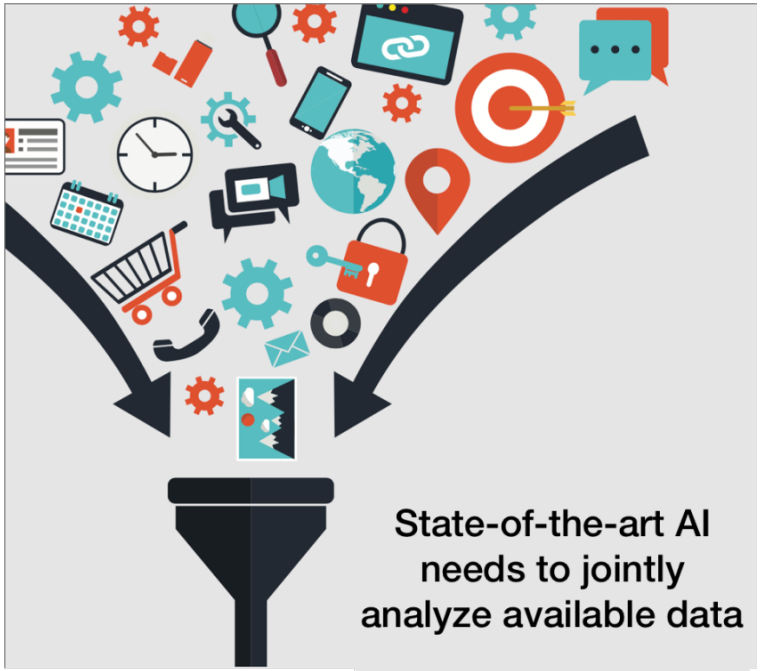
Outline

- Part I. Introduction
- Part II. ML for DI
- Part III. DI for ML
  - Training data creation
  - Data cleaning
- Part IV. Conclusions and research directions

Successful ML requires Data Integration

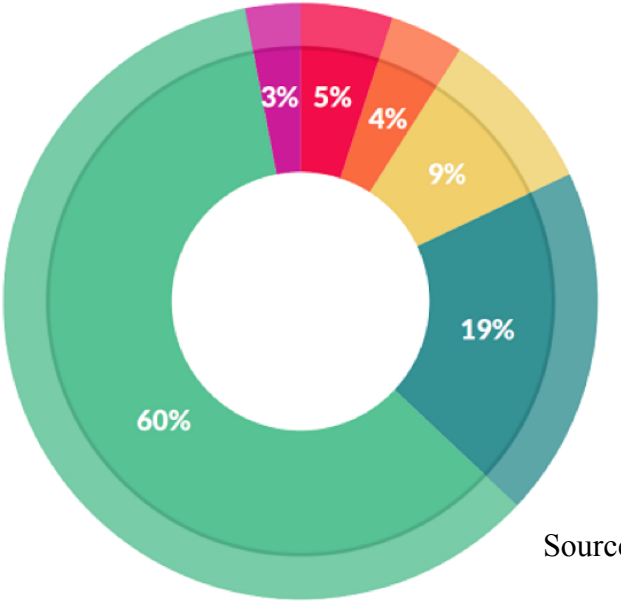


IMAGENET MovieLens

COCO is a large-scale object detection, segmentation, and captioning dataset.

Large collections of **manually curated** training data are necessary for progress in ML.

Noisy data is a bottleneck



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

Source: Crowdflower

Cleaning and organizing data comprises 60% of the time spent on an analytics of AI project.

50 Years of Data Cleaning

1970s (Nulls)

E. F. Codd

- Understanding relations (installment #7). *FDT - Bulletin of ACM SIGMOD*, 7(3):23–28, 1975.
- Null-related features of DBs

1980s (Normalization)

Integrity Constraints

- Normal forms to reduce redundancy and integrity
- FDs, MVDs etc.

1990s (Warehouses)

Data transforms

- Part of ETL
- Errors within a source and across sources
- Transformation workflows and mapping rules; domain-knowledge is crucial

2000s (Data Repairs)

Constraints and Probabilities

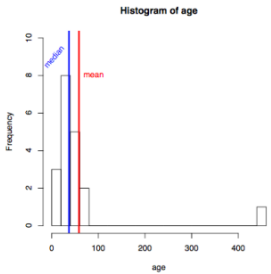
- Dichotomies for consistent query answering
- Minimality-based repairs to obtain consistent instances
- Statistical repairs
- Anomaly detection

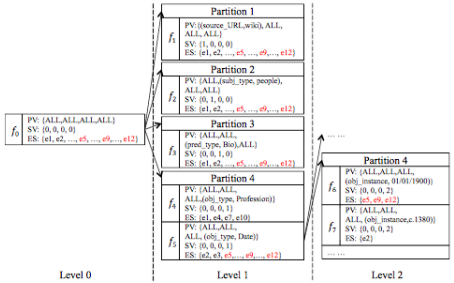
Where are we today?

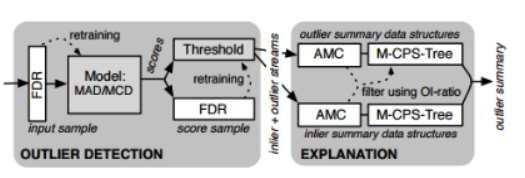
Machine learning and statistical analysis are becoming more prevalent.

Error detection (Diagnosis)

- Anomaly detection [Chandola et al., ACM CSUR, 2009]
- Bayesian analysis (Data X-Ray) [Wang et al., SIGMOD'15]
- Outlier detection over streams (Macrobase) [Bailis et al., SIMGOD'17]





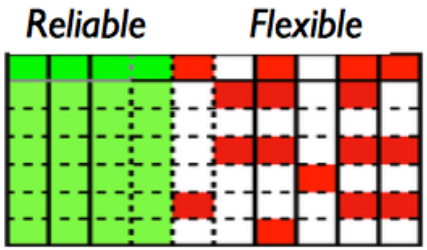


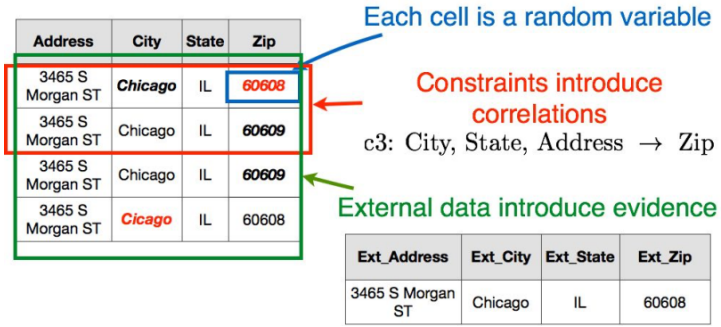
Where are we today?

Machine learning and statistical analysis are becoming more prevalent.

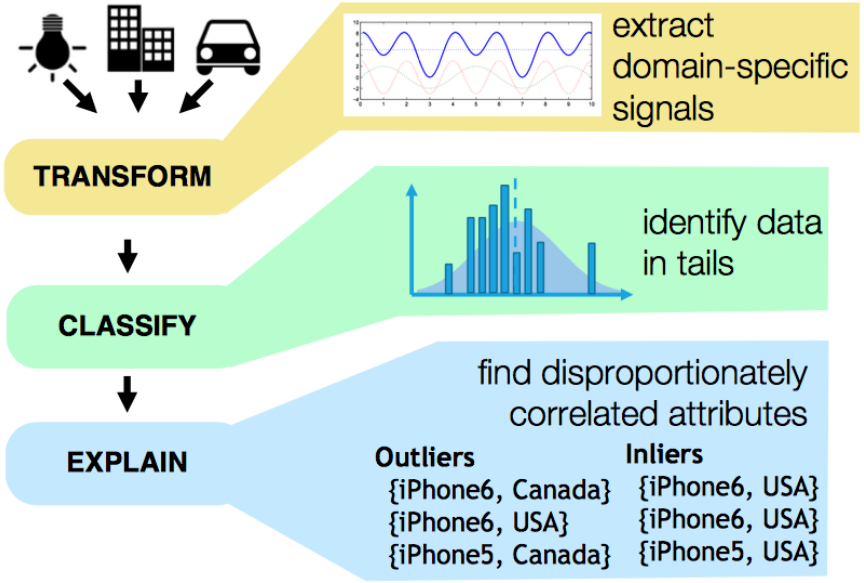
Data Repairing (Treatment)

- Classical ML (SCARE, ERACER) [Yakout et al., VLDB'11, SIGMOD'13, Mayfield et al., SIGMOD'10]
- Boosting [Krishan et al., 2017]
- Weakly-supervised ML (HoloClean) [Rekatsinas et al., VLDB'17]





Error Detection: MacroBase [Bailis et al., SIGMOD'17]



[Figure by Kai Sheng Tai]

Streaming Feature Selection

Setup: Online learning of a classifier (e.g., LR)

Goal: Return top-k discriminative features

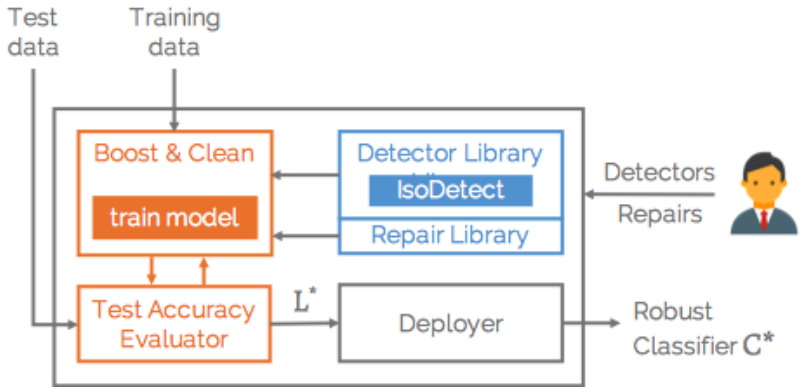
Weight-Median Sketch

Sketch of a classifier for fast updates and queries for estimates of each weight and comes with approximation guarantees

A data analytics tool that prioritizes attention in large datasets.

Code at: macrobase.stanford.edu

Data Repairing: BoostClean [Krishnan et al., 2017]



Ensemble learning for error detection and data repairing.

Relies on domain-specific detection and repairing.

Builds upon boosting to identify repairs that will maximize the performance improvement of a downstream classifier.

On-demand cleaning!

Scalable machine learning for data enrichment

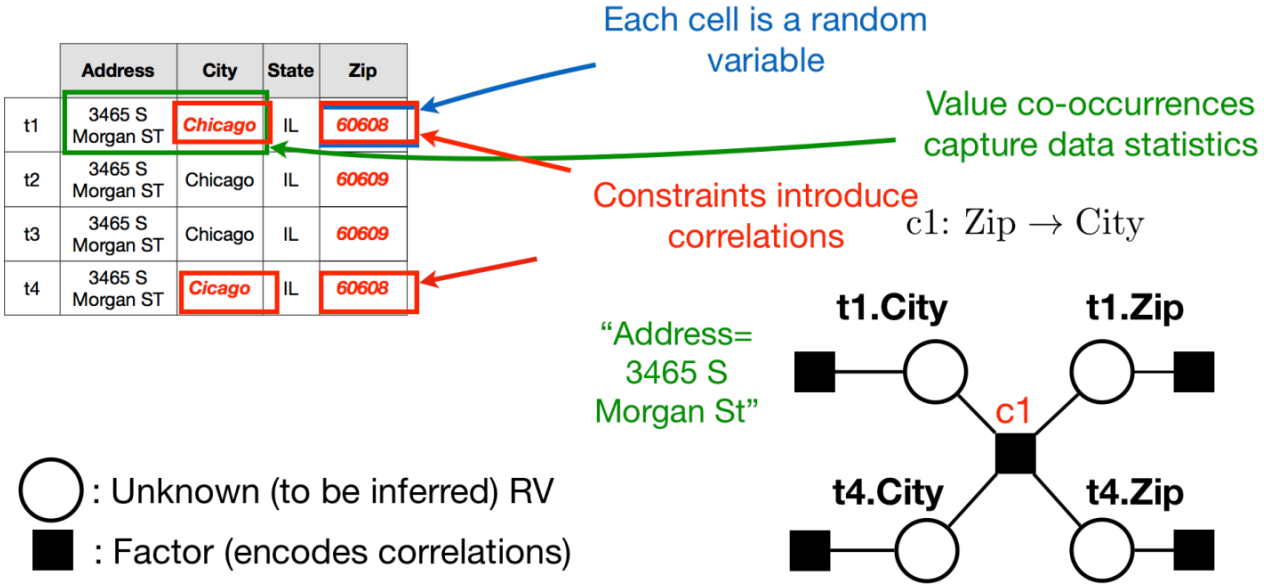


HoloClean

Code available at:

http://www.holoclean.io

Data Repairing: HoloClean [Rekatsinas et al., VLDB'17]

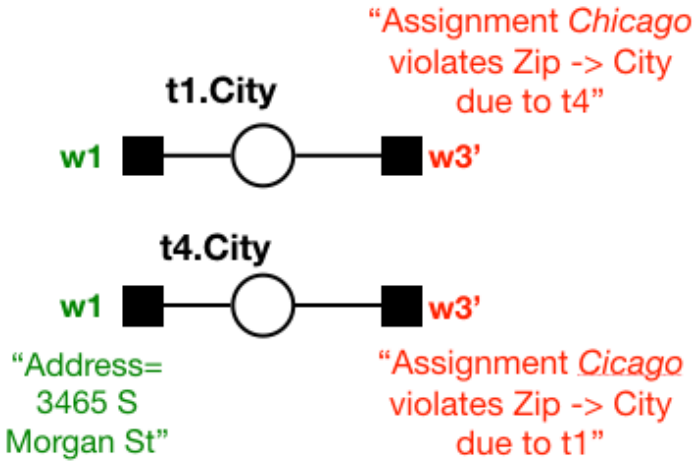


	Address	City	State	Zip
t1	3465 S Morgan ST	Chicago	IL	60608
t2	3465 S Morgan ST	Chicago	IL	60609
t3	3465 S Morgan ST	Chicago	IL	60609
t4	3465 S Morgan ST	Cicago	IL	60608

Holistic data cleaning framework: combines a variety of heterogeneous signals (e.g., integrity constraints, external knowledge, quantitative statistics)

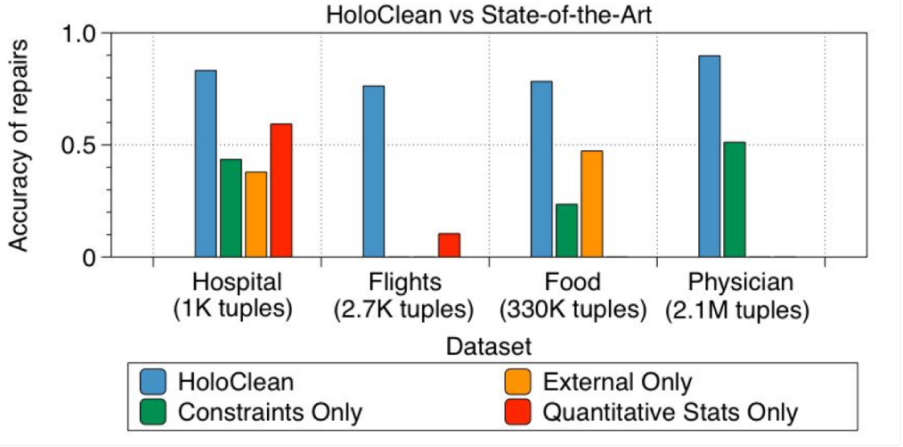
Data Repairing: HoloClean [Rekatsinas et al., VLDB'17]

	Address	City	State	Zip
t1	3465 S Morgan ST	*Chicago*	IL	*60608*
t2	3465 S Morgan ST	Chicago	IL	*60609*
t3	3465 S Morgan ST	Chicago	IL	*60609*
t4	3465 S Morgan ST	*Cicago*	IL	*60608*



Scalable learning and inference: Hard constraints lead to complex and non-scalable models. Novel relaxation to features over individual cells.

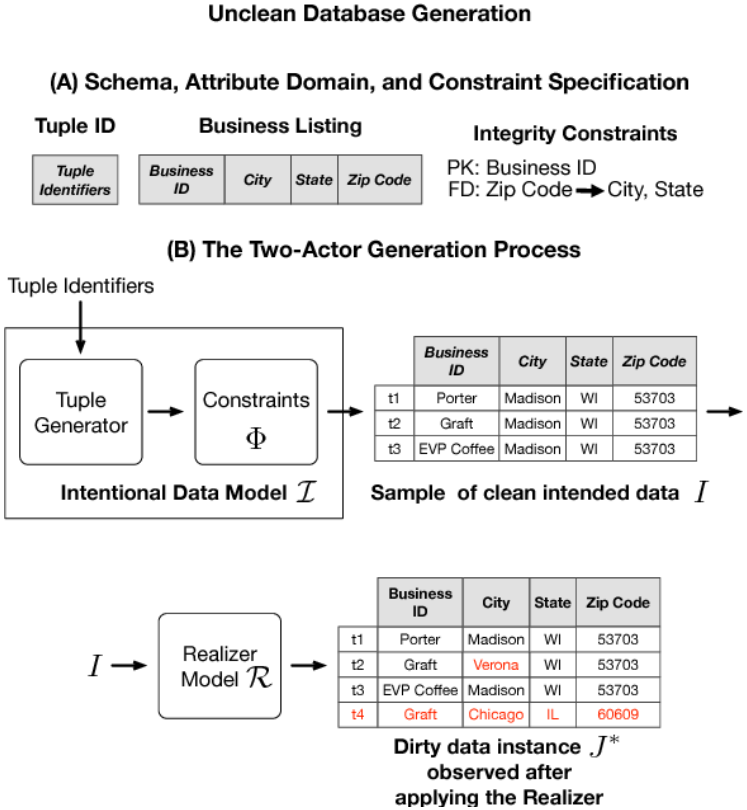
Data Repairing: HoloClean [Rekatsinas et al., VLDB'17]



HoloClean: our approach combining all signals and using inference
Holistic[Chu,2013]: state-of-the-art for constraints & minimality
KATARA[Chu,2015]: state-of-the-art for external data
SCARE[Yakout,2013]: state-of-the-art ML & qualitative statistics

HoloClean is 2x more accurate. Competing methods either do not scale or perform no correct repairs.

Probabilistic Unclean Databases [De Sa et al., 2018]



A two-actor noisy channel model for managing erroneous data.

Preprint: *A Formal Framework For Probabilistic Unclean Databases*

https://arxiv.org/abs/1801.06750

Challenges in Data Cleaning

- Error detection is still a challenge. To what extent is ML useful for error detection? Tuple-scoped approaches seem to be dominating. Is deep learning useful?

- We need a formal framework to describe when automated solutions are possible.

- A major bottleneck is the collection of training data. Can we leverage weak supervision and data augmentation more effectively?

- Limited end-to-end solutions. Data cleaning workloads (mixed relational and statistical workloads) pose unique scalability challenges.

Recipe for Data Cleaning

- Problem definition: **Detect and repair erroneous data.**
- Short answers
  - **ML can help partly-automate cleaning. Doma expertise is still required.**
  - **Scalability of ML-based data cleaning methods is a pressing challenge. Exciting systems research!**
  - **We need more end-to-end systems!**