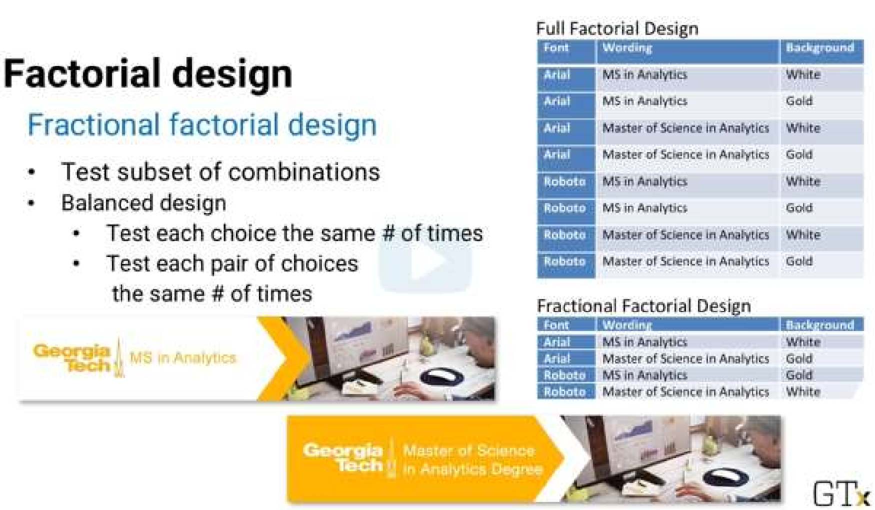
# Midterm 2 Prep

Monday, November 2, 2020 7:12 AM

By: Chris Dedow

- 11.2 Models for Variable Selection
    - Good for initial analysis, often don't perform as well on other data
        - Forward selection
        - Backward elimination
        - Stepwise Regression (most common out of three)
    - Slower but better prediction
        - Lasso - add constraint to standard regresssion equation (basically limit alotment for sum of coefficients
            - Some coefficients forced to 0 to simplify model
            - Scale data if you're going to be limiting coeffiecients (units of data will otherwise skew data)
            - If the data is not scaled, the coefficients can have artificially different orders of magnitude, which means they'll have unbalanced effects on the lasso constraint. At each step, the stepwise regression fits a different model. However, different lasso models can be found by varying T, and R has a function to automatically generate multiple lasso models.
        - Elastic Net - constrain absolute value of coefficients <u>and their squares</u> (underline is different than Lass)
            - Advantages:
                - Variable selection benefits of Lasso
                - Predictive benefits of Ridge Regression
            - Disadvantages:
                - Arbitrarily rules out some correlated variables like Lasso
                - Underestimates coefficients of very predictive variables like Ridge Regression
        - Ridge Regression - constrain squares of coefficients (doesn't do variable selection but can lead to better predictive models
            - Coefficients shrink toward 0 to reduce variance in estimate
            - Ridge regression will choose smaller (in an absolute sense) non-zero coefficients for both models. By nature, it may underestimate the effect of the factors.
- 11.3 Choosing a Variable Selection Model
- 12.1 Introduction to Design of Experiments
    - Which ad campaign to run? Which product to show?
    - Survey
    - Comparison and Control
    - Bloocking - something that could cause variation (sports cars more likely to be red)
- 12.2 A/B Testing
    - Choose between two alternatives
        - Collect data quickly
        - Data must be representative
        - Amount of data is small compared to the whole population
- 12.3 Factorial Designs
    - Full factorial design
        - 2 fonts * 2 wordings * 2 backgrounds = 8 combinations
        - ANOVA (analyiss of variance)
        - Can add up quickly if 7 factors each with 3 choices = 3^7 = 2,187 combinations
    - Fractional factorial design



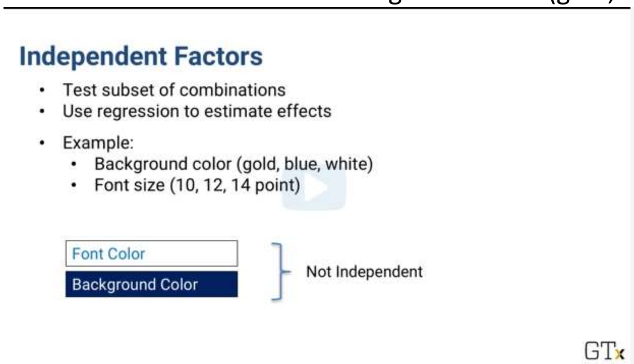

Full Factorial Design

| Font | Wording | Background |
|---|---|---|
| Arial | MS in Analytics | White |
| Arial | MS in Analytics | Gold |
| Arial | Master of Science in Analytics | White |
| Arial | Master of Science in Analytics | Gold |
| Roboto | MS in Analytics | White |
| Roboto | MS in Analytics | Gold |
| Roboto | Master of Science in Analytics | White |
| Roboto | Master of Science in Analytics | Gold |

Fractional Factorial Design

| Font | Wording | Background |
|---|---|---|
| Arial | MS in Analytics | White |
| Arial | Master of Science in Analytics | Gold |
| Roboto | MS in Analytics | Gold |
| Roboto | Master of Science in Analytics | White |

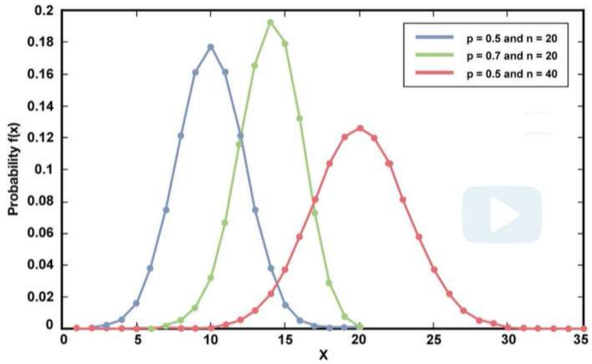
- Independent Factors
  - Test subset of combinations
  - Use regression to estimate effects
    - Ex: background color (gold, blue, white) and font size (10,12,14)




- Factorial design summary
  - Use before collecting data
  - Determine effects of factors
  - Full factorial design
    - Test all combinations
  - Partial factorial design
    - Estimate all effects by comparing some combinations

- 12.4 Multi-Armed Bandits
  - Exploration vs. Exploitation
    - Multi-armed bandit models use the best answer (exploitation) the more they're sure it's best. If the model is less sure what's best, it's more likely to concentrate on trying many options (exploration)
  - Start testing with k alternatives
    - Update probabilities with new information from testing
  - Help you learn faster on the fly and create more value along the way
- 13.1 Introduction to Advance Probability Distributions
- 13.2 Bernoulli, Binomial and Geometric Distribution
  - Bernoulli - probability mass Function P(X=1) = p

## Binomial Distribution



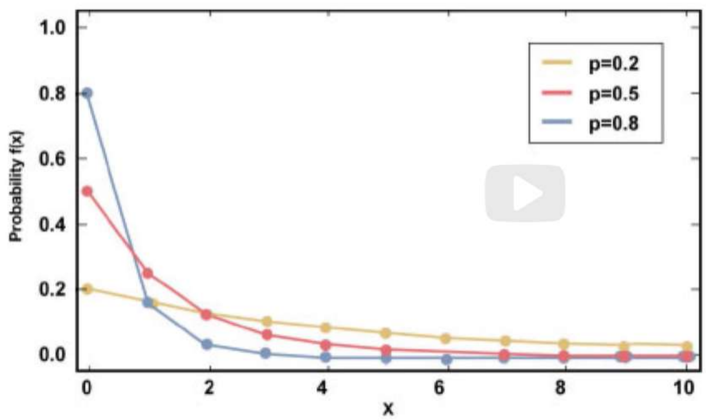

Probability Mass Function:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$
$$= \left(\frac{n!}{(n-x)!x!}\right) p^x (1-p)^{n-x}$$

- Probability of getting $x$ successes out of $n$ independent identically distributed Bernoulli(p) trials
- Large n
  - Binomial distribution converges to Normal distribution

GTx

## Geometric Distribution




Probability Mass Function:

$$P(X = x) = (1-p)^x p$$

- Probability of having $x$ Bernoulli(p) failures until first success?
  - Or, having x Bernoulli(1-p) successes until first failure

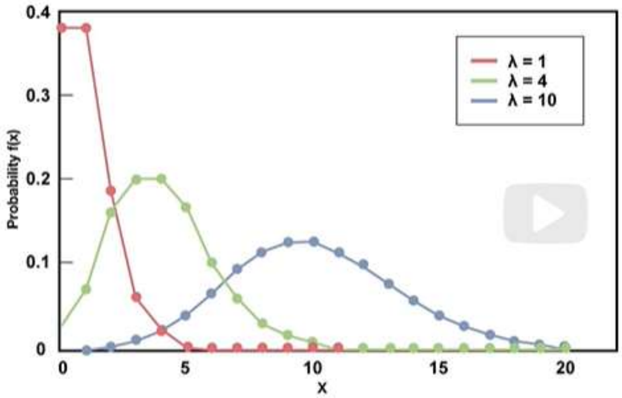
GTx

  - According to the geometric distribution, the probability of having 5 successful sales calls before the first unsuccessful call is (1−p)5p .
  - According to the geometric distribution, the probability of having 5 successful sales calls before the first unsuccessful call is p5(1−p) .
- 13.3 Poisson, Exponential and Weibull Distributions
  - Won't cover (do on your own)
    - Distribution's probability, mass index, density function, expectation, variance

## Poisson Distribution


Probability Mass Function:

# Poisson Distribution




Probability Mass Function:

$$f_x(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

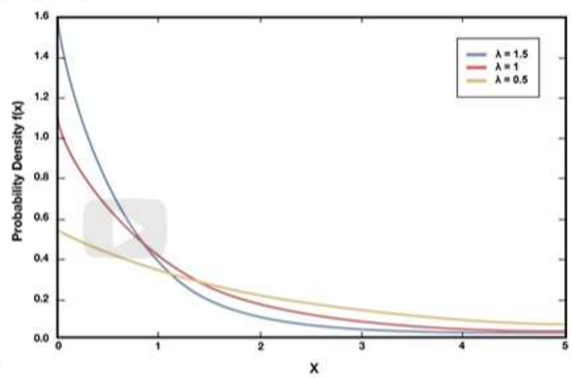
- Good at modeling random arrivals
- λ – Average number of arrivals / time period
- Arrivals are independent and identically distributed (i.i.d)

GTx

# Exponential Distribution

Probability Mass Function:

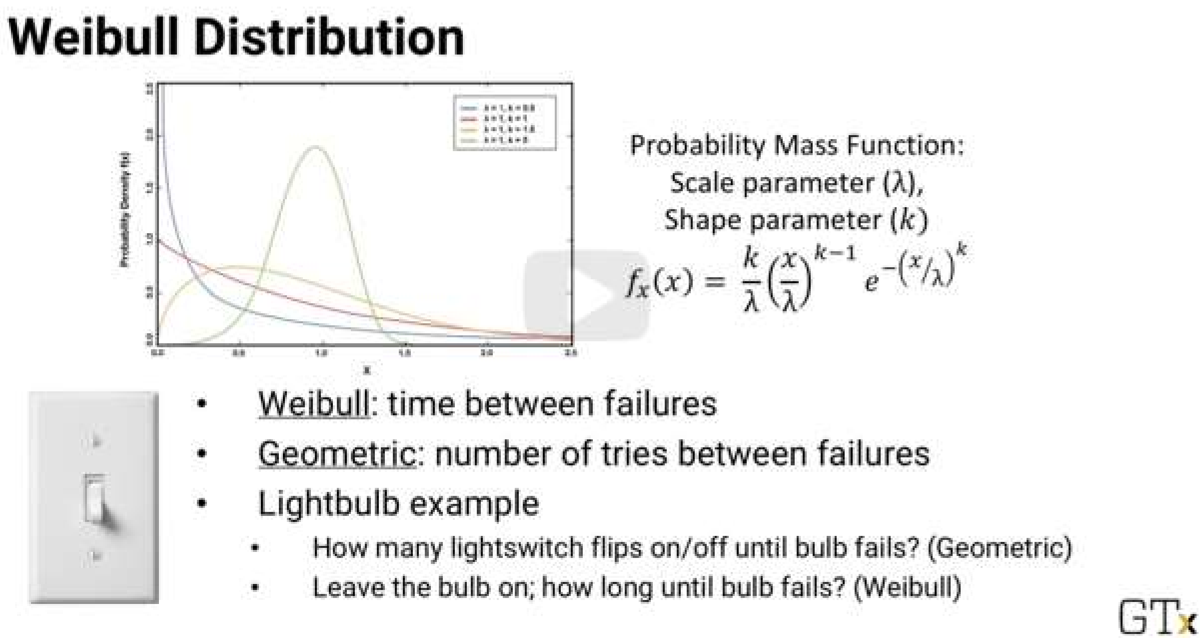
$$f_x(x) = \lambda e^{-\lambda x}$$

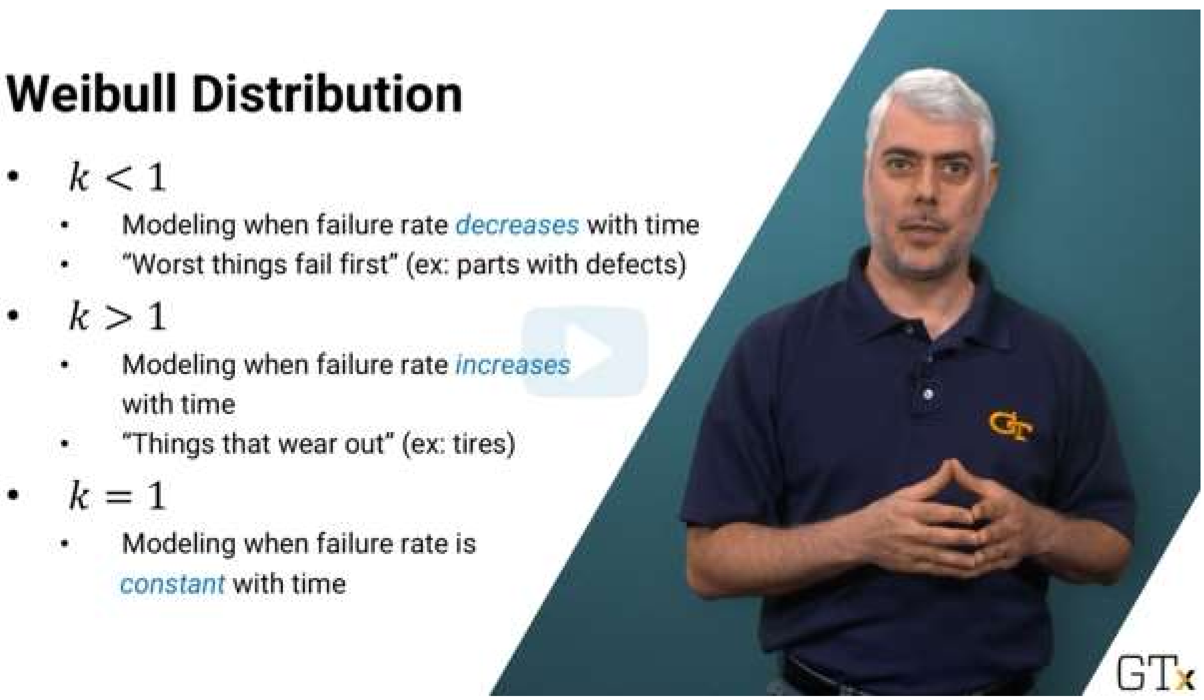



- Relation to Poisson
  - If arrivals are Poisson(λ), then time between successive arrivals is exponential(λ) distribution

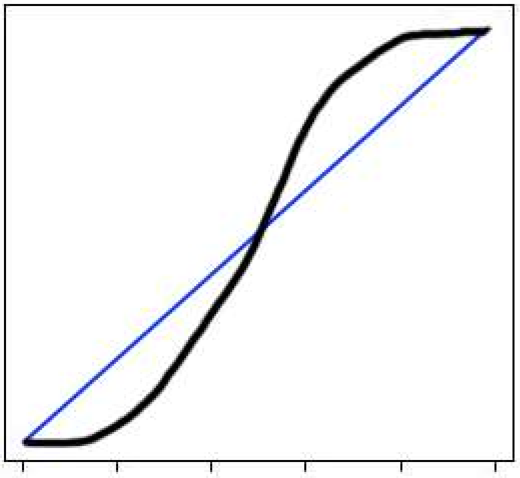
Poisson arrivals ⇔ exponential inter-arrival time

GTx



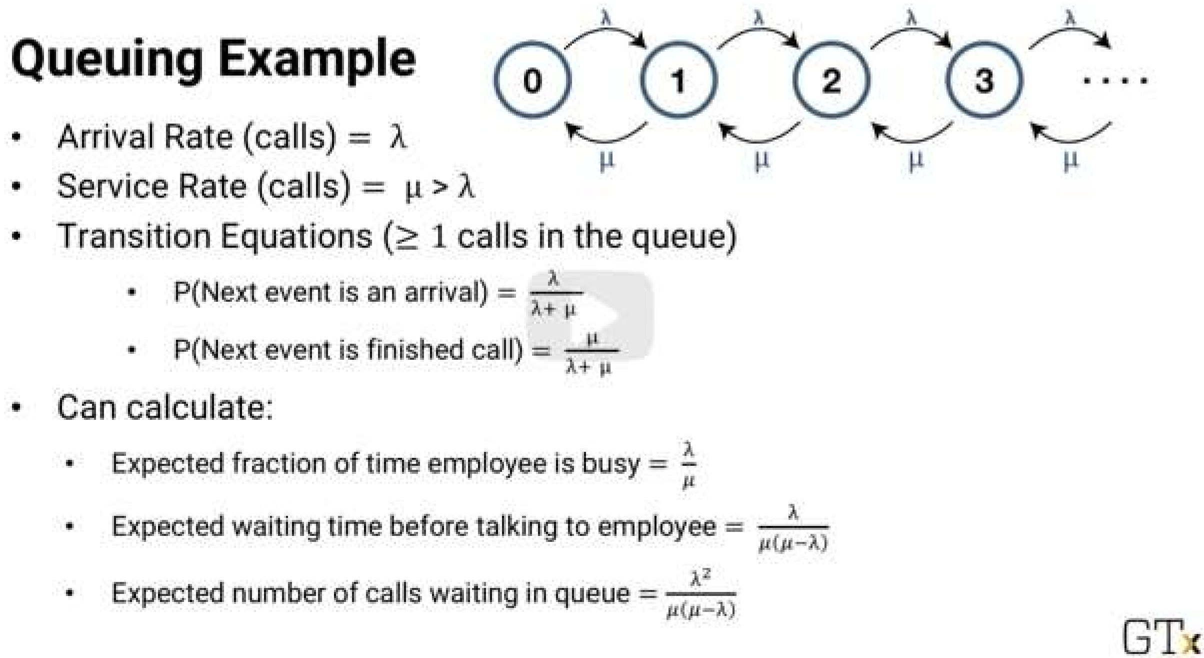




- 13.3 Poisson, Exponential and Weibull Distributions
    - Here's how these two are related. If arrivals are Poisson, with arrival rate lambda, then the time between arrivals, known as the inter-arrival time follows the Exponential distribution, with one over lambda as the average inter-arrival time.
    - If times between arrivals are exponentially distributed with average inter-arrival time one over lambda, then the arrivals follow the Poisson distribution with an average of lambda arrivals per unit time.
      The geometric distribution models how many tries it takes for something to happen, while the Weibull distribution models how long it takes.
- 13.4 QQ Plots



- Heavy tailed distribution

• 13.5 Queing



## Memoryless Property

- Memoryless exponential distribution
  - Distribution of *remaining* call time = *initial* distribution of call time
- Memoryless Poisson distribution
  - Distribution of time to next arrival = initial distribution of time to next arrival
  - Poisson interarrival times are exponentially distributed

- Data fits exponential distribution → memoryless
- Not memoryless → not exponential

GTx

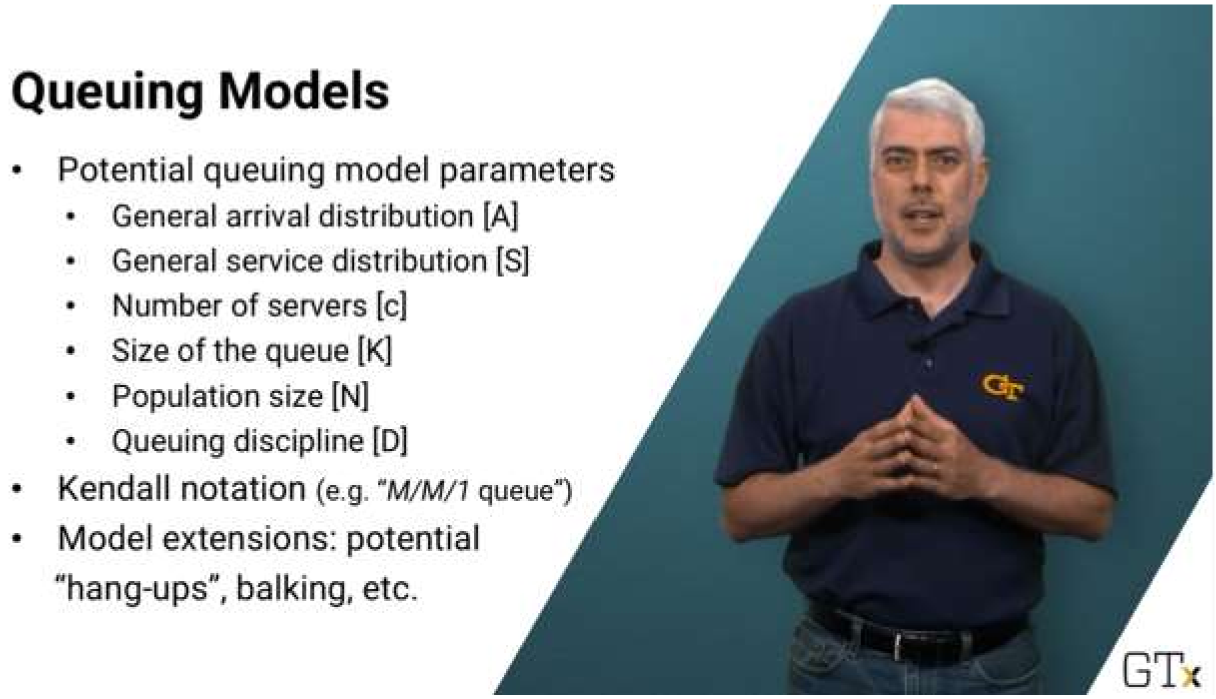
---

## Memoryless Property

- Law firm example
  - Should tire manufacturer pay damages for accident that happened when tire with 10,000 miles failed?

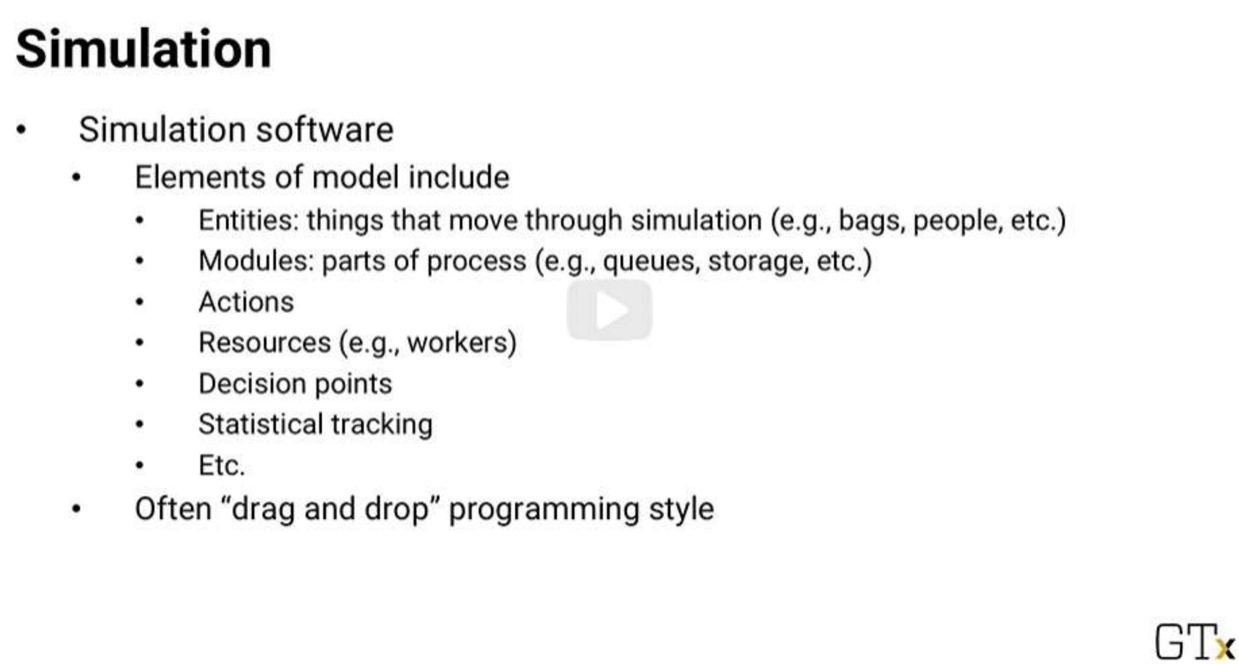
  - Probability(tire fails at 10,000 miles) = ?
    - Tires are more likely to fail, the more worn out they are
    - Not memoryless
    - Cannot model with the exponential distribution
    - Maybe try Weibull with k>1

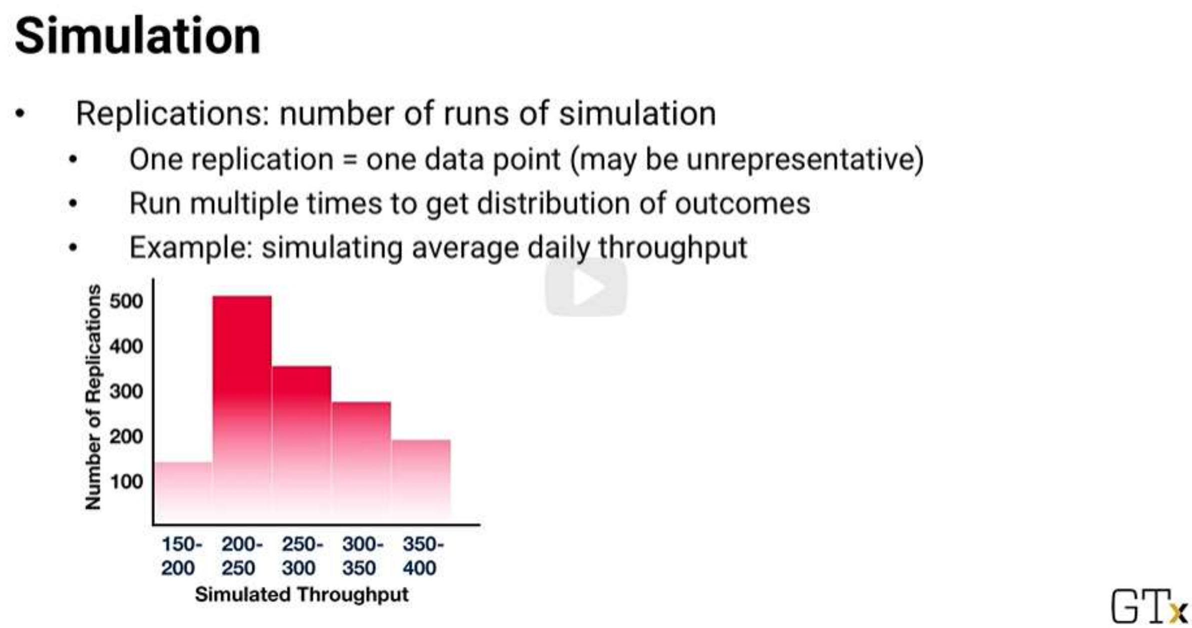
GTx

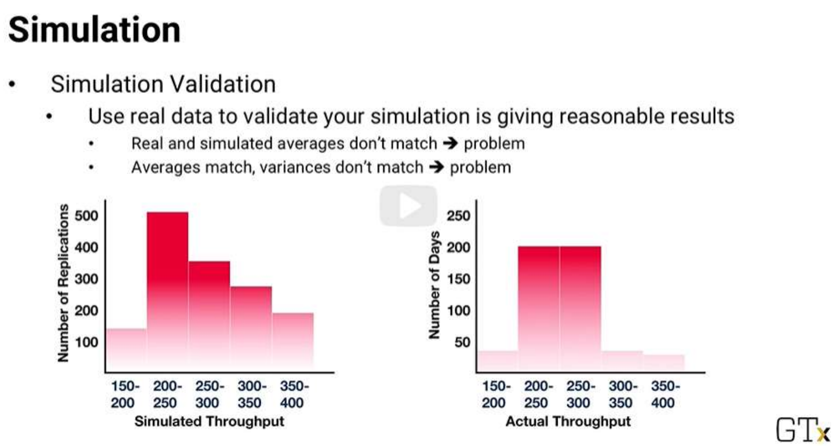



- Memoryless Property Distribution
- Distribution of remaining call time = initial distribution of call time

- 13.6 Simulation Basics
  - **Deterministic simulations** do not incorporate elements of randomness and so their output for given inputs does not vary.
  - **Stochastic simulations** incorporate randomness and thus might produce different outputs. They are generally more useful in analytics. Stochastic simulations must be replicated because one result may not be characteristic.
  - In **continuous-time simulations**, changes can happen continuously. In **discrete event simulations**, changes only happen at discrete time points when something happens.






- Stochastic simulation should be run many times because one random outcome might not be representative of system performance in the range of different situations that could arise
- A stochastic simulation is meant to show the performance of a system over a range of random events that could happen.
- If the simulation isn't a good reflection of reality, then any insights we gain from studying the simulation might not be applicable in reality.
- This is such an important point about simulation that I wanted to make sure you all clicked on it. I've overseen a lot of analytics projects that included simulation, and my experience is that it's easy to rely too much on simulated insights that might not be true in reality. It's critical to make sure that the simulation is a good-enough model of reality that insights from the simulation can effectively be transferred to reality

- 13.7 Prescriptive Simulation
- 13.8 Markov Chains

The next state of a 'memoryless' process doesn't depend on previous states, but it does depend on the current state

## Markov Chains

- Long-run probability of rainy days?
  - Calculate $(((\pi P)P)P)P \ldots = \pi P^{\infty}$ ?
  - Instead, use "steady state"
    - Apply P, and get initial vector back: $\pi^* P = \pi^*$
  - Solve for such a $\pi^*$ vector
    - $\pi^* P = \pi^*$, and $\sum_i \pi_i^* = 1$
  - $\pi^*$ might not always exist
    - Can't have cyclic behavior
    - Every state must be reachable from all others

| P | Sunny | Cloudy | Rainy |
|---|---|---|---|
| Sunny | .75 | .15 | .10 |
| Cloudy | .20 | .40 | .40 |
| Rainy | .40 | .30 | .30 |

GTx

- 14.1 Missing Data
- 14.2 Meothods That Do Not Require Imputation
  - Two different ways dealing with missing data
    - Remove
    - Add categorical variables
  - Avoide estimating what missing data may be

## Categorical variable approach

### Quantitative variable has missing data

- All missing values = 0
- New categorical variable
  - Missing data could be biased
  - Include interactions

| ID# | Height | Missing? | # kids | #kids,missing |
|---|---|---|---|---|
| 1 | 5'6" | No | 2 | 2 |
| 2 | 6'1" | No | 3 | 3 |
| 3 | 5'11" | No | 5 | 5 |
| 4 | 0 | Yes | 4 | 0 |
| 5 | 5'10" | No | 1 | 1 |
| 6 | 0 | Yes | 5 | 0 |

**Split into two models**

Height data

| ID# | Height | # kids |
|---|---|---|
| 1 | 5'6" | 2 |
| 2 | 6'1" | 3 |
| 3 | 5'11" | 5 |
| 5 | 5'10" | 1 |

No height data

| ID# | # kids |
|---|---|
| 4 | 4 |
| 6 | 5 |

GTx

- 14.3 Imputation Methods
  - Imputation - estimating missing values
    - Mean, median (numeric) or mode (categorical)
      - Advantage
        - Hedge against being too wrong
        - Easy to compute
      - Biased imputation
        - Ex: people with high income less likely to answer survey
        - Mean/median will underestimate the missing value
    - Second approach: regression

- True! It's usually not recommended to impute more than 5% of values, and advanced methods like multivariate imputation by chained equations (MICE) can impute multiple factor values together

## Imputation Approaches

- Second approach: regression
  - Reduce or eliminate the problem of bias
  - Income example
    - Factors: zip code, profession, number of cars owned, etc.
  - Disadvantages
    - Complex: build, fit, validate, test to estimate missing value
    - Does not capture all the variability

| Income | Zip code | Profession | Cars Owned |
|---|---|---|---|
| $15,000 | 24001 | Fast food server | 0 |
| $31,000 | 24330 | Telemarketer | 1 |
| $59,000 | 17330 | Executive Assistant | 1 |
| $98,000 | 24335 | Plumber | 1 |
| $120,000 | 30228 | Car Salesman | 2 |
| $240,000 | 34509 | Financial Advisor | 3 |
| $403,000 | 67767 | Dentist | 4 |

GTx

## Imputation with variability

- Imputation does not capture all variability
  - Even with regression model
- Impute with added variability
  - Add perturbation to each imputed value
    - Ex: normally-distributed variation
    - Less accurate on average
    - More-accurate variability

| More accurate on average<br>Less accurate variability | Less accurate on average<br>More accurate variability |
|---|---|
| No perturbation | Perturbation distributed according to actual variability |

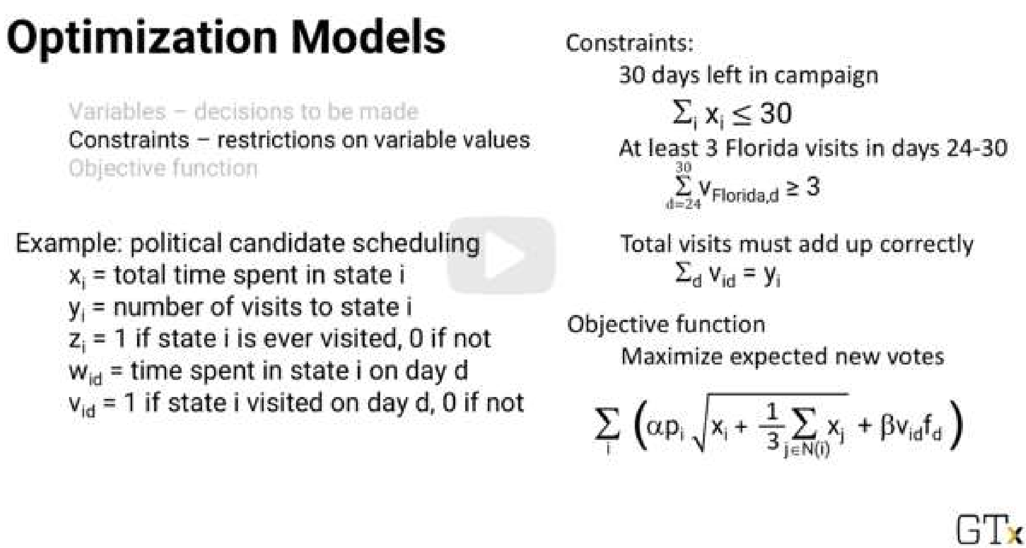
GTx

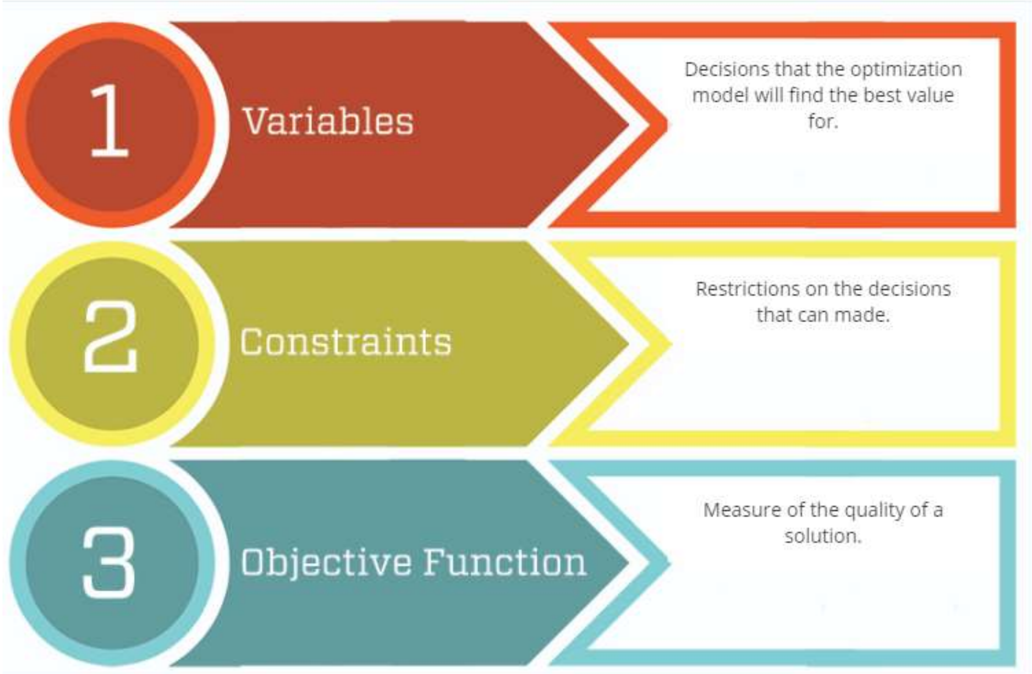
- 15.1 Indtroduction to Optimization
  - Regression classification, etc.
    - Software automates solution and model building
  - Optimization
    - Sotware automates solution
    - Model building is up to you!
  - Statistical software can both build and solve regression models. Optimization software only solves models; human experts are required to build optimization models.

## Optimization for prescriptive analytics (examples)

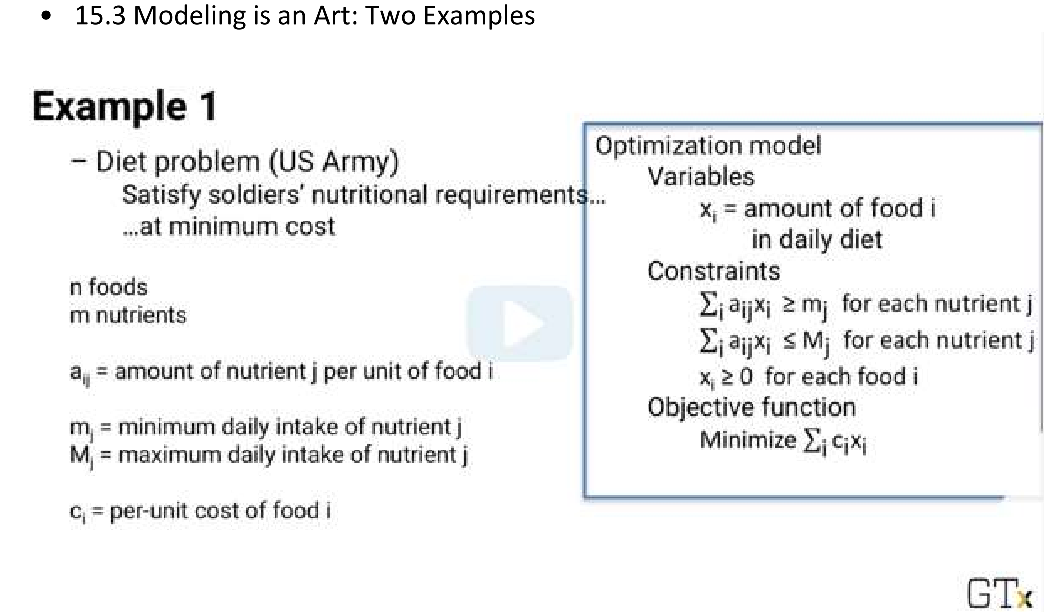
- Airplane mechanic scheduling
- Crude oil shipment planning

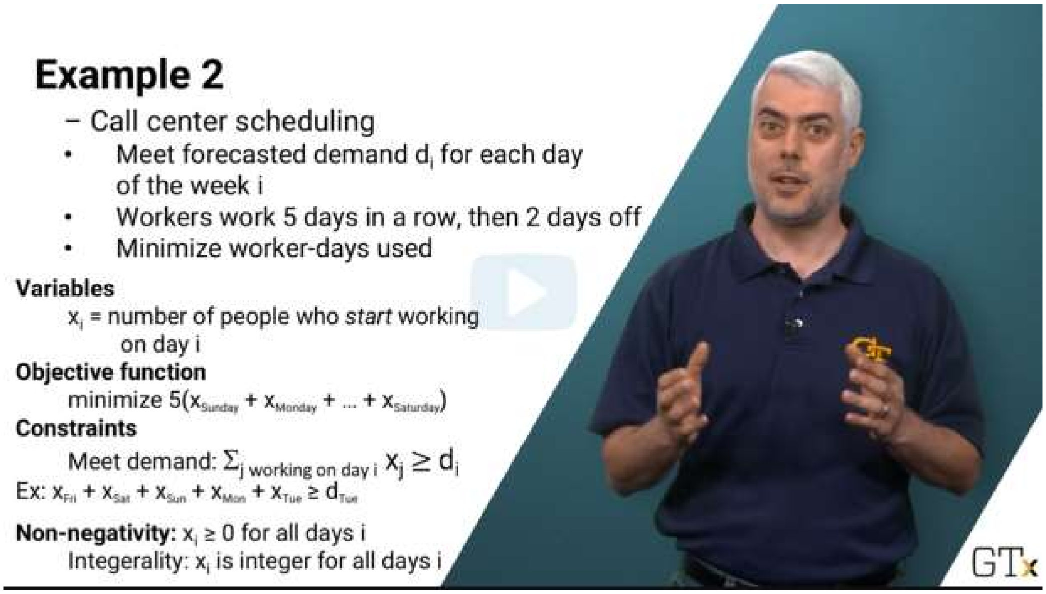

- 15.2 Elements of Optimization Models
  - Variables - decisions to be made
  - Constraints - restrictions on variable values
  - Objectives function - solution quality measure
  - Solution - values for each variables
  - Feasible solution - variable values that satisfy all constraints
  - Optimal solution - feasible solution with the best objective value






- 15.3 Modeling is an Art: Two Examples






- 15.4 Modeling with Binary Variables

# Example: Stock Market Investment

Invest to balance return and risk

B = investment budget
n = number of stocks available
$r_i$ = expected return of stock i relative to market
$Q_{ij}$ = covariance of returns of stocks i and j

Linking constraints

Transaction fees
Fixed charge
Could be $t_i$ if vary by stock
Can be weighted as one-time expense

Variables
$x_i$ = amount invested in stock i
$y_i$ = 1 if invest in stock i, 0 if not

Constraints
$\sum_i x_i \le B$
$x_i \ge 0$ for all stocks i
$x_i \le By_i$ for all stocks i

$y_i = 0$ means $x_i \le 0$ (and $x_i \ge 0$ forces $x_i = 0$)
$y_i = 1$ means $x_i \le B$

Objective function
Maximize $\sum_i r_i x_i - \theta \sum_i \sum_j Q_{ij} x_i x_j - \sum_i t y_i$

GTx

# Example: Stock Market Investment

Invest to balance return and risk

B = investment budget
n = number of stocks available
$r_i$ = expected return of stock i relative to market
$Q_{ij}$ = covariance of returns of stocks i and j
$m_i$ = minimum dollar amount for each stock

Personal constraints (examples)
Invest in Tesla
Invest in Amazon, Google, or Apple
Invest in both or neither of FedEx, UPS
Invest in exactly one of Coca-Cola and Pepsico

Variables
$x_i$ = amount invested in stock i
$y_i$ = 1 if invest in stock i, 0 if not

Constraints
$\sum_i x_i \le B$
$x_i \ge 0$ for all stocks i
$x_i \le By_i$ for all stocks i
$x_i \ge m_i y_i$ for all stocks i
$y_{Tesla} = 1$
$y_{Amazon} + y_{Google} + y_{Apple} \ge 1$
$y_{FedEx} = y_{UPS}$
$y_{Coca\text{-}Cola} = 1 - y_{PepsiCo}$

Objective function
Maximize $\sum_i r_i x_i - \theta \sum_i \sum_j Q_{ij} x_i x_j - \sum_i t y_i$

GTx

# Example: Stock Market Investment

Invest to balance return and risk

B = investment budget
n = number of stocks available
$r_i$ = expected return of stock i relative to market
$Q_{ij}$ = covariance of returns of stocks i and j
$m_i$ = minimum dollar amount for each stock

Personal constraints (more examples)
If invest in energy, invest in at least 5

Option 1: $\sum_{j \in energy} y_j \ge 5 y_i$ for all energy stocks i

Variables
$x_i$ = amount invested in stock i
$y_i$ = 1 if invest in stock i, 0 if not

Constraints
$\sum_i x_i \le B$
$x_i \ge 0$ for all stocks i
$x_i \le By_i$ for all stocks i
$x_i \ge m_i y_i$ for all stocks i
$y_{Tesla} = 1$
$y_{Amazon} + y_{Google} + y_{Apple} \ge 1$
$y_{FedEx} = y_{UPS}$
$y_{Coca\text{-}Cola} = 1 - y_{PepsiCo}$

Objective function
Maximize $\sum_i r_i x_i - \theta \sum_i \sum_j Q_{ij} x_i x_j - \sum_i t y_i$

GTx

## Example: Stock Market Investment

Invest to balance return and risk

B = investment budget
n = number of stocks available
$r_i$ = expected return of stock i relative to market
$Q_{ij}$ = covariance of returns of stocks i and j
$m_i$ = minimum dollar amount for each stock

Personal constraints (more examples)
If invest in energy, invest in at least 5

Option 2: $z_{Energy}$ = 1 if invest in energy, 0 if not

$$\sum_{j \in \text{energy}} y_j \ge 5 z_{Energy}$$

$z_{Energy} \ge y_i$ for all energy stocks i

Variables
- $x_i$ = amount invested in stock i
- $y_i$ = 1 if invest in stock i, 0 if not

Constraints
- $\sum_i x_i \le B$
- $x_i \ge 0$ for all stocks i
- $x_i \le By_i$ for all stocks i
- $x_i \ge m_i y_i$ for all stocks i
- $y_{Tesla} = 1$
- $y_{Amazon} + y_{Google} + y_{Apple} \ge 1$
- $y_{FedEx} = y_{UPS}$
- $y_{Coca\text{-}Cola} = 1 - y_{PepsiCo}$

Objective function
- Maximize $\sum_i r_i x_i - \theta \sum_i \sum_j Q_{ij} x_i x_j - \sum_i t y_i$

GTx

## Integer variables in optimization

- Fixed charges in objective function
- Constraints to choose among options
- Constraints requiring same/opposite decisions
- If-then constraints

Optimization modeling is an art!

GTx

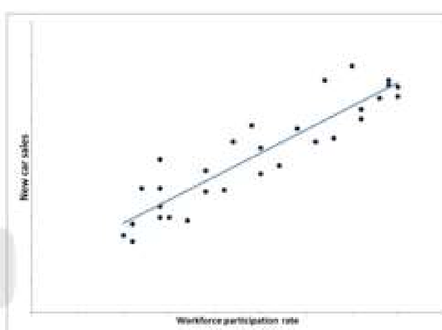
- 15.5 Optimization for Statistical Models

## Linear Regression Model

Variables
- $a_0, a_1 \dots a_m$

Constraints
- none

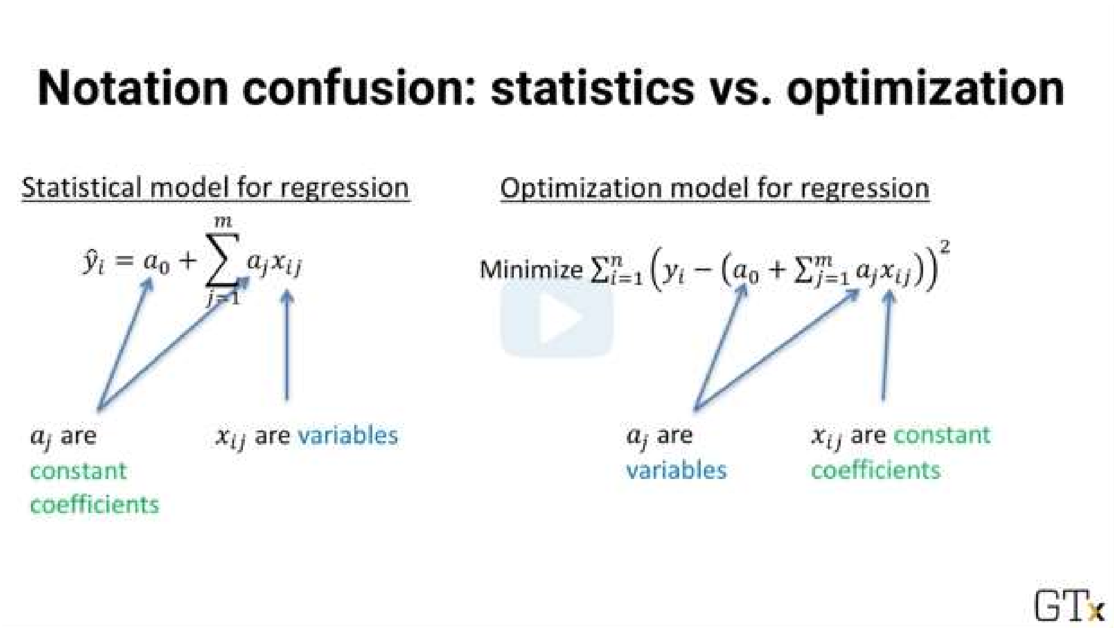
Objective function
- Minimize $\sum_{i=1}^{n} \left(y_i - \left(a_0 + \sum_{j=1}^{m} a_j x_{ij}\right)\right)^2$

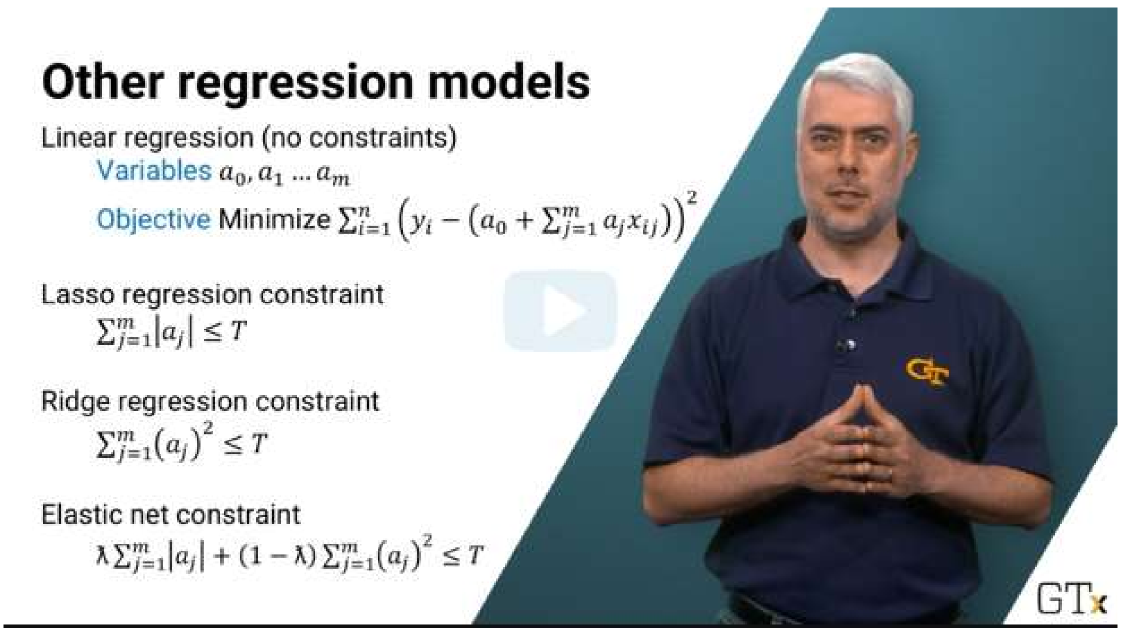



Given $n$ data points
- $x_{ij}$ = $j$th factor for data point $i$
- $y_i$ = response for data point $i$

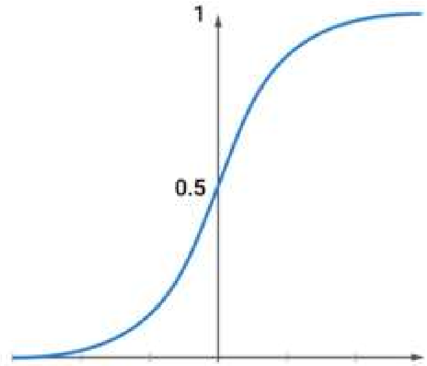
Find coefficients $a_0, a_1 \dots a_m$ to best fit data

GTx


Notation confusion: statistics vs. optimization
Statistical model for regression
$\hat{y}_i = a_0 + \sum_{j=1}^{m} a_j x_{ij}$
$a_j$ are constant coefficients
$x_{ij}$ are variables
Optimization model for regression
Minimize $\sum_{i=1}^{n} \left( y_i - \left( a_0 + \sum_{j=1}^{m} a_j x_{ij} \right) \right)^2$
$a_j$ are variables
$x_{ij}$ are constant coefficients
GTx


Other regression models
Linear regression (no constraints)
Variables $a_0, a_1 \ldots a_m$
Objective Minimize $\sum_{i=1}^{n} \left( y_i - \left( a_0 + \sum_{j=1}^{m} a_j x_{ij} \right) \right)^2$
Lasso regression constraint
$\sum_{j=1}^{m} |a_j| \leq T$
Ridge regression constraint
$\sum_{j=1}^{m} (a_j)^2 \leq T$
Elastic net constraint
$\lambda \sum_{j=1}^{m} |a_j| + (1 - \lambda) \sum_{j=1}^{m} (a_j)^2 \leq T$
GTx

## Logistic Regression




Variables
- $a_0, a_1 \dots a_m$

Constraints
- none

Objective function
- Maximize

$$\prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} \left(1 - p(x_i)\right)$$

where $p(x_i) = \frac{1}{1+e^{-\left(a_0+\sum_{j=1}^{m} a_j x_{ij}\right)}}$

Given $n$ data points
- $x_{ij}$ = $j$th factor for data point $i$
- $y_i$ = response for data point $i$

Find coefficients $a_0, a_1 \dots a_m$ to best fit data

GTx

## Support vector machine models

Hard classification

Variables
- $a_0, a_1 \dots a_m$

Constraints
- $\left(a_0 + \sum_{j=1}^{m} a_j x_{ij}\right) y_i \geq 1$ for each $i$

Objective function
- Maximize $\sum_{j=1}^{m} \left(a_j\right)^2$

Soft classification

Variables
- $a_0, a_1 \dots a_m$

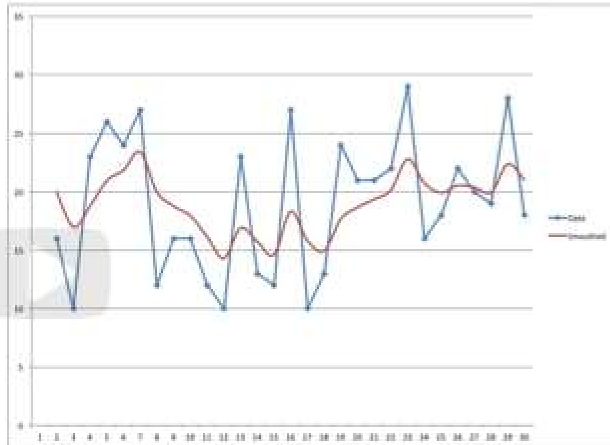
Constraints
- none

Objective function
- Minimize

$\sum_{i=1}^{n} \max\{0, 1 - (a_0 +$
$\sum_{j=1}^{m} a_j x_{ij}) y_i\} + \lambda \sum_{j=1}^{m} \left(a_j\right)^2$

GTx

# Time series models



Exponential smoothing
Variables
- $\alpha, \beta, \gamma$

Constraints
- $0 \leq \alpha \leq 1$
- $0 \leq \beta \leq 1$
- $0 \leq \gamma \leq 1$

Objective function
- Minimize

$$\sum_{t=1}^{n} (x_t - \hat{x}_t)^2$$

# Time series models

ARIMA
Variables
- $\mu, \varphi_i, \theta_i$

Constraints
- none

Objective function
- Minimize

$$\sum_{t=1}^{n} (x_t - \hat{x}_t)^2$$

where

$$D_{(d)t} = \mu + \sum_{i=1}^{p} \alpha_i D_{(d)t-i} - \sum_{i=1}^{q} \theta_i (\hat{x}_{t-i} - x_{t-i})$$

GARCH
Variables
- $\omega, \beta_i, \gamma_i$

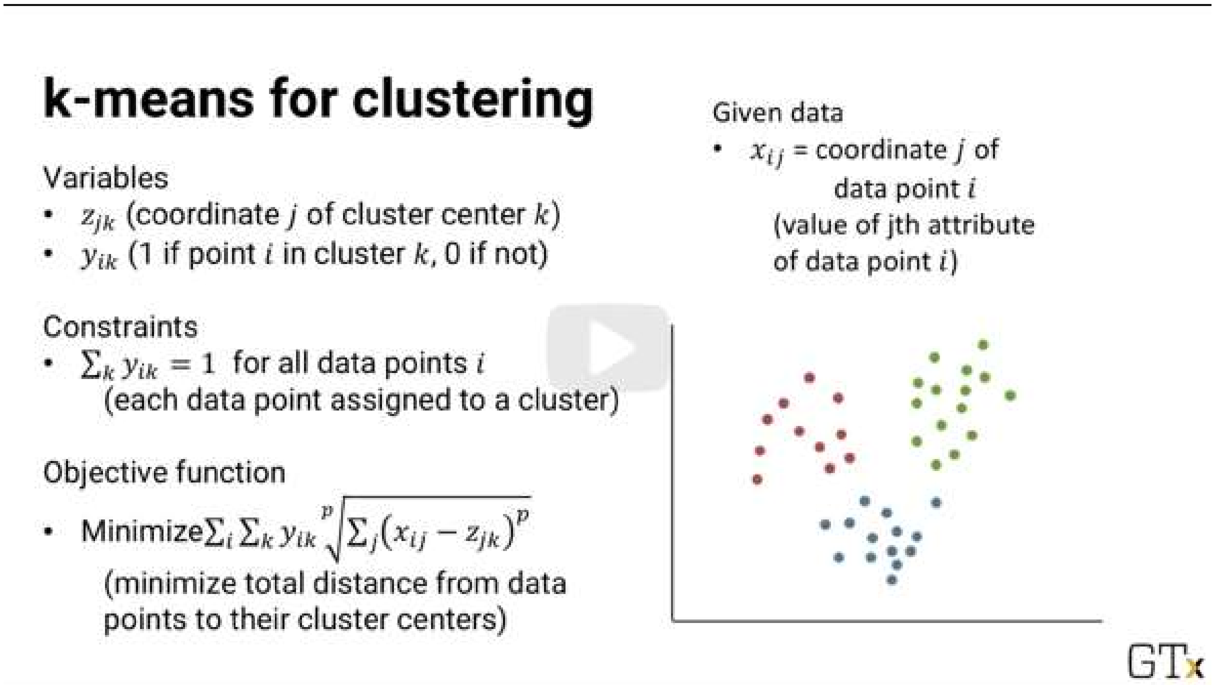
Constraints
- none

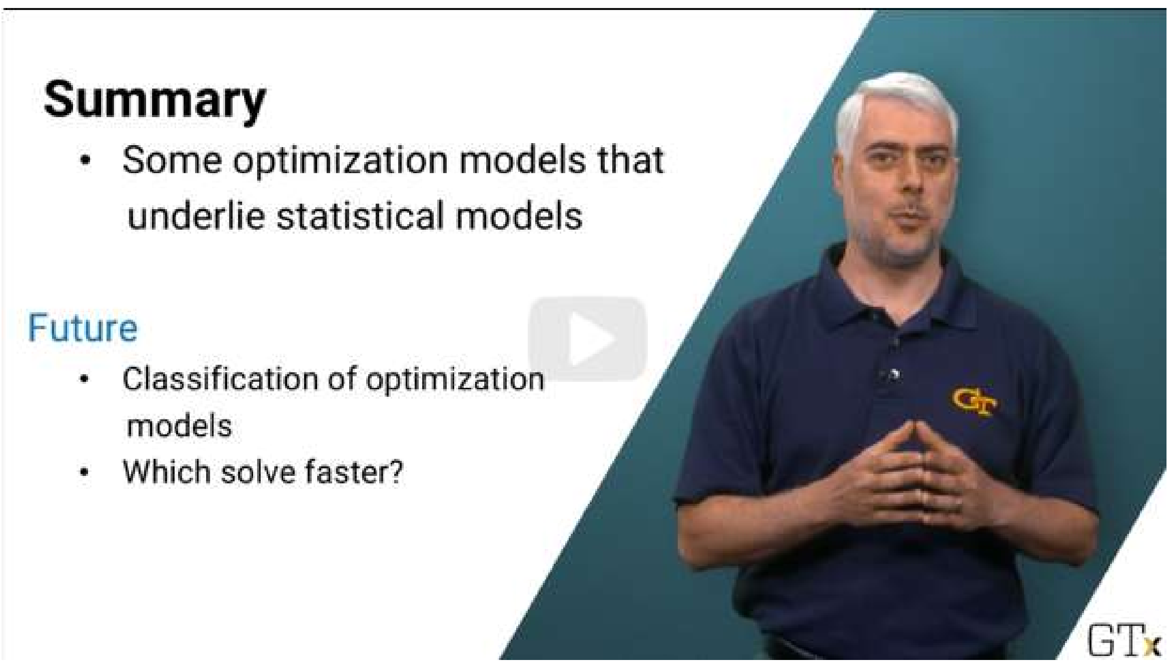
Objective function
- Minimize

$$\sum_{t=1}^{n} (\sigma_t^2 - \hat{\sigma}_t^2)^2$$

where

$$\hat{\sigma}_t^2 = \omega + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 - \sum_{i=1}^{p} \gamma_i \epsilon_{t-i}^2$$




- All of the following have optimization models to find best fit:
  - Linear regression
  - Logistic regression
  - Lasso regression
  - Exponential smoothing
  - k-means clustering

- 15.6 Classification of Optimization Models

## Optimization Models

$x$ = vector of variables

Minimize or maximize objective function
  Minimize $f(x)$ or Maximize $f(x)$

Variables $x$ must belong to a set $X$
  Subject to $x \in X$

GTx

## Linear program

- $f(x)$ is a linear function

  Minimize or Maximize $C + \sum_{i=1}^{n} c_i x_i$

- Constraint set $X$ is defined by linear equations and inequalities

  $\sum_{i=1}^{n} a_{ij} x_i \le b_j$ or
  $\sum_{i=1}^{n} a_{ij} x_i = b_j$ or
  $\sum_{i=1}^{n} a_{ij} x_i \ge b_j$
  for each constraint $j$

Easy/fast to solve, even for very large instances

GTx

# Convex quadratic program

- $f(x)$ is a convex quadratic function

  Minimize $f(x)$ or Maximize $-f(x)$

- Constraint set $X$ is
  defined by linear equations and
  inequalities

Easy/fast to solve, but
not as quickly as linear programs

# Optimization Model

Convex Optimization Program

- Objective function $f(x)$ is
  - Concave (if maximizing)
  - Convex (if minimizing)
- Constraint set $X$ is a convex set
- Easy to solve, but solutions can take
  a lot longer

# Optimization Model

**Convex Optimization Program**

- Objective function $f(x)$ is
  - Concave (if maximizing)
  - Convex (if minimizing)
- Constraint set $X$ is a convex set
- Easy to solve, but solutions can take a lot longer

**Integer Program**

- Linear program, plus
- Some (or all) variables restricted to take only integer values
  - Variables could be binary
    - Either 0 or 1
- More difficult to solve even with good software packages

**General Non-Convex Program**

- Optimization problem is not convex
  - Hard to find optimal solutions

GTx

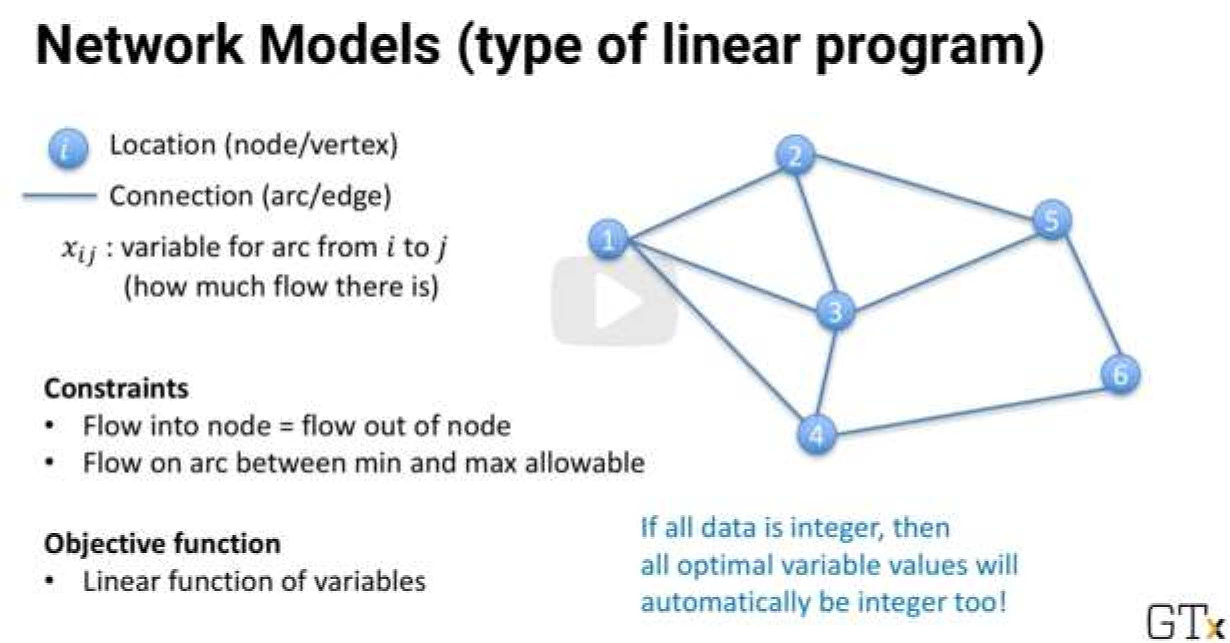
# Optimization Problem Types

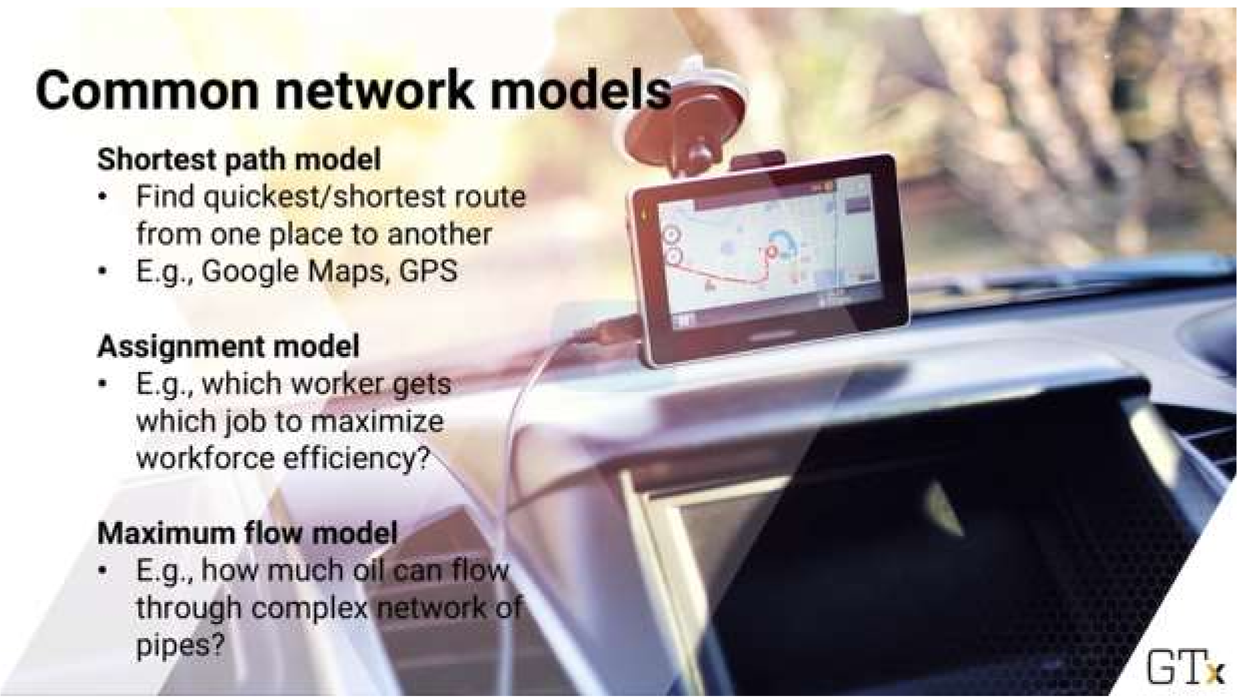
**From quickest/easiest to slowest/hardest**

1. Linear Programs
2. Convex Quadratic Programs
3. Convex Programs
4. Integer Programs
5. General Non-Convex Programs

**What if our problem is too hard?**

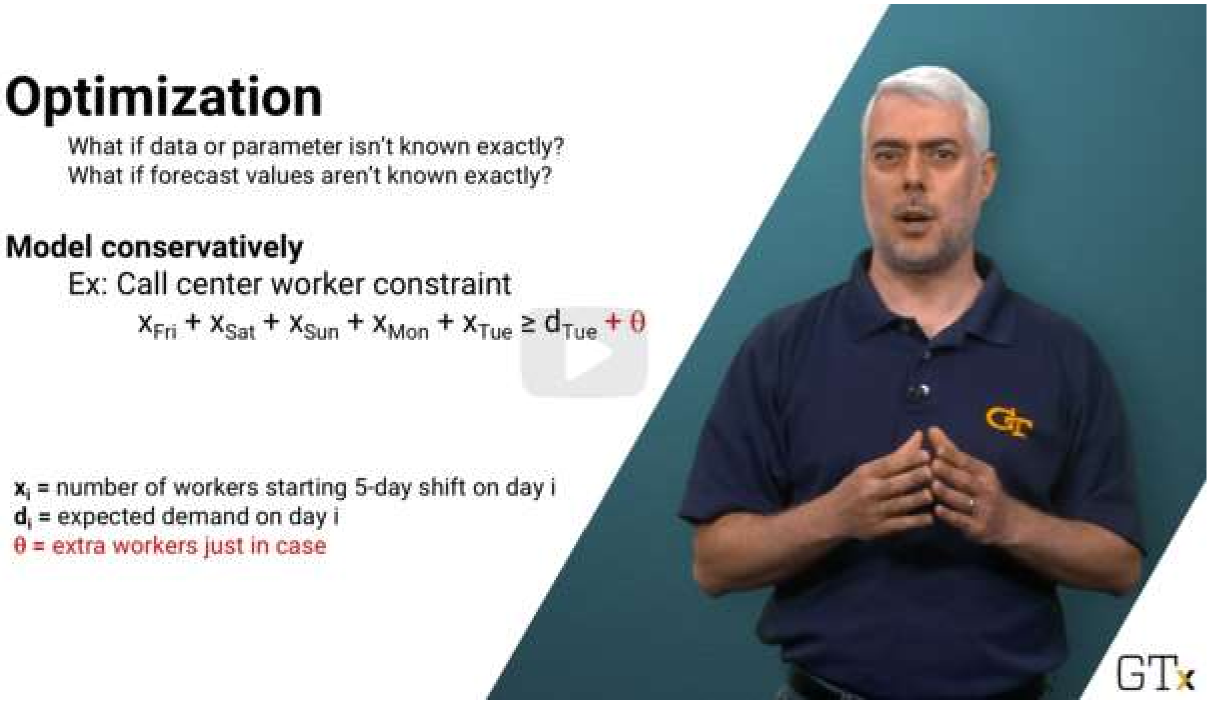
- Heuristic
  - Rule-of-thumb process
  - Usually gives good solutions

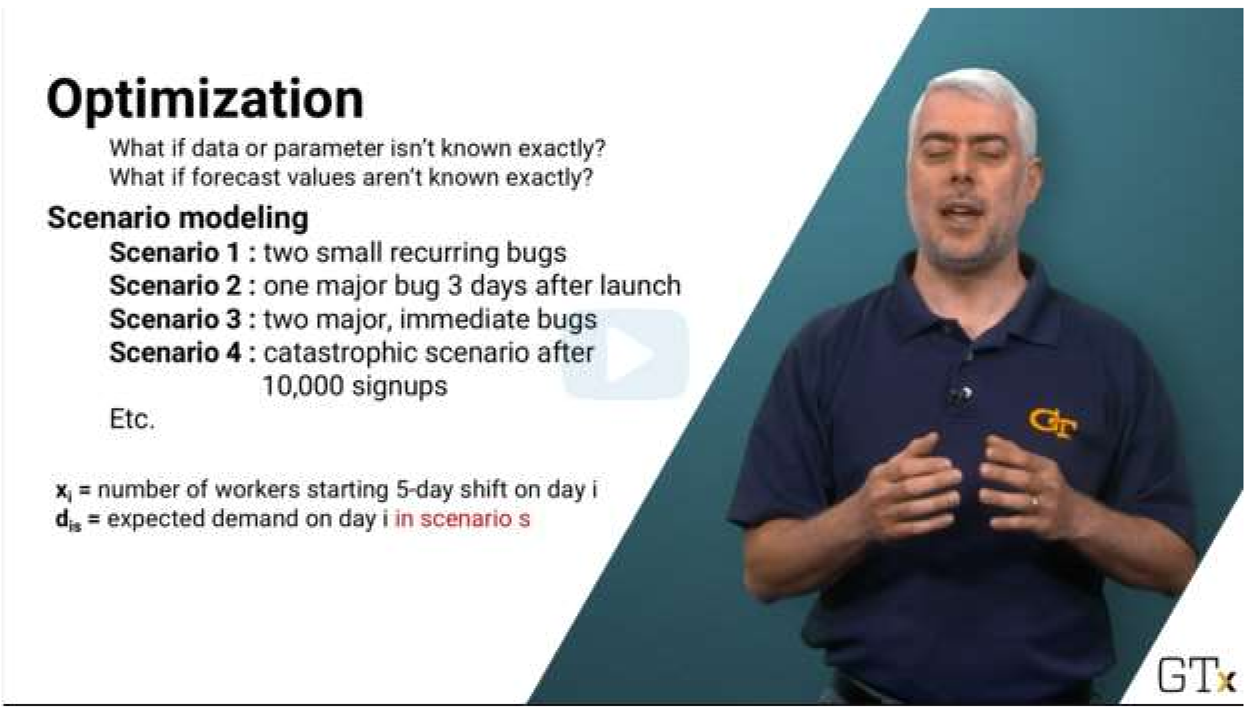



- ○ Adding integer variables moves the model from a linear program, which usually solves very quickly, to an integer program, which sometimes takes a long time to solve.

- 15.7 Stocastic Optimization


Optimization
What if data or parameter isn't known exactly?
What if forecast values aren't known exactly?
Model conservatively
Ex: Call center worker constraint
$x_{Fri} + x_{Sat} + x_{Sun} + x_{Mon} + x_{Tue} \geq d_{Tue} + \theta$
$x_i$ = number of workers starting 5-day shift on day i
$d_i$ = expected demand on day i
$\theta$ = extra workers just in case
GTx


Optimization
What if data or parameter isn't known exactly?
What if forecast values aren't known exactly?
Scenario modeling
Scenario 1 : two small recurring bugs
Scenario 2 : one major bug 3 days after launch
Scenario 3 : two major, immediate bugs
Scenario 4 : catastrophic scenario after 10,000 signups
Etc.
$x_i$ = number of workers starting 5-day shift on day i
$d_{is}$ = expected demand on day i in scenario s
GTx

# Optimization

What if data or parameter isn't known exactly?
What if forecast values aren't known exactly?

## Scenario modeling

**Scenario 1 :** two small recurring bugs
**Scenario 2 :** one major bug 3 days after launch
**Scenario 3 :** two major, immediate bugs
**Scenario 4 :** catastrophic scenario after 10,000 signups

## Robust solution

$$\left.\begin{aligned}
x_{Fri} + x_{Sat} + x_{Sun} + x_{Mon} + x_{Tue} &\ge d_{Tue,1}\\
x_{Fri} + x_{Sat} + x_{Sun} + x_{Mon} + x_{Tue} &\ge d_{Tue,2}\\
x_{Fri} + x_{Sat} + x_{Sun} + x_{Mon} + x_{Tue} &\ge d_{Tue,3}\\
x_{Fri} + x_{Sat} + x_{Sun} + x_{Mon} + x_{Tue} &\ge d_{Tue,4}
\end{aligned}\right\}\ \text{Satisfy } \textit{all} \text{ scenario demands}$$

**$x_i$** = number of workers starting 5-day shift on day i
**$d_{is}$** = expected demand on day i in scenario s

---

# Optimization

What if data or parameter isn't known exactly?
What if forecast values aren't known exactly?

## Scenario modeling

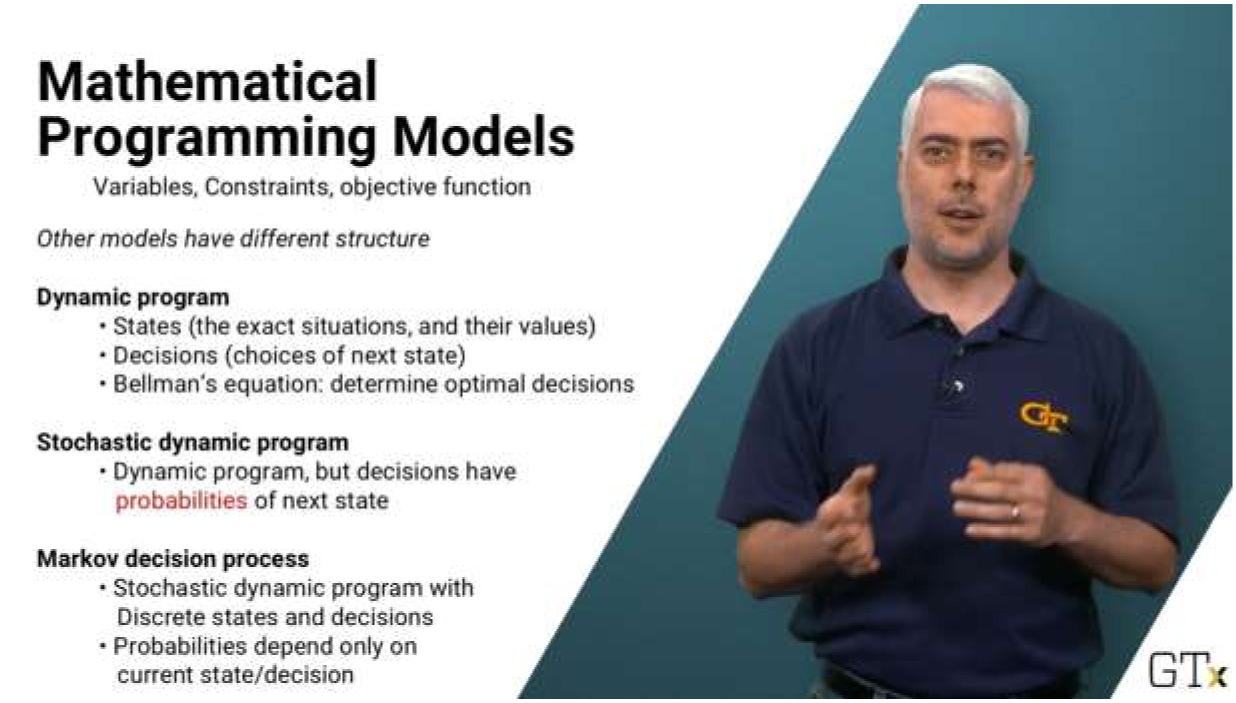
**Scenario 1 :** two small recurring bugs
**Scenario 2 :** one major bug 3 days after launch
**Scenario 3 :** two major, immediate bugs
**Scenario 4 :** catastrophic scenario after 10,000 signups

## Optimize expected cost

minimize $5(x_{Sun} + x_{Mon} + \dots + x_{Sat}) + \sum_s c(p_{Sun,s}y_{Sun,s} + p_{Mon,s}y_{Mon,s} + \dots + p_{Sat,s}y_{Sat,s})$

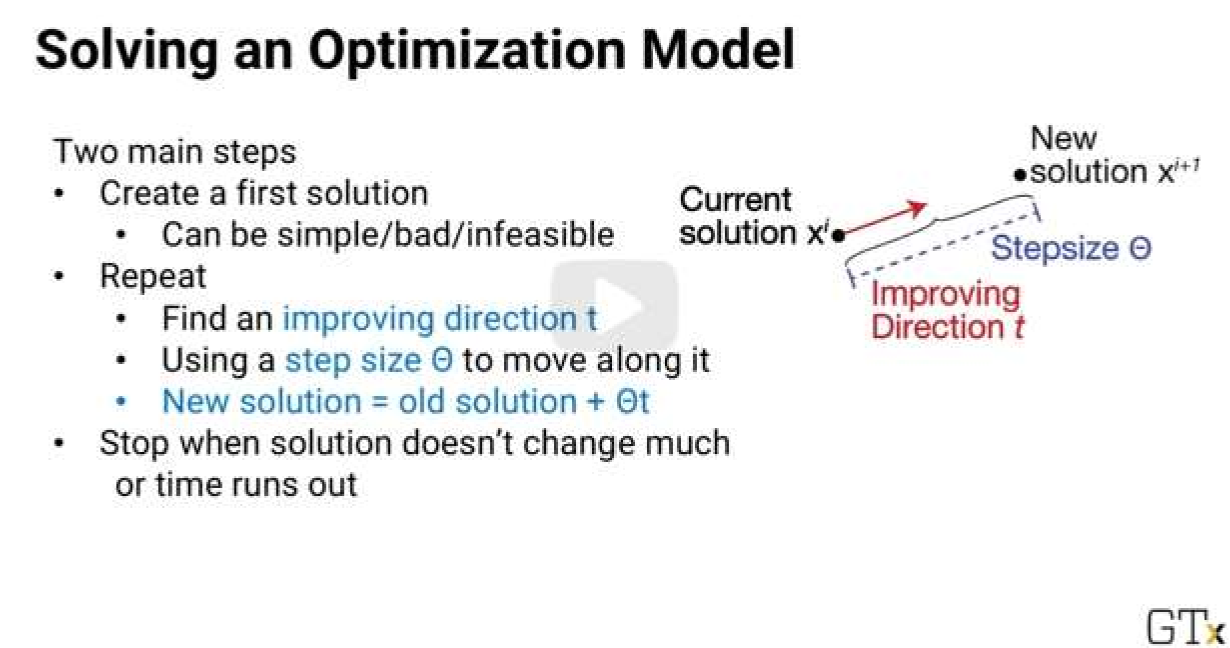
Example constraints: $x_{Fri} + x_{Sat} + x_{Sun} + x_{Mon} + x_{Tue} + y_{Tue,1} \ge d_{Tue,1}$ and $y_{Tue,1} \ge 0$

**$x_i$** = number of workers starting 5-day shift on day i
**$d_{is}$** = expected demand on day i in scenario s
**c** = cost for each worker below demand level
**$y_{is}$** = expected worker shortfall on day i in scenario s
**$p_{is}$** = probability of scenario s occurring




  - Optimization models treat all of the data as known exactly.
- 15.8 Basic Optimization Algorithms

---

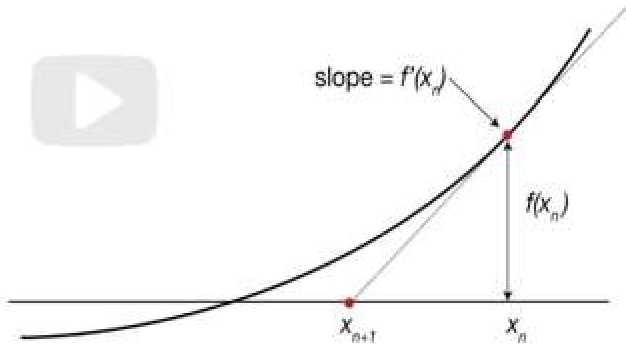


# Algorithm from Calculus

- Newton's method: finding root of f(x)
- Current solution $x_n$ at step n
- $x_{n+1} = x_n - 1\,\frac{f(x_n)}{f'(x_n)}$

Step size

Improving direction



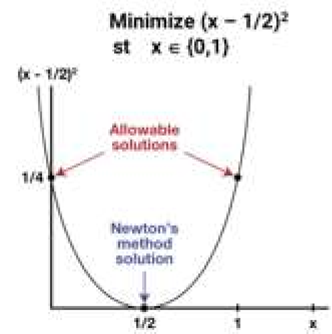

GTx

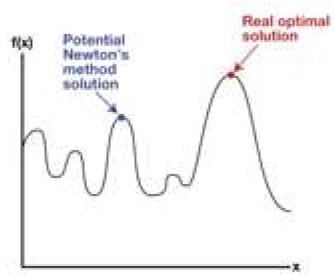
# Convex Optimization Problem

**Convex optimization problem**
- Guaranteed to find optimal solution

**Non-convex optimization problem**
- Not guaranteed to find optimal solution
  - Ex: converge to infeasible solution
  - Ex: converge to local optimum

**Running time**
- Integer programs: can be long
- Linear programs: often fast



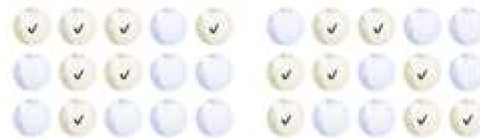

GTx

- The two main steps of most optimization algorithms are:
  - Find a good direction to move from the current solution
  - And determine how far to go in that direction.
  - Many optimization algorithm follow the pattern of finding an improving direction and a step size, make the move, and repeat.

- 16.1 Non-Parametric Methods

## McNemar's Test – Example

- Two competing treatments for a virus
  - A: successful on 61/100
  - B: successful on 68/100



Scenario 1

32 cases: neither worked
61 cases: both A and B worked
7 cases: B worked, A did not
0 cases: A worked, B did not

- Conclude that B is better

p=0.02 (accept)

Scenario 2

12 cases: neither worked
41 cases: both A and B worked
27 cases: B worked, A did not
20 cases: A worked, B did not

- Conclude that B is *not* better

p=0.38 (reject)

McNemar's (binomial) test: only consider where A and B are different

GTx

## Wilcoxon Signed Rank Test for Medians

- Assumption:
  - Distribution is continuous and symmetric
- Question
  - Is the median of the distribution different from m?

GTx

# Wilcoxon Signed Rank Test

**Is the median different from m?**

Given responses $y_1, \ldots, y_n$

1. Rank $|y_1 - m|, \ldots, |y_n - m|$ from smallest to largest
2. $W = \sum_{y_i > m} rank(y_i - m)$ = sum of all ranks where $y_i > m$
3. p-value test for $W$

**Do two sets of paired samples have the same median?**

Given pairs $(y_1, z_1), \ldots, (y_n, z_n)$ from observations $y$ and $z$

- Use $|y_1 - z_1|, \ldots, |y_n - z_n|$ for rank test

**Comparing paired samples**

- Numeric data: use Wilcoxon
- Yes/no data: use McNemar

GTx

# Mann-Whitney test

Two data sets, but not paired samples

- Given independent observations $y_1, \ldots, y_n$, and $z_1, \ldots, z_m$

1. Rank all observations together: $y_1, \ldots, y_n, z_1, \ldots, z_m$
2. $U$ = smaller of two adjusted rank sums:

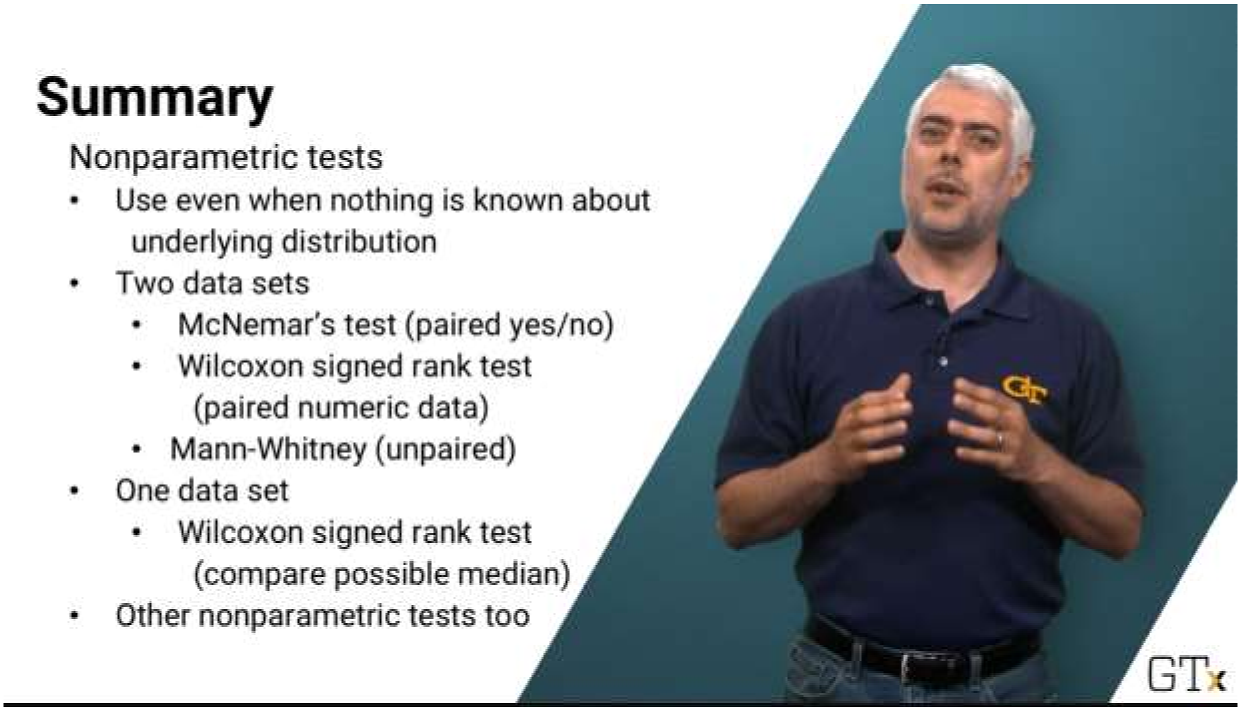
$$U = \min\{U_y, U_z\}$$

$$U_y = \sum_{i=1}^{n} rank(y_i) - \frac{n(n+1)}{2}$$

$$U_z = \sum_{j=1}^{m} rank(z_j) - \frac{m(m+1)}{2}$$

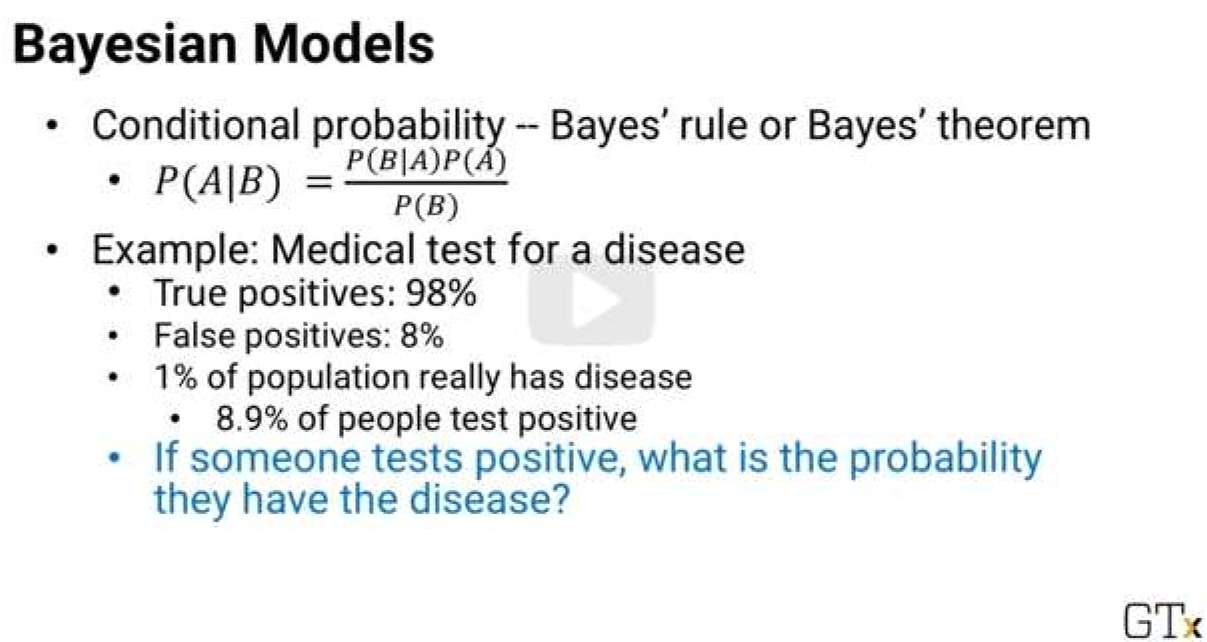
3. Find significance of $U$ (need software or a table)

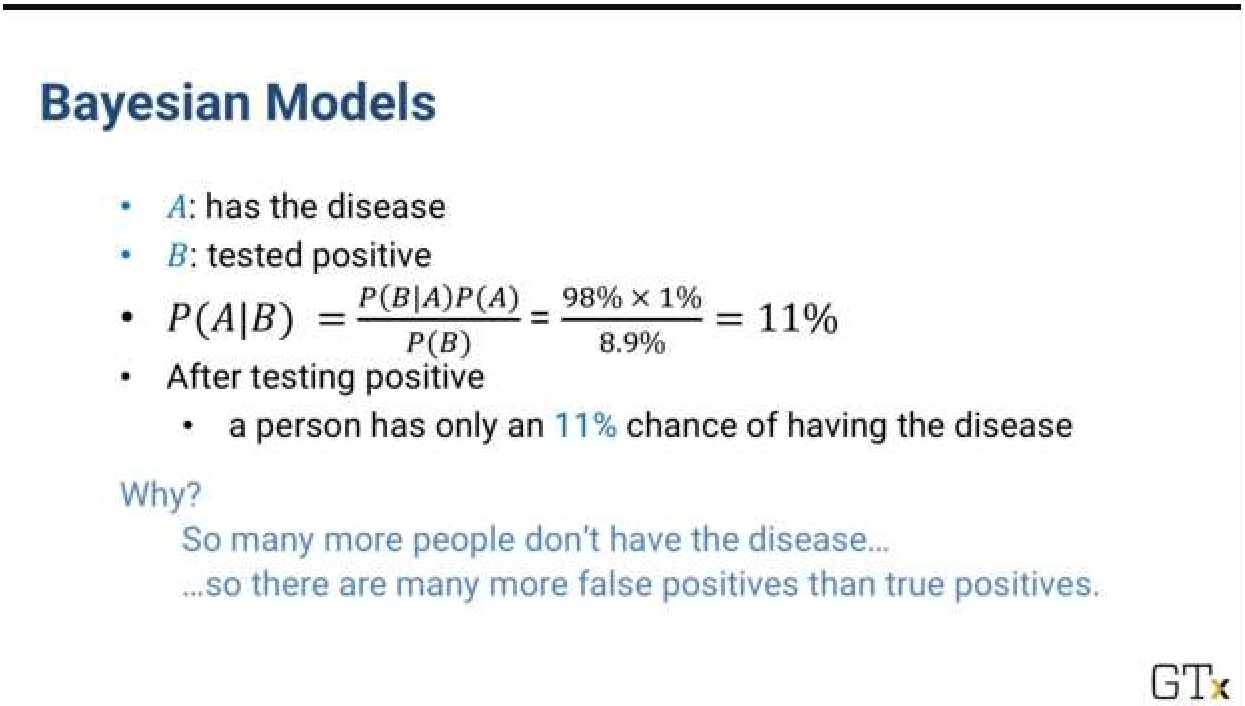
GTx



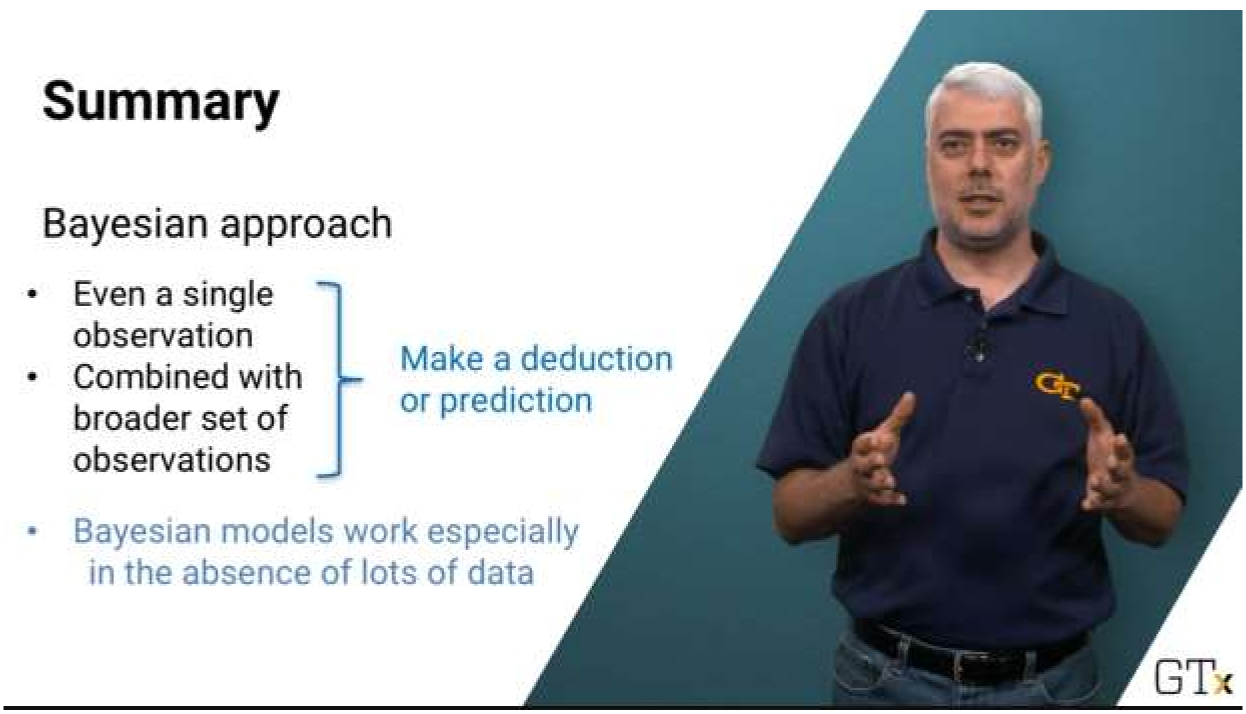

- ○ Because we don't have a good distribution to fit parameters to, a nonparametric test is useful. Many nonparametric tests focus on the median, and they can be used even when we do not know the form of the underlying distribution.}}
- ○ This one is a little tricky! By focusing on the median, nonparametric tests make it less important whether a small data set includes the right distribution and range of results. All a nonparametric test needs is enough data to figure out approximately where the middle value is.}}

- 16.2 Bayesian Modeling






- Baysian Models
    - P(B|A) = probability of testing positive given A = 98%
    - P(A) = probability of having the disease = 1%
    - P(B) = probability of testing positive = 8.9%
- Empirical Bayes Modeling




- Expert opinion can be used to define the initial distribution of P(A) , and observed data about B can be used with Bayes' theorem to obtain a revised opinion P(A|B) .
- The initial distribution assumed for P(A) is called the 'prior distribution' and the revised distribution P(A|B) is called the 'posterior distribution'

• 16.3 Communities in Graphs

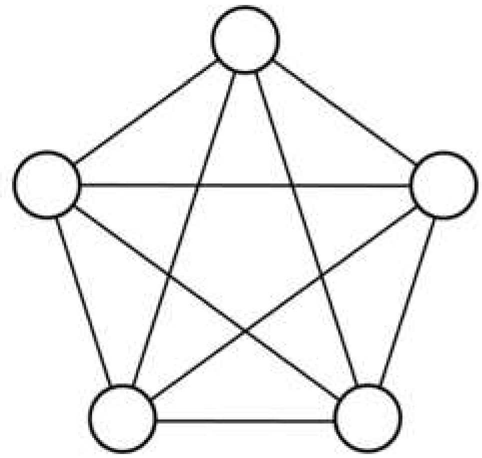
## Communities

**Community**

- a set of circles that's highly connected within itself

**Graph**

- Circles = nodes/vertices
- Lines = arcs/edges
- Clique = a set of nodes that all have edges between each other



Louvain algorithm: decomposing a graph into communities

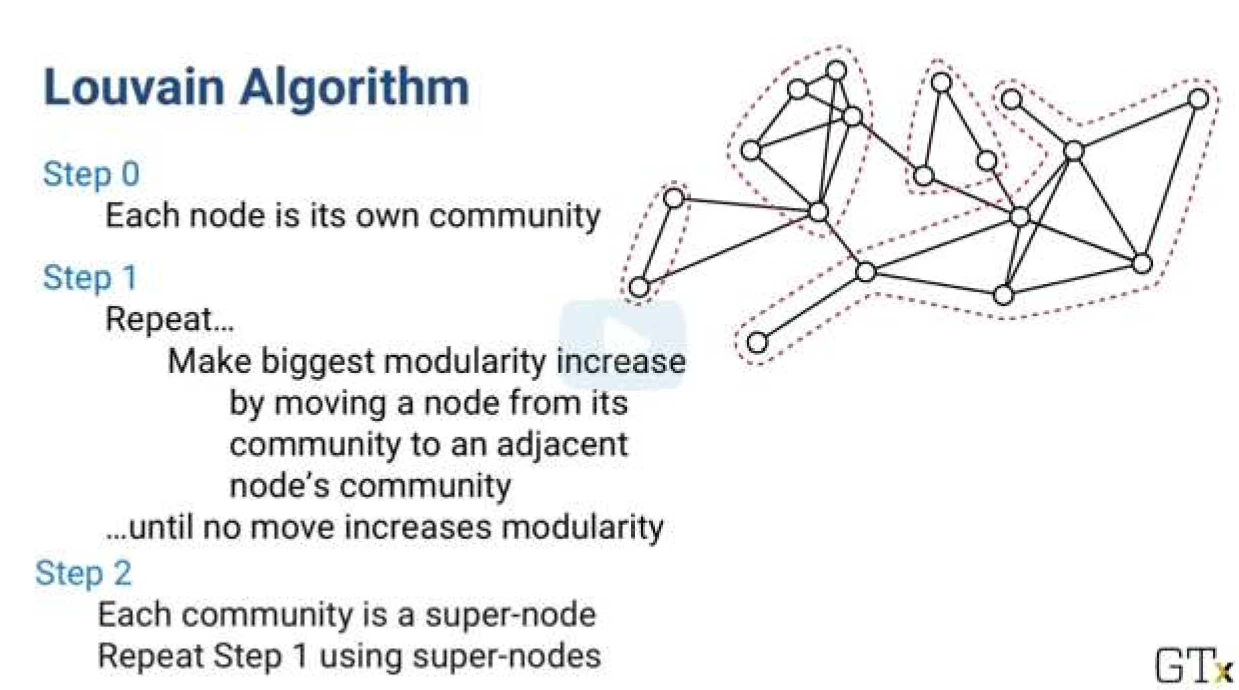
GTx

## Louvain Algorithm

Maximize the modularity of a graph

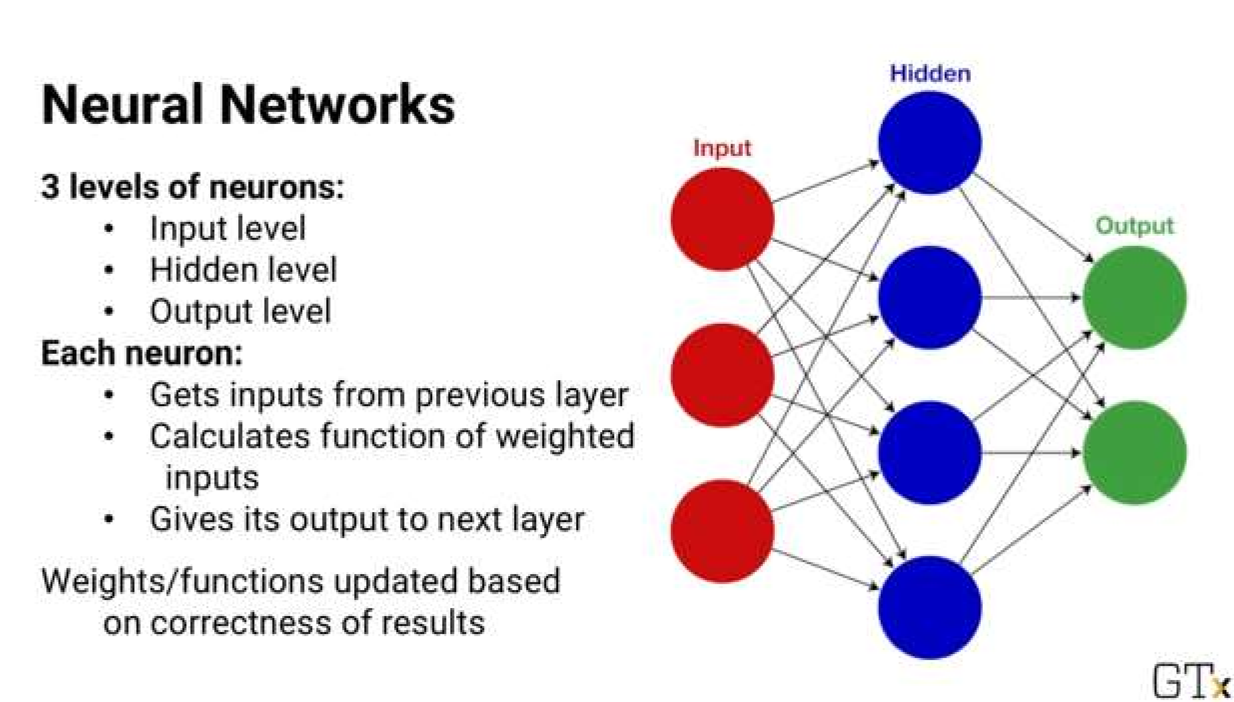
- $a_{ij}$: weight on the arc between nodes $i$ and $j$
- $w_i$: total weight of arcs connected to $i$
- $W$: total weight of all the arcs
- Modularity = $\frac{1}{2W} \sum_{i,j \text{ in same community}} \left(a_{ij} - \frac{w_i w_j}{2W}\right)$

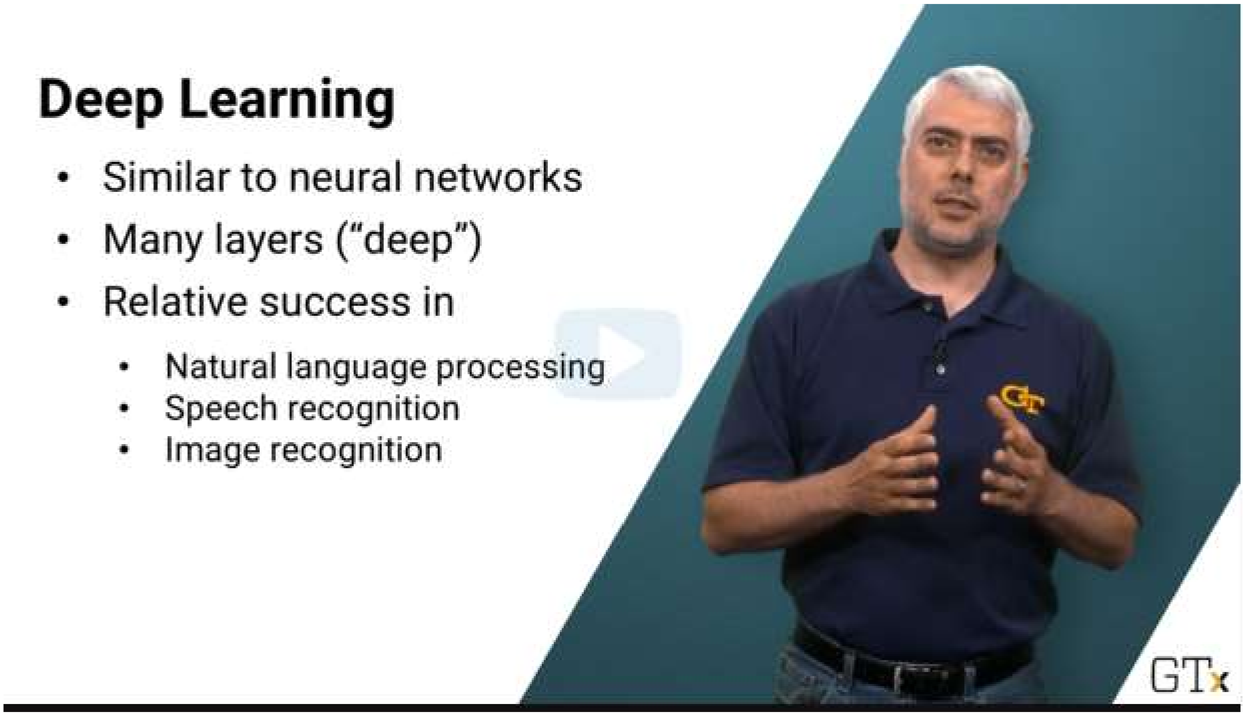
GTx




- Modularity is a measure of how well the graph is separated into communities or modules that are connected a lot internally, but not connected much between each other.
- Louvain is a heuristic (not guaranteed to find the absolute best partition of the graph into communities, it often gives very good solutions though very quickly.
  - Used somewhat often when we want to find communities inside a large network, especially social media networks, and networks of people, computers, etc.
- [Quick jargon break: just like a set of nodes with edges between each pair is called a 'clique', a set of nodes without any edges between them is called an 'independent set'.]

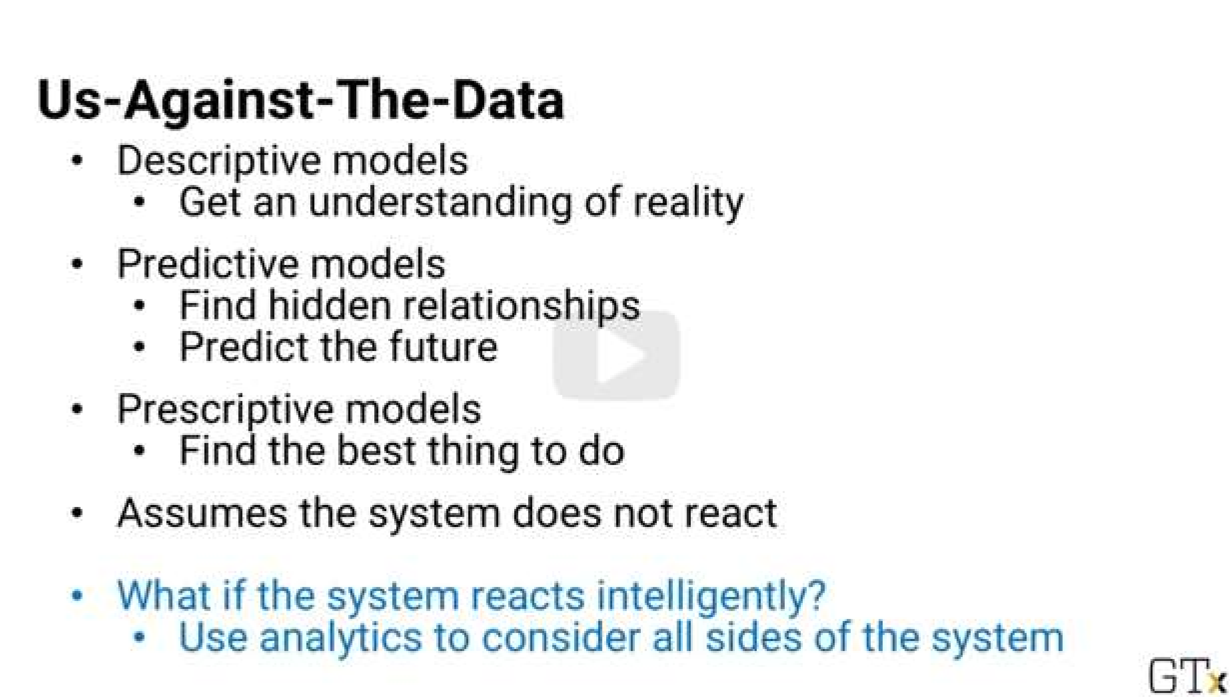
- 16.4 Neural Networks and Deep Learning
  - Basic way to adjust weigths within the model is to use gradient descent using the slope of a function

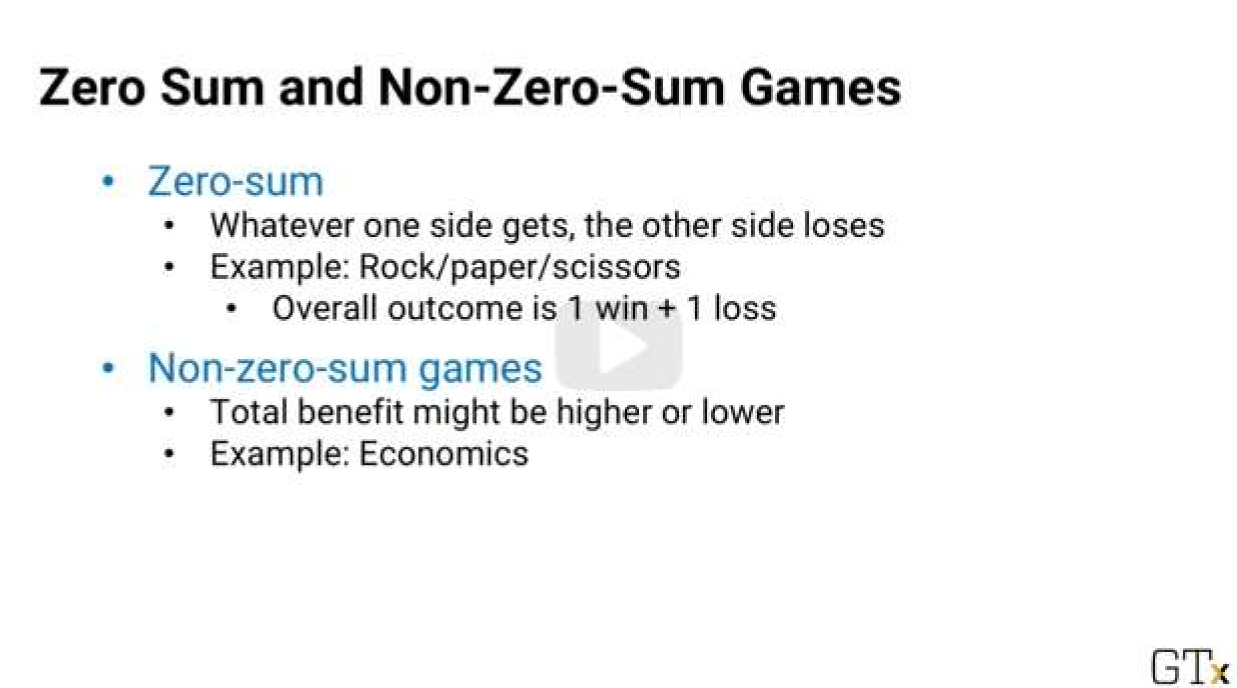




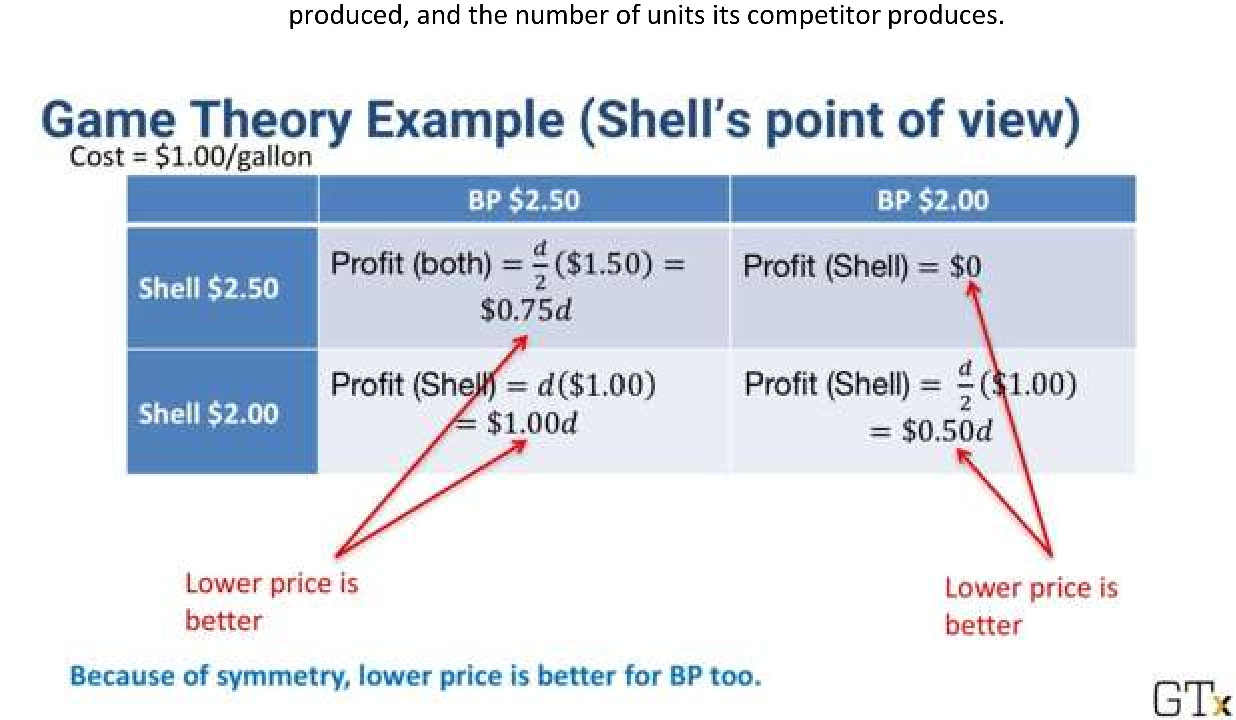

- Deep learning is currently one of the best approaches for recognizing images, speech, writing, and language.

- 16.5 Competitive Models






- Competitive decision making (game theoretic model) appropriate
  - A company wants to optimize its production levels, based on production cost, price, and demand. The company already has estimated a function to give predicted selling price and demand as a function of the number of units produced, and the number of units its competitor produces.
    - A company wants to optimize its production levels, based on production cost, price, and demand. The company already has estimated a function to give predicted selling price and demand as a function of the number of units produced, and the number of units its competitor produces.



# Same Example with Smaller Profit Margin

Cost = $1.75/gallon

| | BP \$2.50 | BP \$2.00 |
|---|---|---|
| Shell \$2.50 | Profit (both) = $\frac{d}{2}(\$0.75) = \$0.375d$ | Profit (Shell) = \$0 |
| Shell \$2.00 | Profit (Shell) = $d(\$0.25) = \$0.25d$ | Profit (Shell) = $\frac{d}{2}(\$0.25) = \$0.125d$ |

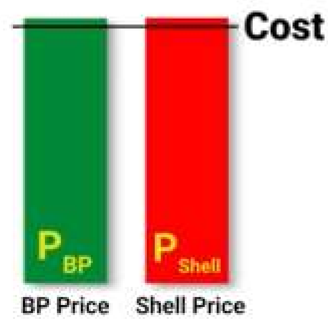
Higher price is better

Higher price is better

Now, both are better off charging higher price.

# Extension of Example

- Choose any price they want
- BP price = $p_{BP}$
  - Shell: If $p_{Shell} > p_{BP}$, then profit = $\$0$ (sell nothing)
    If $p_{Shell} = p_{BP}$, then profit = $\frac{d}{2}(p_{Shell} - \text{cost})$
    If $p_{Shell} < p_{BP}$, then profit = $d(p_{Shell} - \text{cost})$
  - So, Shell might price slightly lower price than BP
    - Then, BP might price slightly lower than Shell
      - Then Shell might price slightly lower than BP
        - Etc.




- Both keep lowering prices until price is about equal to the cost