



# MAKING SENSE OF DATA VISION

Author	Professor Bob Williamson
Subject	Making Sense of Data Vision
Document Version #	1.0
Date Finalised	1 August 2008

## **Overview**

The purpose of this paper is to provide an overview of the Making Sense of Data Research Theme Strategic Plan.

## **Making Sense of Data**

Information and Communication Technology has enabled the systematic gathering of large quantities of data. Making sense of this data remains a challenge. From financial transactions to document analysis to problems arising in life sciences and environmental management, there are many complex challenges to make use of the data gathered including detection and classification of patterns of activity, filtering, detection of structure, summarisation, and model building.

## **Vision**

Our Vision is to develop platform technologies that enable us to solve large scale challenging problems in making sense of data more effectively.

## **Capability**

NICTA has world-class capability in Machine Learning; Computer Vision; Natural Language Technologies; Multimodal user interfaces and Cognitive Systems Engineering.

## **Current Activities**

Key current activities within the theme include face recognition using very low quality surveillance video; inference of traffic flows on roads; inference of human movement patterns from body sensor data; mining bioinformatic data for cancer diagnostics and improved crop breeding; development of generic new machine learning techniques for extremely large data sets; computer vision techniques for dynamic scene analysis and cognitive load modelling for the better design of complex control rooms.

# NICTA Research Theme Strategic Plan

## Making Sense of Data

Beware of the man of one method or one instrument, either experimental or theoretical. He tends to become method oriented rather than problem oriented. The method-oriented man is shackled: the problem-oriented man is at least reaching freely toward what is most important. [7]

### Overview

Information and Communication Technology has enabled the systematic gathering of large quantities of data. Making sense of this data remains a challenge. From financial transactions to document analysis to problems arising in life sciences and environmental management, there are many complex challenges to make use of the data gathered including detection and classification of patterns of activity, filtering, detection of structure, summarisation, and model building.

### Vision

Our Vision is to develop platform technologies that enable us to solve large scale challenging problems in making sense of data more effectively.

### Capability

NICTA has world-class capability in Machine Learning; Computer Vision; Natural Language Technologies; Multimodal user interfaces and Cognitive Systems Engineering.

### Current Activities

Key current activities within the theme include face recognition using very low quality surveillance video; inference of traffic flows on roads; inference of human movement patterns from body sensor data; mining bioinformatic data for cancer diagnostics and improved crop breeding; development of generic new machine learning techniques for extremely large data sets; computer vision techniques for dynamic scene analysis and cognitive load modelling for the better design of complex control rooms.

### Longer Term Goals

- **Principles** These key principles will influence our MSOD research:
  - *Composability* – developing methods that can be composed together in order that more complex MSOD systems can be developed effectively and sub-components re-used.
  - *Embeddability* – embedding MSOD technology pervasively.
  - *Layerability* – building higher level representations of knowledge.

- *Reliability* – developing methods for better assessing the quality of conclusions reached using MSOD techniques.
- *Plurality* – we reject the notion of “one true way”. We are problem focussed rather than technique focussed.
- ***Domain Focus*** Rather than making a universal toolkit, we will focus on particular application domains in order to get critical mass and to build critical linkages:
  - Bioinformatics which includes natural language and documents
  - Sport and human performance
  - Video Surveillance, especially face detection and recognition
  - Road traffic monitoring and estimation for control purposes
  - Human Implant technologies MSOD needs
- ***Platforms*** There will be a systematic attempt to share tools through common platforms. In order to do so there will be a focussed effort on platform design. We aim to achieve impact from this in two ways: by providing competitive advantage to our end-user focussed projects, and to provide global impact by open sourcing large parts of the platform.
- ***Development versus Use*** We will consciously manage the tension between the *development* of the core MSOD technology and the provision of a service to aid the *deployment* of MSOD technology.

## Context

As ICT advances, more data is created, and this trend is accelerating [12]. The data does not just live in data warehouses or databases – it is pervasive, embedded, streaming, structured, fluid, and demanding. There is thus an increasing need to make sense of it. This does not mean simple statistical interpretations of data provided directly to a human as a consumer; it includes problems as diverse as inferring how to direct biomedical experimentation to automatically detecting faults in telecommunication networks.

Making Sense of Data (MSOD) is an enabling technology within ICT, which serves as an enabler across many other sectors. It embraces a range of traditional disciplines (machine learning, statistics, signal processing, computer vision, natural language processing, information retrieval, databases, visualization, and others), but NICTA does not want to view MSOD as merely a container for those research disciplines with their existing (and typically narrow) research agendas.

While ultimately people are consumers of technology, in shaping a research agenda for MSOD we do not want to focus solely on problems that have people as the immediate consumer of the data. Much real value is created by complex sequences of transformations and inferences on data; the plan below consciously addresses this “MSOD stack”. In many cases the data is entirely consumed within a system that nevertheless performs useful functions for people: consider the inference of traffic patterns in a city-wide traffic control system. People consume the control actions so generated; not the data itself. This last point illustrates the boundary between making sense of data, and *using* it to *act*, is at best fuzzy.

The purpose of this strategic plan is to set a long term (10+ years) research agenda for MSOD within NICTA that allows (indeed *enables*) the development and use of MSOD technologies to solve real problems, and to set a global MSOD research agenda beyond those of the disciplines that form the (current) basis for it.

Some of the key drivers are

- Embedded Sensing
- Connectedness of data
- Burgeoning of data sources; scale
- Complexity of data (as opposed to scale); structuredness
- Desire for composability (only reliable way to build large systems is by composition)
- Open data [6]
- Connection with the hardware (confer computer vision / computer graphics)

We want to:

- Focus on platforms and technologies that are both broader and deeper than other offerings
- Find new conceptualisations and unifications of the field as a whole to move it towards an Engineering discipline (as opposed to a craft) by mimicking the implementational compositionality that the nascent

software platforms provide by developing conceptual compositionality that allows the building of a hierarchy. [It is not expected that such work will be complete after 10 years; it is a clear direction to head]

- Become problem driven instead of technique driven: some current research in MSOD is heavily technique driven. Real problems do not line up neatly with existing solutions.

The plan has at its core an MSOD platform, that is, collection of tools and codebases that are intended to be mutually composable. As explained below it is not merely a piece of software, although it is intended to consist of multiple pieces of coherent software, but it is developed within a conceptual framework that will allow the development of solutions to some of the problems alluded to above.

The majority of the work in the MSOD theme will be on focused end-use driven projects. The purpose of this theme plan is to articulate some higher level overarching goals to exploit the scale of the MSOD effort and to try to consciously steer the development of the field as a whole. This plan is not meant to be prescriptive of *everything* we will do in MSOD, but will influence most of it.

## Positioning – competitive advantage

NICTA's key competitive advantages in MSOD are **scale**, **depth**, **breadth** and **focus**.

- **Scale**

MSOD is around 30% of NICTA's activities and is thus a \$20M/yr operation. This plan is intended to look out 10 years. There are few other groups in the world that are in a situation to take such a large and long view of these challenges.
- **Depth**

The depth is two-fold: we have some deep research capability in terms of its profundity or quality; and we have research capability within many aspects of MSOD. Most big real problems require significant advance at many levels of abstraction – we are in a position to do that.

  - Software architecture for MSOD
  - Core machine learning algorithms
  - Optimization
  - Theory of MSOD
  - End user interfaces, cognitive load measurement and cognitive systems engineering – this bridges the gap between the machine intelligence aspect of MSOD and the “human in the loop” aspect.
  - Reasoning and representation of knowledge
  - Particular end-user expertise in computer vision and bioinformatics (broadly construed)
- **Breadth**

Looking outside the MSOD theme, there are complementary skills in each of NICTA's other three themes: these are allied disciplines that combine to make MSOD a success – goes with the fact that MSOD is largely an *enabler*.

**Embedded Systems:** Component architectures; composable design; embeddability of MSOD algorithms;

**Networked Systems:** A consumer of MSOD (e.g. the Monitoring and Managing the Internet project) and an enabler/driver – think of distributed inference; the network as a database.

**Managing Complexity:** Constraints, optimization, modeling languages, knowledge representation, reasoning, control

- **Focus**
  - We will focus on a small number of end use areas (noting that it is necessary to develop domain expertise):
    - Bioinformatics
    - Documents
    - Computer Vision
  - We will focus on particular components of MSOD
    - Core ML algorithms
    - Data representations
    - MSOD viewed through the specific principles articulated below, especially composability.

## Vision

Our Vision is to develop platform technologies that enable us to solve large scale challenging problems in making sense of data more effectively.

We aim to set a global research agenda in MSOD as an integrated set of scientific/technological disciplines and execute the agenda by developing a broad, deep, and composable platform that enables solutions to diverse MSOD problems in each of our business areas.

## Objectives

Our objective is to develop a next generation platform for making sense of data that turns MSOD from a craft to more of an engineering discipline. One aspect of this is to better understand and catalog the toolkit of techniques available. In doing so we want to set a new research agenda around composable MSOD. As a consequence of the construction of our technologies we want to develop compelling new solutions to challenging end user problems that will have a global impact. We aim to position ourselves at the top of the value chain in MSOD.

The way we will do this is multi-dimensional. We will

- Work at all levels of the MSOD stack and integrate the relevant parts;
- Exploit capability elsewhere in NICTA
- Aim to commoditize the lower layers via open source software
- Be problem-driven rather than technique-driven
- Aim for declarative rather than procedural conceptualizations - that is separate the problem statement from the solution technique. (An analogy is the modeling language versus solver components of the G12 project)
- Only work on problems where our unique *research* capability and linkage opportunities give us a compelling advantage.

## Principles

There are 7 core principles we will adopt in the development of these objectives. These principles are an attempt to view the problems of MSOD at a higher level of abstraction. The majority of the research will be largely driven by end-use problems. We believe that by simultaneously considering some of these principles, better solutions will be developed.

### P1. Composability

- This is software reuse, algorithm/method reuse, and conceptualisation of problems in a modular way. Building of software components that can be glued together and to software that does other things. It is also reuse at the problem level.
- One wants to be able to shrinkwrap the components – describe in terms of the problems they solve: requires a taxonomy and structure of *problems*.
- Compare the Declarative versus procedural approach: in order to be able to compose, one needs to clearly describe what each component does – suggest the need for improved languages and frameworks for doing this.
- Composability is perhaps the most important design principle used to build any large system<sup>1</sup> and it sits at the centre of notions of peer-to-peer (see the Peer-to-peer manifesto [19] for how far this idea can be taken)
- The component architecture needs to occur at multiple levels of abstraction: individual software components and data components; aggregations of software (via middleware etc) through to a componentised way of viewing algorithms and problems.
- The composability should reach through to control of sensing – choice of sensing.
- *We want composability to inform all design decisions in constructing the software platform and in addition to influence the very way any new MSOD problem is approached.*

### P2. Scalability

- Massive data sets are becoming more prevalent. A major bottleneck for many applications is the computation required. Thus there are challenges in developing effective means for making sense of data that exploit the complex computational hardware that is available and are effective at avoiding computation that is unlikely to be effective.
- The data may not all be in one place: distributed and streaming data
- In order to address (competitively!) it requires one to fully exploit the available hardware (which can vary significantly). It is not sufficient to

---

<sup>1</sup> Confer the principles of UNIX programming from [11]:

1. small is beautiful → **composability**
2. make each program do one thing well → **reliability, composability**
4. choose portability over efficiency → **plurality** (but not scalability)
6. use software leverage to your advantage → **layerability, composability**
8. avoid captive user interfaces → **composability, plurality**

Lesser Tenets:

6. think parallel → **scalability**
7. the sum of the parts is greater than the whole → **composability**
10. think hierarchically → **layerability**

just develop parallelisable algorithms, although that is a start. There are deep interactions with issues such as composability when one considers parallel implementations [18].

**P3. Embeddability**

- In order to have a large impact, MSOD technology needs to be embedded pervasively. This connects to Embedded Systems but is broader. It captures the notion of system in the loop too (control). And more generally is all about making the technology easy to use.
- Faster processing *allows* embeddability, but creates a challenge too – how to explore architectures (FPGA, GPU, multicore) in a manner that does not lock in particular techniques
- Choice of languages and information architecture is crucial
- It is related to the notion of distributed inference or taking the inference to the data rather than the data to the inferrer.

**P4. Layerability**

- By this I mean the ability to build hierarchical systems. It is related to composability but is different. At present much of MSOD starts from scratch each time. Furthermore the notion is that typically you have a database, you make sense of it, and then you stop. But one will want to often make databases of stuff you have already made sense of. There are many challenges to do with how to *represent* the outputs of a MSOD system. How to combine with *reasoning* systems and notions such as ontologies.
- *Not* merely building multi-layer representations [9,10] although they are somehow relevant
- Bridge the symbolic – sub-symbolic gap; take the *output* of some MSOD technology as the *input* for others (e.g. reasoning with probability distributions (or other uncertainty calculi primitives) rather than raw data.
- This suggests a need to consider alternative uncertainty calculi in MSOD (hardly studied to date [16])
- This is related to the issue of working at the *right* level of abstraction

**P5. Reliability**

- In order to achieve larger impact, MSOD technology needs to be made reliable. The challenges here are to develop better understanding of performance, new ways of determining performance etc.
- Furthermore, many real problems are plagued with imperfections that affect reliability (missing data, noise, misleading and mislabelled data etc)
- At present performance assessment is typically considered something done after the construction of a solution; but there are frameworks for estimating how well you are doing as you do the inference. Indeed the whole issue of performance assessment cries out for systematization. This would allow a much richer set of possibilities to be deployed for any particular problem.
- Making performance assessment itself composable and layerable is a particular challenge. With the exception of a pure Bayesian approach, none of the standard methods compose at all.

**P6. Plurality**

- Real problems are not solved by techniques of just one flavour. But much MSOD research is heavily technique driven. A challenge is to turn the field around to become more problem driven and technique eclectic and pluralistic. In order to do that some languages for comparing and relating problems as well as techniques need to be developed.
- Plurality can occur at all different levels of abstraction:
  - Platform: it is *not* necessary to impose a rigid monolithic coding framework to allow different software components to interoperate. Whilst understood generally for a long time (consider for example the .NET framework [17], the principle design features of which include interoperability, common runtime engine, language independence and portability / platform agnosticism<sup>2</sup>), most MSOD solutions (it is probably misleading to describe them as “platforms”) require you to change too much of what you currently do.

**P7. Auditability**

- This is not the same as reliability. It concerns being able to *understand* the inductive inference that has occurred. It is related to reasoning systems. But typically inductive and deductive problems are addressed by distinct communities.
- How should reasoning be structured so that inductive inference can best plug into it? Conversely, how should the inductive step be structured (and evaluated) such that it fits best into a higher order reasoning system.
- These sorts of questions are hardly even asked at present because there is not the common task language to allow it to be posed precisely enough to solve it.

**Specific Objectives**

Distilling the above to specific objectives, we aim to:

**O1. Build a research agenda around the 7 future MSOD principles**

- Influence the way MSOD is taught and researched by reconceiving the separate activities as a related field, focusing on composability and the other principles listed
- Use the principles as an organizing schema for development of concrete research questions that can lift MSOD to a higher level
- Focus on
  - Core machine learning
  - Dynamic Scene Analysis (providing a focus for computer vision)
  - Applied algorithmics for large-scale data processing
  - The bridge between natural language processing, information retrieval and machine learning
  - Natural Language Processing
- Making sense of data versus *using* Data – NICTA has relatively little control engineering. Integrate the need for control into MSOD.

---

<sup>2</sup> Ironically those attributes (or more precisely, the developer of .NET) do not include one of the most important principles for composability: openness, which NICTA embraces as one of its core values.

- Actively promulgate the agenda (as well as executing it ourselves) to other researchers around the world and influence them to embrace it.

## O2. Carry out research at all levels of the stack to execute the agenda

- **Implementational level** – ways of exploiting complex computational architectures. This may include some aspects of programming languages, and also includes the information architecture necessary to have a platform as envisaged
- **Underpinning computational science level**, optimization techniques *particularly focused* on MSOD problems (*not* generic optimization)
- **Algorithmics** – this is what would traditionally be seen to be core MSOD – the development of new algorithms for solving particular problems
- **Reliability** – development of improved methods of performance evaluation and control
- **Connection to reasoning** – understanding the bidirectional interface between the inductive and deductive sciences underpinning MSOD; [don't understand this previous point] alternative problem representations, including effect of uncertainty calculi on MSOD solutions
- **Application level** – deeply engaging in a limited number of end use problems so that 1) we solve the real problem, and not a faulty abstraction of it and 2) to motivate development of new generic problems elsewhere in the stack
- **Problem taxonomies and relationships** – this is a meta-level that overarches everything else and provides a theoretical framework in which the above issues can sit in a composable and layerable way.

## O3. Build coherent software platforms that embrace these principles

- We need to actually build these platforms. Needs to be best in the world in many (not all) dimensions
- Not just a single piece of code. They are a framework that allows interoperability of many components in a manner that respects the principles articulated above
- Commoditise suitable parts; retain the most valuable top of the value chain parts
- Not just “implementing algorithms” – there has to be a special edge: preferable a “multiplicative” one: both an edge in terms of the algorithms being developed (based on internationally calibrated research) *and* an edge in terms of the implementation (for example bleeding edge exploitation of modern and future computational platforms)
- Exploit future computational platforms – Field Programmable hardware, embeddable hardware, multi-core processing
- Ensure very well packaged and document in order to ensure widespread use
- Open source much of the platform to encourage a community of interest
- Make it as futureproof as possible

- O4. Exploit the platform to build globally impactful solutions to challenge real problems within each of our business areas.** [Expect commercial as well as scientific impact]
- Create economic impact by
    - Assisting top tier partners in targeted vertical markets
    - Potentially creating a spinout
  - Value add to other NICTA activities – repeat the MAMI experience
  - Focus on NICTA’s business areas<sup>3</sup>
- O5. As a consequence of O1-O4 build the reputation of NICTA as a lead group in MSOD world-wide**
- Strong and coherent visibility to industry for MSOD problems
  - Destination of choice for high profile researchers wishing to do sabbaticals
  - Strong engagement with the core conferences: bring ICML, CVPR, ICCV etc to Australia
  - Website
    - Containing tools and methods and scientific content; open source software
  - Summer schools
    - Continue the tradition, but focus on the strategic vision
    - Broaden the target market
- O6. Provide support to other themes within NICTA**
- Create a platform that makes it easier for other researchers to use it
  - Partner with researchers from other themes on particular end-use problems
    - Use of platforms as *the* tool for ensuring the effective exploitation of MSOD technology within NICTA. That is, use the platforms as the way to achieve the “statistical consulting service” without removing the research motivators for core MSOD staff. Challenge: repeat the success of MAMI
  - Work out how to manage the tension between future technology development and making that technology available to assist other parts of NICTA.

---

<sup>3</sup> We have plans for some business areas that are more developed. For example, in biomedical and life sciences our goal is to help derive biologically and medically significant information from the data being collected in the biomedical domain. The volume of such data is growing at an unprecedented rate, due to factors including the appearance of new, high-throughput sequencing and analysis technologies; scaling-up of public projects for collecting research data from biological samples; large-scale initiatives linking research and clinical data held at individual institutions; and the ongoing growth in numbers of research publications, which contain concrete results and are increasingly supported by online publication of supporting data.

This data presents many challenges. It can be difficult to extract and use, particularly when it contains free text. An analysis of a single sample may consist of millions of data points, and thousands of samples may be analyzed in the course of a research project. Search and comparison may not be well understood in genomic or proteomic data. Current volumes of data are in some cases magnitudes greater than can be handled by current algorithms, and within only a year or two current volumes will seem small.

We are tackling several aspects of these problems, from text methods for extracting and collating information from free text to inference over large collections of uncertain data. We plan to develop scalable algorithms, embodied in a robust codebase, for easy deployment on a wide range of biomedical research applications.

## Strategies

### S1. Platforms

- a. Build large parts of the platforms open source, especially the “lower layers” in order to commoditize them
  - o Choose a licence that allows wrapping commercialisable code on top;
  - o Be more systematic about ensuring the provenance of the contributed code (warranties)
- b. Define an overall information architecture that allows integration and interworking of multiple components in order to ease the migration of existing codebases by
  - o Carefully defining the interfaces
  - o Being language agnostic as possible
  - o Not single layered
- c. Focus the hardware connections into a small set of targets (initially FPGAs and multicore)
- d. Put explicit effort into the mechanics of making a platform easy to use by explicit provisioning of information architect and software librarian and potentially resourcing one-off migration efforts
- e. Build an open source community by provision of solution code and tutorials and engagement with community efforts (e.g. MLOSS <http://mloss.org/software/> )
- f. Attempt to future-proof the platforms by developing a roadmap of future evolution of the platform
- g. Be pluralistic – not monistic, since if we enforce “one true way” then no-one else will adopt
- h. Build the platforms in such a manner that they allow for open *data* as well as open *source code* (cf. [3,6])

### S2. Research

- o Focus on research that differentiates by
  - Supporting the platform development
  - Looking up and down the MSOD stack
  - Is grounded in the needs of end users, but do not engage in projects that do not have a strong research component
  - Is problem driven rather than technique driven
- o Develop an overarching framework: a language to describe and classify components and to describe their relationships.
- o Address the 7 principles articulated earlier: [this needs further fleshing out]
  - **Composability**
    - Relating different problems to each other (how to solve a complex problem by combining solutions to simpler ones)
  - **Scalability**
    - Developing the optimization / algorithmic core of MSOD solutions that exploit available hardware (not generic optimization, but particular requirements for MSOD solutions)

- Developing generic tools to allow the exploitation of such hardware
- **Embeddability**
  - Develop suitable infrastructure for embedding MSOD
    - NOSA platform
    - Computer vision embedded platform
- **Layerability**
  - Design for layers
  - Discovery of layers
  - Connection with reasoning and uncertainty
  - Design of databases that allow aggregated entities as first class objects
- **Reliability**
  - Error models that stack. Bayesian inference does automatically, but they suffer from deficiencies
  - Study layerable performance measures; develop a framework for considering different performance measures in the problem framework (break down the apparent gap between problems and performance evaluation)
- **Plurality**
  - Developing ways to exploit multiple inferential frameworks
- **Auditability**
  - Building auditing functions in at the start
- Explicitly determine the research needed elsewhere and strategically build the linkages necessary to utilize the results
  - Algorithms and complexity
  - Architecture
  - Data Acquisition – Especially Sensor networks. Databases. Connection with networked systems vision of the network as a database. Challenges of different views of data (more general than traditional views problems in data warehouses – confer M-Context)
  - Database representations – build connections with the VRL databases expertise
  - Representation and Reasoning with aggregated data
  - Natural Language technology

### **S3. Linkages**

- Design linkage for particular purposes (see categories in following section)
- Explicitly record what we want to achieve from each linkage
- Aim for a small number of linkages with world-class partners that are rich in interaction
- Ensure each linkage has an assigned relationship manager

### **S4. End user engagement**

- Driven by business areas and leveraging off existing partnerships

**S5. IP Strategy**

- Open source to commoditize lower levels
- Extract value from top of the food chain
  - Shrink-wrapped software
  - Partnering with verticals
  - Plan how to create a spin out

**S6. Reputation**

- Deliberately target some high profile reputation building activities and resource them

**S7. Talent Management**

- Students: summer schools, and projects that cohere – Resolve the tension between students aligning with NICTA's projects and their need to be able to have their thesis done and examined by conceiving of student projects aligned with the overall vision (rather than just 'in the area').
- Staff: mentoring, recruitment, sabbatical as a recruitment device

**Actions****A1. Meta**

- Add timelines to all actions
- Lock in mechanisms to regularly review the plan
- Add performance indicators

**A2. Reputation**

- Target some high profile visitors to be *theme* visitors
- Agenda setting internationally. Organize Dagstuhl perspectives series on challenges in Making Sense of Data
- Focus the summer schools around the vision
- Focussed effort on MSOD web presence

**A3. Platform**

- Develop a technology roadmap for the platform. It will not be a single monolithic piece of software. Work out when and how to embrace functional languages etc.
- Develop migration plans for existing codebases
  - Integration of existing software base

**A4. Research**

- Develop analysis of opportunities in each of our business areas (existing engagements plus new ones)
- Develop more detailed research plans for the underpinning research; projectisation of the plan

**A5. Mapping of existing activities**

- Build a Matrix of projects and activities including end dates
- Flesh out connections / overlaps with other themes

- Map each researcher with an interest in MSOD onto a refined graphical representation of the plan

#### **A6. Talent**

- Mentoring of researchers within the theme
- Maintain a register of suitable projects for students that underpin the overall plan

#### **A7. Linkages**

- Audit existing research linkages
- Study business area strategic plans for targets

### **Linkages**

All linkages will have explicitly noted the following *attributes*:

- *Purpose* (as per following list, with explicit explanation of what we expect to get out of the linkage)
- *Type* (international, national, local, internal)
- *Relationship manager* (who looks after the relationship)
- *Status* (past, current, developing, planned)
- *Priority* (Core, desirable, other)

Linkages for this plan have a variety of *purposes*, which we have classified as follows:

#### **Underpinning science and technology we rely upon**

- Ex-NOSA group @ VRL – underlying middleware for distributed MSOD problems [internal; to be developed]
- Yahoo! Or some other web /cloud computing provider [a possibility]
- Microsoft Research Cambridge [international; potential; underpinning computer languages for our vision]

#### **Partnering to carry out the core research underpinning the platform**

- MPI for Biological Cybernetics [International; formally in place; staff exchange; student exchange; focused on underlying machine learning problems and problem representation; good leverage for EU 8<sup>th</sup> Framework project?]
- CMU [international; potential; weak links in place at present]
- ANU Computer Sciences Lab [local; strong source of students; focus on problem taxonomy]

#### **Partnering to construct the platform (the software artifacts)**

- Georgia Tech group [international; formal agreement?; sharing of code and development effort]
- SAIL group? [reasoning; knowledge representation; Natural Language interfaces?]

#### **End user engagement**

(as consumers of solutions and generators of domain driven problems)

- Australian Institute of Sport
- Peter MacCallum Cancer Institute

- Diversity Arrays (DART)
- DSTO? (aspects of BANESH?)
- RTA

**Research groups working on similar domain problems**

- Tsinghua University (traffic sensing) [under development]

## References

- [1] M. Garofalakis, K.P. Brown, M.J. Franklin, J.M. Hellerstein, D.Z. Wang, E. Michelakis, L. Tancau, E. Wu, S.R. Jeffery, and R. Aipperspach, "Probabilistic Data Management for Pervasive Computing: The Data Furnace Project," *Data Engineering Bulletin*, **29**(1), 2006. See also Kate Green, "The Future of Computing, According to Intel," *Technology Review*, September 26, 2007.
- [2] H.H. Bock and E. Diday, *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer, 2000.
- [3] J.M. Hellerstein, *Bricolage: Data at Play*, Keynote Talk, ICDM 2007.  
<http://db.cs.berkeley.edu/jmh/talks/icdm07-keynote.pdf>
- [4] Committee on the Fundamentals of Computer Science: Challenges and Opportunities, *Computer Science – Reflections on the Field, Reflections from the Field*, National Research Council, The National Academies Press, Washington, 2004.
- [5] Lillian Lee, "I'm Sorry Dave, I'm Afraid I Can't do That": Linguistics, Statistics and Natural Language Processing Circa 2001, in [4] pages 111-117.
- [6] Editorial, "Compete, Collaborate, Compel," *Nature Genetics*, **39**, 931 (2007)  
<http://npg.nature.com/ng/journal/v39/n8/full/ng0807-931.html>
- [7] J.R. Platt. Strong inference. *Science*, 146(3642):347–353, October 1962.
- [8] R.A. Brooks, "Intelligence without Representation," *Artificial Intelligence*, **47**, 139-159 (1991)
- [9] G.E. Hinton, "Learning multiple layers of representation", *Trends in Cognitive Sciences*, **11**(10), 428–434 (2007)
- [10] P.E. Utgoff and D.J. Stracuzzi, "Many-Layered Learning", *Neural Computation*, **14**(10), 2497-2529 (2002)
- [11] E.S. Raymond, *The Art of UNIX Programming*, Addison-Wesley, 2004.
- [12] Peter Lyman and Hal R. Varian, "How Much Information", 2003.  
<http://www.sims.berkeley.edu/how-much-info-2003>
- [13] F. Conway and J. Siegelman, *Dark Hero of the Information Age: In Search of Norbert Wiener, The Father of Cybernetics*, Basic Books, New York, 2005. (especially chapter 11)
- [14] CALO: Cognitive Agent that Learns and Organizes,  
<http://caloproject.sri.com/>

- [15] A. Hargadon and R.I. Sutton, "Technology Brokering and Innovation in a Product Development Firm," *Administrative Science Quarterly*, **42**(4), 716-749 (1997)
- [16] Giorgio Corani and Marco Zaffalon. "Naive credal classifier 2: a robust approach to classification for small and incomplete data sets." Technical Report IDSIA-08-07, IDSIA / USI-SUPSI, Dalle Molle Institute for Artificial Intelligence, September 2007.
- [17] T.L. Thai and H.Q. Lam, *.NET Framework Essentials*, O'Reilly, 2003.
- [18] Tim Harris, Simon Marlow, Simon Peyton Jones, and Maurice Herlihy. "Composable memory transactions," *ACM Conference on Principles and Practice of Parallel Programming 2005 (PPoPP'05)*  
<http://research.microsoft.com/~simonpj/papers/stm/stm.pdf>
- [19] Michel Bauwens, *The Peer to Peer Manifesto: The Emergence of P2P Civilization and Political Economy*, November 3, 2007,  
[http://www.masternewmedia.org/news/2007/11/03/the\\_peer\\_to\\_peer\\_manifesto.htm](http://www.masternewmedia.org/news/2007/11/03/the_peer_to_peer_manifesto.htm)

## Appendix

The future evolution of this plan will require detailed competitor analysis. Below is an example.

### **Competitor Platforms: Text and Language Processing**

A (small) number of competitor platforms exist, mainly in research labs, to what we propose here. However, these are generally platforms developed by a homogeneous research group exploiting expertise in a single sphere of research, such as machine learning, language processing, or image analysis. What is more rare (to our knowledge, non-existent) is a common platform that deploys basic functionality from these different areas in a way that enables new advances and solutions by drawing on components from a common framework.<sup>4</sup> Such platforms miss the opportunity to exploit synergistic co-development of components from the different research areas that could lead to richer solutions to certain problems.

In the natural language processing area, the University of Sheffield's GATE<sup>5</sup> (General Architecture for Text Engineering) is the most extensive platform. GATE is an open-source toolkit for text mining and information extraction, providing a number of low-level (e.g. sentence-splitting, POS-tagging) and mid-level (e.g. entity recogniser) components for text processing, as well as higher-level tools (e.g. information extraction). As well as a comprehensive list of text processing components, GATE is also compatible with the UIMA architecture,<sup>6</sup> and is designed to integrate with related tools and platforms, such as the Lucene information retrieval engine and the WEKA machine learning toolkit.

OpenNLP<sup>7</sup> is another popular NLP toolkit; however, it is a much looser collection of open-source language-processing tools and not a common integrated platform in the same way as GATE. Within the biomedical text processing research world, the JULIE (Jena University Language and Information Engineering) Lab<sup>8</sup> has recently released a collection of NLP components as UIMA packages, specifically for processing biomedical documents. These include more generic components, such as tokenisers and sentence splitters, as well as taggers for genes and other biomedical entities. While the JULIE toolkit is not as comprehensive as GATE, the focus on the biomedical domain is pertinent, as a primary NICTA Business Theme.

For text-based information retrieval, there are a number of open-source or free retrieval engines. One of the most popular is the Lucene engine, which is maintained as an Apache project. The Nutch project builds on Lucene by providing auxiliary

---

<sup>4</sup> Note that platforms in, say, text processing may utilize machine learning techniques for certain tasks, such as entity recognition, but would use machine learning components developed by an external party.

<sup>5</sup> <http://gate.ac.uk/>

<sup>6</sup> Originally developed at IBM, UIMA (Unstructured Information Management Architecture) is now an Apache project and is becoming a *de facto* standard framework for text processing pipelines.

<sup>7</sup> <http://opennlp.sourceforge.net/>

<sup>8</sup> <http://www.julielab.de/>

components, such as web crawlers and the Carrot2 clustering engine for organizing search results. While Nutch is much narrower than even the text-specific component of the platform we envisage, its widespread use in applications demonstrates its power.

A more flexible information retrieval platform is the CMU/UMass Lemur toolkit, an open-source toolkit designed to support research and development in information retrieval, language modeling, and text mining. Lemur provides a number of tools and components, including IR necessities such as indexing (token and passage) and retrieval (ad hoc or language-model based, document or passage, relevance feedback). Lemur also contains components for related text tasks, such as summarisation and clustering.