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.**

WPI DATA SCIENCE

Have you heard of DATA SCIENTISTS?

WPI DATA SCIENCE

SMART COLLEGE KIDS LIKE YOU
WHO FIND PATTERNS IN DATA

WPI DATA SCIENCE

THE HUMAN MIND IS CLEVER AT SEEING PATTERNS IN THINGS…



BUT A TRUE SCIENTIST DOESN'T JUST RELY ON HOW THINGS LOOK…

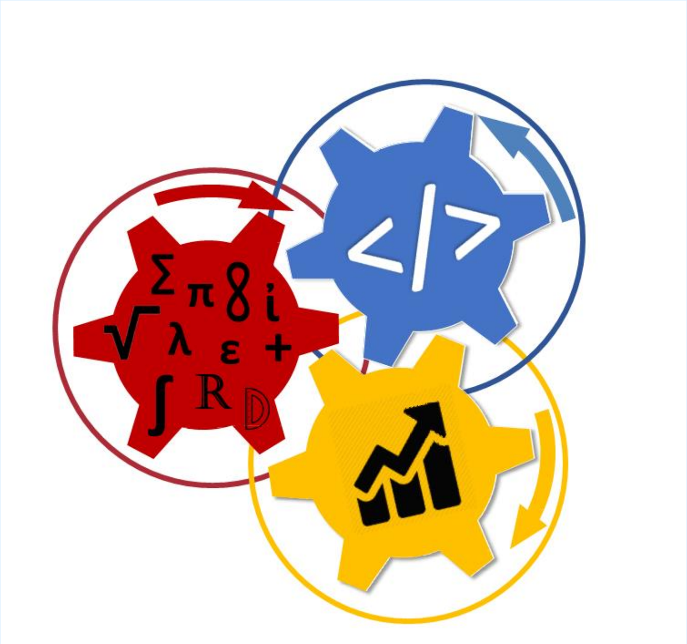
WPI DATA SCIENCE

SCIENTISTS COULD USE A SIMPLE METHOD TO FIND RESULTS…

HE LOVES ME,
HE LOVES ME NOT

WPI DATA SCIENCE

WE USE A TRIFECTA APPROACH TO TRAIN OUR DATA SCIENTISTS.



Also known as a triple threat!

WPI DATA SCIENCE

WAIT!

WHERE DO ALL THE NUMBERS

OR *DATA*

COME FROM?

WPI DATA SCIENCE

EVERY TIME YOU LOG ON, PLUG IN OR CLICK BUY, YOUR 'DATA' IS SAVED.

EVERY PURCHASE RECORDED AND SHARED

YOUR TWEETS ARE STORED

MEDICAL RECORDS ARE STORED ELECTRONICALLY

SOCIAL MEDIA KEEPS YOUR EVERY MOVE

EVERY SEARCH IS SAVED

WPI DATA SCIENCE

ALL OF THAT INFORMATION OR *DATA* IS COLLECTED AND STORED.
SO MUCH DATA THAT IT'S NOW IN PETABYTES (10^15)

WPI DATA SCIENCE

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.

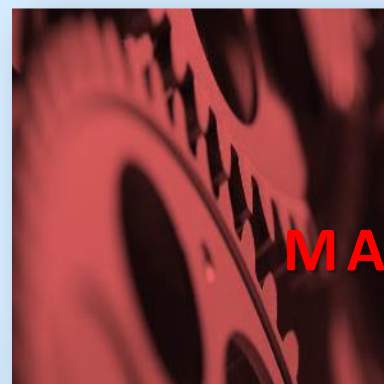
WPI DATA SCIENCE

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

WPI DATA SCIENCE

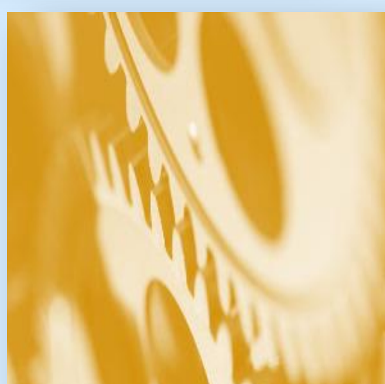
WE COMBINE COMPUTER SCIENCE SKILLS WITH MATHEMATICS, AND BUSINESS SKILLS, AND A DATA SCIENTIST CAN MAKE *SENSE* OF DATA!

COMPUTER SCIENCE



MATHEMATICS

BUSINESS



WPI DATA SCIENCE

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…

WPI DATA SCIENCE

SHOW ME
THE MONEY!

So what is a Data Scientist paid?

DATA SCIENCE NEWS ROUNDUP: BECOMING A PROFESSION AT $300/HOUR –

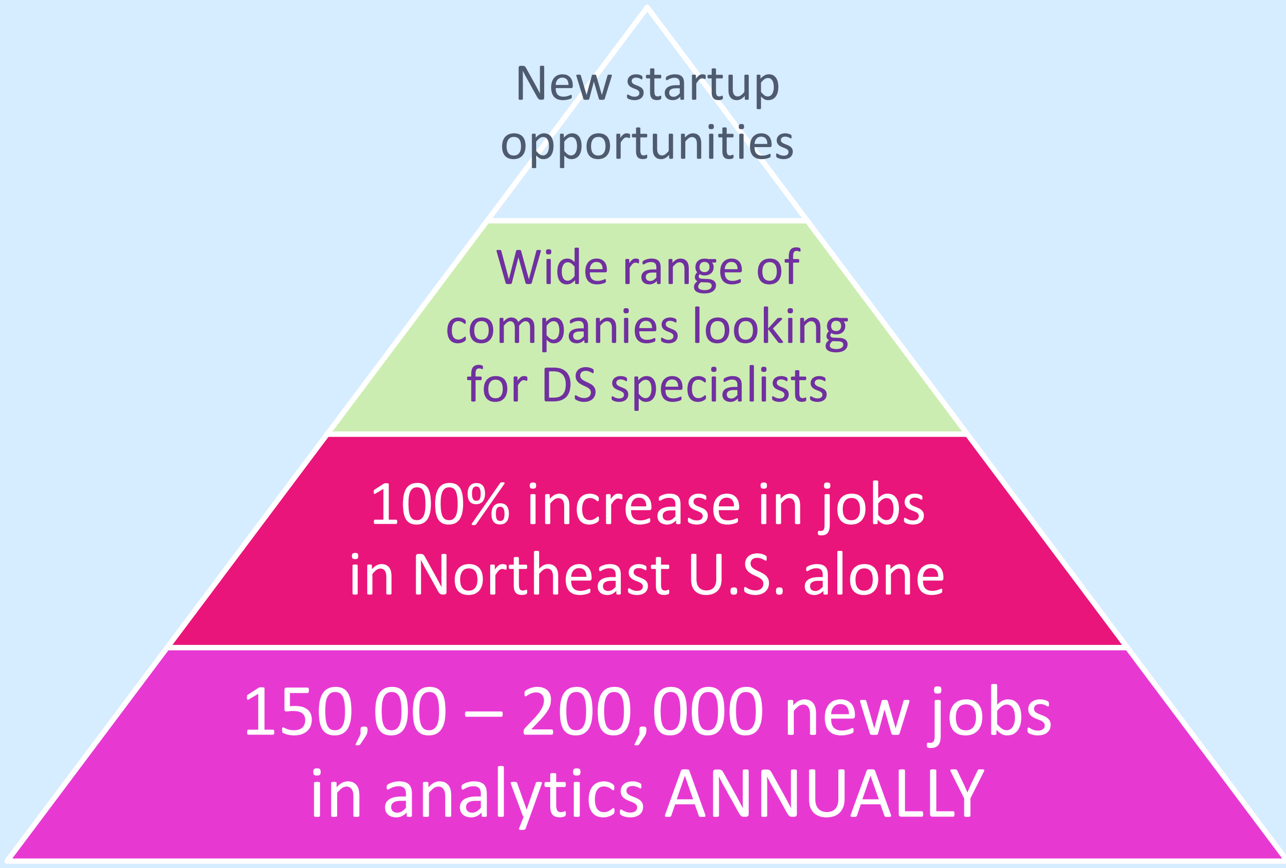
Forbes.com

WPI DATA SCIENCE

Are there jobs out there?

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

Big Data - Big Opportunity



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

DATA SCIENCE NEWS ROUNDUP: BECOMING A PROFESSION AT $300/HOUR –
Forbes.com

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!**

**** Harvard Business Review, Oct 2012.***

MY

RESEARCH

PROJECTS

MATTERS:
Economic Analytics Dashboard
For Massachusetts

Massachusetts Technology, Talent and Economy Reporting System: MATTERS

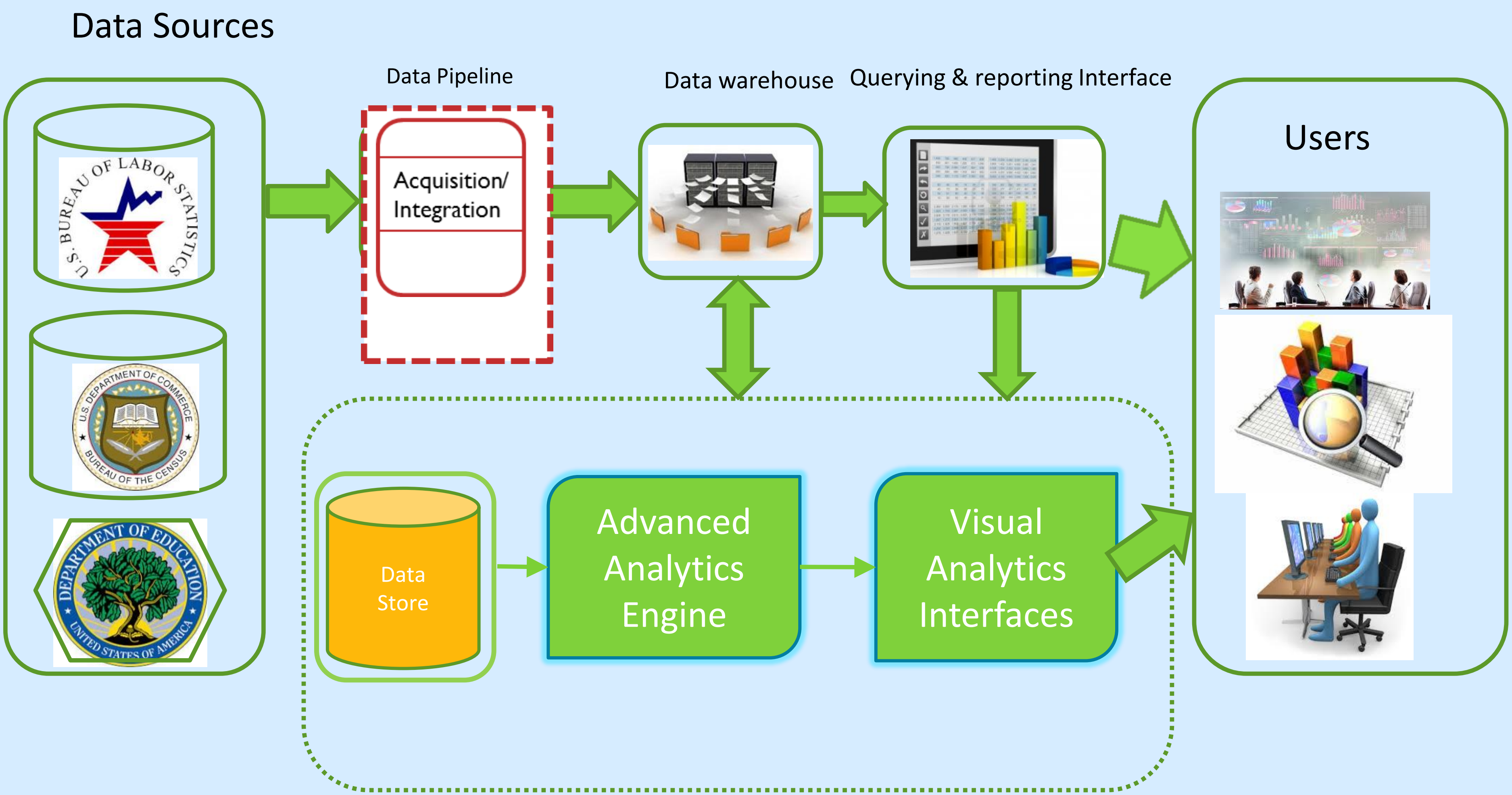
For Massachusetts High Tech Council

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 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

United States Census Bureau

U.S. Department of Commerce Bureau of Economic Analysis

Bureau of Labor Statistics

ies Institute of Education Sciences

NSF

tpc Tax Policy Center Urban Institute and Brookings Institution

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

Year	State	Wages Subject to Tax	Minimum Rate	Maximum Rate
2013	MA	$14,000	1.26%	12.27%
2013	NH	$14,000	2.60%	7.00%
2013	NY	$8,500	0.90%	8.90%
2013	OH	$9,000	0.70%	9.10%

State name initials → Full state names

Numbers with signs → Numbers without signs

Percentages → Decimals

Year	State	Wages Subject to Tax	Minumum Rate	Maximum Rate
2013	Massachusetts	14000	0.0126	0.1227
2013	New Hampshire	14000	0.026	0.07
2013	New York	8500	0.009	0.089
2013	Ohio	9000	0.007	0.091

Diversity of Data Sets

State Government Tax Collections: 2012

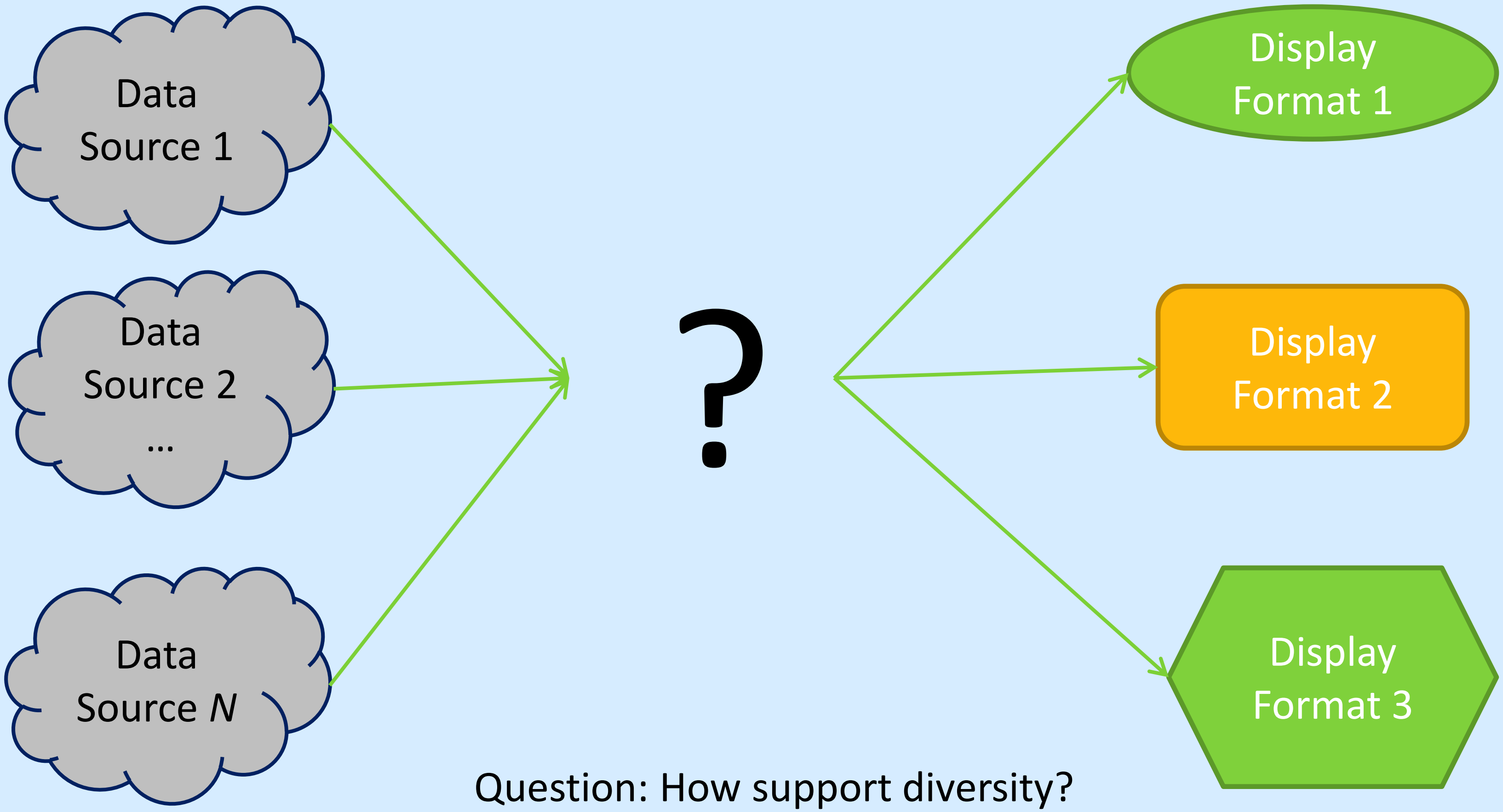
Government	Total Taxes	Property Taxes	Sales and Gross Receipts Taxes	License Taxes	Income Taxes	Other Taxes
United States	798,221,675	13,104,336	377,541,72	54,090,961	322,654,16	30,830,487
Alabama	9,049,294	321,530	4,626,357	517,676	3,430,690	153,041
Alaska	7,049,398	215,407	248,432	135,055	663,144	5,787,360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

SA1-3 Personal income summary

Bureau of Economic Analysis
State or DC

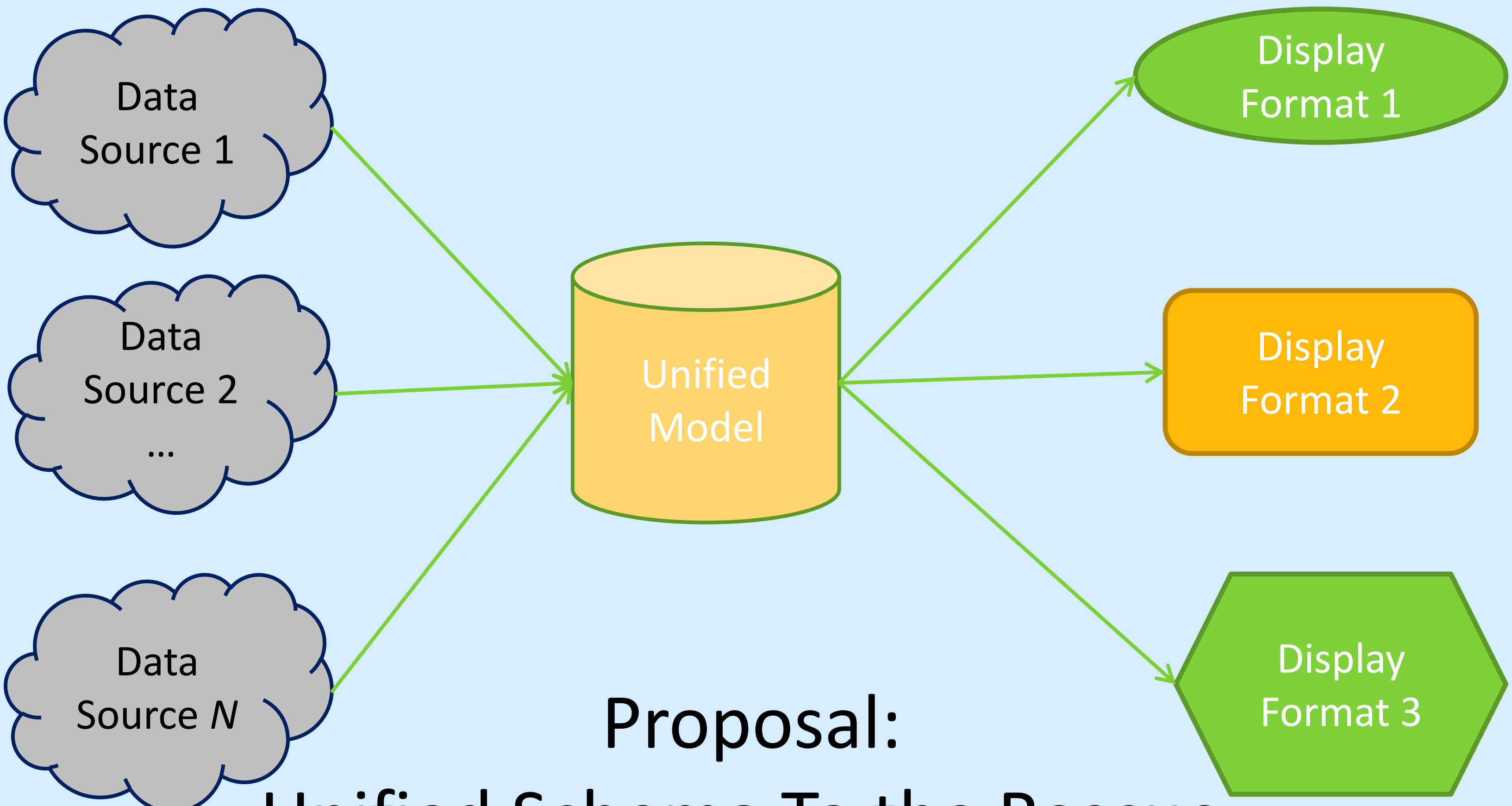
Area	Description	2008	2009	2010	2011	2012
California	Personal income (thousands of dollars)	1.596E+09	1.536E+09	1579148473	1683203700	1768039281
California	Population (persons) 1/	36604337	36961229	37334410	37683933	38041430
California	Per capita personal income (dollars) 2/	43609	41569	42297	44666	46477
Colorado	Personal income (thousands of dollars)	212243112	206422648	210607673	226031916	237461494
Colorado	Population (persons) 1/	4889730	4972195	5048472	5116302	5187582
Colorado	Per capita personal income (dollars) 2/	43406	41515	41717	44179	45775
Connecticut	Personal income (thousands of dollars)	198981824	191312735	197839341	207161731	214297085
Connecticut	Population (persons) 1/	3545579	3561807	3576616	3586717	3590347
Connecticut	Per capita personal income (dollars) 2/	56121	53712	55315	57758	59687

Explosion of Formats: Tame the Diversity ?



Question: How support diversity?

Tame the Diversity: Towards a Unified Schema

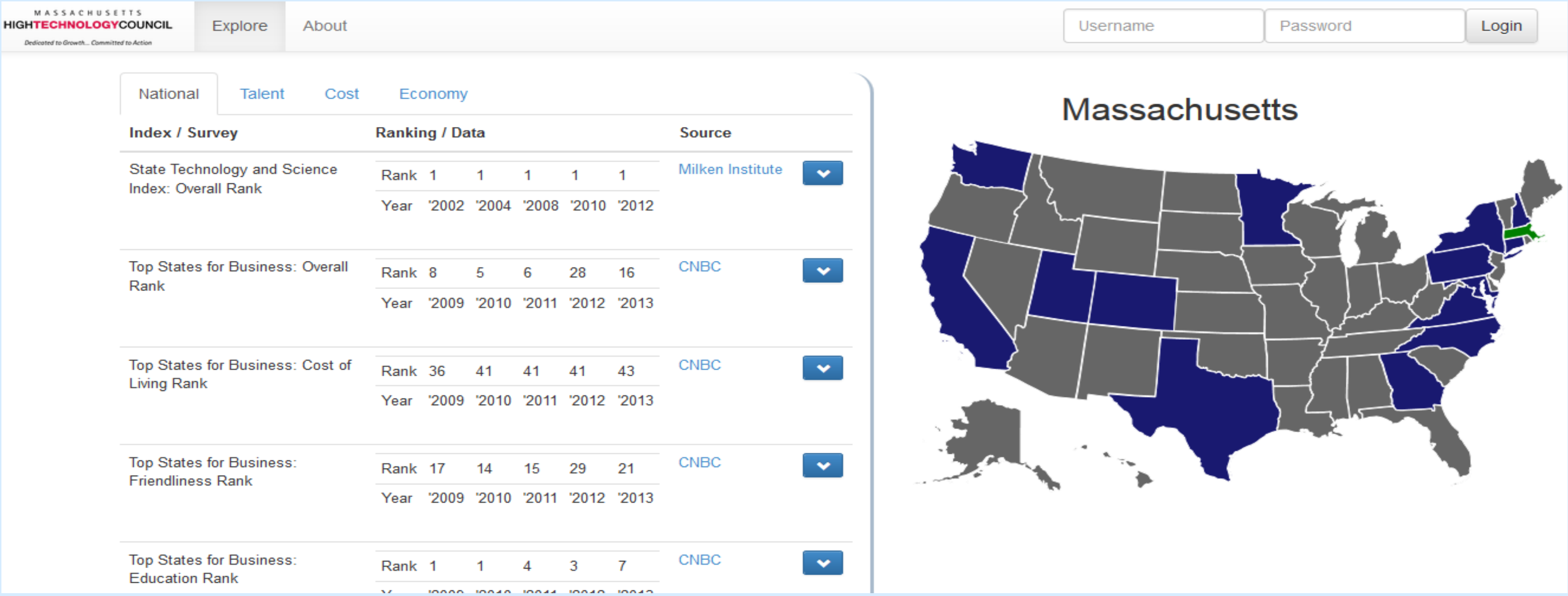


Proposal:

Unified Schema To the Rescue.

(diversity, generality, extensibility)

MATTERS Dashboard:
Make an Impact on your Community



MASSACHUSETTS HIGHTECHNOLOGYCOUNCIL
Proven Leadership in Action

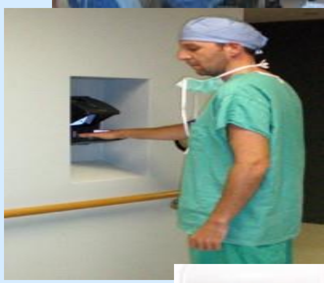
Learn Technologies

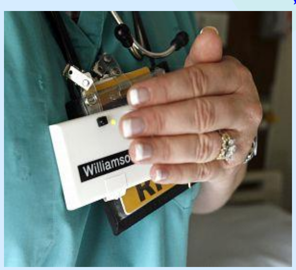
- Framework:
- Data stores:
- Web charting packages:
- Shared development repository:
- Shared document management:
- Project management:

UMASS Medical Center: Tracking for Infection Control

With Di Wang, Prof. Ellison, Mo Liu, Medhabi Ray, etc.

Health Care Application : Infection Control





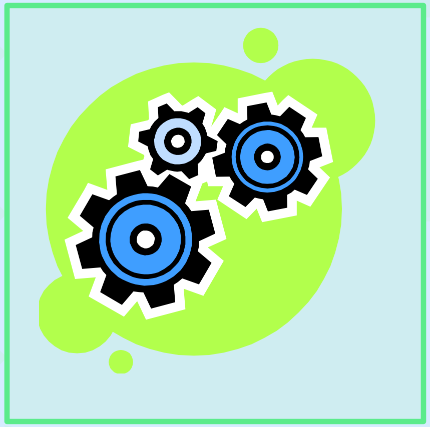
Data
Sources

RFID Input

RFID Input

RFID Input

RFID Input



Track workers

Detect hygiene violations

Put on mask for H1N1
contagious patients

Wash your hands before
touching next patients

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

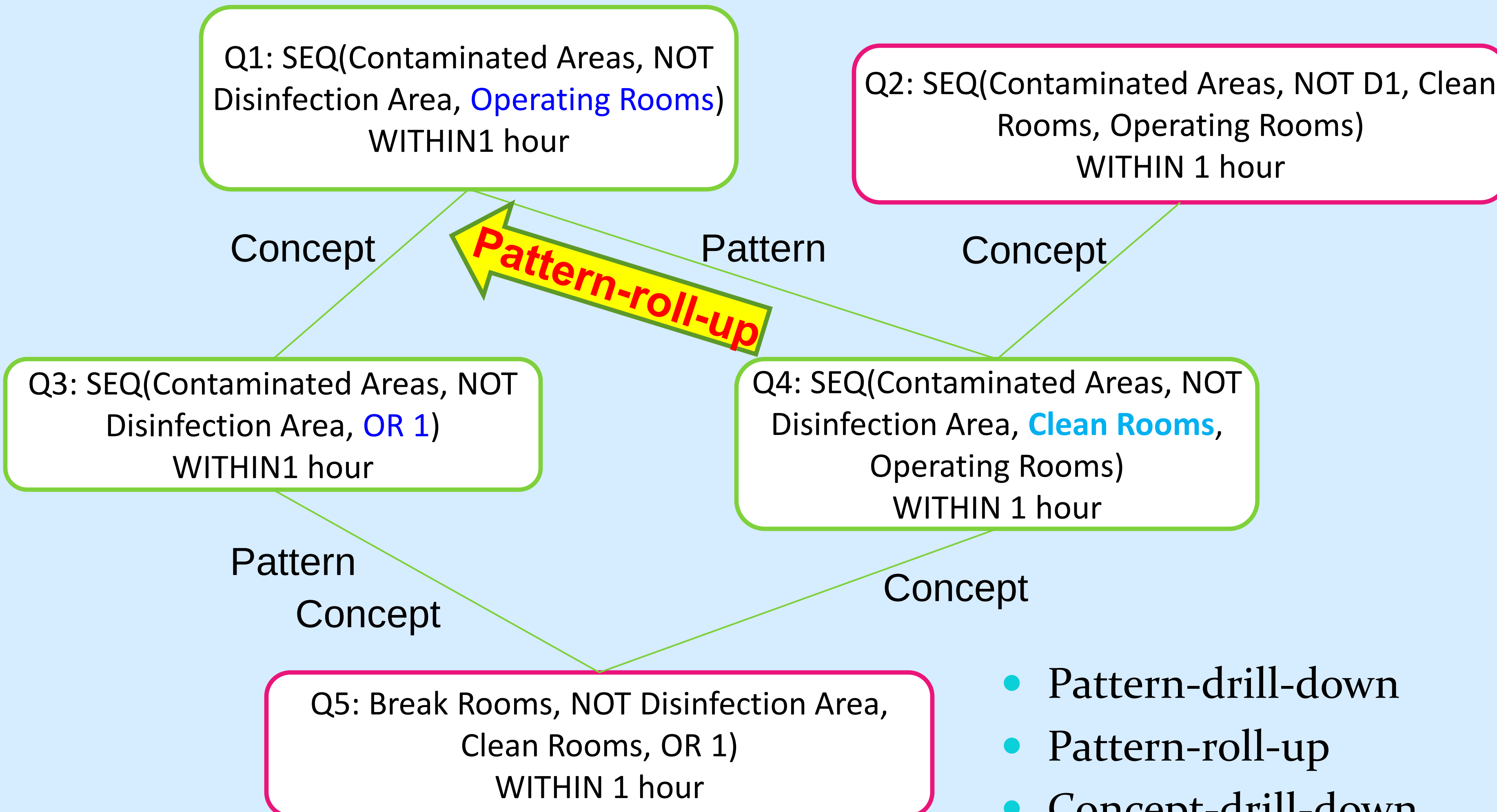
- Event Stream: Continuous stream of event instances
- Sequence patterns: matched against event stream

Example:

PATTERN SEQ(OpRoom1, ! Disinfection Area, OpRoom2)[id]
WITHIN 5 minutes

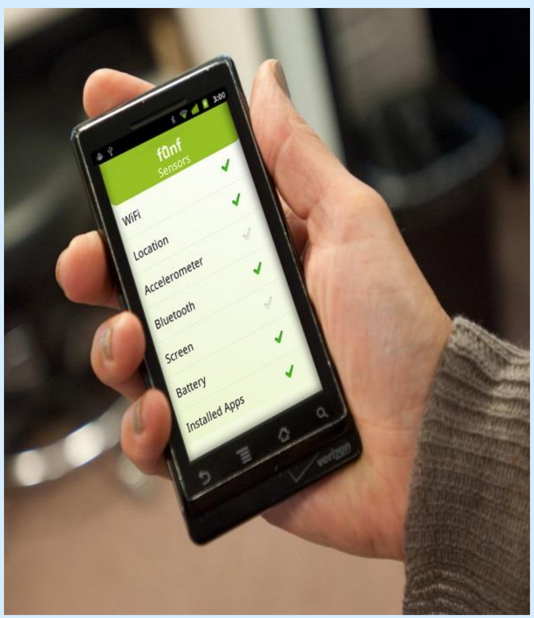
Time

Ecube Hierarchy



- Pattern-drill-down
- Pattern-roll-up
- Concept-drill-down
- Concept-roll-up

Emotex : Emotion Detection in Social and Smartphone Sensors



Tweets
#Hashtags

With Maryam Hasan, Prof. Agu and others

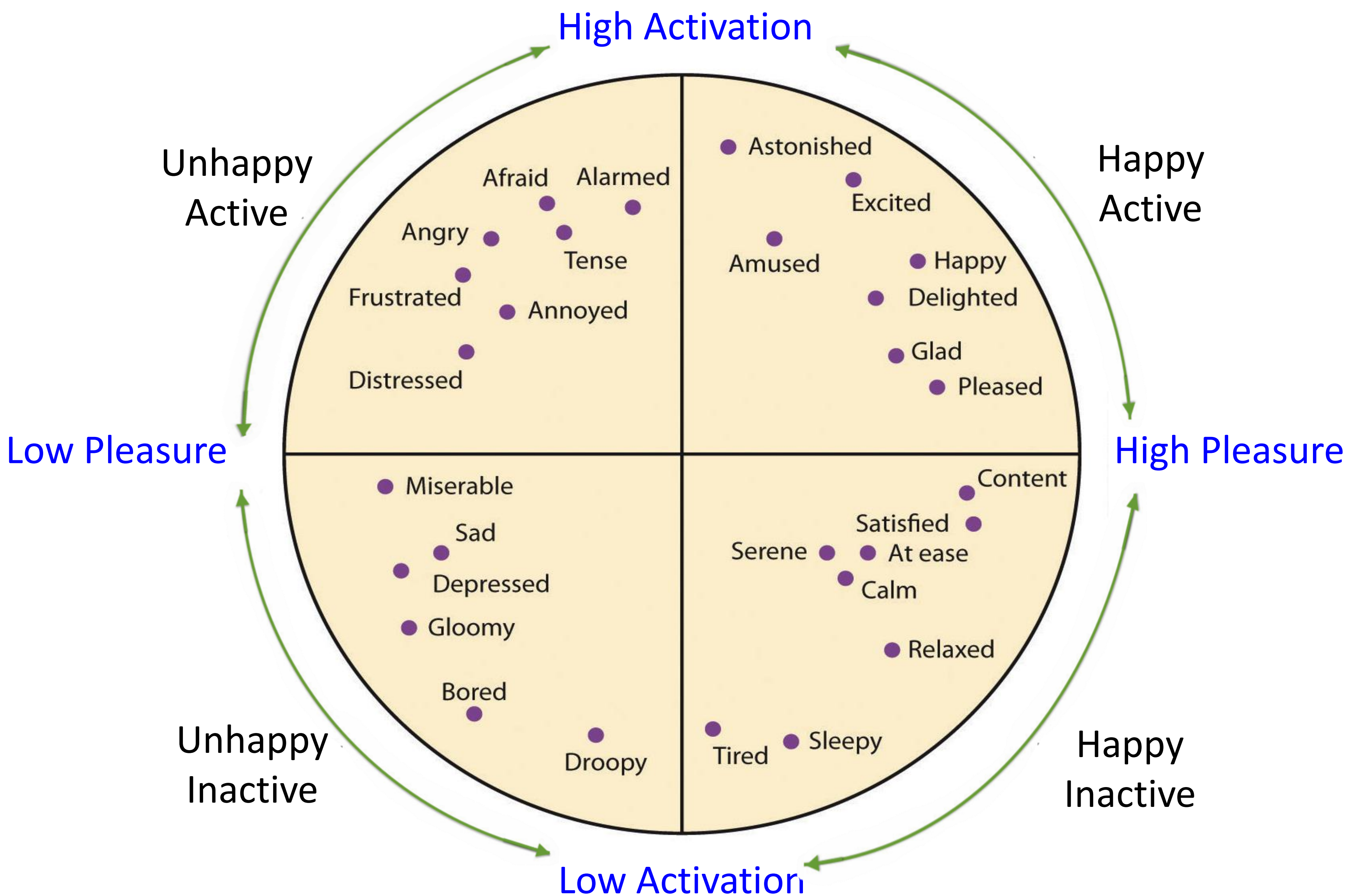
Textual Sensors: Twitter

- Microblog tools such as Twitter express their feelings and opinions in the form of short text messages.

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

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

Emotional States: Circumplex Model



Circumplex model (Posner and Russell 2005).

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 happyyy 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

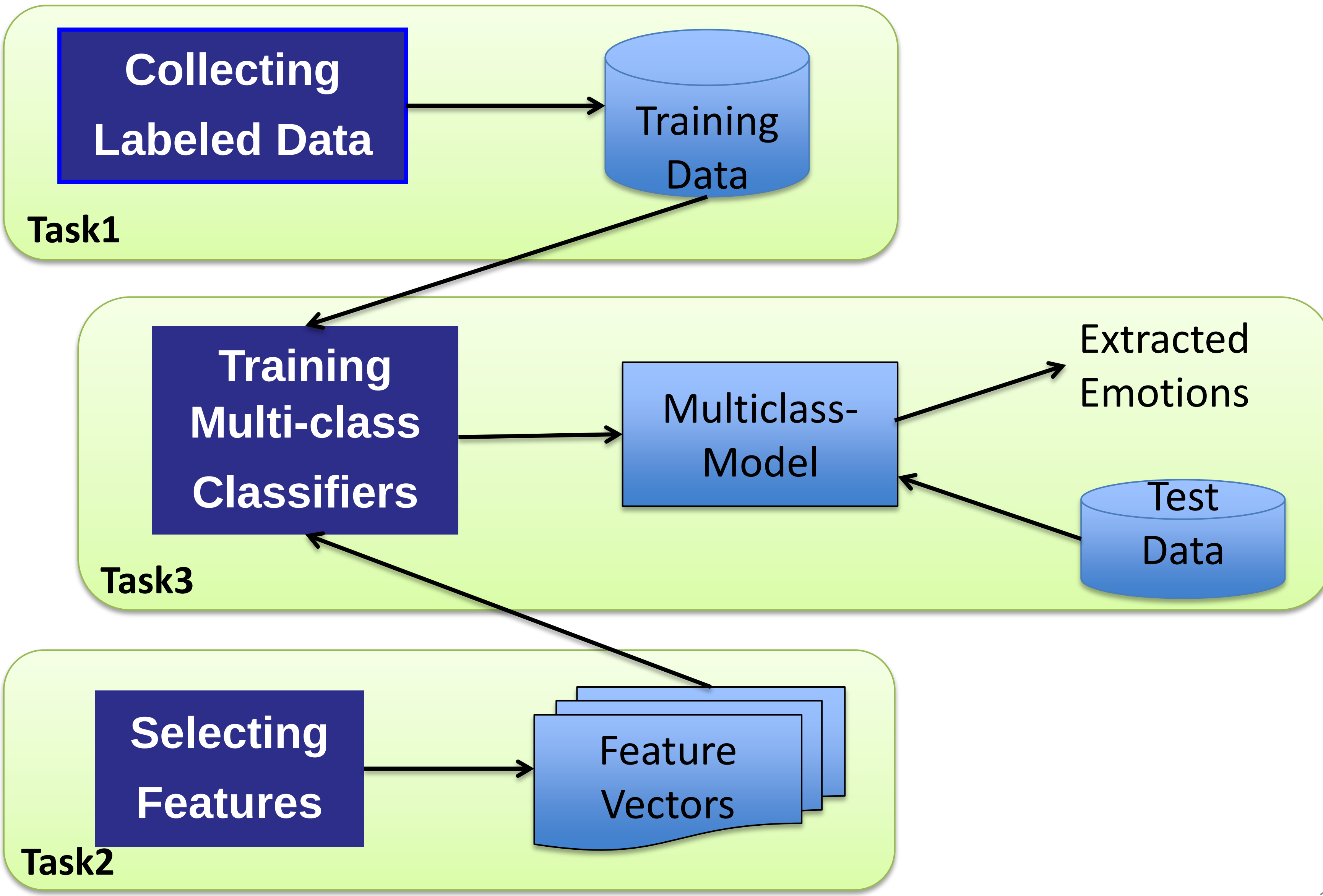


Figure 2- Model of EMOTEX

Data Cleaning

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

Replace:

- Http links with URL,
- User Names with UID (e.g.@Marilyn)
- Repeated characters with two characters (e.g. happyyyyy)

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 :

$$F = (f_1, \ldots, f_D) \in R^D$$

- Mark each message by a label
- Train learning algorithm on labeled messages

A. Neviarouskaya, H. Prendinger, and M. Ishizuka. "Affect analysis model: novel rule-based approach to affect sensing from text." *Natural Language Engineering* 17.1 (2011): 95-135.

Selecting Features

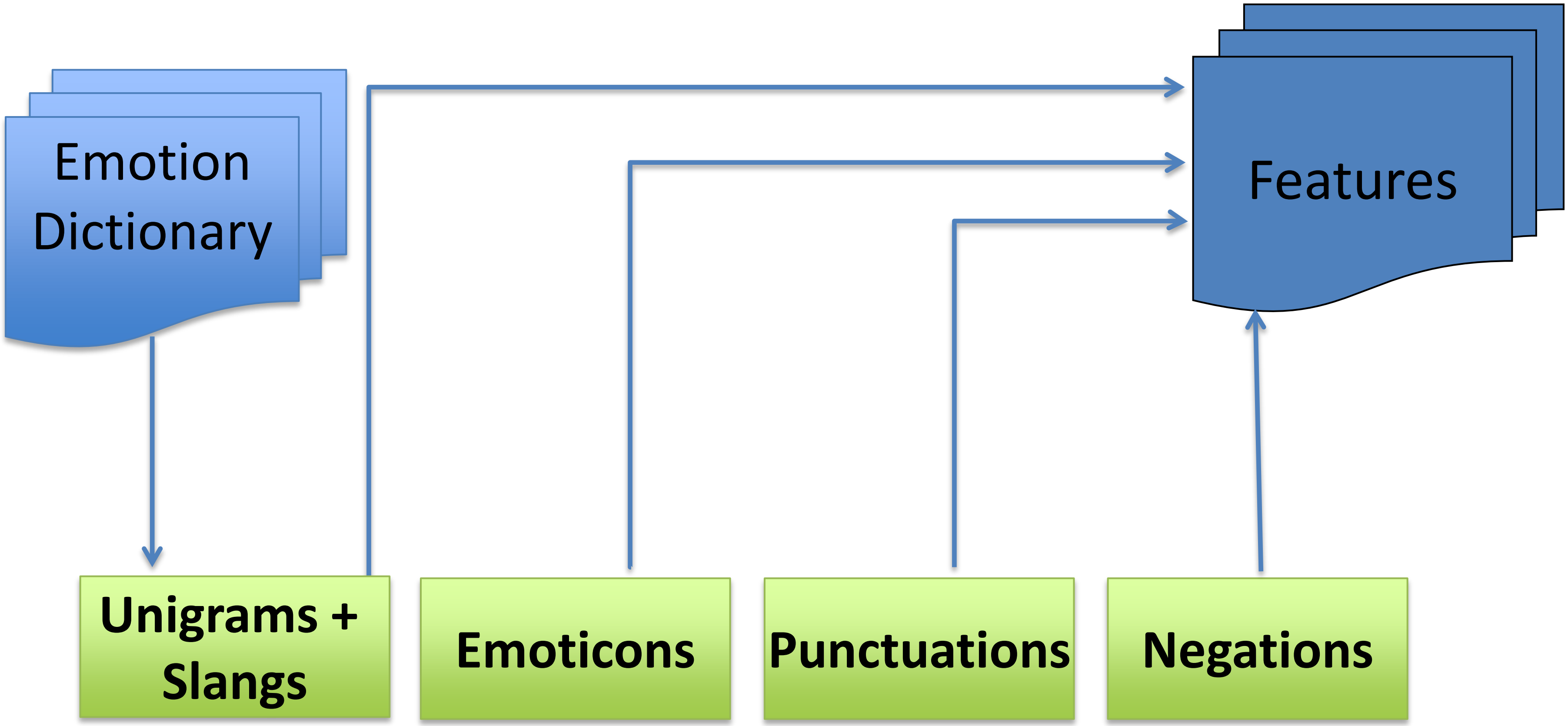


Figure 4- Feature Selection

Emoticon Features

Category	Emoticons
Happy Emoticons	:) ;) =) :] :p ;p :D ;D :> :3 :-) ;-) :^) :o) :~) ;^) ;o) :') :-D :->
Sad Emoticons	:(=(:-(:^(:o(:^(:'(:-<
Angry Emoticons	>:S >:{ >: x-@ :@ :-@ :-/ :-\
Afraid/Surprised Emoticons	:-o :-O o_O O_o :$
Sleepy Emoticons	-_- ~_~

Table2- 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.

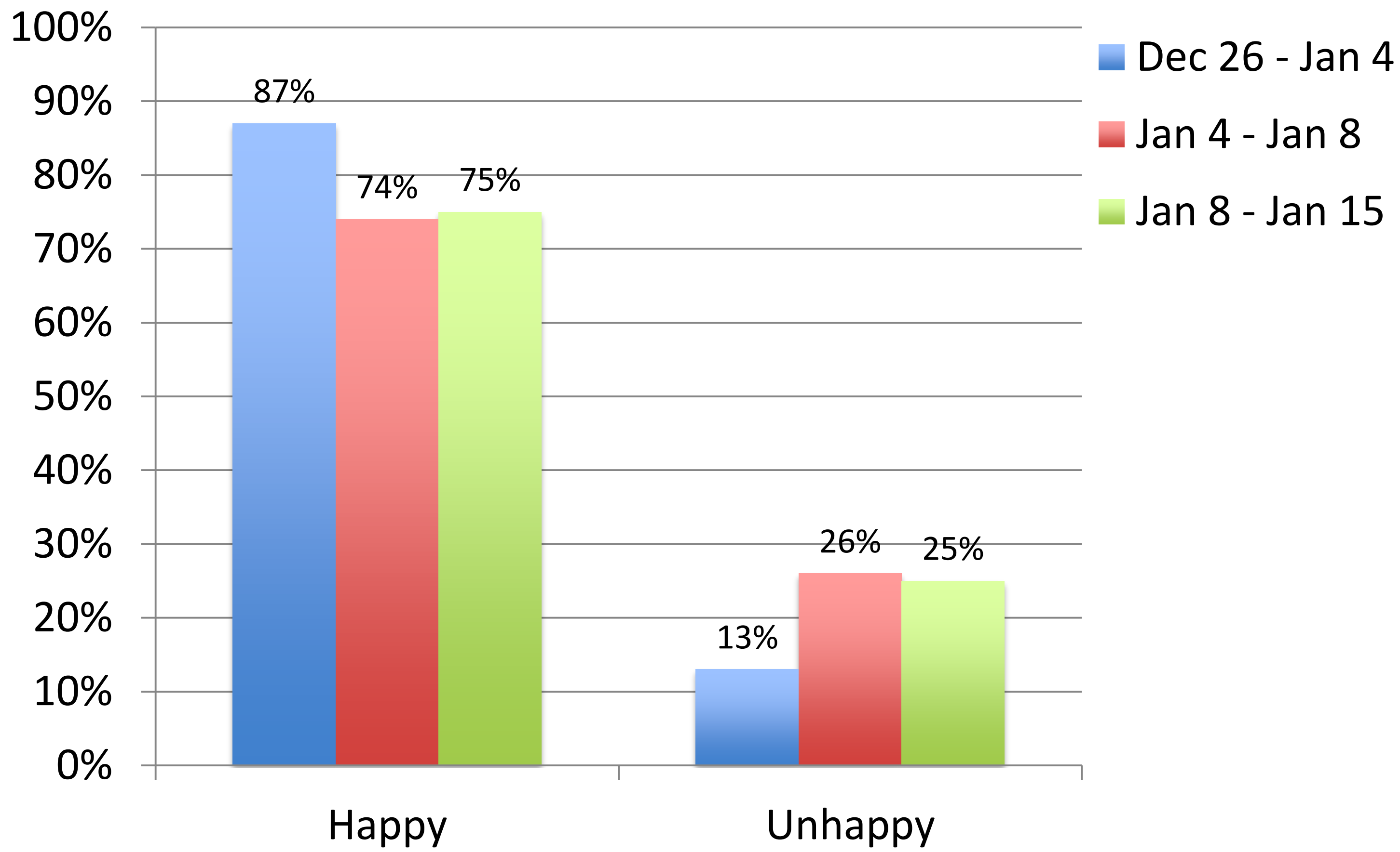


Figure 5 - Distribution of the emotions during new year vacation and after it

Results: Classification accuracy of SVM, KNN, Naïve Bayes, Decision Tree

	Unigram	Unigram, Emoticon	Unigram, Punctuation	Unigram, Negation	All Features
SVM	**89.86**	88.92	89.59	88.97	89.36
Naïve Bayes	86.27	86.40	86.61	86.91	**86.95**
Decision Tree	89.48	89.59	89.72	89.62	**89.93**
KNN	90.10	90.07	90.10	**90.14**	90.13

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
THANK