IBM **Research**

Towards consumable analytics: Challenges and recent advances

T. Ide (Ide-san)
IBM T. J. Watson Research Center

The 2015 IEEE ICDM Workshop on Data Mining for Services

Contents

The hype curve

Gartner's 2015 Hype Cycle for Emerging Technologies

- Machine learning (data mining) is at the peak

- Is it really ready for real business?

Success of deep learning gives rise to much optimism over AI technologies

Speech recognition

Text analysis

Image recognition

[Abdel-Hamid et. al, 2014]

Factors that make deep learning work

The task is well-defined and well-accepted.

Results are easily verified by humans

Huge amount of labeled training data is available

Very rare in practice

Trivial atomic representation is known

(next page)

Example. Comparison between text and sensor data. Ambiguity in anomic representation makes representation learning challenging

Sensor data

- No obvious atomic representation
- Pre-process is mostly problem-dependent; general-purpose tools do not help

What if sensors have different time resolutions?

What if there are missing observations?

Should we use low-pass filter for noise removal?

Text data

"President Obama announced that he had rejected the request from a Canadian company to build the Keystone XL oil pipeline"

- Obvious atomic representation ("president," "announce," etc.)
- Well-established preprocess (stemming, PoS tagging, etc.)
- Clearly-defined task pipeline (UIMA)

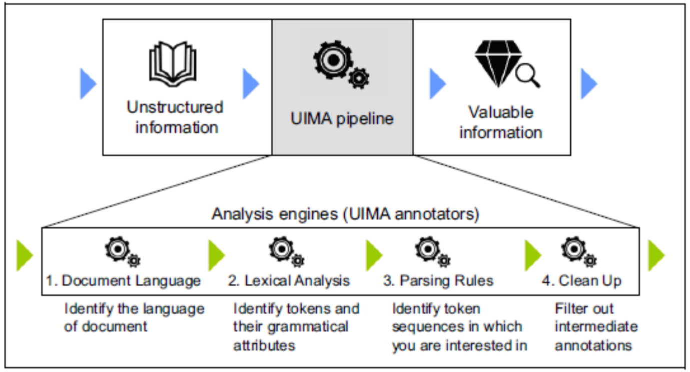


Figure 11-1 UIMA pipeline

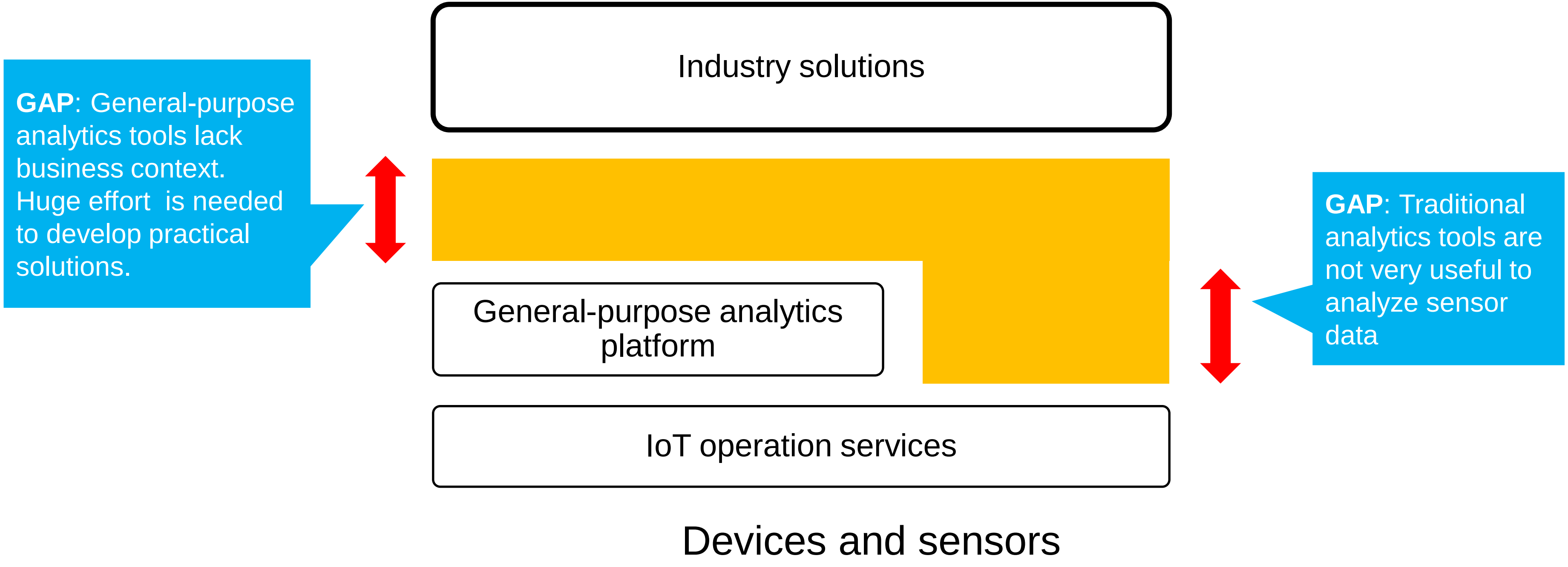
Two areas major gaps exist towards consumable analytics

- Sensor data analytics
  - Interpretability matters
    - ✓ Especially in critical applications such as anomaly detection of manufacturing plants
  - No clearly-defined atomic representation
    - ✓ Time resolution? Low-pass filtering? Fourier domain?

- Human behavior modeling
  - Interpretability matters
    - ✓ e.g. marketing, employee evaluation
  - No clearly-defined atomic representation
    - ✓ What quantity we should use as the feature?

Contents

Observation: Two gaps in IoT technology stack



SROM (Smarter Resource and Operations Management): Solution-oriented analytics library

Data

- Asset/Equipment Attributes
- Failure/Repair History
- Operations Data
- Process Data
- Product Attributes
- Environmental Data (e.g., weather)
- Sensors / Devices
- Meters
- Grid Energy Price (TOD)
- Resources
- Costs / Budget
- …

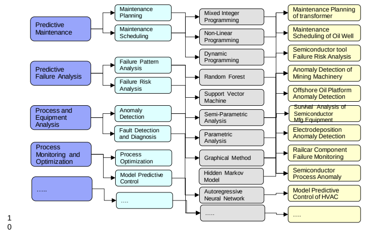
Analytics Solution Library

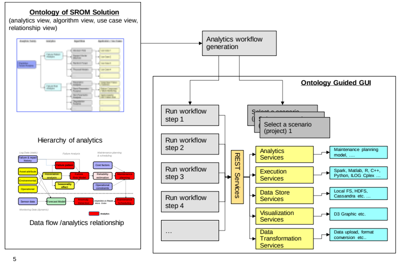
- Maintenance Planning
- Maintenance Scheduling
- Failure Risk Analysis of Assets
- Failure Pattern Analysis of Assets
- Anomaly Detection
- Fault Detection and Diagnosis
- Process Optimization
- Model Predictive Control
- ….

State-of-the-art algorithms (Key Differentiator)

- Mixed Integer Programming
- Nonlinear Programming
- Dynamic Programming
- Random Forest
- Support Vector Machine
- Neural Network
- Parametric Analysis
- Semi-Parametric Analysis
- Graphical Methods
- Cohort Analysis
- Support Vector Machine
- Hidden Markov Model
- ….

Ontology-Guided Analytics Workflow and Reusable APIs





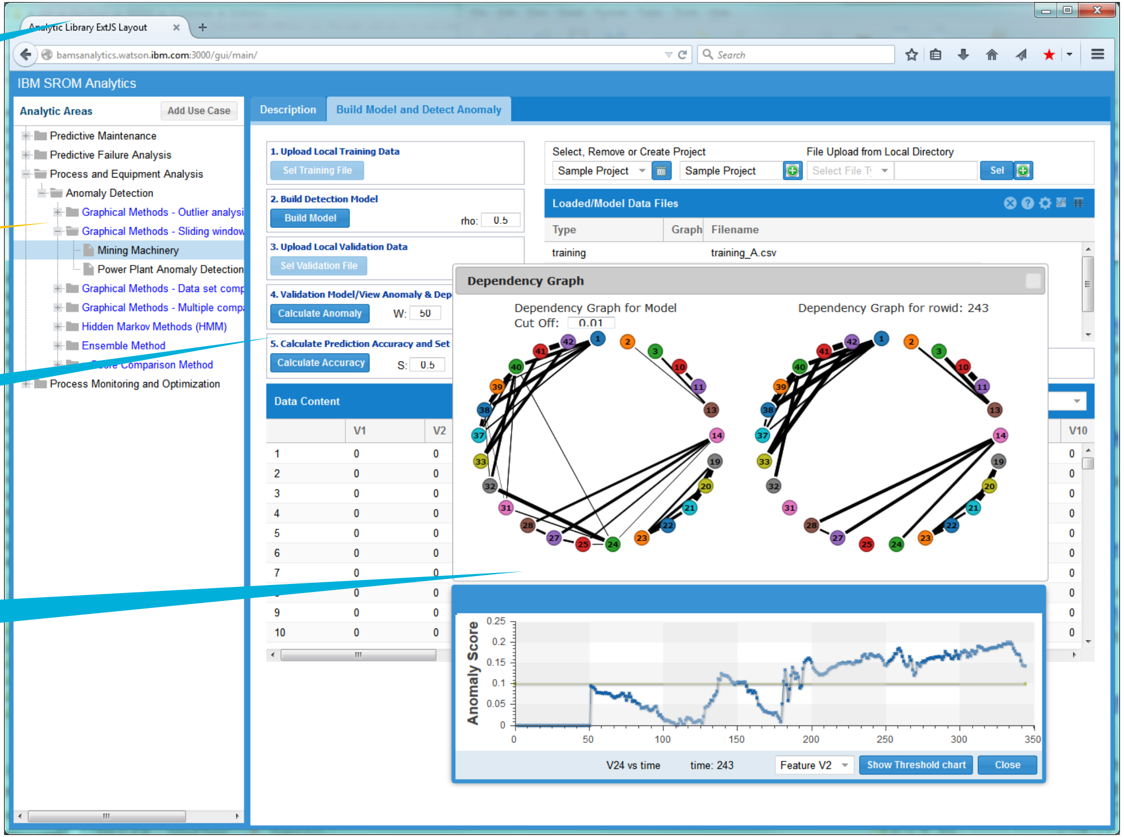
SROM solution library at-a-glance

Web-based system allowing cloud-based offerings

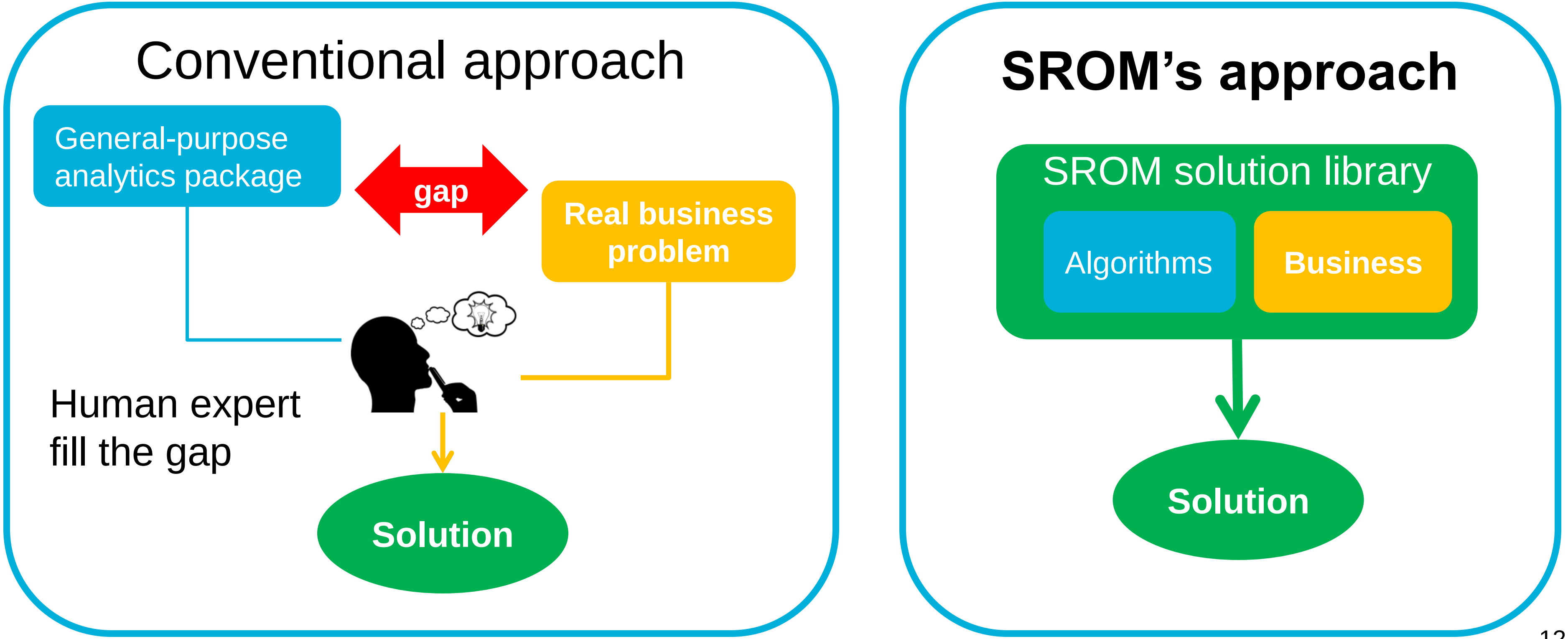
World's most comprehensive solution library

Easy-to-follow analysis steps that reproduces experts' workflow

State-of-the-art algorithms beyond standard packages



SROM reduces time-to-business in analytics solution development by integrating real business use-cases



Solution-oriented architecture is the next generation design principle of analytics libraries

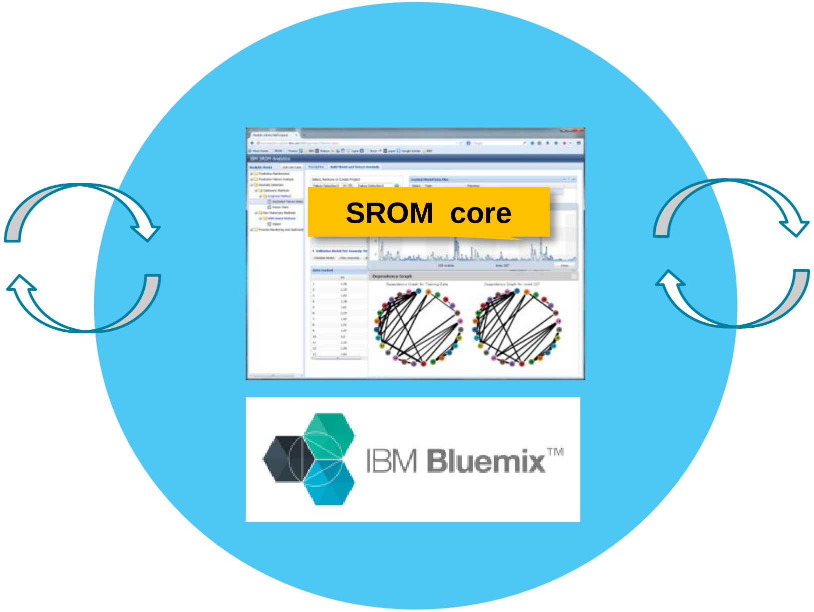
Fancy GUI does not solve the issue

1st gen	2nd gen	**3rd gen**
Bare statistics / ML library S, R, Weka, etc.	ML library armed with composable GUI + enterprise system SAS, SPSS, etc.	Cloud-based Solution-oriented Function based-APIs **SROM**

Key technology ingredients of SROM

- Implementation
  - Live on cloud
  - Well-defined workflow
- Algorithm
  - Equipped with state-of-the-art algorithms developed by IBM researchers
  - Covers entire areas in asset monitoring and management

Ecosystem to enhance the coverage



IBM

- Develop solution using SROM core
- Develop new algorithms if needed
- Provide the solution as a service on cloud

Client

- Provide business problems
- Provide data
- Use developed solution approach to provide feedback

Real example of recycling solutions

- Quick prototyping for mining machinery anomaly detection
- Power transformer maintenance scheduling
- Optimal demand response of building HVAC system

Real example of recycling solutions

- Quick prototyping for mining machinery anomaly detection

Quick prototyping for mining machinery anomaly detection

Customer's pain: too many false alerts

Sensor data is highly noisy & dynamic; Traditional methods do not work .

Underground coal mining

Applied a similar usecase of SROM

High accuracy false positive filter in a few weeks

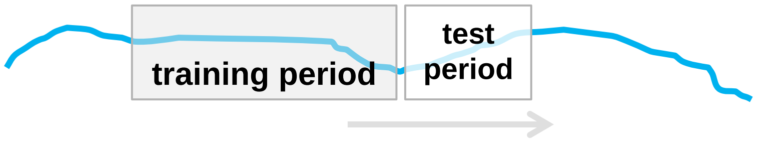
Newly added usecase

General features of sensor data

- Cutter current data of a cutting machine
  - Red: failure episode
  - Blue: 6min prior to failure
- Data is highly dynamic and noisy
  - Traditional statistical approaches do not work

Tackling too many false positives

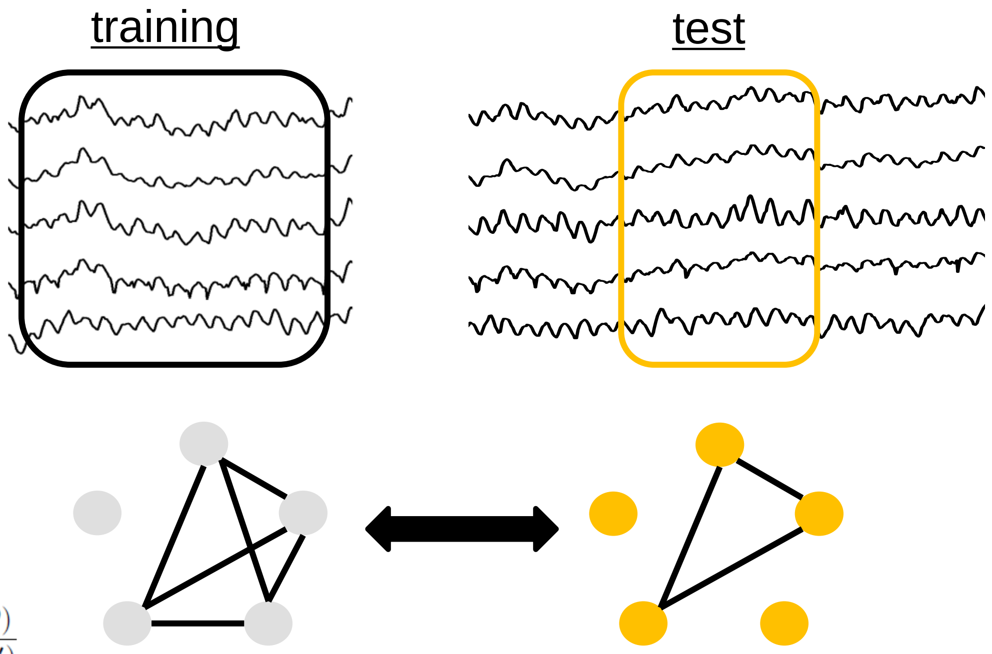
- Formulated the problem as online multivariate change detection
  - Compute distance between previous and present situations
- Classifying false and true positives
  - True positive: Change = yes
    - ✓ if the test window is in the failure episode
  - False positive: Change = no
    - ✓ otherwise
- Multivariate treatment is required



Adopted advanced machine learning algorithm based on probabilistic graphical model

- Compute dependency graph in the training region
- Do the same for the test region
- Compare the graphs in an information-theoretical fashion
  - i-th variable's change score

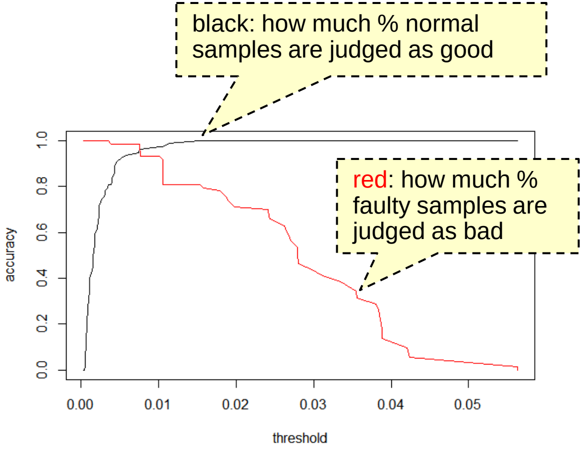
$$a_i \equiv \int \mathrm{d}\boldsymbol{x}_{-i}\, p(\boldsymbol{x}_{-i} \mid \mathcal{D}) \int \mathrm{d}x_i\, p(x_i \mid \boldsymbol{x}_{-i}, \mathcal{D}) \ln \frac{p(x_i \mid \boldsymbol{x}_{-i}, \mathcal{D})}{p(x_i \mid \boldsymbol{x}_{-i}, \mathcal{D}')}$$



[Ide, et al., SDM 09]

Result: Achieved 90+% accuracy in false/true positive classification

- Achieved 90+% prediction accuracy for both faulty and normal sample accuracies
  - ✓ Shown as a function of the detection threshold (→)
- No tuning parameters: The detection model was built fully automatically.
  - ○ c.f. handcrafted rules



Real example of recycling solutions

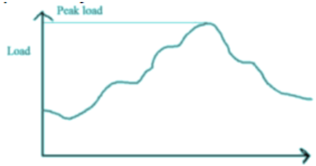
- Power transformer maintenance scheduling

Power transformer maintenance scheduling

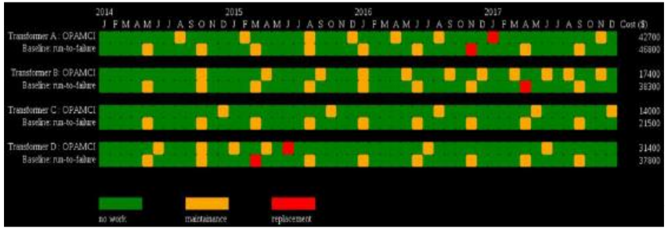
- Solution approach: marriage of physics, analytics, and optimization

basic load time series data

electric network



Electrical simulation → Failure prediction → **Maintenance scheduler**



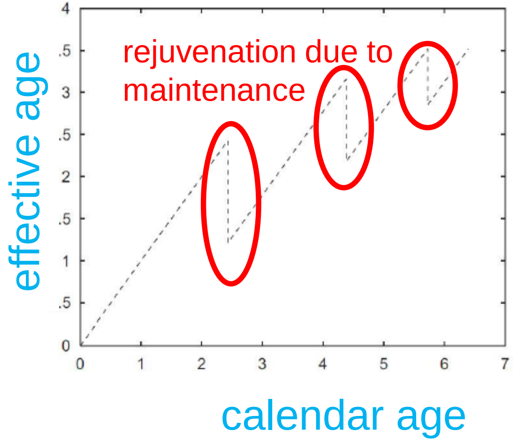
Power transformer maintenance scheduling

- Output example
  - For the 1st transformer, you should perform
    - ✓ maintenance at 11th, 17th, 22nd, 25th months
    - ✓ replacement at 28th month
- Things to consider
  - Lifetime of transformers is extended by maintenance
  - Trade-off between risk of failure and maintenance cost
  - Complex business constraints
- Reduced to solving nonlinear optimization problem

blue: maintenance

Trans.	1st	2nd	3rd	4th	5th	6th	7th
1	11	17	22	25	28		
2	7	13	18	22	24		
3	14	23	29	34	38	42	45
4	6	10	13				

yellow: replacement



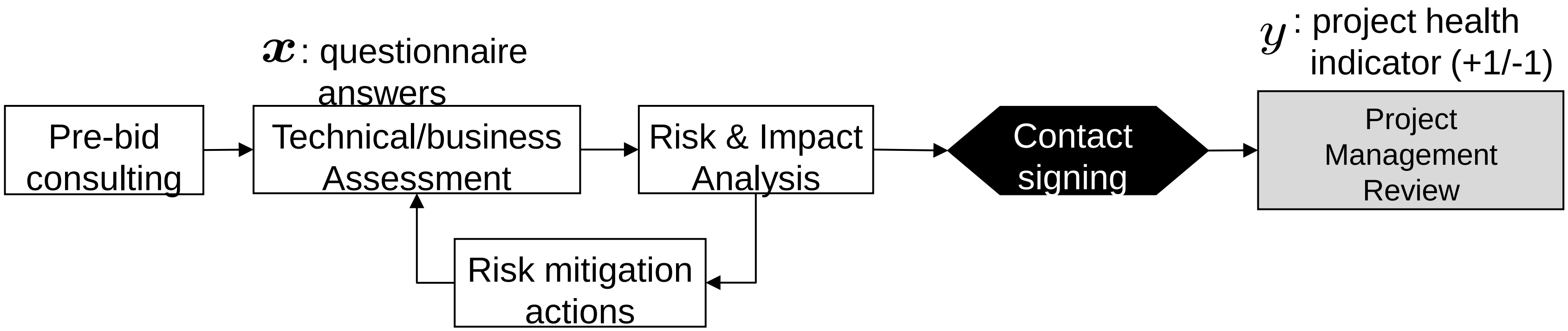
Contents

- Introduction: optimism over AI technologies
- Solution-oriented framework for sensor data analytics
- Tackling cognitive biases in project risk management

[Ide-Dhurandhar, ICDM 15]

Motivating real problem: Predict how much likely a project is going to fail

- Input data, x : questionnaire answers
  - Surveyor asks about the project status
  - Project manager answers to the questions
- Outcome value, y : failure or success (after contract signing)

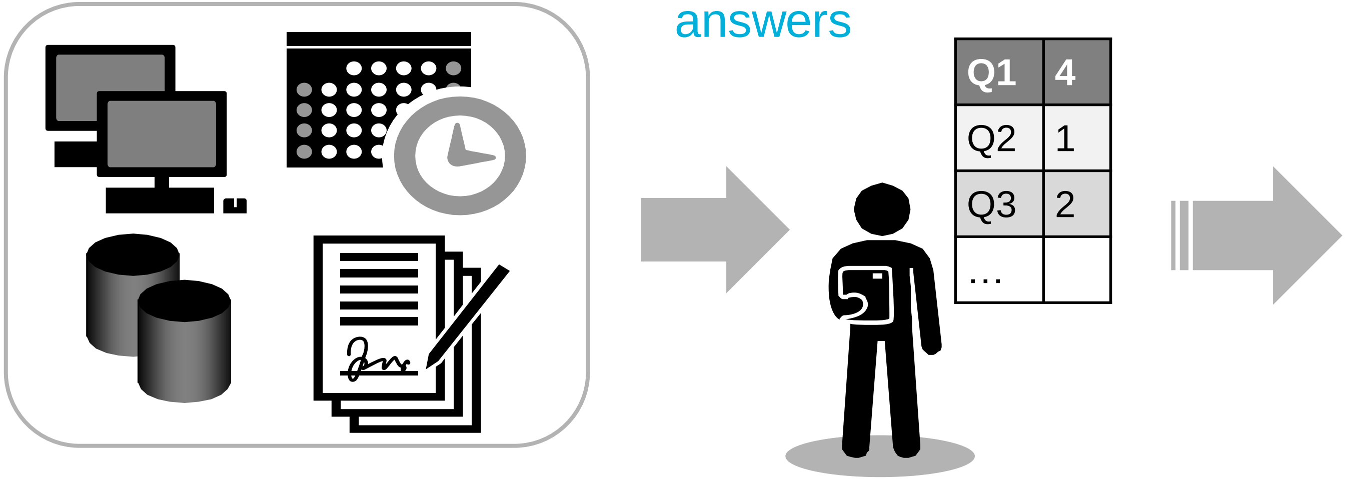


Motivating real problem: Predict how much likely a project is going to fail

- Input data, x : questionnaire answers
  - Surveyor asks about the project status
  - Project manager answers to the questions
- Outcome value, y : failure or success (after contract signing)

IT system development project

x : questionnaire answers

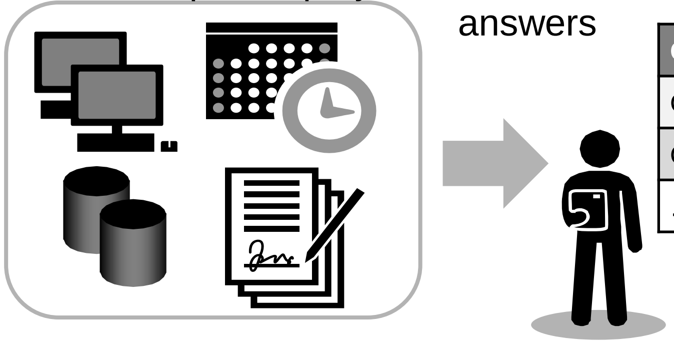


Q1	4
Q2	1
Q3	2
…	

y : failure or success?

(For ref.) What questionnaire looks like

- Major topics covered
  - Communication issues with the client
  - Well-definedness of the project scope
  - Issues related to subcontractors and internal teams
  - Project management issues



Q1	**4**
Q2	1
Q3	2
…	

Another problem: employee evaluation

- Input data, x : questionnaire answers on employee’s performance
  - Questions are like “Has he/she made good enough contributions to teamwork?”
  - Managers put evaluation on individual questions
- Outcome value, y : termination or not

Q1	4
Q2	1
Q3	2
…	

y :
stay or leave?

Interpretability really matters

Managers have to be clear on the rationale of their decision:

- What is the difference between lay-off and no lay-off groups?
- How can you justify your weighting? Why are some questions important?
- Some question may be easily achievable. How can we quantify between yes and no?

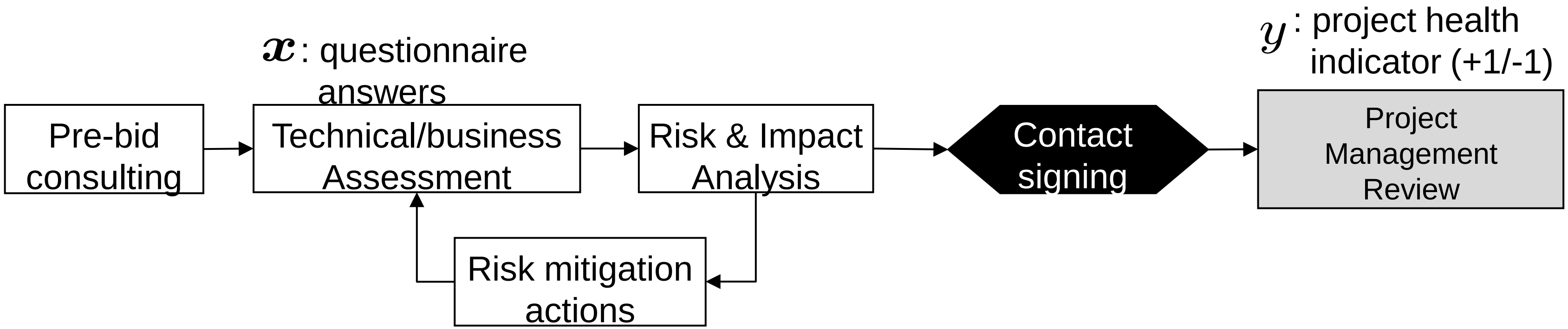
Comparison to other instances

Comparison between different questions

Comparison between different question choices

Challenge (1): No evidently bad answers. Need to discover indications of failures from apparently good answers

- Iterative review process allows removing all evident risk factors
  - This is actually a prerequisite to get into the final review right before contract signing
- However, some of them might be “pretending” as good
- Wish to discover such indications



Challenge (2): Interpretability really matters. We have to make the model fully interpretable

- Fully interpretable predictive model (in the questionnaire setting) must allow
  - quantitative comparison between subjects in terms of their importance,
  - quantitative comparison between question items in terms of their importance,
  - quantitative comparison between answer choices in terms of probability of choosing each option,
- while maintaining a comparable accuracy to other less interpretable methods.

Comparison to other instances

Comparison between different questions

Comparison between different question choices

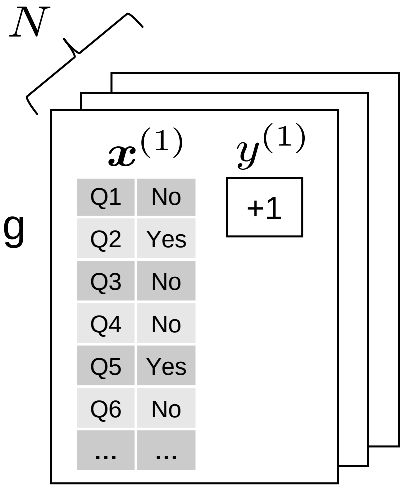
Problem summary: Informative prediction on questionnaire

- Build a fully interpretable predictive model for project failure/success (*y*) given a new set of questionnaire answers (***x***)
- Compute the informativeness of the question items

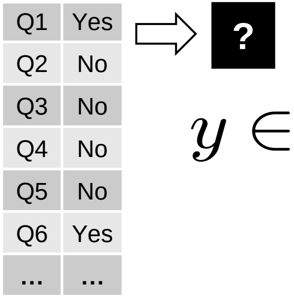
M **binary** questions
- 1: at-risk
- 0: no risk

$$\boldsymbol{x} \in \{0,1\}^M$$

Training data



Prediction model



$$y \in \{-1,+1\}$$

Key idea: Assume *x* is stochastically generated by a latent variable that is more faithful to the truth

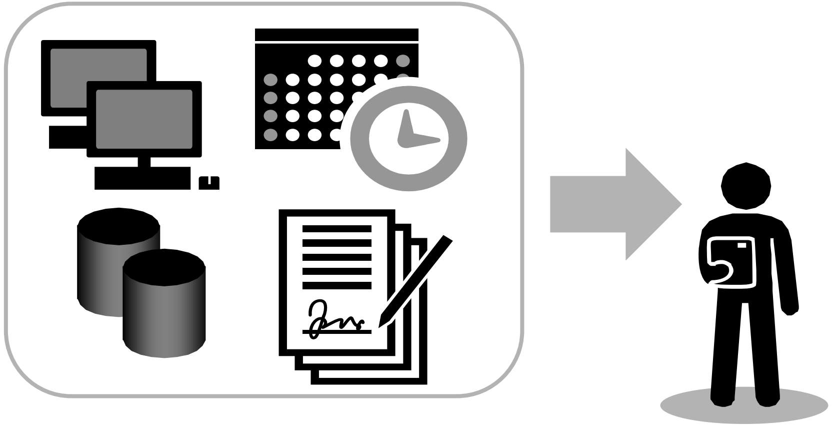
latent failure tendency

$$\theta \longrightarrow x$$

questionnaire answer

Should be more faithful to the truth

May be biased

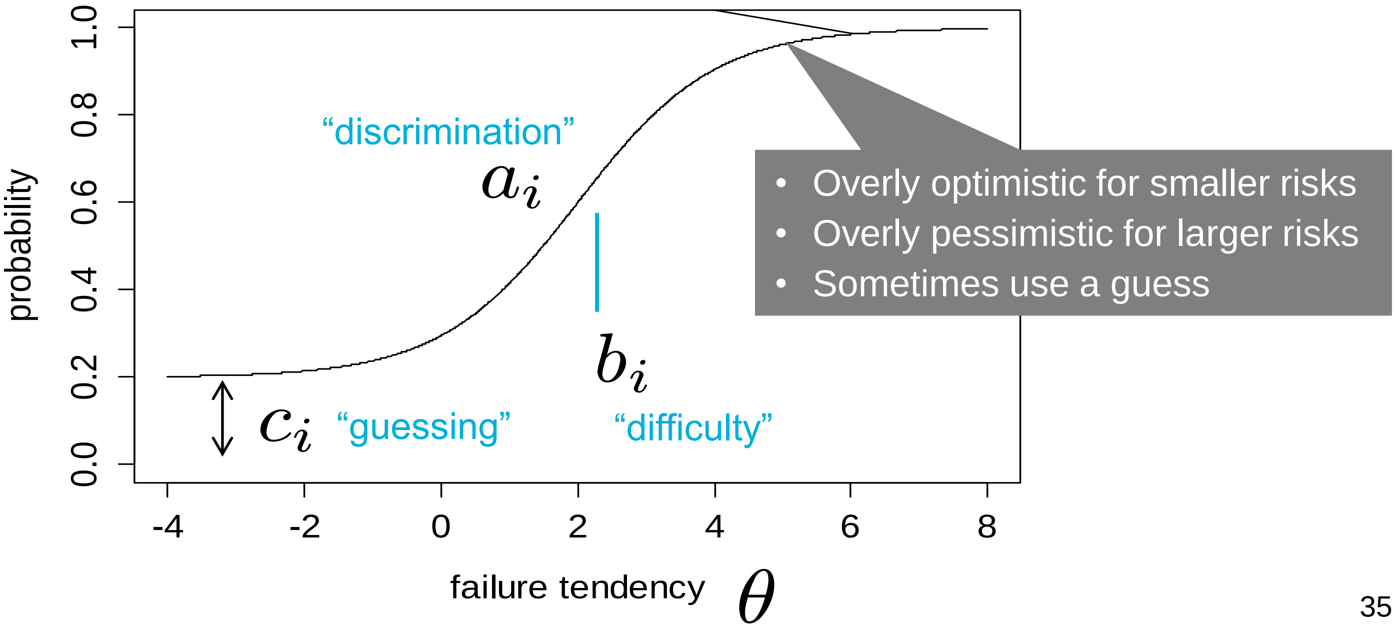


Item Response Theory: Using a "shifted S-curve" as a natural model of representing cognitive bias

- Represents nonlinear relationship of $\theta \rightarrow \boldsymbol{x}$

Prob. of answering as at-risk for the *i*-th question

$$P(\theta, a_i, b_i, c_i) \equiv c_i + \frac{1 - c_i}{1 + \mathrm{e}^{-a_i(\theta - b_i)}}$$



We are extending Item Response Model (IRT) in psychometrics

Use the same "shifted S-curve" to take account of cognitive bias

Extend the original IRT in **the supervised learning setting**

- IRT is the standard method to analyze academic tests
  - SAT is a well-known example
- IRT is unsupervised. We are developing a supervised version of IRT by including the outcome variable, y

Structure of the model: (1) Learn probabilistic model for the S-curve. (2) Learn distance metric for k-NN prediction

Historical record

training phase

$$(\boldsymbol{x}^{(n)}, y^{(n)})$$
$$n = 1, \ldots, N$$

Generative model for x based on the shifted S-curve

$$p(\boldsymbol{x}|\theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \prod_{i=1}^{M} P(\theta, a_i, b_i, c_i)^{\delta(x_i,1)} [1 - P(\theta, a_i, b_i, c_i)]^{\delta(x_i,0)}$$

Prior distribution for theta

$$f(\theta|y) = \begin{cases} \frac{\gamma}{\sqrt{2\pi}} \exp\left(-\frac{\gamma}{2}\theta^2\right) & \text{for} \quad y = -1, \\ \frac{\gamma}{\sqrt{2\pi}} \exp\left(-\frac{\gamma}{2}(\theta - \omega)^2\right) & \text{for} \quad y = +1, \end{cases}$$

model parameters

$$\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{c}}$$

Metric learning to give a distance metric, ***A***

Use numerical integration technique (Gauss-Hermite quadrature) to maximize marginalized likelihood

$$L(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}|\mathcal{D}) = \sum_{n=1}^{N} \ln\left[\pi(y^{(n)}) p(\boldsymbol{x}^{(n)}|\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},y^{(n)})\right]$$

$$p(\boldsymbol{x}^{(n)}|\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},y^{(n)}) \equiv \int_{-\infty}^{\infty} \mathrm{d}\theta^{(n)}\; p(\boldsymbol{x}^{(n)}|\theta^{(n)},\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})\; f(\theta^{(n)}|y^{(n)})$$

$$(\boldsymbol{a}^*,\boldsymbol{b}^*,\boldsymbol{c}^*) = \arg\max_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}} L(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}|\mathcal{D})$$
$$\text{subject to}\quad 0 \le c_i \le 1 \quad (i = 1,\ldots,M)$$

$$\int_{-\infty}^{\infty} \mathrm{d}\theta\; f(\theta|y)\; p(\boldsymbol{x}|\theta,\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}) \approx \sum_{i=1}^{N_h} w_i\, p\left(\boldsymbol{x} \,\middle|\, \sqrt{\frac{2}{\gamma}}\theta_i + \omega\delta(y,1),\ \boldsymbol{a},\boldsymbol{b},\boldsymbol{c}\right), \quad (8)$$

where practically good enough approximation is obtained by taking $N_h \approx 20$. The coefficients $\{w_i\}$ are defined by

$$w_i \equiv \frac{2^{N_h-1} N_h!}{N_h^2 [H_{N_h-1}(\theta_i)]^2},$$

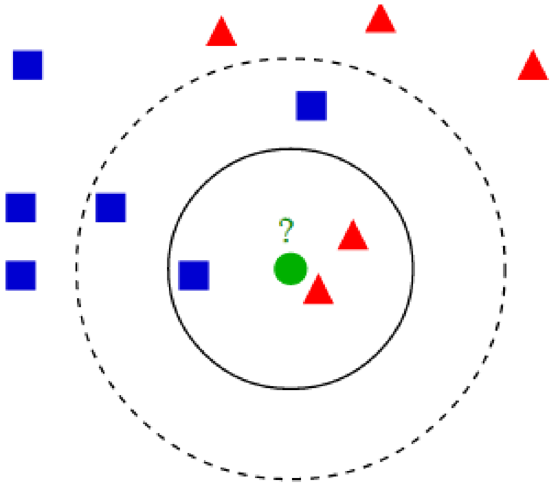
and the position of break points $\{\theta_i\}$ is determined by the roots of the Hermite polynomial $H_{N_h}(\theta)$, which are tabulated [26]. The approximation (8) means that the integration is readily performed by performing summation over about 20 terms for arbitrary values of $\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}$. Thus the use of gradient method for solving the optimization problem (7) should not be a problem.

Making prediction using estimated supervised IRT model

New question answer $\boldsymbol{x}$ → **prediction phase** → estimated binary label $\hat{y}$

Use k-NN classification
based on the distance metric

$$d(\boldsymbol{x}, \boldsymbol{x}^{(n)}) = (\boldsymbol{x} - \boldsymbol{x}^{(n)})^\top \mathsf{A} (\boldsymbol{x} - \boldsymbol{x}^{(n)})$$



The metric *A* can be found from
the supervised IRT model

Toy example: binary classification on bi-variate binary inputs

- Compared with regularized logistic regression
- Took the diagonal metric as informativeness of each variable
- Proposed method gives much richer and more informative results

TABLE I
SUMMARY OF SYNTHETIC DATA.

x	$(y = +1)$	$(y = -1)$
(0,0)	8	9
(0,1)	6	16
(1,0)	20	20
(1,1)	16	16

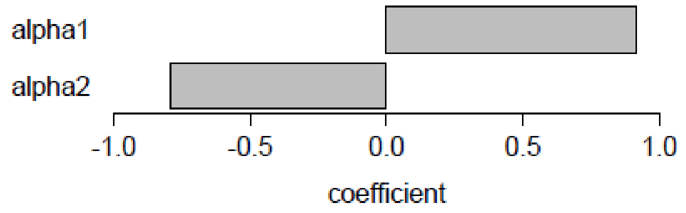


Fig. 5. Learned coefficients of regularized logistic regression for the synthetic data.

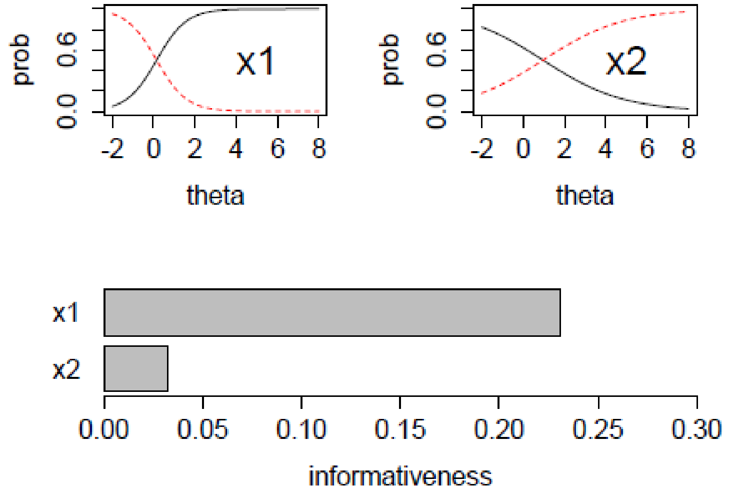
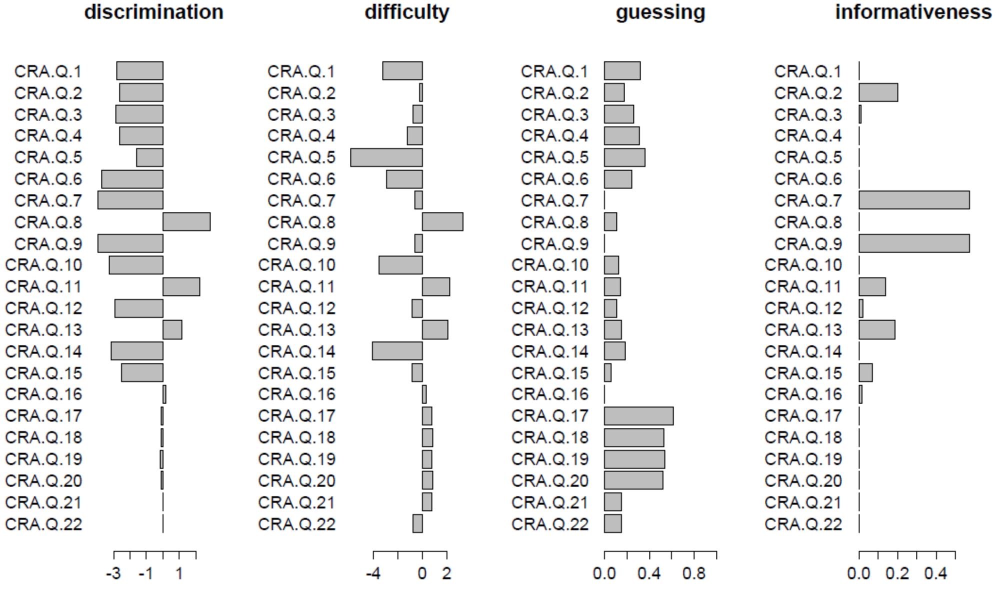


Fig. 6. Item characteristic curves and informativeness score for the synthetic data.

Experiment: Using service provider's real data in IT system development

- Questionnaire called CRA (contract risk assessment)
- M = 22 rather qualitative questions
- N = several hundred
- Each question is yes (at-risk) or no (no-risk)
- Final project evaluation is failure (y=+1) or non-failure (y= -1)

Result (1): Estimated S-curves and model parameters providing practical information on the usefulness of each question



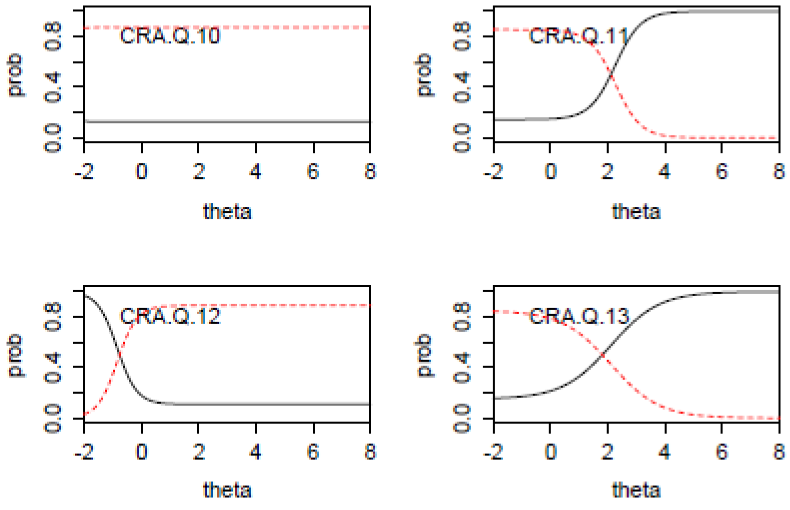


Fig. 8. Examples of ICCs for the CRA data.

Result (2): Achieved comparable or even better accuracy, while maintaining high interpretability

- Compares F-value
  - harmonic mean between troubled project accuracy and non-troubled project accuracy
- Clearly outperform the baseline
  - Baseline is based only on **x**
    - ✓ Logistic regression
    - ✓ Simple k-NN
  - Our approach uses theta instead of **x**

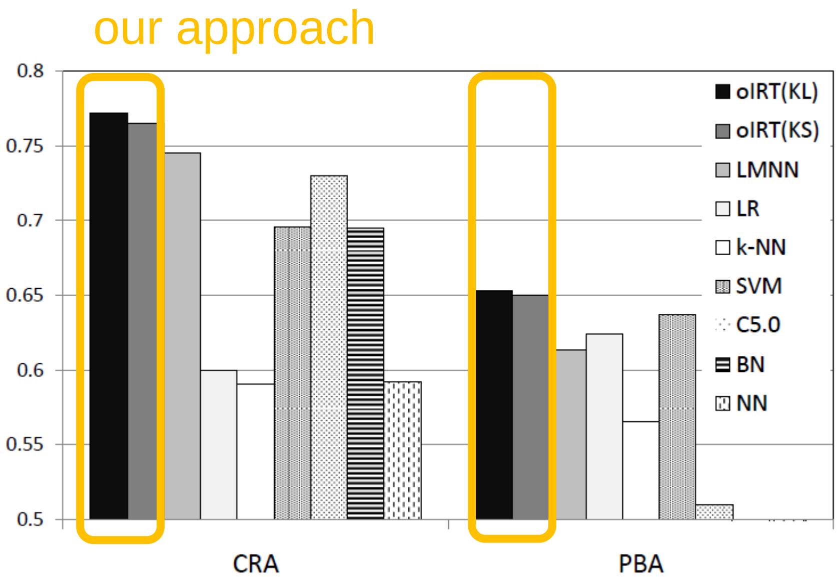


Fig. 9. Comparison of F-values (BN and NN are not visible for PBA).

Conclusion

- Proposed the notion of “full-interpretability” in the context of questionnaire data analysis
- Extended the item response theory in psychometrics to a supervised setting
- Developed a method for metric learning on the supervised IRT

Reference:

- T. Ide, A. Dhurandhar,
  - "Informative Prediction based on Ordinal Questionnaire Data“
  - Proceedings of 2015 IEEE International Conference on Data Mining (ICDM 15), 2015, to appear.
- T. Ide, et al.,
  - "Latent Trait Analysis for Risk Management of Complex Information Technology Projects"
  - Proceedings of the 14th IFIP/IEEE International Symposium on Integrated Network Management (IM 2015), 2015, pp.305-312.

Thank you!