Data for Machine Learning

Prasanna Balaprakash
Computer Scientist
Mathematics and Computer Science Division & Leadership Computing Facility
Argonne National Laboratory
Types of data

Regular domain
- Spatial
- Temporal
- Spatio temporal
- Tabular

Irregular domain
- Graphs
- Temporal on graphs
- Point cloud
Regular domain

- Spatial (2D, 3D)
  - Images
    - Large data set
      - Deep learning
        » Conv. Neural Nets
    - Small data set
      - Z + Shallow learning
        » Z is domain-specific
        » Gaussian Process, Support Vector, Bagging, Boosting
      - Z + stats (geographical spatial data)
        » \texttt{https://cran.r-project.org/view=Spatial}
Regular domain

• Temporal
  – Time series
• Large data set
  – Deep learning
    » Temporal Conv
    » Recurrent NNs
• Small data set
  – Z + Stats
    » Z is domain-specific
    » https://cran.r-project.org/view=TimeSeries
Regular domain

- Spatiotemporal
  - Large data set
    - Deep learning
      » Conv NN + {TCN, RNN}
  - Small data set
    - Z + Stats
      » Z is domain-specific
      » https://cran.r-project.org/view=SpatioTemporal
Regular domain

- Uni/Multivariate/Tabular
  - $Y = f(X)$ type relationship
    - Large data set
      - Deep learning
        » Multi Layer Perceptron
    - Small data set
      - $Z +$ Shallow learning
        » $Z$ is domain-specific
        » Gaussian Process, Support Vector, Bagging, Boosting
Irregular domain

• Graphs
  – $Y = f(V, E)$ type relationship
    • $V$: vertices; $E$: edges
  • Large data set
    – Deep learning
      » Graph Conv NN
  • Small data set
    – $Z$ + Shallow learning
      » $Z$ is domain-specific
      » Gaussian Process,
      Support Vector, Bagging, Boosting
Irregular domain

- Temporal on graphs
  - $Y = f(V, E, T)$ type relationship
    - $V$: vertices; $E$: edges; $T$: time series
  - Large data set
    - Deep learning
      » Graph CNN + RNN
  - Small data set
    - $Z$ + Shallow learning
      » $X$ is domain-specific
      » Gaussian Process, Support Vector, Bagging, Boosting
Irregular domain

• Point clouds
  – \( Y = f(3d \text{ coordinates in space}) \)
    • Large data set
      – Deep learning
        » Pointnet
    • Small data set
      – \( Z \) + Shallow learning
        » \( Z \) is domain-specific
        » Gaussian Process, Support Vector, Bagging, Boosting
Data requirement
Bias variance tradeoff

- All supervised learning algorithms seek to reduce bias and variance in a different way
- No free lunch: no single algorithm will work well on all data sets