# Data Integration and Large-Scale Analysis

## 06 - Data Cleaning and Fusion

**Dr. Lucas Iacono - 2025**
Know Center Research GmbH & Graz University of Technology

# Agenda

- Motivation and Terminology
- Data Cleaning and Fusion
- Missing Value Imputation

# Motivation and Terminology

# Recap: Corrupted/Inconsistent Data

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

# Recap: Corrupted/Inconsistent Data

**Uniqueness & duplicates** | **Contradictions & wrong values** | **Missing Values** | **Ref. Integrity**

[**Credit:** Felix Naumann]

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

| Zip | City |
|---|---|
| 98120 | San Jose |
| 90001 | Lost Angeles |

**Typos**

# Examples (aka errors are everywhere)

```
- US,DFW,LIT,ER4;M83;M83
+ US,DFW,LIT,ER4;M83
```

- Duplicates
- Formatting
- Data Entry Errors
- Encoding errors
- Missing values
- Date-time encoding

```
- Beni Airport,Beni,Congo (Kinshasa),BNC,FZNP,0.575,2
+ Beni Airport,Beni,Democratic Republic of Congo,BNC,
```

```
- RAF St Athan,4Q,STN,UNited Kingdom,N
+ RAF St Athan,4Q,STN,United Kingdom,N
```

```
- Oyo Ollombo Airport,Oyo,Congo (Brazzaville),O
+ Oyo Ollombo Airport,Oyo,Republic of Congo,OLL
```

```
ID,NAME,RATING,PHONENUMBER,NO_OF_REVIEWS,ADDRESS
1445980000001,1,5,"(800) 586-5735",38,"867 N Hermitage Ave, Chicago, IL 60622"
1445980000002,326,3.5,"(323) 549-2156",33,"6333 3rd St, Los Angeles, CA 90036"
1445980000003,1760,4,"(415) 359-1212",454,"1760 Polk St, San Francisco, CA 94109"
1445980000004,"□□ ",4,"(773) 866-9898",185,"2977 N Elston Ave, Chicago, IL 60618"
1445980000005,"□□□Disiac Lounge ",3.5,"(212) 586-9880",164,"402 W 54th St, New York, NY 10019"
1445980000006,"□□□Ģ T.'s review of Belly Good Cafe & Crepe",4.5,"(415) 346-8383",843,"1737 Post St,
1445980000007,"□□ireTrea ",4,"(415) 967-2726",63,"San Francisco, CA 94109"
1445980000008,"10e Restaurant ",4,"(213) 488-1096",166,"811 W 7th St, Los Angeles, CA 90017"
1445980000009,"10th & Wood ",4,"(510) 645-1955",275,"945 Wood St, Oakland, CA 94607"
```

| src | flight | scheduled_dept | actual_dept |
|---|---|---|---|
| ua | 2011-12-01-UA-2708-EWR-CLT | Thu- Dec 1 2:55 PM | Thu- Dec 1 2:55 PM |
| airtravelcenter | 2011-12-01-UA-2708-EWR-CLT | | 12/1/11 3:04 PM (-05:00) |
| myrateplan | 2011-12-01-UA-2708-EWR-CLT | | 12/1/11 3:04 PM (-05:00) |
| helloflight | 2011-12-01-UA-2708-EWR-CLT | | 12/1/11 3:04 PM (-05:00) |
| flytecomm | 2011-12-01-UA-2708-EWR-CLT | | 12/1/11 3:04 PM (-05:00) |
| flights | 2011-12-01-UA-2708-EWR-CLT | | 2011-12-01 02:52 PM |
| businesstravellogue | 2011-12-01-UA-2708-EWR-CLT | | 2011-12-01 02:52 PM |
| flylouisville | 2011-12-01-UA-2708-EWR-CLT | | 2011-12-01 02:52 PM |
| flightstats | 2011-12-01-UA-2708-EWR-CLT | 2011-12-01 2:55 PM | 2011-12-01 2:52 PM |
| quicktrip | 2011-12-01-UA-2708-EWR-CLT | 2011-12-01 2:55 PM | 2011-12-01 2:52 PM |
| flightview | 2011-12-01-UA-2708-EWR-CLT | | 3:04 PMDec 01 |
| panynj | 2011-12-01-UA-2708-EWR-CLT | | 3:04 PMDec 01 |
| gofox | 2011-12-01-UA-2708-EWR-CLT | | 3:04 PMDec 01 |

# Terminology

- **#1 Data Cleaning (aka Data Cleansing)**
  - **Detection** and **repair** of data errors
  - **Outliers/anomalies:** values or objects that do not match normal behavior
    (different goals: data cleaning vs finding interesting patterns)
  - **Data Fusion:** resolution of inconsistencies and errors
    (e.g., entity resolution **see Lecture 05**)
- **#2 Missing Value Imputation**
  - **Fill missing info** with “best guess”
  - Difference between NAs and 0 (or special values like NaN) for ML models
- **#3 Data Wrangling**
  - Automatic cleaning unrealistic? → Interactive data transformations
  - Recommended transforms + user selection

# Express Expectations as Validity Constraints

- Manual Approach: **"Common Sense"**

Age=9999?

- (Semi-)Automatic Approach: **Expectations!**
  - PK → Values must be **unique and defined (not null)**
  - Exact PK-FK → **Inclusion** dependencies
  - **Semantics** of attributes→ Value ranges / # distinct values
  - Invariant to **capitalization**
    → **Duplicates** that differ in capitalization
- Formal Constraints
  - Denial constraints

```
- US,DFW,LIT,ER4;M83;M83
+ US,DFW,LIT,ER4;M83
```

2019-11-15 vs Nov 15, 2019

```
- RAF St Athan,4Q,STN,UNited Kingdom,N
+ RAF St Athan,4Q,STN,United Kingdom,N
```

# Express Expectations as Validity Constraints

- **Formal Constraints**
  - E.g. Denial constraints

$$\forall t_\alpha t_\beta \in R: \neg(t_\alpha.Role = t_\beta.Role \wedge t_\alpha.City = 'NYC' \wedge t_\beta.City \neq 'NYC' \wedge t_\alpha.Salary < t_\beta.Salary)$$

# Data Cleaning and Fusion

# Data Validation

**Validation checks on expected shape before training first model**

[Neoklis Polyzotis, Sudip Roy, Steven Euijong Whang, Martin Zinkevich: Data Management Challenges in Production Machine Learning. Tutorial, **SIGMOD 2017**]


(Google Research)

- Check a feature's **min, max,** and **most common value**
  - Ex: Latitude values must be within the range [-90, 90] or [-π/2, π/2]
- The **histograms** of continuous or categorical values **are as expected**
  - Ex: There are **similar numbers of positive and negative** labels
- Whether **a feature** is **present in enough examples**
  - Ex: Country code must be in **at least 70%** of the examples
- Whether a **feature** has the **right number of values** (i.e., cardinality)
  - Ex: There **cannot be more than one age of a person**

# Data Validation, cont.

[Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Bießmann, Andreas Grafberger: Automating Large-Scale Data Quality Verification. **PVLDB 2018**]

(**Amazon Research**)

- **Constraints and Metrics for quality check UDFs**

| constraint | arguments |
|---|---|
| dimension *completeness* | |
| isComplete | column |
| hasCompleteness | column, udf |
| dimension *consistency* | |
| isUnique | column |
| hasUniqueness | column, udf |
| hasDistinctness | column, udf |
| isInRange | column, value range |
| hasConsistentType | column |
| isNonNegative | column |
| isLessThan | column pair |
| satisfies | predicate |
| satisfiesIf | predicate pair |
| hasPredictability | column, column(s), udf |
| statistics (can be used to verify dimension *consistenc* | |
| hasSize | udf |
| hasTypeConsistency | column, udf |
| hasCountDistinct | column |
| hasApproxCountDistinct | column, udf |
| hasMin | column, udf |
| hasMax | column, udf |
| hasMean | column, udf |
| hasStandardDeviation | column, udf |
| hasApproxQuantile | column, quantile, udf |
| hasEntropy | column, udf |
| hasMutualInformation | column pair, udf |
| hasHistogramValues | column, udf |
| hasCorrelation | column pair, udf |
| time | |
| hasNoAnomalies | metric, detector |

| metric |
|---|
| dimension *completeness* |
| Completeness |
| dimension *consistency* |
| Size |
| Compliance |
| Uniqueness |
| Distinctness |
| ValueRange |
| DataType |
| Predictability |
| statistics (can be used to |
| Minimum |
| Maximum |
| Mean |
| StandardDeviation |
| CountDistinct |
| ApproxCountDistinct |
| ApproxQuantile |
| Correlation |
| Entropy |
| Histogram |
| MutualInformation |

**Organizational Lesson:**
benefit of shared vocabulary/procedures

**Technical Lesson:**
fast/scalable; reduce manual and ad-hoc analysis

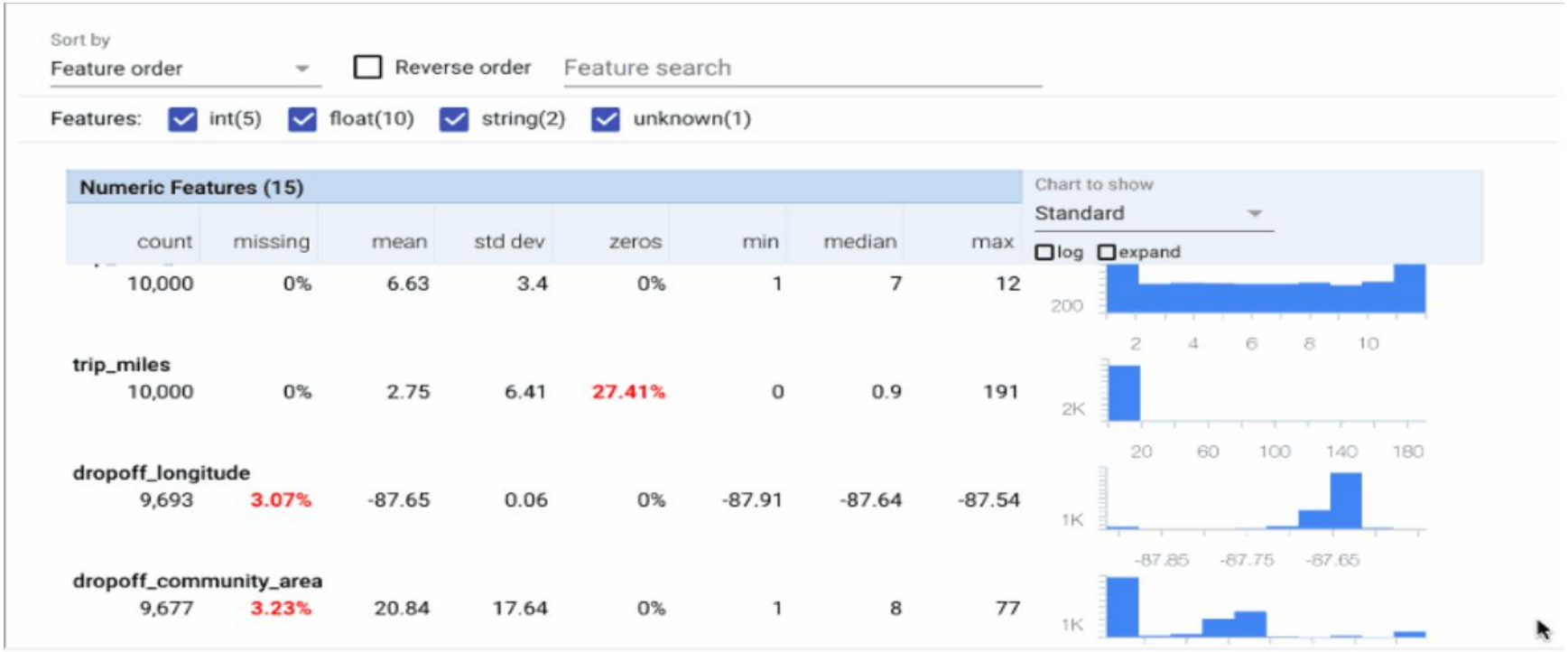
- **Approach**
  - **#1** Quality checks on basic metrics, computed in **Apache Spark**
  - **#2 Incremental maintenance** of metrics and quality checks

# Data Validation, cont.

[Mike Dreves; Gene Huang; Zhuo Peng; Neoklis Polyzotis; Evan Rosen; Paul Suganthan: From Data to Models and Back. **DEEM 2020**]

(**Google**)

- **TensorFlow Data Validation (TFDV)**
  - Library or TFX components
  - Provides functions for stats computation, validation checks and anomaly detection



# Standardization and Normalization

- **#1 Standardization**
  - Centering and scaling to mean 0 and variance 1
  - **Ensures well-behaved training**
  - **Densifying operation**
  - Awareness of NaNs
  - Batch normalization in DNN: standardization of activations

```
X = X - colMeans(X);
X = X / sqrt(colVars(X));
```

```
X = replace(X, pattern=NaN,
  replacement=0); #robustness
```

# Standardization and Normalization

- **#1 Standardization**
  - Centering and scaling to mean 0 and variance 1
  - **Ensures well-behaved training**
  - **Densifying operation**
  - Awareness of NaNs
  - Batch normalization in DNN: standardization of activations

```
X = X - colMeans(X);
X = X / sqrt(colVars(X));

X = replace(X, pattern=NaN,
  replacement=0); #robustness
```

Wo/Standardization = Age Values [10, 20, 30, 40, 50]

Mean = 30, Scale [50 - 10 = 40] → **bad-behaved training**

W/ Standardization = Age Values [-1.26, -0.63, 0, 0.63, 1.26]

Mean = 0, Scale [1,26 -(-1,26) = 2,52] → **well-behaved training**

# Standardization and Normalization

- **#1 Standardization**
  - Centering and scaling to mean 0 and variance 1
  - **Ensures well-behaved training**
  - **Densifying operation**
  - Awareness of NaNs
  - Batch normalization in DNN: standardization of activations

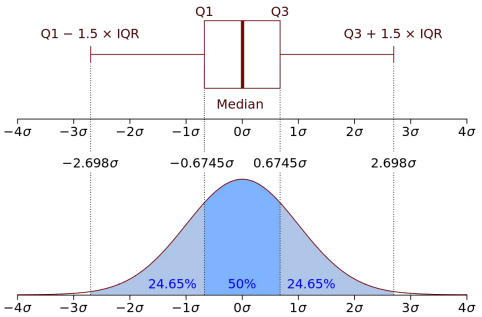
```
X = X - colMeans(X);
X = X / sqrt(colVars(X));

X = replace(X, pattern=NaN,
  replacement=0); #robustness
```

- **#2 Normalization**
  - Aka min-max normalization
  - Rescale values into common range [0,1]
  - Does not handle outliers [10, 20, 30, 40, 1000] → Norm [0.00, 0.01, 0.02, 0.03, 1.00]

```
X = (X - colMins(X))
  / (colMaxs(X) - colMins(X));
```

# Winsorizing and Trimming




[**Credit:** https://en.wikipedia.org]

- **Method to detect and correct outliers (errors)**
- **Winsorizing: Replace** tails of data distribution at user-specified threshold

```
# compute quantiles for lower and upper
ql = quantile(X, 0.05);
qu = quantile(X, 0.95);

# replace values outside [ql,qu] w/ ql and qu
Y = ifelse(X < ql, ql, X);
Y = ifelse(Y > qu, qu, Y);
```

- **Truncation/Trimming: Remove** tails of data distribution at user-specified threshold

```
# remove values outside [ql,qu]
I = X < qu | X > ql;
Y = removeEmpty(X, "rows", select = I);
```

**Instead of a constant, you can also use mean/median/mode**

HCC Institute of HUMAN-CENTRED COMPUTING

# Outliers and Outlier Detection

- **Types of Outliers**

  [Varun Chandola, Arindam Banerjee, Vipin Kumar: Anomaly detection: A survey. **ACM Comput. Surv. 2009**]

  - **Point outliers:** single data points far from the data distribution
  - **Contextual outliers:** noise or other systematic anomalies in data
  - **Sequence (contextual) outliers:** sequence of values w/ abnormal shape/agg

- **Types of Outlier Detection**

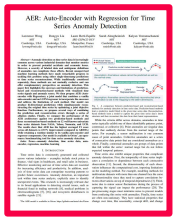
  [Victoria J. Hodge, Jim Austin: A Survey of Outlier Detection Methodologies. **Artif. Intell. Rev. 2004**]

  - **Type 1 Unsupervised:** No prior knowledge of data, similar to unsupervised **clustering** → **expectations:** distance, # errors
  - **Type 2 Supervised:** Labeled normal and abnormal data, similar to supervised **classification**
  - **Type 3 Normal Model:** Represent normal behavior, similar to **pattern recognition** → **expectations:** rules/constraints

# Time Series Anomaly Detection

[Lawrence Wong, et al: AER: Auto-Encoder with Regression for Time Series Anomaly Detection. **IEEE BigData 2022**]



- **Basic Problem Formulation**

*Given a regular time series **detect anomalous subsequences** where the **observed pattern** cannot be well reconstructed **nor well predicted** by **models trained on normal data.***

- **Anomaly Detection**

  **#1 Reconstruction Error** (Auto-Encoder)

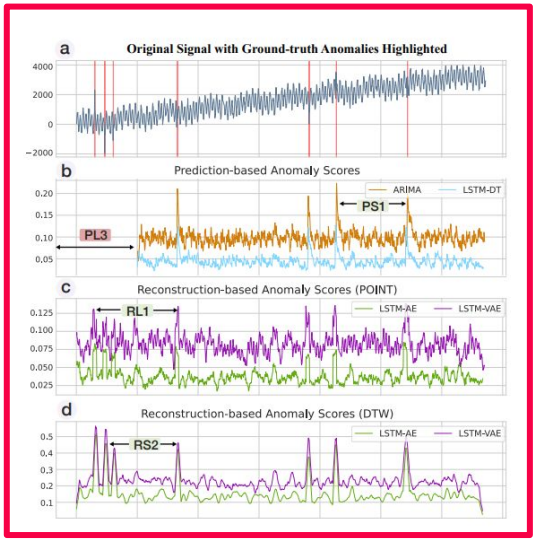
  Train an auto-encoder on normal time series windows

  **#2 Prediction Error** (Regression Model)

  Train a model (e.g., LSTM) to predict the next value.

  **#3 Combined Anomaly Score** (Error Auto-Encoder + Error LSTM)

  High score → anomalous subsequence.




HCC Institute of HUMAN-CENTRED COMPUTING

# Time Series Anomaly Detection

- **Recommended:**

# Automatic Data Repairs

- **Overview Repairs**
  - Question: Repair data, rules/constraints, or both?
  - General principle: **"minimality of repairs"**

[Xu Chu, Ihab F. Ilyas: Qualitative Data Cleaning. Tutorial, **PVLDB 2016**]


- **Example Data Repair**
  - Functional dependency **A→ B**
  - 

| A | B |
|---|---|
| 1 | 2 |
| 1 | 3 |
| 1 | 3 |
| 4 | 5 |

**OK, dist=1**

| A | B |
|---|---|
| 1 | 3 |
| 1 | 3 |
| 1 | 3 |
| 4 | 5 |

vs

| A | B |
|---|---|
| 1 | 2 |
| 1 | 2 |
| 1 | 2 |
| 4 | 5 |

vs

| A | B |
|---|---|
| 1 | 5 |
| 1 | 5 |
| 1 | 5 |
| 4 | 5 |

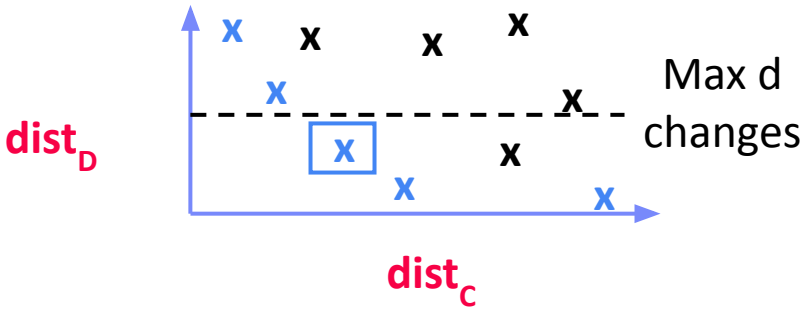
# Automatic Data/Rule Repairs, cont.

- **Example**
  - Expectation: **City → Country**; new data conflicts

[George Beskales, Ihab F. Ilyas, Lukasz Golab, Artur Galiullin: On the relative trust between inconsistent data and inaccurate constraints. **ICDE 2013**]

| IATA | ICAO | Name | City | Country |
|---|---|---|---|---|
| MEL | YMML | Melbourne International Airport | Melbourne | Australia |
| MLB | KMLB | Melbourne International Airport | **Melbourne** | USA |

- **Relative Trust: {FName, LName} → Salary**
  - **Trusted FD:** → change salary according to {FName, LName} → Salary
  - **Trusted Data:** → change FD to {FName, LName, DoB, Phone} → Salary
  - **Equally-trusted:** → change FD to {FName, LName, DoB} → Salary AND data accordingly



# Data Wrangling

- **Data Wrangler Overview**
  - **Interactive data cleaning** via spreadsheet-like interfaces
  - Iterative structure inference, recommendations, and data transformations
  - **Predictive interaction** (infer next steps from interaction)

[Vijayshankar Raman, Joseph M. Hellerstein: Potter's Wheel: An Interactive Data Cleaning System. **VLDB 2001**]

[Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, Jeffrey Heer: Wrangler: interactive visual specification of data transformation scripts. **CHI 2011**]

[Jeffrey Heer, Joseph M. Hellerstein, Sean Kandel: Predictive Interaction for Data Transformation. **CIDR 2015**]

- **Commercial/Free Tools**
  - **Trifacta** (from Data Wrangler)
  - **Google BigQuery + Looker Studio**
  - Microsoft Power BI
  - Tableau

TRIFACTA

PowerBI

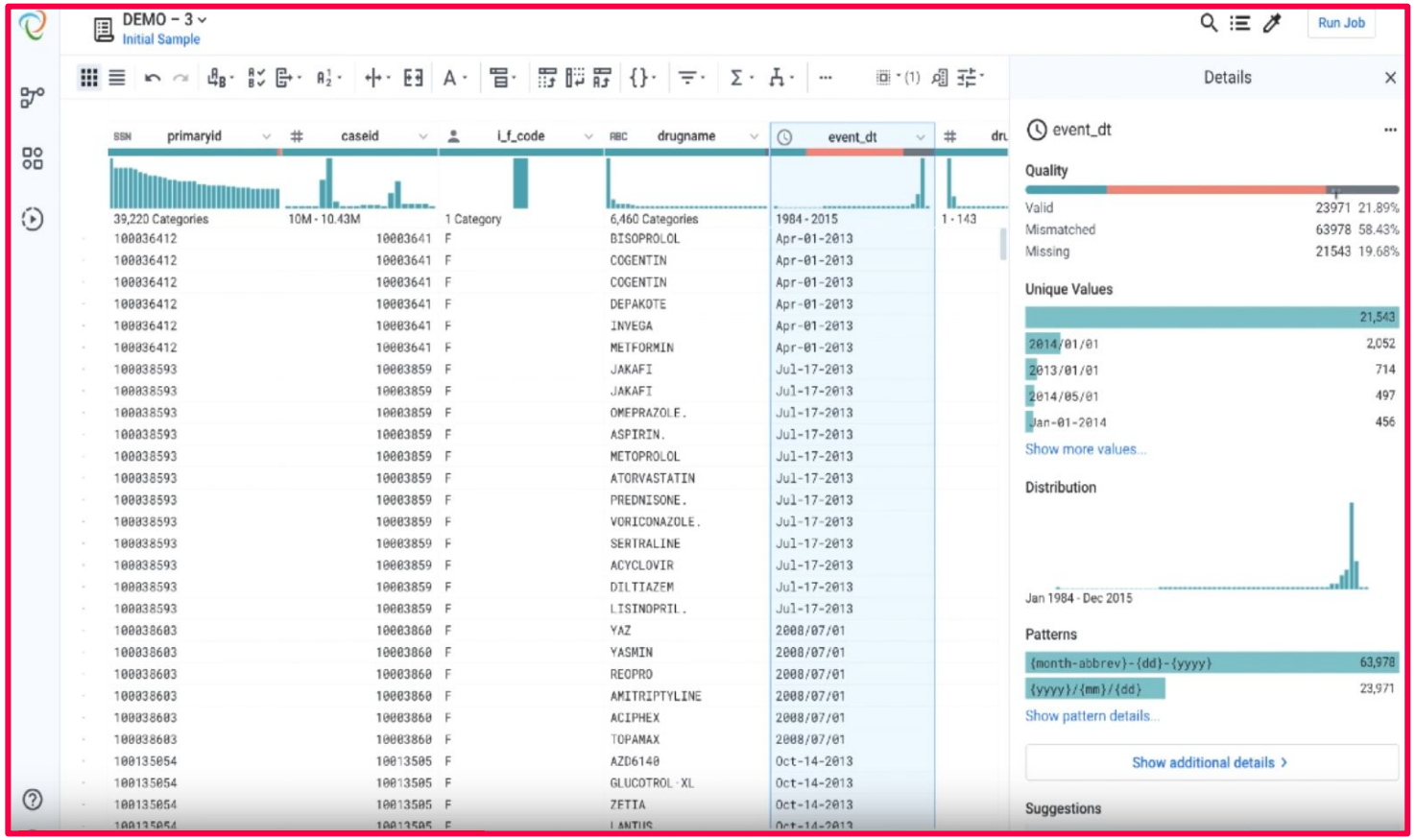
tableau

HCC Institute of HUMAN-CENTRED COMPUTING

# Data Wrangling, cont.

- Example: Trifacta Smart Cleaning

[**Credit:** Alex Chan (Apr 2, 2019) https://www.trifacta.com/blog/trifacta-for-data-quality-introducing-smart-cleaning/]



# Missing Value Imputation

# Basic Missing Value Imputation

- **Missing Value**
  - **Application context** defines if **0 is missing value or not**
  - If **differences between 0 and missing values**, use **NA or NaN**?
  - Could be a number outside the domain or symbol as **'?'**

- **Relationship to Data Cleaning**
  - Missing value is error, need to generate **data repair**
  - **Data imputation techniques** can be used as **outlier/anomaly detectors**

- **Recap: Reasons**
  - #1 **Heterogeneity of Data Sources**
  - #2 **Human Error**
  - #3 **Measurement/Processing Errors**

**MCAR:** Missing **Completely** at Random
**MAR:** Missing at Random
**MNAR:** Missing **Not at** Random

**missing data mechanisms (aka missingness)**

# Basic Missing Value Imputation

- **Missing Completely at Random**
  - Missing values are randomly distributed across all records (independent from recorded or missing values)
- **Missing at Random**
  - Missing values are randomly distributed within o**ne or more sub-groups of records**
  - Missing values depend on the recorded but not on the missing values, and **can be recovered**
- **Not Missing at Random**
  - Missing data depends on the missing values themselves
  - E.g., missing low salary, age, weight, etc.

| ID | Position | Salary ($) | |
|---|---|---|---|
| 1 | Manager | **null** | (3500) |
| 2 | Secretary | 2200 | |
| 3 | Manager | 3600 | |
| 4 | Technician | **null** | (2400) |
| 5 | Technician | 2500 | |
| 6 | Secretary | **null** | (2000) |

| ID | Position | Salary ($) |
|---|---|---|
| 1 | Manager | 3500 |
| 2 | Secretary | 2200 |
| 3 | Manager | 3600 |
| 4 | **Technician** | **null** |
| 5 | **Technician** | **null** |
| 6 | Secretary | 2000 |

| ID | Position | Salary ($) |
|---|---|---|
| 1 | Manager | 3500 |
| 2 | Secretary | **null** |
| 3 | Manager | 3600 |
| 4 | Technician | **null** |
| 5 | Technician | 2500 |
| 6 | Secretary | **null** |

<= 2400
missing

HCC Institute of HUMAN-CENTRED COMPUTING

# Basic Missing Value Imputation, cont.

- **Basic Value Imputation (for MCAR)**
  - **General-purpose:** replace by user-specified constant, or drop records, or one-hot encode as separate column
  - **Continuous variables:** replace by mean, median
  - **Categorical variables:** replace by mode (most frequent category)

- **Iterative Algorithms (chained-equation imputation for MAR)**
  - Train ML model on available data to predict missing information
    - Initialize with basic imputation (e.g., mean)
    - One dirty variable at a time
    - Feature k → label, split data into training: observed / scoring: missing
    - Types: categorical → classification, continuous → regression
  - Noise reduction: train models over feature subsets + averaging

[Stef van Buuren, Karin Groothuis-Oudshoorn: mice: Multivariate Imputation by Chained Equations in R, **J. of Stat. Software 2011**]

# Basic Missing Value Imputation, cont.

[More details: 1:15' to 1:23' last year lecture]

- **MICE example**
  - Initialization: **fill in the missing values with column mean (w/ or w/o NAs)**
  - Iterations: **each column per iteration**

| V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|
| 1 | 56 | 2 | 2 | 2 |
| 2 | 23 | 0 | 0 | 0 |
| 1 | NA | 0 | 0 | 2 |
| 2 | 24 | -1 | 2 | NA |
| NA | 22 | 1 | 2 | 0 |

| V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|
| 1 | 56 | 2 | 2 | 2 |
| 2 | 23 | 0 | 0 | 0 |
| 1 | 25 | 0 | 0 | 2 |
| 2 | 24 | -1 | 2 | 0.8 |
| 1.2 | 22 | 1 | 2 | 0 |

train(y) train(x)

| V1 | V2 | V3 | V4 | V5 |
|---|---|---|---|---|
| 1 | 56 | 2 | 2 | 2 |
| 2 | 23 | 0 | 0 | 0 |
| 1 | 25 | 0 | 0 | 2 |
| 2 | 24 | -1 | 2 | 0.8 |
| ? | 22 | 1 | 2 | 0 |

← test(x)

# DNN Based MV Imputation

[More details: 1:15' to 1:23' last year lecture]

[Felix Bießmann et al: DataWig: Missing Value Imputation for Tables, **J. of ML Research 2019**]

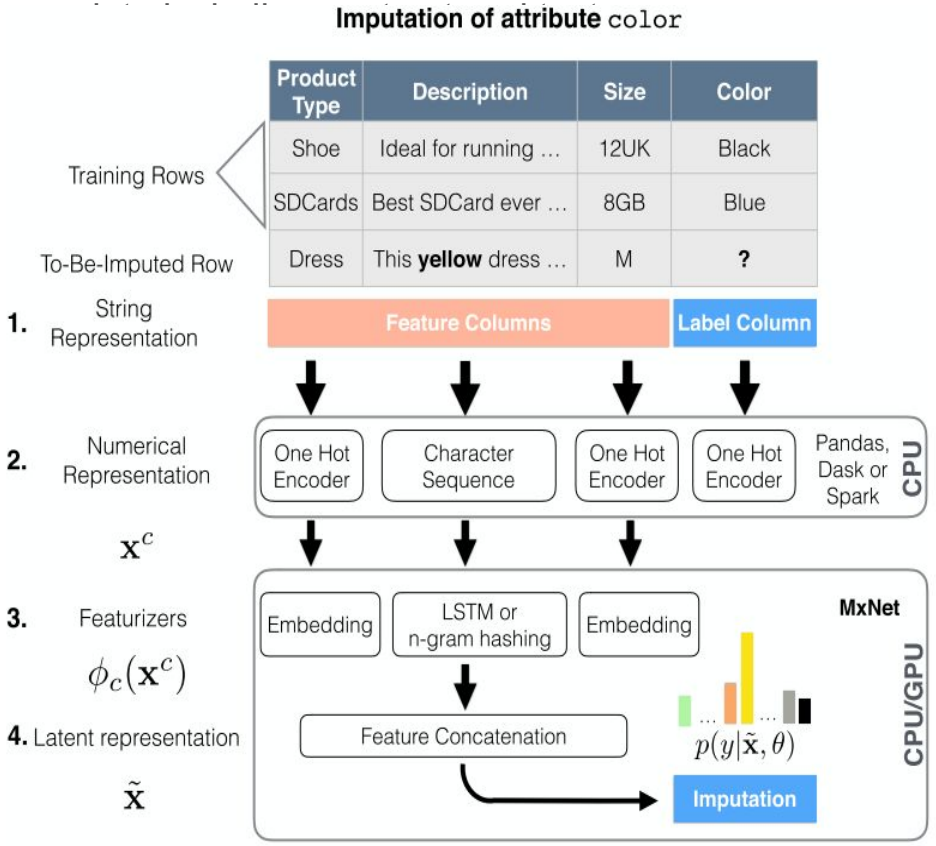
- DataWig
  - Missing values imputation for heteroge

| Data Type | Featurizers | Loss |
|---|---|---|
| Numerical | Normalization Neural Network | Regression |
| Categorical | Embeddings | Softmax |
| Text | Bag-of-Words LSTM | N/A |

```
table = pandas.read_csv('products.csv')
missing = table[table['color'].isnull()]

# instantiate model and train imputer
model = SimpleImputer(
          input_columns=['description',
                         'product_type',
                         'size'],
          output_columns=['color'])
        .fit(table)

# impute missing values
imputed = model.predict(missing)
```



# Time Series Imputation

[More details: 1:15' to 1:23' last year lecture]

[Steffen Moritz and Thomas Bartz-Beielstein: imputeTS: Time Series Missing Value Imputation in R, **The R Journal 2017**]

- Example R Package imputeTS

| Function | Option | Description |
|---|---|---|
| na.interpolation | | |
| | linear | Imputation by Linear Interpolation |
| | spline | Imputation by Spline Interpolation |
| | stine | Imputation by Stineman Interpolation |
| na.kalman | | |
| | StructTS | Imputation by Structural Model & Kalman Smoothing |
| | auto.arima | Imputation by ARIMA State Space Representation & Kalman Sm. |
| na.locf | | |
| | locf | Imputation by Last Observation Carried Forward |
| | nocb | Imputation by Next Observation Carried Backward |
| na.ma | | |
| | simple | Missing Value Imputation by Simple Moving Average |
| | linear | Missing Value Imputation by Linear Weighted Moving Average |
| | exponential | Missing Value Imputation by Exponential Weighted Moving Average |
| na.mean | | |
| | mean | MissingValue Imputation by Mean Value |
| | median | Missing Value Imputation by Median Value |
| | mode | Missing Value Imputation by Mode Value |
| na.random | | Missing Value Imputation by Random Sample |
| na.replace | | Replace Missing Values by a Defined Value |

# Summary and Q&A

- **Motivation and Terminology**
- **Data Cleaning and Fusion**
- **Missing Value Imputation**
- **Recommended extra-reading:**
  - **SAGA: A Scalable Framework for Optimizing Data Cleaning Pipelines for Machine Learning Applications [S. Siddiqi et al 2023]**

- Next Lectures (**Part B**)
  - **08 Cloud Computing Foundations** [Nov 21st]