



SciBot: A Data Science Research Bot

— Data Science Course Project

Meng Jiang
Data Science



Dr. Zordon: *Alpha, find me a data scientist!*

<http://powerrangers.wikia.com/wiki/Zordon>

Salaries in \$ (USD)

Average

Min

Max



Data Scientist
Facebook
61 salaries

\$134,715
per year

\$100k

\$180k



Data Scientist
Microsoft
41 salaries

\$123,328
per year

\$94k

\$152k



Data Scientist
IBM
36 salaries

\$109,864
per year

\$81k

\$144k



Data Scientist
Apple
13 salaries

\$145,974
per year

\$120k

\$177k



Data Scientist
Google
11 salaries

\$152,856
per year

\$103k

\$210k



Data Scientist
Uber
17 salaries

\$122,432
per year

\$103k

\$152k



Data Scientist
Civis Analytics
16 salaries

\$76,284
per year

\$61k

\$95k



Data Scientist
LinkedIn
11 salaries

\$132,196

\$111k

\$4261k

*They are
too expensive!*



*How many data science papers have you read?
(#papers cited in your PhD thesis)*

256

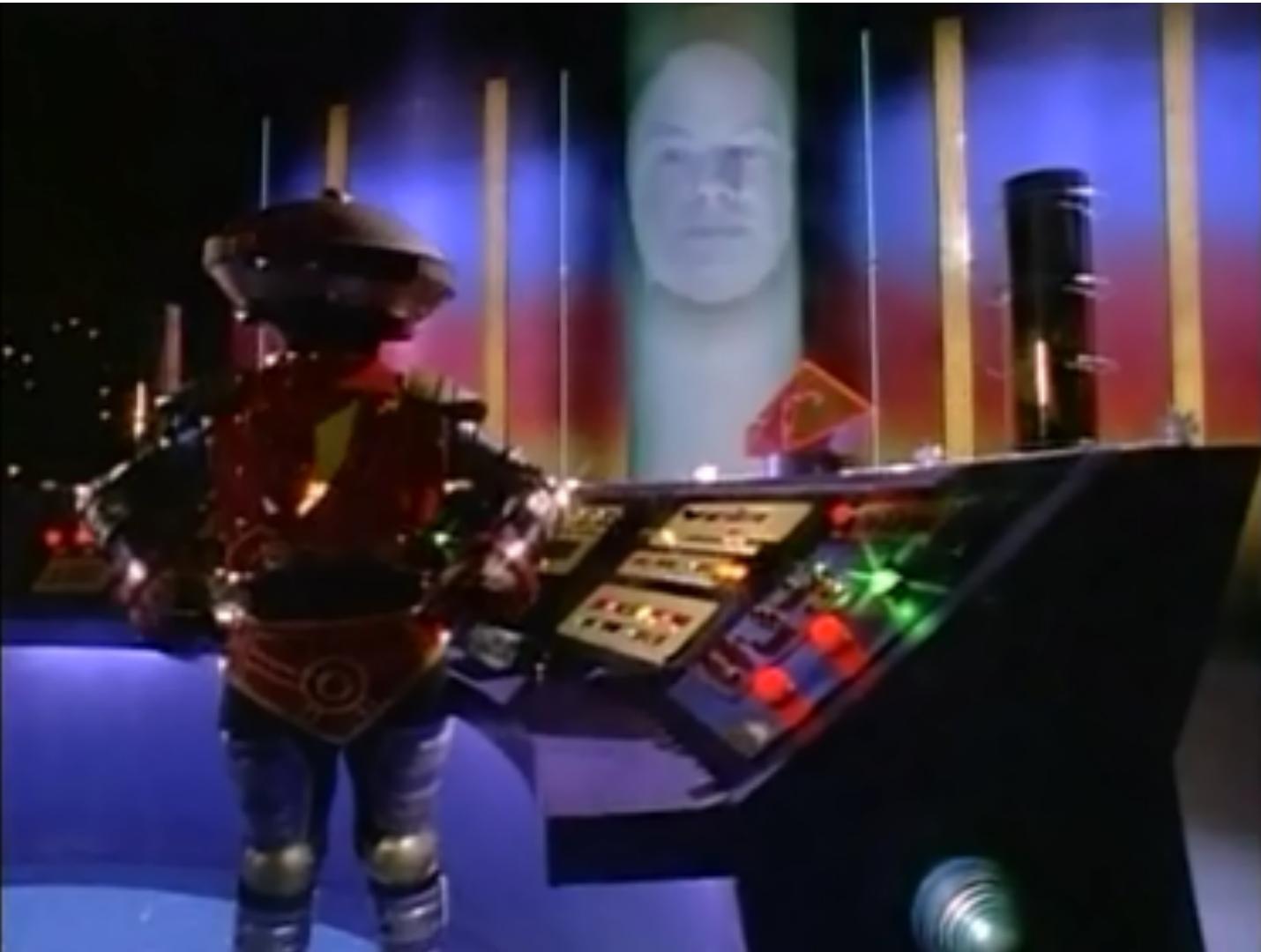
172

185

120

210

How many papers we grabbed from the web?



5,853 papers from 4 data science conferences!

KDD: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (August)

1994 – 2015: collected 1,646 papers

'16: San Francisco; '17: Halifax, Canada; '18: London

Due: February



ICDM: IEEE International Conference on Data Mining (December)

2001 – 2015: collected 881 papers

'16: Barcelona; '17: New Orleans

Due: June



WWW: International World Wide Web Conference (April)

2001 – 2015: collected 2,975 papers

'16: Montreal, Canada; '17: Perth, Australia; '18: Lyon, France

Due: October



WSDM: ACM International Conference on Web Search and Data Mining (February)

2008 – 2015: collected 351 papers

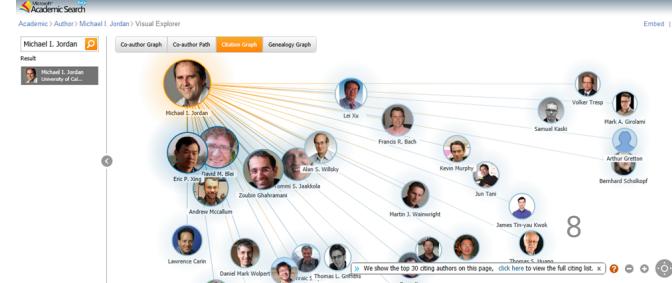
'16: San Francisco; '17: Cambridge, UK; '18: Los Angeles

Due: August



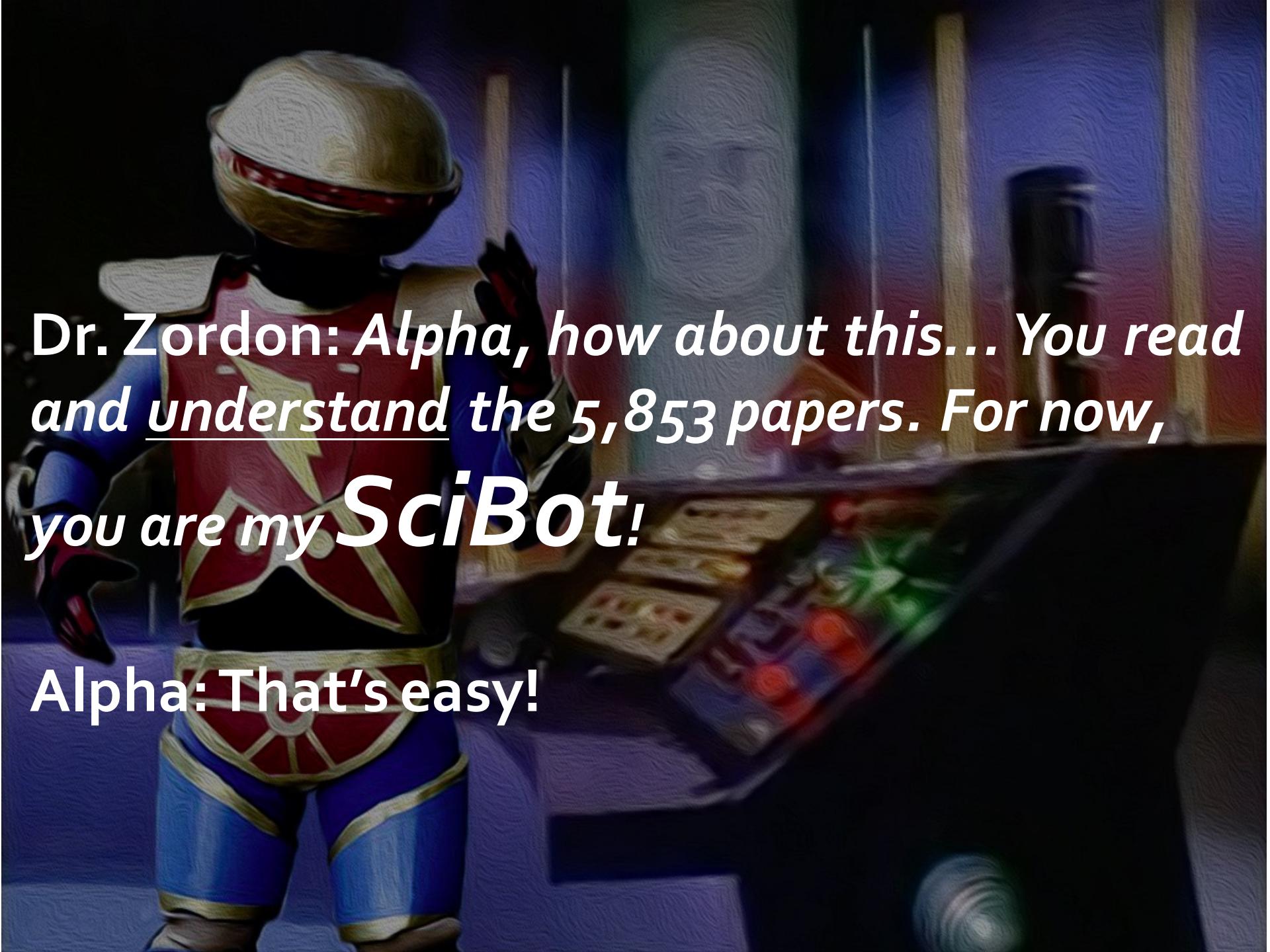
Dataset Introduction

- pdf (5.5G): <https://www.dropbox.com/s/46oh772tpuceew5/pdf.zip?dl=0>
 - “icdm01-d3.pdf”
 - Not recommended
- text (270M): <https://www.dropbox.com/s/oof7qjb5mobmfvt/text.zip?dl=0>
 - “icdm01-d3.txt”
 - **Recommended to skip the REFERENCES section.**
- microsoft (100M): <https://www.dropbox.com/s/0gqzhbddopmk5wm/microsoft.zip?dl=0>
 - **index.txt:** PID, PDFID
 - **Papers.txt:** PID, TITLE, YEAR, CONF
 - **PaperKeywords.txt:** PID, KEYWORD
 - **PaperAuthorAffiliation.txt:** PID, AID, FID, AFF, SID
 - **Authors.txt:** AID, AUT



How long does it take you to read all the five thousand papers?

$$5853 / 365 / 5 = 3.207 \text{ years...}$$

A Teenage Mutant Ninja Turtles character, Dr. Zordon, is shown from the waist up. He is wearing his signature metallic suit with red and blue panels and gold trim. He is holding a small, glowing electronic device with a screen and several buttons. The background is dark and slightly blurred.

*Dr. Zordon: Alpha, how about this... You read
and understand the 5,853 papers. For now,
you are my SciBot!*

Alpha: That's easy!

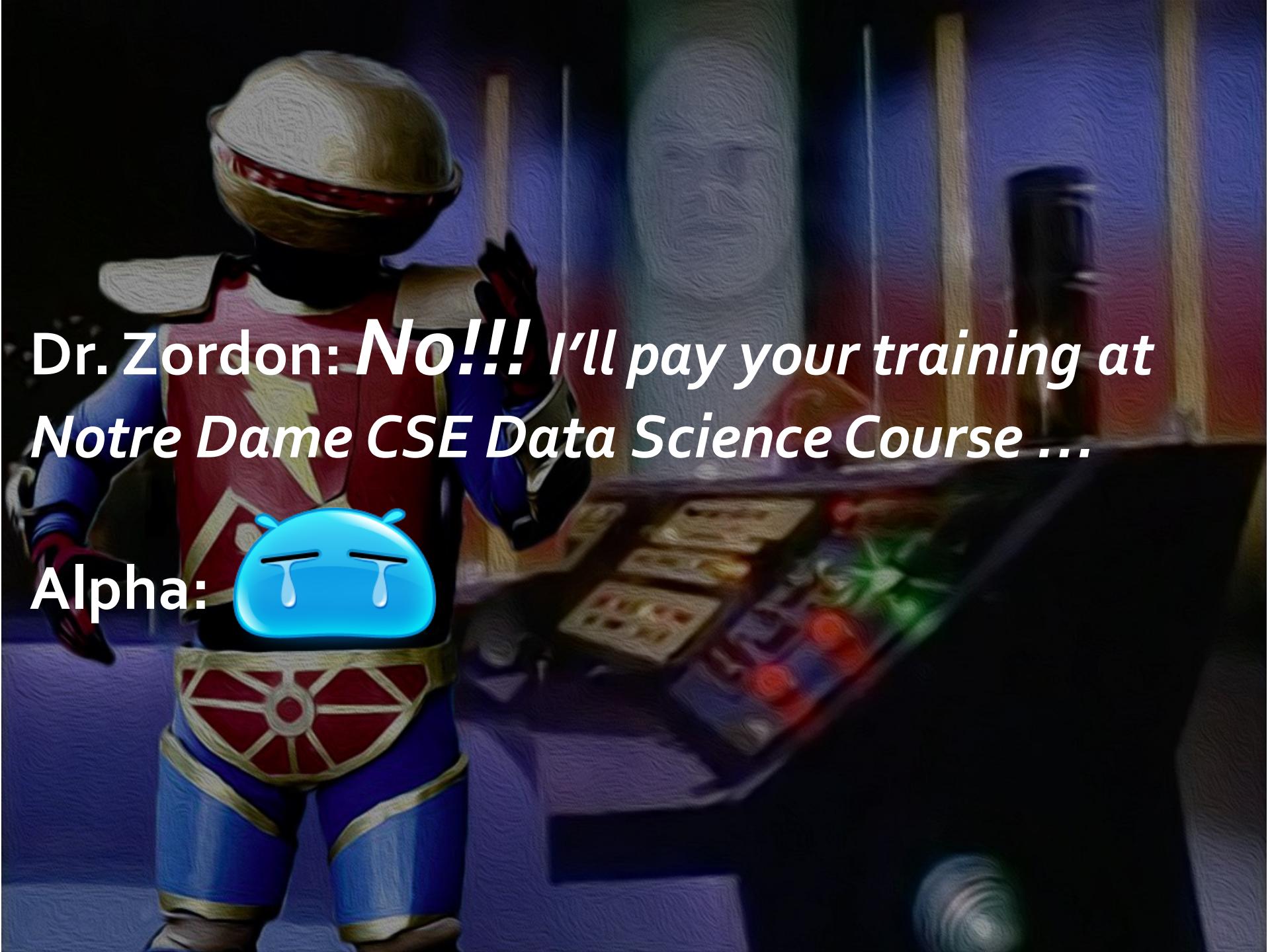
Author	#paper
jiawei han	94
philip s yu	82
christos faloutsos	64
jieping ye	42
ravi kumar	41
jure leskovec	40

Word	#
the	2,154,841
of	1,127,177
and	818,732
a	797,101
in	727,085
to	718,857

Char	#
e	20,953,911
t	14,791,644
i	12,864,269
a	12,716,160
o	11,753,274
n	11,673,479

Am I a smart Bot?

abc

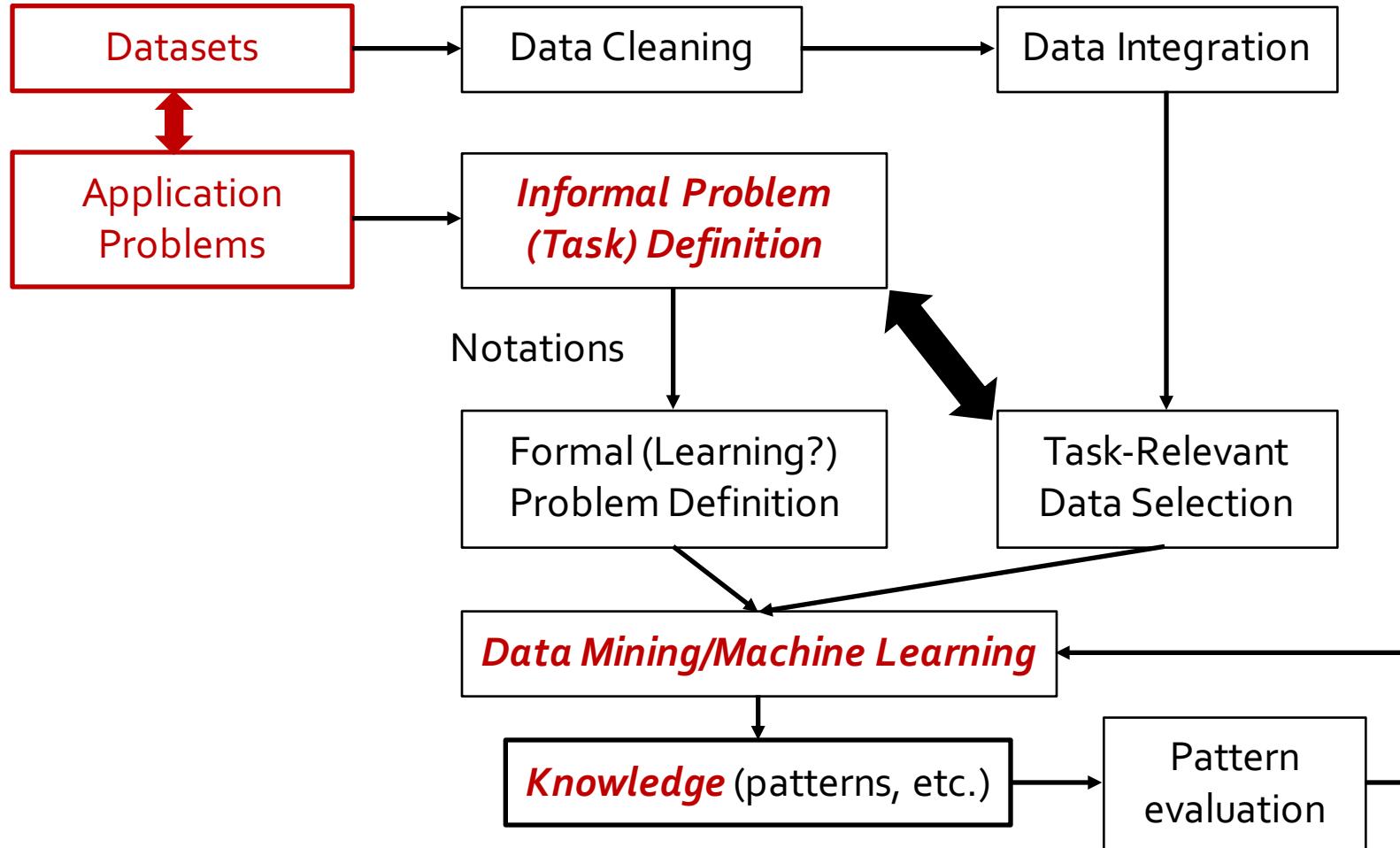
A Power Ranger in blue and red armor is holding a glowing blue orb with a face. The background is dark and textured.

Dr. Zordon: **No!!!** I'll pay your training at
Notre Dame CSE Data Science Course ...

Alpha:



Data Science Research



Seven Required Tasks: Task 4-7

Concepts and Methods:

Task 4: Collaboration discovery → Frequent pattern mining

Task 5: Problem-method
association mining → Association rule mining

Task 6: Problem/method/author-
to-conference classification → Classification

Task 7: Paper clustering → Clustering

Task 4

Concepts and Methods:

Task 4: Collaboration discovery → Frequent pattern mining

Task 5: Problem-method
association mining → Association rule mining

Task 6
to-con

Task 7

Given the paper/keyword/conference-author data,
find frequent author-sets (as patterns): which
two/three/four authors often collaborate together?

Techs: Frequent pattern mining (Apriori, FP-Growth).

Hints: Here each paper is considered as a transaction.
Each author is an item.

Task 5

Concepts and Methods:

Task 4: Collaboration discovery → Frequent pattern mining

**Task 5: Problem-method
association mining** → Association rule mining

Task 6: Problem
to-conference

Task 7: Paper
of high support and confidence.

Given the paper-problem-method data, find strong association rules:
problem X → method Y, or
method X → problem Y,

Techs: Association rule mining.

Task 6

Concepts and Methods:

Task 4: Collaboration discovery → Frequent pattern mining

Given a problem/method/author, predict if a conference has papers of it.

Tas
ass
Techs: Binary classification (Naïve Bayes, Decision Tree).

**Task 6: Problem/method/author- → Classification
to-conference classification**

Hints:

1. What are the attributes (features) you want to use?
2. How to set up training and testing? Please evaluate the performance on different features, different models, and different setups.

Task 7

Concepts and Methods:

Task Given a set of papers, cluster them into K groups.

Techs: K-partitioning clustering methods (K-Means).

Task Hints:

1. What are the attributes (features) you want to use?
2. Suppose $K = 4$ and the ground-truth is the conference.
Please evaluate the performance on different features
and different methods.

Task 7: Paper clustering

→ Clustering

Seven Required Tasks: Task 1-3

- Process raw data into *clean, multi-dimensional* tabulated data.

PID	YEAR	CONF	TITLE	KEYWORDS
776E2648	2010	kdd	new perspectives	machine learning networks supervised learning variance reduction
784B7EF4	2014	kdd	improving management	clustering data mining invasive species networks risk assessment
7E395F14	2008	icdm	start globally optimal	data mining global optimization learning artificial intelligence optimization probabilistic

SEQ_AUTHOR_AFFS

1:109A673C:ryan n lichtenwalter:066A71BC:university of notre dame|2:7DA9ABBD:jake t lussier:066A71BC:university of notre dame
1:7E5C680D:jian xu:066A71BC:university of notre dame|2:5DAE606C:thanuka l wickramarathne:066A71BC:university of notre dame
1:7776FD94:david a cieslak:066A71BC:university of notre dame|2:76014D6E:nitesh v chawla:066A71BC:university of notre dame

PROBLEMS	METHODS	DATASETS	METRICS
???	???	???	???
???	???	???	???
???	???	???	???

From structured tables

From unstructured paper text

Large-Scale Distributed Bayesian Matrix Factorization using Stochastic Gradient MCMC

Sungjin Ahn *
University of California, Irvine
sungjia@ics.uci.edu

Anoop Korattikara †
Google
kbanoop@google.com

Nathan Liu
Yahoo Labs
nanliu@yahoo-inc.com

Suju Rajan
Yahoo Labs
suju@yahoo-inc.com

Max Welling ‡
University of Amsterdam
m.welling@uva.nl

ABSTRACT

Despite having various attractive qualities such as high prediction accuracy and the ability to quantify uncertainty and avoid overfitting, Bayesian Matrix Factorization has not been widely adopted because of the prohibitive cost of inference. In this paper, we propose a scalable distributed Bayesian matrix factorization algorithm using stochastic gradient MCMC. Our algorithm, based on Distributed Stochastic Gradient Langevin Dynamics, can not only match the prediction accuracy of standard MCMC methods like Gibbs sampling, but at the same time is as fast and simple as stochastic gradient descent. In our experiments, we show that our algorithm can achieve the same level of prediction accuracy as Gibbs sampling an order of magnitude faster. We also show that our method reduces the prediction error as fast as distributed stochastic gradient descent, achieving a 4.1% improvement in RMSE for the Netflix dataset and an 1.8% for the Yahoo music dataset.

Categories and Subject Descriptors

G.4 [Mathematical Software]: Algorithm design and analysis;
Parallel and vector implementations

Keywords

Large-Scale, Distributed, Matrix Factorization, MCMC, Stochastic Gradient, Bayesian Inference

1. INTRODUCTION

Recommender systems have become a popular tool to understand customers and their interest range between music recommendation (Pandora), movie recommendation (Amazon), news recommendation (Yahoo) to partner recommendation. Recommender systems represent a personalized help filter at an individual level the enormous amount of information that is available to us. Given the exponential growth of information, recommender systems are likely to play a major role to manage our information streams.

During 2006-2011 Netflix [6] ran a competition around the world could develop and test new technology on Netflix movie rating data. A few lessons learnt from that exercise. First, matrix factorization is very well compared to nearest neighbor type of methods.averaging over many different models pays off in terms of accuracy. One particularly effective model is probabilistic matrix factorization (BPBMF) [26] which averaged over samples from the posterior distribution proved prediction accuracy, a full Bayesian approach with additional advantages such as probability intervals, robustness against overfitting, prior knowledge and side-information [2].

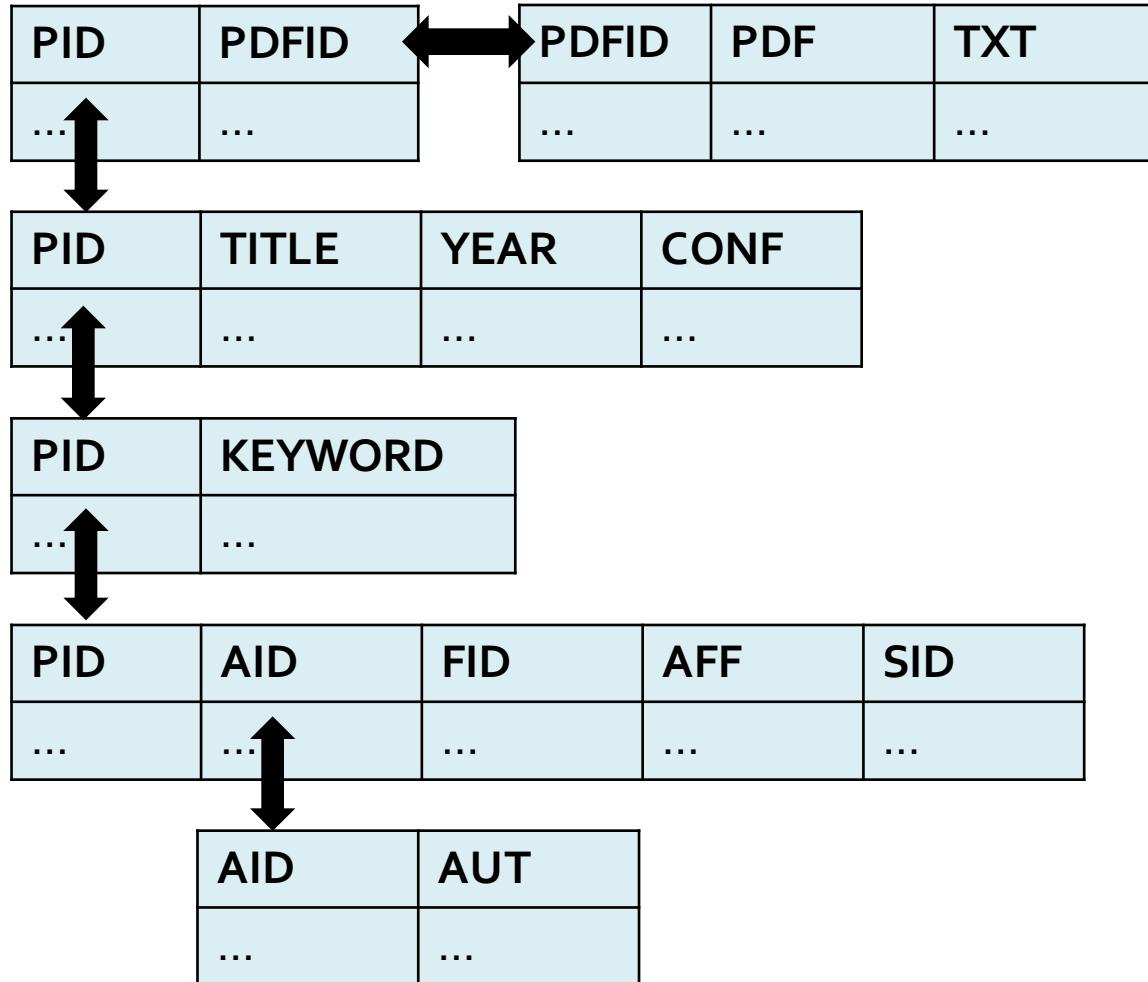
Unfortunately, since the number of user easily run into the billions, posterior inference

Dimension	Value	Dimension	Value
PROBLEM	prediction	DATASET	netflix
METRIC	prediction accuracy	DATASET	yahoo music
PROBLEM	uncertainty	PROBLEM	recommender systems
PROBLEM	over-fitting	PROBLEM	music recommendation
METHOD	Bayesian matrix factorization	PROBLEM	book recommendation
PROBLEM	prohibitive cost of inference	PROBLEM	movie recommendation
METHOD	scalable distributed Bayesian matrix factorization	PROBLEM	news recommendation
METHOD	stochastic gradient mcmc	PROBLEM	partner recommendation
METHOD	distributed stochastic gradient langevin dynamics	PROBLEM	recommender systems
METHOD	mcmc	DATASET	netflix
METHOD	gibbs sampling	PROBLEM	recommender
METHOD	stochastic gradient descent	DATASET	netflix movie rating
PROBLEM	prediction	METHOD	bayesian probabilistic matrix factorization
METRIC	accuracy	METHOD	posterior distribution
METHOD	gibbs sampling	METHOD	bayesian analysis
METRIC	prediction error	METRIC	confidence intervals
METHOD	distributed stochastic gradient descent	METRIC	robustness
METRIC	rmse		



Labeling 5,853 papers?
You must be kidding me!!!

Task 1: Data Cleaning and Integration



Task 2: Entity Name Recognition

- 2.1 Candidate generation + 2. 2 Quality assessment

PHRASE	QUALITY SCORE AS AN ENTITY NAME
latent_dirichilet_allocation	0.999
support_vector_machine	0.999
support_vector_machines	0.999
singular_value_decomposition	0.998
information_retrieval	0.998
mean_average_precision	0.997
collaborative_filtering	0.995
...	...
is_a	0.001

T2-1: Entity Name Candidate Generation

- You are recommended to try both methods.
- Method 1: N-Grams
 - 2-grams: “matrix factorization”
 - 3-grams: “bayesian matrix factorization”
 - ...
- Method 2: Frequent pattern mining
 - Minimum support (frequency)
 - Apriori property
- Method 3: Hand-crafted rules
 - “... Support Vector Machines ...”
 - “... support vector machines (SVMs) ...”
 - “... non-negative matrix factorization (NMF) ...”

Examples (Method 3): Rule Matching

```
SVM      345      Support Vector Machine:137|Support Vector Machines:78|support vector machine:74|support vector machines:43|Support vector machine:7|Support vector machines:4|Support Vectors Machine:2
LDA      289      Latent Dirichlet Allocation:173|latent Dirichlet allocation:58|Linear Discriminant Analysis:19|linear discriminant analysis:15|Latent Dirichlet allocation:12|latent dirichlet allocation:3|Linear discriminant analysis:2|linear discriminative analysis:2|linear discrimination algorithm:1|Latent Dirchlet Allocation:1|latent Dirichlet Allocation:1|latent Drichlet association:1|latent dirichelet allocation:1
MAP      202      Mean Average Precision:61|mean average precision:60|maximun a posterio ri:47|maximun a posterior:10|Maximum a Posteriori:8|Maximum A Posterior:6|Maximum A Posteriori:4|Mean average precision:3|Maximum a posteriori:3
IR       164      information retrieval:86|Information Retrieval:53|implicit relatedness :5|instance reduction:2|Information retrieval:2|Internal Representation:2|internal representation:2|Intermediate Representation:2|imbalance ratio:1|individually rational:1|Instance Ranking:1|Interactions Rank:1|Identifier Renaming:1|indirect retweet:1|implicated relatedness:1|indirect Retweet:1|image rank:1|intermediate representation:1
SVD      159      singular value decomposition:74|Singular Value Decomposition:73|Singular value decomposition:3|Singular Value Decompositions:3|singular value decompositions :2|singular value decomposing:1|Single Value Decomposition:1|singular vector decomposition:1|Singular vector decomposition:1
WWW      145      World Wide Web:140|World wide web:4|world wide web:1
PCA      136      Principal Component Analysis:62|principal component analysis:35|Principal Component Analysis:11|Principal Components Analysis:7|Principal component analysis :7|principle component analysis:6|principal components analysis:5|principle components analysis:2|partial completeness assumption:1
CF       122      collaborative filtering:51|Collaborative Filtering:37|Collaborative filtering:22|concept feedback:2|Collaborating Filtering:1|Column Fusing:1|click feedback :1|consistent framework:1|Content Fragments:1|clarification forms:1|Certainty Factor:1 |clustering feature:1|cluster feature:1|certainty factor:1
EM       112      Expectation Maximization:80|expectation maximization:24|Ensemble Media n:2|Expected Maximization:1|Expectation maximization:1|Entity Matching:1|Expectation Maximisation:1|edit map:1|Exact Match:1
RDF      96       Resource Description Framework:82|Resource description framework:9|resource description framework:2|Resource Description framework:2|Resource Definition Format:1
```

Examples (Method 3): Re-Matching for Support

latent_dirichlet_allocation	247	LDA:247
support_vector_machine	218	SVM:218
support_vector_machines	214	SVM:125 SVMs:89
singular_value_decomposition	150	SVD:150
world_wide_web	145	WWW:145
information_retrieval	141	IR:141
mean_average_precision	124	MAP:124
collaborative_filtering	110	CF:110
expectation_maximization	105	EM:105
principal_component_analysis	104	PCA:104
resource_description_framework	95	RDF:95
neural_information_processing_systems	93	NIPS:93
stochastic_gradient_descent	91	SGD:91
minimum_description_length	78	MDL:78
natural_language_processing	77	NLP:77
normalized_mutual_information	76	NMI:76
mean_reciprocal_rank	76	MRR:76
document_object_model	71	DOM:71
mean_absolute_error	69	MAE:69
logistic_regression	67	LR:67
normalized_discounted_cumulative_gain	66	NDCG:66
latent_semantic_indexing	63	LSI:63
maximum_a_posteriori	62	MAP:62
mean_squared_error	62	MSE:62
receiver_operating_characteristic	58	ROC:57 RoC:1
dynamic_time_warping	58	DTW:58
markov_chain_monte_carlo	57	MCMC:57
naive_bayes	57	NB:57
transactions_on_information_systems	56	TOIS:56
probabilistic_latent_semantic_analysis	52	PLSA:52
directed_acyclic_graph	51	DAG:51
open_directory_project	50	ODP:50
non-negative_matrix_factorization	50	NMF:50
national_science_foundation	50	NSF:50
hidden_markov_models	47	HMMs:26 HMM:21
maximum_likelihood_estimation	47	MLE:47
root_mean_square_error	47	RMSE:47
hidden_markov_model	45	HMM:45
conditional_random_fields	45	CRFs:26 CRF:19
matrix_factorization	44	MF:44
information_extraction	42	IE:42
named_entity_recognition	42	NER:42
root_mean_squared_error	39	RMSE:39
cumulative_distribution_function	39	CDF:39
nonnegative_matrix_factorization	39	NMF:39

T2-2: Entity Name Quality Assessment

- You are recommended to try all the measures.
- Measure 1: Support (Frequency)
 - Sentence, Paragraph, Document
- Measure 2: Outlier-ness measures (t-score, Z-score, etc.)

Church et al. <https://pdfs.semanticscholar.org/ccd5/898c79f90c25e24222410d5613edfd007985.pdf>

$$\text{sig}(P_1, P_2) \approx \frac{f(P_1 \oplus P_2) - \mu_0(P_1, P_2)}{\sqrt{f(P_1 \oplus P_2)}}$$

"bayesian matrix factorization"?

Pedersen. <http://www.d.umn.edu/~tpederse/Pubs/scsug96.pdf>

El-Kishky et al. <http://www.vldb.org/pvldb/vol8/p305-ElKishky.pdf>

- Other measures:
 - "in_this", "at_the" ...
 - "turn_out_to_be", "increased_by" ...
 - "in_this_paper" ...
 - "data sets", "real data" ...

Thresholds?

Classification?

Task 3: Entity Typing

- Trigger words
 - METHOD: method algorithm model approach framework process scheme implementation procedure strategy architecture
 - PROBLEM: problem technique process system application task evaluation tool paradigm benchmark software
 - DATASET: data dataset database
 - METRIC: value score measure metric function parameter
- Classification using contextual features and **Evaluation**

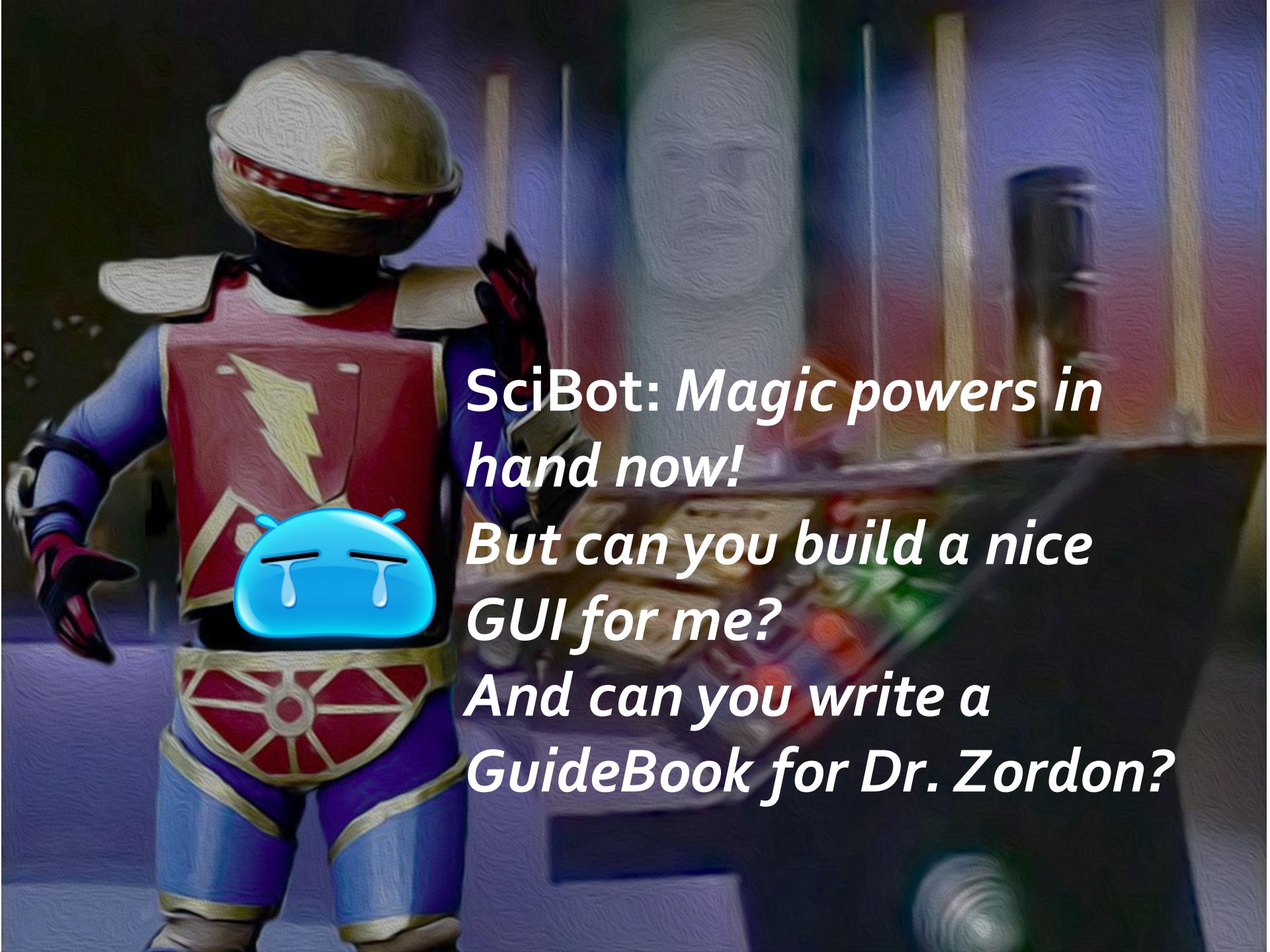
collaborative_filtering	967	METHOD: 729 PROBLEM: 217 DATASET: 18 METRIC: 3
feature_selection	895	METHOD: 708 PROBLEM: 138 METRIC: 48 DATASET: 1
link_prediction	641	PROBLEM: 348 METHOD: 204 METRIC: 89
active_learning	588	METHOD: 492 PROBLEM: 96
latent_dirichlet_allocation	538	METHOD: 530 PROBLEM: 6 DATASET: 2
matrix_factorization	518	METHOD: 354 PROBLEM: 164
supervised_learning	503	METHOD: 347 PROBLEM: 153 METRIC: 2 DATASET: 1
logistic_regression	490	METHOD: 443 PROBLEM: 31 METRIC: 16
expectation_maximization	434	METHOD: 426 PROBLEM: 5 METRIC: 3
social_network	391	DATASET: 280 METHOD: 68 PROBLEM: 40 METRIC: 3
binary_classification	391	PROBLEM: 328 METHOD: 48 DATASET: 11 METRIC: 4
resource_description_framework	367	DATASET: 267 METHOD: 92 PROBLEM: 8
random_walk	367	METHOD: 313 DATASET: 27 METRIC: 14 PROBLEM: 13

Examples (Entity Typing)

METHOD	latent_dirichlet_allocation	247	LDA:247
METHOD	support_vector_machine	218	SVM:218
METHOD	support_vector_machines	214	SVM:125 SVMs:89
METHOD	singular_value_decomposition	150	SVD:150
DATASET	world_wide_web	145	WWW:145
PROBLEM	information_retrieval	141	IR:141
METRIC	mean_average_precision	124	MAP:124
METHOD	collaborative_filtering	110	CF:110
METHOD	expectation_maximization	105	EM:105
METHOD	principal_component_analysis	104	PCA:104
DATASET	resource_description_framework	95	RDF:95
DATASET	neural_information_processing_systems	93	NIPS:93
METHOD	stochastic_gradient_descent	91	SGD:91
METRIC	minimum_description_length	78	MDL:78
PROBLEM	natural_language_processing	77	NLP:77
METRIC	normalized_mutual_information	76	NMI:76
METRIC	mean_reciprocal_rank	76	MRR:76
METHOD	document_object_model	71	DOM:71
METRIC	mean_absolute_error	69	MAE:69
METHOD	logistic_regression	67	LR:67
METRIC	normalized_discounted_cumulative_gain	66	NDCG:66
METHOD	latent_semantic_indexing	63	LSI:63
METHOD	maximum_a_posteriori	62	MAP:62
METRIC	mean_squared_error	62	MSE:62
METRIC	receiver_operating_characteristic	58	ROC:57 RoC:1
METHOD	dynamic_time_warping	58	DTW:58
METHOD	markov_chain_monte_carlo	57	MCMC:57
METHOD	naive_bayes	57	NB:57
ENTITY	transactions_on_information_systems	56	TOIS:56
METHOD	probabilistic_latent_semantic_analysis	52	PLSA:52
METHOD	directed_acyclic_graph	51	DAG:51
DATASET	open_directory_project	50	ODP:50
METHOD	non-negative_matrix_factorization	50	NMF:50
DATASET	national_science_foundation	50	NSF:50
METHOD	hidden_markov_models	47	HMMs:26 HMM:21
METHOD	maximum_likelihood_estimation	47	MLE:47
METRIC	root_mean_square_error	47	RMSE:47
METHOD	hidden_markov_model	45	HMM:45
METHOD	conditional_random_fields	45	CRFs:26 CRF:19
METHOD	matrix_factorization	44	MF:44
PROBLEM	information_extraction	42	IE:42
PROBLEM	named_entity_recognition	42	NER:42
METRIC	root_mean_squared_error	39	RMSE:39
METRIC	cumulative_distribution_function	39	CDF:39
METHOD	nonnegative_matrix_factorization	39	NMF:39
METHOD	vector_space_model	38	VSM:38
ENTITY	defense_advanced_research_projects_agency	37	DARP:
METRIC	information_gain	37	IG:37

Review: Seven Required Tasks

- T₁: Data cleaning and data integration (D.P.)
- T₂: Entity name recognition (D.P.)
- T₃: Entity typing (D.P./Cla.) + Eval.
- T₄: Collaboration discovery (F.P.M.)
- T₅: Problem-method association mining (A.R.M.)
- T₆: Problem/method/author classification (Cla.) + Eval.
- T₇: Paper clustering (Clu.) + Eval.
 - D.P. = Data preprocessing; Cla. = Classification
 - F.P.M. = Frequent pattern mining; Clu. = Clustering
 - A.R.M. = Association rule mining



*SciBot: Magic powers in
hand now!*

*But can you build a nice
GUI for me?*

*And can you write a
GuideBook for Dr. Zordon?*

Policy, Requirement, Grading

before Recommended Tasks

- Students are also required to write a **project report / paper** to describe their achievement including the following points **for each task:** (1) **Motivation and task definition,** (2) **Approach,** (3) **Results,** and (4) **Discussions.**
- Students are required to submit their code package + “readme” (.ZIP) and project report/paper (.PDF). There is no paper template requirement.
- The project due is **Nov 30, 2017**. There will be **NO extension**. **Significant updates** are welcome **before the final exam** – students can send updates to the instructor after the due by e-mail but they **have to submit one version before the due**.
- **HWs will be designed to follow-up your progress** ☺
 - **Milestones**

Milestones!!!

- T₁: Data cleaning and data integration (D.P.)
- T₂: Entity name recognition (D.P.) – **HW3 (Oct. 3)**
- T₃: Entity typing (D.P./Cla.) + Eval.
- T₄: Collaboration discovery (F.P.M.)
- T₅: Problem-method association mining (A.R.M.) – **HW4 (Nov. 9)**
- T₆: Problem/method/author classification (Cla.) + Eval.
- T₇: Paper clustering (Clu.) + Eval.
 - D.P. = Data preprocessing; Cla. = Classification
 - F.P.M. = Frequent pattern mining; Clu. = Clustering
 - A.R.M. = Association rule mining – **Project (Nov. 30)**

~~Individual Project!~~



[updated Sept. 28, 2017]

Team Project

- A team will have at most 2 members.
- Students will give their partner's name (or N/A – if by their own) in HW3.
- Members in the same team will have the same score.
- In last two lectures we tend to grade individual projects/undergraduates better than team projects/graduates if they generate the same results.

Grading

- Students will **volunteer to present** their SciBot (tech and results) **in the last two lectures**. Classmates and the instructor will grade them based on the presentation. For the students **who do not present**, the **instructor** will grade their projects after all the lectures end. Note that we will have **comparative grading** – *finishing* all the **required tasks** cannot make sure that you have all points – but *finishing in high quality yes* – or you can have all points if you do *required tasks in low quality* but *some more* tasks.
- Graders should have **higher expectations on graduates** than undergraduates – not only on the project results (more tasks, better performances) but also **on writing (a workshop-quality paper of strong reasoning)**. Undergraduates will be applied with a **uniform grading policy** no matter what majors they have.

What Are Encouraged?

- Students are **encouraged** to implement algorithms such as Apriori, FP-Growth, and K- Means Clustering by themselves instead of calling Python packages.
- Students are also **encouraged** to use Python packages (e.g., numpy and scipy) when they use advanced techniques (e.g., SVMs, Neural Networks, word2vec) to address challenging tasks.
- Students are **encouraged** to compare different methods on the same task and discuss their advantages and disadvantages. Reasoning is always welcome in the paper.
- Students are **encouraged** to share any annotation data (e.g., labels, hand-crafted rules) but not any segment of codes.
- Students are **encouraged** to make a GUI for the SciBot. They are also encouraged to give a better name to their bots than “SciBot”.

Your SciBot



FAQ

- Q1:
- A1:
- Q2:
- A2:

Next: Recommended tasks

WebUI for Annotation

- From UIUC CS412 Summer'17

dd15-p1005
dd15-p1015
dd15-p1035
dd15-p1045
dd15-p1055
dd15-p1065
dd15-p1075
dd15-p1085
dd15-p109
dd15-p1095
dd15-p1105
dd15-p1115
dd15-p1125
dd15-p1135
dd15-p1145
dd15-p1155
dd15-p1165
dd15-p1175
dd15-p1185
dd15-p119
dd15-p1195
dd15-p1205
dd15-p1215
dd15-p1225
dd15-p1235
dd15-p1245
dd15-p1255
dd15-p1265
dd15-p1275
dd15-p1285
dd15-p129
dd15-p1295
dd15-p1305
dd15-p1315
dd15-p1325
dd15-p1335
dd15-p1345
dd15-p1355
dd15-p1365
dd15-p1375
dd15-p1385
dd15-p139
dd15-p1395
dd15-p1405
dd15-p1415
dd15-p1425
dd15-p1435
dd15-p1445
dd15-p1455
dd15-p1465
dd15-p1475
dd15-p1485
dd15-p149
dd15-p1495
dd15-p150
dd15-p1513
dd15-p1523
dd15-p1533
dd15-p1543
dd15-p1553
dd15-p1563
dd15-p1573
dd15-p1583
dd15-p159
dd15-p1593
dd15-p1603
dd15-p1641
dd15-p1651
dd15-p1661
dd15-p1671

the Social Sciences & University of Koblenz-Landau Martin Becker University of Wurzburg wuerzburg.de Philipp Singer GESIS - Leibniz Institute for the Social Sciences & University of Koblenz-Landau Denis Helic Graz University of Technology Andreas Hotho University of Wurzburg and L3S Hannover wuerzburg.de Markus Strohmaier GESIS - Leibniz Institute for the Social Sciences & University of Koblenz-Landau ABSTRACT We present a new method for detecting **interpretable subgroups** with exceptional **transition behavior** in **sequential data**. Identifying such patterns has many potential applications , e . g . , for studying human **mobility** or analyzing the behavior of internet users . To tackle this **task** , we employ **exceptional model mining** , which is a general approach for identifying **interpretable data subsets** that exhibit unusual interactions between a set of target attributes with respect to a certain model class . Although **exceptional model mining** provides a well-suited framework for our problem , previously investigated model classes cannot capture **transition behavior** . To that end , we introduce first-order **Markov chains** as a novel model class for **exceptional model mining** and present a new **interestingness** measure that quantifies the exceptionality of transition subgroups . The measure compares the **distance** between the Markov transition matrix of a subgroup and the respective matrix of the entire data with the **distance** of random dataset samples . In addition , our method can be adapted to find subgroups that match or contradict given **transition hypotheses** . We demonstrate that our method is consistently able to recover subgroups with **exceptional transition models** from **sequential data** and illustrate its potential in two application examples . Our work is relevant for researchers and practitioners interested in detecting **exceptional transition behavior** in **sequential data** . Keywords : **Subgroup Discovery** ; **Exceptional Model Mining** ; **Markov chains** ; **Transitions** ; **Sequential data** . 1. INTRODUCTION Exceptional Model Mining , a generalization of the classic **subgroup discovery task** , is a framework that identifies patterns which contain unusual interactions between multiple target attributes . In order to obtain operationalizable insights , it emphasizes the detection of easy-to-understand subgroups , i . e . , it aims to find **exceptional subgroups** with descriptions that are directly interpretable by domain experts . In general , exceptional model mining-Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page . Copyrights for components of this work owned by others than the author must be honored . Abstracting with credit is permitted . To copy otherwise , or republish , to post on servers or to redistribute to lists , requires prior specific permission and /or a fee . Request permissions from . KDD '16 , August 13 - 17 , 2016 , San Francisco , CA , USA 2016 Copyright held by the owner / author . Publication rights licensed to ACM . ISBN 978-1-4503-4232-2/16 / 08 . \$ 15.00 DOI : [10.1145/2959100.2959200](https://doi.org/10.1145/2959100.2959200) ing operates as follows : A target model of a given model class is computed once over the **entire dataset** , resulting in a **set of model parameters** . The same parameters are also calculated for each **subgroup** in a large (often implicitly specified) candidate set , using only the instances covered by the respective subgroup . A subgroup is considered as exceptional or interesting if its parameter values differ significantly from the ones of the overall dataset . While **exceptional model mining** has been implemented for a variety of model classes including **classification** , **regression** , **Bayesian network** and **rank correlation** models , it has not yet been applied using models for **sequential data** . In this paper , we aim to apply **exceptional model mining** to discover **interpretable subgroups** with exceptional **transition behavior** . This enables a new analysis method for a variety of applications . As one example , assume a **human mobility** dataset featuring user transitions between locations . The overall transition model could for example show that people either move within their direct neighborhood or along main roads . Detecting subgroups with exceptional **transition behavior** goes beyond this simple analysis : It allows to automatically identify subgroups of people (such as " male tourists from France ") or subsegments of time (such as " 10 to 11 p . m . ") that exhibit unusual movement characteristics , e . g . , tourists moving between points-of-interest or people walking along well-lit streets at night . Other application examples could include subgroups of web-users with unusual **navigational** behavior or subgroups of companies with unusual development over time , cf . The main contribution of this paper is a new method that enables mining subgroups with exceptional **transition behavior** by introducing first-order **Markov chains** as a novel model class for **exceptional model mining** . **Markov chains** have been utilized for studying **sequential data** about , e . g . , **human navigation** and **mobility** , **meteorology** , or **economics** . To apply **exceptional model mining** with this model , we derive an **interestingness** measure that quantifies the exceptionality of a subgroup's transition model . It measures how much the **distance** between the Markov transitions matrix of a subgroup and the respective matrix of the entire data deviates from the **distance** of random dataset samples . This measure can be integrated into any known **search algorithm** . We also show how an adaptation of our approach allows to find subgroups specifically matching (or contradicting) given hypotheses about **transition behavior** (cf .) . This enables the use of **exceptional model mining** for a new type of studies , i . e . , the detailed analysis of such hypotheses . We demonstrate the potential of the proposed approach with synthetic as well as **real-world** data . 965 The remainder of this work is organized as following : We summarize our background in Section 2 . Then , the main approach for mining subgroups with exceptional **transition behavior** is introduced in Section 3 . Section 4 presents experiments and results . Finally , we discuss related work in Section 5 , before we conclude in Section 6 . 2 . BACKGROUND Our solution extends **Exceptional Model Mining** with first-order **Markov Chain** Models . In the following , we give a brief overview of both techniques . 2.1 **Exceptional Model Mining** We formally define a dataset D as a multiset of data instances | I | described by a set of attributes A consisting of describing attributes AD A and model attributes AM A . A subgroup consists of a subgroup description p : D - that is given by a **Boolean function** , and a subgroup cover c , i . e . , the set of instances described by p , i . e . , c = . In principle , our approach works with any **pattern description language** to describe

Ignore:
10 australia
11 ca
12 kdd
13 acm
14 august
15 copyright
16 keywords
17 san francisco
18 introduction
19 isbn
20 experiments

Method:
20 barabasi-albert
21 progressive sampling
22 heuristic
23 monte-carlo, montecarlo
24 non-linear optimization
25 map-reduce, mapreduce
26 poisson process
27 em algorithm
28 cmpp
29 random sampling
30 linear system, linear model

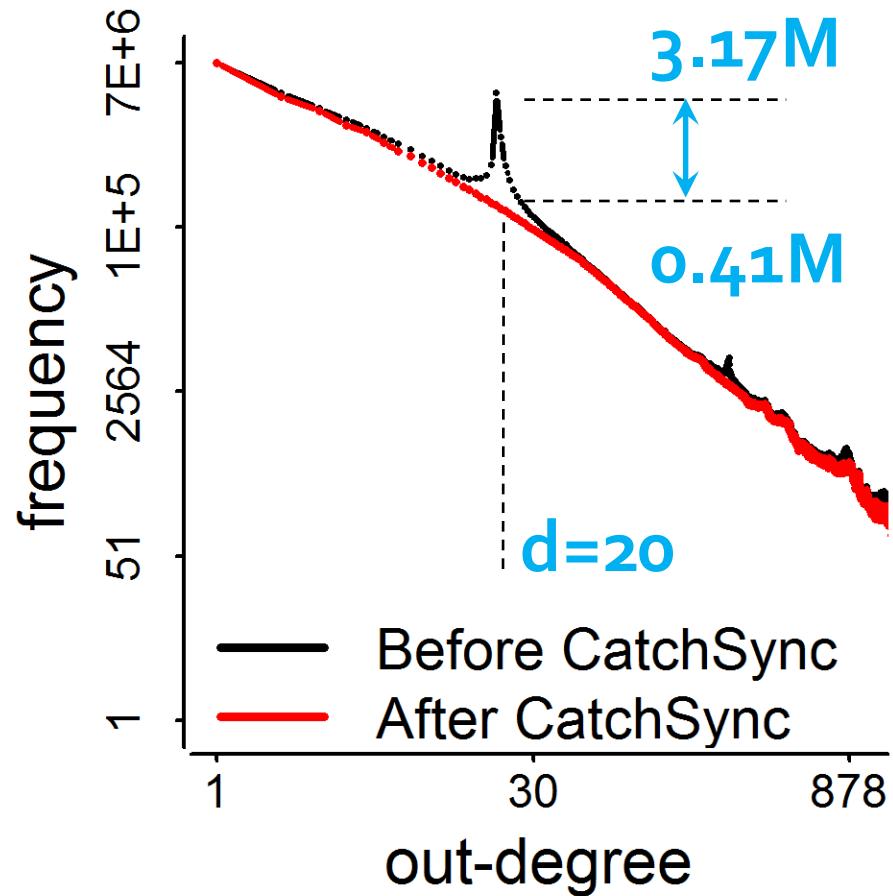
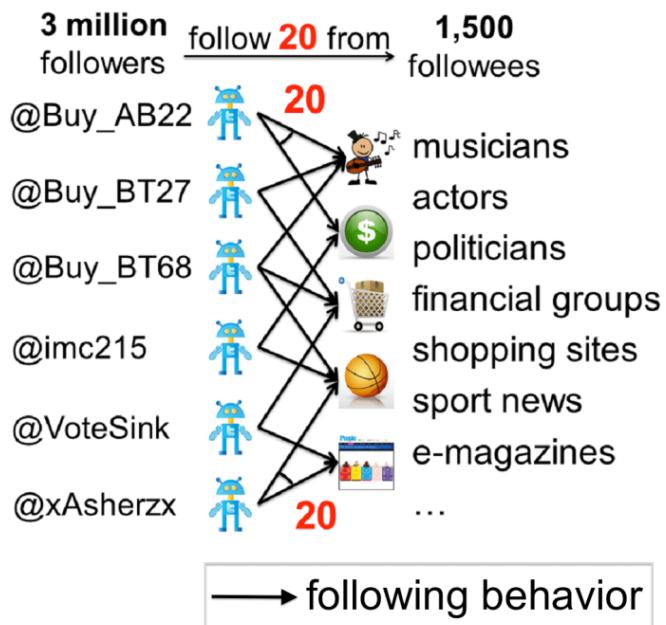
Problem:
31 p...
47 large-scale network
48 minimum-cut
49 motif discovery
50 rule discovery
51 future event
52 co-authorship network
53 modularity
54 paper recommendation
55 eco-centric circle
56 biological network

Topic:
57 ...
8 standard error
9 z-score
10 normality
11 p-value
12 critical value
13 total variation
14 statistical significance
15 quality measure
16 accuracy
17 euclidean norm, euclidean distan
18 ...

Source:
1 wikileaks
2 facebook
3 synthetic data
4 real-world
5 flickr
6 lastfm
7 bms-pos
8 fimi
9 sequential data
10 linkedin

Task 1+: Degree Distributions

- Regression
 - Linear, Power-law
 - Poisson, Gaussian



CatchSync (KDD'14): <http://www.meng-jiang.com/pubs/catchsync-kdd14/catchsync-kdd14-paper.pdf>
KONECT network collection: <http://konect.uni-koblenz.de/networks/>

Task 3+: Entity Cluster Typing

- First entity clustering and then typing
 - High confidence on one entity's type will help to type other entities in the same cluster
 - What are the features for entity clustering?

METHOD	latent_dirichlet_allocation	247	LDA:247
METHOD	support_vector_machine	218	SVM:218
METHOD	support_vector_machines	214	SVM:125 SVMs:89
METHOD	singular_value_decomposition	150	SVD:150
DATASET	world_wide_web	145	WWW:145
PROBLEM	information_retrieval	141	IR:141
METRIC	mean_average_precision	124	MAP:124
METHOD	collaborative_filtering	110	CF:110
METHOD	expectation_maximization	105	EM:105
METHOD	principal_component_analysis	104	PCA:104
DATASET	resource_description_framework	95	RDF:95
DATASET	neural_information_processing_systems	93	NIPS:93
METHOD	stochastic_gradient_descent	91	SGD:91
METRIC	minimum_description_length	78	MDL:78
PROBLEM	natural_language_processing	77	NLP:77
METRIC	minimum_description_length	76	MDL:76

Task 4+: Advisor-Advisee Discovery

- Beyond collaborations
 - First-last-author rule
 - Measuring proximity with *Kulc* measure
 - Correlation between them
 - Ground-truth and evaluation

Task 8: SciCube

- Given enriched structured data, can we construct a **data cube** and compute **iceberg cubes** for query-based applications?
- E.g., expert recommendation: Given a problem, list authors, papers and other information that help related research.
- **Techs:** Data cube, iceberg cube, closed cells, etc.
- **Hints:**
 - Each paper is considered as a transaction. The cell maintains a set of papers. A paper may be in multiple cells. We count the size of paper set for the cube computation.
 - For a list of entities, the attribute types are dimensions (e.g., problem, method, dataset, author, conference); the attribute values are the dimension values (e.g., “naïve bayes”, “decision tree”).
 - More functionalities of the data cube, and efficiency analysis are welcome.

Task 9: Pattern-based Entity Recognition and Typing

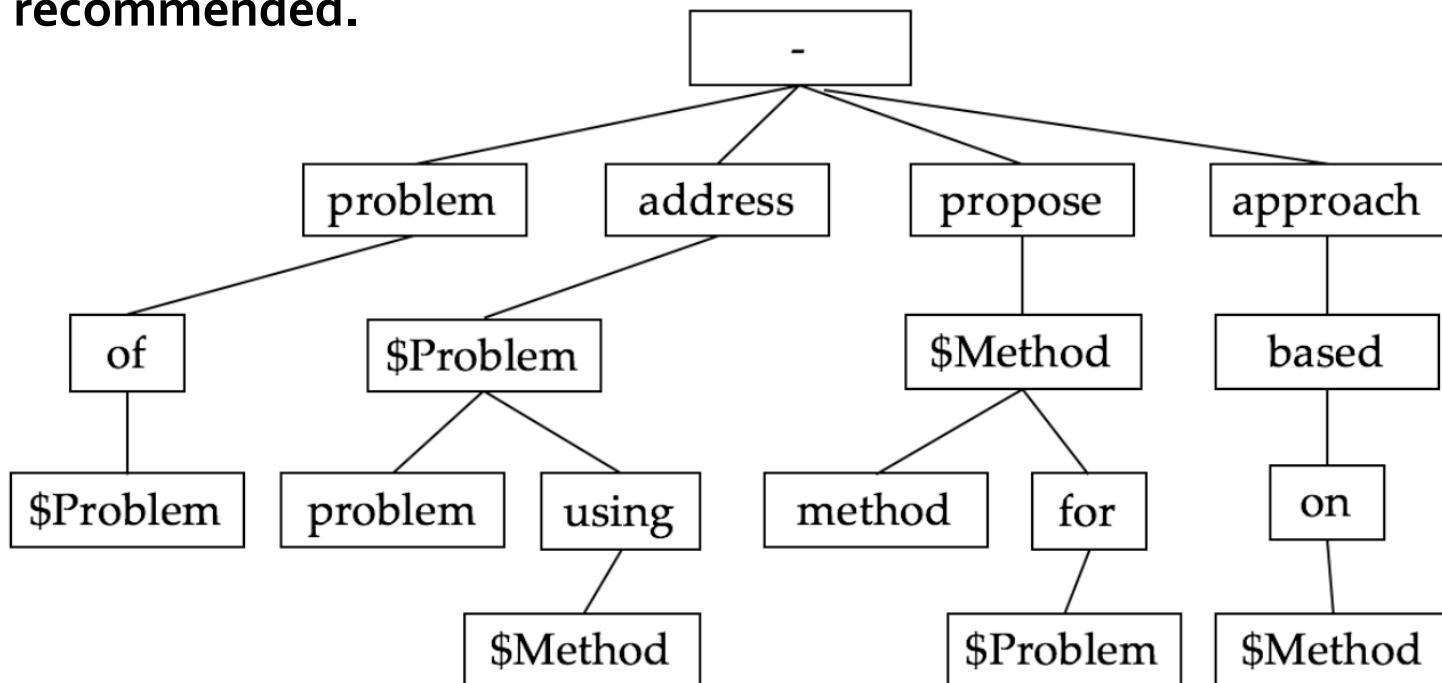
- Given entity names, can we find **frequent patterns** around the entities? We replace concrete entity names as "\$Entity." You can find more entities via **pattern matching**.
- Given seed typed entities (methods, problems, etc.), can we find concrete frequent patterns around the typed entities? We replace concrete method/problem entities as "\$Method"/"\$Problem". Those patterns indicate that you may be able to find more entities of the specific types.
- **Techs:** Constraint-based frequent pattern mining.
- **Hints:**
 - An **iterative process** that first generate and evaluate the support of patterns such as "problem of \$Problem", "address \$Problem problem", "propose \$Method method", "approach based on \$Method", "propose \$Method for \$Problem", "address \$Problem using \$Method" and then recognize more entities and their types by matching the patterns in the text and repeat until convergence.

Hint: Pattern Matching

- **How to match the text with patterns efficiently?**
- Suppose we have 6 patterns: “problem of \$Problem”, “address \$Problem problem”, “propose \$Method method”, “approach based on \$Method”, “propose \$Method for \$Problem”, “address \$Problem using \$Method”.
- **Match string patterns???**

Hint: Pattern Matching

- **How to match the text with patterns efficiently?**
- Suppose we have 6 patterns: “problem of \$Problem”, “address \$Problem problem”, “propose \$Method method”, “approach based on \$Method”, “propose \$Method for \$Problem”, “address \$Problem using \$Method”.
- **The data structure of Trie Tree (<https://en.wikipedia.org/wiki/Trie>) is recommended.**

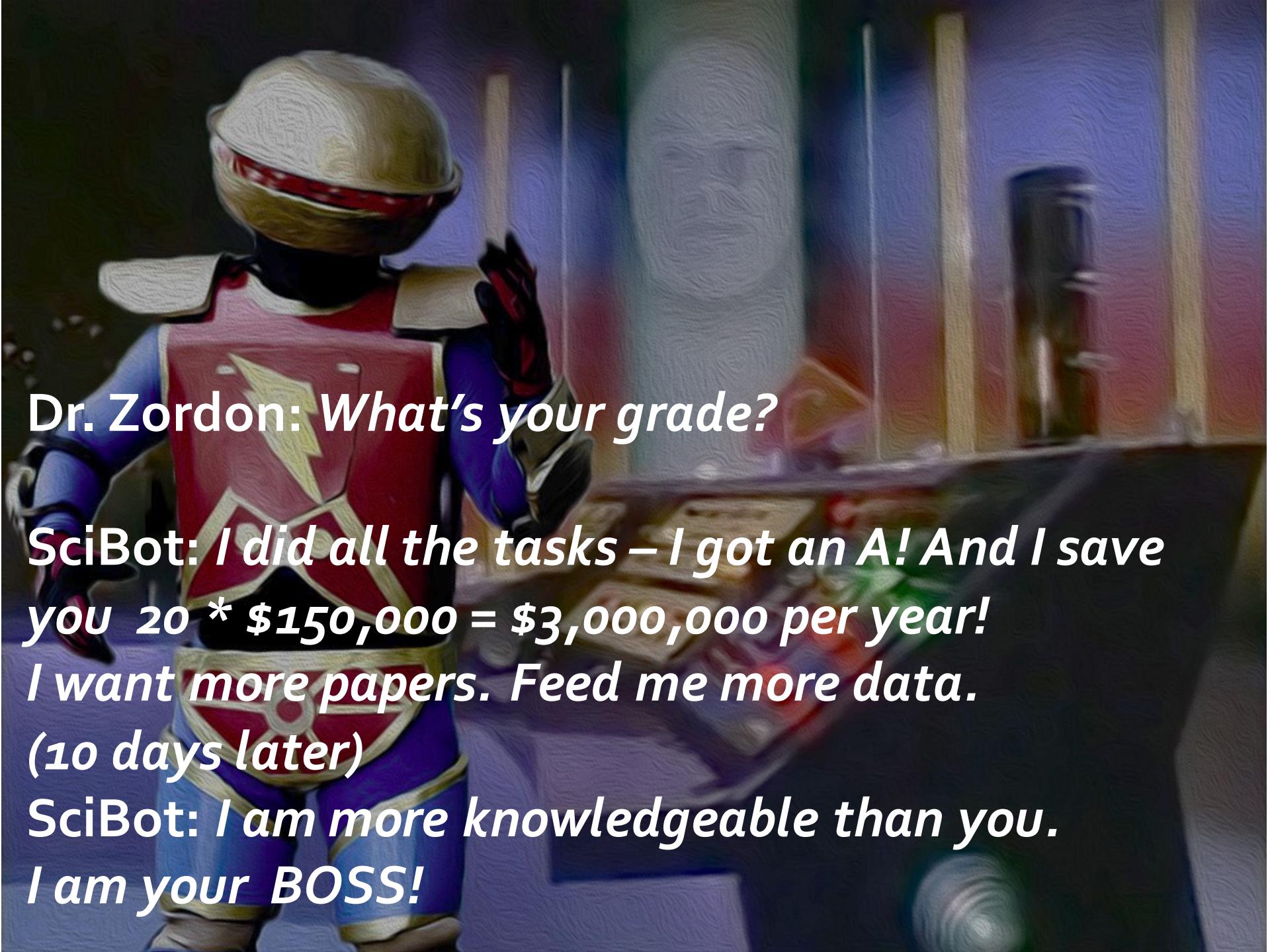


Other Tasks

- Task 10: Problem/method/author clustering
 - Given a set of problems/methods/authors, cluster them into K groups.
Evaluate the clustering results in a proper way.
 - **Techs:** K-partitioning clustering methods (K-Means).
- Task 11: Attribute discovery
 - Suppose we use rules to type digit number as \$Digit. Can we find the size of datasets used in the papers? Can we find the performance of methods?
 - **Techs:** Constraint-based frequent pattern mining.
- Task 12: Ensemble learning
 - Suppose we have multiple models/methods for a specific task (actually you do have if you've finished the required tasks). Can we use ensemble methods to further improve the performance?
 - **Techs:** Ensemble methods (bagging, Adaboost, etc.).

Other Tasks (cont.)

- Task 13: Practice with advanced classification and clustering methods
 - Can you solve the above tasks with advanced classification models (e.g., SVMs, Neural Networks) and clustering methods (e.g., spectral clustering)?
- Task 14: Other interesting tasks related to other data entries/attribute:
 - Affiliation ranking on a specific method/problem
- Task 15: Data visualization is encouraged.



Dr. Zordon: What's your grade?

*SciBot: I did all the tasks – I got an A! And I save
you $20 * \$150,000 = \$3,000,000$ per year!
I want more papers. Feed me more data.
(10 days later)*

*SciBot: I am more knowledgeable than you.
I am your BOSS!*

Enjoy the Power-Rangers Project ☺

