About Me

- **Professional Experience**
  - Data Scientist – *Paychex, Inc.*
  - Public Health Analyst – *Vermont Department of Health*
  - Business Intelligence Engineer – *MyWebGrocer*
  - Business Analyst – *TetonVillageSport.com*
  - Carpenter – *Henderson Construction*

- **Education**
  - MS Statistics – *University of Vermont*
  - BS Management – *Le Moyne College*
  - *McQuaid Jesuit High School*
DATA SCIENCE & PREDICTIVE ANALYTICS AT PAYCHEX
What happened?
How many, how often, where?
Where exactly is the problem?
What actions are needed?
Why is this happening?
What if these trends continue?
What will happen next?
What’s the best that can happen?
What actions are needed?
Where exactly is the problem?
How many, how often, where?
What happened?

Optimization
Predictive Modeling
Forecasting
Statistical Analysis
Alerts
Query Drilldown
Ad hoc Reports
Std. Reports

Proactive Decision Making
Reactive Decision Making
The Data Science & Predictive Analytics Team

- Team Consists of Seven Data Scientists
- Part of the Risk Management Division
- Established in 2007

Extract.....Transform....Load - ETL

Data Transformations Staging Platform

Centralized Data Warehouse

Wage
Transactional
Marketing
Credit

Data Mining

Ad Hoc Analysis
Predictive Models
Reporting
A Data Scientists Toolbox
Important Skills of a Data Scientist

Motivated by the challenge of mastering difficult and complex subjects

- Intellectual Curiosity
- Communicating Results Across Cross-Functional Audiences
- Project Vision & Planning
- Avoiding Rabbit Holes
Math & Stats in Data Science

Regression Modeling

\[ y_i = \beta_0 + \sum_{i=1}^{n} \beta_j x_i + \epsilon_i \]

- Linear (shown), Poisson, Logistic
- Regression Strategies to predict:
  - Potential clients for a product
  - Sales by zip codes
  - Client retention efforts

**Under the Hood:**

- Calculus
  - Minima & Maxima
  - Derivatives
- Linear Algebra
  - Matrix multiplication
- Logarithms & Exponents
Regression Modeling: Example
Objective: Predict number Client Tax Penalties Per Paychex Branch

**Modeling:**
- What distribution best represents the target?
  - Probability Theory
- Data Preparation
  - Exponents & Logarithms
- Variable Reduction
- Model evaluation
- Presentation of Results
- Deployment
Math & Stats in Data Science

Regression Modeling: Variable Transformations

Histogram of $x$

Histogram of $y$

Histogram of $z$

Histogram of $\log(x)$

Histogram of $\log(y)$

Histogram of $\log(z)$
Math & Stats in Data Science

Cluster Analysis

Euclidean Distance Similarity Metric

\[
sim(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}
\]

Cosine Similarity Metric

\[
sim(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sqrt{\sum_{i=1}^{n} B_i^2}}}
\]
Mathematics & Data Science

Data Visualization*

*Example images from R
“The best thing about being a statistician is that you get to play in everyone’s backyard.”

John Tukey
BEYOND DATA SCIENCE
Trigonometry: Law of Cosines
(w/ a little algebra)

- As a carpenter, I used the law of cosines to determine the angle of two walls when installing crown molding.

\[ C^2 = a^2 + b^2 - 2ab \cos \theta \]
Classroom Application(s)

- Encourage the use of computer applications that require programming
  - R, Mathematica, Matlab

- Data exploration: Find a data set online; summarize with analytical & visual tools

- Promote the connection of application and theory through exercise(s):
  - Example: Use __________ (some math/stat task) to solve a problem related to the students personal interests?

  - While exercises like this are common for final projects, there is value in early adoption to maintain interest and keep focus on the big picture.
Thank You!

Questions?

Chip Galusha
Data Scientist
Paychex, Inc.
fgalusha@paychex.com
frankgalushajr@gmail.com