Case Study: Northstar

Tim Kraska <kraska@mit.edu>

This lecture

- Design a system for Interactive Data Science
- Northstar demo
- How does Northstar work
- Problems with making Data Science more accessible and future directions

Data Science Today

Data Science Today

Case Study: A System for Democratizing Data Science

Design a system to make Data Science more accessible to a broader range of users (25min total)

PART I: Key requirements/User interface design (10min)

PART II: Implementation (10min)

PART III: Open challenges (5min)

No clicker today. Instead at the end of the class, you hand-in your final solution to:

Matt Perron <mperron@csail.mit.edu>

Case Study Part I: Design the UI

Design the UI for a system to make Data Science more accessible to a broader range of users (e.g., your parents) - 10min

A. What are key requirements? (List key requirements)

B. User interface design (Sketch a few UIs on how you envision a users would build a predictive model over her sales data) northst*r

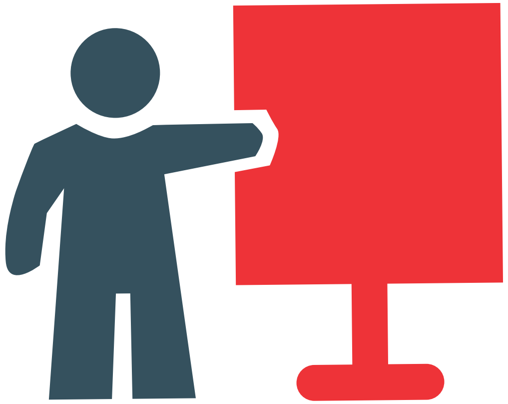
the data science platform

northst*r

the data science platform

Three Core Technical Contributions

Laax
A Novel Interface
for Everyone



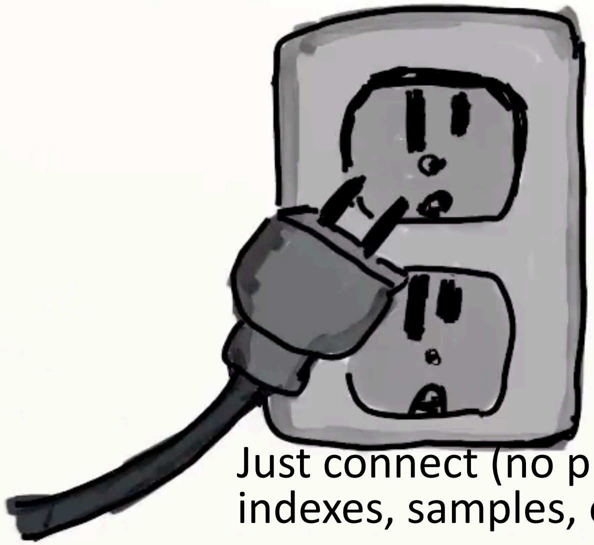
designed for data enthusiast (i.e., people with limited statistics and ML knowledge), domain experts, and data scientist alike.

[1]Laax is the successor of Vizdom, our first user interface

Key System Requirements

– Just connect (no pre-computed indexes, samples, etc.)

Key System Requirements



Just connect (no pre-computed indexes, samples, etc.)

First response <500ms

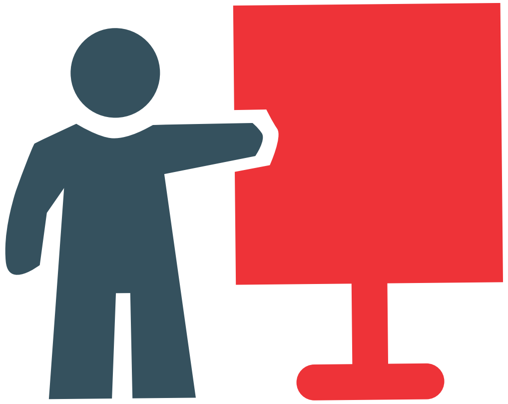
Case Study Part II: Implementation

How would you implement the system powering your interface - 10min

- What are the most important components? How does your architecture look like? – Create an architecture diagram
- How does your system deal with very large data or very compute intensive operations?
- What new techniques are needed? Which existing techniques can be used?

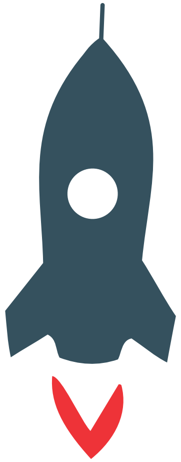
Three Core Technical Contributions

Vizdom
A Novel Interface for Everyone



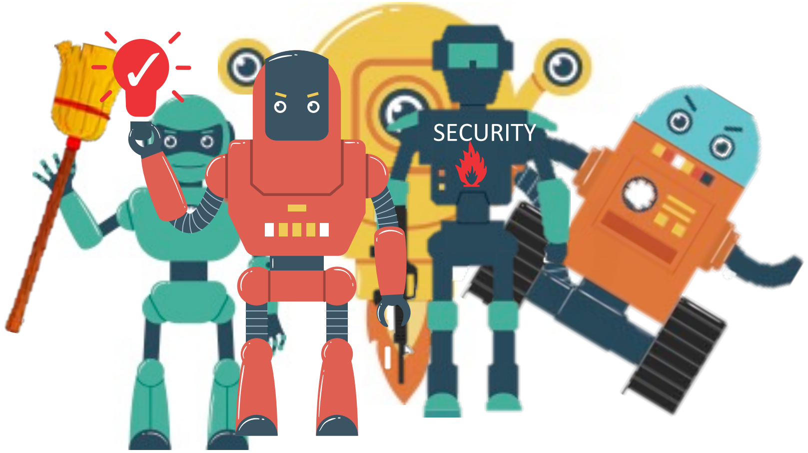
designed for data enthusiast (i.e., people with limited statistics and ML knowledge), domain experts, and data scientist alike.

IDEA
The Data Exploration Accelerator



No waiting: immediately returns visual results for all operations and progressively refines them in the background

Smart Assistance
Towards Data Science Automation



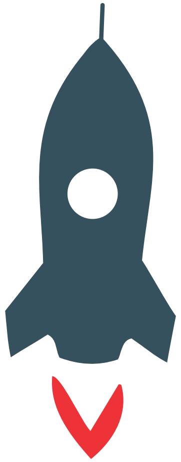
Protect users from common mistakes, point out data cleaning issues, help with building models

Three Core Technical Contributions

Vizdom
A Novel Interface for Everyone

designed for data enthusiast (i.e., people with limited statistics and ML knowledge), domain experts, and data scientist alike.

IDEA
The Data Exploration Accelerator



No waiting: immediately returns visual results for all operations and progressively refines them in the background

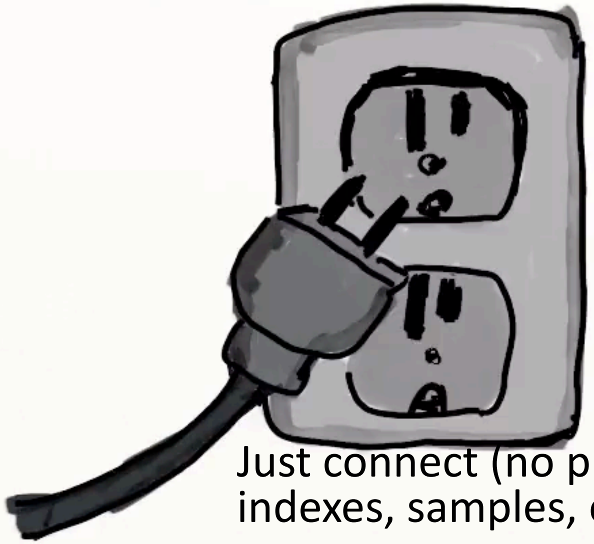
Smart Assistance
Towards Data Science Automation

Protect users from common mistakes, point out data cleaning issues, help with building models

Key System Requirements

- Just connect (no pre-computed indexes, samples, etc.)

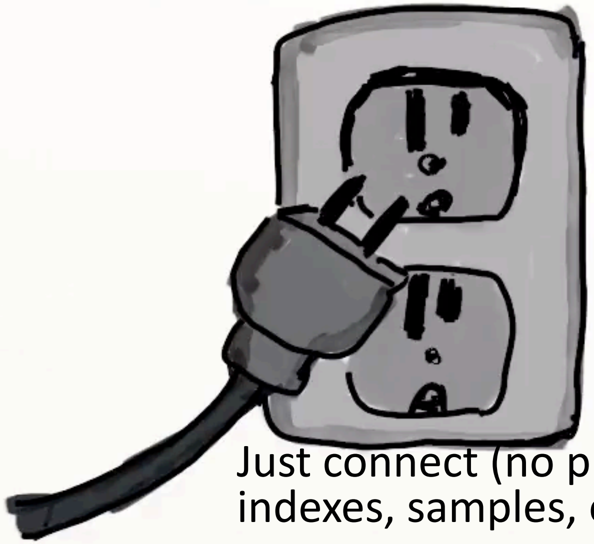
Key System Requirements



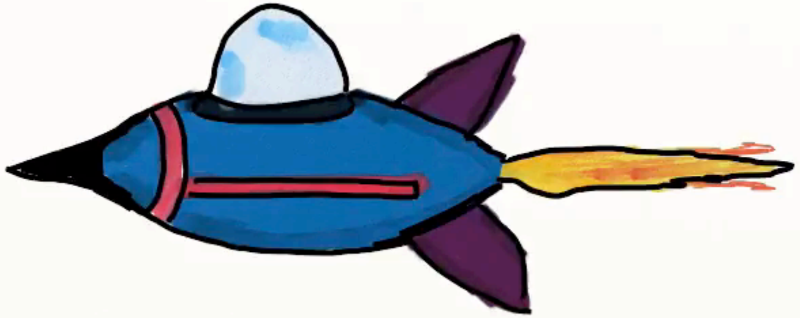
Just connect (no pre-computed indexes, samples, etc.)

First response <500ms

Key System Requirements



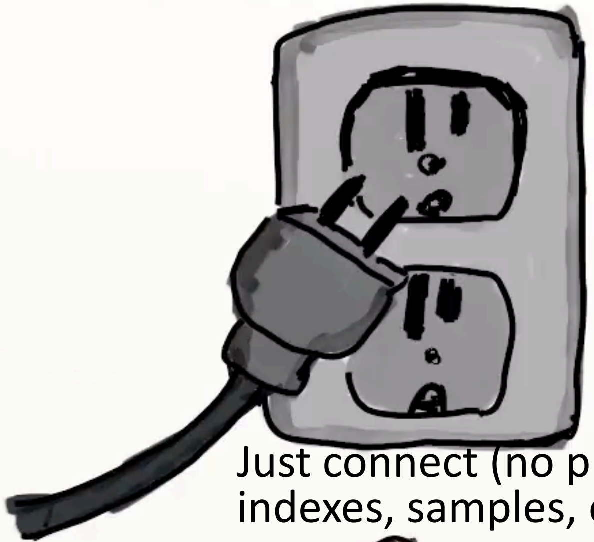
Just connect (no pre-computed indexes, samples, etc.)



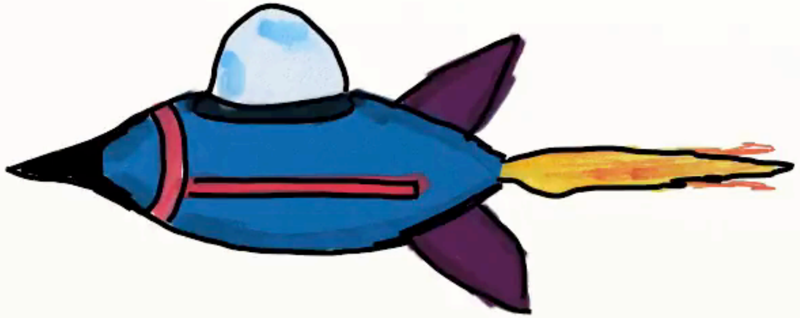
First response <500ms

See results unfold/
Progressive results

Key System Requirements



Just connect (no pre-computed indexes, samples, etc.)

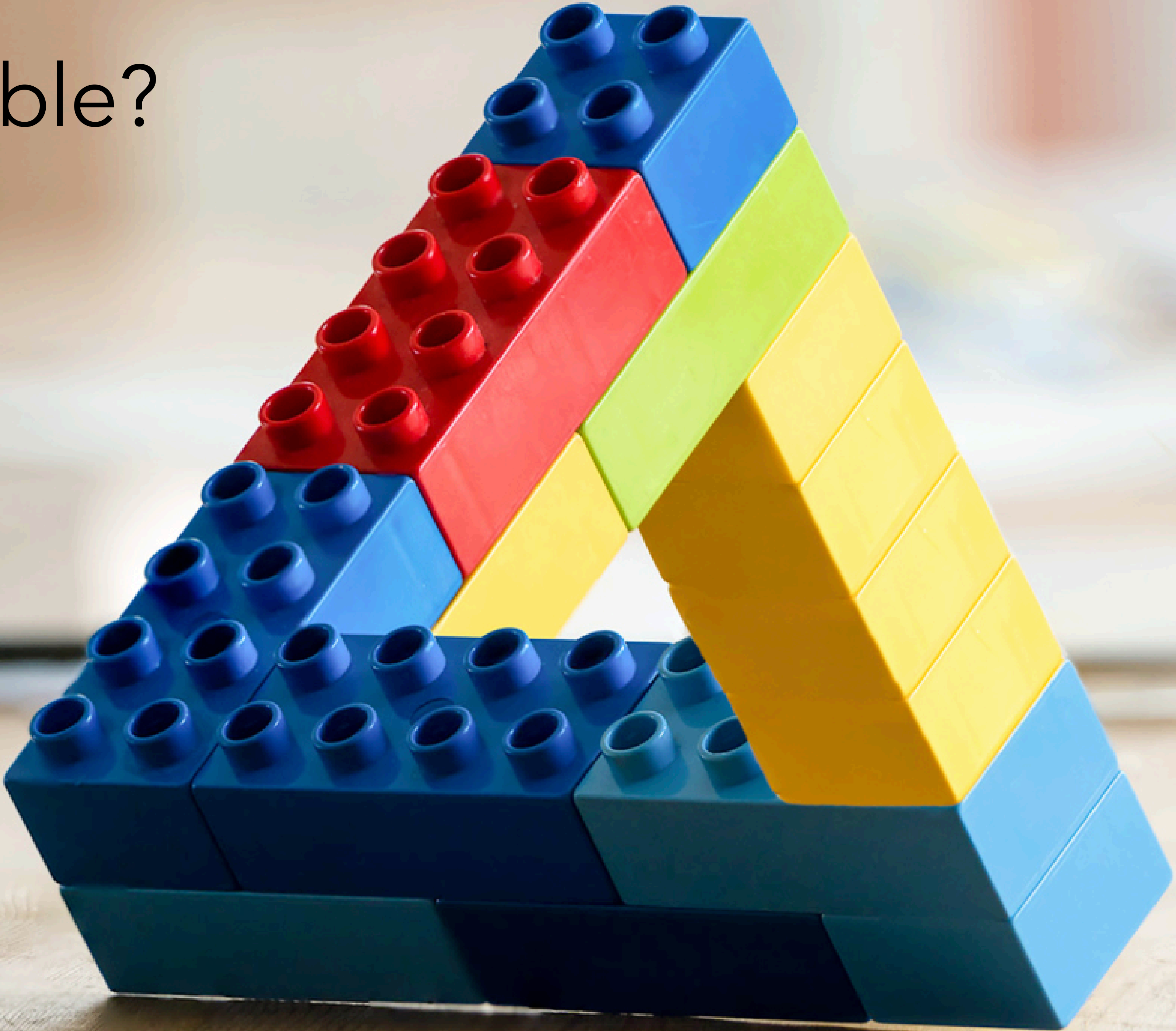


First response <500ms

See results unfold/
Progressive results

Quantify risk

Impossible?



Three Unique Opportunities

Think-Time

Three Unique Opportunities

Think-Time

Queries are built incrementally

Three Unique Opportunities

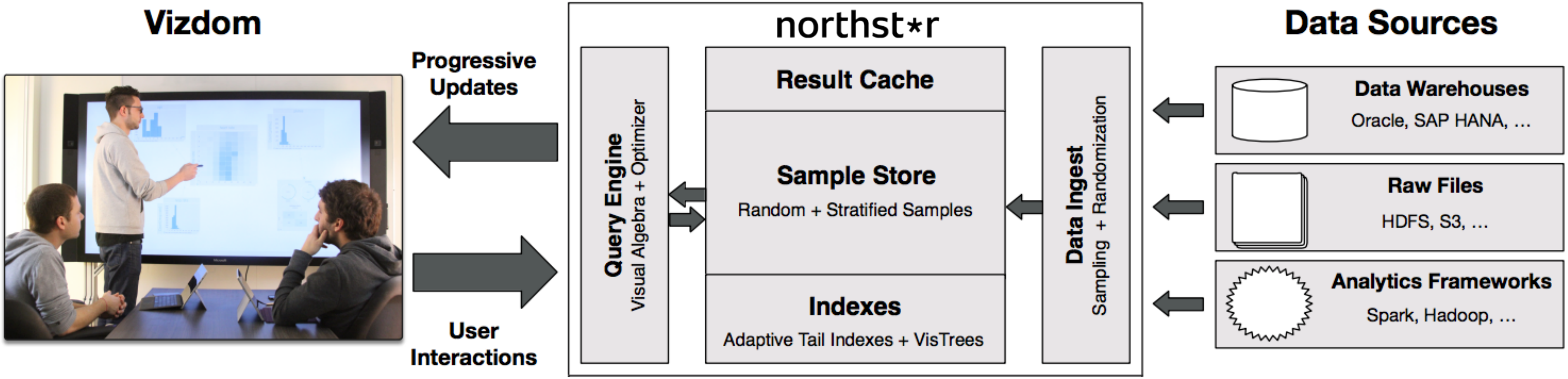
Think-Time

Many interesting **research questions** on how to take **advantage** of these **opportunities**??

Queries are built incrementally

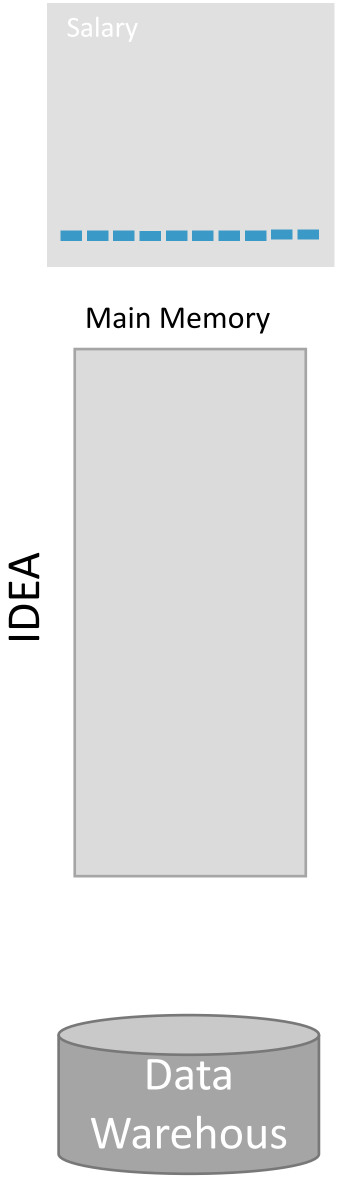
Visualizations

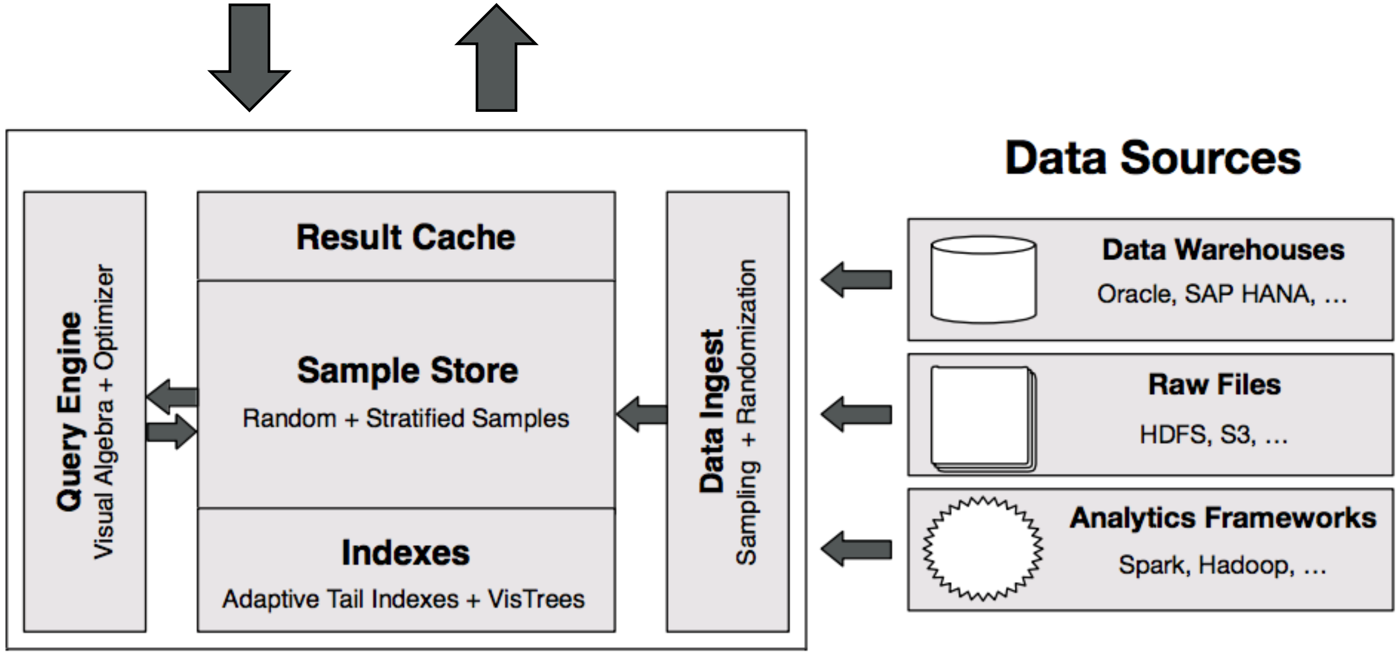
IDEA: The Interactive Data Exploration Accelerator of Northstar



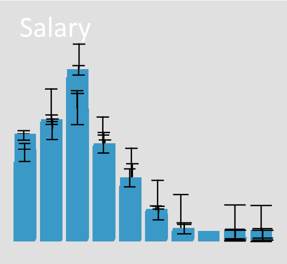
Ensures interactive latencies regardless of the operation (e.g., linking, brushing, model building), data source, and data size through our novel approximate query processing (AQP) techniques for Interactive Data Science.

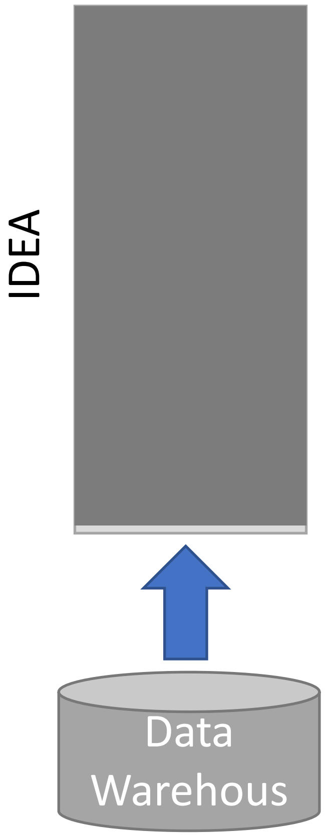
Opportunity I: **Think-Time**





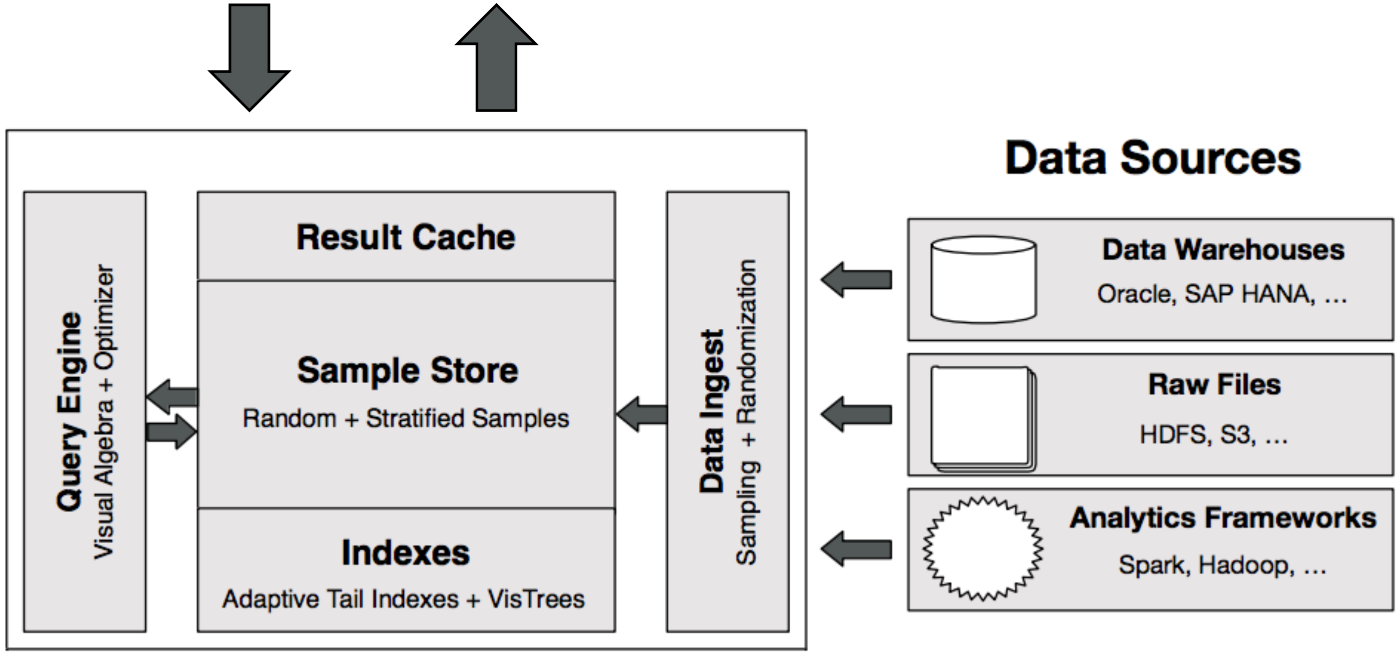
Opportunity I: **Think-Time**



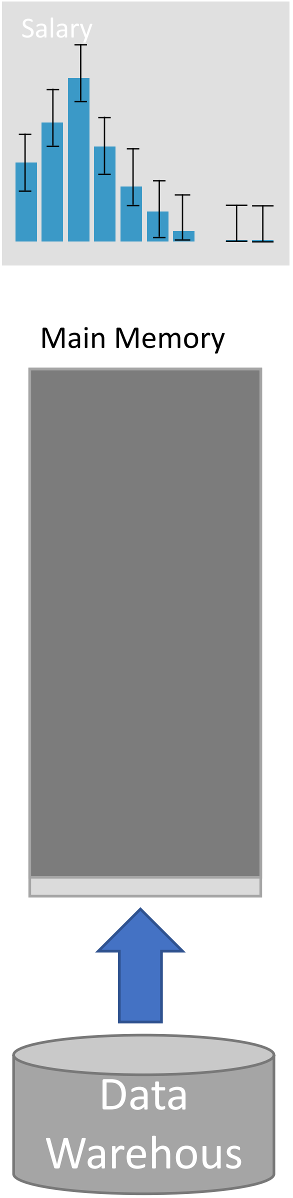


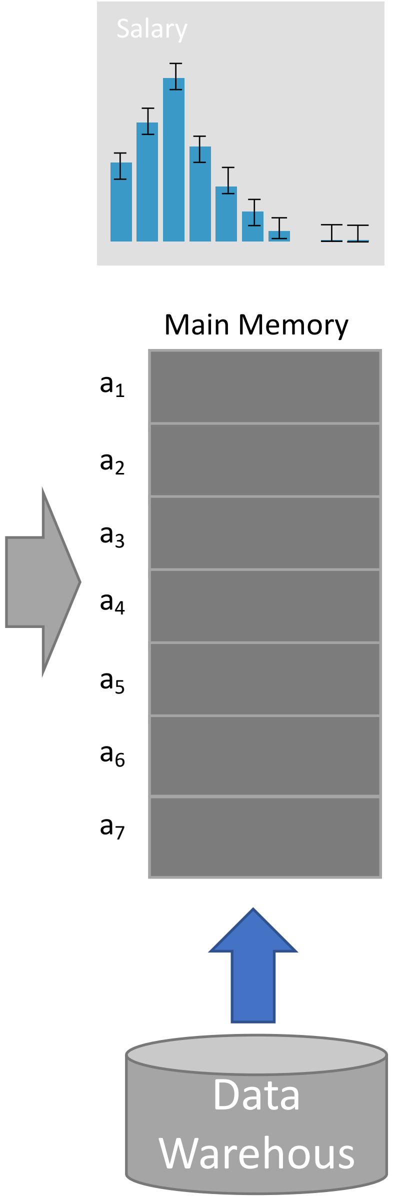
Key ideas:

- **Caching** and **online aggregation**



Opportunity I: **Think-Time**



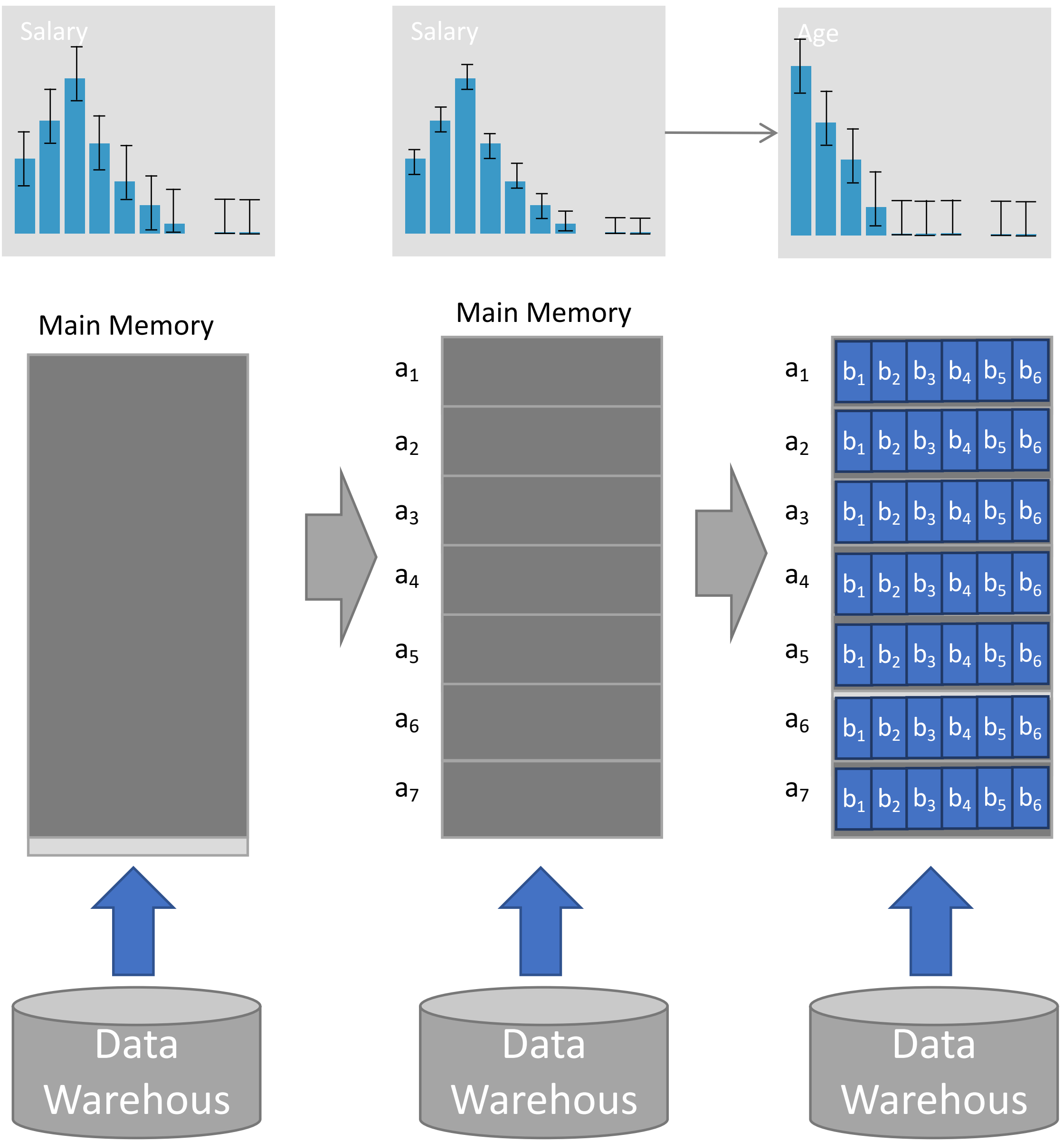


Key ideas:

- **Caching** and **online aggregation**
- If out of memory → **Stratified reservoir sampling**

Why not do database cracking: sorting might destroys the randomness for follow up operations

Opportunity I: **Think-Time**

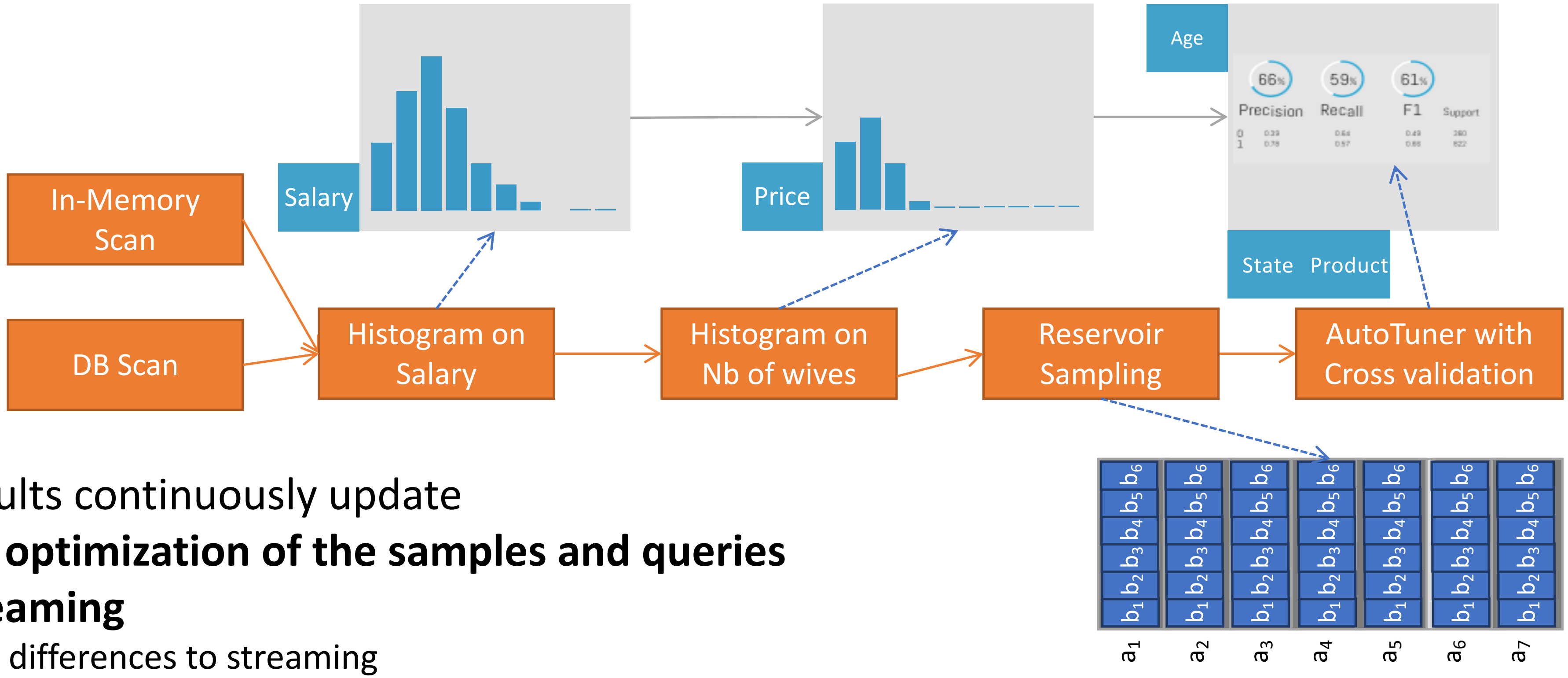


Stratification in 2 Dimensions

(attr1, attr2)

→ Much faster as we already started with a stratified sample for attr1

Requires a New Processing Model



- Pipeline results continuously update
- **Continuous optimization of the samples and queries**
- **Sample Streaming**
  - Important differences to streaming
    - Sample based and queries are added/removed all the time
    - Streams are usually infinite and data sources are “intelligent”

→ **Opens up a whole new set of research challenges. For example:**

- How do you show progress? (new version has two progress indicators and one quality indicator)
- How do you integrate UDFs and UDAs (e.g., AutoML)?
- What happens if the underlying data-source changes?
- How to push down operations?
- …

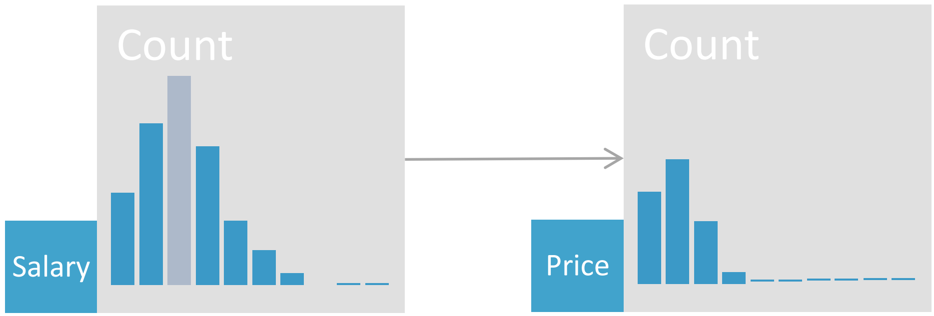
[HILDA16]

Opportunity II: **Queries are built incrementally**

- Huge potential to re-use intermediate results
- But, re-use of approximate results are not sufficiently studied

[VLDB17]

Potential For Re-Use



Reuse for Approximate Query Processing [SIGMOD17]

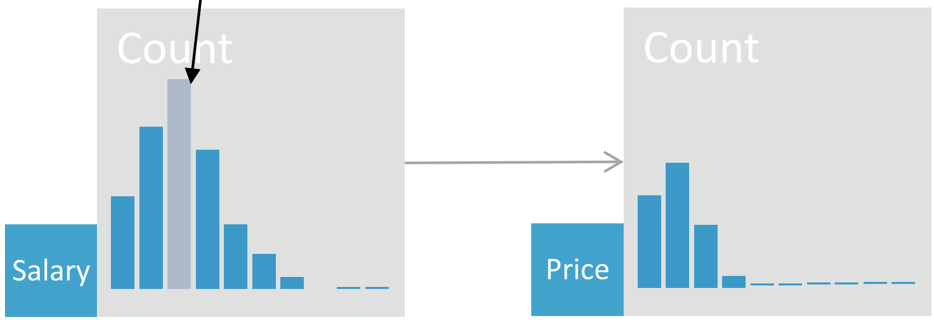
- Visualizations are largely visual representations of statistics.

[VLDB17]

Potential For Re-Use

HeightofBar = P(S="30k-40k") * N

S=Random variable represented the salary
N=Data Size



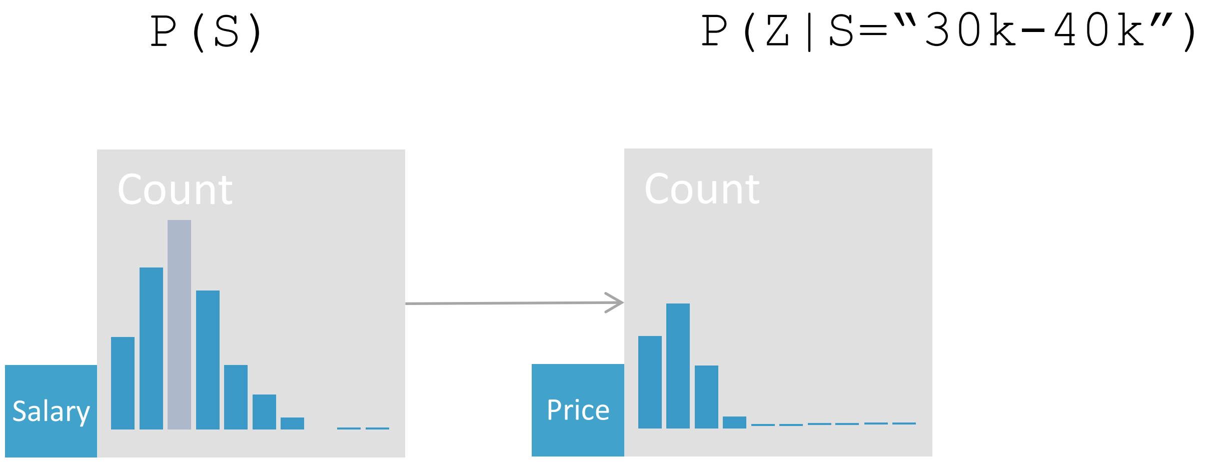
Reuse for Approximate Query Processing [SIGMOD17]

- Visualizations are largely visual representations of statistics.

[VLDB17]

Potential For Re-Use

S=Random variable represented the salary
N=Data Size



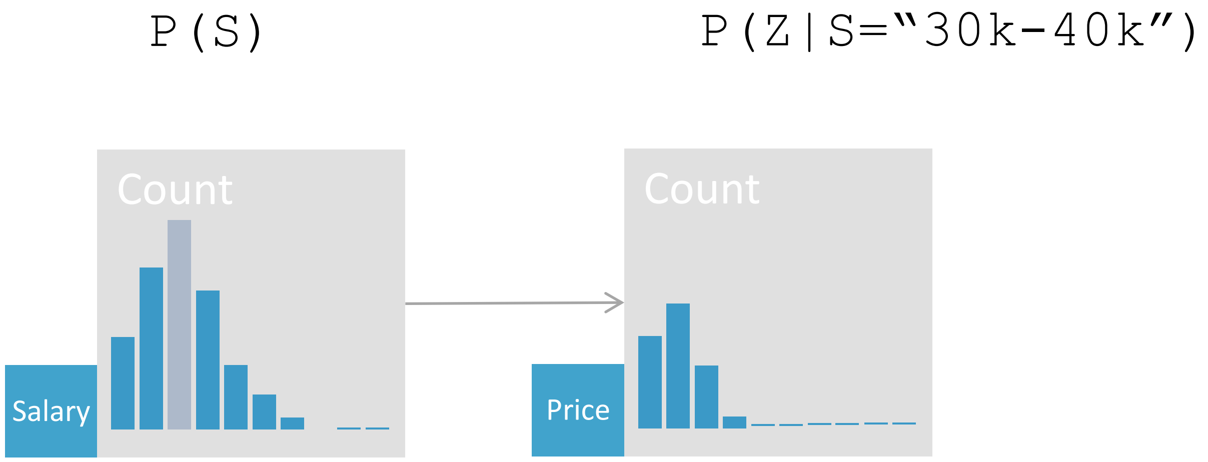
Reuse for Approximate Query Processing [SIGMOD17]

- Visualizations are largely visual representations of statistics.

[VLDB17]

Potential For Re-Use

S=Random variable represented the salary
N=Data Size



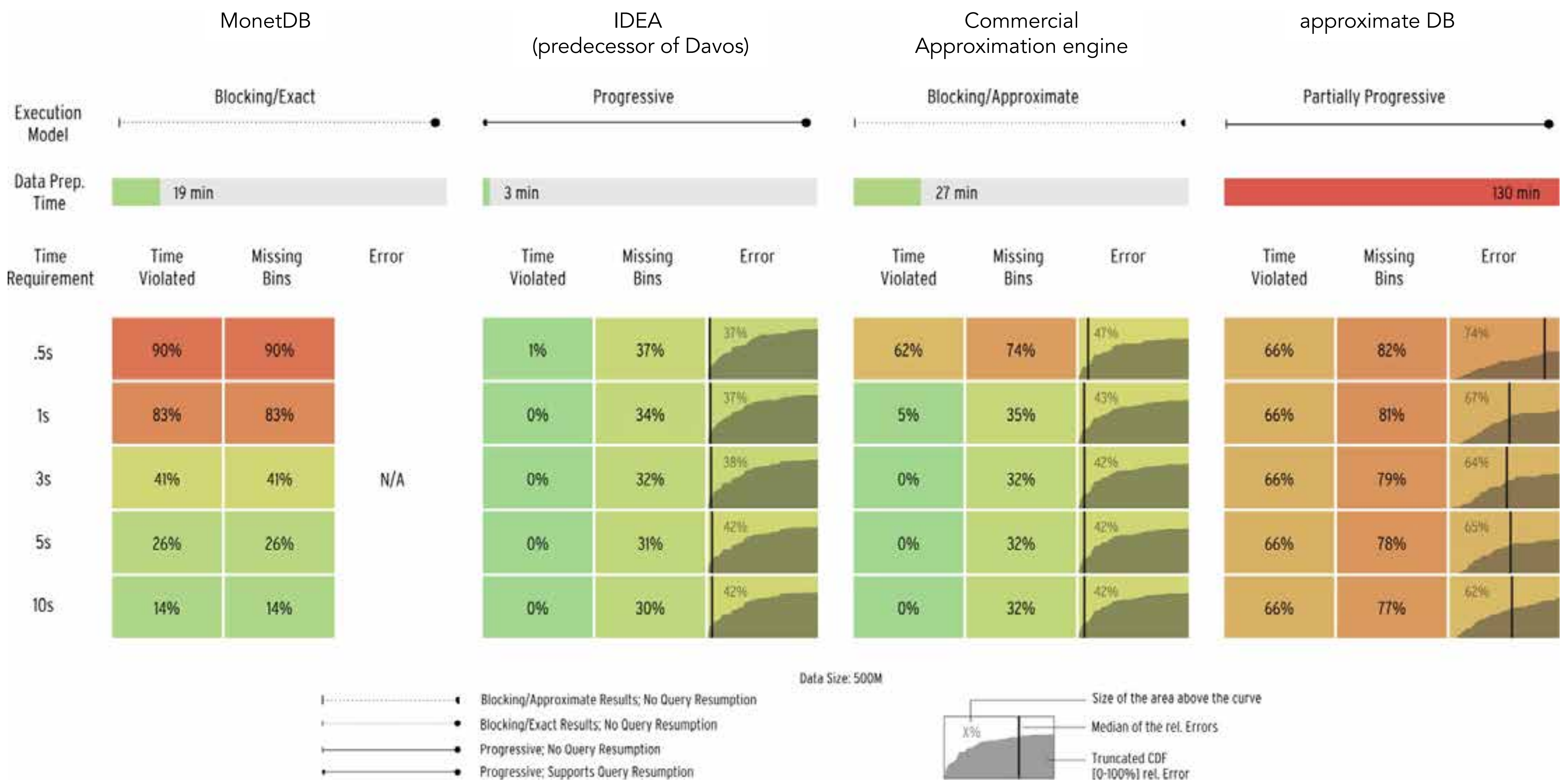
Reuse for Approximate Query Processing [SIGMOD17]

- Visualizations are largely visual representations of statistics.
- Store (inter-)mediate results as random variables
- Query optimization over random variables
  → Enables new optimizations

[VLDB17]

Davos – Faster and more precise results

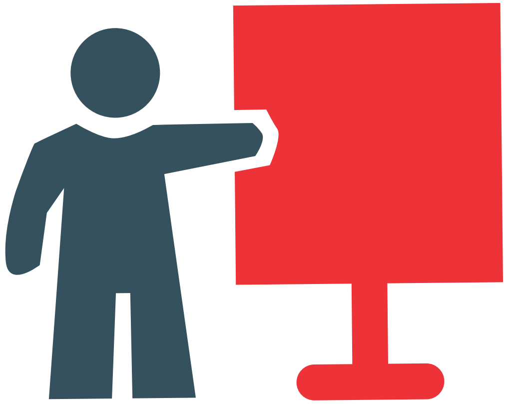
IDEA/Davos is a first progressive query approximation engine. In contrast to alternative system it does require very little pre-processing time and can also approximate the results of Python functions



https://arxiv.org/pdf/1804.02593.pdf

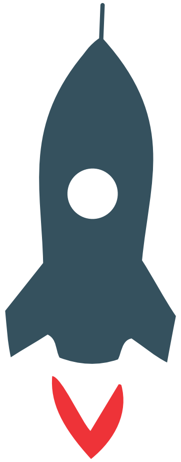
Three Core Technical Contributions

Vizdom
A Novel Interface for Everyone



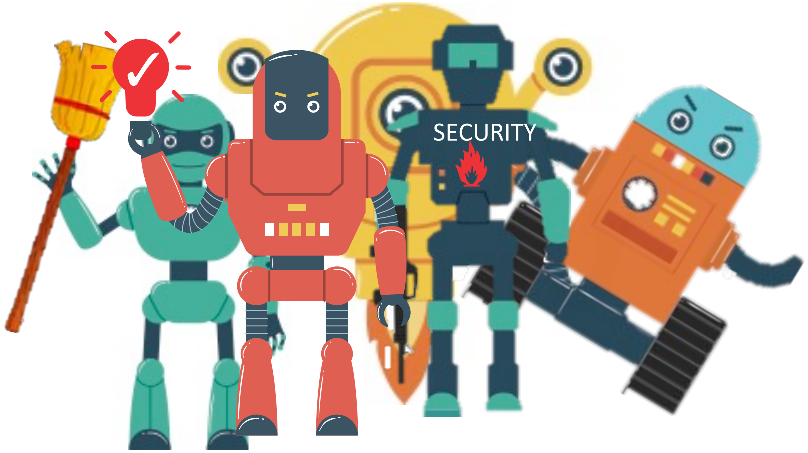
designed for data enthusiast (i.e., people with limited statistics and ML knowledge), domain experts, and data scientist alike.

IDEA
The Data Exploration Accelerator



No waiting: immediately returns visual results for all operations and progressively refines them in the background

Smart Assistance
Towards Data Science Automation



Protect users from common mistakes, point out data cleaning issues, help with building models

Three Core Technical Contributions

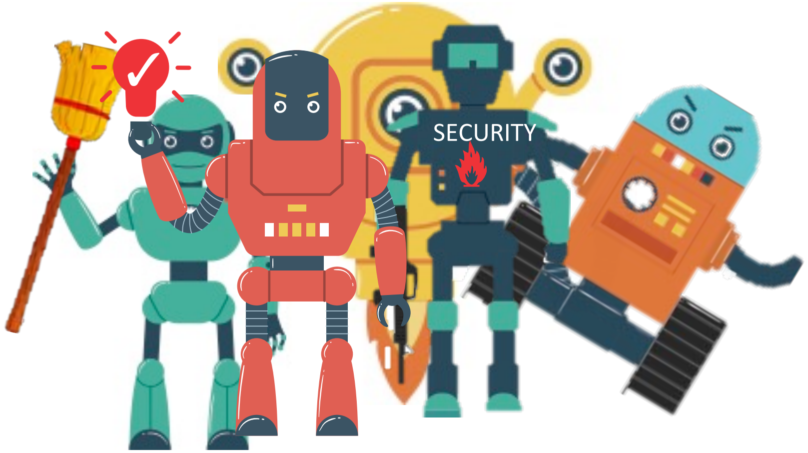
Vizdom
A Novel Interface for Everyone

designed for data enthusiast (i.e., people with limited statistics and ML knowledge), domain experts, and data scientist alike.

IDEA
The Data Exploration Accelerator

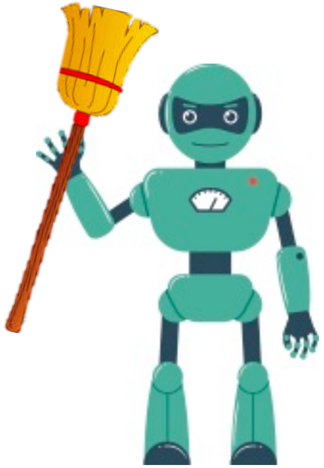
No waiting: immediately returns visual results for all operations and progressively refines them in the background

Smart Assistance
Towards Data Science Automation

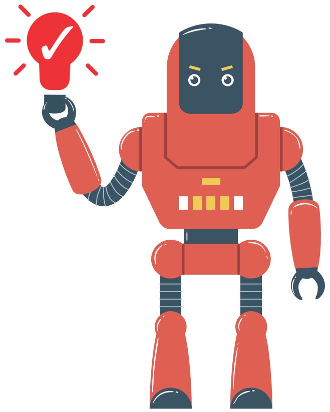


Protect users from common mistakes, point out data cleaning issues, help with building models

ML Assistants Everywhere

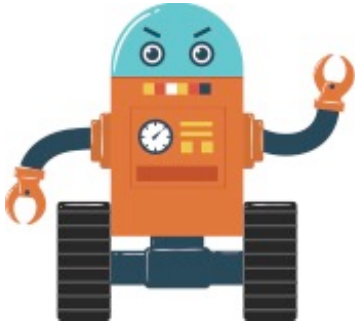


Data Cleaner: automatically bring the data into shape



Insight suggestion: automatically analyze user data for interesting insights

Virtual Data Scientist: given a task find best ML pipeline



Execution Helper: speculatively execute queries

ML Assistants Everywhere

Data Cleaner: automatically bring the data into shape

Insight suggestion: automatically analyze user data for interesting insights

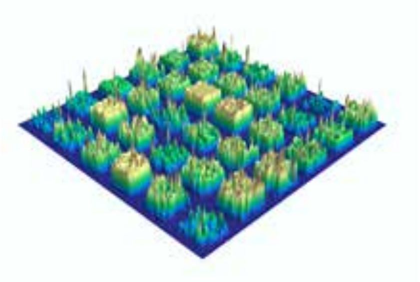
Virtual Data Scientist: given a task find best ML pipeline

Execution Helper: speculatively execute queries

Northstar's Virtual Data Scientist As An Example

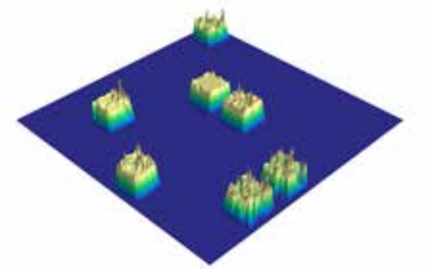
Not Your Normal AutoML–Tool: Build For Interactive Results

What modeling options do I have?



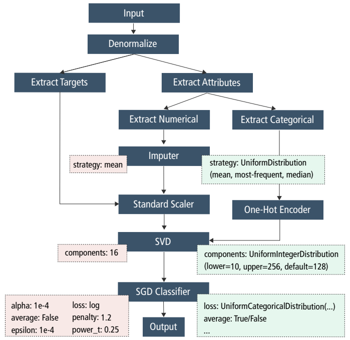
Rule-based Search Space Expansion

What should I try first?



Preselection Based On Past Experience (Learned Knowledge Base)

How can I get some quick results?



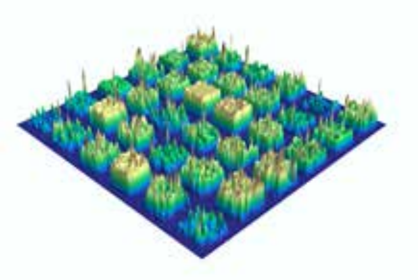
Adaptive sampling-based pruning

ML/System Co–Design: key for achieving **interactivity**

Northstar's Virtual Data Scientist As An Example

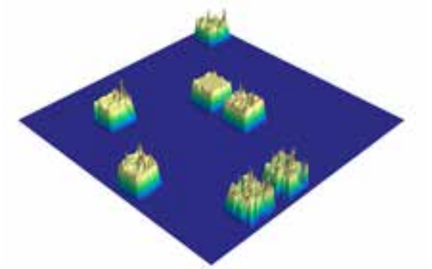
Not Your Normal AutoML-Tool: Build For Interactive Results

What modeling options do I have?



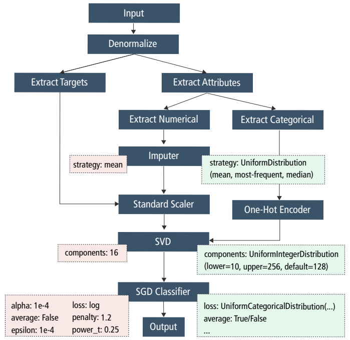
Rule-based Search Space Expansion

What should I try **next?**



Preselection Based On Past Experience (Learned Knowledge Base)

How can I get some quick results?



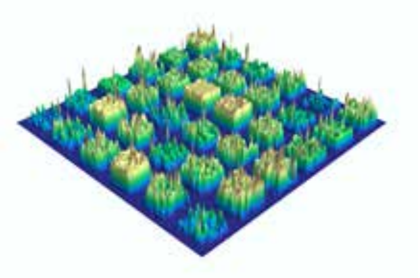
Adaptive sampling-based pruning

ML/System Co-Design: key for achieving **interactivity**

Northstar's Virtual Data Scientist As An Example

Not Your Normal AutoML–Tool: Build For Interactive Results

What modeling options do I have?



Rule-based Search Space Expansion

What should I try first?

Preselection Based On Past Experience (Learned Knowledge Base)

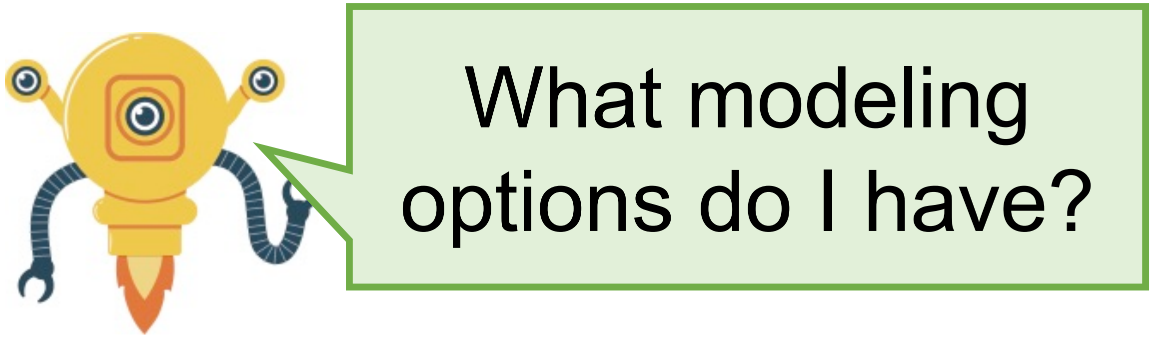
How can I get some quick results?

Adaptive sampling-based pruning

ML/System Co–Design: key for achieving **interactivity**

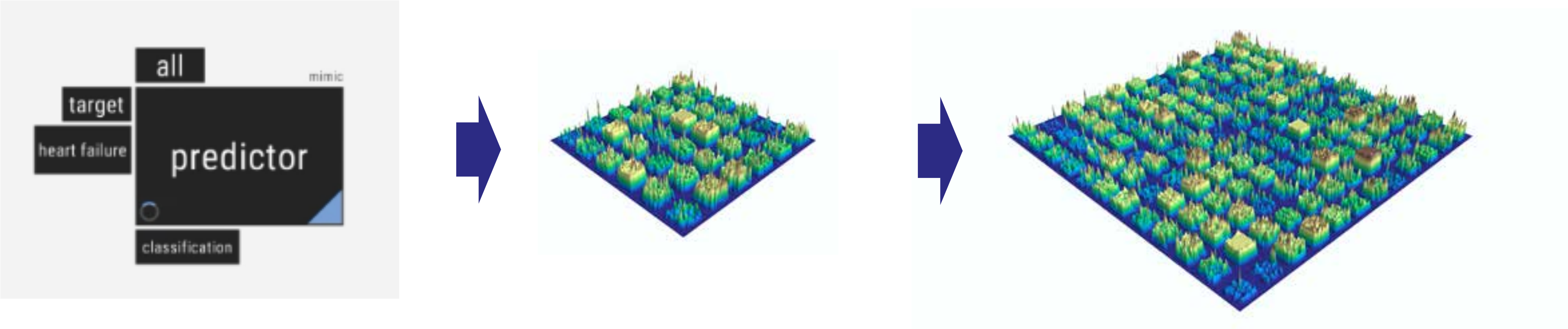
Alpine Meadow

Not Your Normal AutoML–Tool



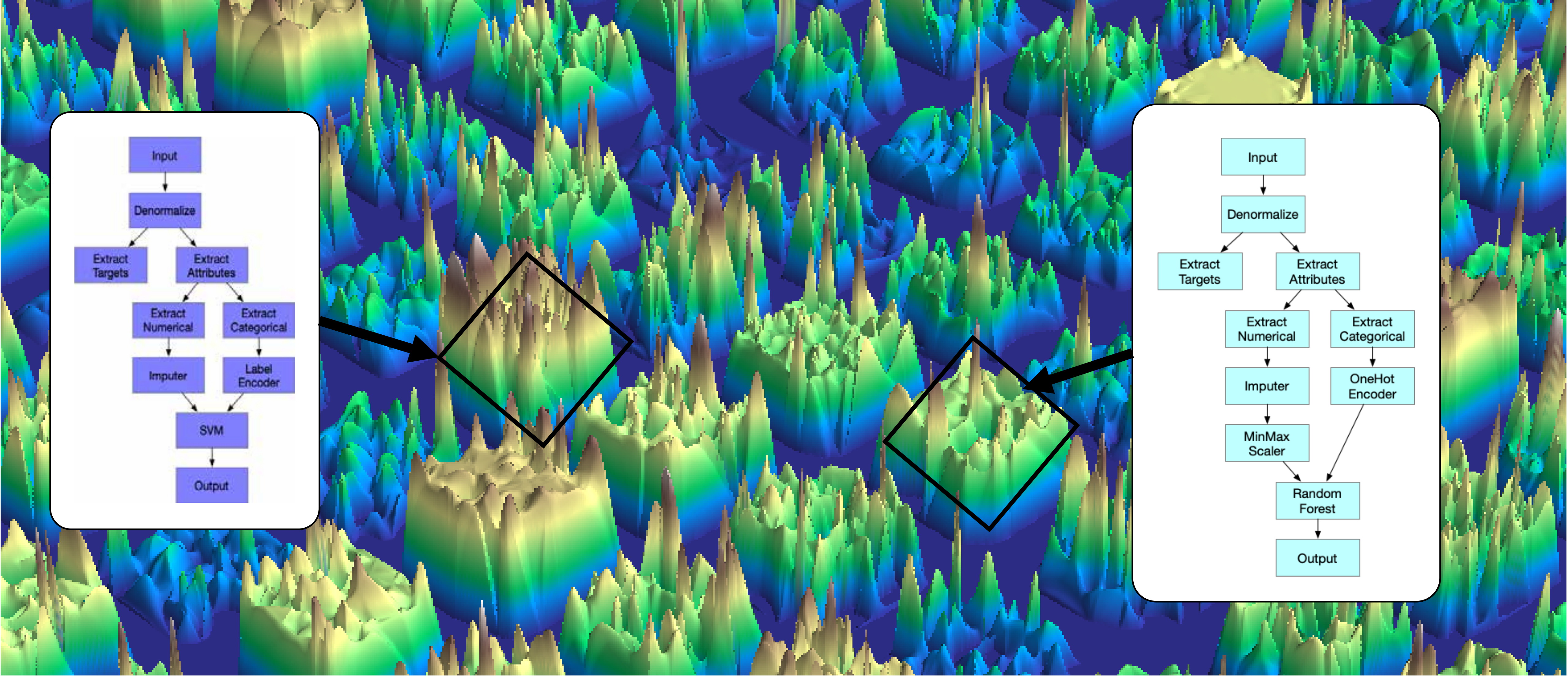
Rule-Based Search Space Expansion

- Rules added by Experts and learned from thousands of publicly available pipelines (Kaggle and OpenML)
- Example rules:
  - unscaled numeric feature → MinMaxScaler, Mean Normalizer
  - categorical feature → use encoder (label or one hot)
  - classification → SVM with default learning rate of 0.001 – 1.00
  - Image classification -> pre-trained neural network (transfer learning)



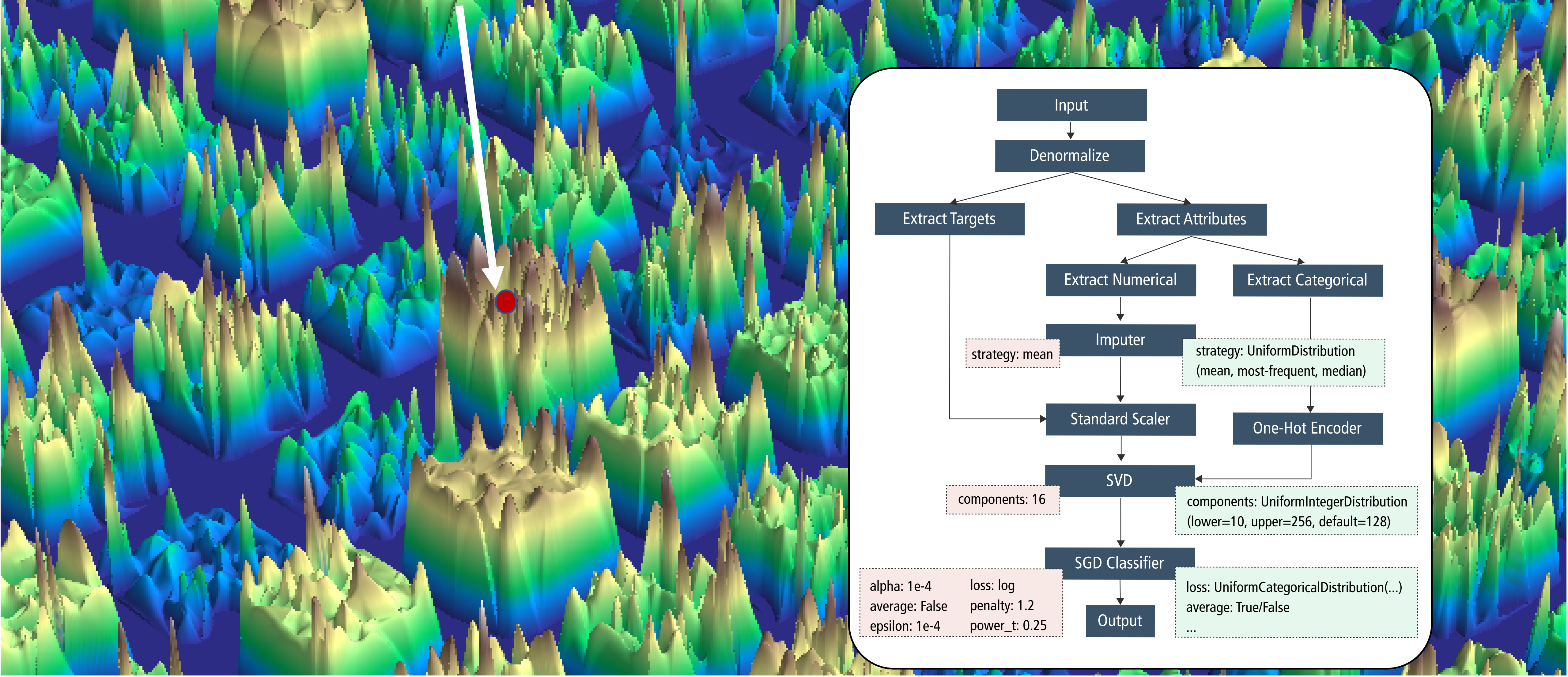
Looking into the Search Space

Every box represents a full logical pipeline
Including feature engineering, preprocessing and model family (e.g., random forest, SVM,…)



Looking into the Search Space

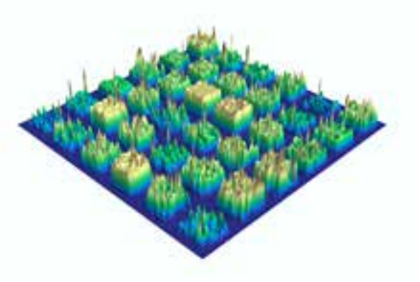
Every point in a box is a physical pipeline including hyperparameters



Northstar's Virtual Data Scientist As An Example

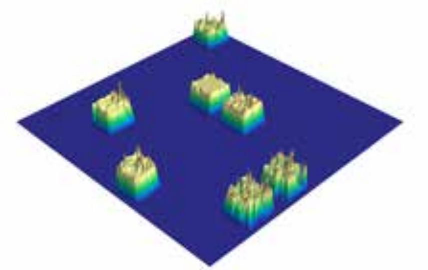
Not Your Normal AutoML–Tool: Build For Interactive Results

What modeling options do I have?



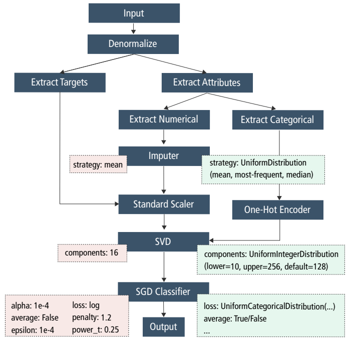
Rule-based Search Space Expansion

What should I try first?



Preselection Based On Past Experience (Learned Knowledge Base)

How can I get some quick results?



Adaptive sampling-based pruning

ML/System Co–Design: key for achieving **interactivity**

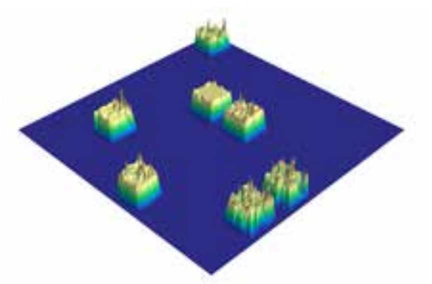
Northstar’s Virtual Data Scientist As An Example

Not Your Normal AutoML–Tool: Build For Interactive Results

What modeling options do I have?

Rule-based Search Space Expansion

What should I try first?



Preselection Based On Past Experience (Learned Knowledge Base)

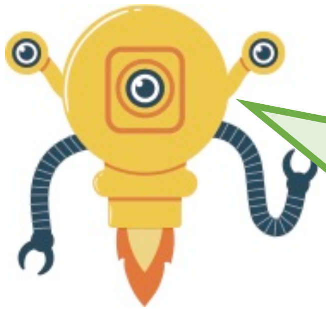
How can I get some quick results?

Adaptive sampling-based pruning

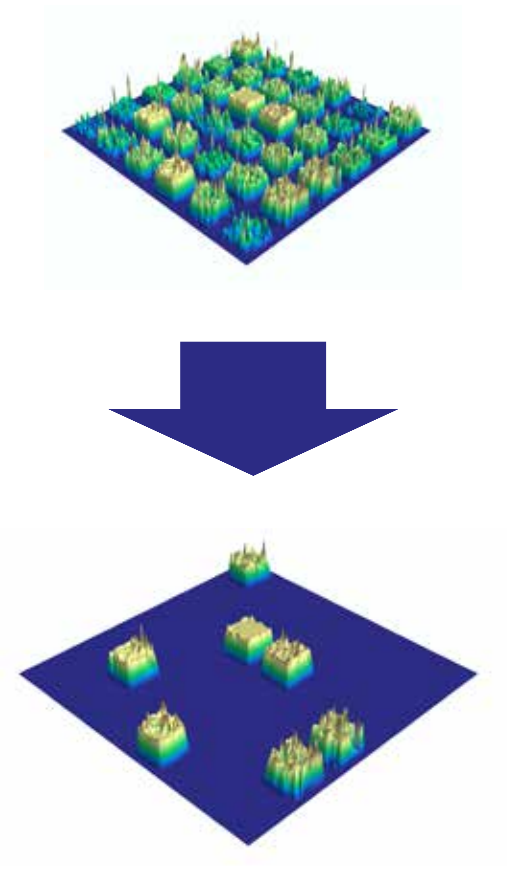
ML/System Co–Design: key for achieving **interactivity**

Alpine Meadow

Not Your Normal AutoML–Tool



"What should I try first?"



Preselection Based On Past Experience

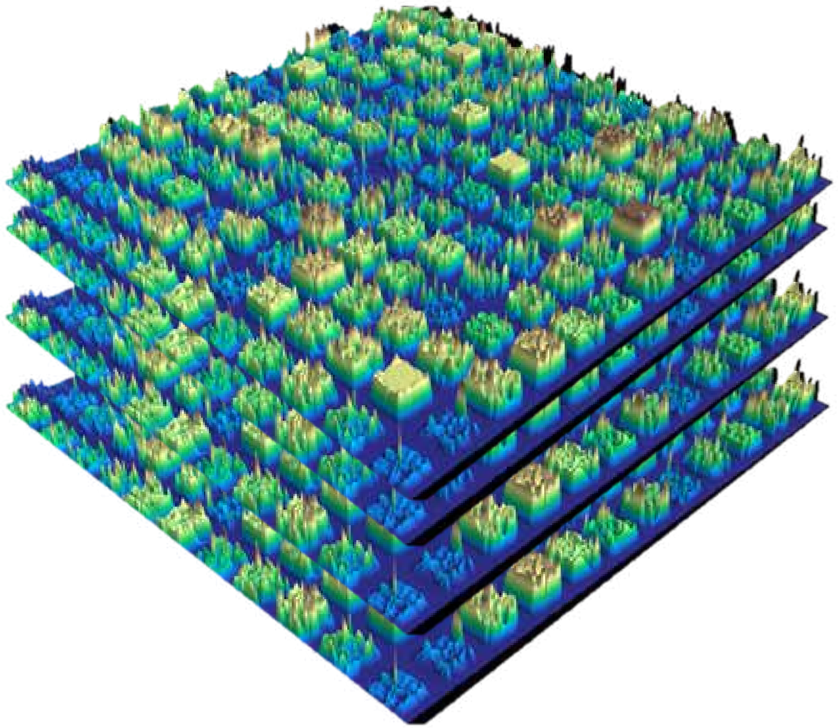
- Expected quality/time trade-off (reliable fast pipelines first, high-risk expensive pipelines later)
- Learned from past experience
- Finally, translate pipeline to python code

For example:

- Gradient Boosting Trees are most-likely a good starting point for the given dataset
- Given the data size, don't even try to use slow models, e.g., neural nets

Alpine Meadow

Not Your Normal AutoML–Tool: Built For Interactive Results



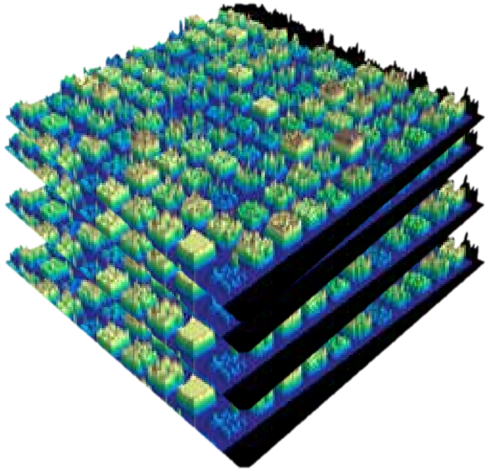
Build knowledge base

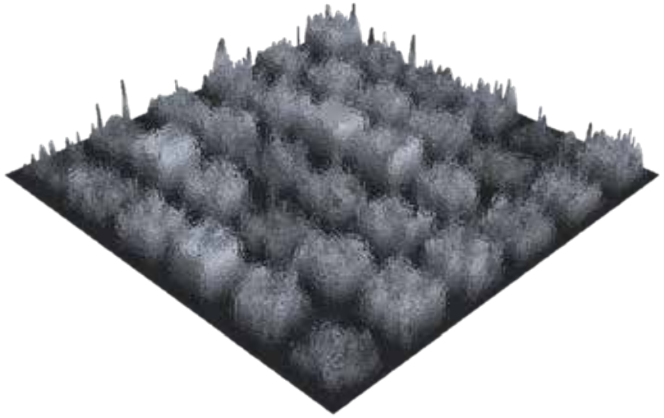
- Run Alpine Meadow on lots of datasets (Kaggle, etc.) and collect all the pipeline traces
- Every time a Alpine Meadow "solves" a new problem add the traces to the knowledge base

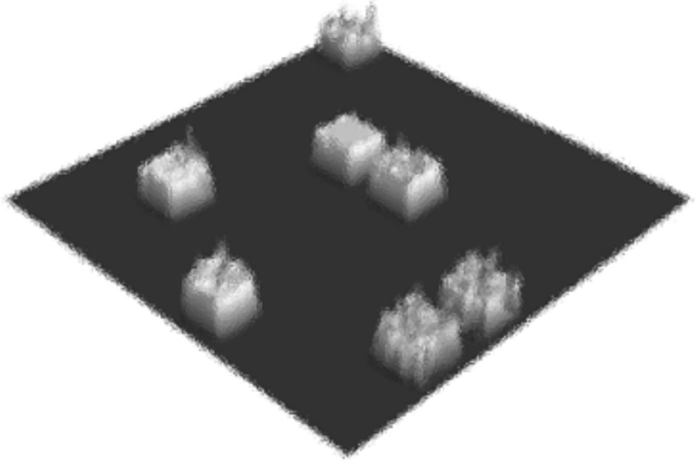
Alpine Meadow

Not Your Normal AutoML–Tool: Built For Interactive Results

Pipeline Selection

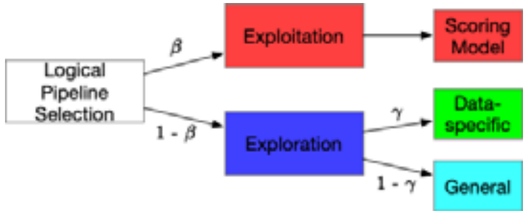






- Find "similar" problems (**meta-learning**) and score logical pipelines based on the past experience and training time
- Similarity is defined through meta-features of a dataset

- Select most promising logical pipelines based on score
- Balance exploration vs exploitation



Alpine Meadow

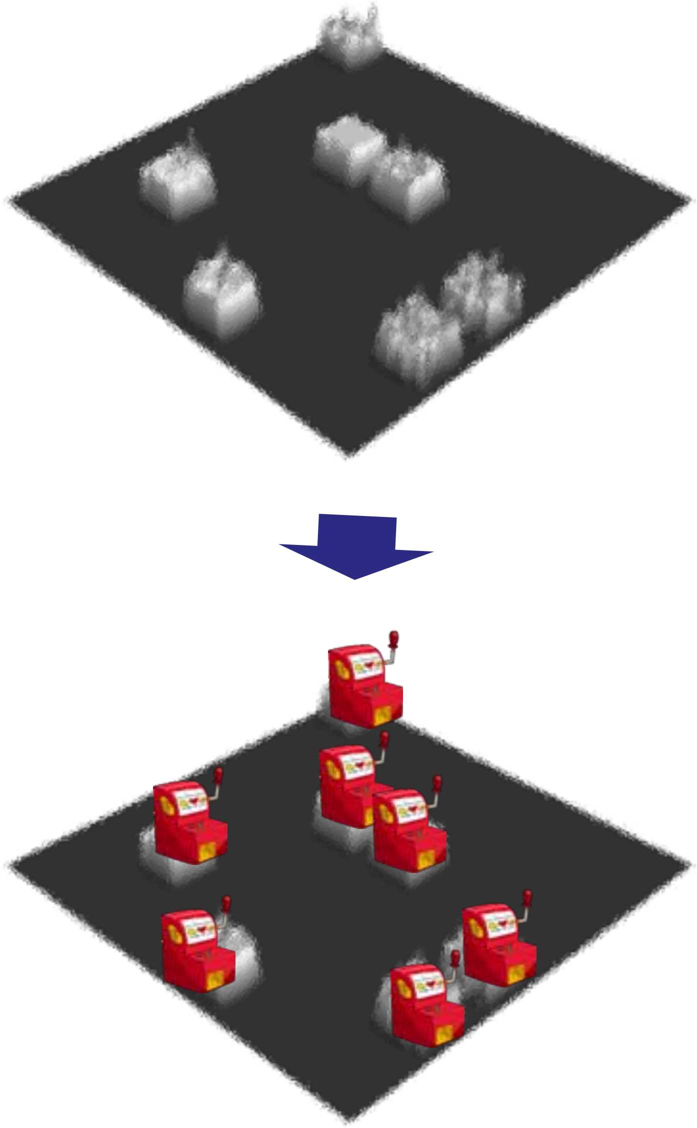
Not Your Normal AutoML–Tool: Built For Interactive Results

Cost-aware Scoring Model

- Multi-armed bandit problem
- Use past history to select promising logical pipelines (warm-starting from the knowledge bases)
- Consider cost and performance at the same time

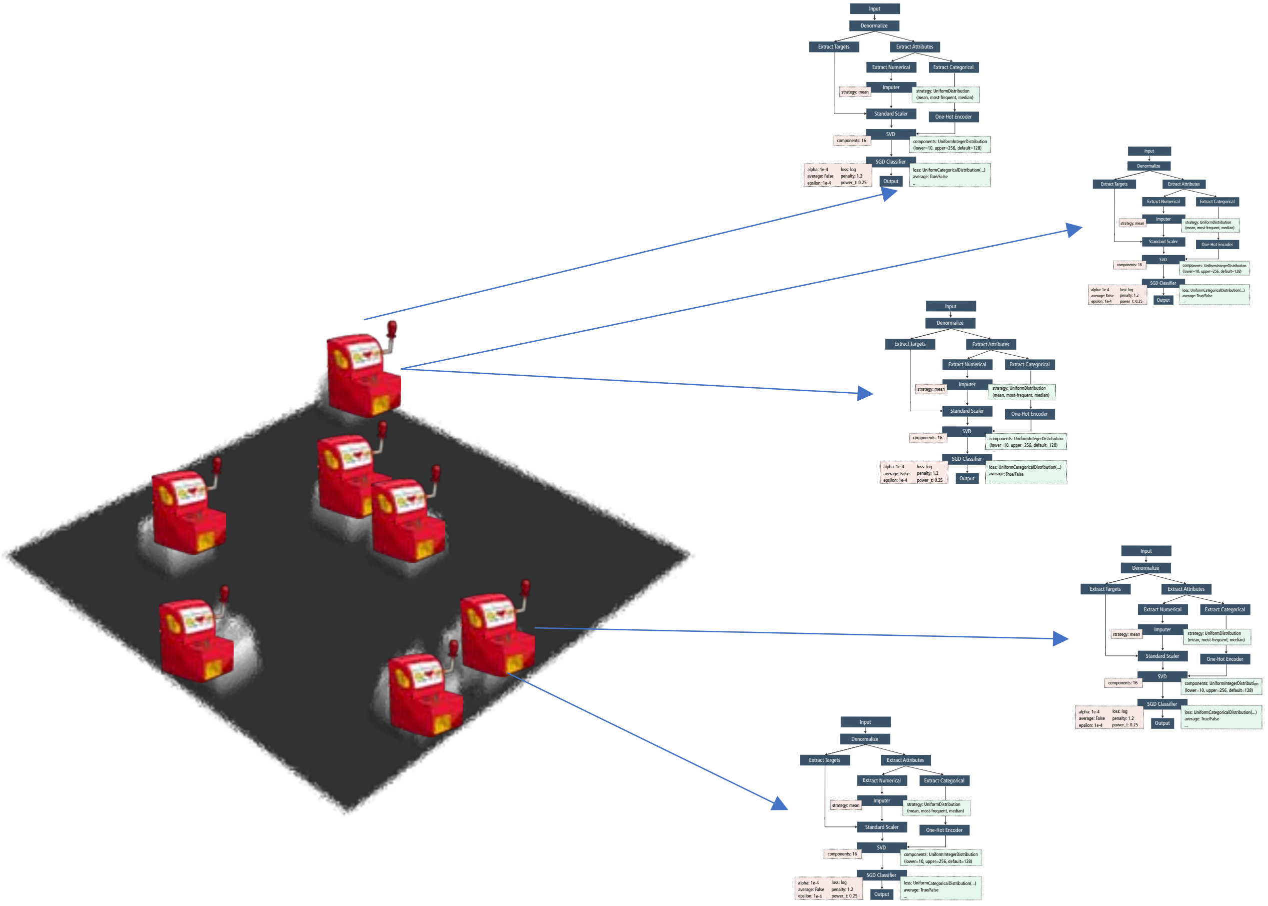
- μ: mean of performance (e.g., accuracy)
- c: mean of cost (e.g., time)
- δ: standard deviation of performance
- **ϴ**: constant to balance risk
- Selecting pipeline with probability proportional to S

$$s = \mu + \frac{\Theta}{c}\delta$$



Alpine Meadow

Not Your Normal AutoML–Tool: Built For Interactive Results



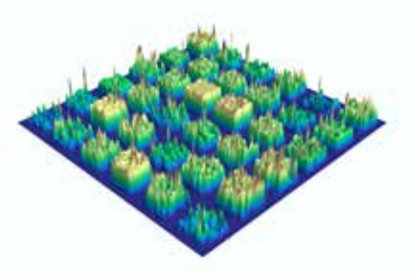
Physical Pipeline Selection

- Hyper-parameter tuning: Bayesian Optimization
- Efficient method for black-box function optimization
- Model the function behavior and select the next promising one

Northstar's Virtual Data Scientist As An Example

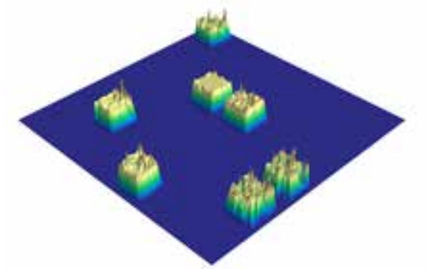
Not Your Normal AutoML–Tool: Build For Interactive Results

What modeling options do I have?



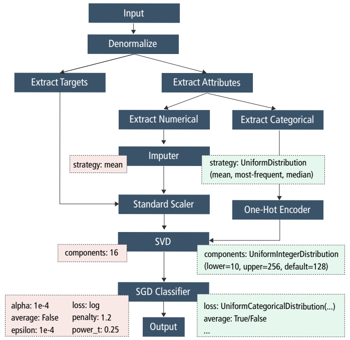
Rule-based Search Space Expansion

What should I try first?



Preselection Based On Past Experience (Learned Knowledge Base)

How can I get some quick results?



Adaptive sampling-based pruning

ML/System Co–Design: key for achieving **interactivity**

Northstar's Virtual Data Scientist As An Example

Not Your Normal AutoML–Tool: Build For Interactive Results

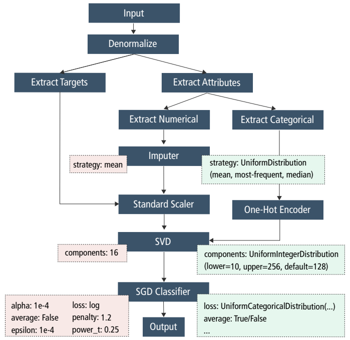
What modeling options do I have?

Rule-based Search Space Expansion

What should I try first?

Preselection Based On Past Experience (Learned Knowledge Base)

How can I get some quick results?

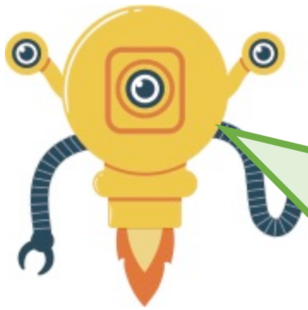


Adaptive sampling-based pruning

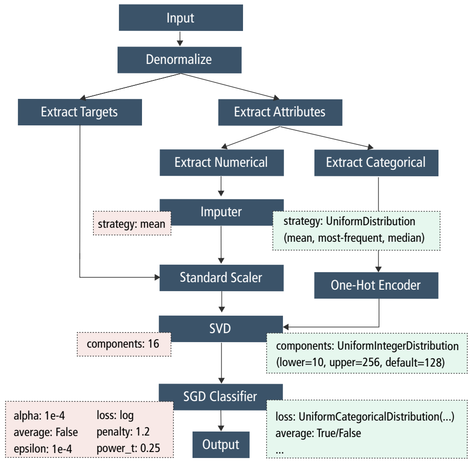
ML/System Co–Design: key for achieving **interactivity**

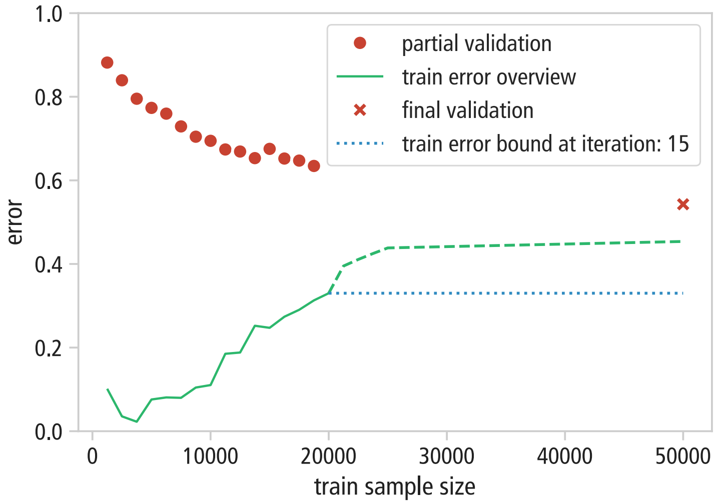
Alpine Meadow

Not Your Normal AutoML–Tool



How can I get some quick results?





Try pipeline first on a small sample

- Observe training and test error
- If pipeline performs well, increase sample size

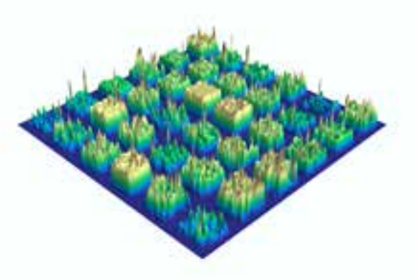
Adaptive Pipeline Selection

- Train error as the lower bound the test error
- Prune if the training error is beyond the current best validation error

Northstar's Virtual Data Scientist As An Example

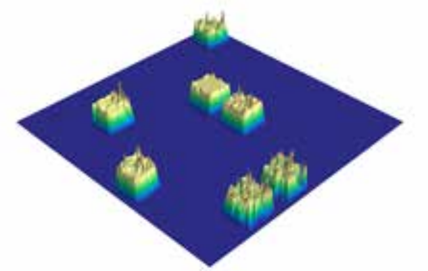
Not Your Normal AutoML–Tool: Build For Interactive Results

What modeling options do I have?



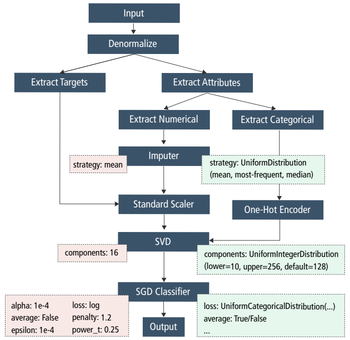
Rule-based Search Space Expansion

What should I try first?



Preselection Based On Past Experience (Learned Knowledge Base)

How can I get some quick results?



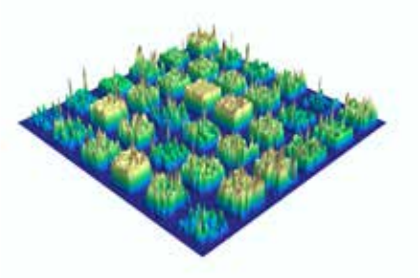
Adaptive sampling-based pruning

ML/System Co–Design: key for achieving **interactivity**

Northstar's Virtual Data Scientist As An Example

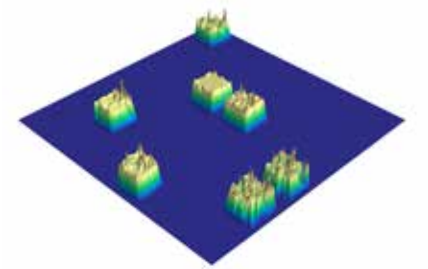
Not Your Normal AutoML–Tool: Build For Interactive Results

What modeling options do I have?



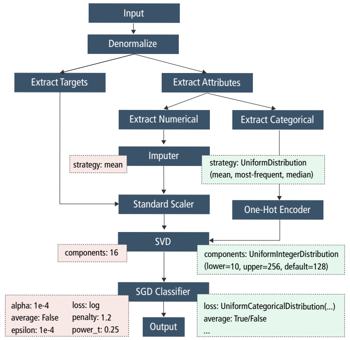
Rule-based Search Space Expansion

What should I try first?



Preselection Based On Past Experience (Learned Knowledge Base)

How can I get some quick results?



Adaptive sampling-based pruning

ML/System Co-Design: key for achieving **interactivity**

DARPA Data-Driven Discovery of Model (D3M)

	Solved Problems	Better Than Baseline	Normalized Score
DARPA Baseline	100%	0%	0.00

- System 2-10 are competing teams from UC Berkeley, Stanford, NYU, ….
- Tested over 300 DARPA datasets
- Includes structured classification and regression task, image classification and measuring, audio transcription, among others

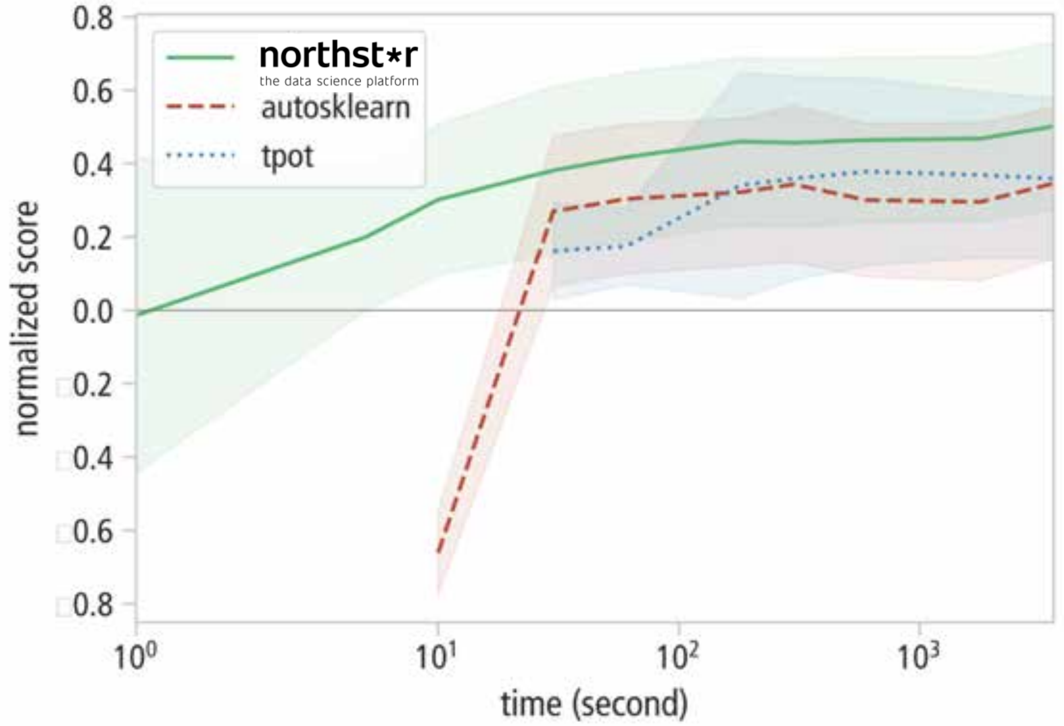
DARPA Data-Driven Discovery of Model (D3M)

	Solved Problems	Better Than Baseline	Normalized Score
northst*r	100%	80%	0.42
System 2	40%	27%	0.09
System 3	40%	13%	0.02
DARPA Baseline	100%	0%	0.00
System 4	20%	7%	-0.07
System 5	87%	47%	-0.16
System 6	27%	7%	-0.22
System 7	60%	20%	-0.59
System 8	87%	53%	-0.75
System 9	60%	20%	-1.14
System 10	60%	20%	-4.57

- System 2–10 are competing teams from UC Berkeley, Stanford, NYU, ….
- Tested over 300 DARPA datasets
- Includes structured classification and regression task, image classification and measuring, audio transcription, among others

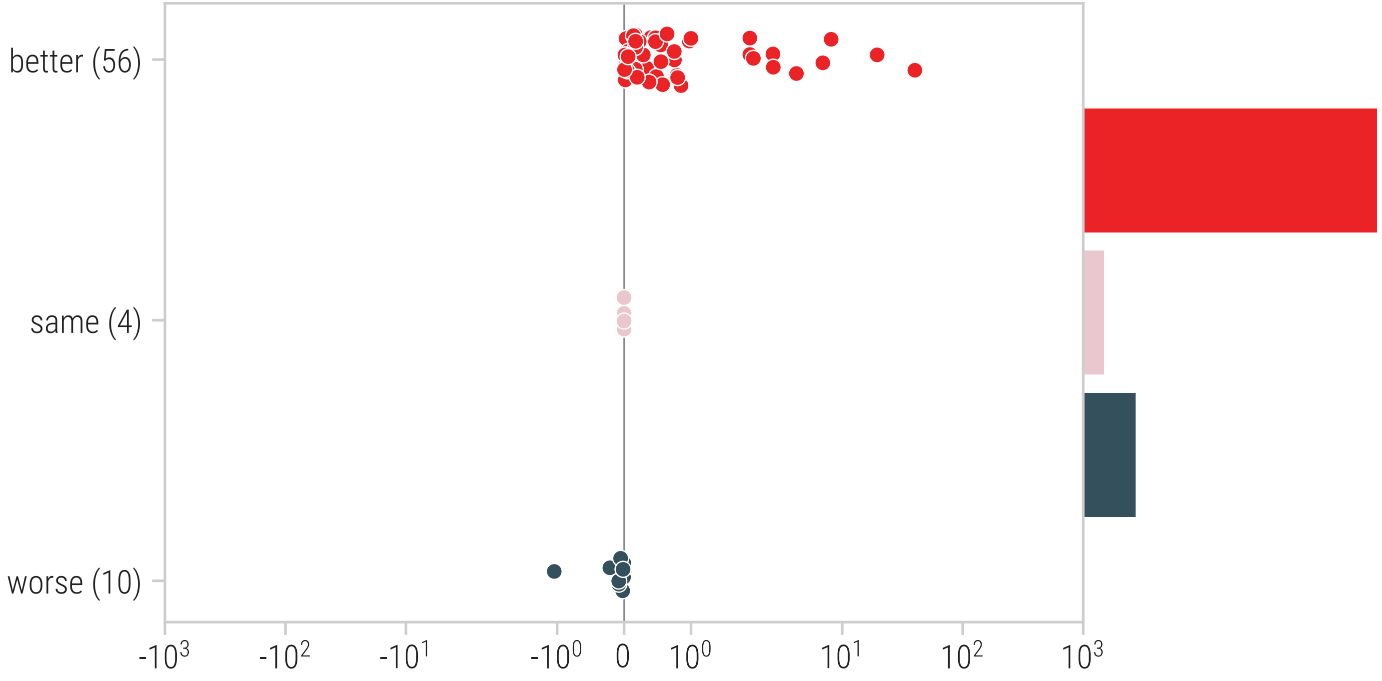
Northstar's Auto-ML: Better Results Faster

	Solved Problems	Better Than Baseline	Normalized Score
northst*r	100%	80%	0.42
System 2	40%	27%	0.09
System 3	40%	13%	0.02
DARPA Baseline	100%	0%	0.00
System 4	20%	7%	-0.07
System 5	87%	47%	-0.16
System 6	27%	7%	-0.22
System 7	60%	20%	-0.59
System 8	87%	53%	-0.75
System 9	60%	20%	-1.14
System 10	60%	20%	-4.57



- System 2–10 are competing teams from UC Berkeley, Stanford, NYU, ….
- Tested over 300 DARPA datasets
- Includes structured classification and regression task, image classification and measuring, audio transcription, among others

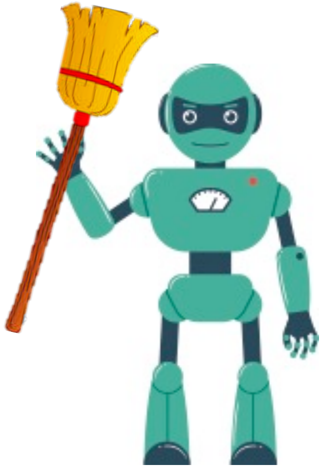
For 86% of the Datasets Better Than Azure's AutoML



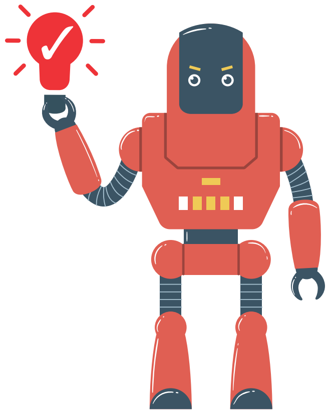
$$\text{normalized score} = \frac{\text{northstar score - Azure ML score}}{\text{Azure ML score}}$$

- Tested on 150 tabular regression and classification datasets for 10 minutes each
- Only able to obtain scores on Azure ML for 70 of the datasets; we started to investigate with the Azure team why the failures happen
- Northstar outperforms Azure AutoML in 86% of the successful runs
- Northstar supports many more problem types than Azure AutoML: Graph Matching, Community Detection, Image Classification, Audio Classification, Collaborative Filtering

ML Assistants Everywhere

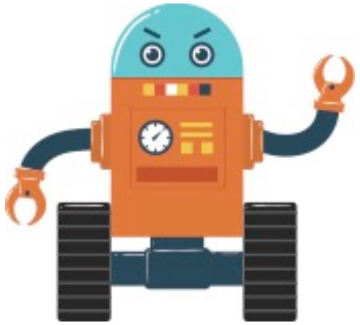


Data Cleaner: automatically bring the data into shape



Insight suggestion: automatically analyze user data for interesting insights

Virtual Data Scientist: given a task find best ML pipeline



Execution Helper: speculatively execute queries

Case Study Part III: Open Challenges (5min)

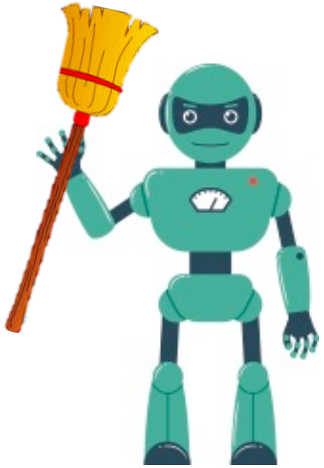
List potential challenges/problems.

For example:

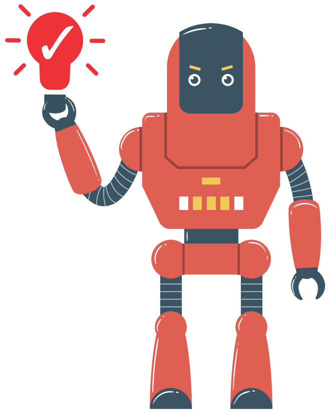
- What problems do you see in letting domain experts without a deep understanding of ML/statistics do complex analytics on their own?
- Are there increase risk factors for experts in ML/statistics?
- What might be potential limitations of such a system?
- Do you see other technical challenges?

When you are done, please hand in your final case study document to Matt Perron <mperron@csail.mit.edu>

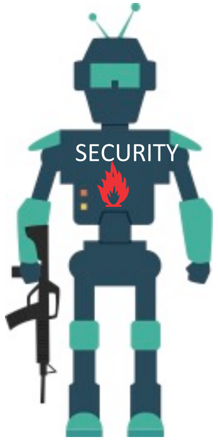
ML Assistants Everywhere



Data Cleaner: automatically bring the data into shape



Insight suggestion: automatically analyze user data for interesting insights



Discovery Protector: protect the user of common mistakes

Virtual Data Scientist: given a task find best ML pipeline

Execution Helper: speculatively execute queries

Skip Discovery Protector

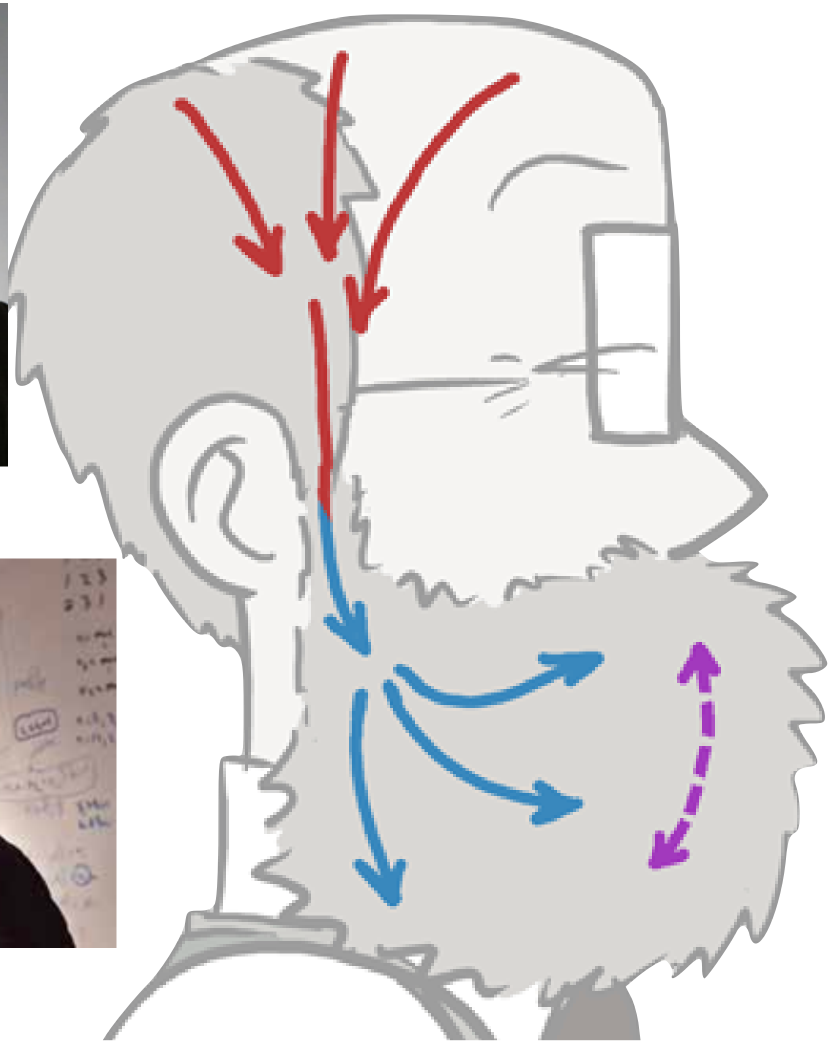
There has been an **Increase** of very

Questionable Findings

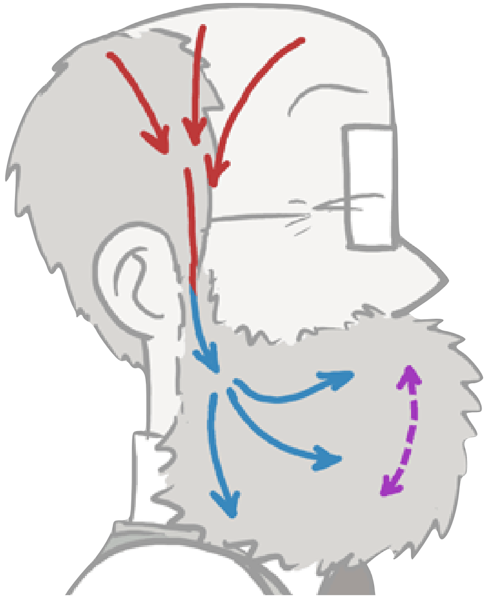
A New Study shows: A Glass Of Red Wine Is The Equivalent To An Hour At The Gym [Fox News 02/15 and others]

http://www.huffingtonpost.co.uk/2016/01/08/a-glass-of-red-wine-is-the-equivalent-to-an-hour-at-the-gym-says-new-study_n_7317240.html

Very concerning hypothesis in the media: The hair migration pattern of male professors




LONG LIFE



Reasons are manifold, but easy to use visual exploration tools contribute to the problem

Why Northstar and systems like it increase the risk of multiplicity

Interactive Data Exploration

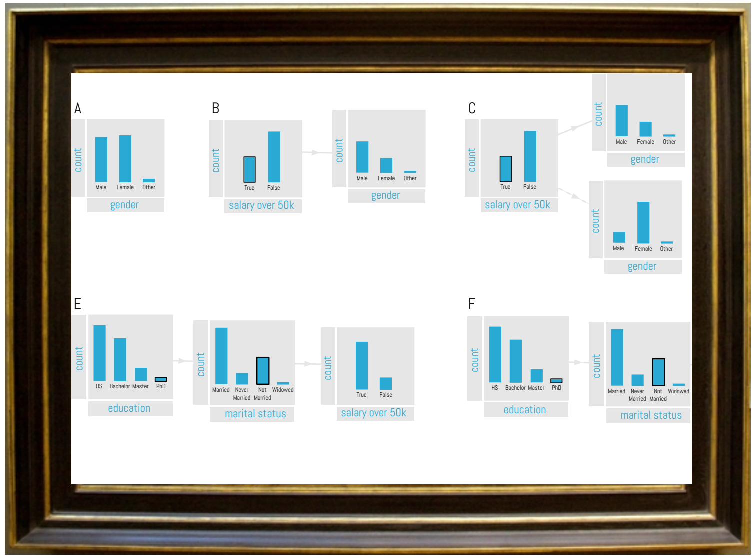
Visualization Recommendation Systems

Hypothesis Generator

Solutions

Interactive Data Exploration Tools

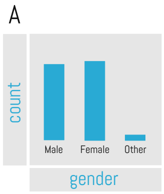
Northstar as an example but also applies to Tableau, PowerBI, etc.



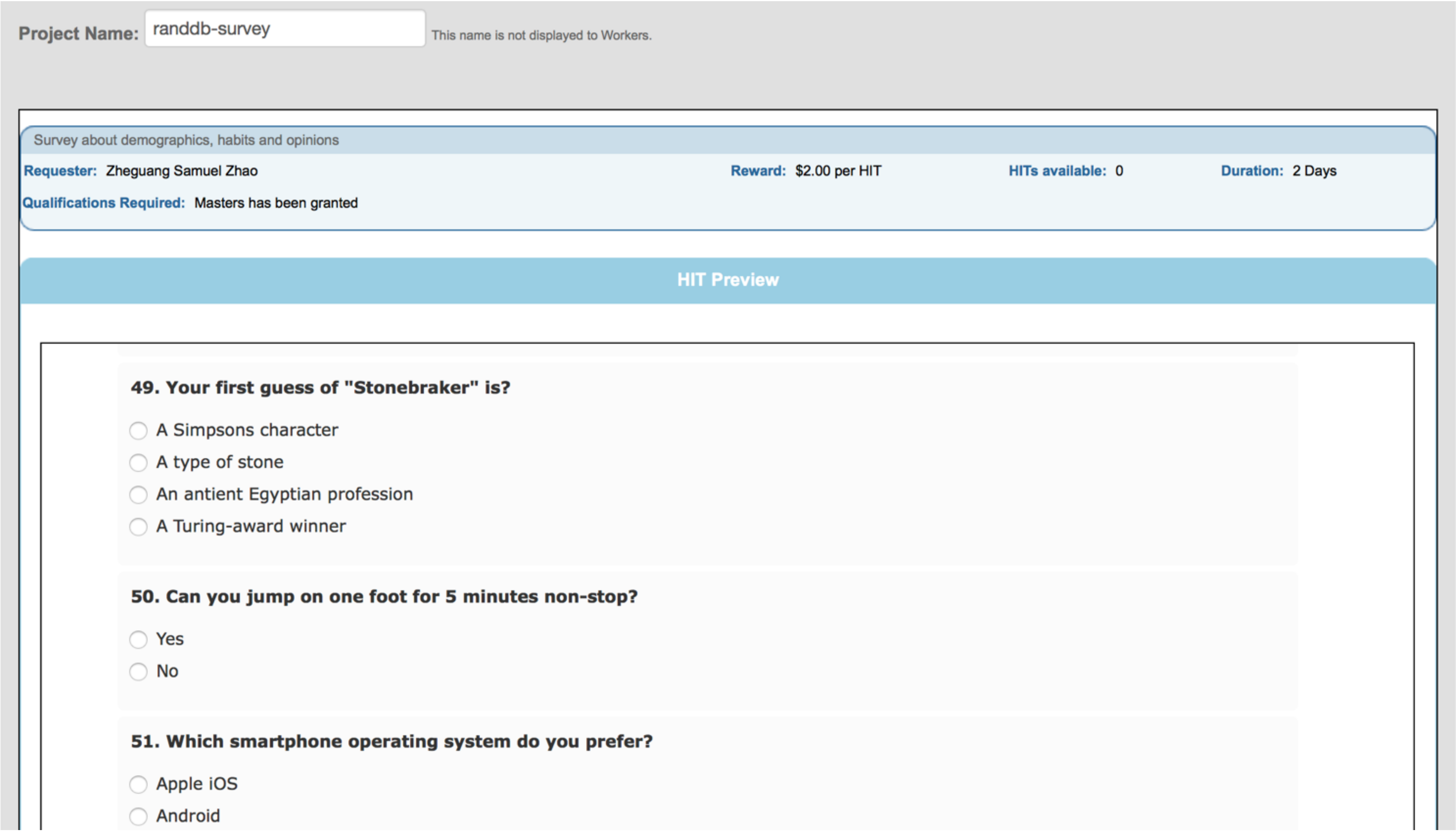
If a visualization provides any insight over a larger population, it is a hypothesis test

Otherwise, visualizations have just to be taken as pretty pictures about (potentially) random facts

If visualizations are used to find something interesting, the user is doing multiple hypothesis testing



Running Example: Survey on Amazon Mechanical Turk



Project Name: randdb-survey This name is not displayed to Workers.

Survey about demographics, habits and opinions

Requester: Zheguang Samuel Zhao **Reward:** $2.00 per HIT **HITs available:** 0 **Duration:** 2 Days

Qualifications Required: Masters has been granted

HIT Preview

49. Your first guess of "Stonebraker" is?

- A Simpsons character
- A type of stone
- An antient Egyptian profession
- A Turing-award winner

50. Can you jump on one foot for 5 minutes non-stop?

- Yes
- No

51. Which smartphone operating system do you prefer?

- Apple iOS
- Android

Our goal: To find good indicators (correlations) that somebody knows who Mike Stonebraker is.

And after searching for a bit, one of my favorites

Pearson correlation significance-level $p < 0.05$

How do interactive data exploration tools contribute

Criticism

Blaming the multiple-comparison problem on fast visualization-generation is like blaming fast cars for child driver casualties due to car accidents…

But…

Why Northstar and systems like it increase the risk of multiplicity

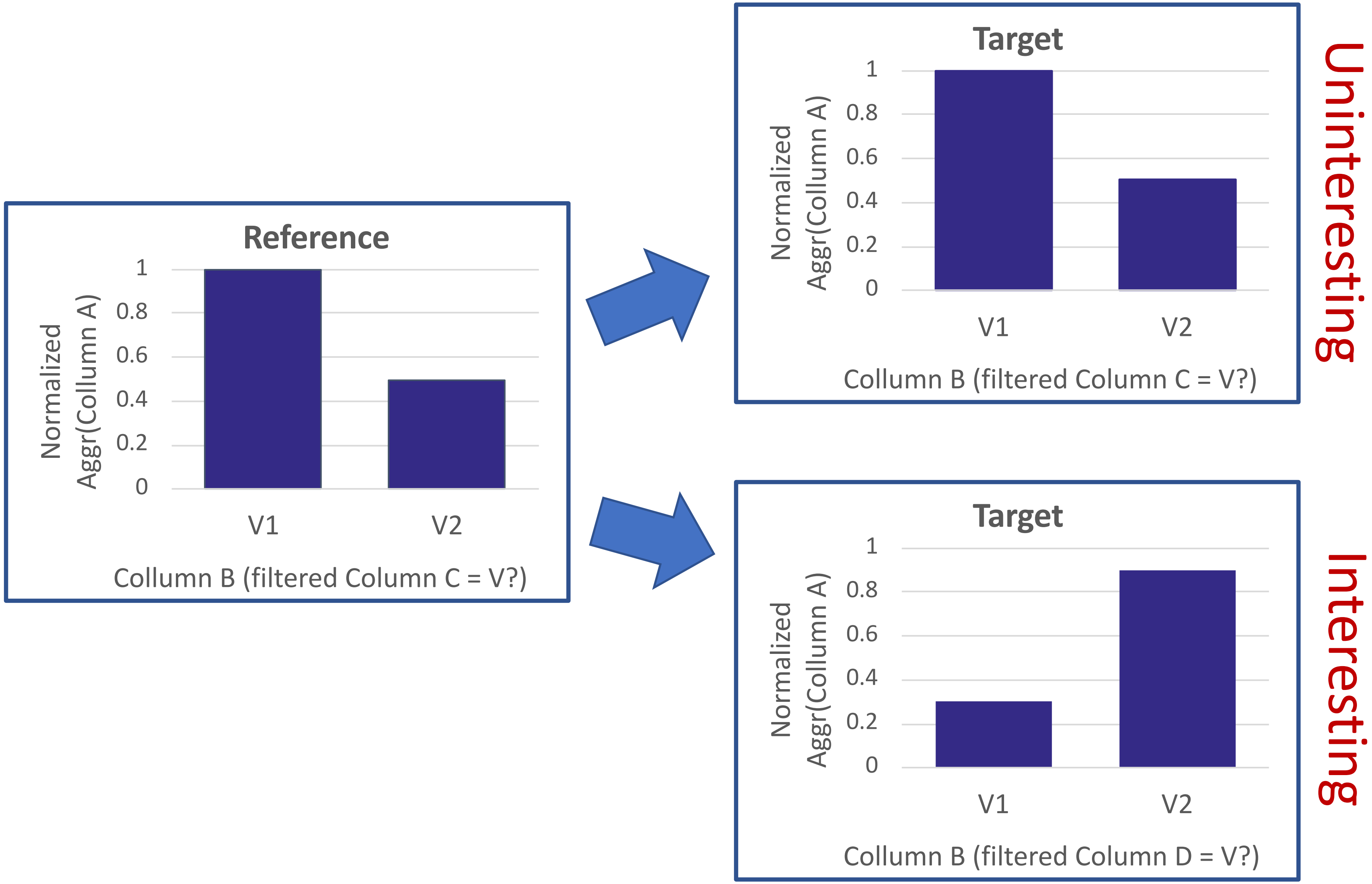
Interactive Data Exploration

Visualization Recommendation Systems

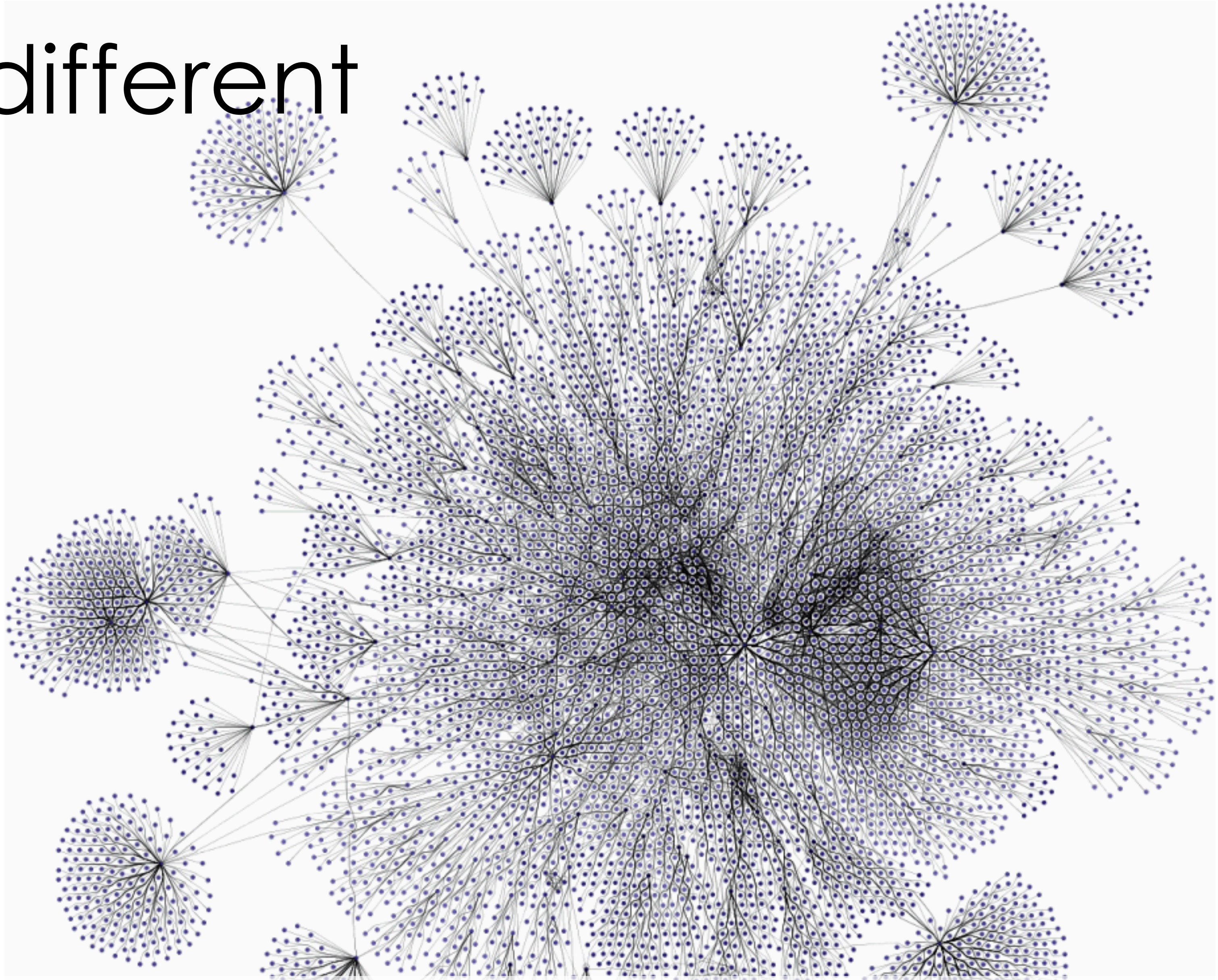
Hypothesis Generator

Solutions

Visual Recommendation Systems (SeeDB as an Example)

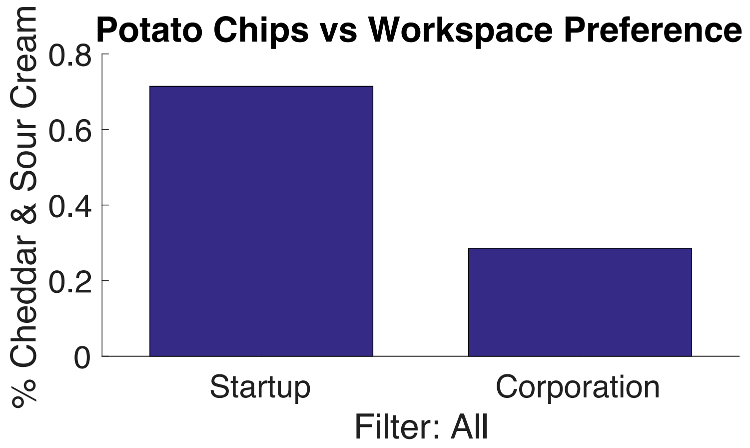


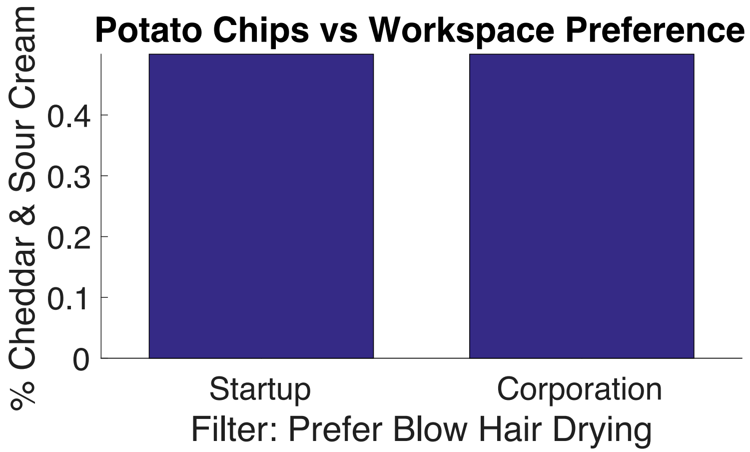
What is different

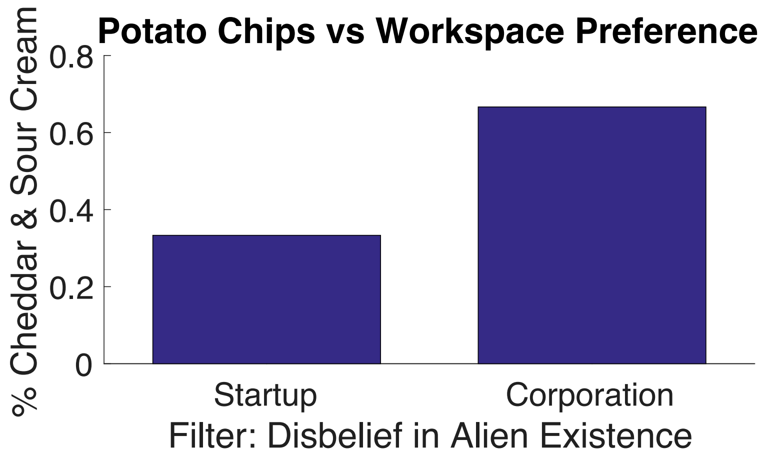


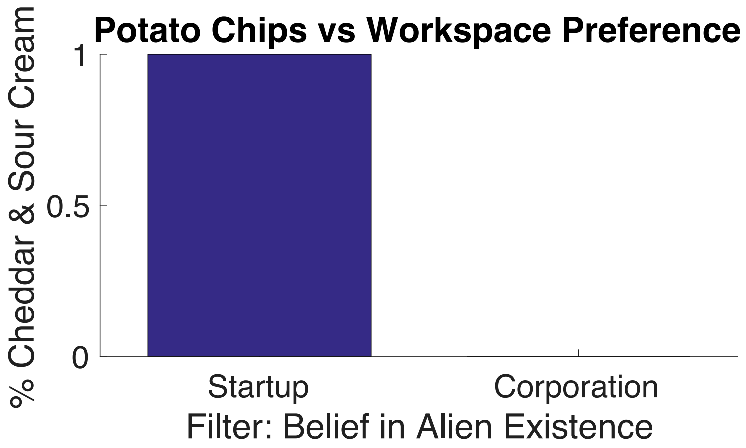
The system automatically generates thousands of visualizations and ranks them somehow (e.g., based on effect size)

SeeDB on Our Survey Data

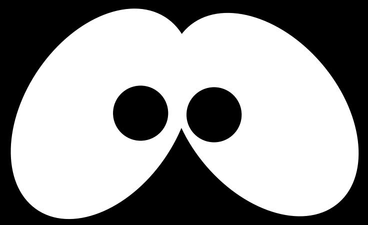








What is the Problem?



The user is in the dark what the system did.
The system might have “tested” thousands of potential visualization, just to find something interesting.

My suggestions, these tools should include a warning like

WARNING

After using the tool,
throw away the data.
It is not safe![1]

[1]To be more precise: you do not have to throw it all away, but you can not use the same data anymore for significance testing

Why Northstar and systems like it increase the risk of multiplicity

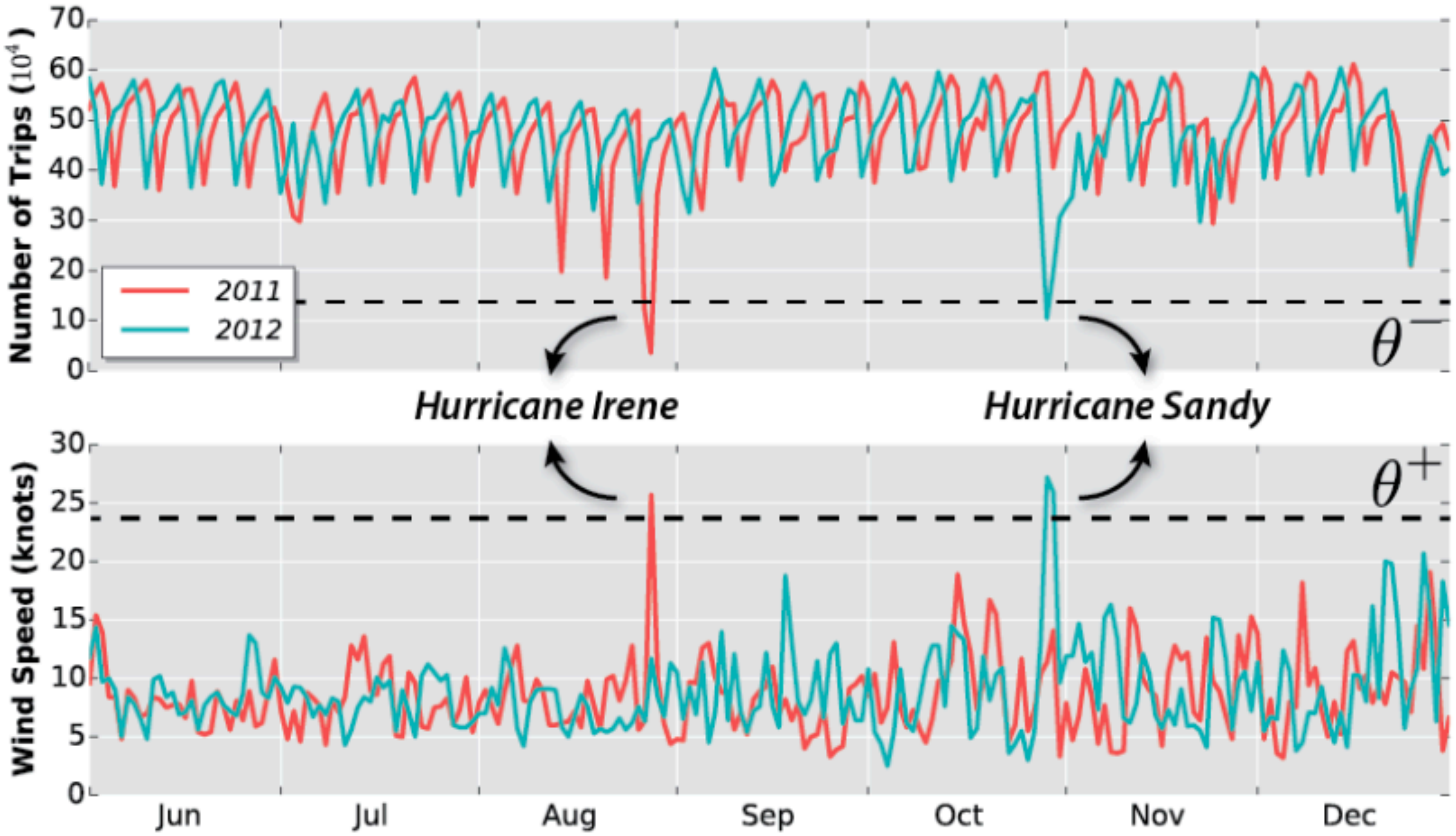
Interactive Data Exploration

Visualization Recommendation Systems

Hypothesis Generator

Solutions

3) Real Hypothesis Generators (Data Polygamy as an Example)



Example Data Polygamy

- We executed Data Polygamy over a (small) randomly generated data set with 11 attributes
- We further injected randomly generated extreme data points sampled from a different distribution
- With this setup Data Polygamy found a total of **43 random relationships in 50 independent trials**
- The problem, like before, you can not use the same data anymore to verify your findings.
- Also note that Data Polygamy is the definition of p-hacking: as described in the paper it searches for a correlation with a p-value smaller than 0.05

Should we stop working on IDE, Recommenders, etc?

NO

- Actively inform the user about the risk factors
- *If possible*, split data into **a exploration and a validation set.**
  - Be aware, **significantly lowers the power**
  - Everything on the validation data set has to be carefully handled (i.e., use multi-hypothesis control)
- *If possible*, use **additional experiments** (e.g., A/B testing)
  - Requires a small number of hypothesis and careful design
  - Might not always be possible or is very expensive

Better: control the multi-hypothesis problem from the start

Why Northstar and systems like it increase the risk of multiplicity

Interactive Data Exploration

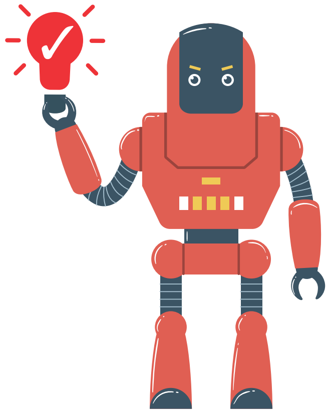
Visualization
Recommendation Systems

Hypothesis Generator

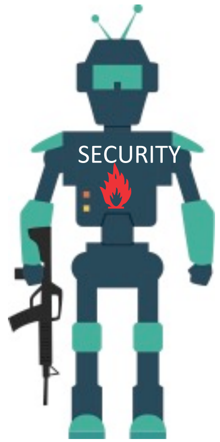
Solutions

ML Assistants Everywhere

Data Cleaner: automatically bring the data into shape

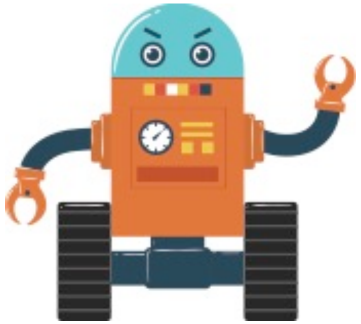


Insight suggestion: automatically analyze user data for interesting insights



Discovery Protector: protect the user of common mistakes

Virtual Data Scientist: given a task find best ML pipeline

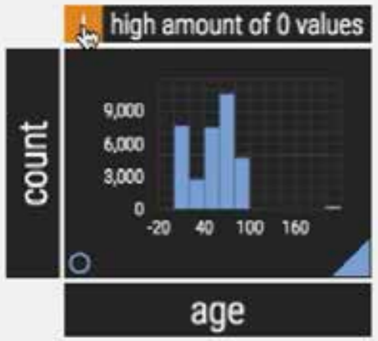


Execution Helper: speculatively execute queries

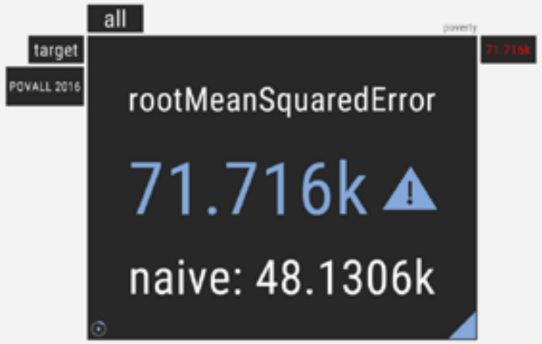
Skip Discovery Protector

Inform the user about potential problems

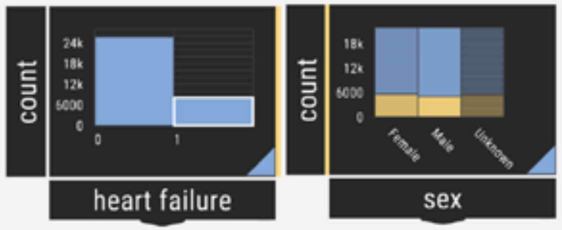
Data Problems



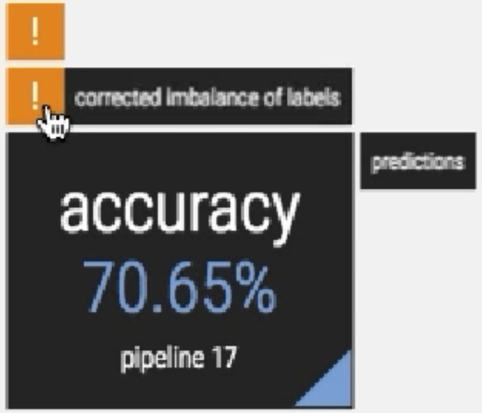
Accuracy Problems



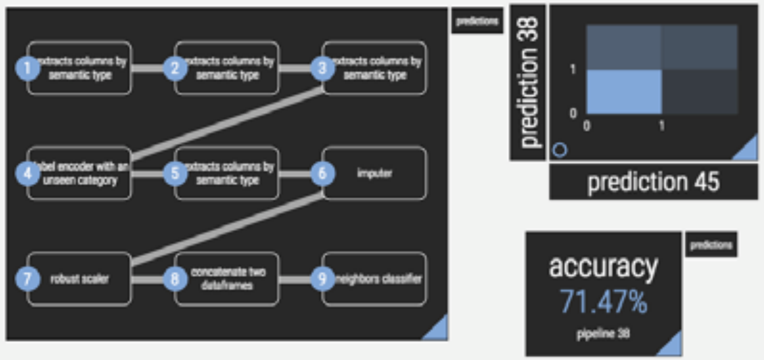
Simpson Paradox



Label Problems

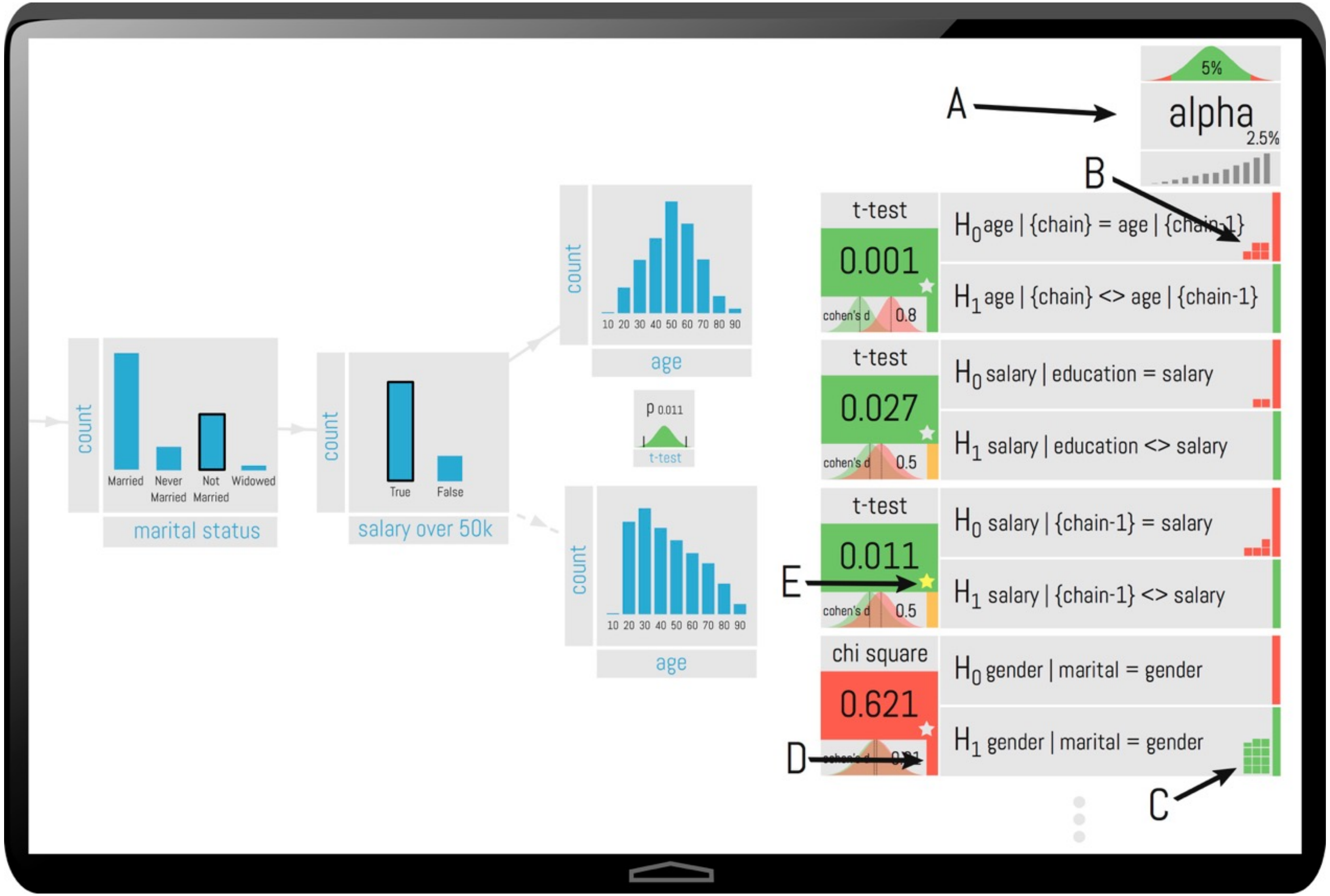


Model Inspection



many more…

QUDE – UI


A
B
C
D
E
5%
alpha
2.5%
t-test
0.001
cohen's d
0.8
H_0 age | {chain} = age | {chain-1}
H_1 age | {chain} <> age | {chain-1}
t-test
0.027
cohen's d
0.5
H_0 salary | education = salary
H_1 salary | education <> salary
t-test
0.011
cohen's d
0.5
H_0 salary | {chain-1} = salary
H_1 salary | {chain-1} <> salary
chi square
0.621
H_0 gender | marital = gender
H_1 gender | marital = gender
count
Married
Never Married
Not Married
Widowed
marital status
count
True
False
salary over 50k
count
10 20 30 40 50 60 70 80 90
age
p 0.011
t-test
count
10 20 30 40 50 60 70 80 90
age

Automatically Derive Hypothesis

- Currently simple heuristic:
    1. *Every visualization without any filter conditions is NOT a hypothesis unless the user makes it one.*
    2. *Every visualization with a filter condition is a hypothesis regarding its correlation*
    3. *If two visualization with the same but some negated filter conditions are put next to each other, it is a test with the null-hypothesis that there is no difference (supersedes 2.)*
- Much more work needed

What multi–hypothesis control technique should we use?

- Hold–out data set / Additional Tests
- Family–wise error (e.g., Bonferroni correction)
- False Discovery Rate (e.g., alpha–investing
- Permutation–based techniques
- Bayesian techniques (e.g., Bayesian FDR)
- Uniform Convergence and (Structural) Risk Minimization (more on that later)

False Discovery Rate

$$\text{FDR} = \mathrm{E}\left[\frac{V}{R}\right]^{*}$$

False discoveries

All discoveries

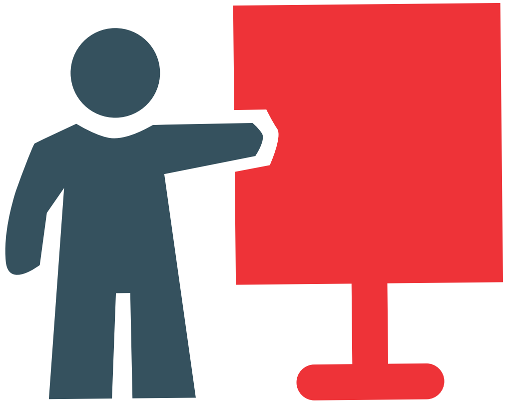
Benjamini-Hochberg procedure(BH)

1. Sort all p-values such that $p_1 < p_2 < \ldots. < p_n$
2. Determine the maximum k, such that $p_k < \frac{k}{m} \cdot \alpha$
3. Reject the null hypotheses corresponding to the p-values $p_1, p_2, \ldots, p_k$

** We define FDR to be zero when R = 0*

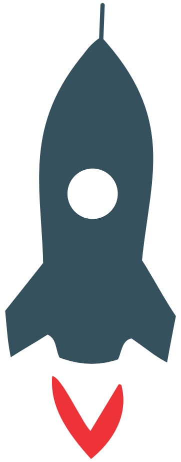
Three Core Technical Contributions

Vizdom
A Novel Interface
for Everyone



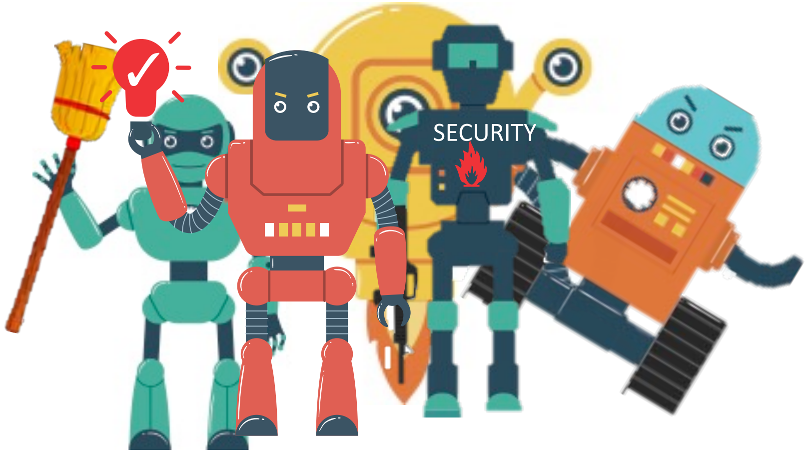
designed for data enthusiast (i.e., people with limited statistics and ML knowledge), domain experts, and data scientist alike.

IDEA
The Data Exploration
Accelerator



No waiting: immediately returns visual results for all operations and progressively refines them in the background

Smart Assistance
Towards Data Science
Automation



Protect users from common mistakes, point out data cleaning issues, help with building models

Beta-Testers

Northstar Publications

http://northstar.mit.edu/

Zeyuan Shang et al: Democratizing Data Science through Interactive Curation of ML Pipelines, SIGMOD 2019

Tim Kraska: Northstar: An Interactive Data Science System. PVLDB 11(12): 2150-2164 (2018)

Yeounoh Chung, Sacha Servan-Schreiber, Emanuel Zgraggen, Tim Kraska: Towards Quantifying Uncertainty in Data Analysis & Exploration. IEEE Data Eng. Bull.41(3): 15-27 (2018)

Zeyuan Shang et al : Towards Interactive Curation & Automatic Tuning of ML Pipelines. DEEM@SIGMOD 2018: 1:1-1:4

Yeounoh Chung, Michael Lind Mortensen, Carsten Binnig, Tim Kraska: Estimating the Impact of Unknown Unknowns on Aggregate Query Results. ACM Trans. Database Syst. 43(1): 3:1-3:37 (2018)

Emanuel Zgraggen, Zheguang Zhao, Robert C. Zeleznik, Tim Kraska: Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. CHI2018: 479

Alex Galakatos, Andrew Crotty, Emanuel Zgraggen, Carsten Binnig, Tim Kraska: Revisiting Reuse for Approximate Query Processing. PVLDB 10(10): 1142-1153 (2017)

Emanuel Zgraggen, Alex Galakatos, Andrew Crotty, Jean-Daniel Fekete, Tim Kraska: How Progressive Visualizations Affect Exploratory Analysis. IEEE Trans. Vis. Comput. Graph. 23(8): 1977-1987 (2017)

Carsten Binnig, Lorenzo De Stefani, Tim Kraska, Eli Upfal, Emanuel Zgraggen, Zheguang Zhao: Toward Sustainable Insights, or Why Polygamy is Bad for You. CIDR 2017

Yue Guo, Carsten Binnig, Tim Kraska: What you see is not what you get!: Detecting Simpson's Paradoxes during Data Exploration. HILDA@SIGMOD 2017: 2:1-2:5

Tim Kraska: Approximate Query Processing for Interactive Data Science. SIGMOD Conference 2017: 525

Zheguang Zhao, Lorenzo De Stefani, Emanuel Zgraggen, Carsten Binnig, Eli Upfal, Tim Kraska: Controlling False Discoveries During Interactive Data Exploration. SIGMOD Conference 2017: 527-540

Zheguang Zhao, Emanuel Zgraggen, Lorenzo De Stefani, Carsten Binnig, Eli Upfal, Tim Kraska: Safe Visual Data Exploration. SIGMOD Conference 2017: 1671-1674

Philipp Eichmann, Emanuel Zgraggen, Zheguang Zhao, Carsten Binnig, Tim Kraska: Towards a Benchmark for Interactive Data Exploration. IEEE Data Eng. Bull. 39(4): 50-61 (2016)

Muhammad El-Hindi, Zheguang Zhao, Carsten Binnig, Tim Kraska: VisTrees: fast indexes for interactive data exploration. HILDA@SIGMOD2016: 5

Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, Tim Kraska: The case for interactive data exploration accelerators (IDEAs).HILDA@SIGMOD 2016: 11

Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, Tim Kraska: Vizdom: Interactive Analytics through Pen and Touch. PVLDB 8(12): 2024-2027 (2015)

Evan R. Sparks, Ameet Talwalkar, Daniel Haas, Michael J. Franklin, Michael I. Jordan, Tim Kraska: Automating model search for large scale machine learning. SoCC 2015: 368-380

Emanuel Zgraggen, Robert C. Zeleznik, Steven M. Drucker: PanoramicData: Data Analysis through Pen & Touch. IEEE Trans. Vis. Comput. Graph. 20(12): 2112-2121 (2014)

Tim Kraska et al: MLbase: A Distributed Machine-learning System. CIDR 2013

Core Team

Prof. Tim Kraska
MIT, *Systems/ML*

Prof. Eli Upfal
Brown U., *Statistics/ML*

Dr. Emanuel Zgraggen
MIT, *HCI*

Benedetto Buratti
Brown U.
PhD student in ML

Philipp Eichmann
Brown U.
PhD student in HCI

Zeyuan Shang
MIT
PhD student in Systems/ML

Tim Kraska
<kraska@mit.edu>

MIT CSAIL

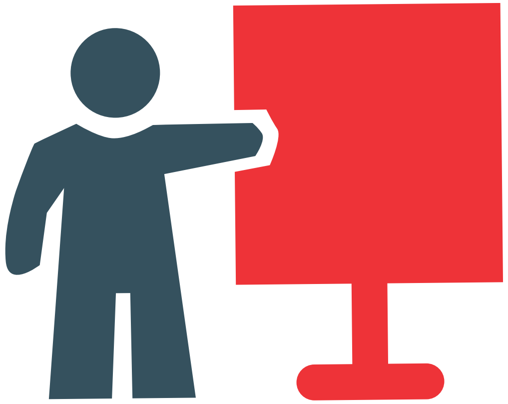
- Supporting Interactive Data Science requires to rethink the entire analytics stack.

- Northstar is a first Interactive Data Science System
  - With **Laax** we put the user experience first
  - **Davos**: an AQP engine for Interactive Data Science
  - **Alpine Meadows**: an Interactive ML-Autotuner (learning to learn)

http://northstar.mit.edu/

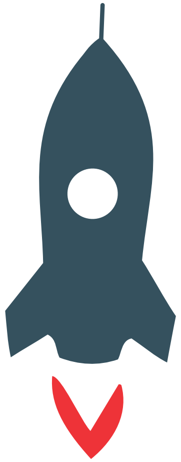
Three Core Technical Contributions

Laax
A Novel Interface for Everyone



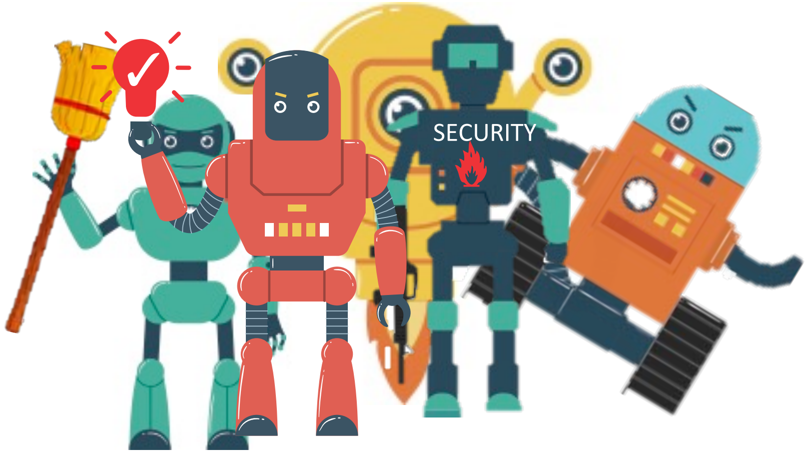
designed for data enthusiast (i.e., people with limited statistics and ML knowledge), domain experts, and data scientist alike.

Davos[2]**:** the first Interactive Data Exploration Accelerator



No waiting: immediately returns visual results for all operations and progressively refines them in the background

Smart Assistance
Towards Data Science Automation



Protect users from common mistakes, point out data cleaning issues, help with building models

[1]Laax is the successor of Vizdom, our first user interface. [2]Davos is the successor of IDEA, our first backend.

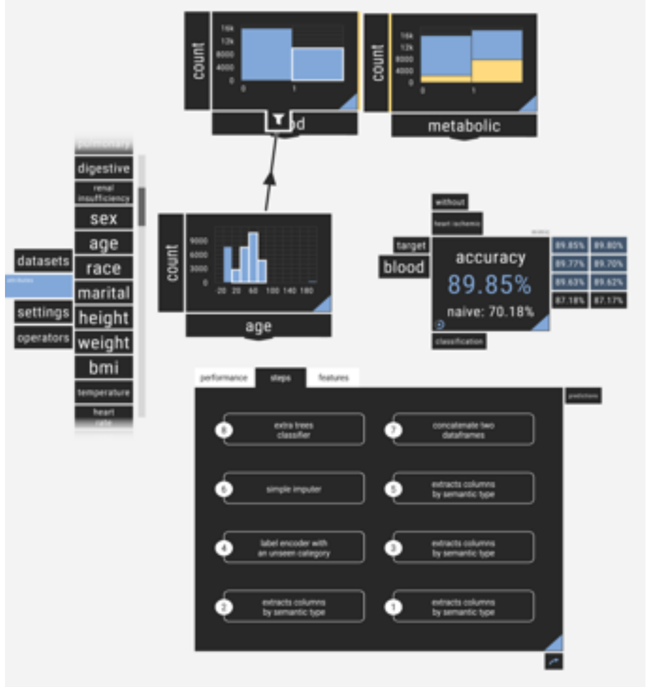
We created these new versions of the front- and backend based on the customer feedback we received from Shell, P&G, IGT, and others.

For a general overview of the different components see: Tim Kraska: Northstar: An Interactive Data Science System. PVLDB 11(12): 2150-2164 (2018) einblick analytics

3 Areas of Innovation

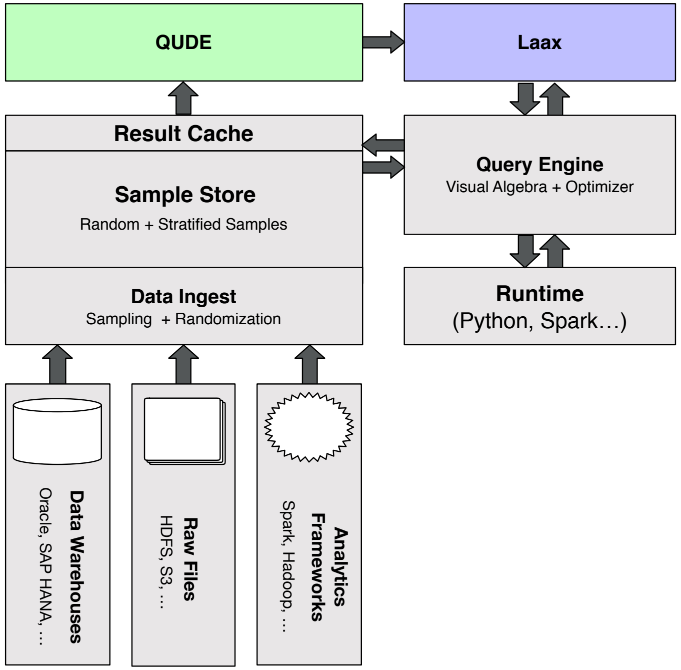
Laax[1]: a new data interaction paradigm

Enables playful interaction with data



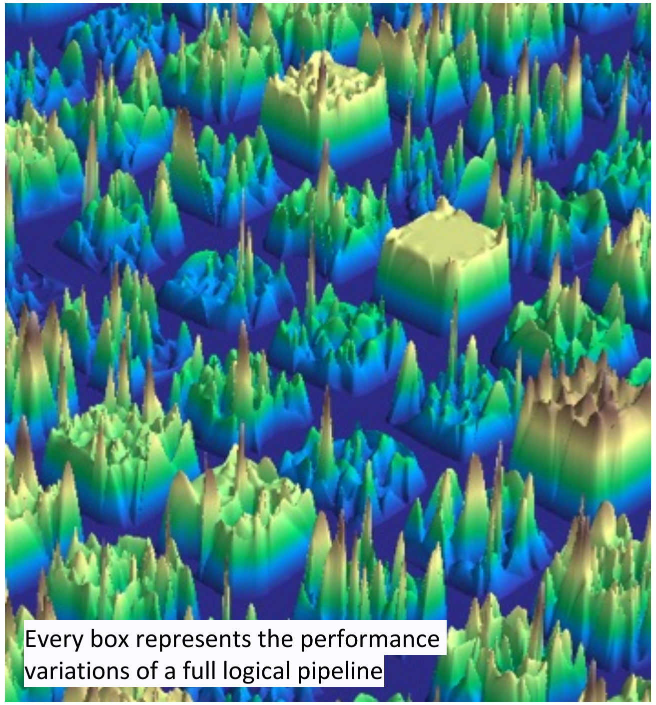
Davos[2]: the first Interactive Data Exploration Accelerator

Ensures interactive response times through progressive computation and approximation



Alpine Meadow: an interactive Data Mining and Auto-ML tool

Enables business analysts to do things only Data Scientist can do now



[1]Laax is the successor of Vizdom, our first user interface. [2]Davos is the successor of IDEA, our first backend.

We created these new versions of the front- and backend based on the customer feedback we received from Shell, P&G, IGT, and others.

For a general overview of the different components see: Tim Kraska: Northstar: An Interactive Data Science System. PVLDB 11(12): 2150-2164 (2018)

Co-designed to guarantee the best end-user experience

What did reviewer 2 say?

FDR simply reports the expected fraction of incorrectly rejected hypothesese, but doesn't tell you which of your accepted hypothesis is in fact reliable! …. **Familywise error rate (FWER) will be far more intuitive and useful to a naive user** as it bounds the probability of making one or more false discoveries (Type I errors).

False Discovery Rate

$$\text{FDR} = \mathrm{E}\left[\frac{V}{R}\right]^{*}$$

False discoveries

All discoveries

Benjamini-Hochberg procedure(BH)

1. Sort all p-values such that $p_1 < p_2 < .... < p_n$
2. Determine the maximum k, such that $p_k < \frac{k}{m} \cdot \alpha$
3. Reject the null hypotheses corresponding to the p-values $p_1 , p_2 , . . . , p_k$

** We define FDR to be zero when R = 0*

False Discovery Rate

F ... eries

FDR = E[...

Problem with the
Benjamini-Hochberg procedure(BH)
for Data Exploration???

Ben ... ure(BH)

1. S ... such that $p_1 < p_2 < \ldots < p_n$
2. De ... mine the maximum k, such that $p_k < \frac{k}{m} \cdot \alpha$
3. Reject the null hypotheses corresponding to the p-values $p_1, p_2, \ldots, p_k$

** We define FDR to be zero when R = 0*

False Discovery Rate

$$\text{FDR} = \mathrm{E}\left[\frac{V}{R}\right] \qquad\qquad \text{mFDR} = \frac{E[V]}{E[R]+\eta}$$

$\eta \text{ is commonly set to } 1 \text{ or } (1-\alpha)$

False Discovery Rate

$$\text{mFDR} = \frac{E[V]}{E[R]+\eta}$$

Under the complete null-hypothesis: E[V]=E[R]

$$\text{E[V]} \leq \frac{\alpha\,\eta}{(1-\alpha)}$$

If we set $\eta\ to\ (1-\alpha)$

$$\text{E[V]} \leq \alpha$$

→ Weak control of FWER

Alpha Investing

Initial alpha wealth $W(0) = \eta\alpha$

1. Set α_i for test t
2. Loose or gain budget

Return

Investment

$$W(t) - W(t-1) = \begin{cases} \boxed{\omega} & \text{if } p_j \leq \boxed{\alpha_j}, \\ \boxed{-\frac{\alpha_j}{1-\alpha_j}} & \text{if } p_j > \alpha_j \end{cases}$$

Loss

with $w < \alpha$

1.E. Aharoni and S. Rosset. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794, 2014.

IDE Alpha Investing Strategies

- **γ-fixed**
  invest a fixed fraction
  (think Bonferroni)

- **β-farsighted**
  at least a fraction **β** of the current **α**-wealth always remains
  (think incremental Bonferroni)

Investing Rule 1 β-farsighted

1: $W(0) = \eta\alpha$
2: **for** $j = 1, 2, \ldots$ **do**
3: $\quad \alpha_j = \min\left(\alpha, \frac{W(j-1)(1-\beta)}{1+W(j-1)(1-\beta)}\right)$
4: $\quad$ **if** $p(H_j) < \alpha_j$ **then**
5: $\quad\quad W(j) = W(j-1) + \omega$
6: $\quad$ **else**
7: $\quad\quad W(j) = W(j-1) - \frac{\alpha_j}{1-\alpha_j} = \beta W(j-1)$
8: $\quad$ **end if**
9: **end for**

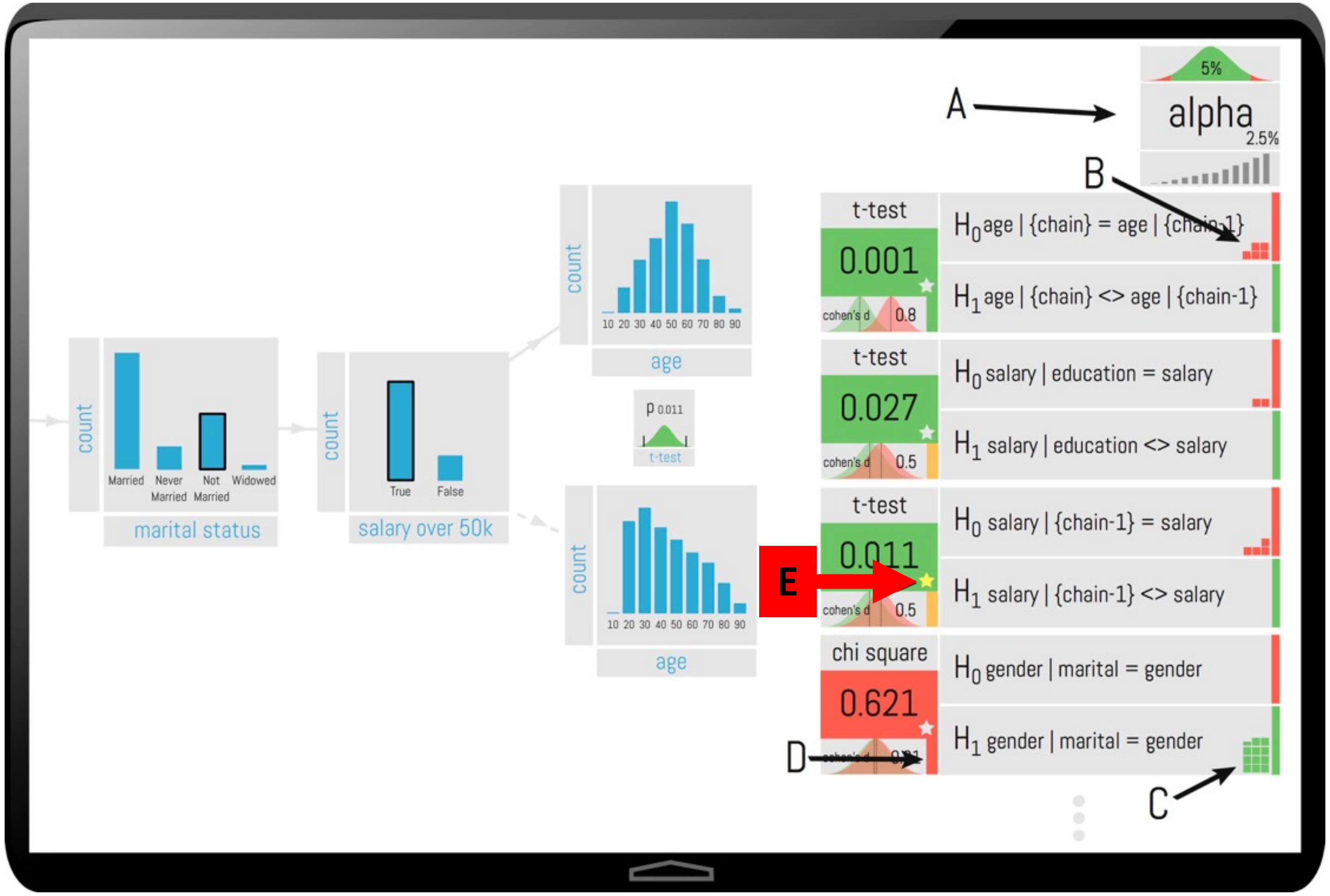
- **δ-Hopeful**
  expects that one of the next **δ** will be rejected

- **ε-Hybrid**
  adjust between **δ**-Hopeful and **γ**-fixed based on the randomness

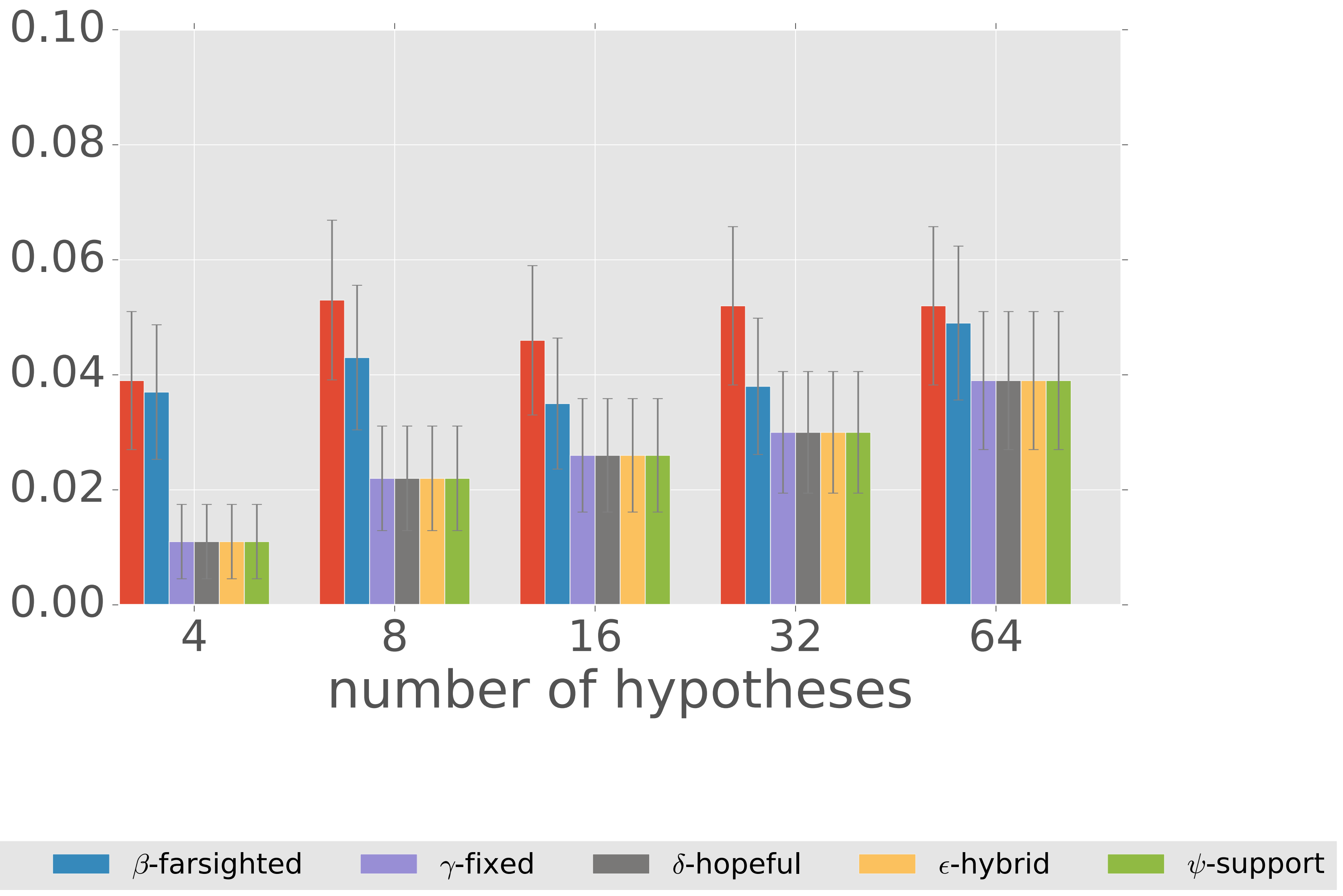
- **ψ-support**
  Invest based on how much support (i.e., records) a test considers

For more details see: [Zheguang Zhao, Lorenzo De Stefani, Emanuel Zgraggen, Carsten Binnig, Eli Upfal, Tim Kraska: **Controlling False Discoveries During Interactive Data Exploration.** CoRR abs/1612.01040 (2016)]

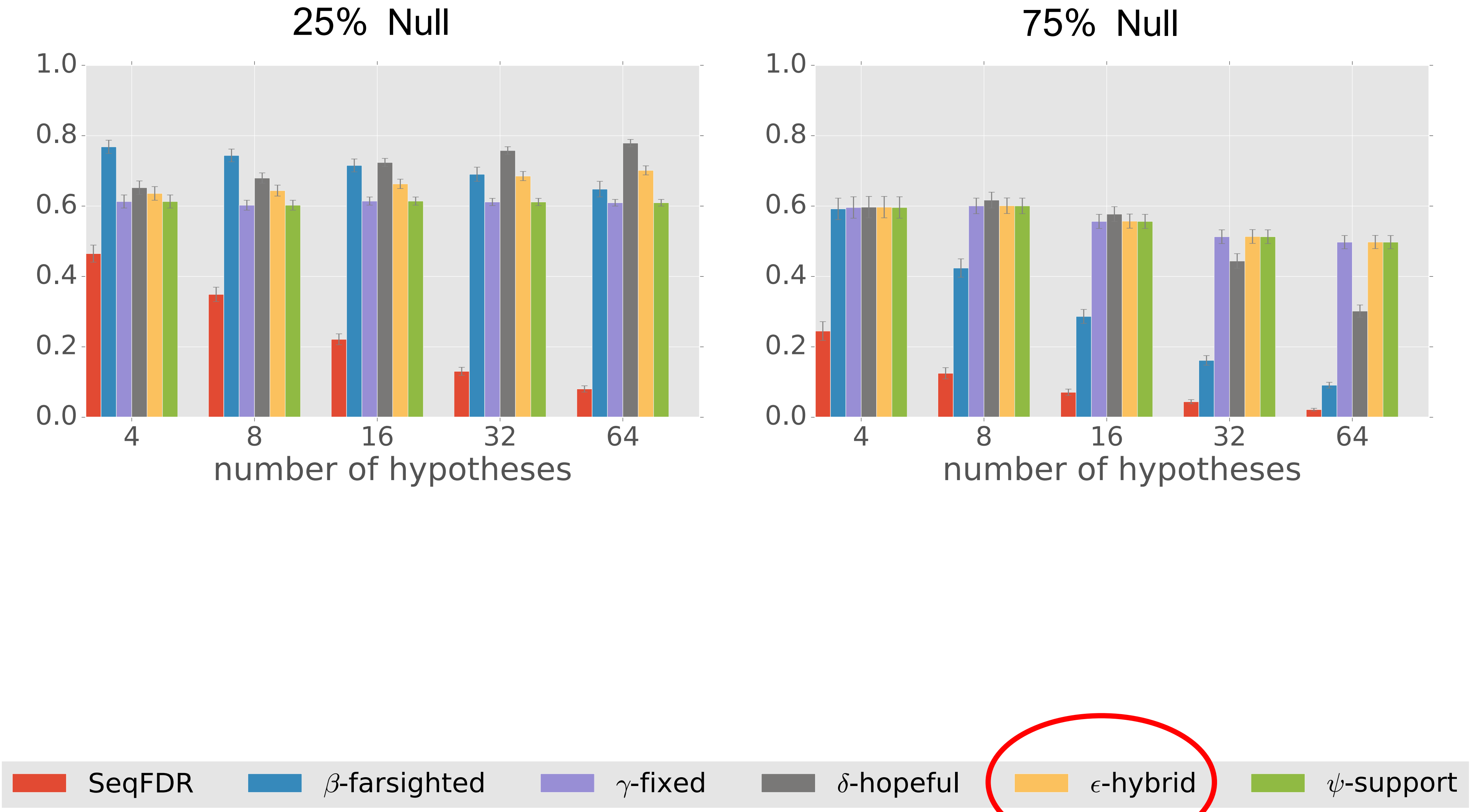
Marking the most important discoveries - what control to we get for them?



Complete Null-Hypothesis



Power


25% Null
75% Null
1.0
0.8
0.6
0.4
0.2
0.0
4
8
16
32
64
number of hypotheses
SeqFDR
β-farsighted
γ-fixed
δ-hopeful
ϵ-hybrid
ψ-support

Many Interesting Open Problems

We are just at the beginning

- Transparent hypothesis testing
  how to automatically derive what the hypothesis is the user is testing
- How to convey the meaning to the user
  (e.g., FDR vs family-wise error)
- Safe recommender techniques
  (we are currently exploring new techniques based VC-dimensions to control the error)
- Incremental multiple-hypothesis control techniques
  (for example, see "Controlling False Discoveries During Interactive Data Exploration" CoRR abs/1612.01040 how we use new alpha-investing policies to do that)
- Dependencies between hypothesis
  (this can safe "hypothesis budget")
- …

Error Types

- Unsufficient Support
- Approximation Error
- Data/Uncertainty Error
- Type I and Type II errors (false positives vs false negatives)
- Multi-Hypothesis Problem (part I of this talk)
- Simpson-Paradox (and related problems)
- Feature vs. data balance
- Unknown data error (part II of this talk)
- …