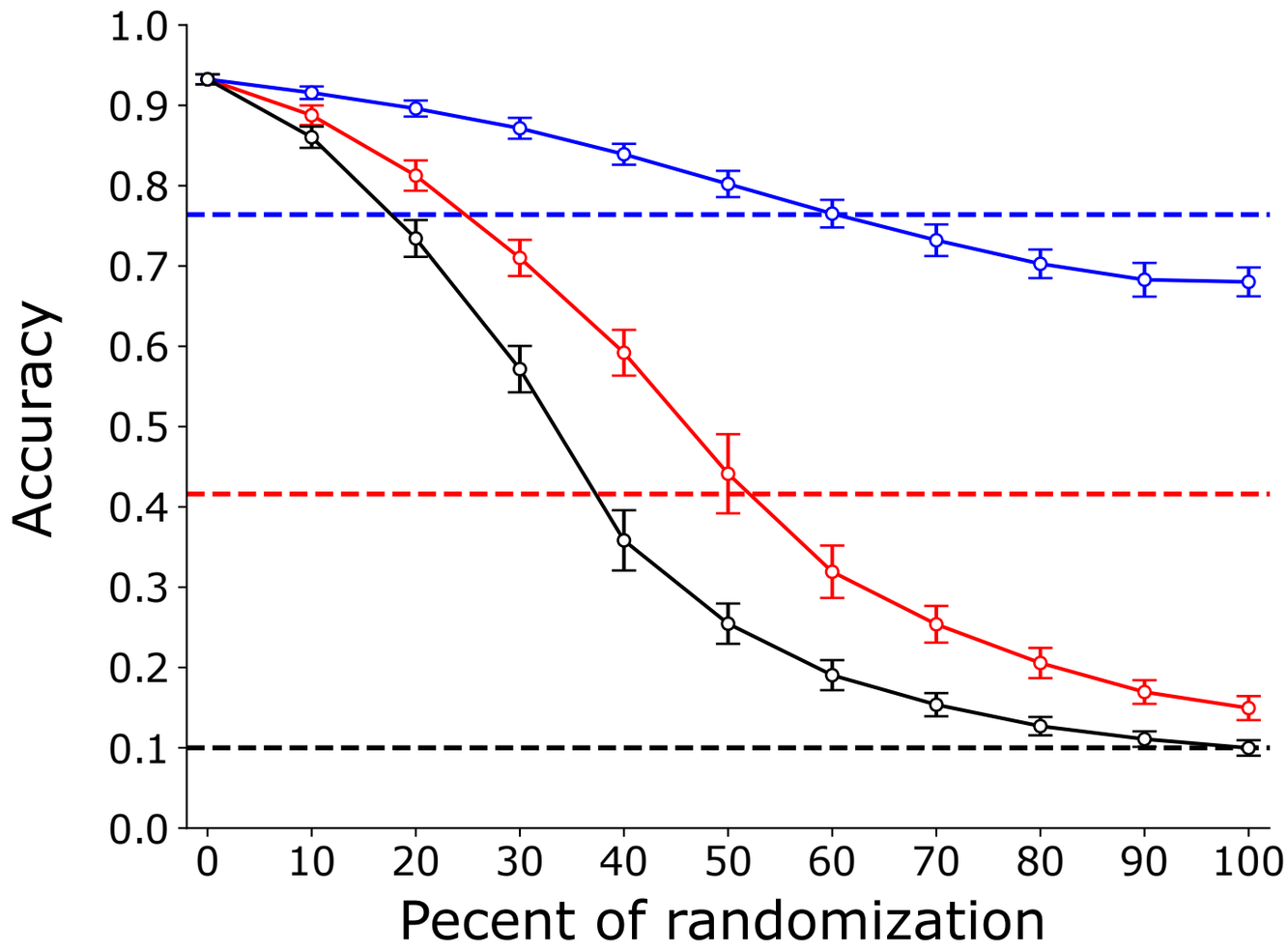


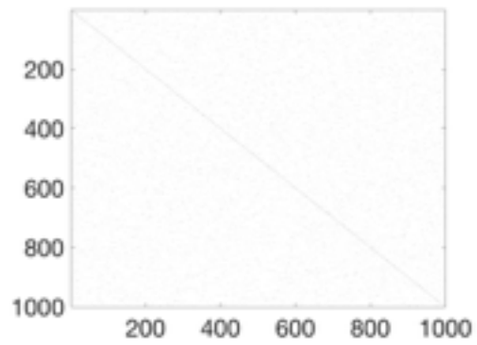
How aligned are *A*, *X* and *Y*?

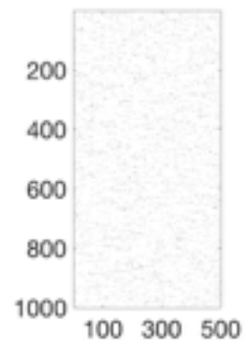
Misalignment by randomisation
Toy model





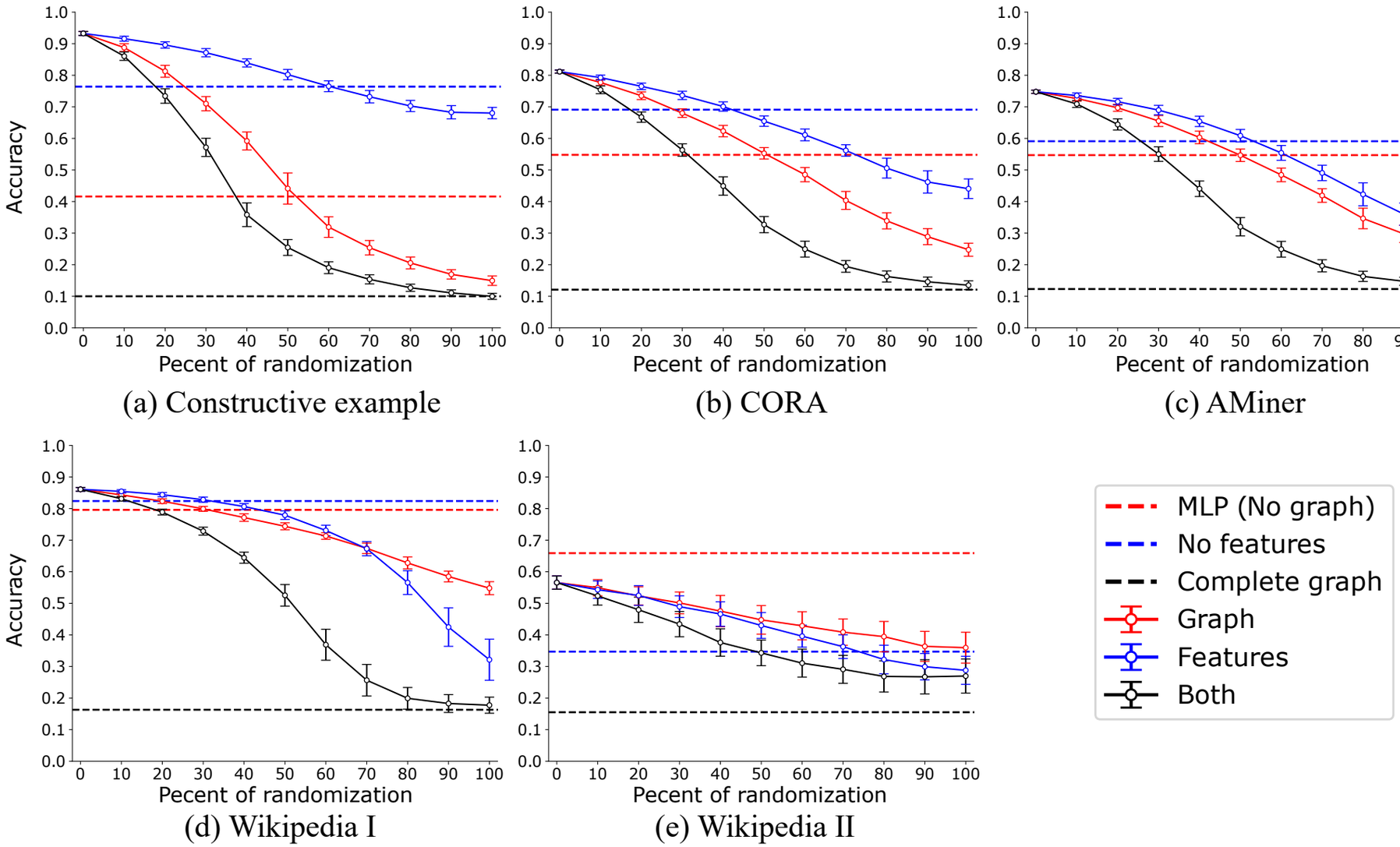






Features: label switching
Graph: configuration model

Randomisation - all datasets



As randomisation increases, information/alignment is lost and performance decreases.

Q: How to measure the alignment of subspaces?

A: Principal angles between $\mathcal{A}, \mathcal{B} \in \mathcal{R}^N$

$$\dim(\mathcal{A}) = \ell \leq \dim(\mathcal{B}) = p \leq N$$

Ye and Lim, « Schubert varieties and distances between subspaces of different dimensions », SIAM J. on Matrix Anal. and Appl., 2016

Q: How to measure alignment?

A: Principal angles between $\mathcal{A}, \mathcal{B} \in \mathcal{R}^N$

$$\dim(\mathcal{A}) = \ell \leq \dim(\mathcal{B}) = p \leq N$$

$$\theta_1 = \min_{a_1 \in \mathcal{A}, b_1 \in \mathcal{B}} \arccos\left(\frac{|a_1^T b_1|}{\|a_1\| \|b_1\|}\right),$$

$$\theta_j = \min_{\substack{a_j \in \mathcal{A}, b_j \in \mathcal{B} \\ a_j \perp a_1, \ldots, a_{j-1} \\ b_j \perp b_1, \ldots, b_{j-1}}} \arccos\left(\frac{|a_j^T b_j|}{\|a_j\| \|b_j\|}\right), \quad j = 2, \ldots, \ell.$$

By how much does $\mathcal{A}$ need to be rotated to « maximally overlap » with $\mathcal{B}$.

Björck and Golub, « Numerical methods for computing angles between linear subspaces », Mathematics of Computation, 1973.

Chordal distance

Any distance between subspaces has to be a function of the principal angles.

$$d(\mathcal{A}, \mathcal{B}) = \sqrt{\sum_{j=1}^{\ell} \sin^2 \theta_j}$$

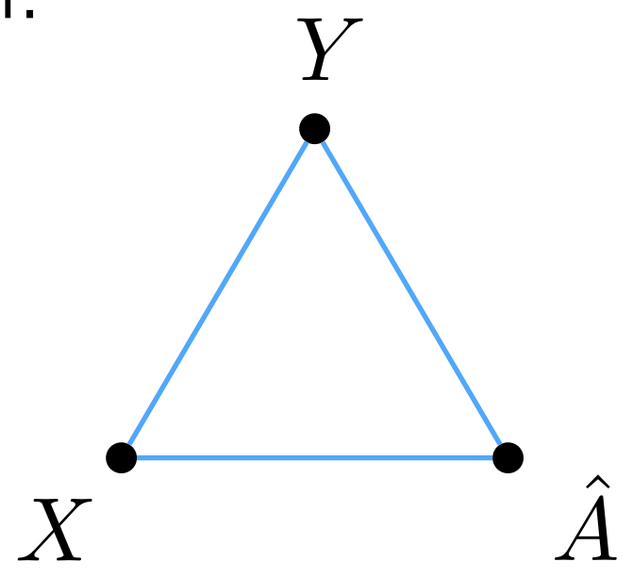
Ye and Lim, « Schubert varieties and distances between subspaces of different dimensions », SIAM J. on Matrix Anal. and Appl., 2016

Pairwise distances

$$D = \begin{bmatrix} 0 & d(X,\hat{A}) & d(X,Y) \\ d(X,\hat{A}) & 0 & d(\hat{A},Y) \\ d(X,Y) & d(\hat{A},Y) & 0 \end{bmatrix}$$

That we summarise with:

$$\|D\|_F = \sqrt{\sum_{i=1}^{3}\sum_{j=1}^{3} D_{ij}^2}$$



The need for dimensionality reduction

For the distance to be non trivial, the dimensions of the subspaces need to be strictly smaller than *N*.

$$\dim(\mathcal{A}) = \ell \leq \dim(\mathcal{B}) = p \boxed{<} N$$

We have:

$$\dim(X) = F < N$$

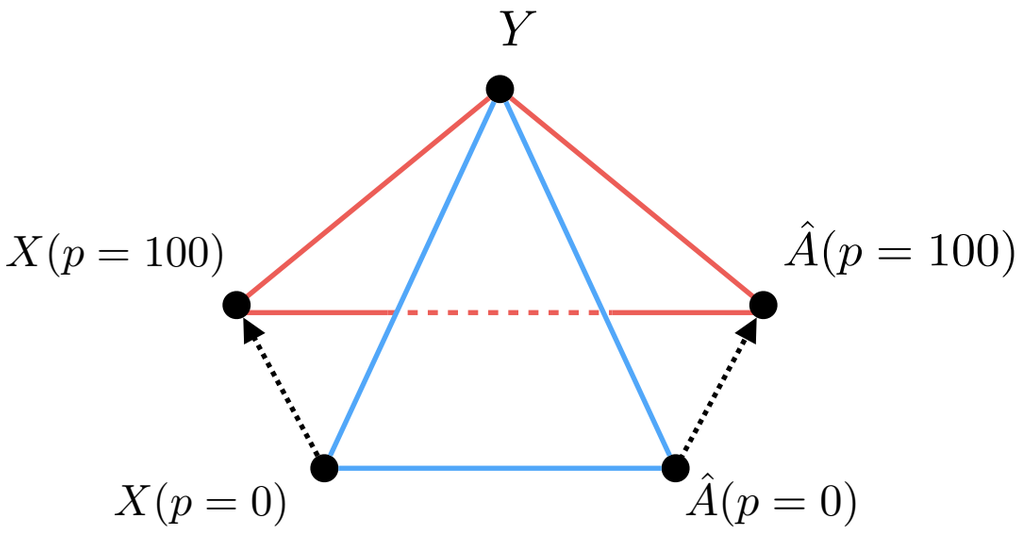
$$\dim(Y) = C < N$$

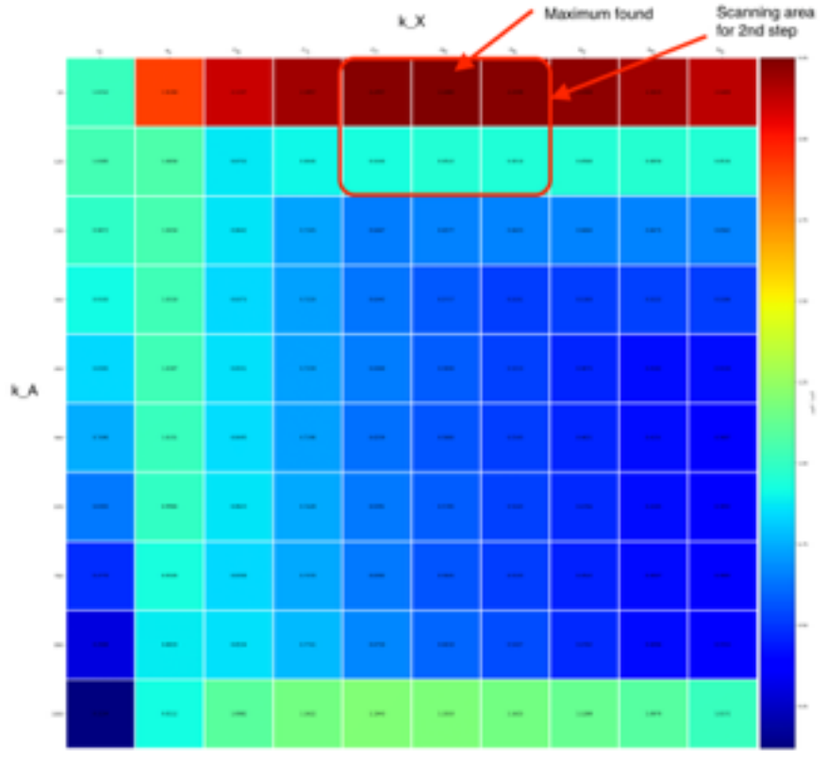
$$\underline{\dim(\hat{A}) = N}$$

How much information can we destroy?

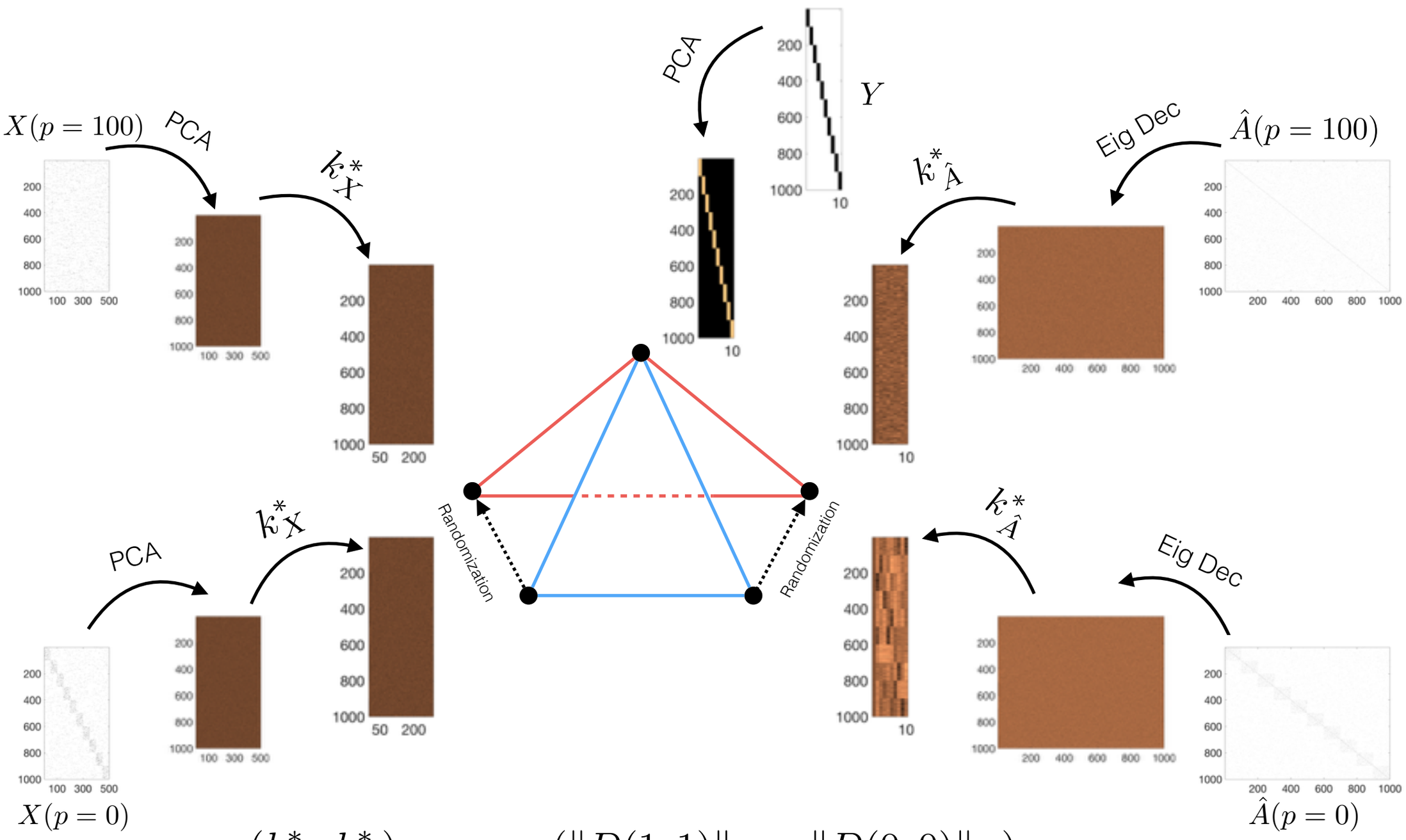
Maximum alignment loss between original and fully randomised data:

$$(k_X^*, k_{\hat{A}}^*) = \max_{k_X, k_{\hat{A}}} \left(\|D(1,1)\|_F - \|D(0,0)\|_F \right)$$



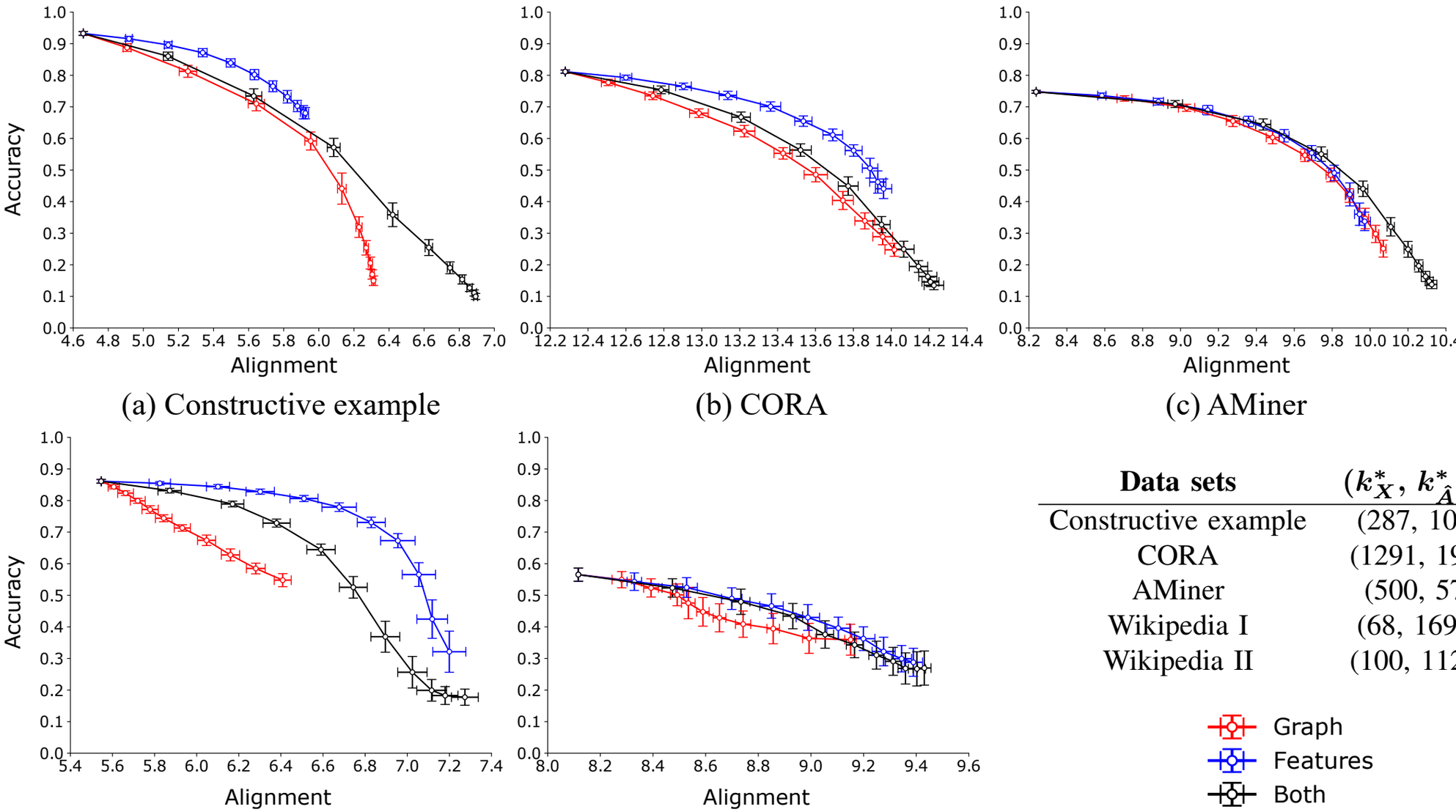


Summary of the pipeline



$$(k^*_X, k^*_{\hat{A}}) = \max_{k_X, k_{\hat{A}}} \left(\|D(1,1)\|_F - \|D(0,0)\|_F \right)$$

Performance versus alignment



(a) Constructive example (b) CORA (c) AMiner

(d) Wikipedia I (e) Wikipedia II

Data sets	$(k^*_X, k^*_{\hat{A}}, k^*_Y)$
Constructive example	(287, 10, 10)
CORA	(1291, 190, 7)
AMiner	(500, 57, 7)
Wikipedia I	(68, 1699, 5)
Wikipedia II	(100, 1125, 5)

Graph
Features
Both

Take home messages

- A degree of alignment (and modular structure) is needed for GCN to perform well.
- Principled way to reduce the dimensionality of the subspaces to have a meaningful notion of distance/alignment.
- General applicability to data alignment measurement.

Preprint: https://arxiv.org/abs/1905.12921

Code: https://github.com/haczqyf/gcn-data-alignment

Yifan Qian, QMUL

Mauricio Barahona, ICL

Tom Rieu, ICL

Pietro Panzarasa, QMUL

Thank you for listening, questions?

Preprint: https://arxiv.org/abs/1905.12921

Code: https://github.com/haczqyf/gcn-data-alignment

Imperial College London

EPSRC Engineering and Physical Sciences Research Council

Queen Mary University of London