IBM **Research**

Towards Cognitive Manufacturing

Tsuyoshi Ide (“Ide-san”)
PhD, Senior Technical Staff Member
IBM Thomas J. Watson Research Center

Invited talk at IEEE International Workshop on Data Mining for Service
(DMS 2017, November 18, 2017), New Orleans, USA

Contents

Cognitive Manufacturing:
Future vision beyond the forth industrial revolution

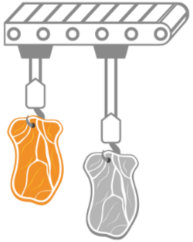
First

Industrial machines driven by steam power

First mechanical loom,1784

Second

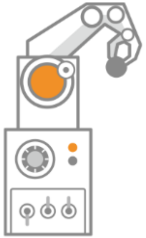
Introduction of **mass production** systems



First conveyor belt, Cincinnati slaughterhouse, 1870

Third

Automation with electronic devices



First programmable logic controller (PLC) Modicon 084,1969

Fourth

AI revolution: incorporation of **learning system**

TODAY

Degree of complexity

Fourth

Third

Second

First

1800

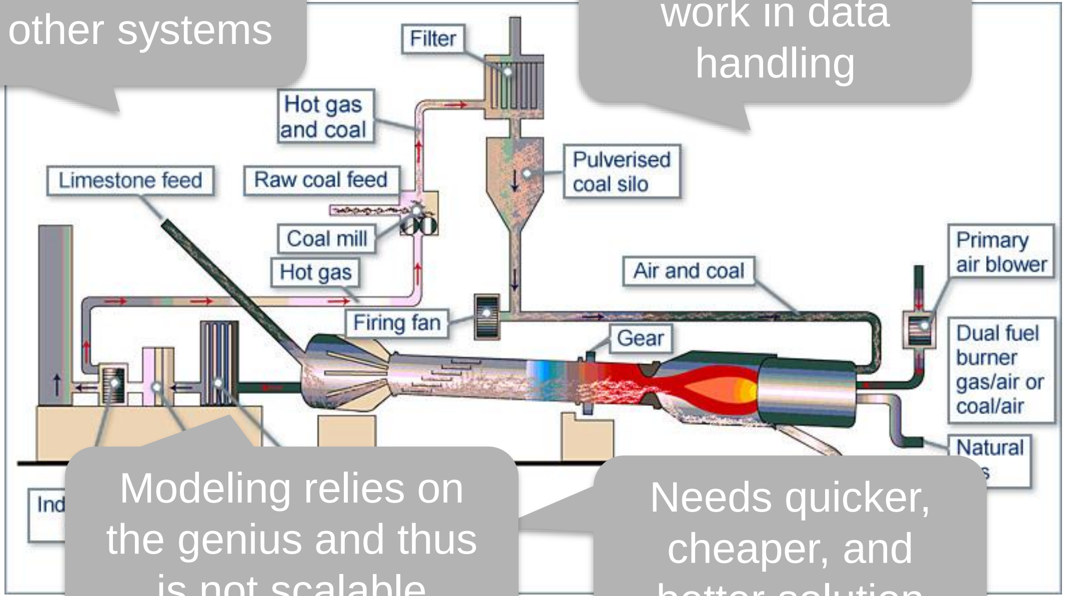
1900

2000

What is the difference from industrial automation in 70s?

- Real-time control of cement manufacturing plant based on time-series prediction*
  - Manual feature selection
  - Fitting autoregressive model
  - Optimal determination of control parameter through state-space modeling

- Mathematical model looks good enough



Isolated from other systems

A lot of manual work in data handling

Modeling relies on the genius and thus is not scalable

Needs quicker, cheaper, and better solution

Image: http://www.britishlime.org/education/

* T.Otomo, T.Nakagawa and H.Akaike, Statistical approach to computer control of cement rotary kilns, Automatica, 8 (1972) 35-48.

There still be technical challenges to transform data into business insights

Industry solutions

Linking existing ML/DM methods to particular business problems

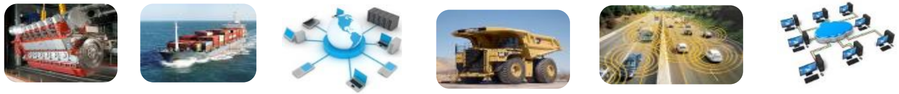
Developing new ML/DM methods for new tasks/data

IBM analytics platform

IBM IoT platform

In the IoT context, sensor data analytics is of particular importance

Devices and sensors



Cognitive technology (≒ statistical machine learning) transforms raw data into actionable rules

Data

knowledge

actionable rules

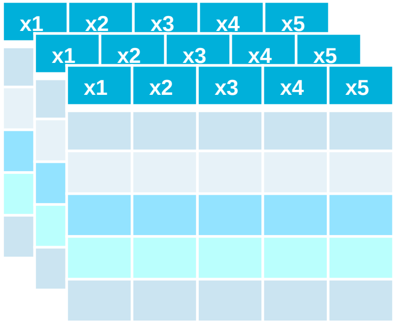
input data

Statistical machine learning

Learned patterns

(probability density or functional relationship)

Anomaly / change score



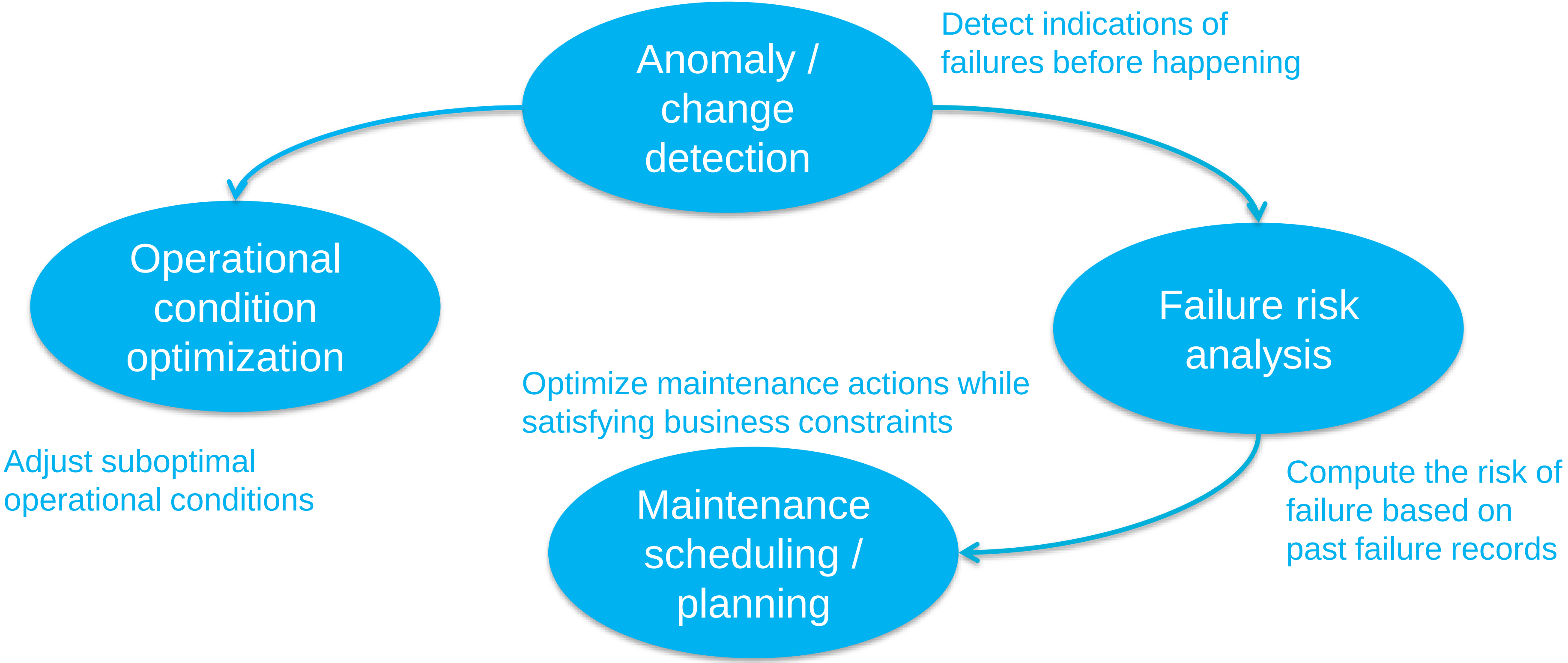
$$p^{(t)}(\boldsymbol{x})$$

$$a^{(t)} = -\ln p^{(t)}(\boldsymbol{x})$$

$$a^{(t)} = -\ln \frac{p^{(t)}(\boldsymbol{x}|y{=}+1)}{p^{(t)}(\boldsymbol{x}|y{=}-1)}$$

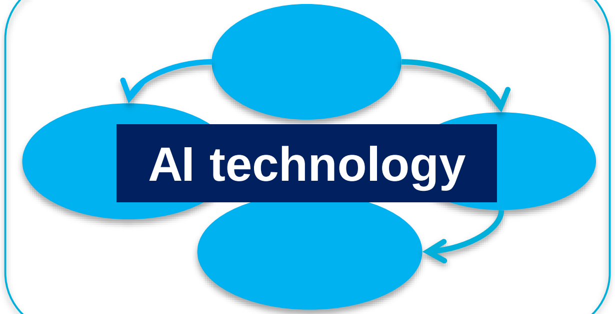
Statistical machine learning is a tool for knowledge acquisition and rule generation

Key technical areas of Cognitive Manufacturing (from an AI perspective)



Key technical areas of Cognitive Manufacturing

Database technology



User interface technology

- Get data ready
- Get algorithms work better

- Convert data into insights
- Obtain actionable rules

- "Dashboarding" data
- Help explore data / algorithms

Contents

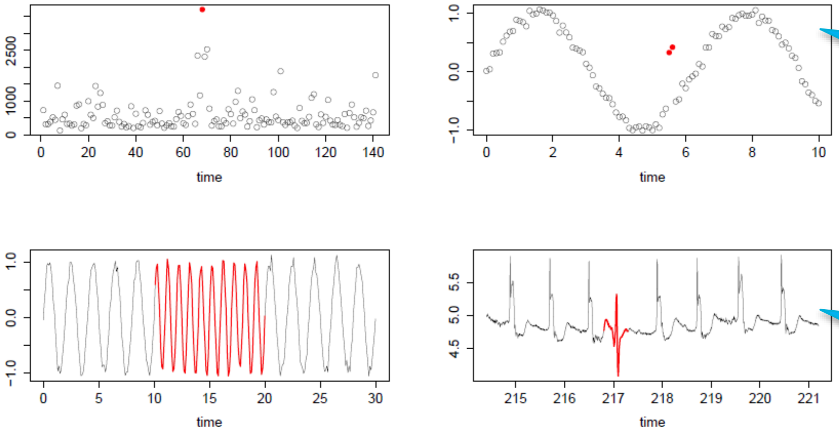
No “one-size-fits-all” algorithm

- Example of anomaly detection
  - “*Happy families are all alike; every unhappy family is unhappy in its own way*.” - *Anna Karenina*, Leo Tolstoy

Examples of anomalies

outliers (from i.i.d. samples)

outliers (from auto-correlated samples)



change points

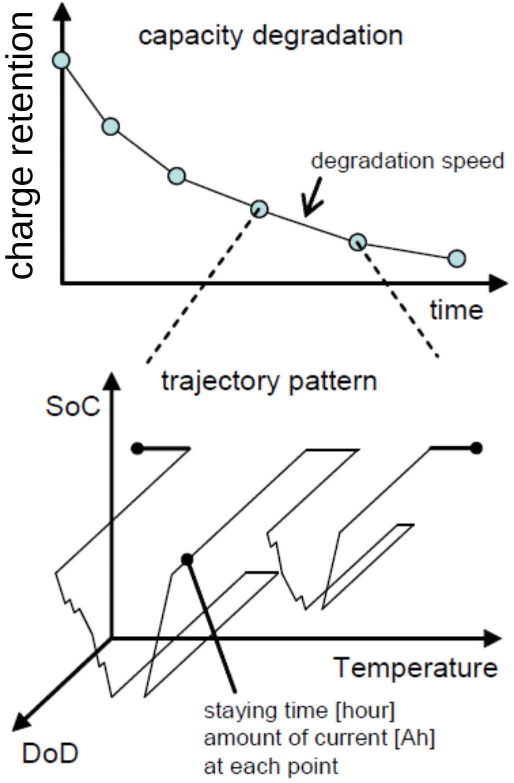
discords

井手, 杉山, 異常検知と変化検知, 講談社, 2015.

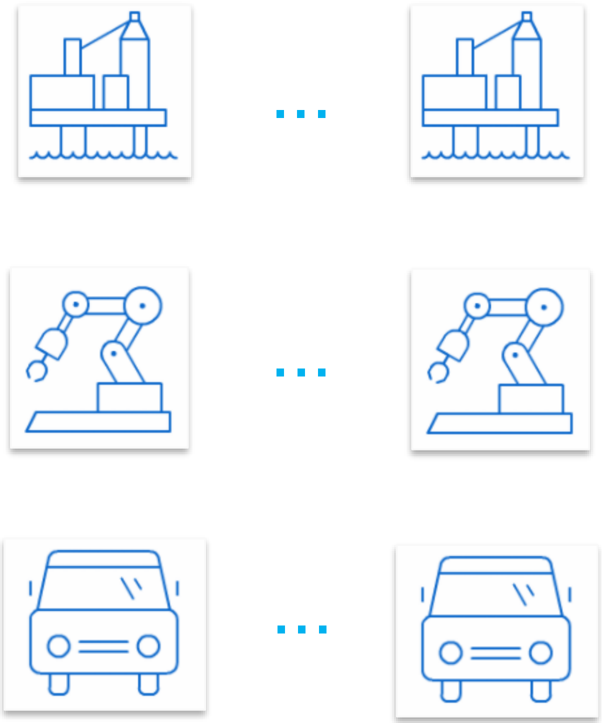
Ready-to-use solution to your problem might not even exist

- Example: Battery life prediction of electric vehicle batteries
  - Depends on the entire history of battery usage
  - Battery usage is represented as a complex trajectory of a multi-dimensional space

- Standard regression / time-series prediction methods are not applicable
  - Just buying a general-purpose ML package doesn't help

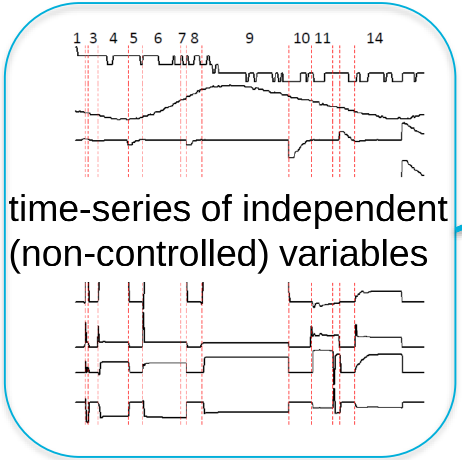


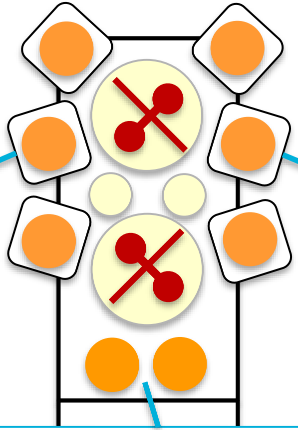
Non-standard setting is everywhere: Need for a collective approach to condition-based management

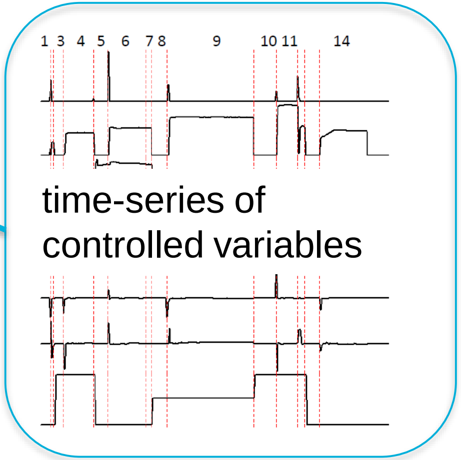


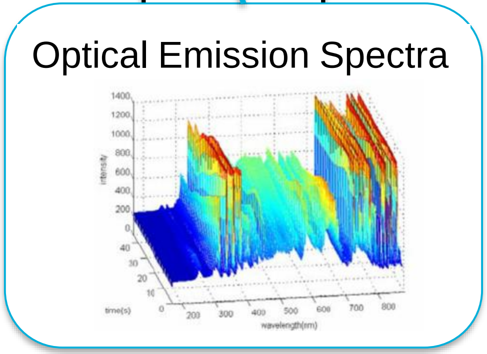
- Typically assets are managed as a cohort
  - 10s of off-shore oil production systems
  - 100s of industrial robots
  - 1000s of electric vehicles in a certain area
- Little is known to systematically to build/manage predictive models for cohort
  - Model building is too time-consuming
  - Model maintenance cost is prohibitive
  - Across-machinery performance comparison is hard

Non-standard setting is everywhere: Etching trace data take a form of higher-order tensor



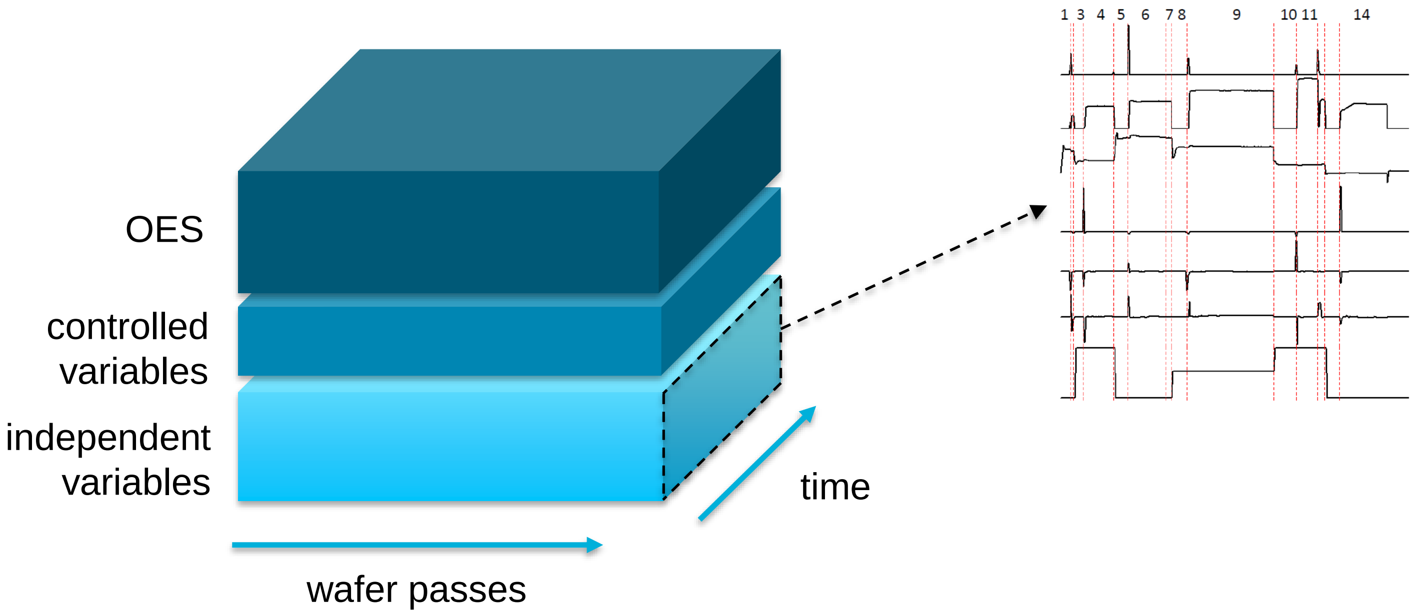






Tracking plasma etch process variations using Principal Component Analysis of OES data. Ma, B.; McLoone, Seán; Ringwood, J. 2007. International Conference on Informatics in Control, Automation and Robotics (ICINCO 2007), Angers, France.

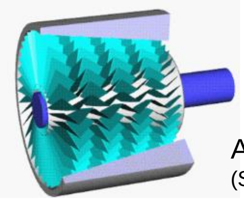
Non-standard setting is everywhere: Etching trace data take a form of higher-order tensor



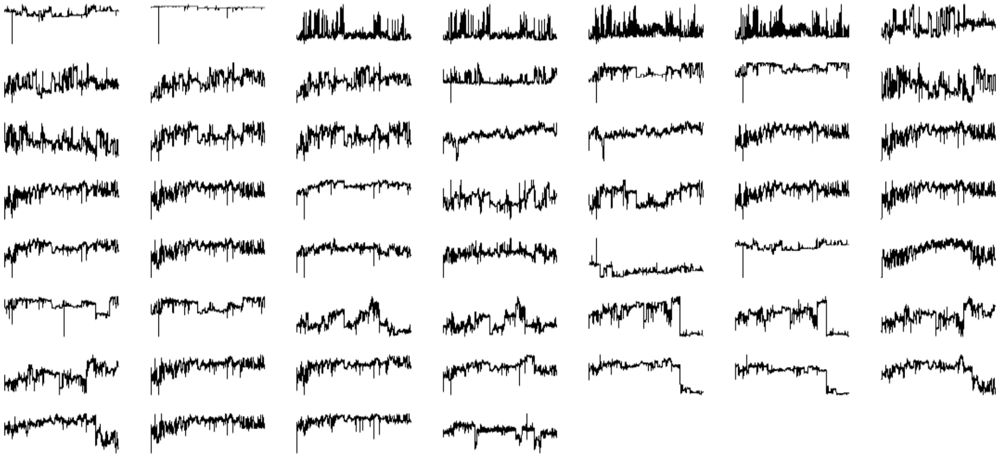
Who knows the ground truth? Who provides labeled data?

- Example: sensor data of a compressor of oil production system
  - Data taken under a normal operational condition
  - Noisy, nonstationary, heterogeneous, high-dimensional …
- Hard to recognize useful patterns by human eye
  - Hard even to experienced engineers

(Image: Wikimedia commons)



Axial compressor
(Source: Wikipedia)

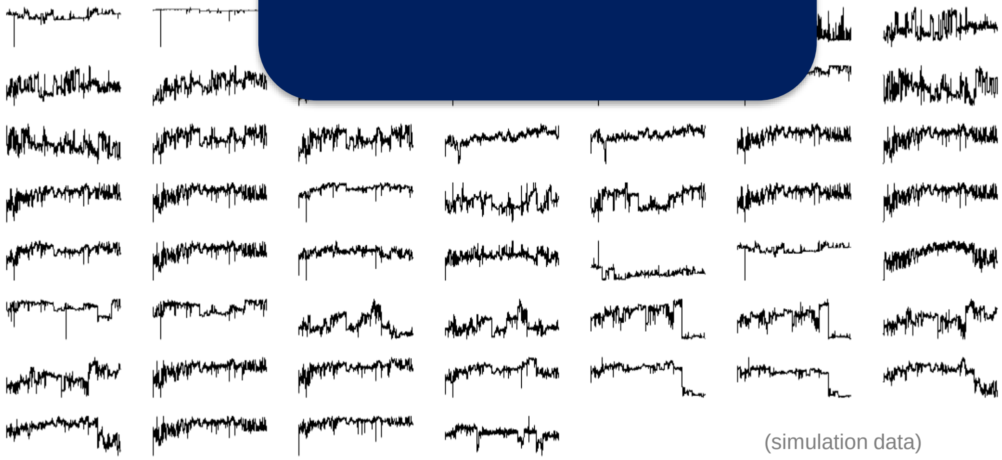


Help me to figure out what’s going on

Very important for sensor data that is NOT human-readable

- Hard to recognize useful patterns by human eye
  - Hard even to experienced engineers
- We really need a good UI to better understand data
  - Easily navigate
  - Provide insights unexpected
- AI should work with UI

User interface technology



(simulation data)

Deep learning. The end of journey? Probably not.
Factors that make deep learning work

Well-defined and well-accepted task

No need to tell why

Huge amount of labeled training data

Typically needs millions labeled samples

Minimum uncertainty in data representation

Pixels, words, log-Mel-filterbank

- Good applications meeting these criteria
  - Image recognition
  - Speech recognition
  - (Some of) natural language processing
- How about industrial dynamic systems?
  - Interesting research topic

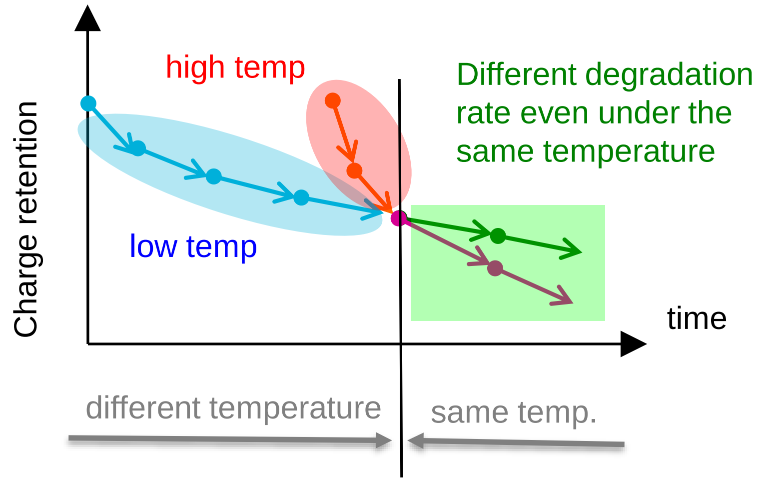
General challenges towards cognitive manufacturing

- AI solutions are diverse. Identifying the right algorithm is challenging
- Often need to develop completely new algorithms
- Real industrial problems are full of unsolved machine learning tasks
- Good UI is needed. Available software tools are far from perfect
- Significant work is needed to make deep learning work for noisy sensor data

Contents

Monitoring battery health of electric vehicles

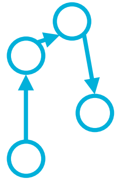
- Battery life measured by charge retention has a strong dependency on the entire usage history
- Time-series (state-space) modeling is challenging due to lack of enough amount of samples
- Developed machine learning algorithm to efficiently predict charge retention
- The core algorithm has been integrated into a “battery traceability system” of a major auto manufacturer



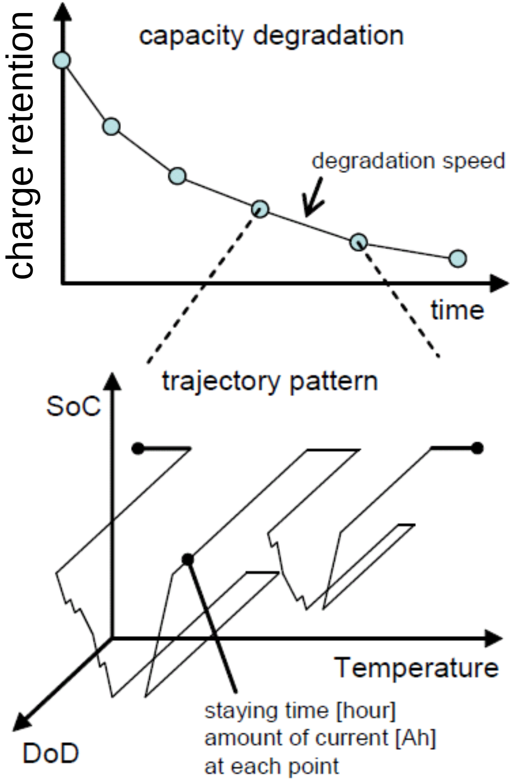
Formalized the task as "trajectory regression"

- Usage history of a battery can be represented as a "trajectory" in a multidimensional space
  - Right: 3-dimensional space spanned by SOC, temperature, and another variable
- The task is to find the function that relates the charge retention (y) with the trajectory (x)

Trajectory regression

charge retention $y = f($  $)$

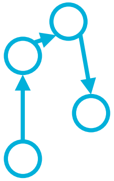
"trajectory"



Technical challenge: Making trajectories comparable

- Simple k-NN prediction doesn't work
  - The length and shape of trajectories vary a lot
  - Unlikely to find a trajectory very similar to a query trajectory
- How can we formalize the notion of "partially similar"?

Partial similarity can be captured by dual (kernel-based) and primal (feature-based) formulations

$$y = f\Big($$

$$\Big)$$

Kernel-based [Ide&Kato SDM 09]

- Represent a trajectory as a symbol sequence
- Use a string kernel [Leslie 02]

Feature-based [Ide&Sugiyama AAAI 11]

Objective function to be minimized

$$\Psi(\boldsymbol{f}|\lambda) = \sum_{n=1}^{N}\left(y^{(n)} - \sum_{e \in \boldsymbol{x}^{(n)}} c_e(f_e)\right)^2 + \frac{\lambda}{2}\sum_{e=1}^{M}\sum_{e'=1}^{M} S_{e,e'}|f_e - f_{e'}|^2$$

“Predicted cost should be close to observed values”

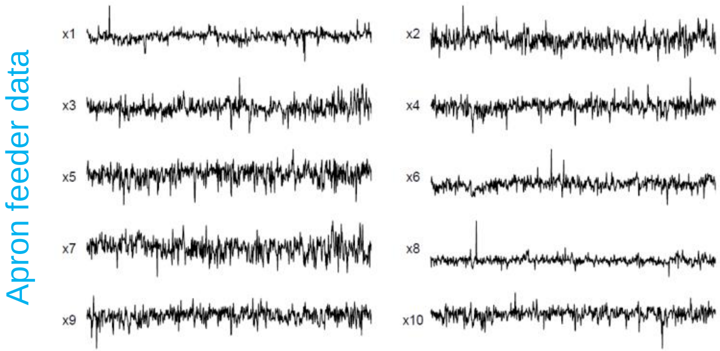
- “Neighboring links should take similar values”
- ***S*** is the similarity matrix between links

Contents

Continuous operation of conveyor systems is critical in the mining industry

- **Business goal**: Ensure continuous operation of conveyor system by detecting early indications of failures
- **Data**: Physical sensor data from conveyors and motors
  - Every several seconds over ~ 1 year
  - Sensors include: Gearbox temperatures, motor power consumptions, apron speed, etc.
- **Challenge**: Conveyor system is subject to significant fluctuation in load. Hard to characterize the normal operation
  - Crude ore conveyed never be uniform

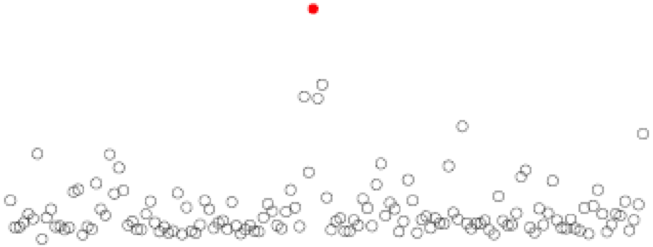

(Image: Wikimedia commons)
(Image: Wikimedia commons)

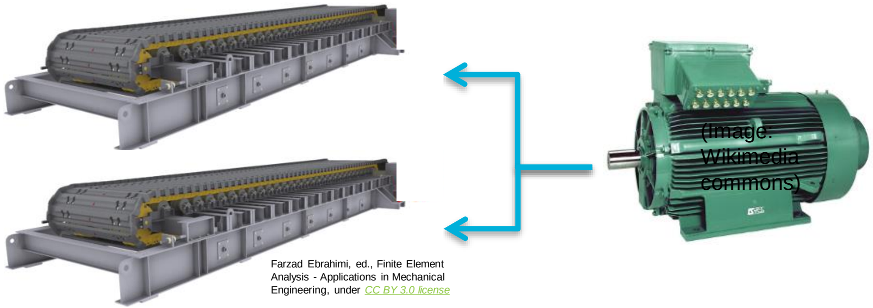


(simulation data)

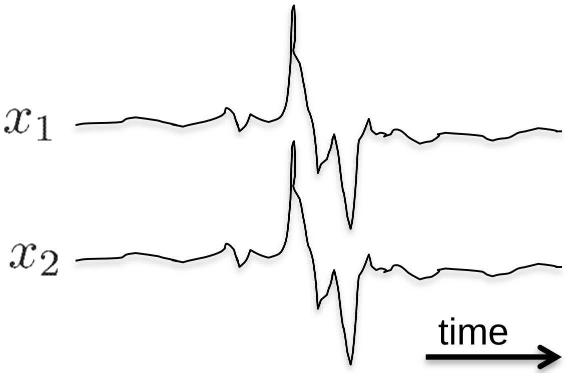
Tackling two different types of noise to achieve practical robustness

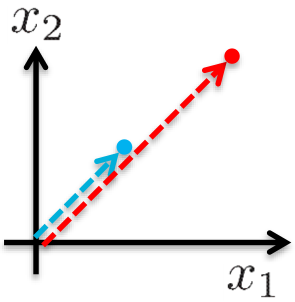
- Independent impulse noise
  - Treated as a contaminated sample





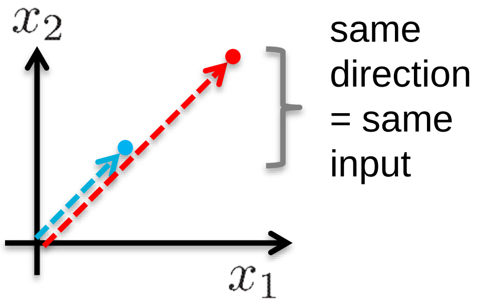
- Multiplicative noise equally applied to correlated variables
  - Commonly observed in conveyor driven by the same motor

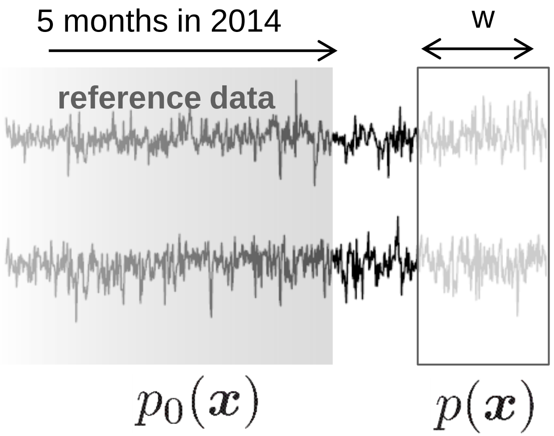




Formalizing as change detection problem for directional data

- To filter out multiplicative noise, we use only the information of the **direction** of ***x***
- Change score is computed as the discrepancy between $p(\mathbf{x})$ and $p_0(\mathbf{x})$
  - $p(\boldsymbol{x})$: Probability density in the window
  - $p_0(\boldsymbol{x})$: Probability density in the reference data
  - $\boldsymbol{x}$: observation (2-dimensional vector in this example)
- Independent impulse noise is handled using the L_1 and L_2 regularization technique





(For ref.) Developed new feature extraction and scoring algorithm based on directional statistics

- Step 1: Solve regularized weighted maximum likelihood for von Mieses-Fisher distribution
  - vMF distribution $$p_0(\boldsymbol{x}) = \frac{\kappa^{M/2-1}}{(2\pi)^{M/2} I_{M/2-1}(\kappa)} \exp\left(\kappa \boldsymbol{u}^\top \boldsymbol{z}\right)$$
  - Optimization problem

$$\{\boldsymbol{u}_i^*, \boldsymbol{w}_i^*\} = \arg\max_{\{\boldsymbol{u}_i, \boldsymbol{w}_i\}} \left\{ \underbrace{\sum_{n=1}^{M} b^{(n)} w_i^{(n)} \ln \mathcal{M}(\boldsymbol{z}^{(n)} | \boldsymbol{u}_i, \kappa)}_{\text{weighted maximum likelihood}} + \underbrace{\sum_{i=1}^{m} \left(\frac{1}{2} \|\boldsymbol{w}_i\|_2^2 + \nu \|\boldsymbol{w}_i\|_1 \right)}_{\text{regularization term}} \right\}$$

$$\text{subject to} \quad \boldsymbol{u}_i^\top \boldsymbol{u}_j = \delta_{i,j} \quad (i, j = 1, \dots, m)$$

- Step 2: Compute the change score as parameterized Kullback-Leibler divergence between the reference and current distributions

$$a^{(t)} = \min_{\boldsymbol{f}, \boldsymbol{g}} \int d\boldsymbol{x}\, \mathcal{M}(\boldsymbol{x} | \mathsf{U}\boldsymbol{f}, \kappa) \ln \frac{\mathcal{M}(\boldsymbol{x} | \mathsf{U}\boldsymbol{f}, \kappa)}{\mathcal{M}(\boldsymbol{x} | \mathsf{U}^{(t)}\boldsymbol{g}, \kappa)}$$

→ computed via singular value decomposition

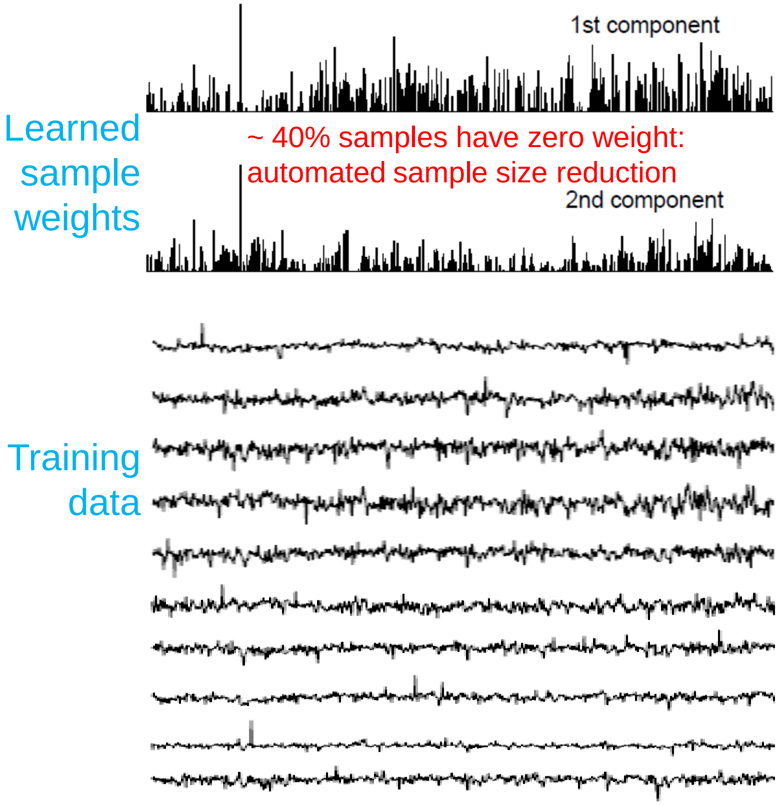
$\mathsf{U} \equiv [\boldsymbol{u}_1^*, \dots, \boldsymbol{u}_m^*]$ for the training data

$\mathsf{U}^{(t)}$ for the window at t

For the detail, see T. Ide et al., “Change Detection Using Directional Statistics”, IJCAI 16.

New algorithm achieves both dimensionality and sample size reduction at once

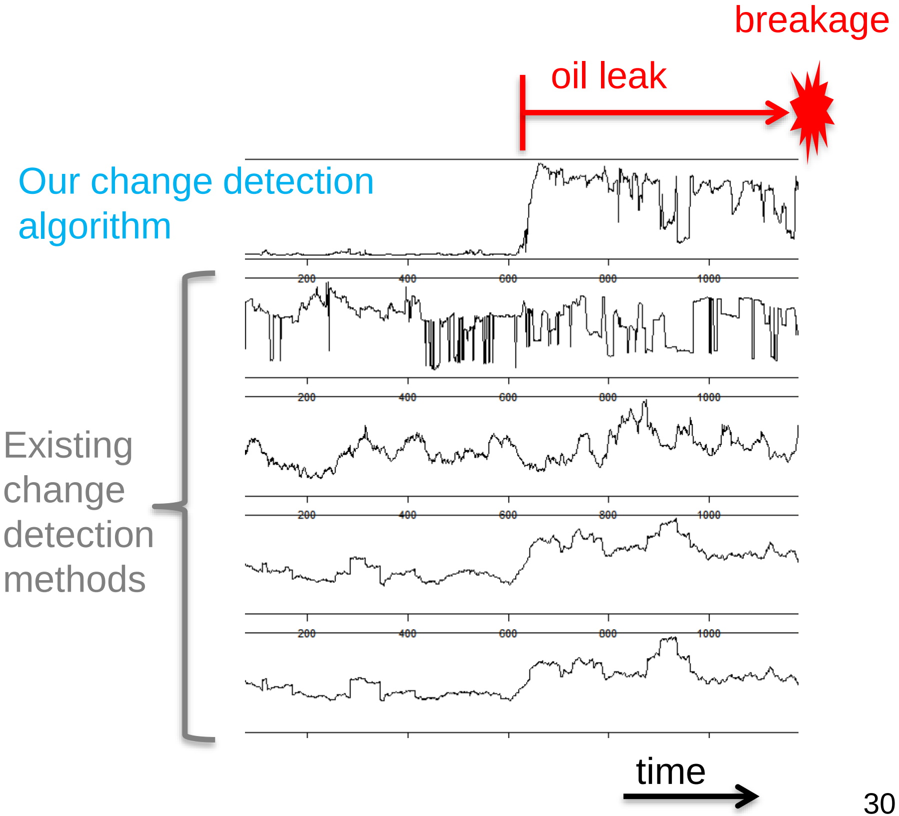
- Random outliers due to noise are automatically removed
  - Automated sample size reduction
- Major directions are automatically found
- The algorithm is (almost) guaranteed to produce a globally optimal solution
  - It is reduced to the convex “trust-region subproblem” in a certain limit



(simulation data)

Detected failure example: Detecting bearing failure in ore transfer system

- Two apron feeders are operated in parallel in this ore transfer system
- A bearing failure due to lube oil leak started showing asymmetric behavior in a few variables across the two conveyors at about *t*=600
  - Variables related to power
- Our change detection method clearly detected the failure
  - Existing methods fail to catch

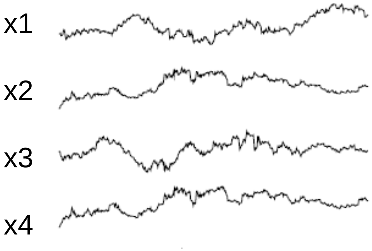


Contents

Automated condition-based monitoring is known to be a hard problem in the maritime industry

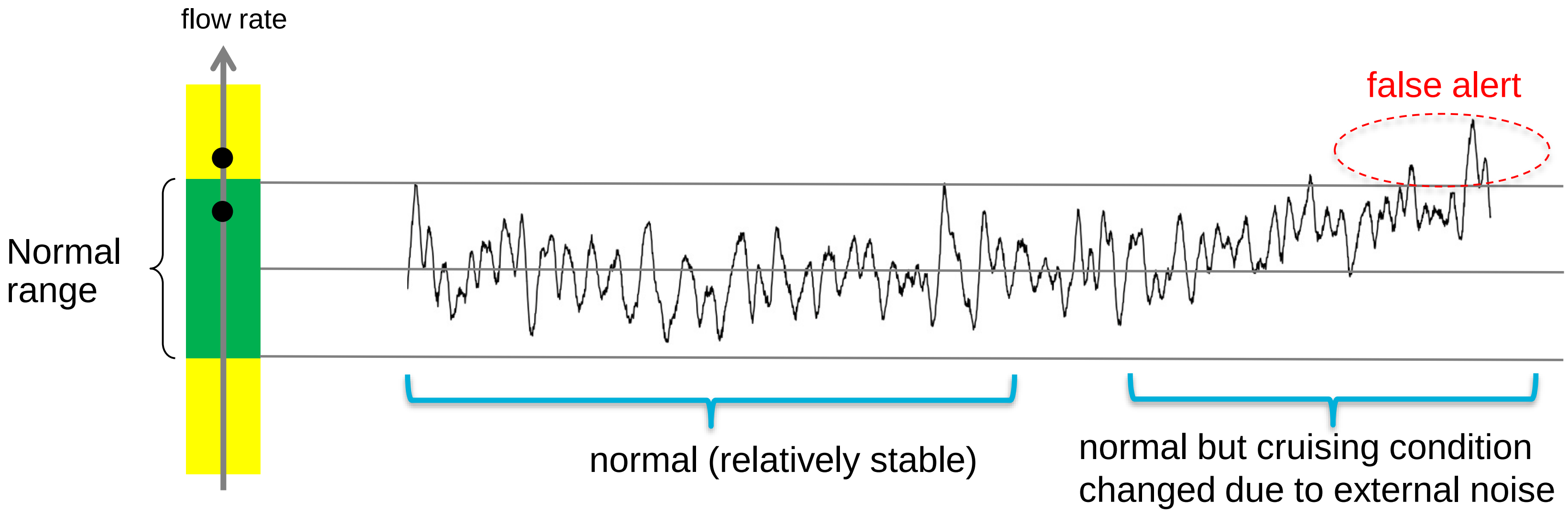
- Preventing failures in main engine is of critical importance for ocean-going vessels
- Many attempts have been made for automated condition-based monitoring, but few succeeded
- Major reason: unpredictable external noise
  - Sea current, waves, weather, wind, etc.
  - Extremely hard to build normal state model


(Image: Wikimedia commons)



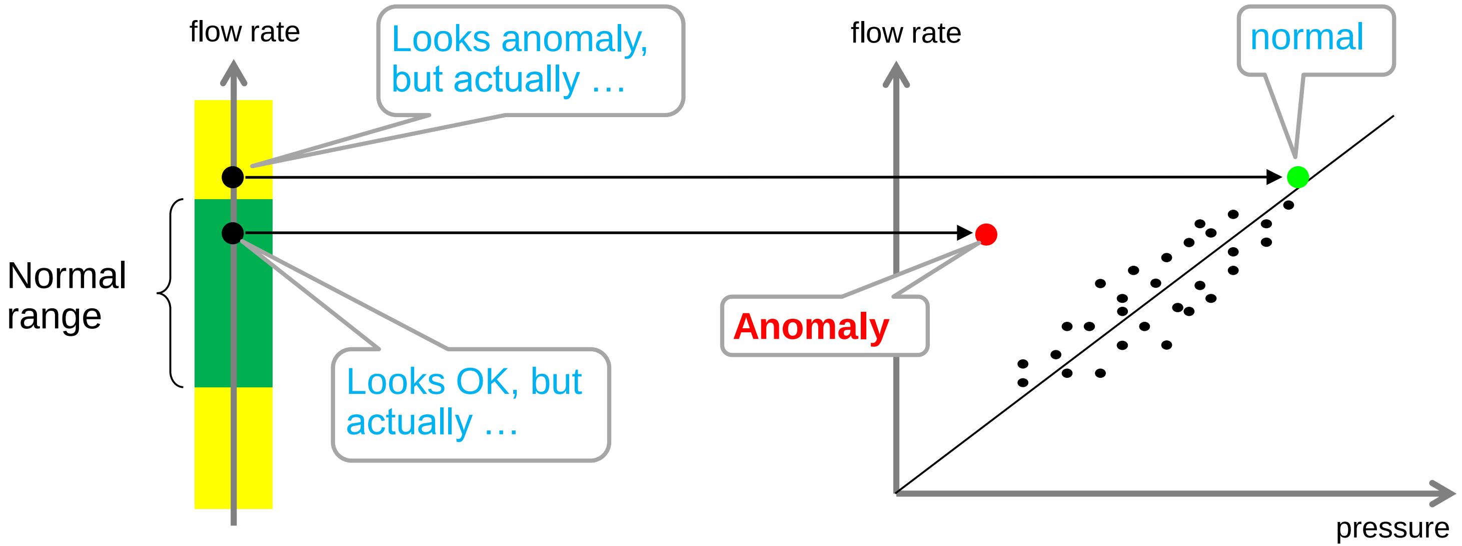
Conventional limit-check approach is of limited use under dynamic unpredictable noise

- Under dynamic noise, monitoring measurement values themselves leads to many false alerts

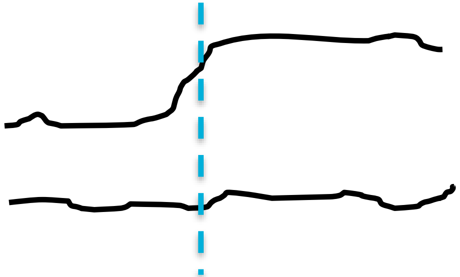


Dependency-based view is useful to remove false alerts of conventional limit-check approach

- Even measurement values are dynamically changing, dependency can be stable in many mechanical systems

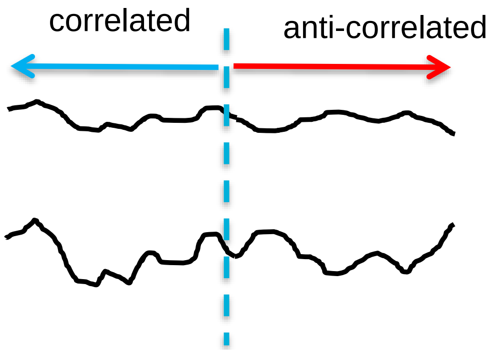


Many nontrivial anomaly detection problems are related to dependency anomalies



Change in the mean

- Easy to detect and quantify
- Classical methods are available

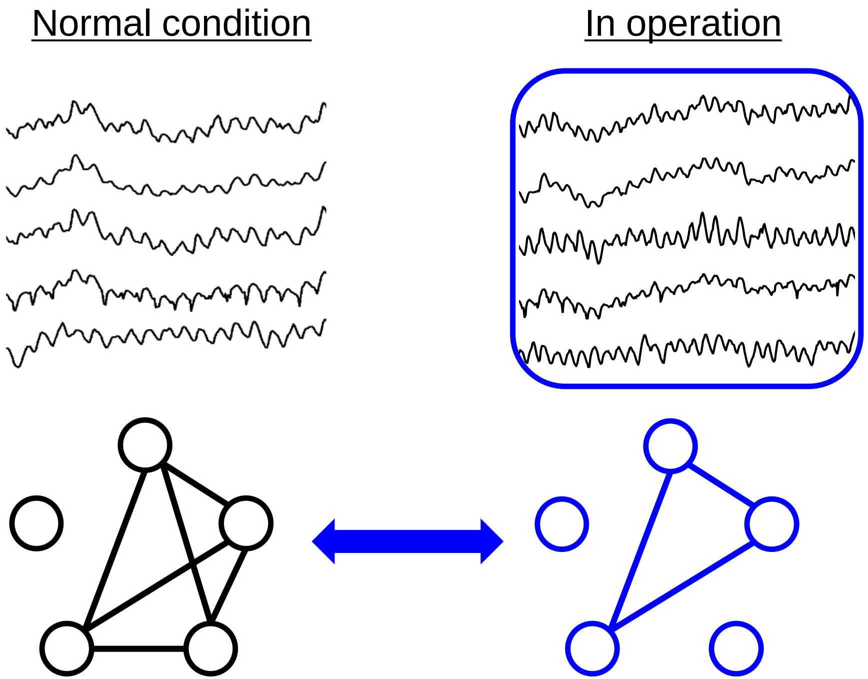


Change in the dependency

- Hard to detect and quantify manually
- Conventional methods cannot handle
- Important in practice

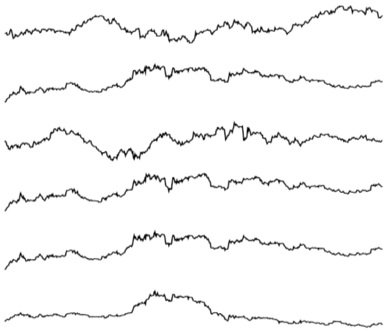
Taking advantage of dependency graph for anomaly detection

- Example: Data set comparison
  - Learn dependency model under the normal condition
  - In operation, check if the dependency significantly changes
- How can we find a precise dependency structure from data?

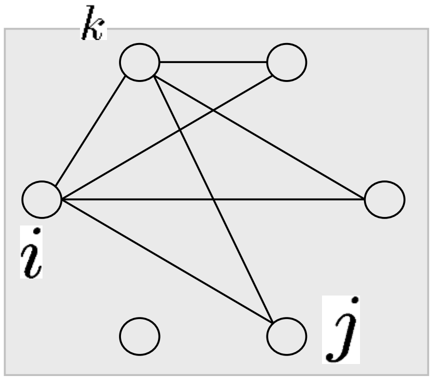


Two major technical problems addressed

- Sparse structure learning
  - How to accurately learn the dependency under heavy noise
- Anomaly scoring
  - How to compute the anomaly score of *individual* variables



Dependency between variables



(For ref.) Algorithm for sparse structure learning

- Assume graphical Gaussian model

$$p(\boldsymbol{x}|\Lambda) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{0}, \Lambda^{-1}) = \frac{\det(\Lambda)^{1/2}}{(2\pi)^{M/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \Lambda \boldsymbol{x}\right)$$

- Put a Laplace prior on Lambda

$$p(\Lambda) = \prod_{i,j=1}^{M} \frac{\rho}{2} \exp\left(-\rho|\Lambda_{i,j}|\right)$$

rho: constant controlling the strength of prior

- MAP (Maximum a posteriori) estimation for Lambda

$$\Lambda^* = \arg\max_{\Lambda} \left\{ \ln p(\Lambda) \prod_{n=1}^{N} p(\boldsymbol{x}^{(n)}|\Lambda) \right\}$$

$$= \arg\max_{\Lambda} \{ \ln\det\Lambda - \mathrm{tr}(S\Lambda) - \rho||\Lambda||_1 \}$$

S: sample covariance matrix

For the detail, see, T. Ide et al., "Proximity-Based Anomaly Detection using Sparse Structure Learning," Proc. SIAM Intl Conf. on Data Mining 2009 (SDM 09).

(For ref.) Anomaly scoring algorithm (for outlier analysis)

- Define the outlier score for the *i*-th variable as

$$\mathrm{score}_i(\boldsymbol{x}|\Lambda) \equiv -\ln p(x_i|x_1, .., x_{i-1}, x_{i+1}, ..., x_M, \Lambda)$$

  - Lambda represents a sparse structure
  - *p* is p.d.f. defined by the graphical Gaussian model

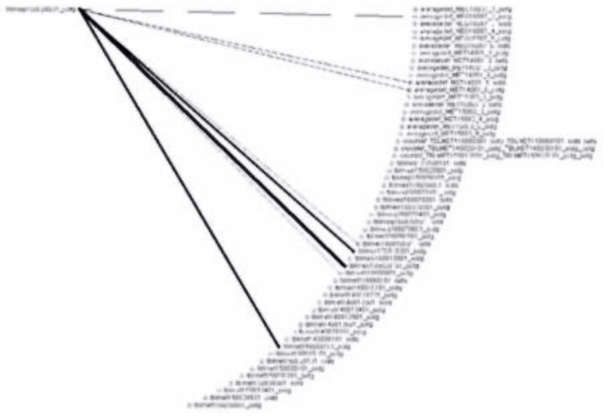
- Final result: Anomaly score of the i-th variable
  - Only variables connected to the i-th variable play a role

$$\mathrm{score}_i(\boldsymbol{x}|\Lambda) = \frac{1}{2}\ln\frac{2\pi}{\Lambda_{i,i}} + \frac{1}{2\Lambda_{i,i}}\left(\sum_{j=1}^{M}\Lambda_{i,j}\,x_j\right)^2$$

Dependency-based anomaly detection provides deeper insights through dependency discovery

- Algorithm was tested using real data from vessels
  - Data: VLCCs and bulk carriers
  - Model construction is done automatically
  - Confirmed better detection accuracy than conventional methods

- Dependency graph provides useful insights for diagnosis



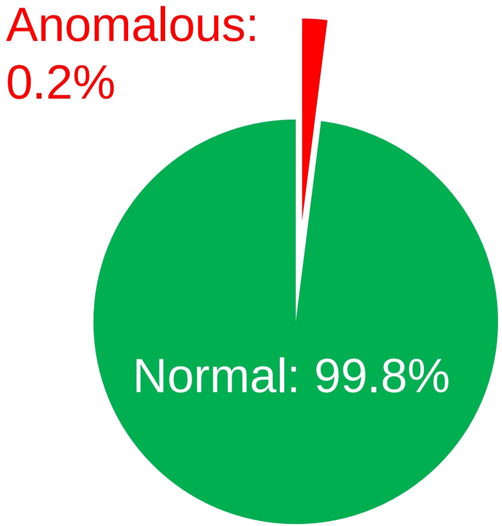
(Image: Wikimedia commons)

Contents

Developing a multi-task learning framework for fleet-level condition-based asset management

- You have many similar but not identical industrial assets
- Management costs can be prohibitive if individual assets are managed independently
- Our framework allows sharing knowledge across different assets

Integrated monitoring tool will allow sharing anomaly data across different assets



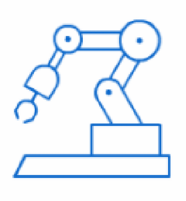
- In condition-based monitoring, big data may not be really big
  - Anomalous samples account for less than 0.2% in a metal smelting process
- Coverage of anomalies and thus accuracy can be limited due to lack of data

Technical challenge:
Multi-modality, heavy noise, interpretability

System 1
(in New Orleans)

⋮

System *s*



⋮

System *S*
(in New York)

- Straightforward solutions have serious limitations
  - 1. Treat the systems separately. Create each model individually
    - ✓ Suffers from lack of fault examples
  - 2. Build one universal model by disregarding individuality
    - ✓ Model fit is not good
- Practical requirements in IoT-related industries
  - Capture both individuality and commonality
  - Automatically capture multiple operational states
    - ✓ Without specifying e.g. # of patterns
  - Be robust to noise
  - Be highly interpretable for diagnosis purposes

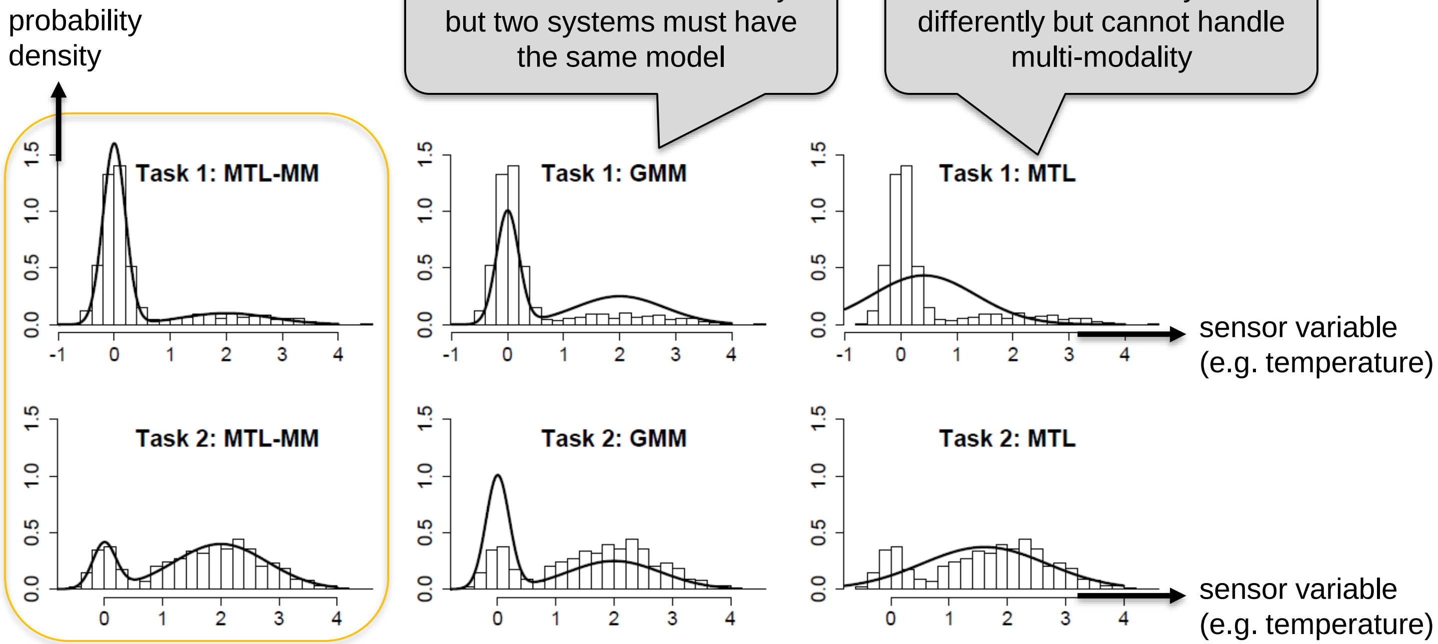
Existing multi-task learning methods cannot handle multi-modality

Can handle multi-modality but two systems must have the same model

Can treat different systems differently but cannot handle multi-modality

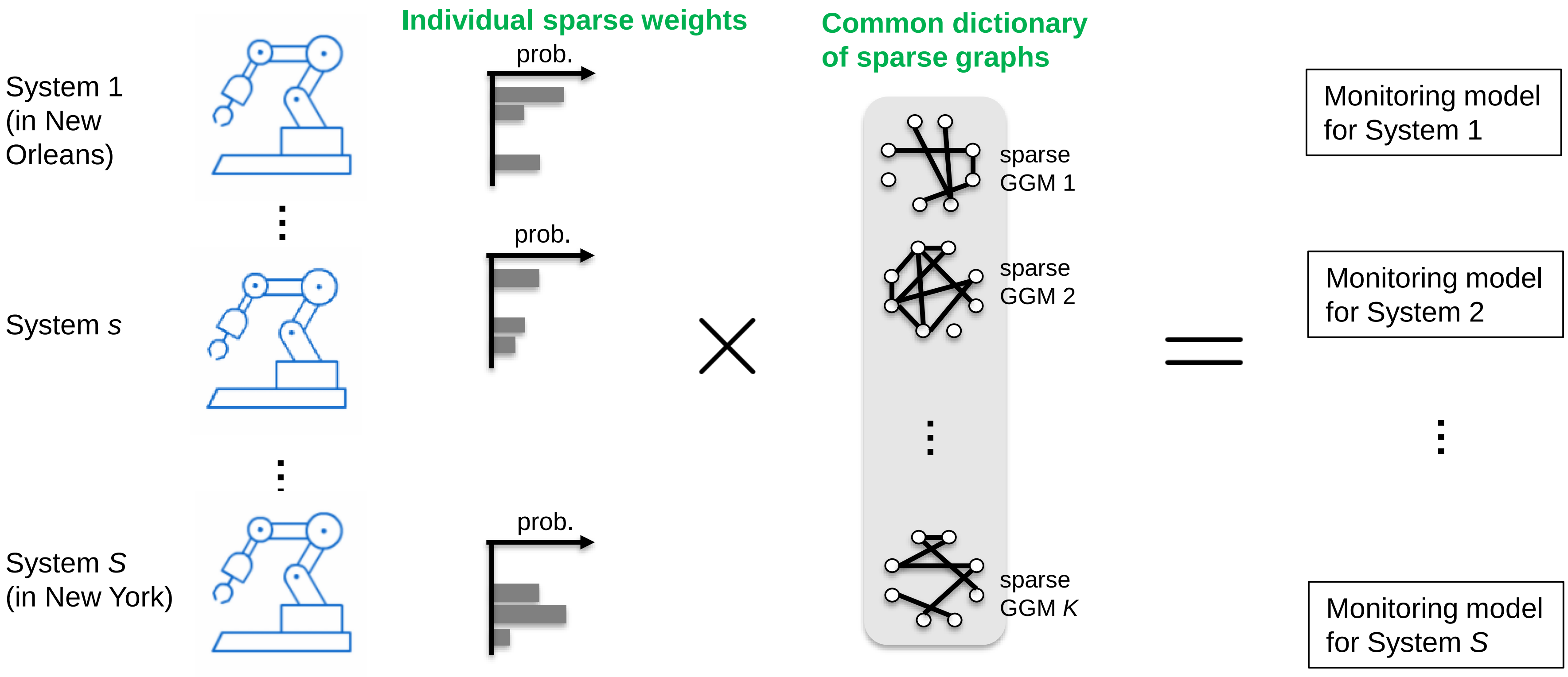
System (task) 1
(in New Orleans)

System (task) 2
(in New York)



Comparing the proposed multi-task multi-modal (MTL-MM) model with standard Gaussian mixture (GMM) and multi-task learning (MTL) models

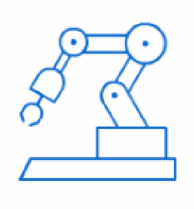
Developed a doubly sparse model representing individuality and commonality of the systems in a fleet

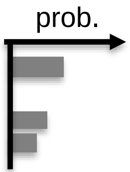


GGM=Gaussian Graphical Model

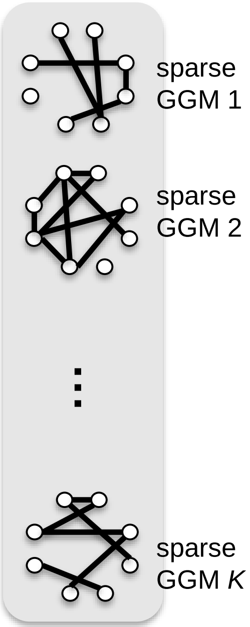
Monitoring model for each asset is represented as a Gaussian mixture model

System *s*





$\times$



GGM=Gaussian Graphical Model

Monitoring model for System s

Gaussian mixture

$$= \sum_{k=1}^{K} \pi_k^s \, \mathcal{N}(\boldsymbol{x}^s \mid \boldsymbol{\mu}^k, (\Lambda^k)^{-1})$$

Sparse mixture weights

Sparse Gaussian graphical model

Overview of probabilistic model

- Observation model
  - Gaussian mixture with task-dependent weight

$$\prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}^s \mid \boldsymbol{\mu}^k, (\Lambda^k)^{-1})^{z_k^s}$$

- Sparsity enforcing priors
  - Laplace prior for the precision matrix
  - Bernoulli prior for the mixture weights

$$p(\Lambda^k) = \left(\frac{\rho}{4}\right)^{M^2} \exp\left(-\frac{\rho}{2}\|\Lambda^k\|_1\right)$$

$$p(\boldsymbol{\pi}) = p_0^{\|\boldsymbol{\pi}\|_0}(1 - p_0)^{G-\|\boldsymbol{\pi}\|_0}$$

- Inference
  - Variational Bayes + convex point estimation

Ide et al., "Multi-task Multi-modal Models for Collective Anomaly Detection", IEEE Intl Conf. on Data Mining 2017 (ICDM 17).

Inference algorithm: Use standard VB framework incorporating two convex optimization problems

- Update sample weights
- Update cluster weights
- Update precision matrices
- Update other parameters

Use new semi-closed form solution

$$\max_{\boldsymbol{\pi}^s} \left\{ \sum_{k=1}^{K} c_k^s \ln \pi_k^s - \tau \|\boldsymbol{\pi}^s\|_0 \right\} \quad \text{s.t.} \quad \|\boldsymbol{\pi}^s\|_1 = 1.$$

Solved by graphical lasso [Friedman 08]

$$\max_{\Lambda^k} \left\{ \ln |\Lambda^k| - \mathrm{Tr}(\Lambda^k \mathsf{Q}^k) - \frac{\rho}{N_k} \|\Lambda^k\|_1 \right\}$$

Ide et al., “Multi-task Multi-modal Models for Collective Anomaly Detection”, IEEE Intl Conf. on Data Mining 2017 (ICDM 17).

Solving the L0-regularized optimization problem for mixture weights

- What is the problem of the conventional approach?
  - Simply differentiate w.r.t. π_k^s
  - Claims to get a sparse solution [Corduneanu+ 01]
  - But mathematically π_k^s cannot be zero due to logarithm

$$\max_{\boldsymbol{\pi}^s} \left\{ \sum_{k=1}^{K} c_k^s \ln \pi_k^s \right\}$$

$$\text{s.t.} \quad \|\boldsymbol{\pi}^s\|_1 = 1.$$

- We re-formalized the problem as a convex mixed-integer programming

- We derived a semi-closed form solution

Ide et al., "Multi-task Multi-modal Models for Collective Anomaly Detection", IEEE Intl Conf. on Data Mining 2017 (ICDM 17).

Comparison with possible alternatives

		Interpretability	Noise reduction	Fleet-readiness	Multi-modality
Our work [Ide et al. ICDM 17]		**Yes**	**Yes**	**Yes**	**Yes**
(single) sparse GGM	[Ide et al. SDM 2009, Ide et al. ICDM 2016]	**Yes**	**Yes**	No	No
Gaussian mixtures	[Yamanishi et al., 2000; Zhang and Fung, 2013; Gao et al., 2016]	Limited	Limited	No	**Yes**
Multi-task sparse GGM	[Varoquaux et al., 2010; Honorio and Samaras, 2010; Chiquet et al., 2011; Danaher et al., 2014; Gao et al., 2016; Peterson et al., 2015].	**Yes**	**Yes**	**Yes**	No
Multi-task learning anomaly detection	[Bahadori et al., 2011; He et al., 2014; Xiao et al., 2015]	No	(depends)	**Yes**	No

Experimental results

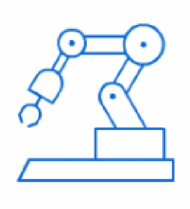
- → See my paper
  - Ide et al., “Multi-task Multi-modal Models for Collective Anomaly Detection”, IEEE Intl Conf. on Data Mining 2017 (ICDM 17).
    - ✓ Available at http://ide-research.net/

Use-case example: Simultaneous monitoring based on on-line computation of anomaly scores

System 1
(in New Orleans)

⋮

System *s*

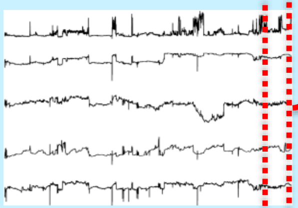


⋮

System *S*
(in New York)

In operation

sensor time-series



window at present

anomaly score
- overall
- variable-wise

⋮

Model construction

- **Doubly sparse mixture model**
- **Automated method to learn the doubly sparse model**

Contents

Summary

- Manufacturing industries are full of unsolved machine learning problems

- Need to go beyond conventional settings
  - Multi-X setting (X = task, modal, view, etc.)
  - Sparsification for better interpretability

- The importance of smart UI / visualization cannot be overemphasized

- Note: rare to encounter peta/exa-scale data in practice
  - A lot of things to do before thinking about distributed, parallelized, and streaming computing settings

Discussion: What is the potential of deep learning in cognitive manufacturing? Image, text, and acoustics

- Image-based analysis can be safely replaced with a DL-based solution
  - If you have a good amount of labeled data

- Text data is tricky
  - Most of maintenance logs, monthly reports, emails, contractual documents are not appropriate for DL-based text analysis
    - ✓ (and conventional text mining methods, either)
  - Mainly due to lack of enough amount of data

- Acoustic data is tricky
  - Sound-based inspection is common in some domains, but DL-based approach may not be very straightforward
  - Mainly due to lack of established preprocess and language models

- Time-series modeling

Thank you!