INFERENTIAL INTERESTS

ECSSN MAGAZINE VOL 3

ECSSC IS HERE!

PETER PAOLA MINH-NGOC KENDRA HELENA

HELLO ECSSC 2021

EDITORS' LETTER

We are so excited the Early Career and Student Statisticians Conference 2021 is here! Kicking off this past weekend, we had short courses in convex optimization and shape analysis. Now onto today - turn to page 4 to 6 to check out the speakers' bios and timetable. We have so many great keynote speakers, as well as information sessions with the BCA (Biostatistics Collaboration of Australia), Survey Design, ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), and the Australian Bureau of Statistics (ABS). Our interviews with the keynote speakers continues this issue with Kendra Vant from Xero, and we also have Andrew Forbes from Monash, who was a co-founder of the BCA.

Make sure you take a look at the 5 tips from our Chair Janan on doing a PhD, and our Secretary Ben has written an article on attending the AMSI Winter School - he learnt a lot on Bayesian statistics and neural networks... and is still reeling from the from the onslaught of information put together by the truly awesome AMSI team. Jordan reviews book "The Art of Statistics" to take a good hard look at the basics... I have ordered my copy already!

Hope you see you at the conference,

Cat and Split

Peter Taylor

Redmond Barry Professor
University of Melbourne

Peter Taylor received a BSc(Hons) and a PhD in Applied Mathematics from the University of Adelaide in 1980 and 1987 respectively. In between, he spent time working for the Australian Public Service in Canberra. After periods at the Universities of Western Australia and Adelaide, he moved at the beginning of 2002 to the University of Melbourne.

In January 2003, he took up a position as the inaugural Professor of Operations Research. He was Head of the Department of Mathematics and Statistics from 2005 until 2010.

Peter's research interests lie in the elds of stochastic modelling and applied probability, with particular emphasis on applications in telecommunications, biological modelling, healthcare, economics and disaster management.

Peter is the Editor-in-Chief of the Applied Probability Trust journals Journal of Applied Probability and Advances in Applied Probability and was the Editor-in-Chief of Stochastic Models between 2002 and 2018. He is also a member of the editorial board of Queueing Systems.

From 2006 to 2008, Peter was Chair of the Australia and New Zealand Division of Industrial and Applied Mathematics (ANZIAM), and from 2010 to 2012 he was the President of the Australian Mathematical Society. In 2013 he was awarded a Laureate Fellowship by the Australian Research Council, and in 2016 he became Director of the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

In 2017, he was awarded the Ren Potts Medal by the Australian Society for Operations Research, in 2018 the George Szekeres Medal by the Australian Mathematical Society and in 2019 the ANZIAM Medal.

ECSSC SPEAKERS

Paola Oliva-Altamirano

Principal Data Scientist. Innovation Lab,
Our Community/SmartyGrants

Paola Oliva Altamirano is a principal data scientist at Innovation Lab, Our Community/SmartyGrants. A research scientist trained in astrophysics, she is currently working on data science projects to help Australian not-for-profit organisations. As a data scientist with the SmartyGrants Innovation Lab, Paola designs algorithms to improve our understanding of social sector data, with the goal being to facilitate human-centred AI solutions.

Paola graduated from Physics in Honduras and later did a PhD and Post-Doc in Astrophysics at Swinburne University of Technology. She transitioned to do Data Science with SmartyGrants in 2018. In 2016 she co-funded the Organisation "Astrophysics in Central America and the Caribbean (Alpha-Cen)" to support students in developing countries pursue careers in science. She currently is Alpha-Cen's Vice-President.

Kendra Vant

EGM - Data, Xero

Kendra is an industry leader in building data-driven products by harnessing emerging artificial intelligence and machine learning techniques to solve problems for businesses and industry globally.

She has had a rich and varied career working in New Zealand, Australia, the US and Malaysia, leading data and engineering teams at companies including SEEK, Telstra, Deloitte and now Xero. At Xero, she heads a globally distributed team of developers, machine learning specialists and data practitioners using emerging practices and technologies to make data work harder for small businesses and their advisors.

After doctoral research in experimental quantum physics at MIT and postdoctoral work in applied quantum computing at Los Alamos National Laboratory, Kendra worked in bespoke software development and then in generating business insights from data before focusing on applying machine learning to create personalised experiences in an increasingly connected and digital world. She gets her greatest satisfaction from working with smart people to solve difficult problems that have a positive impact on the world.

ECSSC SPEAKERS

Minh-Ngoc Tran

Associate Professor / University of Sydney

Minh-Ngoc is currently Associate Professor at the Discipline of Business Analytics, the University of Sydney Business School. He obtained a BSc and a MSc both in Mathematics from the Vietnam National University at Hanoi, and then a PhD in Statistics in 2012 from the National University of Singapore. He then worked as Postdoctoral Research Fellow at the University of New South Wales before joining the University of Sydney as a faculty in 2015.

Minh-Ngoc's main research interest is Bayesian computation with a special focus on Variational Bayes. He is also interested in promoting the use of modern Bayesian computation techniques in Cognitive Science, Consumer Behaviour and Financial Econometrics. Minh-Ngoc's research has been published in many top-tier statistical journals and his research has been well funded including three ARC grants.

Helena Jia

Executive Director, National and International Assessments / ETS

Helena Jia is the Executive Director of National and International Assessments at Educational Testing Service. She received her M.S. and Ph.D. in statistical science at Southern Methodist University in 2004 and in 2007, respectively.

Helena oversees operational psychometric work and research for the centers focused on educational surveys, including the United States National Assessment of Educational Progress (NAEP), the Programme for International Student Assessment (PISA) and the Programme for the International Assessment of Adult Competencies (PIAAC). She has extensive experience in design, analysis, score reporting of large-scale educational survey assessments. Helena's research interests focus on item-response theory latent regression models, linking of test scores in group-score assessment, adaptive testing, and survey sampling.

ECSSC SPEAKERS

SPONSORED BY

ACEMS
AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Australian Bureau of Statistics

BIO STATISTICS COLLABORATION OF AUSTRALIA

AMSI

sdas
Arbutus | QDA Miner | **Stata** | Stat/Transfer | WordStat

ECSSC TIMETABLE

Pre-Conference Schedule

Convex optimization for statistical and machine learning with CVXR

24/07/2021 11:00 am - 3:00 pm
ECSSC2021 Pre-Conference Workshop

Speaker: Anqi Fu

Optimisation plays an important role in fitting many statistical models. In this workshop, Dr Anqi Fu will introduce convex optimization using examples from ordinary least squares and penalized regression. This will be followed by a high-level description of CVXR, how it differs from other packages, and a discussion of the domain specific language that CVXR implements. We will show how CVXR works on different classes of problems, such as linear programs, quadratic programs, and semidefinite programs, and demonstrate its usage with a variety of examples. Finally, we will have a segment for potential developers in which we go over the nuts and bolts of adding new functions to CVXR's library.

Statistical Shape Analysis via Toplogical Data Analysis

25/07/2021 11:00 am - 3:00 pm
ECSSC2021 Pre-Conference Workshop

Speaker: Chul Moon

As modern data applications become complex in size and structure, identifying the underlying shape and structure has become of fundamental importance. The classical approaches such as dimension reduction are challenging for handling these applications. Topological data analysis (TDA) is a rapidly developing collection of methods that focuses on the "shape" of data. TDA can uncover the underlying low-dimensional geometric and topological structures from high-dimensional datasets. TDA has been successfully applied to various areas, including biology, network data, material science, and geology, in recent years. The goal of the lecture is to introduce novel TDA methods that can capture geometric or topological information of data and make statistical inferences. Covered by Assistant Professor Chul Moon, this lecture aims to familiarize these new methods along with their applications to various types of data.

ECSSC TIMETABLE

Conference Schedule

Opening Speech

26/07/2020 11:00 am - 11:15 am
ECSSC2021 Main Conference

Presentation Session 1

26/07/2020 11:15 am - 12:50 am
ECSSC2021 Main Conference

Speaker 1: Michael James Leach

(20 min)
Data visualisation can reveal exposure misclassification in case-crossover studies of medication safety

Speaker 2: Elena Tartaglia

(20 min)
Understanding the role of causal inference from observational datasets in developing government policy

Speaker 3: Atousa Ghahramani

(5 min)
Social media intervention for raising awareness of cardiovascular diseases in the female population of Australia

Speaker 4: Vanessa Pac Soo

(5 min)
Between-centre differences in overall patient outcomes and in trial treatment effects in multicentre perioperative trials

Speaker 5: Raphael Udeh

(5 min)
Calprotectin Metadata in COVID-19

Speaker 6: Ravindi Nanayakkara

(20 min)
Probing Multifractionality of Cosmic Microwave Background Radiation

Speaker 7: Taya Annabelle Collyer

(20 min)
What is statistical expertise?

ECSSC TIMETABLE

Conference Schedule

Survey Design Information Session

26/07/2020 1:00 pm - 1:45 pm
ECSSC2021 Main Conference

Modelling the Bitcoin blockchain: what can probability and statistics teach us?

26/07/2020 2:00 pm - 3:00 pm
ECSSC2021 Main Conference

Speaker: Peter Taylor

In 2009 the pseudononymous Satoshi Nakamoto published a short paper on the Internet, together with accompanying software, that proposed an `electronic equivalent of cash' called Bitcoin. At its most basic level, Bitcoin is a payment system where transactions are verified and stored in a distributed data structure called the blockchain. The Bitcoin system allows electronic transfer of funds without the presence of a trusted third party. It achieves this by making it `very hard work' to create the payment record, so that it is not computationally-feasible for a malicious player to repudiate a transaction and create a forward history with the transaction deleted. The Nakamoto paper contained a simple model used to show that the above-mentioned malicious player would be very unlikely to succeed. Unfortunately, this calculation contained an error, which I shall quickly discuss and show how to correct. As its name suggests, the blockchain is comprised of discrete blocks. Blocks are added to the blockchain by "miners" working across a distributed peer-to-peer network to solve a computationally difficult problem. With reference to historical data, I shall describe some models for the block mining process. I shall finish with some brief comments about how stochastic modelling can be used to address the current concerns that the transaction processing rate of the Bitcoin system is not high enough.

ECSSC TIMETABLE

Conference Schedule

Welcome Event

26/07/2020 3:00 pm - 4:00 pm
ECSSC2021 Main Conference

Presentation Session 2

27/07/2020 11:00 am - 12:30 pm
ECSSC2021 Main Conference

Speaker 1: Chathuri Lakshika Samarasekara

(20 min)
On the Estimation of the Resource Selection Probability Functions from Presence Background Data

Speaker 2: Shih Ching Fu

(20 min)
Comparison of Spatial Models for Analysing On-farm Strip Experiments based on Geographically-Weighted Regression and Kriging

Speaker 3: Josh Jacobson

(5 min)
Multivariate Spatial-Dependence Modelling of Satellite Data

Speaker 4: Jason Whyte

(5 min)
Exploiting scientific publication trends for chemical risk management

Speaker 5: Udani Wijewardhana

(20 min)
Statistical modelling with citizen science data for local shorebirds on the Mornington Peninsula

Speaker 6: John Samuel Warmenhoven

(20 min)
A non-conventional entry into the world of statistics...

ECSSC TIMETABLE

Conference Schedule

BCA Information Session

27/07/2020 12:20 pm - 1:30 pm
ECSSC2021 Main Conference

Metrics in the wild! How to deal with biases when building auto-classification systems

27/07/2020 2:00 pm - 3:00 pm
ECSSC2021 Main Conference

Speaker: Paola Oliva-Altamirano

When designing auto-classification systems we often trust standard metrics in-built in Machine Learning models. How much do you trust them? Or in other words, to which extent do these scores reflect the success of your algorithm? Knowing your data, the decisions that your model will influence, and the role that scoring plays in enhancing or mitigating biases should be essential for Statistic practitioners and product builders. In this talk I will share our lessons learned when building classifiers in the social sector, the biases we have encountered in multilabel text classifications, and our constant battle to design ethical, human centre products.

Virtual Pub

27/07/2020 3:00 pm - 3:30 pm
ECSSC2021 Main Conference

Join the virtual pub to continue discussions with your fellow statisticians.

ECSSC TIMETABLE

Conference Schedule

Presentation Session 3

28/07/2020 11:00 am - 12:30 am

ECSSC2021 Main Conference

Speaker 1: Xu Ning

(20 min)

Penalised Quasi-Likelihood Asymptotics in Generalised Linear Mixed Models

Speaker 2: Muzhi Zhao

(20 min)

Estimation of the Cure Rate for Distributions in the Gumbel Maximum Domain of Attraction Under Insufficient Follow-up

Speaker 3: Phillip Oluwatobi Awodutire

(5 min)

A new two-parameter distribution: Properties, Regression Modelling and Applications to COVID 19 data

Speaker 4: Zhi Yang Tho

(5 min)

Joint Mean and Correlation Regression Modelling for Multivariate Data

Speaker 5: Catriona Croton

(20 min)

Communciating with Clinical Collaborators: a Vet's POV

Speaker 6: Cameron Patrick

(20 min)

Tales from the trenches of statistical consulting: five tips for early career statistical consultants

Round Table Discussion

28/07/2020 12:30 pm - 1:30 pm

ECSSC2021 Main Conference

ECSSC TIMETABLE

Conference Schedule

Career Panel

28/07/2020 1:30 pm - 3:00 pm
ECSSC2021 Main Conference

Virtual Pub

28/07/2020 3:00 pm - 3:30 pm
ECSSC2021 Main Conference

Join the virtual pub to continue discussions with your fellow statisticians.

Bayesian computation: why/when Variational Bayes, not MCMC or SMC?

29/07/2020 11:00 am - 12:00 pm
ECSSC2021 Main Conference

Speaker: Minh-Ngoc Tran

Bayesian inference has been increasingly used in statistics and related areas as a principled and convenient tool for reasoning with uncertainty. Bayesian computation is often a challenging task and modern applications of Bayesian inference, such as Bayesian deep learning, have been called for more scalable Bayesian computation techniques. In this talk, we will give a quick introduction to Variational Bayes for scalable Bayesian inference. We then provide a general discussion on its pros and cons, recent advances and some potential research directions.

ECSSC TIMETABLE

Conference Schedule

ABS Information Session

29/07/2020 12:00 pm - 1:30 pm

ECSSC2021 Main Conference

Presentation Session 4: Bayesian Statistics

29/07/2020 1:30 pm - 3:00 pm

ECSSC2021 Main Conference

Speaker 1: David Allen

(20 min)

Lessons learned from the development and implementation of a Bayesian MIMIC model to predict hearing aid benefit

Speaker 2: Parinaz Mehipour

(20 min)

Bayesian Within-host Modelling of Red Blood Cell Dynamics and Primaquine-induced Haemolysis in G6PD Deficiency

Speaker 3: Gizem Ashraf

(5 min)

Sport-related Eye Trauma Study (SETS): Five-year audit of sports-related eye injuries at a tertiary eye hospital in Australia: 2015-2020

Speaker 4: Puxue Qiao

(5 min)

A bayesian model for inferring intratumour heterogeneity in copy number variation

Speaker 5: Luca Maestrini

(5 min)

Variational approximations for structural equation models

Speaker 6: David Warne

(20 min)

Using model approximations to accelerate Bayesian computation

Speaker 7: Ben Harrap

(20 min)

Things I Didn't Learn at University

ECSSC TIMETABLE

Conference Schedule

Virtual Pub

29/07/2020 3:00 pm - 3:30 pm

ECSSC2021 Main Conference

Join the virtual pub to continue discussions with your fellow statisticians.

GAMES NIGHT

29/07/2020 6:00 pm - 8:00 pm

ECSSC2021 Main Conference

Enjoy an evening of trivia with a twist. Come along to the games night and form a team with your new found colleagues to answer a series of questions and puzzles. Not only will the winnings have bragging rights, there will be prizes. Points will be awarded to the best team name, so start thinking now.

ECSSC TIMETABLE

Conference Schedule

Presentation Session 5

30/07/2020 11:00 am - 12:30 pm
ECSSC2021 Main Conference

Speaker 1: Chun Fung Kwok
(20 min)
Enhancing Markov-chain spatial simulation with web mapping technology

Speaker 2: Fan Cheng
(20 min)
Manifold Learning with Approximate Nearest Neighbors

Speaker 3: Oisin Fitzgerald
(5 min)
Deep learning for glycaemic control in the ICU

Speaker 4: Mohomed Amsar Mohomed Abraj
(5 min)
Modelling of Anisotropic Spatial Random Fields Using Mixture Copulas

Speaker 5: Jeffrey Pullin
(5 min)
Comparing statistical methods for identifying marker genes in single-cell RNA sequencing data

Speaker 6: Sharm Thuraisingam
(20 min)
Surviving a PhD with a toddler during a pandemic

ACEMS Information Session

30/07/2020 12:30 pm - 1:30 pm
ECSSC2021 Main Conference

ACEMS with
SPEAKER: PETER TAYLOR

AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS

ECSSC TIMETABLE

Conference Schedule

Commercial machine learning at scale - the joys and the pitfalls

30/07/2020 2:00 pm - 3:00 pm

ECSSC2021 Main Conference

Speaker: Kendra Vant

The art and science of applying machine learning techniques inside a for profit company is a world away from pursuing algorithm improvement and fundamental in a research setting. I will talk about the end to end process of building smart products within a SaaS company today.

Virtual Pub

30/07/2020 3:00 pm - 3:30 pm

ECSSC2021 Main Conference

Join the virtual pub to continue discussions with your fellow statisticians.

High School Engagement Day

Introduction to programming and statistical analysis workshop

31/07/2020 11:00 am - 2:30 pm

ECSSC2021 Main Conference

ECSSC TIMETABLE

Conference Schedule

Statistical Careers & Education Panel

31/07/2020 2:45 pm - 3:45 pm

ECSSC2021 Main Conference

Statistics, Psychometrics and Data Analytics in Educational Survey Assessments

01/08/2021 11:00 am - 12:00 pm

ECSSC2021 Main Conference

Speaker: Helena Jia

Educational survey assessments are used in measuring and monitoring learning and educational progress for youth as a group, rather than as individuals. This talk offers an overview of how practitioners applying methodologies in statistics, psychometrics and data analytics in the design of educational survey assessments, as well as in the estimation of group scores nationally and internationally. I will describe several statistical approaches that are involved in the group score estimation–item response theory models, latent regression models and missing data imputation/plausible values. Recent and future research and methodology development as survey assessment data collection transitioning from paper-based to digitally-based platform will be discussed as well.

Closing Ceremony

01/08/2020 12:00 pm - 1:00 pm

ECSSC2021 Main Conference

KENDRA VANT

Interview Catriona Croton

One of our keynote speakers for the ECSSC2021, Kendra, joins us to talk about applied data science in the business community. Kendra is the Executive General Manager of Data for Xero, where she leads the use of data, machine learning and augmented intelligence to give small business owners and their accountants data superpowers. Her undergraduate, masters and PhD degree were all in atomic physics, after which she made the leap into software development in the business community. She is focused on the practical process and so is multi-disciplinary; she was Principal Data Scientist at SEEK and has worked in banking, insurance, telecommunications, government and gaming.

On her pathway from academia to an industry career:

I don't know that there is a typical path. I do have an interesting background in that all my professional training is in physics – bachelors degree, masters degree and doctorate. I decided towards the end of my doctorate that as much as I thought the academic life was for me, it wasn't. I liked academia, but what I really loved was solving problems. The academic life changes quite a lot after postdoc to less time in the lab and solving problems, to more time writing grants and supervising research students. I made a really conscious decision to move from academia into industry to keep on solving hard problems.

I took advice that a good way to transition from an academic career to a non-academic career is through a postdoc. In retrospect, it worked really well for me from a life perspective, as I had one child during my graduate degree, and two more while I was a postdoc. By the time I had started my industry career, I had had my three children. It worked well to do something that wasn't a huge departure from what I had done while having babies, and then I had that uninterrupted stretch of time when I started something new.

On her first job in industry:

After my postdoc, I moved back to NZ with my three children under three. Leaving academia, I suffered a massive crisis of confidence as I had worked for a very long time to be good at a very narrow field. I realised that "No, you actually do have a really interesting set of skills that you have honed in a very narrow field". People who make the transition well from academia to industry realise that you have to work out what is transferrable to another field.

For my first job in industry, I was really lucky to be hired and supported by a NZ firm called Datacom, who build bespoke software. They were fantastic, family orientated company who said come join our company, we know that you know nothing about software development, but we will pay you while you learn, because we think you will learn fast. I think that companies do benefit if they are prepared to be flexible and support unusual people from different backgrounds because they get dot connecting thinkers who don't think exactly like others in the same role.

Datacom was a particularly great place to work as I gained a really good grounding in how software is built, from scratch, rather than how to modify an existing package. If you think about it, working in bespoke software build for four years at that time was an incredibly lucky break. As the new field of building machine learning systems and production emerged, I was one of the few people who could build both software and quantitatively complicated systems. And so, I became a specialist in building production machine learning systems.

"First consider a spherical cow..."

The link between atomic physics, machine learning and cows:

What is experimental atomic physics? It is planning over the medium to long term to build complicated systems to solve complicated problems. This is actually the skill set for how to build a machine learning at scale for industry. I really like the joke that engineers make about physicists, which is "First consider a spherical cow..." It is actually engineers being rude about physicists, because the first thing that a physicist tends to do is try to come up with a really simplified model of any system that you can actually solve. And that skill of trying to take a really complicated problem, and going "Cool, I can't solve that, but what is the closest approximation to that problem that I can solve?" Which is invoking that other wonderful quote: All models are wrong, but some are useful. Thinking about if that model is close enough to reality that is useful, that it is a good approximation. That is exactly what you do with building machine learning at scale. I cannot possibly really do what a human does, but if I can get close enough that it is a good approximation, then I can build something that is useful. That is what I have been doing for the past 7 or 8 years.

On a tool, technique or program that is useful:

The thing that I find fascinating and am using it a lot is a concept called Wardley maps. It is a way of thinking about the evolution of capabilities in systems and processes, that really focuses on the idea that companies, technologies and industries are never static. So you should always think about your product, service or company or the geographic area where you are living as constantly evolving. It is systems research thinking and is how everything evolves towards becoming a commodity.

An example is Kodak, who had incredible market spanning power in photographs and they didn't understand that the cell phone was going to completely disrupt their industry, because they had a static view. They needed to think what was the ultimate thing that people are trying to do and how does that evolve through time. Kodak thought people take photos. No - people actually wanted to be able to share their experience. In 1980 the way they did this was by taking a photo, and developing photos and showing it with a piece of paper. But as soon as people could post that on Instagram instead, they didn't want to develop photographs anymore. In Wardley maps, you take into the account the passage of time in what your customer needs are and how you need to adapt with the changing landscape of the market. It is absolutely fascinating.

On falling out of love with a technique:

A technique I have fallen out of love with is word embeddings, which was really famous in natural language processing about 6 or 7 years ago. You could train a word embedding model such that if you said the word "man" maps to the word "woman", then told the model "king" it would be able to tell you that the equivalent mapping was to "queen". As a technique, it can build a useful service, but what I changed my mind about is if it is flexible enough to have longevity. These models are not learning language; they are calculating statistical probability from past languages. And there are many reasons why that is not the same, but perhaps the most problematic is that if there is implicit bias in the language set that the models are trained on, then you are freezing that bias and replicating that bias into the future. If some of the language models were given "man" is to "doctor" as "woman" is to... then it would return "nurse". That is not the model getting it wrong – it is regurgitating what it is trained to do. The model is giving us an accurate reflection of reality, it is just that we don't like the reality.

On running a global team in COVID:

I have a global team in Toronto, Auckland, Wellington and Melbourne and so the biggest change for me with COVID is how I spend my time. Pre-COVID I spent an enormous amount of time travelling, but now I spend an enormous amount of time on Zoom. As long as everyone is on Zoom, the tyranny of distance is removed. If you have a team together in one room in Auckland and one lone person dialling in from Toronto, then is really hard. But if everyone is dialling in, then you forget where they are physically located. Xero has said 100% remote works well and all of our meetings are 100% remote. Everyone dials in, regardless if some of the same team are in the same office – they go to different offices and dial in.

The future for data scientists and statisticians:

The emphasis in industry will shift from who has got the coolest algorithm to who can actually implement the system. So if I was a young person training in those cool things like data science at the moment, I would be learning to program. Because 90% of the jobs that will exist in a decade will be in that. Not who can make the most precise model – those aren't actually valuable – but how to implement it. We are already seeing that shift in industry as research and development jobs are getting fewer and they are only in the elite parts of the biggest companies. People have this romantic idea of "Oh, I can get a job and go do R&D at a company". Do you know that about 10% of the engineers in Google get to work on R&D, and the rest work on bug fixes for the existing product set? While those R&D jobs do exist, they are very elite. More practically for the wider spectrum of people is "Can you make it a reality? Can you come up with a pragmatic end to end system that works, that is bias-free, that isn't going to get the company into legal trouble by implementing it. Can you think about ethics? Can you think about usability? Can you code? Can you practically participate in making the system a reality?" In fact, we no longer call our people data scientists - we call them applied scientists.

Thank you Kendra for the interview! Take a look at an interesting article Kendra wrote for Medium stories called "On Leaving Academia and How To Navigate a Career That Lasts a Lifetime". Make sure you come to her talk at the ECSSC 2021 – she is a great speaker!

Interview Catriona Croton

ANDREW FORBES

Professor Andrew Forbes is one of the founders of the Biostatistics Collaboration of Australia (BCA), which provides an accredited national program of postgraduate courses in biostatistics. He joins us to talk about how he ended up in statistics, starting the BCA and the future of statistics. Andrew is the head of the Biostatistics Unit at Monash University and his research interests are in development of analytical methods for interrupted time series designs, the application of causal modelling principles to practical problems, and latent variable methods.

On starting in engineering:

I started in engineering at university – I did a year of general engineering at Melbourne Uni. I found out I wasn't suited for it at all. I had to make a structure that could hold up the textbook for the subject using two pieces of sticky tape and a sheet of newspaper. So I thought about and I built something small – and then people came in with things three, four times the size of mine and they clearly had hugely practical engineering brains. I didn't and I wasn't good at making things – I enjoyed the theoretical principles of making building bridges and dams, but I didn't have that practical sense at all.
And so I moved into civil engineering in my second year and lasted 3 weeks in that. I was in a lecture about the structural properties of concrete and I just walked out. I thought "I can't take this anymore" and I moved into a straight science degree doing maths and statistics at Monash. I was reasonably good at maths – I wasn't a stellar student – but I was reasonably good. I didn't know exactly what I wanted to do, but I had found something I was interested in.
I remember I had one lecturer who taught probability and I absolutely loved it. In particular when he taught the Gambler's ruin – if a gambler keeps betting he will eventually be ruined, using probability rules. It was fascinating how you could write that down mathematically. I just loved it.

On his early career:

I finished my undergrad degree, and then worked at a cadetship at the Australian Bureau of Statistics, before moved into a PhD overseas at Cornell University in the USA. The system there is you do about 2.5 years of coursework, then you do PhD proposal presentation and an oral exam covering all 2.5 years of coursework, which is very stressful. And only then do you do the actual research component over another 2.5 years. My PhD research was in measurement error in matched case-control studies – I ended up deriving theoretical results using lots of Taylor series' for a method that no one has ever used! But it is the process that is important. After that, I worked for Novartis in drug trials for a year, and then looked to move back to Australia.

On getting a job that he didn't apply for:

While on a visit back to Australia while still working for Novartis, I talked to people in various institutions about possible positions. The Alfred Hospital had a position for a consulting statistician, and I applied for it. When I got back to the US, I received a fax saying "Congratulations, we have appointed you to Lecturer in Biostatistics at Monash University" . I hadn't even applied for that job – I wanted the other job in statistical consulting. But the job I was offered was doing collaborative research and teaching in the Masters in Public Health, and I have to say it was a baptism of fire figuring out what public health students knew and didn't know, coming from my background in mathematical statistics. My first classes were so far beyond the level of the students they went down like a lead balloon (laughs).

On the importance of networks and mentors:

I was the lone statistician at the Alfred hospital for five years. The early career isolation was real for me - I was responsible for the analysis for a lot of major clinical research projects. I didn't have any network or support doing the sorts of projects I was doing. I read a huge amount and quickly realised I had to learn epidemiological methods in detail.

This is why I now say that it is important in your early career to have academic support of other statisticians, especially for those isolated statisticians who are employed in other fields or schools. These statisticians have to answer all the questions that come their way on a range of topics and that is very difficult without support. Unfortunately, after all this time there is still a perception in some areas in medicine (and perhaps elsewhere) that you employ a junior statistician and they can do everything. That is, there is a perception that statistics is easy and all you have to do is follow a recipe and do a t-test and that is it – which isn't true, of course. Countering the lone-soldier aspect is the reason why we have recently established a precinct-wide group here at the Alfred Hospital campus, where we invite statisticians from all research groups across the campus to monthly meetings.

On changes in necessary statistical skills:

Statisticians now have to be able to cope with large amounts of data – and that doesn't mean only using machine learning. I am thinking of the causal inference subject in the BCA, where we talk about the three tasks of data science/statistics: description, prediction and explanation. And so even though data is exploding everywhere, we still have to think what we want to do with it. But in terms of skills, statisticians have to have great programming skills, including being able to simulate scenarios. For example, doing power calculations in a complex scenario, there might not be the theoretical formulae available, and so you have to have strong enough programming skills to be able to simulate it. Also visualisation is important – such as using RShiny – to be able to communicate your complex statistical procedures graphically.

On co-founding the BCA:

The BCA was formed with a federal government grant as there was no masters degree in biostatistics available in Australia. No one university had the capacity to develop and deliver a whole biostatistics masters – there were just not enough biostatisticians around - and so the only way to deliver this was for the universities to work together. There were maybe 30 students in the first year, but now we have trained hundreds of students. Also the BCA has also really allowed biostatisticians across Australia to get know each other, and so allowed much greater collaboration across the different universities. It has been wonderful.

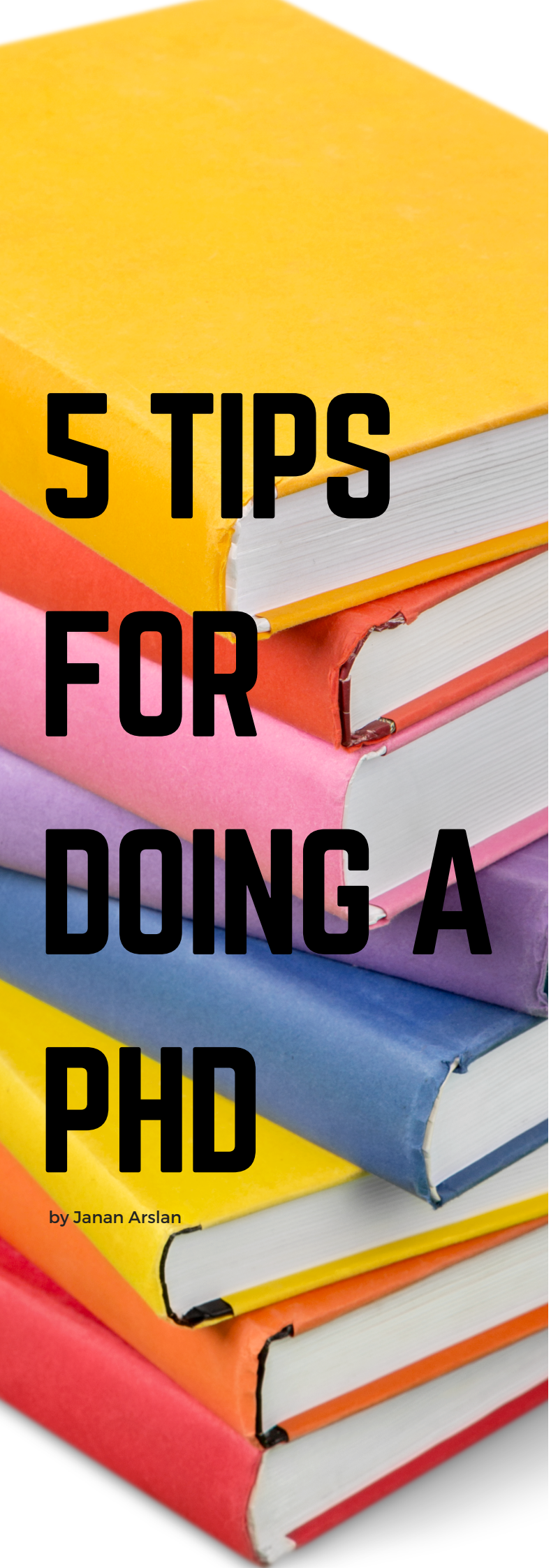
On studying a masters with the BCA:

The BCA program is a specialised program that allows people to become independent biostatisticians themselves or to upskill to do their research better. We take postgraduate students from a range of backgrounds; from health, maths, and engineering. I think biostatistics is better suited as a masters degree, not undergraduate, as you need a level of maturity to really understand the research questions. Some students have done a graduate diploma to prepare for doing PhDs, while others have done the full masters, and even gone on to do PhDs in biostatistics. We offer a very flexible means of studying biostatistics, which is great for people who are juggling work and other areas of their life. Also, we are registered with the Statistical Society of Australia, and so when you graduate with the masters you can become an accredited statistician. I am very proud to say when you think of a postgraduate qualification in biostatistics in Australia, a masters through the BCA has become the de facto standard.

On the future of the BCA:

There are some exciting developments for the BCA program for 2022, with a revised and more flexible program to be announced shortly. I will be able to talk more about that in the ECSSC information booth next week, and if you aren't able to make it, then just reach out to the BCA and we are more than happy to answer your questions.

Thank you Andrew for the interview! Join Andrew for an information session on Zoom about the BCA on Tuesday 27 July from 12:30pm to 1:30pm AEST. Feel free to ask questions about joining the BCA – I am a happy graduate of the program and highly recommend it!



5 TIPS FOR DOING A PHD

by Janan Arslan

1) Find the right supervisors

My number one tip is to make sure you find the right support group for you to avoid unnecessary challenges in your PhD. Be picky and careful - you are going to be working with these people for at least three years! First, choose your Primary Supervisor. This will be your go-to person for resources and admin issues (e.g. funding for publication costs and conferences), but they must also have the expertise to guide your project. Next, choose the Co-Supervisors based on the assistance you will need for the different components of your project. Perhaps one supervisor is an expert in statistical modelling, while another has clinical expertise. Finally, have open and transparent discussions with these potential supervisors before making them an official part of your team, to make sure they share your vision for the project.

2) Ask for help

Doing a PhD can be isolating. Things go wrong and your stress level will rise to new heights, but you are not alone. Feeling anxious at different stages of your PhD is completely normal. You are achieving a lot, and the tasks required can be overwhelming. Please seek help. This is why choosing the right supervisors at the beginning is important. Reach out to your team and let them know when things are going great or bad. Let them know what you need. Too often PhD students believe the onus is on them for a successful completion, but remember, you have a team to support you and ensure your success. If you don't feel comfortable reaching out to your supervisors, reach out to your Student Coordinator, or access University resources, such as counselling. Just know that you have plenty of help available to you whenever you want. This includes the Early Career & Student Statisticians Network at the Statistical Society of Australia! If you need any guidance at any point in your academic career, please reach out to us and we're more than happy to help in any way that we can.

3) Work-life balance

Establish a schedule which includes plenty of breaks. Every 45 minutes, you might take a 10 to 15 minute break. During lunchtimes, don't eat your lunch at your desk - go outside for a 1-hour walk to reset before buckling down again. Make sure that your laptop is shut off after a certain time every night to ensure plenty of rest and include social time with friends and family. The more structured your schedule and the more you adhere to this structure, the easier your PhD.

4) Establish a writing habit

Don't wait until you have results before you start writing. Establish a habit of writing regularly, even if it is only about your research protocol at the beginning. Schedule a regular half-day or one day a week to write something. Trust me on this!

5) Publish early

Again, do not wait until you have results to start writing publications for your PhD. At the beginning, you can put publications out there in the form of commentary articles, perspectives, or even a systematic review of the literature. For example, the first article in my thesis is a perspective article about the application of artificial intelligence to medicine.

LIBRARY(BOOKS)

by Jordan Hedi

This issue we have a book that should be within arm's reach of every person on the planet, not just statisticians. The Art of Statistics: Learning from Data by David Spiegelhalter is a must read, especially at this time in history.

This book is remarkably easy to read and understand – a testament to Spiegelhalter's ability to communicate concepts without being bogged down in jargon. Some critics might argue his style of writing is too 'pedestrian' or 'watered down', but I'd say they are missing the point. He gives everyone an actionable overview of the statistical field, regardless of their background knowledge.

Furthermore, the book takes the time to explain misconceptions about data analysis. Tools like surveys are shown to be less informative due to a lack of a representative sample. The misinterpretation of p-values, point estimates and confidence intervals is dissected. By exploring these issues in depth, the book sculpts a better understanding of the statistical outputs generated by data analysis. Now some might say these areas have already been discussed in length elsewhere, but the reality is many people still misrepresent these figures. So, with the aforementioned ease of understanding that this work offers in its writing, it is a great addition to the overall discourse.

Lastly, I greatly appreciate the final two chapters of this book, which are called "How things go wrong" and "how we can do statistics better". I won't go into detail as you can gauge their content from the chapter titles, but, I do feel that they offer the data analysis profession a much needed breathing space when it comes to talking about issues that matter in the statistical practice.

I learnt a lot from this book; I can't tell if it's due to a deficiency in my statistical understanding or just excellent statistical writing - I'm gonna go with the latter to save face. This book gave me a more comprehensive understanding of statistics than my bachelor's degree in statistics. Check it out in your library or pick up a copy for around twenty dollars – you won't regret it.

AMSI WINTER SCHOOL

The AMSI Winter School for Statistical Data Science has been a whirlwind of statistical methods packed into two fascinating and bamboozling weeks. I am a PhD candidate in applied biostatistics and the topics offered this year ticked many boxes for me – the intro to Bayesian statistics with Professor Gael Martin was a big tick as I didn't get to learn about this area in my degree. This intro was built on with some excellent lectures and tutorials from Associate Professor David Frazier.

Then there were the lectures on all things Markov Chain Monte Carlo by Drs Leah South and Matias Quiroz. Again I felt a little bamboozled, but the idea of MCMC is definitely the most interesting of the topics I've learned at the Winter School. It also seems to be one of the most rapidly developing areas of research, and hopefully I'll get to incorporate Bayesian statistics in my PhD somewhere!

Another box ticked for me was learning more about methods used in data science since it's such a fast-paced environment. Fortunately, Drs Susan Wei and Robert Salomone gave some fascinating, albeit challenging, lectures and tutorials on neural networks and related models. I still haven't quite grasped how everything works, but I feel good about at least knowing the fundamentals and the types of problems they're good for.

Finally, and my personal favourite, were the lectures on dimension reduction by Associate Professor Anastasios Panagiotelis. I think these were my favourite because I found the content the easiest to understand – a fact which I attribute to the quality of the presentation and presenter! I can feel my brain ticking along in the background figuring out how I can make use of dimension reduction in my PhD. I suppose my brain is a literal 'black box neural network', as I have no idea how it works, but I know something useful is going to come out at the end of this Winter School!

So I'd like to say a big thank you to AMSI for not only putting on the Winter School but also making scholarships available. Also, a big thank you to their sponsorship of the ECSSC – don't forget we are live on Monday 26th July 11 am AEST!

by Ben Harrap

THANK YOU TO OUR SPONSORS...

BIO STATISTICS COLLABORATION OF AUSTRALIA

sdas

Arbutus | QDA Miner | **Stata** | Stat/Transfer | WordStat

ACEMS

AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR MATHEMATICAL AND STATISTICAL FRONTIERS

Australian Bureau of Statistics

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

And thank you to our speakers and contributors... without you the conference and the magazine wouldn't be possible.