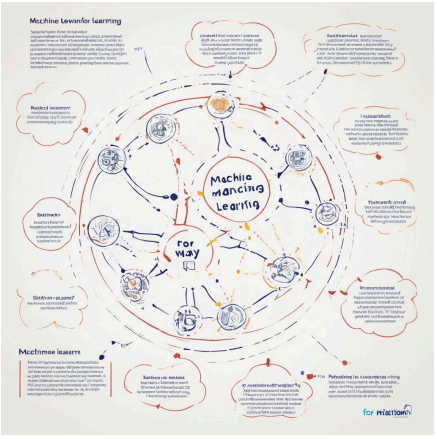
MLMW: Machine Learning My Way

Create a personal study guide for topics in econometrics and machine learning, research design and model productisation, business change and society impacts.

[Part 1. Theory from textbooks](#)
[Part 2. Data types, sources and quality](#)
[Part 3. From research design to model productisation](#)
[Part 4. Software tools](#)
[Part 5. Business change and society impact](#)
[Appendix](#)



Last updated: April 7, 2024.
Version 0.4.1

Target use cases for this publication:

- planning your studies at the beginning (school or college);
- switching fields and learning about adjacent subject areas (e.g. engineering to data science, changing filed in masters degree);
- review and reference for advanced learners (PhD level);
- use as a reader in class when teaching.

Drop me a note if you got an idea how to improve this guide or using it for training or teaching: e.pogrebnyak+mlmw@gmail.com

Links to this document:

- viewer mode: https://t.ly/RcA2Q
- editor mode (restricted): MLMW: Machine Learning My Way

Website: https://epogrebnyak.github.io/mlmw/
Telegram: https://t.me/ml_my_way
Subscribe for updates: https://buttondown.email/mlmw/

- Part 1. Theory from textbooks
  - Intuition and foundational concepts
  - Econometrics and inference
  - Machine learning tasks and methods
  - Deep learning and neural networks
  - Other modelling approaches
    - Departing from frequentism: Bayes and causality
    - Choosing and combining models
    - Interaction, feedback and networks
  - Less obvious or exciting topics
  - Textbook review and prerequisites
- Part 2. Data types, sources and quality
  - Tabular data
  - Not just numbers: text, images and sound
  - Data from business perspective
- Part 3. From research design to model productisation
  - Steps in analysis
  - ML in production
- Part 4. Software tools
  - Programming languages and statistical software
  - Databases and storage
  - Orchestration
- Part 5. Business change and society impact
  - Technology companies as machine learning market players
  - Adoption in broad economy and society
  - Selected industry domains and use cases
- Appendix
  - More personal skills
  - Personas
  - Glossary
  - Changelog: timeline of this document
  - Reader and peer feedback

Part 1. Theory from textbooks

Intuition and foundational concepts

1. Probability and randomness.

    - Probability as repeated events (Bernoulli) vs plausibility estimate (ET Janes).
    - Random variable and its distribution. Distribution as a mass of one.
    - Sequence of events and conditional probability.
    - Joint distribution and marginal probability.
    - Axioms of probability and measure theory.
    - Generating random numbers practically, pseudorandom and seed.

2. Sample vs population. Sampling techniques and bootstrap.

3. Data generating process (DGP). Non-parametric methods.

    We rarely know the true DGP, but might guess its functional form and try to estimate the parameters in that functional form. If we succeed the inference problem is complete.

4. Inference (econometrics) vs prediction and forecasting (machine learning).

5. Correlation, causality, common drift and spurious regressions.

    Example: German Cheeses and a directed acyclic graph (DAG).

6. Observation, experiment and experiment design.

7. Measurement errors and missing data.

8. Model performance metrics and model evaluation

Econometrics and inference

Survey article: Undergraduate Econometrics Instruction: Through Our Classes, Darkly.

9. Linear regression and ordinary least squares (OLS).

10. Violation of OLS assumptions.

    Note: Kennedy textbook is built on listing the violations, very clear to follow.

11. Regression discontinuity.

12. Time series. Seasonal adjustment, smoothing, filtering.

Reference text: Hamilton.

13. Systems of equations.

- Klein's model in Cowles Commmision papers (1950).
- Goldberger (1972). Structural Equation Methods in the Social Sciences.
- Heckman and Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation.

Inference methods

14. Methods of estimation.

    OLS and extensions.
    Maximum likelihood.
    Bayesian estimation
    MCMC.

15. OLS extensions.

    logit/probit
    GMM
    2- and 3- stage OLS
    Quantile regressions
    Lasso, ridge

Machine learning tasks and methods

16. Clustering.

17. Classification.

18. Dimensionality reduction.

19. Decision trees.

20. Support vector machines (SVM).

Deep learning and neural networks

21. How does a simple neural network like a perceptron work? How do more complex networks train and operate?

    Andrew Ng course.
    Brandon Rohrer: How Neural Networks Work

22. Neural network architectures

    Feed-forward Neural Network
    Convolutional Neural Network (CNN)
    Recurrent Neural Network (RNN)
    Generative Adversarial Network (GANs)
    Transformers

23. GPT models. Interaction, dialogue, prompt engineering. Retrieval augmented generation (RAGs). Model fine-tuning.

24. One-shot, federated, transfer learning. Artificial general intelligence (AGI).

    Economist: How to define artificial general intelligence.

Other modelling approaches

Departing from frequentism: Bayes and causality

25. Bayes theorem and Bayesian modeling. Probabilistic programming.

    Probability and Bayesian Modeling (2020)

26. Causality and do-notation.

    Book of Why by Judea Pearl.
    Causal Inference The Mixtape by Scott Cunningham.
    Causal ML Book by Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, Vasilis Syrgkanis (2024)

    Software: PyWhy

Choosing and combining models

27. Ensembles and forecast combination.

28. AutoML.

Interaction, feedback and networks

29. Agents. Game theory. Reinforcement learning. Auction design.

30. Systems with feedback and system dynamics (SD).

31. Graphs and networks.

Less obvious or exciting topics

32. Combinatorics.
33. Convergence and central limit theorems. Asymptotics in theoretical econometrics.
34. Random distributions and their families.
35. Hypothesis testing.
36. Differentiation and differential equations.
37. Random processes.
38. Optimisation models and solvers.
39. Boolean vs fuzzy logic. Qubit.
40. Probability as part of measure theory.

    Reddit quote:

    > What is necessary however, is to understand measure-theoretic probability. If you have a solid foundation in measure theory, that should be quite straightforward. You will see that only a subset of results/theorems from measure theory make frequent appearences. These include the Fubini/Tonelli theorem, absolute continuity, the Randon-Nikodym derivative, the Borel-Cantelli lemma etc. and you can always refer back to your measure theory book for things.

    > For probabilistic/statistical machine learning, you almost always assume probability measures are dominated by the Lebesgue measure on the underlying Euclidean space and work directly with pdfs. The only area of ML theory that I know of that is measure-theory heavy is PAC-Bayes / advanced statistical learning theory.

Textbook review and prerequisites

41. More beginner-friendly vs more advanced texts (with focus on open source texts).

Subject Area	Beginner-friendly	Advanced	Other Classics
Probability			Wentzel (1982)
Statistics			Casella/Berger

Econometrics			Green, Stock/Watson
Machine Learning	ISLR/ISLP	PRML (Bishop) PML (Murphy)	
Deep Learning and AI	Andrew Ng	A lot here	Russel/Norvig

Reddit Quote:

> For the longest time the best books for a mathematical treatment of ML were Chris Bishop's "Pattern Recognition and Machine Learning" and Kevin Murphy's "Machine Learning: A Probabilistic Perspective". Both authors have written new and updated books, better adapted to the deep learning era. Bishop's new book is "Deep Learning: Foundations and Concepts". Murphy released two books: "Probabilstic Machine Learning: An Introduction" and "Probabilistic Machine Learning: Advanced Topics".
>
> I would say that Murphy's two tome book currently provides the most comprehensive and thorough treatment of probabilistic ML. The first chapters of the introductory book are basically mathematical preliminaries, so it's more accessible than before. Additionally, the most frequently used book for getting a stong mathematical foundation for ML is "Mathematics for Machine Learning" by Deisenroth et al.

42. Mathematical prerequisites.

    Calculus.

    Linear algebra.

    Mathematical statistics.

Part 2. Data types, sources and quality

Tabular data

43. Table in a dataframe and in a relational database. Data types and table schema. Data serialization.

44. Textbook datasets. Kaggle and similar datasets. Official statistics, data search, open data.

Not just numbers: text, images and sound

45. Text as vector. Natural language processing.

46. Images and video. Color representation. Computer vision.

47. Sound, noise, music. Waves.

Data from business perspective

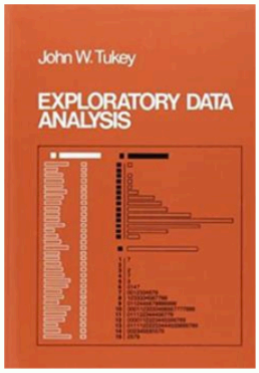
48. Data quality and limitations (incomplete, not granular, not relevant, not yours). Data governance frameworks (DMBOK).

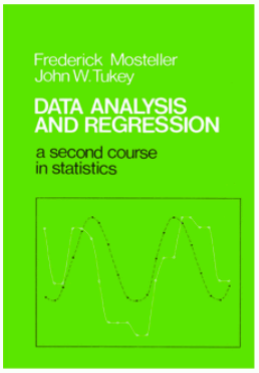
Part 3. From research design to model productisation

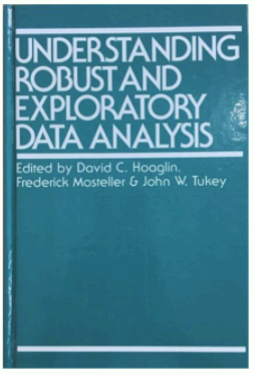
Steps in analysis

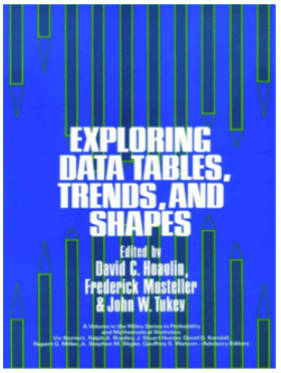
49. Descriptive statistics.

    John Tukey and the Origins of EDA
    Remembrances of Things: EDA (article about John W. Tukey work).
    The 4 R's in EDA (from A Course in Exploratory Data Analysis)









50. Analysis is a DAG.

    Drivendata cookiecutter as a project template example (but not a universal template).

51. Data visualization. Dashboards.

    Economist: Mistakes, We Have Drawn a Few.
    Grammar of Graphics (ggplot)

52. Typical projects: interacting with business or retail customer, controlling our own system or other. Advanced analytics / dashboards vs experiments / decision-making / automation of control.

53. Reproducible research and reproducibility crisis.

ML in production

54. Responsibilities of a data engineer, data scientist or modeller, machine learning engineer, business analyst and other roles.

55. Model life cycle, model drift vs data drift.

56. Machine learning pipelines and MLOps.

    https://github.com/EthicalML/awesome-production-machine-learning

Part 4. Software tools

Programming languages and statistical software

57. Programming languages (R, Python, Julia, Mojo) and machine learning libraries.

    R for statistical packages and classic statistics
    Python for machine learning (scikit-learn) and deep learning (PyTorch and TensorFlow).

    Exercise: Compare a dataframe implementation in R, Python and Julia.

58. Stattistical software, proprietary vs open source and documentation.

    Who won in the statistics market, our hearts and pockets: proprietary (SAS, MATLAB) vs open source (R, Julia).

    Software documentation as learning tools (great to read even if you are not using the package):
    - MATLAB
    - gretl
    - eviews
    - scikit-learn lectures

59. Open source viability and funding models

Example: Streamlit.

60. Cleaner code. Version control. Notebooks vs plain files.

Databases and storage

61. Disk: HDD, SDD and cloud (S3) storage. Disk costs and time to access. File systems (HDFS).

62. Database management systems (DBMS). Relational databases and SQL.

    Other types of databases and NoSQL (key-value, graph, column, vector, time series).

    Processing large data in parallel: MapReduce, Hadoop (HDFS+Yarn+MapReduce), Spark.

    Search databases (ElasticSearch, Splunk, Solr)

    Database popularity:
    - https://db-engines.com/en/ranking
    - https://www.jetbrains.com/lp/devecosystem-2023/databases/
    - https://survey.stackoverflow.co/2023/#section-admired-and-desired-databases

63. Data warehouses and DW architectures. Decoupling storage and compute.

64. Cloud providers (AWS, GCP, Azure). New data solution providers (Snowflake, Databricks).

65. Mergers and acquisitions, venture financing and forks:

    Sun buys MySQL (2008), Oracle buys Sun (2010), EU and US antitrust approval.
    SAP's acquisition of Sybase (2010).
    Cloudera and Hortonworks merger (2019).
    Valkey, a Redis fork after licence change (2024).

    Exercise: find venture-funded database projects here and explain their valuations.

66. More elements of database theory and implementation. Relational algebra and ER-diagrams. ACID, CAP, BASE. DDL and DML. Normalization and normal forms. Hashing and B-trees. OLAP and OLTP.

    Extra video: The ancient art of data management (2023) from DuckDB co-founder.
    Extra reading: Lecture notes on database engineering (VSUUT, India).

Orchestration

67. Workflow and orchestration tools (Airflow, Prefect, Luigi, etc).

Part 5. Business change and society impact

Making money vs making change

Technology companies as machine learning market players

68. Companies that sell ML tools and solutions, their valuations and strategy.
69. Cloud providers.
70. Hardware providers.
71. Internet-scale data owners (Google).

Adoption in broad economy and society

72. Fairness, biases, equity, human loop.
73. Economics, cost and payoffs of applying ML. Business value of ML.
74. Job market: in-house data modeller, consultant or a vendor?
75. Who’s got more data? Data privacy and data protection. Markets and pricing of data.
76. What gets to be regulated. Does a national AI or data strategy make sense?
77. Why the hype: what makes the corporation play hype and overpromise? Why do investors buy that?

Selected industry domains and use cases

78. Social sciences (sociology, political science, psychology, anthropology).
79. Recommender systems (RecSys).
80. Clinical trials.
81. Quality control and dependability.

82. ML in finance (Halperin textbook + efficient market hypothesis + SSRN most popular papers)

Appendix

More personal skills

83. Common sense, logic and critical reasoning.

84. Writing well, explaining, inquiring, communicating.

On Writing Well: An Informal Guide to Writing Nonfiction. Also American Scholar article: On Writing Well and keeping it up-to-date for 35 years by William Zinsser.

Personas

85. People that strike me as great thinkers and educators who make complex thing easy to follow for the rest of us – through courses, books and personal interaction:

- Will Curt
- Scott Cunningham
- Laura Mayoral

You also may be surprised a textbook or a professor can be reachable on Twitter/X or other social media:

- Jeffrey Wooldridge (@jmwooldridge)
- Paul Goldsmith-Pinkham (@paulgp), repo for Yale Applied Empirical Methods PHD Course and the video list

86. Hall of shame: story of Siraj Raval.

Glossary

87. Very common terms – you hear it everywhere

- Supervised vs unsupervised.
- Structured vs unstructured data.

88. Professional slang

- Feature engineering (glorified variable selection and transform).

89. Buzzwords

- AI
- Data mining

Changelog: timeline of this document

Answered on Reddit

https://www.reddit.com/r/datascience/comments/1bwsdgn/recommend_good_books_courses/

April 6, 2024 (88 topics)
Finalised :
- Databases and storage

Edited:
- SEMs and references from Paul Goldsmith-Pinkham

Marked for review:
- DAG, project flow.

To add next:
- catboost
- New causal ML book

March 31, 2024
The topic count is 77, also organized into textbook, data, project, adoption and cases sections. Good reception of the list in comments on Reddit, but removed by moderators, no specific reason or stated.

March 29, 2024
A way to keep up with data modeling and sort out what you already know. So far it is a list of topics organized by section, perhaps somewhat upside down compared to a traditional textbook or a course, but I hope you like the perspective. Few links added where most appropriate and I remembered good stuff. There are open textbook and blog links at Econometrics Navigator website, my previous work. 33 topics in original post.

March 28, 2024
First published as a Reddit post.

Reader and peer feedback

This is pretty neat. Congratulations on putting together such a great list!

Amazing, thanks man, also it would also be much better to provide resource lists as well, still pretty useful, thanks!

This list is gold, thanks :)

This list is pretty comprehensive. I would have a bit on MLOps side because most advanced practitioners of ML should have some amount of understanding of how models are productionized. Perhaps a few topics on model drift, data drift, understanding how experiments are set up etc can be beneficial. Overall looks pretty good and will probably even use this to brush up on my own skills.