Introduction to Data Science

End to end data science

Dina Machuve

31 May 2018

Data Science Africa 2018 Nyeri, Kenya

What is Data Science?

What is Data Science?

Hal Varian (2009)
The sexy job in the next ten years will be statisticians. The ability to take data– to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it–thats going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids.

Hal Varian on how the Web challenges managers

What is Data Science?

Mike Driscoll (2009)
I believe that the folks to whom Hal Varian is referring are not statisticians in the narrow sense, but rather people who possess skills in three key, yet independent areas: statistics, data munging, and data visualization.

The Three Sexy Skills of Data Geeks

What is Data Science?

Mike Loukides (2010)
Data science enables the creation of data products.
Whether that data is search terms, voice samples, or product reviews, the users are in a feedback loop in which they contribute to the products they use. Thats the beginning of data science.

What is data science?
O'Reilly Radar

What is
Data Science?

The future belongs to the companies
and people that turn data into products

What is Data Science?

Jeff Leek (2013)
The issue is that the hype around big data/data science will flame out (it already is) if data science is only about ‘data’ and not about ‘science’. The long term impact of data science will be measured by the scientific questions we can answer with the data.

The key word in “Data Science” is not Data, it is Science

What is Data Science?

Neil Lawrence (2017)

We define the field of data science to be the challenge of making sense of the large volumes of data that have now become available through the increase in sensors and the large interconnection of the internet. Phenomena variously known as "big data" or "the internet of things". Data science differs from traditional statistics in that this data is not necessarily collected with a purpose or experiment in mind. It is collected by happenstance, and we try and extract value from it later.

What is Machine Learning?

What is Data Science?

What question are you trying to answer with data?

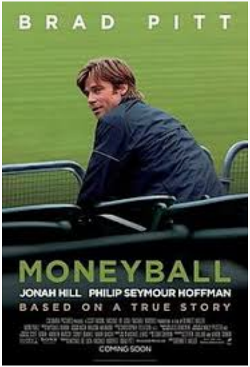
What is Data Science?

What question are you trying to answer with data?

http://bit.ly/1HXvKdW

Data Science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data and communicating the answer to the question to a relevant audience

Moneyball



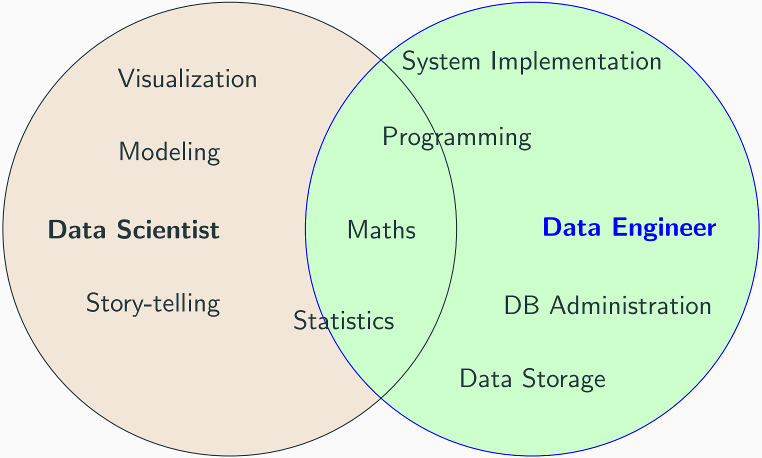
https://en.wikipedia.org/wiki/Moneyball_(film)

Can we build a winning baseball team with a really limited budget?

Trade-Offs in Data Science

Data Scientist vs Data Engineer

http://bit.ly/1XT5nz2



Visualization

Modeling

Data Scientist

Story-telling

System Implementation

Programming

Maths

Statistics

Data Engineer

DB Administration

Data Storage

Trade-offs

https://oreil.ly/2dFGmIG

The “Best” Machine Learning Method

Interpretable

Simple

Accurate

Fast (to train and test)

Scalable

The Data Scientist Challenge

Value of big data in life sciences

- Novel discoveries for healthcare, agriculture, food security (over 1.2 million species of plants & animals)
- Disease surveillance and response
- Management of health data (EHRs and experimental data) can inform diagnosis and treatment Precision Medicine
- Challenges: volume, velocity, variety

Statistics

Statistics

Statistics is the discipline of analyzing data. It intersects heavily with data science, machine learning and, of course, traditional statistical analysis.

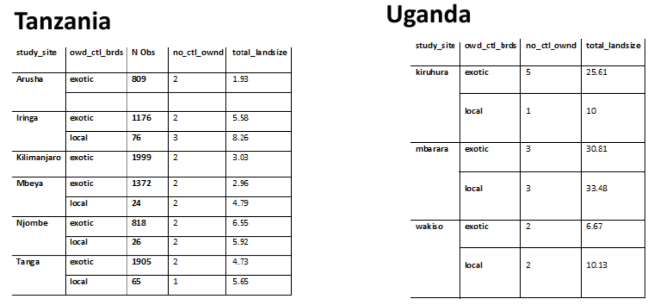
Key activities that define the field:

1. Descriptive statistics (EDA, quantification, summarization, clustering)
2. Inference (estimation, sampling, variability, defining populations)
3. Prediction (machine learning, supervised learning)
4. Experimental Design (the process of designing experiments)

Descriptive statistics

PEARL Project (2017)

Descriptive statistics on livestock dynamics for small scale farmers in Tanzania and Uganda



Tanzania

study_site	owd_ctl_brds	N Obs	no_ctl_ownd	total_landsize
Arusha	exotic	809	2	1.93
Iringa	exotic	1176	2	5.58
	local	76	3	8.26
Kilimanjaro	exotic	1999	2	3.03
Mbeya	exotic	1372	2	2.96
	local	24	2	4.79
Njombe	exotic	818	2	6.55
	local	26	2	5.92
Tanga	exotic	1905	2	4.73
	local	65	1	5.65

Uganda

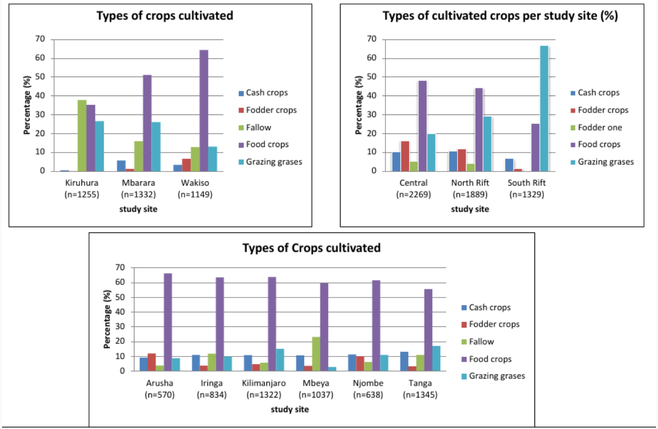
study_site	owd_ctl_brds	no_ctl_ownd	total_landsize
kiruhura	exotic	5	25.61
	local	1	10
mbarara	exotic	3	30.81
	local	3	33.48
wakiso	exotic	2	6.67
	local	2	10.13

- Land size not proportional to number of cattle owned by farmers
- Farmers are engaged in other activities other than dairy

Descriptive statistics

PEARL Project (2017)

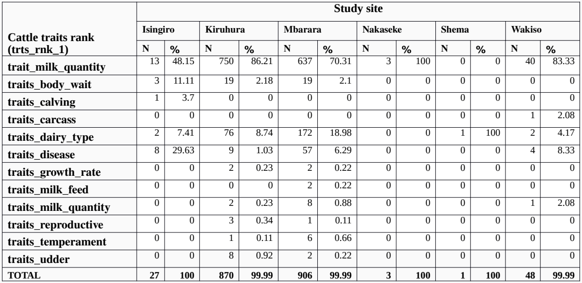
Descriptive statistics on crops cultivated for small scale farmers in Kenya, Tanzania and Uganda



Descriptive statistics

PEARL Project (2017)

Descriptive statistics on traits of dairy cows for large scale farmers in Uganda



Cattle traits rank (trts_rnk_1)	Study site											
	Isingiro		Kiruhura		Mbarara		Nakaseke		Shema		Wakiso	
	N	%	N	%	N	%	N	%	N	%	N	%
trait_milk_quantity	13	48.15	750	86.21	637	70.31	3	100	0	0	40	83.33
traits_body_wait	3	11.11	19	2.18	19	2.1	0	0	0	0	0	0
traits_calving	1	3.7	0	0	0	0	0	0	0	0	0	0
traits_carcass	0	0	0	0	0	0	0	0	0	0	1	2.08
traits_dairy_type	2	7.41	76	8.74	172	18.98	0	0	1	100	2	4.17
traits_disease	8	29.63	9	1.03	57	6.29	0	0	0	0	4	8.33
traits_growth_rate	0	0	2	0.23	2	0.22	0	0	0	0	0	0
traits_milk_feed	0	0	0	0	2	0.22	0	0	0	0	0	0
traits_milk_quantity	0	0	2	0.23	8	0.88	0	0	0	0	1	2.08
traits_reproductive	0	0	3	0.34	1	0.11	0	0	0	0	0	0
traits_temperament	0	0	1	0.11	6	0.66	0	0	0	0	0	0
traits_udder	0	0	8	0.92	2	0.22	0	0	0	0	0	0
TOTAL	**27**	**100**	**870**	**99.99**	**906**	**99.99**	**3**	**100**	**1**	**100**	**48**	**99.99**

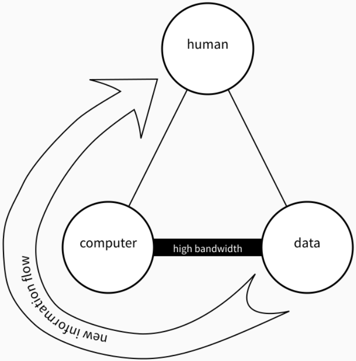
Machine Learning

What is Machine Learning?

Neil Lawrence (2017)
Machine learning is the principle technology underpinning the recent advances in artificial intelligence. Machine learning is perhaps the principal technology behind two emerging domains: data science and artificial intelligence. The rise of machine learning is coming about through the availability of data and computation, but machine learning methdologies are fundamentally dependent on models.

$$\textit{data} + \textit{model} + \textit{compute} \longrightarrow \textit{prediction}$$

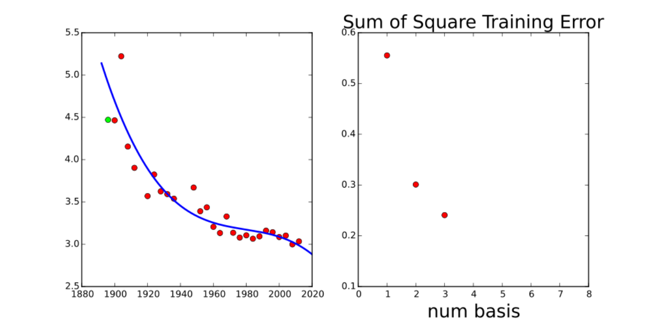
What is Machine Learning?



Large amounts of data and high interconnection bandwidth $\implies$ much of our information about the world around us received through computers

Example of prediction

The Olympic gold medalist in the marathons pace is predicted using a regression fit. In this case the mathematical function is directly predicting the pace of the winner as a function of the year of the Olympics.



Machine Learning Approaches

Neil Lawrence (2017)
Machine learning takes the approach of observing a system in practice and emulating its behavior with mathematics. One of the design aspects in designing machine learning solutions is where to put the mathematical function. Obtaining complex behavior in the resulting system can require some imagination in the design process.

The Machine Learning classical approaches:

- supervised learning
- unsupervised learning
- reinforcement learning

Machine Learning Approaches

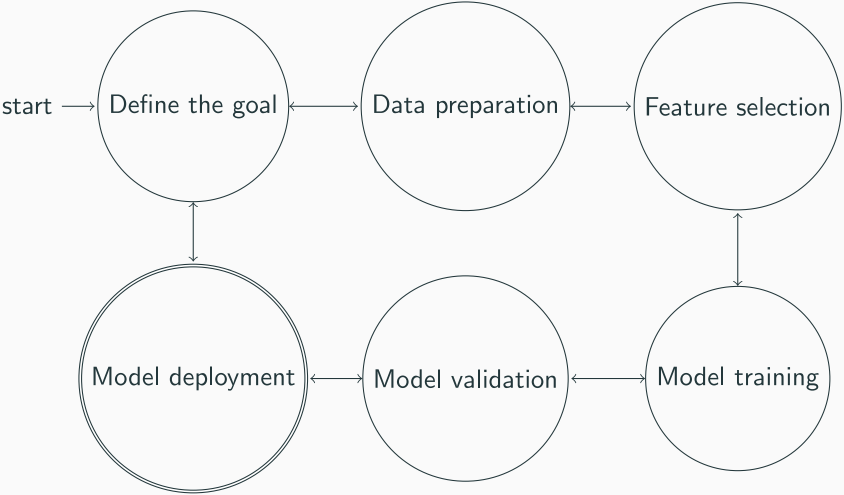
1. Supervised Learning
    - Learn a model from a given set of input-output pairs, in order to predict the output of new inputs.
    - Further grouped into **Regression** and **classification** problems.
2. Unsupervised Learning
    - Discover patterns and learn the structure of unlabelled data.
    - Example **Distribution modeling** and **Clustering**.
3. Reinforcement Learning
    - Learn what actions to take in a given situation, based on rewards and penalties
    - Example consider teaching a dog a new trick: you cannot tell it what to do, but you can reward/punish it.

Machine learning vs traditional statistical analyses

Machine learning	Traditional statistical analyses
Emphasize predictions	Emphasizes superpopulation inference
Evaluates results via prediction performance	Focuses on a-priori hypotheses
Concern for overfitting but not model complexity per se	Simpler models preferred over complex ones (parsimony), even if the more complex models perform slightly better
Emphasis on performance	Emphasis on parameter interpretability
Generalizability is obtained through performance on novel datasets	Statistical modeling or sampling assumptions
Concern over performance and robustness	Concern over assumptions and robustness

End to End Data Science

End to end Data Science



End-to-end predictive analytics approach

STEP 1: Define the goal

STEP 2: Data understanding and preparation

- Importing, cleaning, manipulating and
- Visualizing your data

STEP 3: Building your machine learning model

- Feature selection
- Model training
- Model validation

STEP 4: Model deployment