Big Data Analytics
--- CIS's Perspective on Big Data

1 October 2014

Xin Yao
http://www.cs.bham.ac.uk/~xin
President, Computational Intelligence Society (CIS)
CERCIA, School of Computer Science
University of Birmingham, UK

Overview

- Publications
  - An illustrative example
- Conferences
- Technical activities

IEEE Computational Intelligence Society

CIS Publications

IEEE TRANSACTIONS ON

NEURAL NETWORKS AND LEARNING SYSTEMS

A PUBLICATION OF THE IEEE COMPUTATIONAL INTELLIGENCE SOCIETY

http://cis.ieee.org/tnnls

SEPTEMBER 2014 VOLUME 25 NUMBER 9 ITNNEP (ISSN 2162-237X)

SPECIAL ISSUE ON COMPLEX- AND HYPERCOMPLEX-VALUED NEURAL NETWORKS

IEEE

IF 4.37

IEEE TRANSACTIONS ON

FUZZY SYSTEMS

A PUBLICATION OF THE IEEE COMPUTATIONAL INTELLIGENCE SOCIETY

www.ieee-cis.org/pubs/tfs

AUGUST 2014 VOLUME 22 NUMBER 4 IEFSEV (ISSN 1063-6706)

(Contents Continued on Back Cover)

IF 6.30

IEEE

IEEE TRANSACTIONS ON

EVOLUTIONARY COMPUTATION

A PUBLICATION OF THE IEEE COMPUTATIONAL INTELLIGENCE SOCIETY

http://cis.ieee.org/ieee-transactions-on-evolutionary-computation.html

AUGUST 2014 VOLUME 18 NUMBER 4 ITEVF5 (ISSN 1089-778X)

IF 5.54

IEEE

- **IEEE Computational Intelligence Magazine**
- **IEEE Transactions on Autonomous Mental Development**
- **IEEE Transactions on Computational Intelligence and AI in Games**
- ***And other co-sponsored Transactions***

IEEE

Special Issue of TNNLS

- *IEEE Transactions on Neural Networks and Learning Systems* **Special Issue on "Learning in Nonstationary and Evolving Environments" in January 2014.**
- **Tackling Veracity (uncertainty of data) and Velocity (analysis of streaming data) in particular.**

Some of the SI Papers

- Active Learning With Drifting Streaming Data
- PCA Feature Extraction for Change Detection in Multidimensional Unlabeled Data
- Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm
- Mining Recurring Concepts in a Dynamic Feature Space
- Learning in the Model Space for Cognitive Fault Diagnosis
- Dealing With Concept Drifts in Process Mining
- Learning Geotemporal Nonstationary Failure and Recovery of Power Distribution
- Continuous Dynamical Combination of Short and Long-Term Forecasts for Nonstationary Time Series

Some of the SI Papers

- Active Learning With Drifting Streaming Data
- PCA Feature Extraction for Change Detection in Multidimensional Unlabeled Data
- Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm
- Mining Recurring Concepts in a Dynamic Feature Space
- Learning in the Model Space for Cognitive Fault Diagnosis
- Dealing With Concept Drifts in Process Mining
- Learning Geotemporal Nonstationary Failure and Recovery of Power Distribution
- Continuous Dynamical Combination of Short and Long-Term Forecasts for Nonstationary Time Series

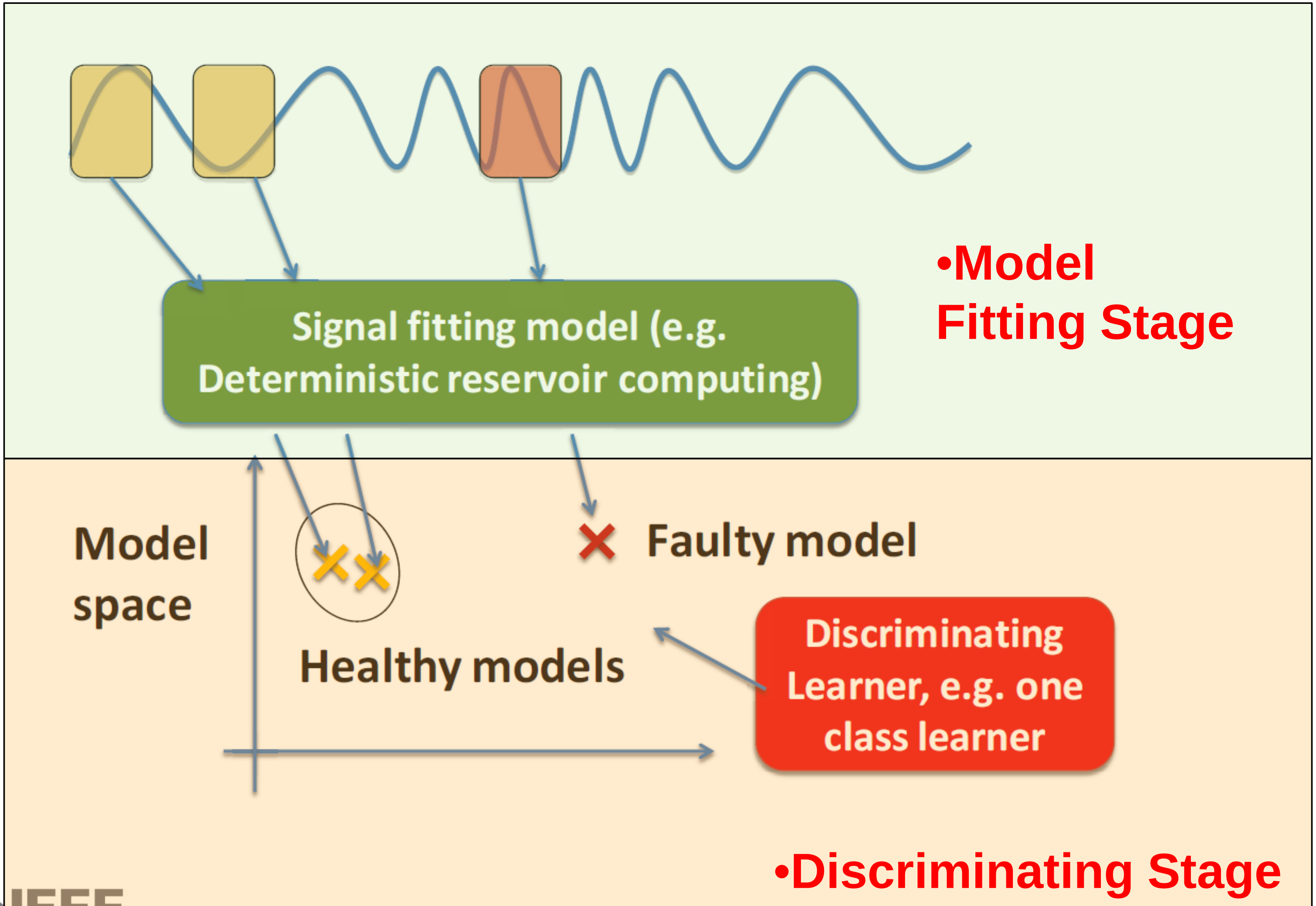
Overview

- **Publications**
  - **An illustrative example**
- **Conferences**
- **Technical activities**

An Example

- Traditional machine learning： in the data space
- Learning in the model space (3 steps)
  – Model generation: fit multiply models, e.g., generative models, to the data
  – Model measurement: define the distance between these fitted models
  – Model employment: develop learning algorithms in the model space

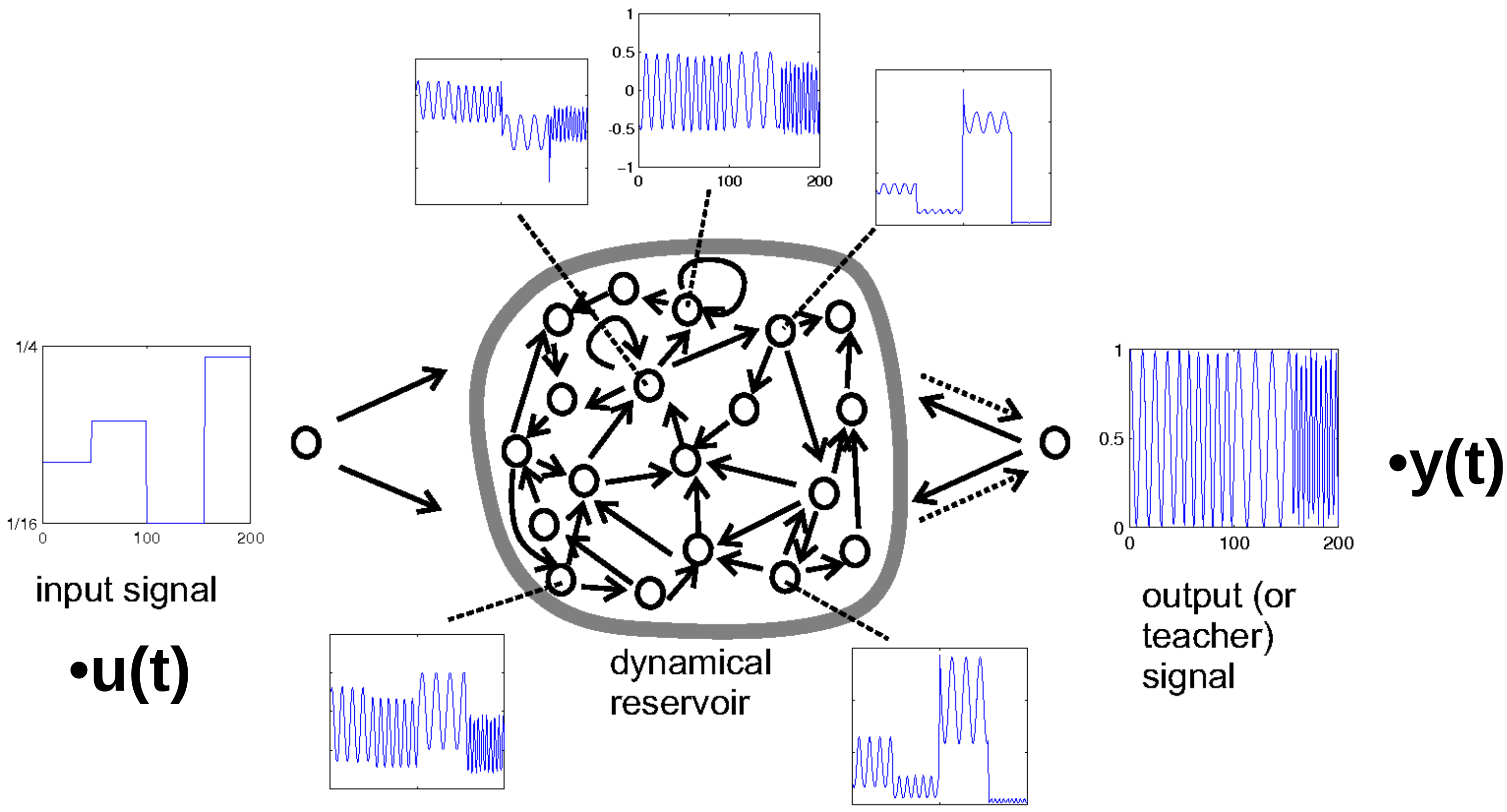
Learning in the model space



Why Model Space?

- **Disadvantages in data space**
  - **Missing values**
  - **High dimensionality**
  - **time-varying and *uncertain* environments**
  - **…**
- **Model (function) space is relatively smooth and easy to understand**

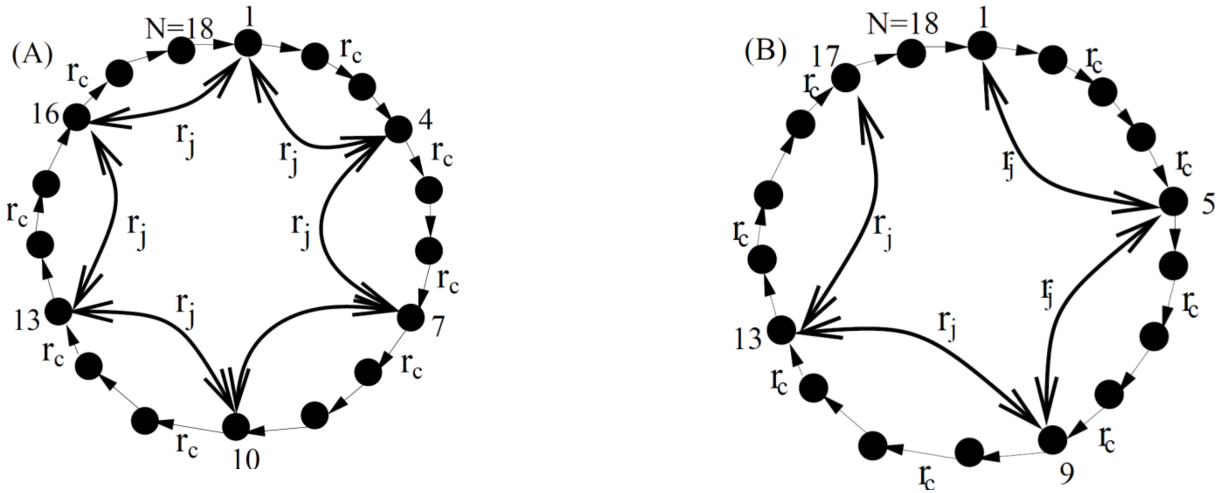
Reservoir Model



- **Solid arrows indicate fixed, random connections; dotted arrows trainable connections.**

Deterministic Reservoir Model

- Motivations to use the reservoir model :
  - input-output function with inner memory
  - The model can be trained fast and run in real-time
- Limitations of reservoir computing
  - Rely on random initialisation: unstable and dynamic
- Deterministic reservoir construction



•Simple cycle topology with regular jumps

Distance in the Model Space

- In the model space, the *m*-norm distance between models $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ is defined as follows:

$$L_m(f_1, f_2) = \left(\int_C D_m\left(f_1(\mathbf{x}), f_2(\mathbf{x})\right) d\mu(\mathbf{x}) \right)^{1/m},$$

  where $D_m\left(f_1(\mathbf{x}), f_2(\mathbf{x})\right) = \|f_1(\mathbf{x}) - f_2(\mathbf{x})\|^m$ is a function to measure the difference, $\mu(x)$ is the probability density function of the input domain, and *C* is the integral range.

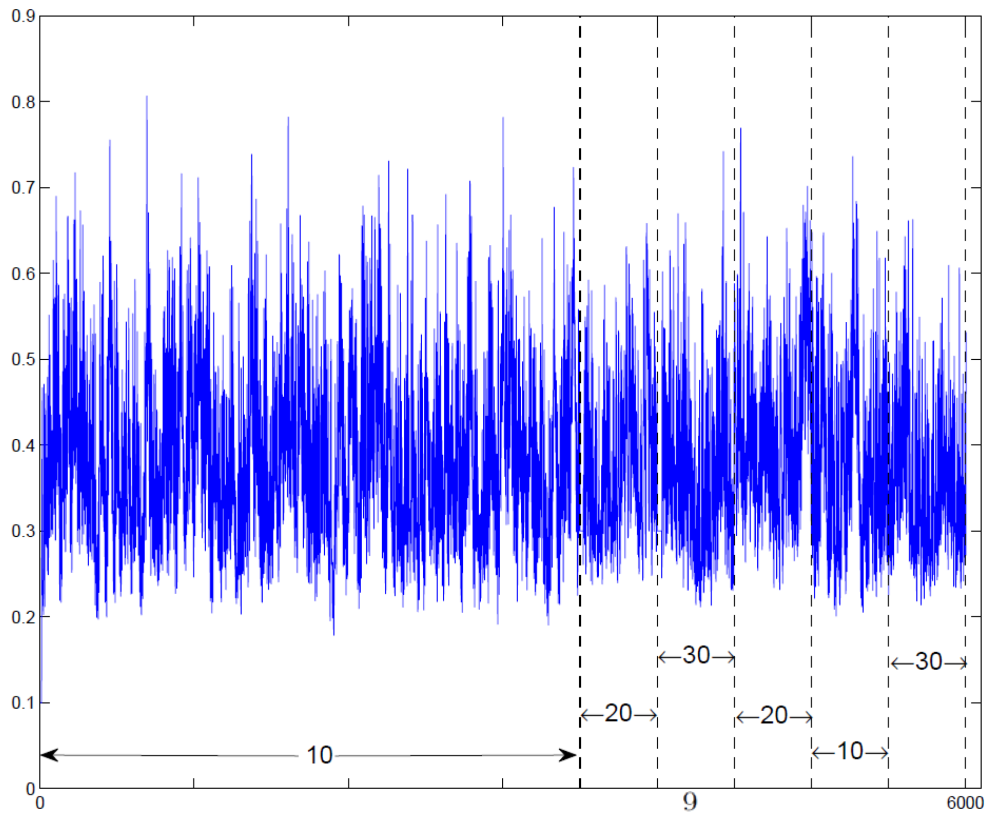
Learning Algorithms (in the Model Space)

- **Any distance based learning techniques can be used in the model space**
  - **Supervised learning: kernel approaches, kNN, etc**
  - **unsupervised learning: manifold learning, kernel clustering, etc**
  - **…**

- **For one class SVMs in the model space, the data distance is replaced by the *model distance***

$$\phi_\sigma(f_i, f_j) = \exp\{-\sigma \cdot L_2(f_i, f_j)\}$$

Nonlinear Auto-Regressive Moving Average (NARMA): An Example



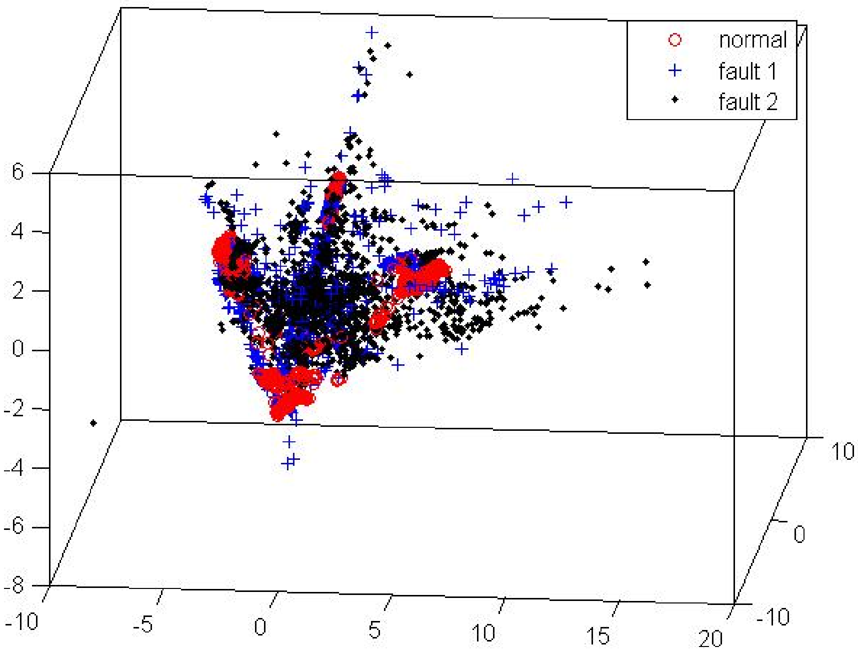
- 10 order
- 20 order
- 30 order

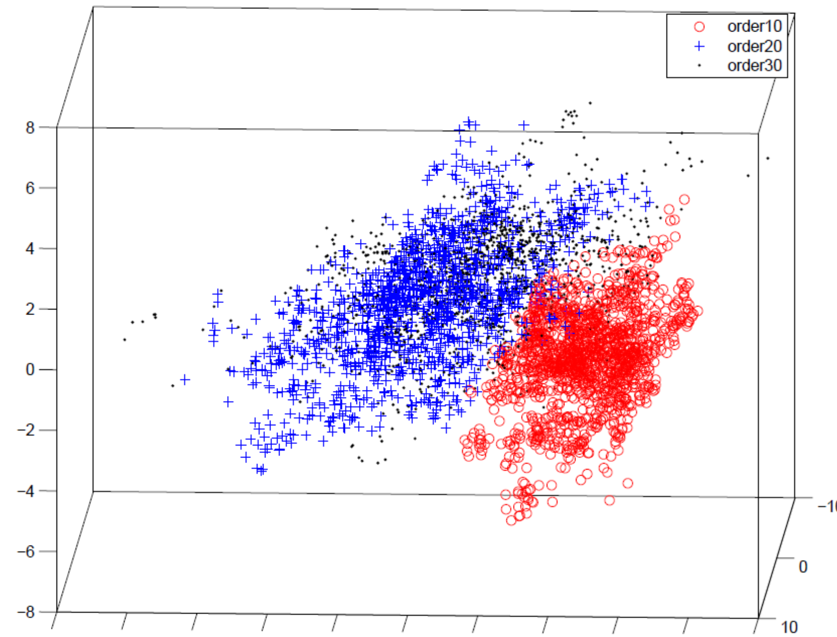
$$y(t+1) = 0.3y(t) + 0.05y(t)\sum_{i=0}^{9} y(t-i) + 1.5u(t-9)u(t) + 0.1$$

$$y(t+1) = \tanh(0.3y(t) + 0.05y(t)\sum_{i=0}^{19} y(t-i) + 1.5u(t-19)u(t) + 0.01) + 0.2$$

$$y(t+1) = 0.2y(t) + 0.004y(t)\sum_{i=0}^{29} y(t-i) + 1.5u(t-29)u(t) + 0.201$$

Signal vs. Model Spaces

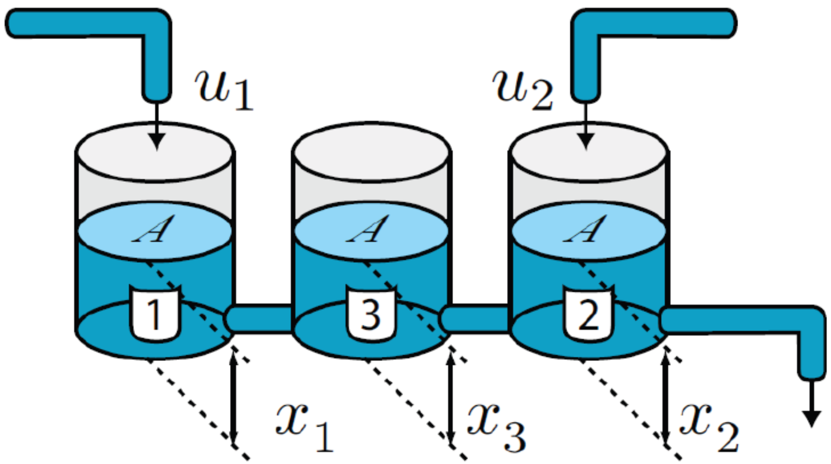




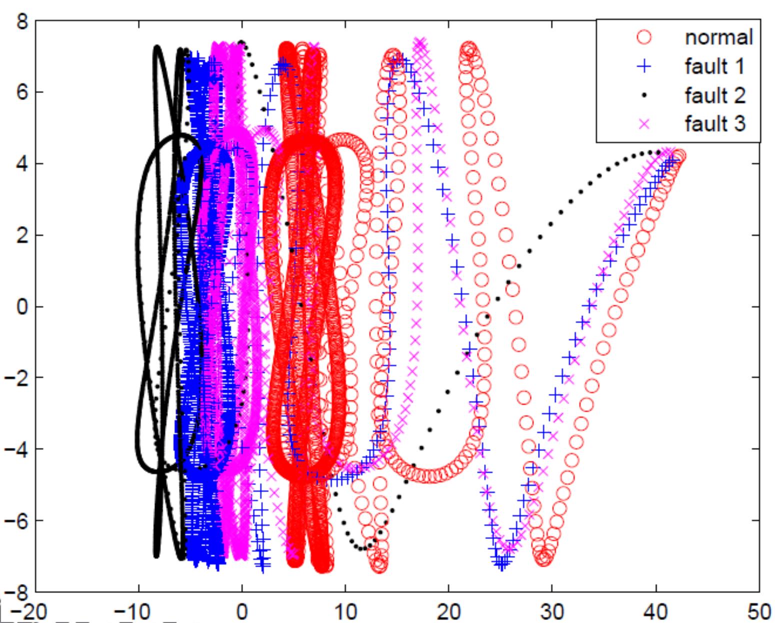
•Signal Space

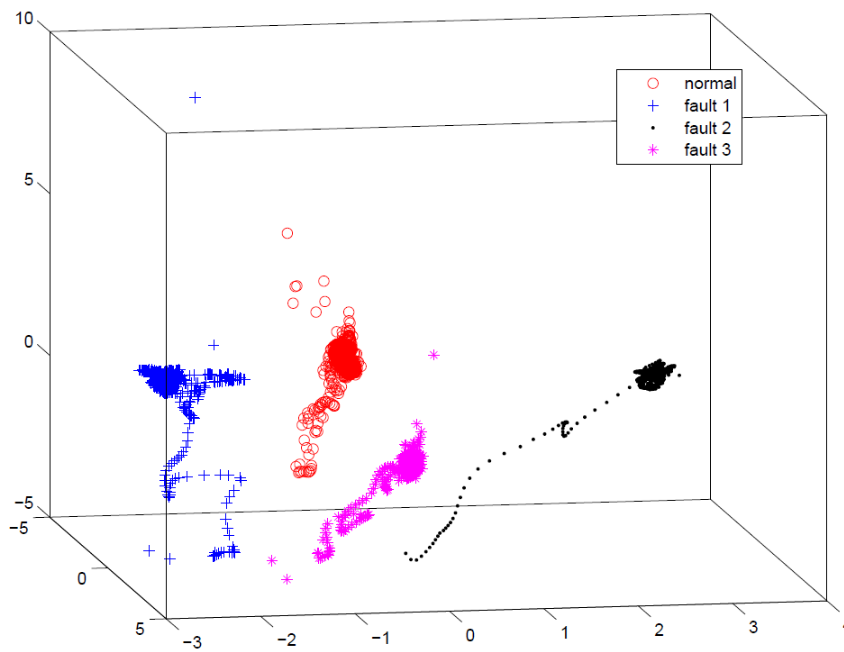
•Reservoir Model Space

Three Tank System

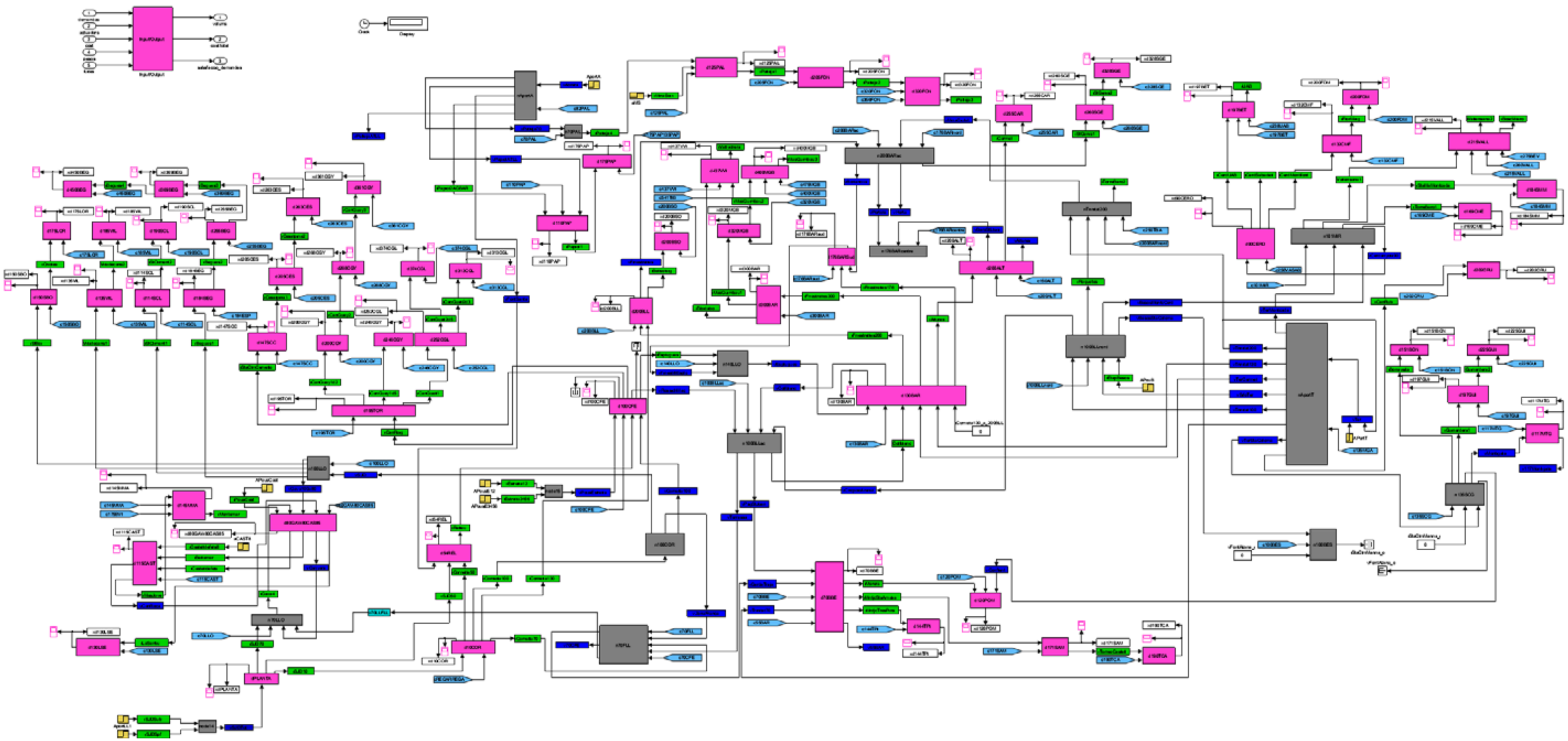


- Three faults:
  - ➢ Actuator fault in pump 1
  - ➢ Pump2 blocked
  - ➢ Actuator fault in pump 2

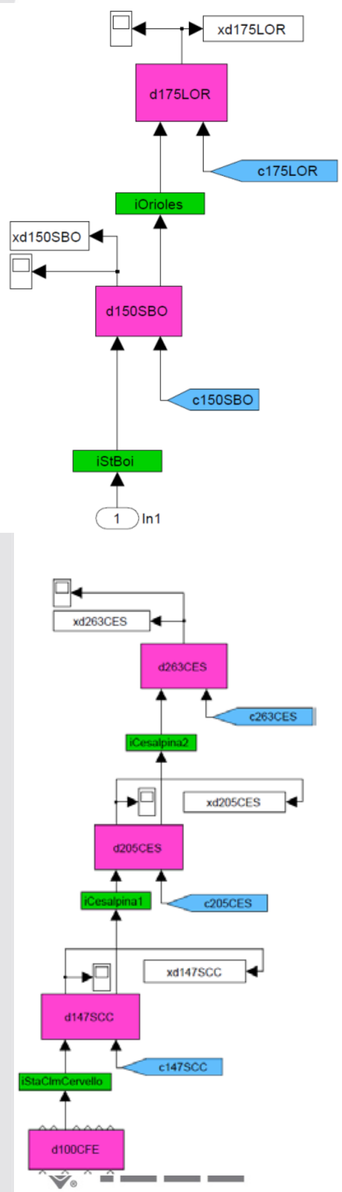




Case Study: Barcelona Water Distribution System



Barcelona Water Distribution Network: Faults



ID	Faulty Element	Type	Magnitude	ID	Faulty Element	Type	Magnitude
1	iOrioles	1	-25%	17	iStaClmCervello	3	0.01%
2	iOrioles	2	-25%	18	iStaClmCervello	4	0.5%
3	iOrioles	2	-10%	19	iStaClmCervello	5	-
4	iOrioles	3	0.001%	20	iStaClmCervello	6	4
5	iOrioles	3	0.1%	21	iCesalpina1	1	10%
6	iOrioles	4	10%	22	iCesalpina1	2	-15%
7	iOrioles	4	1%	23	iCesalpina1	3	0.01%
8	iOrioles	5	-	24	iCesalpina1	4	0.75%
9	iOrioles	6	2	25	iCesalpina1	5	-
10	c175LOR	1	-20%	26	iCesalpina1	6	0.75
11	c175LOR	2	-15%	27	c263CES	1	30%
12	c175LOR	3	0.01%	28	c263CES	2	-15%
13	c175LOR	4	1%	29	c263CES	3	0.025%
14	c175LOR	5	-	30	c263CES	4	0.5%
15	iStaClmCervello	1	-15%	31	c263CES	5	-
16	iStaClmCervello	2	-7.5%				

Type	Details & Parameter	Type	Details & Parameter
1	Additive offset (%MFD)	4	Additive drift (%MFD)
2	Additive incipient offset (%MFD)	5	Abrupt freezing (-)
3	Noise (variance %MFD)	6	Multiplicative offset (divided by)

- **31 faults in two sub-systems**
- **2 deterministic reservoir models**

Fault Detection Rate

	NARMA		Van der Pol		Three Tank		Barcelona Water	
Algorithm	FDR	FAR	FDR	FAR	FDR	FAR	FDR	FAR
T2	0.9072	0.1000	0.3009	0.0998	0.2311	0.0999	0.2316	0.1384
DBscan	1	0.0917	0.9146	0.2317	0.8958	0.0683	0.7981	0.1368
OCS-Model	1	0.1102	0.9310	0.0509	0.8521	0.1082	0.9313	0.2683
OCS-Signal	0.7042	0.2097	0.7686	0.2104	0.7521	0.2082	0.4920	0.3796
AP-Model	**1**	**0**	1.0000	0.3405	0.8407	0.1128	0.9014	0.2678
AP-Signal	1	0.5427	1.0000	0.7405	0.7155	0.2387	0.8879	0.2458
ARMAX-OCS	0.9882	0.0517	0.8727	0	0.9776	0	0.7369	0.1588
RC-OCS	0.9747	0.0558	0.9762	0.0158	0.8387	0	0.8271	0.1079
DRC-OCS(Sampling)	0.9789	0	0.9804	0	**0.9926**	**0**	0.9327	0.0817
DRC-OCS	0.9921	0	**0.9818**	**0**	0.9919	0	**0.9762**	**0.0473**

- **Comparisons of several algorithms in terms of fault detection ability, i.e. fault detection rate (FDR) and false alarm rate (FAR).**

Cognitive Fault Isolation

	NARMA (3 classes)				Van der Pol (4 classes)			
Algorithm	Classes	Precision	Recall	Specificity	Classes	Precision	Recall	Specificity
DBscan	4	0.6690	0.7650	0.8825	10	0.7629	0.6842	0.8018
AP-Model	271	0.9699	0.9698	0.9899	367	0.8778	0.8757	0.9585
ARMAX-OCS	5	0.9354	0.9229	0.9615	2	0.4309	0.4880	0.7868
RC-OCS	3	0.9637	0.9615	0.9808	6	0.9606	0.9583	0.9861
DRC-OCS(Sampling)	3	0.9683	0.9692	0.9914	5	0.9617	0.9726	0.9819
DRC-OCS	3	**0.9861**	**0.9858**	**0.9929**	5	**0.9736**	**0.9731**	**0.9910**
	Three Tank (4 classes)				Barcelona Water (32 classes)			
Algorithm	Classes	Precision	Recall	Specificity	Classes	Precision	Recall	Specificity
DBscan	14	0.8742	0.7561	0.9253	61	0.8019	0.7326	0.8654
AP-Model	272	0.9713	0.9704	0.9901	654	0.9366	0.9428	0.9751
ARMAX-OCS	5	0.9914	0.9923	0.9984	57	0.7826	0.7419	0.8237
RC-OCS	9	0.9182	0.8788	0.9596	44	0.8913	0.8942	0.9263
DRC-OCS(Sampling)	7	**0.9940**	**0.9949**	**0.9988**	39	0.9219	0.9310	0.9513
DRC-OCS	10	0.9931	0.9931	0.9977	48	**0.9538**	**0.9640**	**0.9871**

- **Comparisons of several algorithms in terms of fault isolation ability.**

Some Questions

- Raw data may be Big in terms of volume, but could they be represented by good models?
- Variety, Veracity, Variability in the raw data pose challenges, could generative models help to “smooth” them?
- Raw data might be unstructured, could we learn to structure them at runtime?

Overview

- Publications
  - An illustrative example
- Conferences
- Technical activities

Special Issue of CIM

- **August 2014 Special Issue of *IEEE Computational Intelligence Magazine* on "Computational Intelligence in Big Data"**
  - **The Emerging "Big Dimensionality"**
  - **Computational Intelligence Challenges and Applications on Large-Scale Astronomical Time Series Databases**
    - **"… The LSST will stream data at rates of 2 Terabytes per hour …"**

Upcoming CIM Special Issue

- Special Issue of *IEEE Computational Intelligence Magazine* on "New Trends of Learning in Computational Intelligence"
- Z. H. Zhou, N. V. Chawla, Y. Jin, and G. J. Williams. "Big data opportunities and challenges: Discussions from data analytics perspectives," *IEEE Computational Intelligence Magazine*, 2014 (to appear).

Special Issue of TCIAIG

- *IEEE Transactions on Computational Intelligence and AI in Games (T-CIAIG)* **Special issue on "Game Data Mining and Player Behavior Analysis Using in-Game Data"**
- **Big Data in games.**

Overview

- Publications
  - An illustrative example
- Conferences
- Technical activities

IEEE Computational Intelligence Society

Conferences

- **Panel discussions on Big Data** at WCCI 2014 in Beijing, China, 7-12 July 2014.
- **First IEEE Symposium on Computational Intelligence in Big Data**, 9-12 December 2014, Orlando, Florida, USA, as part of SSCI'2014.
- **2015 International Conference on Data Science and Advanced Analytics (IEEE DSAA2015)**, 19-21 October 2015, Grenoble, France.

IEEE

Overview

- Publications
  - An illustrative example
- Conferences
- Technical activities

Technical Activities

- The DMTC (Data Mining Technical Committee) has been changed to Data Mining and Big Data Analytics Technical Committee
- Two Big Data related task forces:
  - Big Data
  - Data Science and Advanced Analytics

Other TCs in CIS

- Neural Networks TC
- Fuzzy Systems TC
- Evolutionary Computation TC
- Computational Finance and Economics TC
- Games TC
- Adaptive Dynamic Programming and Reinforcement Learning TC
- Emergent Technologies TC
- Intelligent Systems Applications TC
- Bioinformatics and Bioengineering TC
- Autonomous Mental Development TC

Looking Forward
--- CIS Perspective

- **Outreach to non-academic communities in promoting and supporting our members in BD related activities.**
  - **One outreach workshop in Lima in March 2014**
  - **Another workshop in URI on October 3.**
- **Close collaborations with other IEEE societies in joint publications, conferences, technical activities, education activities, …**

Looking Further --- CIS Perspective (I)

- There appears to be more talks on Volume than other Vs, e.g., veracity, variability, and velocity.
- One reason might be that volume is easier to measure.
- CIS has been keen on other Vs as well, e.g., our TNNLS special issues.

Looking Further
--- CIS Perspective (II)

- **Is there a unique set of core knowledge that defines and underpins the Big Data?**
  - Where does it different from existing fields other than a vague word “Big”?
  - What are the new scientific questions that are unique to the Bid Data?
  - How can we capture such core knowledge?
  - How can we provide education programs to our members?

Looking Even Further --- CIS Perspective (I)

- What could and should be added to the core knowledge?
- What new technologies could be developed based on such core knowledge?

Looking Even Further --- CIS Perspective (II)

- How could CIS promote the fundamental advances of the BD field, especially in algorithms and analytics?
- How could CIS support its members in their pursue of the field? What would be appropriate education programs?
- What are the new opportunities for our members?

Thank you!