ELKE RUNDENSTEINER

Professor, Computer Science
Director, Data Science
Worcester Polytechnic Institute
rundenst@cs.wpi.edu
Oct 2014

Thank you to NSF #1018443, #1018443, #097017, HP Inv., WPI/UMASS, IBM, CRA.
Have you heard of DATA SCIENTISTS?

HELLO
WPI DATA SCIENCE

SMART COLLEGE KIDS LIKE YOU WHO FIND PATTERNS IN DATA
THE HUMAN MIND IS CLEVER AT SEEING PATTERNS IN THINGS...

BUT A TRUE SCIENTIST DOESN’T JUST RELY ON HOW THINGS LOOK...
SCIENTISTS COULD USE A SIMPLE METHOD TO FIND RESULTS...
WE USE A TRIFECTA APPROACH TO TRAIN OUR DATA SCIENTISTS.

Also known as a triple threat!
WAIT!
WHERE DO ALL THE NUMBERS OR DATA COME FROM?
EVERY TIME YOU LOG ON, PLUG IN OR CLICK BUY, YOUR ‘DATA’ IS SAVED.

EVERY PURCHASE RECORDED AND SHARED

YOUR TWEETS ARE STORED

SOCIAL MEDIA KEEPS YOUR EVERY MOVE

MEDICAL RECORDS ARE STORED ELECTRONICALLY

EVERY SEARCH IS SAVED
ALL OF THAT INFORMATION OR DATA IS COLLECTED AND STORED. SO MUCH DATA THAT IT’S NOW IN PETABYTES ($10^{15}$)
THAT’S HOW COMPANIES TRACK YOUR PURCHASES AND ADVERTISE WHAT YOU LIKE ON YOUR Facebook PAGE.
AND HOW LIFE INSURANCE COMPANIES KNOW WHO TO INSURE.
AND HOW GOOGLE KNOWS WHAT YOU’RE SEARCHING...
WPI DATA SCIENCE

PRACTICALLY INFINITE AMOUNT OF DATA FROM ALL OVER THE WORLD BEING STORED.
WPI DATA SCIENCE

SO HOW DOES A DATA SCIENTIST MAKE SENSE OF IT ALL?
WPI DATA SCIENCE

We combine COMPUTER SCIENCE skills with MATHEMATICS, and BUSINESS skills, and a DATA SCIENTIST can make SENSE OF DATA!
WHERE DO I WORK AS A DATA SCIENTIST?
ANYWHERE AND EVERYWHERE

Healthcare Companies
Social Media—Google, Yahoo, Yahoo, Bing, FB
Gaming and Video - FUN
Education – higher and lower
Trains, Planes and Automobiles
All transit companies – World wide
Telecommunications – world wide
Security Companies – I Spy...
Banks and Brokerage Firms - NYC!!
Target – Gap – All retail stores
Bottom line...
So what is a Data Scientist paid?

DATA SCIENCE NEWS

ROUNDUP:

BECOMING A PROFESSION AT $300/HOUR –

Forbes.com
Are there jobs out there?

BIG DATA SCIENTISTS GET 100 RECRUITER EMAILS A DAY – Networkworld.com
Big Data - Big Opportunity

- New startup opportunities
- Wide range of companies looking for DS specialists
- 100% increase in jobs in Northeast U.S. alone
- 150,000 – 200,000 new jobs in analytics ANNUALLY

by MGI and McKinsey's Business Technology Office
WPI DATA SCIENCE

https://www.facebook.com/pages/WPI-DATA-Science/

DATA SCIENTIST...
SEXIEST JOB OF THE 21ST CENTURY.
Harvard Business Review

NOW MOBILE ANALYTICS STARTUPS ARE HIRING DATA SCIENTISTS.
HERE'S WHY – VentureBeat.com

BIG DATA: CAREER OPPORTUNITIES ABOUND IN TECH'S HOTTEST FIELD - Mashable

BIG DATA SCIENTISTS GET 100 RECRUITER EMAILS A DAY – Networkworld.com

All recent news articles posted on our FB page, written by industry leaders!
WPI DATA SCIENCE

IT’S THE SEXIEST JOB OF THE 21st CENTURY!*  

MY
RESEARCH
PROJECTS
MATTERS:
Economic Analytics Dashboard
For Massachusetts
Massachusetts Technology, Talent and Economy Reporting System: MATTERS

For Massachusetts High Tech Council
By WPI Team composed of over 10 students including Ramoza Ashan, Rodica Neamtu, and Caitlin Kuhlman, and many others
Project Goals

• Create and host an analytics platform that:
  – Represents an *integrative data* resource on high fidelity cost and talent competitive metrics,
  – Provides ease of access via *web-based* dashboard
  – Offers *data-driven analysis* capabilities supported by descriptive and predictive modeling,

• to:
  – help MHTC advocate for Massachusetts becoming a state attractive for business
MATTERS Overview

Data Sources

Acquisition/Integration

Data Pipeline

Data warehouse

Querying & reporting Interface

Advanced Analytics Engine

Visual Analytics Interfaces

Data Store

Users

5/9/2014
Data Metrics

1. State and Local Tax Burden "per capita" and "% of personal income"

2. Economy: Total Employment

3. Economy: Tech Employment

4. Economy Unemployment Rate

5. Talent Development Metrics

6. Unemployment Insurance Payroll Tax
Challenges with Data Sources

- Diversity of data sources & formats
- Excel files containing unstructured text
- Data not available for contiguous years
- Inconsistent data representations
- Some metrics composed across multiple sources
- Data must be transformed to be integrated
- Sources update data sporadically
- Data extraction is a complex custom process
## Data Cleaning: Uniformity & Consistency

### Table 1: Wages Subject to Tax, Minimum & Maximum Rates

<table>
<thead>
<tr>
<th>Year</th>
<th>State</th>
<th>Wages Subject to Tax</th>
<th>Minimum Rate</th>
<th>Maximum Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>MA</td>
<td>$14,000</td>
<td>1.26%</td>
<td>12.27%</td>
</tr>
<tr>
<td>2013</td>
<td>NH</td>
<td>$14,000</td>
<td>2.60%</td>
<td>7.00%</td>
</tr>
<tr>
<td>2013</td>
<td>NY</td>
<td>$8,500</td>
<td>0.90%</td>
<td>8.90%</td>
</tr>
<tr>
<td>2013</td>
<td>OH</td>
<td>$9,000</td>
<td>0.70%</td>
<td>9.10%</td>
</tr>
</tbody>
</table>

### Table 2: State Names Conversion

- **State name initials**
- **Numbers with signs**
- **Percentages**
- **Numbers without signs**
- **Decimals**

### Diagram:

- **Full state names**
- **Numbers with signs**
- **Percentages**
- **Numbers without signs**
- **Decimals**

### Table 3: Wages Subject to Tax, Minimum & Maximum Rates

<table>
<thead>
<tr>
<th>Year</th>
<th>State</th>
<th>Wages Subject to Tax</th>
<th>Minimum Rate</th>
<th>Maximum Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>Massachusetts</td>
<td>14000</td>
<td>0.0126</td>
<td>0.1227</td>
</tr>
<tr>
<td>2013</td>
<td>New Hampshire</td>
<td>14000</td>
<td>0.026</td>
<td>0.07</td>
</tr>
<tr>
<td>2013</td>
<td>New York</td>
<td>8500</td>
<td>0.009</td>
<td>0.089</td>
</tr>
<tr>
<td>2013</td>
<td>Ohio</td>
<td>9000</td>
<td>0.007</td>
<td>0.091</td>
</tr>
</tbody>
</table>
### State Government Tax Collections: 2012

<table>
<thead>
<tr>
<th>Government</th>
<th>Total Taxes</th>
<th>Property Taxes</th>
<th>Sales and Gross Receipts Taxes</th>
<th>License Taxes</th>
<th>Income Taxes</th>
<th>Other Taxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>798,221,675</td>
<td>13,104,336</td>
<td>377,541,72</td>
<td>54,090,961</td>
<td>322,654,16</td>
<td>30,830,487</td>
</tr>
<tr>
<td>Alabama</td>
<td>9,049,294</td>
<td>321,530</td>
<td>4,626,357</td>
<td>517,676</td>
<td>3,430,690</td>
<td>153,041</td>
</tr>
<tr>
<td>Alaska</td>
<td>7,049,398</td>
<td>215,407</td>
<td>248,432</td>
<td>135,055</td>
<td>663,144</td>
<td>5,787,360</td>
</tr>
<tr>
<td>Arizona</td>
<td>12,973,265</td>
<td>754,428</td>
<td>8,066,124</td>
<td>370,222</td>
<td>3,741,713</td>
<td>40,778</td>
</tr>
<tr>
<td>Arkansas</td>
<td>8,284,500</td>
<td>1,008,707</td>
<td>3,982,832</td>
<td>355,768</td>
<td>2,805,985</td>
<td>131,208</td>
</tr>
<tr>
<td>California</td>
<td>115,089,654</td>
<td>2,079,878</td>
<td>41,341,188</td>
<td>8,658,041</td>
<td>62,973,435</td>
<td>37,112</td>
</tr>
</tbody>
</table>

#### SA1-3 Personal income summary

<table>
<thead>
<tr>
<th>Area</th>
<th>Description</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>California</td>
<td>Personal income (thousands of dollars)</td>
<td>1.596E+09</td>
<td>1.536E+09</td>
<td>1579148473</td>
<td>1683203700</td>
<td>1768039281</td>
</tr>
<tr>
<td>California</td>
<td>Population (persons) 1/</td>
<td>36604337</td>
<td>36961229</td>
<td>37334410</td>
<td>37683933</td>
<td>38041430</td>
</tr>
<tr>
<td>California</td>
<td>Per capita personal income (dollars) 2/</td>
<td>43609</td>
<td>41569</td>
<td>42297</td>
<td>44666</td>
<td>46477</td>
</tr>
<tr>
<td>Colorado</td>
<td>Personal income (thousands of dollars)</td>
<td>212243112</td>
<td>206422648</td>
<td>210607673</td>
<td>226031916</td>
<td>237461449</td>
</tr>
<tr>
<td>Colorado</td>
<td>Population (persons) 1/</td>
<td>4889730</td>
<td>4972195</td>
<td>5048472</td>
<td>5116302</td>
<td>5187582</td>
</tr>
<tr>
<td>Colorado</td>
<td>Per capita personal income (dollars) 2/</td>
<td>43406</td>
<td>41515</td>
<td>41717</td>
<td>44179</td>
<td>45775</td>
</tr>
<tr>
<td>Connecticut</td>
<td>Personal income (thousands of dollars)</td>
<td>198981824</td>
<td>191312735</td>
<td>197839341</td>
<td>207161731</td>
<td>214297085</td>
</tr>
<tr>
<td>Connecticut</td>
<td>Population (persons) 1/</td>
<td>3545579</td>
<td>3561807</td>
<td>3576616</td>
<td>3586717</td>
<td>3590347</td>
</tr>
<tr>
<td>Connecticut</td>
<td>Per capita personal income (dollars) 2/</td>
<td>56121</td>
<td>53712</td>
<td>55315</td>
<td>57758</td>
<td>59687</td>
</tr>
</tbody>
</table>
Explosion of Formats: Tame the Diversity?

Question: How support diversity?
Tame the Diversity: Towards a Unified Schema

Proposal: Unified Schema To the Rescue. *(diversity, generality, extensibility)*
MATTERS Dashboard: Make an Impact on your Community
Learn Technologies

- Framework:
- Data stores:
- Web charting packages:
- Shared development repository:
- Shared document management:
- Project management:
UMASS Medical Center: Tracking for Infection Control

With Di Wang, Prof. Ellison, Mo Liu, Medhabi Ray, etc.
Health Care Application: Infection Control

Data Sources

- RFID Input
- RFID Input
- RFID Input
- RFID Input
- RFID Input
- RFID Input
- RFID Input

Put on mask for H1N1 contagious patients

Wash your hands before touching next patients

Put on surgical gloves

Track workers

Detect hygiene violations

Real-time Reminder:
Put on surgical gloves

Aggregate statistics for a hospital

Complex Event Analysis:
Large number of related complex CEP queries over stream data at different abstractions

Patent filed, EPTS Principles Award, Deploy & Clinical Trial at UMASS Memorial Hospital
Complex Event Processing

– Event Stream: Continuous stream of event instances
– Sequence patterns: matched against event stream

Example:

\[
\text{PATTERN SEQ(OpRoom1, ! Disinfection Area, OpRoom2)[id] WITHIN 5 minutes}
\]
Q1: SEQ(Contaminated Areas, NOT Disinfection Area, Operating Rooms) WITHIN 1 hour

Q2: SEQ(Contaminated Areas, NOT D1, Clean Rooms, Operating Rooms) WITHIN 1 hour

Q3: SEQ(Contaminated Areas, NOT Disinfection Area, OR 1) WITHIN 1 hour

Q4: SEQ(Contaminated Areas, NOT Disinfection Area, Clean Rooms, Operating Rooms) WITHIN 1 hour

Q5: Break Rooms, NOT Disinfection Area, Clean Rooms, OR 1) WITHIN 1 hour

- Pattern-drill-down
- Pattern-roll-up
- Concept-drill-down
- Concept-roll-up
Emotex: Emotion Detection in Social and Smartphone Sensors

With Maryam Hasan, Prof. Agu and others
Microblog tools such as Twitter express their feelings and opinions in the form of short text messages.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Tweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>• So many weddings coming up how exciting is that</td>
</tr>
<tr>
<td></td>
<td>• Excited to see him in Texas in two weeks</td>
</tr>
<tr>
<td>Relaxed</td>
<td>• I feel so at peace right now</td>
</tr>
<tr>
<td></td>
<td>• The sound of rain always puts me to sleep</td>
</tr>
<tr>
<td>Stressed</td>
<td>• Presentation? I'm feeling like I'm waiting to get an injection</td>
</tr>
<tr>
<td></td>
<td>• Seriously stressed over this final.</td>
</tr>
<tr>
<td>Depressed</td>
<td>• RIP Grandpa, you will be missed.</td>
</tr>
<tr>
<td></td>
<td>• I'm just so #depressed and on the verge of crying</td>
</tr>
</tbody>
</table>

Objective: Learn about Emotional State of the Author of a Message.
Emotional States: Circumplex Model

Circumplex model (Posner and Russell 2005).

Unhappy Active

Low Pleasure

High Activation

Unhappy Inactive

Low Activation

High Pleasure

Happy Active

Happy Inactive
Challenges of Analyzing Microblogs

- What are Microblogs:
  - short terse textual messages
  - casual style of expression
  - grammatical and spelling errors

- Examples of Microblogs:
  - I'm soo happenyy I have such wonderful people in my life!
  - Its always a good feeling to know dat the person ur friend has a crush on, actually likes u.

- Challenges:
  - Requires labeled data required for training.
  - Must handle high dimensional and sparse feature vectors
  - Use Twitter #hash-tags as noisy labels
Model of EMOTEX

Collecting Labeled Data

Training Multi-class Classifiers

Training Data

Extracted Emotions

Test Data

Task 1

Task 2

Task 3

Selecting Features

Feature Vectors

Figure 2- Model of EMOTEX
**Data Cleaning**

**Data Cleaning** → **Resolving Conflicts** → **Removing Hash-tags**

**Replace:**
- Http links with **URL**, e.g. @Marilyn
- User Names with **UID**, e.g. @Marilyn
- Repeated characters with two characters, e.g. happyyyyyy

**Resolves:**
- Hash-tag conflicts, e.g. #sleepy #happy
- Emoticon conflicts, e.g. :) :-(():
- Tag-Emoticon conflicts, e.g. :(( #excited )

**Removes:**
- Hash-tags from the end of Tweets
Supervised Learning Approach

• Represents each message by a D-dimensional feature vector:
  \[ F = (f_1, \ldots, f_D) \in \mathbb{R}^D \]

• Mark each message by a label

• Train learning algorithm on labeled messages

Selecting Features

Figure 4- Feature Selection
# Emoticon Features

<table>
<thead>
<tr>
<th>Category</th>
<th>Emoticons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy Emoticons</td>
<td>:) :) =) :) ] :p ;p :D ;D :&gt; :) :D :&gt; :3</td>
</tr>
<tr>
<td></td>
<td>:-) ;-) ^) :o) :~) ^) :o) :' ) :D :&gt;</td>
</tr>
<tr>
<td>Sad Emoticons</td>
<td>:( =(:-( :^)( :o( :^)( :!( :&lt;-</td>
</tr>
<tr>
<td>Angry Emoticons</td>
<td>&gt;:S &gt;:{ &gt;: x-@ :@ :-@ :-/ :-\</td>
</tr>
<tr>
<td>Afraid/Surprised Emoticons</td>
<td>:-o :-O o_O O_o $</td>
</tr>
<tr>
<td>Sleepy Emoticons</td>
<td>__ __ <del>__</del></td>
</tr>
</tbody>
</table>

Table 2- Emoticon Features
Unigram Features

- Single Words in our training data such as:
  - excited, sad, hope, hate,…

- Problem:
  - Huge number of unigrams in training data
  - Sparse feature vector of each tweet

- Solution:
  - Using emotional unigrams from Emotion lexicons: LIWC (Linguistic Inquiry and Word Count)

James W. Pennebaker, Roger J. Booth, and Martha E. Francis, University of Texas, 2007, http://www.liwc.net/
Twitter Data

200,000 tweets collected before 2014 and after 2014.

Figure 5 - Distribution of the emotions during new year vacation and after it
# Results: Classification accuracy of SVM, KNN, Naïve Bayes, Decision Tree

<table>
<thead>
<tr>
<th></th>
<th>Unigram</th>
<th>Unigram, Emoticon</th>
<th>Unigram, Punctuation</th>
<th>Unigram, Negation</th>
<th>All Features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SVM</strong></td>
<td>89.86</td>
<td>88.92</td>
<td>89.59</td>
<td>88.97</td>
<td>89.36</td>
</tr>
<tr>
<td><strong>Naïve Bayes</strong></td>
<td>86.27</td>
<td>86.40</td>
<td>86.61</td>
<td>86.91</td>
<td><strong>86.95</strong></td>
</tr>
<tr>
<td><strong>Decision Tree</strong></td>
<td>89.48</td>
<td>89.59</td>
<td>89.72</td>
<td>89.62</td>
<td><strong>89.93</strong></td>
</tr>
<tr>
<td><strong>KNN</strong></td>
<td>90.10</td>
<td>90.07</td>
<td>90.10</td>
<td></td>
<td>90.13</td>
</tr>
</tbody>
</table>

Table 4- Classification accuracy of different methods using different features
Conclusion

• Cool Science and Engineering to be done
• Data-driven projects are here to stay
• Learn something new – rewarding personally
• Impactful on community, economy, health...
Thank you to My Collaborators

Work produced in collaboration with students and colleagues, including Mo Liu, Di Wang, Medhabi Ray, Kara Greenfield, Tonje Stolpestad, Dick Ellison, Dan Dougherty, Yingmei Qi, Dazhi Zhang, Chetan Gupta, Ismail Ari, Song Wang, Abhay Mehta, Matt Ward, Di Yang, Abhishek Mukherji, Mohamed Eltabakh, Avani Shastri, Mani Murali, Karen Works, Chuan Lei, Lei Cao, Yingmei Qi, Prof. Agu, Maryam Hasan, Ramoza Ashan, Rodica Neamtu, and many others...

The credit for the work all goes to them! I am just the messenger!
THANK YOU