JMLR: Workshop and Conference Proceedings 6: 277-288 NIPS 2008 workshop on causality

Causality Challenge: Benchmarking relevant signal components for effective monitoring and process control

Michael McCann MCCANN-M15@EMAIL.ULSTER.AC.UK

Yuhua Li Y.LI@ULSTER.AC.UK

Liam Maguire LP.MAGUIRE@ULSTER.AC.UK

Adrian Johnston A.JOHNSTON@EMAIL.ULSTER.AC.UK

Intelligent Systems Research Centre
University of Ulster
Derry
N.Ireland

Editor: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

A complex modern manufacturing process is normally under consistent surveillance via the monitoring of signals/variables collected from sensors. However, not all of these signals are equally valuable in a specific monitoring system. The measured signals contain a combination of useful information, irrelevant information as well as noise. It is often the case that useful information is buried in the latter two. Engineers typically have a much larger number of signals than are actually required. If we consider each type of signal as a feature, then feature selection may be used to identify the most predictive signals. Once these signals have been identified causal relevance may then be investigated to try and identify the causal features. The Process Engineers may then use these signals to ensure a small scrap rate further downstream in the process, increase the throughput and reduce the per unit production costs. Working in partnership with industry we aim to address this complex problem as part of their process control engineering in the context of wafer fabrication production and enhance current business improvement techniques with the application of causal feature selection as an intelligent systems technique.

Keywords: Causal discovery, feature selection, semi-conductor manufacturing, industry, business improvement techniques

1. Introduction

In high volume manufacturing close control and monitoring of production processes are required to ensure quality control and efficiency (Jeong and Cho 2006). Considering the number of process steps in wafer fabrication, typically over 500, and the amount of data recorded during the entire production process, this produces a vast amount of monitoring data. However not all of this data is equally relevant for process control monitoring. Within this environment industry standard business improvement techniques are the tools that are used to try and solve this complex problem. Currently within industry Six Sigma is one of the main business improvement strategies employed to improve the manufacturing process, although this is a well proven technique throughout industry there are a number of weaknesses inherent within its approach (Johnston 2007). The application of new computational intelligence techniques is now being introduced in manufacturing environments. The introduction of feature selection techniques is proposed as an intelligent systems approach to solving this issue. These techniques are prevalent in high volume data environments, in this application domain they may be deployed to identify the desired Key Process Input Variables (KPIVs) and assess their causal relevance. Once identified process engineers may then use these KPIVs to significantly reduce the time required to reach mature product target line yield figures for new product integration with an overall impact on bottom line production costs. The aims and objectives are to investigate and understand the nature of the complex process control issues faced on a daily basis by the semi conductor industry particularly in high volume manufacturing, with a view to research and develop a causal feature selection methodology that can be combined in a hybrid approach to business improvement in this domain. This solution will address the impact of KPIVs on production line yield figures failure rates and hence improve efficiency specifically in the area of new product development (NPD).

Section 2 provides an introduction to how intelligent systems are being deployed within industry to enhance their current business improvement strategies. Section 3 gives an overview of feature selection and causal relevance. Section 4 describes the SECOM dataset that has been put forward for the challenge along with some baseline results and Section 5 outlines conclusions and future work proposing how feature selection and causal relevance may be applied to process control engineering within the semi conductor industry.

2. Business Improvement and Intelligent Systems

Within an industrial context there has been a growing requirement for the introduction of intelligent system techniques over the past 10 years to assist process engineers with their decision-making (Johnston 2007, Peretto 1999). The advances in hardware automation and control systems have impacted the overall importance of utilizing these new techniques within manufacturing (Harrison and Petty 2002). One of the issues faced by engineers in a modern manufacturing environment is how experiential knowledge is utilized within the decision making process. The use of intelligent systems to aid in this decision-making process helps to overcome this problem, current techniques include Fuzzy Logic (FL), Artificial Neural Networks (ANNs) and Genetic Algorithms (GAs). Cus and Balic (2003) propose the use of GAs for use in metal cutting processes to optimize parameters in machine operation and FL combined with ANNs are proposed for grinding processes by Chen and Kumara (1998) for automation of design. As each of these intelligent techniques have different advantages and disadvantages, see Table 1, hybrid combinations are often used to

address complex systems. An example of a hybrid system is proposed by Guh et al (1999) for use in Statistical Process Control (SPC) combining neural networks and expert systems.

Table 1. Intelligent System Techniques Properties (Johnston 2007)

	Properties				
	Reasoning	Generalisation	Decision-making	Adaption	Rule Visibility
Neural Networks	✗	✓	✗	✓	✗
Fuzzy Logic	✓	✗	✓	✗	✓
Genetic Algorithms	✗	✗	✓	✓	✗
Expert Systems	✓	✗	✓	✗	✓
Case Based Reasoning	✓	✓	✓	✓	✓

Six Sigma is one of the main business improvement strategies employed in the manufacturing process. The determining factor within Six Sigma is its aim to identify causal KPIVs and therefore ensure that process outputs remain in control (Flott 2000, Card 2000, Rao et al. 2000, Schmidt et al. 1998). One of the major issues in applying Six Sigma as an improvement strategy within a high volume production environment, where time to full production for integration products is such a critical milestone, is that due to the nature of Six Sigma projects they tend to be time consuming and project centric (Johnston 2007). Thus although it is an industry standard technique in certain circumstances it is not always a feasible solution. The Six Sigma process flow for project implementation is shown in Figure 1. Once a project has been defined the initial measure phase is typically conducted by a project team consisting of all parties that have the relevant expertise and a stake hold in the overall project definition.

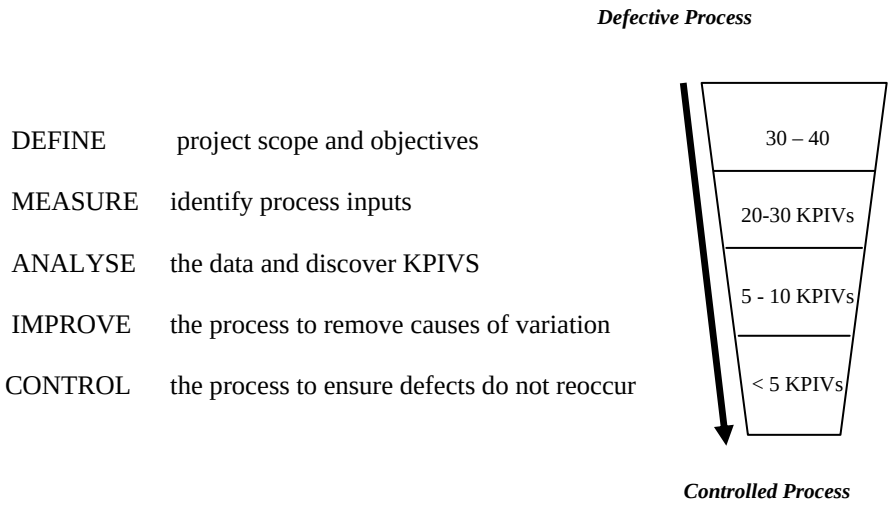

Figure 1. Six Sigma Process Flow (Johnston 2007)

Therefore this phase and hence the overall success of the project is highly dependent on project team experiential knowledge, which unfortunately can be lost, forgotten or invalid for new projects. The advantage of considering intelligent systems such as causal feature selection methods to solve a similar problem is the fact that it does not rely on this experiential knowledge as much to narrow down the processes that are under consideration (Patterson et al., 2005). This allows all the data that is relevant to the overall scope of the

project definition to be considered when trying to discover the desired KPIVs. This also overcomes another issue known as the "anchoring effect" wherein project teams tend to focus on impact processes that have previously displayed concerns within similar project types. Hence this form of human conditioning can lead to previously undiscovered KPIVs being excluded from investigation. During the analysis phase of the project statistical tools are employed to analyze the data that has been identified from the measure phase. Once again this phase is dependent upon the engineer applying the appropriate statistical analysis techniques for the extraction and interpretation of the data such as hypothesis testing on individual process input variables. This entire phase is extremely time and labour intensive and therefore is not always appropriate for time critical projects (Johnston 2007). The improvement phase then requires the consideration of implementing the appropriate actions from these findings to be integrated into the current process flow. This may require optimisation with procedures such as design of experiment (DOE) and potentially failure modes and effects analysis (FMEA). Unfortunately this type of procedure is practically unfeasible in a high volume manufacturing environment because of the amount of data and time required to run trials which has an impact on production and scrap rates. It would be much more desirable to introduce an intelligent systems approach that was able to identify causal KPIVs and apply this methodology in tandem with overall business improvement strategies.

3. Feature Selection

In recent years the nature of feature selection has changed in terms of the complexity of the application. For example in 1997 the applications explored in this field seldom contained more than 40 features (Chiang and Pell 2004, Kohavi and John 1997), whereas in recent years this has changed as feature selection methods are required for domains with in excess of tens of thousands of features such as in gene selection (Guyon et al., 2002), text categorisation (Liu et al., 2005) and other various engineering applications (Guyon and Elisseeff 2003). The selection of relevant features, and the elimination of irrelevant ones, is one of the central problems in machine learning (Blum and Langley 1997). There have been significant advances in feature selection development in recent years and there are a significant number of methods that can be utilised to try and achieve the optimum results. In pattern recognition, the goal of feature selection is to find a feature subset that has the most discriminative information from a given set of a candidate features (Abe and Kudo 2006).

Data representations tend to be very domain specific (Guyon and Elisseeff 2003). Once data is available for machine learning it is often required to manipulate this "raw" data into a format that is conducive to the methodology that is to be applied. This is known as feature construction and may involve simple data manipulation or the application of data transformations. This is often achieved through what is known as pre-processing steps some simple examples of which are (Guyon et al., 2006):
Standardisation e.g. measurements that have different scales
Normalisation e.g. pixel intensity values in image processing
Signal enhancement e.g. smoothing or sharpening
Principal component analysis and multidimensional scaling projecting data into a lower dimensional space whilst retaining the information.

Feature selection then is primarily performed to select the most informative features but other motivations include (Guyon et al., 2006):
General data reduction for storage requirements and processing speed
Feature set reduction to save resources

Performance improvement to gain predictive accuracy
Data understanding to gain knowledge of the process that generated the data or visualisation

3.1 Causal Considerations

Although feature selection on its own is mainly concerned with making accurate predictions with as few variables as possible it does not follow that these variables are necessarily causal within a specific domain. The issue faced by semi conductor manufacturing is not a typical predictive or classification one, it has a large causal element to the problem. For high volume manufacturing the key requirement is to determine which of the variables selected prove causal in terms of affecting failure rates on the factory line yield. So the optimum results would involve identifying these KPIVs giving process engineers insight into the hidden causal relationships within individual manufacturing process steps and overall line yield pass/fail rates. Obviously in real life terms validation of results is not always feasible because of the financial impact of experimental alterations on production processes and the associated unknowns on yield excursions. For this reason it is proposed that any intelligent systems approach to process control be sanitised by inclusion in existing business improvement techniques such as Six Sigma.

4. SECOM: SEmiCOnductor Manufacturing dataset

This challenge aims to investigate a range of feature selection techniques and how appropriate they are to identifying the causal effects faced by process control engineering in semi conductor manufacturing. “In the manufacturing process of semiconductor products one deals with a great number of production steps that involve many different machines. Malfunctions can usually not be ruled out or identified in each processing step” (Pfingsten et al., 2007). Operating conditions can change frequently in a process control environment both intentionally and unintentionally identification of the KPIVs allows rapid recovery, optimisation and control (Chiang and Pell 2004). The goal of this case study is to develop a causal feature selection approach that applies to this domain, helps to solve process control issues and enhance overall business improvement strategies.

Consider in more detail at the nature of the wafer fabrication production process. In the case of integration products it takes time to tweak the processes to achieve target yield figures. Feature selection techniques may be applied to the production process to provide the process control engineers with the necessary intelligence to decrease this integration time and achieve target yield figures earlier in the product life cycle and hence proceed into full production quicker. As highlighted earlier current strategies depend heavily on experiential knowledge which limits the data under investigation and is time consuming. Figure 2 shows “time to yield” baseline trends for integration products i.e. the time required to get new products up to target yield figures hence improving time to market. Good line yields mean:

- Low cost per product
- Predictable schedule adherence and starts planning
- Can run the factory leaner (fewer starts)
- Better throughput at critical tools
- Better quality downstream
- Better product predictability
- No ‘firefighting’ – more resource for project work
- Less waste – less use of consumables

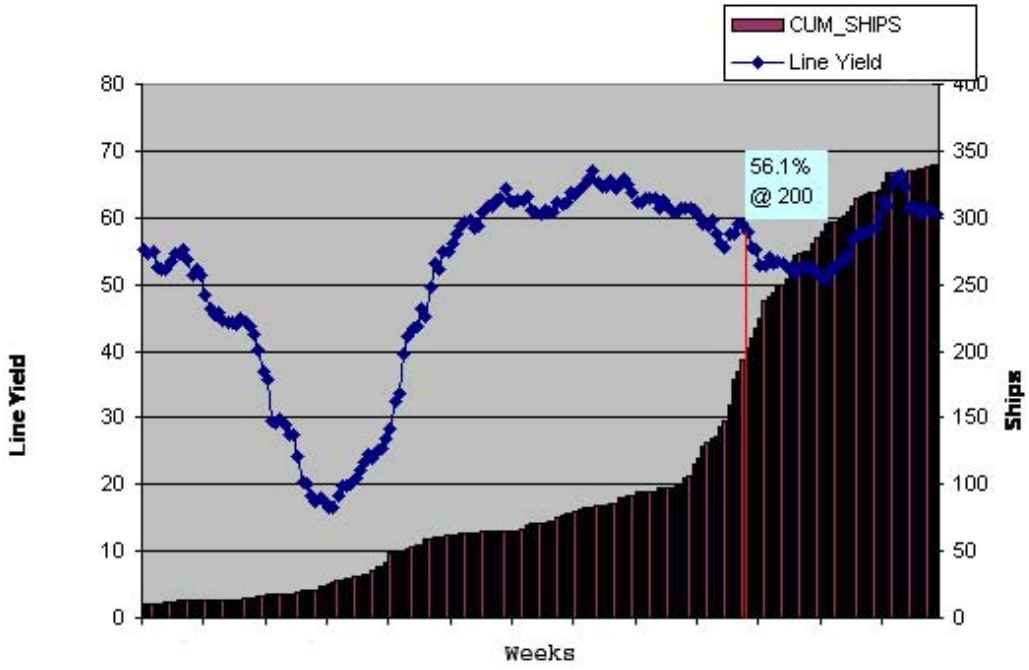

Figure 2: Line Yield Trends

By enabling process engineers to identify KPIVs earlier in the production process it should enable them to affect yield figures more accurately and increase productivity using a more efficient strategy and hence achieving target yield figures for integration products.

4.1 Data Structure

The SECOM dataset presented in this paper, (for a summary see Appendix A), represents a selection of process related data taken from a production line. The dataset is presented with features in columns each representing a recorded measurement and product examples in rows. Within the production cycle there are several major check points for in house line testing to ensure product functionality as demonstrated in Figure 3. The labels file then represents a simple pass/fail classification corresponding to each row in the dataset, where –1 corresponds to a pass and 1 corresponds to a fail. A date-time stamp for each pass/fail is also provided in the labels file corresponding to a selected functionality test.

The data consists of 2 files, the dataset file SECOM consisting of 1567 examples each with 591 features, a 1567 x 591 matrix, and a labels file containing the classifications and date time stamp for each example. As with any real life data situations this data contains null values varying in intensity depending on the individual features corresponding to data-points with no recorded measurement in the original data. This may be taken into consideration when investigating the data either through pre-processing or within the technique applied. Using feature selection techniques it is desired to obtain a sub-set of the most predictive features and then consider the causal relationships within these features and how they impact on the overall pass/fail rates for the product. It is suggested that cross validation be used for generalization performance. Some baseline results are given below.

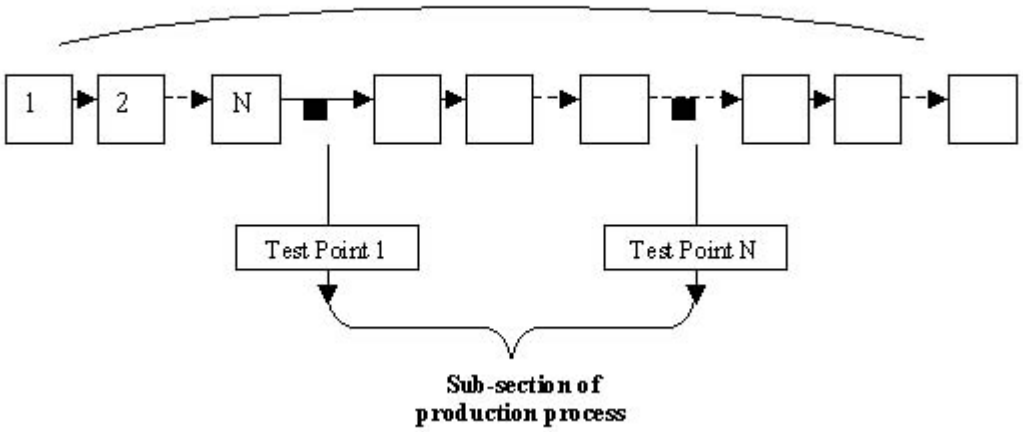

Figure 3. Production Cycle

Baseline Results: Preprocessing objects were applied to the dataset simply to standardize the data and remove the constant features. Then a number of simple statistical feature-ranking techniques were applied with a simple Naïve-Bayes classifier to achieve some initial baseline results. 40 features were selected in each case. 10 fold cross validation was used and the balanced error rate (BER) generated as an initial performance metric to help investigate this dataset. The results are shown in Table 2 below. The desired goals at this stage are to improve upon these error rates for models selecting no more than 40 features and investigate the causal relationships with the target values.

Table 2. SECOM Dataset: 1567 examples 591 features, 104 fails

FSmethod (40 features)	BER %	True + %	True - %
No feature selection	36.9 ±2.4	43.8 ±4.7	82.4 ±1.5
S2N (signal to noise)	34.5 ±2.6	57.8 ±5.3	73.1 ±2.1
Ttest	33.7 ±2.1	59.6 ±4.7	73.0 ±1.8
Relief	40.1 ±2.8	48.3 ±5.9	71.6 ±3.2
Pearson	34.1 ±2.0	57.4 ±4.3	74.4 ±4.9
Ftest	33.5 ±2.2	59.1 ±4.8	73.8 ±1.8
Gram Schmidt	35.6 ±2.4	51.2 ±11.8	77.5 ±2.3

Initial findings and baseline results suggest it may be desirable to increase the size of the dataset significantly to improve performances and allow for separate final tests sets.

5. Conclusion and Future Work

Introducing intelligent system techniques such as causal feature selection within a high volume manufacturing environment would overcome many of the difficulties that have been outlined. Research by Pfingsten et al suggests the use of feature selection to consider the complete assembly line and detect key processes that affecting yield (Pfingsten et al., 2007). Previously undiscovered KPIVs could then potentially be identified earlier in the product integration life cycle where time is of critical consideration. Although intelligent techniques

have seen significant advances in deployment, feature selection has not been seen wide spread use within the semi conductor industry. By investigating how causal feature selection can be deployed within a process control environment, it is proposed that a hybrid approach employing the appropriate feature selection techniques and existing business improvement techniques be designed. This enhanced business improvement strategy may then be deployed to achieve more effective monitoring and process control. This should allow engineers to consider all of the possible KPIVs across the complete production process and overcome some of the disadvantages associated with current methods.

References

ABE, N. and KUDO, M., 2006. Non-parametric classifier-independent feature selection. *Pattern Recognition,* 39(5), pp. 737-46.

BEKKERMAN, R., EL-YANIV, R., TISHBY, N. and WINTER, Y., 2003. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research,* **3**(7-8), pp. 1183-208.

BENJAMINI, Y. and HOCHBERG, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B,* **57**(1), pp. 289-300.

BLUM, A.L. and LANGLEY, P., 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence,* **97**(1-2), pp. 245-71.

CARD, D.N., 2000. Sorting out Six Sigma and the CMM. *IEEE Software,* **17**(3), pp. 1-13

CHEN, Y.T. and KUMARA, S.R.T., 1998. Fuzzy logic and neural networks for design of process parameters: a grinding process application. *International Journal of Production Research,* **36**(2), pp. 395-415.

CHIANG, L.H. and PELL, R.J., 2004. Genetic algorithms combined with discriminant analysis for key variable identification. *Journal of Process Control,* **14**(2), pp. 143-55.

CUS, F. and BALIC, J., 2003. Optimization of cutting process by GA approach. *Robotics and Computer Integrated Manufacturing,* **19**(1-2), pp. 113-121.

DAS, S., 2001. Filters, wrappers and a boosting-based hybrid for feature selection. *Proc.ICML,* .

DREYFUS, G., 2005. Assessment methods. *Feature Extraction, Foundations and Applications, Springer, Berlin,* .

DUCH, W., 2006. Filter Methods. *Feature extraction, foundations and applications,* , pp. 89–118.

FLOTT, L.W., 2000. Six-Sigma Controversy. *Metal Finishing 0026-0576 , **98**,* pp. 43–48.

GILAD-BACHRACH, R., NAVOT, A. and TISHBY, N., 2004. Margin based feature selection - Theory and algorithms, *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004,* Jul 4-8 2004 2004, Association for Computing Machinery, New York, NY 10036-5701, United States pp337-344.

GUH, R.S., TANNOCK, J.D.T. and O'BRIEN, C., 1999. IntelliSPC: a hybrid intelligent tool for on-line economical statistical process control. *Expert Systems with Applications,* **17**(3), pp. 195-212.

GUYON, I., GUNN, S., HUR, A.B. and DROR, G., 2005. Result analysis of the nips 2003 feature selection challenge. *Advances in Neural Information Processing Systems,* **17**, pp. 545–552.

GUYON, I., GUNN, S., NIKRAVESH, M. and ZADEH, L.A., Jul 2006. Feature Extraction: Foundations and Applications (Studies in Fuzziness & Soft Computing) Springer-Verlag Berlin and Heidelberg GmbH & Co. K; Har/Cdr edition (Jul 2006).

GUYON, I. and ELISSEEFF, A., 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research,* **3**(7), pp. 1157-82.

GUYON, I., WESTON, J. and BARNHILL, S., 2002. Gene selection for cancer classification using support vector machines. *Machine Learning,* **46**(1-3), pp. 389-422.

HARRISON, D.K. and PETTY, D.J., 2002. Systems For Planning And Control In Manufacturing. Butterworth-Heinemann Ltd., ISBN 0750649771.

JEONG, B. and CHO, H., 2006. Feature selection techniques and comparative studies for large-scale manufacturing processes. *International Journal of Advanced Manufacturing Technology,* **28**(9), pp. 1006-11.

JOHNSTON, A., 2007. Integrating Business Improvement and Intelligent Systems in high volume manufacturing, *PhD Thesis, University of Ulster* pp. 58-82

KOHAVI, R. and JOHN, G.H., 1997. Wrappers for feature subset selection. *Artificial Intelligence,* **97**(1-2), pp. 273-324.

LAL, T.N., CHAPELLE, O., WESTON, J. and ELISSEEFF, A., 2006. Embedded methods: Feature Extraction, Foundations and Applications. pp. 137–165.

LIU, L., KANG, J., YU, J. and WANG, Z., 2005. A comparative study on unsupervised feature selection methods for text clustering, *Proceedings of the 2005 12th IEEE International Conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE '05),* 30 Oct.-1 Nov. 2005 2005, IEEE pp597-601.

PACELLA, M. and SEMERARO, Q., 2007. Using recurrent neural networks to detect changes in autocorrelated processes for quality monitoring. *Computers & Industrial Engineering,* **52**(4), pp. 502-520.

PATTERSON, A., BONISSONE, P. and PAVESE, M., 2005. Six Sigma Applied Throughout the Lifecycle of an Automated Decision System. *Quality and Reliability Engineering International,* **21**(3), pp. 275-292.

PERETTO, P.F., 1999. Industrial development, technological change, and long-run growth. *Journal of Development Economics,* **59**(2), pp. 389-417.

PFINGSTEN, T., HERRMANN, D.J.L., SCHNITZLER, T., FEUSTEL, A. and SCHOLKOPF, B., 2007. Feature selection for troubleshooting in complex assembly lines. *IEEE Transactions on Automation Science and Engineering,* **4**(3), pp. 465-9.

RADKOWSKI, S. and DYBALA, J., 2007. Geometrical method of selection of features of diagnostic signals. *Mechanical Systems and Signal Processing,* **21**(2), pp. 761-79.

RAO, M., SUN, X. and FENG, J., 2000. Intelligent system architecture for process operation support. *Expert Systems with Applications,* **19**(4), pp. 279-288.

SCHMIDT, D.C., HADDOCK, J., MARCHANDON, S., RUNGER, G.C., WALLACE, W.A. and WRIGHT, R.N., 1998. A methodology for formulating, formalizing, validating, and evaluating a real-time process control advisor. *IIE Transactions*, **30**(3), pp. 235-245.

STOPPIGLIA, H., DREYFUS, G., DUBOIS, R., OUSSAR, Y., GUYON, I. and ELISSEEFF, A., 2003. Ranking a Random Feature for Variable and Feature Selection. *Journal of Machine Learning Research*, **3**(7-8), pp. 1399-1414.

WESTON, J., PEREZ-CRUZ, F., BOUSQUET, O., CHAPELLE, O., ELISSEEFF, A. and SCHOLKOPF, B., 2003. Feature selection and transduction for prediction of molecular bioactivity for drug design. *Bioinformatics*, **19**(6), pp. 764-771.

XING, E.P., JORDAN, M.I. and KARP, R.M., 2001. Feature selection for high-dimensional genomic microarray data. *Proceedings of the Eighteenth International Conference on Machine Learning*, , pp. 601–608.

Appendix A. Pot-luck challenge: FACT SHEET .

Repository URL: http://www.causality.inf.ethz.ch/data/SECOM.zip

Title: SEmi COnductor Manufacturing
Authors: Michael McCann, Yuhua Li, Liam Maguire, Adrian Johnston
Contact name, email and website: Michael McCann, mccann-m15@email.ulster.ac.uk, www.isrc.ulster.ac.uk

Key facts: The data consists of 2 files the dataset file SECOM consisting of 1567 examples each with 591 features a 1567 x 591 matrix and a labels file containing the classifications and date time stamp for each example. The dataset is presented with features in columns each representing a recorded measurement and product examples in rows. The labels file then represents a simple pass/fail classification corresponding to each row in the dataset where –1 corresponds to a pass and 1 corresponds to a fail. A date-time stamp for each pass/fail is also provided in the labels file corresponding to a selected functionality test. The data contains null values varying in intensity depending on the individuals features corresponding to data-points with no recorded measurement in the original metrology data.

Keywords: Causal discovery, feature selection, semi-conductor manufacturing, industry, business improvement techniques