COMP9321

Data Services Engineering

Term 1, 2019

Week 2 Lecture 1

Notes for Assignment 1

- For assessment reason, printed data in table rows should be separated by one single space;

- Some inconsistent names in files don’t need to be changed in this assignment(e.g. “**Sant Martí**” and “**Sant Marti**”, also “**Horta-Guinardó**” and “**Horta-Guinardo**”, “**Meridiana**” and “**Av Meridiana**”);

- Names like “**ARAGÓ**” shall be converted to “**Aragó**” ;

- Names in dataset may be in BLOCK LETTERS(Upper cased) or lower cased, they should be corrected to Title Style, except for “**la**”, “**de**”, “**d’** ” and “**l’** ”.
  - e.g. “**El Camp de l’Arpa Del Clot**” ;
  - This is not the real case. This is a simplified rule.

Notes for Assignment 1

- For Q2 - Q5, invalid data like “**Unknown**”, “**–**” shall be removed;

- Multiple street values shall not be changed and be kept as it is;

- In Q4, you need only match ***by hour, day, month*** and ***district names*** of stations and accidents; You can assume all data are in the same year;

- Accident outside the range of air stations can be ignored;

- We are planning an update sample testing and release a less buggy version for Q2 - Q4.

- We expect q1() – q5() to be able to run independent.
- Generally If you want to reuse your code, feel free to add more methods/fuctions

Readings for this course

Recommended readings:

Some are available on GitHub

- Designing Data Intensive Applications, Martin Kleppmann, O'Reilly
- Flask Web Development, Miguel Grinberg, O'Reilly
- Restful Web Clients: Enabling Reuse Through Hypermedia, Mike Amundsen, O'Reilly
- Python Data Science Handbook: Essential Tools for Working with Data, Jake VanderPlas O'Reilly
- Introduction to Machine Learning with Python: A Guide for Data Scientists, Andreas C. Müller and Sarah Guido, O'Reill

Data Cleansing and Integration

COMP9321 2019T1

Credit to ***John Canny*, CS 194** Fall 2014 at ***UC Berkeley***

DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View, CA	$406Bn
Microsoft	Redmond, WA	$392Bn
Intl. Business Machines	Armonk, NY	$194Bn

```
SELECT Market_Cap
From Companies
Where Company_Name = “Apple”
```

```
Number of Rows: 0
```

Problem:
Missing Data

DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View, CA	$406Bn
Microsoft	Redmond, WA	$392Bn
Intl. Business Machines	Armonk, NY	$194Bn

```
SELECT Market_Cap
From Companies
Where Company_Name = “IBM”
```

```
Number of Rows: 0
```

Problem:
Entity Resolution

DB-hard Queries

Company_Name	Address	Market Cap
Google	Googleplex, Mtn. View, CA	$406
Microsoft	Redmond, WA	$392
Intl. Business Machines	Armonk, NY	$194
Hogwarts School of Witchcraft and Wizardry	Scotland, UK	$460

```
SELECT MAX(Market_Cap)
From Companies
```

Number of Rows: 1

Problem:
Unit Mismatch

Who’s Calling Who’S Data Dirty?

Dirty Data

The Statistics View:

- There is a process that produces data
- We want to model ideal samples of that process, but in practice we have non-ideal samples:
  - **Distortion** – some samples are corrupted by a process
  - **Selection Bias** - likelihood of a sample depends on its value
  - **Left and right censorship** - users come and go from our scrutiny
  - **Dependence** – samples are supposed to be independent, but are not (e.g. social networks)
- You can add new models for each type of imperfection, but you can't model everything.
- What's the best trade-off between accuracy and simplicity?

Dirty Data

The Database View:

- I got my hands on this data set
- Some of the values are missing, corrupted, wrong, duplicated
- Results are absolute (relational model)
- You get a better answer by improving the quality of the values in your dataset

Dirty Data

The Domain Expert’s View:

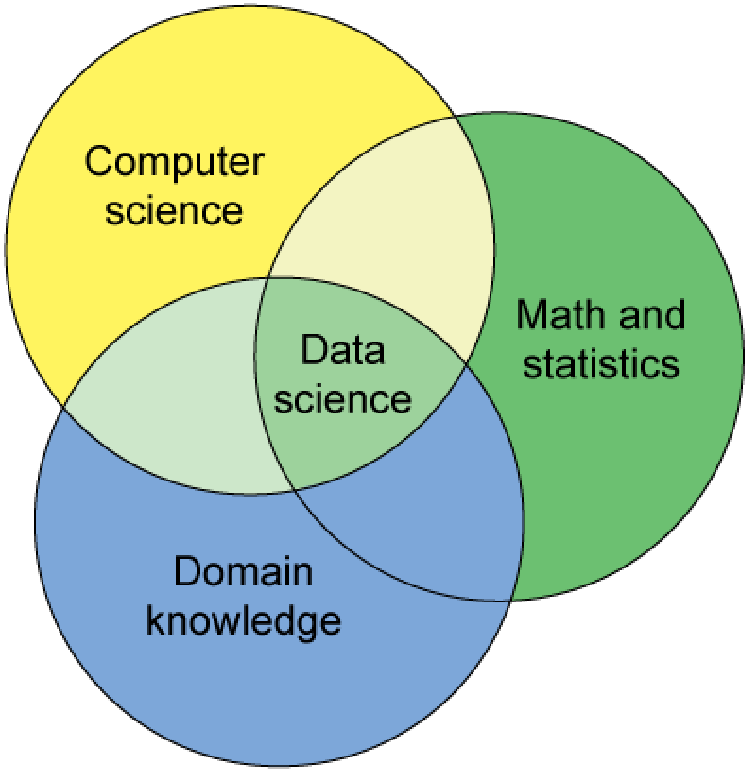
- This Data Doesn’t look right
- This Answer Doesn’t look right
- What happened?

Domain experts have an implicit model of the data that they can test against…

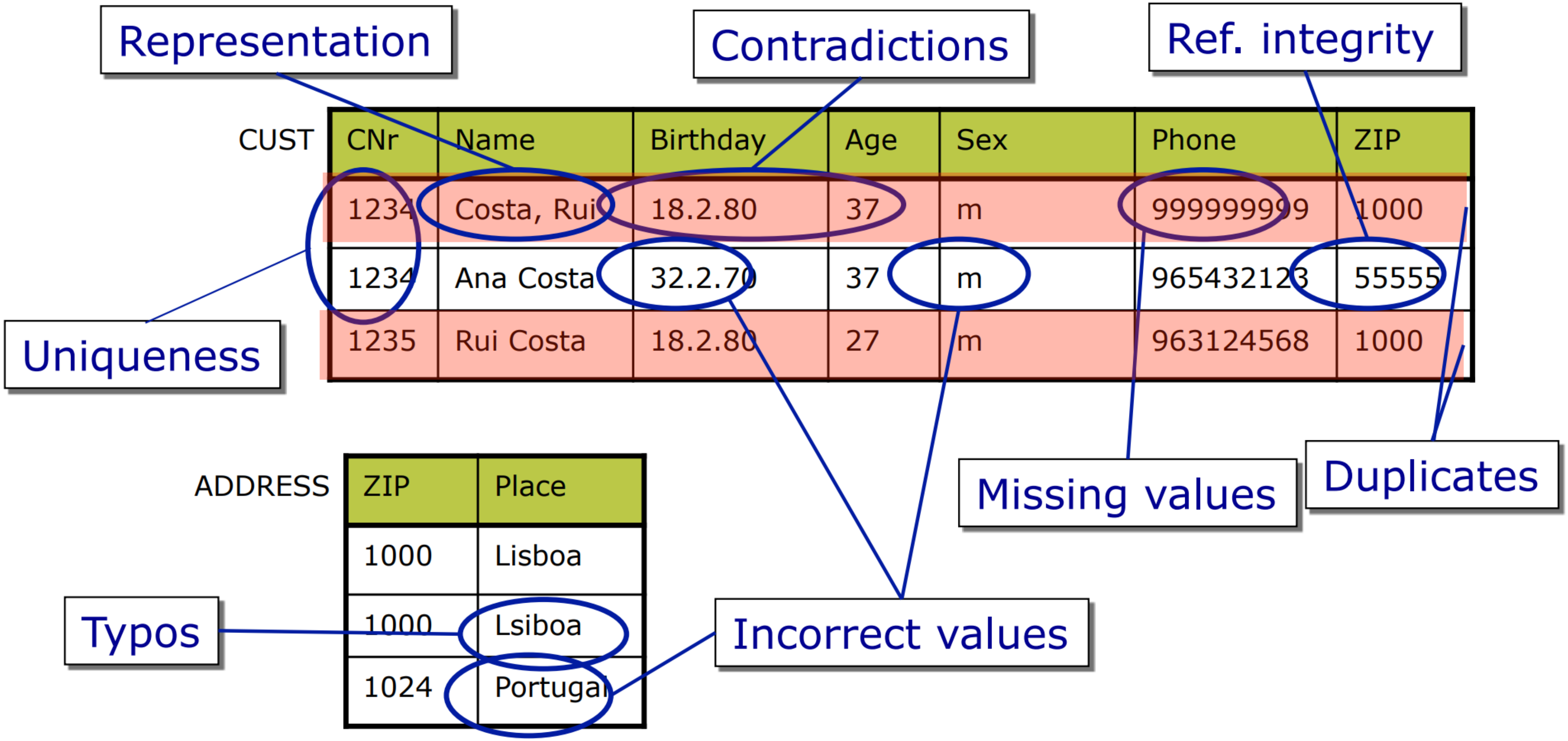
Dirty Data

The Data Scientist’s View:

- Some Combination of all of the above



Example: Data Quality Problems



Representation

Contradictions

Ref. integrity

CUST

CNr	Name	Birthday	Age	Sex	Phone	ZIP
1234	Costa, Rui	18.2.80	37	m	999999999	1000
1234	Ana Costa	32.2.70	37	m	965432123	55555
1235	Rui Costa	18.2.80	27	m	963124568	1000

Uniqueness

Missing values

Duplicates

ADDRESS

ZIP	Place
1000	Lisboa
1000	Lsiboa
1024	Portugal

Typos

Incorrect values

Data Quality Problems

- (Source) Data is dirty on its own.
- Transformations corrupt the data (complexity of software pipelines).
- Data sets are clean but integration (i.e., combining them) screws them up.
- “Rare” errors can become frequent after transformation or integration.
- Data sets are clean but suffer “bit rot”
- Old data loses its value/accuracy over time
- Any combination of the above

Why Data Quality Problems Matter

Incorrect prices in inventory retail databases

□ Costs for consumers 2.5 billion $

□ 80% of barcode-scan-errors to the disadvantage of consumer

IRS 1992: almost 100,000 tax refunds not deliverable

□ 50% to 80% of computerized criminal records in the U.S. were found to be inaccurate, incomplete, or ambiguous. [Strong et al. 1997a]

US-Postal Service: of 100,000 mass mailings up to 7,000 undeliverable due to incorrect addresses [Pierce 2004]

… …

How Data Quality Problems Happen

Incomplete data comes from:

□ non available data value when collected

□ different criteria between the time when the data was collected and when it is analyzed

□ human/hardware/software problems □

Noisy data comes from:

□ data collection: faulty instruments

□ data entry: human or computer errors

□ data transmission

Inconsistent (and duplicate) data comes from:

□ Different data sources, so non-uniform naming conventions/data codes

□ Functional dependency and/or referential integrity violation

Dirty Data Problems

From Stanford Data Integration Course:

1) parsing text into fields (separator issues)
2) Naming conventions: ER: NYC vs New York
3) Missing required field (e.g. key field)
4) Different representations (2 vs Two)
5) Fields too long (get truncated)
6) Primary key violation (from un- to structured or during integration
7) Redundant Records (exact match or other)
8) Formatting issues – especially dates
9) Licensing issues/Privacy/ keep you from using the data as you would like?

Application Scenarios

Integrate data from different sources

□ E.g., populating data from different operational data stores or a mediator-based architecture

Eliminate errors and duplicates within a single source

□ E.g., duplicates in a file of customers

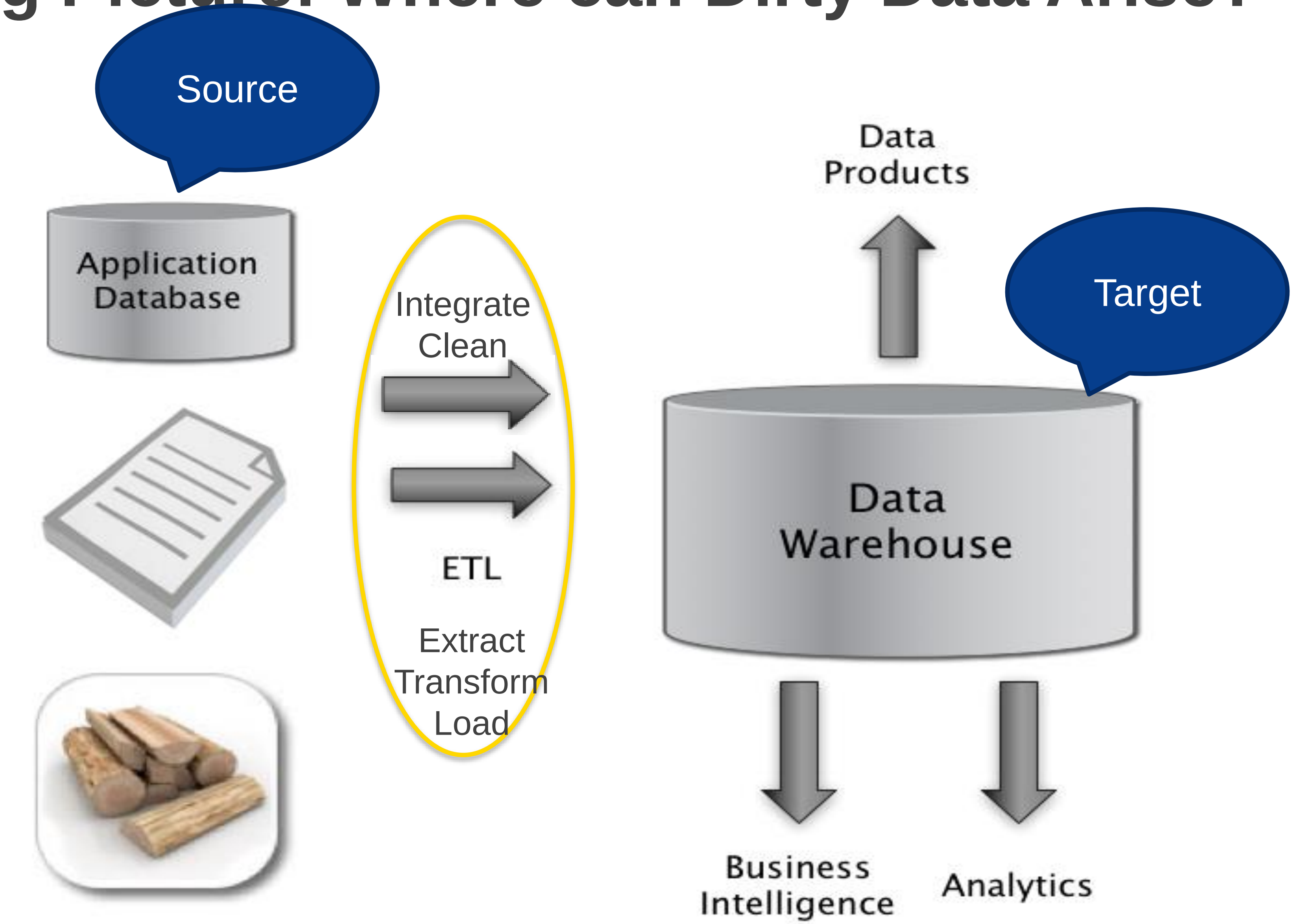
Migrate data from a source schema into a different fixed target schema

□ E.g., discontinued application packages

Convert poorly structured data into structured data

□ E.g., processing data collected from the Web

Big Picture: Where can Dirty Data Arise?



Why Data Cleaning is Important

Activity of converting source data into target data without errors, duplicates, and inconsistencies, i.e., Cleaning and Transforming to get…

- **High-quality data!**

No quality data, no quality decisions!

□ Quality decisions must be based on good quality data (e.g., duplicate or missing data may cause incorrect or even misleading statistics)

Data Quality Problems

Schema level data quality problems

- prevented with better schema design, schema translation and integration.

Instance level data quality problems

- errors and inconsistencies of data that are not prevented at schema level

Data Quality Problems

Schema level data quality problems

- Avoided by an RDBMS
  - Missing data – *product price not filled in*
  - Wrong data type – *“abc” in product price*
  - Wrong data value – *0.5 in product tax (iva)*
  - Dangling data – *category identifier of product does not exist*
  - Exact duplicate data – *different persons with same ssn*
  - Generic domain constraints – *incorrect invoice price*
- Not avoided by an RDBMS
  - Wrong categorical data – *countries and corresponding states*
  - Outdated temporal data – *just-in-time requirement*
  - Inconsistent spatial data – *coordinates and shapes*
  - Name conflicts – *person vs person or person vs client*
  - Structural Conflicts - *addresses*

Data Quality Problems

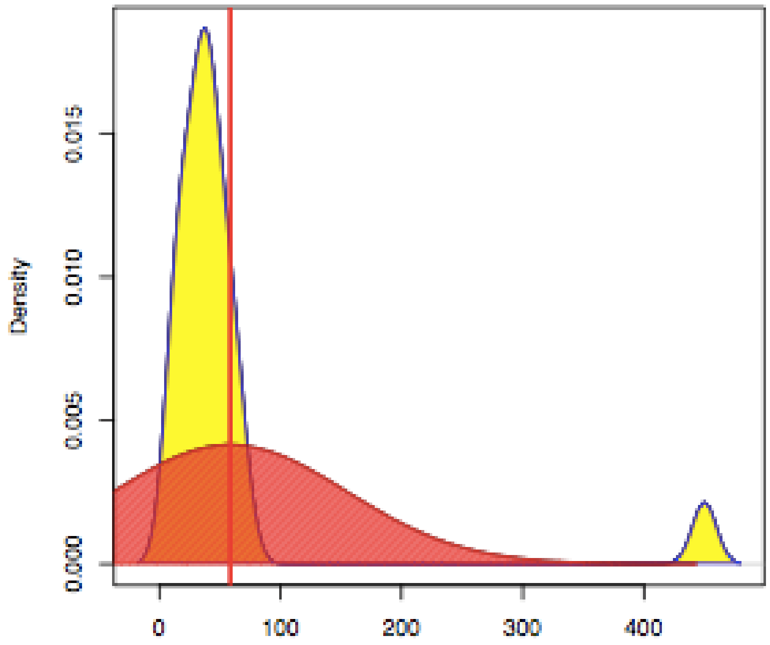
Instance level data quality problems

- Single record
  - Missing data in a *not null* field – *ssn:-9999999*
  - Erroneous data – *price:5 but real price:50*
  - Misspellings: *José Maria Silva vs José Maria Sliva*
  - Embedded values: *Prof. José Maria Silva*
  - Misfielded values: *city: Portugal*
  - Ambiguous data: *J. Maria Silva; Miami Florida,Ohio*
- Multiple records
  - Duplicate records: *Name:Jose Maria Silva, Birth:01/01/1950 and Name:José Maria Sliva, Birth:01/01/1950*
  - Contradicting records: *Name:José Maria Silva, Birth:01/01/1950 and Name:José Maria Silva, Birth:01/01/1956*
  - Non-standardized data: *José Maria Silva vs Silva, José Maria*

Numeric Outliers

12	13	14	21	22	26	33	35	36	37	39	42	45	47	54	57	61	68	450

ages of employees (US)



- median 37
- mean 58.52632
- variance 9252.041

Adapted from Joe Hellerstein's 2012 CS 194 Guest Lecture

Integration error

Data 1

…	Date(mm/dd/yyyy)	…
…	08/02/2019	…
…	09/02/2019	…

Data 2

…	Date(dd/mm/yyyy)	…
…	08/08/2019	…
…	09/08/2019	…

…	Date(mm/dd/yyyy)	…
…	08/02/2019	…
…	09/02/2019	…
…	08/08/2019	…
…	09/08/2019	…

ETL error

…	Address	…
…	Google, 1600 Amphitheatre Parkway, Mountain View, CA	…
…	LINUS MEDIA GROUP INC. 104-18643 52ND AVE, SURREY, BC, CA	…

↓

…	Address	…
…	Google, 1600 Amphitheatre Parkway, Mountain View, California	…
…	LINUS MEDIA GROUP INC. 104-18643 52ND AVE, SURREY, BC, California	…

ETL error

…	Address	…
…	**Building K17, UNSW Sydney, NSW 2052**	…
…	**Building k17, UNSW, Sydney, NSW 2052**	…

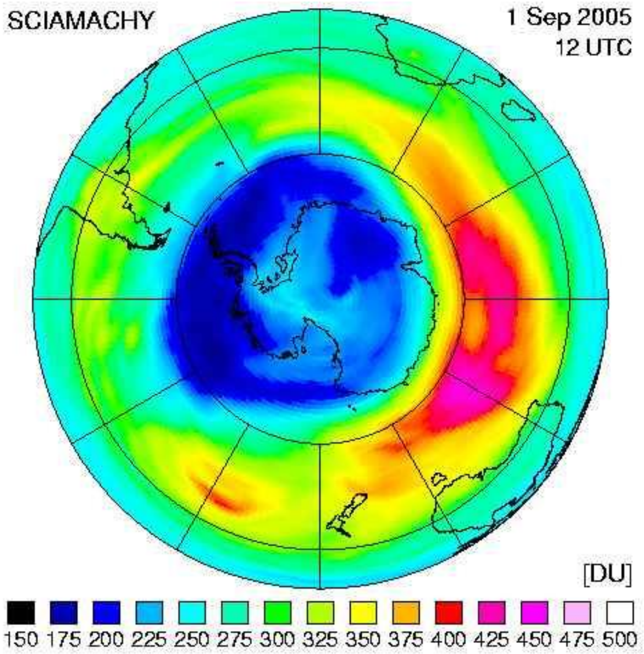
↓

Addr1	Bldg K17	Bldg K17
Addr2		UNSW
City	UNSW Sydney	Sydney
State	NSW	NSW
Postcode	2052	2052

Data Cleaning Makes Everything Okay?

The appearance of a hole in the earth's ozone layer over Antarctica, first detected in 1976, was so unexpected that scientists didn't pay attention to what their instruments were telling them;

they thought their instruments were malfunctioning.



In fact, the data were rejected as unreasonable by data quality control algorithms

National Center for Atmospheric Research

Conventional Definition of Data Quality

Accuracy

- The data was recorded correctly.

Completeness

- All relevant data was recorded.

Uniqueness

- Entities are recorded once.

Timeliness

- The data is kept up to date.
  - Special problems in federated data: time consistency.

Consistency

- The data agrees with itself.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Accuracy

Closeness between a value v and a value v'

considered as the correct representation of the realworld phenomenon that v aims to represent.

- Ex: for a person name "John", v'=John is correct, v=Jhn is incorrect

Syntatic accuracy

closeness of a value v to the elements of the corresponding definition domain D

- Ex: if v=Jack, even if v'=John , v is considered syntactically correct because it is an admissible value in the domain of people names.
- Measured by means of comparison functions (e.g., edit distance) that returns a score

Accuracy

Semantic accuracy

closeness of the value v to the true value v’

- Measured with a <yes, no> or <correct, not correct> domain
- Coincides with correctness
- The corresponding true value has to be known

Ganularity of accuracy definition

Accuracy may refer to:

- a single value of a relation attribute
- an attribute or column
- a relation
- the whole database

Completeness

“The extent to which data are of sufficient breadth, depth, and scope for the task in hand.”

Three types:

- **Schema completeness**: degree to which concepts and their properties are not missing from the schema
- **Column completeness**: evaluates the missing values for a specific property or column in a table.
- **Population completeness**: evaluates missing values with respect to a reference population

Completeness of relational data

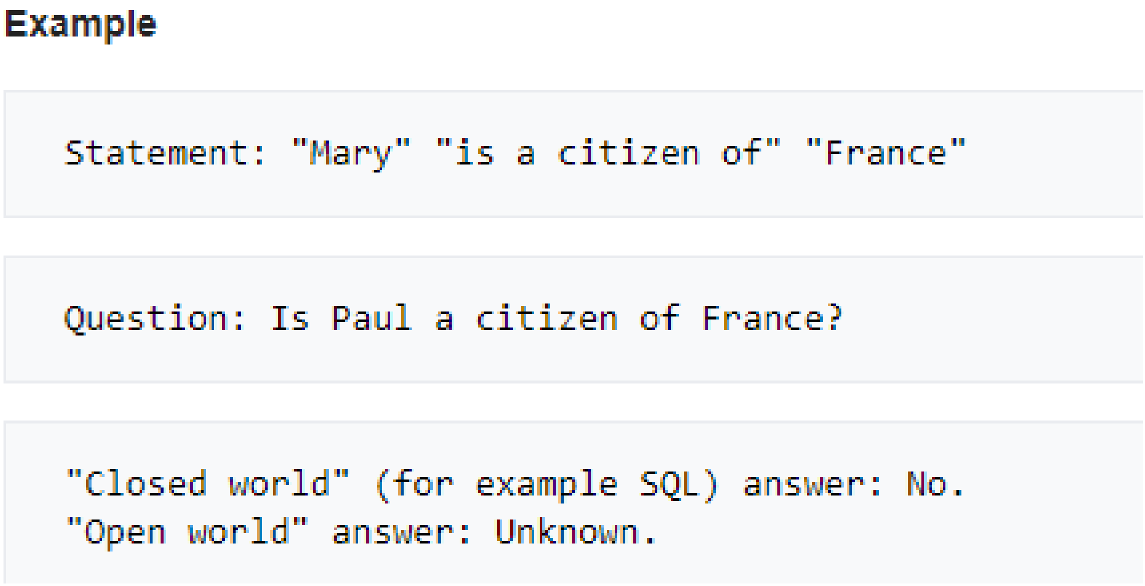
The **completeness of a table** characterizes the extent to which the table represents the real world.

- **The presence/absence and meaning of null values**

Example: Person(name, surname, birthdate, email), if email is null may indicate the person has no mail (no incompleteness), email exists but is not known (incompletenss), is is not known whether Person has an email (incompleteness may not be the case)

Completeness of relational data

- **Validity of open world assumption (OWA) or closed world assumption (CWA)**
  - –OWA: cannot state neither the truth or falsity of facts not represented in the tuples of a relation
  - –CWA: only the values actually present in a relational table and no other values represent facts of the real world.



Example

```
Statement: "Mary" "is a citizen of" "France"
```

```
Question: Is Paul a citizen of France?
```

```
"Closed world" (for example SQL) answer: No.
"Open world" answer: Unknown.
```

Time-related Dimensions

Currency:

concerns how promptly data are updated

- Example:
  - –if the residential address of a person is updated (it corresponds to the address where the person lives) then the currency is high

Volatility:

characterizes the frequency with which data vary in time

- Example:
  - –Birth dates (volatility zero) vs stock quotes (high degree of volatility)

Time-related Dimensions

Timeliness

- expresses how current data are for the task in hand

- Example:

  – The timetable for university courses can be current by containing the most recent data, but it cannot be timely if it is available only after the start of the classes.

Consistency

Captures the violation of semantic rules defined over a set of data items, where data items can be tuples of relational tables or records in a file

- Integrity constraints in relational data
  - Domain constraints, Key, inclusion and functional dependencies
- Data edits: semantic rules in statistics

Others

□ Interpretability: concerns the documentation and metadata that are available to correctly interpret the meaning and properties of data sources

□ Synchronization between different time series: concerns proper integration of data having different time stamps.

□ Accessibility: measures the ability of the user to access the data from his/her own culture, physical status/functions, and technologies available.

Problems …

Unmeasurable

- Accuracy and completeness are extremely difficult, perhaps impossible to measure.

Context independent

- No accounting for what is important. E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.

Vague

- The conventional definitions provide no guidance towards practical improvements of the data.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Finding a modern definition

We need a definition of data quality which

- Reflects the **use** of the data
- Leads to **improvements in processes**
- Is **measurable** (we can define metrics)

First, we need a better understanding of how and where data quality problems occur

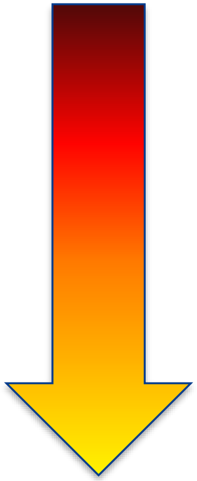
- The data quality continuum

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

The Data Quality Continuum

Data and information is not static, it flows in a data collection and usage process

- Data gathering
- Data delivery
- Data storage
- Data integration
- Data retrieval
- Data mining/analysis



Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Gathering

How does the data enter the system?

Sources of problems:

- Manual entry
- No uniform standards for content and formats
- Parallel data entry (duplicates)
- Approximations, surrogates – SW/HW constraints

software - hardware

- Measurement or sensor errors.

Adapted from Ted Johnson's SIGMOD 20[illegible] Tutorial

Data Gathering - Solutions

Potential Solutions:

- Preemptive:
  - –Process architecture (build in integrity checks)
  - –Process management (reward accurate data entry, data sharing, data stewards)
- Retrospective:
  - –Cleaning focus (duplicate removal, merge/purge, name & address matching, field value standardization)
  - –Diagnostic focus (automated detection of glitches).

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Delivery

Destroying or mutilating information by inappropriate pre-processing

- Inappropriate aggregation
- Nulls converted to default values

Loss of data:

- Buffer overflows
- Transmission problems
- No checks

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Delivery - Solutions

Build reliable transmission protocols

- Use a relay server

Verification

- Checksums, verification parser
- Do the uploaded files fit an expected pattern?

Relationships

- Are there dependencies between data streams and processing steps

Interface agreements

- Data quality commitment from the data stream supplier.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Storage

You get a data set. What do you do with it?

Problems in physical storage

- Can be an issue, but terabytes are cheap.

Problems in logical storage

- Poor metadata.
  - Data feeds are often derived from application programs or legacy data sources. What does it mean?
- Inappropriate data models.
  - Missing timestamps, incorrect normalization, etc.
- Ad-hoc modifications.
  - Structure the data to fit the GUI.
- Hardware / software constraints.
  - Data transmission via Excel spreadsheets, Y2K

Adapted from Ted Johnson’s SIGMOD 2003 Tutorial

Data Storage - Solutions

Metadata

- Document and publish data specifications.

Planning

- Assume that everything bad will happen.
- Can be very difficult.

Data exploration

- Use data browsing and data mining tools to examine the data.
  - Does it meet the specifications you assumed?
  - Has something changed?

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Retrieval

Exported data sets are often a view of the actual data. Problems occur because:

- Source data not properly understood.
- Need for derived data not understood.
- Just plain mistakes.
  - Inner join vs. outer join
  - Understanding NULL values

Computational constraints

- E.g., too expensive to give a full history, we' ll supply a snapshot.

Incompatibility

- Ebcdic? Unicode?

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Data Mining and Analysis

What are you doing with all this data anyway?

Problems in the analysis.

- Scale and performance
- Confidence bounds?
- Black boxes and dart boards
- Attachment to models
- Insufficient domain expertise
- Casual empiricism： use an arbitrary number to support a pre-conception

Adapted from Ted Johnson’s SIGMOD 2003 Tutorial

Retrieval and Mining - Solutions

Data exploration

- Determine which models and techniques are appropriate, find data bugs, develop domain expertise.

Continuous analysis

- Are the results stable? How do they change?

Accountability

- Make the analysis part of the feedback loop.

Adapted from Ted Johnson's SIGMOD 2003 Tutorial

Taxonomy of Data Quality Problems

- Value-level
- Value-set (attribute/column) level
- Record level
- Relation level
- Multiple relations level

Source: Oliveira 2009

Value-Level

Missing value: value not filled in a not null attribute

□ Ex: birth date = ‘’

Syntax violation: value does not satisfy the syntax rule defined for the attribute

□ Ex: zip code = 27655-175; syntactical rule: xxxx-xxx

Spelling error

□ Ex: city = ‘Sidney’, instead of ‘Sydney’

Domain violation: value does not belong to the valid domain set

□ Ex: age = 240; age: {0, 120}

Value-set and Record levels

Value-set level

□ Existence of synonyms: attribute takes different values, but with the same meaning

Ex: emprego = ‘futebolista’; emprego = ‘jogador futebol’

□ Existence of homonyms: same word used with diff meanings

Ex: same name refers to different authors of a publication

□ Uniqueness violation: unique attribute takes the same value more than once

Ex: two clients have the same ID number

□ Integrity contraint violation

Ex: sum of the values of percent attribute is more than 100 Record level

□ Integrity constraint violation

Ex: total price of a product is different from price plus taxes

Record level

- Integrity constraint violation

Ex: total price of a product is different from price plus taxes

Relation Level

Heterogeneous data representations: different ways of representing the same real world entity

- Ex: name = ‘John Smith’; name = ‘Smith, John’

Functional dependency violation

- Ex: (2765-175, ‘Estoril’) and (2765-175, ‘Oeiras’)

Existence of approximate duplicates

- Ex: (1, André Fialho, 12634268) and (2, André Pereira Fialho, 12634268)

Integrity constraint violation

- Ex: sum of salaries is superior to the max established

Multiple Tables Level

Heterogeneous data representations

□ Ex: one table stores meters, another stores inches

Existence of synonyms: **Fair**: just, objective, impartial, unbiased

Existence of homonyms: **Address** - to speak to / location

Different granularities: same real world entity represented with diff. granularity levels

□ Ex: age: {0-30, 31-60, > 60}; age: {0-25, 26-40, 40-65, >65}

Referential integrity violation

Existence of approximate duplicates

Integrity constraint violation

Data Quality Process

Data Quality Auditing (Assessment)

□ Data Profiling

□ Data Analysis

Data Quality Improvement

□ Data Cleaning

□ Data Enrichment

Data Quality Auditing

Constituted by:

□ Data profiling – analysing data sources to identify data quality problems

□ Data analysis – statistical evaluation, logical study and application of data mining algorithms to define data patterns and rules

Main goals:

□ To obtain a definition of the data: metadata collection

□ To check violations to metadata definition

□ To detect other data quality problems that belong to a given taxonomy

□ To supply recommendations in what concerns the data cleaning task

Data Profiling

Data source discovery

□ Metadata

Schema discovery

□ Schema matching and mapping

□ Profiling for metadata (keys, foreign keys, data types, …)

Data discovery

□ Column-level: Null-values, domains, patterns, value distributions / histograms

□ Table-level: Data mining, rules

Methodology for Data Cleaning

1. Extraction of the individual fields that are relevant
2. Standardization of record fields
3. Correction of data quality problems at value level, e.g., missing values, syntax violation, etc
4. Correction of data quality problems at value-set level and record level, e.g., Synonyms, homonyms, uniqueness violation, integrity constraint violation, etc
5. Correction of data quality problems at relation level, e.g., Violation of functional dependencies, duplicate elimination, etc
6. Correction of data quality problems problems at multiple relations level □ Referential integrity violation, duplicate elimination, etc
7. User feedback □ To solve instances of data quality problems not addressed by automatic methods
8. Effectiveness of the data cleaning and transformation process must be always measured for a sample of the data set

Data Cleaning Activities

1. Extraction from sources

□ Technical and syntactic obstacles

2. Transformation

□ Schematic obstacles

3. Standardization

□ Syntactic and semantic obstacles

4. Duplicate detection

□ Similarity functions

□ Algorithms

5. Data fusion / consolidation /integration

□ Semantic obstacles

6. Loading into warehouse / presenting to user

Data Cleaning Activities

1. Extraction from sources

□ Technical and syntactic obstacles

2. Transformation

□ Schematic obstacles

3. Standardization

□ Syntactic and semantic obstacles

4. Duplicate detection

□ Similarity functions

□ Algorithms

5. Data fusion / consolidation /integration

□ Semantic obstacles

6. Loading into warehouse / presenting to user

Standardization/normalization

- Data read from source may not have the correct format (e.g., reading integer as a string)
- Some strings in the data have spacing which might not play well with your analysis at some point.
- The date/time format may not appropriate for your analysis
- Some times the data is generated by a computer program, so it probably has some computer-generated column names, too. Those can be hard to read and understand while working.

Standardization/normalization

- Example1 (change data type on read):

```
df = pd.read_csv(‘mydata.csv’,
dtype={‘Integer_Column’: int})
```

- Example2 (change data type in dataframe)

```
df['column'] = df['column'].to_numeric()
df['column'] = df['column'].astype(str)
```

- Example3 (Spacing within the values):

```
data[‘Column_with_spacing’].str.strip()
```

Standardization/normalization

- Example4 (unnecessary time item in the date field):

```
df[‘MonthYear’] =
pd.to_datetime(df[‘MonthYear’])
df[‘MonthYear’] = df[‘MonthYear’].apply(lambda
x: x.date())
```

- Example5 (rename columns)

```
 data = data.rename(columns =
{‘Bad_Name1’:Better_Name1’,
‘Bad_Name2’:’Better_name2’})
```

Data integration

- Task of presenting a unified view of data owned by heterogeneous and distributed data sources
- Two sub-activities:
  - **Quality-driven query processing**: task of providing query results on the basis of a quality characterization of data at sources
  - **Instance-level conflict resolution**: task of identifying and solving conflicts of values referring to the same real-world objects.

Merging Data

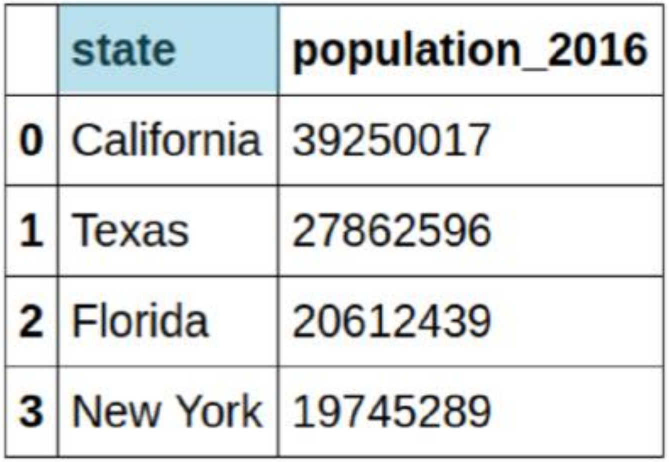
- Sometimes in order to have complete dataset you need to Concatenate two datasets

Example:

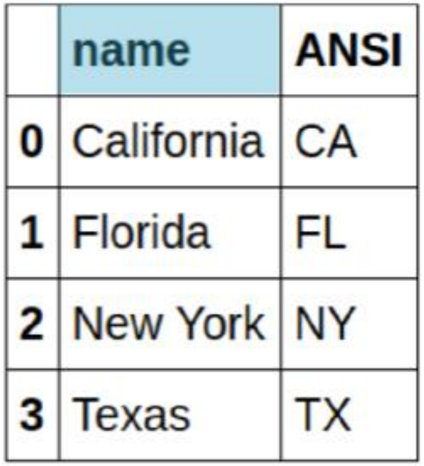
```
Dataset1=pd.read_csv(‘datasets/project1/dataset1
.csv’)
Dataset2=pd.read_csv(‘datasets/project1/dataset2
.csv’)


Full_data=pd.concat[Dataset1, Dataset2] axis=0,
ignore_index=True)
```

Merging Data (Cont’d)

- Sometimes in order to have complete dataset you need to merge two Dataframes



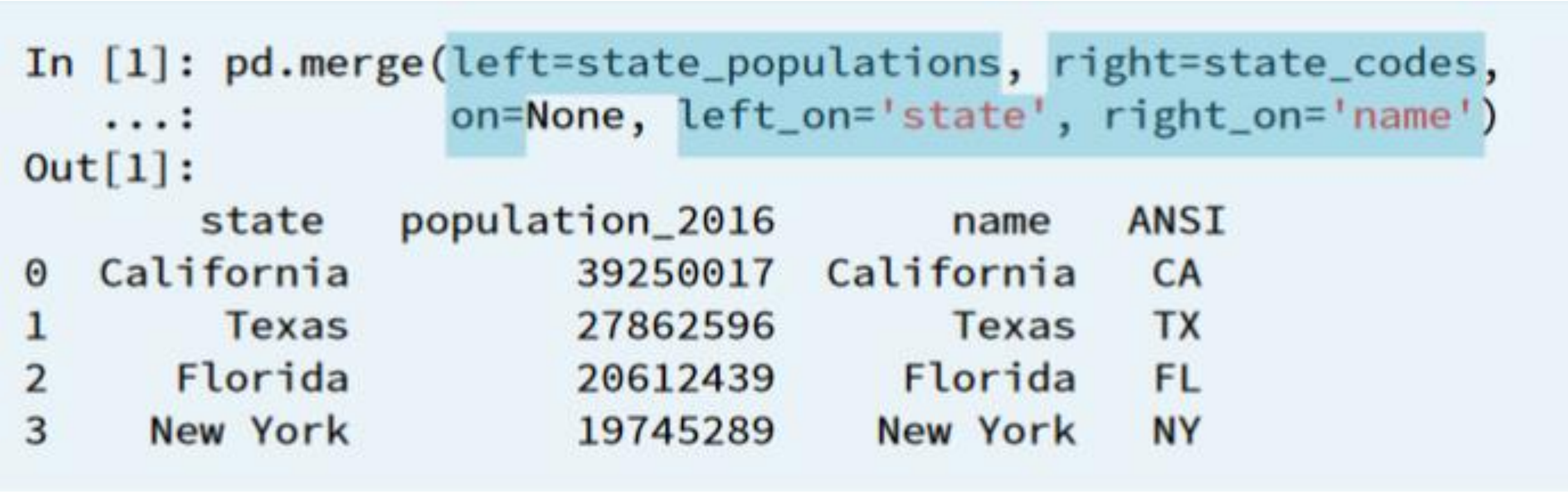
	state	population_2016
0	California	39250017
1	Texas	27862596
2	Florida	20612439
3	New York	19745289



	name	ANSI
0	California	CA
1	Florida	FL
2	New York	NY
3	Texas	TX

https://s3.amazonaws.com/assets.datacamp.com/production/course_3639/slides/ch3_slides.pdf

Merging Data (Cont'd)



```
In [1]: pd.merge(left=state_populations, right=state_codes,
   ...:          on=None, left_on='state', right_on='name')
Out[1]:
       state  population_2016        name ANSI
0 California         39250017  California   CA
1      Texas         27862596       Texas   TX
2    Florida         20612439     Florida   FL
3   New York         19745289    New York   NY
```

https://s3.amazonaws.com/assets.datacamp.com/production/course_3639/slides/ch3_slides.pdf

Duplicate Record Detection

Resolve multiple different mentions:

- Entity Resolution
- Reference Reconciliation
- Object Identification/Consolidation

Remove Duplicates

- Merge/Purge

Record Linking (across data sources)

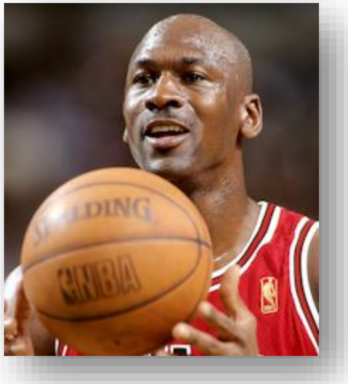
Householding (interesting special case)

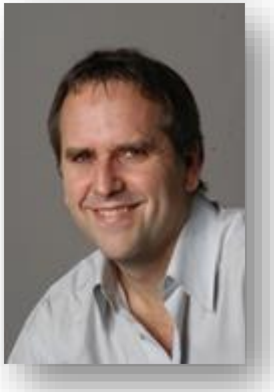
Approximate Match (accept fuzziness)

…

Example:Duplications

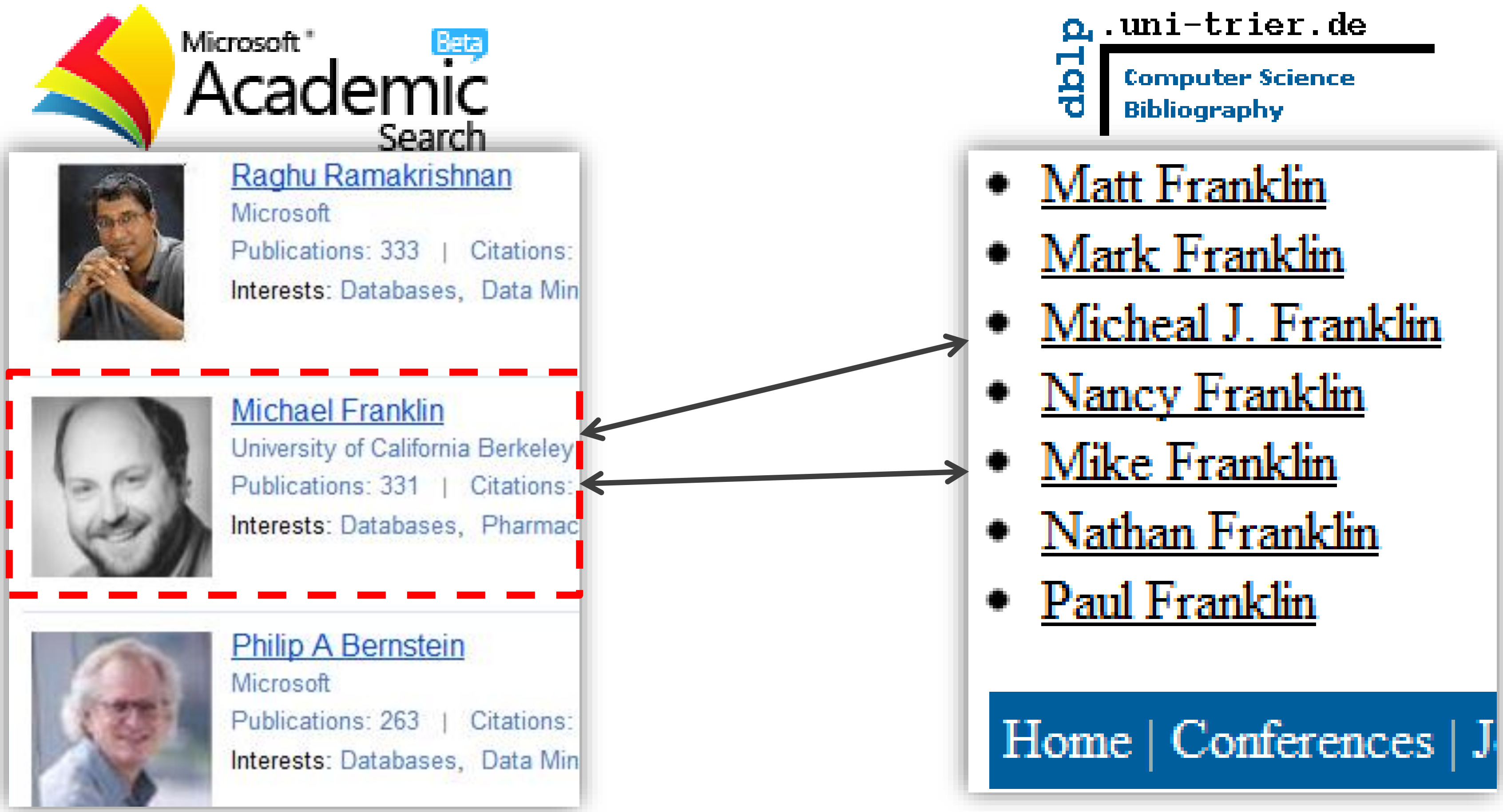
Michael Jordan





Apple

Example: Data Integration



Microsoft Academic Search Beta

Raghu Ramakrishnan
Microsoft
Publications: 333 | Citations:
Interests: Databases, Data Min

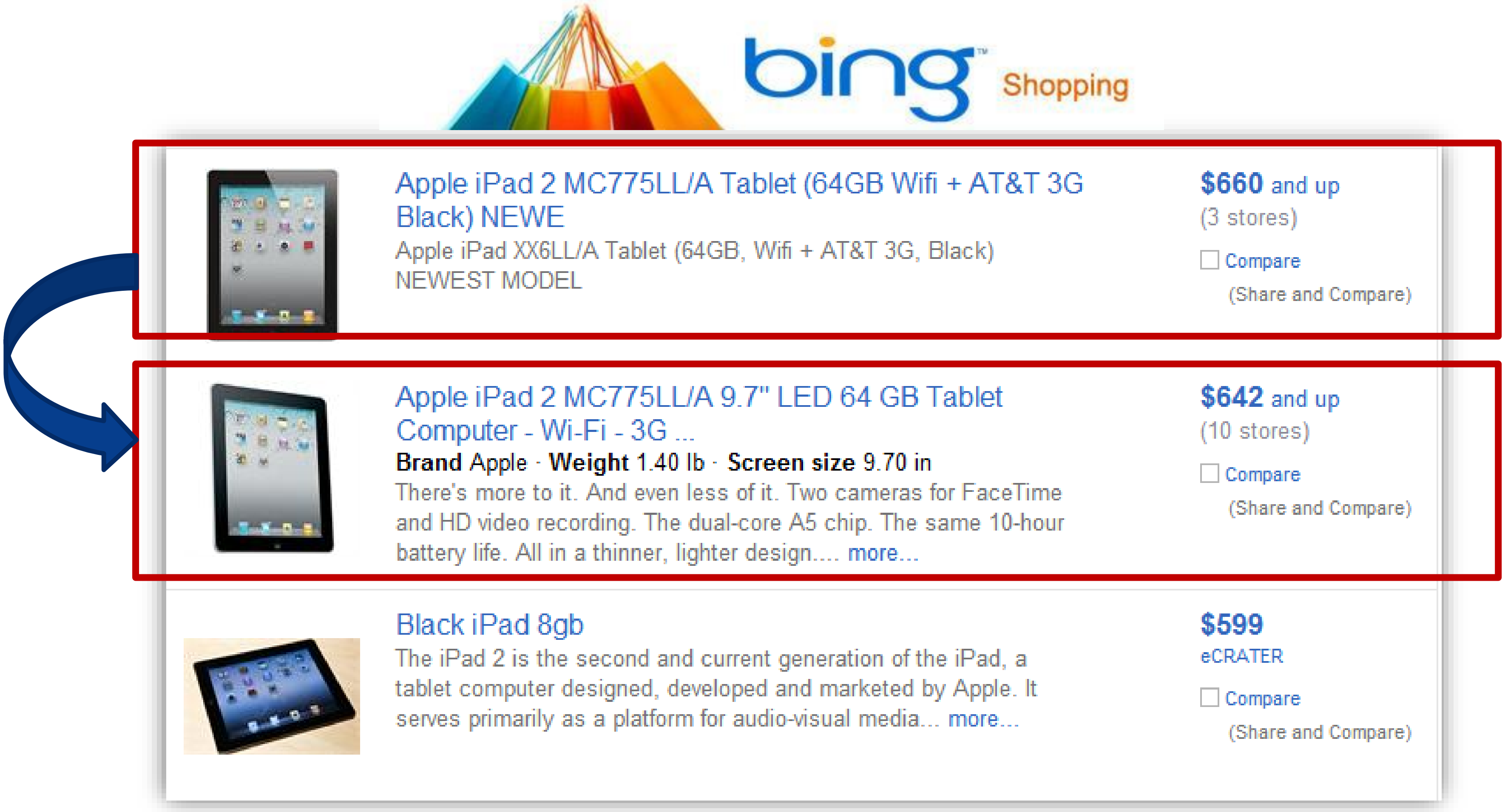
Michael Franklin
University of California Berkeley
Publications: 331 | Citations:
Interests: Databases, Pharmac

Philip A Bernstein
Microsoft
Publications: 263 | Citations:
Interests: Databases, Data Min

dblp.uni-trier.de
Computer Science Bibliography

- Matt Franklin
- Mark Franklin
- Micheal J. Franklin
- Nancy Franklin
- Mike Franklin
- Nathan Franklin
- Paul Franklin

Example: DeDup/Cleaning



Apple iPad 2 MC775LL/A Tablet (64GB Wifi + AT&T 3G Black) NEWE

Apple iPad XX6LL/A Tablet (64GB, Wifi + AT&T 3G, Black) NEWEST MODEL

$660 and up
(3 stores)
□ Compare
(Share and Compare)

Apple iPad 2 MC775LL/A 9.7" LED 64 GB Tablet Computer - Wi-Fi - 3G ...

Brand Apple · **Weight** 1.40 lb · **Screen size** 9.70 in

There's more to it. And even less of it. Two cameras for FaceTime and HD video recording. The dual-core A5 chip. The same 10-hour battery life. All in a thinner, lighter design.... more...

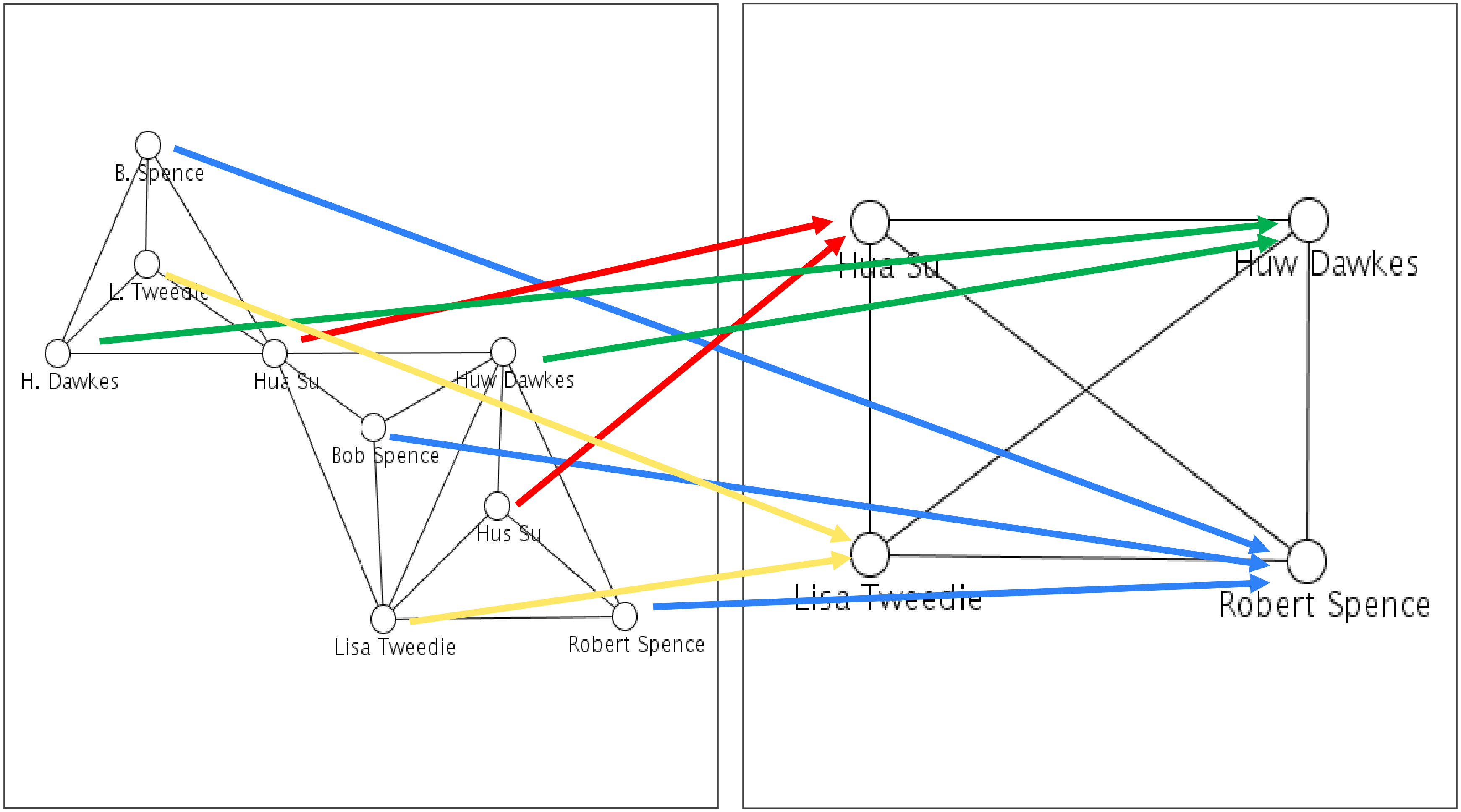
$642 and up
(10 stores)
□ Compare
(Share and Compare)

Black iPad 8gb

The iPad 2 is the second and current generation of the iPad, a tablet computer designed, developed and marketed by Apple. It serves primarily as a platform for audio-visual media... more...

$599
eCRATER
□ Compare
(Share and Compare)

Example: Network Analysis



before

after

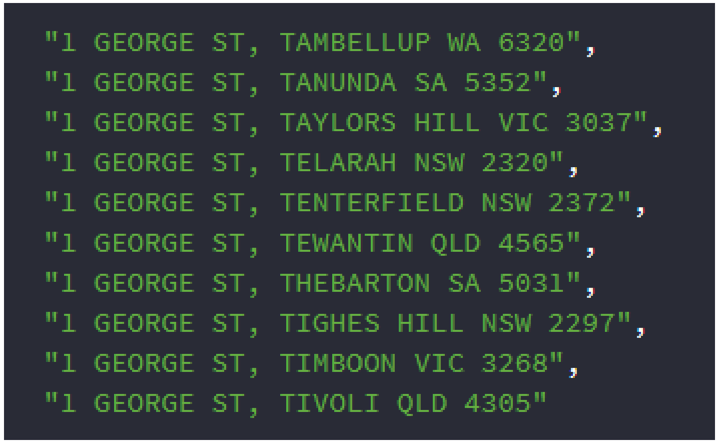
From: Getoor & Machanavajjhala: “Entity Resolution Tutorial”, VLDB 2012

Preprocessing/Standardization

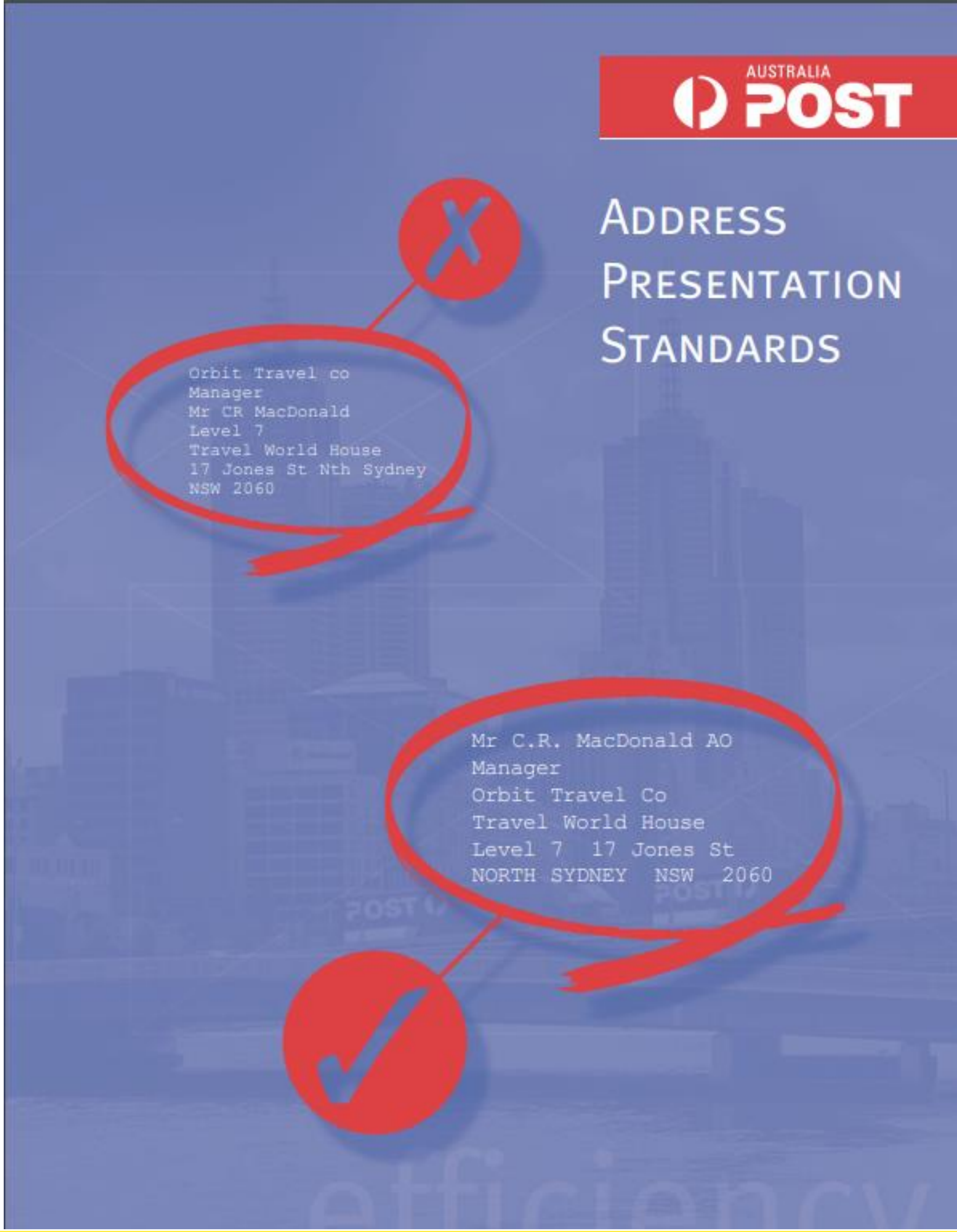
Simple idea:

Convert to canonical form

e.g. addresses



```
"1 GEORGE ST, TAMBELLUP WA 6320",
"1 GEORGE ST, TANUNDA SA 5352",
"1 GEORGE ST, TAYLORS HILL VIC 3037",
"1 GEORGE ST, TELARAH NSW 2320",
"1 GEORGE ST, TENTERFIELD NSW 2372",
"1 GEORGE ST, TEWANTIN QLD 4565",
"1 GEORGE ST, THEBARTON SA 5031",
"1 GEORGE ST, TIGHES HILL NSW 2297",
"1 GEORGE ST, TIMBOON VIC 3268",
"1 GEORGE ST, TIVOLI QLD 4305"
```



AUSTRALIA POST

ADDRESS PRESENTATION STANDARDS

Orbit Travel co
Manager
Mr CR MacDonald
Level 7
Travel World House
17 Jones St Nth Sydney
NSW 2060

Mr C.R. MacDonald AO
Manager
Orbit Travel Co
Travel World House
Level 7 17 Jones St
NORTH SYDNEY NSW 2060