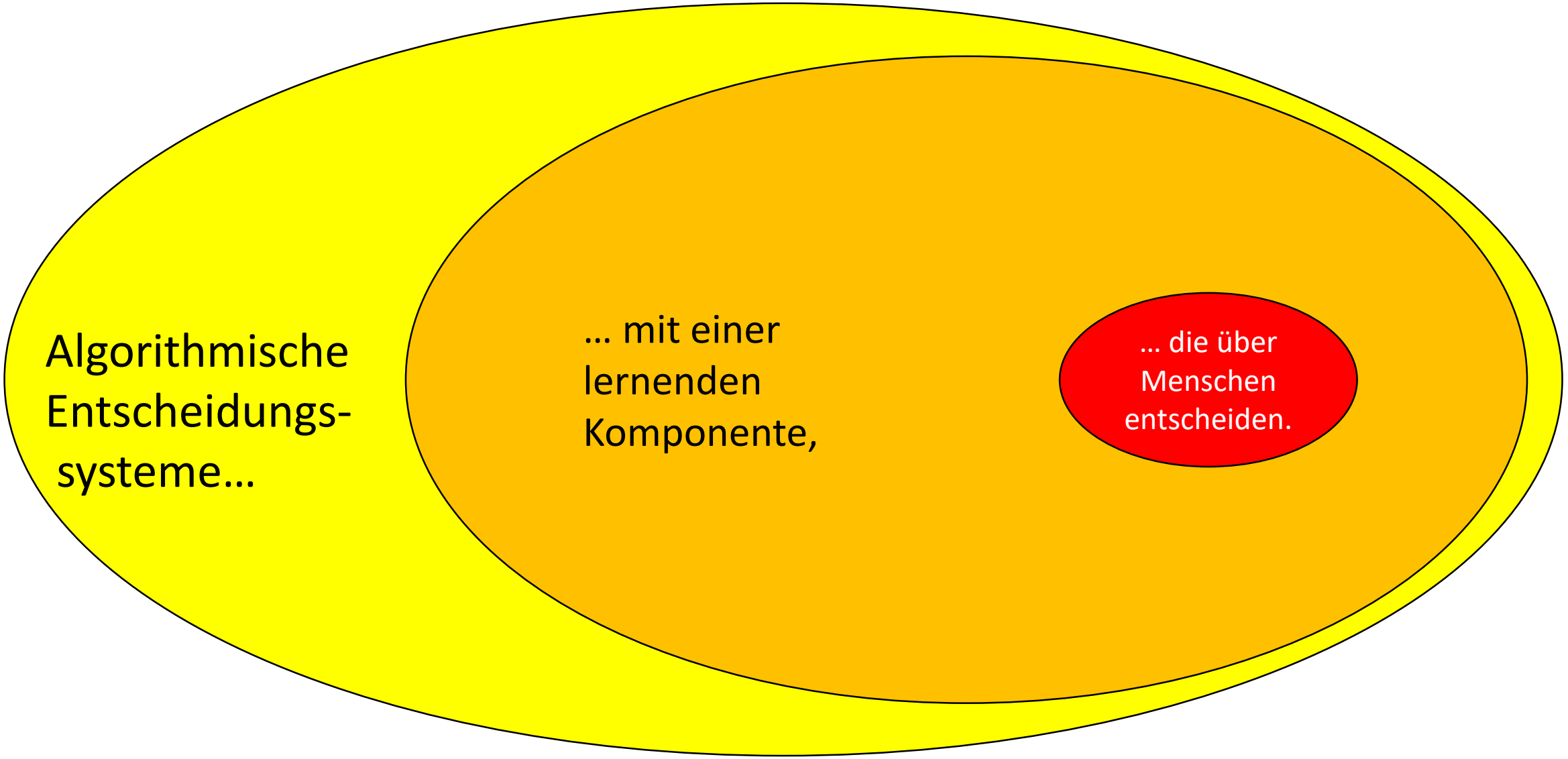
TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Black-Box-Analyse-Methoden

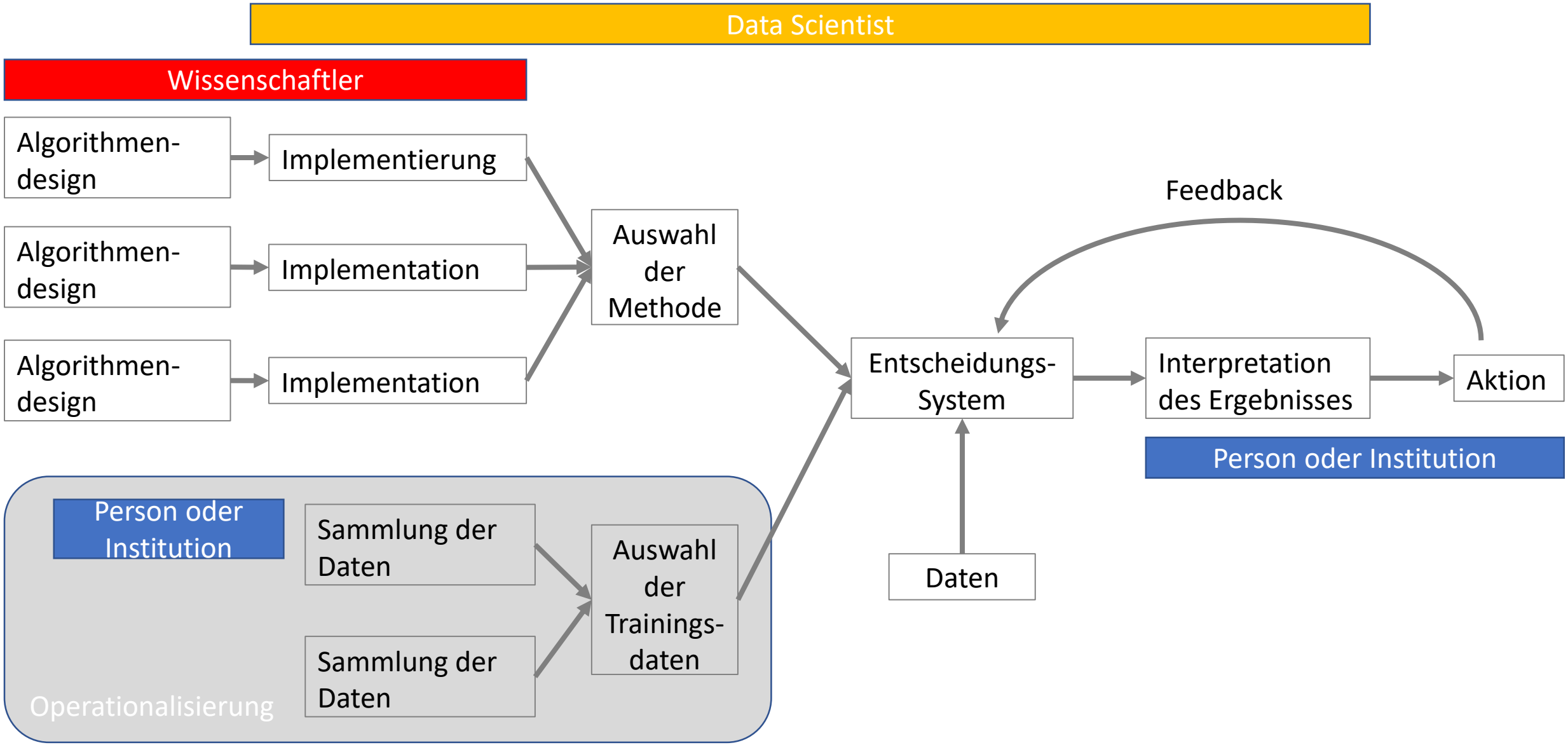
Prof. Dr. Katharina A. Zweig

TU Kaiserslautern

„Algoskop“


Algorithmische Entscheidungs-systeme…
… mit einer lernenden Komponente,
… die über Menschen entscheiden.

Lange Kette der Verantwortlichkeiten

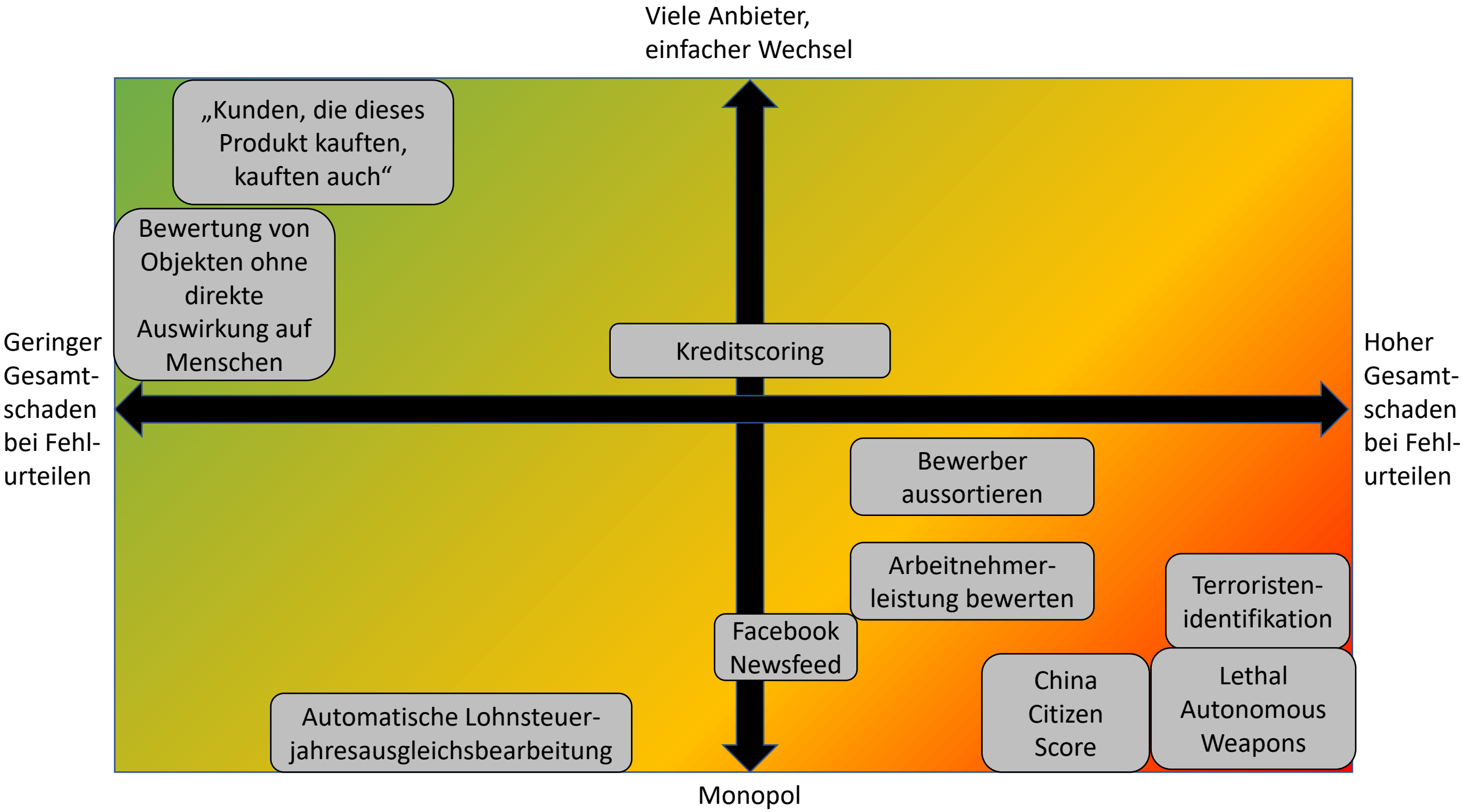


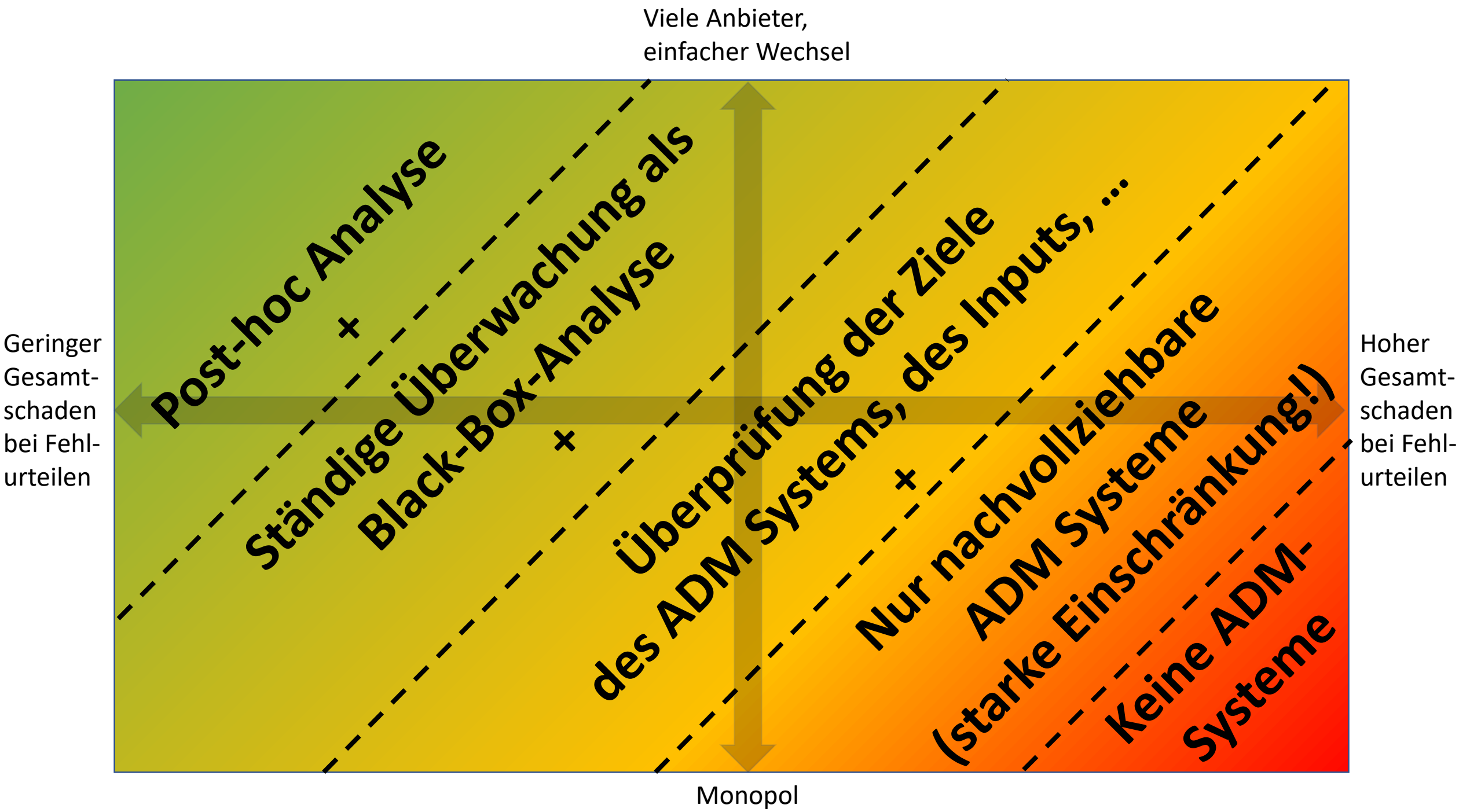
Einordnung auf Risikomatrix

1. Schadenstiefe

$$\sum \begin{array}{c} \text{Schaden für Individuum(Fehlurteil)} \\ +\text{Schaden für Gesellschaft(Fehlurteil)} \end{array}$$

2. Anbietervielzahl, Wechselmöglichkeiten, Möglichkeiten der Anfechtbarkeit, Revisionen durch Menschen, etc.


Viele Anbieter,
einfacher Wechsel
„Kunden, die dieses Produkt kauften, kauften auch“
Bewertung von Objekten ohne direkte Auswirkung auf Menschen
Kreditscoring
Geringer Gesamt-schaden bei Fehl-urteilen
Hoher Gesamt-schaden bei Fehl-urteilen
Bewerber aussortieren
Arbeitnehmer-leistung bewerten
Terroristen-identifikation
Facebook Newsfeed
China Citizen Score
Lethal Autonomous Weapons
Automatische Lohnsteuer-jahresausgleichsbearbeitung
Monopol


Viele Anbieter,
einfacher Wechsel
Geringer Gesamt-schaden bei Fehl-urteilen
Hoher Gesamt-schaden bei Fehl-urteilen
Monopol
Post-hoc Analyse
+
Ständige Überwachung als Black-Box-Analyse
+
Überprüfung der Ziele des ADM Systems, des Inputs, …
+
Nur nachvollziehbare ADM Systeme (starke Einschränkung!)
Keine ADM-Systeme

veröffentlicht

Viele Anbieter,
…er Wechsel

Geringer Gesamt-schaden bei Fehl-urteilen

Hoher Gesamt-schaden bei Fehl-urteilen

Bis 2001

Google Suchmaschine

2017:
Datenspende #btw17

Monopol

Scott's Search

Egypt

About 360,000,000 results (0.24 seconds) Advi

Crisis in Egypt
Voices in **Egypt** have been muted but will not be silenced. Listen.
humanrightsfirst.org/Egypt

Egypt - Wikipedia, the free encyclopedia
Egypt officially the Arab Republic of **Egypt**, is a country mainly in North Africa, with the Sinai Peninsula forming a land bridge in Southwest Asia. ...
Hosni Mubarak - Ancient Egypt - Female genital cutting - History of modern Egypt
en.wikipedia.org/wiki/**Egypt** - Cached - Similar

Egypt News - The Protests of 2011 - The New York Times
World news about **Egypt** and the protests of 2011. Breaking news and archival information about its people, politics and economy from The New York Times.
topics.nytimes.com › World › Countries and Territories - Cached - Similar

Egypt Travel, Tours, Vacations, Ancient Egypt from Tour Egypt
Information for travelers, resources on history, monuments and activities.
www.tour**egypt**.net/ - Cached - Similar

News for Egypt


Globe and Mail

Why Lara Logan Was Eager to Return to **Egypt**
1 hour ago
By Charlotte Triggs AP Lara Logan had already had one troubling experience in **Egypt** before last Friday's "brutal and sustained" sexual assault, ...
People Magazine - 1658 related articles - Shared by 20+

In **Egypt**, renewed hope for gender equality
USA Today - 24874 related articles - Shared by

Realtime updates for **Egypt** (390)

Egypt Daily News, Egypt News

Daniel's Search

Egypt

About 321,000,000 results (0.15 seconds) Advi

Egypt - Wikipedia, the free encyclopedia
Egypt officially the Arab Republic of **Egypt**, is a country mainly in North Africa, with the Sinai Peninsula forming a land bridge in Southwest Asia. ...
Hosni Mubarak - Ancient Egypt - Female genital cutting - History of modern Egypt
en.wikipedia.org/wiki/**Egypt** - Cached - Similar

Egypt Travel, Tours, Vacations, Ancient Egypt from Tour Egypt
Information for travelers, resources on history, monuments and activities.
www.tour**egypt**.net/ - Cached - Similar

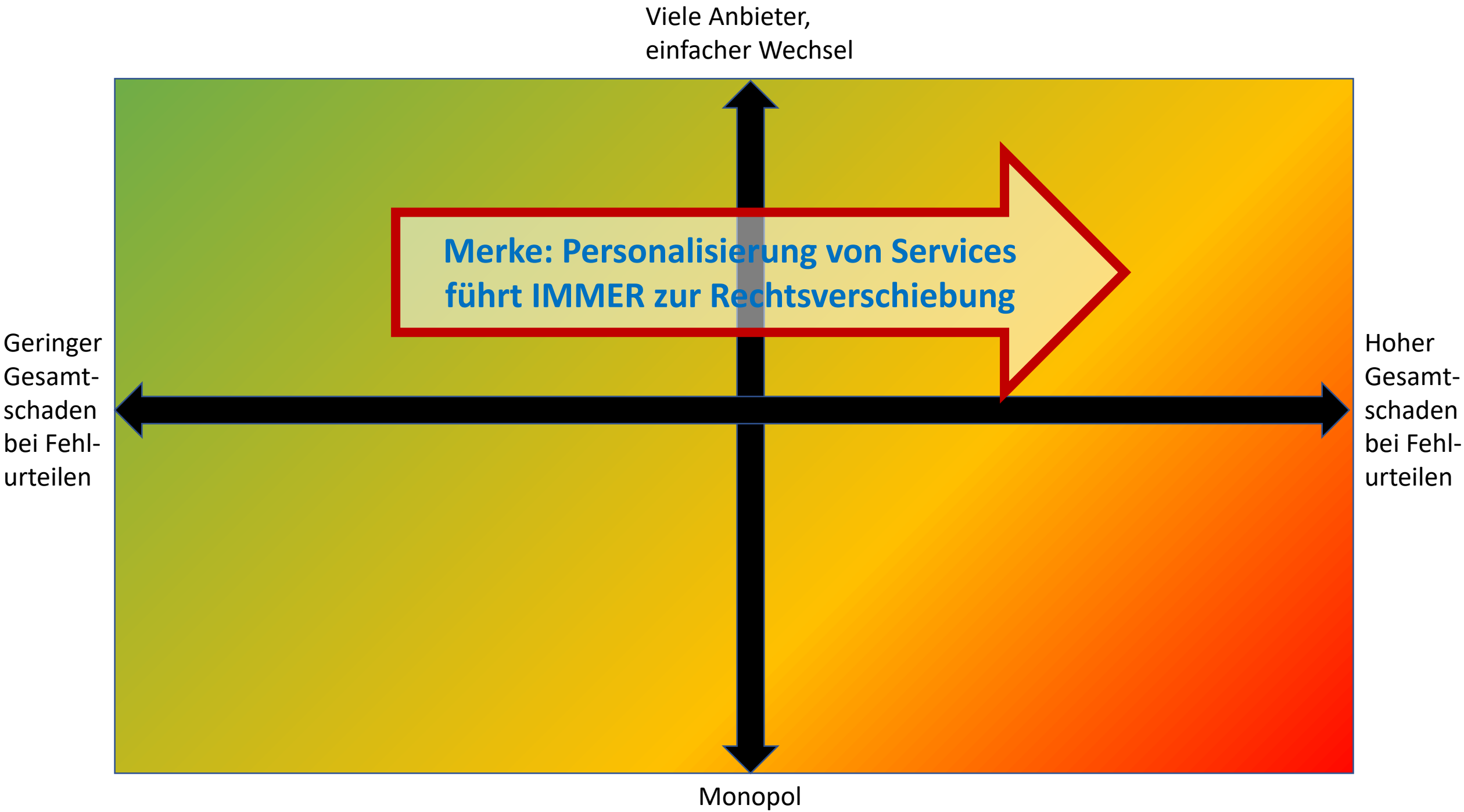
Egypt Daily News, Egypt News
Egypt Daily News, covering **Egypt** News, Arab news, Middle East news and World news. Egyptian Gu..., egyptian recipes, egyptian food, egyptian airforce, ...
www.egy...dailynews.com/ - Cached - Similar

Images for egypt - Report images

Egypt - CIA - The World Factbook
Feb 1, 2011 ... Features a map and brief descriptions of geography, economy, government, and people.
https://www.cia.gov/library/publications/the-world.../eg.html - Cached - Similar

Egypt State Information Service
Official government site provides information on the country's government, politics, culture, history, economy and tourism. [Arabic, English, French]

1 https://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles


Viele Anbieter,
einfacher Wechsel
Merke: Personalisierung von Services
führt IMMER zur Rechtsverschiebung
Geringer
Gesamt-
schaden
bei Fehl-
urteilen
Hoher
Gesamt-
schaden
bei Fehl-
urteilen
Monopol

Datenspende: BTW17

+

+

Landesmedienanstalten der Länder:

Bayern (BLM)
Berlin und Brandenburg (mabb)
Hessen (LPR Hessen)
Rheinland-Pfalz (LMK)
Saarland (LMS)
Sachsen (SLM)

Medienpartner war Spiegel Online.

https://datenspende.algorithmwatch.org/

Datenspende: BTW17

Browserplugin

Zu festen Suchzeitpunkten

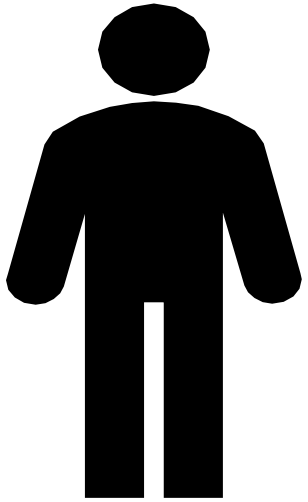
- (4, 8, **12, 16, 20**, 24 Uhr)

Feste Suchbegriffe:

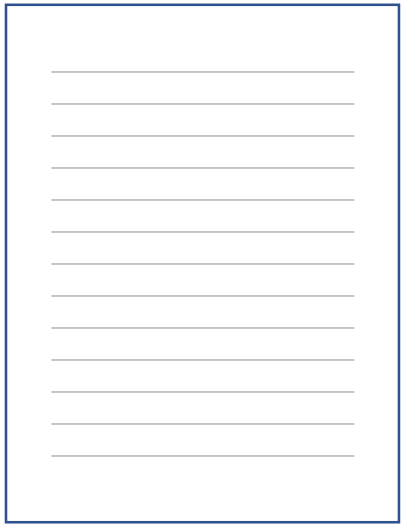
Personen
Alexander Gauland
Alice Weidel
Angela Merkel
Cem Özdemir
Christian Linder
Dietmar Bartsch
Katrin Göring-Eckhardt
Martin Schulz
Sahra Wagenknecht

Parteien
AfD
CDU
CSU
Bündnis 90/Die Grünen
Die Linke
FDP
SPD

Datenspende: BTW17



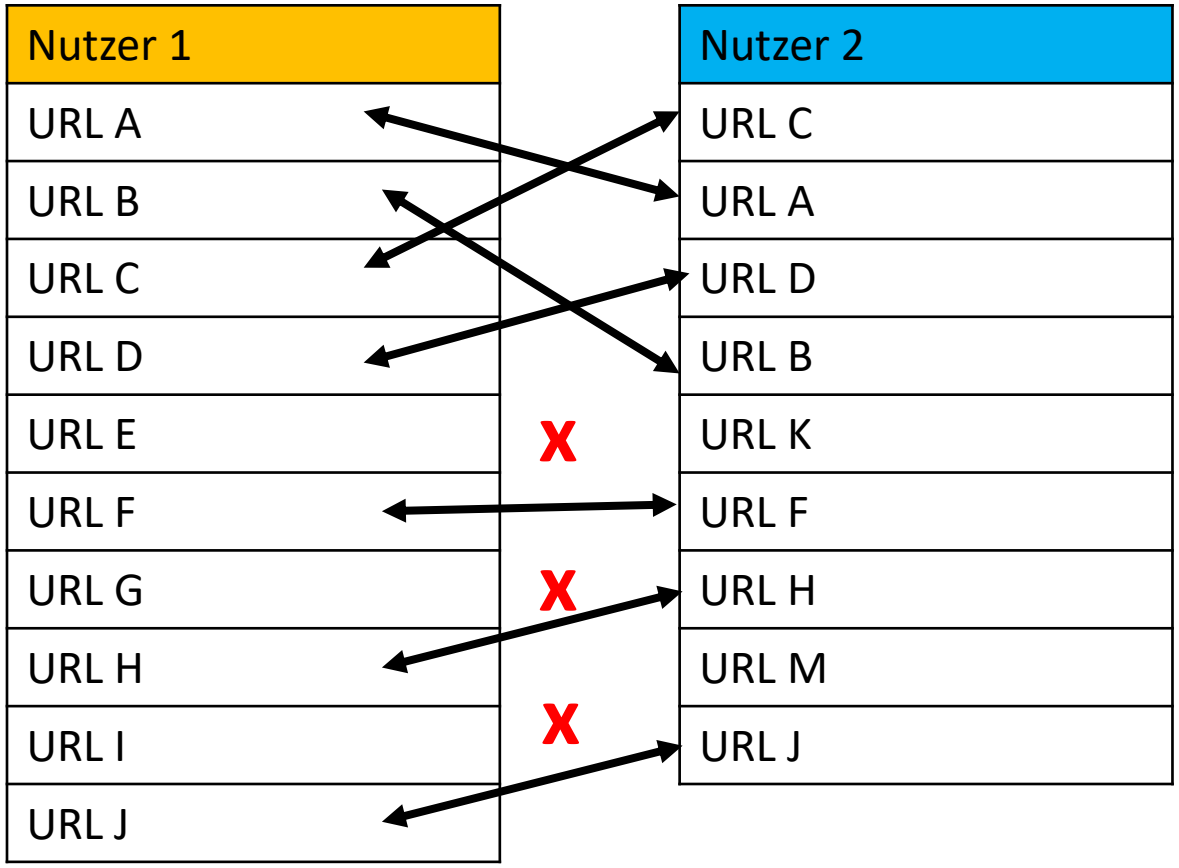
4384 TeilnehmerInnen



5.991.500 (!) gespendete Ergebnislisten

Messung der Personalisierung

- Für alle Paare von Nutzern:
  - Bestimme Anzahl nicht-geteilter Links
  - Im Beispiel:
    - Nutzer 1 teilt drei URLs nicht mit Nutzer 2
    - Nutzer 2 teilt zwei URLs nicht mit Nutzer 1



Busted Filterbubble

- Die Grundlage für eine Personalisierung ist weit kleiner als gedacht.
- Bei den Politikern waren im Durchschnitt für je zwei Nutzer **nur 1-2 Links** **nicht** geteilt von 9-10 Ergebnissen.
- Auf news.google.com sind es 3-4 Links auf 20 Ergebnisse.

	Anzahl nicht geteilter Links
Katrin Göring-Eckardt	**0.9**
Dietmar Bartsch	**1.0**
Angela Merkel	**1.0**
Sahra Wagenknecht	**1.1**
Cem Özdemir	**1.1**
Alexander Gauland	**1.2**
Alice Weidel	**1.4**
Christian Lindner	**1.7**
Martin Schulz	**1.8**

Busted Filterbubble

- Für **Parteien** gibt es **weniger Überlappung.**
- Webseiten der **Ortsverbände**
- Eher **Regionalisierung**

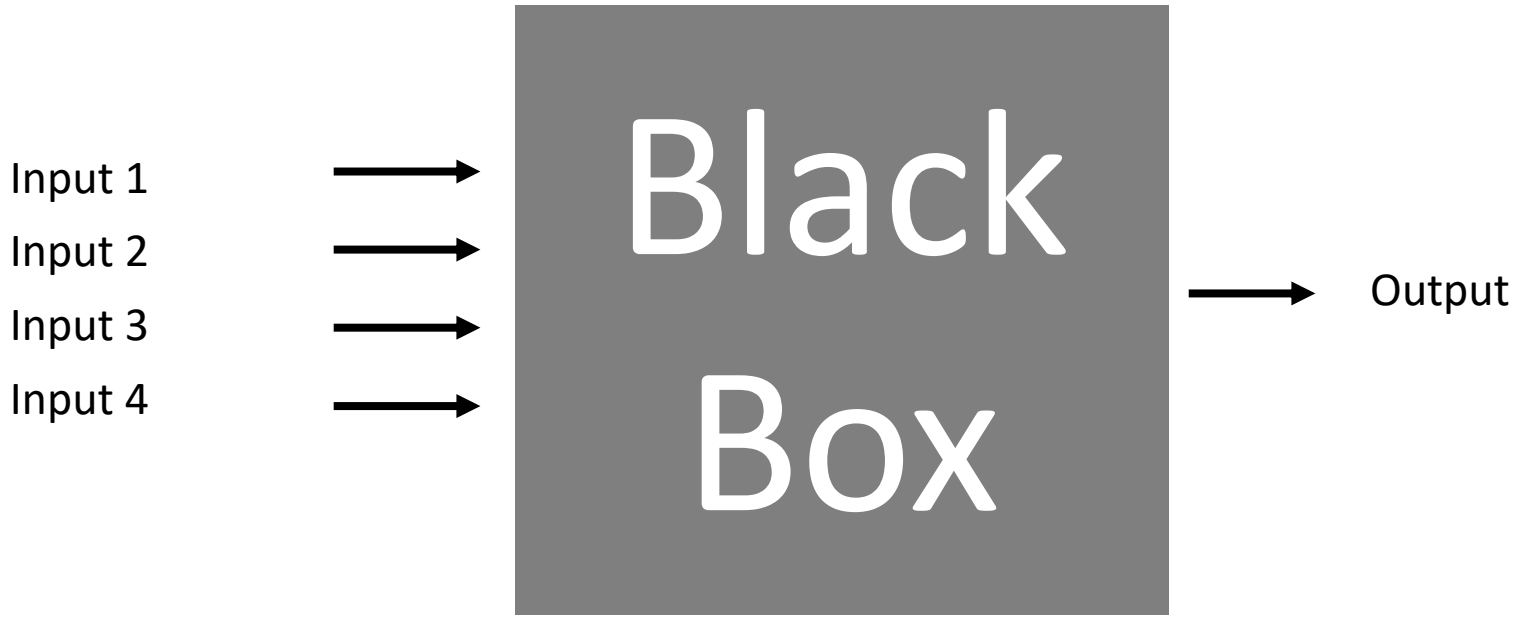
	Durchschnittliche Anzahl nicht-geteilter Links
AfD	2.6
Die Linke	3.1
Bündnis 90/Die Grünen	3.3
CSU	3.4
SPD	3.4
FDP	3.6
CDU	3.7

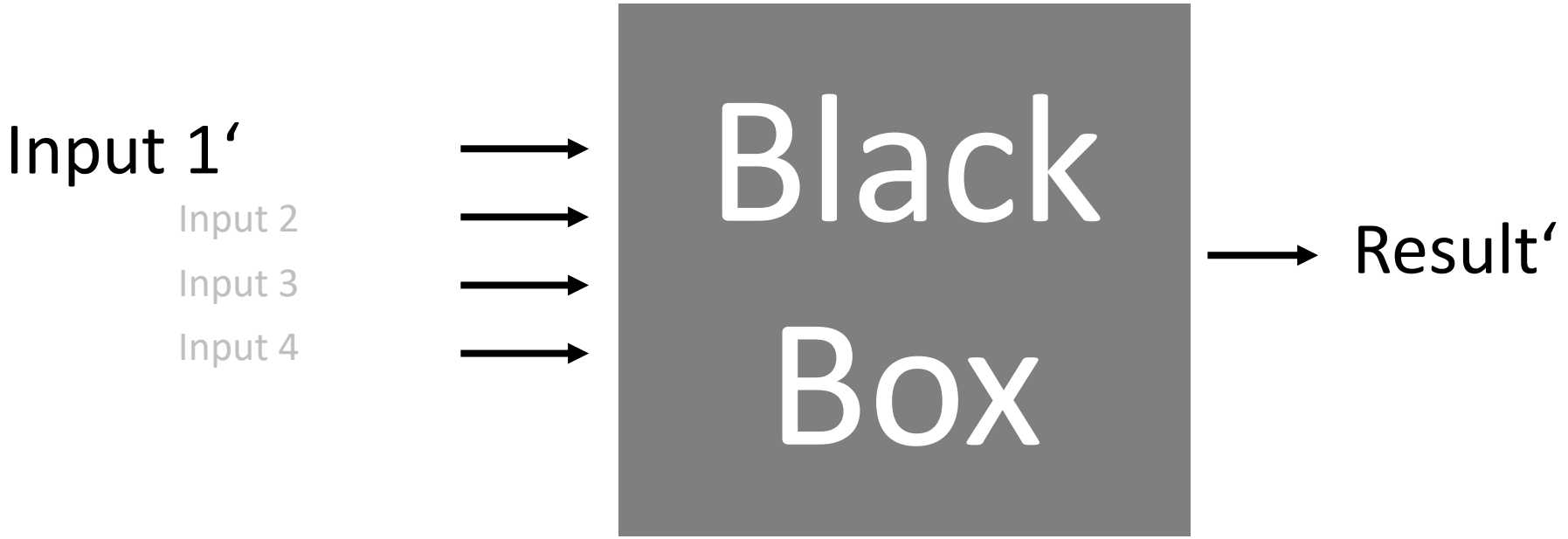
Alle Details hier: https://www.blm.de/aktivitaeten/forschung/datenspende.cfm

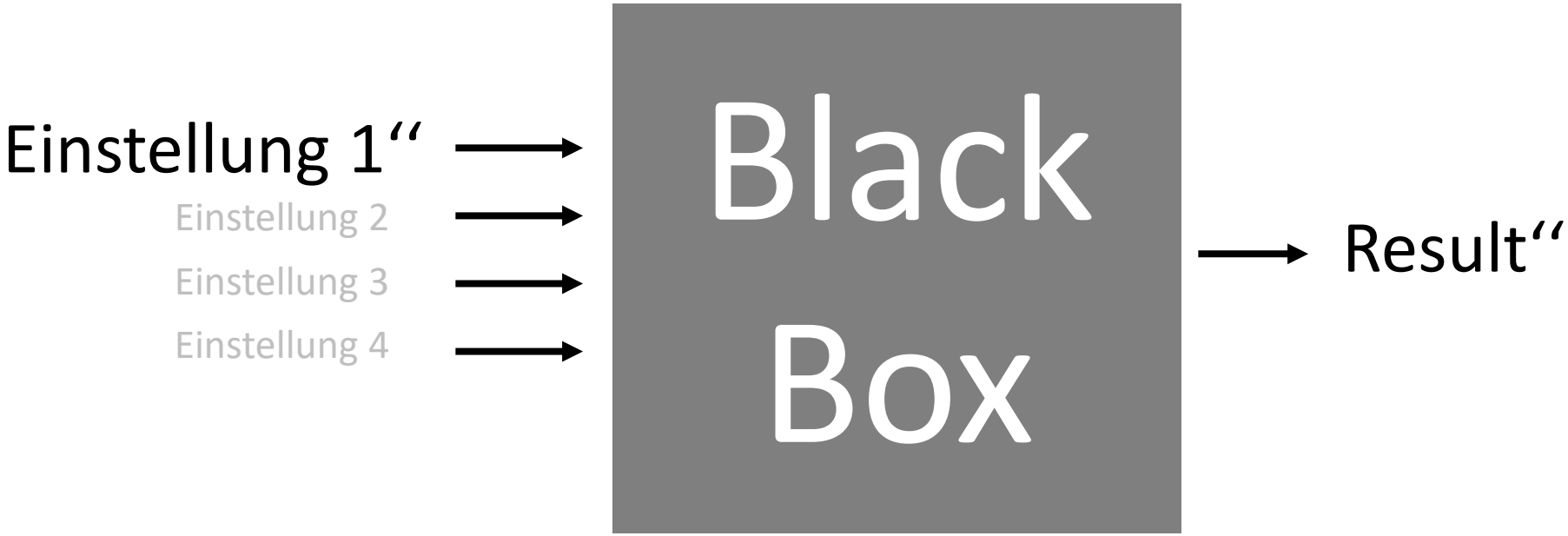

Google

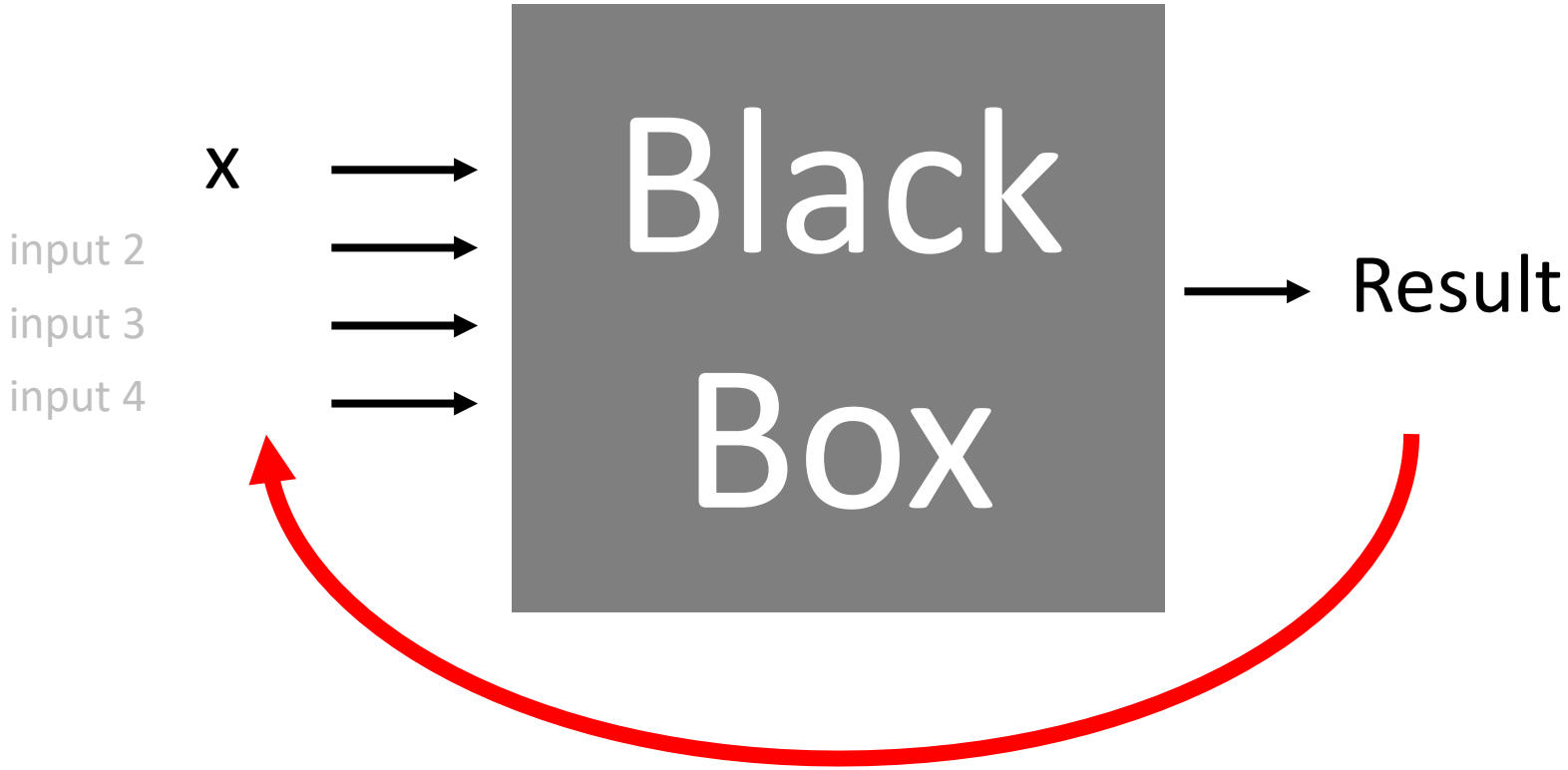
Black-Box-Methoden

- Variieren systematisch Input und beobachten Output.
- Inferieren Verhältnis zwischen Input und Output.
- Klassische Methodik aus den Naturwissenschaften, aber auch im Softwaretesting etabliert.




Input 1‘
Input 2
Input 3
Input 4
Black Box
Result‘


Einstellung 1“
Einstellung 2
Einstellung 3
Einstellung 4
Black Box
Result“



Result = Function (x, input 2, input 3, input 4)

Was kann damit sonst noch analysiert werden?

- Test auf Diskriminierung im Sinne von „*disparate impact*“ (siehe Prof. Basts Vortrag).
  - Geringerer Durchschnittslohn von Jobanzeigen für Frauen als für Männer[1].
  - Rückfälligkeitsvorhersage Kriminelle im COMPAS Algorithmus, der vor Gericht verwendet wird[2].
  - Diskriminierende Werbungsanzeigen bei Personensuche mit Namen afroamerikanischen Ursprungs[3].
  - Test auf Diskriminierung bei durch AI unterstütztem Bewerbungsprozess denkbar.
- Test auf Medienvielfalt, Verbreitung illegalen Contents, Überprüfung Netz-DG: z.B. Anteil Löschungsgrad.
- Test auf Personalisierungsausmaß bei allen personalisierten ADM-Systemen, z.B. politische Nachrichten im NewsFeed bei facebook.
- …

1) Datta, A.; Tschantz, M. C. & Datta, A.: „Automated Experiments on Ad Privacy Settings“, *Proceedings on Privacy Enhancing Technologies, Proceedings on Privacy Enhancing Technologies,* **2015***, 2015*, 92-112
2) https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3) Sweeney, L.: “Discrimination in Online Ad Delivery”, *ACM Queue,* **2013***, 56*, 44-54

Was ist dazu notwendig?

- Unlimitierter und „anonymer“ Zugang zum AI-System
  - Der Anbieter darf nicht „wissen“, dass dies die Testanfragen sind
- Kenntnis über genaue Input-Struktur und Output-Struktur.
- Bei Personalisierungsgradanalysen, u.U. Datenspenden von ‚echten‘ Nutzern notwendig wegen Datenhistorie (**Typ „Datenspende“**).
  - Kann oft durch „gefakte“ Nutzer simuliert werden.
  - Dann: Unlimitierte Anzahl von nicht erkennbaren Fake-Accounts notwendig.

Zusammenfassung

- Black-Box-Methode: Oft der einfachste Weg, um Verhalten von Algorithmen zu verstehen (es sei denn, Code ist sehr, sehr kurz).
- Transparenz von Code (=„Offenlegung“) nicht zielführend und oftmals schädlich (wirtschaftlich, aber auch wegen Manipulationsmöglichkeiten).
- Weitere Forschung zu Anwendungsbedingungen und Grenzen der Methodik notwendig.