Data Engineering for Data Analytics

Chris Williams

The Alan Turing Institute

January 2020

Figure credit: Aaron Bacall

Towards Automating the Data Analytics Process

- Common view that up to 80% of work on a data mining project is involved in **data understanding** and **data preparation**
- The Artificial Intelligence for Data Analytics (AIDA) project asks: What can we do to try to improve that?
- AIDA team: Taha Ceritli, James Geddes, Ernesto Jiménez-Ruiz, Ian Horrocks, Alfredo Nazabal, Tomas Petricek, Charles Sutton, Gerrit Van Den Burg

CRISP-DM Methodology

Cross Industry Standard Process for Data Mining

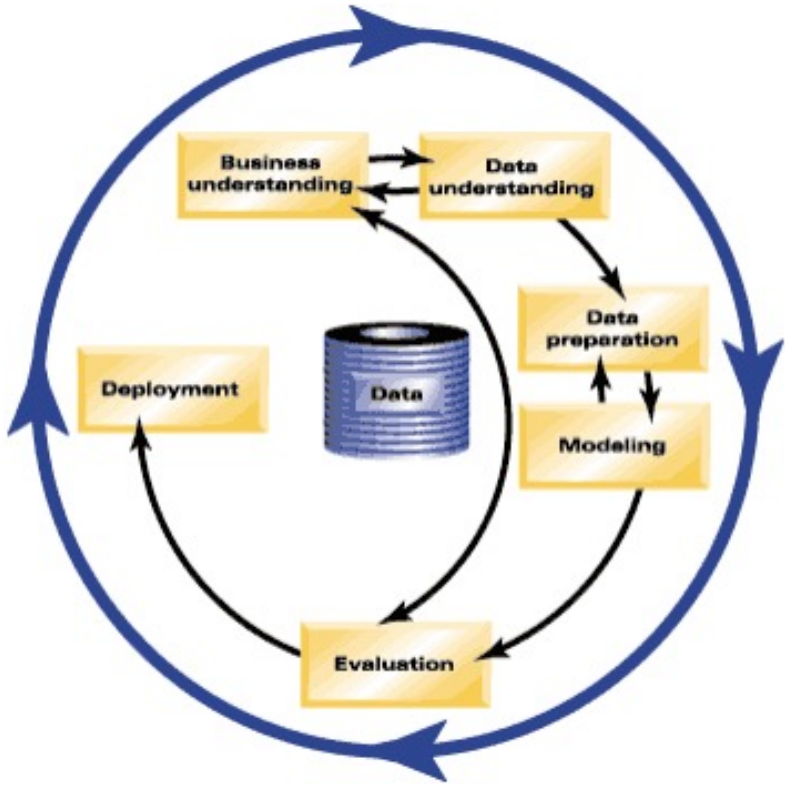


Figure credit: CRISP-DM 1.0 Step-by-step data mining guide, Chapman et al, 2000

Data Wrangling Challenges

1. Data Organization
   - 1.1 **Data parsing**, e.g. converting csv's or tables
   - 1.2 Obtaining (or inferring) a **data dictionary**: basic types + semantics
   - 1.3 **Data integration**: Combining data from multiple sources
   - 1.4 **Data transformation**: e.g. wide vs tall format.
2. Data Quality
   - 2.1 **Canonicalization**: e.g. format variability for dates, but also in names (e.g. IBM, I.B.M.)
   - 2.2 **Missing data**: identification and repair
   - 2.3 **Anomaly detection** and repair
   - 2.4 **Non-stationarity**, e.g. changepoints
3. Feature Engineering

AIDA overview

Multidisciplinary work from machine learning, semantic technologies and programming languages, to:

- Build AI tools for individual steps
- Build an open source platform prototype (Wrattler)
- Integrate components into the prototype
- Provide exemplar use cases with a clear analysis task, along with raw and cleaned-up data

1.1 Data Parsing

- Identifying the structure of the raw data source so it can be read properly
- Detecting CSV formats (van den Berg, Nazabal, Sutton 2019)
- Format detection: Column delimiter, quote character, escape character

Mango; £365,14; £1692,64
Apple; £2568,62; £1183,78
Lemon; £51,65; £685,67
Orange; £1760,75; £128,14
Maple; £880,86; £323,43

- Choosing semicolon, space, comma or £ as the field separator above results in different tables, all with three columns
- Dataset of $\sim$ 10,000 CSV files "in natural habitat"

Search for format to maximize

- Pattern score: uniformity of rows by pattern
  Mango; £365,14; £1692,64
  CDCDC
- Format score: uniformity of columns by type

Method achieves 97% overall accuracy on a large corpus of real-world CSV files, and improves the accuracy on messy CSV files by almost 22% compared to existing approaches

Code: `https://github.com/alan-turing-institute/CSV_Wrangling`

1.2 Data Dictionary

- Understanding the data format: names and types of each field in a table
- The *Data Dictionary* should provide this information, but in reality that this information is often out-of-date or incomplete
- Information can be obtained both from the data in the table, and external sources (e.g. wikidata)
- AIDA work: ptype – a probabilistic method for inferring the syntactic type of a column
- AIDA work: CoINet – column semantic type
- Important to learn (and carry over knowledge) from previous datasets

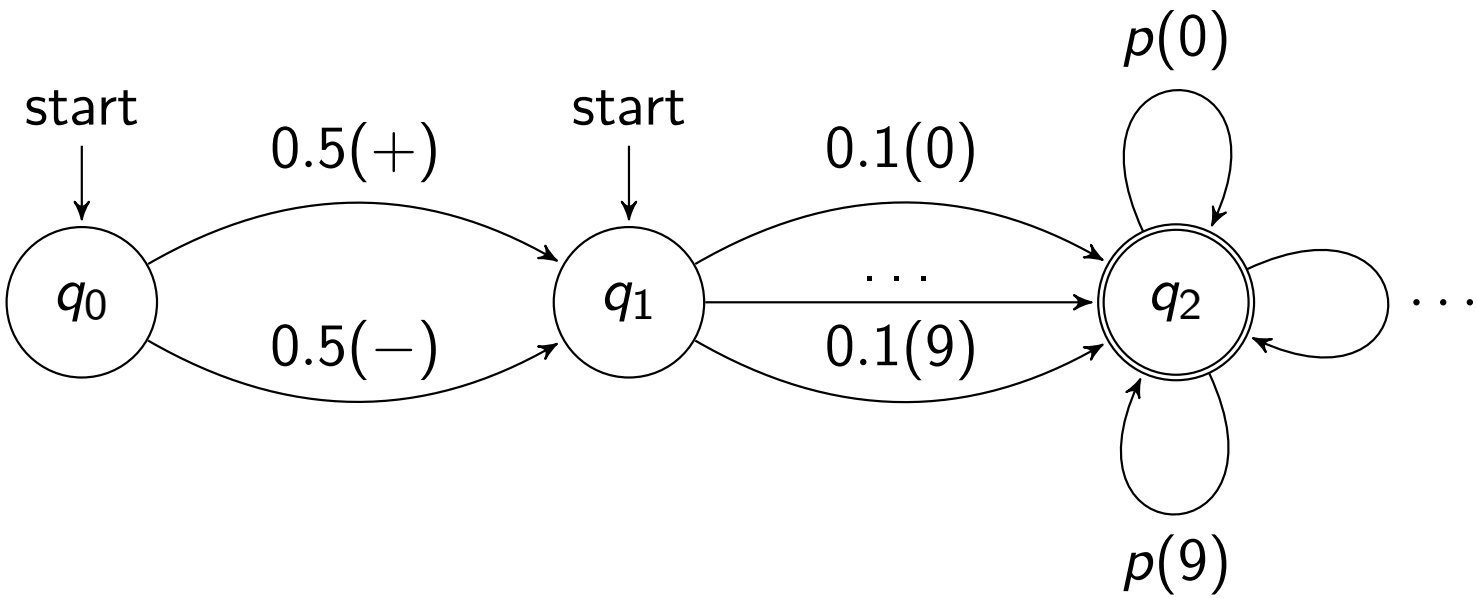
ptype

Ceritli, Williams and Geddes (2019)
Detecting data types in the presence of missing and anomalous data

Normal, missing, and anomalous values are denoted by green, yellow, and red, respectively.

- ptype builds a Probabilistic Finite-State Machine (PFSM) for each type (e.g. boolean, date, float, integer, string)
- It also has types for missing data (common identifiers like NA etc) and anomalous data (any string)
- We then do probabilistic inference for the type of each column, and identify missing/anomalous entries

PFSM for integers



ColNet

Chen, Jiménez-Ruiz, Horrocks, Sutton (AAAI, 2019)

- Task is to predict the semantic type of a column of data
- E.g. a column consisting of "Mute swan", "Eider duck" and "Wandering albatross" is annotated as dbo:Species and dbo:Bird, two classes of DBpedia
- First step: *lookup*. Retrieve column cells' corresponding entities in the KB, and return the classes of the matched entities as a set of candidate classes for annotation
- Second step: *prediction*. Calculate a score for each candidate class of the given column using a customized binary CNN classifier
- ColNet outperforms state-of-the-art approaches

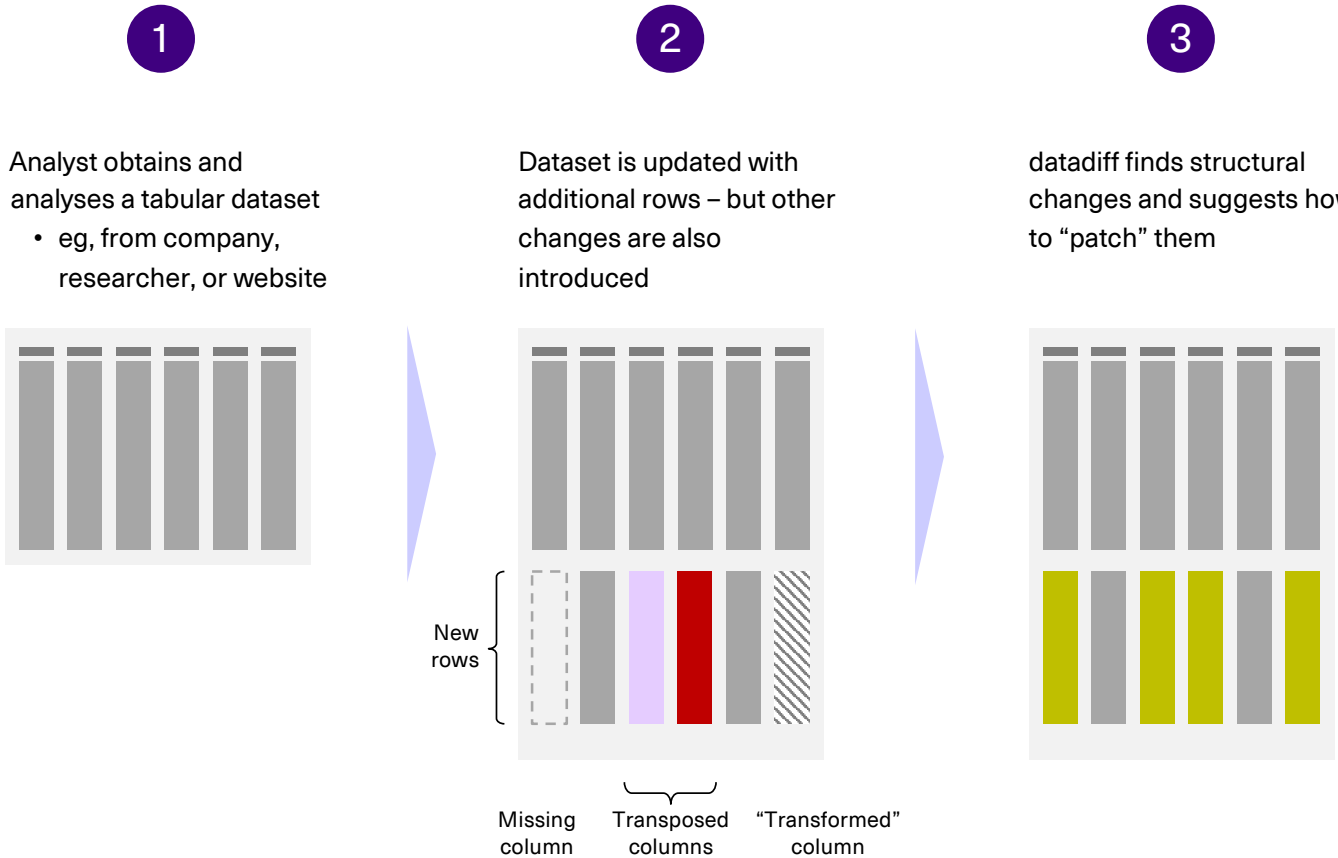
1.3 Data Integration

Combining data from multiple sources

- Record linkage and table joining (column-wise aggregation)
  - Relatively straightforward if primary keys exist in both tables
  - If explicit keys are not available, may face a *probabilistic* record linkage problem
- Table unioning (row-wise aggregation)
  - Problems may arise if table structure changes between installments, e.g. datadiff
- Heterogenous integration, e.g. tables, time series etc.

Datadiff

Caruana, Hobson, Geddes, Sutton (KDD 2018)



- Is this month's data *really* the same format as last month?
- Very often not. Small niggling changes break data wrangling scripts

diffs between data sets

A **diff** is a simple transformation which makes two *distributions* the same, e.g.:

- Swap columns
- Insert or delete columns
- Linear transformation (e.g. change of units)
- Recode categorical data
  e.g. perhaps coding of special data values (like NA) has changed
- Distribution change in a column

Goal: interpretable summary of differences

Choosing the patch

Search through space of patches P to minimize

$$\min_P \ \mathrm{KS}(D_1, P(D_2))$$

Algorithm

- For all pairs of columns (i in D_1, j in D_2) find the best single variable patch from i to j with weight w_{ij}
- Permutation: find the best bipartitie matching with weights W (Hungarian algorithm)

Threshold to prune small differences, calibrated by bootstrap

Example: UK broadband performance in 2014 & 2015

Data: https://data.gov.uk/dataset/uk-fixed-line-broadband-performance

Urban.rural <chr>	Market <int>	Download.speed..Mbit.s..24.hrs <dbl>	Upload.speed..Mbit.s.24.hour <dbl>	DNS.resolution..ms.24.hour <dbl>	Latency..ms.24.hour <dbl>	Web.page..ms.24.hour <dbl>
Urban	3	46.9844	3.115	15.712	20.101	281.602
Urban	3	45.5636	14.967	24.337	23.882	328.972
Urban	3	6.9502	0.806	37.816	40.412	973.295
Urban	3	5.1803	0.893	22.514	21.828	1077.498
Urban	3	32.6965	5.638	13.418	12.891	236.983
Urban	2	36.9341	9.503	11.042	7.669	197.890
Urban	3	59.4303	11.455	7.638	6.569	168.984
Rural	2	23.0231	6.813	24.871	24.711	366.380
Urban	3	62.1146	3.014	12.194	14.267	237.197

LLUvsNon <chr>	MarketClass <chr>	URBAN2 <chr>	DL24hrmean <dbl>	UL24hrmean <dbl>	DNSRTT24hr <dbl>	Latency24hr <dbl>	Web24hr <dbl>	iPlayerStartupDelay24hr <chr>
Y	1	Rural	19.0244719	1.0805485	19.212286	14.098829	279.7595	1225
Y	3	Urban	41.5836412	10.1946693	16.820604	12.383474	192.4834	1455
Y	3	Urban	27.7508452	1.8494538	11.994468	13.608900	430.3546	1033
Y	1	Semi-urban	12.6604958	0.7403905	55.853831	48.619955	1052.2461	2782
Y	2	Semi-urban	28.1385701	8.2658529	23.069865	18.367009	382.6215	#NULL!
Y	2	Semi-urban	36.9236555	9.2734170	14.922790	10.379316	179.8799	1109
Y	2	Semi-urban	37.3616813	9.3963233	14.476586	9.650034	246.3469	849
Y	1	Rural	35.9497322	9.2386704	20.976396	15.670669	246.3126	1299
N	Hull	Urban	18.5523579	0.7483720	28.181700	28.474457	397.8410	#NULL!

- Output patch: $p = p_4 \circ p_3 \circ p_2 \circ p_1$
  - p_1 recodes column 2
  - p_2 deletes column 9
  - p_3 deletes column 1
  - p_4 swaps columns 1 & 2

1.4 Data Transformation

- The tables we have may not be what we want, need to re-format
- Tidy data (Hadley Wickham, 2014)
  - Each variable forms a column
  - Each observation forms a row
  - Each type of observational unit forms a table
- Tidy Data, Hadley Wickham, J. Statistical Software 59(10), 2014
- May need to carry out *information extraction* from text to obtain features, e.g. named entity recognition

Tidy data: example

religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $75–100k, $100–150k and >150k, have been omitted.

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10–20k	34
Agnostic	$20–30k	60
Agnostic	$30–40k	81
Agnostic	$40–50k	76
Agnostic	$50–75k	137
Agnostic	$75–100k	122
Agnostic	$100–150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Figure credit: H. Wickham

2. Data Quality

2.1 Canonicalization

- Converts entities that have more than one possible representation into a *canonical* format
- Example: "UK", "U.K." and "United Kingdom" refer to the same entity
- "18/7/16" vs "18 July 2016"
- Data may contain entires like 70 litres, 4 cu. ft. ?
- Above examples need string matching, and special purpose handling of dates etc.
- Openrefine `openrefine.org` provides some good functionality ("clustering") for such tasks
- More advanced methods have knowledge of the types of abbreviations etc. used in a given domain
- Can also make use of reference data sources (e.g. wikidata)

2.2 Missing data

- Lack of a value for a variable within an observation.
- **Detection**: Might be coded as "NA" or a value that is inconsistent with the type of the attribute (e.g. "NaN")
- But can be coded as e.g. 0 or -99 when it is not clear if this value is inconsistent ... DANGER!! (disguised missing values, Pearson, 2006)
- **Understanding** Why is data missing? Is it *missing at random* (MAR) or is there a systematic reason for its absence?

Repair of missing data

Let X_m denote those values missing, and X_p those values that are present in the data matrix X

If MAR, some "solutions" are to *impute* the missing data

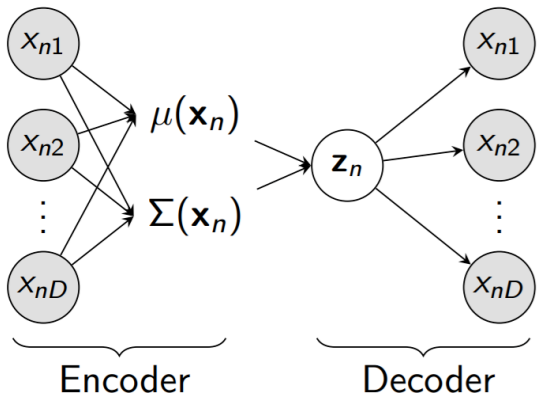
- Model $p(X_m|X_p)$ and average (correct, but hard)
  Can e.g. use sophisticated autoencoder models
- Replace missing data in each column with its global mean or modal value (crude)
- Look for similar (close) input patterns and use them to infer missing values (crude version of density model)
- How important is imputation for the analysis task?
- Reference: *Statistical Analysis with Missing Data* R. J. A. Little, D. B. Rubin, Wiley (1987)

2.3 Anomaly Detection

- "An anomaly is defined as a pattern that does not conform to expected normal behavior" (Chandola, Banerjee and Kumar, 2009)
- Can be at the level of the whole record, or a field in the record
- May arise because an error has occurred in data measurement or transmission; but may also arise from correct measurement of an unusual situation
- Usually handled by building a model of normality, and detecting low probability events/records/observations (outliers)

Robust Variational Autoencoders for Outlier Detection in Mixed-Type Data

Variational Autoencoder (Kingma and Welling, 2014)



Data row $\mathbf{x}_n \in \mathbb{R}^D$

Latent code $\mathbf{z}_n \in \mathbb{R}^K$

- Probabilistic model of the data (decoder)

$$p_\theta(\mathbf{x}_n, \mathbf{z}_n) = p(\mathbf{z}_n) \prod_{d=1}^{D} p_\theta(x_{nd}|\mathbf{z}_n), \qquad p(\mathbf{z}_n) \sim N(\mathbf{0}, I)$$

- Approximate posterior (encoder)

$$q_\phi(\mathbf{z}_n|\mathbf{x}_n) = \mathcal{N}(\mathbf{z}_n|\mu(\mathbf{x}_n), \Sigma(\mathbf{x}_n))$$

Robust variational auotencoder (RVAE)

Eduardo, Nazabal, Williams, Sutton (2019)

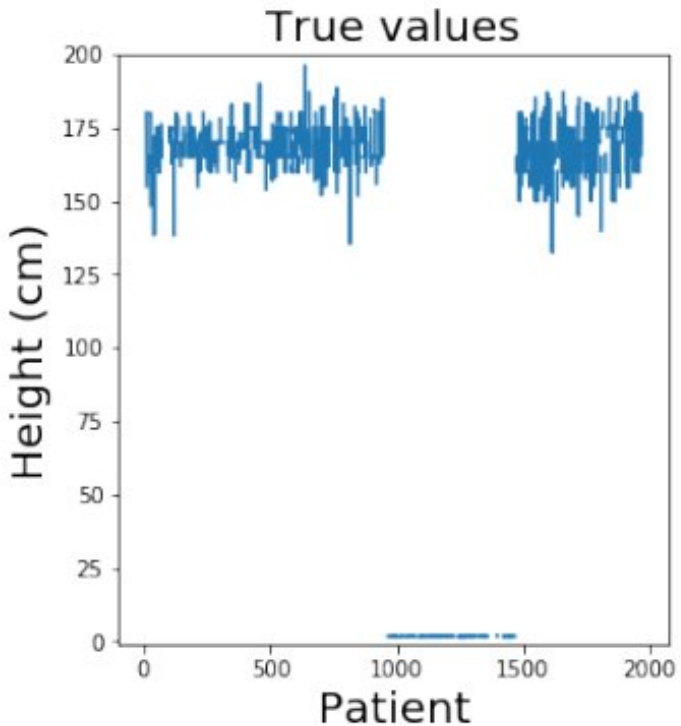
Introduce a Bernoulli switch variable w_{nd} to specify clean vs noisy

$$p_\theta(\mathbf{x}_n, \mathbf{z}_n, \mathbf{w}_n) = p(\mathbf{z}_n)p(\mathbf{w}_n) \prod_{d=1}^{D} p_\theta(x_{nd}|\mathbf{z}_n)^{w_{nd}} p_0(x_{nd})^{1-w_{nd}}$$

- $p_0(x_{nd})$ is a simple noise model (uniform or broad Gaussian)
- Factorized q: $q_{\phi,\pi}(\mathbf{w}_n, \mathbf{z}_n|\mathbf{x}_n) = q_\phi(\mathbf{z}_n|\mathbf{x}_n) \prod_{d=1}^{D} q_\pi(w_{nd}|\mathbf{x}_n)$
- Cell outlier score is obtained as $\hat{\pi}_{nd} = \mathbb{E}[w_{nd}|\mathbf{x}_n]$
- RVAE performs better than several state of the art methods in cell outlier detection, while providing comparable or better results for row outlier detection

2.4 Non-stationarity

- Need to watch out for changes (drifts or jumps) in the data distribution
- Detect change points, possibly due to protocol changes
- What is going on in this HCM column of the CleanEHR dataset?



Example Analyses

- Tundra Traits
- Household Electricity Survey
- CleanEHR: Electronic Health Records
- Broadband

Example: Tundra Traits (Bjorkman et al, *Nature*, 2018)

- data dictionary, e.g. meaning of fields in the data tables
- data integration of climate and plant data
- identification and removal of anomalous data
- handling of missing data

Discussion

- The messiness of real data is a huge barrier to unlocking its potential
- There are a diverse set of issues covering data acquisition and transformation, data understanding and ingestion, data quality and cleaning
- Issues cover tabular data, but also time series, images, graph data etc
- We need to address these problems to build robust ML systems

AIDA References

Available from
`https://www.turing.ac.uk/research/research-projects/artificial-intelligence-data-analytics`

- Data Diff: Interpretable, Executable Summaries of Changes in Distribution s for Data Wrangling. C. Sutton, T. Hobson, J. Geddes and R. Caruana. KDD 2018.
- ColNet: Embedding the Semantics of Web Tables for Column Type Prediction. Jiaoyan Chen, Ernesto Jimenez-Ruiz, Ian Horrocks and Charles Sutton. AAAI, 2019
- G. J. J. van den Burg, A. Nazabal, C. Sutton, Wrangling Messy CSV Files by Detecting Row and Type Patterns. *Data Mining and Knowledge Discovery*, 33(6) 1799–1820, 2019.
- Taha Ceritli, Christopher K. I. Williams, James Geddes. ptype: Probabilistic Type Inference, arXiv 1911.10081, 2019
- S. Eduardo, A. Nazabal, C. K. I. Williams, C. Sutton. Robust Variational Autoencoders for Outlier Detection in Mixed-Type Data, arXiv 1907.06671, 2019