Data Science Concepts and Analysis

Week 0: Welcome to PSTAT 100!

- Course introduction
- Course structure
- Getting started

This week

- Course introduction
    - Perspectives on data science
    - Scope and topics
- Course structure
    - Format and schedule
    - Materials and resources
    - Assignments and assessment
    - Course policies

Course introduction

- Perspective on data science: "lifecycle"
- Course scope

What's data science?

Currently understood, "data science" encompasses a wide range of activities that involve *uncovering insights from quantitative information*.
Data scientists typically combine specific interests ("domain knowledge", *e.g.*, biology) with computation, mathematics, and statistics and probability to contribute to knowledge in their communities.

- Skills combine in different proportions -- no singular background among practitioners.
- Diverse communities -- science, industry, government, medicine, academia, etc.

Data science is a nascent field -- many of you will play a role in shaping its development in years to come!

Data science lifecycle

There is an emerging consensus that doing data science involves proceeding through a **lifecycle**: *a repeated sequence of steps*.

- Less consensus at the moment about how many steps and what they are (google 'data science lifecycle' and check out all the flowcharts).

Most versions of the 'data science lifecycle' involve a few categories of steps:

- Project planning
- Data collection and organization
- Exploration
- Analysis
- Communication and interpretation

(Perhaps it is this idea of a lifecycle that characterizes data science as distinct from other quantitative fields.)

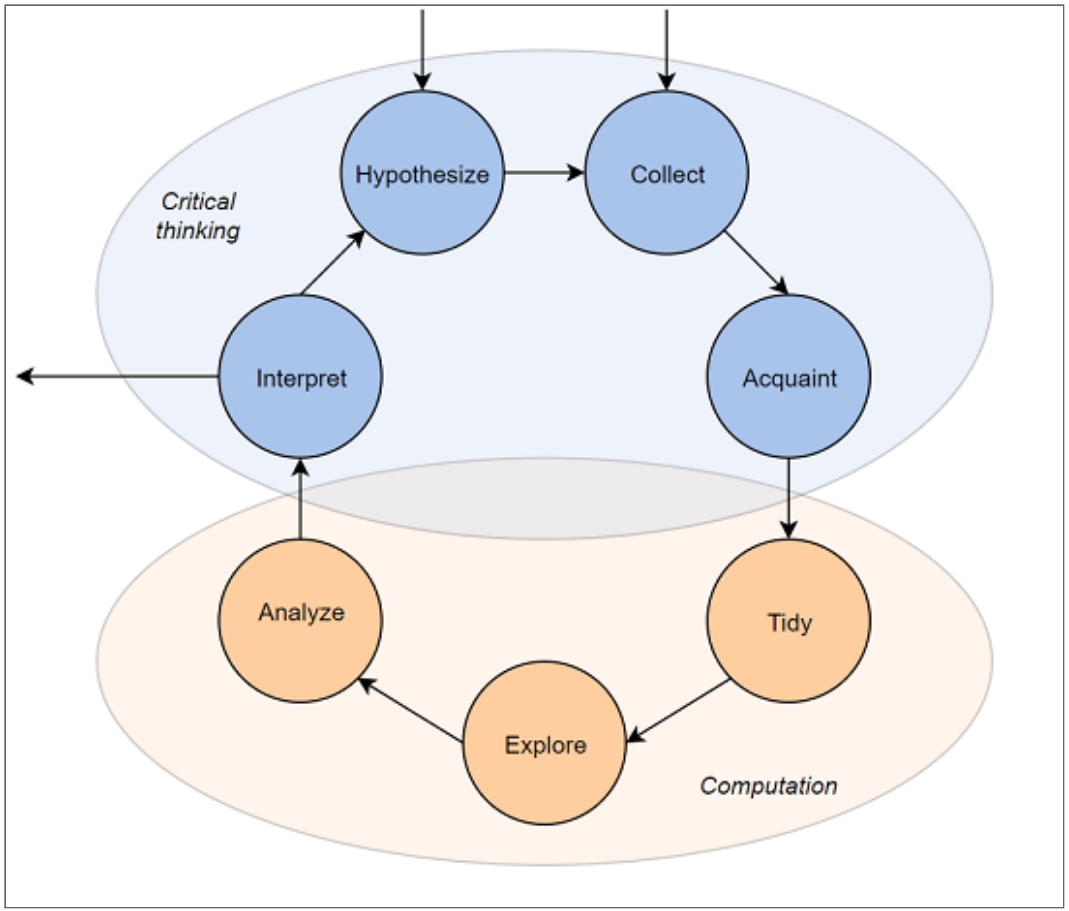
PSTAT 100 lifecylce

In this course, we'll articulate the lifecycle in terms of the following steps.

1. **H**ypothesize: question formulation/refinement.
2. **C**ollect: go out and sample or acquire data 'second-hand'.
3. **Ac**quaint: get to know your dataset; make friends!
4. **T**idy: clean up and organize your data.
5. **E**xplore: search for patterns and structure.
6. **An**alyze: seek to understand.
7. **I**nterpret: explain the meaning of your analysis.

PSTAT 100 lifecycle

No data science lifecycle would be complete without a flowchart!



Notice the multiple entry points -- some projects start with a focused question; others, with a dataset.

Illustrating the cycle

We'll walk through a *very* simple example to get a concrete idea of how the cycle works.

Question: *How do animals' brains scale with their bodies?*

You will see some codes displayed as we walk through the example. Don't worry about understanding them -- that's what this course is for!

Focus on the process.

Step 0: Hypothesize

Question formulation or refinement

There are lots of datasets out there with brain and body weight measurements, so let's make the question a bit more specific:

- *What is the relationship between an animal's brain and body weight?*

It might sound simple, but the relationship is thought to contain clues about evolutionary patterns pertaining to intelligence.

Step 1: Collect

Gather or acquire data

In this case, we won't directly gather data. Instead, we'll acquire a publicly available dataset comprising average body and brain weights for 62 mammals.

In [17]:

```
# import brain and body weights
bb_weights = pd.read_csv('data/allison1976.csv').iloc[:, 0:3]
bb_weights.head()
```

Out[17]:

	species	body_wt	brain_wt
0	Africanelephant	6654.000	5712.0
1	Africangiantpouchedrat	1.000	6.6
2	ArcticFox	3.385	44.5
3	Arcticgroundsquirrel	0.920	5.7

	species	body_wt	brain_wt
4	Asianelephant	2547.000	4603.0

Units of measurement

- body weight in kilograms
- brain weight in grams

Step 2: Acquaint

Get to know the data

Especially because we didn't collect this data ourselves, we should do a little background reading and reflection to understand where the data came from (Allison *et al.* 1976) and what limitations might exist:

- Information about mammals only → no information about birds, fish, reptiles, etc.
- Species weren't chosen to represent mammalia → probably shouldn't seek to generalize
- Averages measured → 'aggregated' data (not individual-level)

So we can only explore the question narrowly for this particular group of animals using the data at hand -- we don't stand to learn anything generalizable.

- Not a bad thing! We can still see what the data suggest and use results for hypothesis generation.

Step 3: Tidy

Clean up and organize

This dataset is already impeccably neat: each row is an observation for some mammal, and the columns are the two variables (average weight).
So no tidying needed -- we'll just check the dimensions and see if any values are missing.

In [3]:

```
# dimensions?
bb_weights.shape
```

Out[3]:

```
(62, 3)
```

In [4]:

```
# missing values?
bb_weights.isna().sum(axis = 0)
```

Out[4]:

```
species      0
body_wt      0
brain_wt     0
dtype: int64
```

Step 4: Explore

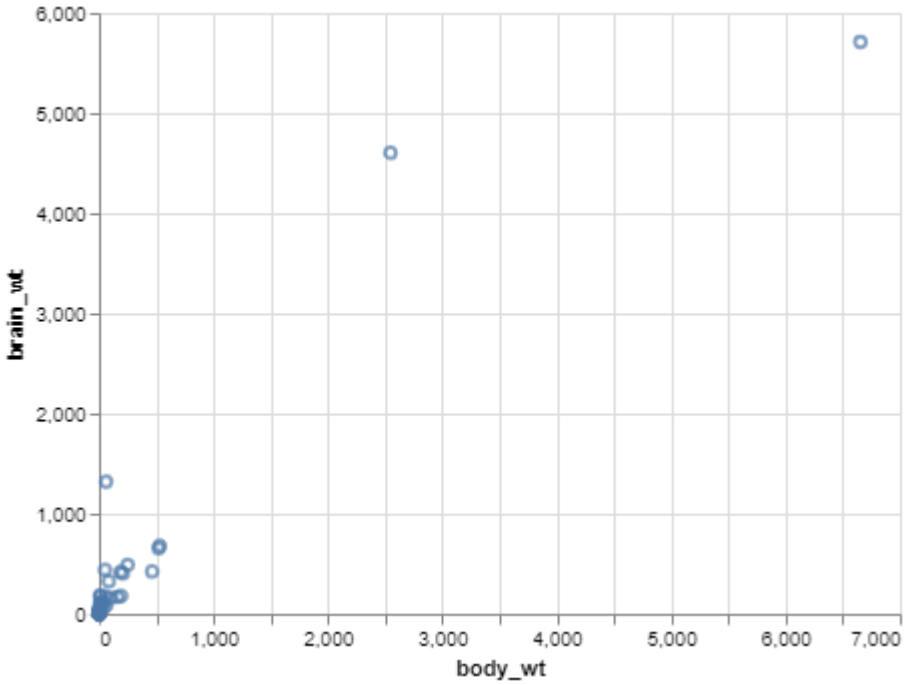
Look for patterns, structure, properites

Visualization is usually a good starting point. Below is a simple scatterplot.

In [5]:

```
# plot
alt.Chart(bb_weights).mark_point().encode(x = 'body_wt', y = 'brain_wt')
```

Out[5]:



Notice the apparent density of points near $(0, 0)$ -- that suggests we shouldn't look for a relationship on the scale of kg/g.

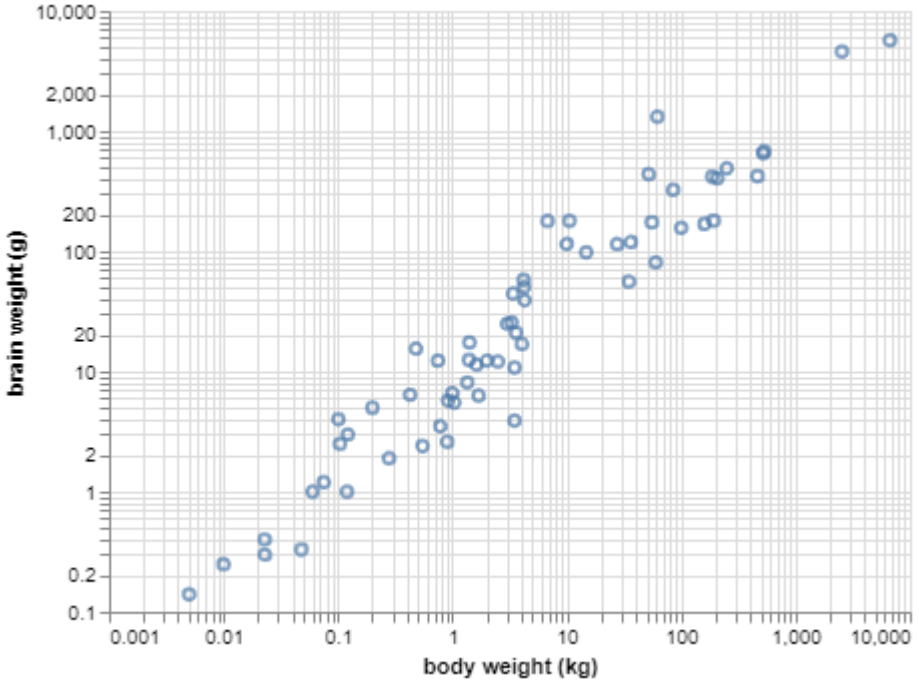
Step 4: Explore

A simple transformation of the axes reveals a clearer pattern.

In [6]:

```
# plot
alt.Chart(bb_weights).mark_point().encode(
    x = alt.X('body_wt', scale = alt.Scale(type = 'log'), title = 'body weight (kg)'),
    y = alt.Y('brain_wt', scale = alt.Scale(type = 'log'), title = 'brain weight (g)')
)
```

Out[6]:


brain weight (g)
10,000
2,000
1,000
200
100
20
10
2
1
0.2
0.1
0.001
0.01
0.1
1
10
100
1,000
10,000
body weight (kg)

Step 5: Analyze

The plot shows us that there's a roughly linear relationship on the log scale:

$$\log(\text{brain}) = \alpha \log(\text{body}) + c$$

Step 6: Interpret

So what does that mean in terms of brain and body weights? A little algebra and we have:

$$(\text{brain}) \propto (\text{body})^{\alpha}$$

This is known as a "power-law" relationship: brain weight changes in proportion to a power of body weight.
So it appears that for these 62 mammals, the brain-body scaling is well-described by a power law. (Notice: no generalization/extrapolation!)

Step 0: Hypothesize

We can now engage in question refinement. Do other classes of animal exhibit the same power law relationship? Is it the same or different from animal to animal?
To investigate, we need richer data.

Step 1: Collect

A number of authors have compiled and published 'meta-analysis' datasets by combining the results of multiple studies.
Below we'll import a few of these for three different animal classes.

In [18]:

```
# import metaanalysis datasets
reptiles = pd.read_csv('data/reptile_meta.csv')
birds = pd.read_csv('data/bird_meta.csv', encoding = 'latin1')
mammals = pd.read_csv('data/mammal_meta.csv', encoding = 'latin1')
```

Step 2: Acquaint

Where does this data come from? It's kind of a convenience sample of scientific data:

- Multiple studies → possibly different sampling and measurement protocols
- Criteria for inclusion unknown → probably neither comprehensive nor representative of all such measurements taken

So these data, while richer, are still relatively narrow in terms of generalizability.

Step 3: Tidy

These datasets are still quite neat, but have a few minor things out of order.

In [8]:

```
# variable names and positions don't quite match up
reptiles.iloc[0:3, [0, 1, 2, 3, 7, 9]]
```

Out[8]:

	Order	Family	Genus	Species	Sex	Body weight (g)
0	Crocodilia	Alligatoridae	Alligator	mississippiensis	m	205000.0
1	Crocodilia	Alligatoridae	Alligator	mississippiensis	m	173000.0
2	Crocodilia	Crocodylidae	Crocodylus	acutus	f	134000.0

In [9]:

```
birds.iloc[0:3, [0, 1, 2, 3, 6, 11]]
```

Out[9]:

	Order	Family	Genus	Species	Sex	Body mass (g)

	Order	Family	Genus	Species	Sex	Body mass (g)
0	Accipitriformes	Accipitridae	Accipiter	cirrocephalus	f	202.7
1	Accipitriformes	Accipitridae	Accipiter	cirrocephalus	f	203.9
2	Accipitriformes	Accipitridae	Accipiter	cirrocephalus	f	200.5

Step 3: Tidy

In order to comine the datasets:

- Select columns of interest;
- Put in consistent order;
- Give consistent names;
- Concatenate.

I've suppressed the detail, but we can now inspect the result.

In [12]:

```
# missing values?
data.isna().mean(axis = 0)
```

Out[12]:

```
Order      0.00000
Family     0.00000
Genus      0.00000
Species    0.00000
```

```
Sex         0.00000
body        0.00000
brain       0.57404
class       0.00000
dtype: float64
```

This dataset has a number (actually quite a lot) of missing brain weight measurements: many of the studies combined to form these datasets did not include that particular measurement.

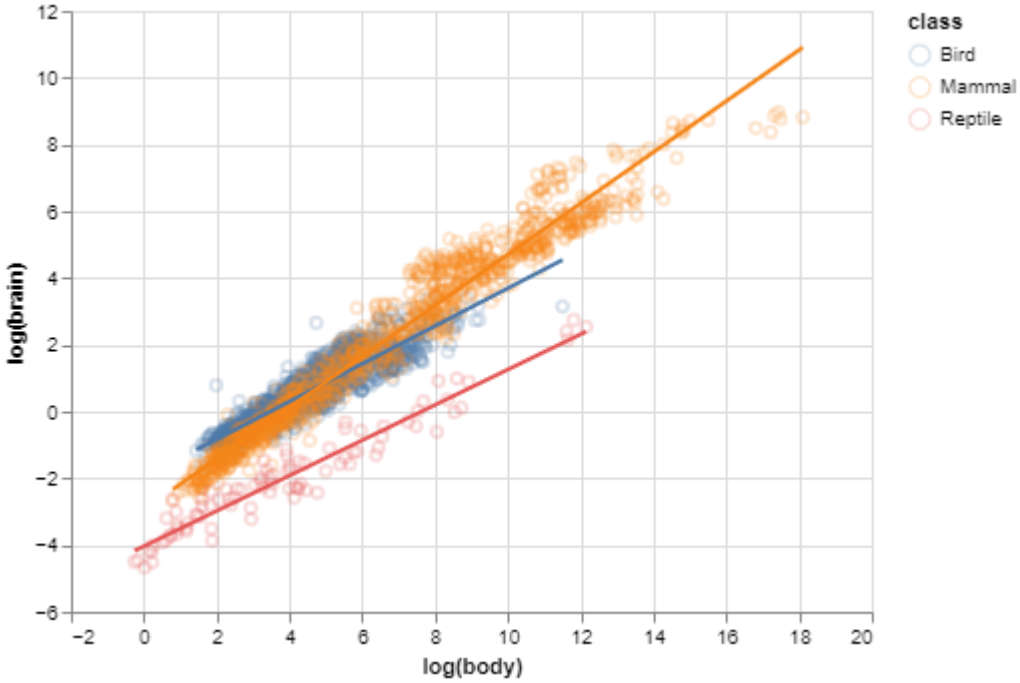
Step 4: Explore

Looking at a similar plot and overlaying trend lines, we see the same power law relationship but with *different* proportionality constants for the three classes of animal.

In [16]:

```
scatter + trend
```

Out[16]:


log(brain)
log(body)
class
Bird
Mammal
Reptile

Step 5: Analyze

So in this case there are three different linear relationships on the log scale that depend on animal class:

$$\begin{aligned}
(\text{brain}) &= \beta_1(\text{body})^{\alpha_1} \quad (\text{mammal}) \\
(\text{brain}) &= \beta_2(\text{body})^{\alpha_2} \quad (\text{reptile}) \\
(\text{brain}) &= \beta_3(\text{body})^{\alpha_3} \quad (\text{bird}) \\
\beta_i \neq \beta_j, \alpha_i &\neq \alpha_j \quad \text{for } i \neq j
\end{aligned}$$

Step 6: Interpret

It seems that the brain and body weights of the birds, mammals, and reptiles measured in these studies exhibit distinct power law relationships.
What would you investigate next?

- Explore further?
    - Seek data on additional animal classes
    - Seek data on correlates of body weight
    - Seek data on other variables (lifespan, habitat, predation, etc.)
- Inference and prediction?
    - Find better generalizable data
    - Estimate the α_i's and β_i's
    - Find a way to predict brain weights for unobserved species

- Something else?

Lather, rinse, repeat

Hopefully you can see how we could go through multiple iterations of the cycle, continuing to refine the question and produce more detailed analyses each time, until we arrive at a fuller understanding of the subject under study.

A comment

Notice that I did not mention the word 'model' anywhere!

This was intentional -- it is a common misconception that analyzing data always involves fitting models.

- Models are not not always necessary or appropriate
- We learned a lot from plots alone

Scope for PSTAT 100

This term we'll work on developing your data science toolkit with foundational skills:

- Programming in Python data science libraries
- Critical thinking about sampling and generalizing from data
- Visualization and exploratory analysis
- Basic statistical models

Throughout, we'll explore applications of these tools to case studies.

Course structure

- Schedule and format
- Course pages and materials
- Assignments and assessment
- Course policies

Weekly Pattern

We'll follow a simple weekly pattern:

- **Mondays at 9am PST** (release date)
    - Weekly announcement posted
    - Reading, lectures, and lab released
- **Fridays at 5pm PST** (due date)
    - Lab due
    - Homeworks released/due biweekly

Course pages & materials

Pages

Course page	Primary use
Gauchospace	Announcements and links to content
Nectir	Communication (preferred over email!)
pstat100.lsit.ucsb.edu	Computing
Gradescope	Assignment submission and grade tracking

Materials

- **Python Data Science Handbook** (DSH)
- **Principles and Techniques of Data Science** (TDS)
- Articles and excerpts as assigned

Tentative schedule

I prefer to count weeks from 0 to 9. Here's what we'll aim to cover:

Week	Topic	Subjects	Lifecycle
0	Introduction	What's data science?	
1	Tidy data	Import and organization	Collect/Acquiant/Tidy
2	Sampling	Informative vs. uninformative data	Collect/Acquaint/Tidy
3	Visualization	Plot types, aesthetics, principles	Explore
4	Exploratory analysis	Density estimation and descriptive statistics	Explore/Analyze
5	Exploratory analysis	Principal components analysis	Explore/Analyze

Week	Topic	Subjects	Lifecycle
6	Linear models	Simple linear regression	Analyze/Interpret
7	Linear models	Multiple linear regression	Analyze/Interpret
8	Classification	Logistic regression	Analyze/Interpret
9	TBD	TBD	

The last week is flex time to explore other topics or extend coverage of previous topics.

Assessments

- **Labs** (40% final grade weight, 10 pts each)
    - Short/moderate-length guided programming assignments
    - Given weekly through week 8
    - Walkthrough videos provided
    - Collaboration encouraged; individual submissions required

***Lab objective**: introduce and develop core skills with data science libraries in Python.*

Assessments

- **Homeworks** (40% final grade weight, 50 pts each)
    - Applications of course ideas and lab skills to analyses of real datasets
    - 4 assignments, released/due biweekly
    - Collaboration encouraged and group submissions allowed

***Homework objective**: practice workflow and explore case studies.*

Assessments

- **Project** (20% final grade weight, 50 pts each)
    - Open-ended data analysis based on your interests
    - 2 stages
        - interim preparation and planning report due mid-quarter
        - final report due end-of-quarter
    - Collaboration expected

Project objective: *apply learned skills to a problem of your choosing.*

Policies

- **Deadlines and late work**
    - One-hour grace period on all deadlines
    - Two free lates on any assignment (except final project report)
    - 75% partial credit thereafter
    - No late work beyond 72 hours after deadline without instructor permission
- **Communication**
    - Nectir, office hours, appointments, email
    - 48 weekday-hour response policy for written communication
        - Remind us if a message is missed!
- **Accomodations**

- In light of remote teaching and learning, I am willing to work with DSP students on additional accommodation needs

Getting started

Here's a short to-do list to get started in the course:

- Check access to course pages
- Fill out welcome survey (GS -> General)
- Introduce yourself on Nectir
- Week 0 reading and lab 0 (due Friday)
    - Don't worry, the first lab is short!