Index

A/B experimentation 21, 27
abuse
  copyright 78
  definition and examples 145–146
  motivations for 172
  music recommendation systems 67
  reproducibility in scientific applications 170
  resistance to 145–149
  traffic speed estimation 75
  see also dependability
academia, place of data science within 268–270
  see also transdisciplinarity of data science
access control 135
accountability
  ethics 222–223
  Health Insurance Portability and Accountability Act 128
action, bias for 185
ad brokers 201
addition, bias for 185–186
adversarial change 119
adversarial data attacks 147–148
advertising
  balancing benefit across parties 191–193
  confidentiality policies 134
  targeted or personalized 80–82
  targeting subpopulations 116–117
  via the Internet 35
advertising click fraud 146
alcohol–vehicle accidents study 165–166
algorithmic fairness 98, 101
algorithmic modeling culture 15, 26
algorithms
  definition 42
  foundational fields of data science 19
  using to solve problems 271–273
AlphaFold2 software 54–55
American Statistical Association (ASA) 37–38
amino acids 54–55
Analysis Rubric
  applying 74–75, 266–267
  challenges 109
  ethical, legal, and societal implications (ELSI) element 62
  implementation-oriented elements 61
  purpose of 2, 61
  quality and care 259–260
  requirements-oriented elements 61–62
  in summary 72
Analysis Rubric applications
  COVID-19 mortality rates 71–72
  financial services 90–95
  healthcare records 69–71
  medicine and public health 84–88
  music recommendation 66–68
  protein folding 68–69
  science-oriented 89–90
  social, political, and governmental 95–99
  speech recognition 64–66
  spelling correction 62–64
  transport and mapping 75–78
  web and entertainment 78–84
anchoring bias 180
anomaly detection 149
anonymized data 103–104
  see also privacy issues
anti-money-laundering 93
antibody tests 115–116
app use patterns 222
Apple 141
applications *see* data science applications
artificial intelligence (AI)
  data science as distinct from 20
  definition 20, 42
  “from algorithms to data to needs” 271–273
  regulatory landscape 216
  terminology 27
AT&T 132
attrition controls 162
auditability 154–158
authority, appeals to 183–184
automated speech recognition (ASR) 49–51
automatic data, as term 27
automatic translation 208–209

automation
  impact on being human 202–203
  of jobs 219, 233
  societal impacts 219
autonomous cars *see* self-driving cars
availability bias 172
averages, setting objectives 191

Babbage, Charles 18
bagging 122–123
base rate fallacy 180–181
baseline treatment 163
batch mode 19
behavioral retargeting 201
behavioral targeting 196
Belmont Principles
  acting ethically 222
  application to data science 39–40
  commission and *Belmont Report* 38–39
  principlist approach 100–105
  see also beneficence; justice; respect for persons
beneficence
  criminal sentencing and parole decision-making 101
  definition 43
  mobility reporting 104
  news feed recommendations 102
  as principle 38, 39–40
  underwriting and insurance 105
  vaccine distribution optimization 103
bias
  cognitive biases 172–173, 180–186
  criminal sentencing and parole decision-making 98–99, 101–102
  inductive bias 121–124
  role of information consumers 180–186
  selection bias 115–116
  social biases 121, 193–196
  underwriting and insurance 105
big data
  computing hardware 22
  data quality 115–116
  definition 42
  "from algorithms to data to needs" 271–273
  and machine learning 113–114
  scale 13
  terminology 27
biometric identification 131, 132, 245
Bitcoin 149
Boeing 737 Max crashes 150
boosting 123
bots 173
bottom-up approach 62
Breiman, Leo 15–16

Cambridge Analytica/Facebook scandal 141
cars *see* alcohol–vehicle accidents study; self-driving cars
casualty insurance 93–94
causal relationships 8–9
  definition 42
  statistics 121
  understandability 158–168
  variable selection 167–168
  visualization 17
census data 136
chain of trust 140
child pornography 113, 141
city noise levels monitoring 190
classifications
  definition 11
  machine learning 24
clear box 12
climate change 166, 235
clinical trials 130
clustering
  definition 12
  machine learning 24
clustering illusion 183
cognitive biases 172–173, 180–186
collaborative filtering 52–53
communicating data science results 171–186
  examples 173–174
  organizational ethics 225–226
  role of information consumers 179–186
  role of journalists 177–179
  role of scientists 174–177
COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) 194–195
competitive development 235–236
computer engineering 20
computer security 142–145
  see also privacy issues
computer vision (CV) 25
computing
  in the curriculum 237–240
  definition 14
  as foundational field of data science 19–24
  key terms 43
  machine learning 24–26
  transdisciplinary research 255
computing hardware 22, 124, 138
concept drift 118–120, 207
conclusions (defining data science)
  data science as field 20
  definition 42
  music recommendation systems 53–54
  types of 7, 11–12
confidence interval 207
confidentiality 126–127, 134, 233
  see also privacy issues
confirmation bias 173, 175
conflict of interest 221–222
confounding variables 159–160
conjunction fallacy 182
consensus-building institutions 251–254

content regulation 212
contextual integrity 127
continual optimization 19
control group 161
copyrighted material identification 78–79
corner cases 191
corner condition 76
corpus data 48–49
correlation 8–9, 42
  see also causal relationships
counterfactual 158
country-wide economic predictions 94–95
courses *see* data science courses
COVID-19
  Analysis Rubric 71–72
  communicating results 225–226
  healthcare records 56
  location data 141–142
  predicting mortality 58–60
  role of science 30–32
  visualization 9–11
COVID-19 vaccinations
  communicating results 173–174
  media reporting 178–179
  vaccine efficacy 173–174
  visualization 9–11
criminal sentencing 98–99
  fairness 194–195
  legal regulation 245
  principlist approach 101–102
crowdsourcing 56–57
cryptography 137–138
curriculum, data science in 237–240
  see also education
cybersecurity 142–145
CyberSecurity Bill of Rights 143
cyclic models 121–122

dark patterns 222
data analysis 27
  see also Analysis Rubric
data attacks 146–149
data deluge 234
data generation and collection
  privacy issues 126, 128–131
  sources 111–112
data-intensive approaches 48–49
data minimization 133
data modeling culture 15, 26
data poisoning 148
data privacy *see* privacy issues
data processing 112–114
data provenance 115
data publication norms 253
data quality 115–117
data retention 131–134
data science
  AI as distinct from 20
  “as coherent field” 265–266
  conclusions, types of 7, 11–12
  definitions 7–8, 42
  foundational fields 2, 13–14
  see also computing; historical perspective; operations research; statistics
  “from algorithms to data to needs” 271–273
  future of 276–277
  insights 7, 8–11
  models 7–8
  opportunities and challenges 266
  “post-modern Prometheus” 274–276
  “rethinking responsibility and success” 270–271
  scale 12–13
  terminology 27, 41–43
  “where does data science fit in academia?” 268–270
  see also transdisciplinarity of data science
data science applications 1–4, 45–46
  academic research areas 30–34
  challenges 109, 275–276
  COVID-19 mortality rates 58–60, 71–72
  economic sector 30
  financial services 90–95
  healthcare records 55–58, 69–71
  legal challenges 214–216
  medicine and public health 84–88
  music recommendation 51–54, 66–68
  protein folding 54–55, 68–69
  science-oriented 89–90
  social, political, and governmental 95–99
  speech recognition 49–51, 64–66
  spelling correction 48–49, 62–64
  teaching 240–241
  technical, contextual, and societal challenges 3
  transport and mapping 75–78
  web and entertainment 78–84
  see also Analysis Rubric; societal concerns
data science courses
  demand for 22–24
  post-secondary 238–240
  primary and secondary curriculum 237–238
data science skills 218–219
data storage 114–115, 126
data volume
  scale 12–13
  virtuous cycle 12
data wrangling 112
datasheets 114–115
de-identification 134
de jure regulation 244–250
  see also legal issues in data science
deep fakes 146
deep learning 20, 26, 27
DeepMind, AlphaFold2 software 54–55
democratizing data 140–142
demographic data 136

denominator neglect 173–174
dependability 126
  Analysis Rubric 61
  autonomous cars 77
  balancing corporate, individual, and government concerns 131–134
  COVID-19 mortality rates 71
  genome-wide association studies (GWAS) 87
  healthcare records 70
  location data 139–140
  music recommendation systems 66–67
  and privacy 126–127
  privacy regulation: unintended consequences 140–142
  privacy versus usage rights 128–131
  protein folding 69
  resilience 149–153
  resistance to abuse 145–149
  security 142–145
  speech recognition 65
  spelling correction 63
  targeted advertising 81
  technologies for privacy 134–139
designers 36
dictionary approach (spelling) 48
differential privacy 136
disciplinary foundations of data science 2, 13–14
  see also computing; operations research; statistics
discounts, anchoring bias 180
discrimination 193–196
  see also fairness
discriminative models 24
disease diagnosis 85
disease outbreaks 85
diseases, causes of 88
disparate impact 193
disparate treatment 193
divisiveness 234
DNA evidence example 181–182
double-blinding 162
dynamic programming 18

earthquake prediction 89–90
economic impacts (data science) 216–220, 233
economic sector
  algorithmic investment 92–93, 120
  data generation and collection 112–113
  data science applications 30
  financial services 90–95
economics, contribution to data science 36, 256
economies of scale 248–250
education
  data science applications 97
  data science courses 22–24
  data science in the curriculum 237–240
  importance of 237
  school testing results 175
  transdisciplinary research 256
  use of data science in improving education 240–241
  vocabulary/definitions 241–243
efficiency *see* optimization
email spam detection 113
empirical computing 20–21, 22–24
employment, automation of jobs 219, 233
enclave-enabled services 138
encryption 135–136
  homomorphic 138–139
  end-to-end 138
energy usage data 134–135
engineering 33
ensemble models 122–123
entertainment
  application of data science 78–84
  music industry 78, 257
  music recommendation systems 51–54, 66–68
  video game personalization 79, 198
environment
  climate change data 166
  data science applications 235
  nudging 198
errors *see* failures, toleration of; uncertainty
estrogen replacement therapy study 166
ethical, legal, and societal implications (ELSI) 62, 212
  acting ethically 220–226
  COVID-19 mortality rates 72
  economic impacts 216–220
  genome-wide association studies (GWAS) 88
  healthcare records 71
  legal issues 212–216
  music recommendation systems 68
  news feed recommendations 83
  protein folding 69
  speech recognition 65–66
  spelling correction 64
  targeted advertising 81–82
  traffic speed estimation 76
ethical principles 4
  acting ethically 220–226
  addressing societal concerns 260–261
  Belmont, application to data science 39–40
  Belmont Commission 38–39
  frameworks 37
  key terms 43
  philosophers 35
  professional ethics 37–38
  success criteria 267
  see also principlist approach
ETL (Extract, Transform, and Load) 111
European Union
  Ethics Guidelines for Trustworthy AI 39

European Union (cont.)
  General Data Protection Regulation of 2016 (GDPR) 128–129, 133, 215
  Right to Be Forgotten 215, 251
experiments
  causality 158–168
  communicating results 171–186
  design of 161–167
  reproducibility 168–171
experts, trust in 183–184, 260–261
explainability 154–158
  see also communicating data science results

Facebook
  Cambridge Analytica scandal 141
  influence of positive and negative posts 197
  user data 146
facial recognition 131, 245
failures, toleration of 206–207
  Analysis Rubric 61–62
  COVID-19 mortality rates 72
  healthcare records 71
  liability 209–211
  music recommendation systems 68
  protein folding 69
  risk 208–209
  route finding 77
  speech recognition 65
  spelling correction 64
  traffic speed estimation 76
  uncertainty quantification 207–208
  web search 83
fair selection of subjects 39–40
fairness
  criminal sentencing 98, 101
  objectives 193–196
  societal concerns 233
  statistics 120–121
  terminology 242
false positives 169, 180–181, 189
federated learning 136–137
filter bubbles 199–200
financial services
  data generation and collection 112–113
  data science applications 90–95
firms, self-regulation 251
fractional factorial test 163
framing bias 180
fraud detection 91
freedom of expression 234
full factorial test 163

gambler's fallacy 182–183
game theory 35
General Data Protection Regulation of 2016 (GDPR) 128–129, 133, 215
generalization (inductive bias) 121–124
generative models 24
genome-wide association studies (GWAS) 86–88, 167–168
geofence warrants 132
GitHub 131
Global Positioning System (GPS) 139–140, 141–142
global warming 166
Goodhart's Law 184
Google
  Bitcoin mining 149
  data requests to 132
  failure tolerance 83
  healthcare records 131–132
  mobility reporting 85, 103–104
  personal data 129
  privacy regulations 141
  release of data logs 170–171
Google Flu Trends 133, 171
Google Research 34
Google Takeout 129
Google Translate 209
governance
  impacts of data science on 234
  innovation in data science 235–236
  large companies 250
government
  data requests 132
  data science applications 95–99
  dependability and privacy 131–134
  legal challenges from 212–214
gradual change 119
graphs
  historical perspective 16
  insights into COVID-19 data 9–11
  see also information visualization
Gray, Jim 32
greenhouse gas emissions 166

hardware 22, 124, 138
Health Insurance Portability and Accountability Act (HIPAA) 128
healthcare
  clinical trials 130
  data science applications 84–88
  Google Flu Trends 133, 171
  hormone replacement therapy study 166
  liability 210–211
  medical records 55–58, 69–71, 131–132
  pharmaceutical treatment effects 167–168
  transdisciplinary research 256
  vaccine distribution optimization 85, 103
heterogeneity of subjects 168
heterogeneous treatment effects 167–168
higher education *see* universities
historical perspective
  emergence of data science 27–28
  foundational fields of data science 2

graphs 16
legal precedents 212
History Lab Project 34
homomorphic encryption 138–139
hormone replacement therapy study 166
house prices 93–94
principlist approach 104–105
quantifying uncertainty 208
HTTPS (HyperText Transfer Protocol Secure) 135–136
human–computer interaction (HCI) 21–22
humanities 34, 256
humanity, impact of data science on being human 202–203
humility 153

IBM Research 111
image classification 189
incentives 221–222
individual protections 213
see also privacy issues
inductive bias 121–124
inferences 14–15, 42
information cascades 184
information consumers, role in understanding results 179–186
information retrieval 21
information visualization
as field 16
historical perspective 16
insights into COVID-19 data 9–11
journalists' role 178
in statistics 16–17
understandability 156
informed consent 39–40
news feed recommendations 102
vaccine distribution optimization 103
innovation
competition between countries 235–236
fostering 256–258
societal concerns 235–236
use of data science in improving education 240–241
insights 7, 8–11
COVID-19 case study 9–11
data science as field 20
definition 42
visualization 17
institutional review boards (IRBs) 221–222, 223
institutional scale 233
instrumentation 111–112
insurance 93–94, 104–105
intellectual property (IP) 130–131
international governance 235–236
Internet
advancing data science 35
advertising 35
data science applications 78–84
data volume 13
online shopping 220
value of web services 129–130
interpretability 154–158
see also communicating data science results
interventions 159
introspection, lack of 182
inverse uncertainty quantification 207
investment, algorithmic approaches 92–93, 120

journalism
communicating data science results 177–179
data publication norms 253
novelty versus robustness in experiments 169
role of journalists 36
transdisciplinary research 256
judicial decisions *see* criminal sentencing
judicial interpretation 247–248
jurisdictional issues 214
justice
criminal sentencing and parole decision-making 101
definition 43
mobility reporting 104
news feed recommendations 102
as principle 38–39, 40
underwriting and insurance 105
vaccine distribution optimization 103

know-your-customer (KYC) compliance regulations 93
known-unknowns 151

L_1 regularization 123
Lake Woebegone effect 175
language use 175
see also communicating data science results
law enforcement 247
lawyers 35
leadership skills 218
legal issues in data science 212–216, 244
Analysis Rubric 62
anti-money-laundering 93
balancing corporate, individual, and government privacy concerns 131–134
de jure regulation 244–250
explainability of results 157–158
financial services 91
General Data Protection Regulation of 2016 (GDPR) 128–129, 133, 215
other guiding forces 250–254
privacy 128–131, 214–215
privacy: unintended consequences of regulation 140–142
Right to Be Forgotten 215
underwriting 93–94
see also ethical, legal, and societal implications (ELSI)

Level-5 (fully autonomous) cars *see* self-driving cars
liability 157–158, 209–211
linear models 121–122, 123
linear programming 18
location data 139–140, 141–142

machine learning 24–26
  and big data 113–114
  definition 20, 42
  “from algorithms to data to needs” 271–273
  internships and courses in 22–24
  statistics as distinct from 26
  terminology 27
  see also models
maintenance needs 97
management skills 218
mapping, data science applications 75–78
  see also Global Positioning System (GPS); self-driving cars
mastery learning 97
McNamara’s Fallacy 184, 190–191
measurability 184, 190–191
media reporting *see* journalism
medical records 55–58, 69–71, 131–132
  see also healthcare
meritocracy 218–219
metadata 114–115
military technology 245
mobility reporting 85, 103–104
model stealing 148–149
models
  building and deploying 118
  clear box or opaque box 12
  COVID-19 mortality rates 58–60
  data modeling culture 15
  definition 42
  healthcare records 57–58
  inductive bias 121–124
  operations research 19
  practical considerations 124–125
  theoretical limitations 118–121
money laundering 93
Moore’s Law 124
multi-armed bandit experiment 163–164
multi-party computation 137–138
multidisciplinarity *see* transdisciplinarity of data science
multivariate testing 163
music industry 78, 257
music recommendation systems 51–54, 66–68

national security 212
national sovereignty 234
natural experiments 165
natural language processing (NLP) 25
nearest-neighbor models 121–122
Netflix 66, 134
neural network models 26
neural networks 22, 27
news feed recommendations 83, 102–103, 196–198
news reporting *see* journalism
newsvendor problem 18–19
noise levels monitoring 190
non-governmental institutions 251
non-stationarity 118–120, 207
non-parametric models 123–124
normal distributions 123
nudging 198
numerical errors 207
NYU’s SONYC (Sounds of New York City) 190

objective function 14, 43, 96
objectives
  Analysis Rubric 61–62
  autonomous cars 78
  balancing benefit across parties 191–193
  challenges 203–205
  clarity of 188–191
  concerns to the individual 196–203
  COVID-19 mortality rates 72
  fairness 193–196
  healthcare records 70
  how to set 187–188
  music recommendation systems 67–68
  protein folding 69
  route finding 77
  speech recognition 65
  spelling correction 64
  targeted advertising 81–82
  traffic speed estimation 76
  transparency 203
objectivity 175
  see also communicating data science results
observational studies 165
online communities *see* usage communities
online mode 19
online shopping 220
opaque box 12
operations research
  definition 14
  as foundational field of data science 17–19
  key terms 43
  large-scale problems 125
  optimization 14, 17
  transdisciplinary research 255
optimization
  definition 43
  economic benefits 216, 219–220
  in machine learning 120
  meaning of 11
  operations research 14, 17
optimizer’s curse 164
ordering bias 182
organizational ethics
  example challenges 225–226

governance 222–223
incentives 221–222
principled decision-making 261
from principles to policies 221–222
outcome measures (design of experiments) 161
overfitting 122

p-values 176–177
pair-matching 162
parametric models 123
Pareto frontier 190
parole decision-making 98–99
fairness 194–195
legal regulation 245
principlist approach 101–102
passwords 143
pattern detection 183
periodic change 119
personal data 126–127
see also privacy issues
personal implications to data 233–234
see also societal concerns
personalization
in advertising 80–82, 116–117, 196–199, 201
filter bubbles 199–200
news feed recommendations 83, 102–103, 196–198
organizational ethics 225–226
privacy issues 201
pros and cons 196–199
reading tutors 97
video games 79, 198
personally identifiable information (PII) 128
pharmaceutical treatment effects 167–168
philosophy, contribution to data science 35, 256
polarization 199–200
political polling 174
political scientists 35
politicians 35
politics
criminal sentencing and parole decision-making 98–99, 101–102, 194–195
filter bubbles 199–200
news feed recommendations 83, 102–103, 196–198
targeting in political campaigns 96–97
Twitter bots 173
polynomial models 121–122
post-modern Prometheus 274–276
post-secondary education 238–240
pre-screen 167
precision 189
prediction, meaning of 11
pricing of property *see* house prices
primary education 237–238
principle of least privilege 135
principlist approach 100–105
criminal sentencing and parole decision-making 101–102
mobility reporting 103–104
news feed recommendations 102–103
organizational ethics 223–225, 261
role of data scientists 260–261
underwriting and insurance 104–105
vaccine distribution optimization 103
privacy issues
balancing corporate, individual, and government concerns 131–134
data generation and collection 113
differential privacy 136
explainability of results 157
healthcare records 57–58, 59
legal regulation 128–131, 214–215, 245–246
location data 139–140
mobility reporting 85, 103–104
personalized advertising 201
reproducibility 170–171
statistics 120–121
technologies for 134–139
terminology 242
traffic speed estimation 75
unintended consequences of privacy regulation 140–142
versus usage rights 128–131
user-generated data 117
Private Aggregation of Teacher Ensembles (PATE) 148
professional ethics 37–38
see also ethical principles
Prometheus (Greek mythology) 274–276
property prices *see* house prices
prosecutor’s fallacy 181–182
protein folding 54–55, 68–69
provenance (metadata) 115
psychologists 36
public health 84–88
see also COVID-19; healthcare

quality in data science 115–117, 259–260
quant trading 92–93
quasi-experiments 165

race
census data 136
fairness 194–195
randomized controlled trial (RCT) 161–167
randomness 207
reading tutors 97
recall 189
recency bias 172–173
recommendations
filter bubbles 199–200
meaning of 11
news feed 83, 102–103, 196–198

recommendations (cont.)
- videos 83
- *see also* personalization

record keeping 215
recruitment (design of experiments) 161–162
regime change 119
regression
- insights into COVID-19 data 9–11
- machine learning 24

regression discontinuity analysis 165
regularization 122, 123
regulation *see* legal issues in data science
reinforcement learning 24, 53
reliability 208–209
replication crisis 169
representativeness 182
reproducibility
- Analysis Rubric 61–62
- in scientific applications 168–171

reputational risk 208
research and development (R&D)
- competition between countries 235–236
- focused and transdisciplinary research 255–256
- fostering innovation 256–258

resilience 149–153
- *see also* dependability

resource allocation problem 18
respect for persons
- criminal sentencing and parole decision-making 101
- definition 43
- mobility reporting 104
- news feed recommendations 102
- as principle 38, 39–40
- underwriting and insurance 104
- vaccine distribution optimization 103

response time 190
responsibilities, as data scientists 270–271
restriction (design of experiments) 162
retail, impact of optimization 220
retention controls 162
reversion 163
Right to Be Forgotten 215, 251
risk
- autonomous cars 77
- COVID-19 mortality rates 72
- healthcare records 71
- terminology 242
- toleration of failures 208–209
- underwriting and insurance 93–94, 104–105

robotics
- automation of jobs 219, 233
- impact on being human 202–203
- machine learning 26
- societal impacts 219

robustness 152
route finding 76–77

sampling 115–116
sampling change 119–120
scale 12–13
- challenges due to 125
- cons to 249–250
- data storage 114–115
- economic benefits 216–218
- economies of scale 248–250
- healthcare records 55–58
- legal precedents 212
- machine learning 113–114
- Moore’s Law 124
- research preferences 169

scarcity 192
sciences 30–32
- data science applications 89–90
- protein folding 54–55
- transdisciplinary research 256

scientific experiments *see* experiments
scientists, communicating data science results 174–177
search engines
- bias 172–173
- clarity of objectives 189
- as data science application 82–83
- optimization 120

secondary education 237–238
secure aggregation 137
secure enclaves 138
secure multi-party computation 137–138
security 142–145, 234
- *see also* national security; privacy issues

selection bias 115–116
self-critical approach 260
self-driving cars
- clarity of objectives 188
- as data science application 77–78
- legal regulation 245
- liability 157–158, 209–210

self-regulation 251
sensitive personally identifiable information (SPII) 128
sentencing decisions (criminal justice) 98–99
- fairness 194–195
- principlist approach 101–102

shopping, use of the Internet 220
Sinclair, John 14
single-subject experiment 164–165
skills, for data science 218–219
Sloan Digital Sky Survey (SDSS) 89
small data 13
social bias 121, 193–196
social issues
- Analysis Rubric 62

data science applications 95–99
legal protections 213
see also ethical, legal, and societal implications (ELSI)
social sciences 32–33
societal concerns 3–4, 229–232
recommendations and approaches 235–236, 267–268
summary of issues 232–235
sociologists 36
software engineering 19
software engineering skills 218
spam filtering 119, 189
speech recognition 49–51
Analysis Rubric 64–66
federated learning 136–137
mistakes in 118–121
spelling correction 48–49, 62–64
SSL (Secure Sockets Layer) 135–136
stationary models 118–120
statistics
causal relationships 121
in the curriculum 237–240
definition 13
as foundational field of data science 14–16
key terms 42
machine learning as distinct from 26
privacy and fairness 120–121
scientists' role in communicating results 174–177
terminology 242
transdisciplinary research 255
visualization 16–17
stochastic optimization 53
stock market investment selection 92–93
stopping criteria 162
storytelling, explaining results 156–157
structural causal model 159
sudden change 119
sufficient power 161
sunk cost fallacy 183
superstar effects 219
supervised learning 24
survivorship bias 184–185
synthetic control method 165

Tacoma Narrows Bridge disaster 151
Target 134
targeted advertising 80–82, 116–117
teaching, use of data science in 240–241
see also education
technical approach
Analysis Rubric 61
COVID-19 mortality rates 71
healthcare records 70
music recommendation systems 66
protein folding 69
speech recognition 65
spelling correction 63
traffic speed estimation 75
technical concerns 274–275
technological change
economic benefits 216, 218–219
economies of scale 248–250
legal regulation 244–246, 247–248
"post-modern Prometheus" 274–276
for privacy 134–139
see also innovation
terms-of-service, liability 209–211
theoretical computer science 19
third-party data 112
TLS (Transport Layer Security) 135–136
top-down approach 62
TOR Project 141
tractable data 61, 111
appropriate use of user-generated data 117
COVID-19 mortality rates 71
data generation and collection 111–112
data quality 115–117
data storage 114–115
healthcare records 70
music recommendation systems 66
processing data 112–114
protein folding 68
speech recognition 64
spelling correction 63
traffic speed estimation 75–76
transdisciplinarity of data science 29
advancing data science 35
building coalitions 35–36
new application areas 30–34
research and development (R&D) 255–256
transformation, meaning of 12
transistors 22, 124
translation systems 208–209
transparency, in objectives 203
transport, data science applications 75–78
see also self-driving cars
traveling salesperson problem (TSP) 18, 125
treatment group 161
trust
chain of trust 140
data science applications 235
Ethics Guidelines for Trustworthy AI 39
for journalists 177–179
reproducibility of results 170–171
for scientists 174–177
secure multi-party computation 137–138
transparency 203
understandability 154–158
truth 191, 260
Tukey, John 9, 27–28
Turing, Alan 27
Twitter bots 173

uncertainty 118–121
  quantification of 207–208
  resilience 151
  terminology 242
  toleration of failures 206–207
underfitting 123
understandability 154
  Analysis Rubric 61–62
  causality 158–168
  communicating data science results 171–186
  COVID-19 mortality rates 72
  healthcare records 70
  interpretability, explainability, and auditability 154–158
  music recommendation systems 67
  protein folding 69
  reproducibility in scientific applications 168–171
  speech recognition 65
  spelling correction 63–64
underwriting 93–94, 104–105
universe, temperature of 89
universities
  data science courses 238–240
  transdisciplinary research 256
unknown-unknowns 151
unsupervised learning 24
usage communities 231
usage rights, versus privacy 128–131
user-generated data 117
users, data generation 112
utilities, energy usage data 134–135

V-Safe system 56–57
vaccine distribution optimization 85, 103
  see also COVID-19 vaccinations
validation 152
value, of web services 129–130
variable selection (causal relationships) 167–168
variance 122
verification 152
video game personalization 79, 198
video recommendations 83
virtuous cycle phenomena 248–250
visualization *see* information visualization
vocabulary, clarity in 241–243

walled gardens (personal information) 201
warfare 245
weather prediction 89–90
web search *see* search engines
wicked problems 277
World Wide Web *see* Internet

Zillow 208