Visual Studio LIVE!
XPERT SOLUTIONS FOR ENTERPRISE DEVELOPERS
Las Vegas

Solving the 80% Problem Data Prep for Microsoft AI

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

Level: Intermediate

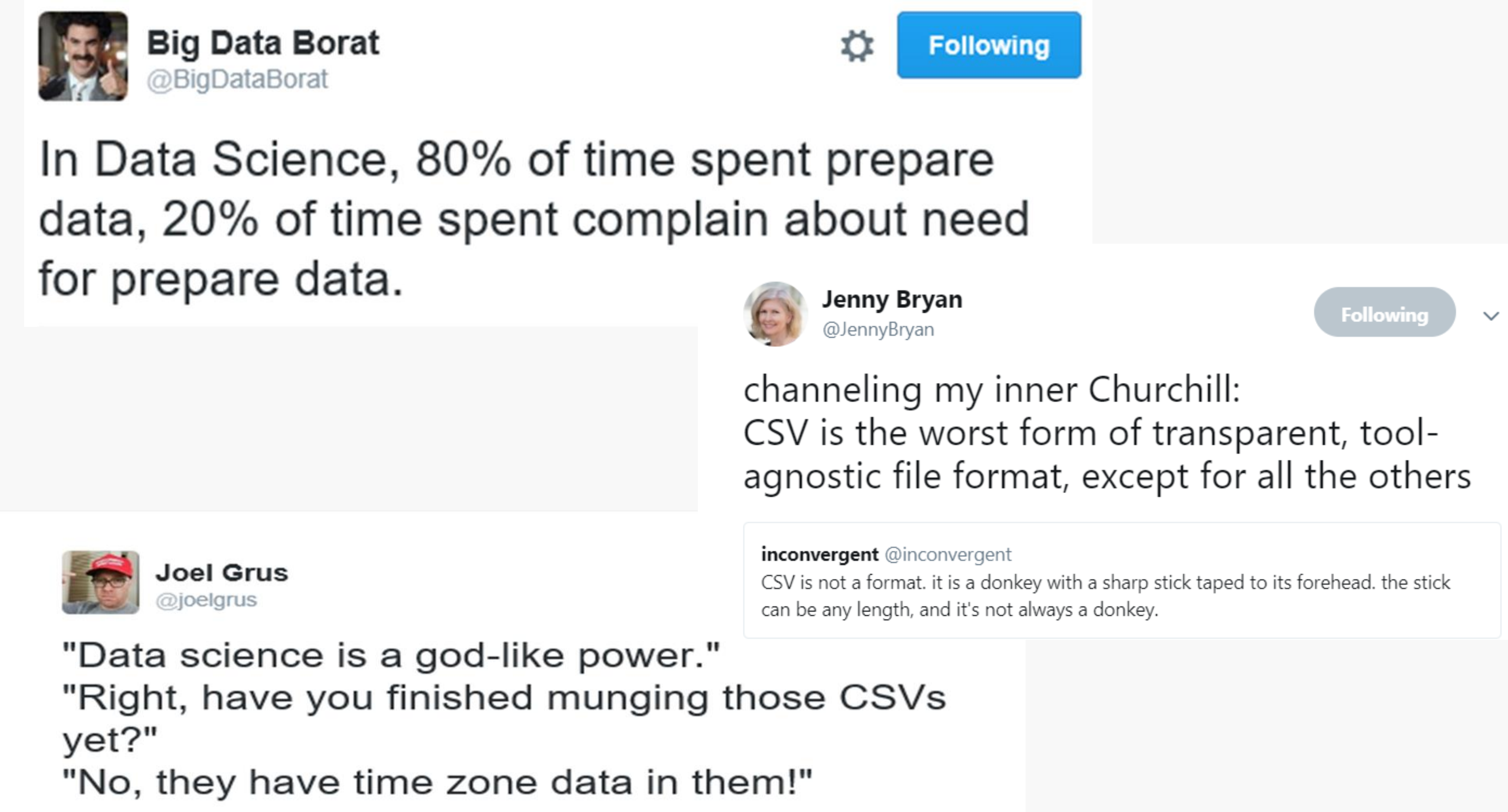
#VSLIVE

Goals

- Scare you a little!
- Provide a framework for thinking about the problem and the solution
- Add tools and techniques to your toolbox

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

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

Jenny Bryan
@JennyBryan

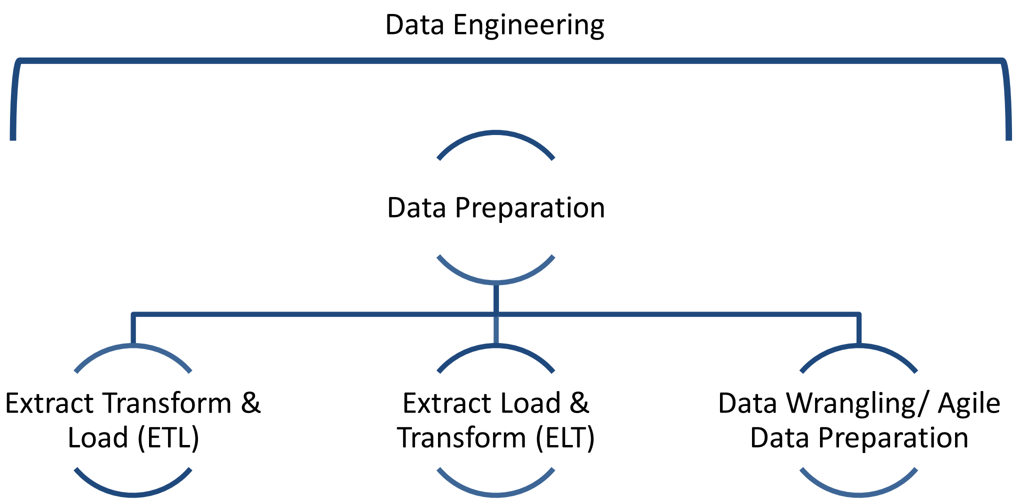
channeling my inner Churchill:
CSV is the worst form of transparent, tool-agnostic file format, except for all the others

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



Data Engineering

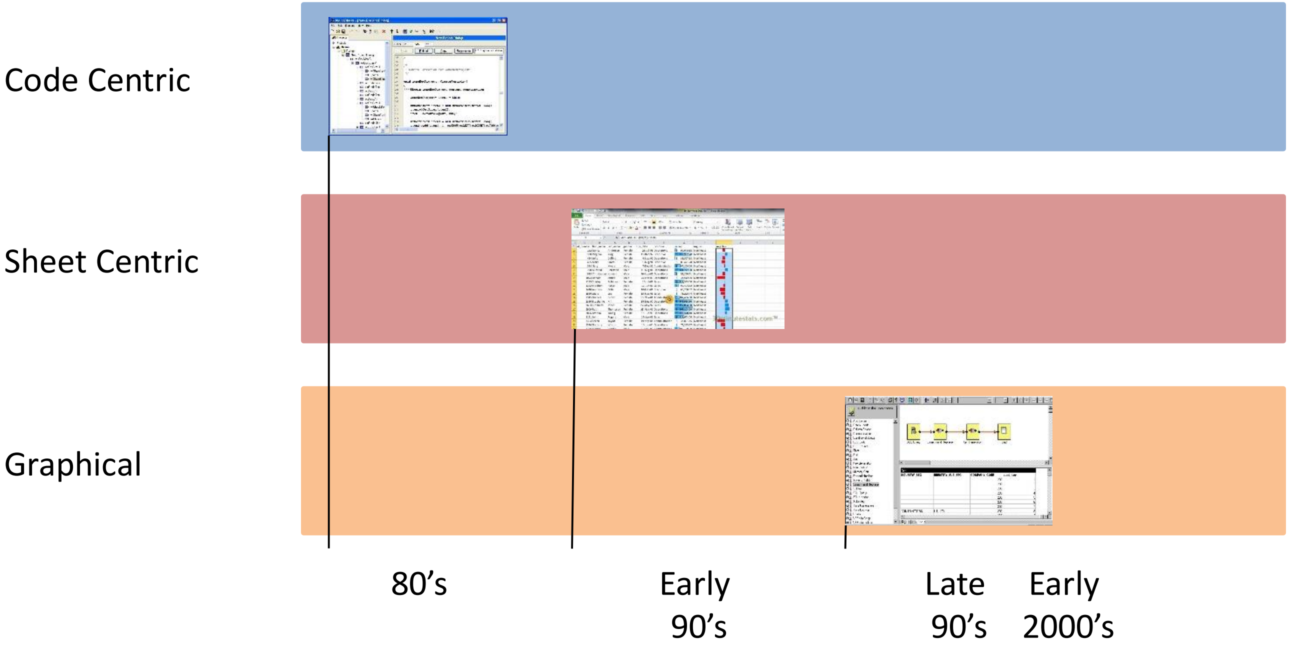
Data Preparation

Extract Transform & Load (ETL)

Extract Load & Transform (ELT)

Data Wrangling/ Agile Data Preparation

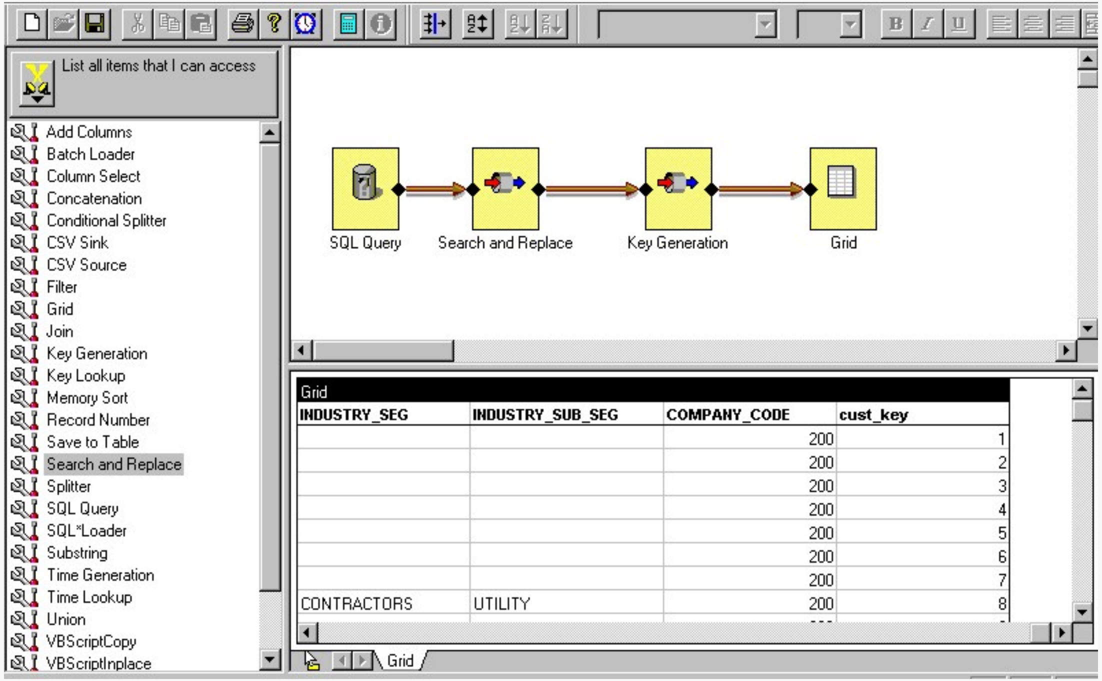
A Data Preparation Journey



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

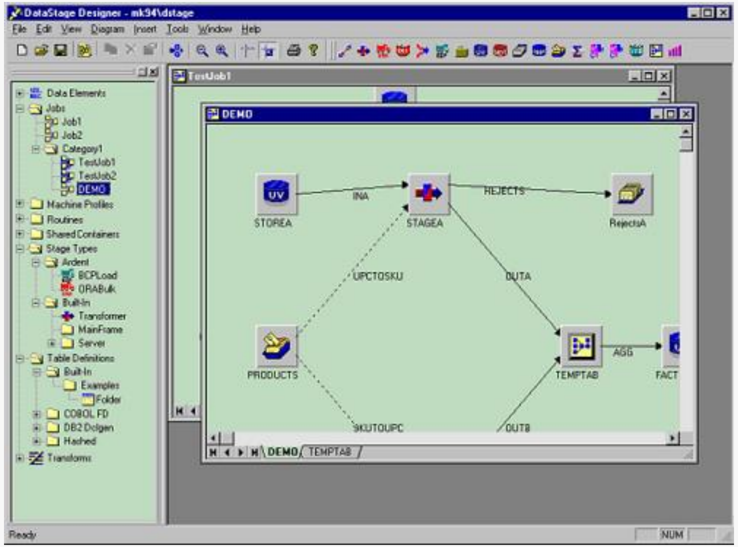
Data Preparation experience has not changed in 20+ years

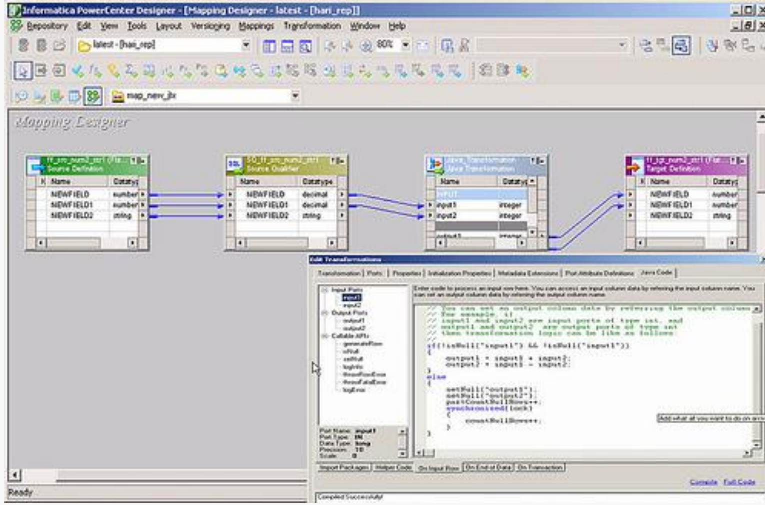
- "Box and Line" ETL
- Command line
- Discover problems using other tools, fix using ETT/ETL/ELT tools
- Known schema to known schema



Grid

INDUSTRY_SEG	INDUSTRY_SUB_SEG	COMPANY_CODE	cust_key
		200	1
		200	2
		200	3
		200	4
		200	5
		200	6
		200	7
CONTRACTORS	UTILITY	200	8

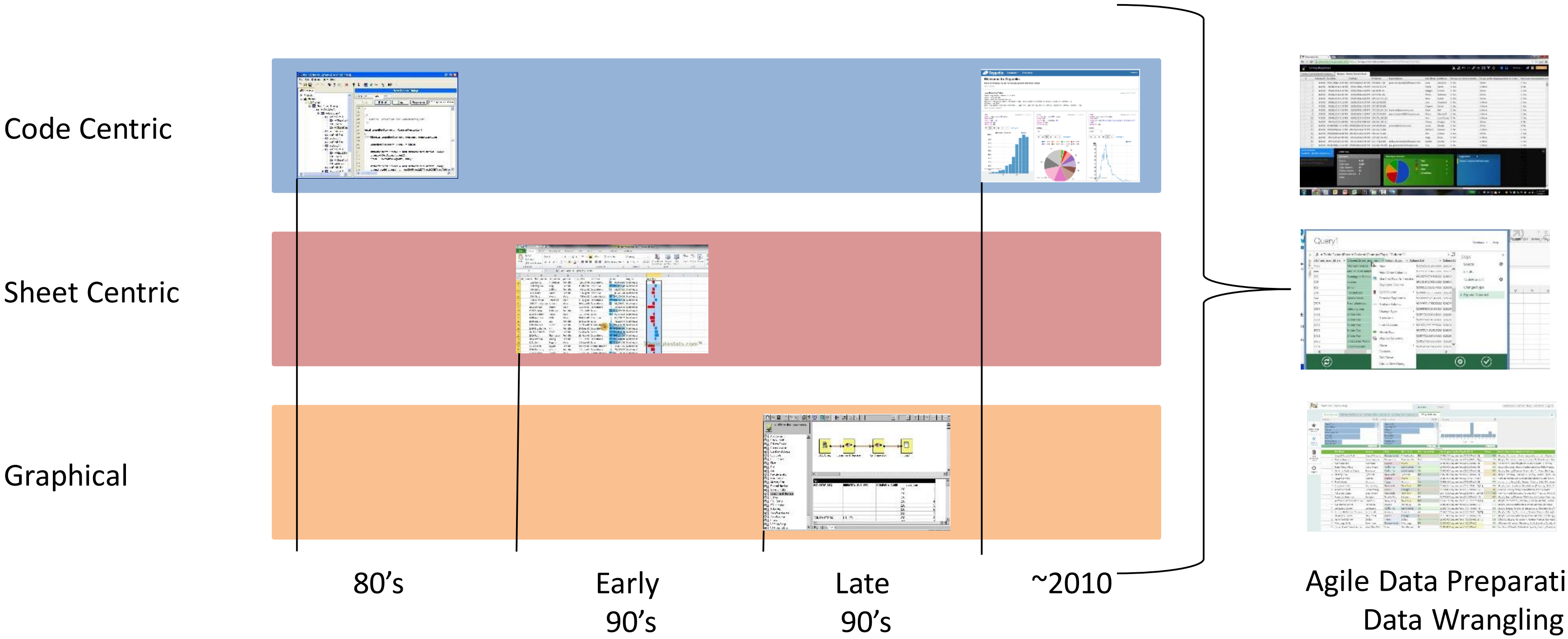




Challenges with historic approaches

- 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



Code Centric

Sheet Centric

Graphical

80's

Early 90's

Late 90's

~2010

Agile Data Preparation / Data Wrangling

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

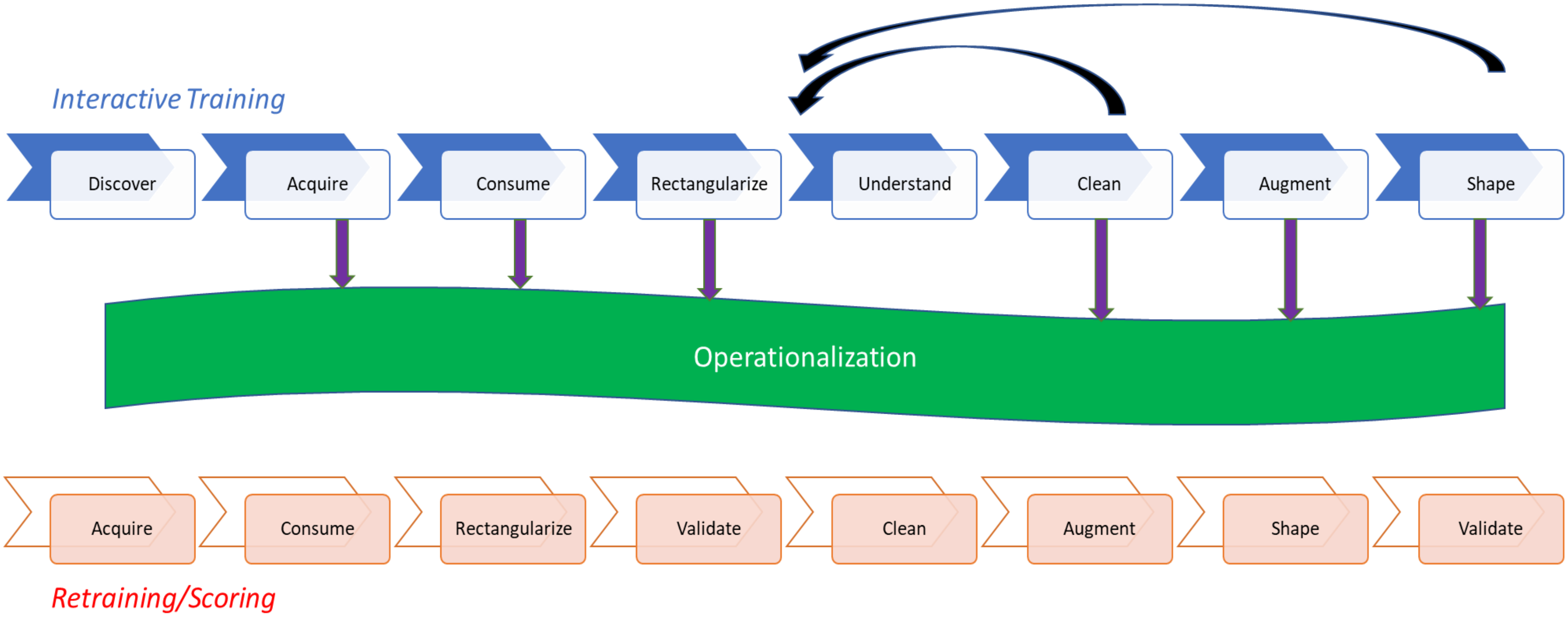
80-90% of Analytic Apps budget on Data prep
- Forrester et al

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

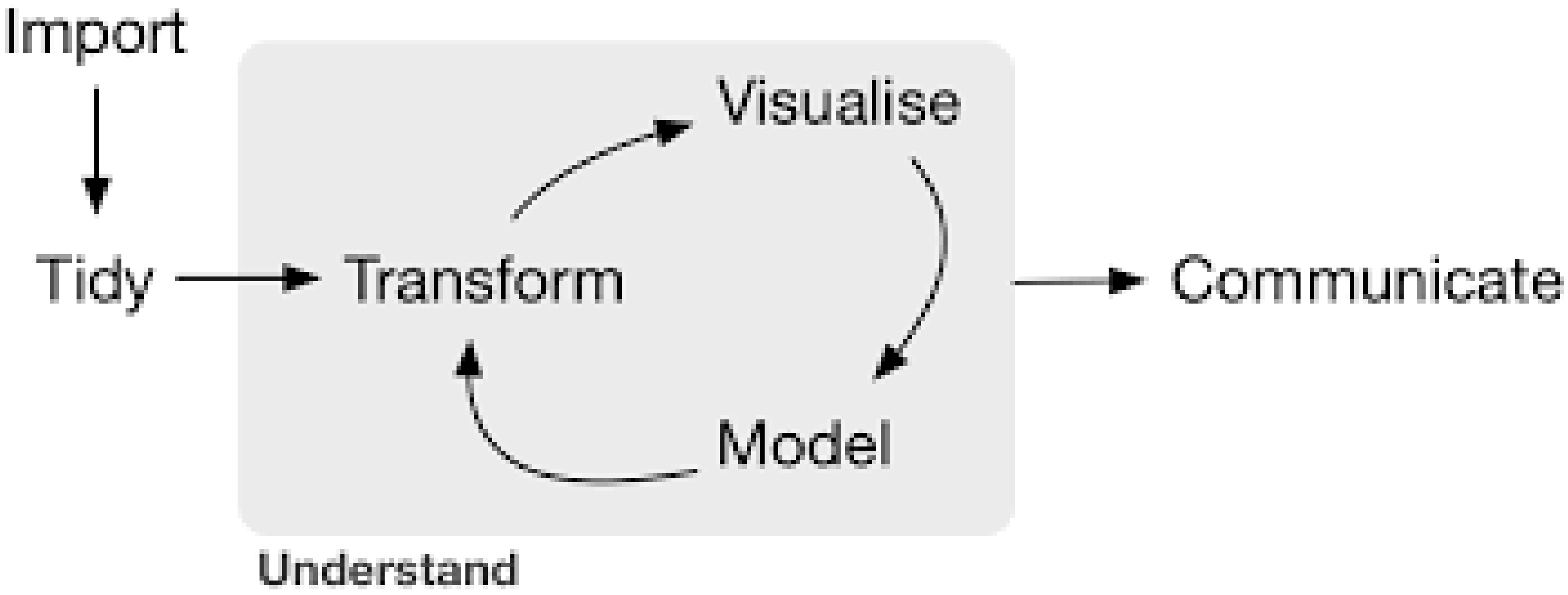
Customer Challenges and Pain Points

- 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

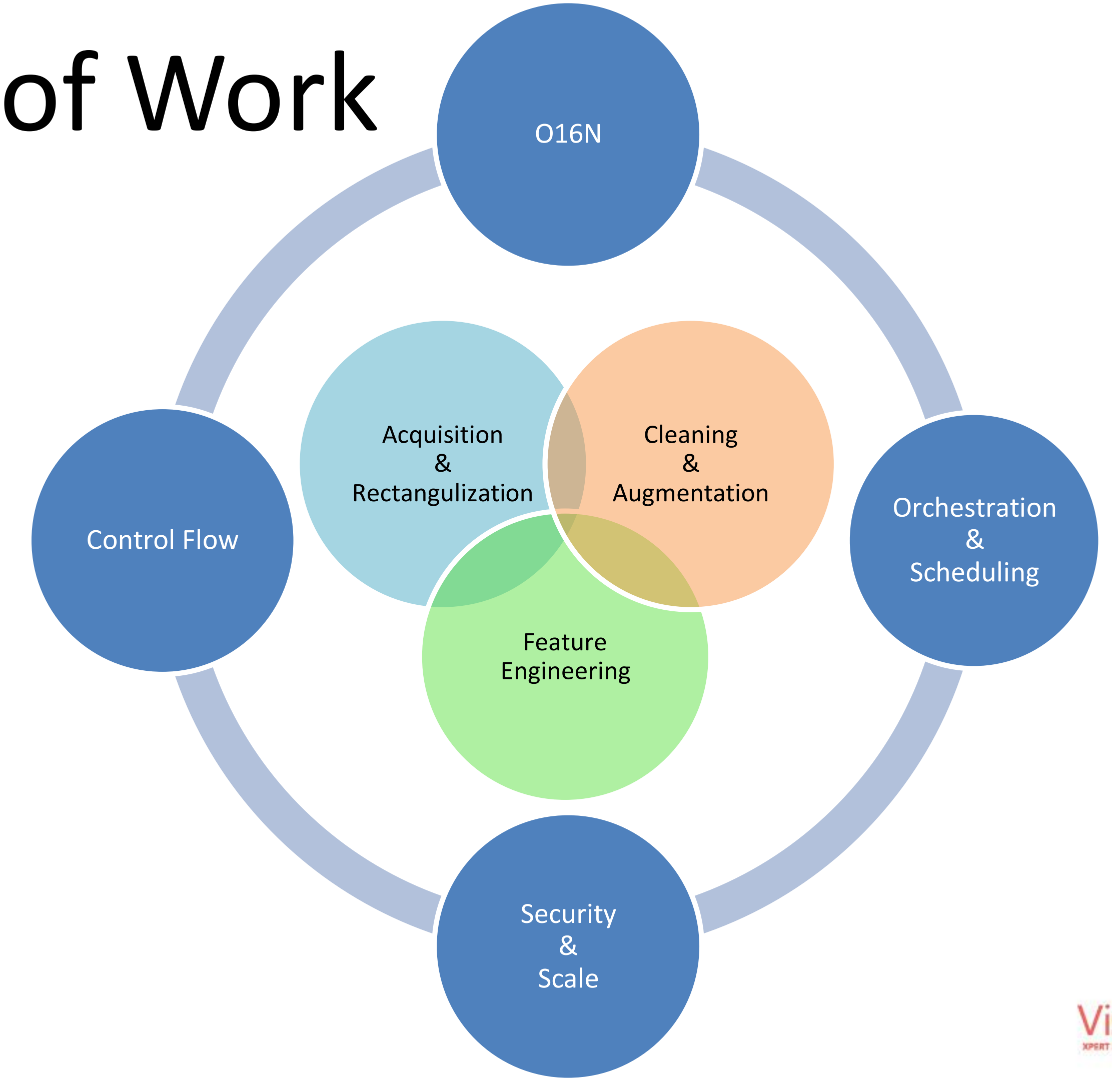
Data lifecycle



R Community (“TidyVerse”) View



Taxonomy of Work



An EDA/Data Prep Checklist

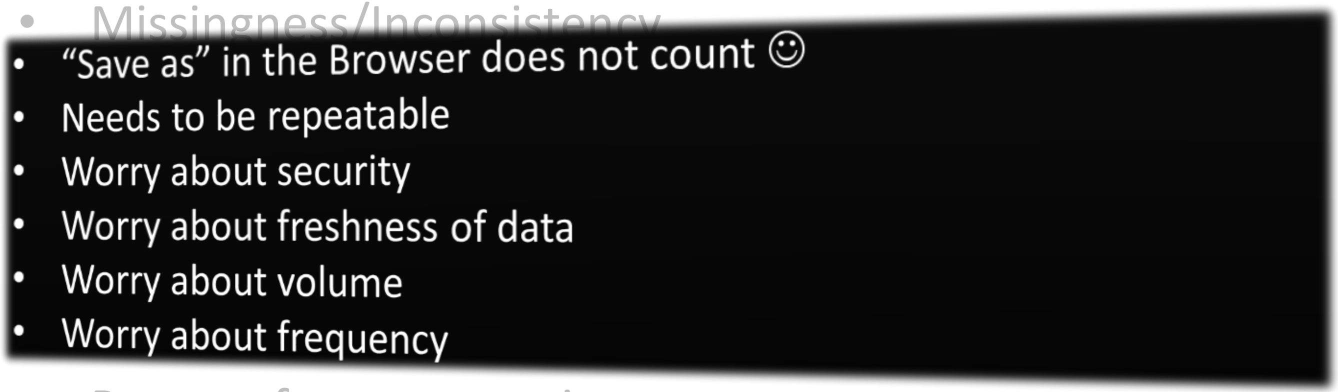
- Acquire
- Rectangularize
- Data Type & Format/Range verification and assertion
- Explore & Understand
- Missingness/Inconsistency
- Outliers
- Derived Columns
- Augmentation & Aggregation
- ML Specific Feature engineering
- Prepare for consumption

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*

An EDA/Data Prep Checklist

- Acquire
- Rectangularize
- Data Type & Format/Range verification and assertion
- Explore & Understand
- Missingness/Inconsistency

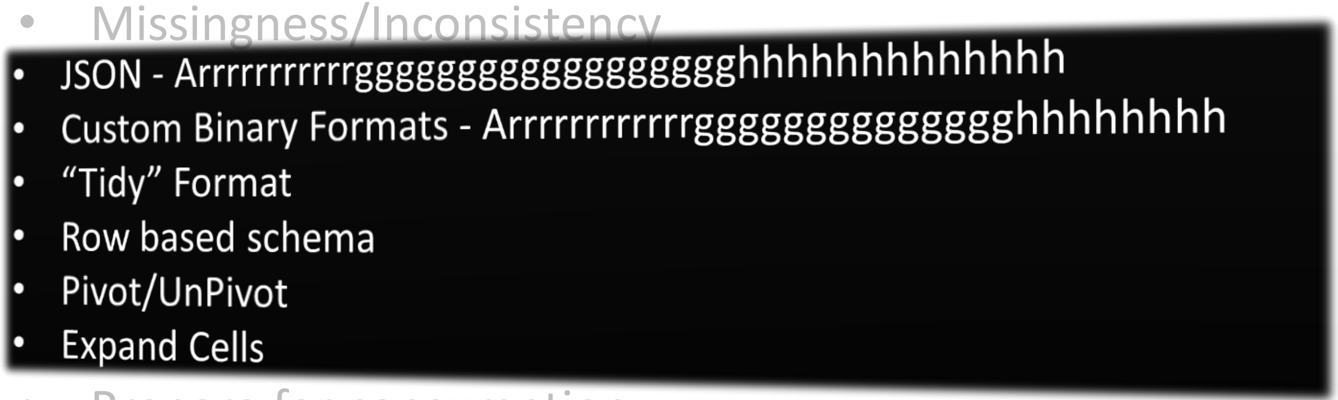


- “Save as” in the Browser does not count ☺
- Needs to be repeatable
- Worry about security
- Worry about freshness of data
- Worry about volume
- Worry about frequency

- Prepare for consumption

An EDA/Data Prep Checklist

- Acquire
- **Rectangularize**
- Data Type & Format/Range verification and assertion
- Explore & Understand
- Missingness/Inconsistency



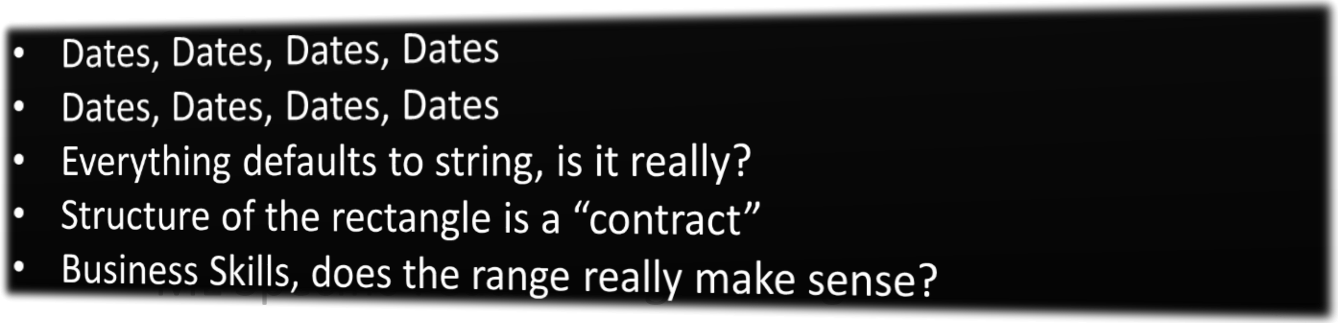
- JSON - Arrrrrrrrrrgggggggggggggggghhhhhhhhhhhh
- Custom Binary Formats - Arrrrrrrrrrrgggggggggggggghhhhhhhh
- “Tidy” Format
- Row based schema
- Pivot/UnPivot
- Expand Cells

- Prepare for consumption

Demo Time

An EDA/Data Prep Checklist

- Acquire
- Rectangularize
- **Data Type & Format/Range verification and assertion**
- Explore & Understand
- Missingness/Inconsistency

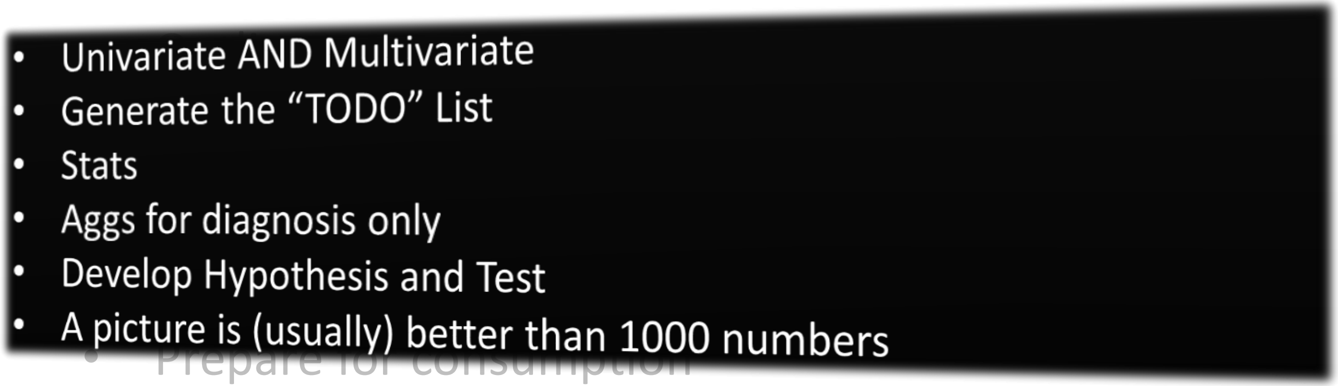


- Dates, Dates, Dates, Dates
- Dates, Dates, Dates, Dates
- Everything defaults to string, is it really?
- Structure of the rectangle is a “contract”
- Business Skills, does the range really make sense?

- Prepare for consumption

An EDA/Data Prep Checklist

- Acquire
- Rectangularize
- Data Type & Format/Range verification and assertion
- **Explore & Understand**
- Missingness/Inconsistency

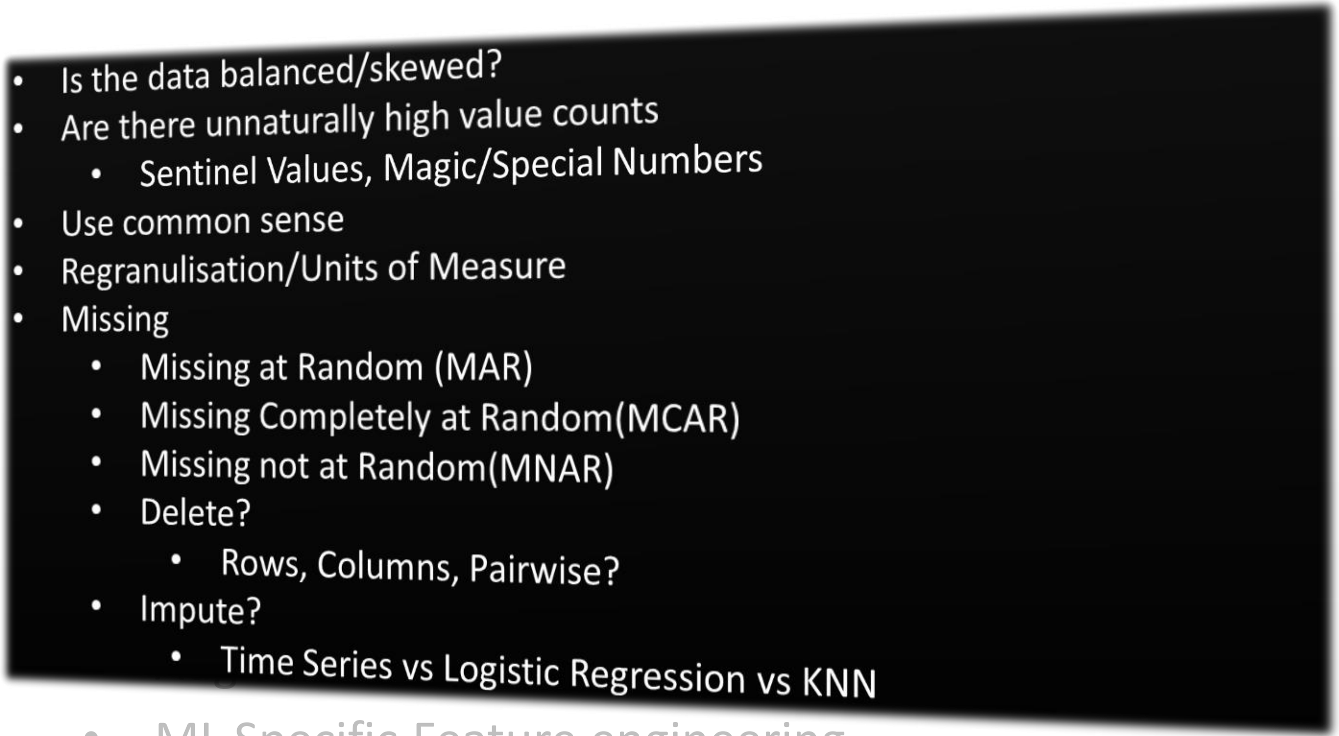


- Univariate AND Multivariate
- Generate the “TODO” List
- Stats
- Aggs for diagnosis only
- Develop Hypothesis and Test
- A picture is (usually) better than 1000 numbers

- Prepare for consumption

Demo Time

An EDA/Data Prep Checklist



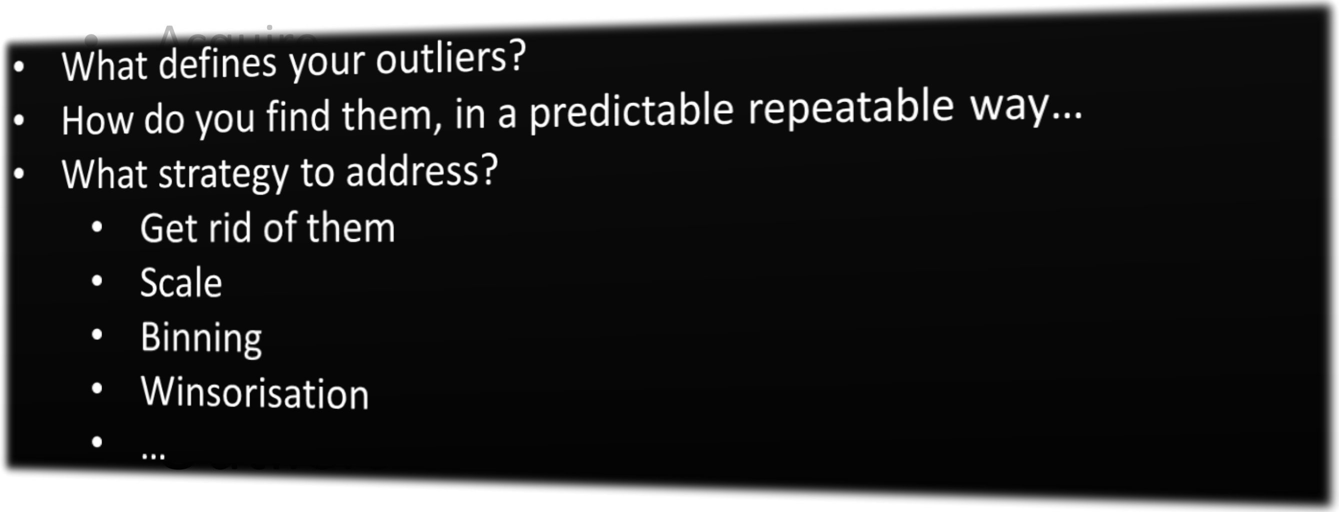
- Is the data balanced/skewed?
- Are there unnaturally high value counts
  - Sentinel Values, Magic/Special Numbers
- Use common sense
- Regranulisation/Units of Measure
- Missing
  - Missing at Random (MAR)
  - Missing Completely at Random(MCAR)
  - Missing not at Random(MNAR)
  - Delete?
    - Rows, Columns, Pairwise?
  - Impute?
    - Time Series vs Logistic Regression vs KNN

- ML Specific Feature engineering
- Prepare for consumption

Demo Time

An EDA/Data Prep Checklist

- Acquire



- What defines your outliers?
- How do you find them, in a predictable repeatable way…
- What strategy to address?
  - Get rid of them
  - Scale
  - Binning
  - Winsorisation
  - …

- Derived Columns
- Augmentation & Aggregation
- ML Specific Feature engineering
- Prepare for consumption

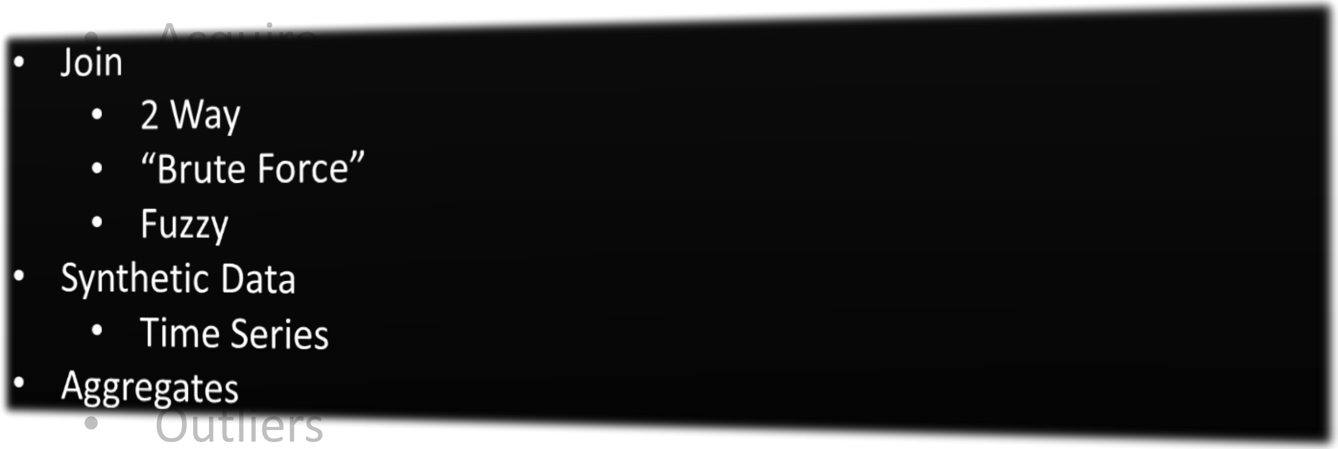
Demo Time

An EDA/Data Prep Checklist

- Acquire
- Rectangularize
- Data Type & Format/Range verification and assertion
- Explore & Understand
- Missingness/Inconsistency
- Outliers
- **Derived Columns**
- Augmentation & Aggregation
- ML Specific Feature engineering
- Prepare for consumption

An EDA/Data Prep Checklist

- Acquire

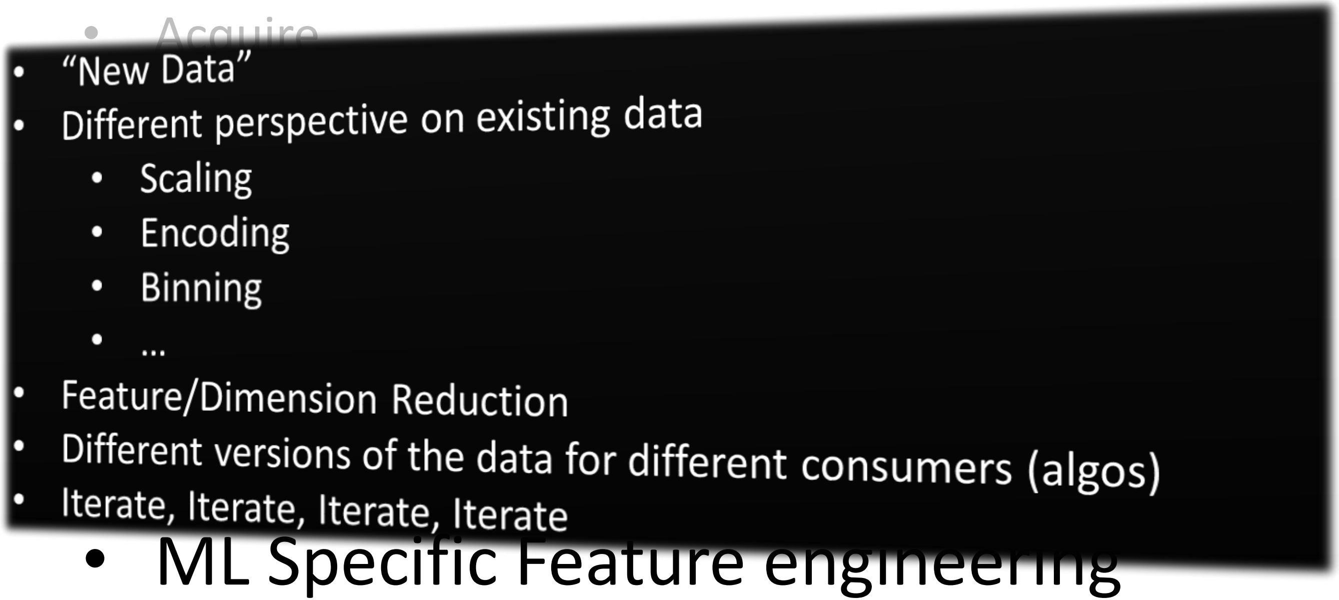


- Join
  - 2 Way
  - “Brute Force”
  - Fuzzy
- Synthetic Data
  - Time Series
- Aggregates

- Outliers
- Derived Columns
- **Augmentation & Aggregation**
- ML Specific Feature engineering
- Prepare for consumption

An EDA/Data Prep Checklist

- Acquire

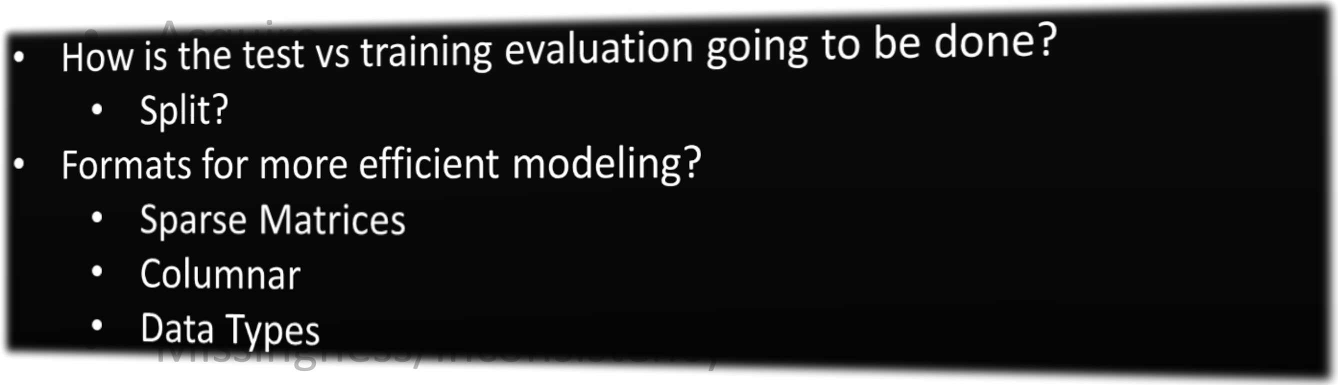


- “New Data”
- Different perspective on existing data
  - Scaling
  - Encoding
  - Binning
  - …
- Feature/Dimension Reduction
- Different versions of the data for different consumers (algos)
- Iterate, Iterate, Iterate, Iterate

- ML Specific Feature engineering
- Prepare for consumption

An EDA/Data Prep Checklist

- Acquire



- How is the test vs training evaluation going to be done?
  - Split?
- Formats for more efficient modeling?
  - Sparse Matrices
  - Columnar
  - Data Types

- Missingness/Inconsistency
- Outliers
- Derived Columns
- Augmentation & Aggregation
- ML Specific Feature engineering
- **Prepare for consumption**

Demo Time

The Winner is…

	Raw Data	Prepped Data
Logistic Regression	0.65	0.78
Random Forest	0.63	0.78
Decision Tree	0.56	0.79
Support Vector Machine	0.66	0.64

But what about…

- Scale
  - Sampling
  - Stats vs Actual Data
  - Visualization
  - Parallelism
- Operationalisation
  - Be defensive
    - Package versioning, especially Python
  - Orchestration/Pipelining
  - Training Data Prep <> Inferencing/Scoring Data Prep
  - Monitor for Drift/Divergence
    - “But it has worked just fine for the last few months…”
  - Dev Ops

@euanga

https://github.com/euanga/VS_Live_0319