

Grid-enabling Social Scientists: the FINGRID infrastructure (e-Infrastructure)

Lee Gillam

Department of Computing, University of Surrey



Financial Decision Making

Challenge:

analysis of streaming financial (time serial) data **and** financial and political news

At the interface of quantitative and qualitative?

FINGRID Project

- ◆ aimed at information management/ processing challenge in social sciences: analysis and fusion of distributed quantitative and qualitative data and programs.
- ◆ 12-month eSS PDP (Oct 03-Sept 04) involving econometrics (Essex) and computing academics, particularly in grid computing and artificial intelligence (Surrey)
- ◆ Third project at Surrey dealing with qualitative data (news and reports) and quantitative data (time series) → EU Projects ACE (ESPRIT 1996-99), GIDA (FP5: 2001-03).

Motivation

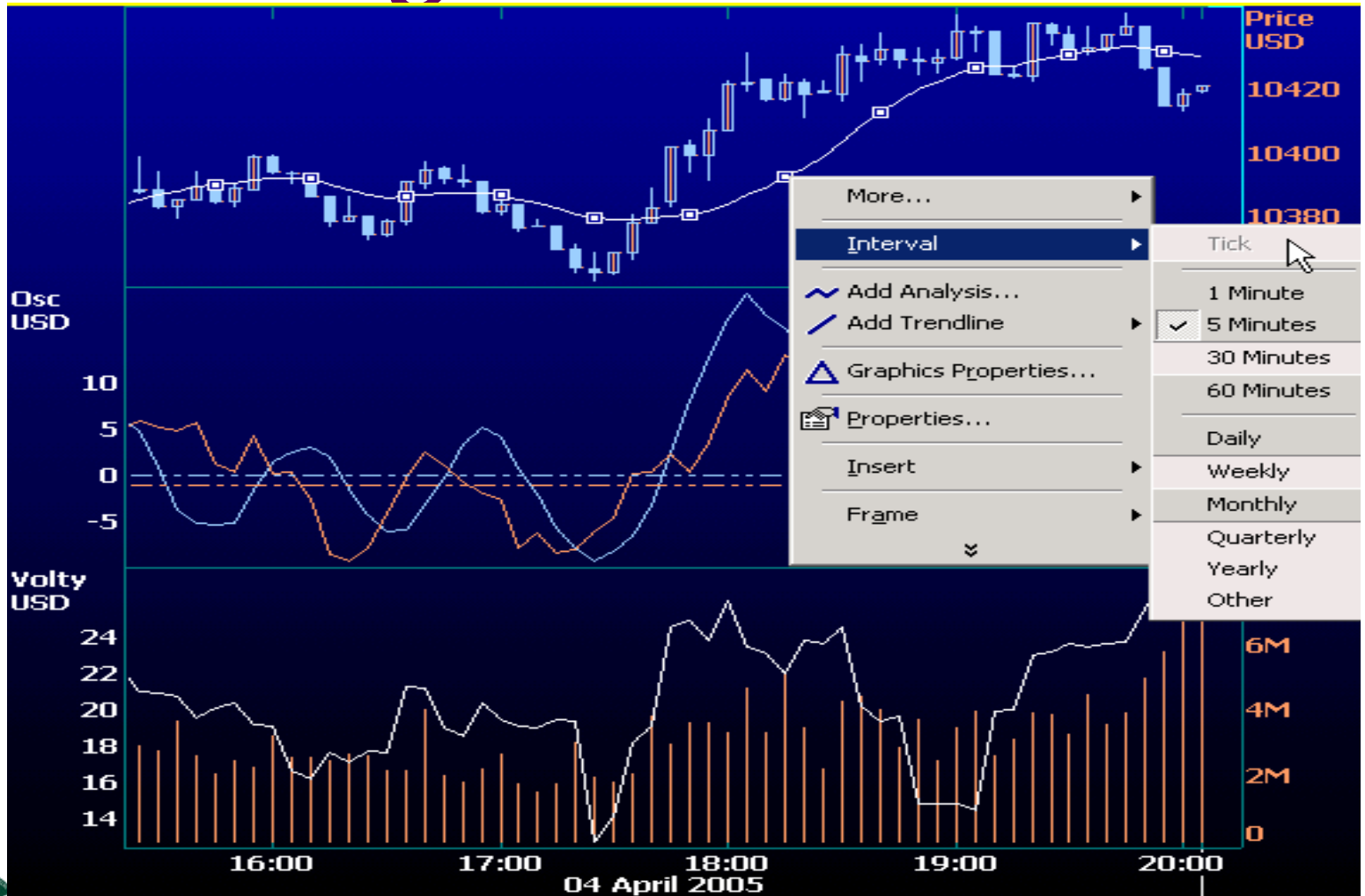
- ❖ **Market sentiment** - quantifying effects of news in the Efficient Market Hypothesis? The effects of news are included in the current price – but effects of *what* news?
 - ❖ Technicalists (chart patterns, stats) and fundamentalists (intrinsic – book - value) locked away from the outside world - no CNN? Challenge of treating multiple data sources
- ❖ Bounded rationality (Simon 1972, Kahneman 2002)
 - ❖ Self-deception of investors rejecting new evidence in favour of prior (incorrect) information (Lakonishk, Lee & Poteshman 2003, Kindlberger 2001) - e.g. “.com” bubble
- ❖ Data “arrival” unpredictable – next item of data may contradict analysis of entire dataset.
- ❖ Data are human patterns of activity: responses to perceived changes, possible patterns and investment hypotheses result in another item of data - Buy/sell - human (re-)action is documented in the dataset

Streaming Data (Reuters)

◆ FOREX (GBP/USD) tick data

09:56:46	↓ 1.8756		1	1.8756	1.8761
09:56:45	↑ 1.8757		1	1.8757	1.876
09:56:44	↓ 1.8753		1	1.8753	1.8763
09:56:43	↓ 1.8756		1	1.8756	1.8761
09:56:43	↓ 1.8756		1	1.8756	1.8759
09:56:42	↓ 1.8757		1	1.8757	1.8759
09:56:41	↑ 1.8758		1	1.8758	1.8761
09:56:40	↓ 1.8756		1	1.8756	1.876
09:56:40	↓ 1.8756		1	1.8756	1.8761
09:56:40	↓ 1.8756		1	1.8756	1.876
09:56:40	↑ 1.8758		1	1.8758	1.876
09:56:39	↑ 1.8756		1	1.8756	1.8762
09:56:39	↓ 1.8755		1	1.8755	1.8759
09:56:38	↑ 1.8757		1	1.8757	1.876
09:56:36	↑ 1.8757		1	1.8757	1.8763
09:56:34	↓ 1.8756		1	1.8756	1.8758
09:56:34	↓ 1.8757		1	1.8757	1.8767
09:56:32	↑ 1.8759		1	1.8759	1.8764
09:56:30	↑ 1.8756		1	1.8756	1.8759
09:56:30	↑ 1.8756		1	1.8756	1.876
09:56:28	↓ 1.8754		1	1.8754	1.8759
09:56:28	↓ 1.8755		1	1.8755	1.8761
09:56:28	↑ 1.8756		1	1.8756	1.8761

Streaming Data (Reuters)



Streaming Data (Reuters)

```
11:12 RTRS-Palestinian gunman wounds Israeli in Gaza-army
11:11 RTRS-IRAQ WRAPUP 2-U.S, Iraqi troops battle dozens of insurgents
11:06 RTRS-*TOP NEWS* Front Page
11:03 RTRS-Former Parmalat execs ask judge to plea bargain
11:02 RTRS-UPDATE 1-Blair meets Queen to set UK election
11:02 RTRS-UPDATE 1-Volkswagen expects profit from new Fox subcompact=2
11:01 RTRS-UPDATE 3-Swiss Life 2004 net soars, proposes dividend=2
11:00 RTRS-SNAPSHOT - Death of Pope John Paul - 1000 GMT
11:00 RTRS-POPE WRAPUP 6-Crowds salute Pope as cardinals consider future
11:00 RTRS-Big rate rises could be risk to fin stability-IMF
11:00 RTRS-IMF-DOUBTS OVER C.BANKS' WILLINGNESS TO HOLD DOLLARS COULD TRIGGER
      FURTHER SIGNIFICANT DLR DECLINES
11:00 RTRS-IMF-STABILITY RISK FROM FURTHER RISES IN COMMODITY, OIL PRICES
      FEEDING THROUGH TO INFLATION
11:00 RTRS-IMF URGES CENTRAL BANKS TO CONTINUE GRADUALLY RAISING RATES TO A
      NEUTRAL LEVEL
11:00 RTRS-IMF-LARGER THAN EXPECTED, ABRUPT INTEREST RATES RISES COULD POSE
      RISK TO GLOBAL FINANCIAL STABILITY
10:58 RTRS-Blair leaves Downing St home to set UK election
10:56 RTRS-UK's Allied Domecq says in bid talks with Pernod
10:54 RTRS-PERNOD RICARD <PERP.PA> SHARES FALL 2.2 PCT AFTER ALLIED DOMEQ
      CONFIRMS TALKS
10:52 RTRS-ALLIED DOMEQ <ALLD.L> SHARES LEAP 14 PCT AFTER SAYS IN OFFER TALKS
      WITH PERNOD
10:50 RTRS-ALLIED DOMEQ SAYS PERNOD RICARD WORKING WITH FORTUNE BRANDS
10:50 RTRS-ALLIED DOMEQ SAYS IN TALKS WITH PERNOD RICARD
10:50 RTRS-ALLIED DOMEQ SAYS IN TALKS ABOUT OFFER
10:48 RTRS-Austria chancellor: govt goes on despite Haider switch
10:46 RTRS-UPDATE 2-Thailand urges vigilance in Bangkok after blasts
```

Streaming Data (Reuters)

```
11:51 RTRS-U.S. stocks seen little changed at open; oil steady
11:50 RTRS-REUTERS CONTACTS: TRAINING, SUPPORT, PRODUCT AND ACC.TEAMS
11:49 RTRS-Jordan government quits, king appoints new PM
11:45 RTRS-JORDAN GOVERNMENT QUILTS, KING APPOINTS NEW PRIME MINISTER -
      GOVERNMENT OFFICIALS
11:42 RTRS-Blair confirms May 5 British election
11:39 RTRS-FACTBOX-Monarchs, presidents to join faithful for Pope funeral
11:39 RTRS-UPDATE 1-Former Parmalat executives seek plea bargaining
11:36 RTRS-BoE moves British May rate decision to May 9
11:34 RTRS-Blair confirms May 5 British election
11:33 RTRS-BANK OF ENGLAND MOVES MAY INTEREST RATE DECISION TO 1100 GMT MONDAY
      MAY 9 BECAUSE OF UK ELECTION
11:30 RTRS-UPDATE 2-Japan approves revised 'nationalist' textbook
11:27 RTRS-BLAIR CONFIRMS BRITISH ELECTION TO BE HELD ON MAY 5
11:26 RTRS-*TOP NEWS* Front Page
```

◆ RSS News feeds from some sources

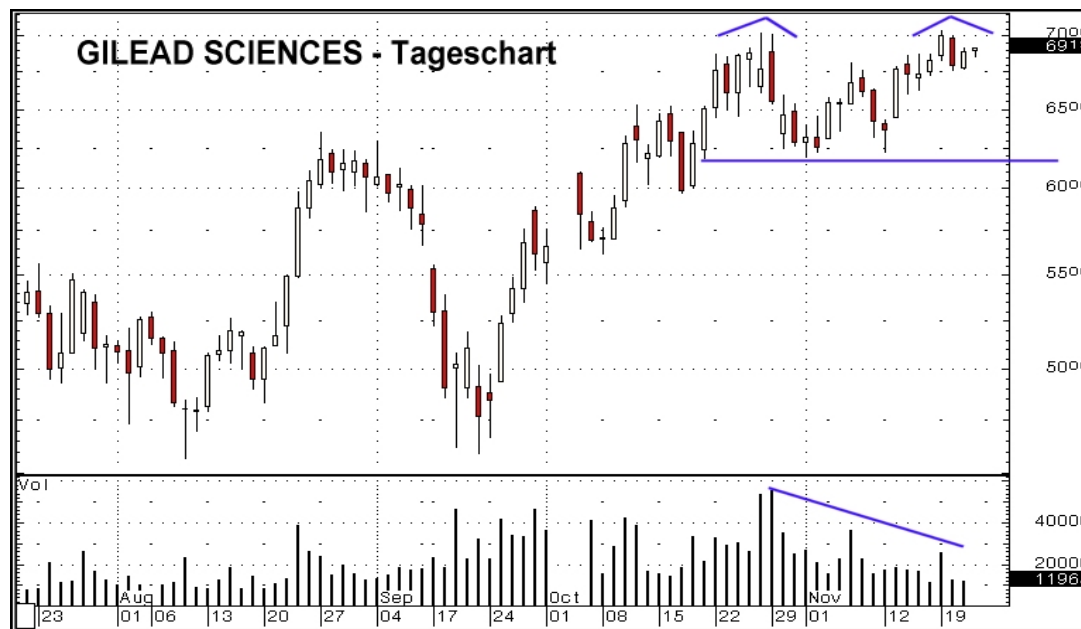
Datasets

Numerical data	<p style="text-align: center;"><u>Time series</u></p> <p style="text-align: center;">price/value movement of financial instruments;</p>	<p>c. 5MB/day, per instrument (XML) - including sources of quote (> 1GB/year/instrument)</p>
Textual data	<p style="text-align: center;"><u>Text streams</u></p> <p style="text-align: center;">news items; financial reports; company brochures; government documents....</p>	<p>c. 40MB/day (> 10GB/year)</p>

- ◆ HFDF data (O&A)
 - e.g. 5 minutes compression GBP/USD 1992 to 2003 inclusive; 1.25M datapoints ($12 \times 24 \times 365 \times 12$) approximates 4MB.
- ◆ Text corpora
 - RCV 1 (over 800000 news stories in 12 of 1996-7); RCV 2 (13 languages)
- ◆ Copyrights/contracts?

Predictability?

- ◆ Encyclopedia of Chart patterns
- ◆ Japanese Candlestick Charting techniques
- ◆ *If price increases and demand decreases.....?*



FINGRID methods/techniques

- ◆ sentiment analysis: automatic terminology extraction; ontology learning; local grammars.
 - Learning the rules for Information Extraction (IE).
 - Patterns derived from a corpus (MB → GB) of texts (arbitrary domain)
- ◆ time series analysis (bootstrapping, wavelet analysis)
- ◆ both types of analysis are computationally intensive – data to computation ratio?
- ◆ Additionally: visualization of large volume time series and texts
- ◆ Look to capabilities of Grid technologies - Globus, Condor, OGSA-DAI, SRB

FINGRID e-Infrastructure

24 computers (dual-proc, hyperthreaded,) with RedHat Linux [v7.3] each provide (at least): **Globus Toolkit [GT3 (3.0.2)]**, Java and FORTRAN software compilers, **Java Commodity Grid kit (CogKit)**, Oracle DB [v8.1.5]

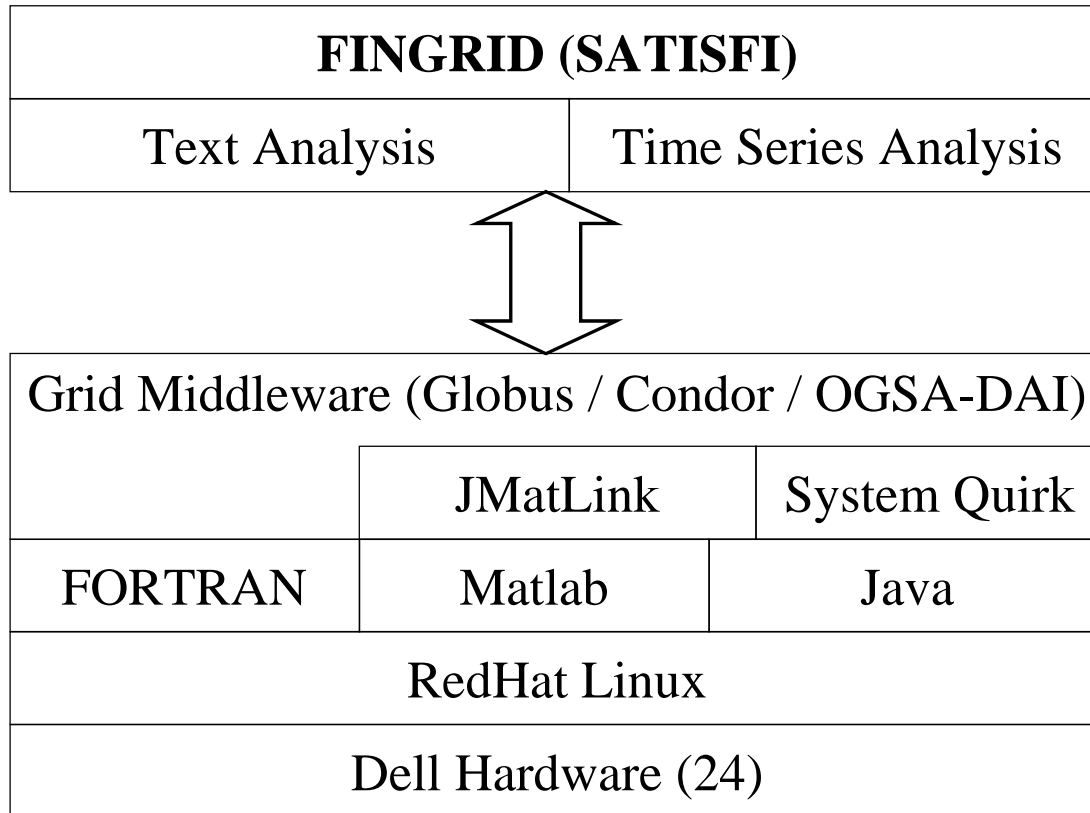
FINGRID uses the Java CogKit to integrate:

- (i) the MATLAB wavelet toolbox via JMatLink;
- (ii) **Reuters data via the Reuters SSL SDK**;
- (iii) bootstrap simulation written in FORTRAN; and
- (iv) **System Quirk components via the Quirk Java SDK**.

Also Condor [v6.6.6] (pool of 76 procs – aiming longer term at Campus Grid); OGSA-DAI [v3.11] Storage Resource Broker [v3.2.1];

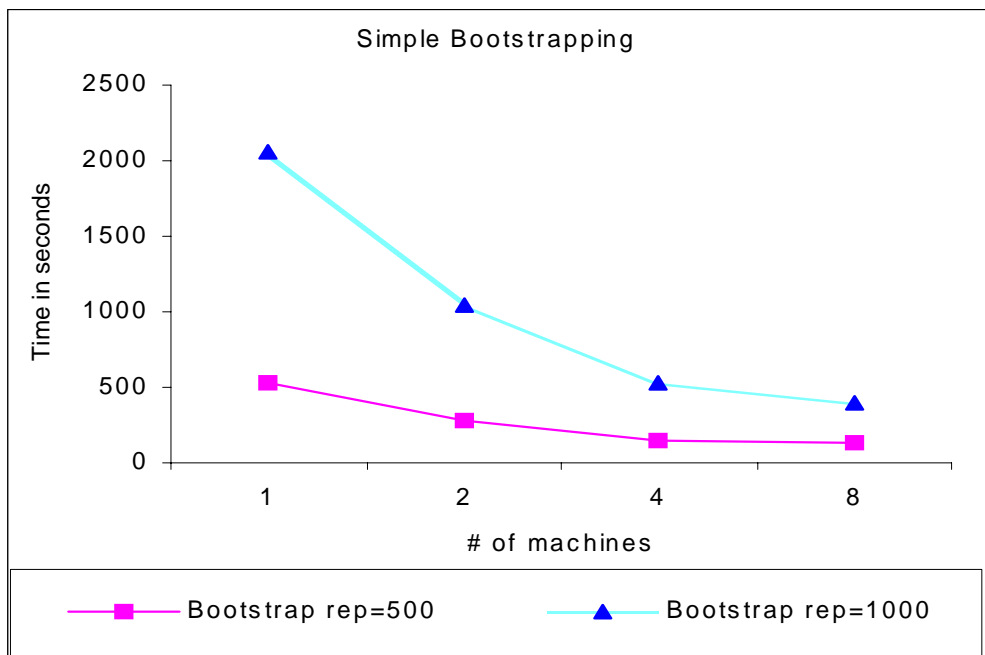
Used for both research and teaching (MSc module in Grid)

FINGRID e-Infrastructure



FINGRID Methods - Bootstrap

- ◆ Bespoke FORTRAN implementation of bootstrapping [Nankervis] algorithm (Globus, Java CoGKit – Grid service)



1000 bootstrap replications:

2 nodes: 1050 seconds (17.5 mins)

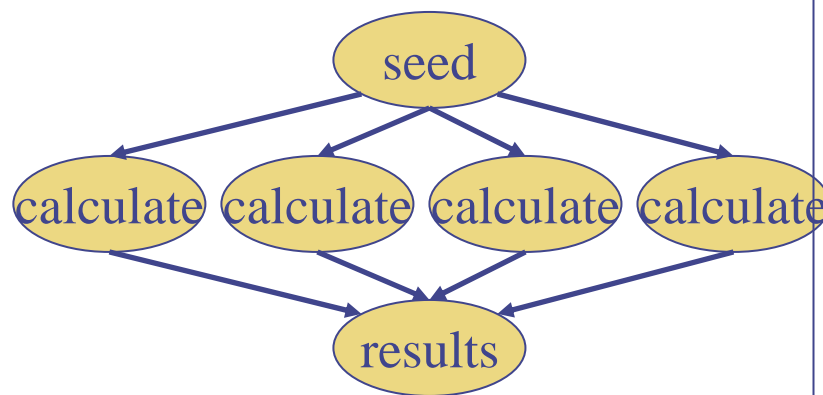
8 nodes: 404 seconds (6.73 mins)

10000+ replications? Linear speedup?

Hypothesis testing – dismiss bad ideas more quickly?

FINGRID Methods - Bootstrap

- ◆ Condor and Condor DAGs (compose metalevel description)
- ◆ Bootstrap is partially parallelizable:
 - **Amdahl's law**: the fraction of code f , which cannot be parallelised, affects speedup factor - replication seeds, results.



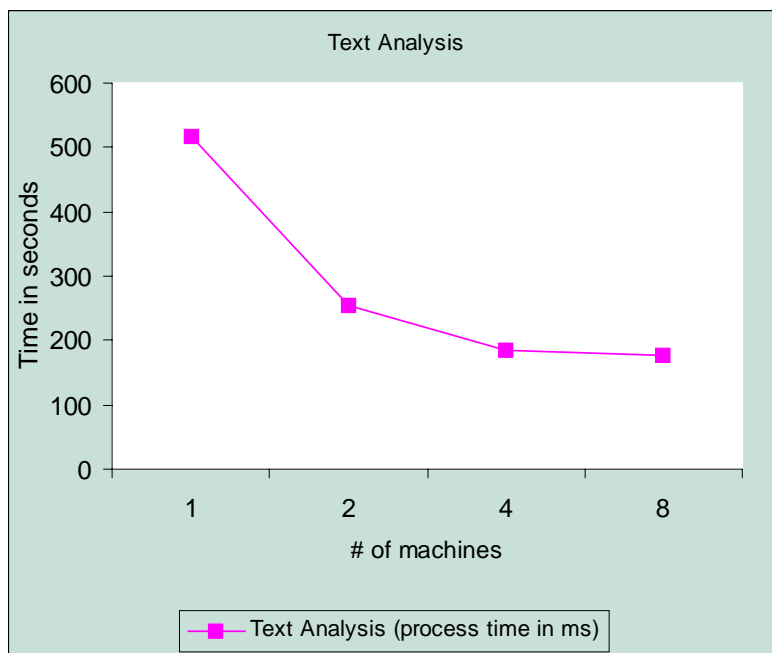
```
Job A seed.cmd
Job B calculate1.cmd
Job C calculate2.cmd
Job D calculate3.cmd
Job E calculate4.cmd
Job F results.cmd

PARENT A CHILD B C D E
PARENT B C D E CHILD F
```

```
executable = calculate.exe
input =
output = calculate.1.out
error = caculate.1.err
transfer_input_files = outs_aa
transfer_files = ALWAYS
log = calculate.1.log
arguments = outs_aa 250
queue
```

FINGRID Methods – Text analysis

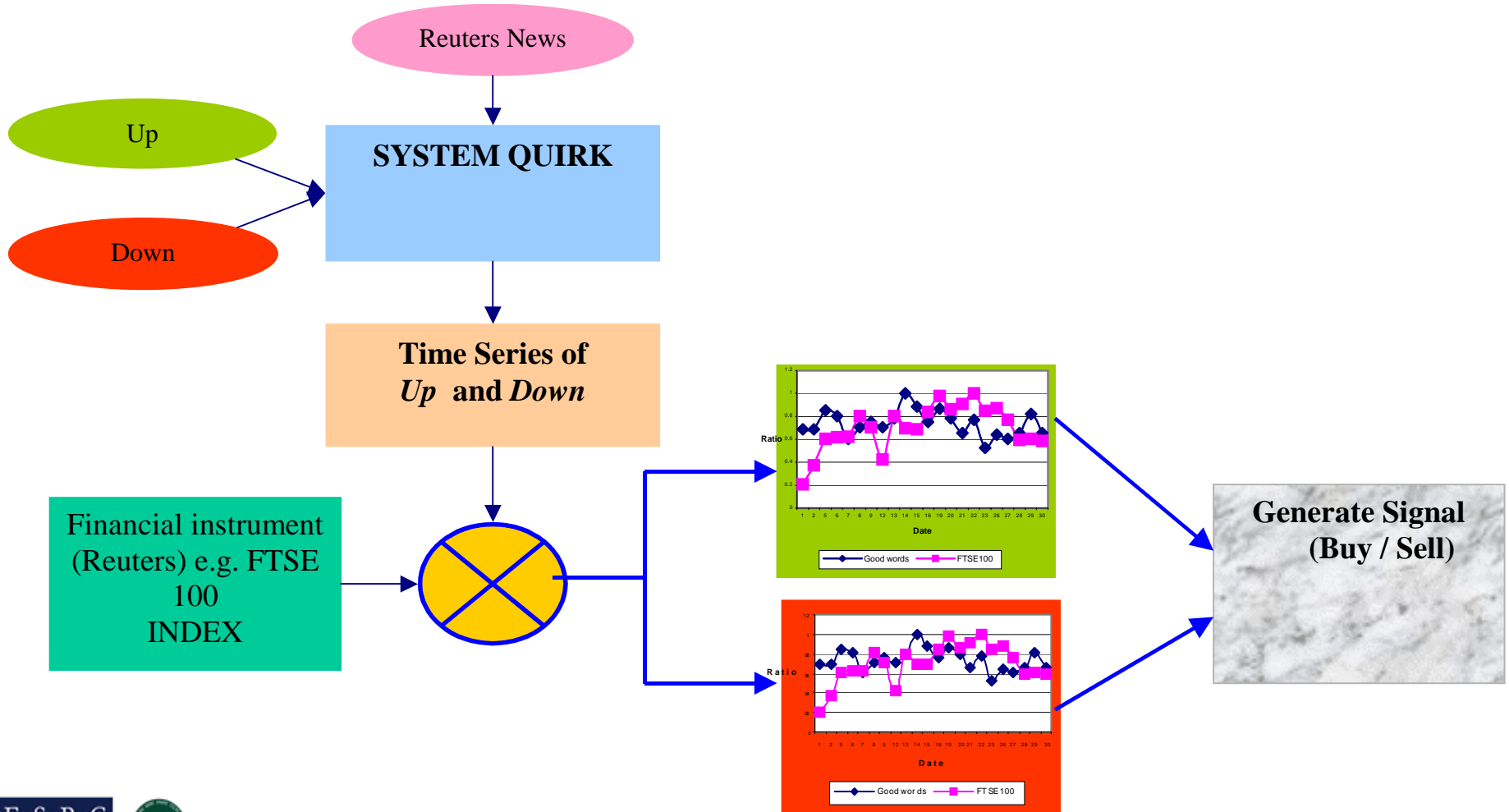
- ❖ Throughput tested with various sizes of corpora – against benchmark (frequency lists – Hughes et al 2004, 2 procs only)



Time required to process one month's news.

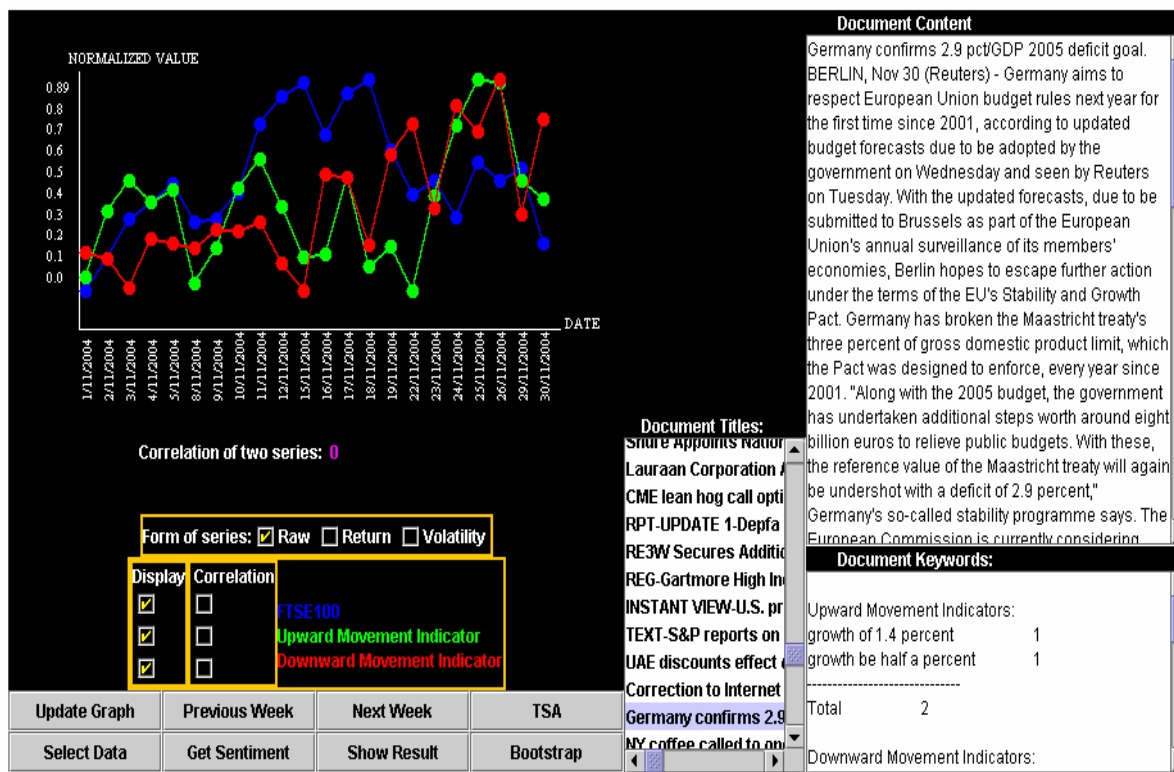
RCV1 takes about 95 minutes on 16 machines. Further experiments to be reported at e-Science AHM 2005

FINGRID - integration



FINGRID – integration and visualisation

- ❖ FINGRID's Sentiment and Time Series: Financial analysis system (SATISFI): for visualising and correlating the sentiment and instrument time series



FINGRID - integration

- ◆ Financial investment rules investigated in EU GIDA project
- ◆ Decision Matrix / confidence of decision

Market (security)	JRC Fractal present?	Divergence exists?	Momentum changed	Volume increased	Reversal exists?	Overbought/oversold	Leverage sufficient	Other, e.g. UniS Sentiment	confidence level
EUR/USD	+	+	+	+	+		+	up	4

FINGRID technologies

- System Quirk
 - Used: text management + terminology&ontology extraction + local grammars +
 - Yet to use: Neural network classifiers (Hebbian networks, Websom); Case-based and fuzzy reasoning; Terminology databases; content workflows; Automatic Text Summarisation; Text alignment.....
- *SRB*: Metadata - Development of Metadata Registries (ISO 11179 conformant):
 - EU eContent project LIRICS (MDR for terminologies, lexicons and syntactic markup) – <http://lirics.loria.fr>
 - ISO 639 - codes for the names of languages – development of part 6 for language variants (21000+ identifiers), in collaboration with the Linguasphere Observatory (Wales) – <http://www.langtags.com>
- *OGSA-DAI* – issues identified in paper. Relies on (about 10) other technologies, with various versions of these also.

Recap

- ◆ sentiment analysis: automatic terminology extraction; ontology learning; local grammars.
 - Learning the rules for Information Extraction (IE).
 - Patterns derived from a corpus (MB → GB) of texts (arbitrary domain)
- ◆ time series analysis (bootstrapping, wavelet analysis)
- ◆ both types of analysis are computationally intensive – data to computation ratio?
- ◆ Additionally: visualization of large volume time series and texts
- ◆ Looked to capabilities of Grid technologies - Globus, Condor, OGSA-DAI, SRB

Outlook

- ◆ Our local *e-Infrastructure* is soon to be increased, courtesy of HEFCE, by 100 processors, terabytes of storage ... and an AGN.
- ◆ Discussions at institution-level w.r.t e-Science certification and Campus Grids
- ◆ *Money + motivated people?*
- ◆ The impending upgrade cycle:
 - “standards”? – version compatibility in Grid middleware? How far back are we? How many versions of the same software should we maintain? Which other resources can we make use of if we locally use version X of technology Y? What if we want to make use of the NGS?

Outlook (= Look out!)

- ◆ RedHat Linux [v7.3] ... RHEL 4
- ◆ Globus Toolkit [GT3 (3.0.2)] GT4
- ◆ Oracle DB [v8.1.5] ... Oracle 10
- ◆ Condor [v6.6.6] ... V6.6.10 (stable), 6.7.8 ("unstable") – being "broken" by certain Windows updates
- ◆ OGSA-DAI [v3.11] ... Release 6?
- ◆ Storage Resource Broker [v3.2.1];
- ◆ OMII

Outlook

- ◆ "Anything that is in the world when you're born is normal and ordinary and is just part of the way the world works.
- ◆ Anything that's invented between when you're fifteen and thirty five is new and exciting and revolutionary and you can probably get a career in it.
- ◆ Anything invented after you're thirty-five is against the natural order of things."
- ◆ "The idea that Bill Gates has appeared like a knight in shining armour to lead all customers out of a mire of technological chaos neatly ignores the fact that it was he who, by peddling second-rate technology, led them into it in the first place."

Douglas Adams

Further information

- ◆ <http://www.computing.surrey.ac.uk/grid/fingrid>
- ◆ <http://www.computing.surrey.ac.uk/courses/csm23>
- ◆ l.gillam@surrey.ac.uk