

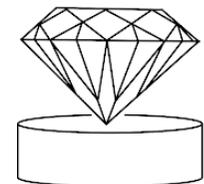
# Machine Learning-Based Data Cleaning : Current Solutions and Challenges

**Laure Berti-Equille**

IRD Montpellier  
Aix-Marseille University, CNRS, LIS, DIAMS  
France

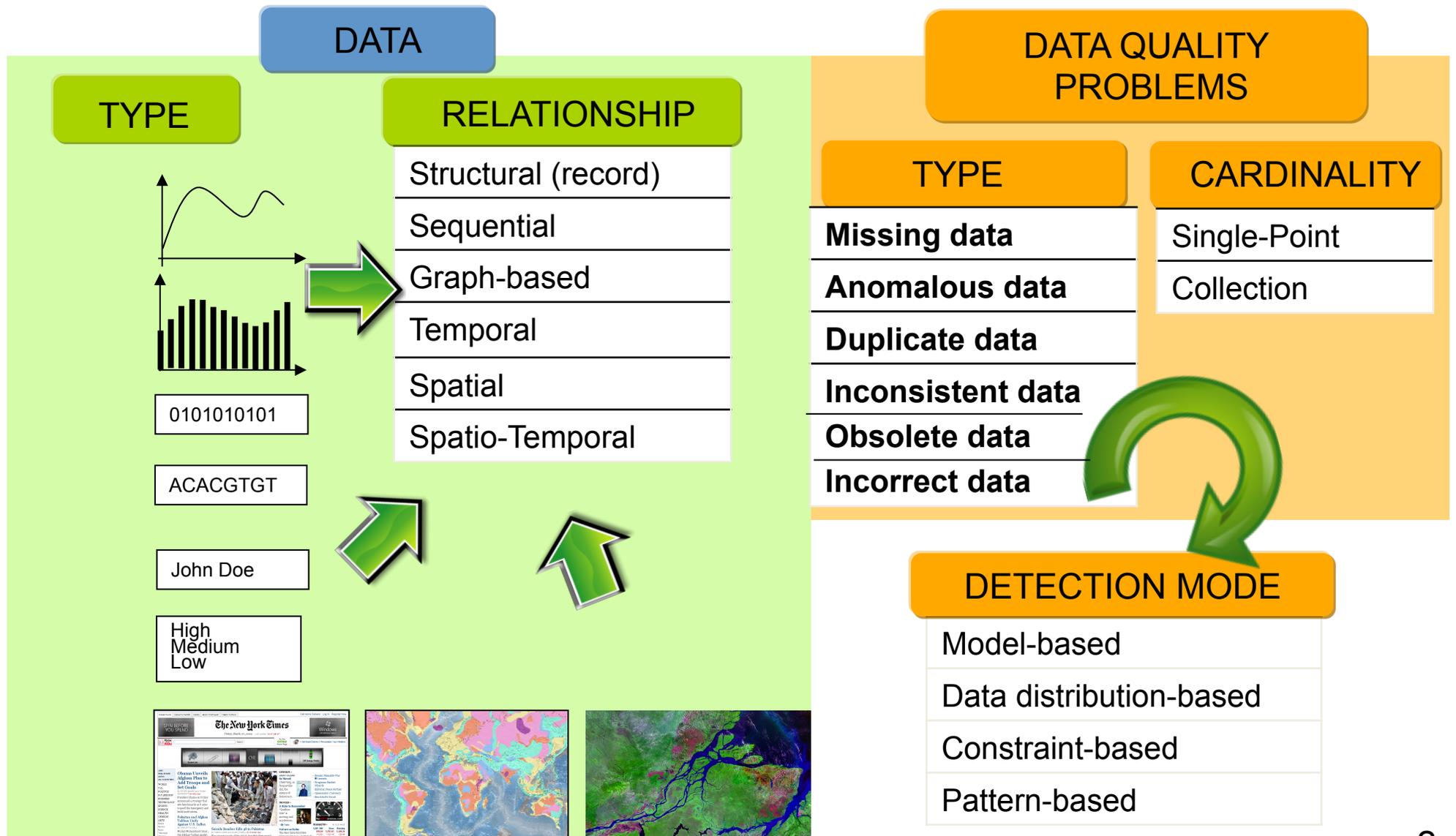
[laure.berti@ird.fr](mailto:laure.berti@ird.fr)

<http://pageperso.lif.univ-mrs.fr/~laure.berti/>



<https://diams.lis-lab.fr/>

# Data Quality Problems



# Example I

Relational data : CiDE.21 committee

Nom	Etablissement	Ville	Tel
Prof. B. JACQUEMIN	Univ. Lille GERiCO	Lyon	+33 (0) 3 20 41 66 38
Malek GHENIMA	ESC Tunis	Tunis	+216 71600615
Anis BEN MAMI	ESC Tunis	Tunis	74415567
M. GHENIHA	Tunis	Univ. de la Manouba	+216 71600615
Mehdi BEN GHANEM	NULL	Tunis	NULL
Hamida AMDOUN		ESEN-14009	00000000

**Representation** (points to the table header)

**Misfielded Value** (points to the 'Lyon' cell)

**Duplicates** (points to 'M. GHENIHA' and 'Tunis' in the 4th row)

**Typos** (points to 'AMDOUN' in the 6th row)

**Inconsistencies** (points to 'ESEN-14009' in the 6th row)

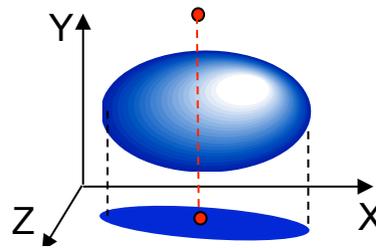
**Obsolete Value** (points to 'ESEN-14009' in the 6th row)

**Incorrect Values** (points to '74415567' in the 3rd row)

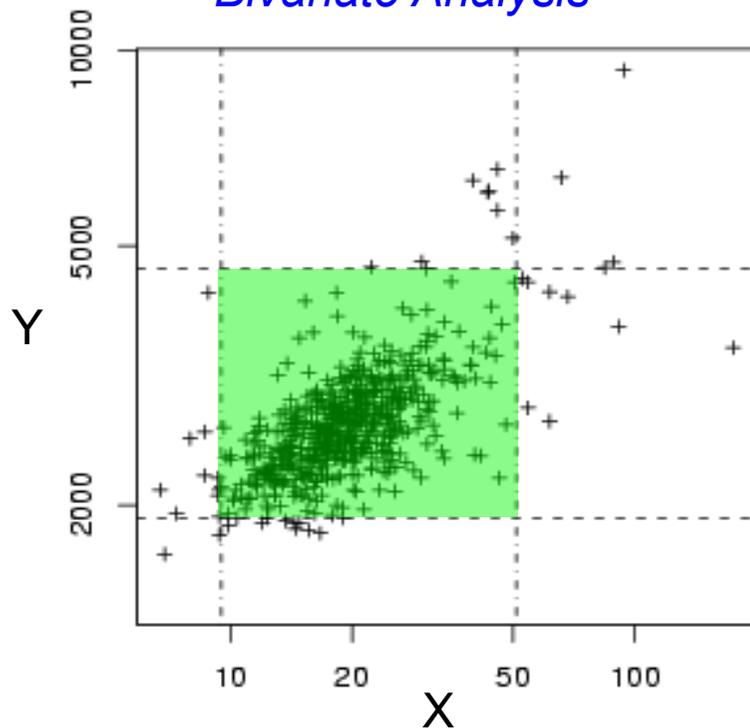
**Missing Values** (points to 'NULL' in the 5th row)

# Example 2

*Outliers*

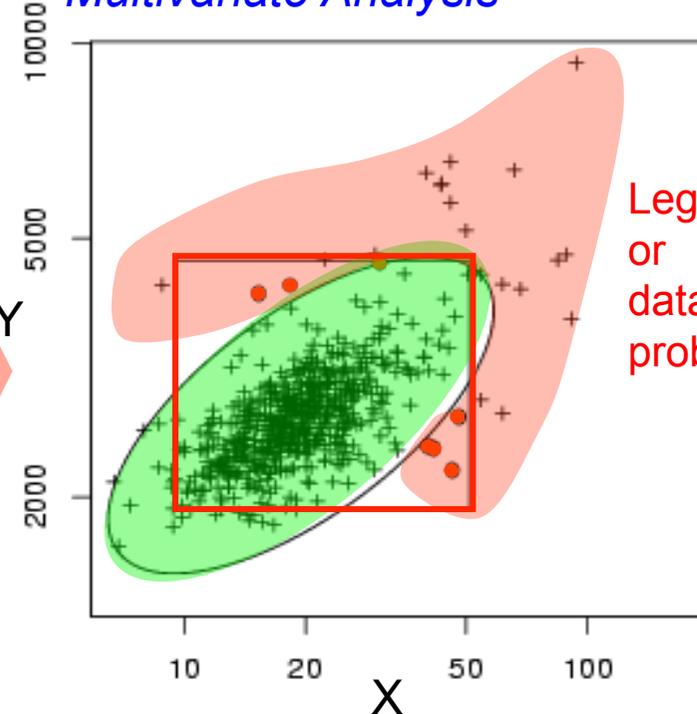
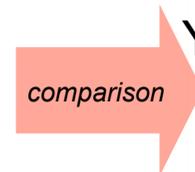


*Bivariate Analysis*



Rejection area: Data space excluding the area defined between 2% and 98% quantiles for X and Y

*Multivariate Analysis*



Rejection area based on:  
 $\text{Mahalanobis\_dist}(\text{cov}(X,Y)) > \chi^2(.98,2)$

# Example 3

## Disguised missing data

Some are obvious...

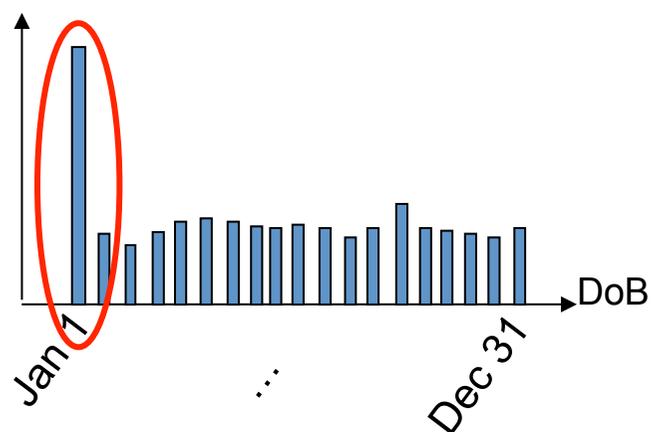
Detectable with syntactical or domain constraints

Phone number: 999-999-9999

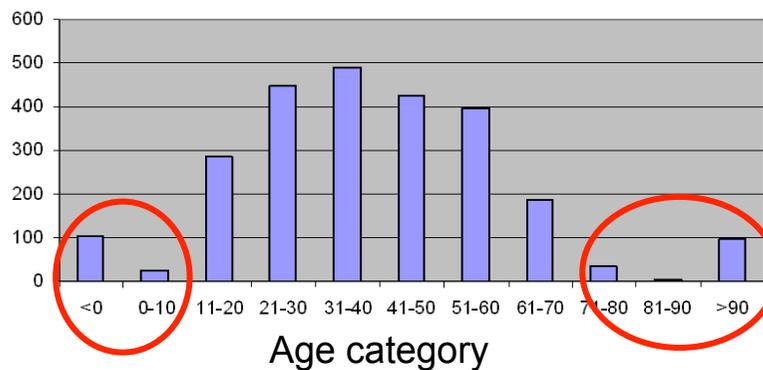
Others are not....

Could be suspected because the data distribution doesn't conform to the expected model

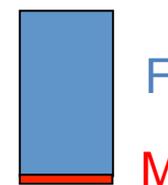
Histogram of DoBs per day of the year



Histogram of online shopping customers per age category



2% patients in the obstetrical emergency service are *male*...



# Example 4

*Are the information sources equally accurate, up-to-date, and trustworthy?*

## AFP apologises to French industrialist after death reported



February 28, 2015 2:42 PM



© REUTERS/ BENOIT TESSIER

**French TV Denies Reports of Bouygues Conglomerate CEO's Death**

AFP issued an apology to French industrialist Martin Bouygues, chairman and CEO of the conglomerate Bouygue...

# Example 5

## Rumors: Celebrity Death Hoaxes



**成龍 Jackie Chan**  
June 21

Hi everybody! Yesterday, I got on a 3am flight from India to Beijing. I didn't get a chance to sleep and even had to clean my house when I got home. Today, everybody called to congratulate me on my rumored engagement. Afterward, everybody called me to see if I was alive.

If I died, I would probably tell the world! I took a photo with today's date, just in case you don't believe me! However, thank you all for your concern. Kiss kiss and love you all!

P.S. My dog is healthy, just like me! He doesn't need surgery! By the way, my dogs are golden retrievers, not Labradors.

Irene Ennenbach, Kimyong Fu Fu, Daniel K... others like this.

10,816 shares

View previous comments

**Damian Lulko** oje  
See Translation  
30 minutes ago

**Rose Quayle** Long live the hero for a thought u turned into chuck norris  
16 minutes ago

**Travis Taylor** shh iackie



DWAYNE JOHNSON died while filming a dangerous stunt for FAST & FURIOUS 7



**R.I.P Morgan Freeman**  
860,689 likes · 972,460 talking about this

Like Message

Community  
At about 5 p.m. ET on Thursday, our beloved actor Morgan Freeman passed away due to an artery rupture. Morgan was born on June 1, 1937. He will be missed but not forgotten. Please show your sympathy and condolences by commenting on and liking this page.



# ML Revolutionizes Industry

## Security and Surveillance

Facial and character recognition, automatic fraud detection, plagiarism detection, DDoS detection, etc.



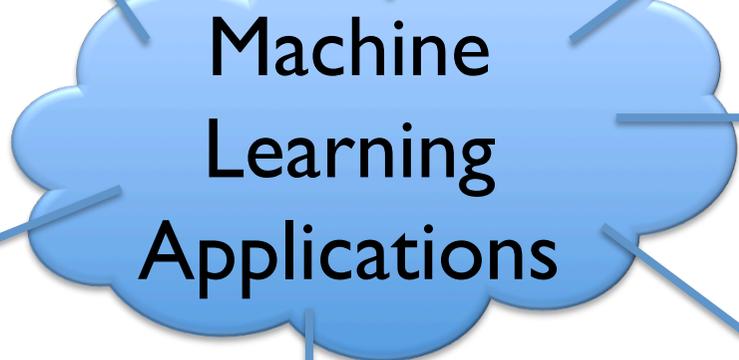
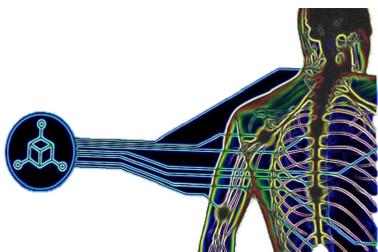
## Manufacturing

optimizing fab operations, automating quality testing, inventory, asset, and supply chain management, predictive maintenance, etc.



## eHealth

Automate screening tool for medical imagery diagnostics, bio-augmentation, etc.



## Autonomous vehicles



## Smart eCommerce

Product recommendations, demand forecasting, search, classification, matching, etc.



## Digital Marketing

User conversion prediction, Ad scoring, customer targeting, brand tracking, viral marketing analysis, etc.



## Personal assistant

Predictive help, automatic speech recognition, dialog management, etc.

# Hot Topic for DB community

[VLDB'17 Keynote]

## Deep Learning (m)eats Databases

(shortened)

Jens Dittrich

### Machine Learning and Databases: The Sound of Things to Come or a Cacophony of Hype?

Divy Agrawal  
University of Washington  
dagrawal@cs.washington.edu

Magdalena Balazinska  
University of Michigan  
magda@cs.washington.edu

Michael Cafarella  
University of Michigan  
michjo@umichigan.edu

Tim Kraska  
Brown University  
tim.kraska@brown.edu

Michael Jordan  
UC Berkeley  
jordan@cs.berkeley.edu

Raghu Ramkrishnan  
Microsoft  
raghu@microsoft.com

Christopher Ré  
Stanford  
chrismre@cs.stanford.edu

Categories and Subject Descriptors  
H.2.0 (Information Systems): Database Management

General Terms  
Database Research, Machine Learning

Keywords  
Database Research, Machine Learning, Panel

#### 1. INTRODUCTION

Machine learning seems to be eating the world with a new breed of high-value data-driven applications in image analysis, search, voice recognition, mobile, and office productivity products. To paraphrase Mike Stonebraker, machine learning is no longer a zero-billion-dollar business. As the leader of high-value, data-driven applications for over four decades, a natural question for database researchers to ask is: what role should the database community play in these new data-driven machine-learning-based applications?

The last few years have seen increasing crossover between database research and machine learning. But is this crossover a wise choice for database research? What are the opportunities and the costs of this approach to industry to the future of database research, and to academics? Do database researchers have something to contribute to this trend? These two areas have distinctive traditions in both research, intellectually, and in industry, so bridging the gap between the fields is likely to require considerable effort. Is it worth it?

• What are the most interesting research problems at this intersection? Are there core intellectual problems in machine learning that can only be solved with researchers from both sides? Or are the problems all data/analytics work? If it is data/analytics work, is it sufficiently interesting/justifiable to examine in research?

• Is there anything fundamentally different about building database systems that use machine learning or are designed to support machine learning? Or are those new systems just the same old thing, rebranded with water packaging?

• To attract partners in the machine learning side of the world, we need to be viewed as providing intellectual value. What do database people know that is useful to machine learning? At which level is our knowledge useful? Should we regard machine learning as a black box? Should we apply our ideas inside the black box? Should we build systems that make the black box happy? Where is the most bang for the buck?

• Do we need a new conference on ML+Databases? Or is SIGMOD or KDD the right place?

• What is the risk to the database community if database people build machine learning tools? Could this lead to us becoming a "me-too" community, i.e., a lagging—rather than a leading—indicator? Or is this risk higher if we don't jump on the machine learning bandwagon like other fields, notably NLP and Computer Vision?

• Can we teach old dogs new tricks? Does working at the intersection of machine learning and databases require that database researchers learn an entirely new set of skills? In contrast, while Database research is applied to and often driven by business, there are few

[ICDE'18 Tutorial]

[SIGMOD'17 Tutorial]

### Database Meets Deep Learning: Challenges and Opportunities

Wei Wang<sup>1</sup>, Meihui Zhang<sup>1</sup>, Gang Chen<sup>1</sup>,  
H. V. Jagadish<sup>2</sup>, Beng Chin Ooi<sup>3</sup>, Kim-Loe Tan<sup>4</sup>  
<sup>1</sup>National University of Singapore <sup>2</sup>Singapore University of Technology and Design  
<sup>3</sup>Zhejiang University <sup>4</sup>University of Michigan  
†{wangwei, oobc, tanw}@comp.nus.edu.sg †meihui.zhang@sutd.edu.sg  
cg@zju.edu.cn jag@umich.edu

#### ABSTRACT

Deep learning has recently become very popular on account of its incredible success in many complex data-driven applications, including image classification and speech recognition. The database community has worked on data-driven applications for many years, and therefore should be playing a lead role in supporting this new wave. However, databases and deep learning are different in terms of both techniques and applications. In this paper, we discuss research problems at the intersection of the two fields. In particular, we discuss possible improvements for deep learning systems from a database perspective, and analyze database applications that may benefit from deep learning techniques.

#### 1. INTRODUCTION

In recent years, we have witnessed the success of numerous data-driven machine-learning-based applications. This has prompted the database community to investigate the opportunities for integrating machine learning techniques in the design of database systems and applications [29]. A branch of machine learning, called deep learning [22, 18], has attracted worldwide interest in recent years due to its excellent performance in multiple areas including speech recognition, image classification and natural language processing (NLP). The foundation of deep learning was established about twenty years ago in the form of neural networks. Its recent resurgence is mainly fueled by three factors: immense computing power, which reduces the time to train and deploy new models, e.g., Graphic Processing Unit (GPU) enables the training systems to run much faster than those in the 1990s; massive (labeled) training datasets (e.g., ImageNet) enables more comprehensive

optimization and large scale data-driven applications since 1970s, which are closely related to the first two factors. It is natural to think about the relationships between databases and deep learning. First, are there any insights that the database community can offer to deep learning? It has been shown that larger training datasets and a deeper model structure improves the accuracy of deep learning models. However, the side effect is that the training becomes more costly. Approaches have been proposed to accelerate the training speed from both the system perspective [5, 19, 9, 28, 11] and the theory perspective [45, 12]. Since the database community has rich experience with system optimization, it would be opportune to discuss the applicability of database techniques for optimizing deep learning systems. For example, distributed computing and memory management are key database technologies. They are also central to deep learning.

Second, are there any deep learning techniques that can be adapted for database problems? Deep learning emerged from the machine learning and computer vision communities. Recently, it has been successfully applied to other domains, like NLP [18]. However, few studies have been conducted using deep learning techniques for database problems. This is partially because traditional database problems—like indexing, transaction and storage management—involve less uncertainty, whereas deep learning is good at predicting over uncertain events. Nevertheless, there are problems in databases like knowledge fusion [10] and crowdsourcing [27], which are probabilistic problems. It is possible to apply deep learning techniques in these areas. We will discuss specific problems like querying interfaces, knowledge fusion, etc. in this paper.



[workshop@SIGMOD]

[SIGMOD Record 2016]

[SIGMOD'15 Panel]

### Data Management in Machine Learning: Challenges, Techniques, and Systems

Arun Kumar  
UC San Diego  
La Jolla, CA, USA

Matthias Boehm  
IBM Research – Almaden  
San Jose, CA, USA

Jun Yang  
Duke University  
Durham, NC, USA

#### ABSTRACT

Large-scale data analytics using statistical machine learning (ML), popularly called advanced analytics, underpins many modern data-driven applications. The data management community has been working for over a decade on tackling data management-oriented challenges that arise in ML workloads, and has built several systems for advanced analytics. This tutorial provides a comprehensive review of

focus is on analyzing the technical challenges and on explaining the key ideas, architecture, strengths, and limitations of major systems that address these challenges. This tutorial aims to provide data management researchers and systems developers with a survey of effective techniques and open issues, and to help identify systems they could build upon or compare with. It could also help data scientists understand the assumptions, pros, and cons of different systems and make more informed choices for their applications.

ACM SIGMOD Blog

Azza Abouzied and Paolo Papotti

FEBRUARY 14, 2018

COURTING ML: WITNESSING THE MARRIAGE OF RELATIONAL & WEB DATA SYSTEMS TO MACHINE LEARNING

Big Data, Databases, Machine Learning

The web is an ever-evolving source of information, with data and knowledge derived from it powering a great range of modern applications. Accompanying the huge wealth of information, web data also introduces numerous challenges due to its size, diversity, volatility, inaccuracy, and contradictions. This year's WebDB 2018 theme emphasizes the challenges and opportunities that arise at the intersection of web data and machine learning research. On one hand, a large portion of web data fuels ML, with novel applications such as predictive analytics, Q&A chat bots, and content generation. On the other hand, the new wave of ML technology found its way into traditional Web data challenges, with contributions such as web data extraction with deep learning, and using ML to optimize data processing pipelines.

To kick start the conversation on research at the cross hairs of ML and data, we interviewed Luna Dong (Amazon Research), Alkis Polyzotis (Google), Jens Dittrich (Saarland University), Arun Kumar (University of California, San Diego) and Peter Bailis (Stanford University). Below you will find their bios. We selected this diverse set of academic and industrial, systems and theoretical researchers to better understand the quickly evolving research field of Machine Learning and Database Systems. We asked them about their motivation for working in this field, their current work and their view on the future. We summarize our interviews along the following four questions.

[SIGMOD Blog, Feb. 2018]

# Introduction : DB perspective

Many problems in data management need precise knowledge and reasoning about information content and linkage for tasks as:

- Information and structure extraction

- Data curation



Our focus

- Data integration

- Querying & DB administration

- Privacy preservation

- Data storage

Many DM tasks can be reformulated as a classification or an optimization problem.

# Goals

- Offer an overview of ML applications to specific areas of data curation
- Analyze when and how ML might be leveraged for developing new areas of data management
- Analyze how data management could help ML workflows and data pipelines and contribute to ML advances
- Discuss about our ML journey in DB research community and how this can apply to yours

# Disclaimer

- Not specific to ML pipelines, systems or techniques
  - [Kumar, Boehm, Yang, Tutorial SIGMOD'17]  
[Polyzotis et al., Tutorial SIGMOD'17]
- Not trying to cover all domain-specific methods
- Not specific to data integration
  - [Dong, Rekatsinas, *coming* Tutorial SIGMOD'18]
- Not specific to “Deep Learning” nor “Big Data”
- Not exhaustive for the sake of conciseness

# Outline

## Introduction

- Motivations
- SWOT Analysis

## ML-Powered Data Curation

- Record Linkage, Deduplication, Entity Resolution
- Error Repair and Pattern Enforcement
- Concluding Remarks and Open Issues

# Outline

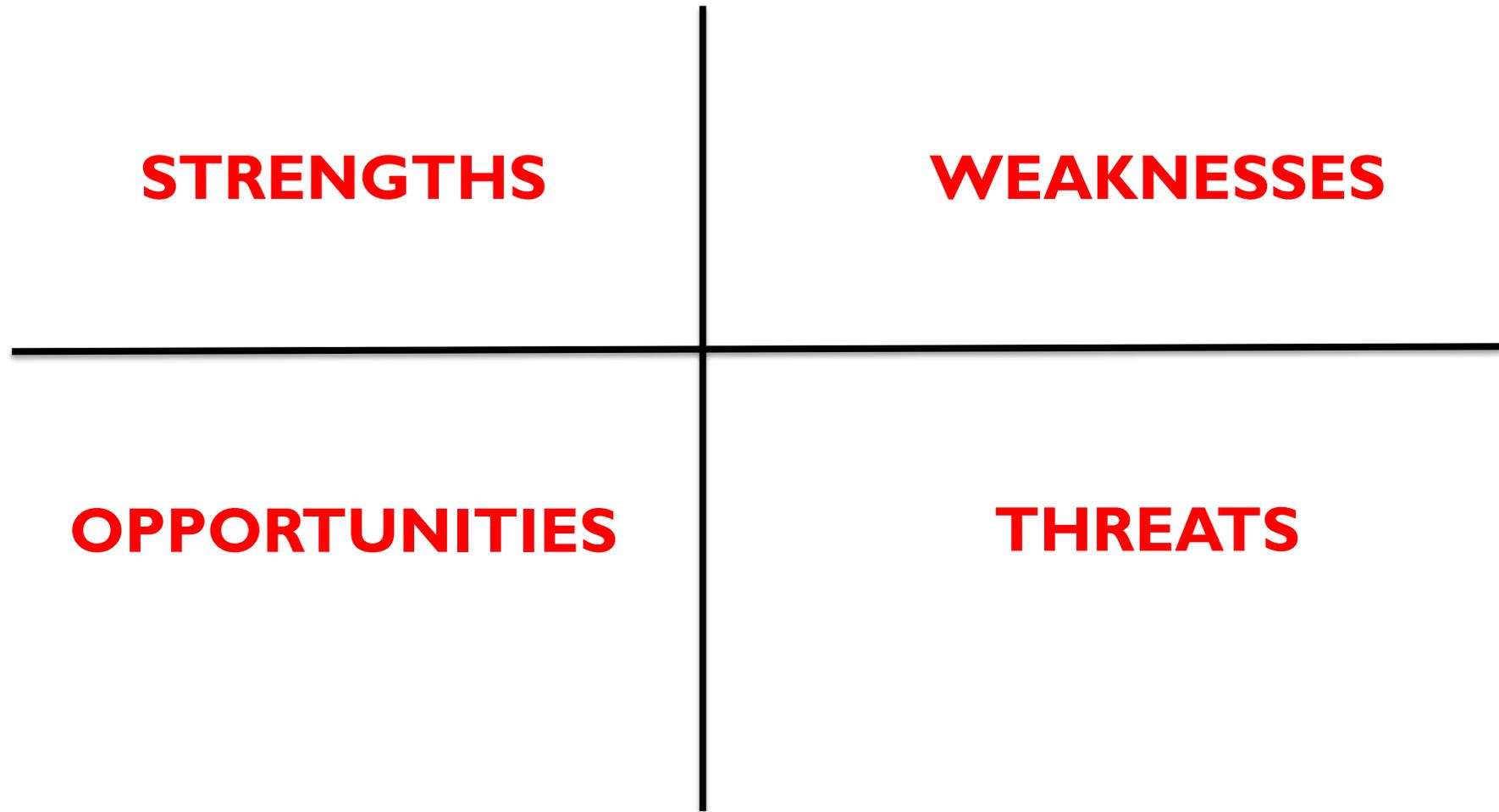
## Introduction

- Motivations
- **SWOT Analysis**

## Part I- ML-Powered Data Curation

- Record Linkage, Deduplication, Entity Resolution
- Error Repair and Pattern Enforcement
- Concluding Remarks and Open Issues

# SWOT Analysis (I)



# SWOT Analysis (2)

## STRENGTHS

**1. Leverage diverse signals/  
data with semantically  
rich representations**

**2. Various techniques for  
learning representations**

## EXAMPLES

***To manage multimedia and cross-modal data:***

- Information extraction, Slot Filling, KB Construction [Shin et al., 2015][Wu et al., SIGMOD'18]
- Cross-modal information retrieval
- Complex event summarization
- Cross-modal synthesis of medical images
- Automatic image/video labeling

***Embeddings, multiple views, hierarchical representations***

- Large-scale networks representation [Tang, KDD'17 tutorial]
- Text representation and classification
- Recommendation
- Link prediction
- Visualization

# SWOT Analysis (3)

## STRENGTHS

### 3. Optimization

### 4. Cost reduction

### 5. Good alternative to heuristics

## EXAMPLES

### *To deduplicate, repair, or fuse data:*

- SCARE [Yakout et al., 2013]
- HoloClean [Rekatsinas et al., 2017]
- SLiMFast [Joglekar et al., 2017]

### *To build large-scale knowledge graph:*

- ML-based relation extraction can automatically generate large amount of annotated data and extract features via distant supervision [Mintz et al., 2009] reducing annotating cost

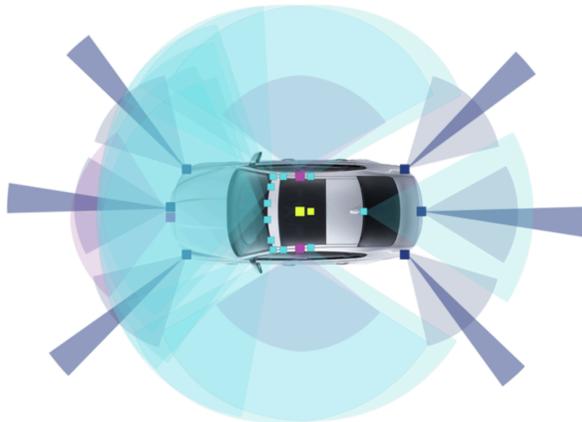
### *To optimize queries & tune DB:*

- Complicated heuristics for estimating selectivity and query plan cost could be replaced and learn dynamically
- Regression-based automatic profiling/tuning (demo Dione [Zacheilas et al., ICDE'18])

# SWOT Analysis (4)

## WEAKNESSES

### I. Obtaining training data is costly



## EXAMPLES

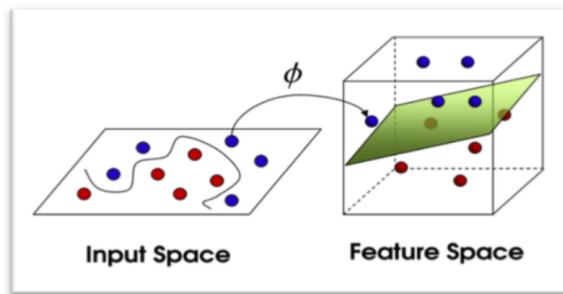
- **Data annotation and preprocessing bottlenecks:** For self-driving cars, 3 million miles of driving data have to be annotated.

Assumptions	Very Conservative estimate
Fleet size	100
Duration of data collection	1 working year / 8h
Volume of data generated by a single car	1TB / h
Data reduction due to preprocessing	0.0005
Research team size	30
Proportion of the team submitting jobs	20%
Target training time	7 days
Number of epochs required for convergence	50
<b>Calculations</b>	
Total raw data volume	203.1 PB
Total data volume after preprocessing	104 TB
Training time on a single DGX-1 Volta system (8 GPUs)	166 days (Inception V3) 113 days (ResNet 50) 21 days (AlexNet)
Number of machines (DGX-1 with Volta GPUs) required to achieve target training time for the team	142 (Inception V3) 97 (ResNet 50) 18 (AlexNet)

# SWOT Analysis (5)

## WEAKNESSES

1. Obtaining training data is costly
2. Finding or coding evidences into features is hard
3. Scaling to Terabytes-size datasets with millions of variables is not easy
4. Model interpretability is limited



## EXAMPLES

- **Data annotation and preprocessing bottlenecks**
  - *Training data generation*: Snorkel [Ratner et al., NIPS'17]
  - *Crowdsourcing automation for labeling training data* suffers from inconsistent quality because expertise is hard to get.
  - *Data integration and curation* are required but generally ad-hoc to get clean training data with well-defined features relevant for the ML models.
- **Deep model training is computationally-expensive.** Techniques for “Learning to learn”, and hyper-parameter optimization can multiply training computation by 5-1000X. [Marcus, Arxiv, 2018]
- **Understand the decisions of Convolutional Neural Network is not straightforward**

Human beings usually cannot fully trust a network, unless it can explain its logic for decisions (NIPS 2017 Interpretable ML Symposium: <http://interpretable.ml/> )

# SWOT Analysis (6)

## OPPORTUNITIES

1. **Revisit DBMS design, techniques and the whole “DBMS abstraction”** [Dittrich, Keynote VLDB’17]

*“ML hardware is at its infancy.”*

[Dean, NIPS 2017]

<http://learningsys.org/nips17/assets/slides/dean-nips17.pdf>

What about ML DBMS?

2. **Apply core-DB technologies to ML workloads**

## EXAMPLES

***To improve components of a DB system:***

- Learned Index structure [Kraska et al., 2017]
- NoDBA project [Sharma et al., 2018] using reinforcement learning to tune a database as a virtual database administrator

***Automated testing of DB applications:***

ETL regression testing [Dzakovic, XLDB’18]

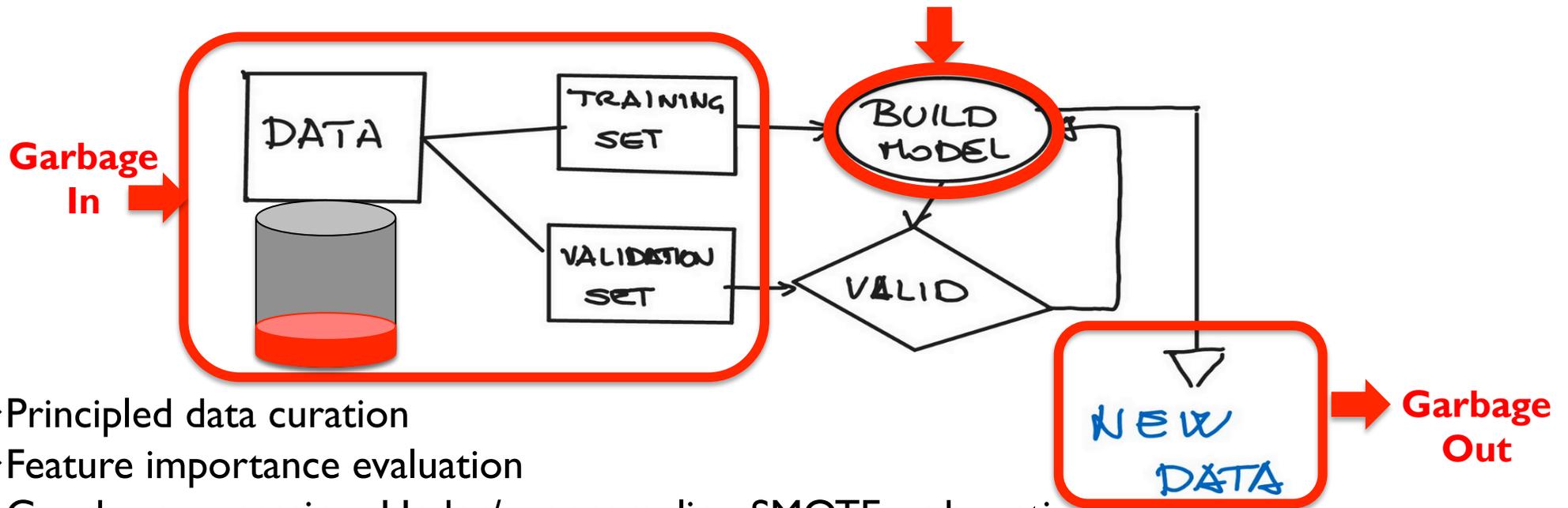
When releasing ETL upgrades, the stakes are high: a single defect can spoil the data in the DB, and the worst-case recovery from a backup would take days

***Principled data curation and preprocessing for ML***

# SWOT Analysis (7)

## THREATS

1. Learning from dirty data is risky
2. Bad feature engineering
3. Minority class problem in unbalanced dataset

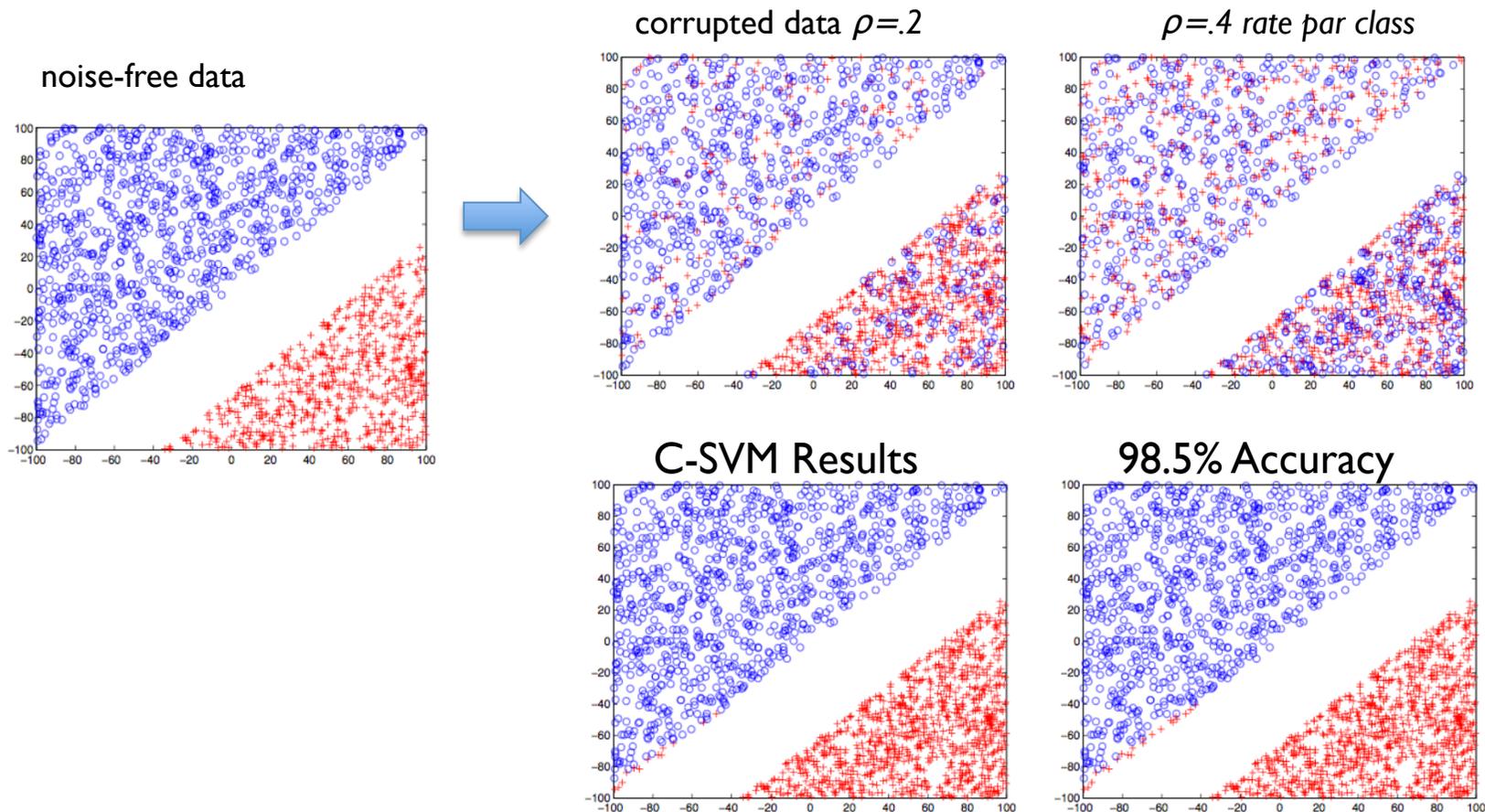


- Principled data curation
- Feature importance evaluation
- Good preprocessing : Under/over-sampling, SMOTE or boosting

# SWOT Analysis (8)

Learning from noisy labels is a hot topic in ML

[Natarajan et al., NIPS'13]



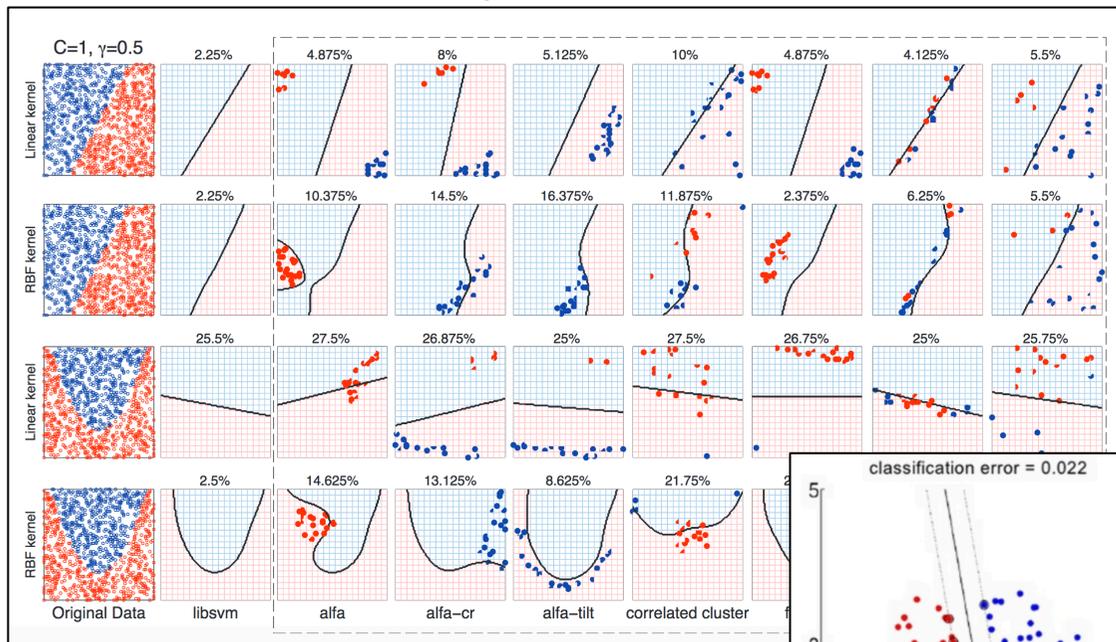
# SWOT Analysis (9)

## THREATS

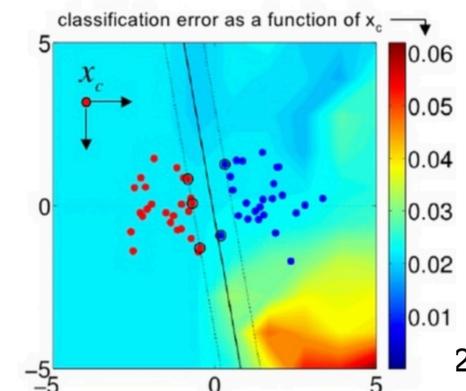
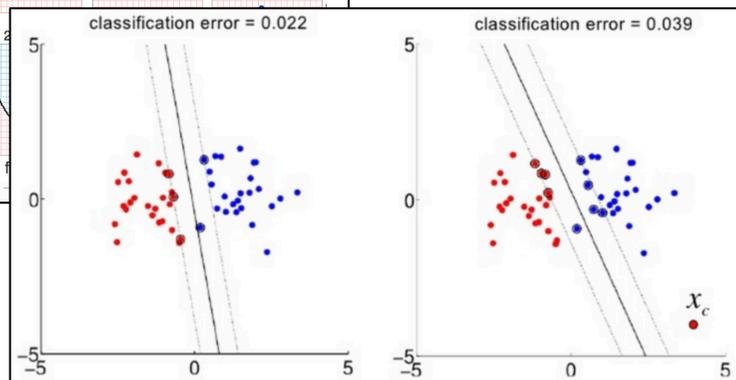
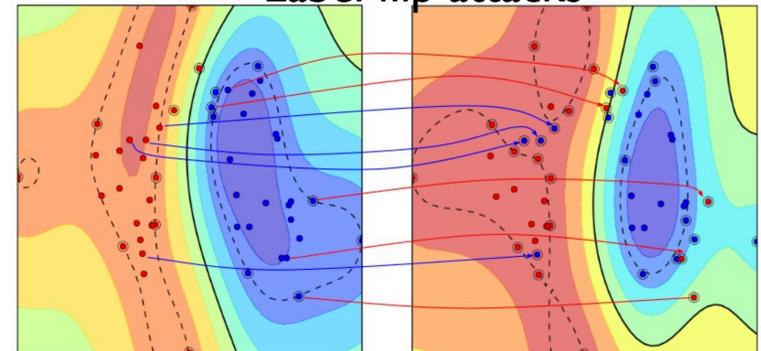
### 4. Adversarial Learning

[Xiao et al., Neurocomputing 2014][Biggio et al., ICML'12]

Poisoning Attacks on SVM



Label flip attacks



# SWOT Analysis: A Summary (10)

## STRENGTHS

1. Leverage diverse signals/data with semantically rich representations
2. Various techniques for learning representations
3. Good alternative to heuristics
4. Optimization with objective functions
5. Reduction of annotating cost

## WEAKNESSES

1. Training data annotation and preprocessing is costly
2. Finding/coding evidences into features is hard
3. Scaling to TB-size datasets with millions of variables is challenging
4. Model interpretability can be limited

## OPPORTUNITIES

1. Revisit design, techniques, and “DBMS abstraction”
2. Apply core-DB technologies to ML workloads

## THREATS

1. Learning from dirty data is risky
2. Bad feature engineering
3. Minority class problem in unbalanced dataset
4. Adversarial Learning

# Outline

## Introduction

- Motivations
- SWOT Analysis

## **ML-Powered Data Curation**

- **Record Linkage, Entity Resolution, Deduplication**
- Error Repair and Pattern Enforcement
- Concluding Remarks and Open Issues

# Record Linkage (RL): Generic Workflow

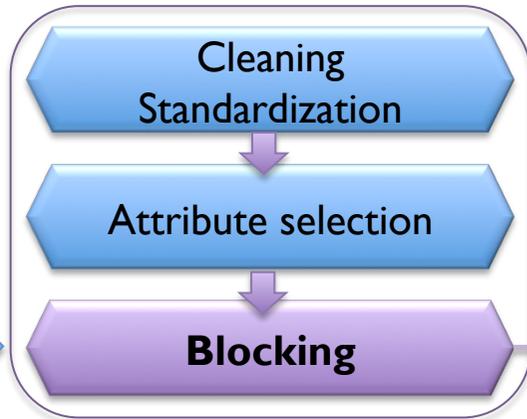
Database R

Name	SSN	Addr
Will Forth	354-564-339	Ada Bd
Jacky Khan	435-232-129	Marple Street
Dom Hack	235-575-689	Main Street
...	...	...

Database S

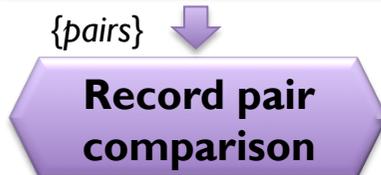
Name	SSN	Addr
Jack Khan	435-223-129	Marple St
Hans Ford	354-564-339	Clover Bd
Tom Hack	235-557-689	Main St
...	...	...

**R X S**

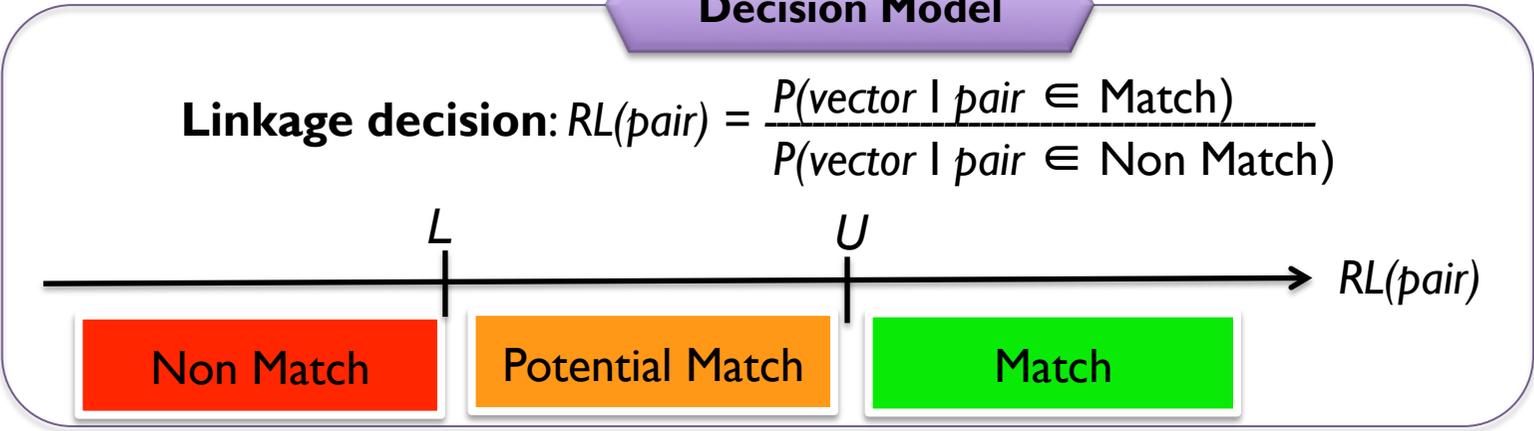


[Fellegi, Sunter, 1969]  
[Christen, 2012]

- Hashing
- Sorted keys
- Sorted NN
- (Multiple) Windowing
- Clustering



- Token-based : N-grams...
- Distance-based: Jaro, Edit, Levenshtein, Soundex
- Domain-dependent



# Pioneer ML-based Deduplication

[Sarawagi, Bhamidipaty, KDD'02]

[Koudas, Srivastava, Sarawagi, Tutorial SIGMOD'06]

Training examples

Customer 1	D
Customer 2	
Customer 1	N
Customer 3	
Customer 4	D
Customer 5	

$f_1$	$f_2$	...	$f_n$	
1.0	0.4	...	0.2	1
0.0	0.1	...	0.3	0
0.3	0.4	...	0.4	1

← Similarity distance functions



Classifier

Unlabeled list

Customer 6
Customer 7
Customer 8
Customer 9
Customer 10
Customer 11

0.0	0.1	...	0.3	?
1.0	0.4	...	0.2	?
0.6	0.2	...	0.5	?
0.7	0.1	...	0.6	?
0.3	0.4	...	0.4	?
0.0	0.1	...	0.1	?



Learnt Rule: All-Ngrams\*0.4  
 + CustomerAddressNgrams\*0.2  
 - 0.3EnrollYearDifference  
 + 1.0\*CustomerNameEditDist  
 + 0.2\*NumberOfAccountsMatch - 3 > 0

Learners:

SVMs: high accuracy with limited data [Christen, 2008]

Decision trees: interpretable, efficient to apply

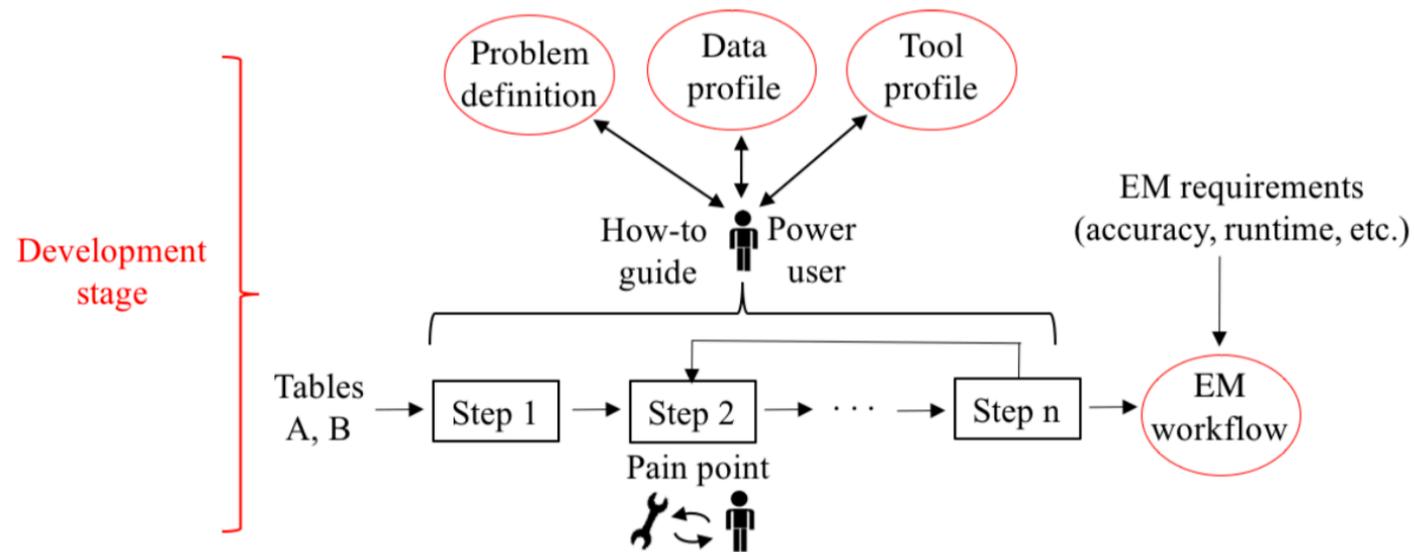
Perceptrons: efficient incremental training

[Bilenko et al., 2005]

# Human-In-The Loop for Entity Matching

[Doan et al., HILDA@SIGMOD'17]

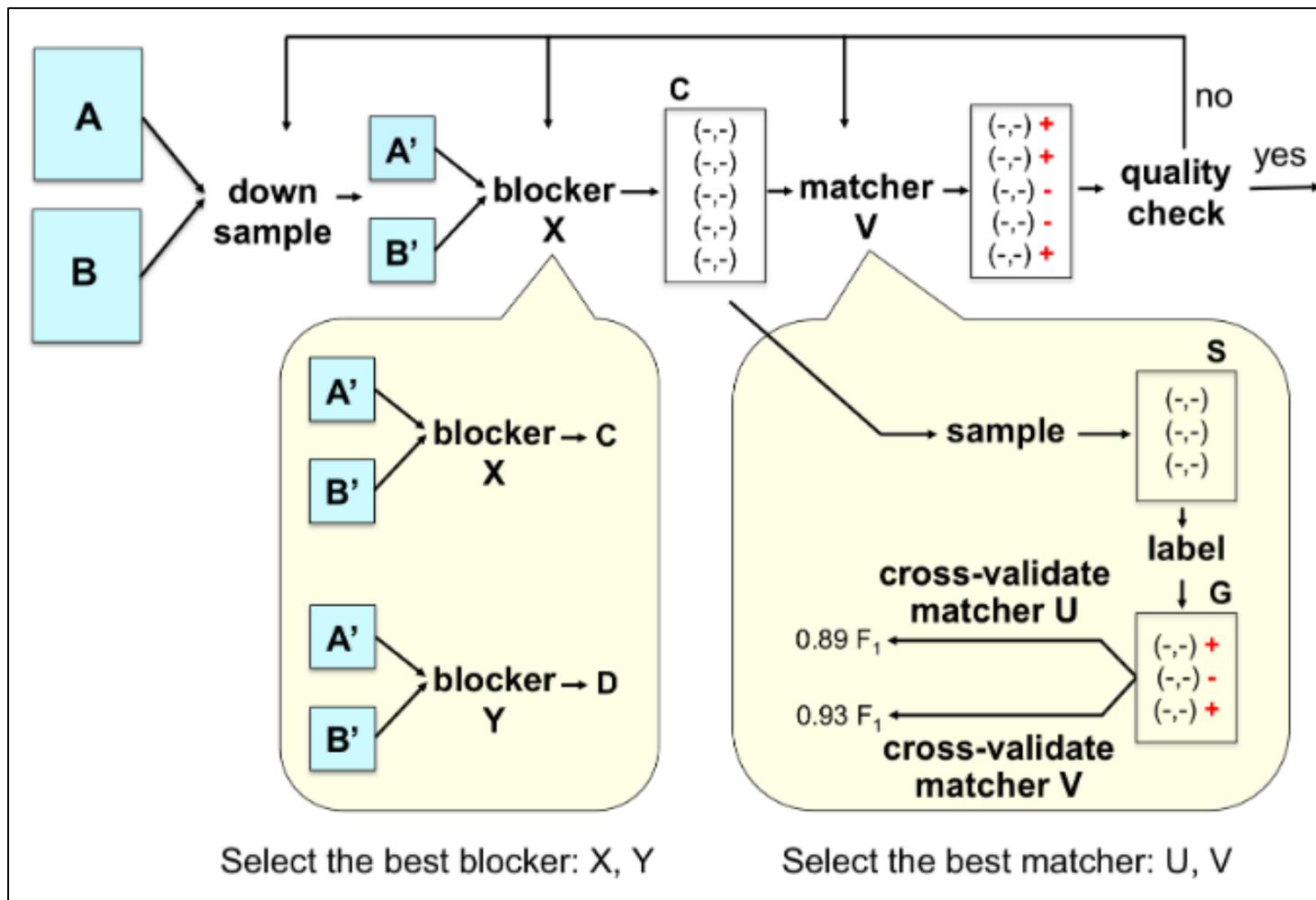
## Magellan project: Lessons learnt for How-to Guide for EM



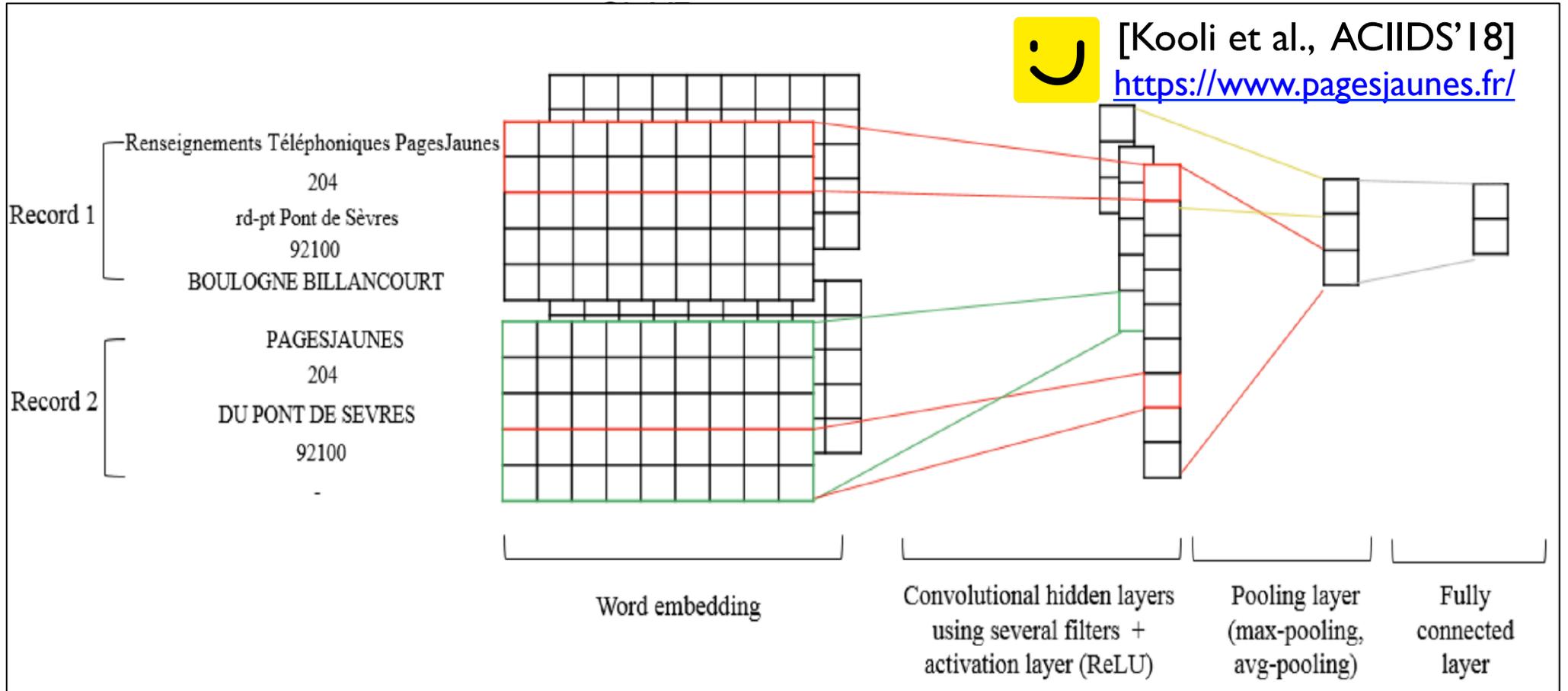
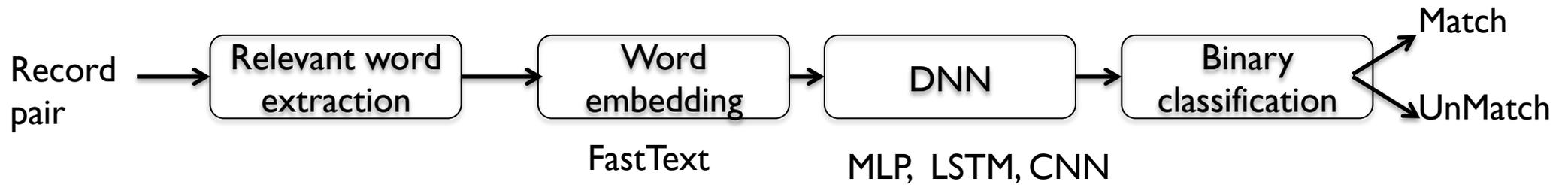
# Human-In-The Loop for Entity Matching

[Doan et al., HILDA@SIGMOD'17]

## Magellan project: Lessons learnt for How-to Guide for EM



# Deep learning for ER



# Outline

## Introduction

- Motivations
- SWOT Analysis

## **ML-Powered Data Curation**

- Record Linkage, Entity Resolution, Deduplication,
- **Error Repair and Pattern Enforcement**
- Concluding Remarks and Open Issues

# ML-Based Repairing

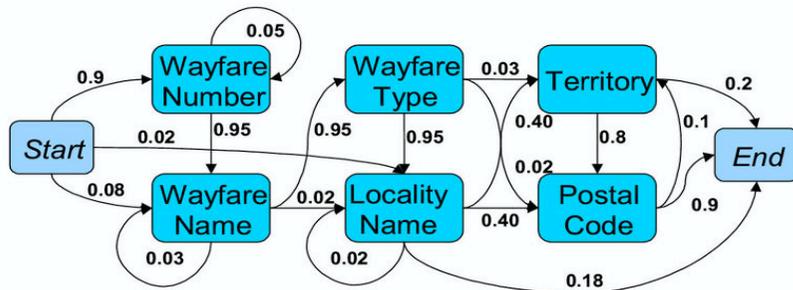
Semi-automatic techniques for:

- **Pattern enforcement**
  - Syntactic patterns (date formatting)
  - Semantic patterns (name/address)
- **Value update** to satisfy a set of rules, constraints, FDs, CFDs, Denial Constraints (DCs), Matching Dependencies (MDs) with minimal number of changes. [Ilyas, Chu, 2015]
- **Value imputation** with statistical methods to replace outliers or missing values
- **Data fusion**

# Febrl: Data standardization with HMM

[Churches et al., 2002]  
[Christen et al., 2002]

## HMM for Address Standardization



	To state							
From state	Start	Wayfare Number	Wayfare Name	Wayfare Type	Locality Name	Territory	Postal Code	End
Start	0	0.9	0.08	0	0.02	0	0	0
Wayfare Number	0	0.05	0.95	0	0	0	0	0
Wayfare Name	0	0	0.03	0.95	0.02	0	0	0
Wayfare Type	0	0	0	0	0.95	0.03	0.02	0
Locality name								
Territory								
Postal Code								
End								

	State								
Observation Symbol	Start	Wayfare Number	Wayfare Name	Wayfare Type	Locality Name	Territory	Postal Code	End	
NU	-	0.9	0.01	0.01	0.01	0.01	0.01	0.1	-
WN	-	0.01	0.5	0.01	0.1	0.01	0.01	0.01	-
WT	-	0.01	0.01	0.92	0.01	0.01	0.01	0.01	-
LN	-	0.01	0.1	0.01	0.8	0.01	0.01	0.01	-
TR	-	0.01	0.07	0.01	0.01	0.94	0.01	0.01	-
PC	-	0.04	0.01	0.01	0.01	0.01	0.85	0.01	-
UN	-	0.02	0.31	0.03	0.06	0.01	0.01	0.01	-

**Selection of representative training data**  
"17 Epping St Smithfield New South Wales 2987"

**Tokenization based on Look-up Tables**  
['17', 'epping', 'street', 'smithfield', 'nsw', '2987']

**Tagging**  
['NU', 'LN', 'WT', 'LN', 'TR', 'PC']  
number-locality name-wayfare type-locality name-territory-postal code

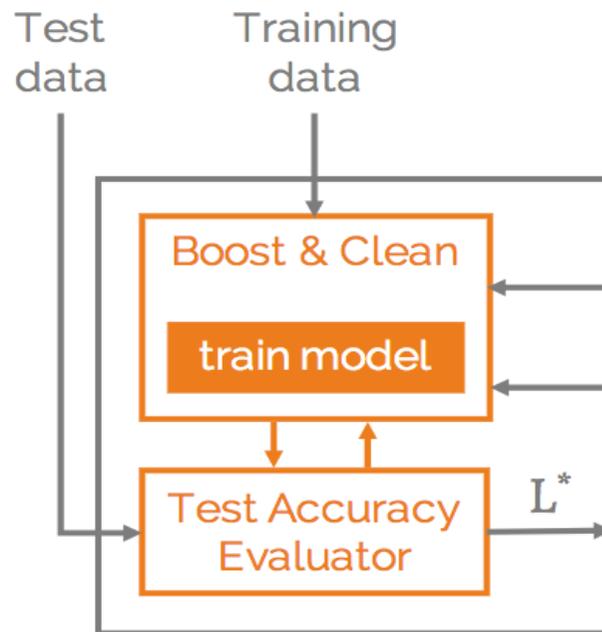
**Frequency-based Maximum Likelihood Estimates**  
 $8^6 = 262,144$  possible combinations of hidden states

- $Start \rightarrow$  Wayfare Name (NU)  $\rightarrow$  Locality Name (LN)  $\rightarrow$  Postal Code (WT)  $\rightarrow$  Territory (LN)  $\rightarrow$  Postal Code (TR)  $\rightarrow$  Territory (PC)  $\rightarrow End$   
 $0.08 \times 0.01 \times 0.02 \times 0.8 \times 0.4 \times 0.01 \times 0.1 \times 0.01 \times 0.8 \times 0.01 \times 0.1 \times 0.01 \times 0.2 = 8.19 \times 10^{-17}$
- $Start \rightarrow$  Wayfare Number (NU)  $\rightarrow$  Wayfare Name (LN)  $\rightarrow$  Wayfare Type (WT)  $\rightarrow$  Locality (LN)  $\rightarrow$  Territory (TR)  $\rightarrow$  Postal Code (PC)  $\rightarrow End$   
 $0.9 \times 0.9 \times 0.95 \times 0.1 \times 0.95 \times 0.92 \times 0.95 \times 0.8 \times 0.4 \times 0.94 \times 0.8 \times 0.85 \times 0.9 = 1.18 \times 10^{-2}$

# BoostClean

[Krishnan et al., 2017]

BoostClean selects an ensemble of methods (statistical and logic rules) for error detection and for repair combinations using statistical boosting.



## Algorithm 2: Boost-and-Clean Algorithm

**Data:**  $(X, Y)$

- 1 Initialize  $W_i^{(1)} = \frac{1}{N}$
- 2  $\mathcal{L}$  generates a set of classifiers  $\mathcal{C}\{C^{(0)}, C^{(1)}, \dots, C^{(k)}\}$  where  $C^{(0)}$  is the base classifier and  $C^{(1)}, \dots, C^{(k)}$  are derived from the cleaning operations.
- 3 **for**  $t \in [1, T]$  **do**
- 4      $C_t = \text{Find } C_t \in \mathcal{C}$  that maximizes the weighted accuracy on the test set.  $\epsilon_t = \text{Calculate weighted classification error on the test set}$   $\alpha_t = \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$   
    $W_i^{(t+1)} \propto W_i^{(t)} e^{-\alpha_t y_i C_t(x_i)}$ : down-weight correct predictions, up-weight incorrectly predictions.
- 5 **return**  $C(x) = \text{sign}\left(\sum_t^T \alpha_t C_t(x)\right)$

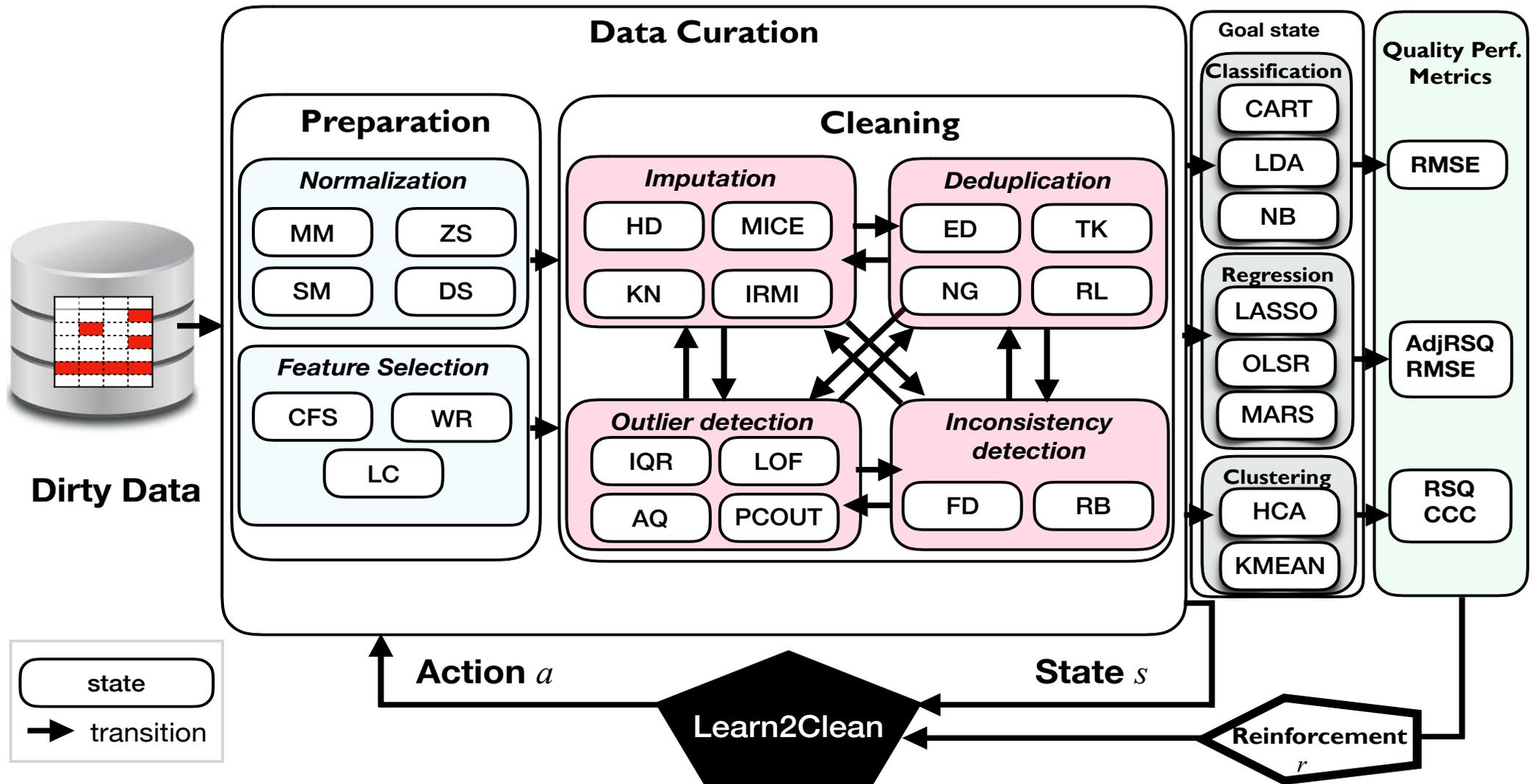
# A Condensed View

<b>Repair System</b>	<b>ML Approach</b>	<b>Goal</b>
<b>Febri</b> [Churches et al., 2002]	HMM and MLE	Standardizing loosely structured texts (e.g., name/address) based on the probabilistic model learnt from training data
<b>SCARE</b> [Yakout, Berti-Equille, Elmagarmid, SIGMOD'13]	Multiple ML models used to capture data dependencies across multiple data partitions	Find the candidate repair that maximizes the likelihood repair benefit under a cost threshold of the update
<b>Continuous Cleaning</b> [Volkovs et al., ICDE'14]	Logistic classifiers	Learning from past user repair preferences to recommend next more accurate repairs
<b>Lens</b> [Yang et al., VLDB'15]	Various ML models encoded in Domain Constraints	Declarative on-Demand ETL with prioritized curation tasks based on probabilistic query processing and PC-Tables
<b>HoloClean</b> [Rekatsinas et al., VLDB 2017]	Probabilistic inference on factor graphs with SGD and Gibbs sampling	Mixing statistical and logical rules, DCs, MDs, etc. to infer candidate repairs in a scalable way with domain pruning and constraint relaxation
<b>BoostClean</b> [Krishnan et al., 2017]	AdaBoost	Mixing statistical and logical rules, domain constraints for detection and repair combinations to maximize the predictive accuracy over test data

# Reinforcement learning for data cleaning

Learn2Clean: Optimizing the Sequence of Tasks for Data Preparation

[The Web Conference 2019]



# Outline

## Introduction

- Motivations
- SWOT Analysis

## **ML-Powered Data Curation**

- Record Linkage, Deduplication, Entity Resolution
- Error Repair and Pattern Enforcement
- Data and Knowledge Fusion
- **Concluding Remarks and Open Issues**

# Concluding Remarks

- ML provides a principled framework and efficient tools for optimizing many Data Management tasks
- ML crucially needs principled data curation
- However, some tasks require **Humans in the loop**
- There are many opportunities for:
  - Cool ML applications to data management
  - Revisiting DB technology **with** and **for** ML
  - Managing and orchestrating human/machine resources

# Open Issues

- **Usability:**
  - To consider Humans as resources
  - To be understood, interpreted, and trusted by Humans
  - To ease/self-adapt the design, tuning, and use
- **Efficiency:**
  - Runtime
  - Incremental
- **Accuracy:**
  - Reduce impact of dirty data
  - Augmenting the training set
  - Ensembling

# Usability (I): Humans as Resources

## Challenge I: Adjusting “Human-in-the-Loop”

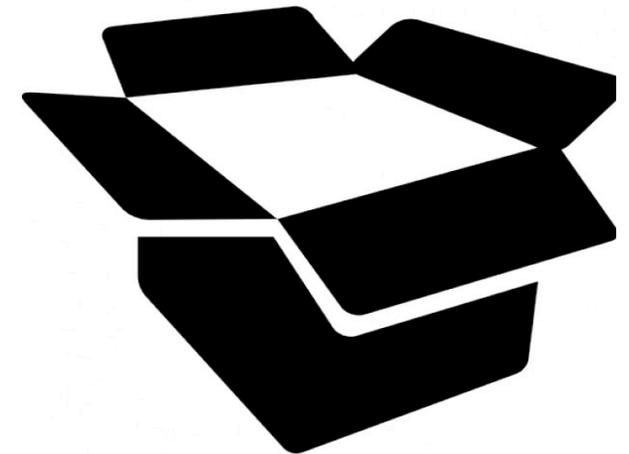
- Seamless integration of humans as resources for ML-powered DM
- “Taskify” and minimize the amount of interactions with the users while, at the same time, maximize the potential “ML benefit” for selecting/cleaning/labeling training data and other data management tasks
- **Current efforts: Crowdsourcing and active learning**
  - Data cleaning with oracle crowds [Bergman et al., SIGMOD’15]
  - Entity resolution: CrowdER [Wang et al., VLDB’12], Corleone [Gokhale, et al., SIGMOD’14]
  - Data fusion and truth inference [Zheng et al., VLDB’17]
- **Direction:**
  - Adaptive and quality-driven orchestration of Humans and Tools for ML-powered DM



# Usability (2): Building trust

## Challenge 2: Open the “Black-Box” and customize it

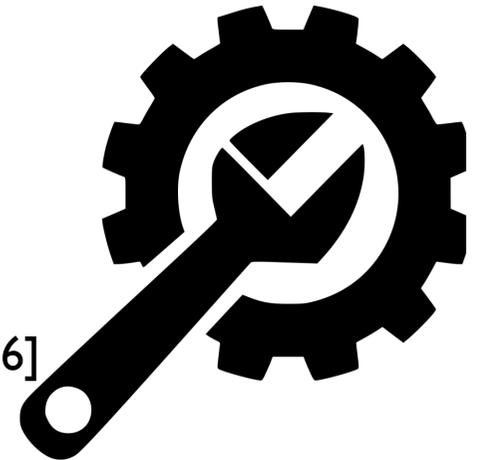
- Improve the interpretability of ML-based decisions
- Build the trust: ML-based decisions should be interpretable, explainable, reproducible to be trusted
- Adapt ML-based DM to on-demand, incremental, progressive tasks
- **Current efforts:**
  - Trusted Machine Learning [Ghosh et al., AAAI'17]
  - Model-Agnostic Explanations [Ribeiro et al., KDD'16]
  - On-demand ETL [Yang et al., VLDB'15]
  - ActiveClean [Krishnan et al., VLDB'16]
  - Continuous cleaning for considering incremental changes to the data and to the constraints [Volkovs et al., ICDE'14]
- **Directions:**
  - Causality and explanations in ML-based DM and their effective representation
  - Reversibility and repeatability
  - Data privacy/security: What if adversarial learning is applied ?



# Usability (3) : Easy to build, tune, and test

## Challenge 3: Engineering ML-based DM applications

- Model building and feature selection
- Model interoperability and model selection
- **Current efforts:**
  - Systematizing/optimizing model selection  
[Kumar, Boehm, Yang, SIGMOD'17 Tutorial],  
MSMS [Kumar et al., SIGMODRec'15], Zombie [Anderson et al., 2016]
  - Declarative ML tasks
  - Interactive model building: Ava [John et al., CIDR'17], Vizdom [Crotty et al., VLDB'15]
  - Meta-learning, bandit techniques
  - PMML, ONNX, PFA for model interoperability
- **Directions:**
  - Analysis of dependability of models
  - Model debugging, versioning, and management (e.g., for large models)
  - Managing ML model provenance and elicitation
  - Transfer pre-trained models from task-/domain-agnostic to \*-specific DM



# Efficiency

- **Challenge 4: Incremental ML application to DM**
  - When we have more training data or refresh/delete some data (obsolete), shall we retrain ML model from scratch? Can we do incremental training/learning? For what cost/trade-off?
- **Challenge 5: Runtime ML-based DM**
  - Could we orchestrate and optimize data annotation and preprocessing tasks? Design cost models, candidate plans?
  - To what extent could we use transfer learning to reduce training data collection/preprocessing cost ?



# Accuracy (I)

- **Challenge 6: Reduce the impact of dirty data**

Glitch types and their distributions can be very different in the datasets used for training, testing, and validation and they affect accuracy of ML models in different ways:

- How could we capture the good, the bad and the ugly combinations?
- Should we robustify the ML algorithms or/and the data curation? Would both be inevitably better/necessary?
- **Find optimal data cleaning strategies for a given ML-based DM application**
  - Can we predict the  $\pm\delta$  in ML accuracy that a given data curation strategy brings to the model?



# Accuracy (2)

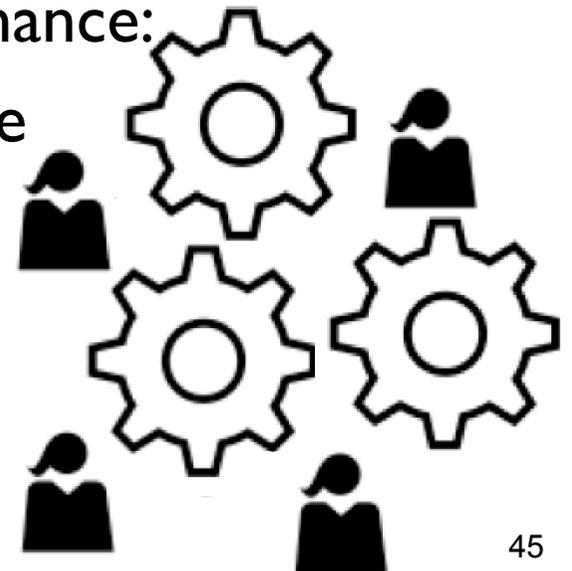
- **Challenge 7: Synthetic training data generation**

Copy/Transform existing labeled data to augment the training set  
[Ratner et al., NIPS'17]

- **Challenge 8: Model/Feature recommendation and ensembling**

Many ML models can be parameterized, applied and combined in different ways leading to various quality performance:

- Could we define a predictive scoring of the models and their ensembles ?
- Would ensembling be (inevitably) better?



**Thanks!**



# References - Part I (I)

- [Anderson et al., 2016] <http://www.vldb.org/pvldb/vol7/p1657-anderson.pdf>
- [Arasu et al., SIGMOD'10] <https://dl.acm.org/citation.cfm?id=1807252>
- [Assadi, Milo, Novgorodov, WebDB'18] <http://slavanov.com/research/webdb18.pdf>
- [Bahmani et al., SUM'15] <https://arxiv.org/pdf/1602.02334.pdf>
- [Battacharya, Getoor, TKDD'07] <https://dl.acm.org/citation.cfm?id=1217304>
- [Bellare et al., KDD'12] <http://ilpubs.stanford.edu:8090/1036/1/main.pdf>
- [Bergman et al., SIGMOD 2015] <http://www.vldb.org/pvldb/vol8/p1900-bergman.pdf>
- [Berti-Equille, Encyclopedia 2018] Encyclopedia of Big Data Technologies, Springer (To Appear), 2018
- [Biggio et al., ICML'12] <https://icml.cc/Conferences/2012/papers/880.pdf>
- [Bilenko et al., ICDM'06] <http://ieeexplore.ieee.org/document/4053037/>
- [Bilenko, Mooney, KDD'03] <https://dl.acm.org/citation.cfm?id=956759>
- [Chaudhuri et al., ICDE'05] <http://ieeexplore.ieee.org/document/1410199/>
- [Chaudhuri et al., VLDB'07] <http://www.vldb.org/conf/2007/papers/research/p327-chaudhuri.pdf>
- [Chen et al., SIGMOD'09] <https://dl.acm.org/citation.cfm?id=1559869>
- [Christen et al., 2002] <http://users.cecs.anu.edu.au/~christen/publications/adm2002-cleaning.pdf>
- [Christen, 2012] <http://ieeexplore.ieee.org/document/5887335/>
- [Churches et al., 2002] <http://www.biomedcentral.com/1472-6947/2/9/>
- [Crotty et al., VLDB'15] <http://www.vldb.org/pvldb/vol8/p2024-crotty.pdf>
- [Dean, NIPS 2017] <http://learningsys.org/nips17/assets/slides/dean-nips17.pdf>
- [Doan et al., HILDA@SIGMOD'17] <http://pages.cs.wisc.edu/~anhai/papers17/hil-in-em-hilda17.pdf>
- [Dzakovic, XLDB'18] <https://conf.slac.stanford.edu/xldb2018/event-information/lightning-talks>
- [Ebraheem et al., Arxiv 2017] <https://arxiv.org/pdf/1710.00597.pdf>
- [Fellegi, Sunter, 1969] A theory for record linkage. J. Am. Stat. Assoc. 1969;64(328):1183–210.
- [Fisher et al., KDD'15] <https://dl.acm.org/citation.cfm?id=2783396>
- [Getoor, Machanavajjhala, Tutorial VLDB'12] [http://legacydirs.umiacs.umd.edu/~getoor/Tutorials/ER\\_VLDB2012.pdf](http://legacydirs.umiacs.umd.edu/~getoor/Tutorials/ER_VLDB2012.pdf)
- [Gokhale et al., SIGMOD'14] <https://dl.acm.org/citation.cfm?id=2588576>
- [Gosh et al., AAI'17] <https://aaai.org/ocs/index.php/WI/AAAIWI7/paper/download/15206/14765>

# References - Part I (2)

- [Gupta, Sarawagi, VLDB'09] <https://dl.acm.org/citation.cfm?id=1687627.1687661>
- [Hall, 1992] Mathematical Techniques in Multisensor Data Fusion, ArtechHouse, 1992
- [Hassanzadeh et al., PVLDB'09] <http://www.vldb.org/pvldb/2/vldb09-1025.pdf>
- [Hu et al, 2017] <http://users.cecs.anu.edu.au/~u5170295/papers/pakdd-hu-2017.pdf>
- [Ilyas, Chu, 2015] <https://cs.uwaterloo.ca/~ilyas/papers/IlyasFnTDB2015.pdf>
- [John et al., CIDR'17] <http://pages.cs.wisc.edu/~jignesh/publ/Ava.pdf>
- [Joglekar, et al., SIGMOD'17] <https://dl.acm.org/citation.cfm?id=3035951>
- [Kooli et al., ACIIDS'18] [https://link.springer.com/chapter/10.1007%2F978-3-319-75420-8\\_1](https://link.springer.com/chapter/10.1007%2F978-3-319-75420-8_1)
- [Köpcke et al., VLDB'10] <http://www.vldb.org/pvldb/vldb2010/papers/E04.pdf>
- [Koudas, Srivastava, Sarawagi, Tutorial SIGMOD'06] <http://www.cs.toronto.edu/~koudas/docs/aj.pdf>
- [Kraska et al. 2017] <https://arxiv.org/abs/1712.01208>
- [Krishnan et al., VLDB'16] <http://www.vldb.org/pvldb/vol9/p948-krishnan.pdf>
- [Krishnan et al., 2017] <https://arxiv.org/pdf/1711.01299.pdf>
- [Kumar et al., SIGMODRec'15] [https://adalabucsd.github.io/papers/2015\\_MSMS\\_SIGMODRecord.pdf](https://adalabucsd.github.io/papers/2015_MSMS_SIGMODRecord.pdf)
- [Kumar, Boehm, Yang, SIGMOD'17 Tutorial] [https://adalabucsd.github.io/papers/2017\\_Tutorial\\_SIGMOD.pdf](https://adalabucsd.github.io/papers/2017_Tutorial_SIGMOD.pdf)
- [Luyang et al., Sensors'17] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5335931/>
- [Marchand, Rubinstein, VLDB'17] <http://www.vldb.org/pvldb/vol10/p1322-rubinstein.pdf>
- [Marcus, Arxiv, 2018] <https://arxiv.org/ftp/arxiv/papers/1801/1801.00631.pdf>
- [Mintz et al., 2009] <https://dl.acm.org/citation.cfm?id=1690287>
- [Natarajan et al., NIPS'13] <https://papers.nips.cc/paper/5073-learning-with-noisy-labels.pdf>
- [Papadakis et al., TKDE 2013] <http://ieeexplore.ieee.org/document/6255742/>
- [Papadakis, Palpanas, Tutorial ICDE'16] <http://www.mi.parisdescartes.fr/~themisp/publications/tutorialicde16.pdf>
- [Polyzotis et al., SIGMOD'17] <https://dl.acm.org/citation.cfm?id=3035918.3054782>
- [Qian et al., CIKM'17] <https://dl.acm.org/citation.cfm?id=3132949>

# References - Part I (3)

- [Ratner et al., NIPS'17] <https://arxiv.org/pdf/1709.01643.pdf>
- [Rekatsinas et al., VLDB'17] <http://www.vldb.org/pvldb/vol10/p1190-rekatsinas.pdf>
- [Ribeiro et al., KDD'16] <http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf>
- [Sarawagi, Bhamidipaty, KDD'02] <https://dl.acm.org/citation.cfm?id=775087>
- [Sharma et al. 2018] <https://arxiv.org/abs/1801.05643>
- [Shin et al., 2015] <http://www.vldb.org/pvldb/vol8/p1310-shin.pdf>
- [Singla, Domingos, PKDD'05] <http://alchemy.cs.washington.edu/papers/pdfs/singla-domingos05b.pdf>
- [Tang, KDD'17 tutorial] <https://sites.google.com/site/pkujiantang/home/kdd17-tutorial>
- [Tejada et al. KDD'02] <https://dl.acm.org/citation.cfm?id=775099>
- [Vesdapunt et al., VLDB'14] <http://www.vldb.org/pvldb/vol7/p1071-vesdapunt.pdf>
- [Volkovs et al., ICDE'14] [http://www.cs.toronto.edu/~mvolkovs/icde14\\_data\\_cleaning.pdf](http://www.cs.toronto.edu/~mvolkovs/icde14_data_cleaning.pdf)
- [Wang et al., SIGMOD'13'] <https://dl.acm.org/citation.cfm?id=2465280>
- [Wang et al., VLDB'12] [http://vldb.org/pvldb/vol5/p1483\\_jiannanwang\\_vldb2012.pdf](http://vldb.org/pvldb/vol5/p1483_jiannanwang_vldb2012.pdf)
- [Wu et al. SIGMOD'18] <https://arxiv.org/pdf/1703.05028.pdf>
- [Xiao et al., Neurocomputing 2014] <https://www.sciencedirect.com/science/article/pii/S0925231215001198>
- [Yakout, Berti-Equille, Elmagarmid SIGMOD'13] <https://dl.acm.org/citation.cfm?id=2463706>
- [Yang et al., VLDB'15] <http://www.vldb.org/pvldb/vol8/p1578-yang.pdf>
- [Zheng et al., VLDB'17] <http://www.vldb.org/pvldb/vol10/p541-zheng.pdf>