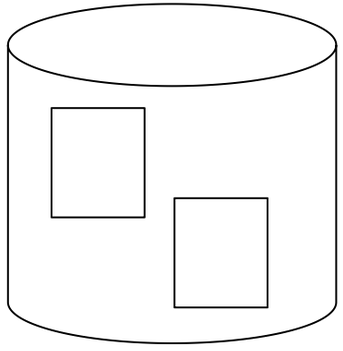
Data Cleaning, Transformation, Enrichment

AnHai Doan

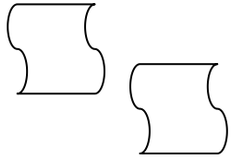
Motivation



X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

news articles



data extraction

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

data cleaning: GE revenue: 351 ⟶ 35.1

schema matching: name = cname
loc = address

schema merging: X(name, loc)
Y(cname, address, rev)
} Z(name, loc, rev)

data matching:

M

xid	yid
x_1	y_2
x_2	y_1

data merging: for name, return the longer string from X.name and Y.cname
for loc, return X.loc

schema mapping: Z = select merge_name(X.name, Y.cname), X.loc, Y.rev
from X, Y, M
where X.id = M.xid and Y.id = M.yid

X

id	name	loc
x_1	Apple	CA
x_2	IBM	NY

Y

id	cname	address	rev
y_1	IBM Corp	CA	25
y_2	Apple Inc	CA	51
y_3	GE	NY	351

Z

name	loc	rev
Apple Inc	CA	51
IBM Corp	NY	25

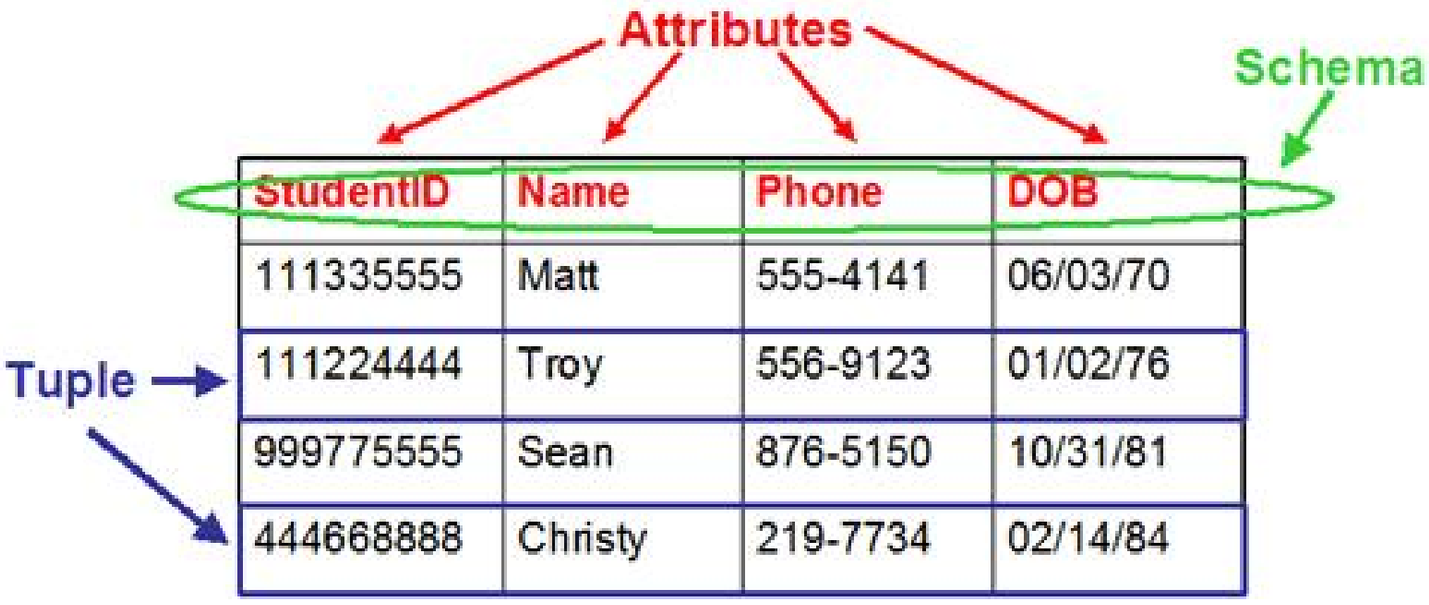
Three Goals for Data Cleaning

- Detect errors (aka data quality problems)
- Decide which errors to fix
- Fix errors

- Often do the above using dictionaries, rules, ML

Data

- Typically taken to mean schema + data instances

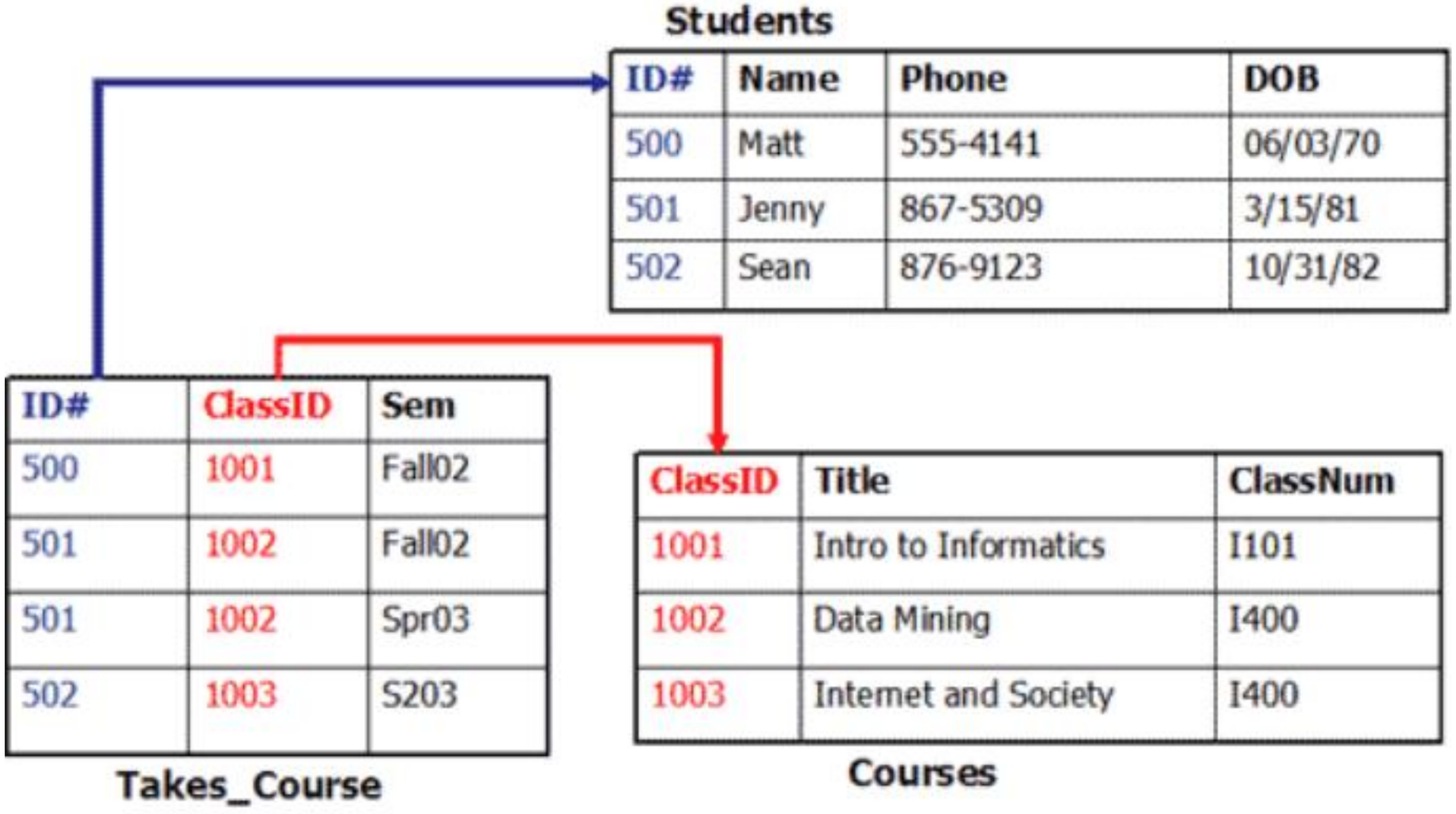


StudentID	Name	Phone	DOB
111335555	Matt	555-4141	06/03/70
111224444	Troy	556-9123	01/02/76
999775555	Sean	876-5150	10/31/81
444668888	Christy	219-7734	02/14/84

- Ideally we should use “schema” and “data instances”
- But often we will say “schema” and “data”

Schema Often Has Many Constraints

- Key, uniqueness, functional dependencies, foreign keys



Students

ID#	Name	Phone	DOB
500	Matt	555-4141	06/03/70
501	Jenny	867-5309	3/15/81
502	Sean	876-9123	10/31/82

Takes_Course

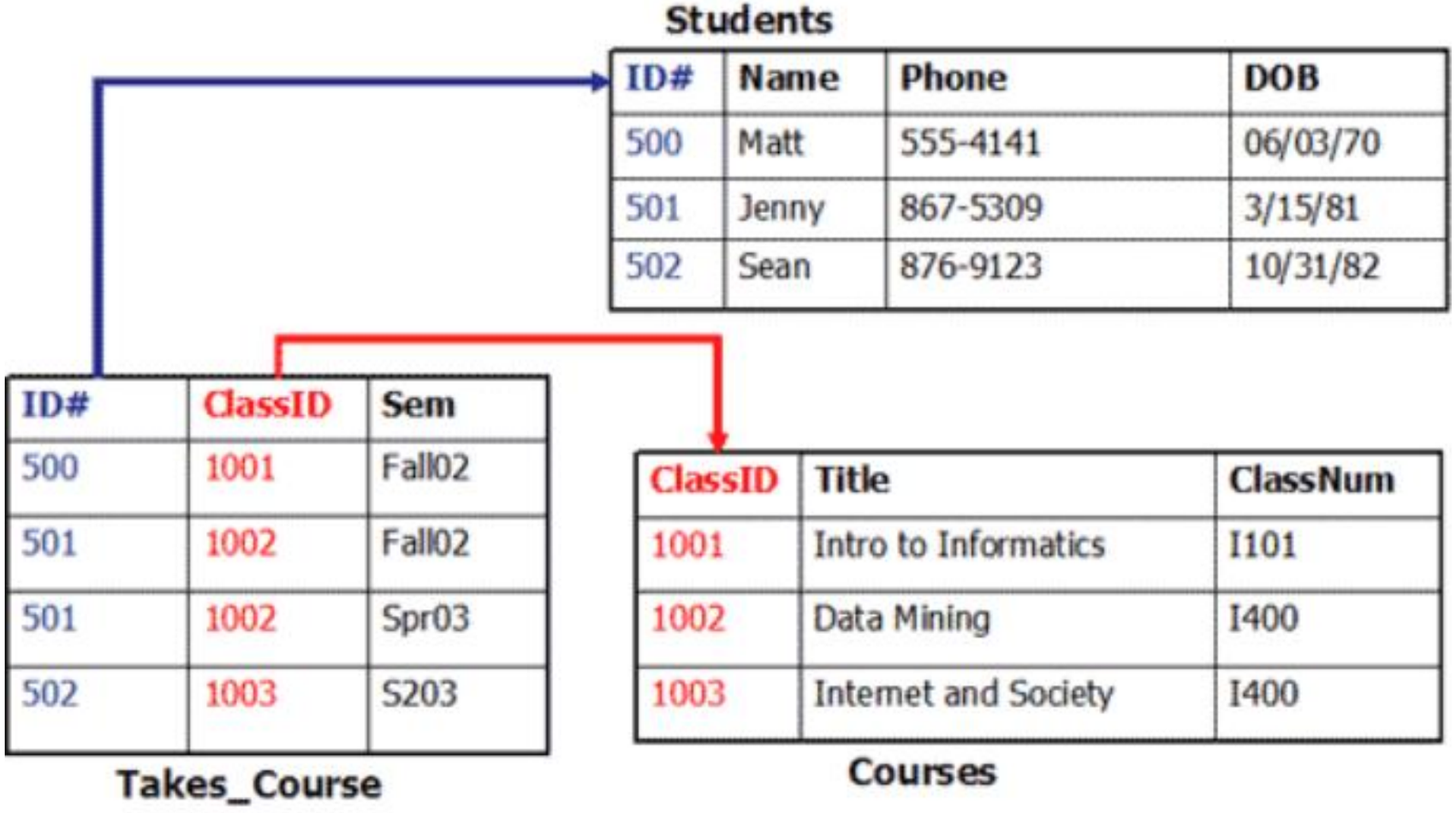
ID#	ClassID	Sem
500	1001	Fall02
501	1002	Fall02
501	1002	Spr03
502	1003	S203

Courses

ClassID	Title	ClassNum
1001	Intro to Informatics	I101
1002	Data Mining	I400
1003	Internet and Society	I400

Data Often Has Many Constraints Too

- value range, format, etc.



Detecting the Problems

- **Schema problems**
  - mispelt names
  - violating constraints (key, uniqueness, foreign key, etc)
- **Data problems**
  - missing values
  - incorrect values, illegal values, outliers
  - synonyms
  - mispellings
  - conflicting data (eg, age and birth year)
  - wrong value formats
  - variations of values
  - duplicate tuples

Deciding Which Problems to Fix

- **Do not have to fix all**
- **Only those that are necessary for the business purpose**
- **Example**
  - extract brand, color, and weight from product descriptions
  - purpose: to allow users to browse products based on brand
  - so brand values should be correct, ideally 100%
  - color and weight values do not have to be entirely correct

Fixing the Problems

- **Good tools exist for certain types of attributes**
  - names, addresses
- **But in general no real good generic tools out there**
- **Much research has been done**
- **People mostly roll their own set of tools**
  - lot of these use dictionaries (aka reference data), rules, ML
- **Using dictionaries (aka reference data)**
  - typically to detect standardization problems and then fix those
- **Using ML**
  - try to predict what a value in a column should be, then compare with the existing value
    - if different, then could be an error

Dirty Data

FirstName	Surname	CompanyName	Address1	Town
peter	jones	jones café	80 riverways	manchester
lisa sefton			76 the avenue	leicester
a baker		bakery baker ltd	7 main road	reading berkshire
Richard	Evans1	Richard's Treats	9 charles Street	Bracknel
Alex		The Alex Centre	13-15 athol street	Bournmouth
Derren	Knight0	Derrens' Delights		Gillingham
Janine		The Janine Way	10 Fleet Place	Brackneill
Katherine	Bolton	Bolton Foods	bond Street	
Emma	Wright	The Write Way Pld	260 Bath road	Birmingham
emma	w	The Write Way	260 Bath rd	Birmingham
David	Smith	Dave's Gifts	PO BOX 21	Leigh
Dave	Smith	Dave's Gift	po box	Leigh Lancs

Un-Standardised

Missing or misspelled

Duplications

Clean Data

FirstName	Surname	CompanyName	Address1	Town
Peter	Jones	Jones Café	80 Riverways	Manchester
Lisa	Sefton		76 The Avenue	Leicester
A	Baker	Bakery Baker Ltd	7 Main Road	Reading
Richard	Evans	Richard's Treats	9 charles Street	Bracknell

Correctly Standardise

Emma	Wright	The Write Way Pld	280 Bath road	Birmingham
emma	w	The Write Way	280 Bath rd	Birmingham
David	Smith	Dave's Gifts	PO BOX 21	Leigh
Dave	Smith	Dave's Gift	po box	Leigh Lancs

Duplications

Clean Data

FirstName	Surname	CompanyName	Address1	Town
Peter	Jones	Jones Café	80 Riverways	Manchester
Lisa	Sefton		76 The Avenue	Leicester
A	Baker	Bakery Baker Ltd	7 Main Road	Reading
Richard	Evans	Richard's Treats	9 charles Street	Bracknell
Alex	Froy	The Alex Centre	13-15 athol street	Bournmouth
Derren	Knight0	Derrens' Delights	25 Camel Lane	Gillingham
Janine	Hutton	The Janine Way	10 Fleet Place	Bracknell
Katherine	Bolton	Bolton Foods	bond Street	London
Emma	Wright	The Write Way Pld	280 Bath road	Birmingham
David	Smith	Dave's Gifts	PO BOX 21	Leigh

Correctly Standardise

Populated and Corre

Duplications Remove

Examples in Industry (see Google Doc)

Additional Transformations

- **These are not to correct something wrong in schema/data per se**
- **Not data cleaning**
- **But rather transformations of schema/data into something better suited for our purposes**
- **Examples**
  - split a field (eg full name)
  - concat of multiple values/fields
  - schema transformation

Example of Splitting an Attribute

Name
John Smith
Henry R. White
Dr. Andy Brown
Steve D. Brook Jr.

Title	First	Middle	Last	Suffix
	John		Smith	
	Henry	R.	White	
Dr.	Andy		Brown	
	Steve	D.	Brook	Jr.

Some Other Possible Steps

- **Data enrichment using additional data sources**

First	Last	Income
John	Smith	$ 32,000
Henry	White	$ 88,000
Andy	Brown	$120,000
Steve	Brook	$ 54,000

Income L	Income U	Target
20000	39999	A
40000	59999	B
60000	79999	C
80000	99999	D
100000	119999	E
120000	139999	F

First	Last	Income	Target
John	Smith	$ 32,000	A
Henry	White	$ 88,000	D
Andy	Brown	$120,000	F
Steve	Brook	$ 54,000	B

Another Example

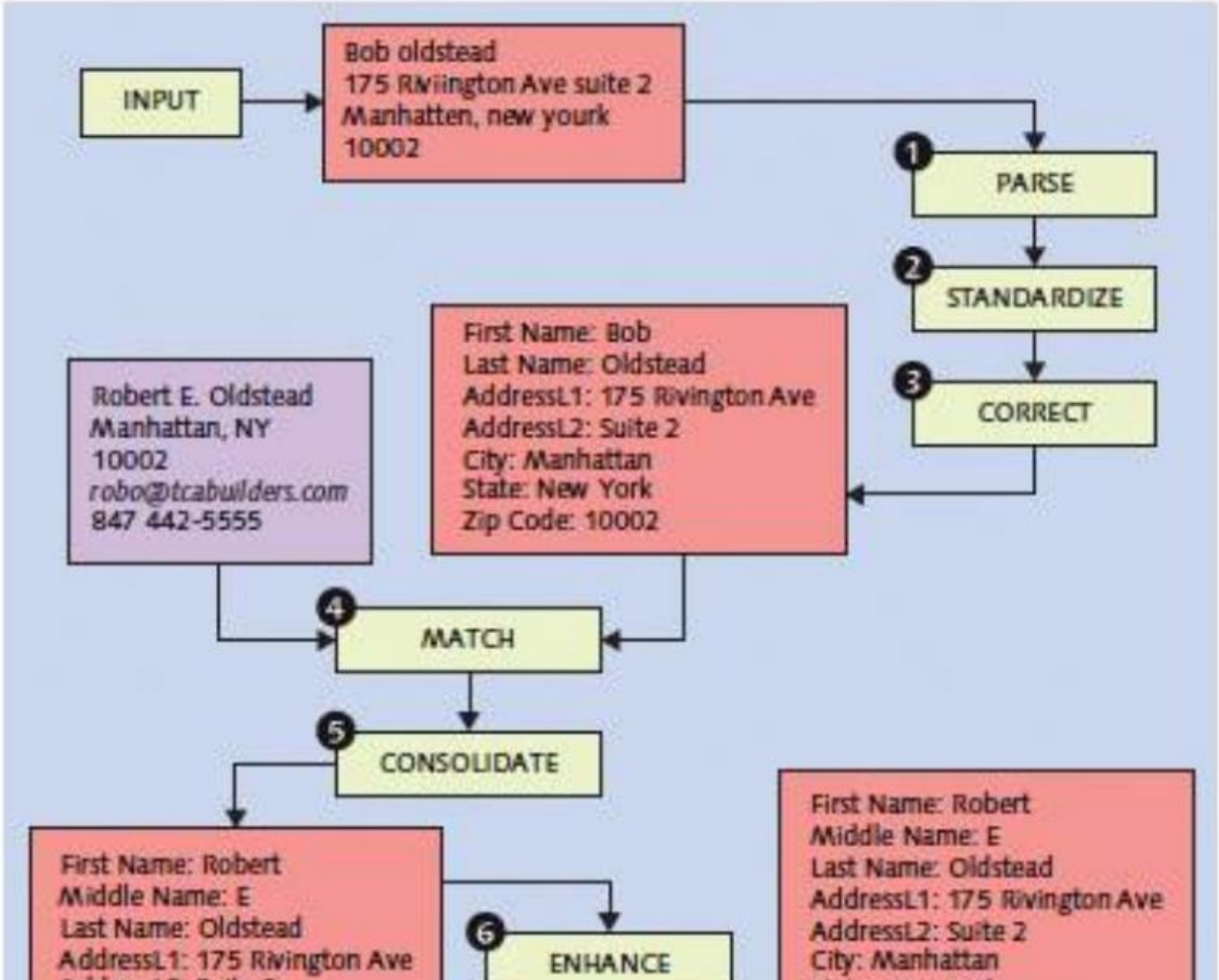
Data Source A

Title	First	Middle	Last	Suffix
	John		Smith	
	Henry	R.	White	
Dr.	Andy		Brown	
	Steve	D.	Brook	Jr.

Data Source B

First	Last	Age
John	Smith	34
Henry	White	19
Andy	Brown	45
Steve	Brook	67

Title	First	Middle	Last	Suffix	Age
	John		Smith		34
	Henry	R.	White		19
Dr.	Andy		Brown		45
	Steve	D.	Brook	Jr.	67


INPUT
Bob oldstead
175 Rivington Ave suite 2
Manhatten, new yourk
10002
1
PARSE
2
STANDARDIZE
3
CORRECT
First Name: Bob
Last Name: Oldstead
AddressL1: 175 Rivington Ave
AddressL2: Suite 2
City: Manhattan
State: New York
Zip Code: 10002
Robert E. Oldstead
Manhattan, NY
10002
robo@tcabuilders.com
847 442-5555
4
MATCH
5
CONSOLIDATE
First Name: Robert
Middle Name: E
Last Name: Oldstead
AddressL1: 175 Rivington Ave
6
ENHANCE
First Name: Robert
Middle Name: E
Last Name: Oldstead
AddressL1: 175 Rivington Ave
AddressL2: Suite 2
City: Manhattan

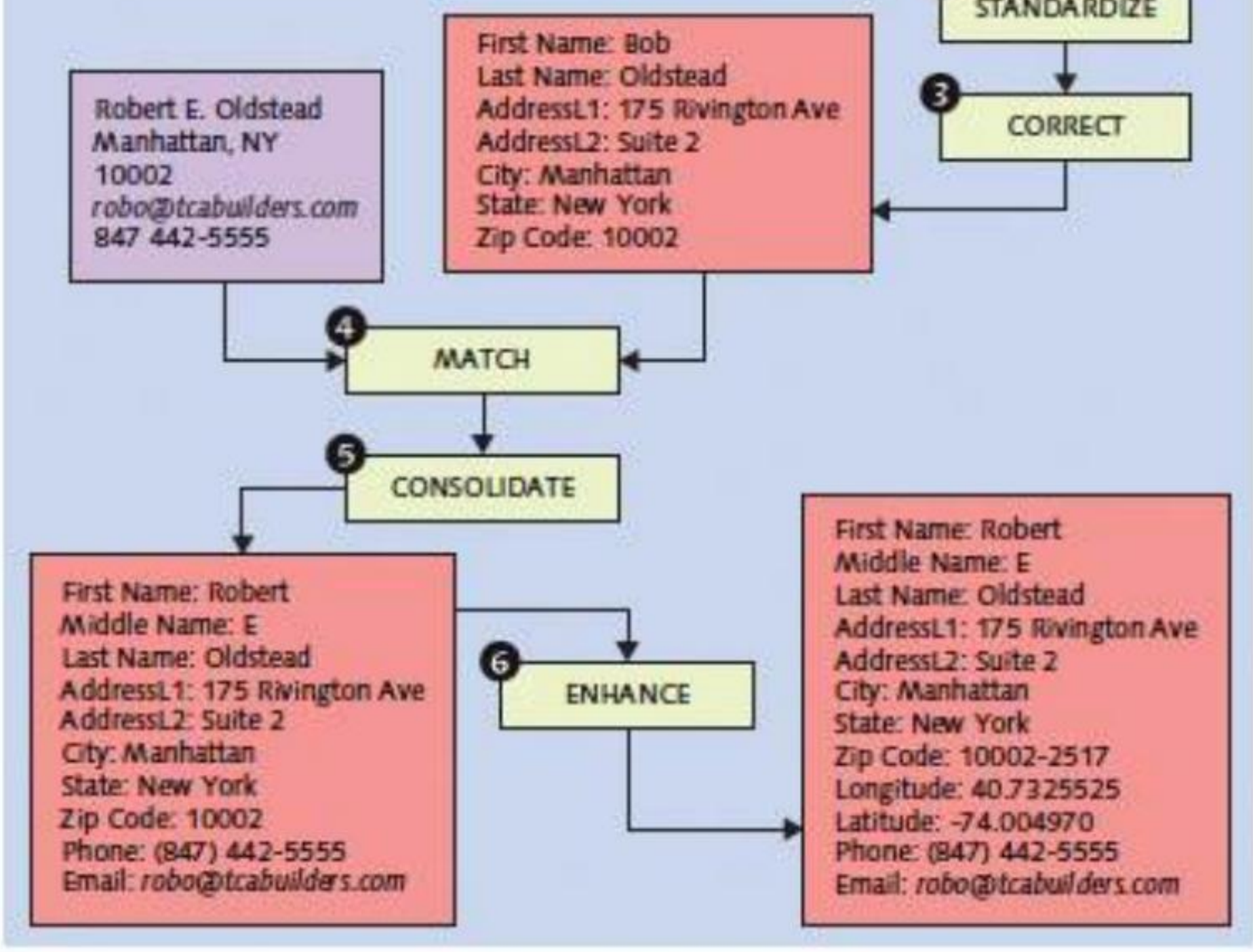


Figure 4.6 Example of the Data Quality Process

Discussion

- **Exploration and cleaning are often required in many steps of the DS pipeline**
  - after integrating multiple data sets, may have to explore and clean again
  - profiling and cleaning are often built into data transformation steps
- **Data quality is a huge topic**
  - the notion of data quality covers many additional aspects
    - freshness, coverage, understandability, trustworthiness, etc.
  - can be addressed within a single DS pipeline, or long-term within a department or the entire company, as a part of data governance
    - e.g., assign an owner to each data set, whose job is to keep the data quality high