DATA+AI SUMMIT 2022

einblick

Data Boards:

A Collaborative and Interactive Space for Data Science

ORGANIZED BY databricks

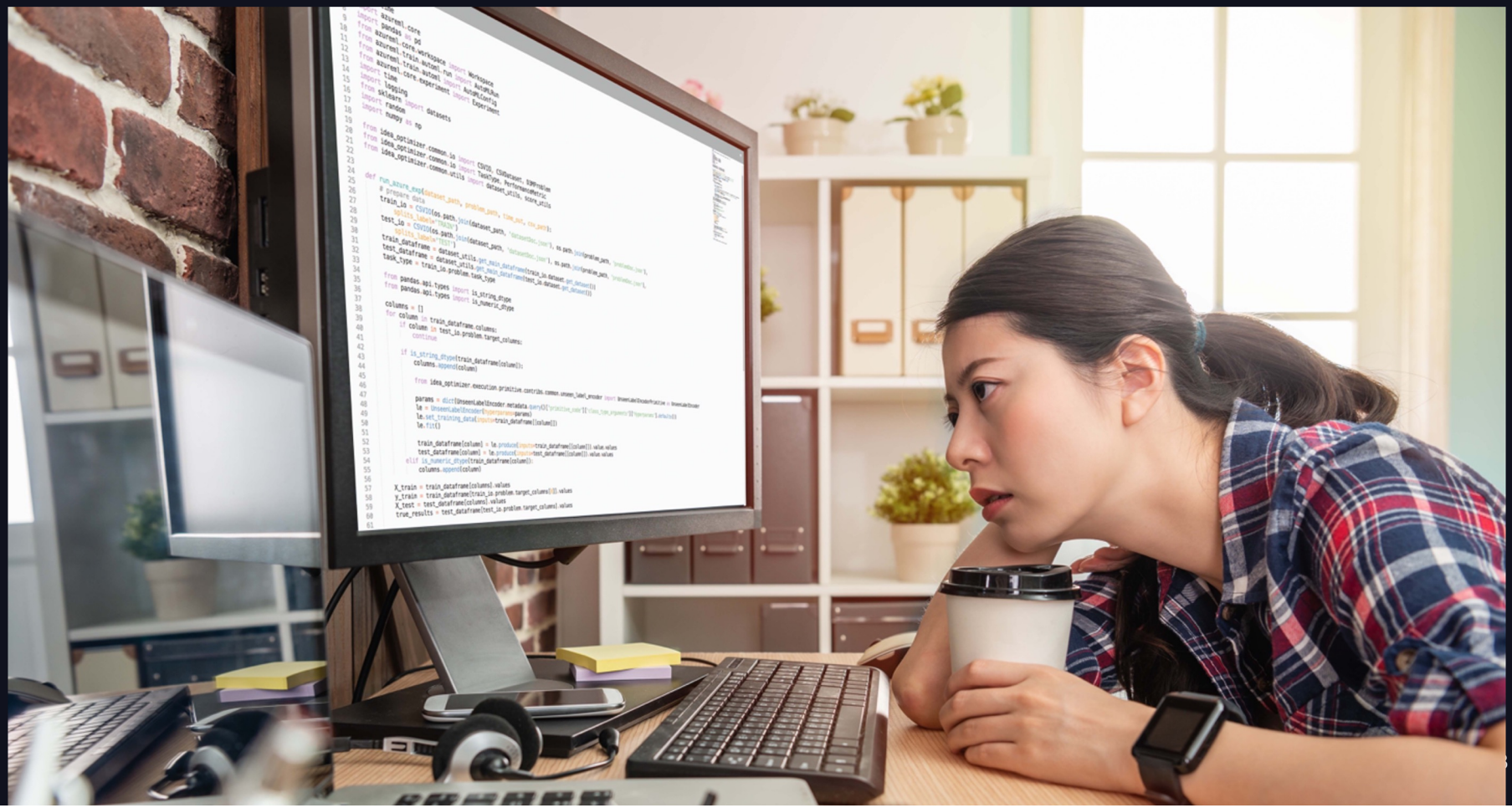
Tim Kraska
Professor, MIT & Co-founder, Einblick

Paul Yang
Data Scientist, Einblick

How Data Science should be

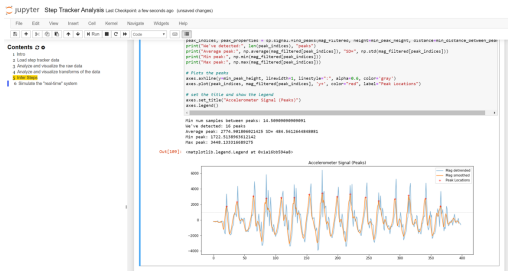
Vision video produced by Microsoft in 2015

Data Science Today



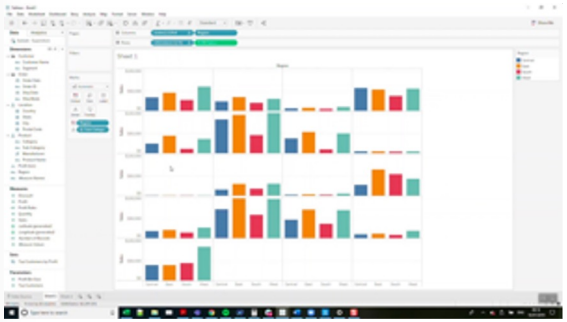
Data Science Today: Good Tools, a Bit Stuck

Notebooks



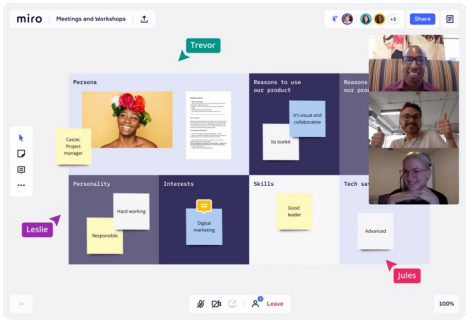
- Allows complex workflows
- No interactive feedback
- Hard to quickly iterate
- Not as easy to use
- Not collaborative

BI tools



- Easy to use
- Durable
- Hard to see the big picture
- Easy to get stuck
- Not collaborative

Collaboration tools



- Collaborative
- Not designed for analytics

Data Science Today: Microsoft Office is King

Data Engineers

"The table is ready, good luck!"

Domain Expert

"Here are a few bullets, good luck!"

Data Scientists

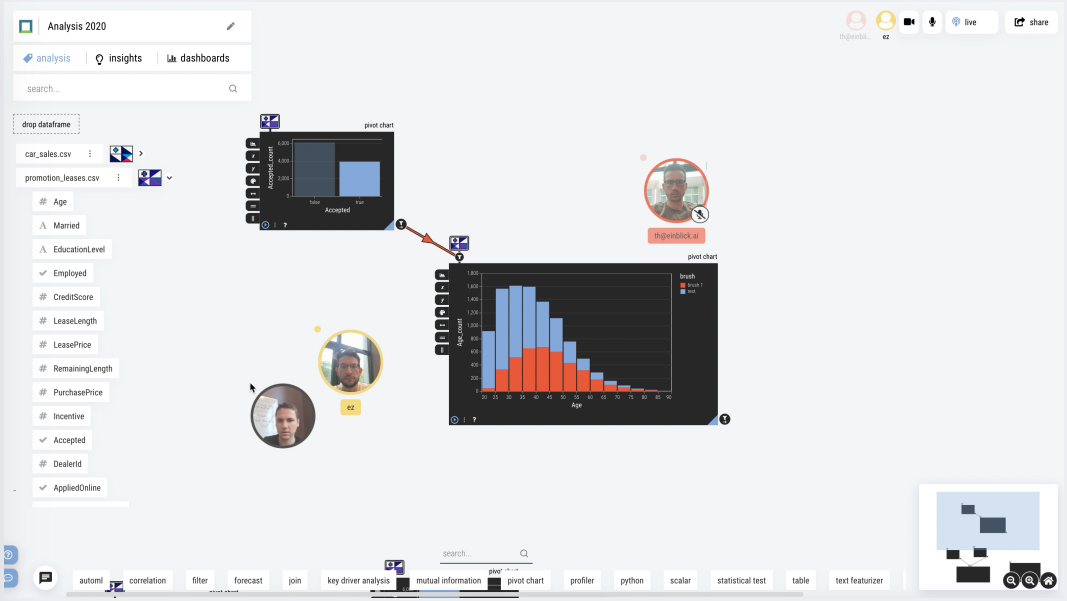
"Flip your powerpoint to slide 19"

Stakeholders

Data Boards: A new interaction paradigm

Data Scientists	Data Engineers	Decision Makers	Domain Experts

- **Rapid**: immediate feedback for all operations
- **Integrated**: covers all aspects from data ingest, over exploration, cleaning, to modeling
- **Limitless**: allows to build complex workflows with a few clicks and integrate new custom visual operators when needed
- **Collaborative**: designed for in-meeting and remote collaboration



northst*r

the data science platform

Northstar

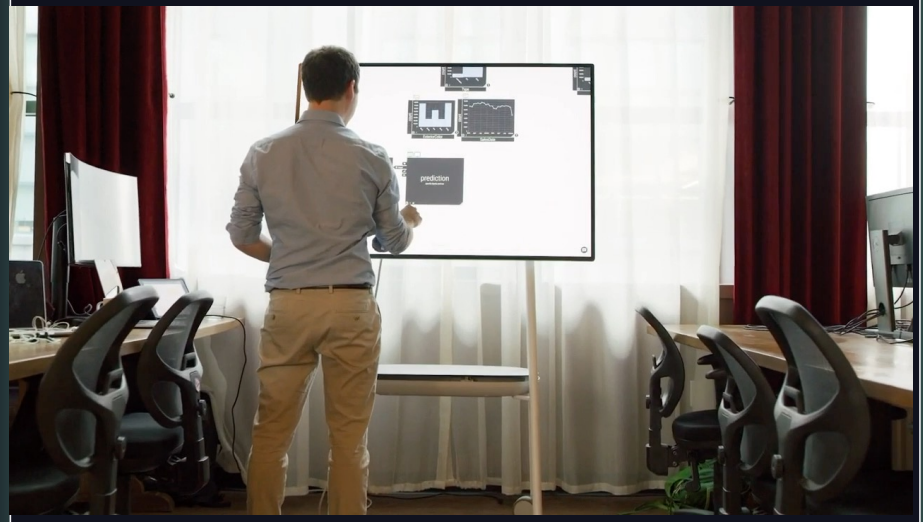
Built from the ground up for everyone involved in data science

We started with targeted User Experience

Data Boards

A new data interaction paradigm to collaborate

in person

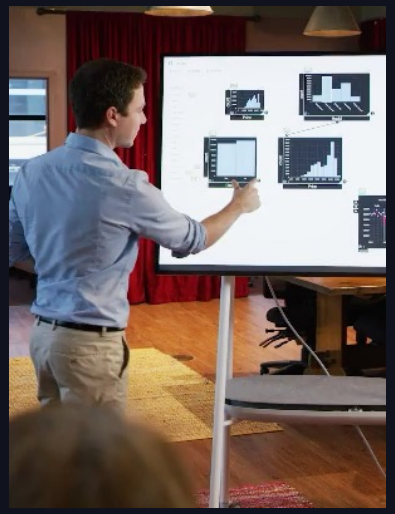


or remote

Northstar

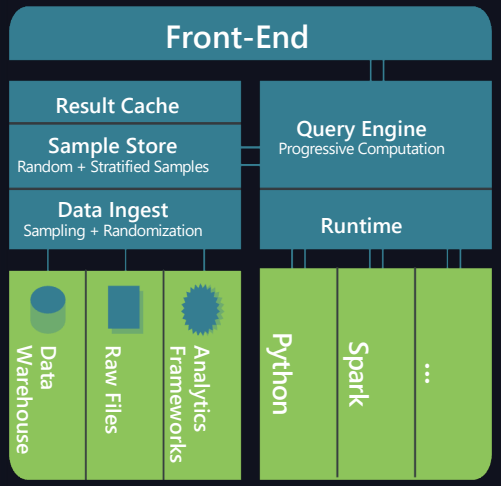
Built from the ground up for everyone involved in data science

We started with targeted User Experience



Data Boards

this required us to re-design the entire backend



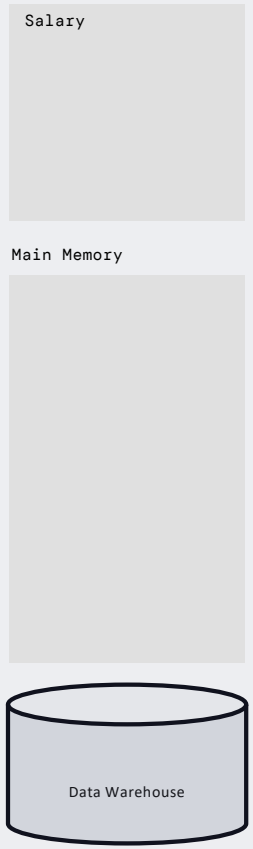
Progressive Approximation Engine

“Delays of 500ms incurred significant costs, decreasing user activity and data set coverage while reducing rates of observation, generalization and hypothesis.”

Zhicheng Liu, Jeffrey Heer:
The Effects of Interactive Latency on Exploratory Visual Analysis.
IEEE Trans. Vis. Comput. Graph. 20(12): 2122-2131 (2014)

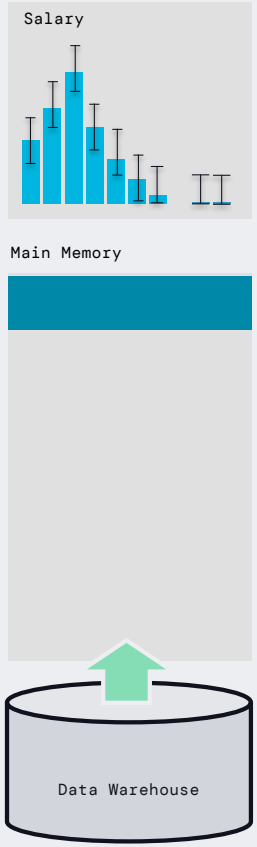
Engine: The first interactive data exploration accelerator

- Traditional storage, database, and ML systems are not designed for interactivity.
- Davos, Einblick's backend, is a **Sample Management Systems** with the goal to accelerate queries over any data source to enable Interactive Data Science
- Davos acts as an intelligent sampling, progressive computation, and caching layer



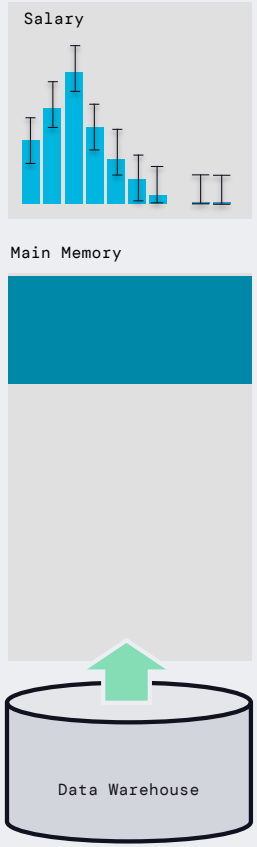
Engine: The first interactive data exploration accelerator

- Traditional storage, database, and ML systems are not designed for interactivity.
- Davos, Einblick's backend, is a **Sample Management Systems** with the goal to accelerate queries over any data source to enable Interactive Data Science
- Davos acts as an intelligent sampling, progressive computation, and caching layer



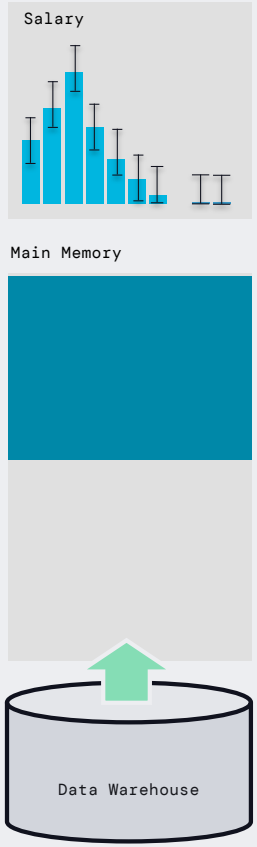
Engine: The first interactive data exploration accelerator

- Traditional storage, database, and ML systems are not designed for interactivity.
- Davos, Einblick's backend, is a **Sample Management Systems** with the goal to accelerate queries over any data source to enable Interactive Data Science
- Davos acts as an intelligent sampling, progressive computation, and caching layer



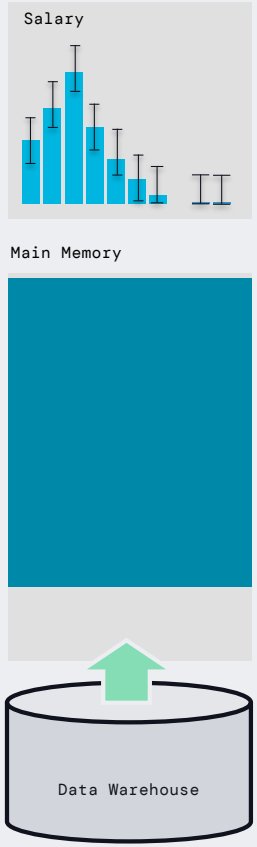
Engine: The first interactive data exploration accelerator

- Traditional storage, database, and ML systems are not designed for interactivity.
- Davos, Einblick's backend, is a **Sample Management Systems** with the goal to accelerate queries over any data source to enable Interactive Data Science
- Davos acts as an intelligent sampling, progressive computation, and caching layer



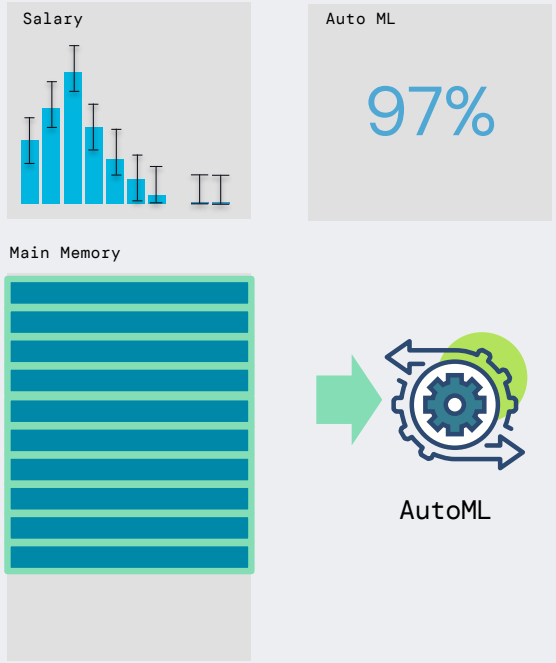
Engine: The first interactive data exploration accelerator

- Traditional storage, database, and ML systems are not designed for interactivity.
- Davos, Einblick's backend, is a **Sample Management Systems** with the goal to accelerate queries over any data source to enable Interactive Data Science
- Davos acts as an intelligent sampling, progressive computation, and caching layer



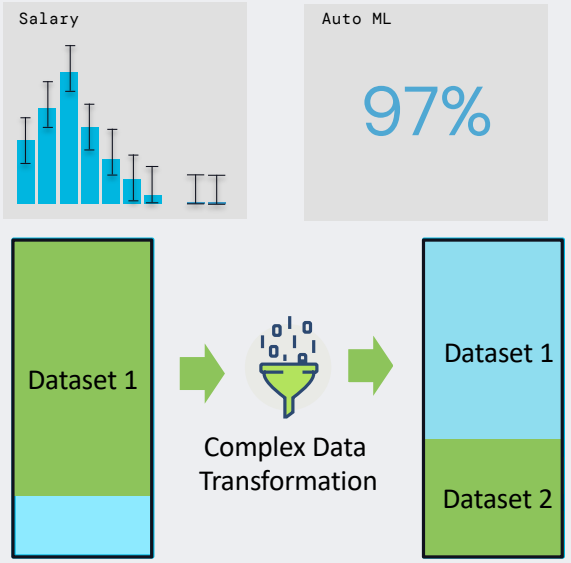
Engine: The first interactive data exploration accelerator

- Traditional storage, database, and ML systems are not designed for interactivity.
- Davos, Einblick's backend, is a **Sample Management Systems** with the goal to accelerate queries over any data source to enable Interactive Data Science
- Davos acts as an intelligent sampling, progressive computation, and caching layer
- Custom Python is also made progressive in the same way



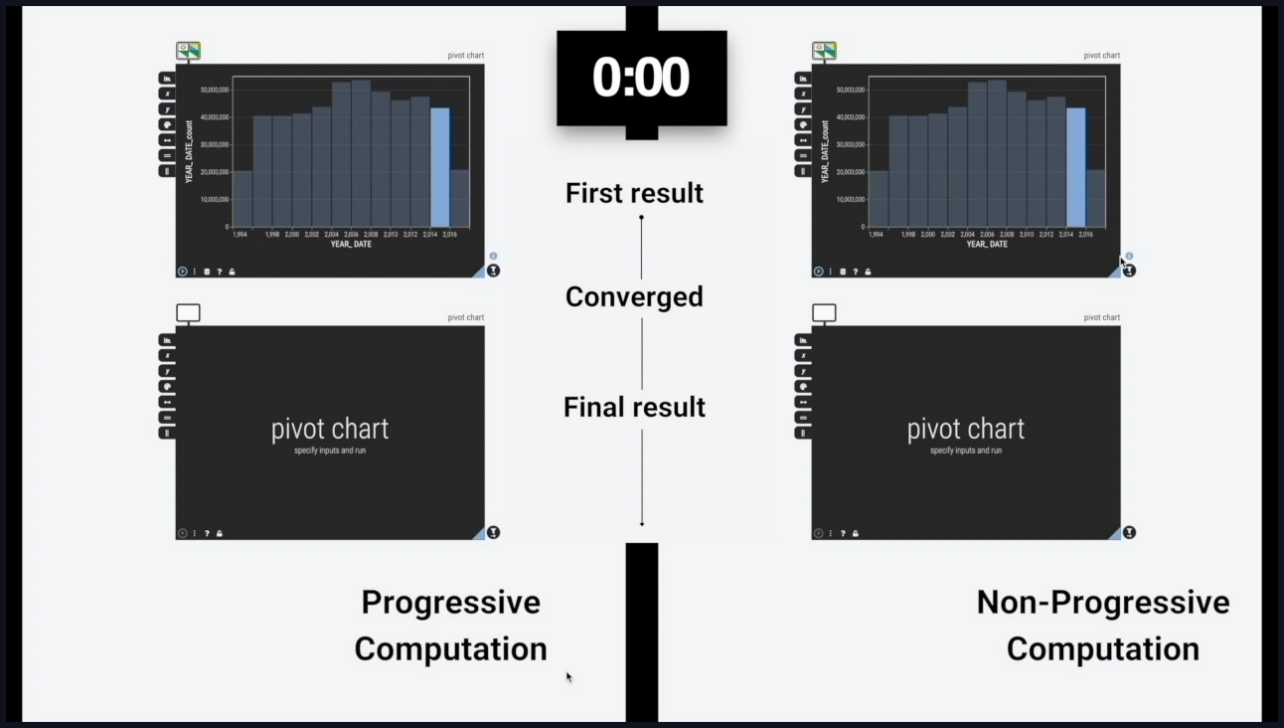
Engine: The first interactive data exploration accelerator

- Traditional storage, database, and ML systems are not designed for interactivity.
- Davos, Einblick's backend, is a **Sample Management Systems** with the goal to accelerate queries over any data source to enable Interactive Data Science
- Davos acts as an intelligent sampling, progressive computation, and caching layer
- Custom Python is also made progressive in the same way
- Guarantees interactive speeds over very large datasets and complex operations



Result: a new type of processing model combines streaming, with sampling, and the traditional query processing model

The impact of our Progressive Approximation Engine

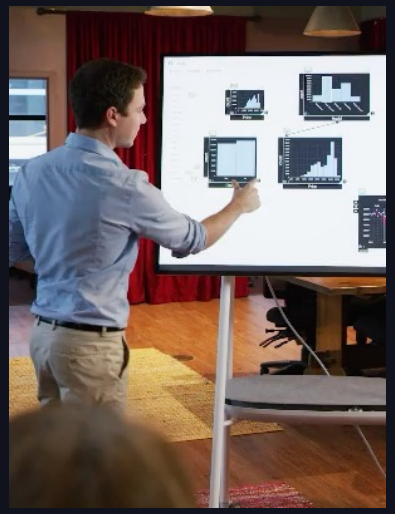


Flight dataset (scaled to 30GB)

Northstar

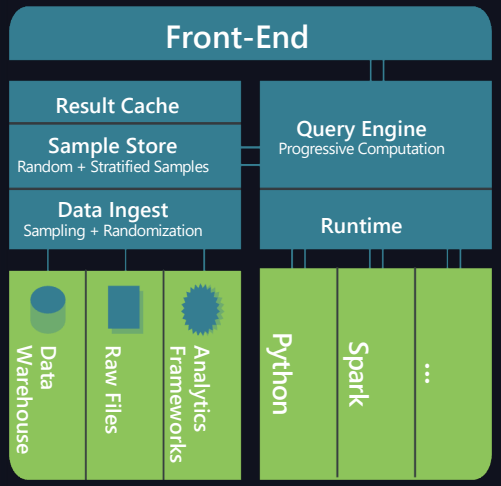
Built from the ground up for everyone involved in data science

We started with targeted User Experience



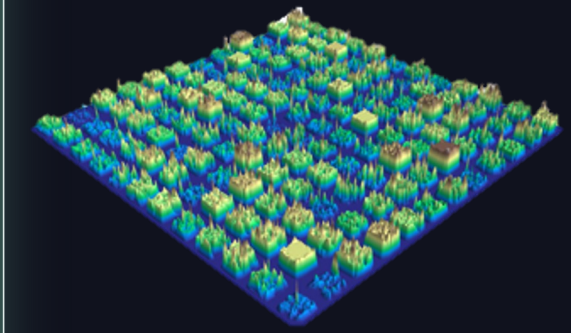
Data Boards

this required us to re-design the entire backend



Progressive Approximation Engine

and integrate new algorithms to accelerate workflows



Interactive KDA and AutoML

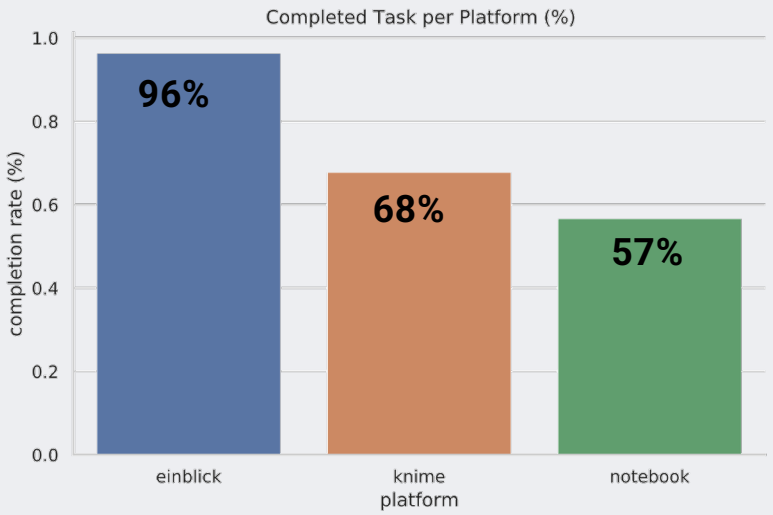
Why does it all matter?
Productivity Increase

Competitive User Study

- Recruited 21 Data Scientist with at least 1 year of experience in Data Science and Python knowledge.
- Each participant was trained in the respective tool for 1h.
- Each one was tasked to solve 3 different task using 3 different tools
  - Tools: **Einblick, Python Notebook, Visual No-Code Workflow Tool**
  - Task types: **Customer Churn**, **Employee Retention**, and **Predictive Healthcare**
  - Each task required to do data ingestion, data transformation, data visualization, data mining and modeling.
  - Each task had a **1h time limit**

Task completion rate per platform

Percentage of completed task per Platform (higher is better)



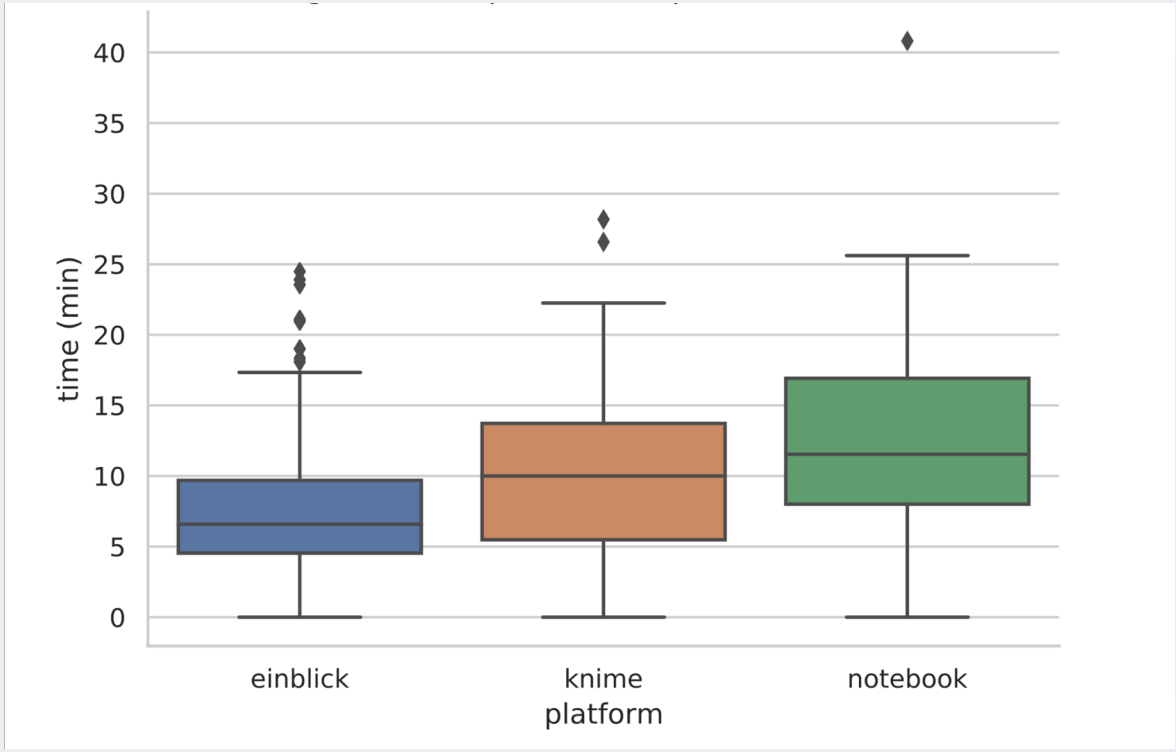
Improvement per step type over competing platforms

	Notebook	Data Science Workflow Tool
Profiling Data	**2.9x**	**1.8x**
Predictive Modeling	**2.1x**	**2.3x**
Pattern Mining	**#N/A***	**3.1x***

* One participant was able to complete the data mining task using the workflow tool, none using Python notebook.

1.5 – 3x increase in capabilities depending on the task

Time to finish task (lower is better)



~1.5-2x faster to transform data, create visualization, build models, etc.

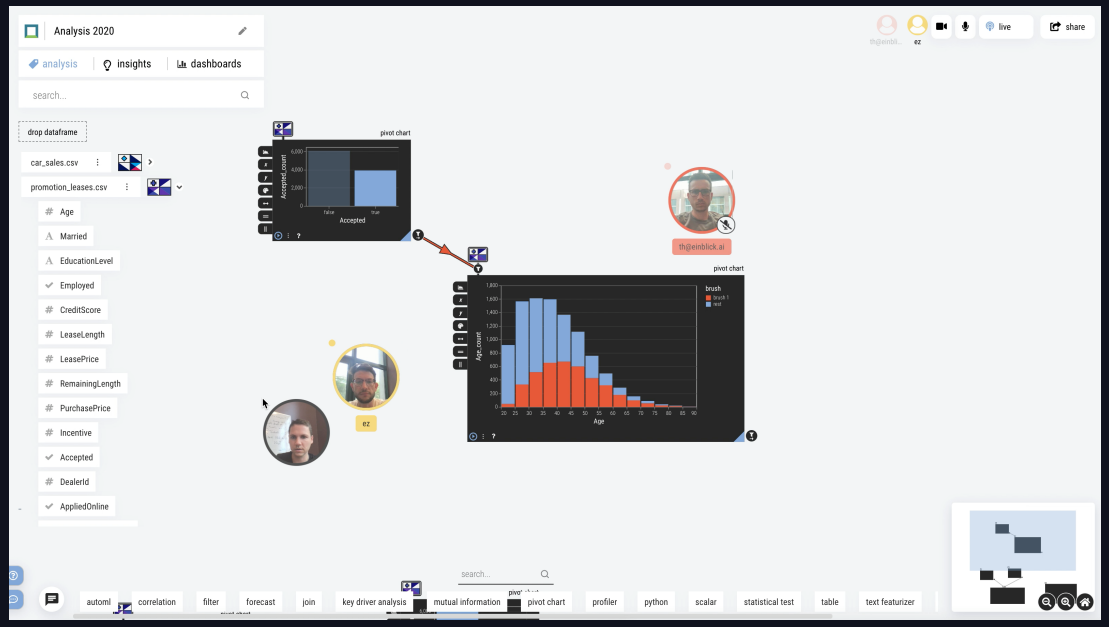
Jul 25 - 28, 2022

No Code Analytics and AI

https://professional.mit.edu/course-catalog/no-code-analytics-and-ai

Date(s): **Jul 25 - 28, 2022**
Registration Deadline: **Jul 11, 2022**
Location: **Live Virtual**
Course Length: **4 half-days**

Demo

Try it yourself: https://einblick.ai

Email me: py@einblick.ai