Basic topics covered in Machine Learning and Data Mining during the 5 weeks of: 30 April, 7 May, 14 May, 21th May and 28th May. Due to time constraints I may not be able to cover all the topics.

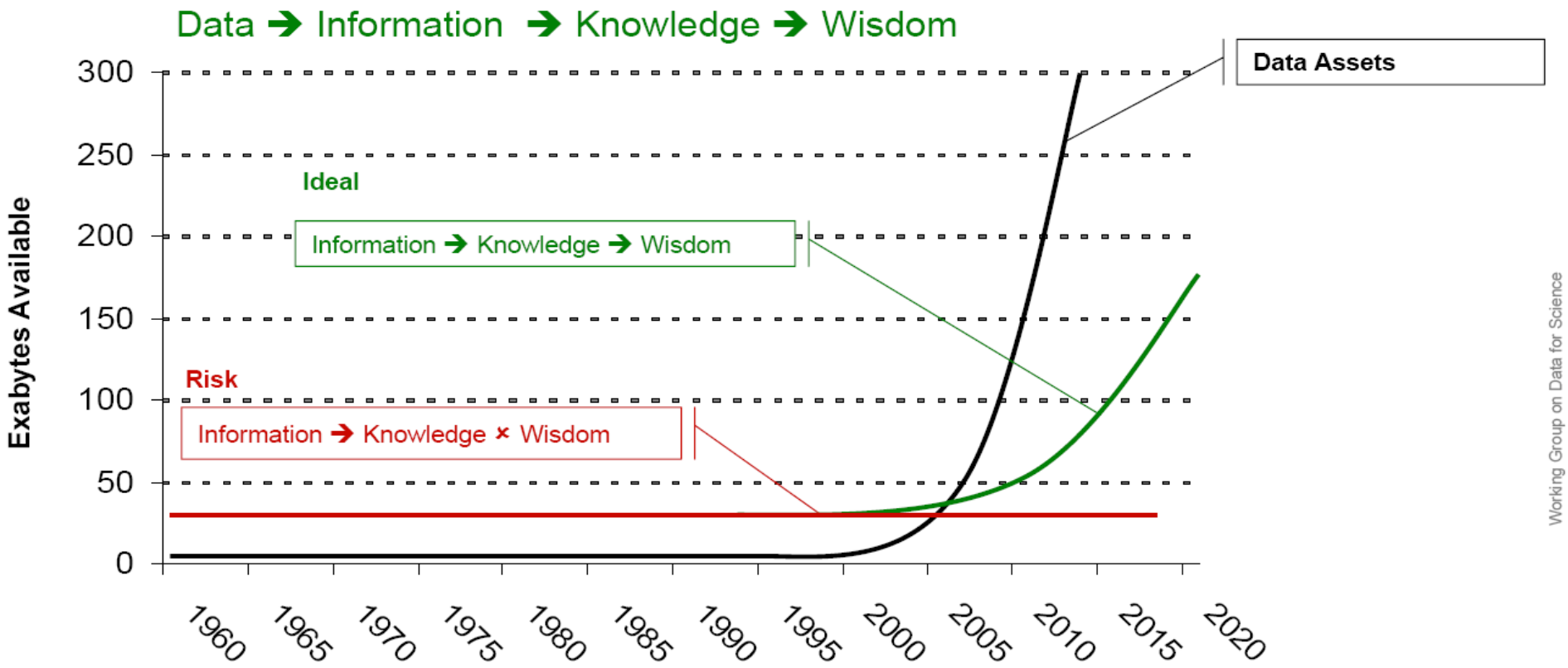
The course content will be as follows:

- Introduction to ML/Data mining motivating when you have to use such techniques
- Basic introduction to probability theory
- Basic machine leaning techniques
  - NB
  - Decision Trees
  - SVMs
- Pattern mining and how they can be used
  - Frequent pattern mining
  - Association rule
  - Feature selection
- Clustering techniques
- Evaluation measures of various machine learning techniques: E.g. Accuracy, F measure, AUC and Cross validation methods.

FROM DATA TO WISDOM:
Pathways to Successful Data Management for Australian Science

Working Group on Data for Science
Report to PMSEIC
December 2006

Reference: www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/documents/Data_for_Science_pdf.htm



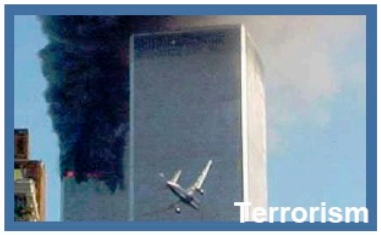
Based on: RioTinto – SETAC Asia Pacific Conference

Exabyte = 10^{18} bytes

" Information is not knowledge.
Knowledge is not wisdom.
Wisdom is not truth.
Truth is not beauty.
Beauty is not love.
Love is not music.
Music is THE BEST."
-- Frank Zappa

Frank Vincent Zappa[1] (December 21, 1940 – December 4, 1993) was an American composer, electric guitarist, record producer, and film director [Wikipedia].

Need for data collection and Knowledge acquisition



Diversity of data sources

Tackling the challenge of knowledge management and discovery at a massive scale

- Database modelling and integration has long been a focus of Information Technology research and development. Classic example being the application of RDBMs for commercial apps.
- A major and accelerating trend is the focus of data integration from business and enterprise applications to scientific and personal applications.
- Exponential growth of data with the spread of the Internet, Web and the multitudes of automatic data generation and collection devices.
- This trend is expected to continue in the foreseeable future.

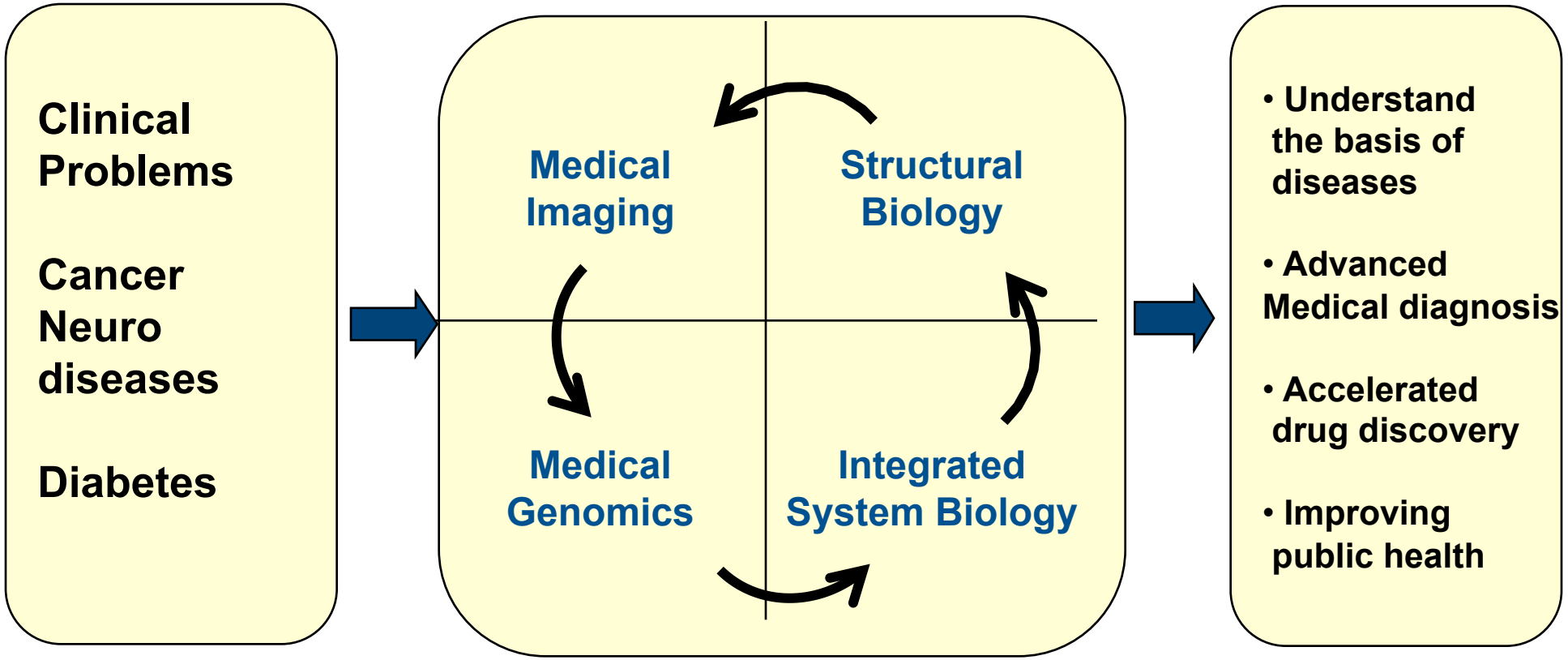
Life sciences research generating extremely rich and complex datasets

- Terabytes/day (and growing) MRI (magnetic resonance Imaging), next generation sequencing, metabolomics, large public projects linking research and clinical data
- Life Technologies announced in 2012 that they can already scan the full genome for $1000, and another company called Geniachip claims to go beyond this and deliver the same results for just $100 cheaper than MRI!
- Very high dimensionality, several hundreds of millions of data points
- Different formats: spatial/temporal/sequence/graph/streams/medical images/video/audio/…

Importance of Problem

- Current computational methods cannot handle magnitude and dimensionality of the data
- Decision makers and Scientists need techniques to help form hypotheses and make evidence based decisions

Example



Agri-science – Plant Genomics – Environmental Science

ref:Geoff Taylor

Research issues

- Data management
  - Issues: Scale, Data Quality: incomplete, inconsistent, multi source, …
  - Provide answers to user queries.

- Machine Learning and Data mining
  - Issues: Scale, dimensionality, data modalities (relational, graphs, text, in addition to problems faced by data management systems, etc.)
  - Discover models that fit the data, explain underlying processes that generated the in information.

Some Research issues

- Databases and information retrieval

  Issues: Scale, Query Languages,

  Data modalities (relational, graphs, text, etc.)

- Distributed computing: grid, peer-to-peer and cloud computing models

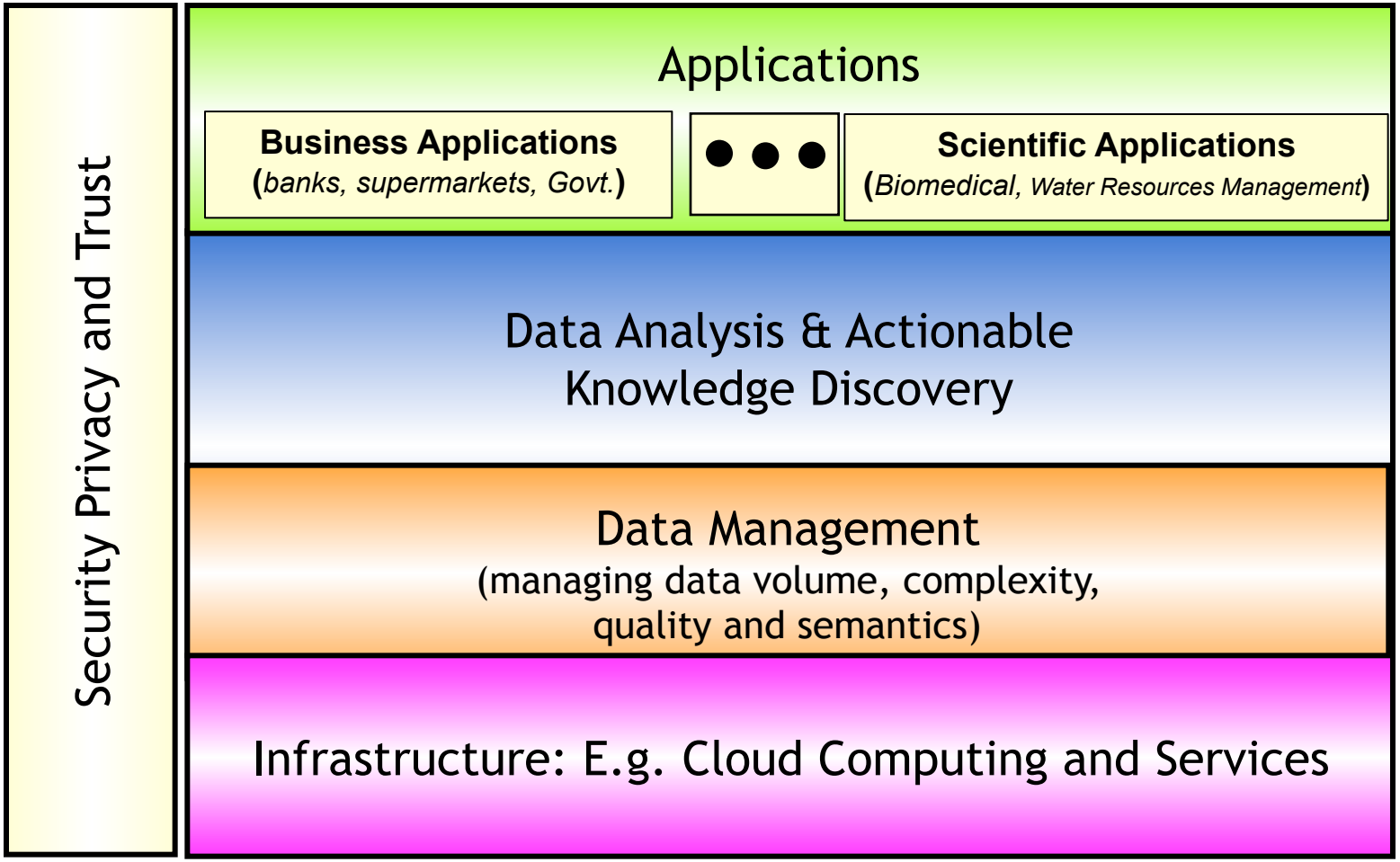
  Issues: parallelization (algorithmic aspects), distribution,

  load balancing, robustness, recovery, …

  Cloud computing is already making inroads and is going to dominate in the immediate future

- Security, Privacy and Trust

  Issues: scale, management, policy



Big pictures of Knowledge Technologies

Multi Disciplinary Scope

Cross disciplinary across many fields:
Engineering, Medicine, Commerce, etc.

Grand Challenges in Computational Biology Joint BSC - IRB Barcelona Conference Barcelona, 2-4 June 2008

> "There are very few cases of human activities growing faster than Moore's law (doubling number of transistors every 18 months), and biology is one of them. The exponential growth of sequence and structural databases, and the discovery of the complexity of most biomolecular interactions are giving rise to computational challenges …"

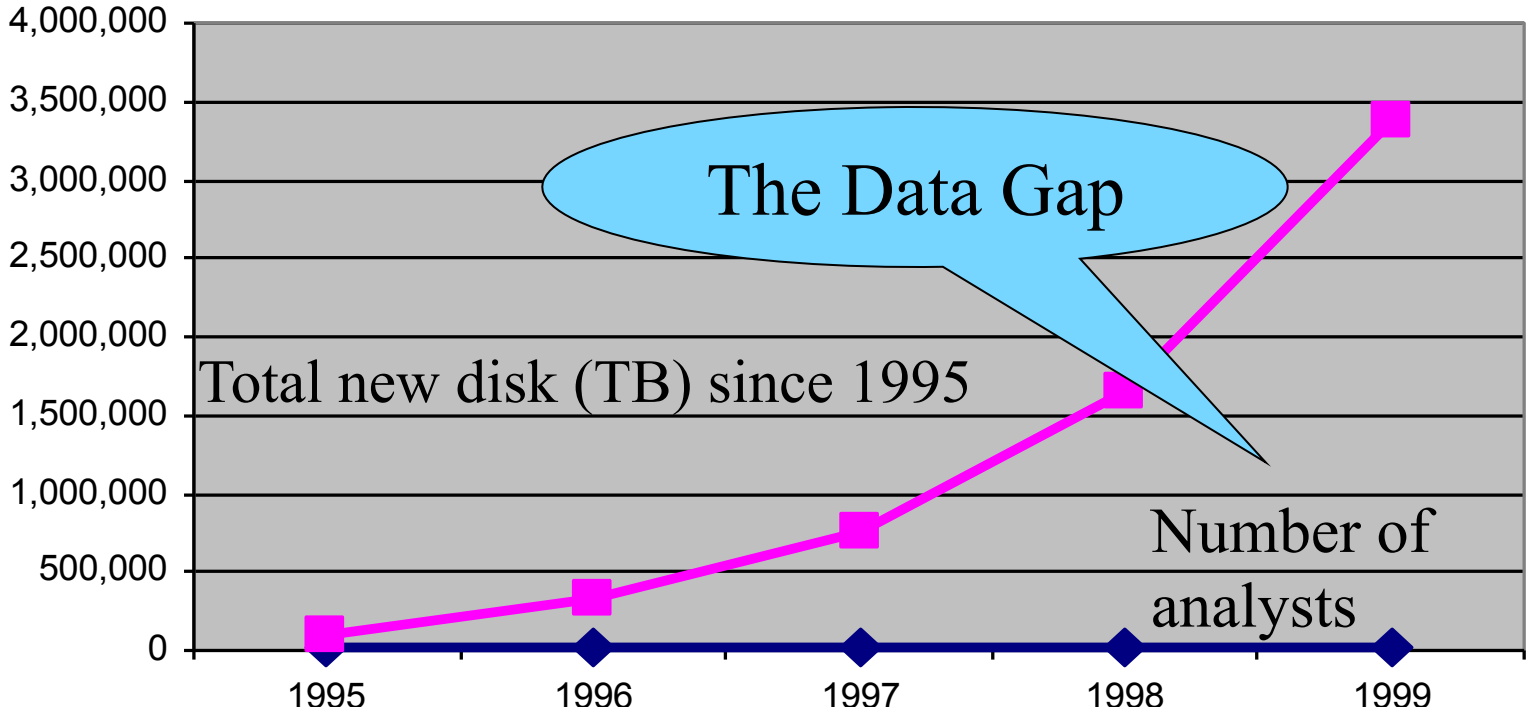
"These problem can be tackled by employing computational knowledge discovery tools."

Mining Large Data Sets - Motivation

There is often information "hidden" in the data that is not readily evident

Human analysts may take weeks to discover useful information

Much of the data is never analyzed at all



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Application areas of Machine learning and Data Mining:

- Bioinformatics
- Commerce
- Health
- Engineering
- Information retrieval (IR)
- Modelling complex systems
- Decision support systems
- Spam detection
- Language Technologies
- Almost all fields of science, engineering, life sciences, social sciences, …

Summary

- Exponential growth in collection and availability of data
- Society enforces decision makers to make decisions based on evidence
- Need effective tools to manage and process information
- Trust worthiness of information will become a major issue. Can computational methods ever determine trust?
- Data mining/Machine learning tools are required for processing a variety of complex data for aiding in decision making