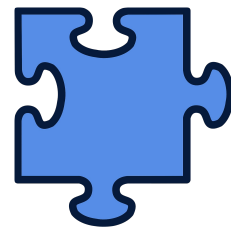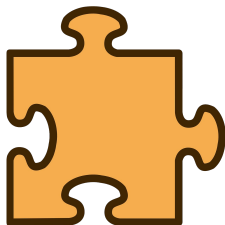# Completing The Multi-Authorship Jigsaw Puzzle

PAN'21 Invited Presentation
Harry Scells - h.scells@uq.edu.au

Who am I?

- Collaborating with group at Weimar University (https://webis.de/)

  - Investigating scientific multi-authorship, PHOENIX project

- Based at The University of Queensland, Australia (https://ielab.io/)

  - Information Retrieval - systematic reviews

Multi-Authorship at PAN

- PAN has lead the way for **multi-authorship analysis**
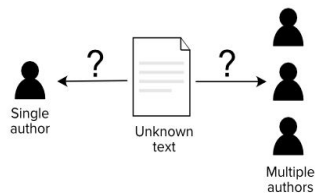
- PAN'16-PAN'21

  - Multi-Author Identification
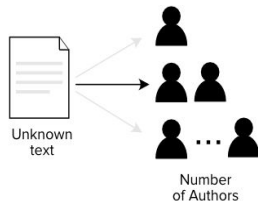
- PAN'21

  - Multi-Author Profiling

2016	2017	2018	2019	2020	2021
Author Diarisation / Style Change Detection					
		Profiling Hate Speech Spreaders on Twitter			

Background: Multi-Author Analysis at PAN

Multi-Authorship Identification

Single author ← ? ─ Unknown text ─ ? → Multiple authors

Multi-Author Detection

Unknown text → Number of Authors

Author Count Prediction

Multi-authored text → Style Change Detection

Unknown text → Known texts/authors

Multi-Author Attribution

Multi-Authorship Profiling

Unknown text → Communities

Multi-Authorship Obfuscation

Multi-Authorship - Completing the Jigsaw

Multi-Authorship Identification
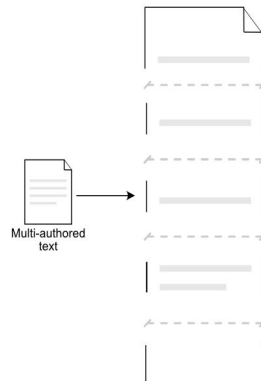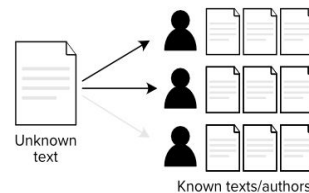
Multi-Authorship Profiling

Multi-Authorship Obfuscation

Tasks

Datasets

Evaluation

Multi-Authorship - Completing the Jigsaw

Tasks

Datasets

Evaluation
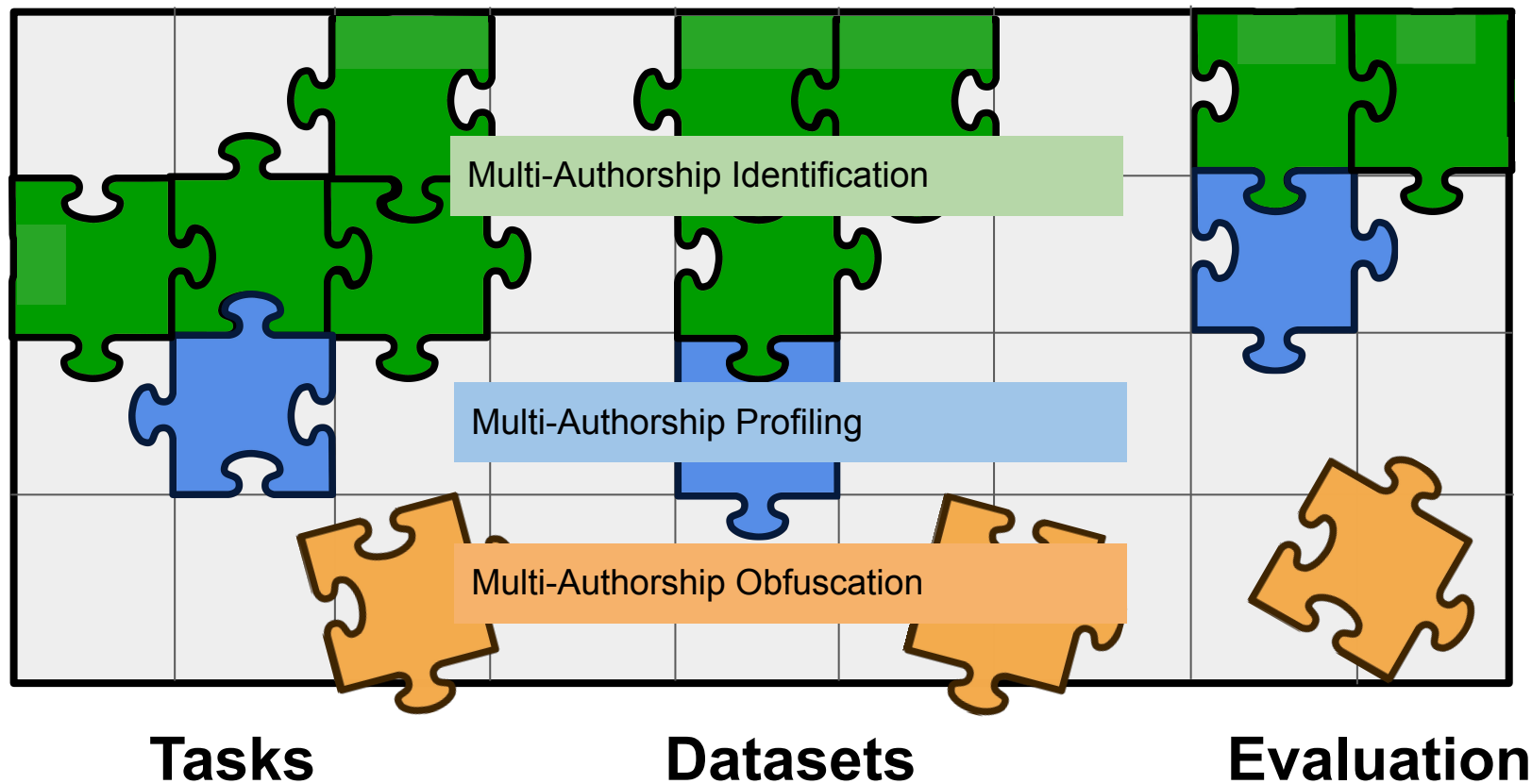
Tasks are ordered in increasing difficulty

Multi-Authorship Identification

Single author ← ? ← Unknown text → ? → Multiple authors

Multi-Author Detection

Unknown text → Number of Authors

Author Count Prediction

Multi-authored text → Style Change Detection

Unknown text → Known texts/authors

Multi-Author Attribution

Multi-Authorship Profiling

Unknown text → Communities

Multi-Authorship Obfuscation
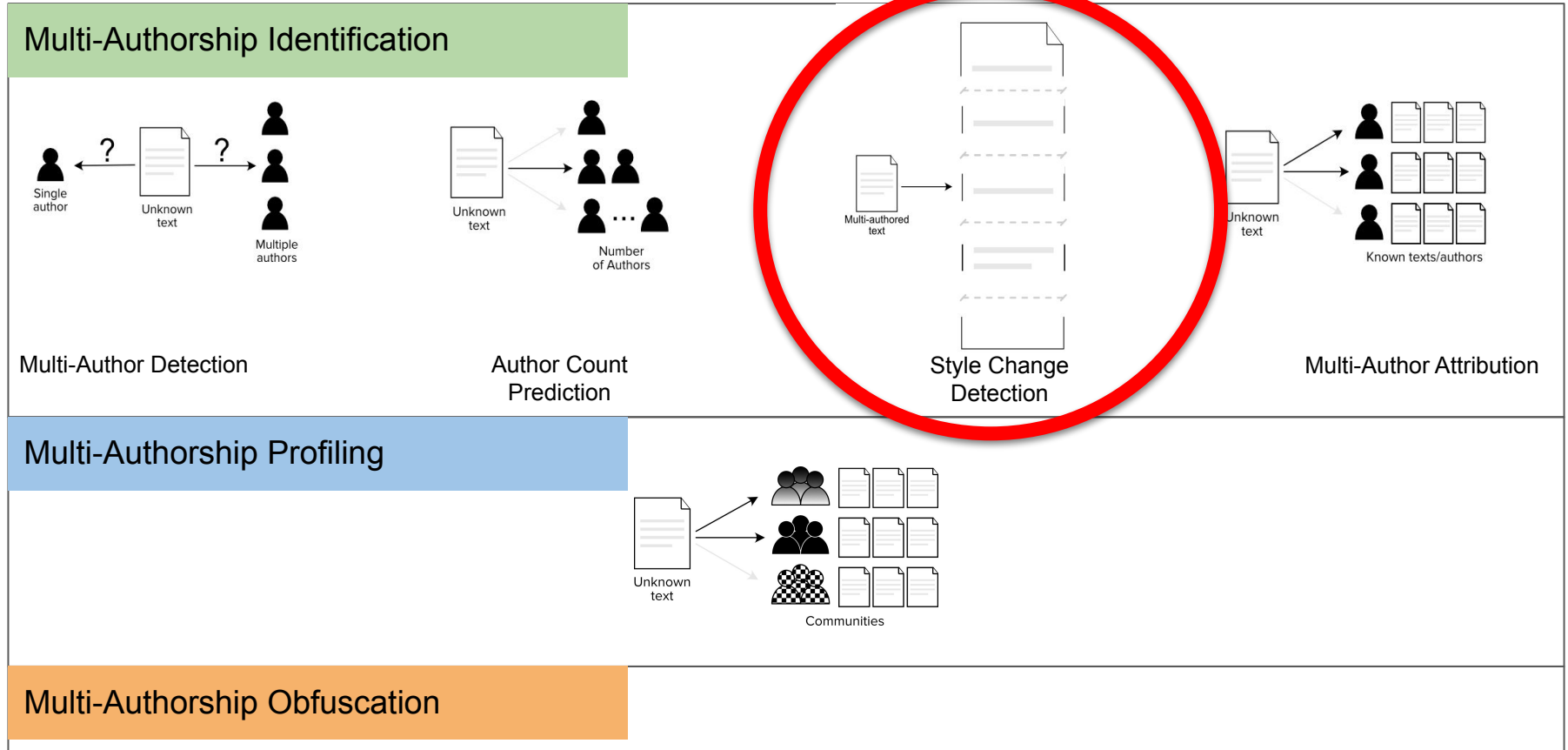
Lesson learnt: start simple

Multi-Authorship Identification

Single author ← ? — Unknown text — ? → Multiple authors

Unknown text → Number of Authors

Multi-authored text →

Unknown text → Known texts/authors

Multi-Author Detection

Author Count Prediction

Style Change Detection
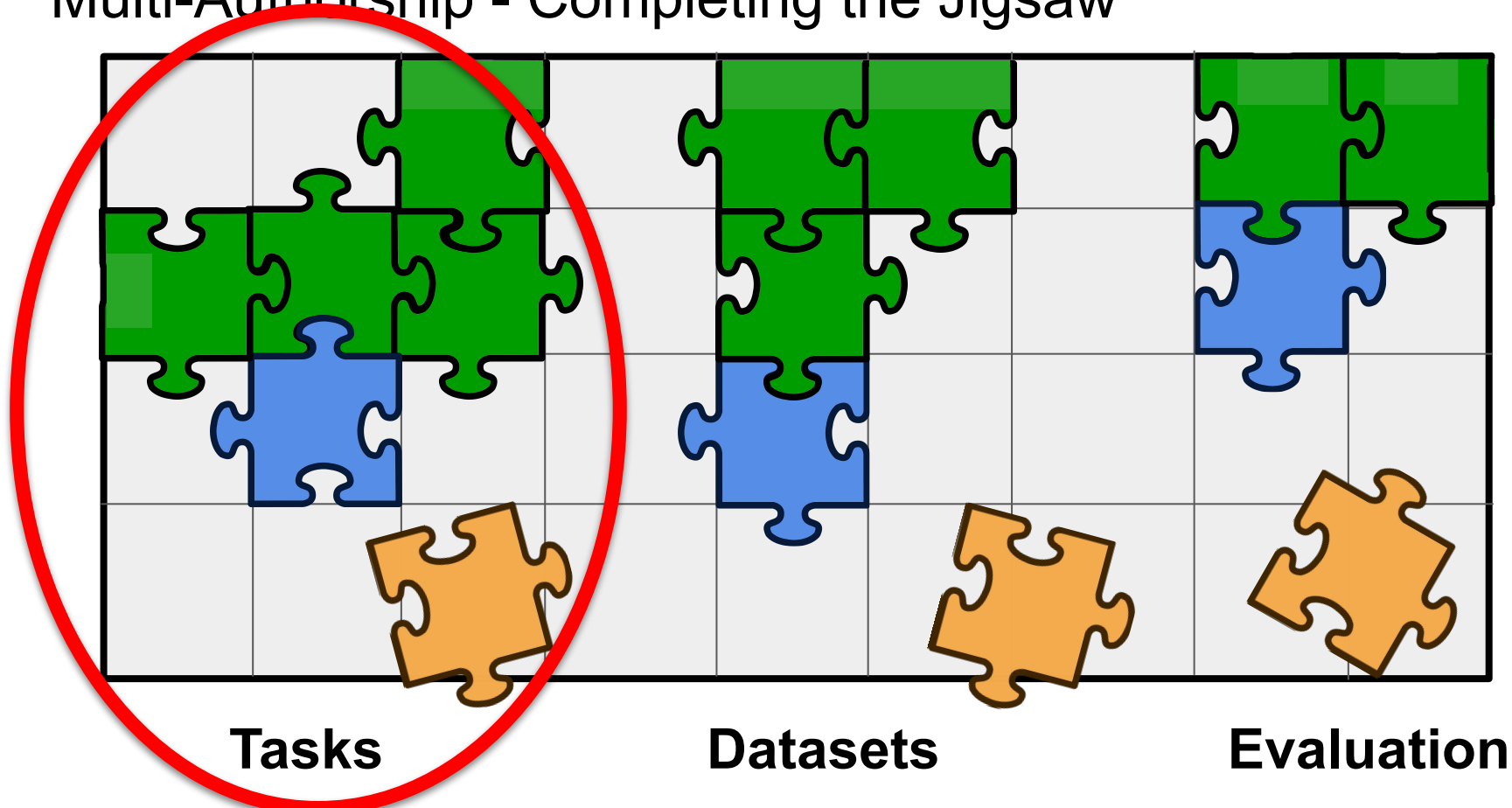
Multi-Author Attribution

Multi-Authorship Profiling

Unknown text → Communities

Multi-Authorship Obfuscation

Reflecting before next challenges
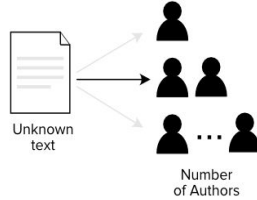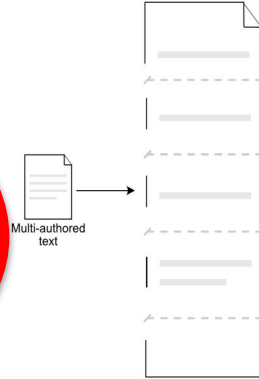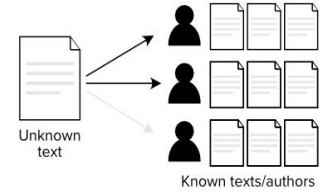
Multi-Authorship Identification

Multi-Author Detection

Author Count Prediction

Style Change Detection

Multi-Author Attribution

Single author
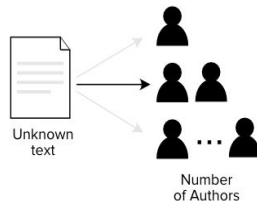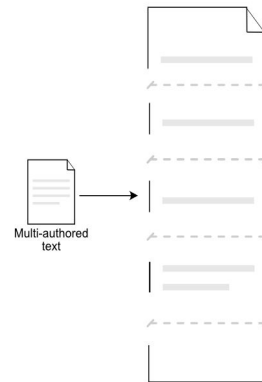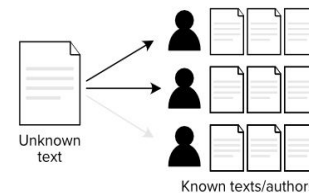
Unknown text

Multiple authors

Unknown text

Number of Authors

Multi-authored text

Unknown text

Known texts/authors

Multi-Authorship Profiling

Unknown text

Communities

Multi-Authorship Obfuscation

Datasets and Evaluation

Tasks Datasets Evaluation
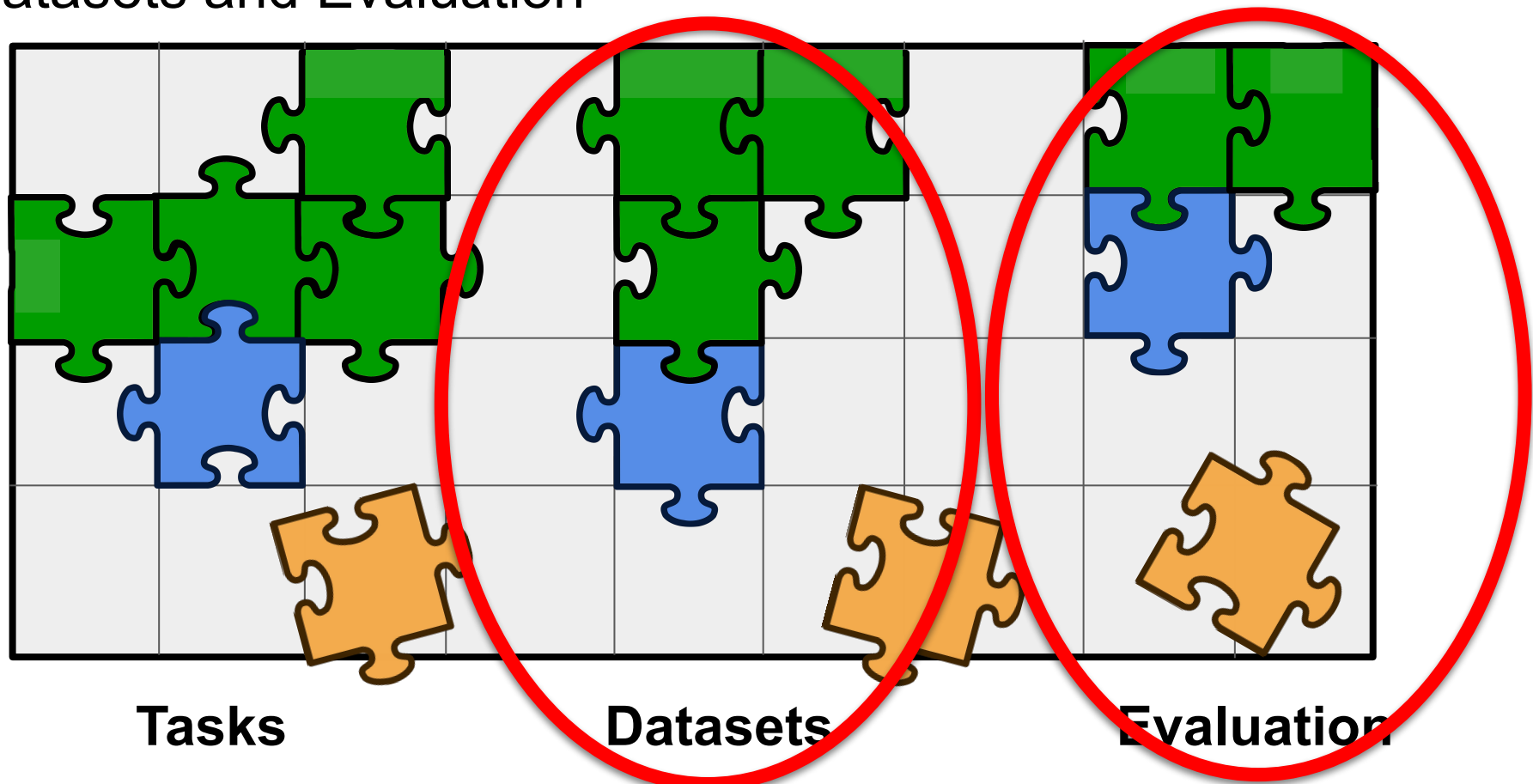
Speculation: Connecting the Larger Parts of the Jigsaw

Tasks   **Datasets**   **Evaluation**

Speculation: Datasets

Tasks   **Datasets**   **Evaluation**
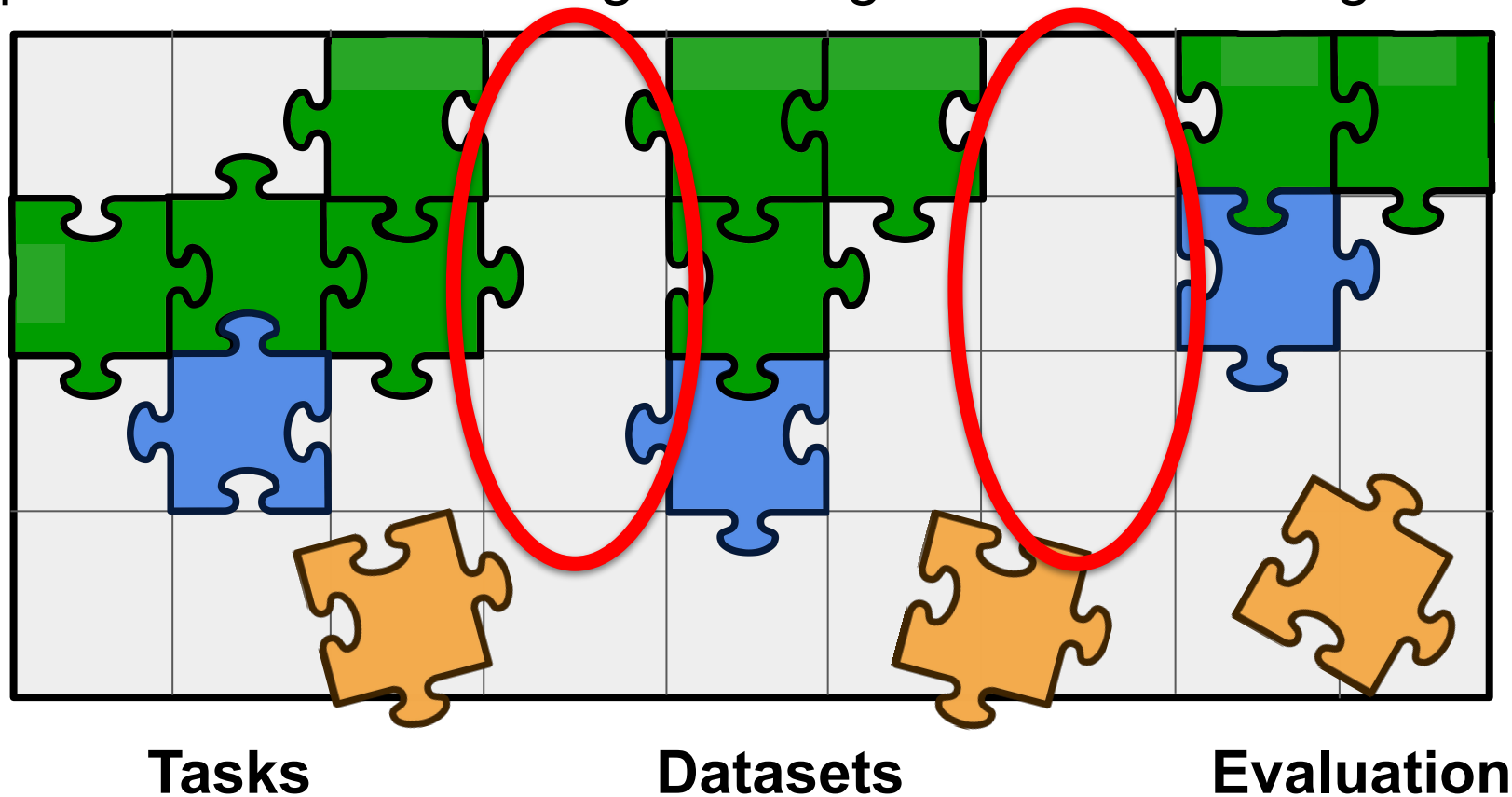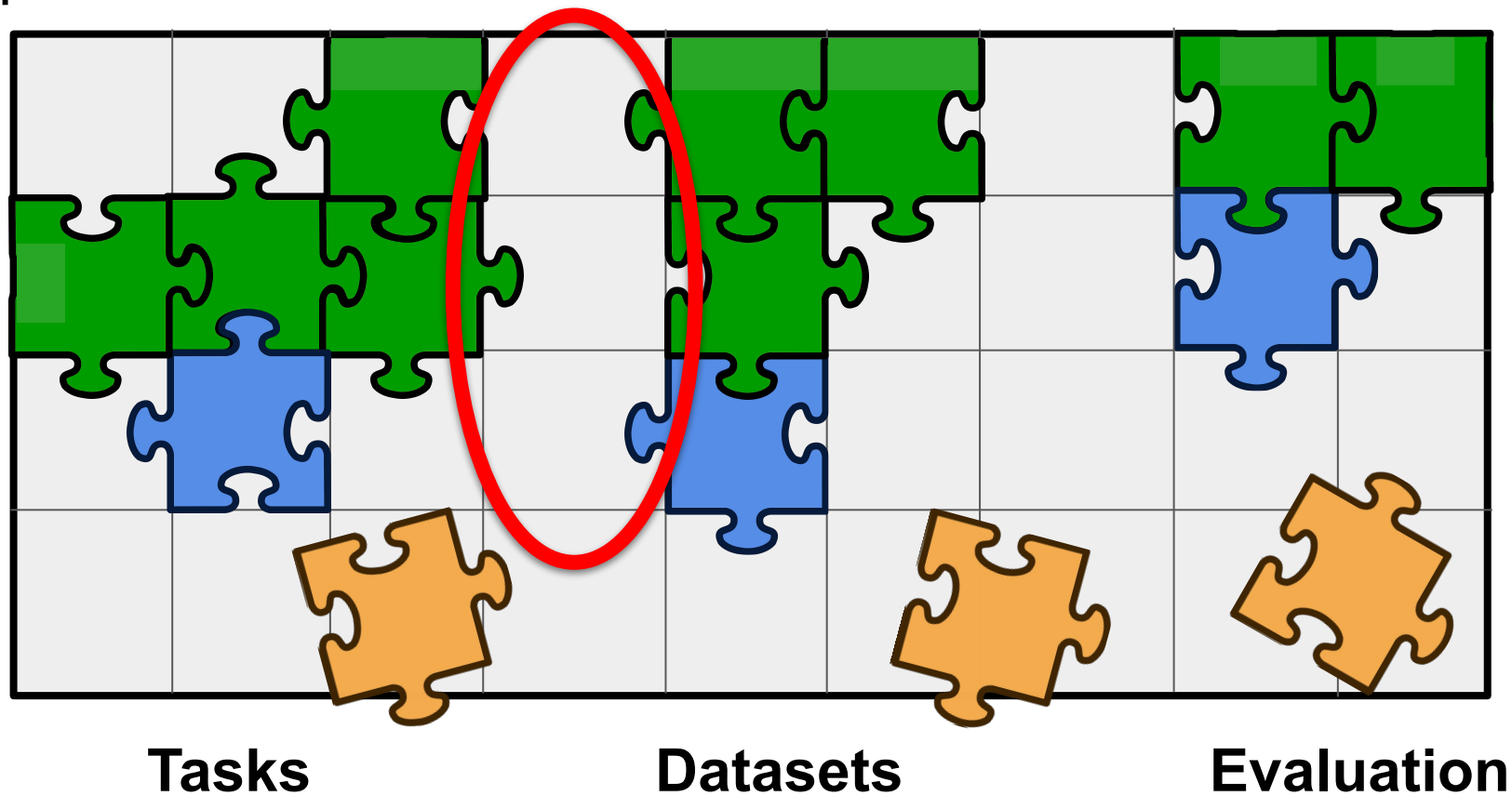
Multi-Authorship Datasets

- Created by **combining** single-author texts into larger synthetic, multi-authored text

- *Horizontal Division* [1]

- **Why** create datasets in this way?

[1] Lowry, P.B. et al. 2004. Building a taxonomy and nomenclature of collaborative writing to improve interdisciplinary research and practice. *The Journal of Business Communication (1973)*. 41, 1 (2004), 66–99.

Speculation: Datasets and Collaborative Writing Style
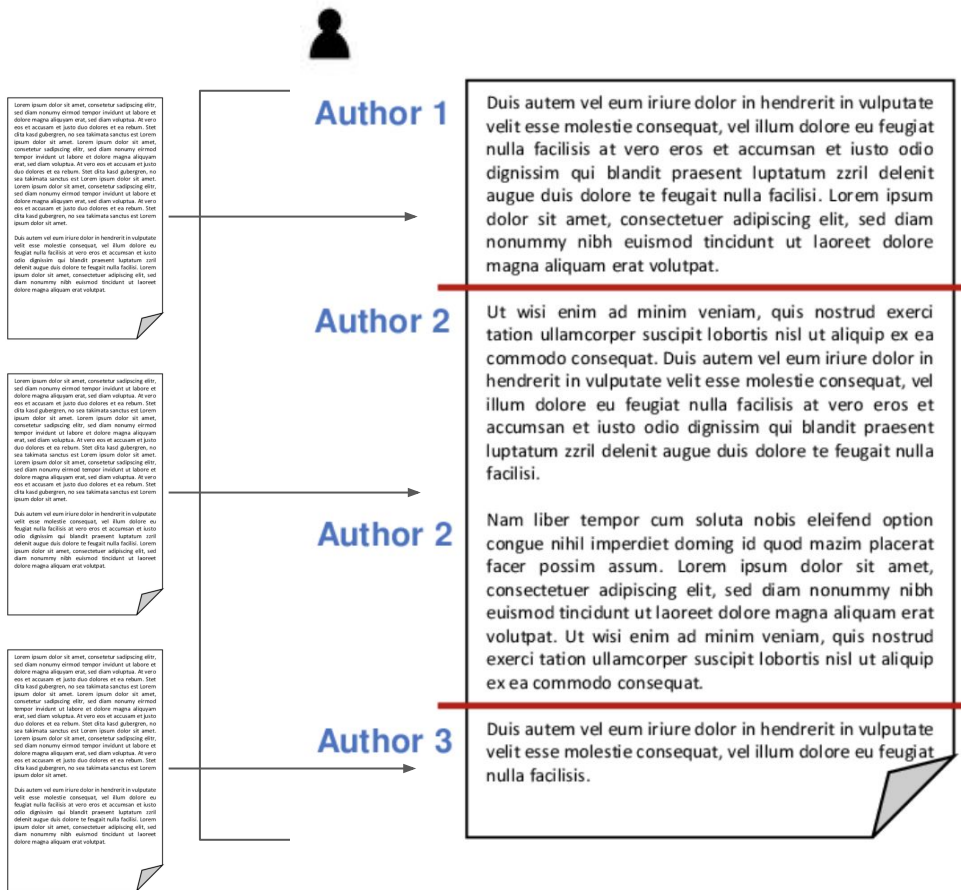
- Many ways to **write collaboratively**

- Are models **robust** to different collaborative writing styles?

- Let's look at some **intuitions** for why this is worth investigating

Collaborative Writing Styles

- Group-Single
- Sequential
- *Horizontal Division*
- Stratified Division
- Reactive

Intuition: Sequential Writing Style

- Each author writes at a given time

- Independently, until text is complete

- Clearly demarcated author styles

- For SCD, **boundaries are large**

Intuition: Group-Single Writing Style

- Many authors come up with ideas

- Single author compiles them

- Consistent authorship style

- Yet, multiple authors were involved

- May not be able to perform SCD: **are there style boundaries?**
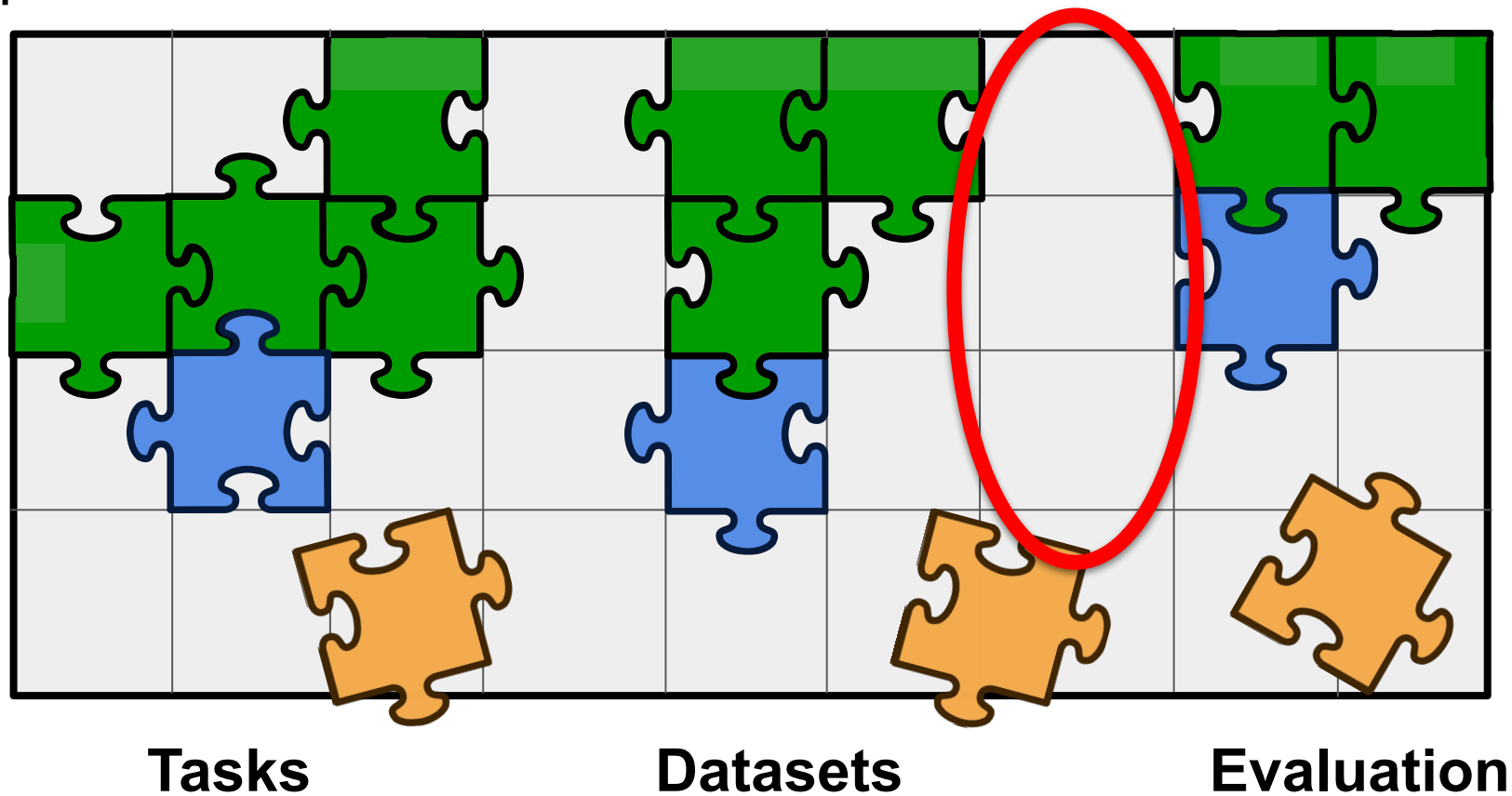
- Might be able to perform other tasks
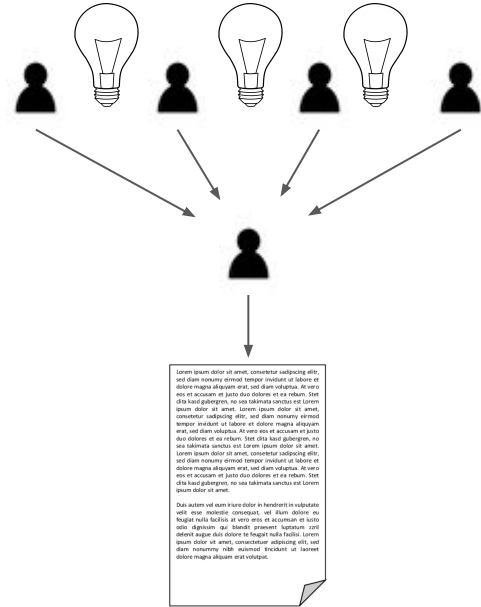
Intuition: Reactive Writing Style

- Many authors all write sync. in real-time

- Adjusting writing of others

- Blurred boundaries of author style

- SCG may be difficult: **boundaries change often and may be very small**

Speculation: Evaluation
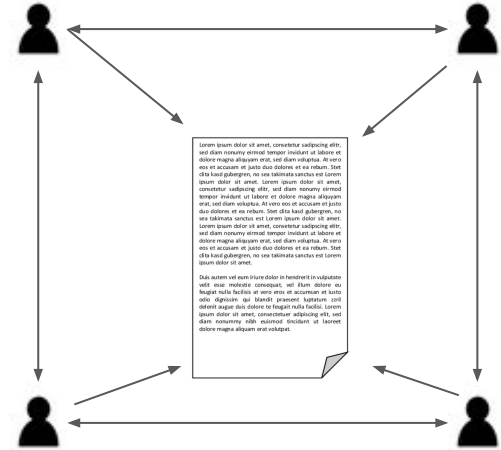
Tasks   **Datasets**   **Evaluation**

Speculation: Evaluation

- 2016: micro/macro F1, BCubed F-Score
  - Character-level SCD
- 2017: WindowDiff, WinPR
  - Character-level SCD
- 2018: Accuracy
  - Multi-Author Detection
- 2019: Accuracy, Ordinal Classification Index
  - Multi-Author Detection, Multi-Author Count Prediction
- 2020: micro-F1
  - Multi-Author Detection, Pre-segmented text SCD
- 2021: micro-F1
  - Multi-Author Detection, Multi-Author Count Prediction, Pre-segmented text SCD

- What is the **most appropriate** evaluation measure?

- Weighting measures depending on **contribution amount**?

- Weighted based on **number of authors**?

- Different measures for different collaborative writing styles?

Why are we working on these problems?
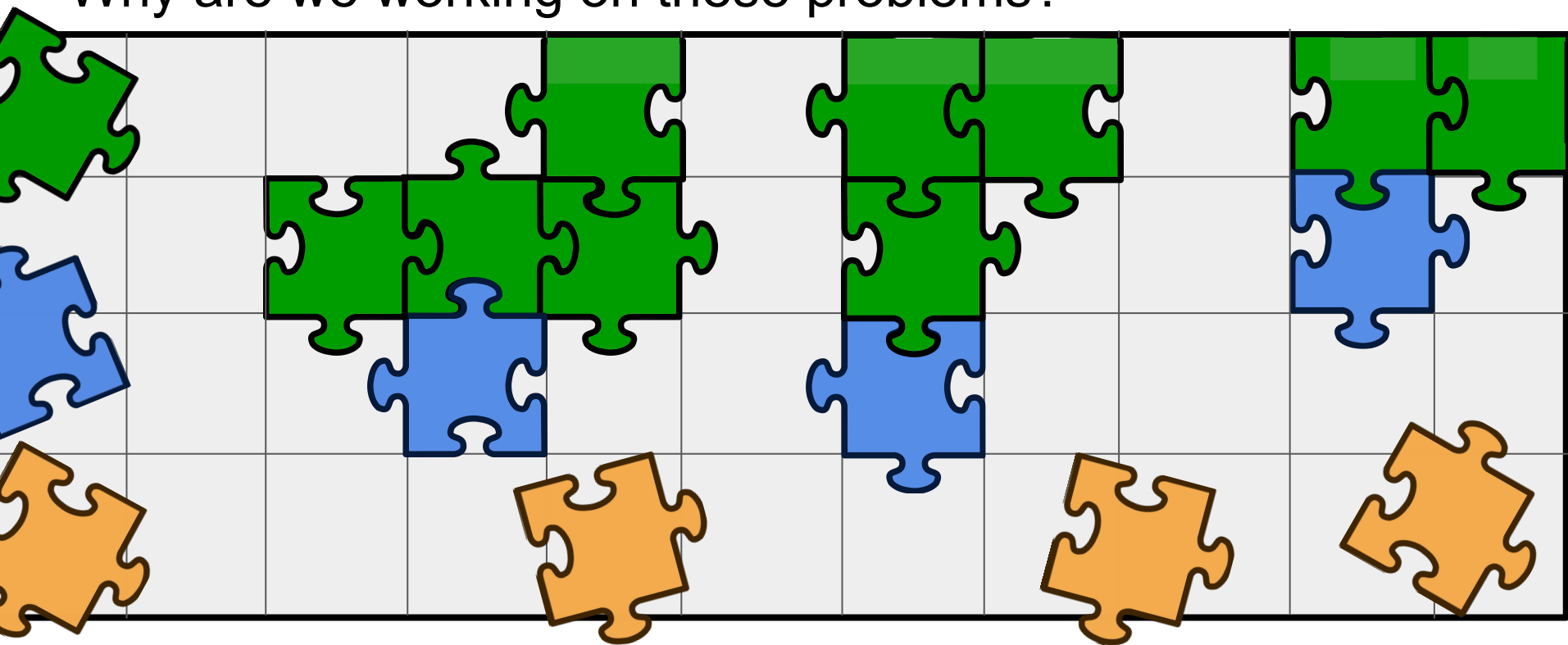
Applications   **Tasks**   **Datasets**   **Evaluation**

[1] Potthast, M. et al. 2016. Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. (Mar. 2016), 393–407.

Summary

- Large parts of the multi-authorship jigsaw are in place

  - Datasets that involve different **collaborative writing styles**

  - Evaluation that **reflects data and tasks**

- **Applications** are important, need to be aware of **why we are working on these problems**

NEW

- **Authorship Toolkit**
  - Common authorship baselines
  - Stylometric features
  - Formats for datasets/experiments
  - Common evaluation measures

People

Harry Scells, Erik Körner, Martin Potthast, Benno Stein