Cutting the Georgian Knot

The knot

Synthetic data ---> Simulator ---> Model ---> Needs (synthetic) data

Possible places to cut the knot

- **Synthetic data comes from a simulator.**
  - We could find a source with similar patterns or conditions, and use those in data modeling (source: Wikipedia).
  - Can we use nonlinear dynamics and chaos theory, i.e., nonlinear differential equations to model the domain and generate data?
  - If little data is available, try MCMC, e.g., Gibbs sampling, and then a human the loop approach such as SMO (generate some data, have a model check which examples are the most dubious, ask human).
  - Review of methods in this large paper.
    - Examples: domain randomization, compositing real data, etc. followed by methods in Section 6.
  - **Disadvantages:**
    - You may start with crap data that has no relevance to the actual data
      - **Rebuttal:** (i) it's still a best-effort estimate (ii) you could try slowly generating better data, like what GANs do
    - Very difficult with very novel domains
      - **Rebuttal:** In such cases, the problem was already expensive to model. We're *reducing* the cost by involving the domain expert less.
    - Some of these may take a while to run
      - **Rebuttal:** Longer is not always worse. We can wait a couple of days if it means saving money that would otherwise be spent on expensive domain experts. We don't need to use the longest/best method, either. Start with a "good enough" approach, then improve on it.
- **Simulators are models.**
  - Not much argument here.
- **Models need synthetic data to learn.**
  - See examples of "domain randomization".
  - **Disadvantages:**

- May not model real data well
    - **Rebuttal:** If we have no real data, we have no standard to compare against. If we have little data, use statistical methods.
- Model may not transfer well
    - **Rebuttal:** Initially, it may not; however, if we *know* it hasn't transferred well, we probably have *some* real data by now. Then, use a data generation approach that is guided by this real data. This is less challenging than blindly trying to get synthetic data.

Clearly, the most promising place to cut the knot is in the first step, i.e., the assumption that synthetic data comes from a simulator. Need to read the review paper for a better understanding.