Wieso sammeln wir so viele Daten?

Donald Kossmann
Microsoft Research
Redmond, USA

Wieso „Data Science“?

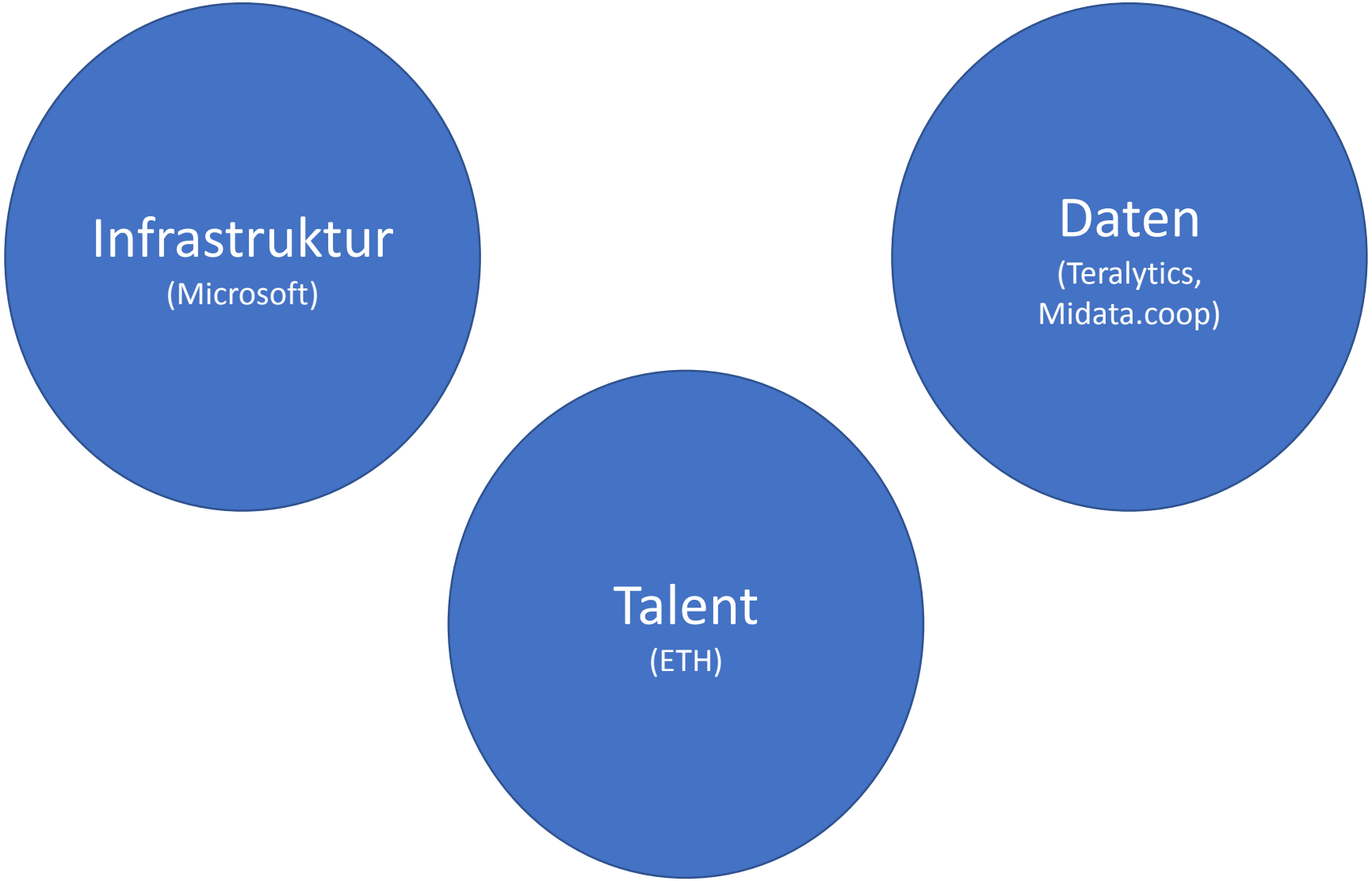
- Wir möchten die brennendsten **Probleme** der Welt **verstehen** und **lösen**
  - Klimawandel, Ernährung, Gesundheit, Kriminalität, Stabilität der Finanzmärkte, ...
- Manchmal ist **Erfahrung** (Daten) besser als ein **Modell** (Mathematik)
  - Die Welt ist zu komplex, um sie in ein Modell zu packen
- **Data Science = Automatisierung von Erfahrung** (= Big Data)
  - “Operationalize Intelligence”

Beispiele

- Übersetze diese Präsentation ins Französische.
- Wie lange braucht man von Zürich nach Aarau Dienstag Nachmittag?
- Wo sollte die Migros den nächsten Supermarkt bauen?
- Sollte ich Donald einen Kredit gewähren?

- ***Wie würden Sie diese Aufgaben angehen?***
  - Daten geben Ihnen Vertrauen in Ihre Antwort
  - Je mehr Daten, desto besser: Abdeckung von Extremfällen

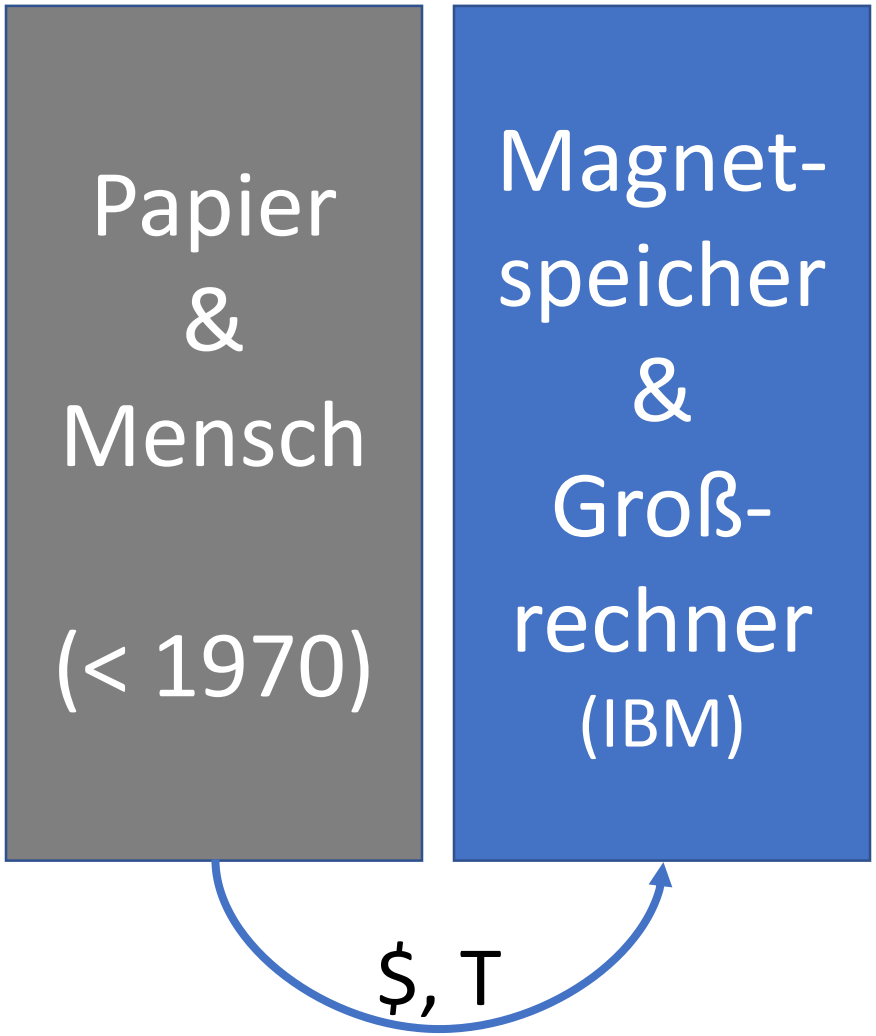
Über mich


Infrastruktur
(Microsoft)
Daten
(Teralytics,
Midata.coop)
Talent
(ETH)

Übersicht: Wieso sammeln wir Daten?

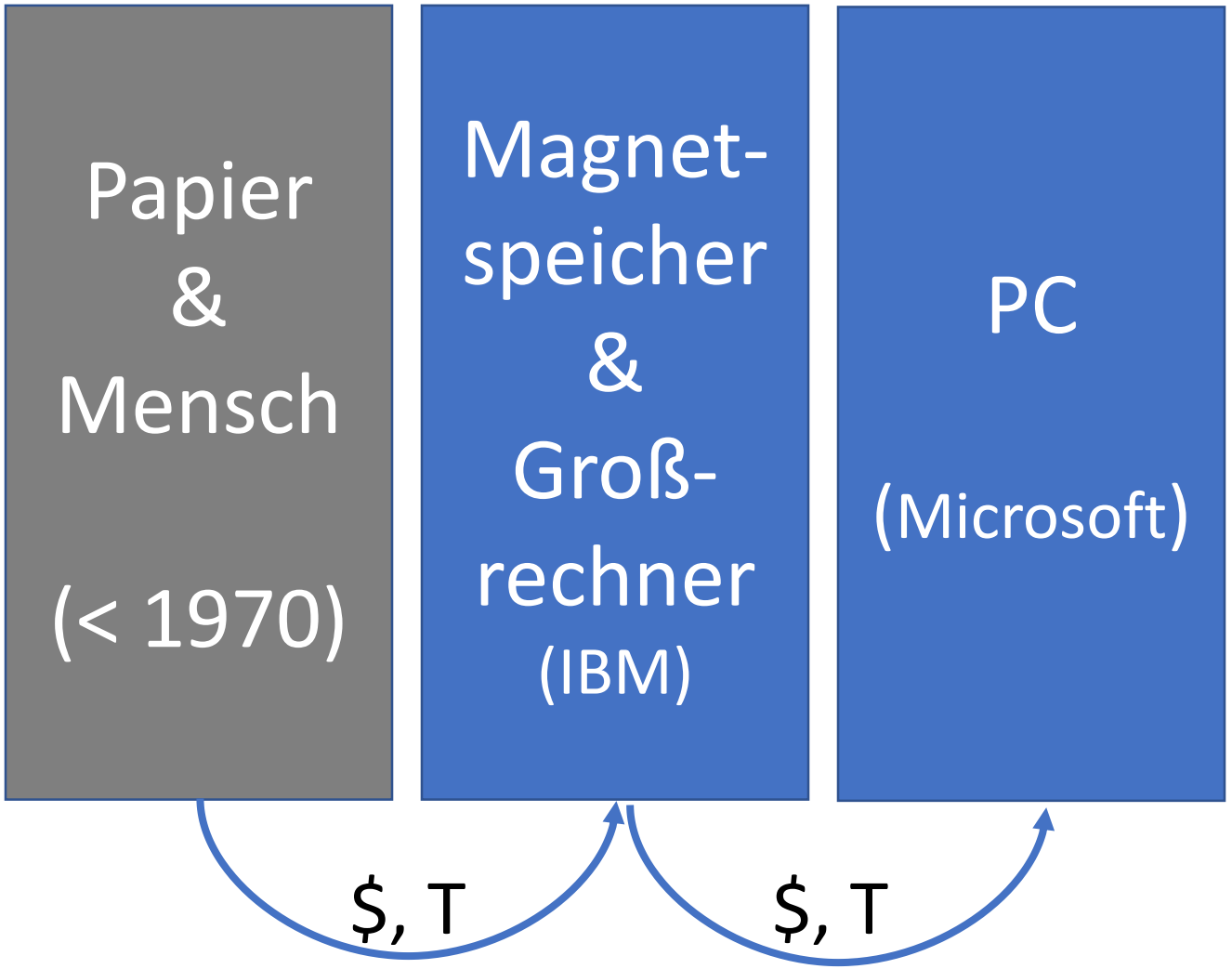
- ... weil wir es können!
  - von der Kuh zum Supermarkt der IT
- ... weil es nützlich ist!
  - vom Supermarkt zur Bank
- Schlussüberlegungen

Geschichte der Digitalisierung (1970)



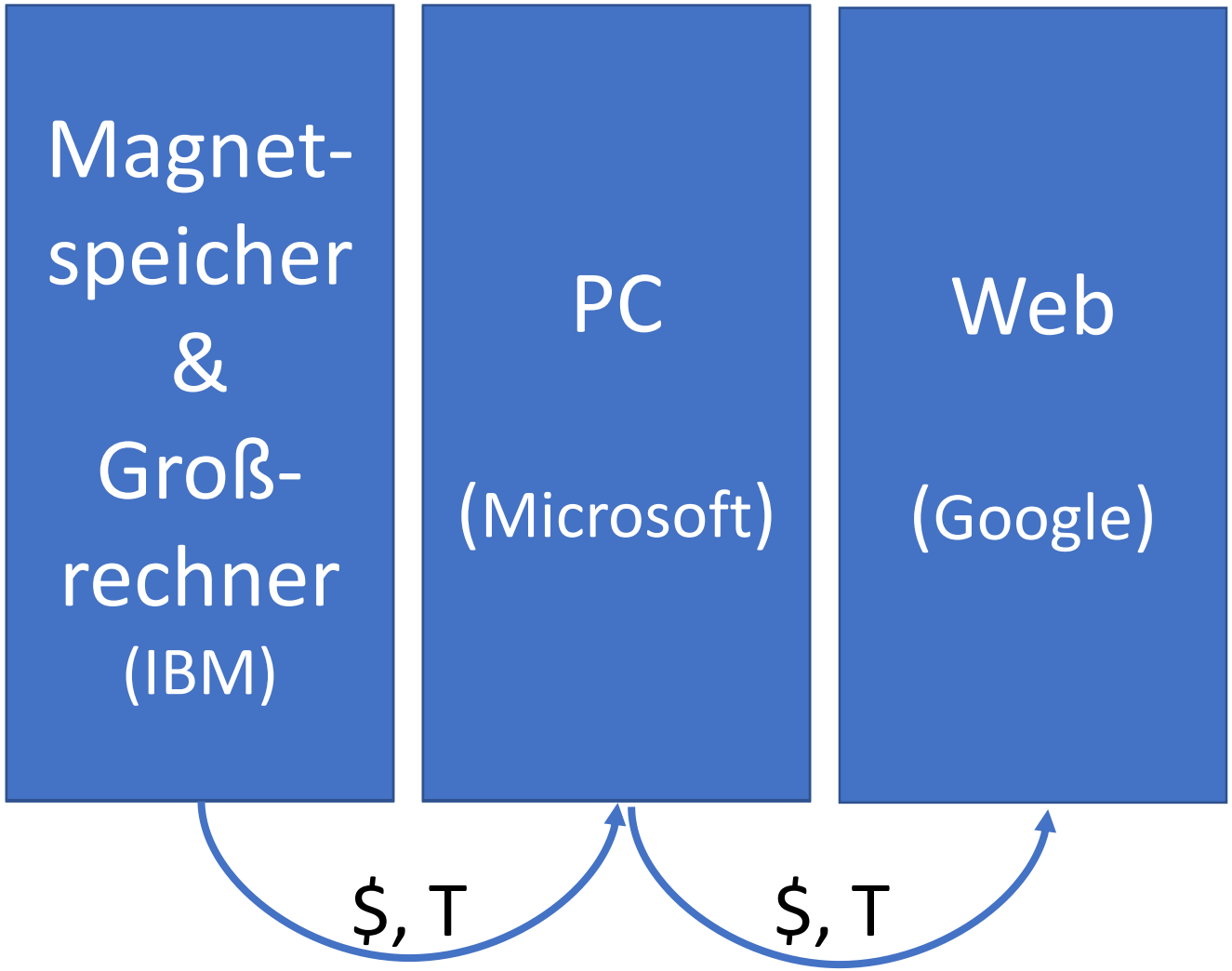
- Kostenersparnisse ($) für Industrie
  - Editierbarkeit der Daten
  - weniger Fehler bei +, -, *, /
  - Datenarchivierung (seit ca. 1995)
- Time to Market (T) für Wissenschaft
  - Beispiel: „Manhattan Projekt“
- Anwendungen
  - Buchhaltung, wissenschaftliches Rechnen

Geschichte der Digitalisierung (1980)



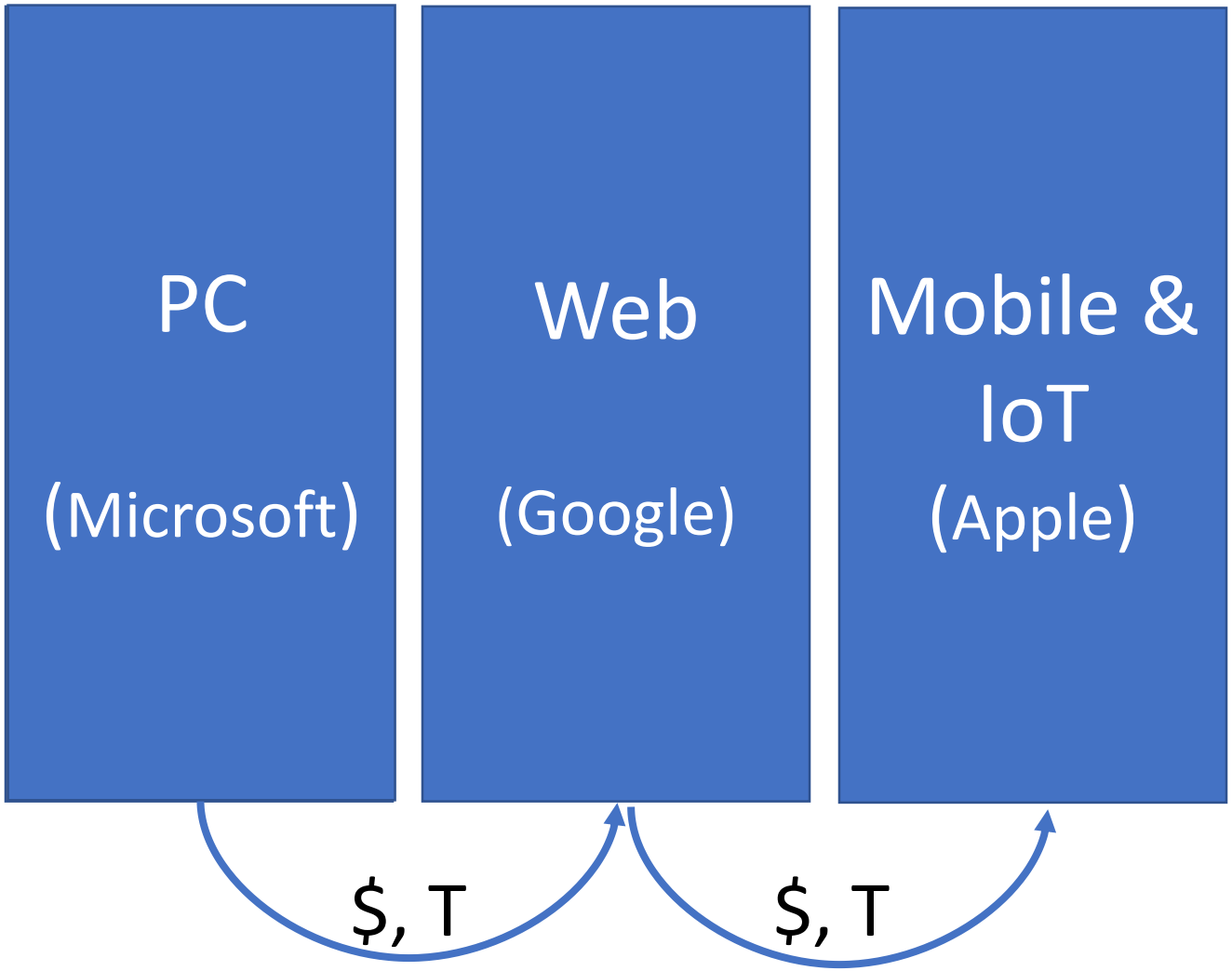
- Demokratisierung der Technologie
  - $, T Vorteile für alle
  - Mission: „PC in jedem Haushalt“
- Demokratisierung des Ecosystems
  - viele Softwareanbieter: Plattform
  - globaler IT Talentpool
- Neue Anwendungen
  - Office, Spiele

Geschichte der Digitalisierung (1990)



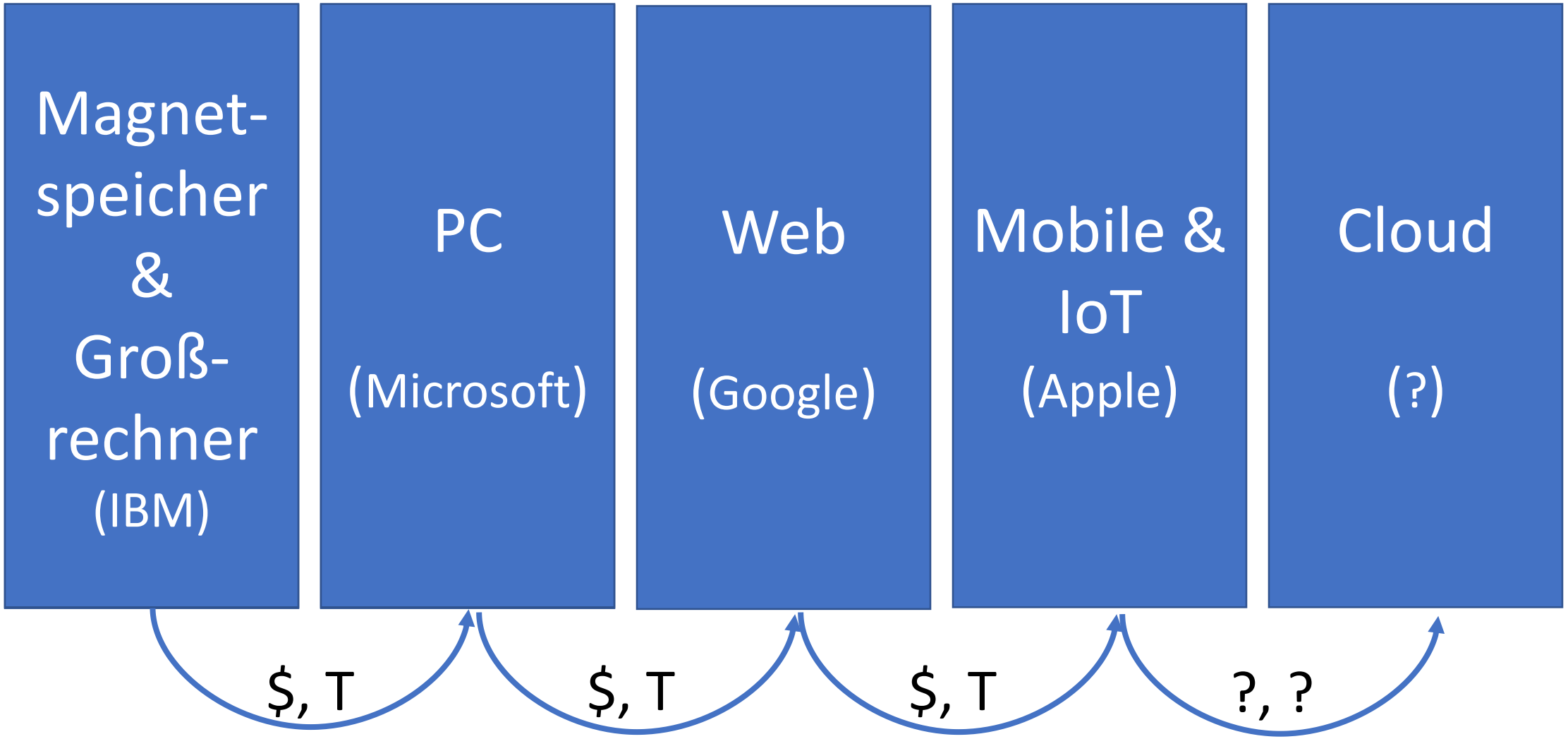
- $ und Time to Market (T)
  - Teilen von Information
  - Ändern von Software (SaaS)
- Neue Anwendungen
  - Beispiel: Soziale Netzwerke (FB)

Geschichte der Digitalisierung (2000)



- Dateneingabe ($)
  - Sensoren im Smartphone
  - Touchscreen
- Time to Market (T)
  - Nutzung immer und überall
- Neuen Anwendungen
  - Navigation, Wellness, …

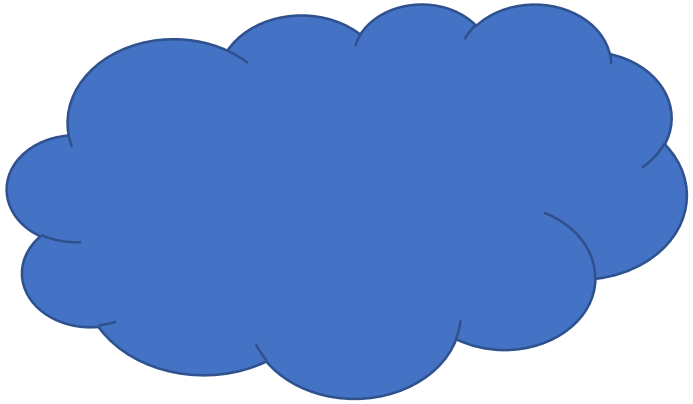
Geschichte der Digitalisierung (1970-2010)



Kuh vs. Supermarkt

Kuh	Supermarkt: Milchflasche
Hohe Investition (ganze Kuh)	Kosten je nach Bedarf
Produziert mehr als man braucht	Man kauft genau so viel, wie man braucht
Ressourcen: Stall, Heu, …	Ressourcen: Kühlschrank
Wartung: Tierarzt, Medikamente, …	Keine Wartung
Abfallprodukte, Behörden, Tierschutz, …	Keine „hidden costs“

Rechner vs. Cloud



Eigener Rechner	Cloud
Hohe Investition (ganze Rechner)	Kosten je nach Bedarf
Produziert mehr als man braucht	Man kauft genau so viel, wie man braucht
Ressourcen: Netzwerk, Strom, …	Ressourcen: Internet
Wartung: Austausch von Festplatten, …	Keine Wartung
Hitze, Compliance / Security, …	Keine „hidden costs“

Übersicht: Wieso sammeln wir Daten?

- ... weil wir es können!
  - von der Kuh zum Supermarkt der IT
- **... weil es nützlich ist!**
  - **vom Supermarkt zur Bank**
- Schlussüberlegungen

Wieso bringe ich mein Geld zur Bank?

Wieso bringe ich mein Geld zur Bank?

- **Bank ist sicher**

- **Bank gibt mir Zinsen**

- Bank hat niedrige Gebühren

Wieso bringe ich meine *Daten* in die *Cloud*?

- **Cloud ist sicher**

- **Cloud gibt mir Zinsen**

- Cloud hat niedrige Gebühren

Data Science: Zinsen in der Cloud

- PC: Armee von hungrigen Informatikern (Talent)
- Web (Forschung): Algorithmen, wie man aus Daten Wert erzielt
- IoT (Smartphone): Automatische, digitale Erfassung von Daten
- Cloud: Zentralisierung aller Daten, Verarbeitungskapazität (Supermarkt)
  - Cloud bringt die Daten zusammen. Die Daten von Einzelnen sind wertlos!

Zürich

Montag, 03:31 Uhr

Teralytics: Eine typische Data Value Chain

Signal

Partner

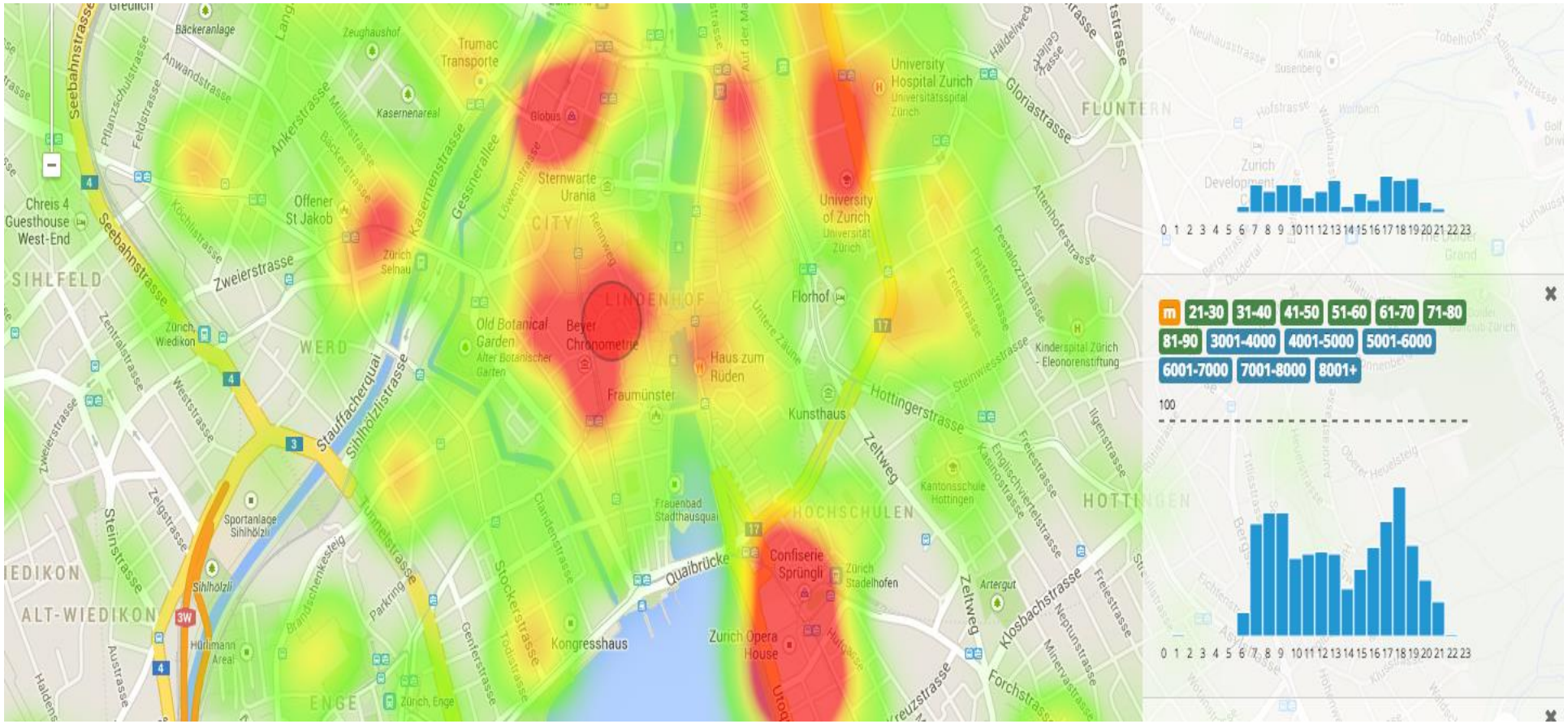
Insight

Teralytics

Action

Customer

Wo soll die neue Bibliothek gebaut werden?



TERALYTICS

TERALYTICS

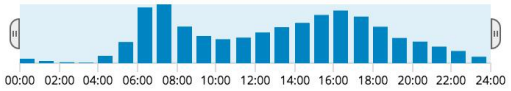
Mode of transport

Private

Number of movements by mode of transport.

Time of day

Number of movements by time of day.

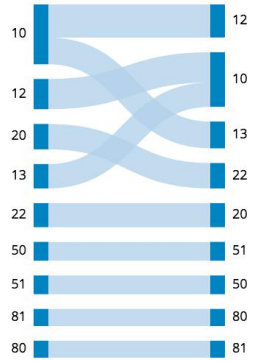


Distance

Movements by distance.

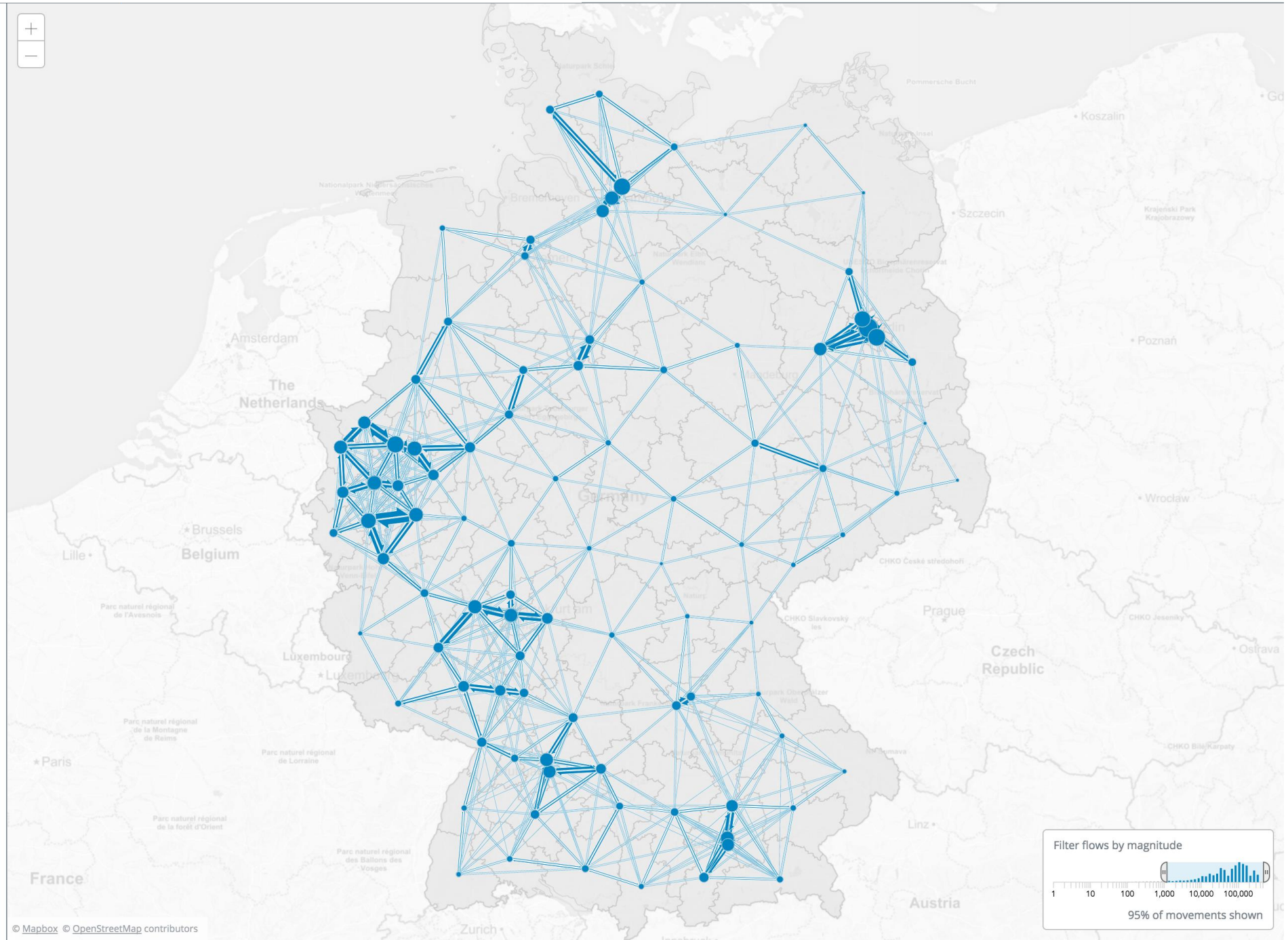
Top Flows

The flows with the most movements.



Top Origins

Top Destinations



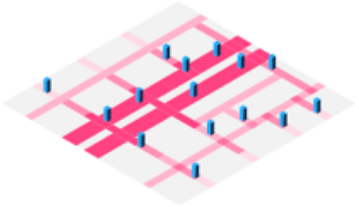
Use Case: Smart City Projekt in Singapur

ST Breaking News / Singapore

LTA seeks more precise picture of travel patterns

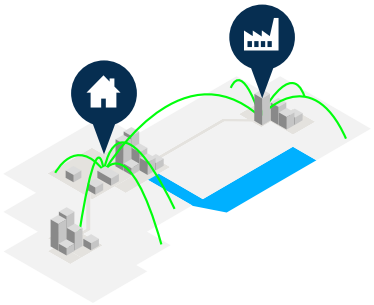
It wants to be able to zoom in on demand in localised regions

Transport planners at the Land Transport Authority (LTA) are looking for a more precise way of analysing and predicting travel patterns in order to optimise road capacity. – ST



Bewegungsströme auf Strassenebene

“Wie viele Autos passieren die River Valley Strasse in Richtung eines Shopping Centers an einem Dienstag während der Mittagszeit?”



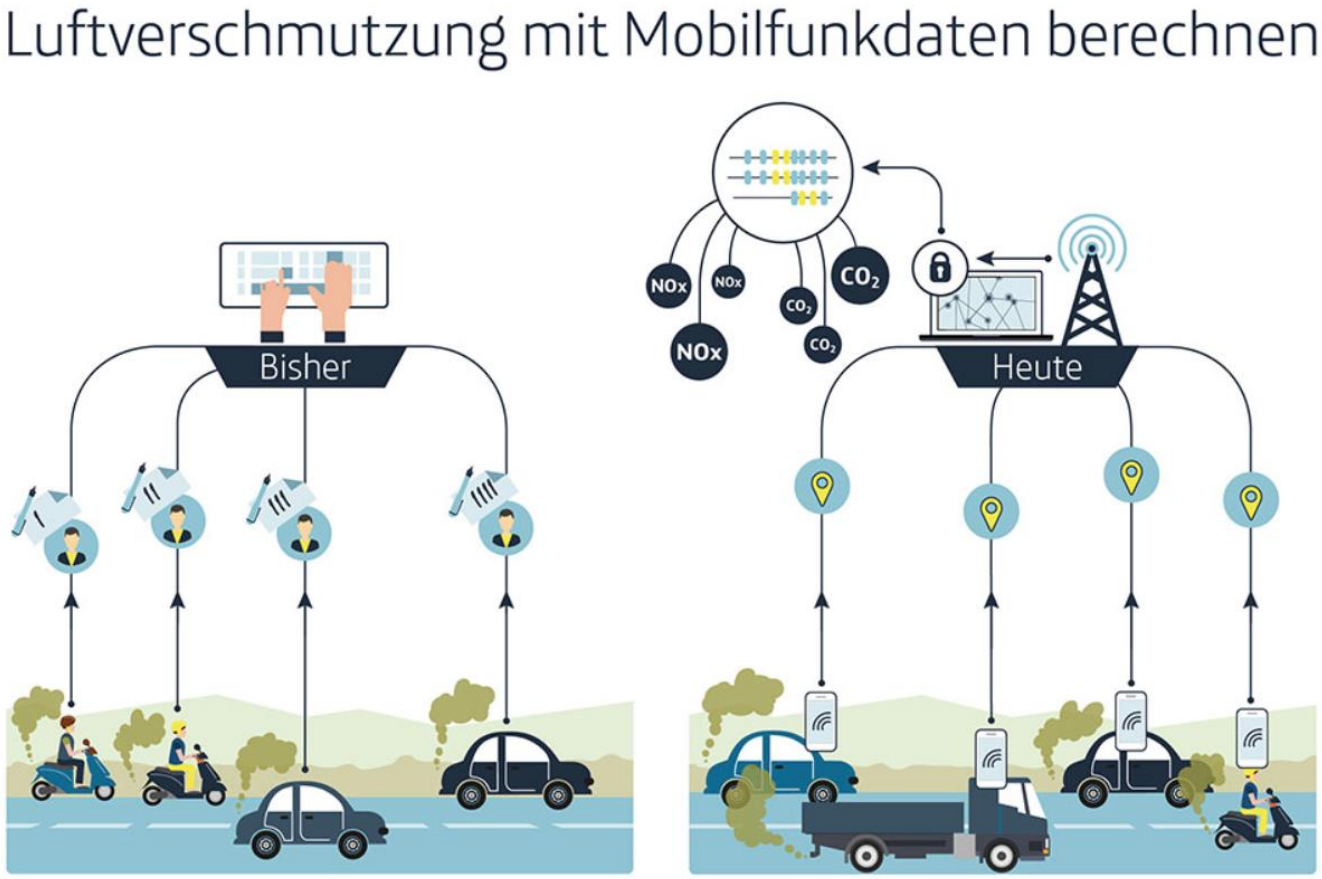
Quell–/ Ziel-Matrix & Transportart

“Wie viele Personen benutzen ihr Auto von Marina Bay zum Flughafen? Wie verhalten sie sich am Ziel- bzw. Herkunftsort?“

Entwicklung der Mobilität in Städten

“Wie hoch ist der Marktanteil des lokalen Nahverkehrsanbieters bei Touristen?”

Data Science für den Klimaschutz



TERALYTICS

NÜRNBERG

south pole group

Telefónica

eit Climate-KIC

Data Science im Gesundheitswesen

- Modell vs. Erfahrung
  - Arzt: “Bitte nehmen Sie dieses Medikament.”
  - Patient: “Wieso?”
  - Doctor: “Weil es funktioniert.”

Data Science im Gesundheitswesen

- Modell vs. Erfahrung
  - Arzt: “Bitte nehmen Sie dieses Medikament.”
  - Patient: “Wieso?”
  - Doctor: “Weil es funktioniert.”

- Precision Medicine
  - individuelle Therapie für Patienten gemäß Anamnese, Diagnose, …
    anhand Erfahrungen mit Patienten mit ähnlicher Anamnese und Diagnose

Data Science: Weitere Beispiele

- Automatische Diashow, Playlists, etc.
  - die 100 schönsten Bilder meiner Griechenlandreise
- Precision Medicine (Midata.coop)
  - Therapien anpassen anhand Erfahrungen mit anderen Patienten
- CRM Lead Prediction
  - Welche Kunden sollte ich für mein neues Produkt anschreiben
- Reiseverhalten
  - Welche Trip Advisor Berichte sind für mich besonders relevant?
- Formel 1
  - Wann sollte Sebastian Vettel seine Pit Stops im Rennen machen?
- …
- Braucht Daten, Algorithmen, Cloud und kreative Leute

Gemeinsamkeiten von Geld und Daten

- Haben einen großen Wert: Je mehr man hat, desto wertvoller
  - Daten des Einzelnen nicht wertvoll, aber Daten von Vielen sehr wertvoll
  - man braucht eine Institution, die bündelt.
- Man kann sie kopieren
  - Primärnutzung vs. Sekundärnutzung; wozu gesammelt vs. wozu verwendet
- Man kann sie verlieren
  - physischer Verlust
  - Verlust der Kontrolle und Transparenz, wozu sie eingesetzt werden
- Regulierung ist notwendig und schwierig
  - komplexe Wertschöpfungsketten; schnelle Innovation.
  - Regulator weiß nicht, was er anrichtet.

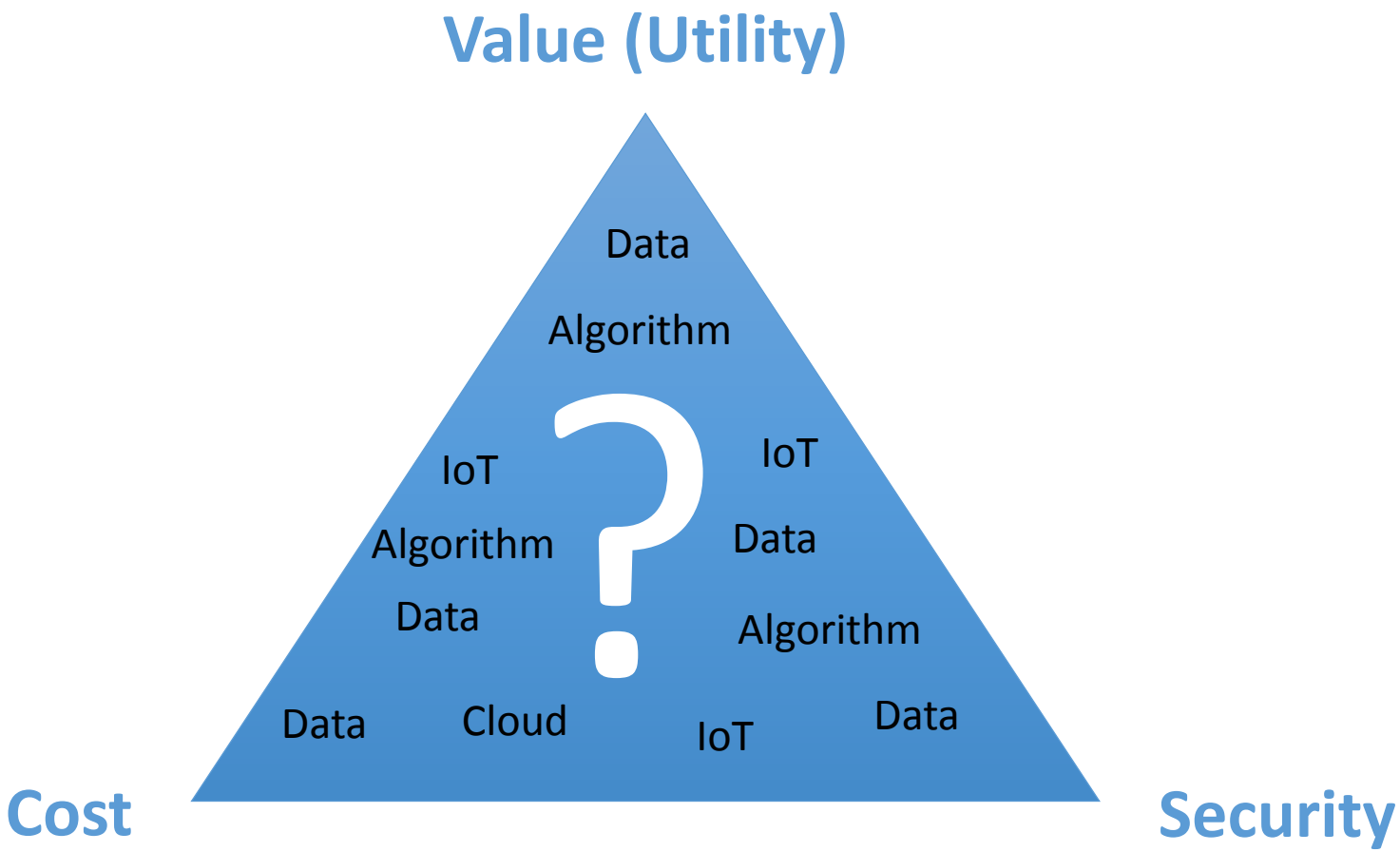
Übersicht: Wieso sammeln wir Daten?

- ... weil wir es können!
  - von der Kuh zum Supermarkt der IT
- ... weil es nützlich ist!
  - vom Supermarkt zur Bank
- **Schlussüberlegungen**

Geschichte der Wissenschaften

- 1960 (Wigner): Unreasonable Effectiveness of *Mathematics*
  - Beispiele: Newton, Einstein, ...
- 2009 (Norvig): Unreasonable Effectiveness of *Data*
  - Beispiel: Google Suche
- Heute: Unreasonable Effectiveness of *Data Science*
  - Beispiele: Formel 1, Medizin, Trip Advisor, Skype Translate, CRM, ...
  - Bausteine: Infrastruktur, Daten (Sensoren), Algorithmen, Talent (Kreativität)
  - Außer Talent alle Bausteine in Massen vorhanden

Die Herausforderung: Integration



Demokratisierung von Data Science

- Data Science ist zu mächtig, um sie in der Hand von Wenigen zu lassen
  - War die US Präsidentenwahl manipuliert?
  - Dominanz von Amazon, Facebook, Google? (Alibaba, Tencent?)

- Woran wir arbeiten müssen
  - Digitale Infrastruktur
  - Talent
  - Daten
  - *und vor allem an der Integration all dieser Komponenten*

  - ***Die Schweiz ist überall gut. Aber nirgendwo top!***