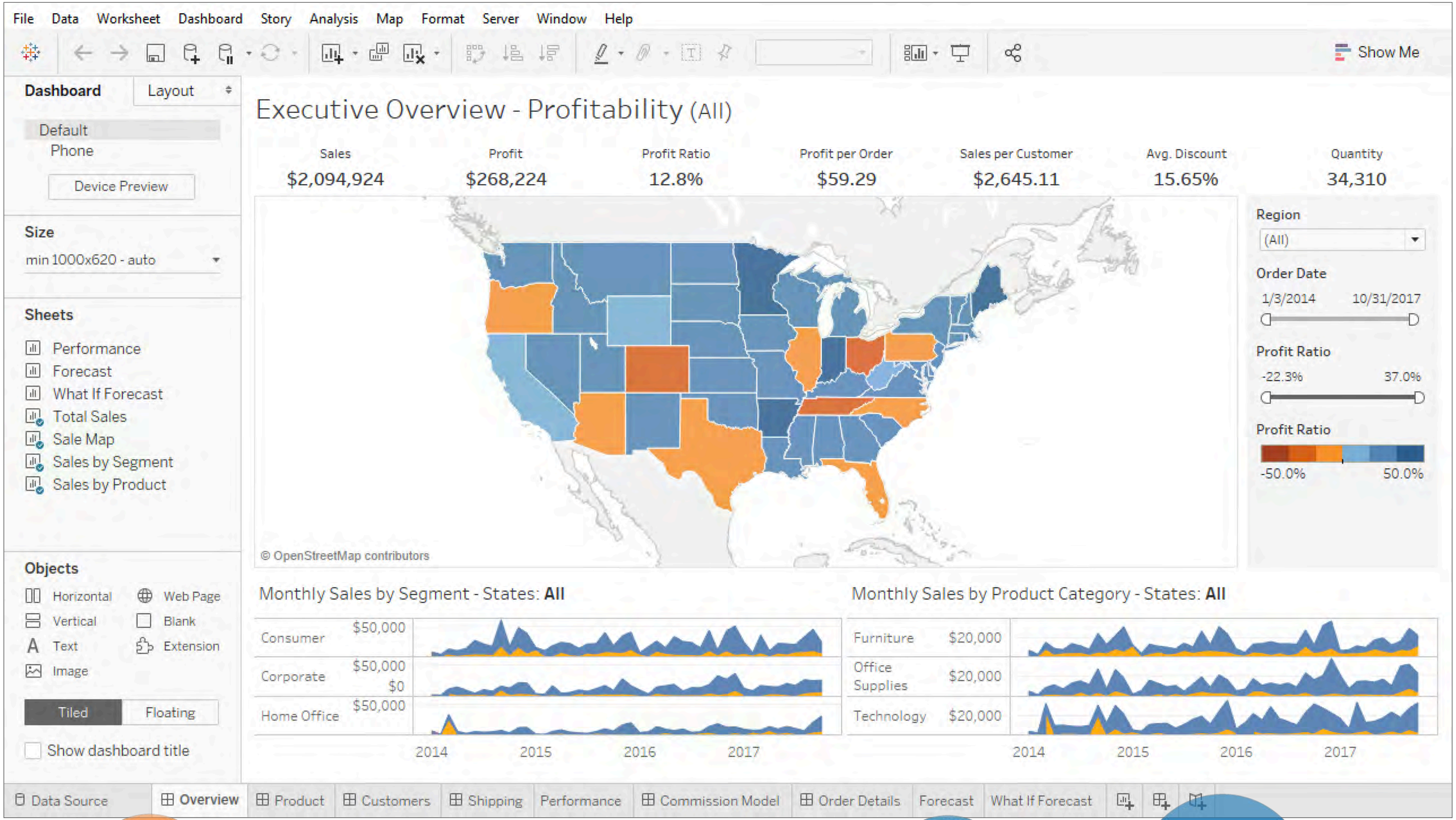
Get real:
How Benchmarks Fail to Represent the Real World

Adrian Vogelsgesang, Michael Haubenschild
Jan Finis, Alfons Kemper, Viktor Leis, Tobias Muehlbauer, Thomas Neumann, Manuel Then
{avogelsgesang, mhaubenschild, jfinis, …}@tableau.com

June 15th 2018

(and other BI tools)



Size: min 1000x620 - auto

□ Show dashboard title

Executive Overview - Profitability (All)

Sales	Profit	Profit Ratio	Profit per Order	Sales per Customer	Avg. Discount	Quantity
$2,094,924	$268,224	12.8%	$59.29	$2,645.11	15.65%	34,310

Region: (All)

Order Date: 1/3/2014 – 10/31/2017

Profit Ratio: -22.3% – 37.0%

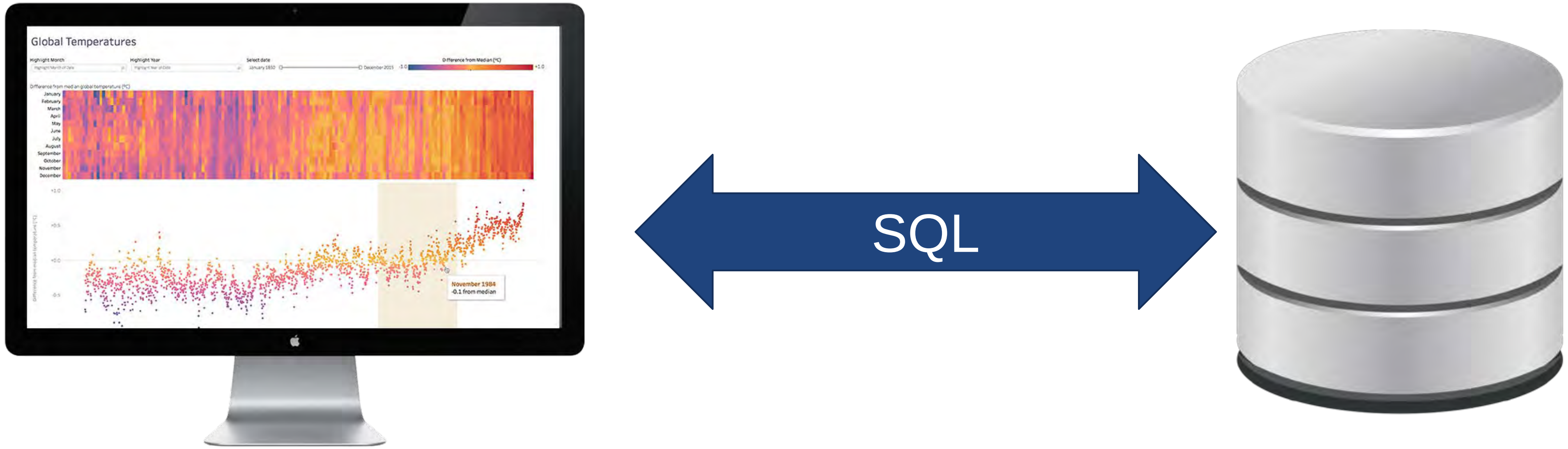
Profit Ratio: -50.0% – 50.0%

Monthly Sales by Segment - States: **All**

Monthly Sales by Product Category - States: **All**

The actual data crunching…

… is delegated to an actual database



The actual data crunching…

… is delegated to an actual database

This could be your system

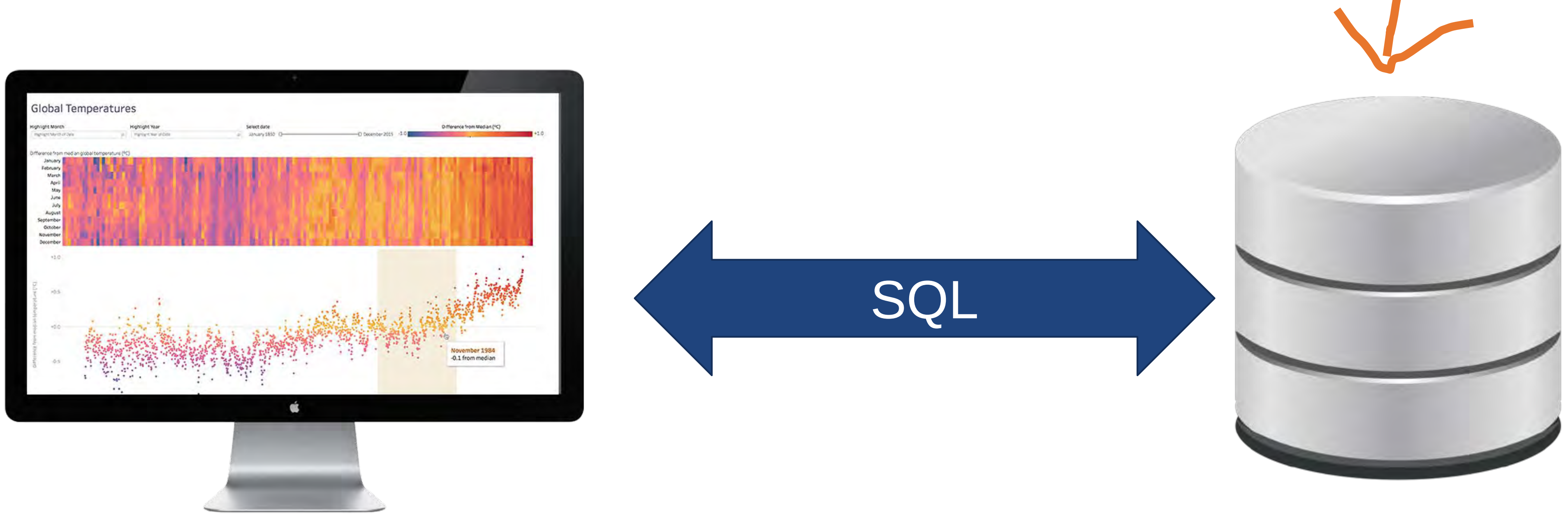


Tableau Public

- Free cloud hosting for visualizations
- Including both visual specification and raw data

For us: a huge repository of test data

This talk: statistics about 60k visualizations **only** from Public

Biased towards small datasets, but we can share our findings with you ☺

Over 1 million queries

Our Insights

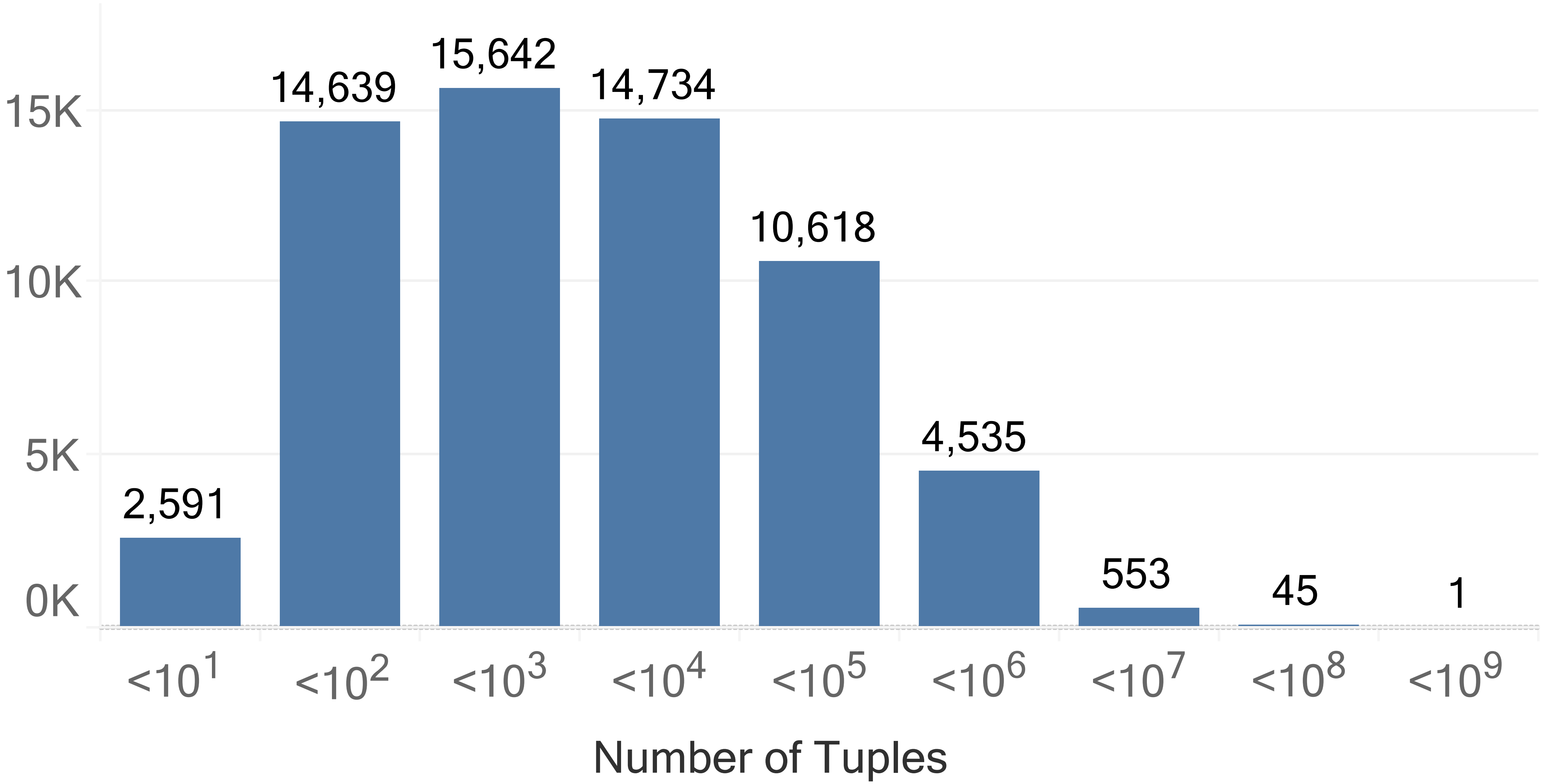
Many meta data queries

- Column names, data types, … (“SELECT tablename FROM pg_tables;”)
- Current server time (“SELECT NOW;”)
- Feature testing (“Let’s see, can I create a temporary table?”)

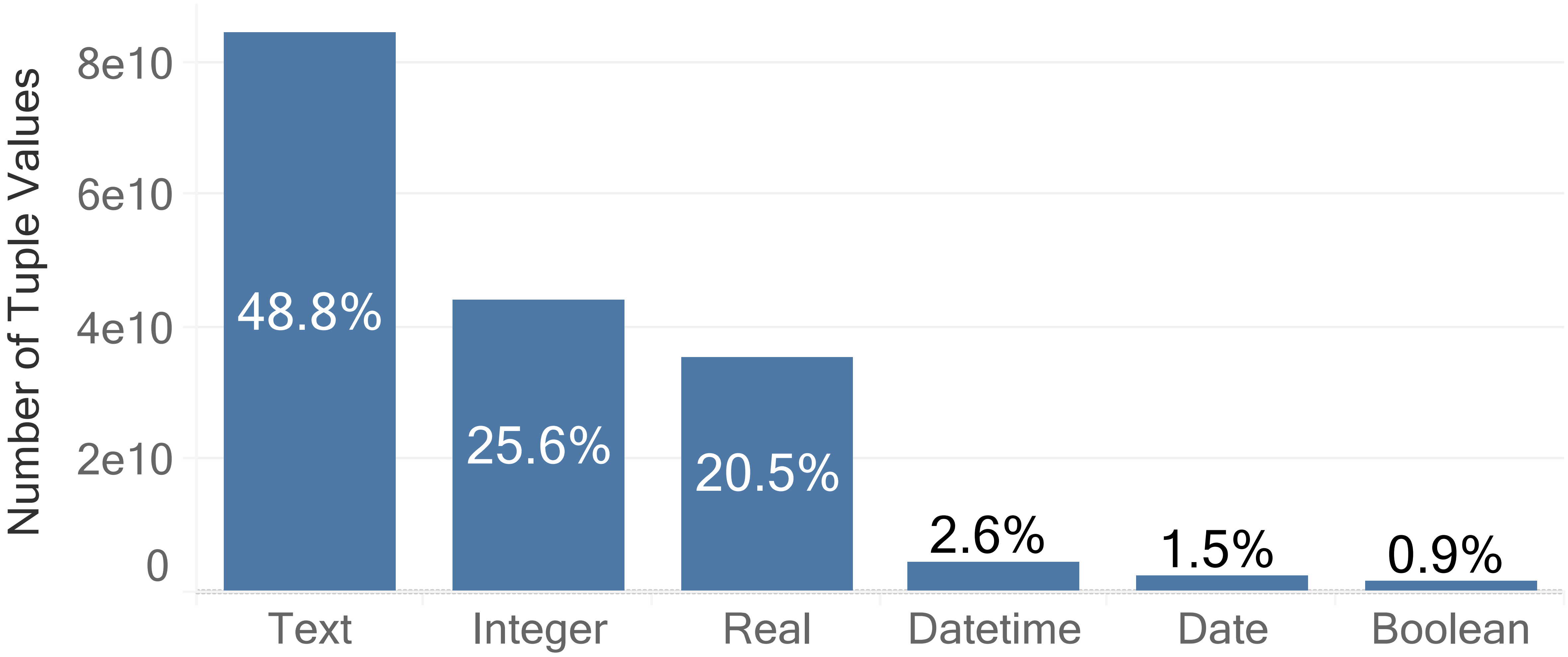
All in all: 75% of the queries

Make metadata queries efficient!

Data set sizes



Strings are everywhere



Strings are everywhere

- ISO country codes, IANA airport codes, ISBNs for books, UUIDs
- Boolean encoded as ”0”/”1” (60% of single-character-strings!)
- “male”/”female”

People don’t care about a clean schema – but:
they do care about performance

International strings & collation support

- 0.64% of the strings contain non-ASCII characters
- Small fraction, but nevertheless must be supported
- Not covered by benchmarks

Even worse: collations   (“A” = “â”)

- 85% of the string columns have a collation
- 70% case- or accent-insensitive
- Makes query optimization harder
- Collations are expensive to evaluate

HashJoin? Anyone?

The queries

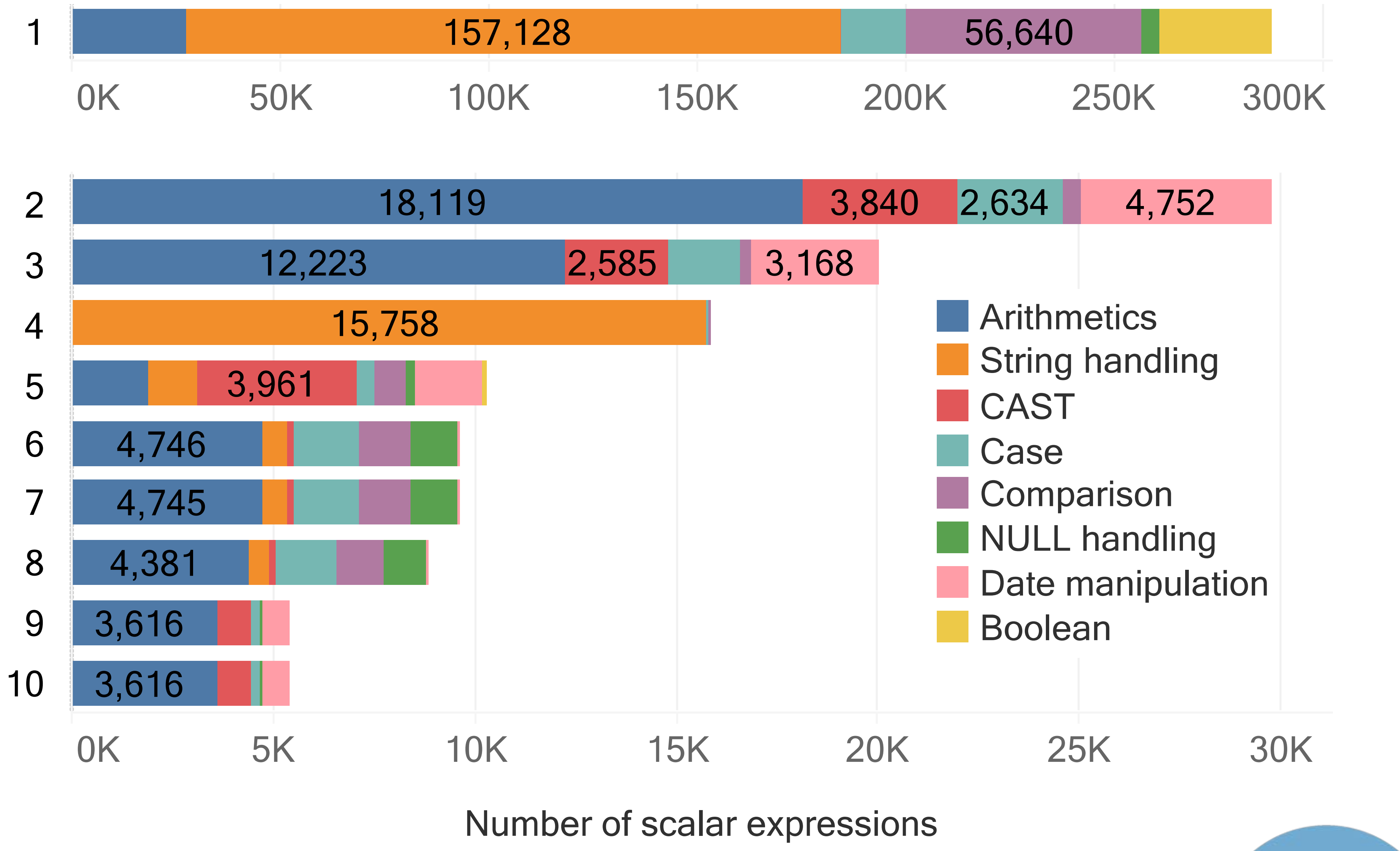
- Most are small: Only 0.5% larger than 5KB
- But: Huge outliers

The queries

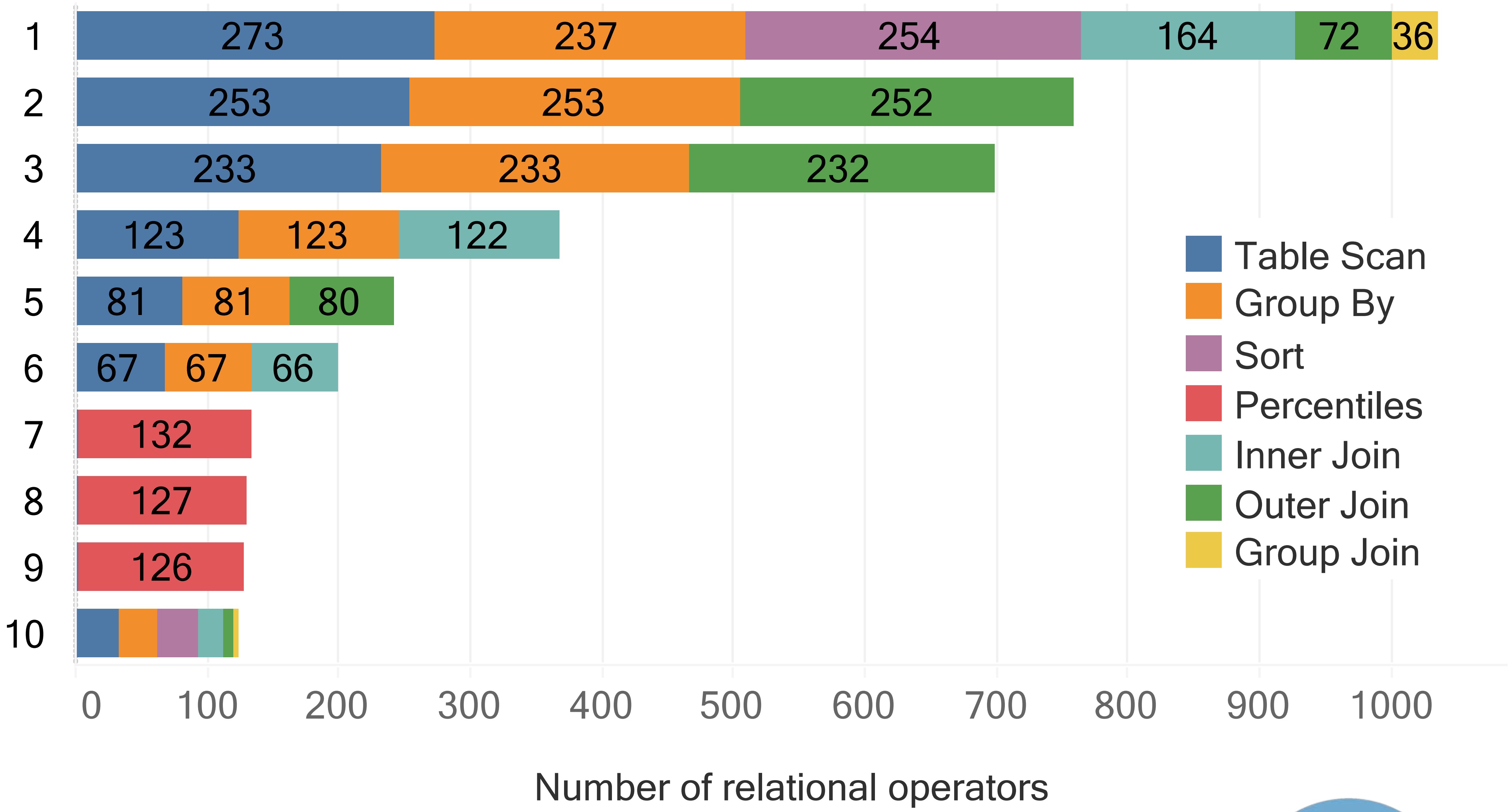
- Most are small: Only 0.5% larger than 5KB
- But: Huge outliers
- Largest query in our data set: 6.7MB
- Largest query I saw so far: 27MB

And that’s not all due to constant strings…

Expression-heavy queries



Operator-heavy queries



Incomplete queries

- Interactive exploration
- Not all queries make sense
- Missing filters, missing join conditions, …

And Benchmarks?

What do benchmarks do? (TPC-H/DS)

- Meaningful queries returning useful results
- Handwritten queries
- Well-designed schema
- Scale the data set size

What does that mean for benchmarks?

1. Include meta-data queries
2. Do bad schema design, use strings more often
3. Include Unicode & collations
4. Benchmark on tiny data sets, too
5. Scale query complexity, not only data size
6. Take into account incomplete/incorrect queries

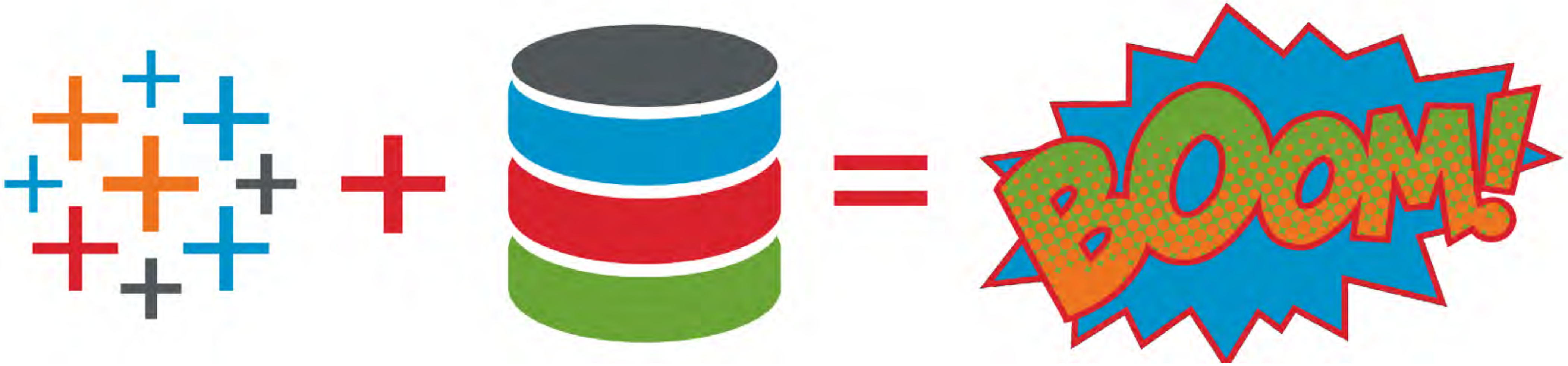
Questions?

Get real:
How Benchmarks Fail to Represent the Real World

Adrian Vogelsgesang, Michael Haubenschild
Jan Finis, Alfons Kemper, Viktor Leis, Tobias Muehlbauer, Thomas Neumann, Manuel Then
{avogelsgesang, mhaubenschild, jfinis, …}@tableau.com

June 15th 2018

Hyper



(Tableau)

(Hyper)