Data Analysis Roadmap

Phase 0: Preparation

Being like an accountant.

- **Studying coding basics and best practices**
- **Gathering unprocessed datasets**
  - Calibration & quality control
  - Despiking as needed (visual/subjective; global or running $n\sigma$ deviation, median, 2nd difference, etc.; iterative exclusion; fancier methods)
  - Interpolation to uniform grid if needed (linear, quadratic, or spline; optimal interpolation, local polynomial fit, or spline basis; error assessment)
  - Dataset organization & documentation
  - ⇒ Working dataset version
    (! With all processing code in a script !)
- **Gathering other relevant processed data**
- **Meta-organization**

Phase 1: Exploratory

Being like an explorer getting a first glimpse of new lands.

- Looking at the data
- Making a bunch of plots (! Don't be afraid to make hardcopies !)
- Following intuition & having fun
- Clearing your mind of preconceptions
- Not paying attention to the literature at all
- Sticking with very simple methods (simple statistics, 1D and 2D histograms and statistics, line plots, simple smoothing)
- Sidestepping minor technical problems; flag these instead
- Keeping eyes out for suspicious or intriguing features
- Refactoring: abstracting figure types, code blocks, etc.
- ⇒ Data report document / figure stack
  (! With all processing code in a script / Jupyter notebook!)
- ⇒ Qualitative, intuitive assessment of noise vs. signal
- ⇒ List of features or aspect worthy of further investigation
- ⇒ ? Possibly iterate to Phase 0 with new information

Do not work directly on the command line without saving your code! You will be stuck in data purgatory for all eternity!

Phase 2: Investigating

Being like a detective patiently building a case.

- Brainstorming (with pen and paper) interpretations of the data
- Forming **multiple** hypotheses to explain interesting features
- Countering physical hypotheses with a suitable null hypothesis (e.g. noise or artifacts)
- Sticking with simple methods (see next slide)
- Sidestepping roadblocks (! Do not stop when you hit obstacles!)
- Gathering evidence in support of / opposing these hypotheses
- Setting aside personal preferences (yours and your advisor's)
- Setting aside what everyone believes to be true
- Maintaining a curious, open mind
- Building a case through plots, argumentation, and analysis
- Keeping in mind the limitations of the dataset
- Asking: Is the evidence conclusive?
- Asking: What other datasets / perspectives could be helpful?
- Asking: What other methods may be called for?

Phase 2 Methods

- Clarifying variability at relevant timescales: diurnal, tidal, inertial, annual, etc.
- Separating variability with simple smoothing, harmonic fits, formation of composite cycle (e.g. annual), etc., *and residual*
- Examining all of the Phase 1 aspects on separated components
- Studying theory of relevant processes to familiarize yourself
- Forming simple conceptual, kinematic, or statistical model for the observed features
- Gathering ancillary or environmental data for potential forcing, causative, or associated processes
- Correlations, EOFs, etc.
- Higher-order or circular statistics

Phase 3: Forensics

Bringing in specialize methods to help the investigation.

- Be aware that this is often not necessary!
- Talk to colleagues with more experience
- Ask: do I want to learn this, or instead, find a collaborator?
- Be prepared to sit down and study for weeks or months
- Learn the method thoroughly *before* applying it to your dataset!

Some possibilities:

- Fourier spectral analysis
- Wavelet analysis
- Stochastic modeling
- Interpolation methods: OI, local polynomial fitting, spline
- Correlation analysis: CCA, MCA, MLR, SVD
- Clustering methods
- Statistical hypotheses testing

- ⇒ Definitive evidence *or* inconclusive results
- ⇒ Proceed to Phase 4 *or* iterate with Phase 2

Phase 4: Closing the Case

Being like a lawyer presenting the case to the jury.

- Understanding your results within the context of the literature
- Assessing what makes a unit of scientific progress
- Putting together a case with figures, equations, and arguments
- Considering all possible objections
  (! Use Thinking Hats and role-playing!)
- Iterating to earlier phases as needed
- Building a watertight case
- Being honest about the limitations of your results
- Seeing unanswered questions as future possibilities
- Knowing when to stop
- Remembering to not get personally involved with your client

Also!

- Learning about data formats and conventions; iterate to 0 and 1

⇒ A scientific paper; a finished, shared dataset; open software