

---

# **CSE 519: Data Science**

## **Steven Skiena**

### **Stony Brook University**

---

Lecture 13: Building Models

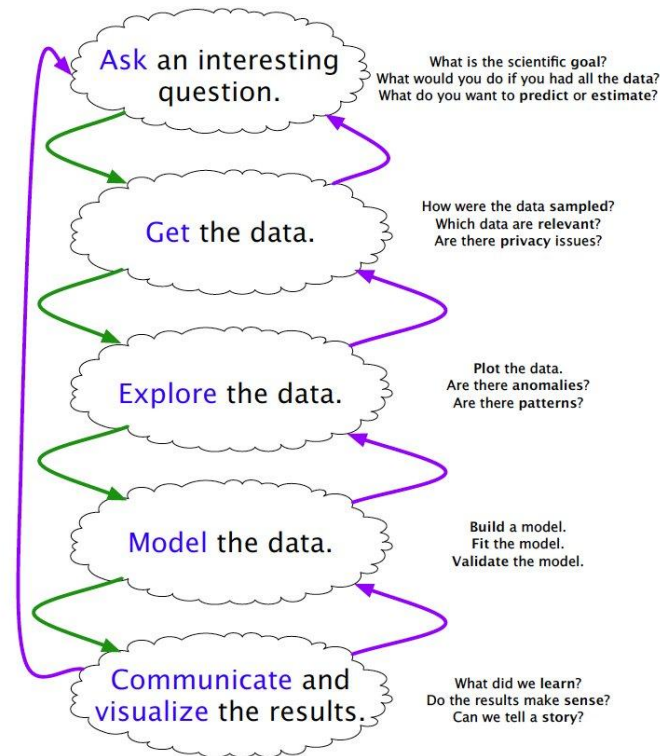
---

# The Data Science Analysis Pipeline

---

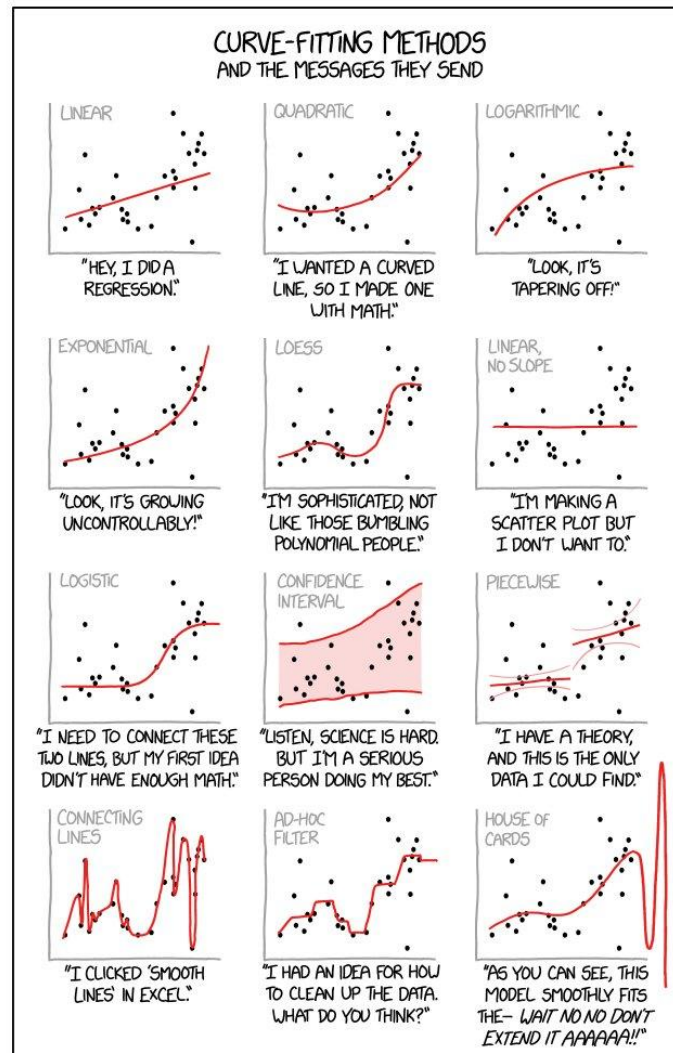
Modeling is the process of encapsulating information into a tool which can make forecasts/predictions.

The key steps are building, fitting, and validating the model.



# Which is Best?

There are many ways to model any given data set. How can we decide which approach is better?



# Philosophies of Modeling

---

We need to think in some fundamental ways about modeling to build them in sensible ways.

- Occam's Razor
  - Bias-Variance trade offs
  - Nate Silver: The Signal and Noise
-

# Occam's Razor

---

This philosophical principle states that “the simplest explanation is best”.

With respect to modeling, this often means minimizing the parameter count in a model.

Machine learning methods like LASSO/ridge regression employ penalty functions to minimize features, but also do a “sniff test”.

---

# Bias-Variance Tradeoffs

---

“All models are wrong, but some models are useful.”

– George Box (1919-2013)

- *Bias* is error from erroneous assumptions in the model, like making it linear. (underfitting)
- *Variance* is error from sensitivity to small fluctuations in the training set. (overfitting)

First-principle models likely to suffer from bias, with data-driven models in greater danger of overfitting.

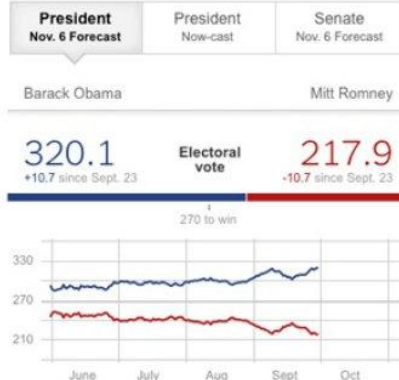
---

# What would Nate Silver do?



## FiveThirtyEight Forecast

Updated 12:27 AM ET on Oct. 1



## State-by-State Probabilities



## Electoral Vote Distribution

The probability that President Obama receives a given number of Electoral College votes.



# Principles of Nate Silver

---

- Think probabilistically
  - Change your forecast in response to new information.
  - Look for consensus
  - Employ Bayesian reasoning
-



# The Output of Your Models

---

Demanding a single deterministic “prediction” from a model is a fool’s errand.

Good forecasting models generally produce a probability distribution over all possible events.

Good models do better than baseline models, but you could get rich predicting if the stock market goes up/down with  $p > 0.55$ .

---

# Properties of Probabilities

---

- They sum to 1.
- They are never negative.
- Rare events do not get probabilities of zero.

Probabilities are a measure of humility in the accuracy of the model, and the uncertainty of a complex world.

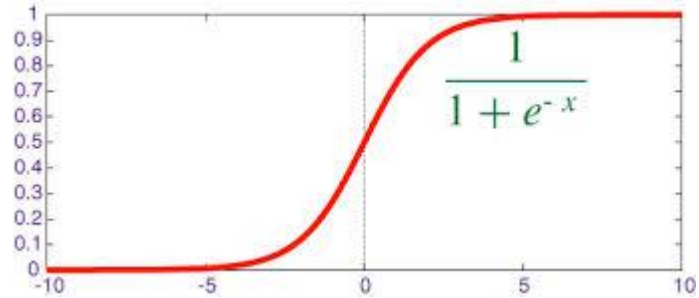
**Models must be honest in what they do/don't know.**

---

# Scores to Probabilities

---

The logit function maps scores into probabilities using only one parameter.



Summing up the “probabilities” over all events  $s$  defines the constant  $1/s$  to multiply each so they sum up to 1.

---

# Live Models

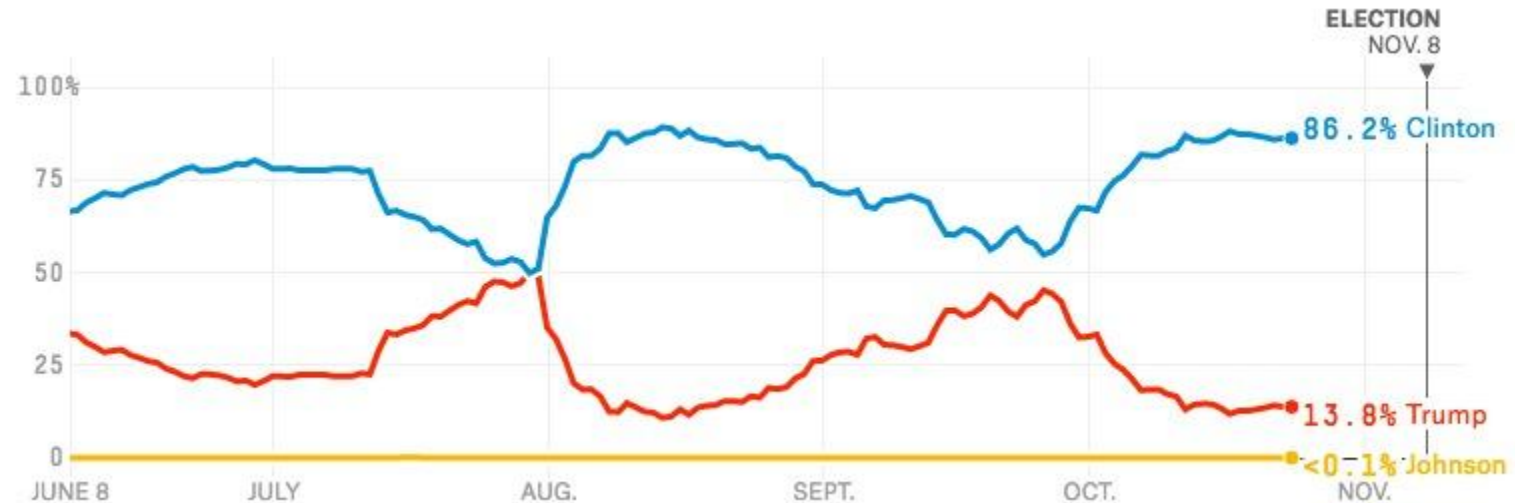
---

A model is *live* if it continually updating predictions in response to new information.

- Does the forecast ultimately converge on the right answer?
  - Does it display past forecasts so the user can judge the consistency of the model?
  - Does the model retrain on fresher data?
-

# Presidential Election Forecast, 2016

---



# Look for Consensus

---

- Are there competing forecasts you can compare to, e.g. prediction markets?
- What do your baseline models say?
- Do you have multiple models which use different approaches to making the forecast?

Boosting is a machine learning technique which explicitly combines an ensemble of classifier.

---

# Google Flu Trends

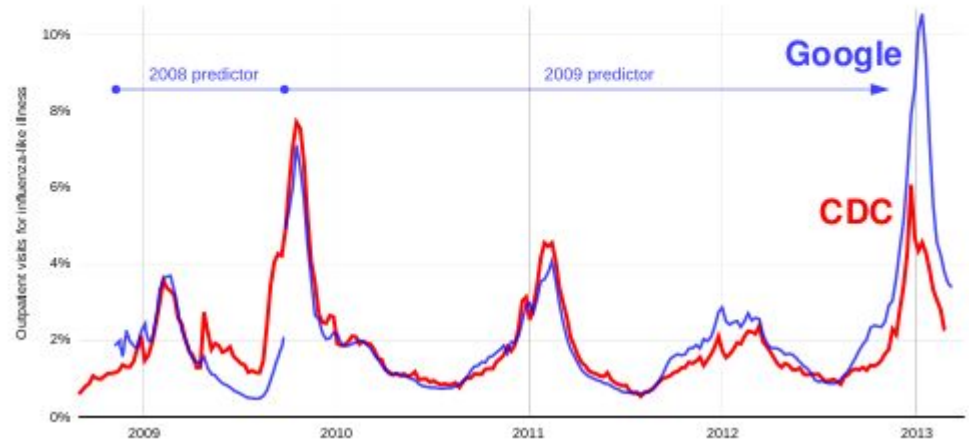
---



Predicted flu outbreaks using query frequency of illness terms.

The model failed after Google added search suggestions

Second divergence in 2012–2013 for U.S.



# Modeling Methodologies

---

- **First principle models**: based on a **theoretical** explanation of how the system works (like simulations, scientific formulae)
- **Data-driven models**: based on observed **data** correlations between input parameters and outcome variables.

Good models are typically a mixture of both.

---



# Principled or Data Driven? (Projects)

---

- Miss Universe?
  - Movie gross?
  - Baby weight?
  - Art auction price?
  - Snow on Christmas?
  - Super Bowl / College Champion?
  - Ghoul Pool?
  - Future Gold / Oil Price?
-

# Baseline Models

---

“A broken clock is right twice a day.”

The first step to assess whether your model is any good is to build **baselines**: the simplest *reasonable* models to compare against.

Only after you decisively beat your baselines can your models be deemed effective.

---

# Representative Baseline Models

---

- Uniform or random selection among labels.
- The most common label in the training data.
- The best performing single-variable model.
- Same label as the previous point in time.
- Rule of thumb heuristics.

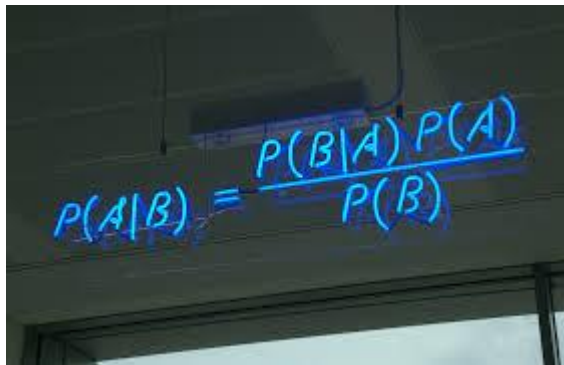
Baseline models must be fair: they should be simple but not stupid.

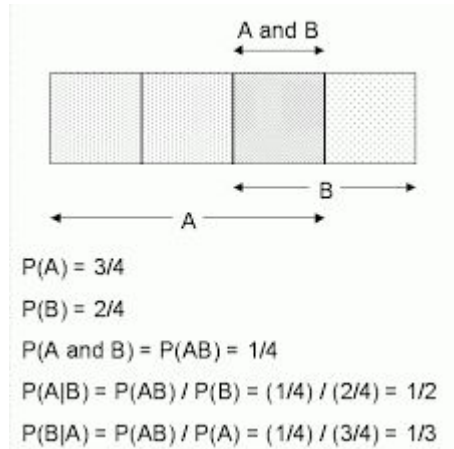
---

# Bayesian Reasoning

Bayes' Theorem lets us update our confidence in an event in response to fresh evidence.

Bayesian reasoning reflects how a **prior** probability  $P(A)$  is updated to given the **posterior** probability  $P(A|B)$  in the face of a new observation  $B$  according to the ratio of the **likelihood**  $P(B|A)$  and the **marginal** probability  $P(B)$


$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$



# Project Baseline Models

---

- Miss Universe?
  - Movie gross?
  - Baby weight?
  - Art auction price?
  - Snow on Christmas?
  - Super Bowl / College Champion?
  - Ghoul Pool?
  - Future Gold / Oil Price?
-

# Taxonomy of Models

---

Models have different properties inherent in how they are constructed:

- Linear vs. nonlinear
  - Discrete vs. continuous models
  - Black box vs. descriptive
  - Stochastic vs. deterministic
  - Flat vs. hierarchical
-

# Discrete vs. Continuous Models

---

**Discrete models** manipulate discrete entities. Representative are discrete-event simulations using randomized (Monte Carlo) methods.

**Continuous models** forecast numerical quantities over reals. They can employ the full weight of classical mathematics: calculus, algebra, geometry, etc.

---

# General vs. Ad Hoc Models

---

Machine learning models for classification and regression are **general**, meaning they employ no problem-specific ideas, only specific data.

**Ad hoc models** are built using domain-specific knowledge to guide their structure and design.

Data science generally seek general models, but I think ad hoc models can be better.

---



# Steps to Build Effective Models

---

- Identify the best output type for your model, likely a probability distribution.
  - Develop reasonable baseline models.
  - Identify the most important levels to build submodels around.
  - Test models with out-of-sample predictions.
-