Beyond the Query Log

Detecting Missions in Personal Web Archives

Ludwig Lorenz

March 7, 2024

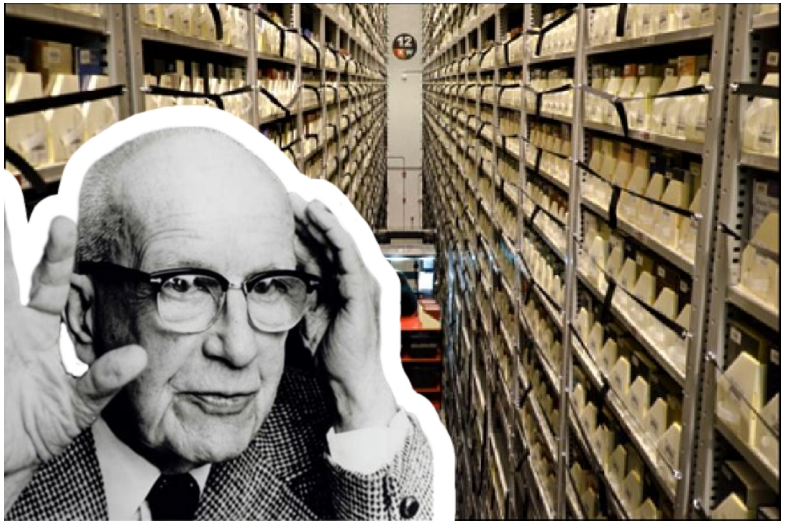
Research Group for Intelligent Information Systems, Bauhaus-Universität Weimar

Motivation

Personal Knowledge - The Past

Zettelkasten

A box with connected index cards was used by sociologist Niklas Luhmann from 1952 until 1997 to organize personal knowledge. He referred to it as a tool for thought.



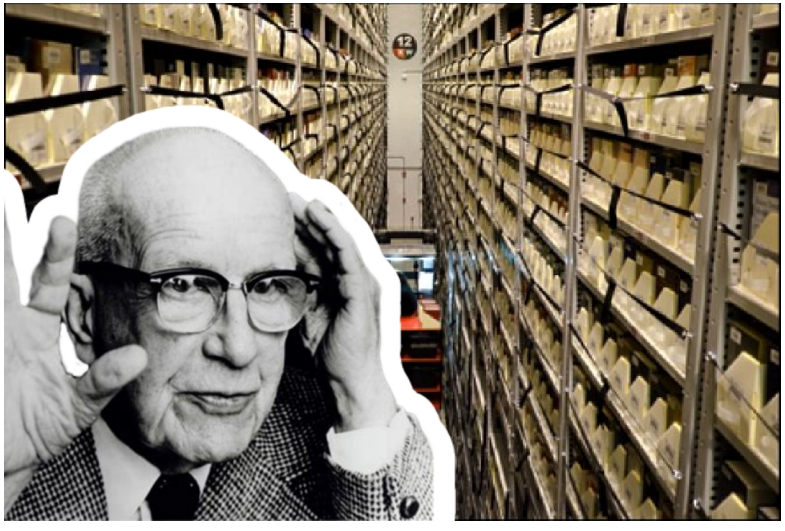
Dymaxion Chronofile

... is a collection of scrapbooks by R. Buckminster Fuller that documents his life from 1917 to 1983 detailed up to the hourly level. There is no judge of what is valid to put in or not.

Personal Knowledge - The Past

Zettelkasten

A box with connected index cards was used by sociologist Niklas Luhmann from 1952 until 1997 to organize personal knowledge. He referred to it as a tool for thought.



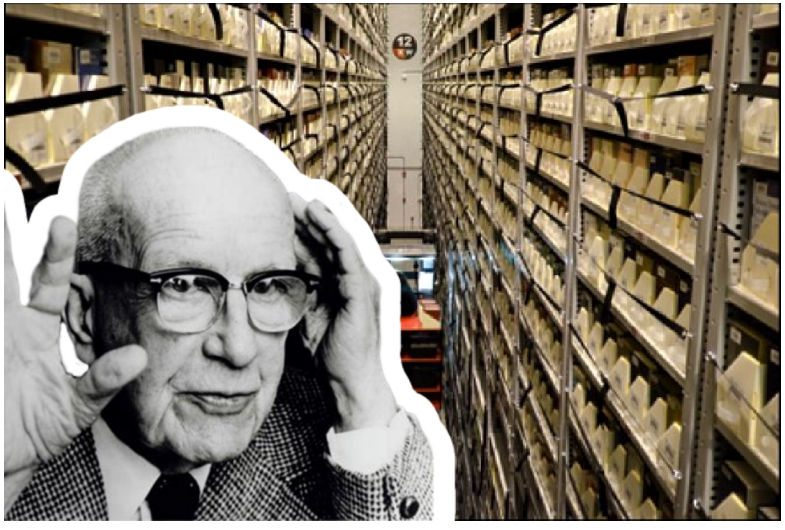
Dymaxion Chronofile

... is a collection of scrapbooks by R. Buckminster Fuller that documents his life from 1917 to 1983 detailed up to the hourly level. There is no judge of what is valid to put in or not.

Personal Knowledge - The Past

Zettelkasten

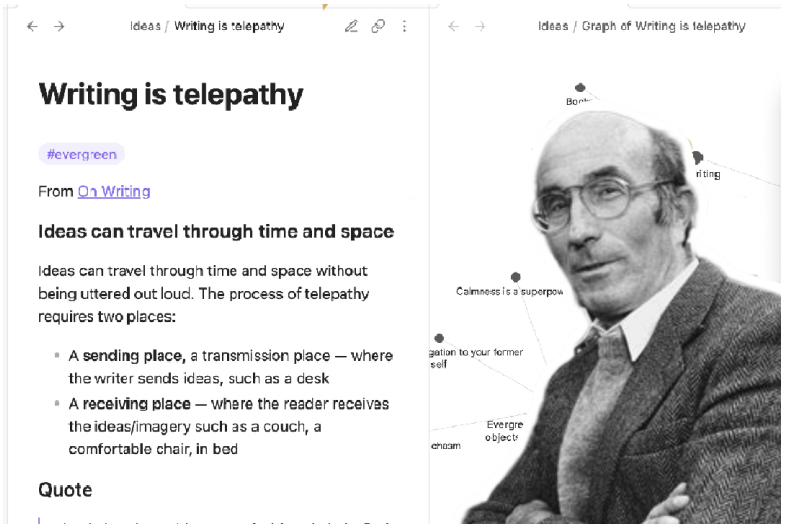
A box with connected index cards was used by sociologist Niklas Luhmann from 1952 until 1997 to organize personal knowledge. He referred to it as a tool for thought.



Dymaxion Chronofile

... is a collection of scrapbooks by R. Buckminster Fuller that documents his life from 1917 to 1983 detailed up to the hourly level. There is no judge of what is valid to put in or not.

Personal Knowledge - The Present



Second Brain

Advanced personal knowledge management tools like *obsidian.md* or *Notion* allow to create a digital Zettelkasten of connected personal notes.

Lifelogging

… attempts to create personal records of one person's life in all it's aspects. E.g. heart rate, location, step count etc.

Personal Knowledge - The Future?

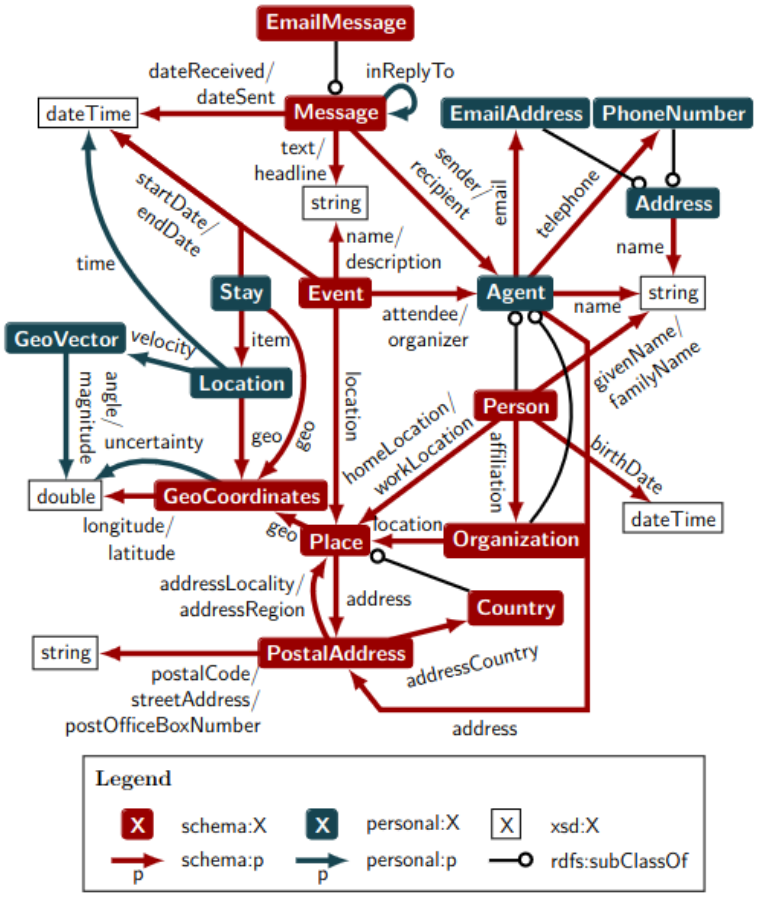


Figure 1: *Thymeflow* system architecture by David Montoya. [5]

- merging personal knowledge and the different streams of digital traces we produce every day
- *Thymeflow* combines personal data and personal knowledge with a central query interface
- multiple sources
  - location
  - emails
  - calendar
  - contacts
- what about your web history?

Personal Knowledge - The Future?

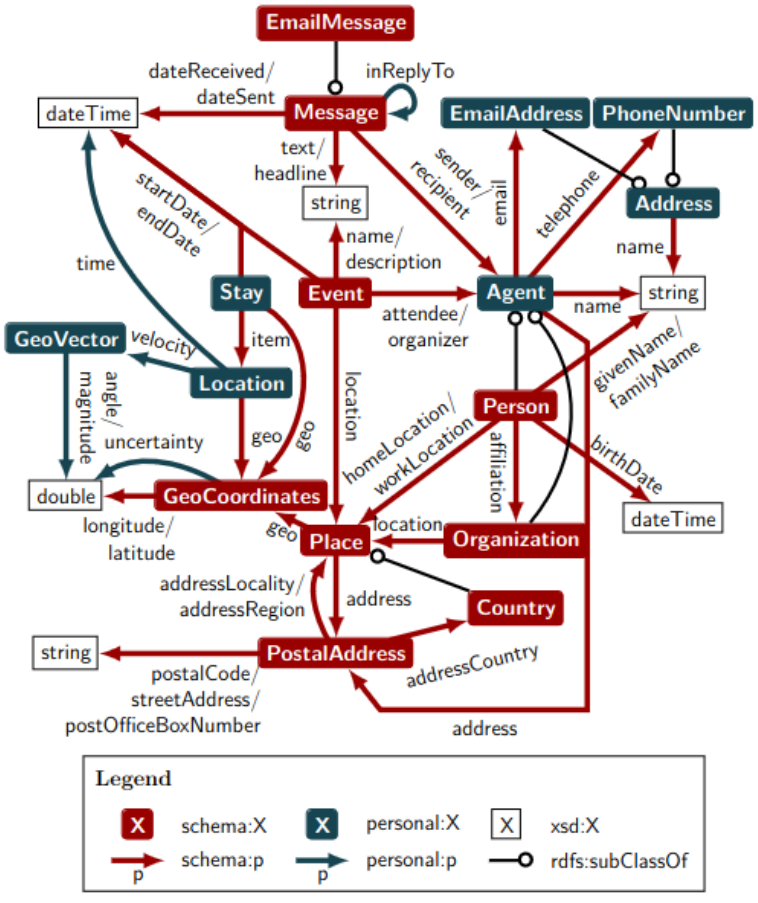


Figure 1: *Thymeflow* system architecture by David Montoya. [5]

- merging personal knowledge and the different streams of digital traces we produce every day
- *Thymeflow* combines personal data and personal knowledge with a central query interface
- multiple sources
  - location
  - emails
  - calendar
  - contacts
- what about your web history?

Personal Knowledge - The Future?

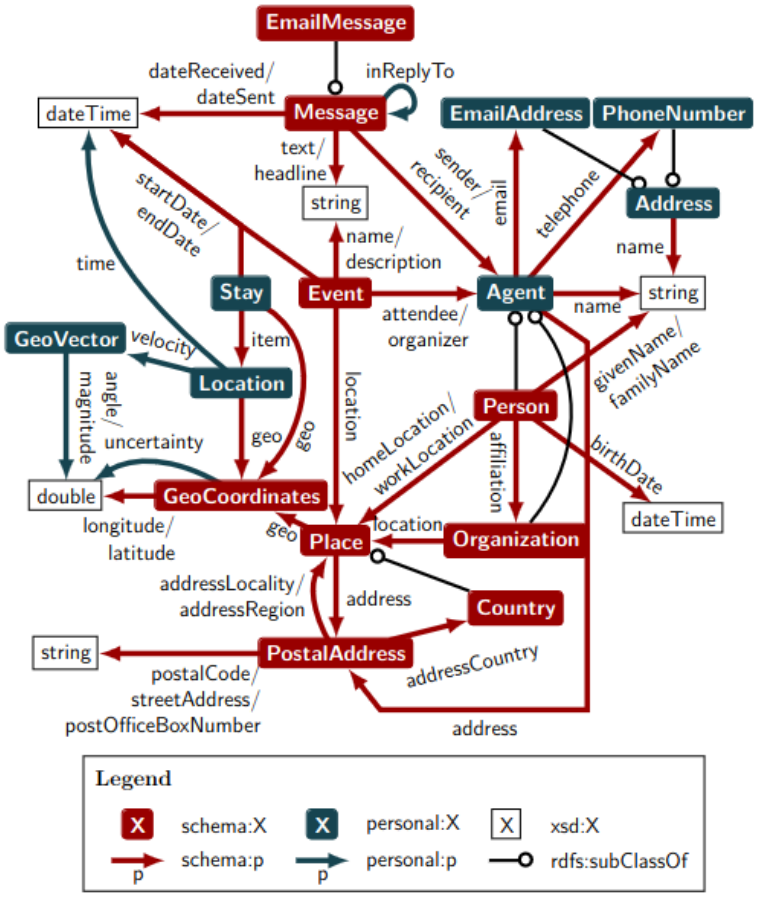


Figure 1: *Thymeflow* system architecture by David Montoya. [5]

- merging personal knowledge and the different streams of digital traces we produce every day
- *Thymeflow* combines personal data and personal knowledge with a central query interface
- multiple sources
  - location
  - emails
  - calendar
  - contacts
- what about your web history?

Thymeflow abstracted stays from a log of location data.

How can we abstract the task behind browsing the web from a log of visited websites?

Thymeflow abstracted stays from a log of location data.

How can we abstract the task behind browsing the web from a log of visited websites?

Table of contents

Important Concepts

From a Query Log ...

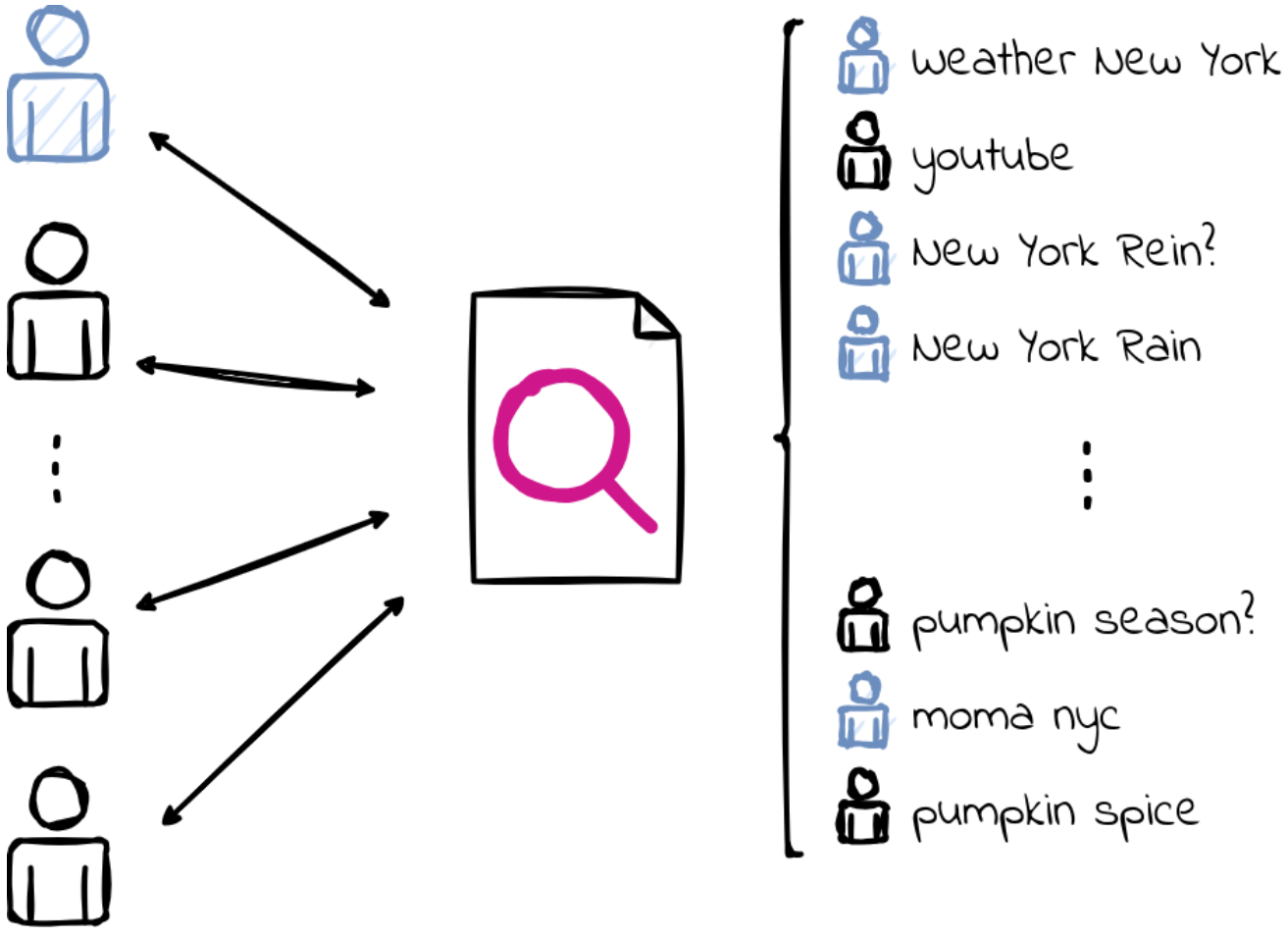


Figure 2: Visual explanation of a query log.

... to a Visit Log

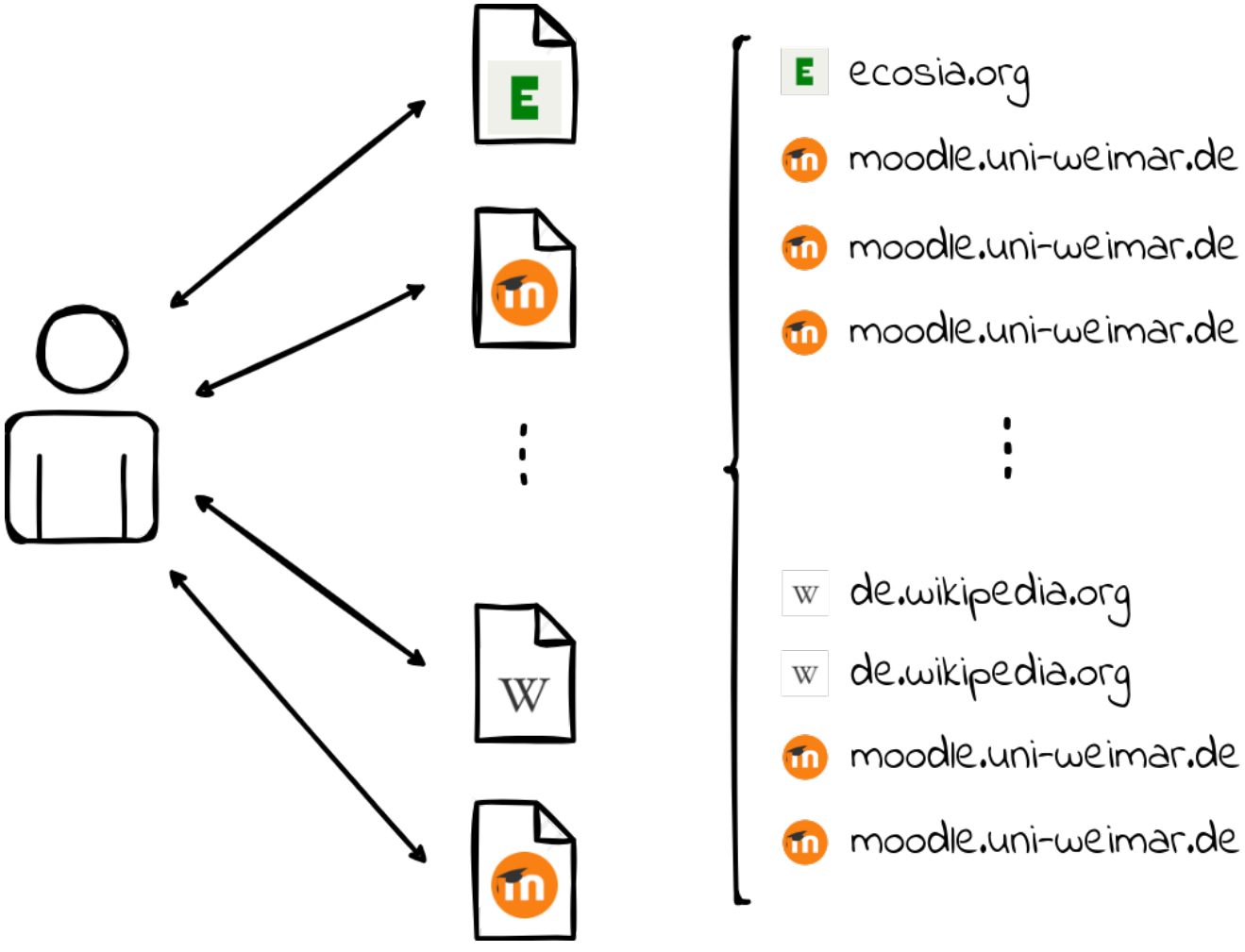
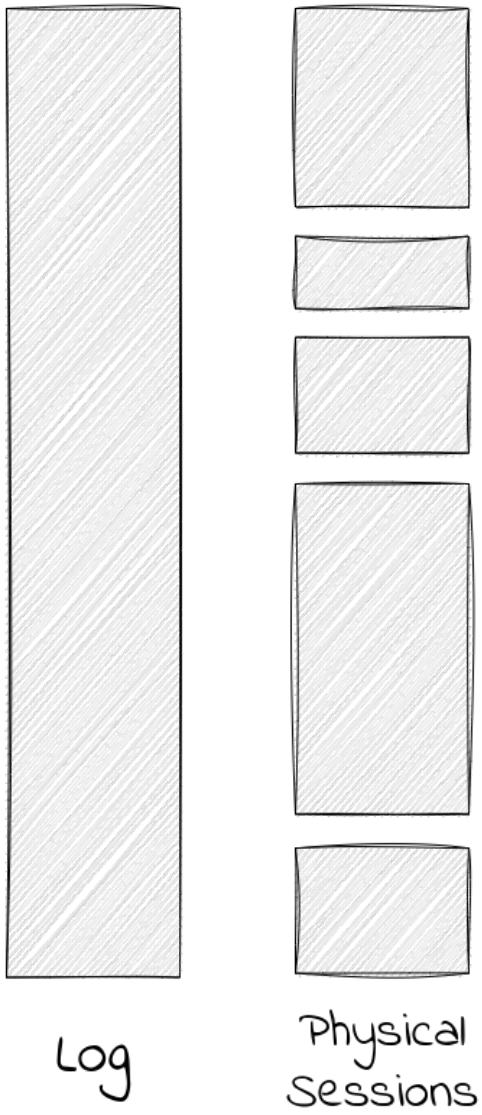
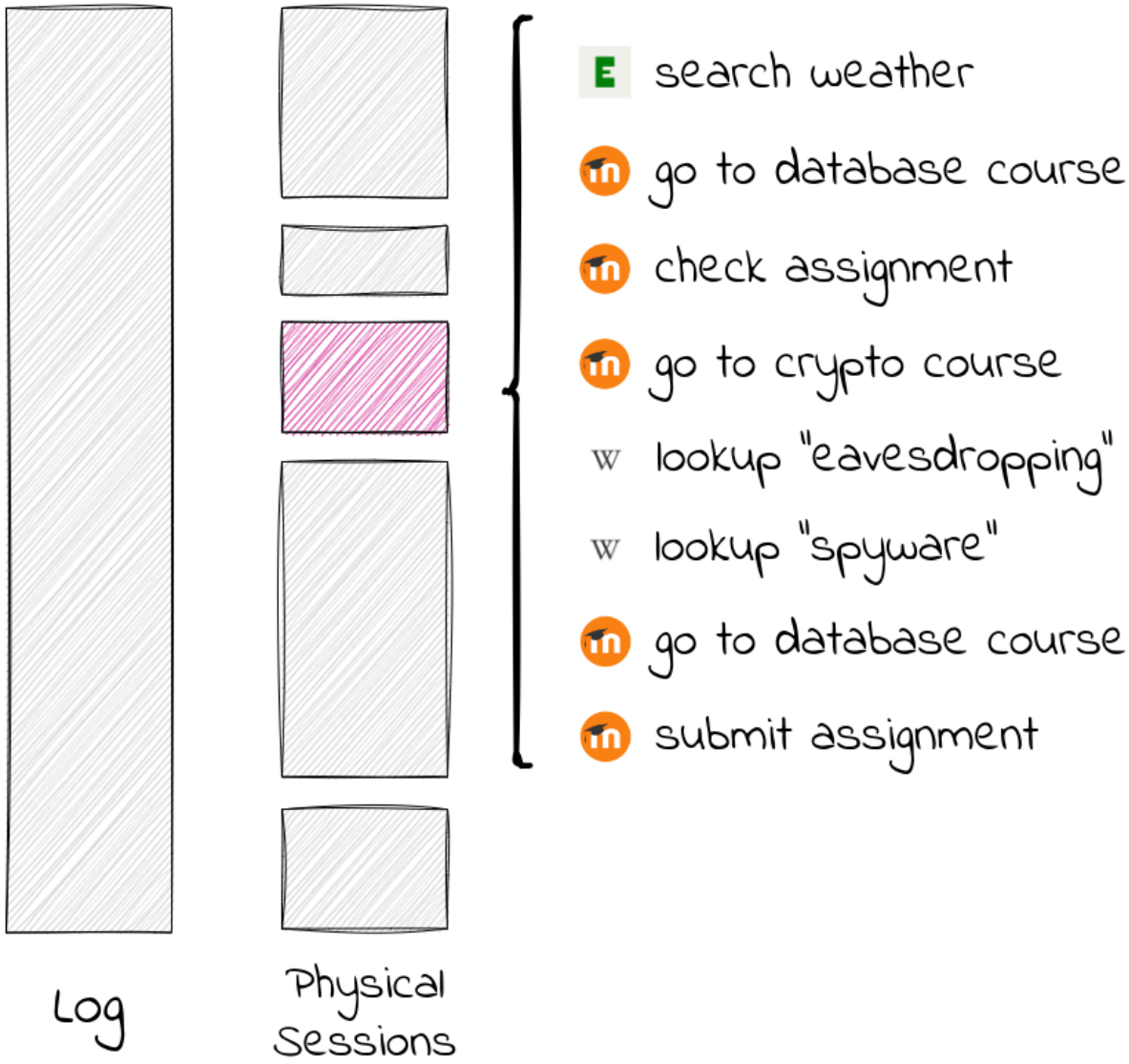


Figure 3: Visual explanation of a visit log.

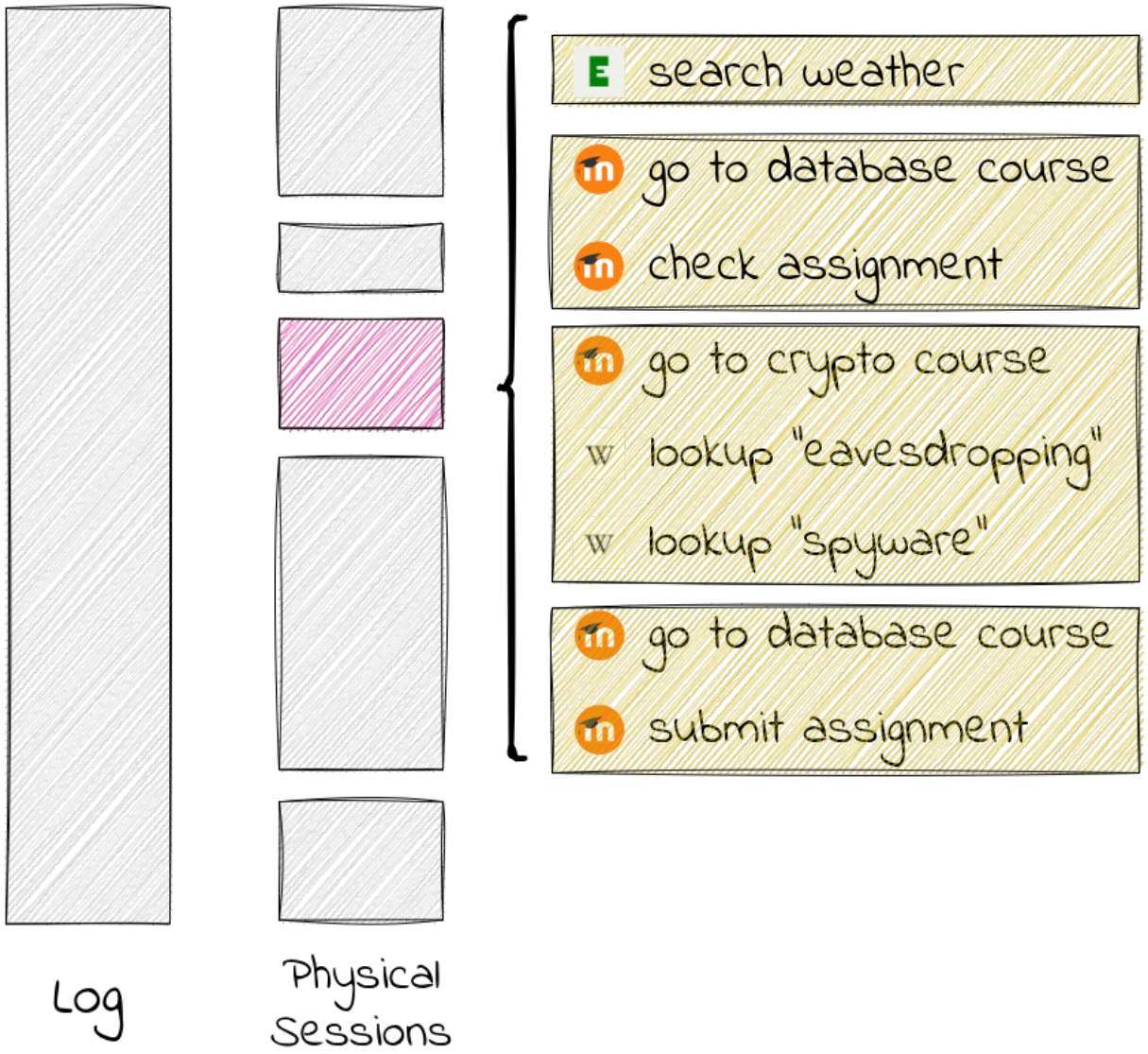
What is a Physical Session? (ϕ)



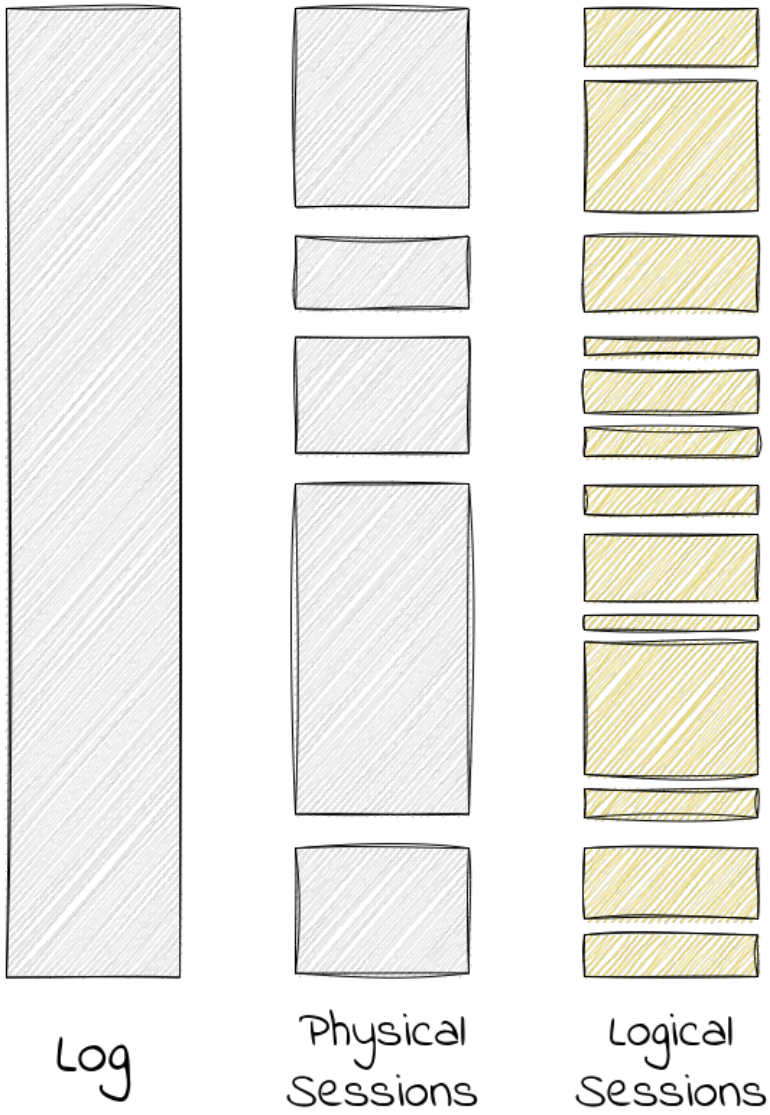
What is a Physical Session? (ϕ)



What is a logical session? (λ)



What is a logical session? (λ)



What is a Mission? (μ)

- $\mathcal{VL}$ denotes a visit log
- λ_i is a logical session, Λ is the set of all logical sessions on the visit log $\mathcal{VL}$.

Definition (Mission μ)

Let $\lambda_i, \lambda_j \in \Lambda$ be any two logical sessions in the visit log. λ_i, λ_j are subsets of the same mission $\mu \subseteq \mathcal{VL}$ iff they were issued for the same task or goal. The set of all missions is denoted as M.

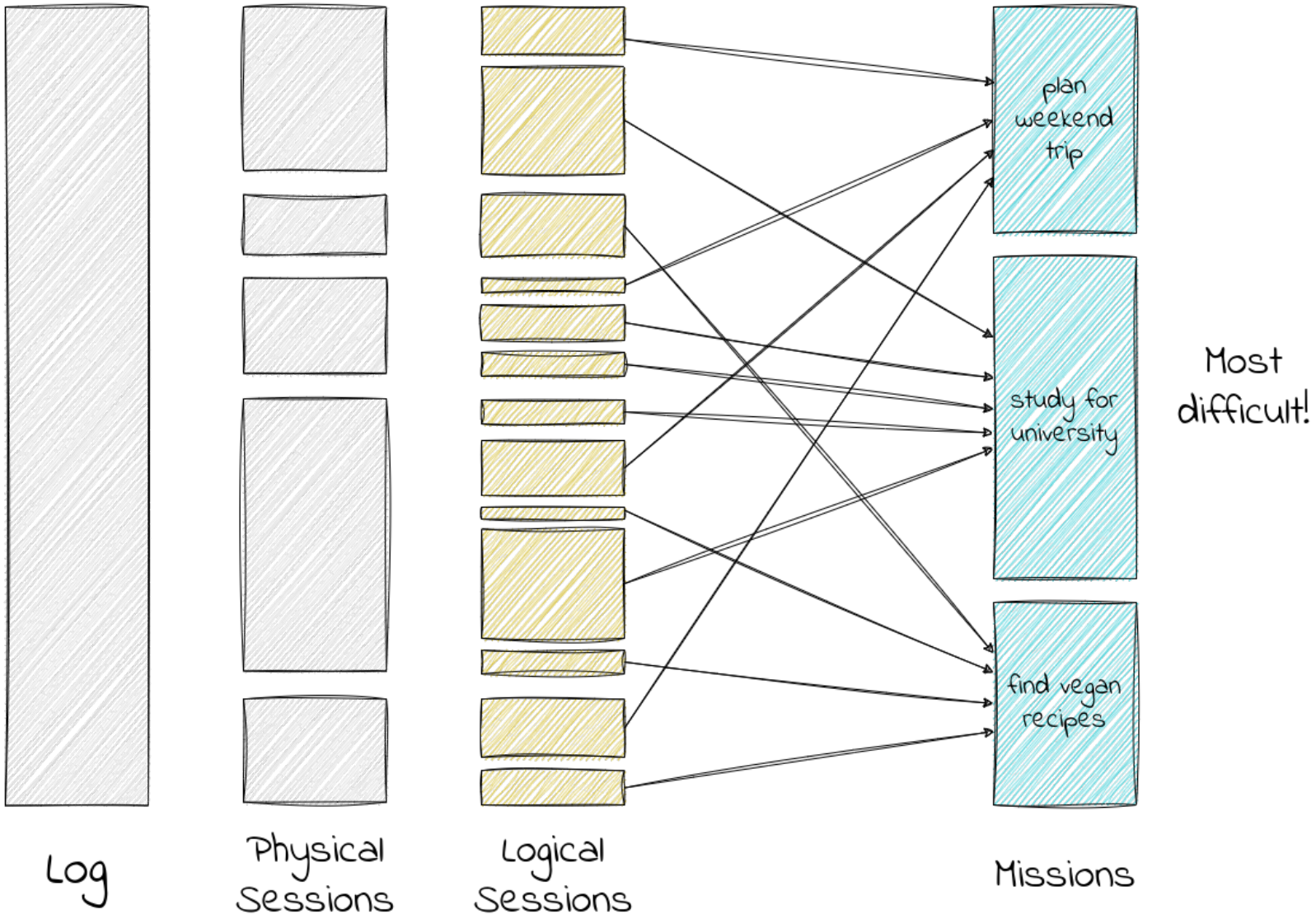
What is a Mission? (μ)

- $\mathcal{VL}$ denotes a visit log
- λ_i is a logical session, Λ is the set of all logical sessions on the visit log $\mathcal{VL}$.

Definition (Mission μ)

Let $\lambda_i, \lambda_j \in \Lambda$ be any two logical sessions in the visit log. λ_i, λ_j are subsets of the same mission $\mu \subseteq \mathcal{VL}$ iff they were issued for the same task or goal. The set of all missions is denoted as M.

What is a Mission? (μ)



Hierarchical Missions

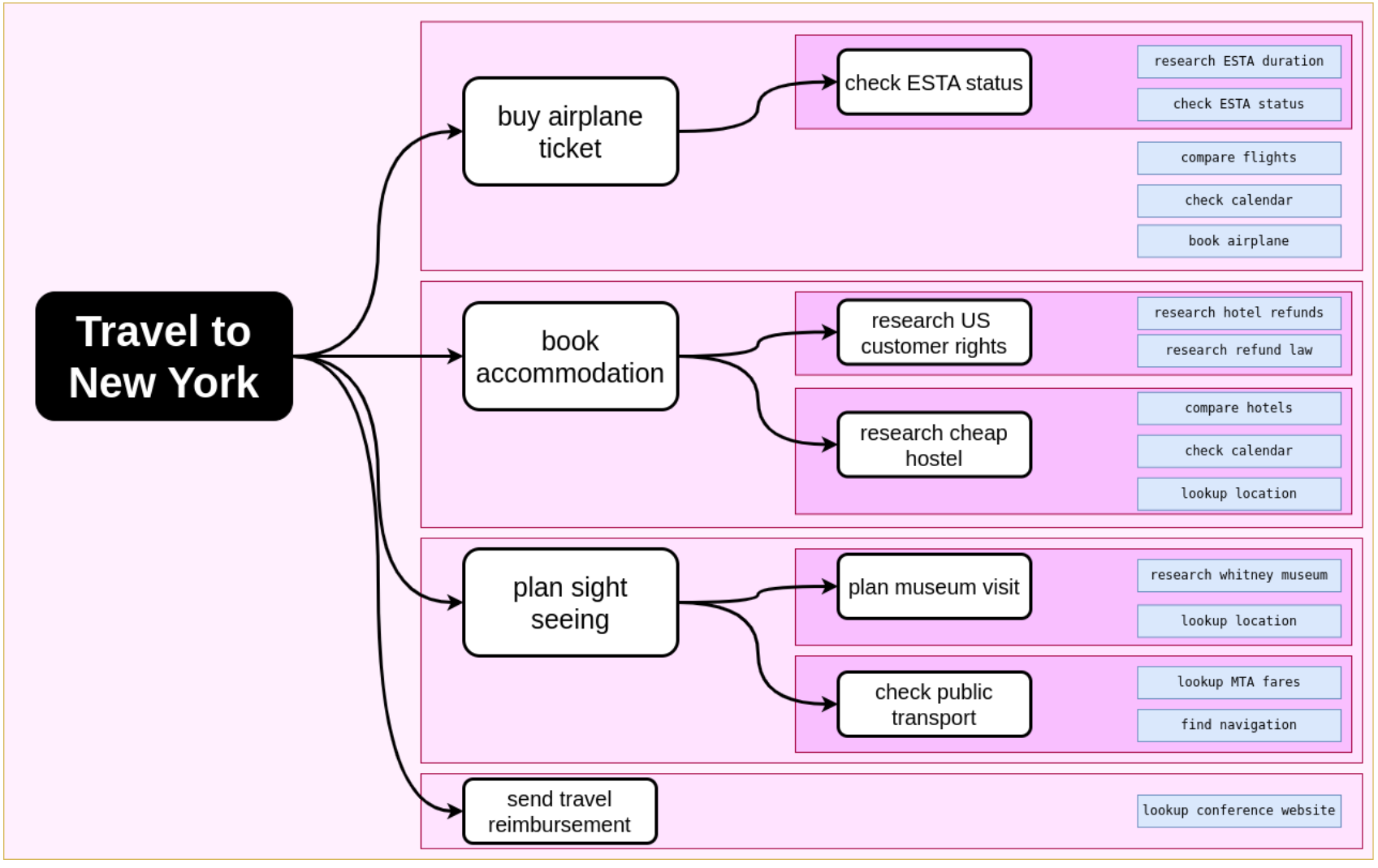


Figure 4: Missions can be defined with varying granularity.

Main Challenges

1. Record!

Record a personal web archive over a time span of one month as a dataset for the detection.

2. Annotate!

Annotate logical sessions and (hierarchical) missions as a ground truth for evaluation.

3. Detect!

Construct and run algorithms to segment the visit log in logical sessions and missions.

4. Evaluate!

Define evaluation measures to compare the results with the ground truth.

Dataset Recording

Experimental Setup

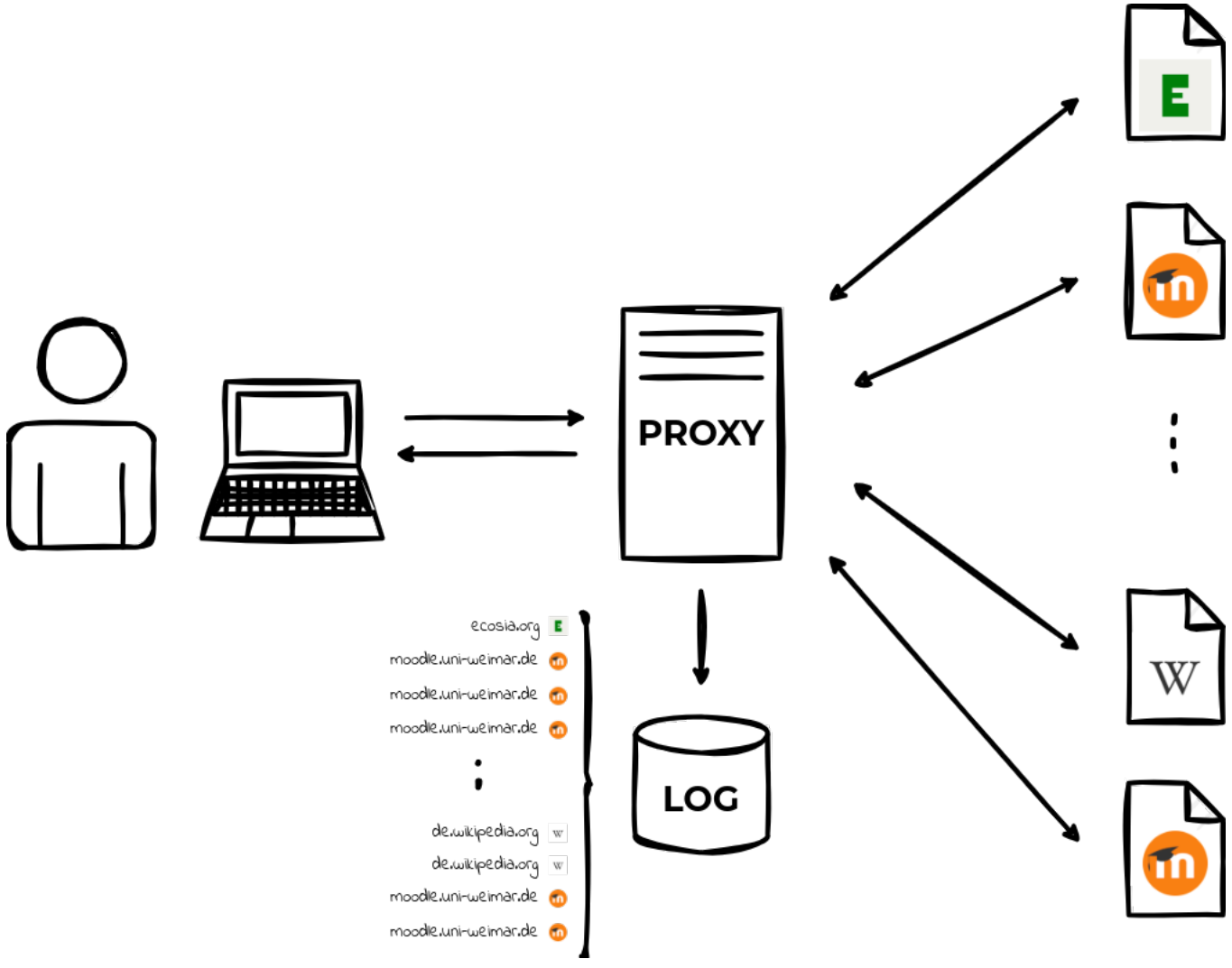


Figure 5: Experimental setup for dataset recording. The *WASP project* [3] is used as a proxy server and to archive the visited web pages.

Annotation

Annotation for Logical Sessions

TorontoView - Session # 1670516996

Time	Domain	Title	Payload	Session	Mission
16:29:56	luftlinie-berechnen.de	Luftlinie berechnen - Online Rechner mit Karte	26.96kB	0	
16:30:18	luftlinie-berechnen.de	Redirecting to https://luftlinie-berechnen.de/berlin/weimar	809B	0	
16:30:20	luftlinie-berechnen.de	Luftlinie zwischen Berlin und Weimar	21.46kB	0	
16:30:45	luftlinie-berechnen.de	Redirecting to https://luftlinie-berechnen.de/montreal/toronto	821B	0	
16:30:49	luftlinie-berechnen.de	Luftlinie zwischen Montreal und Toronto	20.33kB	0	
17:19:38	www.ecosia.org	css snowflake pattern - Ecosia - Web	168.38kB	1	
17:20:15	github.com	dmolsen/CSS3-Snowflakes: Some simple CSS for creating snowflakes as well as some JavaScript for quickly adding them to your website.	280.12kB	1	
17:20:19	github.com	dmolsen/CSS3-Snowflakes: Some simple CSS for creating snowflakes as well as some	280.12kB	1	

Figure 6: Screenshot of the annotation interface for *logical sessions*.

Annotation for Missions

Figure 7: Screenshot of the annotation interface for *missions*.

Annotation

- in a timespan of a month 1418 intentional website visits [a] were recorded
- during annotation, the log was segmented into 84 physical sessions, 294 logical sessions and 75 (hierarchical) missions

Interesting Observations

During annotation various phenomena e.g. ”information need digression” or ”session entry points” were identified.

[a] 4561 unintentional visits were cleared from the log during annotation

Annotation

- in a timespan of a month 1418 intentional website visits [a] were recorded
- during annotation, the log was segmented into 84 physical sessions, 294 logical sessions and 75 (hierarchical) missions

[a] 4561 unintentional visits were cleared from the log during annotation

Interesting Observations

During annotation various phenomena e.g. ”information need digression” or ”session entry points” were identified.

Detecting Logical Sessions

Algorithmic Approach

- different features are considered and measured, if the sum of all measurings exceeds a certain ”continuation threshold” a session break is detected and a it continues with a new session.
- time-based feature
- content-based feature
  - Jaccard coefficient of keywords
  - Jaccard coefficient of links

Basic Features

Visit logs are a novel data structure. Assessing the effectiveness of basic features will help to construct more complex features.

Algorithmic Approach

- different features are considered and measured, if the sum of all measurings exceeds a certain ”continuation threshold” a session break is detected and a it continues with a new session.
- time-based feature
- content-based feature
  - Jaccard coefficient of keywords
  - Jaccard coefficient of links

Basic Features

Visit logs are a novel data structure. Assessing the effectiveness of basic features will help to construct more complex features.

Logical Session Detection Evaluation

Precision: How many of the *detected* breaks are actually session breaks?

- Here, the time feature scores best. (Precision: 0.507)

Recall: How many of the actual session breaks are *detected*?

- Here, the content-based features score best. (Recall: 1.0)

Simple is effective

Although it has no perfect recall, the time feature scores best when taking both - precision and recall - into account. ($F_{1.5}$: 0.636)

Logical Session Detection Evaluation

Precision: How many of the *detected* breaks are actually session breaks?

- Here, the time feature scores best. (Precision: 0.507)

Recall: How many of the actual session breaks are *detected*?

- Here, the content-based features score best. (Recall: 1.0)

Simple is effective

Although it has no perfect recall, the time feature scores best when taking both - precision and recall - into account. ($F_{1.5}$: 0.636)

Logical Session Detection Evaluation

Precision: How many of the *detected* breaks are actually session breaks?

- Here, the time feature scores best. (Precision: 0.507)

Recall: How many of the actual session breaks are *detected*?

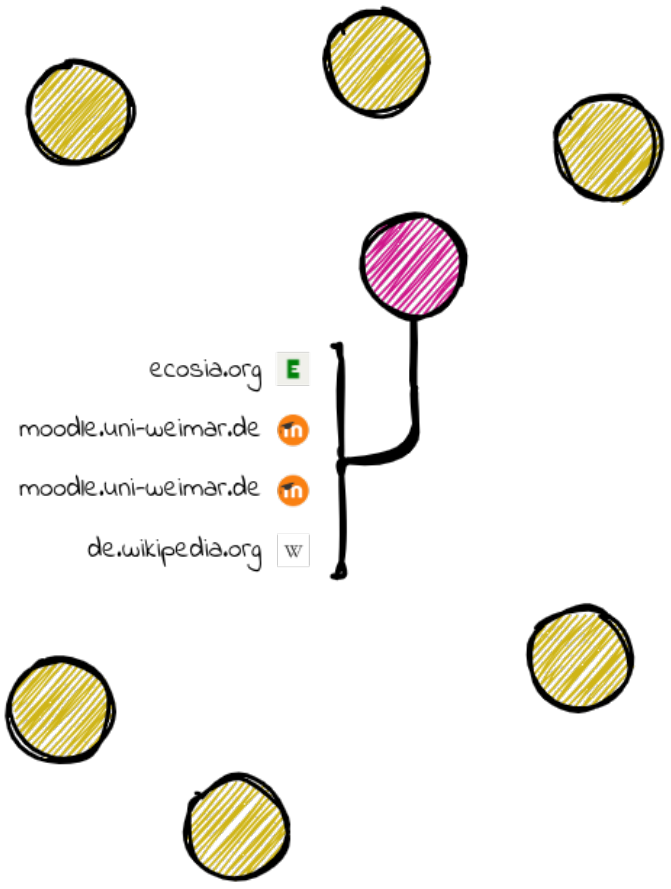
- Here, the content-based features score best. (Recall: 1.0)

Simple is effective

Although it has no perfect recall, the time feature scores best when taking both - precision and recall - into account. ($F_{1.5}$: 0.636)

Detecting Missions

Algorithmic Approach for Mission Detection

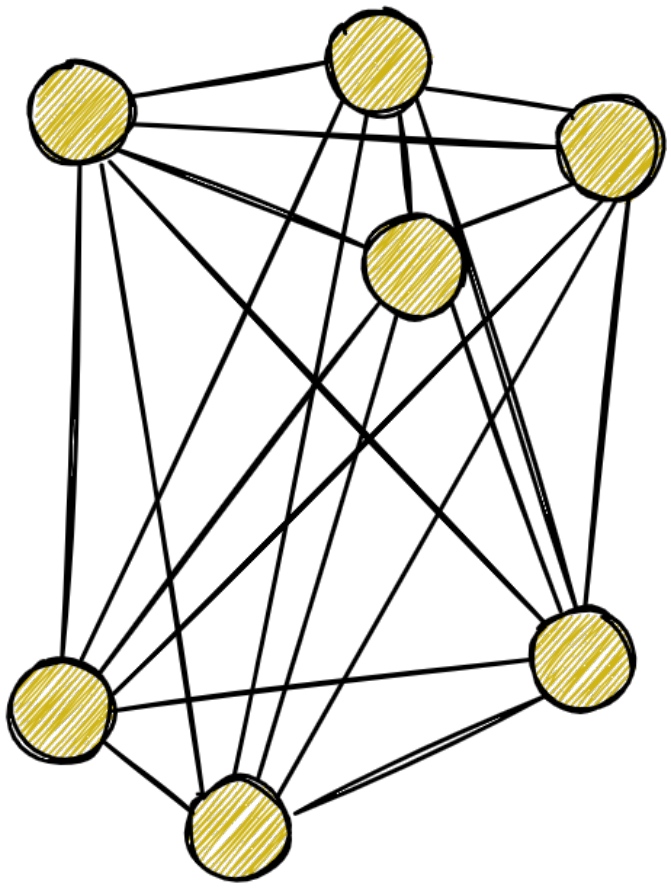


Algorithm 1 WCC Clustering

Require: $\Lambda \subseteq \mathcal{P}(\mathcal{VL})$

1: $V = \Lambda$
2: $E' = \Lambda \times \Lambda$
3: **for all** $\lambda_i \in V$ **do**
4: **for all** $\lambda_j \in V$ **do**
5: $d_j = \text{distance}(\lambda_i, \lambda_j)$
6: **if** $d_j > c_\mu$ **then**
7: $E' = E' \setminus \{(\lambda_i, \lambda_j)\}$
8: **end if**
9: **end for**
10: **end for**
11: $M' = \text{Component-DFS}(V, E')$

Algorithmic Approach for Mission Detection

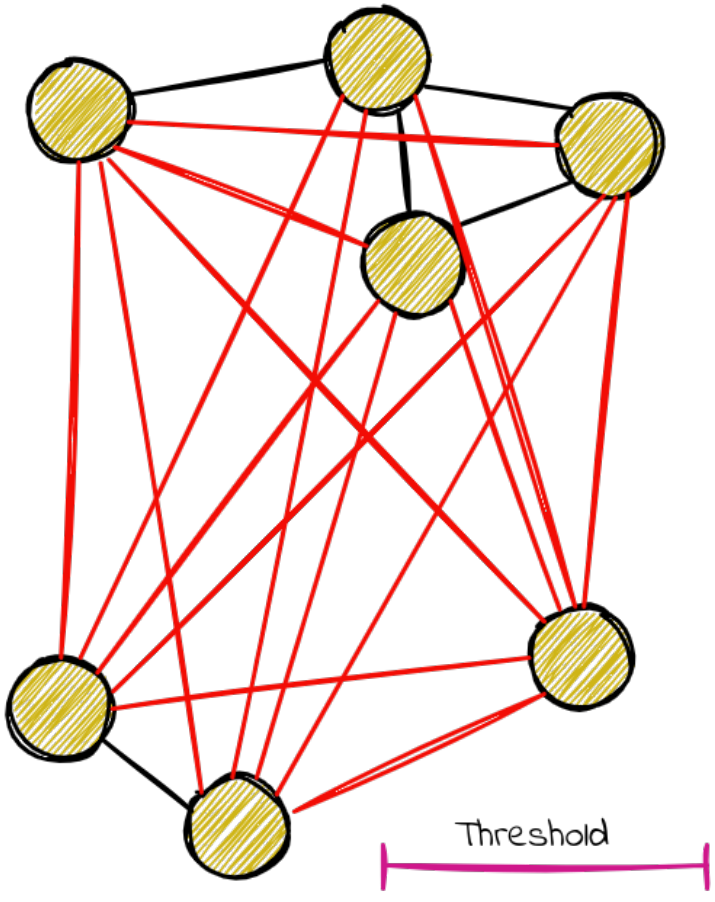


Algorithm 1 WCC Clustering

Require: $\Lambda \subseteq \mathcal{P}(\mathcal{VL})$

1: $V = \Lambda$
2: $E' = \Lambda \times \Lambda$
3: **for all** $\lambda_i \in V$ **do**
4: **for all** $\lambda_j \in V$ **do**
5: $d_J = \text{distance}(\lambda_i, \lambda_j)$
6: **if** $d_J > c_\mu$ **then**
7: $E' = E' \setminus \{(\lambda_i, \lambda_j)\}$
8: **end if**
9: **end for**
10: **end for**
11: $M' = \text{Component-DFS}(V, E')$

Algorithmic Approach for Mission Detection

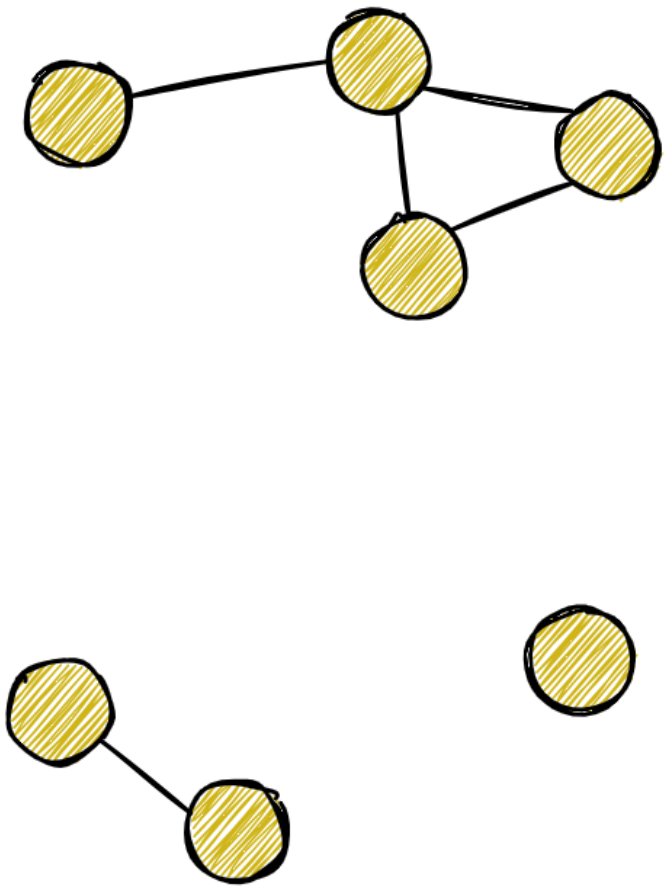


Algorithm 1 WCC Clustering

Require: $\Lambda \subseteq \mathcal{P}(\mathcal{VL})$

1: $V = \Lambda$
2: $E' = \Lambda \times \Lambda$
3: **for all** $\lambda_i \in V$ **do**
4: **for all** $\lambda_j \in V$ **do**
5: $d_J = \text{distance}(\lambda_i, \lambda_j)$
6: **if** $d_J > c_\mu$ **then**
7: $E' = E' \setminus \{(\lambda_i, \lambda_j)\}$
8: **end if**
9: **end for**
10: **end for**
11: $M' = \text{Component-DFS}(V, E')$

Algorithmic Approach for Mission Detection

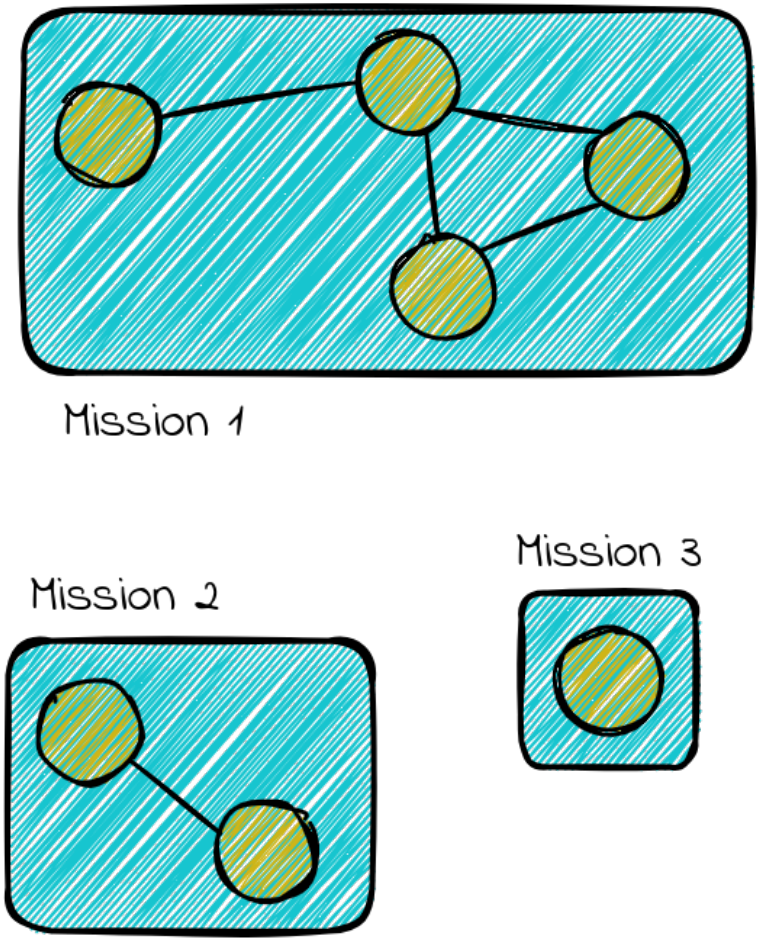


Algorithm 1 WCC Clustering

Require: $\Lambda \subseteq \mathcal{P}(\mathcal{VL})$

1: $V = \Lambda$
2: $E' = \Lambda \times \Lambda$
3: **for all** $\lambda_i \in V$ **do**
4: **for all** $\lambda_j \in V$ **do**
5: $d_J = \text{distance}(\lambda_i, \lambda_j)$
6: **if** $d_J > c_\mu$ **then**
7: $E' = E' \setminus \{(\lambda_i, \lambda_j)\}$
8: **end if**
9: **end for**
10: **end for**
11: $M' = \text{Component-DFS}(V, E')$

Algorithmic Approach for Mission Detection



Algorithm 1 WCC Clustering

Require: $\Lambda \subseteq \mathcal{P}(\mathcal{VL})$

1: $V = \Lambda$
2: $E' = \Lambda \times \Lambda$
3: **for all** $\lambda_i \in V$ **do**
4: **for all** $\lambda_j \in V$ **do**
5: $d_J = \text{distance}(\lambda_i, \lambda_j)$
6: **if** $d_J > c_\mu$ **then**
7: $E' = E' \setminus \{(\lambda_i, \lambda_j)\}$
8: **end if**
9: **end for**
10: **end for**
11: $M' = \text{Component-DFS}(V, E')$

The *distance*[1] is calculated as the harmonic mean of the Jaccard coefficients for the aggregated keywords and links in a logical session.

$$\text{distance}(\lambda_i, \lambda_j) = 1 - \frac{J(keywords(\lambda_i), keywords(\lambda_j)) + J(links(\lambda_i), links(\lambda_j))}{2}$$

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

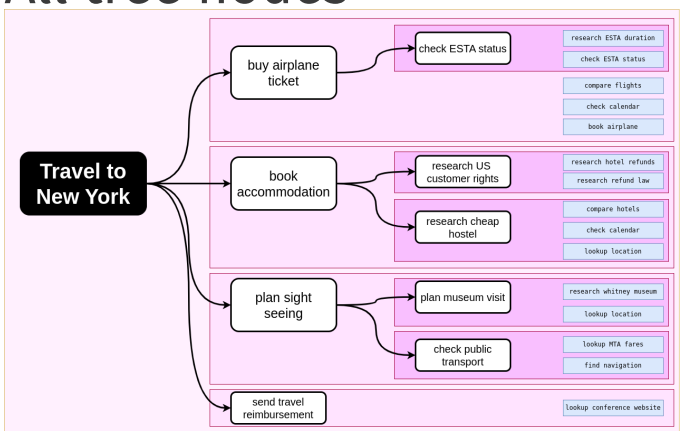
[1] this is not a distance in the notion of a metric space since the triangle inequality is not fulfilled

Algorithmic Approach for Mission Detection

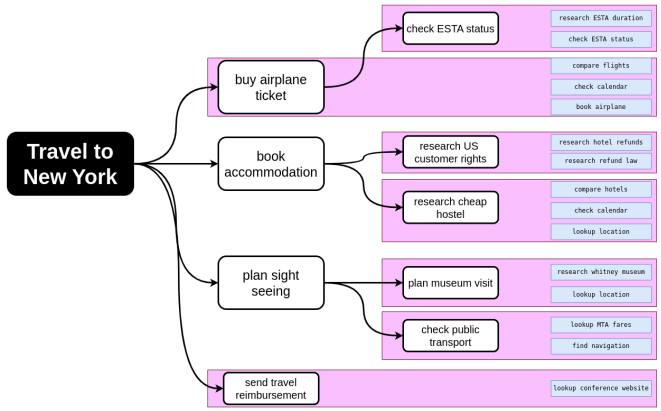
To evaluate the set of detected missions M' we calculate the weighted contribution of each mission's best jaccard index with the ground truth M. [2]

$$E(M', M) = \sum_{\mu' \in M'} \left(\frac{|\mu'|}{|\Lambda|} \cdot \max_{\mu \in M} J(\mu', \mu) \right)$$

All tree nodes



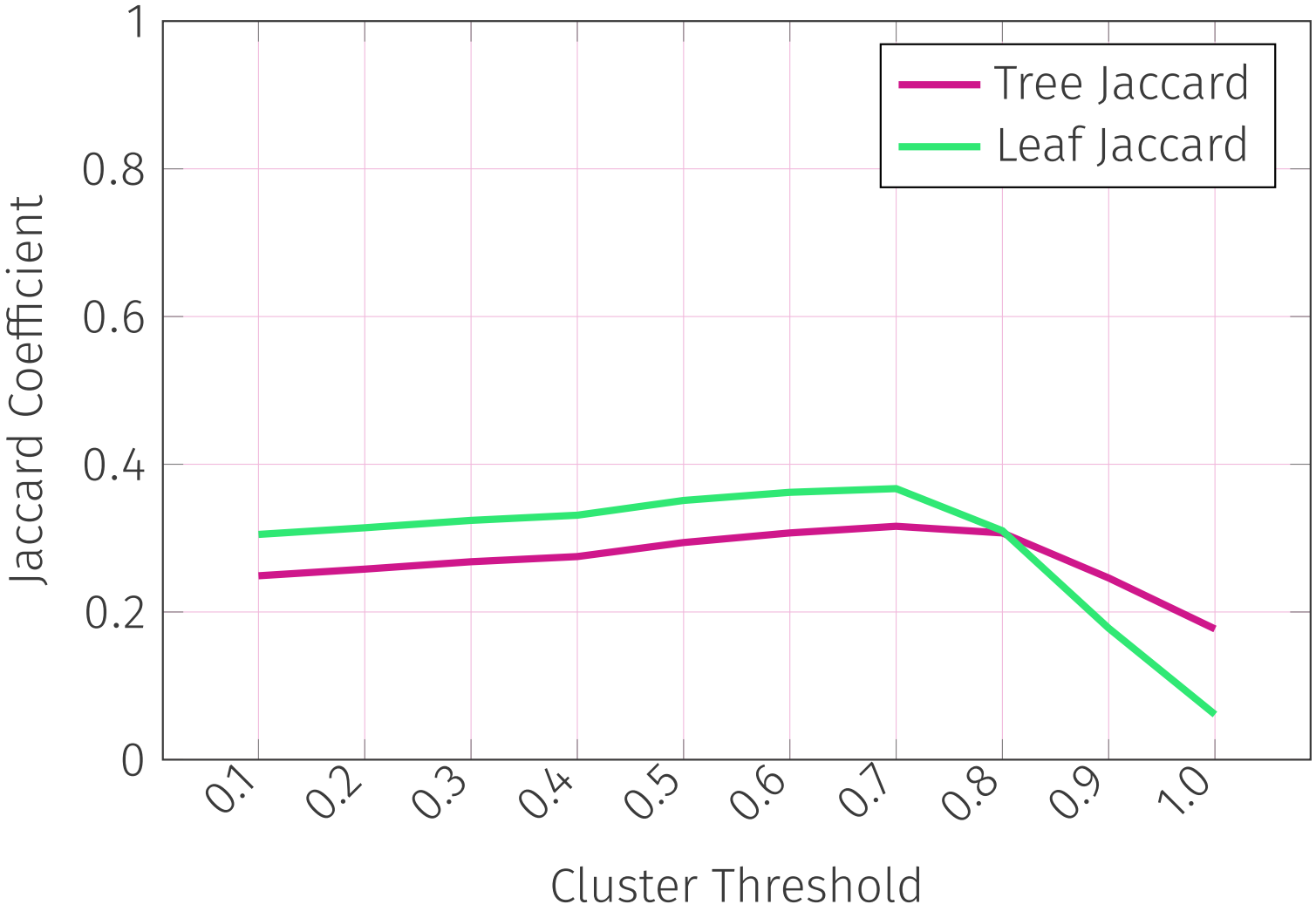
Leaf nodes



[2] fixed typo from thesis version, simplified formula

Mission Detection Evaluation

Mission Detection Evaluation



Conclusion

Main Challenges

1. Record!

Record a personal web archive over a time span of one month as a dataset for the detection.

2. Annotate!

Annotate logical sessions and (hierarchical) missions as a ground truth for evaluation.

3. Detect!

Construct and run algorithms to segment the visit log in logical sessions and missions.

4. Evaluate!

Define evaluation measures to compare the results with the ground truth.

How can the detection be improved?

- **Dataset**: connect additional device e.g. smartphone and tablet to the WARC proxy
- **Dataset**: add other information sources to the dataset like the personal calendar
- **Algorithm:** combine time and content in a complex feature, e.g. geometric method [1], [2]
- **Algorithm**: make use of more advanced clustering methods to separate only sparsely connected components [4]

Personal Calendar

During annotation I often referred to my personal calendar to identify ambiguous cases. *The dataset is already there.*

Dissimilarity Function

Visits that often appear in different missions should be weighted less in the dissimilarity function.

How can the detection be improved?

- **Dataset**: connect additional device e.g. smartphone and tablet to the WARC proxy
- **Dataset**: add other information sources to the dataset like the personal calendar
- **Algorithm:** combine time and content in a complex feature, e.g. geometric method [1], [2]
- **Algorithm**: make use of more advanced clustering methods to separate only sparsely connected components [4]

Personal Calendar

During annotation I often referred to my personal calendar to identify ambiguous cases. *The dataset is already there.*

Dissimilarity Function

Visits that often appear in different missions should be weighted less in the dissimilarity function.

How can the detection be improved?

- **Dataset**: connect additional device e.g. smartphone and tablet to the WARC proxy
- **Dataset**: add other information sources to the dataset like the personal calendar
- **Algorithm:** combine time and content in a complex feature, e.g. geometric method [1], [2]
- **Algorithm**: make use of more advanced clustering methods to separate only sparsely connected components [4]

Personal Calendar

During annotation I often referred to my personal calendar to identify ambiguous cases. *The dataset is already there.*

Dissimilarity Function

Visits that often appear in different missions should be weighted less in the dissimilarity function.

Your Questions

Image Sources

- Slide 2, Left: ”Zettelkasten Oerlinghausen” by Niklas Luhmann Archiv
- Slide 2, Right: ”Dymaxion Chronofile” by Sam Green, (Atlas Obscura)

References

[1] D. Gayo-Avello. A survey on session detection methods in query logs and a proposal for future evaluation. *Information Sciences*, 179, 5 2009.

[2] M. Hagen, B. Stein, and T. Rüb. Query session detection as a cascade. *Proceedings of the 20th ACM international conference on Information and knowledge management*, 10 2011.

References II

[3] J. Kiesel, A. de Vries, M. Hagen, B. Stein, and M. Potthast. WASP: Web Archiving and Search Personalized. In O. Alonso and G. Silvello, editors, *1st International Conference on Design of Experimental Search & Information Retrieval Systems (DESIRES 2018)*, volume 2167 of *CEUR Workshop Proceedings*, pages 16–21, Aug. 2018.

[4] C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei. Identifying task-based sessions in search engine query logs. *Proceedings of the fourth ACM international conference on Web search and data mining*, 2 2011.

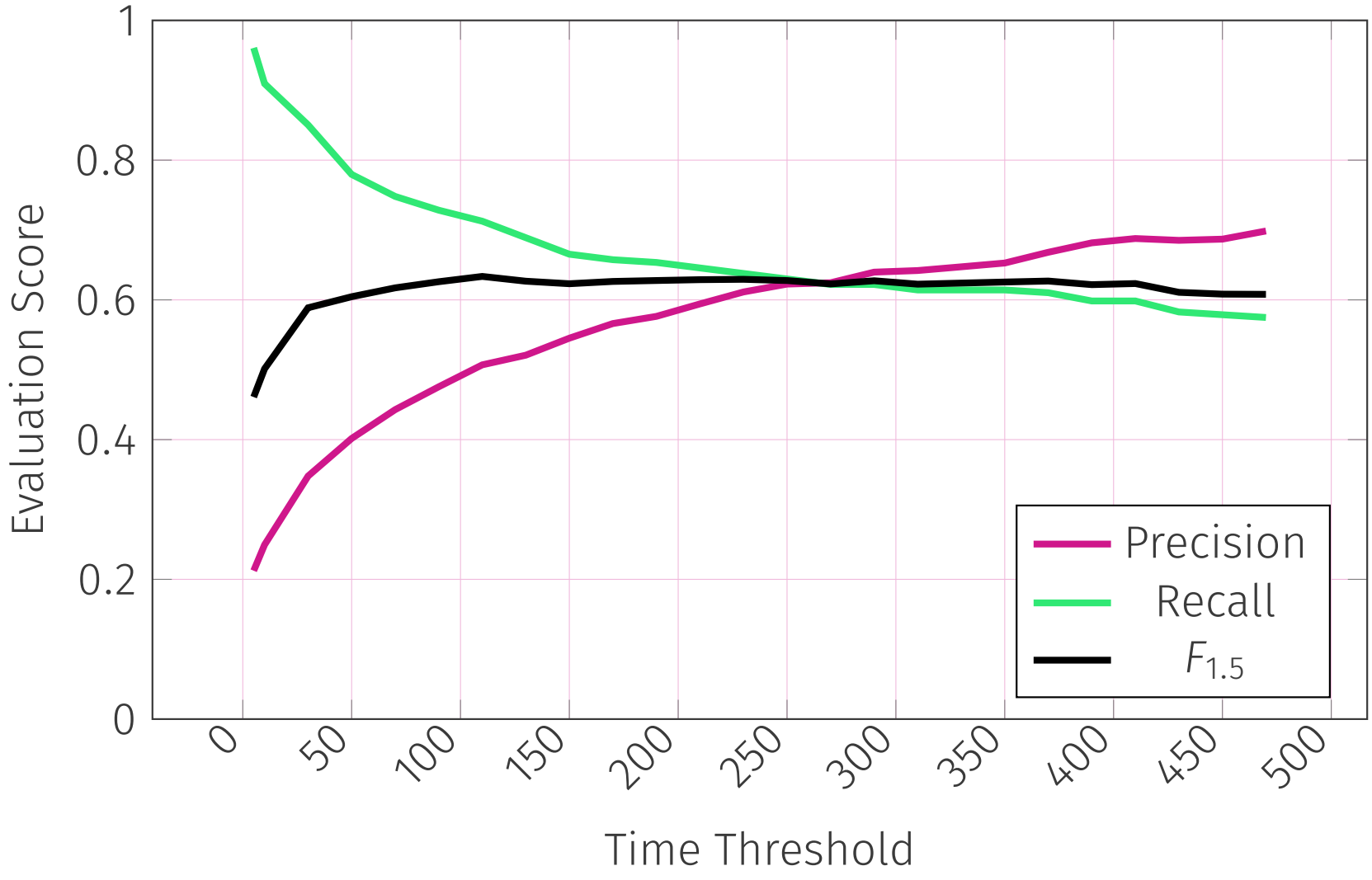
[5] D. Montoya, T. P. Tanon, S. Abiteboul, and F. M. Suchanek. Thymeflow, a personal knowledge base with spatio-temporal data. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, 10 2016.

Backup Slide: Session Detection Evaluation

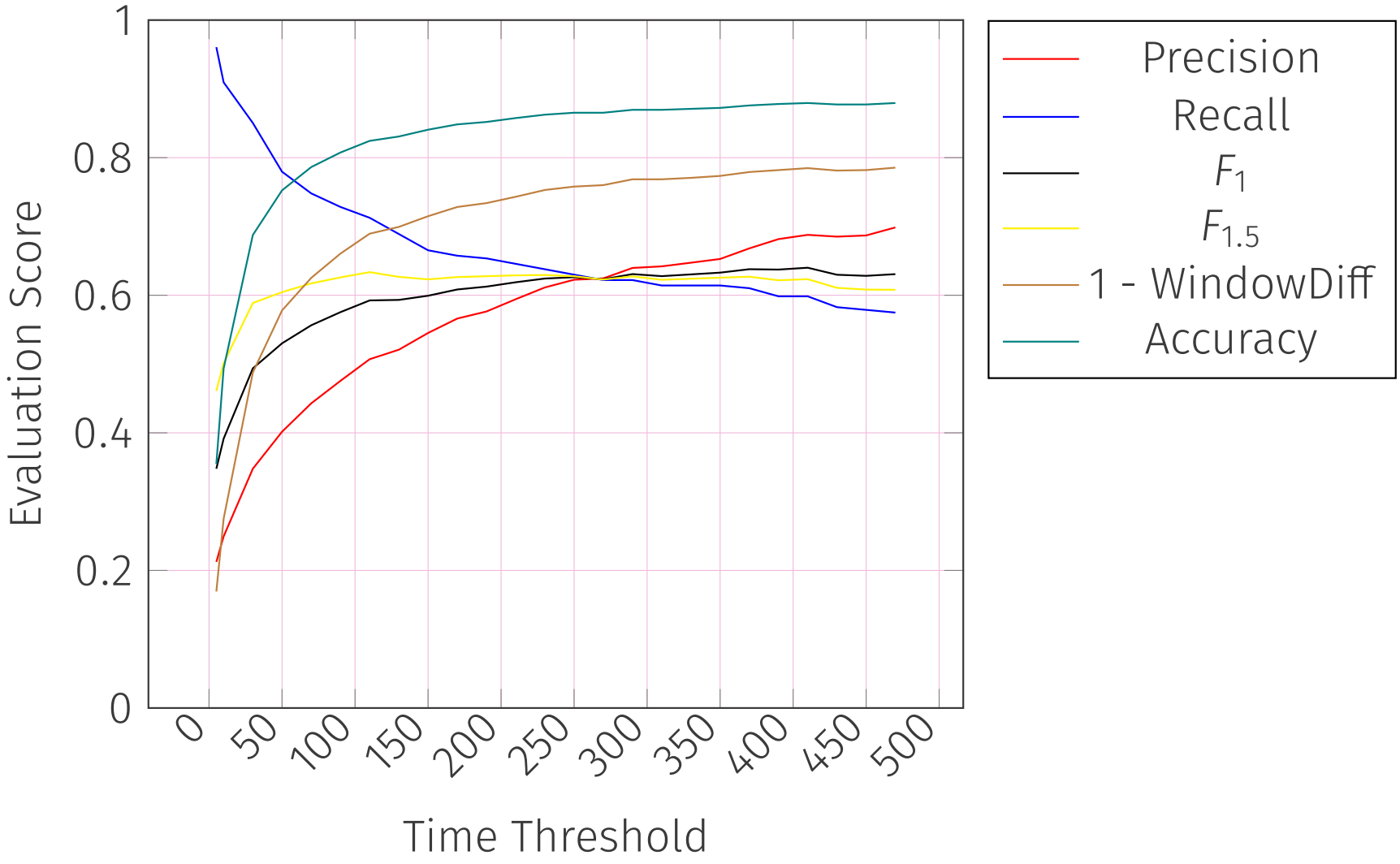
: Computed evaluation measures for each feature. Best score per column is coloured in green. Worst score per column is coloured in red.

Feature	Precision	Recall	F_1	$F_{1.5}$	Inverse Win-dowdiff	Accuracy
Time	0.507	0.717	0.594	0.636	0.689	0.824
Domain	0.28	0.925	0.43	0.541	0.407	0.561
URL Keywords (Intersection)	0.281	0.972	0.436	0.554	0.375	0.55
URL Keywords (Jaccard)	0.215	1.0	0.354	0.471	0.179	0.347
Title Keywords (Intersection)	0.179	1.0	0.304	0.415	0.053	0.179
Title Keywords (Jaccard)	0.179	1.0	0.304	0.415	0.053	0.179
Joined Similarity (Intersection)	0.281	0.972	0.436	0.554	0.375	0.55
Joined Similarity (Jaccard)	0.215	1.0	0.354	0.471	0.179	0.347
Linkage	0.22	0.992	0.361	0.477	0.186	0.37

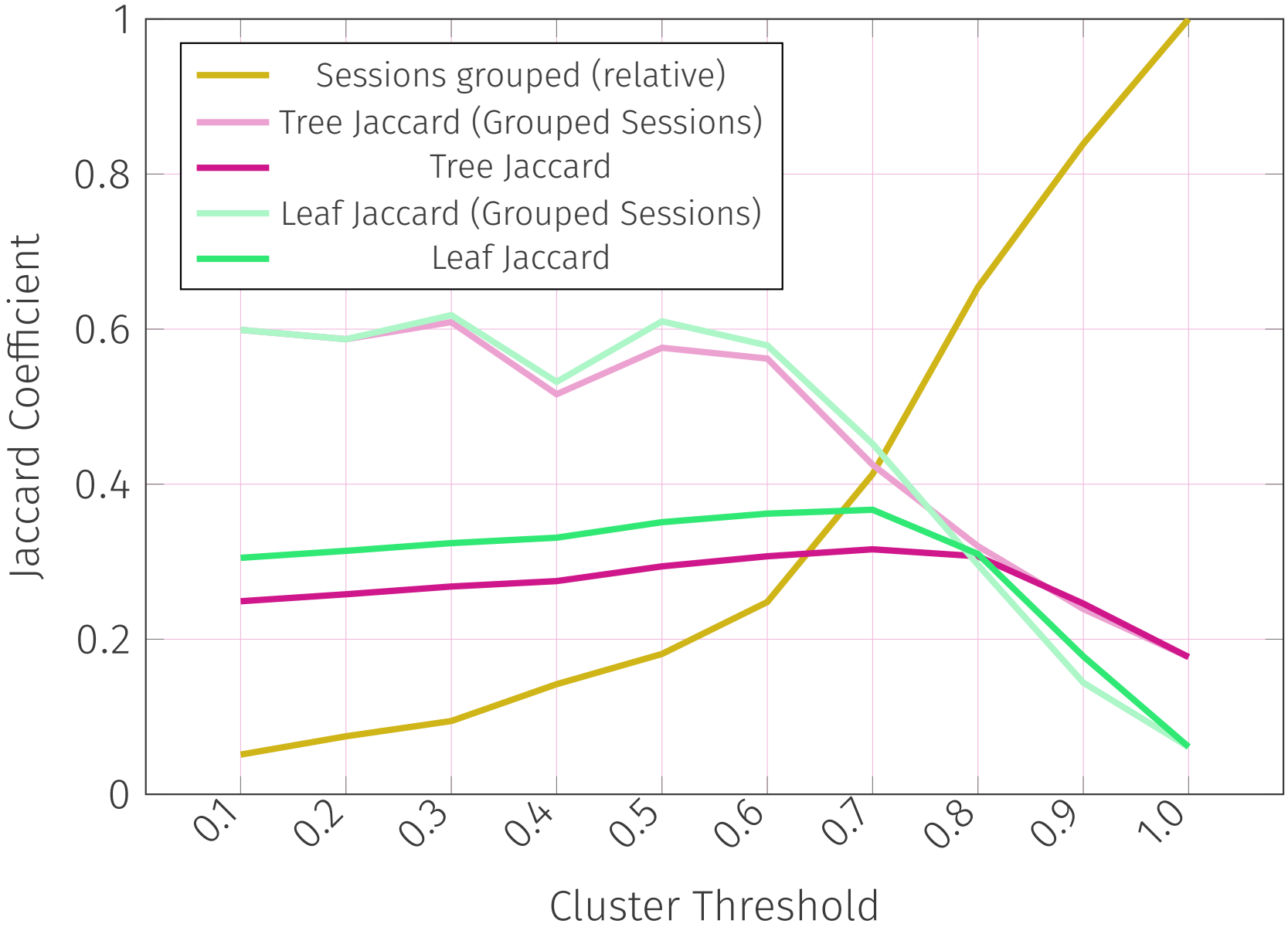
Backup Slide: Session Detection Evaluation



Backup Slide: Session Detection Evaluation



Backup Slide: Mission Detection Evaluation



Backup Slide: Precision, Recall and Accuracy

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$accuracy = \frac{TP + TN}{P + N}$$

$$F_{1.5} = (1+1.5^2) \cdot \frac{precision \cdot recall}{(1.5^2 \cdot precision) + recall}$$

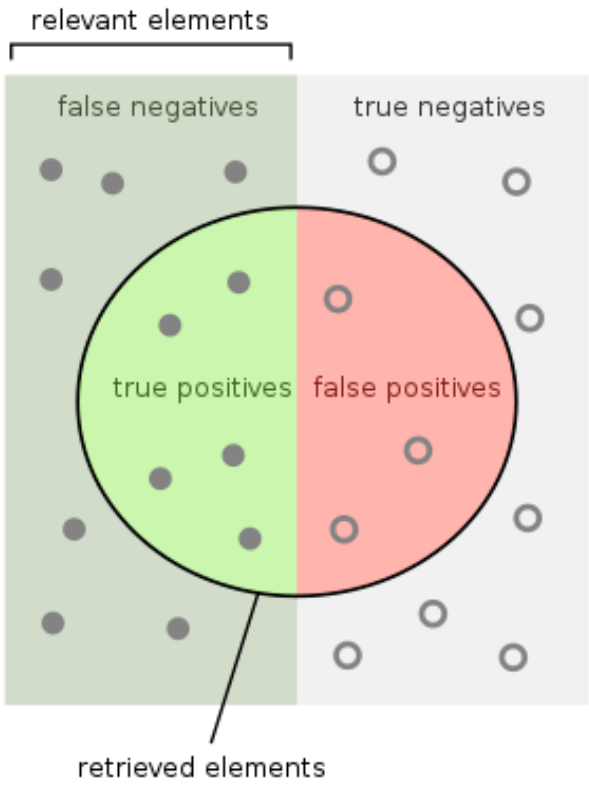
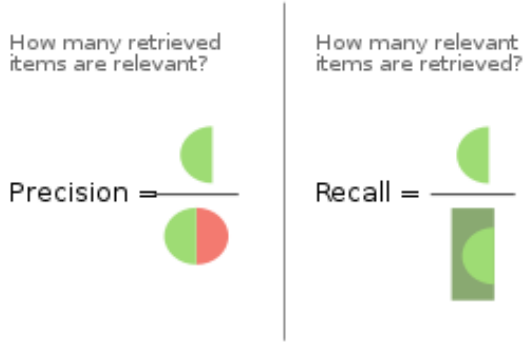


Image Source: Visual Explanation for Precision and Recall. By Walber - Own work, CC BY-SA 4.0



Backup Slide: Limitations of Precision and Recall

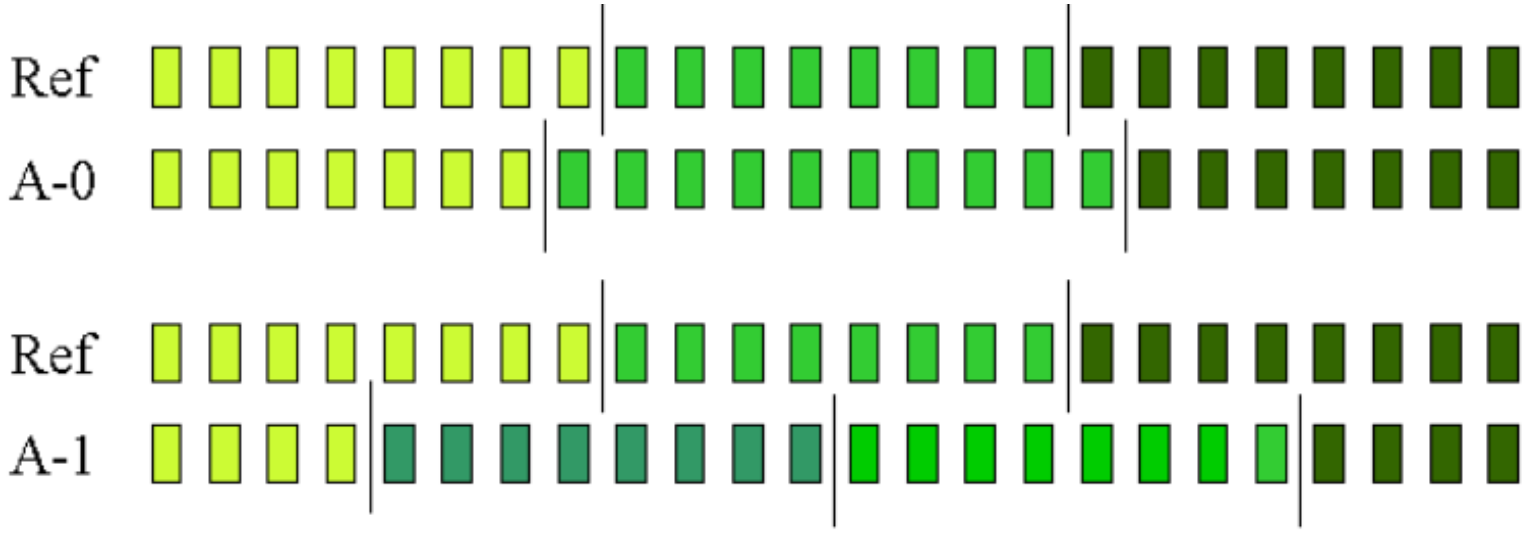


Figure 8: Ground truth and outputs of two segmentation algorithms with equal precision and recall.

Image Source: Assembly AI, use cases for topic segmentation