Data Driven Discovery of Models (D³M)

Wade Shen
D³M: Data-driven discovery of models

Today: Manual

- Model: representation of a real-world system
  - Inferring locations of images
  - Prediction of election outcomes
  - Estimation model for disease outbreaks
- Manual process: 10-1000s of person-years
- Teams of experts required to develop the model

Tomorrow: Automated

- Automatically select problem-specific model primitives
- Extend the library of modeling primitives
- Automatically compose complex models from primitives
- Facilitate user interaction with composed models
D³M: Accelerate scientific discovery and data analysis

- Discover empirical models having complexity beyond current human comprehension
  - Humans can search only a tiny fraction of model space
  - Machines can search a much larger fraction much more rapidly

- Fast, automated model discovery enables:
  - Accelerated scientific discovery
  - Rapid data analysis w/o embedded data scientists

<table>
<thead>
<tr>
<th>Year</th>
<th>Cost (Person-months)</th>
<th>Avg. time to solution (Months)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>As-performed</td>
<td>with D³M (estimated)</td>
</tr>
<tr>
<td>2009</td>
<td>432</td>
<td>4</td>
</tr>
<tr>
<td>2009-2011</td>
<td>126</td>
<td>5</td>
</tr>
<tr>
<td>2014-2015</td>
<td>102</td>
<td>3</td>
</tr>
<tr>
<td>2015-2016</td>
<td>83</td>
<td>4</td>
</tr>
</tbody>
</table>

Average cost of model construction per analytical problem posed to Nexus 7, XDATA, Memex and QCR

Approved for Public Release, Distribution Unlimited.
Automated discovery of complex models with non-expert curation

- **TA1**: A library of selectable primitives
  - Create a “vocabulary” of modeling primitives
  - Make primitives automatically selectable

- **TA2**: Automatically compose complex models
  - Mine corpora of complex models to learn the “syntax” of primitive composition
  - Find optimal compositions
  - Predict additional data requirements

- **TA3**: Curation of models by non-experts
  - Decompose and formalize questions
  - Explain data and models to enable selection and editing
Program goals

**Phase 1:** Reproduce/improve models for existing problems without a data scientist

<table>
<thead>
<tr>
<th>Problem</th>
<th>Example</th>
<th>Pre-(D^3M) Effort (1st - Opt.)</th>
<th>(D^3M) Effort</th>
</tr>
</thead>
</table>
| 1. Simple social/bio-med problems  
Linear/categorical models, flat hierarchy, structured data | Smoking Factors, genetic species classification | 2-200 hrs (data science) | 0.5-2 hrs (SME) |
| 2. Multi-source prediction problems  
Multi-fused models, complex hierarchy, mixed data | Netflix Prize, Kaggle-PTSD, XDATA problems | 2000-15000 hrs (data science) | 1-10 hrs (SME) |

**Phase 2:** Synthesize models for unsolved problems, propose data augmentation

<table>
<thead>
<tr>
<th>Problem classes</th>
<th>Examples</th>
<th>(D^3M) Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Multi-modal predictive models with supplied data</td>
<td>Predict political instability or uprising, riot, conflict, donations to terrorist groups; predict causors/spread of disease; capabilities prediction from designs; optimize manufacturing process (OM)</td>
<td>5-40 hrs (SME)</td>
</tr>
<tr>
<td>2. Multi-modal predictive models with automated data collection</td>
<td>Multi-player games predict strategy/team formation, market/GDP forecasting, weather/ecology/environmental interaction, genetic factors for disease, predict mass shooting events</td>
<td>30-100 hrs (SME)</td>
</tr>
</tbody>
</table>
Some early results

- Problem: given image, identify objects (CIFAR-10)
D³M mining, evaluation and transition platform

- USG supplied infrastructure
  - Federated ML/data analysis corpus drawn from OSDC, Dataverse, Kaggle, UCI, etc.
  - Integration platform for TA1-3 performers
  - Performers deploy systems during integration events

External Data sources
- Kaggle
- OSDC
- Dataverse
- UCI
- Mloss
- Etc.

Independent contributions from empirical modeling/ML communities

Public-facing service:
1. Model service for social/bio scientists and transition partners
2. Human-in-the-loop eval (NIST)