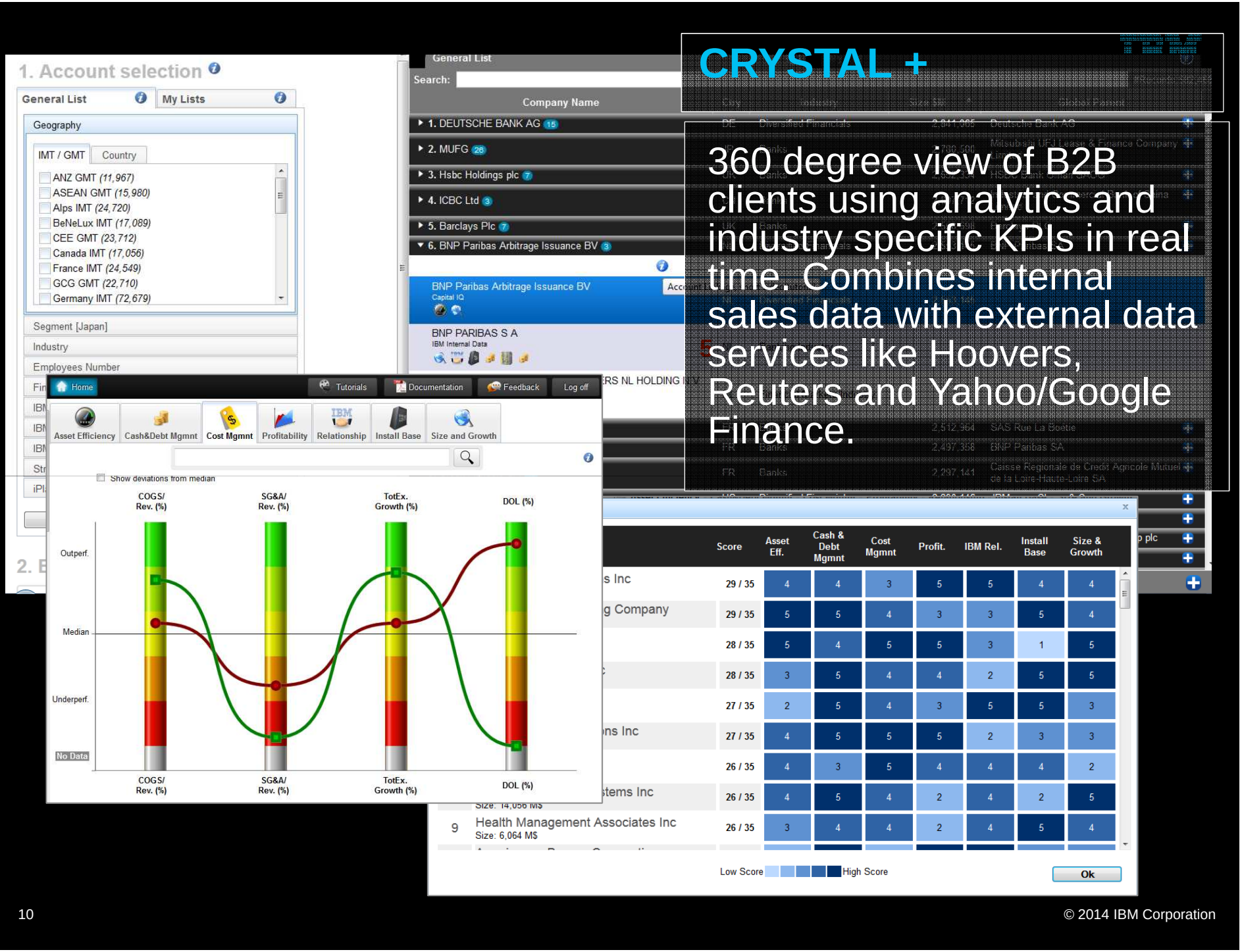
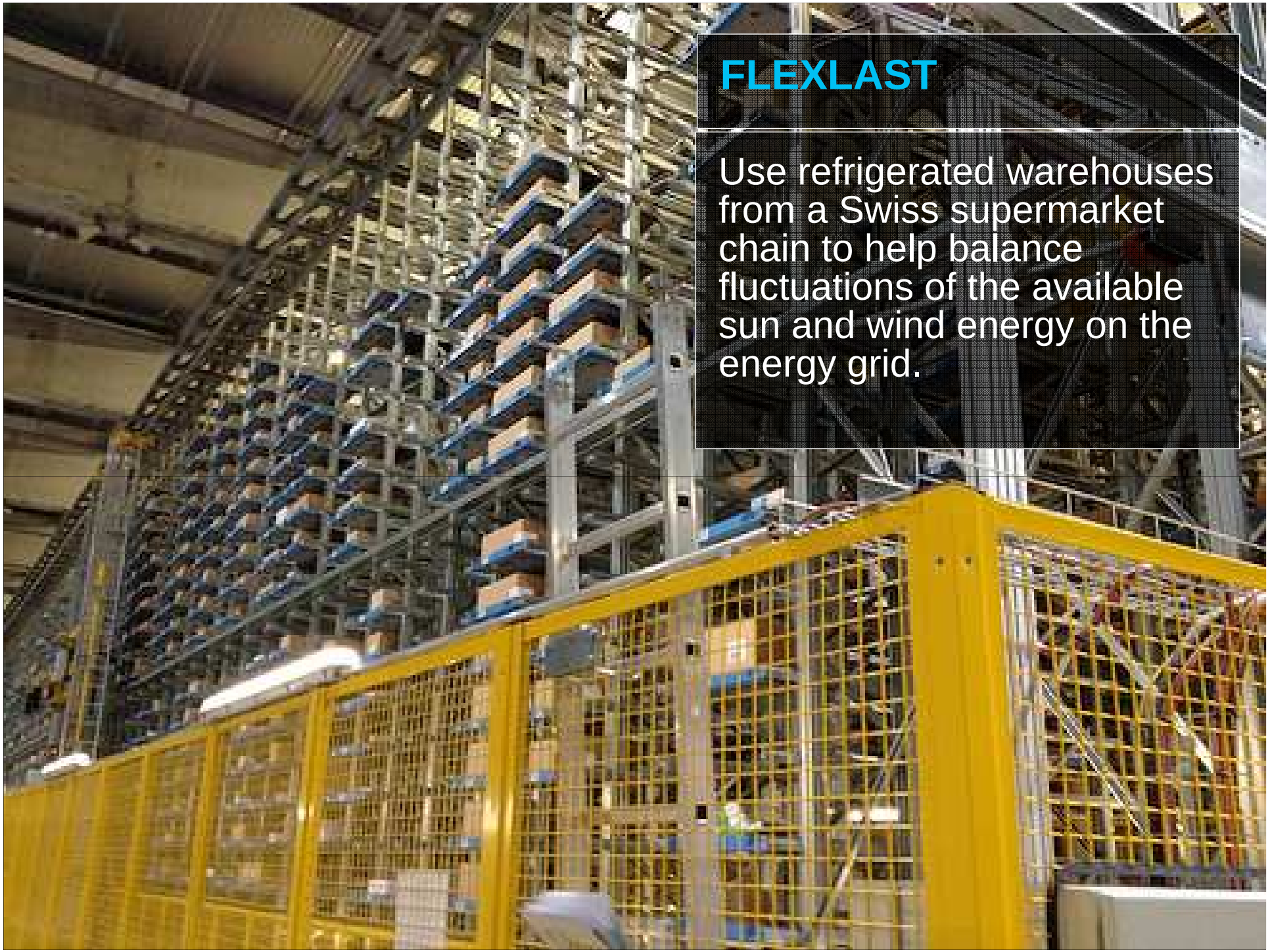
Dorothea Wiesmann, IBM Research – Zurich

Machine Learning & Predictive Analytics for IT Services

Broad Application of Analytics and Data Science

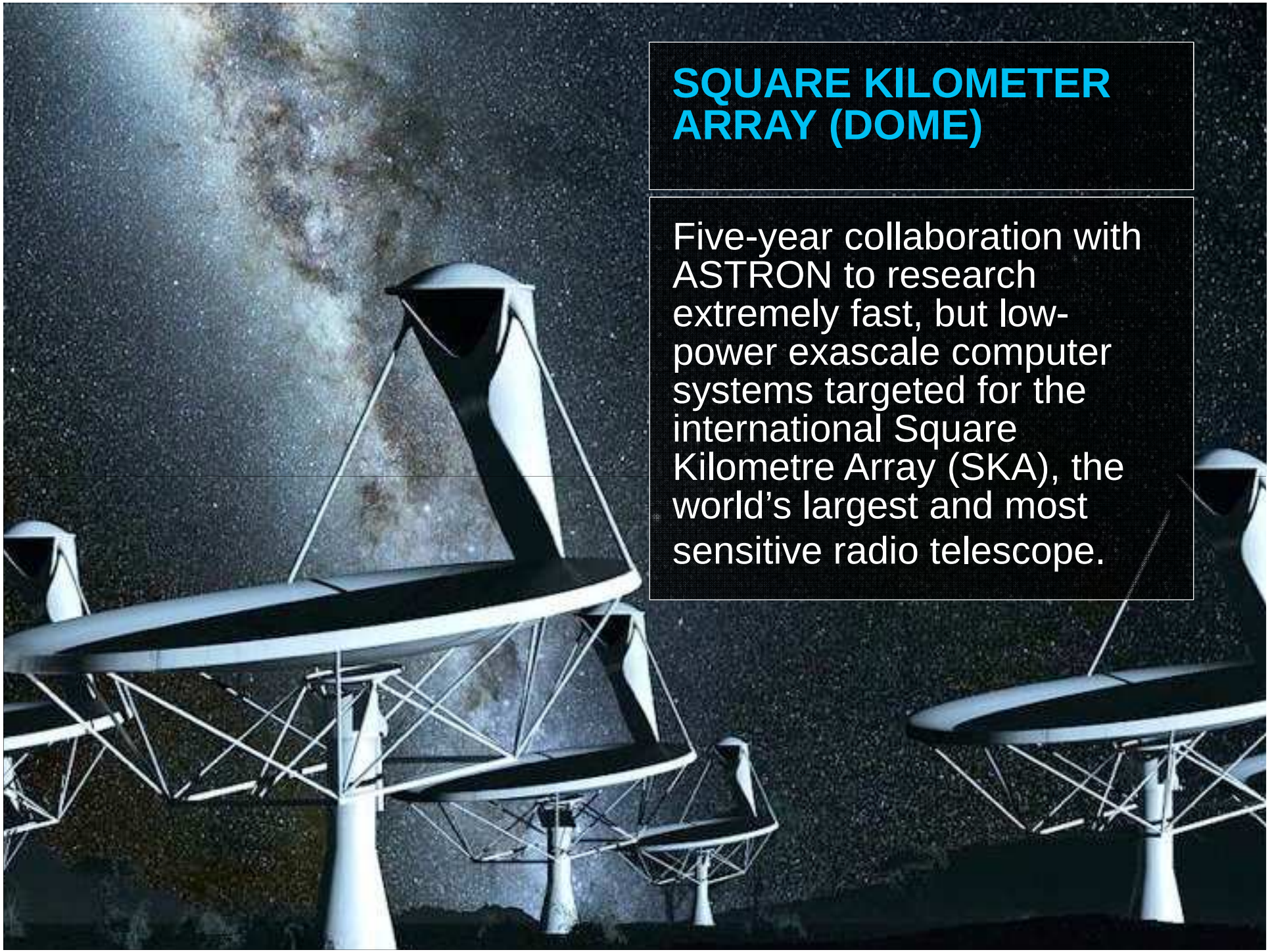
- Various projects in IBM Research – Zurich applying statistical analysis and machine learning methods in the area of sales and marketing, electrical energy management, science ….





FLEXLAST

Use refrigerated warehouses from a Swiss supermarket chain to help balance fluctuations of the available sun and wind energy on the energy grid.



SQUARE KILOMETER ARRAY (DOME)

Five-year collaboration with ASTRON to research extremely fast, but low-power exascale computer systems targeted for the international Square Kilometre Array (SKA), the world's largest and most sensitive radio telescope.

Broad Application of Analytics and Data Science

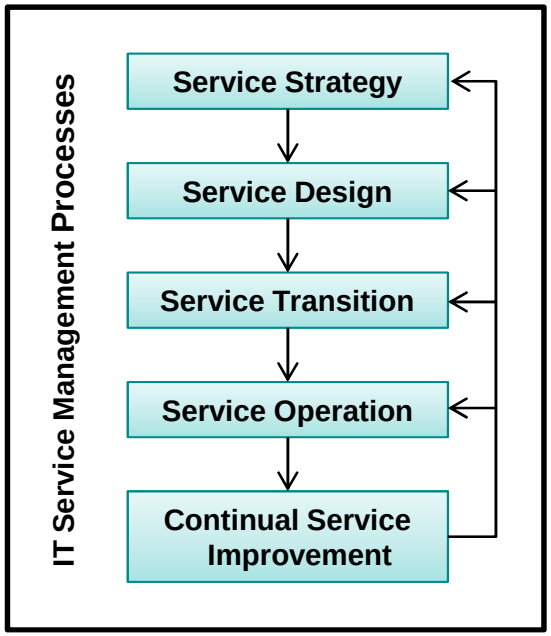
- Various projects in IBM Research – Zurich applying statistical analysis and machine learning methods in the area of sales and marketing, electrical energy management, science ….

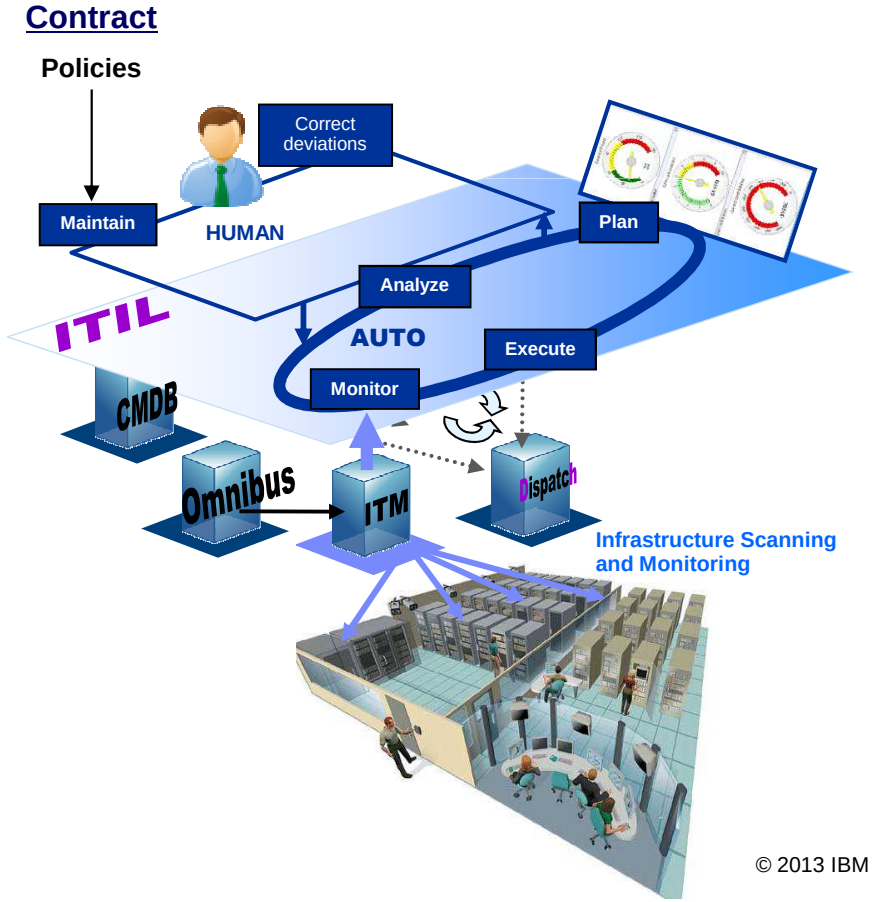
- …. focus of today's presentation on improving IT Services with machine learning and predictive analytics:

  – Introduction to IT Service Delivery
  – Service Analytics Methods and application examples

IBM Services IT Infrastructure Landscape at a Glance

- Over 400 data centers
- 8 million square feet of space
- 1,235 Mainframes
- 1,447,833 MIPS
- 100,000 Terabytes
- 330,000 images on 200,000 Servers
- 13,000,000 managed desktops
- 175,000 network devices for top 900 clients





Service Analytics Method Workload Characterization: Concept and Methods

- Goal: determine workload arrival patterns and service efforts related to different service feature

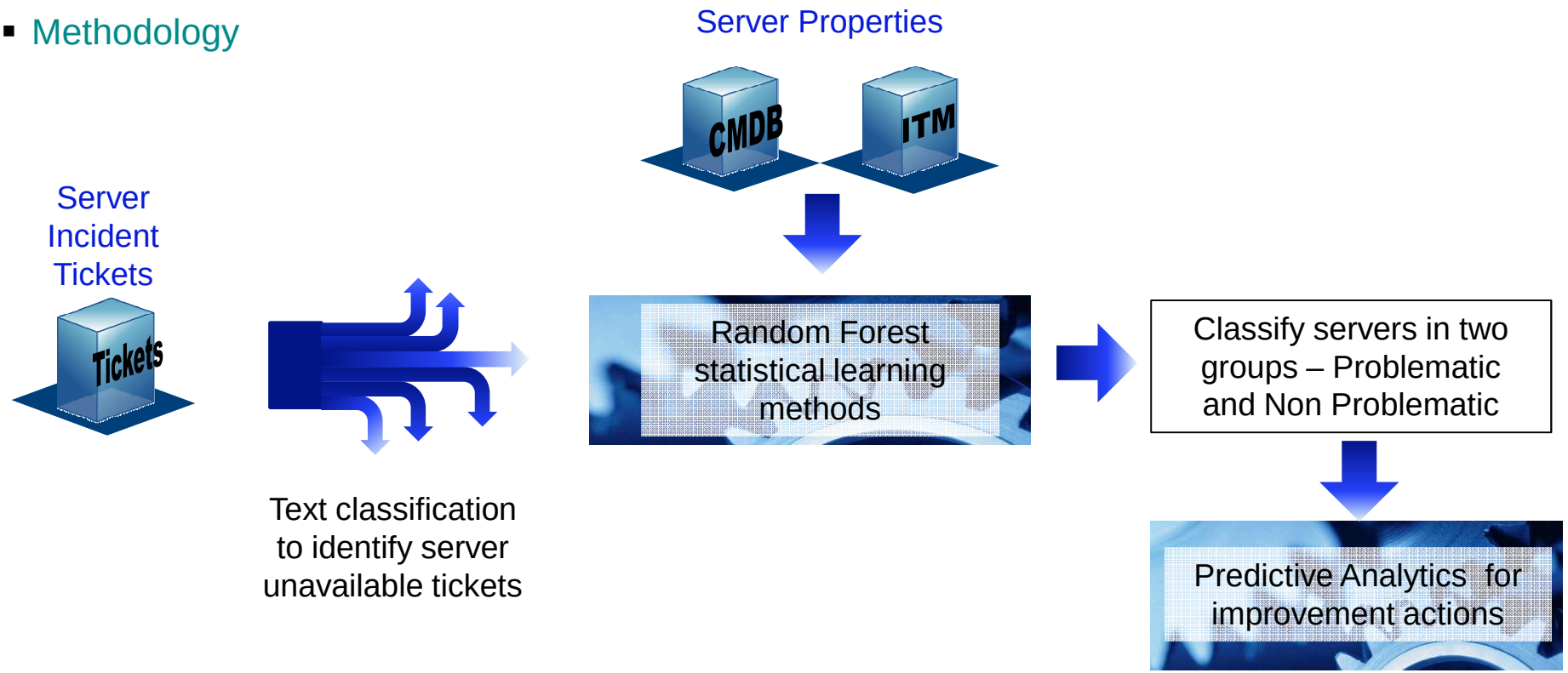
- Use case:
  - Reduce occurrence of prevalent failure types by classifying incident failure types into a priori known failure classes to prioritize root cause analysis for large-volume types
  - Reduce violation of service agreement levels (SLA) by assigning tickets to Subject Matter Experts who have handled similar tickets in the past
  - Reducing problem determination effort by recommending relevant solutions from similar previously solved incidents

- Methods:
  - Workload classification: gradient boosted model with unigram decomposition and stop word removal, random forest
  - Request clustering: hierarchical clustering and k-means clustering

Workload Characterization: Problematic Server Configurations (1/4)

- Objective: Through analytics of combined incident ticket and server configuration information data optimize IT Service Management by
  - identifying high risk/high workload server assets,
  - quantifying workload and risk driver pervasive across different IT environments

- Methodology



Workload Characterization: Problematic Server Configurations (2/4)

- Ticket Text Classification Method: Gradient boosted model
  - Ensemble (weighted sum) of weak learners (decision/regression trees with K terminal nodes)
  - Steepest decent iterative procedure where each step selects the tree that minimizes the loss function at most (most closely approximates its gradient) on a random subsample of the training set
  - The corresponding multiplier of the chosen tree is then selected using line search
- Ticket Text Classification Results

Labeled Set

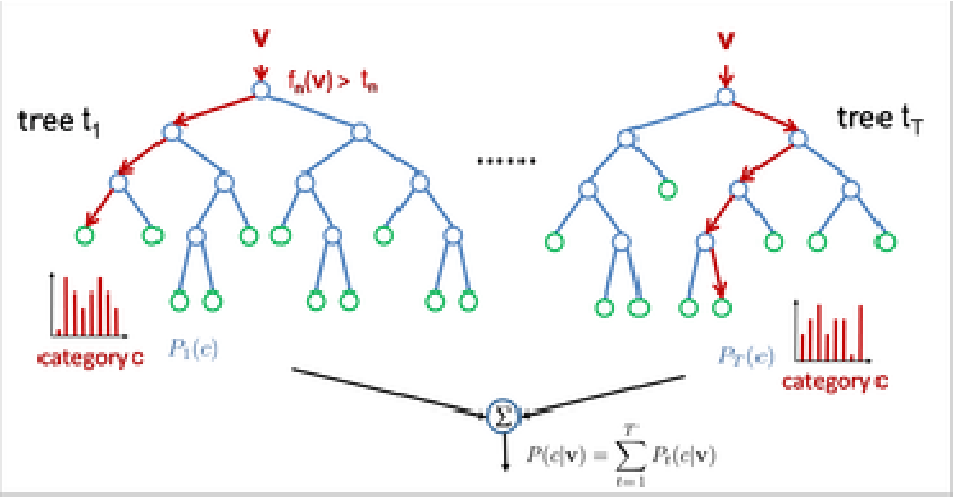
Ticket Class	Count	%
Non-actionable Tickets	1391	71.3%
Server Unavailable	105	5.4%
Disk/FS Capacity	214	11.0%
Other	238	12.2%
CPU/Memory	4	0.2 %

Ticket Class	Recall	Precision
Non-actionable Tickets	0.983	0.947
Server Unavailable	0.741	0.903
Disk/FS Capacity	0.941	0.987
Other	0.730	0.789
CPU/Memory	0	0

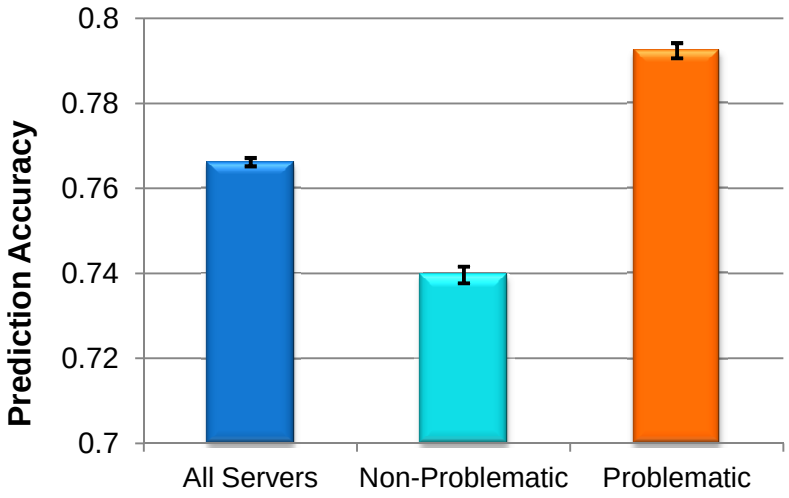
Workload Characterization: Problematic Server Configurations (3/4)

Random Forest statistical learning methods

- Data Set: 10k servers (samples) and their OS-level incident tickets for 12 months
- Model: Random Forest
  - Able to handle categorical predictor readily
  - Captures nonlinear relationships
  - Robust against over-fitting and thus generalizes better
  - Not very sensitive to outliers in training data



Source: http://dovgalecs.com/blog/matlab-random-forest-classifier/



Workload Characterization: Problematic Server Configurations (4/4)

Improvement options | **Ticket details**

		Remediation actions	Improvement forecast (?)
1	Best single	OS refresh (to latest)	34%
2		Memory increase (3x)	9%
3		Disk capacity increase (2x)	8%
4		Disk capacity increase (3x)	8%
5		Disk capacity increase (4x)	8%
6	Best combined	OS refresh (to latest) + Disk capacity increase (2x)	40%
7		OS refresh (to latest) + Disk capacity increase (3x)	40%
8		OS refresh (to latest) + Disk capacity increase (4x)	40%

Current incident behaviour

Study period: 2013-05-01 - 2013-12-10
Ticket classes: Server unavailable

High-severity incidents (severity 1&2 tickets)	Problematic	**11.57 per month**
All incidents (total tickets)	Problematic	**14.57 per month**

Hardware & OS

Hardware architecture	**IBM x**
Age	**8 years**
Operating System	**MICROSOFT WINDOWS 2000**
Purpose	Application
Virtualization level	N/A

Utilization measures

Avg CPU utilization	**14.31%**
Max CPU utilization	N/A
Memory utilization	**69.51%**
Disk utilization	**15.16%**

Service Analytics Method Performance Prediction

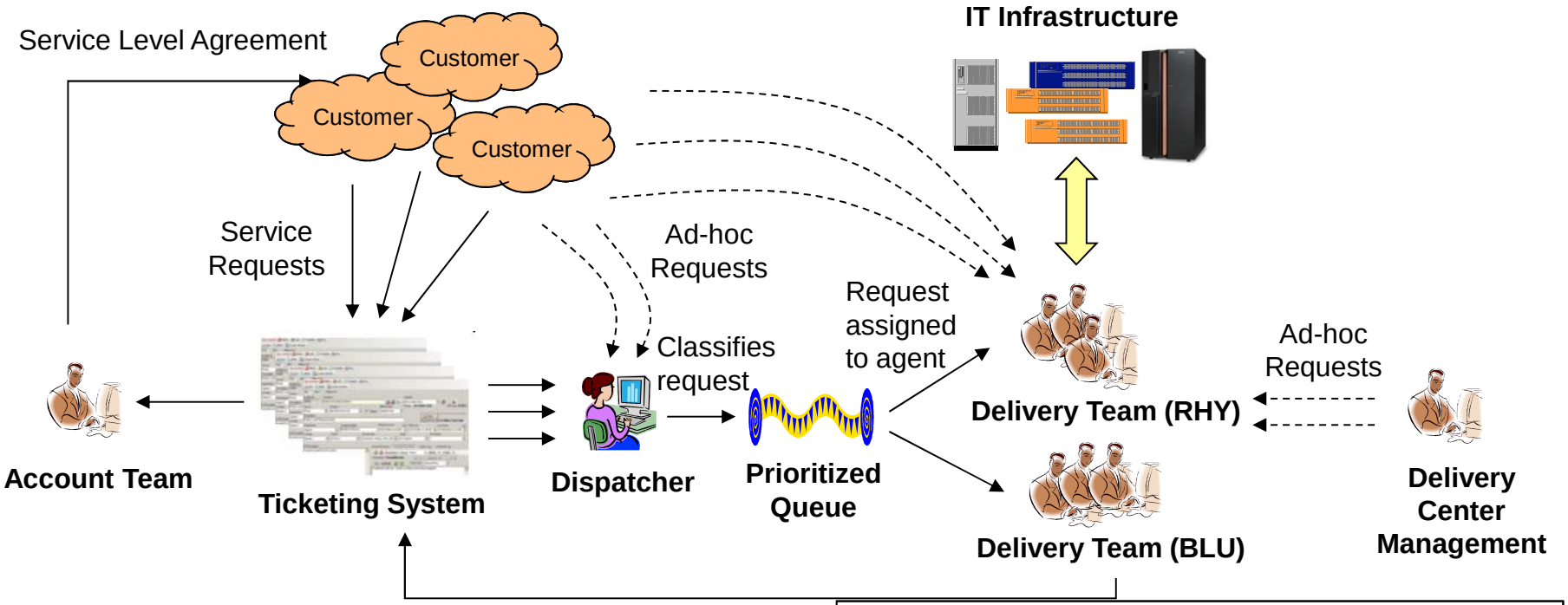
- Goal: leverage the insight on factors that impact IT Service delivery to predict service cost, i.e. volume and effort to perform service activities

- Use case:
  - Predict volumes expected for certain types of services activities. In combination with per-activity effort estimation this allows to tackle business problems pertaining to staff planning, SLA feasibility assessment, and return of investment of automation projects.

- Methods:
  - Multivariate time series forecasting
  - Longitudinal methods

Service Analytics Method Workforce Optimization (1/3)

- Goal: improve IT management effectiveness through resource planning and staffing recommendation. Two main categories: workforce optimization and workload optimization
- Example: Optimize staffing of IT Service teams



To date, model has been applied to more than **450** Delivery Teams, located in over **20** delivery centers from **8** service function areas, including over **15,000** delivery personnel

Large variability in work volumes, request types, skill requirements, and service times

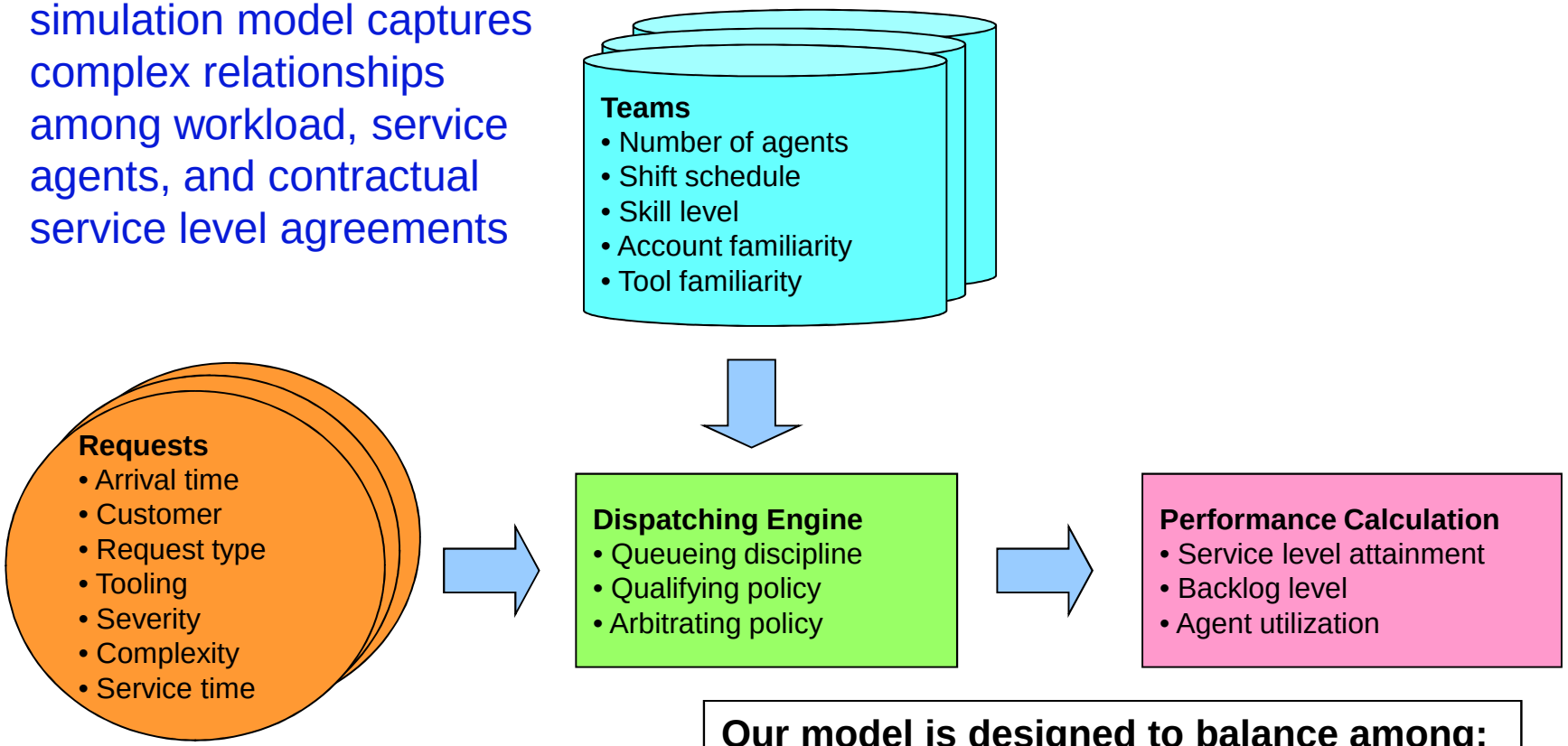
Lack of standards in work data recording (e.g., no telephony switches to record data)

Strict requirements on service level targets

Source: Y. Diao et al., WSC 2011 and CNSM 2011

Service Analytics Method Workforce Optimization (2/3)

A detailed discrete event simulation model captures complex relationships among workload, service agents, and contractual service level agreements



Our model is designed to balance among:

Flexibility for unique team situations

Accuracy for capturing various scenarios

Simplicity for easy parameterization

Source: Y. Diao et al., WSC 2011 and CNSM 2011

Service Analytics Method Workforce Optimization (3/3)

Stage 1

Minimize the staffing cost as the # agents per shift and per skill level

$$C_{\min} = \min \sum_{j=1}^{J} \sum_{k=1}^{K} c_k x_{jk}$$

$$s.t.$$

Meet contractual service level constraints and operation dynamics as captured in the simulation model

$$f_{ip}\left(v_{ipc}^{t}, s_{ipc}, y_{jl}, x_{jk}, w_{ip}, m_{ip}\right) \le \alpha_{ip}$$

Meet staffing constraints regarding skills, shift structures, contracts, local regulation, etc.

$$\underline{x}_k \le \sum_{j=1}^{J} x_{jk} \le \bar{x}_k$$

Ensure system constraints such as agent utilization and stable backlog

$$\underline{x}_j \le \sum_{k=1}^{K} x_{jk} \le \bar{x}_j$$

$$\sum_{j=1}^{J} \sum_{k=1}^{K} a_{kr} x_{jk} = b_r$$

$$\ldots$$

$$x_{jk} \ge 0$$

A two-stage optimization approach minimizes the labor cost and the disruption to current business operations

Stage 2

$$\min \sum_{j=1}^{J} \sum_{k=1}^{K} d_{jk}\left(x_{jk} - x_{jk}^{0}\right)$$

$$s.t.$$

Minimize the deviation from current staffing levels

Ensure minimum system cost from Stage 1

$$\sum_{j=1}^{J} \sum_{k=1}^{K} c_k x_{jk} = C_{\min}$$

$$\ldots$$

Source: Y. Diao et al., WSC 2011 and CNSM 2011

Thank You!