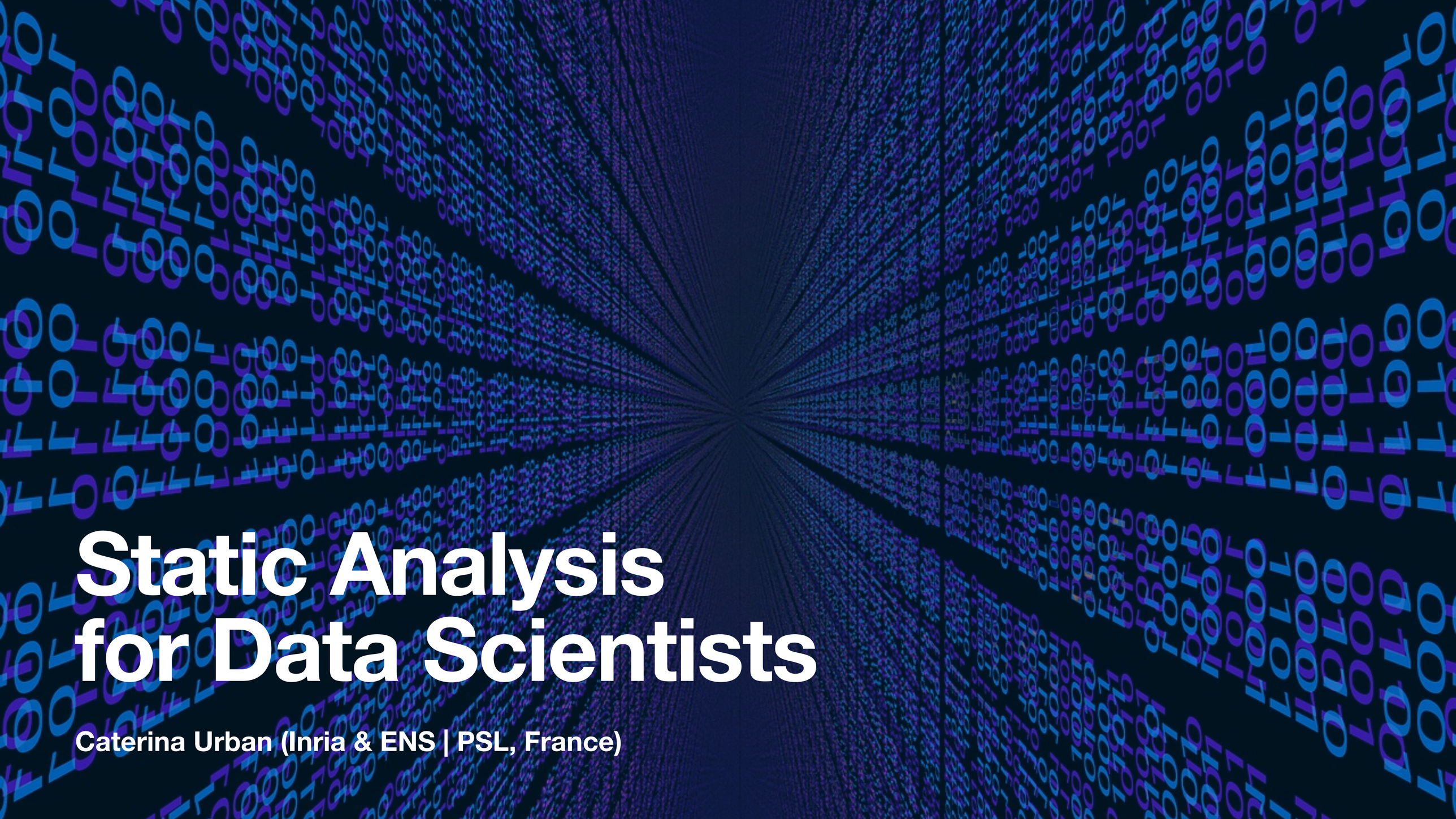
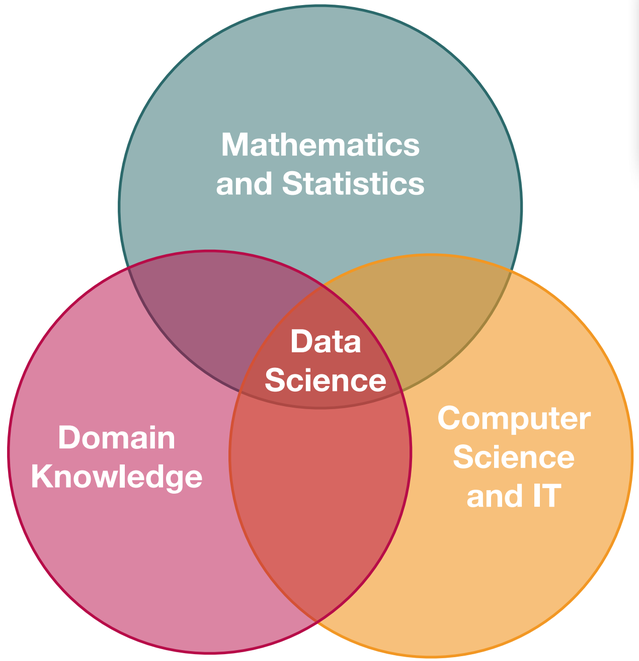


Static Analysis for Data Scientists

Caterina Urban (Inria & ENS | PSL, France)

How It Started



DATA SCIENTISTS ≠ SOFTWARE DEVELOPERS
...BUT THEY **DEVELOP SOFTWARE** NONETHELESS

https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

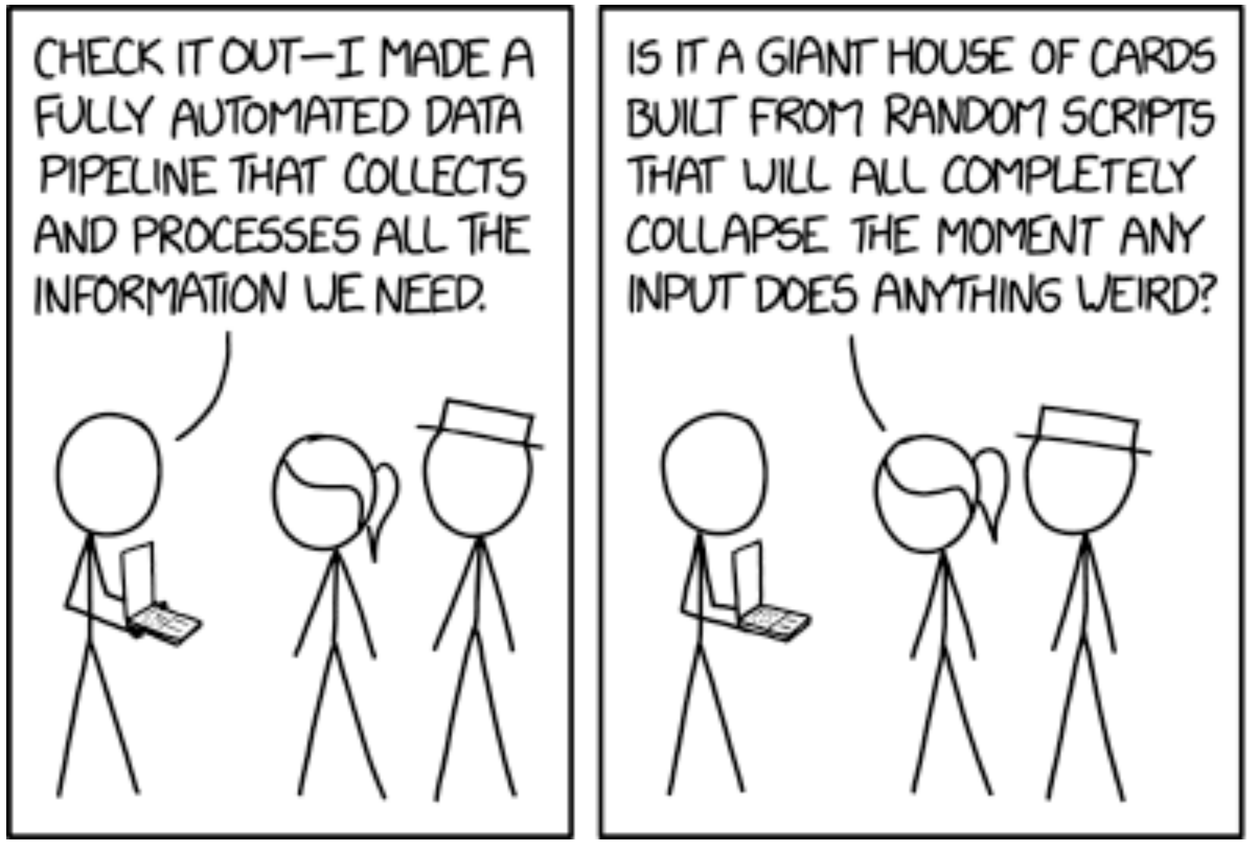
Data Scientist: The Sexiest Job of the 21st Century

Andrew McAfee and Erik Brynjolfsson

Andrew J Buboltz, silk screen on a page from a high school yearbook, 8.5" x 12", 2011 Tamar Cohen

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know

How It Is Going



https://xkcd.com/2054/

WHERE THERE IS **SOFTWARE**, THERE ARE **BUGS**

SCIENCE | US & WORLD | TECH

Excel spreadsheet error blamed for UK's 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT

WHERE THERE IS **DATA SCIENCE SOFTWARE**
THERE ARE **INSIDIOUS SILENT BUGS**

NEWS

Check for updates

The BMJ
Cite this as: *BMJ* 2020;371:m3891
http://dx.doi.org/10.1136/bmj.m3891
Published: 06 October 2020

Covid-19: Only half of 16 000 patients missed from England's official figures have been contacted

Elisabeth Mahase

Details of nearly 16 000 cases of covid-19 were not transferred to England's NHS Test and Trace service and were missed from official figures because of an error in the process for updating the data.

England's health and social care secretary, Matt Hancock, told the House of Commons on Monday 5 October that after the error was discovered on Friday 2 October "6500 hours of extra contact tracing" had been carried out over the weekend. But as at Monday [illegible] (51%) of the people had been

data and furthermore have issued guidance on validation and risk management for these products if they are to be used in such a safety critical manner."

The error came as the Labour Party's leader, Keir Starmer, said that the prime minister had "lost control" of covid-19, with no clear strategy for beating it. Speaking to the *Observer*, Starmer set out his five point plan for covid-19, which starts with publishing the criteria for local restrictions, as the German government did. Secondly, he said public health messaging should be improved by adding a feature to the NHS covid-19 app so people can search their [illegible] find out their local restrictions.

[illegible] the contact tracing

BMJ: first published as 10.1136/bmj.m3891 on 6 October

Jupyter Notebooks

localhost

jupyter House Prices (unsaved changes)

Logout

File Edit View Insert Cell Kernel Widgets Help

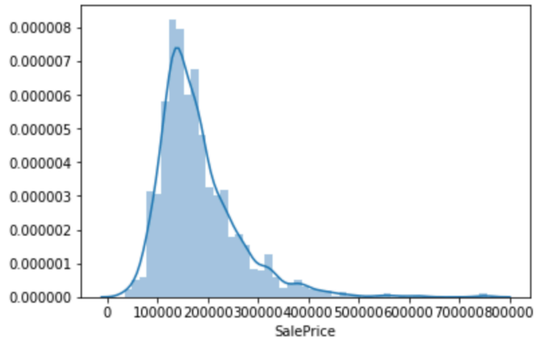
Not Trusted | Python 3

Run Code

In [2]:
```
df_train = pd.read_csv('../input/HousePrices.csv')
```

Sale Prices

In [5]:
```
sns.distplot(df_train['SalePrice']);
```



Missing Data

In [14]:
```
total = df_train.isnull().sum().sort_values(ascending=False)
percent = (df_train.isnull().sum()/df_train.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
```

...

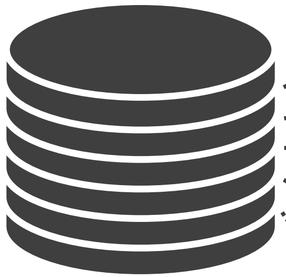
In [15]:
```
df_train = df_train.drop((missing_data[missing_data['Total'] > 1]).index,1)
df_train = df_train.drop(df_train.loc[df_train['Electrical'].isnull()].index)
df_train.isnull().sum().max()
```

Out[15]: 0


Jupyter Notebooks
NOT JUST A PROTOTYPING TOOL
localhost
jupyter House Prices (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3
Run
Code
In [2]: df_train = pd.read_csv('../input/HousePrices.csv')
Private
netflixtechblog.com
Netflix Technology Blog
Follow
Aug 16, 2018 · 13 min read · Listen
Open in app
Get started
Save
Beyond Interactive: Notebook Innovation at Netflix
By Michelle Ufford, M Pacer, Matthew Seal, and Kyle Kelley
Notebooks have rapidly grown in popularity among data scientists
facto standard for quick prototyping and explo
pushing the boundaries eve
use it, and what they can do
this vision a reality.
In this post, we'll share our m
compelling. We'll also introdu
explore some of the novel ways
If you're short on time, we sugge
Motivations
u/Desperate-Walk1780 • Commented on 12 months ago
I can tell you that my job with 120+ data scientists + data analysts on our team, we use jupyter on centos in prod. It actually is working out very well for us. Everyone knows how to use it, we can let jr devs work in prod immediately. We also have a very wide range of analysis types running from basic sql and pandas to spark based machine learning. All in jupyter. Also jupyter is easy to configure to work in security guidelines.
24 upvotes 5 replies
u/EnricoT0 • Commented on 12 months ago
My former employer uses notebooks in production for all DS tasks. My current employer does not. They are both big companies with large teams.
databricks
Databricks for Data Science
An open and unified platform to collaboratively run all types of analytics workloads, from data preparation to exploratory analysis and predictive analytics, at scale.
Loan Analysis

Unexpected Data

TESTING REQUIRES AWARENESS OF EXPECTATIONS: **HARD**!



missing values

extra values

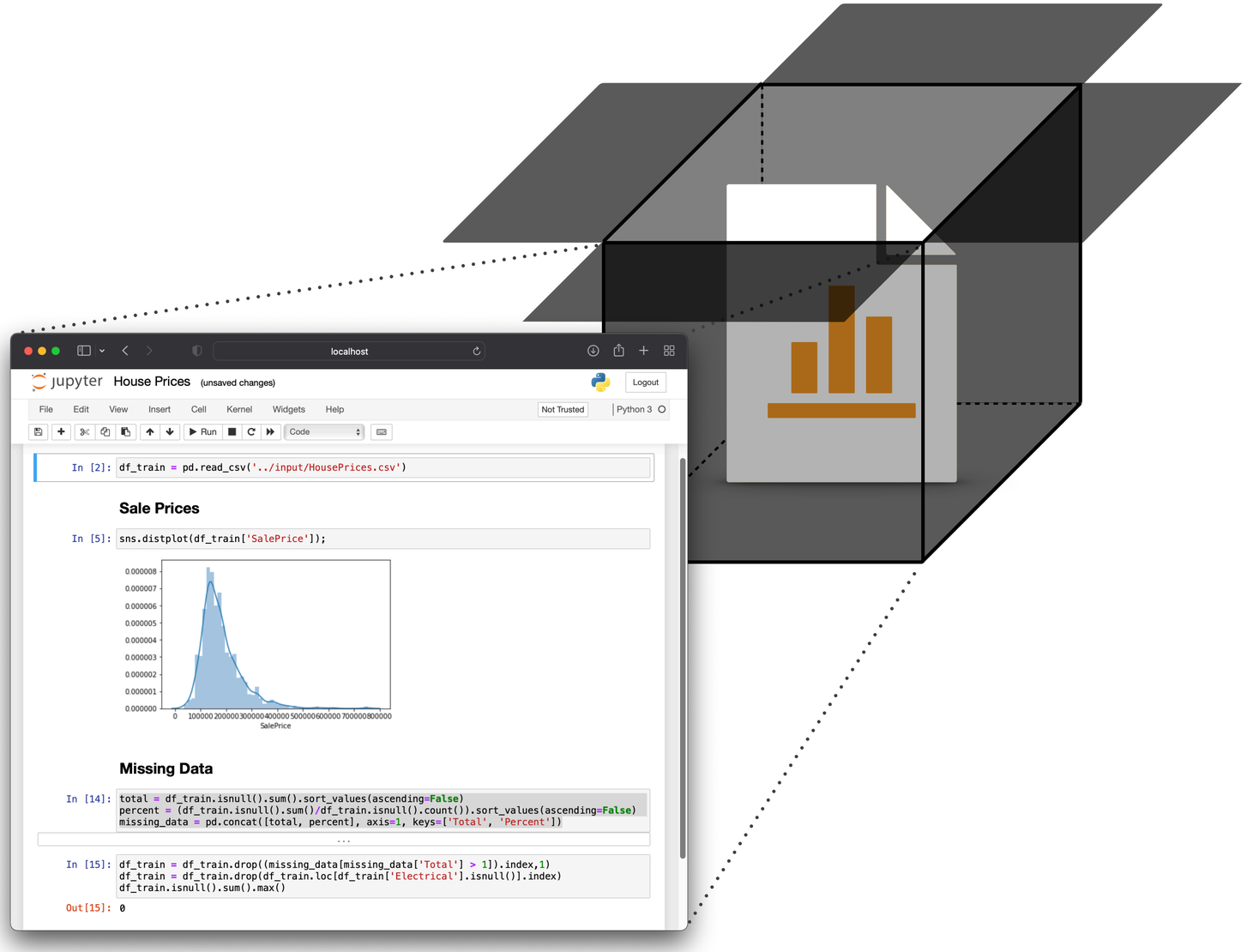
duplicate values

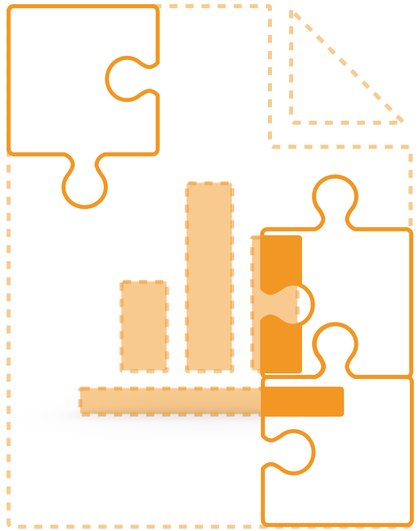
different **format**

different data

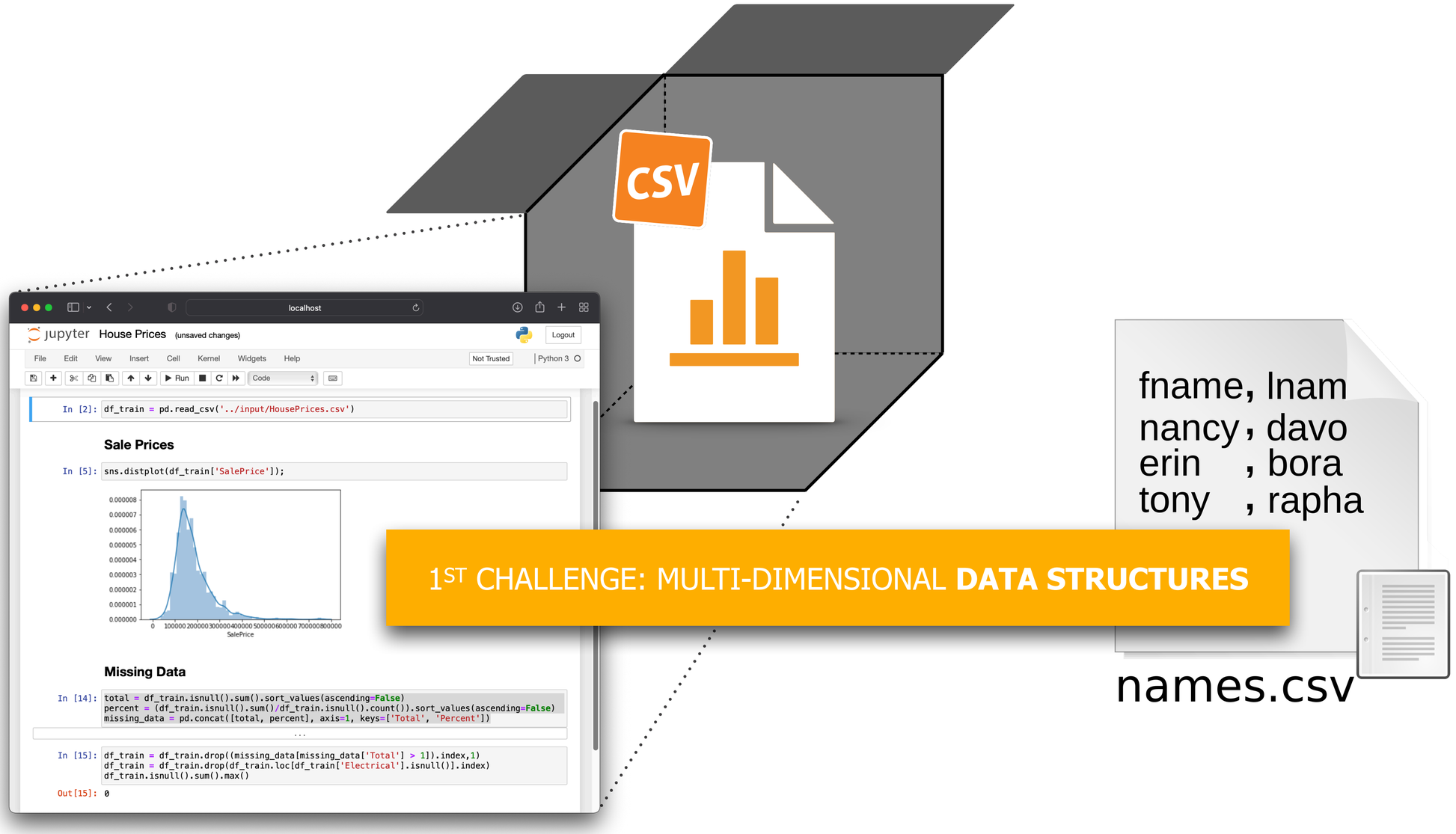

Data-Expectations Static Analysis

Data Expectations Analysis





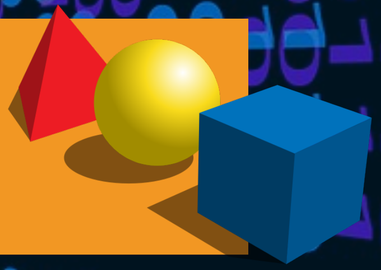
Tabular Data



Abstract Interpretation Recipe

practical tools
targeting specific programs

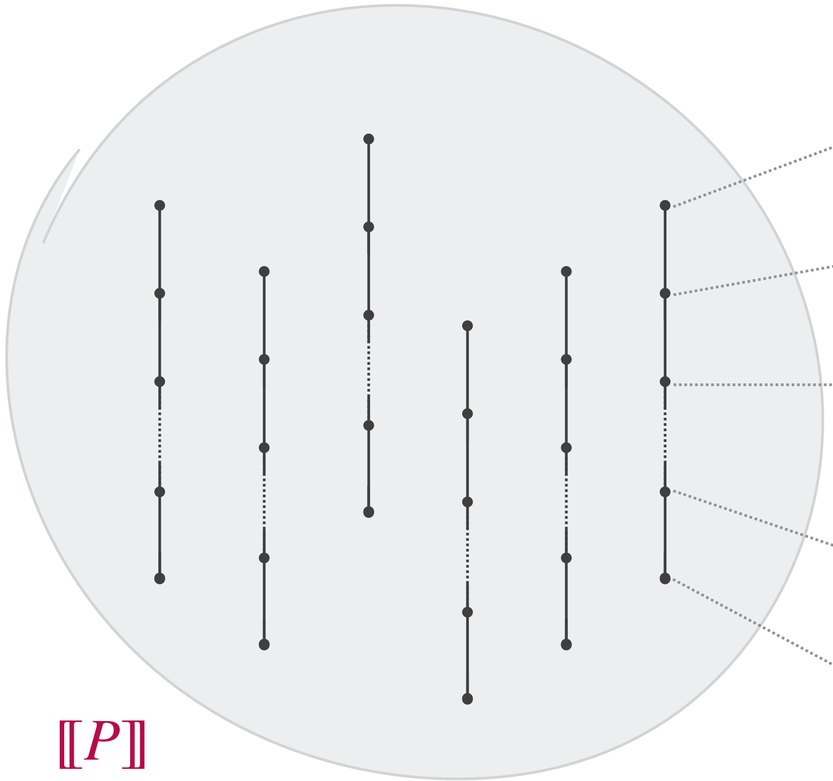
algorithmic approaches
to decide program properties



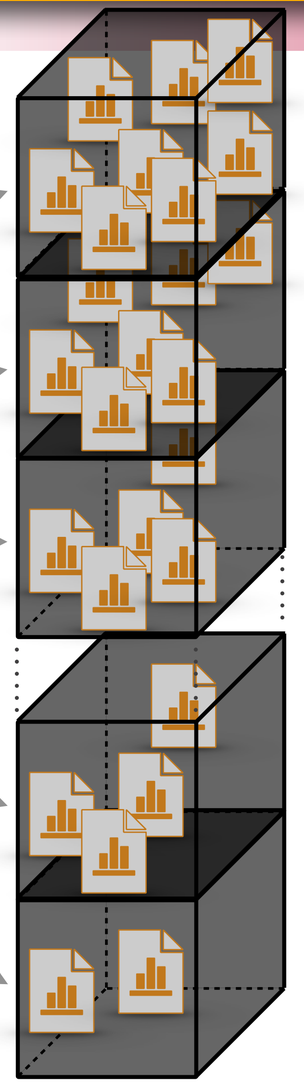
mathematical models
of the program behavior

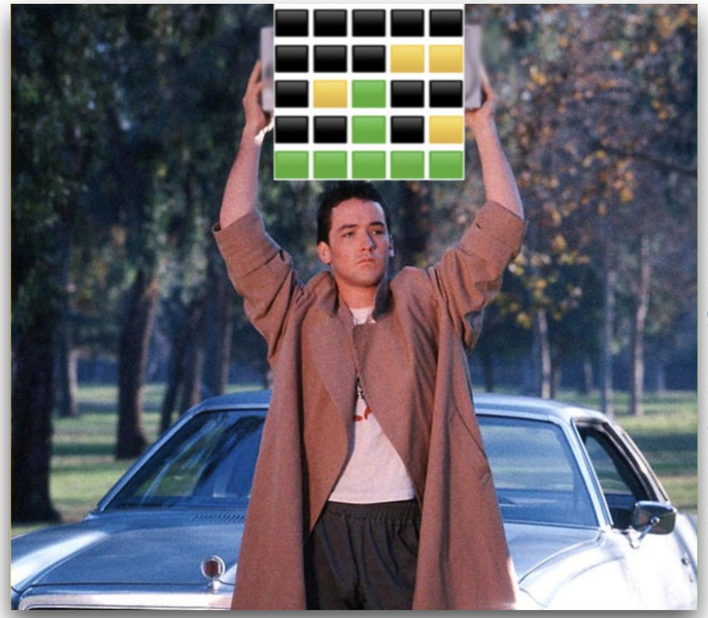
Concrete Semantics

2ND CHALLENGE: INDIRECT **REASONING**



$[\![P]\!]$





Michael J. Seidlinger/@mjseidlinger on Twitter

Abstract Semantics

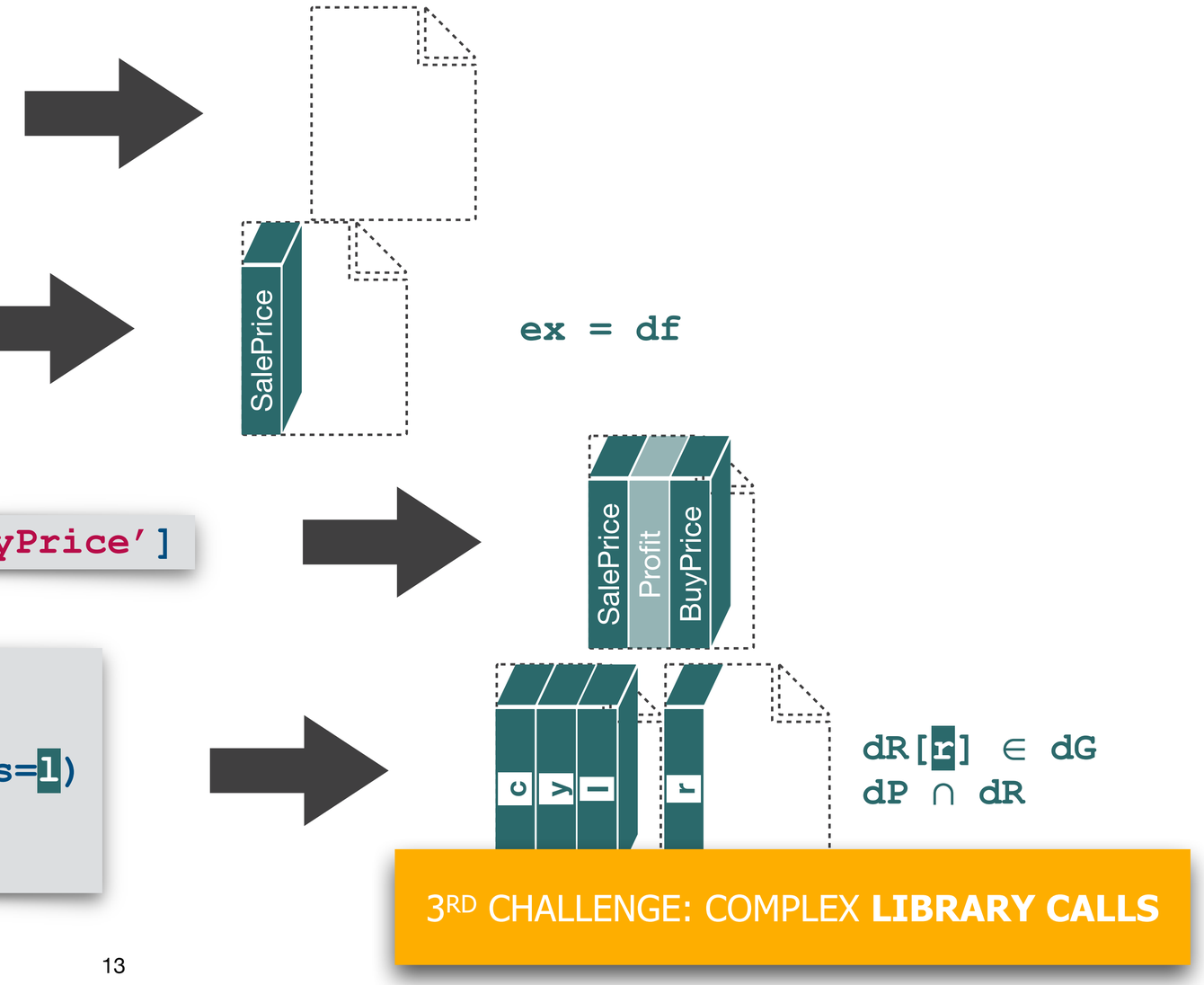
Column/Field Names Expectations

```
import pandas as pd
df = pd.read_csv("HousePrices.csv")
```

```
ex = df[df.SalePrice >= 1000000]
```

```
ex['Profit'] = ex['SalePrice'] - ex['BuyPrice']
```

```
⋮
dL = pd.read_csv("L.csv")
dP = dL.pivot(index=c, columns=y, values=l)
dR = pd.read_csv("R.csv")
dG = dP.loc[:, 0:35].groupby(dR[r])
```



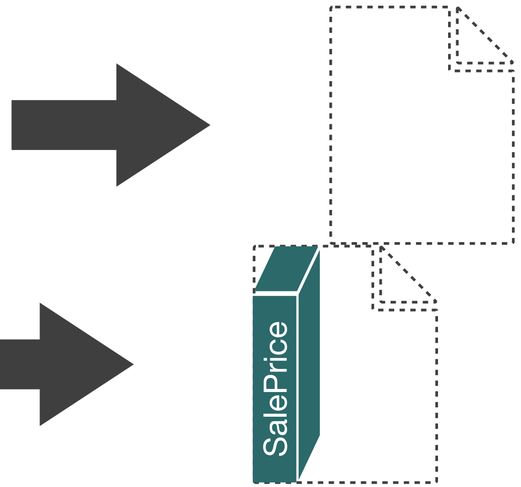
Practical Static Analysis

Necessary vs Sufficient Expectations

```
import pandas as pd
df = pd.read_csv("HousePrices.csv")
```

```
ex = df[df.SalePrice >= 1000000]
```

```
ex['Profit'] = ex['SalePrice'] - ex['BuyPrice']
```



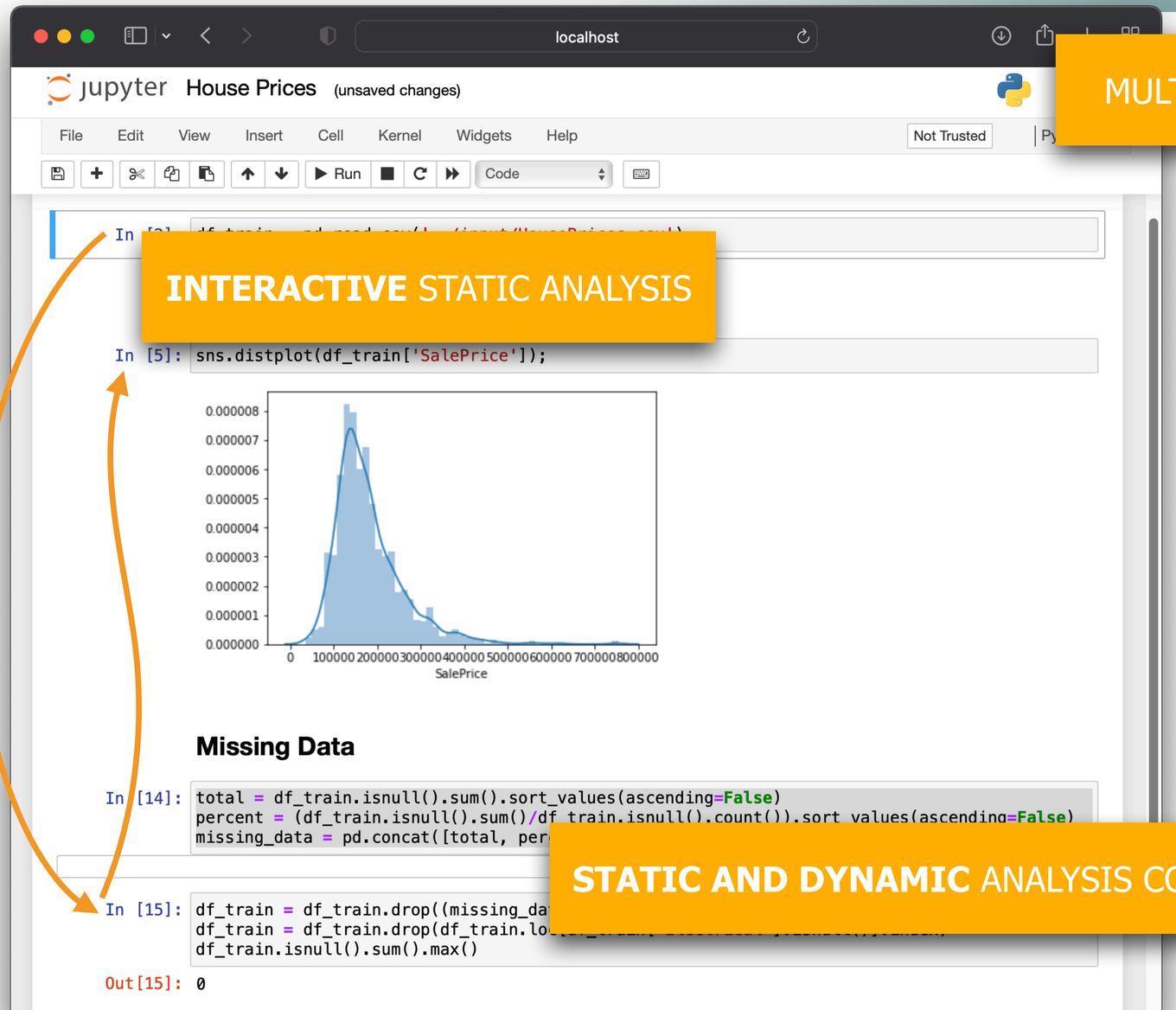
ex = df

VIOLATING NECESSARY EXPECTATIONS **CAUSES TROUBLE** FOR SURE

Implementation

Wish List

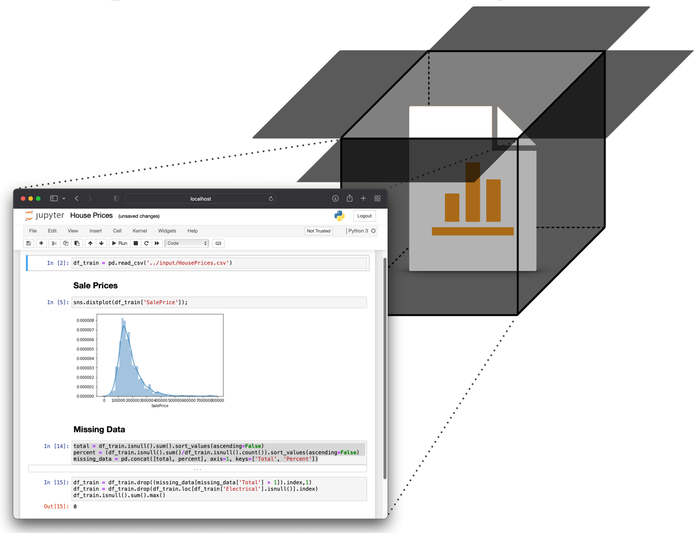
MULTI-LANGUAGE SUPPORT

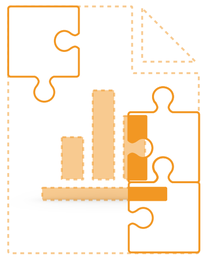


INTERACTIVE STATIC ANALYSIS

STATIC AND DYNAMIC ANALYSIS COMBINATIONS

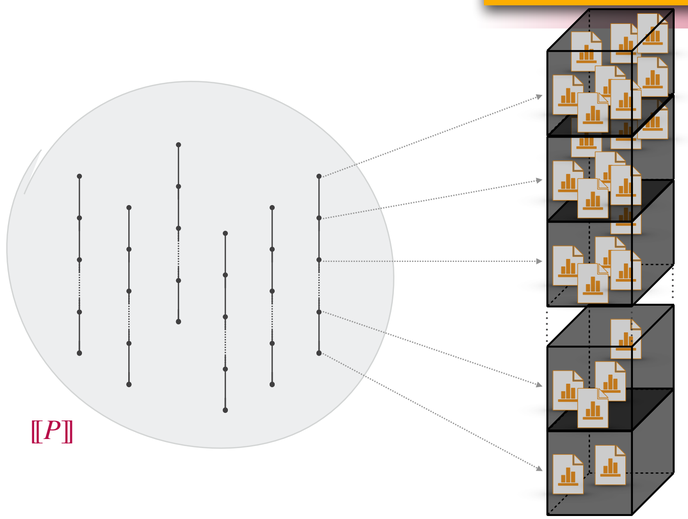
Data Expectations Analysis





Concrete Semantics

2ND CHALLENGE: INDIRECT **REASONING**



Abstract Semantics

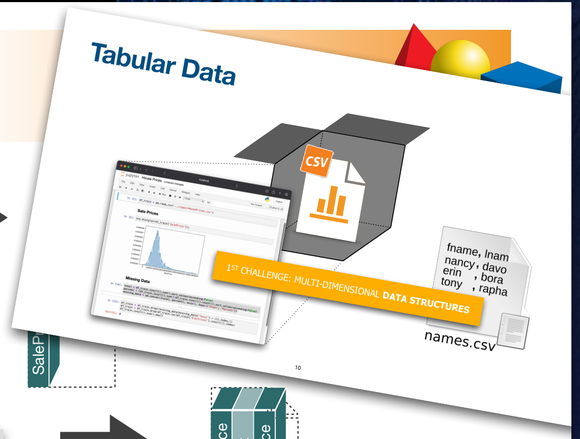
Column/Field Names Expectations

```
import pandas as pd
df = pd.read_csv("HousePrices.csv")
```

```
ex = df[df.SalePrice >= 1000000]
```

```
ex['Profit'] = ex['SalePrice'] - ex['BuyPrice']
```

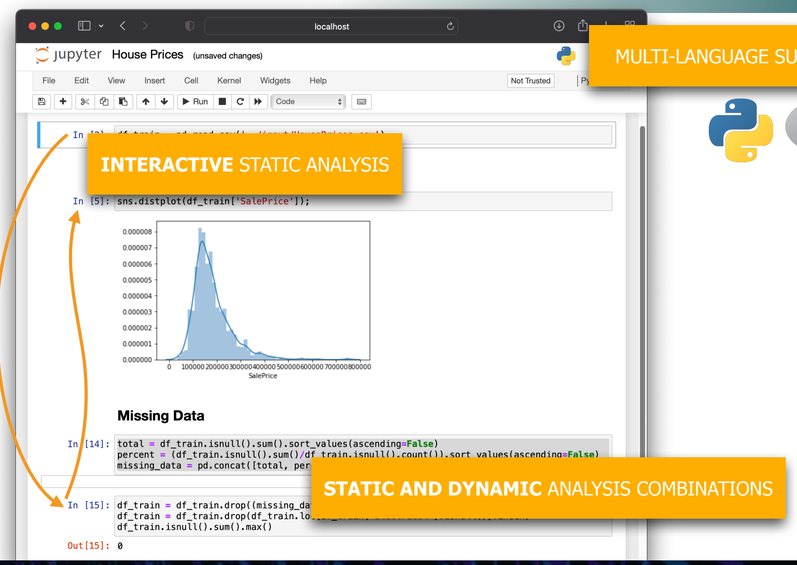
```
⋮
dL = pd.read_csv("L.csv")
dP = dL.pivot(index=c, columns=y, values=l)
dR = pd.read_csv("R.csv")
dG = dP.loc[:, 0:35].groupby(dR[r])
```



3RD CHALLENGE: COMPLEX **LIBRARY CALLS**

Implementation

Wish List



MULTI-LANGUAGE SUPPORT

INTERACTIVE STATIC ANALYSIS

STATIC AND DYNAMIC ANALYSIS COMBINATIONS