Repurposing Administrative Data for Statistical Purposes

Aaron D. Schroeder, Ph.D.
Senior Data Research Scientist
Social & Decision Analytics Lab
Biocomplexity Institute of Virginia Tech

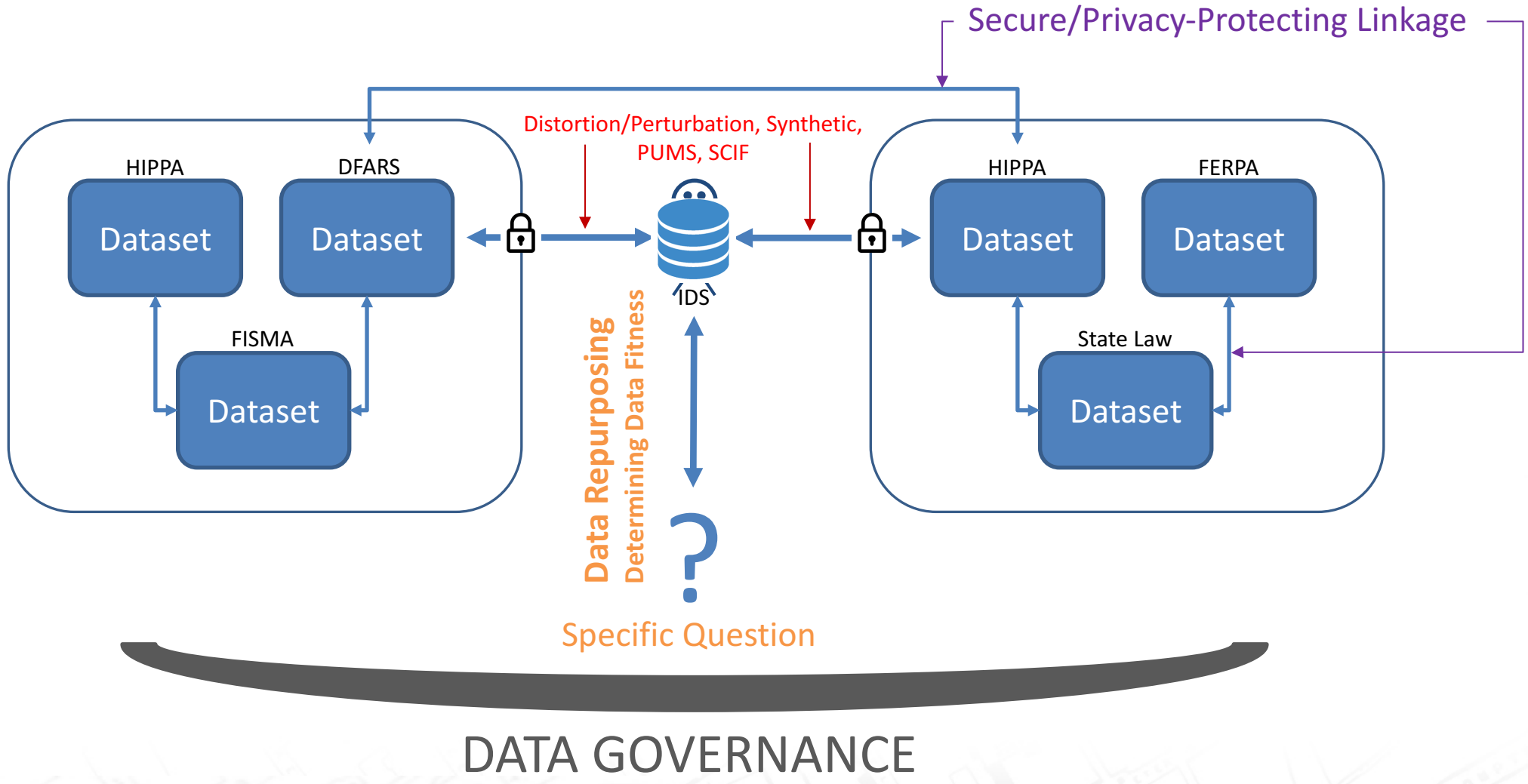
Virginia Tech
Biocomplexity Institute

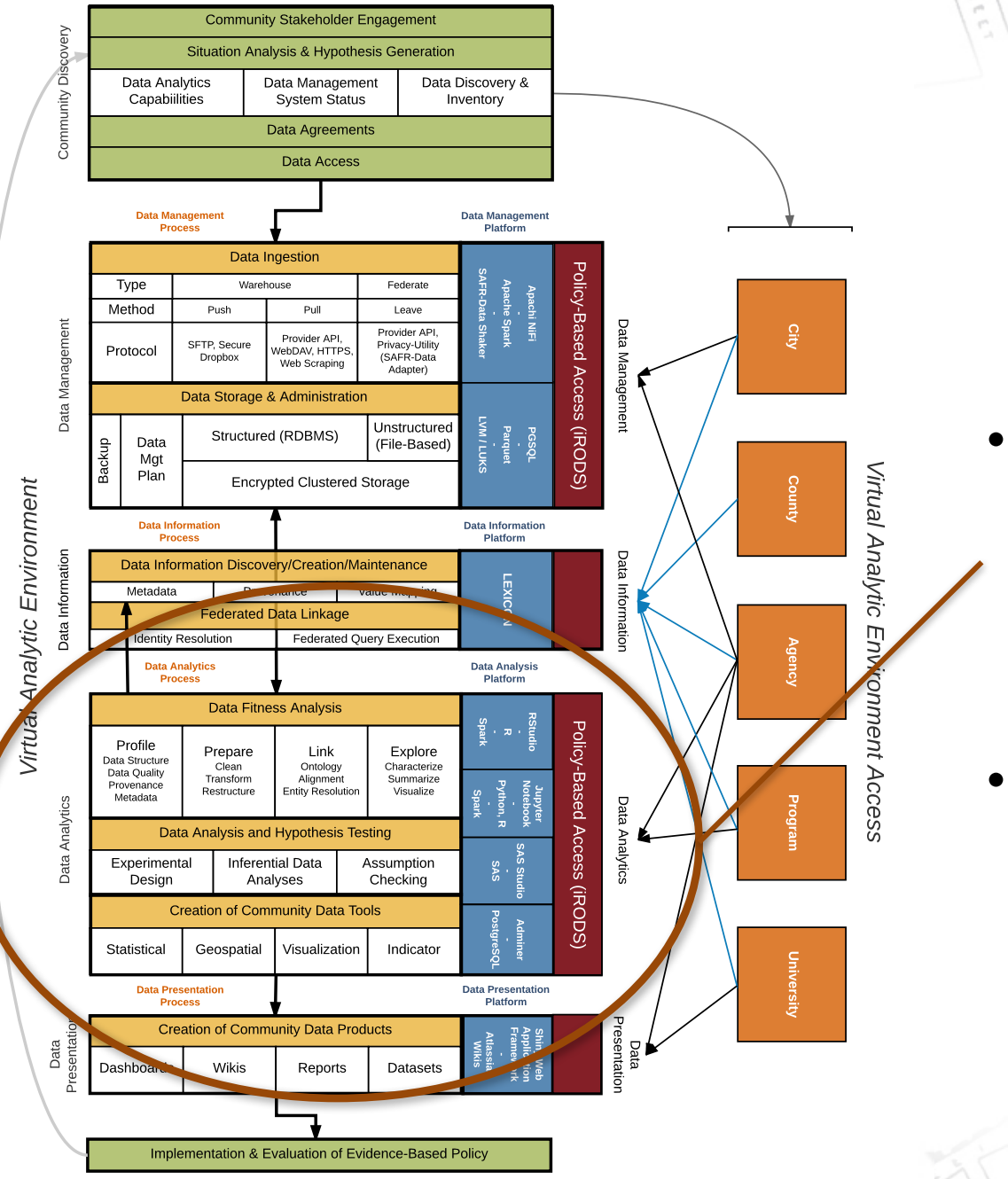
SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Every Repurposing Is a New “Investigation”

- Locating the Data Repurposing Discussion
- Overview of the SDAL “Investigative Process” for Repurposing Data
- Recommendations for Research-Enabling Standards for Integrated Administrative Data Systems to Aid Future Investigations

Data Repurposing
Locating the Discussion





SDAL
Data Science Processes & Platforms for Evidence-Based Policy

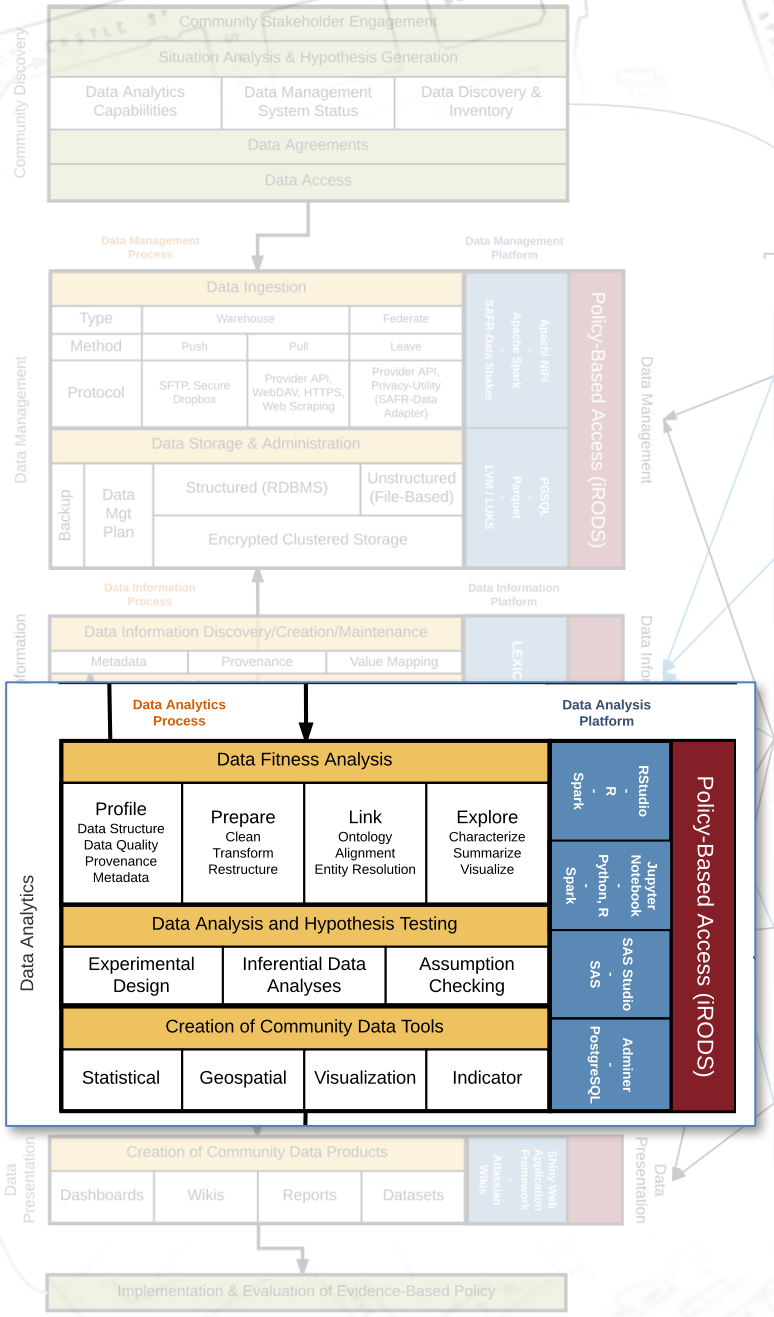
- Data Analytics Process
  - **Data Fitness Analysis**
  - Data Analysis & Hypothesis Testing
  - Creation of Community Data Tools

- Data Fitness Analysis
  - **Profiling**
  - Preparation
  - Linkage
  - Exploration & Assessment

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance



Missing Variables

values in column headers instead of variable names

- e.g. Value-ranges being used as column headers (0-9|10-19|20-29|…)

Combined Variables

more than one variable represented in a attribute (column) value

- e.g. An attribute combining gender and age (m25, f32,...)

Multiple Observation Directions

variables in both columns and rows

- e.g. A dataset with an element(column) for each day of the month (horizontal) and an element(column) for 'month' (vertical)
- note. the messiest and can be dealt with multiple ways according to the needs of the specific analysis

Combined Observation Unit Types

more than one observation unit type per table

- e.g. A table containing both individual demographic data and a periodic measurement like weekly attendance where demographic data and weekly attendance are separate observational units and need to be in separate datasets.

Divided Observation Unit Type

observation unit type is split among multiple tables

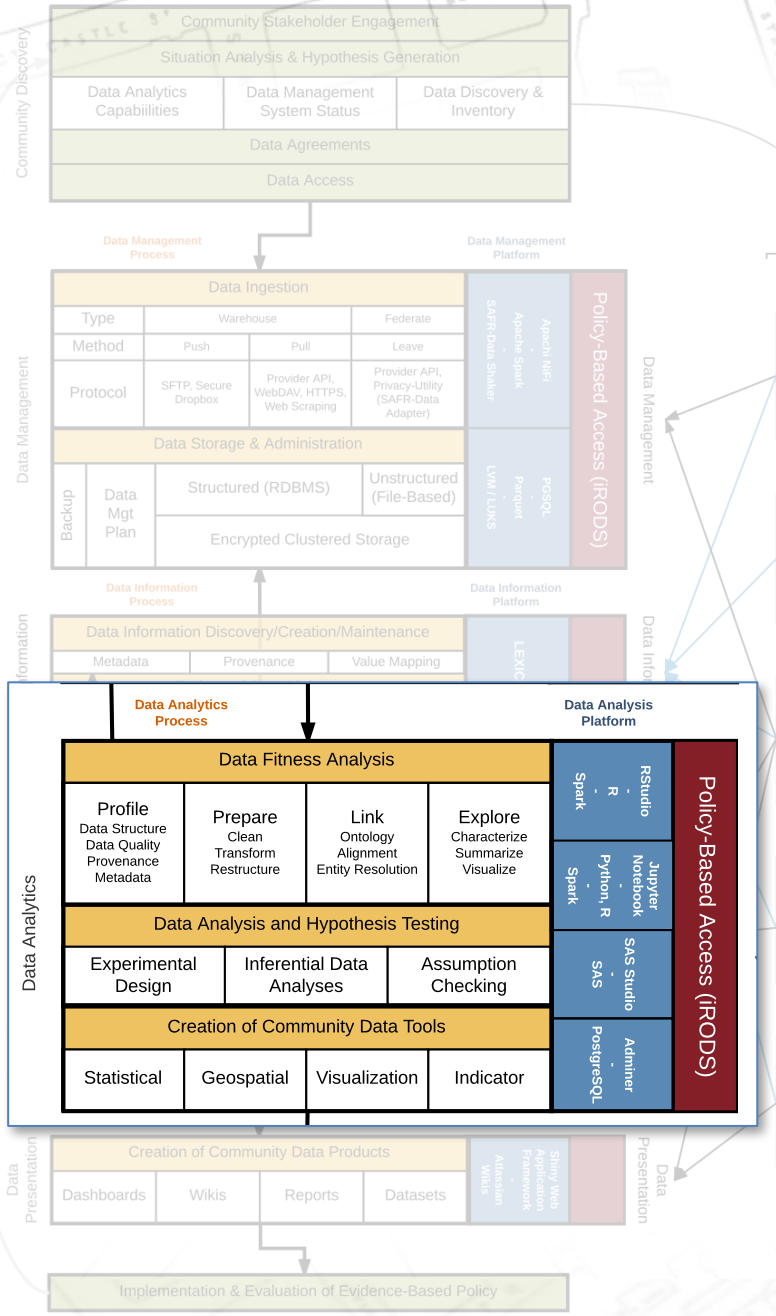
- e.g. Individual demographic information split among several datasets; for example, separate tables for gender, ethnicity, and surname.

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Combined Observation Unit Types



Current Structure of Williamsburg MLS Data

List Number	Agency Name	Agency Phone	Agency Email	Listing Agent	Listing Agent Phone	Listing Agent Email	Co-Listing Agent	Property Type	Card Format
Book Section	Selling Agency	Selling Agency Phone	Selling Agency Email	Selling Agent	Selling Agent Phone	Selling Agent Email	Co-Selling Agent	End Date	book_sec
Listing Date	Sold Date	Under Cont. Date	Fall-thru Date	Status	Status Change	Withdraw Date	Cancel Date	Contingent	Cont. Remarks
Orig. List Price	Price	Sold Price	high_price	Low Price	assessed_val	Partial Tax Assmnt	financing	Area	Relocation
St. #	box_nbr	St. Dir.	Street Name	Address 2	streetdirsuffix	Street Suffix	carrier_route	City	State
county	country	Zip Code	geo_county	Taxes	geo_lat	geo_lon	Est. Fin. SqFt	sqft1	sqft2
sqft3	sqft4	Year Built	2+ Bdroms on 1st Flr	Realtor.com Type	lot_size	Total Acres	Condo Level	sell_broker_comm	Variable Commission
stories	Total Rooms	Total Bedrooms	total_bath	Baths - Full	Baths - Half	baths_3_4	Garage Type	garage_stall	Water Frontage
Zoning	taxes	Tax Year	Subdivision	Public Remarks	Agent Remarks	**Parcel ID**	Legal Description	Directions	Foreclosure
Owner Phone	Owner Name	Neighborhood	mod_timestamp	Ltd Service Agent	Occupied By	Owner/Agent	Mster Bdrm 1st Floor	SqFt Source	Listing Type
# Stories	# Fireplaces	Golf Frontage	IDX Y/N	Supplement Attached	Seller Concession(s)	Special Assmnts	Type	Rollback Taxes	userdefined16
SellingBroker Incent	Ownership	Describe Concession	How Sold	Selling Broker Comp	userdefined22	Assessed Value	Est.Unfinished Sq Ft	Tax Rate	Garage Bays
userdefined27	userdefined28	userdefined29	userdefined30	Est. Closing Date	userdefined32	userdefined33	Lot Description	Short/CompromiseSale	userdefined36
userdefined37	userdefined38	userdefined39	userdefined40	userdefined41	userdefined42	userdefined43	userdefined44	userdefined45	userdefined46
userdefined47	userdefined48	userdefined49	userdefined50	userdefined51	userdefined52	userdefined53	userdefined54	userdefined55	userdefined56
Photo URL	Days on Market	Rooms	Features						

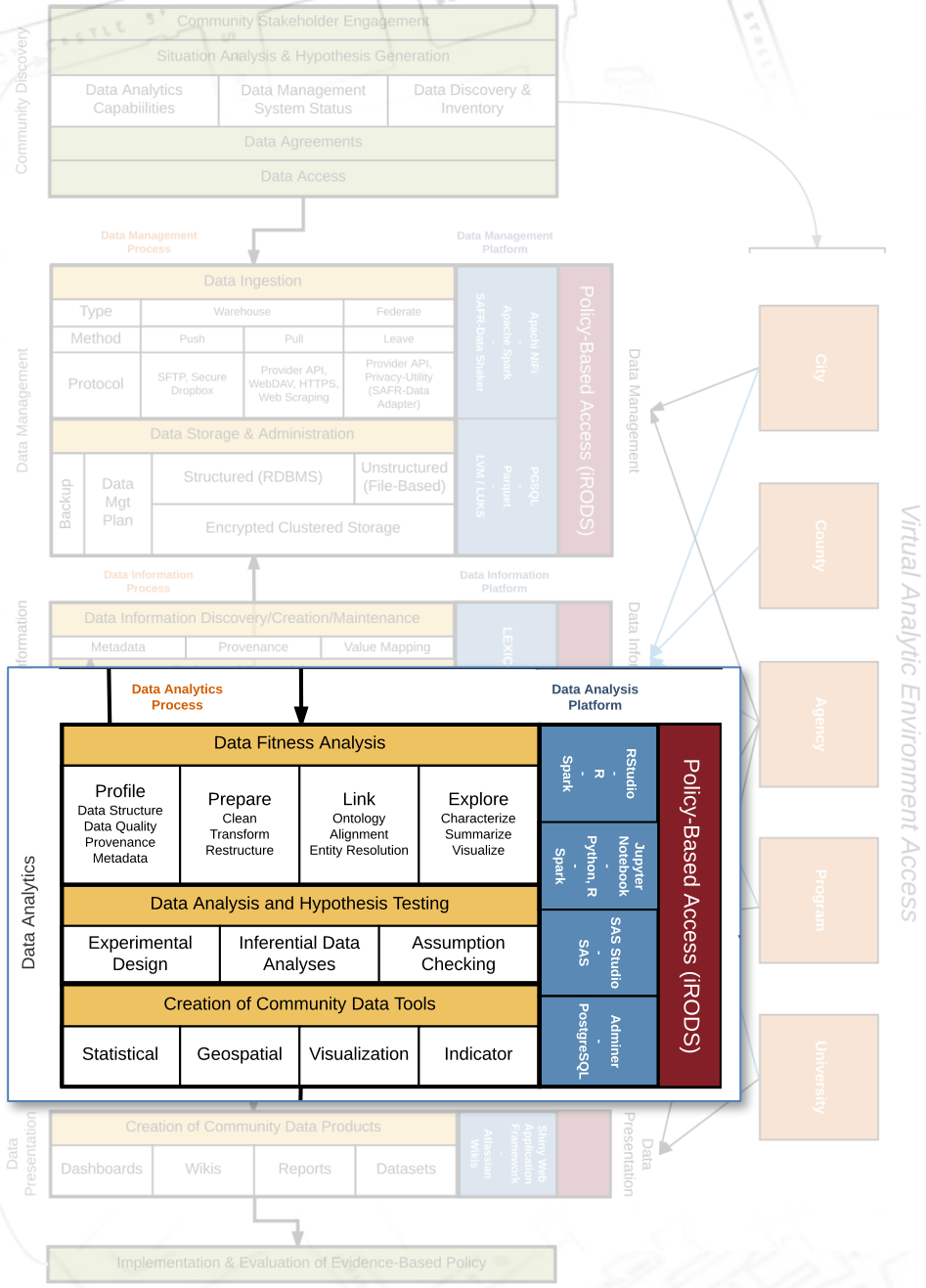
- This is a single record with 128 fields all keyed to the variable “List Number”
- Structured this way, it is not possible to analyze property changes over time
- Pulling out a definitive list of unique properties using “Parcel ID” seems like a possibility
- However, “Parcel ID” is left blank in over 7% of entries – extra work required – perhaps including address, but address is not standardized

Repurposing Data for Statistical Purposes

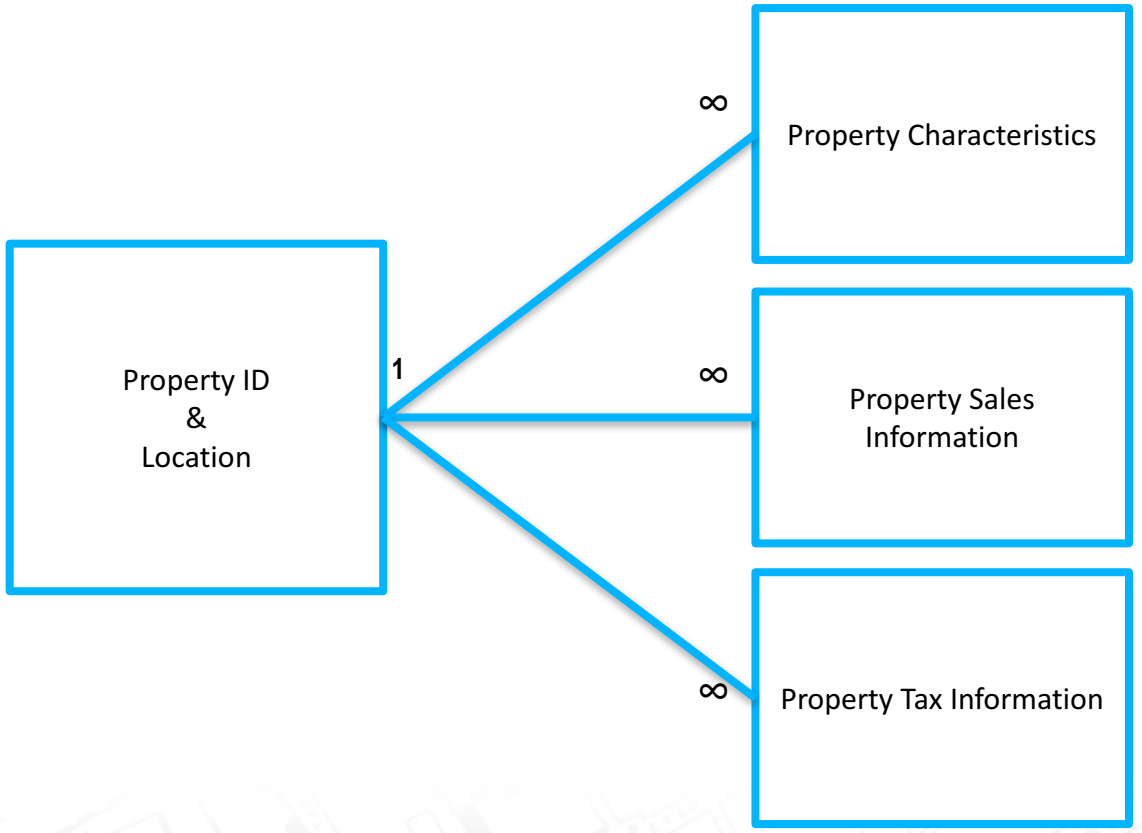
Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Combined Observation Unit Types



Ideal Restructuring of MLS Data

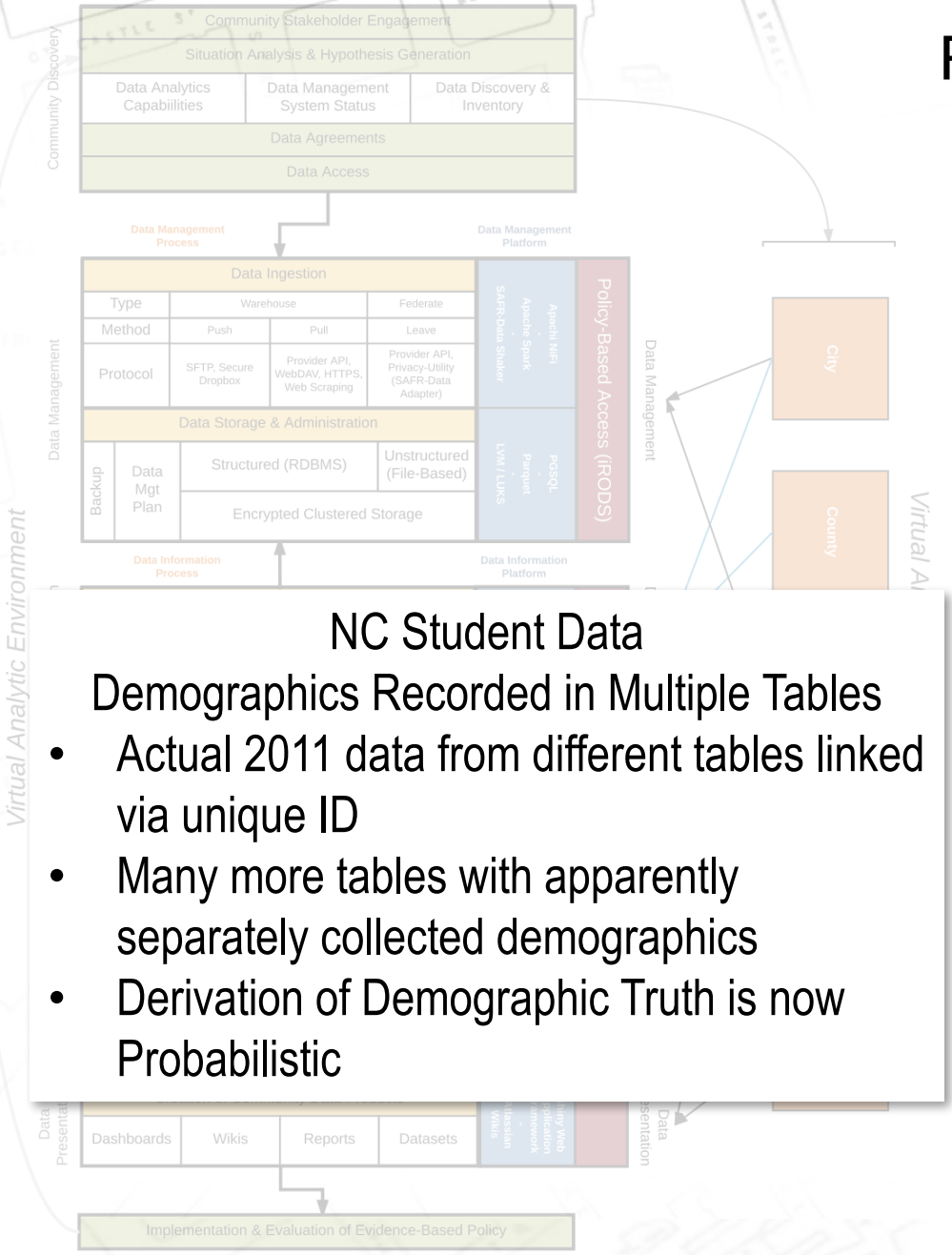


Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**

Structure, Quality, Metadata & Provenance

Divided Observation Unit Types



NC Student Data

Demographics Recorded in Multiple Tables

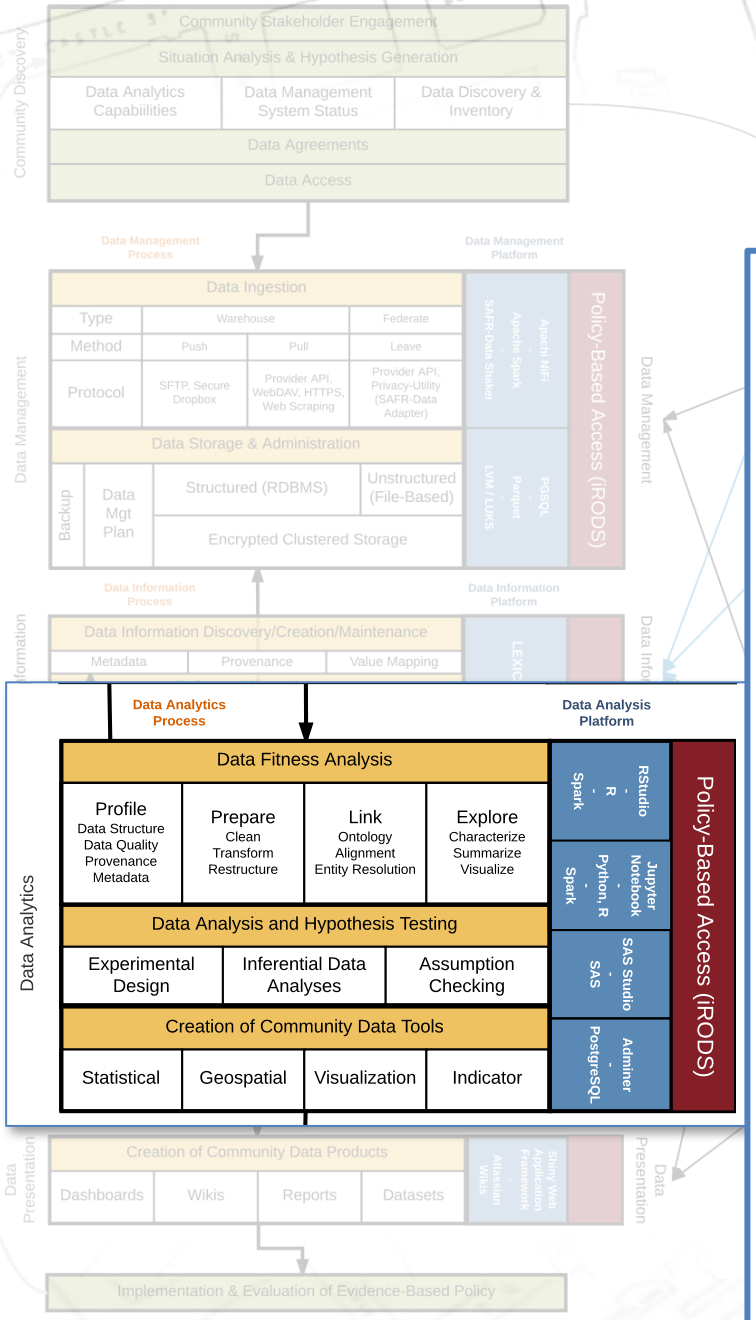
- Actual 2011 data from different tables linked via unique ID
- Many more tables with apparently separately collected demographics
- Derivation of Demographic Truth is now Probabilistic

gender1	id	gender2
F	43XXX13	M
F	43XXX14	M
M	76XXX46	F
F	74XXX98	M
F	76XXX23	M
F	77XXX40	M
M	74XXX98	F
M	78XXX73	F
F	78XXX74	M
M	77XXX84	F
F	79XXX87	M
M	71XXX95	F
M	21XXX96	F
M	71XXX54	F
F	71XXX55	M
F	77XXX86	M
F	80XXX24	M
M	76XXX79	F

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance



Completeness

percentage of elements properly populated

e.g. Testing for NULLs and empty strings where not appropriate

Value Validity

percentage of elements whose attributes possess meaningful values

e.g. A comparison constraint like {male; female} or an interval constraint like age = [0,110]

Consistency

a measure of the degree to which two or more data attributes satisfy a well-defined dependency constraint – relationship validation

e.g. Zip-code – state consistency or gender – pregnancy consistency

Uniqueness

the number of unique values taken by an attribute, or a combination of attributes in a dataset

e.g. Frequency distribution of an element

note. The more homogeneous the data values of an element, the less useful the element is for analysis

Duplication

a measure of the degree of replication of distinct observations per observation unit type

e.g. Greater than 1 registration per student per official reporting period

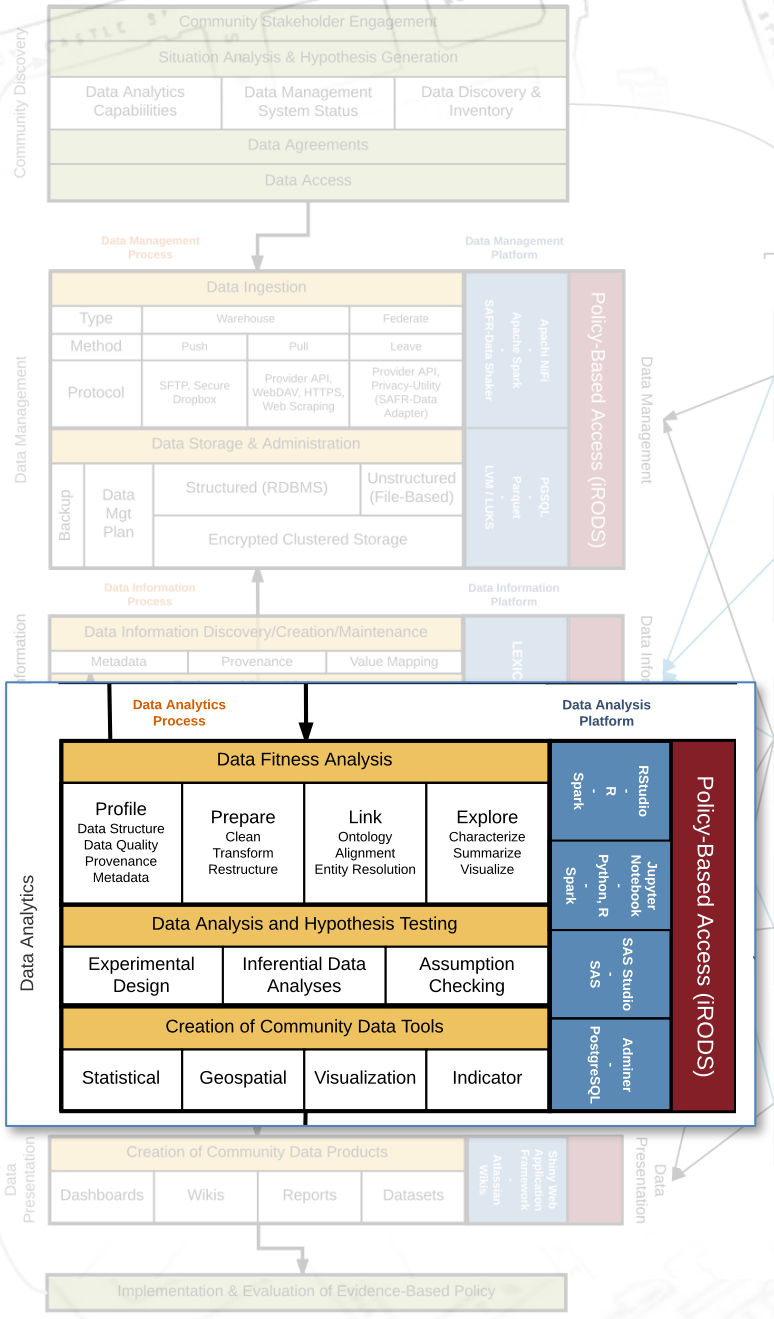
note. Duplication occurs as a result of choice of level of aggregation

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Completeness



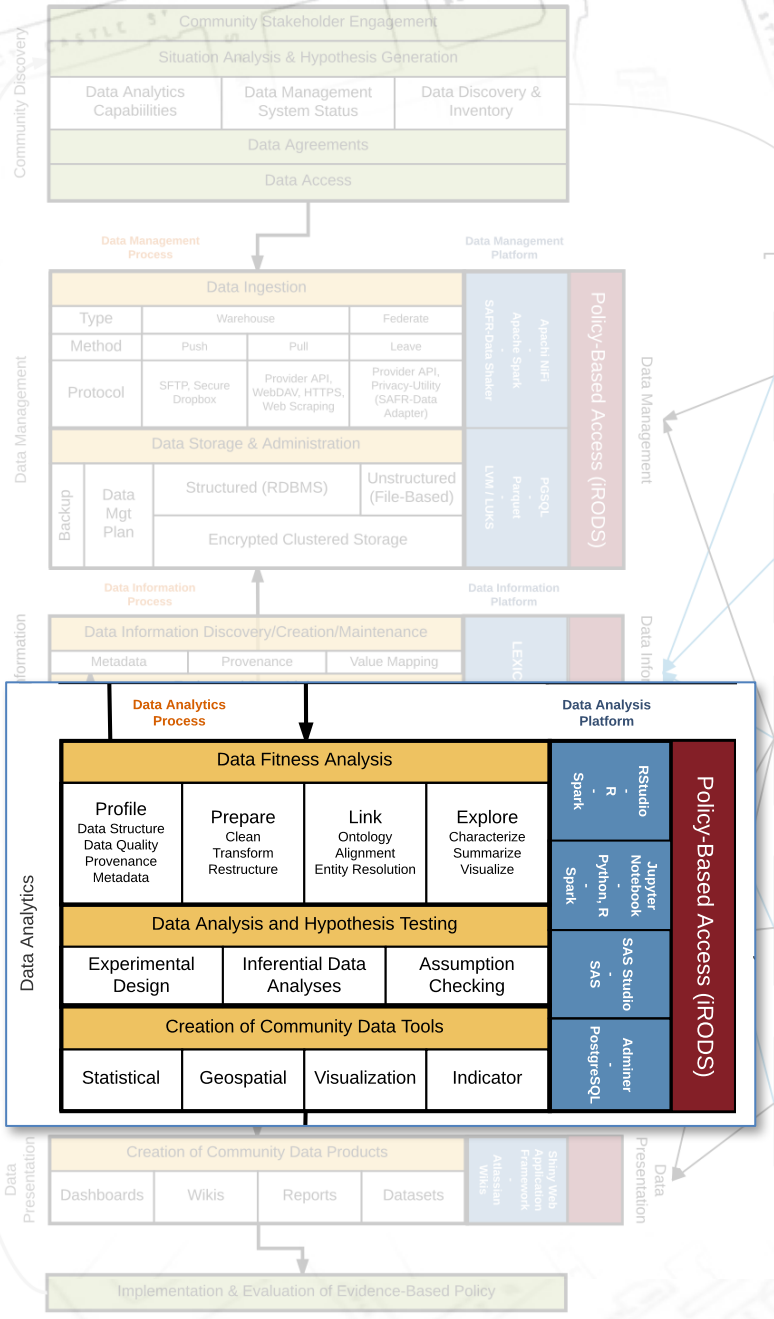
- Seems straight-forward -- Nope
- A set of data is complete with respect to a *given purpose* if the set contains all the relevant data for that purpose
- A common measure is the proportion of data that has values to the proportion that “should” have values.
  - Completeness is *application-specific*
  - Incorrect to simply measure number of missing field values in a record without considering which fields are necessary
    - MLS Data had MANY highly incomplete fields that were not necessary for the study at hand
- Data that are missing can be categorized as:
  - record fields not containing data
  - records not containing necessary fields
  - datasets not containing the requisite records

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Value Validity



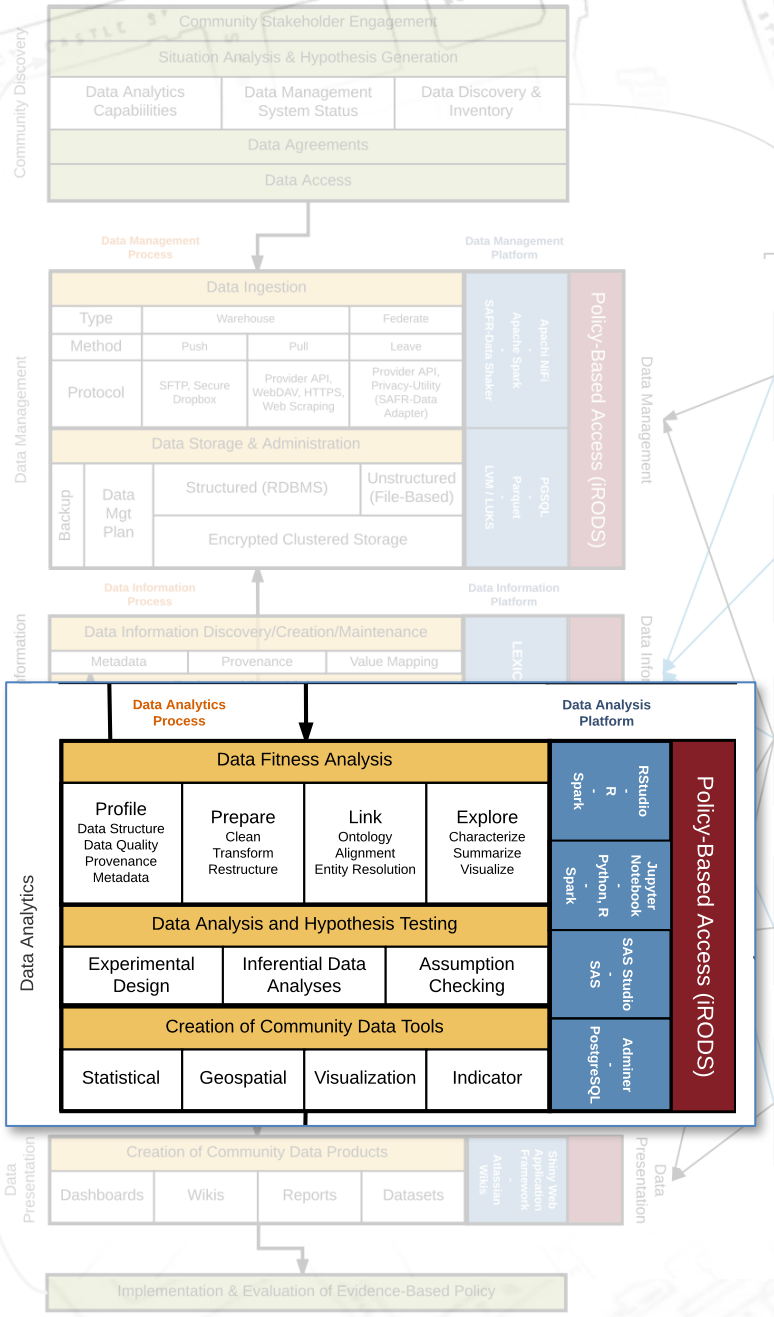
- Data elements with proper values have ***value validity***
- The percentage of data elements whose attributes possess values within the range expected for a legitimate entry is a measure of value validity
- Checking for value validity generally comes in the form of straight-forward domain constraint rules
  - How many entries contain non-valid values for a non-empty text field representing gender?
    - *< count gender where gender is not (male, female) >*
  - How many entries contain non-valid values for a non-empty integer field representing age?
    - *< count age where age is not between* [0, 110] >

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Value Validity



Pulled from current James City County MLS Data

zip_code	area	subdivision	neighborhood	zoning	parcel_id
23185	JCC	Governors Land	River Reach	R-4	4511000022
23188	JCC	Wellington		RESIDENT	1330800178
23188	JCC	Powhatan Secondary		RES	3741600013
23185	JCC	Kingsmill	Padgetts Ordinary	R 4	5041100213
23185	JCC	Pointe @ Jamestown		RES	4640600108
23185	JCC	Paddock Green	Paddock Green	R1	

Comparison constraint: **zoning 2015, James City County** = {A-1, R-1, R-2, R-3, R-4, R-5, R-6, R-7, R-8, LB, B-1, M-1, M-2, RT, PUD, MU, PL, EO}

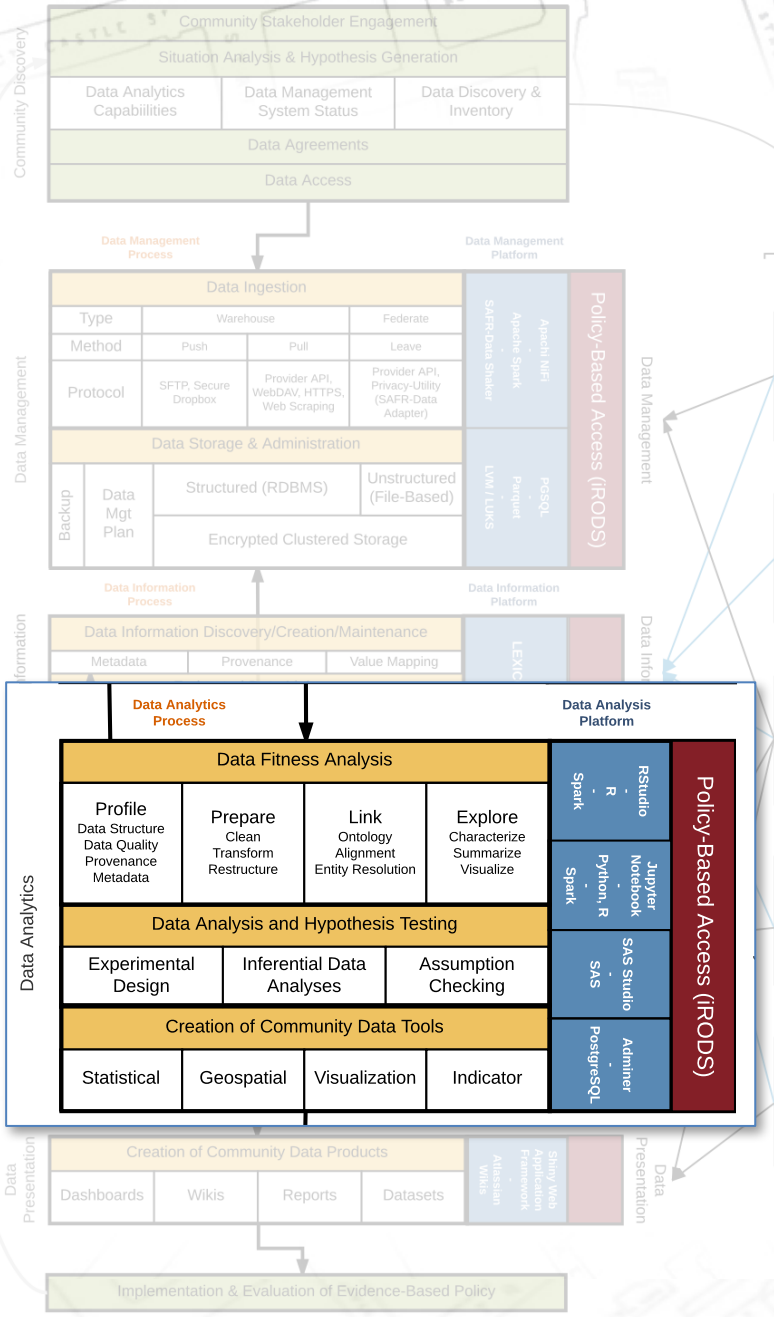
- During Data Profiling issues are described, not “fixed”
- The appropriate fix depends upon the needs of the research
- It may be appropriate to simply normalize all zoning entries to the five major categories of zoning: Residential, Mixed Residential-Commercial, Commercial, Industrial, and Special

Repurposing Data for Statistical Purposes

Data Fitness Analysis: Profiling

Structure, Quality, Metadata & Provenance

Consistency

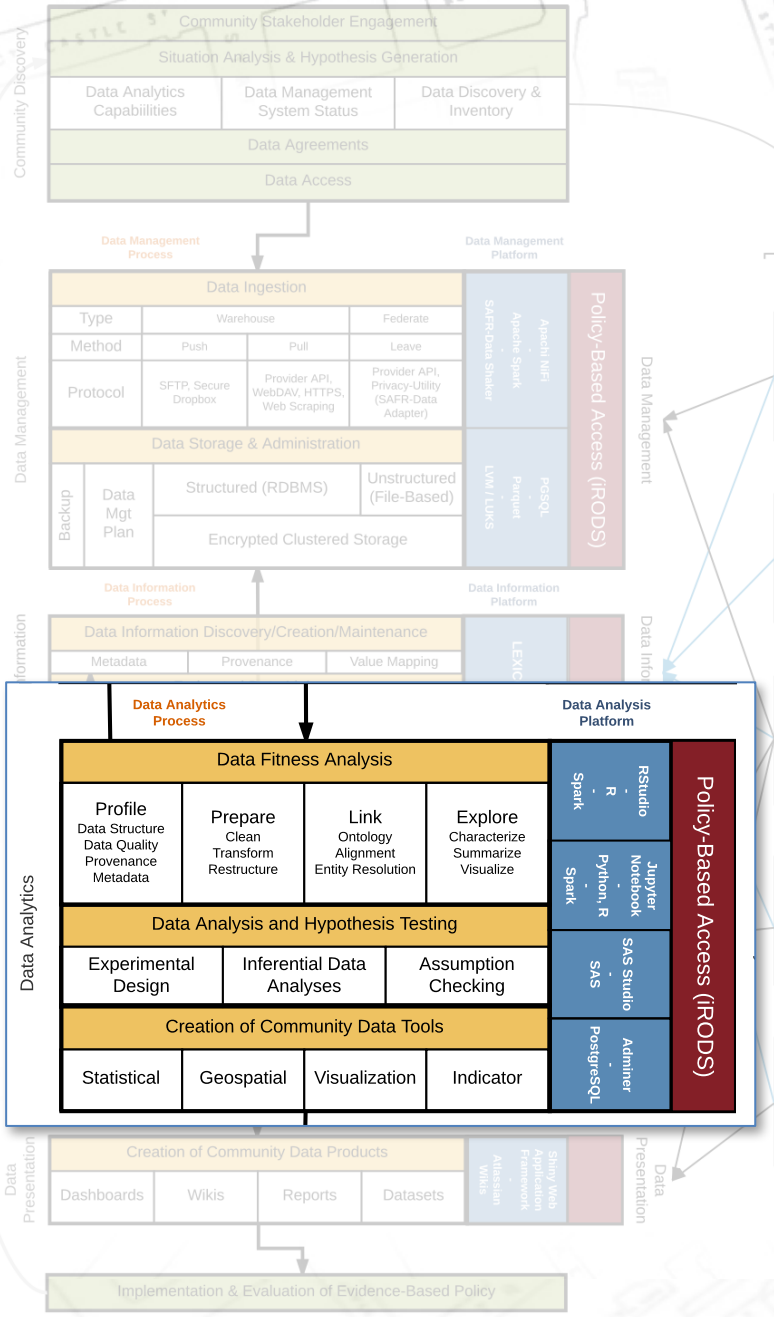


- The Degree to Which Two or More Attributes Satisfy a Dependency Constraint
- Simple example
  - Location disagreements like zip and state (**Record-Level**)
- More complex example (**Longitudinal**)
  - Consistency with locally derived “truth”
  - VDOE Student Record, no definitive list of student demographics
  - Truth must be derived from multiple observations
    - Student Record has multiple observations per school year
    - Query here shows disagreement on gender for some of the observations when Student Record is matched to itself
      - **select count(distinct a.internal_id)**
        from vdoe.student_record a
        join vdoe.student_record b
        on a.internal_id = b.internal_id
        and a.gender <> b.gender
      - **16,310 / 2,346,058 individuals have more than one value for gender**

Repurposing Data for Statistical Purposes

Data Fitness Analysis: **Profiling**

Structure, Quality, **Metadata & Provenance**



Observation Unit Definition
Datasets (tables) without definition and/or non-meaningful/confusing naming

Observation Unit Attributes Definition
Attributes (columns) without definition and/or non-meaningful/confusing naming

Semantic Confusion
Attributes with the same name but different definitions
e.g. An attribute named “Grade” can refer to both a 'score' for a test or the 'level/year’

Multiple Attribute Names
Attributes with different names but the same definition
e.g. Attributes name “Grade” and “Year” both referring to 'level/year' of schooling

Inconsistent Attribute Formats
Attributes of the same type that are formatted differently
e.g. Most commonly an issue when dealing with dates and times

Data Process History
Attributes collected at different locations, with different tools

System of Origin
Where was this data originally collected?

Intermediate Storage Systems
Chain of Custody

Contact Information
Who can I contact with my questions?

Transformation
What happened to the data since collection and why?

Getting this stuff in order is a BIG part of Data Repurposing!

Research-Enabling Standards for Integrated Administrative Data Systems

- Metadata
  - Minimum: Table and Field Definitions, Field Valid Values and Definitions
  - Extra: Valid Value Timing and Relationship Data
- Provenance
  - Minimum: System of Origin/Collection and Contact
  - Extra: Intermediate Storage Systems and Contacts
  - Extra +: Transformations Used and Reasons Why
- Just starting with the Minimums will accomplish quite a bit