Data Preparation for Data Science

Casey Stella
@casey_stella

Hortonworks

2016

Table of Contents

Introduction

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

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/