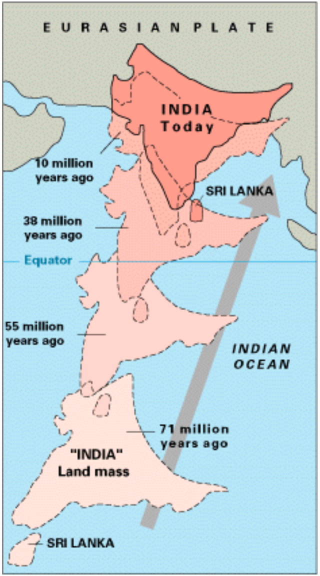
Datascience: A New Field or Just a Rebadging Exercise?

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

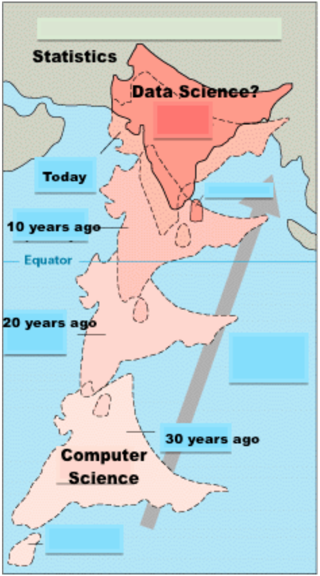
Warwick Statistics

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 ...

$$\mathbf{Y} \in \Re^{n \times p}$$

An Incorrect Model

- Write down our data ...

$$\mathbf{Y} \in \Re^{n \times p}$$

... 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 McCullogh 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=yvDCzhbjYWs&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`.

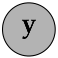
- Social network data, music information (Spotify), exercise.

Not Wrong ... Just Useless

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

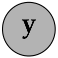
Not Wrong ... Just Useless

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



Not Wrong ... Just Useless

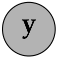
- Here's a model that's not wrong ...



... it's just useless.

Not Wrong ... Just Useless

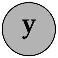
- Here's a model that's not wrong ...



  ... 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

The TT Channel

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

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

  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 $\mathbf{y}^* \in \Re^{n^* \times 1}$ then

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

  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 $\mathbf{y}^* \in \Re^{n^* \times 1}$ then

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

  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

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

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

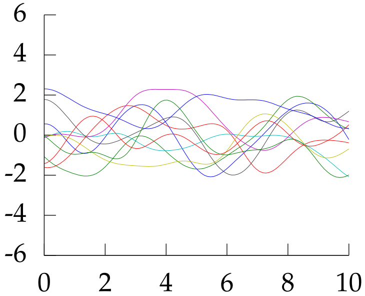
The TT Channel

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

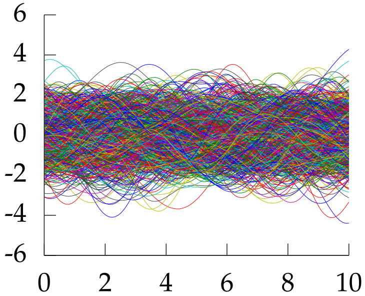
$$p(\mathbf{y}^*|\mathbf{y}) = \int p(\mathbf{y}^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}$$

  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 $\boldsymbol{\theta}$.

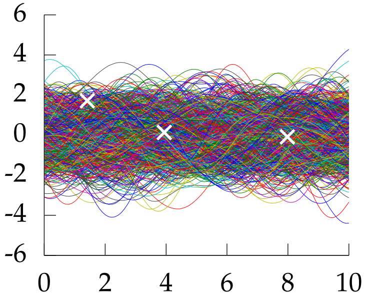
Gaussian Processes: Extremely Short Overview



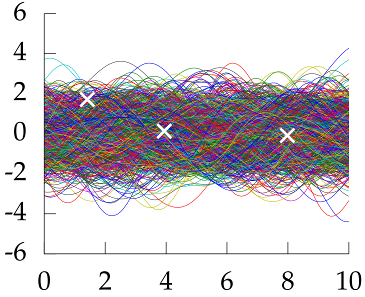
Gaussian Processes: Extremely Short Overview



Gaussian Processes: Extremely Short Overview



Gaussian Processes: Extremely Short Overview



Outline

Mathematically

- 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.

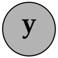
Inducing Variable Approximations

- 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, $\mathbf{u}$.

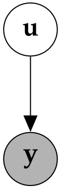
$$p(\mathbf{y}) = \int p(\mathbf{y}, \mathbf{u}) \mathrm{d}\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

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

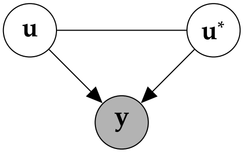
$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{u}$$



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.)

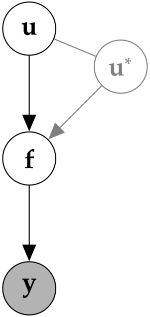
$$p(\mathbf{u}) = \int p(\mathbf{u}, \mathbf{u}^*) \mathrm{d}\mathbf{u}^*$$



Augmented Variable Model: Not Wrong but Useful?

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

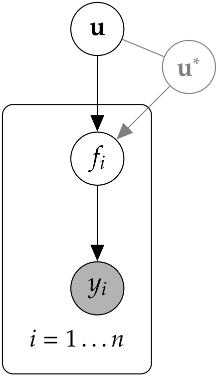
$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

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

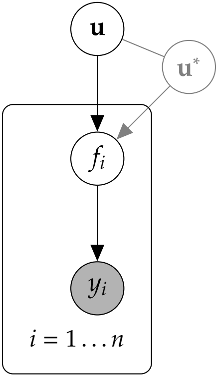
$$p(\mathbf{y}) = \int \prod_{i=1}^{n} p(y_i|f_i)p(\mathbf{f}|\mathbf{u})p(\mathbf{u})\mathrm{d}\mathbf{f}\mathrm{d}\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Focus on integral over $\mathbf{f}$.

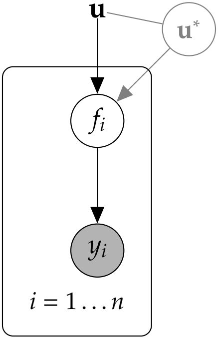
$$p(\mathbf{y}) = \int \int \prod_{i=1}^{n} p(y_i|f_i) p(\mathbf{f}|\mathbf{u}) \mathrm{d}\mathbf{f} p(\mathbf{u}) \mathrm{d}\mathbf{u}$$



Augmented Variable Model: Not Wrong but Useful?

Focus on integral over $\mathbf{f}$.

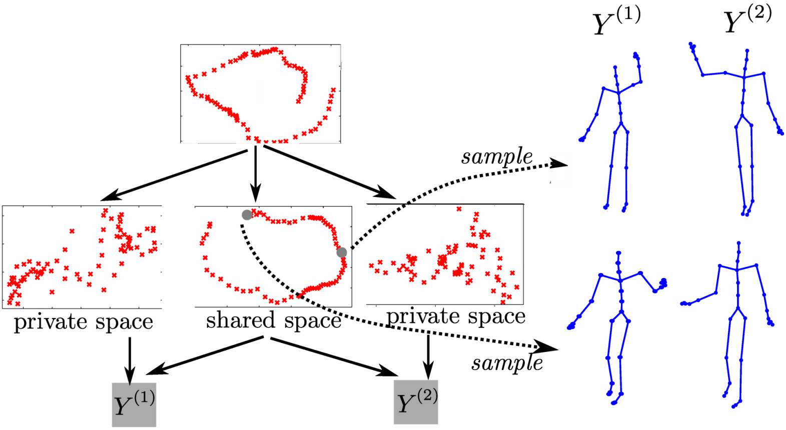
$$p(\mathbf{y}|\mathbf{u}) = \int \prod_{i=1}^{n} p(y_i|f_i)p(\mathbf{f}|\mathbf{u})\mathrm{d}\mathbf{f}$$



Motion Capture

- ‘High five’ data.
- Model learns structure between two interacting subjects.

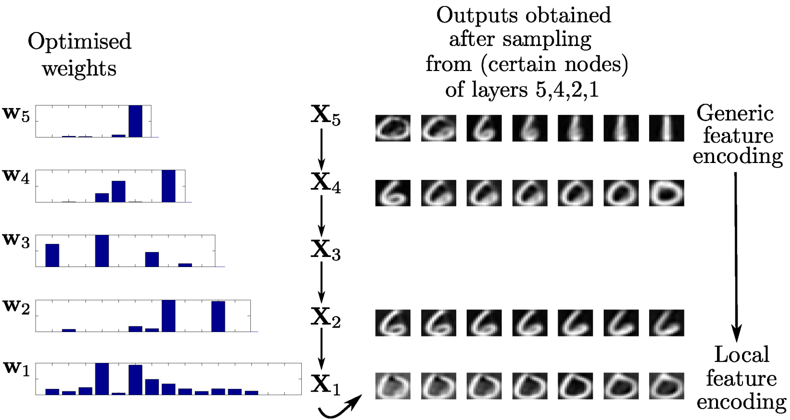
Deep hierarchies – motion capture



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. Tescos 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.

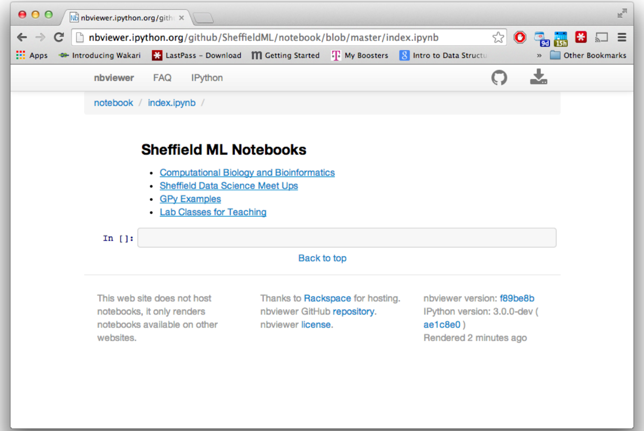
Open Data Science: A Magic Bullet?

- 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.

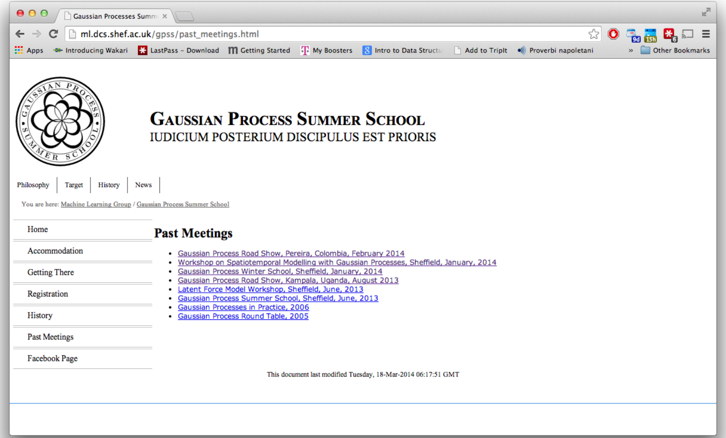
Achieving This

- 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

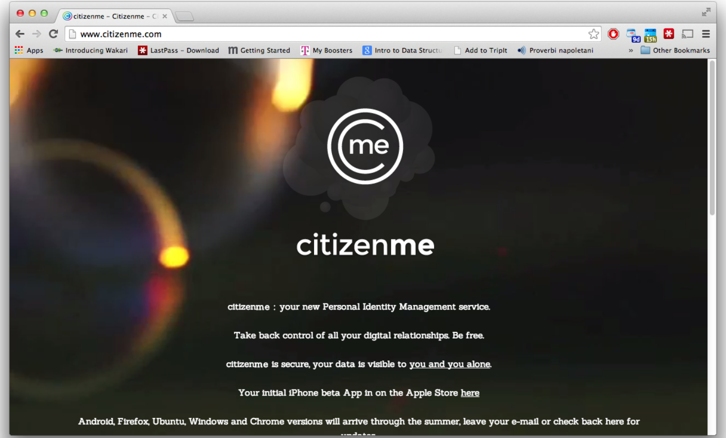


Educating



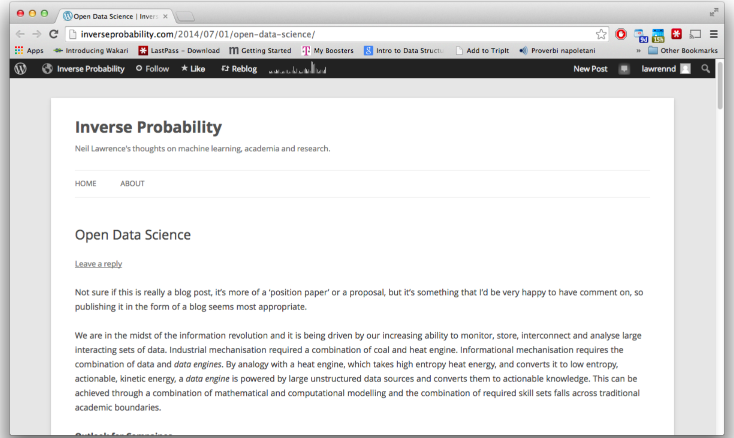
But we need to do much more!

Digital Identity and Data Ownership

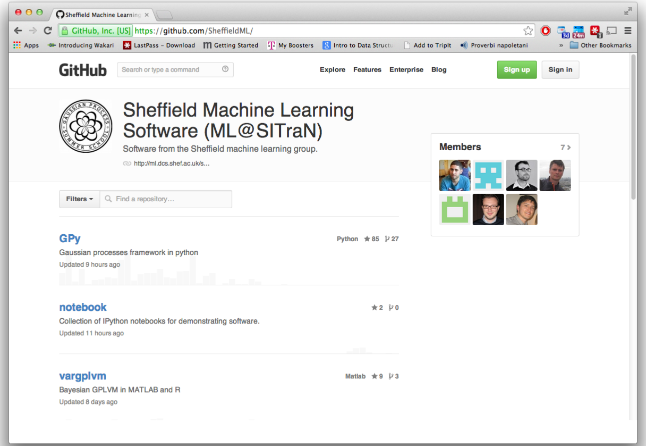


Data Warehousing

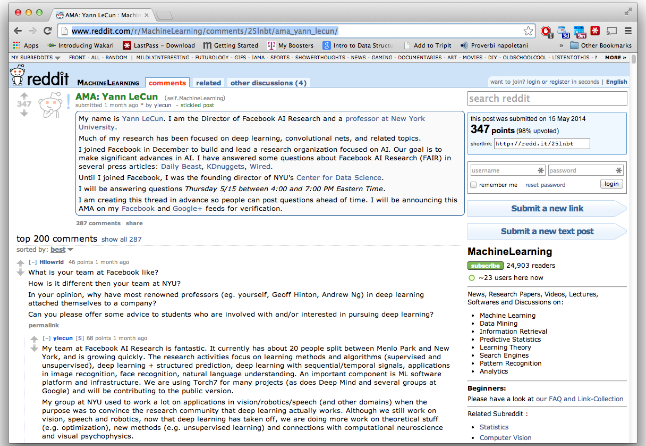
Blog Post


Open Data Science | Invers
inverseprobability.com/2014/07/01/open-data-science/
Apps
Introducing Wakari
LastPass - Download
Getting Started
My Boosters
Intro to Data Structu
Add to Triplt
Proverbi napoletani
Other Bookmarks
Inverse Probability
Follow
Like
Reblog
New Post
lawrennd
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.

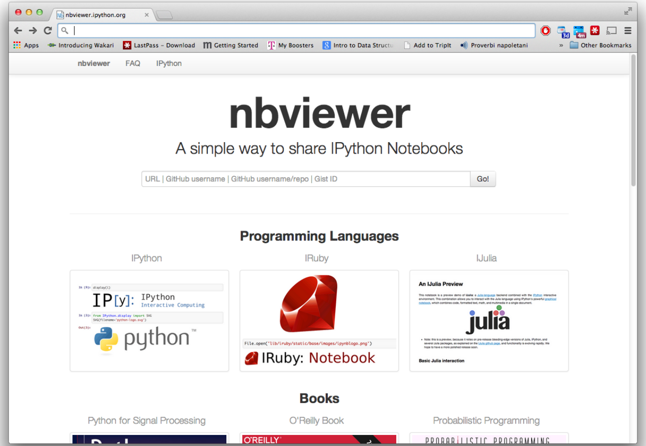
Modern Tools: Github



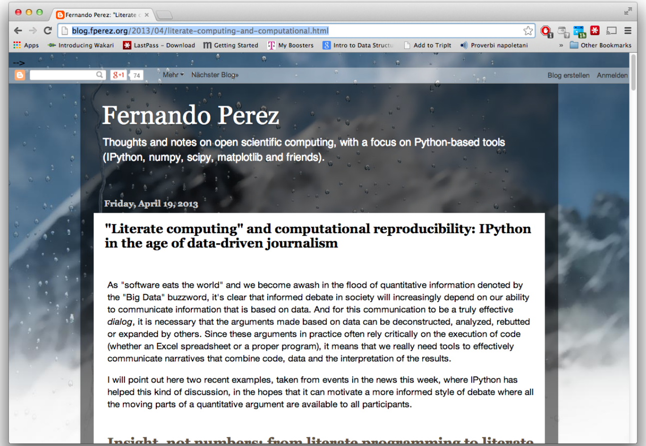
Modern Tools: Reddit



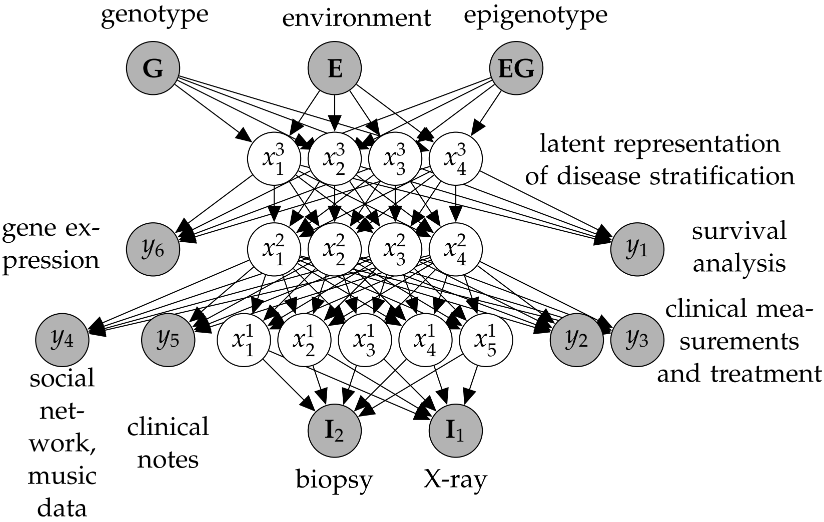
Modern Tools: IPython Notebook



Literate Computing



Deep Health



genotype
environment
epigenotype
G
E
EG
latent representation of disease stratification
x_1^3 x_2^3 x_3^3 x_4^3
gene expression
y_6
x_1^2 x_2^2 x_3^2 x_4^2
y_1
survival analysis
clinical measurements and treatment
y_4 y_5
x_1^1 x_2^1 x_3^1 x_4^1 x_5^1
y_2 y_3
social network, music data
clinical notes
$\mathbf{I}_2$
$\mathbf{I}_1$
biopsy
X-ray

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. Quiñonero 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.