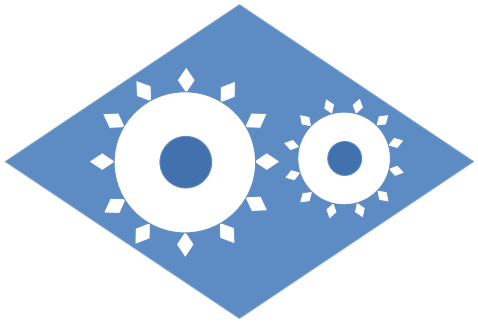
Text and Predictive Analytics with Machine and Deep Learning

Dr. Subrata Das

Adjunct Faculty, Villanova School of Business and Northeastern
Research and Development Consultant, MIT Lincoln Lab
Founder & President, Machine Analytics, Cambridge, MA

sdas@machineanalytics.com

https://www.linkedin.com/in/subrata-das-1293354

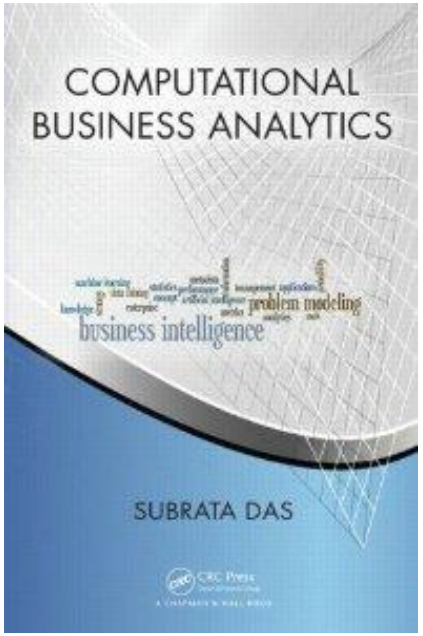


Outline

- Analytics and Unstructured Data
- Computational Modeling of Analytics
- Analytics Landscape

Analytics Defined

"… leverage data in a particular functional process (or application) to enable context-specific insight that is actionable." – *Gartner*

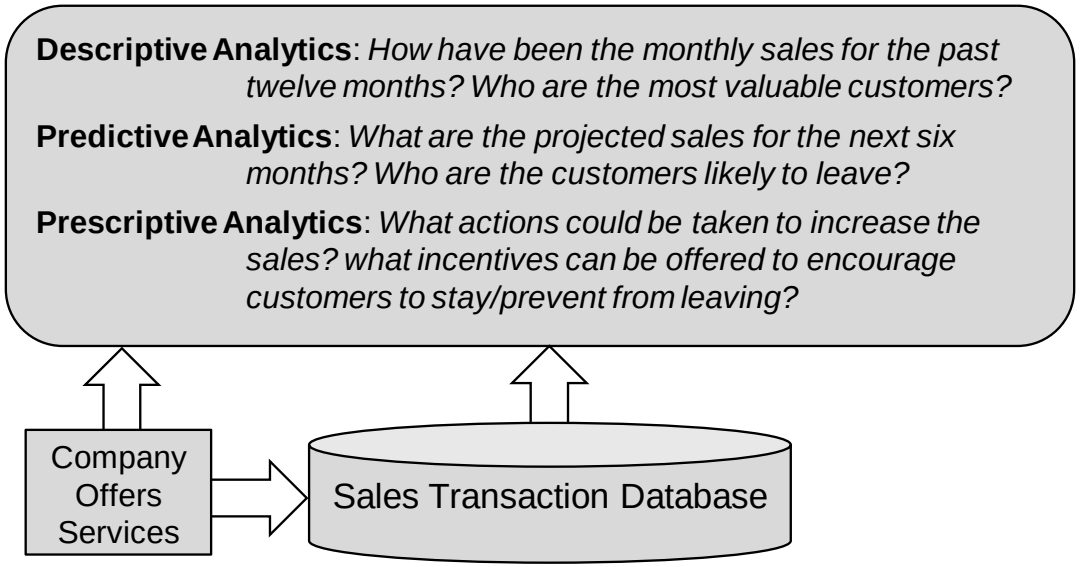


(CRC Press/Chapman & Hall, 2014)

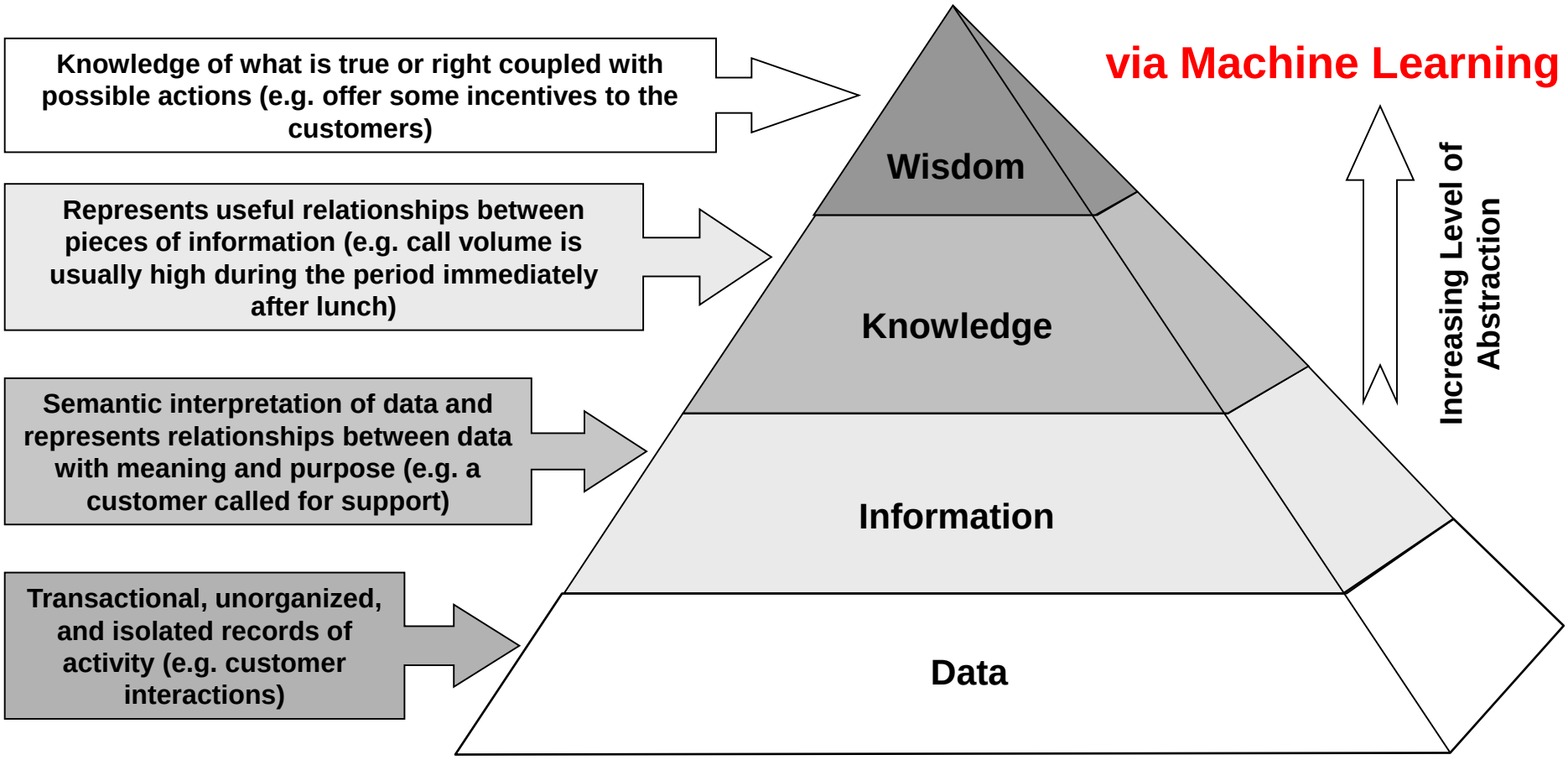
Analytics Categorized

- Descriptive analytics
  - Current and historical look at organizational performance.
- Predictive analytics
  - Predicts future trends, behaviour and events for decision support.
- Prescriptive analytics
  - a.k.a. decision support
  - Determines alternative courses of actions or decisions given the historical, current and projected situations, and a set of objectives, requirements, and constraints.

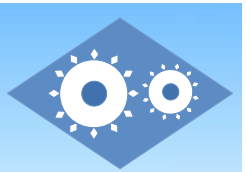
Machine Learning supports all three levels of analytics



Traditional Pyramid and Analytics

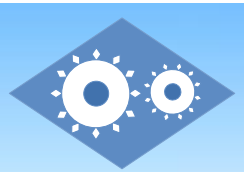


Analytics Examples



- **Customer Relationship Management:** (descriptive) How to best and profitably classify customers into category A (most valuable), B and C? (predictive): How to predict the probability that a customer will be lost within two years?
- **Marketing**: How to compute the likelihood of purchasing a product by each existing customer to launch an advertising campaign for the product?
- **Call Center**: How to assign the most able agent to an incoming call requiring specialized expertise?
- **Insurance**: How to estimate the probability of a claim (e.g. car accident) by an existing customer or by a new application using historical personal data?
- **Telecommunication**: How to clusters customers on the basis of collected historic data points (e.g. calls, text and multi-media messages, navigation, mail exchange) and then offer tailored offers?
- **Banking**: How to determine the credit-worthiness of new clients on the basis of historic data of past clients? How to determine credit card usage fraud based on the usage patterns?
- **Medical and Pharmaceutical**: How to determine possible side-effects of a drug given to a patient and the associated factors? How to determine current and future clinical state of a subject?
- **Logistics Supply Chain**: How to predict the number of goods consumed at different places?
- **Human Resource**: How to predict the financial impact of fundamental strategies such as pay differentiation, pay-at-risk, total rewards mix, and organizational structure?

The 7 Vs of Big Data



- Volume: Enormous volumes of data
- Velocity: Pace at which data flows in from sources like business processes, machines, networks, and human interaction with social media, mobile devices, etc.
- Variety: Many sources and types of data, both structured and unstructured
- Veracity: Biases, noise, and abnormality in data
- Value: Refers to the ability to turn data into value
- Validity: Is the data correct and accurate for the intended usage?
- Volatility: How long do you need to store this data?

Any more Vs?

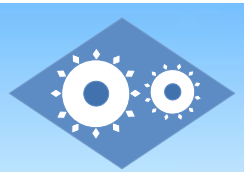
How big is BIG Data?

- The digital universe was 1.2 ZB in 2010 and is projected to be 35 ZB by 2020.
- 90% of the data in the world today has been created in the last 2 years alone.
- Only 1% of the world's data is analyzed today.
- CERN dataflow to permanent storage: 4-6 GB/s.
- Wal-Mart retail database: 2.5 PB.
- Facebook stores 240 billion photos and adds 350 million more everyday.
- Google stores about 30 trillion web pages which requires 100 TB for indexing.
- NASA Center for Climate Simulation stores 32 PB of climate observations and simulations.
- ISR acquisition systems gathered more than 53 TB of data every day in Afghanistan alone.
- Sources: IBM, IDC, Economist, EMC, Bits, Wiki, …

Name	Symbol	Decimal	Binary
Kilobyte	KB	10^3	1024 (2^{10})
Megabyte	MB	10^6	1024^2 (2^{20})
Gigabyte	GB	10^9	1024^3 (2^{30})
Terabyte	TB	10^{12}	1024^4 (2^{40})
Petabyte	PB	10^{15}	1024^5 (2^{50})
Exabyte	EB	10^{18}	1024^6 (2^{60})
Zettabyte	ZB	10^{21}	1024^7 (2^{70})
Yottabyte	YB	10^{24}	1024^8 (2^{80})

OLD!

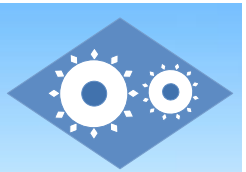
Big Data Scenarios



- Inherently distributed and maintained autonomously
- Stored in a centralized data warehouse
- Reside in a cloud

How do we manage and query?

Structured vs. Unstructured Data



- *Structured* data refer to computerized information which can be easily interpreted and used by a computer program supporting a range of tasks.
- Information stored in a relational database is structured, whereas a web page containing text, video and images is usually *unstructured*.
- Texts are sometimes categorized as *semi-structured*.

Structuring and classification are two fundamental capabilities of text analytics

Text Analytics and Examples

- Extract actionable intelligence from text documents.
  - **Sentiment analysis**: Analyze product reviews from social network data and determine trend.
  - **News**: Extract topic, analyze trend, and summarize and personalize new items.
  - **Litigation**: Categorize documents into relevant and non-relevant to a case.
  - **Human Resource:** Match resumes with a given job description.
  - **Finance:** Analyze analysts mood to determine market trend.
  - **Customer satisfaction**: Extract the most important concept from customer calls and notes from call center agents to input to the prediction model.
  - **Manufacturing**: Car or complex machine manufacturer analyze reports from repair shops to understand causes of frequent failures, providing early warnings to avoid costly product recalls.
  - **Life science**: To study the risk of patients who suffer from heart diseases, using both structured (e.g. blood pressure, cholesterol, age) and unstructured textual information of medical history (e.g. alcoholism) are relevant.
  - **Defense**: Human intelligence processing to extract evidence to assess situation and threat.
- Fundamental challenges are information classification, extraction, and structuring, and machine learning plays a crucial role.

What is Structuring?

⇐ Increasingly suitable for human consumption

Image	Text	Representation
	Homer is sitting on a chair drinking beer. An alternative: Homer is drinking beer while sitting on a chair.	RDF: metadata and triplets (homer, seating, chair) (homer drinking, beer) An alternative: Relational table:

Action	*Person*	*Object*
seating	homer	chair
drinking	homer	beer

Increasingly suitable for machine processing ⇒

Challenge is information extraction and representation

More complex

Audio and Video	Discourse	Representation
	*Steiger (Charley)*: "Oh, I had some bets down for you. You saw some money." *Brando(Terry)*: "You don't understand! I coulda had class. I coulda been a contender. I could've been somebody, instead of a bum, which is what I am, let's face it. It was you, Charley."	Perhaps a declarative logical syntax for representing dialogues and discourse and embedded in it a formalism for handling uncertainty.

Increasingly suitable for machine processing ⇒

Analysis of mood and arguments of the actors involved

Review Sentiment Determination

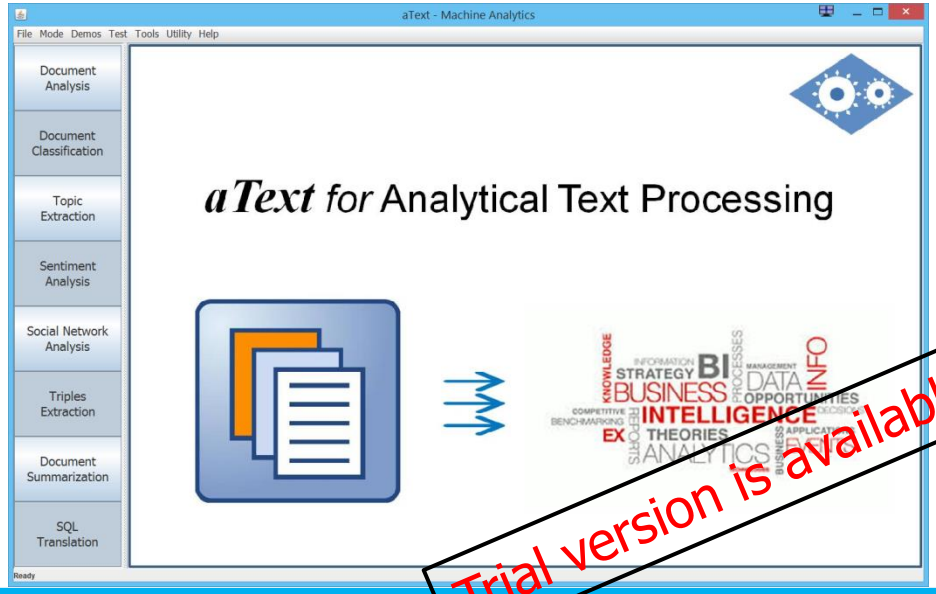
Fantastic For the money the location is fantastic It is about 300 yards from a metro station From the hotel you can walk to the colosseum you can practically see it when you leave the hotel Yes the rooms are small but the hotel has great character I would definitely stay there again and would recommend it to others. **Positive**

Terrible experience The twin room booked turned out to be a single with single bed and camp bed By our third night we had an ant infestation which the management were unwilling to deal with Eventually I had to spray the room and wait till 1 30am before being able to return. **Negative**

Given a set of labeled reviews, can we automatically classify an unlabelled review as positive or negative?

aText: Analysis of Text

- Extracts information from text and represents in the form of RDF triples via NLP, named entity recognition, coreference resolution, stemming, and dependency relation extraction.
- Supervised and unsupervised categorization of text documents.
- Applications:
  - Sentiment Analysis
  - Social Network Analysis
  - Document Summarization
  - Semantic Search



Outline

- Analytics and Unstructured Data
- Computational Modeling of Analytics
- Analytics Landscape

Computational Modeling

Real world problems and observations

Actionable intelligence or recommendations

Mimicking via computational models

Human-in-the-loop

Inference Engine

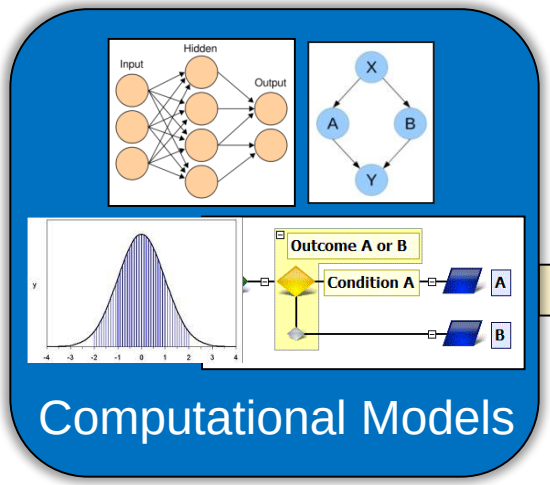
Uses **computational models** – mathematical or statistical description of a system or process

Descriptive Analytics: *How have been the monthly sales for the past twelve months? Who are the most valuable customers?*

Predictive Analytics: *What are the projected sales for the next six months? Who are the customers likely to leave?*

Prescriptive Analytics: *What actions could be taken to increase the sales? what incentives can be offered to encourage customers to stay/prevent from leaving?*

Model-based Analytics



Computational Models

Inference Engine

Analytics

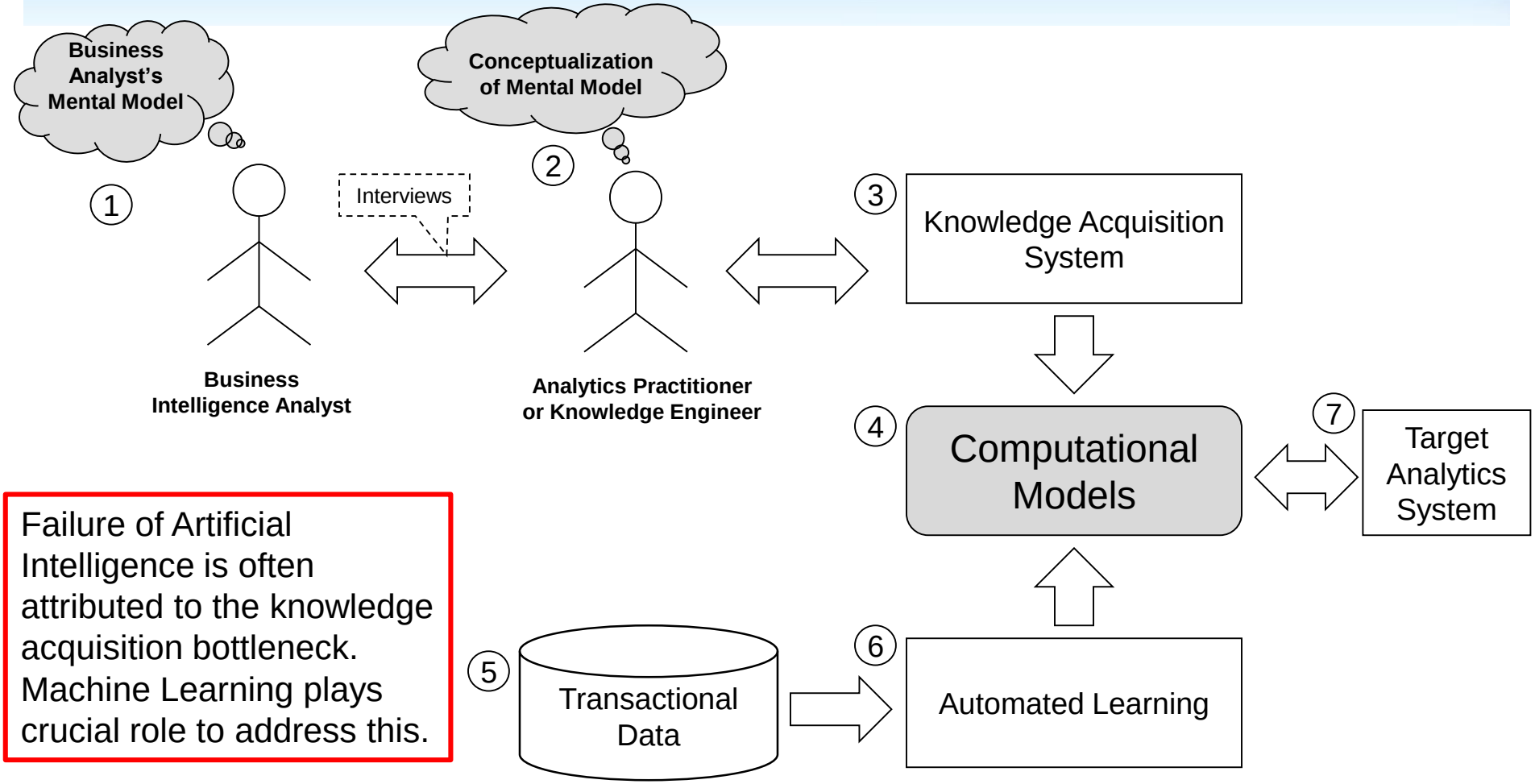
Unstructured Text

Structured Data

Information Collected and Extracted

How do we build models?

Typical Analytics Model Building

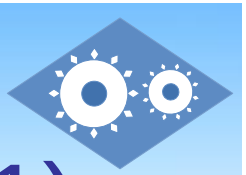


Failure of Artificial Intelligence is often attributed to the knowledge acquisition bottleneck. Machine Learning plays crucial role to address this.

Modeling Technologies

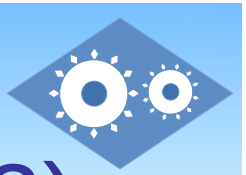
Paradigm	Overall Approach	Technologies
Statistical	Non-deterministic relationships between variables are captured in the form of mathematical equations and probability distributions	Test hypothesis, regression analyses, probability theory, sampling, inferencing, …
Machine Learning (ML)	System input/output behavior is observed, and machine learning techniques extract system behavior models	Clustering, neural networks, and various linear, nonlinear, and symbolic approaches to learning, …
Artificial Intelligence (AI)	Domain experts provide knowledge of system behavior, and knowledge engineers develop computational models using an underlying ontology	Logic-based expert systems, fuzzy logic, Bayesian networks, natural language processing, …
Temporal	Linear/nonlinear equations specify behavior of stochastic processes or of dynamic systems as state transitions and observations	Autoregression, survival analysis, Kalman filters, Hidden Markov Models, Dynamic Bayesian Networks, …

Traditional Statistical Approach (1)



- Descriptive statistics
  - Summarize and describe the information content of data that have been collected
  - Distribution, measure of central tendency, measure of dispersion
- Inferential statistics
  - Draws valid inferences about a population based on a sample
  - Make a generalization, test hypothesis, estimate, prediction or decision

Traditional Statistical Approach (2)



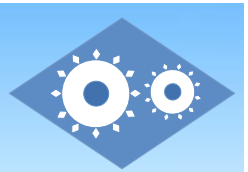
- Dependence methods
  - Use independent variable(s) to predict dependent variable(s)
  - Linear, logistics and kernel regressions, auto-regression, factor analysis, survival analysis
- Interdependence methods
  - Variables involved are independent variables
  - Hierarchical and k-means clustering, linear discriminant analysis, multidimensional scaling

Augmentation and Enrichment

Some examples

Statistical	Machine Learning
k-Means, Hierarchical, kNN Clustering/Segmentation	SVM, LSA, LDA, Probabilistic LSA, Hierarchical PLSA and LDA
Factor Analysis	Subspace methods (PCA, LSA, ICA)
Linear and Logistics Regressions	Bayesian Regression and Networks
ARMA, ARCH, and their extensions	Kalman Filters, HMM, Gaussian Process
Survival Analysis	Dynamic Bayesian Networks
Decision Trees (CART)	C4.5, Forests, Influence Diagrams, Rule Induction
Monte-Carlo Simulations (M-H)	Particle Filter and extensions

Augmentation and Enrichment: Survival Analysis Example



- The time from the beginning of an observation period (e.g. surgery) to an event (e.g. death, end of the study) or loss of contact/withdrawal from the study.
- A censored subject may or may not have an event after the end of observation time; right censoring: at the time of observation, the relevant event had not yet occurred
- Kaplan-Meier (product-limit)/Life Table estimators of the survivor and hazard functions for the sample as a whole, or for separate subgroups; they are not multivariate regression models.

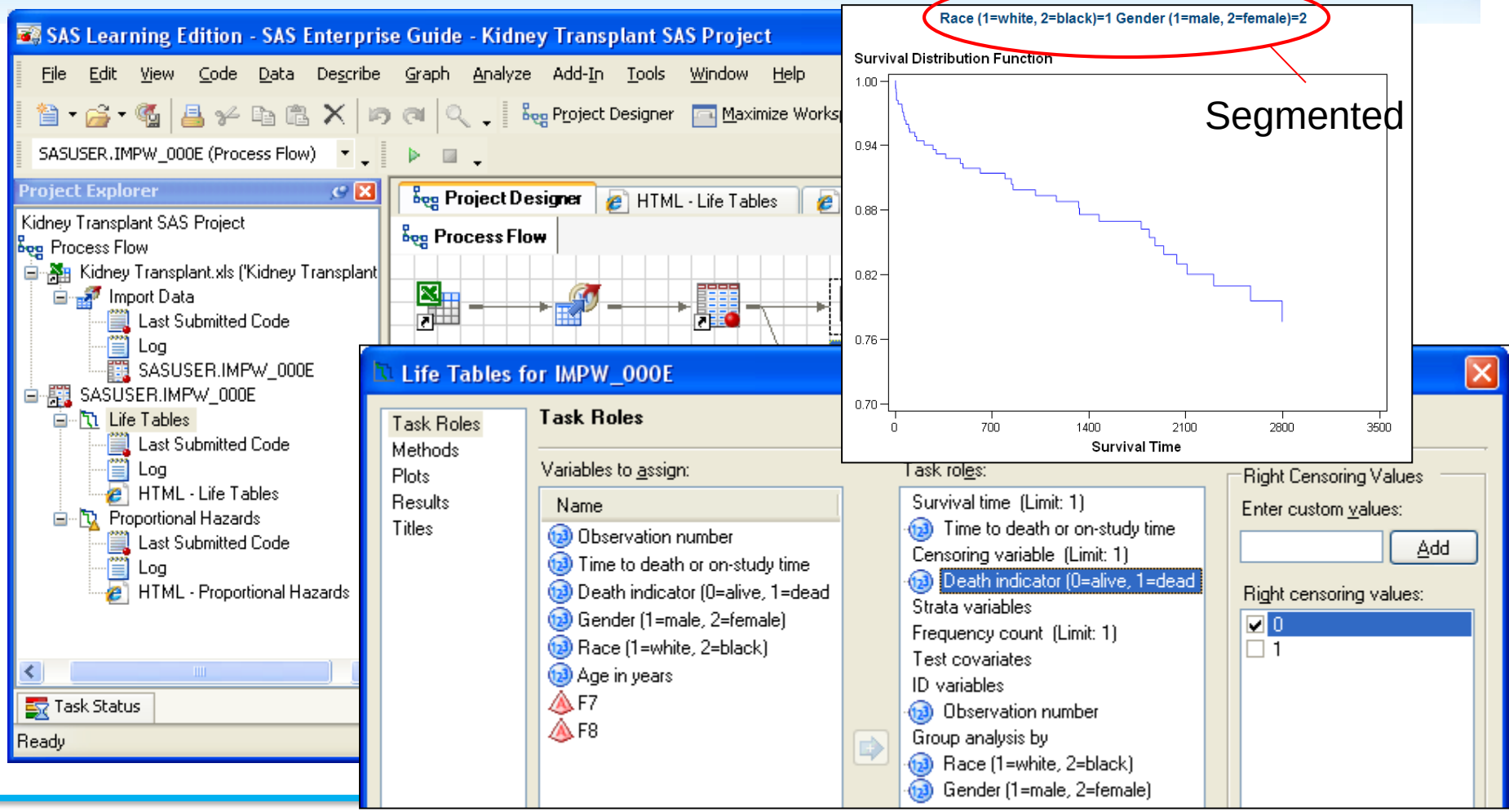
Kidney transplant data: Variables represented in the dataset are as follows (Source: Medical College of Wisconsin):

Observation number
Time to death or on-study time
Death indicator (0=alive, 1=dead)
Gender (1=male, 2=female)
Race (1=white, 2=black)
Age in years

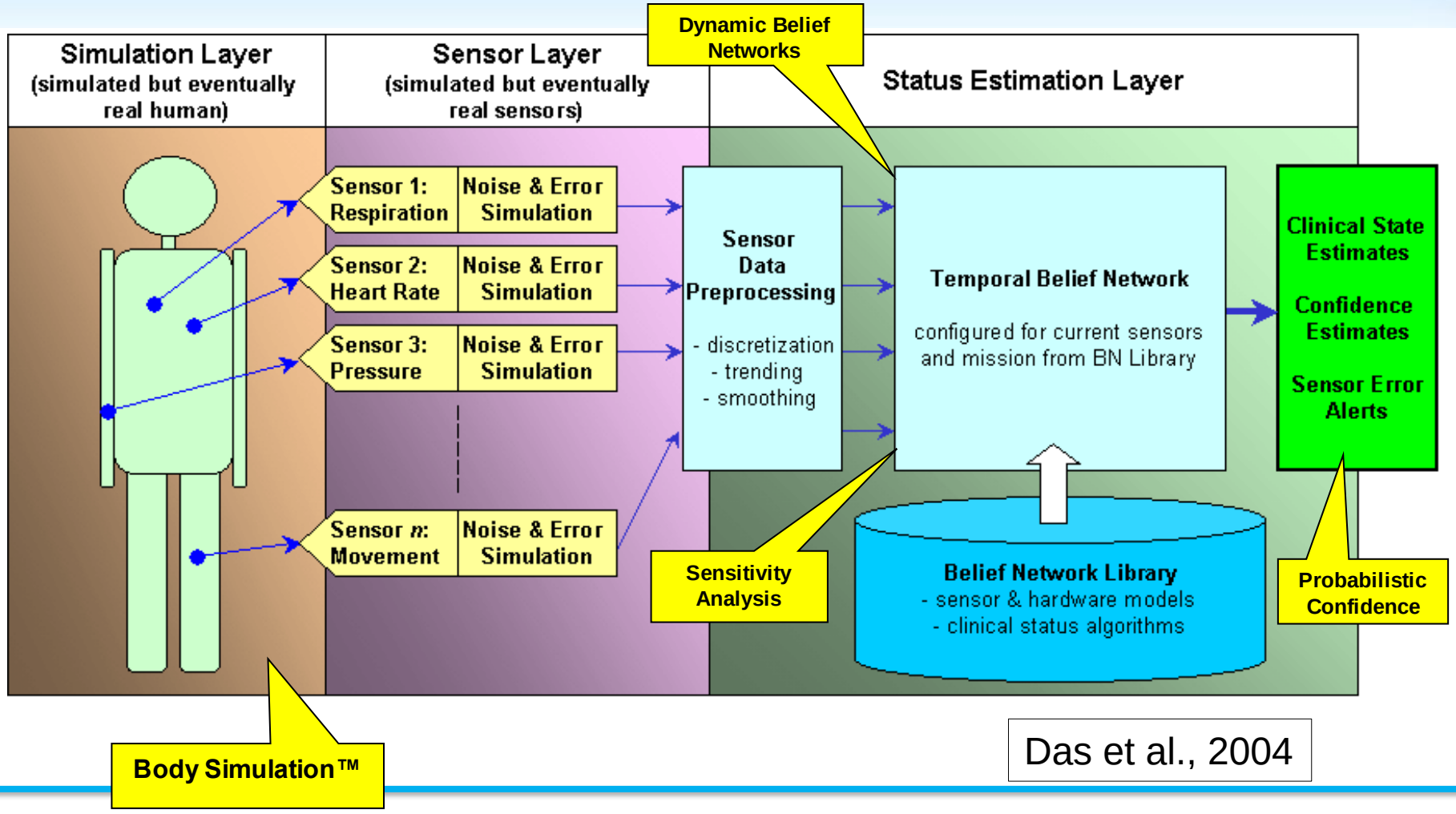
SAMPLE

1	1	0	1	1	46
2	5	0	1	1	51
3	7	1	1	1	55
….					
524	3430	0	1	2	28
525	1	0	2	1	41
526	2	1	2	1	60
…..					

SAS Demo of Survival Analysis

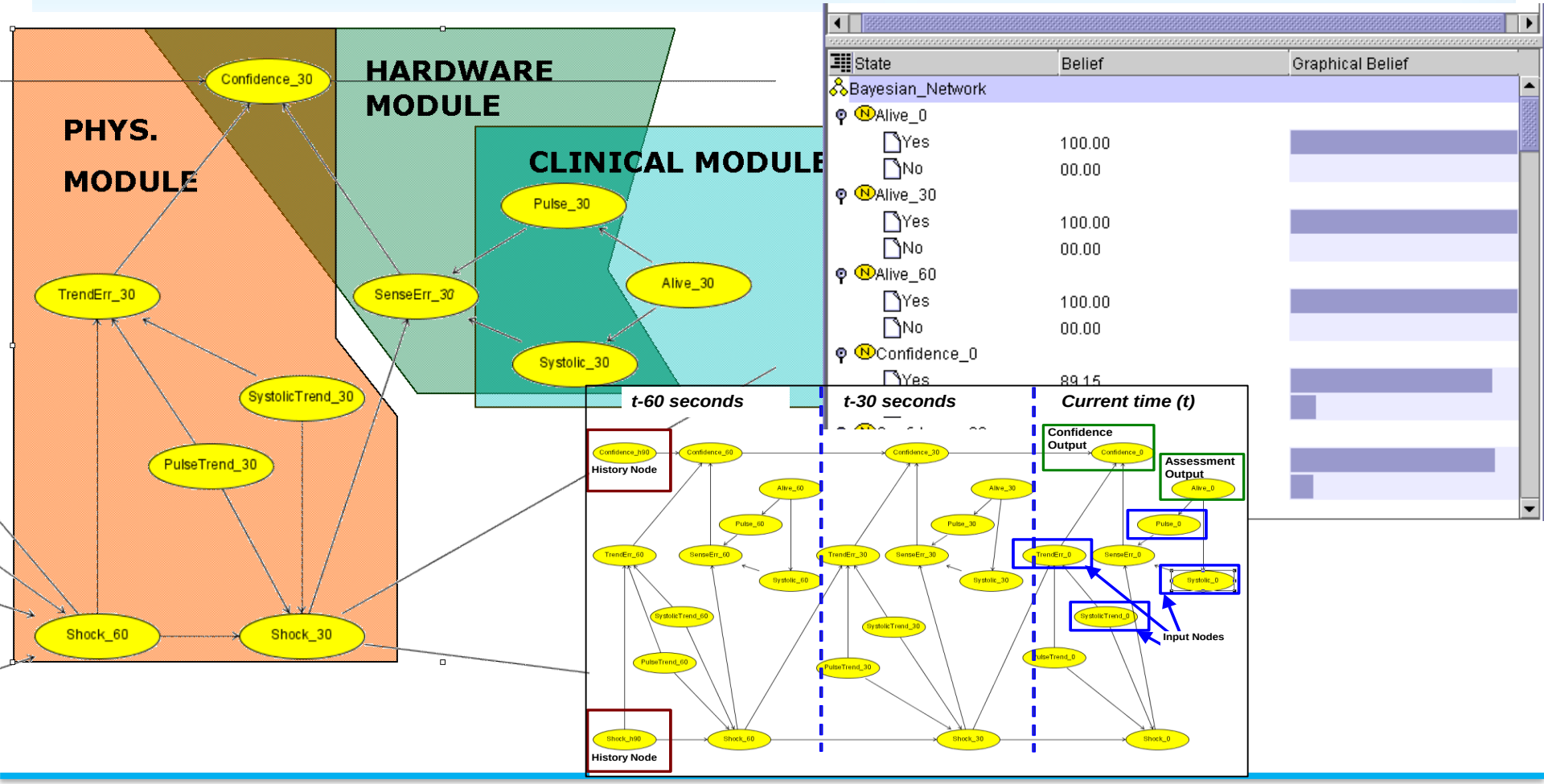


Life Status Estimation

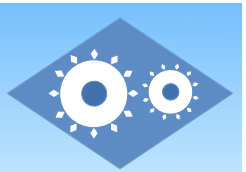


Das et al., 2004

Dynamic Bayesian Network for Life Status Estimation

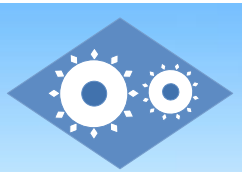


Benefits of Augmentation and Enrichment



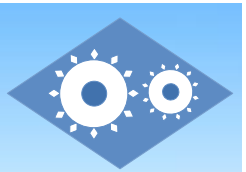
- Copes with inherently noisy, uncertain, and incomplete data; confidence is measured explicitly.
- Computational model naturally combines both discrete (categorical) and continuous (numerical) variables (e.g. shock and blood pressure).
- Both deductive and abductive reasoning (e.g. what actions should I take now if I want to keep the subject at a low shock stage in the next hour).
- Incorporation of well-defined concept (e.g. shock) into the model for which there is no direct measurement available.
- Model is extensible since, for example, a new module corresponding to a sensor device can be added easily.
- Output of a statistical survival analysis can be posted as evidence into the model.

Computational Analytics in Business



- Regression analysis, cluster analysis and decision trees are very common techniques
- Next comes neural networks
- Time series analysis (e.g. auto regression) is widely used
- Association rule mining is frequently talked about but specific techniques are not mentioned
- Techniques like PCA, ICA, SVM and reinforcement learning are upcoming
- Text analytics is popular but hard to understand results
- Lots of small and medium data mining companies providing analytics services based on commercial-off-the-shelf and in-house tools

Analytics Landscape



Blue Sky

Research to be transitioned between 5-10 years (e.g. deep semantical analyses, universal prediction, decision recommendation)

Cutting-Edge

Incubate ongoing research for differentiation (e.g. graphical models, deep learning, generative and markov modelling, probabilistic programming, active and semi-supervised learning, natural language processing, crowd and cloud analytics)

Low-Hanging Fruits

Leverage off-the-shelf well-proven AI/ML technologies (e.g. decision trees, neural networks, Bayesian networks, support vector machine)

Current Practice

Buy commercial-off-the-shelf statistical and business rule packages (e.g. SAS, SPSS, R, Data Miner, JBoss)