Developing your research career as a health data scientist: a 360 degree view

IDPLN Conference, September 2018, Banff, Canada
Professor Colin McCowan, University of Glasgow
Professor George Moulton, University of Manchester
Objectives

1) To identify key skills that are required by a data scientist
2) To assess your proficiency in the core skills and identify potential training needs
3) To align core skills with broader academic researcher development (VITAE)
4) To explore and develop an understanding of the benefits of team data science
5) To create a personalised career development plan
Objectives

• What is a data scientist?
• How do you stack up?
• How does a data scientist develop in an academic environment?
• Why should we use team data science?
• How can I develop in my field?
Agenda

1:00-1:15 – Welcome & Introduction

1:15-2:25 – Data Scientists in the academic environment

2:25-3:25 – Team Data Science

3:25-3:45 - Coffee Break

3:45-4:30 – Planning your development
George Moulton

• Professor of BioHealth Informatics & Education at UoM
• First Degree – BSc Molecular Biology and Biochemistry
• First academic job – Teaching Fellow
• Degrees – BSc, MSc, PhD
• Years in academia – 13
• Post-doc years – 13
Colin McCowan

• Professor of Health Informatics at UoG

• First Degree – Computer Science

• First academic job – Software Developer

• Degrees – BSc, MSc, MSc, PhD

• Years in academia – 25+
  • Post-doc years – 9+
Getting to know you - First Degree (name and Degree)

• Blue – Maths/Stats

• Yellow - Computing

• Red – Clinical or profession allied to health care

• Green - Other
Getting to know you – Current job role

• Blue – Statistician / Analyst

• Yellow - Programmer

• Red – Researcher

• Green - Clinician
Getting to know you - Experience

• Blue – current student

• Yellow – post-doctoral < 5 years

• Red – <5 years

• Green – 5 years +
Getting to know you - Pets

• Blue – Dog

• Yellow - Cat

• Red – None

• Green - Other
What is a data scientist?
MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS
- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE
- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g. R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with IaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION
- Able to engage with senior management
- Story telling skills
- Translate data driven insights into decisions and actions
- Visual art design
- R packages like ggplot2 or lattice
- Knowledge of any of visualization tools e.g. flare, D3.js, Tableau
SAS – What makes a great data scientist?

Largest group reported “traditional” traits (analytical, logical, technical) but strong communication and creativity skills also flagged as important.

Data scientists required to be technically proficient, mathematically agile, business savvy and good at communicating.

55 per cent of data scientists have < three years experience

more than a quarter are adapting their behaviours to fulfil roles that are not well matched to their skills or work personality profiles.
1. The Geeks 41%

The Geeks are the largest group in our sample and have the largest female membership of all the groups at 37 per cent. They have a naturally technical bias, strong logic and analytical skills. Essentially “black and white” thinkers, they like to speak plainly and stick to the point - don’t expect them to be moved by emotionally charged arguments. With their attention to detail and fondness for the rules, the Geeks are well suited to roles such as defining systems requirements, designing processes and programming.
2. The Gurus  11%

The next largest group, the Gurus, has a measure of reactive introversion, like the Geeks, which pre-disposes them to scientific and technical subjects. Yet they also display a diametrically opposite characteristic: the strong presence of proactive extroversion, including solid and often highly persuasive communications and social skills. The Gurus can play a very important role by using their enthusiasm, tact and diplomacy to promote the benefits of the data sciences to those holding the purse strings, or who have the authority to give projects the green light.
3. The Drivers  11%

The Drivers are proactive introverts: highly pragmatic individuals who use their determination and focus to realise their goals. Self-confident and results-oriented, they are ideal project managers and team leaders, who excel at prioritising, monitoring and driving projects to a successful conclusion.
4. The Crunchers 11%

This category is probably one of the least self-promoting groups. Strongly reactive - rather than proactive - personalities, the Crunchers like routine and constancy. They display high technical competence and consistency, making them superb in a range of technically-oriented support roles including data preparation and entry, statistical analysis, monitoring of incoming data and quality control.
5. The Deliverers  7%

Like the Drivers, these individuals are proactive and well suited to project and man management. This is also the group with the largest proportion of men at 80 per cent. However, the Deliverers also have a strong pre-disposition towards acquiring and/or applying technical skills. So, while they are capable of bringing focus and momentum to ensure project success, they are also likely to understand the finer technical details and devise solutions in much greater technical depth.
6. The Voices

The Voices are strong communicators with less apparent detailed technical knowledge than the Gurus. The presence of this group suggests a strong demand for natural promoters who have the ability to generate enthusiasm for the potential of big data and the data sciences at a conceptual level - rather than the practical or technical level. The Voices are strongly valued for their positive outlook, and may be engaged in presenting the results of big data projects as well as supporting their implementation.
7. Other Personalities 13%

A smaller number of respondents displayed a range of other traits. Phase two of our study will give us a bigger sample to analyse, when we'll be able to learn more about how these traits are emerging in the data science community.

- **The Ground Breakers:** offer new approaches, new methods and new possibilities, drawn from a mix of inspiration and dogged logical thinking. Roles include: system design and algorithm development.

- **The Seekers:** combine superb technical knowledge and understanding with inquisitiveness and a drive to find solutions. Roles include: research.

- **The Teachers:** skilled at imparting knowledge and inspiring others to want to learn. Roles include: training and mentoring.

- **The Lynchpins:** important team players who may not have a depth of technical knowledge but provide essential support services. Roles include: co-ordination and administration.
Researcher Development Framework

- Vitae is a development framework that focuses on researcher skills
- Realise the maximum potential at each stage of a researcher career
- Developed originally by a group of HEIs and Research Councils
- It has become the UK standard
- Sub-divide generic research skills into 4 domains

www.vitae.ac.uk
Engagement, influence and impact

The knowledge and skills to work with others and ensure the wider impact of research.

Domain D

Knowledge and intellectual abilities

The knowledge, intellectual abilities and techniques to do research.

Domain A

Research governance and organisation

The knowledge of the standards, requirements and professionalism to do research.

Domain C

Personal effectiveness

The personal qualities and approach to be an effective researcher.

Domain B

Self-management

Professional and career development

Finance, funding and resources

Research management

Communication and dissemination

Working with others
Phases 1 - 5 of the Framework
Vitae framework & Health Data Science

• Looking at Vitae framework, which categories do you think would map onto our understanding of health data science?

• Once you have identified the categories, think about the knowledge and behaviour you will need to operate in health data science?
Team Science

• Team Science is:

“collaborative effort to address a scientific challenge that leverages the strengths and expertise of professionals trained in different fields. ...[] coordinated teams of investigators with diverse skills and knowledge may be especially helpful for studies of complex social problems with multiple causes.”

Data Science Executive

• Ruth Dundas
• Claudia Geue
• Angela Henderson
• Jim Lewsey
• Danny Mackay
To build on our international reputation for excellence in using routine data to support a number of different methodologies including intervention studies, to evaluate public health, for epidemiology and to explore health inequalities.
Areas of work

• Data & Analytical Platforms

• Methodology Development

• Sharing resources

• Dissemination / Communication

• Training
Team Data Science

- Prof John Ainsworth
- Prof Georgina Moulton
- Prof Tjeerd Van Staa
- Prof Niels Peek
- Dr Matthew Sperrin
- Dr Nophar Geifman
- Dr Glen Martin

- Plus, 3 Technical Programme Managers
- 10 Software Engineers
- 3 Project/Programme Managers
- 2 Communication Managers
- >150 researchers across the University who work at the interface of biology and/or medicine and informatics/data science
Pull-through from Research Themes, Industry and Health System Needs
Manchester (Cancer, Musculoskeletal, Respiratory, Mental Health, Skin, Hearing…) + External

Managed Integrative Informatics and Data Sciences

**Digital Phenomarker Factories**
- MRC Farr @ Health eResearch Centre
- DH Connected Health Cities Hub
- NIHR Patient Safety Centre
- NIHR Manchester BRC Imaging Science
- Arthritis Research UK Epidemiology...

**Biomarker Factories**
- MRC/Stoller Biomarker Centre
- MRC Molecular Pathology Innovation Centre
- CRUK Computational Biology
- Genomic Sciences/Medicine
- NIHR Manchester BRC
- Centre for Musculoskeletal Research...

**Digital Methods Factories**
- Data Science Institute
- Mathematics in the Life Sciences
- MRC Farr @ Health eResearch Centre
- Bio-health Informatics in Computer Science
- Imaging Science
- ESRC National Centre for Research Methods

Cross-disciplinary Training & Capacity Building

**Endotype Discovery**
- Scalable Epidemiology
- Efficient Trials
- Precision Medicine
- Citizen Driven Health(care)
- Learning Health Systems

Analytic Commons
(Research; Regulatory; NHS; Civic)
- GM NHS Digital Collaborative, GM Connect, Northern Connected Health Cities, NICE

Algorithm 'Spin-in Laboratory' for External Partnerships
Team Health Data Science

• A community-based longitudinal study on adherence to endocrine therapy and cancer outcomes
• Abstract of research

• Aims & Objectives: To investigate whether patients in the community, prescribed adjuvant endocrine therapy for breast cancer are adherent to their medication and whether adherence influences recurrence and survival. Secondary objectives are to describe the pattern of endocrine therapy within a community population, the characteristics of patients with poor adherence, to examine differences in outcomes between patients receiving aromatase inhibitors and those receiving tamoxifen and to perform an economic evaluation of the benefit of adherence to medication.

• Relevance: Published studies suggest that around 25% of women have low adherence to adjuvant aromatase inhibitors or tamoxifen but whether this influences outcomes is unknown. Recent work by the applicants showed low adherence to tamoxifen increased the risk of all-cause mortality.

• Patients & Methods: An established multidisciplinary team will perform a retrospective cohort study of all women with incident breast cancer in the Tayside region of Scotland between 1993 and 2008. Encashed prescription records will be used to calculate adherence to adjuvant endocrine therapy. Survival analysis will determine the effect of adherence on all-cause and breast cancer mortality and recurrence. Economic modelling will explore the cost-effectiveness of adherence to medication.

• Length of Study: 2 years
• Identify Project Staffing
Team Health Data Science

<table>
<thead>
<tr>
<th>Role</th>
<th>Data science domains used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics Supervisor</td>
<td>Analysis, Communication, Context</td>
</tr>
</tbody>
</table>
Team Health Data Science

• Research Assistant 100%FTE (24 months)
• Data Manager 50%FTE (6 months)
• Health Economist 100%FTE (6 months)
• Project Supervision
• Data Science Supervisor 10%FTE (24 months)
• Clinical Professor 2.5%FTE (24 months)
• Stats Professor 2.5%FTE (24 months)
• Economics Supervisor 22.5%FTE (24 months)
Developing a PDP

1. Identify
   - What do I want to learn or start doing?
2. Plan
   - What do I have to do?
3. Action
   - What support & resources will I need?
4. Record
   - How will I measure success?
5. Review
   - Target date for a review?
PDP Identify

• Need to identify training needs and prioritise them
  • SWOT analysis
  • 5 years Time – actions & outputs
  • List training needs - up to 5
  • Prioritise and pick key element you want to work on first
PDP Plan

• What is your new goal/development need?
  • Is it clearly defined?
  • Is it measurable?

• Consult your supervisor / mentor to discuss this
  • "O wad some Power the giftie gie us To see oursels as ithers see us!“ R Burns

• Formal training
• Informal aspect
• Opportunities to integrate in working practice
  • Barriers
  • Facilitators
• Timescale
• Success criteria
PDP Plan

• SMART Goals / Objectives
  • Specific
  • Measurable
  • Achievable
  • Relevant
  • Time bound
PDP Action

• How do you integrate this into your working life?
  • What resources do you need?
  • What support do you need?
  • Will you change the way you work?
    • What do you need to start doing?
    • What do you need to stop doing?
• Think about small steps
PDP Record

• Keep a training record
• Demonstrate and document when you use new techniques
• Reflective journals / reflective time
• Ask others how things went
PDP Review

• Tie in with other processes

• Don’t leave it to the end – do it often and be regular

• Don’t be afraid of failure – learn from it

• Be honest!
Objectives

• What is a data scientist?
• How do you stack up?
• How does a data scientist develop in an academic environment?
• Why should we use team data science?
• How can I develop in my field?