Business or Organizational Unit
Franchise or Department

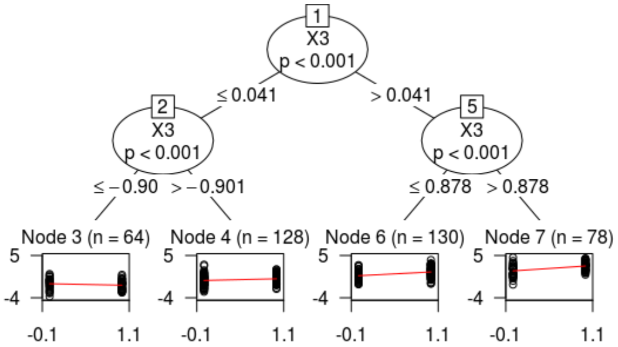
Subgroup Benchmarking Framework

Sophie Sun, AEA/AMDS
Björn Bornkamp, SMC/AMDS
Yao Chen , SMC/AMDS
Jiarui Lu , SMC/AMDS
Kostas Sechidis , AEA/AMDS

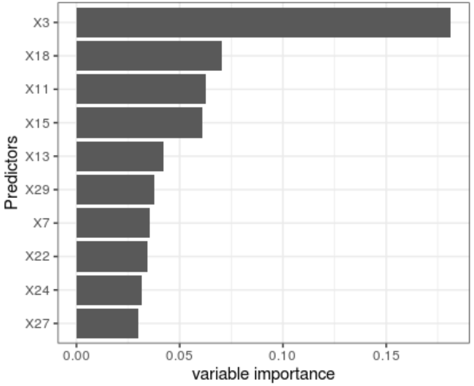
NOVARTIS | Reimagining Medicine

Motivation

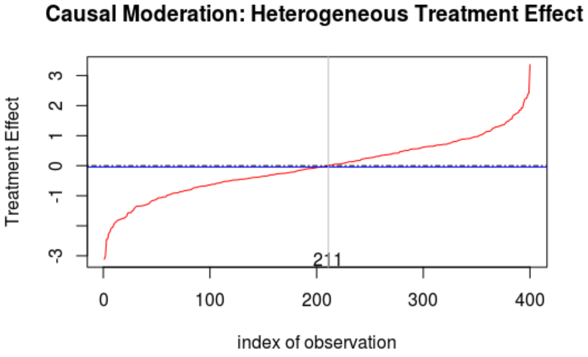
MOB



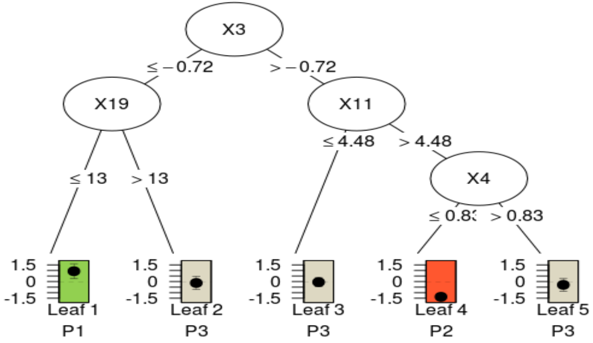
Causal Forest



FindIt



Quint



Challenge

- There has been review work to compare subgroup methods (Lipkovich et al. 2017, Huber et al. 2019, Zhang et al. 2018, Loh et al. 2019)
- Data: simulation data is not a good representation of the real clinical data
  - Mostly use variables generated from parametric distribution with simple dependent/ independent covariates distributions (Lipkovich et al. 2017, Huber et al. 2019, Zhang et al. 2018)
  - Large subgroup effects (Loh et al.2019)
- Methodology:
  - Only consider certain type of response (Loh et al. 2019, Huber et al. 2019)
  - Might tend to consider cases and metrics favoring the proposed methods
- Objective: Build objective & realistic benchmarking framework and considers metrics of practical importance, provide guidelines for subgroup analysis

Data generation: Prognostic and predictive structure

Generate data from: $f(X) = f_{prog}(X) + Trt * (\beta_0 + \beta_1 * f_{pred}(X))$

- Benchtm (go/benchtm) provides two covariate distributions:
  1. Synthetic data generated from a real clinical trials
     - Generated from real clinical trial data using “synthpop”
     - To ensure anonymization
       - Continuous covariates scaled to (0, 1)
       - Covariates relabeled to $X_1, X_2, \dots, X_p$
  2. Generated from parametric distribution with dependence/ independent covariates distributions
- User can provide the structure of $f_{prog}(X)$ and $f_{pred}(X)$ to account for different type of problems (e.g. linear/non-linear, step)

Data generation: different sets of parameters

Generate data from: $f(X) = f_{prog}(X) + Trt * (\beta_0 + \beta_1 * f_{pred}(X))$

Parameter	Setting
Sample size	100, 500, 1000
Number of predictors	30,100
Overall Effect size (determined by power)	Large (power = 0.9) Medium (power = 0.5) Small (power = 0.05)
Standard deviation of treatment effect	Small, medium, large (multiple)
Treatment effect structure	Step, linear
Number of variables defining subgroup	0 (no subgroup), 1, 2
Clinical endpoints	**Continuous, binary, count, survival**

NOVARTIS | Reimagining Medicine

List of methods for simulation

Non-ensembling methods

- Tree-based methods: (suggest subgroup)
  - ❑ Virtual Twins (Foster, Taylor, and Ruberg 2011)
  - ❑ SIDES (Lipkovich et al. 2011)
  - ❑ GUIDE (Loh and Zhou 2020)
  - ❑ Interaction Tree (Su et al. 2008)
  - ❑ MOB (Zeileis and Hothorn 2015)
  - ❑ QUINT (Dusseldorp and Van Mechelen 2014)
  - ❑ STIMA (Dusseldorp, Conversano, and Van Os 2010)
- Linear- Regression-based methods: (estimate interaction effect)
  - ❑ Lasso & Ridge, Glmnet (Hastie and Qian 2014)
  - ❑ FindIT (Imai, Ratkovic, and others 2013)
  - ❑ STIMA (Dusseldorp, Conversano, and Van Os 2010)
  - ❑ OWL (Fu, Zhou and Faries 2016, Yu et al. 2015)

Ensembling methods: (provide variable importance)

- ❑ Casual forest (Athey et al. 2019)
- ❑ GUIDE (Loh and Zhou 2020)
- ❑ MOBFOREST (Garge et al. 2019)
- ❑ Virtual Twins (Foster, Taylor, and Ruberg 2011)
- ❑ BART (Chipman et al. 2010)
- ❑ TSDT (Battioui et al. 2018)
- ❑ subtee (Bornkamp et al. 2017)

Performance Metrics

Ability to reliably determine subgroups with higher or lower (relative to overall) treatment effect (reproducible in new studies)

- Test for existence of differential treatment effects
  - Right/wrong decision
- Variable selection bias based on variable importance rankings (for differential treatment effect)
  - “True treatment effect modifying variates” are most important variables
  - Prognostic variables falsely identified among most important variables
- Estimation of individual treatment effects
  - Bias & MSE overall (and maybe in top 50%, top 20%, top 10% of predicted effects)

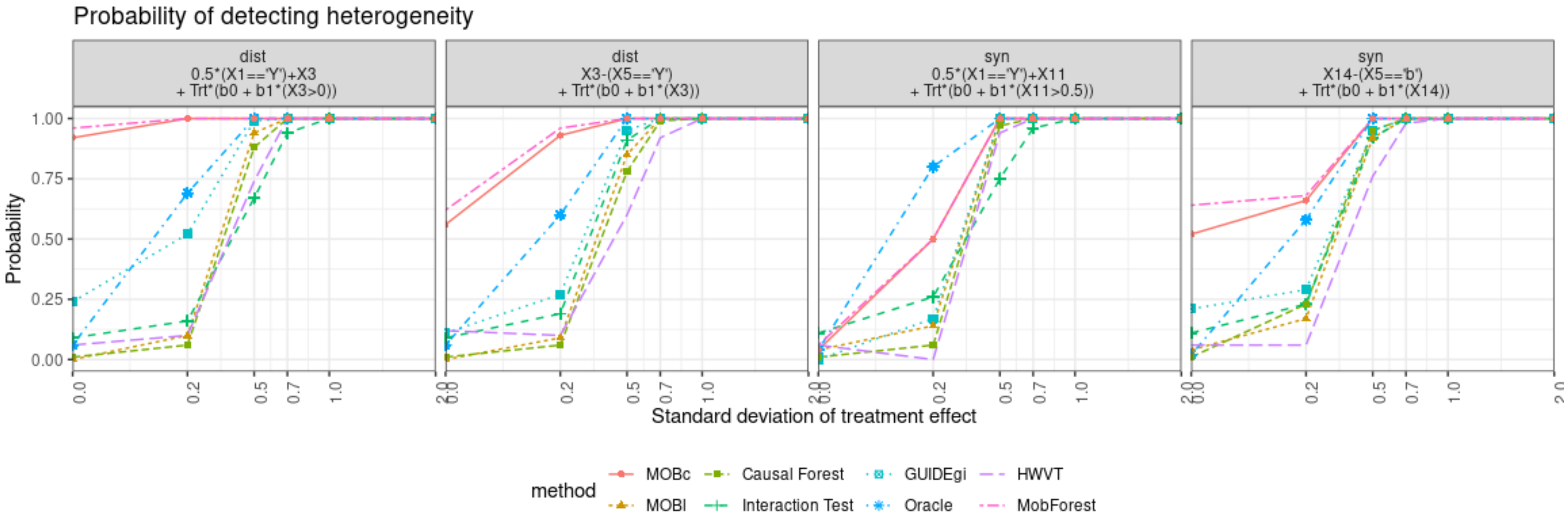
Performance Metrics: Treatment effect heterogeneity?

Setup:

- Continuous response
- Sample size n = 500
- power = 0.5
- Repeat on 1000

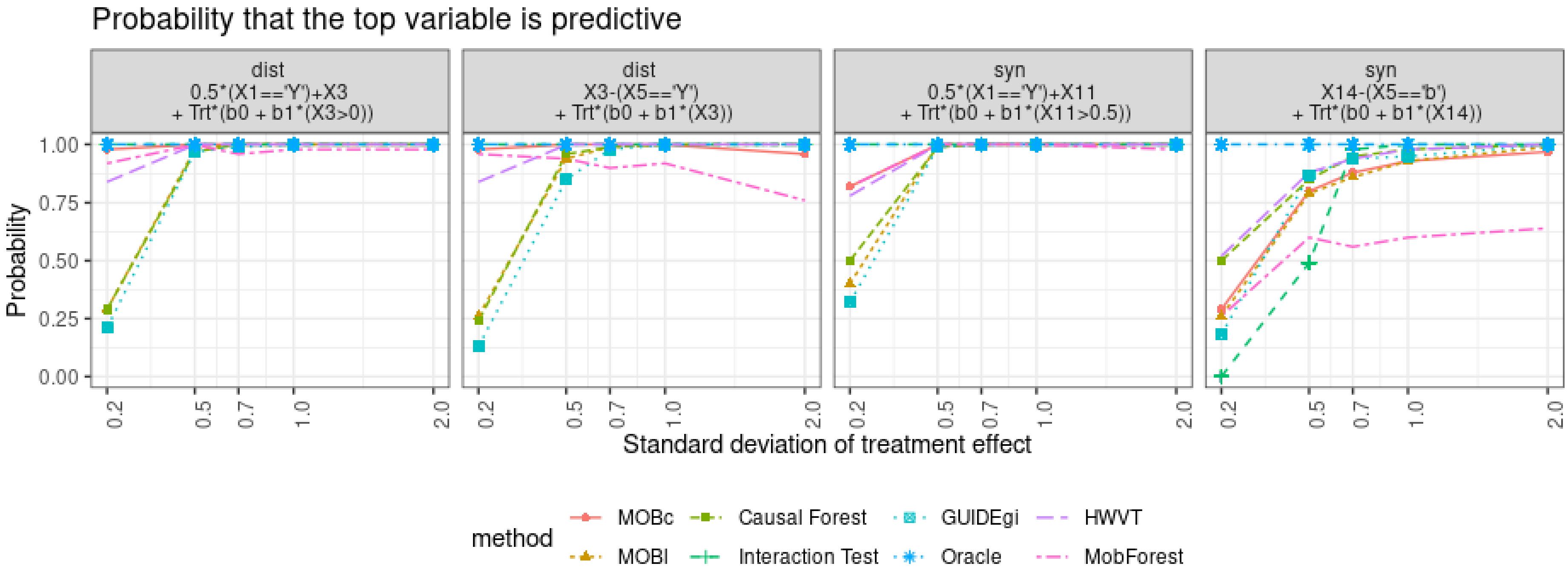
Methods to compare:

- CausalForest: Forest based
- MOB: MOBc with node fit $Y = Trt$, MOBI: $Y = \alpha Trt + \sum_j \beta_j X_j$, MobForest: ensemble MOBc
- GUIDE: tree-based
- Interaction test: Univariate interaction model
- HWVT: Holmes and Watson (2020) + virtual Twins implementation
- Oracle: true model



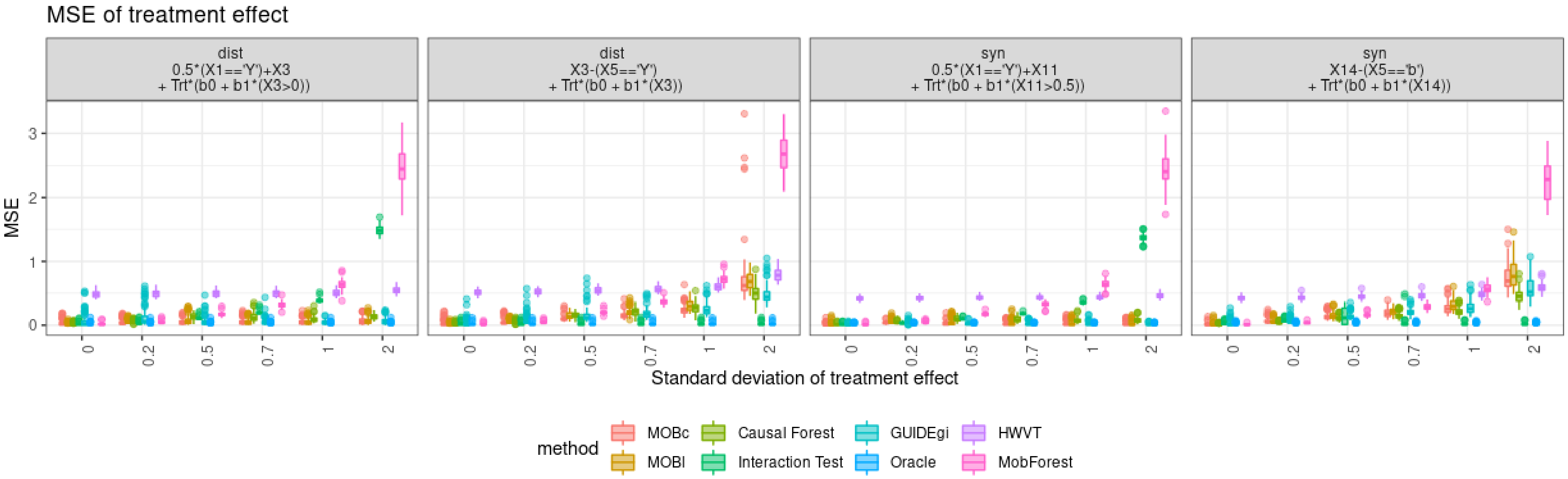
NOVARTIS | Reimagining Medicine

Performance Metrics: Variable selection bias



NOVARTIS | Reimagining Medicine

Performance Metrics: estimation of treatment effect

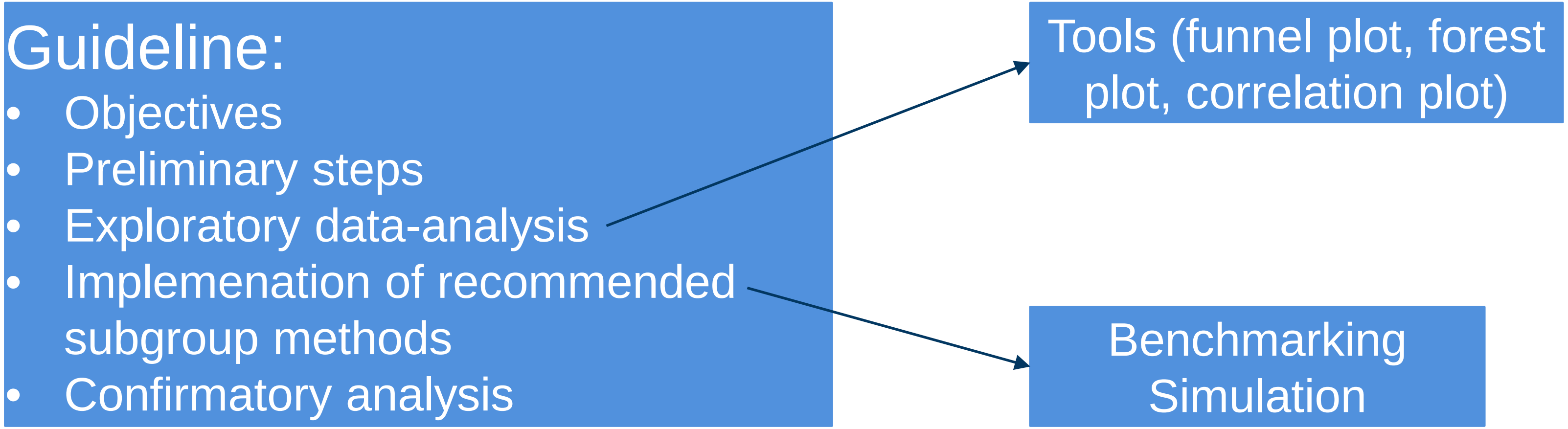


NOVARTIS | Reimagining Medicine

Subgroup identification method recommendation

- Benchmarking framework help us to better understand the performance of each methods for subgroup identification problems on:
  - Test for treatment effect heterogeneity
  - Variable selection bias (important variables are predictive)
  - Prediction bias (predicted treatment effect close to truth)
- Other than the performance part, whether a subgroup identification is recommended also depends on
  - Types of responses it can handle
  - Interpretability of the result
  - Computation cost
  - How easy it is to use the method (whether there is a package, how easy it is to use it in different systems)

Future work (go/subgroup)



Guideline:

- Objectives
- Preliminary steps
- Exploratory data-analysis
- Implemenation of recommended subgroup methods
- Confirmatory analysis

Tools (funnel plot, forest plot, correlation plot)

Benchmarking Simulation

Thank you

NOVARTIS | Reimagining Medicine

Reference

[IT] Su, Xiaogang, et al. "Subgroup analysis via recursive partitioning." Journal of Machine Learning Research 10.Feb (2009): 141-158.

[Quint] Dusseldorp, Elise, Lisa Doove, and Iven Van Mechelen. "Quint: An R package for the identification of subgroups of clients who differ in which treatment alternative is best for them." *Behavior research methods* 48.2 (2016): 650-663.

[FindIt] Imai, Kosuke, and Marc Ratkovic. "Estimating treatment effect heterogeneity in randomized program evaluation." The Annals of Applied Statistics 7.1 (2013): 443-470.

[VT] Foster, Jared C., Jeremy MG Taylor, and Stephen J. Ruberg. "Subgroup identification from randomized clinical trial data." Statistics in medicine 30.24 (2011): 2867-2880.

[GUIDE] Loh, Wei-Yin, et al. "Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables." Statistics in medicine 35.26 (2016): 4837-4855.[GUIDE]

[SIDES] Lipkovich, Ilya, et al. "Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations." *Statistics in medicine* 30.21 (2011): 2601-2621.

[IT] Su, Xiaogang, et al. "Interaction trees with censored survival data." *The international journal of biostatistics* 4.1 (2008).

[STIMA] Dusseldorp, Elise, Claudio Conversano, and Bart Jan Van Os. "Combining an additive and tree-based regression model simultaneously: STIMA." *Journal of Computational and Graphical Statistics* 19.3 (2010): 514-530.

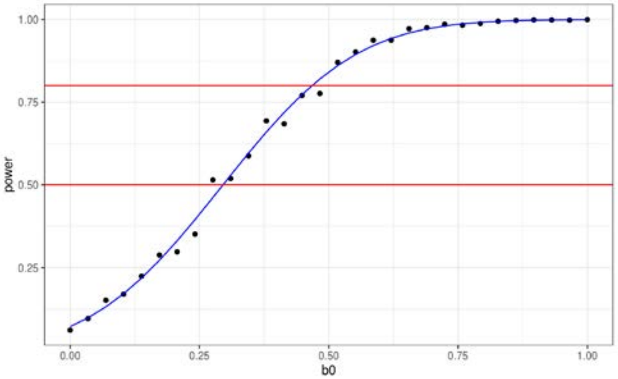
[Holmes Watson] Watson, James A., and Chris C. Holmes. "Machine learning analysis plans for randomised controlled trials: detecting treatment effect heterogeneity with strict control of type I error." Trials 21.1 (2020): 156.

[MOB] Zeileis, Achim, Torsten Hothorn, and Kurt Hornik. "Model-based recursive partitioning." Journal of Computational and Graphical Statistics 17.2 (2008): 492-514.

How to choose $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ in benchtm?

Generate data from: $f(X) = f_{prog}(X) + Trt * (\beta_0 + \beta_1 * f_{pred}(X))$

1. Can specify β_0 and β_1 (difficult in practice)
2. Derive β_0 and β_1 based on the *overall treatment effect power* and *the standard deviation of the treatment effect*
   - Derive β_1 based on $\mathrm{sd}_{\mathrm{TE}}$: $\mathrm{sd}_{\mathrm{TE}} = \beta_1 sd(f_{pred}(X))$
   - Find β_0 for a pre-specified overall power (for the naive unadjusted test)



- Depends on covariate distribution for X and structural form of $f_{pred}(X)$ as well as $f_{prog}(X)$

NOVARTIS | Reimagining Medicine

Acknowledge

AMDS: Janice Branson

- SMC: Mouna Akacha, Björn Bornkamp, Yao Chen, Jiarui Lu
- AEA: David Ohlssen, Kostas Sechidis, Sophie Sun
- SCC: Douglas Robinson, Ardalan Mirshani
- CTS: Mark Baillie

Collaborations:

- PMX: Chong Ma
- NIBR: Pablo Serrano
- DS&AI: Brian Buckley