

Benchmarking & Experimentelle Bewertung

UV+SE AI Werkstatt – Wintersemester 25/26

Roland Kwitt / Christine Bauer / Frank Pallas



Heute:

→ Benchmark / Experiment Design

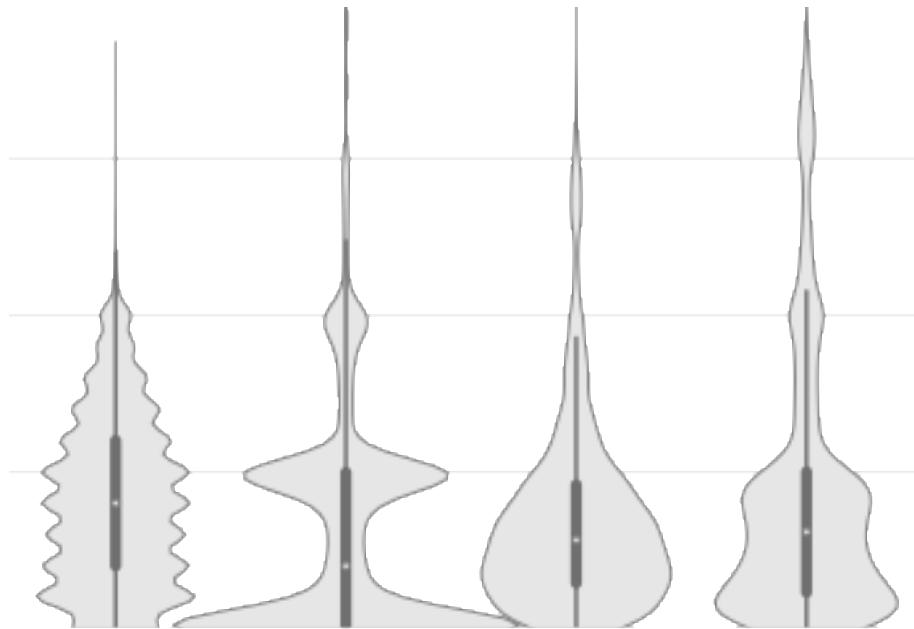
→ Durchführung von Benchmarks / Experimenten

→ Beispiele

Disclaimer:

Viele Beispiele aus dem Benchmarking verteilter Systeme
→ Transfer sollte möglich sein...

Disclaimer 2: Aber nicht nur



Heute:

→ Benchmark / Experiment Design

→ Durchführung von Benchmarks / Experimenten

→ Beispiele

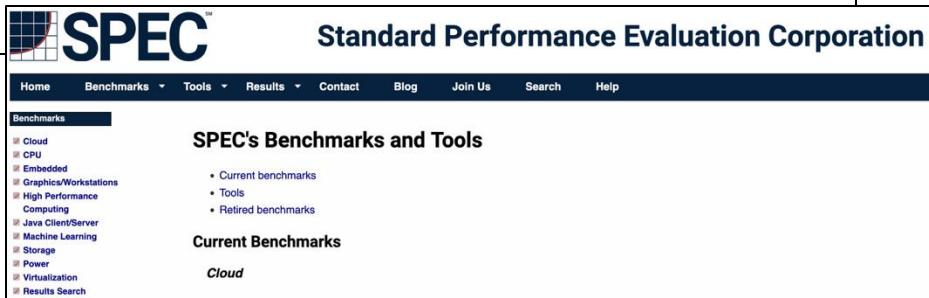
Wozu Benchmarking / Experimentation?

Grund / Ziel: Erfassen von **Qualitätseigenschaften** eines Systems / Dienstes – z.B.

- Performance (result quality – Qualität von Ergebnissen)
- Performance (speed – Durchsatz, Latenz, ...) → Schon Begrifflichkeiten schwer!
- Konsistenz von Daten / Ergebnissen
- Kosten (fixed, variable)
- (diverse Verfügbarkeitsdimensionen)
- ...

Benchmark

Ein/eine **Benchmark** (von englisch *benchmark* oder *bench mark*) ist ein Vergleichsmaßstab. **Benchmarking** (sinngemäß „Maßstäbe vergleichen“) bezeichnet die vergleichende Analyse von Ergebnissen oder Prozessen mit einem festgelegten **Bezugswert** oder Bezugsprozess.



- Benchmarks sollen „fair“ vergleichen, Experimente sollen (u.a.) bewerten
- „Regeln“ für Benchmarks sind strenger, führen aber auch zu „guten“ Experimenten

The Art of Building a Good Benchmark

Karl Hinspler
IBM Corporation
IBM IMS XDR
3605 Highway 52 North
Rochester, MN 55901
hinspler@us.ibm.com

Abstract. What makes a good benchmark? This is a question that has been asked often, answered often, altered often. In the past 25 years, the information processing industry has seen the creation of dozens of "industry standard" performance benchmarks – some highly successful, some less so. This paper will explore the overall requirements of a good benchmark, using existing industry standards as examples along the way.

1. Introduction – Building a Good Benchmark

Why so many benchmarks? The cynic would say "They haven't got it right, yet." The pessimist would say "They'll never get it right, but they keep on trying." The realist knows "The computing industry is so vast and changes so rapidly that new benchmarks are constantly required, just to keep up."

TPC-D SYSmart2007 Storage Performance Council

TPC-A SPECjM2005 SPECjms2007

TPC-B SPECcs2008 TPCE

TPC-C spec TPC-H SPC-2C

TPC-R SPECjvm2006 SPC-2B

TPC-App SPECjvm2003 SPC-2A

BAPCO Real World, Real Benchmarks

SPECint2000 SPECint_rate2000 SPC-1C

SPECint2006 SPECint_rate2007

K. Hinspler und M. Pfeifer (Eds.): TPC/TC 2009, LNCS 5895, pp. 18–30, 2009.
© Springer-Verlag Berlin Heidelberg 2009

Fünf Schlüsseleigenschaften eines “guten Benchmarks”:

- **Relevanz:** Lesende der Resultate glauben, dass der Benchmark etwas Wichtiges betrifft.
- **Wiederholbarkeit:** Es ist Wahrscheinlich, dass der Benchmark ein zweites Mal mit den gleichen Ergebnissen durchgeführt werden kann.
- **Fairness:** Alle zu vergleichenden Systeme und/oder Software werden gleichwertig in den Benchmark eingebunden.
- **Verifizierbarkeit:** Man geht davon aus, dass das dokumentierte Ergebnis echt ist.
- **Ökonomische Durchsetzbarkeit:** Die Geldgebenden können es sich leisten, den Benchmark durchzuführen

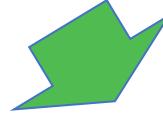
Fünf Schlüsseleigenschaften eines “guten Benchmarks”:

- **Relevanz:** Lesende der Resultate glauben, dass der Benchmark etwas Wichtiges betrifft.
- **Wiederholbarkeit:** Es ist Wahrscheinlich, dass der Benchmark ein zweites Mal mit den gleichen Ergebnissen durchgeführt werden kann.
- **Fairness:** Alle zu vergleichenden Systeme und/oder Software werden gleichwertig in den Benchmark eingebunden.
- **Verifizierbarkeit:** Man geht davon aus, dass das dokumentierte Ergebnis echt ist.
- **Ökonomische Durchsetzbarkeit:** Die Geldgebenden können es sich leisten, den Benchmark durchzuführen

Zusätzliche Kosten durch Anwendung von Mechanismus / Konfiguration X



(Typischerweise) Einfluss der Performance auf System mit X im Vergleich zum
gleichen System ohne X



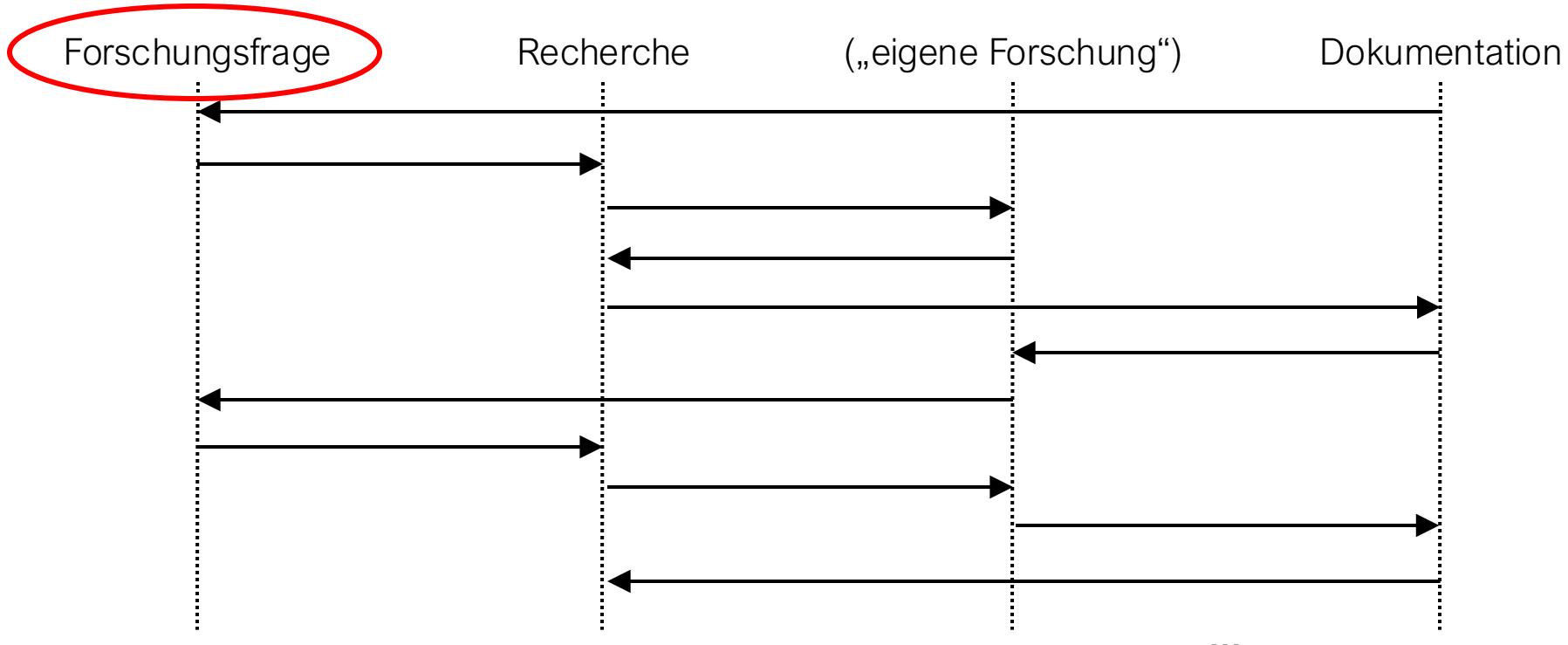
Einfluss der Performance bei
Nutzung gleicher Ressourcen

Einfluss auf die notwendigen
Ressourcen, um die gleiche
Performance zu erhalten

→ Es ist einfacher den Einfluss bei Nutzung gleicher Ressourcen zu messen,
aber der Einfluss auf die notwendigen Ressourcen ist meist interessanter.

- Klar definieren, was Sie experimentell „herausbekommen“ wollen
 - Für wen ist die Frage relevant und warum?
 - Was ist eine dafür geeignete Experimentumgebung?
 - Welche Parameter wollen Sie ändern / vergleichen und warum?
 - Ist das wirklich die interessanteste (gut) oder nur die am leichtesten beantwortbare (schlecht) Frage?

- Klar definieren, was Sie experimentell „herausbekommen“ wollen
 - Für wen ist die Frage relevant und warum?
 - Was ist eine dafür geeignete Experimentumgebung?
 - Welche Parameter wollen Sie ändern / vergleichen und warum?
 - Ist das wirklich die interessanteste (gut) oder nur die am leichtesten beantwortbare (schlecht) Frage?



- Klar definieren, was Sie experimentell „herausbekommen“ wollen
(ja, das kann sich auch ändern)
 - Für wen ist die Frage relevant und warum?
 - Was ist eine dafür geeignete Experimentumgebung?
 - Welche Parameter wollen Sie ändern / vergleichen und warum?
 - Ist das wirklich die interessanteste (gut) oder nur die am leichtesten beantwortbare (schlecht) Frage?

- Klar definieren, was Sie experimentell „herausbekommen“ wollen?
 - Für wen ist die Frage relevant und warum?
 - **Was ist eine dafür geeignete Experimentumgebung?**
- Welche Parameter wollen Sie ändern / vergleichen und warum?
- Ist das wirklich die interessanteste (gut) oder nur die am leichtesten beantwortbare (schlecht) Frage?

Relativ häufige Probleme bei Bewertung verteilter Systeme:

- Benchmarking des Klienten statt des Ziels
- Benchmarking des Netzwerks statt des Ziels
- Benchmarking des Systemstarts
- Benchmarking von völlig anderen Systemkomponenten / -eigenschaften
- ...

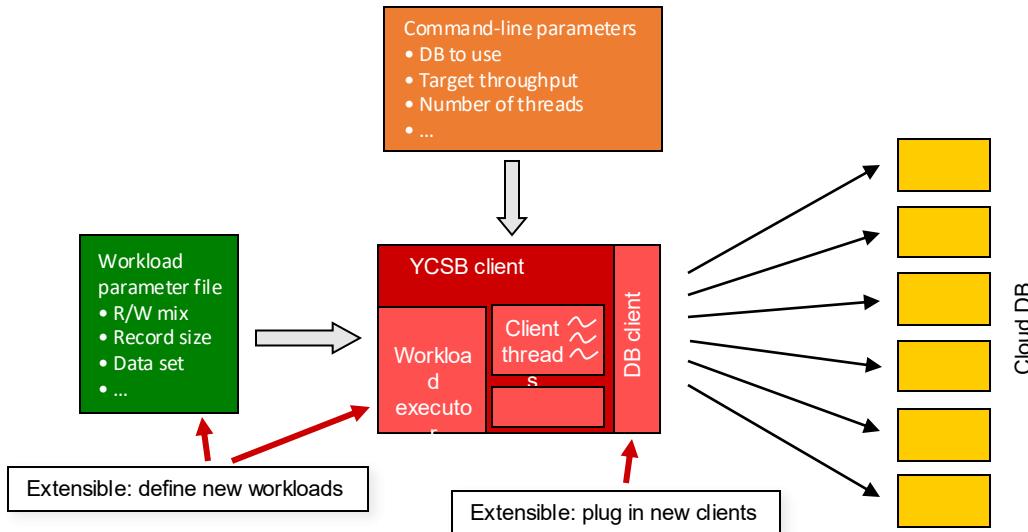


→ Versuchen Sie, derartige Probleme zu vermeiden!

Fünf Schlüsseleigenschaften eines “guten Benchmarks”:

- **Relevanz:** Lesende der Resultate glauben, dass der Benchmark etwas Wichtiges betrifft.
- **Wiederholbarkeit:** Es ist wahrscheinlich, dass der Benchmark ein zweites Mal mit den gleichen Ergebnissen durchgeführt werden kann.
- **Fairness:** Alle zu vergleichenden Systeme und/oder Software werden gleichwertig in den Benchmark eingebunden.
- **Verifizierbarkeit:** Man geht davon aus, dass das dokumentierte Ergebnis echt ist.
- **Ökonomische Durchsetzbarkeit:** Die Geldgebenden können es sich leisten, den Benchmark durchzuführen

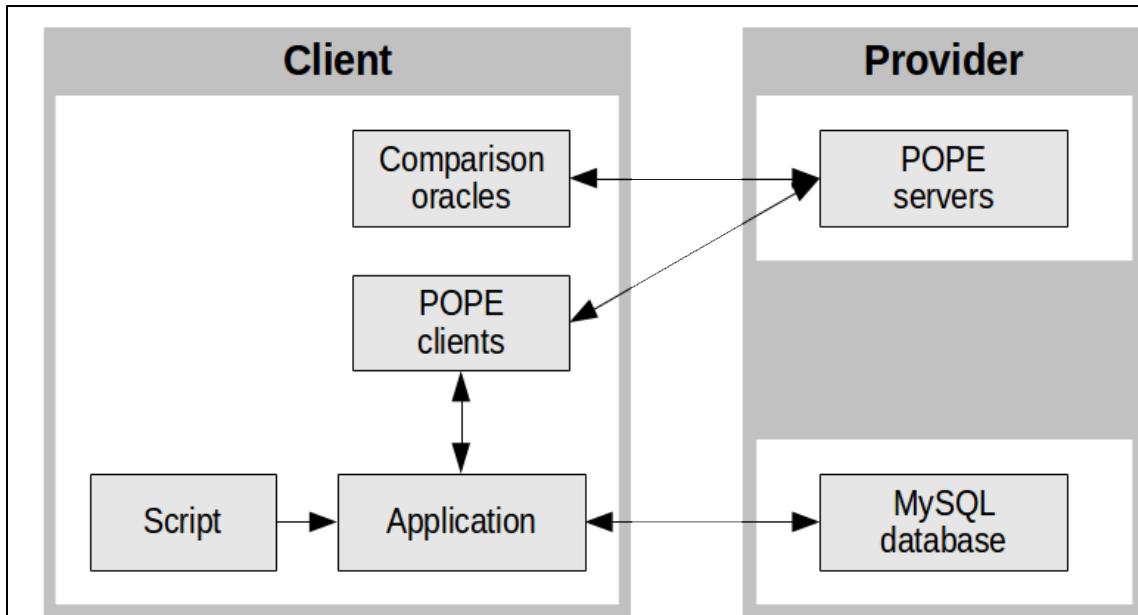
Wiederholbarkeit: Automatisierung



<http://wiki.github.com/brianfrankcooper/YCSB/>

B.F. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, R. Sears. "Benchmarking cloud serving systems with YCSB.", ACM SOCC, 2010

→ Nutzung von existierenden Benchmark Tools und Testdatensätzen, wann immer sie verfügbar und im angedachten System anwendbar sind



Pallas & Grambow (2018): Three Tales of Disillusion

→ Wenn existierende Tools / Datensätze nicht angemessen / möglich sind,
implementieren Sie eigene **und machen Sie diese verfügbar**

Fünf Schlüsseleigenschaften eines “guten Benchmarks”:

- **Relevanz:** Lesende der Resultate glauben, dass der Benchmark etwas Wichtiges betrifft.
- **Wiederholbarkeit:** Es ist Wahrscheinlich, dass der Benchmark ein zweites Mal mit den gleichen Ergebnissen durchgeführt werden kann.
- **Fairness:** Alle zu vergleichenden Systeme und/oder Software werden gleichwertig in den Benchmark eingebunden.
- **Verifizierbarkeit:** Man geht davon aus, dass das dokumentierte Ergebnis echt ist.
- **Ökonomische Durchsetzbarkeit:** Die Geldgebenden können es sich leisten, den Benchmark durchzuführen

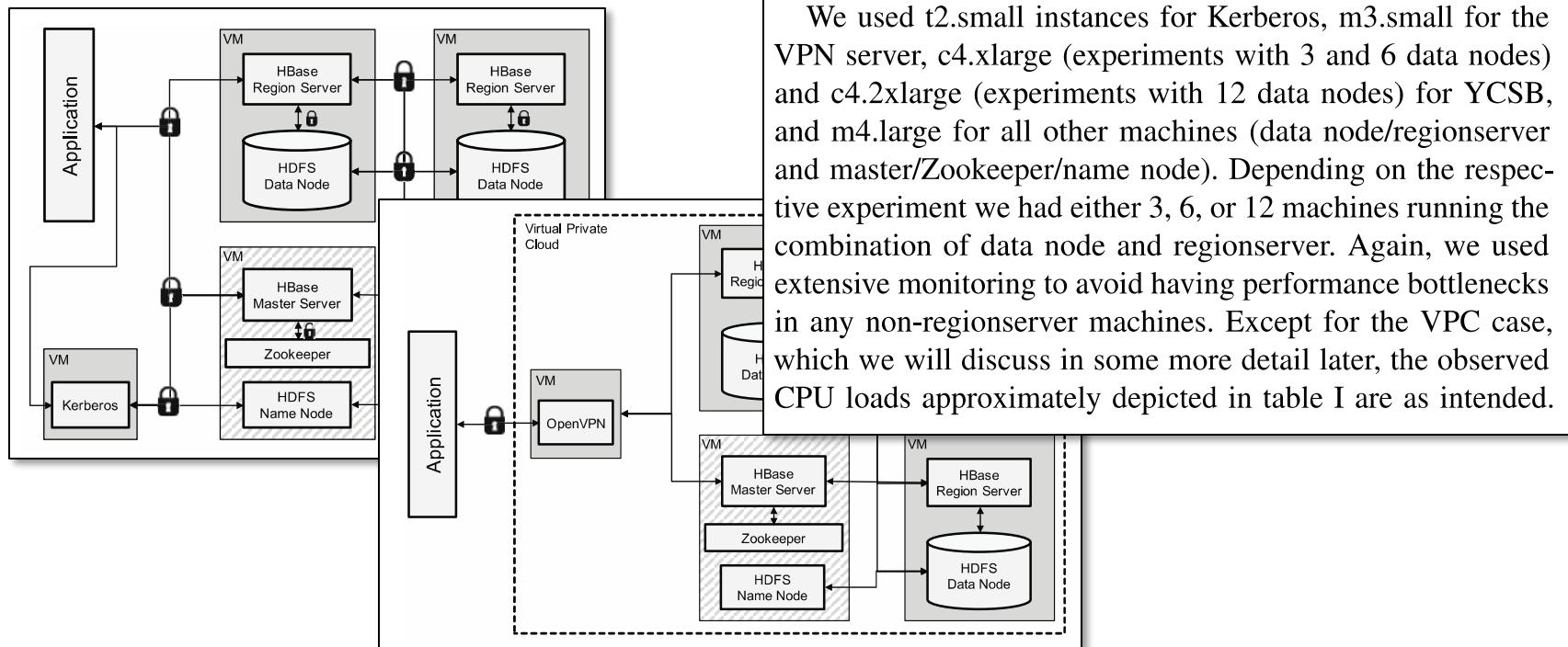
Fairness: Beispiel SPEC

Test Sponsor	System Name	Base Copies	Processor			Results	
			Enabled Cores	Enabled Chips	Threads/ Core	Base	Peak
ASRock Rack Inc.	1U4LW-X570 RPSU AMD Ryzen 7 5800X,3.8GHz	16	8	1	2	61.9	70.4
ASRock Rack Inc.	1U4LW-B650/2L2T RPSU AMD Ryzen 9 7950X	32	16	1	2	127	126
ASRock Rack Inc.	1U4LW-B650/2L2T RPSU AMD EPYC 4464P	12	12	1	2	123	126
ASRock Rack Inc.	1U4LW-B650/2L2T RPSU AMD EPYC 4564P	16	16	1	2	133	142
ASUSTeK Computer Inc.	ASUS RS700-E9(Z11PP-D24) Server System (2.70 GHz, Intel Xeon Gold 6150)	72	36	2	2	199	201
ASUSTeK Computer Inc.	ASUS RS700-E9(Z11PP-D24) Server System (2.10 GHz, Intel Xeon Platinum 8176)	112	56	2	2	233	237
ASUSTeK Computer Inc.	ASUS RS700-E9(Z11PP-D24) Server System (2.70 GHz, Intel Xeon Gold 6150)	72	36	2	2	199	202
ASUSTeK Computer Inc.	ASUS RS700-E9(Z11PP-D24) Server System (2.10 GHz, Intel Xeon Platinum 8176)	112	56	2	2	233	236
ASUSTeK Computer Inc.	ASUS WS C621E SAGE Server System (2.50 GHz, Intel Xeon Platinum 8180)	112	56	2	2	252	257
ASUSTeK Computer Inc.	ASUS RS720Q-E9(Z11PH-D12) Server System (2.70 GHz, Intel Xeon Gold 6150)	72	36	2	2	205	209
ASUSTeK Computer Inc.	ASUS RS720Q-E9(Z11PH-D12) Server System (2.10 GHz, Intel Xeon Platinum 8176)	112	56	2	2	241	244

Fünf Schlüsseleigenschaften eines “guten Benchmarks”:

- **Relevanz:** Lesende der Resultate glauben, dass der Benchmark etwas Wichtiges betrifft.
- **Wiederholbarkeit:** Es ist Wahrscheinlich, dass der Benchmark ein zweites Mal mit den gleichen Ergebnissen durchgeführt werden kann.
- **Fairness:** Alle zu vergleichenden Systeme und/oder Software werden gleichwertig in den Benchmark eingebunden.
- **Verifizierbarkeit:** Man geht davon aus, dass das dokumentierte Ergebnis echt ist.
- **Ökonomische Durchsetzbarkeit:** Die Geldgebenden können es sich leisten, den Benchmark durchzuführen

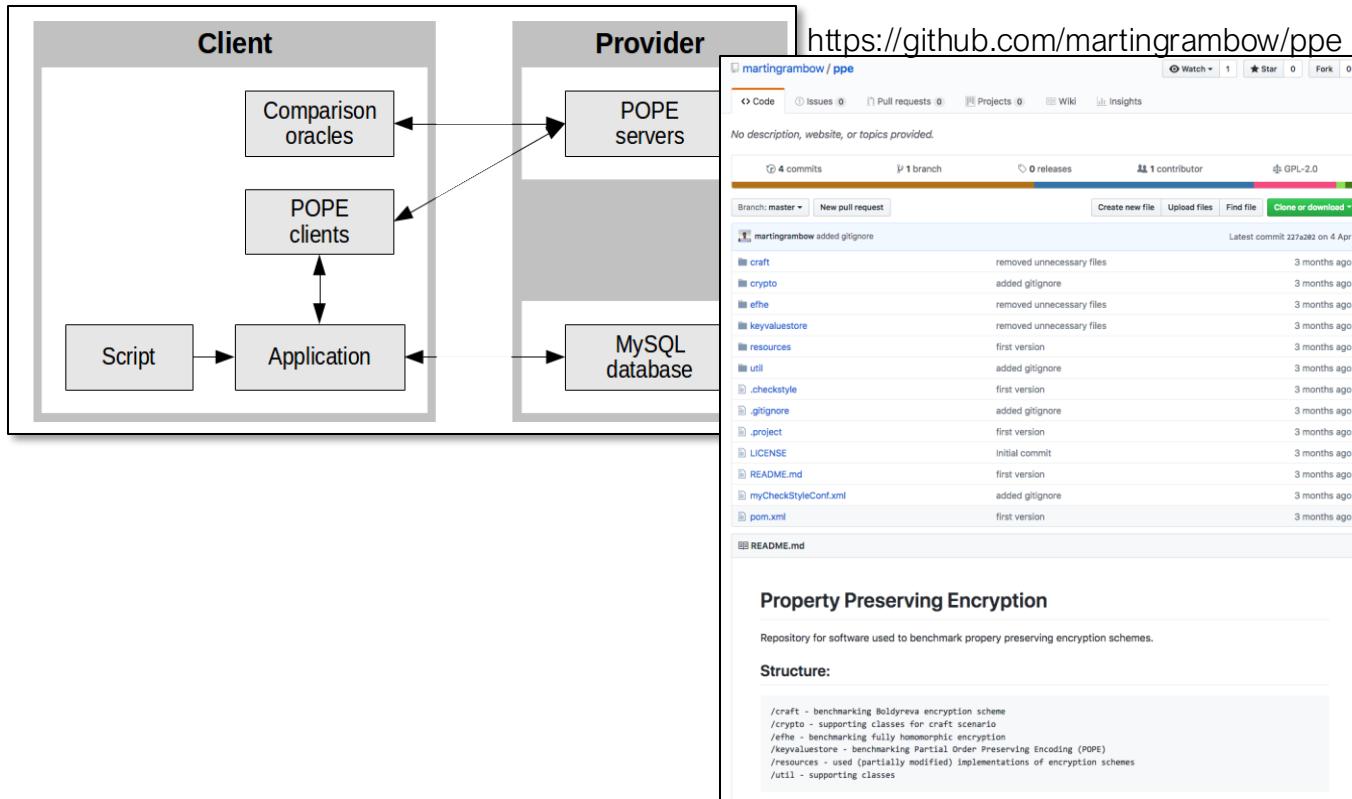
Verifizierbarkeit: Transparente Umgebung



We used t2.small instances for Kerberos, m3.small for the VPN server, c4.xlarge (experiments with 3 and 6 data nodes) and c4.2xlarge (experiments with 12 data nodes) for YCSB, and m4.large for all other machines (data node/regionserver and master/Zookeeper/name node). Depending on the respective experiment we had either 3, 6, or 12 machines running the combination of data node and regionserver. Again, we used extensive monitoring to avoid having performance bottlenecks in any non-regionserver machines. Except for the VPC case, which we will discuss in some more detail later, the observed CPU loads approximately depicted in table I are as intended.

Pallas, Günther, Bermbach (2016): Pick Your Choice in HBase

Verifizierbarkeit: Selbst implementierte Komponenten



Fünf Schlüsseleigenschaften eines „guten Benchmarks“:

- **Relevanz:** Ein Leser der Resultate glaubt, dass der Benchmark etwas Wichtiges betrifft.
- **Wiederholbarkeit:** Es ist Wahrscheinlich, dass der Benchmark ein zweites Mal mit den gleichen Ergebnissen durchgeführt werden kann.
- **Fairness:** Alle zu vergleichenden Systeme und/oder Software werden gleichwertig in den Benchmark eingebunden.
- **Verifizierbarkeit:** Man geht davon aus, dass das dokumentierte Ergebnis echt ist.
- **Ökonomische Durchsetzbarkeit:** Die Geldgeber können es sich leisten den Benchmark durchzuführen

Heute:

→ Benchmark / Experiment Design

→ Durchführung von Benchmarks / Experimenten

→ Beispiele

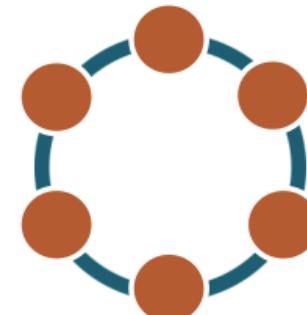
Allgemeines

- Verteilte, extrem skalierbare, ringförmige Datenbank
- Genutzt von Apple (100e PB, 300k Knoten), Netflix (58k Knoten, 20PB), eBay, ...
- Typische Nutzung: Eigene Installation die n öffentliche Cloudinstanzen umfasst



Transportverschlüsselung (Data in Transit Encryption)

- TLS
- Vollständige Liste aller Konfigurationsmöglichkeiten wie sie durch TLS bereitgestellt werden (keylength, op.-mode, HW-support, ...)



Welche Auswirkung haben unterschiedliche Konfigurationen der Transportverschlüsselung auf die „Geschwindigkeit“ von Cassandra?

6-stufiger Ansatz – ursprünglich für Sicherheit, aber auch für andere „Messziele“ anwendbar

Identifizierung relevanter Parameter & Trade-Offs

Reduktion der Parameter im Design Space

Reihenfolge der Präferenzen

Planung des Experiments

Ausführung des Experiments

Analyse der Ergebnisse

Für Details siehe Pallas, Bermbach, Müller, Tai (2017): Evidence-based...

|Channel_classes| (AR, RR, ..)
X

|Sym_algorithms| (AES, Camellia, ..)
X

|Key_lengths| (128, 256, ..)
X

|Op_modes| (CBC, GCM, ..)
X

|MAC_algorithms| (SHA, SHA256, ..)
X

... (SSL impl., HW support, etc.)...

6-stufiger Ansatz – ursprünglich für Sicherheit, aber auch für andere „Messziele“ anwendbar

Identifizierung relevanter Parameter & Trade-Offs

Reduktion der Parameter im Design Space

Reihenfolge der Präferenzen

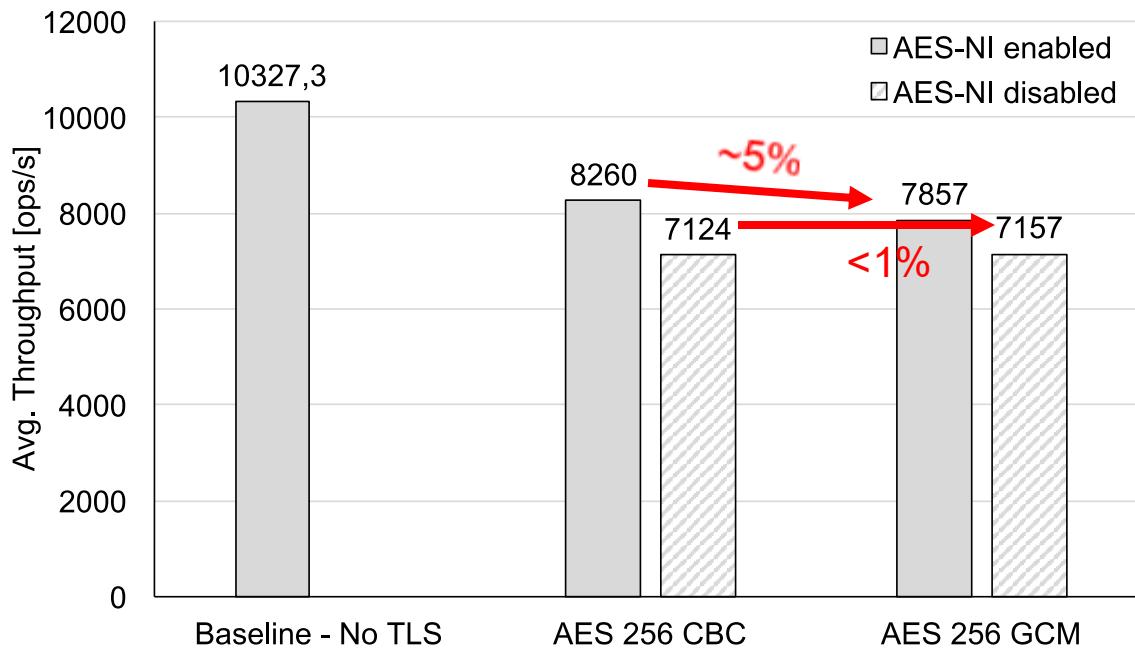
Planung des Experiments

Ausführung des Experiments

Analyse der Ergebnisse

Für Details siehe Pallas, Bermbach, Müller, Tai (2017): Evidence-based...

Cassandra TLS Optionen, 3 Knoten, 50/50 Last



→ Wenn AES-NI verfügbar, ist GCM „die 5% wert“?

→ Wenn AES-NI nicht verfügbar ist gibt es keinen Grund GCM nicht zu nutzen.

Heute:

→ Benchmark / Experiment Design

→ Durchführung von Benchmarks / Experimenten

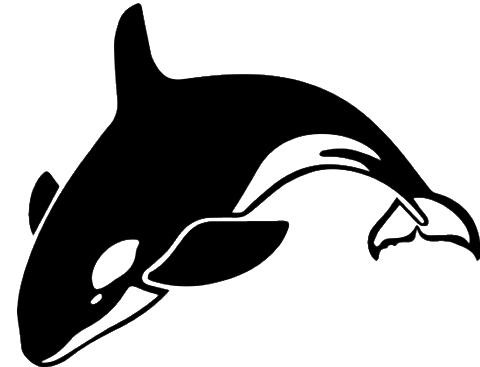
→ (Weitere) Beispiele

Allgemeines

- Zentrale Komponente des Apache Hadoop BigData Ökosystems
- Nutzung: Eigene Installation die n öffentliche Cloudinstanzen umfasst

Data in Transit Encryption

- native Verschlüsselung beruht auf Kerberos, Java SASL etc.
- Praktisch nur an/aus auf 2 „Ebenen“ / „Layern“
- Erwarteter Einfluss auf die Performance: “~10%”



APACHE
HBASE

45. Special Cases

- HBase and MapReduce
- 46. HBase, MapReduce, and the CLASSPATH
- 47. MapReduce Scan Caching
- 48. Bundled HBase MapReduce Jobs
- 49. HBase as a MapReduce Job Data Source and Data Sink
- 50. Writing HFiles Directly During Bulk Import
- 51. RowCounter Example
- 52. Map-Task Splitting
- 53. HBase MapReduce Examples
- 54. Accessing Other HBase Tables in a MapReduce Job
- 55. Speculative Execution
- 56. Cascading
- Securing Apache HBase
- 57. Using Secure HTTP (HTTPS) for the Web UI
- 58. Using SPNEGO for Kerberos authentication with Web UIs
- 59. Secure Client Access to Apache HBase

Once HBase is configured for secure RPC it is possible to optionally configure encrypted communication. To do so, add the following to the `hbase-site.xml` file on every client:

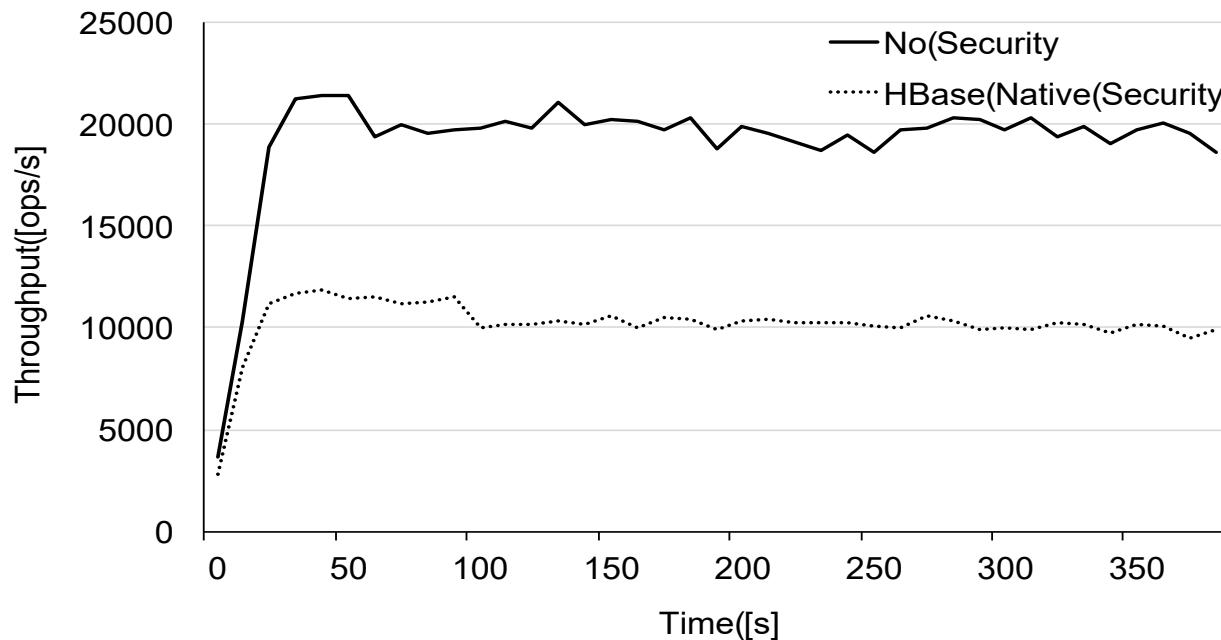
```
<property>
  <name>hbase.rpc.protection</name>
  <value>privacy</value>
</property>
```

This configuration property can also be set on a per-connection basis. Set it in the `Configuration` supplied to `Table`:

```
Configuration conf = HBaseConfiguration.create();
Connection connection = ConnectionFactory.createConnection(conf);
conf.set("hbase.rpc.protection", "privacy");
try (Connection connection = ConnectionFactory.createConnection(conf)) {
    try (Table table = connection.getTable(TableName.valueOf(tablename)) {
        .... do your stuff
    }
}
```

Expect a ~10% performance penalty for encrypted communication.

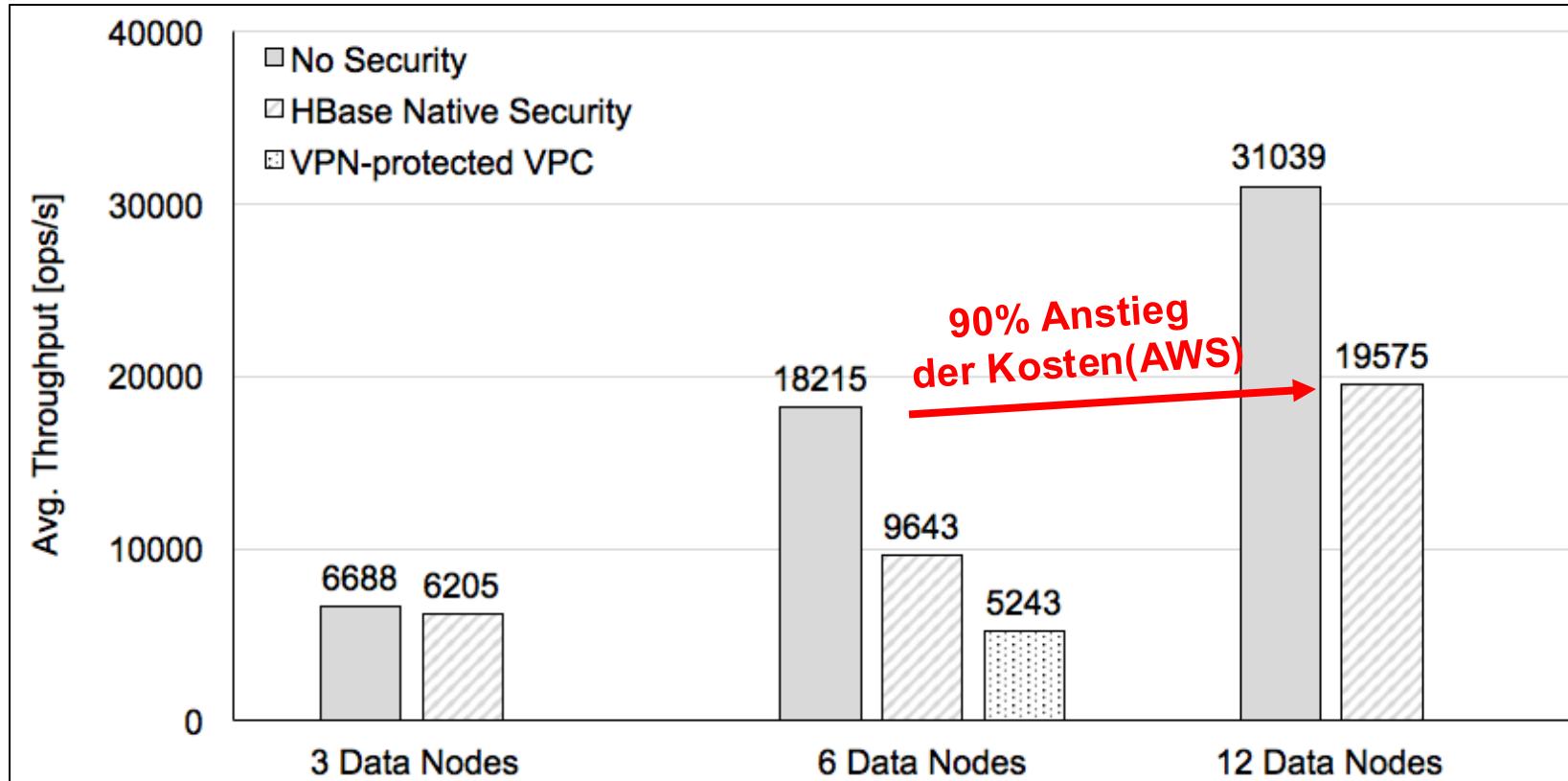
<https://hbase.apache.org/book.html#hbase.secure.configuration>



18,215 ops/s → 9,643 ops/s: durchschn. Durchsatz verschlechtert um ~47%

Experiments are fun!

HBase Verschlüsselung – Einfluss auf die Kosten



Allgemeines

- Hosted, managed Cloud Storage Service (Teil von AWS)
- Nutzung: Abgerechnet pro “Durchsatz Einheiten“ (read + write)



DynamoDB

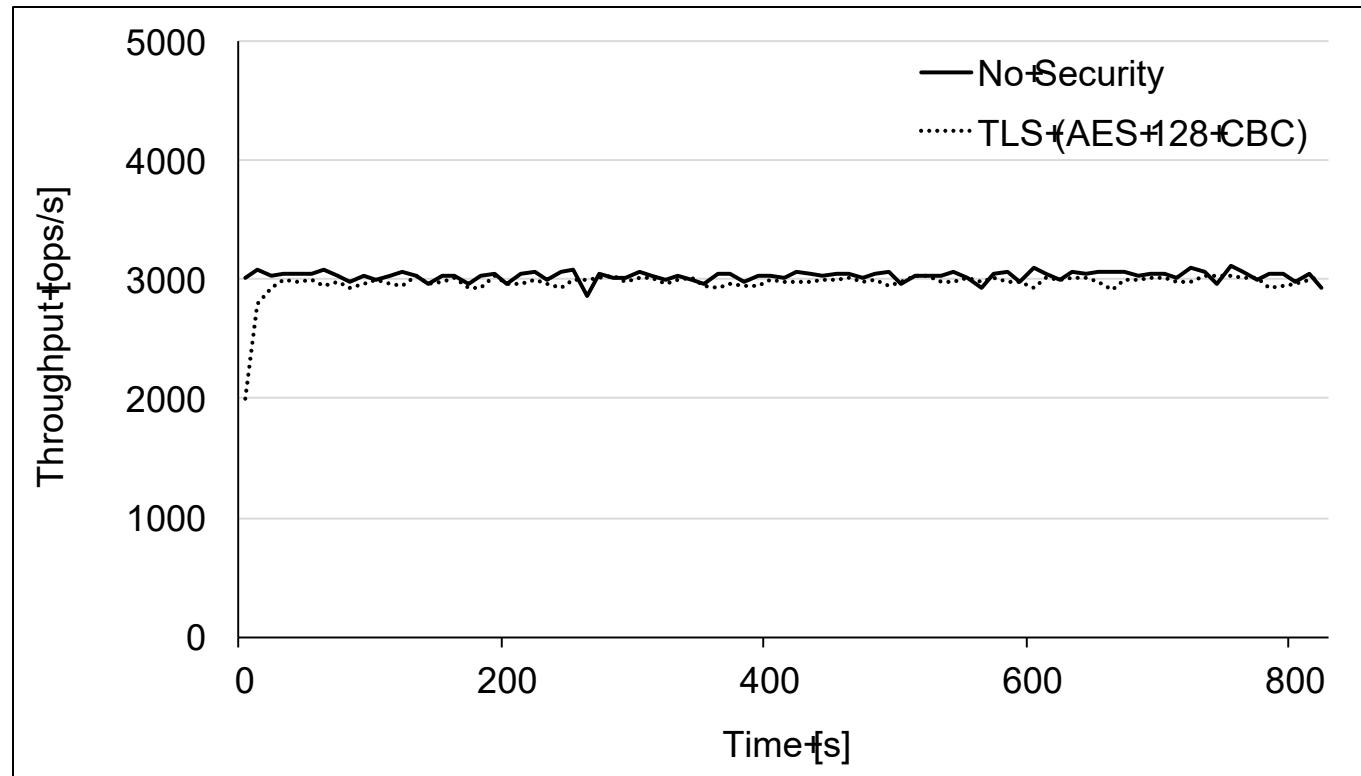


Data in Transit Encryption

- TLS-basiertes HTTPS
- Stark limitierte Teilmenge der TLS Konfigurationen (abhängig von den AWS Regionen)

→ „Konfigurierung“ durch Auswahl der Region

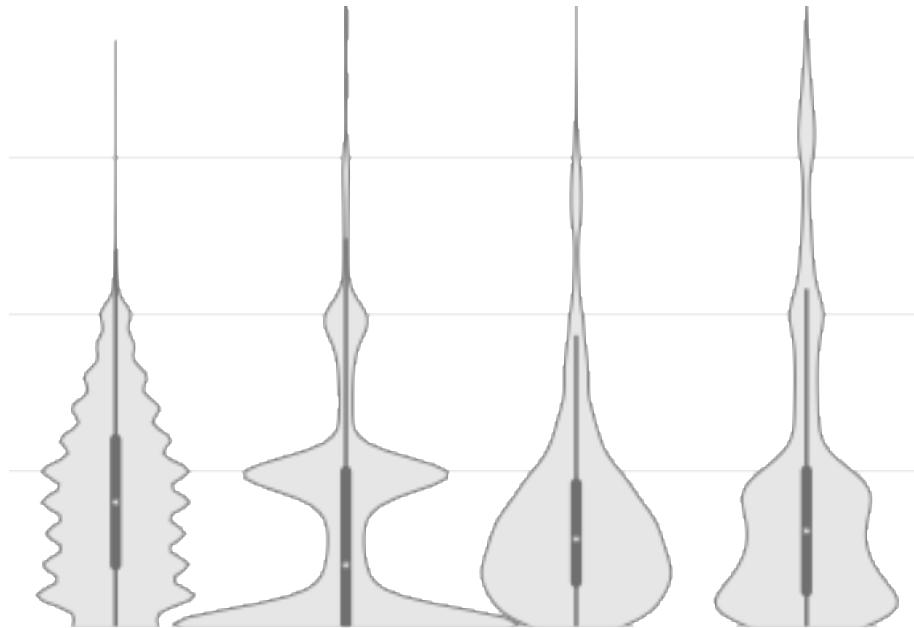
DynamoDB: Ergebnisse des Experiments



DynamoDB: Ergebnisse des Experiments



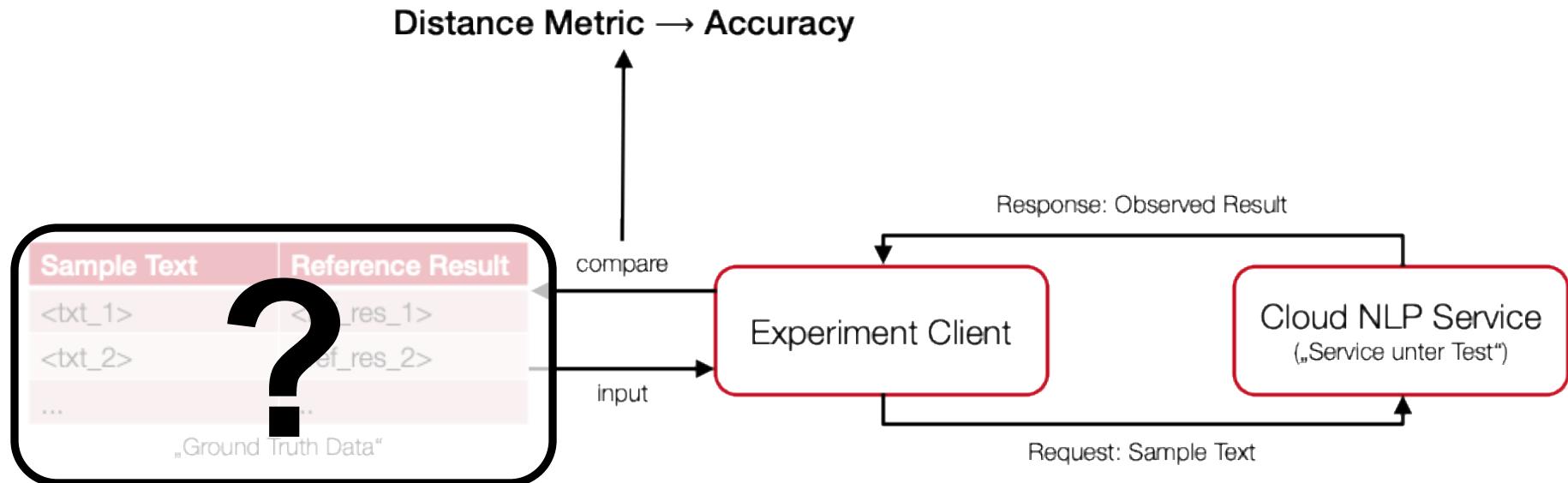
Disclaimer 2: Aber nicht nur



- Klar definieren, was Sie experimentell „herausbekommen“ wollen
 - Für wen ist die Frage relevant und warum?
 - Was ist eine dafür geeignete Experimentumgebung?
 - Welche Parameter wollen Sie ändern / vergleichen und warum?
 - Ist das wirklich die interessanteste (gut) oder nur die am leichtesten umzusetzende (schlecht) Frage?

Wie gut (hinsichtlich erreichter Genauigkeit) sind Cloud-NLP-Services von Google, Amazon (AWS), IBM und Microsoft?
(sentiment analysis, named entity recognition, text classification)

- Klar definieren, was Sie experimentell „herausbekommen“ wollen
 - Für wen ist die Frage relevant und warum?
 - Was ist eine dafür geeignete Experimentumgebung?
- Welche Parameter wollen Sie ändern / vergleichen und warum?
- Ist das wirklich die interessanteste (gut) oder nur die am leichtesten umzusetzende (schlecht) Frage?



Professionell annotierte Datasets

CoNLL-2003, WikiGold, Reuters Corpora

+ hohe Qualität / Verlässlichkeit

– vglw. klein

– häufig veralteter Inhalt

– **hohes Risiko von pre-fitted models!**



→ „Wenn **existierende Tools / Datensätze nicht angemessen / möglich sind, implementieren Sie eigene** und machen Sie diese verfügbar“

Professionell annotierte Datasets

CoNLL-2003, WikiGold, Reuters Corpora

+ hohe Qualität / Verlässlichkeit

– vglw. klein

– häufig veralteter Inhalt

– **hohes Risiko von pre-fitted models!**



„Implizites Crowd-Sourcing“

z.B. Twitter data, Produktreviews, ...

+ solide Qualität / Verlässlichkeit
erreichbar

+ quasi-unendliche Größen möglich

+ aktueller Inhalt

+/- mittl. Risiko von pre-fitted models



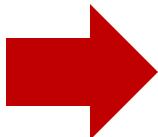
Ground Truth für Sentiment Analysis

★★★★★ 10/9/2020
Overall a great two night stay in the four seasons of Philadelphia. Service was great, rooms are very nice and the infinity edge pool provides views of the entire city. Breakfast at the JG skyhigh was also excellent with great food and large portion sizes. The Covid safety protocols in place were substantial and provided an extra layer of safety for guest and staff. It doesn't quite match up to the Langham in Chicago (best thinking of sp)

★★★ 12/28/2019
So, hubby and I went to the Christmas Market (awesome) and did a little shopping, stopped by Reading Market, etc etc. We then headed over to the new Four Seasons to check out the bar on the 60th floor. Would have loved to look at a room for my mom's upcoming 75 birthday party in a few years but so much for that. We were advised about an hour long wait to go up the elevator for standing room only at the bar. Super glad they are busy. But here's the BIG, HOWEVER. The bar on the ground floor was closed at 4:20 pm on a Saturday!!! What the hell? LOL. Maybe open the empty bar and give CUSTOMER pagers OR A WAIT LIST so PATRONS can sit and drink, spend some money and wait to be called to check out the new bar and lobby!!! Oh yeah and maybe get some apps!! Customers win, hotel makes money and nobody gets turned away!!!! We watched about 30 people walk out the front door.
Might want to think this through Comcast. Too bad especially for people like us that drove from West Chester

Useful 3 Funny 1 Cool 1

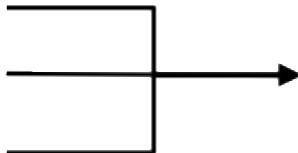
Yelp, Four Seasons Hotel Philadelphia



Sample Text	Reference Result
"Overall a great two night stay..."	5
"So, hubby and I went to the ..."	3
...	...

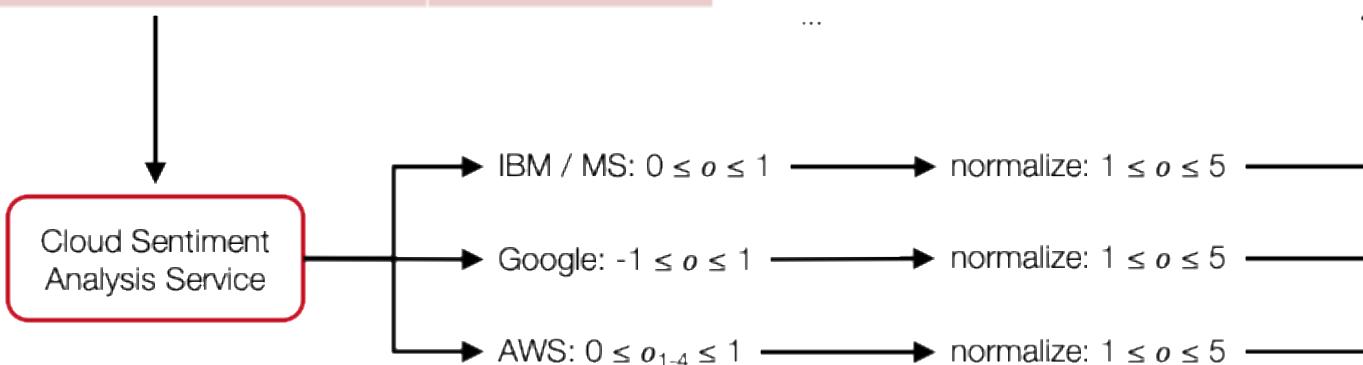
Sentiment Analysis: Experiment-Prozedur

Sample Text	Reference Result
"Overall a great two night stay..."	5
"So, hubby and I went to the ..."	3
...	...

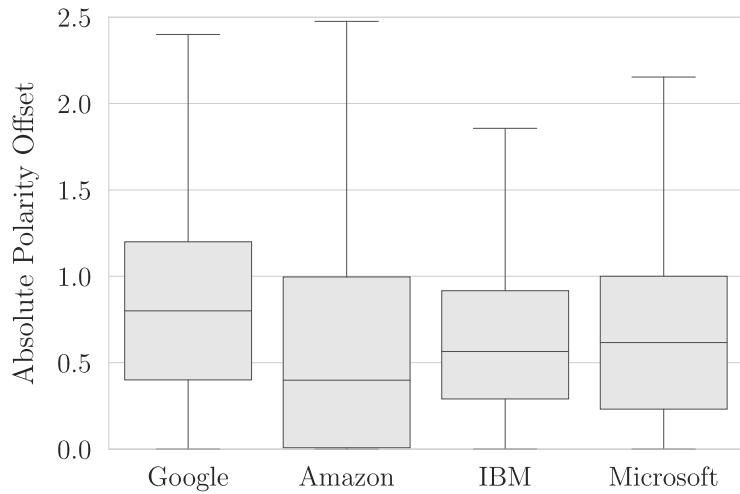


Absolute Polarity Offset (APO)
for each sample

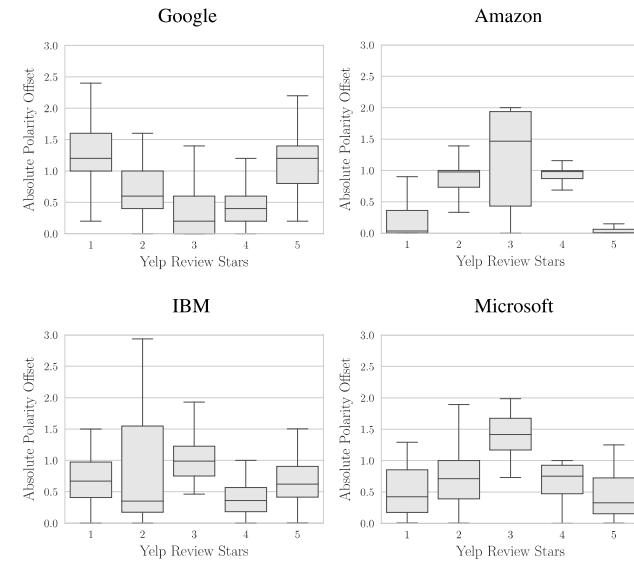
→ Exploratory analysis, also w.r.t. additional factors (star rating, usefulness, ...)



Ergebnisse



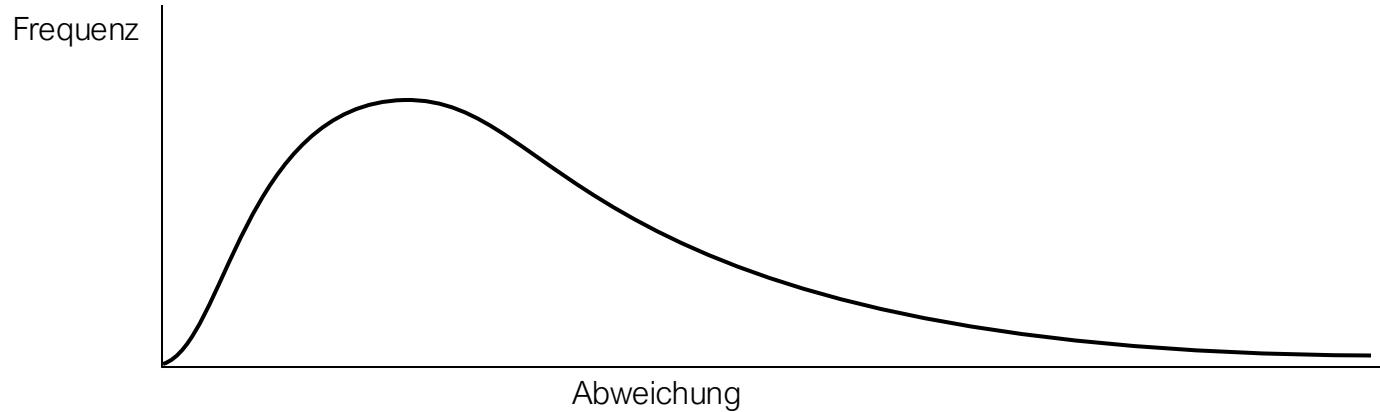
AWS hat geringsten median Offset (besser)
IBM hat konsistenteste Offsets
→ Insgesamt aber wenig signifikante Unterschiede



Detailanalyse nach Stern-Ratings:
→ Google besser für mittlere Ratings
→ AWS (+MS) deutlich besser für extreme Ratings

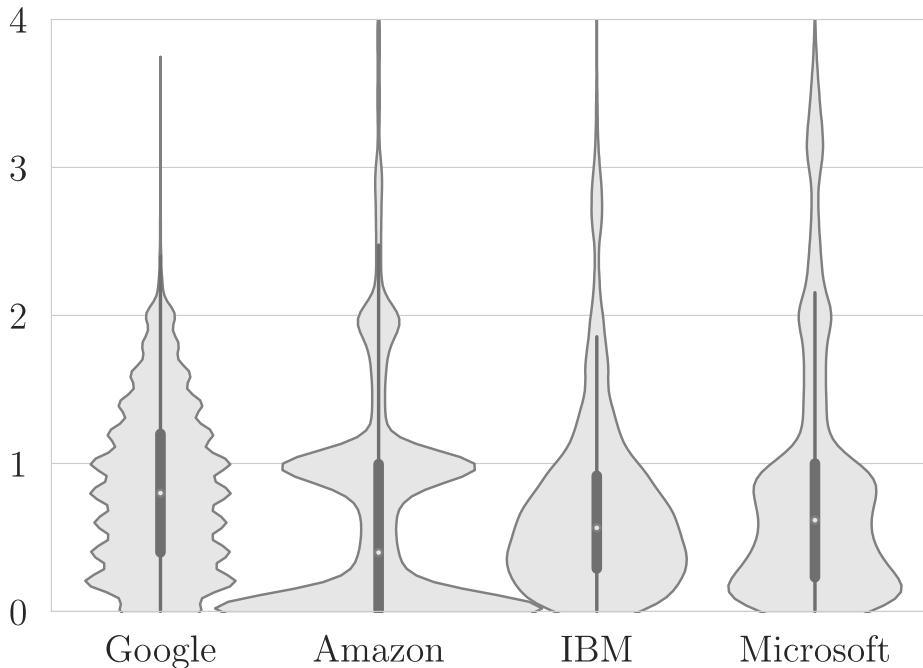
Sentiment Analysis: Abweichungsverteilung

Was zu erwarten gewesen wäre



Was wir tatsächlich beobachteten...

Sentiment Analysis: Abweichungsverteilung



- IBM (+MS) zeigen „nahezu erwartete“ Verteilungen
- Google produziert „schrittweise“ Abweichungen (ca. 0.2)
- AWS Abweichungen häufig sehr nahe an (aber nicht exakt) Ganzzahlen – 3.99, 2.02, ...

→ „Erkennt“ AWS den Text als „Review-Text, der ein ganzzahliges Sentiment haben muss“?

→ Wenn ja: Zufall oder Intention?

Experiments are fun!

Professionell annotierte Datasets

CoNLL-2003, WikiGold, Reuters Corpora

+ hohe Qualität / Verlässlichkeit

– vglw. klein

– häufig veralteter Inhalt

– **hohes Risiko von pre-fitted models!**



„Implizites Crowd-Sourcing“

z.B. Twitter data, Produktreviews, ...

+ solide Qualität / Verlässlichkeit
erreichbar

+ quasi-unendliche Größen möglich

+ aktueller Inhalt

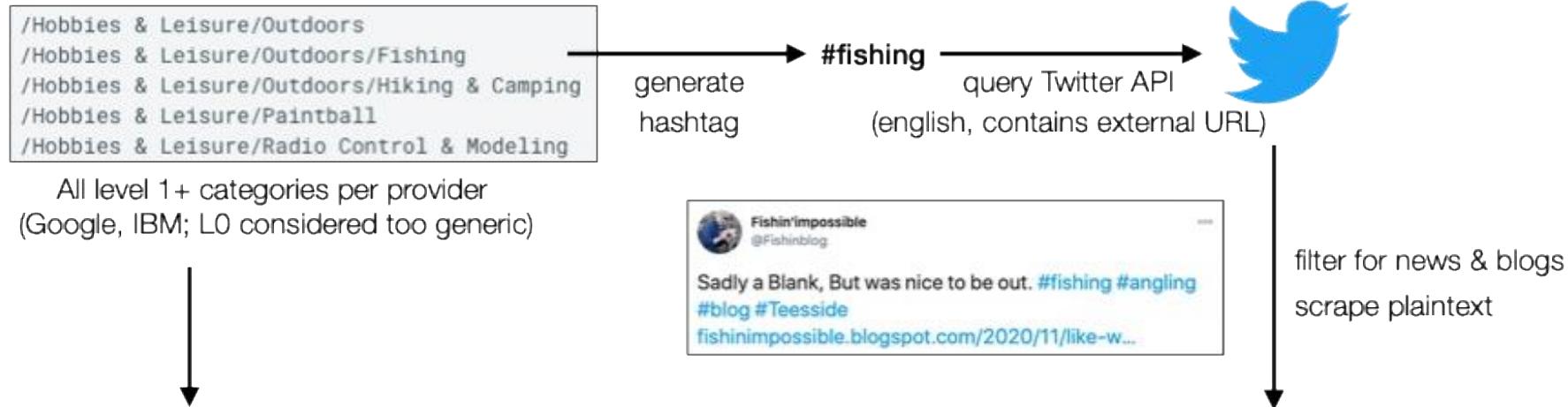
+/- mittl. Risiko von pre-fitted models



- Benötigt: Texte, die sich verlässlich auf ein Thema aus einem von der Classification vorgegebenen Katalog beziehen
- Die Wahrscheinlichkeit, dass der NLP-Service die Texte „schon kennt“, soll gering sein

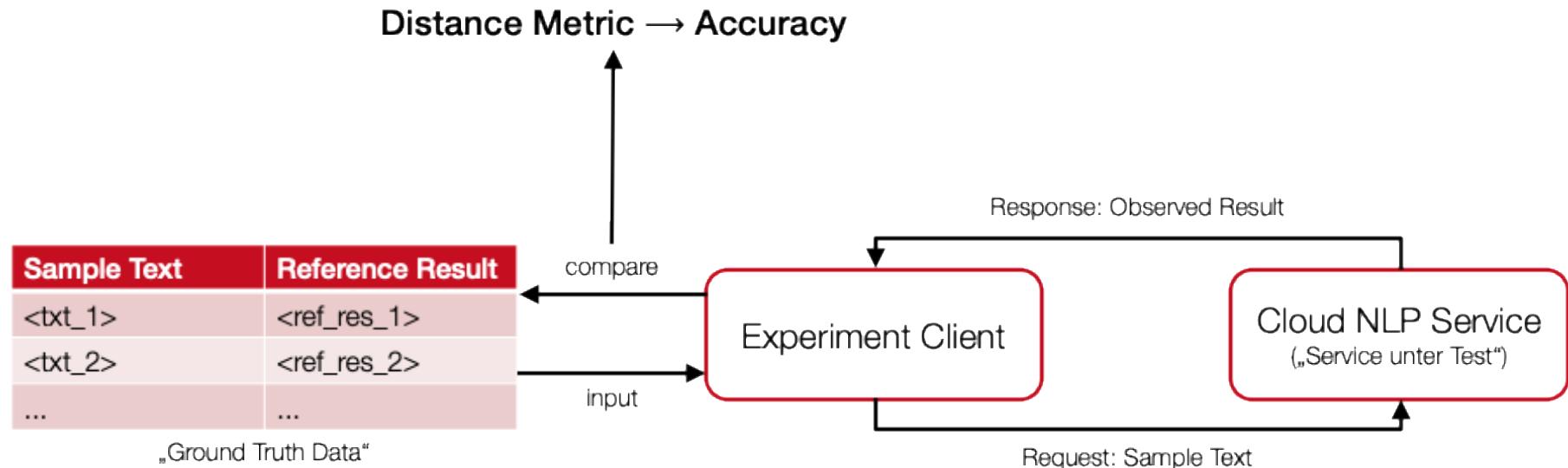
- Klar definieren, was Sie experimentell „herausbekommen“ wollen
 - Für wen ist die Frage relevant und warum?
 - Was ist eine dafür geeignete Experimentumgebung?
 - Welche Parameter wollen Sie ändern / vergleichen und warum?
- Ist das wirklich die interessanteste (gut) oder nur die am leichtesten umzusetzende (schlecht) Frage?

Ground Truth für Text Classification: Indirektion



Reference Result	Sample Text
Hobbies & Leisure/Outdoors/Fishing	
Hobbies & Leisure/Outdoors/Fishing	
...	

→ News / Blogposts, die auf Twitter mindestens einmal zusammen mit einem Referenzkategorie-Hashtag erwähnt wurden



	L-0 Accuracy		L-1+ Accuracy	
	GCP	IBM	GCP	IBM
Overall	0.60	0.59	0.33	0.36
Culture	0.62	0.56	0.40	0.40
Health & Fitness	0.68	0.81	0.39	0.57
Human Activities	0.69	0.65	0.29	0.39
Science	0.59	0.25	0.30	0.17
Society	0.60	0.69	0.29	0.32
Commercial	0.45	0.49	0.29	0.30

Grundsätzlich ähnliche Ergebnisse, aber:

- Google deutlich besser für wissenschaftsbezogene Kategorien
(Profit von Scholar-Aktivitäten?)
- IBM deutlich besser für Gesundheits- und Fitnessthemen
(Profit von Aktivitäten mit Watson Health?)

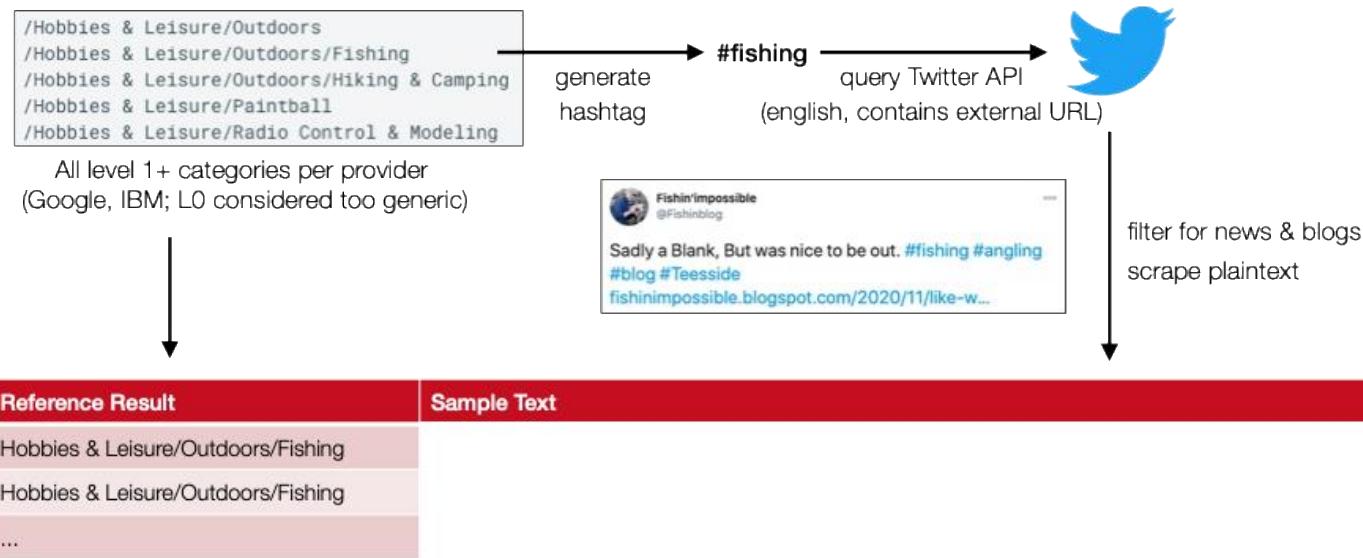
Erkenntnisse, die sich nur durch Experimente „am echten System“
gewinnen lassen

- Klar definieren, was Sie experimentell „herausbekommen“ wollen
 - Für wen ist die Frage relevant und warum?
 - Was ist eine dafür geeignete Experimentumgebung?
 - Welche Parameter wollen Sie ändern / vergleichen und warum?
 - Ist das wirklich die interessanteste (gut) oder nur die am leichtesten umzusetzende (schlecht) Frage?

Fünf Schlüsseleigenschaften eines “guten Benchmarks”:

- **Relevanz:** Lesende der Resultate glauben, dass der Benchmark etwas Wichtiges betrifft.
- **Wiederholbarkeit:** Es ist Wahrscheinlich, dass der Benchmark ein zweites Mal mit den gleichen Ergebnissen durchgeführt werden kann.
- **Fairness:** Alle zu vergleichenden Systeme und/oder Software werden gleichwertig in den Benchmark eingebunden.
- **Verifizierbarkeit:** Man geht davon aus, dass das dokumentierte Ergebnis echt ist.
- **Ökonomische Durchsetzbarkeit:** Die Geldgebenden können es sich leisten, den Benchmark durchzuführen

Früh mit dem Experimentieren beginnen...



... aufhören / pausieren können Sie später immer noch.

fin.