VU VRIJE UNIVERSITEIT AMSTERDAM

VU Research Portal

Patroonherkenning: Tussen tellen en Toetsen

Elzinga, C.H.

2012

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)
Elzinga, C. H. (2012). *Patroonherkenning: Tussen tellen en Toetsen*. VU University.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

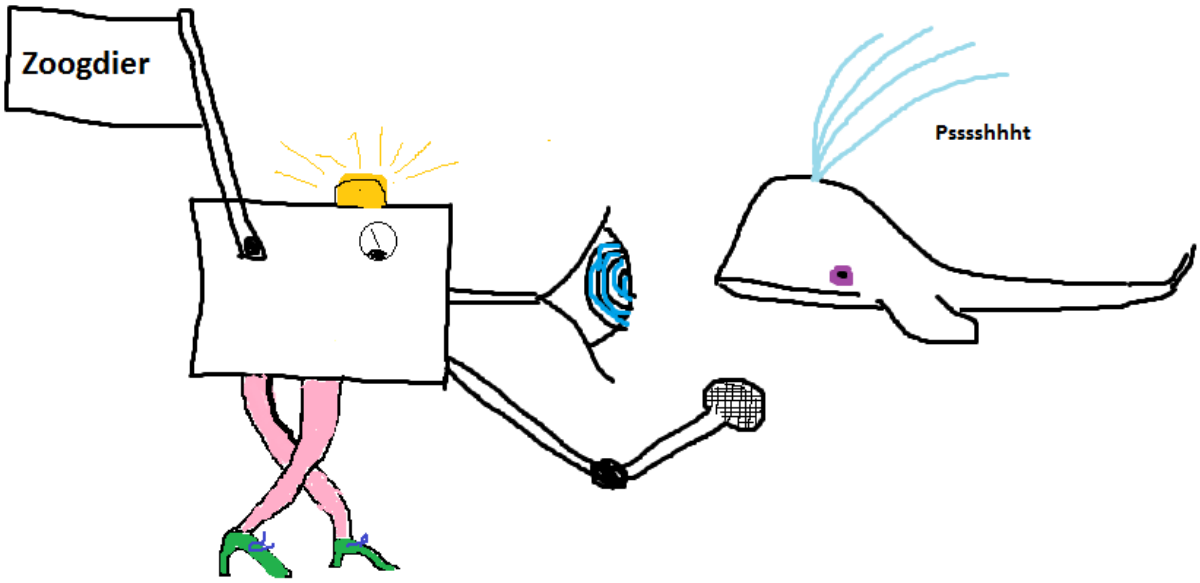
E-mail address:
vuresearchportal.ub@vu.nl

Download date: 26. May. 2021

Patroonherkenning:

Tussen tellen en toetsen

Oratie van C.H. Elzinga dd 20 december 2012



Mijnheer de Rector, Dames en Heren, beste vrienden en collega's,

Dank voor uw komst naar deze openbare les.
Mijn doel is u in begrijpelijke taal uitleggen wat ik deed en nog ga doen. Ik moet hier en daar wat vereenvoudigen; niet omdat er domoren in de zaal zitten maar omdat de nuance en het detail veel te veel tijd kosten. Ik hoop dat u daarvoor begrip wil hebben.

Mijn leeropdracht heet "Patroonherkenning in discrete datastructuren". Ik ga u uitleggen wat dat betekent en wat ik daarmee hoop te bereiken in de sociale wetenschappen. Ik ga u ook vertellen wat de plaats is van die bezigheid temidden van het bonte boeket dat "sociale wetenschappen" heet en op dat laatste heeft het motto van deze lezing betrekking: "Tussen tellen en toetsen".

U heeft allemaal een horloge maar ik wed dat er hier niemand in de zaal is die kan uitleggen hoe dat werkt, dat horloge. Ik ook niet. En dat geeft ook niet want we kunnen allemaal een horloge gebruiken.

Zo ga ik u vandaag niet uitleggen hoe patroonherkenning werkt maar ik ga u wel uitleggen hoe je patroonherkenning kunt gebruiken en hoe dat er dan uitziet.

Laat ik beginnen met "datastructuren". Data zijn de gegevens waarmee de wetenschapper werkt, waarop hij zijn uitspraken baseert. Een datastructuur omvat een hoeveelheid gegevens, opgeslagen in een bestand, een lijst of tabel met een beschrijving van wat er in die lijst te vinden is en hoe je dat kunt vinden.
Hier is een voorbeeld van zo'n bestand: school/werk-carriëres van Noord-Ierse jongeren:

MVAD - Kladblok

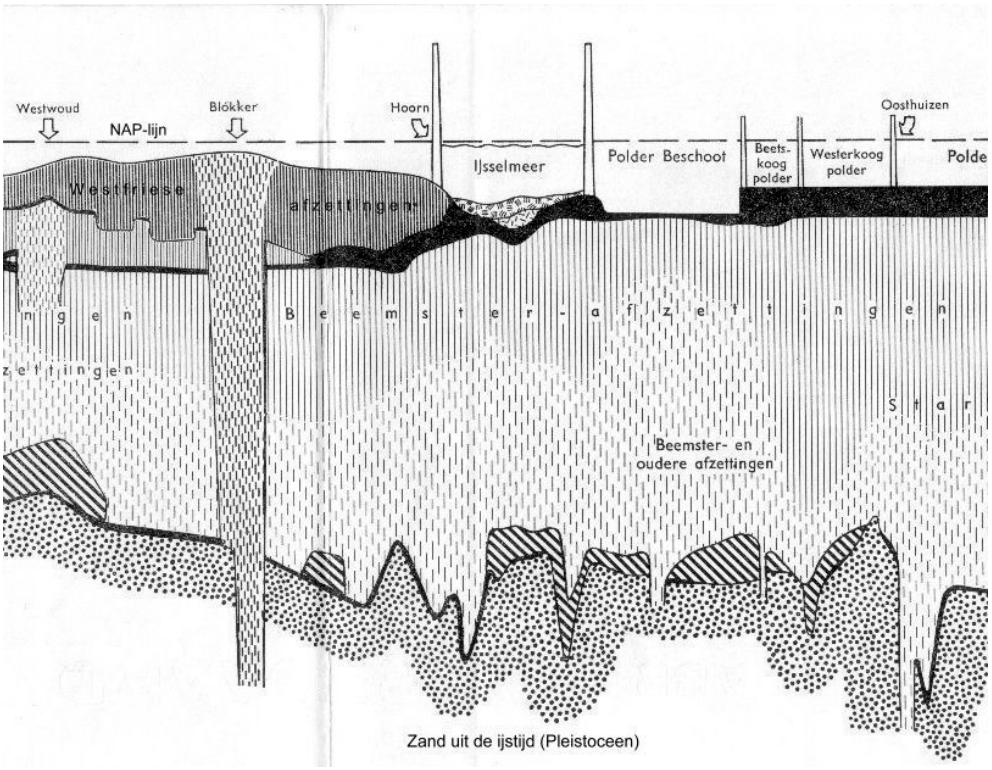
Bestand Bewerken Opmaak Beeld Help

```
id weight male catholic Belfast NEastern SouthRn SEastern Western
1 0.33 0 0 0 0 0 0 1 0 0 0 1 1 T/2 E/4 T/2 E/64
2 0.57 0 0 0 0 0 0 1 0 0 1 0 1 U/2 F/36 H/34
3 1.59 1 1 0 0 0 0 1 0 0 0 0 1 U/2 T/24 F/34 E/10 U/2
4 1.59 0 0 0 0 0 0 1 0 0 0 0 1 T/49 E/14 U/9
5 0.57 1 0 0 0 0 0 1 0 1 0 0 1 U/2 F/25 H/45
6 1.59 1 1 0 0 0 0 1 0 0 0 0 0 U/3 T/33 E/36
7 0.57 1 1 0 0 0 0 1 0 0 0 0 0 U/2 F/30 E/40
8 2.75 1 1 0 0 0 0 1 1 0 0 1 1 E/2 F/22 E/48
9 2 0 0 0 0 0 1 0 0 0 0 0 0 U/2 T/21 E/49
10 3.6 0 0 0 0 0 1 0 0 0 0 0 1 E/2 S/10 U/2 E/46 U/12
11 0.69 1 0 0 0 0 1 0 0 0 0 1 1 U/1 E/1 F/49 E/12 H/9
12 1.1 0 0 0 0 0 0 1 0 0 1 0 1 E/2 S/36 F/9 E/25
13 1.1 1 1 0 0 0 0 1 0 1 1 1 1 U/2 S/24 H/13 E/12 H/21
14 0.57 0 1 0 0 0 0 1 0 0 1 0 0 U/2 F/20 E/33 U/17
15 2 0 1 0 0 0 0 1 1 0 0 0 1 T/12 U/18 E/8 U/34
```

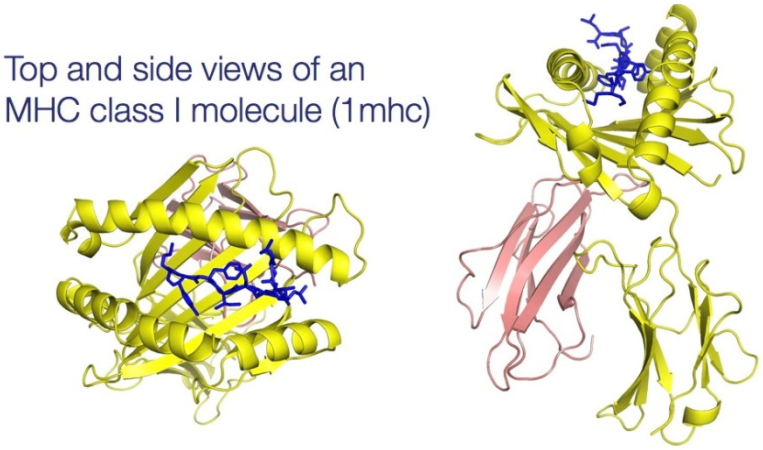
Hier is nog een voorbeeld: gegevens over de sociale netwerken van zwak-begaafden.

46	4	F	Z	1		
47	5	H	F			
48	6	F	SFM			
49	INT358		1			
50	1	F	ego	2	4,7	7
51	2	H	F	3,1	1	3
52	3	H	B	4,2	1,4	2,4
53	4	F	BC	3	3,5,6,1	3
54	5	H	BS	4,3,	6	4,3
55	6	H	BS	4,3	5	4,3
56	7	F	BD		1	
57	INT359		1			
58	1	H	ego	2,3	4	
59	2	H	pCF	3	3,4,	
60	3	F	pCM	2	4	2
61	4	F	pC		1,3,	
62	5	F	Z		16	
63	6	F	D	12	12	
64	7	F	D			
65	8	F	Z	9,13,7	14	
66	9	H	ZC	8		8

Nu denken we bij wetenschappelijke data vaak aan lange reeksen getallen voor en achter de komma: gegevens over snelheden, aantallen nakomelingen, magnetische velden en wat dies meer zij.
Maar dan vergeten we even dat veel wetenschappelijke gegevens helemaal niet uit metingen, uit getallen bestaan maar uit namen, codes, afkortingen en verwijzingen. Zo ziet u in dit voorbeeld allerlei afkortingen die iets zeggen over de relatie tussen de leden van een sociaal netwerk; de getallen zijn verwijzingen en vertellen wie contact heeft met wie. Kortom, codes, afkortingen en verwijzingen. En denkt u nu niet dat zulke rare gegevens alleen in de sociale wetenschappen voorkomen: hier is een voorbeeld uit de geologie – een dwarsdoorsnede van de bodem onder het IJsselmeer in de buurt van Hoorn:



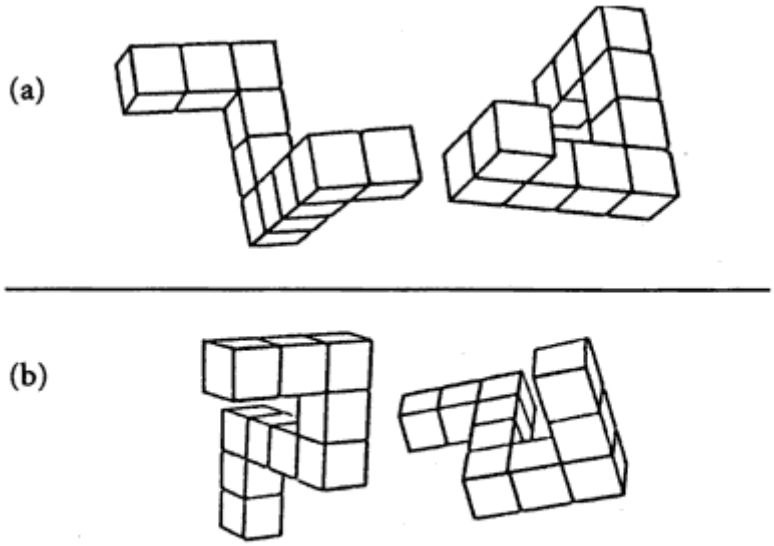
Geen getal te zien en toch heel informatief: namen en codes in de vorm van arceringen – het product van boren, visuele inspectie en chemische analyse.
Hier is nog een voorbeeld:



Schematische voorstellingen van eiwitten als ruimtelijke structuren: verschillend gekrulde en gekleurde linten. In werkelijkheid zijn er geen linten en geen kleuren. Toch vatten die plaatjes op een prachtige manier samen hoe ons begrip van die stoffen opgebouwd is. Ook deze plaatjes kunnen we opslaan in de vorm van lijsten met namen, symbooltjes, richtingen en relaties.
Zo zien we dus dat veel wetenschappelijke gegevens worden opgeslagen in de vorm van letters en codes en helemaal geen “numeriek” karakter hebben. Dat we als codes dan toch vaak weer getallen gebruiken doet niet ter zake: de getallen fungeren als symbolen die makkelijk in computerbestanden zijn op te bergen, net als telefoonnummers.
Zulke niet-numerieke, symbolische gegevens noemen we “discrete data”.

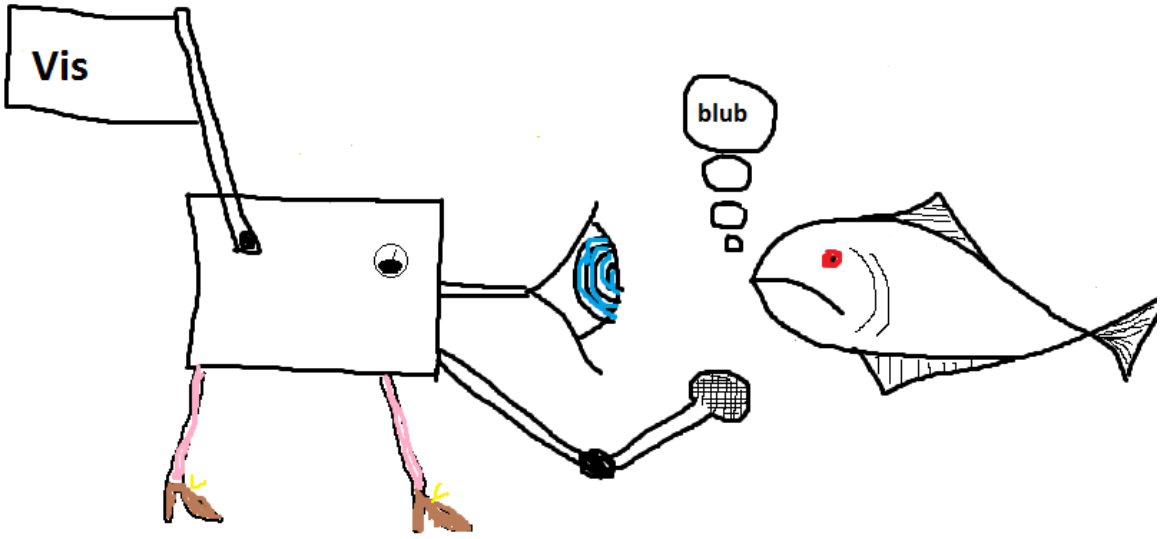
Dat was de laatse helft van “Patroonherkenning in discrete datastructuren”. Nu de eerste helft, de patroonherkenning.

Intuïtief hebben we daar allemaal wel een voorstelling van, van “patroonherkenning”. Het herkennen van regelmaat en structuur, het correct benoemen van allerlei objecten en gebeurtenissen, het sorteren en classificeren van de dingen om ons heen. Wij mensen zijn er heel goed in, we herkennen al heel vroeg complexe beelden zoals de gezichten van onze ouders, we leren al snel gesproken taal begrijpen, liedjes zingen en ritmisch bewegen.



Mental Rotation Test—Are these two figures the same except for their orientation?

Ook dit testje laat zien hoe goed wij kunnen “herkennen”. De meesten van ons kunnen dit soort puzzeltjes wel oplossen maar het kost tijd, het is saai en we maken er allemaal ook gemakkelijk fouten mee.



Ook machines kunnen we “leren” om patronen te “herkennen”. Hier ziet u een prototype dat het al heel aardig doet.
Bij zo’n patroonherkenner denken we dan aan een combinatie van een “waarnemer”, een “herkenner” en een “uitvoerorgaan”. De waarnemer is een apparaat dan kan “zien”, “luisteren” of “voelen”. De herkenner bestaat uit software die het signaal van de waarnemer analyseert en het uitvoerorgaan vertelt ons wat waargenomen is. In de sociale wetenschappen is de wetenschapper meestal degene die de data aan de “herkenner” levert

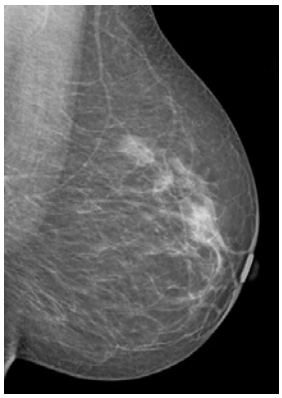
in de vorm van een data-bestand. De taak van de patroonherkenner is om bij de gegeven invoer het best passende “label” of “type” te produceren. U ziet, deze patroonherkenster is al aardig ingewerkt!

Inmiddels zit ons dagelijks leven vol met patroonherkennings-machines. Ik geef een paar voorbeelden.

Het patroon is de bar-code: invoer voor een bar-code scanner. In alle mogelijke rotaties en hoeken kan het apparaat heel snel herkennen om wat voor product het gaat en het proces initiëren dat daarbij hoort: bedrag op de rekening toevoegen, voorraad afboeken, etc.

Een enorme postsorteer-inrichting: herkent tienduizenden adressen per uur, handgeschreven of met verschillende lettertypen en vanuit allerlei hoeken.

Gecompliceerder is spraakherkenning; hier een foto van een zware kraan in een hoogoven die volledig met gesproken commando's, de patronen, bediend wordt door verschillende operators met verschillende stemmen, accenten en intonaties!



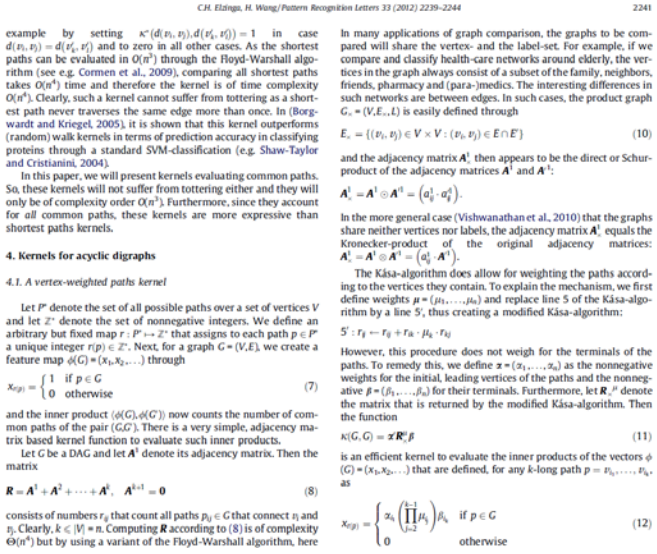
Beeldherkenning aan een mammografie. Ingewikkelde beelden met veel verschillende soorten weefsel. We zijn steeds beter in staat dit soort beelden automatisch te scannen; het aantal gevallen waarin de radioloog of oncoloog er zelf aan te pas moet komen daalt snel dankzij steeds betere patroonherkenning, d.w.z. dankzij steeds geavanceerdere wiskunde om beelden te analyseren.

Tenslotte een rij machines die weinigen onder u zullen herkennen: DNA-sequencers.

Apparaten die een chemische analyse uitvoeren op ons erfelijk materiaal DNA. In essentie bestaat dat materiaal uit een lange keten, opgebouwd uit slechts 4 verschillende basen. Hier zien we een model van zo'n gecompliceerd molecuul.

Als we goed kijken zien we bolletjes met slechts 4 verschillende kleuren om de 4 basen te kunnen voorstellen waaruit het DNA is opgebouwd. De invoer in de sequencer bestaat uit een potje met een vloeistof waarin DNA-houdend materiaal is opgelost. De uitvoer van zo'n sequencer zien we hieronder: een hele lange sequentie opgebouwd uit 4 symbolen A,C, T en G, de afkortingen voor de 4 basen.

Om te kijken of organismen met elkaar verwant zijn of op elkaar lijken kunnen we dit soort patronen met elkaar vergelijken. En zo begrijpen we meteen waarom hier de menselijke patroonherkenner verslagen wordt door een machine: de sequenties zijn heel lang en de verschillen zijn subtiel.
Mensen zijn heel goed met geluid en beeld maar niet zo goed met lange rijen van symbolen, d.w.z. niet zo goed met discrete data. Dus hebben we discrete wiskunde, computers en software nodig om dit soort data snel te analyseren. Werken aan patroonherkenning betekent dus het bedenken van wiskunde om patronen te ontdekken en te vergelijken en die wiskunde te vertalen in bruikbare software. Concreet: artikelen schrijven over rekenen en programmeren met symbolen.
Artikelen zien er zo uit:



C.H. Elzinga, H. Wang / Pattern Recognition Letters 33 (2012) 2239–2244 2241

example by setting $\kappa^*(d(v_i,v_j),d(v'_k,v'_l)) = 1$ in case $d(v_i,v_j) = d(v'_k,v'_l)$ and to zero in all other cases. As the shortest paths can be evaluated in $O(n^3)$ through the Floyd-Warshall algorithm (see e.g. Cormen et al., 2009), comparing all shortest paths takes $O(n^4)$ time and therefore the kernel is of time complexity $O(n^4)$. Clearly, such a kernel cannot suffer from tottering as a shortest path never traverses the same edge more than once. In (Borgwardt and Kriegel, 2005), it is shown that this kernel outperforms (random) walk kernels in terms of prediction accuracy in classifying proteins through a standard SVM-classification (e.g. Shaw-Taylor and Cristianini, 2004).

In this paper, we will present kernels evaluating common paths. So, these kernels will not suffer from tottering either and they will only be of complexity order $O(n^3)$. Furthermore, since they account for *all* common paths, these kernels are more expressive than shortest paths kernels.

4. Kernels for acyclic digraphs

4.1. A vertex-weighted paths kernel

Let P^* denote the set of all possible paths over a set of vertices V and let $\mathbb{Z}^+$ denote the set of nonnegative integers. We define an arbitrary but fixed map $r : P^* \mapsto \mathbb{Z}^+$ that assigns to each path $p \in P^*$ a unique integer $r(p) \in \mathbb{Z}^+$. Next, for a graph $G = (V,E)$, we create a feature map $\phi(G) = (x_1, x_2, \ldots)$ through

$$x_{r(p)} = \begin{cases} 1 & \text{if } p \in G \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and the inner product $\langle \phi(G), \phi(G') \rangle$ now counts the number of common paths of the pair (G,G'). There is a very simple, adjacency matrix based kernel function to evaluate such inner products.

Let G be a DAG and let $\boldsymbol{A}^1$ denote its adjacency matrix. Then the matrix

$$\boldsymbol{R} = \boldsymbol{A}^1 + \boldsymbol{A}^2 + \cdots + \boldsymbol{A}^k, \quad \boldsymbol{A}^{k+1} = \boldsymbol{0} \tag{8}$$

consists of numbers r_{ij} that count all paths $p_{ij} \in G$ that connect v_i and v_j. Clearly, $k \leqslant |V| = n$. Computing $\boldsymbol{R}$ according to (8) is of complexity $\Theta(n^4)$ but by using a variant of the Floyd-Warshall algorithm, here

In many applications of graph comparison, the graphs to be compared will share the vertex- and the label-set. For example, if we compare and classify health-care networks around elderly, the vertices in the graph always consist of a subset of the family, neighbors, friends, pharmacy and (para-)medics. The interesting differences in such networks are between edges. In such cases, the product graph $G_\times = (V, E_\times, L)$ is easily defined through

$$E_\times = \{(v_i,v_j) \in V \times V : (v_i,v_j) \in E \cap E'\} \tag{10}$$

and the adjacency matrix $\boldsymbol{A}^1_\times$ then appears to be the direct or Schur-product of the adjacency matrices $\boldsymbol{A}^1$ and $\boldsymbol{A}'^1$:

$$\boldsymbol{A}^1_\times = \boldsymbol{A}^1 \odot \boldsymbol{A}'^1 = \left(a^1_{ij} \cdot a'^1_{ij}\right).$$

In the more general case (Vishwanathan et al., 2010) that the graphs share neither vertices nor labels, the adjacency matrix $\boldsymbol{A}^1_\times$ equals the Kronecker-product of the original adjacency matrices: $\boldsymbol{A}^1_\times = \boldsymbol{A}^1 \otimes \boldsymbol{A}'^1 = \left(a^1_{ij} \cdot \boldsymbol{A}'^1\right)$.

The Kàsa-algorithm does allow for weighting the paths according to the vertices they contain. To explain the mechanism, we first define weights $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and replace line 5 of the Kàsa-algorithm by a line 5′, thus creating a modified Kàsa-algorithm:

$5' : r_{ij} \leftarrow r_{ij} + r_{ik} \cdot \mu_k \cdot r_{kj}$

However, this procedure does not weigh for the terminals of the paths. To remedy this, we define $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ as the nonnegative weights for the initial, leading vertices of the paths and the nonnegative $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$ for their terminals. Furthermore, let $\boldsymbol{R}_{,j}$ denote the matrix that is returned by the modified Kàsa-algorithm. Then the function

$$\kappa(G,G) = \boldsymbol{\alpha}' \boldsymbol{R}'' \boldsymbol{\beta} \tag{11}$$

is an efficient kernel to evaluate the inner products of the vectors $\phi(G) = (x_1, x_2, \ldots)$ that are defined, for any k-long path $p = v_{i_1}, \ldots, v_{i_k}$, as

$$x_{r(p)} = \begin{cases} \alpha_{i_1} \left(\prod_{j=2}^{k-1} \mu_{i_j} \right) \beta_{i_k} & \text{if } p \in G \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

en een stukje code ziet er zo uit:

```
else//dichotomous!
{
//recode the sequences
    Hashtable htx=new Hashtable();
    Hashtable hty=new Hashtable();
    ArrayList xcode=new ArrayList();
    ArrayList ycode=new ArrayList();
    double x=(double) typear[0];
    string code="";
    for (int iseq=0;iseq<DataSpec.NSeqs;iseq++)
    {
        if(DataSpec.cov[iseq,EntropyData.itype]==x)
        {
            for (int j=0;j<DataSpec.seqlength[iseq];j++)
            {
                code=(string) DataSpec.SeqCode[DataSpec.seq[iseq][j]];
                if(htx.ContainsKey(code))
                {
                    DataSpec.seq[iseq][j]=(int)htx[code];
                }
                else
                {
                    DataSpec.seq[iseq][j]=htx.Count;
                    htx.Add(code,htx.Count);
                    xcode.Add(code);
                }
            }
        }
        else
        {
            for (int j=0;j<DataSpec.seqlength[iseq];j++)
            {
                code=(string)DataSpec.SeqCode[DataSpec.seq[iseq][j]];
                if(hty.ContainsKey(code))
                {
                    DataSpec.seq[iseq][j]=(int)hty[code];
                }
                else
                {
                    DataSpec.seq[iseq][j]=hty.Count;
                     hty.Add(code,hty.Count);
```

Sneller rekenen, nauwkeuriger rekenen, toepassingen ontwikkelen. Dat is wat ik doe. Samen met collega's in Ierland, in Italië, in Zwitserland en in China. Maar ook hier, met collega's in Amsterdam en Den Haag.

Ook in de sociale wetenschappen?

Ja, ook in de sociale wetenschappen. Net als in de micro-biologie, komen in de sociale wetenschappen heel veel data voor die de vorm hebben van lange rijen van een beperkt aantal verschillende symbolen. Van zulke data is dit een mooi voorbeeld

MVAD - Kladblok

Bestand Bewerken Opmaak Beeld Help

```
id weight male catholic Belfast NEastern SouthRn SEastern Western
1 0.33 0 0 0 0 0 0 1 0 0 0 1 1 T/2 E/4 T/2 E/64
2 0.57 0 0 0 0 0 0 1 0 0 1 0 1 U/2 F/36 H/34
3 1.59 1 1 0 0 0 0 1 0 0 0 0 1 U/2 T/24 F/34 E/10 U/2
4 1.59 0 0 0 0 0 0 1 0 0 0 0 1 T/49 E/14 U/9
5 0.57 1 0 0 0 0 0 1 0 1 0 0 1 U/2 F/25 H/45
6 1.59 1 1 0 0 0 0 1 0 0 0 0 0 U/3 T/33 E/36
7 0.57 1 1 0 0 0 0 1 0 0 0 0 0 U/2 F/30 E/40
8 2.75 1 1 0 0 0 0 1 1 0 0 1 1 E/2 F/22 E/48
9 2 0 0 0 0 0 1 0 0 0 0 0 0 U/2 T/21 E/49
10 3.6 0 0 0 0 0 1 0 0 0 0 0 1 E/2 S/10 U/2 E/46 U/12
11 0.69 1 0 0 0 0 1 0 0 0 0 1 1 U/1 E/1 F/49 E/12 H/9
12 1.1 0 0 0 0 0 0 1 0 0 1 0 1 E/2 S/36 F/9 E/25
13 1.1 1 1 0 0 0 0 1 0 1 1 1 1 U/2 S/24 H/13 E/12 H/21
14 0.57 0 1 0 0 0 0 1 0 0 1 0 0 U/2 F/20 E/33 U/17
15 2 0 1 0 0 0 0 1 1 0 0 0 1 T/12 U/18 E/8 U/34
```

Als eerder gezegd, een bestand van school/werk-carrières van Noord-Ierse jongeren. Die data kwamen tot stand door jongeren van 22 jaar te vragen wat ze in de 6 jaar sinds het einde van hun leerplicht gedaan hadden. Deze activiteiten werden gecodeerd als een van 6 "toestanden":

Code	Betekenis
E	Werkend (Employed)
U	Werkloos (Unemployed)
T	Beroepsgericht onderwijs (Training)
S	School
F	Voortgezet onderwijs (Further education)
H	Hoger onderwijs (higher education)

Zo ontstonden uit 712 interviews 712 carrières van 6 jaar ofwel 72 maanden. Voor iedere deelnemer onstond zo een carrière, een patroon zoals hier getoond:

$$\text{U/2 F/36 H/34} = \underbrace{\underbrace{UU}_{2}\,\underbrace{FFF\ldots\ldots\ldots FFF}_{36}\,\underbrace{HHH\ldots\ldots\ldots\ldots\ldots H}_{34}}_{72}$$

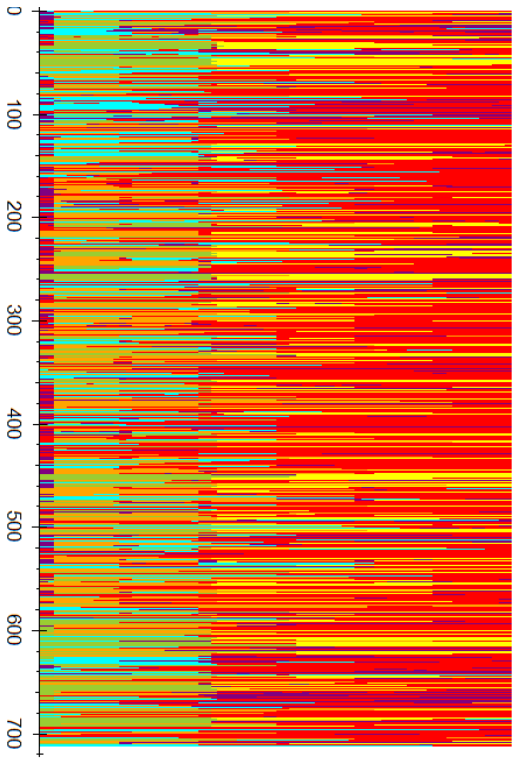
Daarin zien we dat de betreffende jongere eerst 2 maanden werkloos was (U/2), daarna 36 maanden doorbracht in het voortgezet onderwijs (F/36) en tenslotte nog 34 maanden in het hoger onderwijs (H/34). Verder werd informatie verzameld over allerlei achtergrondvariabelen zoals geslacht, religie, de buurt waar ze woonden, het opleidingsniveau, de gezinssamenstelling etc., alles gecodeerd met 0-en en 1-en.

Misschien, zo was de gedachte, zou het mogelijk blijken te voorspellen – op basis van de achtergrondkenmerken van de jongeren – welke jongeren een grote kans lopen in een “mislukt” school/werk-traject terecht te komen. Als dat mogelijk zou blijken zou men kunnen proberen met gericht beleid te voorkomen dat jongeren mislukken bij hun pogingen in de arbeidsmarkt te integreren.
Het doel van dit onderzoek (McVicar & Anyadike-Danes 2002) was nu om na te gaan of met die achtergrondvariabelen het type school/werk-traject kon worden voorspeld.
Het construeren van zo’n typologie van school/werk-trajecten vereist dat we de 712 verschillende trajecten kunnen vergelijken.

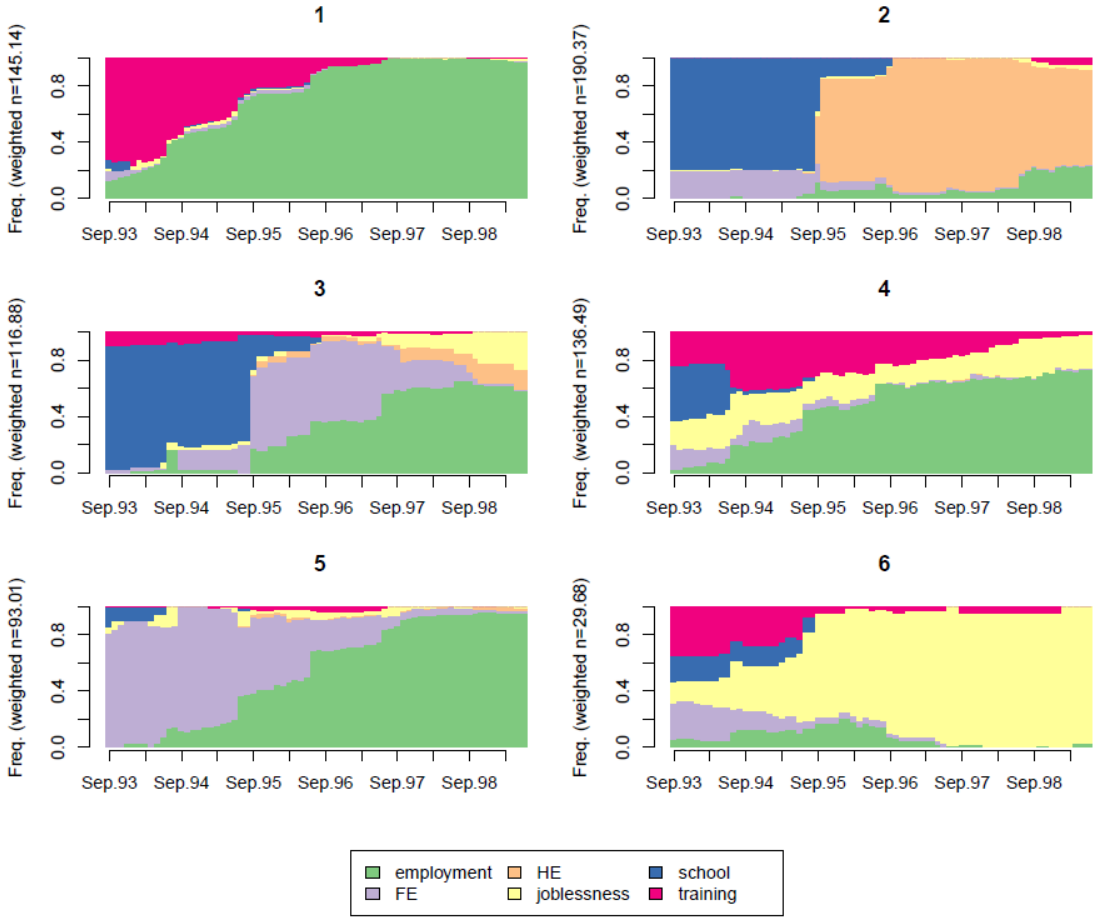
Om u nu uit te leggen wat het resultaat is van zo’n sociaal-wetenschappelijke toepassing van patroonherkenning moet ik u even uitleggen hoe een grafiek van iemands carrière eruit kan zien. Ik geef eerst een eenvoudig, verzonnen voorbeeld en laat u dan dat soort grafiekjes zien die betrekking hebben op de school-werk carrières van de Noord-Ierse jongeren.

In dit plaatje hebben de kleurtjes betrekking op verschillende activiteiten: een blauw staafje betekent dat de betreffende persoon op school zat en de lengte van het blauwe staafje staat voor de tijd. Hoe langer het staafje, hoe langer op school. Zo zien we de totale carrière als een lange staaf met verschillende kleurtjes.

Vervolgens heb ik de staafjes van 3 denkbeeldige jongeren gestapeld. Zo kunnen we een een beeld krijgen van de carrières van 3 jongeren. We kunnen natuurlijk ook veel meer dan 3 van die staafjes stapelen.
Dat heb ik hier gedaan:



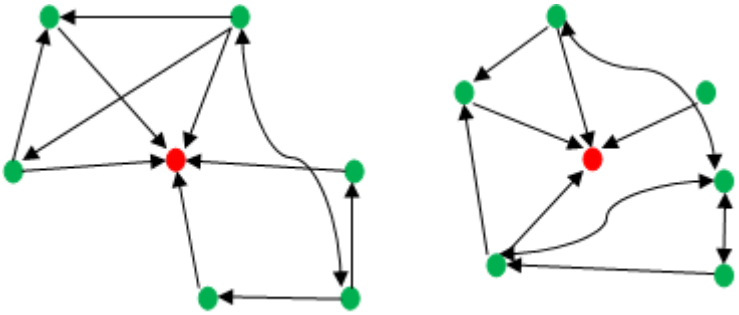
Daar is niet zo heel veel aan te zien, het is een grote warboel van kleurtjes. Met behulp van patroonherkenning kunnen we nu proberen die 712 carrières in een beperkt aantal groepen in te delen zodanig dat carrières binnen groepen meer op elkaar lijken dan carrières tussen groepen
Het resultaat daarvan is precies wat we in het volgend plaatje (Studer, 2012) laten zien: 6 groepen, 6 typen school-werk carrières van in totaal 712 jongeren



Binnen iedere groep is de gelijkenis tussen de jongeren veel groter dan tussen die van jongeren uit verschillende groepen. In Groep 6 vinden we jongeren met een carrière die sterk gedomineerd wordt door werkloosheid en in Groep 1 alle jongeren die direct of na een beroepsgerichte training aan het werk komen. Het resultaat van patroonherkenning in de sociale wetenschappen!
Precies met dit soort toepassingen ben ik in de afgelopen jaren actief geweest, samen met collega's Aart Liefbroer (Elzinga & Liefbroer 2007, Liefbroer & Elzinga 2012), Hilde Bras (Bras et al 2010), Irma Reci, Anna Manzoni (Mooi-Reci et al 2012) en Raffaella Piccarreta (Piccarreta & Elzinga 2013).
Toepassingen van dezelfde wiskunde hebben we ook gezien in de gedragsbiologie, in de meteorologie (Pakalapati et al 2009), in de logistiek (Wilson 2008), in de technologie van bio-electronica (micro-arrays) (Elzinga et al 2008) en in de besliskunde (Elzinga et al 2011, Elzinga & Wang 2012).

Ik zou u gemakkelijk nog veel meer toepassingen kunnen laten zien waar mijn collega's en ik de afgelopen jaren aan gewerkt hebben. Ik volsta met u te melden dat ik nu met Marjolein Broese van Groenou bezig ben met het bestuderen van "zorg-carrières" van ouderen. Ongetwijfeld gaat dat leiden tot nieuwe inzichten in het welzijn van ouderen en de daarmee gepaard gaande kosten van zorg.

Ik wil deze korte inleiding in de sociaal-wetenschappelijke patroonherkenning besluiten met een nieuwe toepassing waarvoor ik onlangs samen met collega's in Belfast en in China nieuwe methoden heb ontwikkeld. Om die toepassing uit te leggen kijken we even naar het volgende plaatje



De twee figuurtjes zijn schematische voorstellingen van sociale netwerken. U mag daarbij denken aan zorgnetwerken rondom ouderen of samenwerkingsverbanden tussen bedrijven. Ieder bolletje stelt een rol, een speler voor en ieder pijltje een relatie. Die rollen en pijltjes zijn zeker niet allemaal hetzelfde maar terwille van de eenvoud van de plaatjes heb ik dat maar niet aangegeven.
Sommige netwerken zijn succesvol en andere niet. Sommige zorgnetwerken, bevolkt door dezelfde rollen zoals verpleegkundige, huisarts, kinderen, thuishulp, etc., slagen erin de ouderen lang, tegen lage kosten en in betrekkelijk welzijn thuis te laten wonen; andere netwerken lukt dat veel minder goed. Hoe komt dat? Ligt dat misschien aan de structuur van de netwerken of aan de aard of intensiteit van de relaties? En waarvan zijn die structuren afhankelijk? Van religie misschien, of van het opleidingsniveau van de zorgontvangers? Een analoge problematiek speelt bij netwerken rond jongeren met gedrags- en aanpassingsproblemen of netwerken van bedrijven die elkaars klanten en leveranciers zijn.
Als we dat soort vragen willen beantwoorden moeten we in staat zijn om netwerkstructuren te vergelijken. Dat is een veel moeilijker probleem dan het vergelijken van rijen symbolen en het probleem wordt al vele tientallen jaren in tal van wetenschappen bestudeerd, o.a. in de chemie voor gelijkenis tussen moleculen, in de logistiek voor optimalisering van routering, in de electronica voor het ontwerpen van schakelingen en in de medische wetenschap voor het vergelijken van radiologische beelden.
Samen met collega Hui Wang heb ik de afgelopen tijd nieuwe methoden bedacht om netwerken te vergelijken. Methoden die nauwkeuriger en veel sneller zijn dan de standaard-techniek. Om dat te illustreren citeer ik uit eigen werk een tabelletje:

	% precisie	snelheid
Subgraphs Isomorphism	53.9	4h 16’ 11”

Dit is wat we met een bekend test-bestand en moderne software, “stand techniek”, kunnen bereiken. Als we echter de door ons ontwikkelde methode gebruiken zien we het volgende resultaat:

	% precisie	snelheid
Subgraphs Isomorphism	53.9	4h 16’ 11”
Paths counting	65.2	3’ 21’’

Het tabelletje laat zien dat we met deze nieuwe methoden veel preciezer kunnen vergelijken en dat we tòch 50 à 100 keer sneller kunnen rekenen. Dat laatste is vooral voor chemici interessant omdat daar de moleculen, c.q. de netwerken zo groot zijn.
Op dit moment ben ik bezig om samen met Eric Widmer en Marlène Sapin van de universiteit van Genève onderzoek te doen waarin we gezinsnetwerken van verstandelijk beperkten vergelijken om hun sociale kwetsbaarheid te beter te begrijpen. Het aardige is nu dat dit type onderzoek veel lijkt op onderzoek aan zorgnetwerken voor ouderen.
Ik hoop dan ook dat ik in de nabije toekomst aan zorgnetwerken kan gaan werken met Marjolein Broese en Theo van Tilburg van mijn eigen afdeling. In de iets verdere toekomst hoop ik met de groep in Genève onder leiding van Eric Widmer, met de groep van Marjolein

in onze eigen afdeling en met de groep van Emily Grundy in Cambridge een consortium te vormen waarmee we op Europees niveau financiering kunnen vinden voor grootschalig vergelijkend onderzoek aan familie- en zorg-netwerken. Die grootschaligheid is nodig vanwege de veelvormigheid van sociale netwerken, het internationaal vergelijkende is nodig vanwege de invloed van cultuur- en zorgstelselverschillen op het functioneren van die sociale netwerken. Ook buiten mijn eigen afdeling is er op deze terreinen kennis genoeg binnen onze Faculteit en in breder VU-verband; ik hoop dat we die in het kader van het Talma-instituut stevig kunnen verankeren.

Dames en heren, ik heb u aan de hand van veel voorbeelden en plaatjes uitgelegd waar mijn leerstoel over gaat en wat ik daarmee wil:

- verder ontwikkelen van methoden voor discrete patroonherkenning en werken aan toepassingen in de sociale wetenschappen,
- internationale samenwerking opzetten om fondsen te verwerven voor grootschalig onderzoek aan zorg- en familie-netwerken,

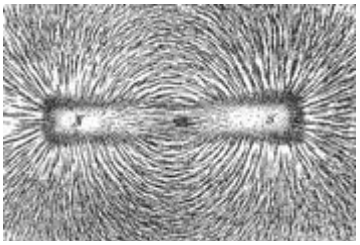
Wat ik nog niet heb uitgelegd is waarom ik het motto van deze oratie heb geformuleerd als “Tussen Tellen en Toetsen”. Dat ga ik nu proberen.
Om een en ander uit te leggen moeten we het eerst even hebben over het doel van wetenschap.

Het is belangrijk dat u dat doel goed voor ogen heeft dus ik heb dat maar even voor u op het bord gezet: opdat wij niet meer bang zijn in het donker. Daarmee bedoel ik dat wetenschap tot doel heeft onze wereld op een heldere manier te beschrijven en alle verschijnselen en samenhangen daarin een plaats te geven. De vraag is natuurlijk hoe dat moet, dat ordenen, dat classificeren.
In onze faculteit vinden we daar twee opvattingen over. Dat is trouwens ook een classificatie en niet zo’n fijnzinnige bovendien.
In de eerste plaats zijn er de collega’s die menen dat kennis, classificatie, ordening, ontstaat door het waarnemen, het meten van de wereld. Voor hen is het probleem te beslissen bij welk wereldbeeld die waarnemingen het beste passen. Daartoe stellen ze dan hypothesen op - veronderstellingen over hoe de wereld in elkaar steekt - en ze beslissen, al dan niet met behulp van de statistiek, of hun gegevens strijdig zijn met de hypothesen. Als dat zo is gaan ze hun wereldbeeld aanpassen en leiden daaruit weer nieuwe hypothesen af. Zo zijn ze constant bezig om hun hypothesen, hun wereldbeeld, te toetsen. Dit denken over kennis stamt al uit de 17de eeuw. Een n bekende representant van die opvatting is Isaac Newton, de “uitvinder” van de zwaartekracht en grondlegger van de moderne optica. Ik zeg wel

“uitvinder” maar ik moet eigenlijk zeggen: de “ontdekker”. In Newton’s gedachten bestonden al die zaken al sinds de schepping; hij “ontsluierde” ze slechts.

“Hypotheses non Fingo”

Met zijn beroemde uitspraak “Hypotheses non fingo” (“Ik verzin geen hypothesen”) bedoelde hij waarschijnlijk te zeggen dat al zijn beweringen te herleiden waren tot waarnemingen. Impliciet óók, dat zijn waarnemingen en wetten betrekking hadden op een zelfstandig bestaande realiteit. Voor Newton viel wetenschap dus samen met het zo eenvoudig mogelijk beschrijven van die op zichzelf bestaande werkelijkheid. Sommige collega’s zullen nu willen aanvoeren dat ze ook zoeken naar “oorzaken” voor hun observaties ze bedoelen dan dat ze hun wereldbeeld, hun beschrijving gewoon zo simpel mogelijk maken. Problematisch is natuurlijk dat zo helemaal niet duidelijk wordt waar toch die hypothesen, die wereldbeelden vandaan komen.
Voor sociale wetenschappers is het probleem met deze opvatting dat ze sterk geënt is op de klassieke natuurkunde waarin naar believen herhaald kon worden. Dat herhalen was belangrijk want zo kon je controleren of de waarnemingen wel correct gedaan waren.



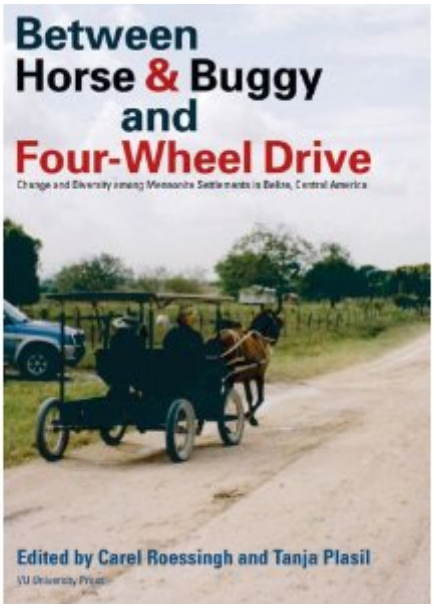
Velen van ons kennen nog het klassieke proefje uit onze natuurkundelessen: een beetje ijzervijlsel, een magneet erbij en je “ziet” het magnetisch veld. Dat proefje kun je eindeloos herhalen want alle magneten en ijzerdeeltjes gedragen zich hetzelfde en dat blijven ze ook doen, ongeacht de proefjes die we met ze uithalen!
Nog steeds is herhaalbaarheid van observaties van groot belang voor allen die hechten aan het primaat van de 17de - en 18de-eeuwse natuurkunde en dat is ook wel begrijpelijk, want die natuurkunde heeft een geweldige invloed gehad op ons dagelijks leven, op de techniek die we dagelijks om ons heen zien en zelf gebruiken. Die natuurkunde was een daverend succes! Het is dan ook geen wonder dat het kennismodel van de natuurkunde alle andere wetenschappen beïnvloed heeft.

Veel verschijnselen die sociale wetenschappers bestuderen, herhalen zich niet zoals de ijzerdeeltjes iedere keer weer netjes op hun plaats gaan liggen als je er een magneet bij houdt. Denkt u maar eens aan de beurscrises van 1929 en die van 2008.

Twee beurscrises maar met grotendeels verschillende oorzaken en verschillende gevolgen. Zo zien we in de economie en de geschiedenis, in de antropologie en ook in de sociologie heel veel gebeurtenissen en verschijnselen die éénmalig lijken te zijn. En als gebeurtenissen eenmalig zijn, is het moeilijk hun oorzaak ondubbelzinnig vast te stellen. We kunnen immers zo'n verklaring, zo'n theorie niet testen want de gebeurtenis herhaalt zich niet of doet zich voor onder totaal verschillende omstandigheden.

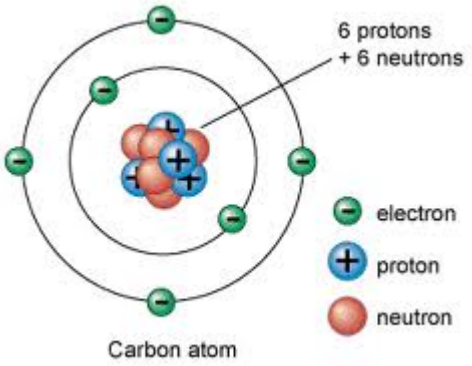
Een tweede voorbeeld is het huwelijk. We kunnen constateren dat mannen aan het eind van de 19de eeuw later trouwden dan rond 1980. Maar wat betekent zo'n observatie als we ons realiseren dat een huwelijk in 1880 voor de betrokkenen, voor de families en voor de samenleving een geheel andere betekenis had dan in 2012. Anders in vrijwel alle opzichten: juridisch, economisch, sociaal en emotioneel. Het huwelijk is sinds 1880 totaal verbouwd,

iets geheel anders geworden doordat *wij* het anders zijn gaan interpreteren (taal) en er anders mee omgaan (handelen).



Een laatste voorbeeld is vervat in een prachtig boekje van mijn collega Carel Roessingh. Het boekje beschrijft de gemeenschap van Mennonieten in Belize, een klein staatje ingeklemd tussen Mexico, Guatemala en de Caraïbische Zee. Het onderzoek waarop het boekje is gebaseerd vond plaats in de periode 2004-2007 en het **kan niet meer opnieuw** gedaan worden: Belize is veranderd, informanten zijn dood, verhuisd of tenminste zelf veranderd; die wereld bestaat niet meer. Maar Carel en zijn collega's hebben voor ons een soort foto gemaakt door hun persoonlijke lens die inmiddels ook al weer gedateerd is. Is het boekje daarom fictie, is het daarom geen waardevolle bijdrage aan ons weten, aan ons begrip van hoe een complexe samenleving zich kan ontwikkelen? Ik dacht het niet!

Zo is er een tweede groep collega's die een ander kennismodel hanteert dan dat van de klassieke natuurkunde. Want voor de toch ernstige en gecompliceerde problemen die zij bestuderen, voor de vragen die zij stellen, is dat klassieke denken nu eenmaal niet goed bruikbaar. Zij begrijpen dat niet-herhaalbare, dat unieke in onszelf en onze samenleving, via onze taal en via ons individueel en collectief handelen. Ook deze opvatting van kennis is niet uniek voor de sociale wetenschappen maar werd al fraai verwoord door Niels Bohr, de grote Deense natuurkundige en filosoof uit de eerste helft van de vorige eeuw, toen de klassieke, 17de -eeuwse natuurkunde was stukgelopen op de interpretatie van de toen nog nieuwe quantummechanica. Ik heb die uitspraak voor u uit het Engels vertaald en het woord "Natuurkunde (Physics)" vervangen door het woord "Wetenschap".



“Wetenschap is niet zozeer het onderzoek van iets dat al buiten ons om bestond maar veeleer de ontwikkeling van het samenvatten en ordenen van de menselijke ervaring.”

Vrij naar Niels Bohr (1885 - 1962)

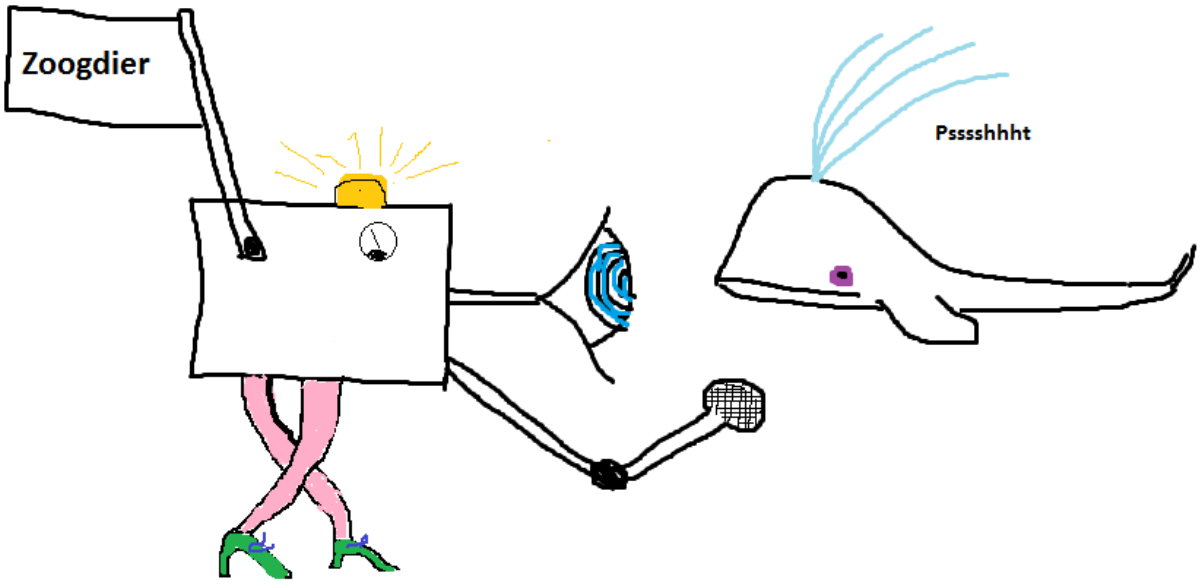
“Samenvatten en ordenen”, dat is een veel ruimere en tegelijk veel bescheidener opvatting van wat wij kunnen kennen en hoe we tot kennis kunnen komen dan de, overigens begrijpelijke, hoogmoed van het 17de-eeuwse “Hypotheses non fingo”. In het samenvatten en ordenen is ook en terecht plaats voor het beschrijven, exploreren en interpreteren. Wie denkt dat het anders is, leze Huizinga’s “Herfsttij der Middeleeuwen” of Ginzburg’s “The Cheese and the Worms”.

Waarom vertel ik u dit allemaal?

Om u duidelijk te maken wat de betekenis is van de alliteratie “Tussen Tellen en Toetsen”.
Tellen is een tamelijk elementaire, beschrijvende handeling, vaak met een exploratief doel. De uitkomst van tellen kan een getal zijn, “273”, en daar houden veel sociale wetenschappers niet van, maar ook een kwalitatieve uitdrukking zoals “veel”, “vaak” of “zelden” en dat mag dan weer wel, naar het schijnt.
Toetsen daarentegen is helemaal niet beschrijvend maar evaluerend: hypothesen opstellen en toetsen hoort in het 17de-eeuws model van wetenschap bedrijven.
In Bohr’s opvatting is er ruimte geschapen voor een breed scala aan benaderingen van sociale verschijnselen. Van heel kwantitatief, model-toetsend onderzoek als bijvoorbeeld van Marja Aartsen naar cognitief functioneren van ouderen, via beschrijvend onderzoek naar Mennonieten in Belize, langs interpretatieve analyse zoals bijvoorbeeld van Kathy Davis en Lorrain Nencel van de vraag waar mensen zich thuis voelen tot, helemaal aan het andere einde van het spectrum, de erudiete speculatie van kunsthistorici over het grotendeels verdwenen oeuvre van Robert Campin.

Je met patroonherkenning bezighouden en technieken uit dat vak gebruiken in toepassingen, betekent enerzijds dat je bezig bent met beschrijving en exploratie van data en anderzijds, maar niet noodzakelijk, met het toetsen of die gevonden patronen lijken op de patronen die je in gedachten had. Patroonherkenning staat vaak precies tussen een louter beschrijvende, interpretatieve aanpak en hypothese-toetsing. Met patroonherkenning vinden we vaak de hypothesen die bij sommigen uit het niets tevoorschijn floepen maar die anderen helemaal niet lijken te willen hebben. Patroonherkenning is een kwalitatieve bezigheid: we genereren kwalitatieve uitspraken, labels, maar we doen dat met veel rekengeweld. Zo sta ik midden tussen al mijn geachte collega's, tussen kwalitatief en kwantitatief, tussen tellen en toetsen.

Ook wel als een horlogemaker die vaak de tijd vergeet.



Dames en heren, ik heb u nu verteld wat ik doe en waarom en ik heb mijzelf een plaats verklaard temidden van mijn collega's in de faculteit. Het wordt nu dus tijd om te bedanken, eerst en vooral u allen voor uw komst en geduld.

Meer in het bijzonder wil ik onze decaan Anton Hemerijck, het College van Bestuur en de leden van de Benoemingsadviescomissie bedanken voor hun instemming.

Lorrain, Ineke, Gerhard, Fleur, Els en Jacqueline; dank voor jullie steun, jullie wijze raad en jullie tolerantie voor mijn veelvuldig feilen in het leiden van een afdeling. Josien, steeds buitengewoon lid van mijn MT's, dank voor je openheid en discretie, dank voor je steun in precaire zaken. Harry, Bert, Theo, Marjolein en Halleh, dank voor jullie vertrouwen. Mareanne, zakelijk en met humor recht door zee, dank daarvoor. Ietje, dank voor je formidabele loyaliteit en inzet.

Lieve vrienden, dank dat jullie me al zo veel jaren hebt willen volgen langs mijn soms wat kronkelige pad naar deze bijzondere dag.

Lieve Martje, jou kan ik hier niet bedanken want daarvoor is de tijd te kort.

Mijnheer de Rector, ik heb gezegd.

Bibliografie

Bohr, Niels (2010 (1958)) *Atomic Physics and Human Knowledge (Dover Books on Physics)* Dover Publications, Mineola (NY).

Bras, Hilde and Liefbroer, Aart C. en Elzinga, Cees H. (2010) Standardization of Pathways to Adulthood? An Analysis of Dutch Cohorts Born Between 1850 and 1900, *Demography*, 47(**4**), 1013-1034.

Buis, Maarten (2010) *Inequality of Educational Outcome and Inequality of Educational Opportunity in the Netherlands during the 20th Century*. Thesis, VU University Press.

Davis, Kathy en Nencel, Lorrain (2011) Border skirmishes and the question of belonging: An auto-ethnographic account of everyday exclusion in multicultural society, *Ethnicities*, 11(4), 467-488.

Day, Mark, (2008) *The Philosophy of History*, Continuum International Publishing Group, London.

Elzinga, Cees H. en Liefbroer, Aart C. (2007) De-Standardization and Differentiation of Family Life Trajectories of Young Adults: A Cross-National Comparison Using Sequence Analysis, *European Journal of Population*, 23(3-4), 225-250.

Elzinga, Cees H., Rahmann, Sven en Wang, Hui (2008) Algorithms for Subsequence Combinatorics, *Theoretical Computer Science*, 409(**3**), 394-404.

Elzinga, Cees H., Wang, Hui, Lin, Zhiwei en Kumar, Yash (2011) Concordance and Consensus, *Infomation Sciences*, 181, 2529-2549.

Elzinga, Cees en Wang, Hui (2012) Kernels for Acyclic Digraphs, *Pattern Recognition Letters*, 33(**16**), 2239-2244.

Elzinga, Cees H. en Wang, Hui (2012) Versatile String Kernels (under review).

Ginzburg, Carlo (1976) *The Cheese and the Worms. The cosmos of a sixteenth-century miller.* Penguin Books, London.

Huizinga, Johan (1919) *Herfsttij der Middeleeuwen. Studie over levens- en gedachtenvorming der veertiende en vijtiende eeuw in Frankrijk en de Nederlanden*. Tjeenk Willink, Haarlem.

Legendre, Frederic en Robillard, Tony en Desutter-Grandcolas, Laureen Whiting, Michael F. en Grandcolas, Philippe (2008) Phylogenetic Analysis of Non-Stereotyped Behavioural Sequences with a Successive Event-Pairing Method, *Biological Journal of the Linnean Society,* 94, 853-867.

Liefbroer, Aart C. en Elzinga, Cees H. (2012) Intergenerational Transmission of Behavioral Patterns: How Similar are Parents' and Children's Demographic Trajectories, *Advances in Life Course Research*, 17(1), 1-10.

Nuechterlein, Jeanne (2004) In search of artistic personality: the case of Robert Campin, *Art History*, 27(2), 312-320.

McVicar, Duncan en Anyadike-Danes, Michael (2002) Predicting Successful and Unsuccessful Transitions from School to Work by using Sequence Methods, *Journal of the Royal Statistical Society. Series A*, **165**(2), 317-334.

Mooi-Reci, Irma en Manzoni, Anna en Elzinga, Cees H. (2012) The Career Disadvantage of Unemployment: Cumulating, Persisting or Accelerating? (under review).

Pakalapati, Swathi, en Beaver, Scott en Romagnoli, Jose A. en Palazoglu, Ahmet (2009) Sequencing Diurnal Air Flow Patterns for Ozone Exposure Assessment around Houston, Texas *, Atmospheric Environment*, **43**(3), 715-723.

Piccarreta, Raffaella en Elzinga, Cees H. (2013) Mining for Associations between Life Course Domains, in McArdle, J.J en Ritschard, G. (Eds) *Contemporary Issues in Exploratory Data Mining (Quantitative Methods Series)*, Chpt. 8, Routledge.

Roessingh, Carel (2009) en Plasil, Tanja (Eds.) (2009). *Between Horse & Buggy and Four-Wheel Drive: Change and Diversity among Mennonite Settlements in Belize, Central America*. Amsterdam, VU University Press.

Studer, Matthias (2012) Étude des inégalités de genre au début de carrière académique à l'aide de méthodes innovatrices d'ánalyse de données séquentielles. Thèse, présentée à la Faculté des sciences économiques et sociales de l'Université de Genève.

Wilson, Clarke (2008) Activity Patterns in Space and Time: Calculating Representative Hagestrand Trajectories, *Transportation*, 35(4), 485-499.