

## Index

- ABC model, 210  
 active pharmaceutical ingredient (API), 24  
 Adis Reactions Database, 54  
 adverse events  
   reporting, 19, 35  
   text mining, 195  
 Alzheimer's disease, 19, 23  
 amantadine, 172  
 Amazon Cloud, 152  
 anacetrapib, 21  
 analoging, 24  
 analytical variation, 83–84  
 Anatomical and Therapeutic Chemical classification system, 192  
 anatomy, text mining, 194  
 anchor verbs, 199  
 animal models, patent mining, 211–212  
 annotations, 53–54  
 ANSI/NISO Z39.19, 189  
 antibiotics, 2  
 anti-coagulants, 22  
 antifungal agents, 160, 161  
 anti-histamines, 28, 29  
 antiretroviral drugs, 156  
 aplaviroc, 156, 159, 161  
   nearest neighbor, 162  
 apolipoproteins, 21  
 Apple iPod/iPhone, 143, 151, 152  
 application algorithms, 148–150  
 ARACNe algorithm, 41  
 area under the curve (AUC), 89, 183  
 ARPANET, 142  
 artificial intelligence (AI), 47  
 assembly line of drug discovery, 3–5  
 atomoxetine, 172  
 atorvastatin, 21, 178  
 Atracsys, 144  
 attrition, 24  
 authors, text mining, 197  
 automatic term recognition (ATR), 189, 190  
 Avastin. *See* bevacizumab  
 Ayasdi Core, 224  
 azathioprine, 172  
 bacteriology, 1  
 BANNER algorithm, 192  
 Bernoulli experiment, 99  
 Bertin, Jacques, 146  
 best-in-class drugs, 57  
 beta-blockers, 20  
 bevacizumab, 74, 76  
 bibliome, 48  
 BioCarta, 224  
 Bioconductor, 223  
 bioinformatics, 3  
 biologics, 27, 64  
 biomarkers, 63–92  
   classification, 73–74  
   competitive advantage, 78  
   concept qualification, 83–84  
   drug relations, 64–70  
     common tools and concepts, 67–70  
     shared characteristics, 64–66  
   drug safety, 76–77  
   efficacy, 76  
   interpretation of results, 84–92  
     experimental design, 84–85  
     net reclassification, 90–92  
     power analysis, 85–86  
     predictive models, 86  
     receiver operator characteristic (ROC), 88–89  
     regression modeling, 87–88  
   patient delimiting, 77–79  
   pharmacodynamic, 75–76  
   target engagement, 74–75  
   text mining, 195  
   translation, 76–77  
 BioMart, 222  
 biomedical databases, 47 *See also specific databases*  
 BioNLP, 48  
 BioPrint database, 21, 95, 102, 159  
   CCR5 antagonists, 165–166  
 Biosis Previews, 50  
 biospectra analysis, 8, 158, 159–161, 167  
 BioSystems, 222  
 BioVia, 223  
 Black, James, 29

- blockbuster drugs, 20, 24, 29  
 Brown, Michael S., 4  
 bubonic plague, 1  
 bupropion, 172
- CAIV, 137–153  
 application algorithm and design, 148–150  
 components of, 138  
 data visualization, 137–138, 146–148  
 framework, 138–140  
 hardware, 140–142  
 input/output devices, 142–145  
 “Magic Wall”, 138, 145  
 security, 152  
 user adaptation, 150–153
- cancer, 20
- CASETOX, 95, 101
- causal relationships, 202
- causality, 24
- CCR5 antagonists, 168  
 background, 156–159  
 biospectra analysis, 159–161, 167  
 CCS values, 158, 160, 162, 166  
 nearest neighbors, 162  
 data analysis, 165–168  
 BioPrint database, 165–166  
 dendrograms, 166  
 expansion of data set, 168  
 human tissue body map, 162–163, 164, 168  
 methodology, 165  
 safety, 156  
 target-based prejudices, 155–156
- CCS. *See* cluster similarity (CCS) values
- Celebrex. *See* celecoxib
- celecoxib, 21, 172
- Cell Line Knowledge Base, 194
- Cell Name Index, 194
- Cellosaurus, 194
- Center for Drug Evaluation and Research (CDER), 54
- Ceox. *See* rofecoxib
- CETP (cholesterol-ester transfer protein)
- ChEBI, 192
- ChemAxon, 193
- CHEMDNER, 193
- Chemical Abstracts, 50
- chemicals, text mining, 192–193
- ChemIDplus, 192
- chemogenomics, 102
- CHEMSPOT, 193
- chlorpromazine, 29, 172
- cholesterol-ester transfer protein (CETP), 21
- cimetidine, 29
- clinical phases, 4
- clinical records, 56–57
- clinical trials, 52–53, 171–180  
 FCCs, 176–177  
 drugs looking for diseases, 177  
*in vitro* methods, 178  
*in vivo* methods, 178–179  
 repositioning, 171–176  
 challenges, 179–180
- ClinicalTrials.gov, 53
- closed discovery paradigm, 211
- closed-ended questions, 210, 213
- clotrimazole, 160
- cluster analysis, 130–133
- cluster similarity (CCS) values, 158, 160, 162, 166
- coexpression analysis, 40
- comparative biology, 2
- complementarity determining regions, 73
- computational biology, 13–14  
 rise of, 2–3
- computational chemistry, 6
- computational phenotype assessment (CPA), 95  
 comparison with other methods, 101–111  
 compounds used for, 103  
 expanded knowledge, 101  
 false positives  
 compounds used for, 107  
 rate of, 100
- isoprenaline, 101  
 toxicological ontology, 103
- learning from, 99
- ontology terms, 97
- results  
 cardiovascular, 99  
 gastrointestinal, 100  
 musculoskeletal, 100
- screening database, 97
- Tanimoto searching, 97
- test sets, 96–97  
 false positive, 96  
 toxicological ontology, 97, 98
- computer-aided information visualization. *See* CAIV
- computer hardware, 140–142
- concanavalin A, 157
- concept unique identifiers (CUIs), 189
- ConceptMapper, 190
- Connectivity Map, 34, 35
- Copyright Clearance Center, 50
- corpus, 48
- cowpox, 1
- CPA. *See* computational phenotype assessment
- Critical Path Initiative, 63
- CTAGs, 3
- CUDA architecture, 141
- curated text, 53–54
- curation, 207
- CYP2C9, 22
- Cytoscape, 69, 126, 223
- dalcetrapib, 21
- dapoxetine, 172
- data visualization, 114–133

- computer-aided, 137–138, 146–148
- functional genomics, 129–133
- biological context, 133
  - gene clusters and expression patterns, 130–133
  - individual genes, 130
  - whole data sets, 129–130
- genomic sequence, 115–118
- molecular networks, 124–128
- pathways, 122–125
- sequence comparisons, 118–122 *See also* CAIV
- data warehouses, 139
- data-driven, 24
- DAVID Bioinformatics Resources, 222
- dbVar, 222
- Deductive Estimation of Risk from Existing Knowledge.  
*See* DEREK
- dendrograms, 159, 160, 166
- Depo-Provera, 172
- depression, 20
- DEREK, 95, 101
- Derwent Drug File, 50, 54
- design of experiments, 84–85
- diabetes, 20
- dictionary matching, 191
- differential connectivity, 40
- digitizers, 143
- diphenhydramine, 172
- disambiguation, 197
- discovery
- browsing, 211 *See also* drug discovery pipeline
- disease
- complexity, 37–42
  - diagnostics, 77
  - text mining, 193
- distributed computing, 140
- Dragon Naturally Speaking, 143
- drug development, 17, 18, 20, 21 *See also* drug discovery pipeline
- drug discovery pipeline (DDP), 1–14, 26, 32
- as assembly line, 3–5
  - clinical phases, 4, 8–10
  - history of drug discovery, 1–2
  - idea to targets, 5–6
  - lead identification, 4, 6–7
  - lead optimization, 4, 7–8
  - potential weaknesses, 11–12
  - preclinical assessment, 4, 8–10
  - rise of computational biology, 2–3
  - target identification, 4, 17
  - translational biomarkers. *See* biomarkers
- drug optimization, 71
- drug properties, 71
- druggability, 71–73
    - large molecule drugs, 73
    - small molecule drugs, 71–73
- drug repositioning. *See* repositioning
- drug text mining, 192–193
- DrugBank database, 26, 35, 192
- druggability, 7, 8, 64, 71–73
- large molecule drugs, 73
  - small molecule drugs, 71–73
- duloxetine, 172
- Dundas Data Visualization, 147
- Dunning-Kruger effect, 14
- efficacy, 19–20
- biomarkers, 76
- eflornithine, 172
- electronic gaming, 148
- electronic medical records, 9
- Embase, 50, 206, 221, 224
- empagliflozin, 14
- Emtree, 54, 206
- Endo, Akira, 3
- ENSEMBL, 116, 225
- Enzyme nomenclature database, 225
- etiology, 24
- Europe PubMed Central, 50
- European Bioinformatics Institute (EBI), 192
- European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL-EBI), 34
- eUtils, 221
- evacetrapib, 21
- event mining, 202
- failed clinical candidates. *See* FCCs
- false positives, 203
- FCCs, 171, 176–177
- drugs looking for diseases, 177
  - positioning of
    - in vitro* methods, 178
    - in vivo* methods, 178–179
- FDA Adverse Event Reporting System (FAERS), 54
- FDA Orange Book, 26
- FDA RSS feeds, 224
- finasteride, 172
- first-in-class drugs, 57
- fit-for-purpose, 83
- Fleming, Alexander, 1, 2, 3
- fluoxetine, 172
- F-measure, 182–185
- follower drugs, 22
- follow-on drugs, 57
- Food and Drug Administration. *See* FDA
- fractalkine receptor (GPR13), 81
- Framingham Heart Study, 20
- Fry, Benjamin, 146
- functional genomics, 114
- fuzzy matching, 190
- g-speak spatial operating environment, 144
- galantamine, 173
- Gapminder Foundation, 147
- GATE, 206

- gene clusters, 130–133  
 gene expression, 12  
 Gene Expression Omnibus (GEO), 34, 53, 222, 225  
 Gene Ontology (GO), 226  
 Gene Ontology Annotations (GOA), 226  
 gene ontology enrichment, 33  
 Gene Reference Into Function (GeneRIF), 54  
 gene sequencing, 2  
 gene set enrichment, 33  
 gene text mining, 192  
 Genentech, 36  
 genetic variation, 2, 21  
 genome sequencing  
   sequence comparisons, 118–122  
   visualization of, 115–118  
 genome-wide association studies, 34  
 genomics, 9–11, 30, 114  
   functional, 114  
 genotyping, 30  
 glycopyrrolate, 173  
 GNAT algorithm, 192  
 Goldstein, Joseph, 4  
 Google, 149, 151  
 G-protein coupled receptors (GPCRs), 72  
 GPU computing, 141  
 grant proposals, 53  
 graphic representations. *See* data visualization  
 grid computing, 140
- Han, Jeff, 144  
 heat maps, 130  
 high-dimensional data, 24, 129  
 high-throughput screening, 5, 26, 198  
 history of drug discovery, 1–2  
   target identification, 28–31  
 HMG-CoA reductase inhibitors, 178  
 HUGO gene symbols, 207  
 human immunodeficiency virus 1 (HIV-1), 156  
 human tissue body map, CCR5 antagonists, 162–163,  
   164, 168  
 Hy's Law, 90  
 hygiene, 1  
 hypercholesterolemia, 4, 9–11, 20  
 hyperlipidemia, 20, 21  
 hypertension, 20  
 hyponyms, 186  
 hypothesis-driven, 24
- ideas to targets, 5–6  
*in silico* approaches, 3, 7, 42, 213  
 information extraction (IE), 47  
 information retrieval (IR), 47  
 information visualization, 138, 139 *See also* CAIV  
 Infotrieve, 50  
 ingenuity pathway analysis, 31, 225  
 input/output devices, 142–145  
 insulin sensitizing agents, 21
- integrated discrimination improvement (IDI), 90  
 intellectual property, 180  
 internal company documents, 55–56  
 International Statistical Classification of Diseases and  
   Related Health Problems (ICD), 188  
 InterPro, 225  
 inverted terms, 188  
 Investigational New Drug application (IND), 19  
 ion channels, 72  
 isoprenaline  
   CPA assessment, 101  
   toxicological ontology, 103  
 IUPAC, 192
- Jardiance. *See* empagliflozin  
 Jenner, Edward, 1, 2  
 Journal Article Tag Suite (JATS), 51
- kinases, 72  
 knockout mice, 36  
 Kyoto Encyclopedia of Genes and Genomes (KEGG),  
   224
- large molecule drugs, 27, 64, 73  
 lead identification, 4, 6–7  
 lead optimization, 4, 7–8  
 lidocaine, 173  
 Linguamatics I2E package, 79  
 link-resolving tools, 50  
 Lipinski “Rule of Five”, 96  
 Lister, Joseph, 1  
 literature holes, 212–213  
 literature mining, 48  
   discovery, 210–211  
 literature-based discovery, 210–211  
 lopinavir, 162
- “Magic Wall”, 138, 145  
 Mammalian Phenotype Ontology, 195  
 maraviroc, 157, 159, 161, 166  
   nearest neighbor, 162, 164  
 MASSIVE software, 148  
 Mathematica, 150  
 “me too” drugs, 23, 28, 36, 57  
 mecamylamine, 173  
 MedDRA, 195  
 Medical Subject Headings. *See* MeSH ontology  
 Medline, 47, 49, 54, 224  
 Medmeme, 50  
 MeSH ontology, 54, 188, 189, 206, 222, 224  
 Message Understanding Conference (MUC), 182, 190  
 metabolomics, 30  
 MetaCore, 225  
 meta-knowledge, 202  
 MetaMap, 190  
 miconazole, 160  
 microarrays, 129

- microbiology, 1  
 Microsoft Excel, 146  
 mifepristone, 173  
 milnacipran, 173  
 minoxidil, 173  
 MLR-1023, 178, 179  
 MMRRC (Mutant Mouse Region Resource Centers), 37  
 molecular biology, 30  
 molecular inhibitors, 5  
 monoclonal antibodies, 73  
 Moore's Law, 6, 11  
 Mouse Genome Informatics (MGI), 225  
 multi-scale data sets, 24  
 multi-touch display devices, 144  
 mutations, text mining, 195–196
- named-entity recognition (NER), 190  
 National Cancer Institute (NCI) Thesaurus, 188, 189  
 National Drug File, 192  
 National Information Standards Organization (NISO), 51  
 National Institutes of Health (NIH), 48  
 National Library of Medicine (NLM), 47  
 natural language processing (NLP), 47, 79  
 NCBI Gene, 192  
 NCBO Annotator, 190  
 net reclassification, 90–92  
 networks, 37–42
  - data visualization, 124–128
  - differential connectivity, 40
- new chemical entities (NCEs), 94  
 New Drug Application (NDA), 19
  - CCR5 antagonists, 168
- new molecular entities (NMEs), 22  
 news reports, 55  
 next-generation sequencing, 5, 30  
 NextMove, 193  
 next-nearest-neighbor, 83  
 Nintendo Wii remote control, 143  
 NSAIDs, 155  
 nuclear magnetic resonance, 7
- Oblong Industries, 144  
 OBO-Edit, 189  
 Oculus Rift, 145  
 off-target events, 21  
 off-target proteins, 94  
 omics data, 24, 42, 72  
 oncogenes, 5  
 Online Mendelian Inheritance in Man (OMIM), 54, 222, 225  
 ontologies, 185–188
  - breadth of, 188
  - managing, 189
  - primary terms, 187
  - qualifier terms, 188
  - synonyms, 186 *See also specific ontologies*
- open access (OA), 50  
 Open Biological and Biomedical Ontologies (OBO), 189  
 Open Researcher and Contributor ID (ORCID), 197  
 open source tools, 221–224  
 open-ended questions, 210  
 optical character recognition (OCR), 51  
 organ-on-a-chip, 32  
 orphan diseases, 23, 26  
 OSCAR, 193  
 overhead, 184, 186
- paclitaxel, 173  
 PANTHER Classification System, 186  
 Pareto principle (80-20 rule), 210  
 Pasteur, Louis, 1  
 patents, 52
  - animal models, 211–212
- PathVisio, 67  
 pathways, 66–67
  - data visualization, 122–125
  - next nearest neighbor analysis, 83
  - text mining, 201
- patient delimiting biomarkers, 77–79  
 PCSK9, 9–11  
 Pencina, Michael, 90  
*Penicillium notatum*, 2  
 Perceptive Pixel, 144  
 perturbations, 202  
 Pfam, 53  
 Pharma Documentation Ring (P-D-R) consortium, 51  
 pharmaceutical industry, 18, 20
  - R&D investment, 27
- pharmacodynamic markers, 75–76  
 PharmGKB, 192  
 phenotyping, 26, 36
  - text mining, 195
- phentolamine, 173  
 phosphatases, 72  
 Pipeline Pilot and Inforsense Suite, 206  
 pleiotrophic effects, 24  
 polyhierarchies, 186  
 polymerase chain reaction (PCR), 12  
 power analysis, 85–86  
 precision, 182, 183, 184  
 preclinical assessment, 4  
 predictive models, 86  
 Predictive Safety Testing Consortium (PSTC), 77  
 pregnancy tests, 78  
 premarin, 173  
 primary terms, 187  
 Procter & Gamble Innovation Network, 152  
 promethazine, 29  
 propranolol, 20  
 proteases, 72  
 Protégé, 189  
 protein B-cell activating factor (BAFF), 80  
 Protein Data Bank (PDB), 225

- protein panels, 165  
protein text mining, 192  
protein–protein interaction (PPI) network, 36, 197  
proteomics, 30  
  druggable proteome, 165  
Pubget, 50  
PubMed, 47, 80  
PubMed Central, 50
- qualifier terms, 188  
quantitative structure–activity relationship (QSAR), 95  
question answering, 203  
Quosa, 50
- raloxifene, 173  
ranking, 184  
reading, 204  
recall, 182, 183, 184, 204  
receiver operating characteristic (ROC), 88–89  
regression modeling, 87–88  
relational ontology, 80  
relevance, 184  
repositioning, 9, 171–176  
  challenges, 179–180  
  drugs looking for diseases, 177 *See also* FCCs  
ReprintsDesk, 50  
retin-A, 173  
rheumatoid arthritis, 20  
Rhône-Poulenc, 29  
RNA splicing, 37  
ROARMAP database, 50  
rofecoxib, 21, 155  
ropinirole, 173  
RxNorm, 192
- Saccharomyces cerevisiae*, 33  
safety, 19  
  data analysis, 54  
  long-term, 21  
safety biomarkers, 76–77  
SAGE (Serial Analysis of Gene Expression), 5  
scatter plots, 130  
scientific publications, 49–51  
scraping data, 24  
screening, 6  
search engines, 149  
secondary metabolites, 24  
Senge, Peter M., 150  
severe adverse reactions, 35  
sibutramine, 174  
Side Effect Resource (SIDER), 54  
sildenafil, 174  
Simple Knowledge Organization System (SKOS), 189  
simvastatin, 178  
single nucleotide polymorphisms (SNPs), 34  
small molecule drugs, 27, 64, 71–73, 94–110  
  computational phenotype assessment. *See*  
  computational phenotype assessment (CPA)  
  rule of five, 71  
  smallpox, 1  
  Smith, Kline & French, 29  
  SNOMED-CT ontology, 79, 188, 195  
  social media, 56  
  Southern blotting, 2  
  species  
  text mining, 193  
  speculative and negative facts, 204  
  Spotfire, 129  
  spurious hits (overgeneration), 184  
  statins, 4, 23, 27, 178  
  structure–activity relationship (SAR), 7  
  supercomputers, 140  
  support vector machine (SVM) algorithm, 212  
  synonymy, 186  
  synthetic biology, 24  
  Systems Biology Graphical Notation (SBGN), 123
- tadalafil, 174  
Tanimoto searching, 97  
target definition, 18–25  
target engagement, 74–75  
target identification, 4, 17, 71  
  current situation, 31–37  
  disease complexity, 37–42  
  efficacy, 19–20  
  history, 28–31  
  methods, 26–28  
  networks, 37–42  
  safety, 19, 21  
Taverna, 223  
taxonomies, 186  
templates, 199  
term classification, 189  
term identification, 189–192, 198  
term mapping, 189  
term recognition, 189  
test sets, 96–97  
  false positive, 96  
tetracyclines, 174  
text analytics, 24, 48, 79–83  
text mining, 47–59, 79, 182–213  
  abbreviations, 197  
  anatomy, 194  
  applications, 209–213  
  discovery questions, 210–211  
  easy and hard questions, 209–210  
  open-ended and closed-ended questions, 210  
  authors, 197  
  biomarkers, 195  
  causal relationships and perturbations, 202  
  chemicals and drugs, 192–193  
  companies and institutions, 196–197

- content diversity, 57–58
- definition, 47–48
- diseases, 193
- events, 202
- experimental methods, 194–195
- genes and proteins, 192
- identification of terms, 189–192
- integration, 207–208
- interoperability, 206
- measurements and parameters, 196
- mutations, 195–196
- ontologies, 185–188
  - managing, 189
- pathways, 201
- phenotypes and adverse events, 195
- post-processing and curation, 207
- practical examples, 211–213
  - literature holes, 212–213
  - patent mining of animal models, 211–212
- pre-processing, 205
- question answering, 203
- relationships, 197–201
- scope definition, 205–206
- sources of content, 48–57
  - clinical records, 56–57
  - clinical trials, 52–53
  - curated text and annotations, 53–54
  - grant proposals, 53
  - internal company documents, 55–56
  - news reports, 55
  - patents, 52
  - safety data, 54
  - scientific abstracts/articles, 49–51
  - social media, 56
  - websites, 55
- species, 193
- structured vs unstructured data, 58–59
- timeliness, 57
- tool evaluation, 182–185
- users, 203–204
- visualization and presentation, 208
  - within pharmaceutical organization, 213
- thalidomide, 174
- theraTRACE® platform, 178
- tioconazole, 160
- tofisopam, 174
- Tool for Approximate LargeE (TALE) graph matching, 33
- tools, 67–70
  - Cytoscape, 69
  - PathVisio, 67
  - topiramate, 174
  - TOPKAT, 95, 101
  - topological overlap map (TOM), 39
  - torcetrapib, 21
  - touchscreens, 143, 144
  - toxicity, 19
  - Toxicity Prediction by Komputer Assisted Technology.
    - See TOPKAT
  - TOXNET databases, 54
  - transcriptomic profiling, 30
  - transcriptomics, 33
  - transitive inference, 210
  - translation medicine, 9
  - translational biomarkers. *See* biomarkers
  - tranSMART, 223
  - TREC Genomics Track, 203
  - Trendalyzer software system, 147
  - Trialtrove, 223
  - Tufts, Edward, 67, 137, 146
- UIMA, 206
- Unified Medical Language System (UMLS), 189, 224
- UniProt, 53
- UniProt Knowledgebase (UniProtKB), 192, 225
- United States Patent and Trademark Office (USPTO), 52
- user adaptation, 150–153
- vaccination, 1
- vascular endothelial growth factor-A (VEGF-A), 74, 76
- vicriviroc, 156, 157, 159, 161
  - nearest neighbor, 162
- Vioxx. *See* rofecoxib
- Vioxx effect, 155–156
- VKORC1, 22
- voice emulators, 145
- voice recognition input, 142
- wall displays, 144
- wearable devices, 145
- Web Ontology Language (OWL), 189
- websites, 55
- Wolfram|Alpha, 149
- workstations, 141
- X-ray crystallography, 7