Slides Credit: Matthias Boehm & Shafaq Siddiqi

Data Integration and Large Scale Analysis

05 Entity Linking and Deduplication

HALLOWEEN SPECIAL

Agenda

- Motivation and Terminology
- Entity Resolution Concepts
- Entity Resolution Tools
- Example Applications

Motivation and Terminology

Recap: Corrupted/Inconsistent Data

- **#1 Heterogeneity of Data Sources**
  - Update anomalies on denormalized data / eventual consistency
  - Changes of app/prep over time (US vs us) → inconsistencies
- **#2 Human Error**
  - Errors in semi-manual data collection, laziness (see default values), bias
  - Errors in data labeling (especially if large-scale: crowd workers / users)
- **#3 Measurement/Processing Errors**
  - Unreliable HW/SW and measurement equipment (e.g., batteries)
  - Harsh environments (temperature, movement) → aging

HALLOWEEN SPECIAL

Recap: Corrupted/Inconsistent Data

Uniqueness & duplicates

Contradictions & wrong values

Missing Values

Ref. Integrity

[**Credit:** Felix Naumann]

ID	Name	BDay	Age	Sex	Phone	Zip
3	Smith, Jane	05/06/1975	44	F	999-9999	98120
3	John Smith	38/12/1963	55	M	867-4511	11111
7	Jane Smith	05/06/1975	24	F	567-3211	98120

Zip	City
98120	San Jose
90001	Lost Angeles

Typos

HALLOWEEN SPECIAL

Terminology

[Douglas Burdick, Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, Wang-Chiew Tan: Expressive power of entity-linking frameworks. **J. Comput. Syst. Sci. 2019**]

- **Entity Linking**
  - "***Entity linking*** is the problem of creating links among **records** representing **real-world entities** that are **related** in certain ways."
  - "As an important special case, it includes **entity resolution**, which is the problem of **identifying or linking duplicate entities**
- **Other Terminology**
  - **Entity Linking** → Entity Linkage, Record Linkage
  - **Entity Resolution** → Data Deduplication, Entity Matching

☺

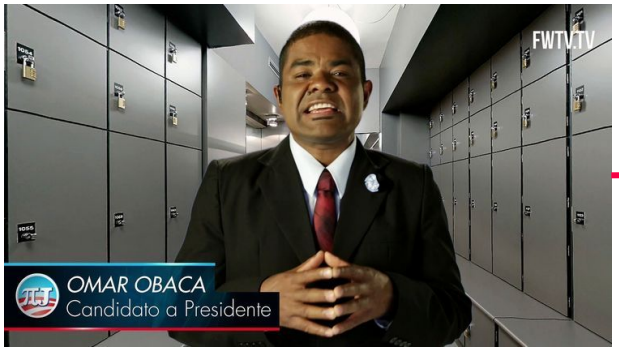
Barack Obama
Barack **Hussein** Obama **II**
Omar Obaca
The US president **(**2016**)**
Barack and Michelle
are married ….

HALLOWEEN SPECIAL

Terminology

[Douglas Burdick, Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, Wang-Chiew Tan: Expressive power of entity-linking frameworks. **J. Comput. Syst. Sci. 2019**]

- **Entity Linking**
  - "***Entity linking*** is the problem of creating links among **records** representing **real-world entities** that are **related** in certain ways."
  - "As an important special case, it includes **entity resolution**, which is the problem of **identifying or linking duplicate entities**
- **Other Terminology**
  - **Entity Linking** → Entity Linkage, Record Linkage
  - **Entity Resolution** → Data Deduplication, Entity Matching



Barack Obama
Barack Hussein Obama II
Omar Obaca
The US president (2016)
Barack and Michelle are married ….

HCC
Institute of HUMAN-CENTRED COMPUTING

HALLOWEEN SPECIAL

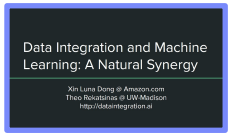
Terminology

[Douglas Burdick, Ronald Fagin, Phokion G. Kolaitis, Lucian Popa, Wang-Chiew Tan: Expressive power of entity-linking frameworks. **J. Comput. Syst. Sci. 2019**]

- **Applications**
  - **Named entity recognition and disambiguation**
  - Archiving
  - Recommenders / social networks
  - Financial institutions (persons and legal entities)
  - Travel agencies, transportation, health care

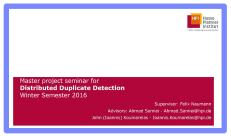
HALLOWEEN SPECIAL

Entity Resolution Concepts



[Xin Luna Dong, Theodoros Rekatsinas: Data Integration and Machine Learning: A Natural Synergy. Tutorials, **SIGMOD 2018**, **PVLDB 2018**, **KDD 2019**]

[Sairam Gurajada, Lucian Popa, Kun Qian, Prithviraj Sen: Learning-Based Methods with Human in the Loop for Entity Resolution, Tutorial, **CIKM 2019**]



[Felix Naumann, Ahmad Samiei, John Koumarelas: Master project seminar for Distributed Duplicate Detection. Seminar, **HPI WS 2016**]

HCC
Institute of HUMAN-CENTRED COMPUTING

HALLOWEEN SPECIAL

Problem Formulation

- **Entity Resolution**
  - "Recognizing those records in two files which represent identical persons, objects, or events"
  - Given two data sets A and B
  - Decide for all pairs of records a_i – b_j in A x B
    if match (**link**), no match (**non-link**), or not enough evidence (**possible-link**)

[Ivan Fellegi, Alan Sunter: A Theory for Record Linkage, J. American. Statistical Assoc., pp. 1183-1210, 1969]

- **Naïve Deduplication**
  - UNION DISTINCT via hash group-by or sort group-by
  - **Problem:** only exact matches

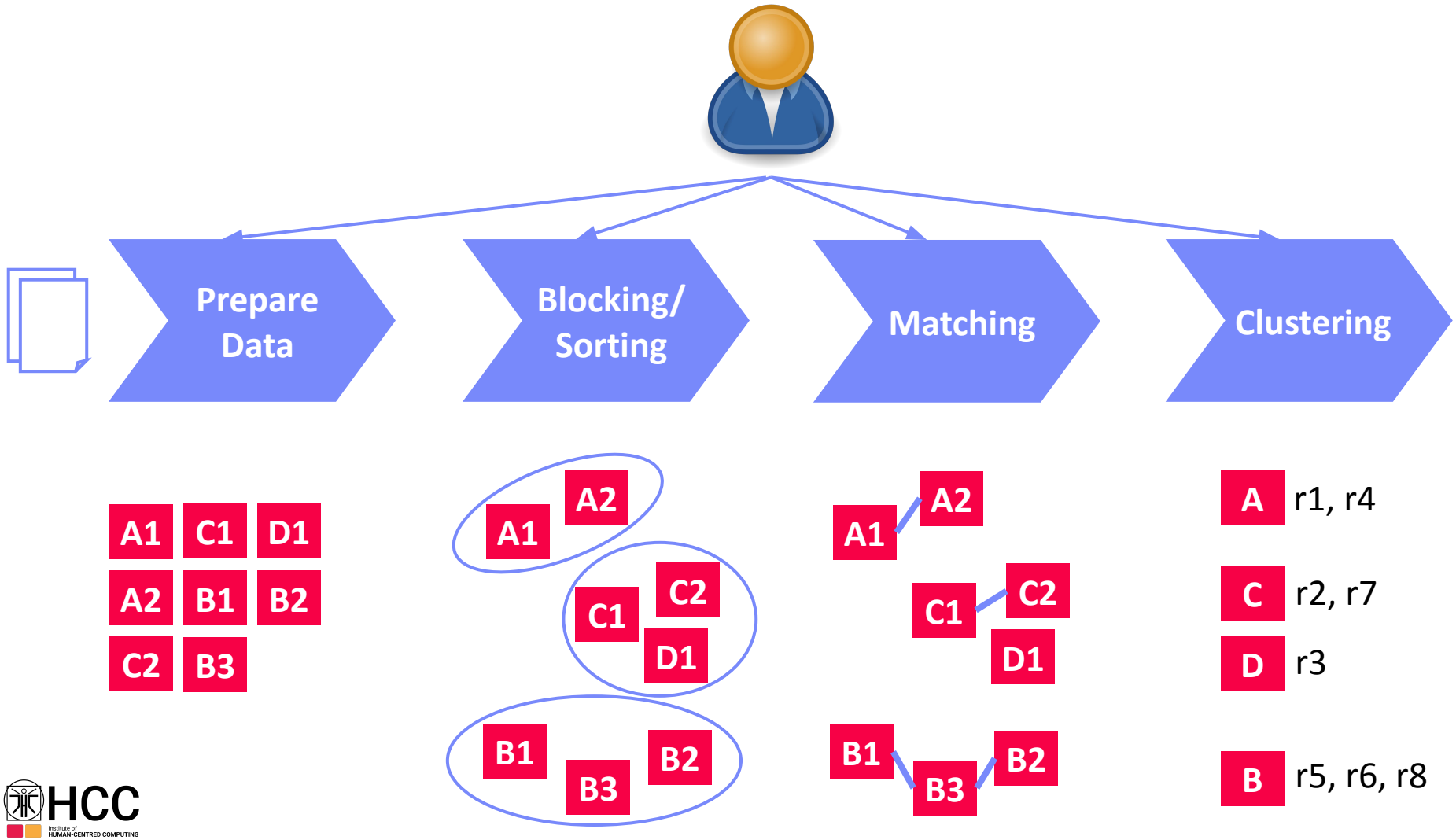
Name	Position	Affiliation	Department
Lucas Iacono	Area Manager	Know Center	DAI
Lucas Lacono	Lecturer	TU Graz	HCC

→ **Similarity Measures**

- Token-based: e.g., Jaccard J(A,B) = (A ∩ B) / (A ∪ B)
- Edit-based: e.g., Levenshtein lev(A,B) → min(replace, insert, delete)
- Phonetic similarity (e.g., soundex, metaphone), **Python lib Jellyfish**

HALLOWEEN SPECIAL

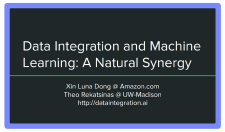
Entity Resolution Pipeline



HALLOWEEN SPECIAL

Entity Linking Approaches

[Xin Luna Dong, Theodoros Rekatsinas: Data Integration and Machine Learning: A Natural Synergy. **PVLDB 2018**]



50 Years of Entity Linkage

Rule-based and stats-based
- Blocking: e.g., same name
- Matching: e.g., avg similarity of attribute values
- Clustering: e.g., transitive closure, etc.

Supervised learning
- Random forest for matching
  F-msr: >95% w. ~1M labels
- Active learning for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

1969 (Pre-ML) — **~2000 (Early ML)** — **~2015 (ML)** — **2018 (Deep ML)**

Sup / Unsup learning
- Matching: Decision tree, SVM
  F-msr: 70%-90% w. 500 labels
- Clustering: Correlation clustering, Markov clustering

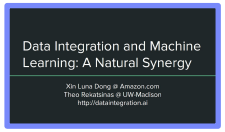
Deep learning
- Deep learning
- Entity embedding

HCC Institute of HUMAN-CENTRED COMPUTING

HALLOWEEN SPECIAL

Entity Linking Approaches

[Xin Luna Dong, Theodoros Rekatsinas: Data Integration and Machine Learning: A Natural Synergy. **PVLDB 2018**]



50 Years of Entity Linkage

Rule-based and stats-based
- Blocking: e.g., same name
- Matching: e.g., avg similarity of attribute values
- Clustering: e.g., transitive closure, etc.

1969 (Pre-ML)

~2000 (Early ML)

Sup / Unsup learning
- Matching: Decision tree, SVM
  F-msr: 70%-90% w. 500 labels
- Clustering: Correlation clustering, Markov clustering

~2015 (ML)

Supervised learning
- Random forest for matching
  F-msr: >95% w. ~1M labels
- Active learning for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

2018 (Deep ML)

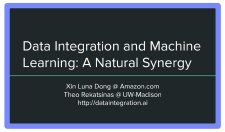
Deep learning
- Deep learning
- Entity embedding

HCC Institute of HUMAN-CENTRED COMPUTING

HALLOWEEN SPECIAL

Entity Linking Approaches

[Xin Luna Dong, Theodoros Rekatsinas: Data Integration and Machine Learning: A Natural Synergy. **PVLDB 2018**]



50 Years of Entity Linkage

Rule-based and stats-based
- Blocking: e.g., same name
- Matching: e.g., avg similarity of attribute values
- Clustering: e.g., transitive closure, etc.

Supervised learning
- Random forest for matching
  F-msr: >95% w. ~1M labels
- Active learning for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

1969 (Pre-ML) — **~2000 (Early ML)** — **~2015 (ML)** — **2018 (Deep ML)**

Sup / Unsup learning
- Matching: Decision tree, SVM
  F-msr: 70%-90% w. 500 labels
- Clustering: Correlation clustering, Markov clustering

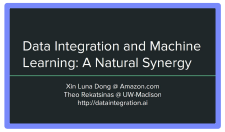
Deep learning
- Deep learning
- Entity embedding

HCC Institute of HUMAN-CENTRED COMPUTING

HALLOWEEN SPECIAL

Entity Linking Approaches

[Xin Luna Dong, Theodoros Rekatsinas: Data Integration and Machine Learning: A Natural Synergy. **PVLDB 2018**]



50 Years of Entity Linkage

Rule-based and stats-based
- Blocking: e.g., same name
- Matching: e.g., avg similarity of attribute values
- Clustering: e.g., transitive closure, etc.

Supervised learning
- Random forest for matching
  F-msr: >95% w. ~1M labels
- Active learning for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

1969 (Pre-ML) — ~2000 (Early ML) — ~2015 (ML) — 2018 (Deep ML)

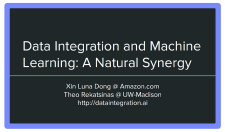
Sup / Unsup learning
- Matching: Decision tree, SVM
  F-msr: 70%-90% w. 500 labels
- Clustering: Correlation clustering, Markov clustering

Deep learning
- Deep learning
- Entity embedding

HALLOWEEN SPECIAL

Entity Linking Approaches

[Xin Luna Dong, Theodoros Rekatsinas: Data Integration and Machine Learning: A Natural Synergy. **PVLDB 2018**]



50 Years of Entity Linkage

Rule-based and stats-based
- Blocking: e.g., same name
- Matching: e.g., avg similarity of attribute values
- Clustering: e.g., transitive closure, etc.

Supervised learning
- Random forest for matching
  F-msr: >95% w. ~1M labels
- Active learning for blocking & matching
  F-msr: 80%-98% w. ~1000 labels

1969 (Pre-ML) — **~2000 (Early ML)** — **~2015 (ML)** — **2018 (Deep ML)**

Sup / Unsup learning
- Matching: Decision tree, SVM
  F-msr: 70%-90% w. 500 labels
- Clustering: Correlation clustering, Markov clustering

Deep learning
- Deep learning
- Entity embedding

HCC Institute of HUMAN-CENTRED COMPUTING

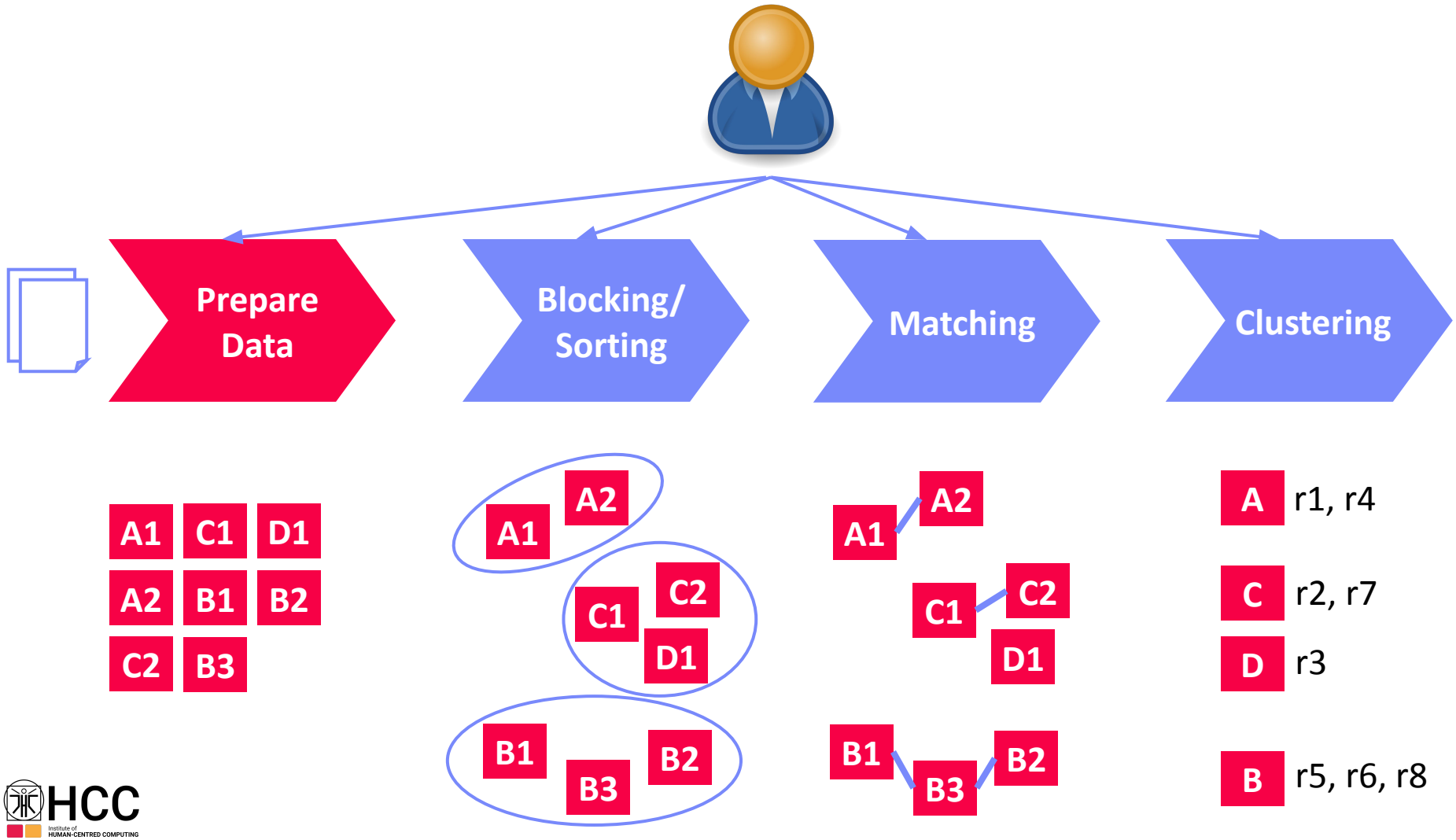
HALLOWEEN SPECIAL

Entity Embeddings

- Each **entity** (“AWS”, “AZURE”) → **vector** (list of numbers).
- **Vector’s numbers** are chosen → **similar entities** → **similar vectors**.
- E.g. “AWS” and “AZURE” vectors might be **close to each other** because they are both Cloud Services (**semantics capture**).
- Machines **learn vectors** automatically by **looking at how entities appear together in data** (similar to how word embeddings like BERT work for words).
- **Once entities are represented as vectors, the system can measure how similar they are using math (e.g., cosine similarity).**

HALLOWEEN SPECIAL

Entity Resolution Pipeline



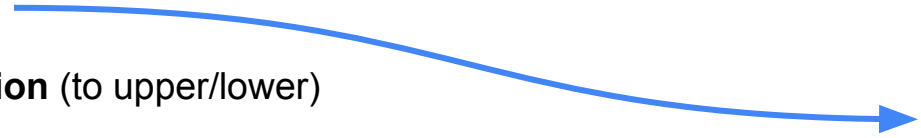
HALLOWEEN SPECIAL

Step 1: Data Preparation

- **#1 Schema Matching and Mapping**
  - See lecture **04 Schema Matching and Mapping**
  - Create **homogeneous schema** for comparison
  - **Split composite attributes**

Autonomous, heterogeneous systems

- **#2 Normalization**
  - Removal of **special characters** and **white spaces**
  - **Stemming**
  - **Capitalization** (to upper/lower)
  - **Remove** redundant words
  - **Resolve** abbreviations

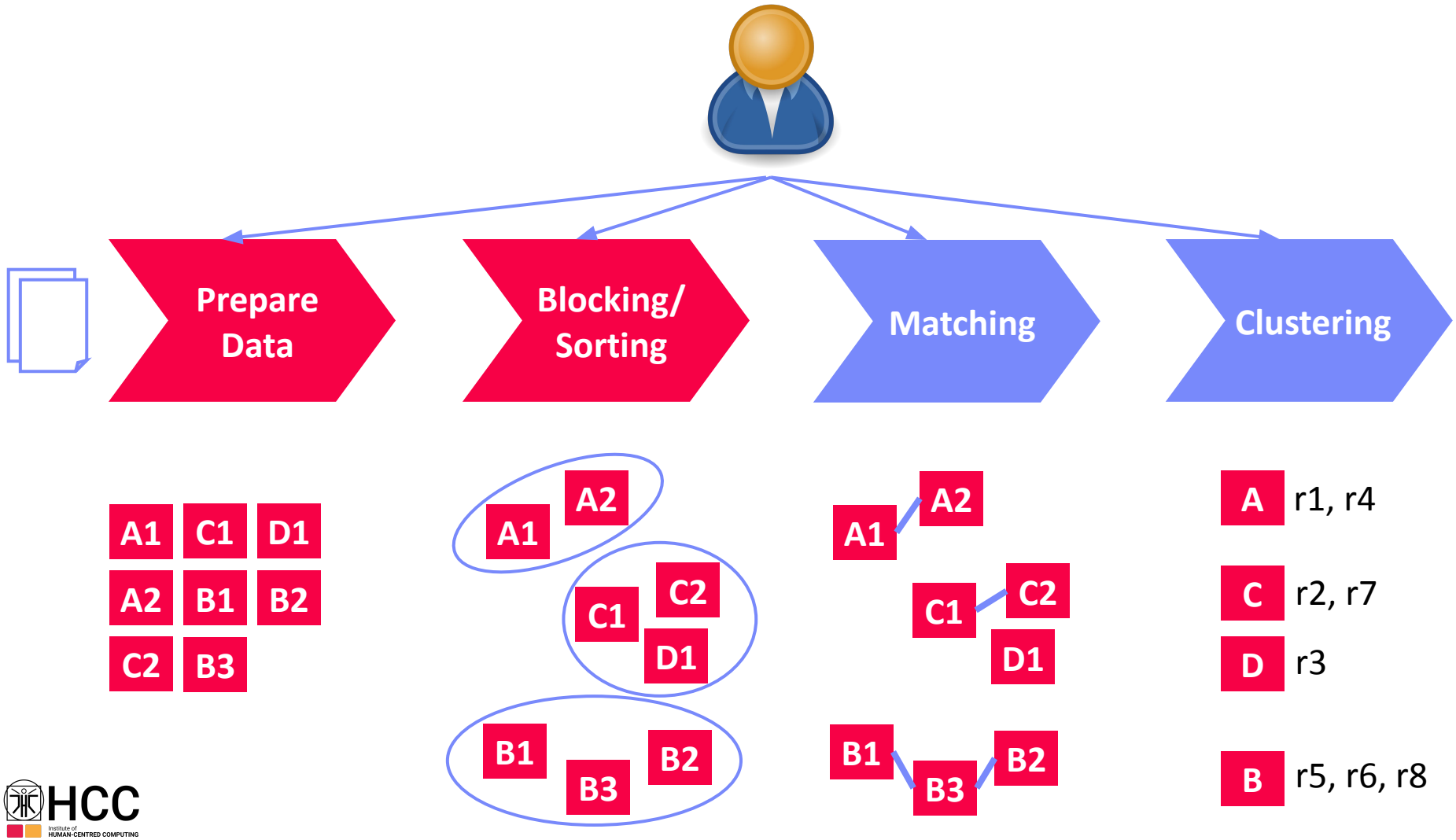


likes/liked/likely/liking
→ like

- **#3 Data Cleaning**
  - See lecture **06 Data Cleaning and Data Fusion**
  - Correct data corruption and inconsistencies

HALLOWEEN SPECIAL

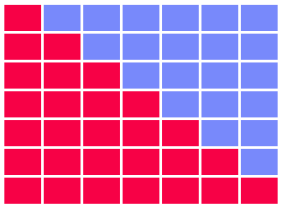
Entity Resolution Pipeline



HALLOWEEN SPECIAL

Step 2: Blocking and Sorting

- **#1 Naïve All-Pairs**
  - Brute-force, naïve approach
    → n*(n-1)/2 pairs → **O(n^2) complexity**



HALLOWEEN SPECIAL

Step 2: Blocking and Sorting

- **#2 Blocking / Partitioning**
  - Efficiently create **small blocks of similar records** for pairwise matching
  - **Basic:** equivalent values on selected attributes (name)
  - **Predicates: whole** field, **token** field, common **integer**, same x char start, **n-grams**
  - **Hybrid:** disjunctions/conjunctions
  - **Blocking Keys:**

→ JR01111

John Roberts	20 Main St	Plainville	MA	01111
Julia Ray	32 Main St	Plainville	MA	01111

  - **Learned:** Minimal rule set via greedy algorithms

→ **Significant reduction:** 1M records → 1T pairs (**Naive**)
→ 1K partitions w/ 1K records → 1G pairs (**1000x**)

[Nicholas Chammas, Eddie Pantrige: Building a Scalable Record Linkage System, **Spark+AI Summit 2018**]

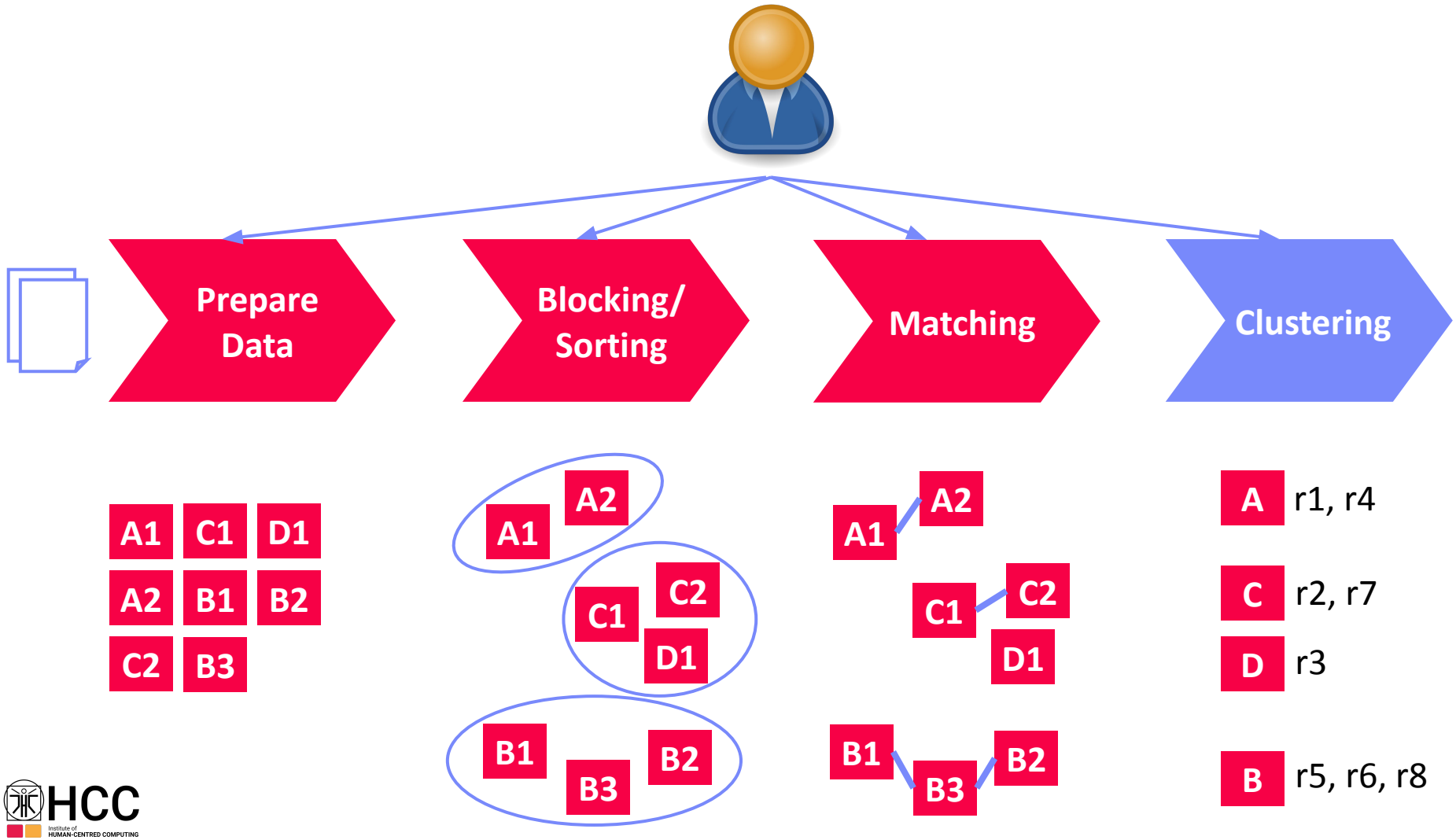
HALLOWEEN SPECIAL

Step 2: Blocking, cont.

- **#3 Sorted Neighborhood**
  - Define **sorting keys** (similar to blocking keys)
  - Sort records by sorting keys
  - Define **sliding window of size m** (e.g., 100) and compute all-pair **matching within sliding window**

HALLOWEEN SPECIAL

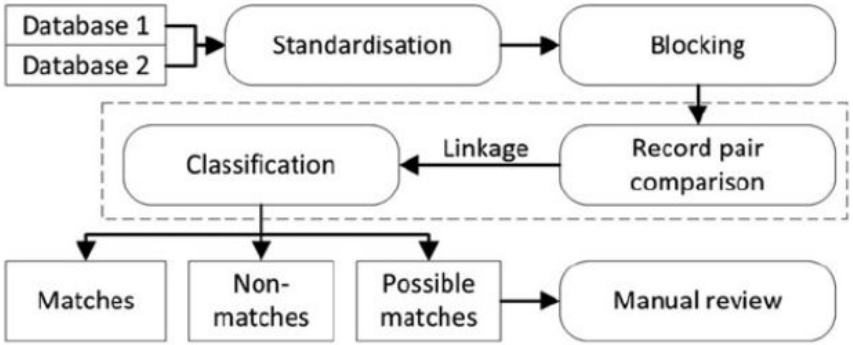
Entity Resolution Pipeline



HALLOWEEN SPECIAL

Step 3: Matching

- **#1 Basic Similarity Measures**
  - Pick **similarity measure** sim(r, r') **and thresholds:** high θ_h (and low θ_l)
  - Record similarity: avg attribute similarity
  - **Match:** sim(r, r') > θ_h **Non-match:** sim(r, r') < θ_l
    possible match: θ_l < sim(r, r') < θ_h



[O'Hare, K.et.al. D. P., & A. Jurek-Loughrey,2019]

HALLOWEEN SPECIAL

Step 3: Matching

- **#2 Learned Matchers (Traditional ML)**
  - Labeling (select samples of pairs between data sets and label them)
    - Match (1) o Non-match (0).
  - Selection of samples for labeling
    - **Sufficient:** enough samples
    - **Suitable:** accurately represent the different types of matches.
    - **Balanced**: cases with similar proportions of matches and non-matches.
  - **Phase 1:** Model Generation (We select a model and train)
    - **SVM** and **decision trees**, **logistic regression**, **random forest**, XGBoost
  - **Phase 2:** Model Application
    - The trained model is applied to new pairs of records to predict whether they match or not.

[Mikhail Bilenko, Raymond J. Mooney: Adaptive duplicate detection using learnable string similarity measures. **KDD 2003**]

[Hanna Köpcke, Andreas Thor, Erhard Rahm: Evaluation of entity resolution approaches on real-world match problems. **PVLDB 2010**]

[Xin Luna Dong: Building a Broad Knowledge Graph for Products. **ICDE 2019**]

HALLOWEEN SPECIAL

Step 3: Matching, cont.

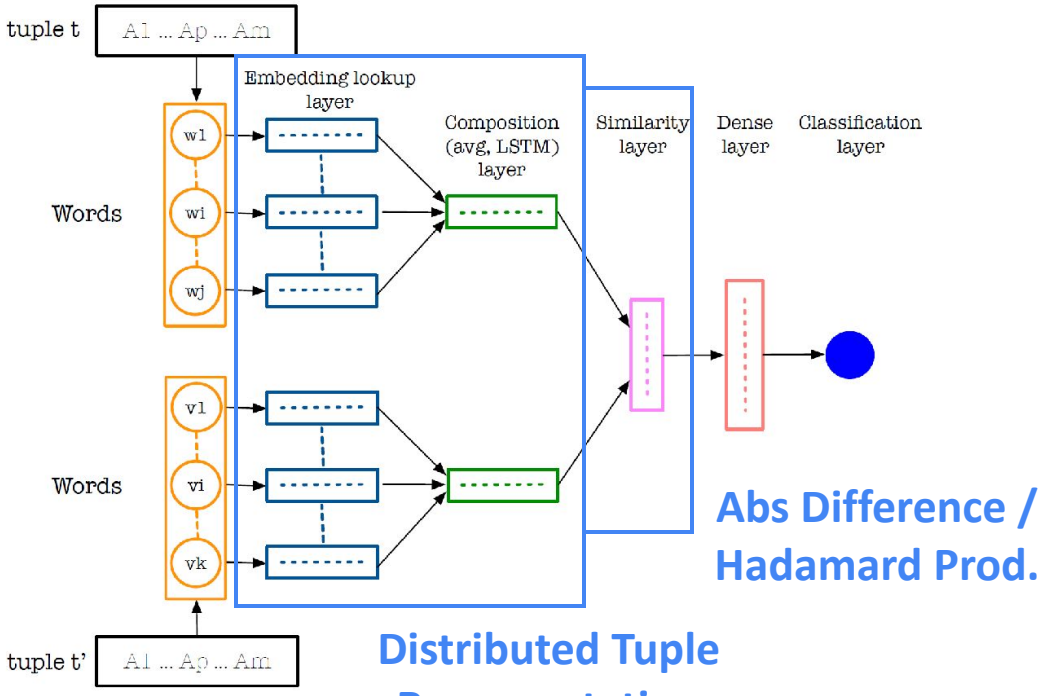
- **Deep Learning for ER**
  - Automatic **representation learning** from text (avoid feature engineering)
  - Leverage pre-trained **word embeddings for semantics** (no syntactic limitations)

- **Example DeepER**

[Muhammad Ebraheem et al: Distributed Representations of Tuples for Entity Resolution. **PVLDB 2018**]

- **Example Magellan**
  - DL for text and dirty data

[Sidharth Mudgal et al: Deep Learning for Entity Matching: A Design Space Exploration. **SIGMOD 2018**]



HCC Institute of HUMAN-CENTRED COMPUTING

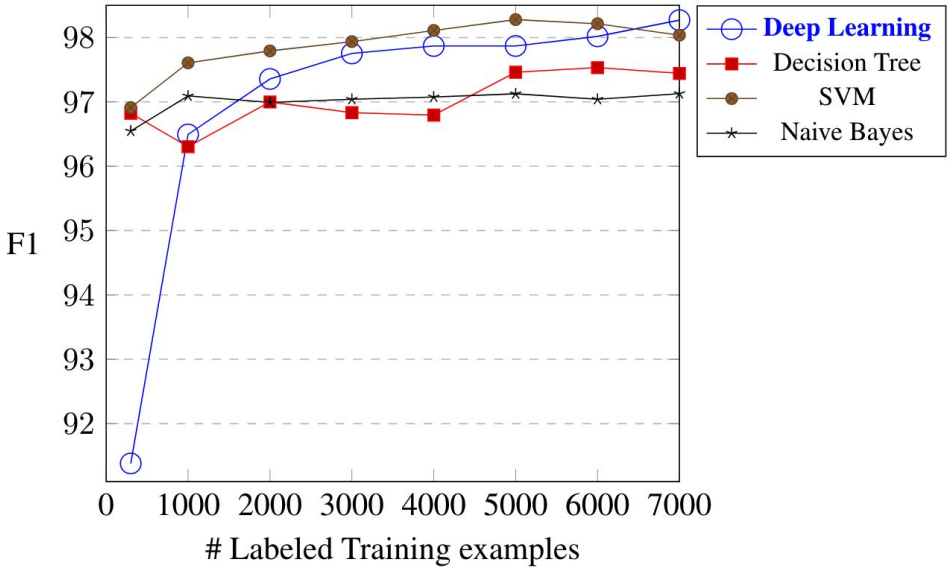
HALLOWEEN SPECIAL

Step 3: Matching, cont.

- **Labeled Data**
  - **Lack of experts**
  - **Class imbalance**
    - Oversampling (++ minor class)
    - Undersampling (-- major class)

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

DBLP-ACM



→ **Transfer Learning**

- Learn model from **high-resource ER scenario**
- Fine-tune using **low-resource** examples

→ **Active Learning**

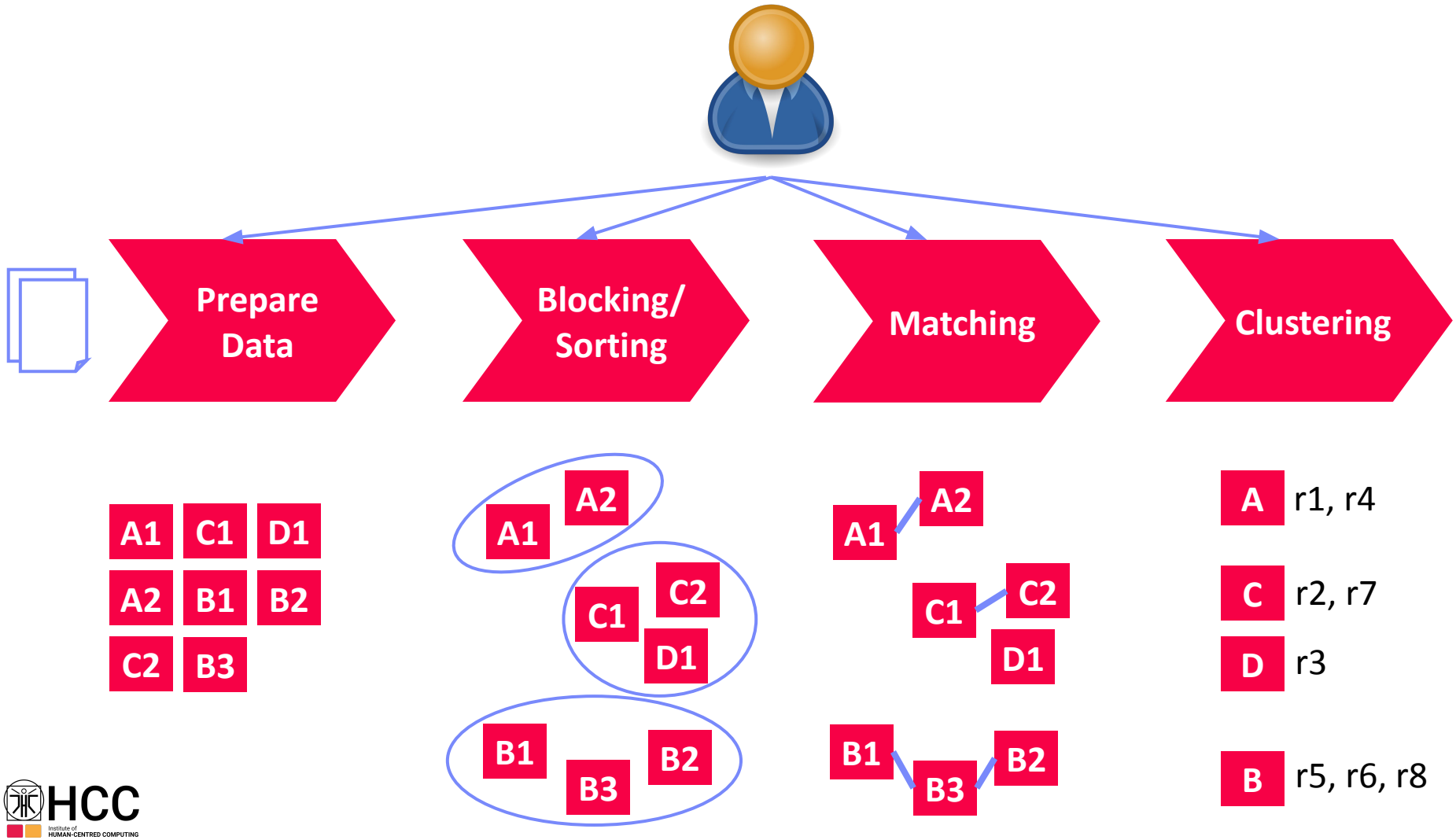
- Select instances for tuning to min labeling

[Sairam Gurajada, Lucian Popa, Kun Qian, Prithviraj Sen: Learning-Based Methods with Human in the Loop for Entity Resolution, Tutorial, **CIKM 2019**]

HCC Institute of HUMAN-CENTRED COMPUTING

HALLOWEEN SPECIAL

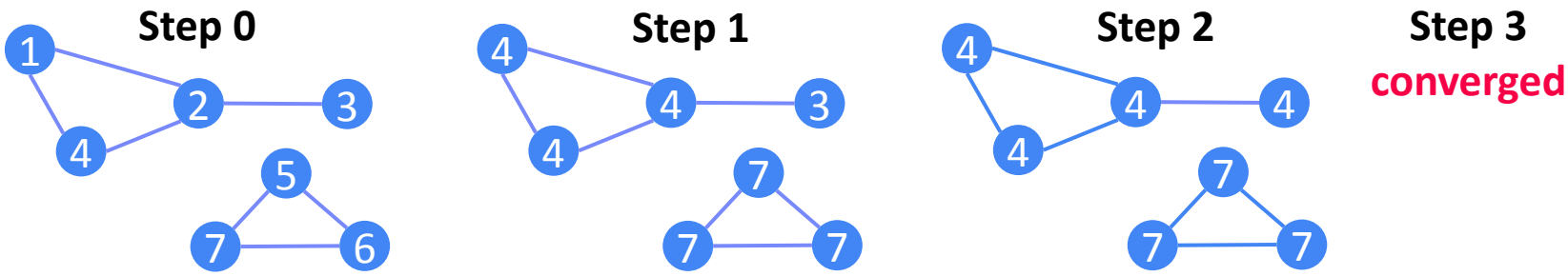
Entity Resolution Pipeline



HALLOWEEN SPECIAL

Step 4: Clustering

- **Recap: Connected Components**
  - Determine connected components of a graph (subgraphs of **connected nodes**)
  - Propagate max(current, msgs) if != current to neighbors, terminate if no msgs



- **Clustering Approaches**
  - **Basic:** connected components (transitive closure) w/ edges sim > θ_h → Issues: **big clusters** and **dissimilar records**

HALLOWEEN SPECIAL

Step 4: Clustering

- **Clustering Approaches**
  - **Correlation clustering:** global optimization based on similarity
    - High-sim → same cluster
    - Low-sim → different cluster
    - NP-hard!!!
  - **Markov clustering:** stochastic flow simulation via random walks
    - More walks between nodes → same cluster
    - Less walks between nodes → different cluster

[Oktie Hassanzadeh, Fei Chiang, Renée J. Miller, Hyun Chul Lee: Framework for Evaluating Clustering Algorithms in Duplicate Detection. **PVLDB 2009**]

HALLOWEEN SPECIAL

Entity Resolution Tools

HALLOWEEN SPECIAL

Python Dedupe

https://docs.dedupe.io/en/latest/API-documentation.html
https://dedupeio.github.io/dedupe-examples/docs/csv_example.html

- **Overview**
  - **Python library for data deduplication** (entity resolution)
  - **By default:** logistic regression matching (and blocking)
- **Example**

```
fields = [
  {'field':'Site name', 'type':'String'},
  {'field':'Address', 'type':'String'}]
deduper = dedupe.Dedupe(fields)

# sample data and active learning
deduper.sample(data, 15000)
dedupe.consoleLabel(deduper)
```

```
Do these records refer
to the same thing?
(y)es / (n)o /
(u)nsure / (f)inished
```

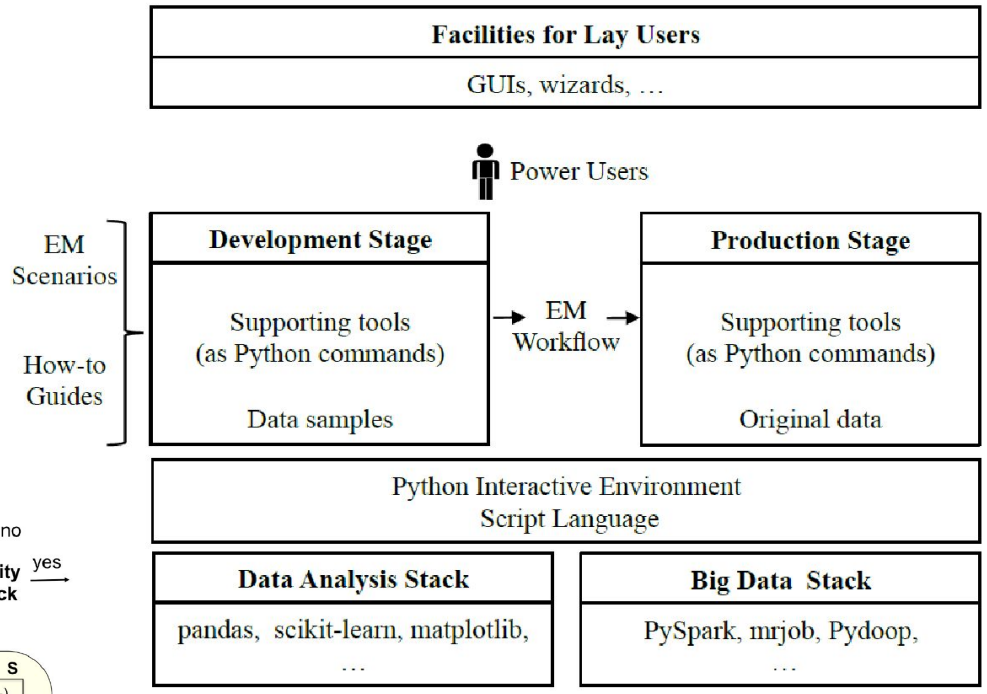
```
# learn blocking rules and pairwise classifier
deduper.train()

# Obtain clusters as lists of (RIDs and confidence)
threshold = deduper.threshold(data, recall_weight=1)
clustered_dupes = deduper.match(data, threshold)
```

HALLOWEEN SPECIAL

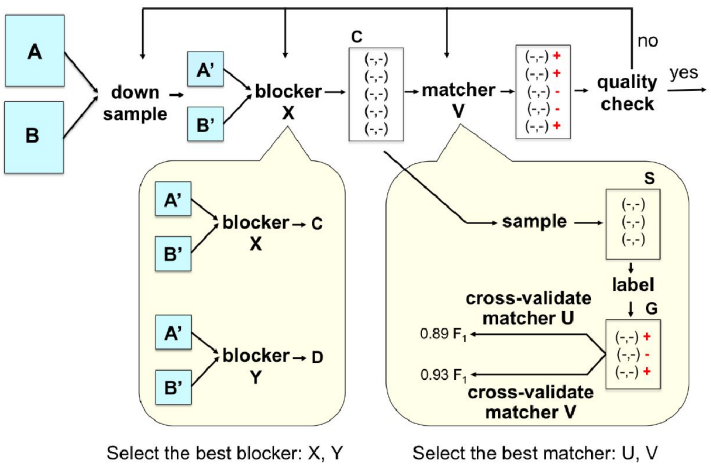
Magellan (UW-Madison)

[Pradap Konda et al.: Magellan: Toward Building Entity Matching Management Systems. **PVLDB 2016**]

- **System Architecture**
  - **How-to guides for users**
  - Tools for individual steps of **entire ER pipeline**
  - Build on top of existing Python/big data stack
  - Scripting environment for power users



User



[Yash Govind et al: Entity Matching Meets Data Science: A Progress Report from the Magellan Project. **SIGMOD 2019**]

HCC Institute of HUMAN-CENTRED COMPUTING

Example Applications

HALLOWEEN SPECIAL

Record Linkage

- **Task: Distributed Entity Resolution on Apache Spark**
  - Uni Leipzig Benchmarks
- **Example 1: DBLP, ACM, Google Scholar Publications**
  - `(title, authors, venue, year)`
  - Basic preprocessing via title capitalization, etc
  - How about leveraging the linked PDF papers?
- **Example 2: Amazon, Google Products**
  - `(name, description, manufacturer, price)`
  - NLP for matching medium and long descriptions, e.g., word embeddings
  - How about leveraging the product images (different angles)
- **SIGMOD Programming Contest 2022**
  - Design blocking scheme for Notebooks specifications dataset

https://dbs.uni-leipzig.de/research/projects/object_matching/benchmark_datasets_for_entity_resolution

In practice:
multi-modal data, and
feature engineering

https://dbgroup.ing.unimore.it/sigmod22contest/task.shtml?content=description

HALLOWEEN SPECIAL

Data Management – Autograding

- Plagiarism Detection via Entity Resolution
  - https://issues.apache.org/jira/browse/SYSTEMDS-3191
  - **Data preparation:** file names/properties, runtime, correctness
  - **Blocking:** by programming language, results sets
  - **Matching**
    - Exact matches via basic diff + threshold
    - Code similarity via SotA embeddings
  - **Clustering**
    - Connected components within each block (min sim threshold)

[Fangke Ye et al: MISIM: An End-to-End Neural Code Similarity System. **CoRR 2020** arxiv.org/pdf/2006.05265.pdf]

HALLOWEEN SPECIAL

Summary and Q&A

- **Motivation and Terminology**
- **Entity Resolution Concepts**
  - **Data Preparation**
  - **Blocking/Sort**
  - **Matching**
  - **Clustering**
- **Entity Resolution Tools**
- **Example Applications**
- Next Lectures (Data Integration Architectures)
  - **November 7. No lecture (Time for project preparation)**
  - **November 14. Data Cleaning and Data Fusion**

Incremental Data Deduplication

- **Goals**
  - Incremental stream of updates
    → previously **computed results obsolete**
  - Same or **similar results** AND **significantly faster** than batch computation

[Anja Gruenheid, Xin Luna Dong, Divesh Srivastava: Incremental Record Linkage. **PVLDB 2014**]

- **Approach**
  - End-to-end incremental record linkage for new and changing records
  - Incremental maintenance of similarity graph and incremental graph clustering
  - Initial graph created by **correlation clustering**
  - Greedy update approach in polynomial time
    - Directly connect components from increment ΔG into Q
    - **Merge** of **pairs of clusters** to obtain better result?
    - **Split** of **cluster into two** to obtain better result?
    - **Move** nodes **between two clusters** to obtain better result?