INTRODUCTION TO DATA SCIENCE: COURSE OVERVIEW AND TEAM INTRODUCTION

INDUCTION AND INTRODUCTION

OUTLINE

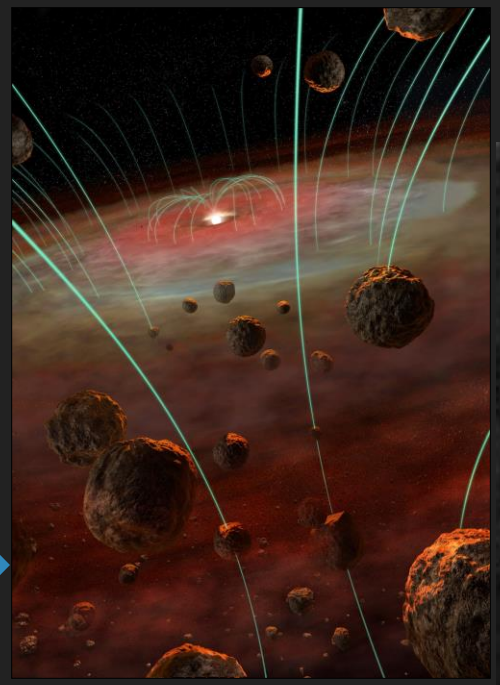
- Introductions
  - Meet your instructor
  - Icebreaker survey
- Course structure
  - Course objectives – Intended Learning Outcomes and overall goal
  - Weekly topics and activities
- Assessment
- Additional resources

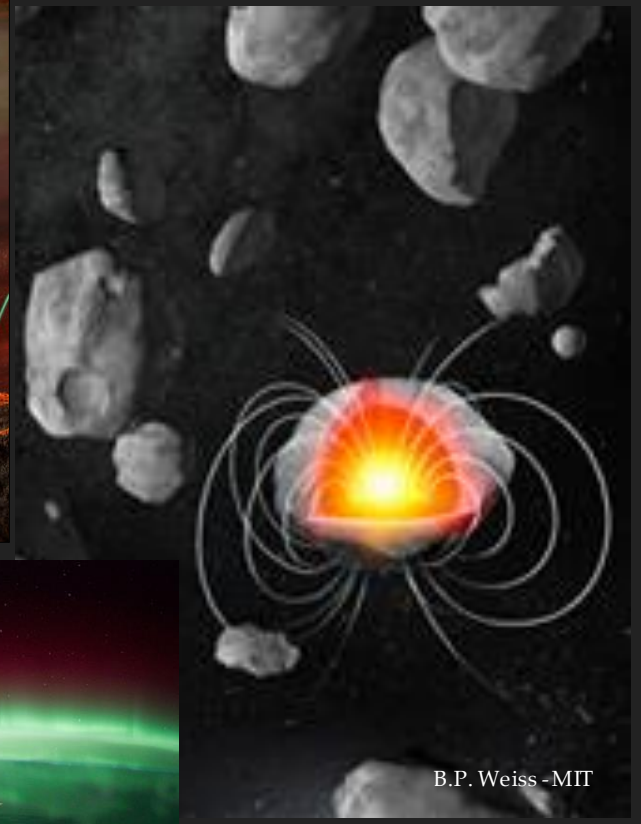
INTRODUCTIONS

WHO I AM

- Currently the Lord Kelvin Adam Smith Research Fellow in Data Science
- Based in the School of Geographical and Earth Sciences
- Hobbies include surfing, skating and rock climbing

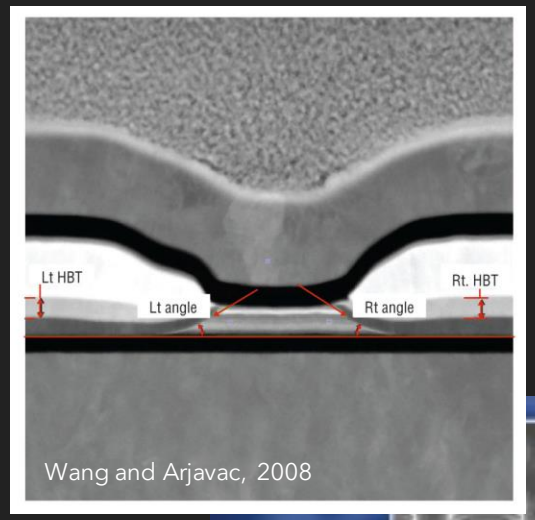
FROM DEVICE PHYSICS TO THE ORIGIN OF THE SOLAR SYSTEM



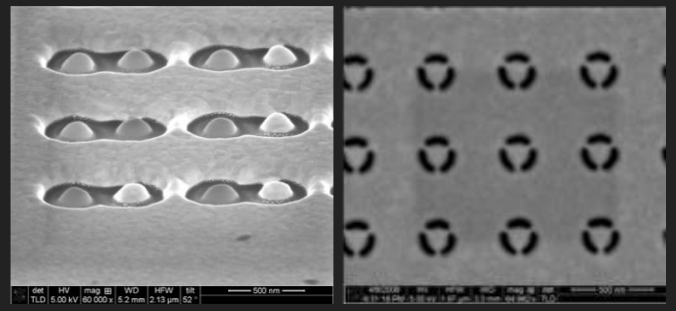


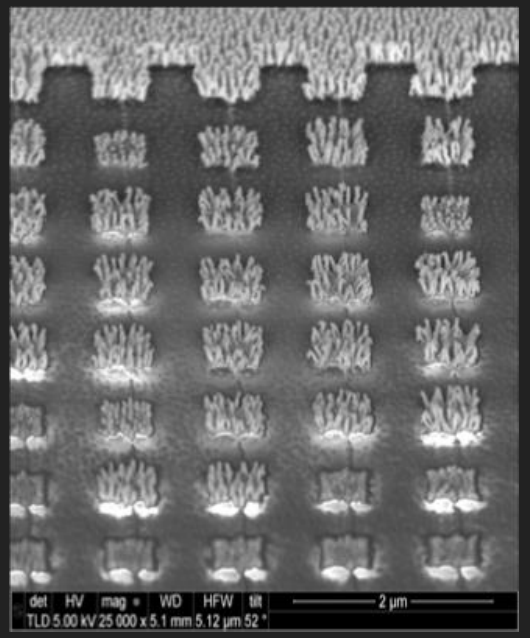
PROCESS CONTROL AND DEVICE PHYSICS

Statistical Process Control

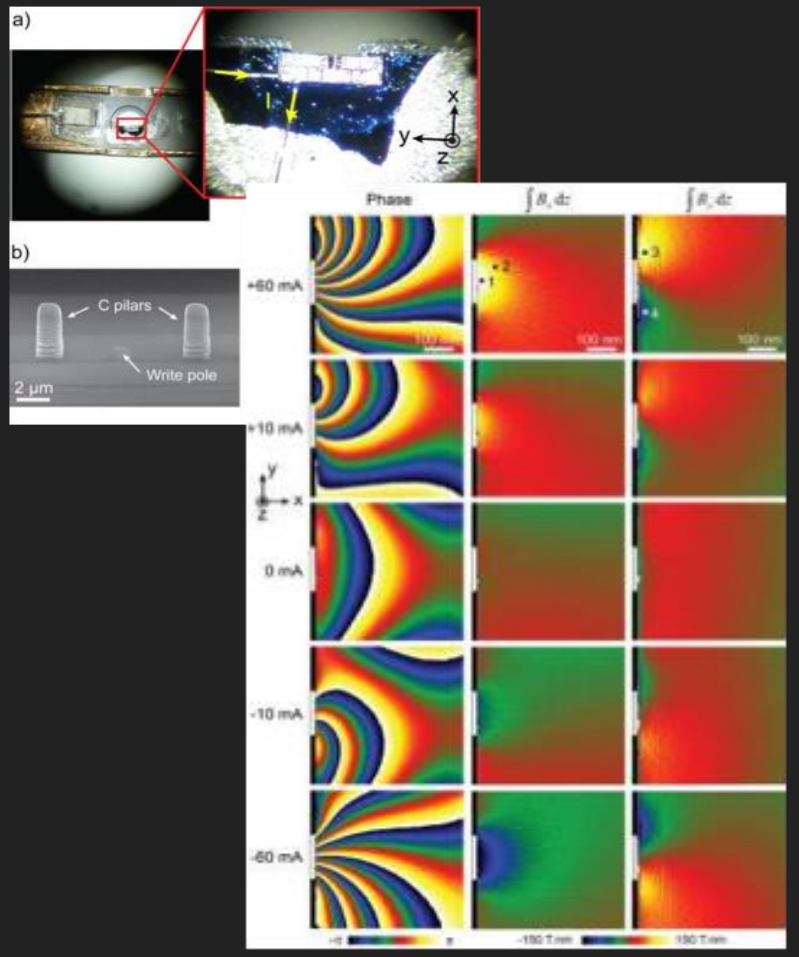


FIB Nanofabrication





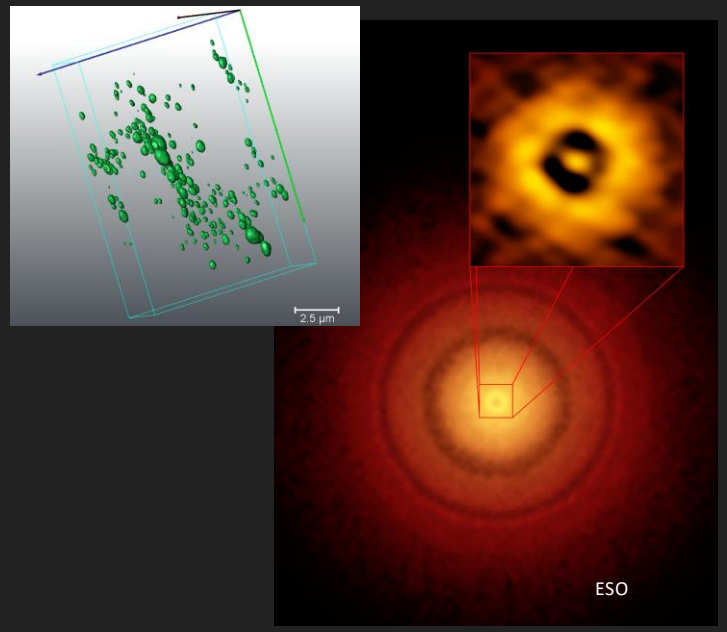
In-Situ Holography of Active Hard Disk Drive Write Pole



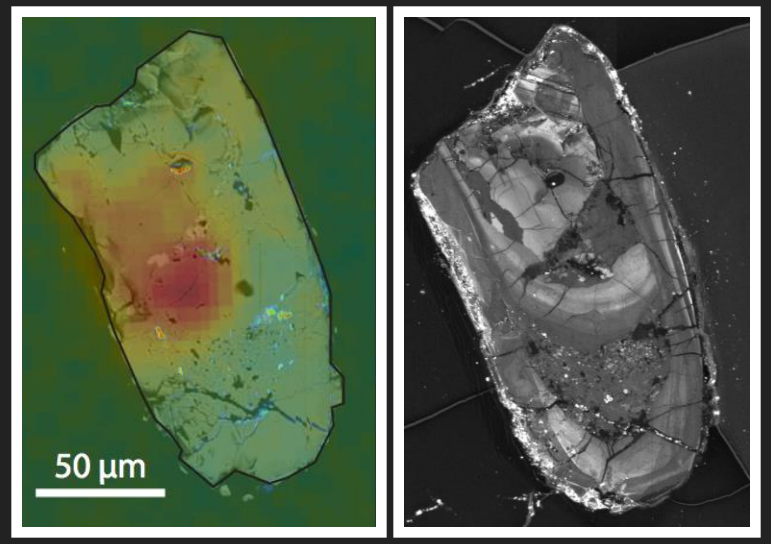
Einsle et al. Nano Research (2015)

CONNECTING MATERIALS PROPERTIES THROUGH DEEP TIME

Fidelity of Nebular Magnetic Fields

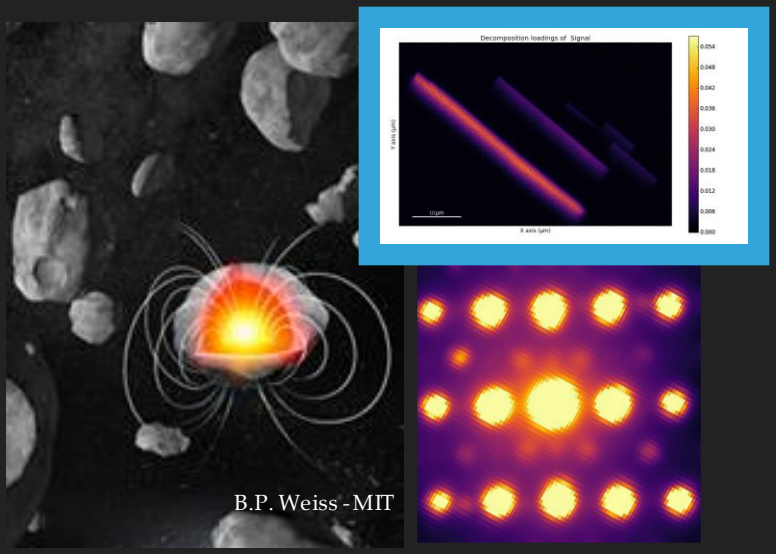


Mineral limits for Hadean Zircons



Weiss et al, Geology 2018

Magnetic Records in Differentiated Palentessimals



- How do you maximise the information available in the oldest and rarest materials?
  - How are we confident in the measurements that we make?
  - How do bulk measurements directly relate to the nanostructures in a sample?

COURSE ADMINISTRATORS

Rosemary (Rosie) McGovern

Upskilling Course Administrator

Anne Dunlop

Learning & Teaching Administrator (GES)

GETTING TO KNOW YOU - ICE BREAKER

- The link below is to an interactive icebreaker.
  - These are all short answer and or multiple choice
- This is to let me get to know you better and generate a little data that we can use in our first practical.
- All data collected is completely anonymous and covered by GDPR
- https://www.menti.com/ch2gfr1ihh

COURSE STRUCTURE

COURSE OBJECTIVES - INTENDED LEARNING OUTCOMES

By the end of this course students will be able to:

1. Demonstrate how to use the data structures, functions, and visualization tools in Python (and several dependent libraries) to explore and analyze multivariate data.
2. Produce summary statistics for exploratory data analysis and multivariate statistical tests.
3. Employ supervised machine learning algorithms to perform classification and prediction tasks on data sets.
4. Apply unsupervised machine learning to perform dimensional reduction, data clustering, and categorization on unlabeled high-dimensional data.

THE BIG OBJECTIVE

The goal of the course is to teach the theory behind machine learning and make it less mystical, so you are equipped to use it appropriately.

This course will:

1. Provide you with a strong foundation in data science and familiarity with statistical tools.
2. You will learn to use Python as **tool** for data analysis.
3. You will learn how to use coding basics to explore data and document your insights with Jupyter notebooks and make informative plots with python tools.

This course is not focused on:

1. Teach you good coding practice (ie rev control, code optimisation) - but some bits will be covered as a 'by the way you might want to consider…'
2. It is not an introduction to coding (you will learn some coding along the way if you have never thought)
3. We will not build dashboards, applications or how to build virtual machines / environments

COURSE STRUCTURE

Week	Topic	Materials Covered
1	intro to python, data structures and tabular data - moving beyond Excel	Course induction (weekly structure, assessment, introductions) introduction to google colab and jupyter notebooks, beyond spread sheets, introduce Pandas Basic plots.
2	exploratory data analysis - missing data and summary statistics	manipulating a dataset: missing values, examining the data, summary statistics, scatter plots , 'For loops'
3	multi-variate statistics -normal distributions and t-tests	exploring normal distributions and using the t-tests for small data populations / non-normal datasets
4	Multivariate statistics II - ANOVA testing	introduce analysis of variance, hypothesis testing, one and two way ANOVA testing also box plots,
5	Data wrangling : normalisation and scaling	exploring why need to re-scale and 'manipulate ' data
6	supervised learning: linear regression	test and train data; least-squares, ridge and lasso
7	supervised learning: nearest neighbour analysis & final project design	NNA for missing data, work with students on designing their final project
8	Unsupervised Machine learning: PCA and Factor analysis	high dimensional data and need for dimensional reduction, scree plots, PCA and VARIMAX, heteroscedastic noise, linearity
9	Unsupervised Machine learning: clustering	k-means clustering, building an unsupervised data pipeline (pre-processing, PCA, clustering)
10	working on BYOD final project	supervised time to work on final project. And preparing report. Evaluation use the tools presented to figure out a way to create an action for solving a problem

WEEKLY STRUCTURE

- Monday – Lectures for week go live
- Tuesday – Live practical data workshop (optional) (6:30 pm to 8:30 pm)
- Wednesday – Question and Answer forum closes (10 PM)
- Thursday – Live (optional) Question and Answer (7-8 pm)

TYPES OF ACTIVITIES

- **Interactive lectures**
  - Call and response format –
    - Recorded lecture will demonstrate a concept
    - then you practice in a pre-formatted notebook
- **Data Workshop(optional)**
  - Live zoom session presenting an opportunity for further practice of topics covered in lecture using new data
  - Opportunity to build up a workflow
- **Weekly Question and Answer (Optional/ live zoom/ recorded)**
  - Provide live discussion of questions submitted on a Moodle forum
- **Forum Discussions**
  - Reflect on readings, and videos shared on Moodle

ASSESSMENT

ASSESSMENT

- 2 Summative marked assessments – Data Analysis Project
  - Both due week 11
  - 5 min video presentation (equivalent to 500 word written report) (30%)
  - 1000-1200 word essay (70%)
- 2 Formative assessments – ungraded/ Generic Feedback
  - Week 3 – Short answer document on what is the most common type of data that you deal with? How can machine learning help?
  - Week 7 – Present a video proposal of what your project is. Voice over power point, 3 slides.

THE FINAL PROJECT (SUMMATIVE ASSESSMENT)

- Objective is to gain insight about a complex problem by leveraging 'large data'
  - Here large can mean lots of data or it can mean high dimensional or both.
    - That said between the constraints of Google Colab and what we cover in the course I would not expect anything more than a few GB in size.
  - You should use informative visualisation tools to demonstrate the insights derived, and the steps used to pre-process and analyse your data.
  - Ideally the dataset should come from your place of work, but if not there are multiple open datasets that would be suitable.
  - The two formative assessments are designed to get you thinking about what you will do for your project.

FURTHER RESOURCES (TOPIC SPECIFIC)

This is a masters level course. As such you should engage with the research and professional literature around the topics. It is also a rapidly expanding and changing field with new tools constantly being produced. Two great books to reference and start your reading would be:

- 'Hands on Machine Learning with Scikit-Learn and TensorFlow' by Aurélien Géron. (O'Reilly)
- 'Effective Pandas' by Matt Harrison.

Both are selected for there practical approach to the topics and provide a great reference to some of the ideas that we will cover over the coming weeks. Also useful for plotting data is:

- 'Scientific Visualization: Python + Matplotlib' by Nicolas P. Rougier (https://github.com/rougier/scientific-visualization-book)