SemEval-2013 Task 13: **Word Sense Induction for Graded and Non-Graded Senses**

David Jurgens
Dipartimento di Informatica
Sapienza Universita di Roma
jurgens@di.uniroma1.it

Ioannis Klapaftis
Search Technology Center Europe
Microsoft
ioannisk@microsoft.com

- Introduction
- Task Overview
- Data
- Evaluation
- Results

John sat on the **chair**.

1. a seat for one person, with a support for the back
2. the position of professor
3. the officer who presides at the meetings of an organization

Which meaning of the word is being used?

John sat on the **chair**.

1. a seat for one person, with a support for the back
2. the position of professor
3. the officer who presides at the meetings of an organization

Which meaning of the word is being used?

This is the problem of
Word Sense Disambiguation (WSD)

What are the meanings of a word?

It was too **dark** to see

I light candles when it gets **dark**

It was **dark** outside

These are some **dark** glasses

The **dark** blue clashed with the yellow

Her dress was a **dark** green

The project was made with **dark** designs

We didn’t ask what **dark** purpose the knife was for

What are the meanings of a word?

It was too **dark** to see

I light candles when it gets **dark**

It was **dark** outside

These are some **dark** glasses

The **dark** blue clashed with the yellow

Her dress was a **dark** green

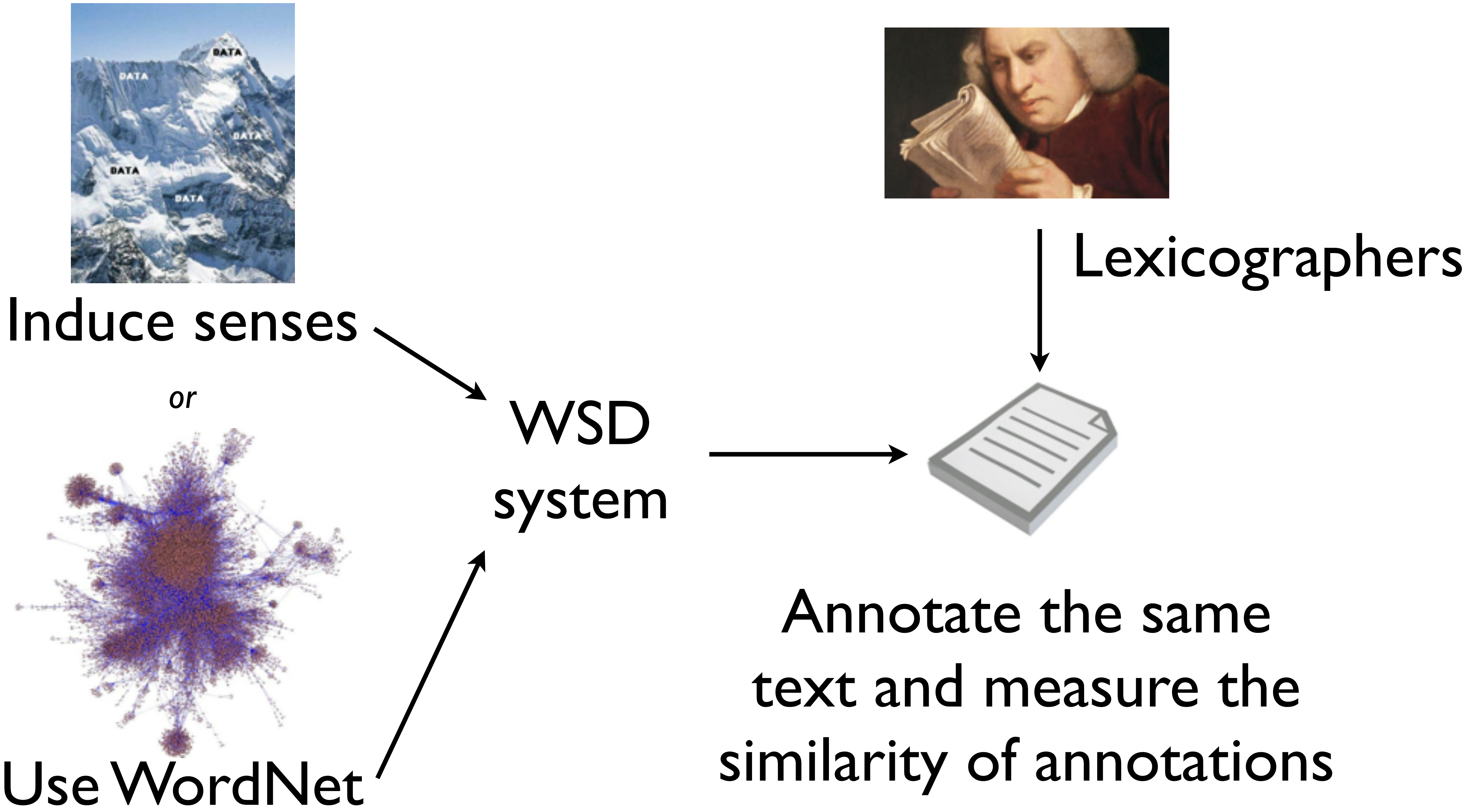
The project was made with **dark** designs

We didn’t ask what **dark** purpose the knife was for

This is the problem of
Word Sense Induction (WSI)

- Introduction
- Task Overview
- Data
- Evaluation
- Results

Task 13 Overview



Induce senses

or

Use WordNet

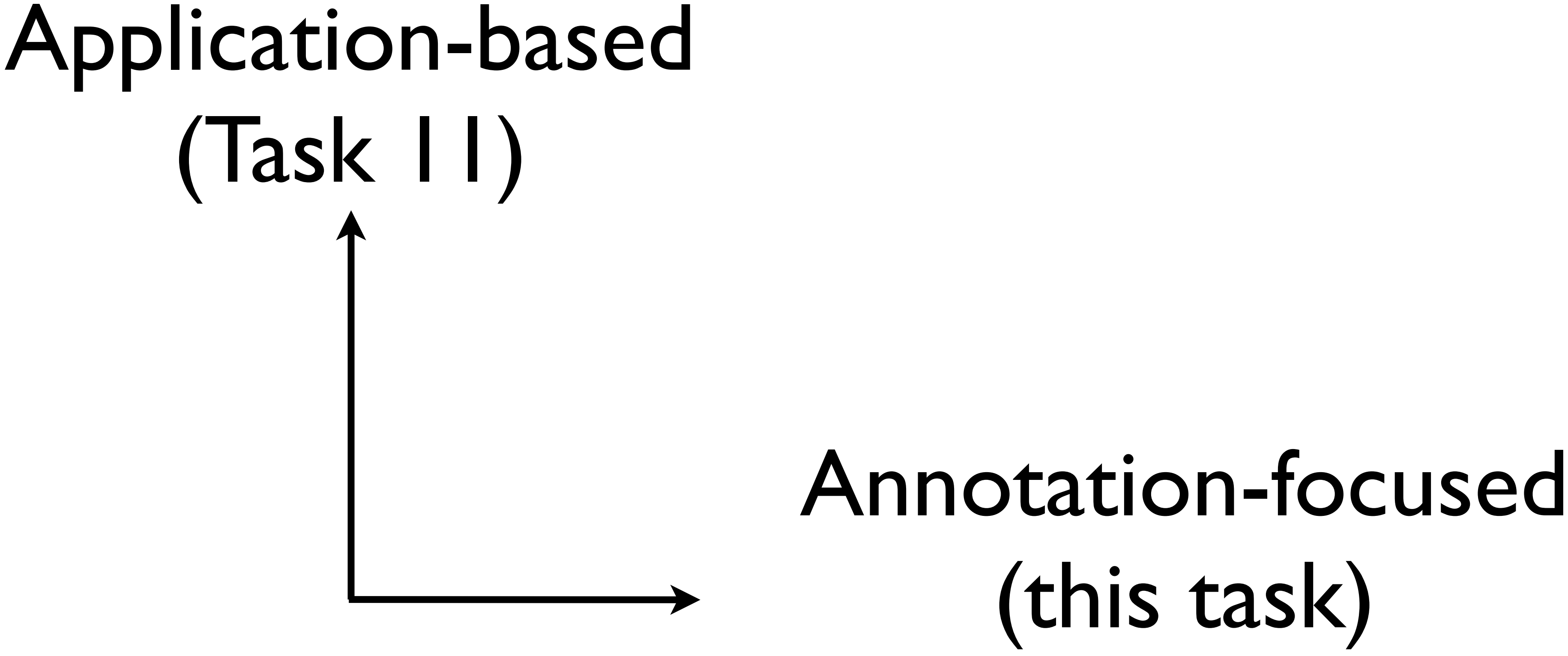
WSD system

Lexicographers

Annotate the same text and measure the similarity of annotations

Why another WSD/WSI task?

Why another WSD/WSI task?

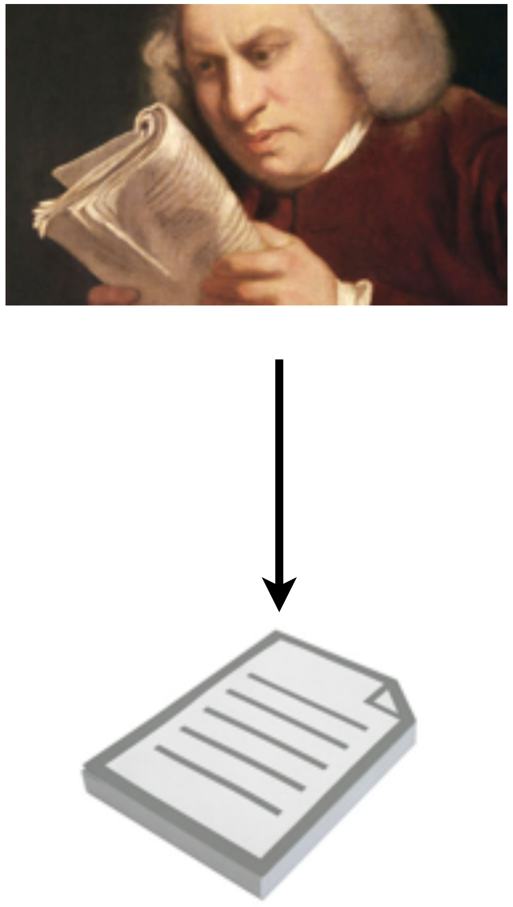


Application-based
(Task 11)

Annotation-focused
(this task)

WSD Evaluation is tied to Inter-Annotator Agreement (IAA)

Lexicographers



If lexicographers can’t agree on which meaning is present, WSD systems will do no better.

Why might humans not agree?

He **struck** them with full force.

He **struck** them with full force.

He’s probably fighting so

strike#v#1“deliver a sharp blow”

He **struck** them with full force.

He’s clearly playing a piano!

strike#v#10 “produce by manipulating keys”

He **struck** them with full force.

I thought he was minting coins the old fashioned way

strike#v#19 “form by stamping”

He **struck** them with full force.

- strike#v#1 “deliver a sharp blow”
- strike#v#10 “produce by manipulating keys”
- strike#v#19 “form by stamping”

Only one sense is correct, but contextual ambiguity makes it impossible to determine which one.

She handed the **paper** to her professor

Multiple, mutually-compatible meanings

She handed the **paper** to her professor

- paper#n#1 - a material made of cellulose
- paper#n#2 - an essay or assignment

Multiple, mutually-compatible meanings

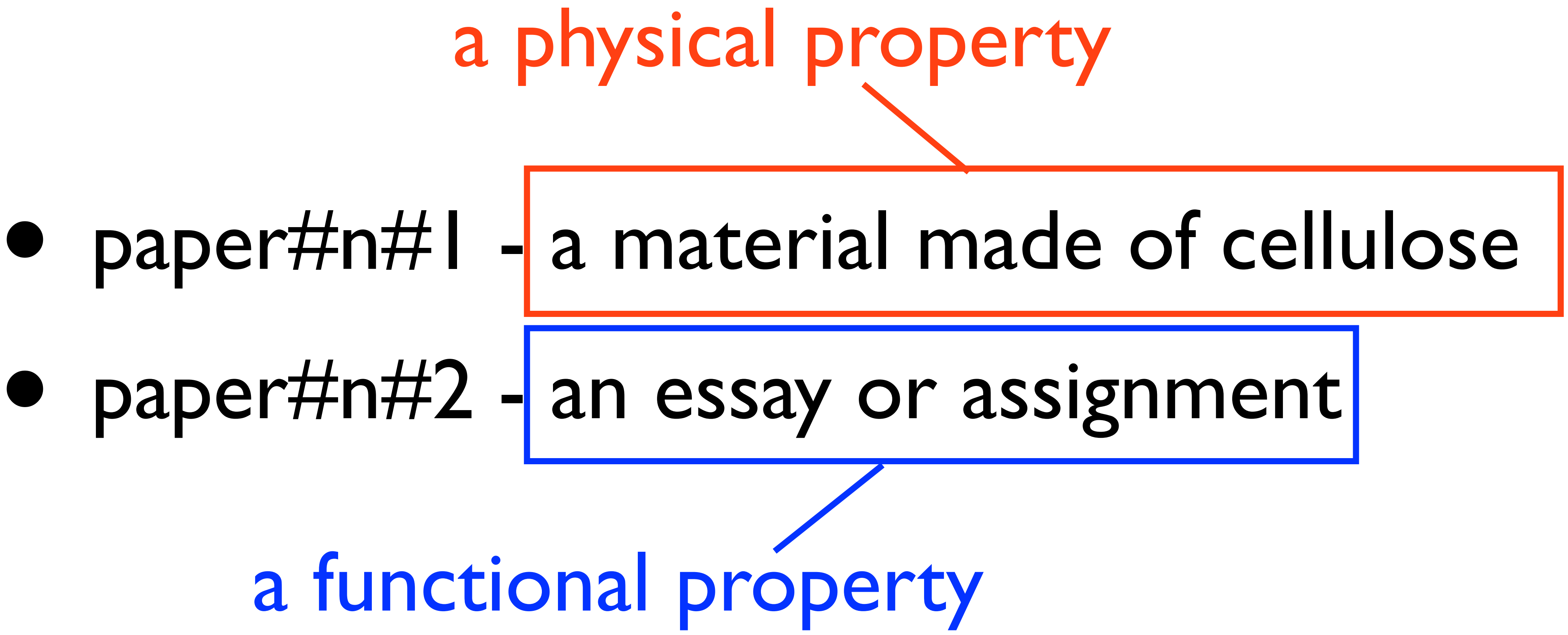
She handed the **paper** to her professor

a physical property

- paper#n#1 - a material made of cellulose
- paper#n#2 - an essay or assignment

Multiple, mutually-compatible meanings

She handed the **paper** to her professor



a physical property

- paper#n#1 - a material made of cellulose
- paper#n#2 - an essay or assignment

a functional property

Parallel literal and metaphoric interpretations

We commemorate our births from out of the **dark** centers of women

- dark#a#1 – devoid of or deficient in light or brightness; shadowed or black
- dark#a#5 – secret

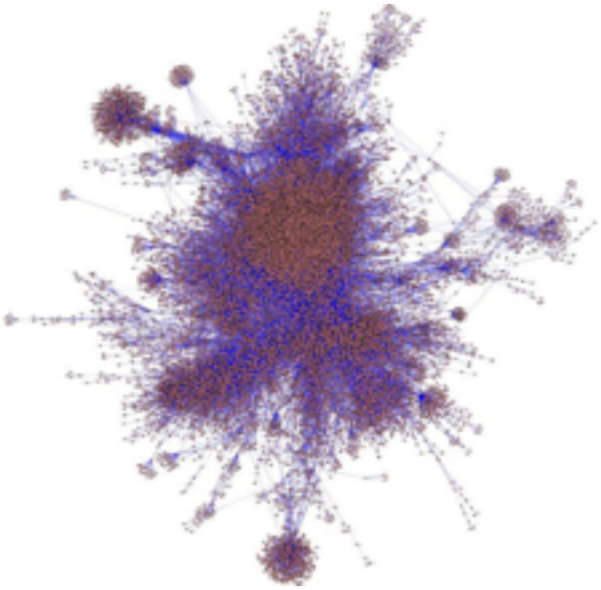
Annotators will use multiple senses if you let them

- Véronis (1998)
- Murray and Green (2004)
- Erk et al. (2009, 2012)
- Jurgens (2012)
- Passoneau et al. (2012)
- Navigli et al. (2013) - **Task 12**
- Korkontzelos et al. (2013) - **Task 5**

New in Task 13: **More Ambiguity**!

Induce senses

or



Use WordNet

WSD system

Lexicographers

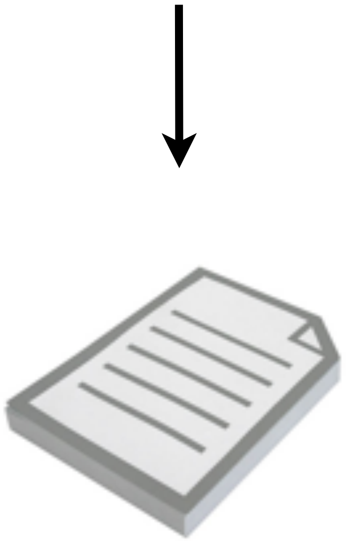
Annotate the same text and measure the similarity of annotations

Task 13 models explicitly annotating instances with...

- Ambiguity
- Non-exclusive property-based senses in the sense inventory
- Concurrent literal and metaphoric interpretations

Task 13 annotation has lexicographers and WSD systems use **multiple** senses with **weights**

The student handed her **paper** to the professor



Task 13 annotation has lexicographers and WSD systems use **multiple** senses with **weights**

The student handed her **paper** to the professor

- paper%1:10:01:: – an essay

Definitely! 100%

Task 13 annotation has lexicographers and WSD systems use **multiple** senses with **weights**

The student handed her **paper** to the professor

- paper%1:10:01:: – an essay

  Definitely! 100%

- paper%1:27:00:: – a material made of cellulose pulp

  Sort of? 30%

Potential Applications

- Identifying "less bad" translations in ambiguous contexts
- Potentially preserve ambiguity across translations
- Detecting poetic or figurative usages
- Provide more accurate evaluations when WSD systems detect multiple senses

- Introduction
- Task Overview
- Data
- Evaluation
- Results

Task 13 Data

- Drawn from the Open ANC
  - Both written and spoken
- 50 target lemmas
  - 20 noun, 20 verb, 10 adjective
- 4,664 Instances total

Annotation Process

amazon mechanicalturk™
Artificial Artificial Intelligence

1 Use methods from Jurgens (2013) to get MTurk annotations

Annotation Process

amazon mechanical turk™
Artificial Artificial Intelligence

1 Use methods from Jurgens (2013) to get MTurk annotations

2 Achieve high (> 0. 8) agreement

Annotation Process

amazon mechanical turk™
Artificial Artificial Intelligence

1 Use methods from Jurgens (2013) to get MTurk annotations

2 Achieve high (> 0. 8) agreement

3 Analyze annotations and discover Turkers are agreeing but are also wrong

Annotation Process

amazon mechanical turk™
Artificial Artificial Intelligence

1 Use methods from Jurgens (2013) to get MTurk annotations

2 Achieve high (> 0. 8) agreement

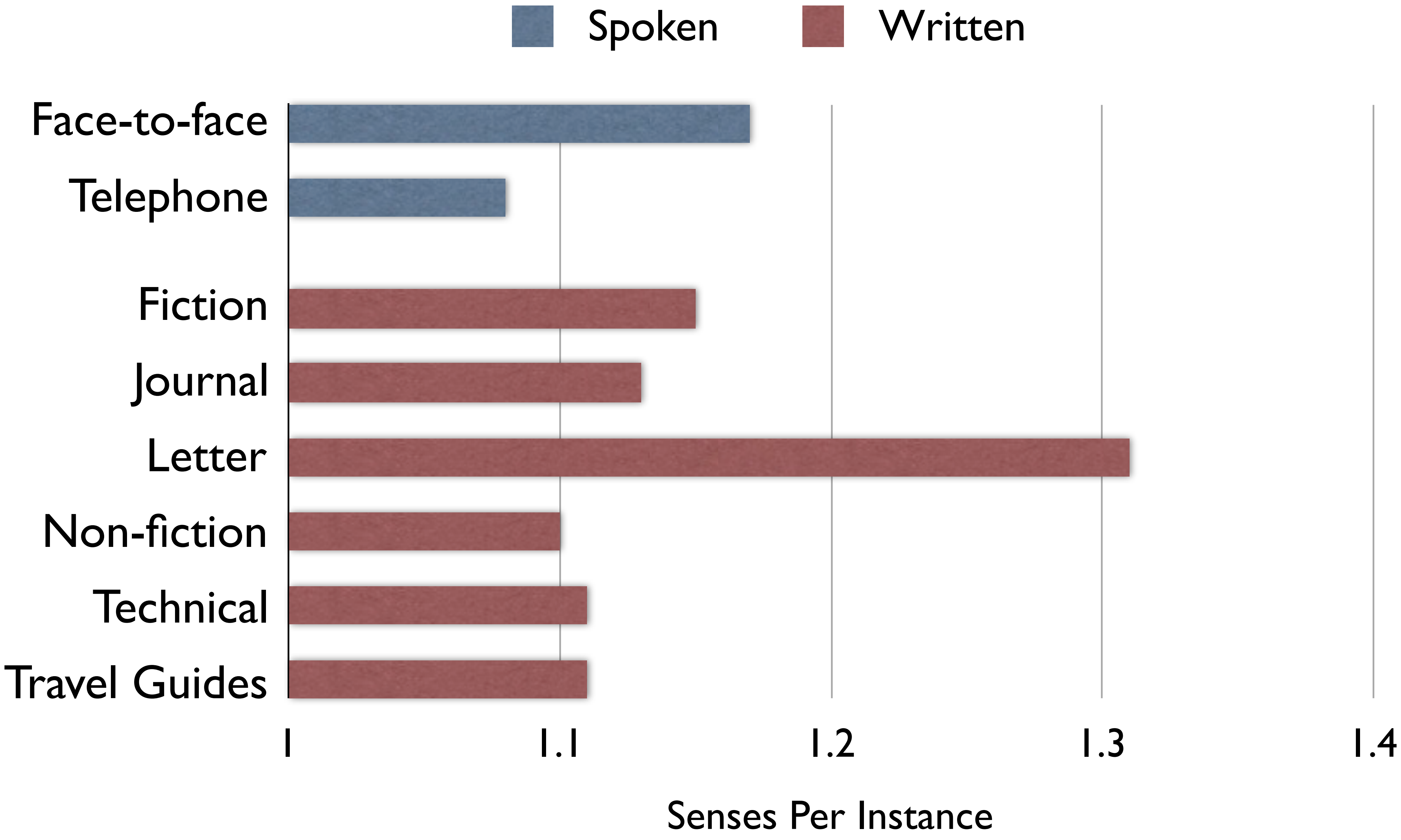
3 Analyze annotations and discover Turkers are agreeing but are also wrong

4 Annotate the data ourselves

Annotation Setup

- Rate the applicability of each sense on a scale from one to five
  - One indicates doesn’t apply
  - Five is exactly applies

Multiple sense annotation rates



- Introduction
- Task Overview
- Data
- Evaluation
- Results

Evaluating WSI and WSD Systems

Lexicographer Evaluation

WSD Evaluation

WSI Evaluations

It was **dark** outside

Her dress was a **dark** green

We didn’t ask what **dark** purpose the knife was for

WSI Evaluations

It was too **dark** to see

I light candles when it gets **dark**

It was **dark** outside

Dark nights and short days

These are some **dark** glasses

The **dark** blue clashed with the yellow

Her dress was a **dark** green

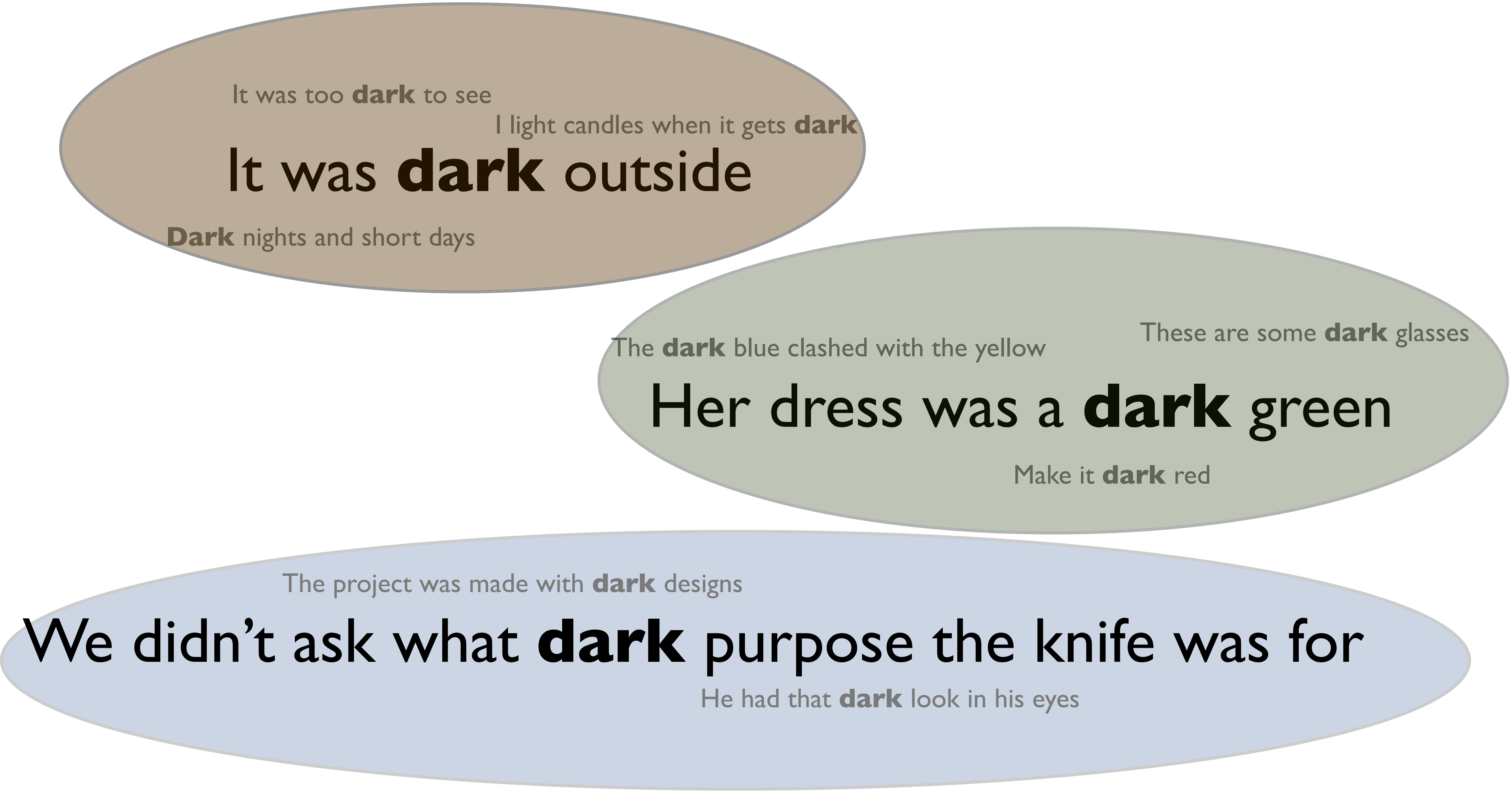
Make it **dark** red

The project was made with **dark** designs

We didn’t ask what **dark** purpose the knife was for

He had that **dark** look in his eyes

WSI Evaluations



It was too **dark** to see

I light candles when it gets **dark**

It was **dark** outside

Dark nights and short days

The **dark** blue clashed with the yellow

These are some **dark** glasses

Her dress was a **dark** green

Make it **dark** red

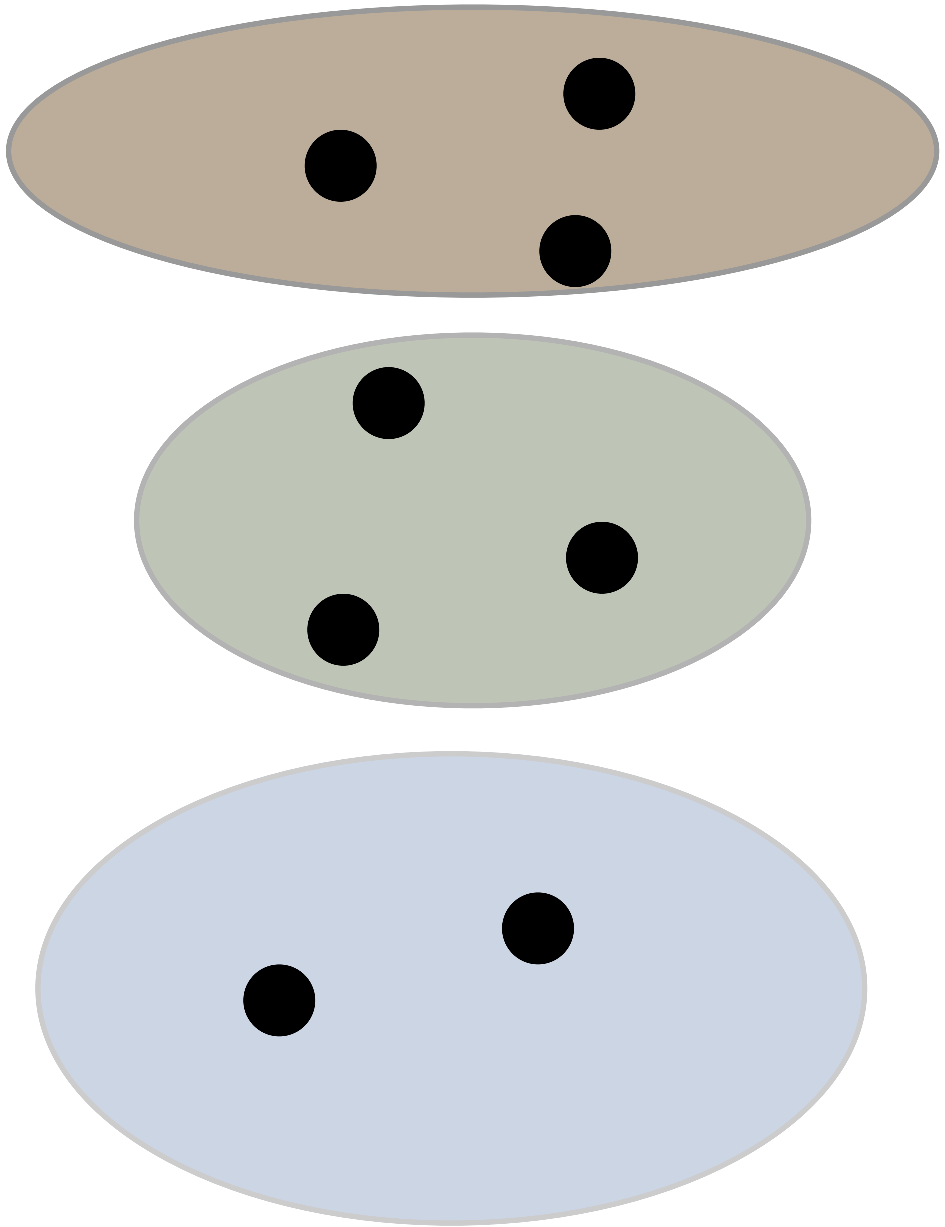
The project was made with **dark** designs

We didn’t ask what **dark** purpose the knife was for

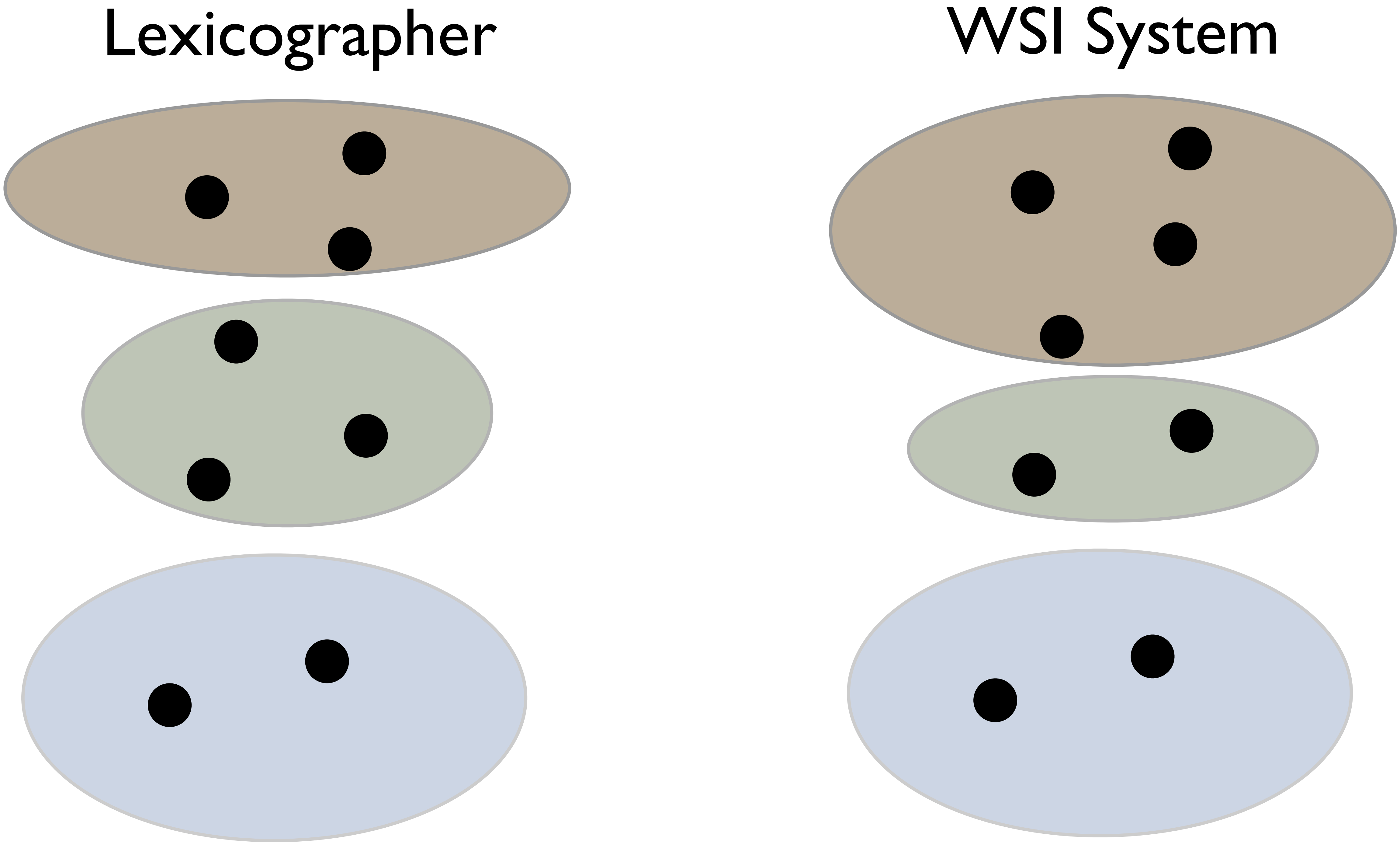
He had that **dark** look in his eyes

WSI Evaluations

Lexicographer



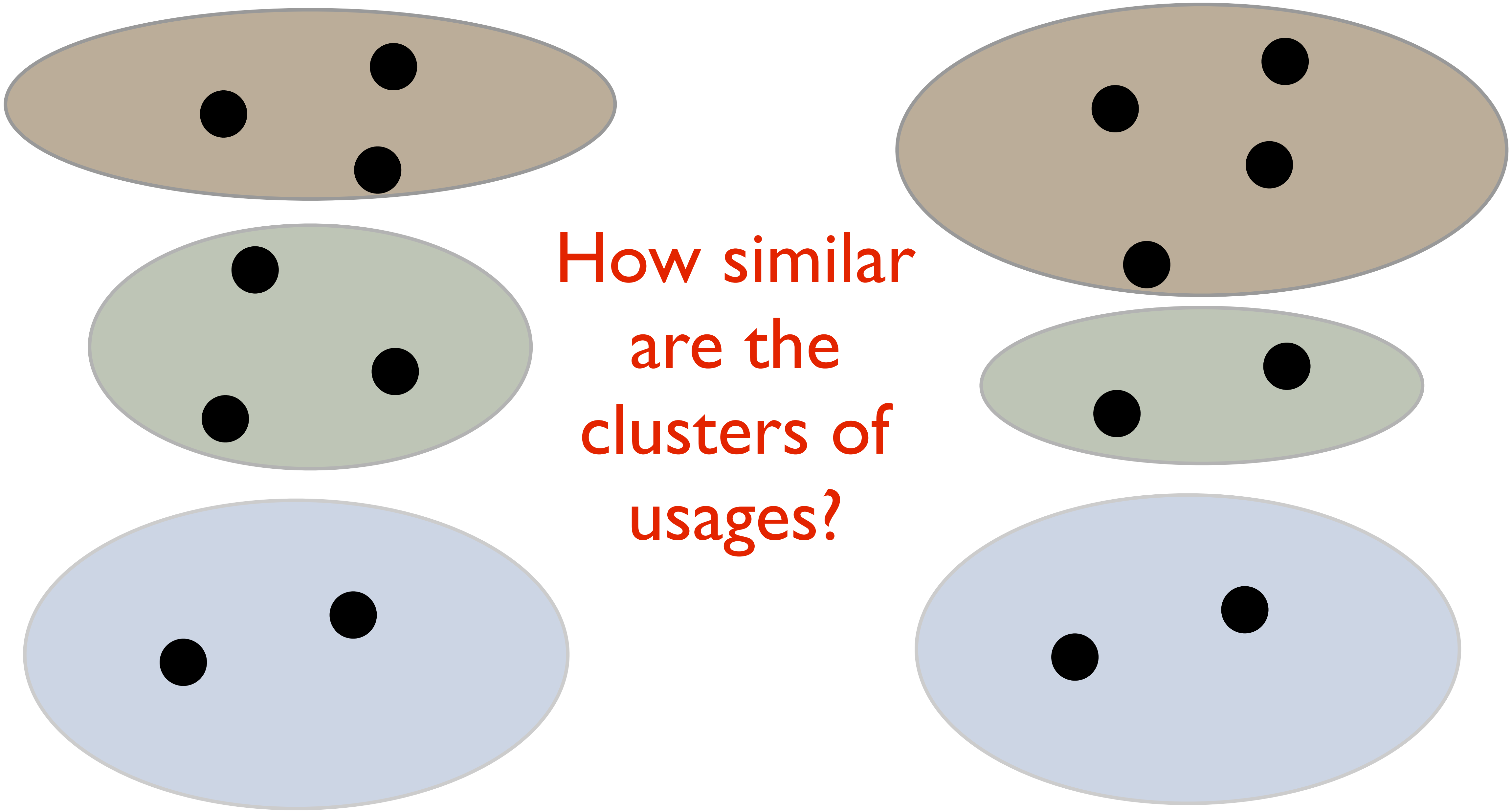
WSI Evaluations



Lexicographer

WSI System

WSI Evaluations



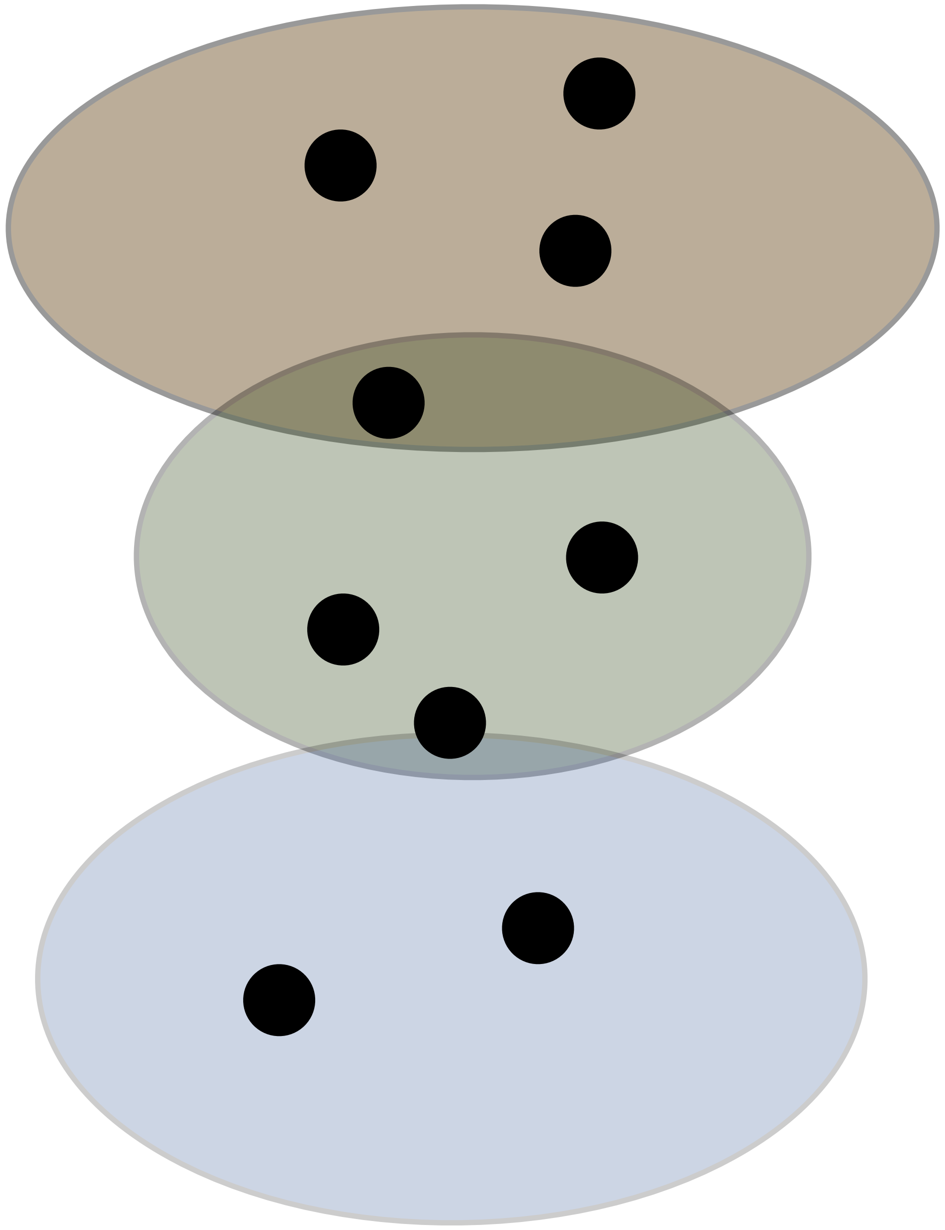
Lexicographer

WSI System

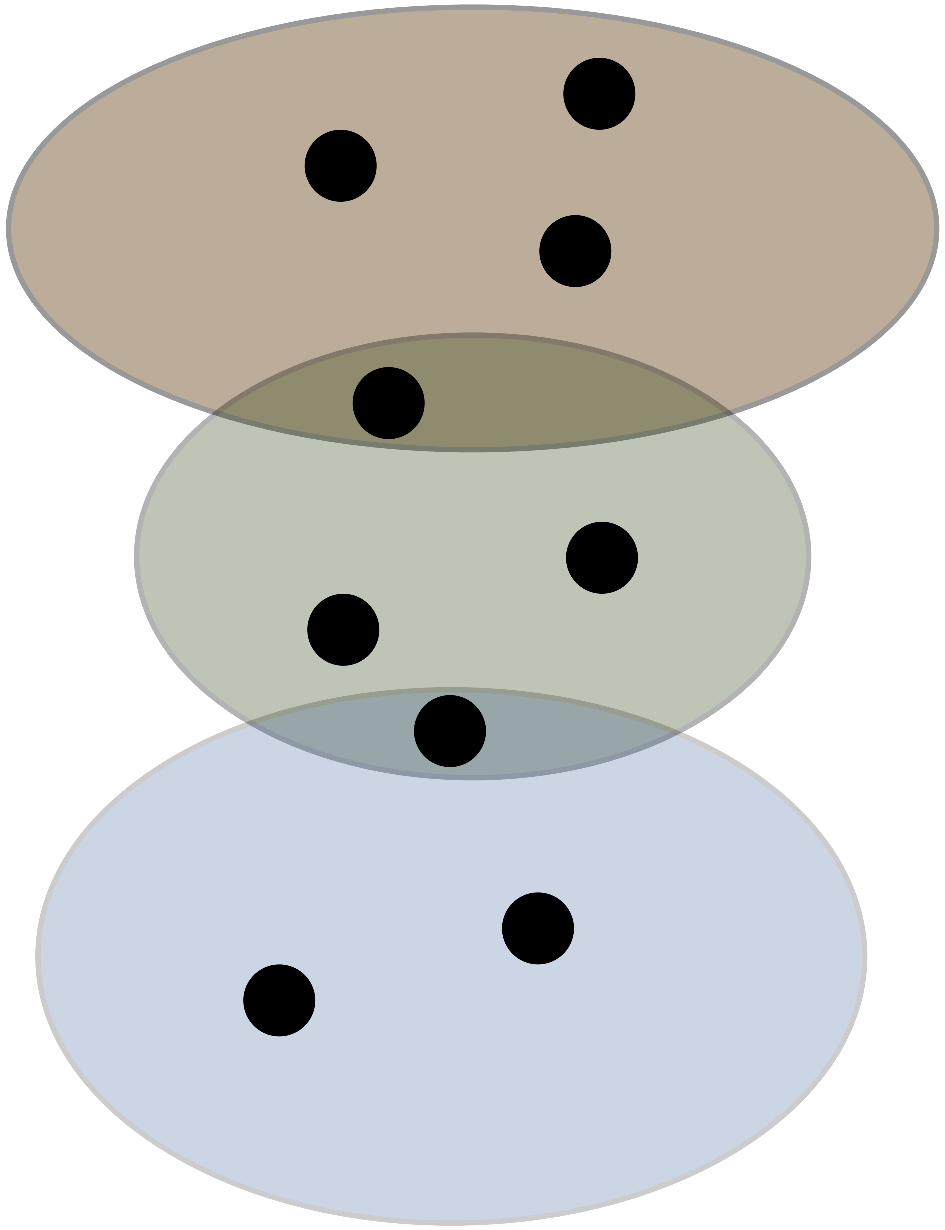
How similar are the clusters of usages?

The complication of fuzzy clusters

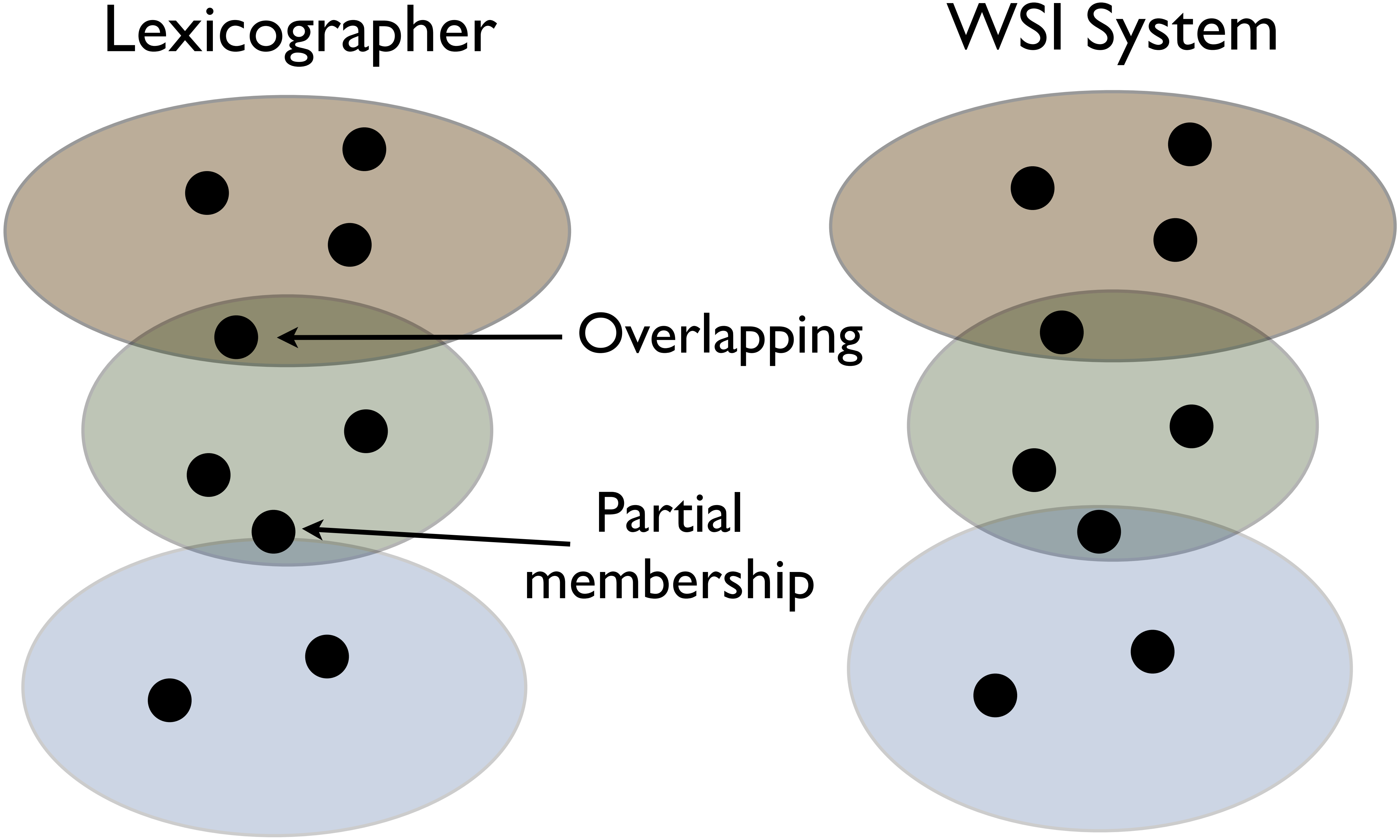
Lexicographer



WSI System



The complication of fuzzy clusters



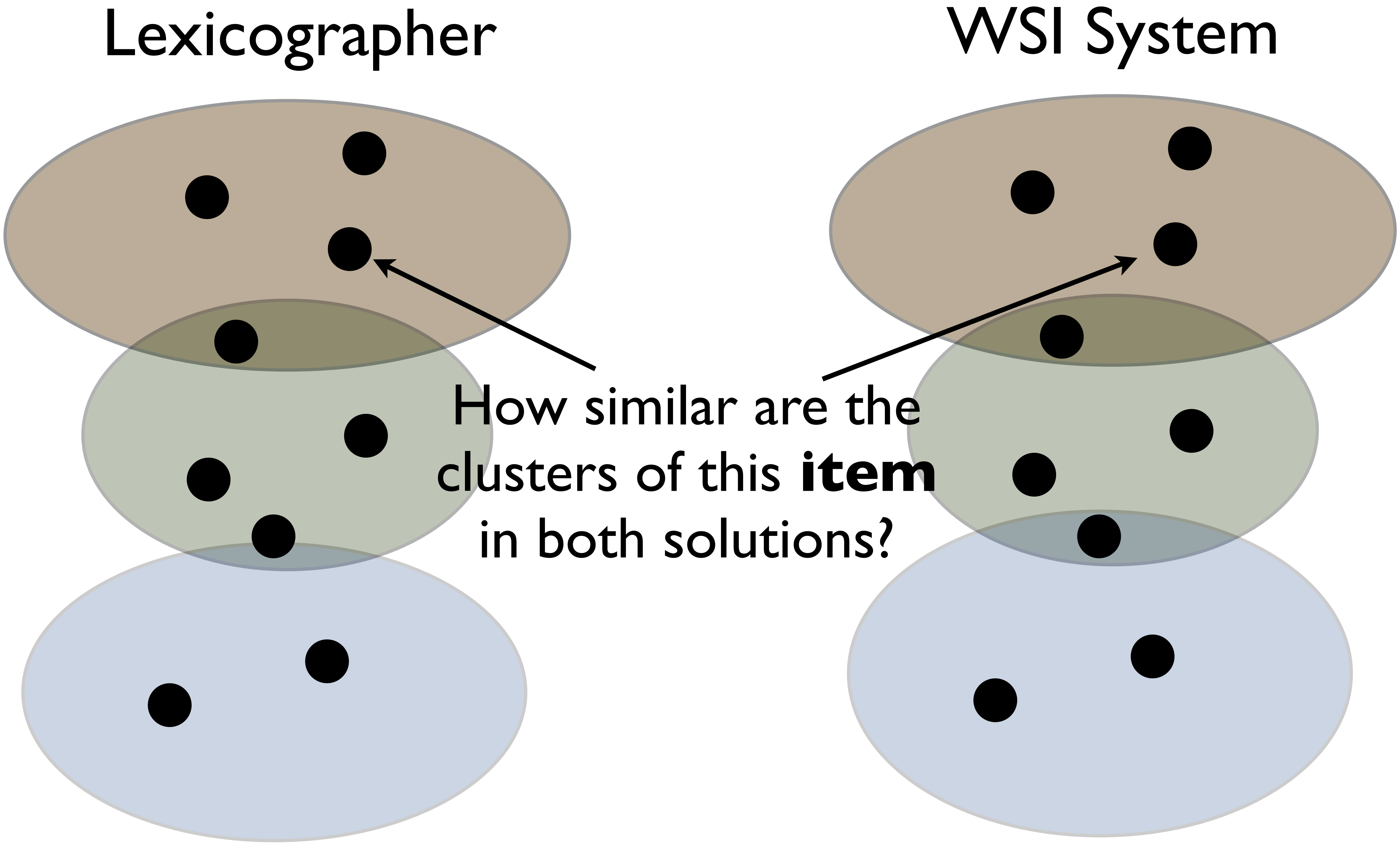
Lexicographer

WSI System

Overlapping

Partial membership

Evaluation 1: Fuzzy B-Cubed

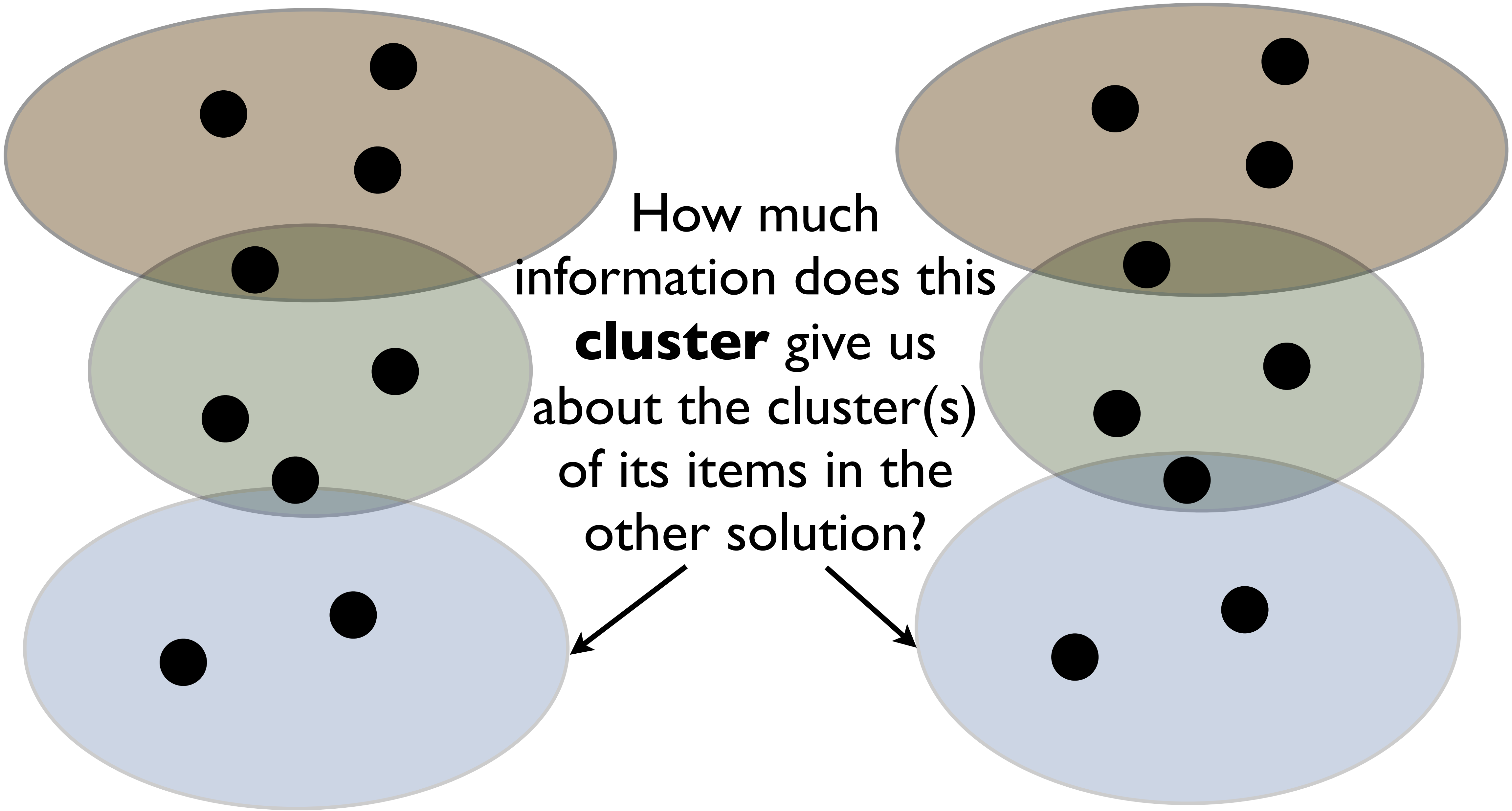


How similar are the clusters of this **item** in both solutions?

Evaluation 1: Fuzzy Normalized Mutual Information

Lexicographer

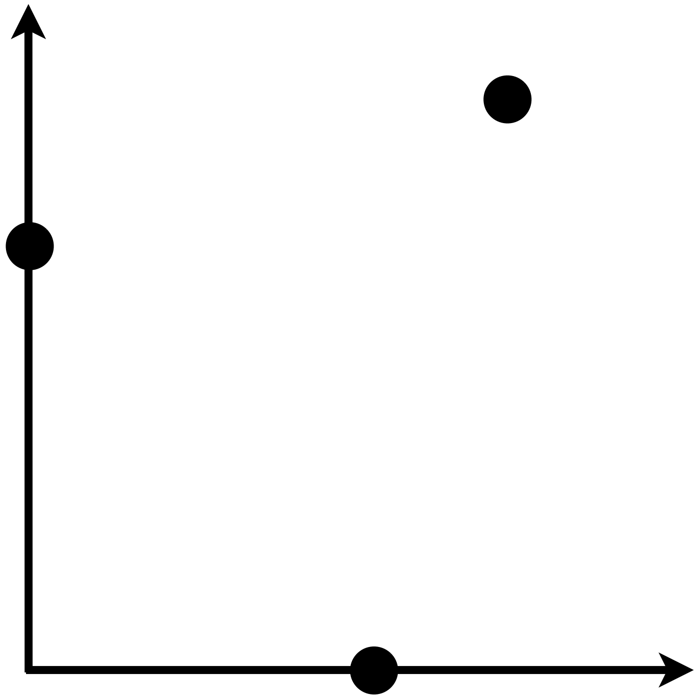
WSI System



How much information does this **cluster** give us about the cluster(s) of its items in the other solution?

Why two measures?

B-Cubed: performance with the same sense distribution



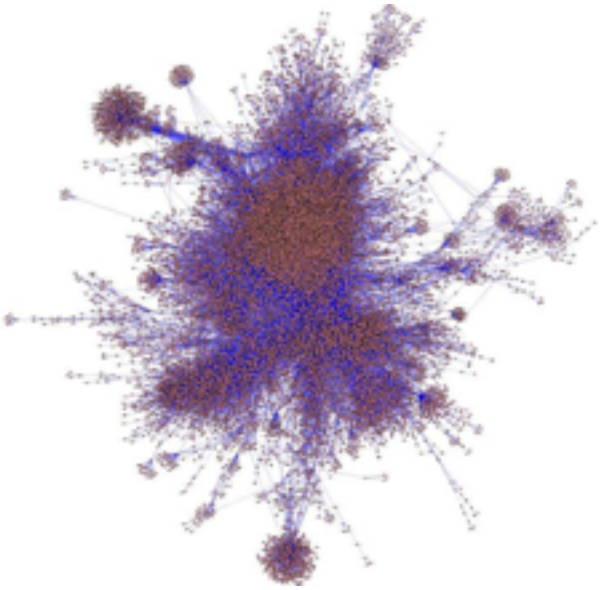
NMI: performance independent of sense distribution

WSD Evaluations

WSD Evaluations

Induce senses

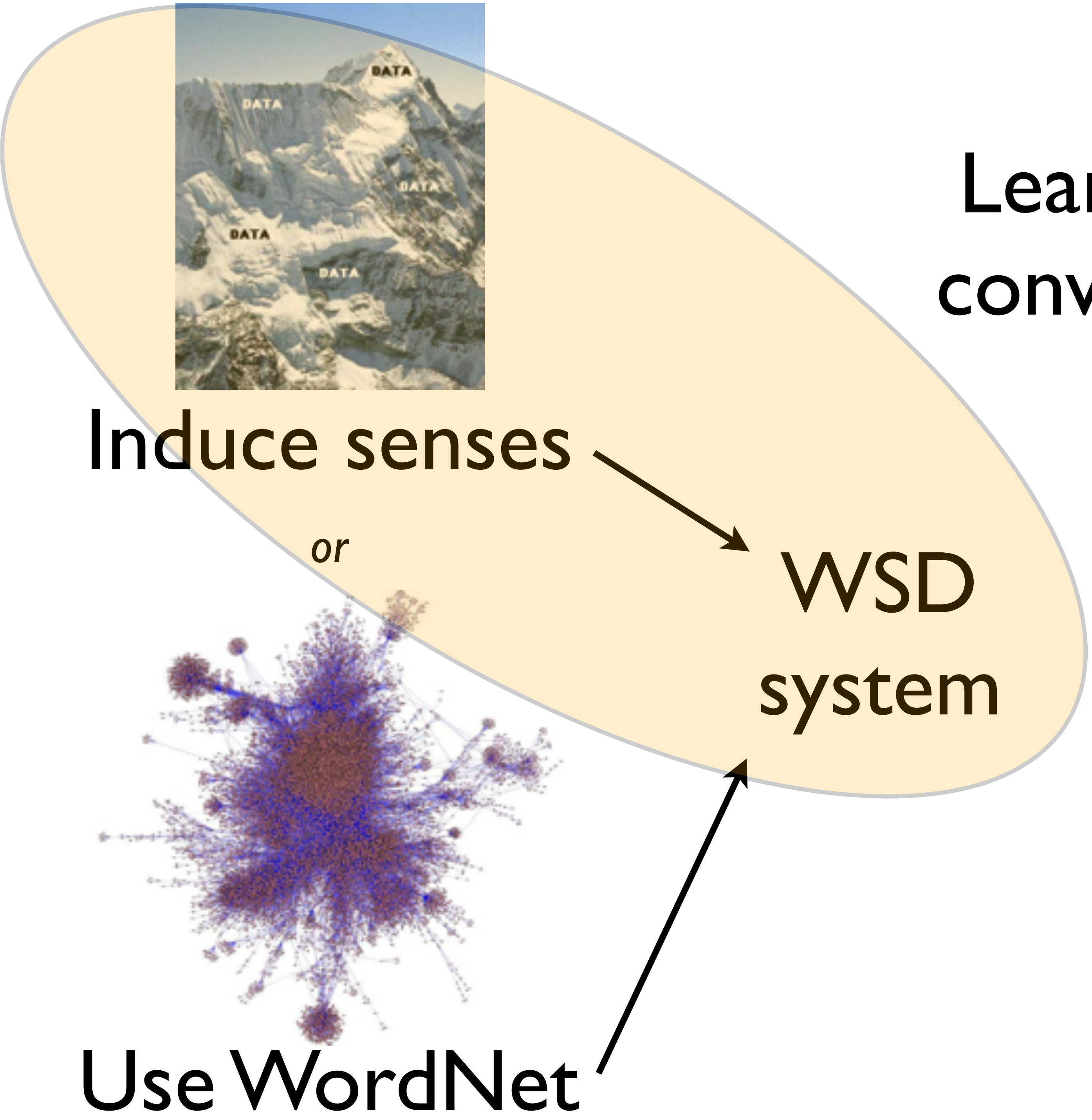
or



Use WordNet

WSD system

WSD Evaluations



Induce senses

or

WSD system

Use WordNet

Learn a mapping function that converts an induced labeling to a WordNet labeling

- 80% use to learn mapping
- 20% used for testing
- Used Jurgens (2012) method for mapping

WSD Evaluations

1 Which senses apply?

2 Which senses apply more?

3 How much does each sense apply?

WSD Evaluations

1 Which senses apply?

Gold = {$\text{wn}_{1},\ \text{wn}_{2}$}

Test = {wn_{1}}

Jaccard Index

$$\frac{|\text{Gold} \cap \text{Test}|}{|\text{Gold} \cup \text{Test}|}$$

WSD Evaluations

2 Which senses apply more?

Gold = {wn_1:0.5, wn_2:1.0, wn_3:0.9} → wn_2 > wn_3: > wn_1

Test = {wn_1:0.6, wn_2:1.0,} → wn_2 > wn_1: > wn_3

Kendall’s Tau Similarity

with positional weighting

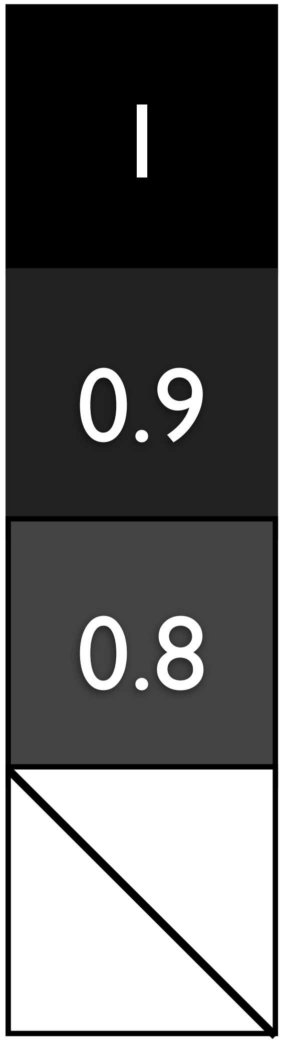
WSD Evaluations

3 How much does each sense apply?

Weighted Normalized Discounted Cumulative Gain

WSD Evaluations

- All measures are bounded in [0,1]

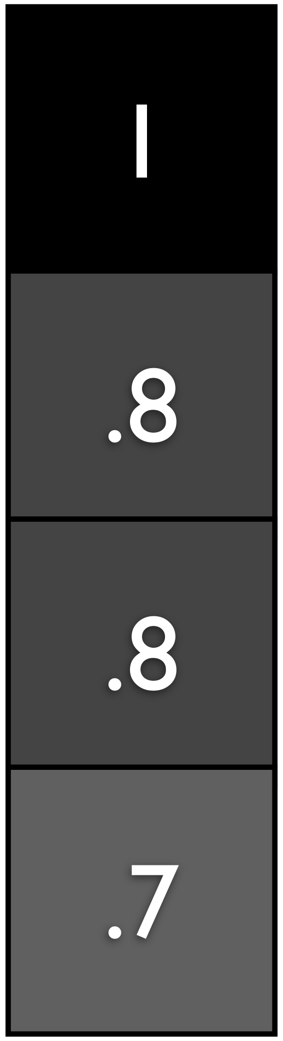


1

0.9

0.8

Avg: 0.9



1

.8

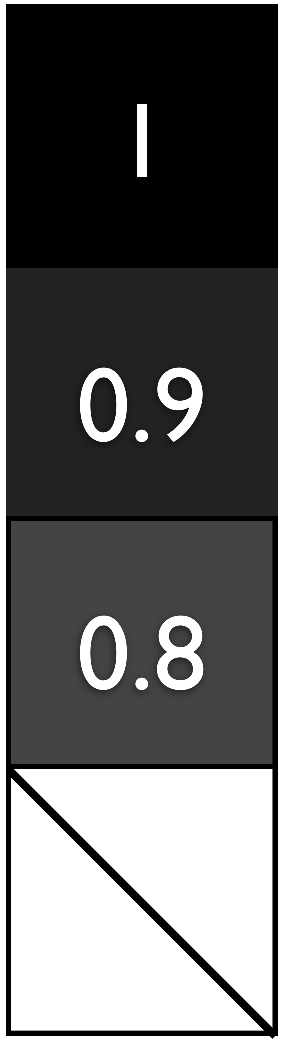
.8

.7

Avg: 0.825

WSD Evaluations

- All measures are bounded in [0,1]
- Extend Recall to be average across all answers



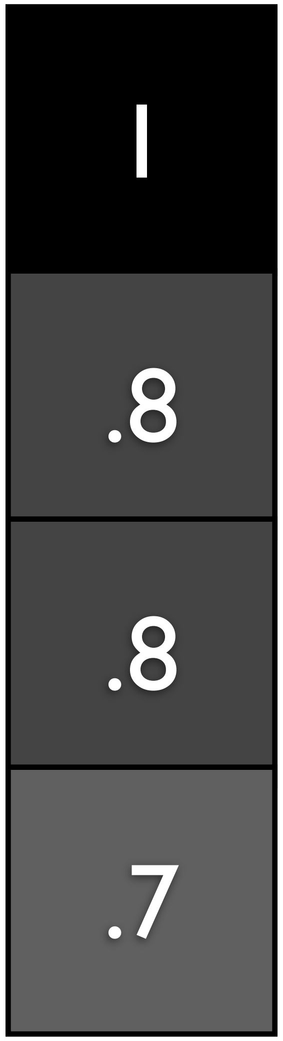
1

0.9

0.8

Avg: 0.9

Recall: 0.675



1

.8

.8

.7

Avg: 0.825

Recall: 0.825

Teams

AI-KU (WSI)

Lexical Substitution
+ Clustering

Teams

AI-KU (WSI)

Lexical Substitution
+ Clustering

Unimelb (WSI)

Topic Modeling

Teams

AI-KU (WSI)

Lexical Substitution
+ Clustering

Unimelb (WSI)

Topic Modeling

UoS (WSI)

Graph Clustering

Teams

AI-KU (WSI)

Lexical Substitution
+ Clustering

Unimelb (WSI)

Topic Modeling

UoS (WSI)

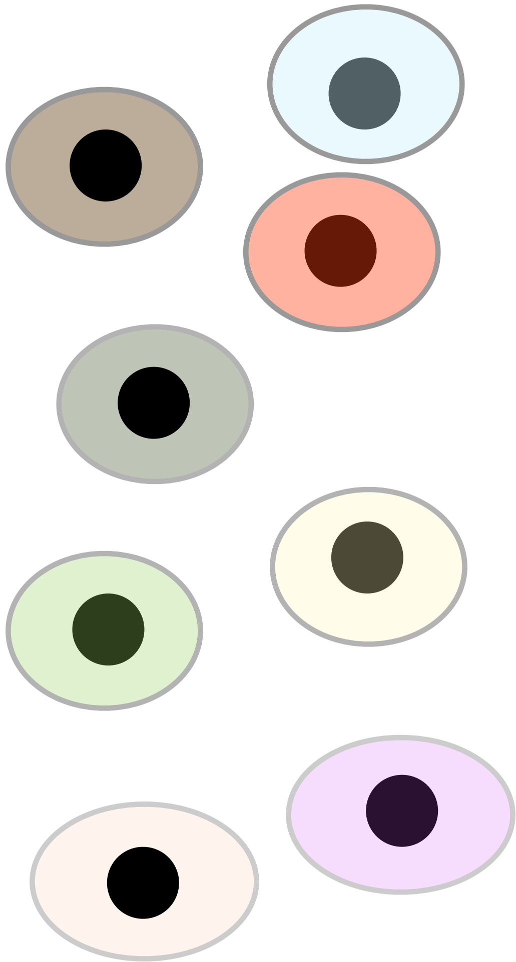
Graph Clustering

La Sapienza (WSD)

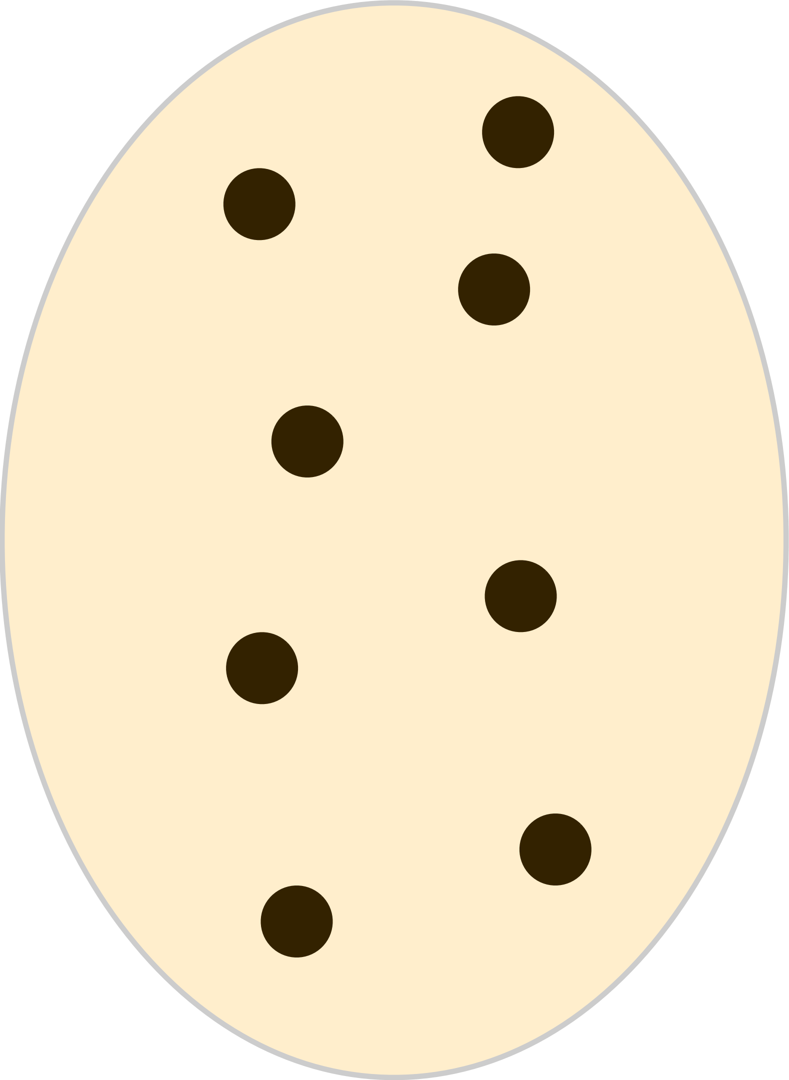
PageRank over
WordNet graph

WSI Baselines

One cluster per instance
(1c1inst)



One cluster

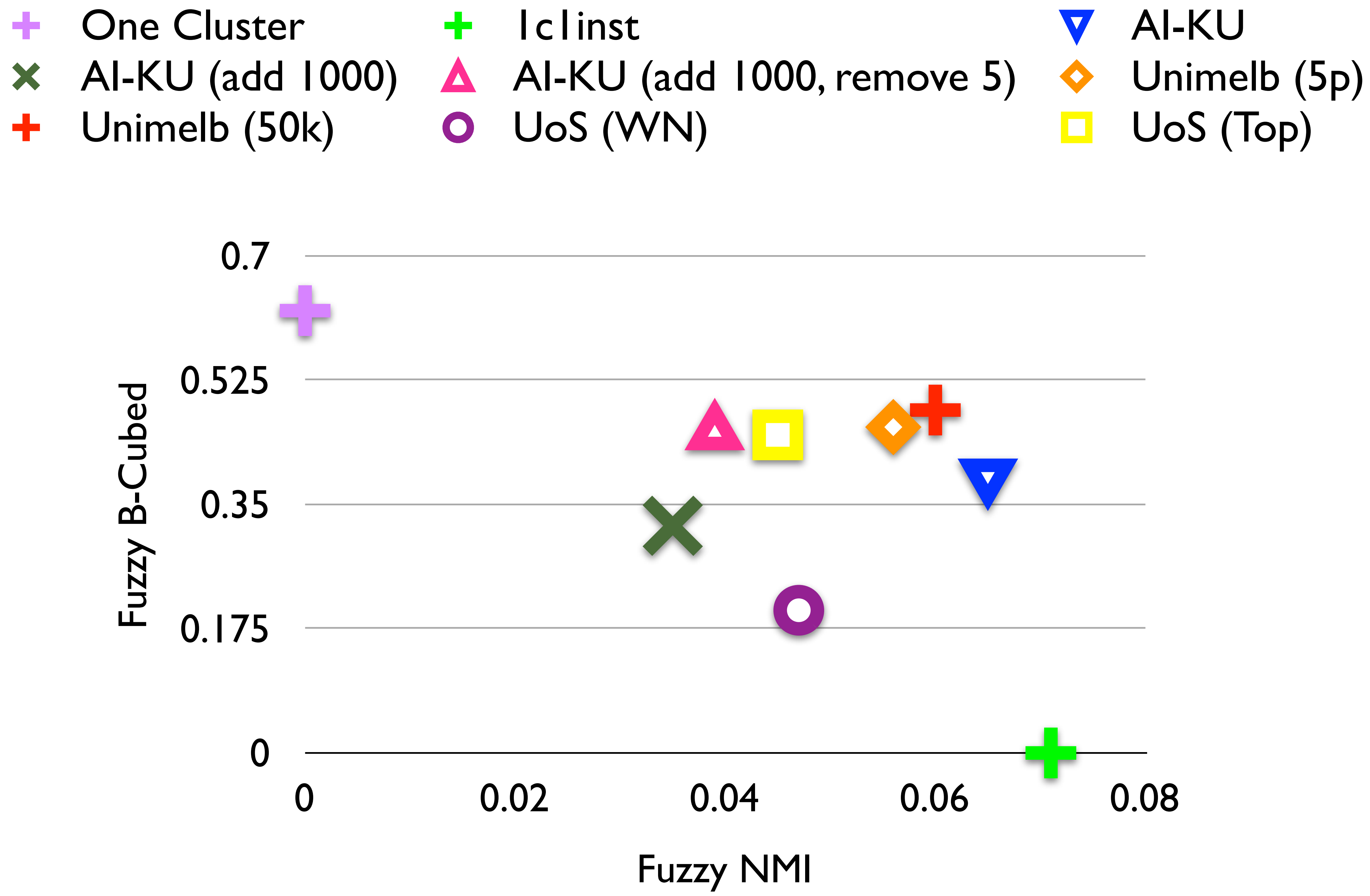


WSD Baselines

- **MFS** - All instances labeled with MFS from SemCor
- **Ranked Senses** - All instances labeled with *all senses*, proportionally weighted by their frequency in SemCor

- Introduction
- Task Overview
- Data
- Evaluation
- Results

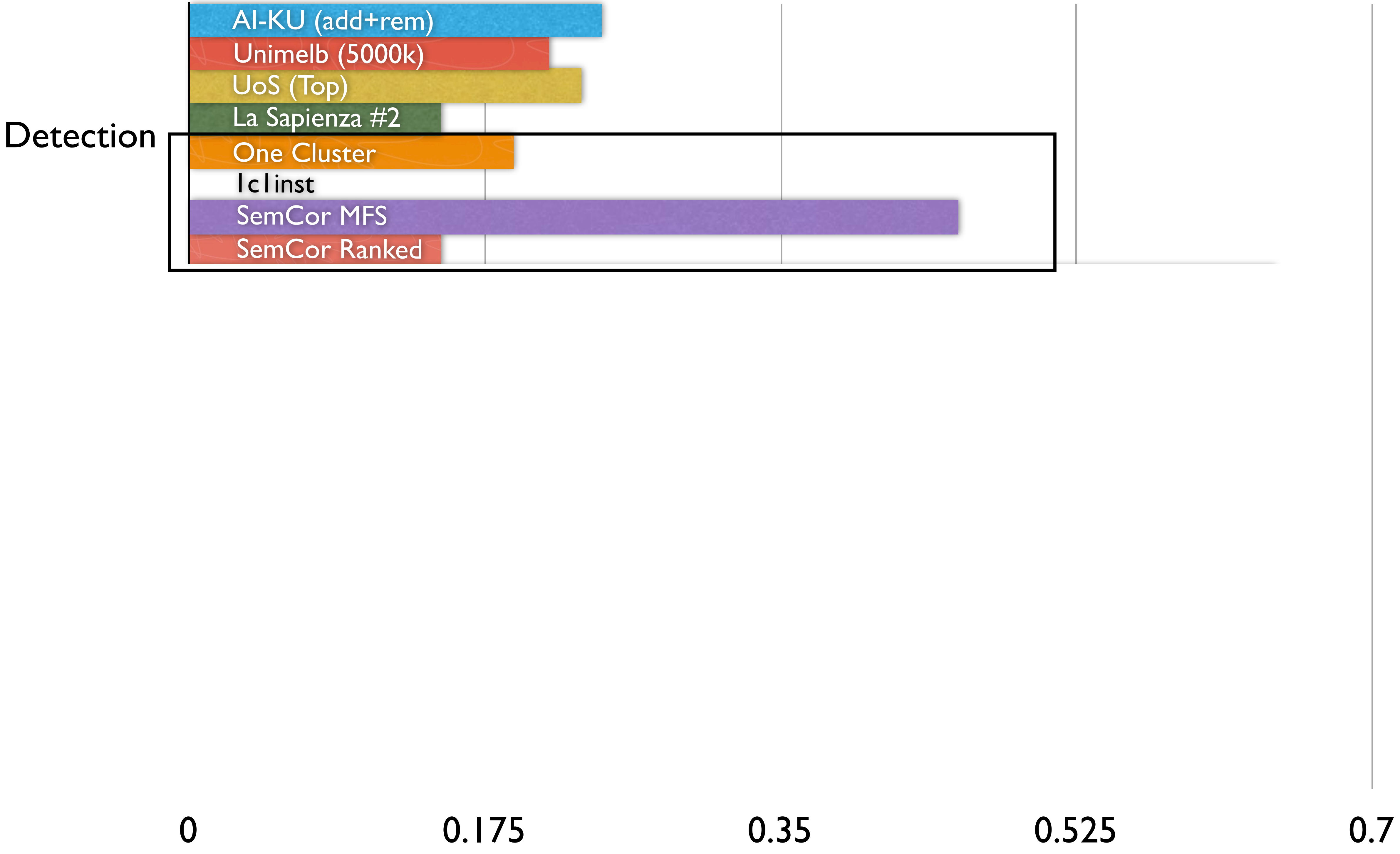
WSI Results



+ One Cluster
✕ AI-KU (add 1000)
+ Unimelb (50k)
+ 1c1inst
△ AI-KU (add 1000, remove 5)
○ UoS (WN)
▽ AI-KU
◇ Unimelb (5p)
□ UoS (Top)
0.7
0.525
0.35
0.175
0
0
0.02
0.04
0.06
0.08
Fuzzy B-Cubed
Fuzzy NMI

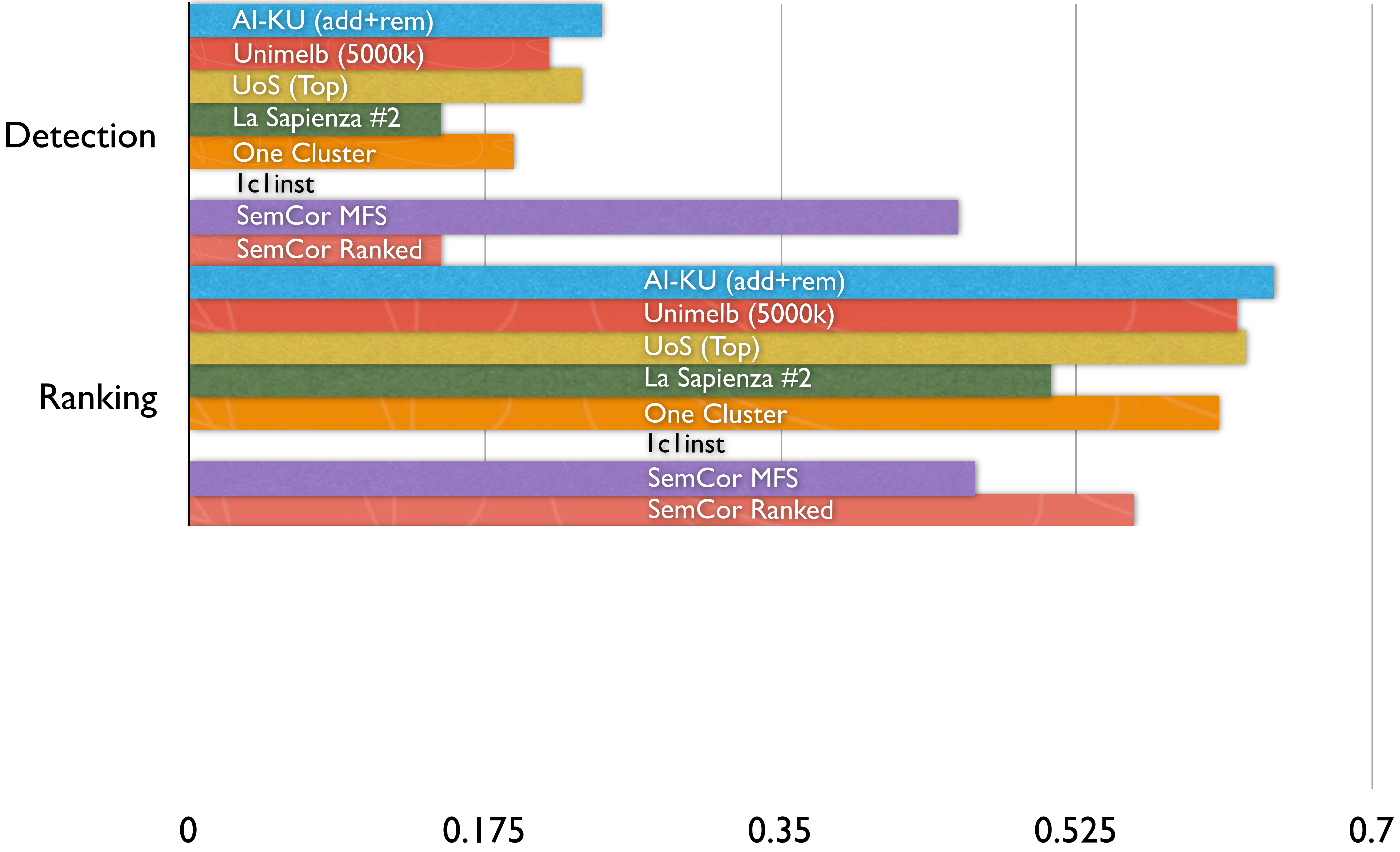
🟦 AI-KU (add+rem) 🟥 Unimelb (5000k) 🟨 UoS (Top) 🟩 La Sapienza #2
🟧 One cluster (WSI) 🟪 1c1inst (WSI) 🟪 Semcor MFS 🟥 Semcor Ranked

WSD Results

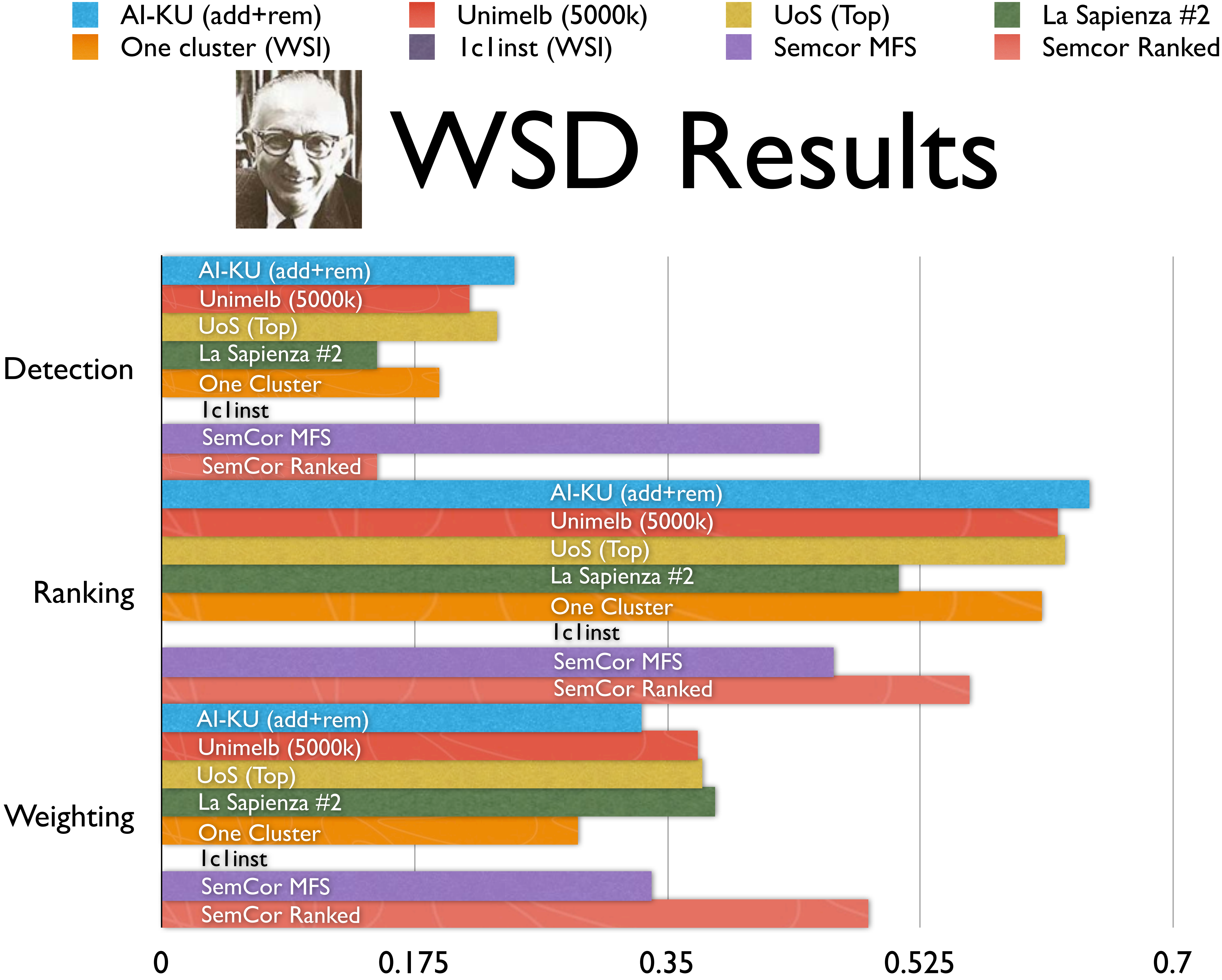


🟦 AI-KU (add+rem) 🟥 Unimelb (5000k) 🟨 UoS (Top) 🟩 La Sapienza #2
🟧 One cluster (WSI) 🟪 1c1inst (WSI) 🟪 Semcor MFS 🟥 Semcor Ranked

WSD Results



WSD Results


AI-KU (add+rem)
One cluster (WSI)
Unimelb (5000k)
1c1inst (WSI)
UoS (Top)
Semcor MFS
La Sapienza #2
Semcor Ranked
Detection
Ranking
Weighting
AI-KU (add+rem)
Unimelb (5000k)
UoS (Top)
La Sapienza #2
One Cluster
1c1inst
SemCor MFS
SemCor Ranked
0
0.175
0.35
0.525
0.7

Issues with Evaluation

Multi-sense Annotation Rate

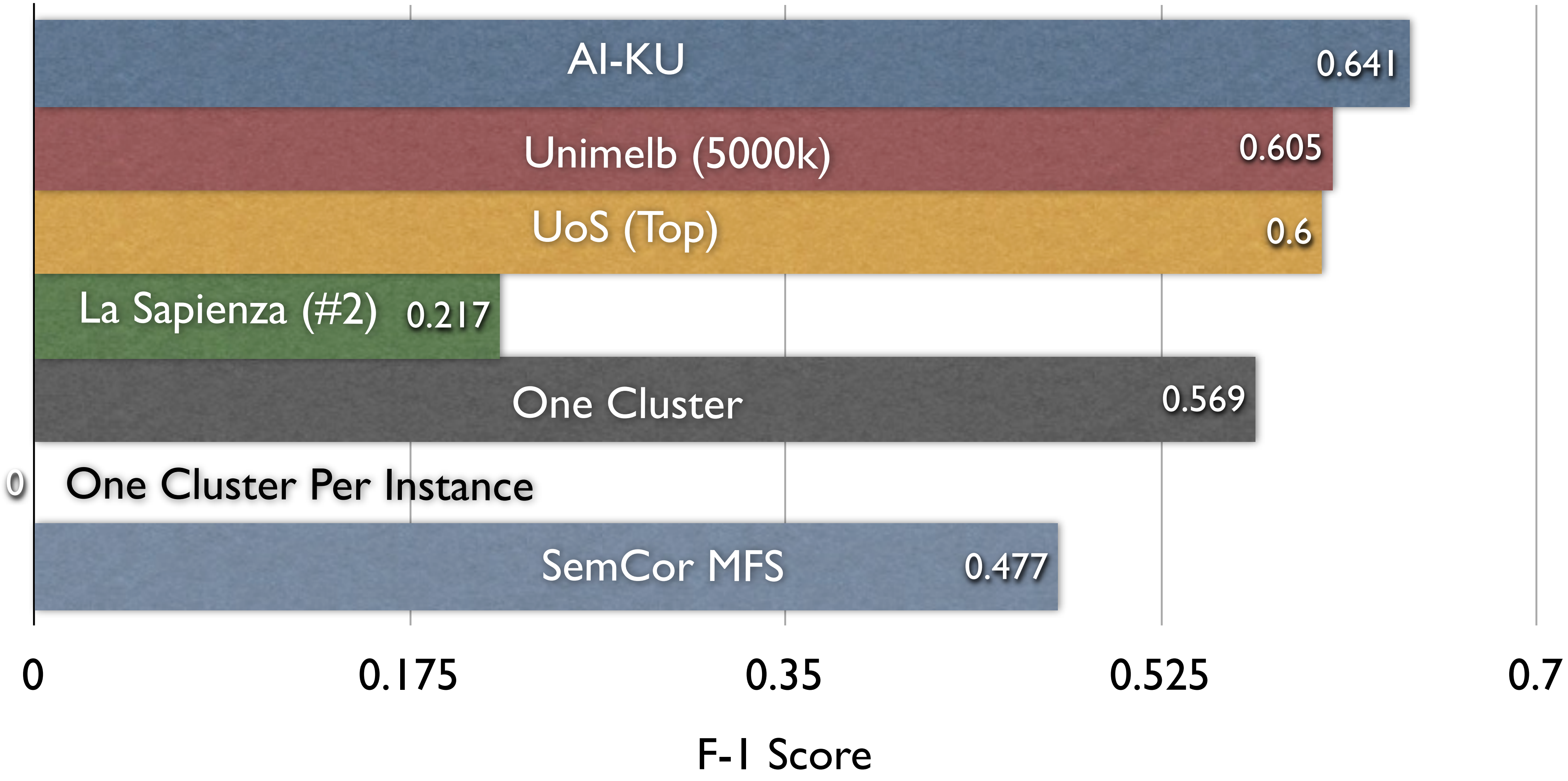
Trial	**Test**
100%	11%

Task 13 evaluation measures specifically designed for multiple senses

Evaluation #2

- Modify the WSI mapping procedure to only produce a single sense
- Modify WSD systems to retain only highest-weighted sense

WSD Results for single-sense instances



Conclusions

- Multiple sense annotations offers a way to improve annotation by making ambiguity explicit
- WSI offer some hope for creating highly accurate semi-supervised systems

Future Work

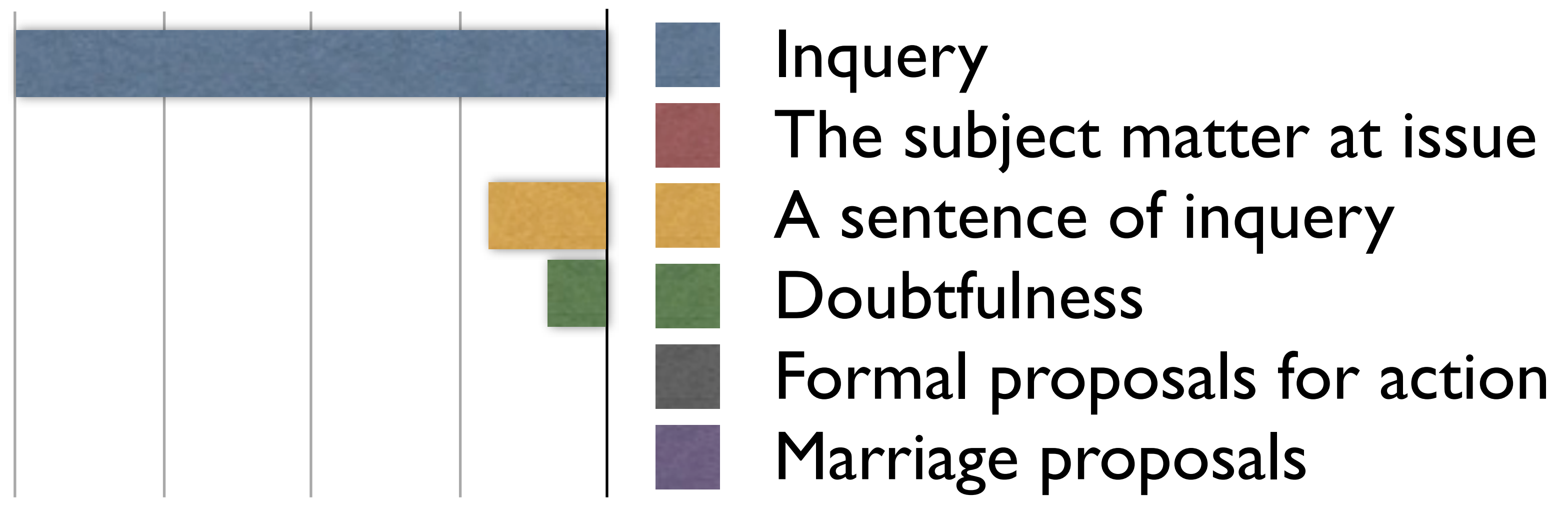
- Embed this application in a task
  - Task 11 extension with multiple labels?
- Have systems annotate *why* an instance needs multiple senses
- Build WSI sense mapping on an external tuning corpus

Summary

- All resources released on the Task 13 website: http://www.cs.york.ac.uk/semeval-2013/task13/
- All evaluation scoring and IAA code is released on Google code https://code.google.com/p/cluster-comparison-tools/
- Annotations (hopefully) being folded into MASC

SemEval-2013 Task 13: **Word Sense Induction for Graded and Non-Graded Senses**

Any **questions**?



David Jurgens and Ioannis Klapaftis

jurgens@di.uniroma1.it ioannisk@microsoft.com