Infrastructure, Data Cleansing and Mining for Scientific Simulations

Yingping Huang

Committee Members:

Dr. Bowyer
Dr. Flynn
Dr. Madey
Dr. Uhran

Agenda

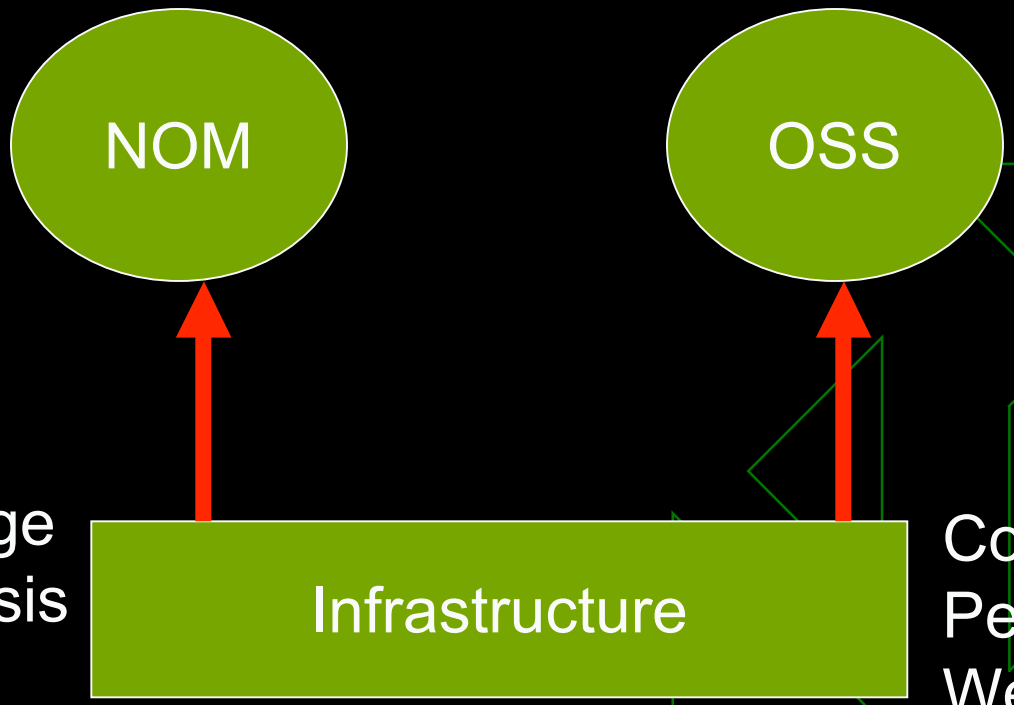
- **Overview**
- Background
- Multi-tier infrastructure
- Data cleansing algorithms
- Data mining applications
- Summarize
- Timeframe

Overview

- Multi-tier infrastructure powers scientific simulations.
- Data cleansing algorithms result in better data quality.
- Data mining applications discover hidden knowledge in environmental and social science.

Motivation

Simulation Anytime and Anywhere



Data Storage
Data Analysis
Reports

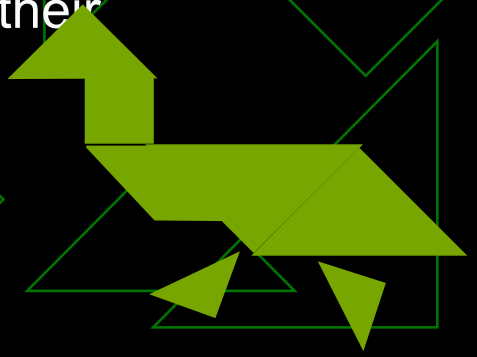
Collaboration
Personalization
Web-based

Agenda

- Overview
- **Background**
- Multi-tier infrastructure
- Data cleansing algorithms
- Data mining applications
- Summarize
- Timeframe

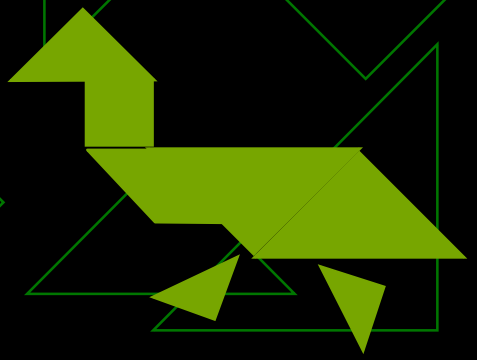
Background

- Projects under way
  - NOM
    - Research on natural organic matter (NOM)
    - Study evolution of NOM over time
    - Joint work of scientists across disciplines including chemists, biochemists, environmental scientists
  - OSS
    - Research on the open source software (OSS) development phenomenon
    - Study the behavior of OSS developers and their motivations
    - Joint work with social scientists



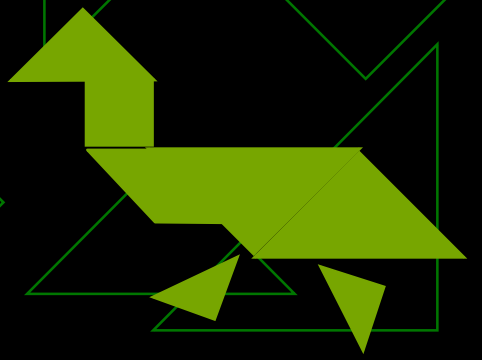
Simulation Models

- Standalone or traditional client-server
  - Software needs to be installed on clients
  - Incompatibility makes installation difficult
- Web-based using applets
  - Security – file permission, firewall
  - Inconvenience – plug-ins download
  - Network traffic – download before executing
  - Incompatibility – Swarm
- What should be done?
  - Web-based server-side simulation models
  - Centralized simulation management
  - Collaboration and personalization



Data Cleansing

- Known approaches
  - Sorted neighborhood (Stolfo 1995/1998)
    - Domain dependent keys for sorting
  - Record matching(Monge, 2000)
    - Edit distance only
  - String mapping (Li, 2003)
    - Potential high dimensional target space
- Our approaches
  - Sample database
  - Lipschitz mapping



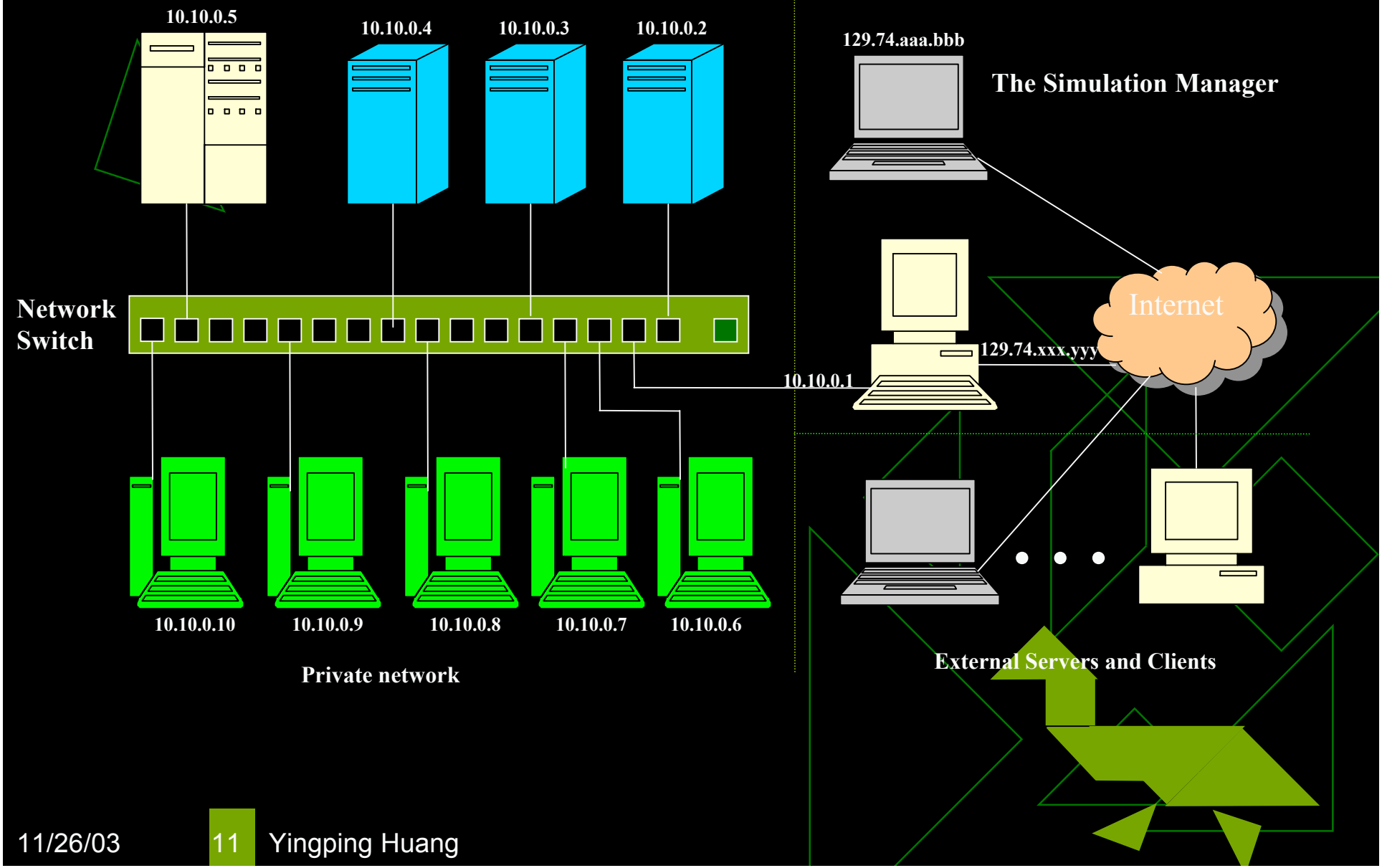
Data Mining

- Data mining in astronomy
  - SKYCAT: star/galaxy classification (Fayyad, 1996)
  - JARTool: detect volcanoes on Venus (Burl, 1998)
  - Sapphire: find galaxies (Kamath, 2001)
- Data mining in biology
  - Bioinformatics
  - SARS diagnosis (ehealth.org)
- What should be done?
  - Data mining for social science (OSS)
  - Data mining for environmental science (NOM)
  - Add intelligence to simulation models by applying data mining results

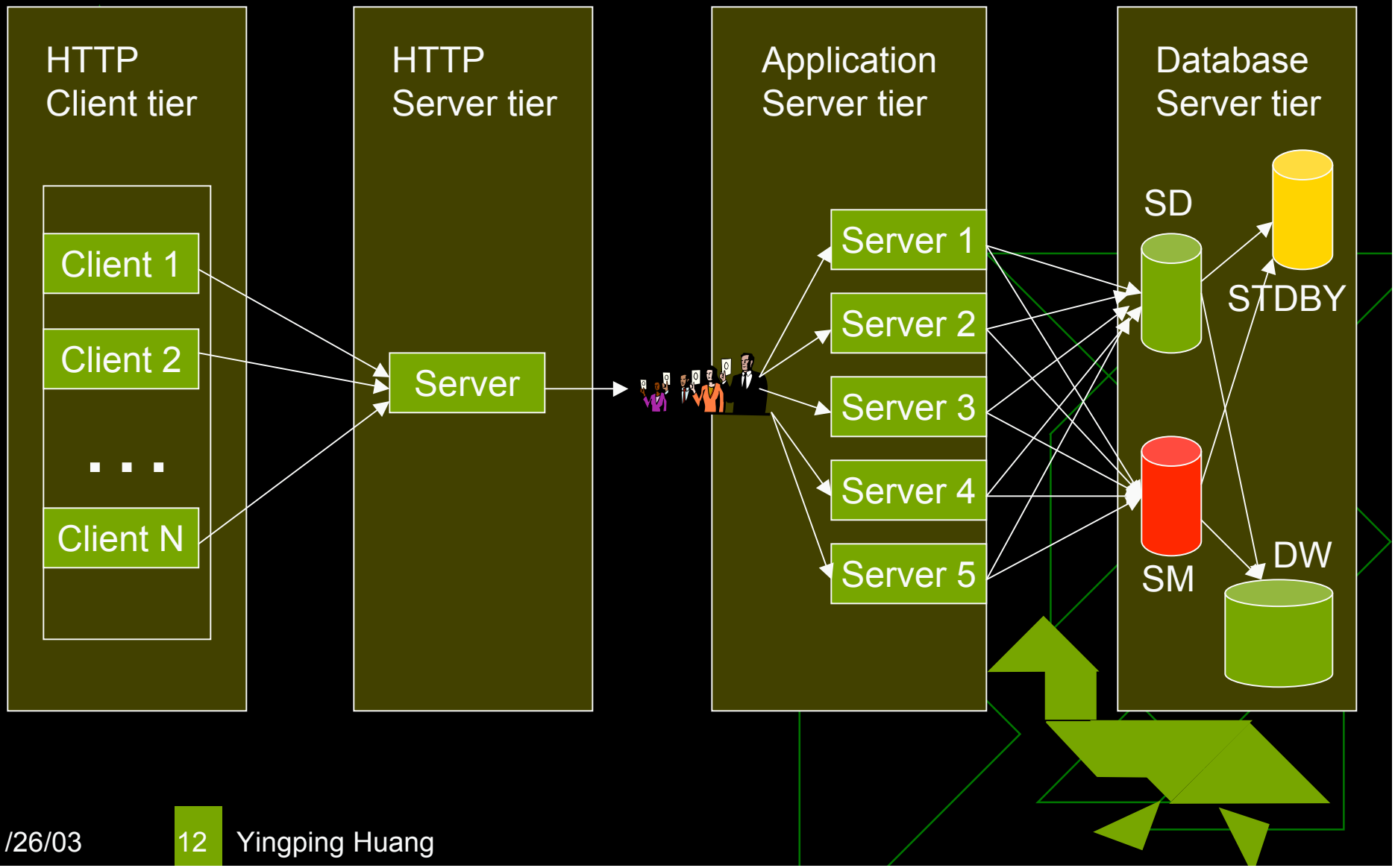
Agenda

- Overview
- Background
- **Multi-tier infrastructure**
- Data cleansing algorithms
- Data mining applications
- Summarize
- Timeframe

Physical Layout



Multi-tier Architecture



Two Features

- Load-balancing
  - Scalability achieved
  - Implementation using JMS, AQ & EJB
  - Implementation using Shell scripts & PL/SQL
- Simulation-resuming
  - Reliability achieved
  - Checkpoint
  - Implementation using JTA/JTS

Load-balancing Using JMS & AQ

job_id	resumed	checkpoint	status
100	0	0	

Job queue

Topic 1

mdb1 mdb2 mdb3 mdb4 mdb5

Invoke simulation & update job queue

Loadavg queue

mdb6 mdb7 mdb8 mdb9 mdb10

Judge bean

Topic 2

Shell Scripts & PL/SQL

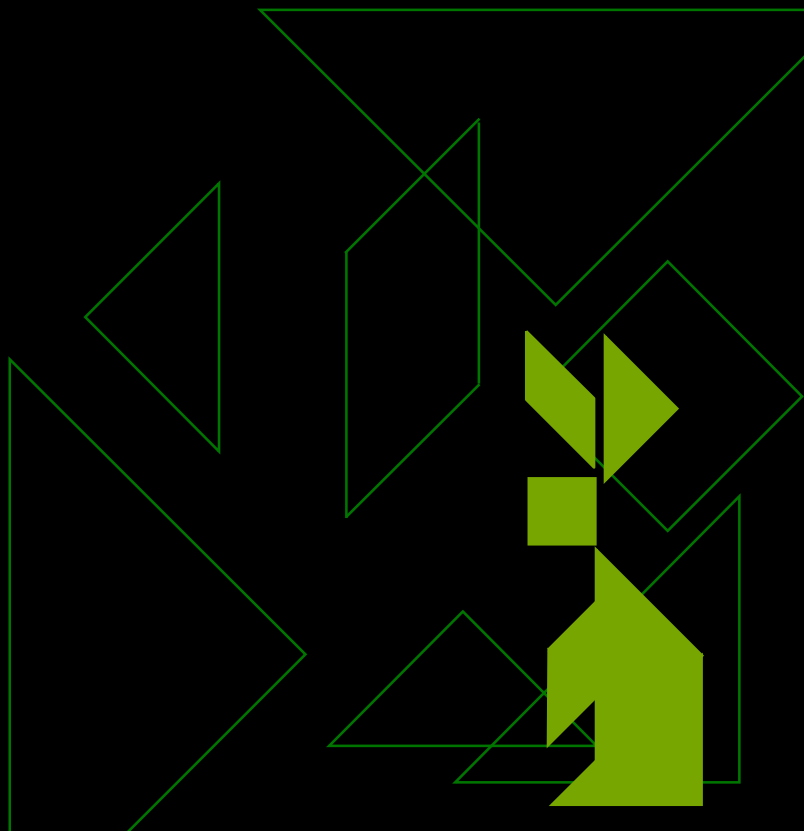
- Dispatcher (HTTP server)
  - Dispatch simulations
  - Send KEEPALIVE messages to running simulations
- Intelligent agent (application server)
  - Upload load averages
  - Check simulations
  - Send ACK to KEEPALIVE messages

Load-balancing Algorithm

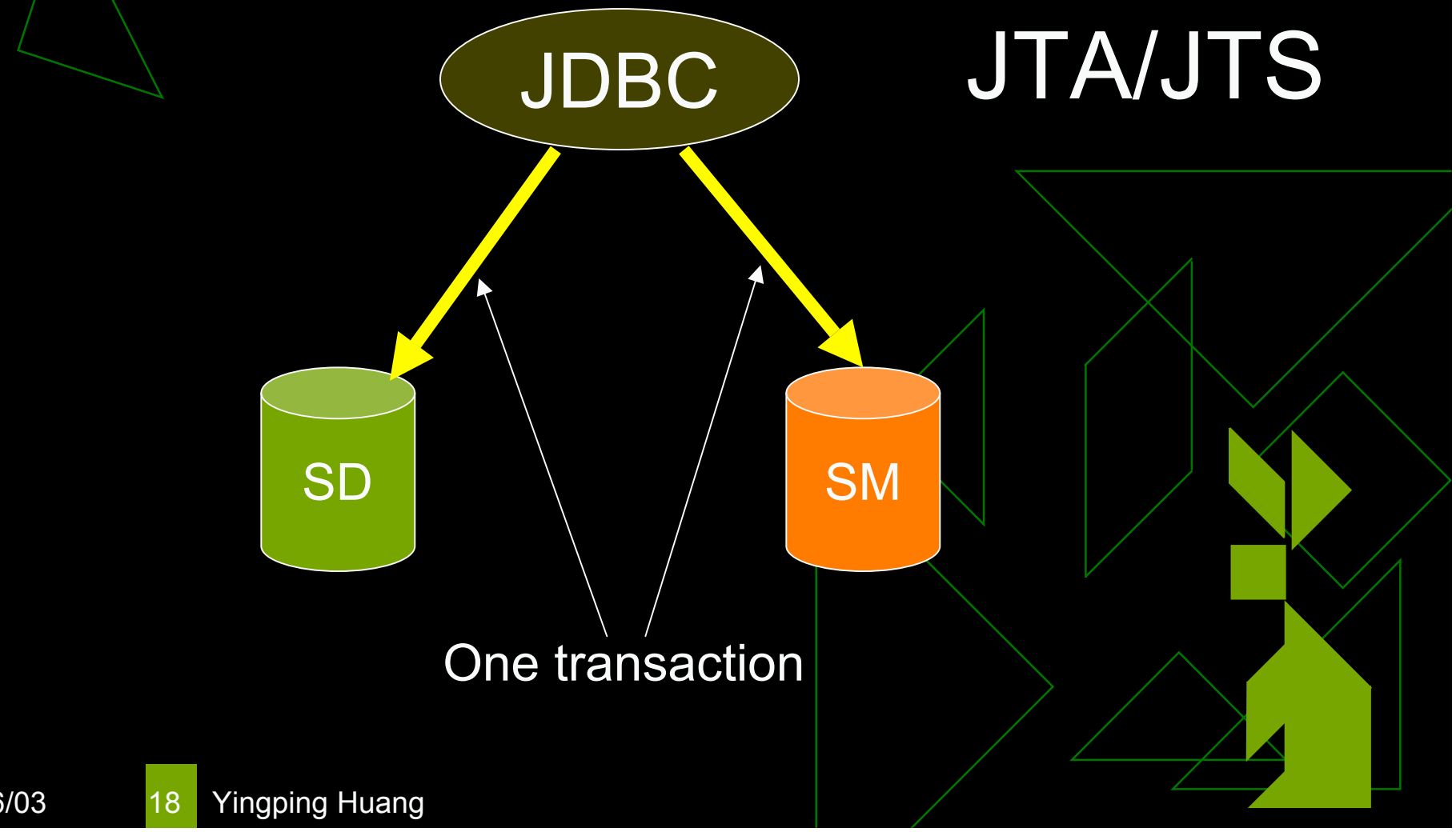
- Instance learning approach
  - Based on completion time prediction
- Two step completion time prediction
  - Completion time estimation
    - Load average
    - Data amount
  - Completion time prediction
    - Nearest neighborhood

Completion Time Estimation

- Completion time estimation formula



Checkpoint



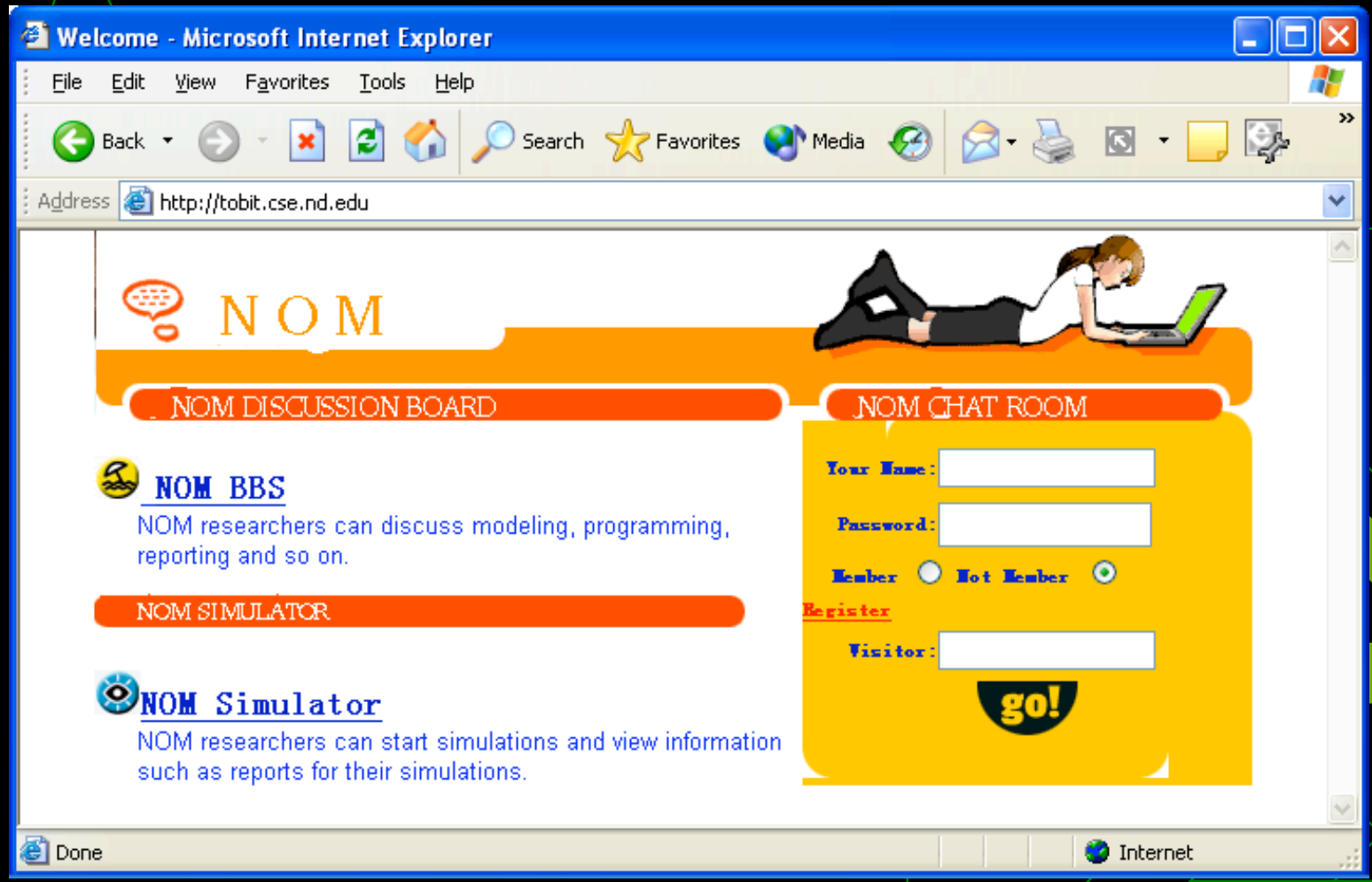
Checkpoint Issues

- Checkpoint data
  - All data for restarting the simulation
  - Size depends on number of agents
- Checkpoint frequency
  - Checkpoint-interval
    - # of MB data
  - Checkpoint-timeout
    - # of minutes

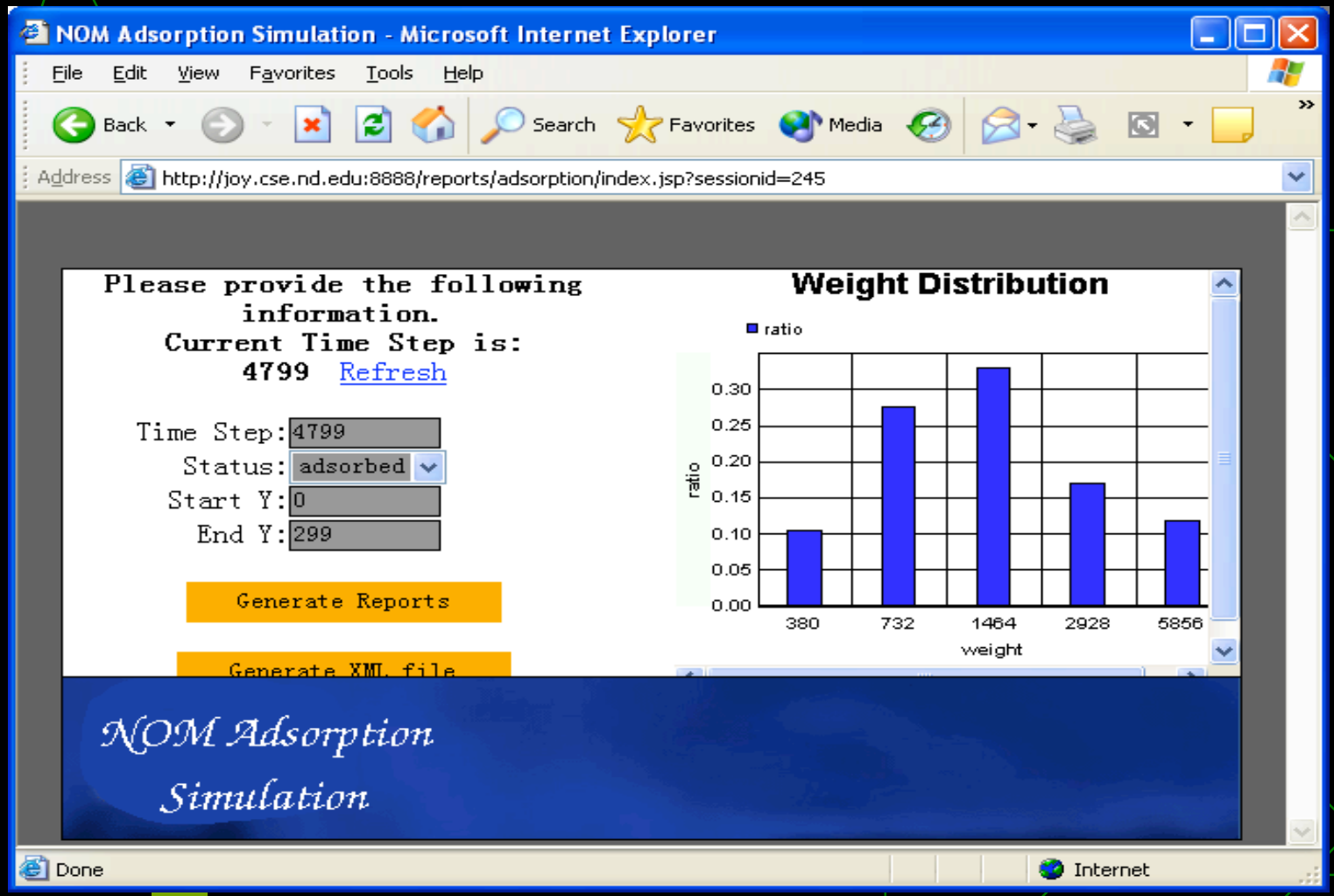
Simulation-resuming

- To restart a terminated simulation
  - A new simulation with same job_id inserted into the job queue
  - A terminated simulation has smaller job_id than new simulations, higher priority
- In case of application server failure
  - All simulations' job_ids inserted into the job queue
  - All simulations will be running on other application servers

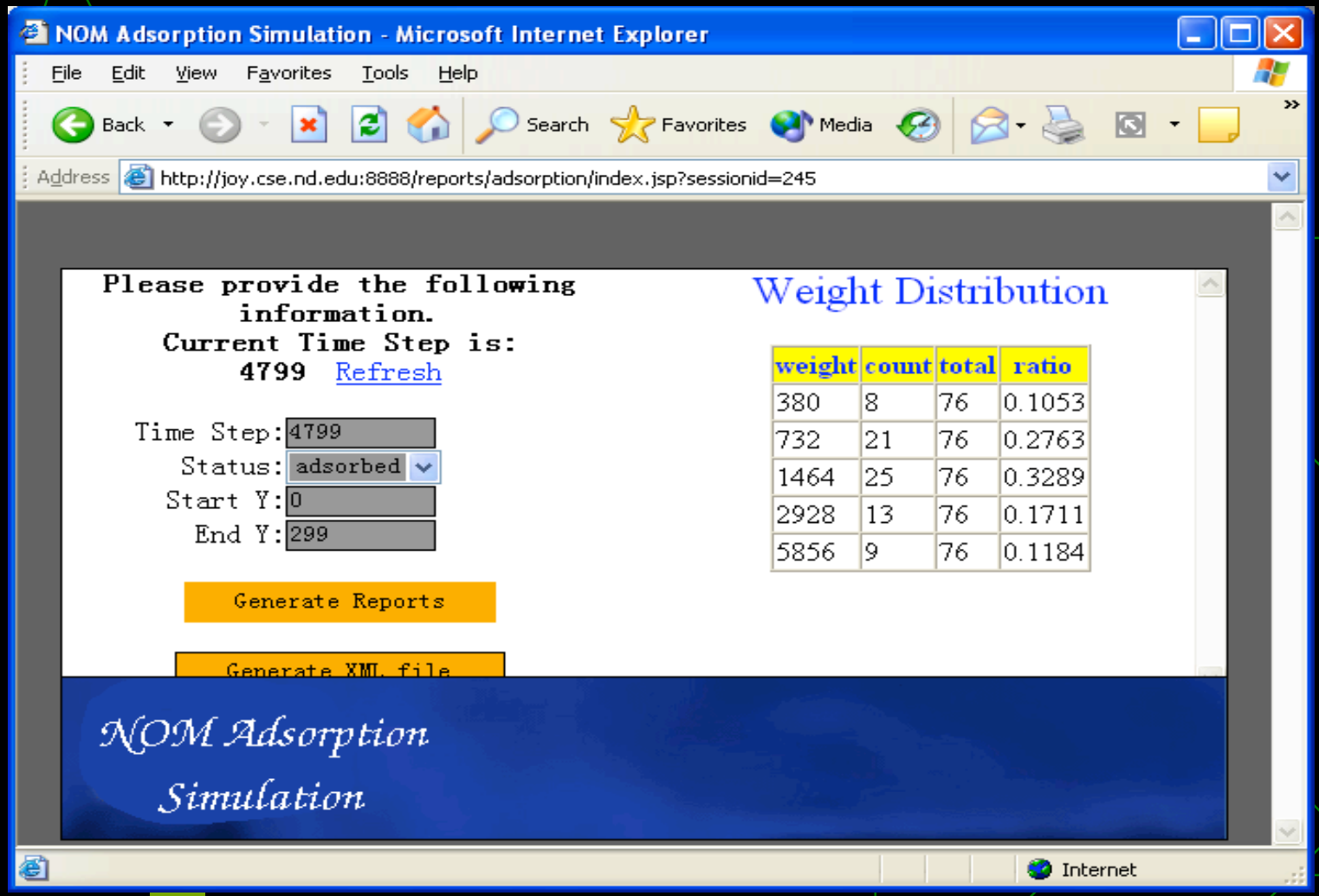
Collaboration Suite



Graphical Reports



XML Reports


NOM Adsorption Simulation - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://joy.cse.nd.edu:8888/reports/adsorption/index.jsp?sessionid=245
Please provide the following information.
Current Time Step is:
4799 Refresh
Time Step: 4799
Status: adsorbed
Start Y: 0
End Y: 299
Generate Reports
Generate XML file
Weight Distribution
weight count total ratio
380 8 76 0.1053
732 21 76 0.2763
1464 25 76 0.3289
2928 13 76 0.1711
5856 9 76 0.1184
NOM Adsorption Simulation
Internet

Agenda

- Overview
- Background
- Multi-tier information system
- **Data mining applications**
- Summarize
- Timeframe

Methodology

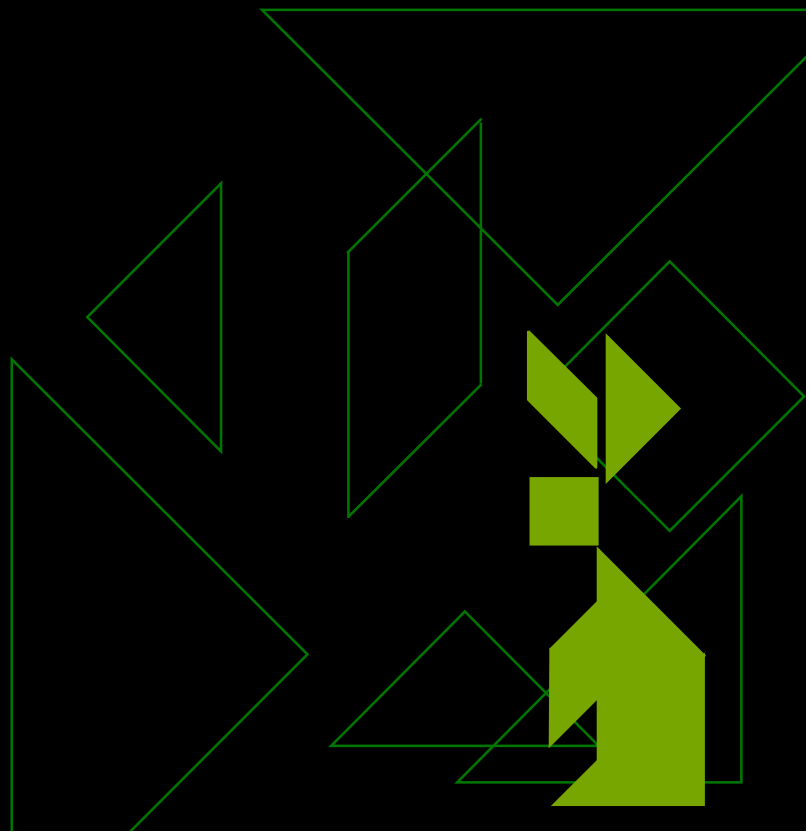
- Traditional approach
  - Form hypotheses
  - Verify hypotheses by finding patterns in data
- Data mining approach
  - Find patterns in data
  - Form hypotheses
  - Design simulation models
  - Verify hypotheses
- U. Fayyad, J. Gray at Microsoft Research

Technology & Software

- Data mining technology
  - Clustering
    - K-means
    - Orthogonal cluster
  - Classification
    - Decision tree
    - Naïve Bayes
  - Association rules
    - Apriori
- Data mining software
  - Oracle Data Mining Suite
  - DM4J
  - JDeveloper

OSS

- Study behavior of open source software (OSS) developers
  - Agent-based
  - Stochastic
- Data mining involving
  - Clustering
  - Classification
    - Churn prediction
    - Acquisition prediction
  - Association rules

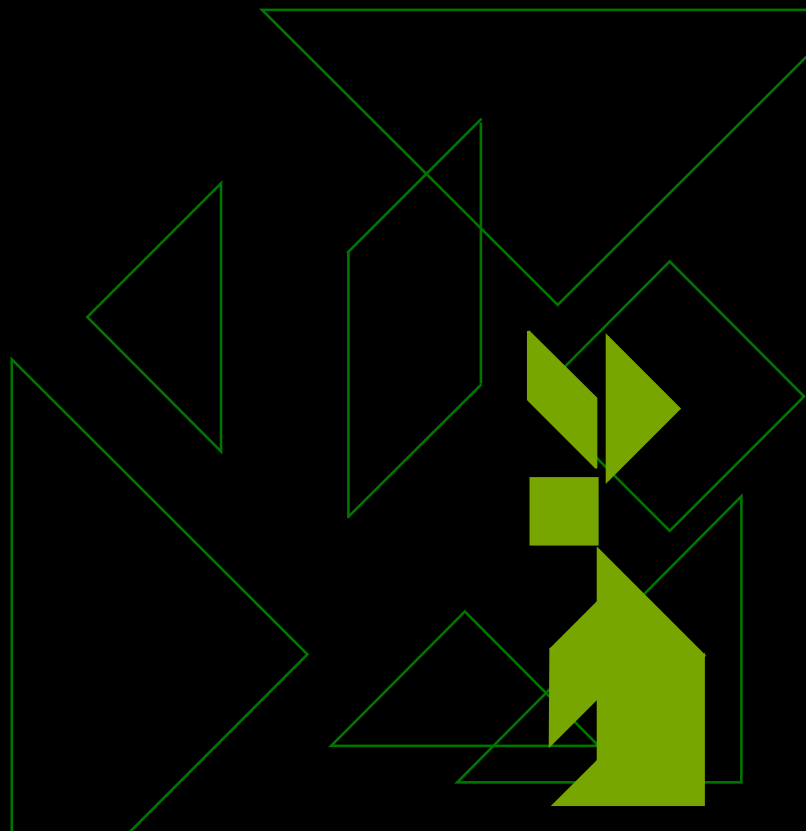


OSS Data Warehousing

- Data from sourceforge.com
  - Developers
  - Projects
- Data warehousing
  - Table partitioning
  - Aggregation
  - Star schema
  - Analysis SQL
  - ETL tools → Warehouse Builder

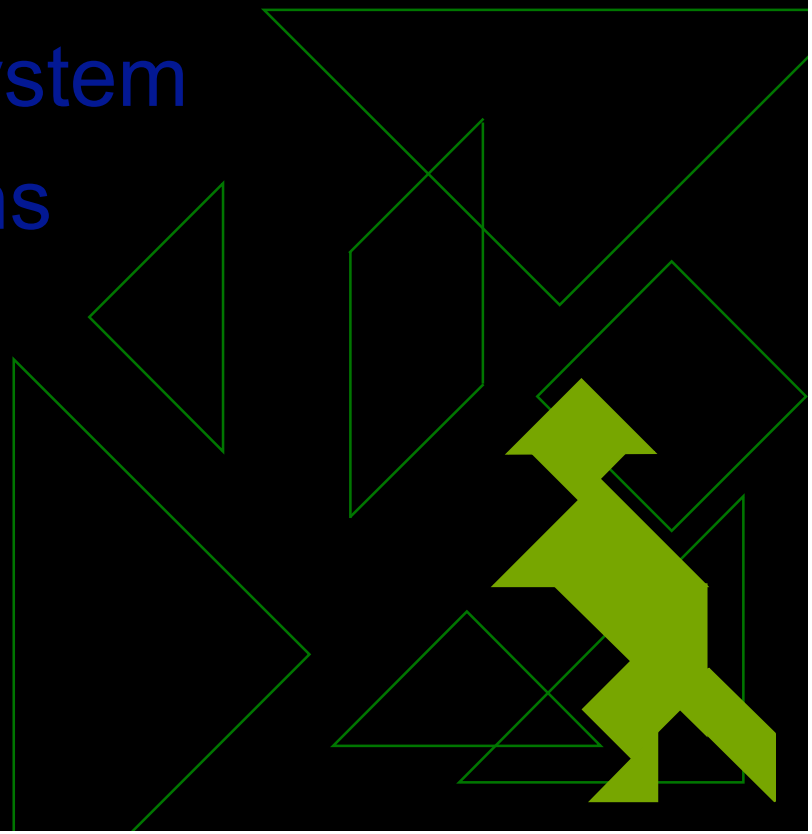
NOM

- Study behavior of natural organic matter (NOM)
  - Agent-based
  - Stochastic
- Data mining involving
  - Clustering
    - Micelle formation
  - Classification
    - Transportation prediction
    - Adsorption prediction
  - Association rules



Agenda

- Overview
- Background
- Multi-tier information system
- Data mining applications
- **Summarize**
- Timeframe



Summarize

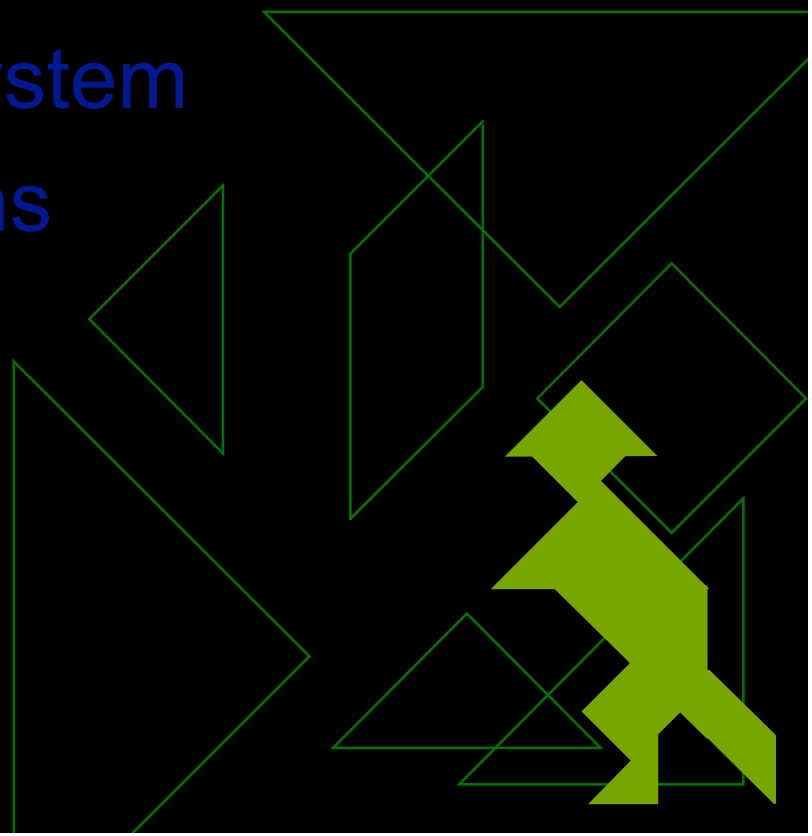
- Multi-tier information system integrates
  - Application servers & reports server
  - Database servers
  - Data warehousing & data mining
  - Swarm
- Collaboration suite
- Data mining guided model-design

Insights & Impacts

- Server-side simulation models
  - Centralized simulation management
  - Centralized data repository
- Collaboration suite
  - Simulation sharing
  - Knowledge sharing
- Data mining applications
  - Find patterns in data
  - Model deployment for simulation-design

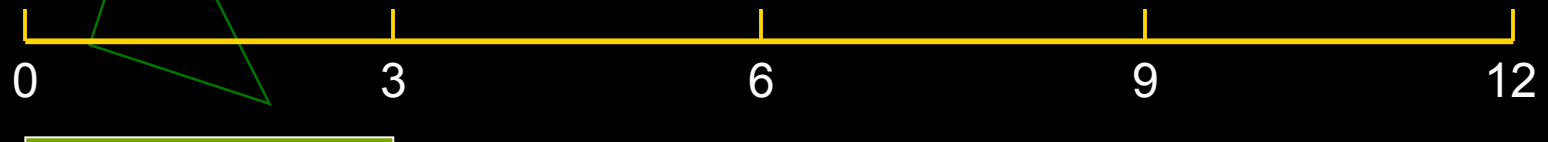
Agenda

- Overview
- Background
- Multi-tier information system
- Data mining applications
- Summarize
- **Timeframe**



Timeframe

May 2003 ~ May 2004



Implement infrastructure

Data collection & statistical analysis

Data mining model design

Data mining model evaluation

Deployment

Writing up

Expected Publications

- Information system design for scientific simulations
  - By August 2003
- Data warehousing for scientific simulations
  - By November 2003
- Data mining for OSS
  - By February 2004
- Data mining for NOM
  - By March 2004

Demo

Demonstration

Finally

Thank you!

Features

- Multi-tier information system
  - HTTP client tier → HTTP server tier →Application server tier →EIS tier
- Scalability at the application server tier
  - Load-balancing
- Reliability at the application server tier
  - Simulation-resuming
- Reliability at the database tier
  - Standby databases

Features (cont.)

- Data mining models
  - Stored in database
  - Stored Java procedures
  - PL/SQL procedure call using JDBC
- Simulation models
  - Agent-based
  - Stochastic
  - Data mining guided