Data Science for Development - the Labs

INAFU6513, Spring 2016

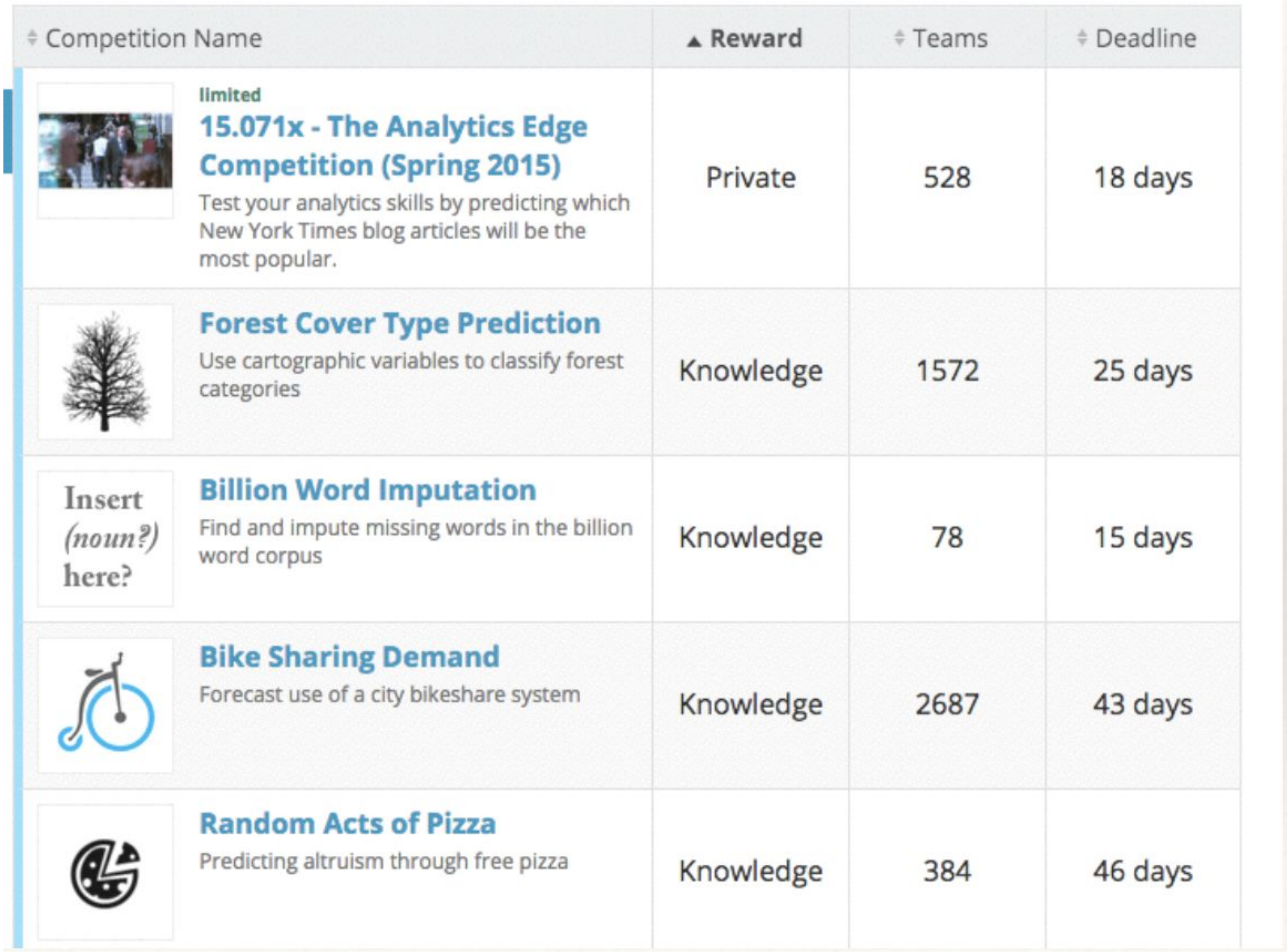
What is Data Science?

- “A data scientist… excels at analyzing data, particularly large amounts of data, to help a business gain a competitive edge.”

- “The analysis of data using the scientific method”

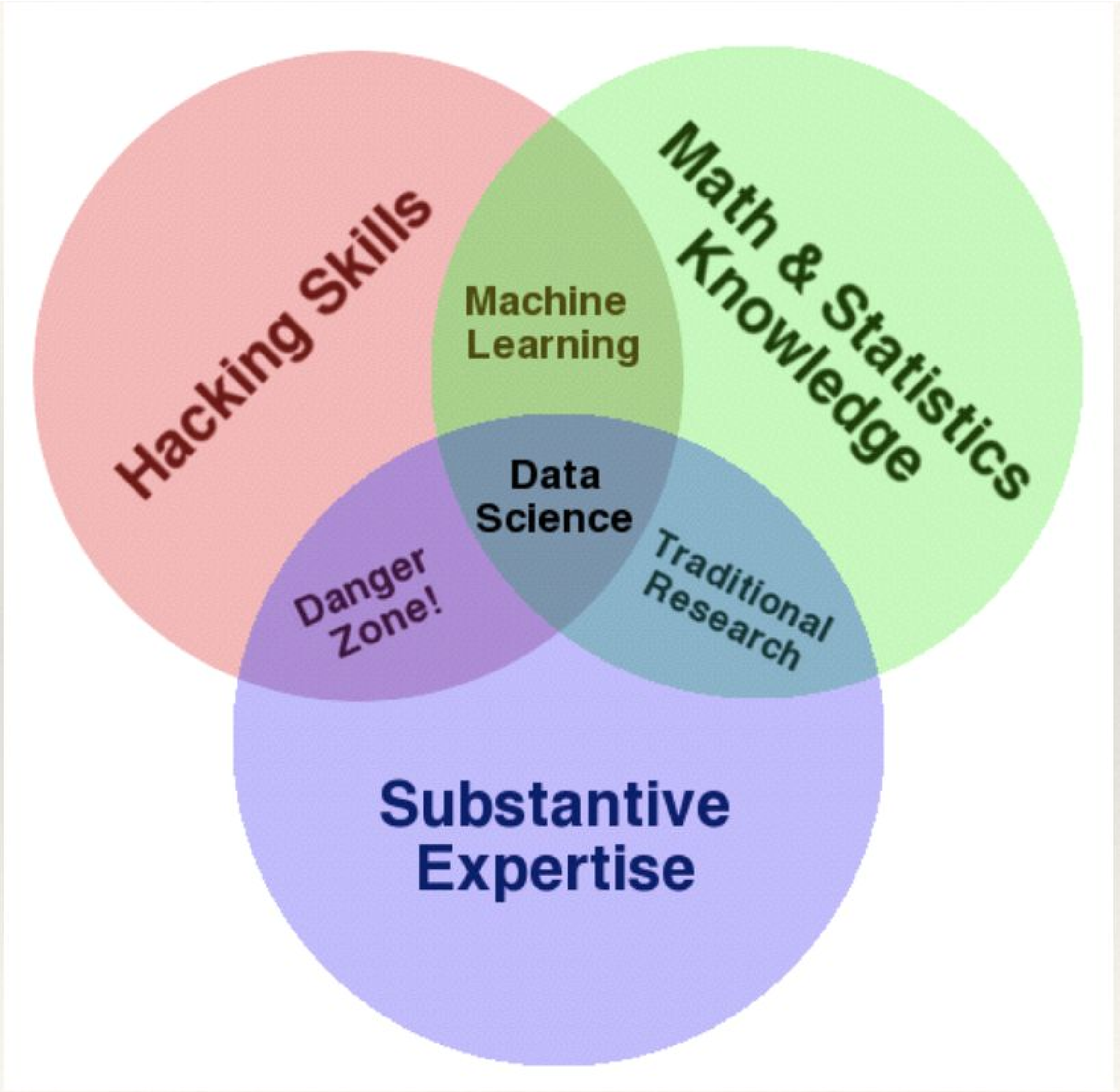
- “A data scientist is an individual, organization or application that performs statistical analysis, data mining and retrieval processes on a large amount of data to identify trends, figures and other relevant information.”

Understanding through Data



Competition Name	▲ Reward	Teams	Deadline
limited **15.071x - The Analytics Edge Competition (Spring 2015)** Test your analytics skills by predicting which New York Times blog articles will be the most popular.	Private	528	18 days
Forest Cover Type Prediction Use cartographic variables to classify forest categories	Knowledge	1572	25 days
Billion Word Imputation Find and impute missing words in the billion word corpus	Knowledge	78	15 days
Bike Sharing Demand Forecast use of a city bikeshare system	Knowledge	2687	43 days
Random Acts of Pizza Predicting altruism through free pizza	Knowledge	384	46 days

What’s a Data Scientist



How do you become a data scientist?

Practice

- Kaggle - online datascience competitions
- Driven Data - social good datascience competitions
- Innocentive - some datascience challenges
- CrowdAnalytix - business datascience competitions
- TunedIt - scientific/industrial datascience challenges
- Your individual and group exercises in this course...

Should you become a data scientist?

- Not necessarily. There are lots of data science students desperate for good problems to work on.
- You might want to become someone who can work **with** data scientists
- Which means learning how to specify data problems well

Data Science Lab Format

- Single topic
- Learn 4-6 concepts
- Try apps/ code related to that topic

- No prerequisites
- Will learn basics of Python, R and data tools

You need to:

- Download and install required tools
  - We’ll give you instructions
  - We can help!
- Do some (light) background reading
- Be playful with data, and have fun!

The Labs

1. Python basics
2. Acquiring data
3. Communicating results
4. Cleaning and exploring data
5. Predicting values from data
6. Handling text data
7. Handling geospatial data
8. Learning relationships from data
9. Working with data science teams
10. Enterprise data tools
11. Learning classes from data
12. Handling big data

The Lab Themes

People

- Working with data science teams
- Communicating results

Tools

- Python basics
- Enterprise data tools

Getting Data

- Acquiring data
- Cleaning and exploring data

Special data types

- Handling text data
- Handling geospatial data
- Handling big data

Learning from data

- Predicting values from data
- Learning relationships from data
- Learning classes from data

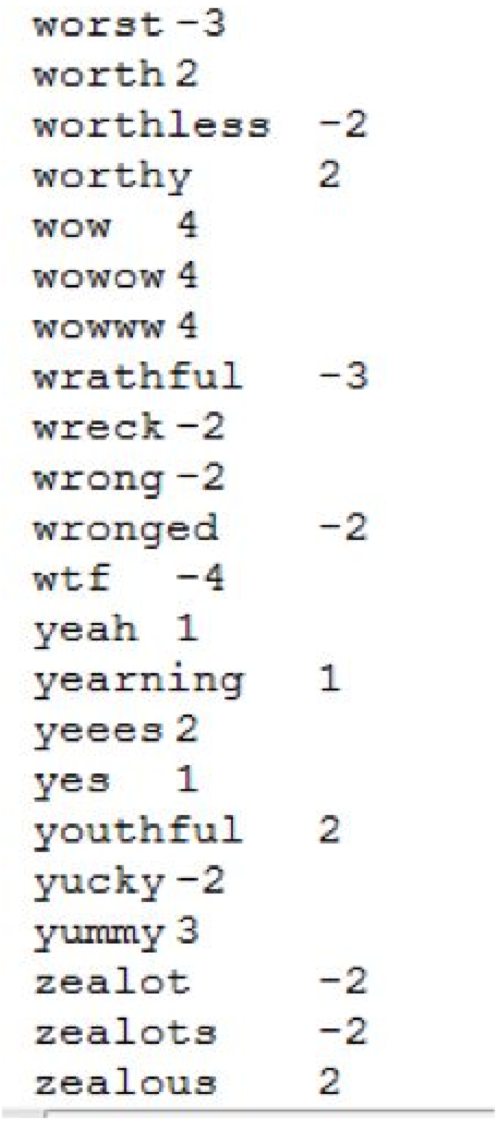
The tools

- Coding (Python, R)
- Scrapers and cleaners (Tabula, OpenRefine)
- Visualisation (Tableau, D3)
- GIS (QGIS, CartoDb)
- Big data (Hadoop)

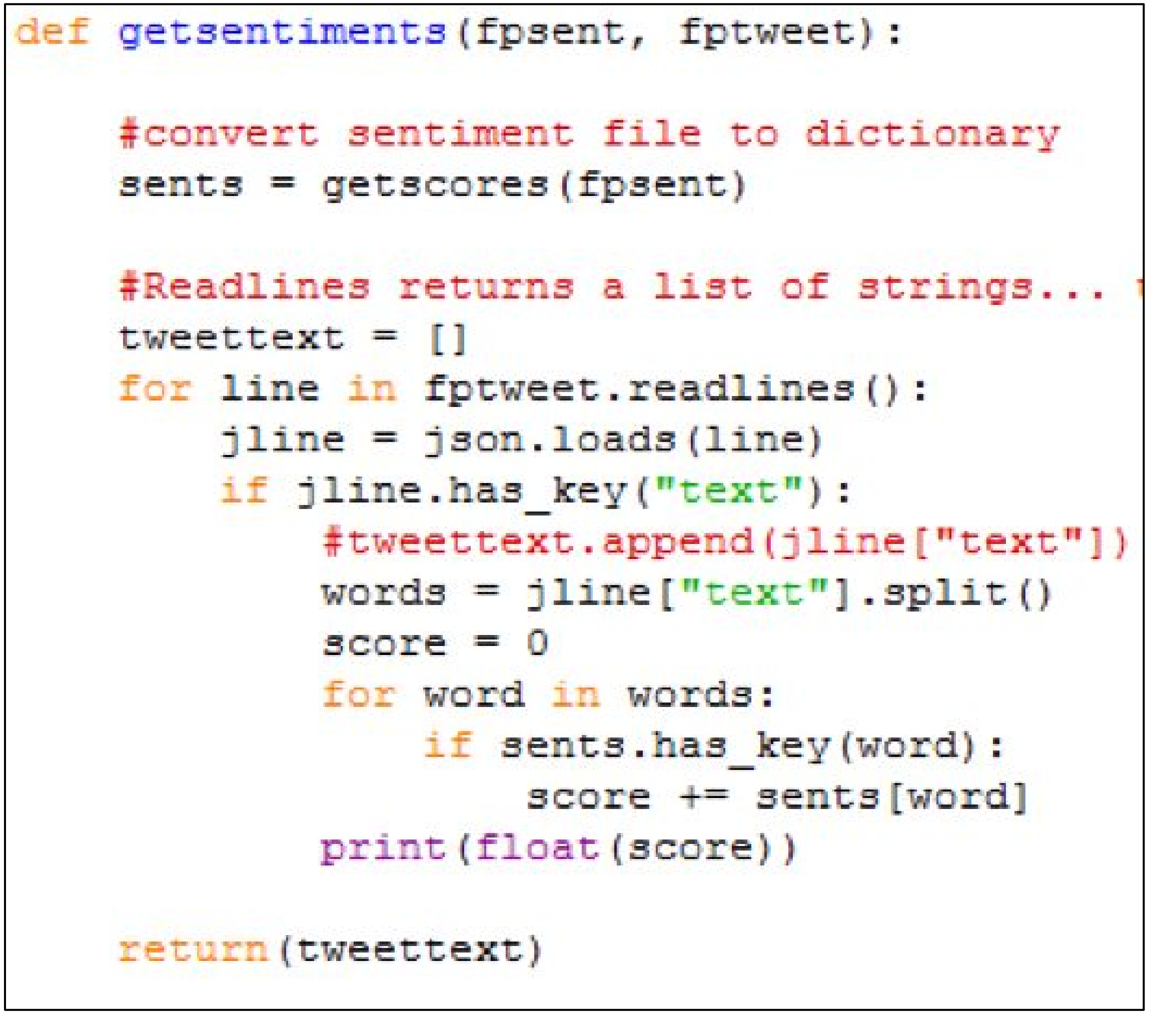
The gory details: Process

- Ask an interesting question
- Get the data
- Explore the data
- Model the data
- <sanity-check the data, in context>
- Communicate and visualize the data

Coding



```
worst	-3
worth	2
worthless	-2
worthy	2
wow	4
wowow	4
wowww	4
wrathful	-3
wreck	-2
wrong	-2
wronged	-2
wtf	-4
yeah	1
yearning	1
yeees	2
yes	1
youthful	2
yucky	-2
yummy	3
zealot	-2
zealots	-2
zealous	2
```



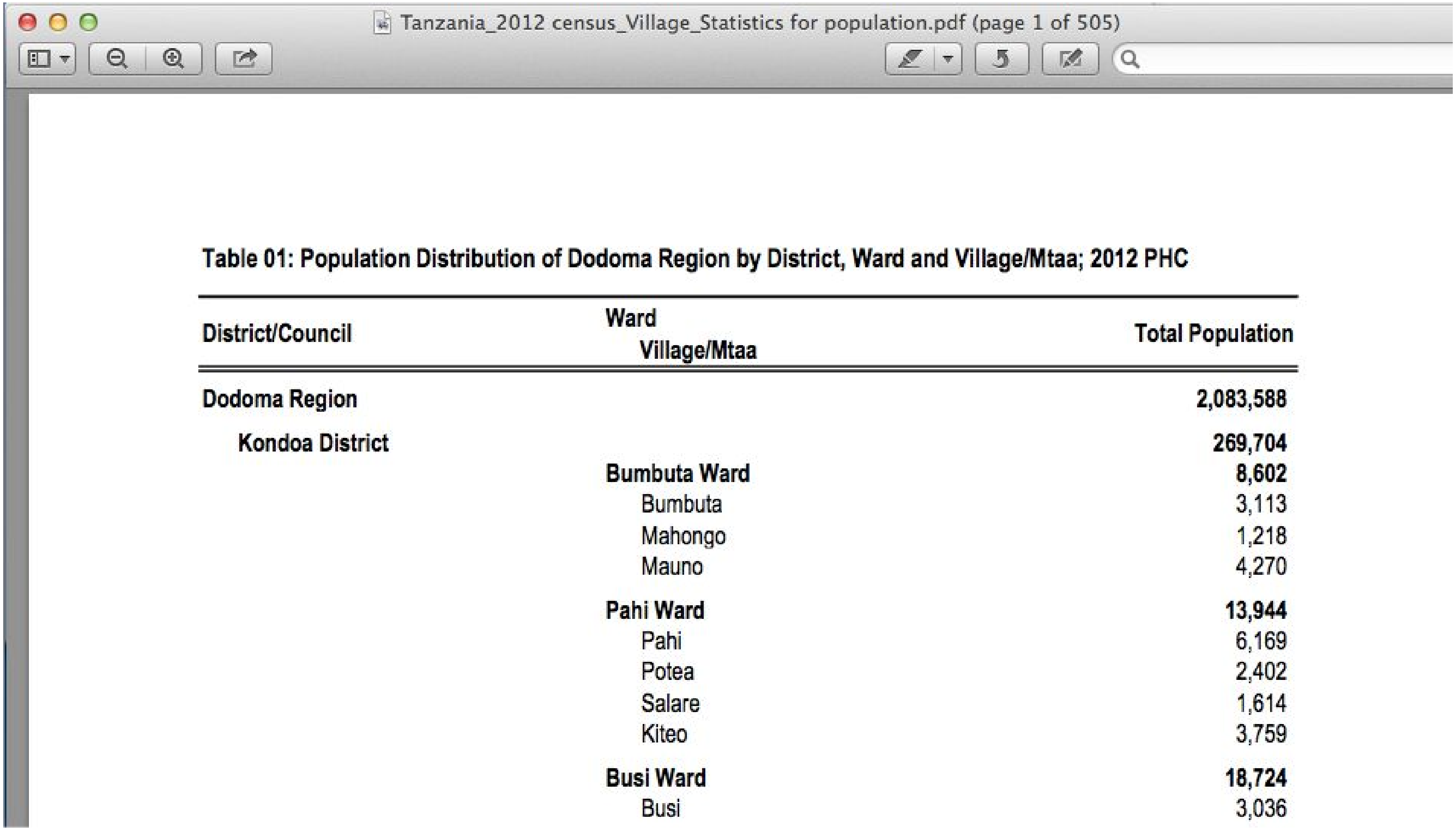
```
def getsentiments(fpsent, fptweet):

    #convert sentiment file to dictionary
    sents = getscores(fpsent)

    #Readlines returns a list of strings...
    tweettext = []
    for line in fptweet.readlines():
        jline = json.loads(line)
        if jline.has_key("text"):
            #tweettext.append(jline["text"])
            words = jline["text"].split()
            score = 0
            for word in words:
                if sents.has_key(word):
                    score += sents[word]
            print(float(score))

    return(tweettext)
```

Acquiring data

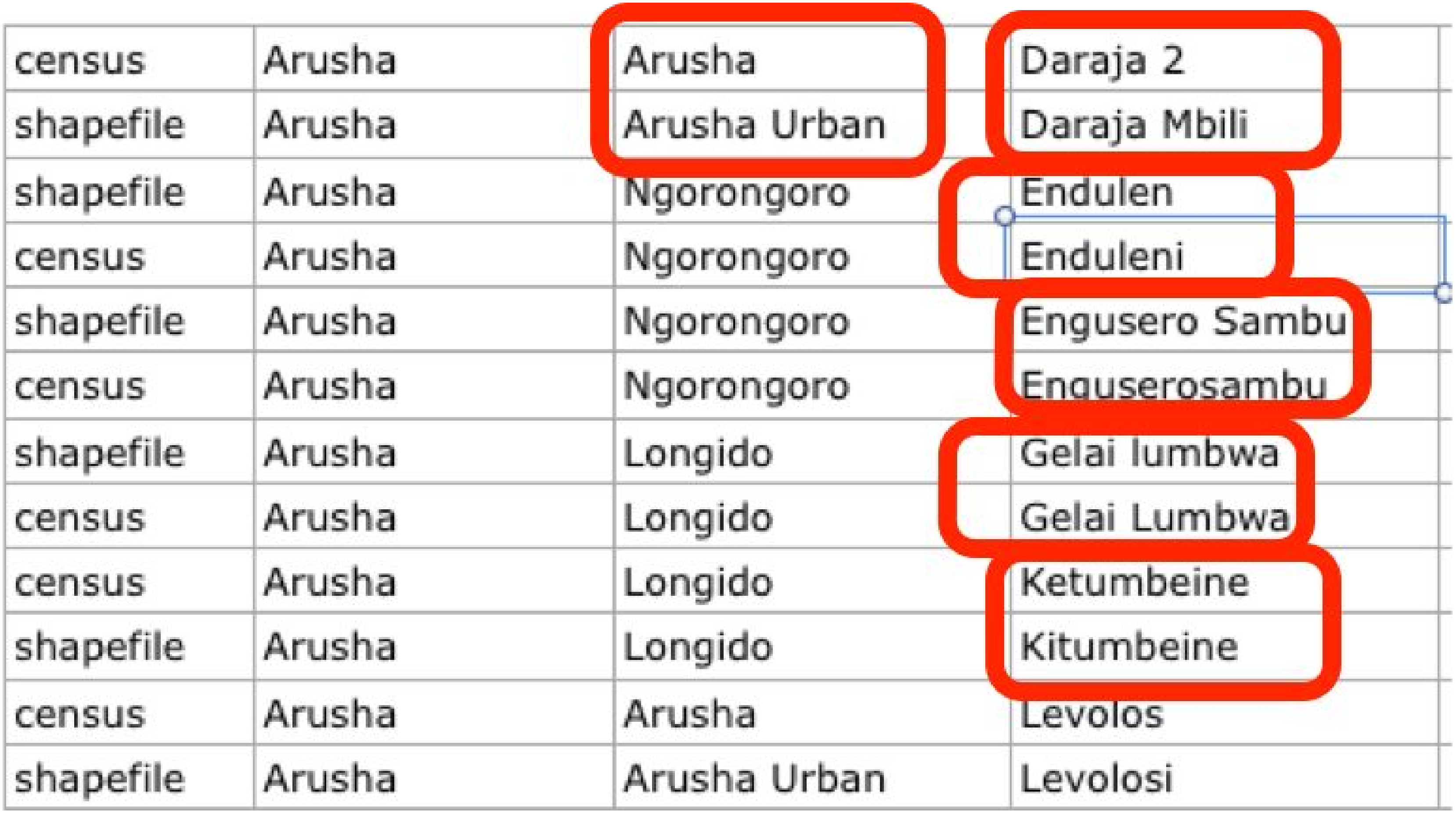


Tanzania_2012 census_Village_Statistics for population.pdf (page 1 of 505)

Table 01: Population Distribution of Dodoma Region by District, Ward and Village/Mtaa; 2012 PHC

District/Council	Ward Village/Mtaa	Total Population
Dodoma Region		2,083,588
Kondoa District		269,704
	Bumbuta Ward	8,602
	Bumbuta	3,113
	Mahongo	1,218
	Mauno	4,270
	Pahi Ward	13,944
	Pahi	6,169
	Potea	2,402
	Salare	1,614
	Kiteo	3,759
	Busi Ward	18,724
	Busi	3,036

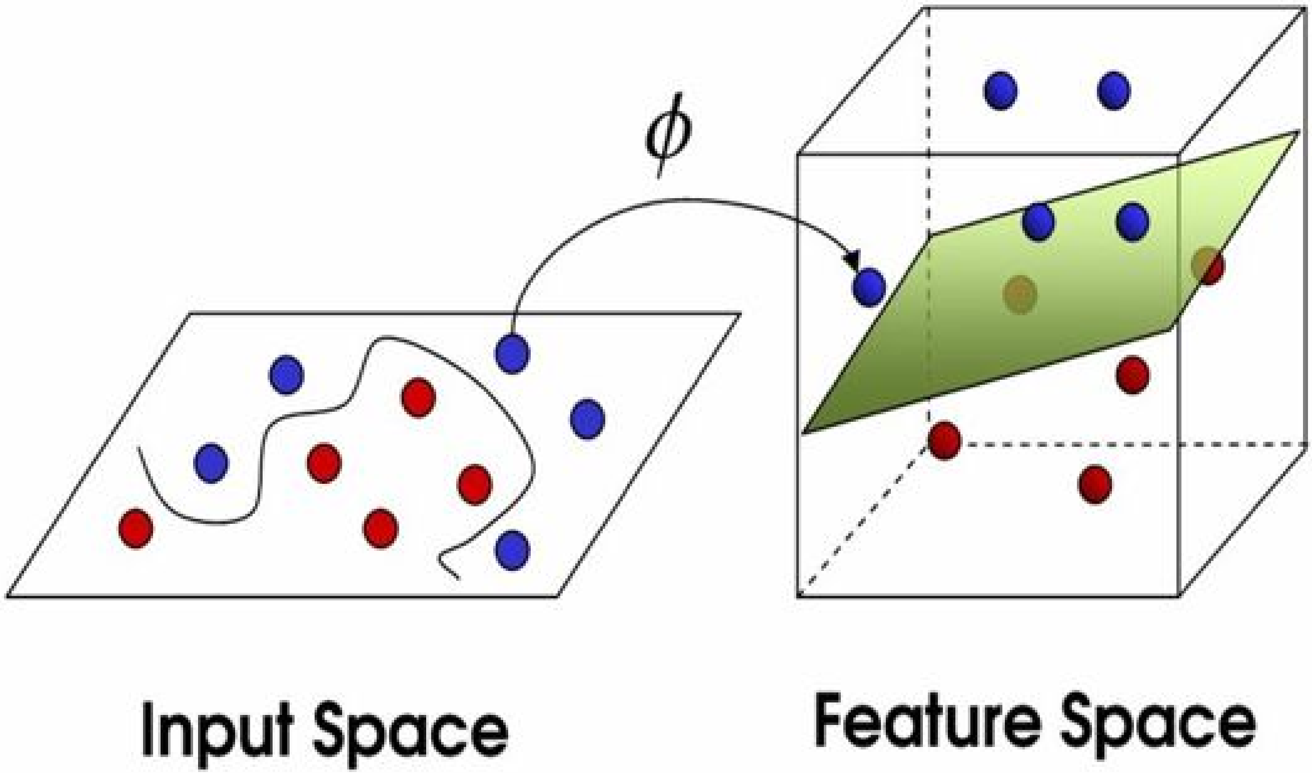
Cleaning and Exploring



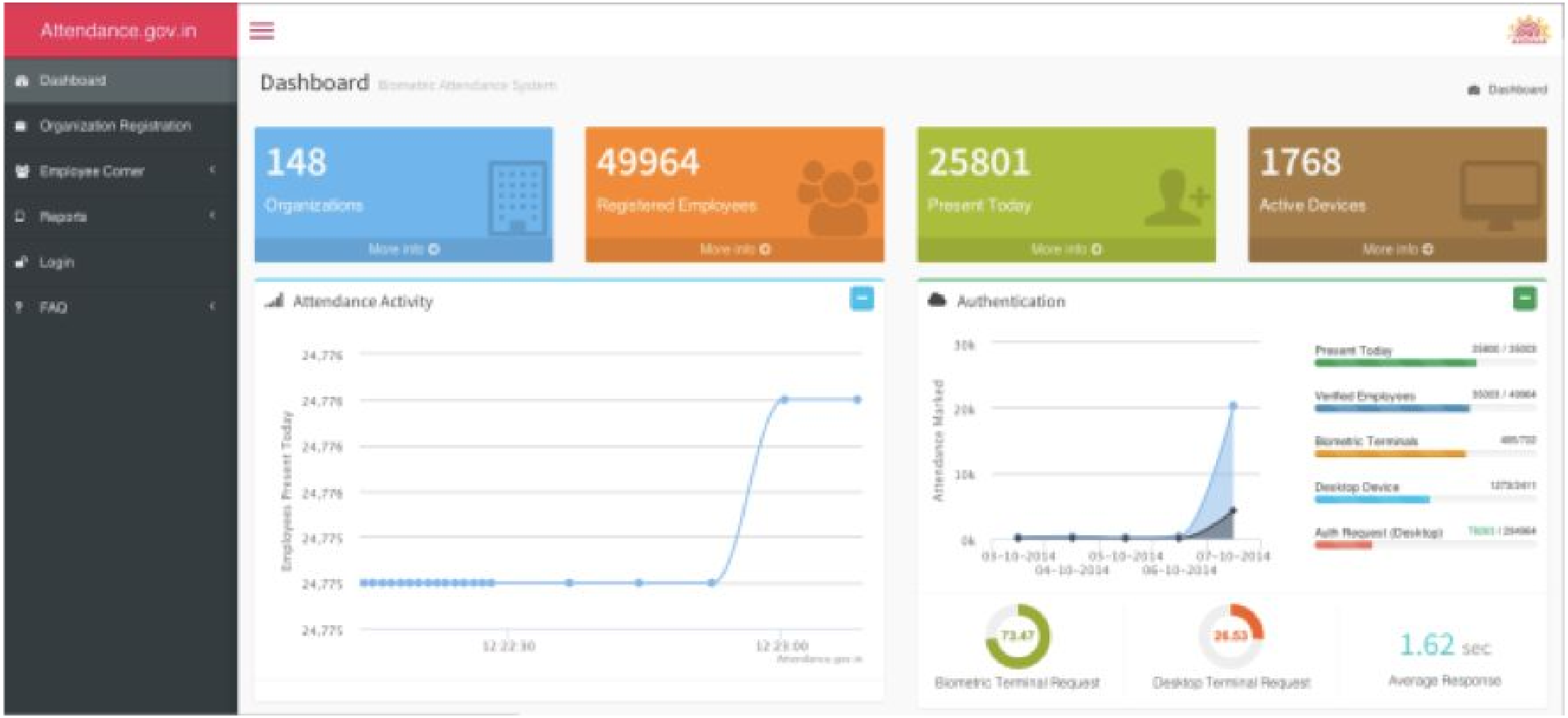
census	Arusha	Arusha	Daraja 2
shapefile	Arusha	Arusha Urban	Daraja Mbili
shapefile	Arusha	Ngorongoro	Endulen
census	Arusha	Ngorongoro	Enduleni
shapefile	Arusha	Ngorongoro	Engusero Sambu
census	Arusha	Ngorongoro	Enguserosambu
shapefile	Arusha	Longido	Gelai lumbwa
census	Arusha	Longido	Gelai Lumbwa
census	Arusha	Longido	Ketumbeine
shapefile	Arusha	Longido	Kitumbeine
census	Arusha	Arusha	Levolos
shapefile	Arusha	Arusha Urban	Levolosi

Modelling and Learning

Principle of Support Vector Machines (SVM)



Communicating results



Attendance.gov.in

Dashboard Biometric Attendance System

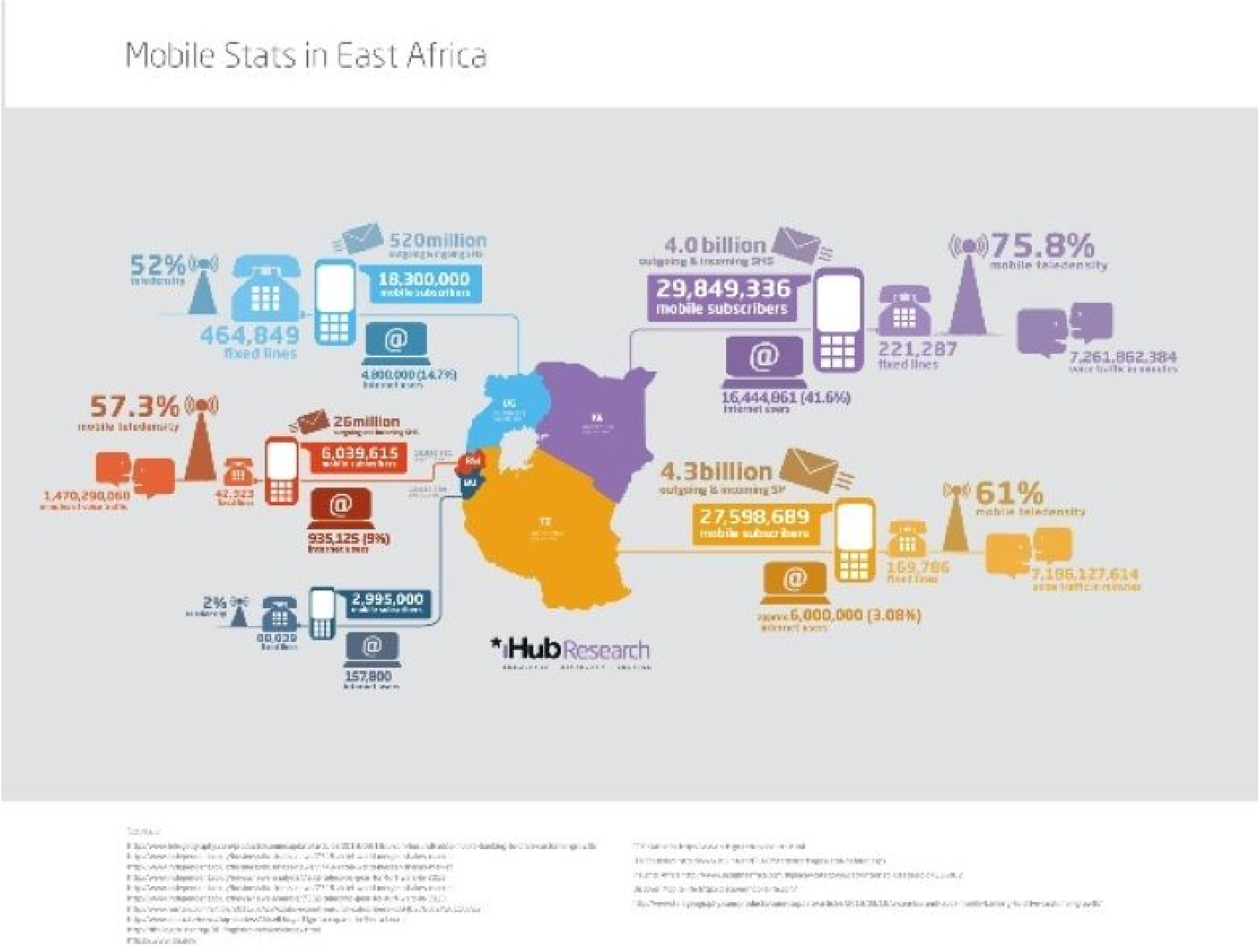
148 Organizations

49964 Registered Employees

25801 Present Today

1768 Active Devices

Big Data


Mobile Stats in East Africa
520million
18,300,000
52%
464,849
fixed lines
4,800,000 (14.7%)
57.3%
mobile teledensity
26million
6,039,615
1,470,290,068
42,523
935,125 (9%)
2%
2,995,000
157,800
UG
KA
RW
BU
TZ
4.0 billion
outgoing & incoming SMS
29,849,336
mobile subscribers
75.8%
mobile teledensity
221,287
fixed lines
7,261,862,384
16,444,861 (41.6%)
internet users
4.3billion
27,598,689
mobile subscribers
61%
mobile teledensity
169,786
fixed lines
7,186,127,614
approx 6,000,000 (3.08%)
*iHub Research

What We’re Aiming For

Ask good questions;
Tell good stories

What you need to do by next week:

Install tools:

- Courseworks folder: Pre-lab instructions
- File: InstallingEverything-READMEFIRST.pdf