

# Empirical Approach to Software Engineering

Kenichi Matsumoto  
 Nara Institute of Science and Technology (NAIST)  
 EASE Project, Ministry of Education, Culture, Sports, Science and Technology (MEXT)  
 Japan

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



# Outline

- Introduction of NAIST
- Current status of Software industry in Japan
- Empirical Software Engineering
- EASE Project
- EPM
- Application of EPM to Open Source Software
- Data Analysis on EPM

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



# Introduction of NAIST

Nara Institute of Science and Technology

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



# Location



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



# Organization

- Graduate School of **Information Science**
- Graduate School of Biological Sciences
- Graduate School of Materials Science
- Digital Library
- Information Technology Center
- Research Center for Advanced Science and Technology
- Administration Bureau
- ...

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



# Information Science



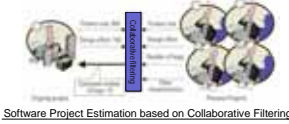
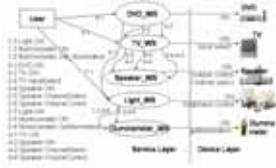
Number of Students	Master's Program	292
	Doctoral Program	124

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Software Engineering Lab.

- Web Service
- Software Protection
- Software Use Support
- **Software Metrics**
- **Software Process**



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.

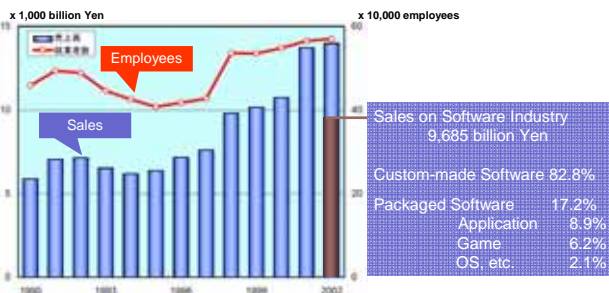


## Current Status of Software Industry in Japan

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## IT Service Industry in Japan: Sales and Employees



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Nikkei Report (2003/11/17)

- Questionnaire based investigation about the **success rate of software development project** in Japan.
- **1,746 companies** replied to the questionnaire.
- Each company was requested to answer about the largest project of software development in 2003.

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



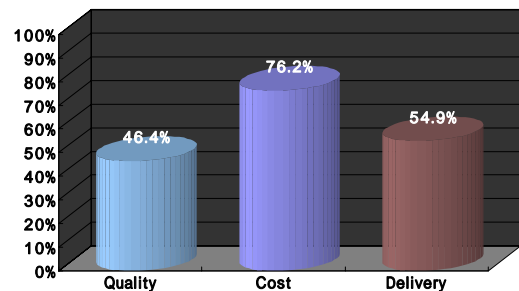
## Assumptions: Project Success

- Success in software **quality**
  - if the company got customer satisfaction for the software developed by the project.
- Success in development **cost**
  - if the company completed software development at less cost than the planned cost.
- Success in **delivery** (time schedule)
  - if the company completed software development before the planned completion date.

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



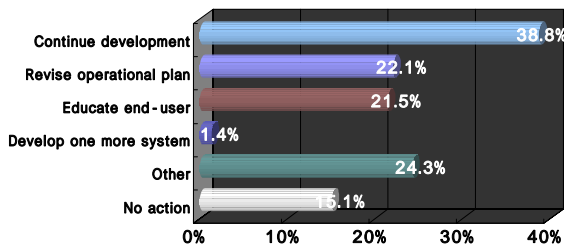
## Success Rate of QCD



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



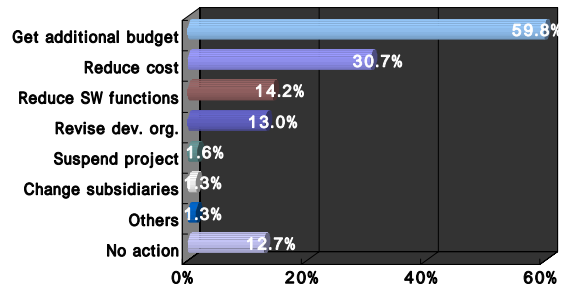
## Corrective Action for Poor Quality



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



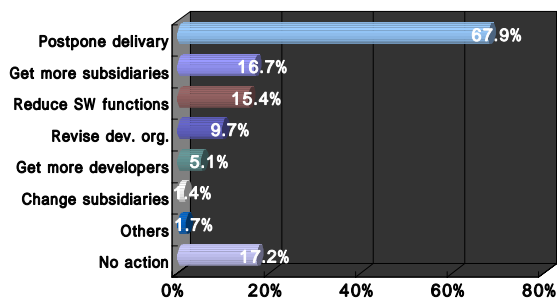
## Corrective Action for Cost Overrun



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



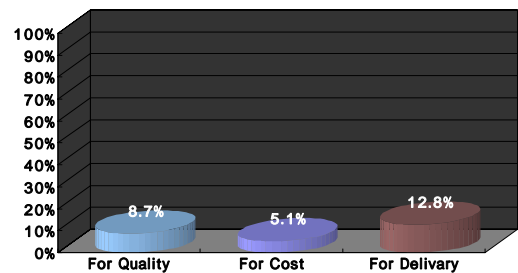
## Corrective Action for Time Overrun



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## How many companies performed "Quantitative Management"?



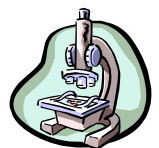
Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Empirical Software Engineering

## Empirical Software Engineering

- Various technologies in Software Engineering based on **empirical data**.
  - Data of process and product collected from software development project in academic or industrial environment.
- SE research requires empirical evidence regarding its effectiveness and applicability.
- Published empirical studies have to be (theoretically) replicatable in order to add to the existing body of knowledge.



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Observations, Laws and Theories

- Empirical observations
  - Facts from individual empirical studies.
  - We can **characterize** phenomena based on them.
- Laws
  - Repeatable observations.
  - We can understand context enough to make prediction about future observations.
  - We can **predict** phenomena by them. (what)
- Theories
  - Cause-effect relationships.
  - We can **explain** phenomena by them. (why)

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Journal by Kluwer Empirical Software Engineering

- **Scope**
  - Cost estimation techniques
  - Analysis of the effects of design methods and characteristics
  - Evaluation of testing methodologies
  - Development of predictive models of defect rates and reliability from real data
  - Infrastructure issues, such as measurement theory, experimental design, qualitative modeling and analysis approaches.



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## International Symposium on Empirical Software Engineering



- ISESE2002 @ Nara, Japan
- ISESE2003 @ Roma, Italy
- ISESE2004 @ California, USA
- ...

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## ISERN

International Software Engineering Research Network

- ISERN is a community that believes software engineering research needs to be performed in an experimental context.
- ISERN was established in 1993 by researchers of software engineering from 12 countries, including USA, Germany, Australia, Italy, Finland and Japan.
- ISERN provides several means of communication between members;
  - Electronic Communication,
  - Annual meetings, and
  - Exchange of researchers.

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Japan Software Engineering Center

- Open in October 2004 supported by Ministry of Economy, Trade and Industry (METI).
- Conduct in-depth practical studies to solve the issues of today's software industry.
- Budget of 2004:
  - 1.48 billion yen



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Aims

- Software Process Improvement methods for the Japanese Industry
- **Software measurement standards**
- Demonstration of the methods and tools in advanced software development projects

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Approach to Measurement Std.

- Conduct research into methods of collecting and analyzing quantitative data to measure the quality of software and the productivity of its development.
  - Gather data from various software development projects underway.
  - Analyze these quantitative data.
- And then
- Promote the use of such measurement standards, providing the means to archive highly qualified software with high productivity.

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## EASE Project

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## What is the EASE project?

- **E**mpirical **A**pproach to **S**oftware **E**ngineering
- One of the leading projects of the Ministry of Education, Culture, Sports, Science and Technology (MEXT).
- 5 years project starting 2003.
- Budget: 200 million yen / year.
- Project leader: Koji Torii, NAIST  
Sub-leader: Katsuro Inoue, Osaka University  
Kenichi Matsumoto, NAIST

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



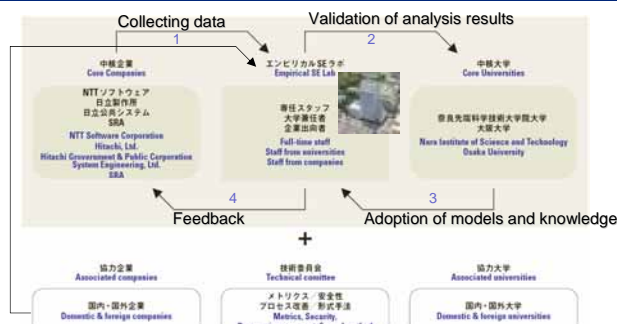
## Aims

- To practice the empirical approach, the same approach adopted by other scientific and engineering fields, including measurement, analysis and evaluation, and feedback for improvement of software quality and productivity.
- MEXT demands the project not only do research, but make an impact on industry.

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Organization



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Advisers

- Prof. Victor R. Basili  
Univ. of Maryland, U.S.A
- Prof. Barry W. Boehm  
Univ. of Southern California, U.S.A
- Prof. Dr. Dieter H. Rombach  
Univ. of Kaiserslautern  
Fraunhofer IESE, Germany
- Prof. Ross Jeffery  
Univ. of New South Wales  
Australian National ICT, Australia

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



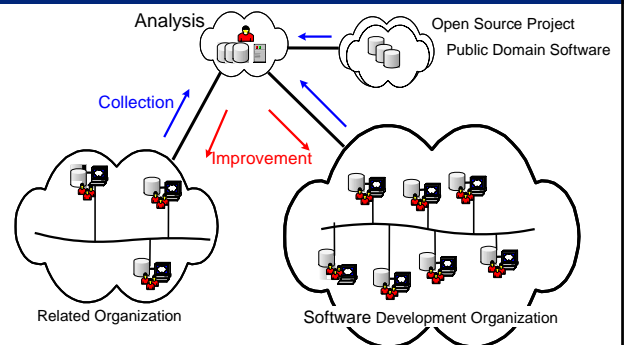
## Approach to Empirical SE

- Construction of an **empirical environment**.
- Distribution of the empirical environment and application to real projects.
- Accumulation of knowledge derived from empirical data and its analysis.

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



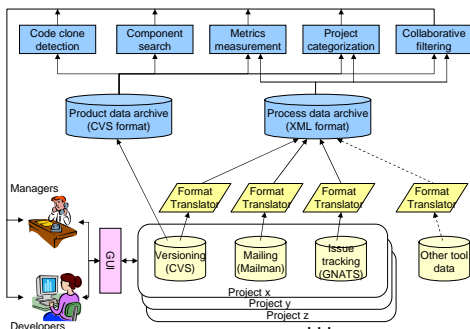
## Concept of Empirical Environment



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Architecture

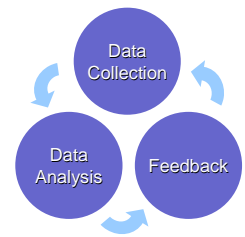


Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Cycle of Empirical Activities

- Data collection in real time
  - configuration management history
  - issue tracking history
  - e-mail communication history
- Analysis with software tools
  - metrics measurement
  - project categorization
  - collaborative filtering
  - software component retrieval
- Feedback to stakeholders for Improvement
  - observations and rules
  - experiences and instances in previous projects



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Merits to Introducing the Empirical Environment

- Easy monitoring of the project in cooperation with the existing development environment.
- Easy accumulation of the knowledge and experience of projects.
- Collection of uniform data for the entire company in real time.
- Automatic integration and reuse of information enabled through integration of empirical data.

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## EPM

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



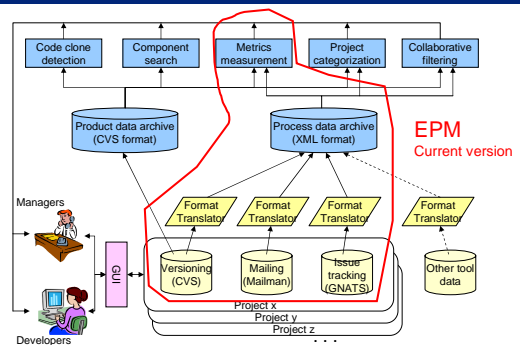
## EPM: Empirical Project Monitor

- A **partial implementation** of Empirical Environment
- EPM **automatically collects** development data accumulated in **open source** development tools through everyday development activities
  - Configuration management system: CVS
  - Issue tracking systems: GNATS, Bugzilla
  - Mailing list managers: Mailman, Majordomo, FML

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Partial implementation of Empirical Environment



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Problems in Document-based Data Collection

- Burden on developers
  - The more managers want to know development status/progress, the more developers need to make documents.
- Information delay
  - Documents cannot be used or shared until they are required.
- Tampering
  - Human (managers/developers) might tamper with documents.

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



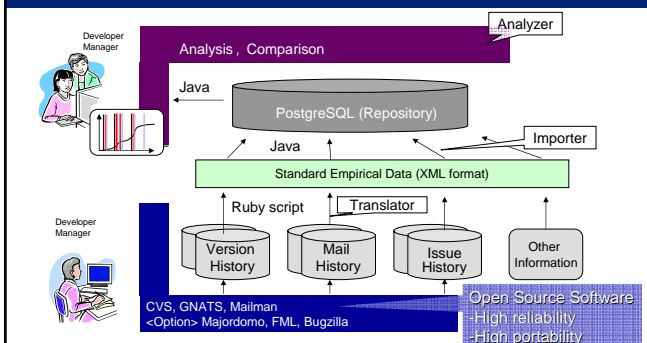
## Automatic Data Collection in EPM

- Burden on developers
  - >>> without additional work for developers
- Information delay
  - >>> in real time
- Tampering
  - >>> using raw (quantitative) data

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Architecture based on OSS



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## GUI

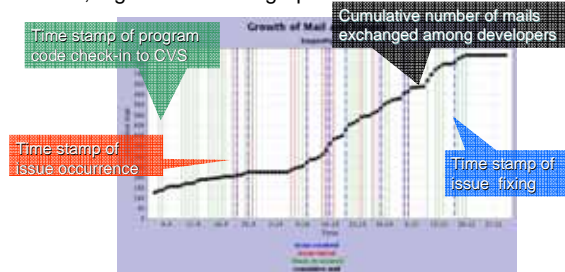


Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Example of Output

- EPM can put data collected by CVS, Mailman, and GNATS, together into one graph.



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.

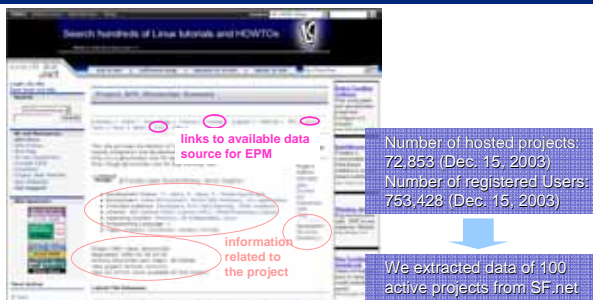


## Application of EPM to Open Source Software

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Data Source: SourceForge.net



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Collaboration Tools on SF.net

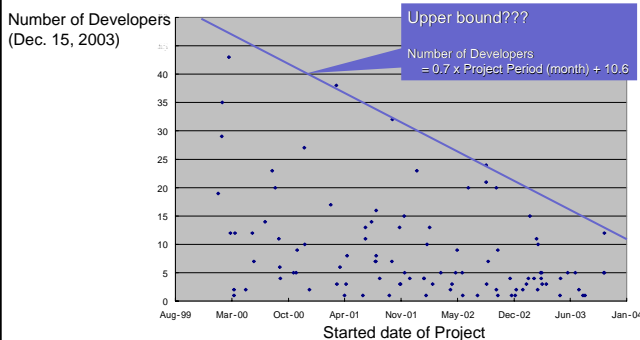
- SourceForge Collaborative Development System (CDS) web tools
- Project Web Server
- Tracker: Tools for Managing Support
- Mailing lists and discussion forums
- MySQL Database Services
- Project CVS Services
- ...

Available data source for EPM

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



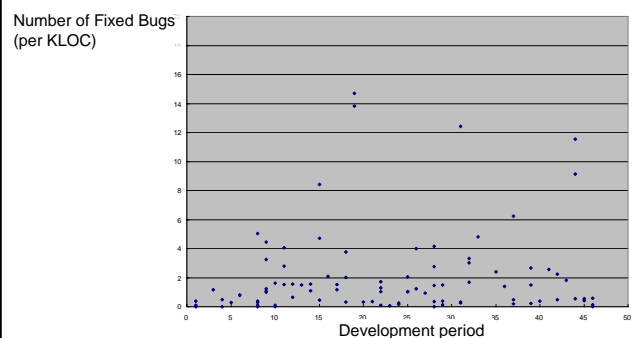
## Summary of 100 Active Projects @SF.net



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



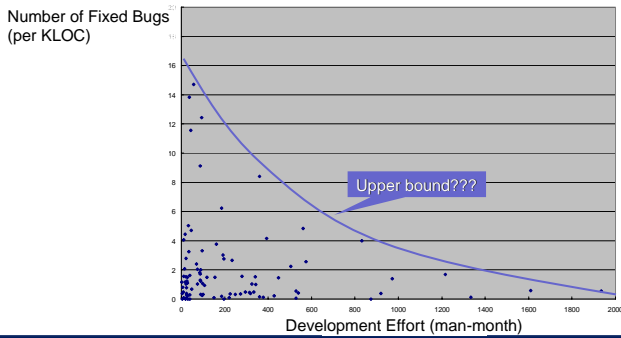
## Number of Fixed Bugs vs. Development period



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



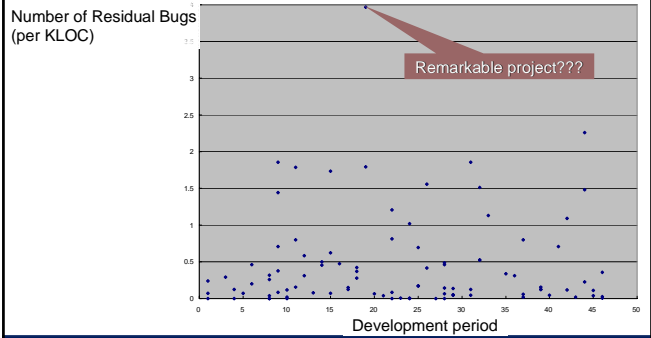
## Number of Fixed Bugs vs. Development Effort



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



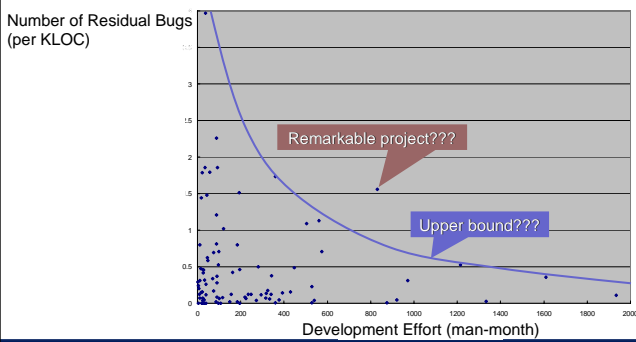
## Number of Residual Bugs vs. Development period



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Number of Residual Bugs vs. Development Effort



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Data Analysis on EPM

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## How can we use such a lot of data?

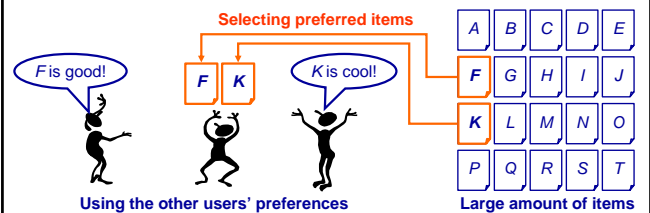


Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Collaborative Filtering

- **Filtering:** means selecting preferred items from a large collection of items.
- **Collaborative:** means using the other users' preferences to filter items.



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Two steps in CF

- Evaluate **similarities** between target user and the other users.
- Estimate the **preference** using the other users' preferences for target item and their similarities.

	Item 1	Item 2	Item 3	Item 4	Item 5	
User A	5 (prefer)	5 (prefer)	1 (not prefer)	3 (even)	5 (prefer)	Estimate
User B	5 (prefer)	5 (prefer)	1 (not prefer)	3 (even)	5 (prefer)	Similar User
User C	5 (prefer)	5 (prefer)	1 (not prefer)	5 (prefer)	5 (prefer)	Similar User
User D	1 (not prefer)	1 (not prefer)	3 (even)	5 (prefer)	1 (not prefer)	Dissimilar User

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Estimation/Prediction by CF

- We can estimate or predict important attributes of an ongoing project using the empirical data of past projects.

	Java	C++	LOC	Complexity	Developers' Skill	# of Faults	
Project A	0.58	-0.60	-0.60	0.82	-1.00	-0.72	Estimate/Predict
Project B	0.58	-0.60	-0.52	0.00	-1.00	-0.56	Similarity: 0.82
Project C	0.58	-0.60	-0.64	0.82	-1.00	-0.84	Similarity: 0.99
Project D	-1.70	1.73	1.73	-1.63	1.7	1.40	Similarity: -0.98

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Similarity Evaluation by CF

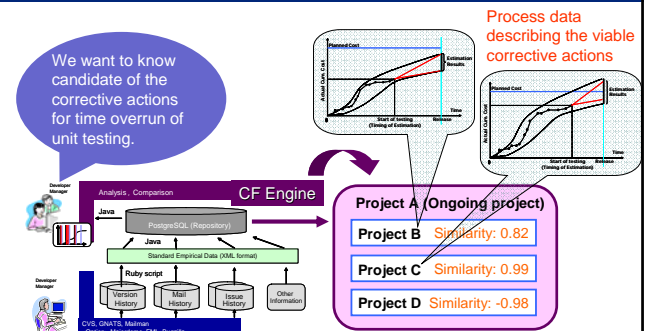
- We can find similar past projects to an ongoing project, based on the similarity estimation.

	Java	C++	LOC	Complexity	Developers' Skill	# of Faults	
Project A	0.58	-0.60	-0.60	0.82	-1.00	? (target)	
Project B	0.58	-0.60	-0.52	0.00	-1.00	-0.56	Similarity: 0.82
Project C	0.58	-0.60	-0.64	0.82	-1.00	-0.84	Similarity: 0.99
Project D	-1.70	1.73	1.73	-1.63	1.7	1.40	Similarity: -0.98

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Clue to More Concrete Project Data



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



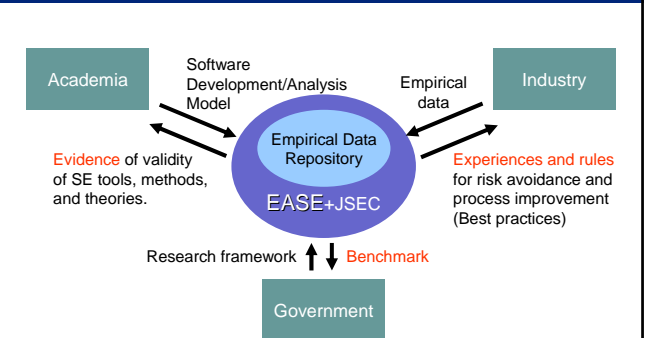
## Conclusion

- This talk is about
  - Current status of software industry in Japan
  - Empirical software engineering
  - EASE project and EPM
- In EASE Forum held in Nov. 2003, more than 140 companies claimed that they would like to use an empirical environment if available.
- 7 companies are now applying EPM alpha-version to software development projects in their organizations, and we will publish the results of these first trials on EPM in March 2005.

Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.



## Vision of EASE project in 2007



Keynote@International Workshop on Computer-Supported Knowledge Collaboration, Shanghai, July 7, 2004.

