Learning analytics and data science

D. Semenovich

- What is "data science"?
- How about "analytics"?
- What are some of the good ways to (independently) learn about analytics and data science?
- A singularly boring presentation template.

What is data science?

Increasing number of industries are affected by "digital transformation" – new business models facilitated by the ubiquitous availability of computing and telecommunication technologies. This "digital transformation" is to a large degree carried out by engineering / software centric "web" companies.

"Data science" in its contemporary usage has originally meant simply "analytics at web companies". The particular setting impacts how the analytics are produced and consumed — e.g. greater degree of automation due to volumes of data and timescales to generate results — but, perhaps, not the underlying concepts.

"Data science" has since evolved (in no small part through aspirations towards "digital transformation") into a vague umbrella term for a very wide range of analytics activities in many industries.

Let's try to define "analytics"

"Analytics" is the process where some manner of systematic measurements and observations are used as a presumably objective basis for making decisions, process improvements or optimising (more or less abstract) performance metrics.

The same theme is repeated with variations across a wide gamut of loosely associated disciplines. Exact definitions have proven notoriously elusive (consider operations research or actuarial science), with few convincing distinctions from the classical notions of rational or scientific inquiry other than through the domain of application.

Analytics landscape is not easy to survey as many movements have overlapped in time and areas of practice, resulting in a layering of ideologies and ideas perhaps impossible to decisively disentangle.

THE

DOCTRINE

OF

CHANCES:

OR,

A METHOD of Calculating the Probabilities of Events in PLAY.

THE THIRD EDITION,
Fuller, Clearer, and more Correct than the Former.

By A. DE MOIVRE,
Fellow of the ROYAL SOCIETY, *and Member of the* ROYAL ACADEMIES OF SCIENCES *of* Berlin *and* Paris.

LONDON:
Printed for A. MILLAR, in the *Strand*.
MDCCLVI.

ANNUITIES

UPON

LIVES:

OR,

The VALUATION of ANNUITIES upon any Number of LIVES; as alſo, of REVERSIONS.

To which is added,

An APPENDIX concerning the EXPECTATIONS of LIFE, and Probabilities of SURVIVORSHIP.

By A. DE MOIVRE. F. R. S.

LONDON,
Printed by *W. P.* and ſold by *Francis Fayram*, at the *South-Entrance* of the *Royal Exchange*; and *Benj. Motte*, at the *Middle Temple Gate, Fleetſtreet*; and *W. Pearſon*, Printer, over-againſt *Wright's-Coffee-Houſe, Alderſgate-Street*. MDCCXXV.

7.

Related fields

Online advertising and website optimisation – online advertising has grown into a massive ecosystem over the last two decades providing revenue for the majority of online services. The nature of the medium is eminently accommodating of tracking and analytics, resulting in one of the more dramatic applications of data science.

Most sophisticated solutions (e.g. Adwords) are deployed by inventory providers and aggregators, such as Google and Facebook. Live A/B testing is also prevalent among online businesses something which is still a rarity in traditional enterprise. This is at present the biggest area of employment for data scientists.

Manufacturing quality control, statistical process control, lean manufacturing, six sigma – this is an area of analytics activity supporting manufacturing activities and has been progressively developed since at

least 1930s. Among the main objectives is monitoring and elimination of variability in manufacturing processes (e.g. part dimensions), ensuring that defect rates are thereby controlled.

Scientific management, management consulting, management accounting – the broad idea of application of "scientific" principles to industrial management is well over 100 years old ubiquitous proliferation of plans, budgets, KPIs and management reports is part of this tradition. Very little attention is usually paid to natural statistical variation in many metrics. Also while many controlled experimental studies have been documented in this setting they have never become a part of the standard methodology.

Operations research, industrial engineering, revenue management, mathematical optimisation, management science. Operations research began as scientific study of military operations (e.g. convoy composition,

bomber interception protocols, logistics) during the Second World War and the principles have been exported to many other industries in the following years. Main tools include mathematical optimisation, stochastic processes. Mathematical optimisation is a somewhat standalone area of research with many close connections to both computer science and statistics.

Statistics, — really requires no introduction, perhaps the main focus of interest in applications has been analysis of government data, polls and surveys and support for evaluation of experimental results in life sciences and medicine.

Applied finance, financial engineering, algorithmic trading, HFT, risk management. There are close parallels between data science and quantitative finance in the 1980s and 1990s. This is not surprising, because in market execution is a key part of any model driven trading strategy, placing a premium on "hacking skills". At present it is perhaps reasonable to view majority of "data scientists" as "quants" of digital advertising.

Engineering control, control theory, signal processing From fly-by-wire systems to cellular networks and synthetic aperture radar. Much less ambitious in scope than AI, but this systems work reliably and are by now absolutely ubiquitous.

Econometrics, mechanism design, causal inference (from observational data) due to the difficulty and costs of real world experiments in economics, econometricians have developed tools and conceptual frameworks for causal inference with observational data. Furthermore mechanism design and the study of auctions have had significant impact on the design of online marketplaces.

Business intelligence, database / warehouse design, dashboards business intelligence is primarily IT vendor driven activity to support management reporting in traditional enterprise with perhaps strongest intellectual links to the academic databases research community. Traditionally little

attention has been paid to statistical aspects of business intelligence or how it is to be actioned.

Machine learning, natural language processing, computer vision, data mining machine learning is a branch of computer science that initially focused on more tractable aspects of artificial intelligence, primarily by constructing models from example data using statistical methods rather than designing them by hand from general principles. Two large application areas are computer vision and natural language processing, including machine translation.

By now the differences between theoretical machine learning and statistics communities is largely superficial, amounting to little more than preferences for different styles of analysis of statistical procedures. Machine learning research has also provided many of the tools used in analytics for online advertising and algorithmic trading. Data mining has originated

from the databases research community and by now has mostly converged with machine learning in terms of both objectives and methodologies.

Analytics (self) education

Mathematical modelling and computing could be argued to be core "data science" skills — increasing number of business processes and low-level operational decisions being subject to automation.

Here I tried to stay away from "flavour of the month" or introductory "data science" offerings that by now unfortunately dominate, focussing instead on those courses where the professors, at times, provide unique perspective on fundamental topics across mathematical modelling and computing.

It is generally a useful heuristic to seek out graduate level courses from reputable North American universities. Video or audio recordings of good lectures can significantly lower the (still considerable) effort required to become familiar with the material compared to self study from textbooks.

Courses: Analytics at web companies

To get an impression of what the future of insurance analytics might look like, it is worthwhile to review some of the courses offered by people with experience implementing analytics solutions for the leading web companies.

Examples include CS281B "Scalable Machine Learning"[1] at UC Berkeley by Alex Smola (formerly of Yahoo) and "Big Data, Large Scale Machine Learning"[2] at NYU by Yan LeCunn (currently at Facebook). In particular the first course offers an interesting insight into the importance of understanding systems, numerical methods and statistics to develop analytics solutions at web scale.

Prerequisites for this material include linear algebra, basic probability and statistics and, ideally, convex optimisation and an introduction to machine learning, as discussed next.

[1] http://www.youtube.com/playlist?list=PLOxR6w3fIHWzljtDh7jKSx_cuSxEOCayP
[2] http://cilvr.cs.nyu.edu/doku.php?id=courses:bigdata:start

Courses: Linear algebra and numerical computing

Numerical linear algebra is the most essential tool in applied mathematics. The majority of computational procedures for solving mathematical models ultimately reduce to iteratively solving systems of linear equations.

An excellent introductory treatment of linear algebra is given by Gilbert Strang in MIT 18.06[3] and 18.086[4], demonstrating a very broad range of applications across engineering subfields.

Another take on the material is given in Stanford EE263[5] taught by Stephen Boyd — in addition to basic linear algebra, the course gives highly intuitive exposition to least squares regression, regularisation, singular value decomposition and linear dynamical systems (which can be viewed as

[3] http://ocw.mit.edu/courses/mathematics/18-06-linear-algebra-spring-2010/
[4] http://ocw.mit.edu/courses/mathematics/18-086-mathematical-methods-for-engineers-ii-spring-2006/
[5] https://see.stanford.edu/Course/EE263

a generalisation of a wide class of time-series models in the actuarial syllabus).

The Fourier transform is one of the most famous special cases of a linear operation — an intuitive introduction to the subject and its multitude of applications, including the Central Limit Theorem, is given in Stanford EE261[6].

[6] https://see.stanford.edu/Course/EE261

Courses: Optimisation

Optimisation based models are pervasive in analytics, whether it be maximum likelihood estimation, "empirical risk minimisation", Neyman-Pearson hypothesis testing, optimal control, Markowitz portfolio theory or option pricing.

Prof. Stephen Boyd's course EE364A Convex Optimization[7] not only gives a solid grounding in the theory but also considers many of the above-mentioned examples. Convex optimisation is widely seen as the foundation of modern statistics, machine learning and signal processing.

There is also an interesting connection between mathematical optimisation and classical algorithms studied in undergraduate computer science courses — many of the problems such as sorting, shortest path, max flow etc turn

[7] https://class.stanford.edu/courses/Engineering/CVX101/Winter2014/about

out to be special cases of linear programming (itself a special case of convex optimisation).

The follow up course EE364B[8] provides more detailed background on scalable and distributed optimization as well as the clearest introduction to the General Equilibrium theory of microeconomics you are likely to find.

[8] https://see.stanford.edu/Course/EE364B

Courses: Machine learning, information theory etc.

There are few unequivocally great introductory probability and statistics courses publically available, at least at the moment. MIT 6.041[9] is a useful probability refresher. A worthwhile follow up is MIT 6.262 "Discrete Stochastic Processes"[10].

When it comes to statistics, or at least a take on the topic that is more attuned to analytics applications, Stanford Statistical Learning[11] is a solid introduction from the authors of the well-known book.

A closely related subject area is machine learning, with the introductory course by Andrew Ng[12] and a much more in depth treatment by Alex

[9] http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-041sc-probabilistic-systems-analysis-and-applied-probability-fall-2013/
[10] http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-262-discrete-stochastic-processes-spring-2011/
[11] https://class.stanford.edu/courses/HumanitiesScience/StatLearning/Winter2014/about
[12] https://see.stanford.edu/Course/CS229

Smola[13]. So called deep networks[14] are a recent hot topic in machine learning, providing state of the art performance for many recognition tasks.

Information theory provides perhaps one of the most successful and widely used applications of probability. There are also important connections to statistics and machine learning (as efficient compression requires effective conditional probability estimation). MIT 6.450 "Principles of Digital Commucations I"[15] is an excellent course. Information theory is an essential foundation of all digital information processing technology.

Another superb discussion of information theory is given in the course taught by the late David MacKay at Cambridge[16], bringing together topics from coding theory, statistics and machine learning.

[13] http://alex.smola.org/teaching/cmu2013-10-701/
[14] http://cilvr.cs.nyu.edu/doku.php?id=courses:deeplearning:start
[15] http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-450-principles-of-digital-communications-i-fall-2006/
[16] http://www.inference.phy.cam.ac.uk/itprnn/Videos.shtml

Courses: Programming

There exists a very wide range of high quality introductory programming courses. Perhaps the Stanford sequence[17,18] deserves a particular mention. Alternatives include the introductory courses at MIT[19,20].

MIT 6.001[21] is the most celebrated introductory programming course of all, with the textbook "Structure and Interpretation of Computer Programs" used in dozens of top universities.

While Scheme, the language that it uses for teaching programming concepts, has for long time been considered less than practical, over the recent years there has been a dramatic resurgence of popularity of the

[17] https://see.stanford.edu/Course/CS106A
[18] https://see.stanford.edu/Course/CS106B
[19] http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-00sc-introduction-to-computer-science-and-programming-spring-2011/
[20] http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-01sc-introduction-to-electrical-engineering-and-computer-science-i-spring-2011/
[21] http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-001-structure-and-interpretation-of-computer-programs-spring-2005/

related body of ideas called functional programming, underpinning many of the latest "big data" technologies.

Beyond the introductory courses, "Programming Paradigms"[22] gives a useful overview of design choices behind a variety of programming languages. Finally, no such list would be complete without an algorithms class[23].

[22] https://see.stanford.edu/Course/CS107

[23] http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-006-introduction-to-algorithms-fall-2011/

Courses: Finance, Economics and social science

While the exact relation between actuarial pricing and financial economics is not clearly set out in popular introductory textbooks, it has been understood in the academic literature for some time as the so called incomplete markets setting.

An introductory discussion of the modern theory of finance (CAPM, option pricing etc) from this more advanced point of view is given in John Cochrane's (University of Chicago) class Asset Pricing on Coursera[24].

A useful generalisation of the concept of an optimisation problem (see e.g. Stanford EE364A) is offered by game theory. Instead of considering a central planning problem where all the decisions are taken by a single agent, game theory looks at situations where there are multiple

[24] http://www.coursera.org/course/assetpricing

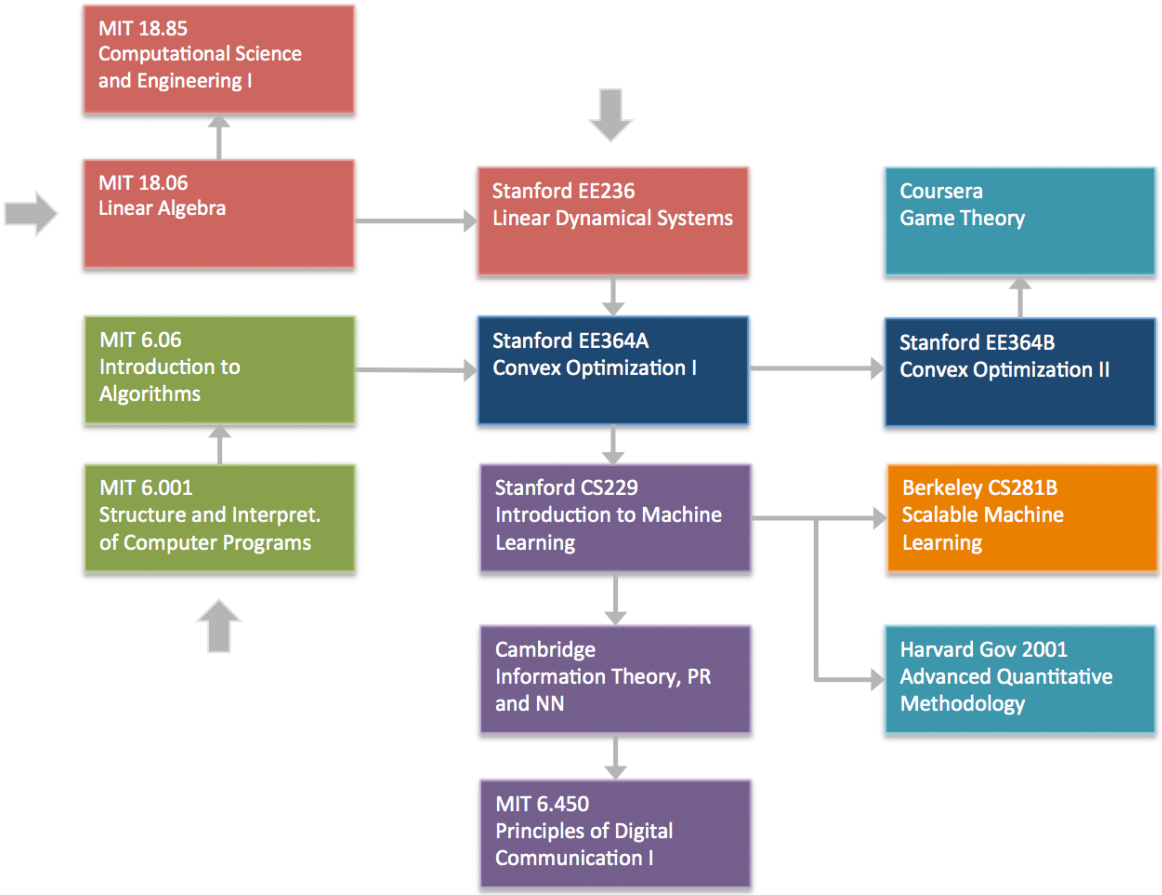
self-interested parties involved. Coursera classes[25,26] provide an introduction to a range of topics, including auctions and mechanism design.

Applications of game theoretic methods to the study of social insurance, optimal taxation and related ideas are given in the Harvard course Public Economics[27].

One example in social science where large-scale experiments have been possible is development economics. The MIT course 14.73[28] offers an in depth discussion of considerations that go into designing a convincing experimental study.

[25] https://www.coursera.org/course/gametheory
[26] https://www.coursera.org/course/gametheory2
[27] http://obs.rc.fas.harvard.edu/chetty/public_lecs.html
[28] http://ocw.mit.edu/courses/economics/14-73-the-challenge-of-world-poverty-spring-2011/

A possible curriculum



But will this be useful to me?

"You have to keep a dozen of your favorite problems constantly present in your mind, although by and large they will lay in a dormant state. Every time you hear or read a new trick or a new result, test it against each of your twelve problems to see whether it helps. Every once in a while there will be a hit, and people will say: 'How did he do it? He must be a genius!'"

R. Feynman via G-C. Rota