

Open Data Science Initiative

Neil D. Lawrence

`data@sheffield`

16th December 2015

Challenges for Companies

- ▶ Trying to dominate the modern interconnected data market (e.g. Amazon, Google, Facebook) — buying up talent and competitors.
- ▶ or trying to exploit current 'data silos' (e.g. Tesco's clubcard, Experian) — monetising our data today (limited shelf life?)
- ▶ or trying to understand their own systems (the internal google search)
- ▶ or new companies with new ideas that will generate data.

Challenges for Companies

- ▶ How do they break the natural data monopoly?
- ▶ How do they access the necessary expertise?

Challenges in Science

Data sharing is more widely accepted but:

- ▶ Most analysis is simple statistical tests or explorative modelling with PCA or clustering.
- ▶ Few scientists understand these methodologies, apply them as black box.
- ▶ There is an understanding gap between the data & scientist and the data scientist.

Challenges in Health

- ▶ Ensure the privacy of patients is respected.
- ▶ Leverage the wide range of data available for wider societal benefit.

International Development

- ▶ Exploit new telecommunications infrastructure to develop a leap-frog developed countries.
- ▶ Needs mechanisms for data sharing that retain the individual's control.
- ▶ Widespread education of *local* talent in code and model development.

Common Strands

- ▶ Improving access to data whilst balancing against individual's right to privacy against societal needs to advance.
- ▶ Advancing methodologies: development of methodologies needed to characterize large interconnected complex data sets.
- ▶ Analysis empowerment: giving scientists, clinicians, students, commercial and academic partners ability to analyze their own data with latest methodologies.

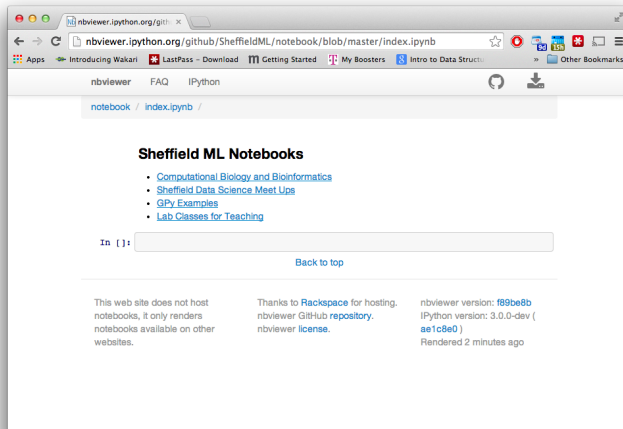
Open Data Science: A Magic Bullet?

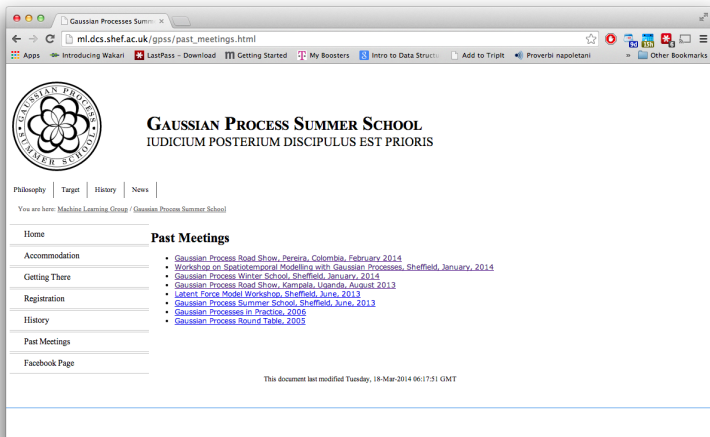
- ▶ Make new methodologies available as widely and rapidly as possible with as few conditions on their use as possible.
- ▶ Educate commercial, scientific and medical partners in use of these methodologies.
- ▶ Act to achieve a balance between data sharing for societal benefit and right of an individual to own their own data.

Achieving This

- ▶ Use BSD-like licenses on software.
- ▶ Educate our partners (summer schools, courses etc).
- ▶ Act to achieve a balance between data sharing for societal benefit and rights of the individual.

Make Analysis Available





The screenshot shows a web browser window with the URL `ml.dcs.shef.ac.uk/gpss/past_meetings.html`. The page features the Gaussian Process Summer School logo on the left, which is a circular emblem with a stylized flower-like design in the center. To the right of the logo, the text "GAUSSIAN PROCESS SUMMER SCHOOL" is displayed in a large, bold, serif font, followed by the Latin motto "IUDICIUM POSTERIUM DISCIPULUS EST PRIORIS" in a smaller, all-caps serif font. Below the logo and motto, there is a navigation bar with links for "Philosophy", "Target", "History", and "News". A breadcrumb trail indicates the current location: "You are here: Machine Learning Group / Gaussian Process Summer School". On the left side of the page, there is a vertical menu with links for "Home", "Accommodation", "Getting There", "Registration", "History", "Past Meetings", and "Facebook Page". The "Past Meetings" link is highlighted. The main content area is titled "Past Meetings" and contains a list of seven past events, each with a bullet point and a link to a specific page. The events are: "Gaussian Process Road Show, Pereira, Colombia, February 2014", "Workshop on Spatiotemporal Modelling with Gaussian Processes, Sheffield, January, 2014", "Gaussian Process Winter School, Sheffield, January, 2014", "Gaussian Process Road Show, Kampala, Uganda, August 2013", "Latent Force Model Workshop, Sheffield, June, 2013", "Gaussian Process Summer School, Sheffield, June, 2013", and "Gaussian Processes in Practice, 2006". At the bottom of the page, a footer note states: "This document last modified Tuesday, 18-Mar-2014 06:17:51 GMT".

GAUSSIAN PROCESS SUMMER SCHOOL
IUDICIUM POSTERIUM DISCIPULUS EST PRIORIS

Philosophy | Target | History | News

You are here: Machine Learning Group / Gaussian Process Summer School

Home
Accommodation
Getting There
Registration
History
Past Meetings
Facebook Page

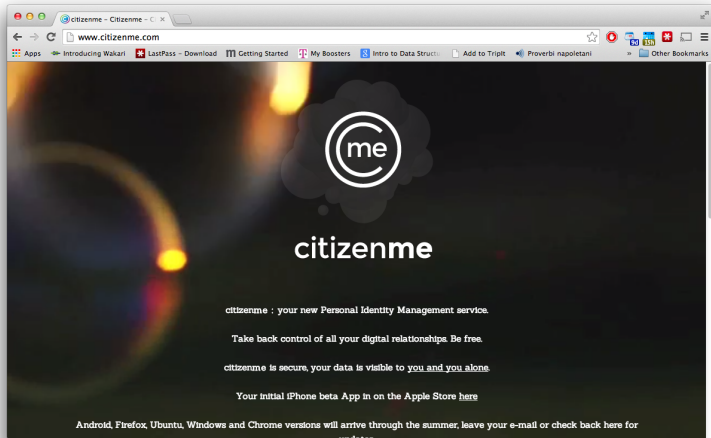
Past Meetings

- [Gaussian Process Road Show, Pereira, Colombia, February 2014](#)
- [Workshop on Spatiotemporal Modelling with Gaussian Processes, Sheffield, January, 2014](#)
- [Gaussian Process Winter School, Sheffield, January, 2014](#)
- [Gaussian Process Road Show, Kampala, Uganda, August 2013](#)
- [Latent Force Model Workshop, Sheffield, June, 2013](#)
- [Gaussian Process Summer School, Sheffield, June, 2013](#)
- [Gaussian Processes in Practice, 2006](#)
- [Gaussian Process Round Table, 2005](#)

This document last modified Tuesday, 18-Mar-2014 06:17:51 GMT

But we need to do much more!

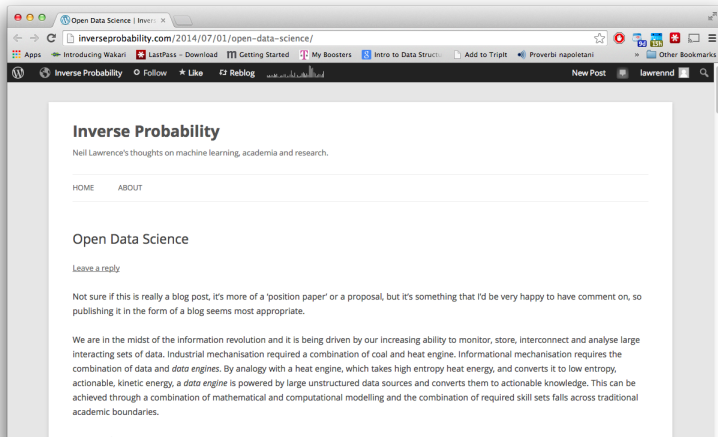
Digital Identity and Data Ownership



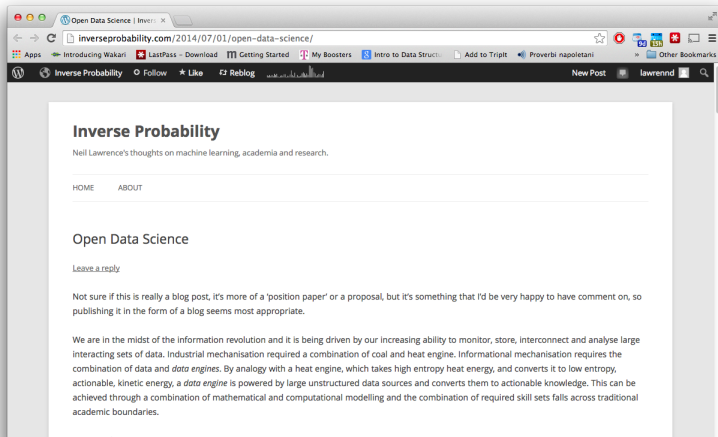
Data Warehousing



Blog Post



Blog Post



[sign in](#)[subscribe](#)[search](#)[jobs](#)[dating](#)[more](#)[UK edition](#)

theguardian

Winner of the Pulitzer prize

[UK](#) [world](#) [sport](#) [football](#) [opinion](#) [culture](#) [economy](#) [lifestyle](#) [fashion](#) [environment](#) [tech](#) [money](#) [travel](#)[all](#)[home](#)

media network

☞ Beware the rise of the digital oligarchy

Neil Lawrence

Powerful algorithms and the concentration of data in the hands means we need better models of data-ownership

0 comments



Eight lessons political parties need to learn to woo young voters

Matthew Hook

2 comments



Mobile World Congress 2015: what it means for marketing pros

James Hilton



How we made MailMen for Royal Mail



PocketHighStreet: linking bricks and clicks at a local level



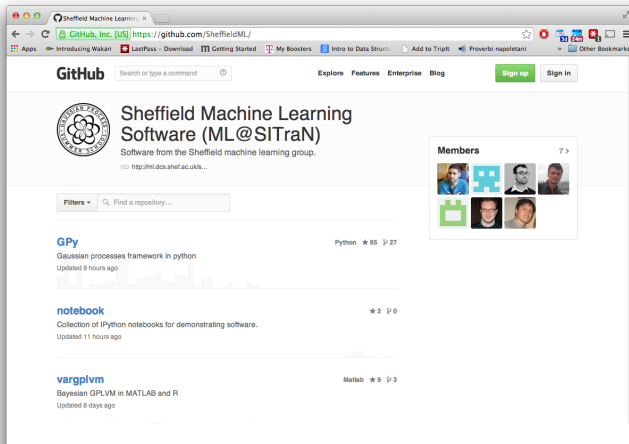
The return of the full-service agency approach

Olly Markeson

0 comments

theguardian.com/media-network/2015/mar/05/political-parties-woo-young-voters-general-election [comments](#)

Modern Tools: Github



Modern Tools: Reddit

The screenshot shows a web browser window displaying a Reddit AMA (Ask Me Anything) thread. The browser's address bar shows the URL: www.reddit.com/r/MachineLearning/comments/25lnbt/ama_yann_lecun/. The browser's bookmark bar contains several items, including 'LastPass - Download', 'Getting Started', 'My Boosters', 'Intro to Data Structures', 'Add to TripIt', 'Proverbs napoletani', and 'Other Bookmarks'. The Reddit interface shows the thread title 'AMA: Yann LeCun' with a self-post icon and a submission time of 'submitted 1 month ago by ylecun'. The post has 347 upvotes and 287 comments. The main content of the post is a text-based AMA where Yann LeCun, Director of Facebook AI Research, discusses his work in deep learning, convolutional nets, and related topics. He mentions his time at NYU's Center for Data Science and his current work at Facebook. The post is sorted by 'best' and has a 'top 200 comments' section visible. The comments section shows a user named 'Hilwerid' asking a question about Yann LeCun's team at Facebook. The post also has a 'MachineLearning' subreddit badge with 24,903 readers and ~23 users online. A sidebar on the right lists related subreddits like 'Statistics' and 'Computer Vision'.

AMA: Yann LeCun : Machine Learning

reddit

AMA: Yann LeCun (self:MachineLearning)
submitted 1 month ago by ylecun · stickied post

My name is Yann LeCun. I am the Director of Facebook AI Research and a professor at New York University.

Much of my research has been focused on deep learning, convolutional nets, and related topics. I joined Facebook in December to build and lead a research organization focused on AI. Our goal is to make significant advances in AI. I have answered some questions about Facebook AI Research (FAIR) in several press articles: Daily Beast, KDNuggets, Wired.

Until I joined Facebook, I was the founding director of NYU's Center for Data Science.

I will be answering questions **Thursday 5/15** between 4:00 and 7:00 PM Eastern Time.

I am creating this thread in advance so people can post questions ahead of time. I will be announcing this AMA on my Facebook and Google+ feeds for verification.

287 comments · share

top 200 comments · show all 287

sorted by: best

[-] Hilwerid 46 points 1 month ago

What is your team at Facebook like?

How is it different then your team at NYU?

In your opinion, why have most renowned professors (e.g. yourself, Geoff Hinton, Andrew Ng) in deep learning attached themselves to a company?

Can you please offer some advice to students who are involved with and/or interested in pursuing deep learning?

permalink

[-] ylecun [S] 68 points 1 month ago

My team at Facebook AI Research is fantastic. It currently has about 20 people split between Menlo Park and New York, and is growing quickly. The research activities focus on learning methods and algorithms (supervised and unsupervised), deep learning - structured prediction, deep learning with sequential/temporal signals, applications in image recognition, face recognition, natural language understanding. An important component is ML software platform and infrastructure. We are using Torch7 for many projects (as does Deep Mind and several groups at Google) and will be contributing to the public version.

My group at NYU used to work a lot on applications in vision/robotics/speech (and other domains) when the purpose was to convince the research community that deep learning actually works. Although we still work on vision, speech and robotics, now that deep learning has taken off, we are doing more work on theoretical stuff (e.g. optimization), new methods (e.g. unsupervised learning) and connections with computational neuroscience and visual psychophysics.

search reddit

this post was submitted on 15 May 2014
347 points (86% upvoted)
shortlink: <http://redd.it/25lnbt>

username password
☐ remember me

Submit a new link

Submit a new text post

MachineLearning

subscribe 24,903 readers
~23 users here now

News, Research Papers, Videos, Lectures, Software and Discussions on:

- Machine Learning
- Data Mining
- Information Retrieval
- Predictive Statistics
- Learning Theory
- Search Engines
- Pattern Recognition
- Analytics

Beginners:

Please have a look at our FAQ and Link-Collection

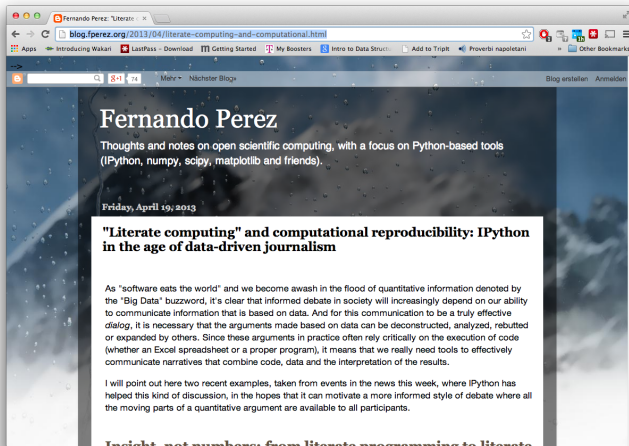
Related Subreddit:

- Statistics
- Computer Vision

Modern Tools: IPython Notebook



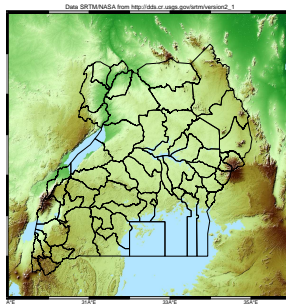
Literate Computing



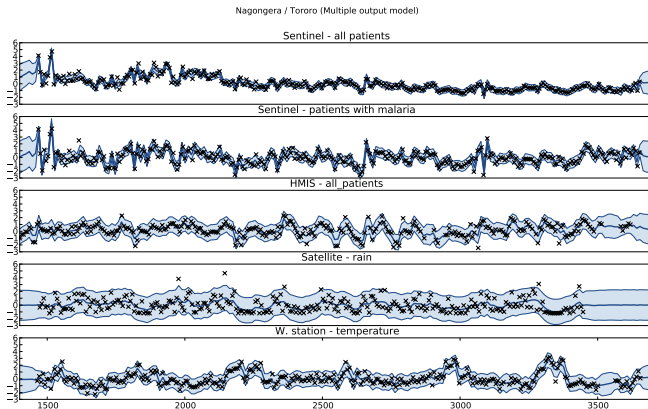
Example: Prediction of Malaria Incidence in Uganda

- ▶ Work with John Quinn and Martin Mubaganzi (Makerere University, Uganda)
- ▶ See <http://air.ug/research.html>.

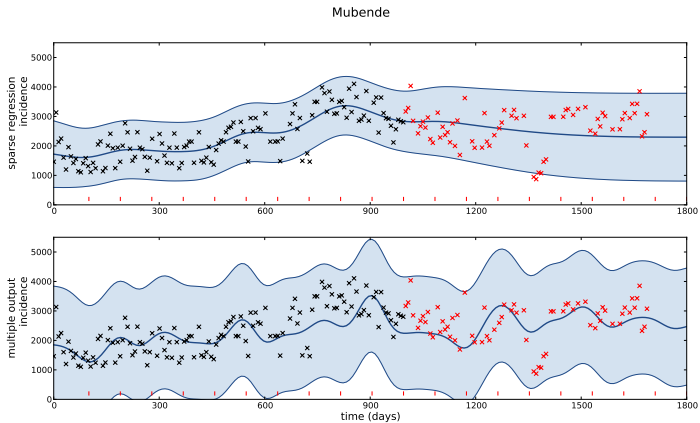
Malaria Prediction in Uganda



Malaria Prediction in Uganda



Malaria Prediction in Uganda



GP School at Makerere



Early Warning Systems

Early Warning Systems