

DataFrame Validation In Python

A Practical Introduction

Yotam Perkal • 04.06.18



WHEN FITTING YOUR MODEL
IS ONLY THE BEGINNING

<https://www.youtube.com/watch?v=UXd0EDy7aTY>

Sounds Familiar?

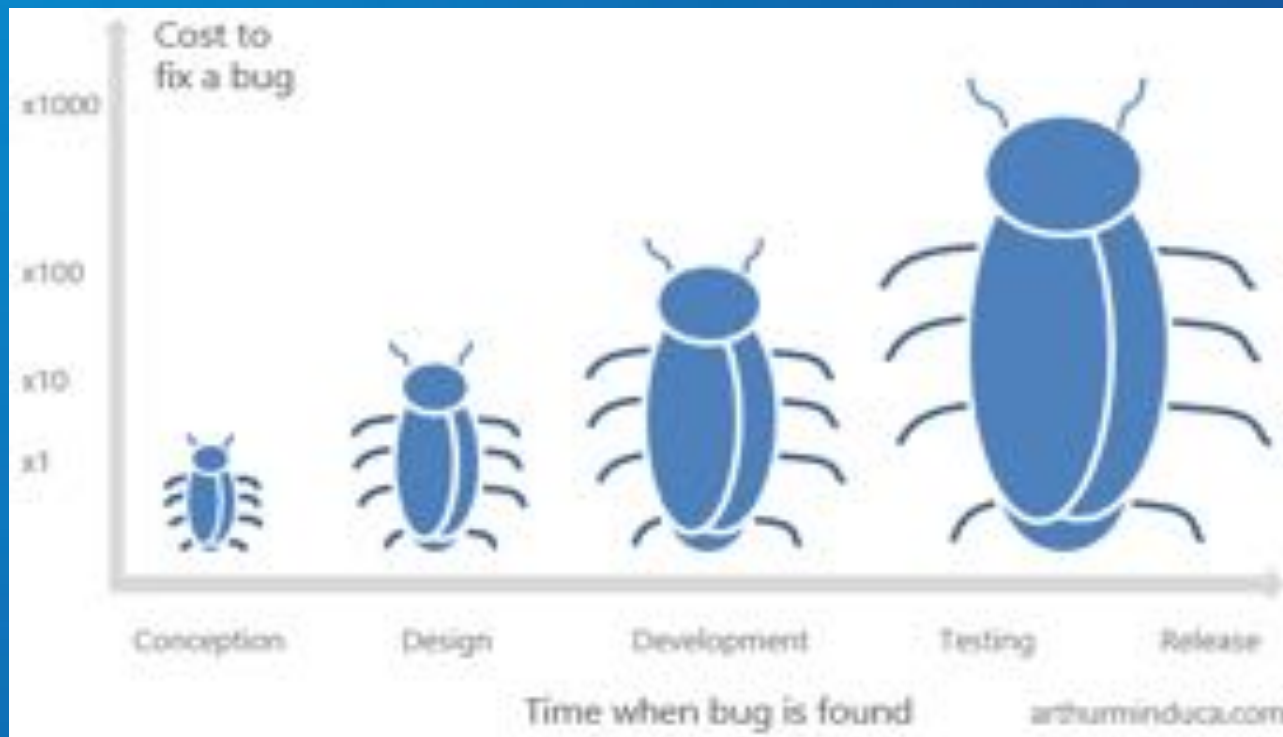
Credit:
Anaconda, Inc.
AnacondaCON 2018

About Me



Motivation

Why do we need data validation?



Data Quality Dimensions

Valid

Accurate

Complete

Consistent


Uniform

Unique

It can happen to all of us

John Lewis / Age at death

-324
1940–1616



Books and overview

The image shows a digital interface for John Lewis. It features a header with the name 'John Lewis / Age at death'. Below this, a large number '-324' is displayed, with the years '1940–1616' underneath it. To the right of the text is a portrait of John Lewis, an African American man in a suit and tie, with the US Capitol building in the background. Below the portrait and text is a circular button with a downward arrow, and at the bottom, the text 'Books and overview' is visible.

1 Perfect World



1 Perfect World

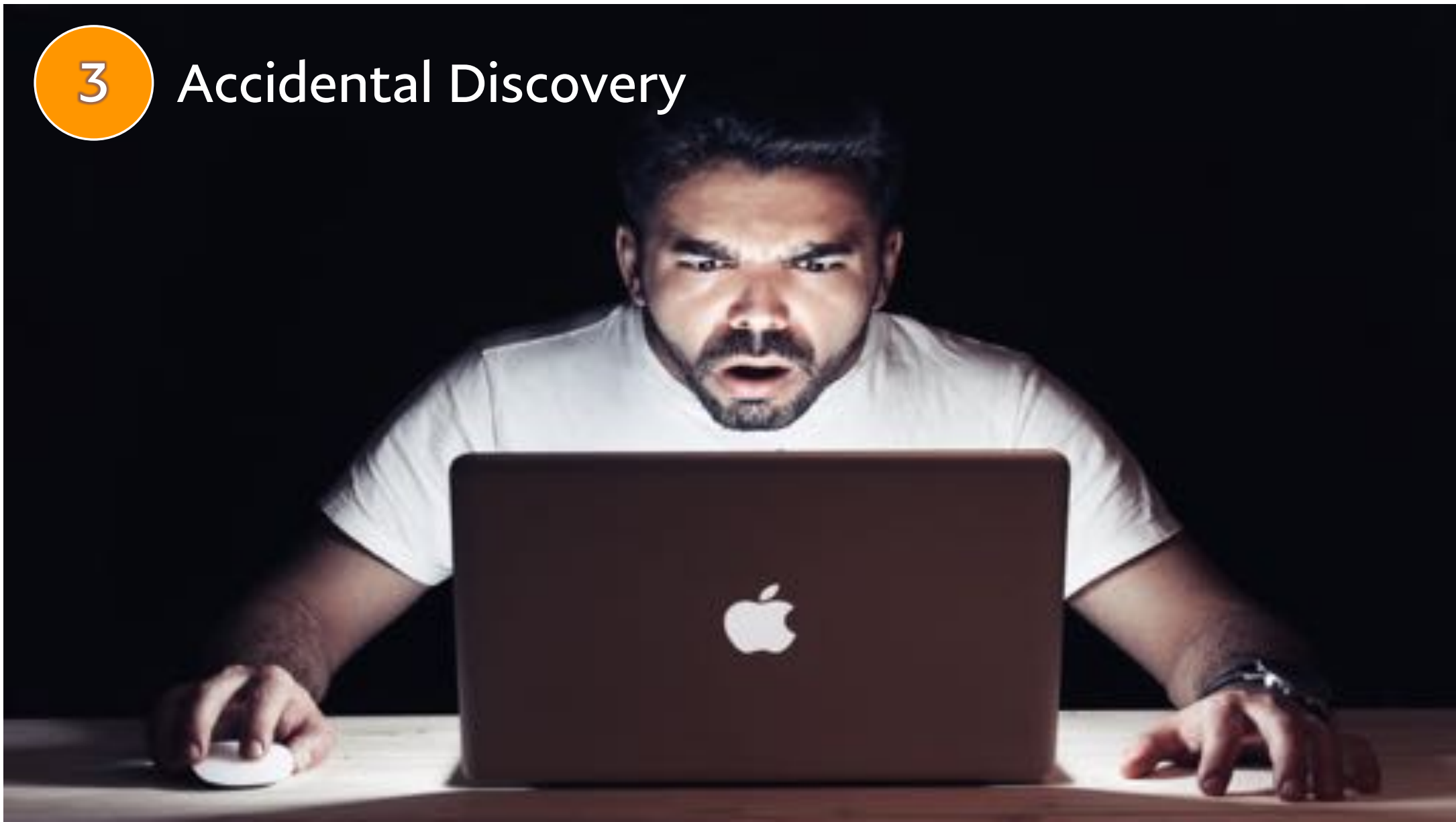


2 Model Deterioration



3

Accidental Discovery



4

Ignorance Is Bliss



Let's See Some Tools?

Voluptuous

Voluptuous is a Python data validation library.

Benefits:

- Simplicity.
- Support for complex data structures.
- Useful error messages.

Similar to Json Schema but works on the data itself.

<https://github.com/alecthomas/voluptuous>

Engarde

Explicitly state your assumptions about the data and check that they're *actually* true.

Benefits:

- Great for flat files like csv
- Supports two methods of execution:
 - As decorators, which are most useful in .py scripts
 - Interactively at the interpreter

<https://github.com/TomAugspurger/engarde>

TDDA

Test Driven Data Analysis

Applying Test Driven Development (TDD) principals to data analysis.

Benefits:

- Correctness
- Regression detection
- Specification, Design and Documentation
- Refactoring
- Portability

<https://github.com/tdda/tdda>

TDDA
Engarde

Voluptuous

A word cloud on a blue gradient background featuring various Python validation libraries. The words are arranged in a roughly triangular shape, with 'Great-Expectations' being the largest and most central. Other prominent words include 'Engarde', 'Voluptuous', 'TDDA', 'Schematics', 'Cerberus', 'PandasSchema', 'Validier', 'JsonSchema', 'MarshmallowSchema', 'Pydantic', 'Pandas-Validator', 'Colander', 'goodtables-py', and 'Hypothesis'.

Validier
PandasSchema
JsonSchema
MarshmallowSchema
TDDA
Engarde
Schematics
Cerberus
Great-Expectations
Voluptuous
Pydantic
Pandas-Validator
Colander
goodtables-py
Hypothesis

Credit - Practical Data Cleaning with Python

KATHARINE JARMUL



<http://kjamistan.com/>

“*Quality* is never an accident;
it is always the result of *intelligent*
effort.”

— John Ruskin



<https://github.com/pyotam/Dataframe-Validation>