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

**DATA**

**TYPE**
- Structural (record)
- Sequential
- Graph-based
- Temporal
- Spatial
- Spatio-Temporal

**RELATIONSHIP**

**DATA QUALITY PROBLEMS**

**TYPE**
- Missing data
- Anomalous data
- Duplicate data
- Inconsistent data
- Obsolete data
- Incorrect data

**CARDINALITY**
- Single-Point Collection

**DETECTION MODE**
- Model-based
- Data distribution-based
- Constraint-based
- Pattern-based
Example 1

Relational data: CiDE.21 committee

<table>
<thead>
<tr>
<th>Nom</th>
<th>Etablissement</th>
<th>Ville</th>
<th>Tel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prof. B. JACQUEMIN</td>
<td>Univ. Lille GERiiCO</td>
<td>Lyon</td>
<td>+33 (0) 3 20 41 66 38</td>
</tr>
<tr>
<td>Malek GHENIMA</td>
<td>ESC Tunis</td>
<td>Tunis</td>
<td>+216 71600615</td>
</tr>
<tr>
<td>Anis BEN MAMI</td>
<td>ESC Tunis</td>
<td>Tunis</td>
<td>74415567</td>
</tr>
<tr>
<td>M. GHENIHA</td>
<td>Tunis</td>
<td>Univ. de la Manouba</td>
<td>+216 71600615</td>
</tr>
<tr>
<td>Mehdi BEN GHANEM</td>
<td>NULL</td>
<td>Tunis</td>
<td>NULL</td>
</tr>
<tr>
<td>Hamida AMDOUN</td>
<td>ESEN-14009</td>
<td></td>
<td>00000000</td>
</tr>
</tbody>
</table>
**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: \( \text{Mahalanobis}_\text{dist}(\text{cov}(X,Y)) > \chi^2(.98,2) \)

Legitimate outliers or data quality problems?
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 issued an apology to French industrialist Martin Bouygues, chairman and CEO of the conglomerate Bouygues Telecom.
Example 5

**Rumors: Celebrity Death Hoaxes**

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.

---

R.I.P. Morgan Freeman

About

Community

At about 5 p.m. ET on Thursday, our beloved actor Morgan Freeman passed away due to a heart attack. 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.

**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.

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

**Autonomous vehicles**
Predictive help, automatic speech recognition, dialog management, etc.

**Personal assistant**
Predictive help, automatic speech recognition, dialog management, etc.
Deep Learning (m)eats Databases (shortened)
Jens Dittrich

Machine Learning and Databases: The Sound of Things to Come or a Cacophony of Hype?
Doug Gemmell
dgemmell@cs.wisc.edu
Metzly Jordan
jordan@cs.wisc.edu
Michael Cyranoski
michael.cyranoski@nytimes.com
Tara Krishnaswamy
tkrishna1@illinois.edu
Chetan Varadarajan
chetanv@illinois.edu

[VLDB’17 Keynote]

ICDE’2016

SIGMOD Record 2016

[VLDB’17 Tutorial]

[ICDE’18 Tutorial]

SIGMOD’15 Panel

[SIGMOD’17 Tutorial]

[SIGMOD Blog, Feb. 2018]

Hot Topic for DB community

DEEM
2nd Workshop on Data Management for End-to-End Machine Learning

SIGMOD Blog, workshop@SIGMOD

VLDB’17 Keynote

ICDE’18 Tutorial

SIGMOD’17 Tutorial

VLDB’17 Tutorial

SIGMOD Record 2016

[ICDE’18 Tutorial]
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

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 1 - ML-Powered Data Curation

• Record Linkage, Deduplication, Entity Resolution

• Error Repair and Pattern Enforcement

• Concluding Remarks and Open Issues
SWOT Analysis (1)
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 [Jogleaker 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.

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Very Conservative estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fleet size</td>
<td>100</td>
</tr>
<tr>
<td>Duration of data collection</td>
<td>1 working year / 8h</td>
</tr>
<tr>
<td>Volume of data generated by a single car</td>
<td>1TB / h</td>
</tr>
<tr>
<td>Data reduction due to preprocessing</td>
<td>0.0005</td>
</tr>
<tr>
<td>Research team size</td>
<td>30</td>
</tr>
<tr>
<td>Proportion of the team submitting jobs</td>
<td>20%</td>
</tr>
<tr>
<td>Target training time</td>
<td>7 days</td>
</tr>
<tr>
<td>Number of epochs required for convergence</td>
<td>50</td>
</tr>
</tbody>
</table>

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]

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/](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]


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
Learning from noisy labels is a hot topic in ML

[Swot Analysis (8)]

- C-SVM Results: 98.5% Accuracy

[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]
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

<table>
<thead>
<tr>
<th>Name</th>
<th>SSN</th>
<th>Addr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will Forth</td>
<td>354-564-339</td>
<td>Ada Bd</td>
</tr>
<tr>
<td>Jacky Khan</td>
<td>435-232-129</td>
<td>Marple Street</td>
</tr>
<tr>
<td>Dom Hack</td>
<td>235-575-689</td>
<td>Main Street</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Database S

<table>
<thead>
<tr>
<th>Name</th>
<th>SSN</th>
<th>Addr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jack Khan</td>
<td>435-223-129</td>
<td>Marple St</td>
</tr>
<tr>
<td>Hans Ford</td>
<td>354-564-339</td>
<td>Clover Bd</td>
</tr>
<tr>
<td>Tom Hack</td>
<td>235-557-689</td>
<td>Main St</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Cleaning

Standardization

Blocking

Attribute selection

Decision Model

\[
RL(pair) = \frac{P(\text{vector} \mid \text{pair} \in \text{Match})}{P(\text{vector} \mid \text{pair} \in \text{Non Match})}
\]

Linkage decision:

- Non Match
- Potential Match
- Match

[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

[Christen, 2008]

[Christen, 2008]

Similarity distance functions

\[ \text{Classifier} \]

Learnt Rule: All-Ngrams*0.4

\[ + \text{CustomerAddressNgrams} \times 0.2 \]

\[ - 0.3 \times \text{EnrollYearDifference} \]

\[ + 1.0 \times \text{CustomerNameEditDist} \]

\[ + 0.2 \times \text{NumberOfAccountsMatch} - 3 > 0 \]

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
Two assumptions:

• A pre-trained word embeddings for all words in the dataset already exists;
• The pre-trained word embeddings that were trained in a task-agnostic manner are sufficient for the ER task.
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

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

Tokenization based on Look-up Tables

Tagging
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 × 0.01 × 0.02 × 0.8 × 0.4 × 0.01 × 0.1 × 0.8 × 0.01 × 0.1 × 0.8 × 0.01 × 0.2 = 8.19 × 10^{-17}

- Start -> Wayfare Number (NU) -> Wayfare Name (LN) -> Wayfare Type (WT) -> Locality (LN) -> Territory (TR) -> Postal Code (PC) -> End
  0.9 × 0.9 × 0.95 × 0.1 × 0.95 × 0.92 × 0.95 × 0.8 × 0.4 × 0.94 × 0.8 × 0.85 × 0.9 = 1.18 × 10^{-2}

http://users.cecs.anu.edu.au/~Peter.Christen/Febrl/febrl-0.3/febrldoc-0.3/node24.html#chapter:hmm-standard
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 \(W^{(1)}_i = \frac{1}{N}\)

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

3. **for** \(t \in [1, T] \)** **do**
   
   4. \(C_t = \text{Find } C_t \in \mathcal{C} \text{ that maximizes the weighted accuracy on the test set. } \epsilon_t = \text{Calculate weighted classification error on the test set } \alpha_t = \ln\left(\frac{1}{\epsilon_t}\right)\)

   
   5. \(W^{(t+1)}_i \propto W^{(t)}_i e^{-\alpha_t y_i C_t(x_i)}\): down-weight correct predictions, up-weight incorrectly predictions.

4. **return** \(C(x) = \text{sign}(\sum^T_t \alpha_t C_t(x))\)


# A Condensed View

<table>
<thead>
<tr>
<th>Repair System</th>
<th>ML Approach</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Febrl</td>
<td>HMM and MLE</td>
<td>Standardizing loosely structured texts (e.g., name/address) based on the probabilistic model learnt from training data</td>
</tr>
<tr>
<td>SCARE</td>
<td>Multiple ML models used to capture data dependencies across multiple data partitions</td>
<td>Find the candidate repair that maximizes the likelihood repair benefit under a cost threshold of the update</td>
</tr>
<tr>
<td>Continuous Cleaning</td>
<td>Logistic classifiers</td>
<td>Learning from past user repair preferences to recommend next more accurate repairs</td>
</tr>
<tr>
<td>Lens</td>
<td>Various ML models encoded in Domain Constraints</td>
<td>Declarative on-Demand ETL with prioritized curation tasks based on probabilistic query processing and PC-Tables</td>
</tr>
<tr>
<td>HoloClean</td>
<td>Probabilistic inference on factor graphs with SGD and Gibbs sampling</td>
<td>Mixing statistical and logical rules, DCs, MDs, etc. to infer candidate repairs in a scalable way with domain pruning and constraint relaxation</td>
</tr>
<tr>
<td>BoostClean</td>
<td>AdaBoost</td>
<td>Mixing statistical and logical rules, domain constraints for detection and repair combinations to maximize the predictive accuracy over test data</td>
</tr>
</tbody>
</table>
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 (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 ±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]  

[Arasu et al., SIGMOD’10]  
https://dl.acm.org/citation.cfm?id=1807252

[Assadi, Milo, Novgorodov, WebDB’18]  

[Bahmani et al., SUM’15]  

[Battacharya, Getoor, TKDD’07]  
https://dl.acm.org/citation.cfm?id=1217304

[Bellare et al., KDD’12]  

[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]  

[Bilenko, Mooney, KDD’03]  
https://dl.acm.org/citation.cfm?id=956759

[Chaudhuri et al., ICDE’05]  

[Chaudhuri et al., VLDB’07]  
https://dl.acm.org/citation.cfm?id=1559869

[Chen et al., SIGMOD’09]  

[Christen et al., 2002]  

[Christen, 2012]  
http://www.biomedcentral.com/1472-6947/2/9/

[Churches et al., 2002]  

[Crotty et al., VLDB’15]  

[Dean, NIPS 2017]  

[Doan et al., HILDA@SIGMOD’17]  
https://conf.slac.stanford.edu/xldb2018/event-information/lightning-talks

[Dzakovic, XLDB’18]  

[Ebraheem et al., Arxiv 2017]  

[Fellegi, Sunter, 1969]  
https://dl.acm.org/citation.cfm?id=2783396

[Fisher et al., KDD’15]  
[Getoor, Machanavajjhala, Tutorial VLDB’12]  

[Gokhale et al., SIGMOD’14]  
https://dl.acm.org/citation.cfm?id=2588576

[Gosh et al., AAAI’17]  
References - Part I (2)

[Gupta, Sarawagi, VLDB’09]
[Hall, 1992]
[Hassanzadeh et al., PVLDB’09]
[Hu et al, 2017]
[Ilyas, Chu, 2015]
[John et al., CIDR’17]
[Joglekar et al., SIGMOD’17]
[Kooli et al., ACIIDS’18]
[Köpcke et al., VLDB’10]
[Koudas, Srivastava, Sarawagi, Tutorial]
[Kraska et al. 2017]
[Krishnan et al., VLDB’16]
[Krishnan et al., 2017]
[Kumar et al., SIGMODRec’15]
[Kumar, Boehm, Yang, SIGMOD’17 Tutorial]
[Luyang et al., Sensors’17]
[Marchand, Rubinstein, VLDB’17]
[Marcus, Arxiv, 2018]
[Mintz et al., 2009]
[Natarajan et al., NIPS’13]
[Papadakis et al., TKDE 2013]
[Papadakis, Palpanas, Tutorial ICDE’16]
[Polyzotis et al., SIGMOD’17]
[Qian et al., CIKM’17]

https://dl.acm.org/citation.cfm?id=1687627.1687661
Mathematical Techniques in Multisensor Data Fusion, ArtechHouse, 1992
http://www.vldb.org/pvldb/2/vldb09-1025.pdf
http://pages.cs.wisc.edu/~jignesh/pub/Ava.pdf
https://dl.acm.org/citation.cfm?id=3035951
https://link.springer.com/chapter/10.1007%2F978-3-319-75420-8_1
https://arxiv.org/abs/1712.01208
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5335931/
https://dl.acm.org/citation.cfm?id=1690287
https://dl.acm.org/citation.cfm?id=3035918.3054782
https://dl.acm.org/citation.cfm?id=3132949
References - Part I (3)

[Sarawagi, Bhamidipaty, KDD’02] https://dl.acm.org/citation.cfm?id=775087
[Singla, Domingos, PKDD’05] https://sites.google.com/site/pkujiangtang/home/kdd17-tutorial
[Tang, KDD’17 tutorial] https://dl.acm.org/citation.cfm?id=775099
[Volkovs et al., ICDE’14] https://dl.acm.org/citation.cfm?id=2465280
[Xia et al., Neurocomputing 2014] https://dl.acm.org/citation.cfm?id=2463706
[Zheng et al., VLDB’17]