Data & Data analyse

Studiekring Gelderland & Overijssel

ERIC DANKAART

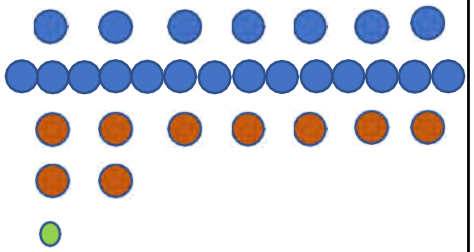
24-11-2022

Even voorstellen

Eric Dankaart

Fiscaal jurist

Loonheffingen inspecteur
Loonheffingen adviseur PwC
Tax Technology PwC
Tax Technology Dankaart BV
Looheffingen Belastingdienst



2

Data & data-analyse

- Data en de opkomst ervan
- Data onderscheiden
- ‘Data-denken’.
- Data-analyse
- Data-analyse, aandachtspunten
- Aan de slag met data
- Trends in data-analyse

Data en de opkomst ervan

Data kan verwijzen naar:

Een verzameling van tijdsaanduidingen, het meervoud van datum; zie Datum (dagtekening)

Data (Star Trek), een androïde uit *Star Trek*

Data Records, een muzieklabel uit Groot-Brittannië

Data (geslacht), een vlindergeslacht

Gegeven, een vastgelegde uitdrukking van een feit: *datum*, meervoud *data* gegevens

Dataset, een gegevensverzameling

De opkomst van data

Digitalisering

Toegenomen rekenkracht computers

Internet & mobiel

Social media

Data onderscheiden

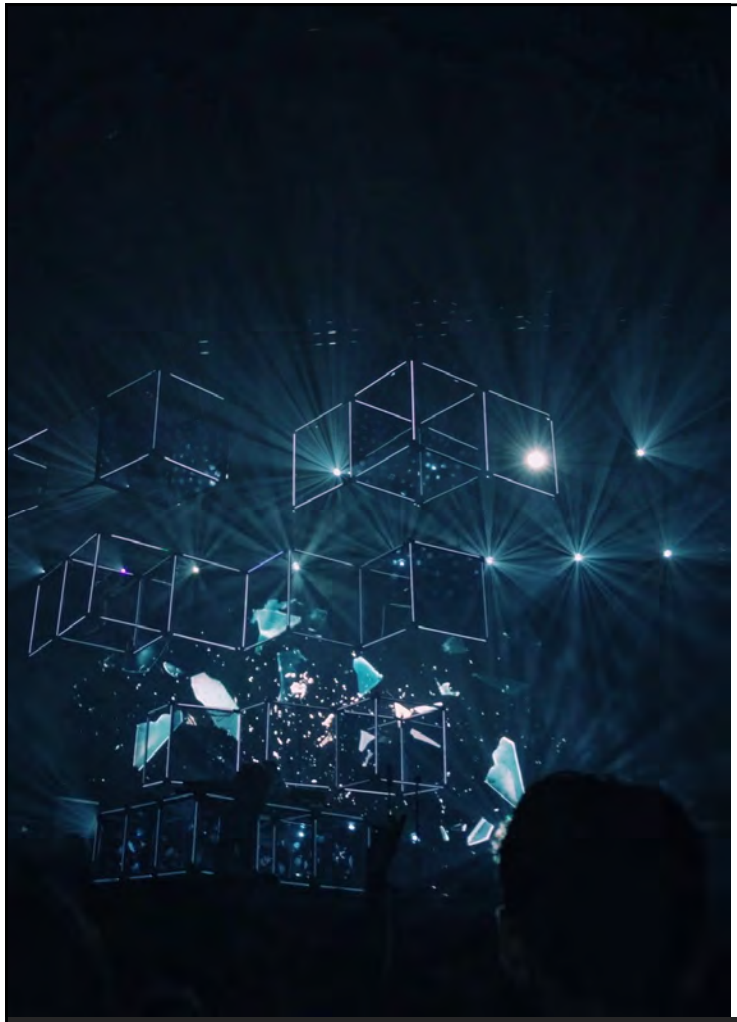
Naar inhoud

Gestructureerd - ongestructureerd

Naar omvang

Naar domein

7



Data naar inhoud

Een verder onderscheid is mogelijk naar bestandstype (fileformat). De meeste formats horen bij een bepaald type inhoud. Sommige formats worden door vele applicaties gebruikt, andere maar door 1 specifieke.

Kwisje! Bij welke applicatie hoort het format en/of welk soort inhoud bevat het?

.xls

MS Excel

.jpg

beeld/foto

.flac

geluid

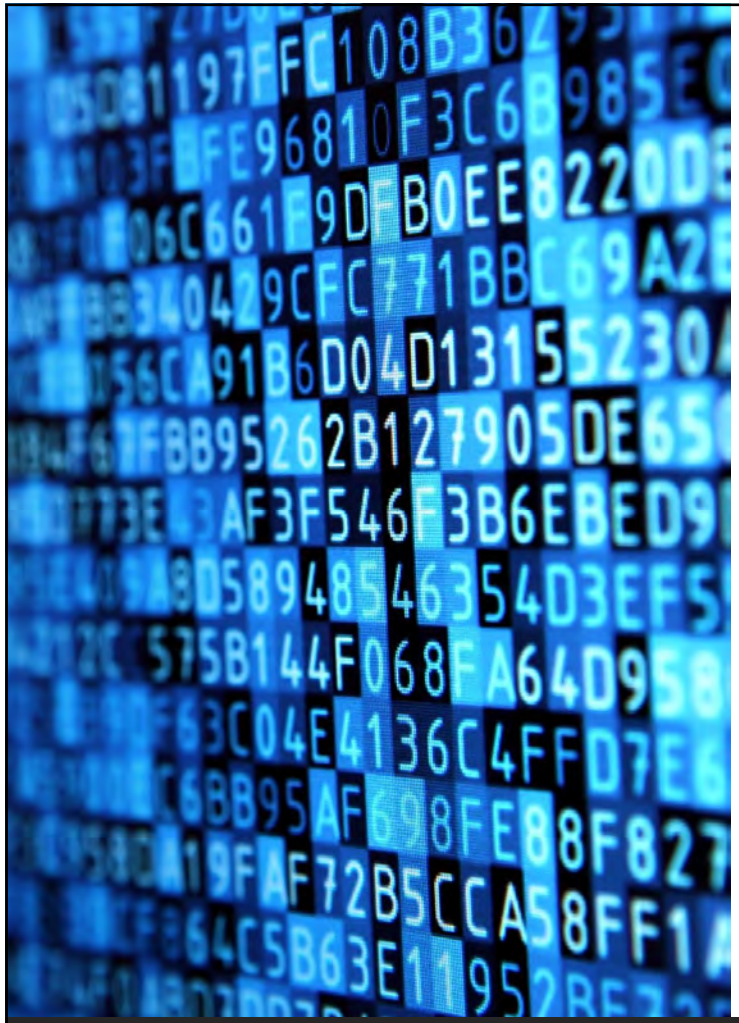
Data, gestructureerd vs. ongestructureerd

Gestructureerd

Noorwegen	56221	Rijst
Zweden	45027	Gerst
Oostenrijk	23992	Aardappel
Nederland	36892	Rijst
Italië	34568	Gerst
IJsland	23245	Rijst
Frankrijk	91102	-
Denemarken	16021	-
Duitsland	45289	Aardappel
België	49921	Rijst

Ongestructureerd

Zweden	12	Nee	28456
Aardappel	Nee	345289	
123	Misschien	Wenen	Gerst
6738	28972	18	Noordzee
Andalusië	1982		Zwart



Data naar omvang

1980: Kb

1990: Mb

2000: Gb

2010: Tb

2020 => Pb, Ex, Zb, Yb...

Data kunnen zeer verschillend in omvang zijn.

Voor de technische mogelijkheden is het van belang te weten hoe groot de databestanden zijn, maar inzicht in fileformat en opbouw van het bestand is voor de analyse belangrijker.

10

Databronnen

Open Data zoals CBS, data.overheid.nl

Commerciële data, company.info, bureau van Dijk

Cliëntendata

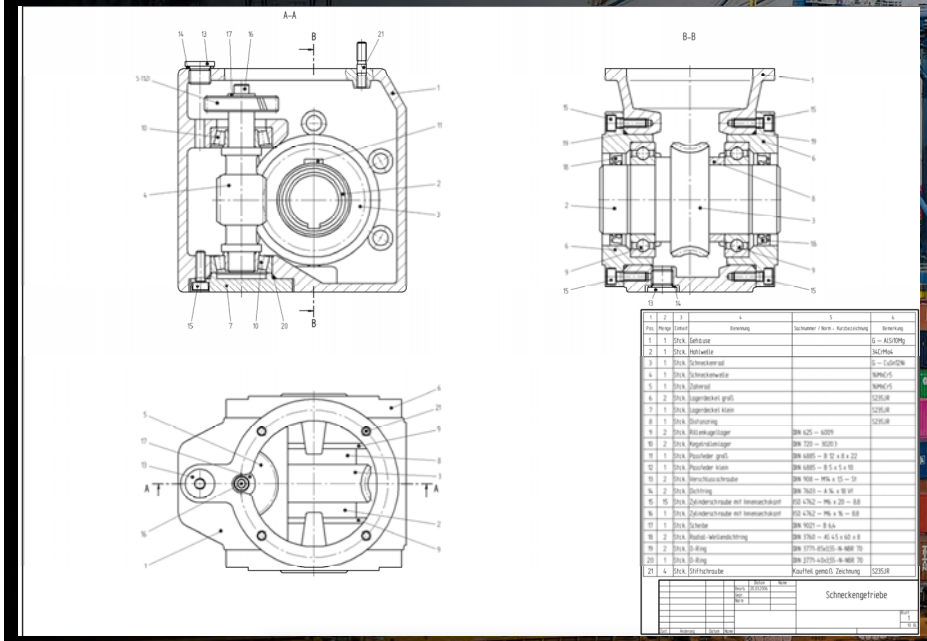
Zelf verkregen data; sensordata, internet

11

12

Logistieke data

HS-codes



14

Beroemd ‘Big data’ voorbeeld

15

16

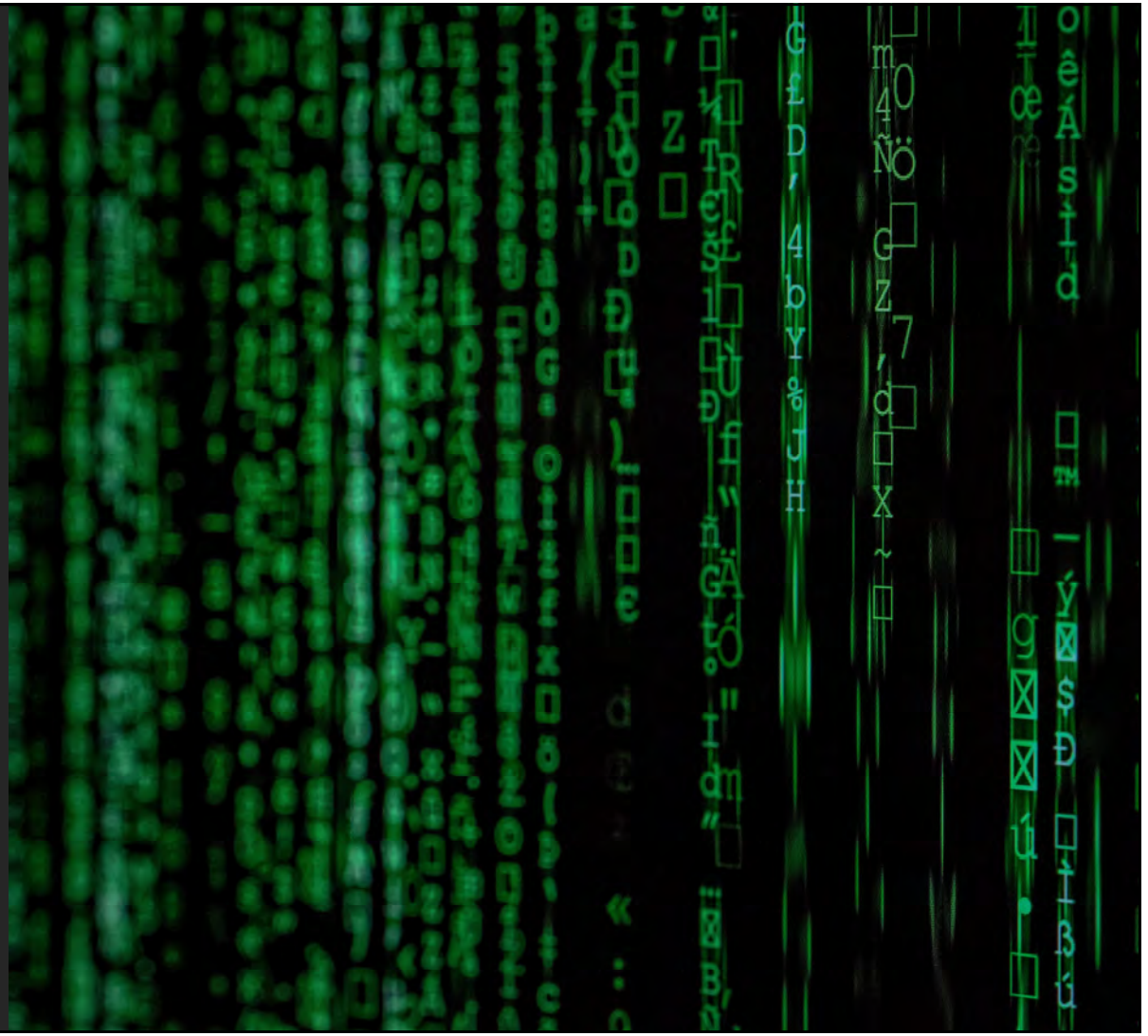
17

Politiek

En natuurlijk...

Datadenken

Om inzicht te krijgen in de mogelijkheden, is het goed om te leren denken in data; ‘datadenken’.



20

Oefening ‘datadenken’.

Bedenk welke data er gebruikt kunnen worden bij het houden van het toezicht op de belastingheffing.

Noem zo precies mogelijk welke data het zijn, bij wie ze opgevraagd kunnen worden en hoe de data geanalyseerd (beoordeeld) worden.

21

1. Auto van de zaak

Data?

Opvragen bij?

Hoe werkt de beoordeling?

NBA: 'Data-analyse is het ontdekken van patronen, afwijkingen, inconsistenties, en het onttrekken van andere nuttige informatie over het object van het onderzoek door middel van analyse, modellering en visualisatie met het oog op de planning of het uitvoeren van de opdracht.'

Data-analyse

Data-analyse is dus het proces van verkrijgen van informatie uit data door gebruik van methoden en technieken.

Voorafgaand aan de analyse zelf moeten de data verkregen worden. Dit proces wordt het ETL-proces genoemd:

Extraction, Transformation, Loading.

ETL-proces

Extraction => het verkrijgen van de data uit (diverse) systemen, bijvoorbeeld een ERP-systeem of een payrollpakket.

Transformation => data moeten zodanig aangepast worden dat het softwarepakket waarmee geanalyseerd gaat worden de data kan interpreteren.

Loading => de te analyseren gegevens worden ingeladen in een database of in de analysesoftware.

24

Data-analyse: 4 typen

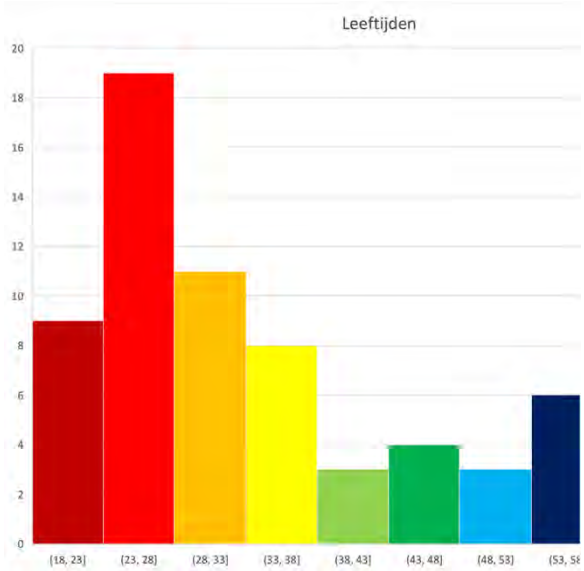
- Descriptive analytics
- Diagnostic analytics
- Predictive analytics
- Prescriptive analytics

25


Data-analyse in soorten

Descriptive analytics

‘Beschrijvende statistiek’, met kengetallen wordt een beschrijving gegeven van de data. Denk aan zaken als het gemiddelde, de mediaan, de modus en de standaarddeviatie. Aanvulling met grafische weergaven, bijvoorbeeld over de verdeling door middel van een histogram.



27

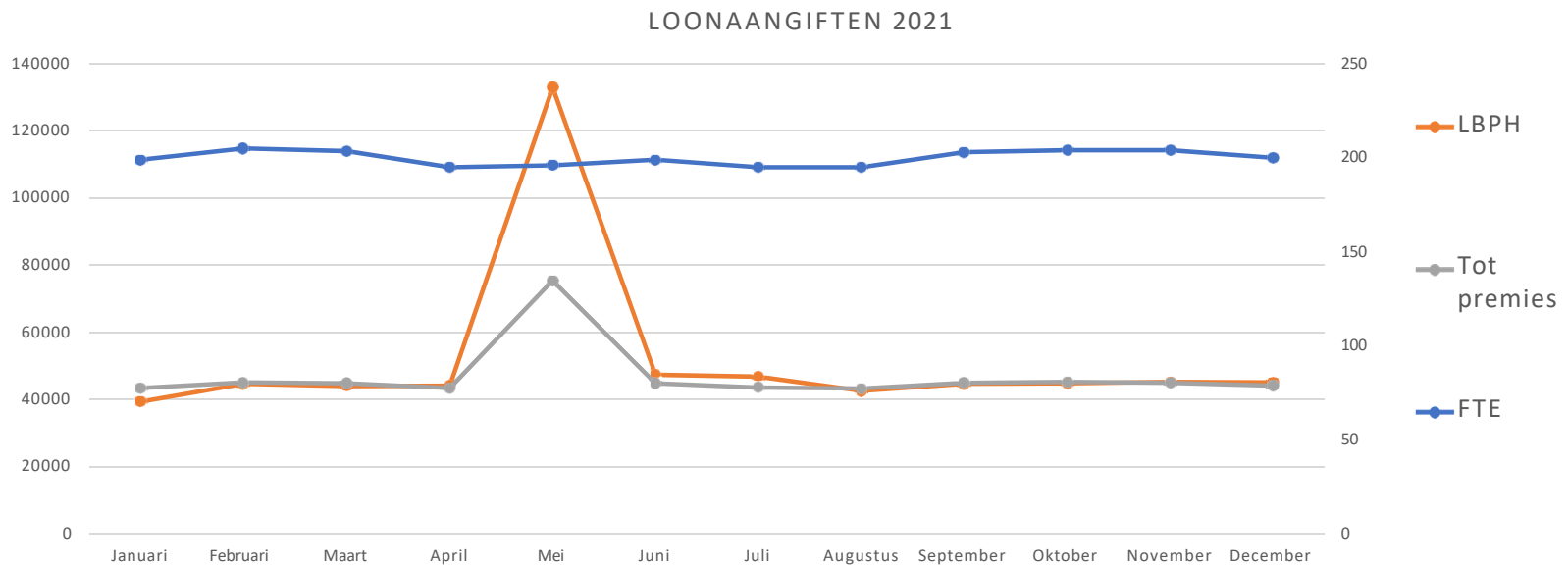
Descriptive analytics in Tax

Maand	FTE	LBPH	Tot premies	ZVW	WIA/WA	WW-Awf	Gem loon
Januari	199	39303	43311	2830	29477	11005	1975
Februari	205	44649	45022	3215	30443	11365	1980
Maart	203,5	44009	44777	3169	30006	11602	1966
April	195	44015	43324	3169	28958	11197	1980
Mei	196	132947	75412	9572	47481	18359	3230
Juni	199	47338	44780	3408	29835	11536	1999
Juli	195	46800	43540	3370	29250	10920	2000
Augustus	195	42510	43231	3061	29250	10920	2000
September	203	44615	44988	3212	30420	11357	1998
Oktober	204	44790	45165	3225	30539	11401	1996
November	204	45197	44991	3254	30539	11198	1996
December	200	45064	44122	3245	29910	10967	1994

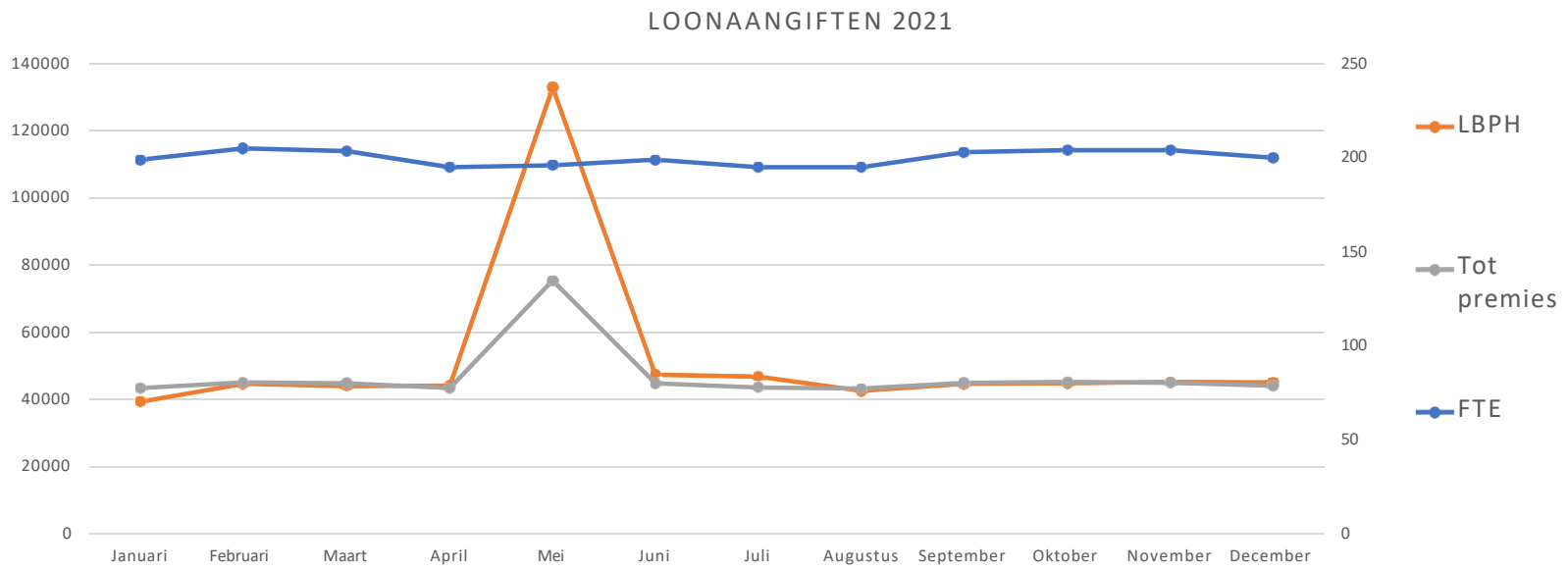
De beschrijving van de data kan worden verrijkt met kengetallen die meer inzicht geven.

Denk bijv. aan toegepaste gemiddelde tarieven, totalen etc.

Descriptive analytics in Tax



Descriptive analytics in Tax

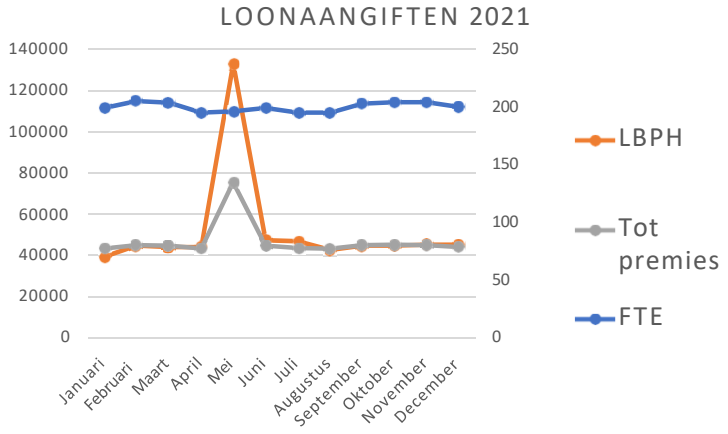


Diagnostic analytics

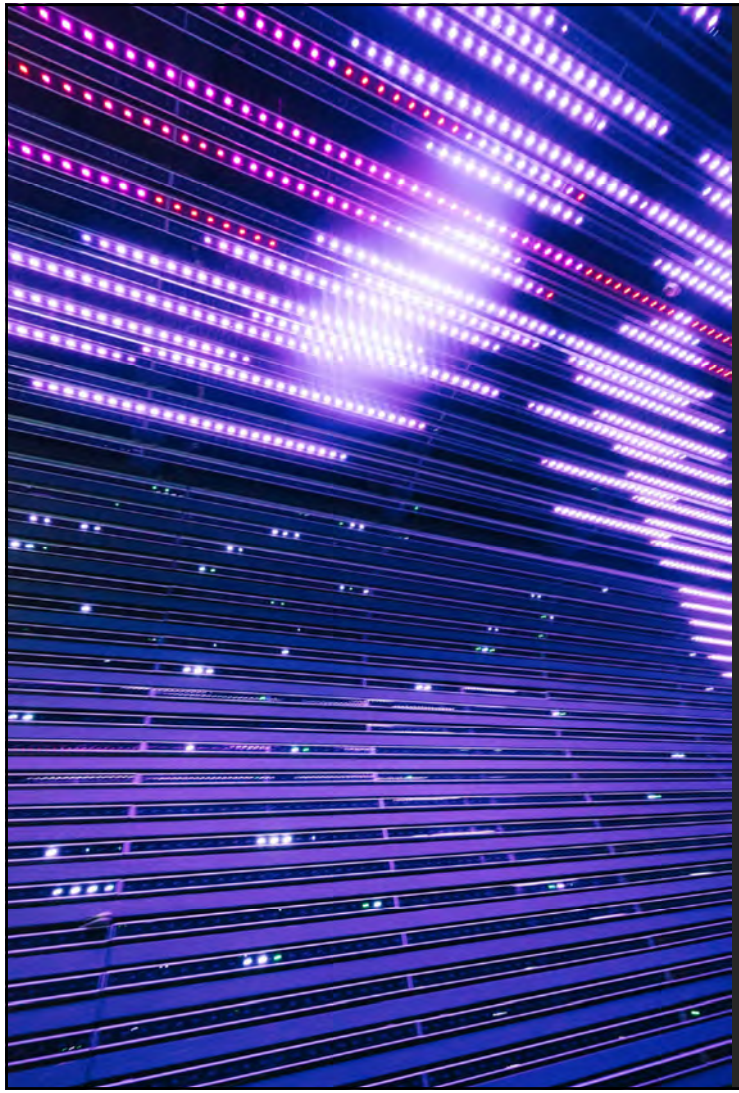
Aan de hand van data proberen te achterhalen wat de oorzaak of oorzaken van een bepaald fenomeen is of zijn. Diagnostic analytics zoekt naar het ‘waarom’ van een bepaalde gebeurtenis. Het beoordelen van de uitstoot van bepaalde gassen en de samenhang met klimaatverandering is een voorbeeld van diagnostic analytics.

Diagnostic analytics in Tax

Wat is de oorzaak van de piek in LBPH en premies in mei?



Maand	FTE	LBPH	Tot premies	ZVW	WIA/WAO	WW-Awf	Gem loon
Januari	199	39303	43311	2830	29477	11005	1975
Februari	205	44649	45022	3215	30443	11365	1980
Maart	203,5	44009	44777	3169	30006	11602	1966
April	195	44015	43324	3169	28958	11197	1980
Mei	196	132947	75412	9572	47481	18359	3230
Juni	199	47338	44780	3408	29835	11536	1999
Juli	195	46800	43540	3370	29250	10920	2000
Augustus	195	42510	43231	3061	29250	10920	2000
September	203	44615	44988	3212	30420	11357	1998
Oktober	204	44790	45165	3225	30539	11401	1996
November	204	45197	44991	3254	30539	11198	1996
December	200	45064	44122	3245	29910	10967	1994



Predictive analytics

Het voorspellen op basis van data.

Vaak is voorafgaand aan het doen van de voorspelling het ontdekken van de wetmatigheid in de data noodzakelijk.

Als de wetmatigheid gevonden is, dan kan die vastgelegd worden in een formule (algoritme.)

Predictive analytics in Tax

Wat zal de volgende maanden de verschuldigde belasting worden? Een analyse van de patronen o.b.v. historische patronen en daarop vastgestelde wetmatigheden kan leiden tot nauwkeurige voorspellingen.

Maand	FTE	LBPH	Tot premies	ZVW	WIA/WAO	WW-Awf	Gem loon
Januari	199	39303	43311	2830	29477	11005	1975
Februari	205	44649	45022	3215	30443	11365	1980
Maart	203,5	44009	44777	3169	30006	11602	1966
April	195	44015	43324	3169	28958	11197	1980
Mei	196	132947	75412	9572	47481	18359	3230
Juni	199	47338	44780	3408	29835	11536	1999
Juli	195	46800	43540	3370	29250	10920	2000
Augustus	195	42510	43231	3061	29250	10920	2000
September	?	?	?	?	?	?	?
Oktober	?	?	?	?	?	?	?
November	?	?	?	?	?	?	?
December	?	?	?	?	?	?	?

Prescriptive analytics

Deze analyse borduurt voort op de predictive analytics en analyseert het ‘recept’ voor het bereiken van een bepaald resultaat.

Door gebruik te maken van meer data, dan in de voorspellende fase is gebeurd, kan met prescriptive analytics het resultaat verbeterd worden.

Samenvatting data-analyse

Descriptive analytics: beschrijving van de data

Diagnostic analytics: beantwoording van het waarom

Predictive analytics: levert een voorspelling

Prescriptive analytics: hoe kan het toekomstig resultaat bereikt (en verbeterd) worden?

Data-analyse, aandachtspunten

Is het doel voldoende duidelijk?

Juridische aspecten (AVG, opdracht)

Technische aspecten (ETL-traject)

Is er voldoende inzicht in de kwaliteit van de data?

Is duidelijk wat de databron is?

Data-analyse,nog meer aandachtspunten

Logging bewerkingen

Opslag ‘deelproducten’

‘Verborgen’ persoonsgegevens

Bewaartermijnen

Overig?

Methoden & Technieken

Methoden en technieken

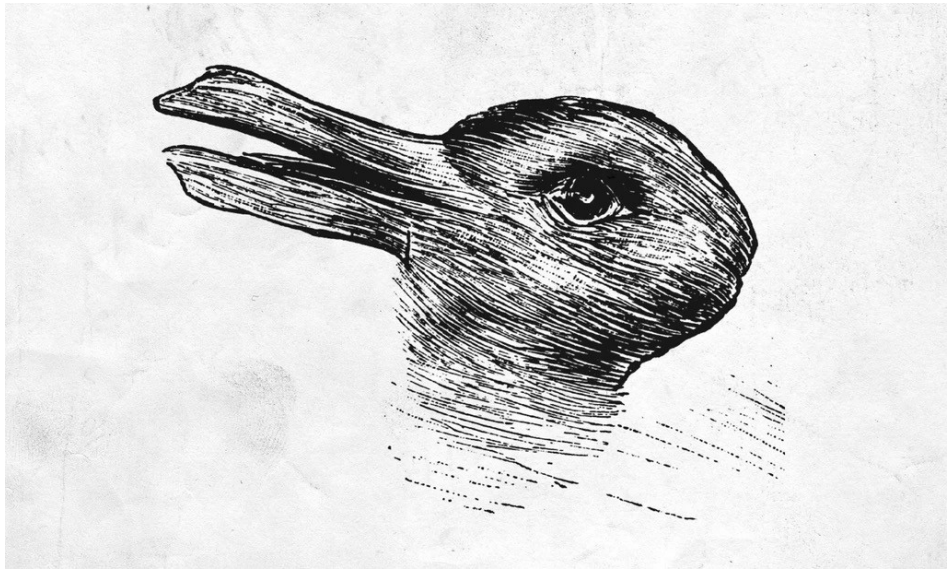
"Lies, damned lies, and statistics"

Mark Twain (Chapters from my Autobiography, 1907).

Ook voor het tijdperk van 'fake news' was de kracht van misleidende cijfers al bekend.

Grafische weergaven kunnen zeer misleidend zijn.

Hoe misinterpretatie voorkomen?



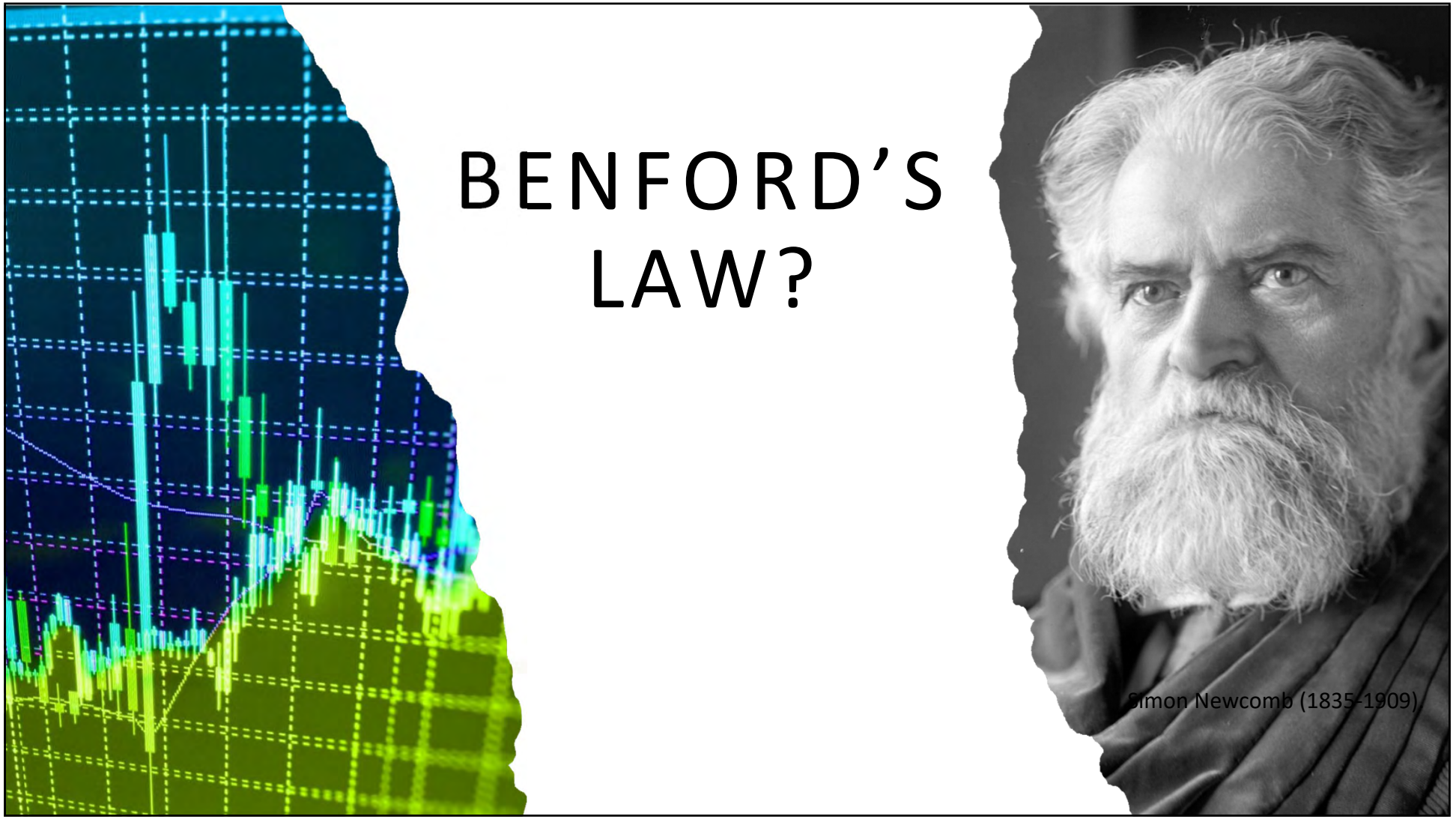
40


STEEKPROEVEN?

CHI² ?

		Cases	Deaths	New Deaths	Total Recovered	Active Cases	Serious, Critical
China	80,881	+21	3,226	+13	68,715	8,940	3,226
Italy	31,506	+3,526	2,503	+345	2,941	26,062	2,060
Iran	16,169	+1,178	988	+135	5,389	9,792	
Spain	11,409	+1,467	510	+168	1,028	9,871	563
Germany	8,639	+1,367	23	+6	67	8,549	2
S. Korea	8,320	+84	81	+6	1,401	6,838	59
France	6,633		148		12	6,473	400
USA	5,704	+1,041	97	+11	74	5,533	12
Switzerland	2,742	+389	27	+8	15	2,700	
UK	1,950	+407	71	+16	52	1,827	20
Netherlands	1,705	+292	43	+19	2	1,660	45
Norway	1,452	+104	3		1	1,448	27

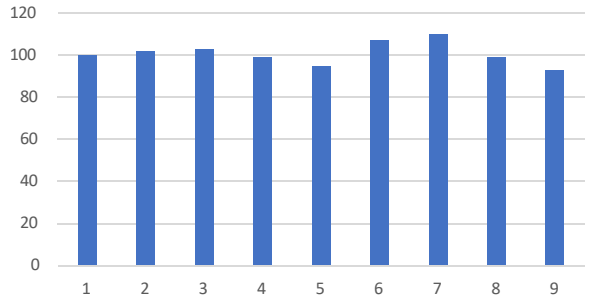
42


BENFORD'S
LAW?
Simon Newcomb (1835-1909)

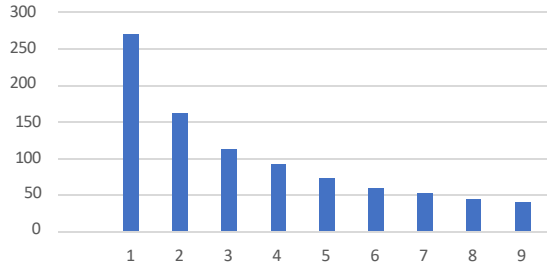
Kwisje!

Als je alle transacties van een onderneming sorteert naar omvang in geldbedrag, hoe ziet de verdeling van de *begincijfers* er dan waarschijnlijk uit?

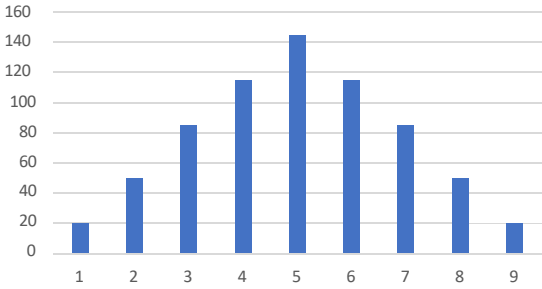
A



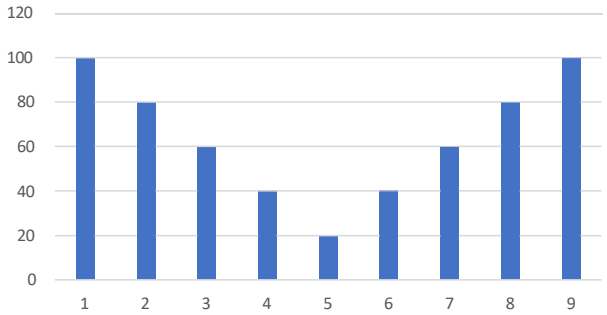
B



C

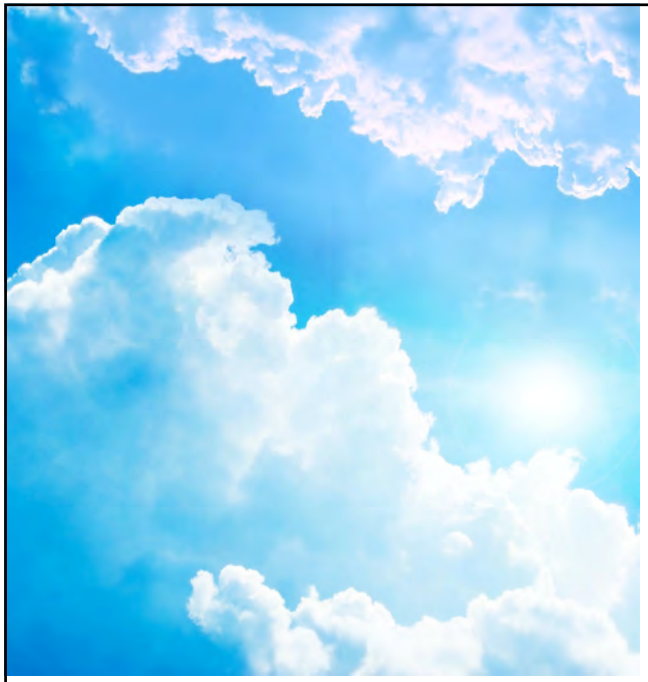


D



44


CORRELATiE
Correlatie is de berekening van de samenhang tussen (numerieke) gegevens.
Niet meer, niet minder.

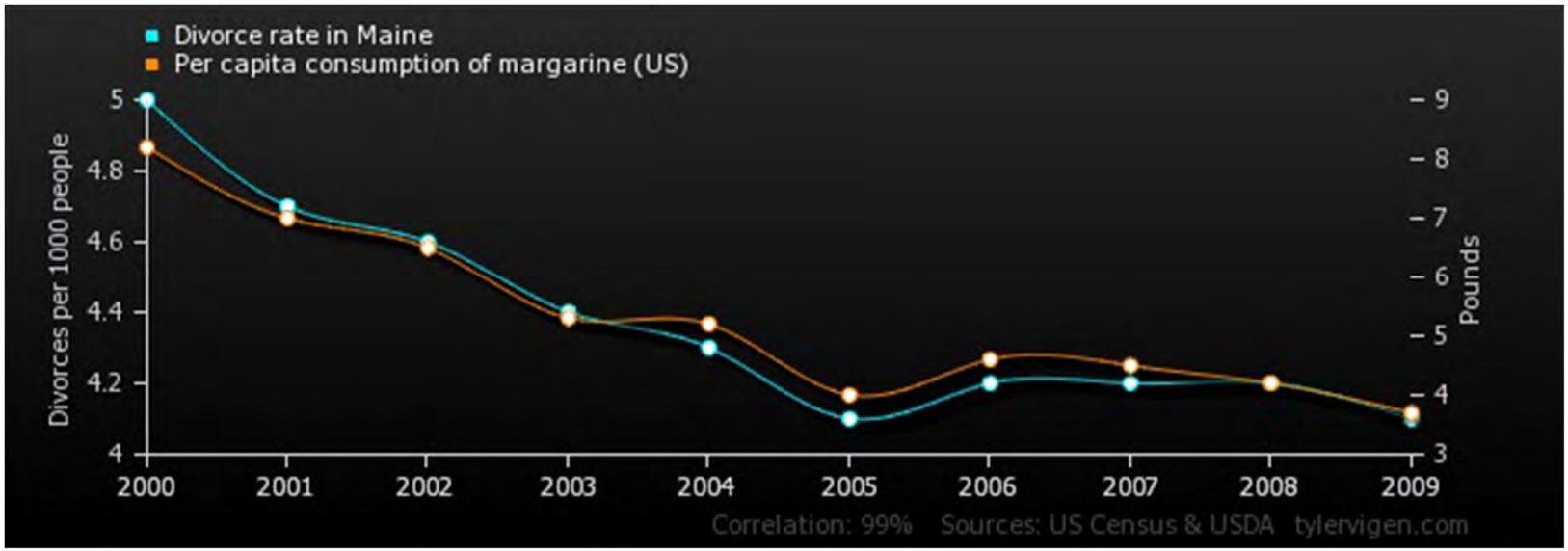



Correlatie..

46


Causatie
Causatie is de wetenschap over oorzaak en gevolg.
1. Er is correlatie
2. Er is een toetsbare theorie die het verband verklaart

47



Correlatie vs causatie

Bron: www.tylervigen.com

Aan de slag met data

Om te voldoen aan de eisen van de fiscale wetten zijn juiste (en tijdige) aangiften van belang. Welke beoordelingen zouden er geautomatiseerd kunnen plaatsvinden? Welke informatie is daarover beschikbaar?

Voor de fiscalist is kennis van de mogelijkheden van belang.

Aan de slag met data, het landschap

Het ‘landschap’ is inzicht in de relevante software en de IT-omgeving waarin deze software functioneert. Denk aan het ERP-systeem, het payrollpakket maar bijvoorbeeld ook een geautomatiseerd toegangssysteem met pasjes.

50

Aan de slag met data.

1. Duidelijke scope; welke belasting(en), welk tijdvak, welke aangifteplichtige, welk doel?
2. Inzicht in het 'landschap', denk ook aan data voor andere toezichthouders!
3. Een overzicht van de 'queries' ofwel, de data-analyses die we willen uitvoeren
4. Welke data hebben we nodig? Aangiften en ?
5. 'Tooling'; software waarme we kunnen analyseren en liefst ook visualiseren

51

Voorbeeld I

1. Scope: maandaangifte loonheffingen 2022. Doel: beoordeling op juistheid
2. Landschap: payrollpakket is AFAS, ERP is AFAS
3. Aangiftedata: XML-loonaangiften (kopie)
4. Queries, t.b.d.
5. Tooling: Excel

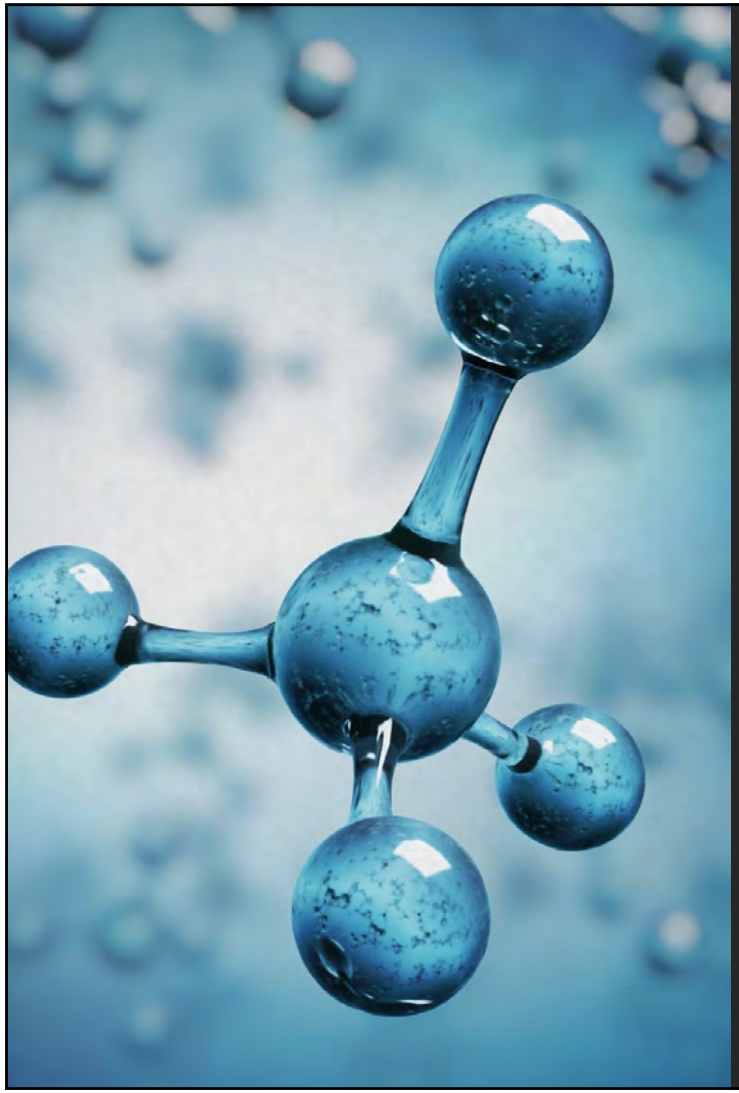
<?xml version="1.0" encoding="ISO-8859-1"?><Loonaangifte xmlns="http://xml.belastingdienst.nl/schemas/Loonaangifte/2020/01" version="2.0"><Bericht><IdBer>ADP012345R0004038071</IdBer><DatTdAanm>2020-04-23T13:52:00</DatTdAanm><ContPers>KC-ADP NL HR DEMO </ContPers><TelNr>0201234567</TelNr><RelNr>swo00012</RelNr><GebrSwPakket>Multipay TWK</GebrSwPakket></Bericht><AdministratieveEenheid><LhNr>112233445L01</LhNr><NmIP>DEMO BV </NmIP><TijdvakAangifte><DatAanvTv>2020-01-01</DatAanvTv><DatEindTv>2020-01-31</DatEindTv><VolledigeAangifte><CollectieveAangifte><TotLnLbPh>0000596598</TotLnLbPh><TotLnSV>0000596598</TotLnSV><TotPrlnAwfAnwLg>0000560577</TotPrlnAwfAnwLg><TotPrlnAwfAnwHg>0000018979</TotPrlnAwfAnwHg><TotPrlnAwfAnwHz>0000000000</TotPrlnAwfAnwHz><PrLnUFO>0000000000</PrLnUFO><IngLbPh>0000150106</IngLbPh><PrWAOAof>0000042133</PrWAOAof><TotPrAwfLg>0000016480</TotPrAwfLg><TotPrAwfHg>0000001507</TotPrAwfHg><TotPrAwfHz>0000000000</TotPrAwfHz><IngBijdrZvw>0000000000</IngBijdrZvw><TotWghZvw>0000038830</TotWghZvw><TotTeBet>0000249056</TotTeBet><TotGen>0000249056</TotGen></CollectieveAangifte><InkomstenverhoudingInitieel><NumIV>0002</NumIV><DatAanv>2011-08-01</DatAanv><PersNr>0001</PersNr><NatuurlijkPersoon><SofiNr>000000001</SofiNr><Voorl>A</Voorl><SignNm>DEMO A</SignNm><Gebdat>1965-06-26</Gebdat><Nat>0001</Nat><Gesl>1</Gesl><AdresBinnenland><Str>DORPSSTRAAT</Str><HuisNr>1</HuisNr><Pc>6412WL</Pc><Woonpl>HEERLEN</Woonpl></AdresBinnenland></NatuurlijkPersoon><Inkomstenperiode><DatAanv>2020-01-01</DatAanv><SrtIV>15</SrtIV><CdAard>01</CdAard><CAO>0824</CAO><IndArbovOnbepTd>J</IndArbovOnbepTd><IndSchriftArbov>J</IndSchriftArbov><IndOprov>N</IndOprov><IndAvrLkvOudrWn>N</IndAvrLkvOudrWn><IndAvrLkvAgWn>N</IndAvrLkvAgWn><IndAvrLkvDgBafSb>N</IndAvrLkvDgBafSb><IndAvrLkvHpAgWn>N</IndAvrLkvHpAgWn><IndLhKort>J</IndLhKort><LbTab>012</LbTab><IndWAO>J</IndWAO>

Data en queries

Het formuleren van de queries, ofwel het beschrijven van hetgeen onderzocht moet worden is natuurlijk essentieel. Dit is de uitwerking van het doel van het onderzoek: is het belang een algemeen beeld krijgen over de juistheid van de aangiften, of wordt er gezocht naar specifieke fouten?

In het voorbeeld is de scope nogal breed: beoordeling op juistheid. Hoe vertalen we dat in queries?

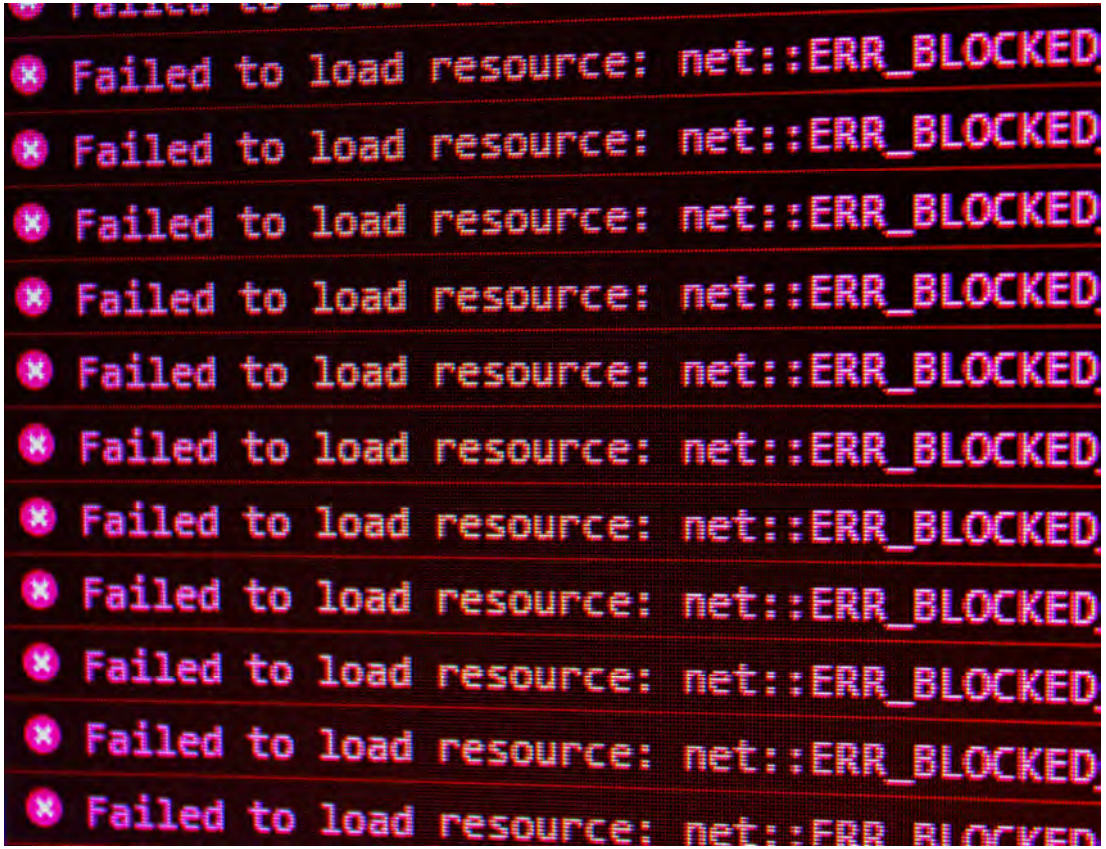
De mogelijke queries hangen af van de beschikbaarheid van de data, dus data en queries is enigszins een kip en ei vraagstuk..



Queries, voorbeeld II.

Queries moeten zo scherp mogelijk geformuleerd worden. Stel u doet een beoordeling van de verkopen op juistheid van de in rekening gebrachte BTW. Dan heeft u deze gegevens nodig:

- Omschrijving geleverde dienst of geleverde goederen
- Datum factuur
- Datum prestatie
- Vestigingsland afnemer



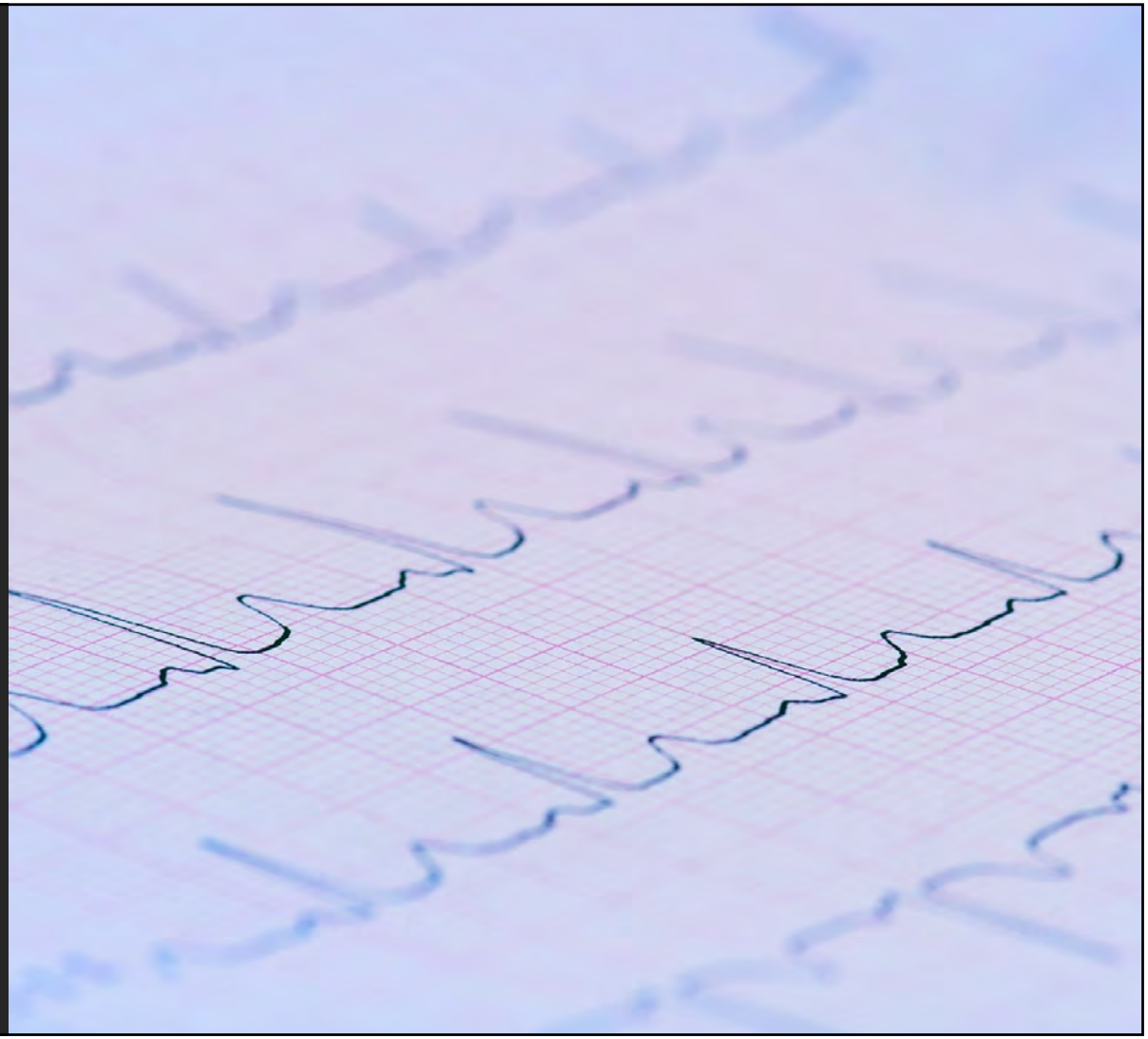
Queries, voorbeeld II.

Maar natuurlijk ook:

- het tijdvak waar de beoordeling op ziet

- hoe er gerapporteerd moet worden:
  - bij afwijking van 21% met omschrijving dienst/levering?
  - of bij 9% met omschrijving dienst/levering?
  - bij 0% maar geen afnemer in buitenland?

Trends in data-analyse



56

Continuous monitoring

Van incidenteel naar frequent naar doorlopend.

Veel organisaties hebben voor hun businessproces een grote informatiebehoefte en richten dan ook dit zelf al in (dashboards).

Voor (financiële) rapportagedoel-einden kan gebruik gemaakt worden van de bestaande informatiestromen.

57

Van geaggreerd naar transacties

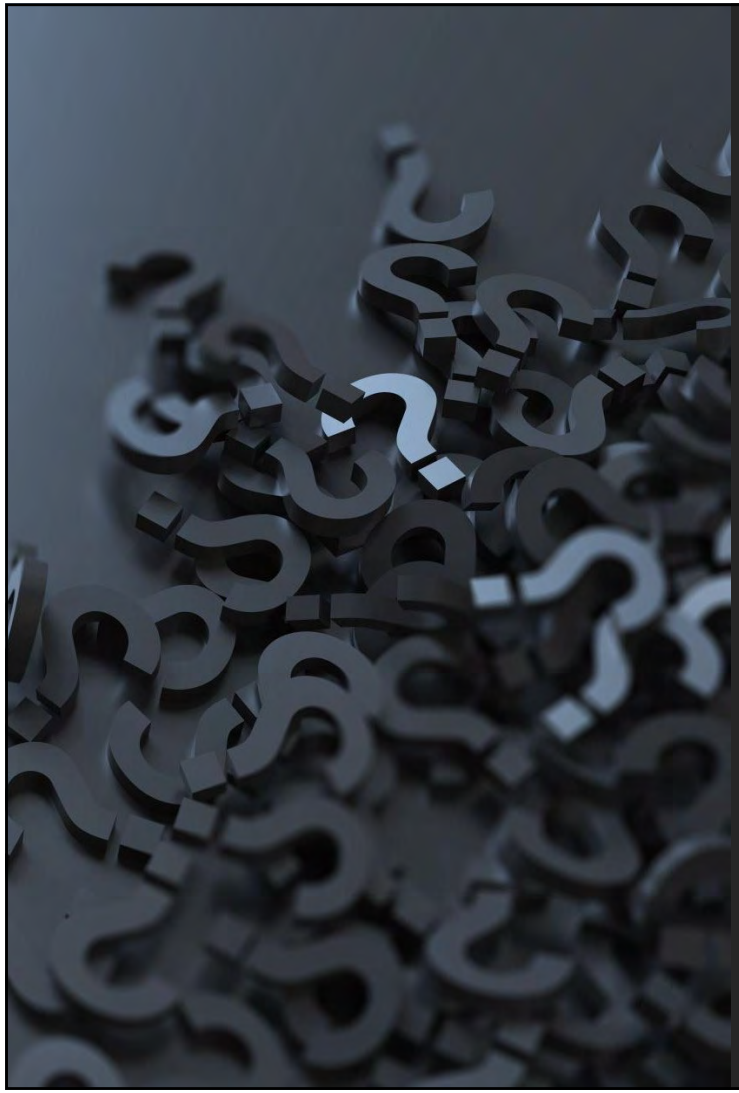
Geaggreerde data zeggen minder dan de bouwstenen van deze data: de transacties. Veel meer informatie kan ontleend worden aan de transacties.

De beoordeling van de transactie-data is wel complexer; de samenhang tussen de data moet in kaart gebracht worden.

Van verzenden naar toegang systeem?

Verzenden van data is niet zonder risico. Een alternatief is het beoordelen van de originele data, dat vereist toegang tot deze data (systemen).

In sommige landen (Hongarije) is er al wetgeving die hiertoe verplicht.



Conclusie

Data is alomtegenwoordig en niet meer weg te denken uit de moderne tijd

De steeds toenemende hoeveelheid data biedt in combinatie met steeds snellere computers en daarmee samenhangende nieuwe technieken een enorm potentieel in alle domeinen.

Om bij te blijven is nodig:

- inzicht in de benodigde data
- inzicht in methoden en technieken (data-analyse)
- inzicht in het juridisch/ethisch kader

60

Einde deel 1

61