How to Analyze Data from Unlisted but Rich Firms

From the Perspective of Data and Analysis

Russell Newman

University of Southampton

Victor Chang

Xi'an Jiaotong-Liverpool University

Robert J. Walters

University of Southampton

Gary Wills

University of Southampton

Abstract

This article illustrates the importance of gathering and analyzing preprocessing data for unlisted but rich firms such as venture capital firms. Using datasets from three major sources, the authors demonstrate how to query and analyze data using both Datastream and SQL.

Introduction

In this article, we aim to demonstrate the importance of preprocessing data, present insights on accessing important sources of data from industry-led solutions, and quantitatively identify the degree to which companies involved with the web sector in recent years have exhibited a repeat of the dot-com bubble observed between 1999 and 2001.

Although business intelligence has been demonstrated as a service to track and present the prices and volatility of selected stocks,[1] there are many companies, particularly venture capital (VC) firms, that are not yet on the stock market. As a result, other types of financial technologies are required to measure their business value.

Our focus is on studying current and historic market data. To do so, we used reliable financial datasets for computational analysis. Datasets from multiple sources were required to analyze the model, which will be detailed later. This necessitated the development of a preprocessing method to sanitize, normalize, and combine the datasets so they could be easily imported into analytical software. The data was then loaded into an analytical tool by running an SQL query through open database connectivity (ODBC). We found that preprocessing was an agile approach to ana- lyzing the data in various dimensions using a self-developed Datastream service.

SOURCING VENTURE CAPITAL ACTIVITY DATA

Our conceptual model called for VC activity data on a per-company basis. However, Bloomberg does not hold data on per-company VC deals because many companies receiving VC support are private and therefore are not required to publish this information.[2] Following discussions with specialists at Bloomberg, we found that this data could be scoured or "scraped" from Bloomberg News articles by checking for certain headlines and keywords, and then extracting the VC data.

However, this approach has four major drawbacks that mitigate its practicality and value.

1. Not all VC deals are announced publically, so these would not appear in the Bloom- berg News service.
2. Scraping the Bloomberg News service for VC-related stories requires writing software programs to perform this task.

3. As an information service (not a data service), Bloomberg News could contain duplicated news stories, resulting in duplicated data.
4. Any data scraped from Bloomberg News would require considerable manual sanitization and verification.

Because per-company VC data cannot be obtained through reliable ways, we found three potential higher-level sources to explore:

1. NESTA's (National Endowment for Science, Technology and the Arts) report *Venture Capital Now and After the Dotcom Crash* (UK data);
2. European Venture Capital Association (EVCA) quarterly reports (European data); and
3. PricewaterhouseCoopers (PWC) and National Venture Capital Association's (NVCA)
4. *MoneyTree Report* (US data).

NESTA provides data on VC activity in the UK in their report,[3] which is compiled using data from the Thomson One service. However, this data lacks the volume and depth required. The EVCA publishes quarterly pan-European reports on VC activity, broken down by sector. The PWC *MoneyTree Report* is a free service that uses data sourced from Thomson Reuters' Datastream to compile a quarterly report on VC activity in the US.[4] This report is prepared on a quarterly basis by PWC on behalf of the NVCA. The data in this report is further grouped by US state, market sector, and funding stage.

An overview of the data sources is shown in Table 1, and a discussion of the data follows.

Table 1. Comparison of venture capitalist (VC) data sources by factor

	NESTA	EVCA	PWC MoneyTree
Subject zone	UK	Europe	US
Data resolution	Yearly	Quarterly	Quarterly
Oldest data	Q1 2000	Q1 2005	Q1 1995
Most recent data	Q4 2009	Q3 2013	Q4 2013
Dimensions	Volume and number of companies	Volume and number of companies	Volume and number of deals
Funding stage grouping	Yes	Yes	Yes
Sector grouping		From Q1 2013	Yes
Geographic area group- ing			Yes
Format	PDF	PDF	Excel
Data definitions	Basic	Basic	Detailed

Data Resolution

Yearly resolution would be adequate, but higher resolutions provide the ability to improve detail. PWC and EVCA offer the highest-resolution data with quarterly reports.

Historic Data

The economic bubbles inherently set the timeframe for which data should be acquired: ideally starting from a period preceding the dot-com bubble (1999 to 2001) and ending as close to the current day as possible. Only

PWC offers data from before, during, and after the dot-com bubble.

Dimensions

All three sources provide VC data quantified by volume of capital invested. NESTA and EVCA provide a further dimension of the number of companies that received investment. Instead of this, PWC provides the number of VC deals. This is an appropriate measure, as it will account for companies that received multiple investments within the time period, and the other sources will not.

Funding Stage Grouping

All sources group their data by funding stage. The VC investment volumes are presented accord- ing to the stage in the VC pipeline (seed, expansion, and so on).

Sector Grouping

Only PWC has consistently grouped their data by market sector since Q1 1995. This enabled us to ignore VC activity outside of the software/Internet sector.

By grouping data by sector, the PWC data clearly exposes funding trends and weightings toward certain sectors. While PWCs data is US-centric, a cursory analysis of per-sector funding reveals strong weightings toward certain sectors. Thus, it would be inaccurate to rely on a dataset that does not break down its data into sectors.

Geographic Area Grouping

Both the NESTA and EVCA data applies to the subject zone as a whole. The PWC data is grouped by US state. However, the relationship of geographic area to VC activity, at a state-level of detail, is not the subject of this work.

Format

Only the PWC data is available in an immediately machine-readable format. Both the NESTA and EVCA would require manual input, remedial work, and verification to convert the data into a machine-readable format.

Data Definitions

The PWC MoneyTree website clearly defines the groupings and categories used to generate and present their data. Their methodology to compile the quarterly report is explained, including nominating types of data excluded for research. PWC provides sufficient details to accurately reproduce their own datasets from base data.

NESTA and EVCA do not provide such detailed information on the sourcing and processing of their data, so it is harder to be sure of the origin and reliability of these datasets.

Selection and Justification

Considering these factors, we decided to use PWC MoneyTree as the source of VC activity data as it is the richest data available and the most suitably defined for this research.

The grouping of VC data by sector is a key differentiator for the datasets. This offers the poten- tial to focus on the industry sectors relevant to this work, eliminating noise from others.

The US data represents a divergence from the assumption that this work would be UK-centric. The result is a more coherent analysis, where data subjects are based in the same geographic area and subject to the same economic conditions.

SOURCING COMPANY DATA

Datastream is a database populated by Thomson Reuters to combine quantitative data from annual and quarterly company reports and other industry sources. It is a single entity where data from otherwise-disparate sources can be queried and compared.[5]

Datastream structures financial data around “levels” representing various entity types in what is called the

"Worldscope." Datastream also features lists of companies and securities, curated around various themes. This allows users to quickly retrieve data for a group of related compa- nies or securities. We use the US Software Companies list in this article—it enumerates 105 companies by the London Stock Exchange.

Datastream Querying Technique

To access Datastream, a graphical mnemonic-based querying tool is provided, manifested as an Excel plugin. Results are delivered directly into a spreadsheet.

Datastream supports various query types. Static queries retrieve data that might not be compared temporally, such as a list of directors of specified companies. Our research used time-series queries to retrieve quantitative data regarding periods of time.[5]

Table 2 shows an example Datastream time-series query. The series is one or more subject ent ties that the query should return data for (these are mnemonic-based representations of companies, securities, and other Worldscope levels). Start and end dates are the time period for which the query should return results, specified as explicit dates or relative dates ("start of year").

Datatypes are dimensions of the series that the query should return, selected from a searchable list (these are mnemonic-based representations of 16,432 financial indicators). Frequency is the time frequency upon which results should be returned (yearly, quarterly, monthly, or weekly).

Table 2. Example Datastream time-series query.

Parameter	Input
Series	@GOOG, @YHOO
Datatypes	WC01201, WC08006
Start and end dates	From: 31/12/2004 To: 14/01/2005
Frequency	Weekly

Queries are carefully constructed to ensure that relevant datatypes are selected. Effectively, the queried attributes can actually be possessed by the entities requested.

Frequency should be specified with attention to the reporting periods of requested datatypes. If an annual datatype is requested (such as WC01001) and a quarterly frequency is specified, the single annual value will be duplicated for each quarter in the results.

Results are returned as a crude crosstab within an Excel spreadsheet with time-series axes, as the query name would suggest. By default, rows represent time periods. Columns represent first the requested series and below those, the datatypes. This effectively concatenates all the requested datatypes into one table.

The following summary shows an example Datastream query to find the research and development spend (mnemonic WC01201) and trading volume (mnemonic WC08006) of Google and Yahoo for the three weeks between 31 December 2004 and 14 January 2005.

Table 3 shows the output from this query and exemplifies poor practice—some of the retrieved data is duplicated because annual datatypes were requested against a weekly frequency. The query should be reformulated using either an annual frequency or different datatypes.

Table 3. Datastream crosstab output.

Name	GOOGLE INC. - RESEARCH & DEVELOPMENT	GOOGLE INC. - TRADING VOLUME	YAHOO! INC - RESEARCH & DEVELOPMENT	YAHOO! INC - TRADING VOLUME
Code	@GOOG(WC01201)	@GOOG(WC08006)	@YHOO(WC01201)	@YHOO(WC08006)
31/12/2004	214289	343760368	368760	153826027
07/01/2005	483978	831408175	547137	178197917
14/01/2005	483978	831408175	547137	178197917

Querying in this way quickly revealed two shortcomings:

- Datastream limits the size of returned result sets. If a query contains too many dimensions, the returned data might be truncated beyond the limit. The truncation manifests as a query error code in the affected cells.
- Datastream does not return data in a true crosstab, as evidenced by company name du- plication in the column headings of Table 3. This prevents easy access and analysis of the data in Excel once Datastream has returned its results, and necessitates the creation of a workflow to convert Datastream output into a truly machine-readable format.

Querying Datastream

Table 4 shows the datatypes that were extracted from Datastream for the purposes of this re- search. Each datatype was fetched in an individual query.

Table 4. Datatypes extracted from Datastream.

Datatype mnemonic	Name	Reporting period	Worldscope level
NOSH	Number of shares	Hourly	
WC01001	Net sales or revenue	Annual	Company
UP	Unadjusted price	Hourly	
MV	Market value (capital)	Hourly	
P	Price (adjusted - default)	Hourly	
WC01201	Research and development	Annual	Company
W08006	Trading volume	Annual	Security

Tables 5 to 7 show samples of the output for the trading volume (mnemonic W08006) query. The four companies shown were selected to exemplify the various types of output that occur. Company mnemonics are shown in parentheses in the header row.

Table 5. Sample output from Datastream query.

	@AKAM(W08006)	U:DDD(W08006)	@ADBE(W08006)	@ALTR(W08006)

1990			4895877	363212
1991		13455	6620124	1054756
1992		12331	7975639	1856235

Table 6. Flattened/linear Datastream output.

Row	Column	Value
1990	@ADBE	4895877
1990	@AKAM	
1990	@ALTR	363212
1990	U:DDD	
1991	@ADBE	6620124
1991	@AKAM	
1991	@ALTR	1054756
1991	U:DDD	13455
1992 1992	@ADBE @AKAM	7975639
1992	@ALTR	1856235
1992	U:DDD	12331

Table 7. Flattened/combined Datastream output.

Year	Company	MV	NOSH	UP	P
1990	@AAPL	4430.25	125681	35.250	8.8125
1990	@ADBE	417.23	20604	20.250	1.2496
1990	@ADI	458.25	47000	9.750	1.6252

This output shows data that reflects three different company circumstances.

1. @AKAM: This company was not publically traded during the requested period, and thus has no data.
2. U:DDD: This company became publically traded during the requested period, so data appears from 1991.
3. @ADBE and @ALTR: These companies were publically traded throughout the entire requested period, so a complete dataset is returned.

Additionally, a company might cease trading publically during the requested period, causing the data to be unavailable following the date the company was taken private.

STRUCTURING AND CONCERTING DATA

The data-sourcing exercise generated nine distinct datasets, which could be combined into one dataset to enable analysis. The work described was carried out in Excel.

Seven of nine sets were generated by similar Datastream queries, where only the datatype was changed. This enabled the sets to be combined by identifying matching company mnemonics and years in each set. By combining each dataset individually into a destination dataset, a combined dataset of all datatypes was

produced.

Datastream's seven crosstab outputs were converted into pivot tables. The following workflow in Excel was found to be most effective and was applied to all seven datasets:

1. Isolate the company mnemonic. Using Excel functions, extract the company mne- monic from cells in the heading row: `=LEFT(A2, FIND("(", A2)-1)`
2. Produce a multiple consolidation range pivot table of the data. This reproduces exactly the same table structure, but in a "live" pivot table.

   The last dataset to be processed (number 7) was additionally flattened into a linear table by fol- lowing Step 3:
3. Double-click the pivot table's Grand Total cell. This creates a new sheet, containing a flattened, linear version of the data in the pivot table. Table 6 shows a sample output from this step, using data in Table 5 as an input.

This exercise produced a flat listing of each company-year (see Row and Column in Table 6), plus the value of datatype W08006 from the last dataset. The pivot table data of datasets 1 to 6 was then imported into this flat listing by appending columns containing GETPIVOTDATA functions similar to the following.

```
=GETPIVOTDATA("Price," 'Share Price PIV'!$A$3, "date," $A4, "company,"
$C4)
```

This resulted in one flat table, containing rows that represent company-years, and one column per dataset. A truncated sample of four datasets is shown in Table 7.

The resultant two-dimensional table contains all the output from Datastream and is suitable for storage and analysis using an SQL database.

VC and online population data was provided in a more usable format. The three datasets all fea- ture annual data, so a column representing the year was extracted from the remaining datasets.

The online population dataset contained data on a per-region, per-year basis. Some regions were actual countries, while others represented groups of countries (such as European Union, high in- come, and world). A list of unique regions was extracted from the dataset and coded according to whether each row represented an individual country or a region.

The VC dataset contained data on a per-phase, per-quarter basis, where "phase" represents phases of the VC process (for example, startup/seed, early stage, later stage, and expansion).

A list of the four phases was extracted from the dataset. The data was provided on a per-quarter basis, so a year column was added and populated from existing data.

A conceptual diagram is shown in Figure 1, describing how the common year fields are used to join the three datasets. Only a sample of columns and attributes are shown in the diagram.

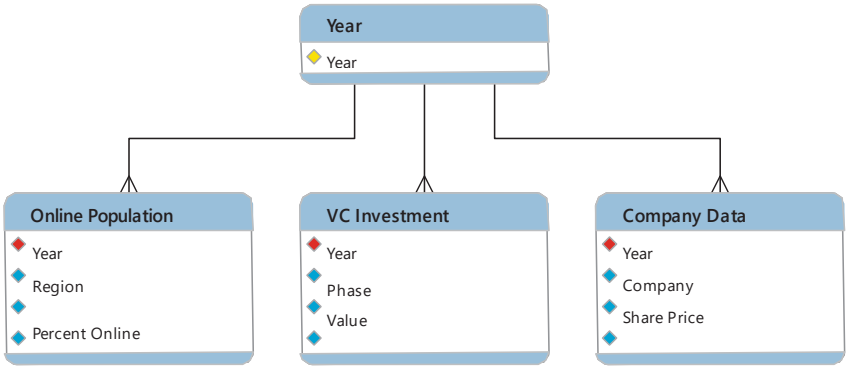

Figure 1. Conceptual diagram showing dataset joins on year columns.

The design in Figure 2 is intended to complement the flat Excel tables by matching the column types and names. This can be seen as the "annualReport" table in the schema. Hence, the Excel tables can be imported directly into the database.

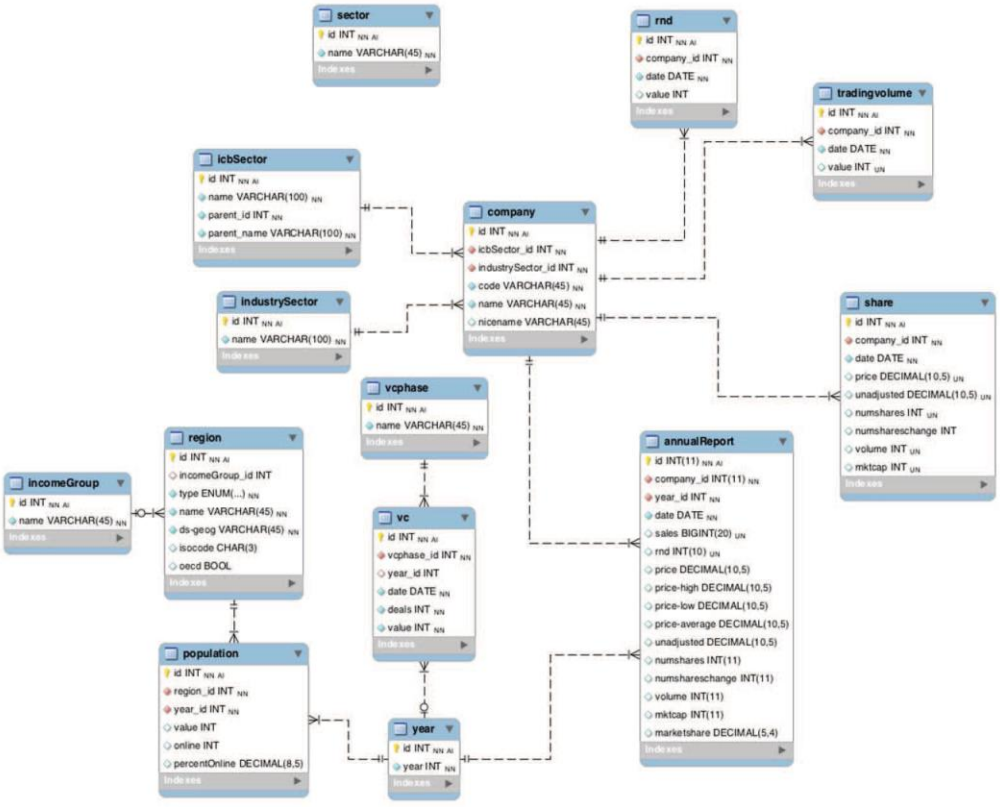

Figure 2. Database schema.

SANITIZING DATA AND CHECKING FOR NORMALITY OF DISTRIBUTION

Before performing any statistical analysis, it is prudent to check that the input data is correct and consistent. The process of importing data into the database was verified by taking samples of data in the database and comparing against the raw output from Datastream, and by comparing columns to one another to ensure that data had been imported into the correct locations and was not erroneously duplicated in other columns.

These verification tasks revealed a pair of datasets with several matching values. This was traced to a regrettable human error while querying Datastream, resulting in one page of results in a dataset for the wrong data series. The erroneous data was identified and replaced with the correct data from Datastream. Both the original and partially cloned datasets were then closely verified.

As a further point of verification, the data should be checked for normality of distribution. This can be accomplished by producing a histogram of the data and comparing it to a normal distribution curve.

Figure 3 shows how raw data from two of the Datastream datasets appear when drawn as a histogram. The distributions are heavily skewed to the lower end of the scale, with long tails and not normally distributed. These results are representative of all the Datastream datasets. Any statistical analysis performed upon data in such a state could suffer and be deemed unreliable.

Figure 4Figure 4 shows histograms for the same data, after the natural logarithm was calculated for each datum.

Following this transformation, the histograms show the datasets as being nor- mally distributed. Hence, results shown in the histograms are representative of all the Datastream datasets.

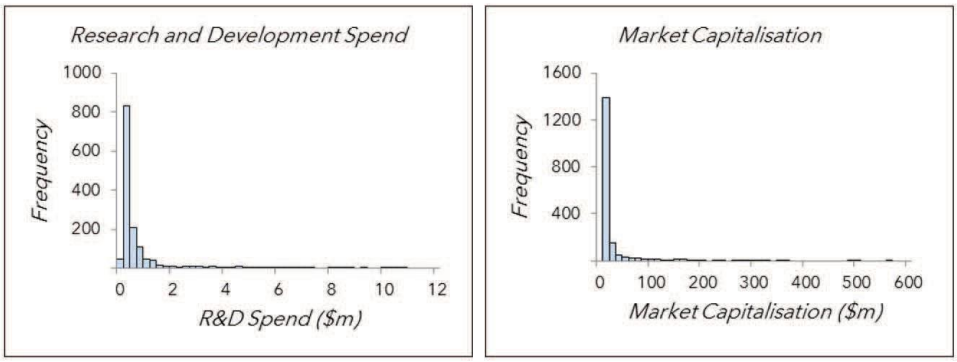

Figure 3. Histograms of raw values for two sample datasets.

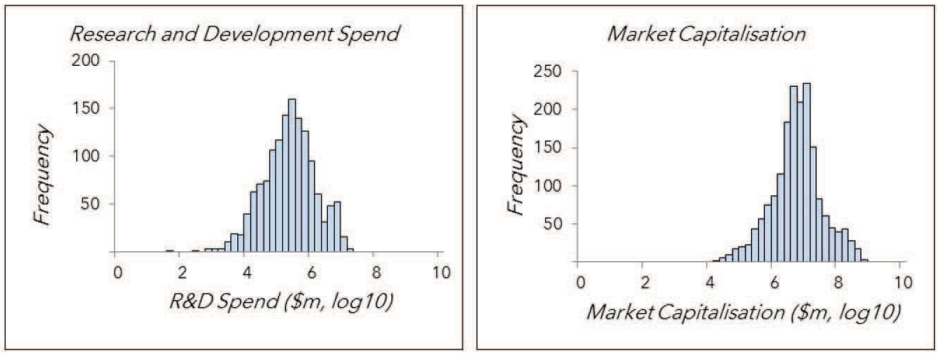

Figure 4. Histograms of natural logarithm values for two sample datasets.

This transformation was applied to all the Datastream datasets, and produced similar output in all seven cases. It was also applied to the VC value columns, which store the financial value of VC activities on a per-phase, per-year basis. In all cases, the original data was left untouched and logarithms were calculated within the views.

ANALYTICAL TECHNIQUES AND APPLICATION TO THE RESEARCH QUESTION

We sourced company and sector data from 1995 through the dot-com bubble until 2015. The data was concerted so that it can be viewed and analyzed as one coherent set. To that end, years in the dataset were coded according to thematic periods. This allowed the data to be analyzed based on time periods, including pre- and post-bubble periods.

We performed factor analysis on the concerted datasets to identify columns that contribute to one another. This output a number of derived factors, showing which columns contribute to each fac- tor (negatively or positively) and by how much. This is a form of dimension reduction, which can be used for exploring relationships between columns of data, or for simplifying wide datasets into a smaller number of columns.

This technique was used to find whether any individual columns of data from the concerted datasets related to one another, and if so, what sort of relationship was observed.

The factor analysis showed the level to which this metric related to other metrics in the dataset. It showed which metrics contributed positively alongside valuations and which contributed negatively. Relationships between other metrics were observed and were explored as potential contributors to the research.

CONCLUSION

We will strive to develop more sophisticated ways to identify unlisted but rich firms involved in activities such as tax avoidance, money laundering, and setting up virtual firms with no real busi- ness activities. We will use artificial intelligence, big data, and financial intelligence techniques to maximize our research outputs.

REFERENCES

1. V. Chang, "The Business Intelligence as a Service in the Cloud," *Future Generation Computer Systems*, vol. 37, 2014, pp. 512–534.
2. M. Meeker, "Internet Trends 2015—Code Conference," *KPCB*, blog, 2015; www.kpcb.com/blog/2015-internet-trends.
3. T. Lux and M. Marchesi, "Volatility Clustering in Financial Markets: A Microsimulation of Interacting Agents," *Int'l J. Theoretical and Applied Finance*, vol. 3, no. 4, 1998; doi.org/10.1142/S0219024900000826.
4. *MoneyTree Report: Q4 2014/Full-year 2014*, report, PricewaterhouseCoopers and National Venture Capital Association, 2014; www.valuewalk.com/wp- content/uploads/2015/04/pwc-moneytree-q4-full-year-2014-summary.pdf.
5. Y. Bai et al., "Efficient Support for Time Series Queries in Data Stream Management Systems," *Stream Data Management: Advances in Database Systems*, Springer, 2005.
6. O.S. Ince and R.B. Porter, "Individual Equity Return Data from Thomson Datastream: Handle with Care!," *J. Financial Research*, vol. 29, no. 4, 2006, pp. 463–479.

ABOUT THE AUTHORS

Russell Newman is a visiting researcher at the University of Southampton. His research interests include IT and finance. Newman received a PhD in computer science from the University of Southampton. Contact him at rn2@ecs.soton.ac.uk.

Victor Chang is an associate professor (reader) and director of PhD and MRes programs at the International Business School Suzhou at Xi'an Jiaotong-Liverpool University. His re- search interests include big data, cloud computing, Internet of Things, and security. Chang received a PhD in computer science from the University of Southampton. He has won nu- merous awards, including the IEEE Outstanding Service Award in 2015 and the Outstanding Young Scientist Award in 2017. Contact him at ic.victor.chang@gmail.com.

Robert J. Walters is a former lecturer in computer science at the University of Southampton. His research interests include grid computing, cloud computing, software engineering, and web science. Walters received a PhD in computer science from the University of Southampton. Contact him at rjw1@ecs.soton.ac.uk.

Gary Wills is an associate professor in computer science at the University of Southampton, a visiting professor at the University of Cape Town, and a research professor at RLabs. His research interests include secure system engineering and applications for industry, medicine, and education. Wills received a PhD industrial hypermedia systems from the University of Southampton. He is a chartered engineer, a member of the Institute of Engineering Technology, and a Principal Fellow of the Higher Educational Academy. Contact him at gbw@ecs.soton.ac.uk.