Ievads datu pseidonimizācijā un anonimizācijā

Līna Lāže

RSU datu kuratore

19.03.2026

Finansē Eiropas Savienība NextGenerationEU

2027 Nacionālais attīstības plāns

Projekta *«Atbalsts atvērtās zinātnes ieviešanai praksē, kā arī izveidoti risinājumi zinātnes datu koplietošanai un dalībai ES atvērtajā zinātnes mākonī»* (ANM 2.1.3.1.i.)

Semināra plāns

- Personas dati un to minimizācijas spektrs
- Pseidonimizācija
- Pseidonimizācijas metodes
- Anonimizācija
- Anonimizācijas raksturlielumi un pielietojamās metodes
- MI drošas lietošanas pamatprincipi

Personas dati

Personas dati ir jebkura informācija, kas attiecas uz identificētu vai identificējamu fizisku personu (datu subjektu). VDAR 4. pants

«Identificēta fiziska persona»

Ārsts medicīnas kartē raksta «Jānis Bērziņš, personas kods 123456-12345, diagnoze — augsts asinsspiediens.» Šeit uzreiz ir skaidrs, par ko ir runa.

«Identificējama fiziska persona»

Tavs telefons sūta atrašanās vietas datus lietotnei. Redzot, ka katru nakti esi vienā adresē un katru dienu citā — var noteikt tavu māju un darbavietu, līdz ar to — tevi.

«Jebkura informācija»

Teksts (tavs vārds anketā)
Attēls (fotogrāfija no balīītes)
Skaņa (balss ieraksts zvanu centrā)
Video (ieraksts no veikala videonovērošanas kameras)
Cipari (tavs telefona numurs)
Atrašanās vieta (GPS koordinātes no tava telefona)

Personas tiešie un netiešie identifikatori

Tiešie identifikatori
nepārprotami identificē konkrētu personu bez papildu informācijas un viennozīmīgi norāda uz vienu konkrētu indivīdu

Personu Datu Saraksts - Tiešie Identifikatori

A	B	C	D	E
Personas kods	Vārds	Uzvārds	Telefona numurs	E-pasta adrese
010190-1XXXX	Jānis	Ozoliņš	+371 29XXXXX	janis.ozolins@firma.lv
010190-1XXXX	Jānis	Ozoliņš	+371 29XXXXX	karis.ozolins@firma.lv
010190-1XXXX	Martas	Ozoliņš	+371 29XXXXX	kan.ozolins@firma.lv
010190-1XXXX	Martan	Ozoliņš	+371 29XXXXX	kom.onus@firma.lv
010190-1XXXX	Jānis	Ozoliņš	+371 29XXXXX	man.solins@firma.lv
010190-1XXXX	Martna	Ozoliņš	+371 29XXXXX	mars.zolins@firma.lv
010190-1XXXX	Vedora	Nelili	+371 29XXXXX	lemiz.omuos@firma.lv
010190-1XXXX	Jānis	Ozoliņš	+371 29XXXXX	jana.cerins@firma.lv
010190-1XXXX	Jānis	Ozoliņš	+371 29XXXXX	bolt.oarins@firma.lv

Pētījuma Dalībnieku Saraksts - Netiešie Identifikatori

A	B	C	D	E
Pētījuma ID	Vecums	Dzimums	Augums (cm)	Saslimšanas
P-001	28	Sieviete	168	Alerģija (putnu spalvas)
P-002	55	Vīrietis	182	Hipertenzija
P-003	41	Sieviete	160	Artrīts
P-004	63	Vīrietis	175	Astma (sezonāla)
P-005	19	Sieviete	155	Nav konstatētu saslimšanu
P-006	35	Vīrietis	190	Muguras sāpes
P-007	70	Sieviete	158	Osteoporoze, Diabēts
P-008	48	Vīrietis	178	Alerģija (putekšņi)
P-009	22	Sieviete	162	Anēmija

Netiešie identifikatori (Kvazi identifikatori)
dati nav pietiekami, lai identificētu personu, bet kombinācijā ar citiem datiem var norādīt uz konkrētu indivīdu

Vispārīgā datu aizsardzības regula

Reizēm tiek saīsināta kā VDAR vai GDPR (General Data Protection Regulation)

Noteikumu kopums, kas nosaka, kā uzņēmumi, organizācijas un, arī pētnieki, drīkst vākt, lietot, glabāt un dzēst cilvēku personīgo informāciju.

Tas stājās spēkā 2018. gadā, lai atjauninātu datu aizsardzības likumus digitālajam laikmetam un sniegtu cilvēkiem lielāku kontroli pār viņu datiem.

Priekšrocības

- Veicina uzticēšanos
- Liek noteikt datu izmantošanas mērķi
- Starptautiska saderība
- Cilvēkiem dod tiesības
- Sniedz datu drošības garantijas

Trūkumi

- Birokrātisks slogs
- Stingrākas piekrišanas prasības
- Prasības datu drošībai
- Nepieciešams datus minimizēt
- Ierobežojumi datu nodošanai
- Sodi par pārkāpumiem

Datu minimizācijas metodes

Anonimizācija

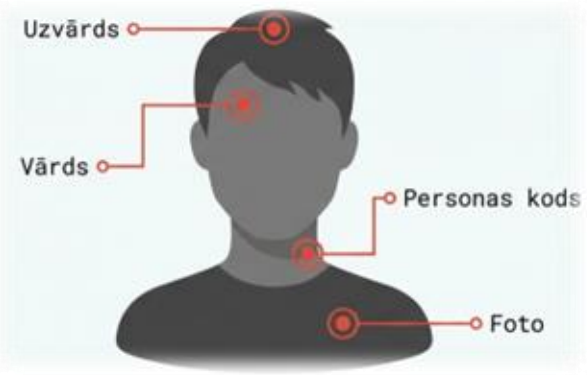
process, kurā personas dati tiek neatgriezeniski pārveidoti tā, lai konkrēto cilvēku (datu subjektu) vairs nekādā veidā - ne tieši, ne netieši, ne izmantojot papildu informāciju - nevarētu identificēt nedz datu apstrādātājs, nedz jebkura cita trešā persona

Pseidonimizācija

process, kurā tiešie identifikatori (piemēram, vārds vai personas kods) tiek aizstāti ar mākslīgu identifikatoru (pseidonīmu, kodu vai numuru). Šādus datus joprojām var sasaistīt ar konkrēto personu, taču tikai tad, ja tiek izmantota papildu informācija (atslēga vai atšifrēšanas tabula), kas tiek glabāta droši un atsevišķi

Datu minimizācijas spektrs

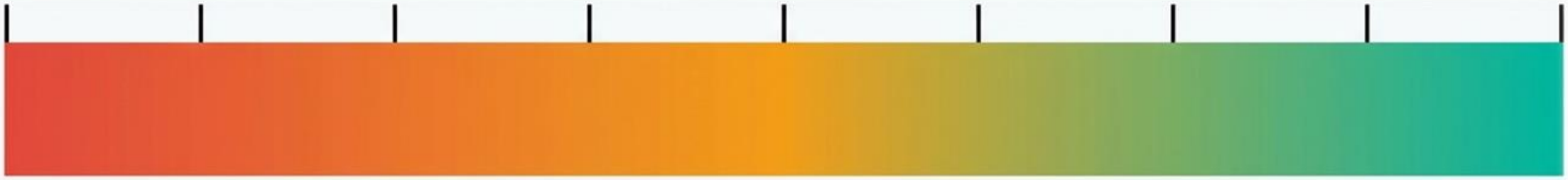
Dati ar tiešiem identifikatoriem



Pseidonimizēti dati

Dati ar re-identifikācijas risku

Anonimizētii dati



Datu drošība nav binära, tas ir process.

Mērķis ir rast līdzsvaru starp lietderību un privātumu.

Kritiskā atšķirība

Pseidonimizācija

- Atgriezeniska
- Jāievēro VDAR (GDPR)
- Pastāv atslēga

≠

Anonimizācija

- Neatgriezeniska
- VDAR vairs netiek piemērots
- Identifikācija nav iespējama

Pseidonimizācija

Pētnieku klasiskā kļūda.

Es aizstāju dalībnieku vārdus ar ID numuriem. Mani dati tagad ir anonīmi, un VDAR (GDPR) vairs nav piemērojams.

Kods nav anonimizēšana!

Pseidonimizācijas realitāte pētniecībā un datu drošībā.

Tā ir bīstama ilūzija.

Personas tiešo identifikatoru aizstāšana ar kodu NAV anonimizēšana.
Datus uzskata par anonimizētiem tikai tad, ja re-identifikācija nav iespējama.

Iepazīstieties: Pseidonimizācija.

Tas ir atgriezenisks process. Personas tiešie identifikatori tiek aizstāti ar kodiem vai pseidonīmiem. Jūs apgrūtināt citu iespējas re-identificēt datus, bet jūs paši saglabājat atslēgu.

Pseidonimizēti dati joprojām ir personas dati.

Kamēr pasaulē eksistē atslēga vai papildu informācija, kas ļauj datus atšifrēt, VDAR (GDPR) prasības ir pilnībā piemērojamas.

Kāpēc vispār pseidonimizēt?

Datu minimizēšana. Lai gan VDAR joprojām darbojas, **pseidonimizācija** ir spēcīgs drošības mehānisms. Tā būtiski samazina riskus datu subjektiem un aizsargā jūsu pētījumu datu noplūdes gadījumā.

Kā tehniski veikt pseidonimizāciju

Pamatmetodes identifikatoru aizstāšanai:

Vienkārša vārdu aizstāšana ar unikāliem ID kodiem (piemēram, “Subjekts 015”).

Tokenizācija

Datu jaukšana (Hashing)

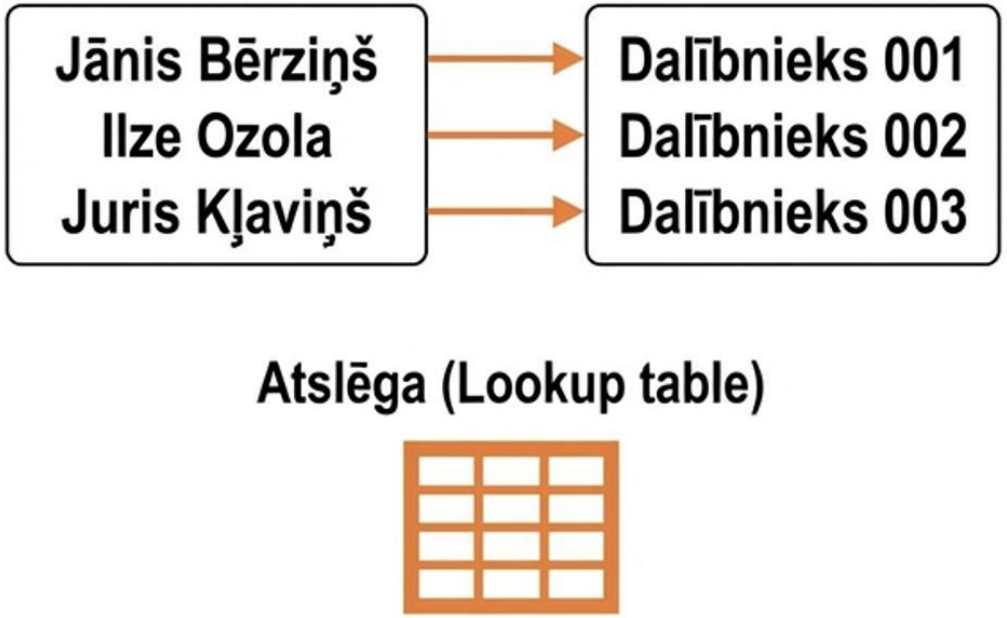
Šifrēšana

Pseidonimizācija ar vienkāršu kodu

Tiešie identifikatori (vārds, uzvārds, personas kods) tiek aizstāti ar vienkāršu burtu un ciparu kombināciju

«Jānis Bērziņš» kļūst par «Dalībnieks 001»

Atslēga ir atsevišķs šifrējumu saraksts jeb atšifrēšanas tabula (Lookup table), kas savieno kodu ar oriģinālo identitāti

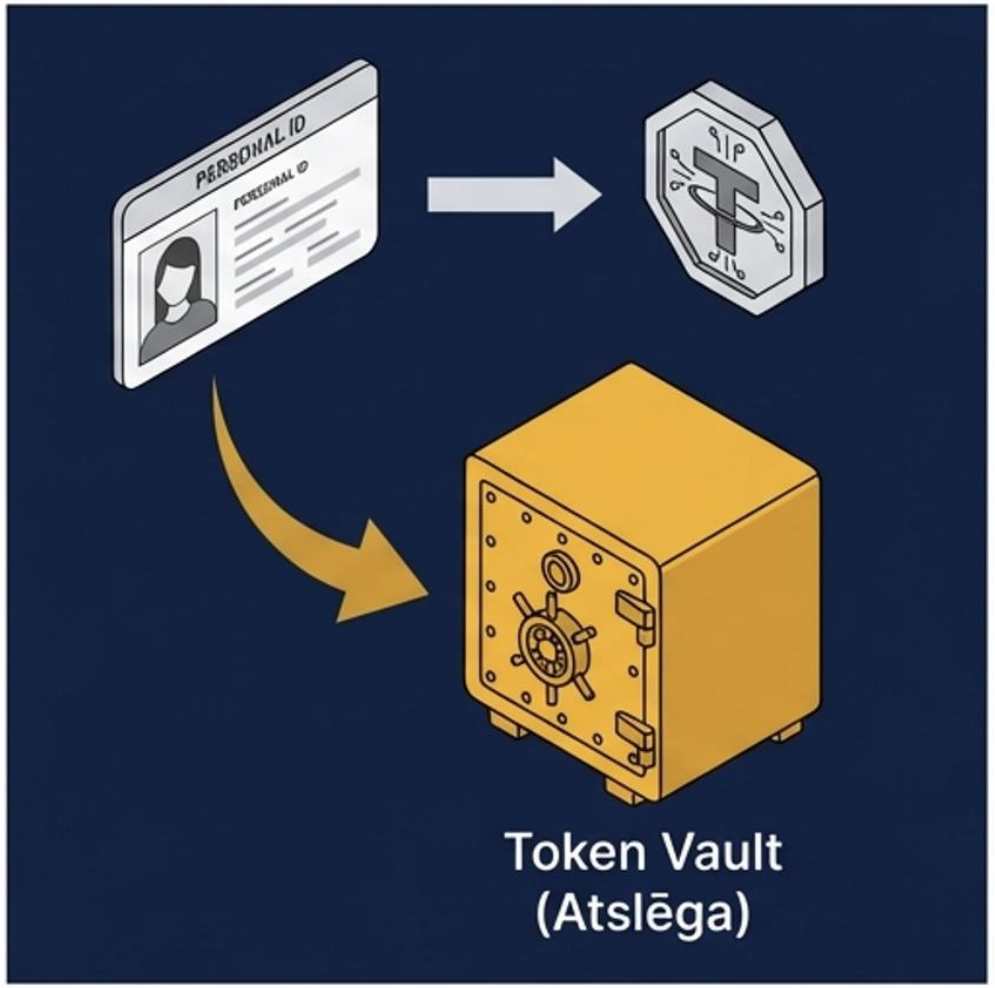


Tokenizācija

Tiešie identifikatori tiek aizstāti ar nejauši ģenerētu vērtību jeb «tokenu» (žetonu), kam pašam par sevi nav nekādas nozīmes vai loģiskas saistības ar datiem

Sistēma/algoritms automātiski katram ierakstam piešķir nejaušu tokenu

Atslēga ir droša ārējā datubāze jeb «token vault» (seifs), kas pārvalda saikni starp tokenu un reālajiem datiem



Jaucējfunkciju izmantošana

Dati tiek apstrādāti ar vienvirziena matemātisku funkciju, radot unikālu digitālo «pirkstu nospiedumu» jeb jaukšanas vērtību (hash), šajā procesā tiek pievienota arī papildus informācija «sāls», kas padara atšifrēšanu grūtāku

Sistēma/algoritms katru ierakstu pārveido, pievienojot «sāli»

Atslēga ir pievienotais slepenais algoritma elements

Visneaizsargātākais posms ir jūsu atslēga.

Pseidonimizācijas drošība ir tieši tik stipra, cik labi jūs sargājat kartēšanas dokumentu jeb "atslēgu".

Zelta likums: Absolūta atdalīšana

Atslēgu un pētniecības datus nedrīkst glabāt kopā. Nekad vienā failā, nekad vienā mapē, un, ideālā gadījumā, fiziski un digitāli atdalītās sistēmās.

Tehniskie un organizatoriskie pasākumi atslēgas aizsardzībai.

- ✓ Glabājiet drošā, šifrētā vidē (nevis uz nešifrētas zibatmiņas vai piezīmju lapiņas).
- ✓ Stingri ierobežojiet piekļuvi: atslēga ir pieejama tikai tiem pētniecības komandas locekļiem, kuriem tas ir absolūti nepieciešams.
- ✓ Pielietojiet institūcijas iekšējos drošības protokolus.

Samazināts risks, saglabāta atbildība.

Pseidonimizācija ir izcils rīks datu minimizēšanai, taču tā neatceļ VDAR. Atcerieties: kods nav anonimizēšana. Aizsargājiet atslēgu par katru cenu.

Kad re-identifikācija ir pētniecībai kritiska

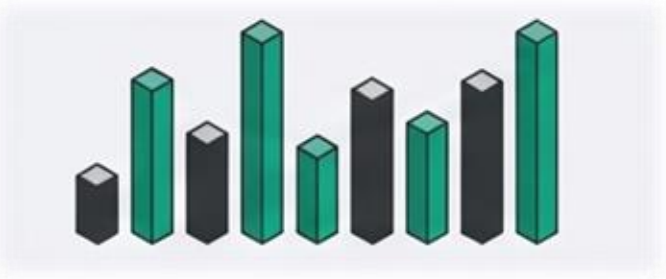
Longitudinālie (garenvirziena) pētījumi: nepieciešamība atkārtoti intervēt to pašu dalībnieku pēc noteikta laika.

Medicīniskie pētījumi un reģistri: nepieciešamība sazināties ar pacientu, ja datos tiek atklātas nejaušas, veselību apdraudošas atradnes.

Anonimizācija

Kuri dati uzskatāmi par anonimizētiem?

Agregēti dati
Vidējie rādītāji un statistika par grupām



Vispārināti dati
Dati, kuros, piemēram, precīza adrese ir aizstāta ar reģionu vai valsti

Pseidonimizācijas atslēga dzēsta- dati anonimizēti?

Pseidonimizācijas atslēgas dzēšana vien nenozīmē, ka dati ir anonimizēti, katra datu kopa ir jāizvērtē atsevišķi

Lai datus uzskatītu par anonīmiem nosaka konkrētus datu raksturlielumus:
K-anonimitāte (k-anonymity)
L-dažādība (l-diversity)
T-tuvums (t-closeness)

Šo lielumu vērtības ir atkarīgas no datu sensitivitātes un potenciāliem riskiem datu subjektiem - katrs gadījums jāvērtē atsevišķi

K-anonimitāte (k-anonymity)

Tas ir lielums, kas raksturo datu kopas privātuma līmeni, nosakot minimālo indivīdu skaitu (k), kurus nav iespējams atšķirt vienu no otra pēc to netiešajiem identifikatoriem (piemēram, vecuma, dzimuma vai dzīvesvietas).

Ko tas nozīmē?

Ja datu kopa ir k-anonīma, jebkura tajā esošā informācija par kādu personu vienmēr sakrīt ar vismaz k-1 citu personu datiem, tādējādi nodrošinot, ka neviens indivīds "neizceļas" un nevar tikt viennozīmīgi identificēts.

Katram ierakstam jābūt neatšķiramam no vismaz k-1 citiem datu kopas ierakstiem.

Slēpšanās pūlī

Ieteicamā vērtība k ≥ 5

Kā noteikt K-anonimitāti?

Datus grupē pēc netiešajiem identifikatoriem
Nosaka, cik ierakstu atbilst konkrētai netiešo identifikatoru kombinācijai

	Netiešie identifikatori		Mērķa atribūts
	Postal Code	Age	Favorite Author
K = 2	22xxxx	21 ... 30	William Shakespeare
	22xxxx	21 ... 30	Agatha Christie
K = 4	54xxxx	41 ... 50	René Goscinny
	54xxxx	41 ... 50	David Baldacci
	54xxxx	41 ... 50	William Shakespeare
	54xxxx	41 ... 50	George Orwell
K = 3	47xxxx	21 ... 30	Osamu Tezuka
	47xxxx	21 ... 30	George Orwell
	47xxxx	21 ... 30	Terry Pratchett

K-anonimitātes plaisas

Homogenitāte:
Ja visiem k cilvēkiem grupā ir viena un tā pati slimība, pūlis vairs neslēpj noslēpumu

Fona zināšanas:
Ja zināms, ka konkrētais pētījums veikts lokācijā, kur konkrētā pazīme ir reta, re-identifikācijas risks pieaug

Dzimums	Vecums	Diagnoze
Vīrietis	50-60	Sirds slimība
Vīrietis	50-60	Sirds slimība
Vīrietis	50-60	Sirds slimība
Vīrietis	50-60	Sirds slimība

Pat ja K=4, visi šajā grupā cieš no vienas un tās pašas slimības, kas ļauj to secināt.

L-dažādība (l-diversity)

Tas ir lielums, kas raksturo to, cik dažādas un daudzveidīgas ir jutīgās vērtības (piemēram, diagnozes vai ienākumu līmeņi) katrā anonimizētajā grupā.

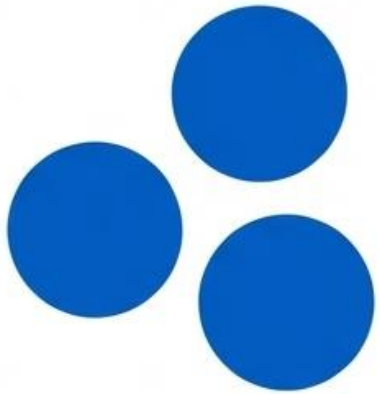
Pasta indekss	Dzimums	Vecums	Diagnoze
LV-10xx	vīrietis	30-40	Plaušu vēzis
LV-10xx	vīrietis	30-40	Hipertensija
LV-10xx	vīrietis	30-40	2. Tipa diabēts

Ko tas nozīmē?

Ja l=3, tad katrā grupā (piemēram, "Vīrieši, 30–40 gadi") ir jābūt vismaz **3 dažādām** diagnozēm. Šādi uzminēt pareizo diagnozi iespēja ir tikai 1 no 3, pat ja kāds ir drošs par personas piederību konkrētajai grupai.

K- anonimitātes un L-dažādības kombinēšana

K-anonimitāte:
Grupai jābūt pietiekami lielai (k).



L-dažādība:
Grupas iekšienē sensitīvajai informācijai jābūt dažādai – vismaz L atšķirīgām, labi pārstāvētām vērtībām.

minimālā vērtība l ≥ 2, atkarībā no datu jutīguma l ≥ 5

L-dažādības ierobežojumi

Iedomāsimies grupu, kur L=3:

- Kuņģa vēzis
- Plaušu vēzis
- Asins vēzis

Pasta indekss	Dzimums	Vecums	Diagnoze
LV-10xx	vīrietis	30-40	Kunģa vēzis
LV-10xx	vīrietis	30-40	Plaušu vēzis
LV-10xx	vīrietis	30-40	Limfoma

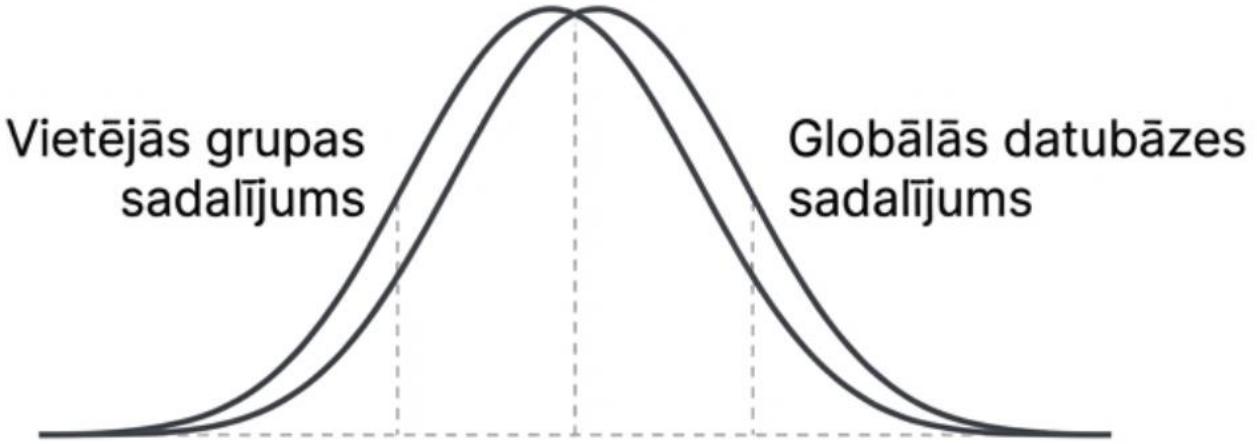
Tehniski L-dažādības nosacījums ir izpildīts, dotajā grupā ir 3 dažādas vērības (3 dažādas diagnozes). Taču tās visas ir smagas onkoloģijas saslimšanas.

Ja kāds zina, ka persona ir šajā grupā, **šī persona atklāj kritisku, sensitīvu informāciju**.

Tāpēc nākamais solis, lai novērstu re-identifikāciju ir **t-tuvums** (t-closeness).

T-tuvums (t-closeness)

Tas ir lielums, kas raksturo to, cik tuvu katras anonimizētās grupas jutīgo datu sadalījums atbilst visas kopējās datu kopas sadalījumam.



Ko tas nozīmē?

Ja slimnīcā veic pētījumu, kur no visiem pacientiem tikai 1% ir noteikta diagnoze, tad pie ideālā t-tuvuma nevienā mazākā apakšgrupā šīs diagnozes īpatsvars nedrīkstētu būtiski pārsniegt šo 1%.

T vērtība norāda pieļaujamo "attālumu" (atšķirību) starp jutīgā atribūta sadalījumu vienā anonimizētā grupā un tā sadalījumu visā sākotnējā datu kopā.

T-tuvuma ierobežojumi un prakse

Informācijas zudums

Jo mazāku t-tuvumu ir mērķis sasniegt, jo vairāk jāpārveido dati, zaudējot analītisko vērtību

Sarežģītība

Lielām datu kopām to t-tuvumu var modificēt tikai ar specifiskiem algoritmiem

T-tuvuma vērtības izvēle

Nav universāla "pareizā" t vērtība - atkarīga no datu jutīguma un konteksta
$t \leq 0.15$ nodrošina augstu aizsardzības līmeni
$0.15 < t \leq 0.25$ nodrošina vidēji augstu aizsardzības līmeni

Tikai t-tuvums neaizsargā pret identitātes atklāšanu

Kombinējot k-anonimitāti + l-dažādību + t-tuvumu, iegūst visaptverošu aizsardzību

Kas jāņem vērā?

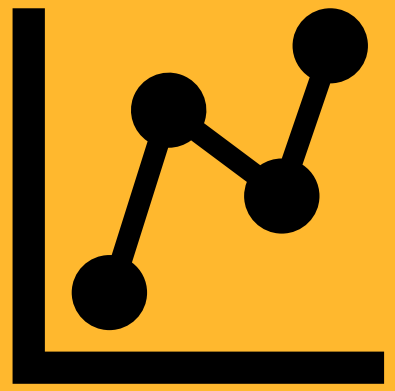
Identificē netiešos identifikatorus un jutīgos atribūtus

Noskaidro vai dati paredzēti iekšējai lietošanai vai ārējai koplietošanai

Sākotnējie mērķi k ≥ 5, l ≥ 2, t ≥ 0.15

Analizē riskus un mērogu, cenšoties rast līdzsvaru starp privātumu un datu lietderību

Datu apstrāde pirms deponēšanas



Anonimizācijas metodes

Mikroagregācija

- aizsargā pret tiešu atpazīšanu, taču iznīcina invidu̇ālo mainību

Vispārināšana

- vienkārši saprotama un ieviešama, taču zūd detaļu precizitāte datu analīzē

Trokšņa pievienošana

- saglabā datu kopējo sadalījumu, taču atsevišķi ieraksti zaudē uzticamību

un citas metodes ...

Mikroagregācija

Ieraksti tiek grupēti mazās kopās un katra indivīda konkrētā vērtība tiek aizstāta ar visas grupas vidējo rādītāju vai skaitlisko intervālu.

Pirms

ID	Vecums	Ienākumi
1	25	30 000
2	27	32 000
3	29	34 000
4	41	50 000
5	43	60 000
6	45	70 000

→

Pēc

ID	Vecums	Ienākumi
1	**27**	**32 000**
2	**27**	**32 000**
3	**27**	**32 000**
4	**43**	**60 000**
5	**43**	**60 000**
6	**43**	**60 000**

Vērtības aizstātas ar 3 cilvēku grupas vidējo rādītāju.

Vispārināšana

Datu precizitātes apzināta samazināšana, kur specifiskas vērtības tiek aizstātas ar plašākām kategorijām, intervāliem vai saīsinājumiem.
Piemērots, ja analīzei nav nepieciešama mikro līmeņa granularitāte.

Pirms

ID	Pasta indekss	Vecums
1	LV-1050	34
2	LV-1052	37
3	LV-1120	42
4	LV-1121	48

→

Pēc

ID	Pasta indekss	Vecums
1	LV-10**	30-40
2	LV-10**	30-40
3	LV-11**	40-50
4	LV-11**	40-50

Precīzi dati pārvērsti par reģioniem un vecuma desmitgadēm.

Trokšņa pievienošana

Oriģinālajām vērtībām tiek pievienotas nejaušas, matemātiski kalibrētas novirzes. Atsevišķi ieraksti vairs nav precīzi, taču visas datubāzes kopējās statistiskās īpašības saglabājas nemainīgas.

Pirms

ID	Alga
1	5 000
2	4 200
3	6 100
Kopā	**15 300**

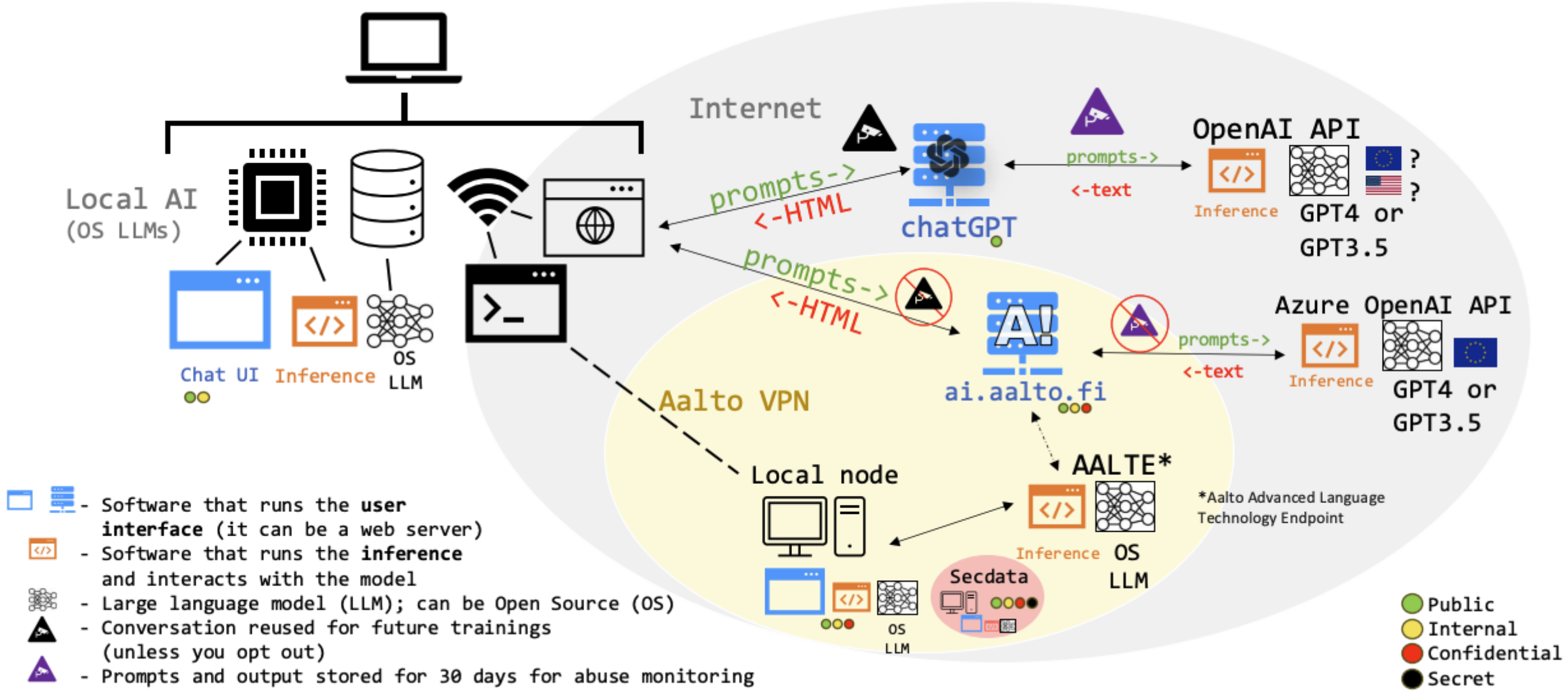
→

Pēc

ID	Alga
1	5 032
2	4 181
3	6 115
Kopā	**15 328**

Nelielas, nejaušas izmaiņas katrā šūnā neļauj atpazīt indivīdu, bet kopsumma gandrīz nemainās.

MI lietojums un datu drošums



Drošs MI pielietojums

Lielisks palīgs skriptu un algoritmu izveidē, lai datus:

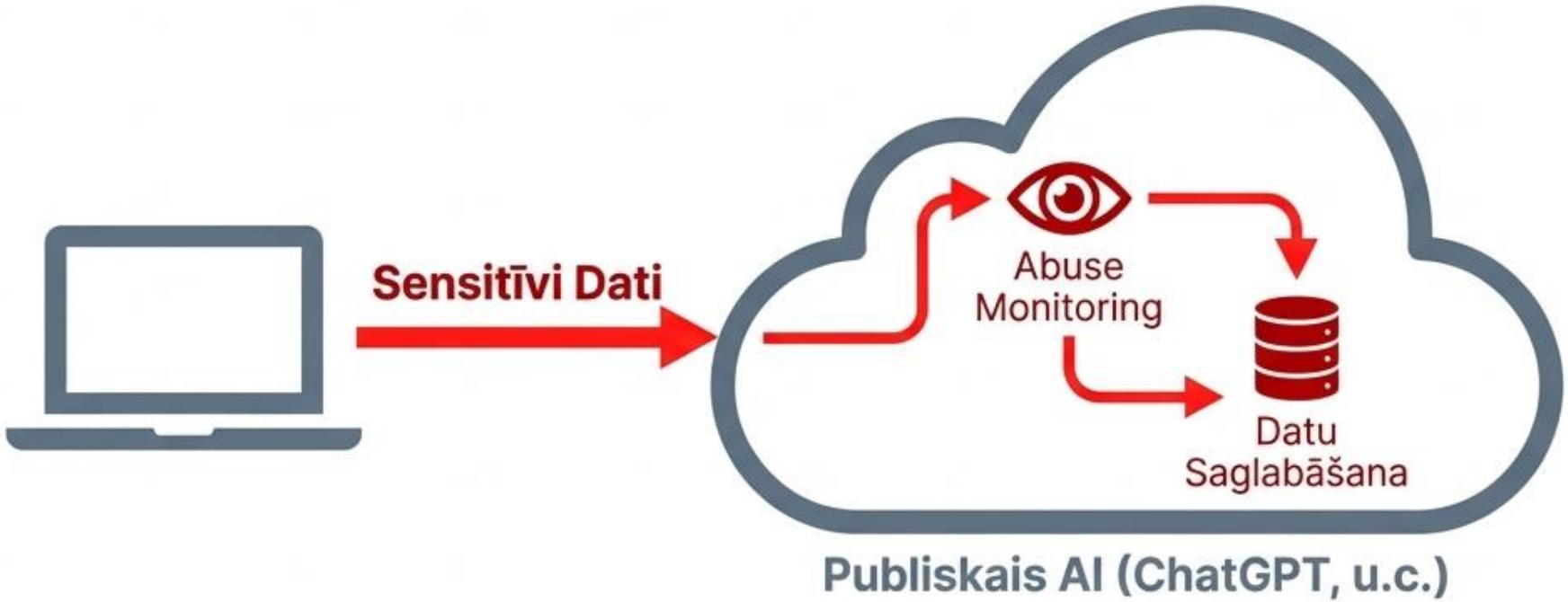
analizētu
pseidonimizētu
anonimizētu
noteiktu anonimizācijas raksturlielumus

Atkarībā no vaicājuma (promta) veidošanas pielāgo atbildes gan R, gan python, gan citu rīku lietotājiem

Nekādā gadījumā nedrīkst MI rīkos augšupielādēt personas datus. Šis ir sevišķi attiecināms uz tiešsaistes MI rīkiem.

Vienīgais izņēmums- universitātes nodrošināts MI rīks, kur katrā solī iespējams izsekot kam dati ir pieejami un kur tie glabājas.

BRĪDINĀJUMS: Publiskie tērzēšanas roboti nav droši



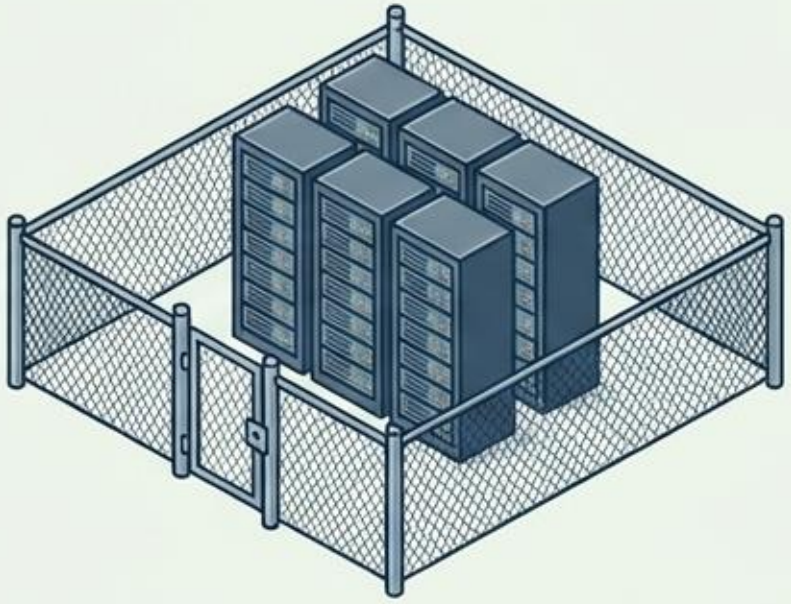
- **Nekad nekopējiet sensitīvus kvalitatīvos datus** (intervijas, medicīnas datus) platformās kā ChatGPT, Claude vai Gemini.
- **Kāpēc?** Dati atstāj drošu jurisdikciju. Tie var tikt saglabāti līdz 30 dienām ļaunprātīgai izmantošanai un potenciāli var tikt izmantoti jaunu modeļu apmācībai.

Drošas alternatīvas anonimizācijai

Lokālie modeļi (Local Inference)

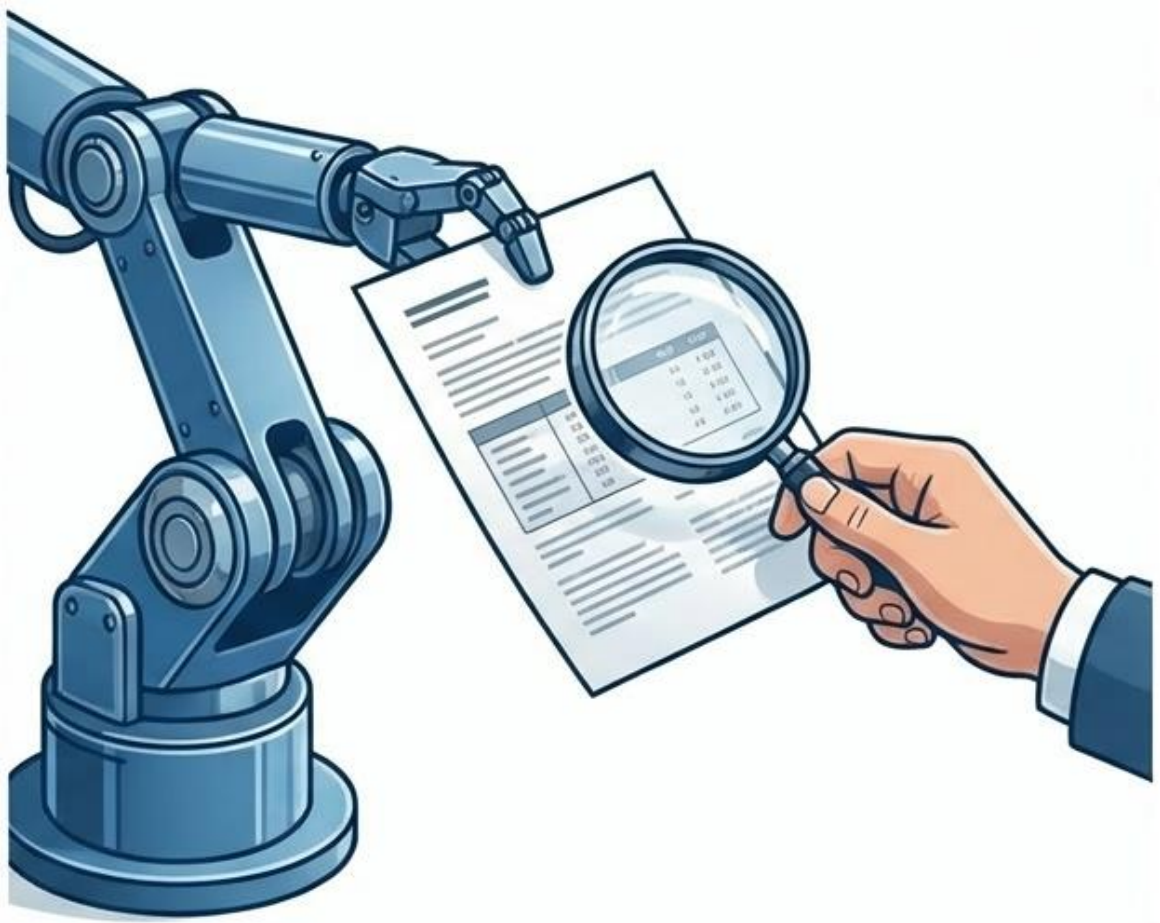
Lejupielādējiet atvērtā pirmkoda modeļus (piemēram, Mistral vai Llama) **drošā datorā**. Jums ir pilnīga kontrole pār saviem datiem.

Droši institucionālie tīkli



Izmantojiet universitātes iekšējos, **slēgtos augstas veiktspējas skaitļošanas (HPC) klasterus** (piemēram, Aalto Triton).

Obligātā cilvēka uzraudzība (Human-in-the-Loop)



AI ir rīks, nevis aizstājējs: AI neatbrīvo pētnieku no ētiskās un tiesiskās atbildības.

Kļūdu līmenis: NER automatizētajai anonimizācijai joprojām ir aptuveni ~1% de-anonimizācijas kļūdu līmenis.

Halucinācijas: Ģeneratīvais AI var radīt neprecizitātes.

Secinājums: Manuāla kvalitātes kontrole un katra rezultāta pārbaude ir absolūti obligāta.

Droša MI darbplūsma pētniecībā

1	**Novērtējiet datus:** Identificējiet sensitīvo informāciju un datu aizsardzības prasības.
2	**Izvēlieties drošu vidi:** Izmantojiet tikai lokālus MI modeļus vai universitātes apstiprinātus, slēgtus serverus.
3	**Pielietojiet MI:** Izmantojiet NER teksta rediģēšanai vai ģeneratīvo MI sintētisko datu izveidei.
4	**Verificējiet rezultātus:** Vienmēr veiciet manuālu datu kvalitātes un privātuma pārbaudi pirms tālākas analīzes vai publicēšanas.

Galvenās atziņas

Datu minimizācija

Vāc un apstrādā tikai tos datus, kas tiešām nepieciešami pētījuma mērķa sasniegšanai

Pseidonimizācija ≠ anonimizācija

- Pseidonimizēti dati joprojām ir personas dati (piemērojams VDAR)
- Anonimizēti dati vairs nav personas dati, bet zaudē daļu informācijas vērtības

Metodes izvēle atkarīga no:

- Pētījuma mērķa un datu analīzes vajadzībām
- Re-identifikācijas riska līmeņa
- Datu tālākizmantošanas plāniem

MI rīku lietošana

Nepārsūti identificējamus personas datus uz ārējiem MI servisiem; izmanto lokālus vai institucionālus risinājumus

Atceries!

Aizsardzības metožu izvēle ir līdzsvars starp privātuma aizsardzību un datu lietderību pētniecībai

Paldies par uzmanību!

Materiāls izstrādāts projekta «Atbalsts atvērtās zinātnes ieviešanai praksē, kā arī izveidoti risinājumi zinātnes datu koplietošanai un dalībai ES atvērtajā zinātnes mākonī» ietvaros (ANM projekta Nr. 2.1.3.1.i) ar Eiropas Savienības Atveseļošanas fonda un Latvijas valsts finansiālo atbalstu

Latvia University of Life Sciences and Technologies

1862 RIGA TECHNICAL UNIVERSITY

UNIVERSITY OF LATVIA

RĪGA STRADIŅŠ UNIVERSITY

Funded by the European Union NextGenerationEU

2027 National Development Plan

VPC