Models: pets and herds

Carlos J. Gil Bellosta
cgb@datanalytics.com

September 2014

This is a pet...

Source: http://jessfalcone.wordpress.com

... and this is a herd

Source: http://bonfirehealth.com/negative-influences-comparisons-social-cues-herd/

Some people treat computers as pets...

Source: aliexpress.com

... an others like herds

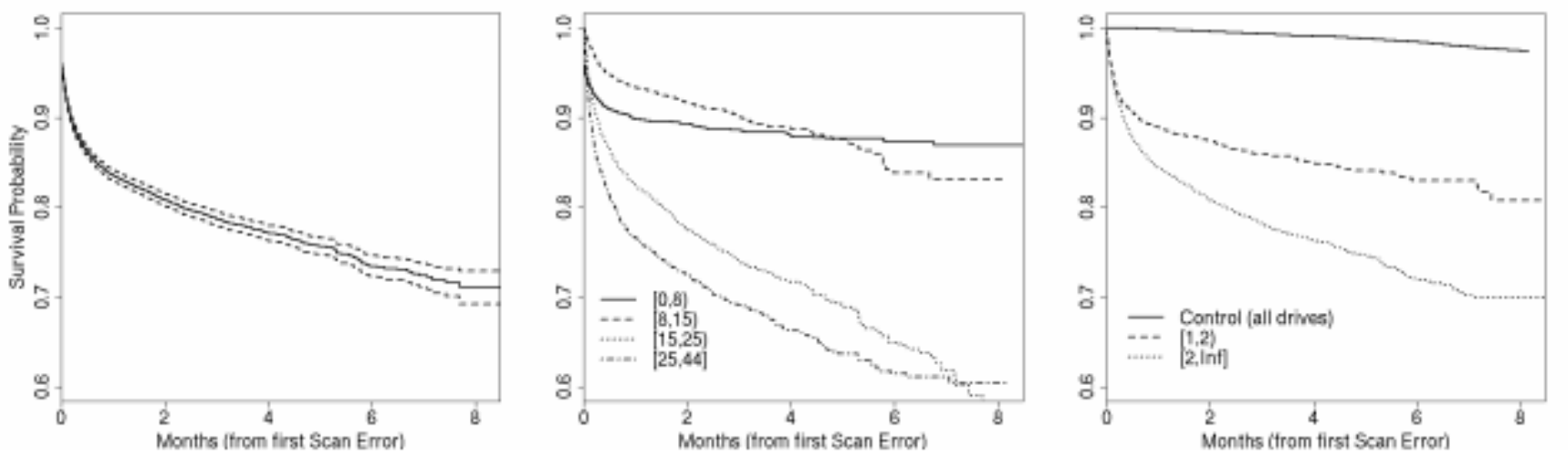


Figure 8: Impact of scan errors on survival probability. Left figure shows aggregate survival probability for all drives after first scan error. Middle figure breaks down survival probability per drive ages in months. Right figure breaks down drives by their number of scan errors.

Source: Failure Trends in a Large Disk Drive Population, Pinheiro et al.

This is a statistical model treated as a pet

```
logit y c.r##c.m cv1, nolog

Logistic regression                                     Number of obs   =        200
                                                        LR chi2(4)      =      66.80
                                                        Prob > chi2     =     0.0000
Log likelihood = -77.953857                             Pseudo R2       =     0.3000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           r |   .4342063   .1961642     2.21   0.027     .0497316    .8186809
           m |   .5104617   .2011856     2.54   0.011     .1161452    .9047782
             |
     c.r#c.m |  -.0068144   .0033337    -2.04   0.041    -.0133483   -.0002805
             |
         cv1 |   .0309685   .0271748     1.14   0.254    -.0222931       .08423
       _cons |  -34.09122   11.73402    -2.91   0.004    -57.08947   -11.09297
------------------------------------------------------------------------------
```

Source: http://www.ats.ucla.edu/stat/stata/seminars/interaction_sem/interaction_sem.htm

Pets are very demanding and require...

1. **variable selection**,
2. checks for **outilers**,
3. assessment of the **goodness of fit**,
4. finding **confidence intervals**,
5. calculating **p-values**,
6. **interpretating** the results,
7. discuss the **generalization**,
8. ...

Models... as herds?

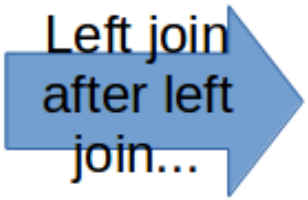
Source: `http://www.gotmedieval.com`

Model construction: population

Source: http://timyeo.wordpress.com/

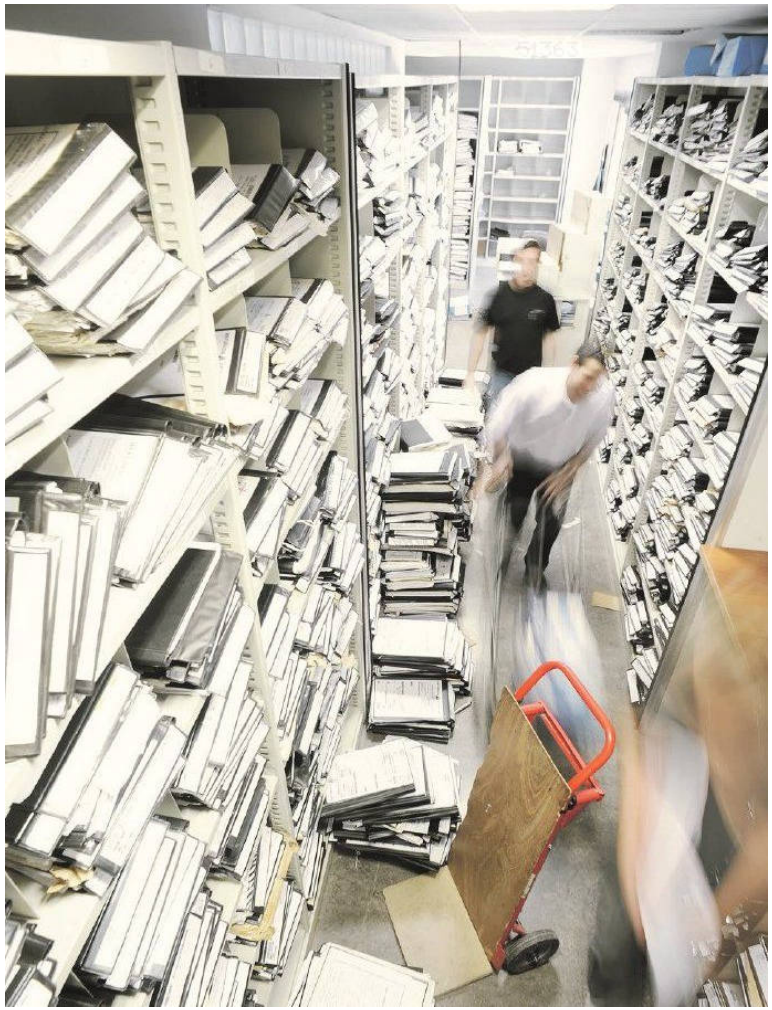
Model construction: data enrichment (aka left join)

id
001
002
003
004
005
006



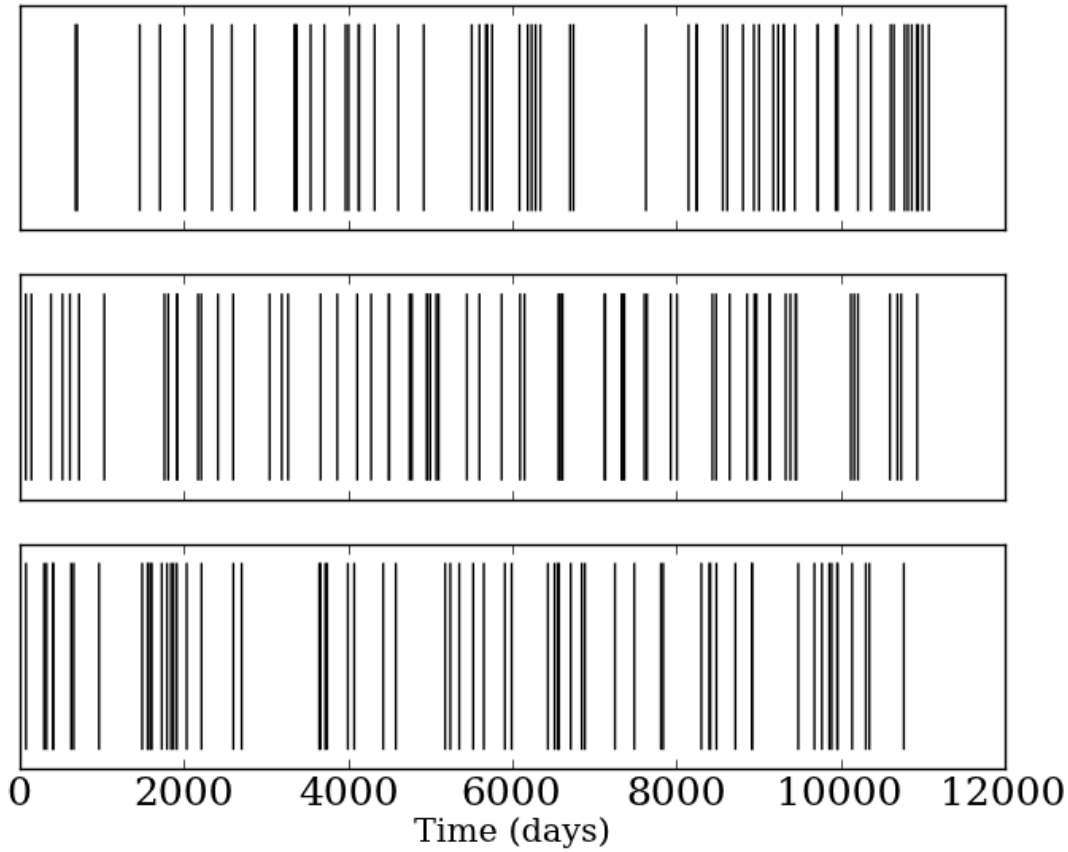
id	sex	age	prov	...
001	H	<18	08	...
002	M	25-30	28	...
003	M	45-50	28	...
004	M	45-50	50	...
005	H	>65	47	...
006	H	30-35	28	...

But subject data is often messy...



Source: `http://arquitectolegista.com.ar/`

... and contains temporal data...



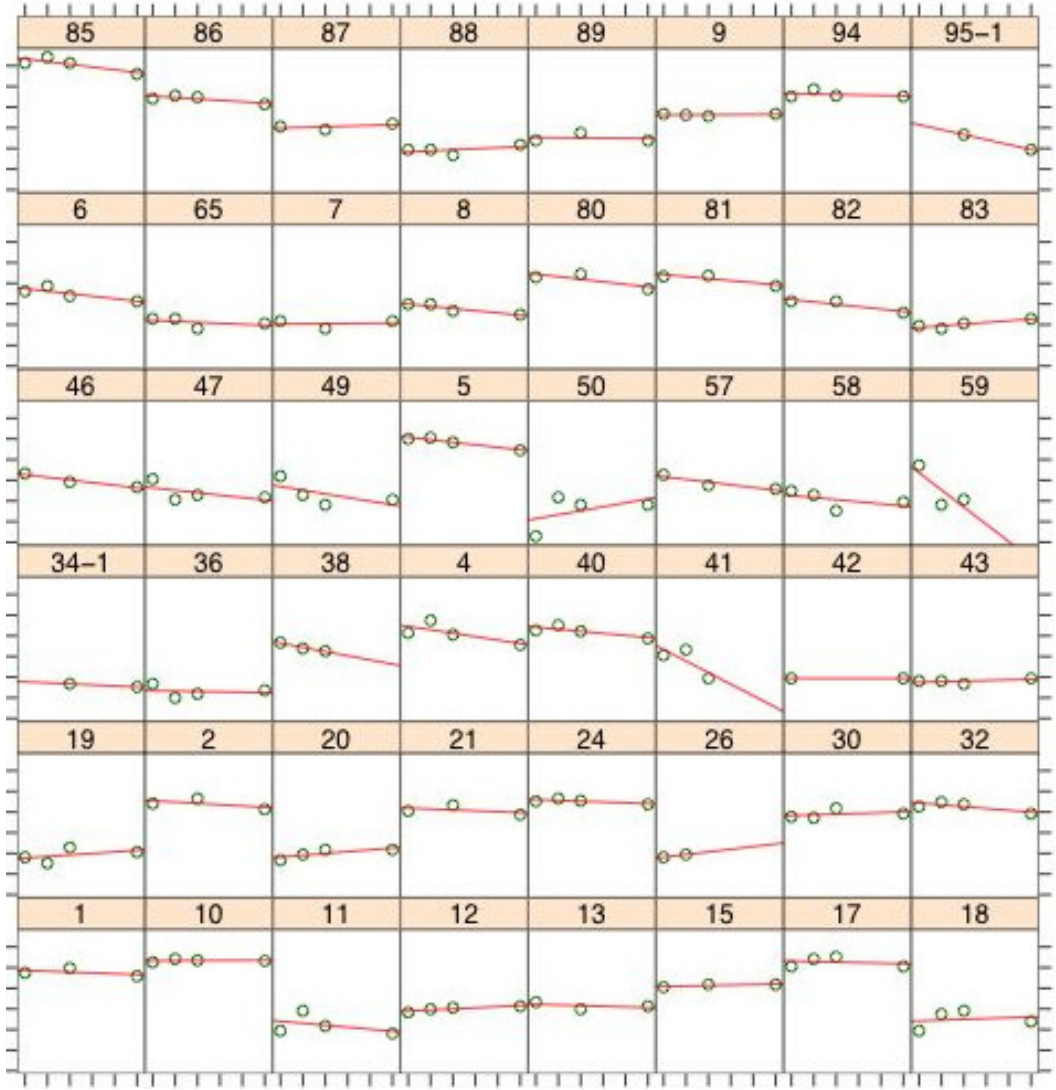
Source: `http://thirdorderscientist.org/`

... that is difficult to fit in a box (table)

Source: http://cutestcatpics.com/cat-trying-to-fit-into-tiny-box/

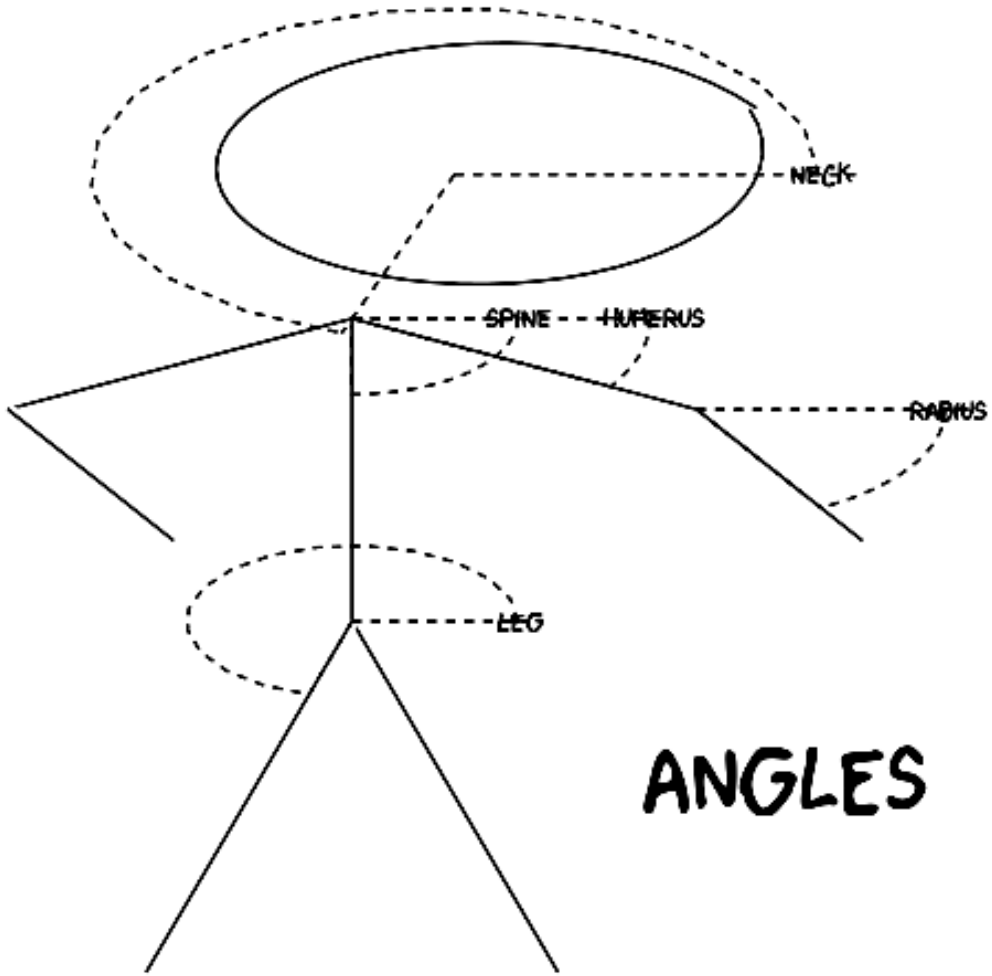
We have a whole dataset per subject!

... and a model per subject?



Source: http://www.unc.edu/

(Most) models are sophisticated summaries of data



Source: http://xkcd.r-forge.r-project.org/

Do you seek α? Build a model per stock!

The Journal of FINANCE

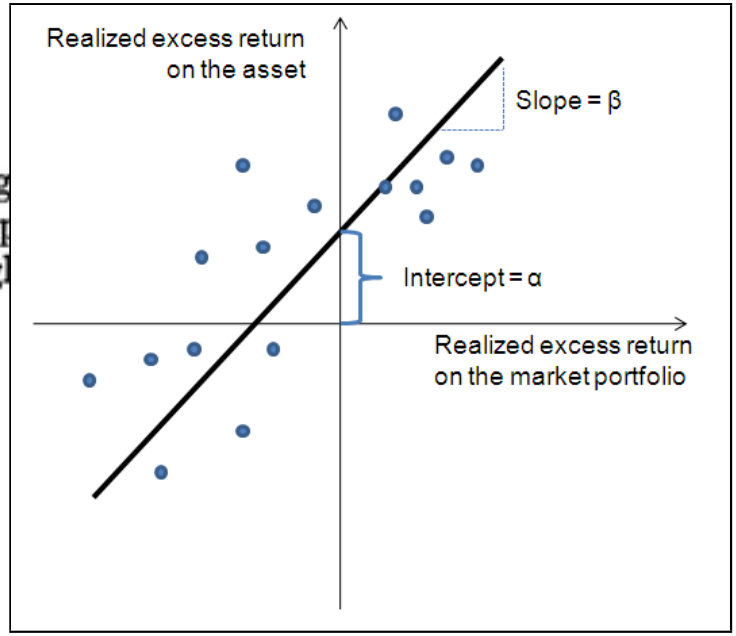
VOL. XIX SEPTEMBER 1964 No. 3

CAPITAL ASSET PRICES: A THEORY OF MARKET EQUILIBRIUM UNDER CONDITIONS OF RISK*

WILLIAM F. SHARPE†

I. INTRODUCTION

ONE OF THE PROBLEMS which has plagued those attempting
behavior of capital markets is the absence of a body of
economic theory dealing with conditions of risk. Althoug



Fitting a million models in the nineties was all of an achievement (for some)

I Just Ran Two Million Regressions

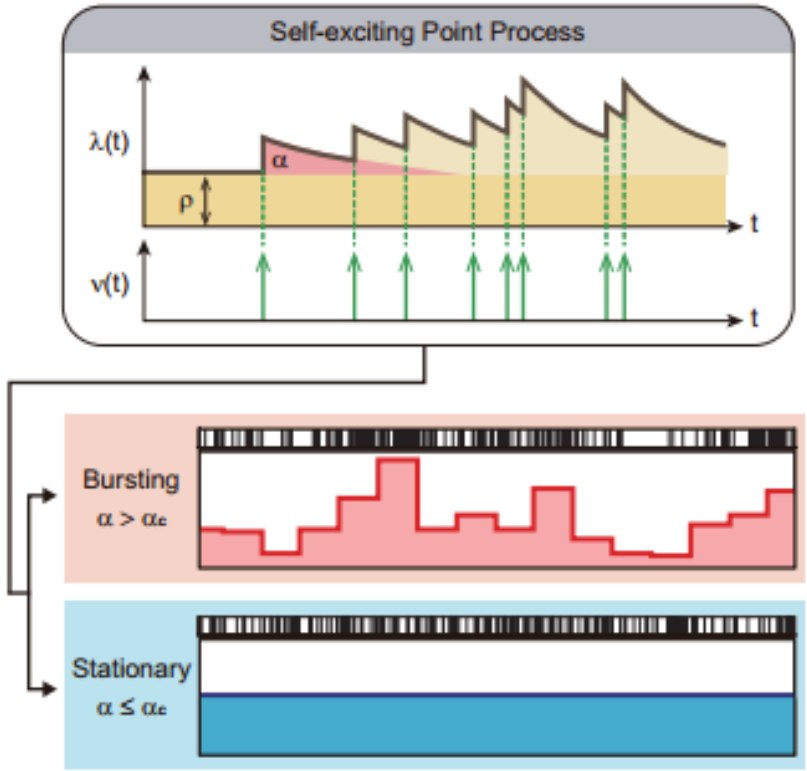
By XAVIER X. SALA-I-MARTIN*

Following the seminal work of Robert Barro (1991), the recent empiri... nomic growth has identifi... An initial answer to this question was given ...David Renelt (1992).[1] ...rd Leamer's (1985)

TABLE 1—MAIN RESULTS OF REGRESSIONS
(DEPENDENT VARIABLE = GROWTH)

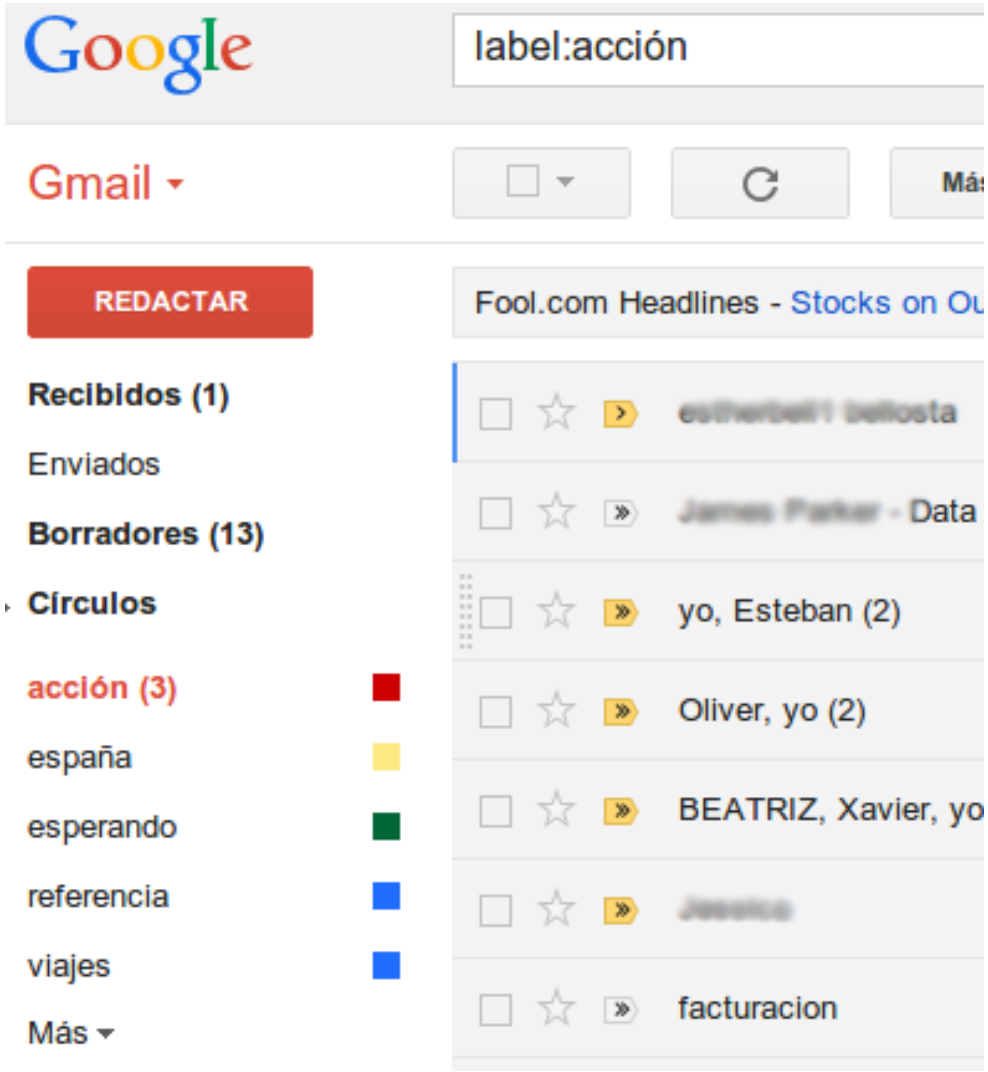
Independent variable	(i) β	(ii) SD	(iii) CDF[a]
Equipment investment	0.2175	0.0408	1.000
Number of years open economy	0.0195	0.0042	1.000
Fraction Confucian	0.0676	0.0149	1.000
Rule of law	0.0190	0.0049	1.000
Fraction Muslim	0.0142	0.0035	1.000
Political rights	−0.0026	0.0009	0.998
Latin America dummy	−0.0115	0.0029	0.998
Sub-Saharan Africa dummy	−0.0121	0.0032	0.997
Civil liberties	−0.0029	0.0010	0.997
Revolutions and coups	−0.0118	0.0045	0.995

Beyond recency and frequency: self exciting processes



Source: *Bursting transition in a linear self-exciting point process*, Onaga, T. et al

One logistic regression per Gmail user...



Challenges: statistical, computational,... and more!

This approach faces many challenges:

1. **Computational:** how do you fit so many models? (but Spark rocks!)
2. **Statistical:** how do you...
    1. perform variable selection?
    2. evaluate the fit?
    3. deal with outliers?
    4. ...
3. And finally, how do you sell these approaches to *business people* (ex-Google)?