

Datascience: A New Field or Just a Rebadging Exercise?

Neil D. Lawrence

Sheffield Institute of Translational Neuroscience and
Department of Computer Science, University of Sheffield,
U.K.

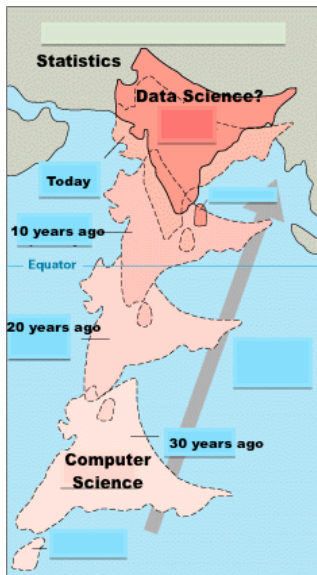
Warwick Statistics

26th November 2014

Shifting Landscapes



Shifting Landscapes



Shifting Landscapes



Shifting Landscapes



Outline

Introduction

Nonparametrics

Process Composition

Conclusions

Box Quote

All models are wrong, but some are useful. (Box, 1976)

Box Quote

All models are wrong, but some are useful. (Box, 1976)

- ▶ Useful quote, but overused.

Box Quote

All models are wrong, but some are useful. (Box, 1976)

- ▶ Useful quote, but overused.
- ▶ Almost become an excuse, my model is wrong so it *might* be useful.

Box Quote

All models are wrong, but some are useful. (Box, 1976)

- ▶ Useful quote, but overused.
- ▶ Almost become an excuse, my model is wrong so it *might* be useful.

*... the scientist must be alert to what is importantly wrong.
It is inappropriate to worry about mice when there are tigers
abroad.* (Box, 1976)

An Incorrect Model

- ▶ Write down our data ...

$$\mathbf{Y} \in \mathcal{R}^{n \times p}$$

An Incorrect Model

- ▶ Write down our data ...

$$\mathbf{Y} \in \mathcal{R}^{n \times p}$$

... this is WRONG!

Is this Separation a Historical Anachronism?

- ▶ A presumption: there is something special and separate about indices over n and p .
- ▶ The subtle difference between features and data points.
- ▶ In practice both n and p could be uncountably large!
- ▶ Standard approach seems to assume that p is fixed.
- ▶ A historic anachronism from the days of collating statistical information?

There is nothing special about p ...

- ▶ Rather ... let's assume each data is indexed by the type of data, as well as location, time, etc.
- ▶ So $y_{17,234}$ is price of a hamburger from McDonald's in Leicester square on 13th April 1984 at 13:34 and $y_{239,201}$ is the price of a chicken wrap from Pret a Manger in Cambridge on 27th December 2001 at 14:34.
- ▶ Further $y_{734,124}$ might be the brand of car my mother currently drives.

Prediction

The answer to any prediction problem is a probability distribution. (Peter McCulloch via Peter Diggle)

- ▶ We assume that we are interested in predicting something about our variables (the likely cost of a burger given the cost of a chicken wrap).

Factorizations

- ▶ Often researchers write down the resulting factorization without a second thought:

$$p(\mathbf{Y}|\boldsymbol{\theta}) = \prod_{i=1}^n p(\mathbf{y}_{i,:}|\boldsymbol{\theta})$$

- ▶ This means that all our information about different data is stored in the parameters.
- ▶ If model is complex, and number of parameters is large, then they will be badly determined when data is few.
- ▶ For me: interesting *research* problems are defined by needing (more) complex models.

Data and Modelling

- ▶ “The Unreasonable Effectiveness of ...
 - ▶ ... Mathematics” (Wigner, 1960)
 - ▶ ...Data” (Halevy et al., 2009)
- ▶ This is a *false* dichotomy.
- ▶ Both are needed for challenging problems of the future.
 - ▶ The relative importance of each is dependent on application.
 - ▶ Norvig also accepts this (see Nando’s question: <http://www.youtube.com/watch?v=yvDCzhhbjYWs&t=54m40s>).
- ▶ Prediction requires model (mathematics) and data.
- ▶ Having better models is particularly important when there’s *uncertainty*.

Open Data

- ▶ Automatic data curation: from curated data to curation of publicly available data.
- ▶ Open Data: `http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M`.

Open Data

- ▶ Automatic data curation: from curated data to curation of publicly available data.
- ▶ Open Data: <http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M>.



Open Data

- ▶ Automatic data curation: from curated data to curation of publicly available data.
- ▶ Open Data: <http://www.openstreetmap.org/?lat=53.38086&lon=-1.48545&zoom=17&layers=M>.



- ▶ Social network data, music information (Spotify), exercise.

Not Wrong ... Just Useless

- ▶ Here's a model that's not wrong ...

Not Wrong ... Just Useless

- ▶ Here's a (graphical) model that's not wrong ...



Not Wrong ... Just Useless

- ▶ Here's a model that's not wrong ...



... it's just useless.

Not Wrong ... Just Useless

- ▶ Here's a model that's not wrong ...



... it's just useless.

- ▶ Does that imply all models that are not wrong are useless?

Not Wrong ... Just Useless

- ▶ Here's a model that's not wrong ...



... it's just useless.

- ▶ Does that imply all models that are not wrong are useless?
- ▶ What is the minimum we can say about our data to get something useful?

Outline

Introduction

Nonparametrics

Process Composition

Conclusions

The TT Channel

- ▶ Objective: predict test data, \mathbf{y}^* , given training data, \mathbf{y} .
- ▶ Parametric models assume

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

for some fixed dimensional vector parameters $\boldsymbol{\theta}$.

- ▶ This looks like a communication channel between training and test data (TT Channel).
- ▶ Capacity of channel given by dimensionality of $\boldsymbol{\theta}$.

Massively Missing Data

- ▶ Michael Goldstein's Maid (via Tony O'Hagan).
- ▶ Let me tell you something unusual about myself ...
- ▶ Large amounts of weak information can give a strong picture.
- ▶ But we must deal with uncertainty when this info isn't present.
- ▶ In real life almost all data is missing almost always.

Kolmogorov Consistency

- ▶ **Claim:** To be 'not wrong' my model must be 'Kolmogorov Consistent'.

Kolmogorov Consistency

- ▶ **Claim:** To be 'not wrong' my model must be 'Kolmogorov Consistent'.
- ▶ Kolmogorov consistency says regardless of future observations, my current marginal model of the data is correct. If $\mathbf{y}^* \in \mathfrak{R}^{n^* \times 1}$ then

$$p(\mathbf{y}|n^*) = \int p(\mathbf{y}, \mathbf{y}^*) d\mathbf{y}^*$$

But if the model is Kolmogorov consistent, $p(\mathbf{y}|n^*) = p(\mathbf{y})$.

Kolmogorov Consistency

- ▶ **Claim:** To be 'not wrong' my model must be 'Kolmogorov Consistent'.
- ▶ Kolmogorov consistency says regardless of future observations, my current marginal model of the data is correct. If $\mathbf{y}^* \in \mathfrak{R}^{n^* \times 1}$ then

$$p(\mathbf{y}|n^*) = \int p(\mathbf{y}, \mathbf{y}^*) d\mathbf{y}^*$$

But if the model is Kolmogorov consistent, $p(\mathbf{y}|n^*) = p(\mathbf{y})$.

- ▶ Here: \mathbf{y} is past observations, \mathbf{y}^* is all possible *future* observations (in either p or n).
- ▶ Models of this type allow us to deal with *massive* missing data because \mathbf{y}^* can even be infinite dimensional.
- ▶ To these models missing data is equivalent to test data.

Nonparametric TT Channel

- ▶ In a non parametric model:

$$p(\mathbf{y}^*|\mathbf{y})$$

Cannot be written as

$$\int p(\mathbf{y}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

for fixed dimensional $\boldsymbol{\theta}$.

The TT Channel

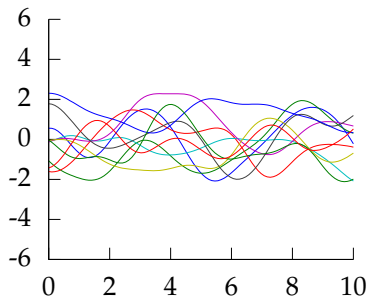
- ▶ Objective: predict test data, \mathbf{y}^* , given training data, \mathbf{y} .
- ▶ Parametric models assume

$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

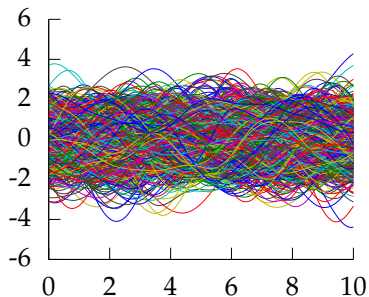
for some fixed dimensional vector parameters $\boldsymbol{\theta}$.

- ▶ This looks like a communication channel between training and test data (TT Channel).
- ▶ Capacity of channel given by dimensionality of $\boldsymbol{\theta}$.

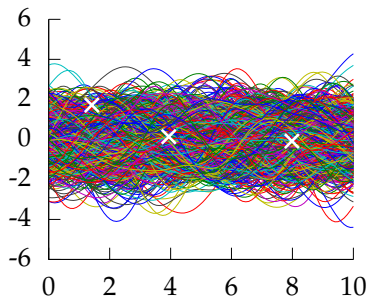
Gaussian Processes: Extremely Short Overview



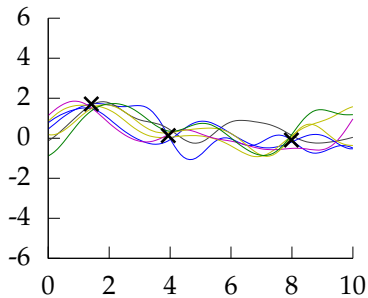
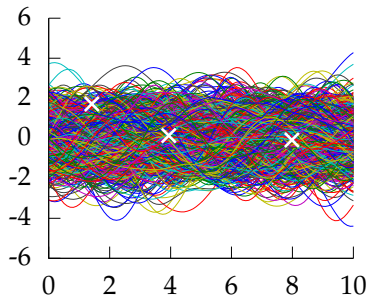
Gaussian Processes: Extremely Short Overview



Gaussian Processes: Extremely Short Overview



Gaussian Processes: Extremely Short Overview



Outline

Introduction

Nonparametrics

Process Composition

Conclusions

- ▶ Composite *multivariate* function

$$\mathbf{g}(\mathbf{x}) = \mathbf{f}_5(\mathbf{f}_4(\mathbf{f}_3(\mathbf{f}_2(\mathbf{f}_1(\mathbf{x}))))))$$

Why Deep?

- ▶ Gaussian processes give priors over functions.
- ▶ Elegant properties:
 - ▶ e.g. *Derivatives* of process are also Gaussian distributed (if they exist).
- ▶ For particular covariance functions they are 'universal approximators', i.e. all functions can have support under the prior.
- ▶ Gaussian derivatives might ring alarm bells.
- ▶ E.g. a priori they don't believe in function 'jumps'.

Process Composition

- ▶ From a process perspective: *process composition*.
- ▶ A (new?) way of constructing more complex *processes* based on simpler components.

Note: To retain *Kolmogorov consistency* introduce IBP priors over latent variables in each layer (Zhenwen Dai).

Analysis of Deep GPs

- ▶ Duvenaud et al. (2014) Duvenaud et al show that the derivative distribution of the process becomes more *heavy tailed* as number of layers increase.

Inducing Variable Approximations

- ▶ Date back to (Williams and Seeger, 2001; Smola and Bartlett, 2001; Csató and Opper, 2002; Seeger et al., 2003; Snelson and Ghahramani, 2006). See Quiñonero Candela and Rasmussen (2005) for a review.
- ▶ We follow variational perspective of (Titsias, 2009).
- ▶ This is an augmented variable method, followed by a collapsed variational approximation (King and Lawrence, 2006; Hensman et al., 2012).

Augmented Variable Model: Not Wrong but Useful?

Augment standard model with a set of m new inducing variables, \mathbf{u} .

$$p(\mathbf{y}) = \int p(\mathbf{y}, \mathbf{u}) d\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Augment standard model with a set of m new inducing variables, \mathbf{u} .

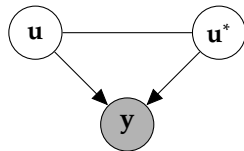
$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})d\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Important: Ensure inducing variables are *also* Kolmogorov consistent (we have m^* other inducing variables we are not *yet* using.)

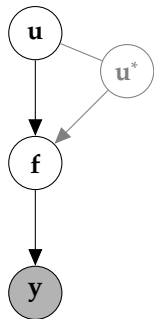
$$p(\mathbf{u}) = \int p(\mathbf{u}, \mathbf{u}^*) d\mathbf{u}^*$$



Augmented Variable Model: Not Wrong but Useful?

Assume that relationship is through \mathbf{f} (represents 'fundamentals'—push Kolmogorov consistency up to here).

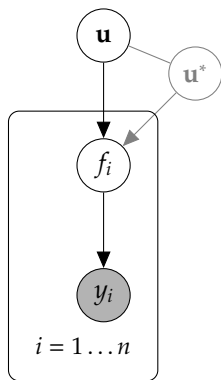
$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}d\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Convenient to assume factorization
(*doesn't* invalidate model—think delta
function as worst case).

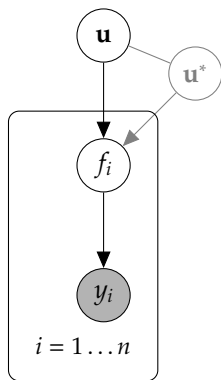
$$p(\mathbf{y}) = \int \prod_{i=1}^n p(y_i|f_i)p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}d\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Focus on integral over \mathbf{f} .

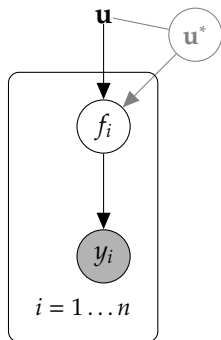
$$p(\mathbf{y}) = \int \int \prod_{i=1}^n p(y_i | f_i) p(\mathbf{f} | \mathbf{u}) d\mathbf{f} p(\mathbf{u}) d\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Focus on integral over \mathbf{f} .

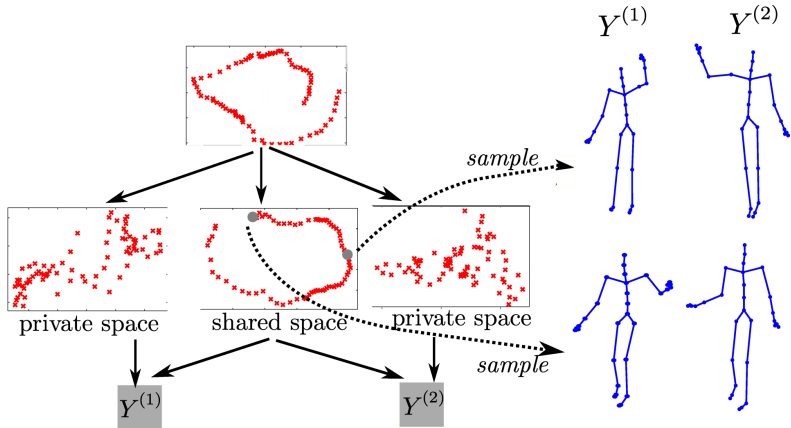
$$p(\mathbf{y}|\mathbf{u}) = \int \prod_{i=1}^n p(y_i|f_i)p(\mathbf{f}|\mathbf{u})d\mathbf{f}$$



Motion Capture

- ▶ 'High five' data.
- ▶ Model learns structure between two interacting subjects.

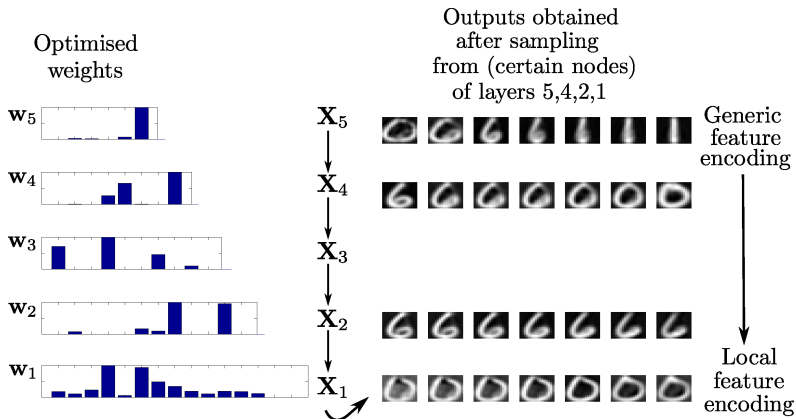
Deep hierarchies – motion capture



Digits Data Set

- ▶ Are deep hierarchies justified for small data sets?
- ▶ We can lower bound the evidence for different depths.
- ▶ For 150 6s, 0s and 1s from MNIST we found at least 5 layers are required.

Deep hierarchies – MNIST



What Can We Do that Internet Giants Can't?

- ▶ Google's resources give them access to volumes of data (or Facebook, or Microsoft, or Amazon).
- ▶ Is there anything for Universities to contribute?
- ▶ Assimilation of multiple views of the patient: each perhaps from a different patient.
- ▶ This may be done by small companies (with support of Universities).
- ▶ A Facebook app for your personalised health.
- ▶ These methodologies are part of that picture.

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.

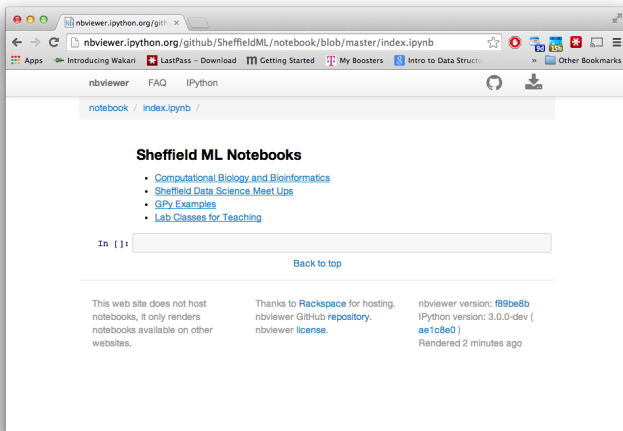
Open Data Science: A Magic Bullet?

- ▶ 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 `nbviewer.ipynb.org/gist/...` and `nbviewer.ipynb.org/github/SheffieldML/notebook/blob/master/index.ipynb`. The browser's address bar and tabs are visible. The page content includes a navigation bar with links for `nbviewer`, `FAQ`, and `IPython`. Below the navigation bar, the breadcrumb `notebook / index.ipynb /` is shown. The main heading is **Sheffield ML Notebooks**, followed by a bulleted list of links: [Computational Biology and Bioinformatics](#), [Sheffield Data Science Meet Ups](#), [GPY Examples](#), and [Lab Classes for Teaching](#). A search bar with the text `In []:` and a `Back to top` link are present. At the bottom, there are three columns of text: a disclaimer about rendering notebooks, a thank you to Rackspace for hosting, and version information for nbviewer (f89be8b) and IPython (3.0.0-dev, ae1c8e0), along with a timestamp: `Rendered 2 minutes ago`.

nbviewer FAQ IPython

notebook / index.ipynb /

Sheffield ML Notebooks

- [Computational Biology and Bioinformatics](#)
- [Sheffield Data Science Meet Ups](#)
- [GPY Examples](#)
- [Lab Classes for Teaching](#)

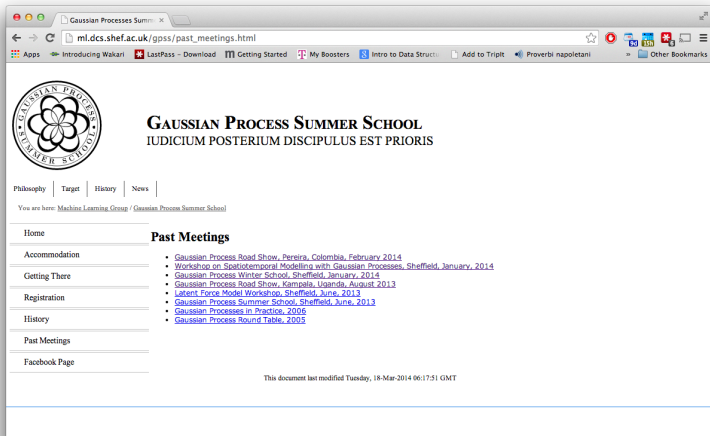
In []:

[Back to top](#)

This web site does not host notebooks, it only renders notebooks available on other websites.

Thanks to [Rackspace](#) for hosting. nbviewer GitHub [repository](#). nbviewer [license](#).

nbviewer version: `f89be8b`
IPython version: `3.0.0-dev (ae1c8e0)`
Rendered 2 minutes ago



The screenshot shows a web browser window displaying the website for the Gaussian Process Summer School. The browser's address bar shows the URL `ml.dcs.shef.ac.uk/gpss/past_meetings.html`. The website features a circular logo on the left with a stylized flower-like pattern and the text "GAUSSIAN PROCESS SUMMER SCHOOL" around the perimeter. To the right of the logo, the title "GAUSSIAN PROCESS SUMMER SCHOOL" is displayed in a large, bold, serif font, with the Latin motto "IUDICIUM POSTERIUM DISCIPULUS EST PRIORIS" underneath it. Below the title, there is a navigation menu 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". A vertical sidebar on the left contains links for "Home", "Accommodation", "Getting There", "Registration", "History", "Past Meetings", and "Facebook Page". The main content area is titled "Past Meetings" and contains a bulleted list of links to various past events, including road shows in Pereira, Colombia (2014), Sheffield (2014), Kampala, Uganda (2013), and a round table in 2005. At the bottom of the page, a small text line states: "This document last modified Tuesday, 18-Mar-2014 06:17:51 GMT".

GAUSSIAN PROCESS SUMMER SCHOOL

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

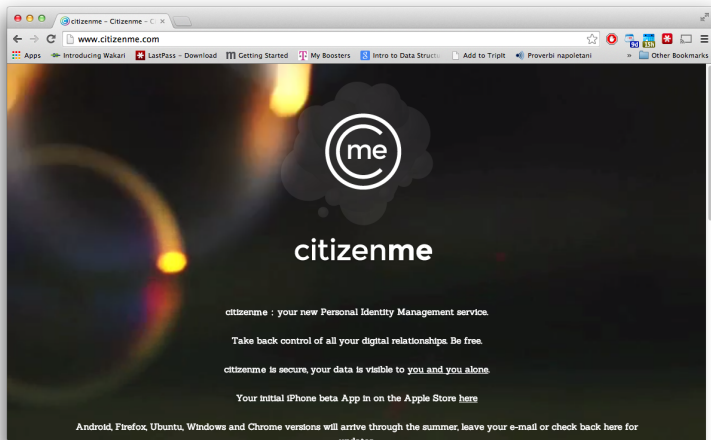
Past Meetings

- [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!

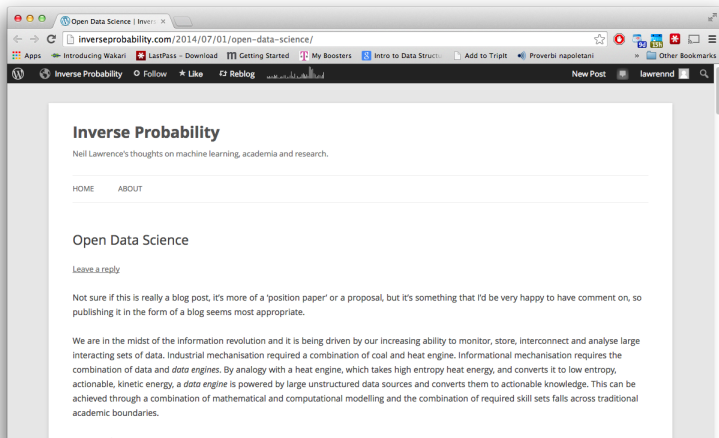
Digital Identity and Data Ownership



Data Warehousing



Blog Post



The screenshot shows a web browser window with the address bar displaying `inverseprobability.com/2014/07/01/open-data-science/`. The browser's bookmark bar contains several items, including 'Inverse Probability', 'Follow', 'Like', and 'Reblog'. The main content of the page is a blog post with the following structure:

Inverse Probability

Neil Lawrence's thoughts on machine learning, academia and research.

[HOME](#) [ABOUT](#)

Open Data Science

[Leave a reply](#)

Not sure if this is really a blog post, it's more of a 'position paper' or a proposal, but it's something that I'd be very happy to have comment on, so publishing it in the form of a blog seems most appropriate.

We are in the midst of the information revolution and it is being driven by our increasing ability to monitor, store, interconnect and analyse large interacting sets of data. Industrial mechanisation required a combination of coal and heat engine. Informational mechanisation requires the combination of data and *data engines*. By analogy with a heat engine, which takes high entropy heat energy, and converts it to low entropy, actionable, kinetic energy, a *data engine* is powered by large unstructured data sources and converts them to actionable knowledge. This can be achieved through a combination of mathematical and computational modelling and the combination of required skill sets falls across traditional academic boundaries.

Modern Tools: Github

The screenshot shows a web browser window displaying the GitHub page for the Sheffield Machine Learning Software (ML@SITraN) repository. The browser's address bar shows the URL `https://github.com/SheffieldML/`. The page header includes the GitHub logo, a search bar, and navigation links for Explore, Features, Enterprise, and Blog. There are also buttons for Sign up and Sign in.

The main content area features the repository's logo, which is a circular emblem with the text "SHEFFIELD HALLAM UNIVERSITY" and "1828". Next to it is the repository name "Sheffield Machine Learning Software (ML@SITraN)" and a description: "Software from the Sheffield machine learning group." Below this is the URL `http://ml.dcs.shef.ac.uk/...`.

A search bar with the placeholder text "Find a repository..." is located below the repository information. To the right, there is a "Members" section showing a grid of profile pictures of the repository's contributors, with a "7 >" indicator.

The repository list shows three items:

- GPpy**: Gaussian processes framework in python. Python. 85 stars, 27 forks. Updated 9 hours ago.
- notebook**: Collection of IPython notebooks for demonstrating software. 2 stars, 0 forks. Updated 11 hours ago.
- vargplvm**: Bayesian GPLVM in MATLAB and R. Matlab. 9 stars, 3 forks. Updated 8 days ago.

Modern Tools: Reddit

The image shows a screenshot of a web browser displaying a Reddit AMA (Ask Me Anything) page. The browser's address bar shows the URL: www.reddit.com/r/MachineLearning/comments/251nbt/ama_yann_lecun/. The page title is "AMA: Yann LeCun" and it has 4 comments. The post is submitted by user "yannlecun" 1 month ago and is a "sticked post".

The main content of the post is as follows:

AMA: Yann LeCun (self:MachineLearning)
submitted 1 month ago by yannlecun - sticked 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, KIDruggets, 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**

[-] HNewsid 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 (eg. 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

[-] yannlecun [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.

On the right side of the page, there is a search bar, a submission box with "Submit a new link" and "Submit a new text post" buttons, and a sidebar for the "MachineLearning" subreddit. The sidebar shows 24,903 readers, ~23 users here now, and a list of related subreddits including Statistics and Computer Vision.

Modern Tools: IPython Notebook

The screenshot shows a web browser window displaying the nbviewer website. The browser's address bar shows the URL `nbviewer.ipython.org`. The website's navigation bar includes links for `nbviewer`, `FAQ`, and `IPython`. The main heading is **nbviewer**, with the subtitle "A simple way to share IPython Notebooks". Below this is a search bar with the text "URL | GitHub username | GitHub username/repo | Gist ID" and a "Go!" button. The page is organized into sections: "Programming Languages" and "Books".

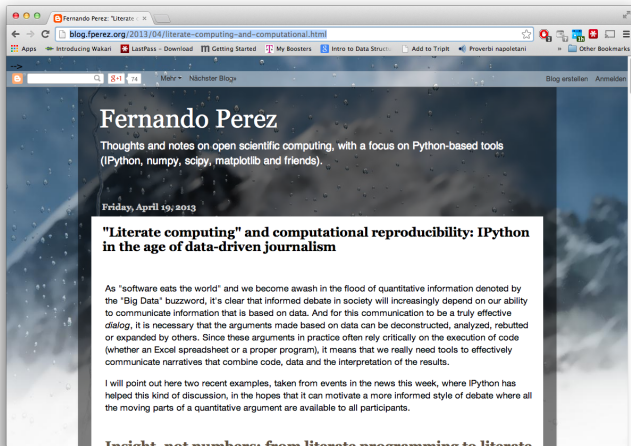
Programming Languages

- IPython**: A card featuring the IPython logo and the text "IP[y]: IPython Interactive Computing".
- IRuby**: A card featuring a red Ruby gem icon and the text "IRuby: Notebook".
- Julia**: A card featuring the Julia logo and the text "An Julia Preview".

Books

- Python for Signal Processing**: A book cover by O'Reilly.
- O'Reilly Book**: A book cover by O'Reilly.
- Probabilistic Programming**: A book cover by O'Reilly.

Literate Computing



A screenshot of a web browser displaying a blog post. The browser's address bar shows the URL `blog.fperex.org/2013/04/literate-computing-and-computational.html`. The page header identifies the author as **Fernando Perez** and describes the blog as "Thoughts and notes on open scientific computing, with a focus on Python-based tools (IPython, numpy, scipy, matplotlib and friends)". The post is dated "Friday, April 19, 2013". The main title of the post is **"Literate computing" and computational reproducibility: IPython in the age of data-driven journalism**. The text of the post begins with a paragraph discussing the "Big Data" buzzword and the need for effective communication of data-based information. It then mentions recent examples from the news where IPython has helped with data-driven journalism. The bottom of the image shows the start of a new paragraph: "Insight, not numbers, from literate programming to literate".

Fernando Perez

Thoughts and notes on open scientific computing, with a focus on Python-based tools (IPython, numpy, scipy, matplotlib and friends).

Friday, April 19, 2013

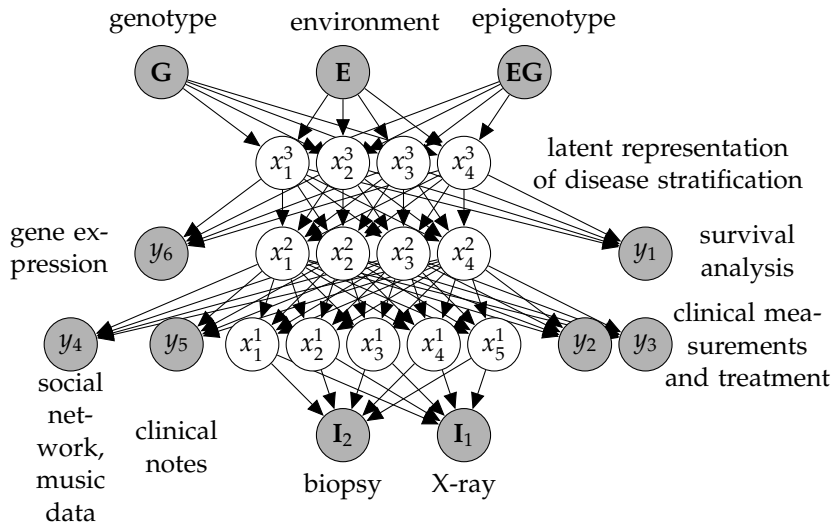
"Literate computing" and computational reproducibility: IPython in the age of data-driven journalism

As "software eats the world" and we become awash in the flood of quantitative information denoted by the "Big Data" buzzword, it's clear that informed debate in society will increasingly depend on our ability to communicate information that is based on data. And for this communication to be a truly effective *dialog*, it is necessary that the arguments made based on data can be deconstructed, analyzed, rebutted or expanded by others. Since these arguments in practice often rely critically on the execution of code (whether an Excel spreadsheet or a proper program), it means that we really need tools to effectively communicate narratives that combine code, data and the interpretation of the results.

I will point out here two recent examples, taken from events in the news this week, where IPython has helped this kind of discussion, in the hopes that it can motivate a more informed style of debate where all the moving parts of a quantitative argument are available to all participants.

Insight, not numbers, from literate programming to literate

Deep Health



Summary

- ▶ 'Big Data' and simple models only takes us so far.
- ▶ Key question: what do we do when 'Big Data' is *small*.
- ▶ Examples include computational biology and personalised health.
- ▶ Our approach is *process composition* (e.g. (Damianou and Lawrence, 2013)).
- ▶ Developing approximate inference algorithms that scale for these models (e.g. (Hensman et al., 2013)).
- ▶ Intention is to deploy these models for assimilating a wide range of data types in personalized health (text, survival times, images, genotype, phenotype).
- ▶ Requires population scale models with millions of features.

References I

- G. E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71(365), 1976.
- L. Csató and M. Opper. Sparse on-line Gaussian processes. *Neural Computation*, 14(3):641–668, 2002.
- A. Damianou and N. D. Lawrence. Deep Gaussian processes. In C. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Workshop on Artificial Intelligence and Statistics*, volume 31, AZ, USA, 2013. JMLR W&CP 31. [PDF].
- D. Duvenaud, O. Rippel, R. Adams, and Z. Ghahramani. Avoiding pathologies in very deep networks. In S. Kaski and J. Corander, editors, *Proceedings of the Seventeenth International Workshop on Artificial Intelligence and Statistics*, volume 33, Iceland, 2014. JMLR W&CP 33.
- A. Y. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009. [DOI].
- J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In A. Nicholson and P. Smyth, editors, *Uncertainty in Artificial Intelligence*, volume 29. AUAI Press, 2013. [PDF].
- J. Hensman, M. Rattray, and N. D. Lawrence. Fast variational inference in the conjugate exponential family. In P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25, Cambridge, MA, 2012. [PDF].
- N. J. King and N. D. Lawrence. Fast variational inference for Gaussian Process models through KL-correction. In *ECML, Berlin, 2006*, Lecture Notes in Computer Science, pages 270–281, Berlin, 2006. Springer-Verlag. [PDF].
- T. K. Leen, T. G. Dietterich, and V. Tresp, editors. *Advances in Neural Information Processing Systems*, volume 13, Cambridge, MA, 2001. MIT Press.
- J. Quiñero Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.
- M. Seeger, C. K. I. Williams, and N. D. Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In C. M. Bishop and B. J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL, 3–6 Jan 2003.
- A. J. Smola and P. L. Bartlett. Sparse greedy Gaussian process regression. In Leen et al. (2001), pages 619–625.
- E. Snelson and Z. Ghahramani. Sparse Gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. C. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18, Cambridge, MA, 2006. MIT Press.

References II

- M. K. Titsias. Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics*, volume 5, pages 567–574, Clearwater Beach, FL, 16-18 April 2009. JMLR W&CP 5.
- E. P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. *Communications on Pure and Applied Mathematics*, 13(1):1–14, 1960. [\[DOI\]](#).
- C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In Leen et al. (2001), pages 682–688.