Data Preparation for Data Science

Casey Stella
@casey_stella

2016

Table of Contents

Hi, I'm Casey Stella!

Garbage In $\Longrightarrow$ Garbage Out

"80% of the work in any data project is in cleaning the data."

— D.J. Patel in *Data Jujitsu*

Data Cleansing $\Longrightarrow$ Data Understanding

There are two ways to understand your data

- Syntactic Understanding
- Semantic Understanding

If you hope to get anything out of your data, you have to have a handle on both.

Syntactic Understanding: True Types

A **true type** is a label applied to data points x_i such that x_i are mutually comparable.

- Schemas type != true data type
- A specific column can have many different types

Syntactic Understanding: Density

Data **density** is an indication of how data is clumped together.

Syntactic Understanding: Density

Data **density** is an indication of how data is clumped together.

- For numerical data, distributions and statistical characteristics are informative
- For non-numeric data, counts and distinct counts of a canonical representation are extremely useful.

Syntactic Understanding: Density

Data **density** is an indication of how data is clumped together.

- For numerical data, distributions and statistical characteristics are informative
- For non-numeric data, counts and distinct counts of a canonical representation are extremely useful.

Canonical representations are representations which give you an idea at a glance of the data format

Syntactic Understanding: Density

Data **density** is an indication of how data is clumped together.

- For numerical data, distributions and statistical characteristics are informative
- For non-numeric data, counts and distinct counts of a canonical representation are extremely useful.

Canonical representations are representations which give you an idea at a glance of the data format

- Replacing digits with the character 'd'
- Stripping whitespace
- Normalizing punctuation

Syntactic Understanding: Density

Data **density** is an indication of how data is clumped together.

- For numerical data, distributions and statistical characteristics are informative
- For non-numeric data, counts and distinct counts of a canonical representation are extremely useful.

Canonical representations are representations which give you an idea at a glance of the data format

- Replacing digits with the character 'd'
- Stripping whitespace
- Normalizing punctuation

Data density is an assumption underlying any conclusions drawn from your data.

Syntactic Understanding: Density over Time

$\frac{\Delta Density}{\Delta t}$ is how data clumps change over time.

Syntactic Understanding: Density over Time

$\frac{\Delta Density}{\Delta t}$ is how data clumps change over time.
This kind of analysis can show

- Problems in the data pipeline

Syntactic Understanding: Density over Time

$\frac{\Delta Density}{\Delta t}$ is how data clumps change over time.
This kind of analysis can show

- Problems in the data pipeline
- Whether the assumptions of your analysis are violated

Syntactic Understanding: Density over Time

$\frac{\Delta Density}{\Delta t}$ is how data clumps change over time.
This kind of analysis can show

- Problems in the data pipeline
- Whether the assumptions of your analysis are violated

$\frac{\Delta Density}{\Delta t} \implies$

- Automation

Syntactic Understanding: Density over Time

$\frac{\Delta Density}{\Delta t}$ is how data clumps change over time.
This kind of analysis can show

- Problems in the data pipeline
- Whether the assumptions of your analysis are violated

$\frac{\Delta Density}{\Delta t} \implies$

- Automation
- Outlier Alerting

Semantic Understanding: "Do what I mean, not what I say"

Semantic understanding is understanding based on how the data is **used** rather than how it is stored.

Semantic Understanding: "Do what I mean, not what I say"

Semantic understanding is understanding based on how the data is **used** rather than how it is stored.

- Finding equivalences based on semantic understanding are often context sensitive.

Semantic Understanding: "Do what I mean, not what I say"

Semantic understanding is understanding based on how the data is **used** rather than how it is stored.

- Finding equivalences based on semantic understanding are often context sensitive.
- May come from humans (e.g. domain experience and ontologies)

Semantic Understanding: "Do what I mean, not what I say"

Semantic understanding is understanding based on how the data is **used** rather than how it is stored.

- Finding equivalences based on semantic understanding are often context sensitive.
- May come from humans (e.g. domain experience and ontologies)
- May come from machine learning (e.g. analyzing usage patterns to find synonyms)

Semantic Understanding: "Do what I mean, not what I say"

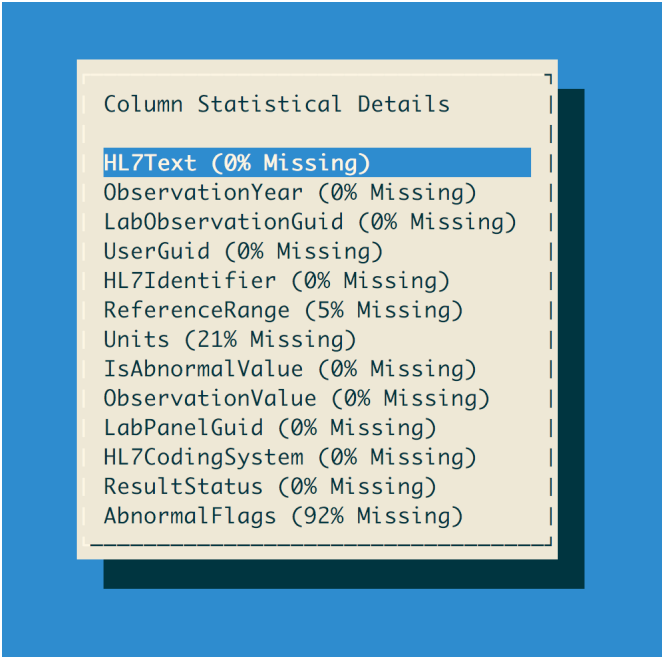
Semantic understanding is understanding based on how the data is **used** rather than how it is stored.

- Finding equivalences based on semantic understanding are often context sensitive.
- May come from humans (e.g. domain experience and ontologies)
- May come from machine learning (e.g. analyzing usage patterns to find synonyms)

Semantic understanding does not imply SkyNet

DEMO

```
usage: SummarizerCLI
 -D <property=value>                                Input properties
 -h,--help                                          This screen
 -i,--input <SOURCE>                                Input source
 -l,--load <JSON>                                   Load an existing
                                                    summary
 -m,--mode <MODE>                                   Type of mode.  One of
                                                    SQL,CSV
 -nns,--non_numeric_sample_size <NUM>               Sample size for
                                                    non-numeric data.
 -ns,--numeric_sample_size <NUM>                    Sample size for
                                                    numeric data.
 -o,--output <SOURCE>                               output location
 -pct,--percentiles <PCTILE1[,PCTILE2]*>            A comma separated
                                                    list of percentiles
                                                    in (0, 100].
 -smo,--similarity_min_occurrance <NUM_OCCURANCES>  Min Occurrances to be
                                                    considered for
                                                    synonyms
 -ssc,--similarity_score_cutoff <SCORE_CUTOFF>      Similarity score
                                                    cutoff.  Scores are
                                                    cosine sim., so they
                                                    range from [0,1],
```


Column Statistical Details
HL7Text (0% Missing)
ObservationYear (0% Missing)
LabObservationGuid (0% Missing)
UserGuid (0% Missing)
HL7Identifier (0% Missing)
ReferenceRange (5% Missing)
Units (21% Missing)
IsAbnormalValue (0% Missing)
ObservationValue (0% Missing)
LabPanelGuid (0% Missing)
HL7CodingSystem (0% Missing)
ResultStatus (0% Missing)
AbnormalFlags (92% Missing)

Summary for HL7Text

Count Statistics

Type	Modifier	Count	Distinct Count
STRING	VALID	29014	348

Canonical Representation Count
VALID STRING

Canonical Value	Count
bilirubin	1679 ▲
protein	1657 ■
hemoglobin	1585 ▒
potassium	1474 ▒
chloride	1472 ▼

Possible Value Synonymns

word	synonym
Albumin, Serum	Bilirubin, Total ▲
Iron Saturation	Prostate Specific Ag... ■
Calcium, Serum	Protein, Total, Seru... ▒
T4,Free(Direct)	eGFR AfricanAmerican ▒
eGFR	eGFR AfricanAmerican ▼

Summary for Units

Count Statistics

Type	Modifier	Count	Distinct Count
STRING	MISSING	6122	1
STRING	VALID	22892	79

Canonical Representation Count
MISSING STRING

Canonical Value	Count
	6122

Canonical Representation Count
VALID STRING

Canonical Value	Count
g/dl	4541 ▲
x{d}{d}e{d}/ul	4180 ■
mg/dl	4032 ▒
mmol/l	3443 ▒
%	2640 ▼

Possible Value Synonymns

word	synonym
%	g/dL ▲
fL	pg ■
Ratio	g/dL ▒
	mL/min/1.73m2 ▒
M/uL	mil/cmm ▼

Implications for Team Structure

To be successful,

Implications for Team Structure

To be successful,

- Your data science teams have to be integrally involved in the data transformation and understanding.

Implications for Team Structure

To be successful,

- Your data science teams have to be integrally involved in the data transformation and understanding.
- Your data science teams have to be **willing** to get their hands dirty

Implications for Team Structure

To be successful,

- Your data science teams have to be integrally involved in the data transformation and understanding.
- Your data science teams have to be **willing** to get their hands dirty
- Your data science teams have to be **allowed** to get their hands dirty

Implications for Team Structure

To be successful,

- Your data science teams have to be integrally involved in the data transformation and understanding.
- Your data science teams have to be **willing** to get their hands dirty
- Your data science teams have to be **allowed** to get their hands dirty
- Your data science teams need software engineering chops.

Questions

Thanks for your attention! Questions?

- Code & scripts for this talk available on my github presentation page.[1]
- Find me at http://caseystella.com
- Twitter handle: @casey_stella
- Email address: cstella@hortonworks.com

[1] http://github.com/cestella/presentations/