Deep Learning 1

2025-2026 – Pascal Mettes

Lecture 7

Graph Deep Learning

Previous lecture

Lecture	Title	Lecture	Title
1	Intro and history of deep learning	2	AutoDiff
3	Deep learning optimization I	4	Deep learning optimization II
5	Convolutional deep learning	6	Attention-based deep learning
7	Graph deep learning	8	From supervised to unsupervised deep learning
9	Multi-modal deep learning	10	Generative deep learning
11	What doesn't work in deep learning	12	Non-Euclidean deep learning
13	Q&A	14	Deep learning for videos

I have to leave a 10:10 on the dot to be able to make it to a PhD committee at 11:00 in the city center.

This lecture

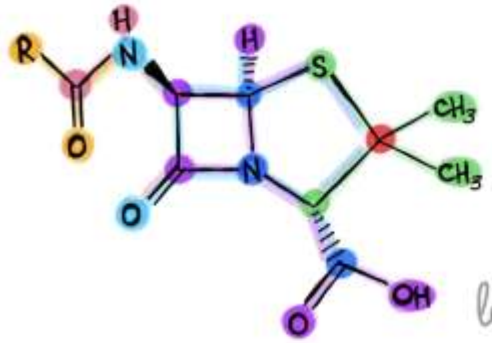
Graphs

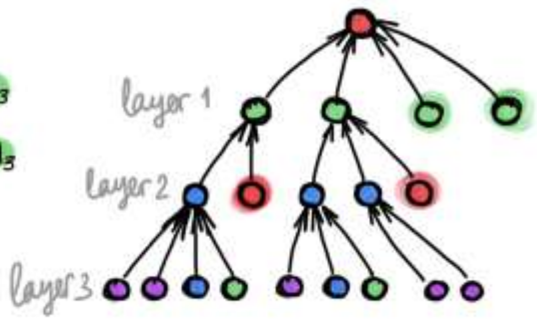
Graph convolutions

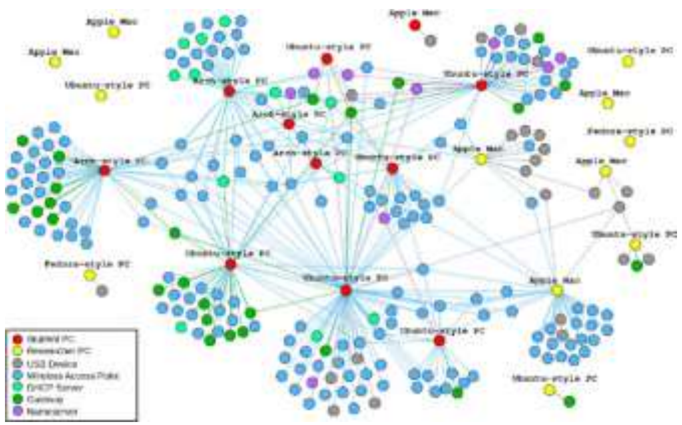
Graph attention

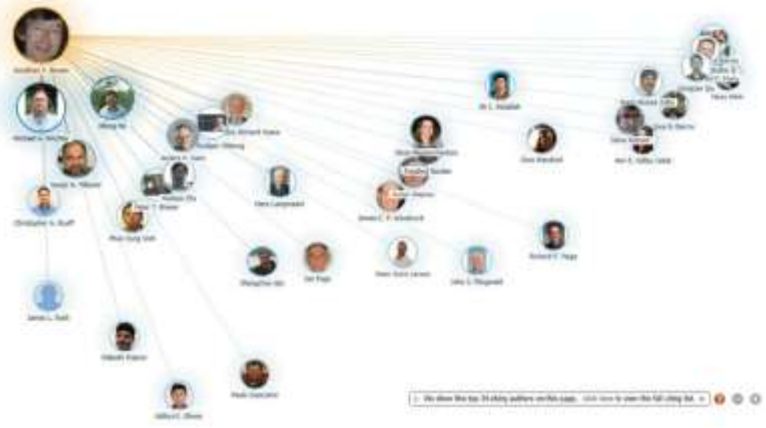
Graph applications

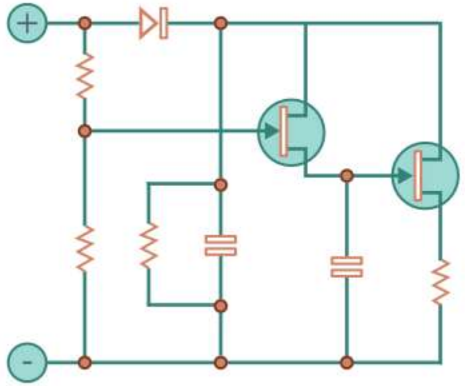
Graphs, more common than you think



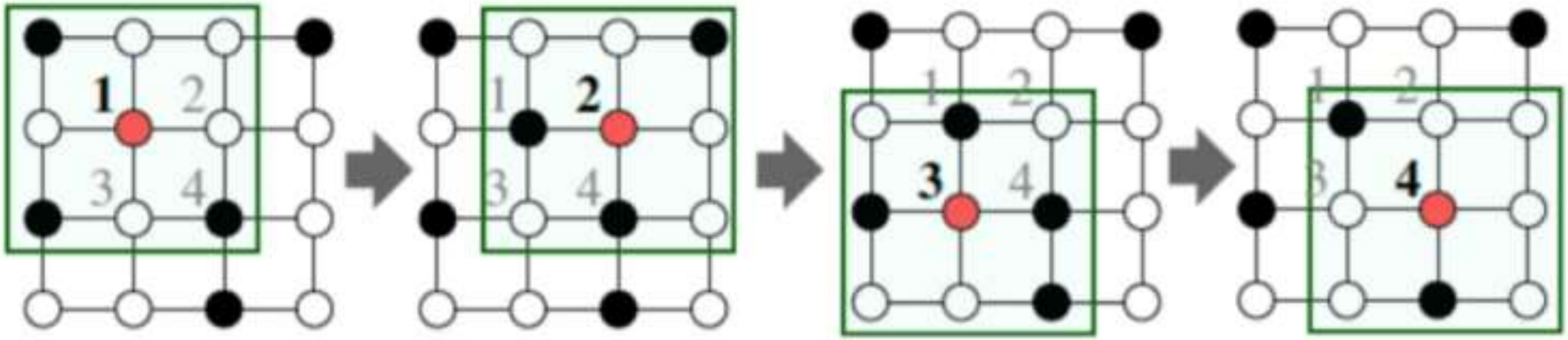




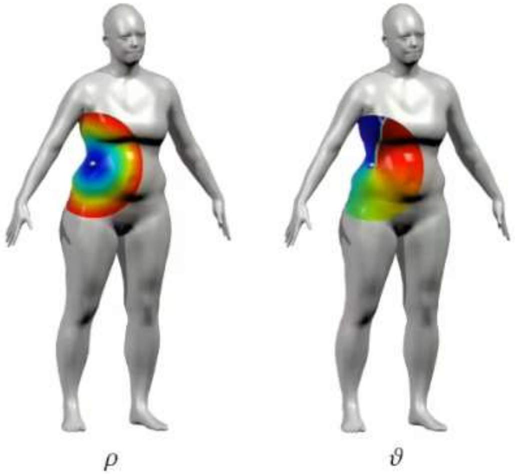


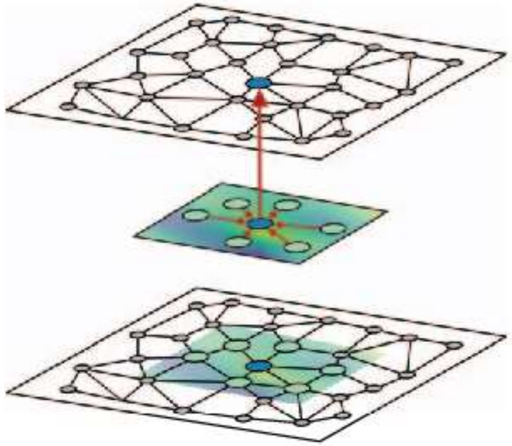


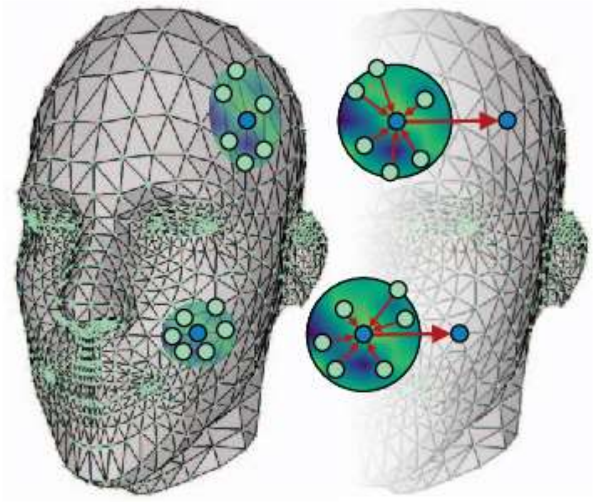
Many structures are special cases of graphs

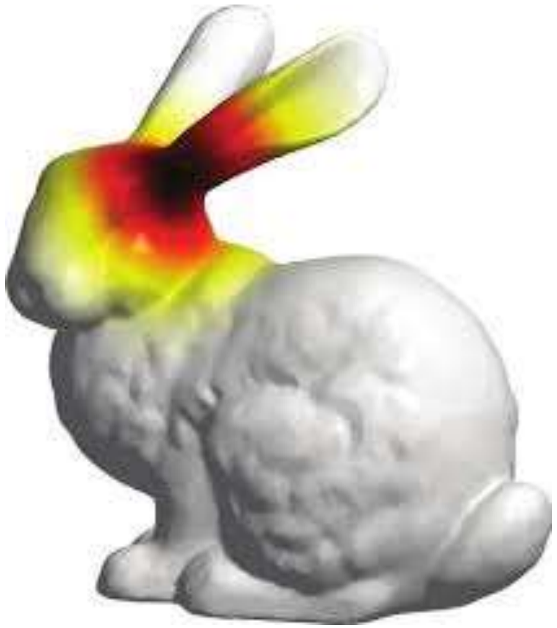


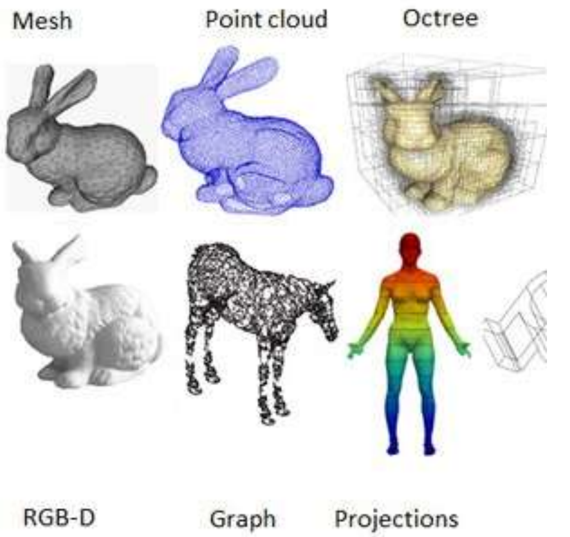
Graphs as geometry





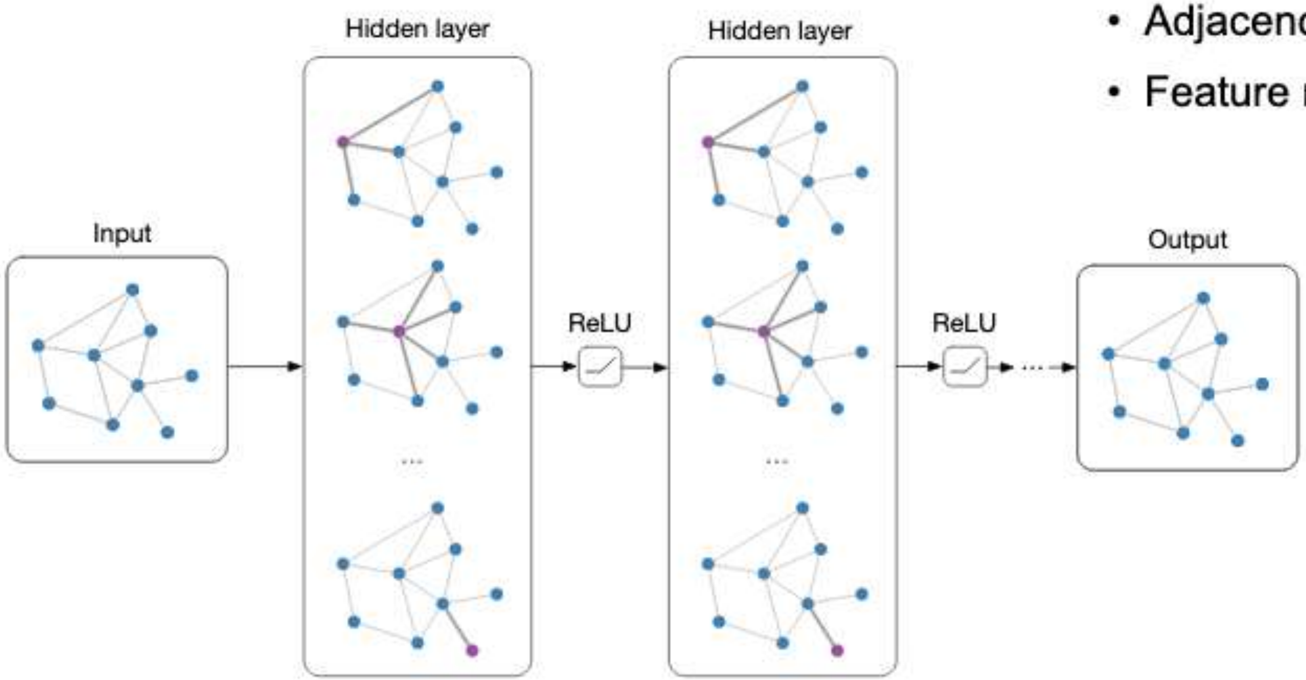






What are graph networks?

The bigger picture:



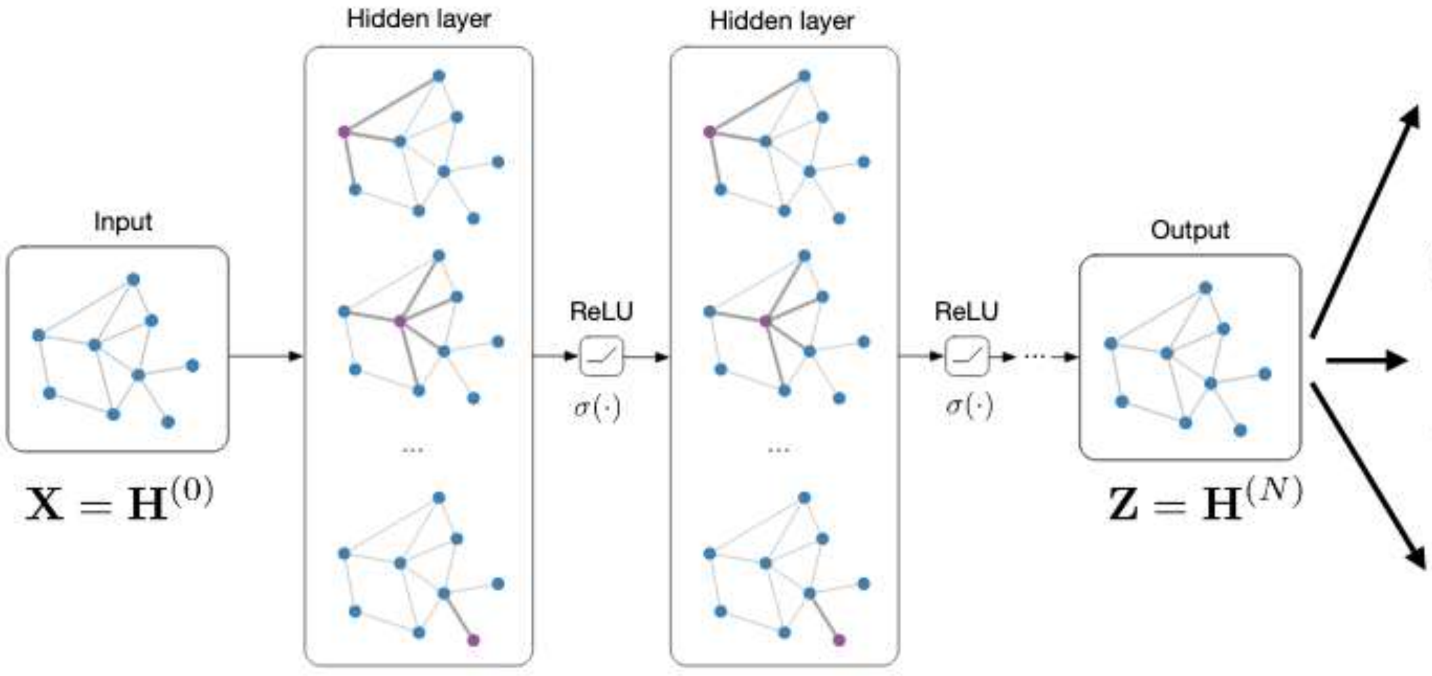
Notation: $\mathcal{G} = (\mathbf{A}, \mathbf{X})$

- Adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$
- Feature matrix $\mathbf{X} \in \mathbb{R}^{N \times F}$

Main idea: Pass messages between pairs of nodes & agglomerate

Graph networks

Input: Feature matrix $\mathbf{X} \in \mathbb{R}^{N \times E}$, preprocessed adjacency matrix $\hat{\mathbf{A}}$



$$\mathbf{H}^{(l+1)} = \sigma\left(\hat{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right)$$

Node classification:

$$\mathrm{softmax}(\mathbf{z_n})$$

e.g. Kipf & Welling (ICLR 2017)

Graph classification:

$$\mathrm{softmax}(\textstyle\sum_n \mathbf{z_n})$$

e.g. Duvenaud et al. (NIPS 2015)

Link prediction:

$$p(A_{ij}) = \sigma(\mathbf{z_i^T z_j})$$

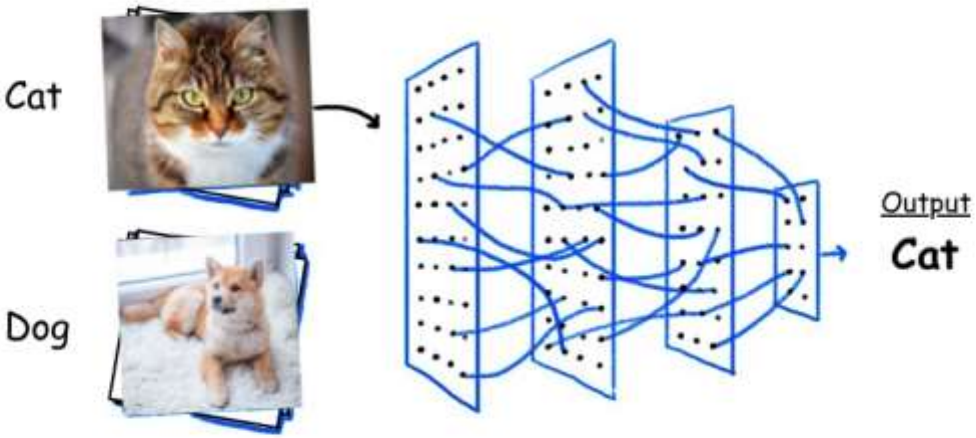
Kipf & Welling (NIPS BDL 2016)

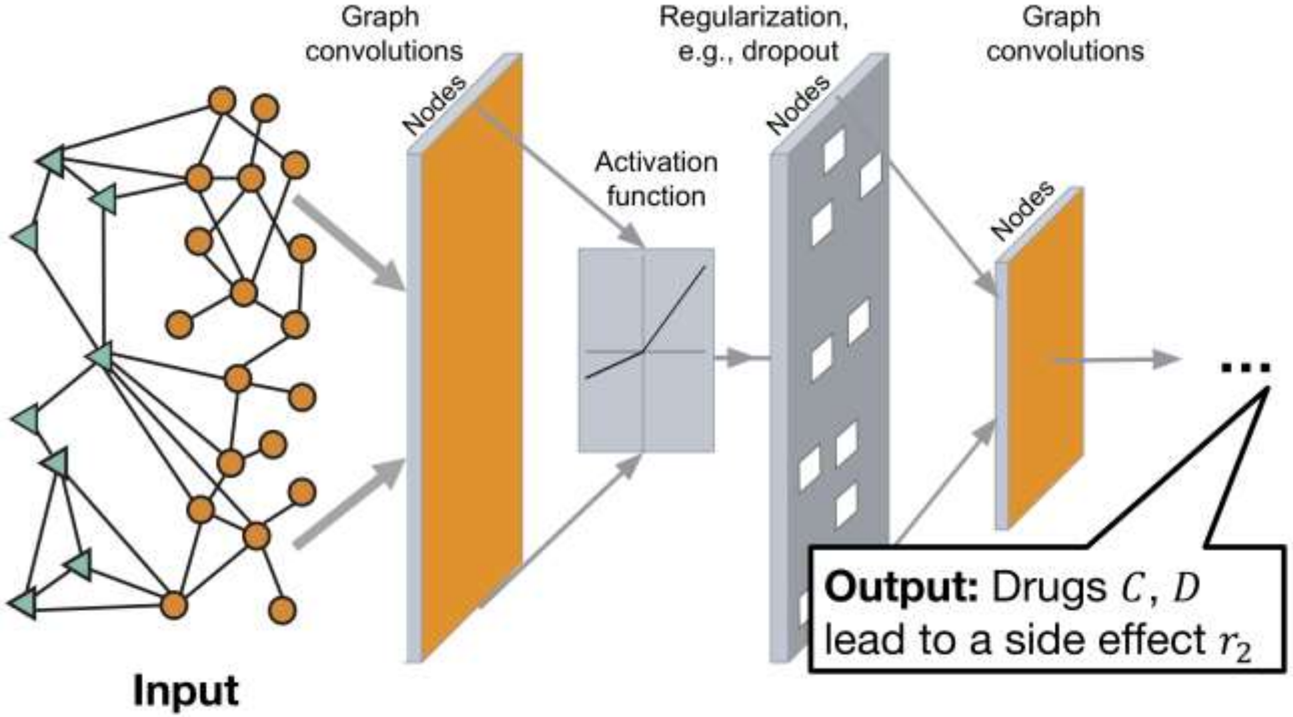
"Graph Auto-Encoders"

1) Graph classification

Make a prediction over the entire graph.

Akin to assigning a label to an entire image.





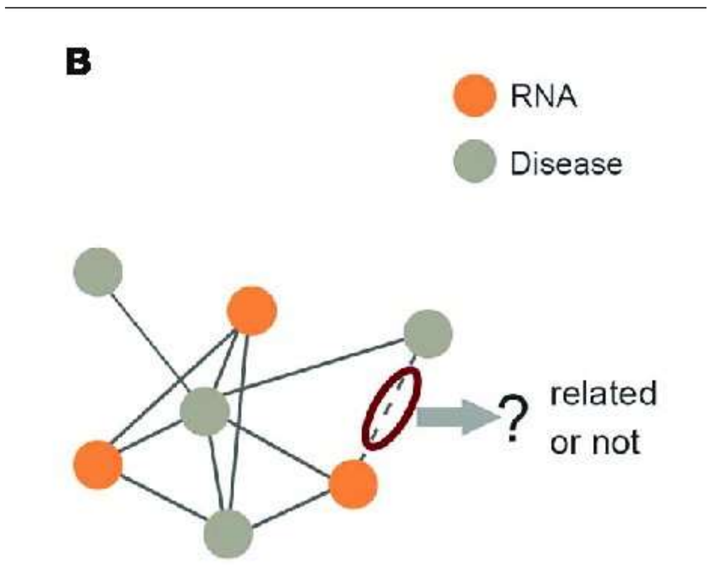
2) Node classification

Make a prediction for each individual node.

Akin to segmentation for images.

3) Link prediction

Make a prediction for each edge between two nodes.

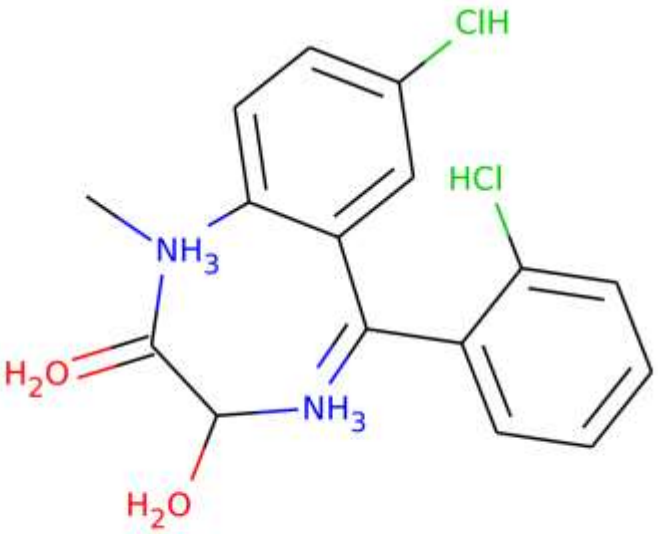


RNA-Disease association network

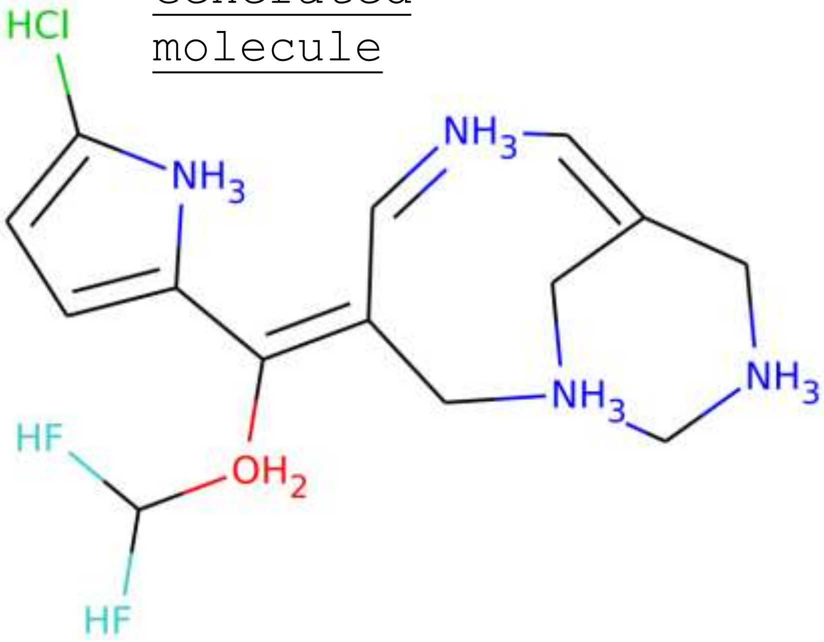
4) Graph generation

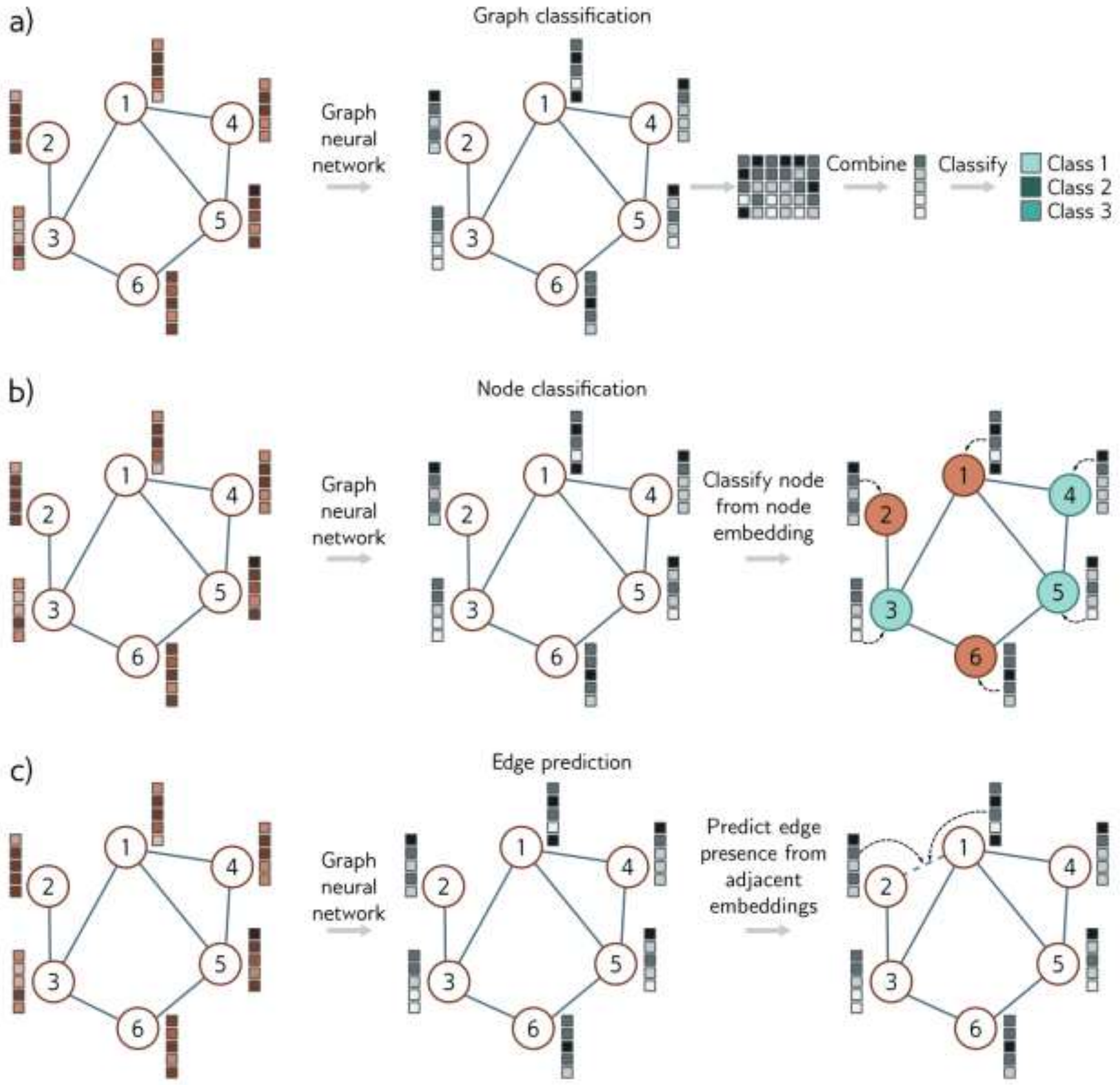
Similar in spirit to image/text generation, topic of next week.

Example molecule



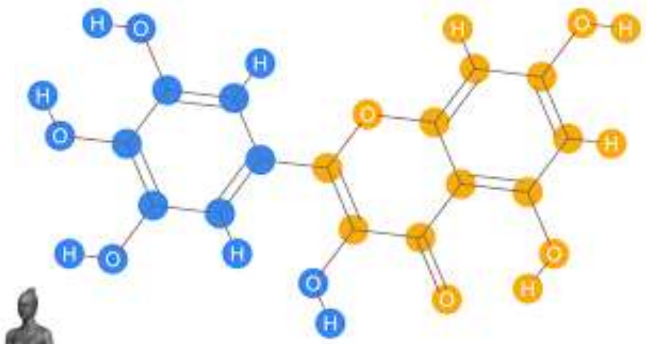
Generated molecule



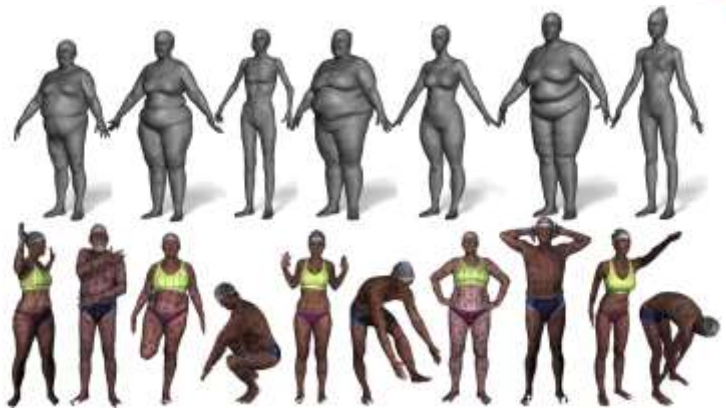

a)
Graph classification
Graph neural network
Combine
Classify
Class 1
Class 2
Class 3
b)
Node classification
Graph neural network
Classify node from node embedding
c)
Edge prediction
Graph neural network
Predict edge presence from adjacent embeddings
1
2
3
4
5
6

Graphs can be dynamic

Graphs have fixed structures.



But many are subject to

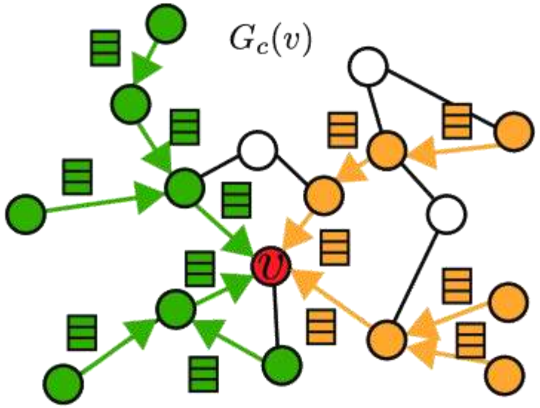


In practice, this change can even be gradual
continuous.

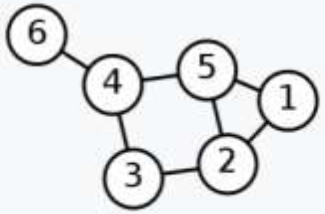
Regular structures and graphs

Regular structures are a subset of graphs. I.e., images are grid graphs.

- Convolution + pooling
- Local neighborhood: fixed window
- Constant number of neighbors
- With fixed ordering
- Translation equivariance



- Message passing + coarsening
- Local neighborhood: 1-hop
- Different number of neighbors
- No ordering of neighbors
- Local permutation equivariance

Labelled graph	Degree matrix	Adjacency matrix	Laplacian matrix
	$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$

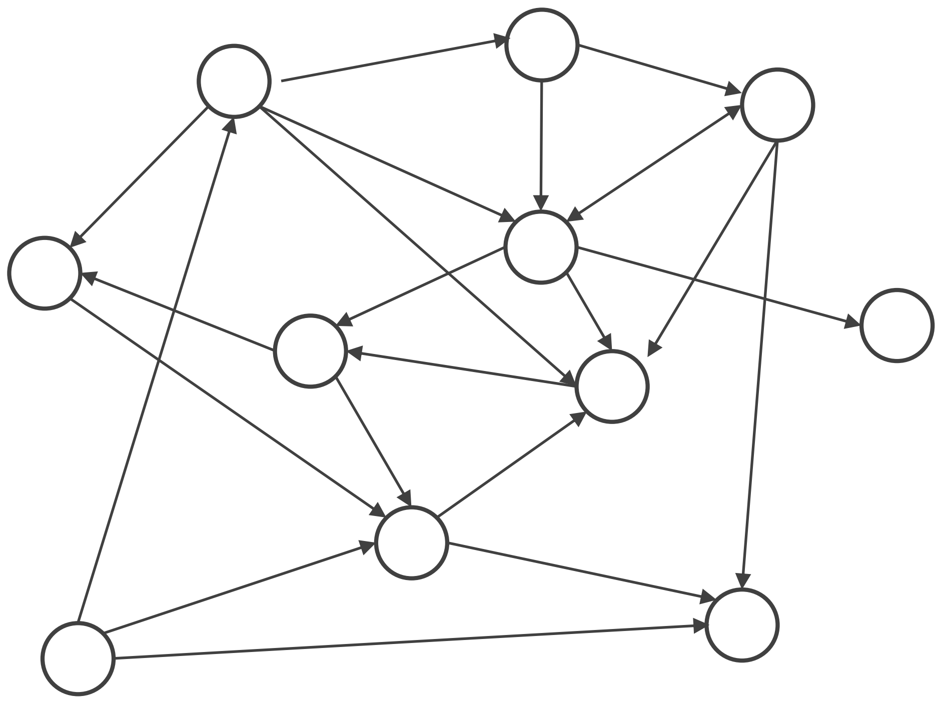
Definition of a graph

(in deep learning)

Directed graphs

Vertices $\mathcal{V} = \{1, \ldots, n\}$, also called “nodes”

Edges $\mathcal{E} = \{(i, j) : i, j \in \mathcal{V}\} \subseteq \mathcal{V} \times \mathcal{V}$ (directed)

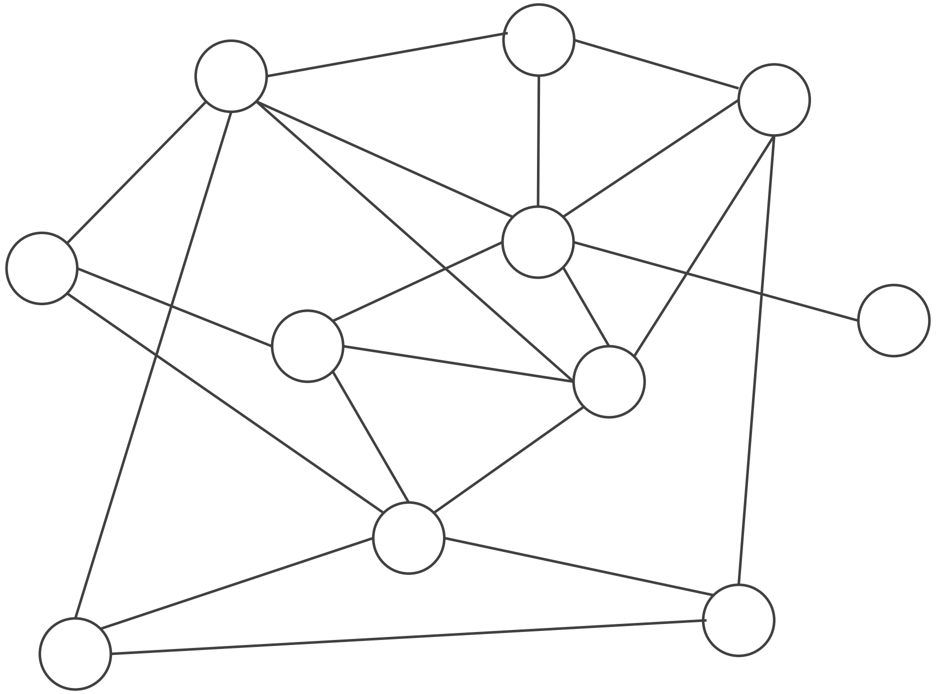


Undirected graphs

Vertices $\mathcal{V} = \{1, \ldots, n\}$

Edges $\mathcal{E} = \{(i, j) : i, j \in \mathcal{V}\} \subseteq \mathcal{V} \times \mathcal{V}$ (directed)

Edges $\mathcal{E} = \{\{i, j\} : i, j \in \mathcal{V}\} \subseteq \mathcal{V} \times \mathcal{V}$ (undirected)



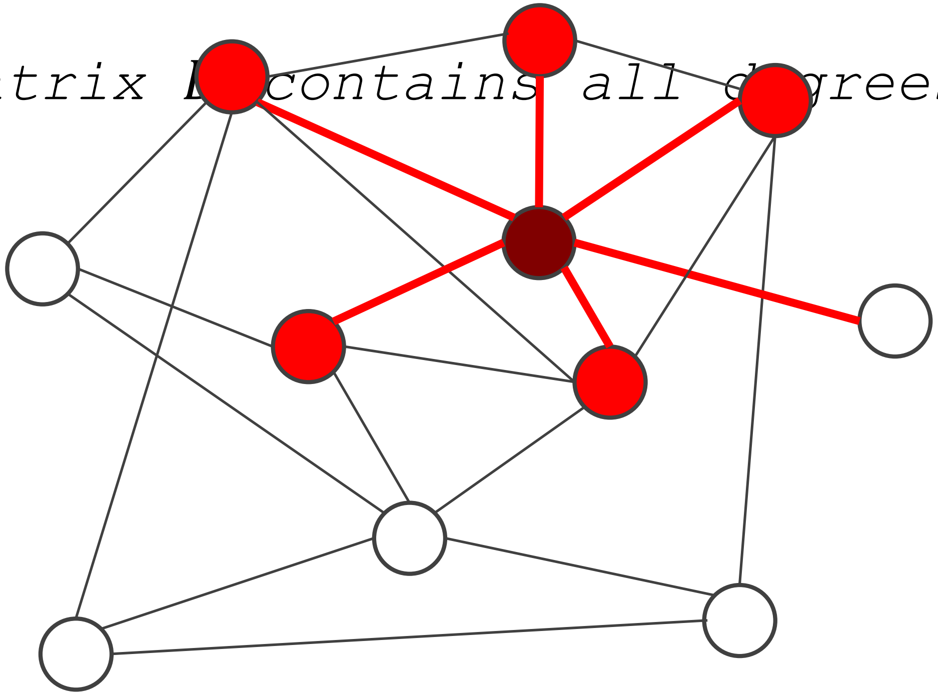
Graph neighborhood

The neighborhood of a node consists of all nodes directly connected to it

$$\mathcal{N}(i) = \{j : (i, j) \in \mathcal{E}\}$$

The **degree** of a node is the number of neighbors: $d_i = |\mathcal{N}(i)|$

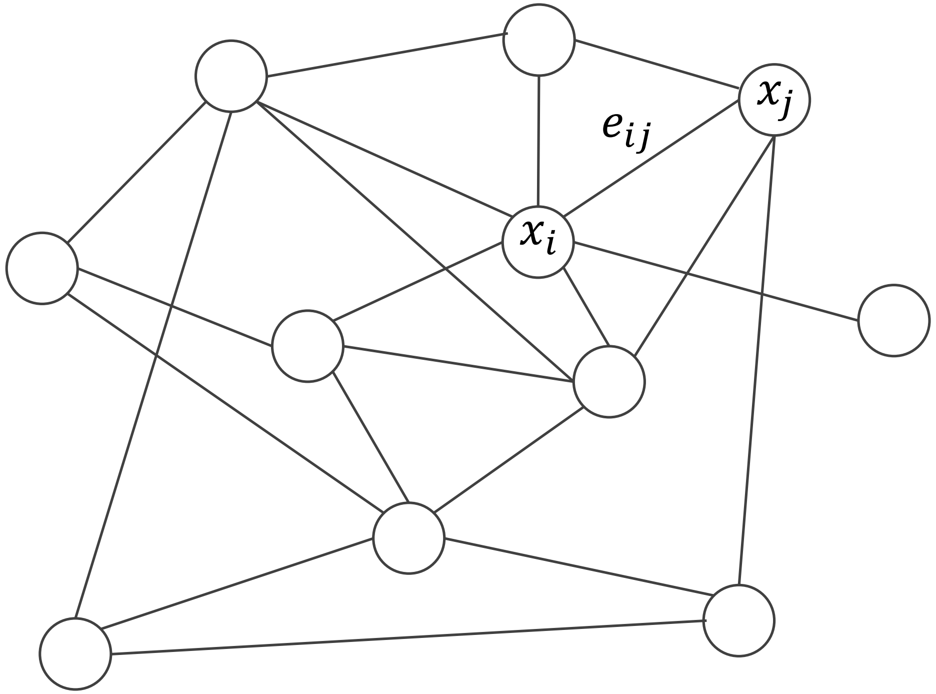
The diagonal matrix D contains all degrees per node



Attributes

Node features $\boldsymbol{x}: \mathcal{V} \to \mathbb{R}^d$, $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$

Edge features $\boldsymbol{e}_{ij}: \mathcal{E} \to \mathbb{R}^{d'}$

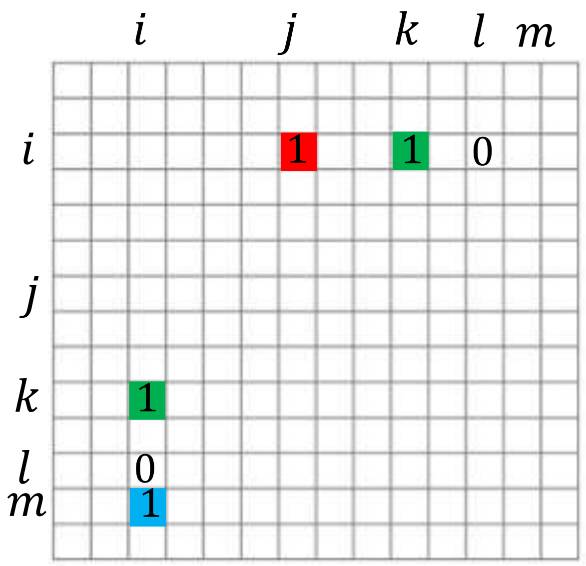


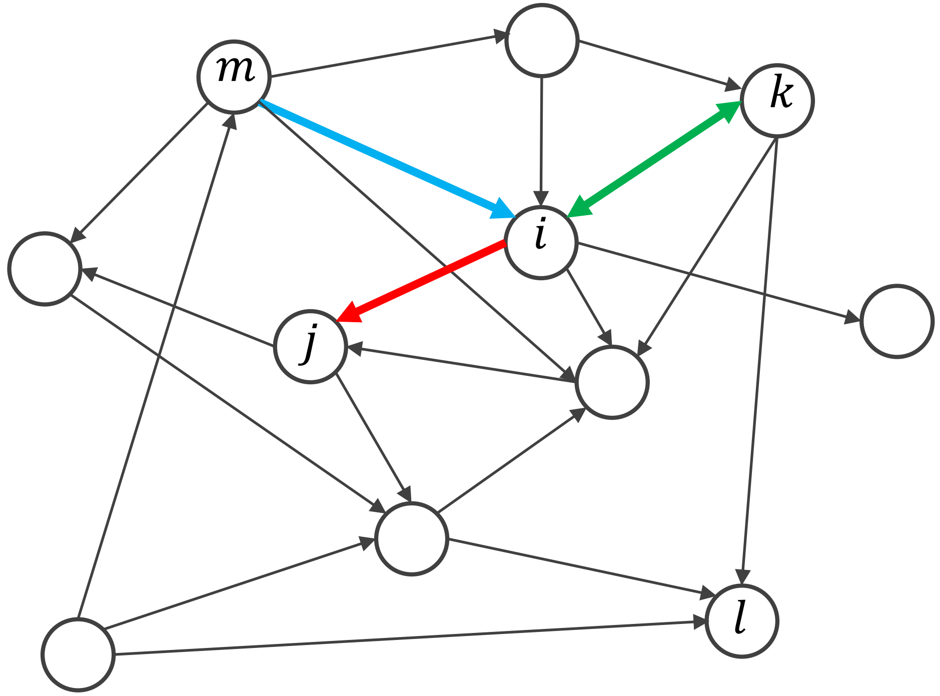
Adjacency matrix

An $n \times n$ matrix A, for n nodes

$$A_{ij} = \begin{cases} 1 \text{ if } (\mathrm{i,j}) \in \mathcal{E} \\ 0 \text{ if } (\mathrm{i,j}) \notin \mathcal{E} \end{cases}$$

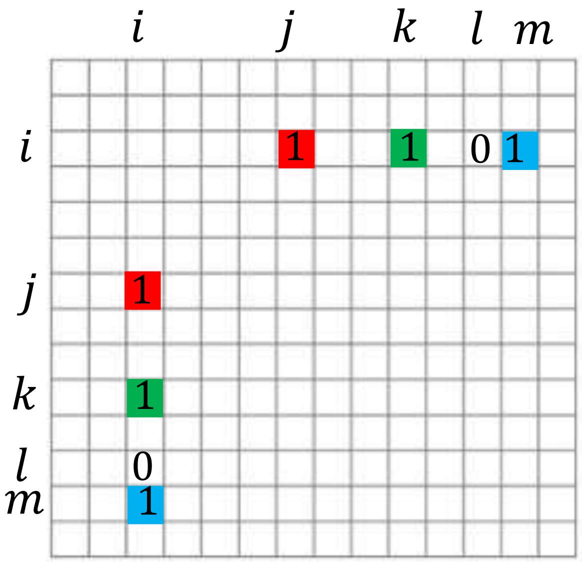
$(A^z)_{ij}$: number of paths that go from i to j in z steps

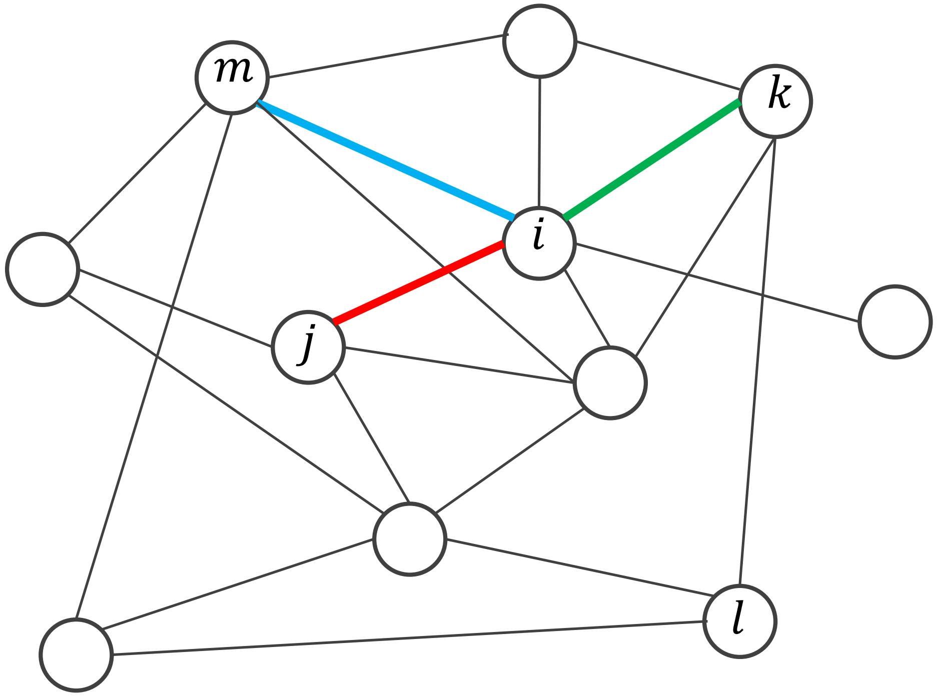




Adjacency matrix for undirected graphs

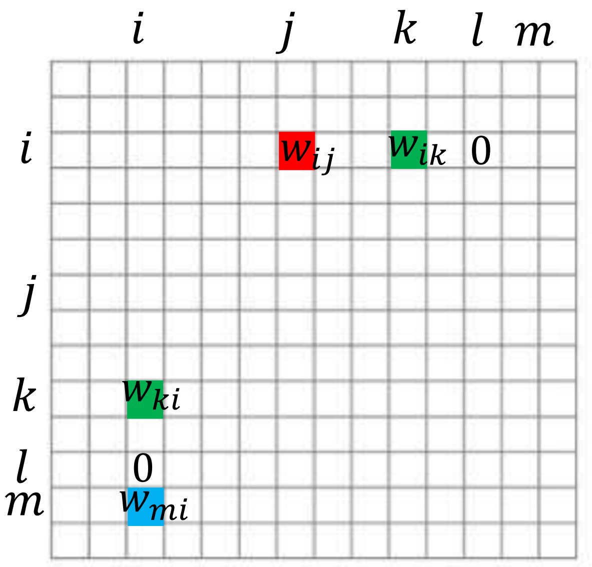
The adjacency matrix is symmetric for undirected graphs.

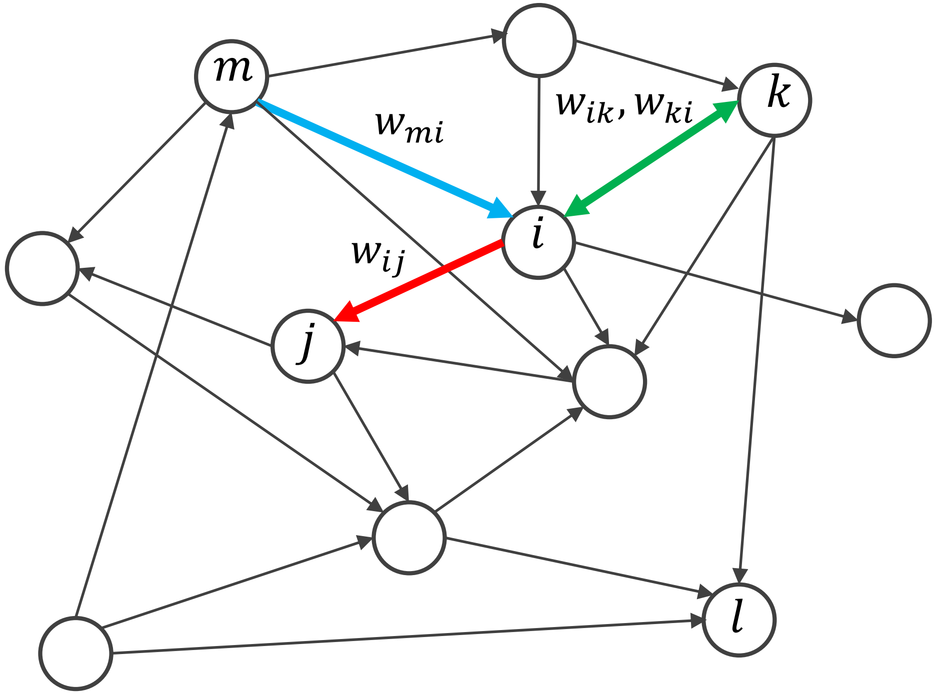




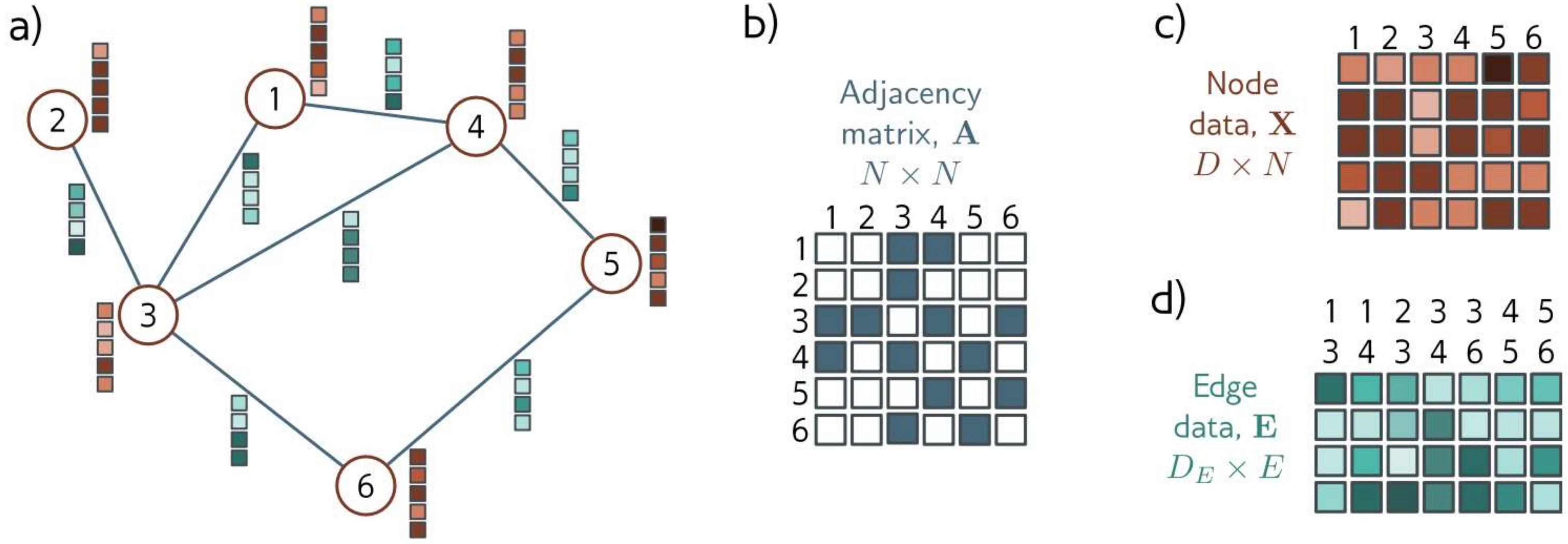
Weighted adjacency matrix

When the edges have weights, so does the adjacency matrix.





Final graph input representation

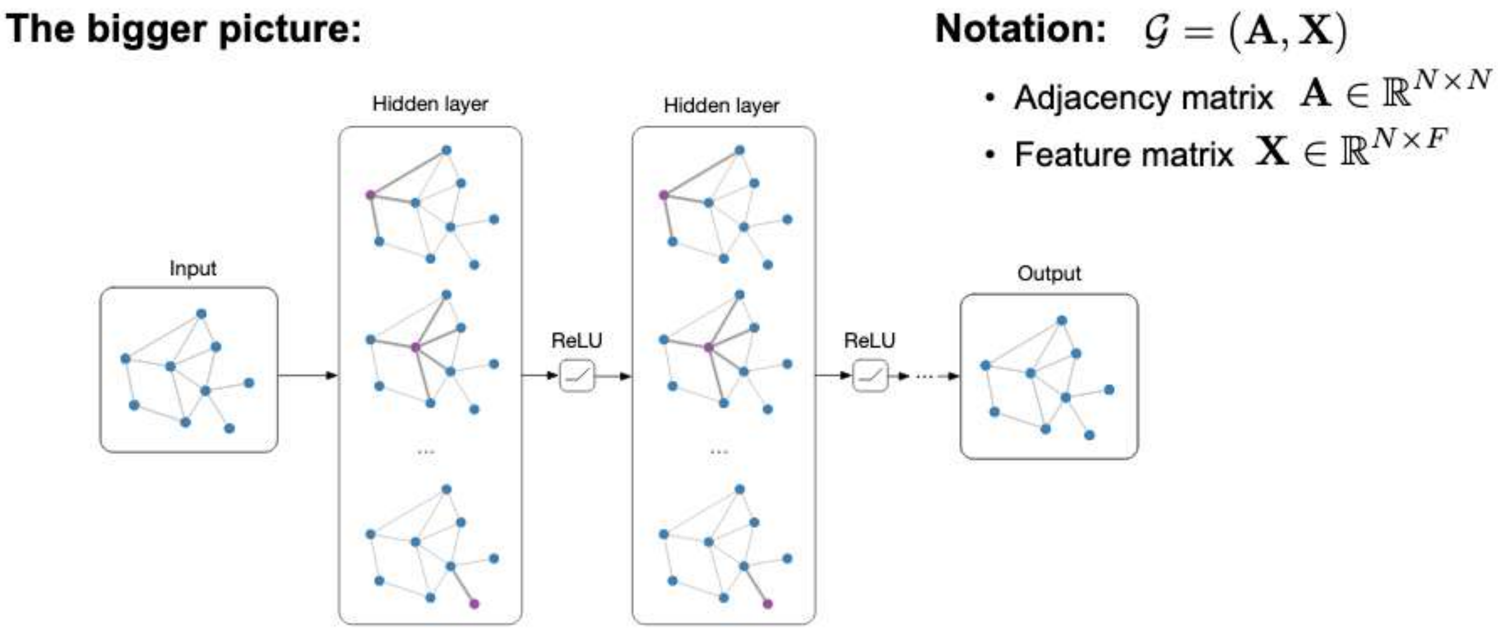


Back to graph networks

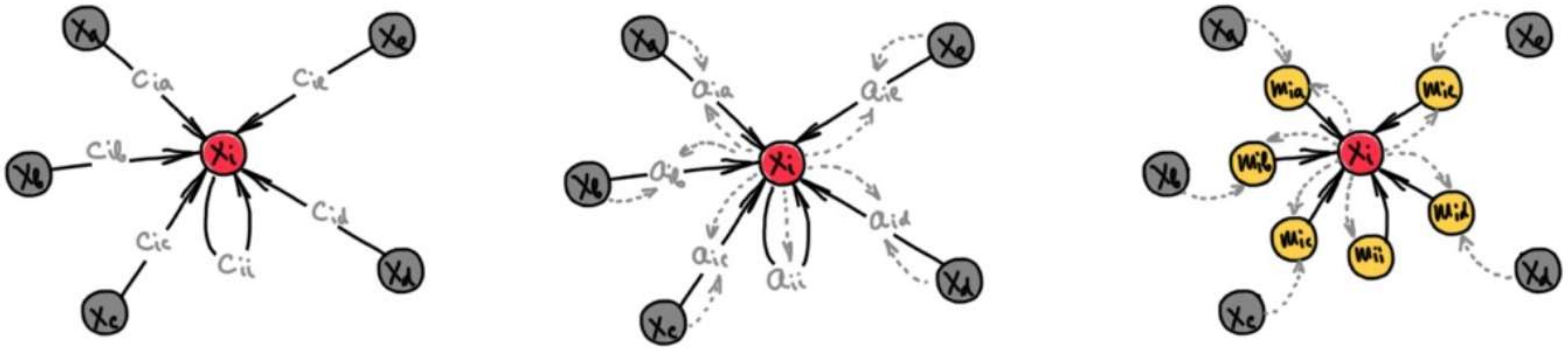
We can do a lot of processing on this data structure.

But the pre-defined features are raw inputs.

Graph networks do 1 thing: transform the feature vector per node over layers.



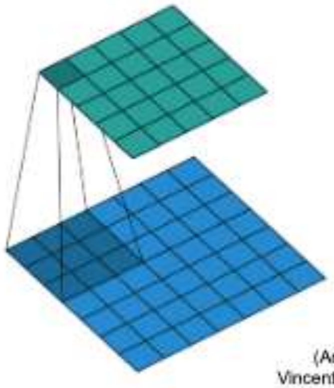
Three perspectives to graph networks



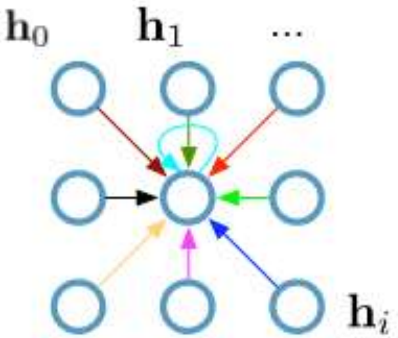
Three "flavours" of GNNs, left-to-right: convolutional, attentional, and general nonlinear message passing flavours. All are forms of message passing. Figure adapted from P. Veličković.

Graph layer as a convolution layer

Single CNN layer with 3x3 filter:



(Animation by Vincent Dumoulin)



Update for a single pixel:

- Transform messages individually $\mathbf{W}_i \mathbf{h}_i$
- Add everything up $\sum_i \mathbf{W}_i \mathbf{h}_i$

$\mathbf{h}_i \in \mathbb{R}^F$ are (hidden layer) activations of a pixel/node

Full update:

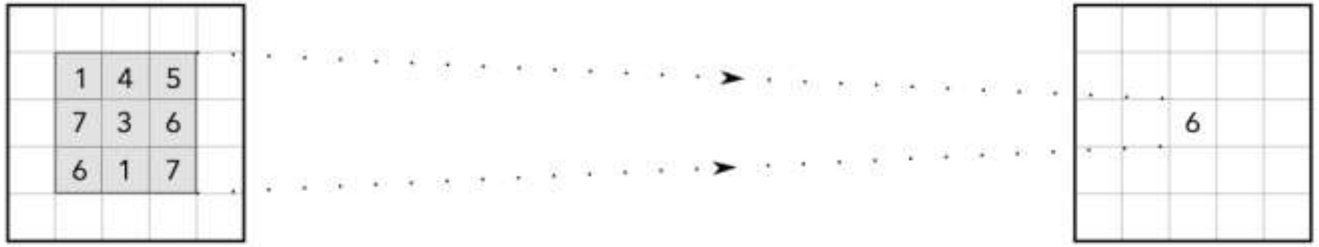
$$\mathbf{h}_4^{(l+1)} = \sigma \left(\mathbf{W}_0^{(l)} \mathbf{h}_0^{(l)} + \mathbf{W}_1^{(l)} \mathbf{h}_1^{(l)} + \cdots + \mathbf{W}_8^{(l)} \mathbf{h}_8^{(l)} \right)$$

Which assumptions from images are no longer valid?

Number of neighbors per node no longer fixed.

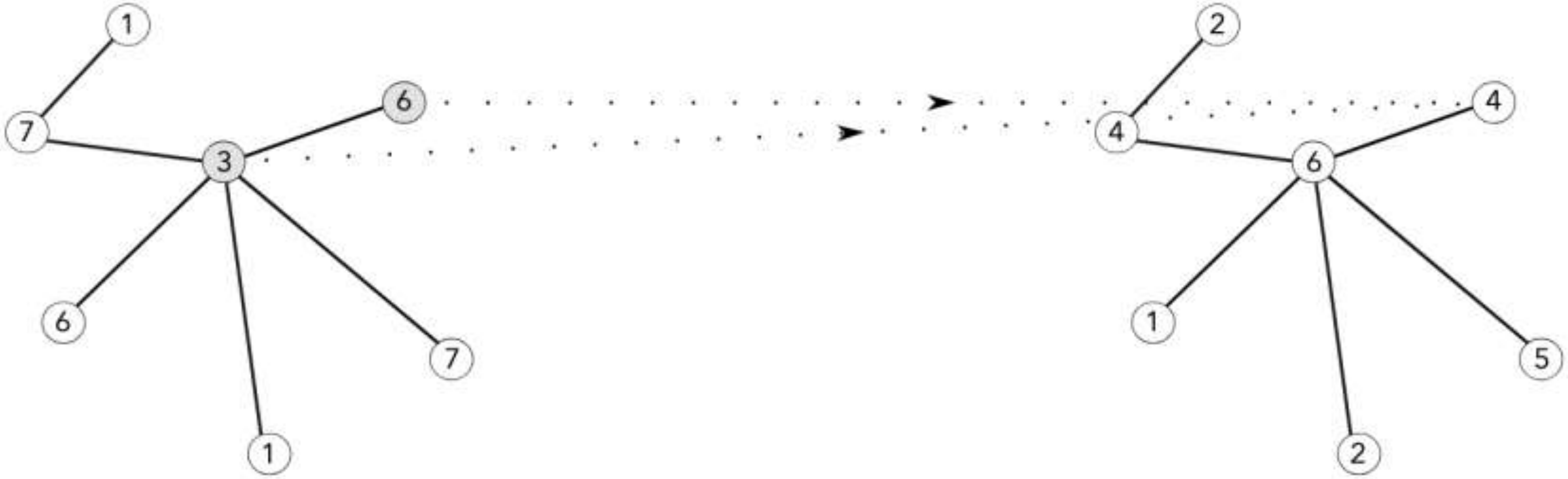
No more ordering between neighbours.

Extending convolutions to graphs



Convolution in CNNs

Convolutions in CNNs are inherently localized.
Neighbours participating in the convolution at the
center pixel are highlighted in gray.

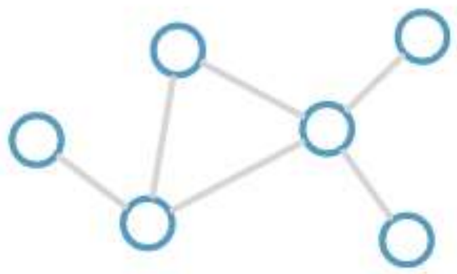


Localized Convolution in GNNs

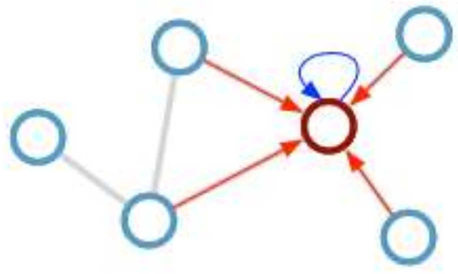
GNNs can perform localized convolutions mimicking CNNs. Hover over
a node to see its immediate neighbourhood highlighted on the left.
The structure of this neighbourhood changes from node to node.

Graph convolution layer

Consider this undirected graph:



Calculate update for node in red:



Stacking graph convolution layers

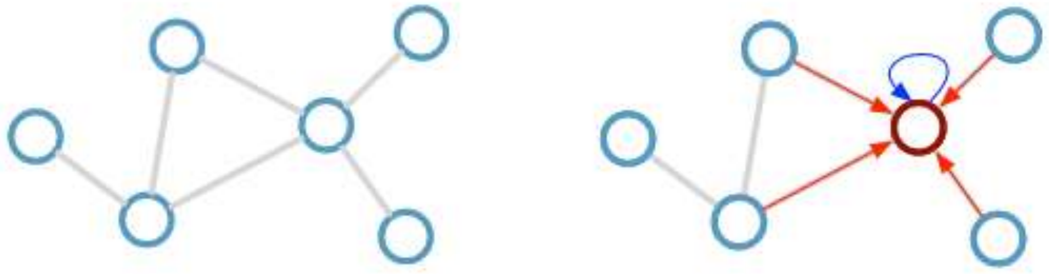
Each layer aggregates information from their direct neighbors.

At the end of each layer, we add a non-linearity such as a ReLU.

We can increase complexity and receptive field simply by stacking multiple layers.

Consider this undirected graph:

Calculate update for node in red:



Update rule:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\mathbf{h}_i^{(l)}\mathbf{W}_0^{(l)} + \sum_{j\in\mathcal{N}_i}\frac{1}{c_{ij}}\mathbf{h}_j^{(l)}\mathbf{W}_1^{(l)}\right)$$

https://mbernste.github.io/posts/gcn/

Graph convolution layer in matrix form

$$f(X, A) := \sigma\left(D^{-1/2}(A+I)D^{-1/2}XW\right)$$

$$
\begin{aligned}
A \in \mathbb{R}^{n\times n} &:= \text{The adjacency matrix} \\
I \in \mathbb{R}^{n\times n} &:= \text{The identity matrix} \\
D \in \mathbb{R}^{n\times n} &:= \text{The degree matrix of } A + I \\
X \in \mathbb{R}^{n\times d} &:= \text{The input data (i.e., the per-node feature vectors)} \\
W \in \mathbb{R}^{d\times w} &:= \text{The layer's weights} \\
\sigma(.) &:= \text{The activation function (e.g., ReLU)}
\end{aligned}
$$

Kipf and Welling (2017)

Let’s break it down

$$f(\boldsymbol{X}, \boldsymbol{A}) \coloneqq \sigma\Big(\underbrace{\underbrace{\underbrace{\boldsymbol{D}^{-1/2}\underbrace{(\boldsymbol{A} + \boldsymbol{I})}_{\text{Add self-loops}}\boldsymbol{D}^{-1/2}}_{\text{Normalize adjacency matrix}}\boldsymbol{X}}_{\text{Aggregate}}\boldsymbol{W}}_{\text{Update}}\Big)$$

Let’s break it down

Add ones to diagonal, needed because each node should pass its own vector through.

$$f(\boldsymbol{X},\boldsymbol{A}) \coloneqq \sigma\Big(\underbrace{\underbrace{\underbrace{\boldsymbol{D}^{-1/2}\underbrace{(\boxed{\boldsymbol{A}+\boldsymbol{I}})}_{\text{Add self-loops}}\boldsymbol{D}^{-1/2}}_{\text{Normalize adjacency matrix}}\boldsymbol{X}}_{\text{Aggregate}}\boldsymbol{W}}_{\text{Update}}\Big)$$

Let's break it down

This step essentially normalizes the adjaceny matrix. I will show how in a few slides.

$$f(\boldsymbol{X},\boldsymbol{A}) \coloneqq \sigma\Big(\underbrace{\underbrace{\underbrace{\boxed{\boldsymbol{D}^{-1/2}}\underbrace{(\boldsymbol{A}+\boldsymbol{I})}_{\text{Add self-loops}}\boxed{\boldsymbol{D}^{-1/2}}}_{\text{Normalize adjacency matrix}}\boldsymbol{X}}_{\text{Aggregate}}\boldsymbol{W}}_{\text{Update}}\Big)$$

Let’s break it down

Just a standard linear layer and a non-linearity.

$$f(\boldsymbol{X},\boldsymbol{A}) \coloneqq \boxed{\sigma}\left(\underbrace{\underbrace{\underbrace{\boldsymbol{D}^{-1/2}\underbrace{(\boldsymbol{A}+\boldsymbol{I})}_{\text{Add self-loops}}\boldsymbol{D}^{-1/2}}_{\text{Normalize adjacency matrix}}\boldsymbol{X}}_{\text{Aggregate}}\boldsymbol{W}}_{\text{Update}}\right)$$

Let's break it down

Just a standard linear layer and a non-linearity.

$$f(\boldsymbol{X},\boldsymbol{A}) \coloneqq \sigma\left(\boxed{\underbrace{\underbrace{\underbrace{\boldsymbol{D}^{-1/2}\underbrace{(\boldsymbol{A}+\boldsymbol{I})}_{\text{Add self-loops}}\boldsymbol{D}^{-1/2}}_{\text{Normalize adjacency matrix}}\boldsymbol{X}}_{\text{Aggregate}}\boldsymbol{W}}_{\text{Update}}}\right)$$

Rewriting into 2 steps

$$\tilde{\boldsymbol{A}} := \boldsymbol{D}^{-1/2}(\boldsymbol{A} + \boldsymbol{I})\boldsymbol{D}^{-1/2}$$

$$\boldsymbol{D} := \begin{bmatrix} d_{1,1} & 0 & 0 & \dots & 0 \\ 0 & d_{2,2} & 0 & \dots & 0 \\ 0 & 0 & d_{3,3} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_{n,n} \end{bmatrix}$$

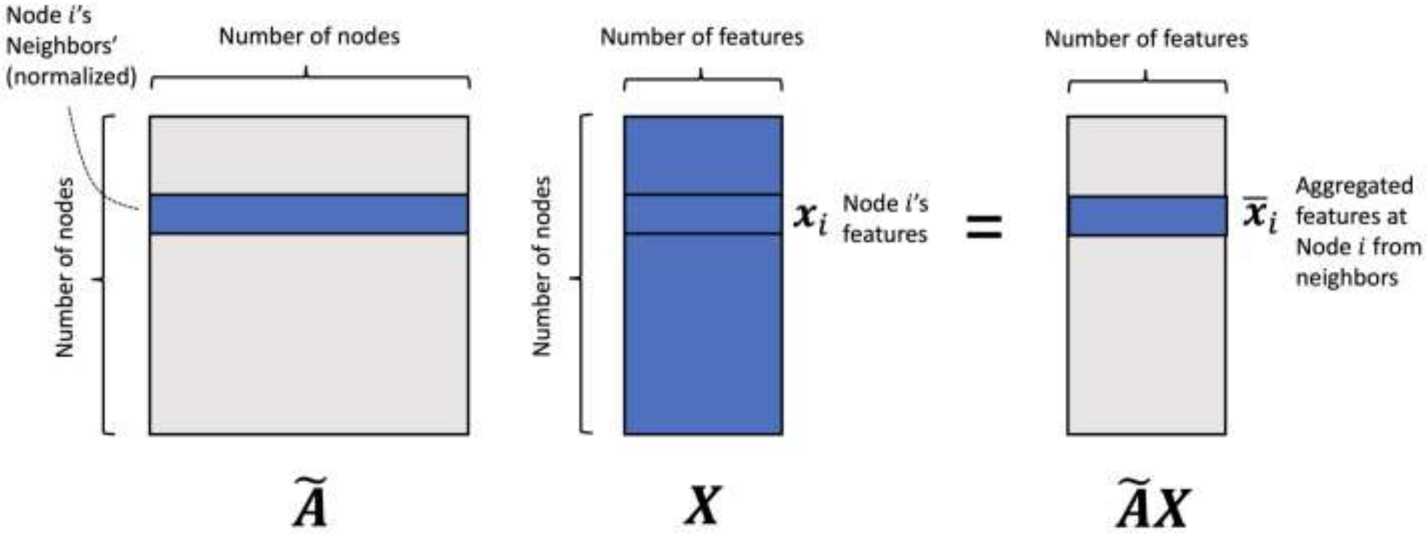
$$\tilde{\boldsymbol{A}}_{i,j} := \begin{cases} \frac{1}{\sqrt{d_{i,i}d_{j,j}}}, & \text{if there is an edge between node } i \text{ and } j \\ 0, & \text{otherwise} \end{cases}$$

$$\boldsymbol{D}^{-1/2} := \begin{bmatrix} \frac{1}{\sqrt{d_{1,1}}} & 0 & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{d_{2,2}}} & 0 & \dots & 0 \\ 0 & 0 & \frac{1}{\sqrt{d_{3,3}}} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \frac{1}{\sqrt{d_{n,n}}} \end{bmatrix}$$

$$f(\boldsymbol{X}, \boldsymbol{A}) := \sigma\left(\tilde{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W}\right)$$

$$f(X, A) := \sigma\left(\boxed{\tilde{A}X}W\right)$$

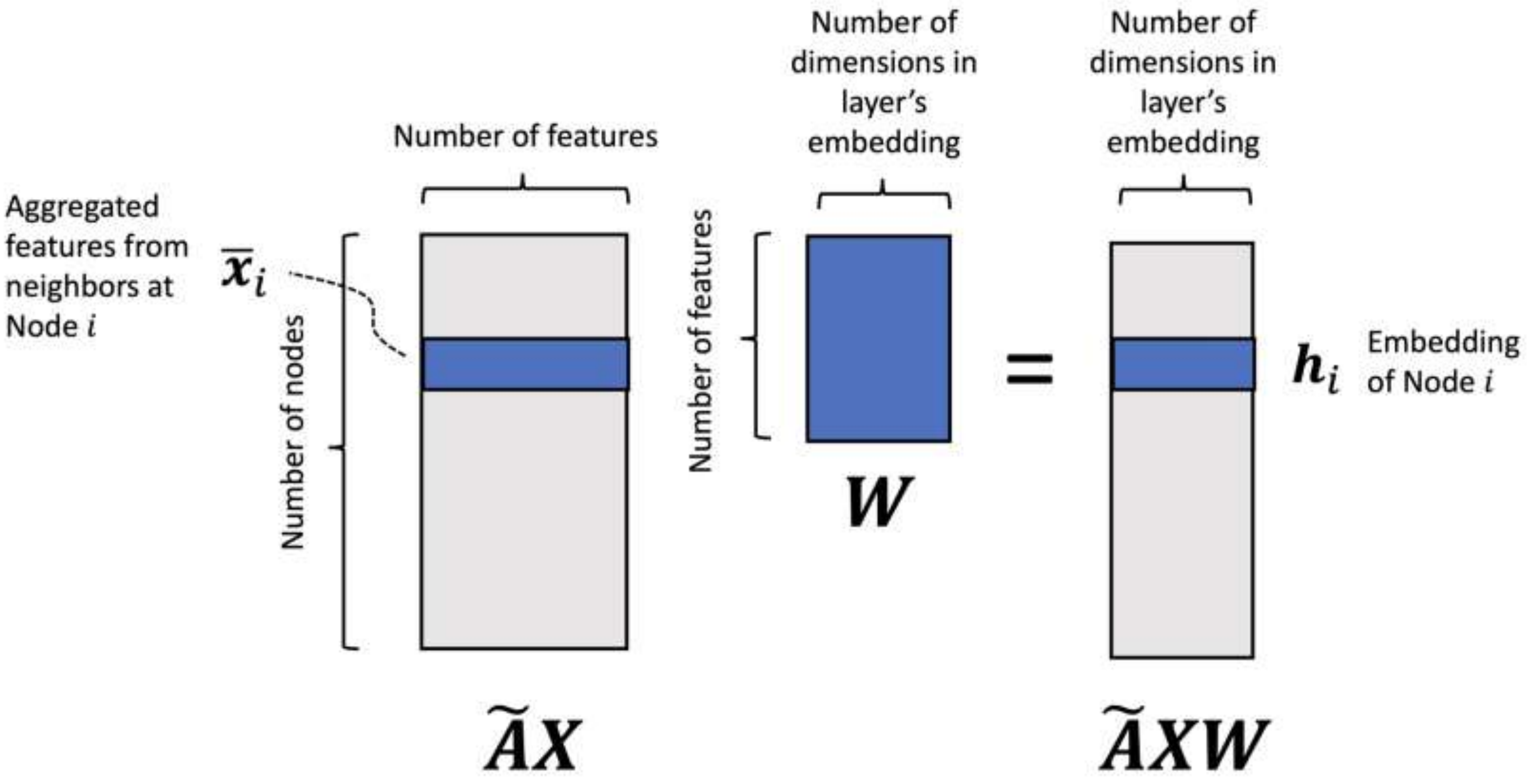
Left side of the equation



$$\begin{aligned} \bar{x}_i &= \sum_{j=1}^{n} \tilde{a}_{i,j} x_j \\ &= \sum_{j \in \mathrm{Neigh}(i)} \tilde{a}_{i,j} x_j \\ &= \sum_{j \in \mathrm{Neigh}(i)} \frac{1}{\sqrt{d_{i,i} d_{j,j}}} x_j \end{aligned}$$

$f(X, A) := \sigma\left(\boxed{\tilde{A}XW}\right)$

Right side of the equation



Why add a normalization step?

Do we even need it? Let's see what happens without it:

$$\hat{\boldsymbol{A}} := \boldsymbol{A} + \boldsymbol{I}$$

Normalization dropped,
only self-loop
retrained

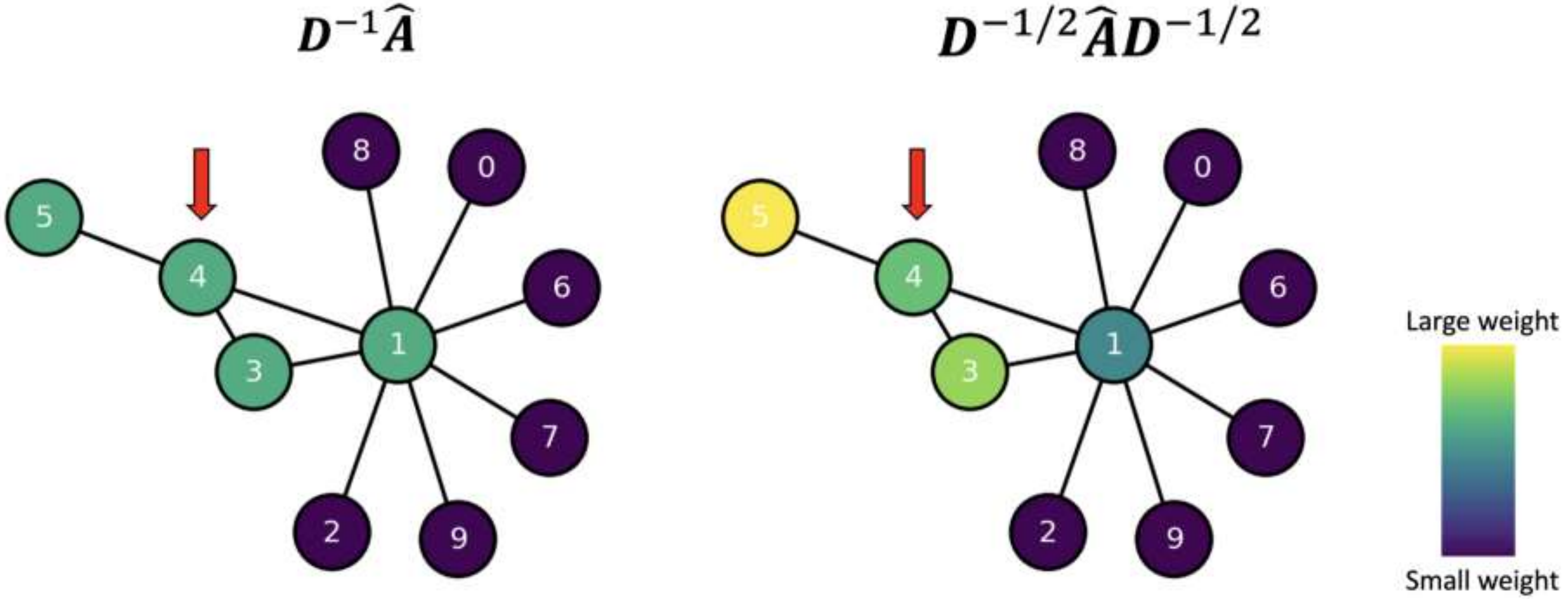
$$f_{\text{unnormalized}}(\boldsymbol{X}, \boldsymbol{A}) := \sigma(\hat{\boldsymbol{A}}\boldsymbol{X}\boldsymbol{W})$$

Layer update remains the
same

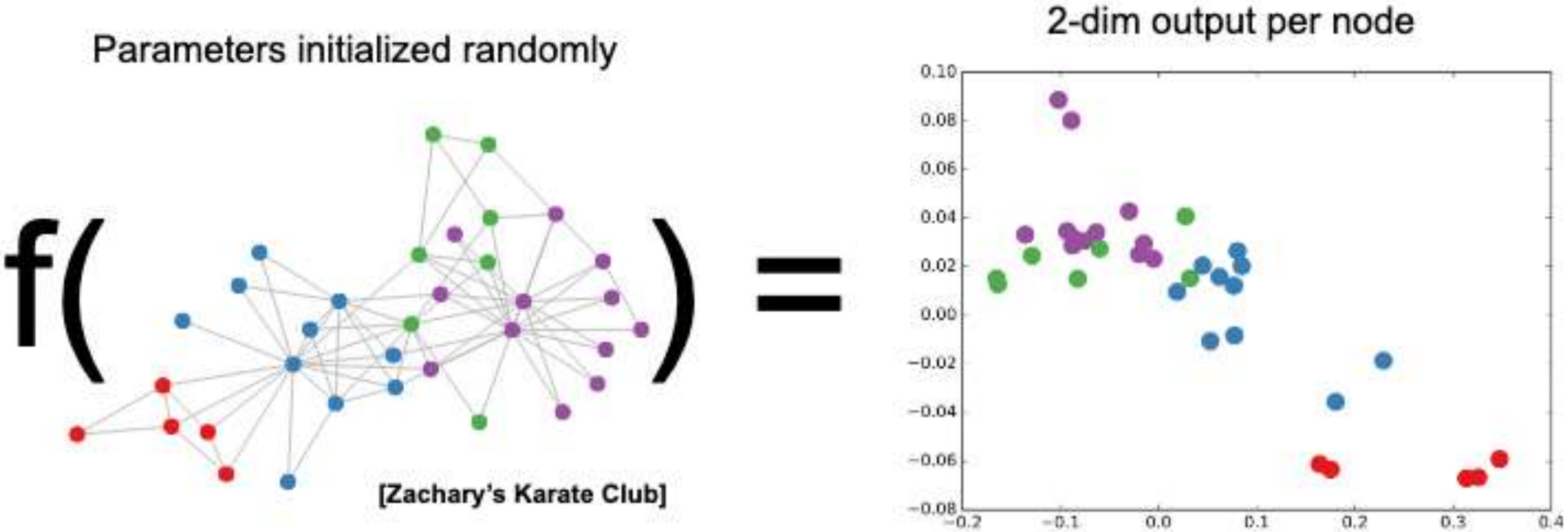
$$\begin{aligned}\bar{\boldsymbol{x}}_i &= \sum_{j=1}^{n} \hat{a}_{i,j}\boldsymbol{x}_j \\ &= \sum_{j=1}^{n} \mathbb{I}(j \in \text{Neigh}(i))\boldsymbol{x}_j \\ &= \sum_{j \in \text{Neigh}(i)} \boldsymbol{x}_j\end{aligned}$$

Problem! More neighbors =
bigger sum.
Huge bias when training graph
networks

Why not simply divide by the node degree?



Visualizing node representations



Alternative: graph layer as attention

Similar but including attention as *aggregation*: $y_i = h\left(\sum_{j\in\mathcal{N}(i)} a_{ij}\boldsymbol{z_j}\right)$

Using self-attention:

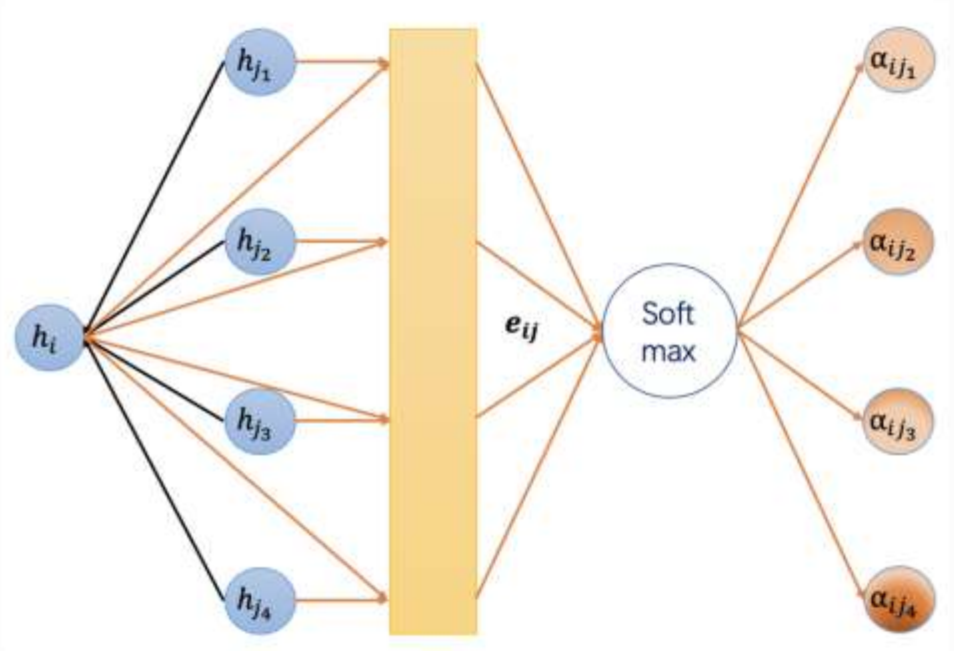
$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}(i)} \exp(e_{ij})},$$

where $\boldsymbol{e_{ij}}$ are the self-attention weights (like query == key)

$$e_{ij} = \mathrm{LeakyRELU}\left([\boldsymbol{x}_i\boldsymbol{W}, \boldsymbol{x}_j\boldsymbol{W}] \cdot u\right)$$

$\boldsymbol{u}$ is a weight vector.

The four steps of a graph attention layer



$$z_i^{(l)} = W^{(l)} h_i^{(l)}, \tag{1}$$

$$e_{ij}^{(l)} = \text{LeakyReLU}(\vec{a}^{(l)^T}(z_i^{(l)} || z_j^{(l)})), \tag{2}$$

$$\alpha_{ij}^{(l)} = \frac{\exp(e_{ij}^{(l)})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik}^{(l)})}, \tag{3}$$

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} z_j^{(l)}\right), \tag{4}$$

Connecting graphs, convolutions, and transformers

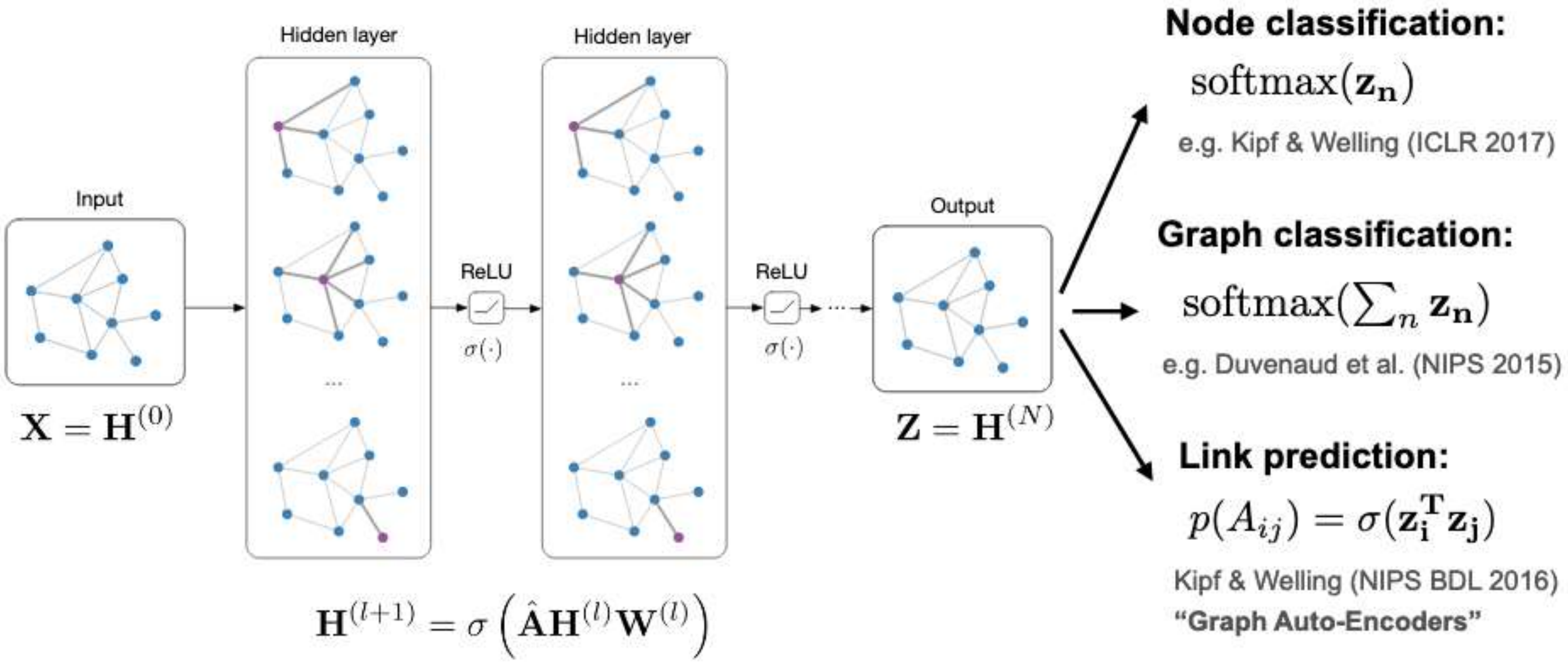
Transformers operate on a complete graph (adjacency matrix with all 1's).

With attention-based GCN, we recover the Transformer.

Architecture	Domain Ω	Symmetry group $\mathfrak{G}$
CNN	Grid	Translation
Spherical CNN	Sphere / SO(3)	Rotation SO(3)
Intrinsic / Mesh CNN	Manifold	Isometry Iso(Ω) / Gauge symmetry SO(2)
GNN	Graph	Permutation Σ_n
Deep Sets	Set	Permutation Σ_n
Transformer	Complete Graph	Permutation Σ_n
LSTM	1D Grid	Time warping

Optimizing graph networks

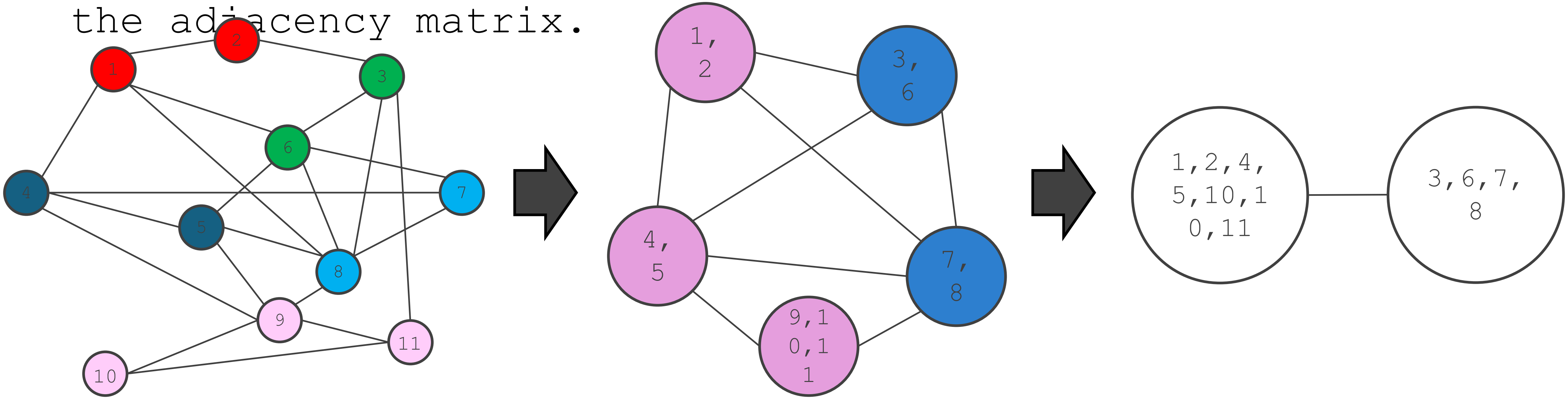
Input: Feature matrix $\mathbf{X} \in \mathbb{R}^{N \times E}$, preprocessed adjacency matrix $\hat{\mathbf{A}}$



Pooling in graph networks

Specifically for graph classification, pooling is an optional operators.

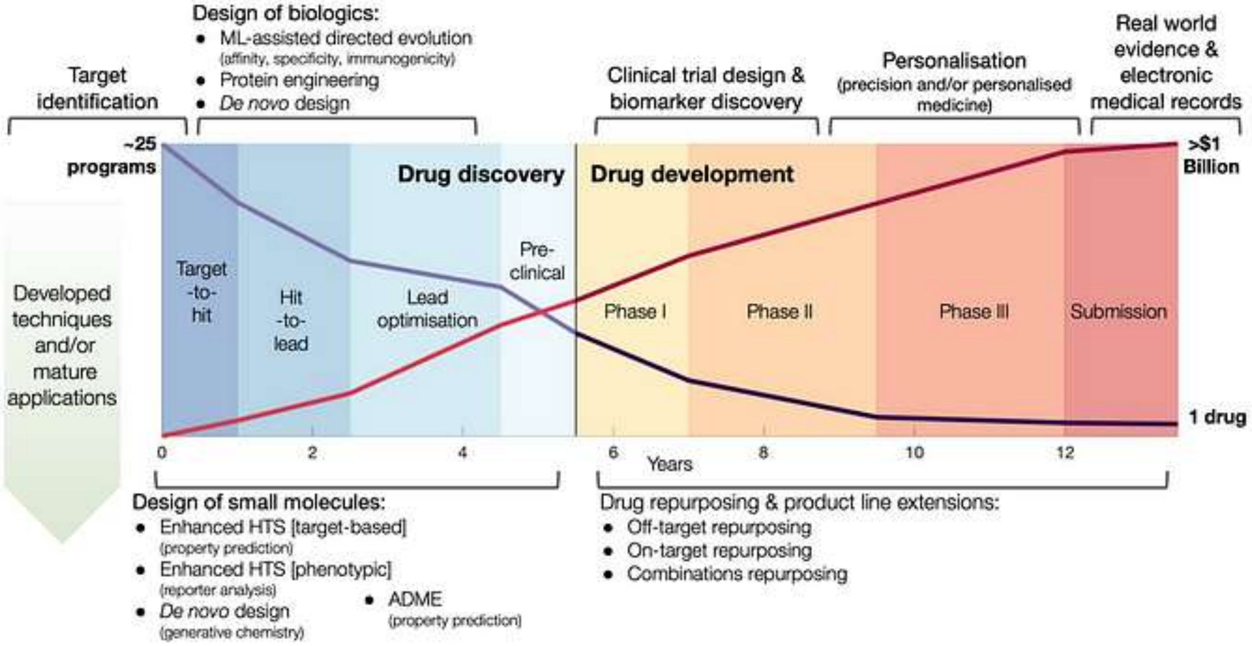
Pool nodes together to save compute, requires updating the adjacency matrix.

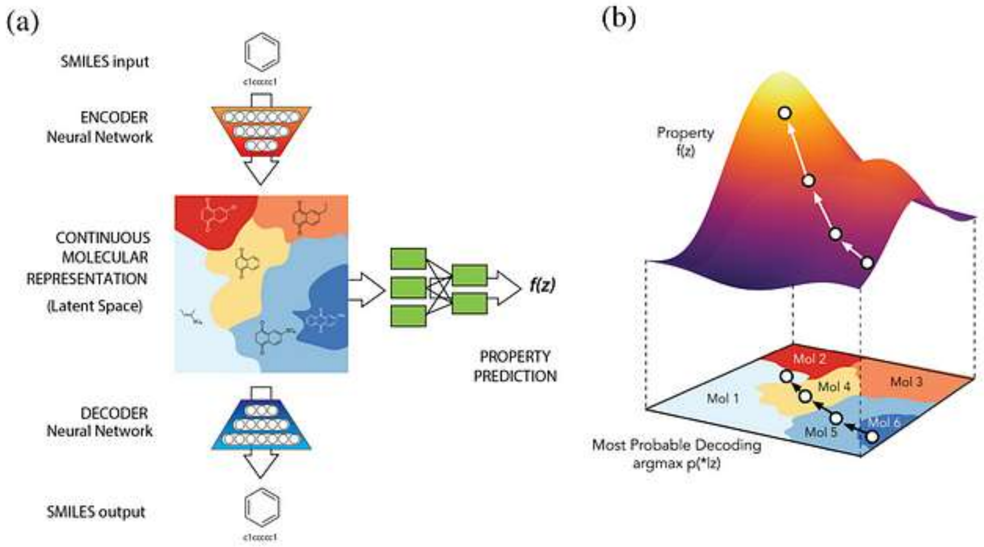


https://jonathan-hui.medium.com/applications-of-graph-neural-networks-gnn-d487fd5ed17d

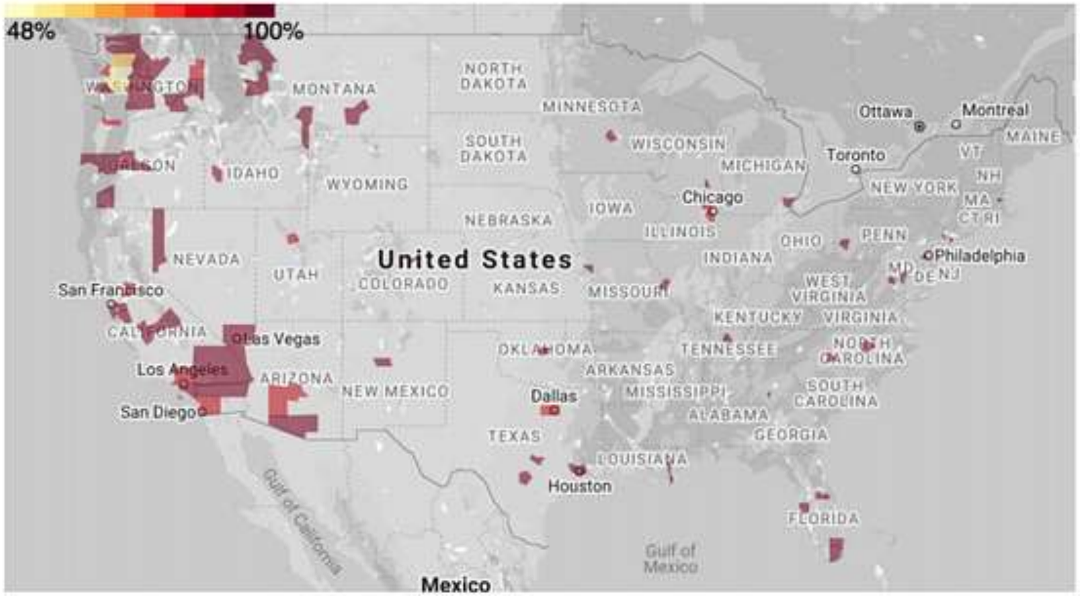
Applications of graph networks

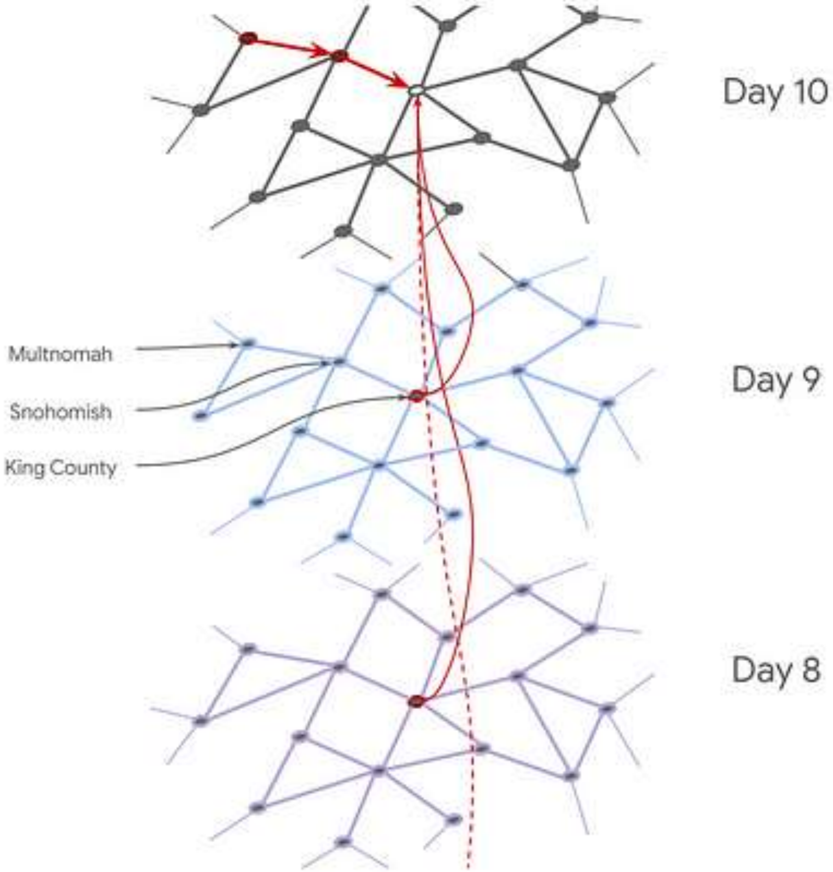
Drug discovery



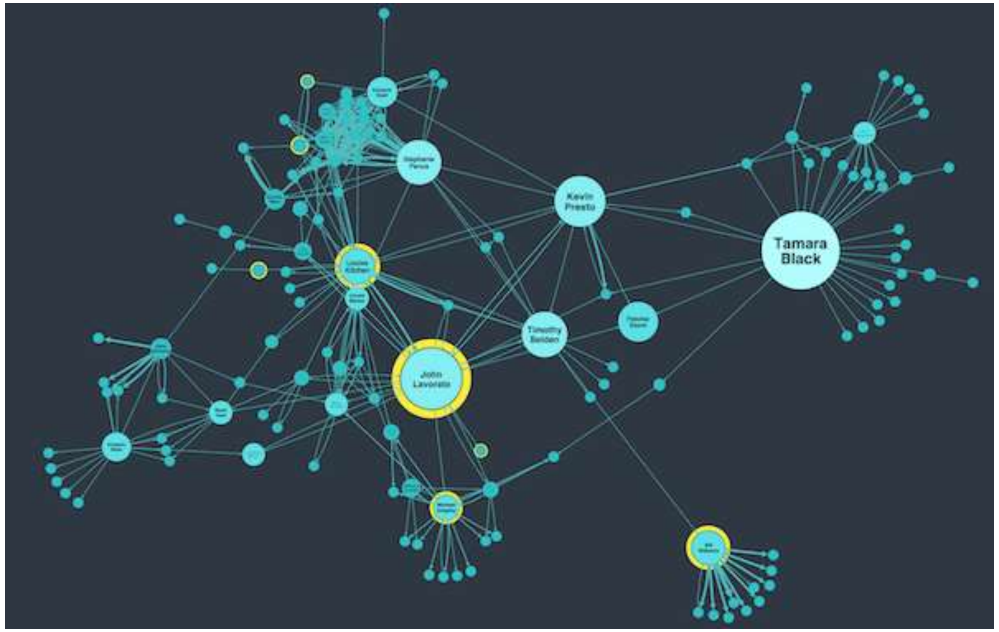


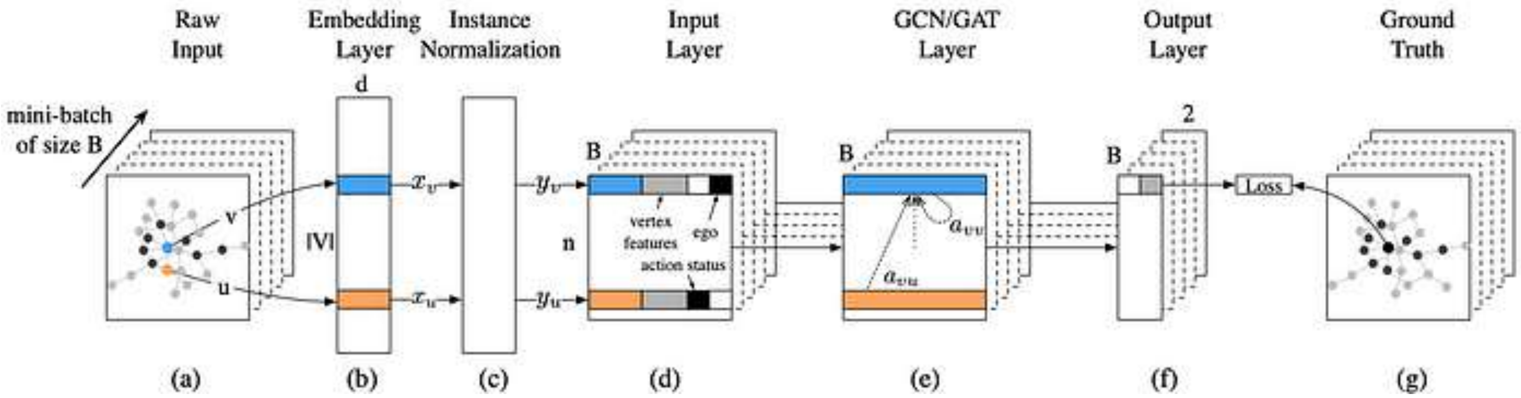
Modeling the spread of deceases



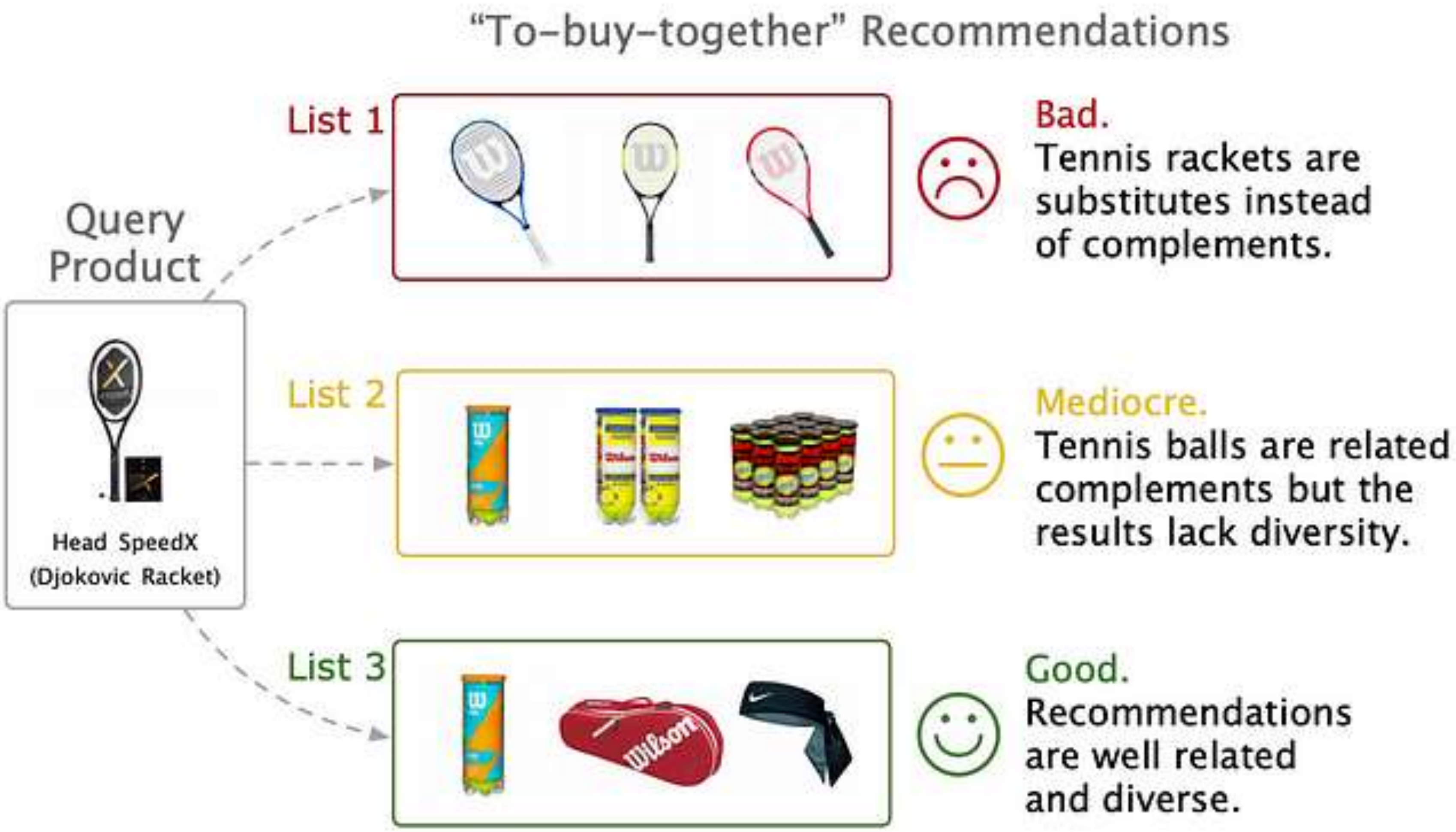


Social networks

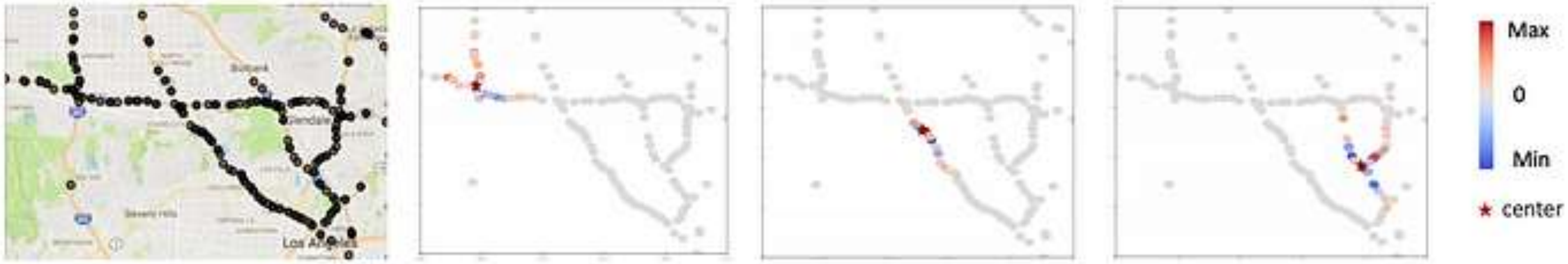


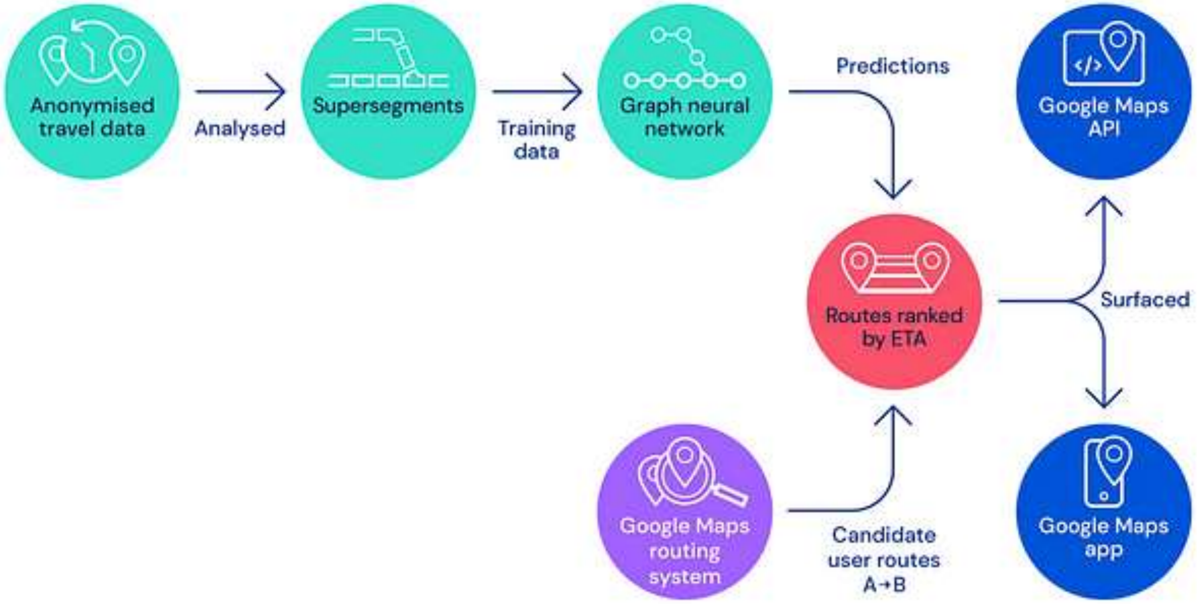


Recommendation



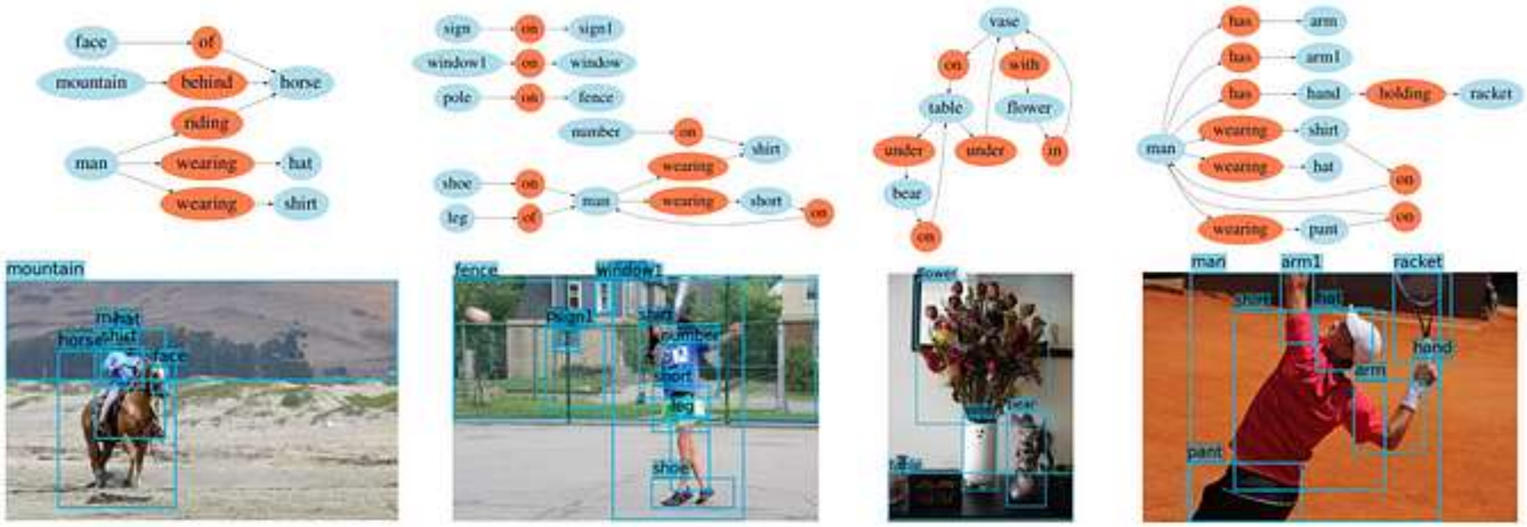
Traffic forecasting

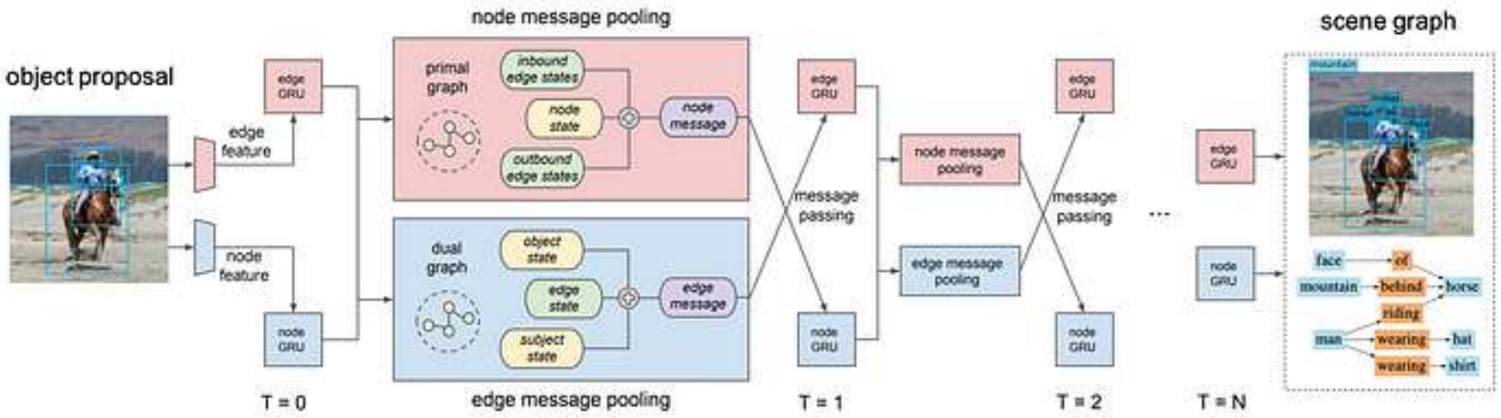




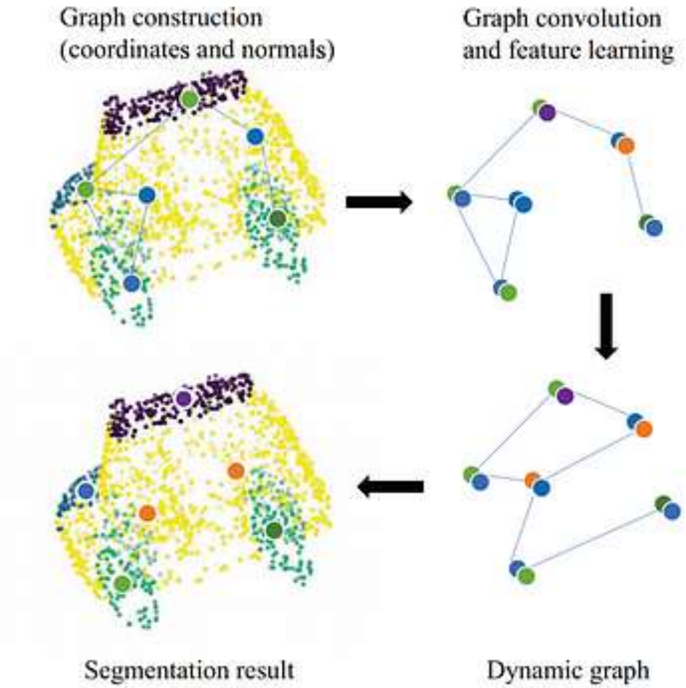
The model architecture for determining optimal routes and their travel time.

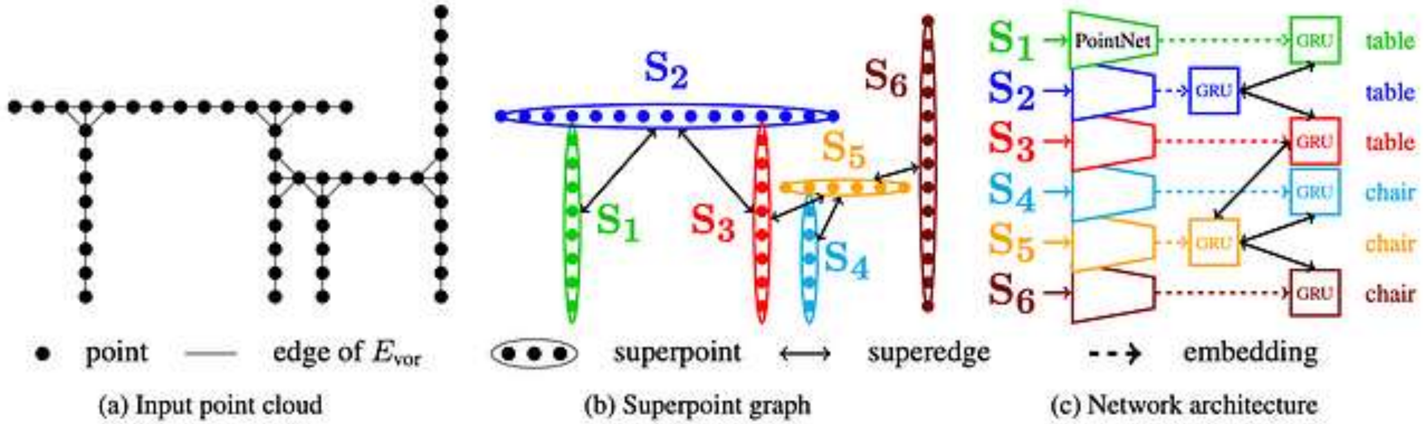
Scene graph generation of visual data



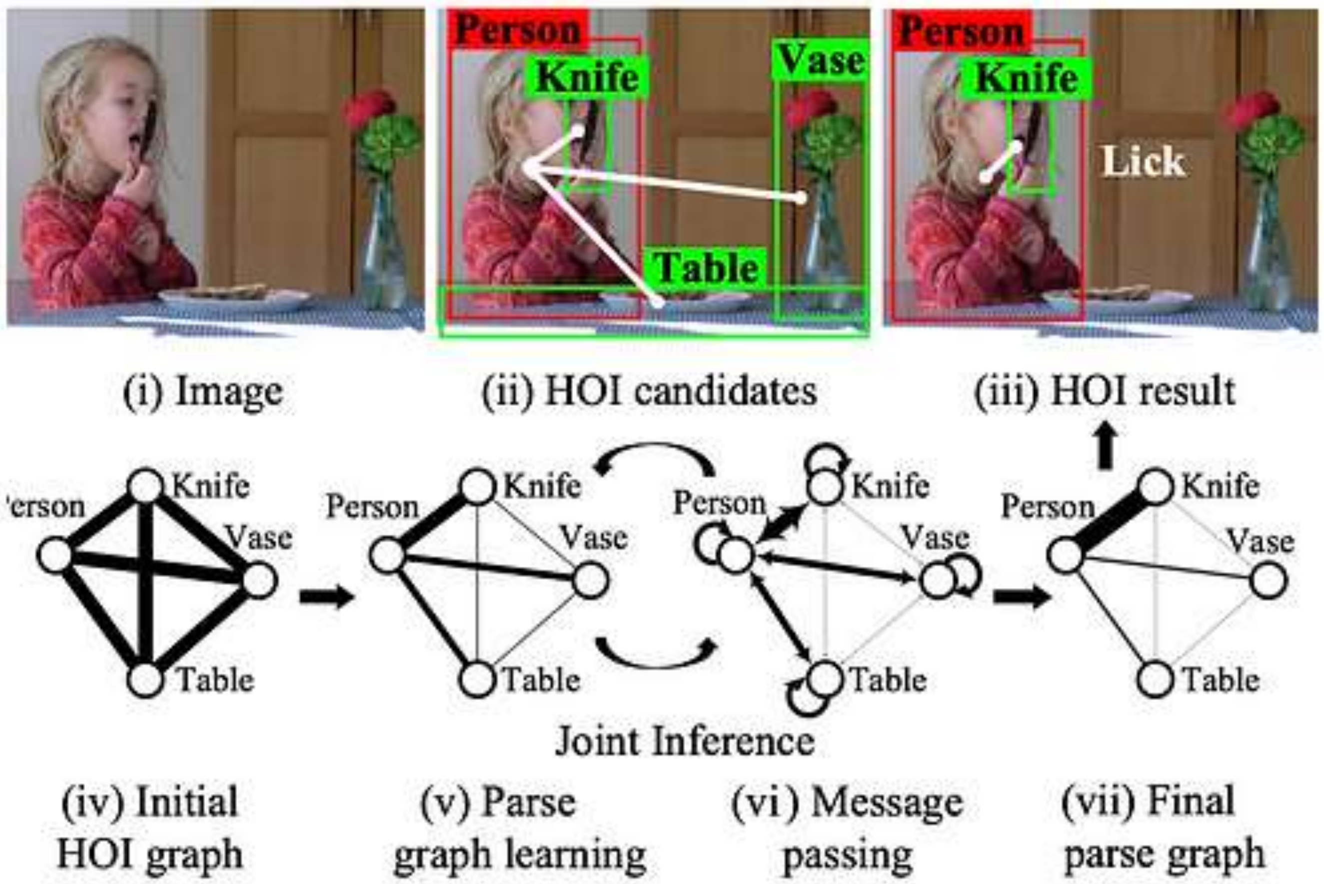


Point cloud classification

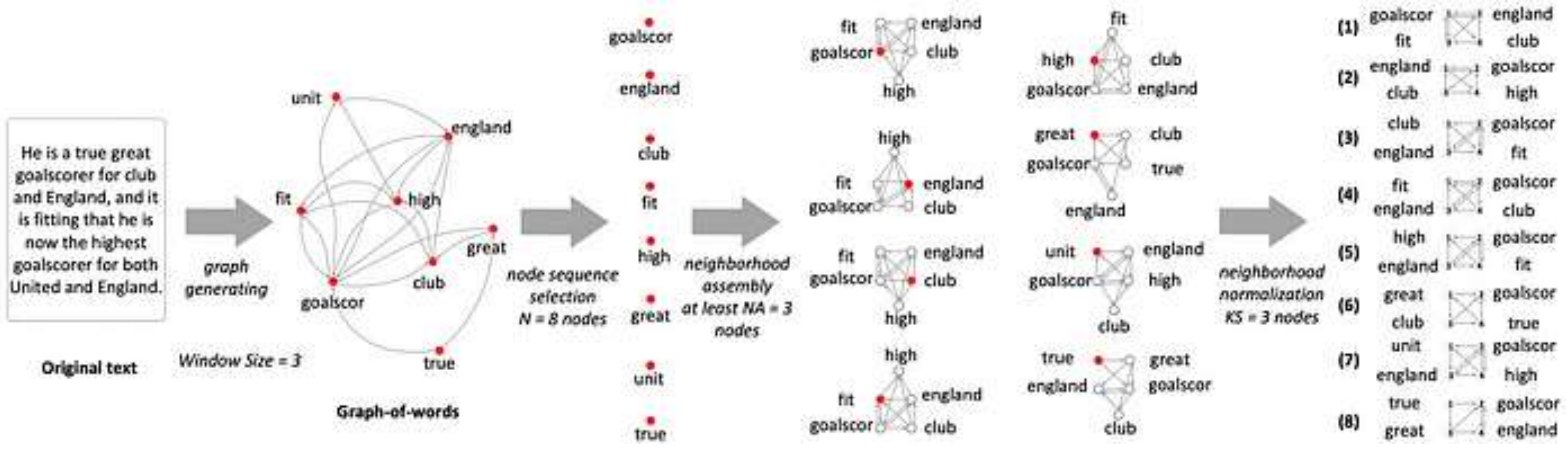




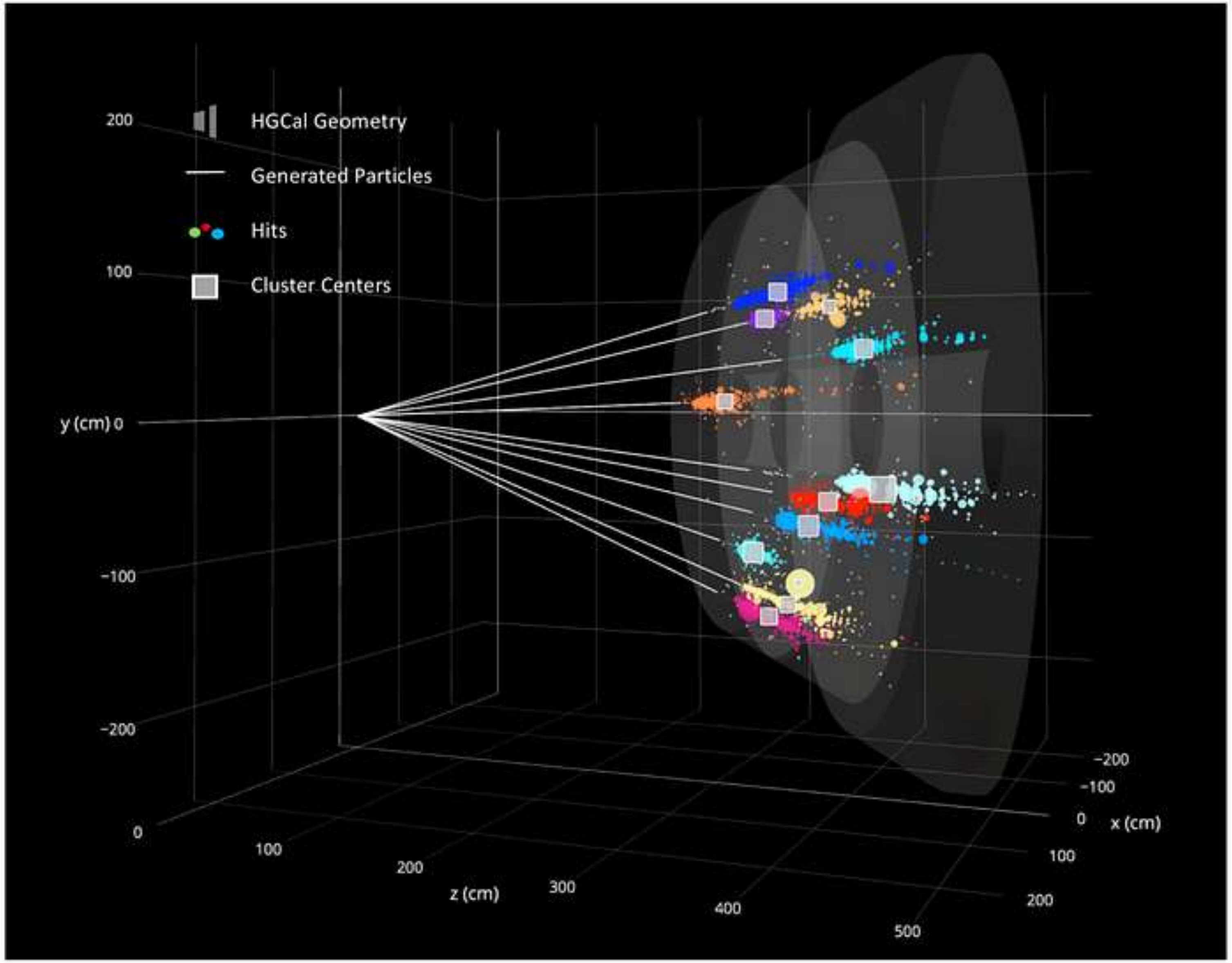
Object interactions



Text classification



Particle physics



Next lecture

Lecture	Title
1	Intro and history of deep learning
3	Deep learning optimization I
5	Convolutional deep learning
7	Graph deep learning
9	Multi-modal deep learning
11	What doesn't work in deep learning
13	Q&A

Lecture	Title
2	AutoDiff
4	Deep learning optimization II
6	Attention-based deep learning
8	From supervised to unsupervised deep learning
10	Generative deep learning
12	Non-Euclidean deep learning
14	Deep learning for videos

Learning and reflection

Understanding Deep Learning: Chapter 13

Thank you!