



# Using Data Mining for Accurate Resource and Skill Demand Forecasting in Services Engagements

Jianying Hu, Moninder Singh, and Aleksandra Mojsilovic

**IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, U.S.A**

KDD 2008 Workshop on Data Mining for Business Applications

# Talk Overview

- Introduction
- OnTheMark (OTM) System Overview
- Problem Description
- Technical Solution
- Evaluation
- Open Issues & Future Plans

# Introduction

- Multiple service projects, competing demand for limited human resources
- Estimate demand for projects/skills & determine optimal staffing levels & resource allocations
- Look at ongoing & in-pipeline projects
- Need to map each project to correct solution category since staffing requirements & cost structures different for different solution categories
- Solution categories themselves change frequently due to evolving business needs and dynamic customer environments

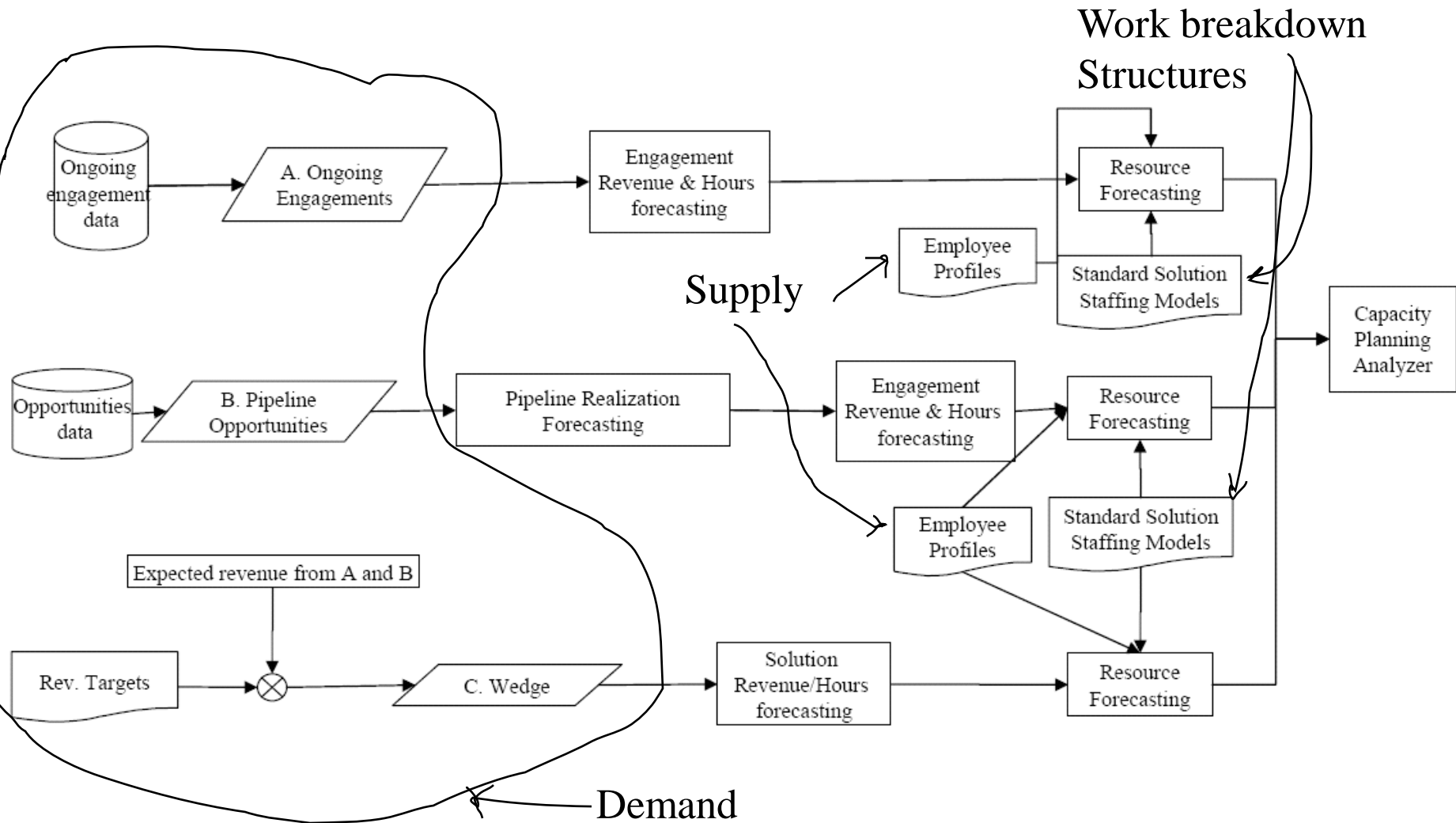
# Introduction

- Due to changing solution categories and large number of projects, manual categorization is limited
- Need automated technique to do so.
- Paper presents new approach to this mapping problem by formulating it in a semi-supervised problem  
and  
describes its application in OnTheMark, a web-based decision support system for demand forecasting & capacity planning as well as computing business metrics for assessing quality of services delivery

# OnTheMark (OTM) System Overview

- Developed for Integrated Technology Services (ITS) line of IBM Global Technology Services (GTS), one of three business units of IBM Global Services.
- Deployed in the US.
  - Resulted in significant reduction in planning cycle times, improved accuracy and increased insight into business & potential for improved profitability.
- Becoming GTS worldwide planning standard.
  - Deploying worldwide in 2008.
- Offerings categorized into solution categories called Service-Product lines (SPLs)
  - Each SPL consists of multiple solutions in specific area, such as business continuity and resiliency services, middleware services, etc.

# OnTheMark (OTM) System Overview



# Problem Description

- Solution categories with associated descriptions
  - Server product services for Microsoft
  - MS Application development and integration services
- Project data with hourly claimed data by job role/skill set
  - PM 30%, S/w Arch 30%, S/w dev 40%
- Project description
  - Optional; noisy, unstructured text; limited by fixed field-width
  - “zOS planning & migration”, “perform review system I”
  - “Hourly svcs”, “application infrastruct”
  - “P001”, “prime bidder”

# Problem Description

- Formulation of problem
  - i. dataset with underlying “basic features” attached to every data point
  - ii. set of “categories” with accompanying “category descriptions”
  - iii. some data points also have textual “data descriptions” generated independently of the category descriptions

# Technical Solution

- Cast problem as a semi-supervised clustering one
- Soft-Seed Generation
  - Use text matching between optional project descriptions and category descriptions to generate “soft” seeds
    - Category labels generated by text matching not guaranteed to be accurate; so soft constraints with varying degrees of strength
    - Seeds do not necessarily provide complete coverage of all categories
  - Incorporate a seed reassignment penalty to handle varying degrees of confidence of such seeds

# Technical Solution

- **Soft-Seeded K-means**
  - Initialize centroids of covered clusters using labeled data points
  - Initialize centroids of remaining clusters through random sampling of unlabeled data points
  - For each data point, assign to cluster that minimizes the distance of point from cluster centroid plus a seed reassignment penalty function
  - Iterate...

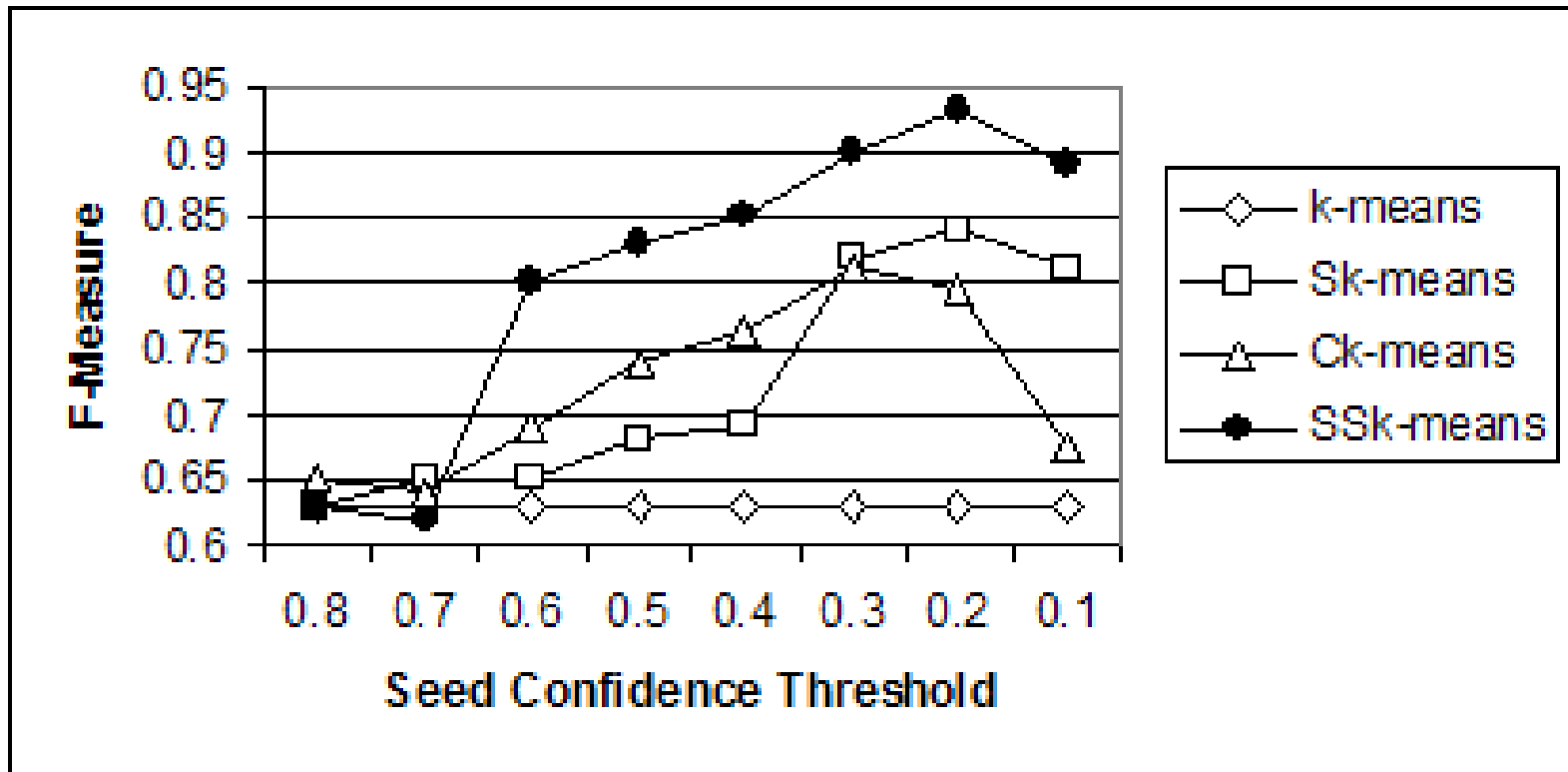
$$\mathcal{P}(j, l_i, s_i) = \begin{cases} 0 & \text{if } j = 0 \text{ or } j = l_i \\ \frac{\gamma}{1+e^{(-\alpha(s_i-\beta))}} & \text{otherwise} \end{cases}$$

# Evaluation

- 302 projects from Server Services Product line
- 8 predefined solution categories
  - validated, re-mapped based on domain experts' input
- 67 dimension skill allocation vector

Conf.	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
% of total	9	10	31	32	33	40	43	59
Coverage	1	2	4	5	5	7	7	7
Accuracy	100	100	100	99	98	96	89	78

# Evaluation



## Open Issues & Future Plans

- Improve seed quality by integrating with WordNet, etc.
- Better model initialization or selection metrics for situations where seed coverage is incomplete
- Evaluation on additional datasets
- Could be applied to other domains, such as stock image databases categorization, etc.
- Integrate fully with OTM.