Data Analysis 4

Banneker Institute Jul 18 - 22

Goal of this week:

To solidify your knowledge of Bayesian statistics and equip you with the tools to tackle new datasets and complex models

My “Open Question” Policy…

My “Open Question” Policy…

Ask!

Where we're heading this week..

1. Review of Bayesian statistics and metrics
2. How do we "actually" analyze data?: Typical workflow
3. Graphical models to represent statistical models
4. Working through a toy problem
5. Wrapping up: Realistic considerations and when to ask for help

Bayesian Statistics and Metrics

Bayes' Theorem Revisited

Bayes' Theorem Revisited

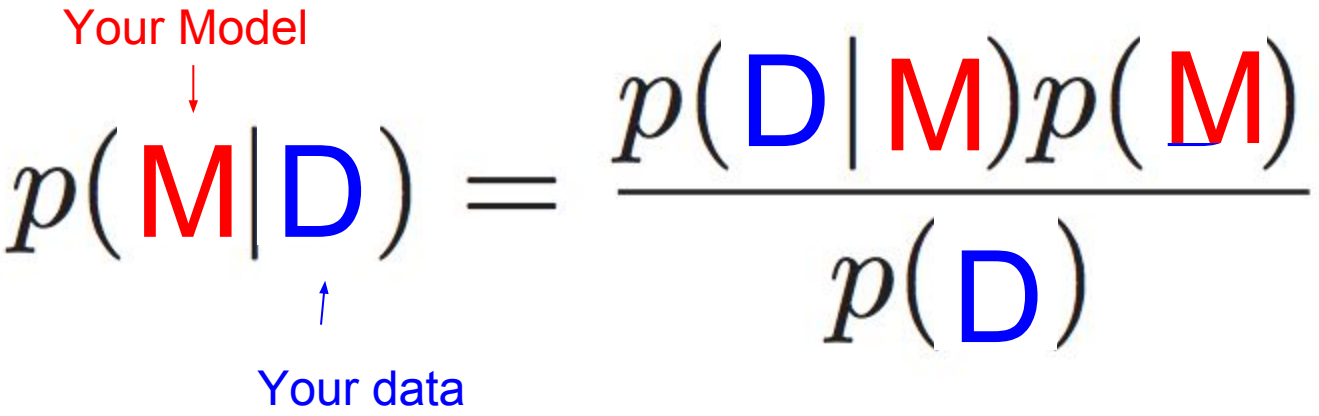
$$p(\mathrm{M}|\mathrm{D}) = \frac{p(\mathrm{D}|\mathrm{M})p(\mathrm{M})}{p(\mathrm{D})}$$

Bayes' Theorem Revisited

Your Model

$$p(\mathrm{M}|\mathrm{D}) = \frac{p(\mathrm{D}|\mathrm{M})p(\mathrm{M})}{p(\mathrm{D})}$$

Bayes' Theorem Revisited



Bayes' Theorem Revisited

Prior

$$p(\mathrm{M}|\mathrm{D}) = \frac{p(\mathrm{D}|\mathrm{M})p(\mathrm{M})}{p(\mathrm{D})}$$

Bayes' Theorem Revisited

Likelihood

$$p(\mathrm{M}|\mathrm{D}) = \frac{p(\mathrm{D}|\mathrm{M})p(\mathrm{M})}{p(\mathrm{D})}$$

Bayes' Theorem Revisited

$$p(\mathrm{M}|\mathrm{D}) = \frac{p(\mathrm{D}|\mathrm{M})p(\mathrm{M})}{p(\mathrm{D})}$$

Posterior

Bayes' Theorem Revisited

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)}$$

The thing we ignore
;]

Practice with Conditional Probability

What is the probability that Dogmeat wins given the fact that it's raining?

	Raining	Not Raining
Dogmeat wins	3	2
Dogmeat loses	1	6

Metrics

What is Chi^2?

Why do we use it?

Metrics

What is Chi^2:

- Geometrically: It's similar to "least squares"
- Mathematically: It's a descendent of a Gaussian Likelihood:

$$p(x_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(x_j - \hat{x}_j)^2 \right),$$

where $\hat{x}_j$ is the predicted value of x_j

Metrics

Why do we use it:

- Almost everything is Gaussian (due to the law of large #s)
- It’s “easy”

Metrics

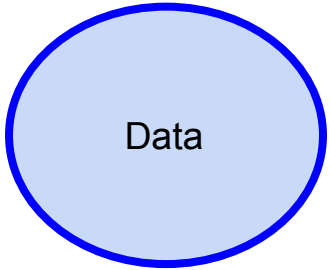
A metric is (typically) a scalar which measures the “goodness” of your model’s fit to the data

It is something you’re trying to optimize

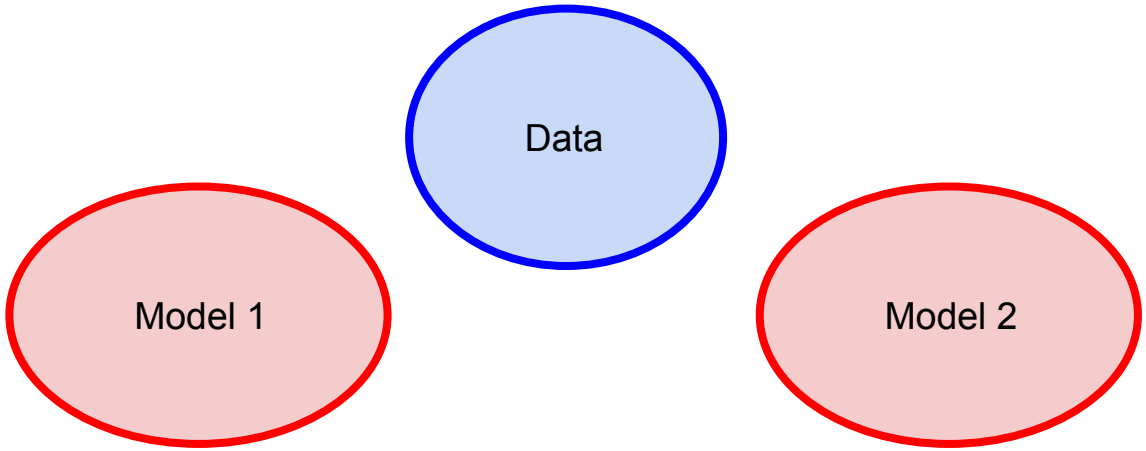
Our Data Analysis Workflow

A Typical Workflow...

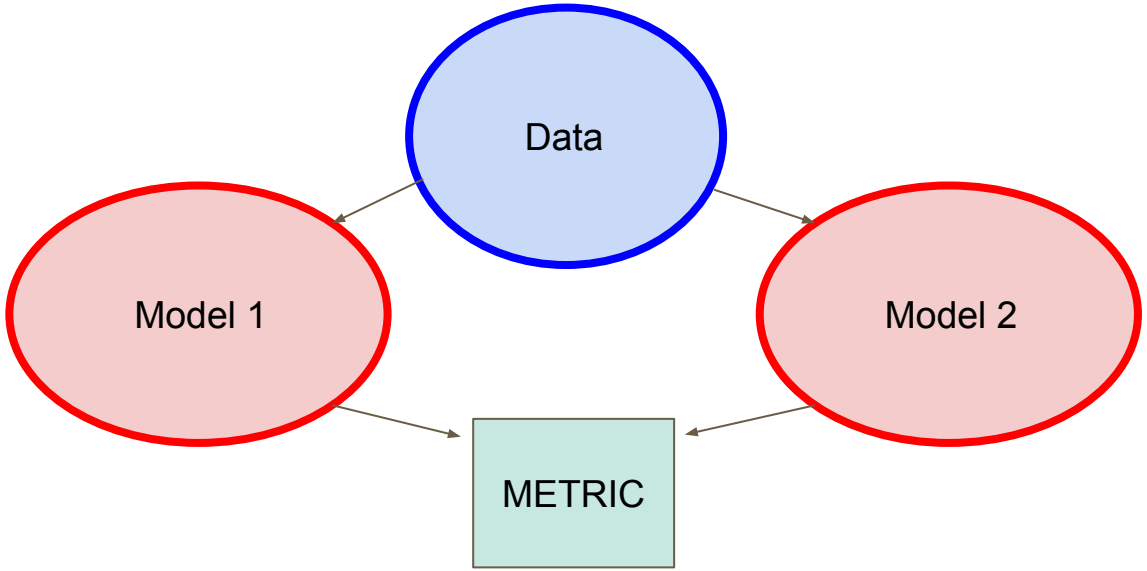
A Typical Workflow...



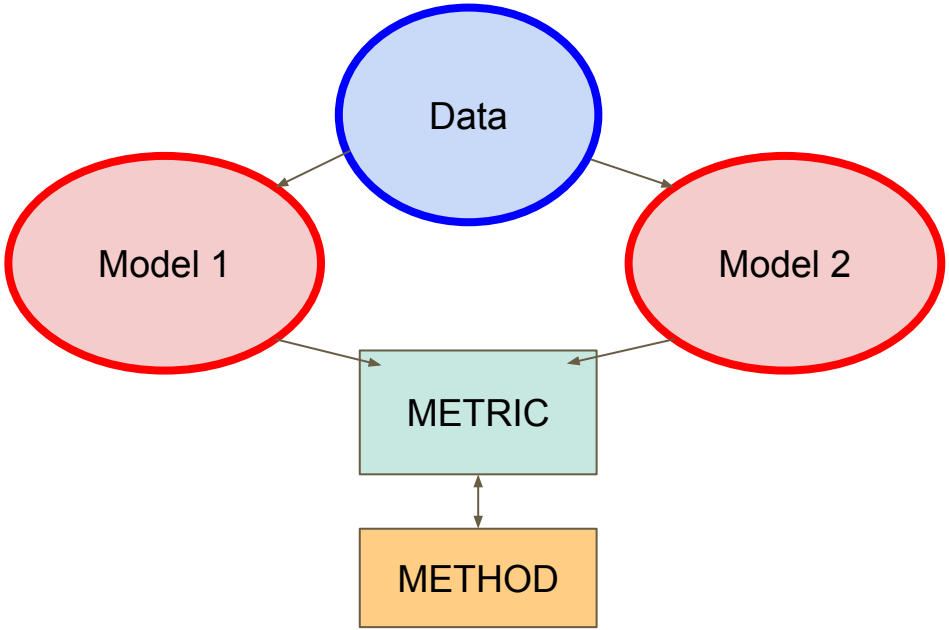
A Typical Workflow...



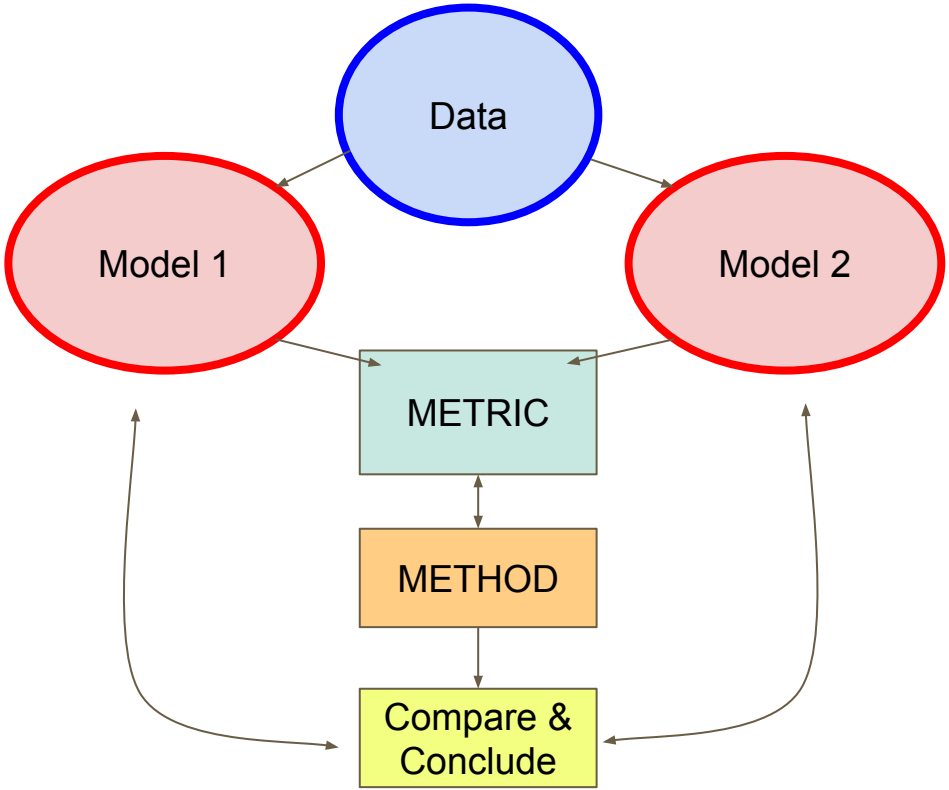
A Typical Workflow...



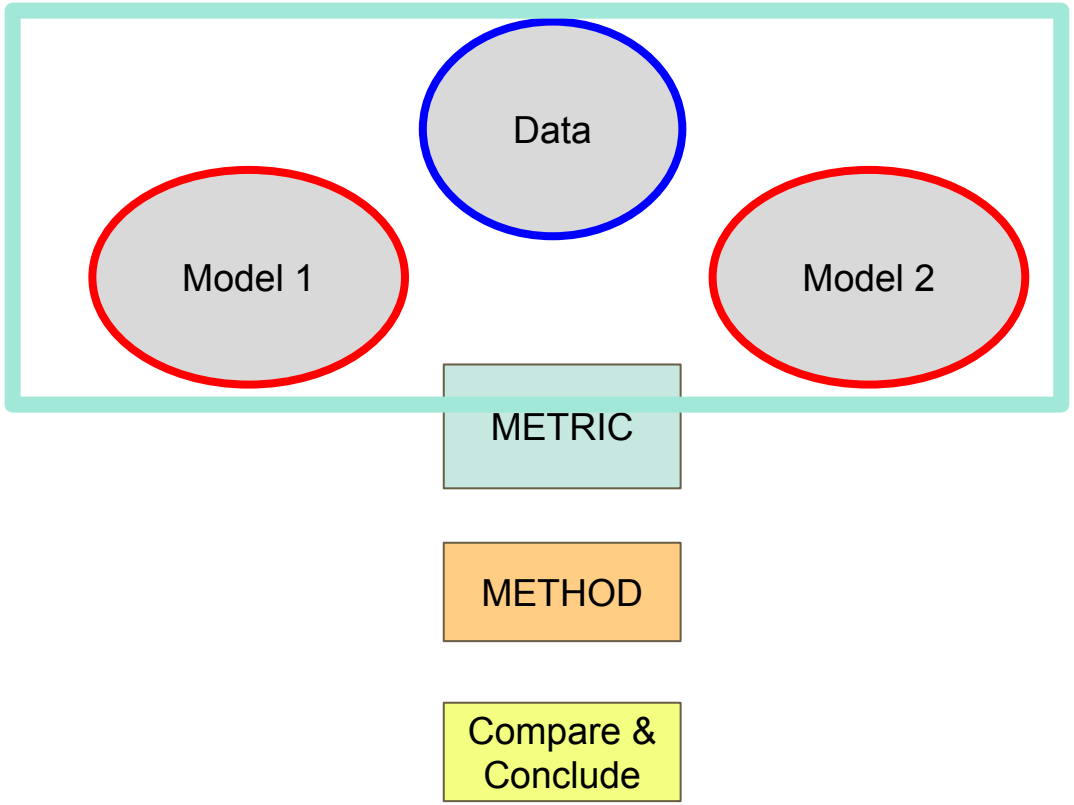
A Typical Workflow...



A Typical Workflow...

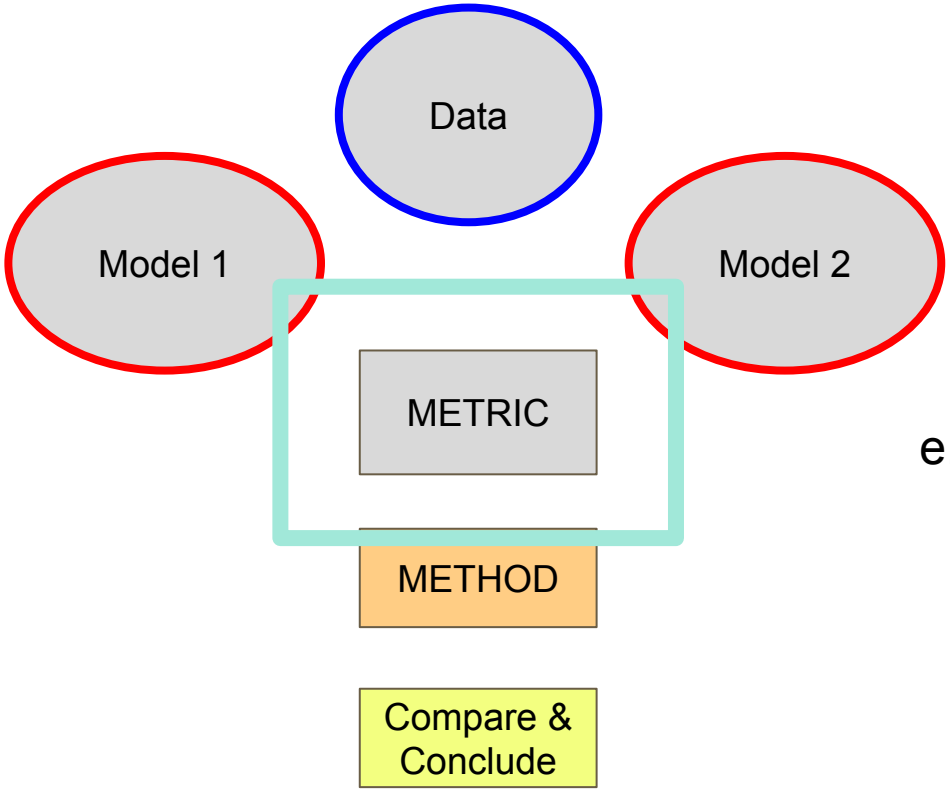


A Typical Workflow...



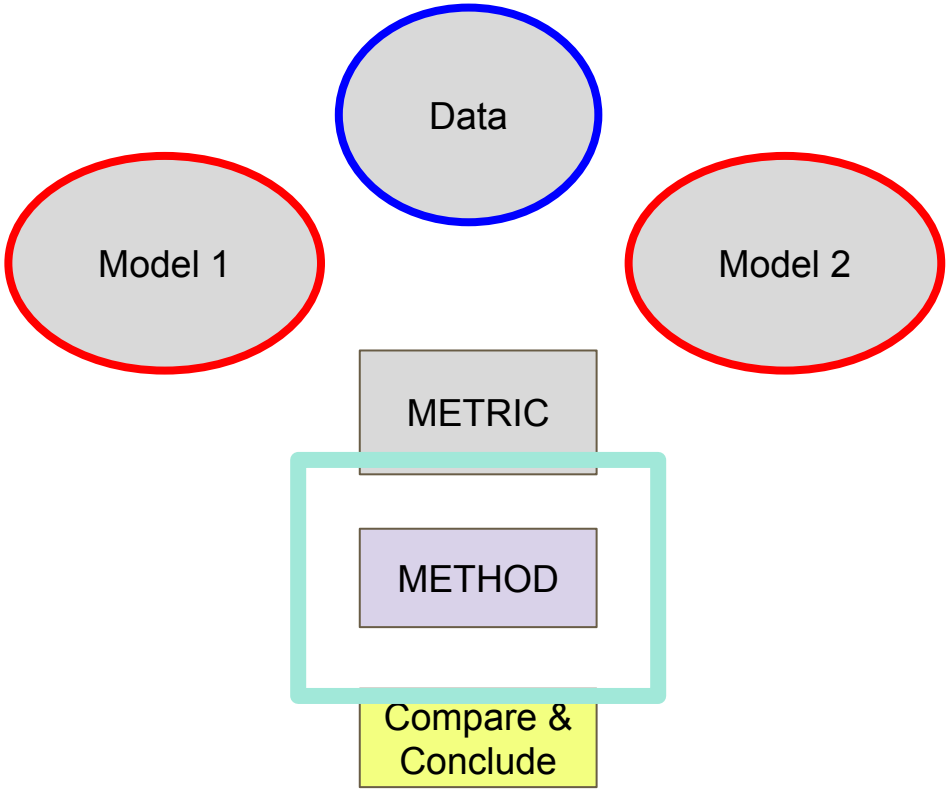
You do this part!
(and map with graphical models)

A Typical Workflow...



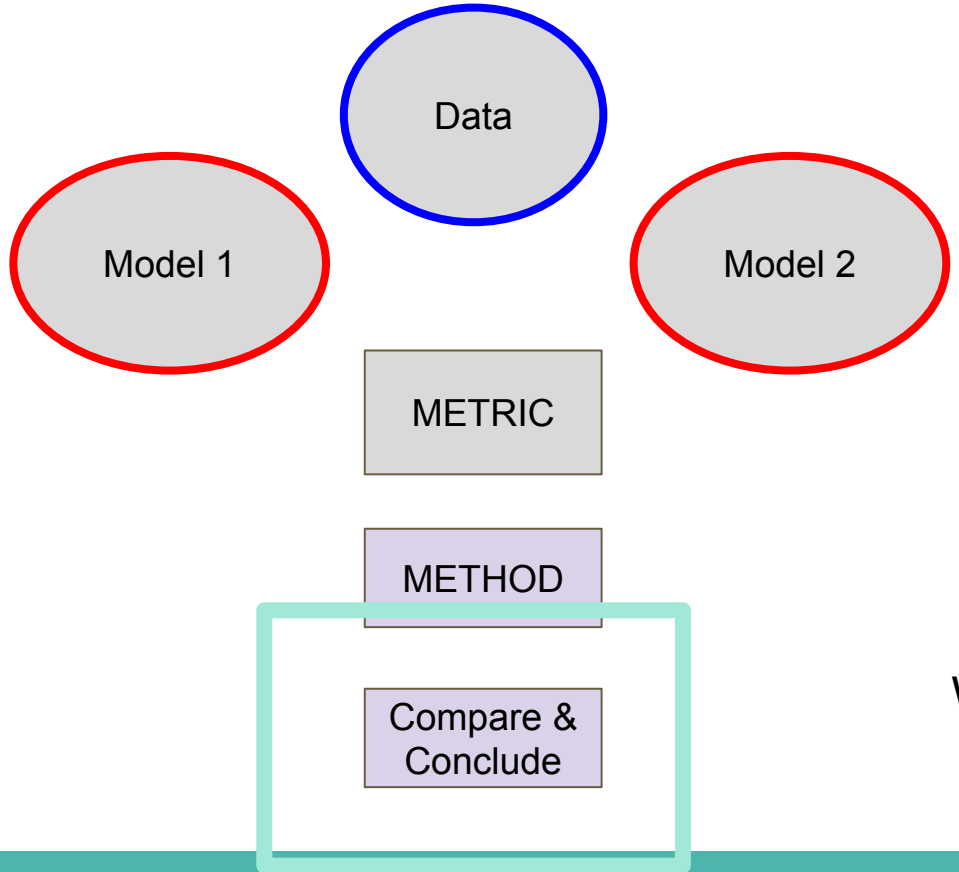
We did this!
e.g. Chi^2 or likelihoods

A Typical Workflow...



We haven't done this!
e.g. Least Squares, MCMC

A Typical Workflow...



We're (probably) NOT doing this!

Questions?