Explaining the Success of Nearest Neighbor Methods in Prediction

Full text available at: http://dx.doi.org/10.1561/2200000064

Other titles in Foundations and Trends® in Machine Learning

Non-convex Optimization for Machine Learningy
Prateek Jain and Purushottam Ka
ISBN: 978-1-68083-368-3

Kernel Mean Embedding of Distributions: A Review and Beyond
Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur and Bernhard Scholkopf
ISBN: 978-1-68083-288-4

Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 1 Low-Rank Tensor Decompositions
Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama and Danilo P. Mandic
ISBN: 978-1-68083-222-8

Tensor Networks for Dimensionality Reduction and Large-scale Optimization: Part 2 Applications and Future Perspectives
Andrzej Cichocki, Anh-Huy Phan, Qibin Zhao, Namgil Lee, Ivan Oseledets, Masashi Sugiyama and Danilo P. Mandic
ISBN: 978-1-68083-276-1

Patterns of Scalable Bayesian Inference
Elaine Angelino, Matthew James Johnson and Ryan P. Adams
ISBN: 978-1-68083-218-1

Generalized Low Rank Models
Madeleine Udell, Corinne Horn, Reza Zadeh and Stephen Boyd
ISBN: 978-1-68083-140-5

Explaining the Success of Nearest Neighbor Methods in Prediction

George H. Chen
Carnegie Mellon University
georgechen@cmu.edu

Devavrat Shah
Massachusetts Institute of Technology
devavrat@mit.edu

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

G. H. Chen and D. Shah. *Explaining the Success of Nearest Neighbor Methods in Prediction*. Foundations and Trends® in Machine Learning, vol. 10, no. 5-6, pp. 337–588, 2018.

ISBN: 978-1-68083-455-0

Foundations and Trends® in Machine Learning
Volume 10, Issue 5-6, 2018

Full text available at: http://dx.doi.org/10.1561/2200000064

Editorial Scope

Topics

Foundations and Trends® in Machine Learning publishes survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2018, Volume 10, 6 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Contents

1 Introduction **3**

1.1 Explaining the Popularity of Nearest Neighbor Methods . . 5

1.2 Nearest Neighbor Methods in Theory 8

1.3 The Scope of This Monograph 9

2 Background **13**

2.1 Regression 14

2.2 Classification 17

2.3 Nearest Neighbor and Kernel Regression 21

2.4 Nearest Neighbor and Kernel Classification 24

3 Theory on Regression **26**

3.1 Overview of Results 27

3.2 Key Definitions, Technicalities, and Assumptions 40

3.3 Theoretical Guarantees for k-NN Regression 45

3.4 Theoretical Guarantees for Fixed-Radius NN Regression . . 61

3.5 Theoretical Guarantees for Kernel Regression 66

3.6 Proofs 68

3.7 Choice of Number of Nearest Neighbors and Kernel Bandwidth 100

4 Theory on Classification **105**

4.1 Converting Guarantees from Regression to Classification . 108

Full text available at: http://dx.doi.org/10.1561/2200000064

4.2 Guarantees Under Weaker Assumptions 112
4.3 Guarantees Under a Margin Bound Condition 128

5 Prediction Guarantees in Three Contemporary Applications 131
5.1 Time Series Forecasting 139
5.2 Online Collaborative Filtering 160
5.3 Patch-Based Image Segmentation 180
5.4 More Training Data, More Problems? 197

6 Computation 200
6.1 Overview 200
6.2 Exact Nearest Neighbors 204
6.3 Approximate Nearest Neighbors 206
6.4 Open Source Software 219

7 Adaptive Nearest Neighbors and Far Away Neighbors 221
7.1 Adaptive Nearest Neighbor Methods 223
7.2 Far Away Neighbors 228

References 237

Full text available at: http://dx.doi.org/10.1561/2200000064

Explaining the Success of Nearest Neighbor Methods in Prediction

George H. Chen[1] and Devavrat Shah[2]

[1] *Carnegie Mellon University; georgechen@cmu.edu*
[2] *Massachusetts Institute of Technology; devavrat@mit.edu*

ABSTRACT

Many modern methods for prediction leverage nearest neighbor search to find past training examples most similar to a test example, an idea that dates back in text to at least the 11th century and has stood the test of time. This monograph aims to explain the success of these methods, both in theory, for which we cover foundational nonasymptotic statistical guarantees on nearest-neighbor-based regression and classification, and in practice, for which we gather prominent methods for approximate nearest neighbor search that have been essential to scaling prediction systems reliant on nearest neighbor analysis to handle massive datasets. Furthermore, we discuss connections to learning distances for use with nearest neighbor methods, including how random decision trees and ensemble methods learn nearest neighbor structure, as well as recent developments in crowdsourcing and graphons.

In terms of theory, our focus is on nonasymptotic statistical guarantees, which we state in the form of how many training data and what algorithm parameters ensure that a nearest neighbor prediction method achieves a user-specified error tolerance. We begin with the most general of such results

George H. Chen and Devavrat Shah (2018), “Explaining the Success of Nearest Neighbor Methods in Prediction”, Foundations and Trends® in Machine Learning: Vol. 10, No. 5-6, pp 337–588. DOI: 10.1561/2200000064.

for nearest neighbor and related kernel regression and classification in general metric spaces. In such settings in which we assume very little structure, what enables successful prediction is smoothness in the function being estimated for regression, and a low probability of landing near the decision boundary for classification. In practice, these conditions could be difficult to verify empirically for a real dataset. We then cover recent theoretical guarantees on nearest neighbor prediction in the three case studies of time series forecasting, recommending products to people over time, and delineating human organs in medical images by looking at image patches. In these case studies, clustering structure, which is easier to verify in data and more readily interpretable by practitioners, enables successful prediction.

1

Introduction

Things that appear similar are likely similar. For example, a baseball player's future performance can be predicted by comparing the player to other similar players (Silver, 2003). When forecasting election results for a U.S. state, accounting for polling trends at similar states improves forecast accuracy (Silver, 2008). In image editing, when removing an object from an image, one of the most successful ways to fill in the deleted pixels is by completing the missing pixels using image patches similar to the ones near the missing pixels (Criminisi *et al.*, 2004). These are but a few examples of how finding similar instances or *nearest neighbors* help produce predictions. Of course, this idea is hardly groundbreaking, with nearest neighbor classification already appearing as an explanation for visual object recognition in a medieval text *Book of Optics* by acclaimed scholar Alhazen in the early 11th century.[1] Despite their simplicity and

[1] A brief history of nearest neighbor classification and its appearance in Alhazen's *Book of Optics* is given by Pelillo (2014). The exact completion date of *Optics* is unknown. Al-Khalili (2015) dates the work to be from years 1011 to 1021, coinciding with much of Alhazen's decade of imprisonment in Cairo, while Smith (2001) claims a completion time between 1028 and 1038, closer to Alhazen's death circa 1040.

age, nearest neighbor methods remain extremely popular,[2] often used as a critical cog in a larger prediction machine. In fact, the machine can be biological, as there is now evidence that fruit flies' neural circuits execute approximate nearest neighbor in sensing odors as to come up with an appropriate behavioral response (Dasgupta *et al.*, 2017).

Although nearest neighbor classification dates back a millennium, analysis for when and why it works did not begin until far more recently, starting with a pair of unpublished technical reports by Fix and Hodges (1951; 1952) on asymptotic convergence properties as well as a small dataset study, followed by the landmark result of Cover and Hart (1967) that showed that k-nearest neighbors classification achieves an error rate that is at most twice the best error rate achievable. Decades later, Cover recollected how his paper with Hart came about:

> Early in 1966 when I first began teaching at Stanford, a student, Peter Hart, walked into my office with an interesting problem. He said that Charles Cole and he were using a pattern classification scheme which, for lack of a better word, they described as the nearest neighbor procedure. This scheme assigned to an as yet unclassified observation the classification of the nearest neighbor. Were there any good theoretical properties of this procedure? (Cover, 1982)

It would take some time for the term “nearest neighbor” to enter common parlance. However, the nearest neighbor procedure spread quickly across areas in computer science. Not long after Cover and Hart's 1967 paper, Donald Knuth's third volume of *The Art of Computer Programming* introduced nearest neighbor search as the *post office problem* (Knuth, 1973), paving the beginnings of computational geometry. In various coding theory contexts, *maximum likelihood decoding* turns out to mean nearest neighbor classification (Hill, 1986). Fast forwarding to present time, with the explosion in the availability of data in virtually all disciplines, architecting database systems that scale to this volume

[2] Not only was the k-nearest neighbor method named as one of the top 10 algorithms in data mining (Wu *et al.*, 2008), three of the other top 10 methods (AdaBoost, C4.5, and CART) have nearest neighbor interpretations.

of data and that can efficiently find nearest neighbors has become a fundamental problem (Papadopoulos and Manolopoulos, 2005). Understanding when, why, and how well nearest neighbor prediction works now demands accounting for computational costs.

1.1 Explaining the Popularity of Nearest Neighbor Methods

That nearest neighbor methods remain popular in practice largely has to do with their empirical success over the years. However, this explanation is perhaps overly simplistic. We highlight four aspects of nearest neighbor methods that we believe have been crucial to their continued popularity. First, the flexibility in choosing what "near" means in nearest neighbor prediction allows us to readily handle *ad-hoc* distances, or to take advantage of existing representation and distance learning machinery such as deep neural networks or decision-tree-based ensemble learning approaches. Second, the computational efficiency of numerous approximate nearest neighbor search procedures enables nearest neighbor prediction to scale to massive high-dimensional datasets common in modern applications. Third, nearest neighbor methods are nonparametric, making few modeling assumptions on data and instead letting the data more directly drive predictions. Lastly, nearest neighbor methods are interpretable: they provide evidence for their predictions by exhibiting the nearest neighbors found.

Flexibility in defining similarity. Specifying what "near" means for a nearest neighbor method amounts to choosing a "feature space" in which data are represented (as "feature vectors"), and a distance function to use within the feature space. For example, a common choice for the feature space and distance function are Euclidean space and Euclidean distance, respectively. Of course, far more elaborate choices are possible and, in practice, often these are chosen in an *ad-hoc* manner depending on the application. For example, when working with time series, the distance function could involve a highly nonlinear time warp (to try to align two time series as well as possible before computing a simpler distance like Euclidean distance). In choosing a "good" feature space (*i.e.*, a good way to *represent* data), features could be manually "hand-engineered" depending on the data modality (*e.g.*, text, images, video, audio) or learned, for example, using deep neural networks (*e.g.*, Goodfellow *et al.* 2016, Chapter 15). Meanwhile, sensor fusion is readily possible as features extracted from multiple sensors (*e.g.*, different data modalities) can be concatenated to form a large feature vector. Separately, the distance function itself can be learned, for example using Mahalanobis distance learning methods (Kulis, 2013) or Siamese networks (Bromley *et al.*, 1994; Chopra *et al.*, 2005). In fact, decision trees and their use in ensemble methods such as random forests, AdaBoost, and gradient boosting can be shown to be weighted nearest neighbor methods that learn a distance function (we discuss this relationship toward the end of the monograph in Section 7.1, building on a previous observation made by Lin and Jeon 2006). Thus, nearest neighbor methods actually mesh well with a number of existing representation and distance learning results.

Computational efficiency. Perhaps the aspect of nearest neighbor methods that has contributed the most to their popularity is their computational efficiency, which has enabled these methods to scale to massive datasets ("big data"). Depending on the feature space and distance function chosen or learned by the practitioner, different fast approximate nearest neighbor search algorithms are available. These search algorithms, both for general high-dimensional feature spaces (*e.g.*, Gionis *et al.* 1999; Datar *et al.* 2004; Bawa *et al.* 2005; Andoni and Indyk 2008; Ailon and Chazelle 2009; Muja and Lowe 2009; Boytsov and Naidan 2013; Dasgupta and Sinha 2015; Mathy *et al.* 2015; Andoni *et al.* 2017) and specialized to image patches (*e.g.*, Barnes *et al.* 2009; Ta *et al.* 2014), can rapidly determine which data points are close to each other while parallelizing across search queries. These methods often use locality-sensitive hashing (Indyk and Motwani, 1998), which comes with a theoretical guarantee on approximation accuracy, or randomized trees (*e.g.*, Bawa *et al.* 2005; Muja and Lowe 2009; Dasgupta and Sinha 2015; Mathy *et al.* 2015), which quickly prune search spaces when the trees are sufficiently balanced. These randomized trees can even be efficiently constructed for streaming data using an arbitrary distance function (Mathy *et al.*, 2015).

Nonparametric. Roughly speaking, nearest neighbor methods being nonparametric means that they make very few assumptions on the underlying model for the data. This is a particularly attractive property since in a growing number of modern applications such as social networks, recommendation systems, healthcare decision support, and online education, we wish to analyze big data that we do not *a priori* know the structure of. A nonparametric approach sidesteps the question of explicitly positing or learning the structure underlying the data. When we posit intricate structure for data, the structure may stray from reality or otherwise not account for the full palette of possibilities in what the data look like. When we learn structure, the computational overhead and amount of data needed may dwarf what is sufficient for tackling the prediction task at hand. Instead of positing or learning structure, nonparametric methods let the data more directly drive predictions. However, being nonparametric doesn't mean that nearest neighbor methods have no parameters. We still have to choose a feature space and distance, and a poor choice of these could make prediction impossible.

Interpretability. Nearest neighbor methods naturally provide evidence for their decisions by exhibiting the nearest neighbors found in the data. A practitioner can use the nearest neighbors found to diagnose whether the feature space and distance function chosen are adequate for the application of interest. For example, if on validation data, a nearest neighbor method is making incorrect predictions, we can look at the nearest neighbors of each validation data point to see why they tend to have incorrect labels. This often gives clues to the practitioner as to how to choose a better feature space or distance function. Alternatively, if the nearest neighbor method is producing accurate predictions, the nearest neighbors found tell us which training data points are driving the prediction for any particular validation or test point. This interpretability is vital in applications such as healthcare that demand a high burden of proof before letting software influence potentially costly decisions that affect people's well-being.

1.2 Nearest Neighbor Methods in Theory

Although nearest neighbor methods for prediction have remained popular, only recently has a thorough theory been developed to characterize the error rate of these methods in fairly general settings. Roughly a millennium after the appearance of nearest neighbor classification in Alhazen's *Book of Optics*, Chaudhuri and Dasgupta (2014) established arguably the most general performance guarantee to date, stating how many training data and how to choose the number of nearest neighbors to achieve a user-specified error tolerance, when the data reside in a metric space.[3] This flavor of result is "nonasymptotic" in that it can be phrased in a way that gives the probability of misclassification for *any* training data set size; we do not need an asymptotic assumption that the amount of training data goes to infinity. Chaudhuri and Dasgupta's result subsumes or matches classical results by Fix and Hodges (1951), Devroye *et al.* (1994), Cérou and Guyader (2006), and Audibert and Tsybakov (2007), while providing a perhaps more intuitive explanation for when nearest neighbor classification works, accounting for the metric used and the distribution from which the data are sampled. Moreover, we show that their analysis can be translated to the regression setting, yielding theoretical guarantees that nearly match the best of existing regression results.

However, while the general theory for both nearest neighbor classification and regression has largely been fleshed out, a major criticism is that they do not give "user-friendly" error bounds that can readily be computed from available training data (Kontorovich and Weiss, 2015). For example, Chaudhuri and Dasgupta's result for nearest neighbor classification depends on the probability of landing near the true decision boundary. Meanwhile, nearest neighbor regression results depend on smoothness of the function being estimated, usually in terms of Lipschitz or more generally Hölder continuity parameters. In practice,

[3] Within the same year a few months after Chaudhuri and Dasgupta's paper appeared on arXiv, Gadat *et al.* posted on arXiv the most general theory to date for nearest neighbor classification in the more restricted setting of finite dimensional spaces, which was finally published two years later in *Annals of Statistics* (Gadat *et al.*, 2016).

these quantities are typically difficult to estimate for a real dataset. Unfortunately, this also makes the theory hard to use by practitioners, who often are interested in understanding how many training data they should acquire to achieve a certain level of accuracy, preferably in terms of interpretable application-specific structure rather than, for instance, Hölder continuity parameters (*e.g.*, in healthcare, each training data point could correspond to a patient, and the cost of conducting a study may scale with the number of patients; being able to relate how many patients should be in the study in terms of specific disease or treatment quantities that clinicians can estimate would be beneficial).

Rather than providing results in as general a setting as possible, a recent trilogy of papers instead shows how clustering structure that is present in data enables enables nearest neighbor prediction to succeed at time series forecasting, recommending products to people, and finding human organs in medical images (Chen *et al.*, 2013; Bresler *et al.*, 2014; Chen *et al.*, 2015). These papers establish nonasymptotic theoretical guarantees that trade off between the training data size and the prediction accuracy as a function of the number of clusters and the amount of noise present. The theory here depends on the clusters being separated enough so that noise is unlikely to cause too many points to appear to come from a wrong cluster. Prediction succeeds when, for a test point, its nearest neighbors found in the training data are predominantly from the same cluster as the test point. That these theoretical guarantees are about clustering is appealing because clusters can often be estimated from data and interpreted by practitioners.

1.3 The Scope of This Monograph

This monograph aims to explain the success of nearest neighbor methods in prediction, covering both theory and practice. Our exposition intentionally strives to be as accessible as possible to theoreticians and practitioners alike. As the number of prediction methods that rely on nearest neighbor analysis and the amount of literature studying these methods are both enormous, our coverage is carefully curated and inexhaustive.

On the theoretical side, our goal is to provide some of the most general *nonasymptotic* results and give a flavor of the proof techniques involved. All key theoretical guarantees we cover are stated in the form of how many training data and what algorithm parameters ensure that a nearest neighbor prediction method achieves a user-specified error tolerance.

On the more practical side, we cover some examples of how nearest neighbor methods are used as part of a larger prediction system (recommending products to people in the problem of *online collaborative filtering*, and delineating where a human organ is in medical images in the problem of *patch-based image segmentation*). We also discuss a variety of approximate nearest neighbor search and related methods which have been pivotal to scaling nearest neighbor prediction to massive, even ever-growing datasets.

Our coverage is as follows, transitioning from theory to practice as we progress through the monograph:

- **Background (Chapter 2).** We anchor notation and terminology used throughout the monograph. Specifically we define the basic prediction tasks of classification and regression, and then present the three basic algorithms of k-nearest neighbor, fixed-radius near neighbor, and kernel regression. These regression methods can in turn be translated into classification methods.

- **Regression (Chapter 3).** We present theoretical guarantees for k-nearest neighbor, fixed-radius near neighbor, and kernel regression where the data reside in a metric space. The proofs borrow heavily from the work by Chaudhuri and Dasgupta (2014) with some influence from the work by Gadat *et al.* (2016). These authors actually focus on classification, but proof ideas translate over to the regression setting.

- **Classification (Chapter 4).** We show how the theoretical guarantees for regression can readily be converted to ones for classification. However, it turns out that we can obtain classification guarantees using weaker conditions. We explain how Chaudhuri and Dasgupta (2014) achieve this for k-nearest neighbor classification and how the basic idea readily generalizes to fixed-radius near neighbor and kernel classification.

- **Prediction Guarantees in Three Contemporary Applications (Chapter 5).** We present theoretical guarantees for nearest neighbor prediction in time series forecasting (Chen *et al.*, 2013), online collaborative filtering (Bresler *et al.*, 2014), and patch-based image segmentation (Chen *et al.*, 2015). Despite these applications seeming disparate and unrelated, the theoretical guarantees for them turn out to be quite similar. In all three, clustering structure enables successful prediction. We remark that the independence assumptions on training data and where clustering structure appears are both application-specific.

- **Computation (Chapter 6).** We provide an overview of efficient data structures for exact and approximate nearest neighbor search that are used in practice. We focus on motifs these methods share rather than expounding on theoretical guarantees, which many of these methods lack. Our starting point is the classical *k-d tree* data structure for exact nearest neighbor search (Bentley, 1979), which works extremely well for low-dimensional data but suffers from the "curse of dimensionality" due to an exponential dependence on dimension when executing a search query. To handle exact high-dimensional nearest neighbor search, more recent approaches such as the *cover tree* data structure exploit the idea that high-dimensional data often have low-dimensional structure (Beygelzimer *et al.*, 2006). As such approaches can still be computationally expensive in practice, we turn toward approximate nearest neighbor search. We describe *locality-sensitive hashing* (LSH) (Indyk and Motwani, 1998), which forms the foundation of many approximate nearest neighbor search methods that come with theoretical guarantees. We also discuss empirically successful approaches with partial or no theoretical guarantees: *random projection* or *partition trees* inspired by k-d trees, and the recently proposed *boundary forest*.

- **Adaptive Nearest Neighbors and Far Away Neighbors (Chapter 7).** We end with remarks on distance learning with a focus on decision trees and various ensemble methods that turn out to be nearest neighbor methods, and then turn toward a new class of nearest neighbor methods that in some sense can take advantage of far away neighbors.

For readers seeking a more "theory-forward" exposition albeit without coverage of Chaudhuri and Dasgupta's classification and related regression results, there are recent books by Devroye *et al.* (2013) (on classification) and Biau and Devroye (2015) (on nearest neighbor methods with sparse discussion on kernel regression and classification), and earlier books by Györfi *et al.* (2002) (on nonparametric regression) and Tsybakov (2009) (on nonparametric estimation). Unlike the other books mentioned, Tsybakov's regression coverage emphasizes fixed design (corresponding to the training feature vectors having a deterministic structure, such as being evenly spaced in a feature space), which is beyond the scope of this monograph. As for theory on nearest neighbor search algorithms, there is a survey by Clarkson (2006) that goes into substantially more detail than our overview in Chapter 6. However, this survey does not cover a number of very recent advances in approximate nearest neighbor search that we discuss.

References

Abraham, C., G. Biau, and B. Cadre (2006). “On the kernel rule for function classification”. *Annals of the Institute of Statistical Mathematics*. 58(3): 619–633.

Ailon, N. and B. Chazelle (2009). “The fast Johnson–Lindenstrauss transform and approximate nearest neighbors”. *SIAM Journal on Computing*. 39(1): 302–322.

Aldous, D. J. (1981). “Representations for partially exchangeable arrays of random variables”. *Journal of Multivariate Analysis*. 11(4): 581–598.

Anandkumar, A., R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky (2014). “Tensor decompositions for learning latent variable models”. *Journal of Machine Learning Research*. 15: 2773–2832.

Anava, O. and K. Y. Levy (2016). “k^*-nearest neighbors: from global to local”. In: *Advances in Neural Information Processing Systems*. 4916–4924.

Andoni, A. and P. Indyk (2008). “Near-optimal hashing algorithms for near neighbor problem in high dimension”. *Communications of the ACM*. 51(1): 117–122.

Andoni, A., P. Indyk, T. Laarhoven, I. Razenshteyn, and L. Schmidt (2015). “Practical and optimal LSH for angular distance”. In: *Advances in Neural Information Processing Systems*. 1225–1233.

Andoni, A., T. Laarhoven, I. Razenshteyn, and E. Waingarten (2017). "Optimal hashing-based time-space trade-offs for approximate near neighbors". In: *Symposium on Discrete Algorithms*. SIAM. 47–66.

Andoni, A. and I. Razenshteyn (2015a). "Optimal data-dependent hashing for approximate near neighbors". In: *Symposium on Theory of Computing*. ACM. 793–801.

Andoni, A. and I. Razenshteyn (2015b). "Tight lower bounds for data-dependent locality-sensitive hashing". arXiv: 1507.04299 [cs.DS].

Audibert, J.-Y. and A. B. Tsybakov (2007). "Fast learning rates for plug-in classifiers". *The Annals of Statistics*. 35(2): 608–633.

Auer, P., N. Cesa-Bianchi, and P. Fischer (2002). "Finite-time analysis of the multiarmed bandit problem". *Machine Learning*. 47(2-3): 235–256.

Austin, T. (2012). "Exchangeable random arrays". *Notes for IAS workshop*.

Bagnall, A., L. Davis, J. Hills, and J. Lines (2012). "Transformation based ensembles for time series classification". In: *International Conference on Data Mining*. SIAM. 307–318.

Bai, W., W. Shi, D. P. O'Regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert (2013). "A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images". *IEEE Transactions on Medical Imaging*. 32(7): 1302–1315.

Barman, K. and O. Dabeer (2012). "Analysis of a collaborative filter based on popularity amongst neighbors". *IEEE Transactions on Information Theory*. 58(12): 7110–7134.

Barnes, C., E. Shechtman, A. Finkelstein, and D. B. Goldman (2009). "PatchMatch: a randomized correspondence algorithm for structural image editing". *ACM Transactions on Graphics*. 28(3): 24.

Batista, G. E. A. P. A., X. Wang, and E. J. Keogh (2011). "A complexity-invariant distance measure for time series". In: *International Conference on Data Mining*. SIAM. 699–710.

Bawa, M., T. Condie, and P. Ganesan (2005). "LSH forest: self-tuning indexes for similarity search". In: *International Conference on the World Wide Web*. ACM. 651–660.

Belkin, M. and K. Sinha (2010). “Polynomial learning of distribution families”. In: *Foundations of Computer Science*. IEEE. 103–112.

Bennett, J. and S. Lanning (2007). “The Netflix Prize”. In: *KDD Cup and Workshop*. 35.

Bentley, J. L. (1975). “A survey of techniques for fixed-radius near neighbor searching”. *Technical report, Stanford Linear Accelerator Center*.

Bentley, J. L. (1979). “Multidimensional binary search trees in database applications”. *IEEE Transactions on Software Engineering*. (4): 333–340.

Bertsimas, D. and J. Dunn (2017). “Optimal classification trees”. *Machine Learning*. 106(7): 1039–1082.

Beygelzimer, A., S. Kakade, and J. Langford (2006). “Cover trees for nearest neighbor”. In: *International Conference on Machine Learning*. ACM. 97–104.

Biau, G. and L. Devroye (2015). *Lectures on the Nearest Neighbor Method*. Springer.

Bigot, J. and S. Gadat (2010). “A deconvolution approach to estimation of a common shape in a shifted curves model”. *The Annals of Statistics*. 38(4): 2422–2464.

Borgs, C., J. Chayes, C. E. Lee, and D. Shah (2017). “Thy friend is my friend: iterative collaborative filtering for matrix estimation”. In: *Advances in Neural Information Processing Systems*. 4718–4729.

Boyd, S., N. Parikh, E. Chu, B. Peleato, and J. Eckstein (2011). “Distributed optimization and statistical learning via the alternating direction method of multipliers”. *Foundations and Trends® in Machine Learning*. 3(1): 1–122.

Boytsov, L. and B. Naidan (2013). “Engineering efficient and effective non-metric space library”. In: *International Conference on Similarity Search and Applications*. Springer. 280–293.

Breiman, L. (2001). “Random forests”. *Machine Learning*. 45(1): 5–32.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and Regression Trees*. Chapman and Hall/CRC.

Bresler, G., G. H. Chen, and D. Shah (2014). “A latent source model for online collaborative filtering”. In: *Advances in Neural Information Processing Systems*. 3347–3355.

Bresler, G. and M. Karzand (2017). "Regret bounds and regimes of optimality for user-user and item-item collaborative filtering". arXiv: 1711.02198 [stat.ML].

Bresler, G., D. Shah, and L. F. Voloch (2016). "Collaborative filtering with low regret". In: *SIGMETRICS Performance Evaluation Review*. Vol. 44. No. 1. ACM. 207–220.

Bromley, J., I. Guyon, Y. LeCun, E. Säckinger, and R. Shah (1994). "Signature verification using a "Siamese" time delay neural network". In: *Advances in Neural Information Processing Systems*. 737–744.

Buades, A., B. Coll, and J.-M. Morel (2005). "A non-local algorithm for image denoising". In: *Computer Vision and Pattern Recognition*. Vol. 2. IEEE. 60–65.

Bubeck, S. and N. Cesa-Bianchi (2012). "Regret analysis of stochastic and nonstochastic multi-armed bandit problems". *Foundations and Trends® in Machine Learning*. 5(1): 1–122.

Bui, L., R. Johari, and S. Mannor (2012). "Clustered bandits". arXiv: 1206.4169 [cs.LG].

Cai, J.-F., E. J. Candès, and Z. Shen (2010). "A singular value thresholding algorithm for matrix completion". *SIAM Journal on Optimization*. 20(4): 1956–1982.

Candes, E. J. and Y. Plan (2010). "Matrix completion with noise". *Proceedings of the IEEE*. 98(6): 925–936.

Candès, E. J. and B. Recht (2009). "Exact matrix completion via convex optimization". *Foundations of Computational Mathematics*. 9(6): 717–772.

Cérou, F. and A. Guyader (2006). "Nearest neighbor classification in infinite dimension". *ESAIM: Probability and Statistics*. 10: 340–355.

Chaovalitwongse, W. A., Y.-J. Fan, and R. C. Sachdeo (2007). "On the time series K-nearest neighbor classification of abnormal brain activity". *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*. 37(6): 1005–1016.

Chatterjee, S. (2015). "Matrix estimation by universal singular value thresholding". *The Annals of Statistics*. 43(1): 177–214.

Chaudhuri, K. and S. Dasgupta (2014). “Rates of convergence for nearest neighbor classification”. In: *Advances in Neural Information Processing Systems*. 3437–3445. The different versions of this paper have different numberings for theorems, equations, etc. The numbering we use is from arXiv: 1407.0067v2 [cs.LG].

Chaudhuri, K. and S. Rao (2008). “Learning mixtures of product distributions using correlations and independence”. In: *Conference on Learning Theory*. Vol. 4. No. 1. 9–20.

Chen, G. H. (2015). “Latent source models for nonparametric inference”. *PhD thesis*. Massachusetts Institute of Technology.

Chen, G. H., S. Nikolov, and D. Shah (2013). “A latent source model for nonparametric time series classification”. In: *Advances in Neural Information Processing Systems*. 1088–1096.

Chen, G. H., D. Shah, and P. Golland (2015). “A latent source model for patch-based image segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 140–148.

Chopra, S., R. Hadsell, and Y. LeCun (2005). “Learning a similarity metric discriminatively, with application to face verification”. In: *Computer Vision and Pattern Recognition*. Vol. 1. IEEE. 539–546.

Clarkson, K. L. (2006). “Nearest-neighbor searching and metric space dimensions”. *Nearest-neighbor methods for learning and vision: theory and practice*: 15–59.

Coupé, P., J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins (2011). “Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation”. *NeuroImage*. 54(2): 940–954.

Cover, T. M. (1982). “This week’s citation classic”. *Current Contents*. 13: 20.

Cover, T. M. and P. E. Hart (1967). “Nearest neighbor pattern classification”. *IEEE Transactions on Information Theory*. 13(1): 21–27.

Criminisi, A., P. Pérez, and K. Toyama (2004). “Region filling and object removal by exemplar-based image inpainting”. *IEEE Transactions on Image Processing*. 13(9): 1200–1212.

Dasgupta, S. and Y. Freund (2008). “Random projection trees and low dimensional manifolds”. In: *Symposium on Theory of Computing*. ACM. 537–546.

Dasgupta, S. and L. Schulman (2007). “A probabilistic analysis of EM for mixtures of separated, spherical Gaussians”. *Journal of Machine Learning Research*. 8: 203–226.

Dasgupta, S. and K. Sinha (2015). “Randomized partition trees for nearest neighbor search”. *Algorithmica*. 72(1): 237–263.

Dasgupta, S., C. F. Stevens, and S. Navlakha (2017). “A neural algorithm for a fundamental computing problem”. *Science*. 358(6364): 793–796.

Datar, M., N. Immorlica, P. Indyk, and V. S. Mirrokni (2004). “Locality-sensitive hashing scheme based on p-stable distributions”. In: *Symposium on Computational Geometry*. ACM. 253–262.

De Finetti, B. (1937). “La prévision: ses lois logiques, ses sources subjectives”. In: *Annales de l’institut Henri Poincaré*. Vol. 7. No. 1. 1–68.

Depa, M., M. R. Sabuncu, G. Holmvang, R. Nezafat, E. J. Schmidt, and P. Golland (2010). “Robust atlas-based segmentation of highly variable anatomy: left atrium segmentation”. In: *MICCAI Workshop on Statistical Atlases and Computational Models of the Heart*. Springer. 85–94.

Deshpande, Y. and A. Montanari (2013). “Linear bandits in high dimension and recommendation systems”. arXiv: 1301.1722 [cs.LG].

Devlin, J., S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick (2015). “Exploring nearest neighbor approaches for image captioning”. arXiv: 1505.04467 [cs.CV].

Devroye, L. (1981). “On the almost everywhere convergence of nonparametric regression function estimates”. *The Annals of Statistics*. 9(6): 1310–1319.

Devroye, L., L. Györfi, A. Krzyżak, and G. Lugosi (1994). “On the strong universal consistency of nearest neighbor regression function estimates”. *The Annals of Statistics*. 22(3): 1371–1385.

Devroye, L., L. Györfi, and G. Lugosi (2013). *A Probabilistic Theory of Pattern Recognition*. Vol. 31. Springer.

Ding, H., G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh (2008). "Querying and mining of time series data: experimental comparison of representations and distance measures". *Proceedings of the VLDB Endowment*. 1(2): 1542–1552.

Erdős, P. and A. Rényi (1959). "On random graphs". *Publicationes Mathematicae*. 6: 290–297.

Feldman, D., M. Faulkner, and A. Krause (2011). "Scalable training of mixture models via coresets". In: *Advances in Neural Information Processing Systems*. 2142–2150.

Fix, E. and J. L. Hodges Jr. (1951). "Discriminatory analysis, nonparametric discrimination: consistency properties". *Technical report, USAF School of Aviation Medicine*.

Fix, E. and J. L. Hodges Jr. (1952). "Discriminatory analysis, nonparametric discrimination: small sample performance". *Technical report, USAF School of Aviation Medicine*.

Freeman, W. T., E. C. Pasztor, and O. T. Carmichael (2000). "Learning low-level vision". *International Journal of Computer Vision*. 40(1): 25–47.

Freund, Y. and R. E. Schapire (1997). "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences*. 55(1): 119–139.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition (2nd ed.)* Academic Press.

Gadat, S., T. Klein, and C. Marteau (2016). "Classification in general finite dimensional spaces with the k-nearest neighbor uule". *The Annals of Statistics*. 44(3): 982–1009. Preliminary version appeared in 2014 (arXiv: 1411.0894 [math.ST]).

Gentile, C., S. Li, and G. Zappella (2014). "Online clustering of bandits". In: *International Conference on Machine Learning*. 757–765.

Gionis, A., P. Indyk, and R. Motwani (1999). "Similarity search in high dimensions via hashing". In: *International Conference on Very Large Data Bases*. Vol. 99. 518–529.

Goldenshluger, A. and O. Lepski (2008). "Universal pointwise selection rule in multivariate function estimation". *Bernoulli*. 14(4): 1150–1190.

Goldenshluger, A. and O. Lepski (2009). "Structural adaptation via $\mathbb{L}_p$-norm oracle inequalities". *Probability Theory and Related Fields*. 143(1-2): 41–71.

Goldenshluger, A. and O. Lepski (2011). "Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality". *The Annals of Statistics*. 39(3): 1608–1632.

Goldenshluger, A. and O. Lepski (2013). "General selection rule from a family of linear estimators". *Theory of Probability & Its Applications*. 57(2): 209–226.

Goldenshluger, A. and O. Lepski (2014). "On adaptive minimax density estimation on R^d". *Probability Theory and Related Fields*. 159(3-4): 479–543.

Goodfellow, I., Y. Bengio, and A. Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Grosse, R. B., R. Salakhutdinov, W. T. Freeman, and J. B. Tenenbaum (2012). "Exploiting compositionality to explore a large space of model structures". In: *Uncertainty in Artificial Intelligence*. AUAI Press. 306–315.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer.

Hall, P., B. U. Park, and R. J. Samworth (2008). "Choice of neighbor order in nearest-neighbor classification". *The Annals of Statistics*. 36(5): 2135–2152.

Hanbury, A., H. Müller, G. Langs, M. A. Weber, B. H. Menze, and T. S. Fernandez (2012). "Bringing the algorithms to the data: cloud–based benchmarking for medical image analysis". In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer. 24–29.

Har-Peled, S., P. Indyk, and R. Motwani (2012). "Approximate nearest neighbor: towards removing the curse of dimensionality." *Theory of Computing*. 8(1): 321–350.

Harper, F. M. and J. A. Konstan (2016). "The MovieLens datasets: history and context". *ACM Transactions on Interactive Intelligent Systems*. 5(4): 19.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)* Springer.

Hill, R. (1986). *A First Course in Coding Theory*. Oxford University Press.

Hoare, C. A. R. (1961). “Algorithm 65: find”. *Communications of the ACM*. 4(7): 321–322.

Hoorfar, A. and M. Hassani (2008). “Inequalities on the Lambert W function and hyperpower function”. *Journal of Inequalities in Pure and Applied Mathematics*. 9(2): 5–9.

Hoover, D. (1981). “Row-column exchangeability and a generalized model for probability”. In: *Exchangeability in Probability and Statistics*. 281–291.

Hsu, D. and S. M. Kakade (2013). “Learning mixtures of spherical Gaussians: moment methods and spectral decompositions”. In: *Innovations in Theoretical Computer Science*. ACM. 11–20.

Iglesias, J. E., E. Konukoglu, D. Zikic, B. Glocker, K. V. Leemput, and B. Fischl (2013). “Is synthesizing MRI contrast useful for intermodality analysis?” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 631–638.

Indyk, P. and R. Motwani (1998). “Approximate nearest neighbors: towards removing the curse of dimensionality”. In: *Symposium on Theory of Computing*. ACM. 604–613.

Johnson, W. B. and J. Lindenstrauss (1984). “Extensions of Lipschitz mappings into a Hilbert space”. *Contemporary Mathematics*. 26(189-206): 1.

Kaján, L., A. Kertész-Farkas, D. Franklin, N. Ivanova, A. Kocsor, and S. Pongor (2006). “Application of a simple likelihood ratio approximant to protein sequence classification”. *Bioinformatics*. 22(23): 2865–2869.

Kamath, G. C. (2015). “Bounds on the expectation of the maximum of samples from a Gaussian”. URL: http://www.gautamkamath.com/writings/gaussian_max.pdf.

Karger, D. R. and M. Ruhl (2002). “Finding nearest neighbors in growth-restricted metrics”. In: *Symposium on Theory of Computing*. ACM. 741–750.

Keshavan, R. H., A. Montanari, and S. Oh (2010a). "Matrix completion from a few entries". *IEEE Transactions on Information Theory*. 56(6): 2980–2998.

Keshavan, R. H., A. Montanari, and S. Oh (2010b). "Matrix completion from noisy entries". *Journal of Machine Learning Research*. 11: 2057–2078.

Al-Khalili, J. (2015). "In retrospect: Book of Optics". *Nature*. (7538): 164.

Kleinberg, J. M. (1997). "Two algorithms for nearest-neighbor search in high dimensions". In: *Symposium on Theory of Computing*. ACM. 599–608.

Kleinberg, R., A. Niculescu-Mizil, and Y. Sharma (2010). "Regret bounds for sleeping experts and bandits". *Machine Learning*. 80(2-3): 245–272.

Kneip, A. and T. Gasser (1992). "Statistical tools to analyze data representing a sample of curves". *The Annals of Statistics*. 20(3): 1266–1305.

Knuth, D. E. (1973). *The Art of Computer Programming - Vol III: Sorting and Searching*. Addison Wesley.

Kohler, M., A. Krzyżak, and H. Walk (2006). "Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data". *Journal of Multivariate Analysis*. 97(2): 311–323.

Kontorovich, A. and R. Weiss (2015). "A Bayes consistent 1-NN classifier". In: *Artificial Intelligence and Statistics*. 480–488.

Konukoglu, E., A. J. W. van der Kouwe, M. R. Sabuncu, and B. Fischl (2013). "Example-based restoration of high-resolution magnetic resonance image acquisitions". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 131–138.

Koren, Y. (2009). "The BellKor Solution to the Netflix Grand Prize". URL: http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf.

Kpotufe, S. (2011). "k-NN regression adapts to local intrinsic dimension". In: *Advances in Neural Information Processing Systems*. 729–737.

Kpotufe, S. and V. K. Garg (2013). “Adaptivity to local smoothness and dimension in kernel regression”. In: *Advances in Neural Information Processing Systems*. 3075–3083.

Krzyżak, A. (1986). “The rates of convergence on kernel regression estimates and classification rules”. *IEEE Transactions on Information Theory*. 32(5): 668–679.

Krzyżak, A. and M. Pawlak (1987). “The pointwise rate of convergence of the kernel regression estimate”. *Journal of Statistical Planning and Inference*. 16: 159–166.

Kulis, B. (2013). “Metric learning: a survey”. *Foundations and Trends® in Machine Learning*. 5(4): 287–364.

Lee, C. E., Y. Li, D. Shah, and D. Song (2016). “Blind regression: nonparametric regression for latent variable models via collaborative filtering”. In: *Advances in Neural Information Processing Systems*. 2155–2163.

Lee, T. and A. Shraibman (2013). “Matrix completion from any given set of observations”. In: *Advances in Neural Information Processing Systems*. 1781–1787.

Lee, Y.-H., C.-P. Wei, T.-H. Cheng, and C.-T. Yang (2012). “Nearest-neighbor-based approach to time-series classification”. *Decision Support Systems*. 53(1): 207–217.

Lepski, O. V. and B. Y. Levit (1998). “Adaptive minimax estimation of infinitely differentiable functions”. *Mathematical Methods of Statistics*. 7(2): 123–156.

Lepski, O. V., E. Mammen, and V. G. Spokoiny (1997). “Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors”. *The Annals of Statistics*. 25(3): 929–947.

Lepski, O. V. and V. G. Spokoiny (1997). “Optimal pointwise adaptive methods in nonparametric estimation”. *The Annals of Statistics*. 25(6): 2512–2546.

Levinthal, C. (1966). “Molecular model-building by computer”. *Scientific American*. 214(6): 42.

Lin, Y. and Y. Jeon (2006). “Random forests and adaptive nearest neighbors”. *Journal of the American Statistical Association*. 101(474): 578–590.

Linden, G., B. Smith, and J. York (2003). “Amazon.com recommendations: item-to-item collaborative filtering”. *IEEE Internet Computing*. 7(1): 76–80.

Liu, T., A. W. Moore, K. Yang, and A. G. Gray (2005). “An investigation of practical approximate nearest neighbor algorithms”. In: *Advances in Neural Information Processing Systems*. 825–832.

Lovász, L. (2012). *Large Networks and Graph Limits*. Vol. 60. American Mathematical Society.

Lovász, L. and B. Szegedy (2006). “Limits of dense graph sequences”. *Journal of Combinatorial Theory, Series B*. 96(6): 933–957.

Mammen, E. and A. B. Tsybakov (1999). “Smooth discrimination analysis”. *The Annals of Statistics*. 27(6): 1808–1829.

Mathy, C., N. Derbinsky, J. Bento, J. Rosenthal, and J. Yedidia (2015). “The boundary forest algorithm for online supervised and unsupervised learning”. In: *AAAI Conference on Artificial Intelligence*. AAAI Press. 2864–2870.

Mazumder, R., T. Hastie, and R. Tibshirani (2010). “Spectral regularization algorithms for learning large incomplete matrices”. *Journal of Machine Learning Research*. 11: 2287–2322.

Moitra, A. and G. Valiant (2010). “Settling the polynomial learnability of mixtures of Gaussians”. In: *Foundations of Computer Science*. IEEE. 93–102.

Motwani, R., A. Naor, and R. Panigrahy (2007). “Lower bounds on locality sensitive hashing”. *SIAM Journal on Discrete Mathematics*. 21(4): 930–935.

Muja, M. and D. G. Lowe (2009). “Fast approximate nearest neighbors with automatic algorithm configuration”. In: *International Conference on Computer Vision Theory and Application*. Vol. 2. No. 331-340. 2.

Muja, M. and D. G. Lowe (2014). “Scalable nearest neighbor algorithms for high dimensional data”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 36(11): 2227–2240.

Nadaraya, È. A. (1964). “On estimating regression”. *Theory of Probability & Its Applications*. 9(1): 141–142.

Nanopoulos, A., R. Alcock, and Y. Manolopoulos (2001). “Feature-based classification of time-series data”. *International Journal of Computer Research*. 10(3): 49–61.

Nikolov, S. and D. Shah (2012). “A nonparametric method for early detection of trending topics”. Interdisciplinary Workshop on Information and Decision in Social Networks.

O’Donnell, R., Y. Wu, and Y. Zhou (2014). “Optimal lower bounds for locality-sensitive hashing (except when q is tiny)”. *ACM Transactions on Computation Theory*. 6(1): 5.

Papadopoulos, A. N. and Y. Manolopoulos (2005). *Nearest Neighbor Search: A Database Perspective*. Springer.

Pelillo, M. (2014). “Alhazen and the nearest neighbor rule”. *Pattern Recognition Letters*. 38: 34–37.

Piotte, M. and M. Chabbert (2009). “The Pragmatic Theory solution to the Netflix Grand Prize”. URL: http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

Recht, B. (2011). “A simpler approach to matrix completion”. *Journal of Machine Learning Research*. 12: 3413–3430.

Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl (1994). “GroupLens: an open architecture for collaborative filtering of netnews”. In: *Conference on Computer Supported Cooperative Work*. ACM. 175–186.

Rodríguez, J. J. and C. J. Alonso (2004). “Interval and dynamic time warping-based decision trees”. In: *Symposium on Applied Computing*. ACM. 548–552.

Rousseau, F., P. A. Habas, and C. Studholme (2011). “A supervised patch-based approach for human brain labeling”. *IEEE Transactions on Medical Imaging*. 30(10): 1852–1862.

Rousseau, F. and C. Studholme (2013). “A supervised patch-based image reconstruction technique: Application to brain MRI super-resolution”. In: *International Symposium on Biomedical Imaging*. IEEE. 346–349.

Sabuncu, M. R., B. T. T. Yeo, K. V. Leemput, B. Fischl, and P. Golland (2010). “A generative model for image segmentation based on label fusion”. *IEEE Transactions on Medical Imaging*. 29(10): 1714–1729.

Sakoe, H. and S. Chiba (1978). “Dynamic programming algorithm optimization for spoken word recognition”. *IEEE Transactions on Acoustics, Speech, and Signal Processing*. 26(1): 43–49.

Sánchez, J. S., R. Barandela, A. Marqués, R. Alejo, and J. Badenas (2003). “Analysis of new techniques to obtain quality training sets”. *Pattern Recognition Letters*. 24(7): 1015–1022.

Silver, N. (2003). “Introducing PECOTA”. *Baseball Prospectus 2003*.

Silver, N. (2008). “Frequently Asked Questions”. *FiveThirtyEight.com*.

Smith, A. M. (2001). “Alhacen’s Theory of Visual Perception”. *Philadelphia: American Philosophical Society*. 1.

Smith, M. R. and T. Martinez (2011). “Improving classification accuracy by identifying and removing instances that should be misclassified”. In: *International Joint Conference on Neural Networks*. IEEE. 2690–2697.

Sridharan, R., A. V. Dalca, K. M. Fitzpatrick, L. Cloonan, A. Kanakis, O. Wu, K. L. Furie, J. Rosand, N. S. Rost, and P. Golland (2013). “Quantification and analysis of large multimodal clinical image studies: application to stroke”. In: *MICCAI Workshop on Multimodal Brain Image Analysis*. Springer. 18–30.

Stone, C. J. (1977). “Consistent nonparametric regression”. *The Annals of Statistics*. 5(4): 595–620.

Sundaram, N., A. Turmukhametova, N. Satish, T. Mostak, P. Indyk, S. Madden, and P. Dubey (2013). “Streaming similarity search over one billion Tweets using parallel locality-sensitive hashing”. *Proceedings of the VLDB Endowment*. 6(14): 1930–1941.

Sutskever, I., R. Salakhutdinov, and J. B. Tenenbaum (2009). “Modelling relational data using bayesian clustered tensor factorization”. In: *Advances in Neural Information Processing Systems*. 1821–1828.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Ta, V.-T., R. Giraud, D. L. Collins, and P. Coupé (2014). "Optimized PatchMatch for near real time and accurate label fusion". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 105–112.

Thompson, W. R. (1933). "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". *Biometrika*. 25: 285–294.

Töscher, A. and M. Jahrer (2009). "The BigChaos Solution to the Netflix Grand Prize". URL: http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf.

Tsybakov, A. B. (2004). "Optimal aggregation of classifiers in statistical learning". *The Annals of Statistics*. 32(1): 135–166.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

Vempala, S. and G. Wang (2004). "A spectral algorithm for learning mixture models". *Journal of Computer and System Sciences*. 68(4): 841–860.

Wachinger, C., M. Brennan, G. C. Sharp, and P. Golland (2017). "Efficient descriptor-based segmentation of parotid glands with nonlocal means". *IEEE Transactions on Biomedical Engineering*. 64(7): 1492–1502.

Wachinger, C., M. Reuter, and T. Klein (2018). "DeepNAT: deep convolutional neural network for segmenting neuroanatomy". *NeuroImage*. 170: 434–445.

Wang, K. and T. Gasser (1997). "Alignment of curves by dynamic time warping". *The Annals of Statistics*. 25(3): 1251–1276.

Wang, Y., S. Jha, and K. Chaudhuri (2017). "Analyzing the robustness of nearest neighbors to adversarial examples". arXiv: 1706.03922 [stat.ML].

Watson, G. S. (1964). "Smooth regression analysis". *Sankhyā: The Indian Journal of Statistics, Series A*. 26: 359–372.

Weinberger, K. Q. and L. K. Saul (2009). "Distance metric learning for large margin nearest neighbor classification". *Journal of Machine Learning Research*. 10: 207–244.

Wilson, D. L. (1972). "Asymptotic properties of nearest neighbor rules using edited data". *IEEE Transactions on Systems, Man, and Cybernetics*. 3: 408–421.

Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg (2008). "Top 10 algorithms in data mining". *Knowledge and Information Systems*. 14(1): 1–37.

Wu, Y. and E. Y. Chang (2004). "Distance-function design and fusion for sequence data". In: *Conference on Information and Knowledge Management*. ACM. 324–333.

Xi, X., E. J. Keogh, C. R. Shelton, L. Wei, and C. A. Ratanamahatana (2006). "Fast time series classification using numerosity reduction". In: *International Conference on Machine Learning*. ACM. 1033–1040.

Zandifar, A., V. Fonov, P. Coupé, J. Pruessner, and D. L. Collins (2017). "A comparison of accurate automatic hippocampal segmentation methods". *NeuroImage*. 155: 383–393.

Zoran, D., B. Lakshminarayanan, and C. Blundell (2017). "Learning deep nearest neighbor representations using differentiable boundary trees". arXiv: 1702.08833 [cs.LG].

Zoran, D. and Y. Weiss (2011). "From learning models of natural image patches to whole image restoration". In: *International Conference on Computer Vision*. IEEE. 479–486.

Zoran, D. and Y. Weiss (2012). "Natural images, Gaussian mixtures and dead leaves". In: *Advances in Neural Information Processing Systems*. 1736–1744.