GNS HEALTHCARE
Driving Intelligent Interventions

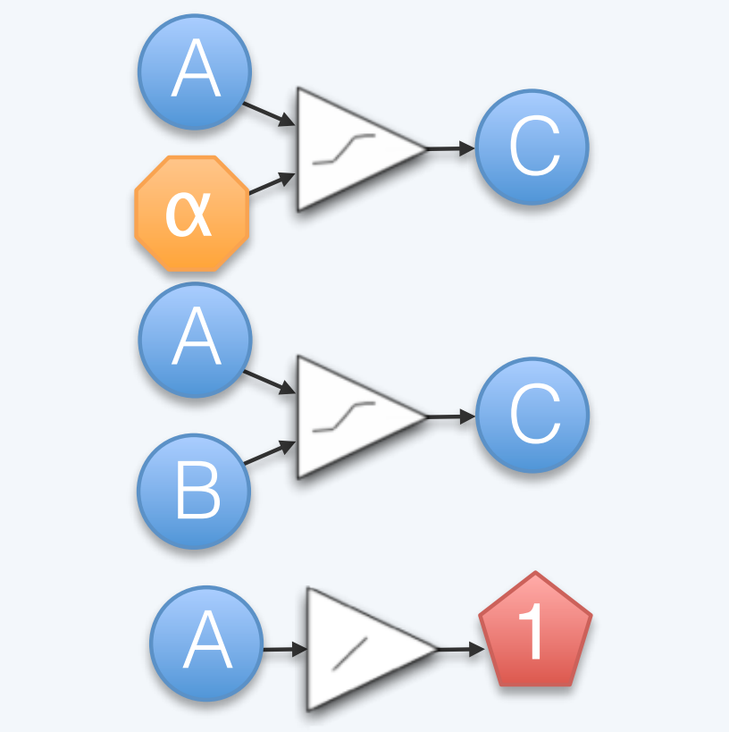
196 Broadway
Cambridge, MA 02139
617-374-2300
www.gnshealthcare.com

Reverse Engineering, Forward Simulation (REFS™)

Machine Learning Causal Inference Platform

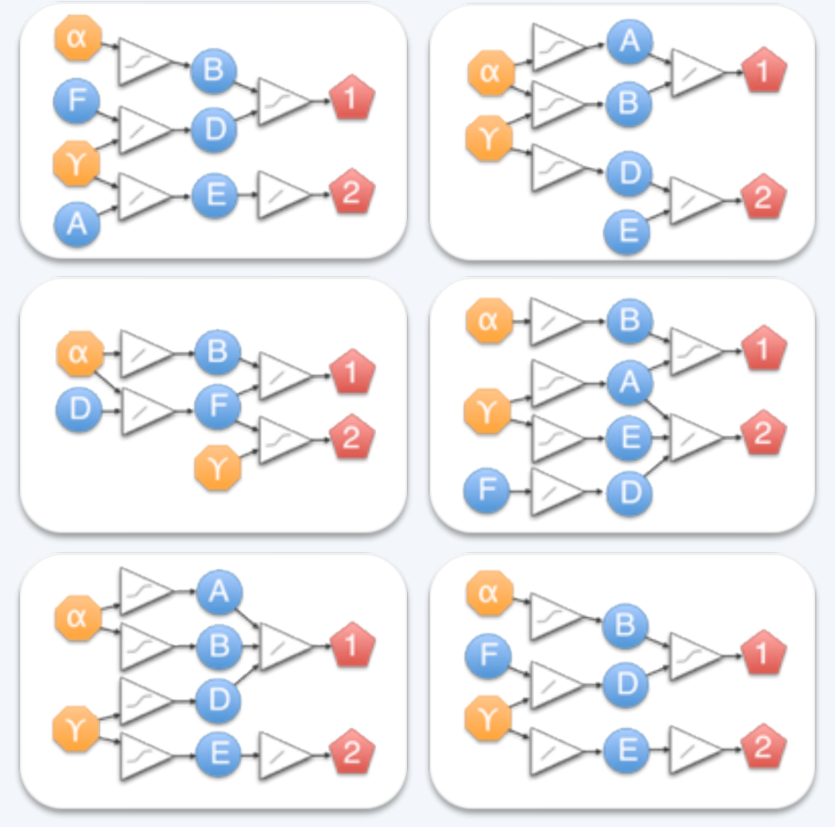
- The REFS™ platform enables unique insights by first learning or Reverse Engineering (RE) an ensemble of models directly from data, without *a priori* hypotheses. Simulations of the learned models can then be employed to make patient-specific predictions and identify the key predictors of health outcomes. REFS™ uses Bayesian network inference to learn models directly from data and subsequently builds an ensemble of models.
- An ensemble of models is built rather than trying to learn a single or 'best' model because the data frame is necessarily under-determined, i.e., the dimension of the space of possible combinations of variables is much higher than the number of observations. A typical ensemble consists of hundreds to thousands of models.
- Interaction forms can consist of any number of variables, and a wide range are available within the REFS™ platform including linear, log-linear, logistic, multinomial interactions, Poisson, Gaussian, and survival models. Interaction forms are available to handle both discrete and continuous variables, as well as combinations of discrete and continuous variables and countless interactions between them. Tens of billions to trillions of models are proposed and scored for each model that eventually is accepted into the ensemble.

Enumeration



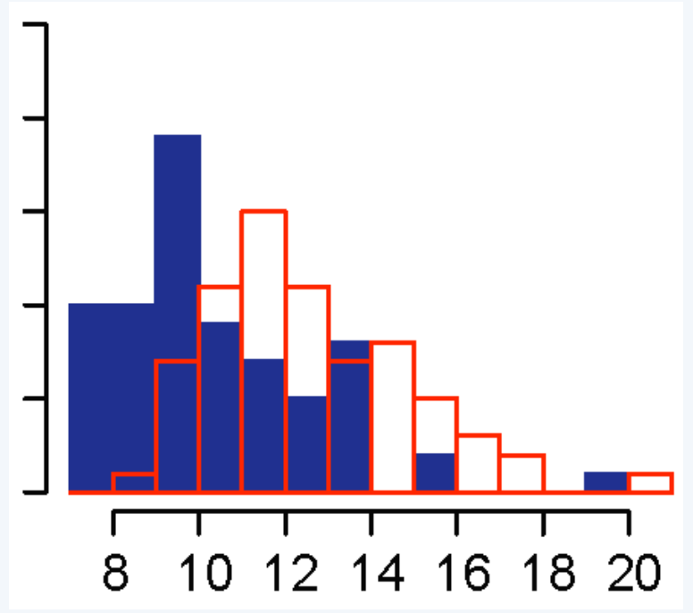
Individual network fragments are scored based on the full distribution of parameter values

Optimization

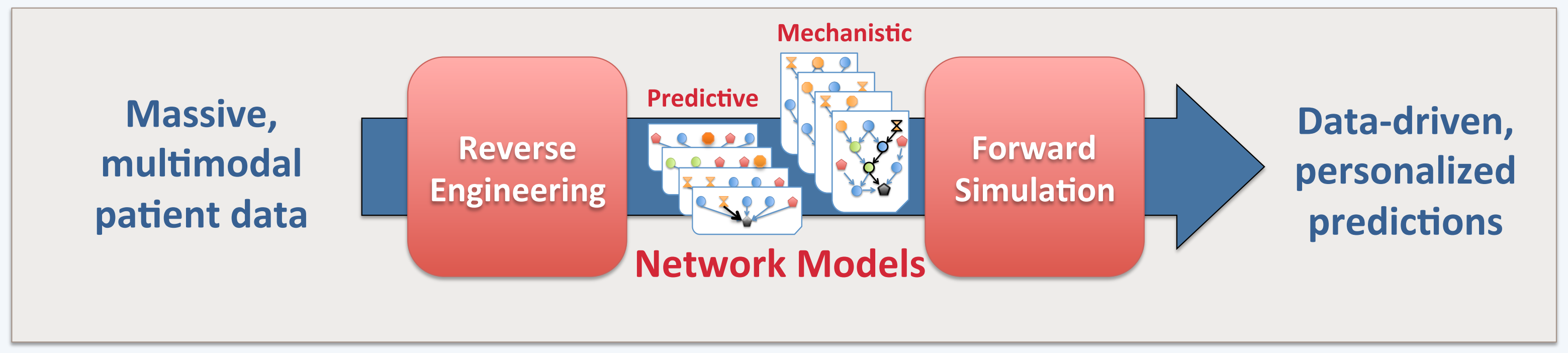


A globally optimal ensemble of networks is found by the Metropolis Monte Carlo algorithm

Simulation

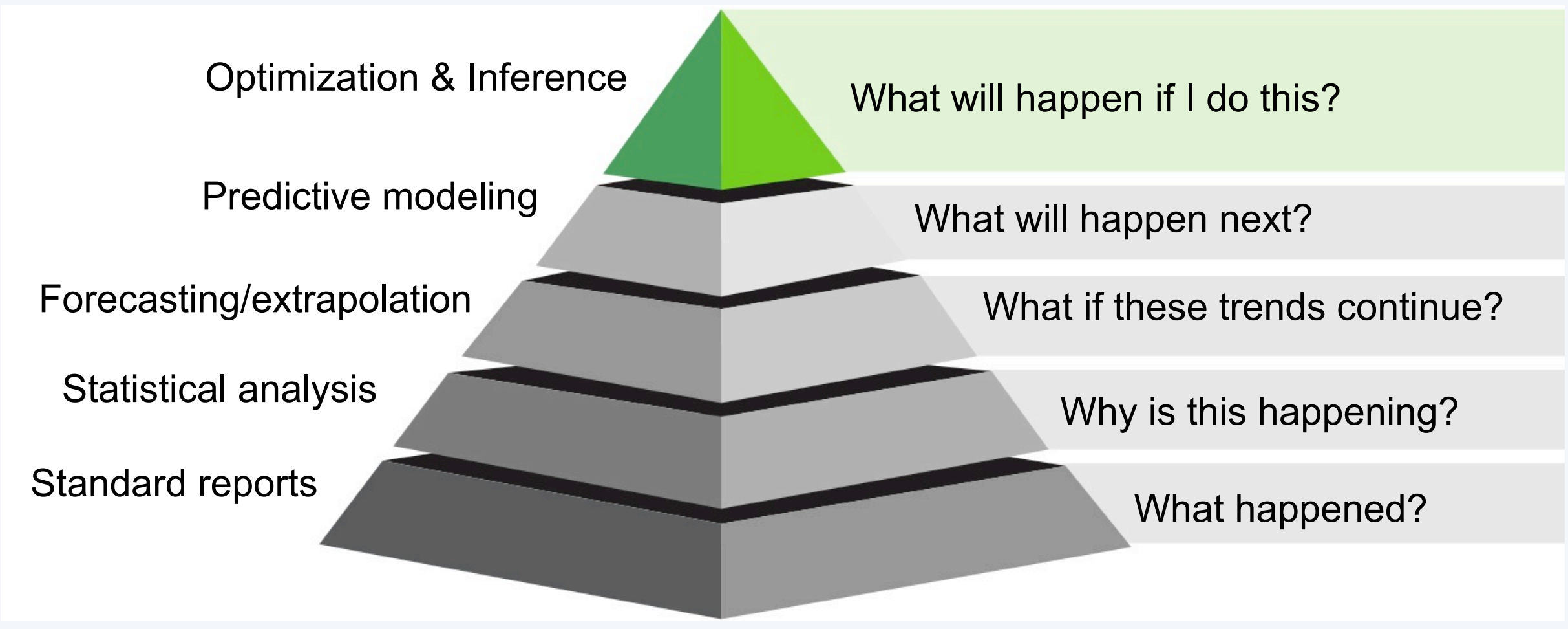


Simulations are run across the ensemble of networks to discover the causal drivers of response



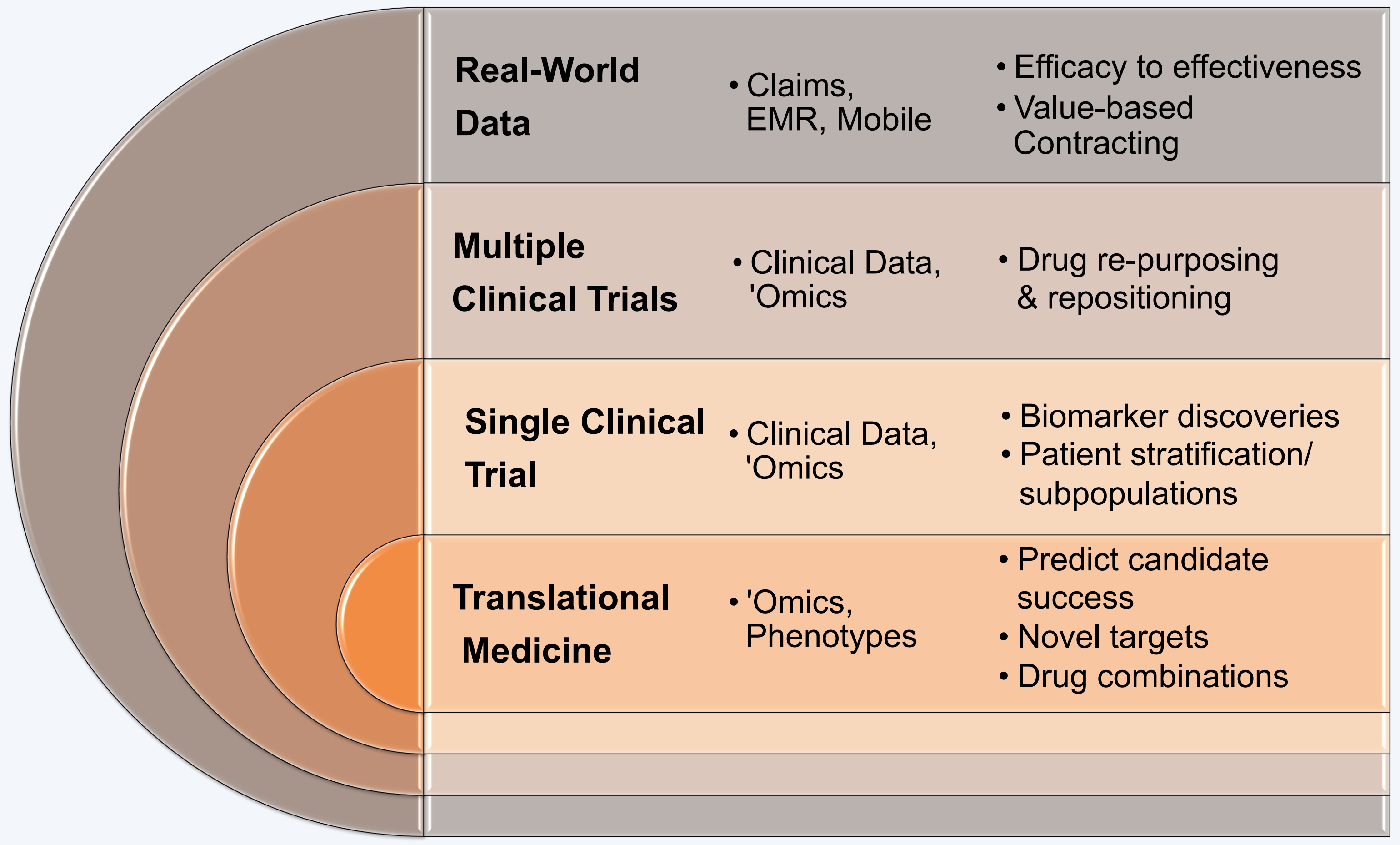
Advantages of REFS™

Explores "What If?" Possibilities



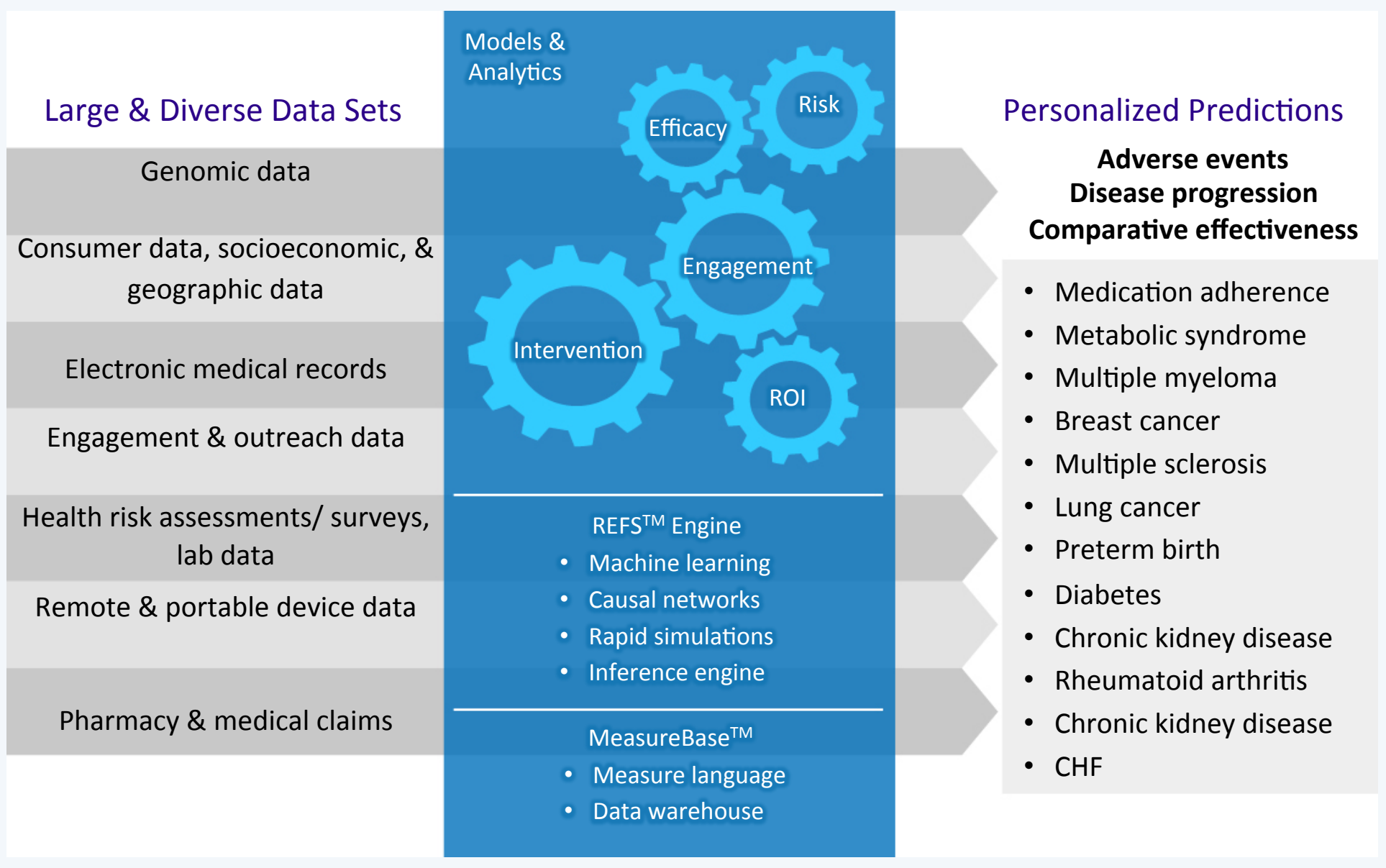
- **Using all the data** – The REFS™ approach is hypothesis-free, so there is no biasing based on literature mining to guess what is important. Therefore, all of the data is considered and REFS™ determines what is important.
- **Novel Discovery** – Potential novel discoveries can be made because REFS™ completes a hypothesis-free search of the data. This is different then a knowledge-based modeling approach (where publications are used to identify important variables before analysis begins), which potentially creates circular research in only discovering what you already know to be true.
- **Scale** – The REFS™ platform can easily explore relationships between a 100,000 or more variables. This scale allows us to create the best set of unified hypothesizes that are not biased by previous research.
- **Quantifying Uncertainty** – For every type of scientific question we attempt to answer, we don't just build one single model, we build 100s. Each one of these 100s of models may arrive at a slightly different answer, and we utilize this "ensemble" of models to understand and quantify the uncertainty around predictions. Therefore REFS™ computes both the specific predictions and the likelihood that they are correct.
- **Generating Personal Predictions** – The REFS™ platform, at its core, can generate a set of predictions for a new patient on an individual level. The REFS™ platform can identify whether the prediction is good, bad, or perhaps even more important, unknown. Therefore, researchers and clinicians have a complete picture on the prediction.
- **Testing Complexity** - REFS™ comprehensively explores the standard main effects of potential predictor variables in addition to the interactions amongst the predictors. These interactions are explored deeply even if the main effects are not present. A more traditional analysis would likely miss effects and not be able to perform hypothesis tests on interactions, causing model accuracy to suffer.
- **Subpopulation Identification** – Each model in REFS™ independently explores the hypothesis space, so a subset of the models may arrive at a specific answer that other models did not learn. This ensemble structure can be exploited using personalized predictions to identify subpopulations in the data and the variables that predict those subpopulations.

Applications Across Healthcare from Discovery to Value-Based Solutions



Real-World Data	• Claims, EMR, Mobile	• Efficacy to effectiveness • Value-based Contracting
Multiple Clinical Trials	• Clinical Data, 'Omics	• Drug re-purposing & repositioning
Single Clinical Trial	• Clinical Data, 'Omics	• Biomarker discoveries • Patient stratification/ subpopulations
Translational Medicine	• 'Omics, Phenotypes	• Predict candidate success • Novel targets • Drug combinations

REFS™ Leverages Multi-Modal Patient Data to Build Disease Models



Large & Diverse Data Sets
- Genomic data
- Consumer data, socioeconomic, & geographic data
- Electronic medical records
- Engagement & outreach data
- Health risk assessments/ surveys, lab data
- Remote & portable device data
- Pharmacy & medical claims

Models & Analytics

REFS™ Engine
- Machine learning
- Causal networks
- Rapid simulations
- Inference engine

MeasureBase™
- Measure language
- Data warehouse

Personalized Predictions

Adverse events
Disease progression
Comparative effectiveness

- Medication adherence
- Metabolic syndrome
- Multiple myeloma
- Breast cancer
- Multiple sclerosis
- Lung cancer
- Preterm birth
- Diabetes
- Chronic kidney disease
- Rheumatoid arthritis
- Chronic kidney disease
- CHF

	Data	Study Outcome
CHDI FOUNDATION	**HD datasets: Lymphocyte gene expression, pre-clinical mouse model (molecular and phenotype), human (clinical and molecular)**	**Connected an established Huntington's disease biomarker to novel genes and recapitulated known biology through a transcriptional network model**
MMRF MULTIPLE MYELOMA Research Foundation This is how cures are found.	**1,000+ patient cohort, including clinical, genomic. Data refreshed every 3 months**	**Built causal models to identify known intervention targets and identified novel targets (molecules and DNA regions) for further research**
INOVA Join the future of health.	**3,000 patient cohort (mother, father, baby), including EMR, genomic. Additional patient ands and timepoints incorporated to validate**	**Identified a unique maternal molecular profile associated with preterm birth outcomes and accurately stratified early preterm births**
G3 GLOBAL | GENOMICS | GROUP	**7500 patients, including Clinical, Blood Based Biomarkers, Coronary CT, Mass-spec, mRNAseq, miRNAseq, WGS, DNA methylation, lipidomic, proteomic**	**Recapitulated molecular signaling driving coronary artery disease and identified novel drivers of CAD endpoints**

Additional Datasets: TRUVEN HEALTH ANALYTICS · DANA-FARBER CANCER INSTITUTE · UCSF · BRIGHAM AND WOMEN'S HOSPITAL · OPTUM · HUMEDICA An Optum Company · ims

Case Study:

Huntington's Disease causal modeling using network ensemble simulations of gene expression data

Identification of gene pairs with causal relationship by simulation

Simulation experiments using;
5-fold perturbation
Sex: Male and female
CAG: Low, medium, and high
Perturbation: knockdown and overexpression

↓

12 result sets
(sex * CAG * perturbation)

↓

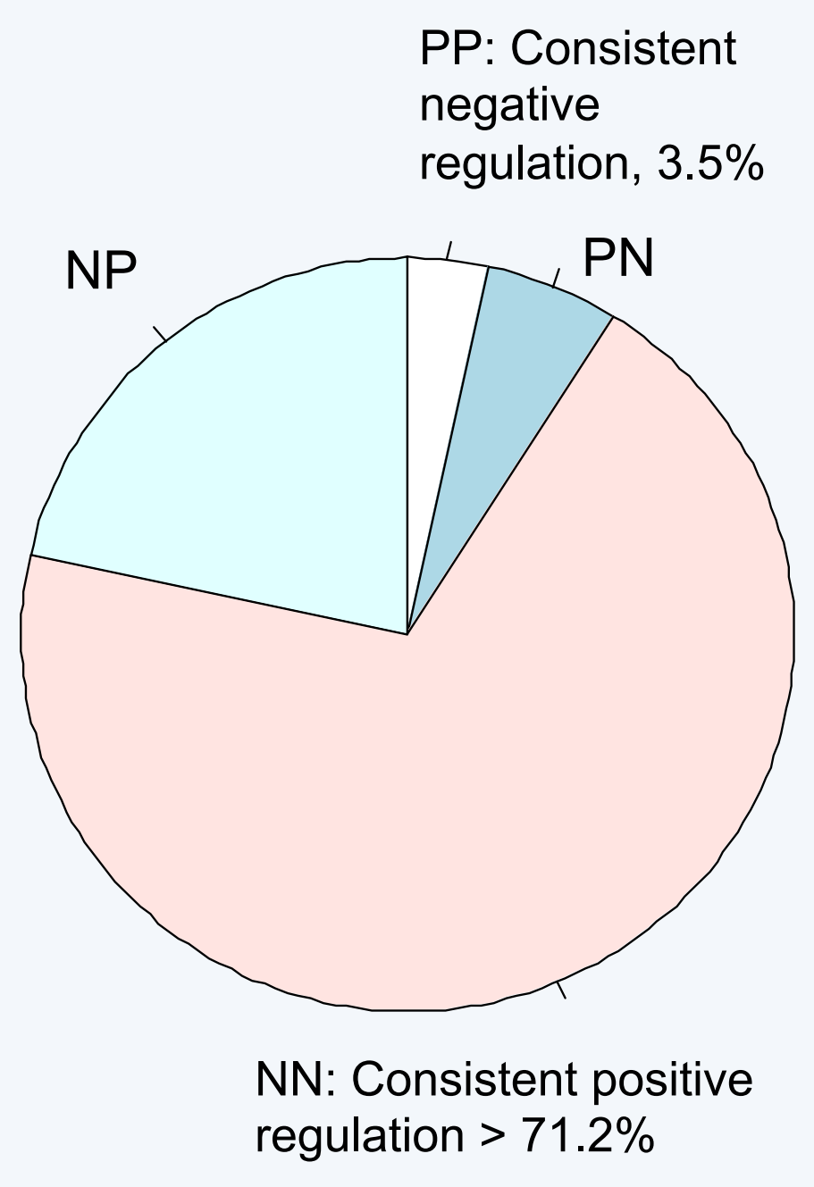
Meta-analysis to combine sex and CAG
12 result files into 2 summary (KD and OE)

Summary of meta-analysis: 5-fold GNS model results

	KD	OE
Total # of pairs	97824	82698
Unique upstream genes	2717	2698
Unique downstream genes	3462	3459
Without gene symbol (upstream)	312	320
Without gene symbol (downstream)	811	851

Concordance between GNS model and LINCS data

Without considering p-value, ~75% of gene pairs consistent directions



> Microarray gene expression data were used to construct GNS Bayesian causal network

> Causal models based on 3537 probes were constructed

> Significant causal relationships were identified by 5-fold knock-down perturbation simulation

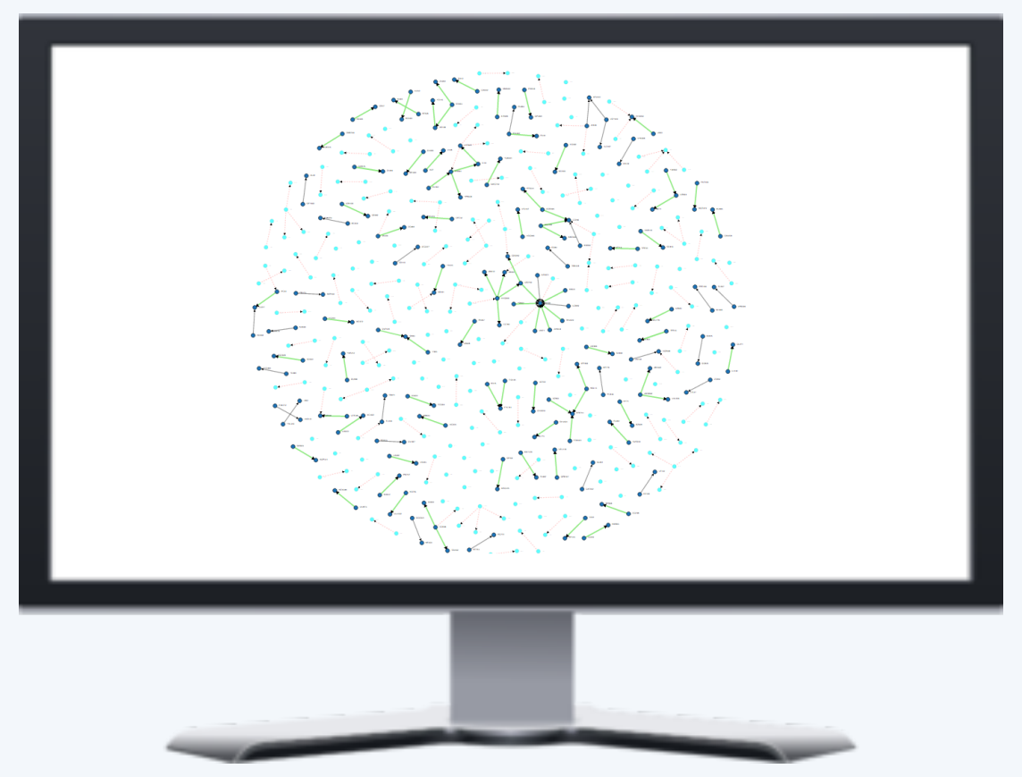
> Those gene pairs were compared to LINCS reliable experiment data

> Between GNS models and LINCS data, approximately 75% concordance rate was observed

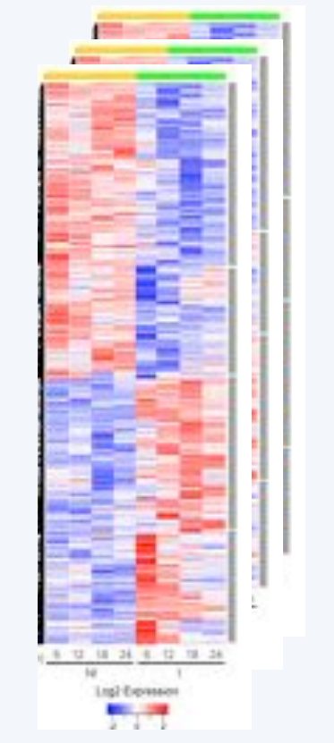
> Effects of CAG on those causal relationship is under investigation

Funded by CHDI Foundation, Inc.

Disease Models as a Platform for Translational Medicine and Clinical Trial Development



+

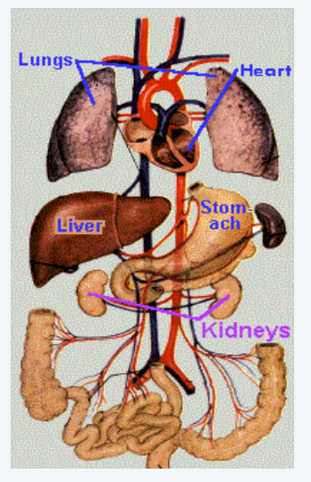


COMPOUND Tx Profile

Biomarkers and Patient Selection

Rational Combinations

New Treatments



New Indications "Repurposing"

Iya Khalil, PhD
Co-Founder and Executive Vice President
iya@gnshealthcare.com

Stacey Wasserman, MBA
Vice President, Business Development, Pharmaceutical Market
stacey@gnshealthcare.com