byte sumo

TrendCalculus

A data science for studying trends.

2014-12-16

Trends?!

Data science seems so focused on the micro scale:

deeper granularity

higher frequency…

Trends?!

My focus is broad patterns; big flock behaviours, and my objective is long range predictions.

Trends are a natural way to think, explain, and forecast.

Yet we lack tools to understand Trends, scientifically.

TrendCalculus is my unfinished research to that end.

What's a trend?

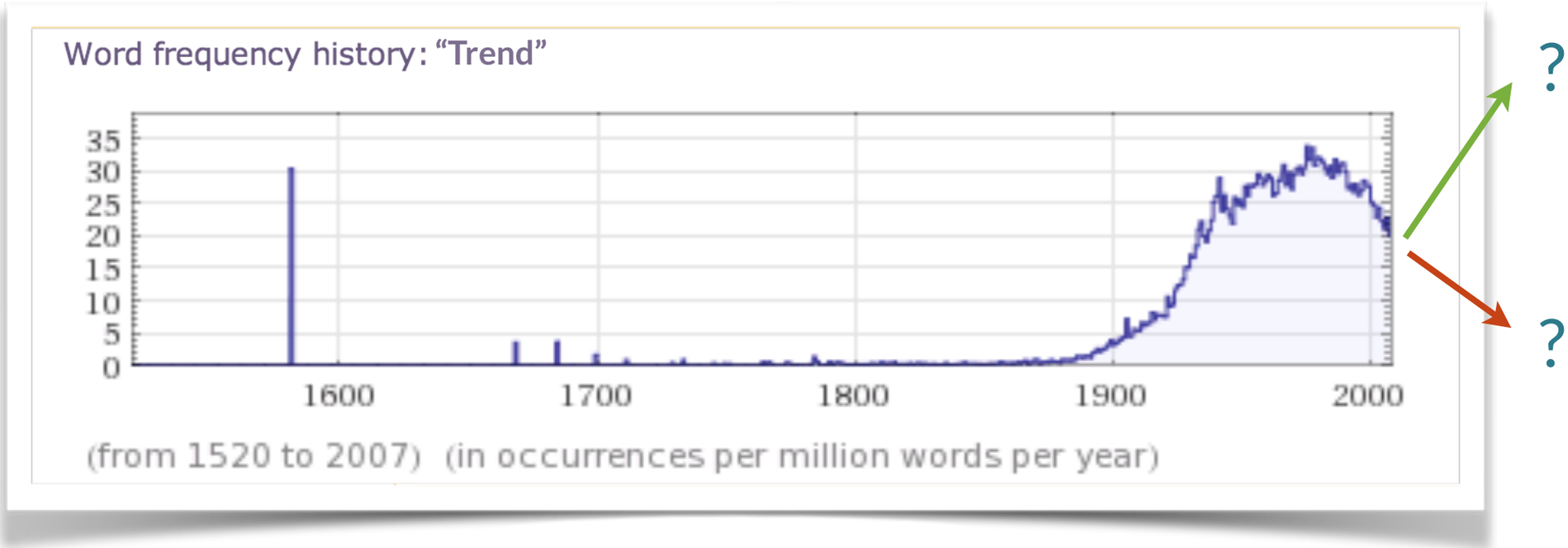
"A Trend is defined by a shift in behaviour or mentality that influences a significant amount of people." - Salomé Areias

"A Trend is the slow variation over a longer period of time, usually several years, generally associated with the structural causes affecting the phenomenon being measured." - Eurostat

400+ years of trend discussion

What do you see?

Perhaps a shift in behaviour or mentality?
Maybe a drift in language use?
How do we quantify and study the trend?



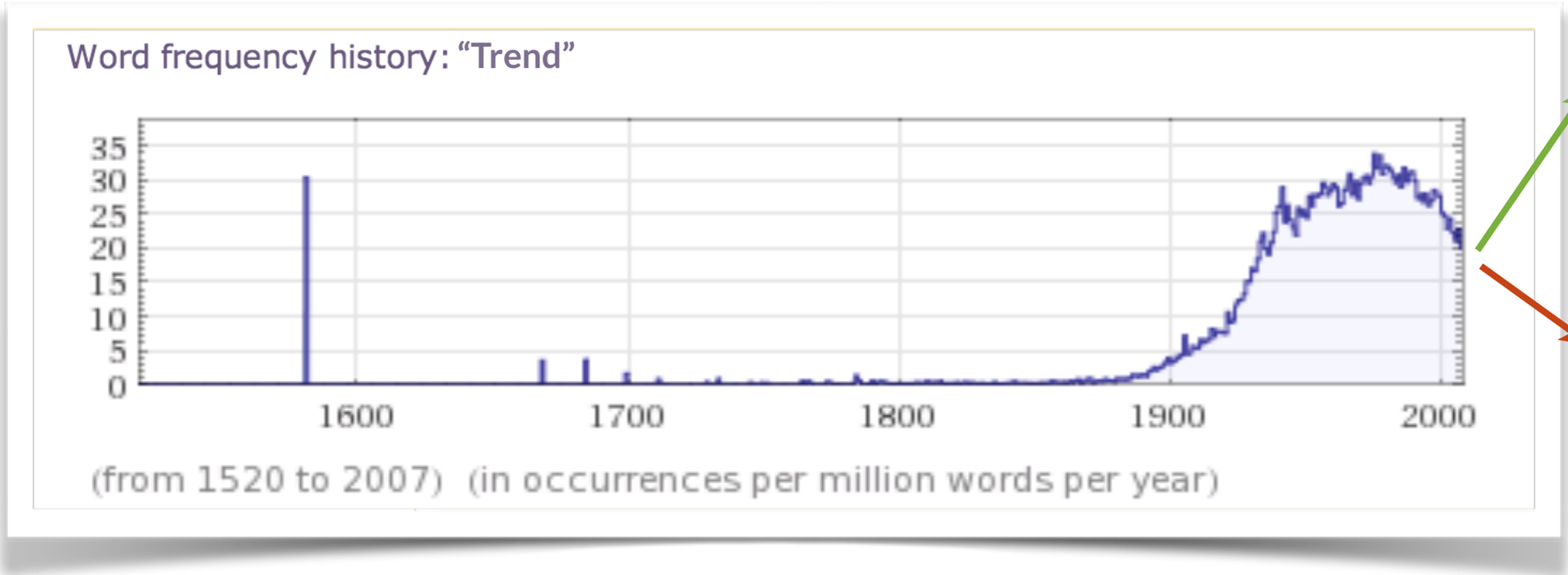
Word frequency history:

based on a Google Books sample of one million books in English; Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. "Quantitative Analysis of Culture Using Millions of Digitized Books." Science 331 (2011)

- Wolfram Alpha

400+ years of trend discussion

What might cause Trend as a topic to be losing popularity?



Word frequency history:

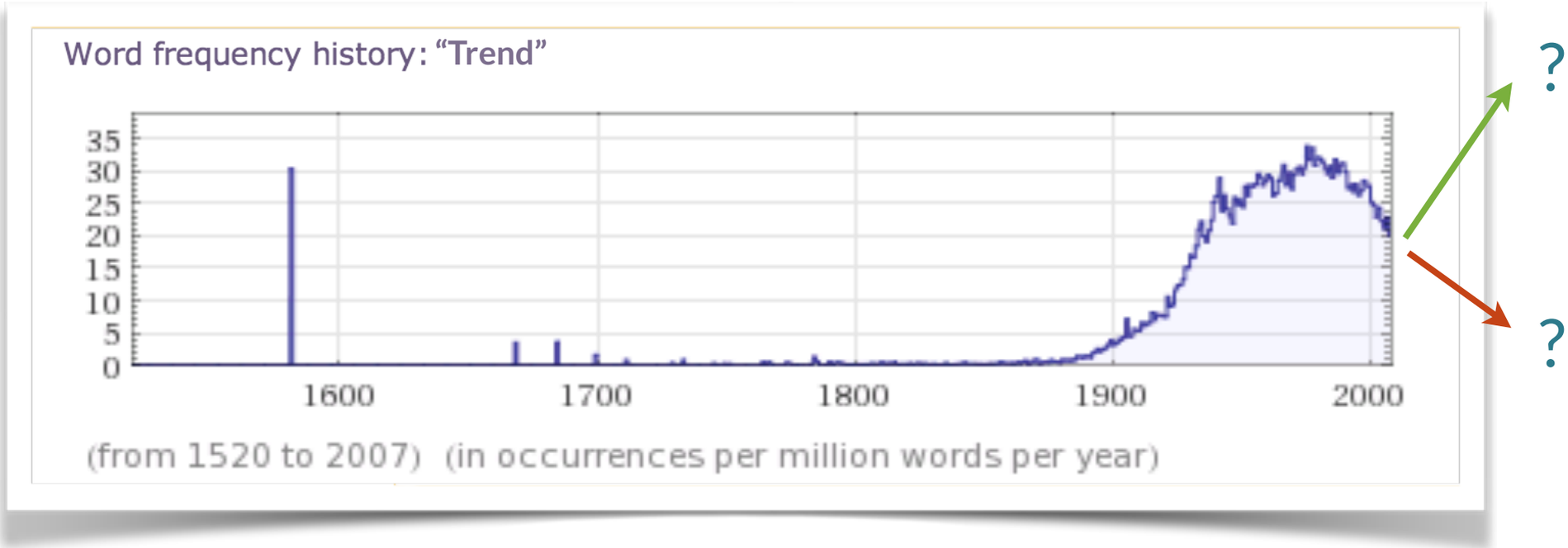
based on a Google Books sample of one million books in English; Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. "Quantitative Analysis of Culture Using Millions of Digitized Books." Science 331 (2011)

- Wolfram Alpha

400+ years of trend discussion

What might cause Trend as a topic to be losing popularity?

Maybe traditional trend analysis is flawed and the collective knows it.



Word frequency history:

based on a Google Books sample of one million books in English; Michel, J.-B., Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. "Quantitative Analysis of Culture Using Millions of Digitized Books." Science 331 (2011)

- Wolfram Alpha

What if we could do better?

What would you do

if you really understood trends

and when they reversed?

Introduction

What is TrendCalculus?

TrendCalculus

is our new, multi-scale

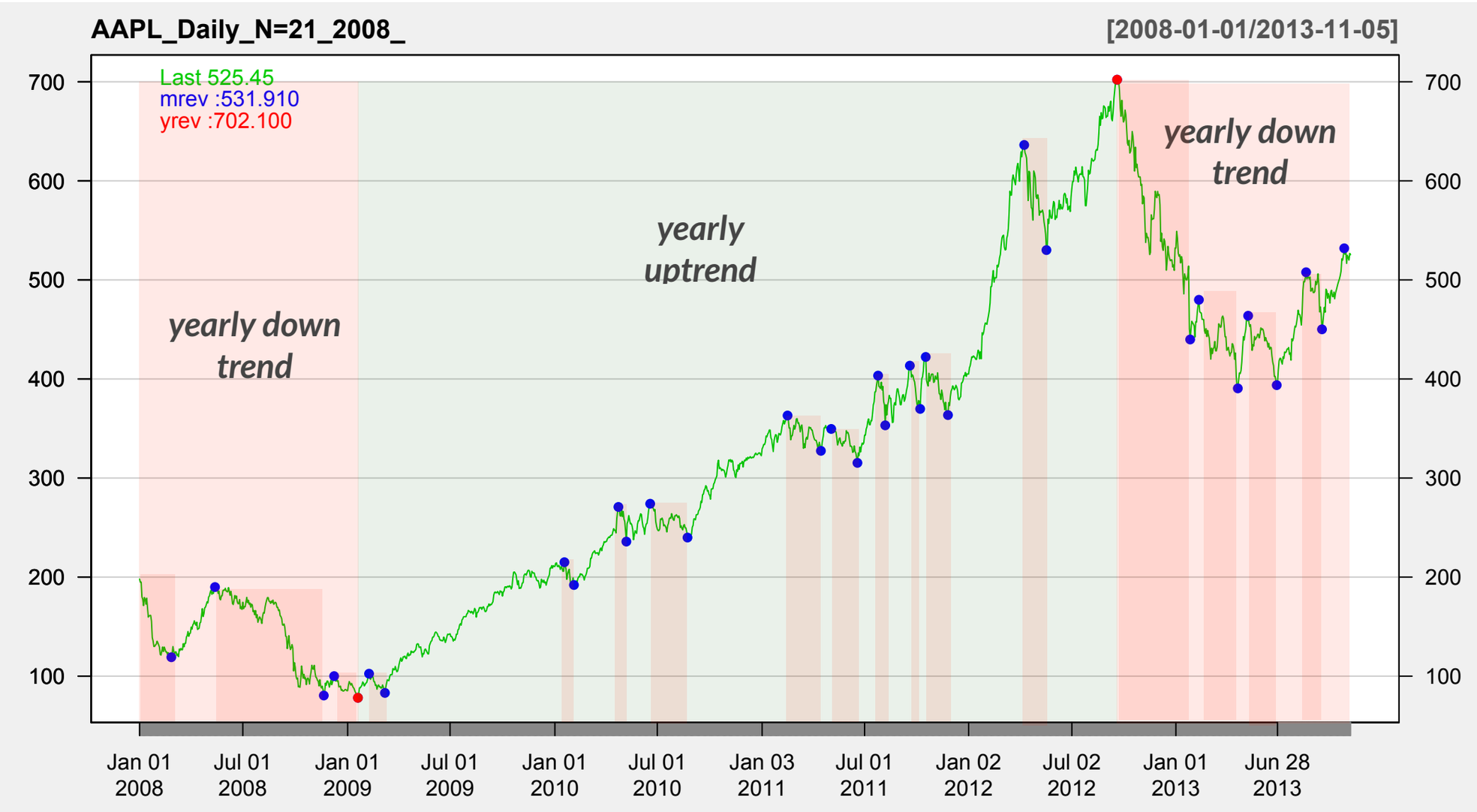
trend reversal detection algorithm

for streamed numeric data

over all timeframes.

It's pretty fast: O(n)

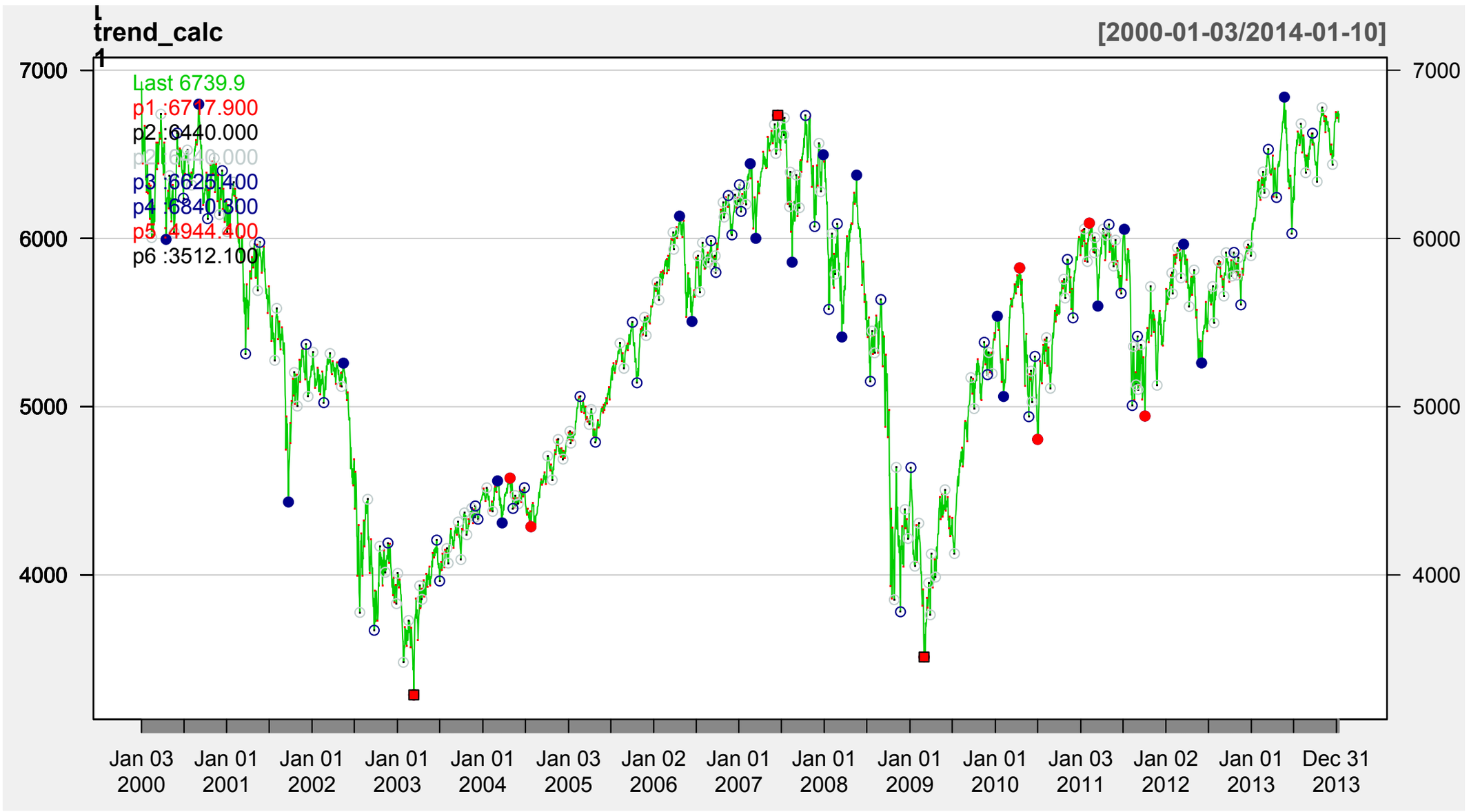
What does output look like?



two timeframes:

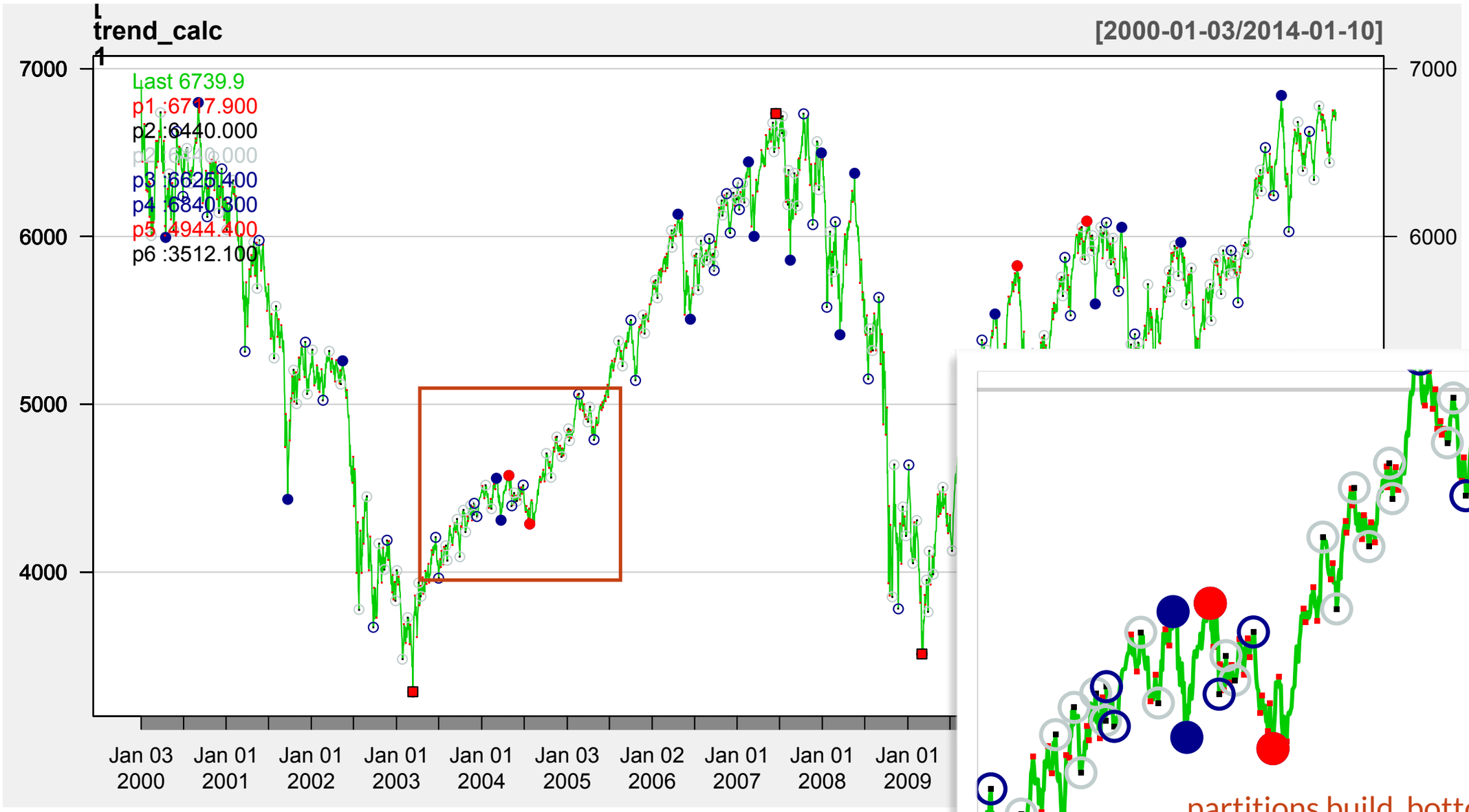
- Yearly Trend Reversal
- Monthly Trend Reversal

What does output look like?



many multiple stacked timeframes
build up into long term structures

What does output look like?



multiple stacked timeframes
build up into long term structures

Where does this fit?

Time Series Representations*

- **Model Based**
  - *Hidden Markov Models*
  - *Statistical Models*
- **Data Adaptive**
  - *Sorted Coefficients*
  - Piecewise Polynomial
    - Piecewise Linear Approximation
      - *Interpolation*
      - *Regression*
    - *Adaptive Piecewise Constant Approximation*
  - *Singular Value Approximation*
  - Symbolic
    - *Natural Language*
    - Strings
      - *Symbolic Aggregate Approximation*
      - Non Lower Bounding
        - *Value Based*
        - *Slope Based*
  - *Trees*
- **Non Data Adaptive**
  - Wavelets
    - Orthonormal
      - *Haar*
      - *Daubechies dbn n > 1*
    - Bi-Orthonormal
      - *Coiflets*
      - *Symlets*
  - *Random Mappings*
  - Spectral
    - *Discrete Fourier Transform*
    - *Discrete Cosine Transform*
    - *Chebyshev Polynomials*
  - *Piecewise Aggregate Approximation*
- **Data Dictated**
  - *Grid*
  - *Clipped Data*

Piecewise Trend Partitioning

TrendCalculus:
Is a multi-scale, bottom up, trend reversal detected, Piecewise Approximation that produces a hierarchical trend partitioning.

* Author:
Eamonn Keogh
Professor
Computer Science & Engineering Department
University of California - Riverside
Riverside, CA 92521

TrendCalculus:

It enables

“Multiscale Trend Analysis”

Multiscale Trend Analysis

What is MTA?

What does MTA offer?

If offers rich time series methods...

to better predict

to correlate timeseries

to index and compress

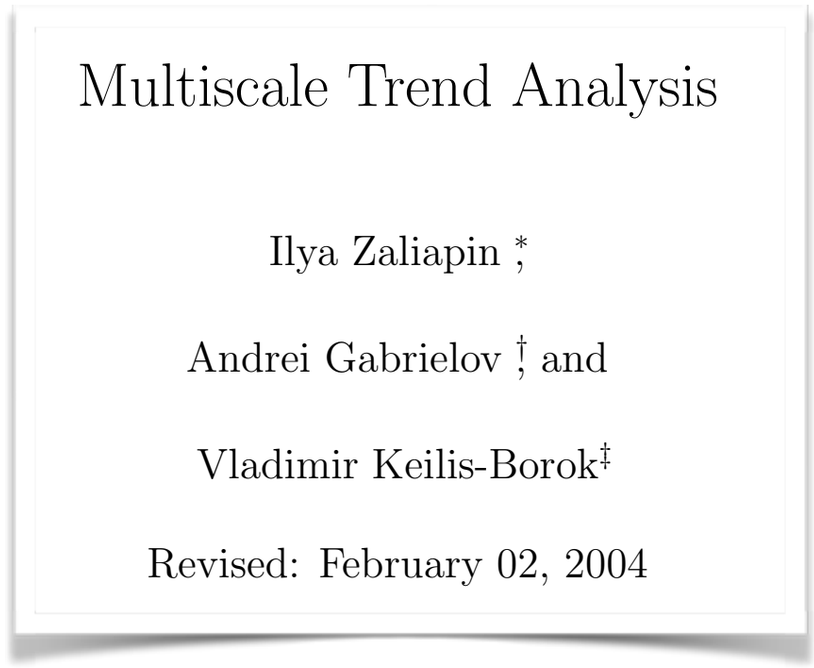
to do cross-scale retrieval of “motifs”

to build ‘episodic memory’ stores

to normalise signal extraction, reduce noise

to convert sub-symbolic data to rich symbolic data

the MTA paper is a good read:



Multiscale Trend Analysis

Ilya Zaliapin [*],

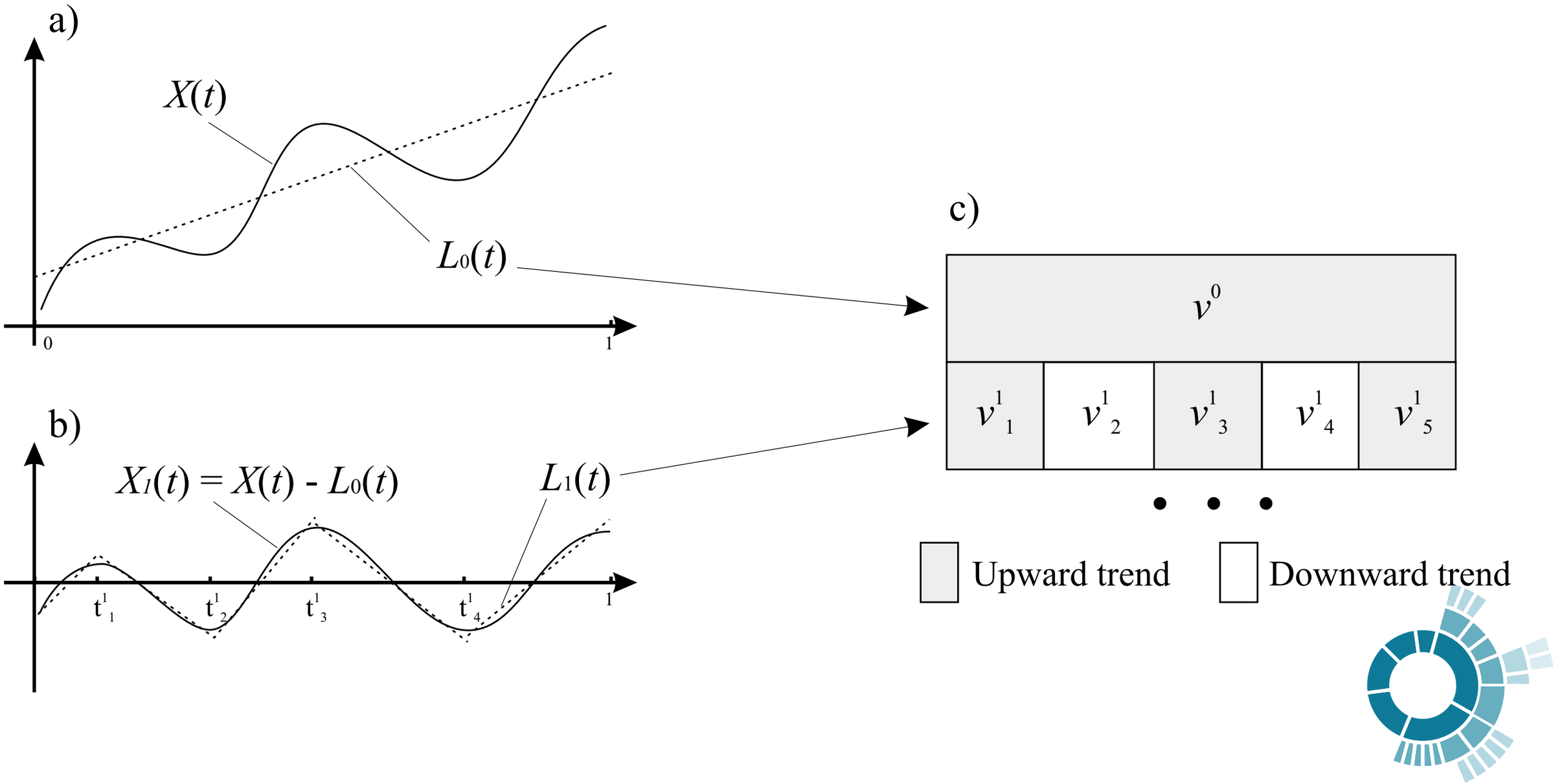
Andrei Gabrielov [†] and

Vladimir Keilis-Borok[‡]

Revised: February 02, 2004

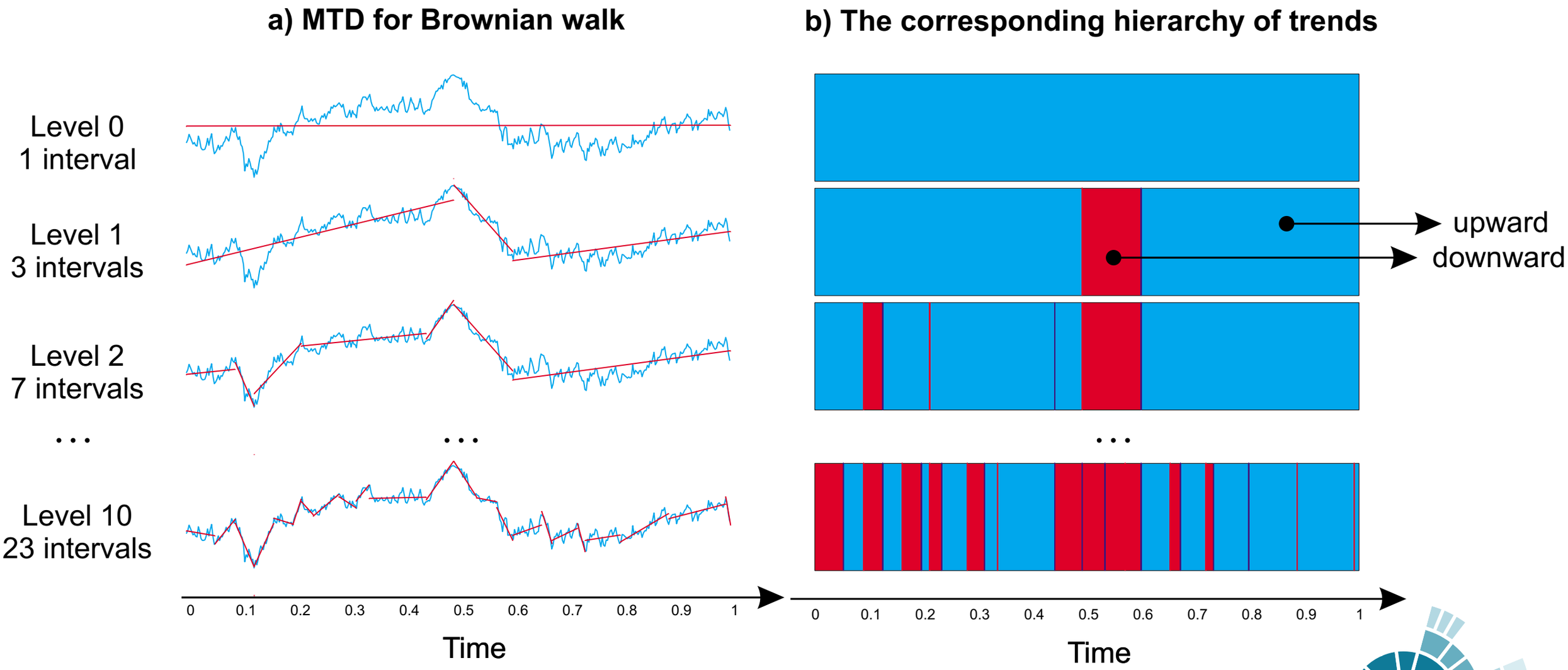
What are multi scale trends?

A time series is decomposed into local linear trends.



* these pictures are from the paper

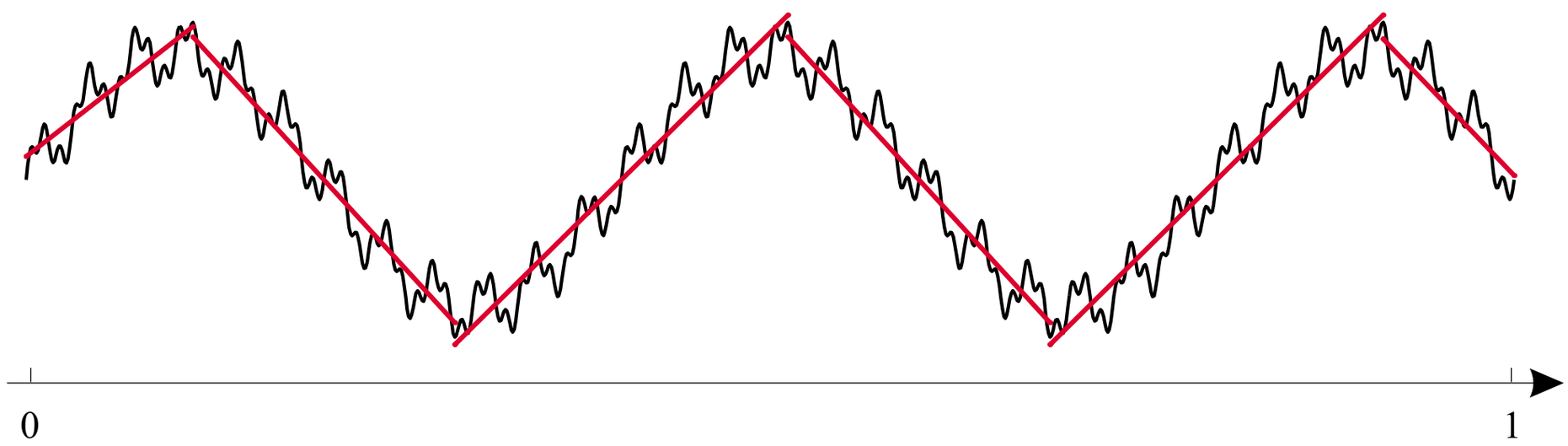
From the MTA paper….



* these pictures are from the paper

The idea is to ignore noise

b) Reversals found on the scale of interest

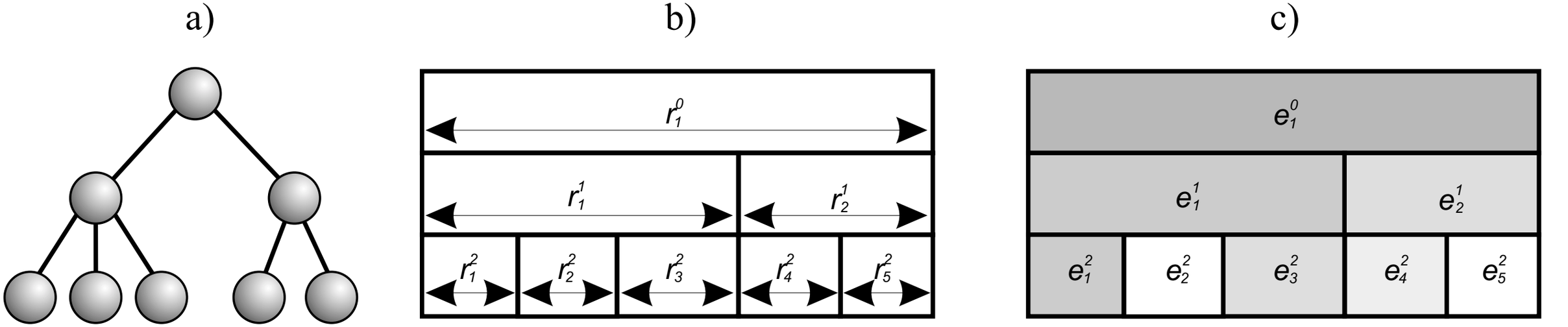


* these pictures are from the paper

Build a tree of local trends

The trends are stacked in a hierarchy.

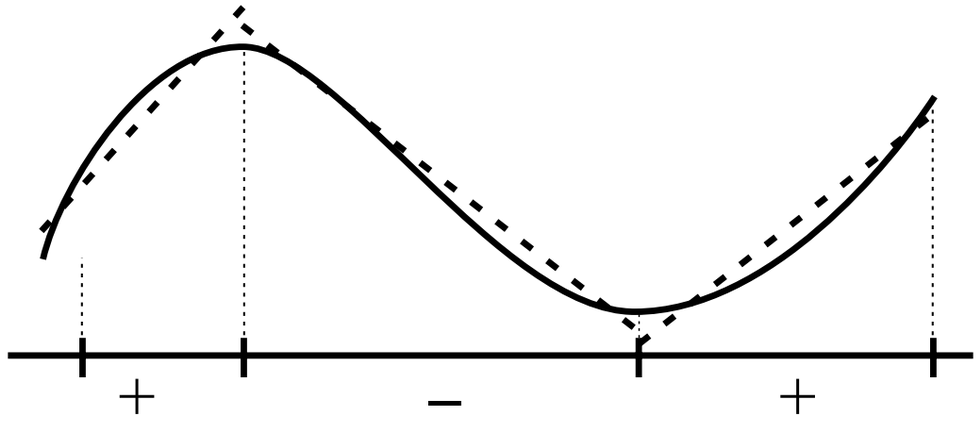
Like a b-tree, we index time series data into a shallow tree which is isn't balanced per se, but partitions are interpretable and meaningful (not necessarily stationary)



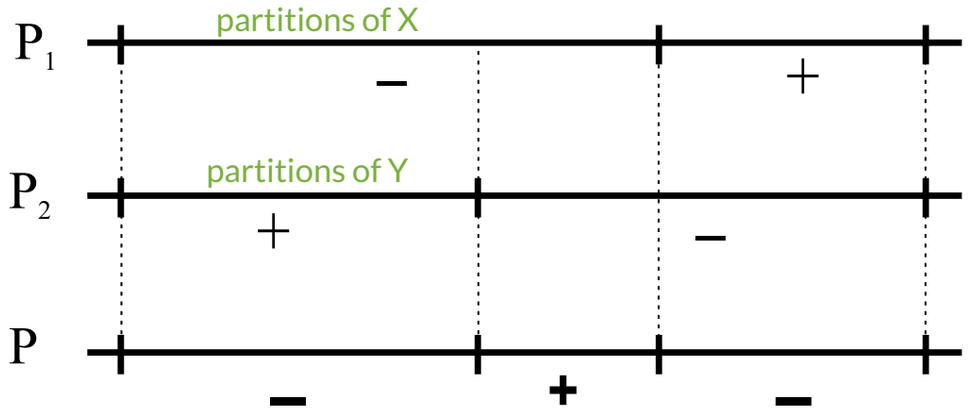
* these pictures are from the paper

Trends are signed integers

a)

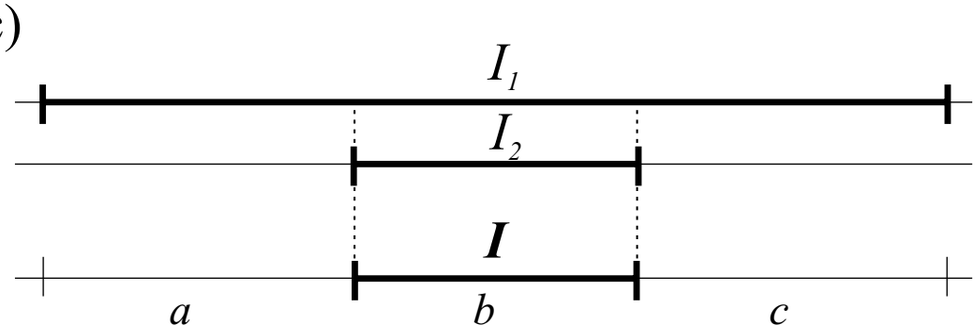


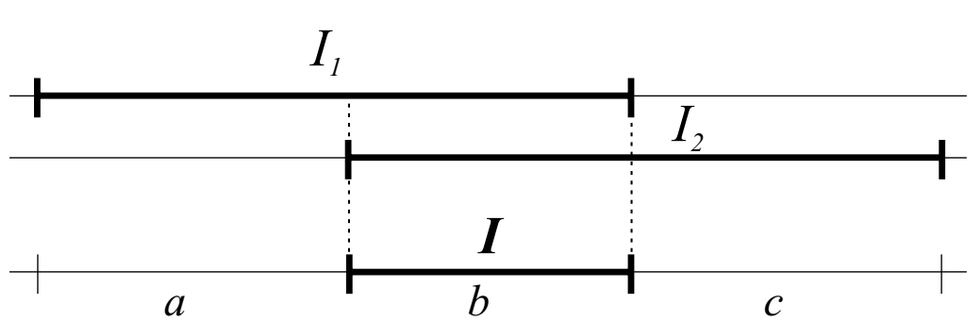
b)



Compare: multiply signs

c)





Compare: ratio of overlapping portions

Trending Together? = a correlation measure.
Multiply the trend signs at time t.
If answer is +1 they are trending together…

* these pictures are from the paper

What does ByteSumo bring?

We created a Bottom Up algorithm, that detects Trend Reversals, aka “Knots”, at a Scale, based on a window, N.

Stacked, it creates multi-scale partitions over a stream of time series data.

It’s fast, because we changed the definition of a Trend (?!)

Yes - We abandoned linear regressions…

Our definition is:

Rising = Higher Highs, Higher Lows

Falling = Lower Lows, Lower Highs

Let's see it in action!

```
Andrews-MacBook-Pro:src andrewmorgan$ cat db_output.csv | wc -l
   21499
Andrews-MacBook-Pro:src andrewmorgan$ time cat db_output.csv |
 lua /Users/andrewmorgan/nix/TrendCalculus/MultiScale/src/tcalc
5.lua -F "," -OFS "," -p 1 -h -H -n 200
ts_id,date,value,info,pass,bardate
.FTSE,1940-09-05,62.2,Bottom,1,1940-12-23
.FTSE,1944-10-17,147.2,Top,1,1944-10-23
.FTSE,1944-11-20,135.5,Bottom,1,1945-06-13
.FTSE,1945-07-09,149.2,Top,1,1946-02-01
.FTSE,1945-10-09,133.3,Bottom,1,1946-02-01
.FTSE,1947-03-17,177,Top,1,1947-05-14
.FTSE,1947-10-28,131.2,Bottom,1,1948-01-02
.FTSE,1948-02-25,162.4,Top,1,1948-08-23
.FTSE,1949-12-29,125.7,Bottom,1,1950-07-24
.FTSE,1951-07-24,176.8,Top,1,1951-11-02
.FTSE,1952-07-31,129.8,Bottom,1,1953-04-03
.FTSE,1955-08-18,281.9,Top,1,1956-04-27
.FTSE,1956-12-28,203.4,Bottom,1,1957-02-01
.FTSE,1957-08-02,261.4,Top,1,1957-11-08
.FTSE,1958-03-24,194.4,Bottom,1,1958-08-15
.FTSE,1960-09-16,433.7,Top,1,1960-12-02
.FTSE,1961-06-05,460.5,Top,1,1961-09-08
.FTSE,1961-01-03,369.5,Bottom,1,1961-09-08
.FTSE,1962-07-11,318.3,Bottom,1,1963-03-22
.FTSE,1964-10-13,475.7,Top,1,1965-07-09
.FTSE,1965-08-06,395.1,Bottom,1,1966-04-15
.FTSE,1966-06-21,471.2,Top,1,1967-01-20
.FTSE,1966-11-11,357.9,Bottom,1,1967-01-20
.FTSE,1968-09-12,657.2,Top,1,1969-05-09
.FTSE,1971-02-16,384.4,Bottom,1,1971-08-27
.FTSE,1972-05-03,684.5,Top,1,1972-06-02
.FTSE,1974-12-06,183.8,Bottom,1,1975-06-27
.FTSE,1976-04-02,529.9,Top,1,1976-04-02
.FTSE,1976-09-28,334.1,Bottom,1,1977-01-07
.FTSE,1977-08-05,691.6,Top,1,1977-10-14
.FTSE,1978-01-20,545.7,Bottom,1,1978-07-21
.FTSE,1979-03-21,703.4,Top,1,1979-04-27
.FTSE,1979-10-04,511.6,Bottom,1,1980-02-01
.FTSE,1981-03-31,752.1,Top,1,1981-08-14
.FTSE,1981-08-27,576.1,Bottom,1,1982-05-21
.FTSE,1987-07-16,2443.3999,Top,1,1987-10-02
.FTSE,1987-11-09,1565.2,Bottom,1,1988-07-08
.FTSE,1990-01-03,2463.7,Top,1,1990-01-19
.FTSE,1990-09-28,1990.2,Bottom,1,1990-10-26
.FTSE,1992-05-08,2725.7,Top,1,1992-05-08
.FTSE,1992-08-25,2281,Bottom,1,1993-02-12
.FTSE,1994-02-02,3520.3,Top,1,1994-08-26
.FTSE,1994-12-12,2943.3999,Bottom,1,1995-06-02
.FTSE,1998-04-06,6105.7998,Top,1,1998-06-26
.FTSE,1998-10-05,4648.7002,Bottom,1,1999-04-02
.FTSE,1999-12-30,6930.2002,Top,1,2000-01-07
.FTSE,2003-03-12,3287,Bottom,1,2003-11-07
.FTSE,2007-06-15,6732.3999,Top,1,2007-09-07
.FTSE,2009-03-03,3512.1,Bottom,1,2009-03-20
.FTSE,2011-02-08,6091.3,Top,1,2011-07-08
.FTSE,2011-10-04,4944.4,Bottom,1,2012-04-13

real    0m12.822s
user    0m12.400s
sys     0m0.412s
Andrews-MacBook-Pro:src andrewmorgan$
```

Let's try the **FTSE 100**, extended back to 1935 via the FTSE 30 data.

Time Series length: **21499 records** (daily closes)

This run uses window size of **n=200** (market days)

The process in Lua creates lots of intermediate calculations for each window size from n down to 1 … so it should be slow….

Total run time is **~13 seconds** on my mac.

Output is shown left: **51** major trend reversals found that approximate the time series.

Alternatively, we could say we have "generalised the time series" into 51 important change points.

it's true luajit can speed this up…

but is how else might we be able to to speed it up?

Let's see it in action!

Let's try another way. Stacking the calculations: i.e. Pipe output back through the algo again x3.

```
Andrews-MacBook-Pro:src andrewmorgan$ cat test.sh
# for fast multiscale reversal finding, I run it on an n=5, and stream output back through the algo 3 more times in a stack.
# this will create a B-Tree depth of max 4.
# will it run faster?

p=5
z=5

cat db_output.csv          |  lua /Users/andrewmorgan/nix/TrendCalculus/MultiScale/src/tcalc5.lua -F "," -OFS "," -p 1 -h -H -n $z \
        | tee reversals1.csv |  lua /Users/andrewmorgan/nix/TrendCalculus/MultiScale/src/tcalc5.lua -F "," -OFS "," -p 2 -h -H -n $p \
        | tee reversals2.csv | lua /Users/andrewmorgan/nix/TrendCalculus/MultiScale/src/tcalc5.lua -F "," -OFS "," -p 3 -h -H -n $p  \
        | tee reversals3.csv | lua /Users/andrewmorgan/nix/TrendCalculus/MultiScale/src/tcalc5.lua -F "," -OFS "," -p 4 -h -H -n $p   > reversals4.csv

Andrews-MacBook-Pro:src andrewmorgan$ time . test.sh

real    0m1.688s
user    0m1.687s
sys     0m0.155s
Andrews-MacBook-Pro:src andrewmorgan$ ls rever*.csv
reversals.csv   reversals1.csv  reversals2.csv  reversals3.csv  reversals4.csv
Andrews-MacBook-Pro:src andrewmorgan$ time . test.sh

real    0m1.359s
user    0m1.377s
sys     0m0.142s
Andrews-MacBook-Pro:src andrewmorgan$ ls -lart reversal*.csv
-rw-r--r--  1 andrewmorgan  staff      0  9 Jan 14:31 reversals.csv
-rw-r--r--  1 andrewmorgan  staff     35  9 Jan 15:35 reversals4.csv
-rw-r--r--  1 andrewmorgan  staff   1210  9 Jan 15:35 reversals3.csv
-rw-r--r--  1 andrewmorgan  staff  10537  9 Jan 15:35 reversals2.csv
-rw-r--r--  1 andrewmorgan  staff  87605  9 Jan 15:35 reversals1.csv
Andrews-MacBook-Pro:src andrewmorgan$ wc -l reversal*.csv
       0 reversals.csv
    2080 reversals1.csv
     250 reversals2.csv
      29 reversals3.csv
       1 reversals4.csv
    2360 total
Andrews-MacBook-Pro:src andrewmorgan$ wc -l db_output.csv
   21499 db_output.csv
Andrews-MacBook-Pro:src andrewmorgan$ head -5 db_output.csv
ric,date,price,info,pass,start_date,start_price
.FTSE,1935-11-13,125.9,raw,0,,
.FTSE,1935-11-14,125.7,raw,0,,
.FTSE,1935-11-15,125.4,raw,0,,
.FTSE,1935-11-16,125.4,raw,0,,
Andrews-MacBook-Pro:src andrewmorgan$
```

There is practically a magnitude improvement in performance when stacking.

With a setting of N=5, I just processed the stack of 4 runs in less than **2 seconds** using straight lua on my mac for 21,499 input records.

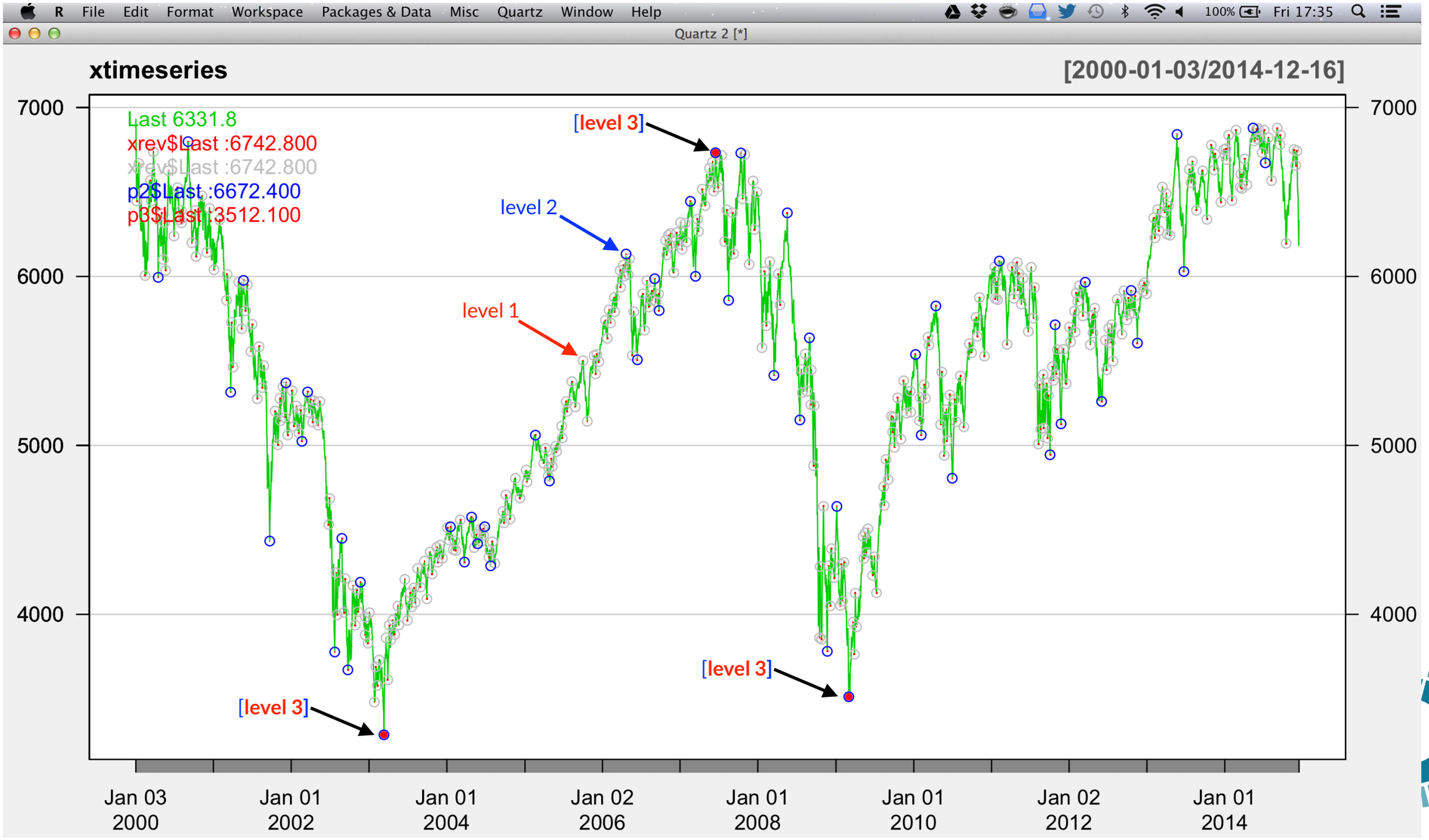
that's ~10k streamed records per second.

With luajit it will drop further! The partitions in the trend tree we calculated are:

level 4	= 0	trend reversals
level 3	= 28	trend reversals
level 2	= 249	trend reversals
level 1	= 2,079	trend reversals.

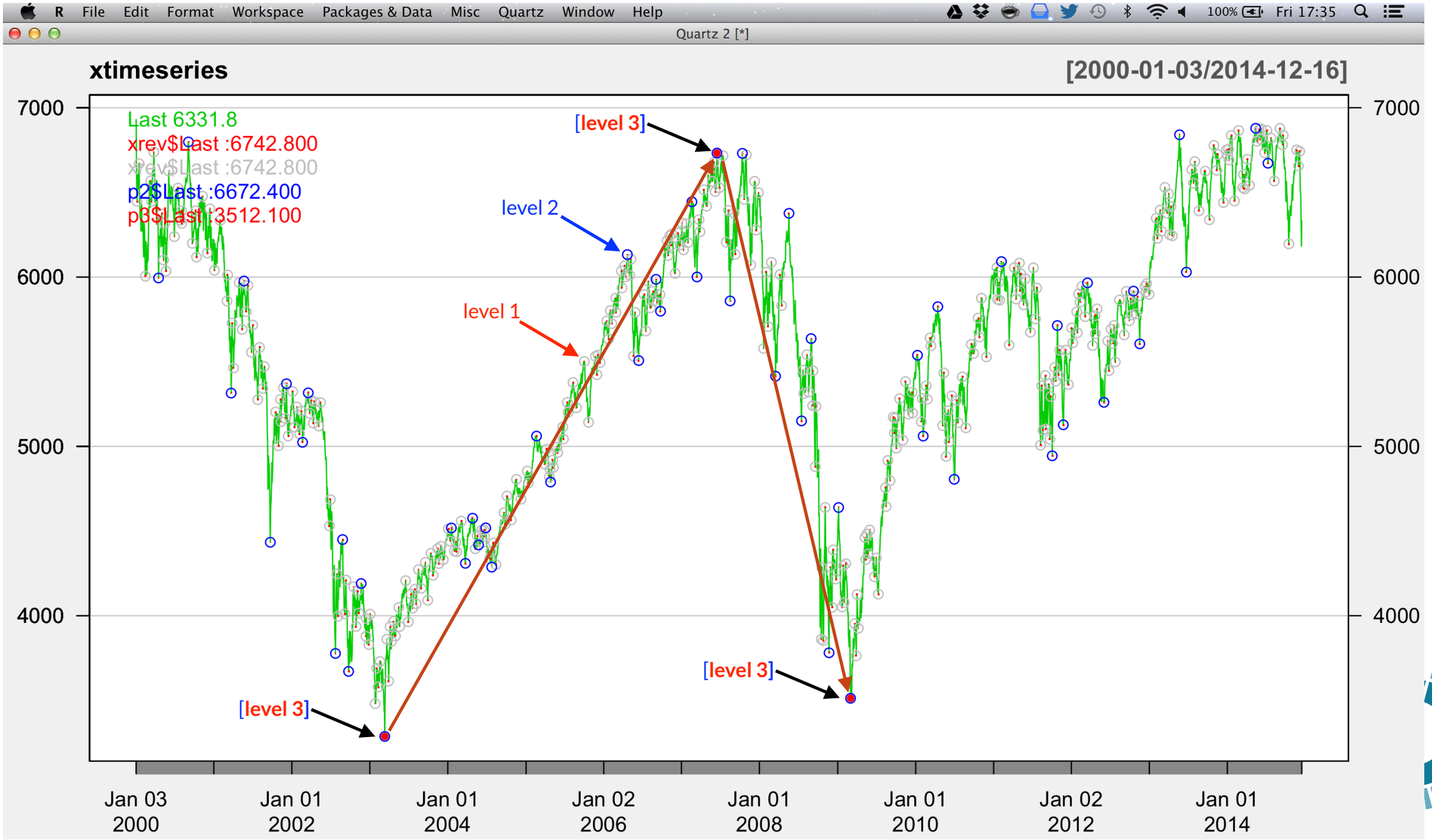
Let's see it in action!

Here is the last 14 years of the stacked output. The 3 levels of partitions are seen nested:



Let's see it in action!

Here is the last 14 years of the stacked output. The 3 levels of partitions are seen nested:

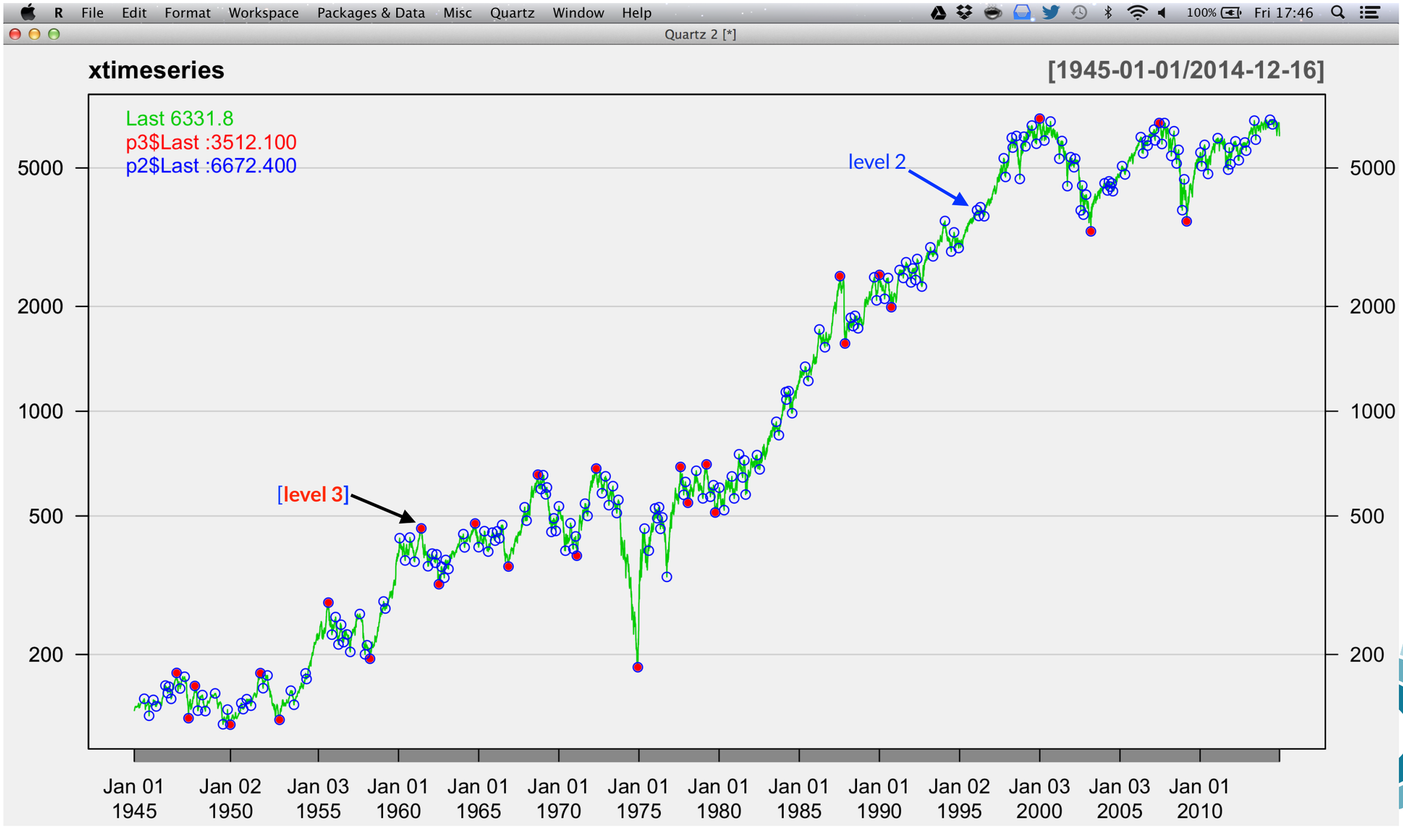


Let's see it in action!

Here is the last 14 years of the stacked output. The 3 levels of partitions are seen nested:

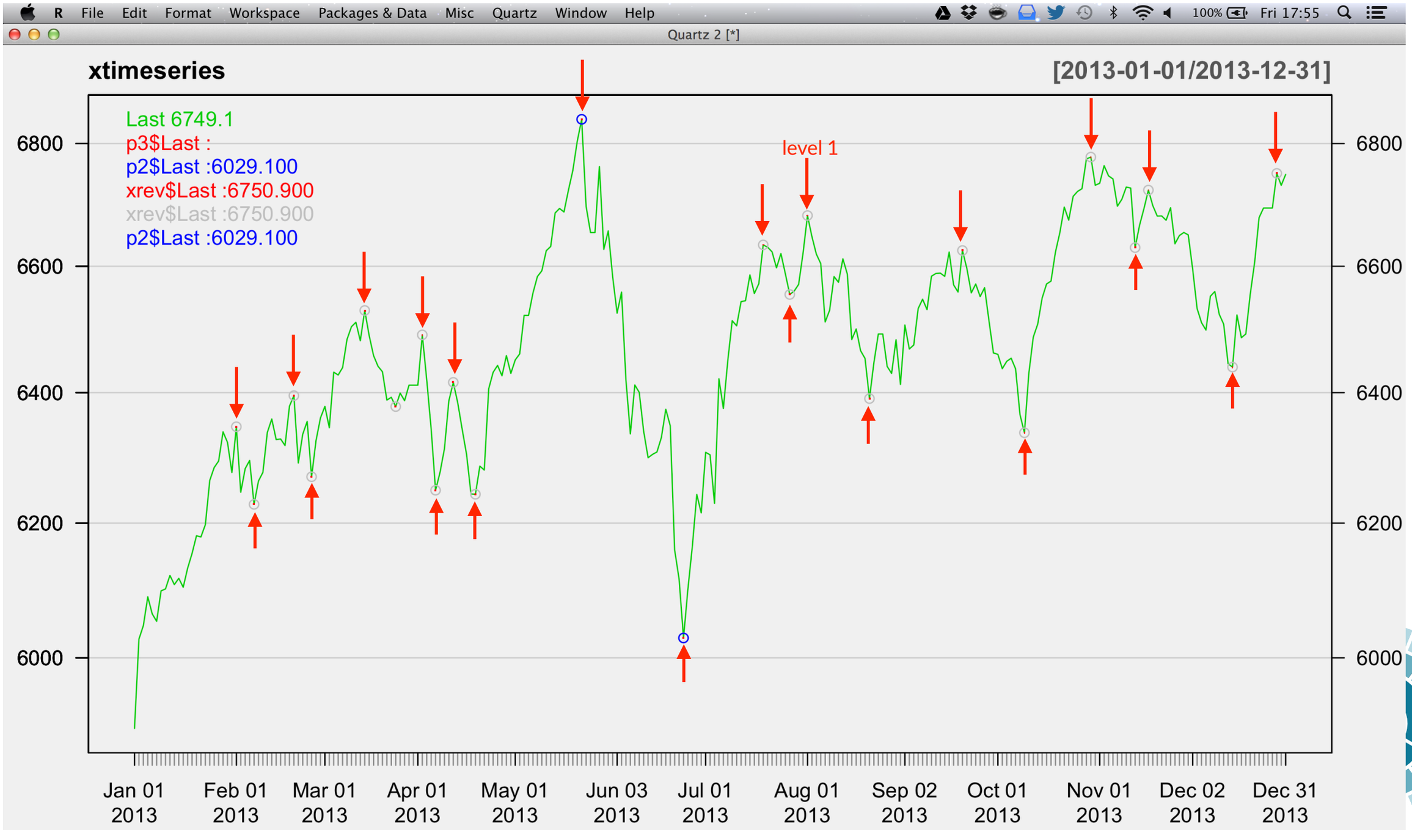
Let's see it in action!

Zoom out. Here is from 1945 to Present. We see the 28 "level 3" partitions as red knots.



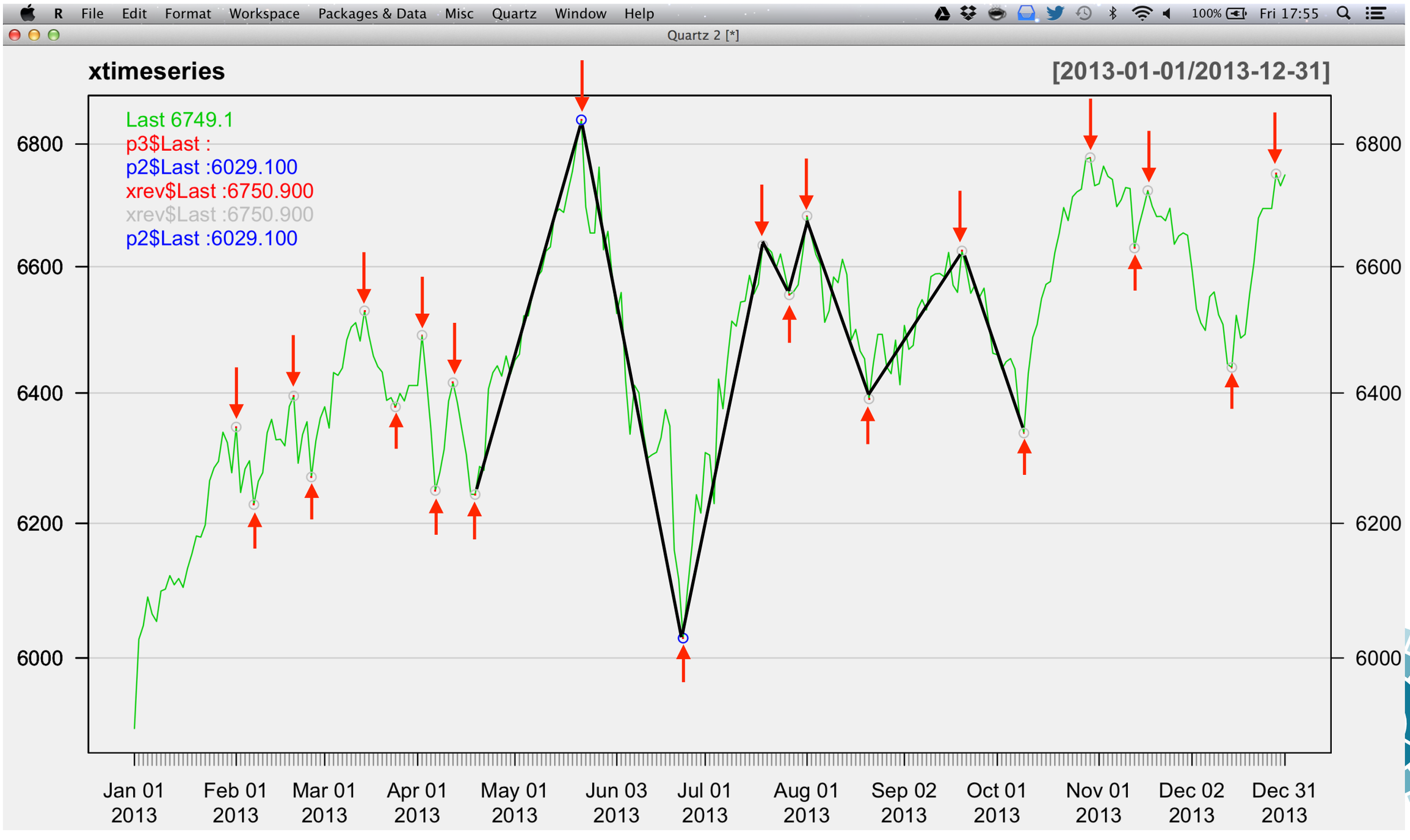
Let's see it in action!

Zoom in. Here is 2013. Here we can see some of the 2,079 fine grain "level 1" reversals up close:



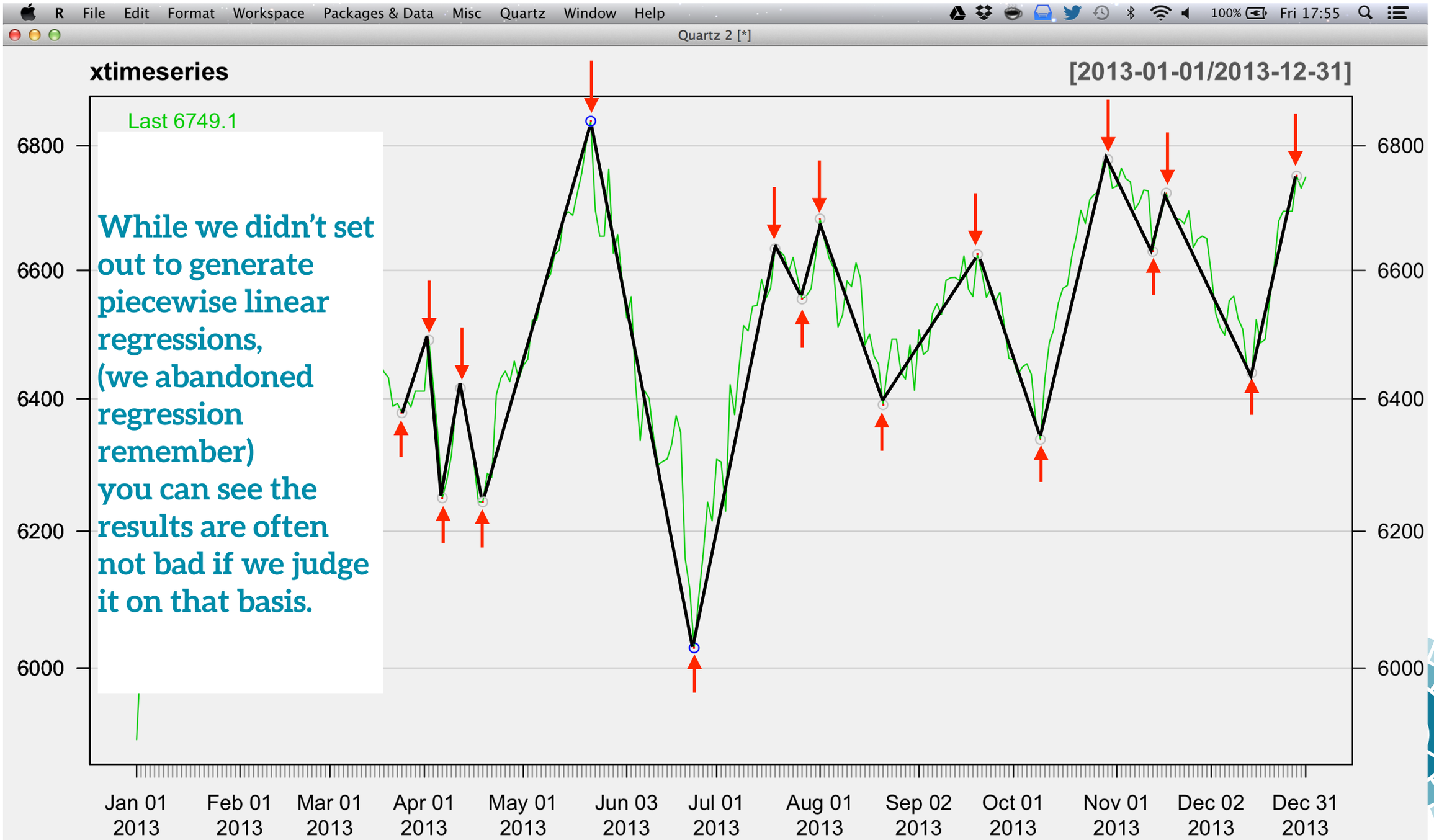
Let's see it in action!

Zoom in. Here is 2013. Here we can see some of the 2,079 fine grain "level 1" reversals up close:



Let's see it in action!

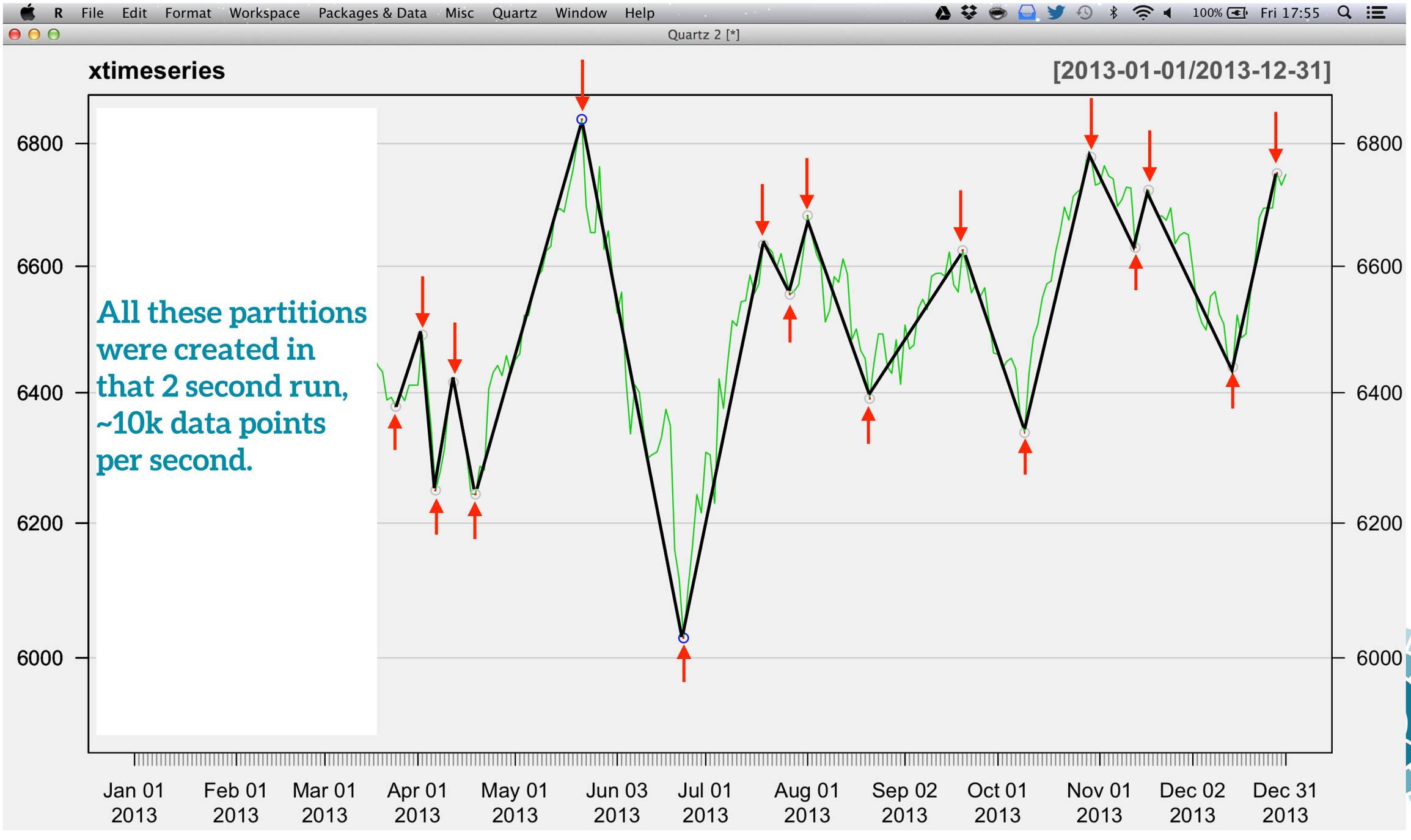
Zoom in. Here is 2013. Here we can see some of the 2,079 fine grain "level 1" reversals up close:



While we didn't set out to generate piecewise linear regressions, (we abandoned regression remember) you can see the results are often not bad if we judge it on that basis.

Let's see it in action!

Zoom in. Here is 2013. Here we can see some of the 2,079 fine grain "level 1" reversals up close:



New Directions for Trend Calculus

Our research is uncovering whole new avenues of study.

A Rolling Trend score?

This involves moving away from fixed windows of N and to rolling arrays for all timeframes to N.

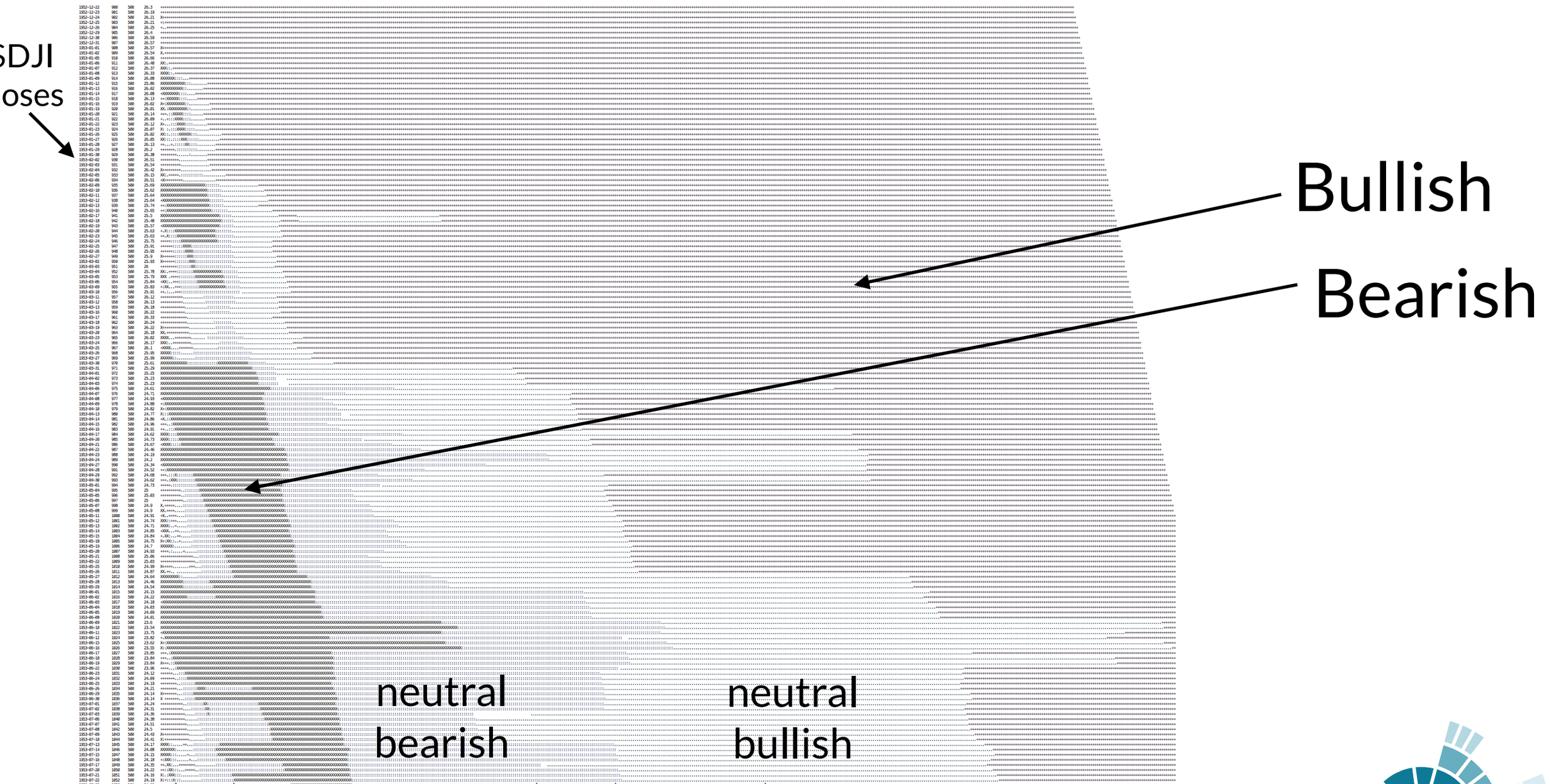
The information revealed is not trend reversals, but the underlying data used in their calculation.

I will output these internal arrays to feed deep learning algorithms as a form of “trend feature generator”.

For display, I turn values into symbols, and we can see rich patterns emerging from the trends across all scales.

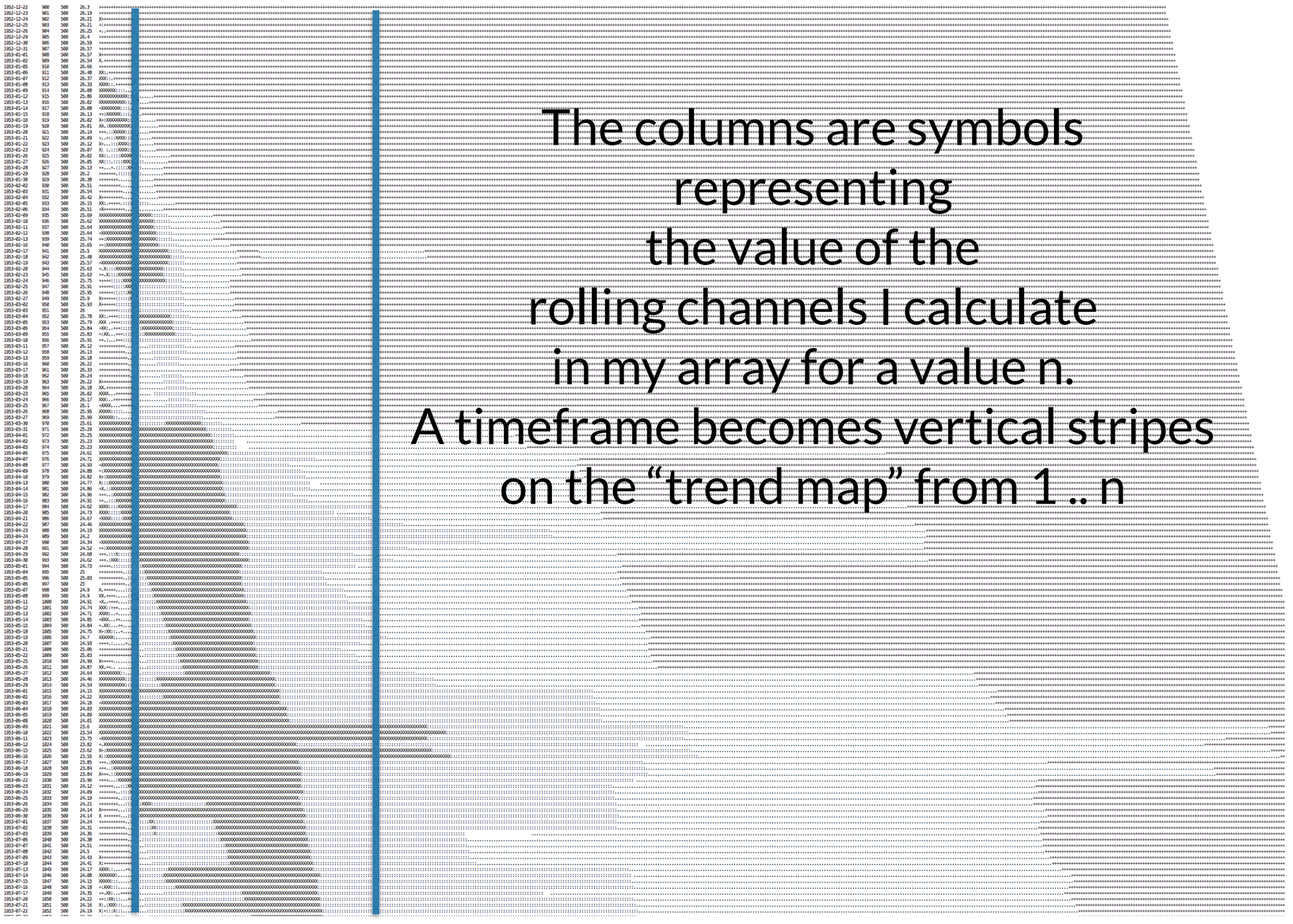
Quants who reviewed this said: “ah, it shows the relationship of the price to the Pivot Points”

Rolling Trends - all timeframes



Here I present the trend score as a symbol for each timeframe to a max N to build a “multi scale trend map”

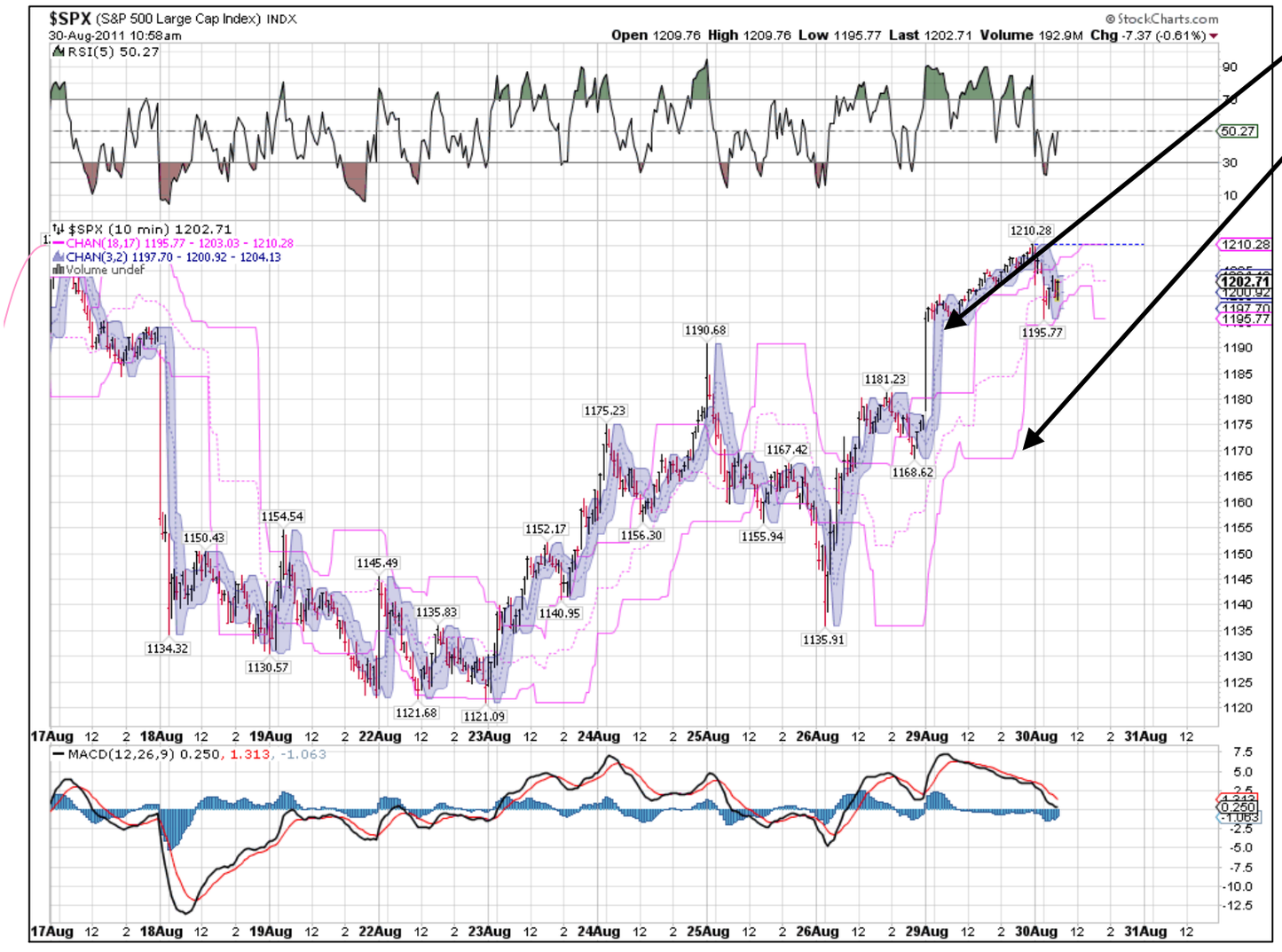
Rolling Trends - all timeframes



What are we seeing?

a small N is a short channel

bigger N is a long channel

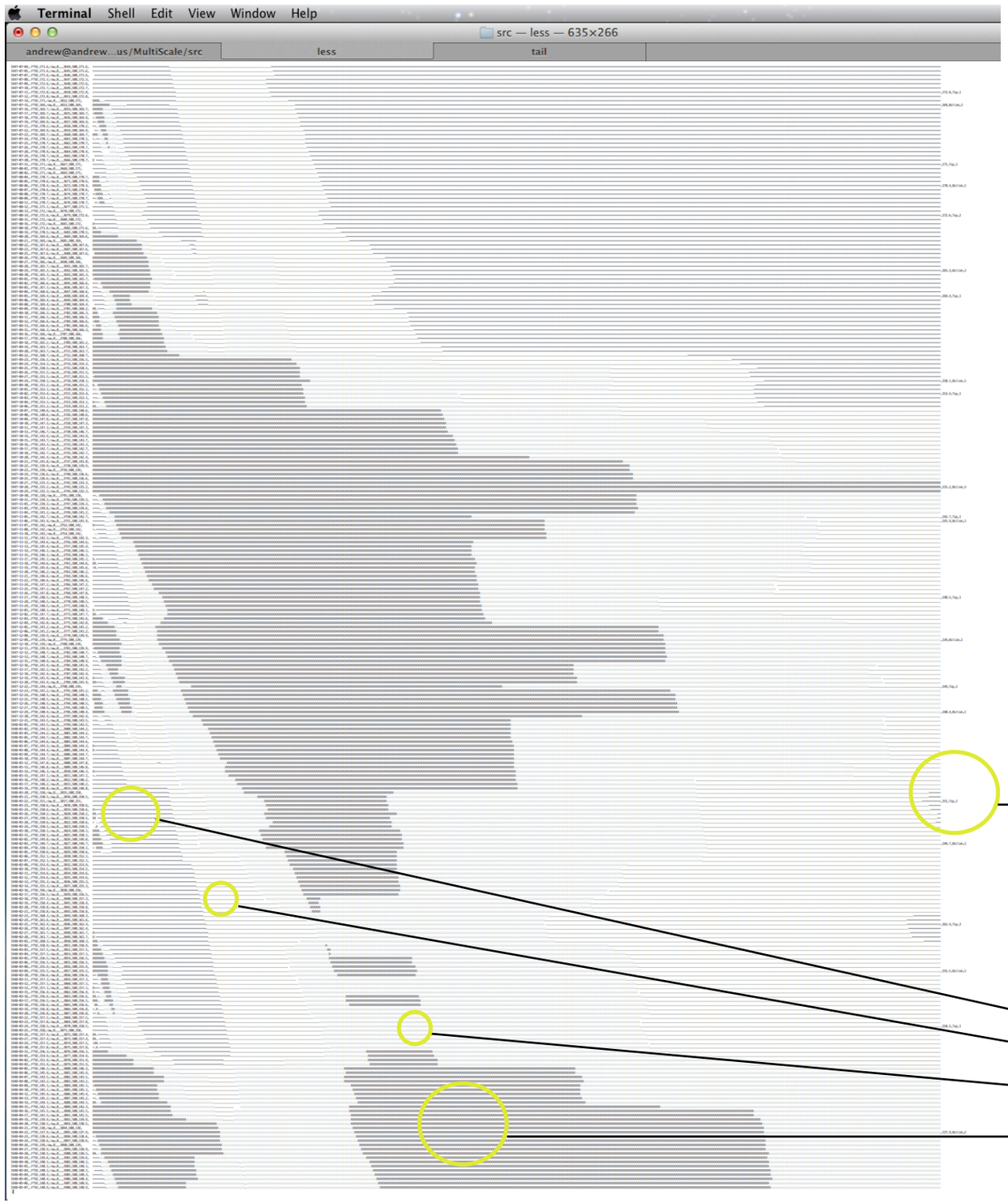


When price above a channel, trend is up.
When below a channel, it's down.

When the price is in a channel
the trend is neutral.

here we see two timeframes.

Trend Maps of all timeframes.



The next steps are to use all these rich inputs to see if we can make long range predictions…

.. by for instance feeding deep learning algorithms with all these trends to predict future trend reversals.

It means I'll use TrendCalculus to generate interesting trend features.

Lots of potential for further work.

Prediction

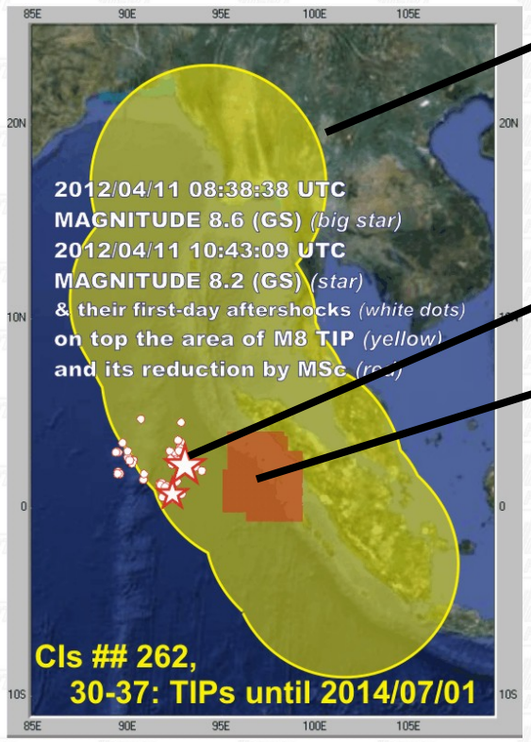
MTA was created by people predicting earthquakes

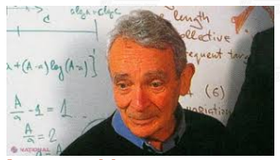
Predicting Earthquakes?!?

The 11 April 2012, M8.6 and M8.2 earthquakes OFF THE WEST COAST OF NORTHERN SUMATRA did confirm an alarm TIP reported in January, in the regular 2010a Update of the M8-MSc predictions of the Global Test of M8 (Healy et al. 1992; password protected URL http://www.mitp.ru/en/restricted_global/2012a/2012am8.html; yellow outline in the attached figure). The earthquake epicenters missed the reduced area of alarm (red outline) diagnosed in the second approximation due to inapplicability of the MSc algorithm outside bulk distribution of seismic activity. Nevertheless, it appears remarkable that the reduced area is about the same as the area of the 11 April 2012 first-day aftershocks located at about the same latitudes.

The 11 April 2012 great earthquakes have ruptured the conjugate faults, about 300 and 500 km each in the oceanic lithosphere of Indo-Australian plate. Both are strike-slip intra-oceanic-plate events with epicenters in an area of sparse seismicity, some 100 km and 200 km to the southwest of the major seismic belt of the subduction zone next to the complex junction of India, Australia, Sunda, and Burma plates. These events continue a series that can be attributed to the 26 December 2004, M9.1 Sumatra-Andaman mega-thrust, followed by the 28 March 2005, M8.6 great Nias earthquake. In course the Global Test of M8 a segment of the subduction zone from Burma to Southern Sumatra was recognized as capable of producing magnitude M8.0+ event starting from July 2005-January 2006, which prediction was already confirmed with a pair of the great 12 September 2007, M8.5 SOUTHERN SUMATRA and M8.1 KEPULAUAN MENTAWAI REGION, INDONESIA earthquakes (http://www.mitp.ru/en/restricted_global/2007b/m8t5confirmed.html).

(Note: The M8 algorithm provides prediction in the first approximation, and the algorithm MSc, if the data permit, narrows down the area covered by alarm. Both apply to the null approximation delivered by identifying earthquake-prone zones, e.g. "active fault zones", "D-intersections or knots", etc.)





Back to "List of predictions"

http://www.mitp.ru/en/index.html

Area of predicted Earthquake

Actual Earthquake

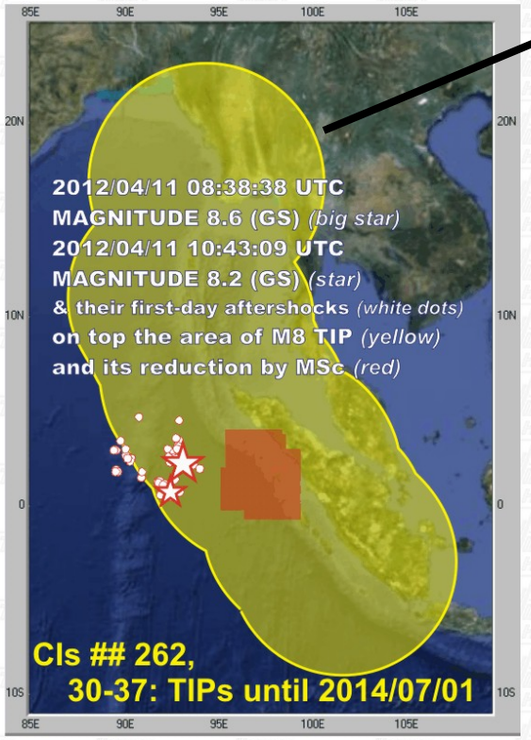
Focus of Prediction

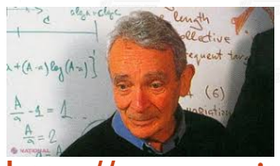
Predicting Earthquakes?!?

The 11 April 2012, M8.6 and M8.2 earthquakes OFF THE WEST COAST OF NORTHERN SUMATRA did confirm an alarm TIP reported in January, in the regular 2010a Update of the M8-MSc predictions of the Global Test of M8 (Healy et al. 1992; password protected URL http://www.mitp.ru/en/restricted_global/2012a/2012am8.html; yellow outline in the attached figure). The earthquake epicenters missed the reduced area of alarm (red outline) diagnosed in the second approximation due to inapplicability of the MSc algorithm outside bulk distribution of seismic activity. Nevertheless, it appears remarkable that the reduced area is about the same as the area of the 11 April 2012 first-day aftershocks located at about the same latitudes.

The 11 April 2012 great earthquakes have ruptured the conjugate faults, about 300 and 500 km each in the oceanic lithosphere of Indo-Australian plate. Both are strike-slip intra-oceanic-plate events with epicenters in an area of sparse seismicity, some 100 km and 200 km to the southwest of the major seismic belt of the subduction zone next to the complex junction of India, Australia, Sunda, and Burma plates. These events continue a series that can be attributed to the 26 December 2004, M9.1 Sumatra-Andaman mega-thrust, followed by the 28 March 2005, M8.6 great Nias earthquake. In course the Global Test of M8 a segment of the subduction zone from Burma to Southern Sumatra was recognized as capable of producing magnitude M8.0+ event starting from July 2005-January 2006, which prediction was already confirmed with a pair of the great 12 September 2007, M8.5 SOUTHERN SUMATRA and M8.1 KEPULAUAN MENTAWAI REGION, INDONESIA earthquakes (http://www.mitp.ru/en/restricted_global/2007b/m8t5confirmed.html).

(Note: The M8 algorithm provides prediction in the first approximation, and the algorithm MSc, if the data permit, narrows down the area covered by alarm. Both apply to the null approximation delivered by identifying earthquake-prone zones, e.g. "active fault zones", "D-intersections or knots", etc.)





Back to "List of predictions"

http://www.mitp.ru/en/index.html

I think they did it by finding unusual divergences "(Un)Correlations" between trends in different geophysical measures and these were found to be precursors to major earthquakes…

see here for detail:

Temporal (Un)correlations Between Coda Q and Seismicity: Multiscale Trend Analysis

http://link.springer.com/article/10.1007%2Fs00024-004-2643-x

THANK YOU

andrew@bytesumo.com

Andrew Morgan is a practicing Senior Enterprise Data Architect and Data Scientist and currently is designing a data science practice and platform for a top 4 audit firm client. He is also the CEO of ByteSumo, a data science Consultancy

He is a specialist in data processing languages, data platform design, emerging data technologies, exotic data structures, data science methods, technical architecture, and data security systems.

Data Science, Data Architecture, Big Data Engineering.

He founded ByteSumo to build a data science led consultancy that has the experts and tools needed to transform and disrupt traditional enterprises.

(curr. client role)	*2014 - 2015*	*Interim Head of Data Science*
ByteSumo	2013 - present	CEO
Capgemini	2010 - 2013	Senior Enterprise Architect, BIM
Thomson Reuters	2006 - 2010	Architect, Senior Technologist
Aprimo (now Teradata)	2005 - 2006	Senior Consultant
Acxiom Corporation	2000 - 2005	Business Solutions Architect
dunnhumby	1999 - 2000	Database Consultant
Elf Gas & Power UK	1995 - 1999	Operational Dev. Executive
Gov't of Ontario	1994 - 1994	Jnr. Planner, GIS systems.

Bachelor of Arts, Geography. University of Toronto. 1994

Attribution

Salomé Areias: *http://salomeareias.com/what-is-a-trend/*

Eurostat: *http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:Trend_cycle*

MultiScale Trend Analysis:

I. Zaliapin, A. Gabrielov, V. Keilis-Borok. ***Multiscale trend analysis for time series.*** *Fractals, v.12, p.275-292, 2004.*
http://www.math.purdue.edu/~agabriel/mta.pdf

Eamonn Keogh:

Time Series Representations - a slide found in the tutorials found here:
http://www.cs.ucr.edu/~eamonn/tutorials.html

byte sumo