An

# Interactive Data Explorer

and

# Synthetic Prediction Tool

for

## Characterizing Allegheny County Opioid Overdoses

Deep Girl Network

Theresa Gebert
Shuli Jiang
Chris Sheng

HackAuton 2018

Carnegie Mellon University

# How to best help policy-makers?

- What exactly is happening right now: who is affected and how much?
- Why and how did this happen: what factors influenced increased opioid usage?
- What can be done to slow and stop this epidemic?
- How can we prevent this from happening again in the future?

# What gets in their way?

- How to manipulate and visualize the data?
- Do we *really* need more data?

# Product Requirements

**INTERACTIVE DATA EXPLORER**

- interactive
- easy and intuitive to use
- lightweight and cheap to maintain

**SYNTHETIC PREDICTION TOOL**

- dynamic
- easy and intuitive to use
- incorporates a measure of uncertainty

# Interactive Data Explorer

- interactive
- easy and intuitive to use
- lightweight and cheap to maintain

## RShiny

# DEMO

# Synthetic Prediction Tool

- Generate synthetic control population
- Extract additional features & match to overdose cases
- Build models

# Synthetic Prediction Tool

- Generate synthetic control population
  - freely available EMR data
  - use population-level statistics to mimic reality

# Synthetic Prediction Tool

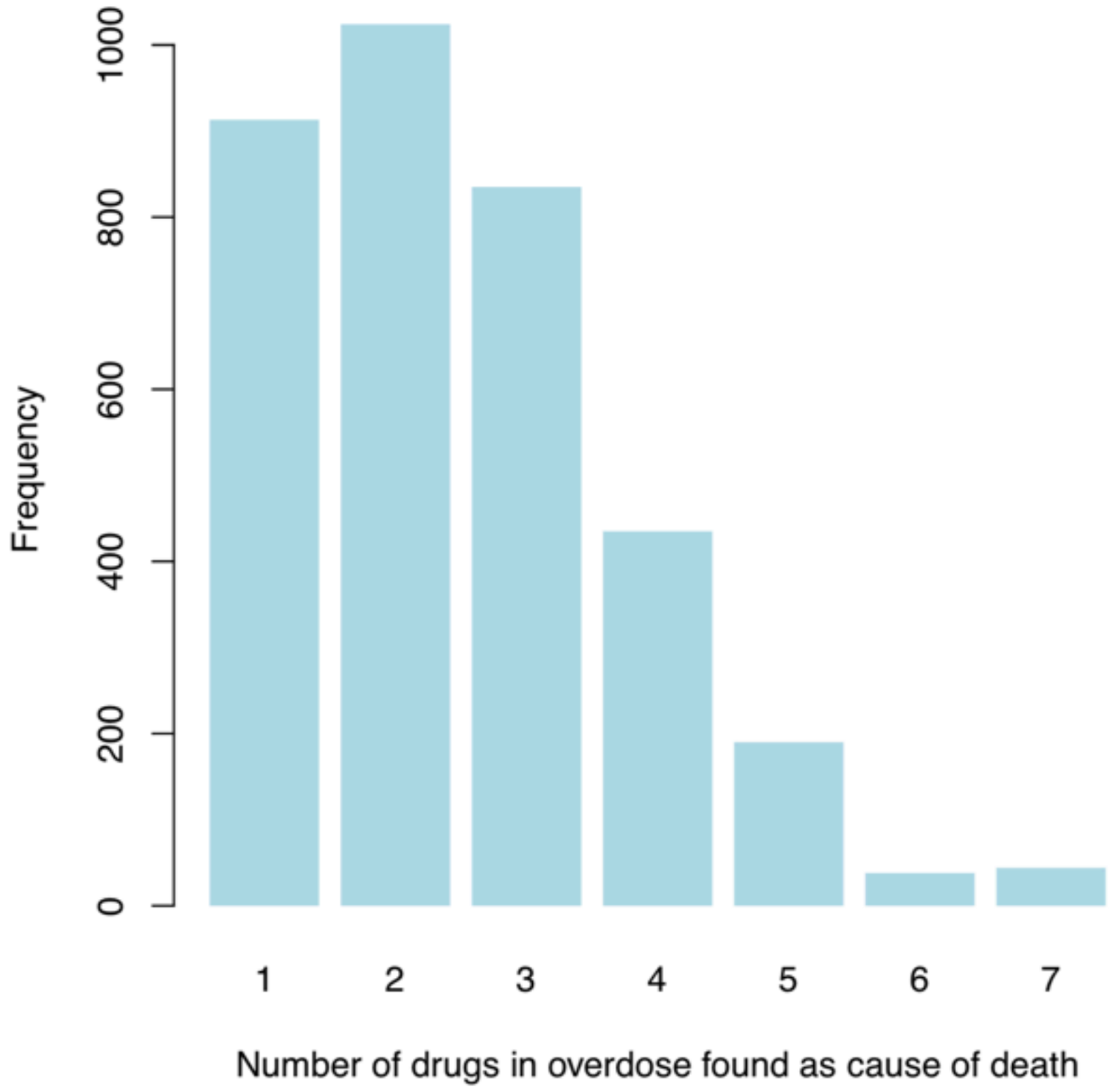
- Extract additional features & match to overdose cases
  - **MATCHING**
    - for each individual that overdosed, match to synthetic person:
      - exactly on **gender**
      - if **race** is white, black, or Asian, then match exactly, otherwise match randomly
      - within 3 years on **age**
  - then remove from candidate pool

# Synthetic Prediction Tool

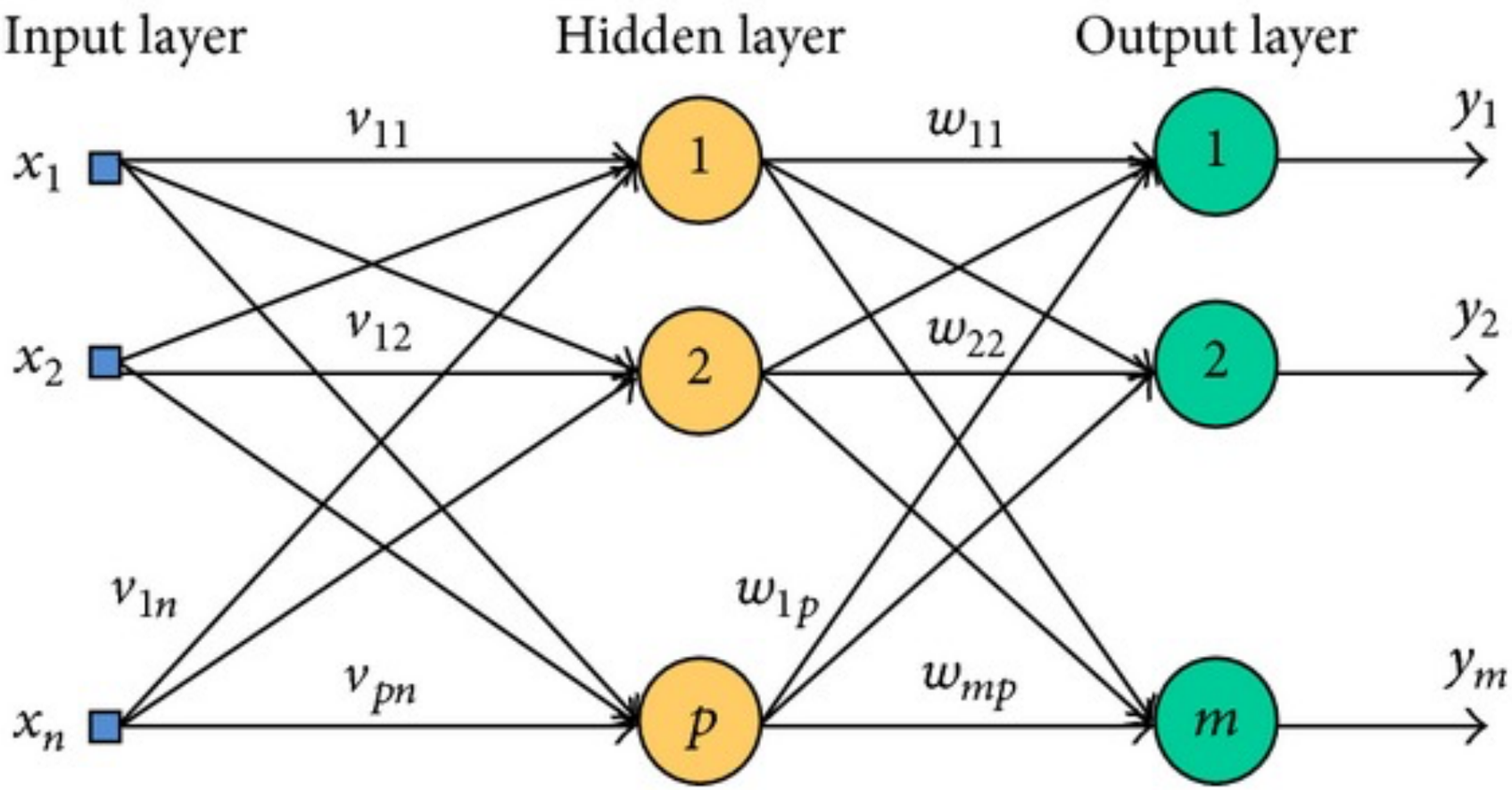
- Extract additional features & match to overdose cases
  - **ADDITIONAL FEATURES**
    - **marital status**
    - **socioeconomic status**
    - **language**
    - **sickliness**: time-discounted days in hospital
    - **disease history**: the number of occurrences in each category of disease as defined by ICD-10

# Synthetic Prediction Tool

- Build models
  - **Poisson generalized linear model**
  - **Random forest**
  - **Neural network**


Frequency
0
200
400
600
800
1000
1
2
3
4
5
6
7
Number of drugs in overdose found as cause of death

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.6871 | 0.0548 | 12.54 | **0.0000** |
| as.factor(Sex)Male | -0.0353 | 0.0230 | -1.54 | 0.1243 |
| as.factor(Race)White | 0.1390 | 0.0329 | 4.23 | **0.0000** |
| as.factor(Race)Hispanic | -0.0946 | 0.2519 | -0.38 | 0.7072 |
| as.factor(Race)Other | -0.0823 | 0.3550 | -0.23 | 0.8167 |
| as.factor(Race)Asian | -0.2169 | 0.2907 | -0.75 | 0.4555 |
| as.factor(Race)Middle Eastern | 0.5753 | 0.2902 | 1.98 | **0.0475** |
| as.factor(Race)Unidentified | -0.8342 | 1.0005 | -0.83 | 0.4044 |
| as.factor(Race)Indian | -0.0258 | 0.7082 | -0.04 | 0.9709 |
| Age | 0.0032 | 0.0009 | 3.66 | **0.0003** |


Input layer
Hidden layer
Output layer
$x_1$
$x_2$
$x_n$
$v_{11}$
$v_{12}$
$v_{1n}$
$v_{pn}$
$w_{11}$
$w_{22}$
$w_{1p}$
$w_{mp}$
1
2
p
1
2
m
$y_1$
$y_2$
$y_m$

# Product Requirements

**INTERACTIVE DATA EXPLORER**

- interactive
- easy and intuitive to use
- lightweight and cheap to maintain

**SYNTHETIC PREDICTION TOOL**

- dynamic
- easy and intuitive to use
- incorporates a measure of uncertainty

# Policy Recommendations

**Collect more data and learn more.**

**Crack down on Fentanyl.**

**Identify dangerous combinations.**

# Future Work

**MATCHING**

- matching on more exhaustive criteria
- more flexible ways of drawing samples

**ANALYSIS**

- Does the significant effect of white and Middle Eastern race on the number of drugs used disappear once socioeconomic status is taken into account?
- Does the significant effect of age on the number of drugs used disappear once health is taken into account?
- What are the bigger killers: hard drugs combined with alcohol, or a dangerous variety of less potent drugs?
- What is the elbow point of drug usage: is there a particular drug or drug combination that leads to a significant spike in drug usage?
- How correlated do additional features need to be to the outcome of interest in order to yield model improvements?
- What masking techniques are most effective at providing high prediction accuracy improvements at low cost to privacy?

# thank you

# References


1. Anna L. Buczak, Steven Babin, and Linda Moniz. *Data-driven approach for creating synthetic electronic medical records.* BMC Medical Informatics and Decision Making. 10(1):59, Oct 2010.
2. Ching Lee Koo, Mei Jing Liew, Mohd Mohamad, and Abdul Hakim Mohamed Sales. *A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology.* 2013:432375, 10, 2013.
3. Nicholas J. Gotelli, Edmund M. Hart, and Aaron M. Ellison. *Co-occurence Analysis.* Cran R-Project. March 8, 2015.
4. R Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/
5. Kudakwashe Dube and Thomas Gallagher. *Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use*. Lecture Notes in Computer Science book series (LNCS, volume 8315).