Let’s make Data Science more like Science

Juliana Freire
Visualization, Imaging and Data Analysis Center (VIDA)
Computer Science & Engineering
Center for Data Science (CDS)

NYU | TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Data Science: A New Scientific Discipline

- What is Data Science?

A new paradigm for research and discovery, integrating approaches from computer science, statistics, applied mathematics, visualization and communication, and many application domains. Data science seeks to extract knowledge and insight from datasets that are often large and/or messy. Innovations in the ***methods for analyzing, visualizing, and interpreting data, and collaborating around data with diverse stakeholders****, have become key to data-intensive discovery in nearly every field.*

https://academicdatascience.org/data-science

- How can we create an environment to facilitate data science and maximize its impact?

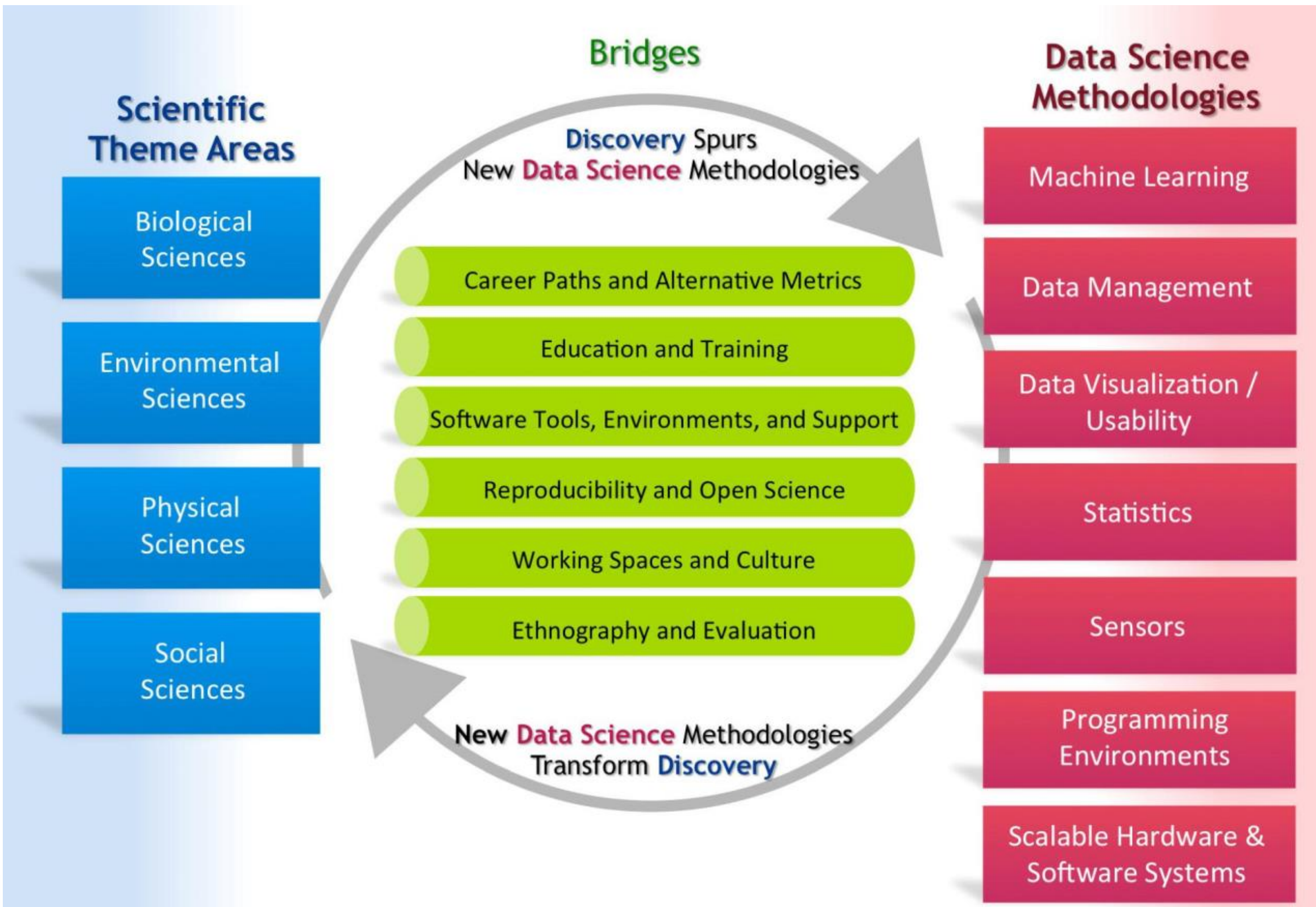
MSDSE
Moore - Sloan
Data Science Environments

http://msdse.org/

Moore Sloan Data Science Environment

How to foster sustainable adoption of data-intensive discovery?

Establish a virtuous cycle: advances in data science methodologies enable advances in discovery, which stimulate further advances in methodologies.



Data Science: A New Scientific Discipline

- Many initiatives and centers, different approaches

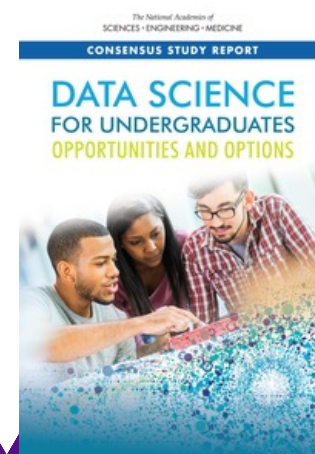
https://academicdatascience.org/

- Educational programs at all levels

NATIONAL ACADEMIES Sciences Engineering Medicine

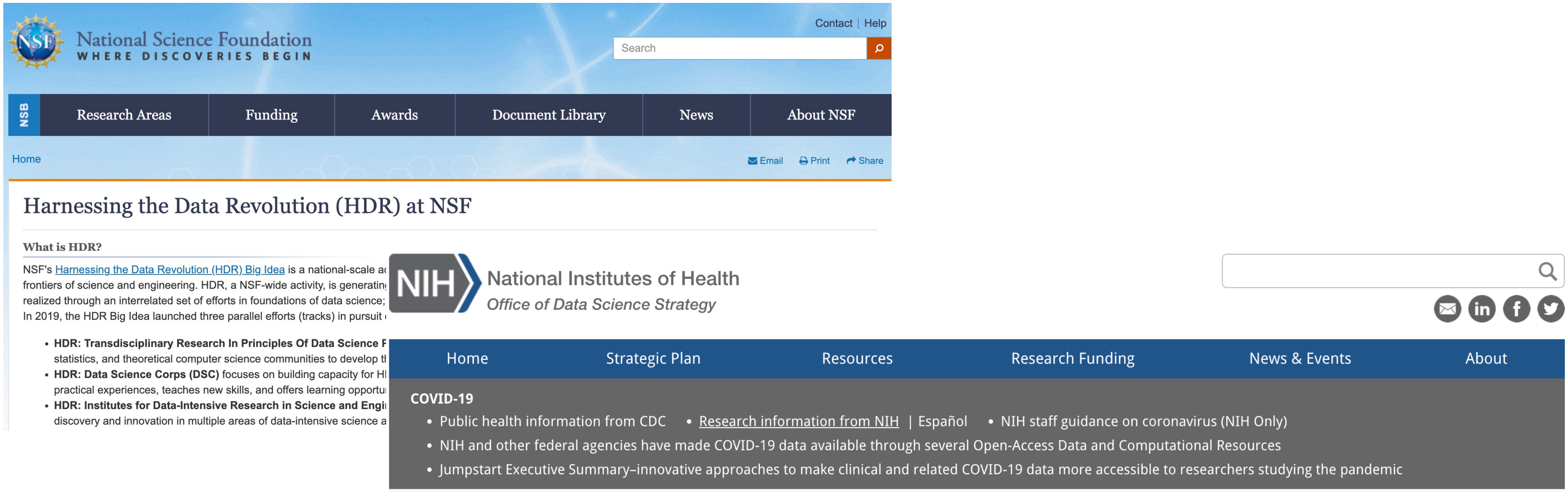
About Us Events Our Work Publications

Foundations of Data Science for Students in Grades K-12: A Workshop



Data Science: A New Scientific Discipline

- Research funding



- Accelerated scientific progress and discoveries – just check!

Google Scholar: (data science method) + domain

"deep learning" physics

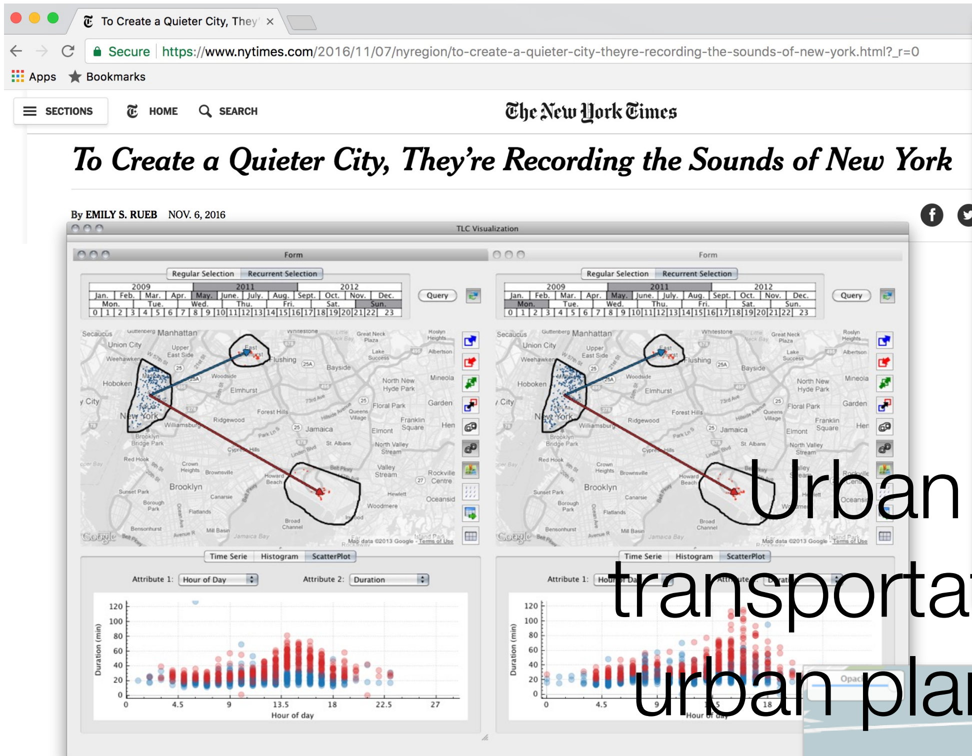
"visualization" biology

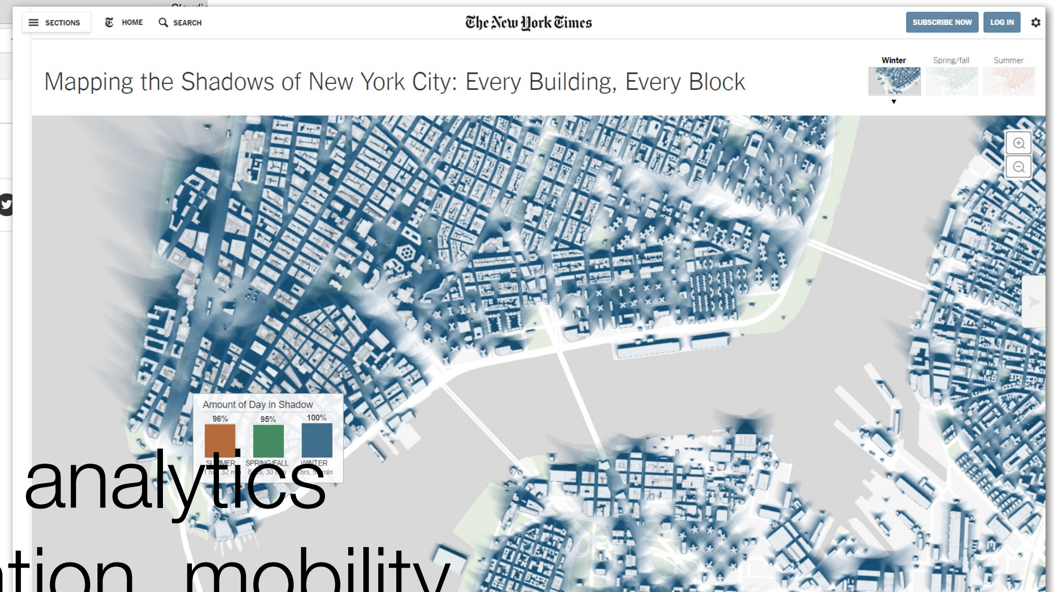
"data management" urban

NYU TANDON SCHOOL OF ENGINEERING

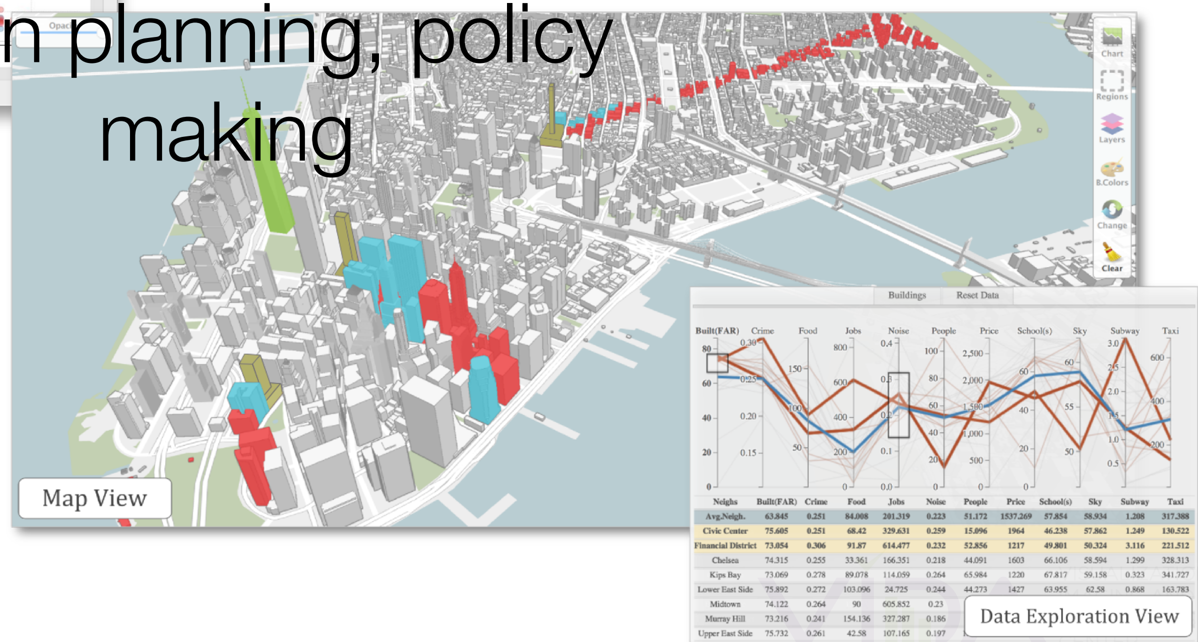
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Data Science fueling a Virtuous Cycle





Urban analytics
transportation, mobility,
urban planning, policy
making




The Economist

Urban planning

Listen to the music of the traffic in the city

Places, like people, have pulses—if only you know how to measure them

Oct 22nd 2016 | From the print edition

THE Rockefeller Centre sprawls across 89,000 square metres of midtown Manhattan. Curiously, Alcatraz, in San Francisco Bay, the island home of America's most famous former prison (see picture), has exactly the same area. That coincidence aside, few might imagine the manicured roof gardens and art deco office buildings of the one have much in common with the brutal crags and blockhouses of the other. But they do. For research by Claudio Silva of New York University and his colleagues suggests that the two have a striking resemblance when it comes to the daily ebb and flow of tourists, as judged from the level of activity on Flickr, a photo-hosting site. Dr Silva thinks the peaks and troughs of Flickr activity that his research has discovered in this and other cases are a measure of an area's "urban pulse". If so, the Rockefeller Centre and Alcatraz share

Like the Rockefeller Centre, but different

Data Science fueling a Virtuous Cycle

- Spatio-temporal data management
- Visual Analytics
- Machine Learning
- Computational Topology
- Data Discovery
- Data Cleaning

Publications

Urban Rhapsody: Large-scale exploration of urban soundscapes. CGF 2022

SPADE: GPU-Powered Spatial Database Engine for Commodity Hardware. IEEE ICDE 2022

A GPU-friendly Geometric Data Model and Algebra for Spatial Queries. ACM SIGMOD 2020

A GPU-based index to support interactive spatio-temporal queries over historical data. IEEE ICDE 2016

Auctus: A Dataset Search Engine for Data Discovery and Augmentation. PVLDB 2021

Interactive Visual Exploration of Spatio-Temporal Urban Data Sets using Urbane. ACM SIGMOD 2018

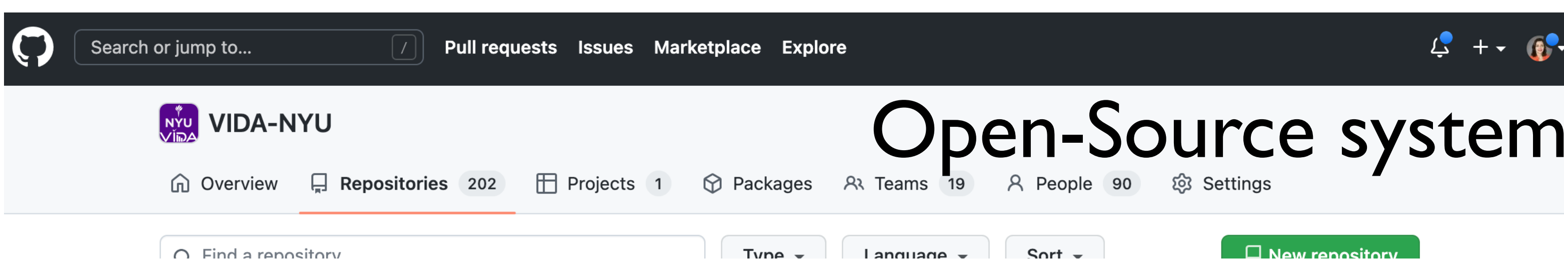
Exploring What not to Clean in Urban Data: A Study Using New York City Taxi Trips. IEEE DEB 2016

Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets. ACM SIGMOD 2016

Exploring Traffic Dynamics in Urban Environments Using Vector-Valued Functions. CG&A 2015

Using Topological Analysis to Support Event-Guided Exploration in Urban Data. IEEE TVCG 2014

Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. IEEE YVCG 2013 . . .

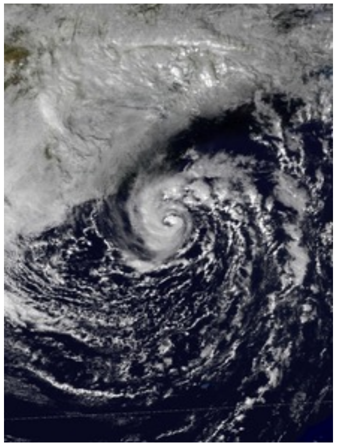


Open-Source systems

NYU

SUALIZATION
AGING AND
ATA ANALYSIS
ENTER

Data-Driven Exploration



- The perfect storm:

  Computing is free

  Storage is free

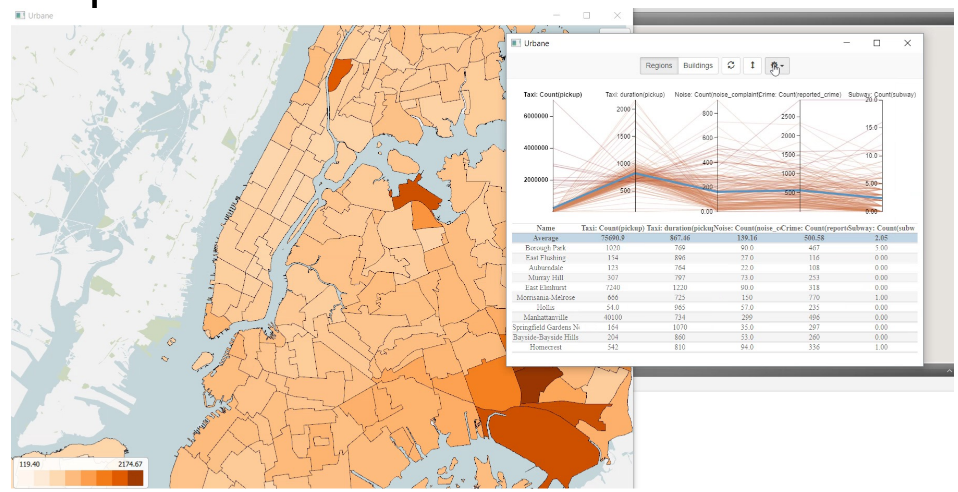
  Data are abundant

- Challenge: The bottlenecks lie with people
  - Complex computational processes are required to extract insight -- hard to assemble and require expertise in a wide range of topics and tools
  - It is difficult for domain experts to explore data
- Solution: Democratize Data Science!

Democratizing Data Science

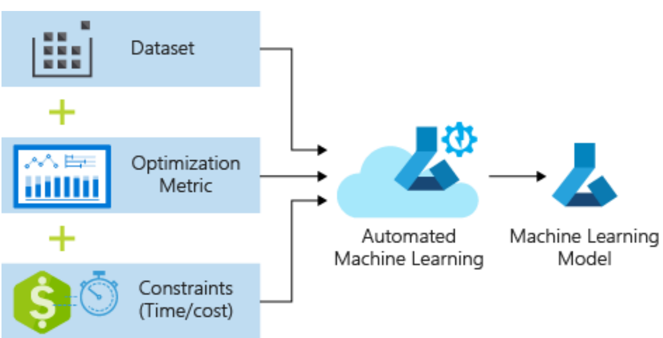
Open-source software

Specialized interactive tools

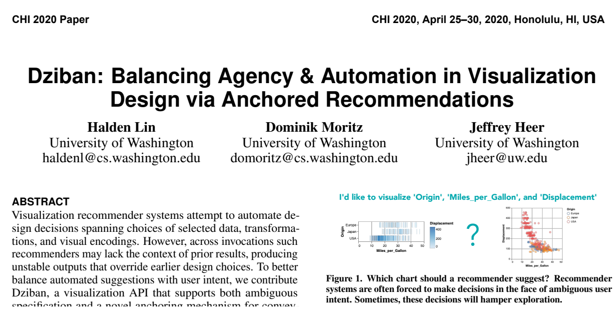


[Ferreira et al., IEEE VAST 2015;
Doraiswamy et al., ACM SIGMOD 2018]

Automated Data Science



AutoML

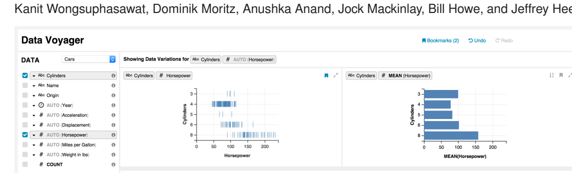

CHI 2020 Paper CHI 2020, April 25–30, 2020, Honolulu, HI, USA

Dziban: Balancing Agency & Automation in Visualization Design via Anchored Recommendations

Halden Lin, University of Washington, haldenl@cs.washington.edu
Dominik Moritz, University of Washington, domoritz@cs.washington.edu
Jeffrey Heer, University of Washington, jheer@uw.edu

ABSTRACT
Visualization recommender systems attempt to automate design decisions spanning choices of selected data, transformations, and visual encodings. However, across invocations such recommenders may lack the context of prior results, producing unstable outputs that override earlier design choices. To better balance automated suggestions with user intent, we contribute Dziban, a visualization API that supports both ambiguous

Figure 1. Which chart should a recommender suggest? Recommender systems are often forced to make decisions in the face of ambiguous user intent. Sometimes, these decisions will hamper exploration.


Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations

Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer

Recommender Systems for Visualization

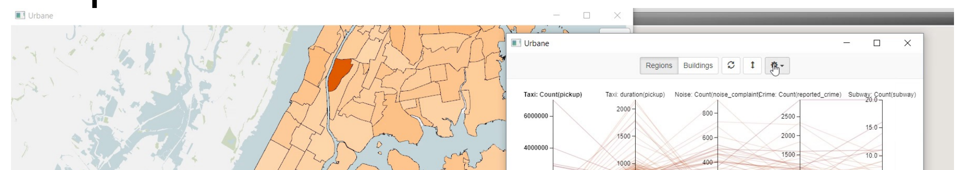
NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

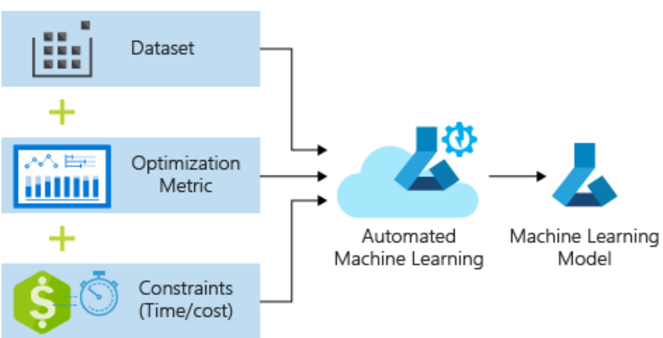
Democratizing Data Science

Open-source software

Specialized interactive tools

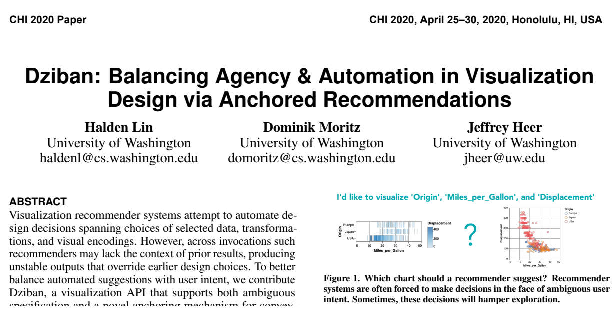


It is becoming increasingly easier to derive results



AutoML

[Ferreira et al., IEEE VAST 2015;
Doraiswamy et al., ACM SIGMOD 2018]



CHI 2020 Paper CHI 2020, April 25–30, 2020, Honolulu, HI, USA

Dziban: Balancing Agency & Automation in Visualization Design via Anchored Recommendations

Halden Lin, University of Washington, haldenl@cs.washington.edu
Dominik Moritz, University of Washington, domoritz@cs.washington.edu
Jeffrey Heer, University of Washington, jheer@uw.edu

ABSTRACT

Visualization recommender systems attempt to automate design decisions spanning choices of selected data, transformations, and visual encodings. However, across invocations such recommenders may lack the context of prior results, producing unstable outputs that override earlier design choices. To better balance automated suggestions with user intent, we contribute Dziban, a visualization API that supports both ambiguous specification and a novel anchoring mechanism for convey-

I'd like to visualize 'Origin', 'Miles_per_Gallon', and 'Displacement'

Figure 1. Which chart should a recommender suggest? Recommender systems are often forced to make decisions in the face of ambiguous user intent. Sometimes, these decisions will hamper exploration.

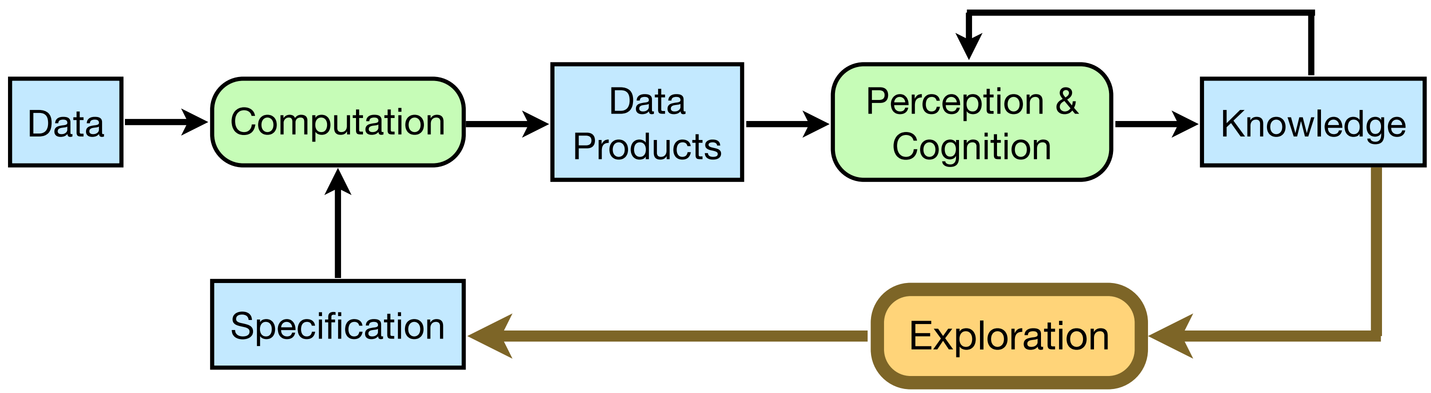
Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations

Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer

Recommender Systems for Visualization

How can Data Science go wrong?

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



[Modified from Van Wijk, Vis 2005]

- After many steps…
  - It is easy to get lost and not remember how a result was derived
  - Did I make any mistakes?
  - Were there any problems with the data, code, computational environment?
  - Results can be hard to understand, interpret and trust

Human Mistakes

American Economic Review: Papers & Proceedings 100 (May 2010): 573–578
http://www.aeaweb.org/articles.php?doi=10.1257/aer.100.2.573

Growth in a Time of Debt

By Carmen M. Reinhart and Kenneth S. Rogoff*

In this paper, we exploit a new multi-country historical dataset on public (government) debt to search for a systemic relationship between high public debt levels, growth and inflation.[1] Our main result is that whereas the link between growth and debt seems relatively weak at "normal" debt levels, median growth rates for countries with public debt over roughly 90 percent of GDP are about one percent lower than otherwise; average (mean) growth rates are several percent lower. Surprisingly, the relationship between public debt and growth is remarkably similar across emerging markets and advanced economies. This is not the case for inflation. We find no systematic relationship between high debt levels and inflation for advanced economies as a group (albeit with individual country exceptions including the United States). By contrast, in emerging market countries, high public debt levels coincide with higher inflation.

Our topic would seem to be a timely one. Public debt has been soaring in the wake of the recent global financial maelstrom, especially in the epicenter countries.
prising, given the expe

especially against the back
ulations and rising social i
sharply elevated public deb
ageable policy challenge?

Our approach here is c
taking advantage of a br
dataset on public debt (in
government debt) first pres
Reinhart and Kenneth S. R
Prior to this dataset, it was
to get more than two or th
lic debt data even for many
virtually impossible for mo
Our results incorporate da
spanning about 200 years.
data incorporate over 3,700
covering a wide range of po
tutions, exchange rate and
ments, and historic circums

We also employ more re
debt, including debt owed t
and by private entities. For emerging markets,
we find that there exists a significantly more

	B	C	I	J	K	L	M
2			Real GDP growth				
3			Debt/GDP				
4	Country	Coverage	30 or less	30 to 60	60 to 90	90 or above	30 or less
26			3.7	3.0	3.5	1.7	5.5
27	Minimum		1.6	0.3	1.3	-1.8	0.8
28	Maximum		5.4	4.9	10.2	3.6	13.3
29							
30	US	1946-2009	n.a.	3.4	3.3	-2.0	n.a.
31	UK	1946-2009	n.a.	2.4	2.5	2.4	n.a.
32	Sweden	1946-2009	3.6	2.9	2.7	n.a.	6.3
33	Spain	1946-2009	1.5	3.4	4.2	n.a.	9.9
34	Portugal	1952-2009	4.8	2.5	0.3	n.a.	7.9
35	New Zealand	1948-2009	2.5	2.9	3.9	-7.9	2.6
36	Netherlands	1956-2009	4.1	2.7	1.1	n.a.	6.4
37	Norway	1947-2009	3.4	5.1	n.a.	n.a.	5.4
38	Japan	1946-2009	7.0	4.0	1.0	0.7	7.0
39	Italy	1951-2009	5.4	2.1	1.8	1.0	5.6
40	Ireland	1948-2009	4.4	4.5	4.0	2.4	2.9
41	Greece	1970-2009	4.0	0.3	2.7	2.9	13.3
42	Germany	1946-2009	3.9	0.9	n.a.	n.a.	3.2
43	France	1949-2009	4.9	2.7	3.0	n.a.	5.2
44	Finland	1946-2009	3.8	2.4	5.5	n.a.	7.0
45	Denmark	1950-2009	3.5	1.7	2.4	n.a.	5.6
46	Canada	1951-2009	1.9	3.6	4.1	n.a.	2.2
47	Belgium	1947-2009	n.a.	4.2	3.1	2.6	n.a.
48	Austria	1948-2009	5.2	3.3	-3.8	n.a.	5.7
49	Australia	1951-2009	3.2	4.9	4.0	n.a.	5.9
50							
51			4.1	2.8	2.8	=AVERAGE(L30:L44)	

countries with debt over 90% of their gross domestic product (GDP) have a negative growth rate

Unexpected Problems: Bugs in Code

ars TECHNICA

OUT OF SORTS —

Researchers find bug in Python script may have affected hundreds of studies

https://arstechnica.com/information-technology/2019/10/chemists-discover-cross-platform-python-scripts-not-so-cross-platform/

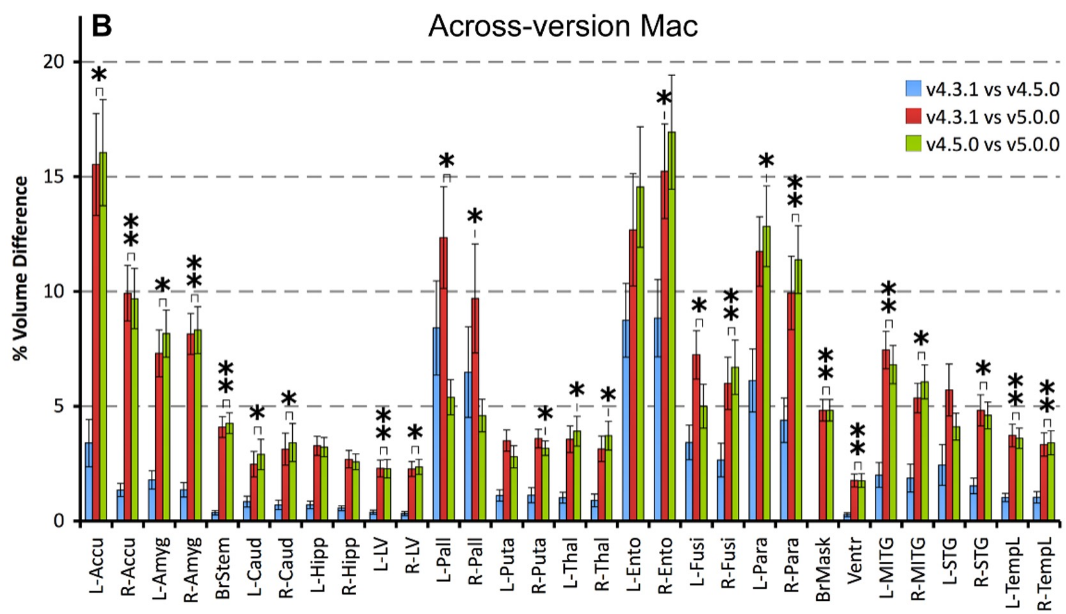
> "The scripts [...] were found to return correct results on macOS Mavericks and Windows 10. But on macOS Mojave and Ubuntu, the results were off by nearly a full percent."

- Scripts used a specific library, *glob*, which returns a different sorted order depending on the OS
- It's not easy to tell this is happening either!

Unexpected Problems: Software

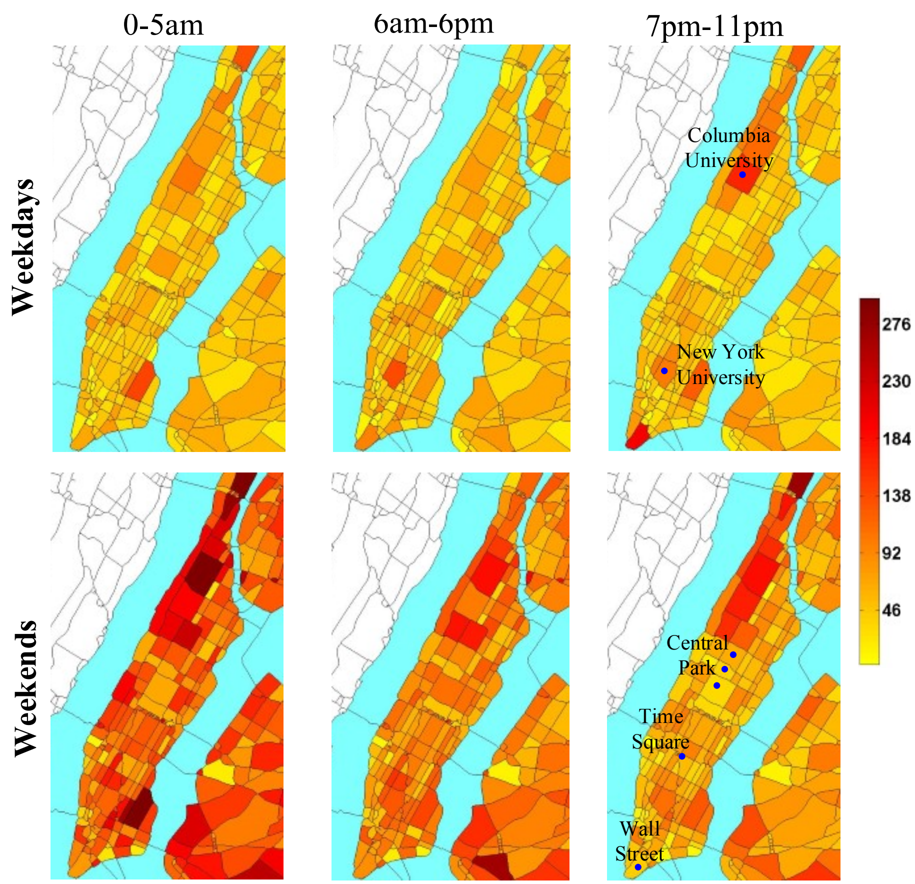
- *The Effects of FreeSurfer Version, Workstation Type, and Macintosh Operating System Version on Anatomical Volume and Cortical Thickness Measurements*. PLOS ONE, June 1, 2012

Significant differences in result of neuro analysis depending on version of software, hardware, and operating system



Unexpected Problems: Data

- Bad or incorrectly used data can lead to incorrect conclusions



Using 311 complaints as a proxy noise sensor [Zheng et al., 2014]

But there are fewer noise complaints in areas with a higher percentage of residents that belong to minorities [Minkoff, 2015]

Lack of Robustness

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht* [1] **Rebecca Roelofs** [1] **Ludwig Schmidt** [1] **Vaishaal Shankar** [1]

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

Conventional wisdom suggests that such drops arise because

CIFAR-10

Orig. Rank	Model	Orig. Accuracy	New Accuracy	Gap	New Rank	Δ Rank
1	`autoaug_pyramid_net_tf`	98.4 [98.1, 98.6]	95.5 [94.5, 96.4]	2.9	1	0
6	`shake_shake_64d_cutout`	97.1 [96.8, 97.4]	93.0 [91.8, 94.1]	4.1	5	1
16	`wide_resnet_28_10`	95.9 [95.5, 96.3]	89.7 [88.3, 91.0]	6.2	14	2
23	`resnet_basic_110`	93.5 [93.0, 93.9]	85.2 [83.5, 86.7]	8.3	24	-1
27	`vgg_15_BN_64`	93.0 [92.5, 93.5]	84.9 [83.2, 86.4]	8.1	27	0
30	`cudaconvnet`	88.5 [87.9, 89.2]	77.5 [75.7, 79.3]	11.0	30	0
31	`random_features_256k_aug`	85.6 [84.9, 86.3]	73.1 [71.1, 75.1]	12.5	31	0

ImageNet Top-1

Orig. Rank	Model	Orig. Accuracy	New Accuracy	Gap	New Rank	Δ Rank
1	`pnasnet_large_tf`	82.9 [82.5, 83.2]	72.2 [71.3, 73.1]	10.7	3	-2
4	`nasnetalarge`	82.5 [82.2, 82.8]	72.2 [71.3, 73.1]	10.3	1	3
21	`resnet152`	78.3 [77.9, 78.7]	67.0 [66.1, 67.9]	11.3	21	0
23	`inception_v3_tf`	78.0 [77.6, 78.3]	66.1 [65.1, 67.0]	11.9	24	-1
30	`densenet161`	77.1 [76.8, 77.5]	65.3 [64.4, 66.2]	11.8	30	0
43	`vgg19_bn`	74.2 [73.8, 74.6]	61.9 [60.9, 62.8]	12.3	44	-1
64	`alexnet`	56.5 [56.1, 57.0]	44.0 [43.0, 45.0]	12.5	64	0
65	`fv_64k`	35.1 [34.7, 35.5]	24.1 [23.2, 24.9]	11.0	65	0

Table 1. Model accuracies on the original CIFAR-10 test set, the original ImageNet validation set, and our new test sets. Δ Rank is the relative difference in the ranking from the original test set to the new test set in the full ordering of all models (see Appendices C.3.3 and D.4.4). For example, $\Delta\text{Rank} = -2$ means that a model dropped by two places on the new test set compared to the original test set. The confidence intervals are 95% Clopper-Pearson intervals. Due to space constraints, references for the models can be found in Appendices C.3.2 and D.4.3.

Lack of Robustness

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht*[1] Rebecca Roelofs[1] Ludwig Schmidt[1] Vaishaal Shankar[1]

Abstract

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense
ing the danger o
test sets. By clo
creation process
classification m
evaluate a broa
racy drops of 3
– 14% on Imag
on the original
on the new test
accuracy drops
by the models'
"harder" image
test sets.

Conventional wisdom suggests that such drops arise because

CIFAR-10		
Orig.	New	Rank
		0
		1
		2
		-1
		0
		0
		0
		Rank
		-2
		3
		0
		-1
		0
		-1
		0
		0

Δ Rank is the
lices C.3.3 and
al test set. The
in Appendices

- Built new test sets for the CIFAR-10 and ImageNet datasets widely in machine learning research
- Assessed to what extent classification models generalize to new data
- Observed significance accuracy drops: 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet

Data Science needs Robustness

When a computational result becomes the basis of policy or may impact human well-being, reliability becomes more than an academic question and has real consequences

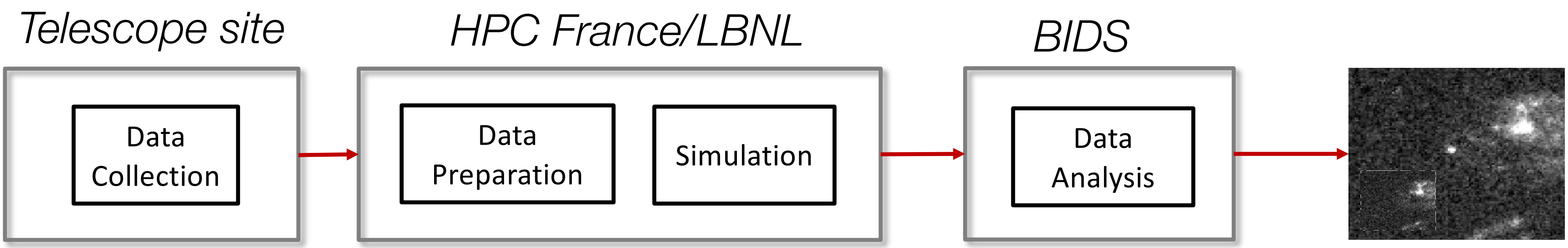
Democratizing Trust and Robustness

- We should learn from science and the scientific method – build trust through replication studies and uncertainty quantification

Repeated findings of consistent results tend to confirm the veracity of an original scientific conclusion, and, by the same token, repeated failures to confirm raise doubts

- Need systematic debugging and testing for data and computations, and explanations for results
  - Some initial steps: explainable AI
- Need to explain *general* computations

Democratizing Trust and Robustness



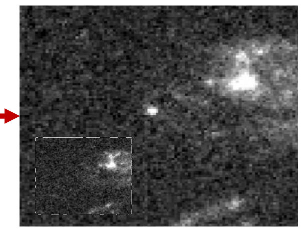
Is the *feature* in the image a discovery or a bug?

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Democratizing Trust and Robustness

Telescope site: Data Collection → *HPC France/LBNL*: Data Preparation, Simulation → *BIDS*: Data Analysis →



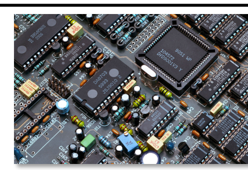
Data

Code/Scripts/Workflows

Libraries

Programming Languages

Operating System



Hardware

Many dependencies – it is difficult to identify the root cause

But we can *experiment – formulate and test hypotheses*

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Provenance and Reproducibility

- Provenance and reproducibility are necessary to verify and build trust in results and to debug data science pipelines

- Opportunity: Machine-assisted debugging through the the *automation of replication* studies [Lourenço et al., ACM SIGMOD 2021]
  - Vary/perturb data,
  - Explore parameter spaces,
  - Compare different methods,
  - Run experiment on different operating systems
  - Test domain specific constraints to flag potential problems
  - …

Call to Action

- Let’s do **reproducible** research

  Need investment in infrastructure

  to support reproducibility

https://sites.nationalacademies.org/sites/reproducibility-in-science

My dream: reproducibility as a standard feature of computational tools and environments

- Let’s **democratize trust and robustness** for data science
  - Many new and challenging research problems!

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Acknowledgments

NSF

Google

DARPA

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION

با تشکر

謝謝

고맙습니다

Merci

Thank you

Obrigada

благодаря

Kiitos

धन्यवाद

Tack

Danke

Ευχαριστω

Bedankt

TANDON SCHOOL OF ENGINEERING

VISUALIZATION IMAGING AND DATA ANALYSIS CENTER