Data Driven Discovery of Models (D³M)

Mr. Wade Shen/I2O
Automate (many of) the methods in data science to create empirical models of real, complex processes

• Enable non-expert users to make predictions from data without the need for data scientists
• Provide expert data scientists with automation that allows them to focus on the hard parts of the problem
D³M: Data-driven discovery of models

**Today: Manual**

- Model: representation of a real-world system
  - Examples: 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

---

Approved for Public Release, Distribution Unlimited
D³M: Accelerate scientific discovery and intelligence 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 intelligence 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 in four DARPA programs

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

Data-Driven Discovery

- **TA1: Discover and develop model primitives**
  - Create a “vocabulary” of modeling primitives
  - Discover key data characteristics that enable automated selection

- **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 questions
  - Explain data and models to enable selection and editing
Example: Prediction of scene location

Scene_location = ?(Data)

Automatically composed model (TA2)

\[ F_1 = NN(SVM(C), NB(R, G, B)) \]
\[ F_2 = SVM(SVM(R), SVM(G)) \]
\[ F_3 = SVM(C) \]
\[ ... \]

Automatically select primitives (TA1)

SVM, NB, GLM

Vocabulary of characterized primitives (TA1)

SVM, LDA, LSTM, NN, GLM, NB, CNN, ...
TA1: Automated selection of model primitives
Via data shape and correlation structure

Data

- Text
  - Zipf-like discrete tokens
  - Naïve Bayes
  - Scene-to-visual-word correlation

Vocab of characterized primitives

Discover

- Image patches
  - Zipf-like visual word dist.
  - Naïve Bayes
  - Scene-to-visual-word correlation

Real world

Characterization

- Data shape
  - Arg max \( p(f|c) \propto \sum_i \lambda_i f_i \)
  - Class-wise correlation with disjoint token subsets

Inputs

Outcome

Topic

Scene type

Approved for Public Release, Distribution Unlimited
TA2: Automatically compose complex models

Real models are complex: need composition
1. **# of models is exponential in complexity**
2. **Find optimum without enumerating all models**
3. **Augment data when underspecified**

Solution 1: Meta-learning to build models
1. Mine corpora of models to learn syntax of primitives:

   - From TA1
   - N possible primitives

<table>
<thead>
<tr>
<th>Problem</th>
<th>Model form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text geolocation</td>
<td>NN(SVM(•), NB(•))</td>
</tr>
<tr>
<td>H1N1 outbreak detection</td>
<td>NN(SVM(•), NB(•))</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Syntax</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P(NN(SVM(•)...)</td>
<td>HIGH</td>
</tr>
<tr>
<td>P(SVM(GMM(•)...)</td>
<td>LOW</td>
</tr>
<tr>
<td>P(GMM(GMM(•)...)</td>
<td>LOW</td>
</tr>
</tbody>
</table>

2. **Best first search, prune low probability paths**
3. **Use shape of missing data inputs to augment data**

**Benefits:** combines any primitives, mine human expertise

Solution 2: Joint model/parameter discovery

**Benefits:** no human expertise required, computationally hard to scale, limited to Bayesian primitives

Approved for Public Release, Distribution Unlimited
• **Formalize questions into models**
  - Decompose sub-problems
  - Natural ways to express questions
  - Elicit annotation of outcomes when missing
  - Automatically propose error metrics

• **Explain data and models to enable selection and editing**
  - Use data shape, correlation structure to show model decisions
  - Explain causal/correlation of variables
  - Editing of variables to constrain models
  - Elimination and re-ranking of models

• **Methods**

<table>
<thead>
<tr>
<th>1. Explain data salience via model introspection</th>
<th>Trace and weight data through complex models</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. Explain importance of primitives</td>
<td>Examine space of generated models to determine most “used” primitives</td>
</tr>
</tbody>
</table>
#### Evaluation metrics

**TA1: Measure efficiency of predicting primitives**

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predict primitives; experts compose</td>
<td>$\Delta$ error of predicted primitives vs. optimal</td>
</tr>
</tbody>
</table>

**TA2: Measure (re)discovery of optimal analyses/models**

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthesize models compare to experts</td>
<td>$\Delta$ error of $D^3M$ model vs. expert models</td>
</tr>
</tbody>
</table>

**TA3: Human curation of models**

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decompose questions, compare with experts</td>
<td>$P_d/P_{fa}$ of automated vs. expert (ROC)</td>
</tr>
<tr>
<td>Compare decisions made by experts with lay users</td>
<td>$\Delta$ error of analyst vs. data scientist in-the-loop</td>
</tr>
</tbody>
</table>

Program-level evaluation (annual integration and evaluation)

<table>
<thead>
<tr>
<th>Protocol</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>TA1-3 form team, work with non-experts to build model, compare with experts</td>
<td>$\Delta$ error of $D^3M$ model vs. expert model</td>
</tr>
</tbody>
</table>
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-D³M Effort (1ˢᵗ - Opt.)</th>
<th>D³M Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Simple social/bio-med problems</td>
<td>Smoking Factors, genetic species classification</td>
<td>2-200 hrs (data science)</td>
<td>0.5-2 hrs (SME)</td>
</tr>
<tr>
<td>Linear/categorical models, flat hierarchy, structured data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Multi-source prediction problems</td>
<td>Netflix Prize, Kaggle-PTSD, XDATA problems</td>
<td>2000-15000 hrs (data science)</td>
<td>1-10 hrs (SME)</td>
</tr>
<tr>
<td>Multi-fused models, complex hierarchy, mixed data</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Problem classes</th>
<th>Examples</th>
<th>D³M Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Multi-modal predictive models with supplied data</td>
<td><em>Predict political instability or uprising, riot, conflict, donations to terrorist groups; predict causors/spread of disease; capabilities prediction from designs; optimize manufacturing process (OM)</em></td>
<td>5-40 hrs (SME)</td>
</tr>
<tr>
<td>2. Multi-modal predictive models with automated data collection</td>
<td><em>Multi-player games predict strategy/team formation, market/GDP forecasting, weather/ecology/environmental interaction, genetic factors for disease, predict mass shooting events</em></td>
<td>30-100 hrs (SME)</td>
</tr>
</tbody>
</table>
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)
Schedule
Execution and phasing

GovTeam: continuous integration, define annual evals

- Model explanation
- Natural specification of questions
- Question formalization
- Question decomposition

- Model mining and synthesis
- Bayesian model discovery
- Deep encoder/decoder models
- Deep topology discovery

- Recommendation
- Unsupervised data discovery
- Unsupervised/semi-supervised machine learning
- Low-rank manifold discovery
- Large-scale factor analysis

TA1: Develop and discover modeling primitives

TA2: Automatically compose complex models

TA3: User <-> model interaction

Summer/winter program hack
Annual evaluation
Integrated toolkit / prelim system
Transition System

Year 1
Year 2
Year 3
Year 4

Approved for Public Release, Distribution Unlimited