# Statistical Society of Canada

STAT 400/ DATA 500 Final Consulting Presentation
December 7th, 2017

Norbert Eke, Mahdi Aziz, Ryan McQueen

# Background

- Every year there is a conference held by the Statistical Society of Canada (SSC) which has speakers within the field of statistics present their work.
    - Student Speakers
    - Invited Speakers
    - Contributed Speakers

- Pre-allocated timeframes are given to the invited speakers, and student speakers must present first

- The details of previous year SSC conferences are provided in LaTeX files

# Problems

- At each table of speakers, there are speakers of unrelated areas present
    - Physics focused statisticians at a table composed primarily of biology statisticians

- Lots of manual work needs to be done to allocate presentation times and groups

- Parallel sessions topics are not as maximized for dissimilarity

# Solution

- Extract author names, abstract titles, abstracts descriptions, and presentation times from the provided LaTeX files

- Process the abstract descriptions, feeding them into deep learning language models to determine similar abstracts, and perform cluster analysis to identify potential groups

- Provide a schedule for the abstracts based on the clusters given such that no talk with the same cluster would happen at the same time and the talks would be presented based on the priorities of speakers.

# External Limitations

- Multi Platform
  - Work on all major operating systems (Linux, MacOS, Windows)

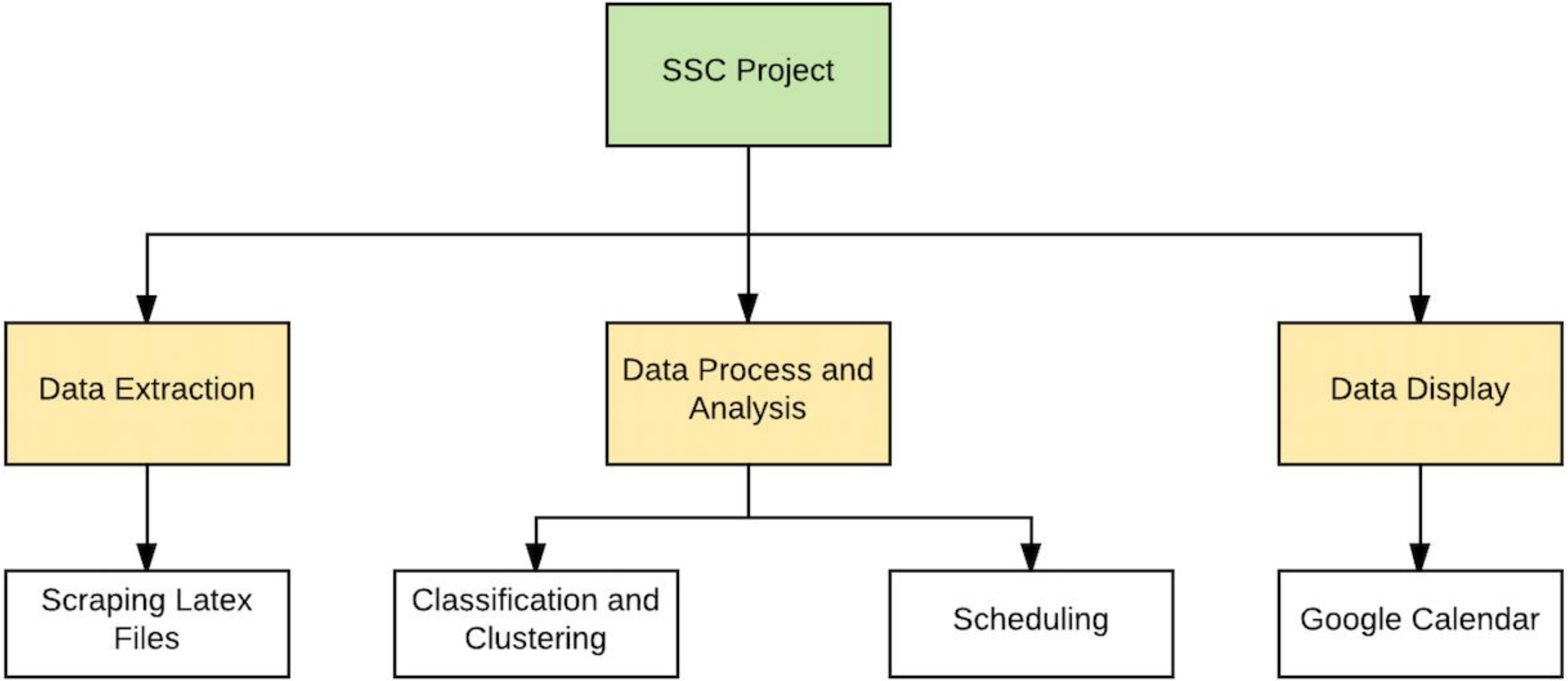
- Code must be written in the R or Python programming languages

# Internal Limitations

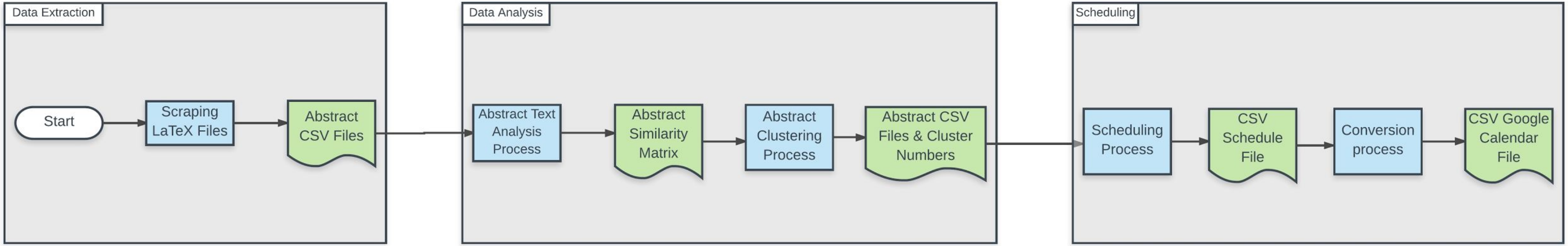
- Do not know who the keynote and student speakers are
    - Need to simulate these speakers to fulfill the project requirements

- Inconsistencies between the abstracts and program outline LaTeX files
    - Some entries are commented in one file and not the other

# High-Level Architecture



# System Box Diagram



# Data Extraction

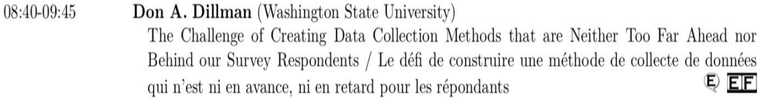
- Data written into two LaTeX files:
  - abstracts.tex: outlines an abstract title, abstract description, and who is presenting it
  - prog.tex: outlines who is an invited, contributed, or poster session speaker

- Data overlap between the abstracts and prog file
  - Which one is the correct file to use?

- Utilize Regular Expressions to handle a majority of the work
  - Allows for patterns to be extracted from text: “\absTime{08:30:00}” => “08:30:00”

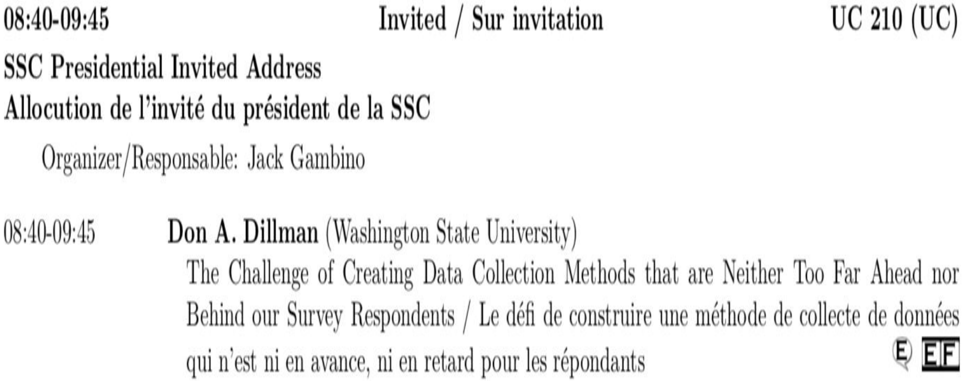
# Data Extraction: File Examples

## abstracts



08:40-09:45 **Don A. Dillman** (Washington State University)
The Challenge of Creating Data Collection Methods that are Neither Too Far Ahead nor Behind our Survey Respondents / Le défi de construire une méthode de collecte de données qui n'est ni en avance, ni en retard pour les répondants E EF

## prog



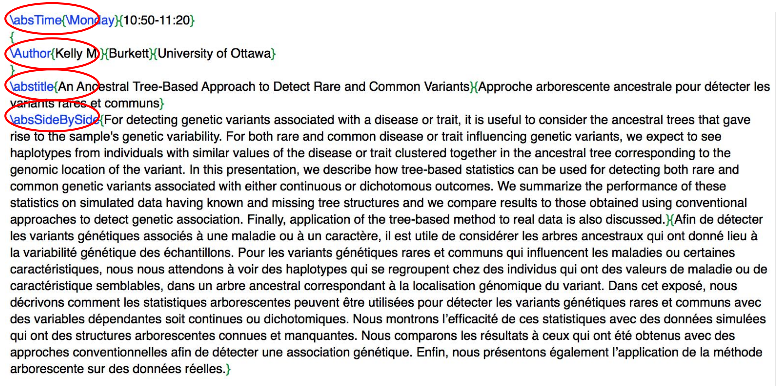
08:40-09:45 Invited / Sur invitation UC 210 (UC)

**SSC Presidential Invited Address**
**Allocution de l'invité du président de la SSC**

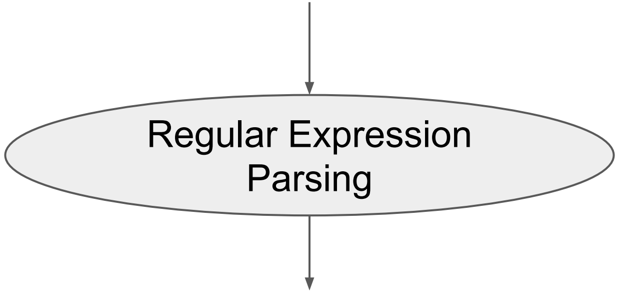
Organizer/Responsable: Jack Gambino

08:40-09:45 **Don A. Dillman** (Washington State University)
The Challenge of Creating Data Collection Methods that are Neither Too Far Ahead nor Behind our Survey Respondents / Le défi de construire une méthode de collecte de données qui n'est ni en avance, ni en retard pour les répondants E EF

# Data Extraction: Abstracts



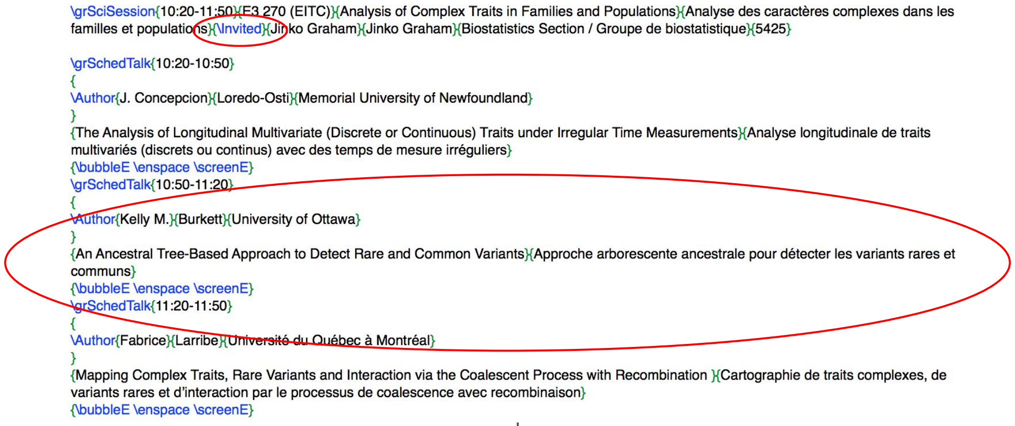
```
\absTime{\Monday}{10:50-11:20}
{
\Author{Kelly M}{Burkett}{University of Ottawa}
}
\abstitle{An Ancestral Tree-Based Approach to Detect Rare and Common Variants}{Approche arborescente ancestrale pour détecter les
variants rares et communs}
\absSideBySide{For detecting genetic variants associated with a disease or trait, it is useful to consider the ancestral trees that gave
rise to the sample's genetic variability. For both rare and common disease or trait influencing genetic variants, we expect to see
haplotypes from individuals with similar values of the disease or trait clustered together in the ancestral tree corresponding to the
genomic location of the variant. In this presentation, we describe how tree-based statistics can be used for detecting both rare and
common genetic variants associated with either continuous or dichotomous outcomes. We summarize the performance of these
statistics on simulated data having known and missing tree structures and we compare results to those obtained using conventional
approaches to detect genetic association. Finally, application of the tree-based method to real data is also discussed.}{Afin de détecter
les variants génétiques associés à une maladie ou à un caractère, il est utile de considérer les arbres ancestraux qui ont donné lieu à
la variabilité génétique des échantillons. Pour les variants génétiques rares et communs qui influencent les maladies ou certaines
caractéristiques, nous nous attendons à voir des haplotypes qui se regroupent chez des individus qui ont des valeurs de maladie ou de
caractéristique semblables, dans un arbre ancestral correspondant à la localisation génomique du variant. Dans cet exposé, nous
décrivons comment les statistiques arborescentes peuvent être utilisées pour détecter les variants génétiques rares et communs avec
des variables dépendantes soit continues ou dichotomiques. Nous montrons l'efficacité de ces statistiques avec des données simulées
qui ont des structures arborescentes connues et manquantes. Nous comparons les résultats à ceux qui ont été obtenus avec des
approches conventionnelles afin de détecter une association génétique. Enfin, nous présentons également l'application de la méthode
arborescente sur des données réelles.}
```



Regular Expression Parsing

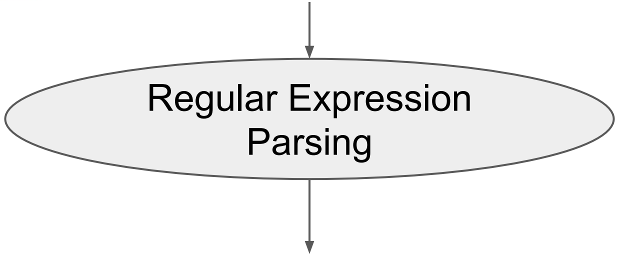
**Time:** “Monday 10:50-11:20”
**Author:** “Kelly M. Burkett, University of Ottawa”
**Abstract Title:** “An Ancestral Tree-Based Approach …”
**Abstract Description:** “For detecting genetic variants …”

# Data Extraction: Program



```
\grSciSession{10:20-11:50}{E3 270 (EITC)}{Analysis of Complex Traits in Families and Populations}{Analyse des caractères complexes dans les
familles et populations}{\Invited}{Jinko Graham}{Jinko Graham}{Biostatistics Section / Groupe de biostatistique}{5425}

\grSchedTalk{10:20-10:50}
{
\Author{J. Concepcion}{Loredo-Osti}{Memorial University of Newfoundland}
}
{The Analysis of Longitudinal Multivariate (Discrete or Continuous) Traits under Irregular Time Measurements}{Analyse longitudinale de traits
multivariés (discrets ou continus) avec des temps de mesure irréguliers}
{\bubbleE \enspace \screenE}
\grSchedTalk{10:50-11:20}
{
\Author{Kelly M.}{Burkett}{University of Ottawa}
}
{An Ancestral Tree-Based Approach to Detect Rare and Common Variants}{Approche arborescente ancestrale pour détecter les variants rares et
communs}
{\bubbleE \enspace \screenE}
\grSchedTalk{11:20-11:50}
{
\Author{Fabrice}{Larribe}{Université du Québec à Montréal}
}
{Mapping Complex Traits, Rare Variants and Interaction via the Coalescent Process with Recombination }{Cartographie de traits complexes, de
variants rares et d'interaction par le processus de coalescence avec recombinaison}
{\bubbleE \enspace \screenE}
```



Regular Expression Parsing

**Abstract Title:** “An Ancestral Tree-Based Approach …”
**Flag:** “Invited”

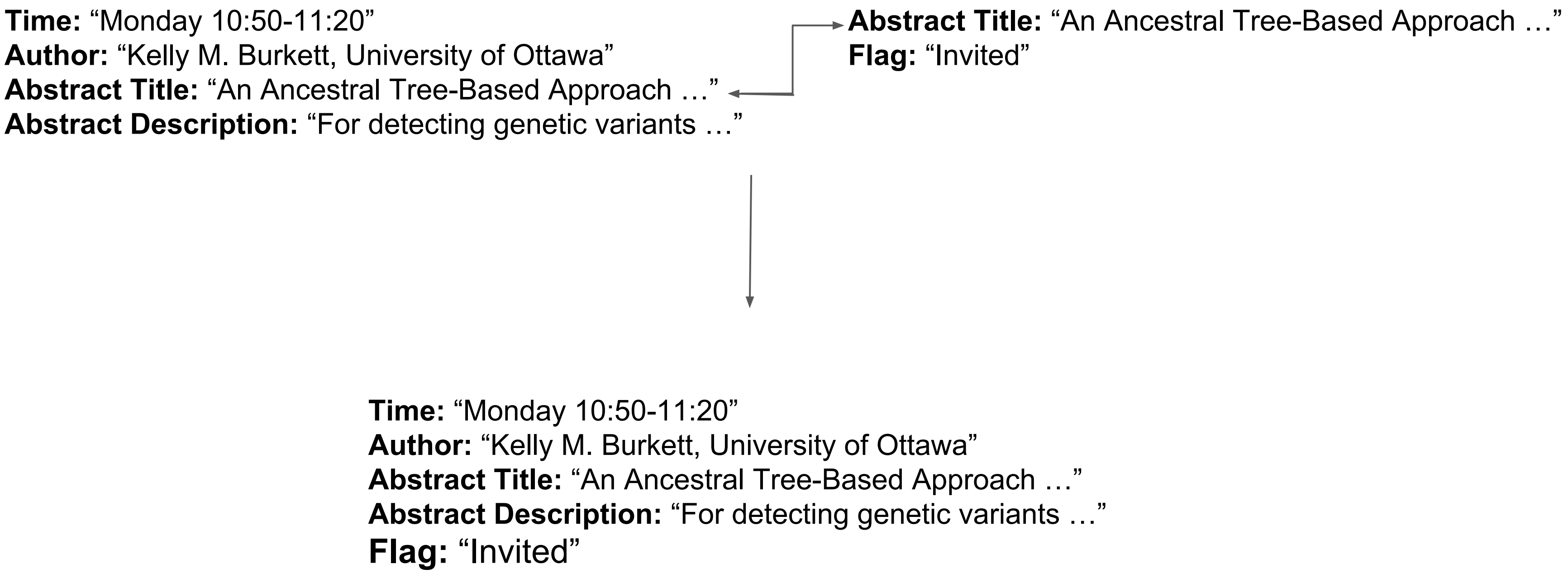
# Data Extraction: Mapping

- Identified which abstract is invited, contributed, or a poster session
    - How can it be related back to the original abstracts information?

- Requires a mapping from the data parsed from abstracts.tex and prog.tex

- Problem: there are some entries present within abstracts.tex, but no prog.tex
    - The curse of LaTeX commented text

# Data Extraction: Mapping



**Time:** “Monday 10:50-11:20”
**Author:** “Kelly M. Burkett, University of Ottawa”
**Abstract Title:** “An Ancestral Tree-Based Approach …”
**Abstract Description:** “For detecting genetic variants …”

**Abstract Title:** “An Ancestral Tree-Based Approach …”
**Flag:** “Invited”

**Time:** “Monday 10:50-11:20”
**Author:** “Kelly M. Burkett, University of Ottawa”
**Abstract Title:** “An Ancestral Tree-Based Approach …”
**Abstract Description:** “For detecting genetic variants …”
**Flag:** “Invited”

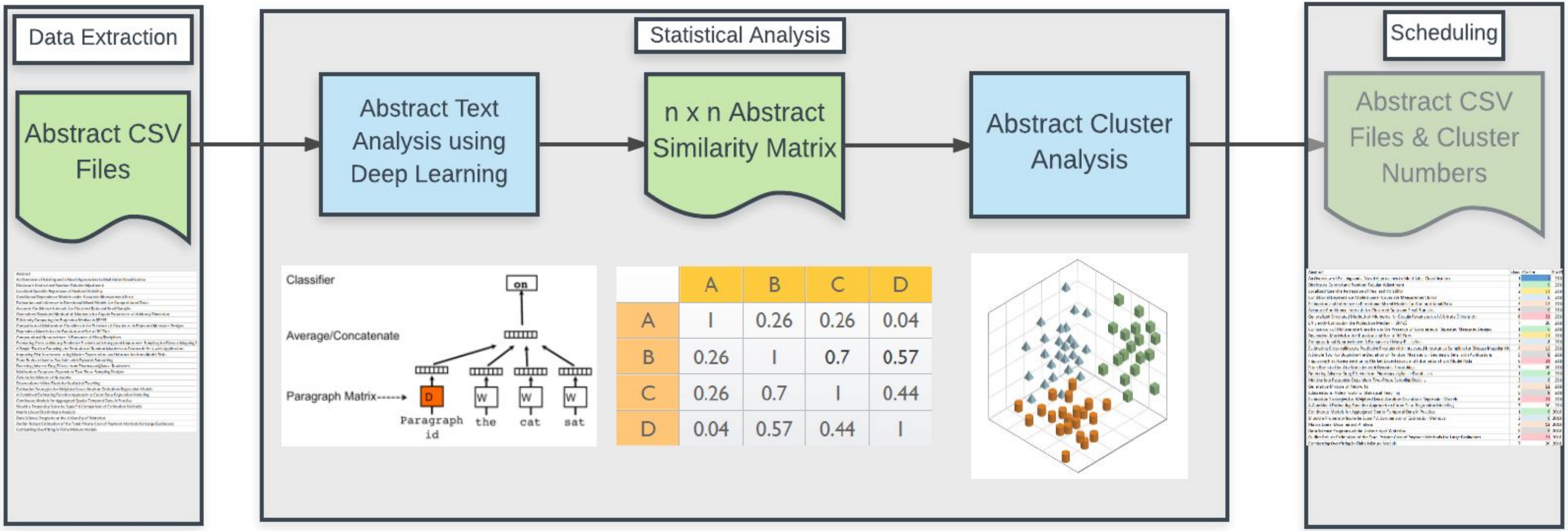
# Data Extraction: Random Labeling

- Due to the lack of student and keynote speaker labels, they needed to be randomly assigned

- Upon discussion with our client, the number of both categories were outlined
  - 3 Keynote speakers, 1 per day. Keynote speakers are invited talks
  - 30 student speakers. Student speakers are contributed talks

- Perform random sampling on the invited and contributed talks
  - 30 random entries with the flag of “Contributed” were assigned the flag of “Student”
  - 3 random entries with the flag of “Invited” were assigned the flag of “Keynote”

# Statistical Analysis



Data Extraction

Abstract CSV Files

Statistical Analysis

Abstract Text Analysis using Deep Learning

n x n Abstract Similarity Matrix

Abstract Cluster Analysis

Scheduling

Abstract CSV Files & Cluster Numbers

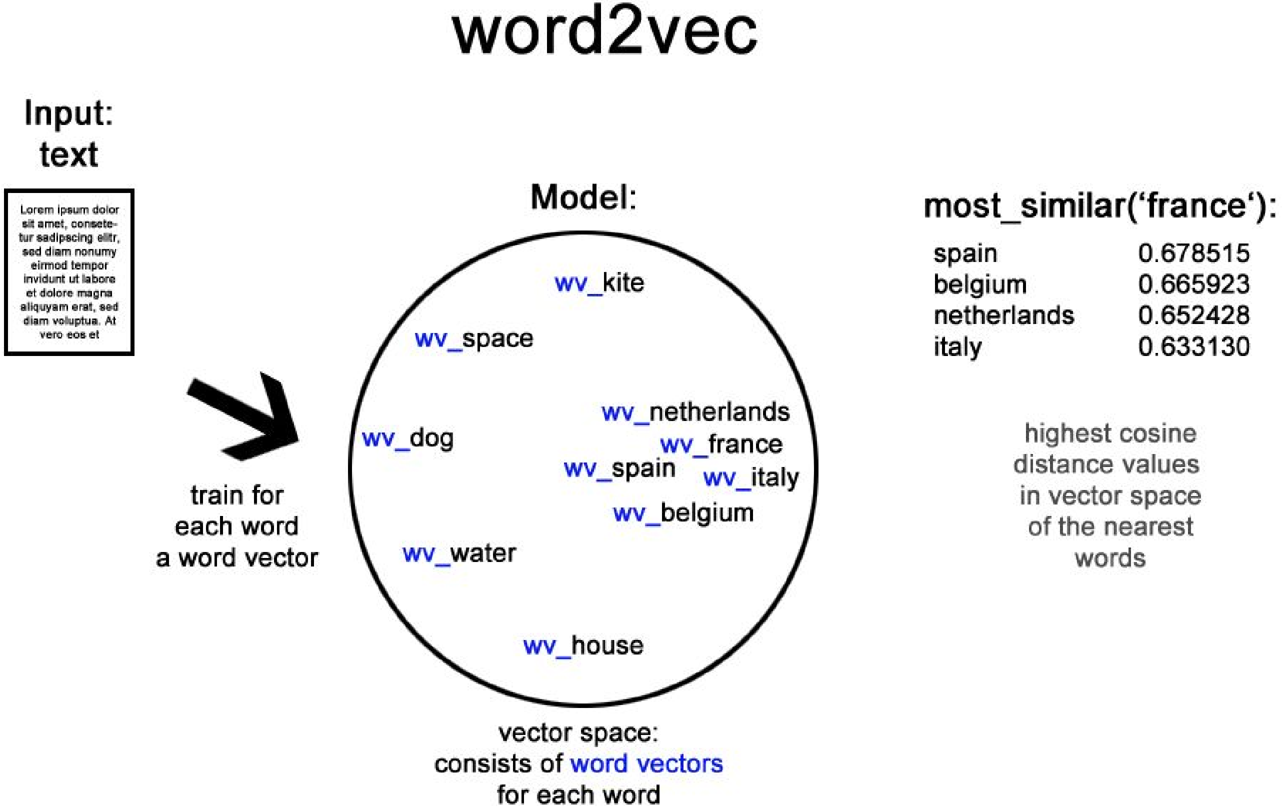
| | A | B | C | D |
|---|---|---|---|---|
| A | 1 | 0.26 | 0.26 | 0.04 |
| B | 0.26 | 1 | 0.7 | 0.57 |
| C | 0.26 | 0.7 | 1 | 0.44 |
| D | 0.04 | 0.57 | 0.44 | 1 |

# Language Models

- Input: words

- Word Vector space

- Output: word cosine similarity



word2vec
Input: text
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et
train for each word a word vector
Model:
wv_kite
wv_space
wv_netherlands
wv_dog
wv_france
wv_spain
wv_italy
wv_belgium
wv_water
wv_house
vector space: consists of word vectors for each word
highest cosine distance values in vector space of the nearest words

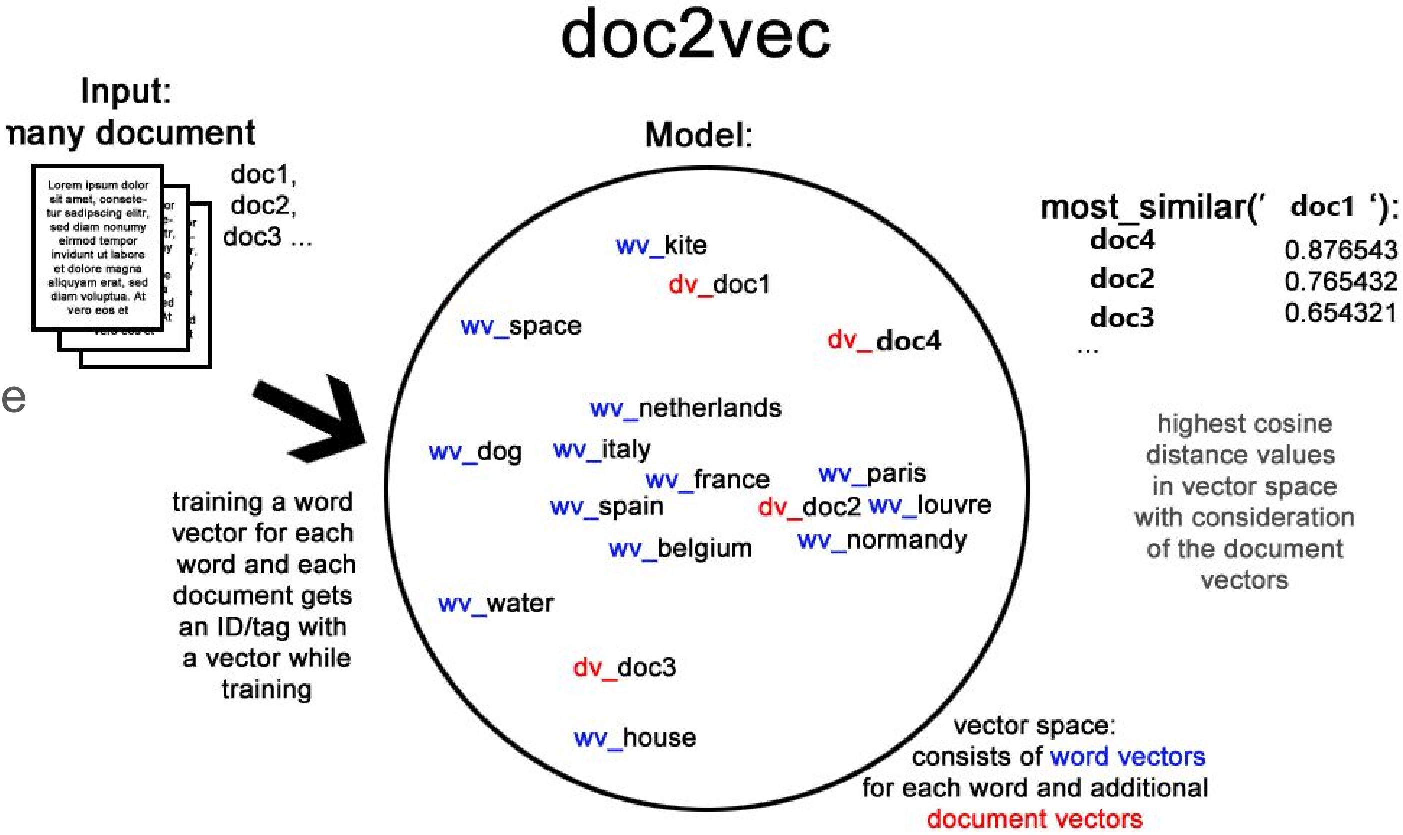
most_similar('france'):

| | |
|---|---|
| spain | 0.678515 |
| belgium | 0.665923 |
| netherlands | 0.652428 |
| italy | 0.633130 |

[2]

# Language Models

- Input: abstracts
- Document Vector space
- Output: Cosine similarity matrix



training a word vector for each word and each document gets an ID/tag with a vector while training

highest cosine distance values in vector space with consideration of the document vectors

vector space: consists of word vectors for each word and additional document vectors

[2]

# Abstract Similarity Cluster Analysis

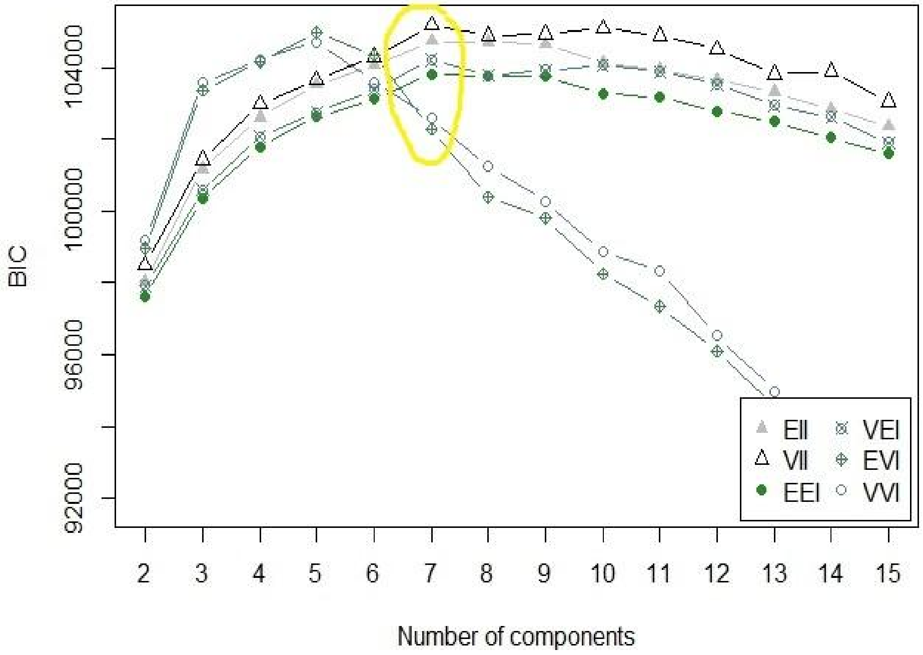
- Input: Abstract Similarity Matrix

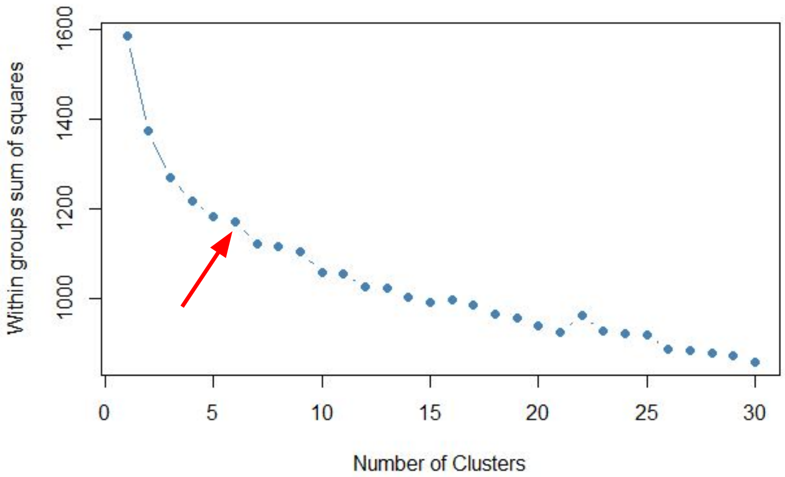
  Output: Abstract Cluster Memberships

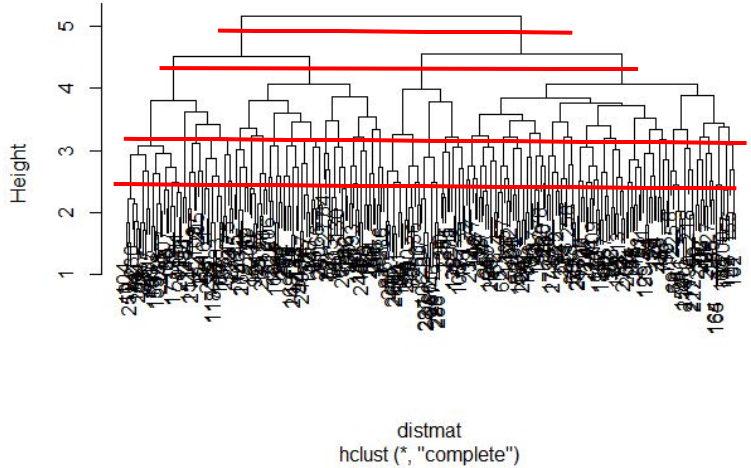
- First Task: Determine Number of Clusters (optimal cluster sizes ?)

- Second Task: Which Clustering algorithm(s) to use ?

# Deciding on Number of Clusters

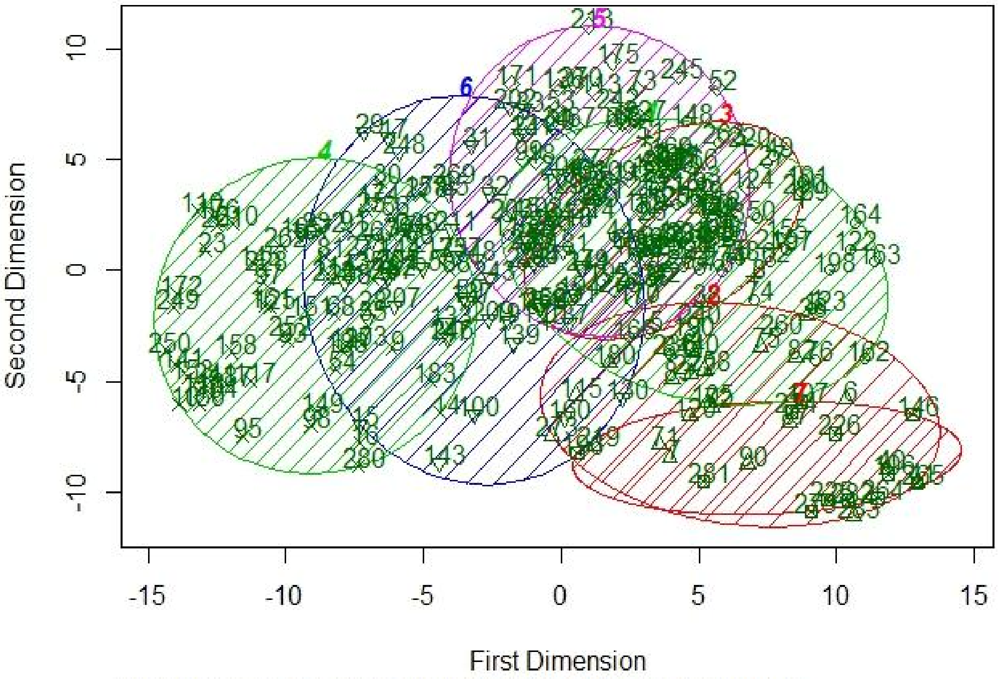

Mixture Models BIC Plot


K-Means Elbow Plot


Hierarchical Clustering Dendrogram

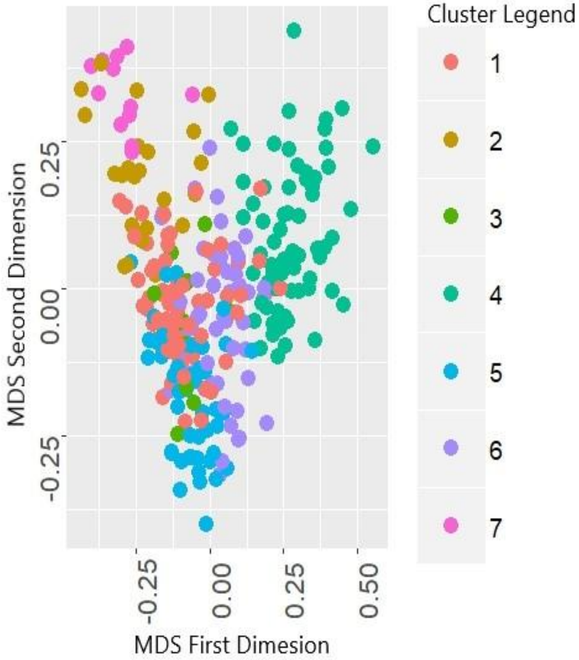
# Abstract Cluster Visualization

Mixture Model Clustering Results on the Dissimilarity Matrix

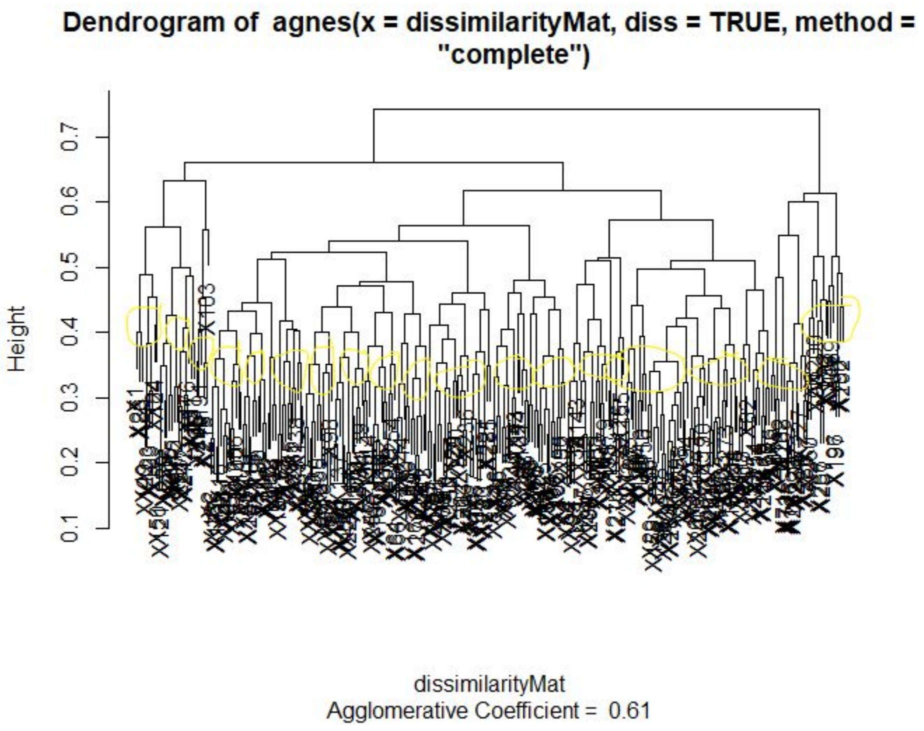


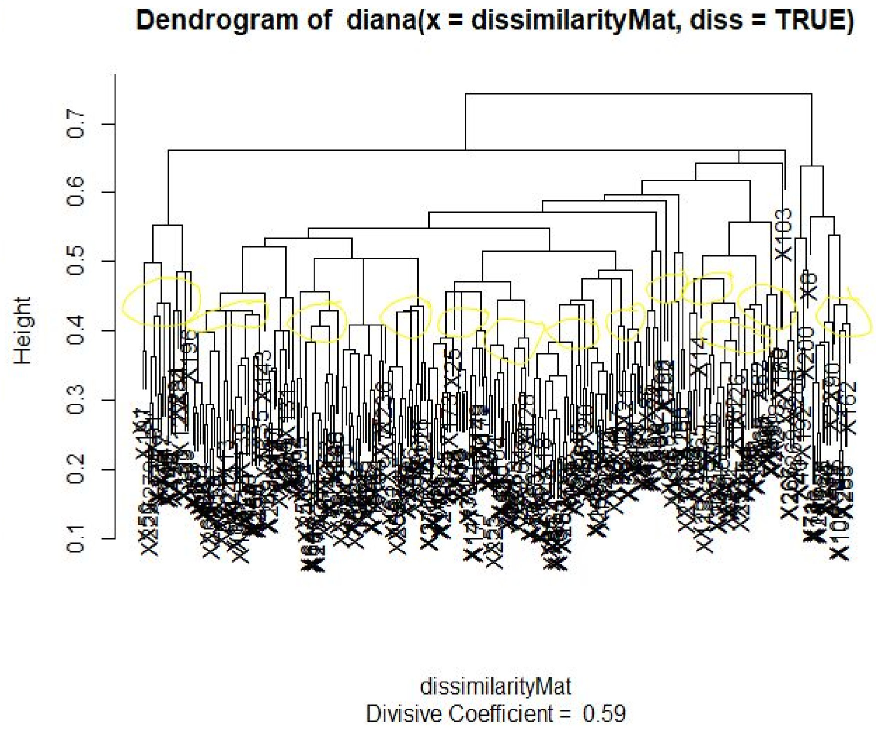
These two components explain 22.02 % of the point variability.

Classical Multidimensional Scaling Visualization
of the Mixture Model Clustering Results



# Exploring Subclusters

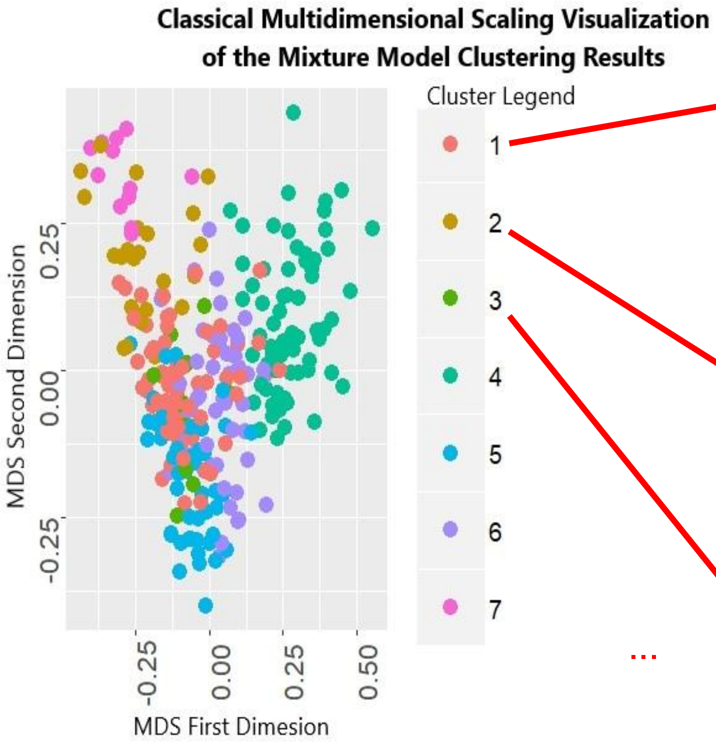




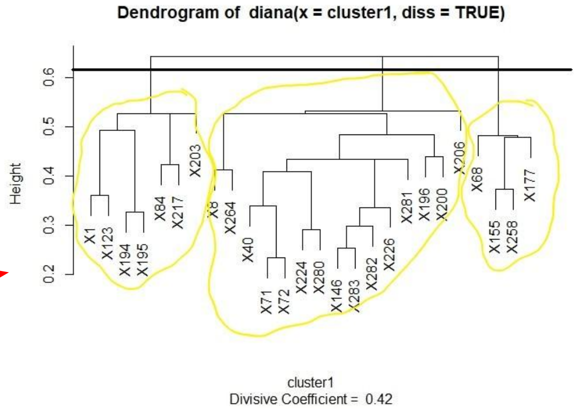
# Best Clustering Approach

1. Perform Mixture Model Clustering to obtain main clusters

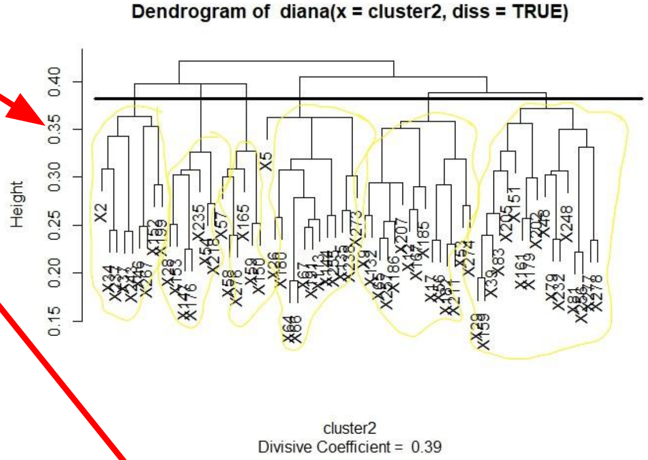
2. Perform divisive clustering to obtain clusters within the main clusters (subclusters)



Classical Multidimensional Scaling Visualization of the Mixture Model Clustering Results



Dendrogram of diana(x = cluster1, diss = TRUE)



Dendrogram of diana(x = cluster2, diss = TRUE)

...

...

# Statistical Analysis Results

- Used 6 different language models to get 6 different abstract similarity matrices:
  - Doc2Vec_300dim, Doc2Vec_apnews, Doc2Vec_Wiki, InferSent_GloVe, DocSim_LSI, WMD_GoogleNews300

- Applied the best clustering approach to each language model
  - Resulted in 6 different cluster analyses

- These 6 different analyses will result in 6 potential schedules
  - Which is the best ? Depends on evaluation criteria.

# Scheduling Problem

- Can be viewed as an optimization problem
  - Hard constraints: the schedule must follow them to be feasible
  - Soft constraints: better schedule if follow them
- Requirements:
  - Code must be written in R
  - Scheduling the student talks
  - Satisfying the hard constraints
  - Exporting the schedule as a CSV file
- Wishes:
  - Scheduling invited,contributed,keynote speakers
  - Scheduling panel discussion
  - Scheduling poster sessions
  - Exporting as a CSV google calendar

# Scheduling Problem

- The conference will be taken place in 3 consecutive days (Mon. to Wed.)
- Each day, three sessions except the last day (2 session)
- Each session can have up to n periods ( n is dependent on the number of talks in each session and the period length)
- 286 talks to be scheduled
- We have up to 10 rooms
- We have four different types (flags) of speakers (listed in order):
  - Keynote
  - Student
  - Invited
  - Contributed
- Each talk has also a priority number

# Hard Constraints [1]

**H1**:  Talks with the same cluster should not happen at the same time.

**H2**: A talk must only appear once in the whole conference.

**H3**: A cluster c is to be assigned Tc talks in the whole conference.

**H4**: A talk cannot be assigned to more than one period.

**H5**: A room cannot be assigned to more than one talk at each time slot.

**H6**:  The keynote talks should happen only once in each day of conference.

# Feasible Schedule

A feasible schedule follows all hard constraints:

$$\sum_{i=1}^{6} H_i = 0$$

# Soft Constraints

**S1**: The bigger size of a cluster, the bigger size of a room should be assigned.

**S2**: The number of times that talks with the same cluster, session, day happen in different rooms.

**S3**:  Talks with higher flag should be taken place first in each session.

**S4**: Keynote talks are prefered to happen in the first period of each day.

**S5**: Keynote talk are prefered to be in the largest rooms.

**S6**: No talk should be parallel with keynote talks. This penalty counts the number of talks that collides with the keynote talks.

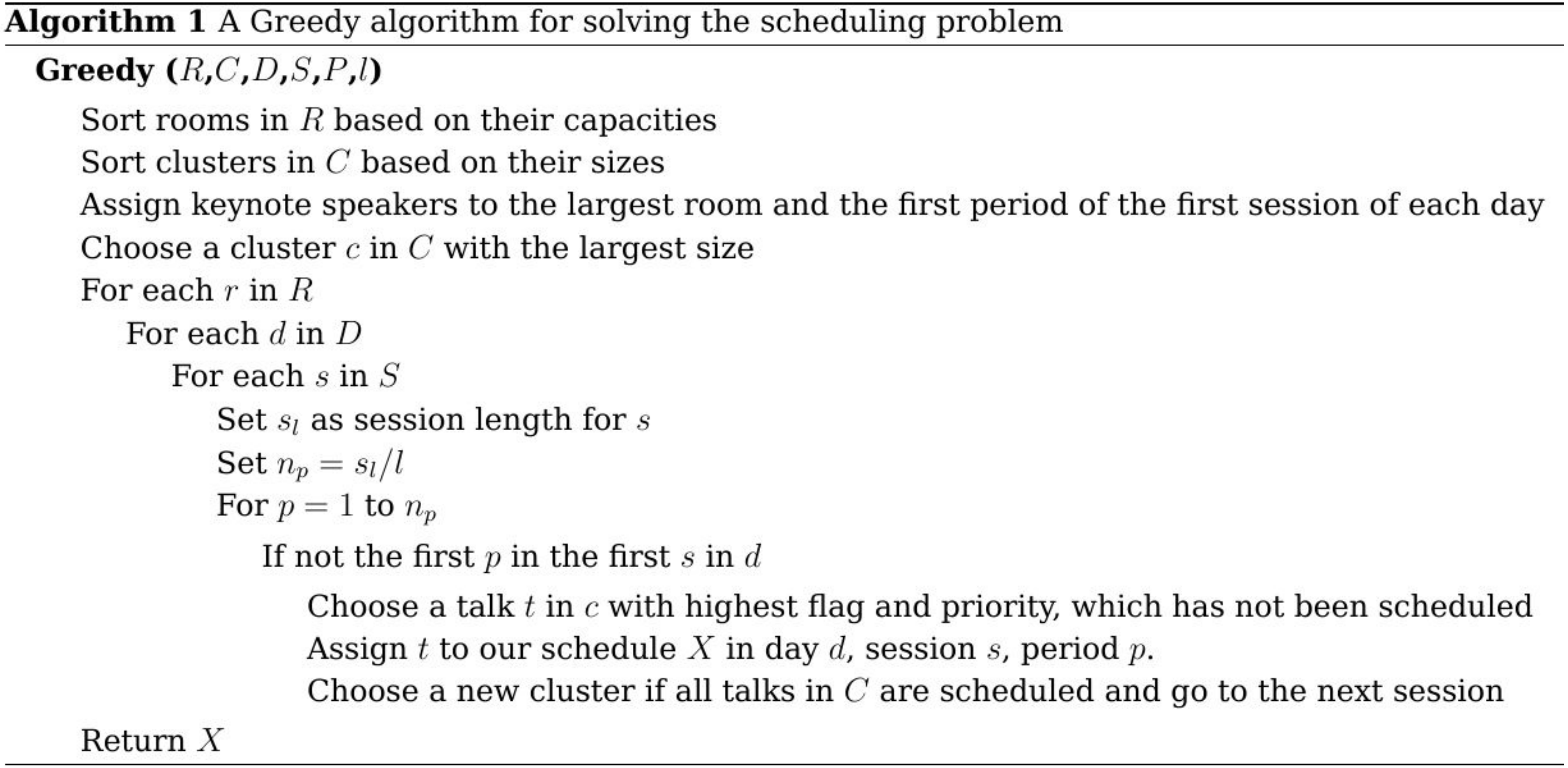
**S7**: In each session, between two speakers with the same flag, the one with higher priority should go first.

# Objective Function

The objective function is formulated as:

$$f = \sum_{i=1}^{7} \theta_i S_i$$

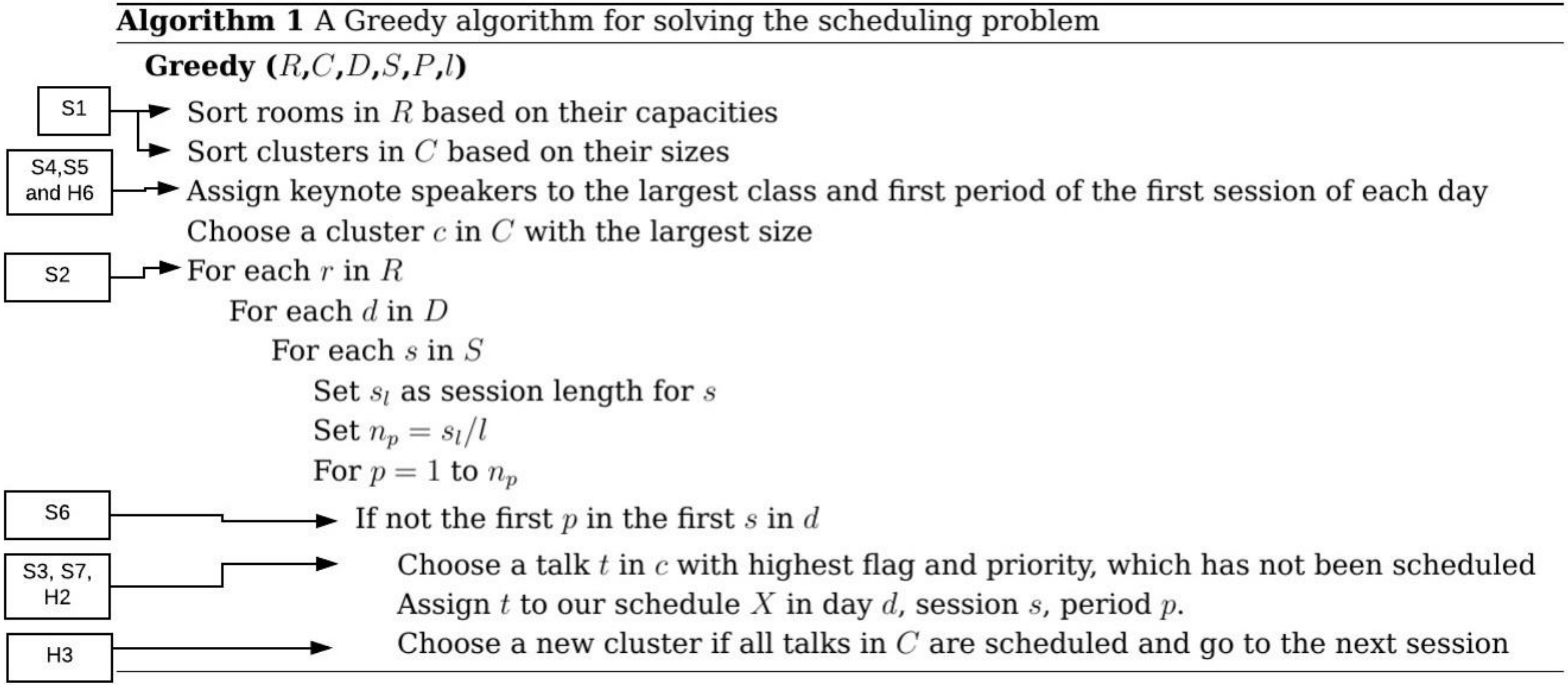
# Greedy Algorithm for Solving the Problem



**Algorithm 1** A Greedy algorithm for solving the scheduling problem

```
Greedy ($R$,$C$,$D$,$S$,$P$,$l$)
    Sort rooms in $R$ based on their capacities
    Sort clusters in $C$ based on their sizes
    Assign keynote speakers to the largest room and the first period of the first session of each day
    Choose a cluster $c$ in $C$ with the largest size
    For each $r$ in $R$
        For each $d$ in $D$
            For each $s$ in $S$
                Set $s_l$ as session length for $s$
                Set $n_p = s_l/l$
                For $p = 1$ to $n_p$
                    If not the first $p$ in the first $s$ in $d$
                        Choose a talk $t$ in $c$ with highest flag and priority, which has not been scheduled
                        Assign $t$ to our schedule $X$ in day $d$, session $s$, period $p$.
                        Choose a new cluster if all talks in $C$ are scheduled and go to the next session
    Return $X$
```

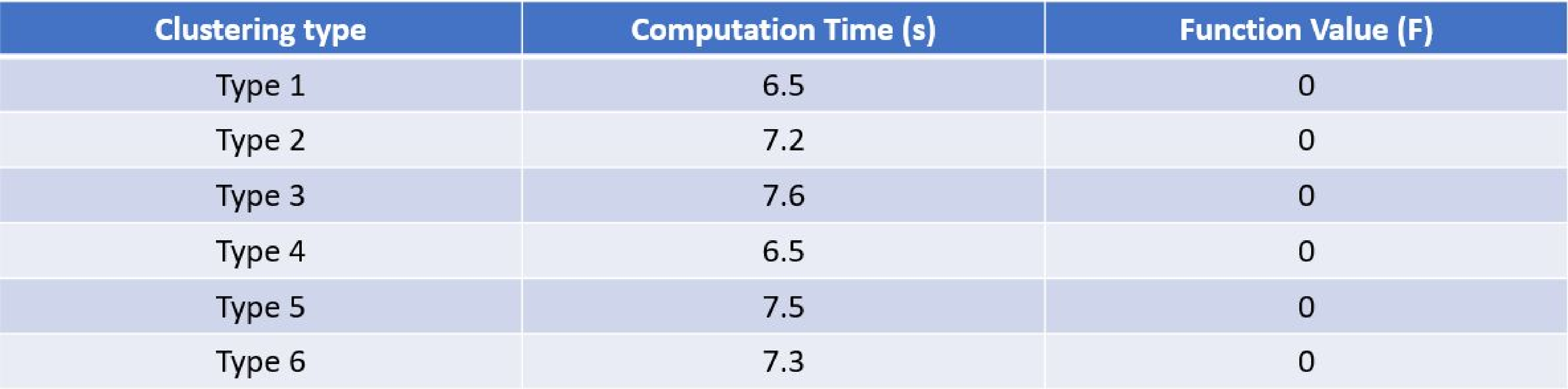
# Greedy Algorithm for Solving the Problem



**Algorithm 1** A Greedy algorithm for solving the scheduling problem

```
Greedy ($R$,$C$,$D$,$S$,$P$,$l$)
  Sort rooms in $R$ based on their capacities
  Sort clusters in $C$ based on their sizes
  Assign keynote speakers to the largest class and first period of the first session of each day
  Choose a cluster $c$ in $C$ with the largest size
  For each $r$ in $R$
    For each $d$ in $D$
      For each $s$ in $S$
        Set $s_l$ as session length for $s$
        Set $n_p = s_l/l$
        For $p = 1$ to $n_p$
          If not the first $p$ in the first $s$ in $d$
            Choose a talk $t$ in $c$ with highest flag and priority, which has not been scheduled
            Assign $t$ to our schedule $X$ in day $d$, session $s$, period $p$.
            Choose a new cluster if all talks in $C$ are scheduled and go to the next session
```

# Results for greedy algorithm



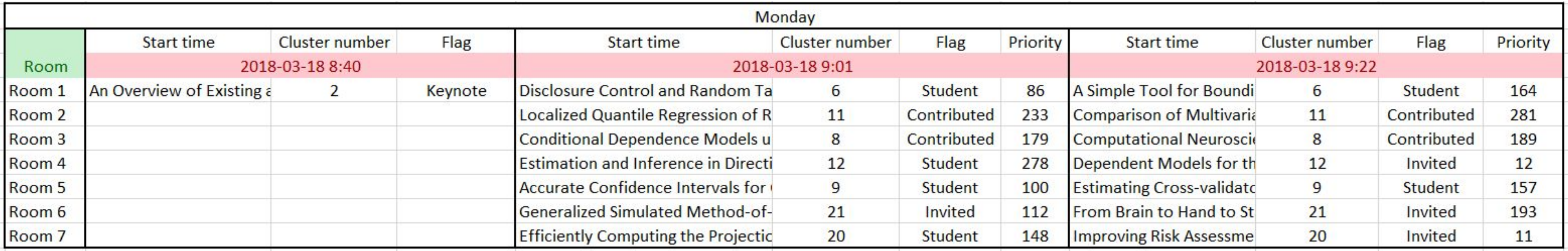
| Clustering type | Computation Time (s) | Function Value (F) |
|---|---|---|
| Type 1 | 6.5 | 0 |
| Type 2 | 7.2 | 0 |
| Type 3 | 7.6 | 0 |
| Type 4 | 6.5 | 0 |
| Type 5 | 7.5 | 0 |
| Type 6 | 7.3 | 0 |

Computation time for a function call: (20 s)

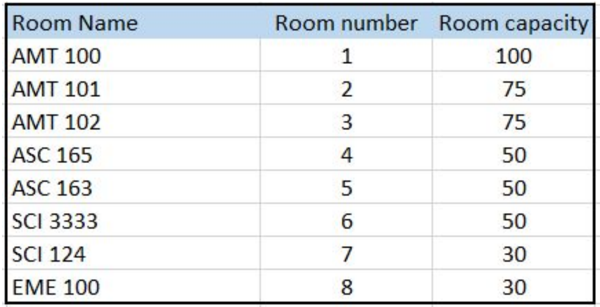
# Why a Heuristic Algorithm is not a Good Idea?

- Objective function is computationally expensive (20 s)
- The greedy algorithm find the schedule in less than 8 seconds.
- Considering the soft constraints mentioned, the greedy algorithm gives us the best schedule.

# Output Visualization (Clustering Method 1)



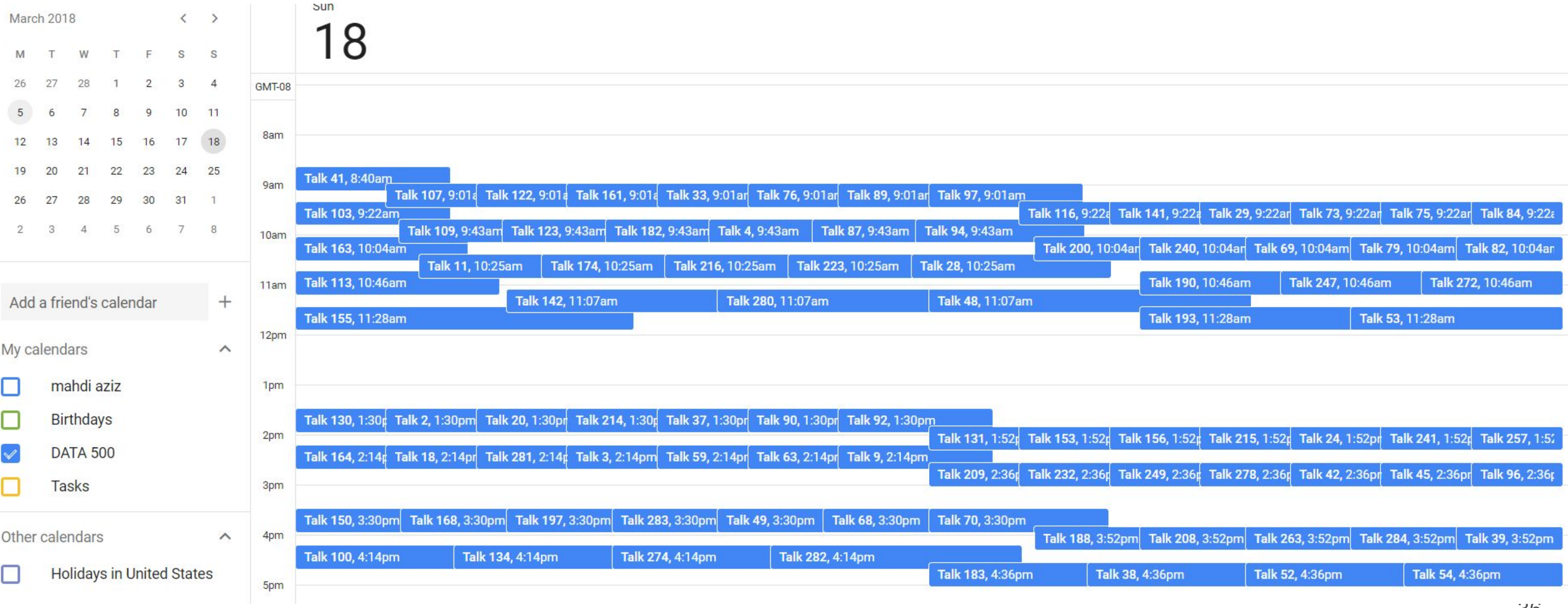
| Monday | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Start time | Cluster number | Flag | Start time | Cluster number | Flag | Priority | Start time | Cluster number | Flag | Priority |
| Room | 2018-03-18 8:40 | | | 2018-03-18 9:01 | | | | 2018-03-18 9:22 | | | |
| Room 1 | An Overview of Existing a | 2 | Keynote | Disclosure Control and Random Ta | 6 | Student | 86 | A Simple Tool for Boundi | 6 | Student | 164 |
| Room 2 | | | | Localized Quantile Regression of R | 11 | Contributed | 233 | Comparison of Multivaria | 11 | Contributed | 281 |
| Room 3 | | | | Conditional Dependence Models u | 8 | Contributed | 179 | Computational Neuroscie | 8 | Contributed | 189 |
| Room 4 | | | | Estimation and Inference in Directi | 12 | Student | 278 | Dependent Models for th | 12 | Invited | 12 |
| Room 5 | | | | Accurate Confidence Intervals for | 9 | Student | 100 | Estimating Cross-validatc | 9 | Student | 157 |
| Room 6 | | | | Generalized Simulated Method-of- | 21 | Invited | 112 | From Brain to Hand to St | 21 | Invited | 193 |
| Room 7 | | | | Efficiently Computing the Projectic | 20 | Student | 148 | Improving Risk Assessme | 20 | Invited | 11 |



| Room Name | Room number | Room capacity |
|---|---|---|
| AMT 100 | 1 | 100 |
| AMT 101 | 2 | 75 |
| AMT 102 | 3 | 75 |
| ASC 165 | 4 | 50 |
| ASC 163 | 5 | 50 |
| SCI 3333 | 6 | 50 |
| SCI 124 | 7 | 30 |
| EME 100 | 8 | 30 |

# Output Google CSV Calendar

Link



| M | T | W | T | F | S | S |
|---|---|---|---|---|---|---|
| 26 | 27 | 28 | 1 | 2 | 3 | 4 |
| 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| 12 | 13 | 14 | 15 | 16 | 17 | 18 |
| 19 | 20 | 21 | 22 | 23 | 24 | 25 |
| 26 | 27 | 28 | 29 | 30 | 31 | 1 |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 |

□ mahdi aziz
□ Birthdays
☑ DATA 500
□ Tasks
□ Holidays in United States

Talk 41, 8:40am
Talk 107, 9:01a Talk 122, 9:01a Talk 161, 9:01a Talk 33, 9:01ar Talk 76, 9:01ar Talk 89, 9:01ar Talk 97, 9:01am
Talk 103, 9:22am Talk 116, 9:22a Talk 141, 9:22a Talk 29, 9:22ar Talk 73, 9:22ar Talk 75, 9:22ar Talk 84, 9:22a
Talk 109, 9:43am Talk 123, 9:43am Talk 182, 9:43am Talk 4, 9:43am Talk 87, 9:43am Talk 94, 9:43am
Talk 163, 10:04am Talk 200, 10:04ar Talk 240, 10:04ar Talk 69, 10:04am Talk 79, 10:04am Talk 82, 10:04ar
Talk 11, 10:25am Talk 174, 10:25am Talk 216, 10:25am Talk 223, 10:25am Talk 28, 10:25am
Talk 113, 10:46am Talk 190, 10:46am Talk 247, 10:46am Talk 272, 10:46am
Talk 142, 11:07am Talk 280, 11:07am Talk 48, 11:07am
Talk 155, 11:28am Talk 193, 11:28am Talk 53, 11:28am
Talk 130, 1:30p Talk 2, 1:30pm Talk 20, 1:30pr Talk 214, 1:30p Talk 37, 1:30pr Talk 90, 1:30pr Talk 92, 1:30pm
Talk 131, 1:52p Talk 153, 1:52p Talk 156, 1:52p Talk 215, 1:52p Talk 24, 1:52pr Talk 241, 1:52p Talk 257, 1:5
Talk 164, 2:14p Talk 18, 2:14pr Talk 281, 2:14p Talk 3, 2:14pm Talk 59, 2:14pr Talk 63, 2:14pr Talk 9, 2:14pm
Talk 209, 2:36p Talk 232, 2:36p Talk 249, 2:36p Talk 278, 2:36p Talk 42, 2:36pr Talk 45, 2:36pr Talk 96, 2:36p
Talk 150, 3:30pm Talk 168, 3:30pm Talk 197, 3:30pm Talk 283, 3:30pm Talk 49, 3:30pm Talk 68, 3:30pm Talk 70, 3:30pm
Talk 188, 3:52pm Talk 208, 3:52pm Talk 263, 3:52pm Talk 284, 3:52pm Talk 39, 3:52pm
Talk 100, 4:14pm Talk 134, 4:14pm Talk 274, 4:14pm Talk 282, 4:14pm
Talk 183, 4:36pm Talk 38, 4:36pm Talk 52, 4:36pm Talk 54, 4:36pm

# Leadership

- Agile software development was utilized due to the alignment of its manifesto
  - Working software
  - Customer Collaboration

- Assign team members to components of the project based on domain knowledge

- Ten minute SCRUM weekly meeting on Thursdays to resolve any issues
  - Mahdi performed the role of SCRUM master

# References

[1] Zhang, D., Liu, Y., M’Hallah, R., & Leung, S. C. H. (2010). A simulated annealing with a new neighborhood structure based algorithm for high school timetabling problems. *European Journal of Operational Research*, *203*(3), 550–558.

[2] Kevin, L, Gordon, M. (2015, June 10). Graphic Representations of word2vec and doc2vec. Retrieved December 07, 2017, from https://groups.google.com/forum/#!topic/gensim/EwK-6JgkWVI

# Thank You.