# MCIS6273 Data Mining / Fall 2017 / Prof. Maull

## LECTURE 1: CLASS POLICIES, TOOLS AND TECHNOLOGIES

|  |  |
| --- | --- |
| Week of 8/30 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes/) |
| **Content** | class policies, class tools, introduction, what this course is about, data mining: tools, technologies and techniques |
| **Expected** **Outcomes** | • overview of course policies • overview of data mining concepts, algorithms, methodologies • installation of Anaconda and Python 3.6 • introduction to Jupyter Notebooks • creation of Github account |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php). → **ch.1** » 2014. Leskovec, Jure and Rajaraman, Anand and Ullman, Jeffrey David; [*Mining of massive datasets*](http://www.mmds.org/). → **ch.1** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.1, ch.2**  **OPTIONAL** › 2012. Downey, Allen; [*Think Python*](http://www.greenteapress.com/thinkpython/thinkpython.html). → **ch.1-ch.3** › (website) -- 2017; *The Periodic Table of Data Science*: <https://www.datacamp.com/community/blog/data-science-periodic-table#gs.TF297Gsm>. → **Familiarize yourself with the entire table.** |
| **Homework** | **DUE:** Monday, 9/6 - midnight Please see the Blackboard/[Github repo](https://github.com/kmsaumcis/mcis6273_f17_datamining/homework) for what to turn in. |

## LECTURE 2: DATA / REPRESENTATION, PREPARATION AND MANIPULATION

|  |  |
| --- | --- |
| Week of 9/6 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | introduction to core concepts in data; data types and representation of data; data formats including structured and unstructured; concepts in pre-processing data including scaling, sampling, normalizing, binning and imputing |
| **Expected** **Outcomes** | • understand data types and common formats • identify cleaning and adjusting scenarios and apply techniques appropriately • utilize and apply the appropriate Python tools (Pandas for data import and cleaning) |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php). → **ch.1** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.1, ch.2** » 2012. McKinney, Wes; [*Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*](https://github.com/wesm/pydata-book). → **ipython/Jupyter notebooks for ch.5, ch.6 and ch.7** » (website) -- 2017; *Distance computations (scipy.spatial.distance)*: <https://docs.scipy.org/doc/scipy/reference/spatial.distance.html>. → **euclidean, cosine, correlation, jaccard**  **OPTIONAL** › 2012. Downey, Allen; [*Think Python*](http://www.greenteapress.com/thinkpython/thinkpython.html). → **ch.1-ch.3** › (website) -- 2017; *Pandas Cookbook*: <https://github.com/jvns/pandas-cookbook>. → **familiarize yourself with this content of this repo** › (Michael Kennedy's Talk Python To Me podcast) -- 11-28-2016; *Episode #90: Data Wrangling with Python*: <http://talkpythontome.fm>. → **listen to the entire episode** |
| **Homework** | **DUE:** Monday, 9/18 - midnight Please see the Blackboard/[Github repo](https://github.com/kmsaumcis/mcis6273_f17_datamining/homework) for what to turn in. |

## LECTURE 3: DATA / DISTANCE, SIMILARITY, STATISTICAL CONCEPTS

|  |  |
| --- | --- |
| Week of 9/13 | [Lecture Notes](https://github.com/kmsaumcis/mcis6123_fa17/tree/master/lecture_notes) |
| **Content** | introduction to comparing data using common metrics; introductory concepts in disorder; introductory statistical concepts; intuitions over data dimensionality and common reduction techniques |
| **Expected** **Outcomes** | • identify common distance metrics and their appropriate contexts • understand similarity (and dissimilarity) in data • develop intuitions of statistical concepts in correlation, distributions and expect value • understand dimensionality reduction via PCA • utilize and apply basic statistical tools in Python (Pandas/Numpy) |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php). → **ch.7** » 2014. Leskovec, Jure and Rajaraman, Anand and Ullman, Jeffrey David; [*Mining of massive datasets*](http://www.mmds.org/). → **ch.11** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.2.5, ch.10.4.2** » 2017. VanderPlas, Jake; [*Python Data Science Handbook*](https://github.com/jakevdp/PythonDataScienceHandbook). → **ch.5.10 (In-depth Principal Components Analysis notebook)** » (website) -- 2017; *sklearn.neighbors.DistanceMetric class*: <http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html>. → **euclidean, cosine, jaccard**  **OPTIONAL** › 1997. Charles M. Grinstead, CM and Snell, JL; [*Introduction to Probability*](http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf). → **nice introductory resource to probability** › (website) -- 2017; *Distance computations (scipy.spatial.distance)*: <https://docs.scipy.org/doc/scipy/reference/spatial.distance.html>. → **cdist, euclidean, cosine, jaccard** › (O'Reilly Data Show podcast) -- 07-06-2017; *A framework for building and evaluating data products*: <https://www.oreilly.com/ideas/a-framework-for-building-and-evaluating-data-products>. → **listen to the entire interview** |
| **Homework** | -- |

## LECTURE 4: ASSOCIATION RULE MINING, PATTERN MINING

|  |  |
| --- | --- |
| Week of 9/20 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | introduction to concepts for rule and pattern mining; introdcution to apriori algorithm for frequent patterns; motivating the market basket analysis context for pattern mining; exploring addition contexts |
| **Expected** **Outcomes** | • understand concepts behind frequent patterns • understand association rule mining, apriori algorithm, FP-growth • apply and compute basic patterns by hand • identify the contexts for applying pattern mining |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php). → **ch.8** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.5**  **OPTIONAL** › (PartiallyDerivative.com podcast) -- 06-13-2017; *The Secret Life Of A Data Scientist*: <http://partiallyderivative.com/podcast/2017/06/13/the-secret-life-of-a-data-scientist>. → **listen to the entire podcast** |
| **Homework** | **DUE:** Monday, 10/2 - midnight Please see the Blackboard/[Github repo](https://github.com/kmsaumcis/mcis6273_f17_datamining/homework) for what to turn in. |

## LECTURE 5: UNSUPERVISED TECHNIQUES / INTRODUCTION TO CLUSTERING

|  |  |
| --- | --- |
| Week of 9/27 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | introduction to cluster analysis and motivations; introduction to unsupervised clustering algorithms; partitioning (k-means, k-mediods); hierarchical agglomerative methods; model-based (expectation-maximization) neural networks (SOM self-organizing maps); visualing with voronoi diagrams |
| **Expected** **Outcomes** | • exposure to unsupervised clustering methods, k-Means • introduction to key clustering algorithms • distinguish between partition and model-based algorithms |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php). → **ch.13, ch.14, ch.15, ch.17** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.7** » (website) -- 2015; *Basic Clustering with k-Means*: <https://nbviewer.jupyter.org/github/tmbdev/teaching-mmir/blob/master/30-kmeans.ipynb>. → **Familiarize yourself with the notebook.**  **OPTIONAL** › (LinearDigressions.com podcast) -- 04-16-2017; *Education Analytics*: <http://lineardigressions.com/episodes/2017/4/16/education-analytics>. → **listen to the entire podcast** › (website) -- --; *Programatically understanding Expectation Maximization*: <https://nipunbatra.github.io/blog/2014/em.html>. → **read this practical explanation (with Python code) of the EM algorithm** |
| **Homework** | **DUE:** Monday, 10/23 - midnight Please see the Blackboard/[Github repo](https://github.com/kmsaumcis/mcis6273_f17_datamining/homework) for what to turn in. |

## LECTURE 6: UNSUPERVISED TECHNIQUES / MORE CLUSTERING

|  |  |
| --- | --- |
| Week of 10/4 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | continued clustering, hierachical algorithms (agglomorative), introduction to density-based algorithms (DBSCAN) |
| **Expected** **Outcomes** | • understand hierarchical and density-based algorithms • develop intuitions for choosing algorithms in various contexts • utilize algorithms on read-world data |
| **Readings &** **Supplemental** | No assigned readings. Please complete readings from previous week if not current. |
| **Homework** | -- |

## LECTURE 7: SUPERVISED TECHNIQUES / CLASSIFICATION AND PREDICTION

|  |  |
| --- | --- |
| Week of 10/11 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | classification and prediction; understanding decision trees, concepts and theory; probabilistic approaches to classification - naïve bayes; introduction to bayesian belief networks |
| **Expected** **Outcomes** | • understand and explain decision trees • develop probabilistic models of classification using naïve Bayes • identify BBNs and their application context • utilize naïve Bayes in real-world applications |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php). → **ch.18, ch.19** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.6.3, ch.6.4**  **OPTIONAL** › (DataSkeptic.com podcast) -- 08-04-2017; *MINI: Bayesian Belief Networks*: <https://dataskeptic.com/blog/episodes/2017/bayesian-belief-networks>. → **explore this light discussion of BBNs** › 2012. Barber, D.; [*Bayesian Reasoning and Machine Learning*](http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/020217.pdf). → **explore ch.3 for in a deeper theoretical treatment of BBNs** |
| **Homework** | **DUE:** Monday, 11/3 - midnight Please see the Blackboard/[Github repo](https://github.com/kmsaumcis/mcis6273_f17_datamining/homework) for what to turn in. |

## LECTURE 8: SUPERVISED TECHNIQUES / CLASSIFICATION AND PREDICTION

|  |  |
| --- | --- |
| Week of 10/18 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | linear regression models for prediction; logistic regression models for prediction; introduction to generalized linear models |
| **Expected** **Outcomes** | • understand and develop linear regression models • understand and interpret logistic regression models • exposure to generalized linear models |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php). → **ch.20** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.6.11**  **OPTIONAL** › (DataSkeptic.com podcast) -- 01-27-2017; *MINI: Logistic Regression on Audio Data*: <https://dataskeptic.com/blog/episodes/2017/logistic-regression-on-audio-data>. → **listen to the entire podcast** › (website) -- --; *Building a logistic regression classifier from the ground up*: <http://inmachineswetrust.com/posts/building-logistic-regression/>. → **this is a nice explanation (and code) in Python** |
| **Homework** | -- |

## LECTURE 9: SUPERVISED TECHNIQUES / CLASSIFICATION AND MODEL EVALUATION

|  |  |
| --- | --- |
| Week of 10/25 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | support vector machines; neural networks and the basic NN model and its relation to learning algorithms; evaluating models and applying techniques to model validation |
| **Expected** **Outcomes** | • understand support vector machines and their strengths • understand neural networks, their basic theory and application • identify and develop intutition around model evaluation and validation |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php). → **ch.21** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.6.6, ch.6.7**  **OPTIONAL** › (DataSkeptic.com podcast) -- 05-27-2017; *Data Science at eHarmony*: <https://dataskeptic.com/blog/episodes/2016/data-science-at-eharmony>. → **listen to the entire podcast** |
| **Homework** | **DUE:** Monday, 11/30 - midnight Please see the Blackboard/[Github repo](https://github.com/kmsaumcis/mcis6273_f17_datamining/homework) for what to turn in. |

## LECTURE 10: ENSEMBLE METHODS

|  |  |
| --- | --- |
| Week of 11/1 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | ensemble methods; introduction to boosting, bagging, random forests and related methods |
| **Expected** **Outcomes** | • understand and identify the need for ensembles • identify and develop intutition around ensemble model evaluation and validation |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php). → **ch.22** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.6.12, ch.6.13, ch.6.14, ch.6.15** |
| **Homework** | -- |

## LECTURE 11: DATA VISUALIZATION: INTRODUCTORY CONCEPTS

|  |  |
| --- | --- |
| Week of 11/8 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | introduction to data visualization; building data narratives |
| **Expected** **Outcomes** | • understand core social mining algorithms • understand concepts in network analysis |
| **Readings &** **Supplemental** | **REQUIRED** » 2015. Knaflic, Cole Nussbaumer; [*Storytelling with data: A data visualization guide for business professionals*](http://www.storytellingwithdata.com/book/). → **ch.8** » (website) -- 2017; *D3.js: Data-Driven Documents*: <http://d3js.org>. → **familiarize yourself with some of the visualizations and capabilities of D3.js**  **OPTIONAL** › 2014. B\"orner, Katy and Polley, David E; [*Visual insights: A practical guide to making sense of data*](https://mitpress.mit.edu/books/visual-insights). → **ch.5** › (website) -- 2017; *Analyzing Scrabble Games*: <http://rpubs.com/jalapic/scrabblr>. → **This is a very interesting exploration in analysis and visualization.** › (website) -- 2017; *World Population Growth*: <https://ourworldindata.org/world-population-growth/>. → **explore some of the data and visualizations** › (website) -- 2017; *RAWGraphs: The missing link between spreadsheets and data visualization*: <http://rawgraphs.io/>. → **explore this site and its galleries** › (website) -- 2016; *Rio 2016 Medals Race: An analysis of the 2016 Olympic Medals*: <http://timesofoman.com/extra/rio_2016_medal_tally/index.html>. → **explore this visualization** |
| **Homework** | -- |

## LECTURE 12: INTRODUCTION TO SOCIAL MINING

|  |  |
| --- | --- |
| Week of 11/15 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | introduction to social mining; introduction to recommendation systems, collaborative and content-based filtering |
| **Expected** **Outcomes** | • understand core social mining algorithms • understand concepts in network analysis • understand core recommender system concepts |
| **Readings &** **Supplemental** | **REQUIRED** » 2014. Leskovec, Jure and Rajaraman, Anand and Ullman, Jeffrey David; [*Mining of massive datasets*](http://www.mmds.org/). → **ch.10** » 2015. Grus, Joel; [*Data science from scratch: First principles with Python*](http://shop.oreilly.com/product/0636920033400.do). → **ch.22**  **OPTIONAL** › 2014. Leskovec, Jure and Rajaraman, Anand and Ullman, Jeffrey David; [*Mining of massive datasets*](http://www.mmds.org/). → **ch.9** › 2014. B\"orner, Katy and Polley, David E; [*Visual insights: A practical guide to making sense of data*](https://mitpress.mit.edu/books/visual-insights). → **ch.5** |
| **Homework** | -- |

## LECTURE 13: INTRODUCTION TO TEXT MINING

|  |  |
| --- | --- |
| Week of 11/29 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | introduction to text mining; concepts in document preparation pipeline (tokenizing, stemming, etc.); TFIDF, cosine similarity; corpus selection |
| **Expected** **Outcomes** | • understand introductory concepts in text mining and information retrieval • understand document preparation tools • apply basic concepts to real-world data |
| **Readings &** **Supplemental** | **REQUIRED** » 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf). → **ch.10.4** » 2008. Manning, Christopher D and Raghavan, Prabhakar and Sch\"utze, Hinrich; [*Introduction to information retrieval*](https://nlp.stanford.edu/IR-book/information-retrieval-book.html). → **ch.6**  **OPTIONAL** › 2008. Manning, Christopher D and Raghavan, Prabhakar and Sch\"utze, Hinrich; [*Introduction to information retrieval*](https://nlp.stanford.edu/IR-book/information-retrieval-book.html). → **ch.13** › (O'Reilly Data Show podcast) -- 07-06-2017; *Language understanding remains one of AI’s grand challenges*: <https://www.oreilly.com/ideas/language-understanding-remains-one-of-ais-grand-challenges>. → **listen to the entire interview** › (LinearDigressions.com podcast) -- 04-30-2017; *Word2Vec*: <http://lineardigressions.com/episodes/2017/4/30/word2vec>. → **listen to the entire podcast** |
| **Homework** | -- |

## LECTURE 14: OPEN DATA, ETHICS IN DATA MINING, THE FUTURE OF DATA SCIENCE

|  |  |
| --- | --- |
| Week of 12/6 | [Lecture Notes](https://github.com/kmsaumcis/mcis6273_f17_datamining/tree/master/lecture_notes) |
| **Content** | open data portals, APIs, tools and technologies; ethics in data mining; anonymization, privacy and data considerations; data science and the future |
| **Expected** **Outcomes** | • exposure to open data portals and open data technologies • exposure to open APIs and tools for open data access • understand data mining ethics and why ethics (and privacy) are critically important • the future to data science, analytics and intelligent systems built on big data |
| **Readings &** **Supplemental** | **REQUIRED** » (DataStori.es podcast) -- 05-18-2016; *74 - Data Ethics and Privacy with Eleanor Saitta*: <http://datastori.es/74-data-ethics-and-privacy-with-eleanor-saitta/>. → **listen to the entire podcast** » (website) -- 2017; *ProgrammableWeb.com: The Journal of the API Economy*: <https://www.programmableweb.com/>. → **familiarize yourself with this site and some APIs**  **OPTIONAL** › (LinearDigressions.com podcast) -- 08-13-2017; *Curing Cancer with Machine Learning is Super Hard*: <http://lineardigressions.com/episodes/2017/8/13/curing-cancer-with-machine-learning-is-super-hard>. → **listen to the entire podcast** |
| **Homework** | -- |

## RESOURCES

1. 2014. Zaki, Mohammed J and Meira Jr, Wagner; [*Data mining and analysis: fundamental concepts and algorithms*](http://www.dataminingbook.info/pmwiki.php).
2. 2014. Leskovec, Jure and Rajaraman, Anand and Ullman, Jeffrey David; [*Mining of massive datasets*](http://www.mmds.org/).
3. 1997. Charles M. Grinstead, CM and Snell, JL; [*Introduction to Probability*](http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf).
4. 2011. Yau, Nathan; [*Visualize this: the FlowingData guide to design, visualization, and statistics*](http://book.flowingdata.com/).
5. 2014. B{\"o}rner, Katy and Polley, David E; [*Visual insights: A practical guide to making sense of data*](https://mitpress.mit.edu/books/visual-insights).
6. 2012. Downey, Allen; [*Think Python*](http://www.greenteapress.com/thinkpython/thinkpython.html).
7. 2012. Conway, Drew and White, John; [*Machine learning for hackers*](http://shop.oreilly.com/product/0636920018483.do).
8. 2015. Grus, Joel; [*Data science from scratch: First principles with Python*](http://shop.oreilly.com/product/0636920033400.do).
9. (website) -- 2017; *The Periodic Table of Data Science*: <https://www.datacamp.com/community/blog/data-science-periodic-table#gs.TF297Gsm>.
10. 2011. Han, Jiawei and Pei, Jian and Kamber, Micheline; [*Data mining: concepts and techniques*](https://ia800300.us.archive.org/5/items/DataMiningConceptAndTechniques2ndEdition/Data.Mining.Concepts.and.Techniques.2nd.Ed-1558609016.pdf).
11. 2012. McKinney, Wes; [*Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*](https://github.com/wesm/pydata-book).
12. 2008. Manning, Christopher D and Raghavan, Prabhakar and Sch{\"u}tze, Hinrich; [*Introduction to information retrieval*](https://nlp.stanford.edu/IR-book/information-retrieval-book.html).
13. 2015. Knaflic, Cole Nussbaumer; [*Storytelling with data: A data visualization guide for business professionals*](http://www.storytellingwithdata.com/book/).
14. 2016. Rose, Doug; [*Data Science: Create Teams That Ask the Right Questions and Deliver Real Value*](http://www.apress.com/us/book/9781484222522).
15. (website) -- 2013; *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*: <https://github.com/ptwobrussell/Mining-the-Social-Web-2nd-Edition/>.
16. 2017. Wexler, Steve and Shaffer, Jeffrey and Cotgreave, Andy; [*The Big Book of Dashboards: Visualizing Your Data Using Real-World Business Scenarios*](http://bigbookofdashboards.com).
17. 2017. VanderPlas, Jake; [*Python Data Science Handbook*](https://github.com/jakevdp/PythonDataScienceHandbook).
18. (website) -- 2015; *Basic Clustering with k-Means*: <https://nbviewer.jupyter.org/github/tmbdev/teaching-mmir/blob/master/30-kmeans.ipynb>.
19. (website) -- 2017; *Distance computations (scipy.spatial.distance)*: <https://docs.scipy.org/doc/scipy/reference/spatial.distance.html>.
20. (website) -- 11-15-2016; *Jupyter Notebook Tutorial: The Definitive Guide*: <https://www.datacamp.com/community/tutorials/tutorial-jupyter-notebook#gs.zExWvMw>.
21. (website) -- 2017; *Pandas Cookbook*: <https://github.com/jvns/pandas-cookbook>.
22. (website) -- 2017; *sklearn.neighbors.DistanceMetric class*: <http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.DistanceMetric.html>.
23. ({Michael Kennedy's Talk Python To Me} podcast) -- 11-28-2016; *Episode #90: Data Wrangling with Python*: <http://talkpythontome.fm>.
24. ({O'Reilly Data Show} podcast) -- 07-06-2017; *A framework for building and evaluating data products*: <https://www.oreilly.com/ideas/a-framework-for-building-and-evaluating-data-products>.
25. ({O'Reilly Data Show} podcast) -- 07-06-2017; *Language understanding remains one of AI’s grand challenges*: <https://www.oreilly.com/ideas/language-understanding-remains-one-of-ais-grand-challenges>.
26. (PartiallyDerivative.com podcast) -- 06-13-2017; *The Secret Life Of A Data Scientist*: <http://partiallyderivative.com/podcast/2017/06/13/the-secret-life-of-a-data-scientist>.
27. (LinearDigressions.com podcast) -- 04-16-2017; *Education Analytics*: <http://lineardigressions.com/episodes/2017/4/16/education-analytics>.
28. (LinearDigressions.com podcast) -- 06-04-2017; *PageRank*: <http://lineardigressions.com/episodes/2017/6/4/pagerank>.
29. (LinearDigressions.com podcast) -- 08-13-2017; *Curing Cancer with Machine Learning is Super Hard*: <http://lineardigressions.com/episodes/2017/8/13/curing-cancer-with-machine-learning-is-super-hard>.
30. (LinearDigressions.com podcast) -- 04-30-2017; *Word2Vec*: <http://lineardigressions.com/episodes/2017/4/30/word2vec>.
31. (DataStori.es podcast) -- 05-18-2016; *74 - Data Ethics and Privacy with Eleanor Saitta*: <http://datastori.es/74-data-ethics-and-privacy-with-eleanor-saitta/>.
32. (website) -- 2017; *ProgrammableWeb.com: The Journal of the API Economy*: <https://www.programmableweb.com/>.
33. (website) -- 2017; *Analyzing Scrabble Games*: <http://rpubs.com/jalapic/scrabblr>.
34. (website) -- 2017; *GSS Data Explorer*: <https://gssdataexplorer.norc.org/>.
35. (website) -- 2017; *World Population Growth*: <https://ourworldindata.org/world-population-growth/>.
36. (website) -- 2017; *RAWGraphs: The missing link between spreadsheets and data visualization*: <http://rawgraphs.io/>.
37. (website) -- 2016; *Rio 2016 Medals Race: An analysis of the 2016 Olympic Medals*: <http://timesofoman.com/extra/rio_2016_medal_tally/index.html>.
38. (website) -- 2017; *D3.js: Data-Driven Documents*: <http://d3js.org>.
39. (DataSkeptic.com podcast) -- 08-04-2017; *MINI: Bayesian Belief Networks*: <https://dataskeptic.com/blog/episodes/2017/bayesian-belief-networks>.
40. (DataSkeptic.com podcast) -- 01-27-2017; *MINI: Logistic Regression on Audio Data*: <https://dataskeptic.com/blog/episodes/2017/logistic-regression-on-audio-data>.
41. (DataSkeptic.com podcast) -- 05-27-2017; *Data Science at eHarmony*: <https://dataskeptic.com/blog/episodes/2016/data-science-at-eharmony>.
42. (website) -- --; *Programatically understanding Expectation Maximization*: <https://nipunbatra.github.io/blog/2014/em.html>.
43. (website) -- --; *Building a logistic regression classifier from the ground up*: <http://inmachineswetrust.com/posts/building-logistic-regression/>.
44. 2012. Barber, D.; [*{Bayesian Reasoning and Machine Learning}*](http://web4.cs.ucl.ac.uk/staff/D.Barber/textbook/020217.pdf).