# Antitrust Notice

- **The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.**

- **Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.**

- **It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.**

Verisk Analytics

# Fueling Innovation From Raw Data

**Rama Duvvuri, FCAS CPCU**
**Vice President – Analytics**
**ISO Innovative Analytics**

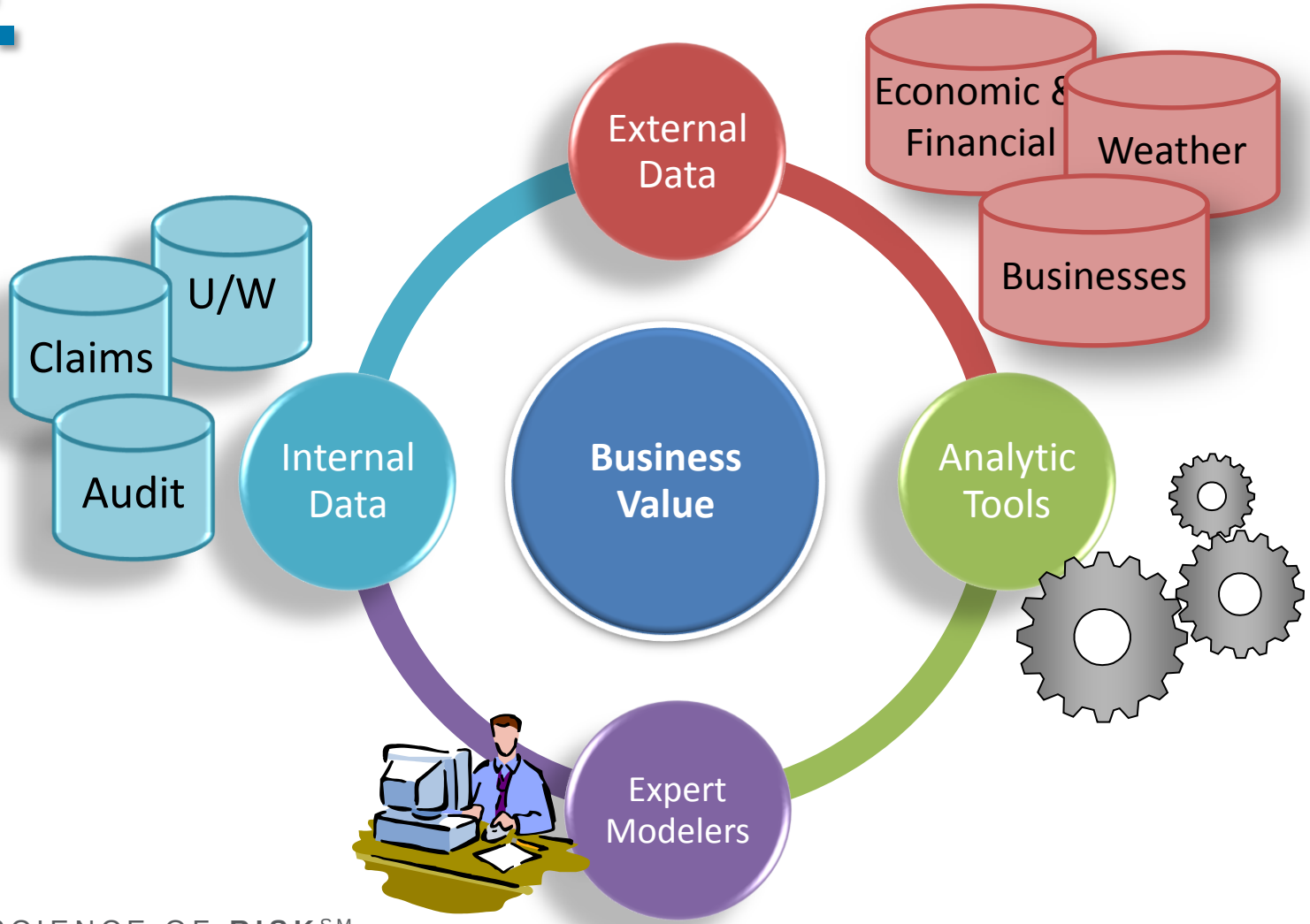
**March 2012**

THE SCIENCE OF **RISK**SM

# Agenda

- **Data**
  - Value, opportunities, and challenges
- **Speed to market?**
  - *Analytic objects* (components) to the rescue
- **Examples of Raw Data**
  - Business locations data in personal auto
  - Weather data in homeowners peril-level modeling
  - Economic growth data in premium audit modeling

# Driving Business Value with Analytics



# Data: Opportunities and Challenges

## Opportunities

- Increasing diversity and volume of useful data
  - Free and fee-based
- Ability to store large data sets
  - Decreasing costs of storage
- Computing capabilities to manipulate large data sets
  - Ever-increasing compute power

## Challenges

- Complex data structures
- Sheer volume
- Multitude of variables
  - Significant preprocessing
- Data management
  - Storage and refresh
- Raw data is rarely useful
  - Need right techniques and tools to extract value

**Considerable value but speed to market remains a concern**

# Business Locations Data

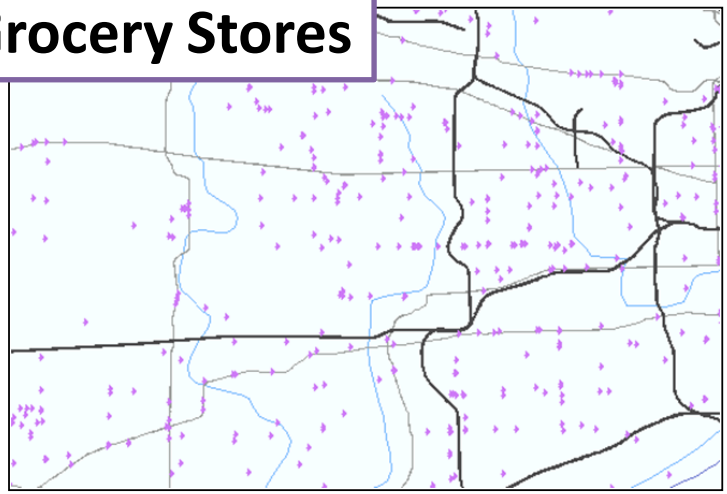
# Problem – Geographic Risk Estimation

# The Data: Business Points

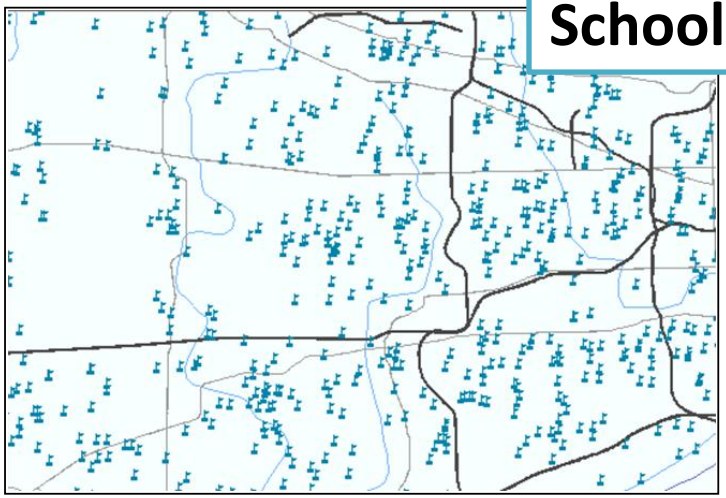
- From a business data vendor (fee-based)
  - 13 million businesses
  - Latitude/longitude
  - 108 distinct SIC codes
- Identify "traffic generators"
  - Businesses that produce traffic in their vicinity
    - e.g., malls/shopping centers, transportation hubs, etc.

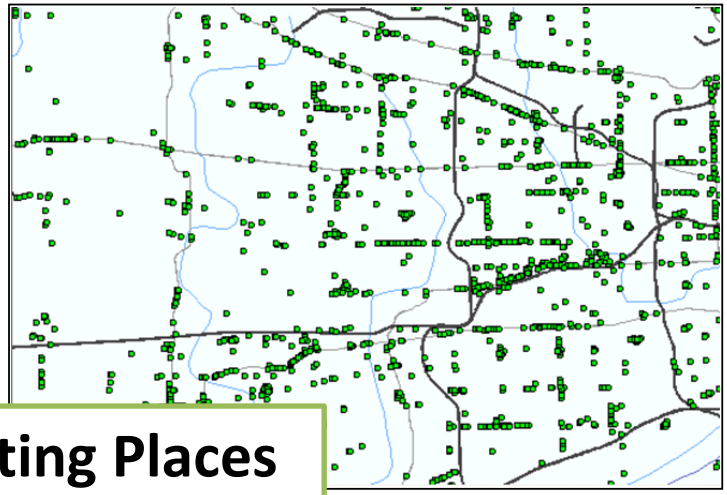
# Examples of Traffic Generators

Grocery Stores

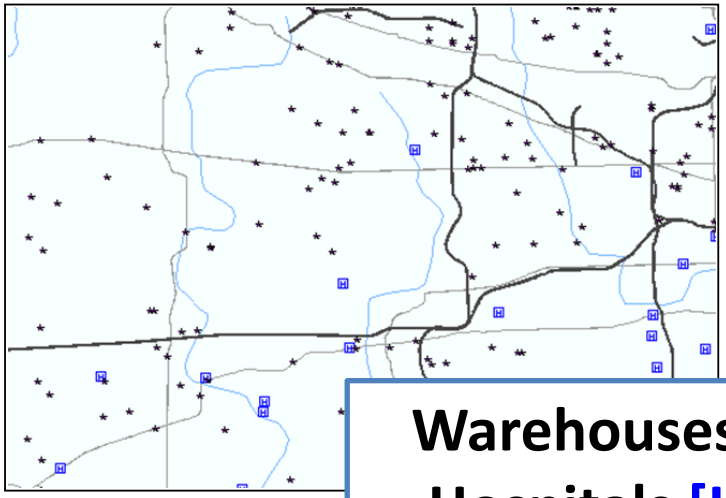


Schools



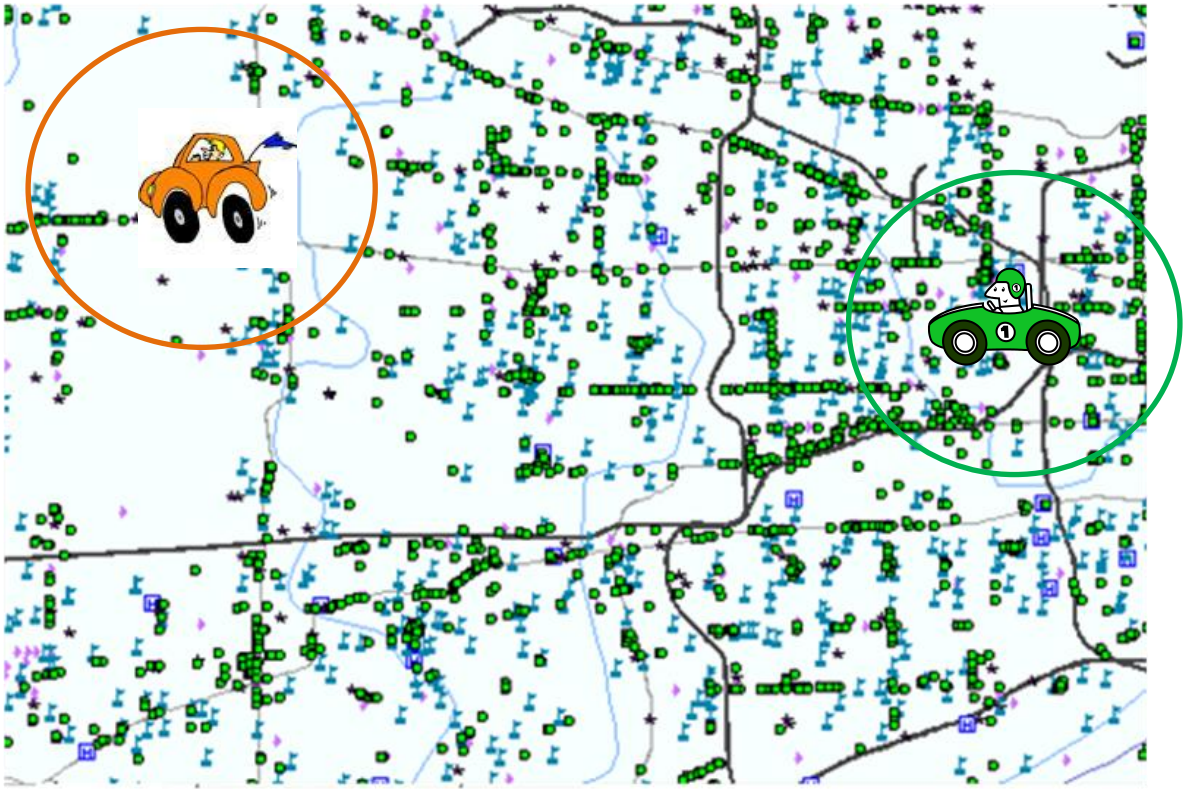


Eating Places



Warehouses *
Hospitals [H]

# Traffic Generators and Auto Losses



- **Distribution of businesses (traffic generators) are different at these two garaging locations**
- **Is there a correlation between these traffic generators and auto losses?**

# Deriving Useful Features

| Count of Grocery Stores within .50 miles | Count of Grocery Stores within 5 miles | Count of Hospitals within .50 miles | Count of Hospitals within 5 miles | Count of Eating Places within .50 miles | Count of Eating Places within 5 miles |
|---|---|---|---|---|---|
| 1 | 48 | 0 | 1 | 9 | 240 |
| 0 | 174 | 0 | 9 | 10 | 970 |

- **Numerous "calculated" dimensions**
  - Distance to nearest business of a certain type
  - Number of businesses of a type within a radius "R"
    - R = 0.25, 0.5, 1, 2, 5, 10, etc.
  - Density estimates
    - Per capita variables, etc.



Area coverage – ½ mile radius = < 1 square mile; 5 mile radius ~ 80 square miles

# Traffic Generators Are Correlated

Correlations among Traffic Generators at ½ Mile

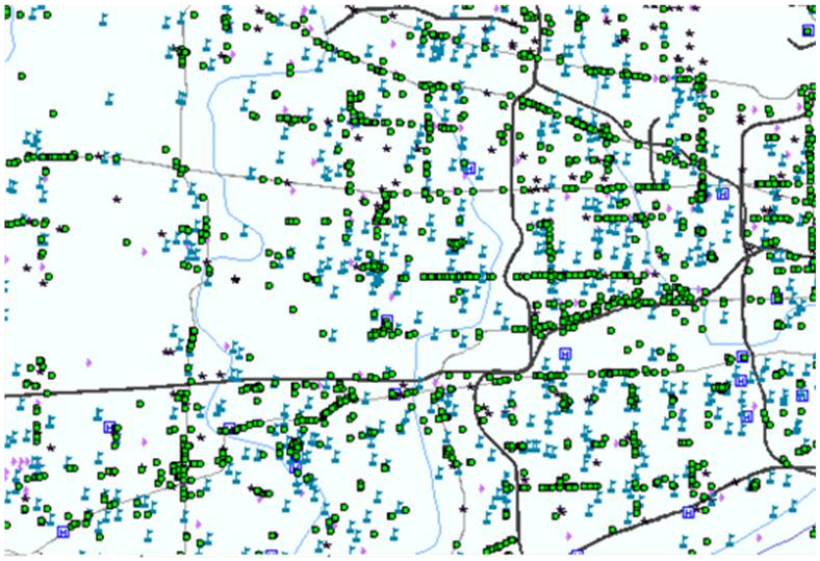
| | Grocery 1/2 Mile |
|---|---|
| Grocery 1/2 Mile | 1 |
| School 1/2 Mile | 0.658 |
| Warehouse 1/2 Mile | 0.564 |
| Hospital 1/2 Mile | 0.338 |
| Eating Places 1/2 Mile | 0.814 |

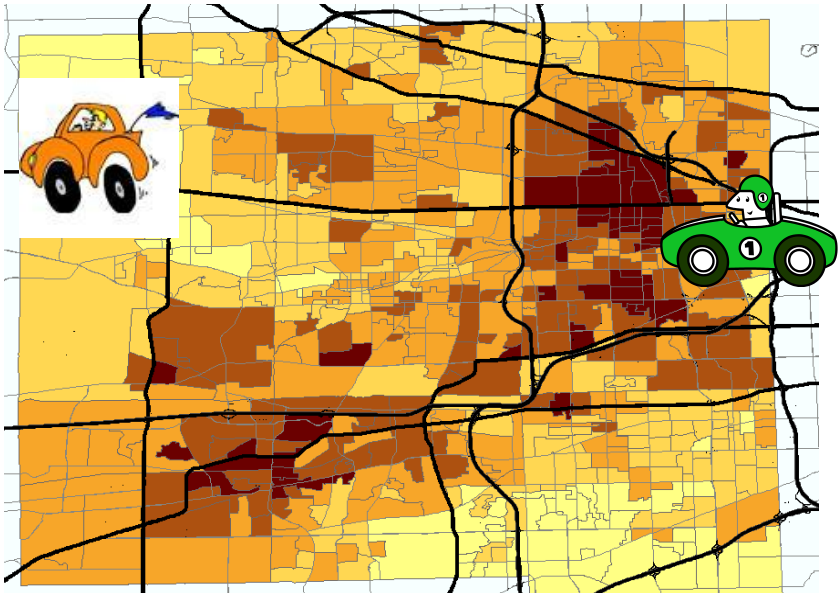
Correlations among Traffic Generators at 5 Miles

| | Grocery 5 Miles | School 5 Miles | Warehouse 5 Miles | Hospital 5 Miles | Eating Places 5 Miles |
|---|---|---|---|---|---|
| Grocery 5 Miles | 1 | | | | |
| School 5 Miles | 0.969 | 1 | | | |
| Warehouse 5 Miles | 0.958 | 0.960 | 1 | | |
| Hospital 5 Miles | 0.871 | 0.859 | 0.879 | 1 | |
| Eating Places 5 Miles | 0.984 | 0.954 | 0.967 | 0.913 | 1 |

- At ½ mile the traffic generators are moderately correlated
- At 5 miles the traffic generators are very highly correlated
- Selecting from highly correlated variables is problematic

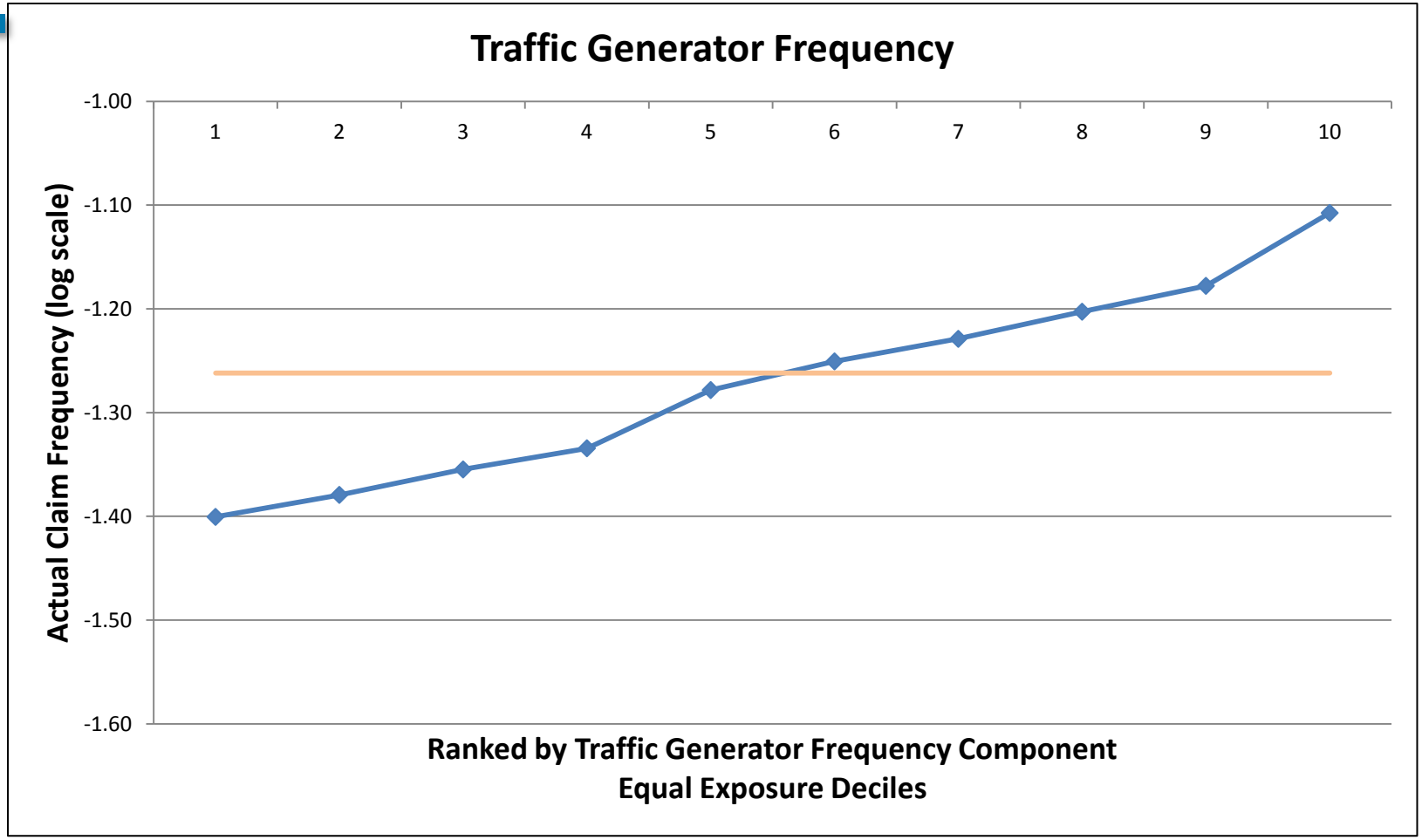
# Collision Frequency Model





- Collision frequency as a function of traffic generators — darker shades represent higher collision frequencies
- This model uses a complex set of traffic generators at various distances

# Components as Predictors



# Personal Auto Environment Model:
## Components and Examples

- **Weather/Terrain**:
  - Measures of snowfall
  - Measures of rainfall
  - Measures of temperature
  - Elevation changes
- **Traffic Density and Driving Patterns**:
  - Commute patterns
  - Public transportation usage
  - Daytime occupancy
  - Speed limits
  - Traffic loads
- **Traffic Composition**:
  - Demographic groups e.g.
    - Household size, home ownership
  - Age distribution
  - Housing occupancy
- **Traffic Generators**:
  - Transportation hubs
  - Shopping centers
  - Hospitals/medical centers
  - Entertainment districts
- **Experience and Trend**:
  - ISO loss cost
  - State frequency/severity trends

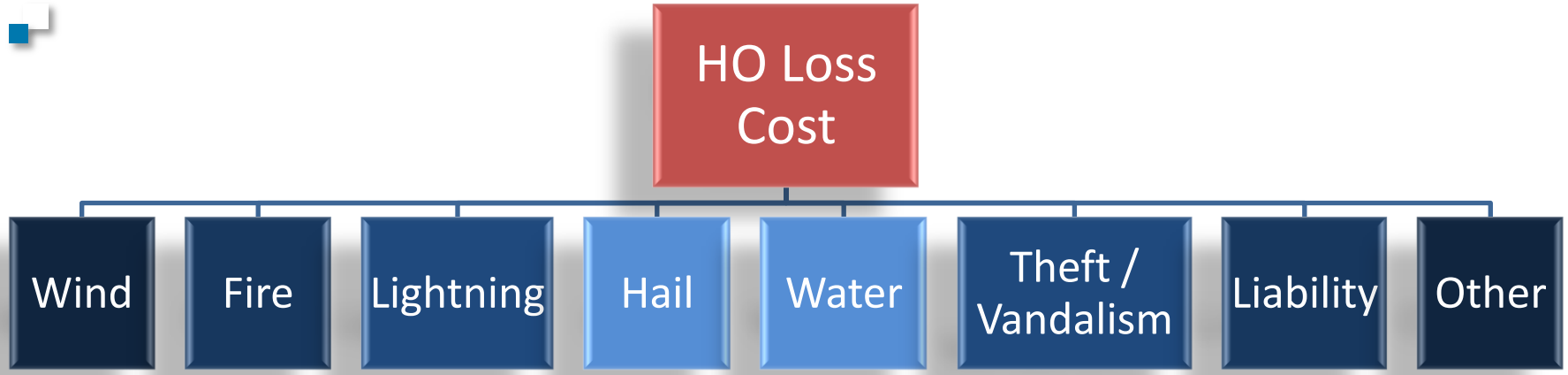
# Increased Segmentation and Value

| Model | Gini Index | Value of Lift |
|---|---|---|
| Current Territories | 8.37% | - |
| Environmental Module -Base Industry Model | 9.49% | $2.75 |
| Insurer Custom Model Using Components | 10.31% | $6.98 |

- Modeling using components doubled incremental lift over current territories

# Weather Data

# Homeowners Peril-Level Risk Estimation



# The Data: NARR Weather Data

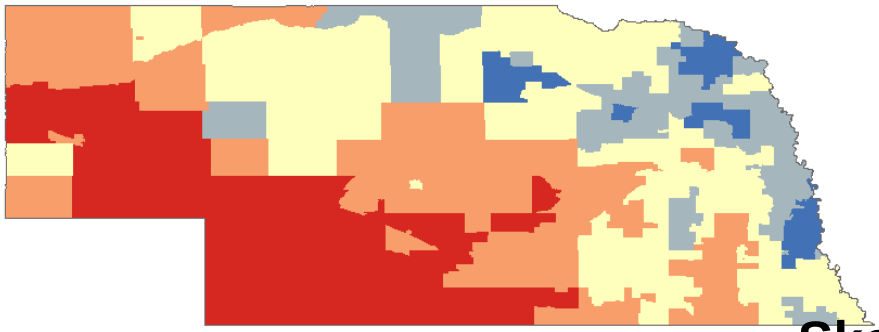
- North American Regional Reanalysis (NARR)
  - "Best/most accurate North American weather and climate data set"
- Data Range – 1979–2007
- Granularity – 32 x 32 km grid
- 8 daily readings (every 3 hours) – raw data
  - Accumulated precipitation
  - Air temperature
  - Rain
  - Wind
  - Relative humidity
  - Snow depth
  - etc.
- Data size ~ 150 GB

# Derive Potentially Useful Features

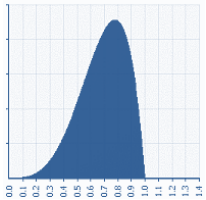
- Temperature
  - Mean
  - Maximum deviation from mean
  - Daily range = Daily max – Daily min
  - Number of consecutive days below freezing, etc.
- Wind
  - Number of days with high wind, etc.
- Precipitation
  - Number of days with severe precipitation
  - Number of days without precipitation, etc.
- Interactions
  - Days without precipitation, high temperature, and high wind, etc.

# Explore Higher-Order Moments

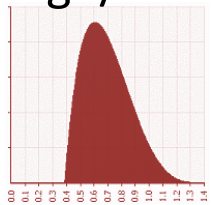
**Mean of Temperature Range**



Low/Neg



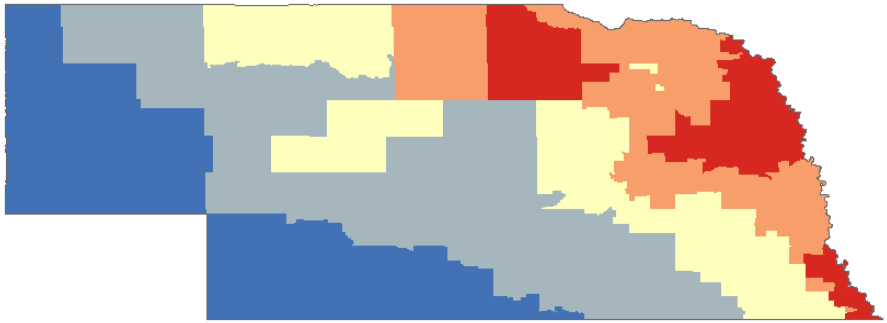
High/Pos



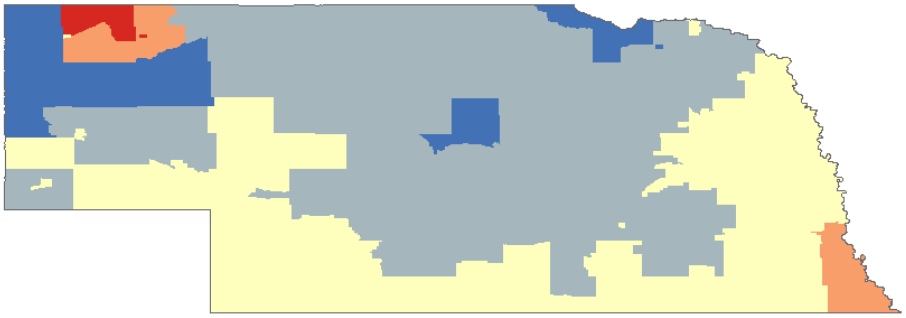
Range = Max – Min

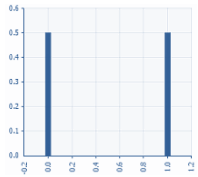
Low  High

**Skewness of Temperature Range**

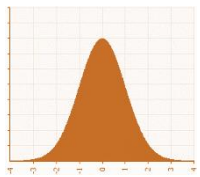


**Kurtosis of Temperature Range**

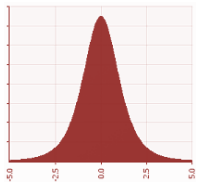




Low/1



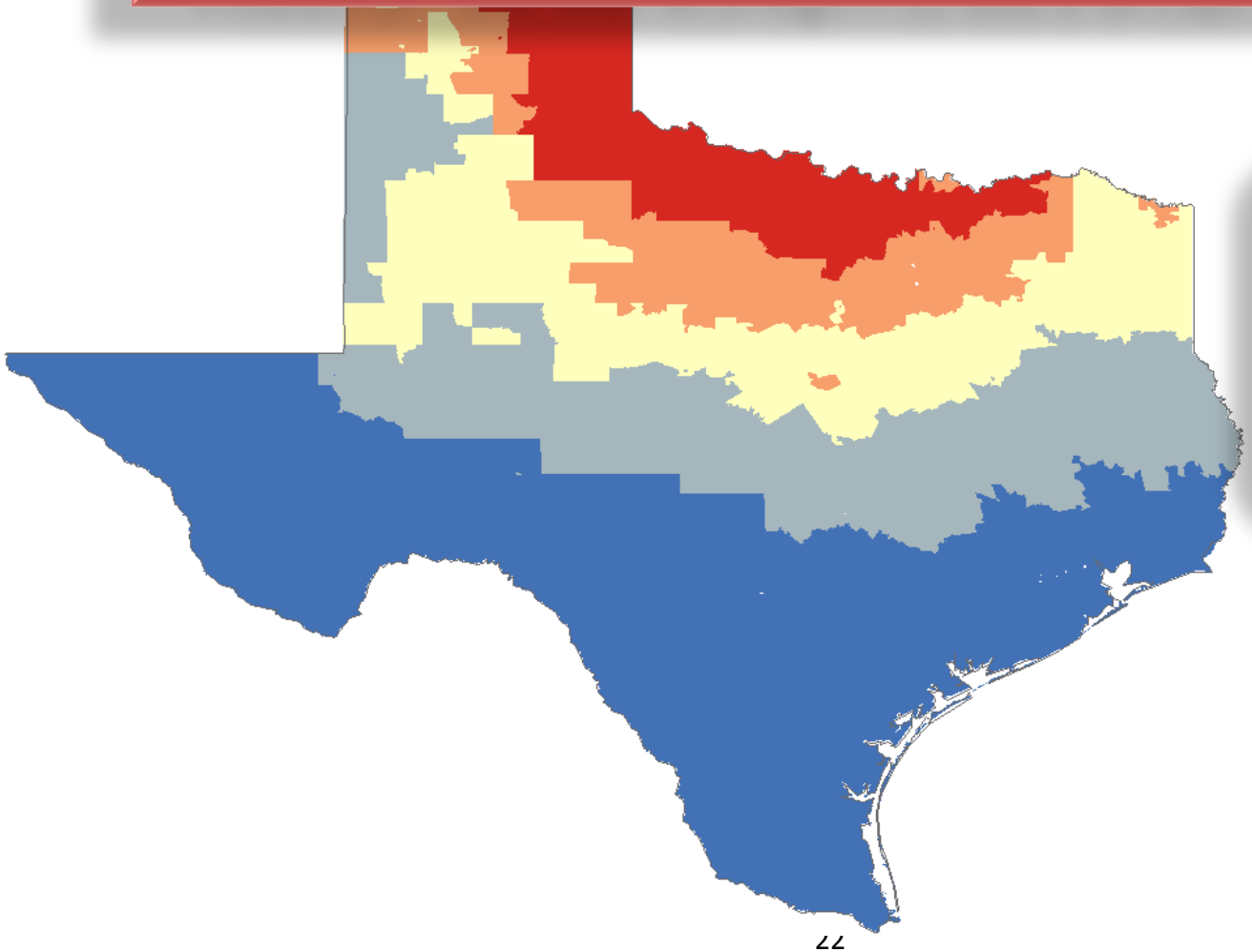
Med/3



High/∞

# Explore Interactions: Use Visualization

Product of - % of days with high < 32 and % of days with low > 72

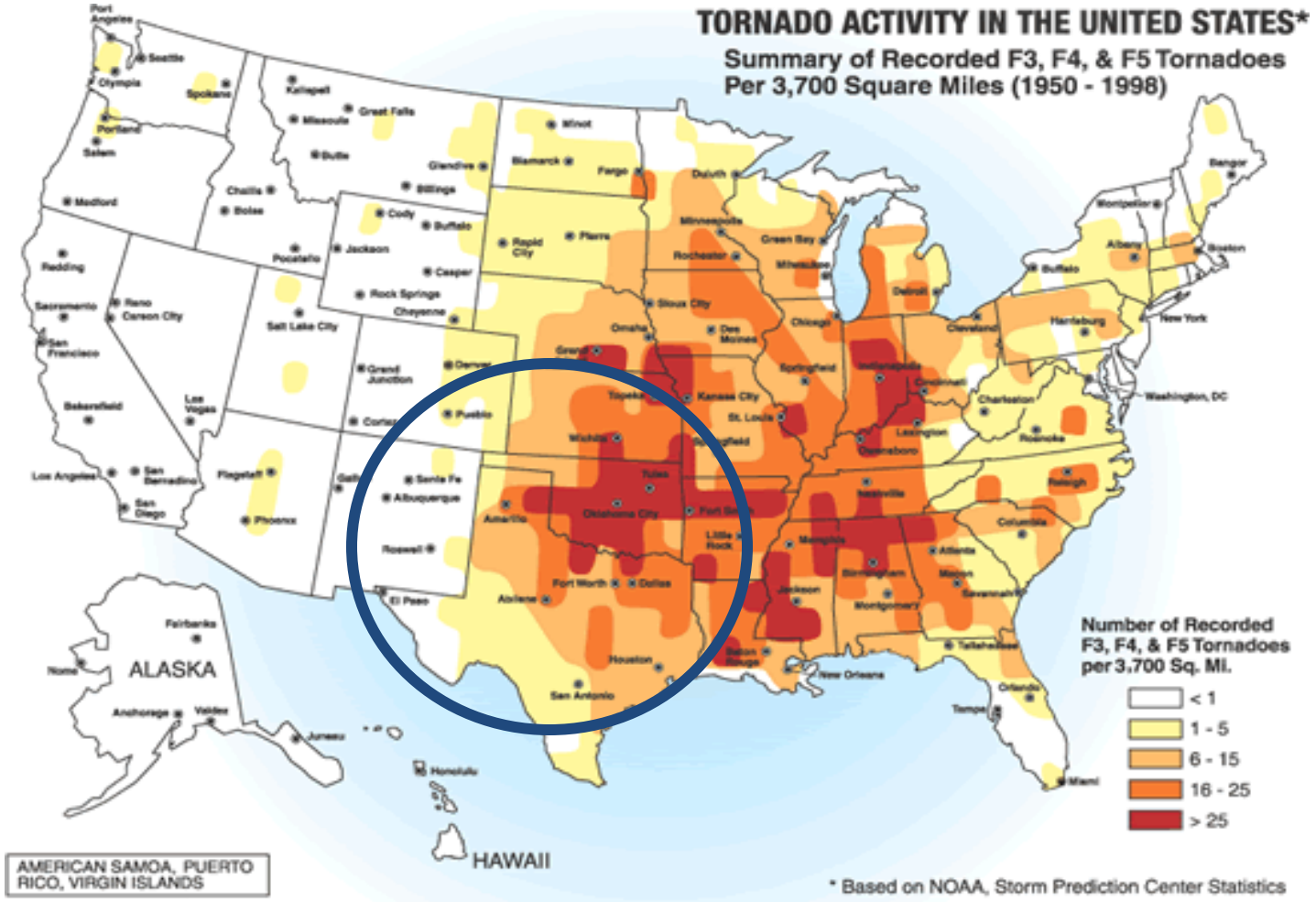


Positive coefficient in wind frequency model – Why?

Low High

T

# Validate Findings Where Possible




* Source: Federal Emergency Management Agency (FEMA) based on National Oceanic and Atmospheric Administration (NOAA) Storm Prediction Center Statistics

# Homeowners Environment Model:

## Components and Examples

- **Weather / Elevation**:
  - Elevation
  - Measures of precipitation
  - Measures of humidity
  - Measures of temperature
  - Measures of wind

- **Proximity :**
  - Commuting patterns
  - Population density
  - PPC

- **Trend / Experience:**
  - Peril's proportion of ISO Loss Cost
  - Trend
  - Amount of insurance

- **Commercial and Geographic Features:**
  - Distance to coast
  - Distance to major body of water
  - Local concentration of types of businesses (e.g., shopping centers)

# Economic Data

# Problem: Identifying Policies to Audit

**What is premium audit?**

- Review of the insured's records to determine the true exposures and correct premium
- Regulatory requirement in WC
- Typically in 30–90 days of policy expiration

**How is it performed?**

- Physical (by a human auditor) – accurate but expensive
- Phone – not very accurate but cost-effective
- Self-report (insured fills and sends a form) – very unreliable

**What are the results?**

- AP – insured owes **additional premium**
- RP – carrier **returns premium** to insured
- Closed-even – no change to premium

# Diverse Data Sources Add Value

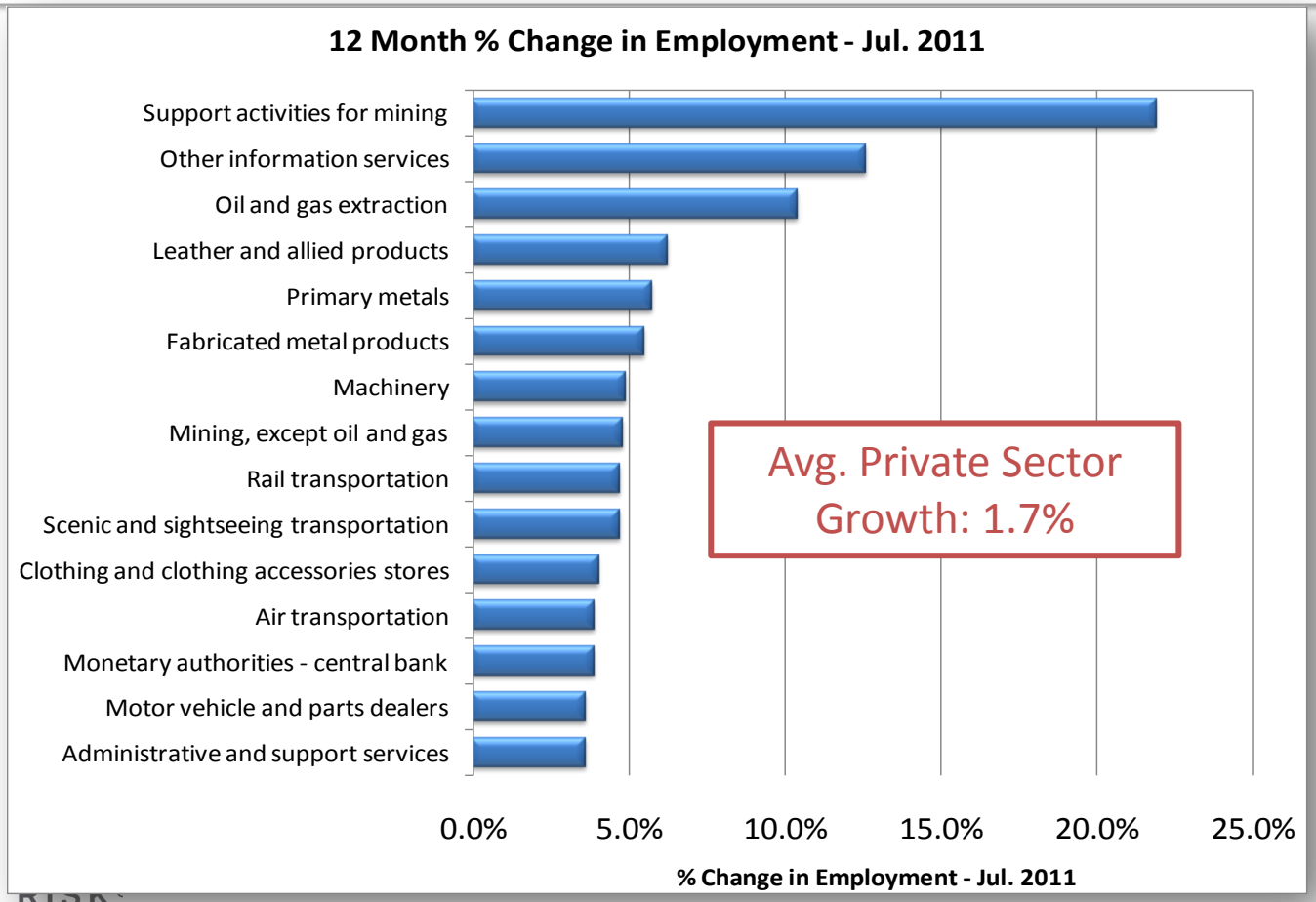
- Federal Reserve
  - Interest rates and money supply
- Bureau of Labor Statistics (BLS)
  - Unemployment statistics
  - Injury, illnesses, and fatalities (IIF)
  - Wages and occupations
- Bureau of Economic Analysis (BEA)
  - Various measures of economic output (GDP)
- Area demographics
  - Census

# Example – Current Employment Survey

- Monthly BLS survey
  - Approximately 140,000 businesses and government agencies representing approximately 410,000 worksites
    - Provides employment, paid hours, and earnings information on a national basis
    - More than 1,100 industries at various levels of aggregation
    - 290 series of seasonally adjusted data
    - 550 special derivative series, such as indexes
- Thousands of features created
  - Ratios, indexes, change over time, change over geography, etc.

# Identify Hot Sectors Early

## Employment growth varies by industry during the economic cycle

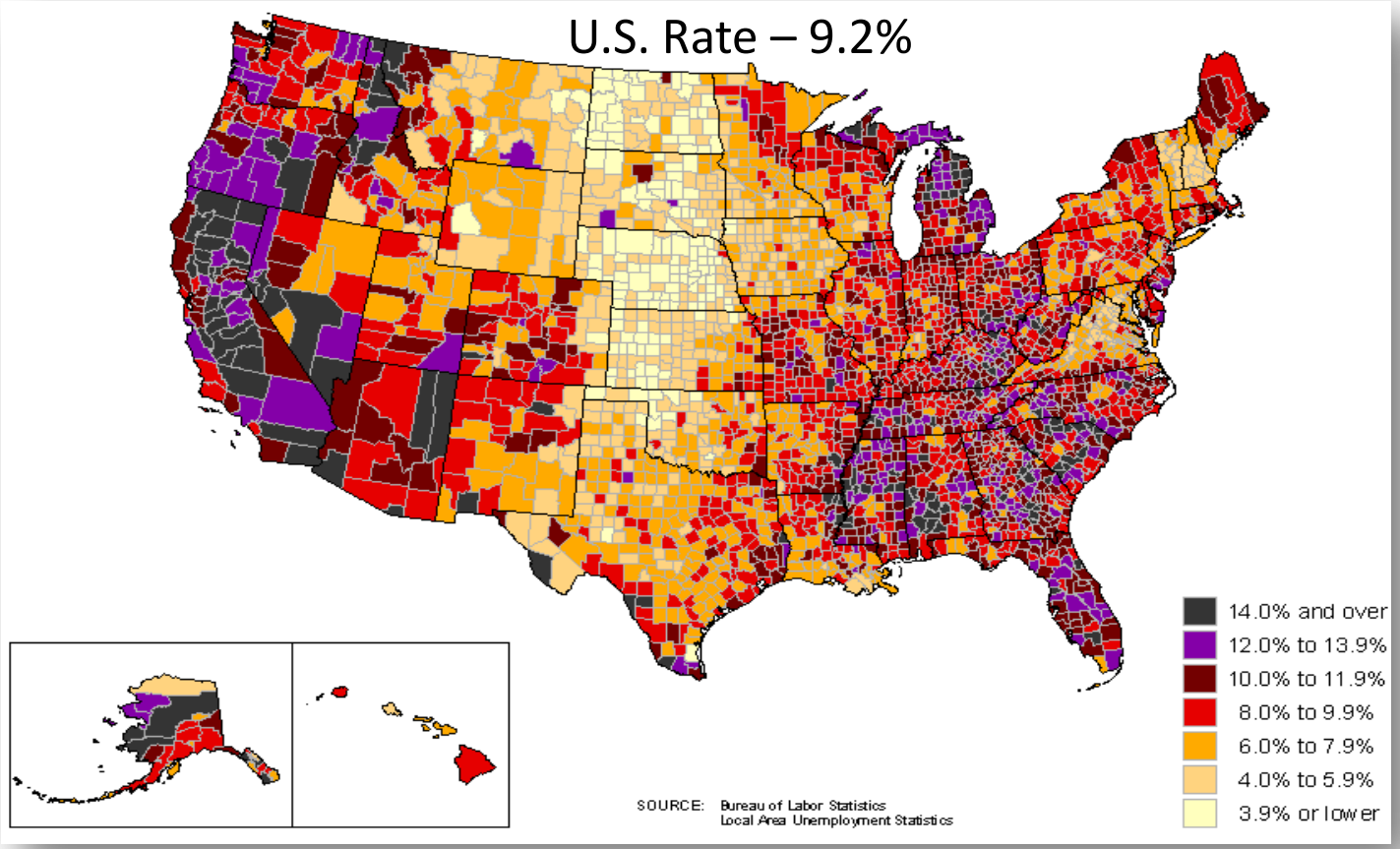



Source: US Bureau of Labor Statistics

# Identify Geographic Effects

Data available at various geographic levels

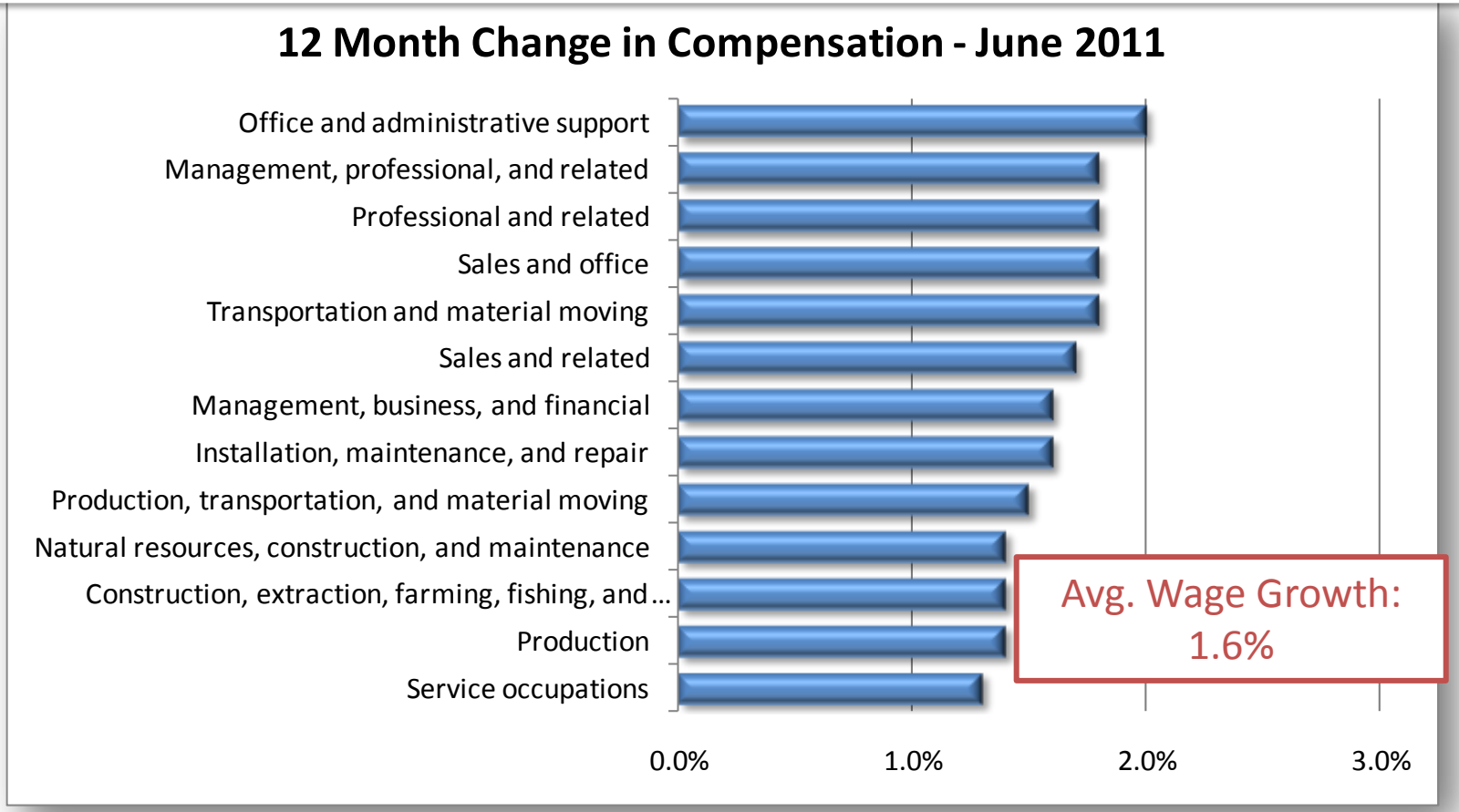
**Unemployment Rate by County – June 2011**




Source: US Bureau of Labor Statistics

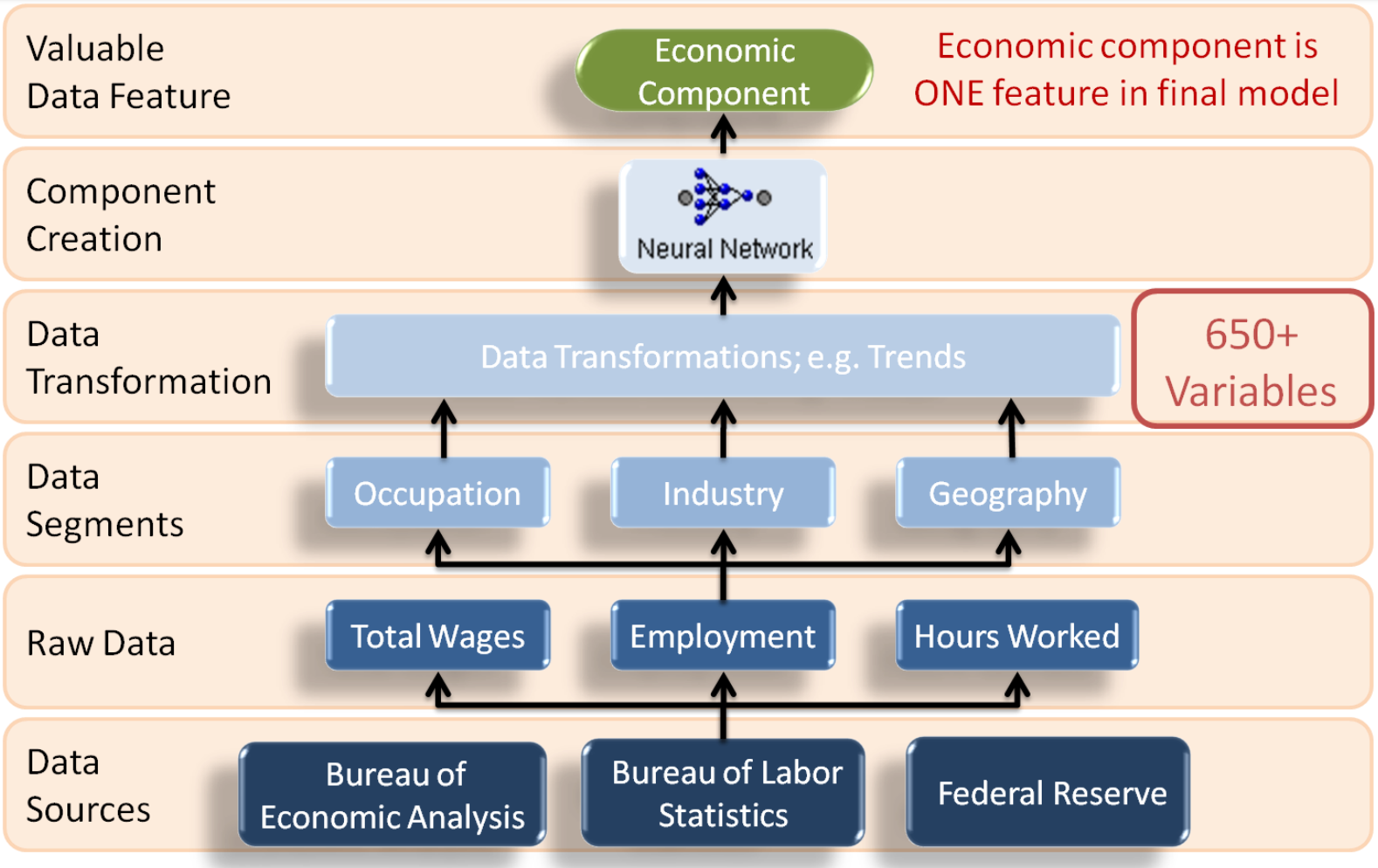
# Wage Growth by Occupation

## Wage growth also varies, adding additional insights

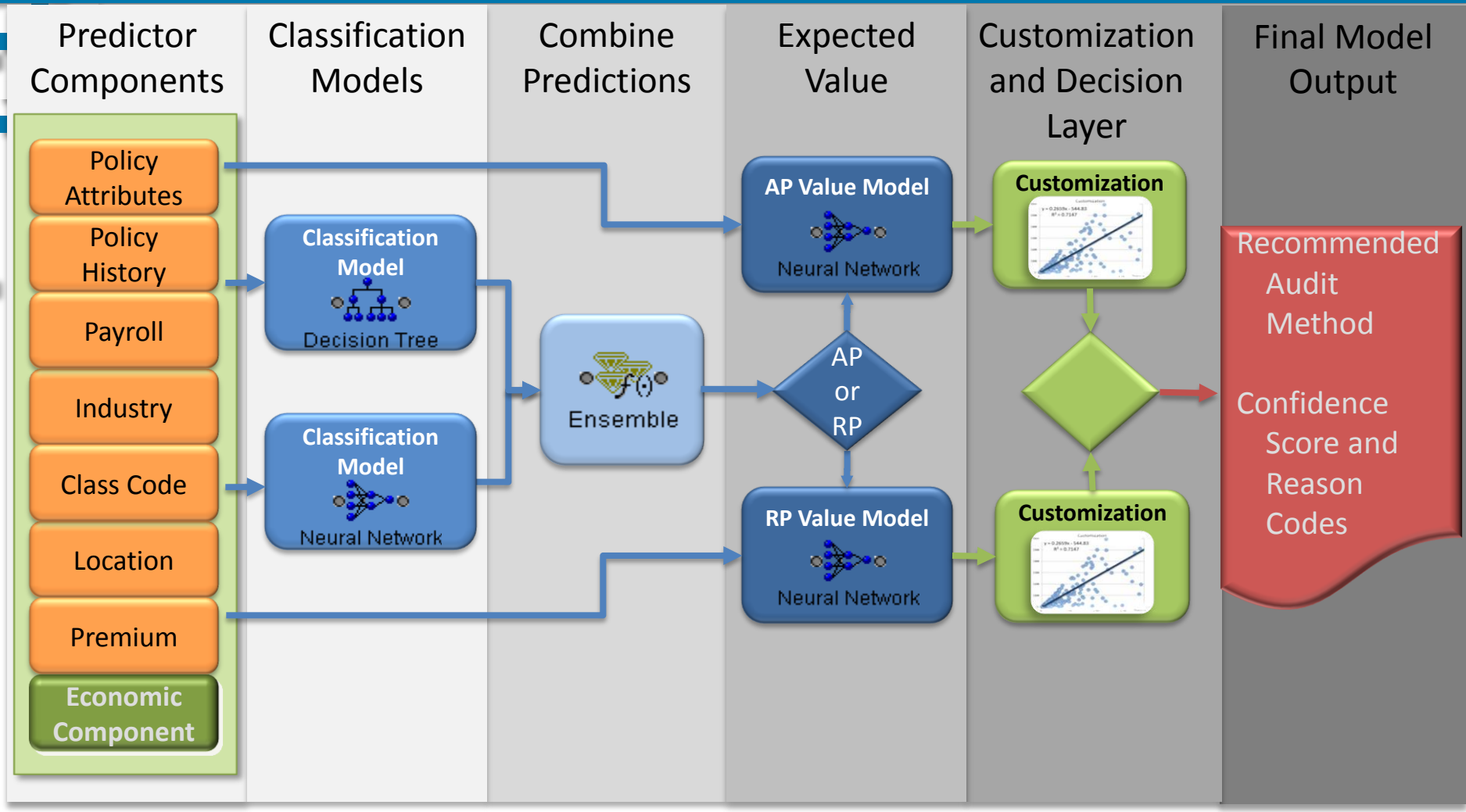



Source: US Bureau of Labor Statistics

# Economic Component



# Audit Selection Model Architecture



# Summary

- Better business decisions are the key to success
  - Analytics can shape decisions, but requires good data
- Raw data is often not very useful
  - Takes time, talent, and tools to extract valuable information from data
- Using packaged *analytic objects* can help fuel innovations quickly and cost-effectively

# Questions?

**Rama Duvvuri, FCAS CPCU**
**Vice President - Analytics**
**ISO Innovative Analytics**
**rduvvuri@iso.com**