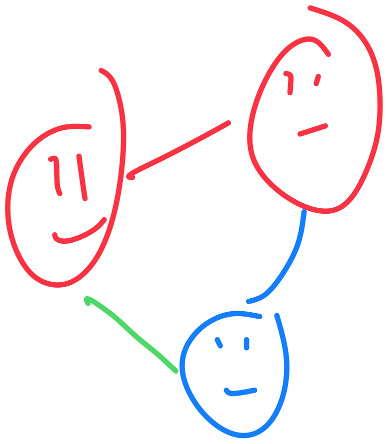
Minhash
&

Data Structures
Graph Fundamentals

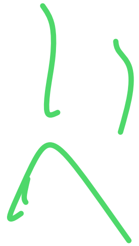
CS 225
Brad Solomon & G Carl Evans

November 10, 2023



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
Department of Computer Science

Plagiarism is not ok



Learning Objectives

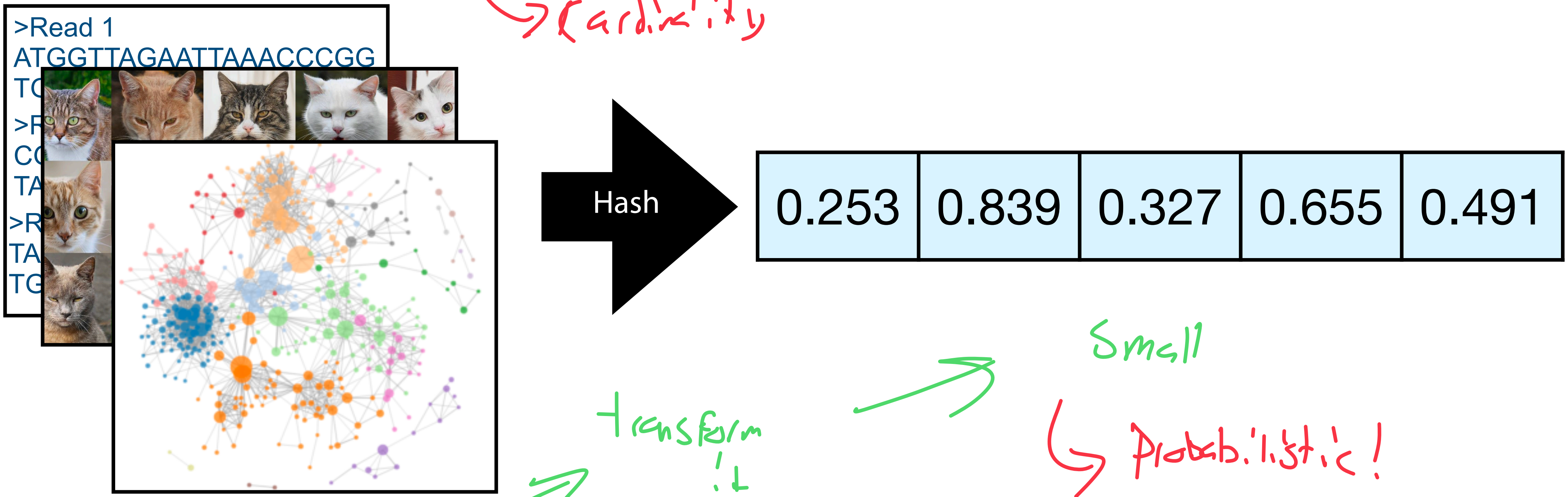
Finish discussing MinHash Sketches

Define graph vocabulary

Discuss graph implementation and storage strategies

Cardinality Sketch

Given any dataset and a SUHA hash function, we can **estimate the number of unique items** by tracking the **k-th minimum hash value**.



Applied Cardinalities

Cardinalities

$$|A|$$

$$|B|$$

$$|A \cup B|$$

$$|A \cap B|$$

\# min value

Set similarities

$$O = \frac{|A \cap B|}{min(|A|, |B|)}$$

$$J = \frac{|A \cap B|}{|A \cup B|}$$

Real-world Meaning

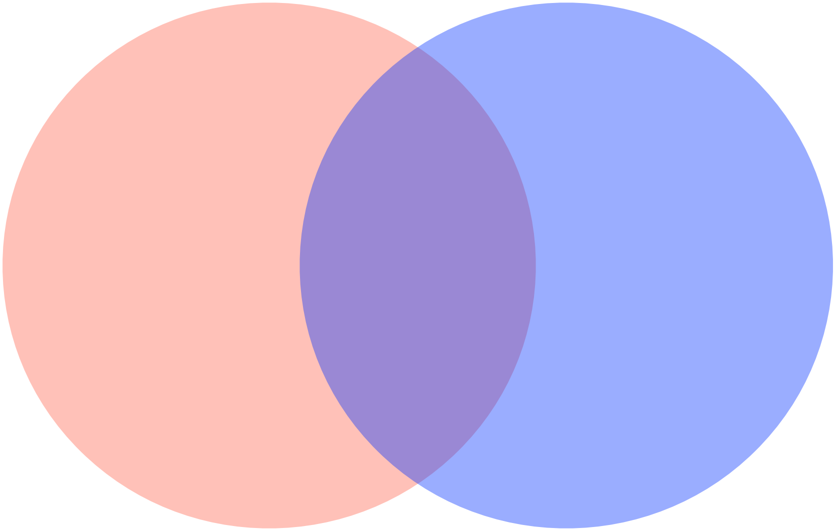
```
AGGCCACAGTGTATTATGACTG
|||||||||||  |||||||||
AGGCCACAGTGAGTTATGACTG

AAAAAAAAAAGATGT-AAGTA
||||||||||||||| |||||
AAAAAAAAAAGATGTAAAGTA

GAGG--TCAGATTCACAGCCAC
||||  ||||||||||||||||
GAGGGGTCAGATTCACAGCCAC
```

Set Similarity Review

To measure **similarity** of A & B, we need both a measure of how similar the sets are but also the total size of both sets.



$$J = \frac{|A \cap B|}{|A \cup B|}$$

J is the ***Jaccard coefficient***

MinHash Sketch

Claim: Under SUHA, set similarity can be estimated by sketch similarity!

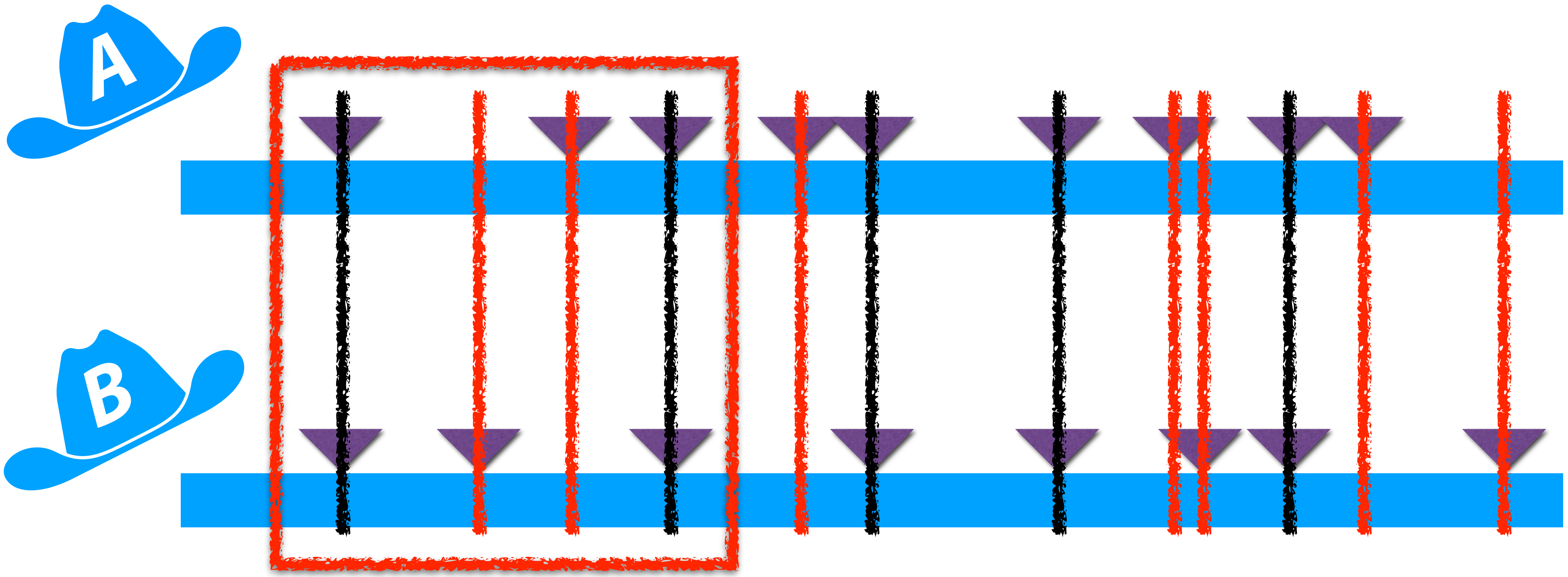


Image inspired by: Ondov B, Starrett G, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. **Mash Screen: high-throughput sequence containment estimation for genome discovery**. *Genome Biol* 20, 232 (2019)

MinHash Jaccard Estimation

Let's assume we have sets A and B sampled uniformly from [0, 100].

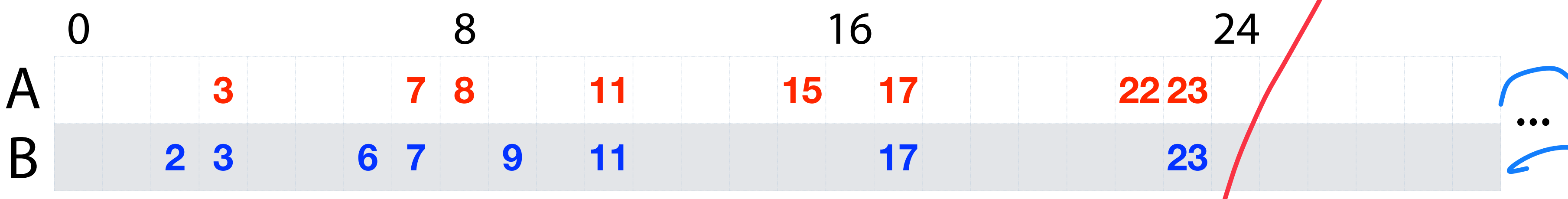
Instead of storing A & B, we store the bottom-8 **MinHash**

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23



MinHash Jaccard Estimation

We dont *know* $|A \cup B|$, but we can estimate it!

Sketch A

3	15
7	17
8	22
11	23

$\cup$

Sketch B

2	9
3	11
6	17
7	23

=

Sketch of $|A \cup B|$

2	8
3	9
6	11
7	15

0 8 16 24

A: 3 7 8 11 15 17 22 23

B: 2 3 6 7 9 11 17 23 …

MinHash Cardinality Estimate

We can estimate the cardinality of the actual sets using our sketches.

Sketch of
$|A \cup B|$

2	8
3	9
6	11
7	15

Our sets sampled from [0, 100].

$$\frac{15}{100} = \frac{8}{N+1}$$

normalize
Kth min

$$N = \left(\frac{800}{15}\right) - 1$$

≈ 52.3 unique items

MinHash Indirect Jaccard Estimation

$$\frac{|A| \cap |B|}{|A| \cup |B|} = \frac{|A| + |B| - |A \cup B|}{|A \cup B|}$$

$k = 8$ MinHash sketches

Our sets sampled from [0, 100]

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23

Sketch of $|A \cup B|$

2	8
3	9
6	11
7	15

$$= \frac{(800/23-1) + (800/23-1) - (800/15-1)}{800/15-1}$$

$$= \frac{34.782 + 34.782 - 53.333 - 1}{53.333 - 1} \approx 0.29$$

Jaccard

MinHash Direct Jaccard Estimate

We can also estimate cardinality directly using our sketches!

Sketch A

3	15
7	17
8	22
11	23

Sketch B

2	9
3	11
6	17
7	23

Intersection

3	23
7	
11	
17	

Union

2	8	17
3	9	22
6	11	23
7	15	

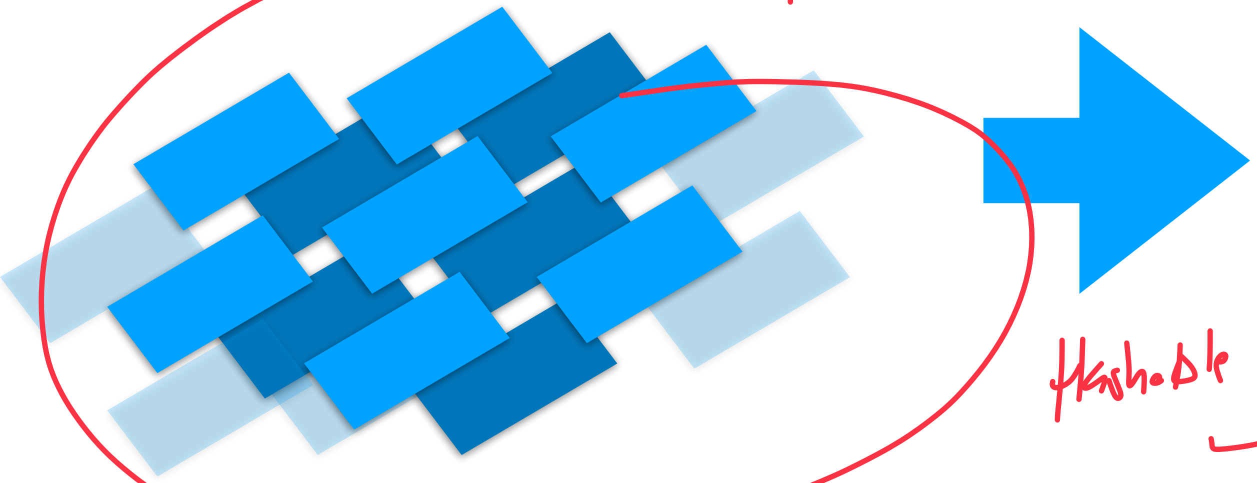
Accurate subsample

SUHA!

5/11 ≈ Jaccard!

MinHash Sketch

We can convert any hashable dataset into a **MinHash sketch**



{1 , 3 , 5}

Hashable → store K hash values/indices

We lose our original dataset, but we can still estimate two things:

1. Cardinality (K min hash is best estimate)

2. Similarity (Directly or indirectly)

Alternative MinHash Sketch Approaches

The **easiest** version of MinHash uses k hashes. How might this work?

Lab: bloom → vector (hash functions)

hash each one & build MH

↳ track global min for each hash

1) I need k hashes → k hash independence very hard

1) Sequence decomposed into **kmers**

S_1 : CATGGACCGACCAG

CAT GAC GAC
ATG ACC ACC
TGG CCG CCA
GGA CGA CAG

GCAGTACCGATCGT : S_2

GTA CGA CGT
AGT CCG TCG
CAG ACC ATC
GCA TAC GAT

2) Multiple hash functions (**Γ**) map kmers to values.

Γ_1	Γ_2	Γ_3	Γ_4	
19	14	57	36	CAT
14	57	36	19	ATG
58	37	16	15	TGG
40	23	2	61	GGA
33	28	11	54	GAC
5	48	47	26	ACC
22	1	60	43	CCG
24	7	50	45	CGA
33	28	11	54	GAC
5	48	47	26	ACC
20	3	62	41	CCA
18	13	56	39	CAG

	Γ_1	Γ_2	Γ_3	Γ_4
GCA	36	19	14	57
CAG	18	13	56	39
AGT	11	54	33	28
GTA	44	27	6	49
TAC	49	44	27	6
ACC	5	48	47	26
CCG	22	1	60	43
CGA	24	7	50	45
GAT	35	30	9	52
ATC	13	56	39	18
TCG	54	33	28	11
CGT	27	6	49	44

3) The smallest values for each hash function is chosen

[5, 1, 2, 15]
Sketch (S_1)

[5, 1, 6, 6]
Sketch (S_2)

4) The Jaccard similarity can be estimated by the overlap in the **Min**imum **Hash**es (**MinHash**)

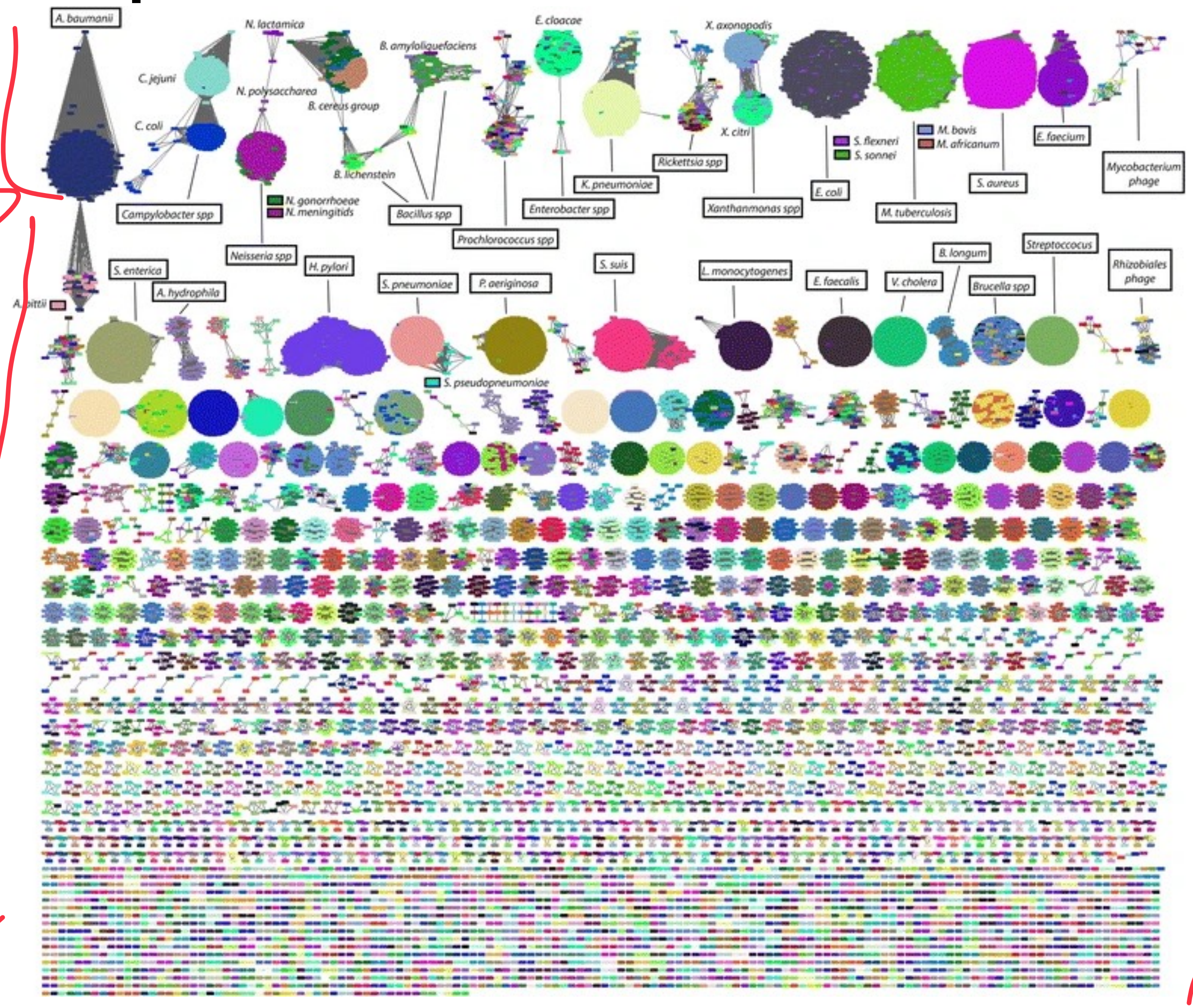
$J(S_1, S_2) \approx 2/4 = 0.5$

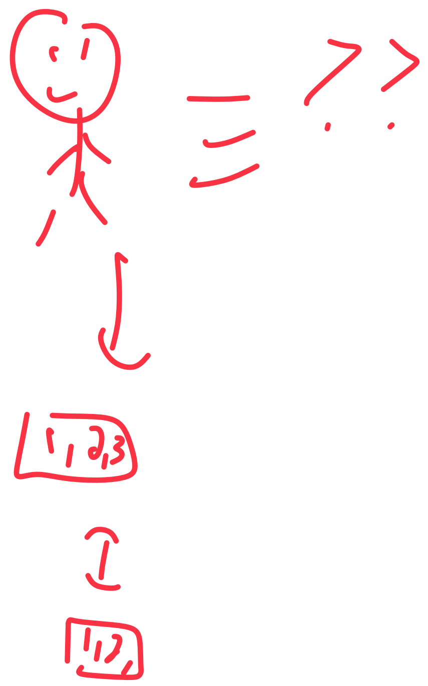
```
S1 : CATGGACCGACCAG
       |  |||||  |
S2 : GCAGTACCGATCGT
```

Assembling large genomes with single-molecule sequencing and locality-sensitive hashing
Berlin et al (2015) *Nature Biotechnology*

MinHash in practice

graphs!





A lot of graphs

Mash: fast genome and metagenome distance estimation using MinHash
Ondov et al (2016) *Genome Biology*

Alternative MinHash Sketch Approaches

What if I have a dataset which is **much** larger than another?

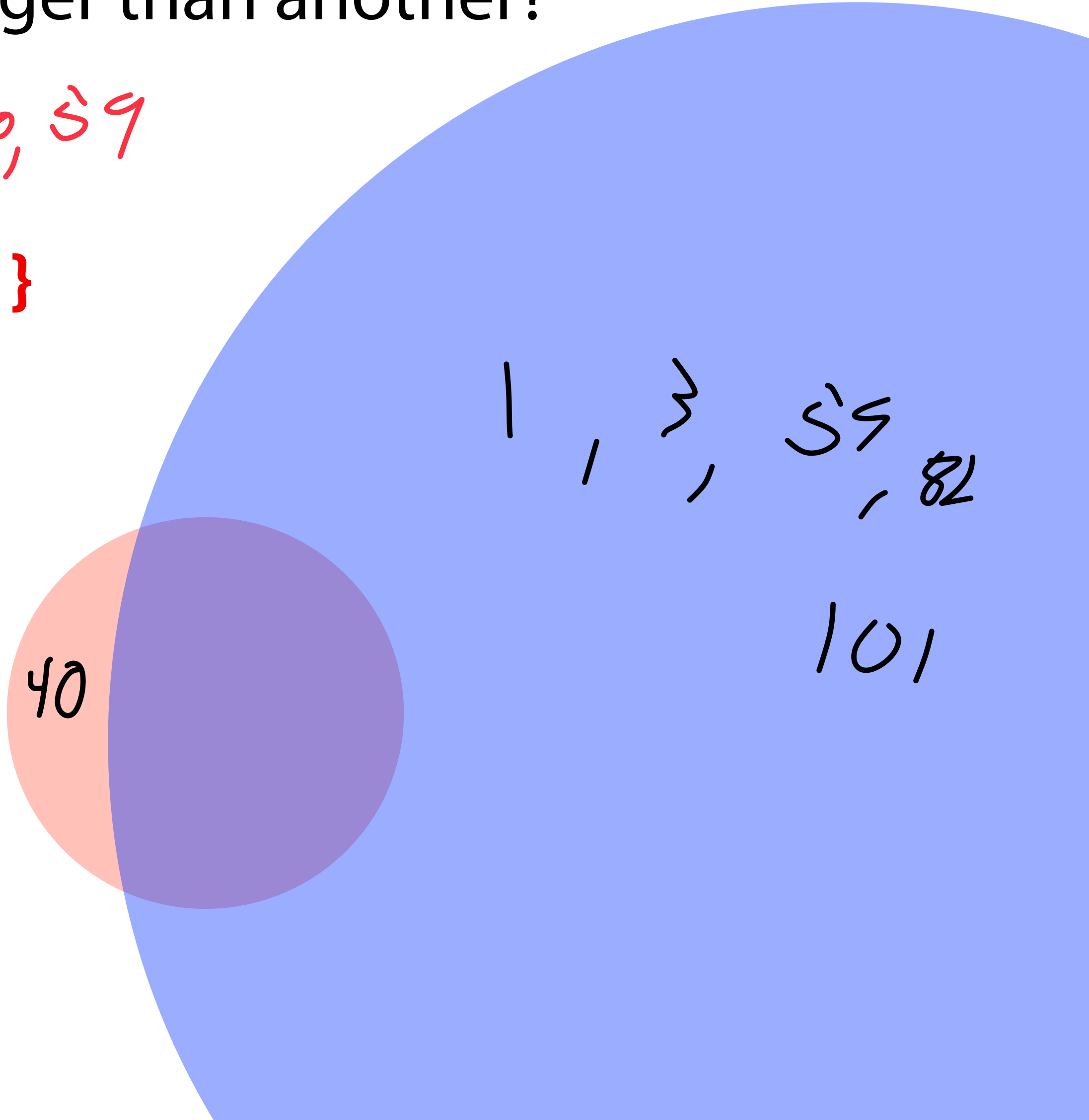
$S_1 = \{ 1, 3, 40, 59, 82, 101 \}$ → 1, 3, 40, 59

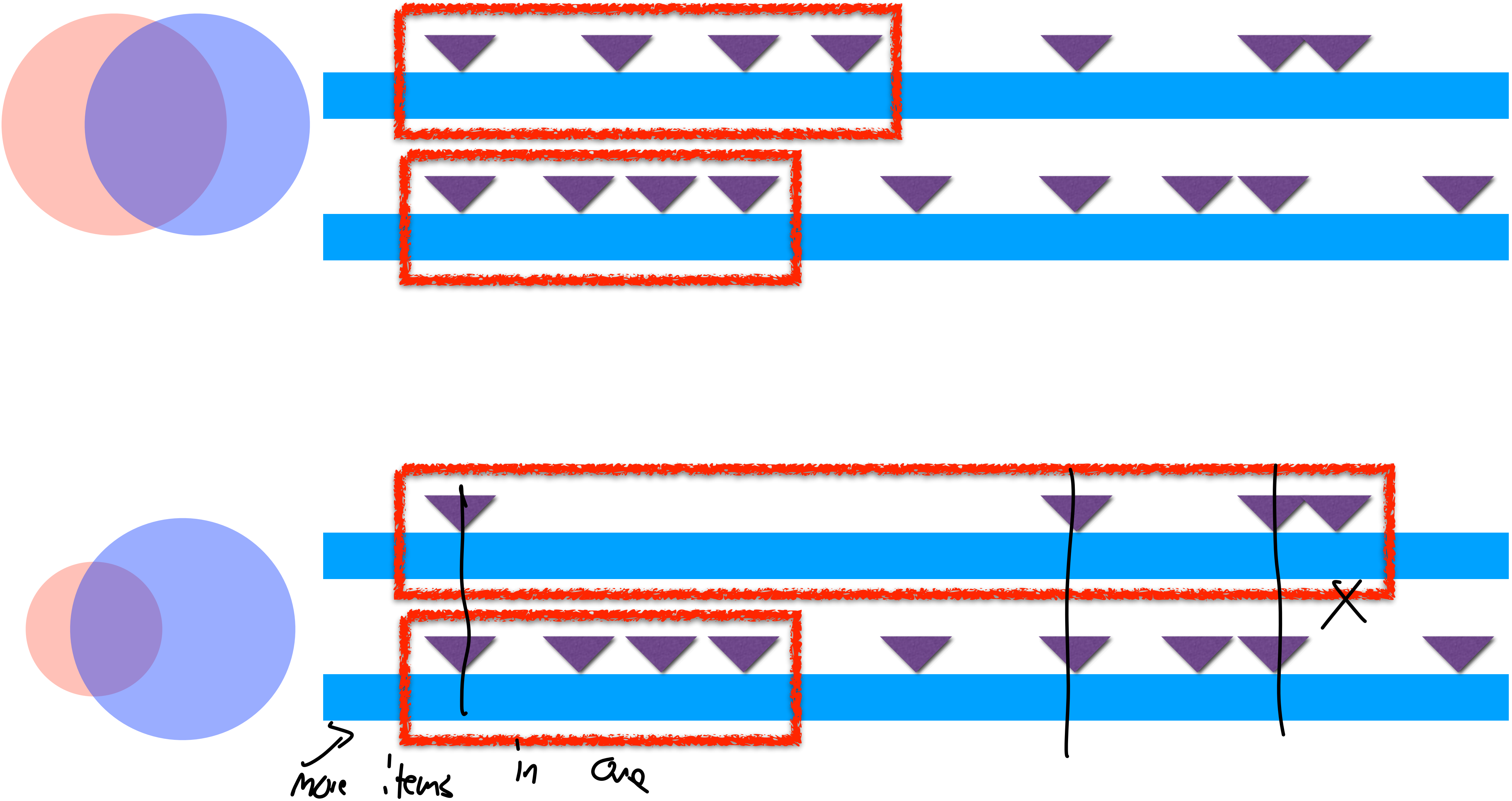
$S_2 = \{ 1, 2, 3, 4, 5, 6, 7, \ldots 59, 82, 101, \ldots \}$

K = 4 MH

S_2 → (1, 2, 3, 4)

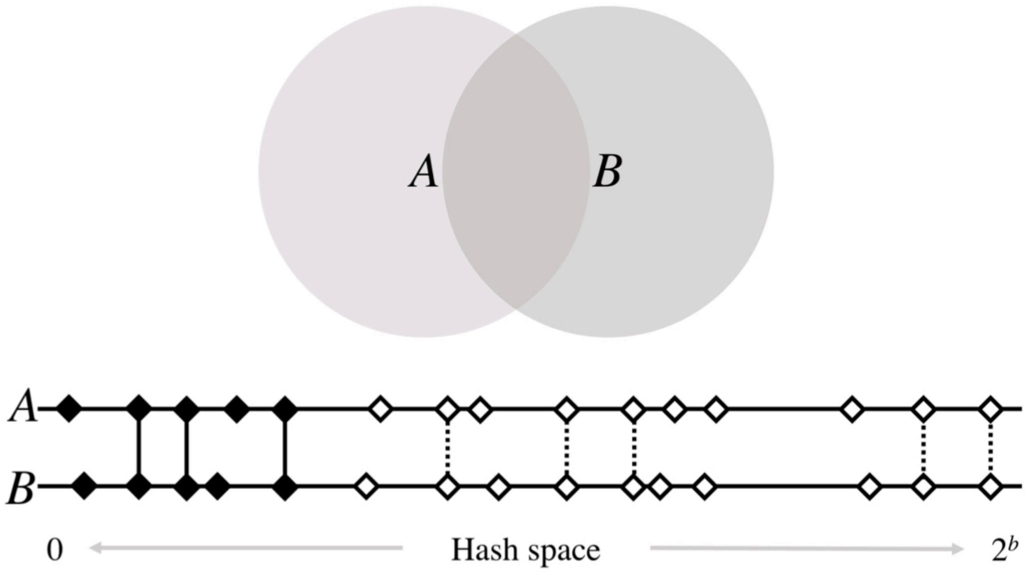
↳ Bottom K minhash!

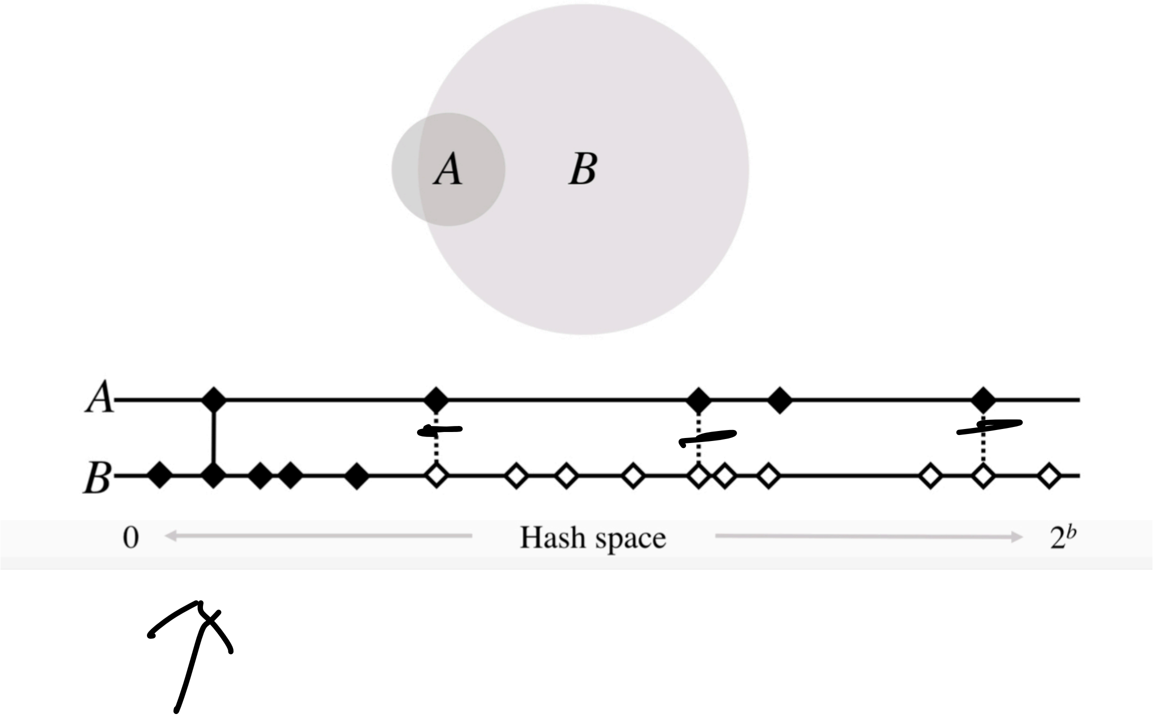



more items in one

Alternative MinHash sketches

Bottom-k minhash has low accuracy if the cardinality of sets are skewed

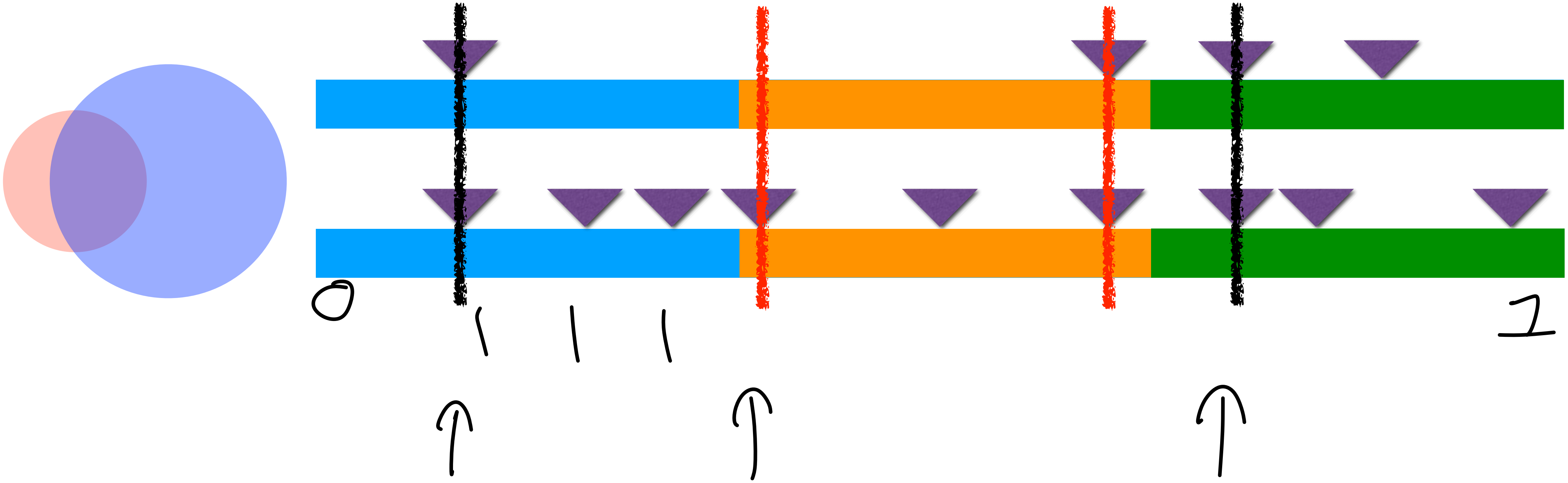




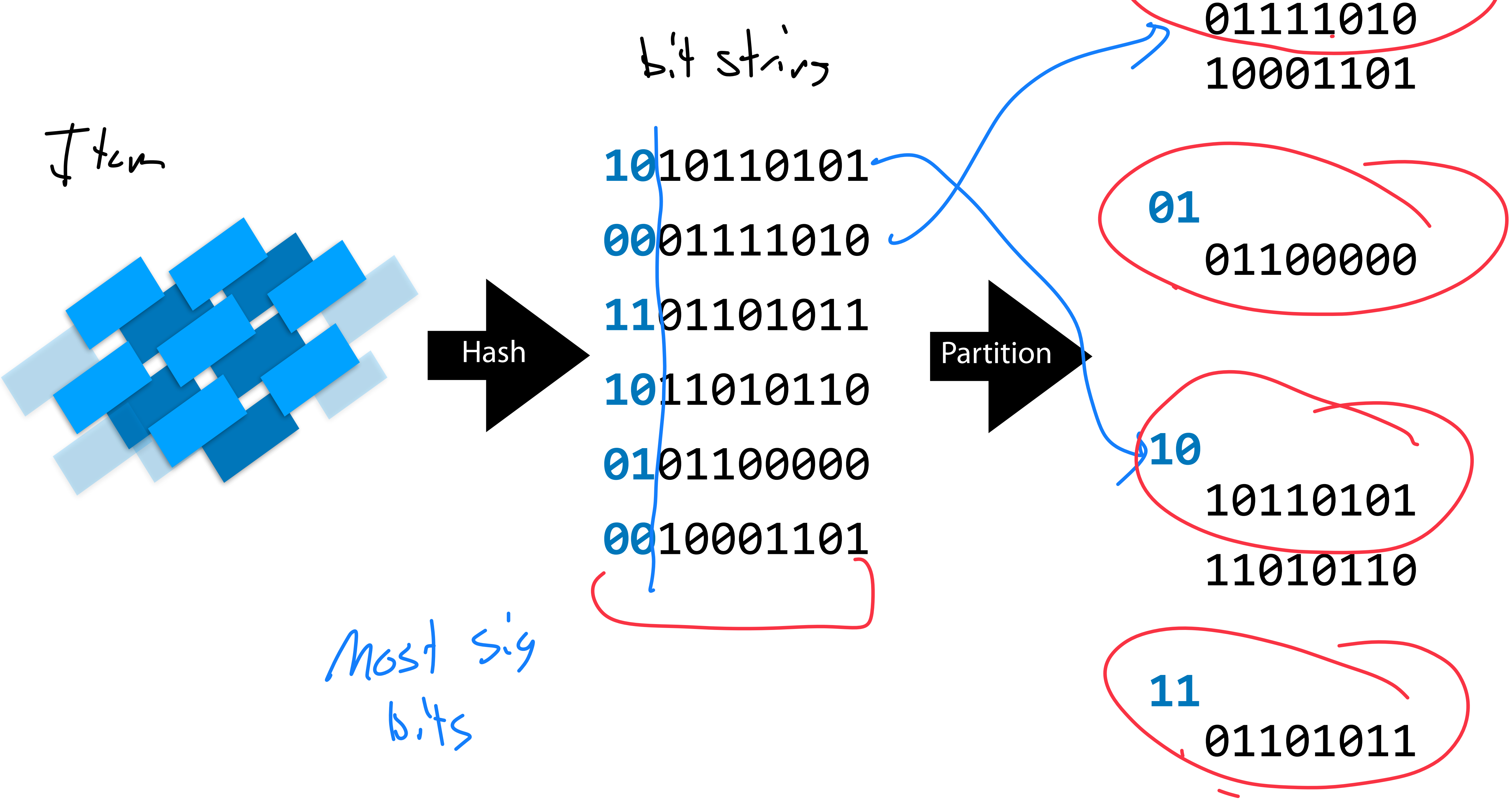
Ondov, Brian D., Gabriel J. Starrett, Anna Sappington, Aleksandra Kostic, Sergey Koren, Christopher B. Buck, and Adam M. Phillippy. **Mash Screen: High-throughput sequence containment estimation for genome discovery**. *Genome biology* 20.1 (2019): 1-13.

Alternative MinHash Sketch Approaches

If there is a large cardinality difference, **use k-partitions!**



K-Partition Minhash



K-Partition Minhash

Hint: What bitwise operator(s) will allow me to do this?

```
1010110101              10
            Partition →
0001111010              00
```

Be careful not to del strings

>> This

3 bits 0001

Bitmasks

<< Much harder

What information do I need to do this in general?

The # of bits in my partition

# of bits total in my bitstring

MP_Sketching: A MinHash experiment

Using legitimate hashes, write MinHash sketch three ways:

```
std::vector<uint64_t> khash_minhash(std::vector<int> inList, std::vector<hashFunction> hv);
```

k different

```
std::vector<uint64_t> kminhash(std::vector<int> inList, unsigned k, hashFunction h);
```

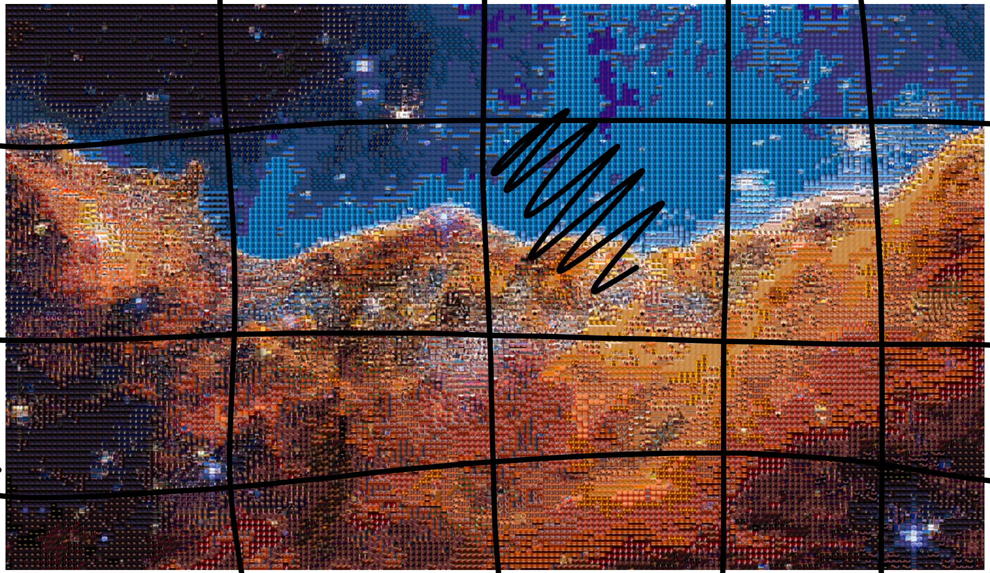
Bottom k minhash (1 hash)

```
std::vector<uint64_t> kpartition_minhash(std::vector<int> inList, int part_bits, hashFunction h);
```

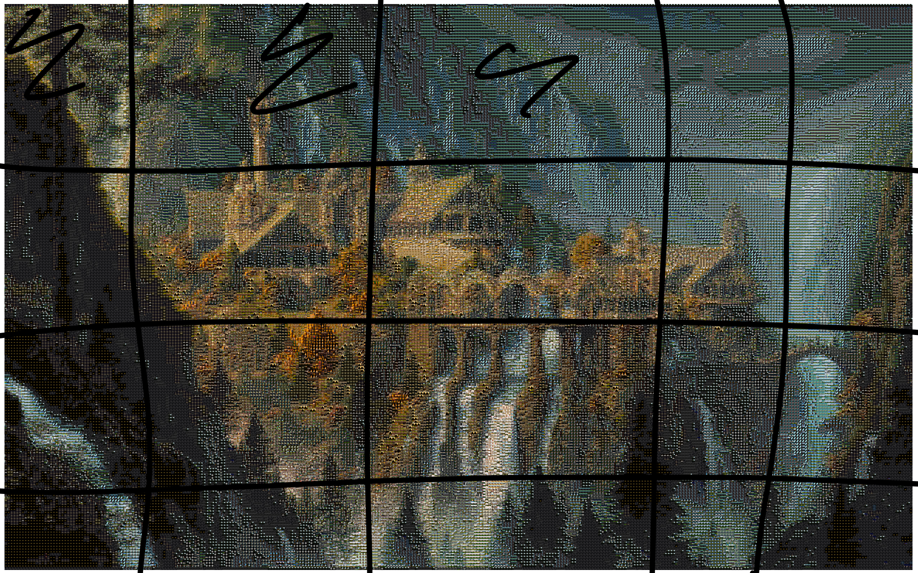
00
01
10
11

MP_Sketching: A MinHash experiment

Use MinHash sketches to estimate PNG similarity



Mosaics (Discord: Bose)

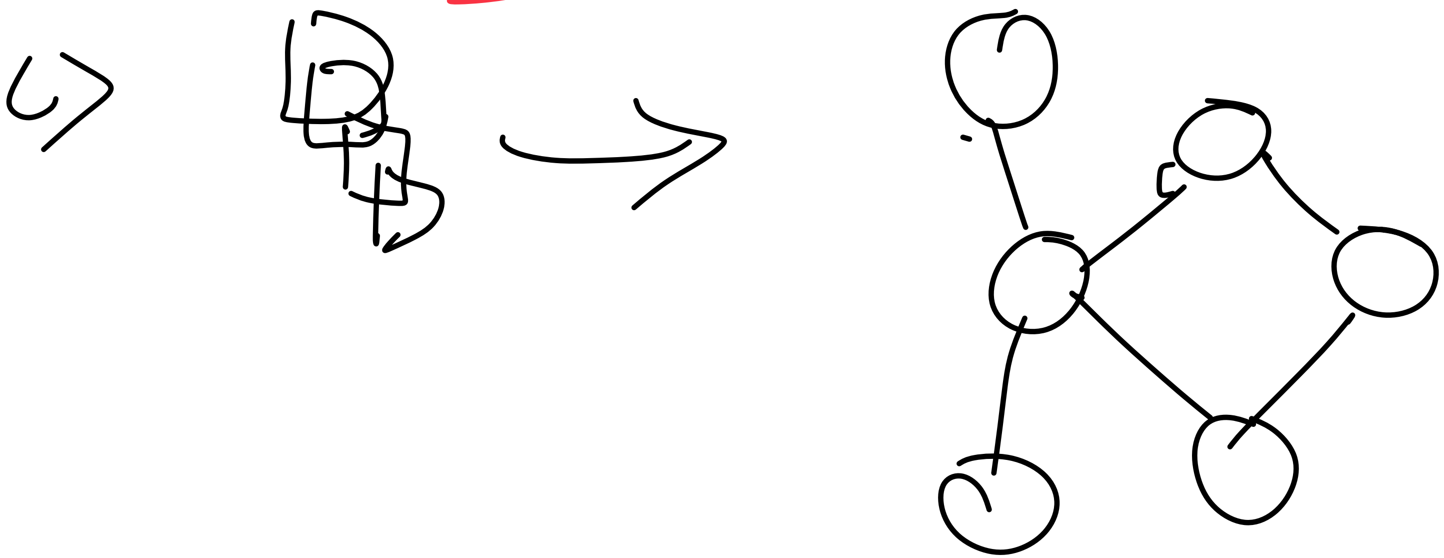


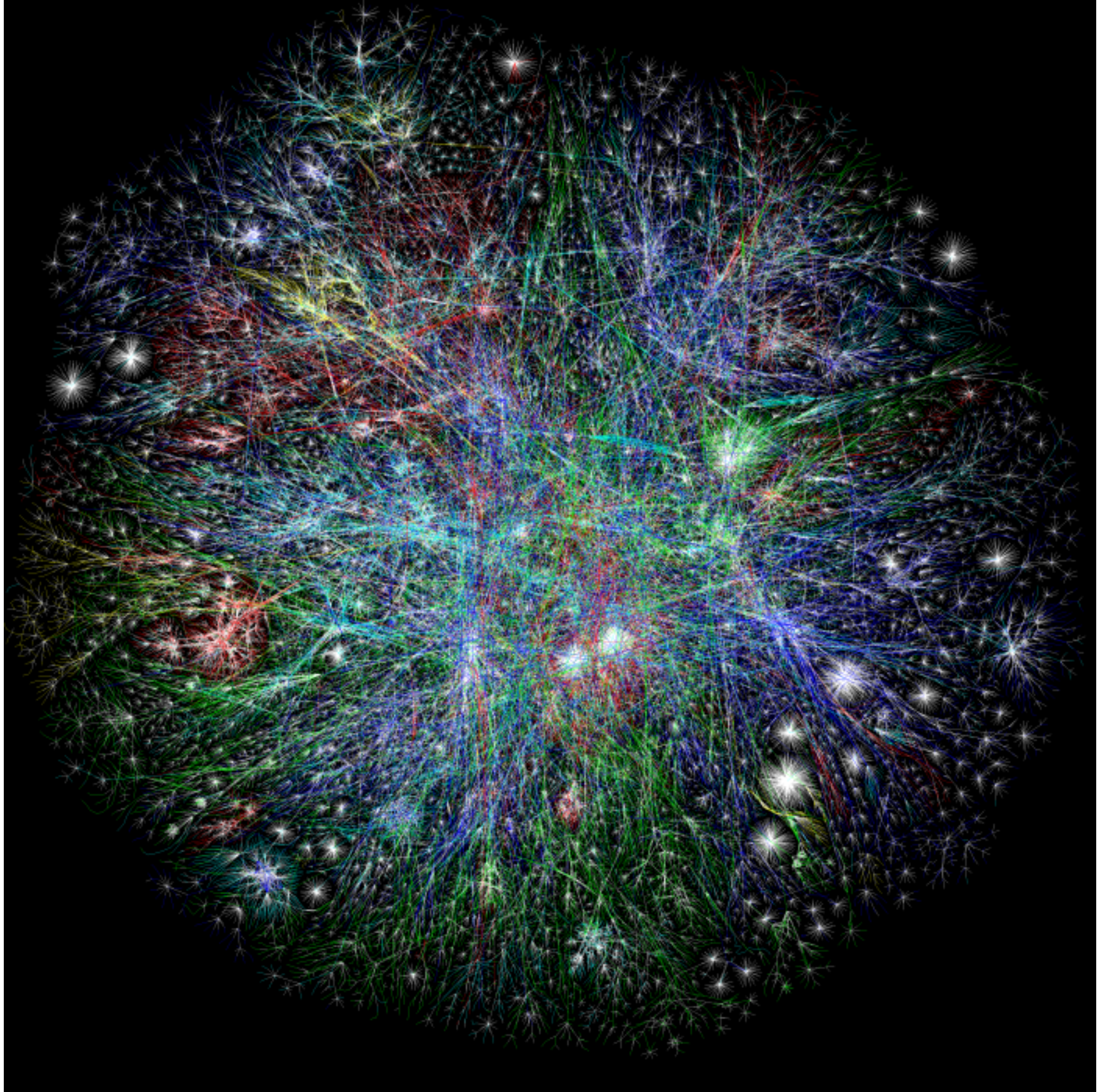
Mosaics (Discord: LightningStorm)

PNGs → RGB → BW Luminance → MinHash sketch

MP_Sketching: A MinHash experiment

Build a weighted graph of every possible pairwise comparison!



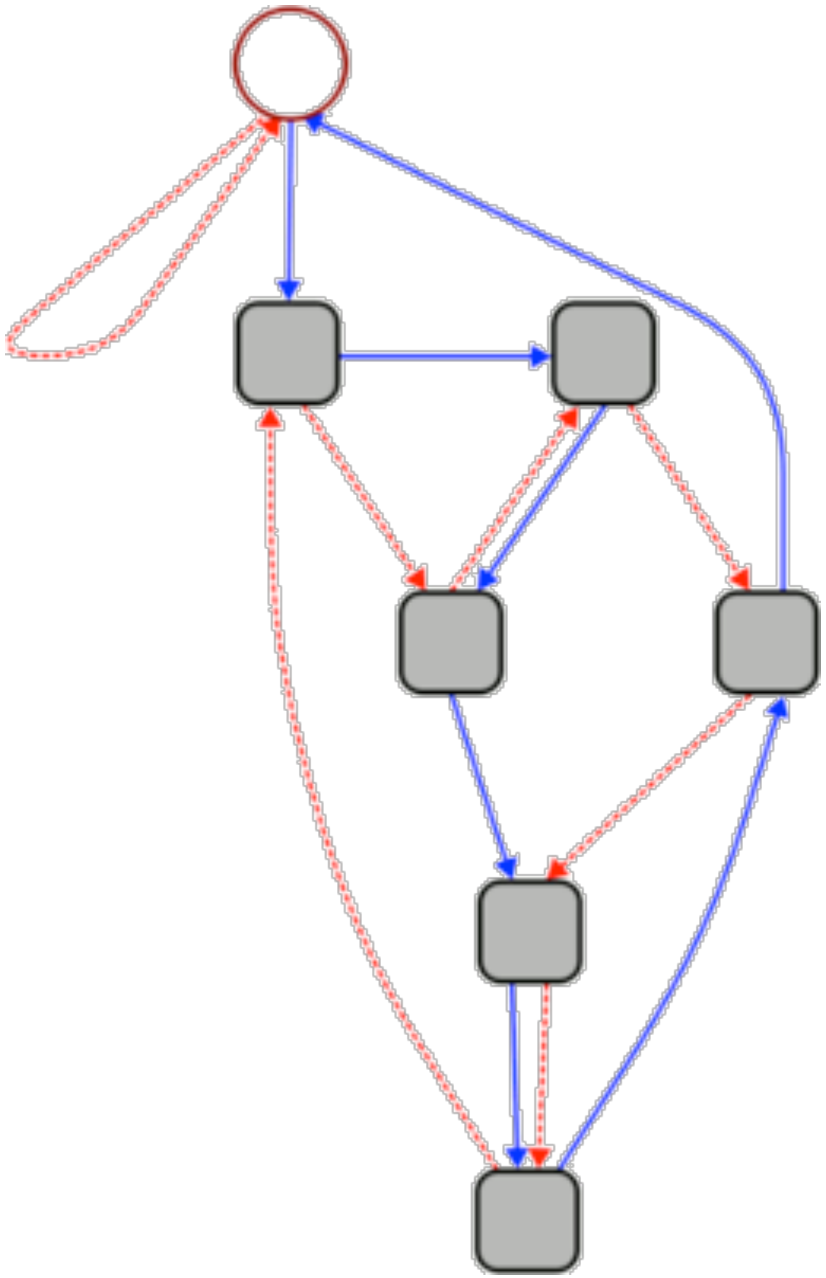


Nodes: Routers and servers

Edges: Connections

The Internet 2003

The OPTE Project (2003)



This graph can be used to quickly calculate whether a given number is divisible by 7.

1. Start at the circle node at the top.
2. For each digit **d** in the given number, follow **d** blue (solid) edges in succession. As you move from one digit to the next, follow **1** red (dashed) edge.
3. If you end up back at the circle node, your number is divisible by 7.

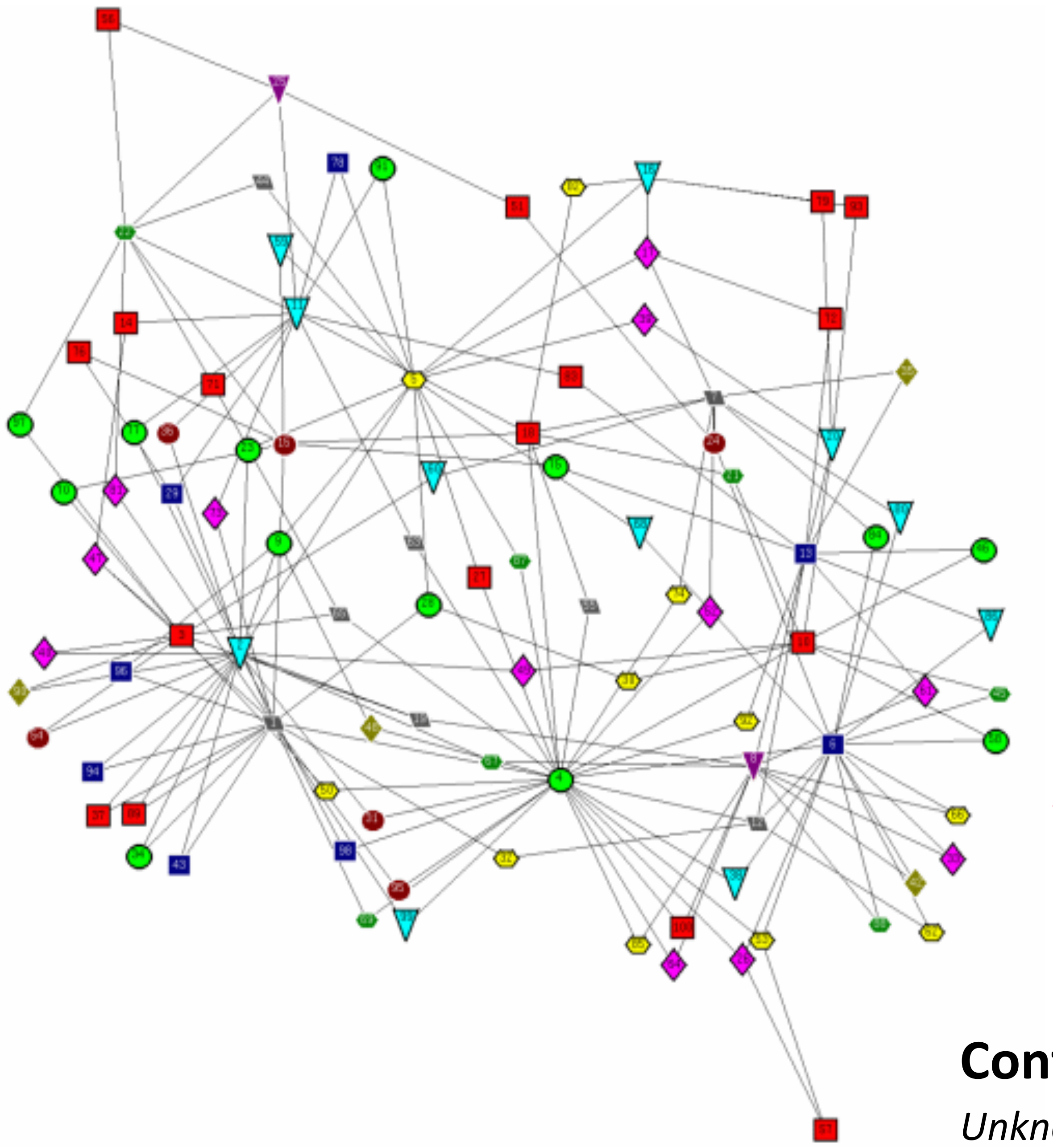
3703

Structure has meaning!

"Rule of 7"

Unknown Source

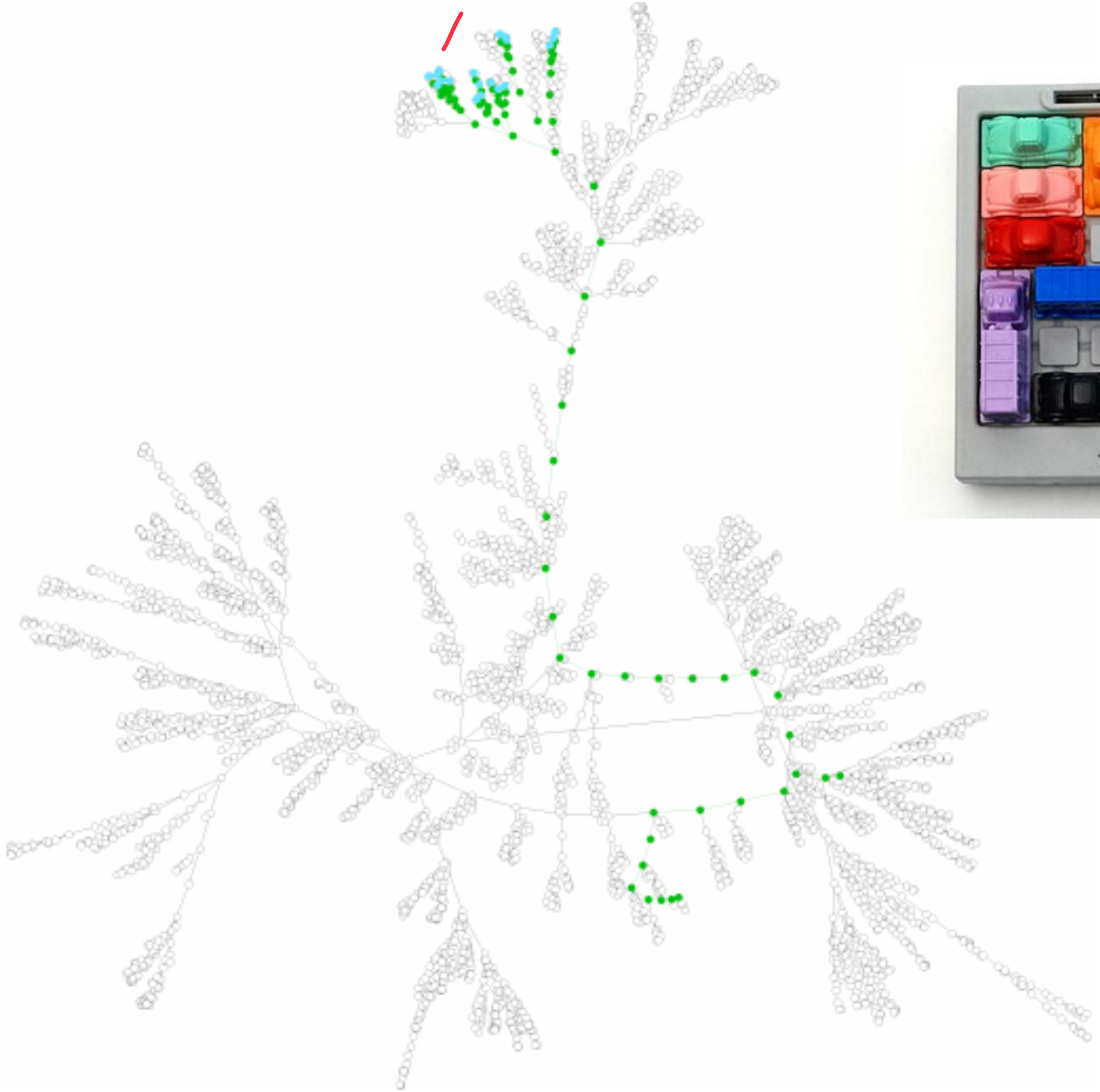
Presented by Cinda Heeren, 2016



Conflict-Free Final Exam Scheduling Graph

Unknown Source

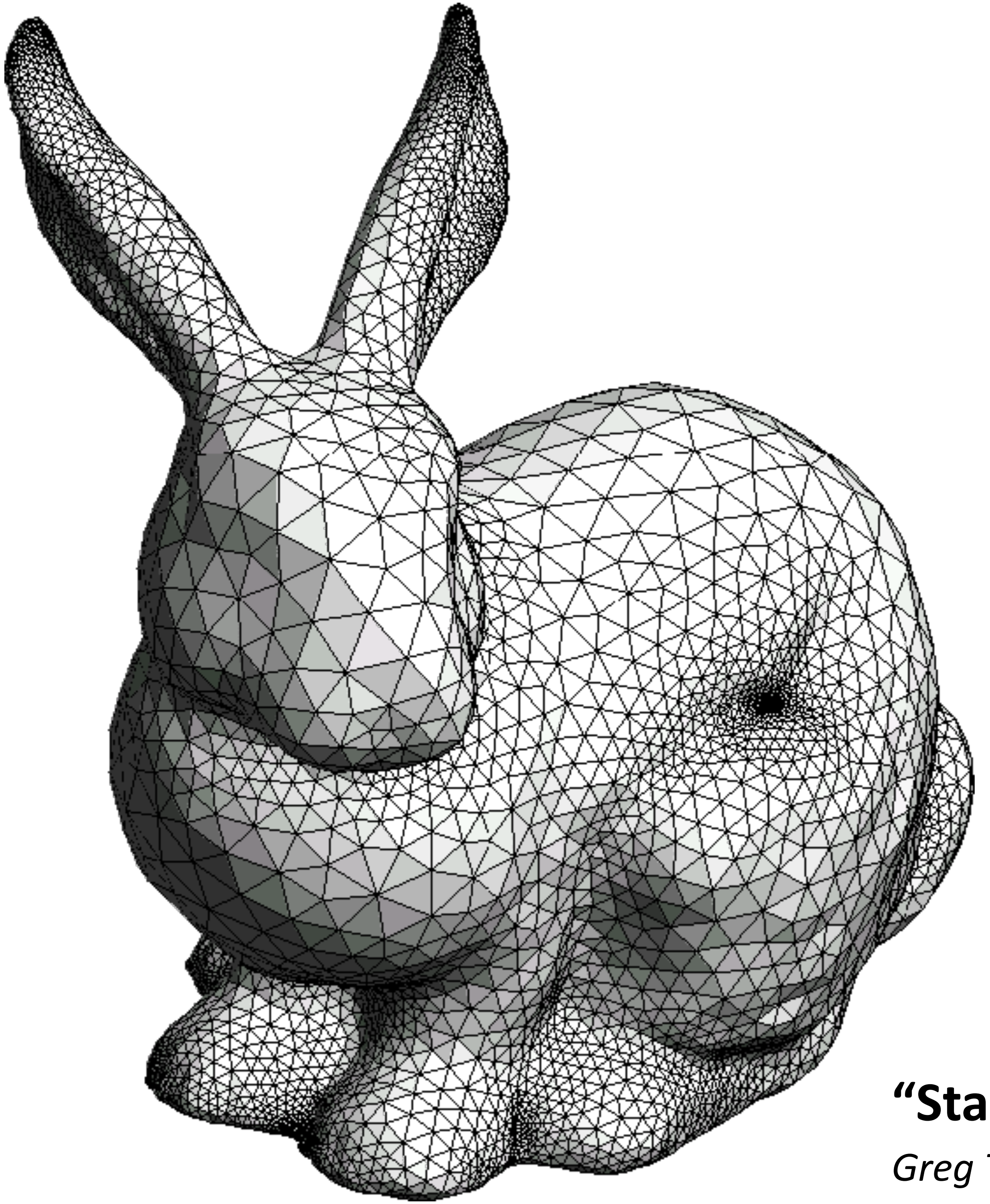
Presented by Cinda Heeren, 2016



“Rush Hour” Solution

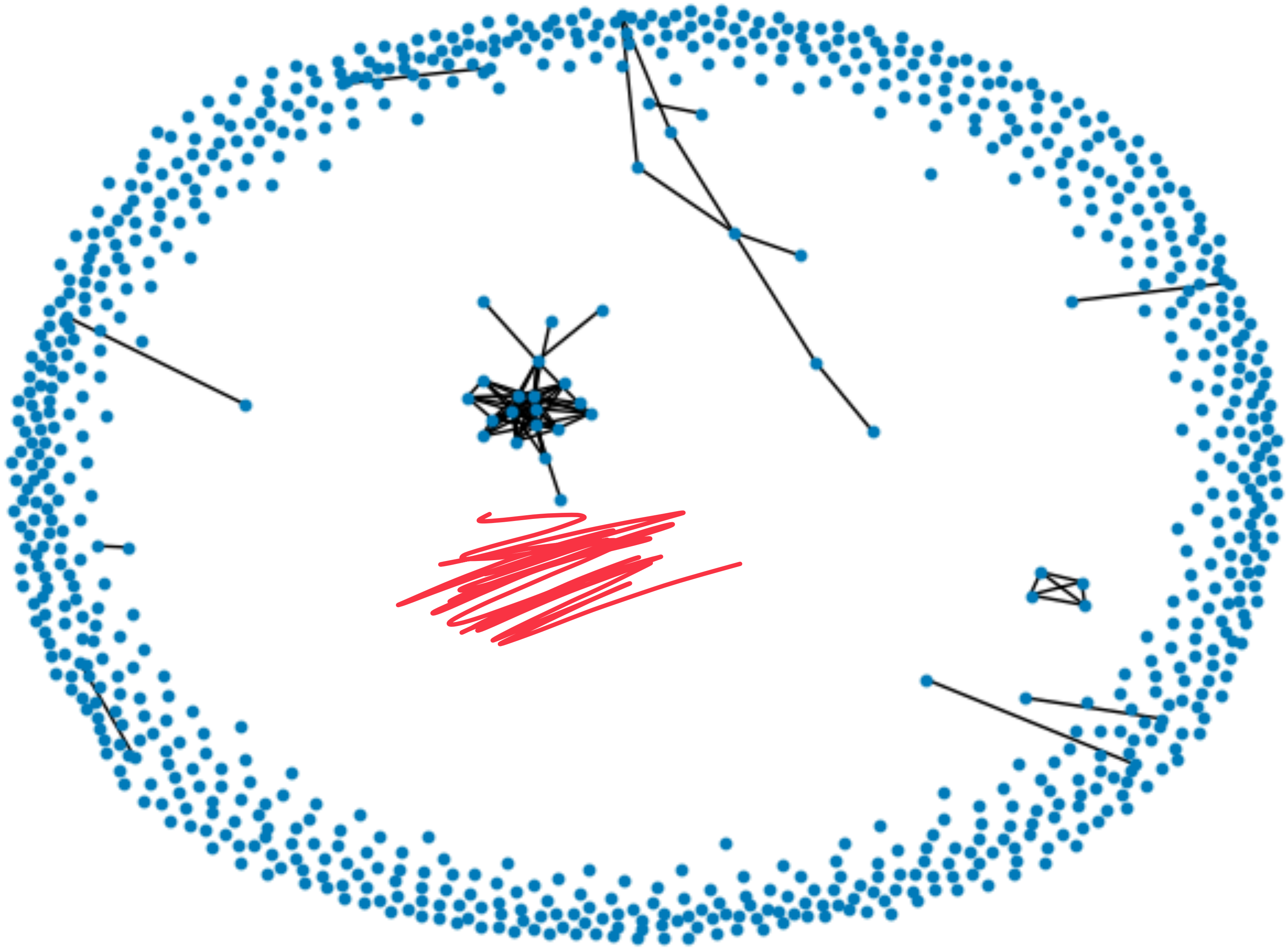
Unknown Source

Presented by Cinda Heeren, 2016

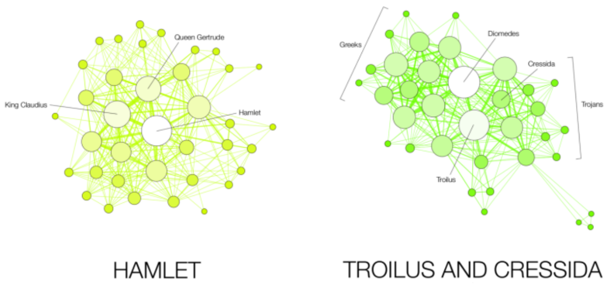


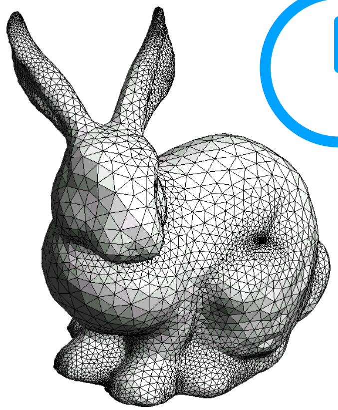
“Stanford Bunny”

Greg Turk and Mark Levoy (1994)



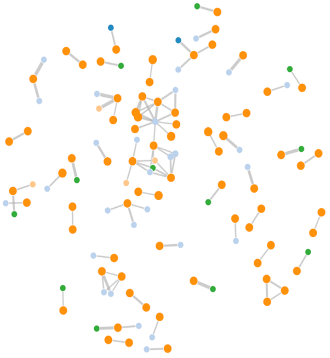
Graphs

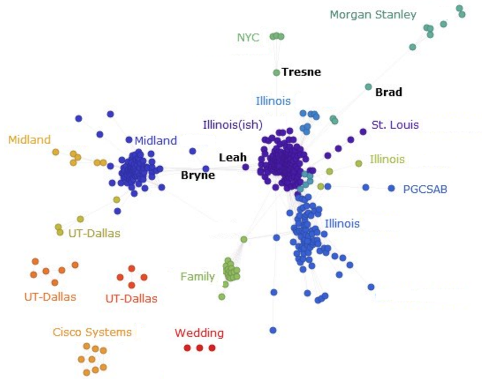


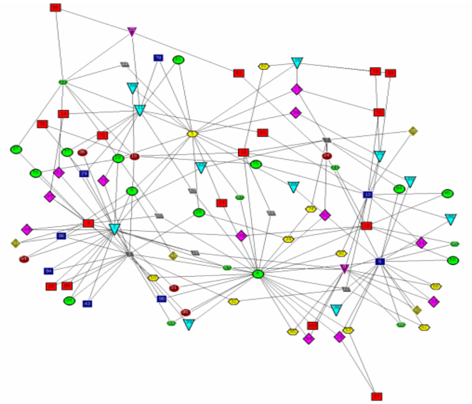


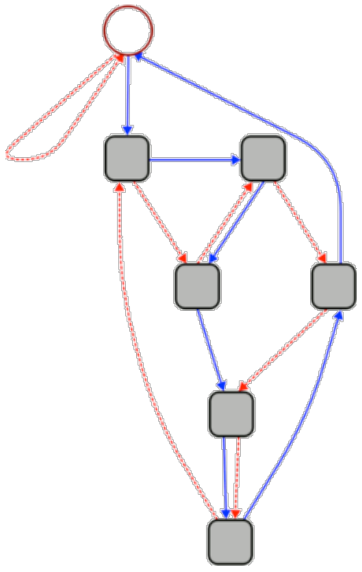
To study all of these structures:

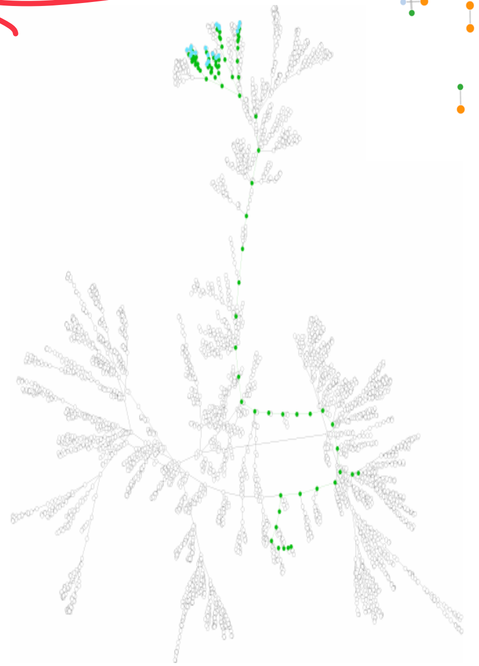
1. A common vocabulary
2. Graph implementations
3. Graph traversals
4. Graph algorithms







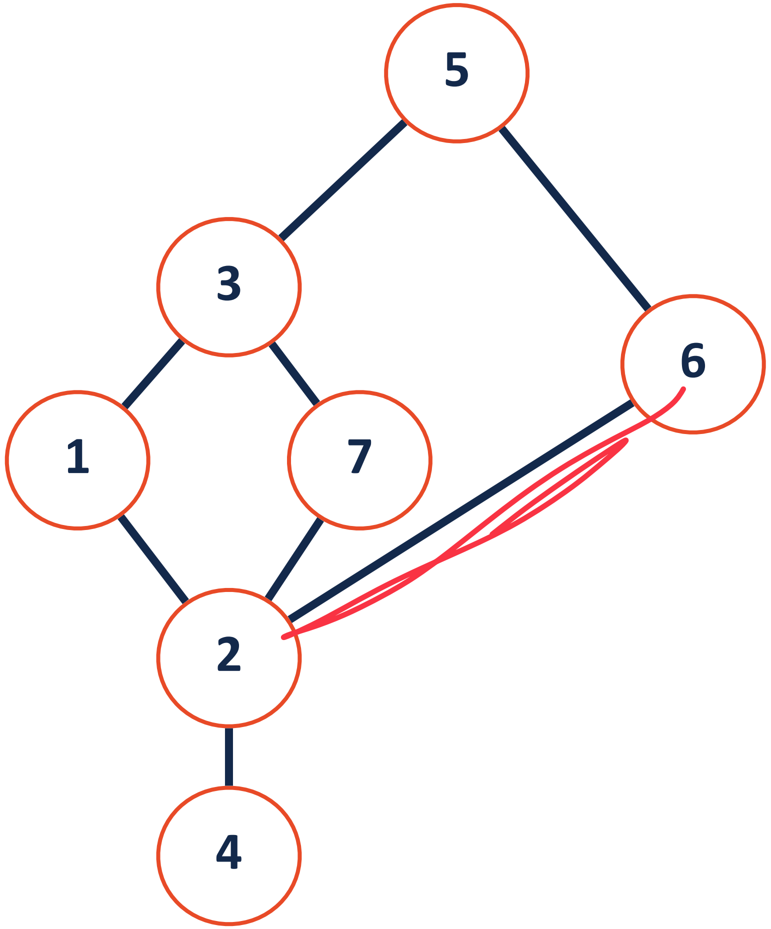




Graph Vocabulary

$G = (V, E)$

A **graph** is a data structure containing a set of vertices and a set of edges



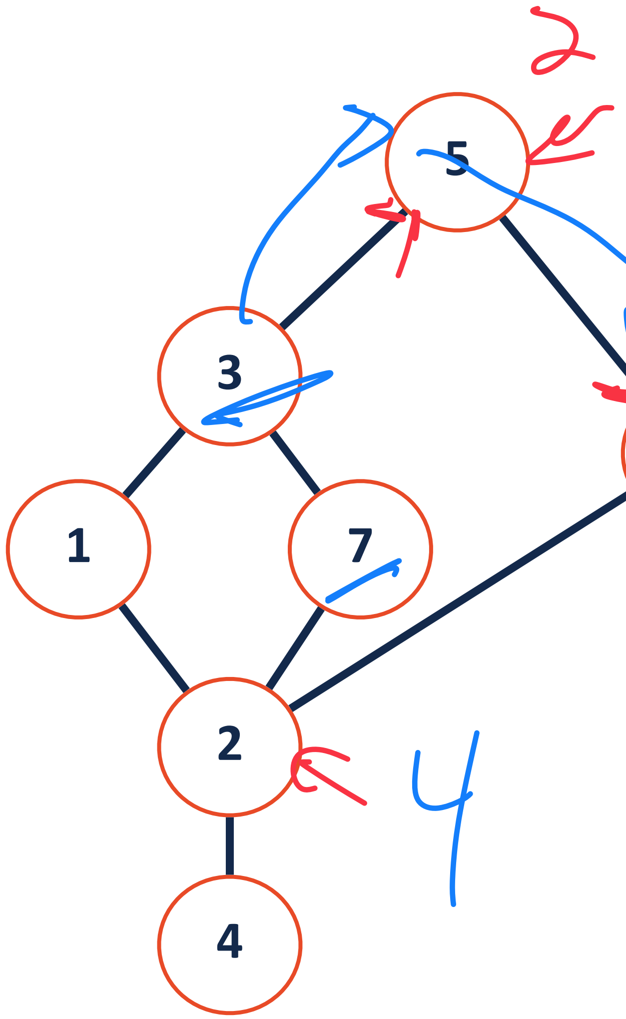
Vertex: Nodes

↳ store value most of the time

Edges: Connections between nodes

2,6 ⟺ 6,2

Graph Vocabulary



Degree: # of edges touching a vertex

↳ Out degree = 2

↳ In degree = 1

Adjacency: Two vertices are adjacent if they are connected by an edge

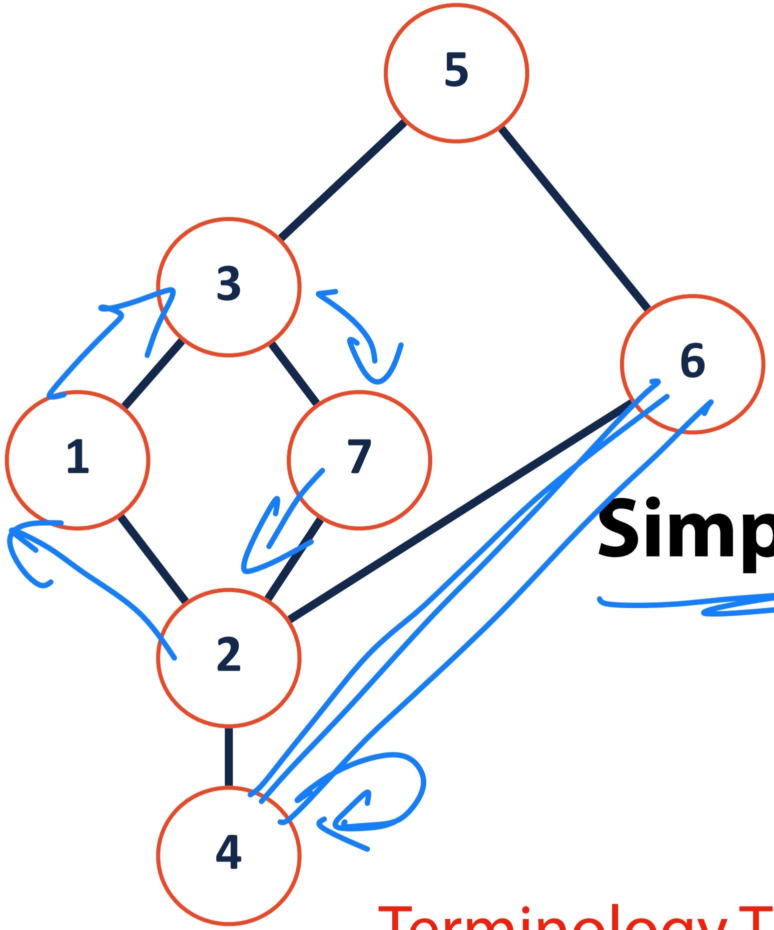
↳ 3,7 adjacent

3,6 Not adjacent

Path: A sequence of vertices (or edges) between two nodes

Graph Vocabulary

A graph has **no root** and **may contain cycles**



Cycle: A path from a node to itself

Simple Graph: No self-loops or multi-edges

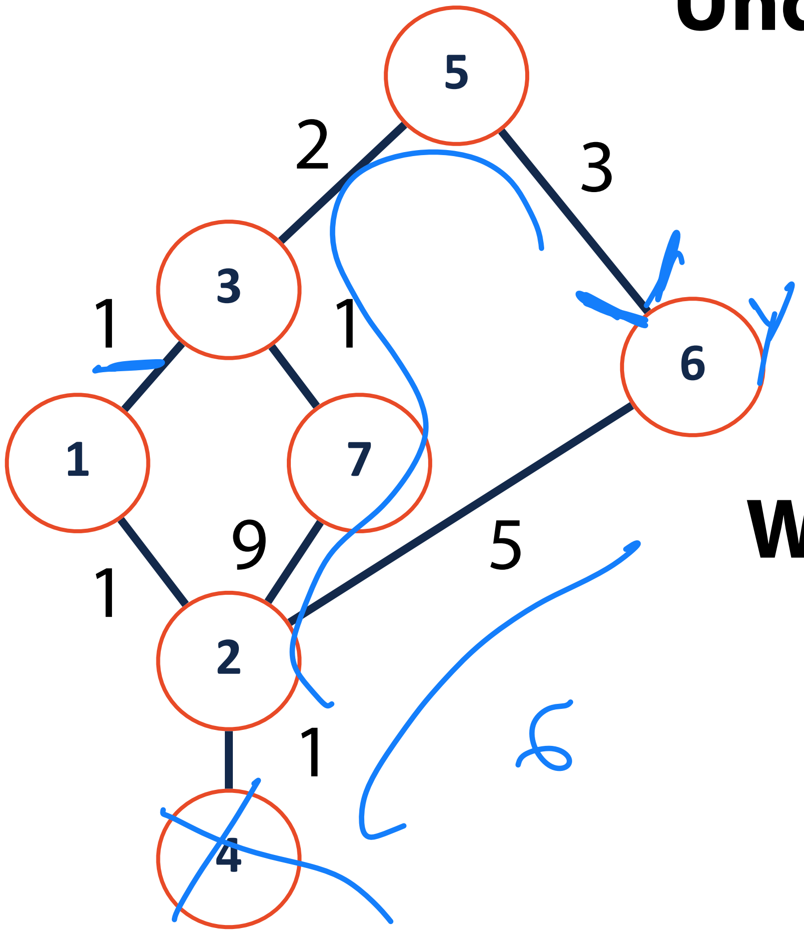
Terminology Trivia: Every tree is a graph but not every graph is a tree

Graph Vocabulary

5 - 6 Not 6 - 5

Directed: Edges are one way connections

Undirected: Traversable in either direction



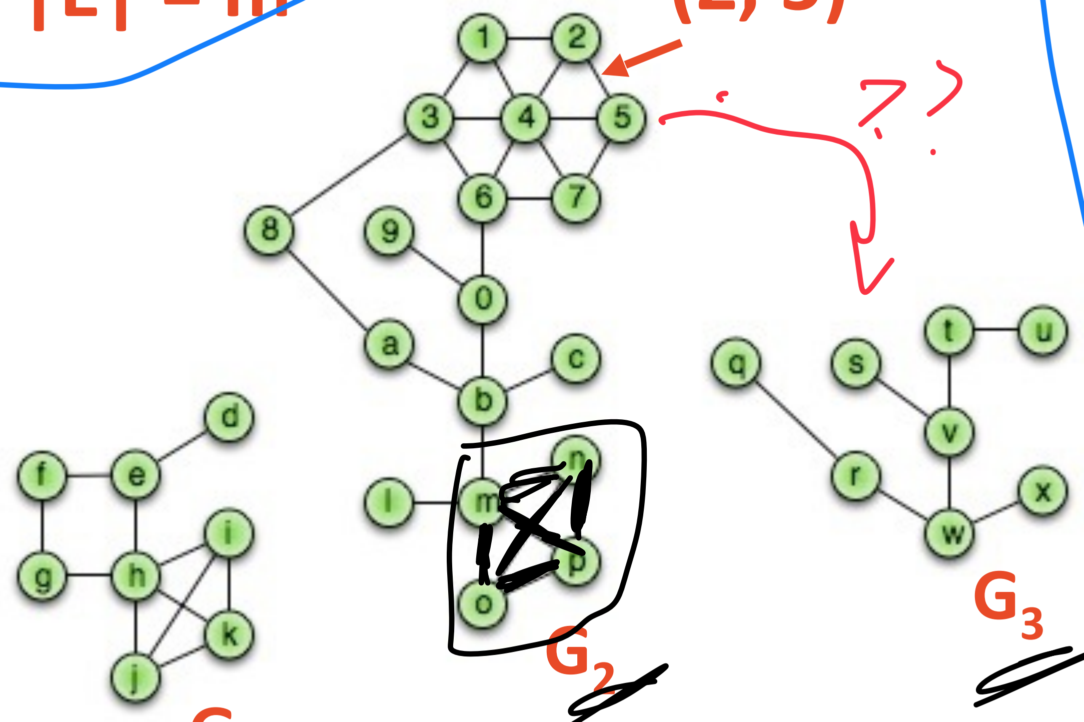
Weighted: A value associated with an edge

Graph Vocabulary

G = (V, E)

|V| = n

|E| = m



Subgraph(G):

G' = (V', E'):

V' ∈ V, E' ∈ E, and

(u, v) ∈ E' → u ∈ V', v ∈ V'

↳ Edges only between vertices in subgraph

Complete subgraph(G)

Connected subgraph(G)

Connected component(G)

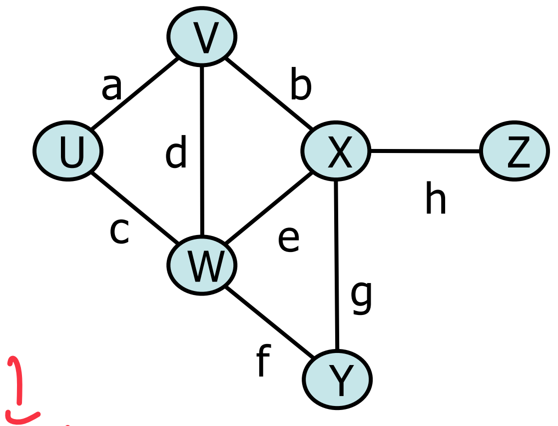
Acyclic subgraph(G)

Spanning tree(G)

↳ Every vertex in graph is connected but no cycles

Running times are often reported by **n**, the number of vertices, but often depend on **m**, the number of edges.

How many edges?



Minimum edges:

Not Connected:

Connected*:

Maximum edges:

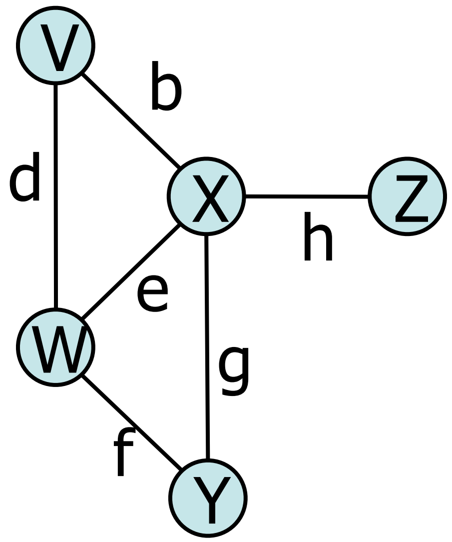
Simple:

$$\sum_{v \in V} deg(v) =$$

Graph ADT

Data:

- Vertices
- Edges
- Some data structure maintaining the structure between vertices and edges.



Functions:

- insertVertex(K key);
- insertEdge(Vertex v1, Vertex v2, K key);

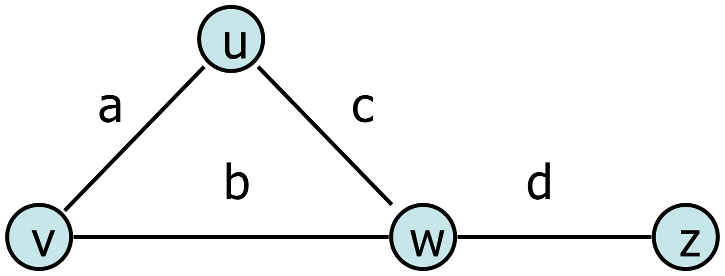
- removeVertex(Vertex v);
- removeEdge(Vertex v1, Vertex v2);

- incidentEdges(Vertex v);
- areAdjacent(Vertex v1, Vertex v2);

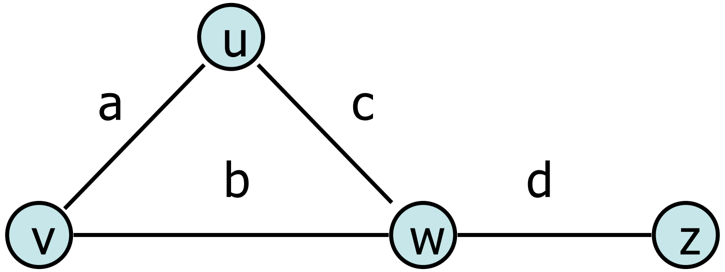
- origin(Edge e);
- destination(Edge e);

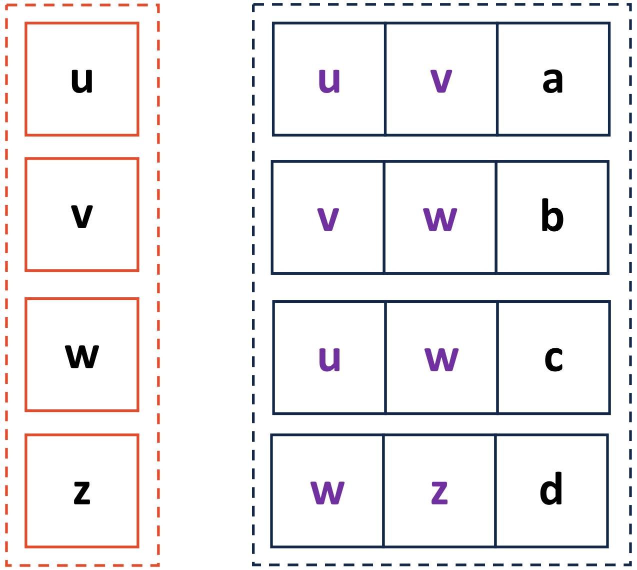
Find()???

Graph Implementation Idea



Graph Implementation: Edge List

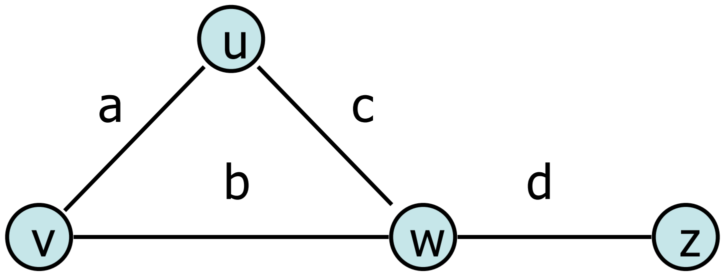




Vertex Collection:

Edge Collection:

Graph Implementation: Edge List



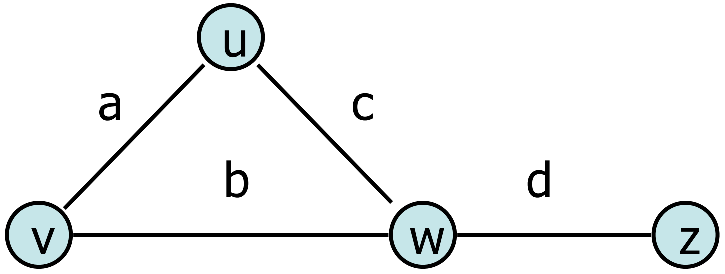
u
v
w
z

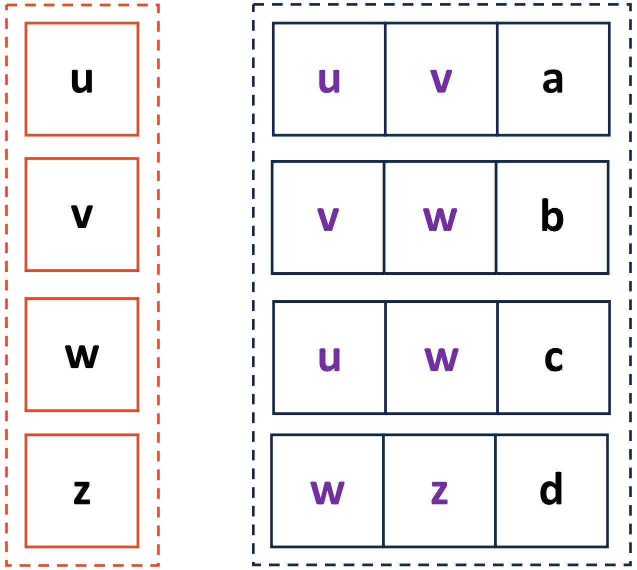
u	v	a
v	w	b
u	w	c
w	z	d

insertVertex(K key):

removeVertex(Vertex v):

Graph Implementation: Edge List



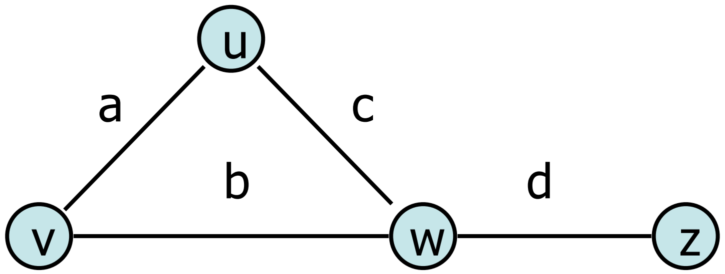


incidentEdges(Vertex v):

areAdjacent(Vertex v1, Vertex v2):

```
G.incidentEdges(v1).contains(v2)
```

Graph Implementation: Edge List



u
v
w
z

u	v	a
v	w	b
u	w	c
w	z	d

insertEdge(Vertex v1, Vertex v2, K key):

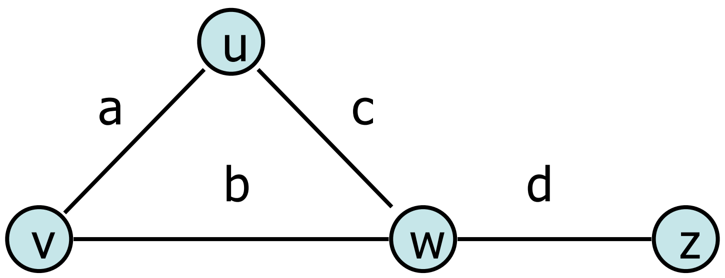
Graph Implementation: Edge List

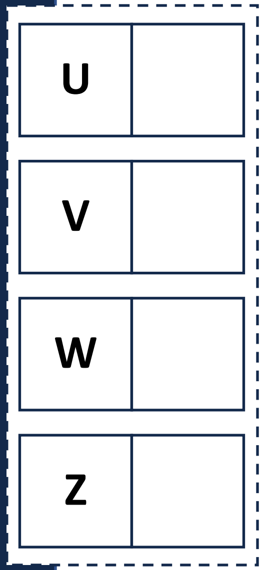
Pros:

Cons:

Graph Implementation: Adjacency Matrix

insertVertex(K key);
removeVertex(Vertex v);
areAdjacent(Vertex v1, Vertex v2);
incidentEdges(Vertex v);





	u	v	w	z
u				
v				
w				
z				