UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

PRIVACY FÜR BIG DATA

PROF. DR. E. RAHM
UND MITARBEITER

WS 2015/16

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

GERMAN CENTERS FOR BIG DATA

Two Centers of Excellence for Big Data in Germany

- ScaDS Dresden/Leipzig
- Berlin Big Data Center (BBDC)

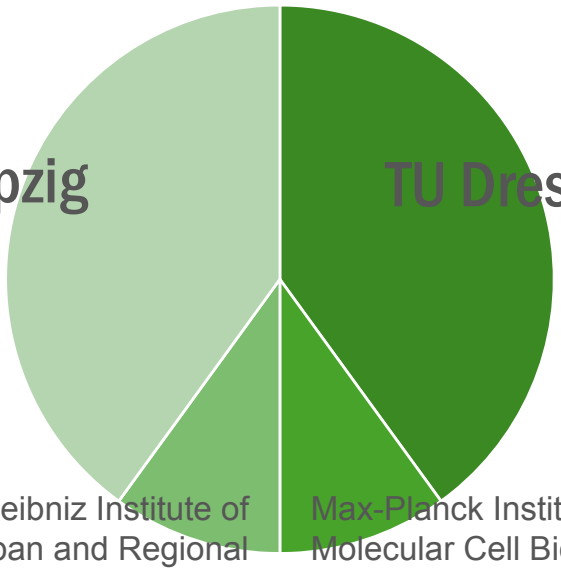
ScaDS Dresden/Leipzig (Competence Center for Scalable Data Services and Solutions Dresden/Leipzig)

- scientific coordinators: Nagel (TUD), Rahm (UL)
- start: Oct. 2014
- duration: 4 years (option for 3 more years)
- initial funding: ca. 5.6 Mio. Euro

ScaDS DRESDEN LEIPZIG GOALS

- Bundling and advancement of existing expertise on Big Data
- Development of Big Data Services and Solutions
- Big Data Innovations

ScaDS DRESDEN LEIPZIG FUNDED INSTITUTES



ScaDS DRESDEN LEIPZIG

STRUCTURE OF THE CENTER

- Life sciences
- Material and Engineering sciences
- Environmental / Geo sciences
- Digital Humanities
- Business Data

Service center

Big Data Life Cycle Management and Workflows

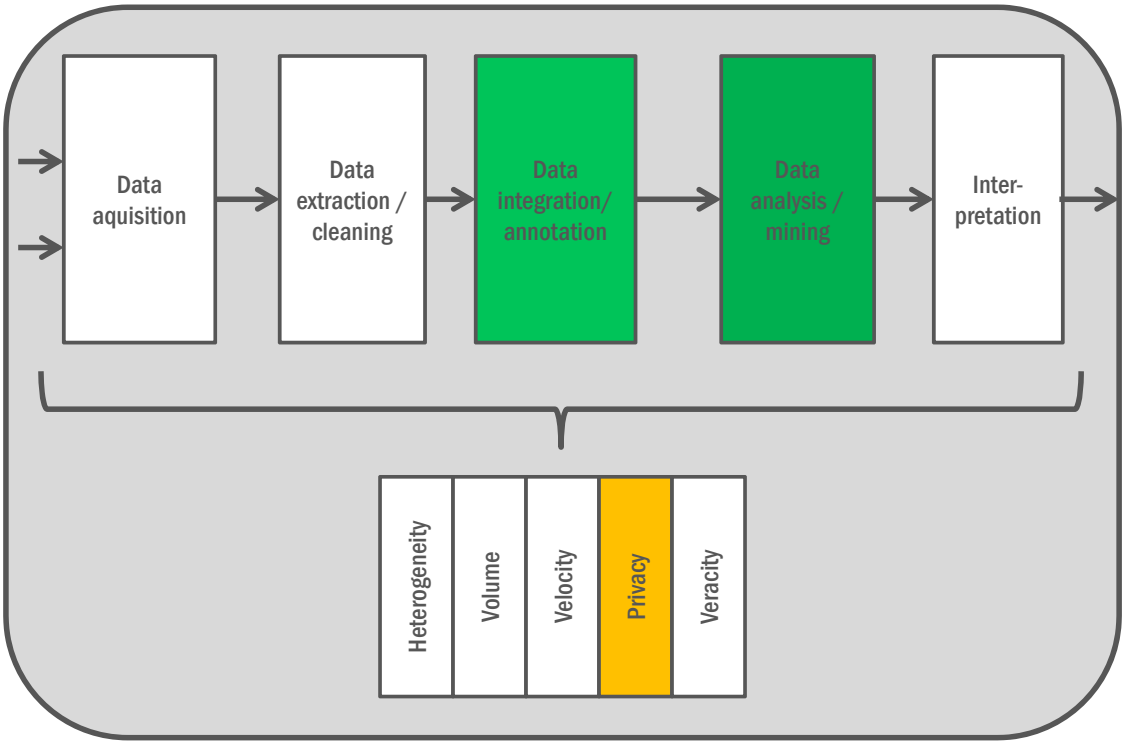
Data Quality / Data Integration	Knowledge Extraktion	Visual Analytics

Efficient Big Data Architectures

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

BIG DATA ANALYSIS PIPELINE



ScaDS DRESDEN LEIPZIG

PRIVACY

Privacy

- right of individuals to determine by themselves when, how and to what extent information about them is communicated to others (Agrawal 2002)

Privacy threats

- extensive collection of personal/private information / surveillance
- Information dissemination: disclosure of sensitive/confidential information
- Invasions of privacy: intrusion attacks to obtain access to private information
- Information aggregation: combining data, e.g. to enhance personal profiles or identify persons (de-anonymization)

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

INFORMATION / DATA PRIVACY

- Protection especially for *personally identifiable information* (PID)
  - name, birthdate, address, email address etc
  - healthcare and genetic records, financial records
  - criminal justice investigations and proceedings ...
- Challenge: preserve maximal privacy despite need to use person-related data for improved analysis / business success (advertisement, recommendations), website optimizations, clinical/health studies, identification of criminals ...
  - extensive tracking and profiling of web / smartphone / social network users (different kinds of cookies, canvas fingerprinting ...)
  - agreeing users are mostly unaware of full consequences

A privacy reminder from Google

To be consistent with data protection laws, we're asking you to take a moment to review key points of Google's Privacy Policy. This is not about a change we've made - but please review the key points below. **Click "I agree" to agree to the terms set out below; you can also explore other options on this page.** You can revoke your consent at any time with effect for the future.

Usage and content data

- When you use Google services to do things like write a message in Gmail or comment on a YouTube video, we store the information you create.
- When you search for a restaurant on Google Maps or watch a video on YouTube, for example, we process information about that activity – including information like the video you watched, device IDs, IP addresses, cookie data, and location.
- Our Privacy Policy contains further descriptions of the data we process.
- We treat all of this as "personal information" when it's associated with your Google Account.
- We also process the kinds of information described above when you use apps or sites that use Google services like ads, Analytics, and the YouTube video player.

We collect information in the following ways: **Information you give us.** for a Google Account. we ask for personal information, like your name, email address, phone number or credit card . public Google Profile may include your name and photo.

•**Information we get from your use of our** services**.** W collect information about how you use services, like when you watch a YouTube video, visit a website that uses our advertising services, or view and interact with our ads and content including

- **Device information**: collect device-specific information (hardware model, operating system version, unique device identifiers, and mobile network information including phone number). Google may associate your device identifiers or phone number with your Google Account.
- **Log information**; we automatically collect and store certain information in server logs. This includes:
  - search queries, your phone number, calling-party number, time and date of calls, duration of calls, SMS routing information and types of calls.
  - Internet protocol address.
  - cookies that may uniquely identify your browser or your Google Account.
- **Location information:** we collect and process information about your actual location. We use various technologies to determine location, including IP address, GPS, and other sensors that may, for example, provide Google with information on nearby devices, Wi-Fi access points and cell towers.

Purposes of the data processing

We process this data for the purposes described in our policy, including to:

- Help our services deliver more useful, customized content such as more relevant search results, based on your interests derived from such data;
- Improve the quality of our services and develop new ones;
- Deliver ads based on your interests, which we can determine based on this data, like ads that are related to things such as search queries or videos you've watched on YouTube;
- Improve security by protecting against fraud and abuse; and
- Conduct analytics and measurement to understand how our services are used.

Combining data

We also combine this data among our services and across your devices for these purposes. For example, we show you ads based on information about your interests, which we can derive from your use of Search and Gmail, and we use data from trillions of search queries to build spell-correction models that we use across all of our services.

Tracking auf Web-Seiten (Bsp.: Huffingtonpost)

COOKIES ERMÖGLICHEN EINE VIELZAHL VON FUNKTIONEN, DIE IHREN BESUCH BEI DER HUFFINGTON POST ANGENEHMER GESTALTEN. INDEM SIE DIESE WEBSITE BENUTZEN, STIMMEN SIE DER NUTZUNG VON COOKIES GEMÄSS UNSEREN RICHTLINIEN ZU. FÜR WEITERE INFORMATIONEN AUCH ZU IHREM WIDERSPRUCHSRECHT, KLICKEN SIE BITTE HIER. ICH STIMME ZU

Cookies und andere Technologien

Ebenso wie dies standardmäßig auf vielen anderen Websites erfolgt, können wir auf Ihrem Computer Cookies, Web-Beacons, Flash-Cookies und andere Technologien einsetzen und auf diese zugreifen.
Mit Ihrer Nutzung der Website www.huffingtonpost.de (die „Seite") stimmen Sie dem Einsatz von Cookies zu.

Bitte beachten Sie, dass Dritte wie zum Beispiel Anzeigennetzwerke, Werbeagenturen, Werbeunternehmen und Anbieter von Zielgruppensegmentierung ebenfalls Cookies, Web-Beacons, Flash-Cookies und andere Technologien auf Ihrem Computer einsetzen und auf diese zugreifen können, wenn Sie die Seite besuchen. Weitere Informationen entnehmen Sie bitte unserer Erörterung von Werbung durch Dritte.

Web-Beacons

Wir können auch Web-Beacons (auch transparente GIFs, Web-Wanzen oder Zählpixel genannt) einsetzen. Bei diesen Technologien handelt es sich um Code-Strings, die ein winziges grafisches Bild auf einer Webseite oder in einer E-Mail bereitstellen. Web-Beacons können bestimmte Arten von Informationen auf Ihrem Computer erkennen, etwa Cookies, Zeit und Datum des Seitenaufrufs sowie eine Beschreibung der Seite, auf der sich das Web-Beacon befindet.

Flash-Cookies

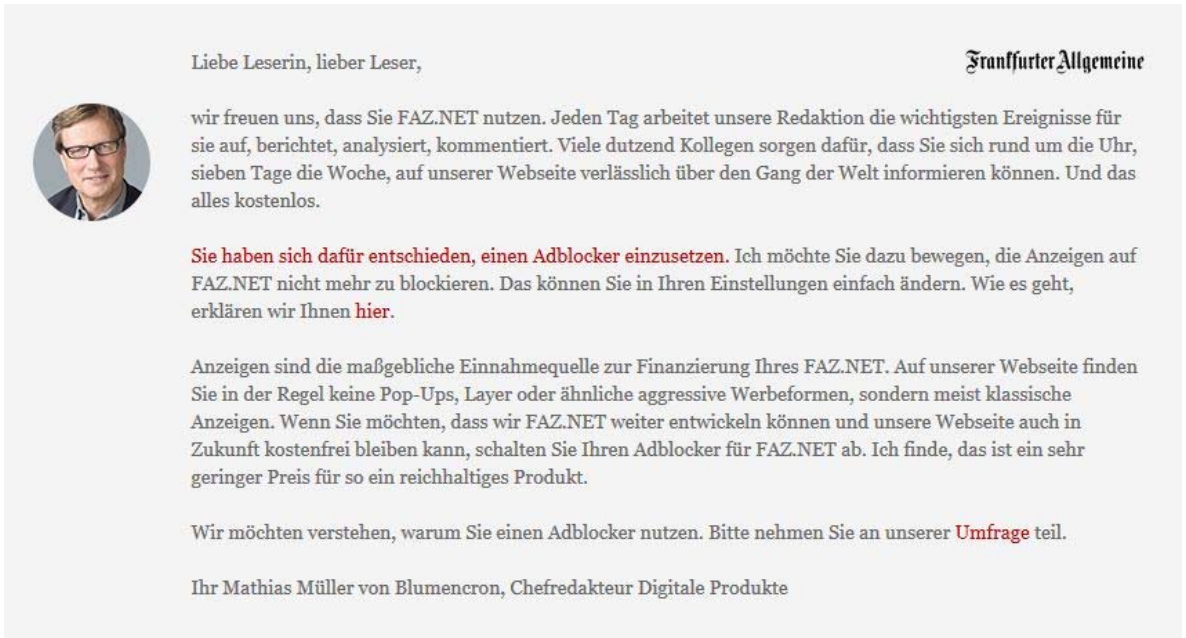
Einige Drittpartner der Seite können Flash-Cookies, auch als Local Shared Objects (LSOs) bekannt, einsetzen. Diese werden verwendet, um Ihre Interessen anhand der Artikel, die Sie lesen, zu bestimmen und allgemeiner um das Benutzerverhalten auf verschiedene Arten zu verfolgen. LSOs speichern Sammlungen Cookie-ähnlicher Daten in einem Verzeichnis auf dem Computer eines Benutzers und werden durch den Flash-Videoplayer von Adobe installiert.

Liste der (68) Cookies

(Huffingtonpost.de)

Unternehmen	Domäne	Cookie-Name	Zweck
[x+1]	.ru4.com	X1ID	ADVERTISING
Adform ApS	.adform.net	uid	ADVERTISING
Adform ApS	track.adform.net	cid	ADVERTISING
Adform ApS	c1.adform.net	cid	ADVERTISING
ADITION technologies AG	adfarm1.adition.com	UserID1	ADVERTISING
Adobe	dpm.demdex.net	dpm	ADVERTISING
Adobe	.demdex.net	demdex	ADVERTISING
AOL Inc.	b.aol.com	MUNAUTHID	ANALYTICS
AOL Inc.	b.aol.com	dlV	ANALYTICS
Omniture	.aol.com	s_vi	ANALYTICS
AOL Inc.	.aol.com	UNAUTHID	ANALYTICS
AOL Inc.	.aol.com	CUNAUTHID	ANALYTICS
AOL Advertising	.adtech.de	JEB2	ADVERTISING
Huffington Post	www.huffingtonpost.de	huffpo_session	ANALYTICS
Huffington Post	www.huffingtonpost.de	mbox	ANALYTICS
Huffington Post	www.huffingtonpost.de	_polar_tu	ANALYTICS
Huffington Post	www.huffingtonpost.de	tfm_rsi_segs	ANALYTICS
Huffington Post	www.huffingtonpost.de	yp_rec	ANALYTICS
Huffington Post	.huffingtonpost.de	s_pers	ANALYTICS
Huffington Post	.huffingtonpost.de	UNAUTHID	ANALYTICS
Huffington Post	.huffingtonpost.de	CUNAUTHID	ANALYTICS
Huffington Post	.huffingtonpost.de	rsi_segs	ANALYTICS
Huffington Post	.huffingtonpost.de	__qca	CONTENT
Huffington Post	.huffingtonpost.de	__qseg	ANALYTICS
Huffington Post	.huffingtonpost.de	__utma	ANALYTICS
Huffington Post	.huffingtonpost.de	__utmb	ANALYTICS
Huffington Post	.huffingtonpost.de	__utmz	ANALYTICS
Huffington Post	.huffingtonpost.de	__utmv	ANALYTICS
Huffington Post	.huffingtonpost.de	__utmt	ADVERTISING
Huffington Post	.huffingtonpost.de	disable_xd	CONTENT
Huffington Post	.huffingtonpost.de	huffpo_type_view	CONTENT
AppNexus	.adnxs.com	uuid2	ADVERTISING
AppNexus	.adnxs.com	sess	ADVERTISING
AppNexus	.adnxs.com	anj	ADVERTISING
AudienceScience	.revsci.net	pudm_AAAA	ADVERTISING
AudienceScience	.revsci.net	NETID01	ADVERTISING
AudienceScience	.revsci.net	rts_AAAA	ADVERTISING
AudienceScience	.revsci.net	rtc_AAAA	ADVERTISING
auFeminin.com SA	.smartadserver.com	pid	ADVERTISING
auFeminin.com SA	.smartadserver.com	TestIfCookieP	ADVERTISING
auFeminin.com SA	.smartadserver.com	csync	ADVERTISING
Company [adscale.de]	.adscale.de	uu	ADVERTISING
Company [adscale.de]	ih.adscale.de	tu	ADVERTISING
Company [krxd.net]	.krxd.net	ServedBy	ADVERTISING
Company [krxd.net]	.krxd.net	_kuid_	ADVERTISING
ComScore	.scorecardresearch.com	UID	ANALYTICS
ComScore	.scorecardresearch.com	UIDR	ANALYTICS
DataXu	.w55c.net	wfivefivec	ADVERTISING
DoubleClick (Google)	.doubleclick.net	id	ADVERTISING
Efficient Frontier	.everesttech.net	gglck	ADVERTISING
Efficient Frontier	.everesttech.net	ev_t	ADVERTISING
Efficient Frontier	.everesttech.net	everest_g_v2	ADVERTISING
Improve Digital BV	ad.360yield.com	tuuid	ADVERTISING
Improve Digital BV	ad.360yield.com	um	ADVERTISING
Improve Digital BV	ad.360yield.com	umeh	ADVERTISING
MediaMath llc	.mathtag.com	uuid	ADVERTISING
MediaMath llc	.mathtag.com	mt_misc	ADVERTISING
MediaMath llc	.mathtag.com	uuidc	ADVERTISING
MediaMath llc	.mathtag.com	mt_mop	ADVERTISING
Parse.ly	parse.ly	parsely_network_uuid	ANALYTICS
Parse.ly	parse.ly	parsely_uuid	ANALYTICS
PubMatic Inc	.pubmatic.com	PUBRETARGET	ADVERTISING
Quantcast	.quantserve.com	mc	ADVERTISING
Shopzilla Inc	.connexity.net	COu	ADVERTISING
SiteScout inc	.sitescout.com	ssi	ADVERTISING
The Nielsen Company	.imrworldwide.com	IMRID	ADVERTISING
Turn Inc.	.turn.com	uid	ADVERTISING
Yahoo	.yahoo.com	B	ANALYTICS

Ad Blocking -> Content Blocking



Liebe Leserin, lieber Leser,

Frankfurter Allgemeine

wir freuen uns, dass Sie FAZ.NET nutzen. Jeden Tag arbeitet unsere Redaktion die wichtigsten Ereignisse für sie auf, berichtet, analysiert, kommentiert. Viele dutzend Kollegen sorgen dafür, dass Sie sich rund um die Uhr, sieben Tage die Woche, auf unserer Webseite verlässlich über den Gang der Welt informieren können. Und das alles kostenlos.

Sie haben sich dafür entschieden, einen Adblocker einzusetzen. Ich möchte Sie dazu bewegen, die Anzeigen auf FAZ.NET nicht mehr zu blockieren. Das können Sie in Ihren Einstellungen einfach ändern. Wie es geht, erklären wir Ihnen hier.

Anzeigen sind die maßgebliche Einnahmequelle zur Finanzierung Ihres FAZ.NET. Auf unserer Webseite finden Sie in der Regel keine Pop-Ups, Layer oder ähnliche aggressive Werbeformen, sondern meist klassische Anzeigen. Wenn Sie möchten, dass wir FAZ.NET weiter entwickeln können und unsere Webseite auch in Zukunft kostenfrei bleiben kann, schalten Sie Ihren Adblocker für FAZ.NET ab. Ich finde, das ist ein sehr geringer Preis für so ein reichhaltiges Produkt.

Wir möchten verstehen, warum Sie einen Adblocker nutzen. Bitte nehmen Sie an unserer Umfrage teil.

Ihr Mathias Müller von Blumencron, Chefredakteur Digitale Produkte

ScaDS DRESDEN LEIPZIG

PRIVACY FOR BIG DATA

- Need for comprehensive privacy support (“privacy by design”)
- Privacy-preserving publishing of datasets
  - Anonymization of datasets
- Privacy-preserving record linkage
  - object matching with encoded data to preserve privacy
- Privacy-preserving data mining
  - analysis of anonymized data without re-identification

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

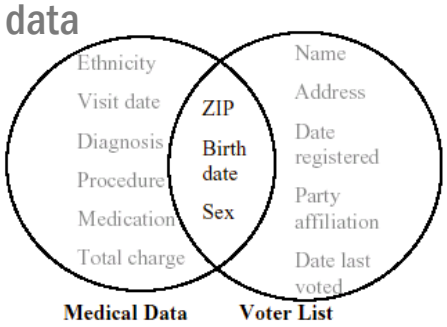
ANONYMIZATION VS PSEUDONYMIZATION

- Anonymization
  - removing, generalizing or changing personally identifying attributes so that people whom the data describe remain anonymous
  - no way to identify different records for same person (e.g., different points in time) or to match/combine records from different sources
- Pseudonymization
  - replace most identifying fields within a data record are replaced by one or more artificial identifiers, or pseudonyms
  - one-way pseudonymization (e.g. one-way hash functions) vs. re-identifiable pseudonymization
  - records with same pseudonym can be matched
  - improved potential for data analysis

ScaDS DRESDEN LEIPZIG

RE-IDENTIFICATION OF „ANONYMOUS DATA“ (SWEENEY 2001)

- US voter registration data
  - 69% unique on postal code (ZIP) and birth date
  - 87% US-wide with sex, postal code and birth data
- Solution: K-Anonymity
  - any combination of values appears at least k times
  - generalize values, e.g., on ZIP or birth date



ScaDS DRESDEN LEIPZIG

K-ANONYMITY EXAMPLE

ID	ZIP	AGE	DISEASE	TREATMENT
1	12345	23	Gastric ulcer	Antacid
2	12345	29	Gastritis	Acid-reducing drug
3	12363	41	Flu	Antipyretic drug
4	12361	43	Stomach cancer	Cytostatic drug
5	12362	59	Pneumonia	Antibiotics
6	12471	52	Bronchitis	Antibiotics
7	12473	55	Flu	Antipyretic drug

(a) Microdata-table

ID	ZIP	AGE	DISEASE	TREATMENT
1	123**	[20-29]	Gastric ulcer	Antacid
2	123**	[20-29]	Gastritis	Acid-reducing drug
3	123**	[40-49]	Flu	Antipyretic drug
4	123**	[40-49]	Stomach cancer	Cytostatic drug
5	123**	[50-59]	Pneumonia	Antibiotics
6	124**	[50-59]	Bronchitis	Antibiotics
7	124**	[50-59]	Flu	Antipyretic drug

(b) 2-anonymous table

from: Nielsen et al: Proc BTW 2015

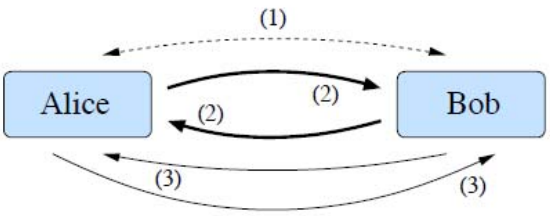
ScaDS DRESDEN LEIPZIG

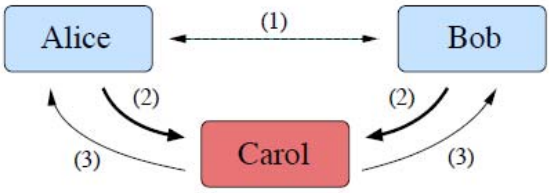
PRIVACY-PRESERVING RECORD LINKAGE (PPRL)

- object matching with encoded data to preserve privacy
  - trivial for exact match on simple pseudonyms / codes (standard practice in clinical studies)
- privacy aspects
  - need to support secure 1-way encoding
  - protection against attacks to identify persons
- conflicting requirements:
  - high privacy
  - match effectiveness (need to support fuzzy matches)
  - scalability to many parties and large datasets

ScaDS DRESDEN LEIPZIG

BASIC PPRL CONFIGURATIONS





- Two-party protocols
  - only two data owners communicate who wish to link their data
- Three-party protocols
  - Use of a trusted third party (linkage unit)
  - LU will never see unencoded data, but collusion is possible
- Multi-party protocols (> 2 data owners)
  - with or without linkage unit

ScaDS DRESDEN LEIPZIG

PRIVACY ASPECTS

- Adversary models (from crytography)
  - Honest-But-Curious: parties follow agreed-on protocols
  - Malicious
- Privacy attacks
  - Crack data encoding: Frequency attack, Dictionary attack, ...
  - Collusion between parties

UNIVERSITÄT LEIPZIG

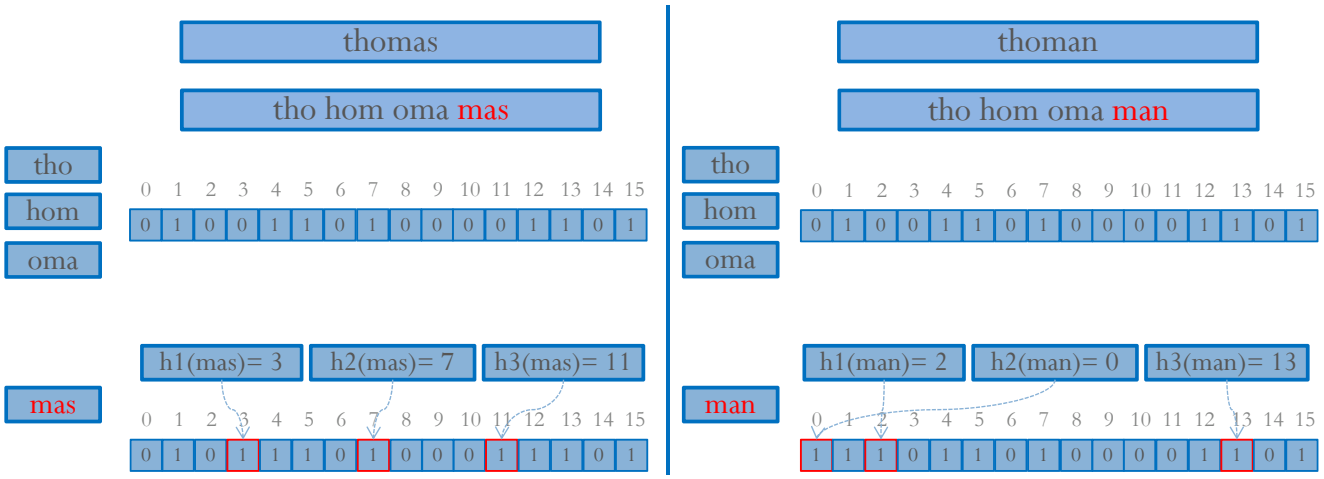
ScaDS DRESDEN LEIPZIG

PPRL WITH BLOOM FILTERS

- effective and simple encoding uses cryptographic bloom filters (Schnell et al, 2009)
- tokenize all match-relevant attribute values, e.g. using bigrams or trigrams
  - typical attributes: first name, last name (at birth), sex, date of birth, country of birth, place of birth
- map each token with a family of one-way hash functions to fixed-size bit vector (fingerprint)
  - original data cannot be reconstructed
- match of bit vectors (Jaccard similarity) is good approximation of true match result

ScaDS DRESDEN LEIPZIG

SIMILARITY COMPUTATION - EXAMPLE

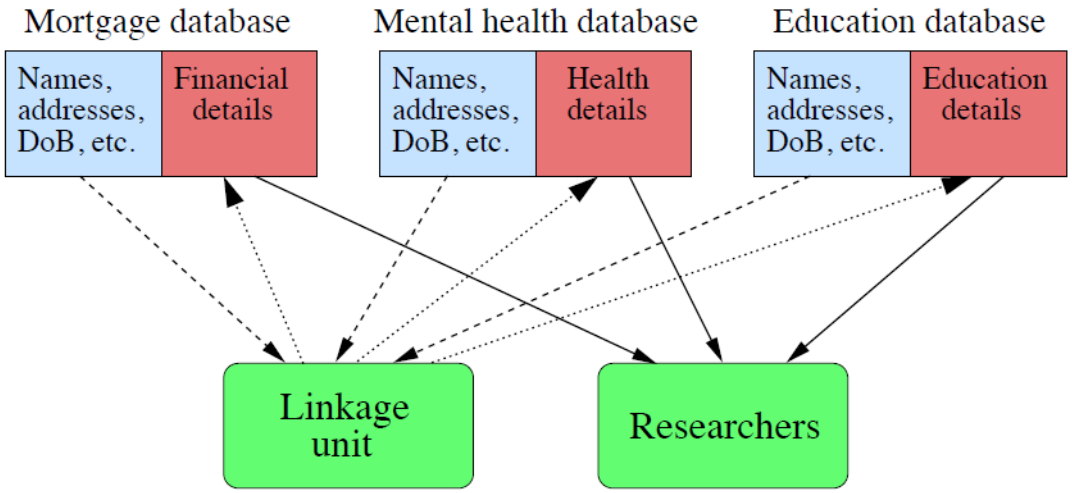


$$Sim_{Jaccard}(r1, r2) = (r1 \wedge r2) / (r1 \vee r2)$$

$$Sim_{Jaccard}(r1, r2) = 7/11$$

ScaDS DRESDEN LEIPZIG

PPRL EXAMPLE IN HEALTH DOMAIN



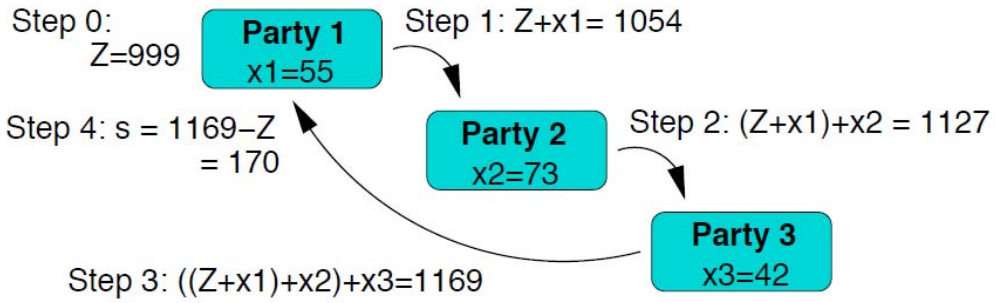
Step 1: Database owners send partially identifying data to linkage unit
Step 2: Linkage unit sends linked record identifiers back
Step 3: Database owners send 'payload' data to researchers

Details given in: Chris Kelman, John Bass, and D'Arcy Holman: *Research use of Linked Health Data – A Best Practice Protocol*, Aust NZ Journal of Public Health, vol. 26, 2002.

ScaDS DRESDEN LEIPZIG

SECURE MULTI-PARTY COMPUTATION (SMC)

- Compute a function across several parties, such as no party learns the information from the other parties, but all receive the final results
- Example 1: millionaire problem
  - two millionaires, Alice and Bob, are interested in knowing which of them is richer but without revealing their actual wealth.
- Example 2: secure summation

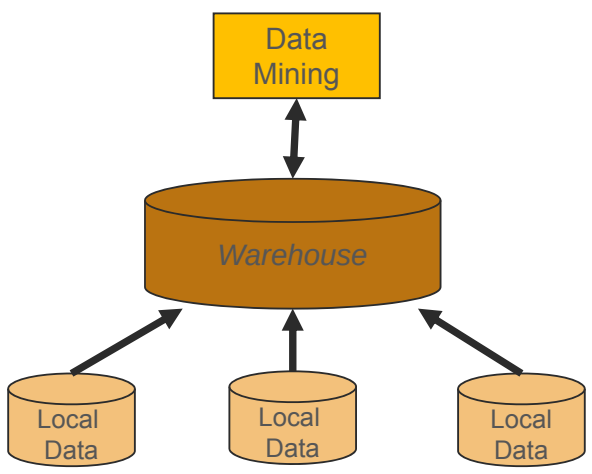


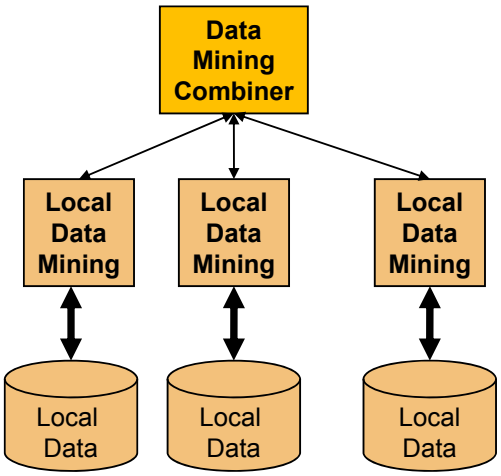
UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

PRIVACY-PRESERVING DATA MINING

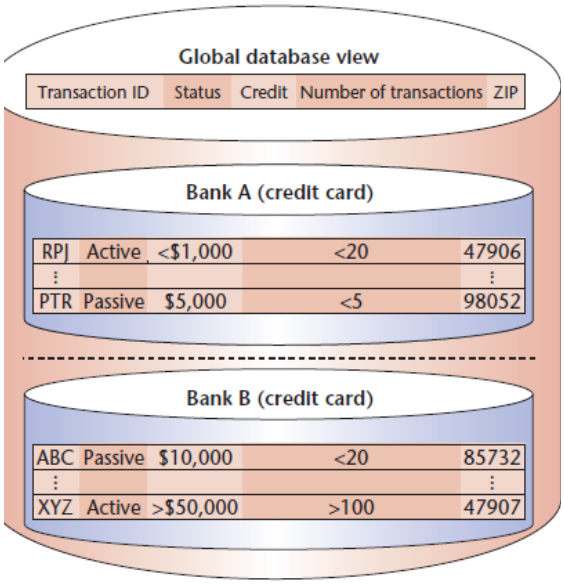
- Physically integrated data (e.g. data warehouse) about persons entails greatest privacy risks
- Data mining over distributed data can better protect personal data by limiting data exchange, e.g. using SMC methods

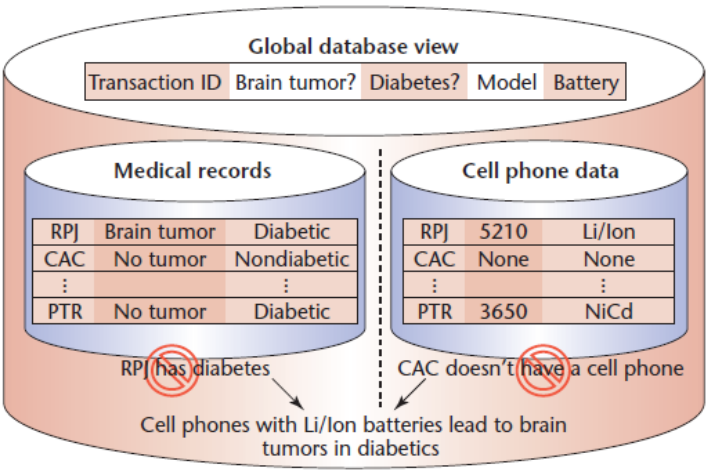




ScaDS DRESDEN LEIPZIG

HORIZONTAL VS VERTICAL DATA DISTRIBUTION





Source: J. Vaidya, C. Clifton: Privacy-Preserving data mining: Why, How, and When. IEEE Security&Privacy 2004

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

SEMINAR

ScaDS DRESDEN LEIPZIG

SEMINARZIELE

- Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
  - kann Grundlage für Abschlussarbeit oder SHK-Tätigkeit sein
- Erarbeitung + Durchführung eines Vortrags unter Verwendung wissenschaftlicher (englischer) Literatur
- Diskussion
- Schriftliche Ausarbeitung zum Thema
- Hilfe und Feedback durch zugeteilten Betreuer

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

SEMINAR: ANRECHNUNGSMÖGLICHKEITEN

- Masterstudium, insbesondere für Schwerpunkt „Big Data“
  - Teil der Module Moderne Datenbanktechnologien
  - Seminarmodul
- Bachelorstudium
  - Seminarmodul
  - evtl. Realisierung von IS

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

SCHEINVERGABE / MODULPRÜFUNG

- selbständiger Vortrag mit Diskussion (ca. 45 Minuten)
  - Abnahme der Folien durch Betreuer
- schriftliche Ausarbeitung (ca. 15 Seiten)
  - Abnahme der Ausarbeitung durch Betreuer
  - Ausarbeitung soll zum Vortragstermin vorliegen (Vorträge ab Januar 2016)
- aktive Teilnahme an allen Vortragsterminen
- Modul-Workload: 30h Präsenzzeit,
  120 h Selbststudium

UNIVERSITÄT LEIPZIG

ScaDS DRESDEN LEIPZIG

SEMINAR (3)

- Themenzuordnung
  - Koordinierungstreffen mit Betreuer bis spätestens 4.11.2015
  - ansonsten verfällt Seminaranmeldung
  - freiwilliger Rücktritt auch bis max. 4.11.2015
- Vortragstermine
  - 5x freitags, P801, ab 8. 1. 2016
  - max. 2 Doppelstunden ab 13:15 Uhr

Komplex	Betreuer	max. #Themen	Termin	Studenten
Einleitung (Begriffe, Gesetze) **Online Privacy** (Grundlagen, Web Tracking)	Nentwig	1 2	8.1.	Landwehrkamp Thamm Frei
Privacy für soziale Netze / Graphen	Junghanns	3	8.1 15.1	Gomez, Otto, Gatz
Anonymisierungstechniken K-Anonymity, differential privacy ...	Peukert	2 (-3)	15.1.	Barcik, Hüsing Blanck
Privacy-preserving Data Mining Grundlagen, Assoc. Rules, graph pattern mining Datenaggr. v. Sensordaten, anonyme Recommendations	Petermann Christen	3 2	22.1.	Meyer, Fahrenberg, Bigl Kießling, Güth
Privacy-preserving Record Linkage Grundlagen, Performance-Optimierungen Symmetr. Verfahren (SMC)	Sehili Christen	2(-3) 1	22.1 29.1.	Karthaus, Hillebrand Noncke Reder
Privacy für biomedizinische Anwendungen	Groß	3	5.2.	Swoboda Anders, Möller
SQL-Ausführung auf verschlüsselten Daten	Peukert	1	5.2.	Lange, Fricke?