Introduction to Applied Statistics

Chris Holmes

Professor of Biostatistics,
Department of Statistics &
Nuffield Department of Clinical Medicine,
University of Oxford

Programme Director for Health
Alan Turing Institute

OxWaSP November 2018

Preamble

- Statistics is the scientific study of uncertainty, randomness, and chance occurrence
  - where uncertainty is measured in units of probability
- Statistics is about being precise about imprecision
- Statisticians study and develop the procedures for learning from data including how to best gather data in order to learn
  - Uncertainty is manipulated using the rules of probability calculus
  - The manipulation of uncertainty is often unintuitive, that's why we need a scientific approach (Statistics) to study it

- Applied statistics is the use of quantitative methods (and visualisation techniques) to assist in inference, the drawing of evidenced based conclusions from data
- Applied statistics is often used to assist optimal decision making under partial information, for example in clinical applications
  - Who to treat?
  - When to treat?
  - What dose to treat?
- Applied statistics puts the scientific problem and the data analysis / data collection first
  - interesting interface between applied stats and machine-learning / AI

$$\widehat{y} = f(x; \theta)$$

  - Crudely speakling.....AI/ML – concentrate of the LHS; Statistics looks to isolate parameters of interest in θ and infer on the RHS

Formal methods: Nine Stages to Applied Statistics

There are broadly nine stages to an analysis

1. Understand the problem
   - the objectives, constraints and criteria for success
2. Understand where the data comes (or will come) from
   - The study design that gave rise to the data, or an optimal design for the experiment
3. Visualisation and data exploration
   - to assist in model formulation
4. Writing an analysis protocol
5. Tentative model fitting
6. Model criticism and refinement
7. Formal assessment of model fit
8. Validation and reporting
9. Documenting of scripts, auditing the analysis, taking responsibility for reproducibility
   - checking against the initial objectives and analysis plan, ensuring repeatability of results

Stage 1: Understand the objectives

- The first step is to understand the problem and the rationale for the analysis

Stage 1: Understand the objectives

- No. I mean *really* understand the problem

Stage 1: Understand the objectives

- No. I mean *really* understand the problem
- It's the most important step. The more effort you put in here the better the analysis, and the lower the risk of solving the wrong problem
  - What are we trying to learn about?
  - Why is it important?
  - Has it been studied before?
  - Is there existing data and has it been used before?

Stage 1: Understand the objectives

- No. I mean *really* understand the problem
- It's the most important step. The more effort you put in here the better the analysis, and the lower the risk of solving the wrong problem
  - What are we trying to learn about?
  - Why is it important?
  - Has it been studied before?
  - Is there existing data and has it been used before?
    - If so. CAUTION!
    - "If you torture the data long enough it will confess to anything", R. Coase

Stage 1: Understand the objectives

- No. I mean *really* understand the problem
- It's the most important step. The more effort you put in here the better the analysis, and the lower the risk of solving the wrong problem
  - What are we trying to learn about?
  - Why is it important?
  - Has it been studied before?
  - Is there existing data and has it been used before?
    - If so. CAUTION!
    - "If you torture the data long enough it will confess to anything", R. Coase
- Keep a data-science lab-book such as Jupyter or Rmarkdown (see pdf in Weblearn)
  - Note your initial ideas and the problem domain; talk the problem back to the data owners – try to explain their problem to them. This will highlight misunderstanding or mis-emphasis

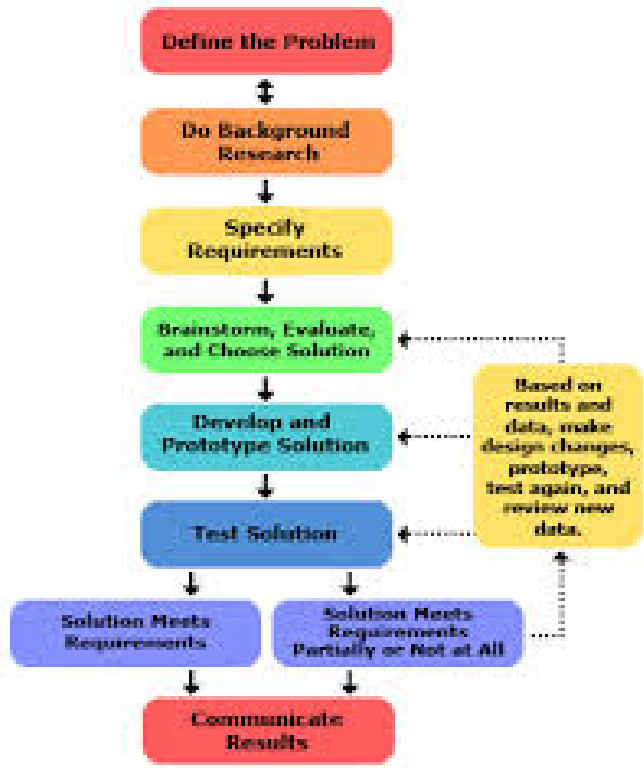
Stage 1: Design criteria and constraints

- Clearly define the criteria for success
  - How will the model or analysis be judged a success?
- What are the constraints (time, money, runtime of analysis, data access, ...)?
  - is the data already in place?
  - or being collected (if so, it will be late!)?
- What is known about the current system under study?
- How will I ensure that the analysis and conclusions are externally transparent and reproducible?
  - current estimates (Freedman et. al., PLOS Biology, 2015) is approximately US$28,000,000,000 (US$28B) per year spent on preclinical research that is not reproducible – in the United States alone

...a brief excursion on the changing nature of Statistics...

- Modern studies are disruptive for Statistics
- Take place in a highly contextual environment
- Provide high-dimensional, heterogeneous data with complex, idiosyncratic noise
- We need to nurture an analysis design process where models, algorithms and implementation are developed holistically alongside domain expertise
  - problem facing models not toys
  - statistics within data science is as an engineering discipling – delivering a solution against stated criteria

- Modern data analysis motivates that we should adopt principles of engineering design process[1] as statisticians; outside of traditional sample size measurement space, (n, p), considerations
  - issues such as robustness, heterogeneity of measurements, contextual domain knowledge, ...etc....
  - The Holmes definition of a data scientist $\equiv$ a statistician who embraces the engineering design process



[1] The "engineering process" is a well defined concept in engineering but I think it perfectly captures modern model building in statistics.

- One illustration of this is the merging or blurring of boundaries between the statistical method (on paper), the computational algorithm (in code), and the hardware platform (in silico)
- It used to be that the statistician first thought of a model, then looked to code it up, and then run it on what ever hardware was available:

Statistics:

$$\prod_{i=1}^{T} f(y_i|x_i^{(k)},\theta)\pi(\theta)$$

$\Longrightarrow$

Code:

```
for i = 1:n
   x_i ← x_i + 1
end
```

$\Longrightarrow$

Hardware:

Emerging theme

- Statisticians need to become computationally expert and domain proficient in order to develop contextual models with algorithmic representations that align to hardware in a holistic manner

Statistics:

$$\prod_{i=1}^{T} f(y_i|x_i^{(k)},\theta)\pi(\theta)$$

Code:

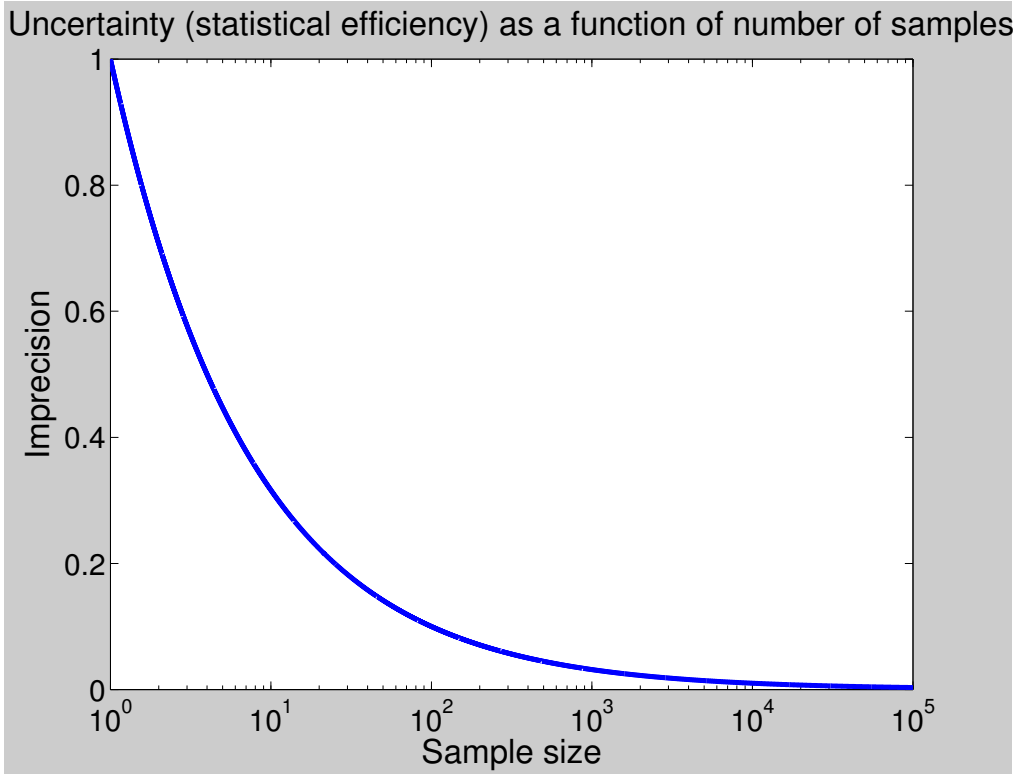
```
for i = 1:n
   x_i ← x_i + 1
end
```

Hardware:

- for example, can I take the data to hardware or should the hardware come to the data?
- does it make sense to even think of a joint model (for all variables), and even if I could think of one, could I work with it?

Once upon a time.....historically....

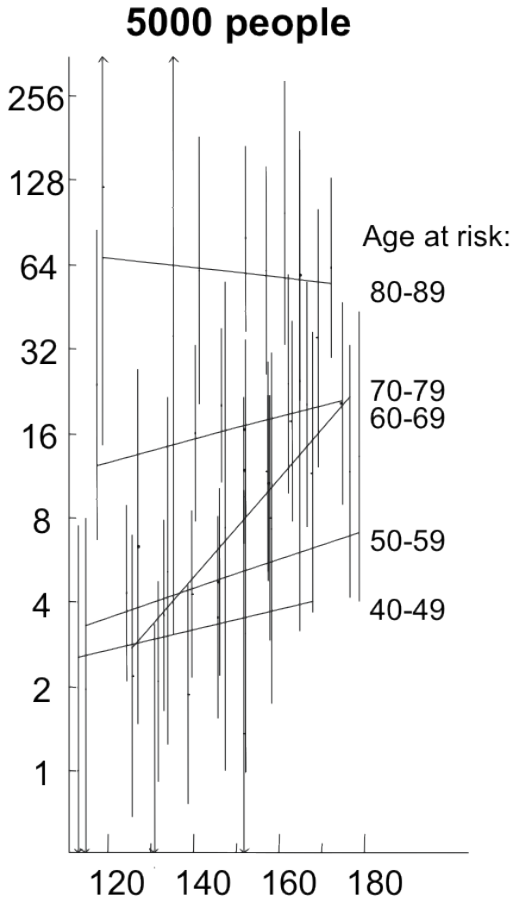
- The origins of statistics concerned itself with the study of precision (statistical efficiency of estimators) as a function of sample size



- Graph illustrating the famous relationship between relative variance (inverse precision) of an estimator such as the sample mean estimating the (unknown and never known) population mean, and $\frac{1}{\sqrt{n}}$

Sample size: illustrating the value of large numbers

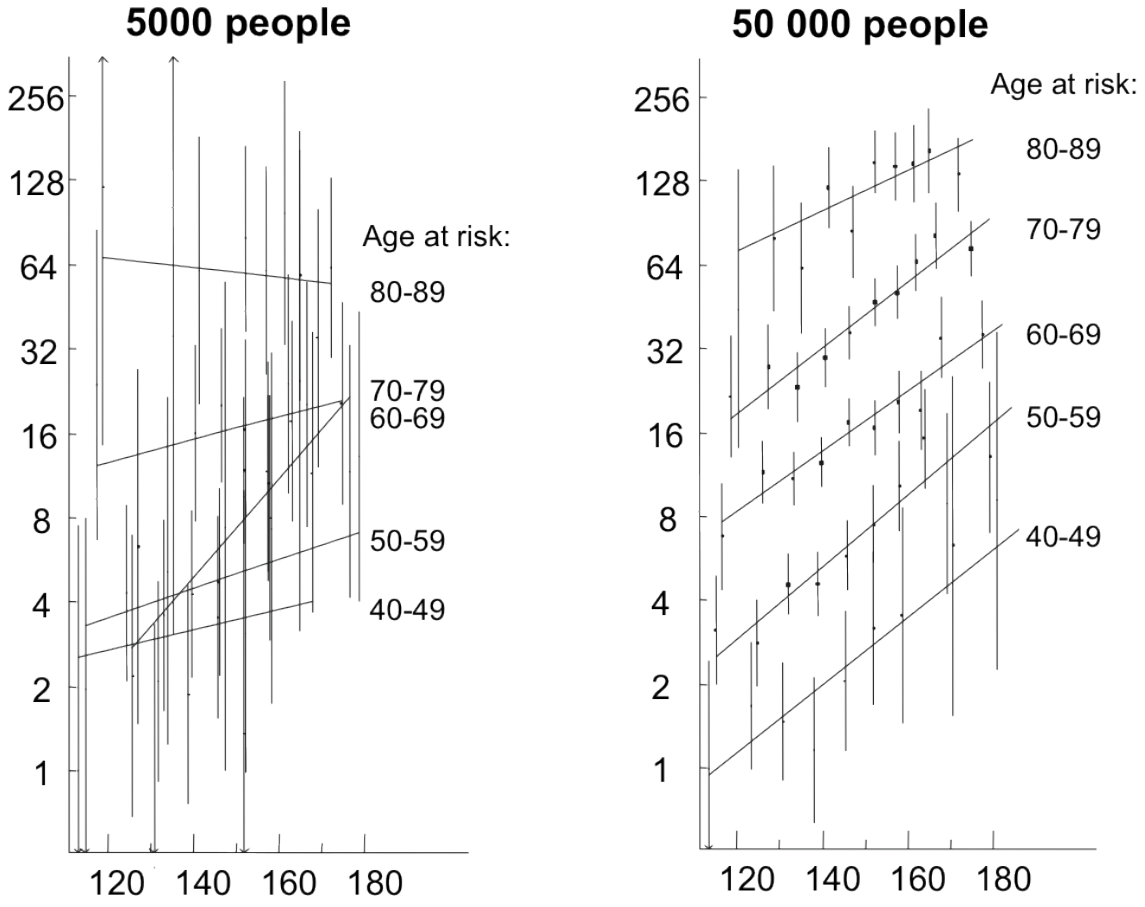
- The beauty of large numbers: separating signal from noise
- Systolic blood pressure vrs heart disease risk, data from UK biobank



- Graphs show Hazard ratio (96% CIs) vrs Usual SBP (mmHg)

Sample size: illustrating the value of large numbers

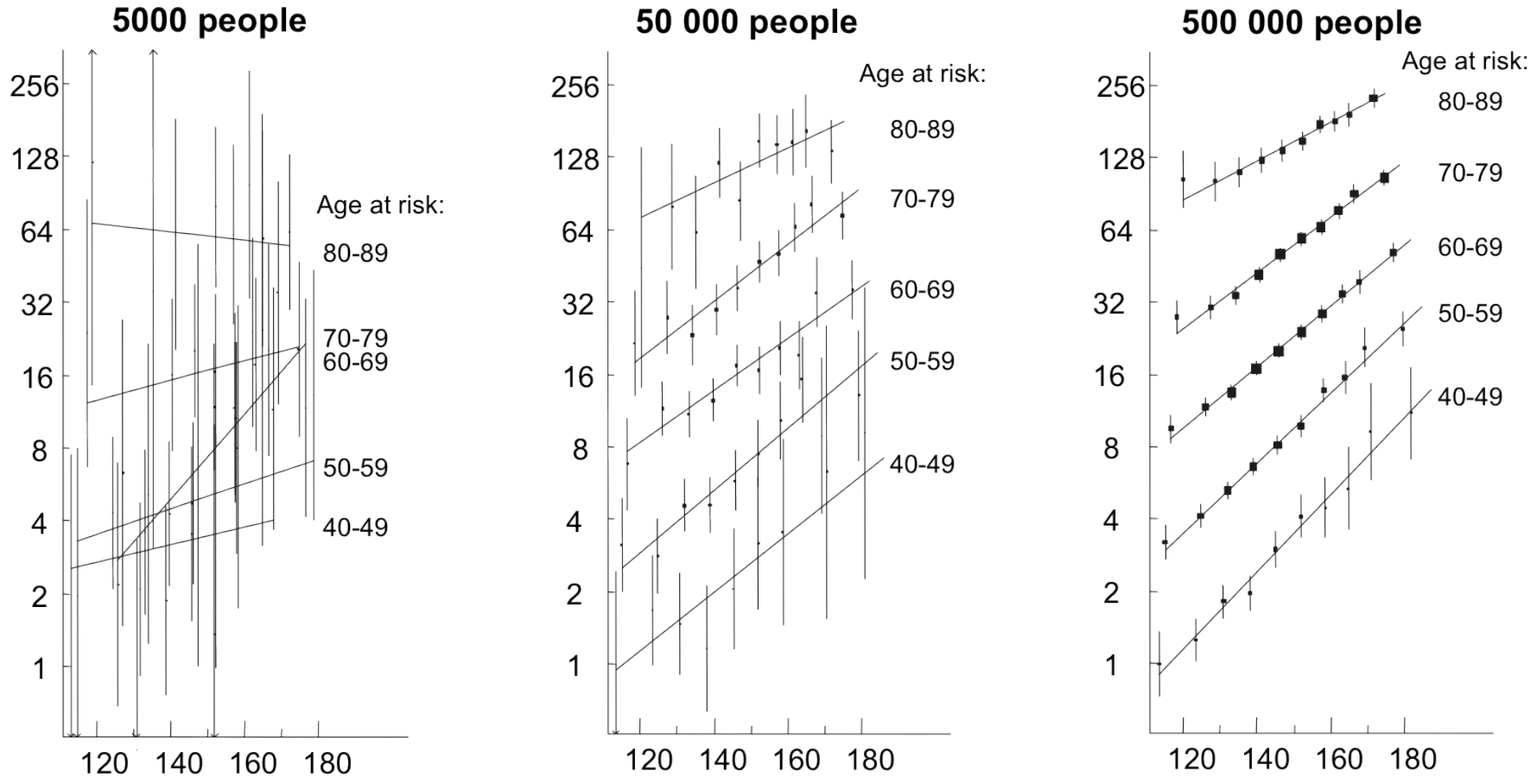
- The beauty of large numbers: separating signal from noise
- Systolic blood pressure vrs heart disease risk, data from UK biobank



- Graphs show Hazard ratio (96% CIs) vrs Usual SBP (mmHg)

Sample size: illustrating the value of large numbers

- The beauty of large numbers: separating signal from noise
- Systolic blood pressure vrs heart disease risk, data from UK biobank

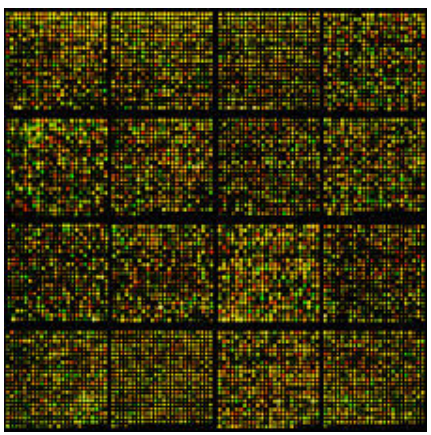


- Graphs show Hazard ratio (96% CIs) vrs Usual SBP (mmHg)

"Large p", a tale (or tail?) from statistical genetics

- As statisticians started to engage with high-throughput genomics in the early 2000s they quickly encountered issues of dimensionality in the number of measurements, p
  - note: to a statistician n is always the sample size and p the number of measurements on each sample[2]
- But crucially, at this point in time, to most statisticians the data was just an "X matrix" – with little context

Data:



$\Longrightarrow$

Math:

$$n(X)^p$$

$\Longrightarrow$

Description:

X is a large data matrix conceptually storable in Excel, with a small number of samples (n rows) and a large number of measurements (p columns) on each subject

[2]Statistical joke: "Let ϵ denote a large negative constant"

First generation genomics

- At this point in time the data was still highly structured with homogeneous measurements
  - the canonical example being microarrays
- Principle task was how to recover low-dimensional "signal" (true associations) embedded in high-dimensional "noise"
- This drove a wave of new statistical theory and new methods (and reinvigorated some older methods)
  - Sparsity
  - LASSO
  - Bayesian variable selection priors
  - False Discovery Rates (FDR)
  - Family Wise Error Rates (FWE)

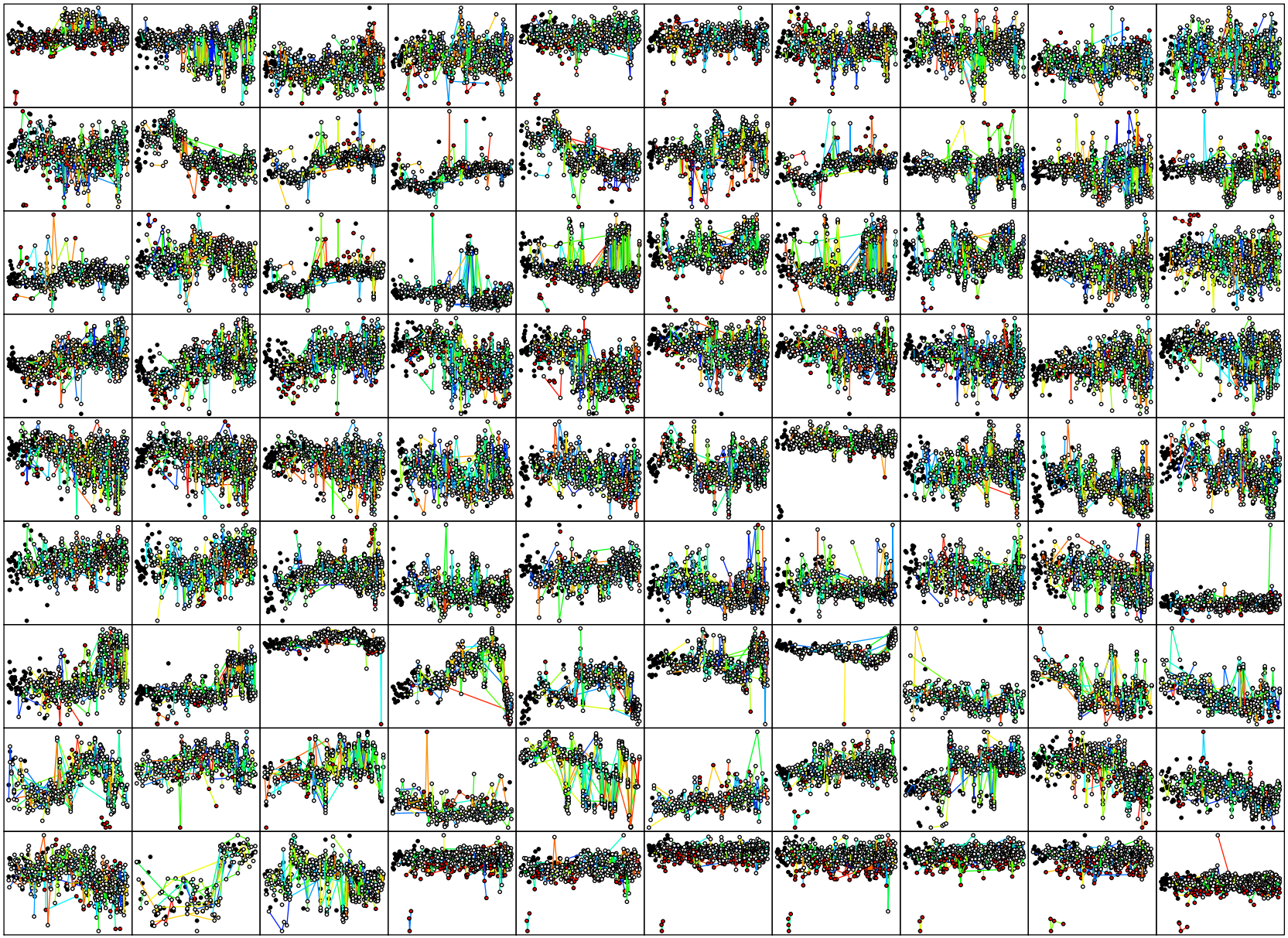
  $\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots$

Modern biomedical data analysis challenges

- In recent times there has been a rapid move towards large-scale data integration from multiple data sources of multi-modalities
  - combining images, eHealth-Records, genomics, biobank data, ...
- These are very different data objects
- This introduces additional complexities to the modelling process
- No longer can we think of the data as homogeneous or exchangeable (it's entirely inappropriate to think of this as representable as an X-matrix)
- To make progress we need carefully tailored contextual models, embedded in an engineering design process

- A good example being the International Mouse Phenotyping Consortium (http://www.mousephenotype.org/)
  - A 10 year study to systematically characterise the functional consequences of 20,000 genes in the mouse genome
  - Recording over 1500 measurements per mouse (leading to around 690 phenotypes), around 7 mice per knockout ($\times$ 2 sexes) and matched controls
  - IMPC will deliver complex multivariate measurements on around 560,000 mice $\times$ 690 dependent phenotypes across 8 Centres

....35,000,000 longitudinal measurements from 700 outcomes.....



Statistical analysis

- Traditional statistical approaches simply won't scale to this
- Moreover the design criteria involve a number of potentially competing factors
- That's why data sets such as these demands engineering design principles
  - involving citeria such as robustness, heterogeneity of measurements, incorporation of domain knowledge, ...etc...., all impact on the optimal statistical approach
- You can't simply run this through a computer program
  - "the data go in at one end and the results come out the other, untouched by human thought" (Doug Altman)

Canonical design criteria

- We can highlight eight common design criteria that impact on modern ("big-data")[3] statistical analyses
- These features help shape the appropriateness of the analysis strategy and models employed
- I'm sure there are more (I'd be keen to hear your opinions)

[3] It's interesting to characterise features of modern data analysis and highlight how genomics differs from many other "big-data" domains. Note: in big-data people talk about the three "V's": Volume of data, Variety of data and Velocity of data (speed it accumulates)

Eight design criteria impacting on statistical models in big-data analysis,....

1. n, p – data size (Volume)
   - the raw data size

Eight design criteria impacting on statistical models in big-data analysis,....

1. n, p – data size (Volume)
   - the raw data size

2. Heterogeneity (Variety) of data
   - images, DNA, RNA, eHR, ...
   - and whether it accumulates (Velocity)

Eight design criteria impacting on statistical models in big-data analysis,....

1. n, p – data size (Volume)
   - the raw data size

2. Heterogeneity (Variety) of data
   - images, DNA, RNA, eHR, ...
   - and whether it accumulates (Velocity)

3. Response-time of model
   - how quickly do we need the model to respond

Eight design criteria impacting on statistical models in big-data analysis,....

1. n, p – data size (Volume)
   - the raw data size

2. Heterogeneity (Variety) of data
   - images, DNA, RNA, eHR, ...
   - and whether it accumulates (Velocity)

3. Response-time of model
   - how quickly do we need the model to respond

4. Security (privacy)
   - how and where data is stored
   - taking algorithms to data, or data to algorithms

Eight design criteria impacting on statistical models in big-data analysis,....

1. n, p – data size (Volume)
   - the raw data size
2. Heterogeneity (Variety) of data
   - images, DNA, RNA, eHR, ...
   - and whether it accumulates (Velocity)
3. Response-time of model
   - how quickly do we need the model to respond
4. Security (privacy)
   - how and where data is stored
   - taking algorithms to data, or data to algorithms
5. Auditable
   - do we need explicit predictions

Eight design criteria impacting on statistical models in big-data analysis,....

1. n, p – data size (Volume)
   - the raw data size
2. Heterogeneity (Variety) of data
   - images, DNA, RNA, eHR, ...
   - and whether it accumulates (Velocity)
3. Response-time of model
   - how quickly do we need the model to respond
4. Security (privacy)
   - how and where data is stored
   - taking algorithms to data, or data to algorithms
5. Auditable
   - do we need explicit predictions
6. Calibrated (probabilistic)
   - do we require a risk score?
   - "do 80% of those we predict at 80% risk get the outcome?"

Eight design criteria impacting on statistical models in big-data analysis,....

1. n, p – data size (Volume)
   - the raw data size
2. Heterogeneity (Variety) of data
   - images, DNA, RNA, eHR, ...
   - and whether it accumulates (Velocity)
3. Response-time of model
   - how quickly do we need the model to respond
4. Security (privacy)
   - how and where data is stored
   - taking algorithms to data, or data to algorithms
5. Auditable
   - do we need explicit predictions
6. Calibrated (probabilistic)
   - do we require a risk score?
   - "do 80% of those we predict at 80% risk get the outcome?"
7. Domain Knowledge
   - known structure (priors)?

Eight design criteria impacting on statistical models in big-data analysis,....

1. n, p – data size (Volume)
   - the raw data size
2. Heterogeneity (Variety) of data
   - images, DNA, RNA, eHR, ...
   - and whether it accumulates (Velocity)
3. Response-time of model
   - how quickly do we need the model to respond
4. Security (privacy)
   - how and where data is stored
   - taking algorithms to data, or data to algorithms
5. Auditable
   - do we need explicit predictions
6. Calibrated (probabilistic)
   - do we require a risk score?
   - "do 80% of those we predict at 80% risk get the outcome?"
7. Domain Knowledge
   - known structure (priors)?
8. Robustness and reproducibility
   - stability of the domain – concept drift?
   - study population – vrs – predictive population?

Stage 2: Understand the study design – how, when and where the data comes from

- You need to understand how the data came about
  - experiment or observation?
  - over what time period?
  - using what equipment?
    - try to go and see the set up
  - is it a mix of studies?
- Define the reference population – who/what does the analysis relate to?
  - is this different to the envisaged application population?
- Are there missing data? or rather there *will* be missing data, how will they be handled?
  - if so, is it missing at random? or structured missingness?
  - see Little and Rubin (2002)
  - .....to be covered in depth on Wednesday....

Stage 3: Visualisation and exploratory analysis

- The starting point of ALL good statistical data analysis begins with graphical plots and summary statistics of the data
- ALWAYS, ALWAYS, ALWAYS, PLOT YOUR DATA!!!
- Why?

Stage 3: Graphical Excellence

"Graphics reveal data, communicate complex ideas and dependencies with clarity, precision and efficiency"

- Edward Tufte: *The Visual Display of Quantitative Information*

Stage 3: Graphical Excellence

- Excellent graphics:
  - Display information in an unbiased and uncluttered manner
  - Have a clear purpose
  - Reveal interesting and pertinent features of the data distribution - and highlight outliers
  - Encourage the viewer to think about the data generating mechanisms (help abductive inference – toward tentative model formulation)

Moreover:

- Graphical plots and summary stats provide a feel for the variation in the data
- They can also highlight unusual results, measurement errors, outliers
  - Such features can severely distort your results if left unchecked!
  - Many formal tests assume that the data follows a certain pattern (a probability distribution such as Normal), if these assumptions are invalid the results will be completely misleading
  - Confidence in these assumptions can be gained through plotting the data

Stage 4 and 5: The analysis plan and tentative model fitting

- Document an initial analysis plan and success criteria
  - what models will be tested, and how will they be judged?
- Tentatively fit some simple models, explore transformations
- Decide on a class of models to adopt, over a range of complexities if possible

Stage 6: Model refinement

- Assess the current method against the design citeria and analysis plan
- Increase the complexity of the model, if warranted by stated design criteria
- Consider the use of robust methods
  - see Huber (2011)
- Clearly note any changes you've made to the design criteria, for example, from lack-of-fit of some class of models

Stage 7: Model assessment

- Check of model fit
  - are the distributional assumptions supported?
- Investigate out-of-sample prediction performance
  - posterior predictive analysis – e.g. Gelman & Hill (2006)
- Residual analysis looking for influential observations with high leverage
- shake the data and look for stability
  - use the bootstrap – arguably the most important method in applied statistics
  - sub-sample and refit
  - Never use just one data split
- shake the models
  - does a small change in the model lead to a big change in conclusions – do robust methods alter things?
  - systematically perturb the model in the worst possible direction and look for robustness of conclusions – see Watson & Holmes (Stat. Sci., 2016)
- Check for "concept drift" if data is indexed by time

Stage 8: Validation and Reporting

- Try your hardest to break any interesting associations you find!
- Try to find truly independent (never seen) data – preferably from another study, and from another lab, to validate on
- What were the deficiencies in the study design – and the way the data was obtained?
  - conceptualise the idealised experiment and compare your data to this. What conclusions are supported by your data? What are the major weaknesses of the design? – see interesting recent paper by Rubin (2008).
  - clearly specify the reference population and the scope of the model's application with caveats
- Entertain a causal model or causal association for the analysis
  - Careful thought in this respect will help in the interpretation of the results, pondering on unmeasured confounders, and weaknesses (or brittleness) in the data and study design. And do Not(!) fall into the trap of association $\equiv$ correlation

Stage 9: Documenting the analysis – Reproducibility

- Document and release all code; and provide scripts to reproduce all results and figures
- Your report should be self-contained, reproducible, refutable, and scientific
  - Lay out clearly how to reproduce your results
  - Link code to Figures and programs with example scripts on how to run them – including random number seeds if using stochastic computation
  - Provide all necessary information for people to refute your conclusions – be scientific
- Audit the final model against the initial design plan and initial objectives
  - what has been learnt from the learning process?
  - have design criteria and objectives changed over the course of the analysis (they almost certainly will have)?
  - if so, how? Be sure this has not inflated optimism in the conclusions

Some References:

- Cleveland, W. S. (1993) *Visualising data*. Hobart Press
  - good overview of graphical stats, perhaps a little old now
- Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* CRC press.
  - every statistician needs to understand the Bootstrap
  - see also Efron, B and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1), 36-48.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
  - good book on regression, advocating the Bayesian approach
- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. 2nd Ed.
  - the leading text on modern statistical methods and interface with machine learning
- Huber, P. J. (2011). *Robust statistics*. Springer Berlin Heidelberg.
  - a thorough overview of robust estimators and models

- Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data*. 2nd ed. John Wiley & Sons.
  - comprehensive book – see also book by Carpenter – how to deal with missing data is rarely taught but so important
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 808-840.
  - interesting idea to compare the idealised randomised experiment to the actual observation study (in the context of causal inference)
- Rubin, Donald B. & Little, Roderick J. A. (2002). *Statistical analysis with missing data (2nd ed.)*. New York: Wiley.
  - covering the foundations of analysis with missing data
- Savage, L, J. (1954). The Foundations of Statistics. Dover.
  - worth reading Chapter 1-8 for those fascinated with foundations of subjective probability (Bayesian stats)
- Tufte, E. (2001) *The Visual Display of Quantitative Information*. 2nd Edn. Graphics Press.
  - you will never look at a graph in the same way again
- Wainer, H. (1984). *How to display data badly*. American Statistician. Vol. 38, No. 2
  - .....because you don't want to!
- Watson, J., & Holmes, C. (2016). Approximate Models and Robust Decisions. *Statistical Science*
  - a discussion paper, and developments, on the formal analysis of model misspecification in decision making