

Reproducibility in Data Science

Juliana Freire

Visualization, Imaging and Data Analysis Center (VIDA)

Computer Science & Engineering

Center for Data Science (CDS)



NYU

TANDON SCHOOL
OF ENGINEERING



Data-Driven Exploration

- Every scientific domain is moving toward data-driven exploration, this has led to great advances and discoveries
- Companies are capitalizing on data
- Government agencies uses data to operate efficiently, make policies, and informed decisions

Computing is free

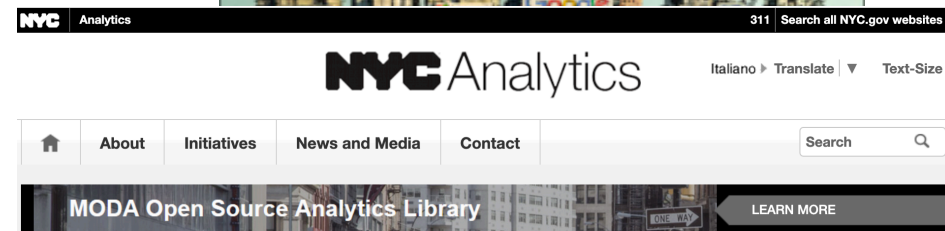
Storage is free

Data are abundant

The bottlenecks lie with people



Federal Data Strategy
Leveraging Data as a Strategic Asset



NYU

**TANDON SCHOOL
OF ENGINEERING**

ATION
AND
ALYSIS

CENTER

Data-Driven Exploration: Challenges

- Data are vast and produced at unprecedented rates
 - Sources are broad, varied, and unreliable
- Computational processes are required to extract insight
 - But they hard to assemble

algorithms machine learning
statistics math
data discovery data curation
data management
data integration provenance
visualization



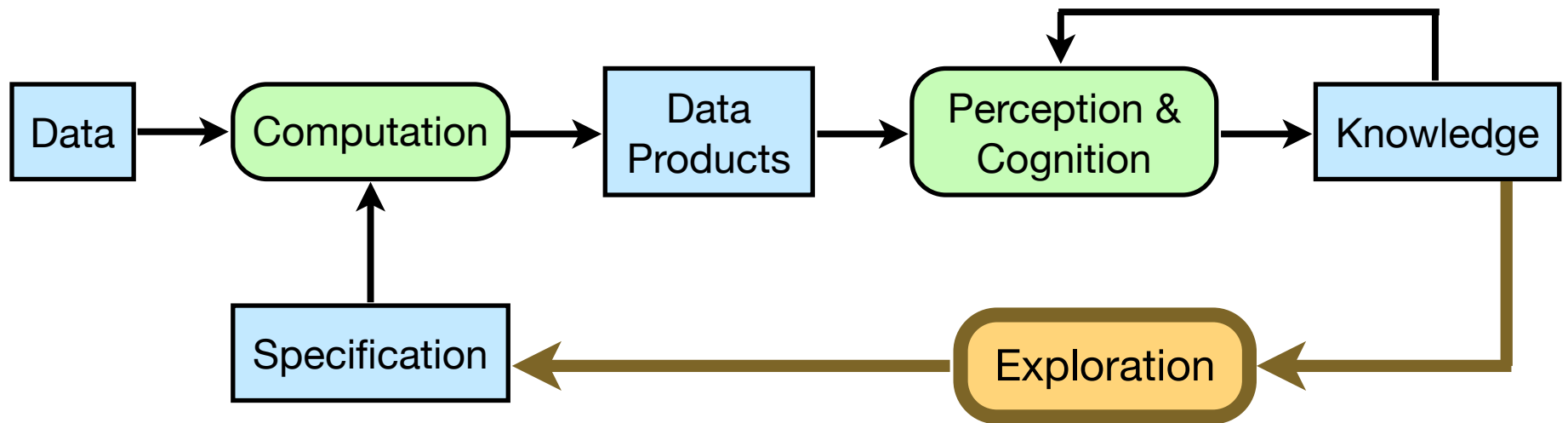
NYU

TANDON SCHOOL
OF ENGINEERING



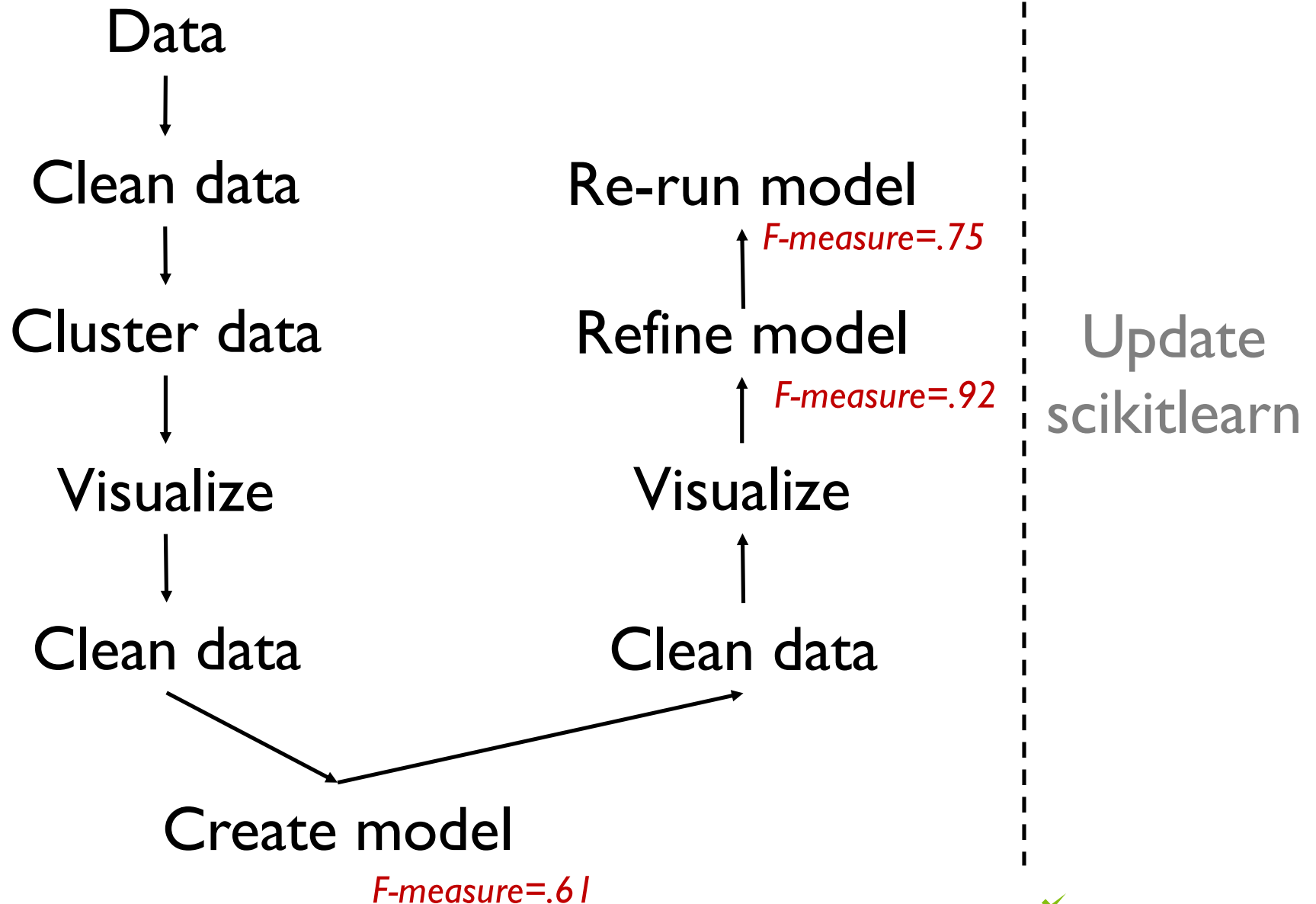
Data-Driven Exploration: Challenges

- Exploratory tasks are inherently iterative as one tests and formulates hypotheses



[Modified from Van Wijk, Vis 2005]

Many Trials and Errors...



Data-Driven Exploration: Challenges

- After many steps...

"An analysis has 30 different steps. It is tempting to just do this then that and then this. You have no idea in which ways you are wrong and your data is wrong" [Kandel et al., VAST 2012]

- It is easy to get lost and not remember what result was derived
- Processes can break or change in unforeseen ways
- Results can be hard to understand, interpret and trust

Need provenance!



Incorrect conclusions can lead to bad decisions!



NYU

TANDON SCHOOL
OF ENGINEERING

Computational Provenance

“Provenance is the source or origin of an object; its history and pedigree; a record of the ultimate derivation and passage of an item through its various owners.”

The Oxford English Dictionary

- **Provenance** is a key ingredient for transparency and reproducibility
- Computational provenance is a causality graph that models process and data dependencies

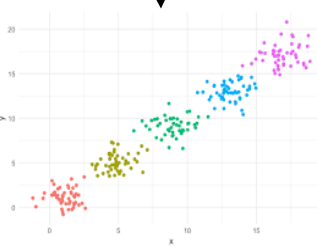
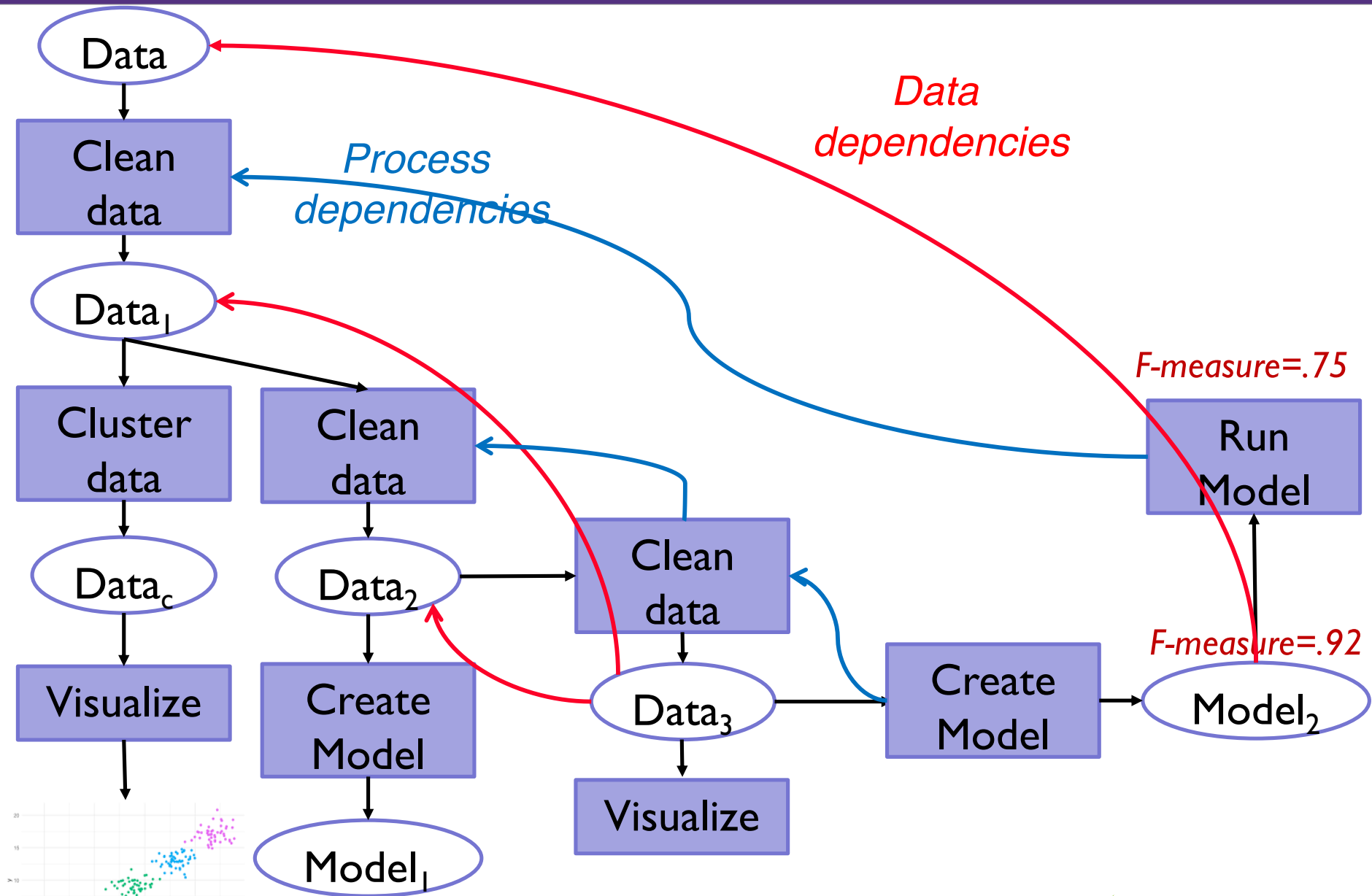


NYU

TANDON SCHOOL
OF ENGINEERING



Computational Provenance = Graph



SCHOOL OF ENGINEERING

F-measure=.61

...



VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Computational Provenance: Benefits

- Interpret and *reproduce* results
- Understand the experiment and chain of reasoning that was used in the production of a result
- Verify that an experiment was performed according to acceptable procedures
- Identify the inputs to an experiment were and where they came from
- Re-run steps, possibly with different settings
- Debug
- Share, re-use and extend results



NYU

TANDON SCHOOL
OF ENGINEERING



Different Flavors of Provenance

- Computations are carried out in a *controlled* environment
 - It is possible to systematically *capture detailed provenance*
- What to capture? Depends on what you will use provenance for:
 - Document computational process
 - Re-execute
 - Enable others to re-execute
 - Extend/modify process



NYU

TANDON SCHOOL
OF ENGINEERING



Capture the Code

Analyzing relationships between NYC taxi trips and weather

```
In [1]: from datetime import datetime
import pandas as pd
from scipy.stats import pearsonr
from scipy.stats import spearmanr
%matplotlib inline
# this makes the output of plotting commands be displayed inline
import matplotlib.pyplot as plt
```

Reading the Taxi Data

```
In [2]: # Order of attributes: time, n. trips, avg miles, avg duration (seconds)
taxi_data = pd.read_csv('Data/taxi_2012.csv', header=0)
# In the original data time is represented in secs since epoch time
# Convert to Python date-time -- provides the ability to analyze data over days, hours, etc.
taxi_data['time'] = pd.to_datetime(taxi_data['time'], unit='s')
# create an index to speed up access
taxi_data.index = taxi_data['time']
# since index already has this information, we can delete the column in the dataframe
del taxi_data['time']
```

```
In [3]: t
Out[3]:
```

What do you get?
Is this enough?



Notebooks and Reproducibility

- Recent study of 1,435,373 notebooks collected from 265,143 GitHub repositories
- 1,029,279 attempted executions of valid notebooks (i.e., notebooks with defined Python version and execution order)
 - Only 25.28% executed without errors, and
 - 4.57% produced the same results
- Problems:
 - No specification of library versions
 - Hard-coded paths
 - Out-of-order cells
 - Hidden states



NYU

TANDON SCHOOL
OF ENGINEERING

[Pimentel et al., MSR2019]



Notebooks: Best Practices

- Use relative paths (or external data repositories)
- Re-run notebook top to bottom before committing
- Declare dependencies and library versions
- Use clean environment to test dependencies

Or use  <https://www.reprozip.org/>



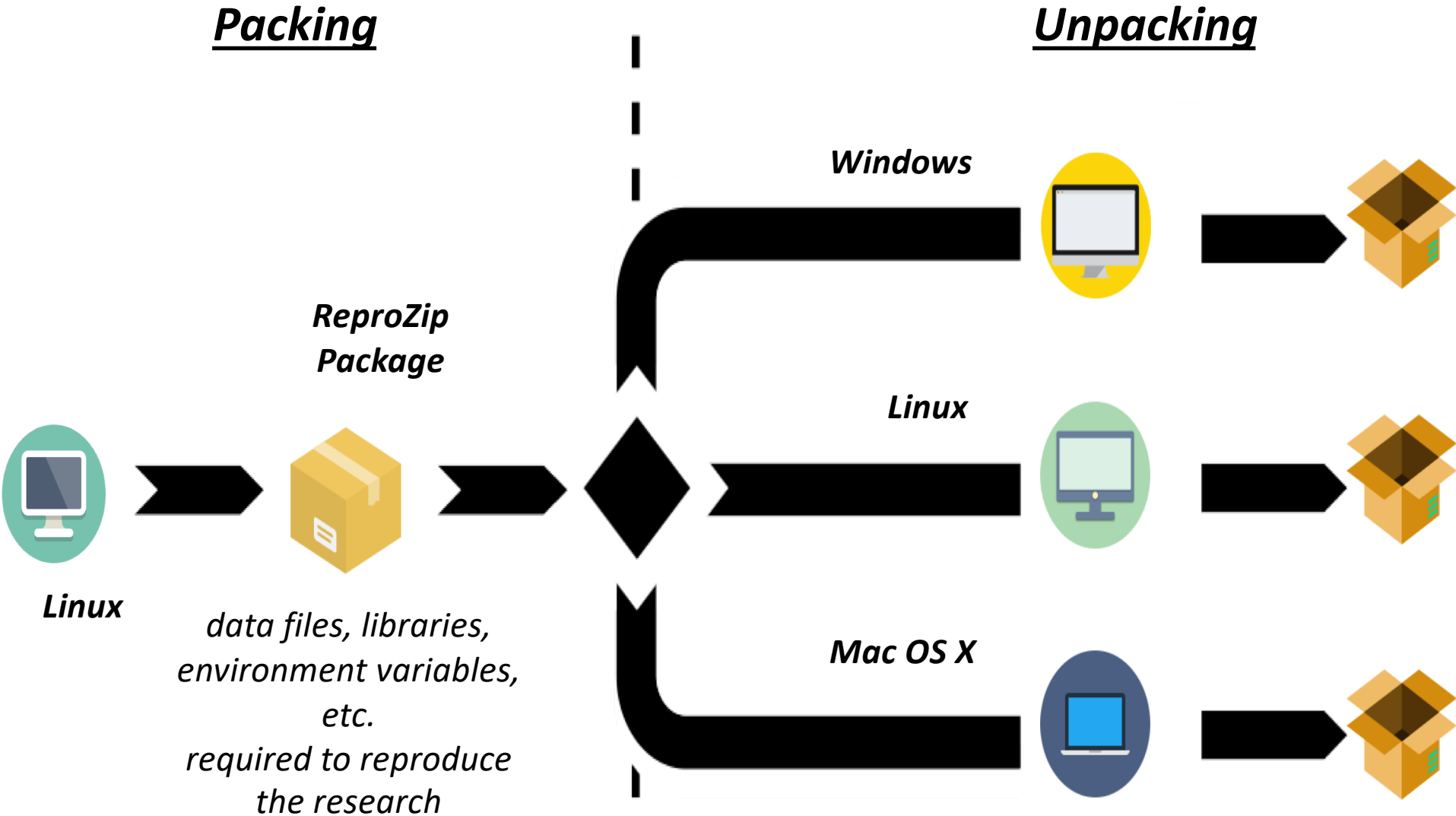
NYU

TANDON SCHOOL
OF ENGINEERING

[Pimentel et al., MSR2019]



ReproZip: Reproducibility in 2 Steps



open, unpack, and reproduce anywhere, anytime!



NYU

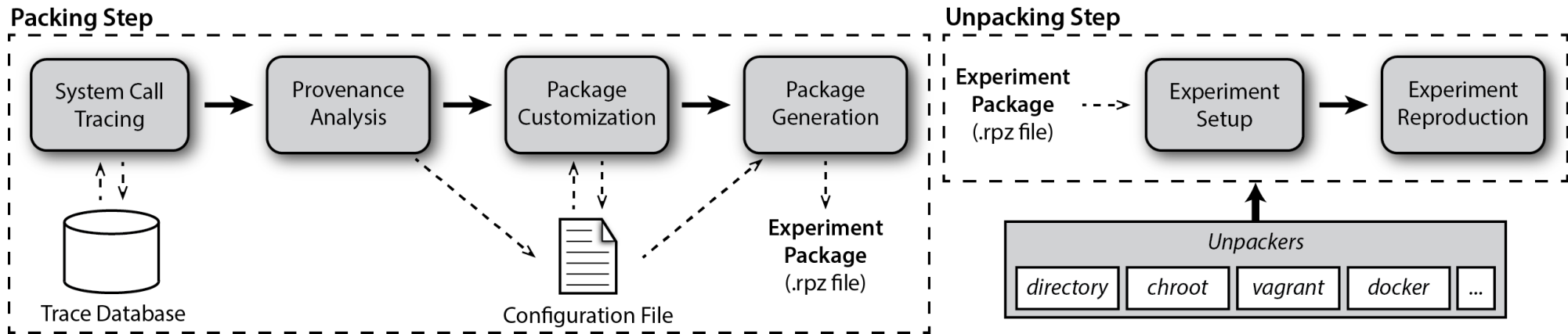
TANDON SCHOOL OF ENGINEERING



ReproZip: Advantages

- Automatically tracks dependencies in an environment and set them up in a different environment – *portability*
- Deals with *variability* in computational environments
- Reproducibility in hindsight
- Very easy (I will show!)

ReproZip: How does it work?



<https://www.youtube.com/watch?v=-zLPuwCHXo0>



NYU

TANDON SCHOOL
OF ENGINEERING

[Chirigati et al., ACM SIGMOD 2013]



Packing a Notebook



Packing

```
Ubuntu Start Page - Mozilla Firefox  
vagrant@ubuntu-1604-amd64: ~/reprozip-examples/visual-white-matter/visual-white-matter 116x36  
vagrant@ubuntu-1604-amd64:~/reprozip-examples/visual-white-matter/visual-white-matter$
```



NYU

TANDON SCHOOL
OF ENGINEERING

VIDA
VISUALIZATION
IMAGING AND
DATA ANALYSIS
CENTER

Reproducing the Notebook

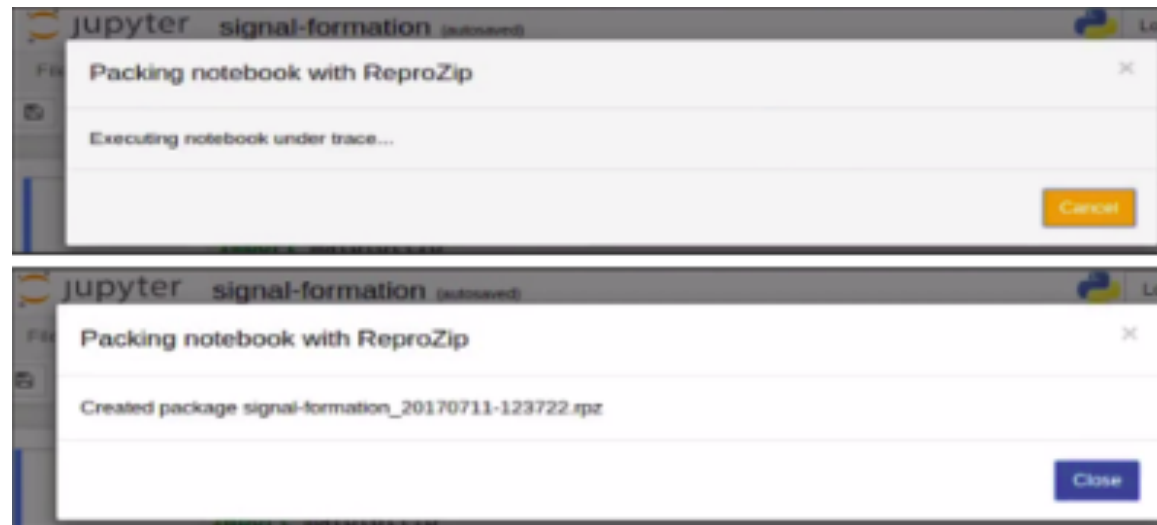
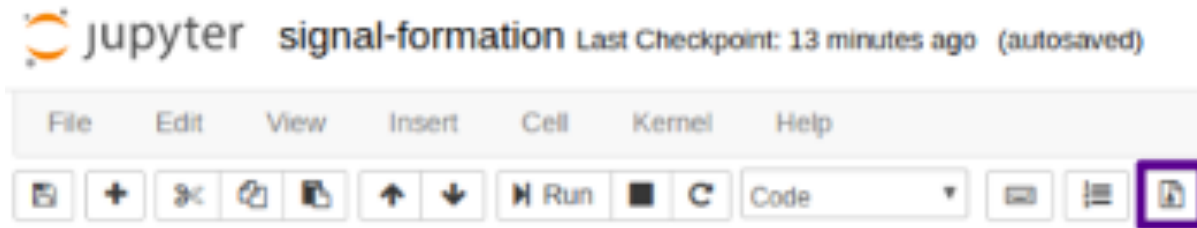


Unpacking

```
about:blank x visual-white-matter -- -bash -- 106x34 Fernando Seabra  
fchirigati@fchirigati-poly:~/projects/VIDA-NYU/reprozip-examples/visual-white-matter/visual-white-matter$
```



ReproZip Jupyter Extension



<https://docs.reprozip.org/en/1.0.x/jupyter.html>



NYU

TANDON SCHOOL
OF ENGINEERING

ReproZip can pack...

Data analysis scripts / software (any language, you name it!)

Graphical tools

Interactive tools

Client-server applications (including databases)

Jupyter notebooks (very soon!)

MPI experiments (setting up the experiment is involved though...)

... and many more!

<https://examples.reprozip.org>

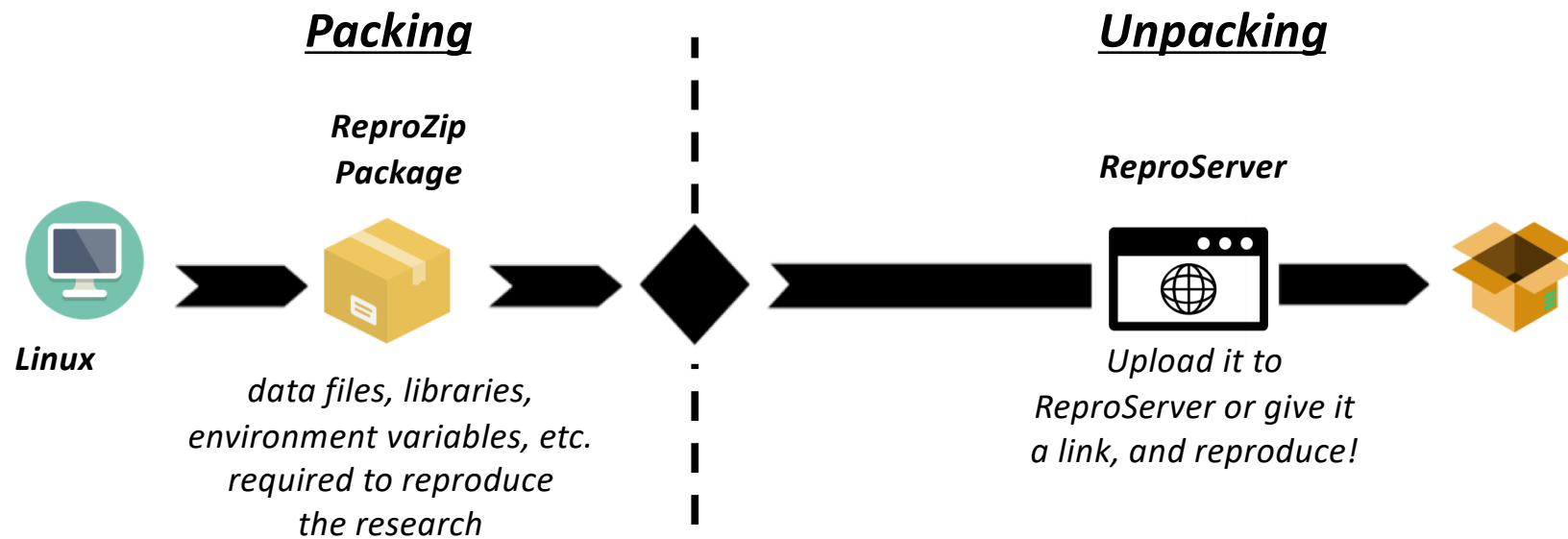


NYU

TANDON SCHOOL
OF ENGINEERING



ReproServer: Unpacking in a Browser



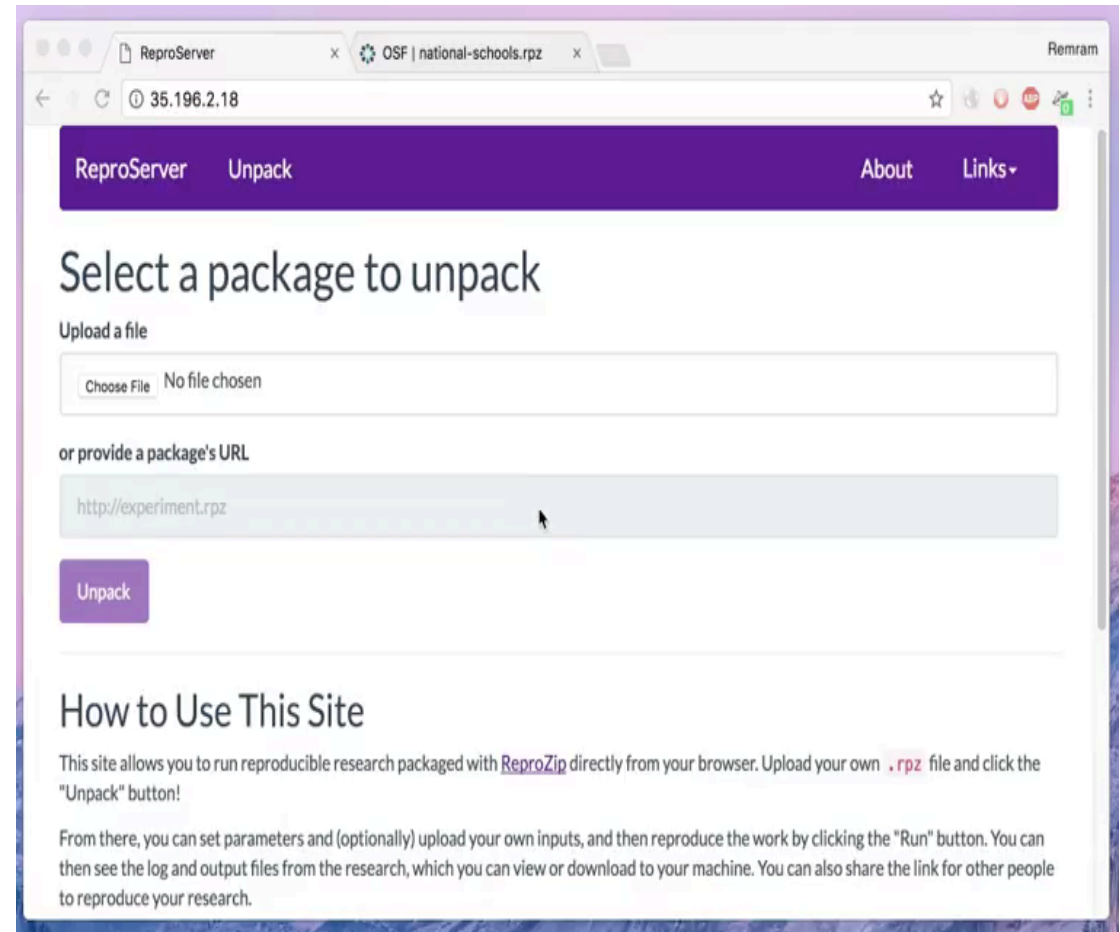
NYU

TANDON SCHOOL
OF ENGINEERING

ReproServer



- Runs ReproZip packages **in the browser**, no local software needed
- Allows **changing** input data, configuration, command-lines
- Gives you a **URL to include in papers/reports** to reproduce your experiment
- **No lock-in**: build on your laptop, pack automatically, reproduce anywhere



[Rampin et al., 2018,
<https://arxiv.org/abs/1808.01406>]

<https://www.youtube.com/watch?v=Ffb-PaVPC58>

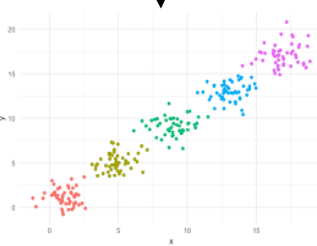
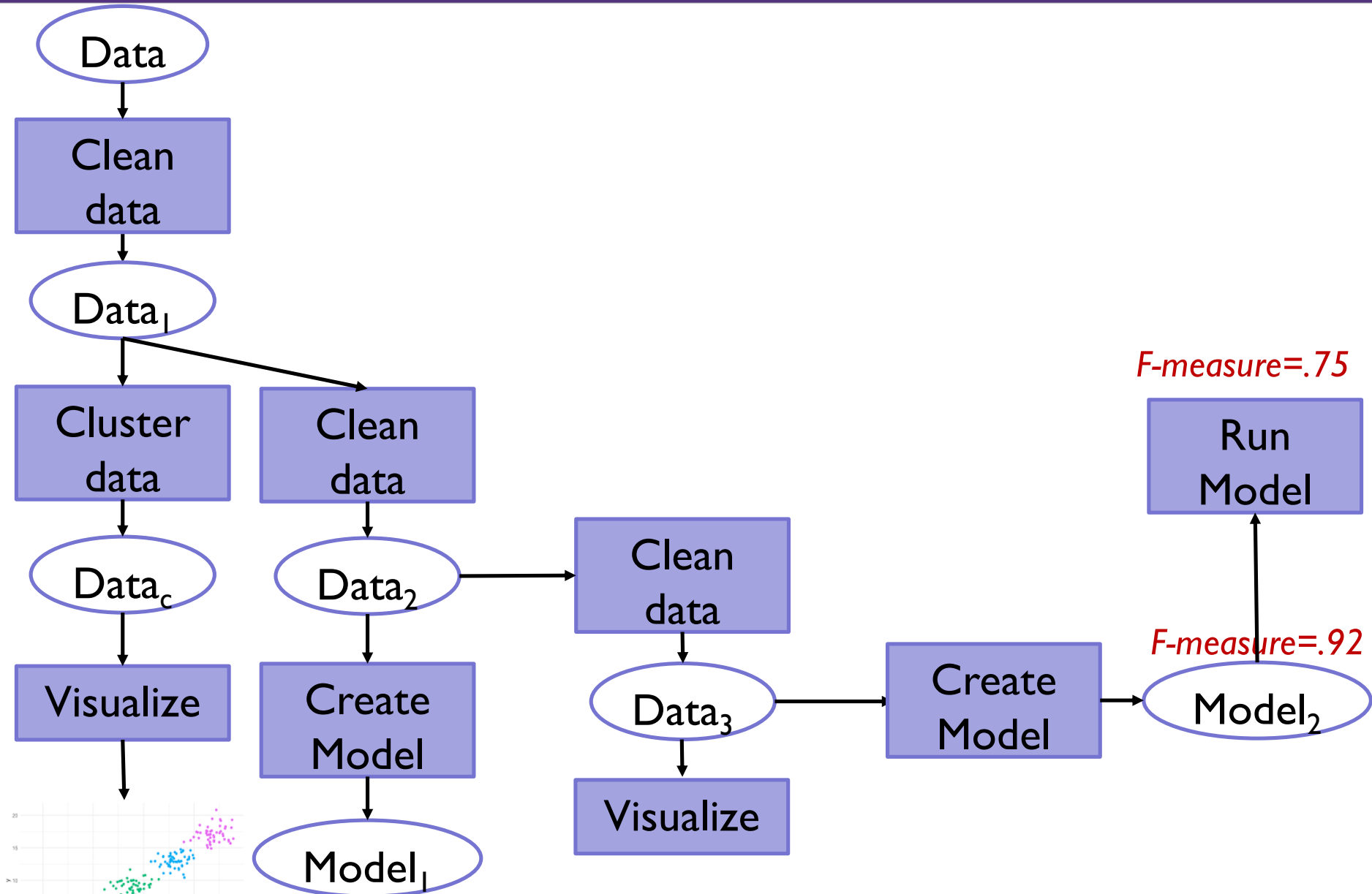


NYU

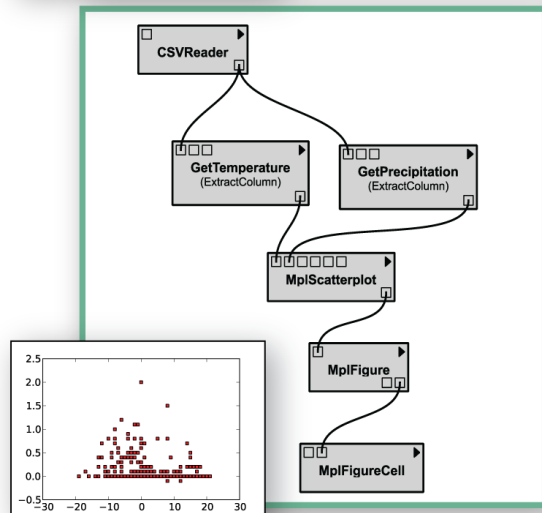
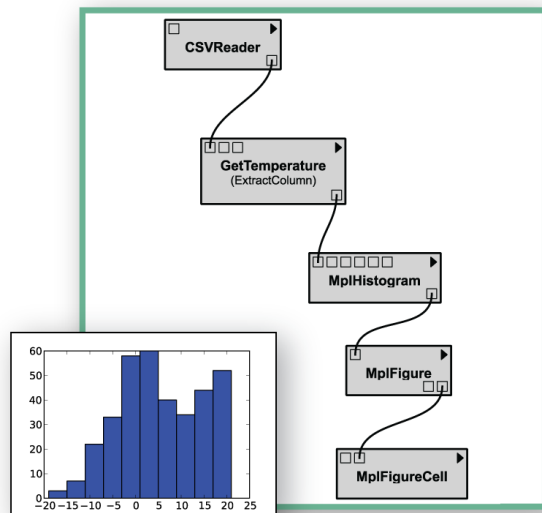
TANDON SCHOOL
OF ENGINEERING

VIDA
VISUALIZATION
IMAGING AND
DATA ANALYSIS
CENTER

Capture the Exploratory Process



Capture the Exploratory Process **Automatically**



The screenshot shows the VisTrails Builder interface for a workflow named "weather_new.vt". The main area displays a "Version Tree View" with a hierarchical structure of nodes: "basic histogram" (root), "precipitation", "scatterplot", "fahrenheit", "colors and title", "simulation", "other scatterplot", "persistent intermediate", and "persistent inputs". A green box highlights the "scatterplot" node. On the right, a "Properties" panel for the "scatterplot" node shows: Tag: scatterplot, User: dakoop, Date: 22 Oct 2010 15:28:49, and Notes: "In this workflow we also extract the precipitation data from the input file to build a scatter plot of precipitation against temperature values." Below the properties is a "Version Metadata" section and a "Preview" section showing a small scatterplot.



NYU

TANDON SCHOOL OF ENGINEERING

<http://www.vistrails.org>



VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Provenance Beyond Reproducibility

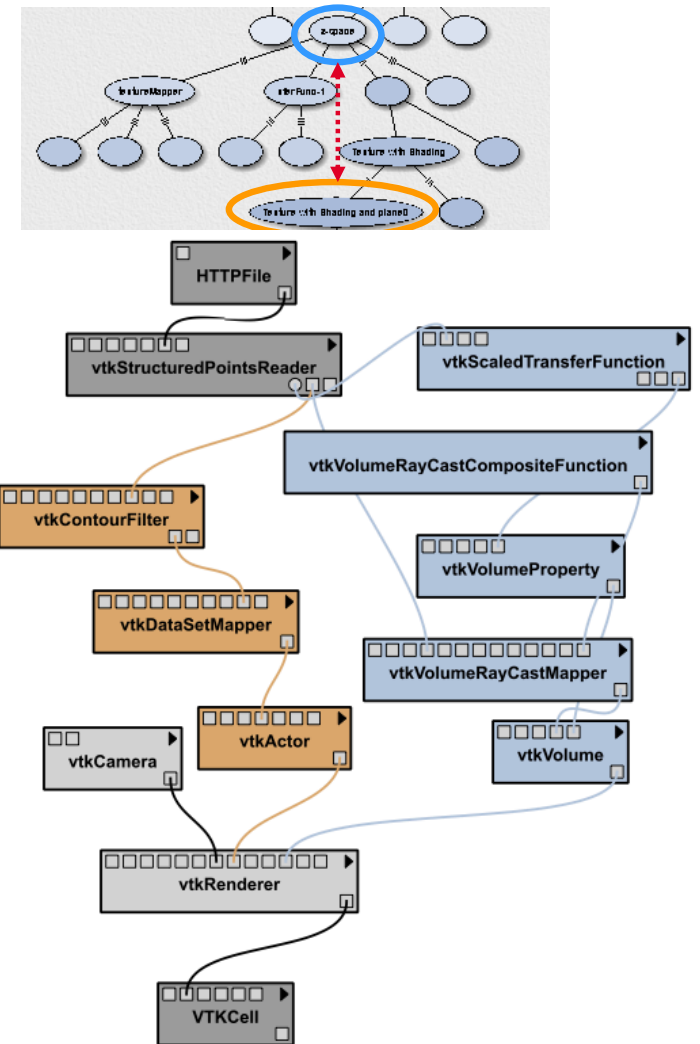
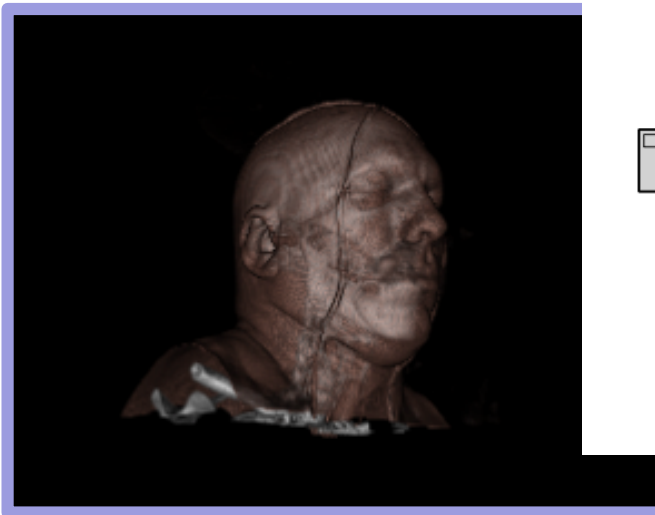
- Support for reflective reasoning
- Ability to compare data products

$$vt_1 = X_j \circ X_{j-1} \circ \dots \circ X_1 \circ \emptyset$$

$$vt_2 = X_j \circ X_{j-1} \circ \dots \circ X_1 \circ \emptyset$$

$$vt_1 - vt_2 = \{X_j, X_{j-1}, \dots, X_1, \emptyset\}$$

$$- \{X_j, X_{j-1}, \dots, X_1, \emptyset\}$$



NYU

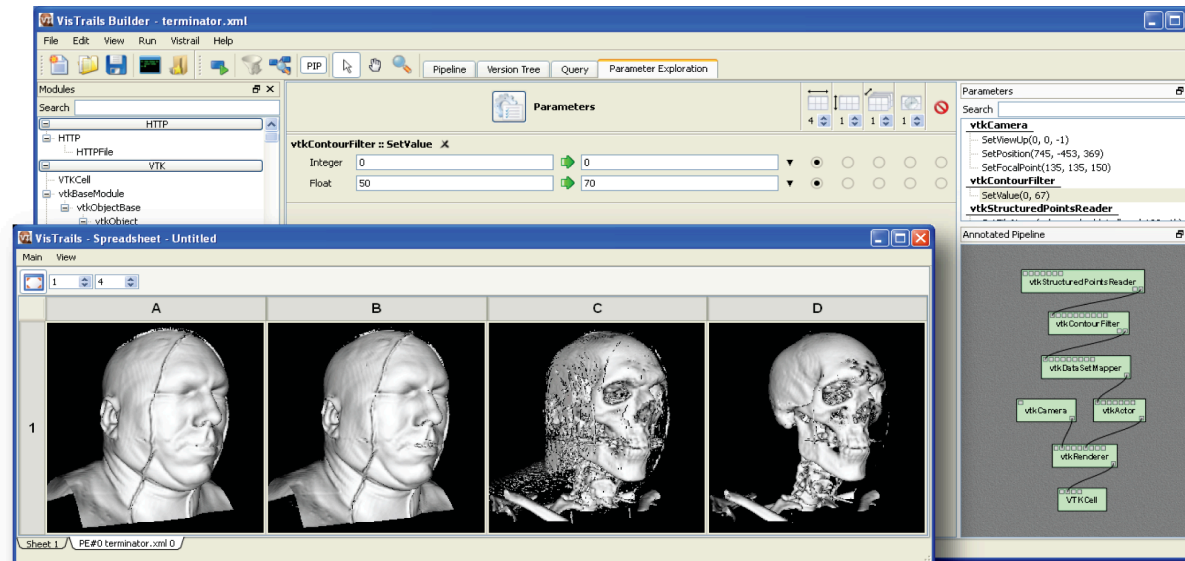
TANDON SCHOOL
OF ENGINEERING

[Freire et al., IPAW 2006]

Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products
- Explore parameter spaces and compare results
 - Also explore alternative computations

$(setParameter(id_n, value_n) \circ \dots \circ$
 $(setParameter(id_1, value_1) \circ \mathbf{V}_t)$



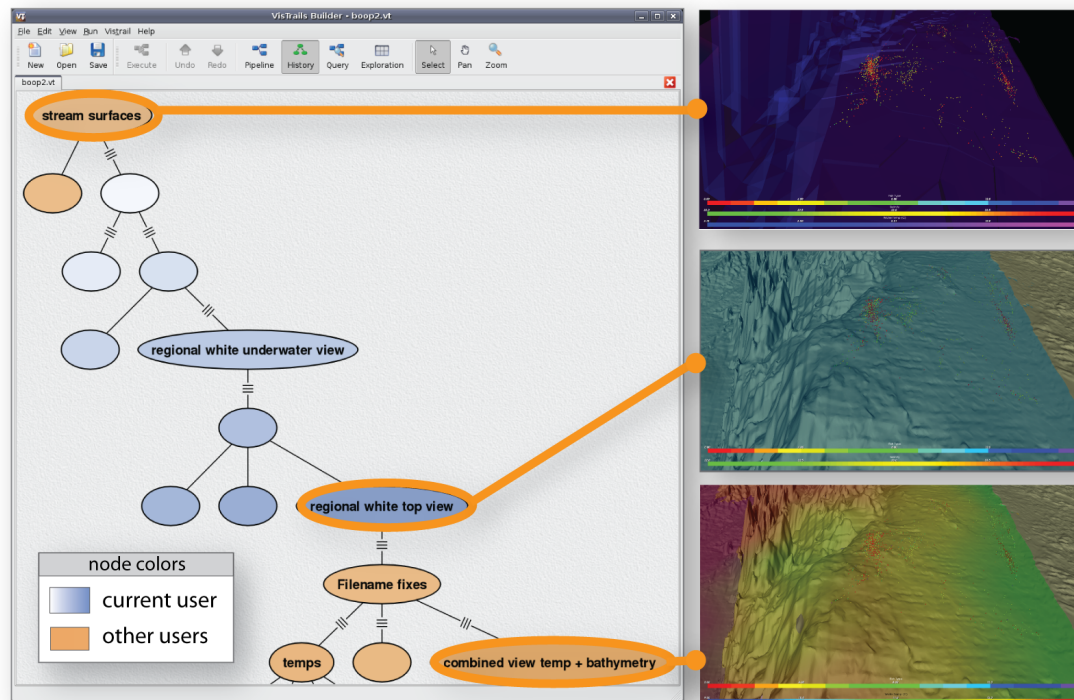
$(addModule(id_i, \dots) \circ (deleteModule(id_j) \circ \mathbf{V}_1)$

\dots
 $(addModule(id_i, \dots) \circ (deleteModule(id_j) \circ \mathbf{V}_n)$

Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products
- Explore parameter spaces and compare results
- Support for collaboration

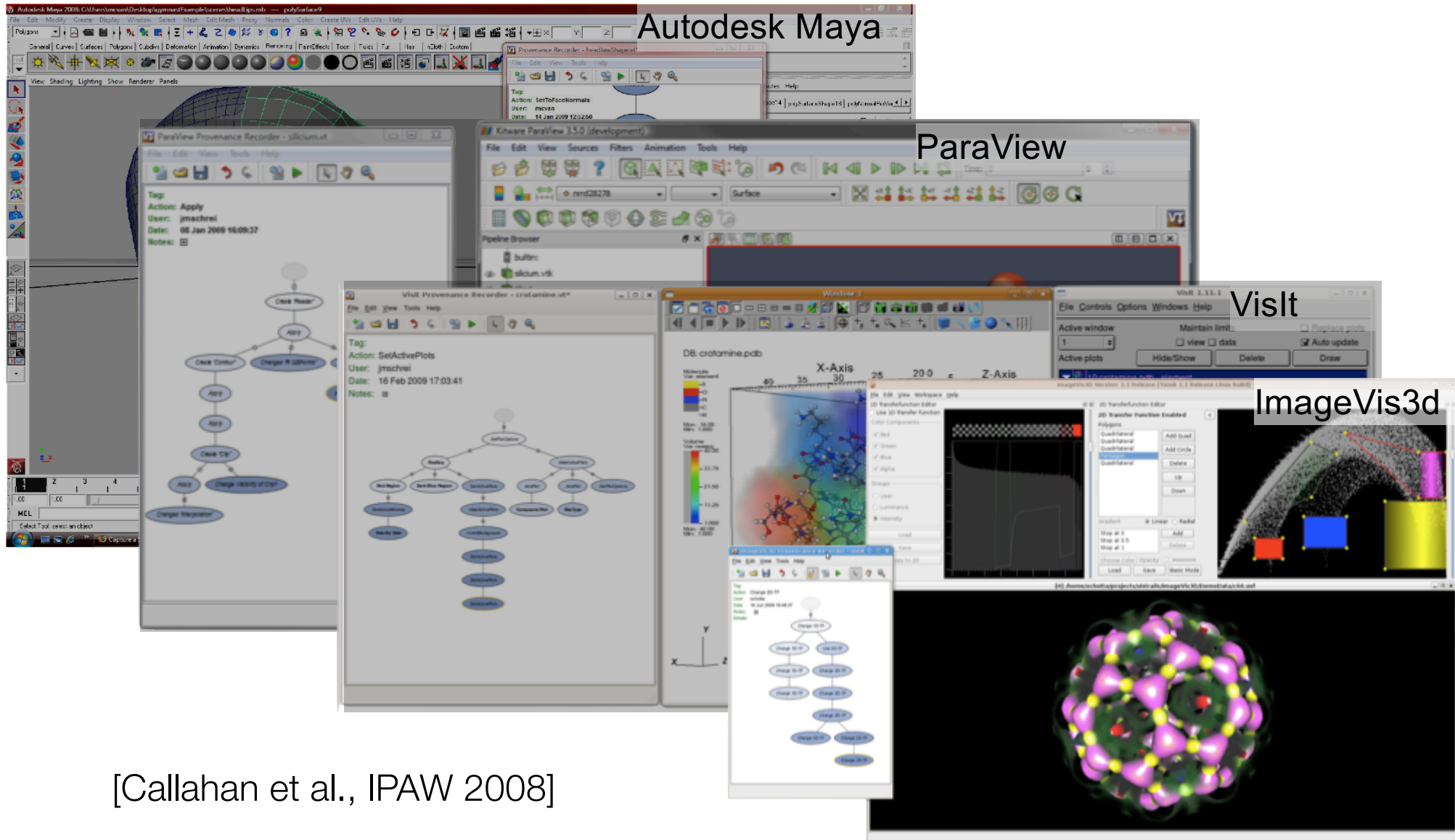
[Ellkvist et al., IPAW 2008]



NYU

TANDC
OF ENGINEERING

Change-Based Provenance: Extensibility



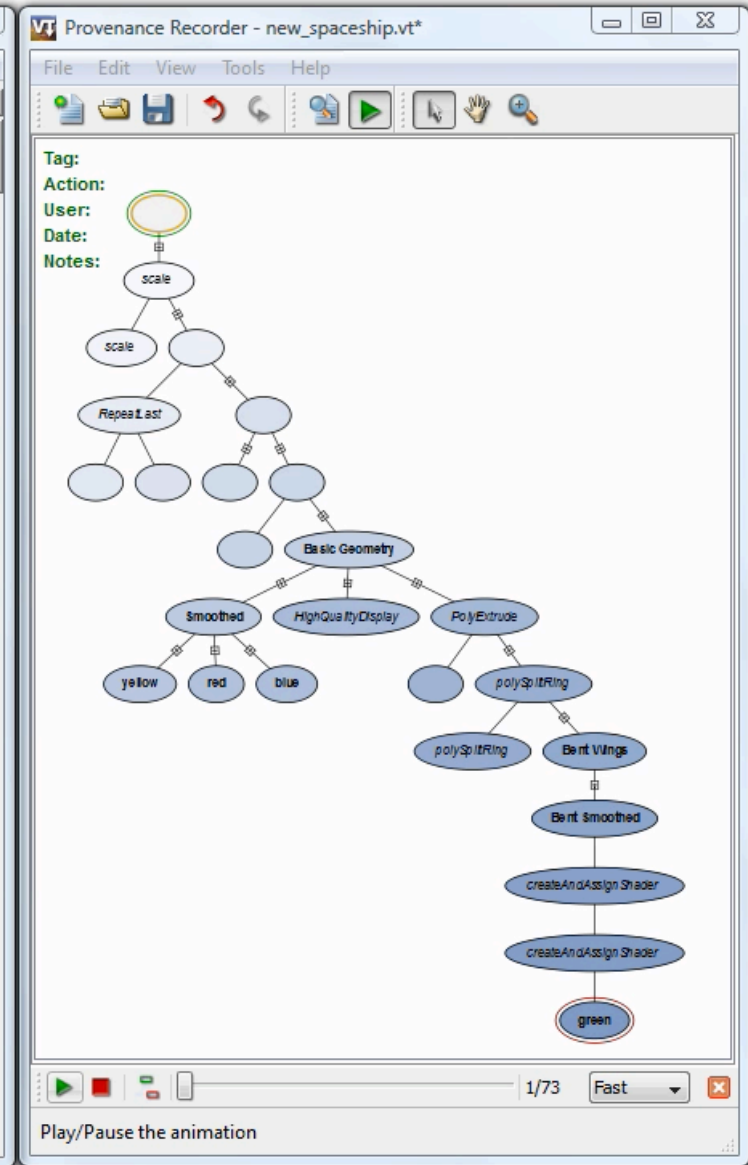
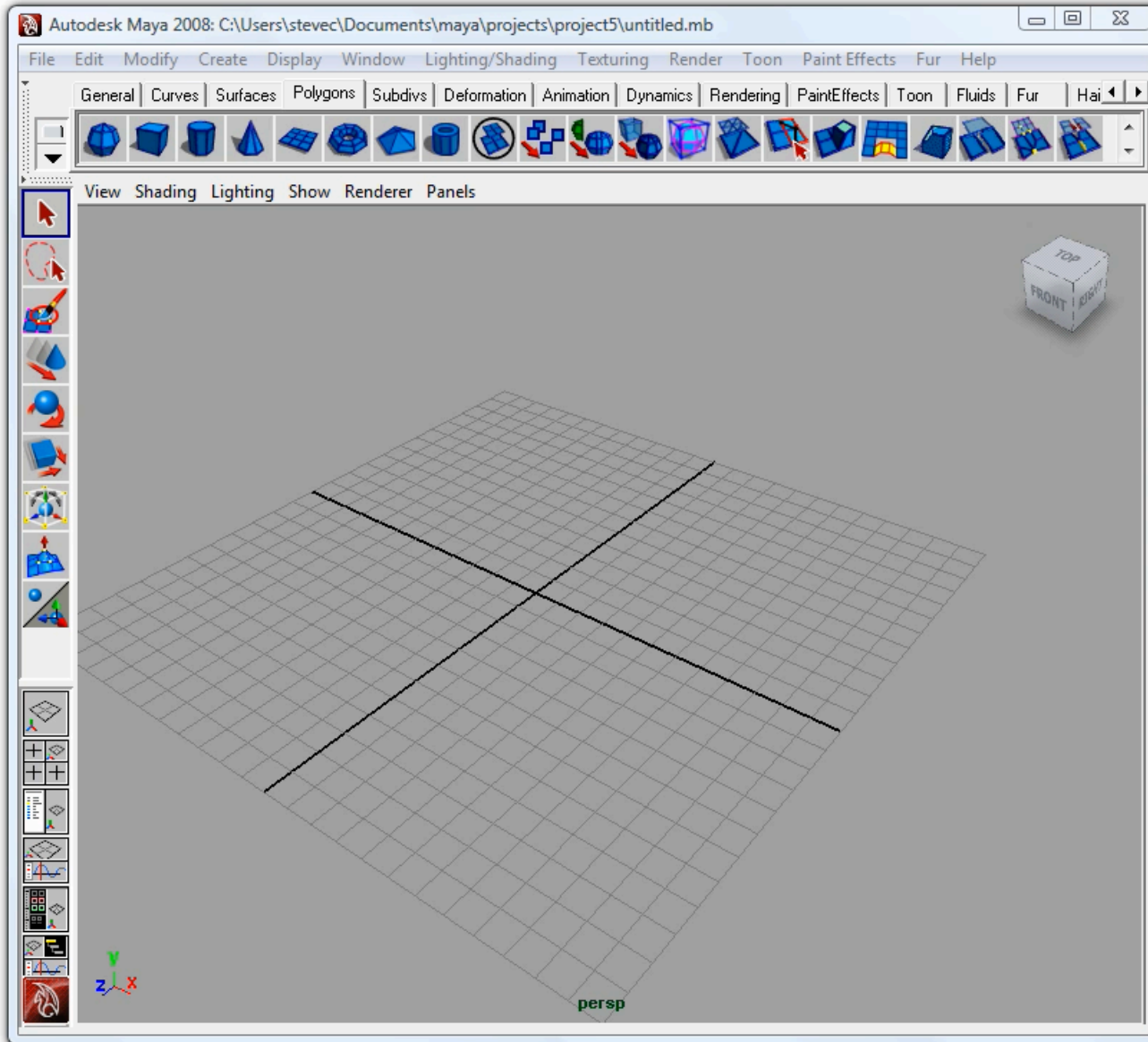
[Callahan et al., IPAW 2008]



NYU

TANDON SCHOOL
OF ENGINEERING

Provenance Plugin for Autodesk Maya



Vizier: Provenance + Notebooks



Vizier

[Features](#) [Video Tour](#) [More Info](#) [Docs](#)

[Download](#)

Data-Centric Notebooks

Vizier is a notebook that puts your data front-and-center.

Whether you prefer to use spreadsheets, notebook scripting, or databases, Vizier makes it easy to to **explore** the data to find out what you have, **validate** that the data makes sense, and **transform** it to fix bugs and mold it into a form your tools can use.

There is a [lot of hate for some popular notebooks](#). Unlike most popular notebooks, Vizier is [multi-lingual](#) and [multi-modal](#), letting you edit your data through the best interface for what you're trying to do. On top of that, it tracks [provenance](#) of your data and automatically [versions](#) your workflows. Vizier also uses dependency analysis to make sure you're never looking at stale outputs.

[Screenshots](#)

[Video Demo](#)

[Learn More](#)

[Install](#)

<https://vizierdb.info/>



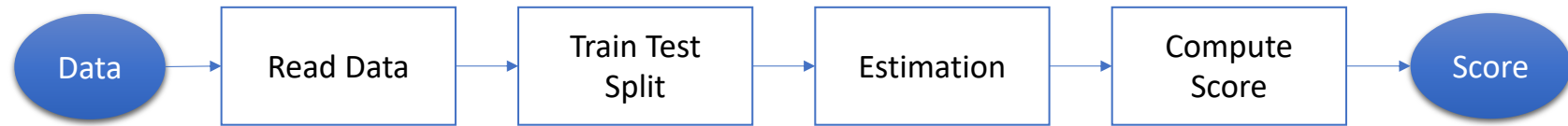
NYU

TANDON SCHOOL
OF ENGINEERING

[Glavic et al., ACM SIGMOD 2019]



Debugging Data Science Pipelines



$P = \{\text{Data, Library, Estimator}\}$
 $U_{\text{data}} = \{\text{Iris, Digits, Images}\}$
 $U_{\text{library}} = \{1.0, 2.0\}$
 $U_{\text{estimator}} = \{\text{Logistic regression, Decision tree, Gradient boosting}\}$
 $E = \text{score} > 0.6$

Instance	Data	Library	Estimator	Score	Evaluation
CP ₁	Iris	1.0	Logistic regression	0.9	Succeed
CP ₂	Digits	1.0	Decision tree	0.8	Succeed
CP ₃	Iris	2.0	Gradient boosting	0.2	Fail
CP ₄	Digits	2.0	Gradient boosting	0.3	Fail
CP ₅	Iris	1.0	Decision tree	0.7	Succeed
CP ₅	Images	1.0	Gradient boosting	0.9	Succeed

- Analyze provenance and explore parameter space to identify root causes



NYU

TANDON SCHOOL OF ENGINEERING

[Lourenço et al., ACM SIGMOD DEEM 2019]



VISUALIZATION
IMAGING AND
DATA ANALYSIS
CENTER

Conclusions

- Provenance and reproducibility are necessary for data science
 - Enable data scientists (and enthusiasts) trust and build on their own work
 - Helps community trust and build on previous work
- Creating reproducible results is not hard: there are tools that help, and best practices too
- Full reproducibility is not always possible
 - E.g., proprietary data and software, special hardware, data that is too large
- But some reproducibility is!
 - Parts of an experiment can be made available and reproduced
- Provenance for explainability and debugging (ongoing research)

Practice reproducibility – it is good for you!



NYU

TANDON SCHOOL
OF ENGINEERING



Acknowledgments

- VisTrails and ReproZip teams
- Funding: Google, National Science Foundation, Moore-Sloan Data Science Environment at NYU, and DARPA.



ALFRED P. SLOAN
FOUNDATION



NYU

TANDON SCHOOL
OF ENGINEERING



謝謝
고맙습니다
Merci
Thank you
Obrigada
благодаря
Kiitos
धन्यवाद
Tack
Danke
Ευχαριστω
Bedankt



NYU

TANDON SCHOOL
OF ENGINEERING

