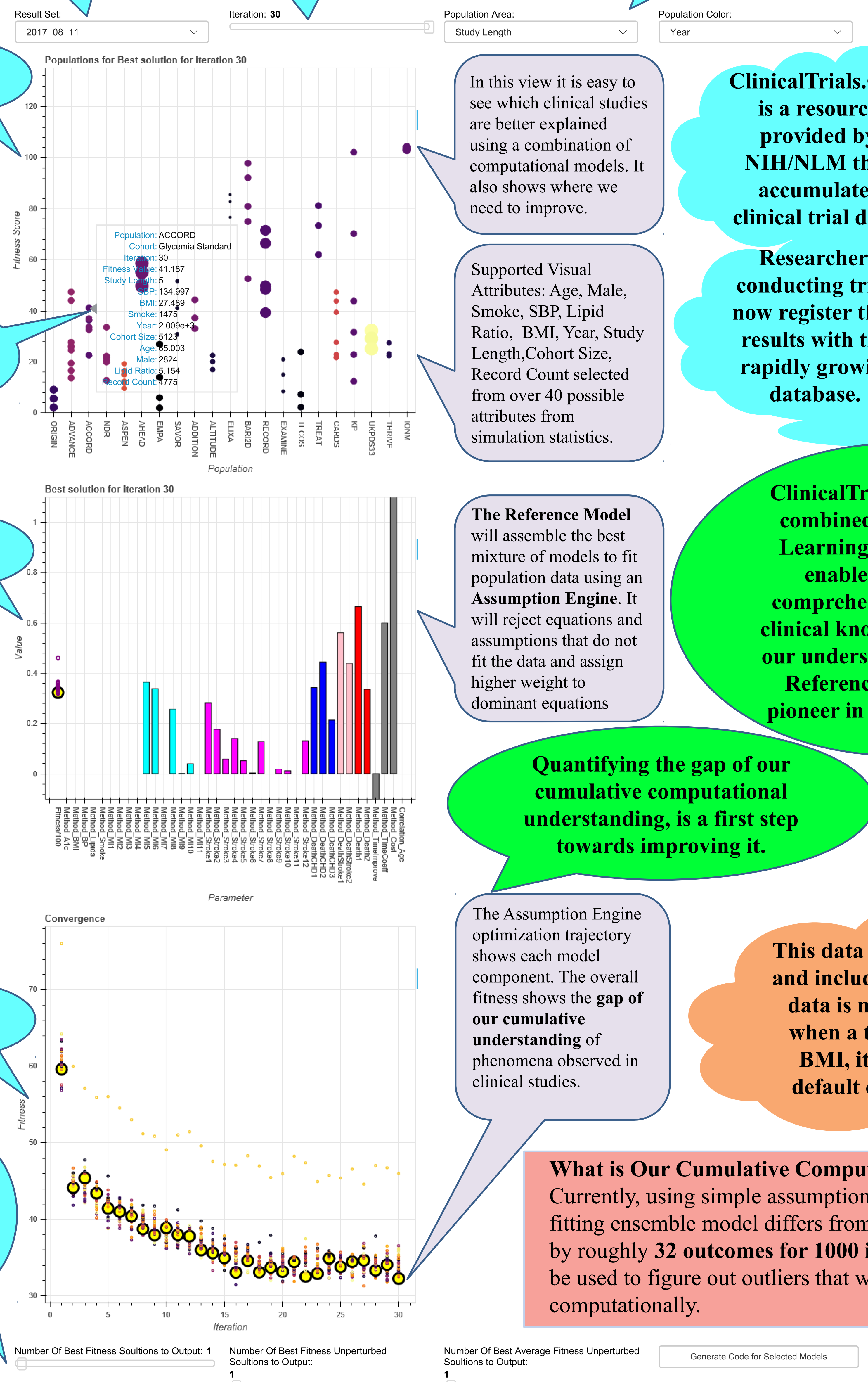
The Reference Model Visualizes Gaps in Computational Understanding of Clinical Trials

In a Nutshell: The Reference Model accumulates knowledge, including models and observed outcomes imported from ClinicalTrials.Gov and shows gaps in our understanding.

Jacob Barhak Ph.D.
http://sites.google.com/site/jacobbarhak/

2018 IMAG Futures Meeting
Moving Forward with the MSM Consortium
21-22 March 2018

New Interactive Interface in a Web Browser



User Can Explore Different Results

User Can Explore Convergence

User Controls Visual Attributes

Population Information

Hover Tool Exposes Detailed Data

Best Fitting Model Mixture

Optimization Convergence

The User Can Extract The Best Models for Additional Simulations

In this view it is easy to see which clinical studies are better explained using a combination of computational models. It also shows where we need to improve.

Supported Visual Attributes: Age, Male, Smoke, SBP, Lipid Ratio, BMI, Year, Study Length,Cohort Size, Record Count selected from over 40 possible attributes from simulation statistics.

The Reference Model will assemble the best mixture of models to fit population data using an **Assumption Engine**. It will reject equations and assumptions that do not fit the data and assign higher weight to dominant equations

The Assumption Engine optimization trajectory shows each model component. The overall fitness shows the **gap of our cumulative understanding** of phenomena observed in clinical studies.

ClinicalTrials.Gov is a resource provided by NIH/NLM that accumulates clinical trial data.

Researchers conducting trials now register their results with this rapidly growing database.

The Reference Model Key Points

- Ensemble model
- Accumulates knowledge from:
  - Existing models
  - Observed outcomes
- Focuses on summary data
  - Avoids individual data restrictions
  - Larger merged population base
- Flexible Import from ClinicalTrials.Gov
- Applicable for other disease processes
- Traceable and reproducible
- Can map our understanding gap
- Currently focuses on diabetic populations

There are more than quarter million trials currently registered with about 10% with results.

ClinicalTrials.Gov growth combined with Machine Learning can eventually enable machines to comprehend accumulated clinical knowledge and close our understanding gap. The Reference Model is one pioneer in this exploration.

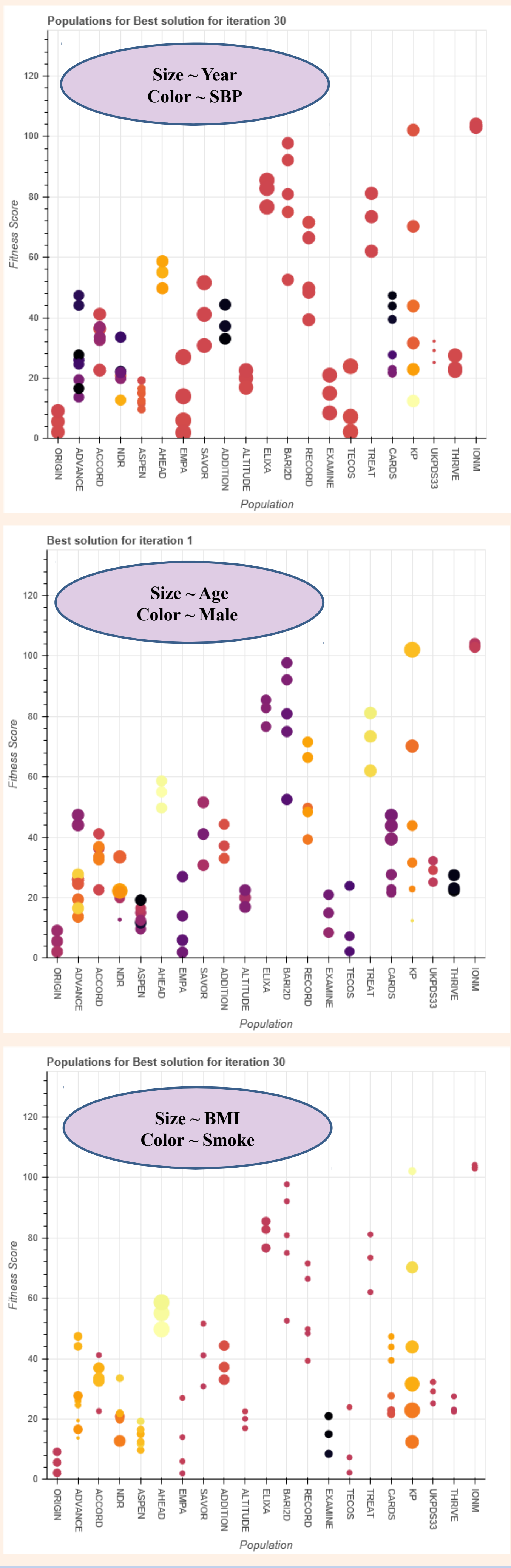
Quantifying the gap of our cumulative computational understanding, is a first step towards improving it.

Human interpretation is still required today and machines will need to learn from humans.

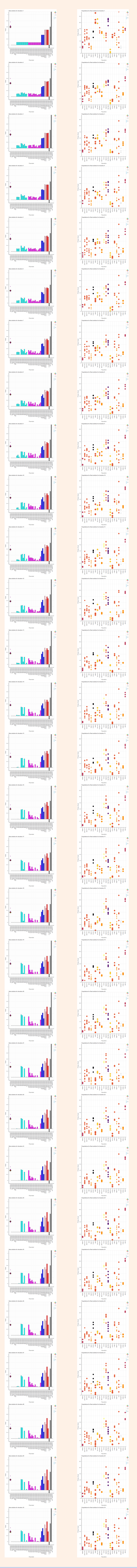
This data is based on simulation and includes assumptions where data is missing. For Example when a trial does not include BMI, it is extracted from a default diabetic population.

What is Our Cumulative Computational Understanding Gap?
Currently, using simple assumptions and data collected, the best fitting ensemble model differs from observed outcomes on average by roughly **32 outcomes for 1000 individuals**. This threshold may be used to figure out outliers that we still cannot explain well enough computationally.

Exploring Population Attributes



Convergence



Abstract:

There Reference Model accumulates knowledge from multiple publicly available sources in two categories. 1) It assembles the best fitting ensemble model from multiple published disease models that attempt to explain cardiovascular disease and mortality. 2) It accumulates observed information from multiple clinical trials for validation. It uses High Performance Computing to optimize the best model mixture and generate synthetic populations to match clinical trial reports.

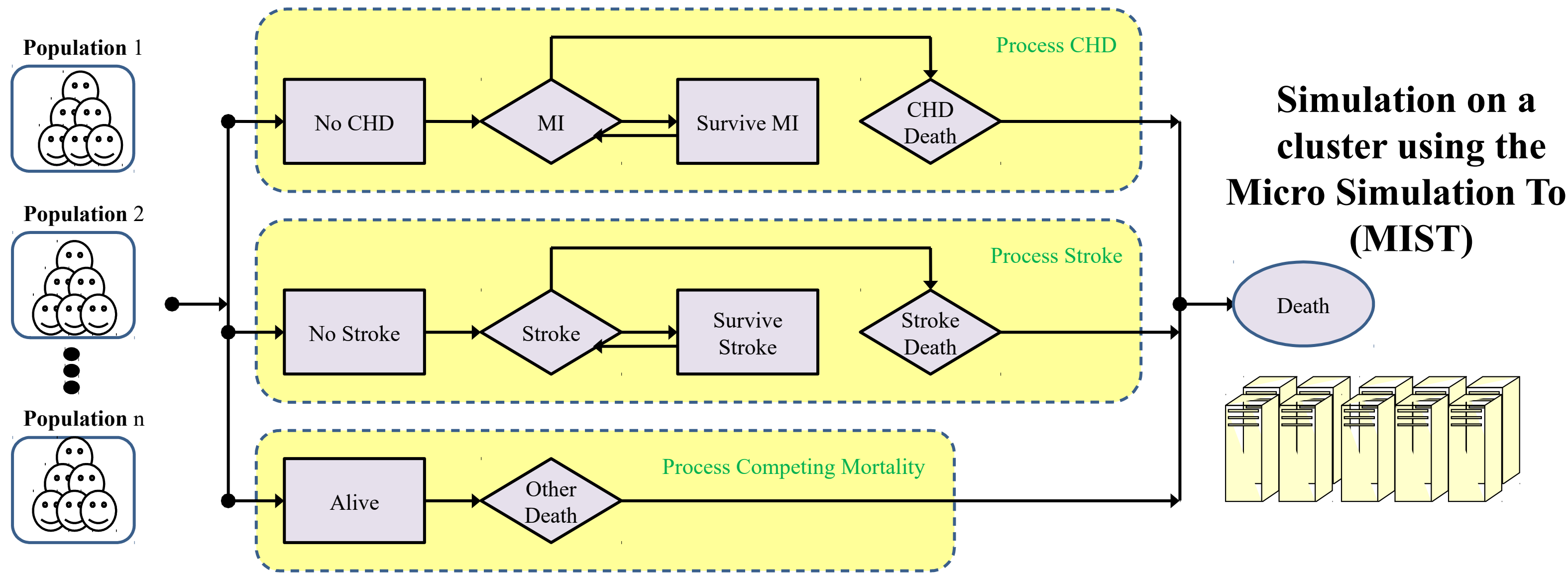
Since the model aggregates two types of knowledge: 1) models and 2) observation data collected from clinical studies, it can show gaps in our cumulative understanding and our ability to explain phenomenon observed. The Reference Model has been accumulating such data globally since 2012 and connected to ClinicalTrials.Gov in 2017 which dramatically increased its access to data with greater future potential.

With the data already accumulated, It is now possible to visualize gaps in our understanding of outcomes reported in 22 diabetic clinical trials with 91 cohorts by showing the fitness of the best model mixture to those clinical trials. The Reference Model showed similar visualization in the past in this forum, then using a color coded fitness Matrix. The advantages in this work, compared to the past, are: **1) The visualization is interactive through a web browser allowing exploration of data. 2) The Reference Model now mixes models, allowing improved fitness and accumulation of assumptions. 3) The size of the current validation effort has passed beyond the largest known validation exercise.** Those changes make it worthwhile presenting the new visualization capabilities and compare those to past work to show our current understanding gap.

The ability to aggregate the data, quantify the gap, and visualize it will aid development of better models to close the computational understanding gap.

The Reference Model

Now allowed as a U.S. Patent



Simulation on a cluster using the Micro Simulation Tool (MIST)

This work builds upon a decade of development with key publications in the following list:

[1] J. Barhak, The Reference Model for Disease Progression uses MIST to find data fitness. PyData Silicon Valley 2014 held at Facebook Headquarters:
Presentation: http://sites.google.com/site/jacobbarhak/home/PyData_SV_2014_Upload_2014_05_02.pptx
Video: https://www.youtube.com/watch?v=vyvxiljc5vA

[2] J. Barhak, A. Garrett, Population Generation from Statistics Using Genetic Algorithms with MIST + INSPYRED. MODSIM World 2014, April 15 - 17, Hampton Roads Convention Center in Hampton, VA.
Paper: http://sites.google.com/site/jacobbarhak/home/MODSIM2014_MIST_INSPYRED_Paper_Submit_2014_03_10.pdf
Presentation: http://sites.google.com/site/jacobbarhak/home/MODSIM_World_2014_Submit_2014_04_11.pptx

[3] J. Barhak, Object Oriented Population Generation, MODSIM world 2015. 31 Mar – 2 Apr, Virginia Beach Convention Center, Virginia Beach, VA.
Paper: http://modsimworld.org/papers/2015/Object_Oriented_Population_Generation.pdf
Presentation: http://sites.google.com/site/jacobbarhak/home/MODSIM2015_Submit_Jacob_Barhak_2015_03_29.pptx

[4] J. Barhak, The Reference Model for Disease Progression Combines Disease Models. I/IITSEC 2016 28 Nov – 2 Dec Orlando Florida.
Paper: http://www.iitsecdocs.com/volumes/2016
Presentation: http://sites.google.com/site/jacobbarhak/home/IITSEC2016_Upload_2016_11_05.pptx

[5] J. Barhak, The Reference Model Models ClinicalTrials.Gov. SummerSim 2017 July 9-12, Bellevue, WA.
Paper: https://doi.org/10.22360/SummerSim.2017.SCSC.022 or http://dl.acm.org/citation.cfm?id=3140087
Presentation: http://sites.google.com/site/jacobbarhak/home/SummerSim2017_Upload_2017_07_09.pptx

[6] J. Barhak, The Reference Model: A Decade of Healthcare Predictive Analytics with Python, PyTexas 2017, Nov 18-19, 2017, Galvanize, Austin TX.
Presentation: http://sites.google.com/site/jacobbarhak/home/PyTexas2017_Upload_2017_11_18.pptx
Video: https://youtu.be/Pj_N4izLmsI

Technology
The Python programming language is the main technological enabler behind the model. The new visualization through a web browser is possible using the bokeh library that allows plotting and user interaction with the data. The Reference Model itself runs simulations using the Micro Simulation Tool (MIST) that runs simulations in parallel on multiple machines on a Cluster. It is possible to run those simulations on the Amazon Elastic Compute Cloud. The free Anaconda Python distribution is used to handle all the packages needed and some versions of MIST are available for download under General Public License through: https://github.com/Jacob-Barhak/MIST

Reproducibility:
The plots in the poster were created using the script ExploreOptimizationResults_2018_03_14.py on Windows 10 environment with bokeh 0.12.10 on python 2.7.14 64 bit based on simulation results stored in: MIST_RefModel_2017_11_08_OPTIMIZE.zip

Acknowledgments:
Thanks to Deanna J. M. Isaman who first introduced me to the the idea of accumulating knowledge from clinical trial summary data. Many thanks to the Nick Ide who developed ClinicalTrials.Gov for the assistance in figuring out how to use the web site. This paper used the MIcro Simulation Tool (MIST) that is based on IEST. The IEST GPL disease modeling framework was initially supported by the Biostatistics and Economic Modeling Core of the MDRTC (P60DK020572) and by the Methods and Measurement Core of the MCDTR (P30DK092926), both funded by the National Institute of Diabetes and Digestive and Kidney Diseases. The modeling framework was initially defined as GPL and was funded by Chronic Disease Modeling for Clinical Research Innovations grant (R21DK075077) from the same institute.
The Reference Model for Disease Progression, the import module, and the MIST framework were developed without financial support.
The author has no affiliation to ClinicalTrials.Gov which is an NIH project and just used here for data extraction.

Download This Poster