Solving the 80% Problem Data Prep for Microsoft AI

Euan Garden
Data & Analytics Nerd
@euanga
https://github.com/euanga/VS_Live_0319

Level: Intermediate

Agenda

- What is data prep and why is it such a pain
- How do we solve it
- How is Microsoft helping you solve it

Big Data Borat
@BigDataBorat

Following

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

Jenny Bryan
@JennyBryan

Following

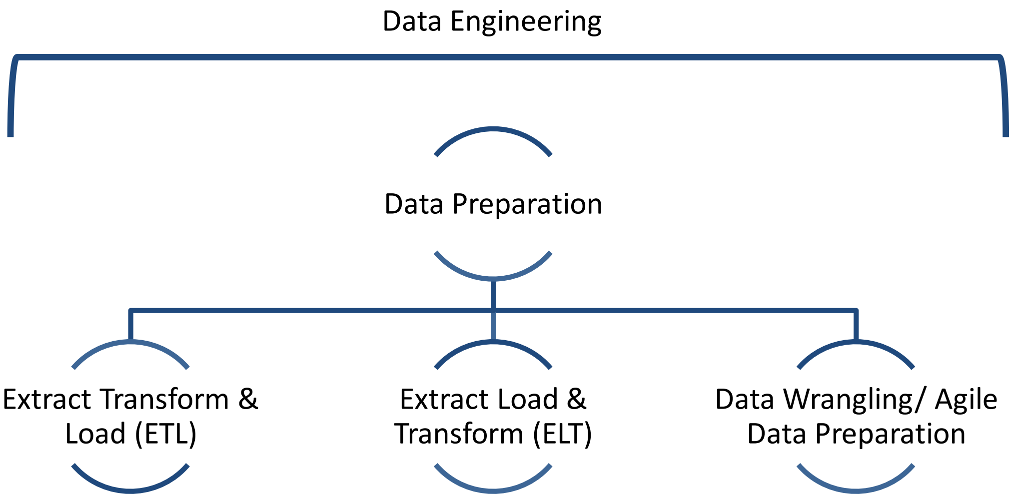
channeling my inner Churchill:
CSV is the worst form of transparent, tool-agnostic file format, except for all the others

> **inconvergent** @inconvergent
> CSV is not a format. it is a donkey with a sharp stick taped to its forehead. the stick can be any length, and it's not always a donkey.

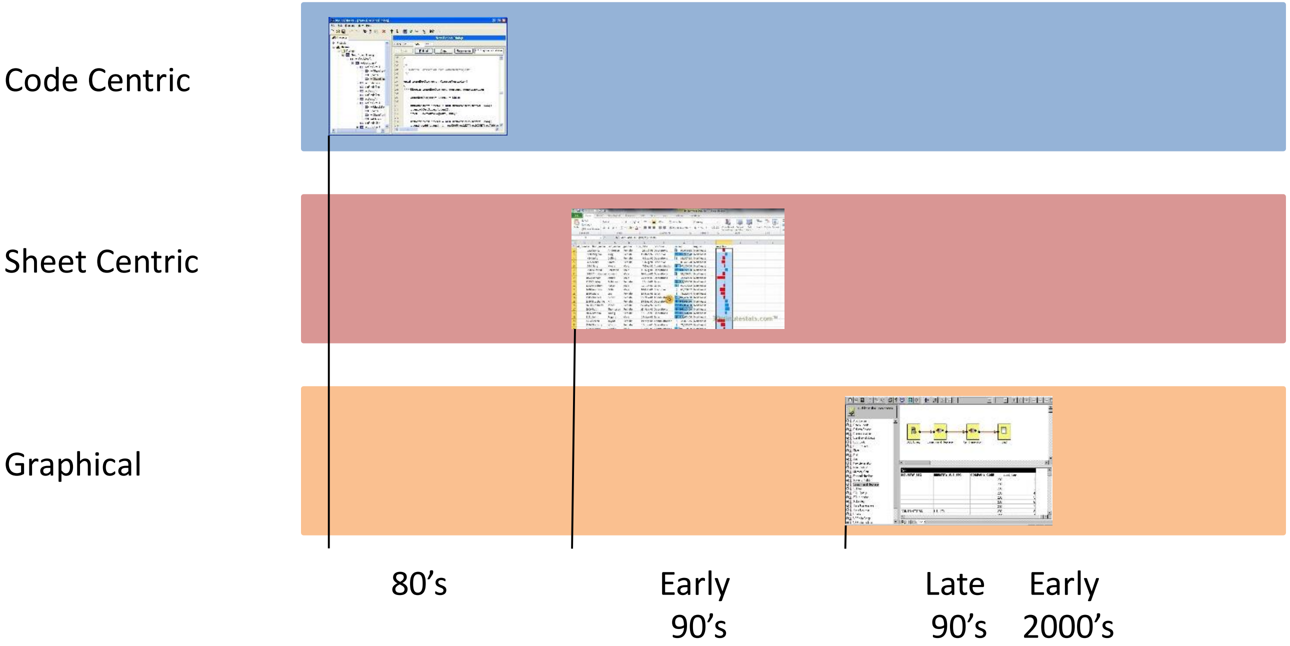
Joel Grus
@joelgrus

"Data science is a god-like power."
"Right, have you finished munging those CSVs yet?"
"No, they have time zone data in them!"

Defining Data Preparation



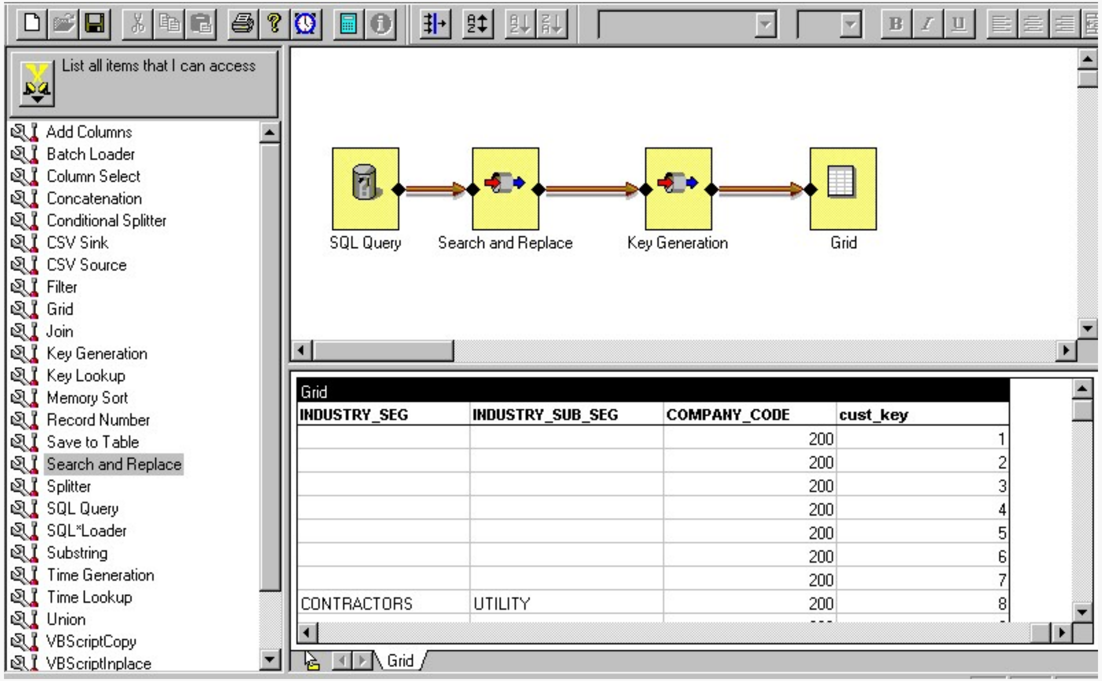
A Data Preparation Journey

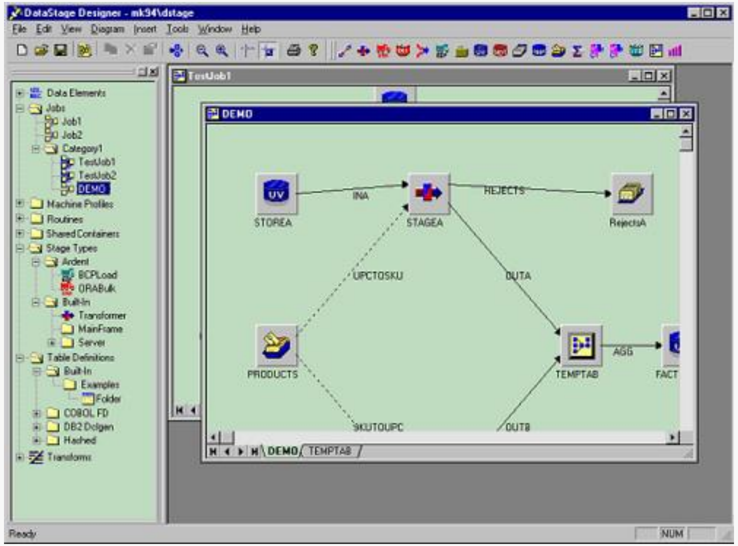


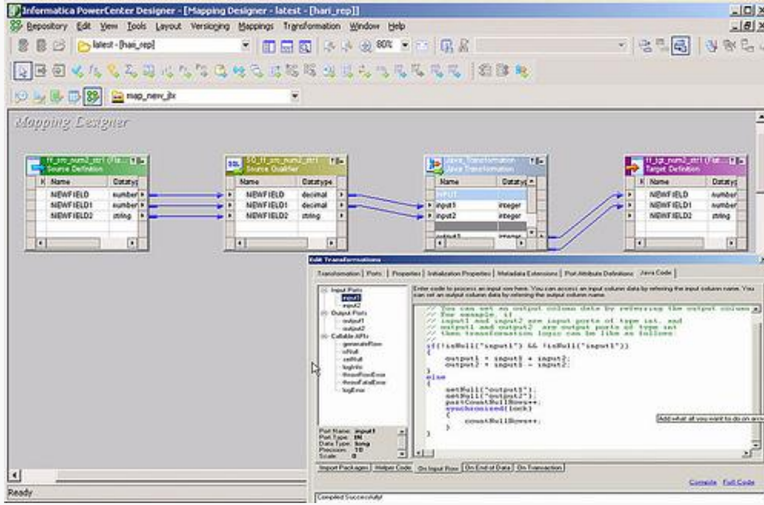
70% of Enterprise DW spent on data integration
- Ralph Kimball/Bill Inmon et al

Data Preparation experience has not changed in 20+ years

- "Box and Line" ETL
- Command line
- Discover problems using other tools, fix using ETT/ETL/ELT tools
- Known schema to known schema



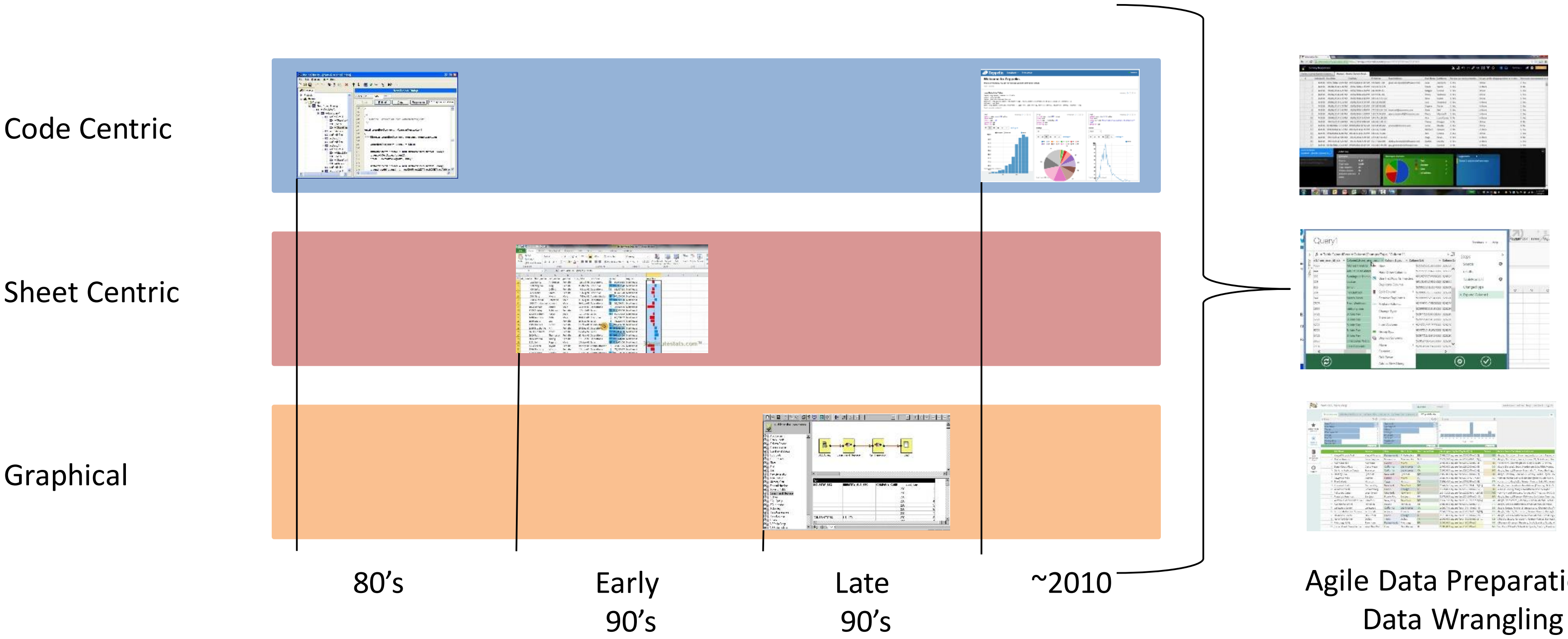




Challenges with historic approaches

- WORK WELL FOR SCHEMA TO SCHEMA, KNOWN TARGET AND SOURCE MAPPING
- INVOLVE LOTS OF CUSTOM CODE
- LIMITED ACCESS TO CUSTOM LIBRARIES
- "BATCHY", RUN, WAIT FOR COMPLETION, DEBUG CYCLE
- ATTACHED TO SCALE UP NOT SCALE OUT ARCHITECTURES
- NOT CLOUD FRIENDLY

Data Preparation today



70% of Enterprise DW spent on data integration
- Ralph Kimball/Bill Inmon et al

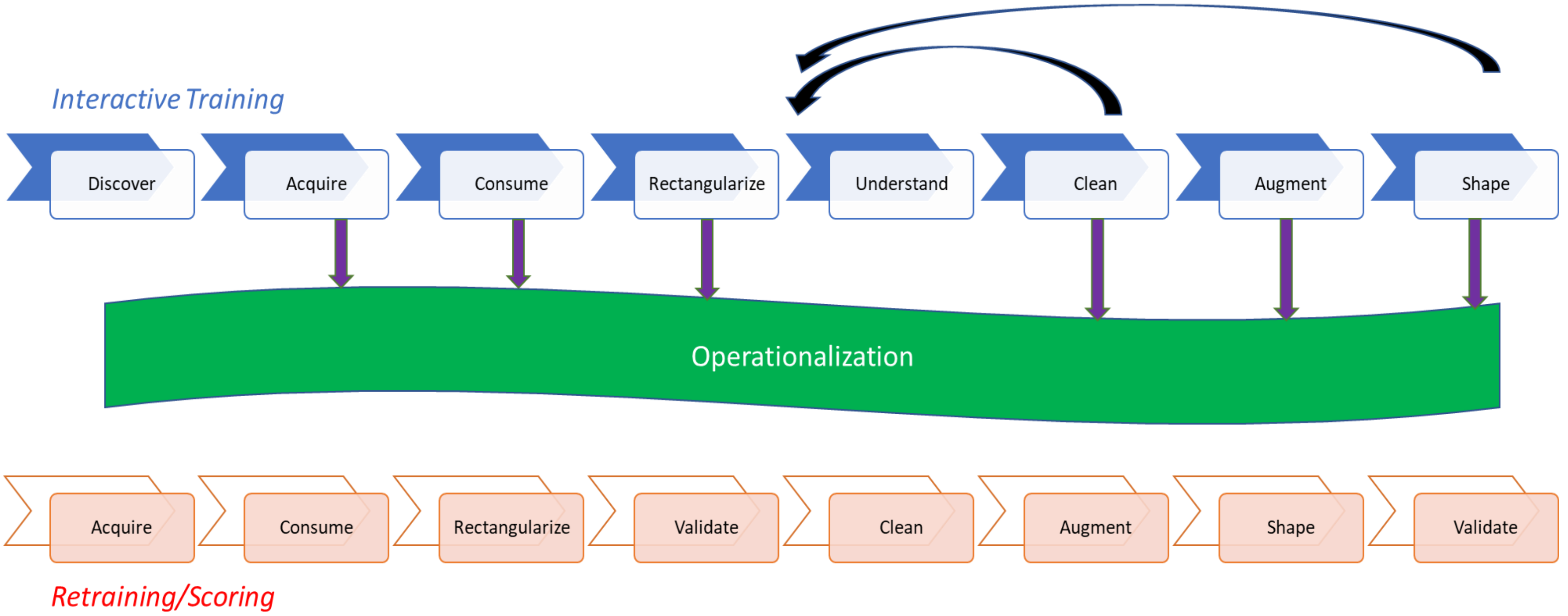
80-90% of Analytic Apps budget on Data prep
- Forrester et al

The act of manipulating raw data into a form that makes it relevant and valuable for consumption by ML algorithms

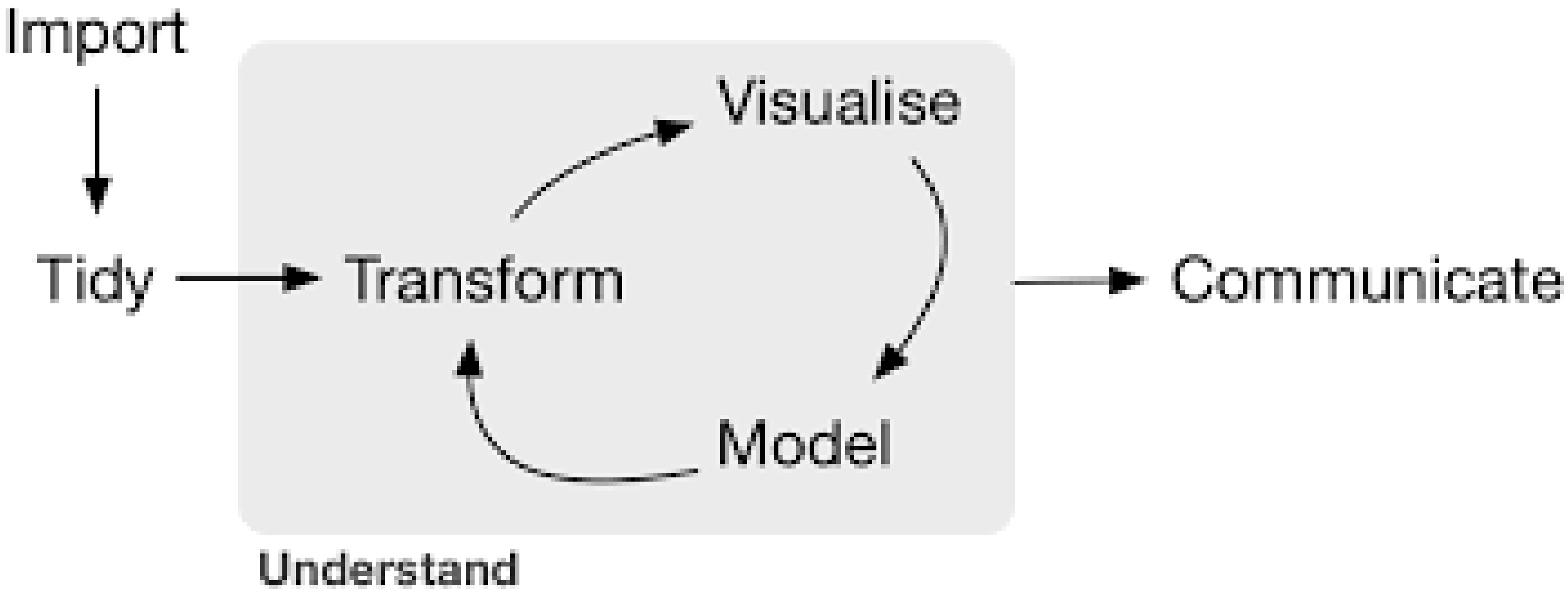
Customer Challenges and Pain Points

- Understanding the semantics of Data is hard and time consuming
- Merging data from different sources is too manual
- Detecting, troubleshooting and fixing errors is a high tax
- Lots of manual, non-scalable work
  - Data Formatting
  - Dealing with Dates
  - “Rectangualising” Data
- Custom code always required
- Operationalization is HARD

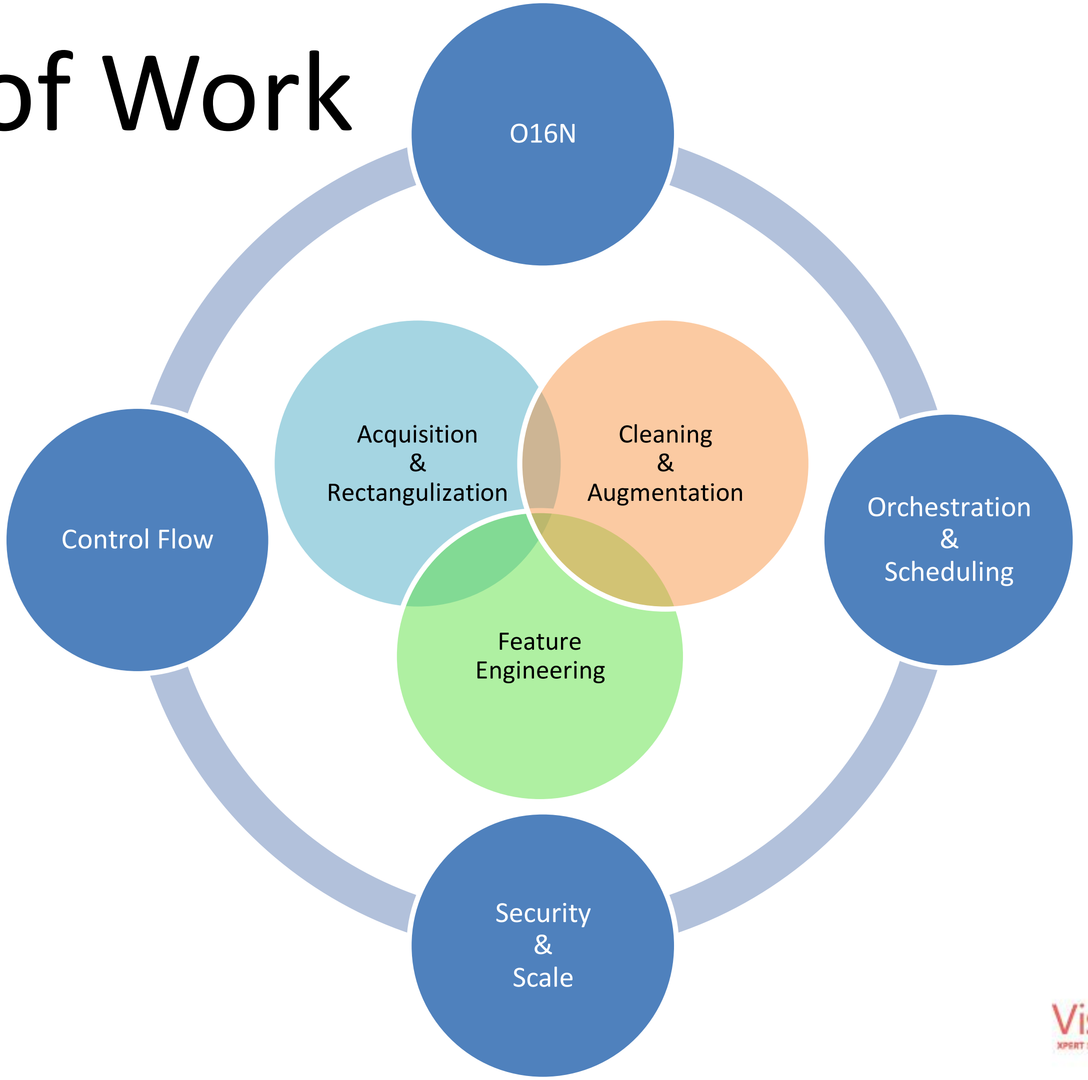
Data lifecycle



R Community ("TidyVerse") View



Taxonomy of Work



An EDA/Data Prep 10 Step Process

- Acquire
- Rectangularize
  - Tidy Format, Pivot/UnPivot, Inconsistent Schema
- Data Type & Format/Range verification and assertion
- Explore & Understand
  - Univariate and Multivariate, ToDo List
- Missingness/Inconsistency
  - Skew, Special/Magic/Sentinel Values, Common Sense
  - Regranularisation, Units of Measure
  - Imputation
- Outliers
- Derived Columns
- Augmentation & Aggregation
  - Join, Synthetic Data
- ML Specific Feature engineering
  - Scaling, Encoding, Binning, Feature reduction, different versions for different algorithmic consumers
- Prepare for consumption
  - Training/Test Split

10 Principles

- *At any/all stages attempt to model & use visualization to check progress,*
- *Use business understanding to review value of data against requirements*
- *Discover the history/journey/lineage of the data you have*
- *Stay iterative & interactive*
- *Filter/Aggregate early*
- *Join/Union late*
- *Drop Columns as early as possible*
- *Drop NA's as late as possible*
- *Trust no-one!*
- *Embrace Experimentation and Failure*

Demo Time

But what about…

- Scale
  - Sampling
  - Stats vs Actual Data
  - Visualization
  - Parralellism
- Operationalisation
  - Be defensive
    - Package versioning
  - Orchestration/Pipelining
  - Training Data Prep <> Inferencing/Scoring Data Prep, potentially
  - Monitor for Drift/Divergence
  - Dev Ops

- *“The idea of imputation is both seductive and dangerous” (R.J.A Little & D.B. Rubin)*
- **Imputation vs Removing Data**
- Before jumping to the methods of data imputation, we have to understand the reason why data goes missing.
- **Missing at Random (MAR):** Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
- **Missing Completely at Random (MCAR):** The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
- **Missing not at Random (MNAR):** Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable’s value (e.g. Let’s assume that females generally don’t want to reveal their ages! Here the missing value in age variable is impacted by gender variable)
- Deletion vs Imputation
  - Deletion
    - Rows, Columns, Pairwise or more decision to delete
  - Imputation
    - Time Series vs Categorical vs Continuous
    - MICE vs LR vs KNN

@euanga

https://github.com/euanga/VS_Live_0319