



คำนำสำหรับการพิมพ์ครั้งที่สอง

ผมภูมิใจอย่างยิ่งกับหนังสือ Data Science from Scratch ฉบับแรก แต่การพัฒนาด้านวิทยาศาสตร์ข้อมูลในช่วงที่ผ่านมาทำให้ระบบนิเวศของ Python มีความก้าวหน้าไปมาก รวมถึงนักพัฒนาและผู้ศึกษาหาความรู้ ซึ่งได้เปลี่ยนแปลงความคิดเกี่ยวกับเนื้อหาของหนังสือเล่มแรกของผม

ดังนั้นผมจึงเขียนโปรแกรมและตัวอย่างทั้งหมดใหม่ โดยใช้ Python 3.6 (และคุณสมบัติใหม่หลายอย่าง) และเน้นการเขียนโปรแกรมที่ ‘คลีน’ มีการเปลี่ยนตัวอย่างที่ดู ‘เล่นๆ’ บางส่วนในฉบับแรกด้วยชุดโปรแกรมที่เหมือนจริงมากขึ้น โดยใช้ชุดข้อมูล “ของจริง” และเพิ่มเนื้อหาใหม่ในหัวข้อ เช่น deep learning และการประมวลผลภาษาธรรมชาติ ให้สอดคล้องกับสิ่งต่างๆ ที่นักวิทยาศาสตร์ข้อมูลในปัจจุบันทำอยู่ (อีกทั้งยังลบบางส่วนที่ไม่ค่อยเกี่ยวข้องออกไป) แก้ไขข้อผิดพลาด และเขียนคำอธิบายให้ชัดเจนกว่าเดิม

แม้ฉบับพิมพ์ครั้งแรกจะเป็นหนังสือที่ยอดเยี่ยม แต่ฉบับนี้จะดียิ่งขึ้น ซึ่งรวมถึงความสนุกด้วย!

Joel Grus

ซีแอตเทิล

2019

สัญลักษณ์ในเล่ม



แสดงถึง เคล็ดลับหรือข้อเสนอแนะ



แสดงถึง บันทึกหรือหมายเหตุ



แสดงถึง คำเตือนหรือข้อควรระวัง

การใช้โปรแกรมในหนังสือเล่มนี้

สำหรับเนื้อหาเพิ่มเติม (ตัวอย่างโปรแกรม, แบบฝึกหัด และอื่นๆ) มีให้ดาวน์โหลดที่ <https://github.com/joelgrus/data-science-from-scratch> ซึ่งโดยทั่วไป คุณสามารถใช้ตัวอย่างโปรแกรมในหนังสือเล่มนี้ในโปรแกรมและเอกสารของคุณ หรือใช้ในการตอบคำถามด้วยการอ้างถึงหนังสือเล่มนี้และโปรแกรมตัวอย่าง โดยไม่จำเป็นต้องขออนุญาต เว้นแต่มีการแก้ไขส่วนที่สำคัญของโปรแกรม อย่างไรก็ตามการขายหรือแจกจ่าย CD-ROM ที่มีตัวอย่างโปรแกรมจากหนังสือของ O'Reilly หรือการนำโปรแกรมตัวอย่างจากหนังสือเล่มนี้ไปใช้ในเอกสารประกอบผลิตภัณฑ์ของคุณ จะต้องได้รับการอนุญาตก่อนเท่านั้น



คุณไม่จำเป็นต้องอ้างอิงแหล่งที่มา แต่เราขอขอบคุณ หากคุณจะอ้างอิง การระบุแหล่งที่มา มักประกอบด้วย ชื่อหนังสือ, ชื่อผู้เขียน, สำนักพิมพ์ และ ISBN ตัวอย่างเช่น “Data Science from Scratch, Second Edition โดย Joel Grus (O’Reilly) Copyright 2019 Joel Grus, 978-1-492-04113-9”

หากต้องการขออนุญาตใช้งานโปรแกรมตัวอย่าง โปรดติดต่อที่ permissions@oreilly.com

คำขอบคุณ

ก่อนอื่นผมขอขอบคุณ Mike Loukides ในการจัดพิมพ์หนังสือเล่มนี้ และขอขอบคุณ บรรณาธิการ Michele Cronin และ Marie Beaugureau ที่ให้คำแนะนำจนเป็นหนังสือที่น่าอ่านกว่า ตอนที่เริ่มต้นฉบับของผมมาก

ผมคงไม่สามารถเขียนหนังสือเล่มนี้ได้ ถ้าไม่เคยเรียนวิชาวิทยาศาสตร์ข้อมูลมาก่อน และการเรียนวิชาวิทยาศาสตร์ข้อมูลของผมได้รับอิทธิพลจาก Dave Hsu, Igor Tatarinov, John Rauser และคนอื่นๆ ในแก๊ง Farecast รวมทั้งผู้คนที่ Coursera และ DataTau ที่สมควรได้รับเครดิตเช่นกัน

ขอขอบคุณผู้อ่านตรวจทานของผม Jay Fundling ซึ่งพบข้อผิดพลาดและชี้ให้เห็นคำอธิบายที่ไม่ชัดเจนหลายแห่ง และทำให้หนังสือเล่มนี้ดีขึ้น (และถูกต้องมากขึ้นด้วย) ขอขอบคุณ Debashis Ghosh ผู้เป็นฮีโร่สำหรับการตรวจทานอย่างรอบคอบ Andrew Musselman ผู้ให้แง่คิดเกี่ยวกับภาษา R และ Python รวมถึง Trey Causey, Ryan Matthew Balfanz, Loris Mularoni, Núria Pujol, Rob Jefferson, Mary Pat Campbell, Zach Geary, Denise Mauldin, Jimmy O’Donnell และ Wendy Grus ผู้ให้คำแนะนำที่มีค่า ขอขอบคุณทุกคนที่อ่านฉบับพิมพ์ครั้งแรก และช่วยทำให้ฉบับพิมพ์ครั้งที่สองเล่มนี้ดีขึ้น และแน่นอนว่า ข้อผิดพลาดที่เหลืออยู่ในหนังสือเล่มนี้ เป็นความผิดชอบของผม

ขอขอบคุณชุมชน #datascience ใน Twitter ที่ให้แนวคิดใหม่ๆ รวมถึงการแนะนำให้รู้จักกับคนเก่งๆ หลายคน ขอขอบคุณเป็นพิเศษกับ Trey Causey (อีกครั้ง) สำหรับการเตือนให้ผมใส่บทพิสูจน์เชิงเส้น (โดยไม่ได้ตั้งใจ) และ Sean J. Taylor ที่ชี้ให้เห็นช่องว่างขนาดใหญ่ในบท “การทำงานกับข้อมูล” (โดยไม่ได้ตั้งใจเช่นกัน)

เหนือสิ่งอื่นใด ต้องขอบคุณ Ganga และ Madeline เพราะสิ่งเดียวที่ยากกว่าการเขียนหนังสือคือการอยู่กับคนที่กำลังเขียนหนังสือ ซึ่งผมคงไม่สามารถทำงานชิ้นนี้ให้สำเร็จได้ หากไม่ได้รับการสนับสนุนจากทั้งสองคน



คำนำสำหรับการพิมพ์ครั้งแรก

วิทยาศาสตร์ข้อมูล

นักวิทยาศาสตร์ข้อมูลถูกขนานนามว่า “เป็นงานที่เซ็กซี่ที่สุดในศตวรรษที่ 21” เป็นสาขาวิชาที่ร้อนแรง แต่วิทยาศาสตร์ข้อมูลคืออะไร? ท้ายที่สุดเราจะไม่สามารถผลิตนักวิทยาศาสตร์ข้อมูลได้ ถ้าไม่รู้ว่าคุณสมบัติของข้อมูลคืออะไร ตามแผนภาพเวนน์ (Venn) ที่มีชื่อเสียง ข้อมูลจะอยู่ที่จุดตัดของ

- ทักษะในการ ‘เจาะ’
- ความรู้ทางคณิตศาสตร์และสถิติ
- แก่นแท้ของความเชี่ยวชาญ

แม้ว่าในตอนแรกผมตั้งใจจะเขียนหนังสือที่ครอบคลุมทั้งหมด แต่ก็ตระหนักได้อย่างรวดเร็วว่าการพูดถึง “แก่นแท้ของความเชี่ยวชาญ” อย่างละเอียดจะต้องใช้เนื้อที่มหาศาล ผมจึงตัดสินใจมุ่งเน้นไปที่สองหัวข้อแรก ด้วยเป้าหมายในการช่วยให้คุณพัฒนาทักษะเริ่มต้นสำหรับการ ‘เจาะ’ ข้อมูล และช่วยให้คุณคุ้นเคยกับคณิตศาสตร์และสถิติที่เป็นแก่นของวิทยาศาสตร์ข้อมูล

การอ่านหนังสือเล่มนี้จะทำให้มีความเข้าใจเกี่ยวกับวิธีที่ผม ‘เจาะ’ ข้อมูลต่างๆ ซึ่งอาจไม่ใช่วิธีที่ดีที่สุดสำหรับคุณ และได้เข้าใจเกี่ยวกับเครื่องมือที่ผมใช้ ซึ่งก็ไม่จำเป็นต้องเป็นเครื่องมือที่ดีที่สุดเช่นกัน อีกทั้งยังได้รับความเข้าใจเกี่ยวกับวิธีเข้าถึงปัญหาของข้อมูล

เราจะไม่สามารถทำงานด้านวิทยาศาสตร์ข้อมูลได้ โดยไม่เข้าใจเรื่องความน่าจะเป็นและสถิติ รวมถึงพีชคณิตเชิงเส้น ซึ่งในจังหวะที่เหมาะสม เราจะลงลึกในสมการทางคณิตศาสตร์เป็นบางครั้ง และผมหวังว่าคุณจะไม่กลัวที่จะดำน้ำกับผมในจังหวะนั้น และยังหวังจะให้ความรู้สีกว่าการเล่นกับข้อมูลเป็นเรื่องสนุก เพราะมันเป็นเรื่องสนุกจริงๆ !

เริ่มจากศูนย์

ปัจจุบันมีไลบรารีและเครื่องมือด้านวิทยาศาสตร์ข้อมูลจำนวนมาก แต่ในหนังสือเล่มนี้เราจะเข้าสู่วิทยาศาสตร์ข้อมูลกันตั้งแต่เริ่มต้น นั่นหมายความว่า เราจะสร้างเครื่องมือและอัลกอริธึมด้วยตัวเอง เพื่อทำความเข้าใจได้ดียิ่งขึ้น แม้ว่าสิ่งเหล่านี้จะทำงานได้ดีกับชุดข้อมูลตัวอย่างขนาดเล็กๆ เท่านั้น

มีการถกเถียงกันถึงภาษาที่ดีที่สุดสำหรับการเรียนรู้วิทยาศาสตร์ข้อมูล หลายคนเชื่อว่าเป็นภาษา R (สำหรับการโปรแกรมทางสถิติ) มีคนแนะนำ Java หรือ Scala บ้าง แต่อย่างไรก็ตามในความคิดของผม Python เป็นตัวเลือกที่ชัดเจนที่สุด ด้วยคุณสมบัติหลายอย่างที่เหมาะสมสำหรับการเรียนรู้ (และการปฏิบัติ) วิทยาศาสตร์ข้อมูล

- ฟรี
- เขียนโปรแกรมค่อนข้างง่าย
- มีไลบรารีที่เกี่ยวข้องกับวิทยาศาสตร์ข้อมูลที่มีประโยชน์มากมาย

ในช่วงหลายปีที่ผ่านมา ผมได้ฝึกฝนนักวิทยาศาสตร์ข้อมูลจำนวนหนึ่ง และผมก็เชื่อว่า ใครก็ตามที่มีความถนัดทางคณิตศาสตร์และทักษะการเขียนโปรแกรมระดับหนึ่ง ถือว่ามีเครื่องมือที่จำเป็นสำหรับการทำงานวิทยาศาสตร์ข้อมูลแล้ว ส่วนที่เหลือคือ จิตใจที่อยากรู้อยากเห็น ความตั้งใจในการทำงานหนัก และหนังสือเล่มนี้



สารบัญ

บทนำ	15
อิทธิพลของข้อมูล	15
วิทยาศาสตร์ข้อมูล (Data Science) คืออะไร?	15
สมมติฐานสร้างแรงบันดาลใจ: DataSciencester	16
ค้นหาตัวเชื่อมต่อที่สำคัญ	16
นักวิทยาศาสตร์ข้อมูลที่คุณอาจรู้จัก	19
รายได้และประสบการณ์	22
ประเมินการชำระเงิน	24
หัวข้อที่น่าสนใจ	25
ก้าวต่อไป	26
 ภาษา Python ฉบับรวบรัด	29
หลักการของ Python	29
เริ่มใช้งาน Python	30
สภาพแวดล้อมเสมือนจริง	30
การจัดรูปแบบช่องว่าง	31
โมดูล	32
ฟังก์ชัน	32
ข้อความ	33
Exception	34
ลิสต์ (List)	34
Tuple	36
Dictionary	36
defaultdict	37
Counter	38
Set	39
ควบคุมการทำงาน	39
ค่าความจริง	40
การเรียงลำดับ	41
การแปลงลิสต์	42
การทดสอบและยืนยัน	42
การเขียนโปรแกรมเชิงวัตถุ (OOP)	43
การทำซ้ำและตัวสร้างค่า	44
การสุ่ม	46
Regular Expression	47
การเขียนโปรแกรมเชิงฟังก์ชัน	47



zip และการแยก Argument	47
args และ kwargs	48
การระบุชนิด	49
วิธีระบุชนิด	51
ยินดีต้อนรับสู่สังคม DataSciencester!	53
สำหรับการค้นคว้าเพิ่มเติม	53
แสดงข้อมูลด้วยภาพ	55
matplotlib	55
แผนภูมิแท่ง	57
แผนภูมิเส้น	60
แผนภูมิกระจาย	61
สำหรับการค้นคว้าเพิ่มเติม	63
พีชคณิตเชิงเส้น	65
เวกเตอร์	65
เมทริกซ์	70
สำหรับการค้นคว้าเพิ่มเติม	73
สถิติ	75
การอธิบายถึงชุดข้อมูล	75
แนวโน้มสู่ส่วนกลาง	77
การกระจายของข้อมูล	78
ค่าสหสัมพันธ์ (Correlation)	80
Paradox ของซิมป์สัน	82
เงื่อนไขอื่นๆ ของความสัมพันธ์	83
ความสัมพันธ์และสาเหตุ	84
สำหรับการค้นคว้าเพิ่มเติม	85
ความน่าจะเป็น	87
การขึ้นต่อกันและความเป็นอิสระ	87
ความน่าจะเป็นที่มีเงื่อนไข	88
ทฤษฎีบทของเบย์	89
ตัวแปรสุ่ม	90
การกระจายแบบต่อเนื่อง	90
การกระจายแบบปกติ	91
ทฤษฎีแนวโน้มเข้าสู่ศูนย์กลาง	94
สำหรับการค้นคว้าเพิ่มเติม	96



สมมติฐานและการอนุมาน	99
การทดสอบสมมติฐานทางสถิติ	99
ตัวอย่าง: การโยนเหรียญ	99
p-Value	102
ช่วงของความเชื่อมั่น	104
p-Hacking	104
ตัวอย่าง: ทำการทดสอบ A/B	105
การอนุมานแบบเบย์	106
สำหรับการค้นคว้าเพิ่มเติม	109
 การเคลื่อนลงตามความชัน	111
แนวคิดของการไล่ระดับลง	111
การประมาณระดับความชัน	112
การใช้งาน	115
การเลือกระยะที่ถูกต้อง	116
ใช้ Gradient Descent เพื่อหาแบบจำลองที่ดี	116
การคำนวณแบบ Minibatch และ Stochastic	117
สำหรับการค้นคว้าเพิ่มเติม	119
 การได้มาซึ่งข้อมูล	121
stdin และ stdout	121
การอ่านไฟล์	123
พื้นฐานของไฟล์ข้อความ	123
ไฟล์แบบมีตัวค้น	125
การเก็บข้อมูลจากเว็บ	126
HTML และการแยกวิเคราะห์	126
ตัวอย่าง: การเก็บข้อมูลของสภาคองเกรส	128
การใช้ API	131
JSON และ XML	131
การใช้ API ที่ไม่มีการตรวจสอบผู้ใช้	132
การค้นหา API	133
ตัวอย่าง: การใช้ Twitter API	133
การลงทะเบียนกับ Twitter	133
สำหรับการค้นคว้าเพิ่มเติม	137
 การทำงานกับข้อมูล	139
สำรวจข้อมูลของคุณ	139
สำรวจข้อมูลมิติเดียว	139
ข้อมูลสองมิติ	141



ข้อมูลหลายมิติ	143
การใช้ NamedTuple	145
Dataclass	146
ทำข้อมูลให้สะอาด	147
การจัดการข้อมูล	149
การปรับมาตราส่วนข้อมูล	152
แสดงความคืบหน้าด้วย tqdm	154
การลดมิติของข้อมูล	155
สำหรับการค้นคว้าเพิ่มเติม	161
Machine Learning	163
การสร้างแบบจำลอง	163
แมชีนเลิร์นนิงคืออะไร?	164
Overfitting และ Underfitting	164
ความถูกต้อง	167
จุดลงตัวของฟังก์ชันการสูญเสียและความแปรปรวน	169
การดึงและคัดเลือกคุณลักษณะ	170
สำหรับการค้นคว้าเพิ่มเติม	171
k-Nearest Neighbors	173
แบบจำลอง	173
ตัวอย่าง: ชุดข้อมูล Iris	175
คำสาปของมิติข้อมูล	178
สำหรับการค้นคว้าเพิ่มเติม	182
Naive Bayes	185
ตัวกรองสแปมแนสท์	185
ตัวกรองสแปมที่ซับซ้อนขึ้น	186
สร้าง Classifier	187
ทดสอบแบบจำลอง	189
เริ่มใช้งานแบบจำลอง	190
สำหรับการค้นคว้าเพิ่มเติม	193
การถดถอยเชิงเส้นอย่างง่าย	195
แบบจำลอง	195
การเคลื่อนลงตามความชัน (Gradient Descent)	198
Maximum Likelihood Estimation	199
สำหรับการค้นคว้าเพิ่มเติม	200



การถดถอยพหุคูณ	203
แบบจำลอง	203
สมมติฐานเพิ่มเติมของแบบจำลองค่ายกกำลังสองที่น้อยที่สุด	204
ปรับแบบจำลองให้เหมาะสม	204
การแปลความแบบจำลอง	206
ความเหมาะสม	207
ขออนอกเรื่อง: Bootstrap	207
ค่าผิดพลาดมาตรฐานของค่าสัมประสิทธิ์การถดถอย	209
การทำให้โมเดลซับซ้อนน้อยลง	210
สำหรับการค้นคว้าเพิ่มเติม	213
 การถดถอยโลจิสติก	215
ปัญหา	215
ฟังก์ชันโลจิสติก	217
การนำแบบจำลองไปใช้	220
ความเหมาะสม	221
Support Vector Machine	222
สำหรับการค้นคว้าเพิ่มเติม	225
 Decision Tree	227
ต้นไม้สำหรับการตัดสินใจคืออะไร?	227
การวัดค่าความไม่แน่นอน	228
เอนโทรปีของพาร์ทิชัน	230
การสร้างดีซิชันทรี	231
รวมทั้งหมดเข้าด้วยกัน	234
โมเดล Random Forest	237
สำหรับการค้นคว้าเพิ่มเติม	238
 โครงข่ายประสาทเทียม	241
Perceptron	241
เครือข่ายประสาทแบบ Feed-Forward	243
Backpropagation	246
ตัวอย่าง: Fizz Buzz	248
สำหรับการค้นคว้าเพิ่มเติม	251
 Deep Learning	253
Tensor	253
โครงสร้างแบบเลเยอร์	255
เลเยอร์เชิงเส้น	257



โครงข่ายประสาทเทียมคืออนุกรมของเลเยอร์	260
การสูญเสีย (Loss) และการเพิ่มประสิทธิภาพ (Optimization)	261
ตัวอย่าง: XOR อีกครั้ง	264
ฟังก์ชันอื่นในการประมวลผล	265
ตัวอย่าง: FizzBuzz อีกครั้ง	267
Softmax และ Cross-Entropy	268
การละทิ้ง (Dropout)	271
ตัวอย่าง: MNIST	272
การเซฟและโหลดแบบจำลอง	276
สำหรับการค้นคว้าเพิ่มเติม	277
การจัดกลุ่ม	279
แนวคิด	279
แบบจำลอง	279
ตัวอย่าง: นิตจัดเลี้ยง	282
การเลือกค่า k	284
ตัวอย่าง: การจัดกลุ่มสี	285
การจัดกลุ่มลำดับชั้นจากล่างขึ้นบน	287
สำหรับการค้นคว้าเพิ่มเติม	293
การประมวลผลภาษาธรรมชาติ	295
กลุ่มคำ	295
โมเดลภาษาแบบ n-Gram	297
ไวยากรณ์	300
การสุ่มตัวอย่างแบบกิบส์	302
การสร้างแบบจำลองของหัวข้อ	303
เวกเตอร์คำ	309
เครือข่ายประสาทที่เกิดขึ้นซ้อนกัน	318
ตัวอย่าง: การใช้ RNN ในระดับตัวอักษร	320
สำหรับการค้นคว้าเพิ่มเติม	324
การวิเคราะห์เครือข่าย	327
ความเป็นศูนย์กลางระหว่างโหนด	327
หาค่าความเป็นศูนย์กลางด้วยเวกเตอร์ลักษณะเฉพาะ	332
การคูณเมทริกซ์	332
ค่าความเป็นศูนย์กลาง	333
กราฟระบุทิศทางและ PageRank	335
สำหรับการค้นคว้าเพิ่มเติม	337



ระบบให้คำแนะนำ	339
ถ้าจะทำได้ด้วยตัวเอง	340
การแนะนำสิ่งที่ได้รับความนิยม	340
การกรองโดยพิจารณาผู้ใช้	341
การกรองโดยพิจารณาหัวข้อ	344
การแยกตัวประกอบเมทริกซ์	346
สำหรับการค้นคว้าเพิ่มเติม	351
 ฐานข้อมูล และ SQL	353
การสร้างตารางและเพิ่มข้อมูล	353
UPDATE	356
DELETE	357
SELECT	358
GROUP BY	360
ORDER BY	363
JOIN	364
การสอบถามย่อย (Subquery)	366
อินเด็กซ์	367
การเพิ่มประสิทธิภาพการสอบถาม	367
NoSQL	368
สำหรับการค้นคว้าเพิ่มเติม	369
 MapReduce	371
ตัวอย่าง: นับจำนวนคำ	371
ทำไมต้องใช้ MapReduce	373
MapReduce ที่มีความทั่วไปมากขึ้น	373
ตัวอย่าง: วิเคราะห์การอัปเดตสเตตัส	375
ตัวอย่าง: การคูณเมทริกซ์	377
ตัวรวบรวม	379
สำหรับการค้นคว้าเพิ่มเติม	379
 จริยธรรมด้านข้อมูล	381
จริยธรรมข้อมูลคืออะไร?	381
จริงๆ แล้วจริยธรรมข้อมูลคืออะไร?	381
เราควรใส่ใจเรื่องจริยธรรมในการใช้ข้อมูลหรือไม่?	382
การสร้างผลิตภัณฑ์ที่ไม่ดี	382
สมดุลของความแม่นยำและความเป็นส่วนตัว	382
การร่วมมือ	384
ความสามารถในการอธิบายการทำงาน	384



คำแนะนำ	384
ข้อมูลที่เอนเอียง	385
การปกป้องข้อมูล	385
สรุป	386
สำหรับการค้นคว้าเพิ่มเติม	386
ก้าวไปกับงานวิทยาศาสตร์ข้อมูล	389
IPython	389
คณิตศาสตร์	389
ไม่ได้มาแต่เริ่มแรก	389
NumPy	390
pandas	390
scikit-learn	390
การแสดงผลด้วยภาพ	390
R	390
Deep Learning	391
ค้นหาข้อมูล	391
โครงการที่สนใจ	391
Hacker News	391
ระดับเพลิง	391
เสื้อเชิ้ต	392
ทวีตบนโลก	392
แล้วคุณล่ะ?	392