Advanced Methodology
and Data Science

Good Data Science Practice

A multi-perspective discussion on data science in the context of drug development

Lukas Widmer, Mark Baillie, Jonas Dorn,
Peter Krusche, Conor Moloney, David Ohlssen

SDS 2021 / June 9th, 2021

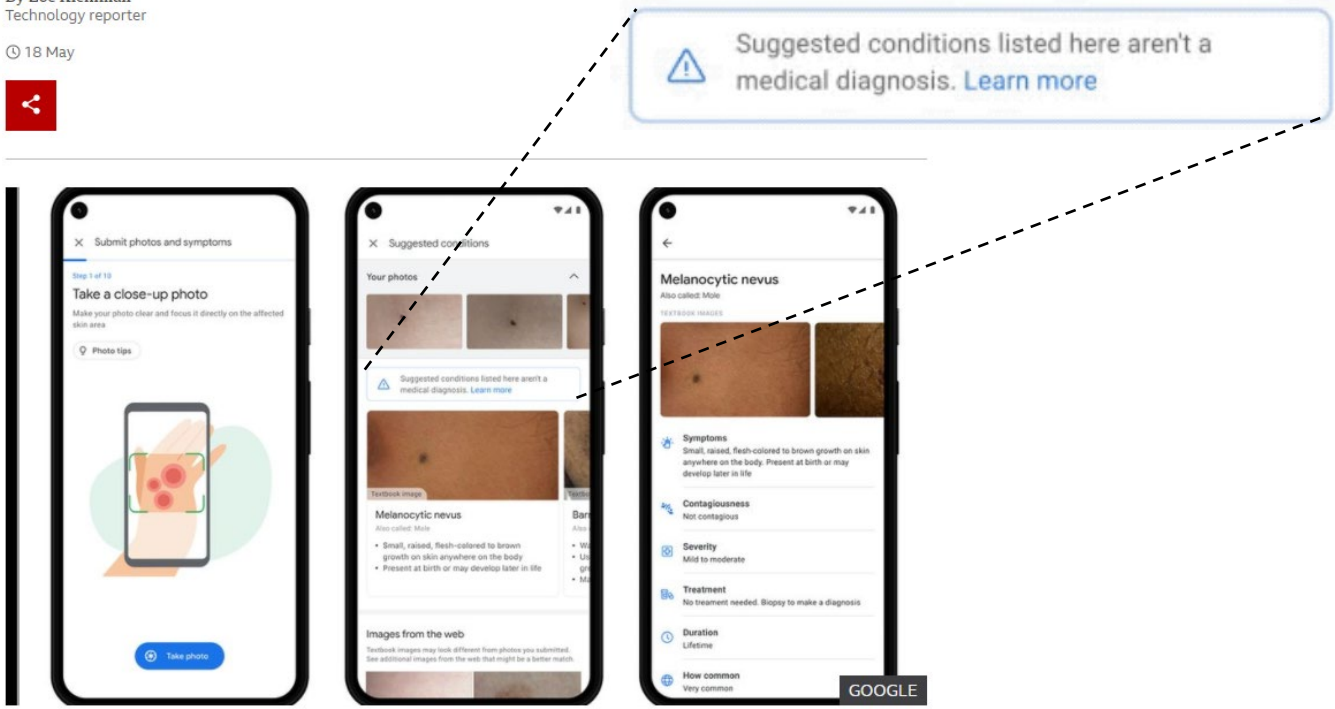
NOVARTIS | Reimagining Medicine

The opinions expressed in this presentation and on the following slides are solely of the presenter, and not necessarily those of Novartis.

Google AI tool can help patients identify skin conditions

By Zoe Kleinman
Technology reporter

18 May



Google has unveiled a tool that uses artificial intelligence to help spot skin, hair and nail conditions, based on images uploaded by patients.

https://www.bbc.com/news/technology-57157566

NOVARTIS | Reimagining Medicine

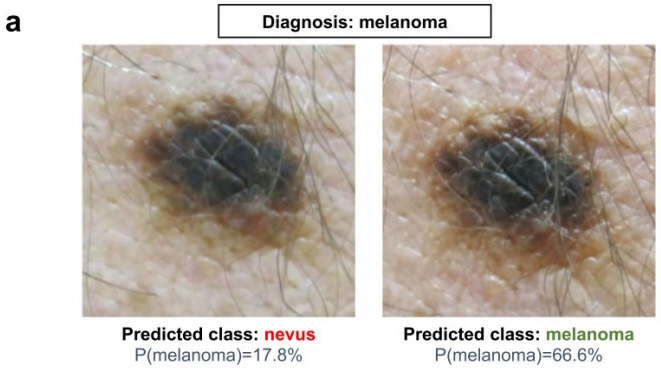


Fig. 3: Dermatologist-level CNN models are not robust across different images taken in the same setting. **a** Representative example of Model A predictions on different images of the same lesion taken sequentially during the same clinic session. The predicted probability of melanoma is shown below the predicted class. (...)

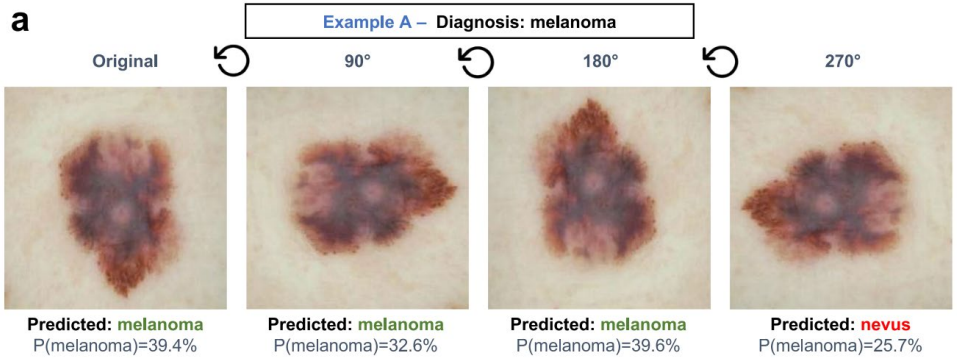


Fig. 4: Dermatologist-level CNN models are not robust to image transformations. **a** Representative example of Model A predictions on a melanoma image from MClass-D. The predicted probability of melanoma is shown below the predicted class. The decision threshold is model confidence >31.0% (...)

Young, A. T. et al (2021). Stress testing reveals gaps in clinic readiness of image-based diagnostic artificial intelligence models. npj Digit. Med. vol. 4 (1).

Good data science practice in drug development?

Aim: learning from existing and future data using advances in science, statistics, machine learning, computation, AI, etc. to:

- increase our understanding of drug, disease and patients,
- accelerate and improve our development projects, and
- inform our decision making.

What do we mean by data science?

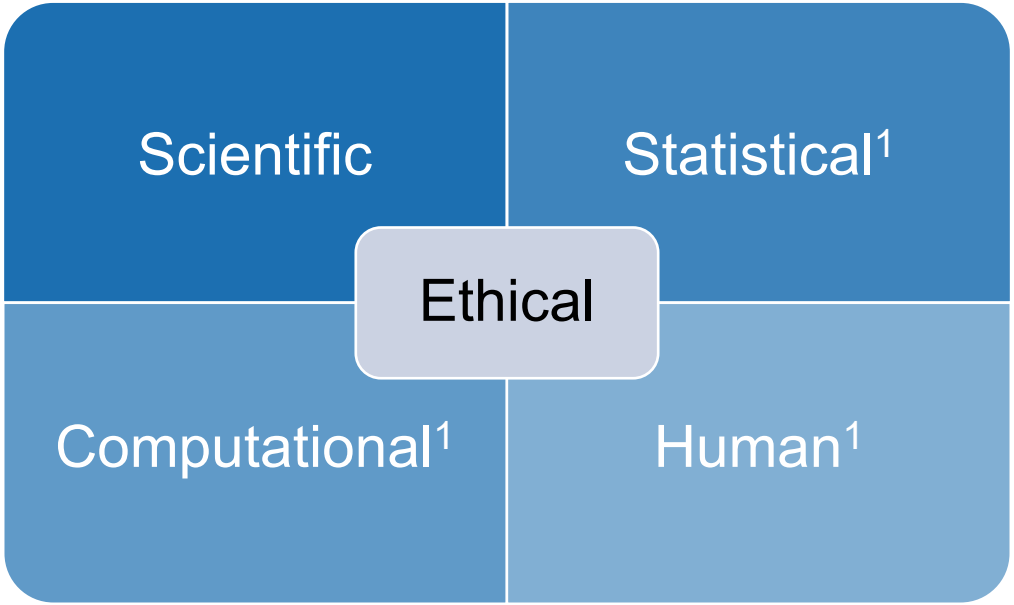
What are the practices of data science?

What do we mean by “good” practices in the context of drug development?

A view of data science for drug development

Good Data Science incorporates multiple perspectives – a holistic view:

Greater statistics tend to be inclusive, eclectic with respect to methodology, closely associated with other disciplines, and practiced by many outside of academia and often outside professional statistics.
(John Chambers, 1993)



[1] *Blei & Smyth (2017) Science and data science. PNAS 114 (33):8689-8692*

HOW A UNIVERSITY GOT ITSELF BANNED FROM THE LINUX KERNEL

The University of Minnesota's path to banishment was long, turbulent, and full of emotion

By Monica Chin | @mcsquared96 | Apr 30, 2021, 10:45am EDT

Illustration by William Joel

https://www.theverge.com/2021/4/30/22410164/linux-kernel-university-of-minnesota-banned-open-source

...

Our community does not appreciate being experimented on, and being "tested" by submitting known patches that [are] either do nothing on purpose, or introduce bugs on purpose. If you wish to do work like this, I suggest you find a different community to run your experiments on, you are not welcome here.

Because of this, I will now have to ban all future contributions from your University and rip out your previous contributions, as they were obviously submitted in bad-faith with the intent to cause problems.

https://lore.kernel.org/linux-nfs/YH%2FfM%2FTsbmcZzwnX@kroah.com/

Statement from Computer Science & Engineering confirming Linux Technical Advisory Board findings - May 9, 2021

We again extend our apologies to the Linux Kernel community for the concerns and extra work caused by our inappropriately designed "hypocrite commits" project. We also want to express our appreciation for the thoughtful report released by the Linux Technical Advisory Board (TAB) on May 5, 2021, and the willingness of the Linux Foundation to meet with us on May 6, 2021.

(…)

We reiterate our apology, and we rededicate ourselves to educating our faculty and students in conducting research that is not only of the highest technical quality, but also follows the highest ethical standards.

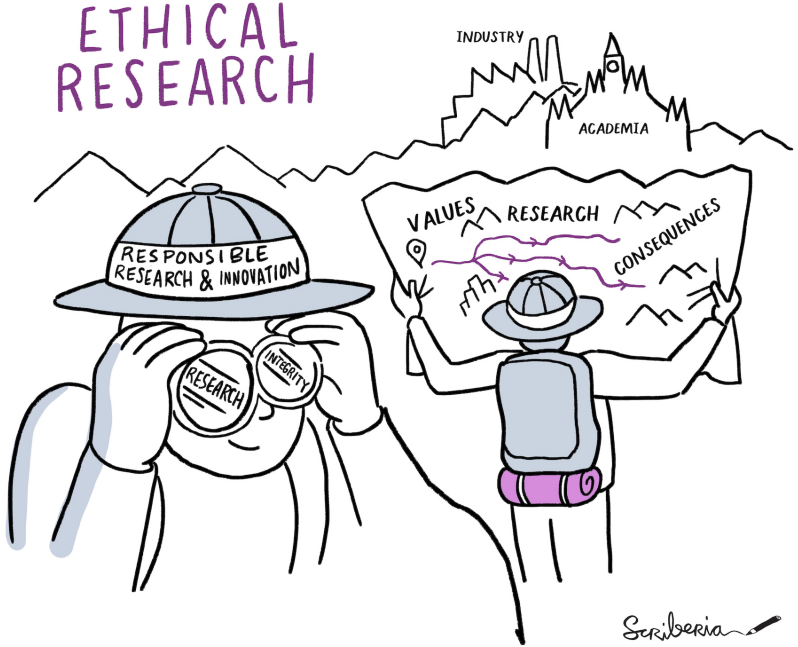
https://cse.umn.edu/cs/statement-computer-science-engineering-confirming-linux-technical-advisory-board-findings-may-9

Ethical perspective

We need less research, better research, and research done for the right reasons
(Doug Altman)

Professional code of conduct[1] – guiding principles:

- Selflessness: Place the needs and concerns of those who depend on us above our own, and prevent harm
- Skill: Continuously aim for excellence in our knowledge and skill
- Trustworthiness: Take responsibility for personal behavior and conduct
- Discipline: Follow prudent procedure and functioning with others



[1] see appendix for pertinent examples

Figure: Turing Way Community (2019) A Handbook for Reproducible Data Science. Zenodo.

Scientific perspective

exploring the context—obtaining sufficient background information to formulate the problem carefully
(Chris Chatfield, 2002)

Provides the context – the why: Subject matter experts critical!

Ensures that prior knowledge can be navigated and leveraged

Ensures clarity on the purpose, outcome, value and impact

Scientific theories determine what to measure, and how

Figure: Farrow & Mathers (2020), Conceptualising Research Methodology for Doctoral Researchers in Open Education (with penguins). Int. J. Manag. Appl. Res. 7, 349–359.

Statistical perspective

There are no routine statistical questions, only questionable statistical routines.
(David R. Cox)

Statistical thinking provides strategies and methods to answer scientific questions:

Translation & precise formulation of the problem – what are we trying to solve?

Question types[1] (e.g., description, prediction, explanation, intervention, ...)

Use of appropriate analytical strategies and methodology, e.g., designing experiments vs strategies for found data (target trial vs reality[2])

Understanding measurements, variation, biases and uncertainty

*1 Hernán et al (2019) A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. CHANCE **32**, 42–49.*

*2 Hernán et al (2016) Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. J. Clin. Epidemiol. **79**, 70–75*

Computational perspective

A big computer, a complex algorithm and a long time does not equal science.
(Robert Gentleman)

"Computational thinking is the thought processes involved in modeling a situation and specifying the ways an information-processing agent can effectively operate within it to reach an externally specified (set of) goal(s)"[1]

Considers algorithmic implementations of methods and understanding their computational footprints.

Design and application of hardware, software, packages, libraries, languages to solve a specific problem

Technical replicability

[1] *Nardelli (2019) Do we really need computational thinking? Commun. ACM 62, 2, 32–35.*

Human perspective

- Strive for balance: prevent one person or perspective from dominating a team and leading to weaker solutions
- Communication across fields: understanding & openness to learn from each other
- Ensure reproducibility, replicability, ...[1]
- Translating, assimilating and operationalizing knowledge

[1] Stark (2018) Before reproducibility must come preproducibility. *Nature* **557**, 613–613.

Figure: Turing Way Community (2019) A Handbook for Reproducible Data Science.

Towards good data science practice

We need to bring our best technology – biological and digital – and our most creative people together to impact diseases in a meaningful way
(Vas Narasimhan)

How can I become a good Data Scientist?

How do we, as an organization, build good Data Science teams?

Good Data Science Principles: a starting point

Has a clear purpose, as well as defined objectives and scope

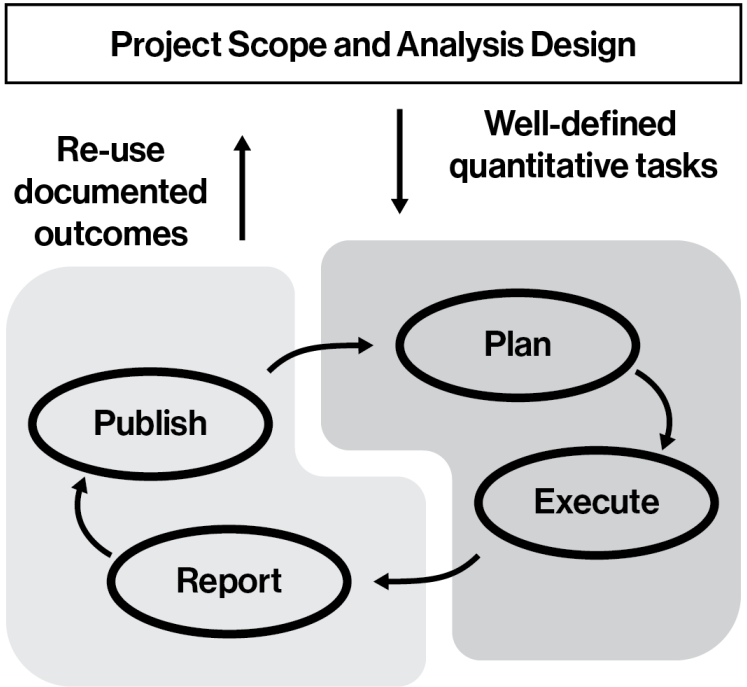
Builds on existing work

Critically thinks about quality, data, methods, code, interpretation, and communication

Is accurate, reproducible, and auditable – all the way from inception to interpretation

Is supported by a learning culture and a collaborative mindset

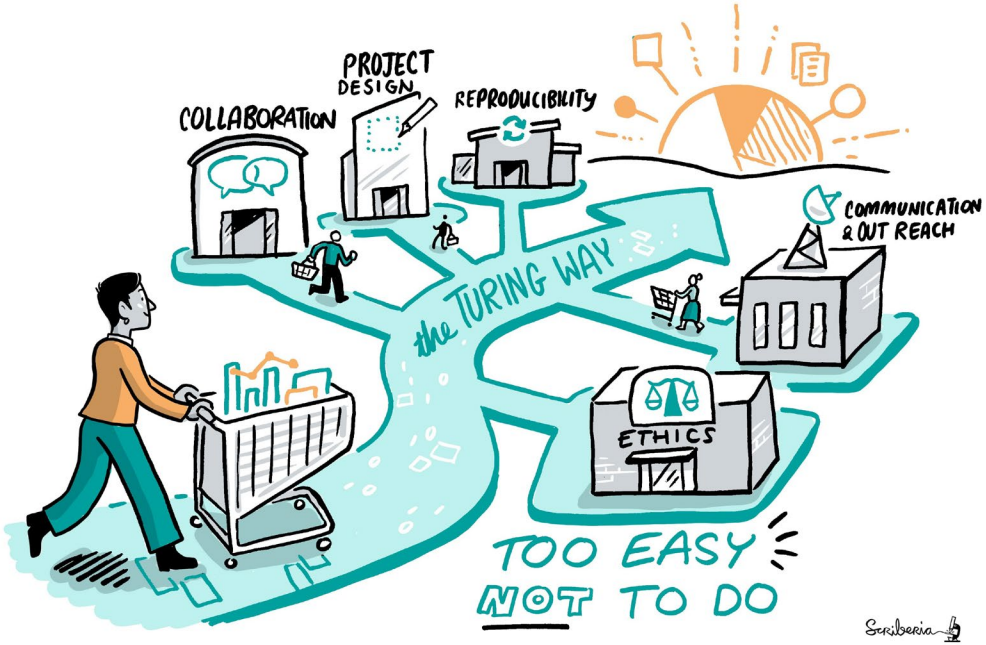
Starting points: frameworks for good data science practice



For example, frameworks for the lifecycle of a project help to identify where the various practices come in:

- Project scoping
- Problem formulation
- Communication
- ...

Starting points: pragmatic & practical guidelines – good practice is a journey



Bad news: there is no trivial solution

Good news: there is low-hanging fruit!

Two resources for getting started:

- The Turing Way: A Handbook for Reproducible Data Science. Zenodo. 10.5281/zenodo.3233986
- Wilson, G. *et al.* Good enough practices in scientific computing. *PLOS Comp. Bio.* 10.1371/journal.pcbi.1005510

Figure: Turing Way Community (2019) A Handbook for Reproducible Data Science. Zenodo.

Summary, learnings & discussion

Good data science practice takes a holistic approach, and includes people that collectively & collaboratively span a range of perspectives

Good data science principles provide a starting point and compass to becoming a good data scientist (team), but good practice is a journey!

An open mind and balancing diverse skillsets are key for successful data science – with a mindset of continuously improving how we work, learn, and communicate as an interdisciplinary team of quantitative scientists.

Practicing good data science is a team sport: the team greatly benefits if we can communicate our deep, diverse skillsets in an understandable manner!

Good Data Science Practice is a community effort – please contact us with your ideas:

lukas_andreas.widmer@novartis.com

NOVARTIS | Reimagining Medicine

Appendix:
Ethics & professional codes of conduct

- ASA: https://www.amstat.org/ASA/Your-Career/Ethical-Guidelines-for-Statistical-Practice.aspx
- ACM: https://www.acm.org/code-of-ethics
- IEEE: https://www.ieee.org/about/corporate/governance/p7-8.html
- For teaching: https://osf.io/preprints/socarxiv/z9uej/
- Declaration of Helsinki: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/
- Novartis: https://www.novartis.com/sites/www.novartis.com/files/code-of-ethics-english.pdf