

## Index

*Note:* Page numbers of article titles are in **boldface** type.

### A

---

- Abstraction methods, in temporal data mining, 91–94
- Affymetrix, 130–131
- Aggregation, of data, 65
- Akaike's Information Criterion, 137
- Allen's temporal language, 95
- Analysis of variance models (ANOVA), 152–153
- Analytical processing systems, 57, 59–60
- Apriori algorithm
  - in clustering analysis, 17
  - in temporal data mining, 94–95
- Association learning methods, 94
- Association rules, 121–123, 125
- Average link, for clustering analysis, 19–20

### B

---

- Backpropagation neural networks, 27–29
- Basic Alignment Search Tool, 132
- Biomarker development, data mining for, **127–143**
  - genomic, 162–163
  - public transcriptomic databases for, 128–131
    - Cancer Genome Anatomy Project, 129–130
    - Massively Parallel Signature Sequencing data, 130
    - SymAtlas V1.2.4, 130–131
    - Unigene, 128–129
  - tissue-specific expression-profiling methods for, 131–139
    - Akaike's Information Criterion, 137
    - ExQuest, 133–134
    - Gene Expression Profiling in silico (GEPIS), 134–135
    - ROKU, 138–139

- selective expression, 131–132
- Shannon Entropy, 135–137
- tissue selectivity, 137–138
- TissueInfo, 132–133

### C

---

- C4.5 tool, 38
- Cancer Biomedical Informatics Grid, 111
- Cancer Genome Anatomy Project, 129–130
- CDC Wonder system, 107
- Centers for Disease Control databases of, 107
  - National Nosocomial Infection Surveillance system, 120
- Chi-square method, 57–58
- Claims data, 102–104
- Classification, of data, 65
- Classification modeling, in genomic data mining, 156–161
- Classification trees, 25–27
- Clementine tool, 39
- Clinical Data Interchange Standards Consortium, 111
- Clinical Data Repository, University of Virginia, 66, 78–81
- Clustering analysis, 17–21, 153–156
- Codes, claim, 102–104
- Collaborative provider data warehouses, 105–106
- Community health information networks, 105
- Complete link, for clustering analysis, 19–20
- Confidentiality, of data warehouses, 65–67
- Cross-validation, for evaluation, 33

Current Procedural Terminology (CPT)  
codes, 102–104

## D

---

Data acquisition and processing, in data  
warehouses, 62–65

### Data mining, 9–35

canonical form lacking in, 5  
characteristics of, 4–6  
data quality in, 5  
databases for. *See* Data warehouses;  
Database(s).  
descriptive, for infection control,  
120–126  
dimensionality of, 4  
discovery techniques for, 11–21  
dual, 76–81  
ethical issues in, 6  
evaluation techniques for, 31–34  
for biomarker development, 127–143  
for genomic data, 145–166  
for infection control, 119–126  
generalization in, 6  
goals of, 10  
heterogeneity of, 4  
history of, 9–10  
imprecision of, 5  
information resources for, 4  
interpretation of, 5  
introduction to, 1–7  
legal issues in, 6  
mathematical characterization of, 5  
multi-database, 73–82  
open-source tools for, 37–54  
overview of, 2–4  
predictive techniques for, 21–31  
social issues in, 6  
temporal, 5–6, 83–100  
warehouses for, 55–71, 101–117

Data Mining Surveillance System, 120–126

### Data warehouses, 55–71

description of, 56–57  
development of, 58–67  
data acquisition and processing,  
62–65  
design, 58–62  
implementation, 62  
security in, 65–67  
examples of, 57–58  
performance issues with, 68–69  
regional and national, 101–117  
textual data in, 67–68  
user interfaces in, 69

Database(s). *See also* Data warehouses.  
multiple, 73–82  
regional and national, 101–117

Dendrograms, 17–20

Descriptive data mining, for infection  
control, 120–126

Diabetes Epidemiology Cohort, 105

Diabetes mellitus, Pima Indian data set for,  
14–15, 20–34

Discovery techniques, 11–21  
for observations, 16–21  
for variables, 12–16

Discrete wavelet transform, in temporal  
data mining, 88–89

Dual data mining, 76–81

Dynamic time warping, in temporal data  
mining, 87–88

## E

---

Ethical issues, in data mining, 6

Euclidean distance, in temporal data  
mining, 86–87

Evaluation techniques, 31–34

Expressed Sequence Tag, 128–136

ExQuest, 133–134

Extensible Markup Language, 111

Extraction, of data, 63

## F

---

False discovery rate, in genomic data  
mining, 148–149

Farr, William, 9–10

File Transfer Protocol, 111

Filtering, of data, 63

Food and Drug Administration database,  
107–108

Fourier transform, in temporal data mining,  
88–89

Frequent set and association rules  
techniques, 121–123, 125

## G

---

GE Healthcare Information Technologies,  
106

Gene Expression Profiling in silico (GEPIS),  
134–135

Gene voting, 158

Genomic data, data mining for, **145–166**  
 biomarkers, 162–163  
 challenges in, 145–148  
 classification modeling, 156–161  
 clustering analysis, 153–156  
 false discovery rate in, 148–149  
 pairwise statistical tests for, 149–151  
 pathway modeling, 161–162  
 statistical modeling on, 151–153

GGobi tool, 52

## H

---

Health Care Common Procedure Coding Systems, 102–103

Health Insurance Portability and Accountability Act, 65–66

Health Plan Employer Data and Information Set, 103

Heterogeneous error model, 153

Hierarchic clustering analysis, 17–21

## I

---

ICD (International Classification of Disease) codes, 102–104

Indexes, for data warehouses, 68–69

Infection control, data mining for, **119–126**  
 descriptive, 120–126  
 operational considerations in, 123–125  
 predictive, 120

Integrative pathway modeling, 162

International Classification of Disease codes, 102–104

Interval databases, 94–95

## K

---

Kaiser Permanente database, 104

Kernel functions, in support vector machines, 29–31, 159

KNIME tool, 48–51

Knowledge discovery in databases, 2

Konstanz Information Miner (KNIME), 48–51

## L

---

Least squares regression, 22–24

Legal issues, in data mining, 6

Linear discriminate analysis, 158, 160

Local-pooled-error test, 150–151

Logical Observation Identifier Names and Codes, 112

Logistic regression, 24–25, 158–159

Ludwig Institute for Cancer Research, 130

## M

---

Massively Parallel Signature Sequencing data, 130

Maximal margin hyperplane, 159

Medical Subject Headings (MeSH), 77–80

MEDLINE database, 77–80

MegaBLAST, 132, 134

Metrics, 33–34

Microarrays, significance analysis of, 149–150

MineSet tool, 39

Misclassification-penalized posterior criterion, 157, 160–161

MLC++ tool, 38–39

Multi-database data mining, **73–82**  
 applications of, 74–75  
 case study of, 78–81  
 for pattern identification, 76–78  
 methods for, 75–76

Multidimensional online analytical processing methods, 60, 62

Multilayer perceptrons, 27–29

## N

---

National databases. *See* Regional and national databases and data warehouses.

Neural networks, 27–29

Nosocomial infection control. *See* Infection control.

Nosocomial Infection Marker, 119–120

## O

---

Observations, discovery techniques for, 16–21

Online analytical processing systems, 57, 59–60

Online transaction processing systems, 56–57, 59–60, 62

Open-source tools, **37–54**  
 advantages and disadvantages of, 40–41  
 evolution of, 38–40  
 GGobi, 52  
 ideal characteristics of, 41–42  
 KNIME, 48–51  
 Orange, 50–52  
 R, 39–40, 42–43  
 Tanagra, 43–44  
 Weka, 39, 45–46  
 YALE, 46–48

Orange tool, 50–52

## **P**

---

Pairwise statistical tests, in genomic data mining, 149–151

Pathway modeling, in genomic data mining, 161–162

Pattern identification, in multi-database data mining, 76–78

Pearson correlation, in temporal data mining, 87

PedCath database, 106

Perceptrons, 27–29

Performance issues, with data warehouses, 68–69

PhenCode database, 73

Pima Indian data set, 14–15, 20–34

Predictive techniques, 21–31  
 classification trees, 25–27  
 for infection control, 120  
 least squares regression, 22–24  
 logistic regression, 24–25  
 neural networks, 27–29  
 support vector machines, 29–31

Principal components, 12–16

Proportion confidence interval, 57–58

Public health databases, 106–109

Public Health Information Network, 107

Public transcriptomic databases, for biomarker development, 128–131  
 Cancer Genome Anatomy Project, 129–130  
 Massively Parallel Signature Sequencing data, 130  
 SymAtlas V1.2.4, 130–131  
 Unigene, 128–129

## **Q**

---

Quadratic discriminate analysis, 158, 160

Qualitative pathway modeling, 161

Quantitative pathway modeling, 162

## **R**

---

R open-source tool, 39–40, 42–43

Receiver operating characteristic curve, 33–34

Regional and national databases and data warehouses, **101–117**  
 challenges with, 109–113  
 communications, 111  
 data linkage, 112  
 data representation  
 reconciliation, 111–112  
 standardization of data, 112–113  
 claims data, 102–104  
 collaborative, 105–106  
 public health, 106–109  
 single-provider, 104–105  
 types of, 102

Regional health information organizations, 105

Relational database management system, 62, 69

Relational online analytical processing methods, 60, 62

Reports, of data mining, 125

ROKU method, 138–139

## **S**

---

Security, of data warehouses, 65–67

Segmentation, in temporal data mining, 91

Selective-expression approach, 131–132

SENIC (Study on the Efficacy of Nosocomial Infection Control), 120

Serial Analysis of Gene Expression, 128–131

Shannon Entropy, 135–137

Significance analysis, of microarrays, 149–150

Single link, for clustering analysis, 19

Single-provider data warehouses, 104–105

Singular value decomposition, 12–16

Sliding window methods, in temporal data mining, 89–91

SNOMED CT (Systematized Nomenclature of Medicine–Clinical Terminology), 112

Snow, John, 9–10

Social issues, in data mining, 6

Software packages. *See also* Open-source tools.  
for biomarker development, 132–139

Statistical modeling, in genomic data mining, 151–153

Structured Query Language model, 84–85

Study on the Efficacy of Nosocomial Infection Control (SENIC), 120

Subjective measures, in multi-database mining, 76

Subsequencing methods, in temporal data mining, 89–94  
segmentation, 91  
sliding window, 89–91  
temporal abstraction, 91–94

Support vector machines, 29–31, 159–161

SymAtlas V1.2.4, 130–131

Systematized Nomenclature of Medicine–Clinical Terminology, 112

---

## T

Tanagra tool, 43–44

Temporal abstraction method, 91–94

Temporal data mining, 5–6, **83–100**  
for temporal patterns, 94–95  
representation of time in, 84–86  
subsequencing methods in, 89–94  
segmentation, 91  
sliding window, 89–91  
temporal abstraction, 91–94  
time series similarity measures in dynamic time warping, 87–88  
transform-based, 88–89

Textual data, in data warehouses, 67–68

Time series similarity measures, in temporal data mining  
dynamic time warping, 87–88  
transform-based, 88–89

Time-Oriented Database model, 84

Tissue specificity analysis, for biomarker development, **127–143**

TissueInfo, 132–133

Tissue-specific expression-profiling methods, for biomarker development, 131–139  
Akaike's Information Criterion, 137  
ExQuest, 133–134  
Gene Expression Profiling in silico (GEPIS), 134–135  
ROKU, 138–139  
selective expression, 131–132  
Shannon Entropy, 135–137  
tissue selectivity, 137–138  
TissueInfo, 132–133

Training set, for evaluation, 32–33

Transaction processing systems, 56–57, 59–60, 62

Transformation, of data, 63–65

Transform-based methods, in temporal data mining, 88–89

Tree classifiers, 25–27

TwinNET database, 74

---

## U

Unigene, 128–129

User interfaces, for data warehouses, 69

---

## V

Vaccine Adverse Event Reporting System, 107

Variables, discovery techniques for, 12–16

Veterans Administration, database of, 105

---

## W

Waikato Environment for Knowledge Analysis (Weka), 39, 45–46

Warehouses, data, **55–71, 101–117**

Weka tool, 39, 45–46

West Nile virus data, 109–110

---

## Y

YALE tool, 46–48

Yet Another Learning Environment (YALE), 46–48