Steps for Approaching ML on Big Data

Gradual Escalation

Effect of Bigness on ML Models

- Accuracy
  - May improve may not (Netflix example)
- Development Time
  - Surely increases

- Conclusion: Before launching into BD ML Modeling make sure you need it.

Incremental Escalation

- Sampling
  - Pick sample size, sample, train test
  - Pick larger sample – repeat
  - If performance flattens – You're done
- Consolation Prize: If you need big data ML, at least you've gotten feel for the problem
  - Feature engineering
  - Algorithm selection

Next step: theater nukes

- For some problems
  - Use large scale clustering to break problem into many smaller (single CPU) parallel problems.
  - Example: Micro-targeting: Cluster customers into small groups.
  - Train independent model on each cluster. (ebay)

Big Data Tools Required

- Sampling algo
- Clustering algo