Analytics Master Class
-more exotic patterns in data

It’s all about me…

Prof. Mark Whitehorn
Emeritus Professor of Analytics
School of Computing
University of Dundee

Consultant
Writer (author)
m.a.f.whitehorn@dundee.ac.uk

It’s all about me…

School of Computing

Teach Masters in:

Data Science

Part time - aimed at existing data professionals

Data Engineering

Agenda

- Dark Data
- Probability calculations
- RFI

and anything else for which we have time

Agenda

These should not be seen as individual, disconnected techniques/approaches

Rather they are synergistic

Agenda

These should not be seen as individual, disconnected techniques/approaches

Rather they are synergistic

(although some connections are simply serendipitous)

Dark Data

“Quite so,” he answered, lighting a cigarette, and throwing himself down into an armchair. “You see, but you do not observe. The distinction is clear.”

Who is describing whom?

Dark Data

“Quite so,” he answered, lighting a cigarette, and throwing himself down into an armchair. “You see, but you do not observe. The distinction is clear.”

Who is describing whom?

Sherlock Holmes describing Dr John Watson

A Scandal in Bohemia, Sir Arthur Conan Doyle

Dark Data

Companies currently collect vast swathes of data; some is analysed, much isn't. The data that isn't analysed is Dark. Clearly the origin of the term is 'Dark Matter'.

Donald Rumsfeld told us about:

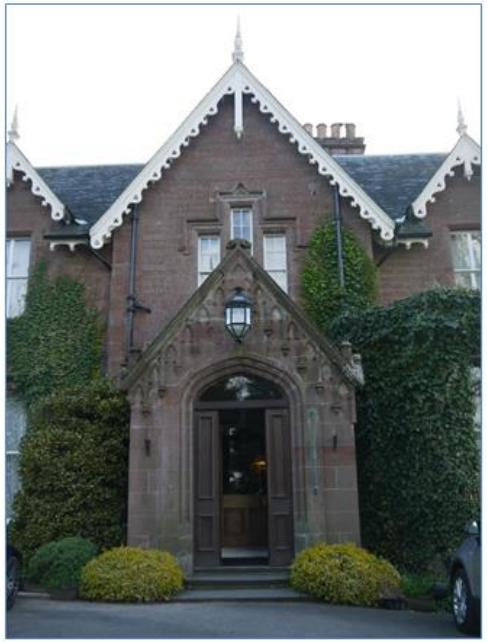
Known knowns

Known unknowns

Unknown unknowns

Dark Data

First usage, modern sense, was at the 2013 Lands of Loyal conference.



Dark Data

First mention:

"Shedding Light on the Dark Data in the Long Tail of Science", P. Bryan Heidorn, LIBRARY TRENDS, Vol. 57, No. 2, Fall 2008 ("Institutional Repositories: Current State and Future," edited by Sarah L. Shreeves and Melissa H. Cragin), pp. 280–299

Heidorn writes that:

'*"Dark data" is not carefully indexed and stored so it becomes nearly invisible to scientists and other potential users and therefore is more likely to remain underutilized and eventually lost.*'

Dark Data

I think we have three identifiable flavours:

Dark Data 1.0 – The data that is in the transactional systems but never makes it to the analytical – we see it but we don't observe

DD 2.0 – Data that is published but which the publisher does not expect to be analysed

DD 3.0 – Data that vanishes and the disappearance is highly significant

Dark Data of the first kind

Is all around us and I believe it is very important.

In the future, we are going to be moving a great deal of this data into analytical systems.

Dark Data of the first kind

The Andria Doria

Dark Data of the first kind

The Andria Doria

Collision between the *Andrea Doria* and the *Stockholm*.

Worst maritime disaster to occur in US waters

25 July 1956.

The Boston Daily Globe

2 Liners Crash Off Nantucket

Andrea Doria Abandoned; Hundreds Are Picked Up; Stockholm's Bow Crushed; Fog Hampers Rescue

$50 Million Project Periled by Toll Road

Over 2500 Passengers on Stricken Vessels

Ile de France at Scene; 3 Injured on Stockholm

$3.7 Billion Compromise On Foreign Aid Wins OK

Nixon-vs-Herter Choice Put Up to Ike

Salem Man Dies, 7 Hurt in Fall O'N.H. Chair Lift

Passengers On Doria

NHRR Surveying 39 Lines; Seeks to Drop Losing Ones

Exchange of Radio Calls

Image from:
http://anotheroldmovieblog.blogspot.co.uk/2010/08/ruth-roman-betsy-drake-and-ss-andrea.html

Dark Data of the first kind

The Andria Doria

Officers on board the *Andrea Doria* did not use the plotting equipment that was available to them to calculate the speed and position of the *Stockholm*.

The navigation officer on board the *Stockholm* thought his radar was set to 15 miles when in fact it was set to 5 miles.

Dark Data of the first kind

The Andria Doria

Both had radar running and yet they collided.

Dark Data of the first kind

The Andria Doria Effect

Bats flying into nets in caves

Griffin, Donald R. *

**Listening in the dark: the acoustic orientation of bats and men*. New Haven, Yale University Press, 1958. 415 p.

Dark Data of the first kind

Bats

"For example, Griffin (1958) describes bats, returning to the roost after an evening's hunt, crashing into a newly erected cave barricade even though it reflected strong echoes. Griffin coins this mishap the "Andrea Doria Effect," because the bats seemed to ignore important information from echo returns to guide their behavior and instead favored spatial memory."

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2927269/

Dark Data of the first kind

You are unlikely to collect data or information from or about bats and/or ocean-going liners.

But your organisation is likely to collect data that is not being analysed (in the sense that the information it contains is not being extracted).

- Patient monitors in intensive care
- Robots in car production

Dark Data of the second kind

DD 2.0 – Data that is published but which the publisher does not expect to be analysed

Data leakage

Amazon publishes the position of a book relative to other books. 1 is good (think J. K. Rowlings).

I collected the position of my books for several years.

Dark Data of the second kind

Amazon positional data (IRDB) for Aug 98 – Mar 00

Dark Data of the second kind

Amazon positional data (IRDB) for Aug 98 – Mar 00

Dark Data of the second kind

I also had the sales figures (number of books sold) from my publisher. I could correlate these with the Amazon position figures.

I also tracked the Amazon position of ‘competing’ books.

So…..

Dark Data of the second kind

I also had the sales figures (number of books sold) from my publisher. I could correlate these with the Amazon position figures.

I also tracked the Amazon position of ‘competing’ books.

So I could estimate the sales figures of my competitors.

Dark Data of the second kind

The estimation of the maximum of a discrete uniform distribution by sampling (without replacement).

Dark Data of the second kind

The estimation of the maximum of a discrete uniform distribution by sampling (without replacement).

A superb example of this is:

The German Tank problem

Dark Data of the second kind

The German Tank problem

The allies were **very** interested in how many tanks the Germans were producing during the second World War.

Intelligence suggested the production was about 1,000 - 1,500 tanks per month.

Dark Data of the second kind

The German Tank problem

The allies started taking the serial numbers from the captured/disabled German tanks.

It turned out that the gearbox numbers were the most useful. (However the chassis numbers and wheel moulds were also used.)

These numbers were also used to estimate production.

Dark Data of the second kind

The German Tank problem

As an example, in June 1941, intelligence suggested tank production was 1,550.

The estimate from serial numbers was 244.

The actual value (verified after the war) was 271.

In August 1942 the numbers were 1,550, 327, 342.

It worked.

Dark Data of the second kind

So how is it done?

Suppose (simplistically) that the actual numbers run from 1 to n.

If we see a number like 24 then at least 24 tanks have been produced.

But we can be smarter. If we see numbers like:

6, 7, 9, 11, 12, 16, 17, 23, 24

then an intelligent guess would be around 30-40.

Dark Data of the second kind

So how is it done?

If we capture 5 tanks (we'll call this value C)

432, 435, 1542, 2187, 3245

then the guess has to be higher than 3,245 – perhaps 4,000.

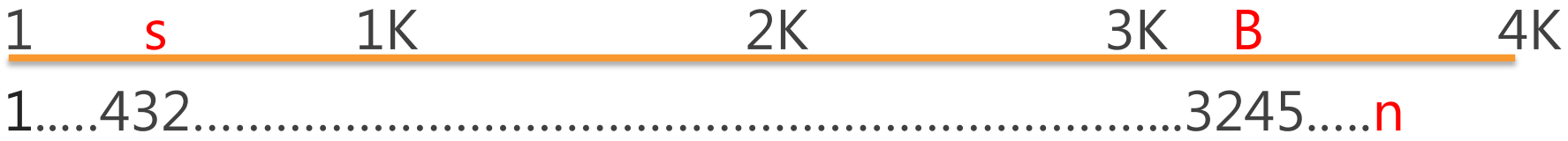
But we can apply logic and a formula.

The biggest number (B) is 3,245, and the smallest (s) is 432. We could reasonably expect as many 'hidden' numbers above 3,245 as there are below 432.

Dark Data of the second kind

So how is it done?

C=5, B=3,245, s=432, n=number of tanks produced



So n is about 3,245 + 432 = 3,677

Dark Data of the second kind

You can also use:

$n = (1+1/C)B$

$n = (1 +1/5)3,245 = 3,894$*

* Roger Johnson of Carleton College in Northfield, Minnesota
Teaching Statistics Vol. 16, number 2 Summer 1994
http://www.mcs.sdsmt.edu/rwjohnso/html/tank.pdf

so the above sum becomes

$$\mathrm{E}(X_{(n)}) = \frac{n}{\mathrm{C}(N,n)} \sum_{k=n}^{N} \mathrm{C}(k,n). \qquad (2)$$

To simplify this, note that the probabilities in (1) must sum to 1. So

$$\sum_{j=n}^{N} \mathrm{C}(j-1,n-1) = \mathrm{C}(N,n)$$

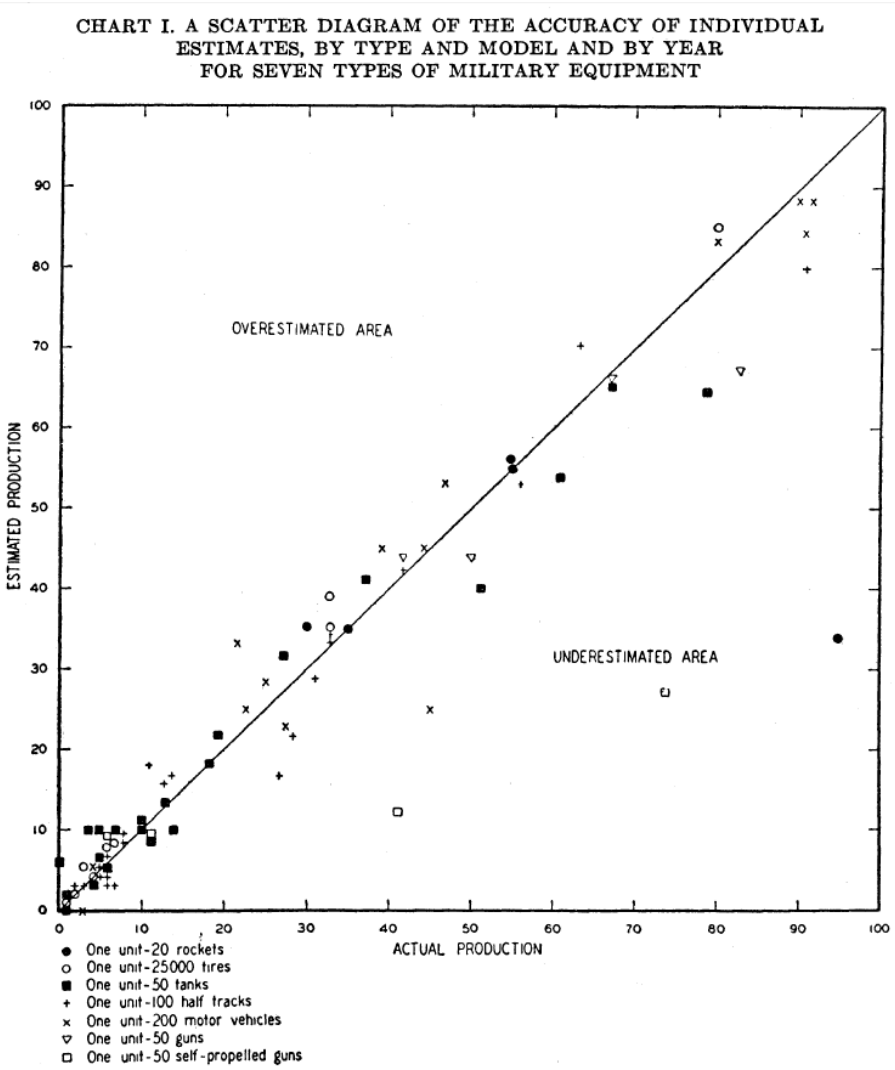
which implies

$$\sum_{k=n}^{N} \mathrm{C}(k,n) = \mathrm{C}(N+1,n+1). \qquad (3)$$

Combining (2) and (3) and going through some simple algebra we see

We close with an anecdote of Colonel Trevor Dupuy (1991). "In the Middle East a few years ago I was given permission by Israeli military authorities to go through the entire Merkava Tank production line. At one time I asked how many Merkavas had been produced, and I was told that this information was classified. I found it amusing, because there was a serial number on each tank chassis."

An Empirical Approach to Economic Intelligence in World War II Author(s): Richard Ruggles and Henry Brodie Source: Journal of the American Statistical Association, Vol. 42, No. 237 (Mar., 1947), pp. 72- 91 Published by: American Statistical Association Stable URL: http://www.jstor.org/stable/2280189



Dark Data of the second kind

Other uses

The serial number data was also used to deduce the:

- number of tank factories and their relative production
- logistics of tank deployment
- changes in production due to, for example, bombing

Dark Data of the second kind

Of course, if the Germans had been aware of this they could have:

- Coded the serial numbers, perhaps using a random surrogate key (boring)
- Created misinformation (a much more interesting option)

Dark Data of the second kind

Again, don’t think of this technique as being limited to tanks.

What data is your company leaking?

What can you do about it?

Which of your competitors is leaking?

(satellite photographs of oppositions parking lot)

(classis car advertisements)

How many complaints does a shipping company receive?

From: DPD Customer Services [mailto:customerservices@dpd.co.uk]
Sent: 20 August 2014 13:12

We've Got Your Email!

Thanks for getting in touch with us. The Consumer team have got your email and will get straight onto it.

We're here Monday to Friday from 08.00 until 18.30 and Weekends from 09.00 until 13.00 and you'll have our response within 24 hours.

Many thanks,

The DPD Consumer Team

Case Reference : 01366108

From: DPD Customer Services [mailto:customerservices@dpd.co.uk]
Sent: 20 August 2014 15:22

We've Got Your Email!

Thanks for getting in touch with us. The Consumer team have got your email and will get straight onto it.

We're here Monday to Friday from 08.00 until 18.30 and Weekends from 09.00 until 13.00 and you'll have our response within 24 hours.

Many thanks,

The DPD Consumer Team

Case Reference : 01367517

Let's see.
7517 – 6108 =
1409 in 130 minutes
= about 10 per minute.

From: DPD Customer Services [mailto:customerservices@dpd.co.uk]
Sent: 21 August 2014 09:47

We've Got Your Email!

Thanks for getting in touch with us. The Consumer team have got your email and will get straight onto it.

We're here Monday to Friday from 08.00 until 18.30 and Weekends from 09.00 until 13.00 and you'll have our response within 24 hours.

Many thanks,

The DPD Consumer Team

Case Reference : 01370836

70836 - 67517 = 3,319 in 1,105 minutes
= about 3 per minute; implying the rate drops overnight.

From: DPD Customer Services [mailto:customerservices@dpd.co.uk]
Sent: 21 August 2014 09:48

We've Got Your Email!

Thanks for getting in touch with us. The Consumer team have got your email and will get straight onto it.

We're here Monday to Friday from 08.00 until 18.30 and Weekends from 09.00 until 13.00 and you'll have our response within 24 hours.

Many thanks,

The DPD Consumer Team

Case Reference : 01370845

Quick test.
70845 followed by 70836 one minute later. 9 in one minute and it looks like they are sequential (but much more testing required).

How many complaints does a shipping company receive?

The solution is simple, it is called a surrogate key.

Key	Surrogate	Type	Location
1	23122	Complaint	DD1 4HN
2	43234	Missing	HR5 3QA
3	4566	Complaint	SW1A 4WW
4	5	Delivery	DD5 4WS
5	45544	Delivery	DD3 2WA

Dark Data of the third kind

The German Bomb problem

DD 3.0 – Data that vanishes and the disappearance is highly significant

My mother worked on explosives and casings from German bombs and shells during the war.

At the start of the war the Germans used very good quality alloys. The allies tracked the elements in the alloy. As the war progressed, certain elements disappeared completely from the mix.

Dark Data of the third kind

The German Bomb problem

If the allies had started analysing in the middle of the war, they would have missed the missing data.

(Would that have made it an unknown, unknown, unknown?)

As it was, they could see what elements were becoming scarce in Germany and plan accordingly.

Dark Data of another kind?

Car club details

The members of a vintage car club often want to know the details that their car club has about their car. So, they give the chassis number to the club and it supplies the details.

So why would anyone be interested in the chassis numbers that have never been requested?

Dark Data of another kind?

So it may be worth monitoring the requests for dark data that are made.

Dark Data

Take home message

Collect data and use it imaginatively.

Using data imaginatively

Analysts often get involved in processing log data.

How many of you have done so?

Log Jam

Logs

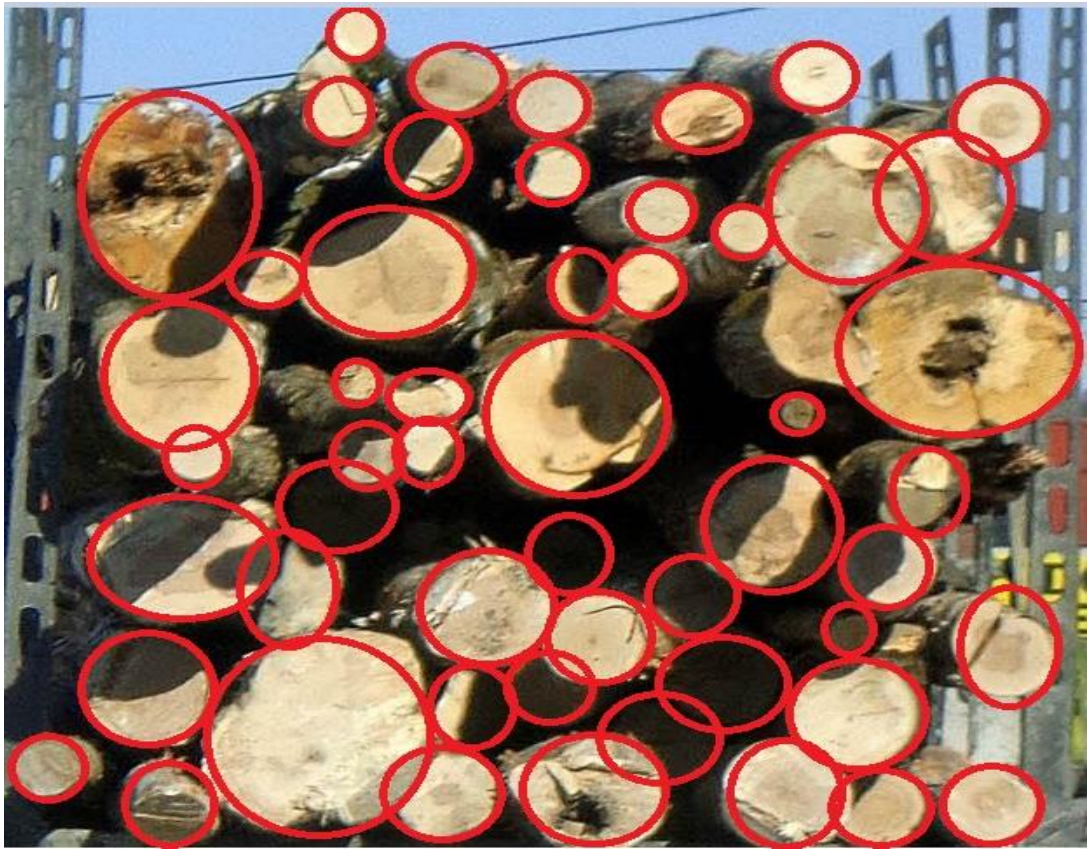
Is this picture big data?
It is if the driver sends the image from his phone to the sawmill.

Logs

Trim
(the picture,
not the
logs).

Logs

Find the logs



Logs

Count (53)

and

measure

diameters

Logs

Driver: Olaf Smithson

Location: Adelsburg

Road distance: 235KM

Journey Time: 6:23

Break Due: 2:30

53 Logs

Load classification: Medium

Schedule: 10:20AM tomorrow

Log No.	Diameter
1	35
2	23
3	68
4	45
5	23
…	…
53	23

Probabilities

Sales people talking to customer

Analytical rules say, not product A, but product B or C

- B Net value £90 – Probability of acceptance 0.54
- C Net value £200 – Probability of acceptance 0.32

Which is the best option (from our point of view) to offer the customer?

Probabilities

Probabilities tell us:

B £90 * 0.54 = £48.6

C £200 * 0.32 = £64.0

The best option is C

When Harry met Sally

Harry and Sally drink at the “Old Row and Column” (a sequel to the pub where people join each other outside at the tables.)

Edgar, the barman, is interested in numbers. He tracks how often Harry and Sally appear.

Probabilities

Harry turns up twice a week - P= 0.3

Sally also turns up twice a week – P= 0.3

How often do they meet?

Probabilities

Harry turns up twice a week - P= 0.3

Sally also turns up twice a week – P= 0.3

How often do they meet? Simply probability calculations tell us it should be about 9% of the time.

	P Harry = 0.3	P Not Harry = 0.7
P Sally = 0.3	0.09	0.21
P Not Sally = 0.7	0.21	0.49

Probabilities

What if Edgar finds that they appear together 40% of the time?

	P Harry = 0.3	P Not Harry = 0.7
P Sally = 0.3	0.09	0.21
P Not Sally = 0.7	0.21	0.49

Probabilities

Perhaps they are walking out together.

Perhaps they tend to appear on Friday and Saturday nights, as do their mutual friends.

	P Harry = 0.3	P Not Harry = 0.7
P Sally = 0.3	0.09	0.21
P Not Sally = 0.7	0.21	0.49

Probabilities

The point is that the initial calculation assumes that the two events (Sally appearing and Harry appearing) are independent (0.3 * 0.3 = 0.09). If the value we get is significantly higher or lower than this, then there may be a dependency between them (there are, of course, other possibilities).

If we are measuring data it can be very useful to calculate the independence (or otherwise) of the variables we have.

Probabilities

This is, of course, the basis of a great deal of:

- Statistical analysis
- Data mining

Probabilities

This is, of course, the basis of a great deal of:

- Statistical analysis
- Data mining

and gossip, obviously

Exercise

Misbehaving dice

Design 4 dice

Each has six sides

Each side has a number between 0 and 6

Dice can have identical numbers on different sides so, for example, 1 2 2 3 3 3 is perfectly possible

Exercise

Die two must, on average, be able to beat die one

Die three must, on average, be able to beat die two

Die four must, on average, be able to beat die three

Exercise

Die two must, on average, be able to beat die one

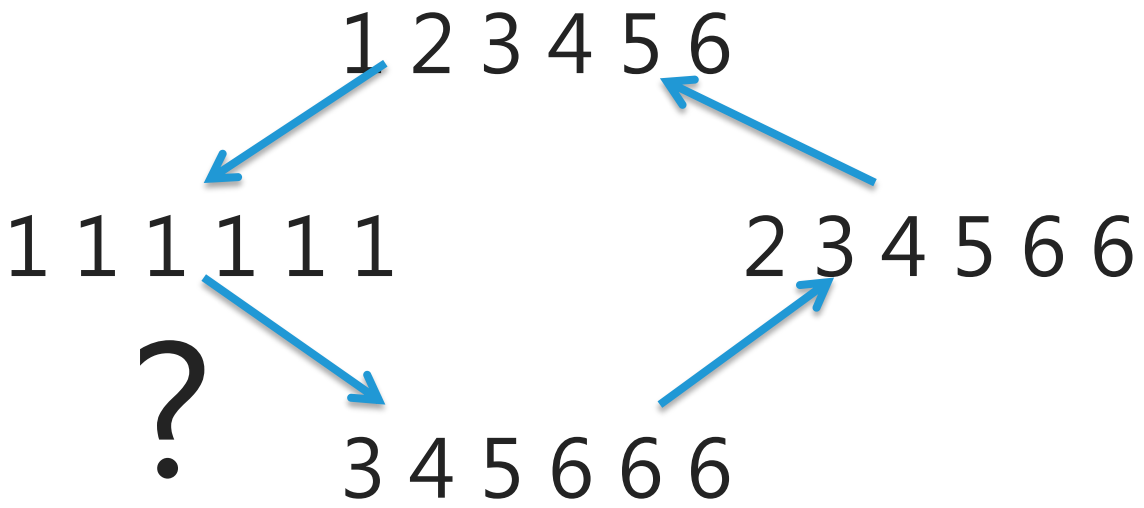
Die three must, on average, be able to beat die two

Die four must, on average, be able to beat die three

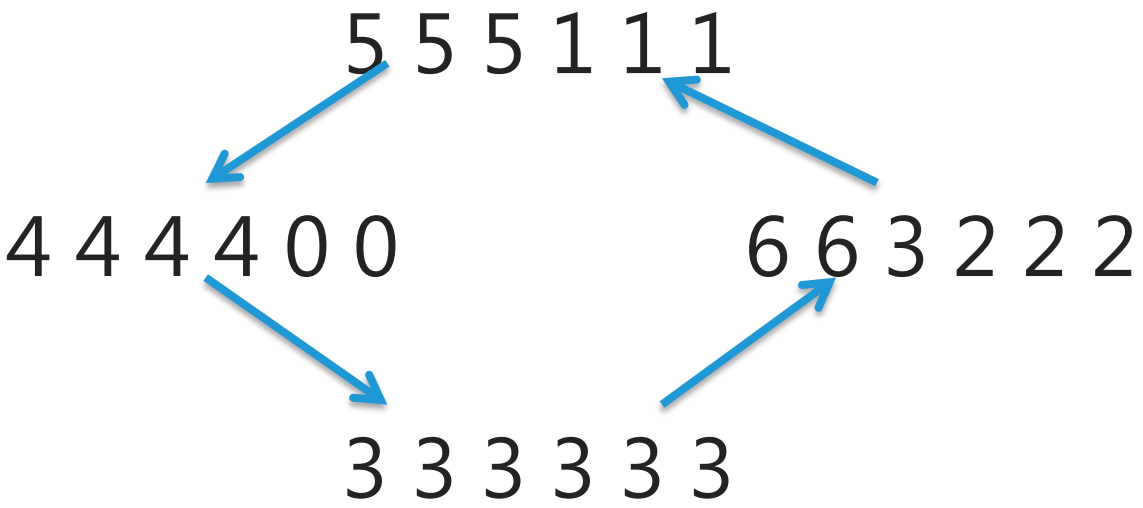
Die one must, on average, be able to beat die four

Exercise

Misbehaving dice



Exercise



A variation of Efron's nontransitive dice. (Efron uses 6 6 2 2 2 2)

Exercise

You can test all of these with a matrix for example, the ‘five’ die against the ‘four’.

	5	5	5	1	1	1
4	W	W	W	L	L	L
4	W	W	W	L	L	L
4	W	W	W	L	L	L
4	W	W	W	L	L	L
0	W	W	W	W	W	W
0	W	W	W	W	W	W

Exercise

You can test all of these with a matrix for example, the ‘five’ die against the ‘four’.

	5	5	5	1	1	1
4	W	W	W	L	L	L
4	W	W	W	L	L	L
4	W	W	W	L	L	L
4	W	W	W	L	L	L
0	W	W	W	W	W	W
0	W	W	W	W	W	W

Note that the Matrix can be a very good friend to you in general.

(This is not a film reference.)

Exercise

	5	5	5	1	1	1
4	W	W	W	L	L	L
4	W	W	W	L	L	L
4	W	W	W	L	L	L
4	W	W	W	L	L	L
0	W	W	W	W	W	W
0	W	W	W	W	W	W

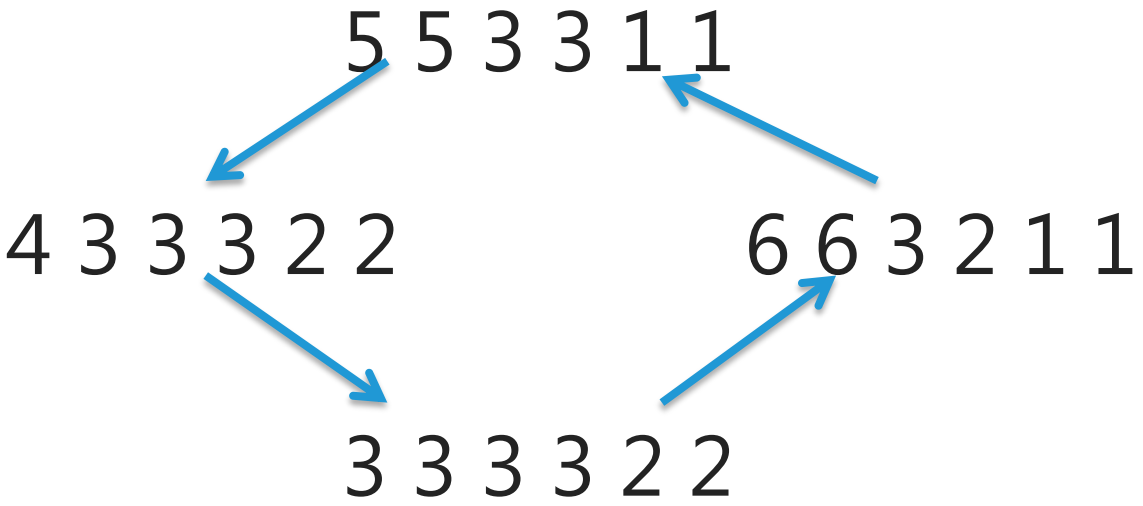
	3	3	3	3	3	3
6	L	L	L	L	L	L
6	L	L	L	L	L	L
3	N	N	N	N	N	N
2	W	W	W	W	W	W
2	W	W	W	W	W	W
2	W	W	W	W	W	W

	4	4	4	4	0	0
3	W	W	W	W	L	L
3	W	W	W	W	L	L
3	W	W	W	W	L	L
3	W	W	W	W	L	L
3	W	W	W	W	L	L
3	W	W	W	W	L	L

	6	6	3	2	2	2
5	W	W	L	L	L	L
5	W	W	L	L	L	L
5	W	W	L	L	L	L
1	W	W	W	W	W	W
1	W	W	W	W	W	W
1	W	W	W	W	W	W

Exercise

Misbehaving dice without using zero



Exercise

	5	5	3	3	1	1
4	W	W	L	L	L	L
3	W	W	N	N	L	L
3	W	W	N	N	L	L
3	W	W	N	N	L	L
2	W	W	W	W	L	L
2	W	W	W	W	L	L

	3	3	3	3	2	2
6	L	L	L	L	L	L
6	L	L	L	L	L	L
3	N	N	N	N	L	L
2	W	W	W	W	N	N
1	W	W	W	W	W	W
1	W	W	W	W	W	W

	4	3	3	3	2	2
3	W	N	N	N	L	L
3	W	N	N	N	L	L
3	W	N	N	N	L	L
3	W	N	N	N	L	L
2	W	W	W	W	N	N
2	W	W	W	W	N	N

	6	6	3	2	1	1
5	W	W	L	L	L	L
5	W	W	L	L	L	L
3	W	W	N	L	L	L
3	W	W	N	L	L	L
1	W	W	W	W	N	N
1	W	W	W	W	N	N

Recency, Frequency and Intensity

Recency, Frequency and Intensity

Analytical technique to distil large quantities of activity

Combined with data mining to predict later activity.

Recency

When did the event last happen ?

- Number of days since the last contact with a customer

Frequency

How often does the event happen?

- How often do we interact with the customer

Intensity

Recency and frequency are indicators of customer loyalty

However, you need intensity to make sure it’s worth while.

Intensity

How big are the interactions?

- Total number of webpages visited by customer

Can also be Monetary

- Total amount of purchases

Can also be Value

How do we do it ?

Compute the RFI for each class of activity in a given time window

- Compute RFI for each customer for each month

This is now “plotted” in 3 dimensional cube

How do we do it?

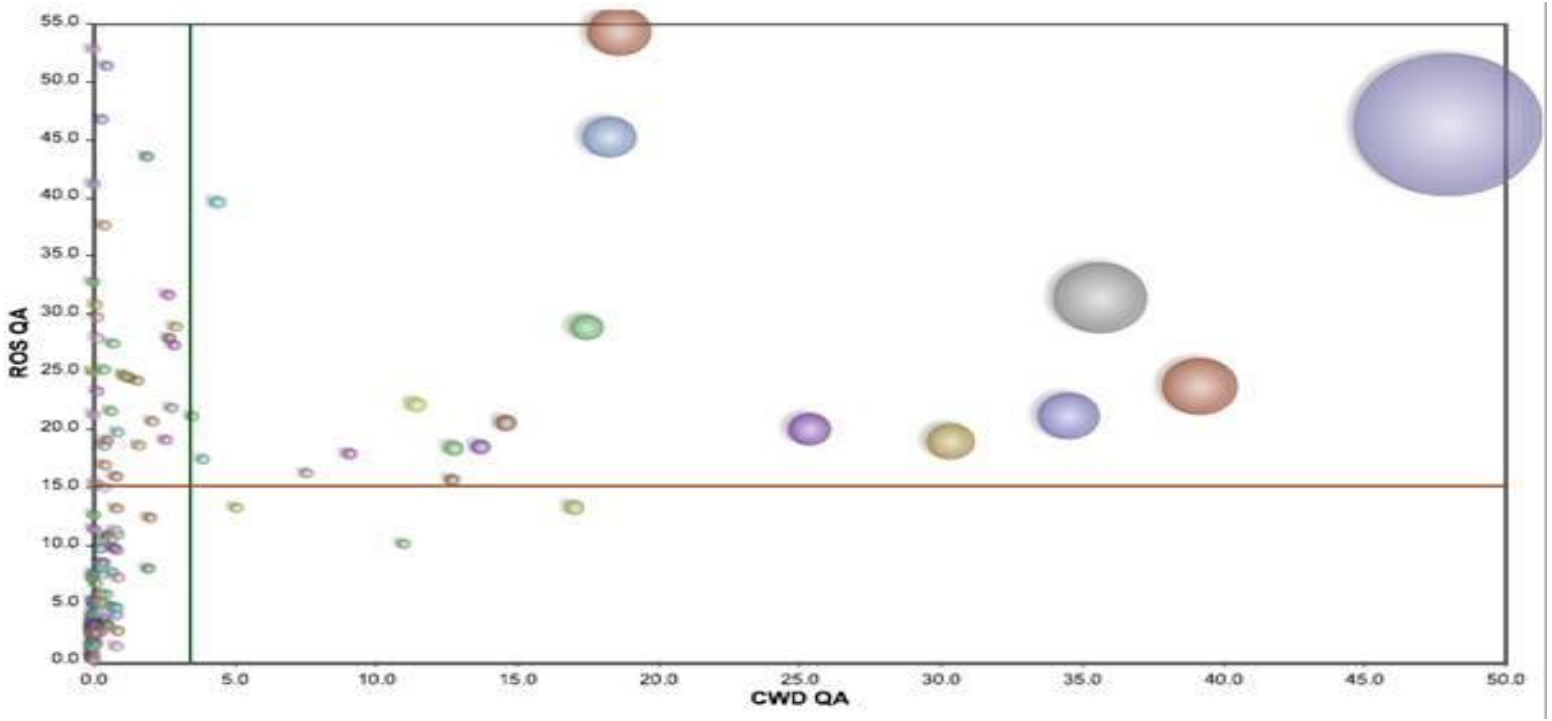
We use data mining to identify clusters in the cube

We may then find N natural clusters

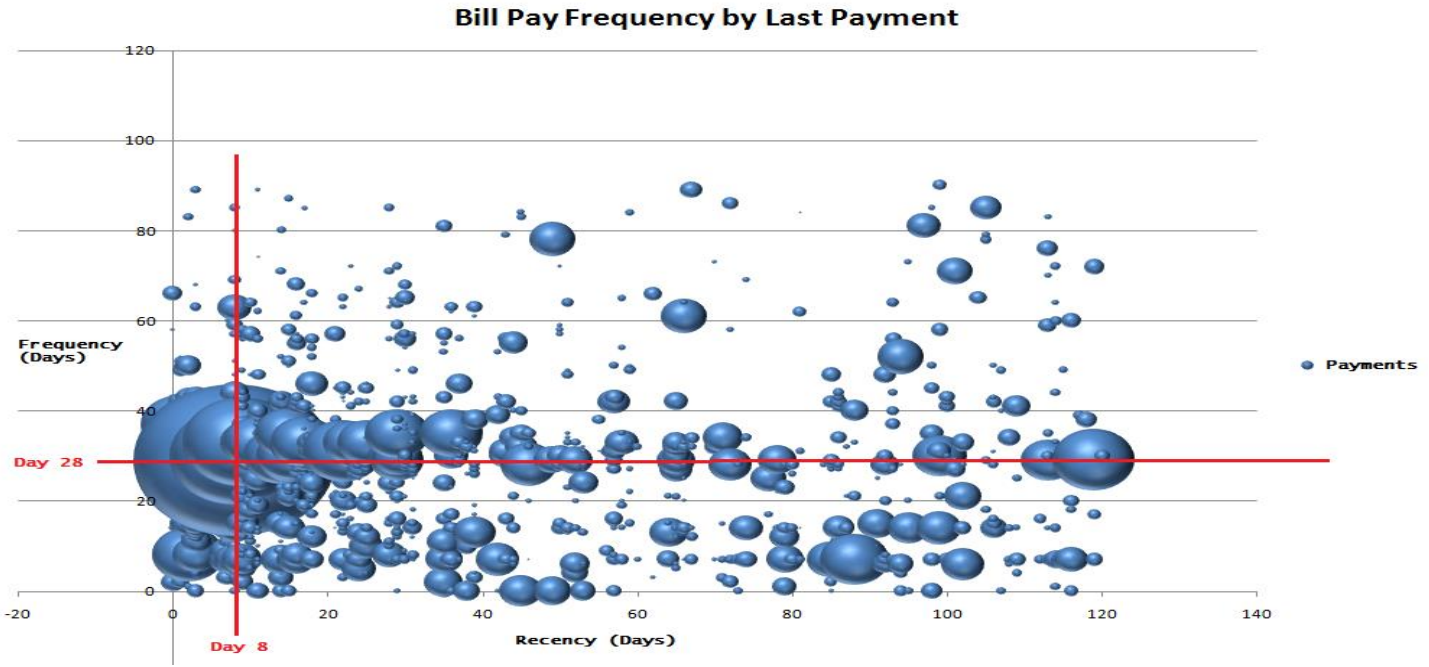
- So, we might find 8 clusters

We then identify what the clusters might mean

We can use bubble charts



Another example



How do we do it ?

Assign a tag to each cluster

Tags can have meaning or not

Tags can be A,B,C,D,E,F for instance

How do we do it ?

Then we develop a time series for the activity

- Activity A A A B C D A D E F

We have now boiled down the very complex activity into a simple, analysable, pattern.

This can now be mined as well.

Example

Suppose the analysis is for customer activity

What might be suitable tags?

What might be good time lines?

Can we see how these can tell customer stories?

Criticism

It may lead to targeting the most attractive prospects, ignoring others that could be usefully developed.

It can lead to oversimplification.

Other Uses

Defects in manufacturing

Scientific research

More info

Kimball design tip 33

Also

Why RFM Works in Predicting Response

http://www.dbmarketing.com/articles/Art245.htm

Martingale

Why doesn’t the Martingale strategy work?

Markov chains

Andrey Markov, professor at St. Petersburg University (when promoted to merited professor he acquired the right to retire and did so immediately and continued to lecture for another 14 years).

Markov chains are an interesting (for ‘interesting’ read ‘useful’) way of modelling the behaviour of both people and systems.

Markov chains

Fred can work from home or he can work in the office. If he works at home he has a tendency to get involved in a project there and so is more likely to work from home the next day.

(P=0.75 he will work at home tomorrow if he worked at home today).

If he works from the office he has a 50/50 chance of working in the office the next day.

Weekends have no effect on Fred's behaviour.

Markov chains

What proportion of Fred's working days are spent at home?

How often will Fred spend 3 or more consecutive days at home?

What are the chances that Fred will work 3 days at home followed by 2 days at work?

Markov chains

Transition Matrix - shows the probability of transitioning from the row state to the column state.

	Home (A)	Work (B)
Home (A)	**P(AIA) = 0.75**	**P(BIA) = 0.25**
Work (B)	**P(AIB) = 0.5**	**P(BIB) = 0.5**

http://setosa.io/ev/markov-chains

Markov chains

Transition Matrix - shows the probability of transitioning from the row state to the column state.

	Home (A)	**Work (B)**
Home (A)	**P(AIA) = 0.75**	**P(BIA) = 0.25**
Work (B)	**P(AIB) = 0.5**	**P(BIB) = 0.5**

If Fred can also work on-site then what do we need to add to the transition matrix?

Markov chains

One row and one column.

	Home (A)	Work (B)	On-Site (C)
Home (A)	**P(AIA) = 0.7**	**P(BIA) = 0.25**	**P(CIA) = 0.05**
Work (B)	**P(AIB) = 0.45**	**P(BIB) = 0.5**	**P(BIC) = 0.1**
On-Site (C)	**P(AIC) = 0.6**	**P(BIC) = 0.1**	**P(CIC) = 0.3**

Spot the two errors in the matrix.

Markov chains

One row and one column.

	Home (A)	Work (B)	On-Site (C)
Home (A)	**P(AIA) = 0.7**	**P(BIA) = 0.25**	**P(CIA) = 0.05**
Work (B)	**P(AIB) = 0.45**	**P(BIB) = 0.45**	**P(CIB) = 0.1**
On-Site (C)	**P(AIC) = 0.6**	**P(BIC) = 0.1**	**P(CIC) = 0.3**

.

Markov chains

Markov chains (MCs) are stochastic processes.

What other stochastic system do you know that also has the initials MCS?

Is this, do you think, a hint about how we can built Markov chains in practice?