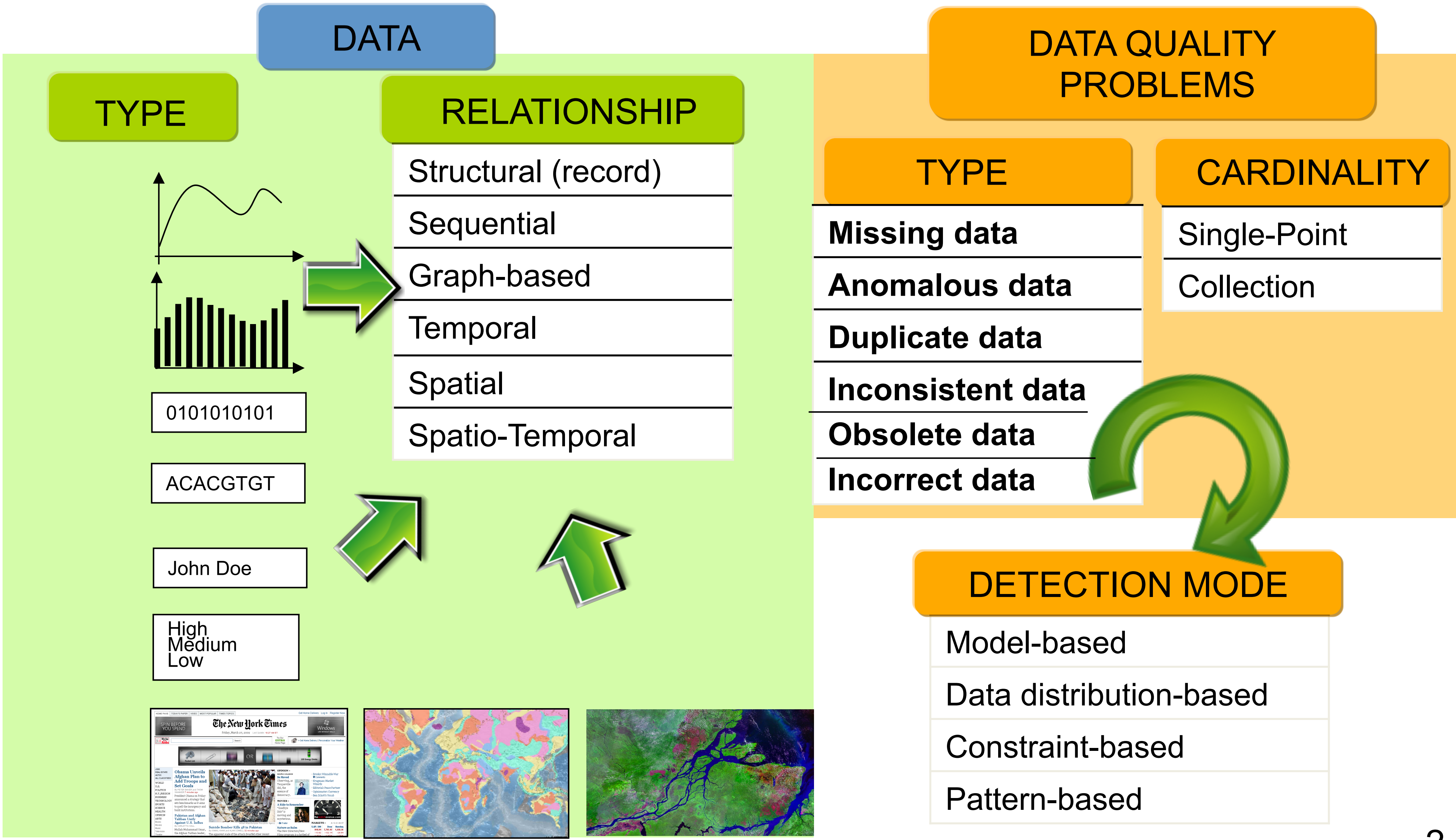
Machine Learning-Based Data Cleaning : Current Solutions and Challenges

Laure Berti-Equille

IRD Montpellier
Aix-Marseille University, CNRS, LIS, DIAMS
France
laure.berti@ird.fr
http://pageperso.lif.univ-mrs.fr/~laure.berti/

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

Data Quality Problems



Example I

Relational data : CiDE.21 committee

Misfielded Value

Representation

Nom	Etablissement	Ville	Tel
Prof. B. JACQUEMIN	Univ. Lille GERiiCO	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

Duplicates

Typos

Inconsistencies

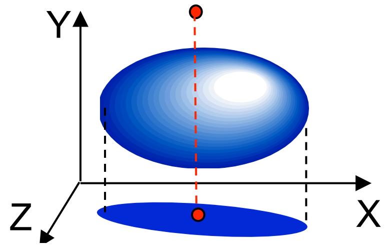
Incorrect Values

Obsolete Value

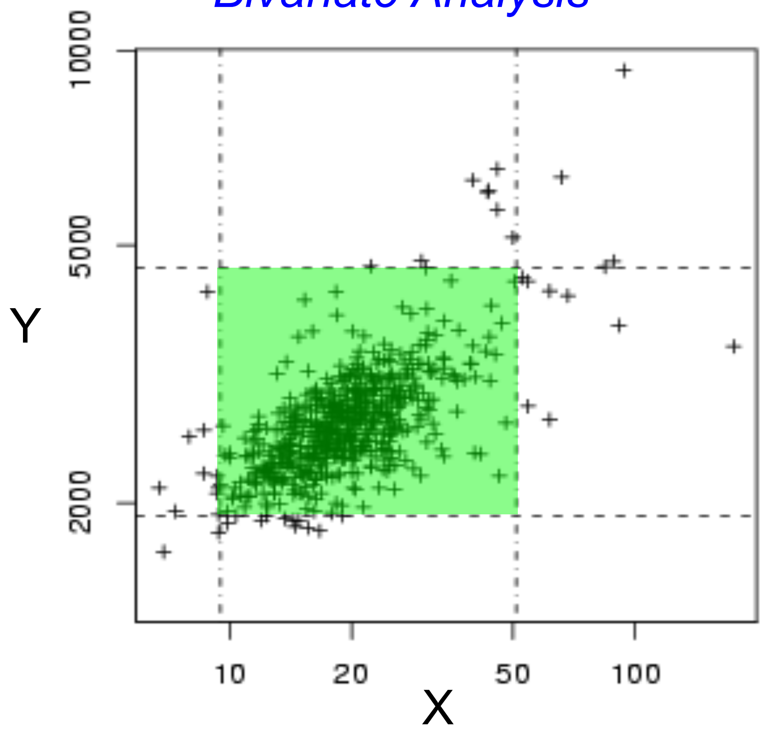
Missing Values

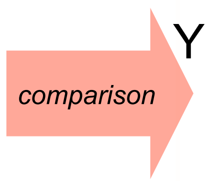
Example 2

Outliers

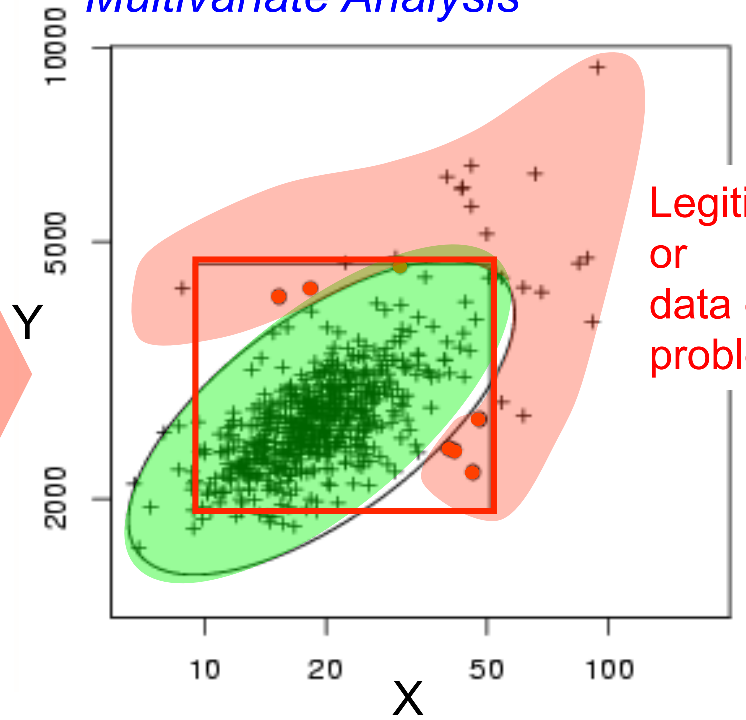


Bivariate Analysis





Multivariate Analysis



Legitimate outliers
or
data quality
problems?

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

Rejection area based on:
Mahalanobis_dist(cov(X,Y)) > $\chi^2(.98,2)$

Example 3

Disguised missing data

Some are obvious...

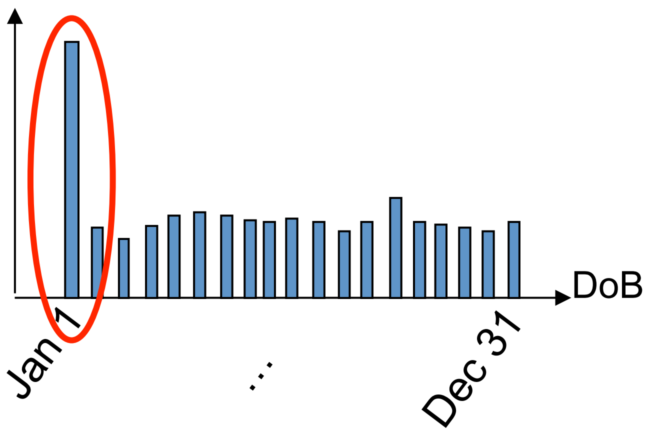
Detectable with syntactical or domain constraints

Phone number: 999-999-9999

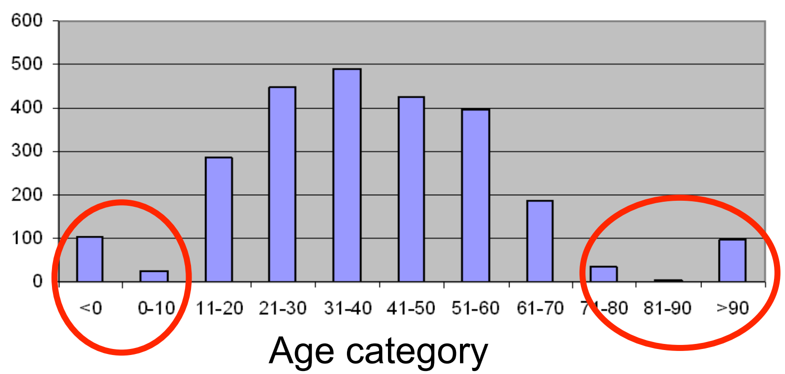
Others are not….

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

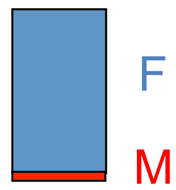
Histogram of DoBs per day of the year



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

AFP February 28, 2015 2:42 PM

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

© REUTERS/ BENOIT TESSIER

French TV Denies Reports of Bouygues Conglomerate CEO's Death

Example 5

Rumors: Celebrity Death Hoaxes

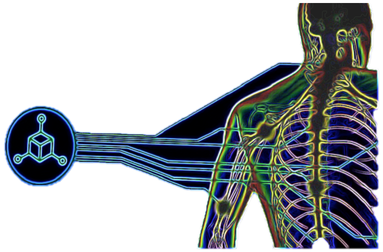
ML Revolutionizes Industry

Security and Surveillance

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

eHealth

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



Manufacturing

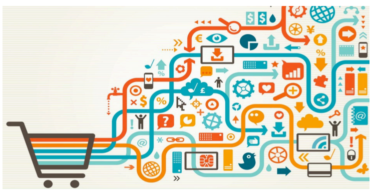
optimizing fab operations, automating quality testing, inventory, asset, and supply chain management, predictive maintenance, etc.

Machine Learning Applications

Autonomous vehicles

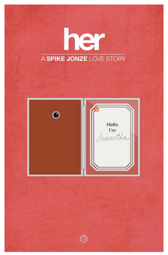
Smart eCommerce

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.

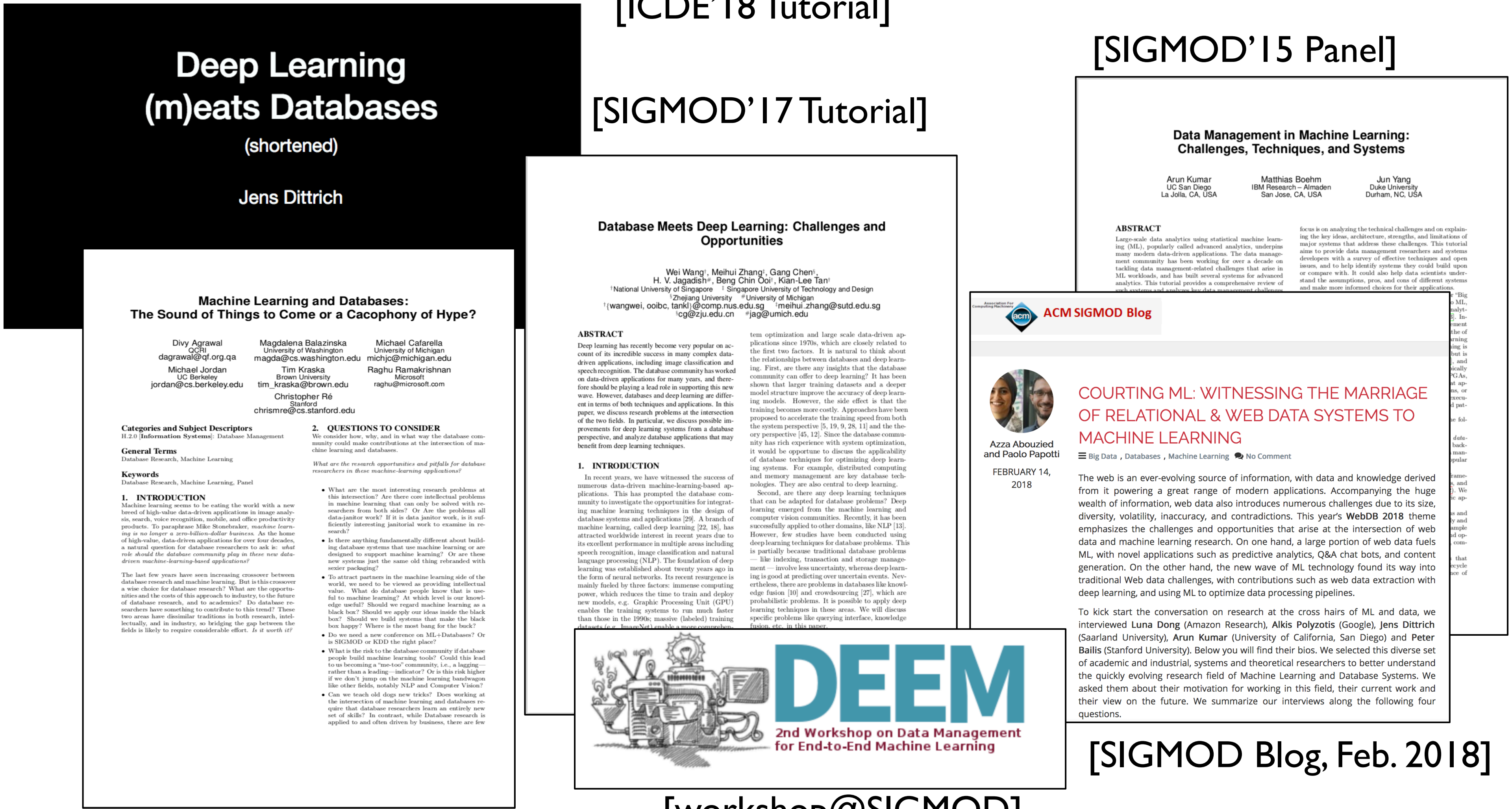
Hot Topic for DB community

[VLDB'17 Keynote]

[ICDE'18 Tutorial]

[SIGMOD'15 Panel]

[SIGMOD'17 Tutorial]



[SIGMOD Blog, Feb. 2018]

[workshop@SIGMOD]

[SIGMOD Record 2016]

Introduction : DB perpective

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

- Information and structure extraction
- Data curation
- Data integration
- Querying & DB administration
- Privacy preservation
- Data storage

Our focus

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 (1)

STRENGTHS	**WEAKNESSES**
OPPORTUNITIES	**THREATS**

SWOT Analysis (2)

STRENGTHS

EXAMPLES

1. Leverage diverse signals/ data with semantically rich representations

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

2. Various techniques for learning representations

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 [Jogleakr 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

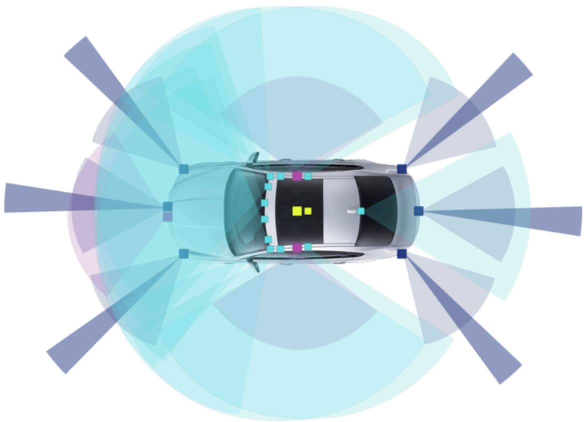
To optimize queries & tune DB:

- 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

1. 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
Calculations	
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)

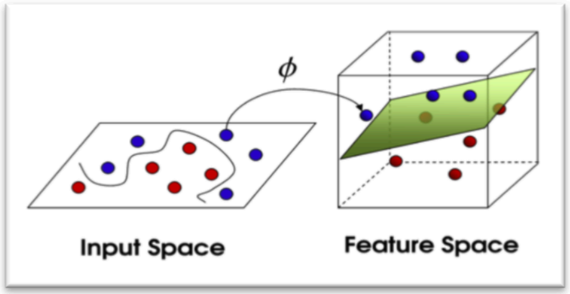
https://devblogs.nvidia.com/training-self-driving-vehicles-challenge-scale

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:

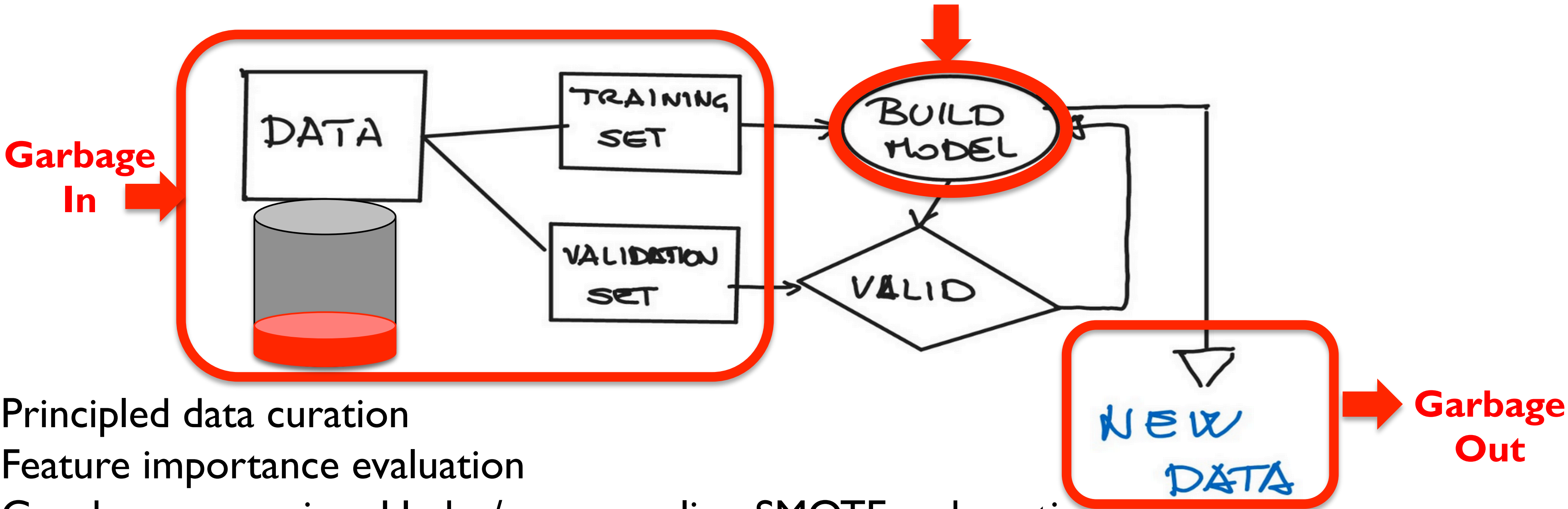
ETL regression testing [Dzakovic, XLDB'18]
When releasing ETL upgrades, the stakes are high: a single defect can spoil the data in the DB, and the worst-case recovery from a backup would take days

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**



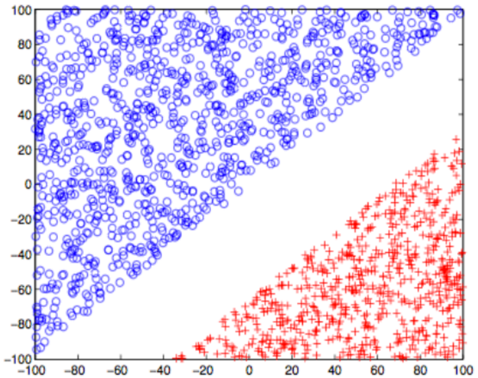
➔Principled data curation
➔Feature importance evaluation
➔Good preprocessing : Under/over-sampling, SMOTE or boosting

SWOT Analysis (8)

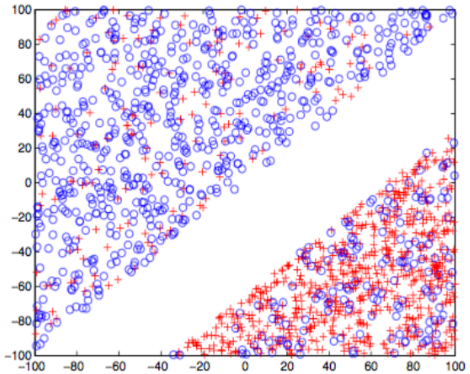
Learning from noisy labels is a hot topic in ML

[Natarajan et al., NIPS'13]

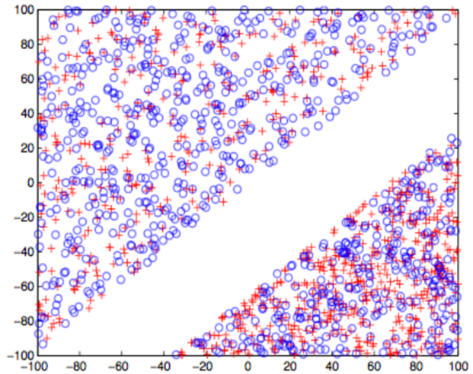
noise-free data



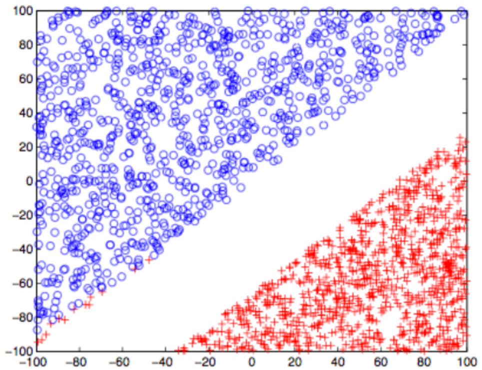
corrupted data ρ=.2



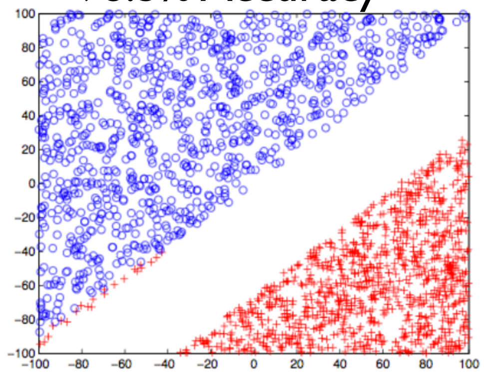
ρ=.4 rate par class



C-SVM Results



98.5% Accuracy



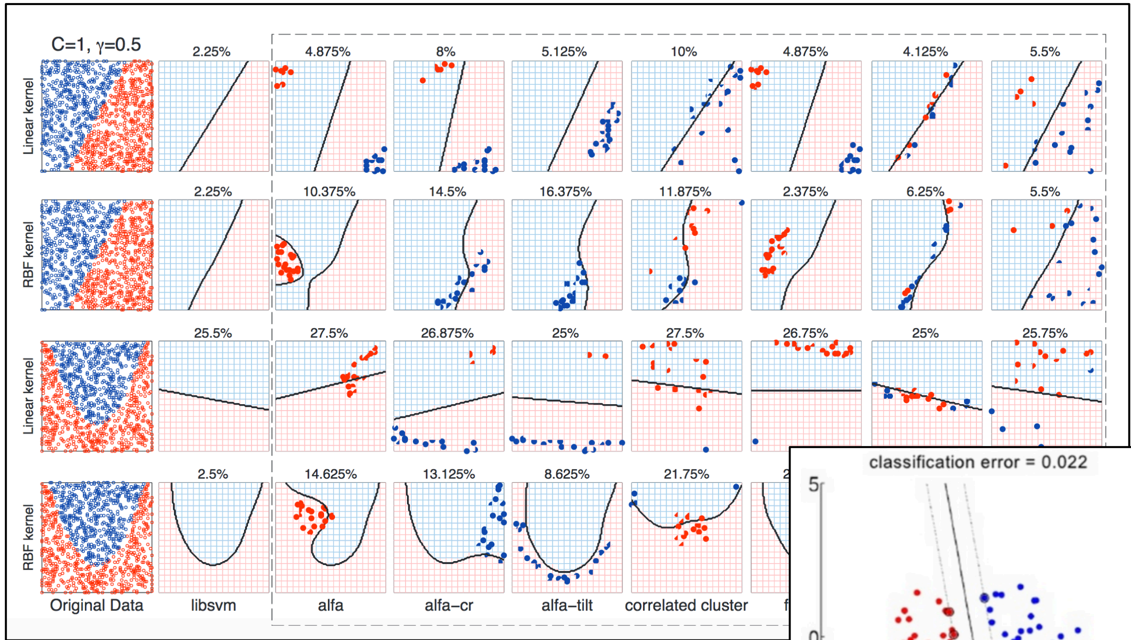
SWOT Analysis (9)

THREATS

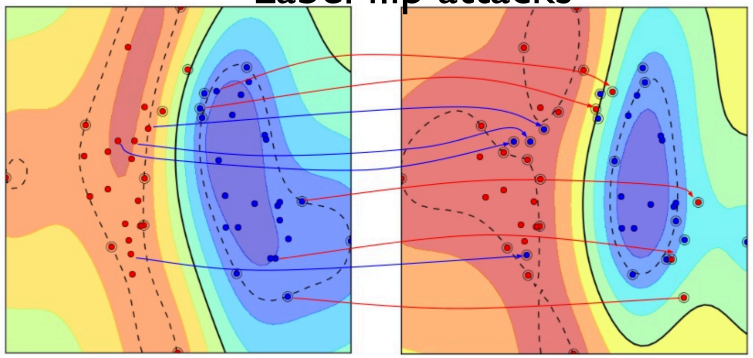
4. Adversarial Learning

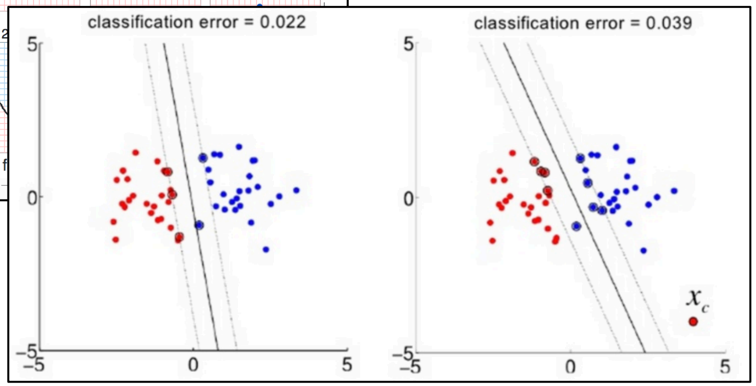
[Xiao et al., Neurocomputing 2014][Biggio et al., ICML'12]

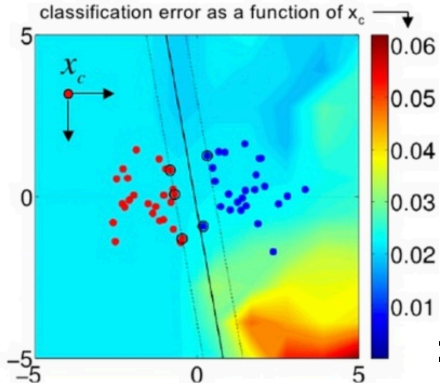
Poisoning Attacks on SVM



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

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

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

Cleaning
Standardization

Attribute selection

Blocking

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

{pairs}

Record pair comparison

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

{comparison vectors }

Decision Model

Linkage decision: $RL(pair) = \dfrac{P(vector \mid pair \in \text{Match})}{P(vector \mid pair \in \text{Non Match})}$

L $\quad U \quad$ $RL(pair)$

Non Match | Potential Match | Match

Pioneer ML-based Deduplication

[Sarawagi, Bhamidipaty, KDD'02]
[Koudas, Srivastava, Sarawagi, Tutorial SIGMOD'06]

Training examples

		f_1	f_2 ... f_n		
Customer 1 Customer 2	D	1.0	0.4	... 0.2	1
Customer 1 Customer 3	N	0.0	0.1	... 0.3	0
Customer 4 Customer 5	D	0.3	0.4	... 0.4	1

← *Similarity distance functions*

Unlabeled list

Customer 6	0.0	0.1	... 0.3	?
Customer 7	1.0	0.4	... 0.2	?
Customer 8	0.6	0.2	... 0.5	?
Customer 9	0.7	0.1	... 0.6	?
Customer 10	0.3	0.4	... 0.4	?
Customer 11	0.0	0.1	... 0.1	?

Classifier

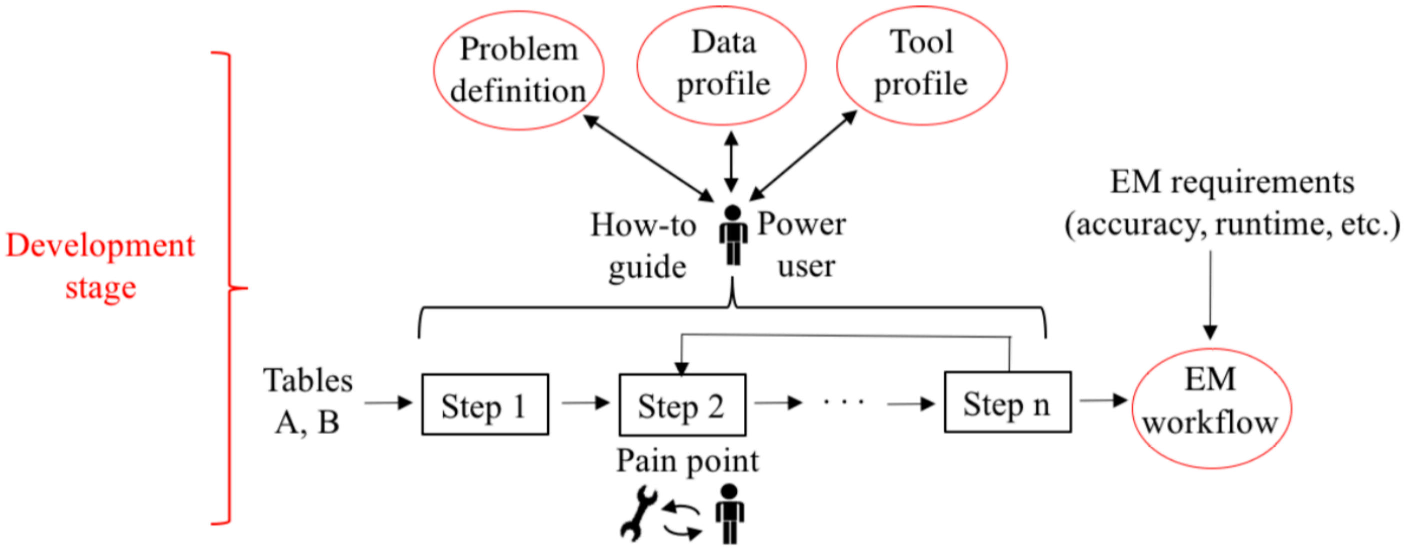
Learnt Rule: All-Ngrams*0.4
+ CustomerAddressNgrams*0.2
– 0.3EnrollYearDifference
+ 1.0*CustomerNameEditDist
+ 0.2*NumberOfAccountsMatch – 3 > 0

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]

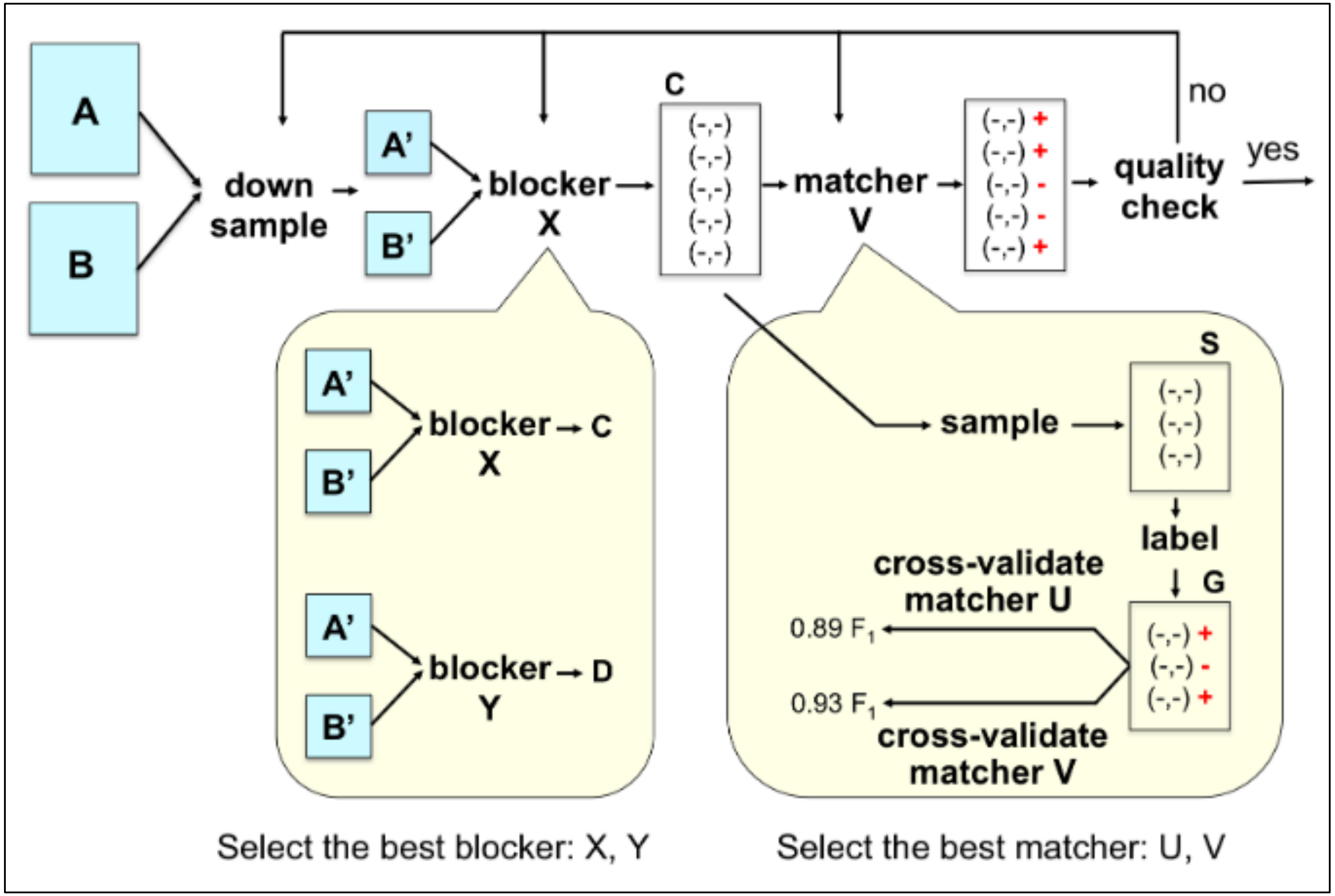
Magellan project: Lessons learnt for How-to Guide for EM



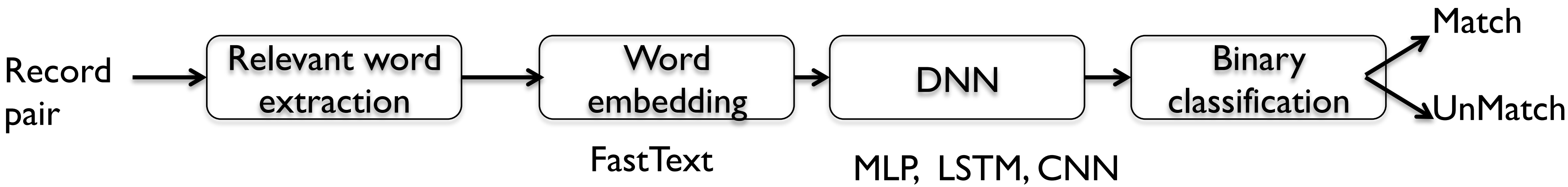
Human-In-The Loop for Entity Matching

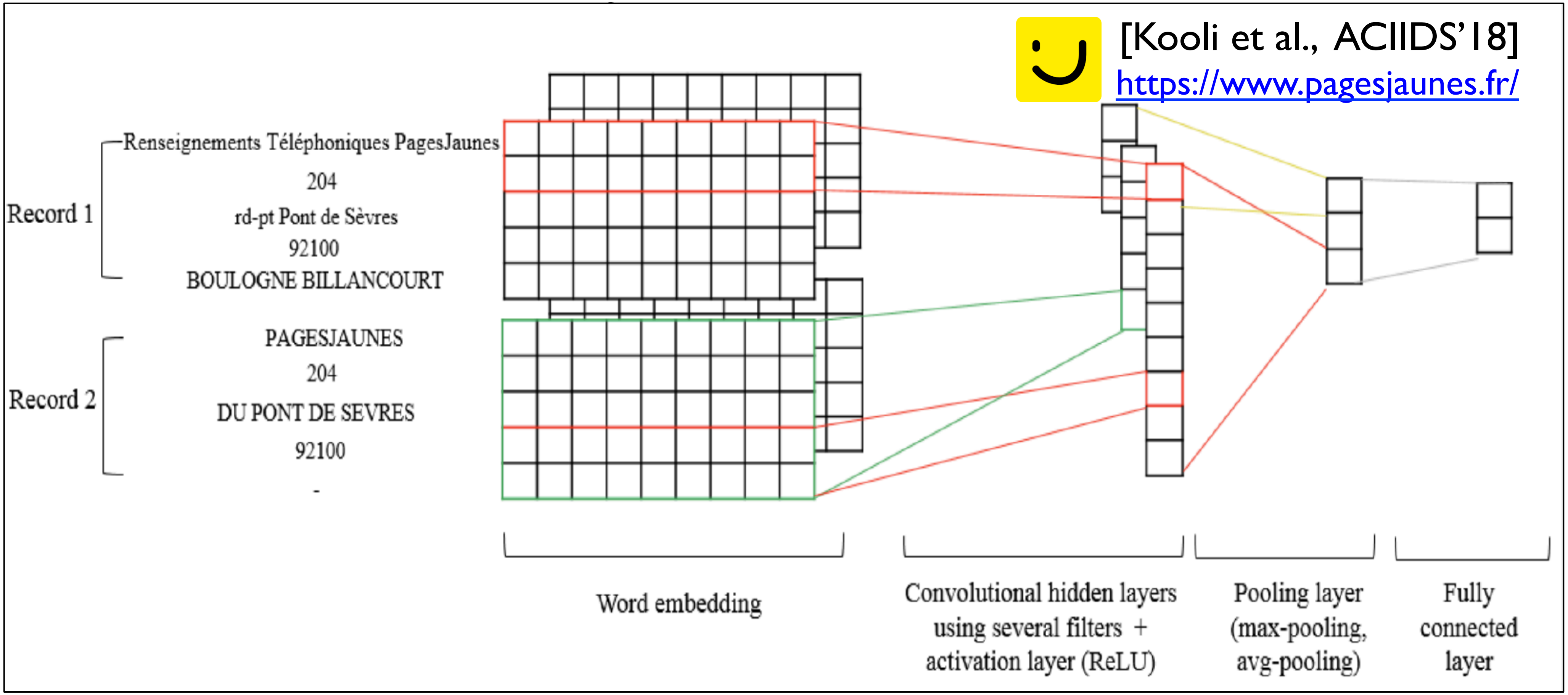
[Doan et al., HILDA@SIGMOD'17]

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

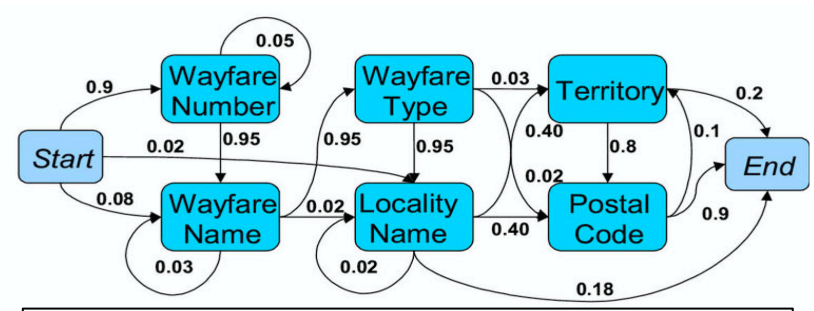
Semi-automatic techniques for:

- **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	0.1	-
WN	-	0.01	0.5	0.01	0.1	0.01	0.01	-
WT	-	0.01	0.01	0.92	0.01	0.01	0.01	-
LN	-	0.01	0.1	0.01	0.8	0.01	0.01	-
TR	-	0.01	0.07	0.01	0.01	0.94	0.01	-
PC	-	0.04	0.01	0.01	0.01	0.01	0.85	-
UN	-	0.02	0.31	0.03	0.06	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

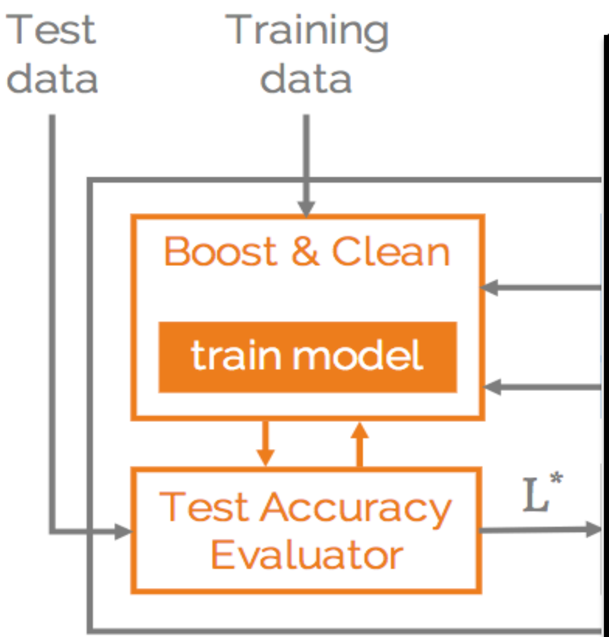
- *Start* -> Wayfare Name (NU) -> Locality Name (LN) -> Postal Code (WT) -> Territory (LN) -> Postal Code (TR) -> Territory (PC) ->*End*
  $0.08 \times 0.01 \times 0.02 \times 0.8 \times 0.4 \times 0.01 \times 0.1 \times 0.01 \times 0.8 \times 0.01 \times 0.1 \times 0.01 \times 0.2 = 8.19 \times 10^{-17}$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

http://users.cecs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/node24.html#chapter:hmm-standard

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)

Initialize $W_i^{(1)} = \frac{1}{N}$

$\mathcal{L}$ generates a set of classifiers $\mathcal{C}\{C^{(0)}, C^{(1)}, ..., C^{(k)}\}$ where $C^{(0)}$ is the base classifier and $C^{(1)}, ..., C^{(k)}$ are derived from the cleaning operations.

for $t \in [1, T]$ **do**

  C_t = Find $C_t \in \mathcal{C}$ that maximizes the weighted accuracy on the test set. ϵ_t = Calculate weighted classification error on the test set $\alpha_t = \ln(\frac{1-\epsilon_t}{\epsilon_t})$ $W_i^{(t+1)} \propto W_i^{(t)} e^{-\alpha_t y_i C_t(x_i)}$: down-weight correct predictions, up-weight incorrectly predictions.

return $C(x) = \text{sign}(\sum_t^T \alpha_t C_t(x))$

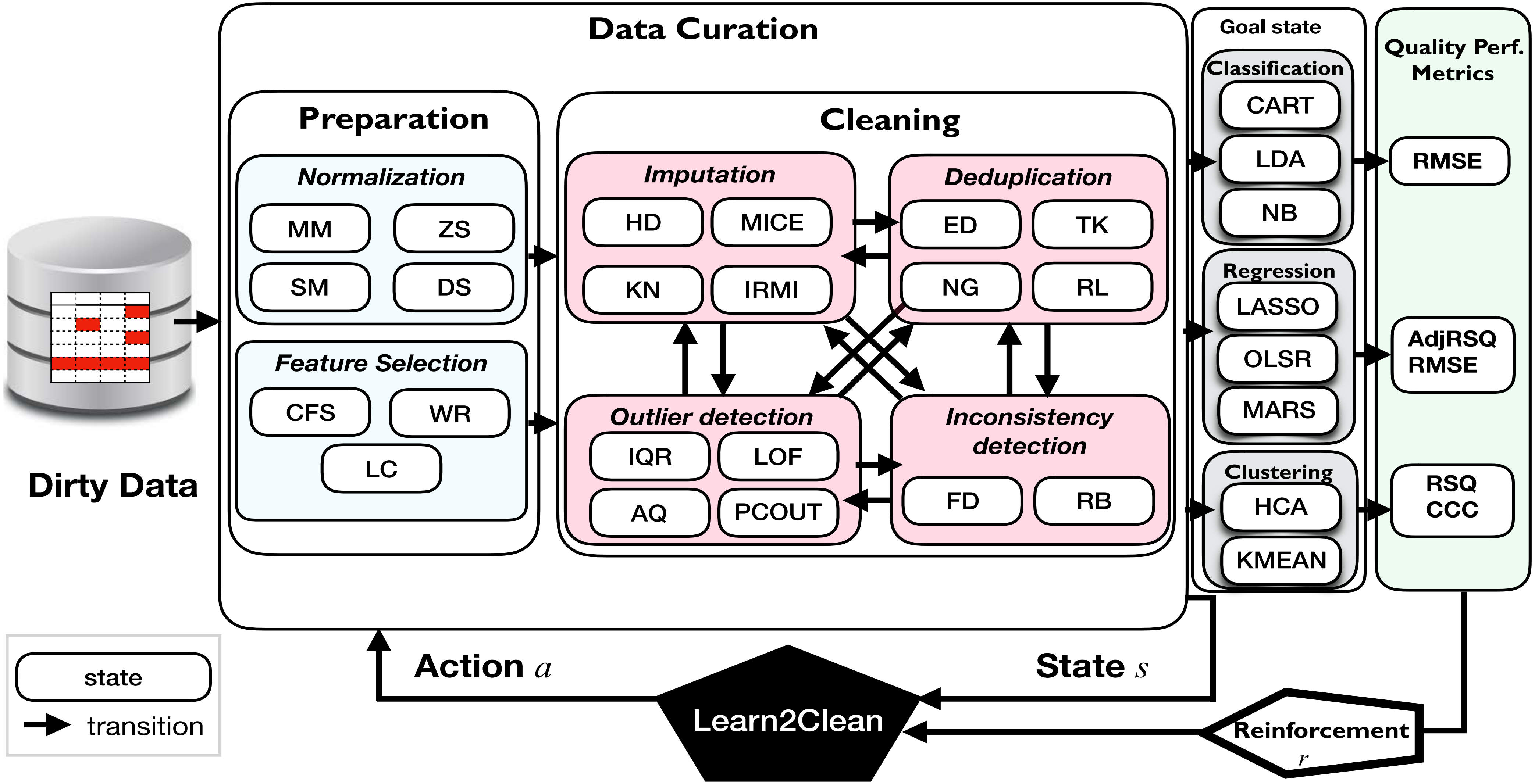
A Condensed View

Repair System	ML Approach	Goal
Febrl [Churches et al., 2002]	HMM and MLE	Standardizing loosely structured texts (e.g., name/ address) based on the probabilistic model learnt from training data
SCARE [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
Continuous Cleaning [Volkovs et al., ICDE'14]	Logistic classifiers	Learning from past user repair preferences to recommend next more accurate repairs
Lens [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
HoloClean [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
BoostClean [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 (1): Humans as Resources

Challenge 1: **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

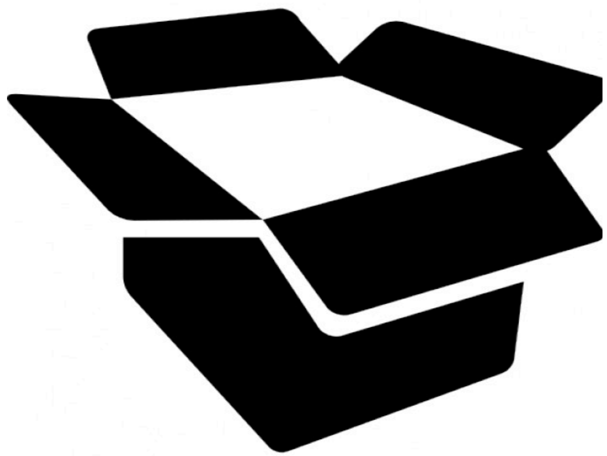
- **Current efforts: Crowdsourcing and active learning**
  - Data cleaning with oracle crowds [Bergman et al., SIGMOD'15]
  - Entity resolution: CrowdER [Wang et al., VLDB'12], Corleone [Gokhale, et al., SIGMOD'14]
  - Data fusion and truth inference [Zheng et al., VLDB'17]
- **Direction**:
  - Adaptive 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]

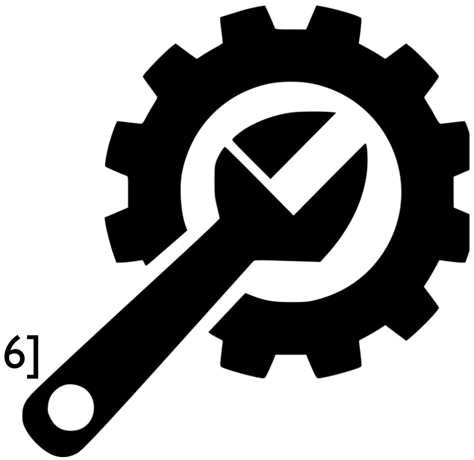


- **Directions**:
  - Causality and explanations in ML-based DM and their effective representation
  - Reversibility and repeatability
  - Data privacy/security: What if adversarial learning is applied ?

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

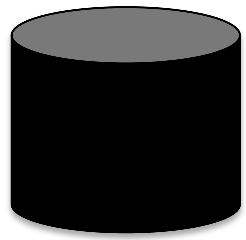
- **Challenge 4**: **Incremental ML application to DM**
  - When we have more training data or refresh/delete some data (obsolete), shall we retrain ML model from scratch? Can we do incremental training/learning? For what cost/trade-off?
- **Challenge 5**: **Runtime ML-based DM**
  - Could we orchestrate 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 (1)

- **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 ±delta in ML accuracy that a given data curation strategy brings to the model?



Accuracy (2)

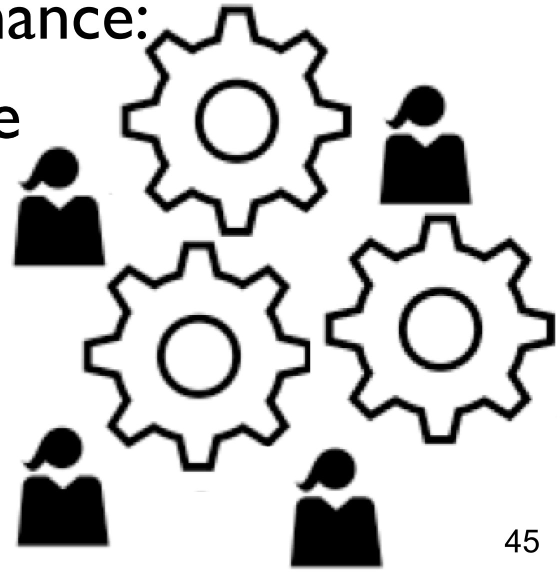
- **Challenge 7**: **Synthetic training data generation**

  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 (1)

[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/papers1/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
[Gokhale et al., SIGMOD'14] https://dl.acm.org/citation.cfm?id=2588576
[Gosh et al., AAAI'17] https://aaai.org/ocs/index.php/WS/AAAIW17/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
[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
[Kumar, Boehm, Yang, SIGMOD'17 Tutorial] 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
[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
[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