POLITECNICO
MILANO 1863

SYSTEMS AND METHODS FOR BIG AND UNSTRUCTURED DATA

Wrangling

Marco Brambilla

marco.brambilla@polimi.it

@marcobrambi

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 (and time consistency is granted).

Consistency
The data agrees with itself.

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.

Incomplete

What about interpretability, accessibility, metadata, analysis, etc.

Vague

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

Isn’t data science sexy?

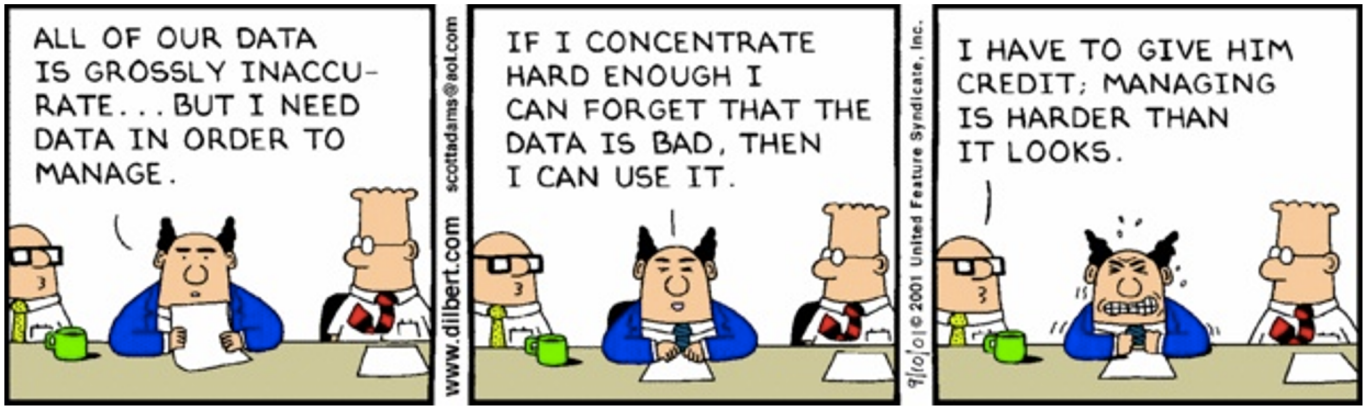
1. When Data Is Wrong

Marco Brambilla.

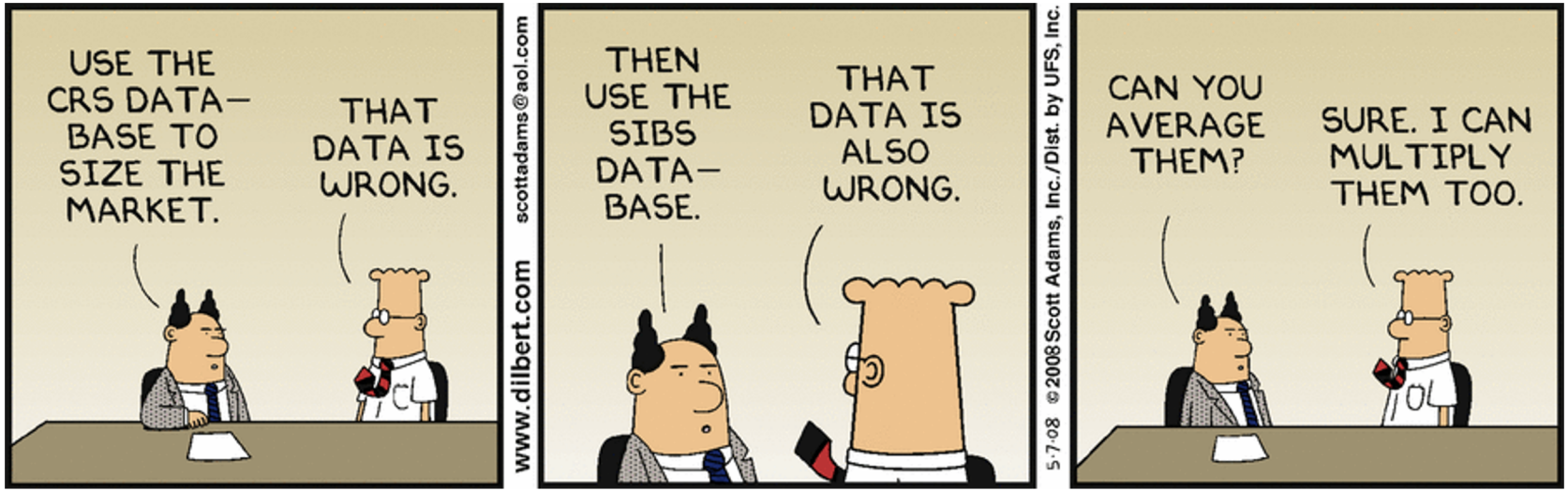
POLITECNICO
MILANO 1863

The skeptic approach

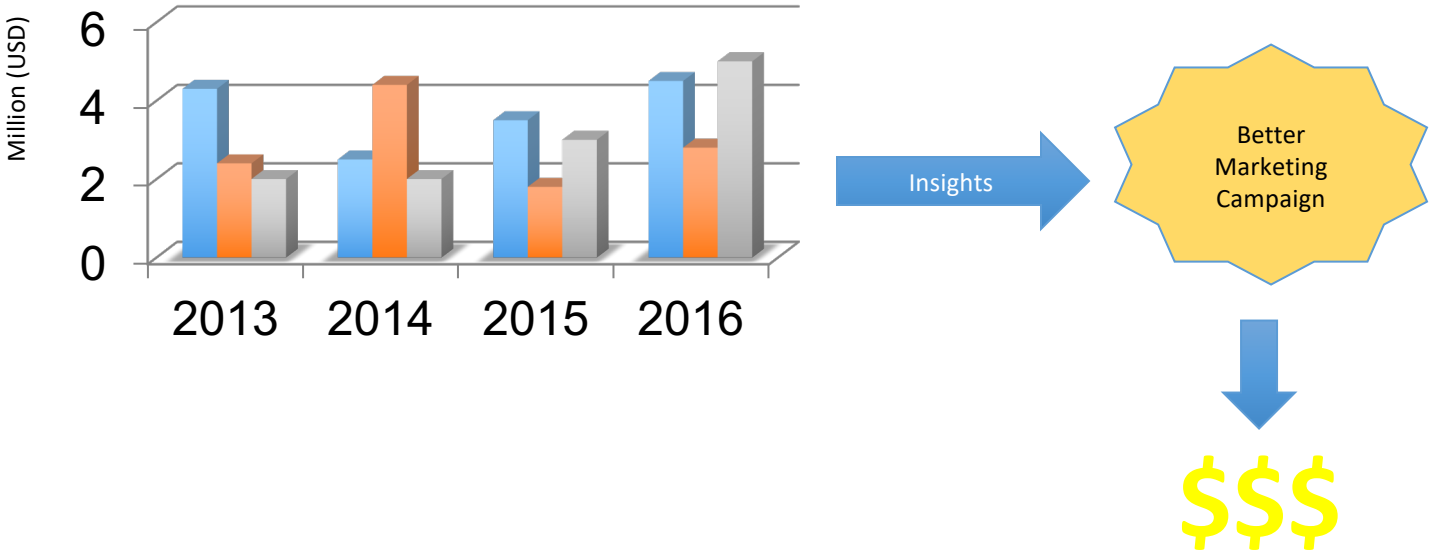
The pragmatic approach



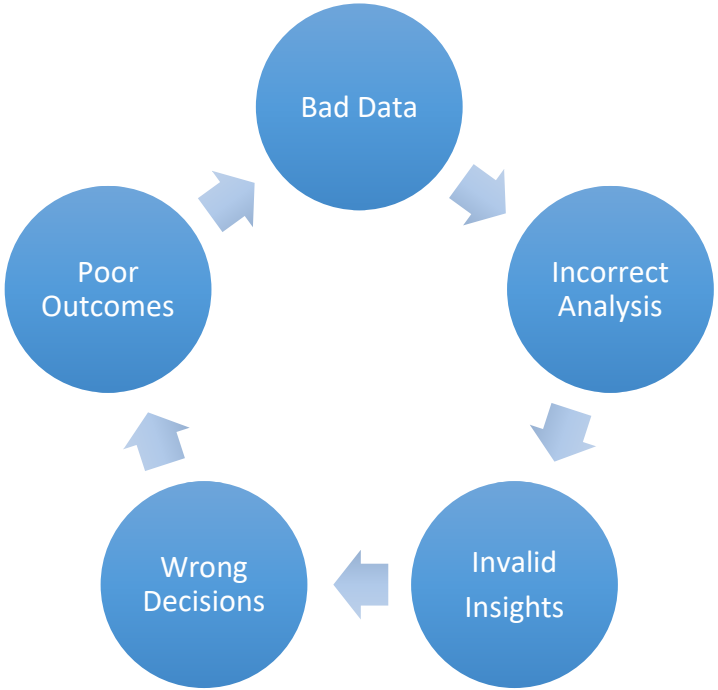
The (pseudo) practioner approach

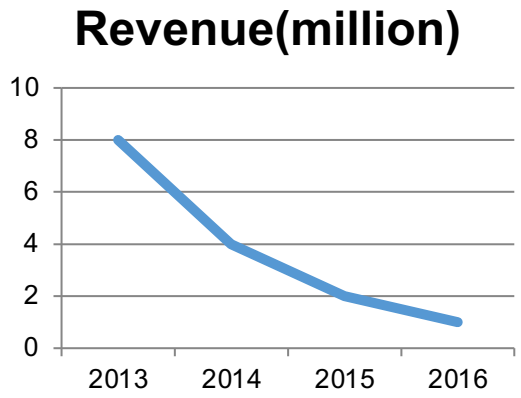


Goal: Better Faster Cheaper!



The Vicious Cycle of Bad Data





Data Quality is an issue…

Data Quality Issue

Gartner Report

By 2017, 33% of the largest global companies will experience an information crisis due to their inability to adequately value, govern and trust their enterprise information.

> *If you torture the data long enough,*
> *it will confess to anything*
>
> *– Darrell Huff*

2. Making a Wrong Right

Marco Brambilla.

POLITECNICO
MILANO 1863

Data Wrangling is …

The process of transforming "raw" data into data that can be analyzed to generate valid actionable insights

Data scientists spend more time
on preparing data than on analyzing it

Data Wrangling a.k.a.

Data Preprocessing
Data Preparation
Data Cleansing
Data Scrubbing
Data Munging
Data Transformation
Data Fold, Spindle, Mutilate...
(good old) ETL

ETL (Extract-Transform-Load)

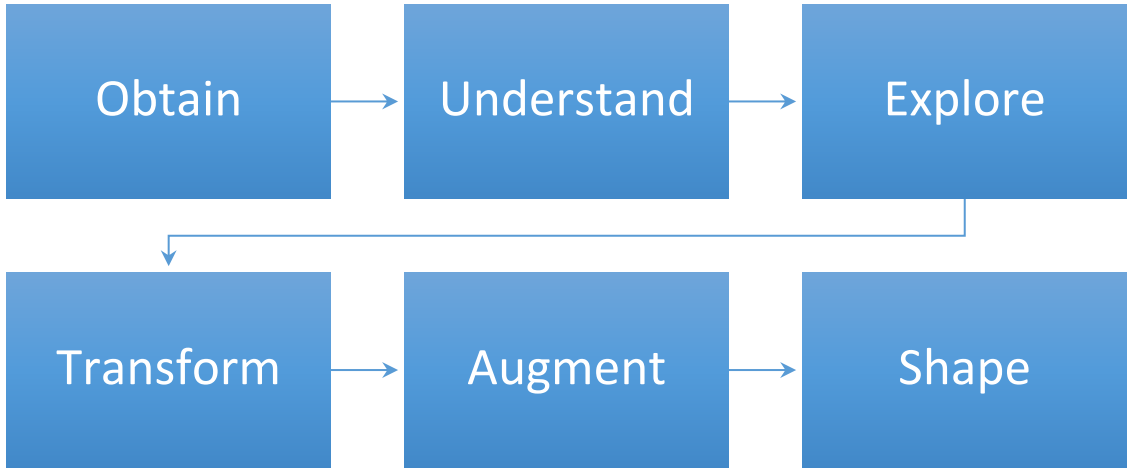
Information architecture pattern for datawarehouse solutions

Batch oriented data integration and delivery

Data Wrangling Steps

Iterative process

- Understand
- Explore
- Transform
- Augment
- Visualize



What is Data Cleansing?

Data cleansing or **data scrubbing** is the act of **detecting and correcting (or removing) corrupt or inaccurate records** from a data set.

The term refers to identifying incomplete, incorrect, inaccurate, partial or irrelevant parts of the data and then replacing, modifying, filling in or deleting this dirty data.

Why is Data “Dirty” ?

Dummy Values,
Absence of Data,
Multipurpose Fields,
Cryptic Data,
Contradicting Data,
Shared Field Usage,
Inappropriate Use of Fields,
Violation of Business Rules,
Reused Primary Keys,
Non-Unique Identifiers,
Data Integration Problems

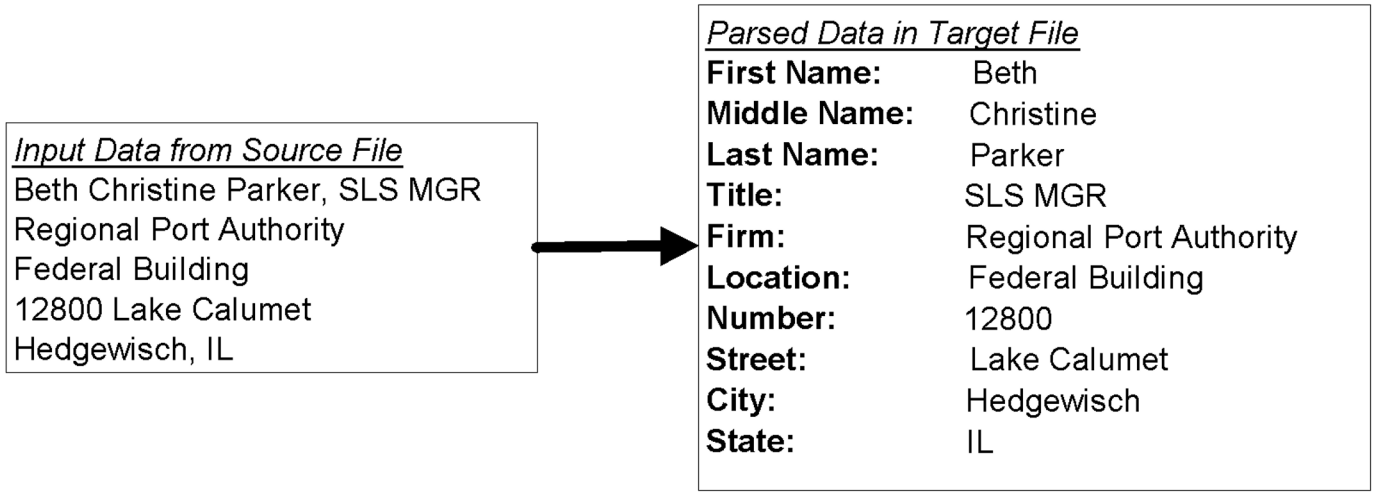
Data Cleansing in Practice

- Parsing
- Correcting
- Standardizing
- Matching
- Consolidating

Parsing

Parsing locates and identifies individual data elements in the source files and then isolates these data elements in the target files.

Parsing


Input Data from Source File
Beth Christine Parker, SLS MGR
Regional Port Authority
Federal Building
12800 Lake Calumet
Hedgewisch, IL
Parsed Data in Target File
First Name: Beth
Middle Name: Christine
Last Name: Parker
Title: SLS MGR
Firm: Regional Port Authority
Location: Federal Building
Number: 12800
Street: Lake Calumet
City: Hedgewisch
State: IL

Correcting

Corrects parsed individual data components using sophisticated data algorithms and secondary data sources.

Correcting

Parsed Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	Lake Calumet
City:	Hedgewisch
State:	IL

Corrected Data

First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Standardizing

Standardizing applies conversion routines to **transform data into its preferred (and consistent) format** using both standard and custom business rules, as well as coherent measurement units, ...

Standardizing

Corrected Data	
First Name:	Beth
Middle Name:	Christine
Last Name:	Parker
Title:	SLS MGR
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	South Butler Drive
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

Corrected Data	
Pre-name:	Ms.
First Name:	Beth
1st Name Match Standards:	Elizabeth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr.
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

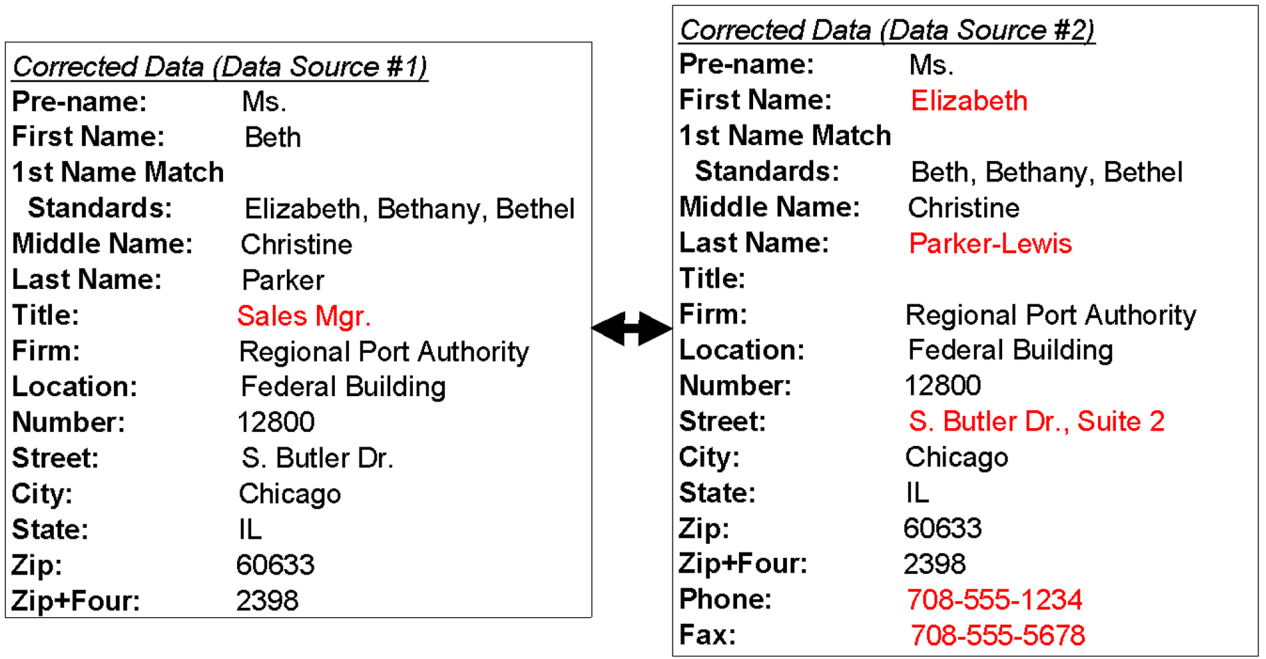
Matching

Searching and **matching records** within and across the parsed, corrected and standardized data based on predefined business rules to **eliminate duplications**.

Match Patterns

Business Name	Street	Branch Type	Customer #/Tax ID	City	Vendor Code	Pattern	Pattern I.D.
Exact	Exact	Exact	Exact	Exact	Exact	AAAAAA	P110
Exact	VClose	Exact	VClose	Exact	Blanks	ABAAA-	P115
Exact	VClose	Exact	Blanks	Exact	Exact	ABA-AA	P120
Exact	VClose	Close	Close	Exact	Exact	ABCCAA	S300
VClose	VClose	Exact	Close	Exact	Exact	BBACAA	S310

Matching



Corrected Data (Data Source #1)

Pre-name:	Ms.
First Name:	Beth
1st Name Match Standards:	Elizabeth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker
Title:	Sales Mgr.
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr.
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398

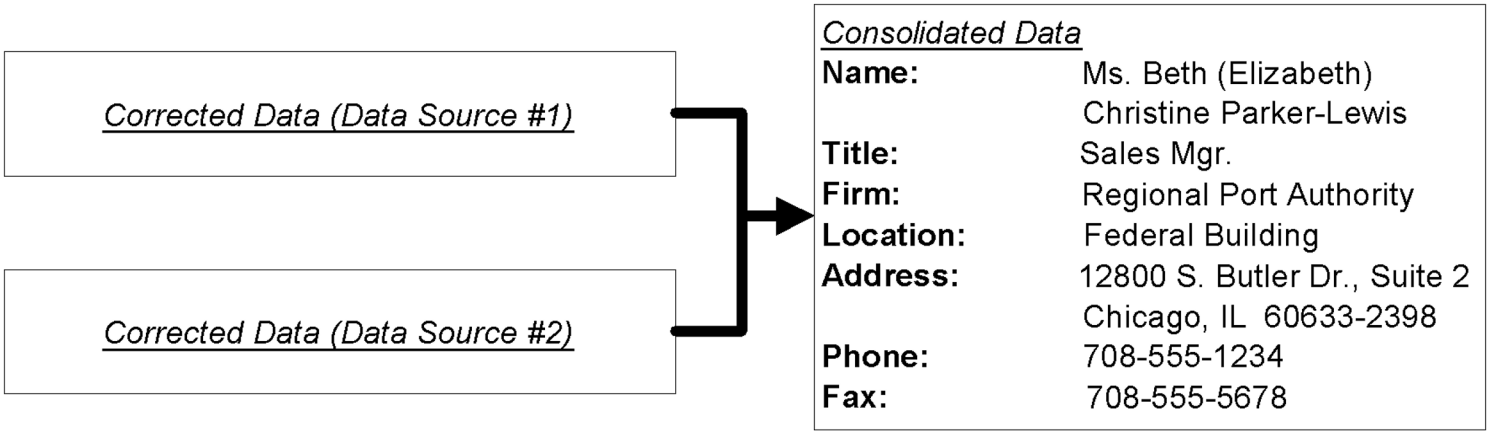
Corrected Data (Data Source #2)

Pre-name:	Ms.
First Name:	Elizabeth
1st Name Match Standards:	Beth, Bethany, Bethel
Middle Name:	Christine
Last Name:	Parker-Lewis
Title:	
Firm:	Regional Port Authority
Location:	Federal Building
Number:	12800
Street:	S. Butler Dr., Suite 2
City:	Chicago
State:	IL
Zip:	60633
Zip+Four:	2398
Phone:	708-555-1234
Fax:	708-555-5678

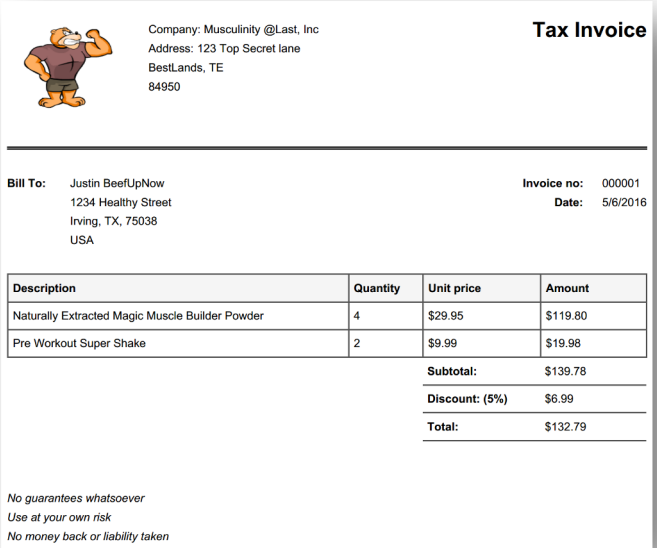
Consolidating

Analyzing and **identifying relationships** between matched records and consolidating/merging them into ONE representation.

Consolidating



Understanding Data: PDF



Company: Musculinity @Last, Inc
Address: 123 Top Secret lane
BestLands, TE
84950

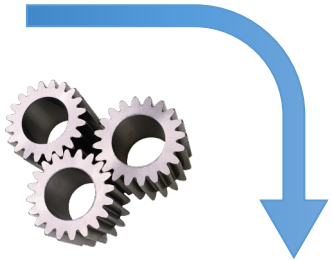
Tax Invoice

Bill To: Justin BeefUpNow
1234 Healthy Street
Irving, TX, 75038
USA

Invoice no: 000001
Date: 5/6/2016

Description	Quantity	Unit price	Amount
Naturally Extracted Magic Muscle Builder Powder	4	$29.95	$119.80
Pre Workout Super Shake	2	$9.99	$19.98
		Subtotal:	$139.78
		Discount: (5%)	$6.99
		Total:	$132.79

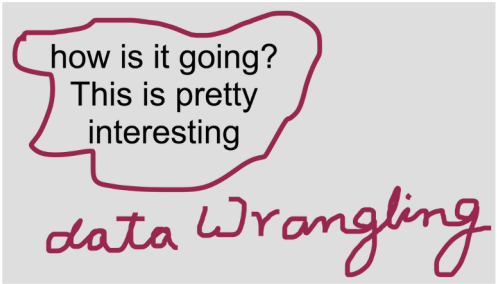
No guarantees whatsoever
Use at your own risk
No money back or liability taken



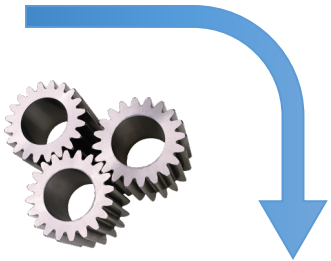
Company	Address	Bill To	Customer Address	Invoice	Date	Subtotal	Discount	Total	Description	Quantity	Unit	Amount
Musculinity @Last, Inc	123 Top Secret lane|BestLands, TE|84950	Justin BeefUpNow	1234 Healthy Street|Irving, TX, 75038|USA	1	5/6/2016	139.78	6.99	132.79	Naturally Extracted Magic Muscle Builder Powder	4	29.95	119.8
Musculinity @Last, Inc	123 Top Secret lane|BestLands, TE|84950	Justin BeefUpNow	1234 Healthy Street|Irving, TX, 75038|USA	1	5/6/2016	139.78	6.99	132.79	Pre Workout Super Shake	2	9.99	19.98

POLITECNICO
MILANO 1863

Understanding Data: Free Text



Python + Textract +Tesseract



```
>>> text = textract.process('/Users/akuntamu/Downloads/Drawing3.png',method='tesseract-ocr')
>>> text
'     \n   \n\nhow is it going?\nThis is pretty\ninteresting\n\n'
>>>
```

Understanding Data: more

"Looks like my V8 Chevy is running low on fuel.
Didn't I fill up just the day before?"

UNSTRUCTURED

STRUCTURED

Owner	Vehicle	Type	Fuel Level	Engine	Last Fill
AK	Chevy	Gas	5%	V8	05/04/16

Understanding Data: more

Decode the following secret message:

DALDFWSFOEWRBOSDCALAXORDJFKMCO

↓

DAL DFW SFO EWR BOS

DCA LAX ORD JFK MCO

Data Munging

Potentially lossy transformations applied to a piece of data or a file

Vague data transformation steps that are not yet completely clear

E.g., removing punctuation or html tags, data parsing, filtering, and transformation.

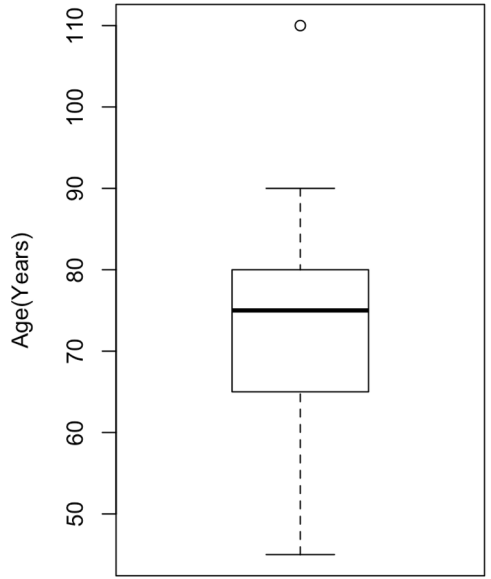
Semantics

???
75
80
65
55
67
78
88
90
45
58
69
80
110

?

Semantics and Outliers

Age(Years)
75
80
65
55
67
78
88
90
45
58
69
80
110



The value stands in the abnormal

Missing Data: Detection

Overtly missing data

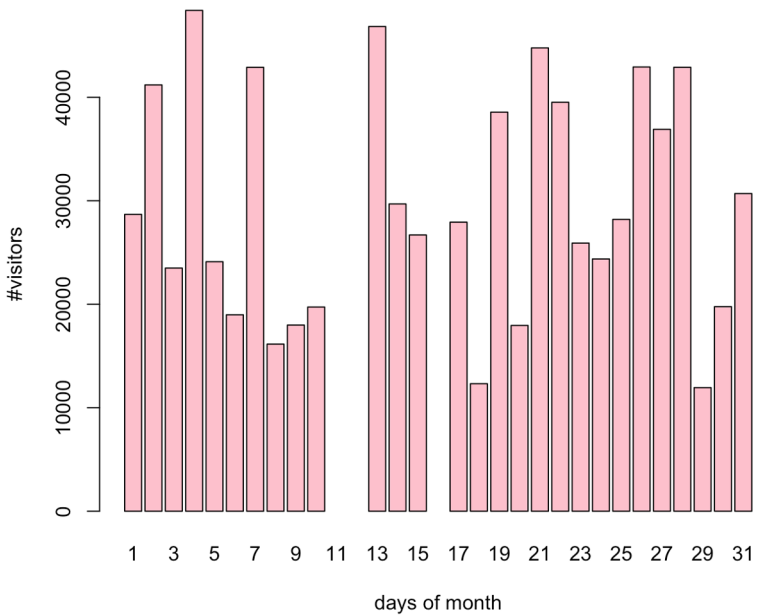
- Match data specifications against data - are all the attributes present?
- Scan individual records - are there gaps?
- Rough checks : number of files, file sizes, number of records, number of duplicates
- Compare estimates (averages, frequencies, medians) with “expected” values and bounds; check at various levels of granularity since aggregates can be misleading

Missing data: Detection (cont.)

Hidden damage to data

- Values are truncated or censored – check for spikes and dips in distributions and histograms
- Missing values and defaults are indistinguishable – too many missing values? metadata or domain expertise can help
- Errors of omission e.g. all calls from a particular area are missing – check if data are missing randomly or are localized in some way

Missing Values: Random



System failures

Complete miss

Missing Values: Wrong Ingestion

CSV to table / excel

Merged fields

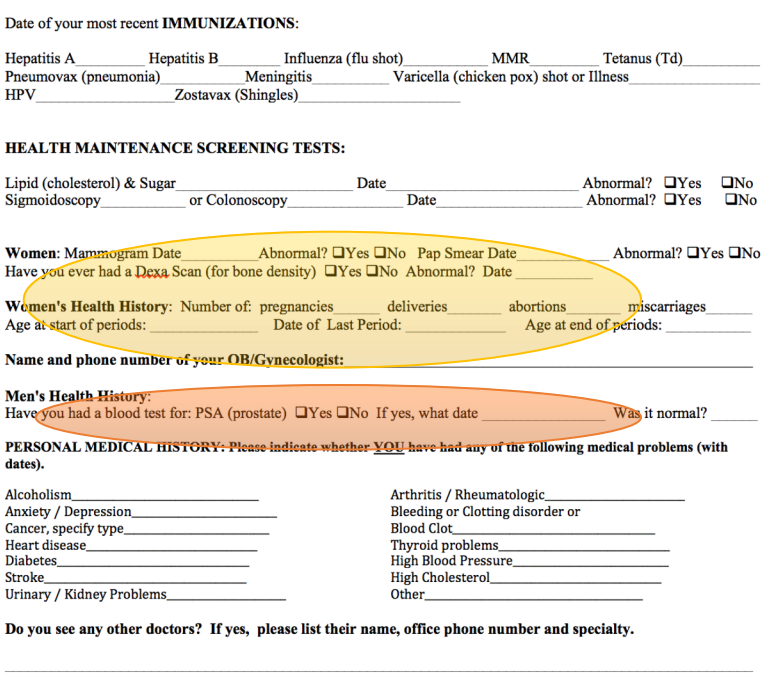
Missing fields

```
Antony Mc James IV, 123 Untidy Cir, Suite# 234, Cumbersome, TX, 76849
Miller Johnson, 1102 Messy Data St, Middletow, OH
Betty Flyier 6483 Phew Lane, Apt A4, FixMe, CA, 91103
```

Missing due to invalid data and ingestion

Customer	Address Line 1	Address Line 2	City	State	Zip
Antony Mc James IV	123 Untidy Cir	Suite# 234	Cumbersome	TX	76849
Miller Johnson	1102 Messy Data St	Middletow	OH		
Betty Flyier 6483 Phew Lane	Apt A4	FixMe	CA	91103	

Missing Values: Inapplicability



Date of your most recent **IMMUNIZATIONS**:

Hepatitis A_____ Hepatitis B_____ Influenza (flu shot)_____ MMR_____ Tetanus (Td)_____
Pneumovax (pneumonia)_____ Meningitis_____ Varicella (chicken pox) shot or Illness_____
HPV_____ Zostavax (Shingles)_____

HEALTH MAINTENANCE SCREENING TESTS:

Lipid (cholesterol) & Sugar_____ Date_____ Abnormal? ❑Yes ❑No
Sigmoidoscopy_____ or Colonoscopy_____ Date_____ Abnormal? ❑Yes ❑No

Women: Mammogram Date_____ Abnormal? ❑Yes ❑No Pap Smear Date_____ Abnormal? ❑Yes ❑No
Have you ever had a Dexa Scan (for bone density) ❑Yes ❑No Abnormal? Date _____

Women's Health History: Number of: pregnancies_____ deliveries_____ abortions_____ miscarriages_____
Age at start of periods: _____ Date of Last Period: _____ Age at end of periods: _____

Name and phone number of your OB/Gynecologist: _____

Men's Health History:
Have you had a blood test for: PSA (prostate) ❑Yes ❑No If yes, what date _____ Was it normal? _____

PERSONAL MEDICAL HISTORY: Please indicate whether YOU have had any of the following medical problems (with dates).

Alcoholism_____
Anxiety / Depression_____
Cancer, specify type_____
Heart disease_____
Diabetes_____
Stroke_____
Urinary / Kidney Problems_____
Arthritis / Rheumatologic_____
Bleeding or Clotting disorder or
Blood Clot_____
Thyroid problems_____
High Blood Pressure_____
High Cholesterol_____
Other_____

Do you see any other doctors? If yes, please list their name, office phone number and specialty.

_____

Partial data by nature

Remember to leave empty slots

Imputing Values to Missing Data

In federated data, between 30%-70% of the data points will have at least one missing attribute - data wastage if we ignore all records with a missing value

Remaining data is seriously biased

Lack of confidence in results

Understanding pattern of missing data unearths data integrity issues

Missing Value Imputation - 1

Standalone imputation

- Mean, median, other point estimates
- Assume: Distribution of the missing values is the same as the non-missing values.
- Does not take into account inter-relationships
- Introduces bias
- Convenient, easy to implement

Missing Value Imputation - 2

Better imputation - use attribute relationships

Assume : all prior attributes are populated

That is, *monotonicity* in missing values.

X1	X2	X3	X4	X5
1.0	20	3.5	4	.
1.1	18	4.0	2	.
1.9	22	2.2	.	.
0.9	15	.	.	.

Two techniques

Regression (parametric),

Propensity score (nonparametric)

Missing Value Imputation –3

Regression method

Use linear regression, sweep left-to-right

X3=a+b*X2+c*X1;
X4=d+e*X3+f*X2+g*X1, and so on

X3 in the second equation is estimated from the first equation if it is missing

Missing Value Imputation - 3

Propensity Scores (nonparametric)

- Let Y_j=1 if X_j is missing, 0 otherwise
- Estimate $P(Y_j = 1)$ based on X_1 through $X_{(j-1)}$ using logistic regression
- Group by propensity score $P(Y_j = 1)$
- Within each group, estimate missing X_js from known X_js using approximate Bayesian bootstrap.
- Repeat until all attributes are populated.

Missing Value Imputation - 4

Arbitrary missing pattern

- Markov Chain Monte Carlo (MCMC)
- Assume data is multivariate Normal, with parameter θ
- (1) Simulate missing X, given θ estimated from observed X ; (2) Re-compute θ using filled in X
- Repeat until stable.
- Expensive: Used most often to induce monotonicity

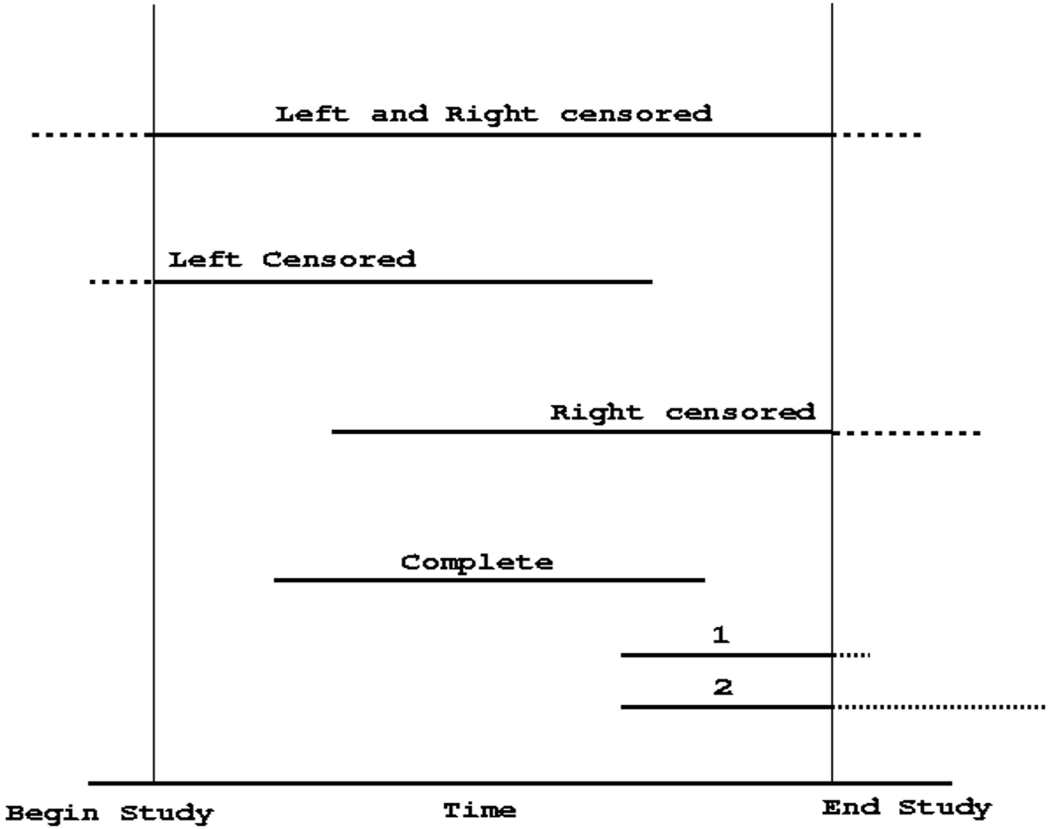
Note that imputed values are useful in aggregates but can't be trusted individually

Censoring and Truncation

Well studied in Biostatistics, relevant to time dependent data e.g. duration

Censored - Measurement is bounded but not precise e.g. Call duration > 20 are recorded as 20

Truncated - Data point dropped if it exceeds or falls below a certain bound e.g. customers with less than 2 minutes of calling per month



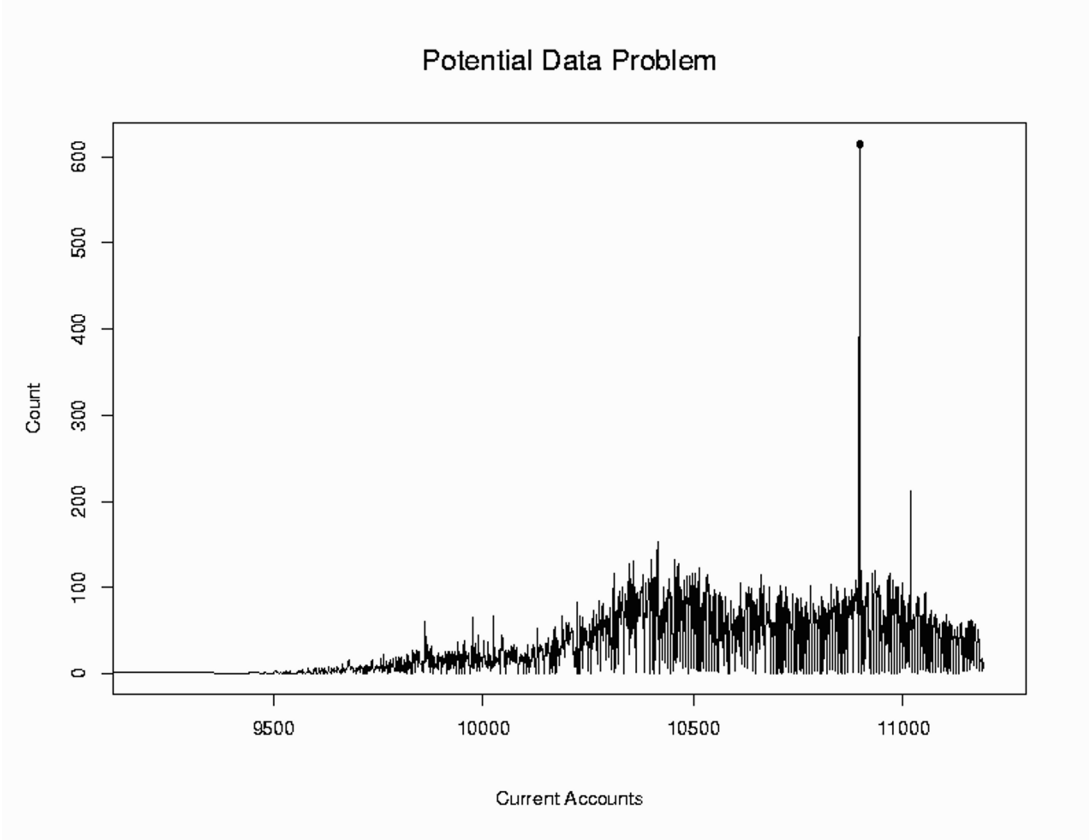
Censored time intervals

Censoring/Truncation (cont.)

If censoring/truncation mechanism not known, analysis can be inaccurate and biased.

But if you know the mechanism, you can mitigate the bias from the analysis.

Metadata should record the existence as well as the nature of censoring/truncation



Spikes usually indicate censored time intervals caused by resetting of timestamps to defaults

Suspicious Data

Consider the data points

3, 4, 7, 4, 8, 3, 9, 5, 7, 6, 92

"92" is suspicious - an *outlier*

Outliers are potentially legitimate

Often, they are data or model glitches

Or, they could be a data miner's dream, e.g. highly profitable customers

Outliers

Outlier – “departure from the expected”

Types of outliers – defining “expected”

Many approaches

- Error bounds, tolerance limits – control charts
- Model based – regression depth, analysis of residuals
- Geometric
- Distributional
- Time Series outliers

Control Charts

Quality control of production lots

Typically univariate: X-Bar, R, CUSUM

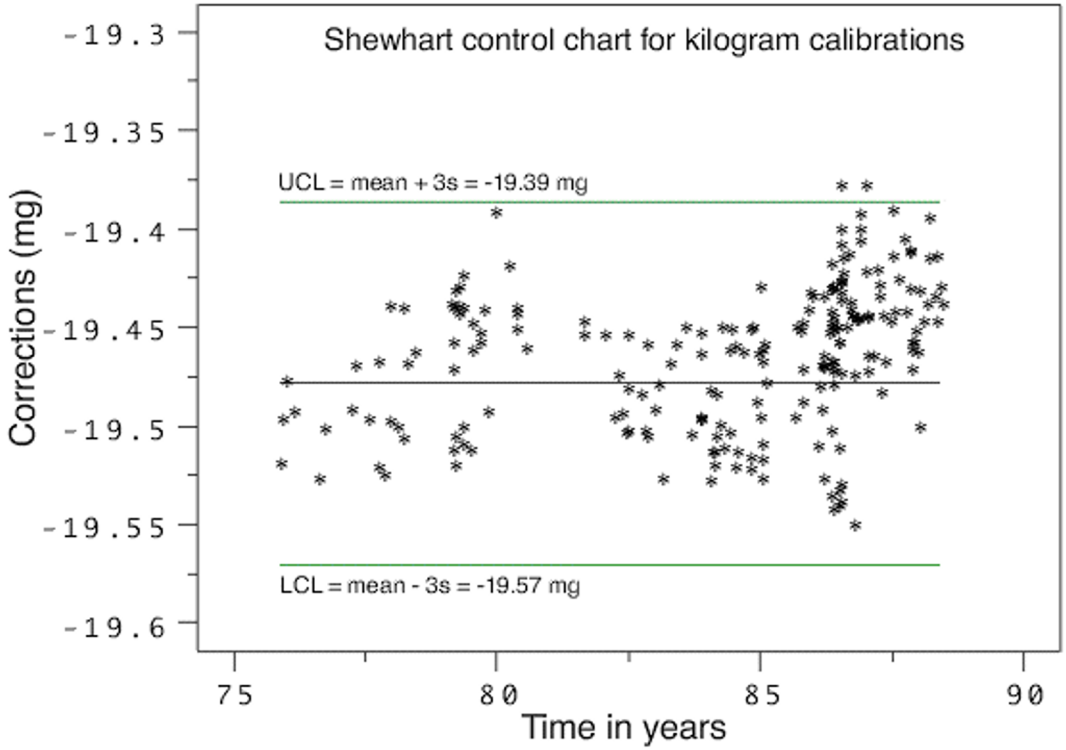
Distributional assumptions for charts not based on means e.g. R–charts

Main steps (based on statistical inference)

- Define “expected” and “departure” e.g. Mean and standard error based on sampling distribution of sample mean (aggregate);
- Compute aggregate each sample
- Plot aggregates vs expected and error bounds
- “Out of Control” if aggregates fall outside bounds

An Example

(http://www.itl.nist.gov/div898/handbook/mpc/section3/mpc3521.htm)



Multivariate Control Charts - 1

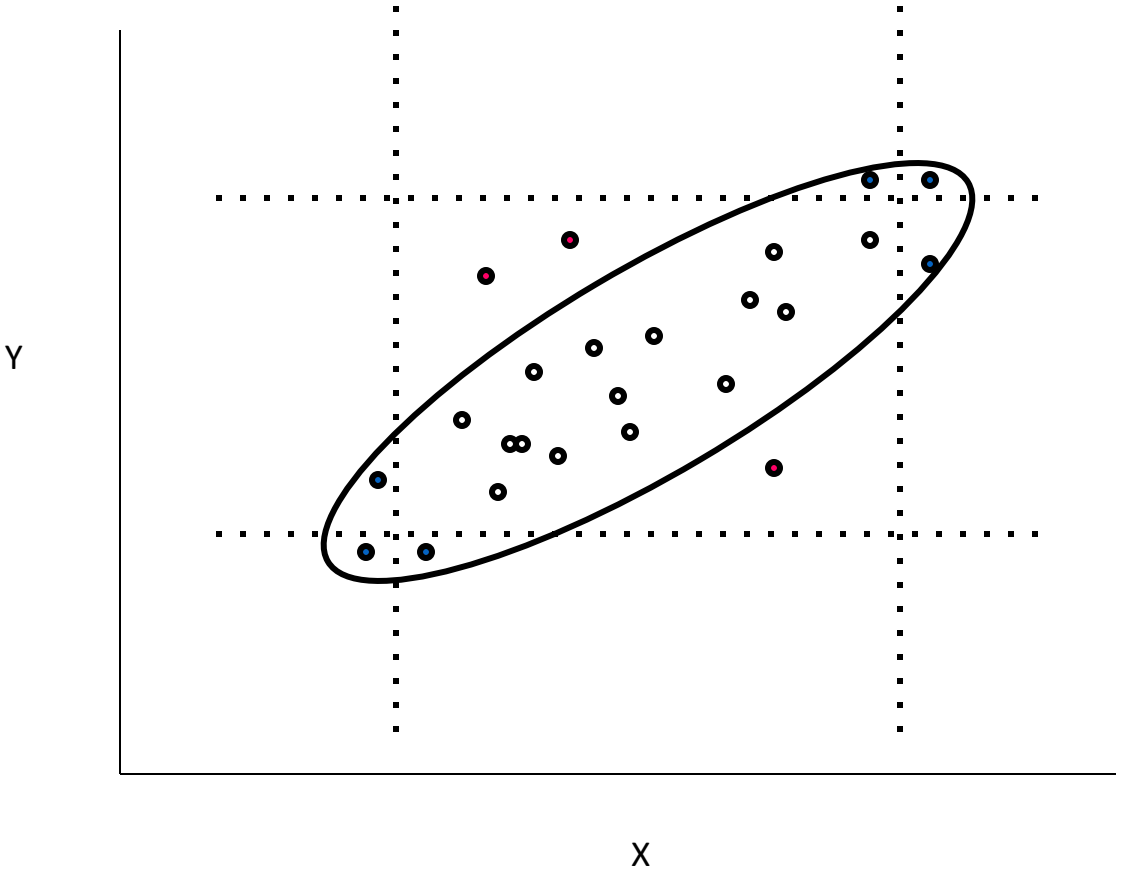
Bivariate charts:

- based on bivariate Normal assumptions
- component-wise limits lead to Type I, II errors

Depth based control charts (nonparametric):

- map n-dimensional data to one dimension using depth e.g. Mahalanobis
- Build control charts for depth
- Compare against benchmark using depth e.g. Q-Q plots of depth of each data set

Bivariate Control Chart



Multivariate Control Charts - 2

Multiscale process control with wavelets:

- Detects abnormalities at multiple scales as large wavelet coefficients.
- Useful for data with heteroscedasticity
- Applied in chemical process control

Model Fitting and Outliers

Models summarize general trends in data
- more complex than simple aggregates
- e.g. linear regression, logistic regression focus on attribute relationships

Data points that do not conform to well fitting models are *potential outliers*

Goodness of fit tests (DQ for analysis/mining)
- check suitableness of model to data
- verify validity of assumptions
- data rich enough to answer analysis/business question?

Set Comparison and Outlier Detection

“Model” consists of partition based summaries

Perform nonparametric statistical tests for a rapid section-wise comparison of two or more massive data sets

If there exists a baseline “good’’ data set, this technique can detect potentially corrupt sections in the test data set

Types of data

Categorical

Qualitative
- Subjective

Quantitative
- Discrete
- Continuous

Types of data

Categorical

Qualitative
- Subjective

Quantitative
- Discrete
- Continuous

Color

Nice, Good, For birthday

6 balloons

Pressure 15 PSI,

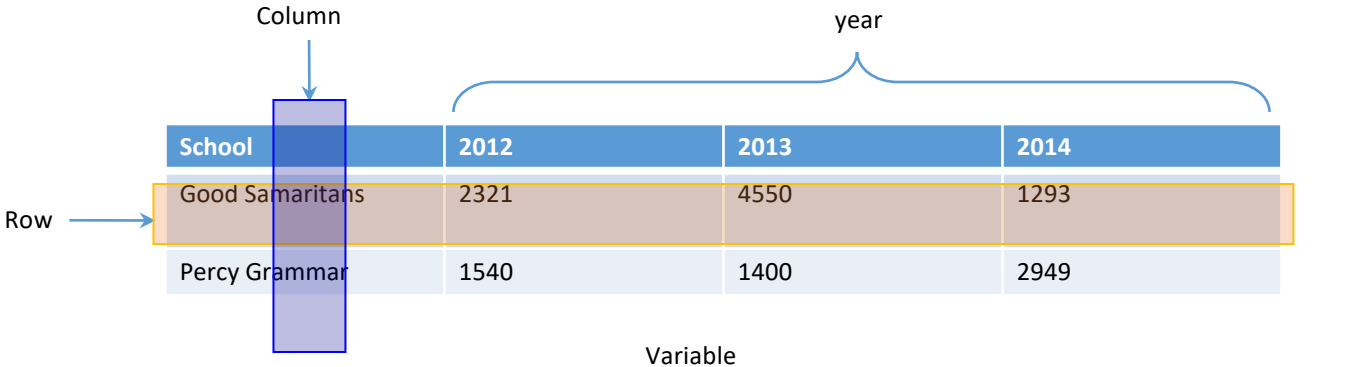
139 m. over sea level

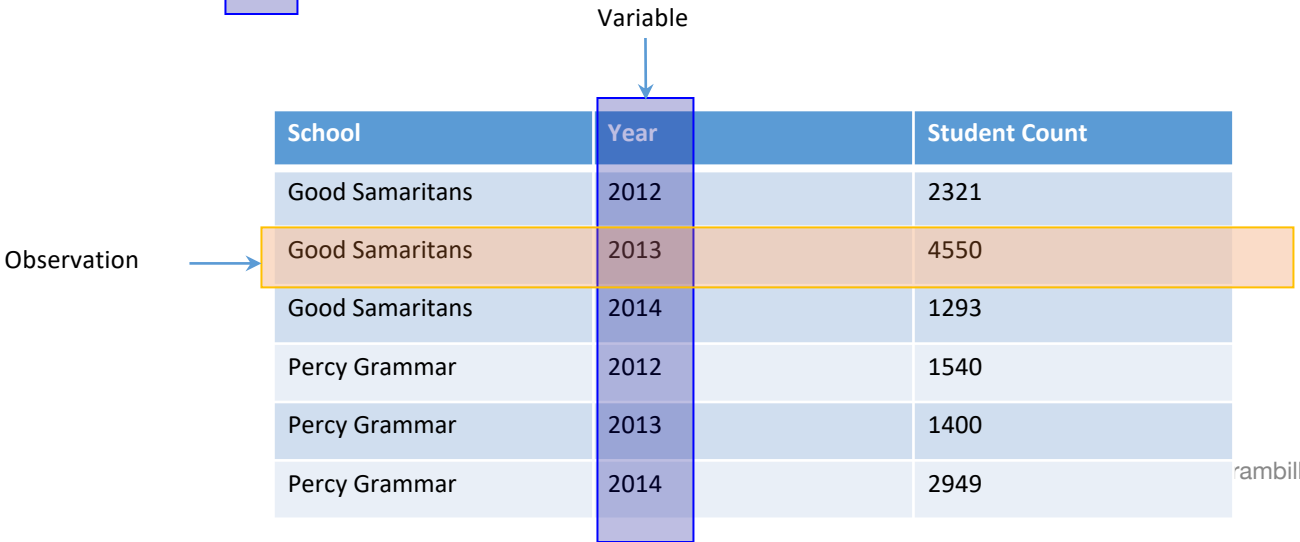
Data Source Selection Criteria

Credibility
Completeness
Accurateness
Verifiability
Currency
Accessibility

Compliance
Cost
Legal issues
Security
Storage
Provenance

Not all tables are created equal





Not all tables are created equal

Year	Comedy-Q1	Thriller-Q1	Action-Q1	…
2014	2	1	0	
2015	0	3	2	

Find total comedy movies in all of 2014? -> Not easy in current form

Category	Quarter	Year	#Hits
Comedy	Q1	2014	2
Thriller	Q1	2014	1
Action	Q1	2014	0
Comedy	Q1	2015	0
Thriller	Q1	2015	3
Action	Q1	2015	2

Find % of hit comedy movies in a 2015?

Very easy to add a new column

Not all tables are created equal

Category	Rating	Q1	Q2	Q3	…
Comedy	Excellent	1	0	1	
Comedy	Good	2	0	2	
Thriller	Excellent	0	1	1	
Thriller	Good	1	0	3	

Very messy data
Variables in both rows and columns

Category	Quarter	Excellent	Good
Comedy	Q1	1	2
Comedy	Q2	0	0
Comedy	Q3	1	2
Thriller	Q1	0	1
Thriller	Q2	1	0
Thriller	Q3	1	3

Each row is complete observation

Not all tables are created equal

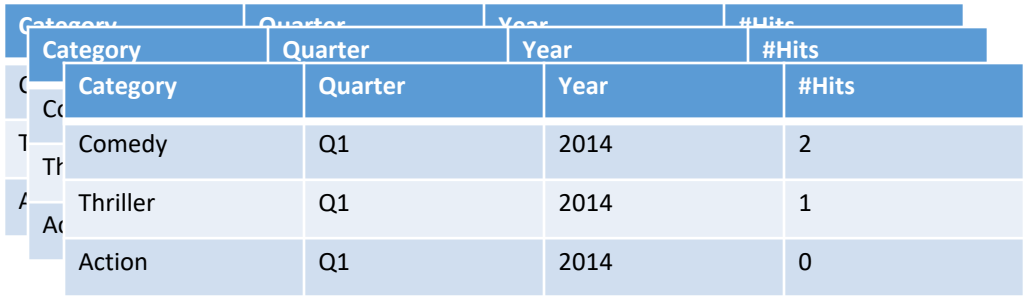
Invoice	Bill To	Sales %	Total($)	SKU#	Item	Qty	Unit Price ($)
1	Jim Jones	8	8.03	A123	Hammer	1	3.55
1	Jim Jones	8	8.03	Q34	Screw Driver	2	2.05
2	Mike Z'Kale	8	97.20	W23	Hair Dryer	1	59.25
2	Mike Z'Kale	8	97.20	E452	Cologne	3	10.25

Normalize to avoid duplication

Invoice	Bill To	Sales %	Total($)
1	Jim Jones	8	8.03
2	Mike Z'Kale	8	97.20

Invoice	SKU#	Item	Qty	Unit Price ($)
1	A123	Hammer	1	3.55
1	Q34	Screw Driver	2	2.05
2	W23	Hair Dryer	1	59.25
2	E452	Cologne	3	10.25

Not all tables are created equal



Category	Quarter	Year	#Hits
Comedy	Q1	2014	2
Thriller	Q1	2014	1
Action	Q1	2014	0

Multiple Tables
Divided by Time

Category	Quarter	Year	#Hits
Comedy	Q1	2014	2
Thriller	Q1	2014	1
Action	Q1	2014	0
Comedy	Q1	2014	2
Thriller	Q1	2014	1
Action	Q1	2014	0

Combine all tables accommodating varying formats

The “Key” (matching) problem

Keys are crucial in DB

Many DBs --> Many keys

How to align?

Identification to certain degree of accuracy likely-identities

e.g., same user match

The “Duplicates” problem

Related to Key problem

Identification to certain degree of accuracy likely-duplicates

e.g., duplicate posts

The “Duplicates” problem

Related to Key problem

Identification to certain degree of accuracy likely-duplicates

Lessons Learnt on Tables

(Multiple) variables in columns

Never values as columns!

Shape may depend on convenience of queries

Matching identities and duplications are crucial in data science!

Lessons Learnt on Tables

Each observation is complete and atomic

Each variable belongs to (only!) one column!

Schema-On-Write Vs Schema-On-Read

Traditional DBMSs enforced writing only data consistent with a pre-designed schema

Today data modeling at design time is a luxury

In schema on read, data is applied to a plan or schema as it is pulled out of a stored location, rather than as it goes in

Popular Open Source Tools

DATASCIENCETOOLKIT

Other Resources

BeautifulSoup

Tesseract

SCHOOL OF DATA

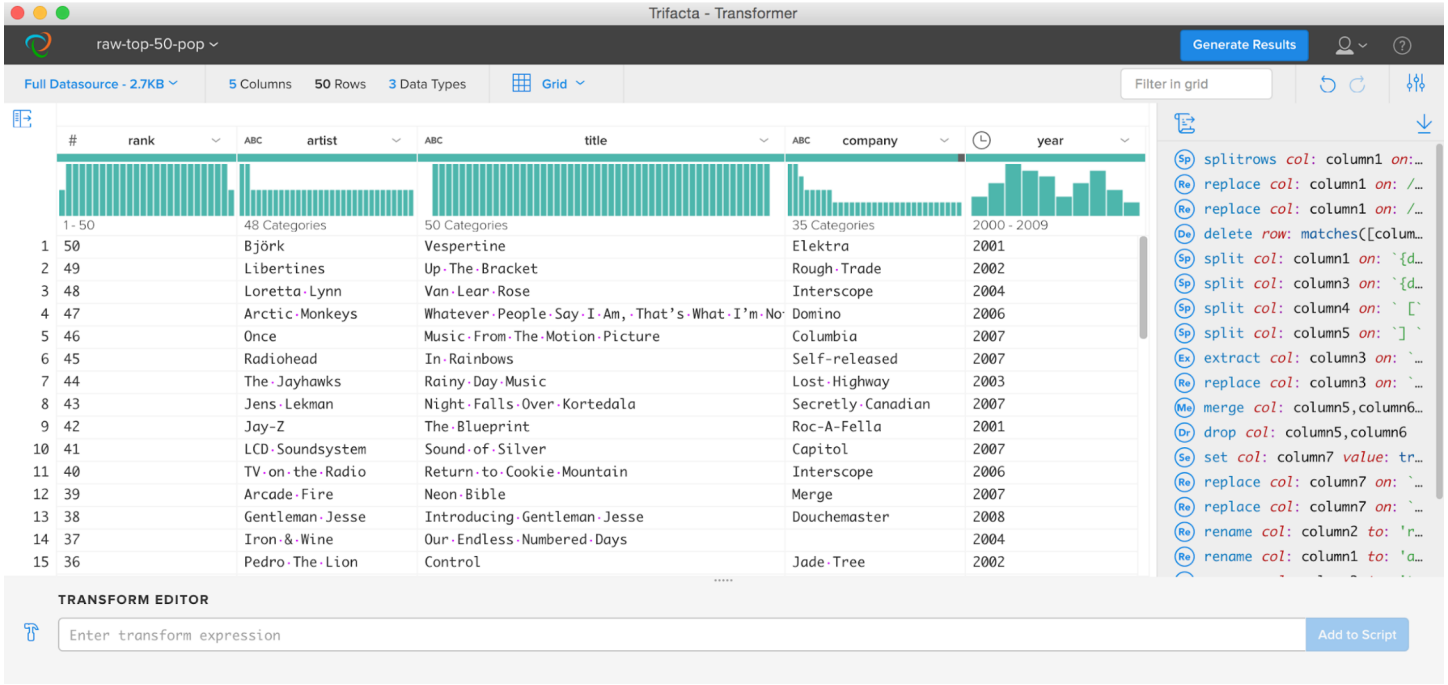
textract

OPEN KNOWLEDGE LABS

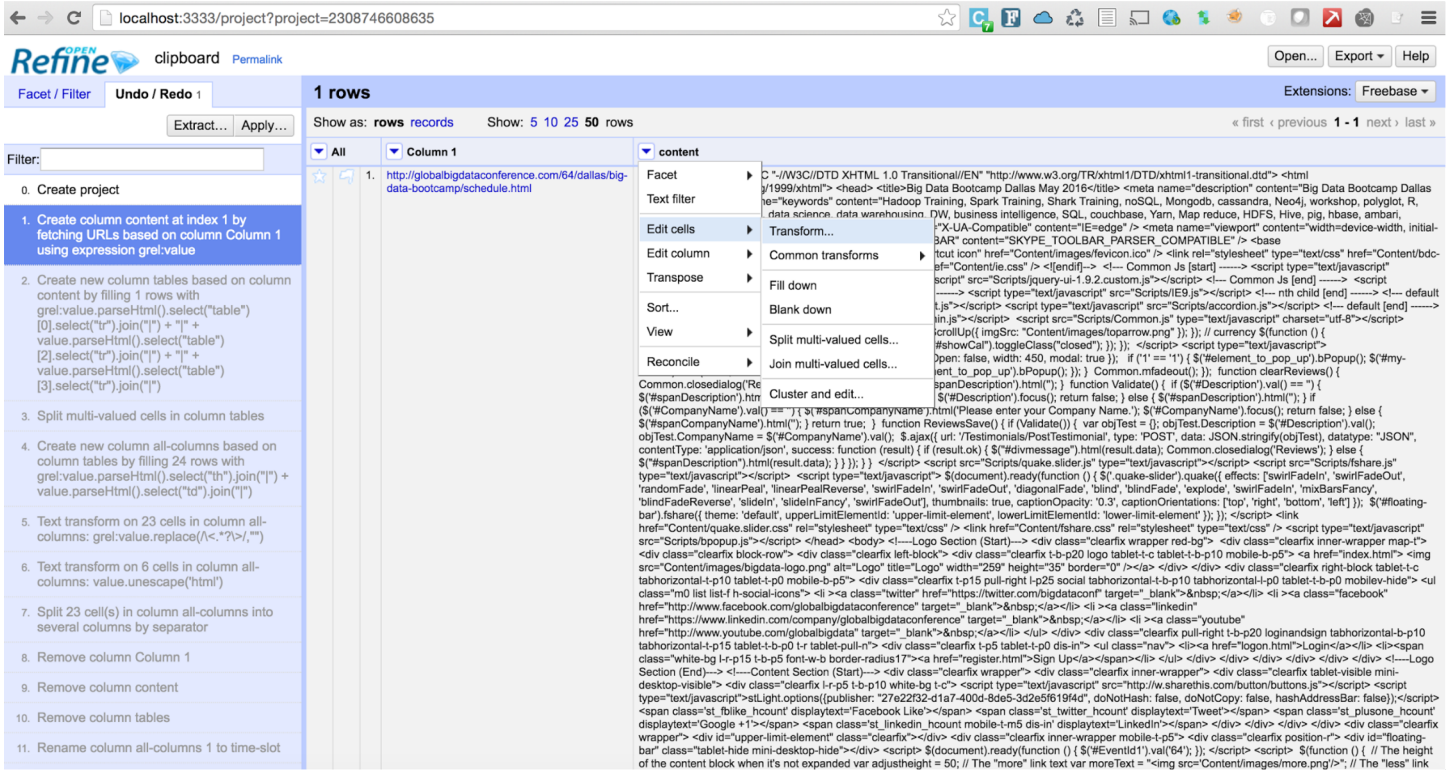
OCRFeeder

Commercial Vendors

Trifacta Wrangler



Google’s Open Refine



Hands on Data Wrangling

Data Ingestion

- CSV
- PDF
- API/JSON
- HTML Web Scraping
- XLS, Access, ...!

Data Exploration

- Visual inspection
- Graphing

Data Shaping

- Tidying Data

Data Cleansing

- Missing values
- Format
- Measurement Units
- Outliers
- Data Errors Per Domain
- Fat Fingered Data

Data Augmenting

- Aggregate data sources
- Fuzzy/Exact match

R Libraries for Data Wrangling

stringr
dplyr
tidyr
readxl, xlsx
lubridate
gtools
plyr
rvest

References – Web and Books

Web:
www2.gbif.org/DataCleaning.pdf
www.webopedia.com/TERM/D/data_cleansing.html

Books:
Data Mining by Ian H. Witten and Eibe Frank
Exploratory Data Mining and Data Quality by Dasu and Johnson (Wiley, 2004)

References – Tools

- Stanford Wrangler
  http://vis.stanford.edu/papers/wrangler
  http://vis.stanford.edu/wrangler
- http://openrefine.org/
- http://okfnlabs.org/
- http://schoolofdata.org/

POLITECNICO
MILANO 1863

SYSTEMS AND METHODS FOR BIG AND UNSTRUCTURED DATA

Wrangling

Marco Brambilla

marco.brambilla@polimi.it

@marcobrambi