What does a Data Scientist Actually Do?

Chris Hillman

Hyper-Hype

Sexiest job of the 21st Century?
Rock Stars?
Ninjas?
Unicorns?

Data science incorporates varying elements and builds on techniques and theories from many fields, including mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products. Data science is a novel term that is often used interchangeably with competitive intelligence or business analytics, although it is becoming more common. Data science seeks to use all available and relevant data to effectively tell a story that can be easily understood by non-practitioners. (wikipedia, Nov 2013)

Data Scientist

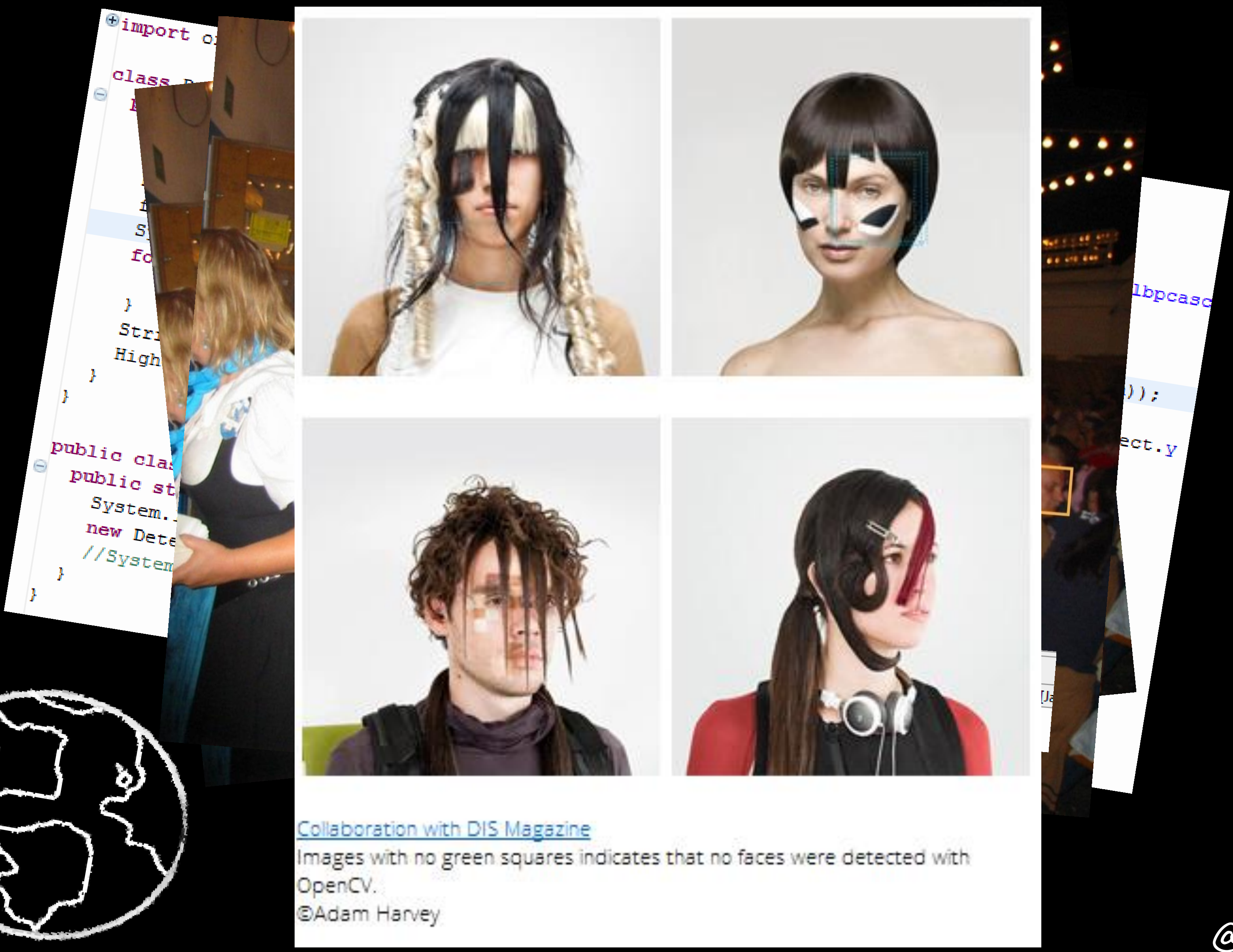
A better programmer than a statistician
A better statistician than a programmer

“Discovery consists of seeing what everybody has seen and thinking what nobody has thought."
Albert Szent-Gyorgy

See what all see, think what none think
quod vide omnia vide, quod cogitare nullus cogitare

Combination of Skills

Face Detection in Images



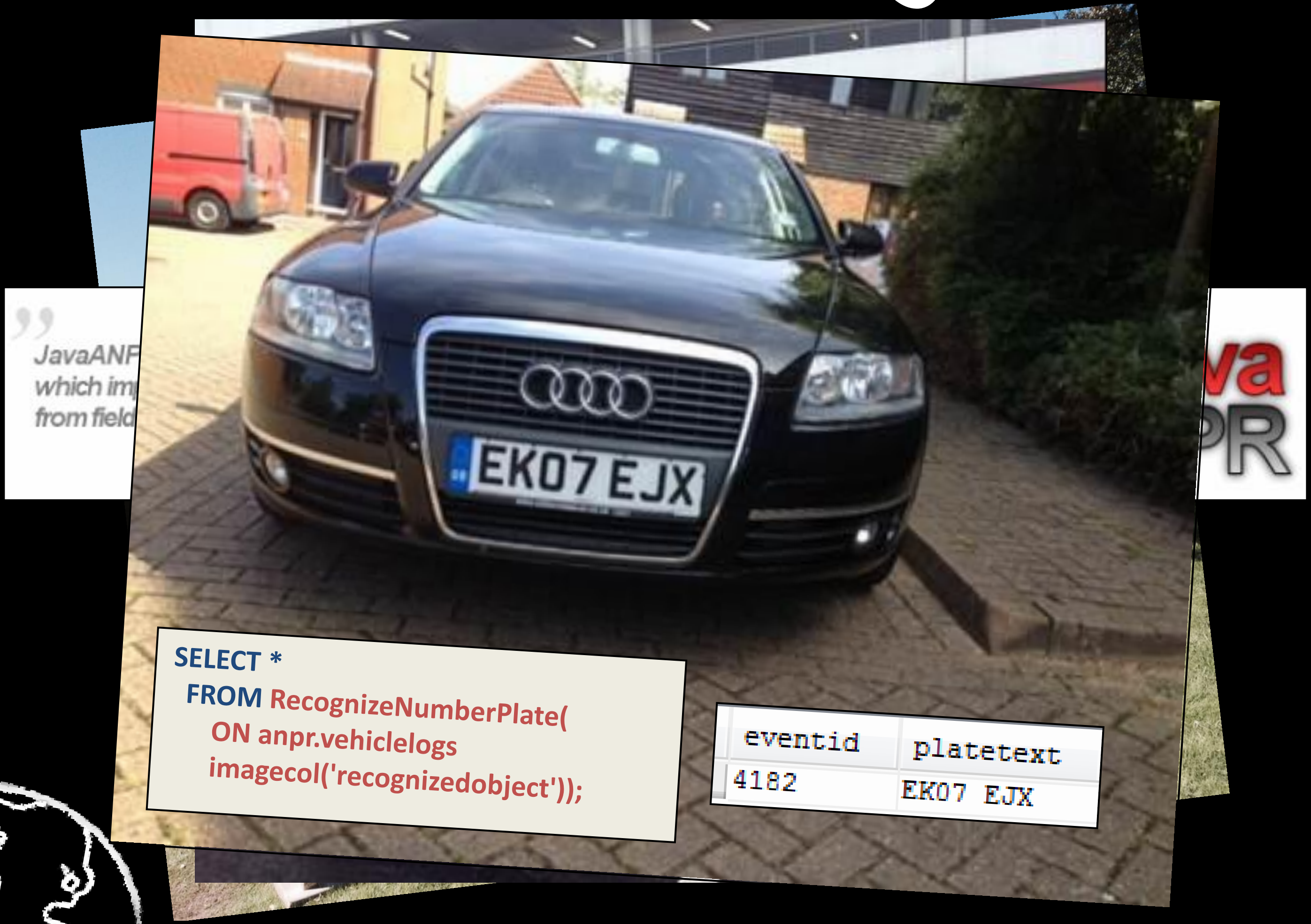
Collaboration with DIS Magazine

Images with no green squares indicates that no faces were detected with OpenCV.

©Adam Harvey

@chillax7

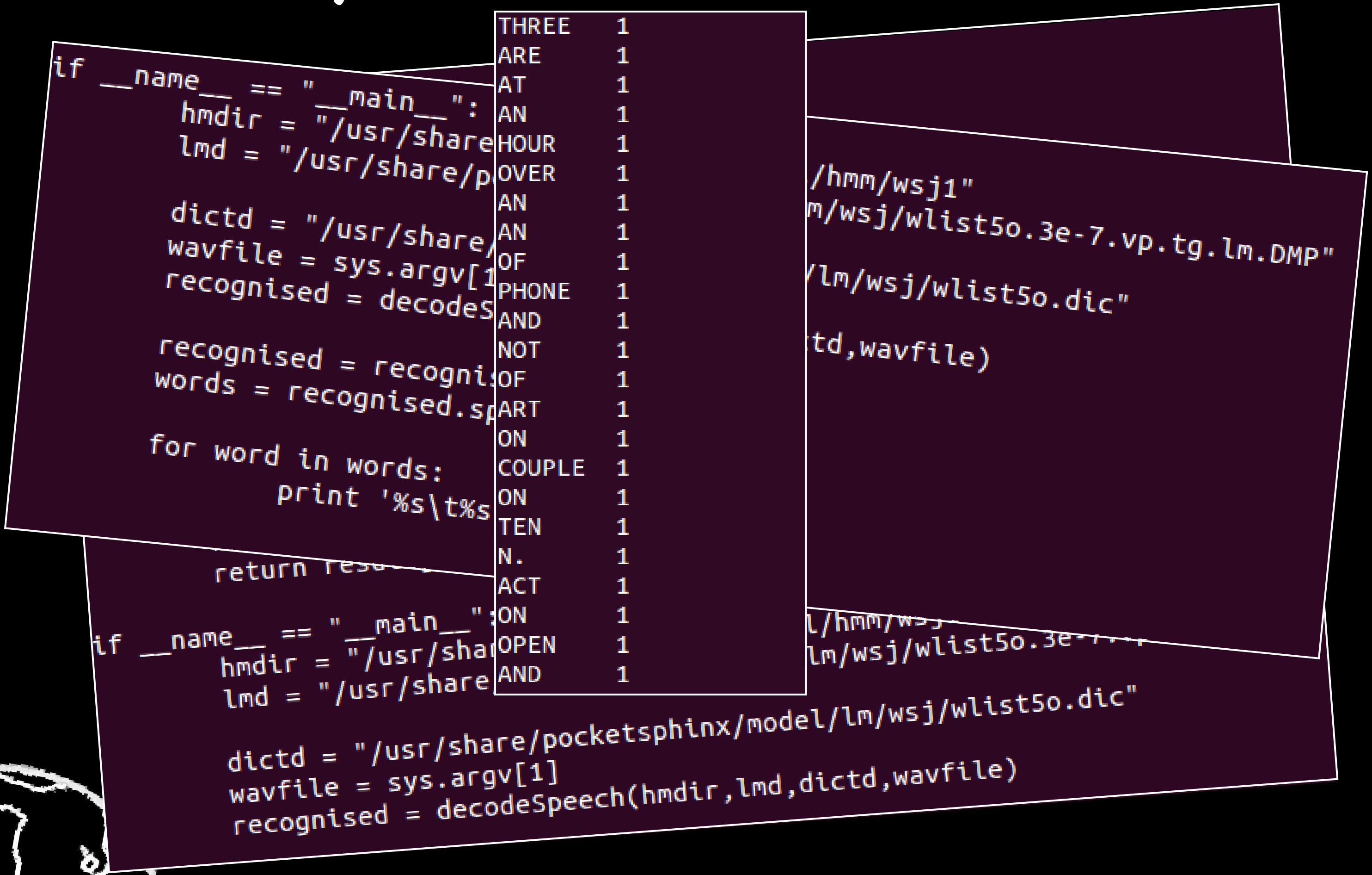
Character Recognition



```
SELECT *
 FROM RecognizeNumberPlate(
   ON anpr.vehiclelogs
   imagecol('recognizedobject'));
```

eventid	platetext
4182	EK07 EJX

Speech to Text

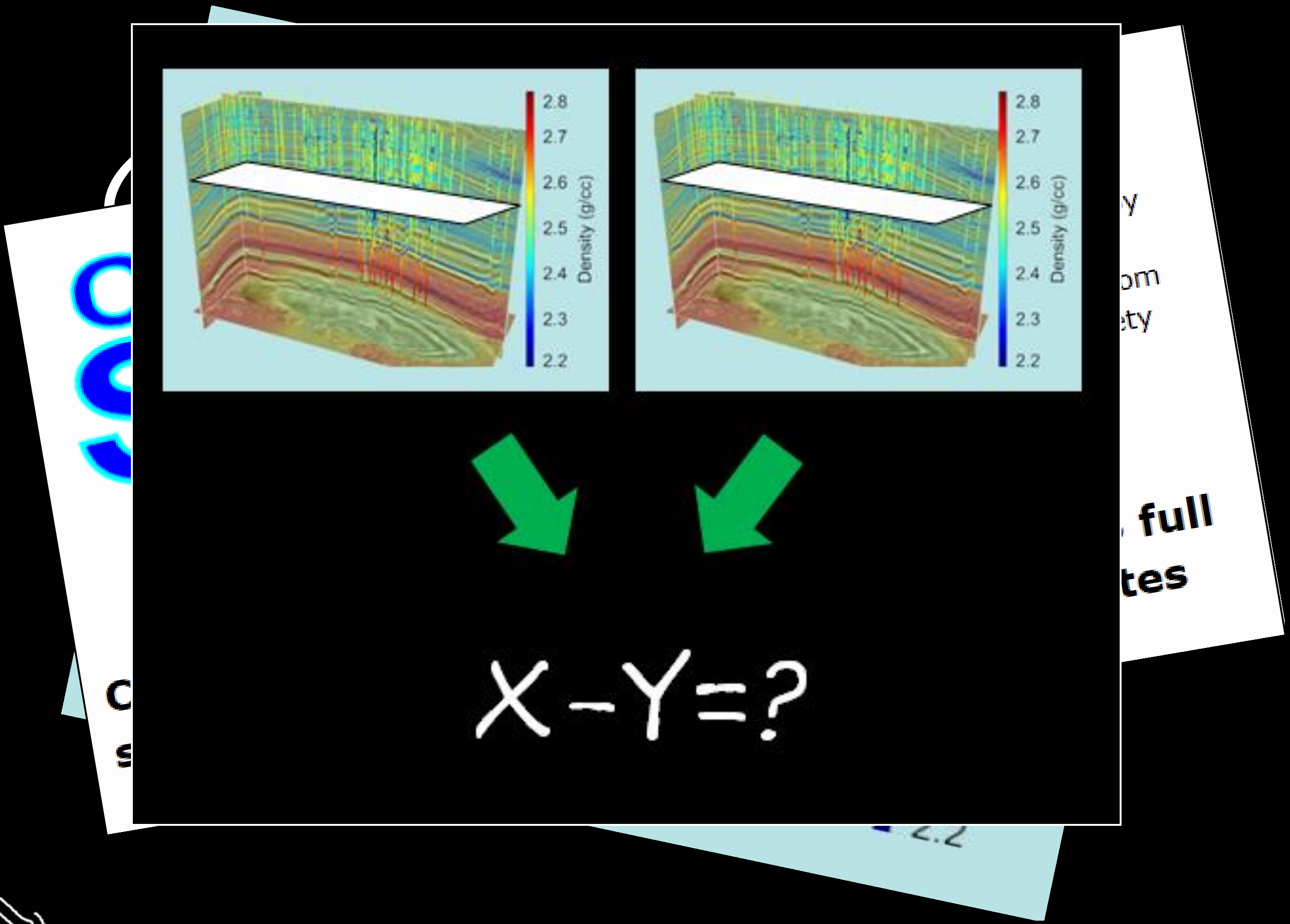


```
if __name__ == "__main__":
        hmdir = "/usr/share /hmm/wsj1"
        lmd = "/usr/share/p m/wsj/wlist5o.3e-7.vp.tg.lm.DMP"

        dictd = "/usr/share/ /lm/wsj/wlist5o.dic"
        wavfile = sys.argv[1
        recognised = decodeS td,wavfile)

        recognised = recogni
        words = recognised.sp

        for word in words:
                print '%s\t%s
```

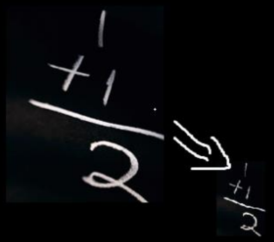
```
THREE    1
ARE      1
AT       1
AN       1
HOUR     1
OVER     1
AN       1
AN       1
OF       1
PHONE    1
AND      1
NOT      1
OF       1
ART      1
ON       1
COUPLE   1
ON       1
TEN      1
N.       1
ACT      1
ON       1
OPEN     1
AND      1
```

```
        return resu

if __name__ == "__main__": l/hmm/wsj
        hmdir = "/usr/shar lm/wsj/wlist5o.3e-7
        lmd = "/usr/share

        dictd = "/usr/share/pocketsphinx/model/lm/wsj/wlist5o.dic"
        wavfile = sys.argv[1]
        recognised = decodeSpeech(hmdir,lmd,dictd,wavfile)
```

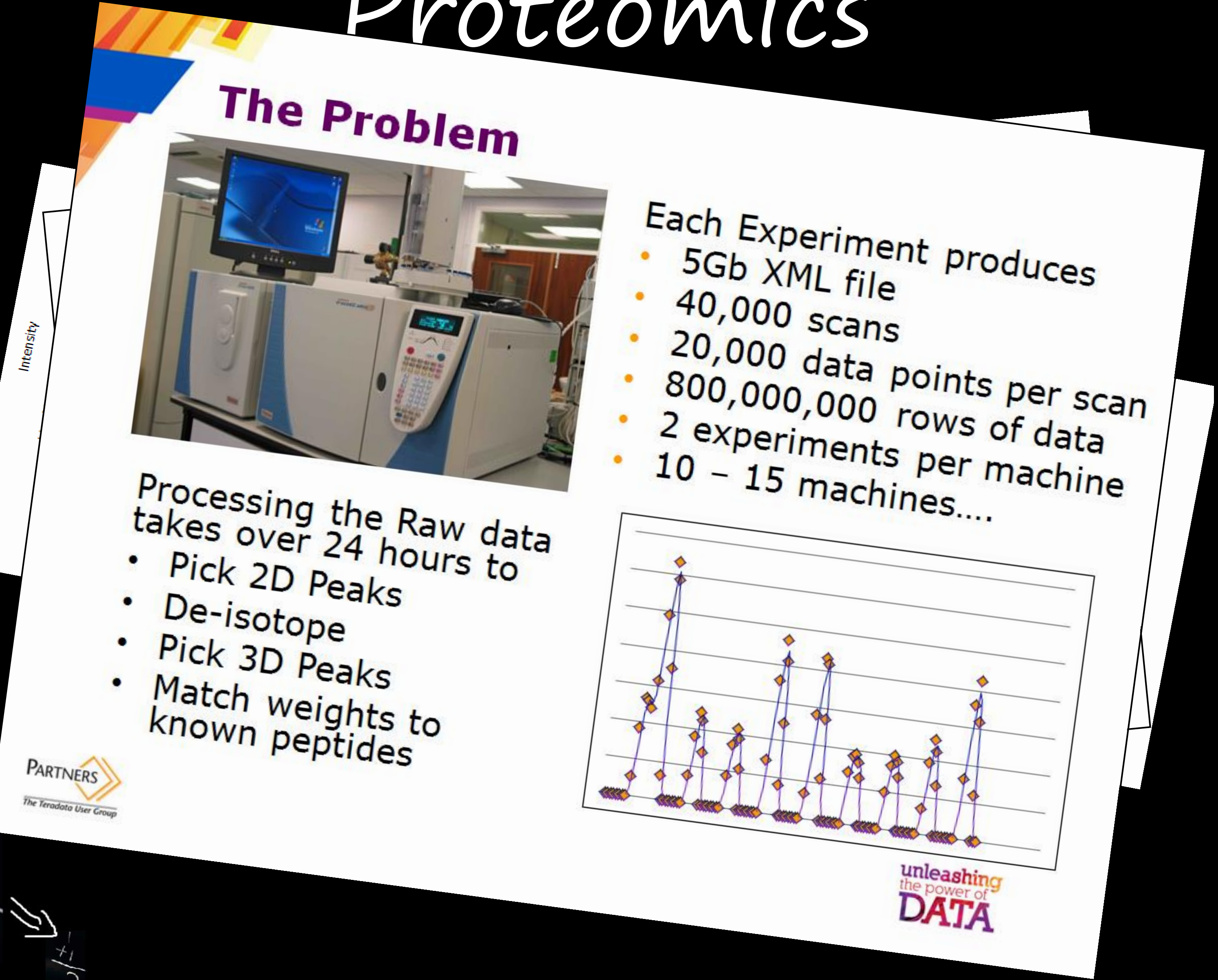
Is Big Data the new oil?



X-Y=?



Proteomics



The Problem

Each Experiment produces

- 5Gb XML file
- 40,000 scans
- 20,000 data points per scan
- 800,000,000 rows of data
- 2 experiments per machine
- 10 – 15 machines….

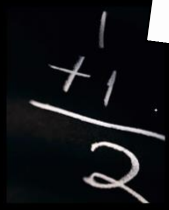
Processing the Raw data takes over 24 hours to

- Pick 2D Peaks
- De-isotope
- Pick 3D Peaks
- Match weights to known peptides

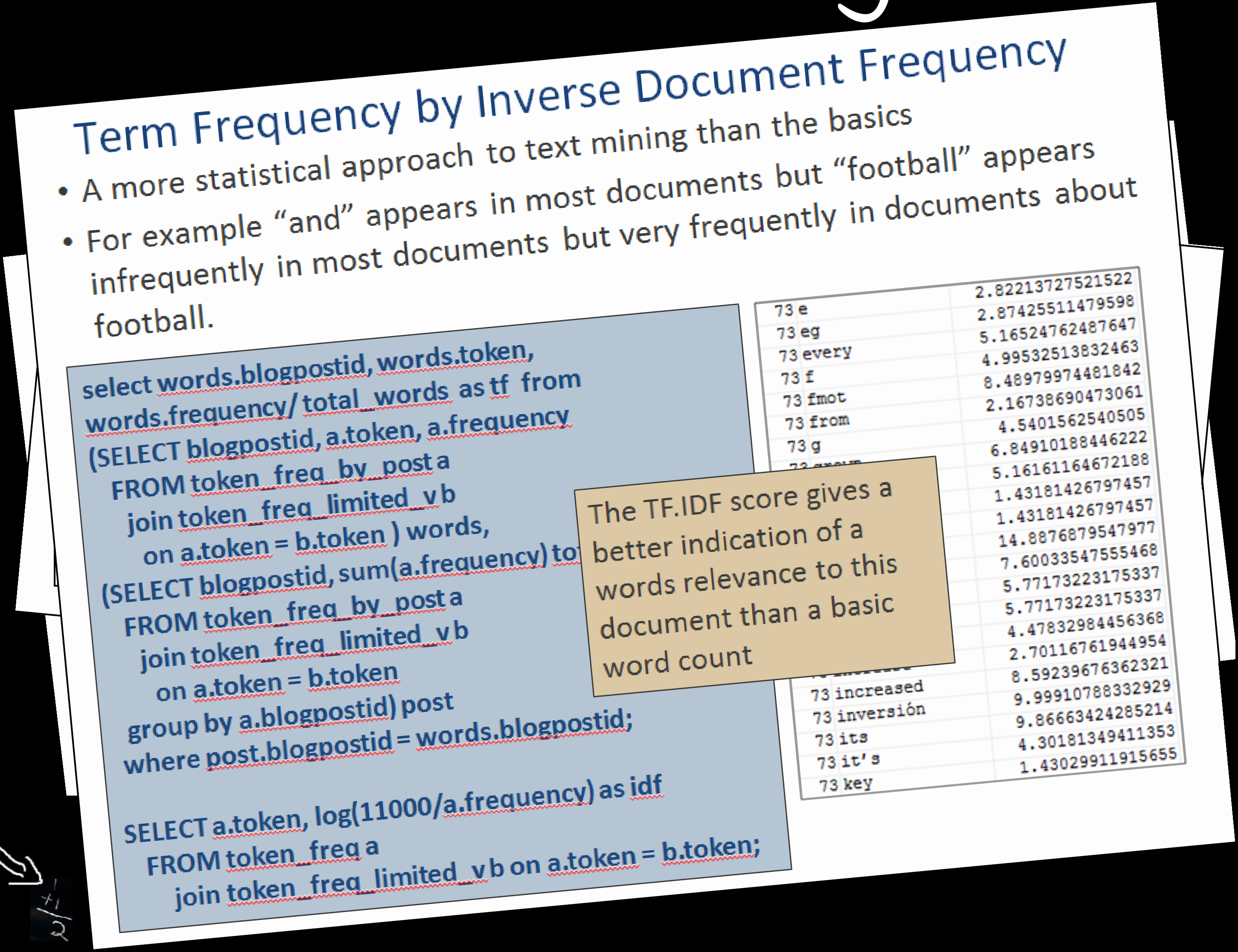
PARTNERS

The Teradata User Group

unleashing the power of DATA



Text Mining



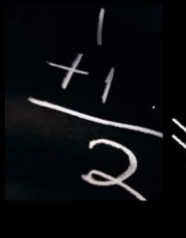
Term Frequency by Inverse Document Frequency

- A more statistical approach to text mining than the basics
- For example "and" appears in most documents but "football" appears infrequently in most documents but very frequently in documents about football.

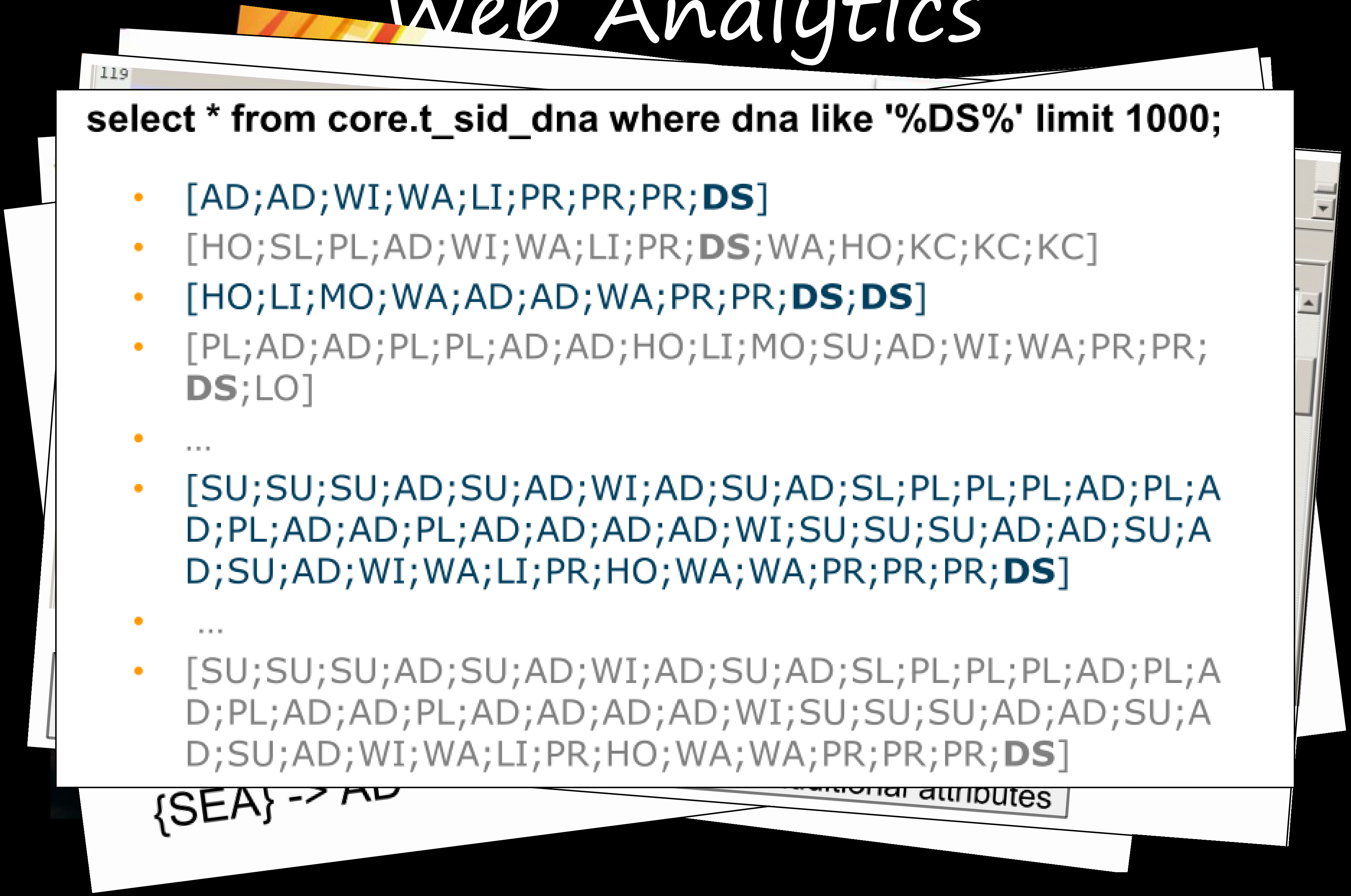
```
select words.blogpostid, words.token,
words.frequency/ total_words as tf from
(SELECT blogpostid, a.token, a.frequency
  FROM token_freq_by_post a
   join token_freq_limited_v b
    on a.token = b.token ) words,
(SELECT blogpostid, sum(a.frequency) to
  FROM token_freq_by_post a
   join token_freq_limited_v b
    on a.token = b.token
 group by a.blogpostid) post
 where post.blogpostid = words.blogpostid;

SELECT a.token, log(11000/a.frequency) as idf
  FROM token_freq a
   join token_freq_limited_v b on a.token = b.token;
```

73	e	2.82213727521522
73	eg	2.87425511479598
73	every	5.16524762487647
73	f	4.99532513832463
73	fmot	8.48979974481842
73	from	2.16738690473061
73	g	4.5401562540505
73	group	6.84910188446222
[illegible]	[illegible]	5.1616116467218
[illegible]	[illegible]	1.43181426797457
[illegible]	[illegible]	1.43181426797457
[illegible]	[illegible]	14.8876879547977
[illegible]	[illegible]	7.60033547555468
[illegible]	[illegible]	5.77173223175337
[illegible]	[illegible]	5.77173223175337
[illegible]	[illegible]	4.47832984456368
[illegible]	[illegible]	2.70116761944954
73	increased	8.59239676362321
73	inversión	9.99910788332929
73	its	9.86663424285214
73	it's	4.30181349411353
73	key	1.43029911915655

The TF.IDF score gives a better indication of a words relevance to this document than a basic word count



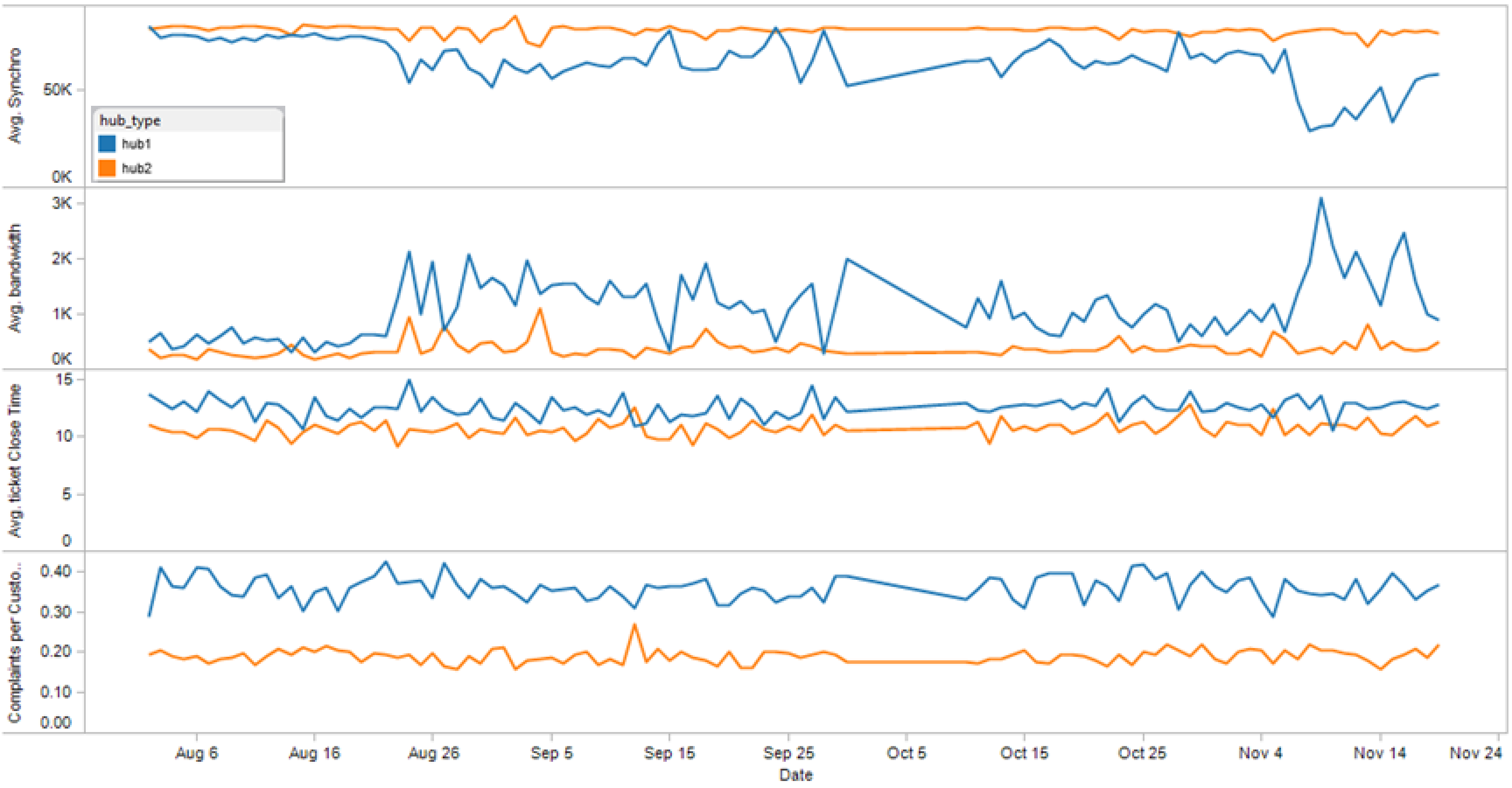
Web Analytics



```
select * from core.t_sid_dna where dna like '%DS%' limit 1000;
```

- [AD;AD;WI;WA;LI;PR;PR;PR;**DS**]
- [HO;SL;PL;AD;WI;WA;LI;PR;**DS**;WA;HO;KC;KC;KC]
- [HO;LI;MO;WA;AD;AD;WA;PR;PR;**DS**;**DS**]
- [PL;AD;AD;PL;PL;AD;AD;HO;LI;MO;SU;AD;WI;WA;PR;PR;**DS**;LO]
- …
- [SU;SU;SU;AD;SU;AD;WI;AD;SU;AD;SL;PL;PL;PL;AD;PL;AD;PL;AD;AD;PL;AD;AD;AD;AD;WI;SU;SU;SU;AD;AD;SU;AD;SU;AD;WI;WA;LI;PR;HO;WA;WA;PR;PR;PR;**DS**]
- …
- [SU;SU;SU;AD;SU;AD;WI;AD;SU;AD;SL;PL;PL;PL;AD;PL;AD;PL;AD;AD;PL;AD;AD;AD;AD;WI;SU;SU;SU;AD;AD;SU;AD;SU;AD;WI;WA;LI;PR;HO;WA;WA;PR;PR;PR;**DS**]

Error and Complaint rates by equipment type



Thank you

Chris Hillman
Chris_hillman@yahoo.com
@chillax7
www.bigdatablog.co.uk