DARPA Overview
&
Data Driven Discovery of Models (D^3M)

Wade Shen
I2O Program Manager

Brian Sandberg
I20 Technical SETA

DARPA Technical Offices

Director (Acting), Steven Walker
Deputy Director (Acting), Stefanie Tompkins

Biological Technologies

Director, Justin Sanchez
Deputy Director, Bradley Ringeisen

- Outpacing Infectious Disease
- Neurotechnology
- Gene Editing & Synthetic Biology

Defense Sciences

Director (Acting), William Regli
Deputy Director, Tyler McQuade

- Math, Modeling & Design
- Physical Systems
- Human-Machine Systems
- Social Systems

Information Innovation

Director, Brian Pierce
Deputy Director, John Everett

- Cyber: Win at Cyber
- Analytics: Understand the World
- Symbiosis: Partner with Machines

Microsystems Technology

Director, William Chappell
Deputy Director, Jay Lewis

- EM Spectrum
- Tactical Information Extraction
- Globalization

Strategic Technology

Director, Thomas Burns
Deputy Director, Daniel Patt

- Win In Any Environment via Composable Kill Chains
  - Sensing
  - Comms, Command, Control
  - Effects

Tactical Technology

Director, Bradford Tousley
Deputy Director, Fred Kennedy

- System Focus Areas
  - Ground, Maritime, Air, Space
- Crosscutting Themes
  - Agile Development
  - Cooperative Autonomy
  - Unmanned Systems
  - Power and Propulsion

May 15, 2017

Example Program

D^3M: Data-driven discovery of models

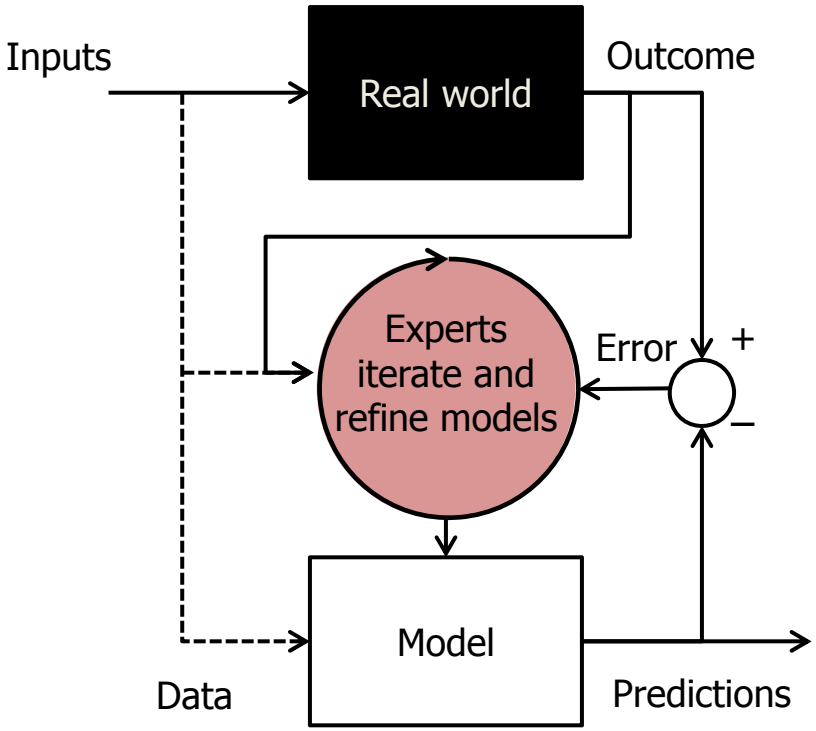
Goal:

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 allow them to focus on the hard parts of the problem

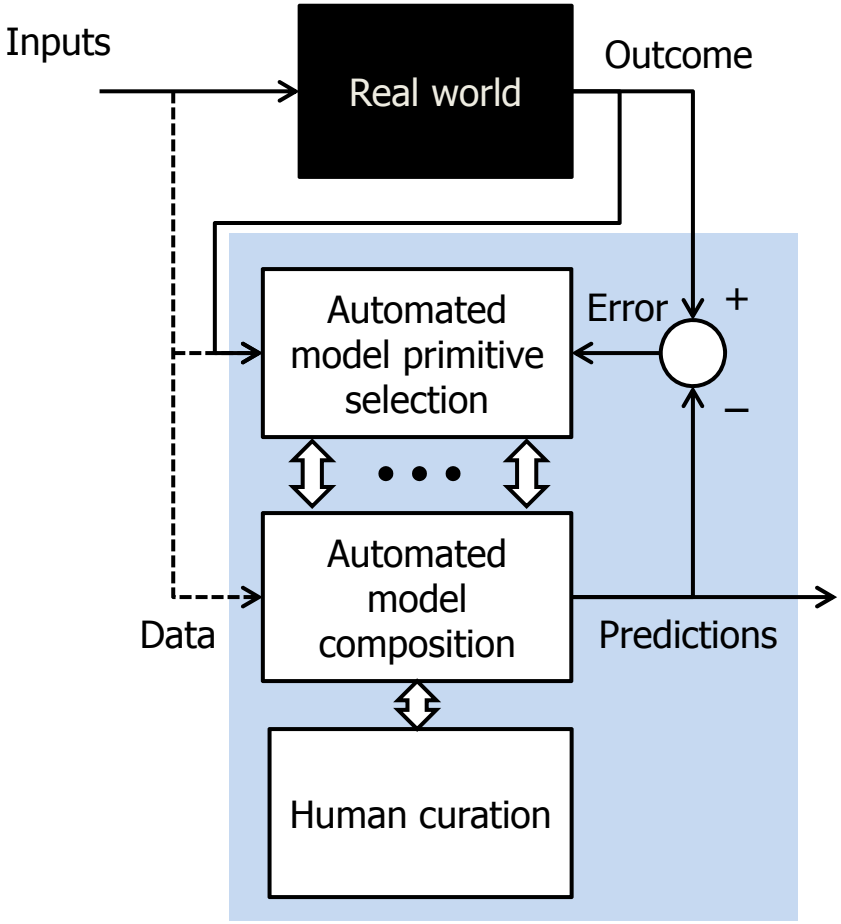
D³M: Data-driven discovery of models

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

D³M: Accelerate scientific discovery and intelligence analysis

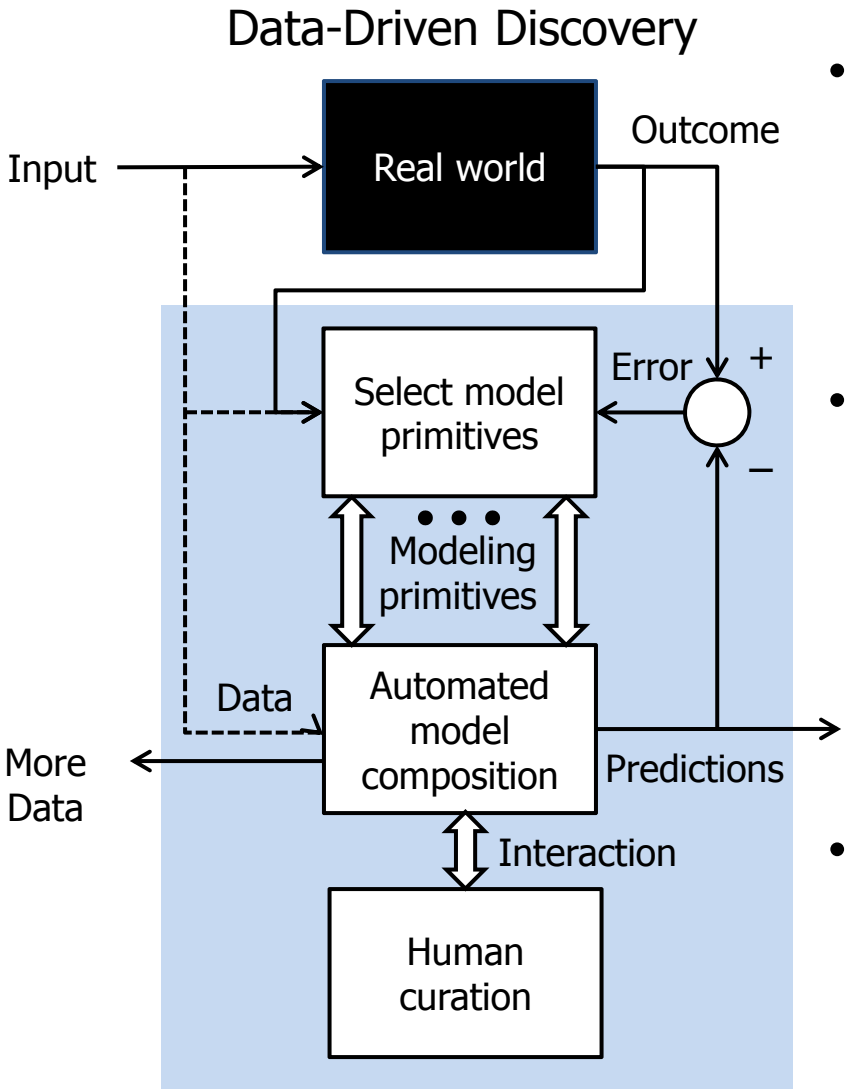
- Discover empirical models having complexity beyond current human comprehension
  - Humans can search only a tiny fraction of model space
  - Machines can search a much larger fraction much more rapidly

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

Year	Cost (Person-months)		Avg. time to solution (Months)	
	As-performed	with D³M (estimated)	As-performed	with D³M (estimated)
2009	432	4	24	0.5
2009-2011	126	5	18	0.25
2014-2015	102	3	6	0.25
2015-2016	83	4	5	1

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

Automated discovery of complex models with non-expert curation



- TA1: Discover and develop model primitives
  - Create a "vocabulary" of modeling primitives
  - Make primitives automatically discoverable
- 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

Evaluation metrics

TA1: Measure efficiency of predicting primitives

Protocol	Metric
Predict primitives; experts compose	Δ error of predicted primitives vs. optimal

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

Protocol	Metric
Synthesize models compare to experts	Δ error of D^3M model vs. expert models

TA3: Human curation of models

Protocol	Metric
Decompose questions, compare with experts	P_d/P_{fa} of automated vs. expert (ROC)
Compare decisions made by experts with lay users	Δ error of analyst vs. data scientist in-the-loop

Program-level evaluation (annual integration and evaluation)

Protocol	Metric
TA1-3 form team, work with non-experts to build model, compare with experts	Δ error of D^3M model vs. expert model

Program goals

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

Problem	Example	Pre-D^{3}M Effort (1st – Opt.)	D^{3}M Effort
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

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

www.darpa.mil