

Exercise: decide which pictures above are computer-generated.

# MLMW: Machine Learning My Way

MLMW is a list of topics and learning resources that should help you identify and fill up knowledge and skill gaps in machine learning.

© Evgeny Pogrebnyak, 2024

[e.pogrebnyak+mlmw@gmail.com](mailto:e.pogrebnyak+mlmw@gmail.com)

Last updated: April 29, 2024.

Version 0.7.0

Get the newest version:

- Mailing list: <https://buttondown.email/mlmw/>
- Reddit: [r/ml\\_my\\_way](https://www.reddit.com/r/ml_my_way)
- Telegram: [https://t.me/ml\\_my\\_way](https://t.me/ml_my_way)

Work in progress (for next release):

- State of AI report.
- Aubrey Clayton videos about ET Janes.
- ITMO ML role model figure.
- Focus on structure of parts 2-6.
- Try more with typst and mdbook.

## [Part 1. Models and methods](#)

[Intuition and foundational concepts](#)

[Econometrics](#)

[Machine learning](#)

[Deep learning and neural networks](#)

[Artificial intelligence](#)

[Other modeling approaches](#)

[Bayes and causality](#)

[Choosing and combining models](#)

[Interaction, feedback, networks and optimisation](#)

[Harder, less obvious or exciting topics](#)

[Mathematical prerequisites](#)

[Textbook review](#)

## [Part 2. Data types, sources and quality](#)

[Tabular data](#)

[Not just numbers: text, images and sound](#)

[Data in business and economic 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 and data engineering tools](#)

### [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](#)

[Interviews](#)

[More personal skills](#)

[Personas](#)

[Glossary](#)

[Other resources](#)

[Changelog – timeline of this document](#)

[Quotes and reader feedback](#)

# Part 1. Models and methods

## Intuition and foundational concepts

### 1. Probability and randomness.

- Probability as repeated events (Bernoulli) vs plausibility estimate (ET Janes).
- Random variables and their distributions.
- Sequence of events and conditional probability.
- Joint distribution of random variables and marginal probability.
- Axioms of probability and measure theory.
- Generating random numbers practically (pseudorandom and seed).

See a listing of probability textbooks at the end of the chapter.

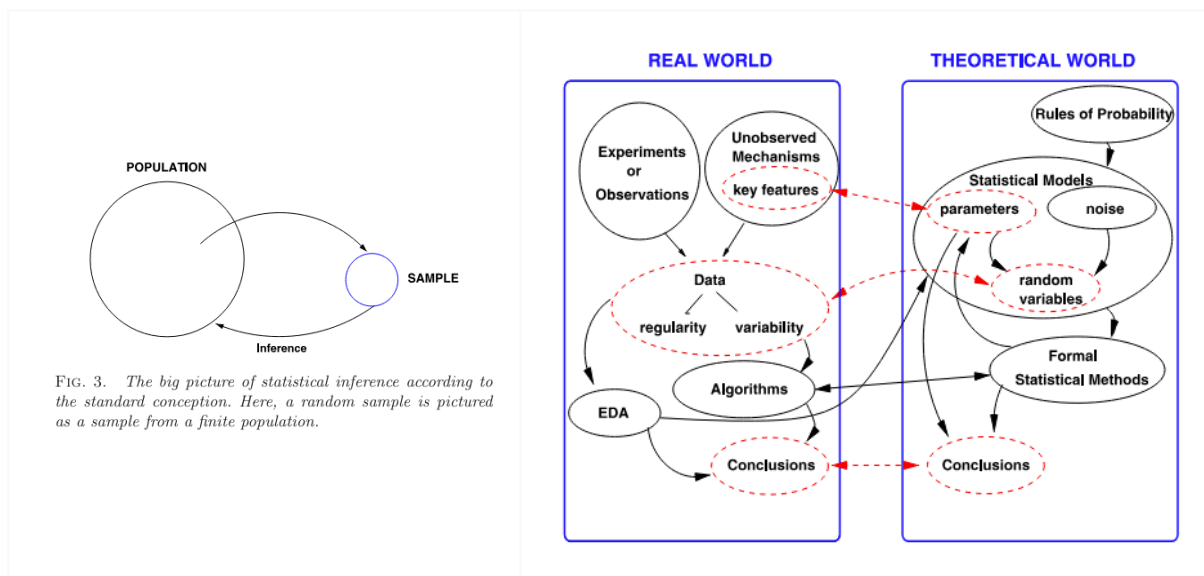
### 2. Data generating process (DGP).

We rarely know the true model of a DGP, but might reason about its functional form based on theory or prior knowledge or take the most simple functional form as a guess. Next we estimate the parameters in that functional form given the observation data points that we have. In general in statistics we rely on a notion there is some

discoverable law, the data generating process, that produces the data that we collect and analyse.

### 3. Sample vs population. Learning from sample about population.

“Statistical pragmatism emphasizes the assumptions that connect statistical models with observed data.” Figures from [Robert E. Kass \(2012\) Statistical Inference: The Big Picture](#).



### 4. Inference (econometrics) and generalization (machine learning).

- Inference and parameter interpretation (statistics and econometrics).
- Generalization and prediction (machine learning or statistical learning).
- Change of behavior (causal inference and heterogeneous treatment).

Readings:

- Presentation: [Hal Varian \(2014\). Machine Learning and Econometrics](#).
- Article: [Susan Athey and Guido Imbens \(2019\). Machine Learning Methods That Economists Should Know About](#).
- Presentation: [The Impact of Machine Learning on Economics and the Economy \(2019\)](#).

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

Example: German Cheeses and a directed acyclic graph (DAG) in [Determining causality in correlated time series](#).

### 6. Observation, experiment and experiment design.

### 7. Measurement errors and missing data.

### 8. Model performance and model evaluation. Modeling trade-offs (eg bias vs variance) and no free lunch.

# Econometrics

Survey article: [Undergraduate Econometrics Instruction: Through Our Classes, Darkly.](#)

Modern introduction course: [Mathematical Econometrics I by Roth and Hall](#)

9. Cross-section, time series, panel and spatial data. Single vs multivariate response variable. Econometrics for macro, micro and finance.
10. Linear regression and ordinary least squares (OLS).
11. Violation of OLS assumptions.

Note: Peter Kennedy textbook (1998) is built on listing the violations, very clear to follow.

12. Difference-in-Differences. Instrumental Variables. Regression Discontinuity.

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

Reference text: Hamilton.

See more textbooks in [Econometrics Navigator: Time Series Section](#).

Extra: [Forecasting: Principles and Practice by Hyndman and Athanasopoulos](#).

14. Systems of equations.

Important part of econometrics for imposing a structure. The rise and fall of large macroeconomic models in 1960s-1970s.

- [Klein's model](#) in Cowles Commission papers (1950).
- Goldberger (1972). [Structural Equation Methods in the Social Sciences](#).
- Heckman and Vytlacil (2005). [Structural Equations, Treatment Effects, and Econometric Policy Evaluation](#). (EP: kind of says the field is alive and well.)

## Estimation

15. Methods of estimation.

OLS and extensions.  
Maximum likelihood.  
Bayesian estimation  
MCMC.

16. OLS extensions.

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

Lasso, ridge

## Machine learning

### 17. Textbooks

ISLR/ISLP is a career starter. Bishop or Murphy for more advanced text with more math, they are also older and not supplemented with code. Do not confuse ISLR/ISLP with Elements of Statistical Learning (ESL), a more mathematically rigorous book.

Acronym	Title	Authors	Latest edition
ISLP	An Introduction to Statistical Learning	Gareth James, Daniela Witten, Trevor Hastie, Rob Tibshirani	2023
PRML	Pattern Recognition and Machine Learning	Christopher Bishop	2006
PML	Probabilistic Machine Learning	Kevin Patrick Murphy	2012. Follow-up books in 2022, 2023.

[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: "Probabilistic 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 strong mathematical foundation for ML is "Mathematics for Machine Learning" by Deisenroth et al.

[CS229 Lecture Notes](#) Fall 2022 by Andrew Ng. Very good structure.

For machine learning (not deep learning), I recommend the Andrew Ng [lecture notes](#) from Stanford's CS229 course. The reason I really like these notes is because you can find past [problem sets](#) that went along with them, and the problem sets are very good: difficult but not impossible, and close to a 50/50 mix of math and programming. I never feel like I've learned a topic just from reading about it, so having good problems to go along with the reading was very important to me ([Reddit quote](#)).

[scikit-learn: machine learning in Python by Gael Varoquaux](#). Part of Scientific Python Lectures. "One document to learn numerics, science, and data with Python."

Two books below are paywalled, but very practical.

[Andreas Mueller](#) and [Sarah Guido](#) (2016). Introduction to Machine Learning with Python.

HOML: [Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow \(3rd edition\)](#) by [Aurélien Geron](#). Repo: <https://github.com/ageron/handson-ml3/>

18. Statistical learning theory. Supervised vs unsupervised learning.

Beginner: Chapter 8 “When Models Meet Data” from Mathematics for Machine Learning by Deisenroth et al (2020).

Advanced: John Shawe-Taylor (2023). Statistical Learning Theory for Modern Machine Learning ([video](#), [slides](#))

19. Classification.

20. Prediction.

21. Clustering.

22. Dimensionality reduction.

23. Decision trees.

24. Support vector machines (SVM) and discriminant analysis.

## Deep learning and neural networks

25. 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](#)

Textbooks (all freely available online):

- Ian Goodfellow, Yoshua Bengio and Aaron Courville. [Deep Learning Book \(DLB\)](#)
- Aston Zhang, Zack C. Lipton, Mu Li and Alex J. Smola. [Dive into Deep Learning](#) (d2l) (with notebooks).
- Simon Prince. [Understanding Deep Learning](#) (UDL) (with notebooks).

Paywalled:

- [Deep Learning: Foundations and Concepts](#) by Bishop and Bishop.

26. Neural network architectures

Feed-forward Neural Network

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

27. GPT models. Interaction, dialogue, prompt engineering. Retrieval augmented generation (RAGs) and model fine-tuning.

28. One-shot, federated, transfer learning and other advances in deep learning.

## Artificial intelligence

Short: Not everything in a NN in AI. AI winter ended with backpropagation and rise of computational power. No AGI (yet), a computer cannot “think”.

29. State of AI Report: [Artificial Intelligence Index Report 2024](#)

Great [summary by a Reddit user](#), beat that clarity: “AI good, getting gooder. Tech > academics. AI costs \$\$\$, gonna cost \$\$\$\$\$. US numba 1. Benchmarks are meh. GenAI is trending. AI regulations increase. People & science can and do benefit from AI. People have begun to pay attention to AI.”

30. History and branches of AI.

[microsoft/AI-For-Beginners: 12 Weeks, 24 Lessons](#) (contains lessons on symbolic approach with Knowledge Representation and reasoning, Genetic Algorithms and Multi-Agent Systems in addition neural nets).

[Annotated History of Modern AI and Deep Learning](#) by Jürgen Schmidhuber.

Textbook: [Artificial Intelligence: A Modern Approach](#) by Norvig and Stuart (a bit old).

31. Artificial general intelligence (AGI).

[Economist: How to define artificial general intelligence.](#)

## Other modeling approaches

For modelling topic classification see [JEL Classification System: Mathematical and Quantitative Methods](#). Compare with [AMS Classification](#) for subjects in mathematics.

## Bayes and causality

Departing from frequentism: Bayesian modelling and causality.

32. Bayes theorem and Bayesian modeling. Probabilistic programming.

[Probability and Bayesian Modeling](#) (2020)

### 33. 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](#), EconML

## Choosing and combining models

### 34. Ensembles and forecast combination.

### 35. AutoML.

## Interaction, feedback, networks and optimisation

### 36. Operations research and statistical decision theory.

### 37. Agents. Reinforcement learning. Game theory. Auction design.

### 38. Systems with feedback and system dynamics (SD). Control theory.

### 39. Graphs and networks. Knowledge graphs.

### 40. Optimisation models and solvers.

Linear programming (LP). PuLP package.

[Convex Optimization textbook by Boyd and Vandenberghe.](#)

Several book suggestions [here](#) in a Reddit post.

## Harder or less obvious

### 41. Combinatorics.

### 42. Random variable distributions and their families.

### 43. Point estimation. Confidence intervals. Hypothesis testing.

[Informal review on hypothesis testing in Notes for Nonparametric Statistics.](#)

### 44. Convergence and central limit theorems. Asymptotics.



#### 45. Sampling methods and techniques.

Sampling techniques: cross-validation, bootstrap, jackknife.

Course (advanced): [Keisuke Hirano and Jack Porter \(2022\). Modern Sampling Methods: Design and Inference.](#)

#### 46. Non-parametric methods.

[Introduction to Nonparametric Statistics by Bodhisattva Sen](#)  
[All of Nonparametric Statistics](#) by Larry Wasserman

Reddit quotes:

- <https://www.reddit.com/r/statistics/comments/obhpte/comment/h3og15r/>
- <https://www.reddit.com/r/statistics/comments/1bya8dd/comment/kyi7i8q/>

#### 47. Differentiation and differential equations.

#### 48. Random processes.

#### 49. Information theory. Entropy.

#### 50. Boolean vs fuzzy logic. Qubit and quantum computing.

#### 51. Knowledge representation and ontologies.

#### 52. 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 appearances. 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.

## Mathematical prerequisites

[Reddit user:](#) Eventually, after years of trying to get in through various "shortcuts" I realised that I actually have to learn maths and statistics like all the other guys.

- **From d2l preface:** Linear Analysis by Bollobás (1999) covers linear algebra and functional analysis in great depth. All of Statistics (Wasserman, 2013) provides a

marvelous introduction to statistics. Joe Blitzstein's books and courses on probability and inference are pedagogical gems.

- **Mathematics for Machine Learning (MML)** by Marc Peter Deisenroth, A. Aldo Faisal and Cheng Soon Ong (2020) - highly recommended. Part 1 is math and part 2 is math application to classic problems of regression, dimensionality reduction, density estimation and classification. Part 2 is focused just on these four problems, which gives a feeling of completeness and achievement. Chapter 8 "When Models Meet Data" is a great introduction to statistical learning theory.

### 53. Calculus.

What is Calculus I, II and III in the US: [Naming of calculus courses](#).  
See [Econometrics Navigator – Mathematic preliminaries – Calculus Real analysis](#) introduction by Hunter (advanced).

### 54. Linear algebra.

See [Econometrics Navigator – Mathematic preliminaries – Linear Algebra](#)

### 55. Probability and mathematical statistics.

- Blitzstein and Hwang and video series. [Statistics 110](#), Wackerly, Mood. Introduction to Probability Theory by Hoel, Port and Stone and some others discussed here: [Best Probability Theory textbook? : r/math](#).
- [Using Julia for Introductory Statistics](#) by John Verzani - you can dismiss the fact it is in Julia, very thoughtful text in stats in general.
- [Why Another Probability Textbook?](#) (2022). [Probability and Statistics Cookbook](#) (2011).
- Larry Wasserman "All of statistics" is the textbook for [10-705 Intermediate Statistics Course](#) (see lectures notes).
- A Reddit comment in [Overlap between Mathematical Statistics and Probability Theory textbooks](#):

An intermediate probability course must include:

1. Probability spaces (axiomatic development of sigma algebras and Probability measure, besides the usual topics)
2. Random variables (as particular cases of measurable functions, probability distribution and density functions with their properties, and the usual intro to common RVs)
3. Random vectors (similar to R variables and criteria of independence, conditional densities)
4. Distribution of R variables and order statistics
5. Moments of RVs (characteristic function and its properties are very important)
6. (most important) Asymptotic theory (convergence in mean squared, in probability and in distribution, and related theorems (Markov, Tchebychev, Slutsky, Helly-Bray, Lévy, etc.))

Now, Wackerly offers a very good introduction to topic 1 (without the axiomatic development of sigma algebras and probability spaces), topics 2 and 3 (without explaining measurability of RVs). It is very clear explaining random vectors. Topic 4 is very well covered and topic 5 (without the characteristic function). Does not touch topic 6, which is essential for inference. Its main advantages are the examples, the clarity of the explanations and the visual organization which makes it very easy to read.

I would suggest you read Wackerly first (till chapter 6) and then read about asymptotics in Hogg's Intro to Math Stats (8th ed.) or Roussas' A Course in Matg Stats (2nd ed) or Mittelhammer's Math. Stats. for Economics and Business (2nd ed.).

That's the probability you are going to need if you want to study statistical inference at an intermediate level. Wackerly is rather elementary in this regard, but can be an excellent introduction to inference if you are new to the field.

## Textbook review

56. More beginner-friendly vs more advanced texts, with focus on open source texts (o) and code examples (c) or Jupyter notebooks (j). Top picks marked with (\*).

Subject Area	Beginner-friendly (undergraduate)	Advanced (graduate)	Reference or other texts
Math prerequisites	Deisenroth* (o, 2020)		
Probability	Blitzstein and Hwang *		Wentzel (1982)
Statistics		Wasserman* (2013) Casella/Berger	
Econometrics	Kennedy* Stock/Watson		Green
Time series			Hamilton
Machine Learning	ISLR/ISLP* (o,c)	PRML by Bishop (o) PML by Murphy (o)	ESL (o) Vapnik (1998)
Deep Learning	Andrew Ng*	UDL(o,j), d2l (o,j), DLB(o)	
Artificial Intelligence			Russel/Norvig (2020)

## Part 2. Data types, sources and quality

### Tabular data

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

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

## Not just numbers: text, images and sound

59. Text as vector. Natural language processing. Deep learning in NLP (ChatGPT).

Textbooks

Jurafsky:

- [Speech and Language Processing by Jurafsky and James H. Martin](#)
- [Big Ideas: Natural Language Processing with MacArthur Fellow Dan Jur...](#)
- [Programming and written exercises](#)

Manning et al (2008) [Introduction to Information Retrieval](#).

[Lewis Tunstall, Leandro von Werra, Thomas Wolf \(2022\). Natural Language Processing with Transformers.](#)

Supplementary:

Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax by Emily M. Bender (2013, paywalled, see [TOC](#))

Courses:

[Lena Voita NLP Course "For You"](#)

Few articles (via Ilya Gusev):

- Word2Vec: Mikolov et al., Efficient Estimation of Word Representations in Vector Space <https://arxiv.org/pdf/1301.3781.pdf>
- FastText: Bojanowski et al., Enriching Word Vectors with Subword Information <https://arxiv.org/pdf/1607.04606.pdf>
- Attention: Bahdanau et al., Neural Machine Translation by Jointly Learning to Align and Translate: <https://arxiv.org/abs/1409.0473>
- Transformers: Vaswani et al., Attention Is All You Need <https://arxiv.org/abs/1706.03762>
- BERT: Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding <https://arxiv.org/abs/1810.0480>

Libraries:

- huggingface
- spacy
- nltk

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

61. Sound, noise, music. Waves.

## Data in business and economic perspective

62. Big vs small data.

Article: [Hal Varian \(2014\). Big Data: New Tricks for Econometrics](#)

Report: [Matthew Harding and Jonathan Hersh \(2018\). Big Data in Economics.](#)

Course: [Melissa Dell and Matthew Harding \(2023\). Machine Learning and Big Data.](#)

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

[Data Governance Is A Top Priority For 65% Of Data Leaders \(Gartner by Atlan\)](#)

## Part 3. From research design to model productisation

### Steps in analysis

64. 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](#))

65. Data visualization. Dashboards.

Economist: Mistakes, We Have Drawn a Few.

Grammar of Graphics (ggplot)

66. Business hypothesis and ways to test it. Business outcomes.

- Interacting with business or retail customer
- Controlling own system or business processes.
- Supply chain, sourcing, inputs control.
- Reporting, regulation, compliance, audit.

Advanced analytics and dashboards loop vs automation of control loop.

67. Analysis as a DAG.

Quote from [drivendata cookiecutter](#) as a starter (but not universal) project template.

68. Reproducible research and reproducibility crisis.

# ML in production

## 69. Responsibilities of a data engineer, data scientist or modeler, machine learning engineer, business analyst and other roles (notes).

Perfect world (bam is the sound like “bang”, popularized by StatQuest videos, an “aha moment”).

Idea what to improve -> Good data -> Known model -> Quick inference -> Bam to production -> Bam cool and unambiguous effect -> Bam business liked it

Practically perfect world:

Many ideas -> Business and modelling hypothesis -> Expected result -> Where to deploy? -> Is there a model for this? Which do we pick? -> Is there data for this? -> Can we estimate/train the model? -> Does it seem to work? -> Can we deploy? -> How is it doing in production? -> Can we improve? -> Is business happy? -> How long would the result persist?

Pick roles who is in charge of what:

- A full-stack “data scientist”
- A modeller
- Data engineer / Data architect
- Machine learning engineer / ML platform engineer
- Software engineers
- Research scientist
- Business analyst
- Business lead or product manager
- Vendor/consultant/ChatGPT

When does the data pipeline become a “product”?

- a) any data pipeline that worked and delivered the business result, however trivial from modeling viewpoint;
- b) a complex system that is not just model+prediction (frontend, hardware, business rule change, etc);
- c) anything that a business or end user wants and has a bit of data or intelligence in it;
- d) something you can sell as a solution in a specific industry.

What can go wrong in a data pipeline? How are companies different with respect to data and machine learning? What are the most aggressive promises about AI in a specific industry? Why is this not happening yet?

Broader perspective:

- system and actors vs control system and desired outcomes
- customer/asset life cycle, person-centricity
- business intelligence, business value and margins
- silos, vested interest, delegation of control/responsibility
- change of business models and company boundary
- immediate ROI, long-term sustainability, business valuation
- data as representation of objects, processes, behaviors

## 70. ML pipelines and roles as told by the companies and experts.

- ITMO University Role model (very detailed roles at the picture):  
<https://github.com/aimclub/ai-competency-model/>

- Exercise: spot the paragraph wrongly placed in a guide  
<https://learn.microsoft.com/en-us/training/modules/leverage-ai-tools/6-understand-machine-learning-lifecycle>
- Exercise: which part do you think is most important in a working data pipeline (extracted from Demystifying AI for the Enterprise book).
  - Ask a Specific Question
  - Start Simple
  - Try Many Algorithms
  - Treat Your Data with Suspicion
  - Normalize Your Inputs
  - Validate Your Model
  - Ensure the Quality of Your Training Data
  - Set Up a Feedback Loop
  - Don't Trust Black Boxes
  - Correlation Is Not Causation
  - Monitor Ongoing Performance
  - Keep Track Of Your Model Changes
  - Don't be Fooled by "Accuracy"

71. Machine learning pipelines and MLOps.

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

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

## Part 4. Software tools

### Programming languages and statistical software

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

R for statistical packages and classic statistics.

Python for statistics (stamodules), machine learning (scikit-learn) and deep learning (PyTorch, TensorFlow, keras).

Differentiation and composability in Julia and JAX.

Exercise: Compare a tabular dataframe implementation in R, Python (pandas or polars) and Julia.

74. Statistical software, proprietary vs open source and documentation.

Who won in the statistics market: 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
- JASP and jamovi

75. Notebooks vs plain files and packages. Version control. Refactoring and cleaner code.

[The Missing Semester of Your CS Education](#) (ignore metaprogramming chapter).

76. Extra topic: open source viability and funding models

VC-driven: streamlit  
Sponsored: PyWhy  
Burnout: curl

## Databases and storage

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

78. 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>

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

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

81. 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.

82. 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 and data engineering tools

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

Andriy Burukov MLOps book and Data Engineering Zoomcamp.

See [MAD@firstmark.com](https://mad@firstmark.com) company landscape.

## Part 5. Business change and society impact

Making money vs making change.

### Technology companies as machine learning market players

84. Companies that sell ML tools and solutions, their valuations and strategy.

Low code: H2O, DataRobot.

85. Cloud providers.

86. Hardware providers.

87. Internet-scale data owners.

### Adoption in broad economy and society

88. Fairness, biases, equity, human loop.

89. Economics, cost and payoffs of applying ML. Business value of ML.

90. Job market: in-house data modeller, consultant or a vendor?

- 91. Who's got more data? Data privacy and data protection. Markets and pricing of data.
- 92. What gets to be regulated. Does a national AI or data strategy make sense?
- 93. Why the hype: what makes the corporation play hype and overpromise? Why do investors buy that?

## Selected industry domains and use cases

- 94. Social sciences (sociology, political science, psychology, anthropology).
- 95. Recommender systems (RecSys).
- 96. Clinical trials.
- 97. Quality control and dependability.
- 98. ML in finance

[Halperin textbook.](#)

"The majority of our findings are kept proprietary. From time to time, however, we decide to publish some of them..." (EP: based on what?)

- 99. Discrete or continuous industrial processes.

## Appendix

### Interviews

**randomlyCoding:** Head of AI at a startup and have been working in the field for over a decade. I certainly don't know everything, but I like to get my feet wet and touch on anything I find interesting. I've trained ML models to do all sorts of tasks and will likely have at least heard of most things.

- 100. Can one summarise a production pipeline as the following: choosing a business and then a modelling hypothesis - dataset - model selection - training - validation - model rollout - business metrics? What are the weak links in this process and where a pipeline may break?

**randomlyCoding:** In general that's about on point. I'd say there's certainly a lot more recursion (eg. you might pick a dataset, build a model and train it, only to realise you've massively overfitting because you don't have enough data; thus you go looking for a bigger/additional dataset). Weak links often occur at either end of the process - you pick a dataset that isn't suited to your problem and thus end up with a solution that solves a

problem you weren't trying to solve or the model is 100% perfect but the business case requires inference to happen in real time and it takes 20 minutes based on the size of the model. I've also seen cases of trying to extend a model to do more than it was initially designed for; this isn't always a bad idea but if the person leading this doesn't understand the underlying model there can often be misalignment between their expectations and reality.

101. What skills would you expect an ML engineer (MLE) to know? How can an decent econometrician upgrade to an MLE?

**randomlyCoding:** I would expect any ML engineer to know one of 3 python packages that are the core of most ML processes (either pytorch, tensorflow, keras), but on top of that I'd expect familiarity with some domain specific packages, that might be NLTK if you're working on natural language processing; it might be scikit-learn if you're looking at random forests. One thing I would say is usually a *\*must\** is familiarity with Linux and a cloud provider (AWS, GCP, Azure). You don't need to know all 3 cloud providers (pick AWS if you don't know any yet - it has 50% market share) but if you don't know any of them it'll be harder to on board you and your first few weeks would be a lot more overwhelming - even knowing a different one to the one you use at a specific job will help as they all have similar functionality.

102. Who puts and ML model into production? You got the weights after training, validation stage passed ok, then it always becomes a small API? Who wraps a notebook into API, a designated engineer?

**randomlyCoding:** Who puts the ML model into production can vary depending on the system in use, it's often an API but not always. I would expect any ML engineer to at least be able to put together a notebook (or similar) that can be used to run inference on the model; in some cases if the organisation is small enough it will be someone who has directly worked on the model; in other cases they may be using a specific orchestration packages that abstracts away this process; in yet more cases it could be hidden behind a message broker. Obviously not all ML models need to be hosted all the time, some are run periodically (thus they might not require anything more than ingesting a CSV file into a single python script).

103. Does a full-stack data scientist role still exist?

**randomlyCoding:** I think the full-stack data scientist role does still exist, it will always exists as long as there are start-ups that have limited budgets and big ideas. If you're in a larger team your remit will often be constrained to a specific task, but depending on the organisation your within that task could change regularly (eg. today you're handling data ingestion because the model we're working on is a transformer and you don't have much experience with transformers, but tomorrow we're building a reinforcement learning system and you're the team's expert in RL). In most teams I'd expect the architect of the model to also do a fair amount of the modelling itself; anyone doing modelling will have to work closing with the data engineers, etc. I think this mean the roles aren't as well defined as in SWE and I think this is because there's a lot of trial and error in ML so it's not as simple as for example ingest the data and pass the process on.

104. What is the most amazing thing in modelling or data analysis you thought would not work but it did?

**randomlyCoding**: Diffusion feels like it shouldn't work. In general it's a multistep process of removing noise from an image until you end up with the image without any noise; but to do that you start with a completely noisy image and then predict a small percentage of the noise that was added (the previous step of noise added) and then subtract that noise from the image. The maths behind it is reasonably simple, but it just feels like it shouldn't work!

## More personal skills

105. Common sense, logic and critical reasoning.
106. Writing well, explaining, inquiring, communicating.

[On Writing Well: An Informal Guide to Writing Nonfiction](#). On Writing Well and keeping it up-to-date for 35 years by William Zinsse in American Scholar.

## Personas

107. 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](#) (PGP)

108. Hall of shame: story of Siraj Raval ([plagiarism in education](#)).

## Glossary

109. Common terms, professional slang and buzzwords.

Common terms:

- Supervised vs unsupervised vs semisupervised learning.

- Structured vs unstructured data.

Professional slang:

- Feature engineering (variable selection and transform).
- ETL vs ELT (data ingestion)

Fading buzzwords:

- Data mining
- Big data

## Other resources

110. Courses, syllabuses and excercises

- [CIS 4190/5190: Applied Machine Learning \(Spring 2023\)](#) – great list of resources.
- <https://deepmleet.streamlit.app> iams to be the leetcode of machine learning.

111. Video series

- [StatQuest with Josh Starmer](#) (This man is a genius.)
- [3Blue1Brown](#) by Grant Sanderson. (Very high quality content!)
- [Machine Learning Street Talk](#) (Suggested by reader: “Sometimes bit too dense for absolute beginners but really good. They list resources, papers, books.”)

## Changelog – timeline of this document

### **v0.7.0 (April 29, 2024):**

- Excellent undergrad econometrics course [Mathematical Econometrics I by Roth and Hall](#).
- Updated econometrics vs machine learning section with papers and courses from Hal Varian and Susan Athley.
- Classic probability textbooks added (Blitzstein and Hwang, Wackerly, Mood) along with modern free websites.
- Links from Econometric Navigator for time series, calculus and linear algebra.
- Total count is 173 links and 111 topics.

### **v0.6.2:**

- Discussion ([r/MachineLearning](#)): [Thoughts on Hands-On Machine Learning with Scikit-Learn, Keras & Tensorflow by Geron](#)

### **v0.6.1:**

- Interview with u/randomlyCoding/
- [Using Julia for Introductory Statistics](#) by John Verzani

### v0.6.0


General interest:

- Overview of metrics in JEL Section C for and AMS classification for math papers.
- Annotated History of Modern AI and Deep Learning by Jürgen Schmidhuber.

Niche:

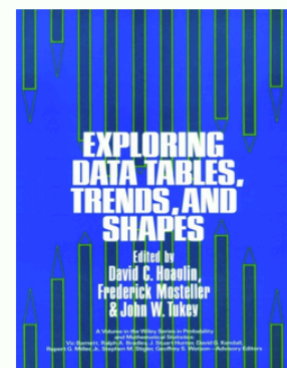
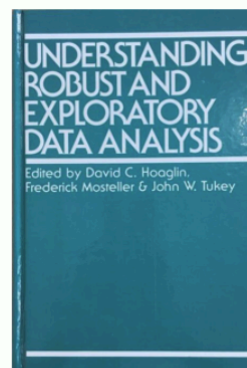
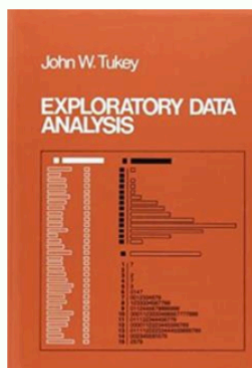
- Bodhisattva Sen and Larry Wasserman on nonparametric statistics.
- Informal review on hypothesis testing (online book appendix chapter).

### v0.5.5:

-  [Big Ideas: Natural Language Processing with MacArthur Fellow Dan Jurafsky](#) (interview, beginner-friendly) and [Lena Voita NLP Course | For You](#)
- [scikit-learn: machine learning in Python by Gael Varoquaux](#). Part of Scientific Python Lectures. One document to learn numerics, science, and data with Python. I think an underappreciated resource (beginner, but programming knowledge required).
- Andrew Ng [lecture notes](#) on machine learning from a 2022 course. EP: Andrew Ng best known for a deep learning course, but the classic machine learning notes are very well structured (intermediate).
- [Deisenroth et al. \(2020\). Mathematics for Machine Learning](#). Chapter 8 “When Models Meet Data” is an accessible introduction to statistical learning (very beginner friendly).
- [Shawe-Taylor \(2023\). Statistical Learning Theory for Modern Machine Learning](#). [has video and slides](#) (advanced).
- [Causal ML Book by Chernozhukov et al. \(2024\)](#) (advanced).

**April 9, 2024**

Added 3x2 table on title page with key topics. Also removed few images.



### April 6, 2024 (88 topics)

- Finalised Databases and storage.
- Edited SEMs and references from Paul Goldsmith-Pinkham
- Marked for review DAG and ML project flow.
- To add next 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](#).

## Quotes and reader 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.*



### Guide roadmap

Models and methods	Pipelines	Tools
Stats and econometrics	Descriptive analysis	Programming languages
Machine learning (ML)	Task design and outcomes	ML and DL libraries
Deep learning (DL)	ML in production	Databases, DE and MLOps
Other methods	Reproducible research	Infrastructure for ML

Model evaluation		
Data	Players and impacts	Economics
Types of data	Technology companies	Cost of (not) doing ML
Sources and ownership	Non-tech business	Markets for data
Data quality and DG	The human user	Rationale for regulation
NLP, CV and Robotics	Society impacts	

ML = machine learning, DE = data engineering, MLOps = devops for ML, DG = data governance

Link to this document: <https://t.ly/RcA2Q>.

Editor mode:  MLMW: Machine Learning My Way and  Prose for MLMW .