
Elegant SciPy

The Art of Scientific Python

*Juan Nunez-Iglesias, Stéfan van der Walt,
and Harriet Dashnow*

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'REILLY®

Table of Contents

Preface.....	vii
1. Elegant NumPy: The Foundation of Scientific Python.....	1
Introduction to the Data: What Is Gene Expression?	2
NumPy N-Dimensional Arrays	6
Why Use ndarrays Instead of Python Lists?	8
Vectorization	10
Broadcasting	10
Exploring a Gene Expression Dataset	12
Reading in the Data with pandas	12
Normalization	14
Between Samples	14
Between Genes	21
Normalizing Over Samples and Genes: RPKM	24
Taking Stock	30
2. Quantile Normalization with NumPy and SciPy.....	31
Getting the Data	33
Gene Expression Distribution Differences Between Individuals	34
Biclustering the Counts Data	37
Visualizing Clusters	39
Predicting Survival	42
Further Work: Using the TCGA's Patient Clusters	46
Further Work: Reproducing the TCGA's clusters	46
3. Networks of Image Regions with ndimage.....	49
Images Are Just NumPy Arrays	50
Exercise: Adding a Grid Overlay	55

Filters in Signal Processing	56
Filtering Images (2D Filters)	63
Generic Filters: Arbitrary Functions of Neighborhood Values	66
Exercise: Conway's Game of Life	67
Exercise: Sobel Gradient Magnitude	68
Graphs and the NetworkX library	68
Exercise: Curve Fitting with SciPy	72
Region Adjacency Graphs	73
Elegant ndimage: How to Build Graphs from Image Regions	76
Putting It All Together: Mean Color Segmentation	78
4. Frequency and the Fast Fourier Transform.	81
Introducing Frequency	81
Illustration: A Birdsong Spectrogram	84
History	90
Implementation	91
Choosing the Length of the DFT	92
More DFT Concepts	94
Frequencies and Their Ordering	94
Windowing	100
Real-World Application: Analyzing Radar Data	105
Signal Properties in the Frequency Domain	111
Windowing, Applied	115
Radar Images	117
Further Applications of the FFT	122
Further Reading	122
Exercise: Image Convolution	123
5. Contingency Tables Using Sparse Coordinate Matrices.	125
Contingency Tables	127
Exercise: Computational Complexity of Confusion Matrices	128
Exercise: Alternative Algorithm to Compute the Confusion Matrix	128
Exercise: Multiclass Confusion Matrix	128
scipy.sparse Data Formats	129
COO Format	129
Exercise: COO Representation	130
Compressed Sparse Row Format	130
Applications of Sparse Matrices: Image Transformations	133
Exercise: Image Rotation	138
Back to Contingency Tables	139
Exercise: Reducing the Memory Footprint	140
Contingency Tables in Segmentation	140

Information Theory in Brief	142
Exercise: Computing Conditional Entropy	144
Information Theory in Segmentation: Variation of Information	145
Converting NumPy Array Code to Use Sparse Matrices	147
Using Variation of Information	149
Further Work: Segmentation in Practice	156
6. Linear Algebra in SciPy	157
Linear Algebra Basics	157
Laplacian Matrix of a Graph	158
Exercise: Rotation Matrix	159
Laplacians with Brain Data	165
Exercise: Showing the Affinity View	170
Exercise Challenge: Linear Algebra with Sparse Matrices	170
PageRank: Linear Algebra for Reputation and Importance	171
Exercise: Dealing with Dangling Nodes	176
Exercise: Equivalence of Different Eigenvector Methods	176
Concluding Remarks	176
7. Function Optimization in SciPy	177
Optimization in SciPy: <code>scipy.optimize</code>	179
An Example: Computing Optimal Image Shift	180
Image Registration with Optimize	186
Avoiding Local Minima with Basin Hopping	190
Exercise: Modify the align Function	190
“What Is Best?”: Choosing the Right Objective Function	191
8. Big Data in Little Laptop with Toolz	199
Streaming with <code>yield</code>	200
Introducing the Toolz Streaming Library	203
k-mer Counting and Error Correction	206
Currying: The Spice of Streaming	210
Back to Counting k-mers	212
Exercise: PCA of Streaming Data	214
Markov Model from a Full Genome	214
Exercise: Online Unzip	217
Epilogue	221
Appendix: Exercise Solutions	225
Index	247