BROWN

TECHNISCHE UNIVERSITÄT DARMSTADT

TOWARDS INTERACTIVE DATA EXPLORATION

CARSTEN BINNIG

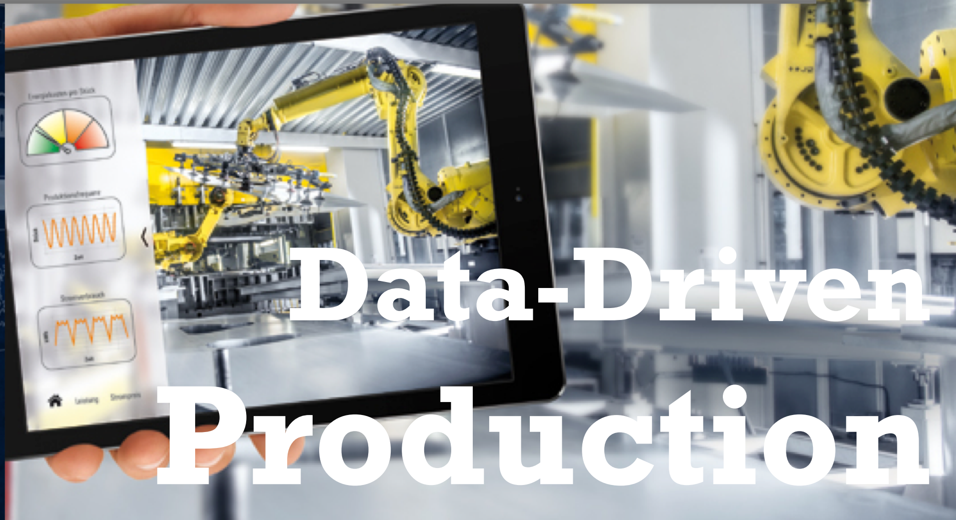
DATA MANAGEMENT LAB

Data-Driven
Marketing

DATA EXPLORATION:

Important first step to understand

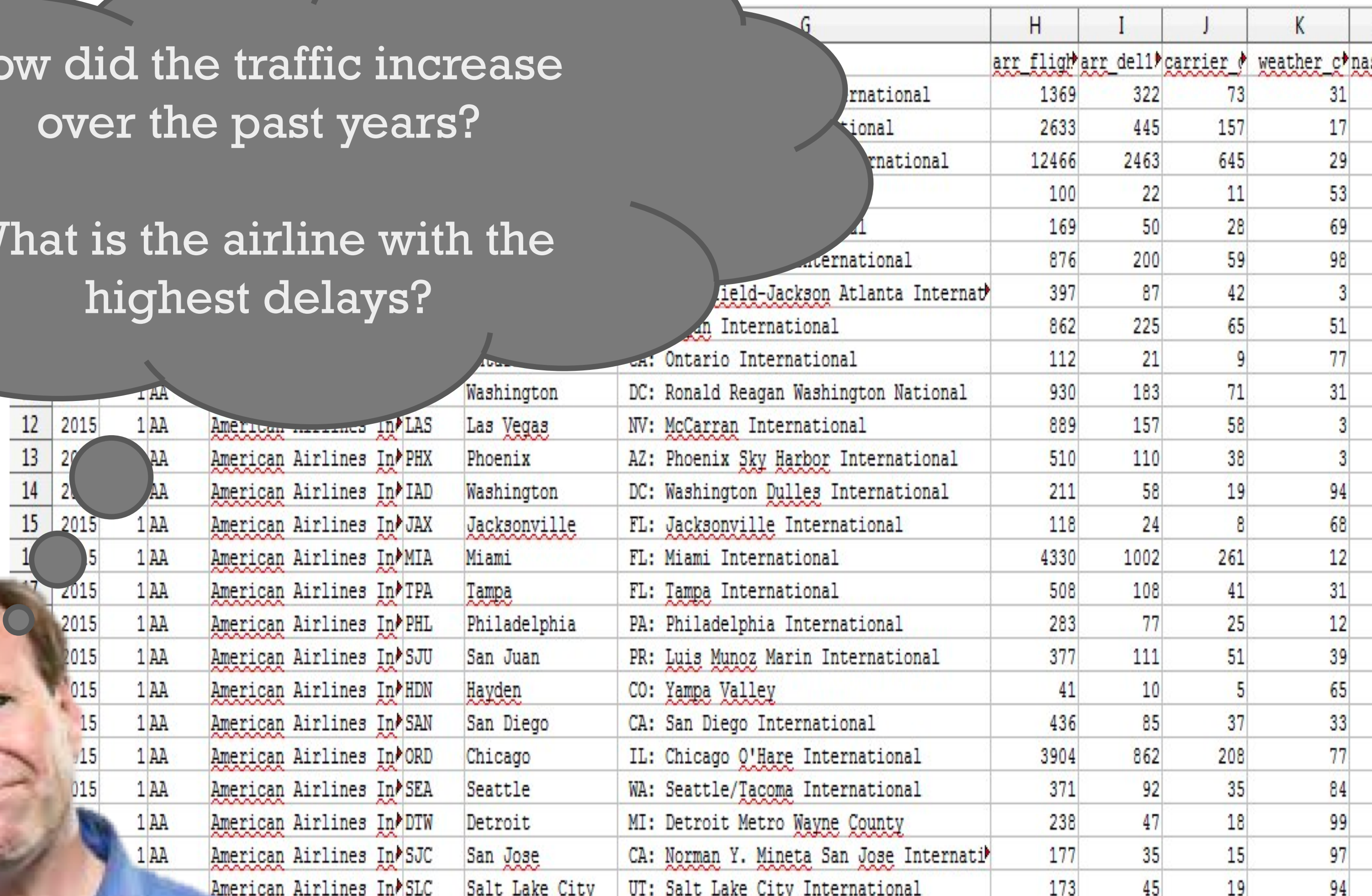
BIG DATA



EXAMPLE: DATA JOURNALISM

How did the traffic increase over the past years?

What is the airline with the highest delays?



							G	H	I	J	K	L
								arr_fligh	arr_del1	carrier_	weather_c	nas_ct
							rnational	1369	322	73	31	8
							ional	2633	445	157	17	25
							rnational	12466	2463	645	29	64
								100	22	11	53	0
							l	169	50	28	69	0
							ternational	876	200	59	98	10
							ield-Jackson Atlanta Internat	397	87	42	3	0
							n International	862	225	65	51	17
							A: Ontario International	112	21	9	77	0
		1	AA			Washington	DC: Ronald Reagan Washington National	930	183	71	31	11
12	2015	1	AA	American Airlines In	LAS	Las Vegas	NV: McCarran International	889	157	58	3	10
13	2		AA	American Airlines In	PHX	Phoenix	AZ: Phoenix Sky Harbor International	510	110	38	3	11
14	2		AA	American Airlines In	IAD	Washington	DC: Washington Dulles International	211	58	19	94	0
15	2015	1	AA	American Airlines In	JAX	Jacksonville	FL: Jacksonville International	118	24	8	68	0
1	5	1	AA	American Airlines In	MIA	Miami	FL: Miami International	4330	1002	261	12	36
	2015	1	AA	American Airlines In	TPA	Tampa	FL: Tampa International	508	108	41	31	6
	2015	1	AA	American Airlines In	PHL	Philadelphia	PA: Philadelphia International	283	77	25	12	0
	2015	1	AA	American Airlines In	SJU	San Juan	PR: Luis Munoz Marin International	377	111	51	39	7
	015	1	AA	American Airlines In	HDN	Hayden	CO: Yampa Valley	41	10	5	65	0
	15	1	AA	American Airlines In	SAN	San Diego	CA: San Diego International	436	85	37	33	6
	15	1	AA	American Airlines In	ORD	Chicago	IL: Chicago O'Hare International	3904	862	208	77	16
	015	1	AA	American Airlines In	SEA	Seattle	WA: Seattle/Tacoma International	371	92	35	84	7
		1	AA	American Airlines In	DTW	Detroit	MI: Detroit Metro Wayne County	238	47	18	99	0
		1	AA	American Airlines In	SJC	San Jose	CA: Norman Y. Mineta San Jose Internati	177	35	15	97	2
				American Airlines In	SLC	Salt Lake City	UT: Salt Lake City International	173	45	19	94	0

Airline Traffic Data

jeff@nytimes.com

Minority Report (2002)

DATA EXPLORATION VISION

TODAY'S USER INTERFACES

TODAY'S USER INTERFACES

… AND THE BIG DATA SYSTEMS?

Flink

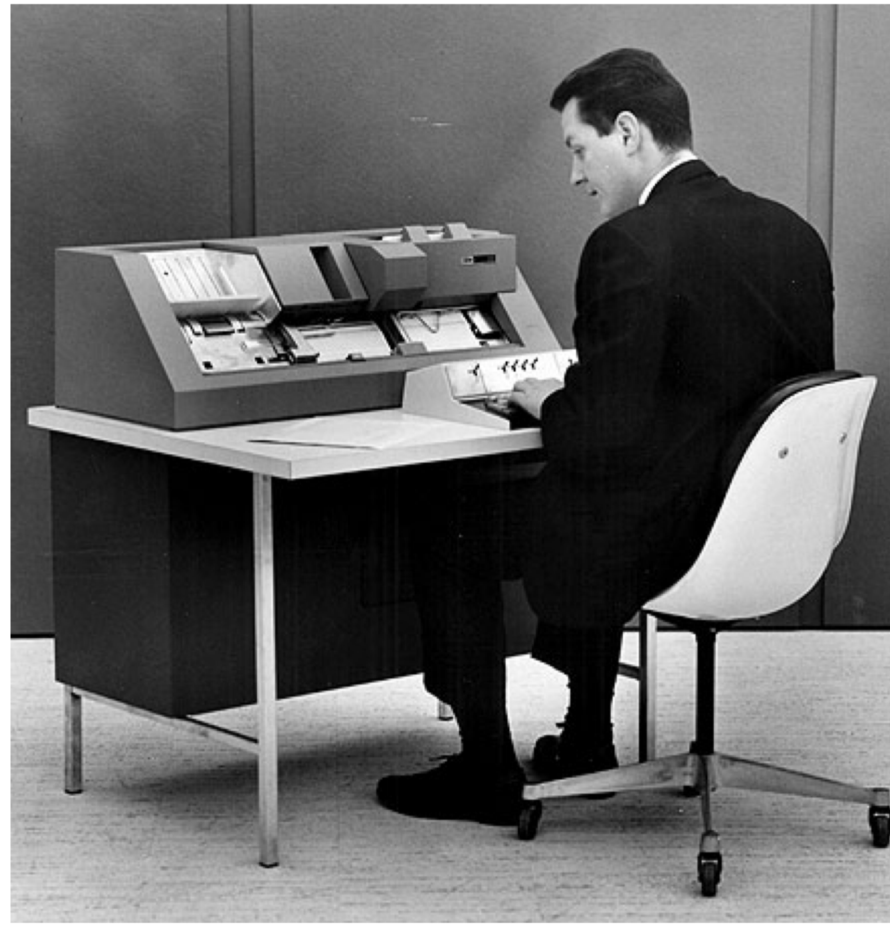
mxnet

Spark

hadoop

TensorFlow

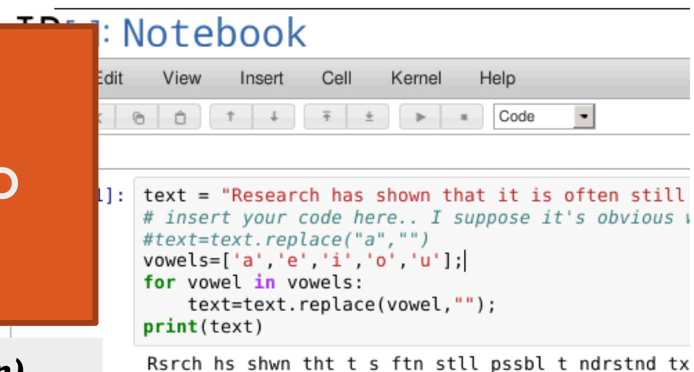
AsterixDB

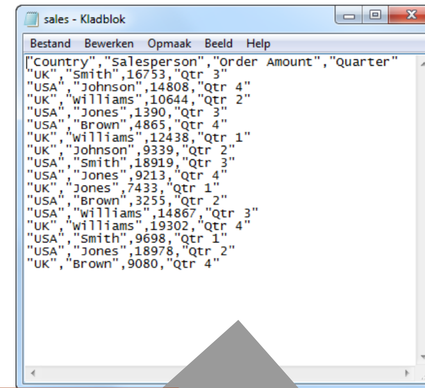


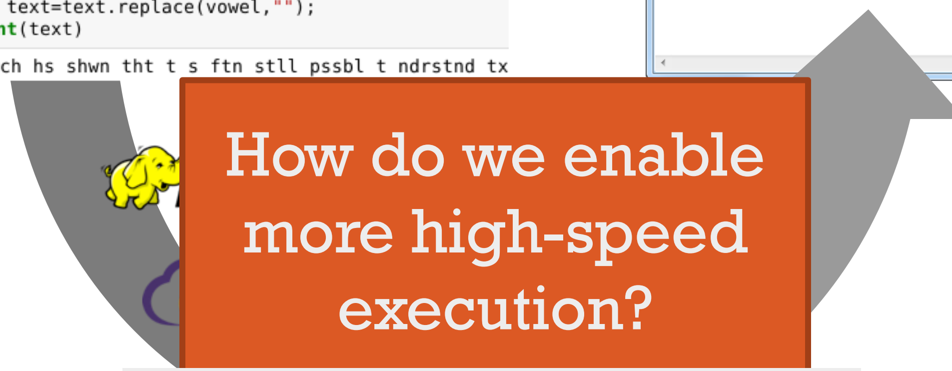
A TYPICAL EXPLORATION PIPELINE

How do query interfaces need to change?

Vizdom (Visual Exploration)
DBPal / EchoQuery (NL Interface)
IDEBench (Benchmarking)

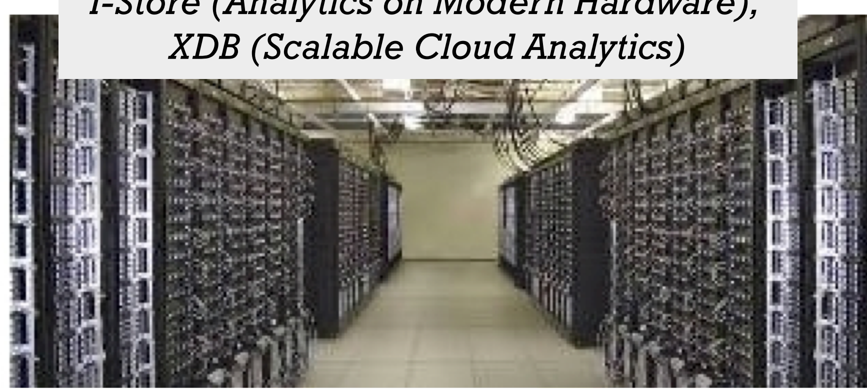






How do we enable more high-speed execution?

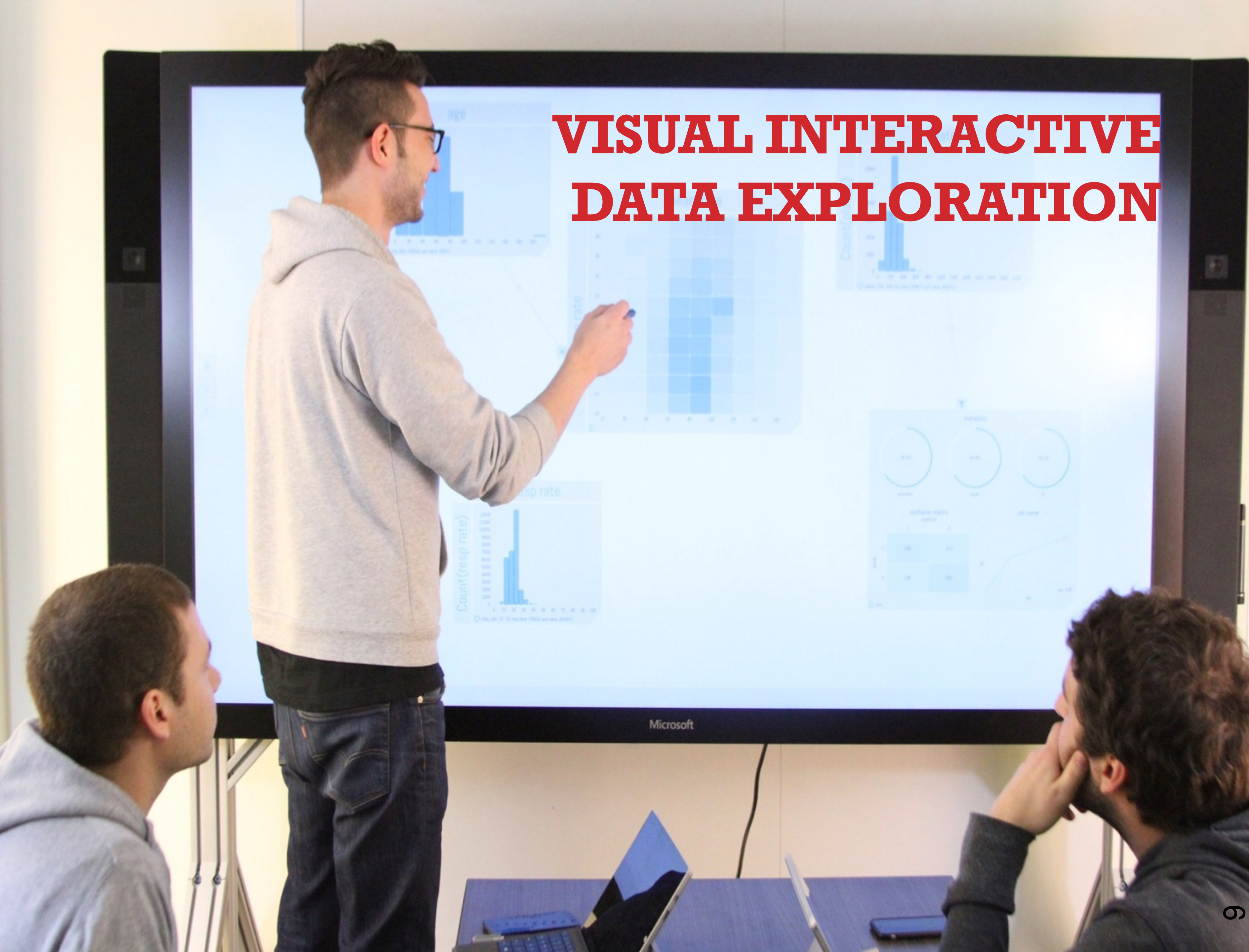
IDEA (Interactive Query Processing)
I-Store (Analytics on Modern Hardware),
XDB (Scalable Cloud Analytics)



How do we reduce data cleaning costs?

UnkownUnkowns (Data Quality)
IncMap (Schema Mapping)
Sherlock (Text Summerization)

VISUAL INTERACTIVE DATA EXPLORATION



BROWN

vizdom

Interactive Analytics through Pen and Touch

Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, Tim Kraska


Challenges &
opportunities

CHALLENGE: INTERACTIVITY

Provide interactive response times for queries even on very large data sets (e.g., <500ms)

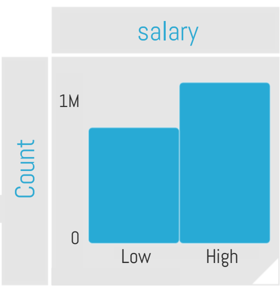
The Effects of Interactive Latency on Exploratory Visual Analysis

Zhicheng Liu and Jeffrey Heer

In this research, we have found that interactive latency can play an important role in shaping user behavior and impacts the outcomes of exploratory visual analysis. Delays of 500ms incurred significant costs, decreasing user activity and data set coverage while reducing rates of observation, generalization and hypothesis. Moreover, initial exposure to higher latency interactions resulted in reduced rates of observation and generalization during subsequent analysis sessions in which full system performance was restored.

CHALLENGE: AD-HOC QUERIES

Provide ad-hoc intuitive query interfaces AND no pre-defined static reports or low-level query interfaces

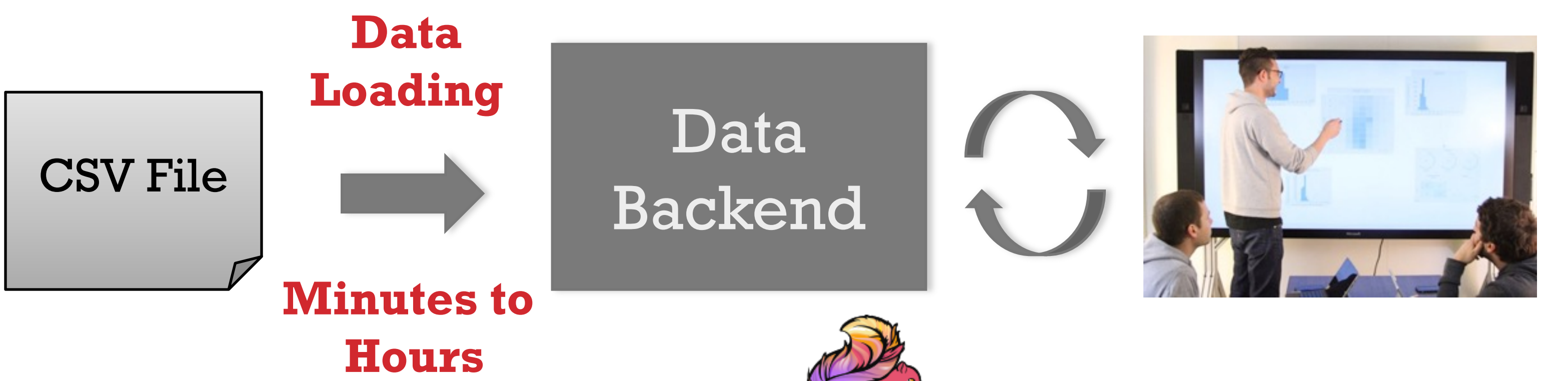


Census Data

Next steps: influence of education, marital status, …?

CHALLENGE: CONNECT & EXPLORE

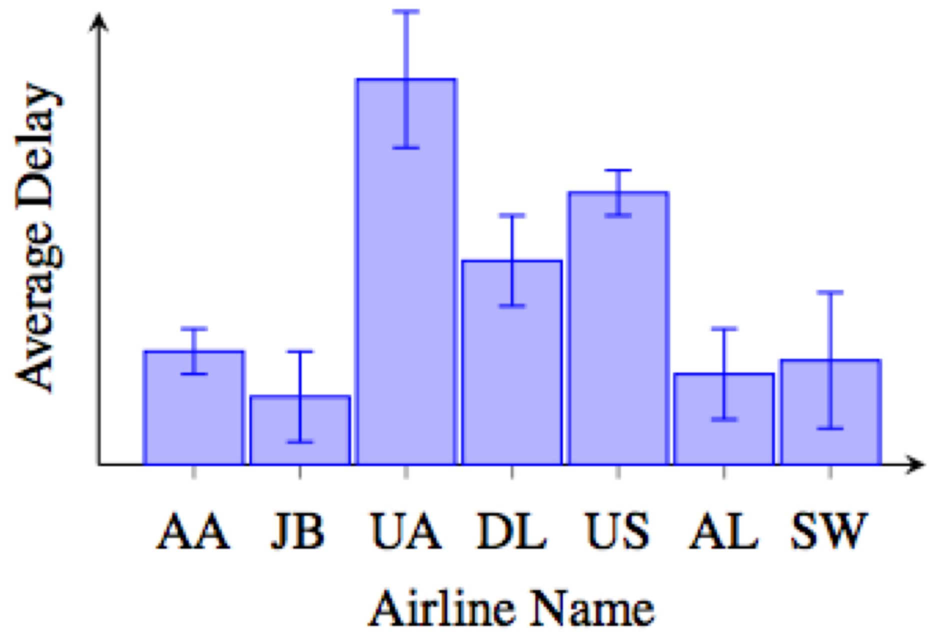
Users want to directly explore new data without waiting for data being loaded (or even cleaned) before



AsterixDB
Flink
hadoop
Spark

OPPORTUNITY: DECISION MAKING

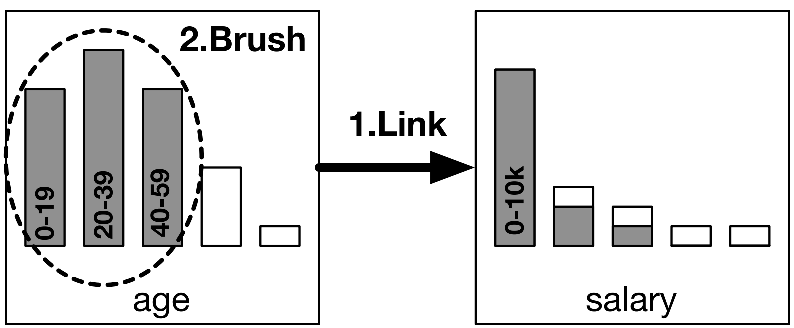
Exact results are often not needed to make decisions



Optimization: Approximate results are good enough

OPPORTUNITY: INCREMENTAL QUERIES


```
SELECT SUM(salary)
FROM census
GROUP BY salary-buckets
```




```
SELECT SUM(salary)
FROM census
WHERE age < 60
GROUP BY salary-buckets
```

Optimization: Reuse results / compute only the diff!

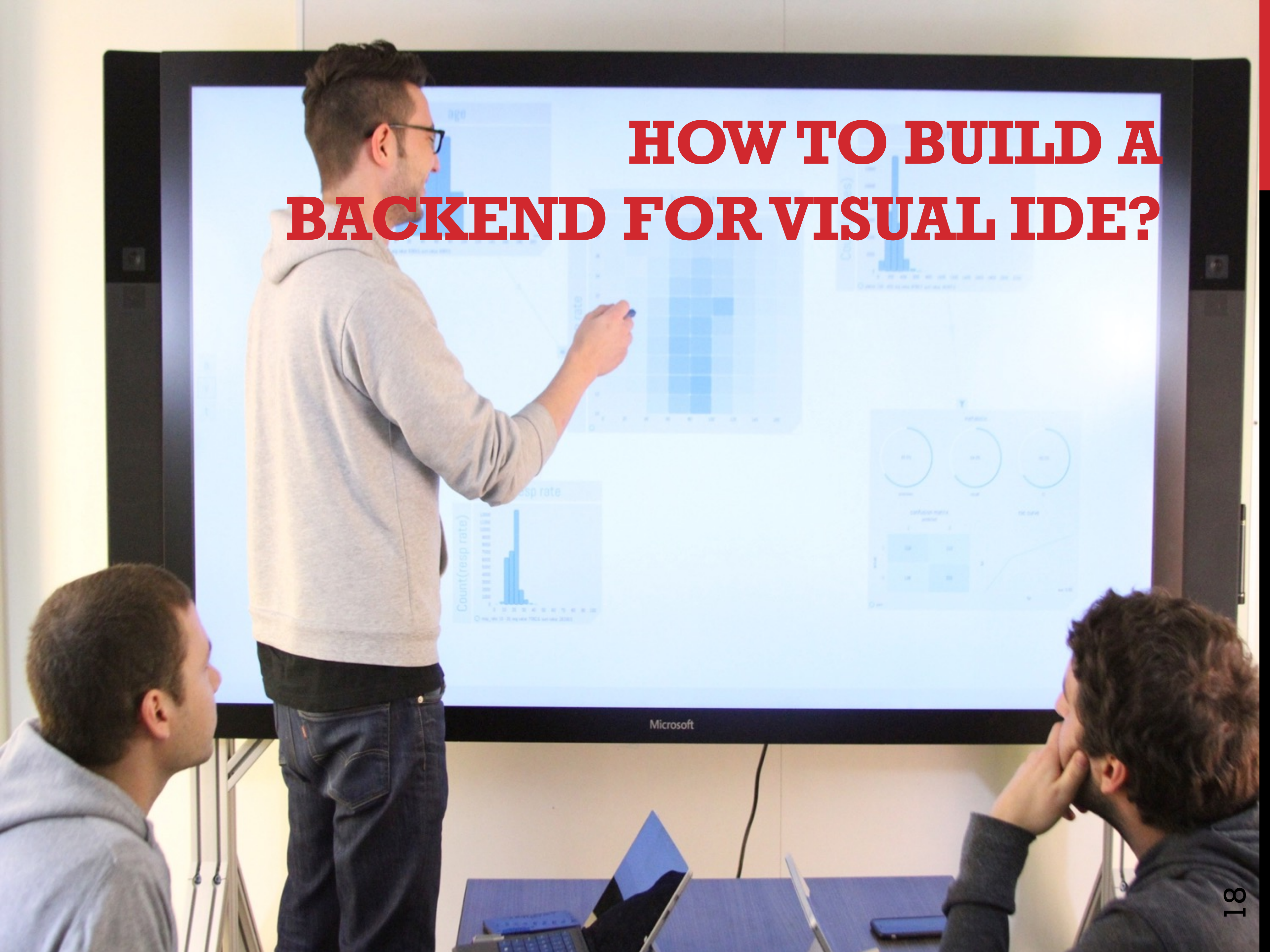
OPPORTUNITY: THINK TIME

User typically look at results for a significant amount of time ("think time") before executing next step

Step A >> 500 ms **Step B**

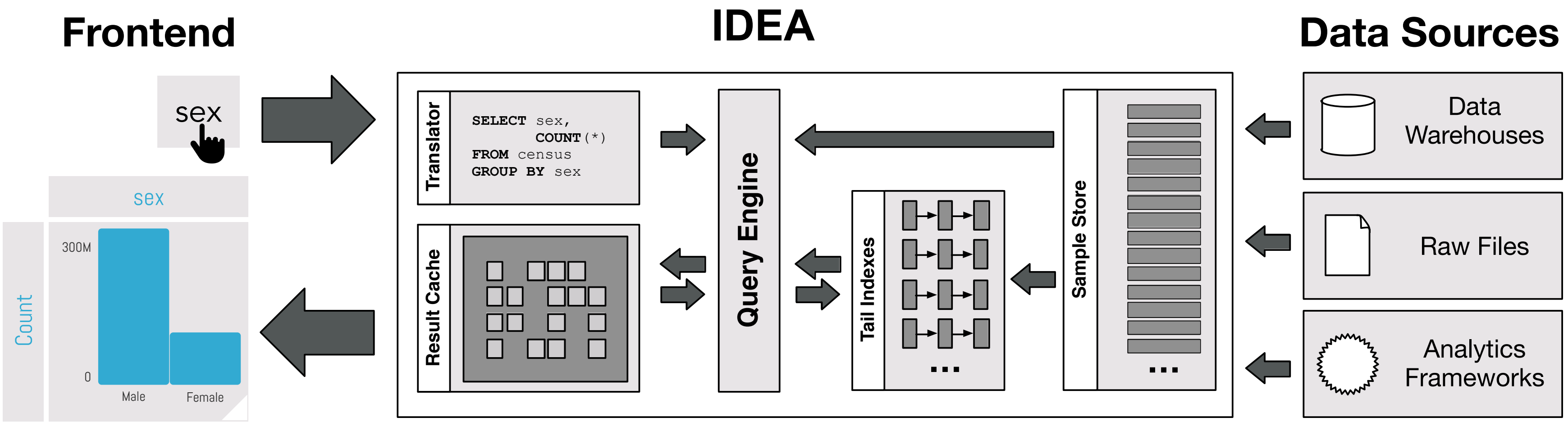
Optimization: Speculative execution while user "thinks"

HOW TO BUILD A BACKEND FOR VISUAL IDE?



OUR APPROACH: IDEA

IDEA = Interactive Data Exploration Accelerator



- Connect & Explore for new Data Sources
- **Progressive & Approximate Query Processing**
- **Incremental Query Building & Reuse**

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

BASIC IDEA OF AQP (FORM THE 90'S)

Sales

Product	Amount
CPU	1
CPU	1
CPU	2
CPU	3
CPU	4
Disk	1
Disk	2
Monitor	1

SalesSample

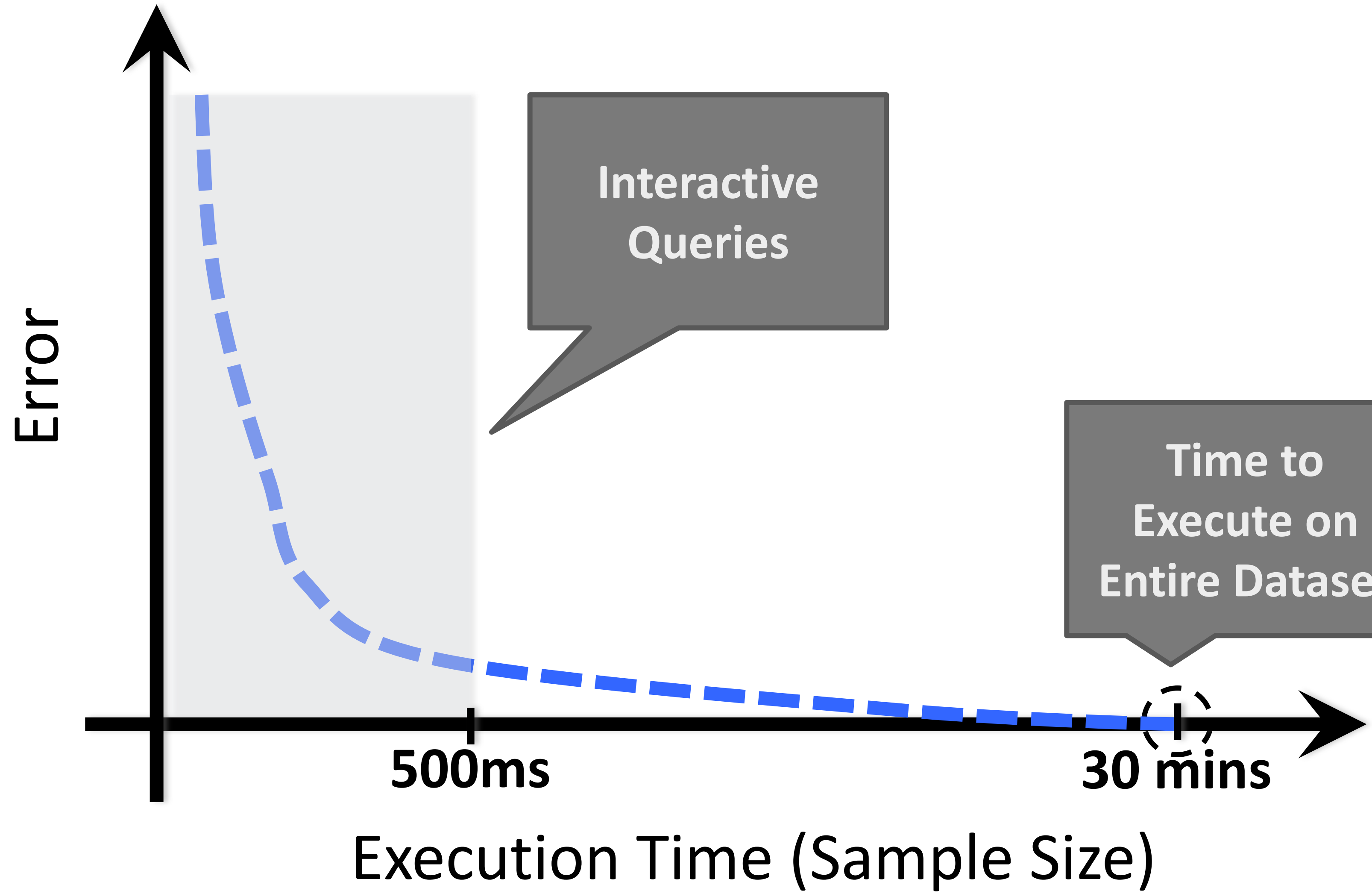
Product	Amount
CPU	1
CPU	2
CPU	3
Disk	2

SELECT SUM(Amount) FROM Sales WHERE Product = 'CPU'

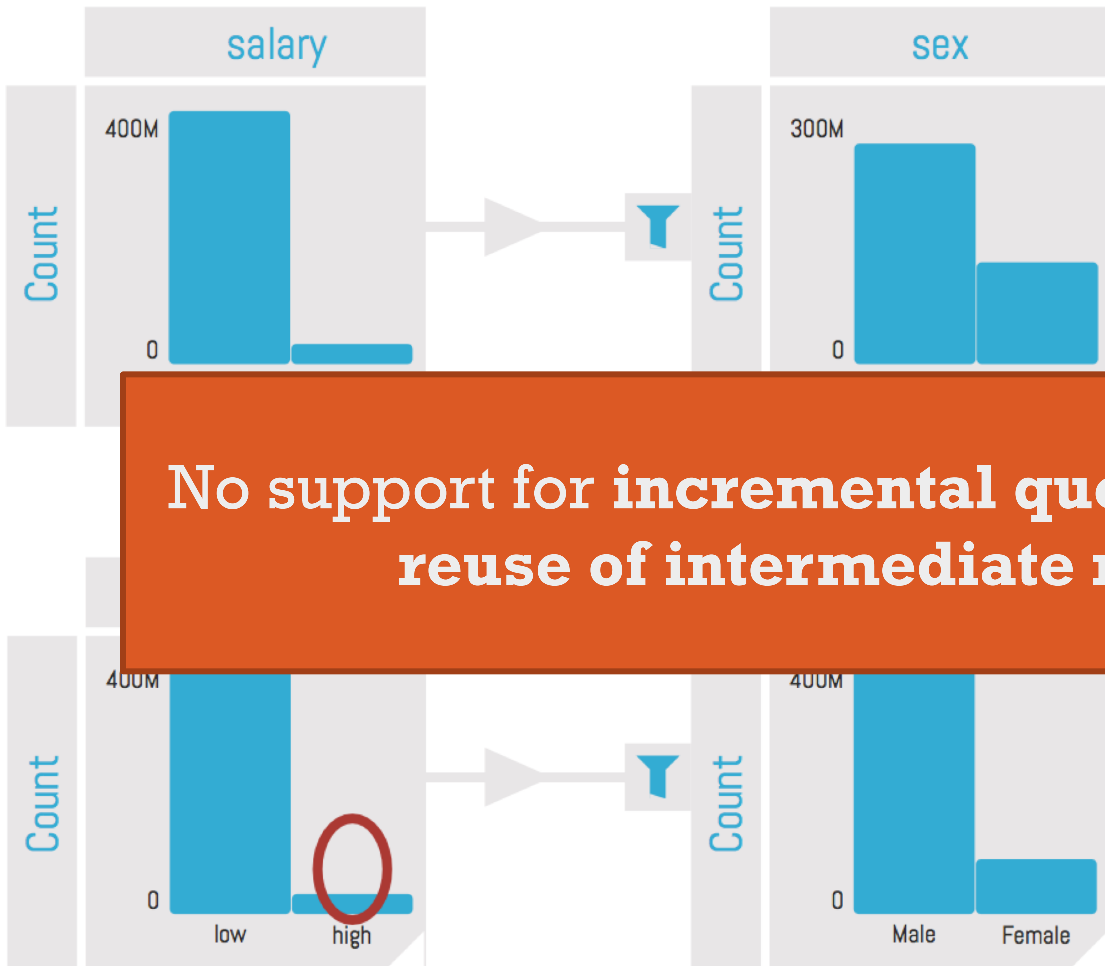
Exact Answer:
1+1+2+3+4 = 11

Approx. Answer:
(1+2+3)*2= 12

AQP: SPEED/ACCURACY TRADE-OFF



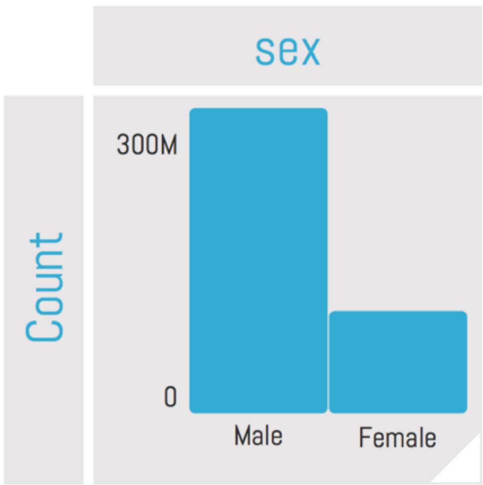
IS CLASSICAL AQP GOOD ENOUGH?



No support for **incremental query building & reuse of intermediate results**

OUR AQP FORMULATION

Main idea: results of prior approximate queries are represented as random variables X



Pr(X=Male)=0.75

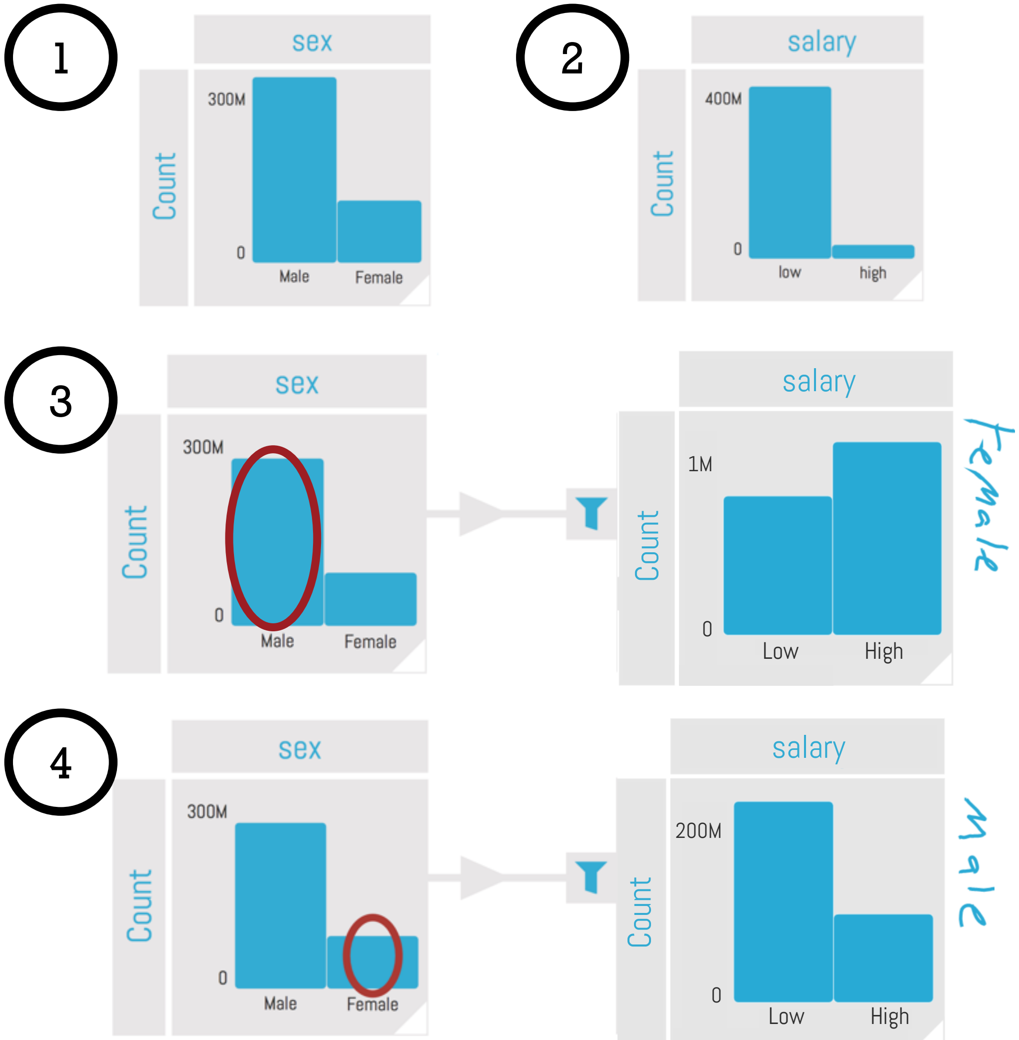
Pr(X=Female)=0.25

Enables reuse of approximate results with error bounds

Alex Galakatos, Andrew Crotty, Emanuel Zgraggen, Carsten Binnig, Tim Kraska: Revisiting Reuse for Approximate Query Processing. PVLDB 2017

AQP: RESULT REUSE

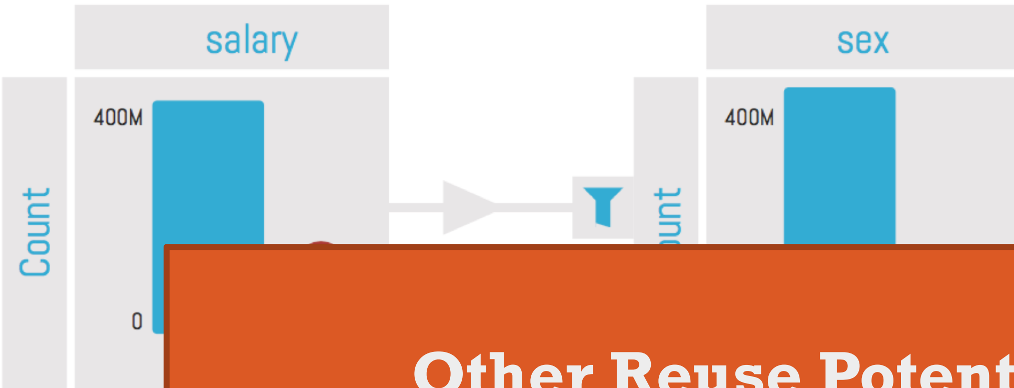
Executed Interactions



Result Cache

P_{male}	$\{0.70, \varepsilon_1\}$
P_{female}	$\{0.30, \varepsilon_2\}$
P_{high}	$\{0.20, \varepsilon_3\}$
P_{low}	$\{0.80, \varepsilon_4\}$
$P_{low\|male}$	$\{0.75, \varepsilon_5\}$
$P_{low\|female}$	$\{0.92, \varepsilon_6\}$
$P_{high\|male}$	$\{0.25, \varepsilon_7\}$
$P_{low\|female}$	$\{0.08, \varepsilon_8\}$

AQP: RESULT REUSE



Result Cache

[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
$P_{low\|female}$	[illegible]
$P_{high\|male}$	$\{0.25, \varepsilon_7\}$
$P_{low\|female}$	$\{0.08, \varepsilon_8\}$

Other Reuse Potentials

- Law of total probability
- Inclusion-exclusion principle

Bay

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$$P_{male|high} = \frac{P_{high|male} * P_{male}}{P_{high}} \approx 0.88$$

IDEA PERFORMANCE RESULTS

Exploration Session (User Study)

#1	sex
#2	education
#3	education **WHERE** sex='Female'
#4	education **WHERE** sex='Male'
#5	sex, education
#6	sex **WHERE** education='PhD'
#7	salary
#8	salary **WHERE** education='PhD'
#9	sex, salary
#10	salary **WHERE** sex='Female'
#11	salary
#12	salary **WHERE** sex='Female'
#13	salary **WHERE** sex<>'Female'
#14	salary **WHERE** sex='Female' **AND** education='PhD', salary **WHERE** sex<>'Female' **AND** education='PhD'
#15	age
#16	salary **WHERE** 20<=age<40 **AND** sex='Female' **AND** education='PhD', salary **WHERE** 20<=age<40 **AND** sex<>'Female' **AND** education='PhD'

Evaluated Systems:

- **MonetDB:** Analytical Column-Store
- **Online Aggregation** (From Hellerstein. 90's)
- **IDEA:** Our System

Data: 500M tuples

Census		#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16
	MonetDB	0.34	0.39	5.40	8.70	0.48	1.20	1.20	0.91	0.53	4.80	0.42	4.70	1.10	5.60	1.60	7.10
	Online Agg	0.05	0.24	0.78	0.59	0.24	0.46	0.04	0.48	0.07	0.11	0.04	0.11	0.08	7.53	0.29	24.3
	IDEA	0.09	0.29	0.42	0.00	0.00	0.00	0.09	0.12	0.00	0.17	0.00	0.00	0.00	0.48	0.37	2.87

MANY OTHER CONSIDERATIONS

Natural Language Interfaces

Benchmarking

Complex Workloads (ML, Text, …)

Hardware Acceleration

…

MANY OTHER CONSIDERATIONS

Natural Language Interfaces

Benchmarking

Complex Workloads (ML, Text, …)

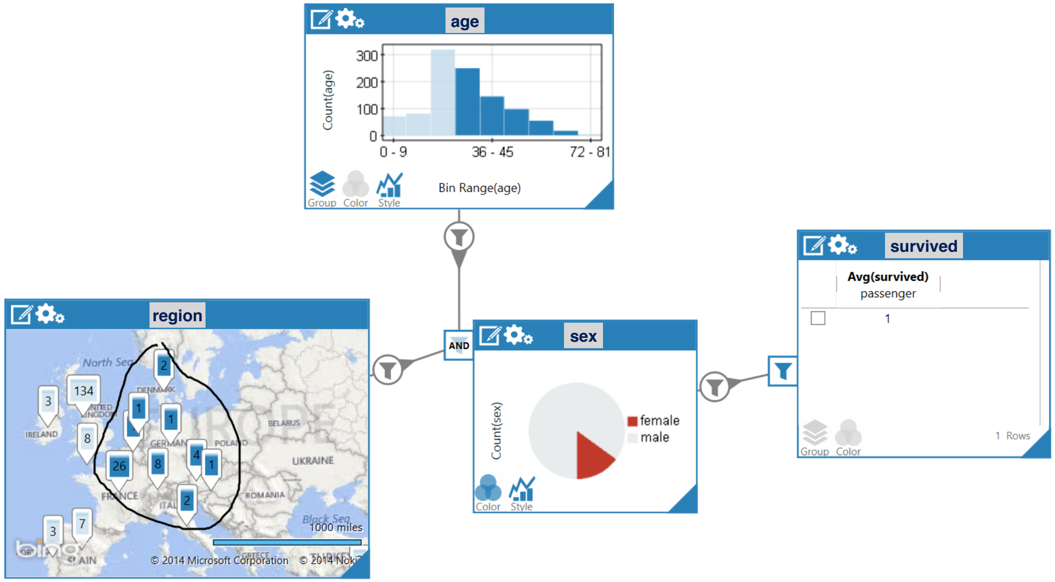
Hardware Acceleration

…

NL INTERFACE FOR DATABASES (NLIDB)

NL Query:

"How many older female people survived the sinking of the Titanic?"



ROBUST QUERY TRANSLATION?

Language Translation Problem

OUR APPROACH: GENERATE TRAINING DATA

Input

DB Schema → **Instantiation using Templates** (Cover variety of SQL) → Simple SQL / NL Pairs → **Paraphrasing, Noising, ….** (Cover variety of NL) → Augmented SQL / NL Pairs

Output

Distant Supervision: Generate large (potentially noisy) training data instead of manually handcrafting it

Fuat Basik, Benjamin Hättasch, Amir Ilkhechi, Arif Usta, Shekar Ramaswamy, Prasetya Utama, Nathaniel Weir, Carsten Binnig, Ugur Çetintemel: DBPal: A Learned NL-Interface for Databases. SIGMOD Conference 2018

OUR APPROACH: GENERATE TRAINING DATA

Input → **Output**

DB Schema → **Generate using Templates** → SQL / Naïve NL Pairs → **Paraphrasing, Noising, …** → SQL / Augmented NL Pairs

Cover variety of SQL

Cover variety of NL

Template(s)

SELECT <att>
FROM <table>
WHERE <filter>

Show me the <att>s of <table>s with <filter>?

name	**age**	**diagno ses**
Carsten	39	fever
Emilie	8	flu
Frederik	4	fever

Patient Table

Naïve Corpus

SELECT ***name***
FROM ***patient***
WHERE ***diagnoses = fever***

Show me the names of patients with diagnoses fever?

…

Paraphrasing

Show me the names of patients diagnosed fever?

Noising

Show the names of patients with ~~diagnosed~~ fever?

…

Millions of different NL/SQL pairs

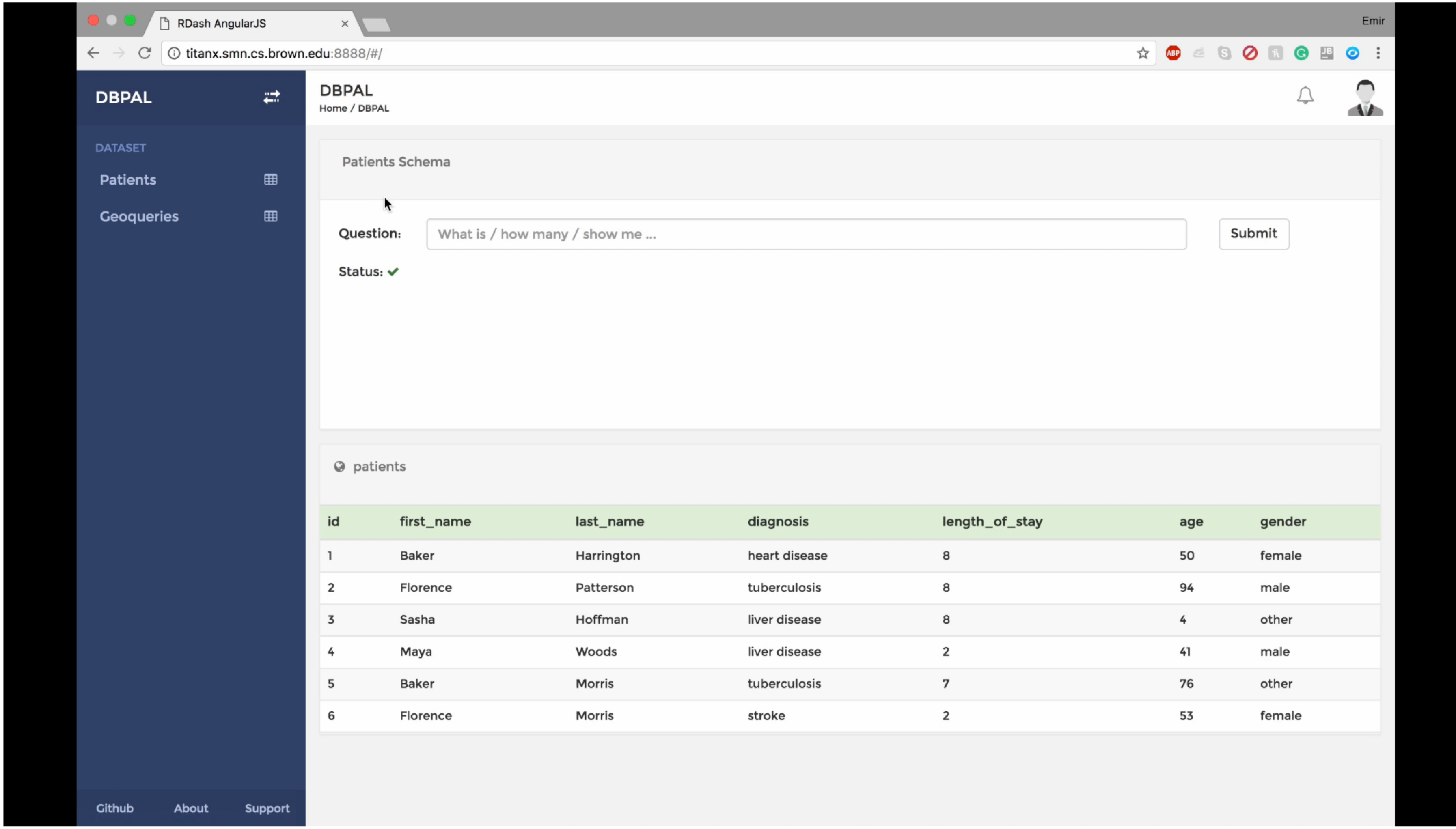
EXPERIMENTAL EVALUATION

Evaluated Systems

- **NaLIR:** Rule-based NLIDB (Best Paper VLDB 2015)
- **Neural Semantic Parser (NSP):** Neural Machine Translation (supervised learning -> **manual effort per database schema**)
- **DBPal:** Our Approach (distant supervision -> **NO manual effort** per database scheme)

	Patients	GeoQuery
NaLIR (w/o feedback)	15.60%	7.14%
NaLIR (w feedback)	21.42%	N/A
NSP++	N/A	**83.9%**
NSP (template only)	10.60%	5.0%
DBPal (w/o augmentation)	74.80%	38.60%
DBPal (full pipeline)	**75.93%**	55.40%

DBPAL IN ACTION



http://titanx.smn.cs.brown.edu:8888/#/patients

MANY OTHER CONSIDERATIONS

Natural Language Interfaces

Benchmarking

Complex Workloads (ML, Text, …)

Hardware Acceleration

…

BENCHMARKING IDE (BIDE)

DB Community

- existing benchmarks (e.g., TPC-H)
- static, non-incremental queries
- report the performance of single queries

Viz Community

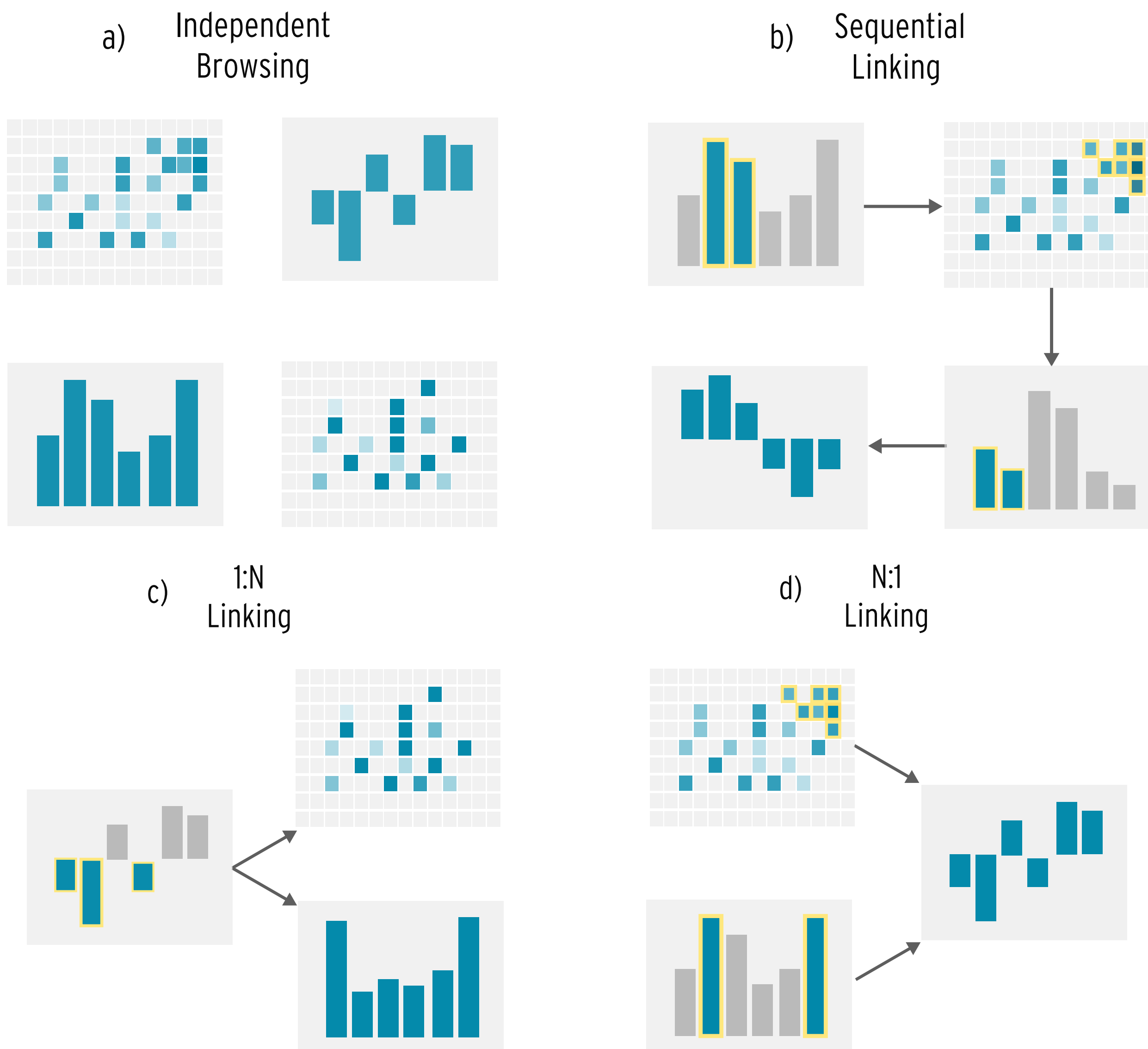
- user studies for evaluation
- use insight-based metrics
- costly, difficult to compare/reproduce

BIDE

- simulates common user behavior (think time, incremental queries, ...)
- introduces new metrics to better measure „interactivity“

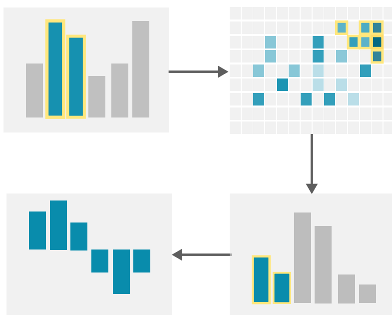
Website: http://idebench.github.io

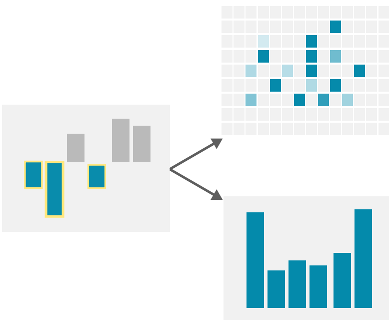
BENCHMARK WORKLOAD



WORKLOAD: OTHER DIFFERENCES

Multiple Concurrent Queries (triggered by one UI interaction)



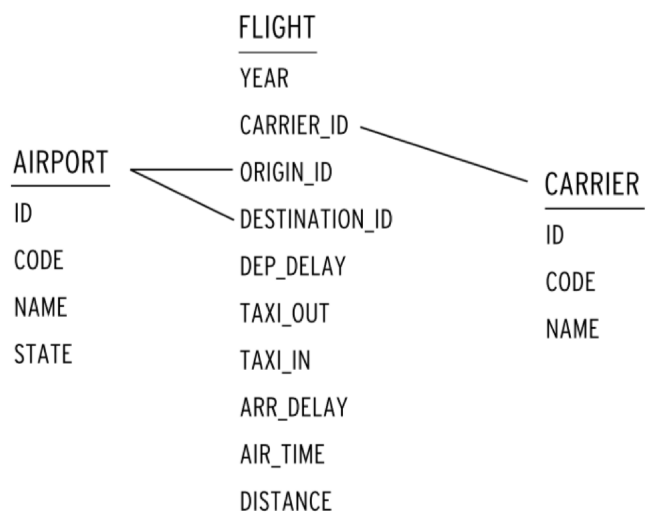


Visualization-Specific Functions (e.g., Binning, Cross-Filter)

Other Parameters (Think Time, …)

BENCHMARK DATA SETS

IDEBench comes with real-world data sets (e.g., Airline)

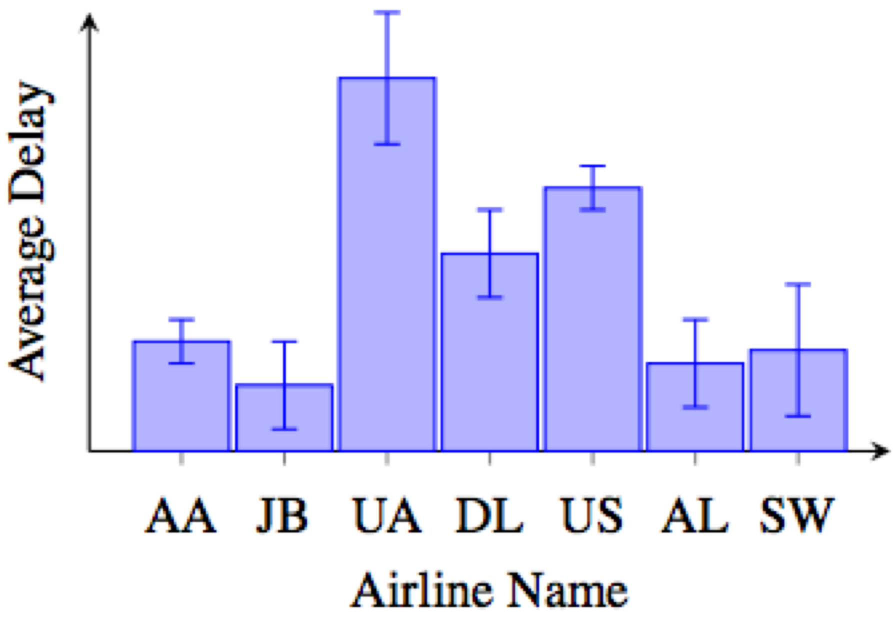


Data Generator:

- **Supports different schema variants (normalized vs. denormalized)**
- **Can be used to scale-up and down data sets**

BENCHMARK METRICS

Result quality (error, completeness) and time are both important metrics for such a benchmark



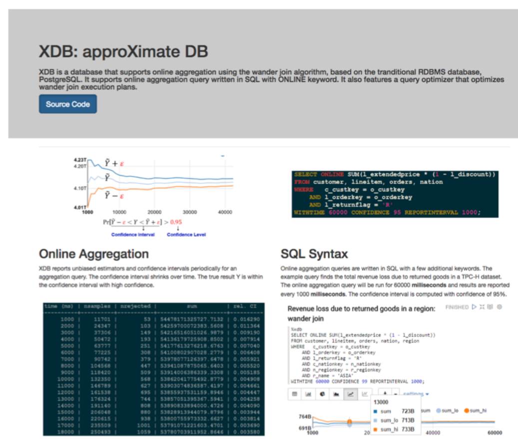
Main Idea: Capture quality of result after time t

BENCHMARKED SYSTEMS

Classical Analytical DBMSs	Approximate DBMSs	Specialized IDE Engines

monetdb

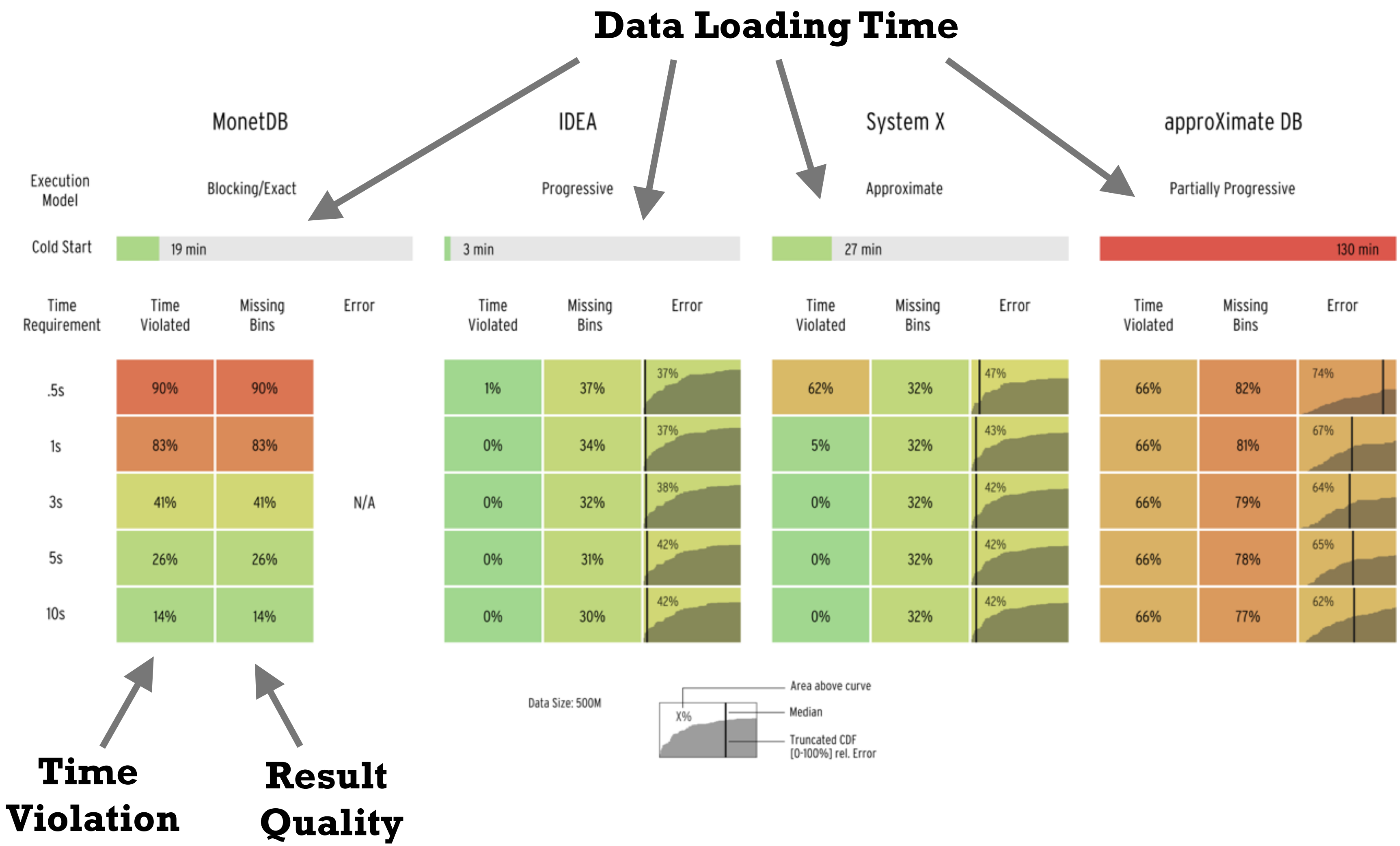
System X



vizdom IDEA

System Y

REPORTED RESULTS



Philipp Eichmann, Carsten Binnig, Tim Kraska, Emanuel Zgraggen: IDEBench: A Benchmark for Interactive Data Exploration. CoRR abs/1804.02593 (2018)

SUMMARY

Interactive Data Exploration is challenging

We need to rethink the full data exploration stack

- Query Interfaces
- Query Execution
- Cleaning / Loading

Other Considerations:

- Complex Workloads (ML, Text, …)
- Hardware Acceleration
- …

COLLABORATORS