Data Science in Echtzeit

für mehr Freude im Job

Universität Oldenburg, Oldenburg, Germany
Wolfram „Wolle“ Wingerath
13. November 2024

Die Folien gibt es auf https://wolle.science

Ich bin Wolle

Research:
- NoSQL Systems
- Stream Processing
- Real-Time Databases
- …

Practice:
- Web Caching
- Big Data Analytics
- Anger Management
- …

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Carl von Ossietzky Universität Oldenburg

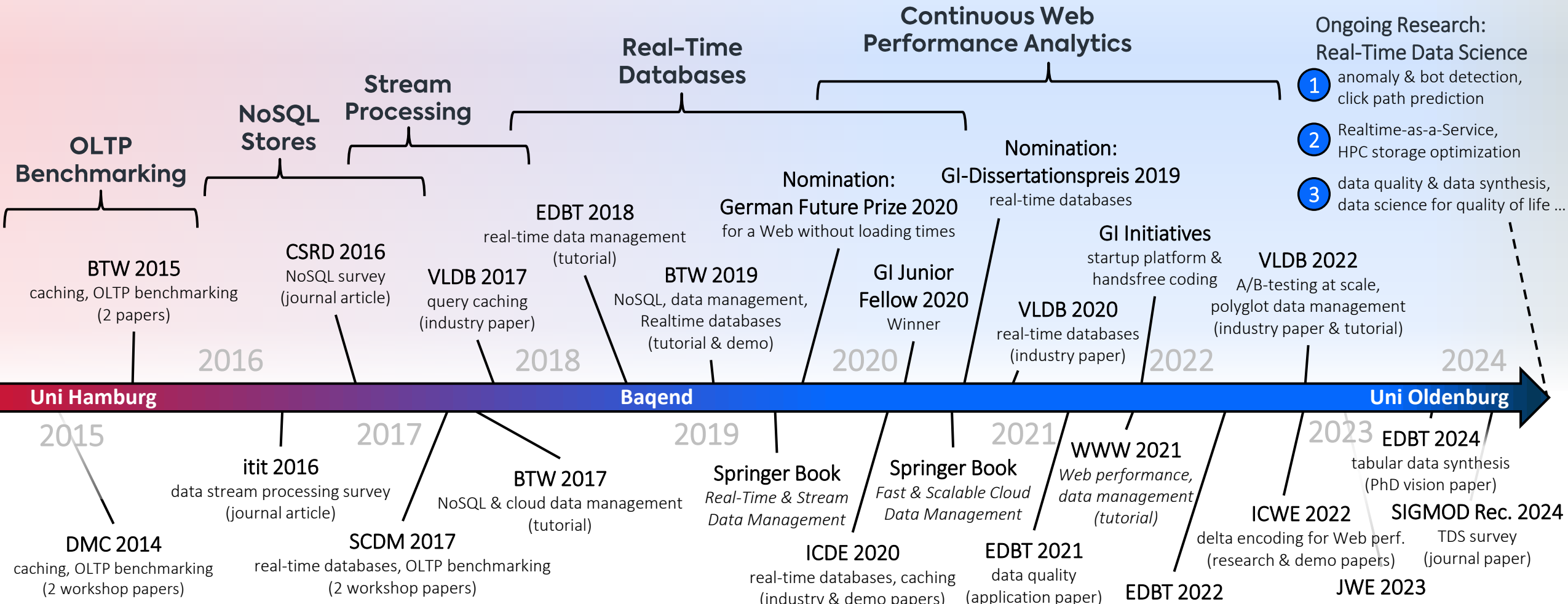
Baqend

Überblick: High-Performance Web Engineering

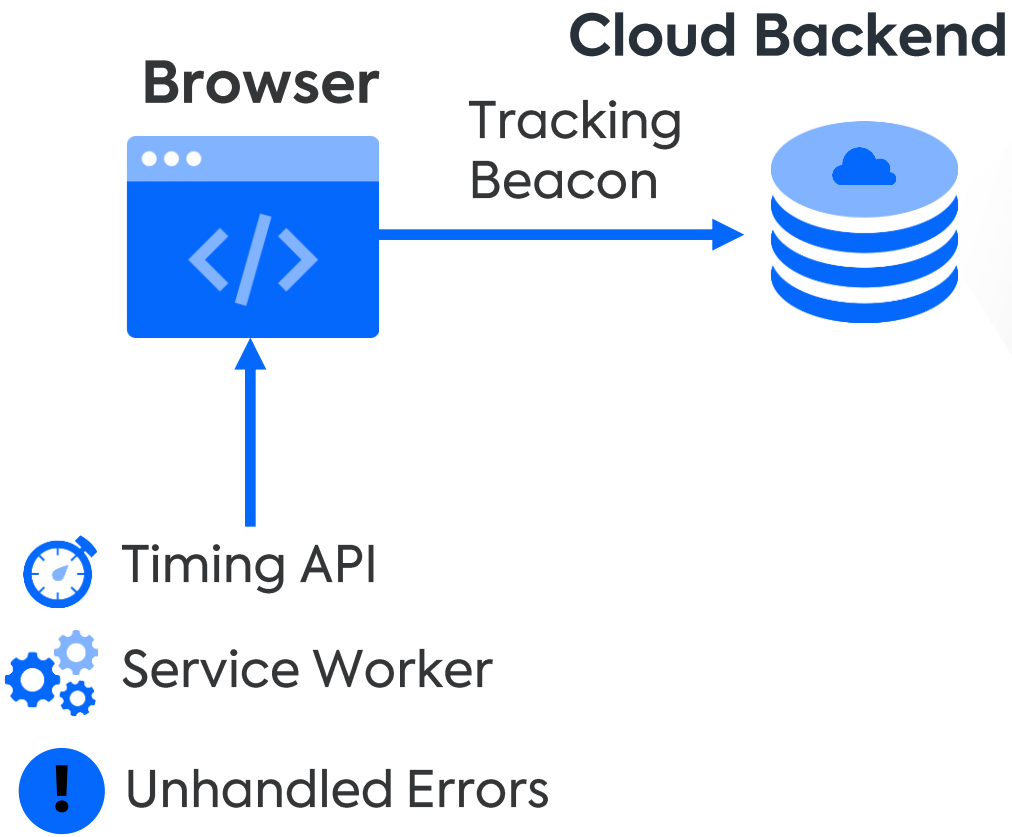
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Baqend

Carl von Ossietzky Universität Oldenburg



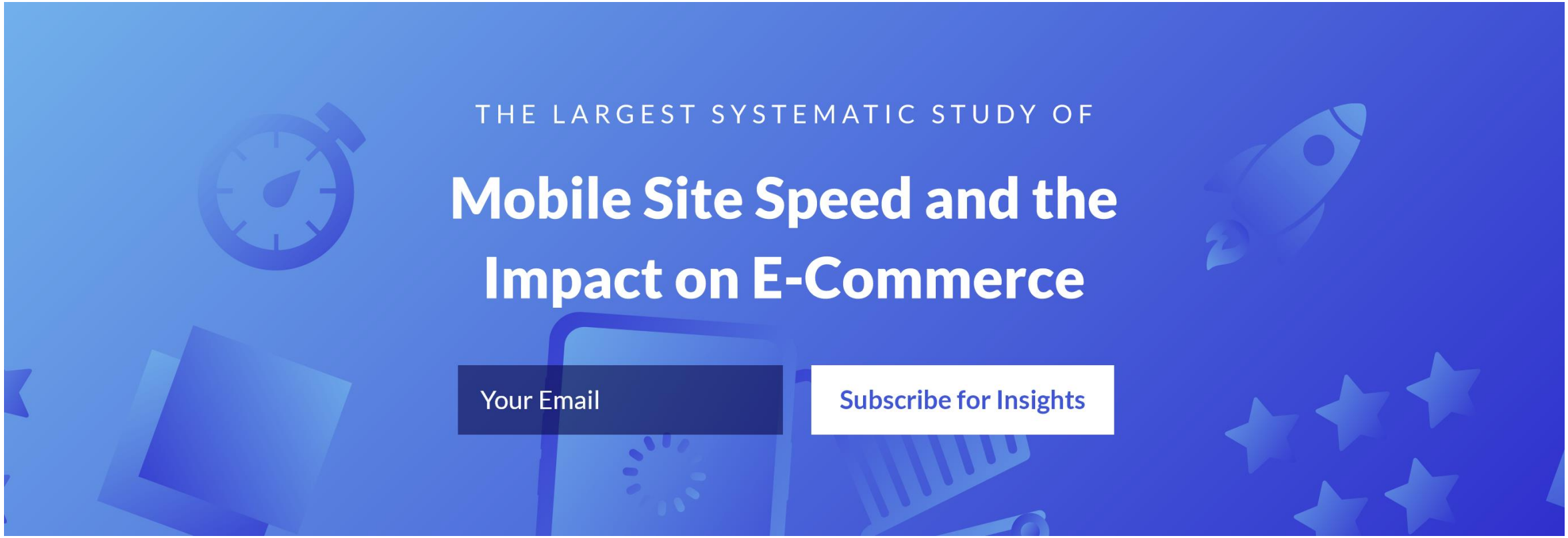
Technische & Business-Performance Messen



▪ Time-to-First-Byte ▪ First (Contentful) Paint ▪ DOM Timer ▪ First Input Delay	Performance
▪ Session Length ▪ Time on Site ▪ First User Interaction ▪ Bounce Rate	User Engagement
▪ Cart Size ▪ Transactions ▪ Conversion Rate ▪ Revenue	Business KPIs
▪ Page Views & Sessions ▪ Browser Distribution ▪ JavaScript Errors ▪ Caching Insights	QA Metadata

Wie & Warum Beschleunigen: speedhub.org

1



Google

Baqend

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

W. Wingerath, B. Wollmer, F. Gessert, S. Succo, N. Ritter. Going for Speed: Full-Stack Performance Engineering in Modern Web-Based Applications, WWW 2021 (Tutorial)

Split Testing für Web Performance

Speed Kit Users vs. Normal Users

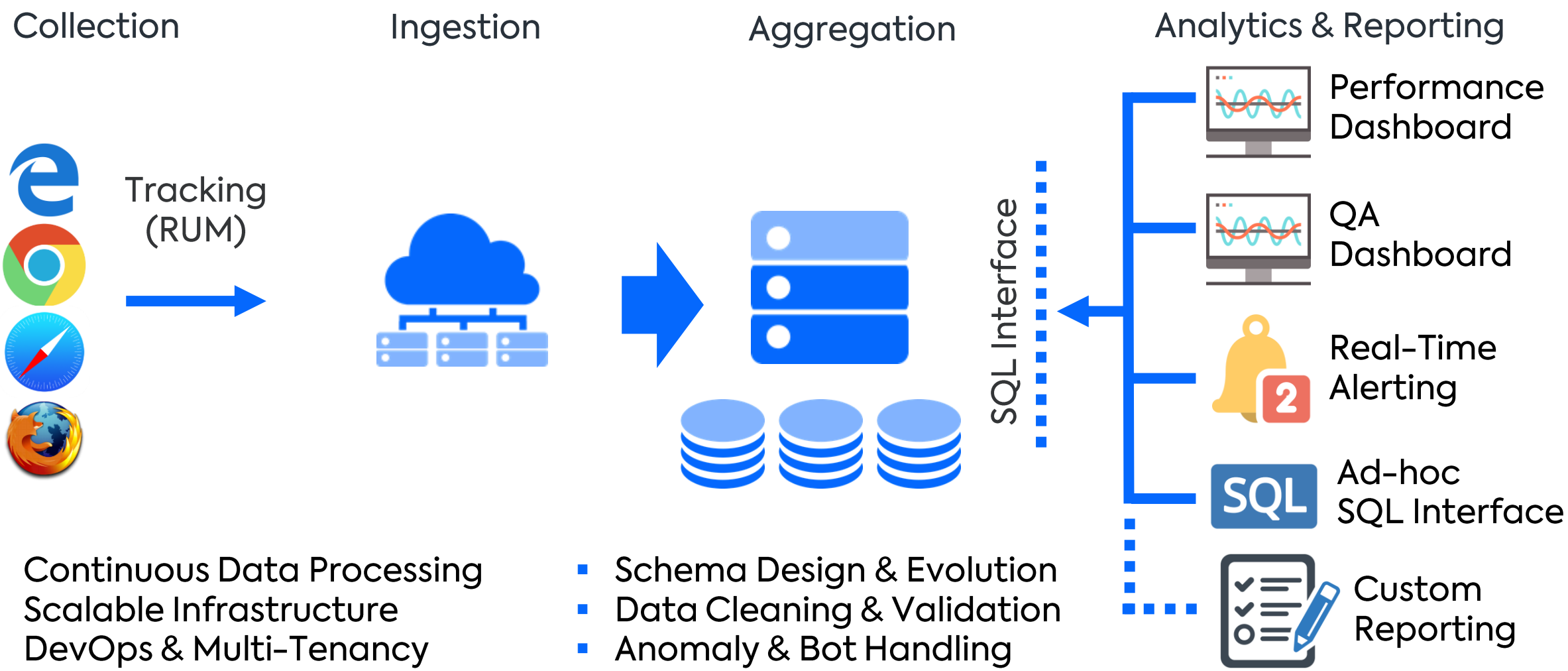
Tracking →

← Tracking

- Speed Kit enabled
- **Measurable uplift**:
  - \+ Performance
  - \+ User engagement
  - \+ Revenue
  - ...
- Speed Kit disabled (no acceleration)

W. Wingerath, F. Gessert, E. Witt, H. Kuhlmann, F. Bücklers, B. Wollmer, N. Ritter. Speed Kit: A Polyglot & GDPR-Compliant Approach For Caching Personalized Content, ICDE 2020

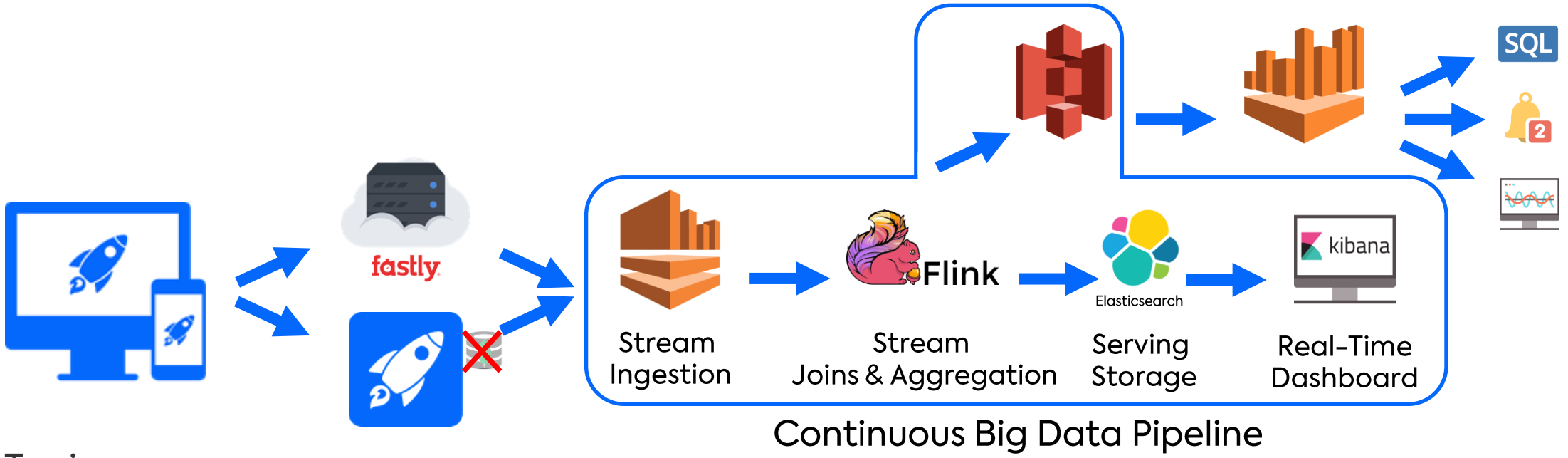
RUM: Real-User Monitoring



- Continuous Data Processing
- Scalable Infrastructure
- DevOps & Multi-Tenancy

- Schema Design & Evolution
- Data Cleaning & Validation
- Anomaly & Bot Handling

Real-Time Data Processing im Großen Stil



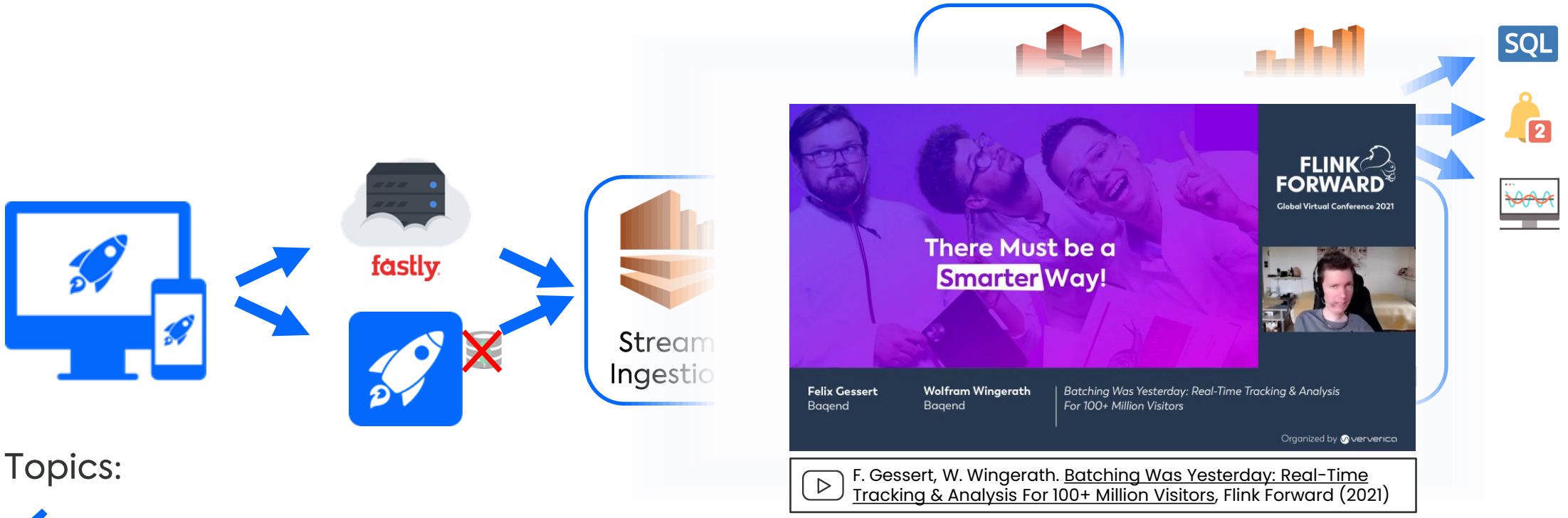
Key Topics:

- ✓ No legacy tech => stability & efficiency
- ✓ Faster ingestion => real-time reporting & analytics
- ✓ Fewer joins => faster analytics

In Collaboration With amazon | science

W. Wingerath, B. Wollmer, M. Bestehorn, S. Succo, F. Bücklers, J. Domnik, F. Panse, E. Witt, A. Sener, F. Gessert, N. Ritter. Beaconnect: Continuous Web Performance A/B-Testing at Scale, VLDB (2022)

Real-Time Data Processing im Großen Stil



F. Gessert, W. Wingerath. Batching Was Yesterday: Real-Time Tracking & Analysis For 100+ Million Visitors, Flink Forward (2021)

Key Topics:

- ✓ No legacy tech => stability & efficiency
- ✓ Faster ingestion => real-time reporting & analytics
- ✓ Fewer joins => faster analytics

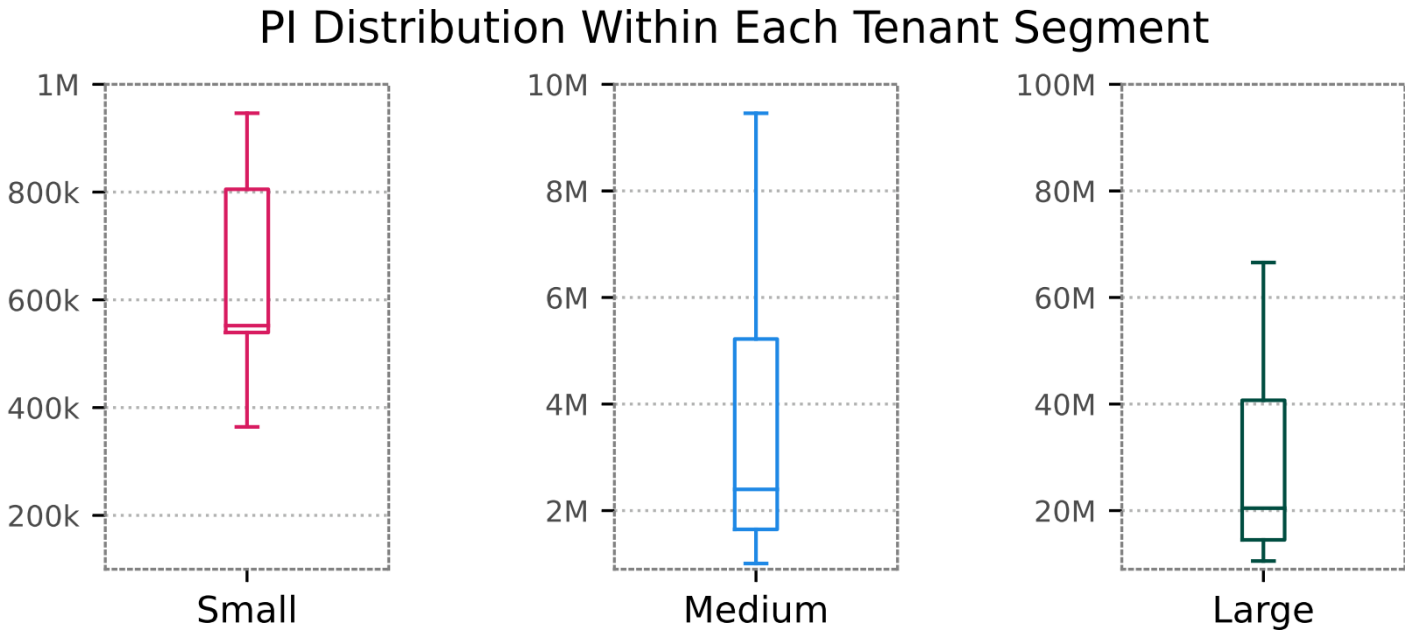
In Collaboration With amazon | science

W. Wingerath, B. Wollmer, M. Bestehorn, S. Succo, F. Bücklers, J. Domnik, F. Panse, E. Witt, A. Sener, F. Gessert, N. Ritter. Beaconnect: Continuous Web Performance A/B-Testing at Scale, VLDB (2022)

Traffic* Über 3 Größenordnungen

Overall Monthly Traffic

- >100M users
- >200M sessions
- >650M PIs
- >3B data beacons



*Monthly Page Impressions (PIs)

In Collaboration With amazon | science

W. Wingerath, B. Wollmer, M. Bestehorn, S. Succo, F. Bücklers, J. Domnik, F. Panse, E. Witt, A. Sener, F. Gessert, N. Ritter. Beaconnect: Continuous Web Performance A/B-Testing at Scale, VLDB (2022)

Incremental Processing: Effizientes Monitoring

Partial Page Impressions (PPIs)
Enhanced Data Beacons

Time	Browser	Device	Test Group	First Contentful Paint (FCP)
11:05:04.578	Firefox	Mobile	Speed Kit	127ms
11:06:48.139	Chrome	Mobile	Original	958ms

1-Min. Time Windows
Immediate Aggregates (Storage)

	Browser	Device	Test Group	First Contentful Paint (FCP)
11:05	Firefox	Mobile	Speed Kit	{200ms: 1, 500ms: 2}
	Firefox	Mobile	Original	{600ms: 2, 800ms: 5}
	Safari	Desktop	Original	{1100ms: 1}
11:06	Firefox	Mobile	Speed Kit	{200ms: 3}
	Chrome	Mobile	Speed Kit	{400ms: 2}
	Opera	Tablet	Original	{700ms: 1, 1300ms: 2}
	Safari	Desktop	Original	{600ms: 4, 900ms}

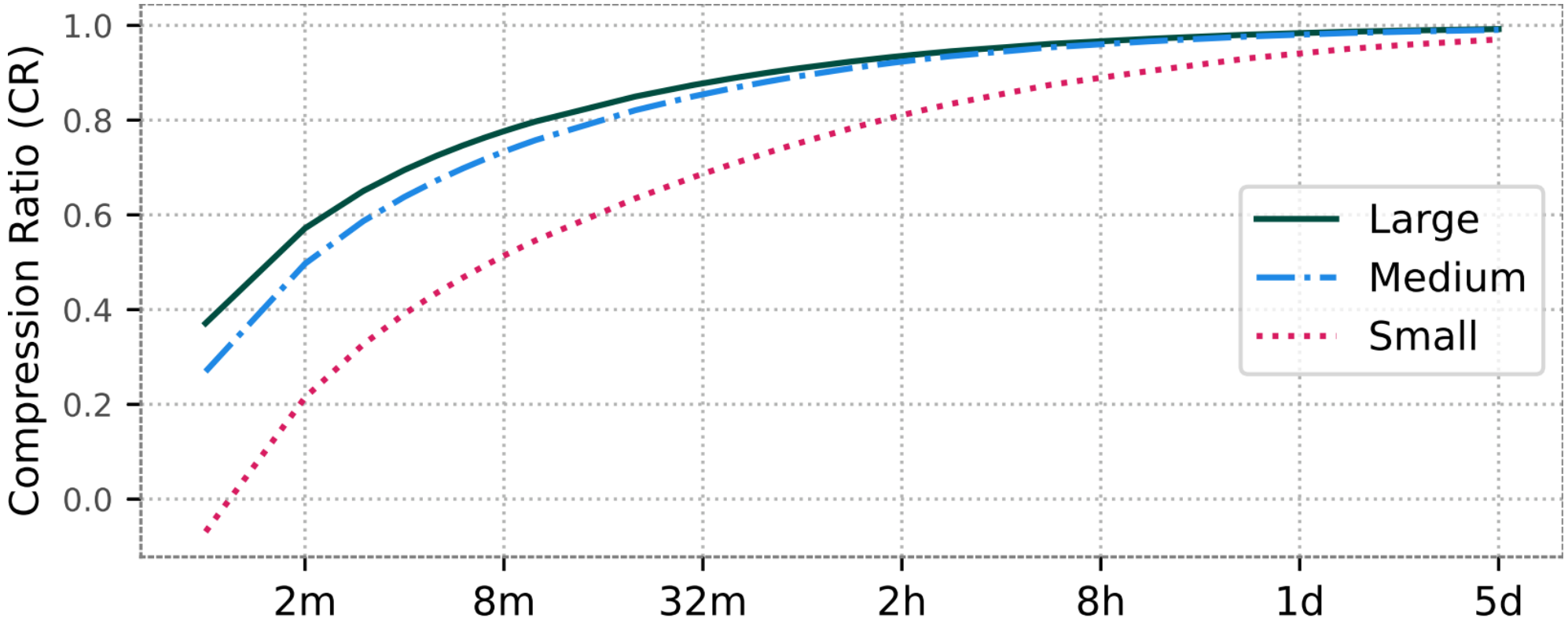
Arbitrary Time Windows
Real-Time Reporting (Dashboard Queries)

11:05 – 11:06	Browser	Device	Test Group	First Contentful Paint (FCP)
	Firefox	Mobile	Speed Kit	{200ms: 4, 500ms: 2}

In Collaboration With amazon | science

W. Wingerath, B. Wollmer, M. Bestehorn, S. Succo, F. Bücklers, J. Domnik, F. Panse, E. Witt, A. Sener, F. Gessert, N. Ritter. Beaconnect: Continuous Web Performance A/B-Testing at Scale, VLDB (2022)

Dashboarding Query Efficiency

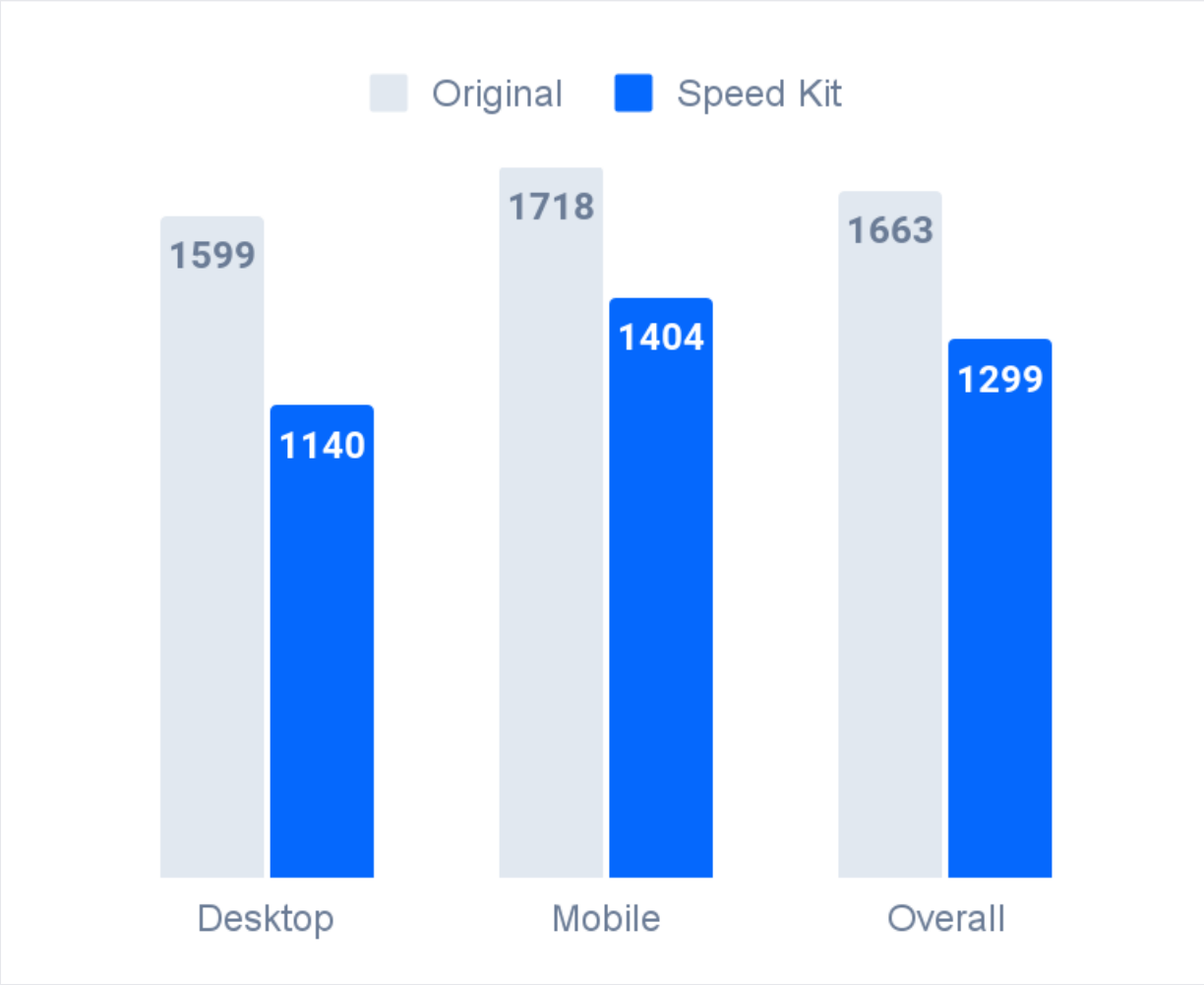


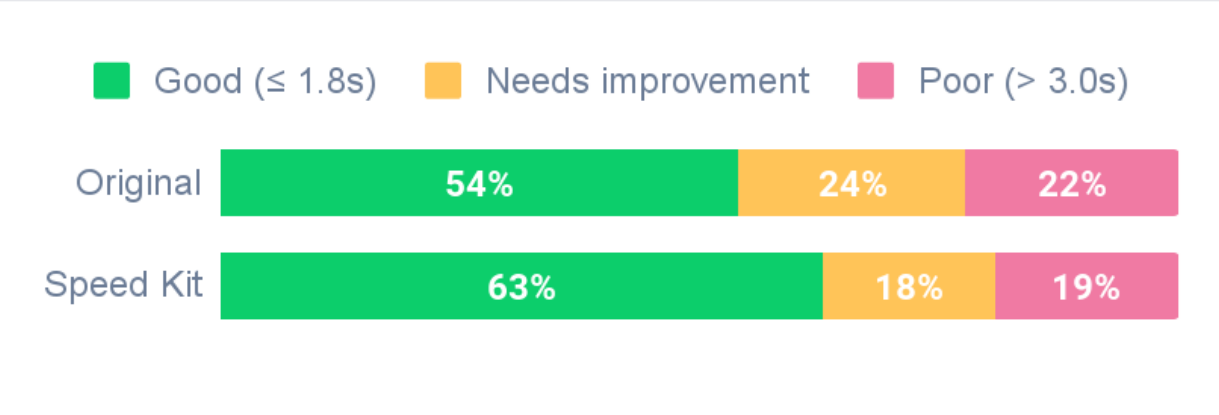
$$\text{Compaction Ratio: } CR = 1 - \frac{|\text{intermediate aggregates}|}{|\text{PIs}|}$$

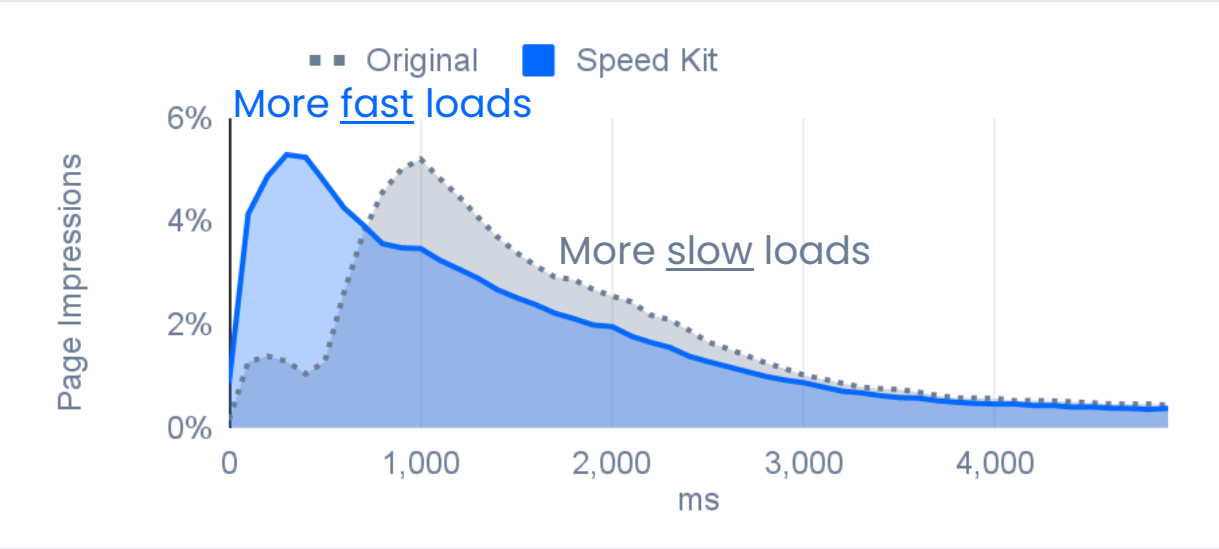
In Collaboration With amazon | science

W. Wingerath, B. Wollmer, M. Bestehorn, S. Succo, F. Bücklers, J. Domnik, F. Panse, E. Witt, A. Sener, F. Gessert, N. Ritter. Beaconnect: Continuous Web Performance A/B-Testing at Scale, VLDB (2022)

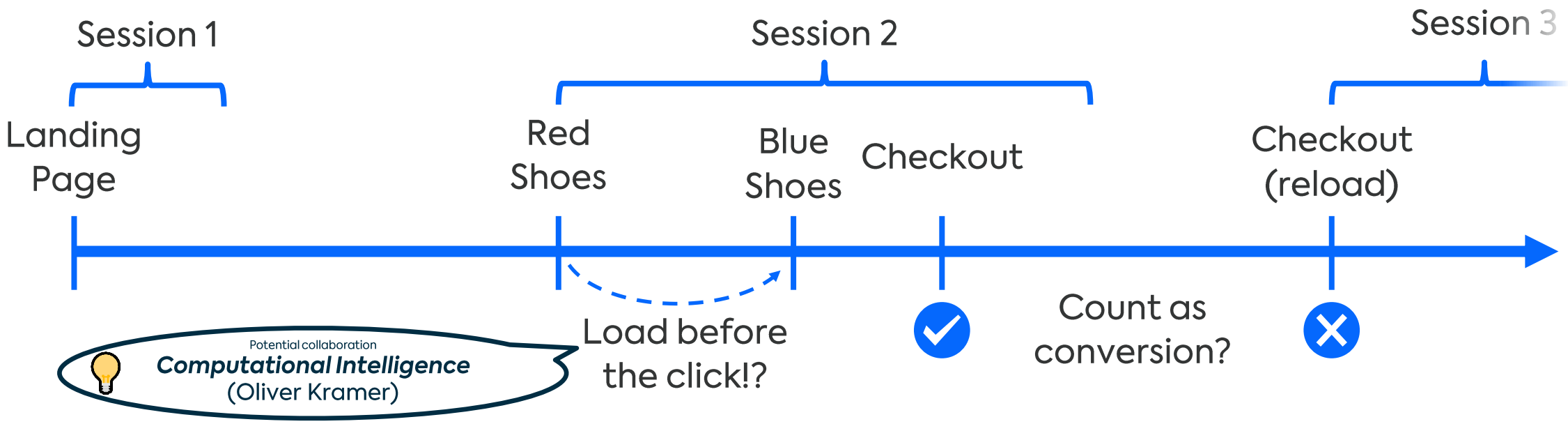
Kontinuierliche Uplift-Visualisierung







Challenges: Optimierung, Vorhersage & Effizienz



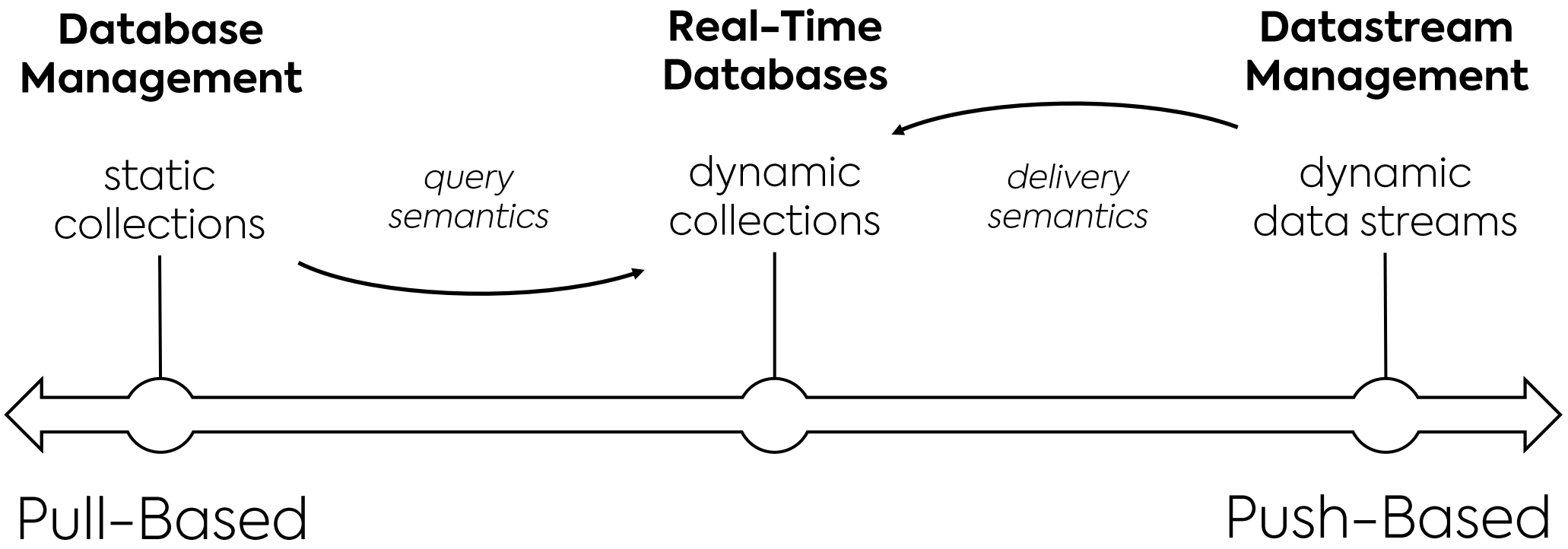
- **User Behavior Analysis**: identify „successful“ user journeys, quantify satisfaction, optimize UX, …
- **Predictive Preloading**: click path prediction (ML), end-to-end system integration, monitoring, …
- **Delta Encoding**: delta computation, caching strategy (dictionary generation, personalized content handling, content invalidation), browser support, preloading efficiency optimization, …

Potential collaboration
Embedded HW/SW Systems
(Verena Kloes)

In Collaboration With fastly Google

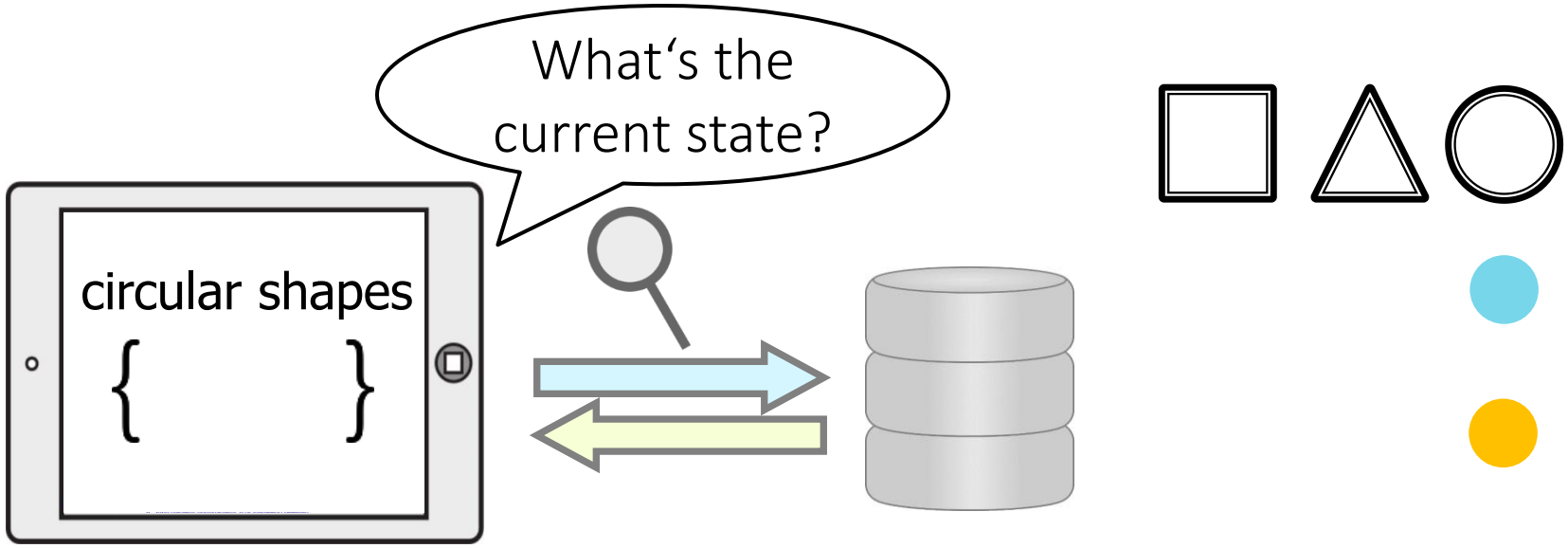
Push vs. Pull: Trade-Offs im Datenmanagement

Perfect match
Informationssysteme
(Marco Grawunder & Ralf Krause)



W. Wingerath, F. Gessert, N. Ritter. Real-Time & Stream Data Management: Push-Based Data in Research & Practice, Springer International Publishing (2019)

Traditionelle DBs: Kein Request, Keine Daten!

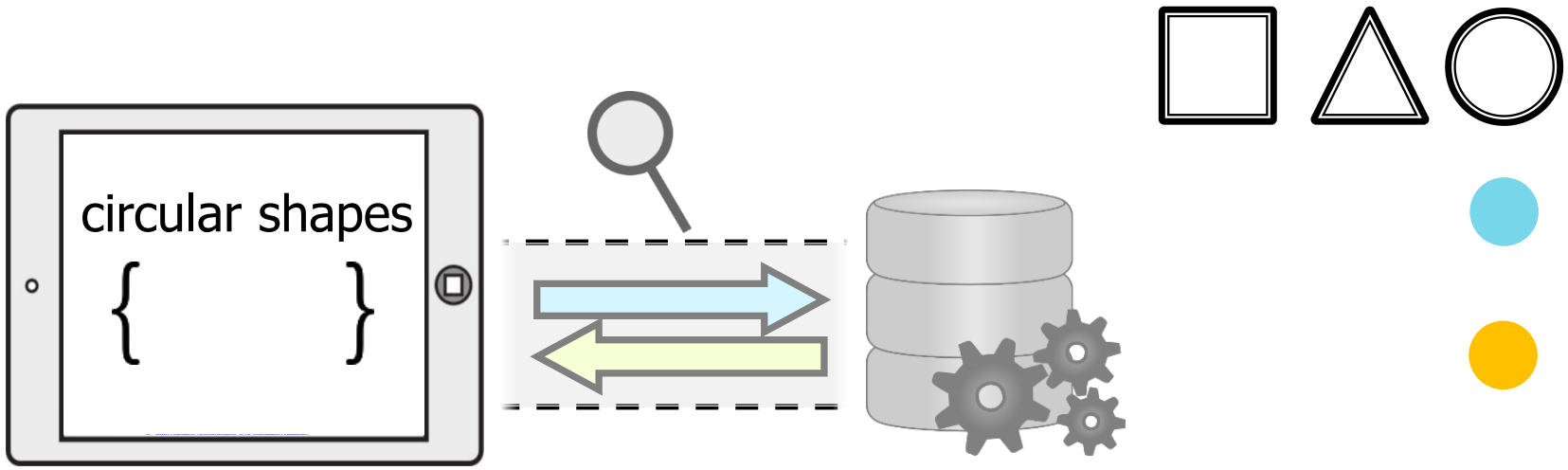


Periodic Polling for query result maintenance:

→ inefficient

→ slow

Echtzeit-Datenbanken: Immer Aktuell



Real-Time Queries for query result maintenance:

→ efficient

→ fast

Echtzeit-Datenbanken: Challenge

	METEOR	RethinkDB	Parse	Firebase
	Poll-and-Diff	Change Log Tailing		Unknown
Write Scalability	✓	✗	✗	✗
Read Scalability	✗	✓	✓	✓	? (100k connections)
Composite Filters (AND/OR)	✓	✓	✓	✓	○ (AND In Firestore)
Sorted Queries	✓	✓	✓	✗	○ (single attribute)
Limit	✓	✓	✓	✗	✓
Offset	✓	✓	✗	✗	○ (value-based)
Self-Maintaining Queries	✓	✓	✗	✗	✗
Event Stream Queries	✓	✓	✓	✓	✓

W. Wingerath, F. Gessert, N. Ritter. InvaliDB: Scalable Push-Based Real-Time Queries on Top of Pull-Based Databases (Extended), VLDB (2020)

Echtzeit-Datenbanken: Challenge

	METEOR	RethinkDB	Parse	Firebase	InvaliDB
	Poll-and-Diff	Change Log Tailing		Unknown	2-D Partitioning
Write Scalability	✓	✗	✗	✗	✗	✓
Read Scalability	✗	✓	✓	✓	? (100k connections)	✓
Composite Filters (AND/OR)	✓	✓	✓	✓	○ (AND In Firestore)	✓
Sorted Queries	✓	✓	✓	✗	○ (single attribute)	✓
Limit	✓	✓	✓	✗	✓	✓
Offset	✓	✓	✗	✗	○ (value-based)	✓
Self-Maintaining Queries	✓	✓	✗	✗	✗	✓
Event Stream Queries	✓	✓	✓	✓	✓	✓

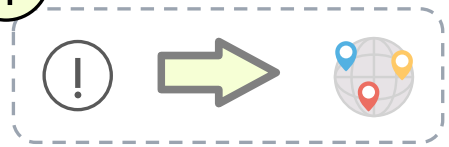
W. Wingerath, F. Gessert, N. Ritter. InvaliDB: Scalable Push-Based Real-Time Queries on Top of Pull-Based Databases (Extended), VLDB (2020)

RaaS: Realtime-as-a-Service

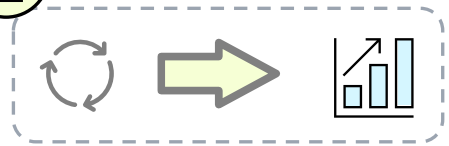
Potential collaboration
System Software & Distributed Systems
(Oliver Theel)

Potential collaboration
Software Engineering
(Andreas Winter)

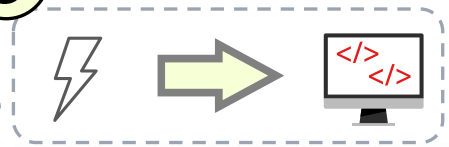
1 (Query) Cache Invalidation, Anomaly Detection & Alerting

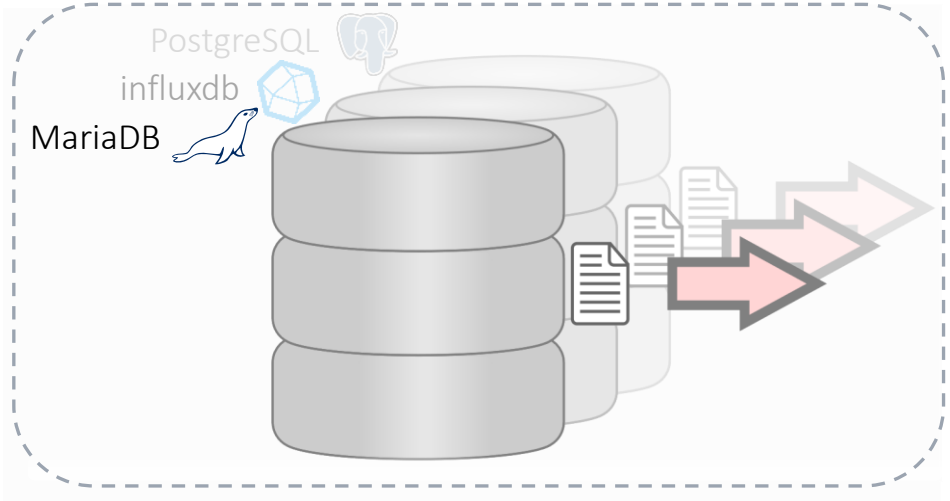


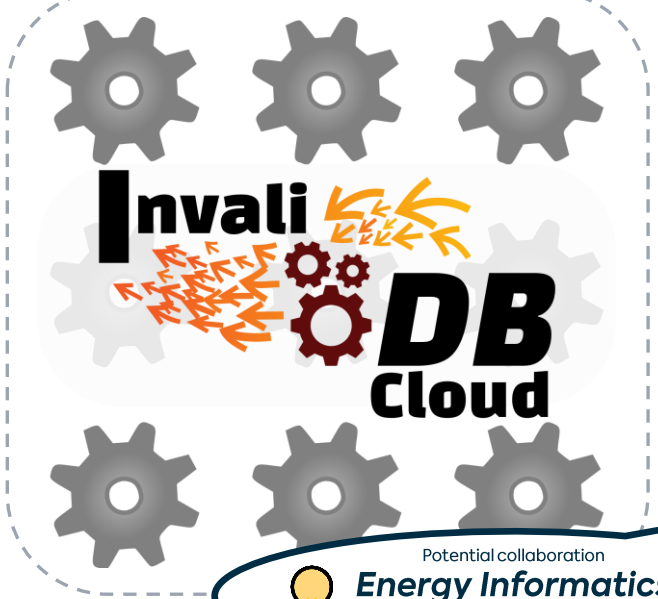
2 Query Maintenance, Continuous Aggregation & Dashboarding



3 Predictive Analytics, Hotspot Forecasting & Handling







Potential collaboration
Energy Informatics
(Sebastian Lehnhoff)

Carl von Ossietzky Universität Oldenburg

Otto von Guericke Universität Magdeburg

Database Systems
(NoSQL & SQL)

InvaliDB Cloud
(Continuous Data Processing)

Use Cases
(Consumers / Plugins)

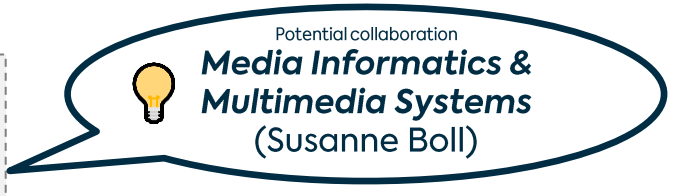
Potential collaboration
NordLink Data Center
(OpenWebSearch)

M. Kuhn, W. Wingerath. Storage System Insights for Predictive Performance and Autonomous I/O Optimizations (SIPPIO), DFG Proposal Draft

RaaS Plugin: Query Result Change Notifications

2

```
var query = DB.Tweet.find()
              .matches('text', /my filter/)
              .descending('createdAt')
              .limit(10)
              .offset(20);
```



Pull-Based Query

```
query.resultList(result => ...);
```

Google

Real-Time Query

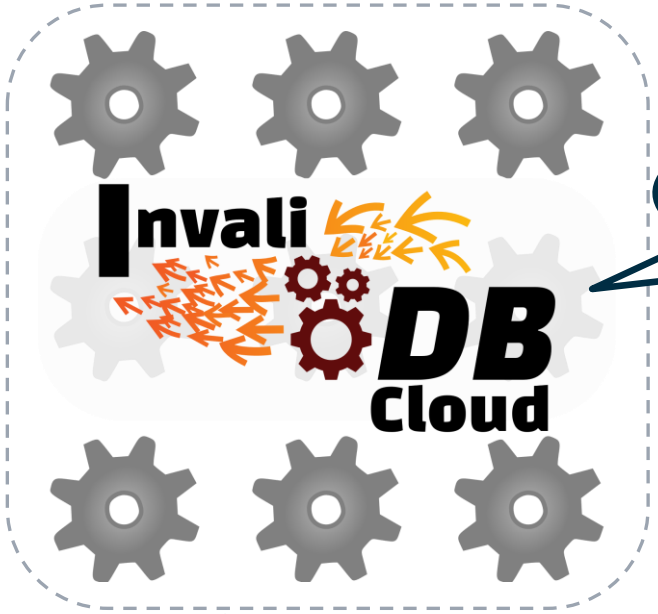
```
query.resultStream(result => ...);
```

Twoogle

W. Wingerath, F. Gessert, N. Ritter. Twoogle: Searching Twitter With MongoDB Queries, BTW (2019)

RaaS Plugin: Query Result Cache Invalidation

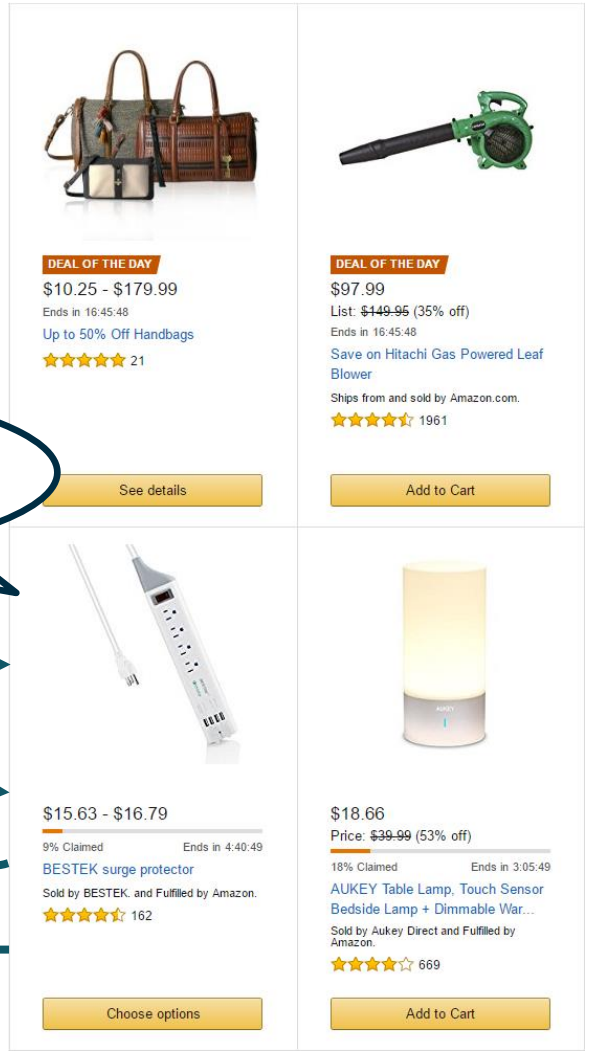
2



Enables **query result caching** by generating cache invalidations with low latency

HTML is rendered from user request: „Give me the most popular products that are in stock.“

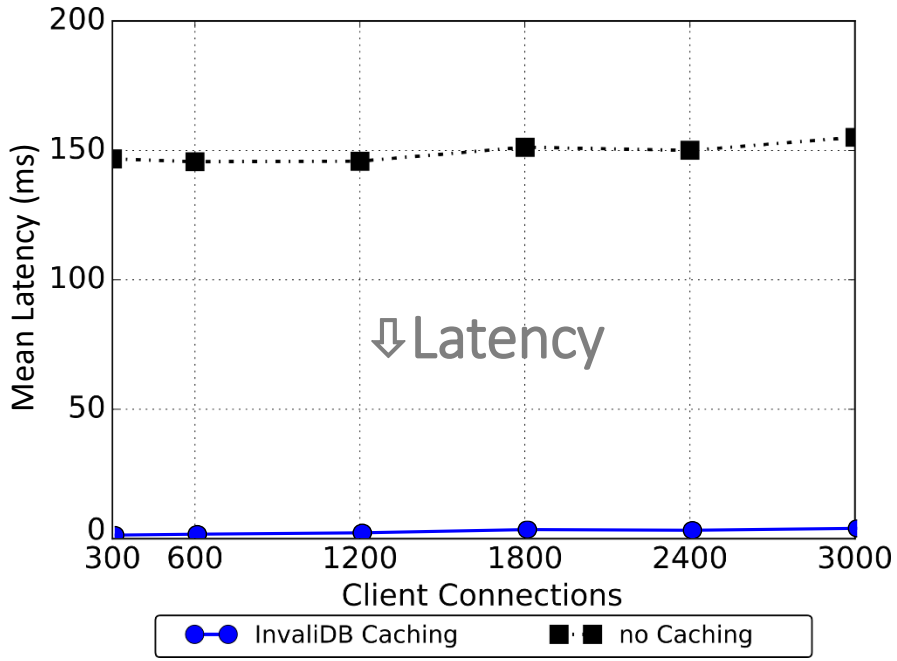
Potential collaboration
Safety-Security Interaction
(Andreas Peter)

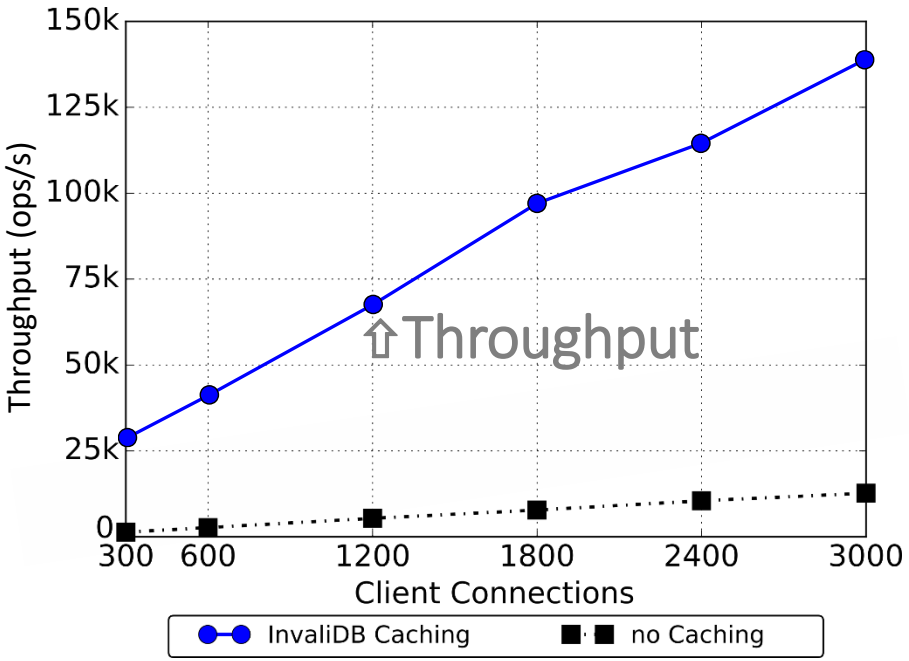


F. Gessert, M. Schaarschmidt, W. Wingerath, E. Witt, E. Yoneki, N. Ritter. Quaestor: Query Web Caching for Database-as-a-Service Providers, VLDB (2017)

RaaS Plugin: Query Result Cache Invalidation

2



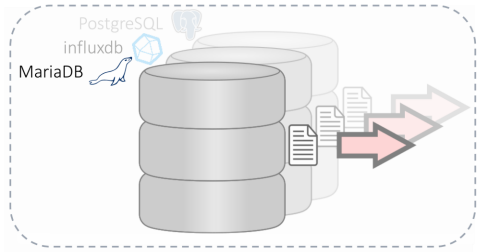


Improvement by 1-2 Orders of Magnitude

F. Gessert, M. Schaarschmidt, W. Wingerath, E. Witt, E. Yoneki, N. Ritter. Quaestor: Query Web Caching for Database-as-a-Service Providers, VLDB (2017)

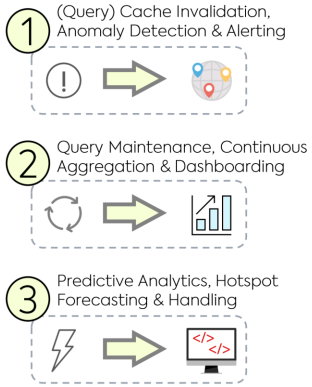
Challenges: Integration, Standardisierung, SQL

	Existing Work	Open Challenges
Integration	• Data sourcing requires all writes to go through application server	• Database extension / plugin with **no external dependencies**
Standardization	• Hard-wired to proprietary Baqend server and client implementations	• Extension of current protocols & integration into **standard tooling**
SQL Expressiveness	• Prototype only supports sorted filter queries with limit / offset in MongoDB	• Full **SQL expressiveness** with features like joins & aggregations



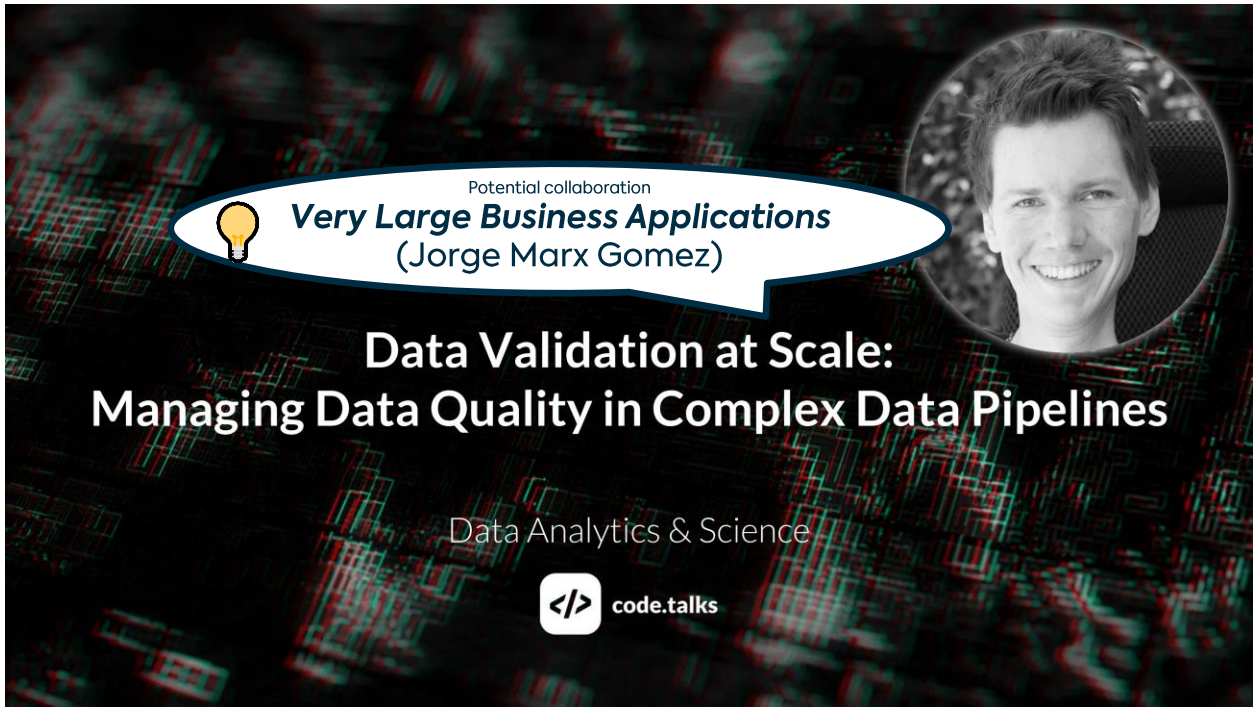
Database Systems
(NoSQL & SQL)

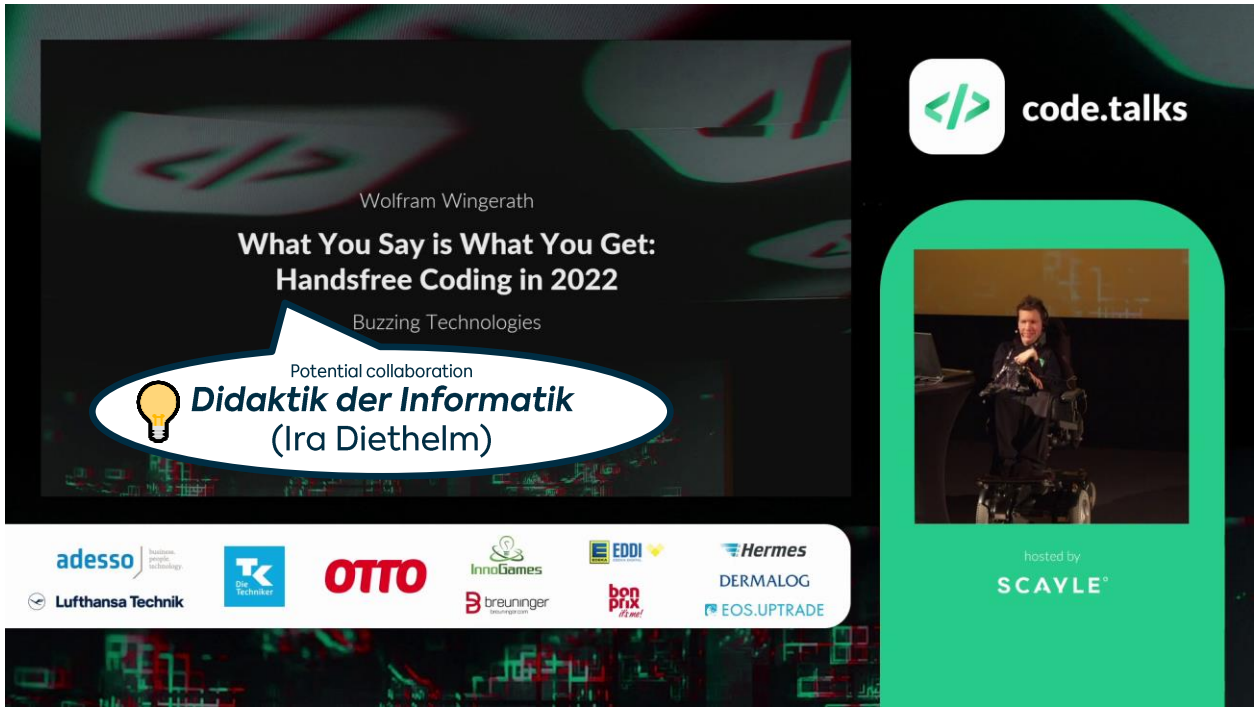
InvaliDB Cloud
(Continuous Data Processing)



Use Cases
(Consumers / Plugins)

Mehr Themen & Aktivitäten





Hamburg, Bamberg, Aachen, ...
(Guest) Lecturer & Co.-Organizer

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

RWTH AACHEN UNIVERSITY

BTW 25 Bamberg

BUCERIUS LAW SCHOOL
HOCHSCHULE FÜR RECHTSWISSENSCHAFT

UNIVERSITÄT BAMBERG

Tech Conferences & Workshops
Regular Speaker at Local Events

aric SPEAKER AT BROWN-BAGS

Meetup

techcamp HAMBURG

code.talks

Wolfram Wingerath. Data Validation at Scale: Managing Data Quality in Complex Data Pipelines, CodeTalks (2023)

Wolfram Wingerath. What You Say is What You Get: Hands-Free Coding in 2022, CodeTalks (2022)

Warum Handsfree Coding!?

Productivity

- Speed up input-heavy tasks
- Faster navigation through easy-to-remember shortcuts

Convenience

- Intuitive interfaces
- Relieve your hands

Accessibility

Compensate handicaps:

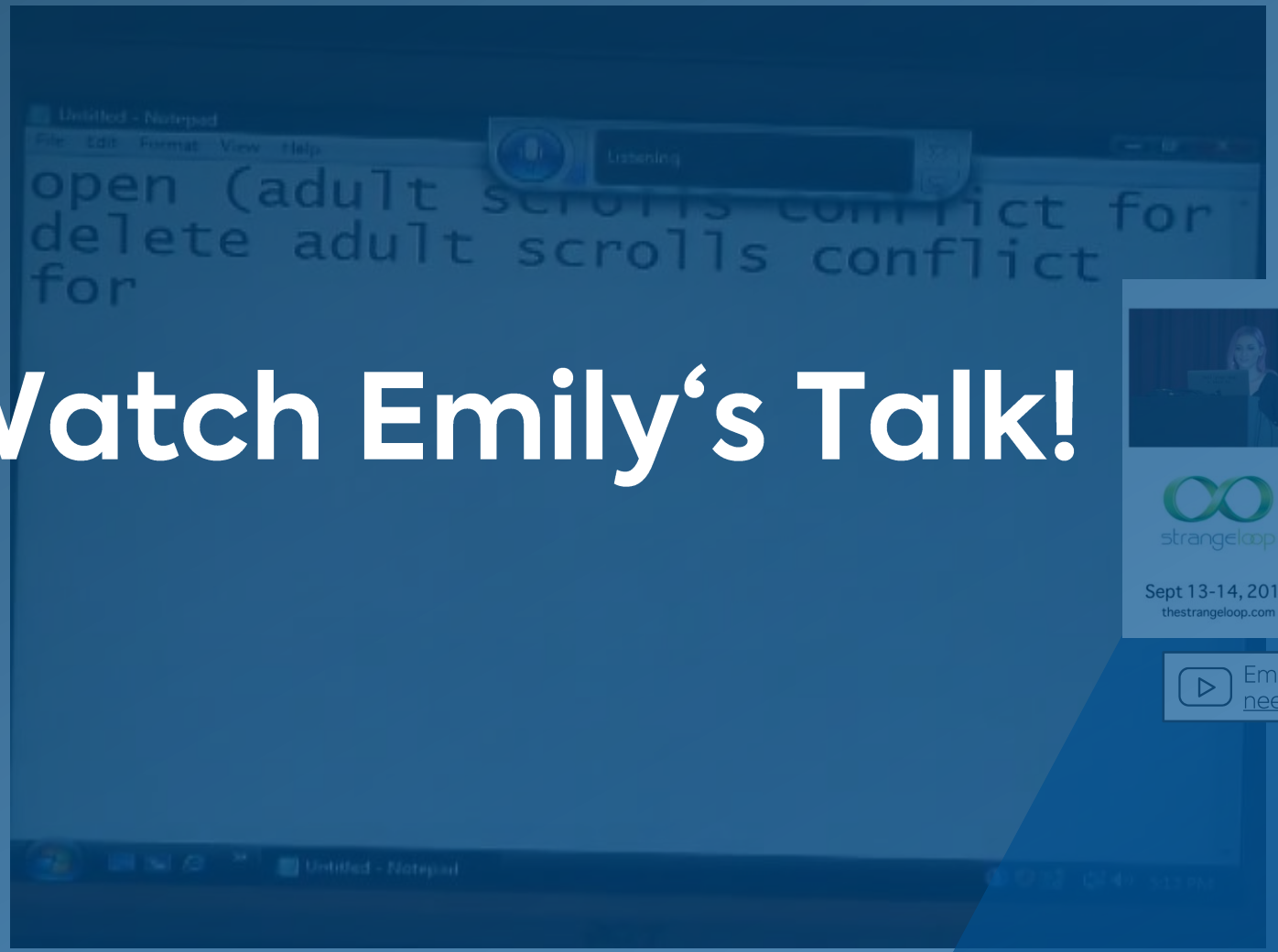
- Injuries (e.g. broken hand)
- Repetitive stress injury (RSI)
- Cubital Tunnel Syndrome
- ...

General Awesomeness

- Talk to your computer!!!

Lasst uns mal schauen, wie easy Easy das ist …

3



Go Watch Emily‘s Talk!

Emily Shea. Voice Driven Development: Who needs a keyboard anyway?, Strange Loop (2019)

scrubadub1. Windows Vista Speech Recognition Tested - Perl Scripting, YouTube, 2007

Idea to use this video blatantly stolen from: Emily Shea. Voice Driven Development: Who needs a keyboard anyway?, Strange Loop, 2019

Und warum ist das jetzt so schwer?

Und warum ist das jetzt so schwer?

WSR, Dragon, …

- **Automatic Speech Recognition (ASR)**: optimized for natural languages
  1. Signal processing extracts features from audio recording
  2. Acoustic model recognizes phonemes
  3. Language model finds a matching sequence of words:

  → Default: Every utterance is interpreted as (spoken) text

  (Commands only through special keywords)

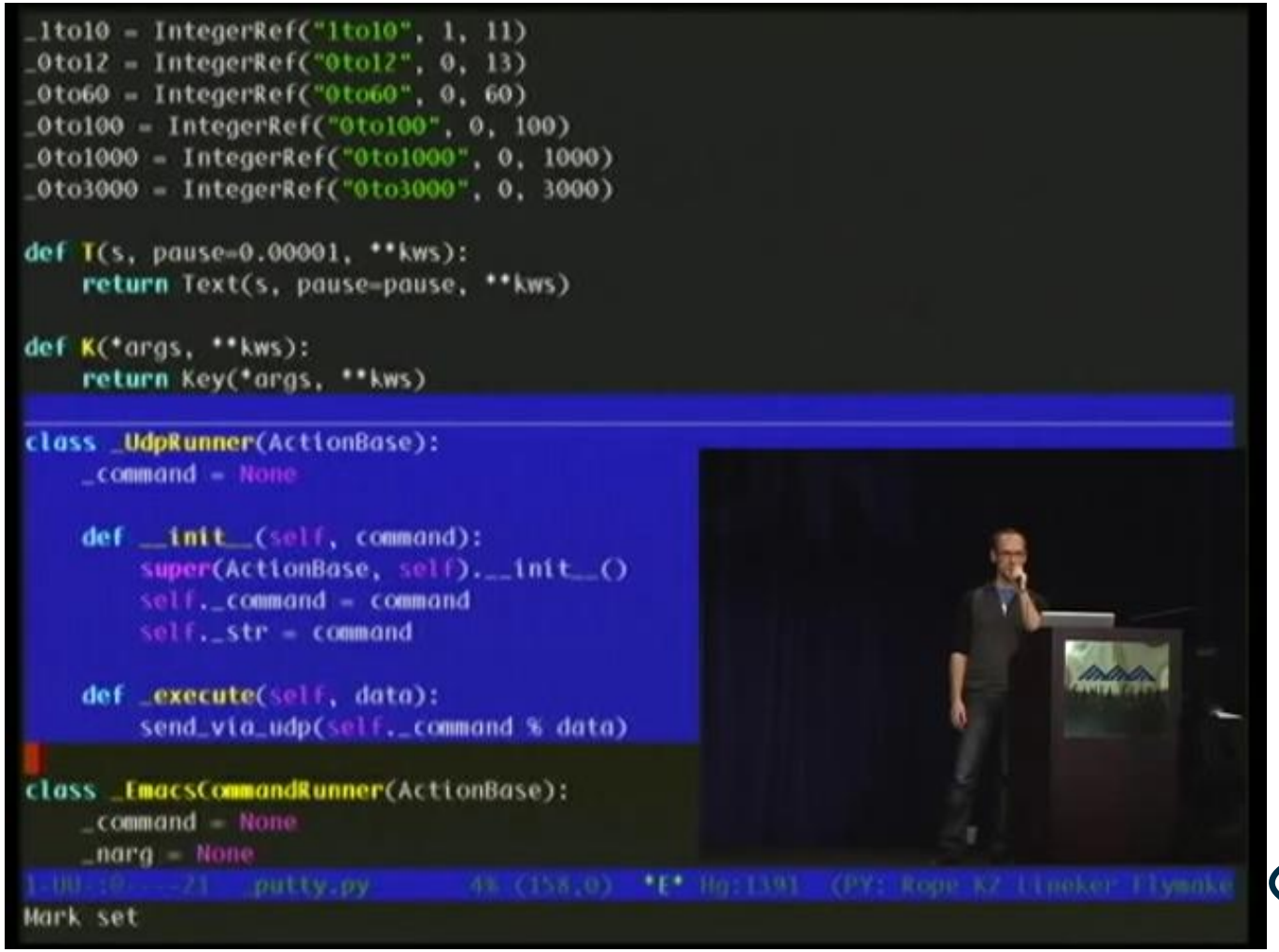
Talon, Dragonfly …

- **Voice Coding**: optimized for actions & programming languages

  → Default: Everything is interpreted as a command

  (Natural language through special keywords, e.g. `say <utterance>`)

Handsfree Coding: Ehrlich Jetzt …

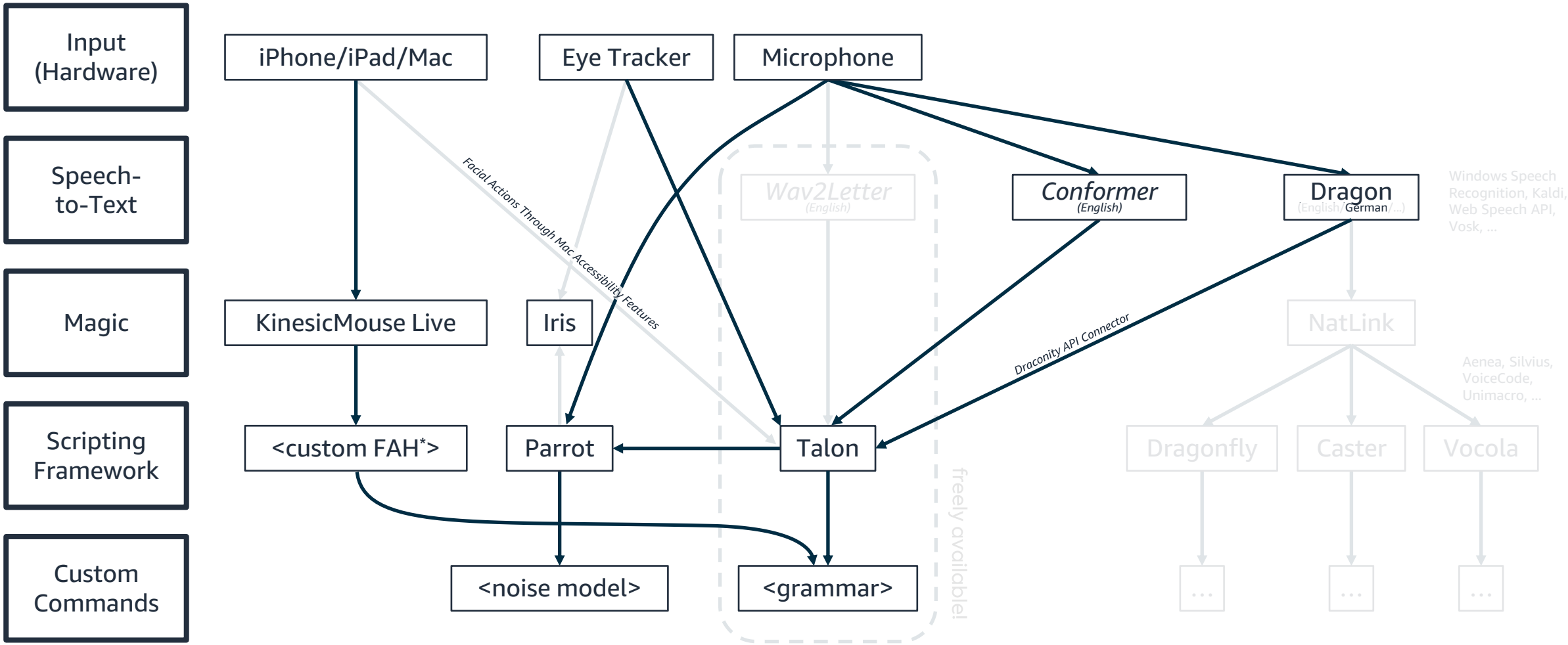


Using Dragonfly!

Tavis Rudd. Using Python to Code by Voice, PyCon US (2013)


The Base Setup

Mein Handsfree Coding Stack: Vereinfacht



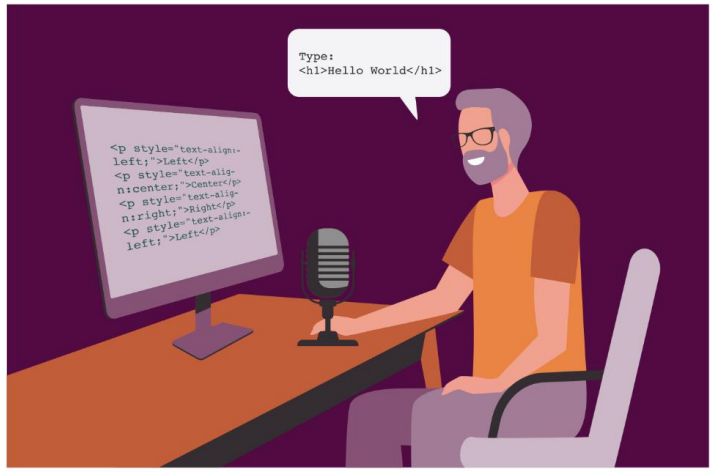
*Facial Action Handling
Please note that this overview is NOT complete: On every level, there are MANY other options!

This overview was inspired by: https://dictation-toolbox.github.io/dictation-toolbox.org/ (accessed: January 4, 2021)

Multi-Computer Setup

Heise-Artikel

REPORT | SOFTWAREENTWICKLUNG



Softwareentwicklung ohne Maus und Tastatur

Sprechen ist das neue Klicken

Dr. Wolfram Wingerath, Michaela Gebauer

Für die Bedienung des Computers brauchte man viele Jahre Maus und Tastatur – heute kann man mit Sprache, Gestik und Mimik sogar programmieren.

zung des Computers ganz ohne Einsatz ihrer Hände."

Wolle ist 33 Jahre alt, Data Engineer und erprobt seit mehr als zehn Jahren Eingabemethoden zur Softwareentwicklung ohne Maus und Tastatur. Inzwischen setzt er fast ausschließlich auf Handsfree Coding, da er damit effizienter arbeitet. „Dadurch muss ich mir keine kryptischen Shortcuts mehr merken und kann ganz bequem mit Sprache, Geräuschen, Mimik oder Gestik den Computer und die Programme steuern", sagt er.

Beim Handsfree Coding spielt das Voice Coding eine zentrale Rolle. Hierbei wird Quellcode per Spracheingabe erstellt. Voice Coding ist jedoch nicht mit handelsüblicher Software zur automatischen Spracherkennung (Automatic Speech Recognition, ASR) vergleichbar. Es gibt zwar einige offensichtliche Parallelen zum Diktieren von Textnachrichten. Mit Standardsoftware zur Spracherkennung kann man aber nicht ohne Weiteres effizient programmieren, da ASR auf die Interpretation und Synthese einer konkreten natürlichen Sprache ausgelegt ist. Sie verwendet dafür jeweils spezifische Modelle, Grammatiken und Optimierungen bei der Ausgabe, etwa, wenn sie automatisch Satzzeichen einfügt oder Substantive großschreibt. Bei typischer ASR-Software sind Befehle stets mit einem Schlüsselwort einzuleiten und durch Sprechpausen abzuschließen. Während sich so einfache Tastenaktionen umsetzen lassen – etwa mit der Aussage „press Enter" zum Drücken der Eingabetaste –, ist die Ausführung von komplexen Aktionen oder Aktionssequenzen eher beschwerlich und ineffizient.

Wolfram Wingerath, Michaela Gebauer: Sprechen ist das neue Klicken, iX 9/2021 (https://wingerath.cloud/2021/ix)


Look,
No Hands!

Handsfree Gaming: Eyes + Face + Voice/Noise

wolle.science/twitch

Und wie geht's jetzt Weiter?

1. **Continuous & Predictive Analytics**
   UX quantification, anomaly & bot detection, user behavior analysis, click path prediction, content preloading, delta encoding, ...

2. **Data Engineering**
   Realtime-as-a-Service for Legacy infrastructure, autonomous storage optimization, query result caching, ...

3. **Quality of Life**
   data quality, validation & synthesis, workflow efficiency, handsfree coding & gaming, ...

Danke für die Aufmerksamkeit!