Data Science and me

Guido Sanguinetti

ANC- School of Informatics, University of Edinburgh
Room IF1.44 gsanguin@inf

October 10, 2016

Positional statement

- I was trained as a physicist/ mathematician
- Emphasis on Science in Data Science
- I'm unconvinced by statements that large-scale data gathering will eliminate the need for theory (i.e. hypothesis driven research), except perhaps in some engineering applications.
- However, science also produces vast amounts of data
- Statistical models and machine learning techniques are increasingly central in turning data into knowledge.

Positional statement

- I was trained as a physicist/ mathematician
- Emphasis on Science in Data Science
- I'm unconvinced by statements that large-scale data gathering will eliminate the need for theory (i.e. hypothesis driven research), except perhaps in some engineering applications.
- However, science also produces vast amounts of data
- Statistical models and machine learning techniques are increasingly central in turning data into knowledge.

Current group interests

- Largish group: 6 post-docs, 5 students, 7 nationalities
- Funding from several sources: ERC, EPSRC, Marie Curie, School of Informatics, CDT/ DTC
- Backgrounds from physics, engineering, CS and maths
- Interests range from analysis of sequencing data to dynamical systems theory

Dynamical systems

- Abstractions of real systems focussing on capturing the mechanisms underlying their time-varying behaviour
- Generally described by a state-vector and some (infinitesimal) transition relationships, e.g. $x_{t+1} = f(x_t) + \epsilon_t$, $dx = f(x)dt + \sigma dW, \ldots$
- Or they can also be defined in terms of agents interacting with each other (sometimes, but not always, equivalent)
- Goal: to determine the probability of the system being in a particular state at a particular time (*single time marginal*)
- Useful when domain knowledge enables us to formulate models grounded in what we understand as the physical reality of the system
- Particularly useful for *prediction* and *understanding*, i.e. they strike a nice balance between explanatory and predictive power

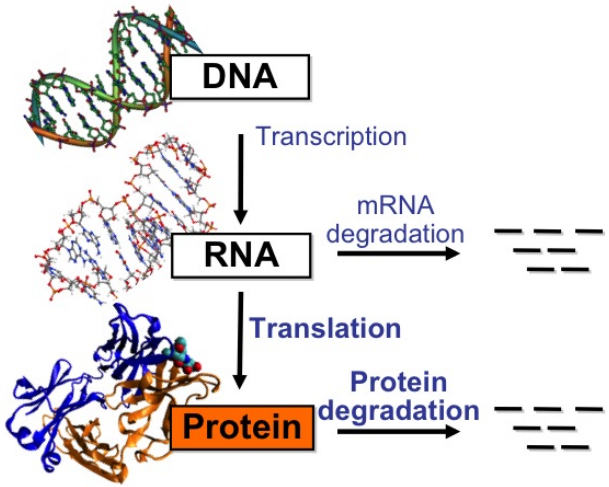
Dynamical systems

- Abstractions of real systems focussing on capturing the mechanisms underlying their time-varying behaviour
- Generally described by a state-vector and some (infinitesimal) transition relationships, e.g. $x_{t+1} = f(x_t) + \epsilon_t$, $dx = f(x)dt + \sigma dW, \ldots$
- Or they can also be defined in terms of agents interacting with each other (sometimes, but not always, equivalent)
- Goal: to determine the probability of the system being in a particular state at a particular time (*single time marginal*)
- Useful when domain knowledge enables us to formulate models grounded in what we understand as the physical reality of the system
- Particularly useful for *prediction* and *understanding*, i.e. they strike a nice balance between explanatory and predictive power

Dynamical systems

- Abstractions of real systems focussing on capturing the mechanisms underlying their time-varying behaviour
- Generally described by a state-vector and some (infinitesimal) transition relationships, e.g. $x_{t+1} = f(x_t) + \epsilon_t$, $dx = f(x)dt + \sigma dW, \ldots$
- Or they can also be defined in terms of agents interacting with each other (sometimes, but not always, equivalent)
- Goal: to determine the probability of the system being in a particular state at a particular time (*single time marginal*)
- Useful when domain knowledge enables us to formulate models grounded in what we understand as the physical reality of the system
- Particularly useful for *prediction* and *understanding*, i.e. they strike a nice balance between explanatory and predictive power

Biology in a slide



Where does variability come into play? What can we measure?
Nice example of a dynamical system with some physical knowledge and a lot of uncertainty.

Systems Biology

- Since late 90s, biologists have been able to measure various biochemical components of cells in a high-throughput fashion
- Also, more precise microscopy-based measurements give time-resolved measurements at single cells
- Each measurement is a noisy readout of one facet of a (set of) complex biological processes
- Interpretable statistical models are (probably) the only way to integrate these disparate data in one coherent mechanistic picture
- Specifically, I work with probabilistic latent variable models (key difference: the latent variables and parameters have physical meanings)

Systems Biology

- Since late 90s, biologists have been able to measure various biochemical components of cells in a high-throughput fashion
- Also, more precise microscopy-based measurements give time-resolved measurements at single cells
- Each measurement is a noisy readout of one facet of a (set of) complex biological processes
- Interpretable statistical models are (probably) the only way to integrate these disparate data in one coherent mechanistic picture
- Specifically, I work with probabilistic latent variable models (key difference: the latent variables and parameters have physical meanings)

Problem

Let's watch a movie!

Problem - in words

- Populations of individual agents of (few) types coexisting in physical space
- Agents move about in space (apparently) randomly
- When agents come into contact (or very close), interactions happen that may result in changes in agents' numbers/ behaviours
- Very frequent scenario in ecology, molecular biology, epidemiology, social sciences, smart cities

Problem - in words

- Populations of individual agents of (few) types coexisting in physical space
- Agents move about in space (apparently) randomly
- When agents come into contact (or very close), interactions happen that may result in changes in agents' numbers/ behaviours
- Very frequent scenario in ecology, molecular biology, epidemiology, social sciences, smart cities

Statistical problem

- The state of the system is determined by the number of particles in each species, and by the positions of the particles
- The state space is a (potentially infinite) union of continuous spaces of different dimensions. The evolution equation for the single time marginal is defined on a Fock space and cannot be solved. No way of getting a likelihood function.
- We found a new representation in terms of Poisson/ Cox point processes which enables us to construct a likelihood surrogate
- Now starting to apply it to the dynamics of disease spreading in Africa

Statistical problem

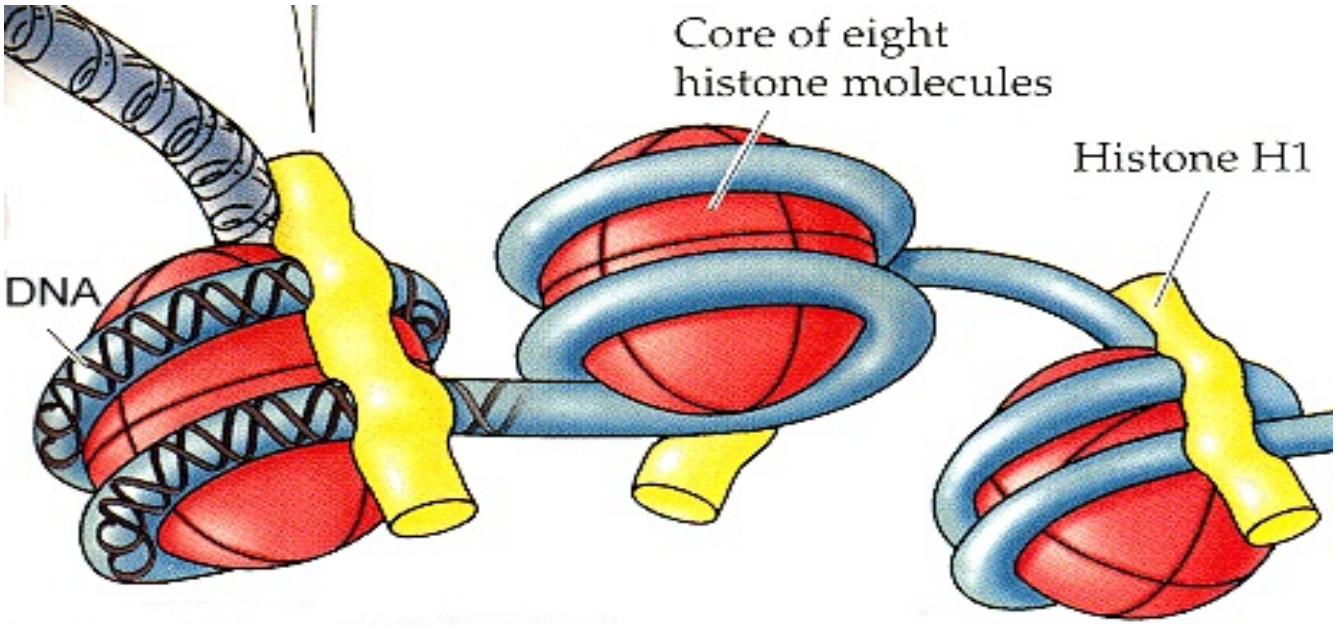
- The state of the system is determined by the number of particles in each species, and by the positions of the particles
- The state space is a (potentially infinite) union of continuous spaces of different dimensions. The evolution equation for the single time marginal is defined on a Fock space and cannot be solved. No way of getting a likelihood function.
- We found a new representation in terms of Poisson/ Cox point processes which enables us to construct a likelihood surrogate
- Now starting to apply it to the dynamics of disease spreading in Africa

Statistical problem

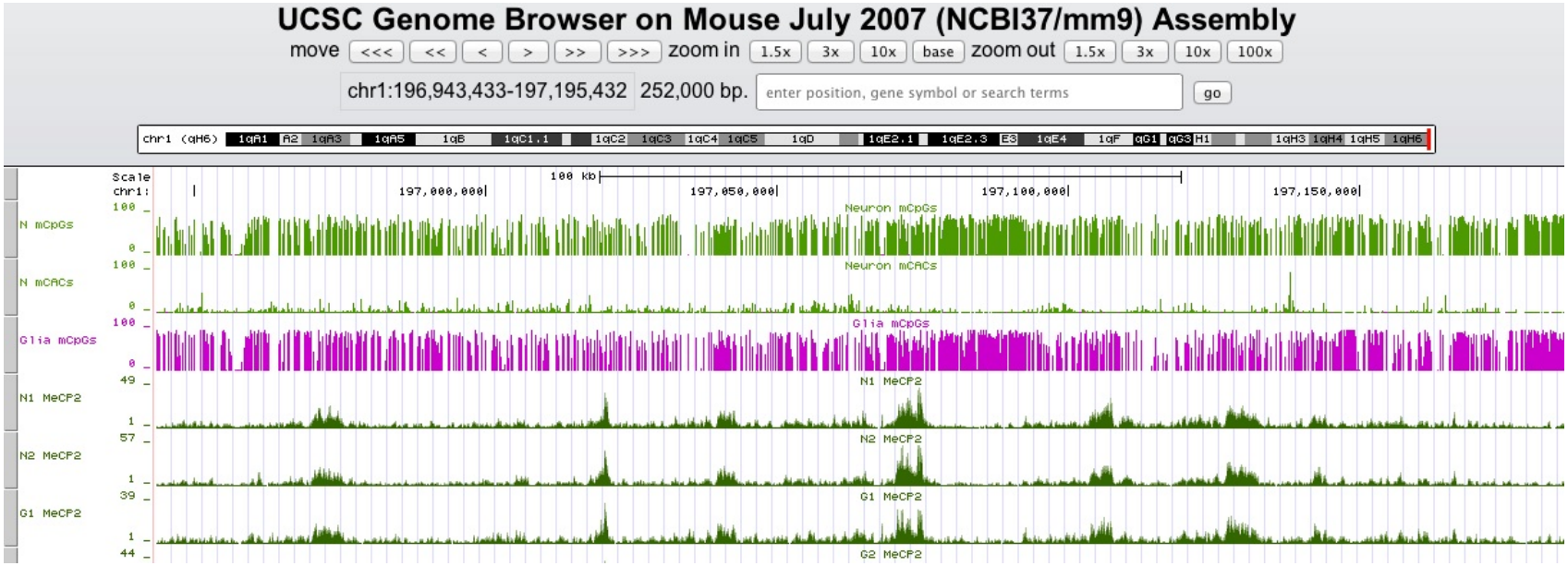
- The state of the system is determined by the number of particles in each species, and by the positions of the particles
- The state space is a (potentially infinite) union of continuous spaces of different dimensions. The evolution equation for the single time marginal is defined on a Fock space and cannot be solved. No way of getting a likelihood function.
- We found a new representation in terms of Poisson/ Cox point processes which enables us to construct a likelihood surrogate
- Now starting to apply it to the dynamics of disease spreading in Africa

Epigenetics

Genetics and transcription cannot be all; spatial organisation of chromosomes plays a role. This is determined by chemical modifications to DNA and histones.



Epigenetics: what the data looks like



Each row is a tiny fraction of a next-generation sequencing experiment's data. Each row $\geq$1GB of data. How do we determine relationships between the rows?

Current results

- Identifying statistically significant differences between the rows is already difficult: some success adapting a kernel method, *Maximum Mean Discrepancy* (Gretton et al 2008), to sequencing data (Schweikert et al, BMC Genomics 2013, Mayo et al, Bioinformatics 2015)
- Predictive models are useful: e.g., given a hypothesis that the green rows are mechanistically determined by the pink rows, we should be able to train a fairly accurate regression model
- Recent success in predicting histone modifications from binding of transcription factor proteins (Benveniste et al, PNAS 2014)
- Technical challenges: large size of the data sets, large number of covariates, inhomogeneities along chromosomes (latent variables?)

Current lines of work

- Develop predictive models to relate sequence and epigenetic marks with each other, based on generalised linear models (T. Mayo)
- Model the interactions between various epigenetic factors and gene expression (consensus clustering, soon to move to more general graphical models) (A. Kapourani, CDT)
- Also important to understand processes downstream of transcription, e.g. RNA folding (A. Selega) and splicing (Y. Huang), and (remarkably) these are often also tied to epigenetics

Looking ahead

- At the moment, the two lines of work appear fairly disjointed, how do we integrate them?
- Technical challenge 1: multi-scale models in spatio-temporal modelling
- Technical challenge 2: (almost) all epigenetic data is a snapshot of a stochastic dynamical process. How do we do inference for (large scale) stochastic dynamical systems from (population/ time) average static measurements?
- Technical challenge 3: how do we identify effective smaller (dynamical) models that match the behaviours observed in data?

Looking ahead

- At the moment, the two lines of work appear fairly disjointed, how do we integrate them?
- Technical challenge 1: multi-scale models in spatio-temporal modelling
- Technical challenge 2: (almost) all epigenetic data is a snapshot of a stochastic dynamical process. How do we do inference for (large scale) stochastic dynamical systems from (population/ time) average static measurements?
- Technical challenge 3: how do we identify effective smaller (dynamical) models that match the behaviours observed in data?

References

- D. Schnoerr, R. Grima and GS, Cox process representation and inference for stochastic reaction-diffusion processes, *Nature Communications* **7**, 11729, 2016
- G. Schweikert, B. Cseke, T. Clouaire, A. Bird and G.S., MMDiff: quantitative testing for shape changes in ChIP-Seq data sets, *BMC Genomics* 14:826, 2013
- D. Benveniste, H.-J. Sonntag, G.S. and D. Sproul, Transcription factor binding predicts histone modifications in human cell lines, *PNAS* 111(37), 13367-13372, 2014
- T. Mayo, G. Schweikert and G.S., M^3D: a kernel-based test for spatially correlated changes in methylation profiles, *Bioinformatics* 31(6), 809-816, 2015
- C-A. Kapourani and GS, Higher order methylation features for clustering and prediction in epigenomic studies, *Bioinformatics* 32(17), i405-i412, 2016 (Proc of ECCB 2016)