

INF105 - 23
DB i kontekst

David Grellscheid

Quizene

- **Quiz 4** - i gruppetimen **torsdag 12.11. og mandag 16.11.**
- catchup Quiz 3: mandag 16.11., kl 16:15 i grupperommet
- catchup Quiz 4: mandag 23.11., kl.16:15 i grupperommet

Prosjekt + Milestone 2

- Prosjektgruppene og Milestone 2 legges ut på torsdag
- Milestone 2 handler om å lage en plan, krever ingen programmering.
- Skal leveres 18.11. og du får tilbakemelding fra oss noen dager etter.
- Det er tid til å jobbe med dette
  i neste gruppetime og i forelesningstiden 17.11.
- om du ikke har fylt ut *Forming a team for a Final Project* på mitt.uib,
  gjør det nå!

Tidsplan

- de **siste obligatoriske gruppene** er Gruppetime 12, 12.11. / 16.11. nå på torsdag og mandag
- etter det er vi der i de vanlige timene for å hjelpe med prosjektet
- de siste forelesningene skjer
  - i morgen
  - neste uke på onsdag 18.11. har vi en siste avsluttende forelesning. Jeg håper at de fleste kan komme fysisk da.

Tidsplan

alle videre gruppe- og forelesningstimene til 30.11. i kalenderen er frivillig drop-in

åpen til alle, kom tidlig!

- onsdag 11.11.:
  - Forelesning 24
- torsdag 12.11.:
  - Gruppetime - A
    (Quiz 4 + Milestone 2 prep)
- mandag 16.11.:
  - Gruppetime - C+B
    (Quiz 4 + Milestone 2 prep)
  - Quiz 3 retake, kl.16:15
- tirsdag 17.1
  - ingen vanlig forelesning, kan få hjelp med Milestone 2 i forelesningssalen 12:15-14:00
- onsdag 18.11.:
  - Milestone 2 levering
  - Forelesning 25, den siste, håper at de fleste kan komme fysisk
- mandag 23.11.:
  - Quiz 4 retake, kl.16:15

Feedback om inf105-nettside

- Karl-Fredrik som har laget INF105-sidene sammen med Nickolay ønsker feedback

- https://form.menti.com/s/aly2w1g8v1a7



Data Quality

- Data quality (DQ) = ‘fitness for use’
  - akseptabel kvalitet in én kontekst kan ses som dårlig kvalitet i en annen
  - DQ bestemmer verdien som data har til driften
  - GIGO: Garbage In, Garbage Out
  - f.eks gamle adresser

Data Quality

- En **DQ framework** kategoriserer ulike dimensjoner til DQ
- f.eks: Wang et al. (1996)
  - 4 categories:
    intrinsic,
    contextual,
    representation,
    access

Data Quality

Category	DQ dimensions	Definitions
Intrinsic	Accuracy	The extent to which data is certified, error-free, correct, flawless and reliable
	Objectivity	The extent to which data is unbiased, unprejudiced, based on facts and impartial
	Reputation	The extent to which data is highly regarded in terms of its sources or content

- Accuracy refers to whether the data values stored for an object are the correct values
  - often correlated with other DQ dimensions

Data Quality

Category	DQ dimensions	Definitions
Contextual	Completeness	The extent to which data is not missing and covers the needs of the tasks and is of sufficient breadth and depth of the task at hand
	Appropriate-amount	The extent to which the volume of data is appropriate for the task at hand
	Value-added	The extent to which data is beneficial and provides advantages from its use
	Relevance	The extent to which data is applicable and helpful for the task at hand
	Timeliness	The extent to which data is sufficiently up-to-date for the task at hand
	Actionable	The extent to which data is ready for use

Data Quality

Category	DQ dimensions	Definitions
Representation	Interpretable	The extent to which data is in appropriate languages, symbols and the definitions are clear
	Easily-understandable	The extent to which data is easily comprehended
	Consistency	The extent to which data is continuously presented in the same format
	Concisely-represented (CR)	The extent to which data is compactly represented, well-presented, well-organized, and well-formatted
	Alignment	The extent to which data is reconcilable (compatible)

Data Quality

Category	DQ dimensions	Definitions
Access	Accessibility	The extent to which data is available, or easily and swiftly retrievable
	Security	The extent to which access to data is restricted appropriately to maintain its security
	Traceability	The extent to which data is traceable to the source

- accessibility <-> security

Data Quality

- Common causes of bad data quality are:
  - **multiple data sources**: multiple sources with the same data may produce duplicates; a problem of *consistency*.
  - **subjective judgment** in data production: data production using human judgment can result in biased information; a problem of *objectivity*.
  - **limited computing resources**: lack of sufficient computing resources may limit the accessibility of relevant data; a problem of *accessibility*.
  - **volume of data:** Large volumes of stored data make it difficult to access needed information in a reasonable time; a problem of *accessibility*.
  - **changing data needs**: data requirements change on an ongoing basis; a problem of *relevance*.
  - **different processes** updating the same data; a problem of *consistency*.

Testdata

dummy data

- kan ikke bruke hva som helst testdata til f.eks
  - ip-adresser
  - telefonnummer
  - bankkonto
  - personnummer
- de er muligens i bruk!

dummy data

- https://nkom.no/telefoni-og-telefonnummer/telefonnummer-og-den-norske-nummerplan/alle-nummerserier-for-norske-telefonnumre

Nummer for TV- og filmproduksjon

Nummerserien for TV- og filmproduksjon kan brukes som fiktive nummer i film- og TV-produksjon. Numrene i serien skal ikke brukes til ordinær bruk.

Nummerserie for TV- og filmproduksjon:

Fra	Til	Antall	Status	Kategori
68 05 00 00	68 06 99 99	10 000	BLOKKERT	TV- og filmproduksjon

dummy data

- https://www.skatteetaten.no/skjema/testdata/

Tenor testdata

Tenor testdatasøk er et arbeidsverktøy for alle som driver systemutvikling og som trenger testdata. Løsningen gir deg enkel tilgang til testdata fra Test-Norge, med testdata fra fler ulike kilder.

Formålet med søkeløsningen er at du skal kunne finne syntetiske testdata på tvers av testmiljøer i Norge.

dummy data


UMOTIVERT TIGER
02866599001
freg
DEILIG DRAGE
16910297256
freg
LEDENDE APE
07876697682
freg
Folkeregisteret (1)
Nøkkelinformasjon
Relasjoner
Kildedata
Gjeldende identifikator: 02866599001
Identifikatorer: 02866599001
Navn: UMOTIVERT TIGER
Fødselsdato: 1965-06-02
Kjønn: kvinne
Personstatus: bosatt
Sivilstand: gift
Adressebeskyttelse: -
Bostedsadresse: Brenneveien 46 H0101 1339 VØYENENGA
Siste hendelse: endringIBostedsadresseUtenFlytting

dummy data

- ip-adresser: 192.0.2.0/24   2001:db8::
- URLs: example.com/.org/.net
- bank: https://www.ecbs.org/iban/norway-bank-account-number.html (?)
- https://decoded.legal/blog/2021/01/alternatives-to-using-genuine-phone-numbers-ip-addresses-domain-names-and-email-addresses-in-your-training-and-documentation/

Design choices

CHRISTOPHER NULL OPINION NOV 5, 2015 5:26 AM

Hello, I'm Mr. Null. My Name Makes Me Invisible to Computers

My last name is "Null," and it comes preloaded with entertainment value.

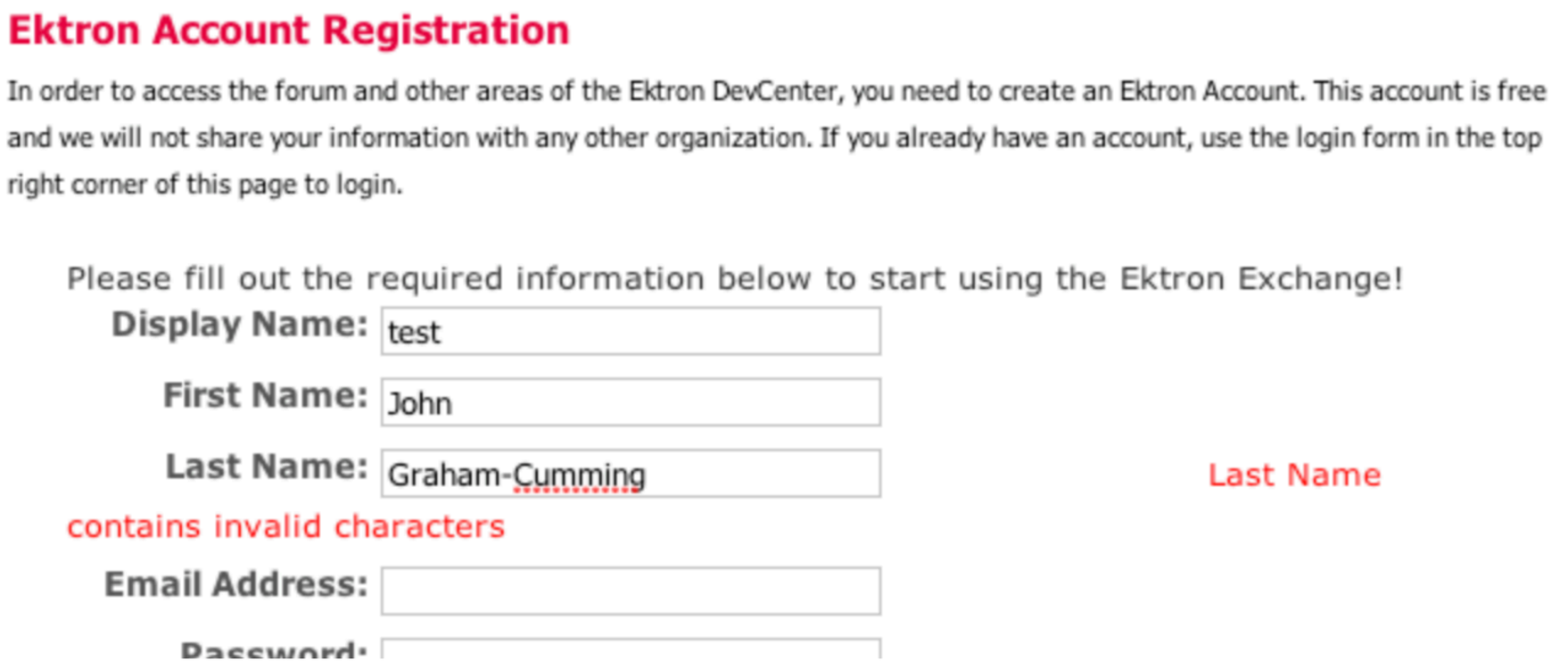
https://archive.is/sQoCh

We have an employee whose last name is Null. (stackoverflow.com)

https://stackoverflow.com/questions/4456438

https://archive.is/XWtty

https://blog.jgc.org/2010/06/your-last-name-contains-invalid.html



https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/

All of these assumptions are wrong. Try to make less of them next time you write a system which touches names.

1. People have exactly one canonical full name.
2. People have exactly one full name which they go by.
3. People have, at this point in time, exactly one canonical full name.
4. People have, at this point in time, one full name which they go by.
5. People have exactly N names, for any value of N.
6. People's names fit within a certain defined amount of space.
7. People's names do not change.
8. People's names change, but only at a certain enumerated set of events.
9. People's names are written in ASCII.
10. People's names are written in any single character set.

All of these assumptions are wrong. Try to make less of them next time you write a system which touches names.

11. People's names are all mapped in Unicode code points.
12. People's names are case sensitive.
13. People's names are case insensitive.
14. People's names sometimes have prefixes or suffixes, but you can safely ignore those.
15. People's names do not contain numbers.
16. People's names are not written in ALL CAPS.
17. People's names are not written in all lower case letters.
18. People's names have an order to them. Picking any ordering scheme will automatically result in consistent ordering among all systems, as long as both use the same ordering scheme for the same name.
19. People's first names and last names are, by necessity, different.
20. People have last names, family names, or anything else which is shared by folks recognized as their relatives.

All of these assumptions are wrong. Try to make less of them next time you write a system which touches names.

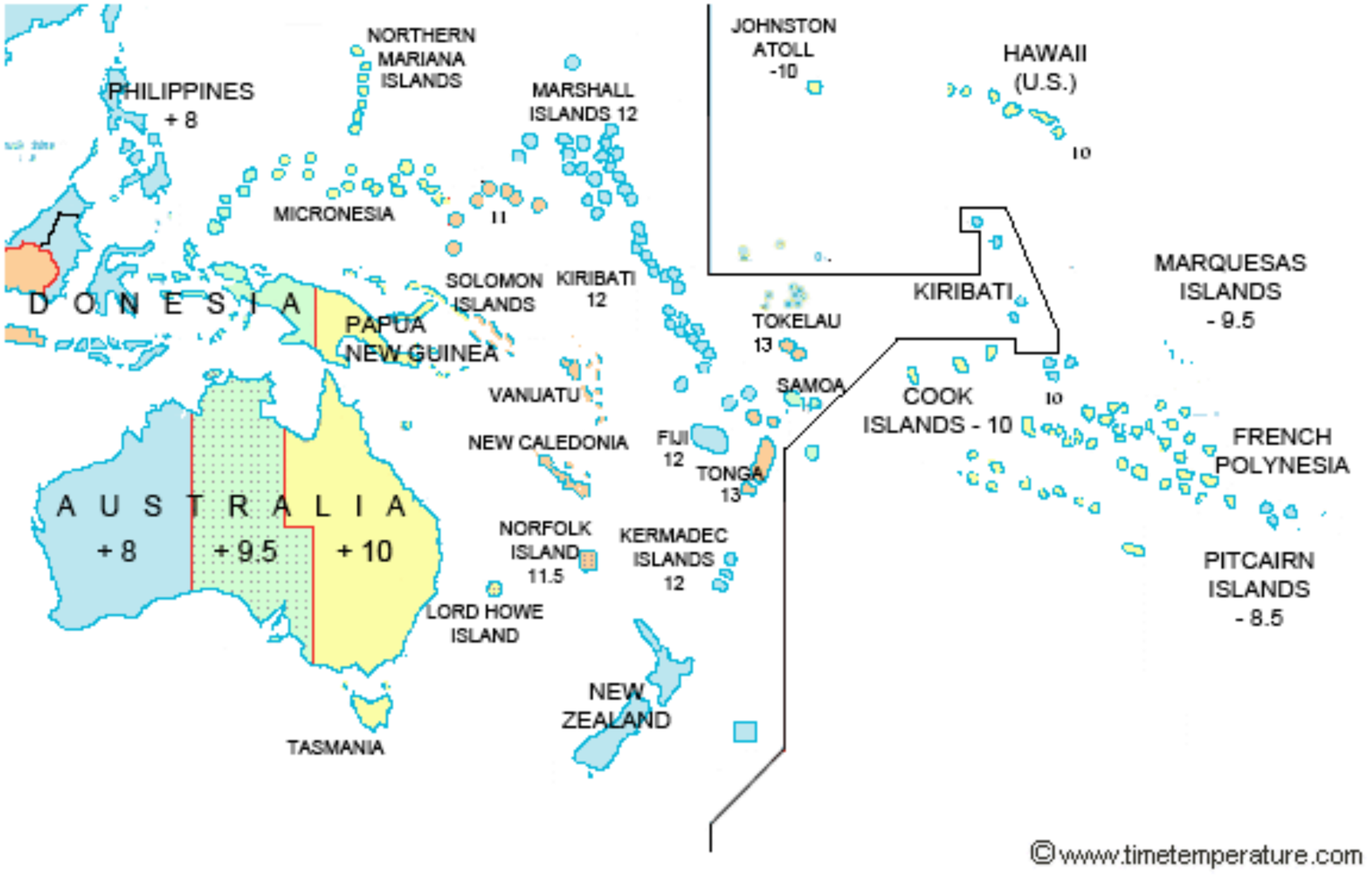
21. People's names are globally unique.
22. People's names are *almost* globally unique.
23. Alright alright but surely people's names are diverse enough such that no million people share the same name.
24. My system will never have to deal with names from China.
25. Or Japan.
26. Or Korea.
27. Or Ireland, the United Kingdom, the United States, Spain, Mexico, Brazil, Peru, Russia, Sweden, Botswana, South Africa, Trinidad, Haiti, France, or the Klingon Empire, all of which have "weird" naming schemes in common use.
28. That Klingon Empire thing was a joke, right?
29. Confound your cultural relativism! People ***in my society***, at least, agree on one commonly accepted standard for names.
30. There exists an algorithm which transforms names and can be reversed losslessly. (Yes, yes, you can do it if your algorithm returns the input. You get a gold star.)

All of these assumptions are wrong. Try to make less of them next time you write a system which touches names.

31. I can safely assume that this dictionary of bad words contains no people's names in it.
32. People's names are assigned at birth.
33. OK, maybe not at birth, but at least pretty close to birth.
34. Alright, alright, within a year or so of birth.
35. Five years?
36. You're kidding me, right?
37. Two different systems containing data about the same person will use the same name for that person.
38. Two different data entry operators, given a person's name, will by necessity enter bitwise equivalent strings on any single system, if the system is well-designed.
39. People whose names break my system are weird outliers. They should have had solid, acceptable names, like 田中太郎.
40. People have names.

https://shinesolutions.com/2018/01/08/falsehoods-programmers-believe-about-names-with-examples/

Tid



https://en.wikipedia.org/wiki/Time_in_Kiribati

UTC+14

https://infiniteundo.com/post/25326999628/falsehoods-programmers-believe-about-time

All of these assumptions are wrong

1. There are always 24 hours in a day.
2. Months have either 30 or 31 days.
3. Years have 365 days.
4. February is always 28 days long.
5. Any 24-hour period will always begin and end in the same day (or week, or month).
6. A week always begins and ends in the same month.
7. A week (or a month) always begins and ends in the same *year.*
8. The machine that a program runs on will always be in the GMT time zone.
9. Ok, that's not true. But at least the time zone in which a program has to run will never change.
10. Well, surely there will never be a change to the time zone in which a program hast to run *in production*.
11. The system clock will always be set to the correct local time.
12. The system clock will always be set to a time that is not wildly different from the correct local time.
13. If the system clock is incorrect, it will at least always be off by a consistent number of seconds.
14. The server clock and the client clock will always be set to the same time.
15. The server clock and the client clock will always be set to *around* the same time.

[...]

Adresser



https://www.mjt.me.uk/posts/falsehoods-programmers-believe-about-addresses/

Genève / Ginevra / Genf / Genevra // Geneva / Ginebra

Mange flere lister:

https://github.com/kdeldycke/awesome-falsehood

Contents

- Meta
- Arts
- Business
- Cryptocurrency
- Dates and Time
- Education
- Emails
- Geography
- Human Identity
- Internationalization
- Management
- Multimedia
- Networks
- Phone Numbers
- Postal Addresses
- Science
- Society
- Software Engineering
- Transportation
- Typography
- Video Games
- Web

Dårlige tekniske valg har konsekvenser

Ektron Account Registration

In order to access the forum and other areas of the Ektron DevCenter, you need to create an Ektron Account. This account is free and we will not share your information with any other organization. If you already have an account, use the login form in the top right corner of this page to login.

Please fill out the required information below to start using the Ektron Exchange!

Display Name: test

First Name: John

Last Name: Graham-Cumming Last Name

contains invalid characters

Email Address:

Password:

Does the web site have any idea how rude it is to claim that my last name contains invalid characters? Clearly not.

What they actually meant is: our web site will not accept that hyphen in your last name. But do they say that? No, of course not.

They decide to shove in my face the claim that there's something wrong with my name.

There's nothing wrong with my name, just as there's nothing wrong with someone whose first name is Jean-Marie, or someone whose last name is O'Reilly.

What is wrong is that way this is being handled. If the system can't cope with non-letters and spaces it needs to say that.

https://blog.jgc.org/2010/06/your-last-name-contains-invalid.html

How about the following error message:

Our system is unable to process last names that contain non-letters, please replace them with spaces.

Don't blame me for having a last name that your system doesn't like, whose fault is that?

Saying "Your last name contains invalid characters" is plain offensive.

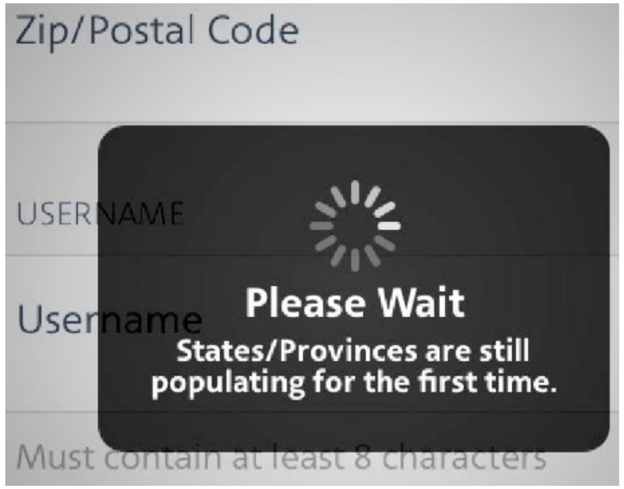
https://blog.jgc.org/2010/06/your-last-name-contains-invalid.html

"Nerdview"

- https://languagelog.ldc.upenn.edu/nll/?cat=106

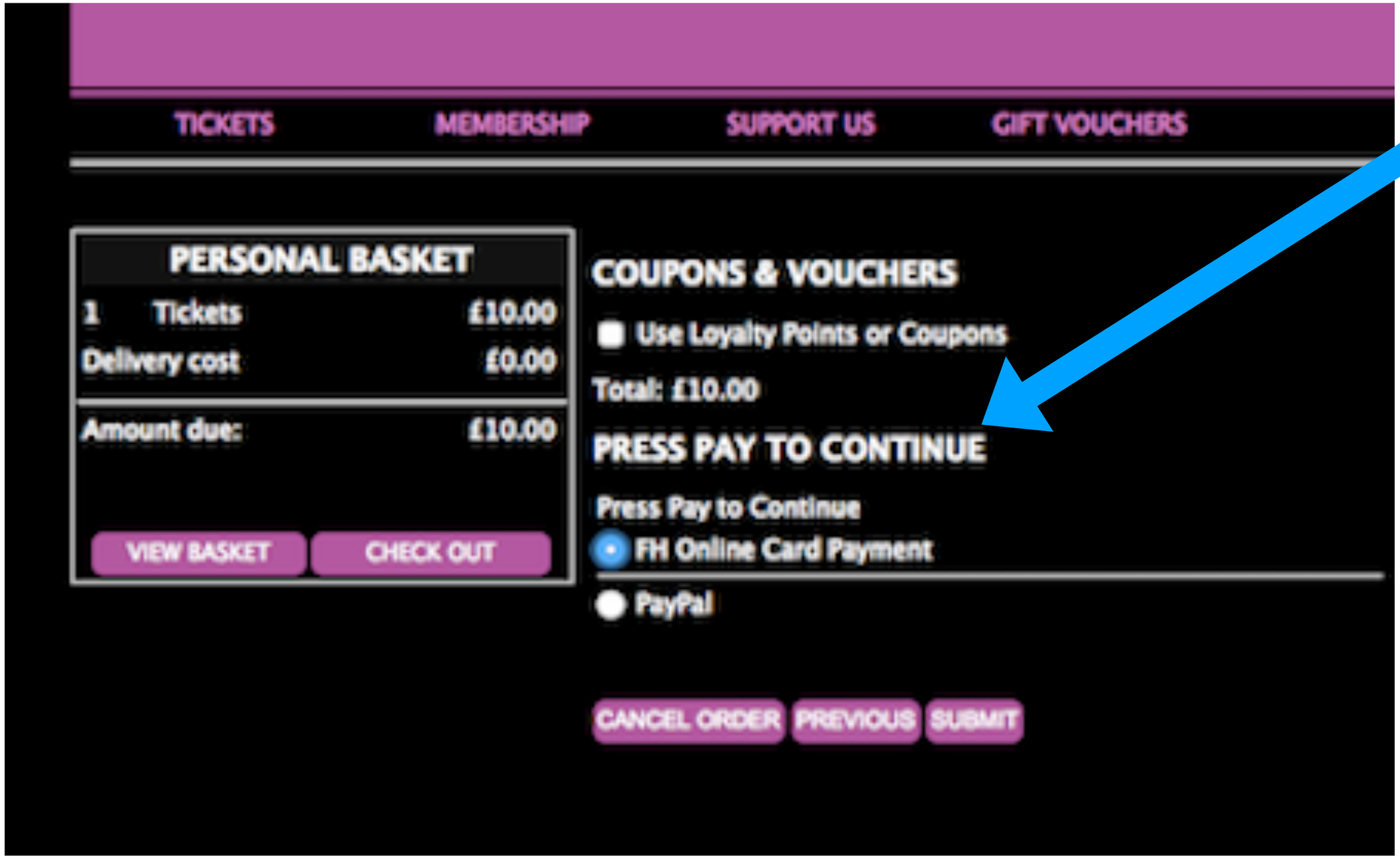
- writing in technical terms from the perspective of the technician or engineer rather than from a standpoint that would seem useful to the customer or reader.

- You may not use the following attribute values for your password:
  - puAccNetID
  - puStaffNo
  - puUserGivenName
  - puUserSurname






Stand B
311 Elizabeth Bay & Millers Point
440 Bronte
M30 Spit Junction
Stand C
338 Clovelly
339 Clovelly
372 Coogee
374 Coogee
376 Maroubra Beach
391 Port Botany
393 Little Bay
395 Maroubra Beach
M10 Maroubra Junction
X93 Little Bay
Stand D
891 UNSW
X10 Eastgardens
X93 Little Bay
Foveaux St
Stand F
338 Railway Square
339 City - Gresham St
374 City - Circular Quay
376 Railway Square
391 Railway Square
892 City - Circular Quay
Chalmers St
Stands G and H
305 Railway Square
308 City - Gresham St
309 Railway Square
310 Railway Square
343 Chatswood
372 Railway Square
393 Railway Square
L23 City - Martin Place
L28 City - Martin Place
L38 City - Martin Place
L39 City - Martin Place
M10 Maroubra Junction
M30 Spit Junction
N10 City - Town Hall
N30 City - Town Hall
N40 City - Town Hall
N50 City - Town Hall
N60 City - Town Hall
N61 City - Town Hall
N70 City - Town Hall
N71 City - Town Hall
N80 City - Town Hall
N100 Bondi Junction


TICKETS
MEMBERSHIP
SUPPORT US
GIFT VOUCHERS
PERSONAL BASKET
1 Tickets £10.00
Delivery cost £0.00
Amount due: £10.00
VIEW BASKET
CHECK OUT
COUPONS & VOUCHERS
Use Loyalty Points or Coupons
Total: £10.00
PRESS PAY TO CONTINUE
Press Pay to Continue
FH Online Card Payment
PayPal
CANCEL ORDER
PREVIOUS
SUBMIT
?

- "we know what users want better than them"

  The spreadsheet software Microsoft Excel, when used with default settings, is known to convert gene names to dates and floating-point numbers. A programmatic scan of leading genomics journals reveals that approximately one-fifth of papers with supplementary Excel gene lists contain erroneous gene name conversions.

SCIENCE

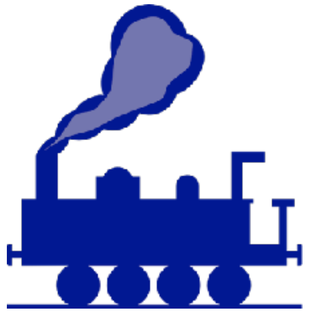
Scientists rename human genes to stop Microsoft Excel from misreading them as dates / Sometimes it's easier to rewrite genetics than update Excel

by James Vincent

Aug 6, 2020, 2:44 PM GMT+2

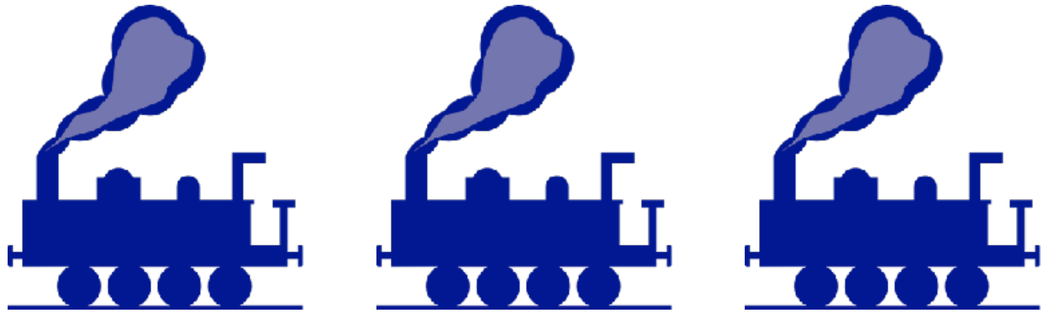
Who is data for?

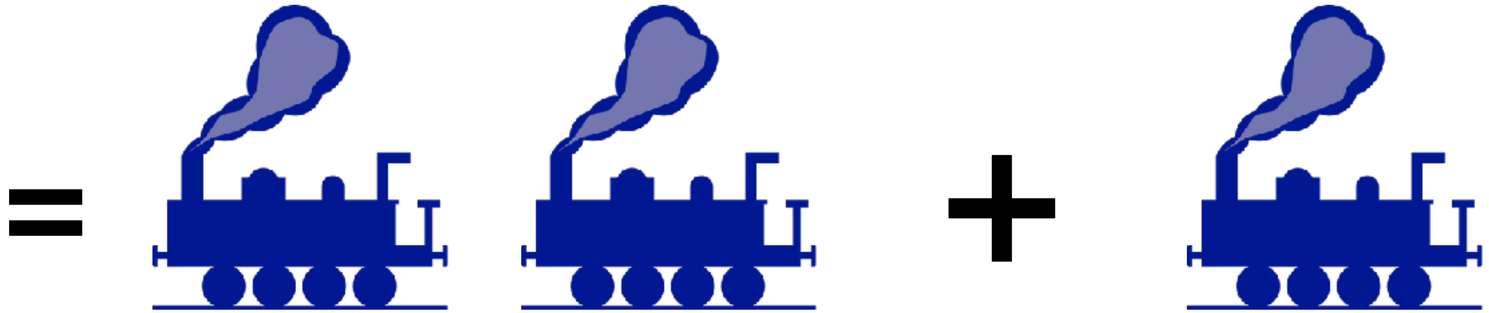
Eksempel: billett

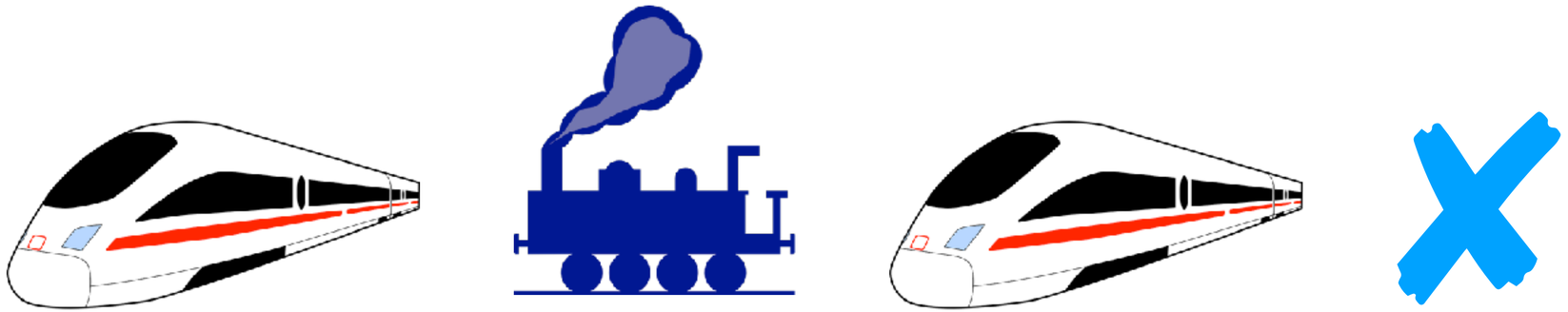


fast pris per km

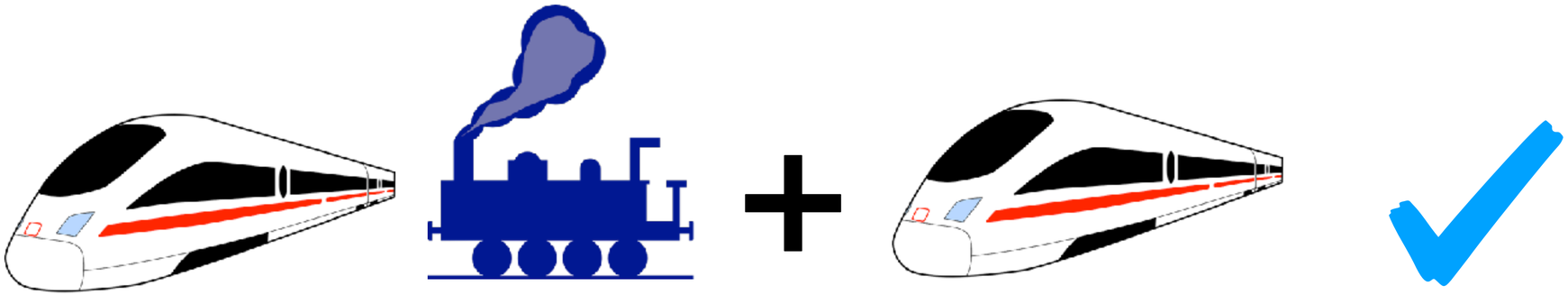
ingen forskjell om man kjøper en lang reise på én eller flere billetter







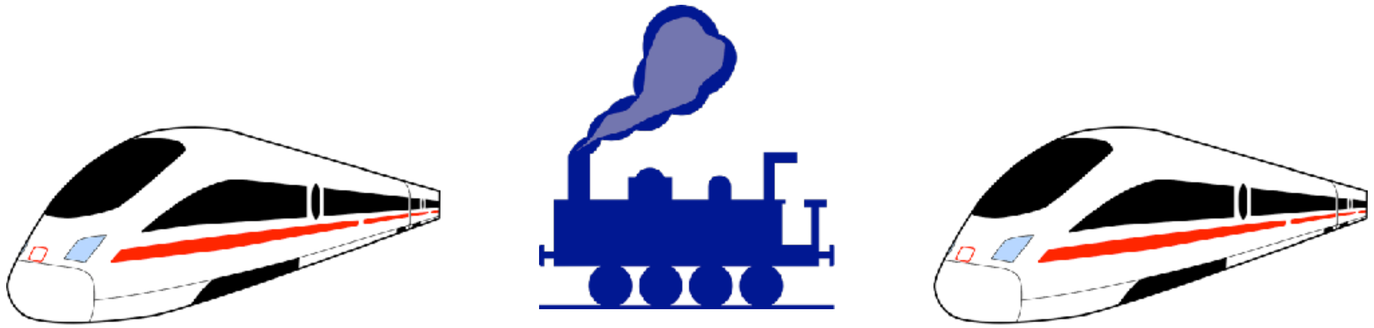
systemet kan ikke selge vanlig togreise i midten av to lyntog på én billett

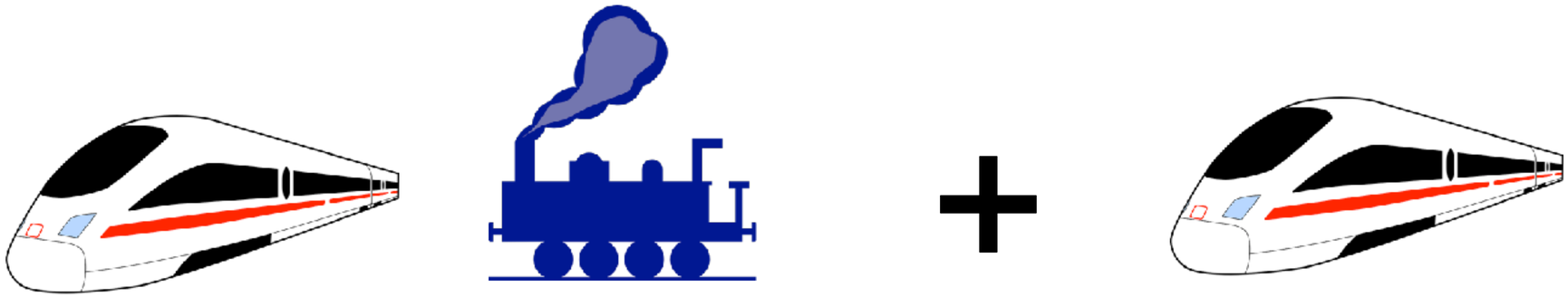


men billettene kan splittes, til samme pris

- **2001 - Nytt prissystem:**
  korte reiser dyrere per km,
  men rabatt jo lenger man reiser
  (senere km blir billigere og billigere)

- Nå blir det prisforskjell når man splitter billett





- Fikse salgssystemet?

- Endring i salgsbetingelser:

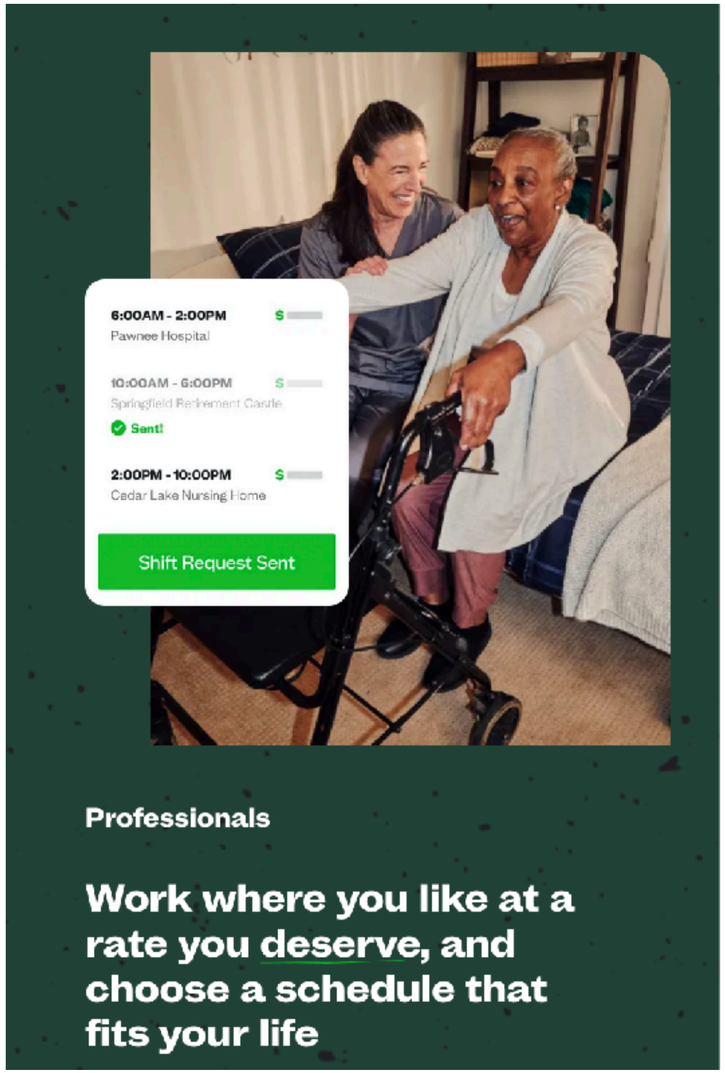
  Reise med lyntog kan ikke selges med vanlig tog i midten

  Det kreves to separate billett

Who is data for?

- Vi mister synet av brukere
- "teknologien" får skylden til hvorfor endringer måtte gjennomføres
- "responsibility dilution" - utvasking av ansvar

"Uber for nurses"



"Before [prod.name] offers a nurse a shift, it purchases that worker's credit history from a data-broker. Specifically, it pays to find out how much credit-card debt the nurse is carrying, and whether it is overdue.

The more desperate the nurse's financial straits are, the lower the wage on offer.

https://pluralistic.net/2025/02/26/ursula-franklin/

"Tech causes more problems than it solves"

- https://www.pewresearch.org/internet/2020/06/30/tech-causes-more-problems-than-it-solves/

- "the value of information"

  Grace Hopper (1982):
  https://youtu.be/si9iqF5uTFk?feature=shared&t=285

  - 4:45 til 8:43

- Anbefaler å lese "Et land på fire hjul" - hvordan en teknologi blir så dominerende at vi ikke ser konsekvensene lenger



GDPR

GDPR

- i norsk lov: Personopplysningsloven
- introdusert i 2018
- implementerer EUs personvernforordning (GDPR)
- https://www.datatilsynet.no/regelverk-og-verktoy/lover-og-regler/om-personopplysningsloven-og-nar-den-gjelder/

Roller

- "***den registrerte***"
  fysisk person

- **behandlingsansvarlig** (BA)
  bestemmer formålet med behandlingen, og midler

- **databehandler**
  behandler opplysninger på vegne av BA

Virkning

- helt/delvis automatisert behandling
  eller manuell behandling i register

- **ikke** når en fysisk person har rent personlige eller familiemessige aktiviteter

- når behandlingsansvarlige i EØS
  eller opplysninger om EØS-borgere knyttet til varer tjenester / deres atferd i EØS

Hva er personopplysning?

- **enhver opplysning om en identifisert eller identifiserbar fysisk person**

- **identifiserbar:**
  som direkte eller indirekte kan identifiseres,
  særlig ved hjelp av en identifikator,
  f.eks. et navn, et identifikasjonsnummer, lokaliseringsopplysninger, en nettidentifikator eller ett eller flere elementer som er spesifikke for nevnte fysiske persons
  fysiske, fysiologiske, genetiske, psykiske, økonomiske, kulturelle eller sosiale identitet

- **Behandling:**
  f.eks. innsamling, registrering, organisering, strukturering, lagring, tilpasning eller endring, gjenfinning, konsultering, bruk, utlevering...

Prinsipper for behandling

- lovlig, rettferdig, åpen
- **spesifikke** angitte formål
  - kan ikke viderebehandles med andre formål
    (unntak: arkivering / forskning / statistikk)
- adekvat, relevant, **begrenset til det som er nødvendig**
- korrekt og oppdatert
- lagringstid begrenset til formålene
  (unntak: arkivering / forskning / statistikk)
- sikret mot uautorisert / ulovlig behandling

BA har ansvar!

Vilkår til behandling

Ett av seks vilkår må oppfylles

- Samtykke
- oppfylle en avtale / registrerte vil inngå en avtale
- BA har rettslig forpliktelse
- verne vitale interesser til registrerte
- oppgave i allmennhetens interesse
- avveining av berettigede interesser

Samtykke

- BA må kunne påvise
- avskilt fra andre forhold
- rett til å trekke tilbake
  Det skal være like enkelt å trekke tilbake som å gi samtykke.
- frivillig
  (ikke frivillig hvis en tjeneste blir nektet som ikke trenger opplysningene)

Norsk tillegg: Fødselsnummer

- Fødselsnummer og andre entydige identifikasjonsmidler kan bare behandles
  - når det er saklig behov for sikker identifisering og
  - metoden er nødvendig for å oppnå slik identifisering

Særlige kategorier

- Behandling av opplysninger om
  - rasemessig eller etnisk opprinnelse
  - politisk oppfatning
  - religion, filosofisk overbevisning
  - fagforeningsmedlemskap
  - genetiske opplysninger,
    biometriske opplysninger med det formålet å entydig identifisere en fysisk person
  - helseopplysninger
  - seksuelle forhold eller seksuelle orientering

er forbudt (med unntak...)

Unntak til behandling av særlige kategorier

Ett av disse må oppfylles

- Uttrykkelig samtykke
- […*BA er selve fagforeningen, rettskrav, medisin, osv.* …]
- åpenbart offentliggjort av registrerte
- arkiv / forskning / statistikk
  skal stå i et rimelig forhold til målet; særlige tiltak som verner registrerte

  samfunnets interesse i at behandlingen finner sted, klart overstiger ulempene for den enkelte

  krever konsekvensvurdering

(EØS-land kan avvike om genetiske / biometriske / helseopplysninger)

Rettigheter til den registrerte

- informasjon
- innsyn
- retting
- sletting
- begrensning
- portabilitet
- protest

Dersom behandlingen ikke (lenger) trenger identifisering, ingen plikt til å bevare info bare for å oppfylle rettighetene

https://www.datatilsynet.no/rettigheter-og-plikter/den-registrertes-rettigheter/

Rettigheter: informasjon

- **kontaktinfo til BA**
- kontaktopplysningene til personvernombudet, dersom dette er relevant,
- **formålene med behandlingen og rettslige grunnlaget for behandlingen**
- dersom basert på interesse:
  de berettigede interessene som forfølges av den behandlingsansvarlige eller en tredjepart,
- **eventuelle mottakere eller kategorier av mottakere av personopplysningene**
- overføring til en tredjestat eller en internasjonal organisasjon
- **tidsrom personopplysningene vil bli lagret**,
  eller kriteriene som brukes for å fastsette dette tidsrommet
- retten om innsyn, retting, sletting, ...
- retten til å trekke samtykke
- forekomsten av automatiserte avgjørelser, herunder profilering

Rettigheter: informasjon

- **Informasjon skal gis før innsamling**

Rettigheter: innsyn

- uformell forespørsel
- gratis
- en kopi eller "vanlig elektronisk form"

Rettigheter: retting / sletting / begrensning

- uformell forespørsel
- **"uten ugrunnet opphold"**
- BA må sikre at alle databehandlere gjør det

- en domstol i belgia ():
  manglende é à æ ø å i navn i database kan anses som uriktig data

Rettigheter til den registrerte

- informasjon
- innsyn
- retting
- sletting
- begrensning
- **portabilitet:**
  få data utlevert i vanlig format for å bytte til en annen tjeneste
- protest

Rettigheter til den registrerte

- informasjon
- innsyn
- retting
- sletting
- begrensning
- portabilitet
- **protest:** mot berettigede interesser / markedsføring

Profilering

- ingen avgjørelser som utelukkende er basert på automatisert behandling som i betydelig grad påvirker vedkommende
- unntak: avtale / samtykke
- rett til menneskelig inngrep fra BA
- til å uttrykke sine synspunkter
- til å bestride avgjørelsen

Plikter til behandlingsansvarlig

- egnede tekniske og organisatoriske tiltak
- løsninger med innebygd personvern som standard
- databehandler:
  - ingen viderefordeling til flere dbh. uten BA
  - må ha avtale
- **behandlingen må protokolleres** (se: informasjon til registrerte)
- krav til melding om brudd

https://www.datatilsynet.no/rettigheter-og-plikter/virksomhetenes-plikter/

Når kameraovervåking vil være i strid med personvernforordningen eller loven her,

er det heller ikke tillatt å benytte uekte kameraovervåkingsutstyr

eller ved skilting, oppslag eller lignende gi inntrykk av at kameraovervåking finner sted

TODO

- Hva er formålene?
- Hvilke personopplysninger trengs?
- Informasjon til registrerte: hvem er BA, rettigheter
- rutiner til databehandling, protokoll

Datatilsynet: virksomhetenes plikter

Datatilsynet: om Cookies