Index

A/A tests, 200
  how to run, 205
  uneven splits and, 204
Above the fold time (AFT), 88
Acquisition, Activation, Retention, Referral, Revenue, 91
Agile software development, 13
  analysis
    automated, 76
    cohort, 241
    edge-level, 234
    logs-based, 129
    post-period, 242
    triggered, 159
analysis results
  review meetings, 62
analysis unit, 168
annotating data, 178
atomicity, 70
automated analysis, 76
backend algorithmic changes, 19
backend delay model, 87
Bayes rule, 186
Bayesian evaluation, 114
Bayesian structural time series analysis, 140
Benjamini-Hochberg procedure, 191
Bernoulli randomization, 231
  bias, 191, 240
  biases, 201
binarization, 197
blocking, 197
Bonferroni correction, 191
bootstrap, 169
bootstrap method, 195
bot filtering, 48

Campbell’s law, 109
  capping, 197
  carryover effects, 74
  cart recommendations, 17
  causal model, 96
  causal relationship, 96
  causality, 8, 137
Central Limit Theorem, 187
  centralized experimentation platform, 181
  churn rate, 8
  click logging, 178
  click tracking, 52
  client crashes metric, 99
  client-side instrumentation, 163
  cohort analysis, 241
  confidence interval, 30, 37, 187, 193
  confidence intervals, 43
  constraints-based design, 76
  constructed propensity score, 143
Control, 6–7
  cooking data, 77
  correlation, 9
  counterfactual logging, 73
  cultural norms, 61
data
  annotating, 178
  analysis pipeline, 151
  collection, 121
  computation, 178
  enrichment, 178
  pipeline impact, 47
  sharing, 65
  visualization, 77
day-of-week effect, 33
Index

deceptive correlations, 145
delayed experience, 237
delayed logging, 157
delta method, 169, 195
density, 199
dependent variable, 7
deploying experiments, 69
designated market area, 138
difference in differences, 143
driver metrics, 91
ecosystem impact, 231
delayed-level analysis, 234
educational processes, 61
empirical evidence, 114
equipose, 118
ethics, 116
anonymous, 123
corporate culture and, 122
equipose, 118
identified data, 123
risk, 118
exclusion restriction, 142
experiments
long-term, 61
experiment
objective, 6
OEC, 6
results, 181
experiment assignment, 71
experiment hypothesis, 112
experiment IDs, 67
experiment lifecycle, 67
experiment platform
performance, 72
experiment scorecard, 179, 216, See also: visualizations
experimentation maturity model, 180
experimentation maturity models
crawling, 59
flying, 59
running, 59
walking, 59
experimentation platform
centralizing, 181
experiments
A/A, 200
analysis, 67
analytics, 177
automated analysis, 76
best practices, 113
bias, 191
browser redirects, 45
channels, 5
client-side, 153
client-side implications, 156
constraints-based design, 76
culture and, 179
data collection, 121
deception, 120
deployment, 69
design, 32
design and analysis, 27
design example, 33
design platform, 58
determining length, 33
duration, 190
delayed-level analysis, 234
evaluation, 128
failure, 226
generating ideas for, 129
historical data retention, 231
holdback, 245
human evaluation and, 130
IDs, 69
impact, 174
infrastructure, 34
instrumentation, 34, 121, 162
interference, 226
interleaved, 141
isolating shared resources, 231
isolation, 231
iterations, 67
just-in-time processes, 61
length, 42
long-term effect, 236
nested design, 76
observation, 127
offline simulation and, 188
organizational goals and, 112
paired, 198
performance testing, 17
platform, 66
platform architecture, 68
platform components, 67
platform for managing, 67
power, 34, 189
power-of-suggestion, 120
production code, 70
randomization, 114
raters, 130
replication, 176
<table>
<thead>
<tr>
<th>Page</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>268</td>
<td></td>
</tr>
<tr>
<td></td>
<td>experiments (cont.)</td>
</tr>
<tr>
<td></td>
<td>replication experiment, 15</td>
</tr>
<tr>
<td></td>
<td>reusing, 231</td>
</tr>
<tr>
<td></td>
<td>reverse, 176, 245</td>
</tr>
<tr>
<td></td>
<td>risk, 118</td>
</tr>
<tr>
<td></td>
<td>sample size, 188, 197</td>
</tr>
<tr>
<td></td>
<td>scaling, 73</td>
</tr>
<tr>
<td></td>
<td>segments, 52</td>
</tr>
<tr>
<td></td>
<td>sensitivity, 28</td>
</tr>
<tr>
<td></td>
<td>server-side, 153</td>
</tr>
<tr>
<td></td>
<td>short-term effect, 235</td>
</tr>
<tr>
<td></td>
<td>side-by-side, 131</td>
</tr>
<tr>
<td></td>
<td>slow-down, 81, 86</td>
</tr>
<tr>
<td></td>
<td>traffic allocation, 33, 192</td>
</tr>
<tr>
<td></td>
<td>trustworthiness, 174</td>
</tr>
<tr>
<td></td>
<td>validation, 135</td>
</tr>
<tr>
<td></td>
<td>vocabulary, 179</td>
</tr>
<tr>
<td></td>
<td>when they are not possible, 137</td>
</tr>
<tr>
<td></td>
<td>external data services, 133</td>
</tr>
<tr>
<td></td>
<td>external validity, 33, 135</td>
</tr>
<tr>
<td></td>
<td>factor. See parameter</td>
</tr>
<tr>
<td></td>
<td>false discovery rate, 42</td>
</tr>
<tr>
<td></td>
<td>first-order actions, 230</td>
</tr>
<tr>
<td></td>
<td>Fisher’s meta-analysis, 192</td>
</tr>
<tr>
<td></td>
<td>focus groups, 132</td>
</tr>
<tr>
<td></td>
<td>gameability, 100, 107</td>
</tr>
<tr>
<td></td>
<td>geo-based randomization, 232</td>
</tr>
<tr>
<td></td>
<td>goal metrics, 91</td>
</tr>
<tr>
<td></td>
<td>goals, 91</td>
</tr>
<tr>
<td></td>
<td>alignment, 93</td>
</tr>
<tr>
<td></td>
<td>articulation, 93</td>
</tr>
<tr>
<td></td>
<td>Goodhart’s law, 109</td>
</tr>
<tr>
<td></td>
<td>granularity, 167</td>
</tr>
<tr>
<td></td>
<td>guardrail metrics, 35, 92, 159, 174, 219</td>
</tr>
<tr>
<td></td>
<td>cookie write rate, 224</td>
</tr>
<tr>
<td></td>
<td>latency, 81</td>
</tr>
<tr>
<td></td>
<td>organizational, 35, 98</td>
</tr>
<tr>
<td></td>
<td>quick queries, 225</td>
</tr>
<tr>
<td></td>
<td>trust-related, 35</td>
</tr>
<tr>
<td></td>
<td>HEART framework, 91</td>
</tr>
<tr>
<td></td>
<td>heterogeneous Treatment effects, 52</td>
</tr>
<tr>
<td></td>
<td>hierarchy of evidence, 9, 138</td>
</tr>
<tr>
<td></td>
<td>holdbacks, 175</td>
</tr>
<tr>
<td></td>
<td>holdouts, 175</td>
</tr>
<tr>
<td></td>
<td>HTML response size per page metrics, 99</td>
</tr>
<tr>
<td></td>
<td>human evaluation, 130</td>
</tr>
<tr>
<td></td>
<td>hypothesis testing, 185, 189</td>
</tr>
<tr>
<td></td>
<td>Type I/II errors, 189</td>
</tr>
<tr>
<td></td>
<td>ideas funnel, 127</td>
</tr>
<tr>
<td></td>
<td>independence assumption</td>
</tr>
<tr>
<td></td>
<td>violation, 203</td>
</tr>
<tr>
<td></td>
<td>independent identically distributed samples, 193</td>
</tr>
<tr>
<td></td>
<td>independently identically distributed, 195</td>
</tr>
<tr>
<td></td>
<td>information accessibility, 180</td>
</tr>
<tr>
<td></td>
<td>infrastructure, 34, 66</td>
</tr>
<tr>
<td></td>
<td>innovation productivity, 114</td>
</tr>
<tr>
<td></td>
<td>institutional memory, 63, 111, 181</td>
</tr>
<tr>
<td></td>
<td>Instrumental Variable method, 231</td>
</tr>
<tr>
<td></td>
<td>Instrumental Variables, 142</td>
</tr>
<tr>
<td></td>
<td>instrumentation, 34, 59–60, 67, 72, 128, 151, 162, 177</td>
</tr>
<tr>
<td></td>
<td>client-side, 163</td>
</tr>
<tr>
<td></td>
<td>corporate culture and, 165</td>
</tr>
<tr>
<td></td>
<td>server-side, 164</td>
</tr>
<tr>
<td></td>
<td>intellectual integrity, 63</td>
</tr>
<tr>
<td></td>
<td>interference, 174</td>
</tr>
<tr>
<td></td>
<td>detecting, 234</td>
</tr>
<tr>
<td></td>
<td>direct connections, 227</td>
</tr>
<tr>
<td></td>
<td>indirect connections, 228</td>
</tr>
<tr>
<td></td>
<td>interleaved experiments, 141</td>
</tr>
<tr>
<td></td>
<td>internal validity, 43</td>
</tr>
<tr>
<td></td>
<td>Interrupted Time Series, 139</td>
</tr>
<tr>
<td></td>
<td>invariants, 35</td>
</tr>
<tr>
<td></td>
<td>isolation, 231</td>
</tr>
<tr>
<td></td>
<td>JavaScript errors, 99</td>
</tr>
<tr>
<td></td>
<td>key metrics, 14</td>
</tr>
<tr>
<td></td>
<td>latency, 99, 135, 156</td>
</tr>
<tr>
<td></td>
<td>layer ID, 75</td>
</tr>
<tr>
<td></td>
<td>leadership buy-in, 59</td>
</tr>
<tr>
<td></td>
<td>learning effect, 243</td>
</tr>
<tr>
<td></td>
<td>least-squares regression model, 142</td>
</tr>
<tr>
<td></td>
<td>lifetime value, 95</td>
</tr>
<tr>
<td></td>
<td>log transformation, 197</td>
</tr>
<tr>
<td></td>
<td>logs, 164</td>
</tr>
<tr>
<td></td>
<td>common identifier, 164</td>
</tr>
<tr>
<td></td>
<td>logs, joining, 177</td>
</tr>
<tr>
<td></td>
<td>logs-based analyses, 129</td>
</tr>
<tr>
<td></td>
<td>long-term effects, 51</td>
</tr>
<tr>
<td></td>
<td>long-term holdbacks, 175</td>
</tr>
<tr>
<td></td>
<td>long-term holdouts, 175</td>
</tr>
<tr>
<td></td>
<td>long-term impact, 173</td>
</tr>
<tr>
<td></td>
<td>long-term Treatment effect, 235</td>
</tr>
<tr>
<td></td>
<td>lossy implementations, 46, 224</td>
</tr>
<tr>
<td></td>
<td>maturity models, 58</td>
</tr>
<tr>
<td></td>
<td>Maximum Power Ramp, 172</td>
</tr>
</tbody>
</table>
mean, 29
measuring impact, 61
meta-analysis, 78, 112
metrics, 14
  analysis unit, 169
  asset, 92
  binary, 197
  business, 92
categorizing, 181
clicks-per-user, 47
client crashes, 99
data quality, 92
debug, 62, 92
defining, 179
developing goal and driver, 94
diagnosis, 92
driver, 91
early indicator, 175
engagement, 92
evaluation, 96
feature-level, 104
gameability, 107
goal, 62, 91
guardrail, 35, 62, 81, 92, 159, 174, 219
how they relate to each other, 114
HTML response size per page, 99
improvements, 60
indirect, 91
invariants, 35
irrelevant metrics significance, 191
JavaScript errors, 99
logs-based, 164
longitudinal stability, 170
negative, 95
normalizing, 105
operational, 92
organizational, 91
organizational guardrail, 98
pageviews-per-user, 99
page-load-time, 18
per-experiment, 179
per-metric results, 181
per-user, 179
predictive, 91
quality, 62
related, 182
revenue-per-user, 99
Sample Ration Mismatch, 219
sensitivity, 103, 114
sessions-per-user, 18
short-term, 239
short-term revenue, 101
sign post, 91
statistical models and, 95
success, 91
surrogate, 91, 104
taxonomy, 90
t-test, 187
true north, 91
user-level, 195
validation, 96
variability, 29
minimum detectable effect, 190
model training, 229
multiple comparisons problem, 42
multiple hypothesis testing, 42
Multivariate Tests (MVTs), 7
nested design, 76
network effects, 237
network egocentric randomization, 233
network-cluster randomization, 233
NHST. See Null hypothesis significant testing
normality assumption, 188
novelty effect, 33, 49, 174
detecting, 51
Null hypothesis, 30, 106, 185, 192
conditioning, 40
Null hypothesis significant testing, 40
Null test. See A/A test
Objectives and Key Results, 90
observational study, 139
limitations of, 144
OEC. See overall evaluation criterion
clicks-per-user, 47
offline simulation, 188
One Metric that Matters, 104
online controlled experiments
website optimization example, 26
backend algorithmic changes, 19
benefits, 10
key tenets, 11
operational concerns, 173
organizational goals, 91
organizational guardrail metrics, 98
orthogonal randomization, 176
orthogonality guarantees, 71
Outcome, Evaluation and Fitness function, 7
outliers, 196
Index

270

overall evaluation criterion, 180
for e-mail, 106
for search engines, 108
revenue-per-user, 109
teams and, 112
triggering and, 212
overall evaluation criterion, 5, 27, 102
definition, 6
purchase indicator, 32

page-load-time (PLT), 88
Page phase time, 88
pageviews per-user metrics, 99
paired experiments, 198
parameter
definition, 7
parameters, 67
system, 70
peeking, 42
perceived performance, 88
percent delta, 194
performance, 135, 156, 179
impact on key metrics, 18, 82
performance testing, 17
per-metric results, 181
permutation test, 188
personal recommendations, 17
PIRATE framework, 91
platform architecture, 68
platform components, 67
platform tools for managing experiments, 67
population segments, 52
post-period analysis, 242
power, 189
primacy and novelty effects, 33
primacy effect, 33, 49, 174
detecting, 51
propensity score matching, 143
p-value, 30, 106, 178, 186, 193, 220
misinterpretation, 40
p-value threshold, 181
p-value thresholds, 77

query share, 108
ramping, 55, 66, 113, 151, 171, 234, 245
Maximum Power Ramp, 172
phase 1
pre-MPR, 174

phase 2
MPR, 174
phase 3
post-MPR, 175
phase 4
long-term effects, 175
randomization, 8
randomization unit, 65, 151, 195
definition, 7
functional, 170
granularity, 167
reading list, 24
Regression Discontinuity Design, 141
regression model, 142
related metrics, 182
replication, 176
replication experiment, 15
Response variable, 7
revenue-per-user metric, 99
reverse experiment, 245
reverse experiments, 176
rings of test populations, 174
risk mitigation, 173
Rubin causal model, 226
sample ratio mismatch, 45
Sample Ratio Mismatch, 215, 219
sample size, 188
sampling, 55
scaling, 73
manual methods, 74
numberline method, 74
single-layer, 73
single-layer method drawback, 74
scorecard, 7
scorecard visualizations, 180
search engine results page, 113
segments, 52, 178, 180
poorly behaving, 180
selection bias, 158
sensitivity, 103, 114, 196
sequential tests, 42
server-side instrumentation, 164
sessionized data, 129
shared goals, 60
shared resources, 44
short-term revenue metric, 101
short-term Treatment effect, 235
side-by-side experiments, 131
significance boundary, 32
Index

Simple Ratio Mismatch
debugging, 222
simple ratio mismatch, 180
Simpson’s paradox, 54
single-layer scaling, 73
skewness coefficient, 187
slow-down experiments, 81, 86
speed, 179
Speed Index, 88
speed, quality, and risk, 172
spurious correlations, 146
SRM. See sample ratio mismatch
Stable Unit Treatment Value Assumption, 43, 168, 226
standard error, 29
statistical power, 30, 185, 190, 198
statistics, 178
confidence interval, 187
practical significance, 189
two-sample t-tests, 185
surrogate metrics, 104
surveys, 132
SUTVA. See Stable Unit Treatment Value Assumption
system parameters, 70
system-learned effect, 243

Taylor-series linear approximation, 83
technical debt, 72
thick client, 153
thick clients, 151
thin client, 153
time to first result, 88
time-based effects, 140
time-based randomization, 232
time-staggered Treatment, 244
time-to-successful-click, 89
timing effects, 47
traffic allocation, 174

Treatment, 6
Treatment effect, 41, 175, 214, 236
learning, 243
system-learned, 243
time-staggered, 244
user-learned, 243
Treatment effect dilution, 240
triggering, 40, 47, 72, 159, 180, 209
attributes, 222
t-statistic, 186
Twymann’s law, 39
type I error rates, 41
Type I errors, 201
Type I/II errors, 189
uber holdouts, 176
user experience research, 95, 131
User Ready Time, 88
user-learned effect, 243
validation, 135
value of ideas, 11
variable. See also parameter
variables
Instrumental, 142
variance, 193
variance estimation, 236
variant
definition, 7
mapping, 7
variants, 6, 67
allocation, 27
assignment, 69
balanced, 198
visualization tool, 181
visualizations, 180
scorecard, 180
web beacons, 163
website performance, 84