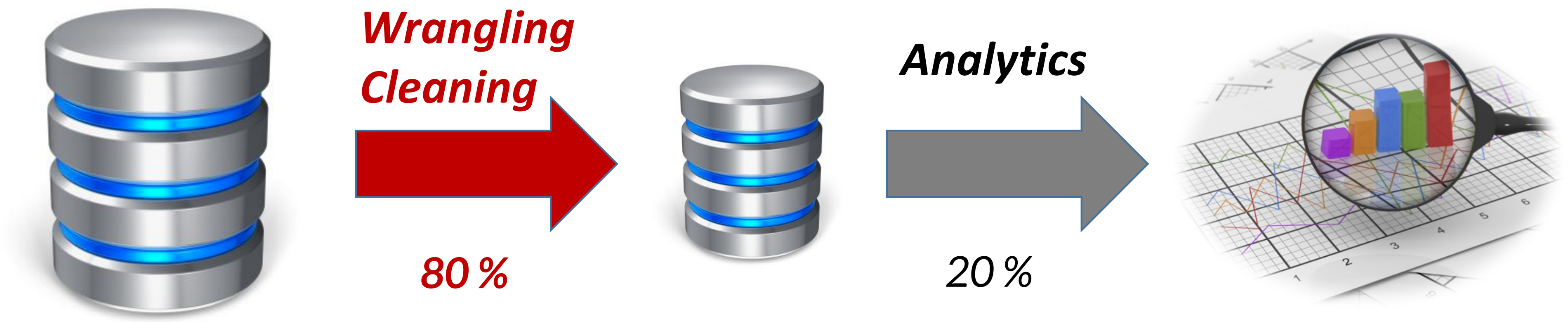
Theory and Practice of Data Cleaning

Overview & Introduction

Data Wrangling vs Analytics



Wrangling
Cleaning

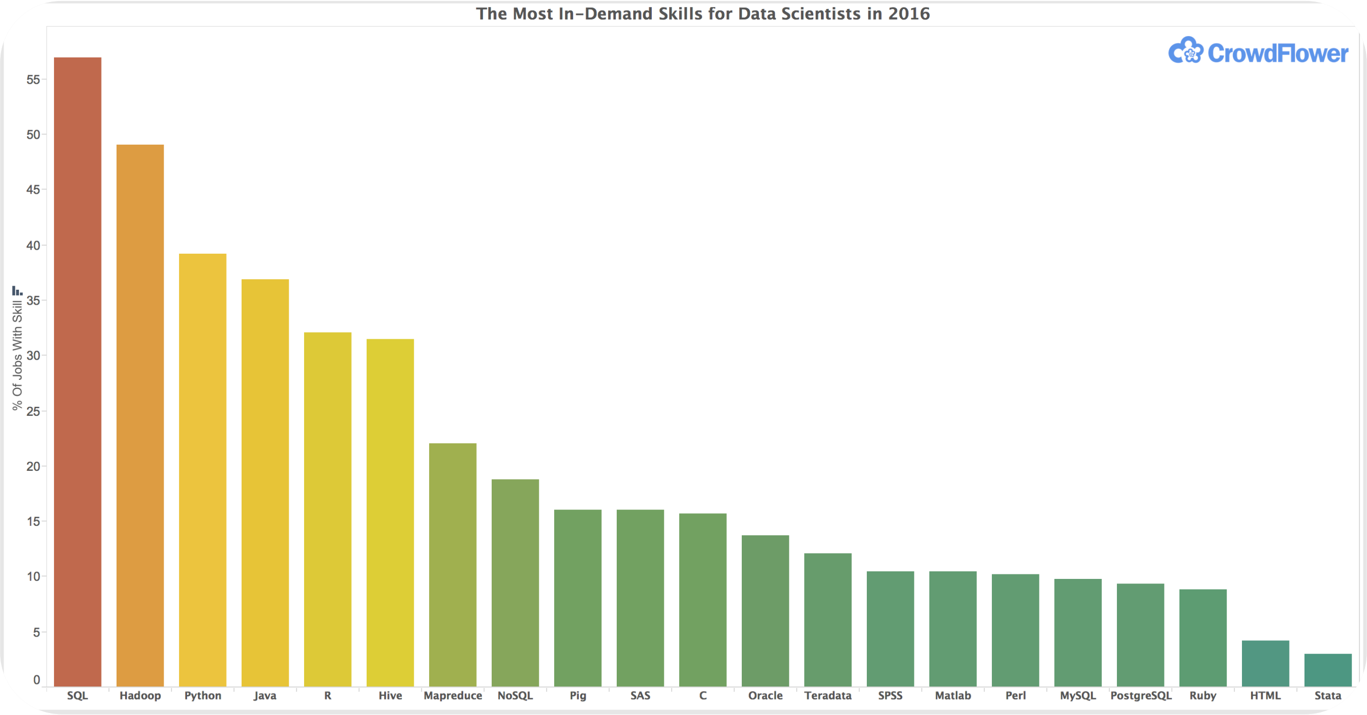
80 %

Analytics

20 %

- **Data wrangling**
  - data processing that allows meaningful analysis to begin
  - extract, transform, load (ETL), integrate, clean, query, repair, …
- **Database** people not always good at public relations
  - … do most of the work
  - ... but most of the attention goes to analytics

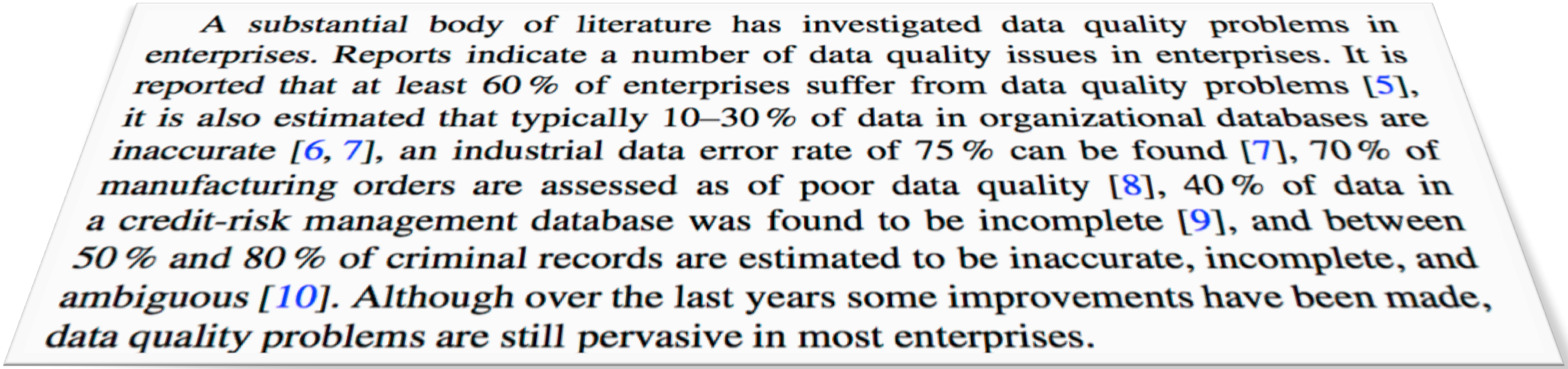
What skills should data scientists have?



Source: Crowdflower.com

- Extracted from 3500 relevant job openings on LinkedIn
- Note: **SQL** (directly and indirectly) is in very high demand!
- Our Focus:
  - Think like a DB person!

Costs resulting from low-quality data

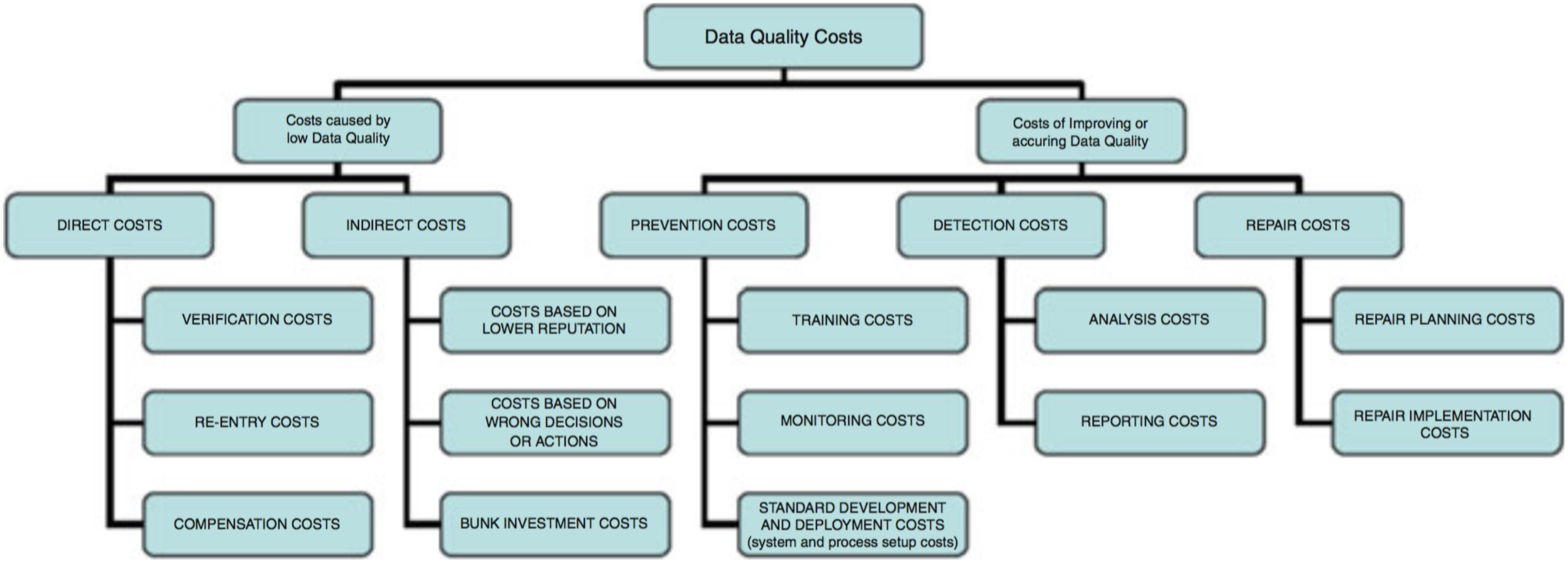


A substantial body of literature has investigated data quality problems in enterprises. Reports indicate a number of data quality issues in enterprises. It is reported that at least 60 % of enterprises suffer from data quality problems [5], it is also estimated that typically 10–30 % of data in organizational databases are inaccurate [6, 7], an industrial data error rate of 75 % can be found [7], 70 % of manufacturing orders are assessed as of poor data quality [8], 40 % of data in a credit-risk management database was found to be incomplete [9], and between 50 % and 80 % of criminal records are estimated to be inaccurate, incomplete, and ambiguous [10]. Although over the last years some improvements have been made, data quality problems are still pervasive in most enterprises.

- Reports indicate cost due to low-quality data to be in the billions of $$$
- ... but also in long-tail of (data) science, data journalism, ..., health, life.

Readings: [GH13] Ge & Helfert. ***Cost and Value Management for Data Quality.***

A Taxonomy of Cost of Types



Data Quality Costs
Costs caused by low Data Quality
Costs of Improving or accuring Data Quality
DIRECT COSTS
INDIRECT COSTS
PREVENTION COSTS
DETECTION COSTS
REPAIR COSTS
VERIFICATION COSTS
RE-ENTRY COSTS
COMPENSATION COSTS
COSTS BASED ON LOWER REPUTATION
COSTS BASED ON WRONG DECISIONS OR ACTIONS
BUNK INVESTMENT COSTS
TRAINING COSTS
MONITORING COSTS
STANDARD DEVELOPMENT AND DEPLOYMENT COSTS (system and process setup costs)
ANALYSIS COSTS
REPORTING COSTS
REPAIR PLANNING COSTS
REPAIR IMPLEMENTATION COSTS

Readings: [GH13] Ge & Helfert. ***Cost and Value Management for Data Quality*.**

Costs resulting from low-quality data

1. Data maintenance costs
2. Personnel costs
3. Data search costs
4. Data quality assessment costs
5. Semantic confusion costs
6. Data re-input costs
7. Wrong data interpretation costs
8. Time costs of viewing irrelevant information
9. Loss of revenue
10. Cost of losing current customer
11. Cost of losing potential new customer
12. Cost of realigning business rules
13. Cost of complicated data integrity
14. “Loss of orders” cost
15. Higher retrieval costs
16. Higher data administration costs
17. General waste of money
18. Cost of system migration and reengineering
19. Costs in terms of lost opportunity
20. Costs due to tarnished image
21. Costs related to invasion of privacy and civil liberties
22. Costs in terms of personal injury and death of people
23. Costs because of lawsuits
24. Process failure costs
25. Information scrap and rework costs
26. Lost and missed opportunity costs
27. Costs due to increased time of delivery
28. Costs of acceptance testing

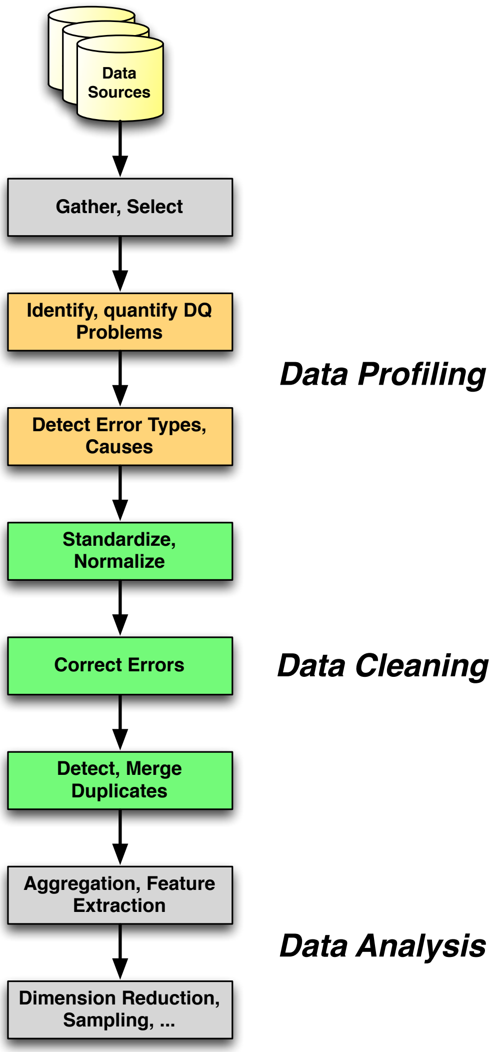
Readings: [GH13] Ge & Helfert. ***Cost and Value Management for Data Quality.***

Data Cleaning – the Big Idea

- **Understanding, assessing, and improving data quality**

- **Quality dimensions** – data should be ...
  - … *accurate, timely, relevant, complete, understood, trusted*, …
- **Questions** we want to ask of data:
  - *Fitness for Use*: Is the quality sufficient to answer my questions?
- **Queries** we want to execute
  - *Data profiling*
  - *Checking Integrity Constraints (ICs)*
  - Answering those questions …
  - ... using Datalog and SQL *queries*!

Data Cleaning in Context



Data Sources

Gather, Select

Identify, quantify DQ Problems

Detect Error Types, Causes

Data Profiling

Standardize, Normalize

Correct Errors

Detect, Merge Duplicates

Data Cleaning

Aggregation, Feature Extraction

Dimension Reduction, Sampling, ...

Data Analysis

- Data from various sources is *gathered*, *selected*

- **Data Profiling**
  - Identify, detect, quantify data quality problems
- **Data Cleaning** (Wrangling)
  - Standardize, normalize data
    - Controlled, reference vocabularies

- **Data Integration & Data Warehousing**
  - Extract, transform, load (ETL tools) → warehouse
  - On demand integration (database mediators)

A Simple Taxonomy of Error Types

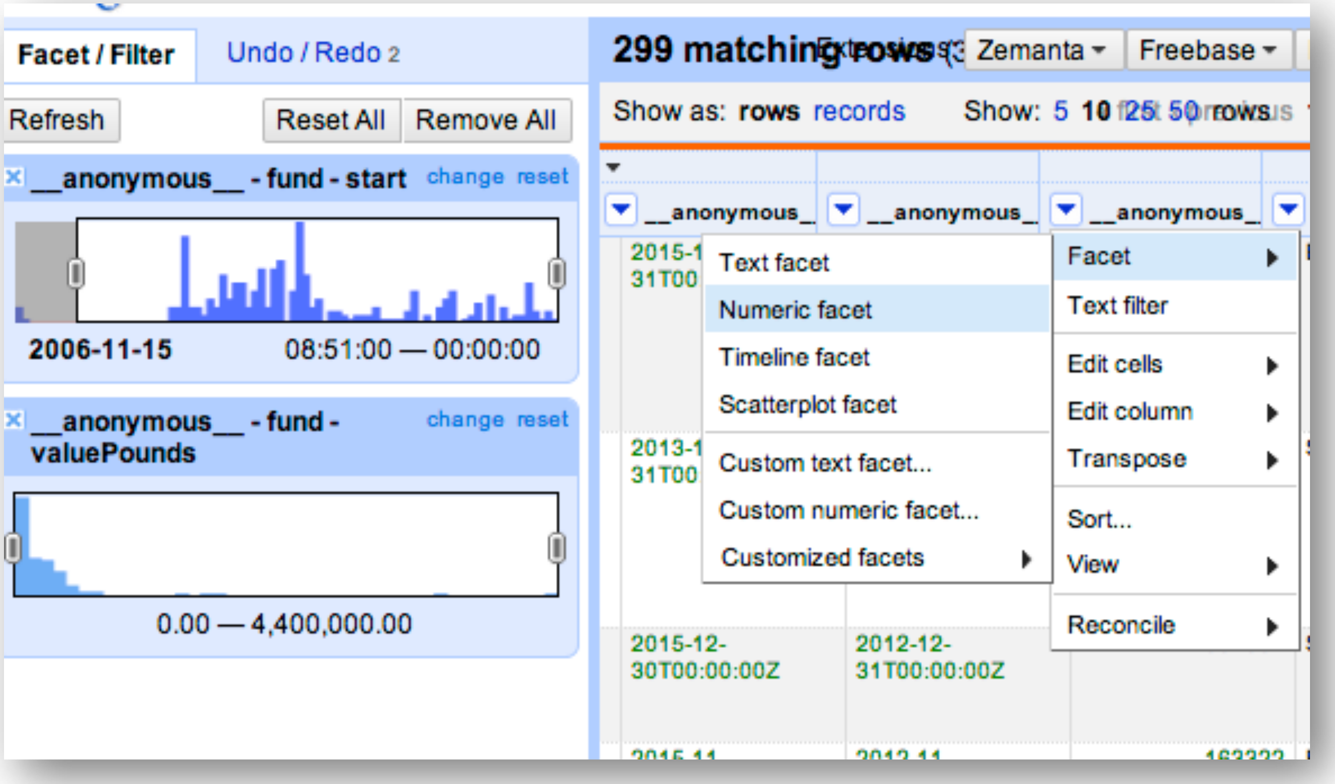
- **Quantitative Errors**
  - Outliers
    - Deviate significantly from the distribution of values
    - Methods from statistics, data mining, machine learning
- **Qualitative Errors**
  - **Syntactic Violations**
    - Pattern violations: variant data formats, spellings, …
  - **Schema / Integrity Constraint (IC) Violations**
    - IC rule violations: Functional or inclusion dependencies
  - Duplicates and other errors
    - Distinct records refer to same real-world entity

Readings: **[ACD+16]** Abedjan et al. ***Detecting Data Errors: Where Are We and What Needs to Be Done?***

Course Themes, Topics, Tools

- Syntax
  - Regular expressions define *patterns* that can be used to *match*, *extract*, and *transform* data, i.e., deal with syntactic variations
  - OpenRefine: open source tool for data wrangling
- Schema & Semantics
  - Using database technologies for **data profiling** (queries); **integrity constraints** (ICs); and repair
  - Datalog and SQL
- Synthesis
  - **Workflow** automation (ETL, scripts)
  - **Provenance** (data lineage and processing history)
  - YesWorkflow: modeling scripts as workflows, provenance

How do you clean data? (*be an OpenRefine hero!*)



… even after OpenRefine, Kurator, Python data cleaning workflows, “dirty data” can make it into our database tables …

PERSON

Id	Name	DOB	Age	Sex	Phone	Zip	Email
43	Doe, Joe	1970-02-27	56	M	(999)-999-999	94102	
43	Jane Dunbar	1.1.1990	26	W	NULL	61820	jdunbar@foobar.com
27	Joe Doe	2/30/70	46	F	+1-530-777-1234	D-6951	joe.doe@gargle.edu

ADDRESS

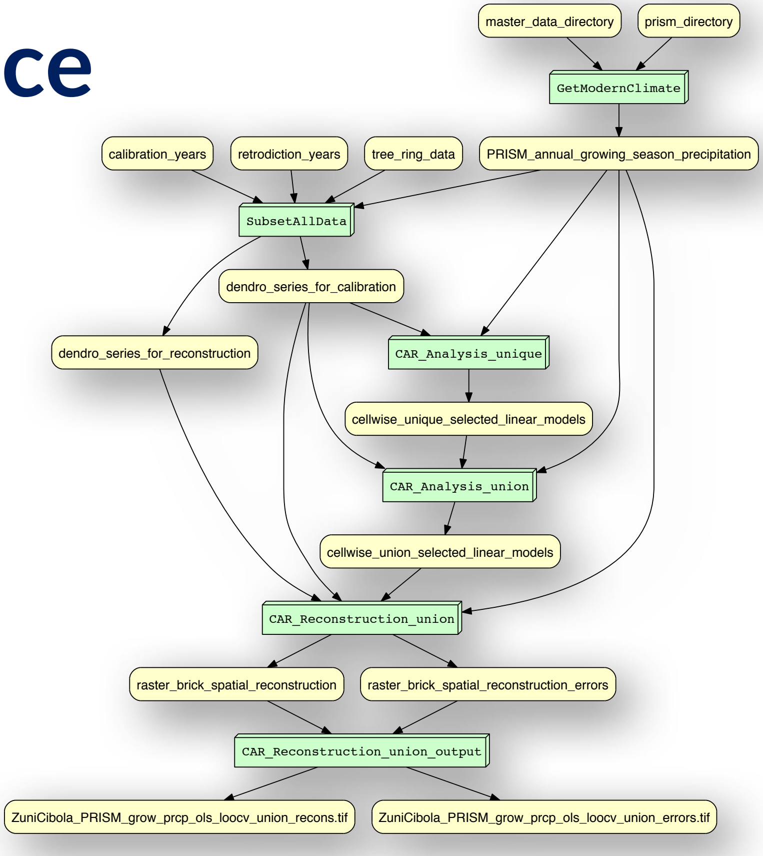
ZIP	City	State
94102	San Franzisco	CA
61821	Champagne	IL
D-6951	Obrigheim	Deutschland

- Errors and IC Violations:
  - Uniqueness (primary key) violation
  - Different representations & formats
  - Contradictions
  - Incorrect values (typos, domain, …)
  - Duplicates
  - Referential Integrity (FK → PK)
    - PERSON.ZIP → ADDRESS.ZIP
  - Incompleteness
  - …

Workflows and Provenance

Kyle B., (computational) archaeologist:

"It took me about 20 minutes to comment. Less than an hour to learn and YW-annotate, all-told."


master_data_directory
prism_directory
GetModernClimate
calibration_years
retrodiction_years
tree_ring_data
PRISM_annual_growing_season_precipitation
SubsetAllData
dendro_series_for_calibration
dendro_series_for_reconstruction
CAR_Analysis_unique
cellwise_unique_selected_linear_models
CAR_Analysis_union
cellwise_union_selected_linear_models
CAR_Reconstruction_union
raster_brick_spatial_reconstruction
raster_brick_spatial_reconstruction_errors
CAR_Reconstruction_union_output
ZuniCibola_PRISM_grow_prcp_ols_loocv_union_recons.tif
ZuniCibola_PRISM_grow_prcp_ols_loocv_union_errors.tif

- Synthesis
  - **Workflow** automation (ETL, scripts)
  - **Provenance** (data lineage and processing history)
  - **YesWorkflow**: modeling scripts as workflows, provenance