PURDUE UNIVERSITY

CS37300: Data Mining and Machine Learning

Data Mining Process

Profs. Steve Hanneke and Chris Clifton

24 March 2023

Thanks to Laura Squier, SPSS for some of the material used

PURDUE UNIVERSITY

Data Mining as a *Process*

- Data mining involves many steps
  - Machine learning is only one aspect
  - Data exploration/understanding, evaluation, etc.
- This needs to be formalized so it is more science than art
  - Steps and tasks involved
- One approach: Process Model
  - Formalize steps
  - Document what is to be done at each step

3

PURDUE UNIVERSITY

Data Mining Process

- Cross-Industry Standard Process for Data Mining (CRISP-DM)
- European Community funded effort to develop framework for data mining tasks
- Goals:
  - Encourage interoperable tools across entire data mining process
  - Take the mystery/high-priced expertise out of simple data mining tasks

4

PURDUE UNIVERSITY

Why Should There be a Standard Process?

The data mining process must be reliable and repeatable by people with little data mining background.

- Framework for recording experience
  - Allows projects to be replicated
- Aid to project planning and management
- "Comfort factor" for new adopters
  - Demonstrates maturity of Data Mining
  - Reduces dependency on "stars"

5

PURDUE UNIVERSITY

Process Standardization

- CRoss Industry Standard Process for Data Mining
- Initiative launched Sept.1996, document released Aug. 2000
- SPSS/ISL, NCR, Daimler-Benz, OHRA
- Funding from European commission
- Peaked at over 200 members of the CRISP-DM SIG
  - DM Vendors - SPSS, NCR, IBM, SAS, SGI, Data Distilleries, Syllogic, Magnify, ..
  - System Suppliers / consultants - Cap Gemini, ICL Retail, Deloitte & Touche, …
  - End Users - BT, ABB, Lloyds Bank, AirTouch, Experian, ...
- Support died out in late 2000''s (IBM purchase of SPSS in 2009)

6

PURDUE UNIVERSITY

CRISP-DM

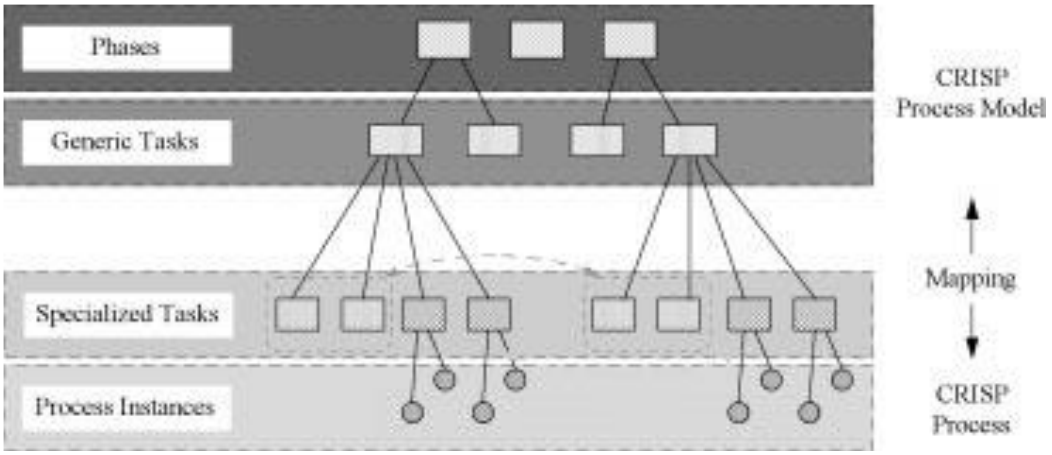
- Non-proprietary
- Application/Industry neutral
- Tool neutral
- Focus on business issues
  - As well as technical analysis
- Framework for guidance
- Experience base
  - Templates for Analysis

7

PURDUE UNIVERSITY

CRISP-DM: Overview

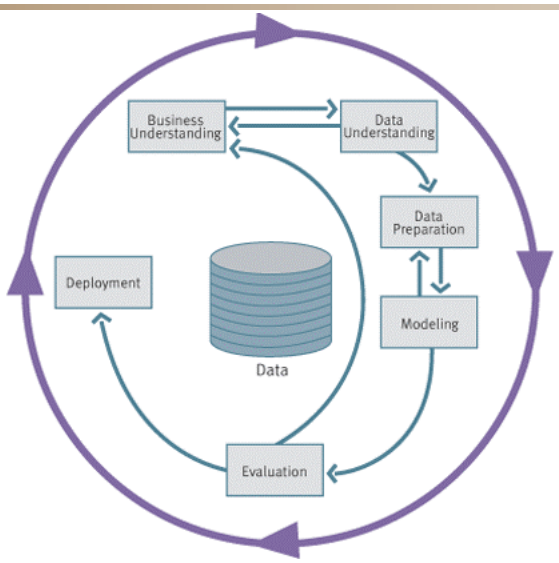
- Hierarchical Model



8

PURDUE UNIVERSITY

CRISP-DM: Phases



9

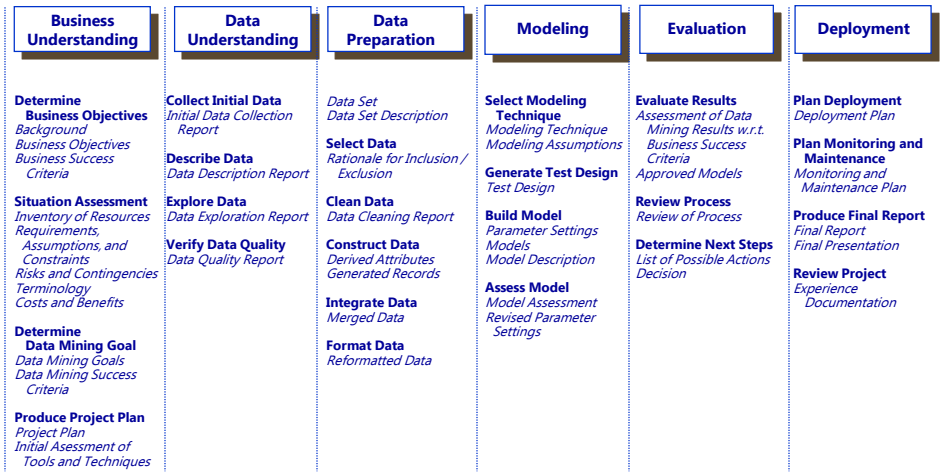
PURDUE UNIVERSITY

CRISP-DM: Phases

1. Business Understanding
   - Understanding project objectives and requirements
   - Data mining problem definition
2. Data Understanding
   - Initial data collection and familiarization
   - Identify data quality issues
   - Initial, obvious results
3. Data Preparation
   - Record and attribute selection
   - Data cleansing
4. Modeling
   - Run the data mining tools
5. Evaluation
   - Determine if results meet business objectives
   - Identify business issues that should have been addressed earlier
6. Deployment
   - Put the resulting models into practice
   - Set up for repeated/continuous mining of the data

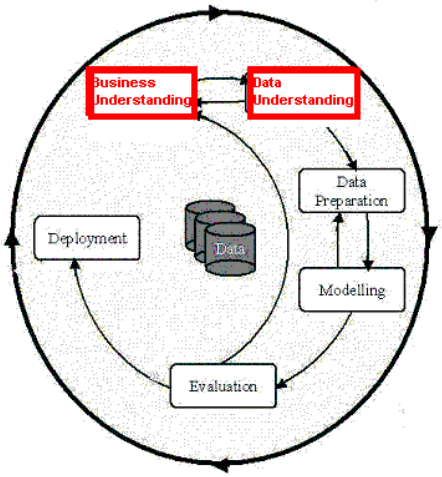
PURDUE UNIVERSITY

Phases and Tasks



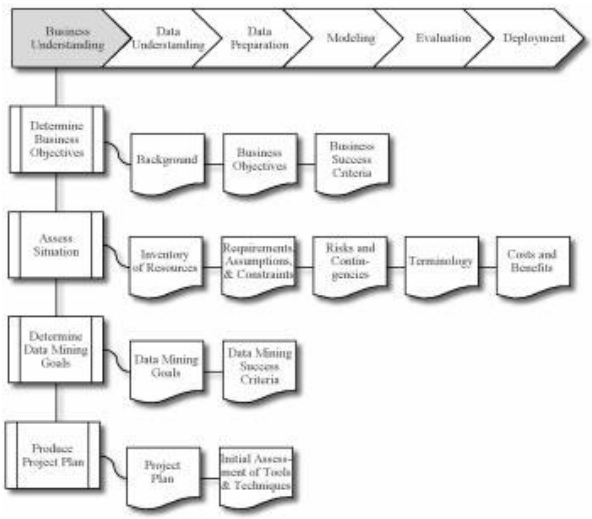
Phase 1: Business Understanding

- Business Understanding:
  - Statement of Business Objective
  - Statement of Data Mining objective
  - Statement of Success Criteria



13

Business Understanding



- Determine Business Objectives
  - Background, Objectives, Success Criteria
- Assess Situation
- Determine Data Mining Goals
  - Success Criteria
- Produce Project Plan

14

PURDUE UNIVERSITY

Business Understanding: Determine Business Objectives

Activities:

- Develop organizational charts identifying divisions, departments and project groups. The chart should also identify managers' names and responsibilities.
- Identify key persons in the business and their roles.
- Identify an internal sponsor (financial sponsor and primary user/domain expert).
- Is there a steering committee and who are the members?
- Identify the business units which are impacted by the data mining project (e.g., Marketing, Sales, Finance)

Current solution

- Describe any solution currently in use for the problem.
- Describe the advantages and disadvantages of the current solution and the level to which it is accepted by the users.

Problem area:

- Identify the problem area (e.g., Marketing, Customer Care, Business Development, etc.).
- Describe the problem in general terms.
- Check the current status of the project (e.g., Check if it is already clear within the business unit that we are performing a data mining project or do we need to advertise data mining as a key technology in the business?).
- Clarify prerequisites of the project (e.g., what is the motivation of the project? Does the business already use data mining?).
- If necessary, prepare presentations and present data mining to the business.
- Identify target groups for the project result (e.g., Do we expect a written report for top management or do we expect a running system that is used by naive end users?).
- Identify the users' needs and expectations.

15

PURDUE UNIVERSITY

Business Understanding: Assess Situation

- Inventory of Resources
- Requirements Assumptions & Constraints
- Risks and Contingencies
- Terminology
- Costs and Benefits

16

PURDUE UNIVERSITY

Business Understanding: Project Plan

- Stages of the project
  - Schedule
  - Resources
  - Dependencies
- Assessment of Tools and Techniques
- "Living Document"
  - Specific points for review/update

17

PURDUE UNIVERSITY

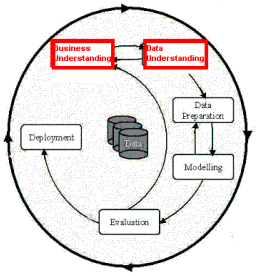
Business Understanding: Phase Report

- Background
- Business objectives and success criteria
- Inventory of resources
- Requirements, assumptions, and constraints
- Risks and contingencies
- Terminology
- Costs and benefits
- Data mining goals and success criteria
- Initial assessment of tools and techniques

18

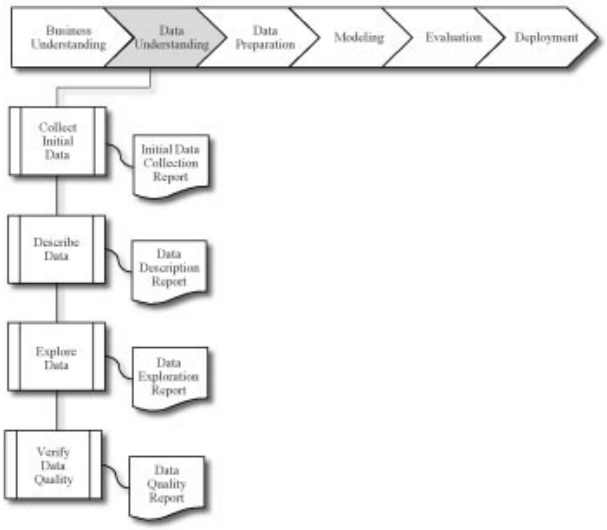
Phase 2: Data Understanding

- Business Understanding:
  - Statement of Business Objective
  - Statement of Data Mining objective
  - Statement of Success Criteria



- Data Understanding
  - Explore the data and verify the quality
  - Find outliers

Data Understanding



- Collect Initial Data
- Describe Data
- Explore Data
- Verify Data Quality

Report at each stage

  - Capture information to ensure repeatability of process

PURDUE UNIVERSITY

Data Understanding: Data Description Report

- Format of data
- Quantity of data
- Identity of fields, other surface features

Does the data acquired satisfy requirements?

22

PURDUE UNIVERSITY

Data Understanding: Explore Data

- We've covered data exploration
  - Distribution, pairwise correlations, sub-populations
- Outcome
  - Need for further transformation/preparation?
  - Is quality sufficient for goals?
  - Initial findings / hypotheses

23

PURDUE UNIVERSITY

Data Understanding: Verify Data Quality

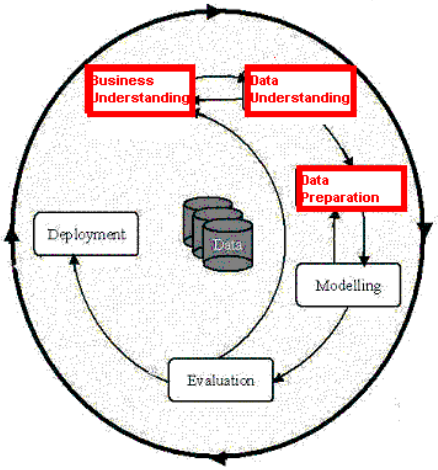
- Completeness
- Correctness
  - Random errors
  - Systematic errors
  - Missing values
- Potential solutions

24

PURDUE UNIVERSITY

Phase 3: Data Preparation

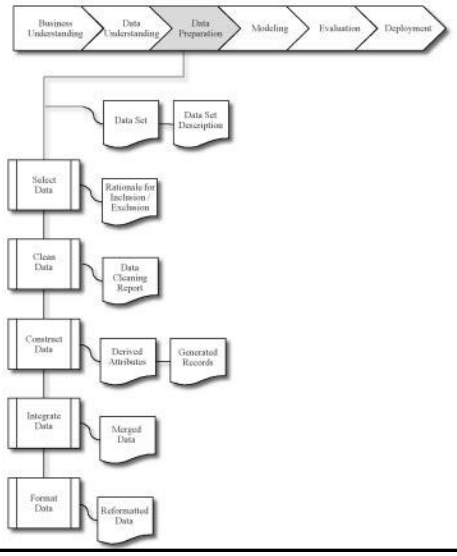
- Data preparation:
- Takes usually over 90% of the time
  - Collection
  - Assessment
  - Consolidation and Cleaning
    - table links, aggregation level, missing values, etc
  - Data selection
    - active role in ignoring non-contributory data?
    - outliers?
    - Use of samples
    - visualization tools
  - Transformations - create new variables



25

PURDUE UNIVERSITY

Data Preparation



- Select Data
- Clean Data
- Construct Data
- Integrate Data
- Format Data

Output: Dataset and Dataset Description

  – Also reports on each stage

PURDUE UNIVERSITY

Data Preparation: Select Data

- Decide what to use for analysis
  - Data mining goals
  - Data quality
  - Technical constraints
- Report: Rationale for inclusion/exclusion

PURDUE UNIVERSITY

Data Preparation: Clean Data

- Where data quality insufficient, improve
  - Select only good subsets
  - Obtain better data
  - Modeling / imputation of values
- Report: Process
  - What has been done
  - How might this impact validity of results?

28

PURDUE UNIVERSITY

Data Preparation: Construct Data

- Feature construction
  - Document how this is done
- Generate records
  - E.g., will modeling technique require records for customers who have made no purchase during a year?

29

PURDUE UNIVERSITY

Data Preparation: Integrate Data

- Data may come from multiple sources
  - Often dissimilar
- Different types of data about same entities
  - Record linkage
- Similar information about different subsets of entities
  - Feature mapping
  - Duplicate elimination
- Data Aggregation

30

PURDUE UNIVERSITY

Data Preparation: Format Data

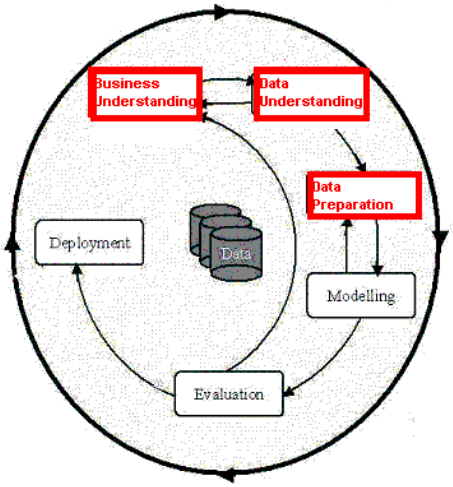
- (Primarily) syntactic modifications to satisfy tool requirements
  - Data format
  - Unique identifiers
- Normalization

31

PURDUE UNIVERSITY

Phase 4: Modeling

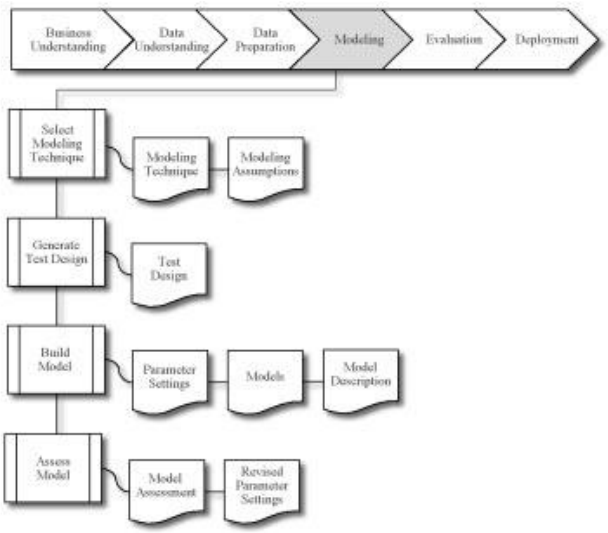
- Model building
  - Selection of the modeling techniques is based upon the data mining objective
  - Modeling is an iterative process - different for supervised and unsupervised learning
    - May model for either description or prediction



33

PURDUE UNIVERSITY

Modeling

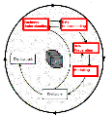


- Select Modeling Technique
- Generate Test Design
- Build Model
  - Capture parameters
- Assess Model

34

PURDUE UNIVERSITY

Types of Models

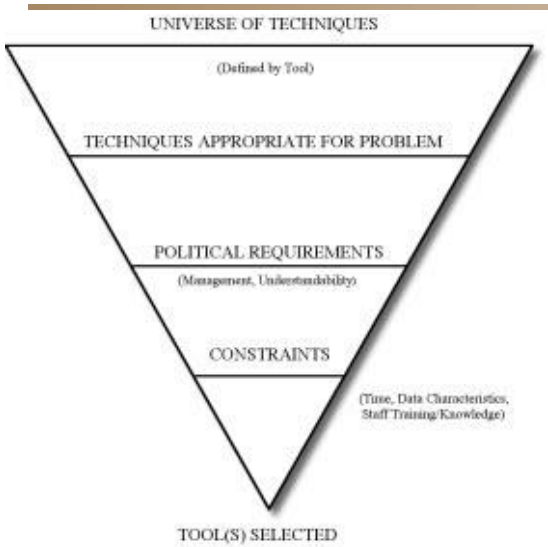


- Prediction Models for Predicting and Classifying
  - Regression algorithms (predict numeric outcome): neural networks, rule induction, CART (OLS regression, GLM)
  - Classification algorithm predict symbolic outcome): CHAID, C5.0 (discriminant analysis, logistic regression)
- Descriptive Models for Grouping and Finding Associations
  - Clustering/Grouping algorithms: K-means, Kohonen
  - Association algorithms: apriori, GRI

35

PURDUE UNIVERSITY

Modeling: Select Modeling Technique



- General task
- Specific tool
- Rationale

36

How to Choose a Data Mining System?

- Commercial data mining systems have little in common
  - Different data mining functionality or methodology
  - May even work with completely different kinds of data sets
- Need multiple dimensional view in selection
- Data types: relational, transactional, text, time sequence, spatial?
- System issues
  - running on only one or on several operating systems?
  - a client/server architecture?
  - Provide Web-based interfaces and allow XML data as input and/or output?

How to Choose a Data Mining System? (2)

- Data sources
  - ASCII text files, multiple relational data sources
  - support ODBC connections (OLE DB, JDBC)?
- Data mining functions and methodologies
  - One vs. multiple data mining functions
  - One vs. variety of methods per function
    - More data mining functions and methods per function provide the user with greater flexibility and analysis power
- Coupling with DB and/or data warehouse systems
  - Four forms of coupling: no coupling, loose coupling, semitight coupling, and tight coupling
    - Ideally, a data mining system should be tightly coupled with a database system

How to Choose a Data Mining System? (3)

- Scalability
  - Row (or database size) scalability
  - Column (or dimension) scalability
  - Curse of dimensionality: it is much more challenging to make a system column scalable that row scalable
- Visualization tools
  - “A picture is worth a thousand words”
  - Visualization categories: data visualization, mining result visualization, mining process visualization, and visual data mining
- Data mining query language and graphical user interface
  - Easy-to-use and high-quality graphical user interface
  - Essential for user-guided, highly interactive data mining

39

Examples of Data Mining Systems

- Python
  - Full programming environment
  - Highly extensible
  - Packages for high performance (e.g., PyTorch)
  - Open source
- R
  - Solid statistical basis
  - Extensive packages and visualization toold
  - Open source (GPL)
- Commercial Tools (rapidmine, SAS, …)
  - Typically more built-in data preprocessing
  - Graphical programming interfaces

40

Modeling: Generate Test Design

- What are the metrics?
  - Success metrics
  - Confidence in that metric
- What data is needed to reliably evaluate?
  - Type
  - Test/validation/?
  - Quantity to satisfy requirements

43

Modeling: Assess Model

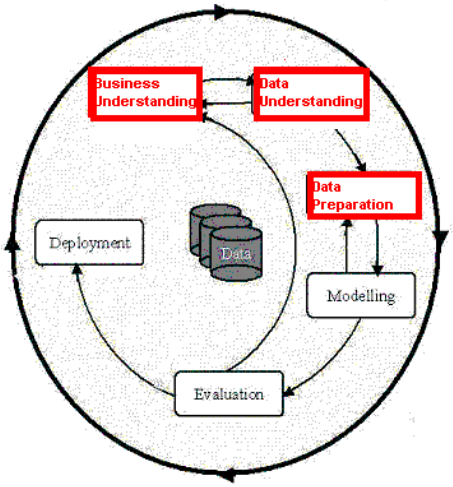
- How does it fair on success metrics?
- Domain expert analysis
  - Does it make sense?
- Rank models
  - What will help business objective?
- *Iterate modeling process*
  - Does this invalidate your success metrics?

44

PURDUE UNIVERSITY

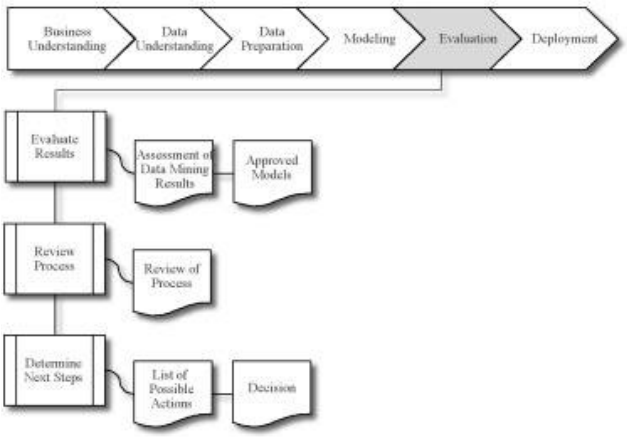
Phase 5: Evaluation

- Model Evaluation
  - Evaluation of model: how well it performed on test data
  - Methods and criteria depend on model type:
    - e.g., coincidence matrix with classification models, mean error rate with regression models
  - Interpretation of model: important or not, easy or hard depends on algorithm



PURDUE UNIVERSITY

Evaluation



- Evaluate Results
- Review Process
  - Anything missed?
  - Quality assurance
  - Compliance
- Determine Next Steps

PURDUE UNIVERSITY

Evaluation: Evaluate Results

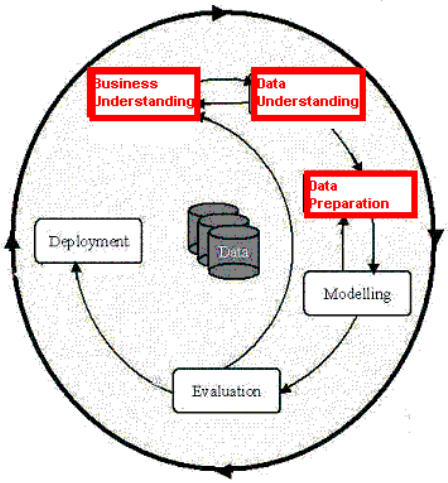
- Does model meet business objectives?
- Test on real applications
- Findings of interest that may not relate to business objectives

48

PURDUE UNIVERSITY

Phase 6: Deployment

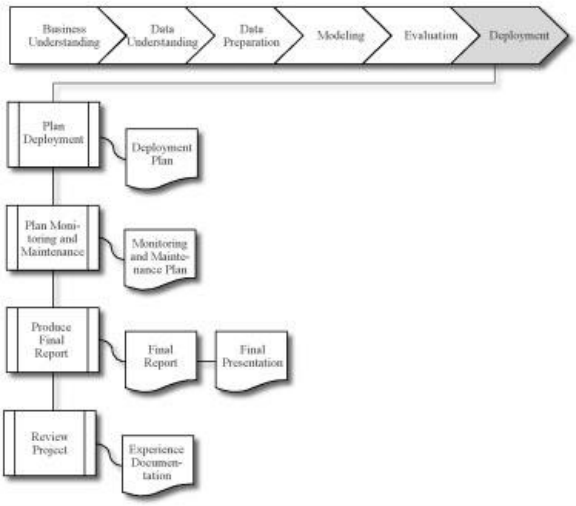
- Deployment
  - Determine how the results need to be utilized
  - Who needs to use them?
  - How often do they need to be used
- Deploy Data Mining results by:
  - Scoring a database
  - Utilizing results as business rules
  - interactive scoring on-line



49

PURDUE UNIVERSITY

Deployment



- Plan Deployment
- Plan Monitoring and Maintenance
- Produce Final Report
  - Written report
    - Include (and update) previous deliverables
  - Presentation
- Review Project
  - Document experience

50

PURDUE UNIVERSITY

Deployment: Plan Deployment

This is where projects typically fail!

- Do outcomes fit within existing business processes?
  - If not, what does it take to change processes?
- What might go wrong?
  - Are contingency plans needed?
- Cost of Deployment

51

PURDUE UNIVERSITY

Deployment: Plan Monitoring and Maintenance

- Model update
  - Process to ensure correctness over time
- Are business objectives being satisfied?
- Unanticipated impacts?

52

PURDUE UNIVERSITY

Why CRISP-DM?

- The data mining process must be reliable and repeatable by people with little data mining skills
- CRISP-DM provides a uniform framework for
  - guidelines
  - experience documentation
- CRISP-DM is flexible to account for differences
  - Different business/agency problems
  - Different data

53