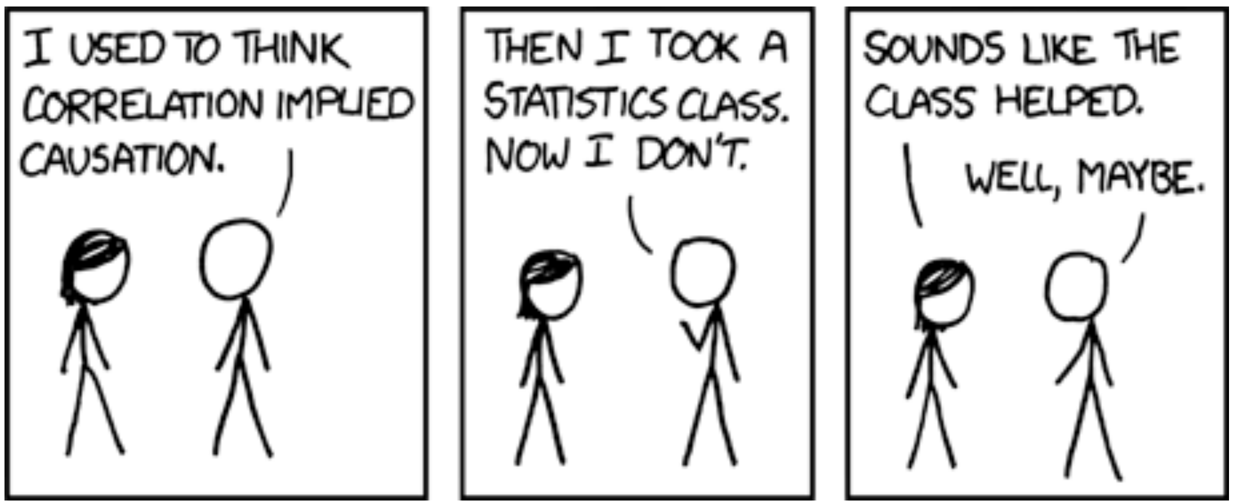
# Causal Inference and Response Surface Modeling

Inference and Representation

DS-GA-1005 Fall 2015

Guest lecturer: Uri Shalit

# What is Causal Inference?




source: xkcd.com/552/

# Causal questions as counterfactual questions

- Does this medication improve patients health?
  - Counterfactual: taking vs. not taking
- Is the new design bringing more customers?
  - Counterfactual: new design vs. old design
- Is online teaching better than in-class?
  - Counterfactual: …

# Potential Outcomes Framework (Rubin's Causal Model)

- Each unit (patient, customer, student, cell culture) has **two** potential outcomes: $(y^0, y^1)$
  - $y^0$ is the potential outcome had the unit not been treated: "***control outcome***"
  - $y^1$ is the potential outcome had the unit been treated: "***treatment outcome***"
- Treatment effect for unit *i*
  $= \quad y_i^1 - y_i^0$
- Often interested in mean or expected treatment effect

# Hypothetical example – effect of fish oil supplement on blood pressure (Hill & Gelman)

| Unit | female | age | treatment | potential outcome $y_i^0$ | potential outcome $y_i^1$ | observed outcome $y_i$ |
|---|---|---|---|---|---|---|
| Audrey | 1 | 40 | 0 | **140** | 135 | 140 |
| Anna | 1 | 40 | 0 | **140** | 135 | 140 |
| Bob | 0 | 50 | 0 | **150** | 140 | 150 |
| Bill | 0 | 50 | 0 | **150** | 140 | 150 |
| Caitlin | 1 | 60 | 1 | 160 | **155** | 155 |
| Cara | 1 | 60 | 1 | 160 | **155** | 155 |
| Dave | 0 | 70 | 1 | 170 | **160** | 160 |
| Doug | 0 | 70 | 1 | 170 | **160** | 160 |

Source: Jennifer Hill

$\text{Mean}(y_i^1 – y_i^0) = -7.5$

$\text{Mean}( (y_i|\text{treatment}=1) - (y_i|\text{treatment}=0)) = 12.5$

# The fundamental problem of causal inference: We only ever observe one of the two outcomes

- How to deal with The Problem:
  - Close substitutes
  - Randomization
  - Statistical Adjustment

# Fundamental Problem (I): Close Substitutes

- Does chemical X corrode material M? Create a piece of material M, break it into. Place chemical on one piece.
- Does removing meat from my diet reduce my weight?
  My weight before the diet is a close substitute to my weight after the diet *had I not gone on the new diet*
- Separated twin studies.

***What assumptions have we made here?***

# Fundamental Problem (II): Randomization

- Assume the outcomes are generated from a distribution.
- Therefore if we sample enough times, we can estimate the mean effect:
  - Obtain a sample of the items of interest. Assign half to treatment and half to control, **at random**
  - This yields two estimates:
    $y_1^0, \dots, y_n^0$
    $y_{n+1}^1, \dots, y_{2n}^1$
  - Average the estimates

# Fundamental Problem (III): Statistical Adjustment

- Sometimes we can't find close substitutes, and can't randomize, for example:
  - Non-compliance: some of the people did not follow the new diet proscribed in the experiment.
  - Ethical: does breathing Asbestos cause cancer?
  - Impractical: do stricter gun laws lead to safer communities?
  - Retrospective: we have data from the past, for example educational attainment and college attendance.

- Control and treatment populations are different

# Fundamental Problem (III): Statistical Adjustment

- Treatment and control group are not similar – what can we do?

- Estimate the outcomes using a model, such as linear regression, random forests, BART (later today). Known as Response Surface Modeling
- Divide the sample into similar subgroups
- Re-weight the units to be more representative

*Today we will focus on statistical adjustment with response surface modeling*

# Response Surface Modeling: Linear Regression

True model:

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \varepsilon_i$$

Fit without **confounding** variable $x_i$:

$$y_i = \beta_0^* + \beta_1^* T_i + \varepsilon_i$$

Represent $x_i$ as a function $T_i$:

$$x_i = \gamma_0 + \gamma_1 T_i + \theta_i$$

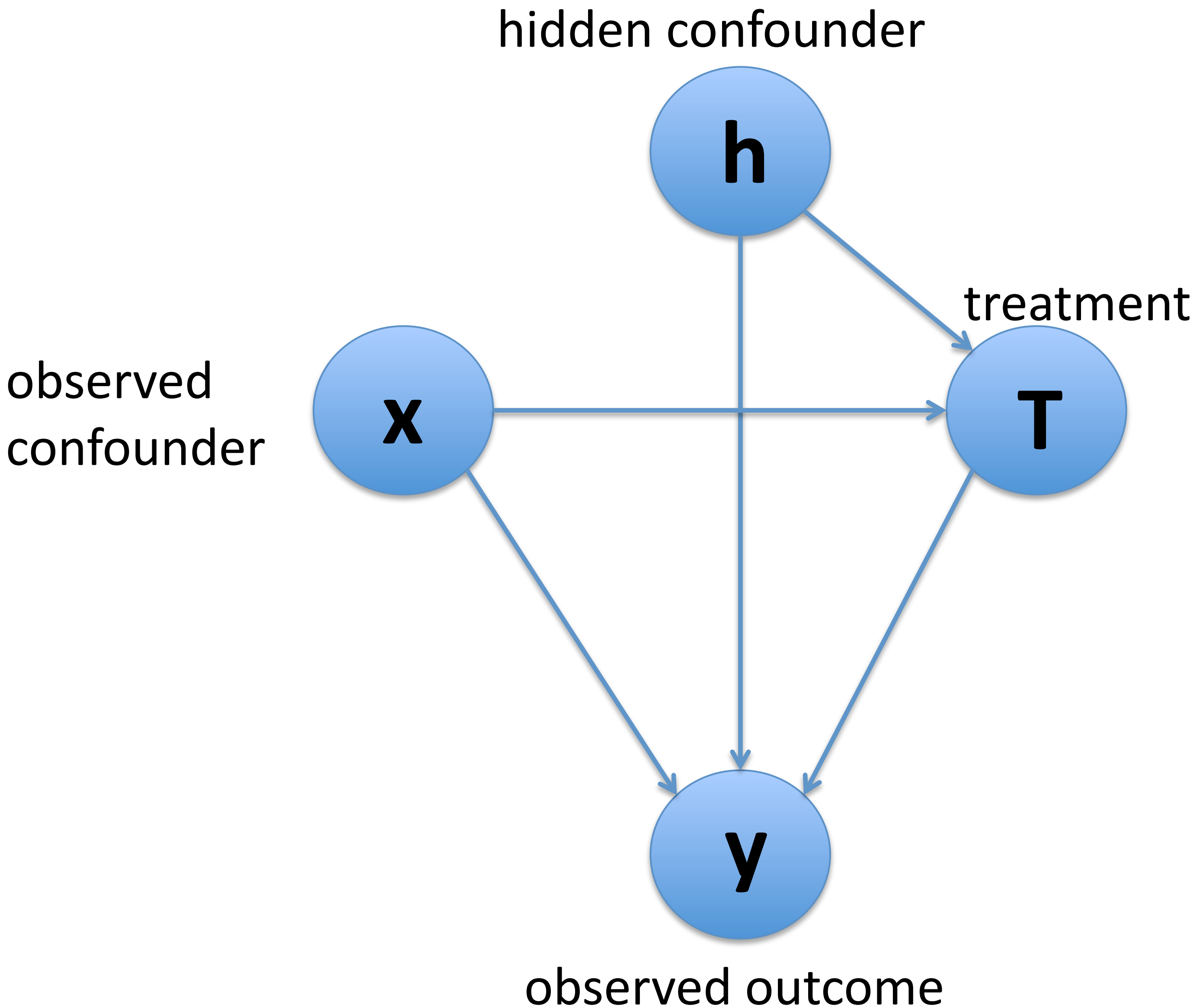
Obtain:

$$\beta_1^* = \beta_1 + \beta_2 \gamma_1$$

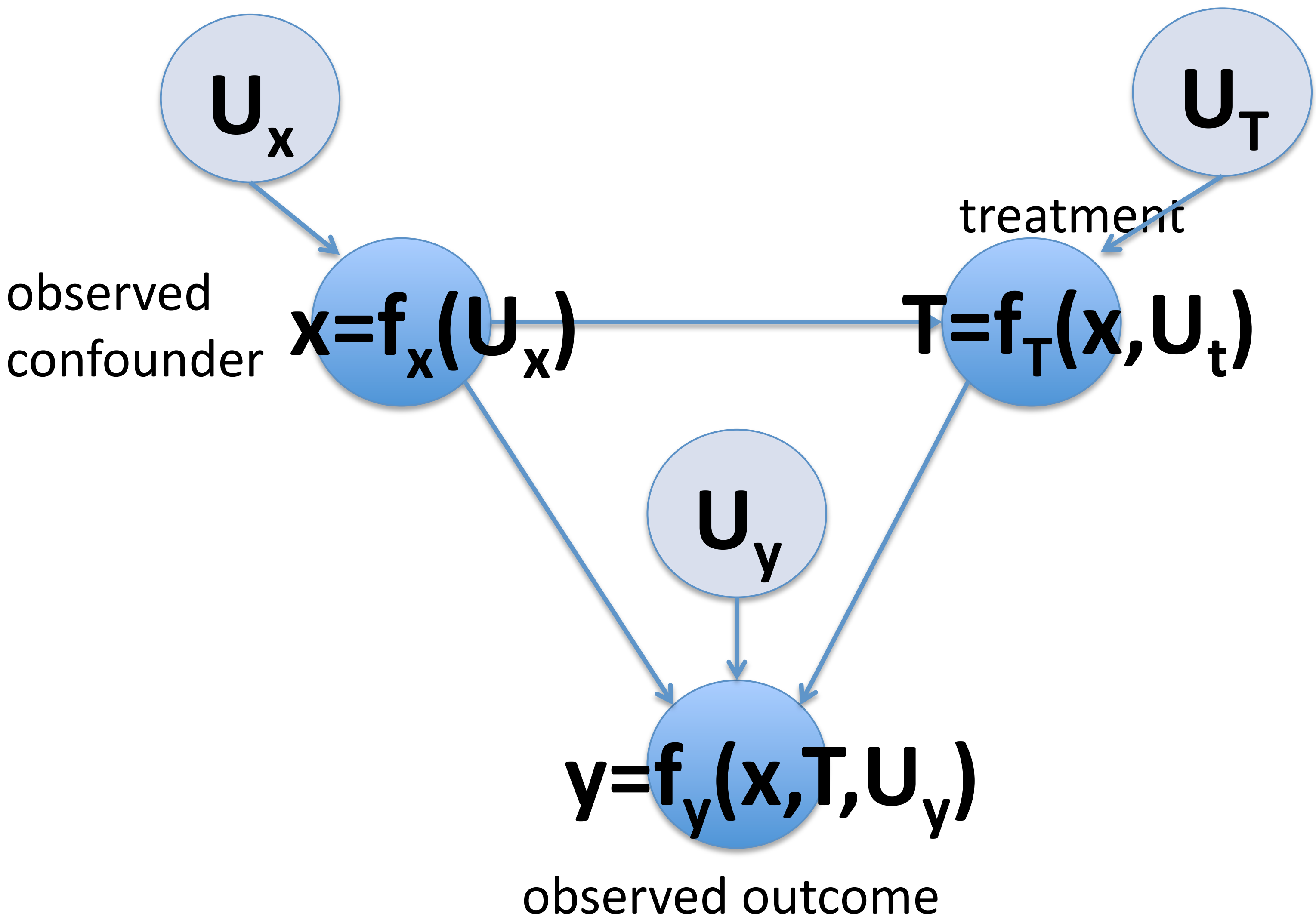
# When will this work?

- No hidden confounders
- Model is correct

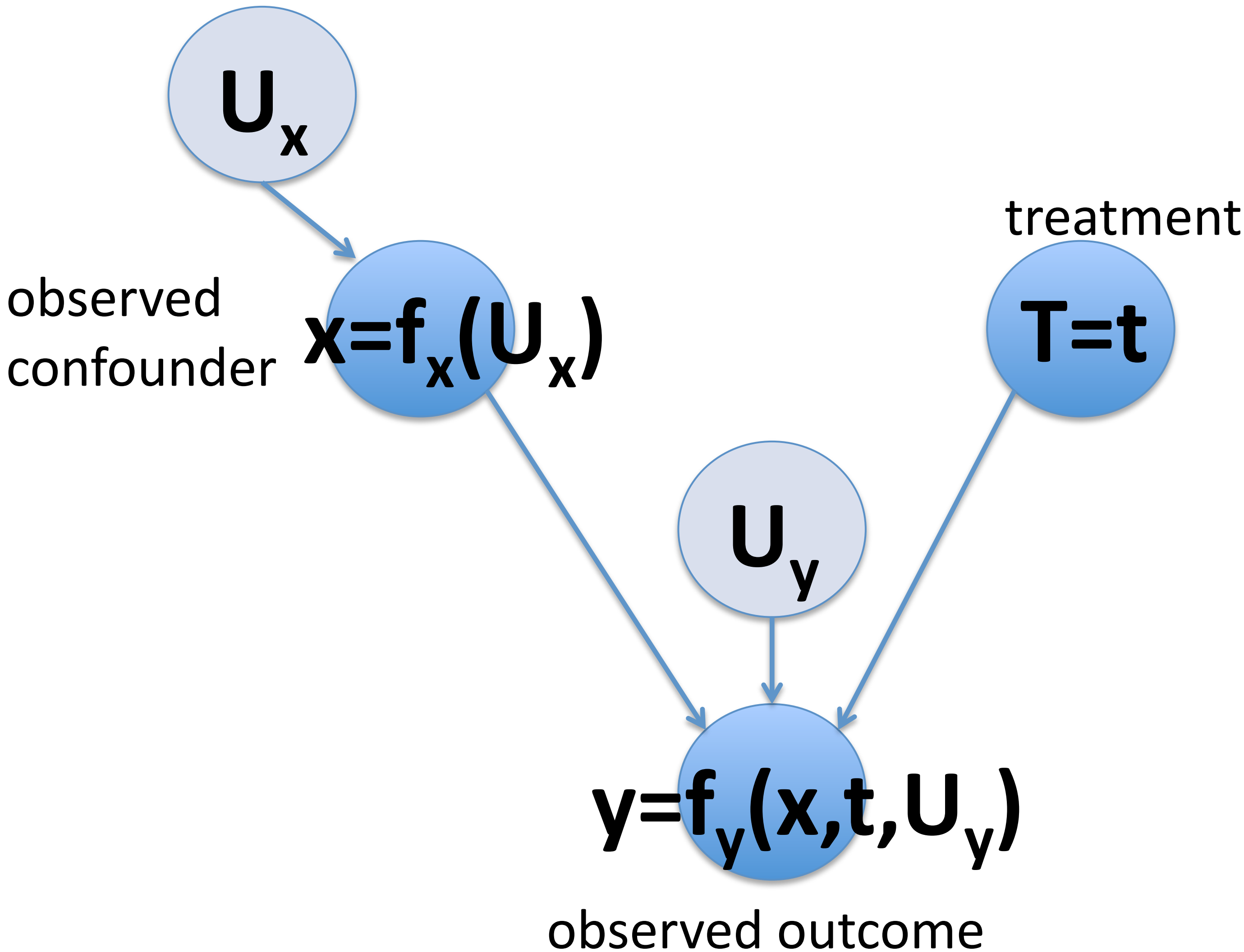
- Both assumptions patently false. How can we make them less false?


hidden confounder
h
treatment
observed
confounder
x
T
y
observed outcome

# Pearl's do-calculus and structural equation modeling



# Pearl's do-calculus and structural equation modeling



# Response Surface Modeling

- We wish to model $U_x$, $f_x(U_x)$, $U_y$, and $f_y(U_y,x,t)$.
- In principle any regression method can work: use $t=T_i$ as a feature, predict for both $T_i=0$, $T_i=1$.
- Linear regression is far too weak for most problems of interest!

# Response Surface Modeling: BART

- In principle ***any*** regression method can work: use $T_i$ as a feature, predict for both $T_i$=0, $T_i$=1.
- In 2008, Chipman, George and McCulloch introduced Bayesian Additive Regression Trees (BART).
- BART is non-linear, yet easy to fit and empirically robust to model misspecification.
- Proven as very successful for causal inference, especially adopted in the social sciences.

# Bayesian Additive Regression Tress (BART)

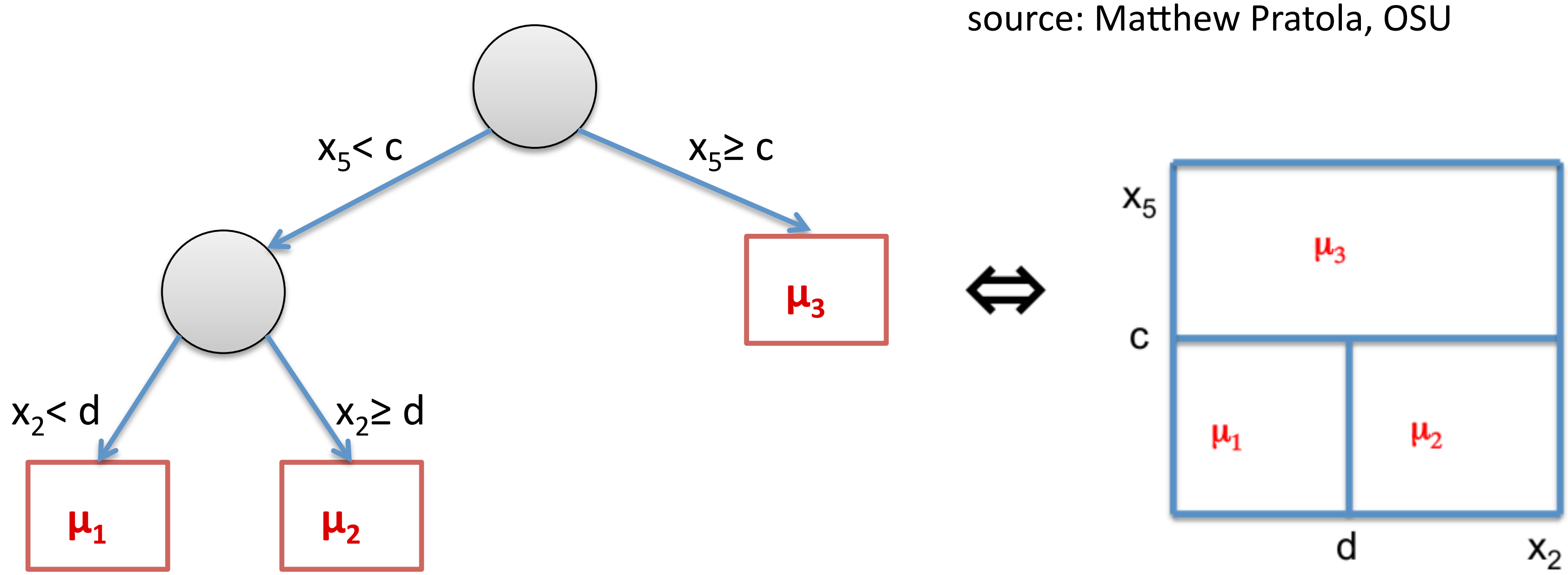
Chipman, H. A., George, E. I., & McCulloch, R. E. (2010).
*BART: Bayesian additive regression trees*.
The Annals of Applied Statistics, 266-298.

# `bartMachine`

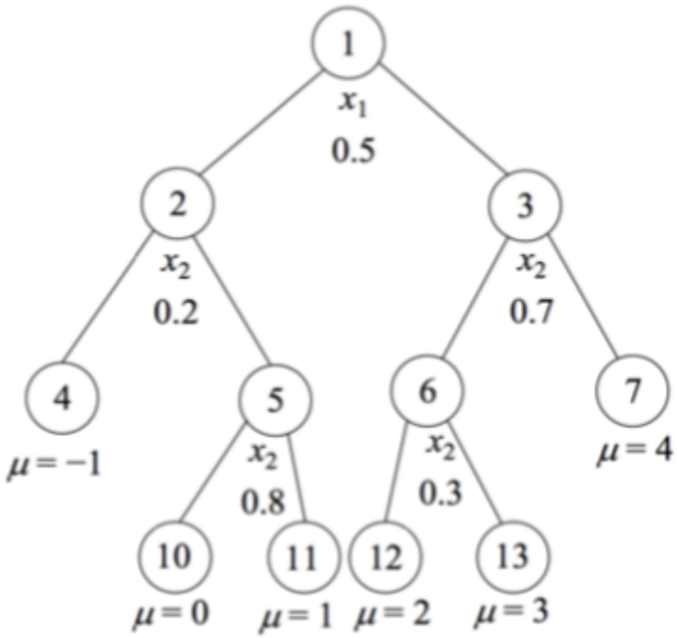
Kapelner, A., & Bleich, J. (2013).
*bartMachine: Machine Learning with Bayesian Additive Regression Trees*.
arXiv preprint arXiv:1312.2171.

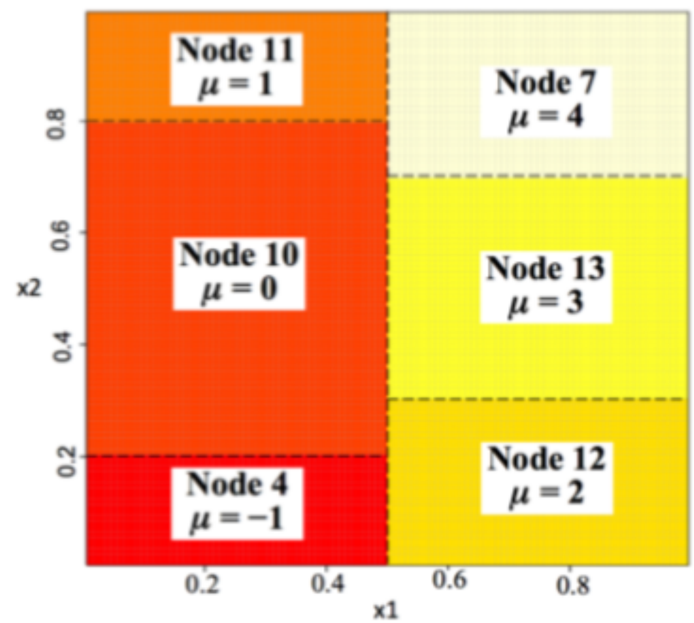
# What's a regression tree?

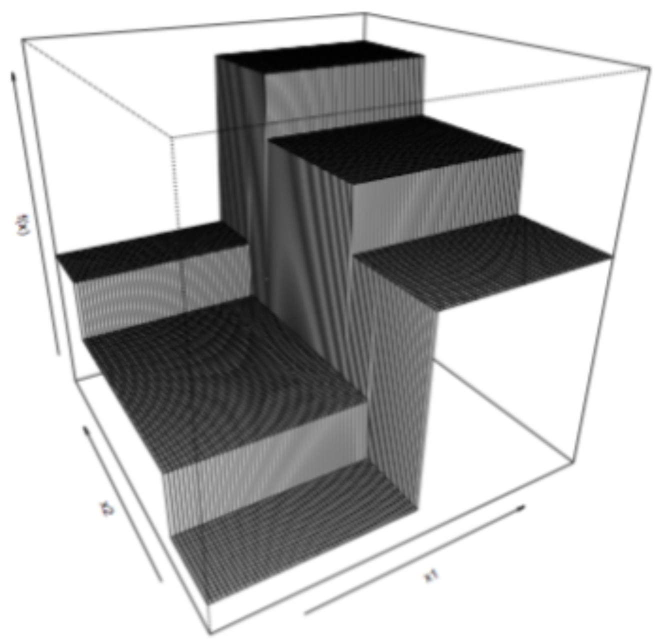
source: Matthew Pratola, OSU




$\mu_k(x)$ can be e.g. linear function, a Gaussian process, or just a constant.



Three different views of a bivariate single tree.

source: Matthew Pratola, OSU

# Bayesian Regression Trees

- Each tree is a function g(· ; T, M) parameterized by:
  - Tree structure T
  - Leaf functions M
- Bayesian framework:
  - Data is generated $y(x) = g(\cdot\ ; T, M) + \varepsilon$, $\varepsilon \sim N(0,\sigma^2)$
  - Prior: $\pi(M,T,\sigma^2) = \pi(M|T,\sigma^2)\pi(T|\sigma^2)\,\pi(\sigma^2)$

# Bayesian *Additive* Regression Trees

- Each tree is a function $g(\cdot\,; T, M)$ parameterized by:
  - Tree structure T
  - Leaf functions M
- Bayesian framework:
  - Data is generated $y(x) = g(\cdot\,; T, M) + \varepsilon$, $\varepsilon \sim N(0,\sigma^2)$
  - Prior: $\pi(M,T,\sigma^2) = \pi(M|T)\,\pi(T)\,\pi(\sigma^2)$
- Additive tress:
  - Data is generated $y(x) = \Sigma_{j=1\ldots m}\, g(\cdot\,; T_j, M_j) + \varepsilon$, $\varepsilon \sim N(0,\sigma^2)$, where each $g(\cdot\,; T_j, M_j)$ is a single tree
  - Prior factorizes: $\pi((M_1,T_1),\ldots,(M_m,T_m),\sigma^2) = (\Pi_{j=1\ldots m}\, \pi\,(M_j|T_j,\sigma^2)\,\pi(T_j|\sigma^2))\,\pi(\sigma^2)$

# Prior over tree structure $\pi(T)$

- Nodes at depth *d* are non-terminal with probability $\alpha(1+d)^{-\beta}$, $\alpha \in (0,1)$, $\beta \in [0,\infty]$
  - Restricts depth
  - Standard implementation: $\alpha=0.95$, $\beta=2$
- Non-terminal node: split on a random variable, choose splitting value at random from multiset of available values at the node

# Prior over leaf functions $\pi(M|T)$

- Leaf functions are constants
- Leaf nodes: i.i.d. $\mu_k \sim N(\mu_\mu , \sigma_\mu^2)$
- $\mu_\mu = (y_{max} - y_{min})/2m$
- $\sigma_\mu^2$ chosen such that $\mu_\mu \pm 2\sigma_\mu^2$ covers 95% of observed y values

# Prior over variance $\pi(\sigma^2)$

- Recall prior: $\pi(M,T,\sigma^2) = \pi(M|T)\,\pi(T)\,\pi(\sigma^2)$
- $\pi(\sigma^2) \sim \mathit{InvGamma}(\nu/2, \nu\lambda/2)$
  where ν, λ are determined using a data guided heuristic

# Likelihood model $p(y|M,T,\sigma^2)$

- Likelihood of outcome at node k:
  $y_k \sim N(\mu_k, \sigma^2)$

# Sampling from the posterior

Gibbs sample from $p((M_1,T_1),\ldots,(M_m,T_m),\sigma^2|y,X)$

Define $\mathbf{R_{-j} = y - \Sigma_{k\neq j}\, g(X;T_k,M_k)}$, *the unexplained response*

$1 : T_1 \mid R_{-1}, \sigma^2$

$2 : M_1 \mid T_1, R_{-1}, \sigma^2$

$3 : T_2 \mid R_{-2}, \sigma^2$

$4 : M_2 \mid T_2, R_{-2}, \sigma^2$

$\vdots$

$2m-1 : T_m \mid R_{-m}, \sigma^2$

$2m : M_m \mid T_m, R_{-m}, \sigma^2$

$2m+1 : \sigma^2 \mid T_1, M_1, \ldots, T_m, M_m$, error

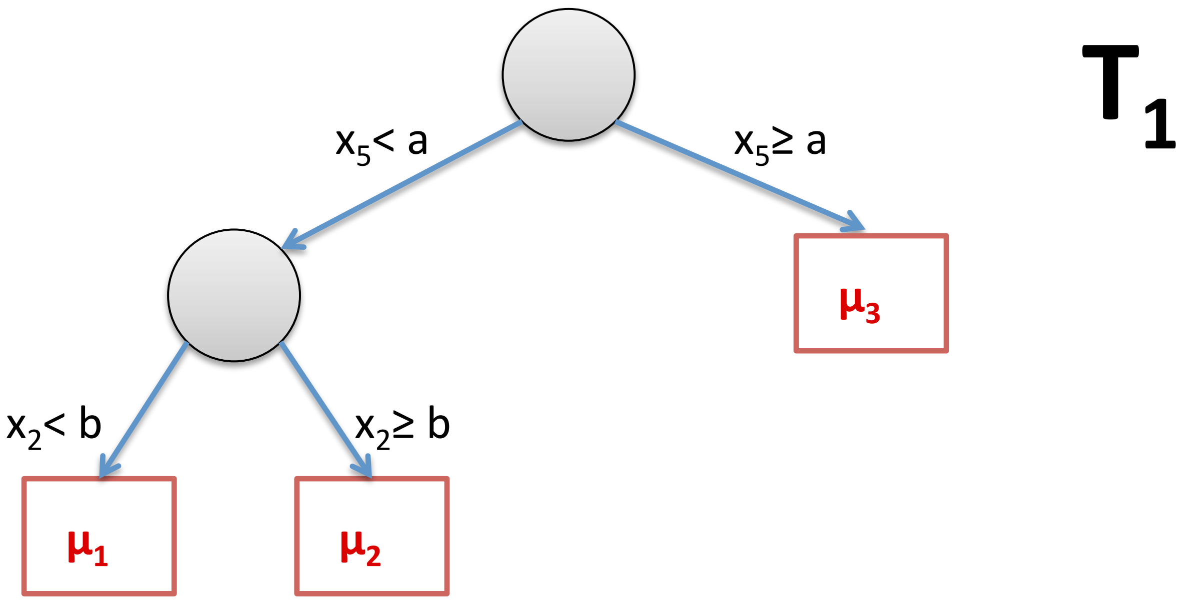
(error $= y - \Sigma_k\, g_k(X;T_k,M_k)$ )

# Sampling

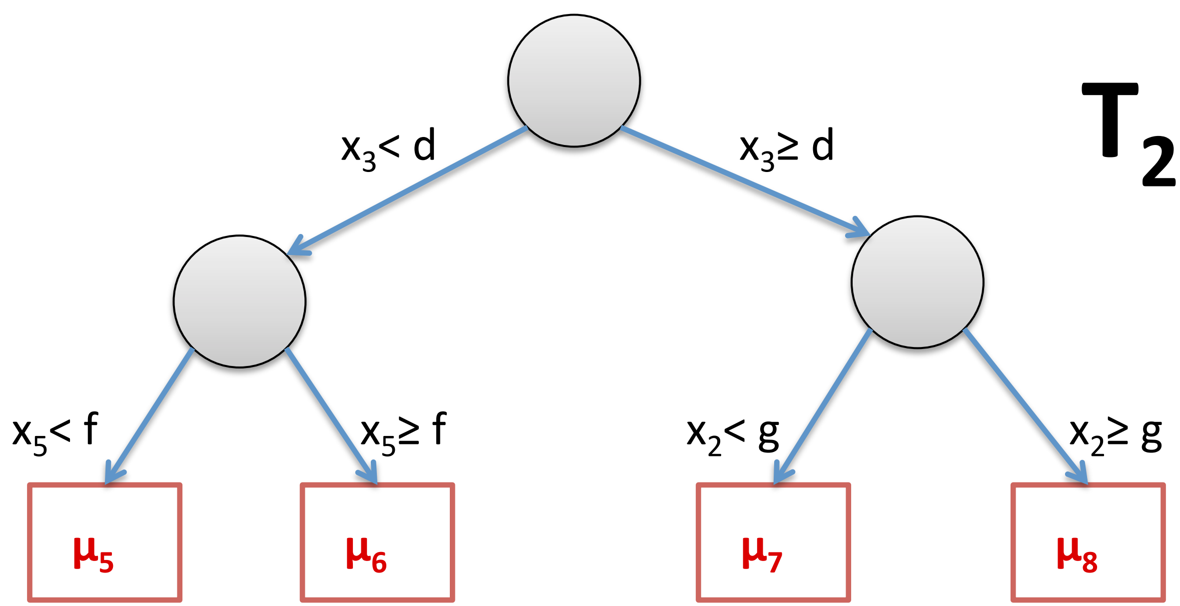
- Leaf node values $M_i|T_i,R_{-i}$ are normally distributed
- $\sigma^2$ is an inverse gamma by conjugacy
- The difficult part is sampling the tree structures

# Metropolis-Hastings sampling of trees I

- Three different "rules":
  - GROW, chosen with probability $p_{grow}$
  - PRUNE, chosen with probability $p_{prune}$
  - CHANGE, chose with probability $p_{change}$
- Each rule potentially changes the probability of the tree and the likelihood of the observations
- GROW: add two child nodes to a terminal node
- PRUNE: prune two child nodes, making their parent a terminal node
- CHANGE: re-sample node splitting rule

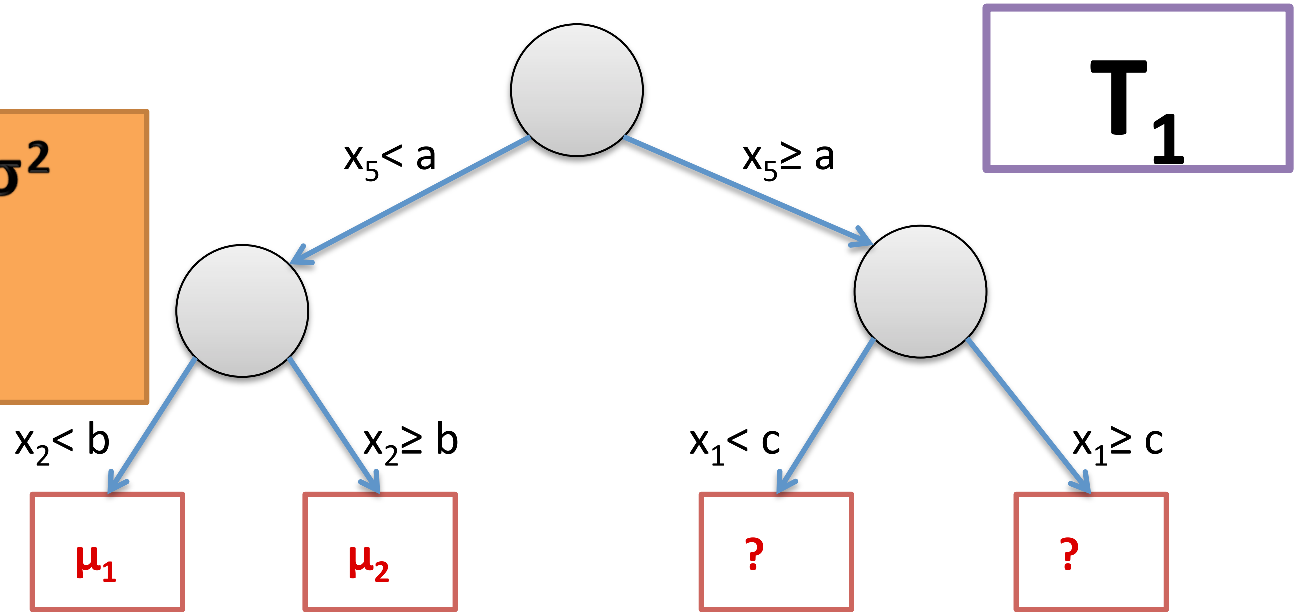


T1
x5< a
x5≥ a
x2< b
x2≥ b
μ1
μ2
μ3

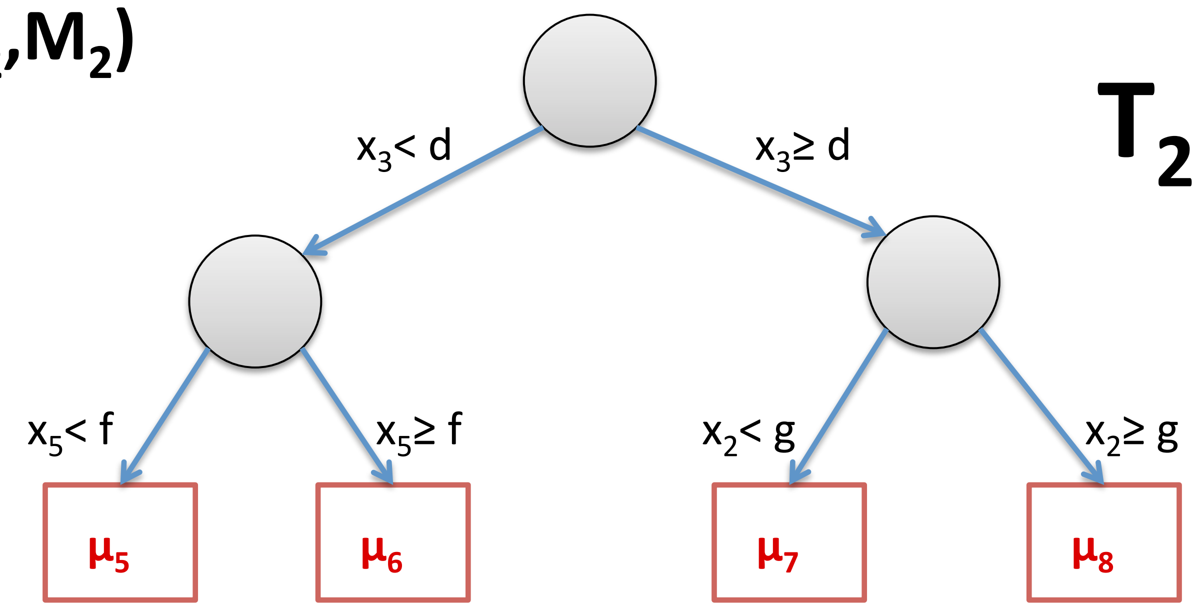



T2
x3< d
x3≥ d
x5< f
x5≥ f
x2< g
x2≥ g
μ5
μ6
μ7
μ8

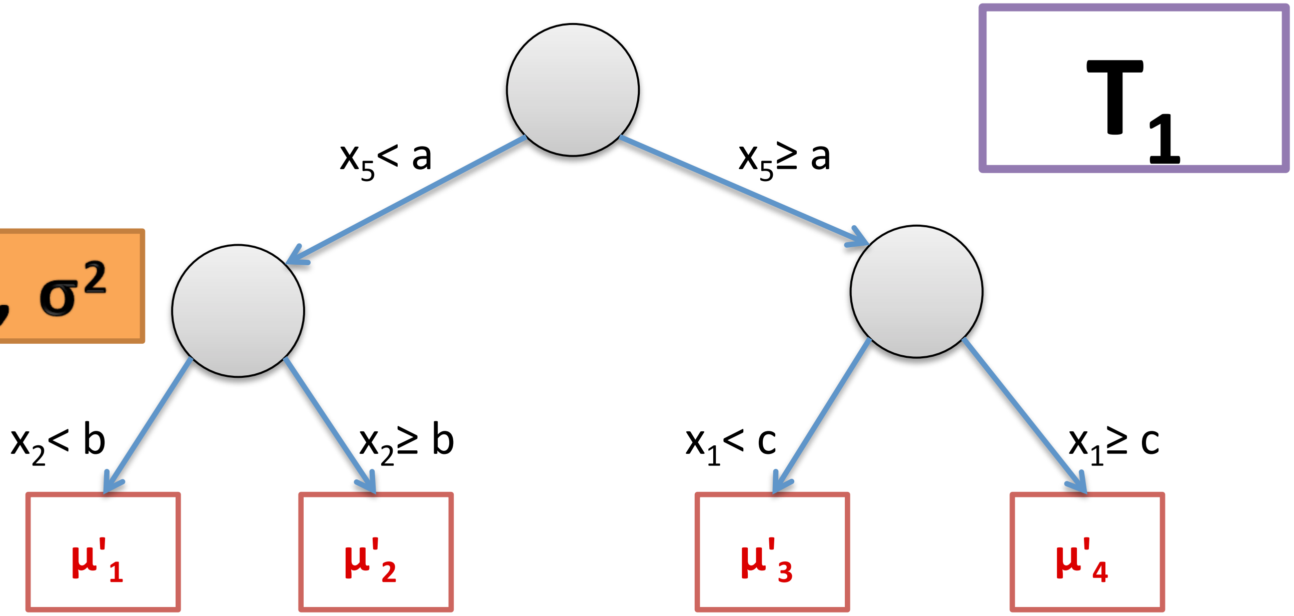
$T_1 \mid R_{-1}, \sigma^2$
"GROW"



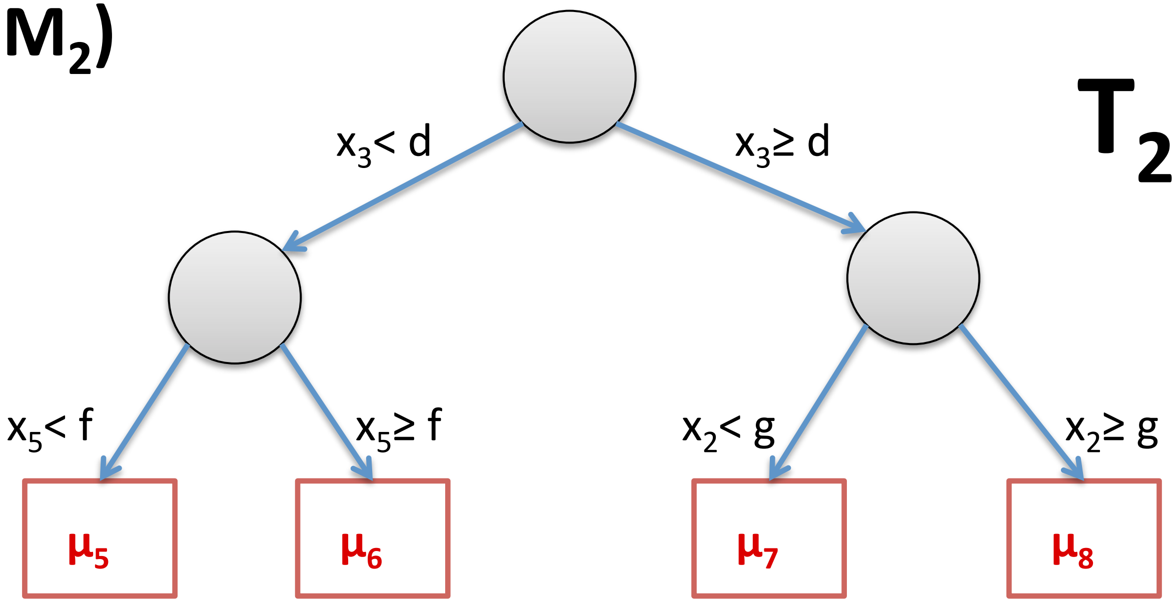

$$R_{-1} = y - g(X;T_2,M_2)$$



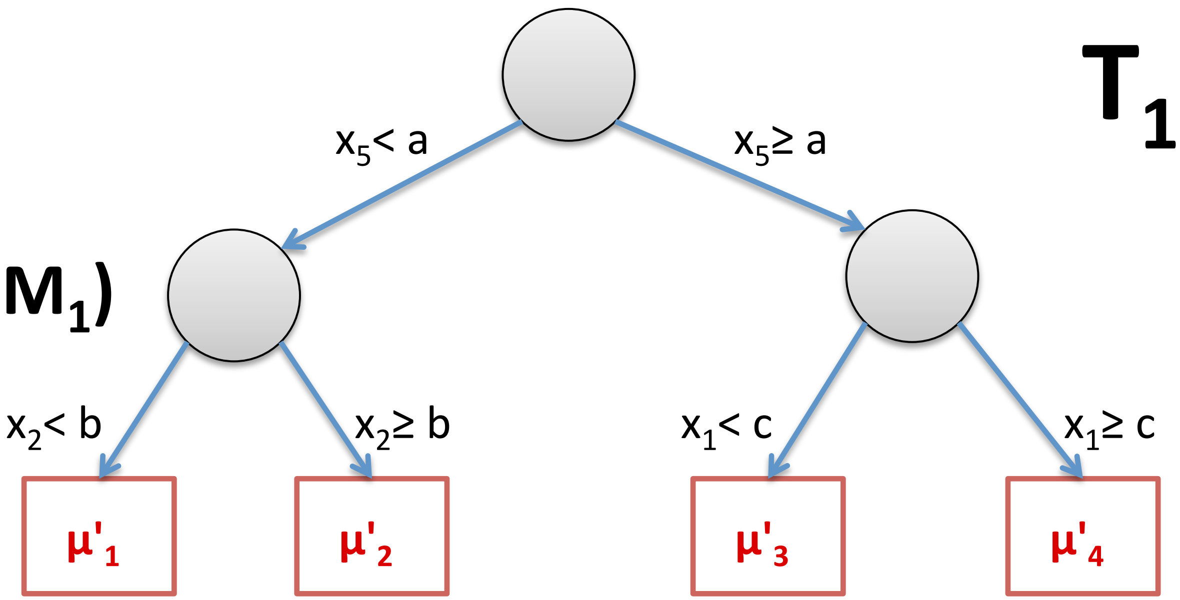
$M_1 \mid T_1, R_{-1}, \sigma^2$



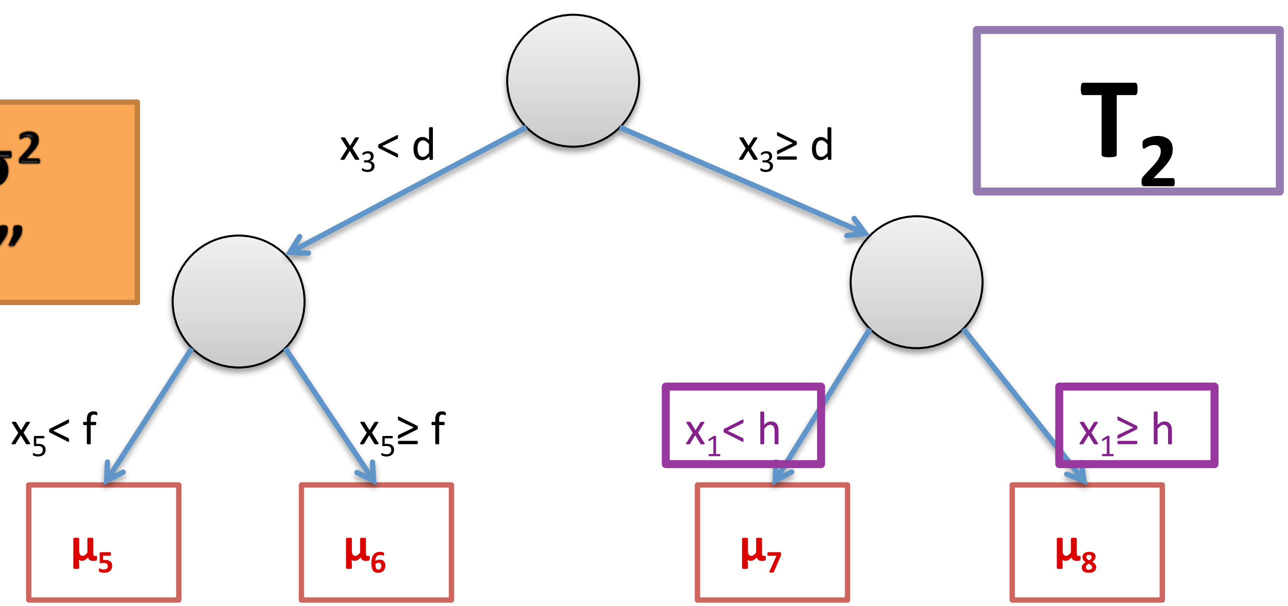

$R_{-1} = y - g(X;T_2,M_2)$



$R_{-2} = y - g(X;T_1,M_1)$

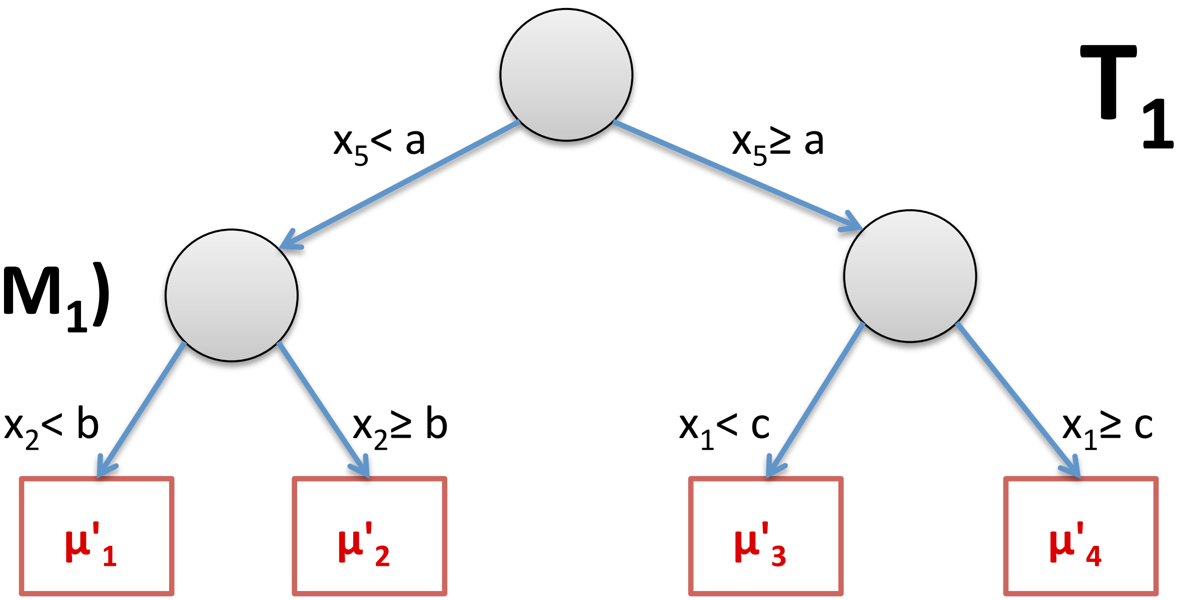



$T_2 \mid R_{-2}, \sigma^2$
**"CHANGE"**



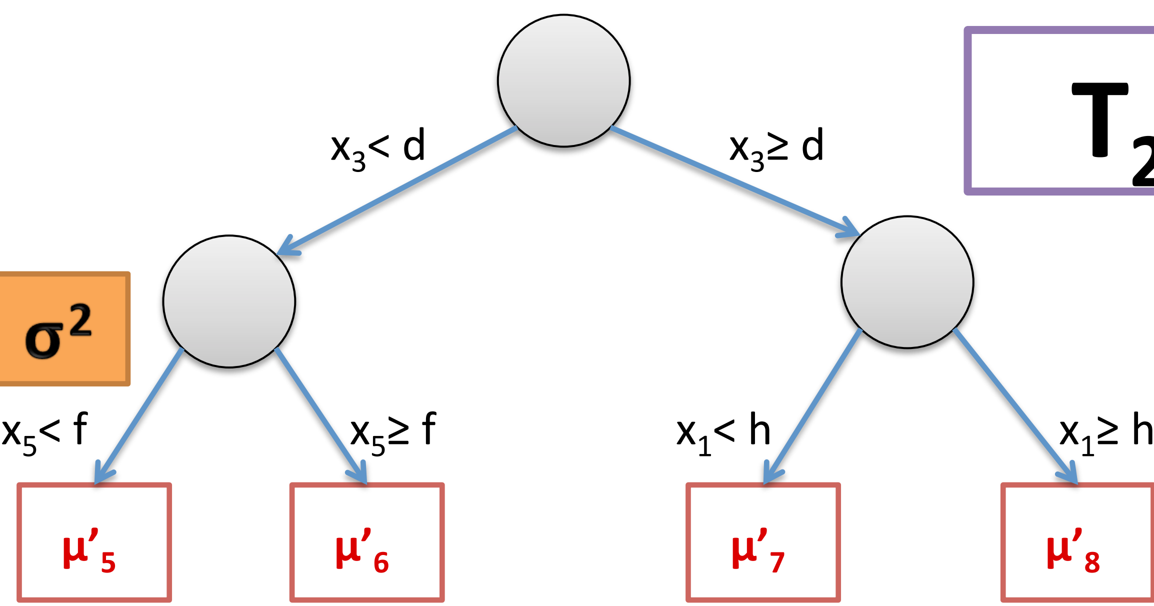
$T_1$

$R_{-2} = y - g(X;T_1,M_1)$




$T_2$

$M_2 \mid T_2, R_{-2}, \sigma^2$



# Metropolis-Hastings sampling of trees II

- Proposal distribution (sometimes denoted Q) ratio, where R is the current unexplained response:

$$r = \frac{p(T_* \rightarrow T)\, p(T_* \mid R, \sigma^2)}{p(T \rightarrow T_*)\, p(T \mid R, \sigma^2)}$$

- Sample u~uniform(0,1)
  if u<min(1,r):
  – update tree to $T_*$
  else:
  – stay with T

# How to calculate the acceptance probability *r*

$$r = \frac{p(T_* \to T)\, p(T_* \mid R, \sigma^2)}{p(T \to T_*)\, p(T \mid R, \sigma^2)}$$

- Calculating p(T|R, $\sigma^2$ ) is hard
- Use Bayes law:

$$p(T \mid R, \sigma^2) = \frac{p(R \mid T, \sigma^2)\, p(T \mid \sigma^2)}{p(R \mid \sigma^2)}$$

- Obtain:

$$r = \frac{p(T_* \to T)}{p(T \to T_*)} \times \frac{p(R \mid T_*, \sigma^2)}{p(R \mid T, \sigma^2)} \times \frac{p(T_*)}{p(T)}$$

# The acceptance probability

$$r = \frac{p(T_* \rightarrow T)}{p(T \rightarrow T_*)} \times \frac{p(R \mid T_*, \sigma^2)}{p(R \mid T, \sigma^2)} \times \frac{p(T_*)}{p(T)}$$

**transition ratio** | **likelihood ratio** | **tree structure ratio**

- Calculate the three terms for each of the updates GROW, PRUNE, CHANGE
- We will only calculate the transition ratio and tree structure ratio for the GROW rule

# GROW rule transition ratio I

$$p(T \to T_*) = p_{grow} \times p(selecting\_node\_\eta) \times$$
$$p(selecting\_j\_feature\_to\_split) \times$$
$$p(selecting\_k\_value\_to\_split) =$$
$$p_{grow} \times \frac{1}{b} \times \frac{1}{f_{adj}(\eta)} \times \frac{1}{n_{j \cdot adj}(\eta)}$$

b=#terminal nodes

$f_{adj}(\eta)$ is number of features left to split on.
Can be smaller than d if a feature has less than two available values at node η)

$n_{j \cdot adj}(\eta)$ is number of *unique* values left to split on in the j-th feature at node η

# GROW rule transition ratio II

$$p(T_* \rightarrow T) =$$

$$p_{prune} \times p(selecting\_node\_\eta\_to\_prune) =$$

$$p_{prune} \times \frac{1}{w_2^*}$$

$w_2^*$=#nodes with 2 terminal child nodes

$$\frac{p(T_* \rightarrow T)}{p(T \rightarrow T_*)} = \frac{p_{prune}}{p_{grow}} \frac{b \cdot f_{adj}(\eta) \cdot n_{j \cdot adj}(\eta)}{w_2^*}$$

# GROW rule transition ratio III

b=#terminal nodes

$f_{adj}(\eta)$ is number of features left to split on.
Can be smaller than d if a feature has less than two available values at node η)

$n_{j\cdot adj}(\eta)$ is number of *unique* values left to split on in the j-th feature at node η

$w_2^*$=#nodes with 2 terminal child nodes

$$\frac{p(T_* \rightarrow T)}{p(T \rightarrow T_*)} = \frac{p_{prune}}{p_{grow}} \frac{b \cdot f_{adj}(\eta) \cdot n_{j\cdot adj}(\eta)}{w_2^*}$$

# GROW rule tree structure ratio

The proposal tree $T^*$ differs from T in two child nodes: $\eta_L$ and $\eta_R$

$$\frac{p(T_*)}{p(T)} = \frac{\left(1-\frac{\alpha}{(1+d_{\eta_L})^{\beta}}\right)\left(1-\frac{\alpha}{(1+d_{\eta_R})^{\beta}}\right)\frac{\alpha}{(1+d_{\eta})^{\beta}}\frac{1}{f_{adj}(\eta)}\frac{1}{n_{j\cdot adj}(\eta)}}{\left(1-\frac{\alpha}{(1+d_{\eta})^{\beta}}\right)}$$

# GROW rule likelihood ratio

- Somewhat tedious math.
- The assumption of normal distributions of the responses and normal priors allows this to be solved analytically.

# BART algorithm overview

- data $X \in R^{d \times n}$, responses $y \in R^n$
- Choose hyperparameters
  – m (number of trees); α, β (tree structure prior); ν, λ (variance prior), and possibly others
- Run Gibbs sampling, cycle over *m* trees:
  – Change tree structure with one of 3 rules (GROW, PRUNE, CHANGE), sample with MH acceptance prob.
  – Sample leaf variables, using normal conjugacy
  – Sample variance σ using inverse Gamma conjugacy
- 1000 burn in iterations over all m trees
- 1000 additional draws to estimate posterior

# Prediction Intervals

- Quantiles of posterior estimate after "burn-in" provide confidence estimates for prediction

# BART use case (semi authentic) – Infant Health and Development Program*

- Population: children who were born prematurely with low weight
- Treatment **T**: give intensive high-quality child care and home visits from a trained provider
- Outcome(s) **y**: IQ test, visual-motor skills test
- Features **X**: birth weight, sex, mother_smoked, mother_education, mother_race, mother_age, prenatal_care, state (overall 25 features)

*Hill, J. L. (2011). *Bayesian nonparametric modeling for causal inference*. Journal of Computational and Graphical Statistics, 20(1).

# BART use case

- Treatment given only to children of nonwhite mothers – race is confounding variable.
  Other confounders as well?
- Fit BART function **g(X,T)** to observed outcomes **y**
- Estimate conditional average treatment effect:

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i,1) - g(x_i,0)$$

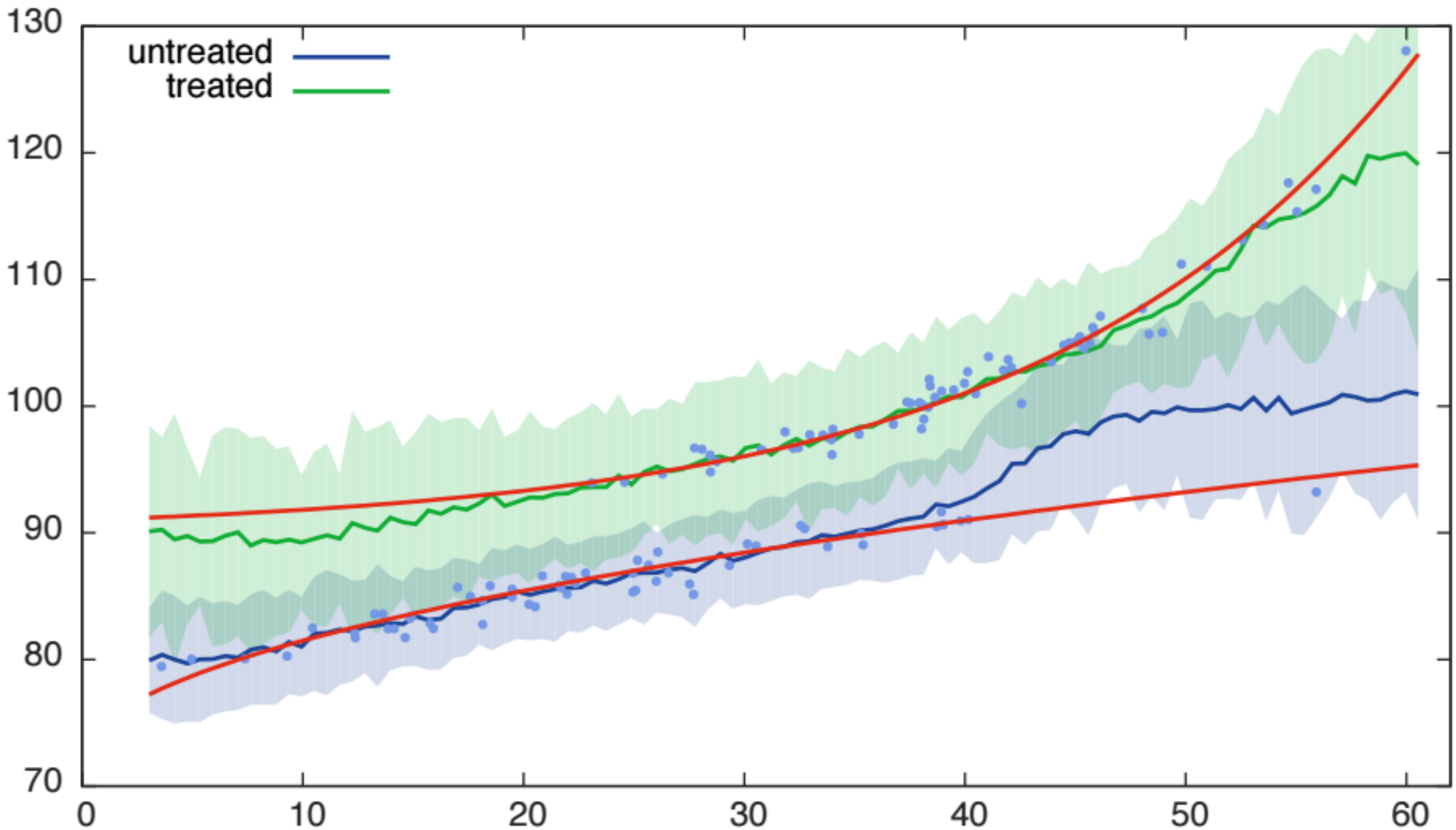
- Estimate  conditional average treatment effect *on the treated:*

$$\frac{1}{n_{treated}}\sum_{i:T_i=1}^{n} g(x_i,1) - g(x_i,0)$$

# BART use case – uncertainty intervals and significance testing

- Let's say we discovered that the conditional average treatment effect is 6, i.e. we estimate the treated population gained 6 IQ points because of the treatment.
- Is this effect significant? Can we trust it? Can we base expensive policy decisions on this results?
- Heady questions… partial answers
- First step: obtain confidence intervals for the effect
  - Use **permutation testing**: permute the treatment variable values between the units to obtain a null distribution of treatment effect, then calculate a p-value
  - Use many **posterior samples** to get uncertainty intervals for predictions

# Confidence intervals: an illustration



# Summary

- Causal inference as counterfactual inference, estimating treatment effect for non-treated and vice-versa
- Difficult in cases where treated and control are different
- One approach – learn a model relating the features, treatment, and outcome
- BART is a successful example of such a model
- Fitting BART by Gibbs and MH sampling