DATA7002:Responsbible Data Science 2021

A/Professor Andrew Crowden
Dr. Slava Vaisman
Dr. Hongzhi Yin
Mr. Hamish MacDonald
Mr. James Bender

Ethics and Data Science: Outline

Week One: Introduction

1. Thinking about Data Science – key questions and definitions
2. Introduction to practical ethics and the nature of moral inquiry and philosophical analysis (dilemmas and conflicts)

Week Two: The theoretical tools of philosophical analysis

3. Approaches to philosophical ethics

Week three: Case study activity

Week Four: Data science governance and regulation

4. What is data and what is information?
5. Collection use and management of data and information
6. Australia's data landscape

Week Five: Decision-making and problem solving in data science case analysis

7. Domain analysis: Research data, Non research data, Algorithm development (machine learning, AI, robotics) and The Practice/s of Data Science.

Data Science case analysis in 4 areas

a) Research **Data**;

b) Non-Research Data
(Collection, Storage and access);

c) **Algorithms** (in machine learning, AI, robotics) and;

d) Data Science **Practices** (responsibilities of data scientists, organisations, data science code of ethics, the characteristics of good data science practice, surveillance)

International Standards and the National Statement

Social/scientific value	Research merit and integrity
Scientific validity	
Fair subject selection	Justice
Favourable risk benefit ratio	Beneficence
Informed consent	Respect for human beings
Respect for potential & enrolled participants	
Independent review	HREC review

VALUES AND PRINCIPLES

- *Research Merit and Integrity*: scientific standard and scholarly merit, integrity of researchers
- *Justice*: distributive justice - benefits, burdens distributed fairly
- procedural justice - recruitment and ethics reviews involve fair processes
- *Beneficence:* weighing of risks and potential harms, sensitivity to welfare and interests of people, social cultural implications of research
- *Respect for people:* intrinsic value of all people, human autonomy, self-determination

Data Science case analysis in 4 areas

a) Research Data;

b) Non-Research Data
(Collection, Storage and access);

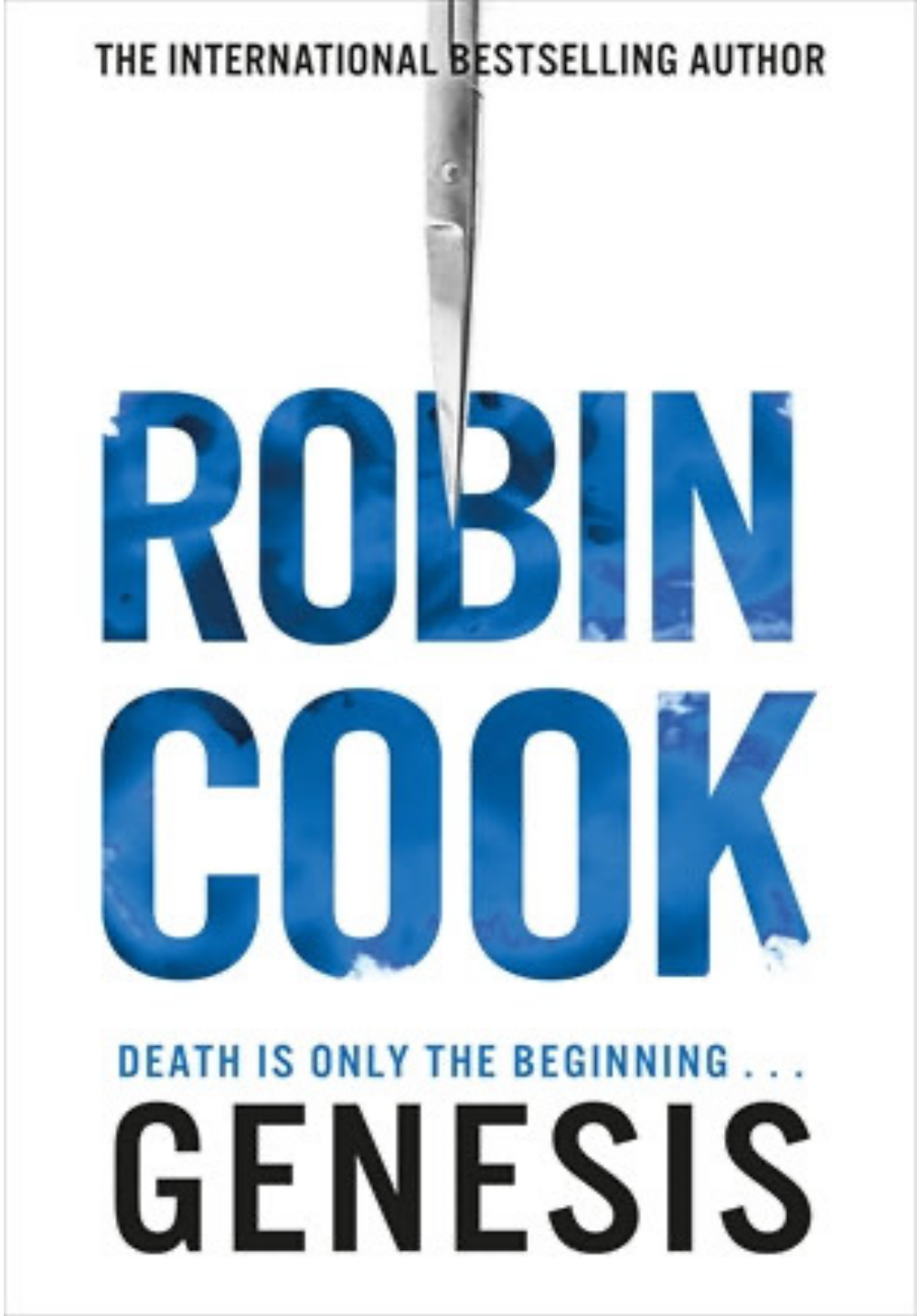
c) Algorithms (in machine learning, AI, robotics) and;

d) Data Science Practices (responsibilities of data scientists, organisations, data science code of ethics, the characteristics of good data science practice, surveillance)

b) Non research data

Forensic DNA

THE INTERNATIONAL BESTSELLING AUTHOR



ROBIN COOK

DEATH IS ONLY THE BEGINNING . . .

GENESIS

Forensic DNA testing is a powerful global tool for criminal investigation and prosecution activities in criminal justice systems

1. Comparing the DNA profiles from suspects to DNA evidence re likelihood of involvement in a crime.
2. Searching for links between the biological material collected at a crime scene to DNA stored in a criminal DNA database.
3. Procedures to search for criminal suspects through their connection with biological relatives.
4. The inference of human externally visible physical features from a biological sample collected at a crime scene.

The Golden State killer

- Joseph James DeAngelo, the alleged Golden State Killer was arrested in 2018, DNA led to DeAngelo's arrest. Authorities used a free genealogy and DNA database called GEDMatch to try to track down the killer who had evaded them for decades.
- They created a profile with crime-scene DNA, and in April 2018, DeAngelo's name emerged in what investigators believed was a pool of possible suspects.
- Detectives then gathered DeAngelo's DNA - some of it taken from the handle of his car door, some from a discarded tissue in his trash -- and found they had a match to evidence.

“Suspected Golden State Killer, East Area Rapist Arrested After Eluding Authorities for Decades.”

- Joseph DeAngelo has plead guilty to the murder of 13 people, the rape of around 50 women and committing burglaries across California during the 1970s and 80s.

- The so-called ‘Golden State Killer’ was arrested in April 2018 after detective work was combined with DNA databases and family trees to identify potential suspects, an approach known as 'genetic genealogy’.

August 21 2020 - sentenced

- DeAngelo was quick to make a deal with prosecutors to avoid execution.
- He pleaded guilty to 13 counts of first-degree murder, as well as 13 counts of kidnapping. While he admitted to dozens of rapes and burglaries, the statute of limitations had passed on those crimes.
- **He will spend the rest of his life in a maximum-security prison after being sentenced to 11 consecutive life sentences without the possibility for parole.**
- DeAngelo had spent his trial sitting silently, staring straight ahead while his victims told the court what he did to them. Loved ones held up photos of the men and women he killed.
- He was not expected to speak, but before his sentencing, DeAngelo rose from a wheelchair, took off his mask and said to the court: "I listened to all your statements, each one of them, and I'm truly sorry for everyone I've hurt."

1. PROFILING THE PERPETRATER

- Investigators collect biological material — such as blood, hair, skin or semen — from a crime scene. Those samples sometimes contain DNA that can be read through genetic sequencing, which involves cutting the DNA into tiny fragments and scattering them over a 'genotyping chip' to see what sticks.
- The chip contains microscopic wells with about 700,000 probes that each match a unique genetic variant — a specific sequence that may (or may not) be found in a person's DNA. Fluorescent dyes are then used to isolate the appropriate probes so a computer can identify the series of DNA letters in a sample, creating a genetic profile.

2. FNDING THE RELATIVES

- The perpetrator's sequence is added to a public database of DNA sequences, such as the genealogy website GEDmatch, which you can search to find similar profiles among its one million users. Those genetic profiles are previously uploaded by consumers who took a DNA test (sold by companies like 23andMe or Ancestry) to learn about variants that might reveal a genetic predisposition to disease and where their (often distant) relatives live around the world.

Law enforcement officials are more interested in whether their perp is closely related to other people in the database, as calculated from number of shared genetic variants. Your DNA is roughly 50% similar to each of your parents, 25% for grandparents. For each generation since two people shared a common ancestor — such as grandparents — their genetic similarity is reduced to one quarter, which means first cousins share about 12.5% of their DNA, second cousins 3.125% and third cousins less than 1%.

3. BUILDING THE FAMILY TREE

Unless all your close relatives are obsessed with their ancestry, it's highly unlikely you would find many first or second cousins in the GEDmatch database, but you should get thousands of third cousins. According to renowned genetic genealogist CeCe Moore, investigators in the Golden State Killer case studied third cousins.

The traditional techniques of genealogy — tracking-down records like birth and marriage certificates, census data and newspaper obituaries — along with modern methods like Facebook stalking are then combined with the DNA profiles to build a huge family tree of people who might be related to the perp. Those family members are painstakingly added to the tree, starting with the twigs of living relatives and then connecting them through branches of distant ancestors.

CeCe Moore calls this process 'reverse genealogy'.

4. IDENTIFYING THE SUSPECT

- After law enforcement agencies have identified some potential suspects via genetic genealogy, they use conventional investigative methods such as comparing present physical features to past eyewitness statements and police sketches. This narrows-down the choices to a few candidates.

- In the Golden State Killer case, some of the victims who survived his attacks described him as a 5'9", 165-pound white male — characteristics that matched the features of Joseph James DeAngelo.

Ethics and Forensic DNA testing

- There are different views about capabilities, benefits and risks.

- *Supporters* – welcome improving efficiency in fighting crime, prevention of miscarriages of justice and deterrence of criminal activity (which is expected to reduce crime and increase public safety and security).

Ethics and Forensic DNA testing

- *Critics* - see potential threats to civil liberties and argue that forensic DNA testing and the storage of profiles with sensitive information on databases (potentially global hybrid infrastructures) may threaten liberty, autonomy, privacy, and the presumption of innocence as well as social stigmatization and racial stereotyping due to the over representation of specific social and ethnic groups in criminal DNA databases, discrimination, mistaken identity and wrongful conviction from erroneous interpretation of data and information.
- Problems re lack of transparency re the use of DNA data, risk of false positives, lack of standardization on DNA analysis among different counties, lack of ethical oversight of the transnational flow of law enforcement information and potential violation of regulations.
- The presentation of DNA information in courts may be impeded by the an perceived lack of DNA literacy from non-experts and over expectation toward the capacity of DNA evidence to solve criminal cases.

The Australian Criminal Intelligence Commission and the Australian National Criminal Investigation DNA Database (NCIDD)

- The NCIDD has recently been enhanced to enable *kinship matching, familial searching and advanced direct matching*

(The Hon Peter Dutton MP Saturday 29 September 2018)

kinship matching, familial searching and advanced direct matching

- Familial searching raises ethical, technical, logistical and efficacy questions related to the economic, temporal and human resources needed to search, review and refine the selection and monitoring of the pool of potential suspects.
- Familial searches expand the net of surveillance to persons whose genetic information would remain private had it not been for the actions of their relatives.
- Need to clarify access to recreational genealogy databases.

We will be ready when we:

- Further clarify and respond to the ethics concerns of critics (need to pay particular attention to processes and pathways of consent).

- Undertake further research into police, other professionals and public DNA literacy and views.

- Facilitate broad police, professional and public engagement to ensure enhanced transparency and accountability.

- Establish National policy/legislation – regulation of DNA databases (criteria for inclusion, retention of data), regulations re familial searching (police access to personal genetic databases), data protection and exchange across borders.

References

Australian Criminal Intelligence Commission, 2019.

Crowden A. Gildersleeve M., Place, Virtue Ethics and Physician-Researcher Dual-Role Consent in Clinical Research, *American Journal of Bioethics*, April 2019 (ID: 1572818 DOI:10.1080/15265161.2019.1572818)

Crowden A., Lamont J., Seven key challenges associated with the ethical governance of new [global] hybrid genomic data infrastructures (genomic biobanks), *Human Genetics Society of Australia, 41st Annual Scientific Meeting*, Brisbane Convention Centre, 5-8 August, 2017

Gildersleeve M., Crowden A., Genetic Determinism and Place, *Nova Prisutnost,* April 2019.

Machado H., & Silva S., What influences public views on forensic DNA testing in the criminal field: a scoping review of qualitative evidence, *Human Genomics* 13;23, 2019.

National Statement on Ethical Conduct in Human Research 2007 (Updated 2018). The National Health and Medical Research Council, the Australian Research Council and Universities Australia. Commonwealth of Australia, Canberra (particularly section 3 on genomic research).

Data Science case analysis in 4 areas

a) Research Data;

b) Non-Research Data
(Collection, Storage and access);

c) Algorithms (in machine learning, AI, robotics) and;

d) Data Science Practices (responsibilities of data scientists, organisations, data science code of ethics, the characteristics of good data science practice, surveillance)

c) Algorithms and automated decision-making (machine learning, AI, robotics)

Some types of data science models can cause harm

Three kinds of data science predictive mathematical modelling that score teachers and students, sort CVs, grant or deny loans, evaluate workers, target voters, monitor our health etc.

1) Competitive sport models – use full healthy, participants interested in outcomes

2) Hypothetical: subjective but in the interest of participants

3) WMD – unfair – opacity (not transparent), scale (effect many), damage (cause harm)

(modified from Cathy 0'Neil ***Weapons of Math Destruction****: how big data increases inequality and threatens democracy*, Penguin Random House USA, 2016)

Four ways we can introduce imperfections into data

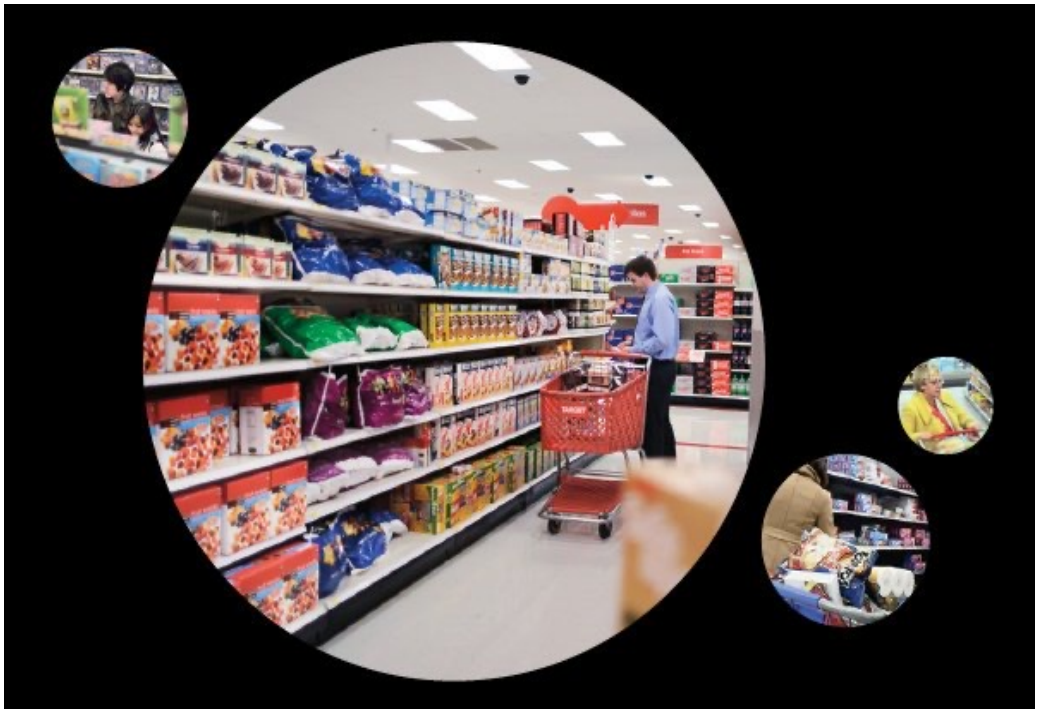
- random errors
- systematic errors
- errors of choosing what to measure
- and errors of exclusion

IMPACT (Sarah Wysocki)

‘Creative . . . Motivating’ and fired.

Target and data mining

Target knew that a teenage girl was pregnant before her father did - and Target proudly disclosed the information to the father!



TARGET

The Facebook and Cambridge Analytica scandal

https://www.vox.com/policy-and-politics/2018/3/23/17151916/facebook-cambridge-analytica-trump-diagram

COMPAS

Northpointe's tool, called COMPAS (which stands for Correctional Offender Management Profiling for Alternative Sanctions), found that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Cases . . . .

https://aiethics.princeton.edu/case-studies/case-study-pdfs/

https://aiethics.princeton.edu/wp-content/uploads/sites/587/2018/10/Princeton-AI-Ethics-Case-Study-1.pdf

Data Science case analysis in 4 areas

a) Research Data;

b) Non-Research Data
(Collection, Storage and access);

c) Algorithms (in machine learning, AI, robotics) and;

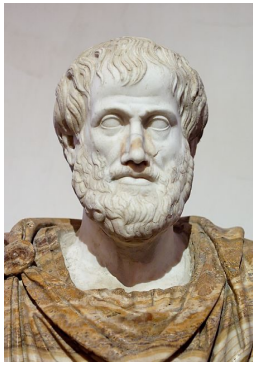
d) Data Science Practices (responsibilities of data scientists, organisations, data science code of ethics, the characteristics of good data science practice, surveillance)

c) Data Science Practices

Three broad approaches (methods) in ethics

1) **Principles.** Non-Consequentialism (deontology)
2) **Consequences.** Consequentialism (utilitarianism)
3) **Virtues.** Virtue Ethics

3) Virtues. Virtue Ethics



- Aristotle (384-322 BC) Nicomachean Ethics

Aristotle connected ethics to “eudaimonia” or flourishing for human beings, which itself was connected to the natural human capacity for rational thought.

- **Virtue ethics focuses on character, rather than actions.**

- **Virtues are character traits, involving habits of action, attitudes and the disposition to certain feelings**. Becoming virtuous requires practice, reflection, and cultivation of those character traits which contribute to a flourishing life.

- **The virtuous individual is someone who, without relying completely on rules or principles, is sensitive and intelligent enough to perceive what is noble or right as it varies from circumstance to circumstance.**

Courage, Generosity, Honesty, Compassion, Self-Respect are examples.

Aristotle offered detailed accounts of each of these virtues as a “mean” between two extremes of excess and deficiency

Phronesis

Aristotle's term for practical wisdom,
refers to the capacity to choose wisely how to act with virtue in particular contexts, where this requires experience, maturity, sensitivity, careful reflection, for example, where the virtues seem to conflict.

- Virtue ethics holds that it is important not only to do the right thing but also to have the proper dispositions, motivations and emotions in being good and in doing right.

- Virtues are excellences in character - acculturated behavioural dispositions that result in habitual acts.

For the consequentialist, a data scientist doing X would be right, in the end, because doing X is expressive of a mode of practice which, when engaged in collectively, maximises utility. For the virtue ethicist, however, a data scientist doing X would be right because doing X is expressive of what the profession of data science actually is, considered as an activity which is committed to one important substantive human goods (goals of data science), which itself is partly constituitive of a humanly flourishing life.

(adapted from doctoring analogy Oakley J & Cocking D., Virtue Ethics and Professional Roles, Cambridge University Press Cambridge, UK 2001, pp 114-15.)

What are the virtues of Data Science?

What are the virtues of data science?

- Virtue ethics focuses on character, rather than actions.
- Virtues are character traits, involving habits of action, attitudes and the disposition to certain feelings.
- Becoming virtuous requires practice, reflection, and cultivation of those character traits which contribute to a flourishing life.

The virtuous individual is someone who, without relying completely on rules or principles, is sensitive and intelligent enough to perceive what is noble or right as it varies from circumstance to circumstance.

Virtues of Science

Communality

Universalism

Disinterestedness

Organised scepticism

Robert Merton (1942)

- ***Communality*** (sometimes referred to as communalism) addresses common ownership of scientific discoveries and the need for scientists to publicly share their discoveries. This could be seen as a precursor to modern initiatives such as open science;

- ***Universalism*** is the idea that everyone can do science, regardless of race, nationality, gender or any other differences, and that everyone's scientific claims should be scrutinized equally. In science, it's all about your arguments, line of evidence and methodology, regardless of who you are;

- ***Disinterestedness*** expresses the idea that scientists should work only for the benefit of science;

- ***Organized scepticism*** expresses the idea that the acceptance of all scientific work should be conditional on assessments of its scientific contribution, objectivity and rigor.

Activity . . .

First, identify the intellectual virtues of data science – those that can be taught about data science . . .

Then, identify the moral virtues, the regulative ideals of data science practice

(those that are learned or inculcated from habit (from doing data science) . . .

Data Science Association Code of Professional conduct

1 – Terminology

Data Scientist – client relationship

2 – Competence

3 – Scope of data science professional services between client and data scientist

4 – Communication with clients

5 – Confidential information

6 – Conflicts of interest

7 – Duties to prospective client

8 – Data science evidence, quality of data and quality of evidence

Maintaining the Integrity of the Data Science Profession

9 – Misconduct

(we need to embed privacy, consent, confidentiality, discrimination, ownership, commercialisation, intellectual property and the importance of fair benefit sharing, conflicts of interest and *the need to ensure equity, reciprocity and respect for cultural diversity.*

Aristotle's formula for moral deliberation

Define the problem – is there a choice? A problem well defined is a problem half solved (John Dewey who admired Aristotle).

1. Don't deliberate in haste
2. Verify all information
3. Consult and listen to an expert advisor
4. Consult or at least look at the situation from the perspective of all parties who will be affected.
5. Identify and examine all known precedents
6. Try to calibrate the *likelihood* of different outcomes and prepare for every single one you think is possible

A practical ethics decision-making framework for data science

Don't deliberate in haste

1. Define the problem - who does/may it affect? *who* may be involved/affected What facts are available? Are there any precedents?
2. Verify all information. Is there any other information you would request that has not been provided? Is expert advice needed?
3. Identify and consider ethical principles/rules/virtues relevant to this case and explain why you think they apply – consider specific ethics frameworks, Identify and explain if there are different perspectives that should be considered?
4. Consider the law. Broadly, what areas of law could be called upon to help with this case? Consider relevant ethical guidelines/policy.
5. Based on your analysis derived from answering questions 1-6, explain what you think is the most ethically and/ or legally sound response?
6. Calibrate the *likelihood* of different outcomes and prepare for every single one you think is possible. Evaluate your choice.