Data science in IT security

Eszter Windhager-Pokol

Agenda

- About me
- Data science in IT security
- Recommender system for anomaly detection
- R coding

About me

- MSc Applied Mathematics
- PhD Studies (absolutorium – no degree)
- 7 years Data Mining Analyst (consultant) – Clementine Consulting, I-insight
- 3 years Senior Data Scientist (IT security product development) – Balabit (One Identity)
- 1 year Lead Data Scientist – Vodafone
- Since 2018 Head of Data Science (consultant) – Starschema
- + 2 years organizer of R-Ladies Budapest Meetup Group

Sector comparison

Multinational Company

Consultancy

Product development (Startup)

IT Security

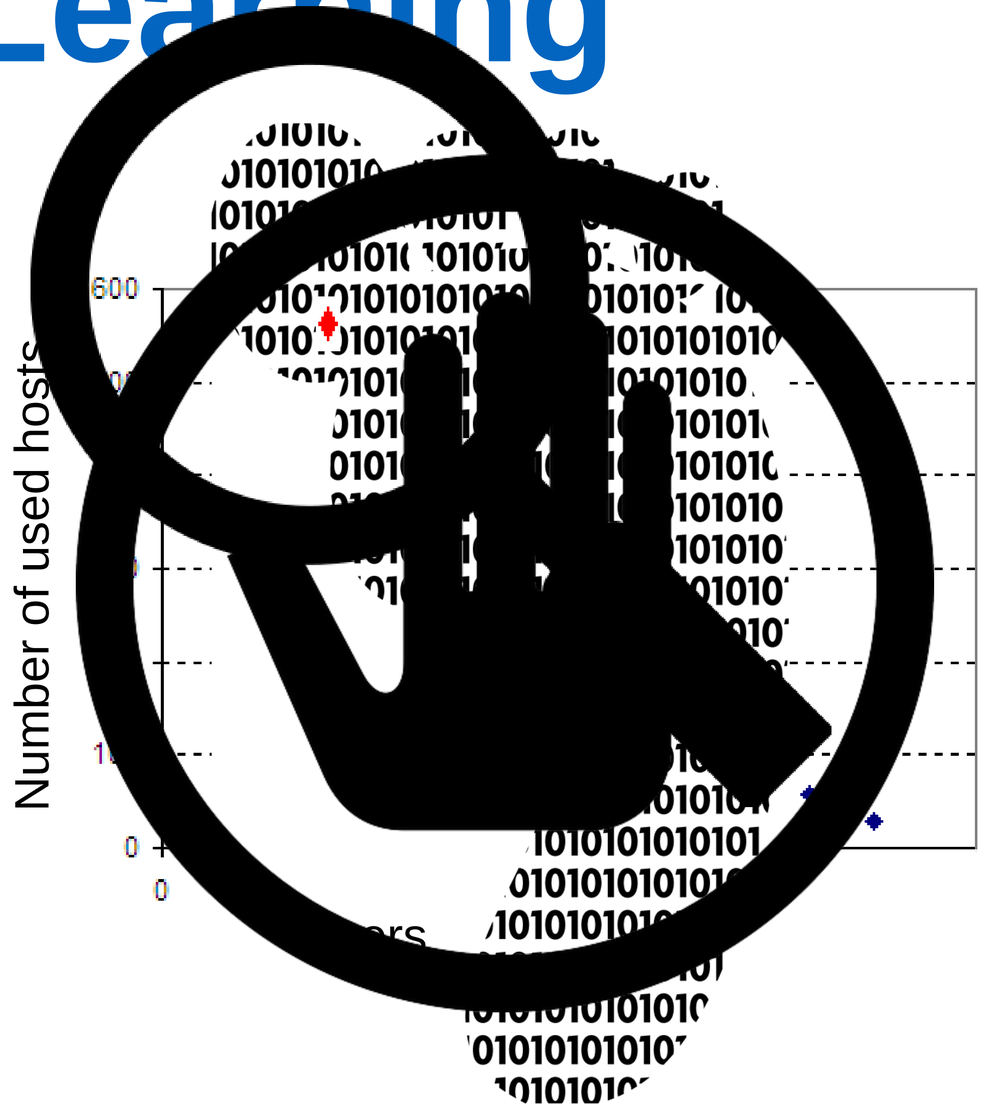
SIEM - Security Information and Event Management

Fixed rules

- Setting up rules is challanging
  - Limit the business
  - Prevent data breach
- Maintenance is almost impossible
  - The environment and the roles are changing rapidly
- **Only for known attack types**

Machine Learning

- Gather users' digital footprints
- Build a baseline for „normal" behavior
- Identify unusual activities in real-time
- Act immediately to prevent data breaches

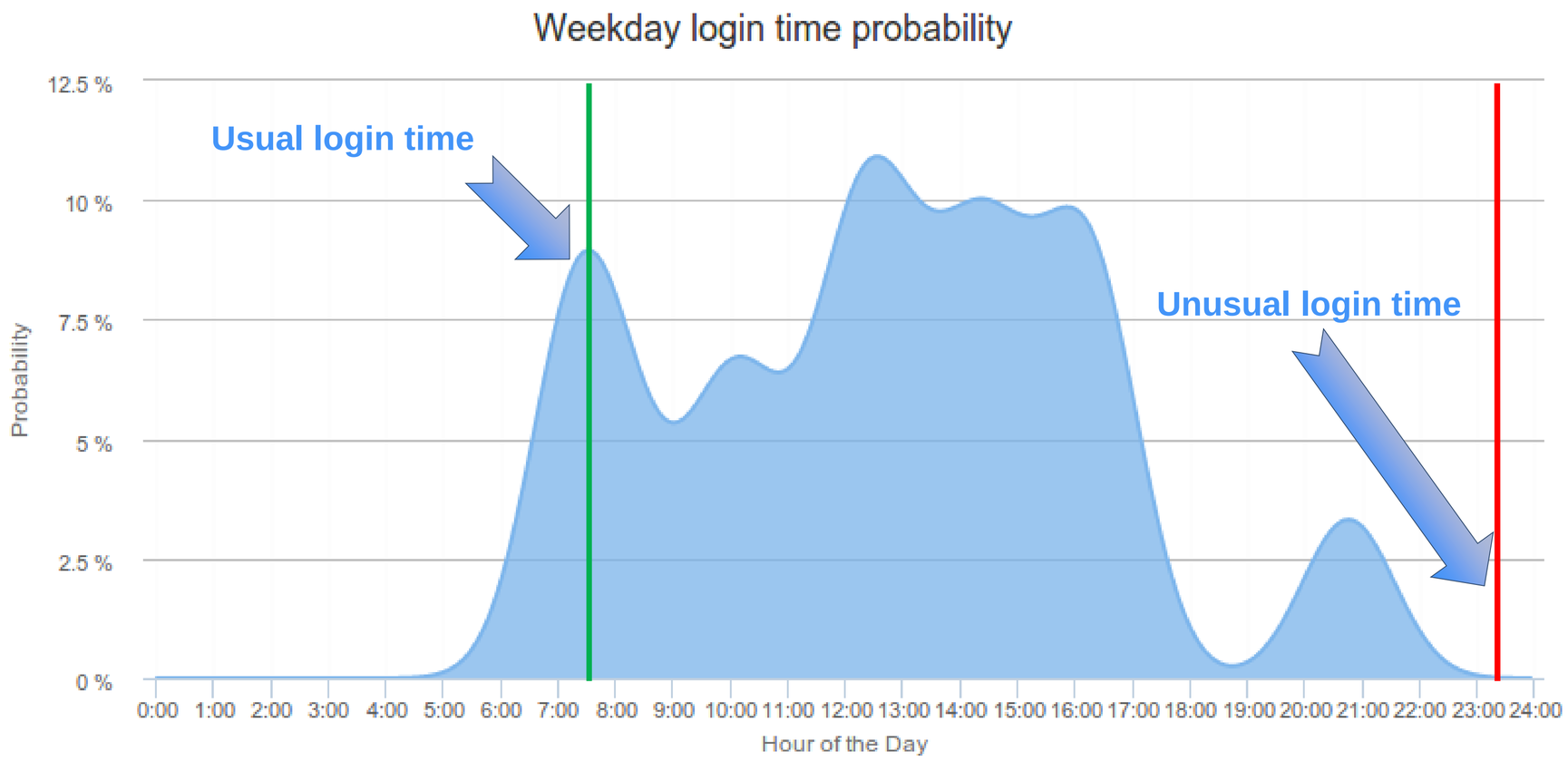


Time for decision making

User Behavior Analysis

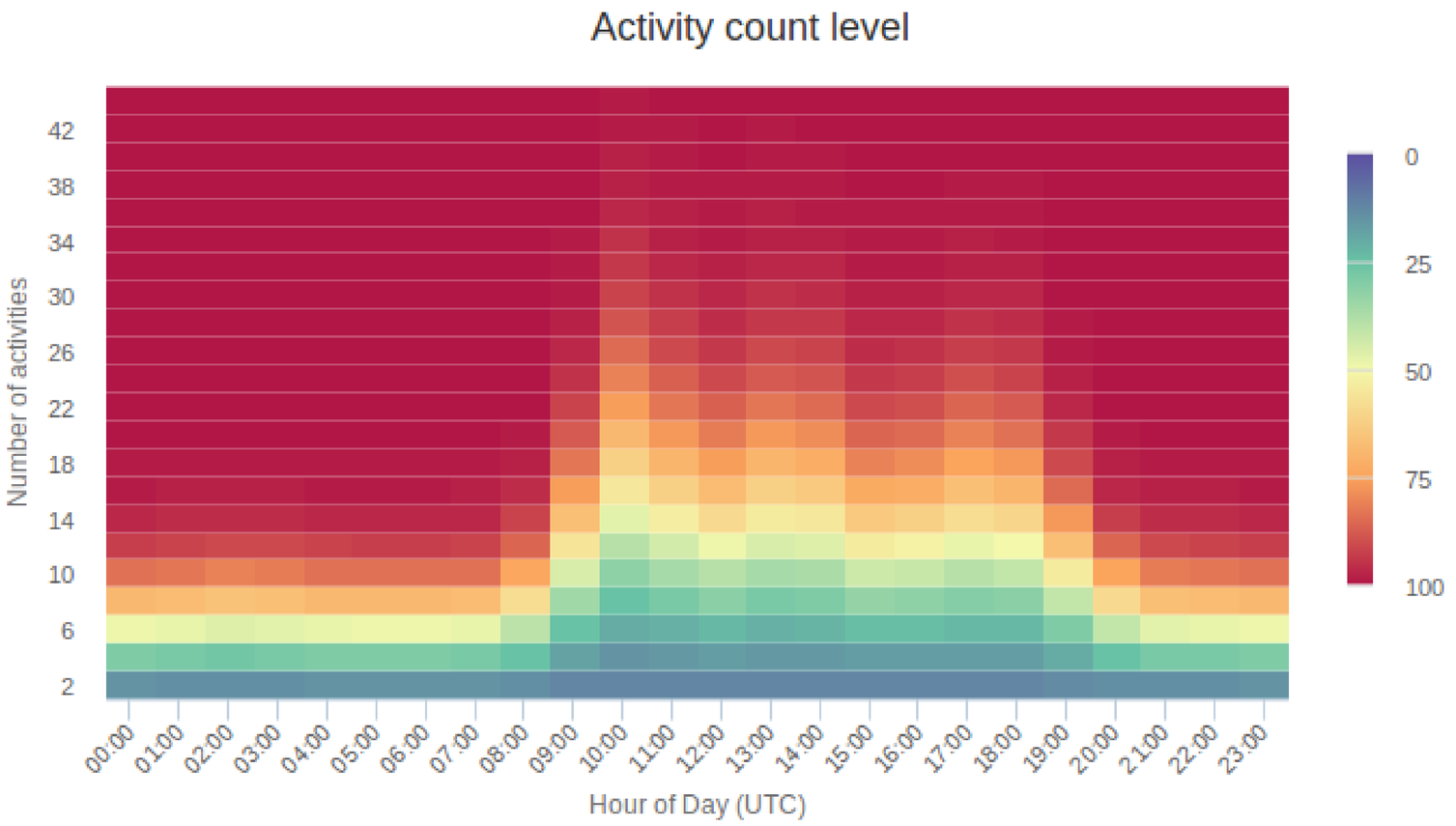
One dimension	Multidimension	Biometric identification
•**Login Times** •**Activity counts** •**Commands** •**Hosts** •**Protocols** •**Window titles**	•**Frequent ItemSet** •**Principal Components Classifier**	•**Keystroke dynamics** •**Mouse movement** •Pointing device •User identification

Logintime distribution



Activity count

Typical activity counts over time of the day



Frequent ItemSet

Find frequent co-occurencies

start_time_is_workday	**true**

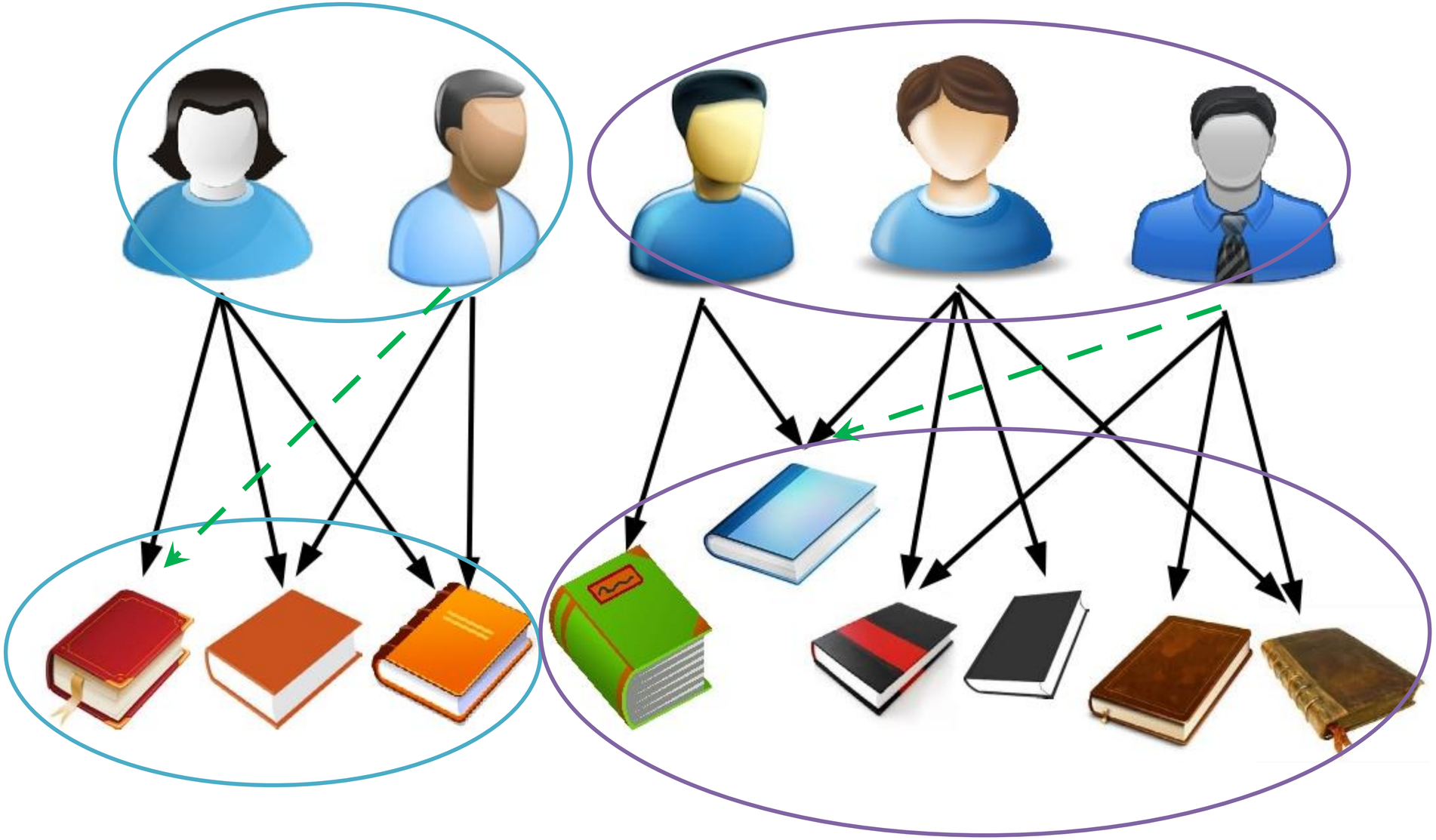
start_time_is_workday	**true**	auth_method	**0**	protocol	**rdp**	src_ip	**209.156.29.226**	port	**3389**

start_time_hour	**8-13**	start_time_is_workday	**true**

channels->verdict.ACCEPT	**true**	start_time_is_workday	**true**

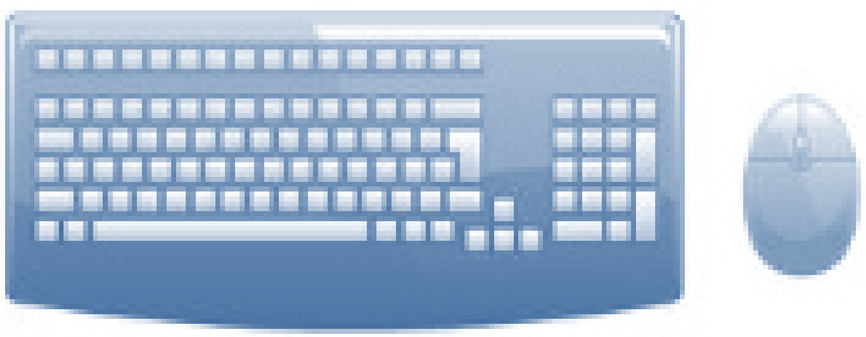
auth_method	**0**	start_time_is_workday	**true**	start_time_hour	**8-13**	src_ip	**209.156.29.226**	port	**3389**	protocol	**rdp**

Recommender system



Multifactor authentication

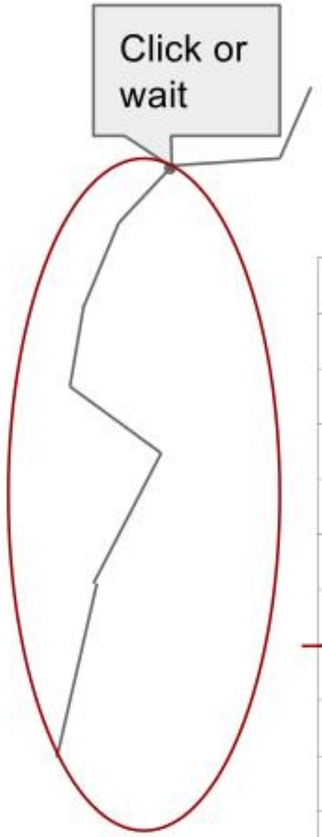
Biometric authentication



- Pointing device recognition
  - Mouse
  - Touchpad
- User identification based on mouse movement
- User identification based on keystroke dynamics

Mouse movement analytics

Separate gestures



ID	X	Y	Timestamp
1234	234	102	12569537329
1235	237	99	12569537335
1236	242	87	12569537342
1237	267	64	12569537354
1238	253	77	12569537360
1239	244	83	12569537370
1240	256	95	12569538123
1241	287	98	12569538131
1242	378	110	12569538139
1243	400	134	12569538142

Descriptive statistics

ID	X	Y	Timestamp	speed	curva-ture	...
1234	234	102	12569537329			
1235	237	99	12569537335			
1236	242	87	12569537342			
1237	267	64	12569537354			
1238	253	77	12569537360			
1239	244	83	12569537370			
1240	256	95	12569538123			
1241	287	98	12569538131			
1242	378	110	12569538139			
1243	400	134	12569538142			

Aggregate to gestures

Movement_ID	speed_min	speed_max	straightness_mean	...
001				
002				
003				
004				
005				
006				
...				

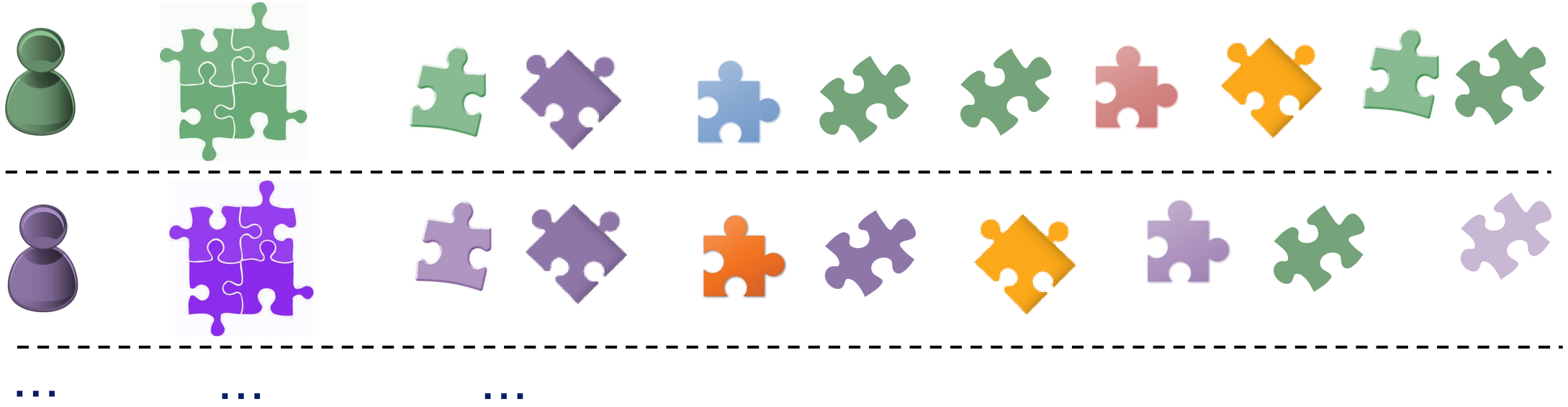
https://github.com/balabit/Mouse-Dynamics-Challenge

Model evaluation

1. Build baseline

2. Future activities

+ 3. Fake activities

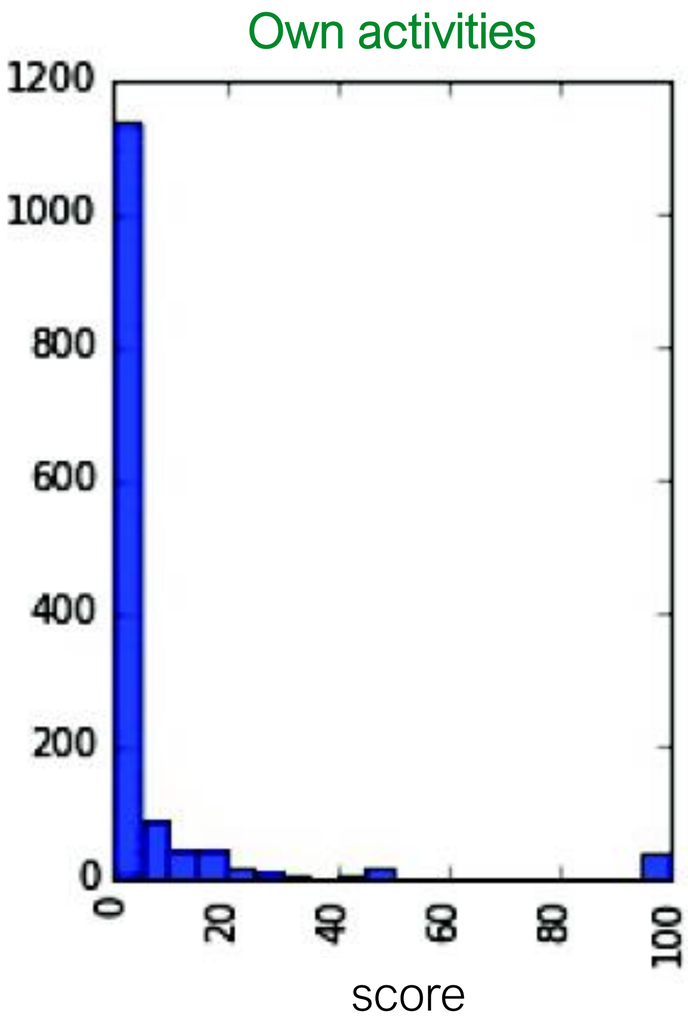


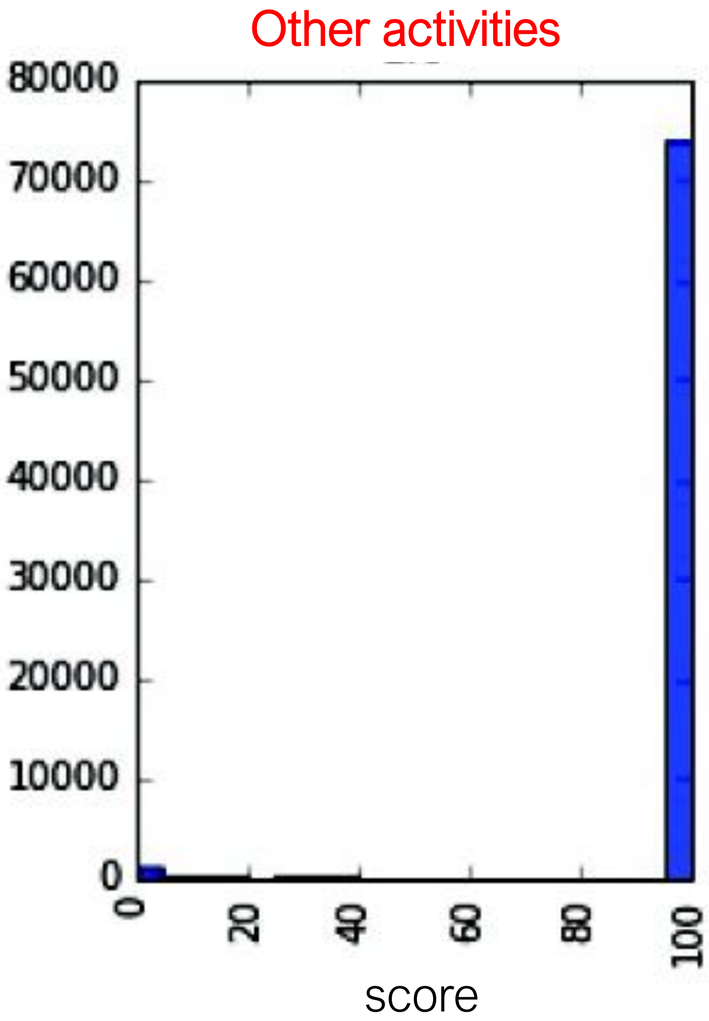
4. Score activities

5. Calculate AUC

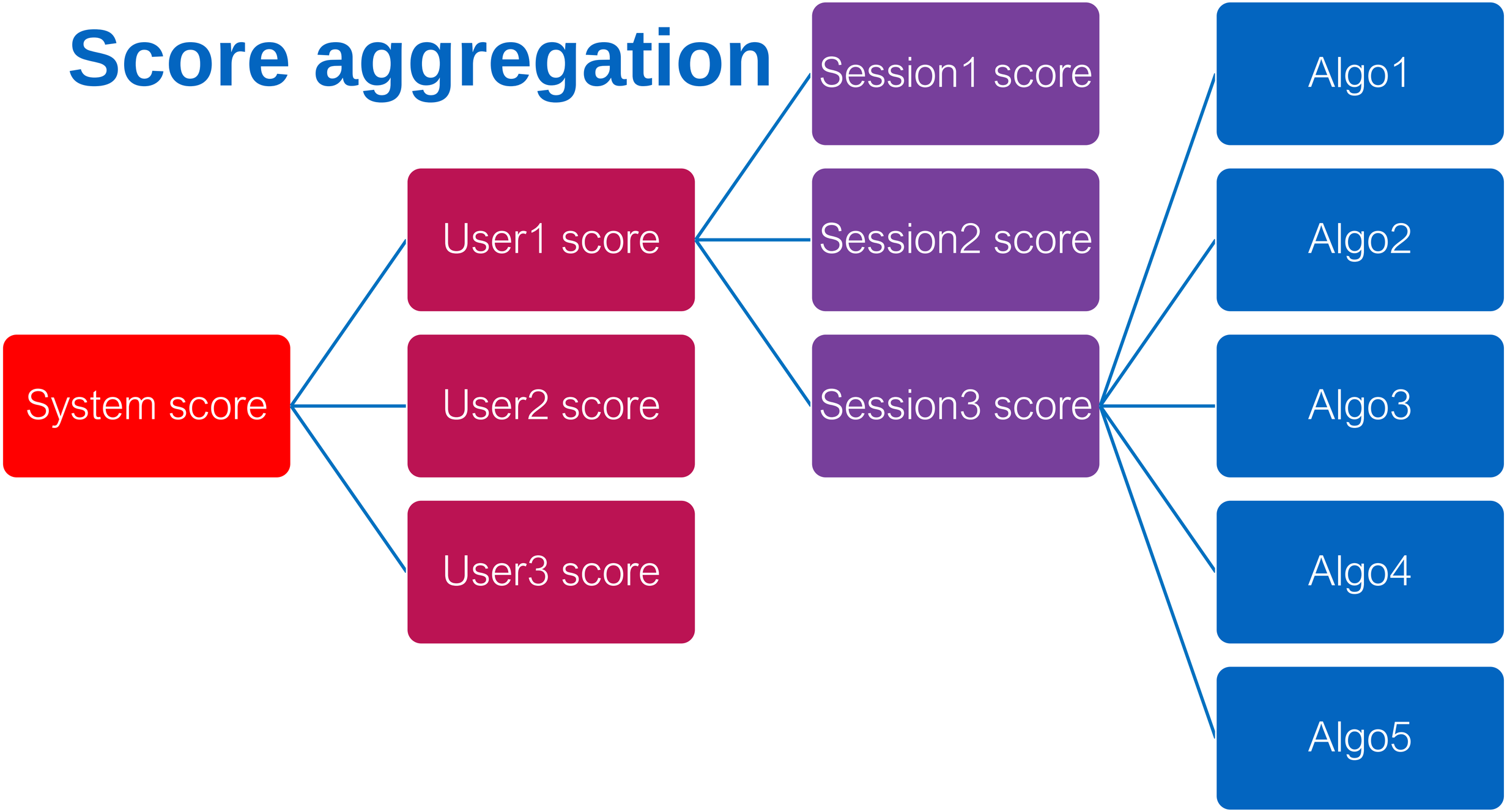
(Area Under the ROC curve)

Evaluation


Own activities
1200
1000
800
600
400
200
0
0
20
40
60
80
100
score


Other activities
80000
70000
60000
50000
40000
30000
20000
10000
0
0
20
40
60
80
100
score

Score aggregation

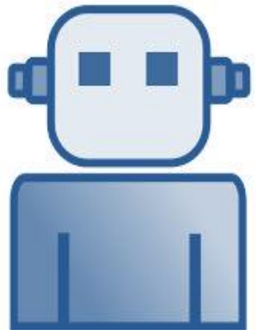


Further analytics

User risk

Peer group

Script detection



User risk

z-score

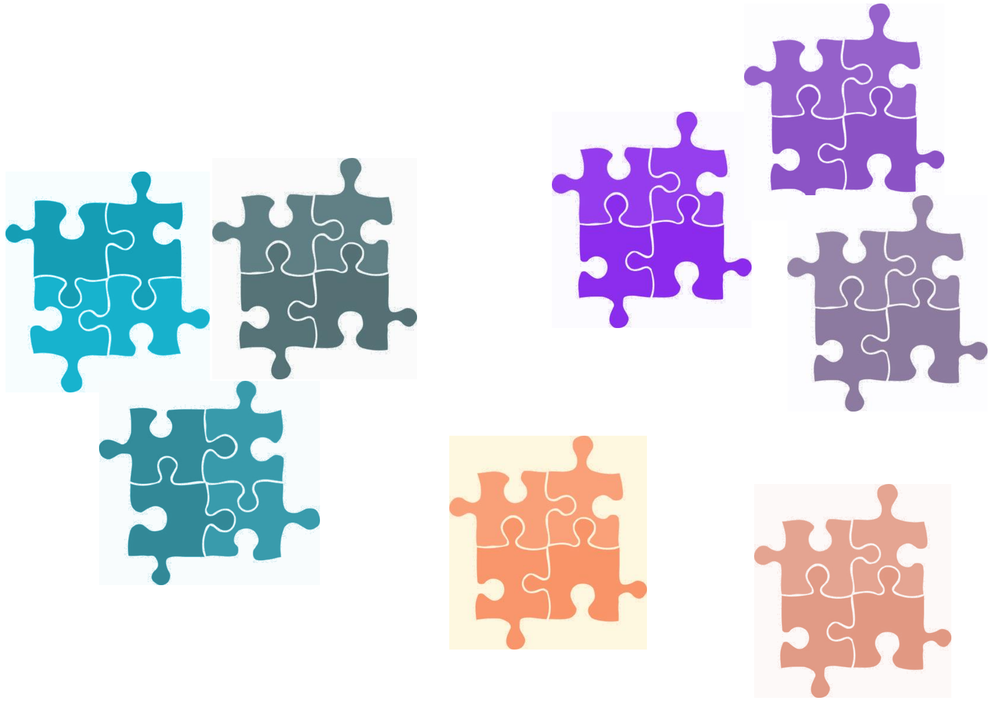


			…	
	X	X		X
	√	X		X
	√	√		X
	√	√		X
…	…	…		…
	√	√		√

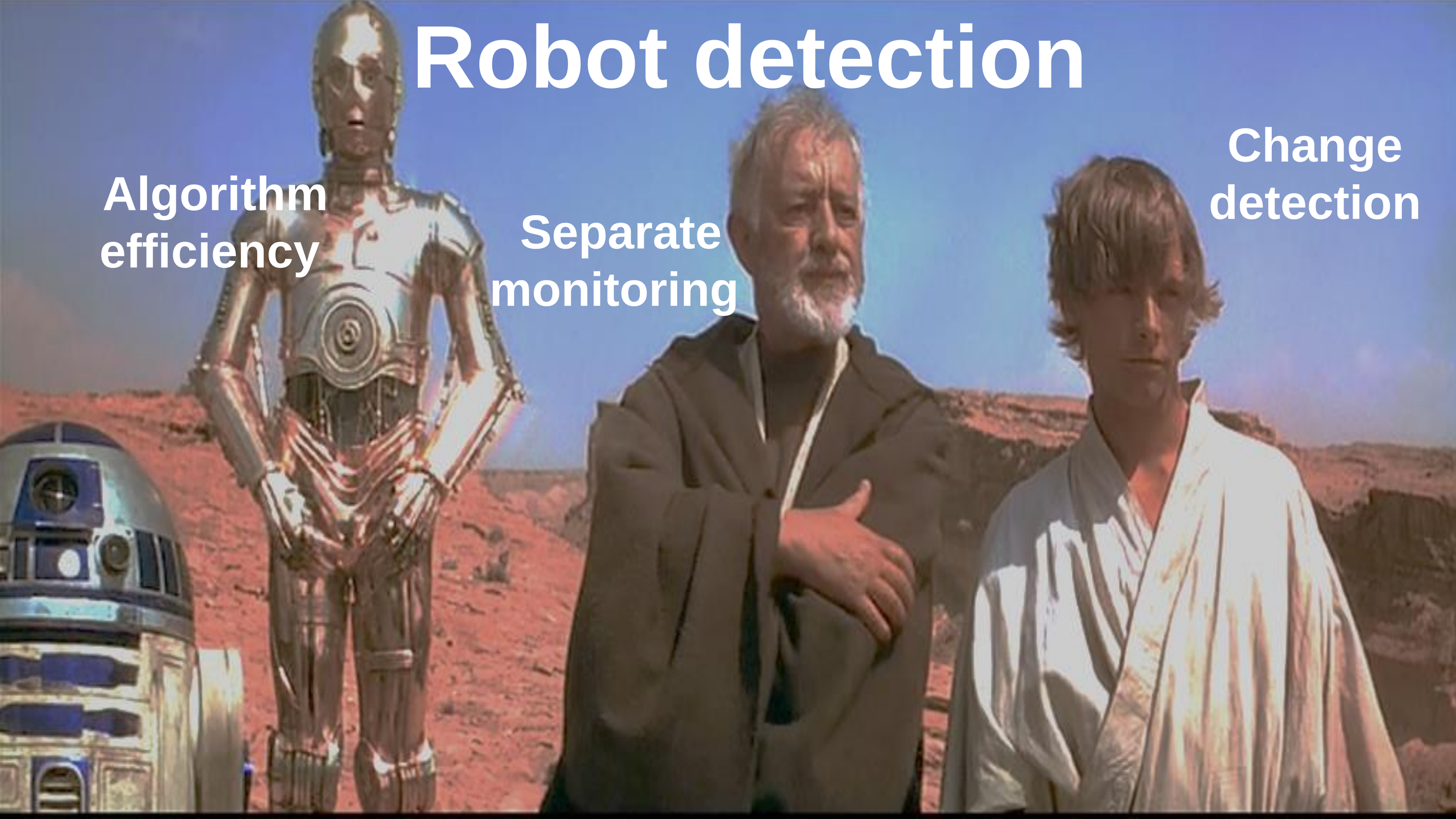
Peer group

Compare activities to user patterns

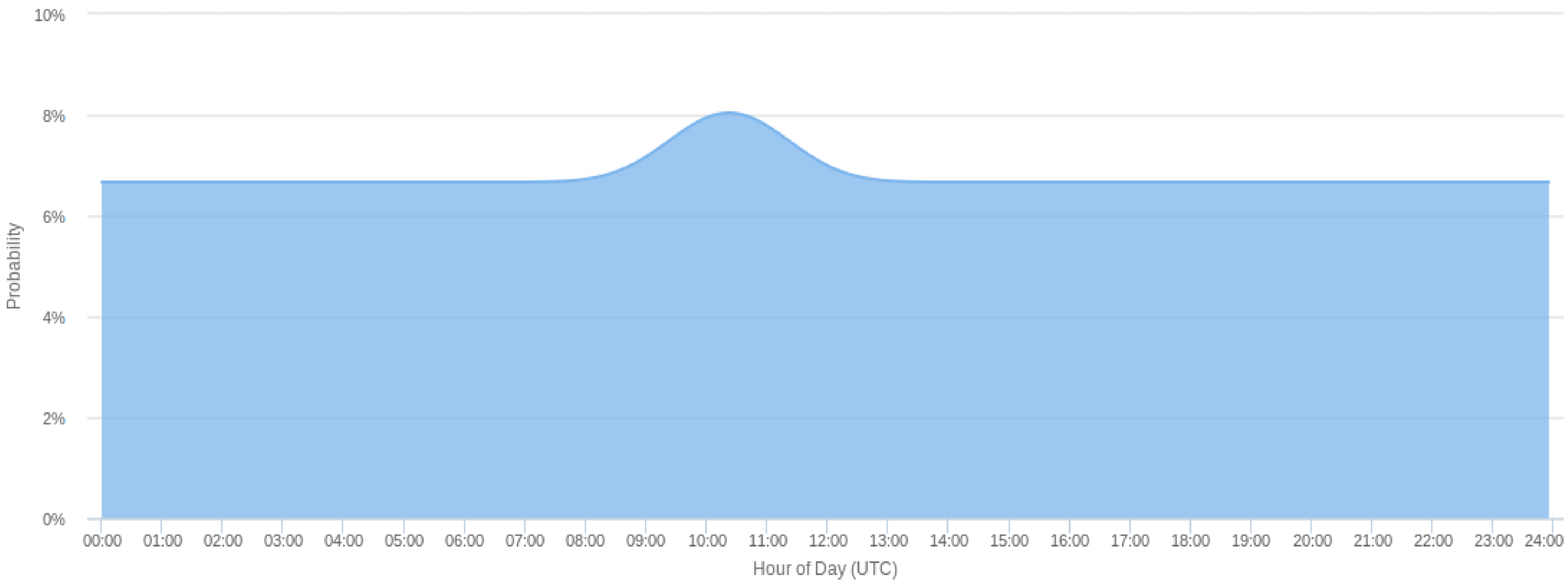
Compare user patterns to each other



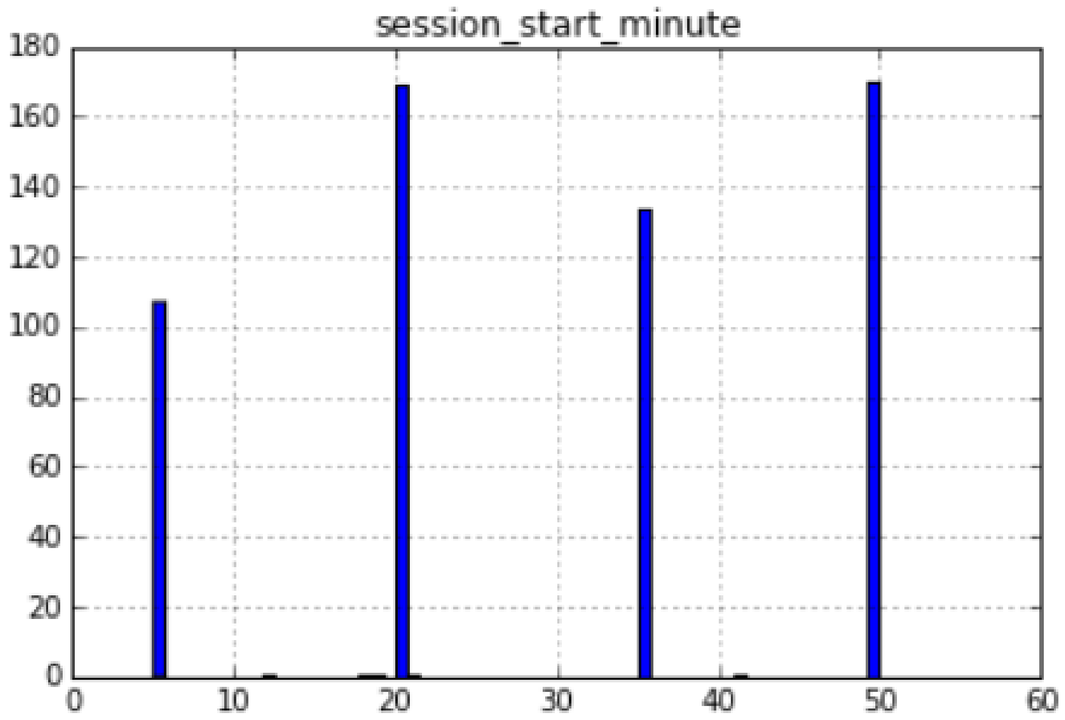
Robot detection

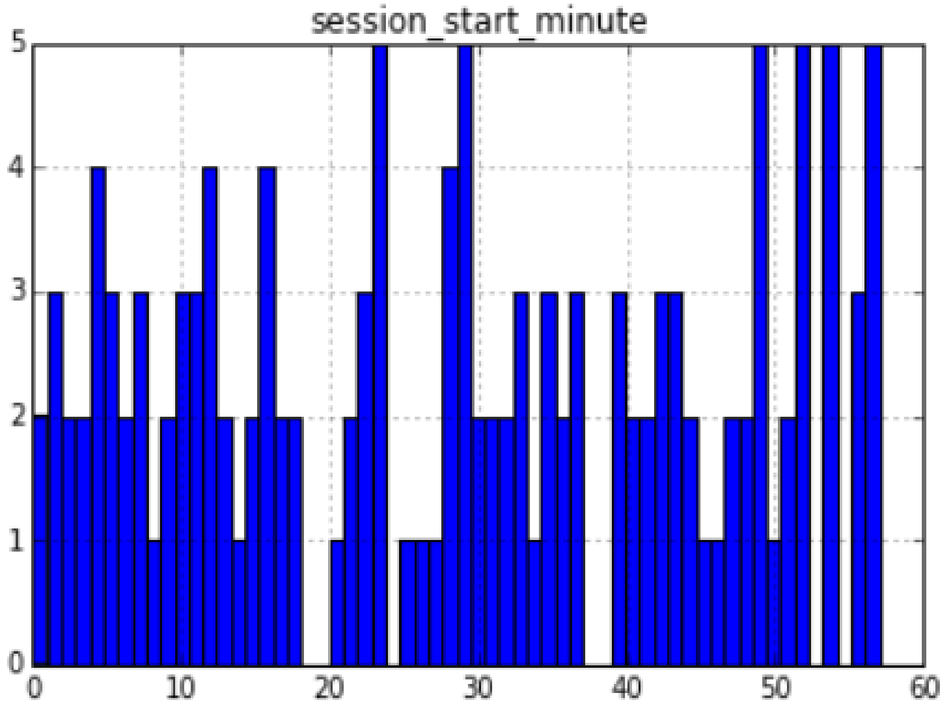


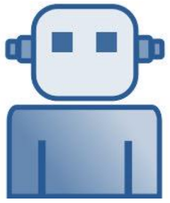
Logintime example



Human or script?





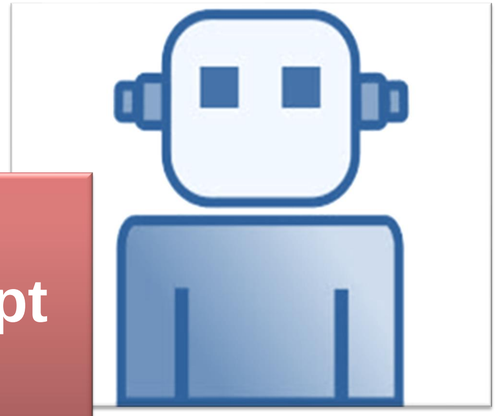


Summary

Change

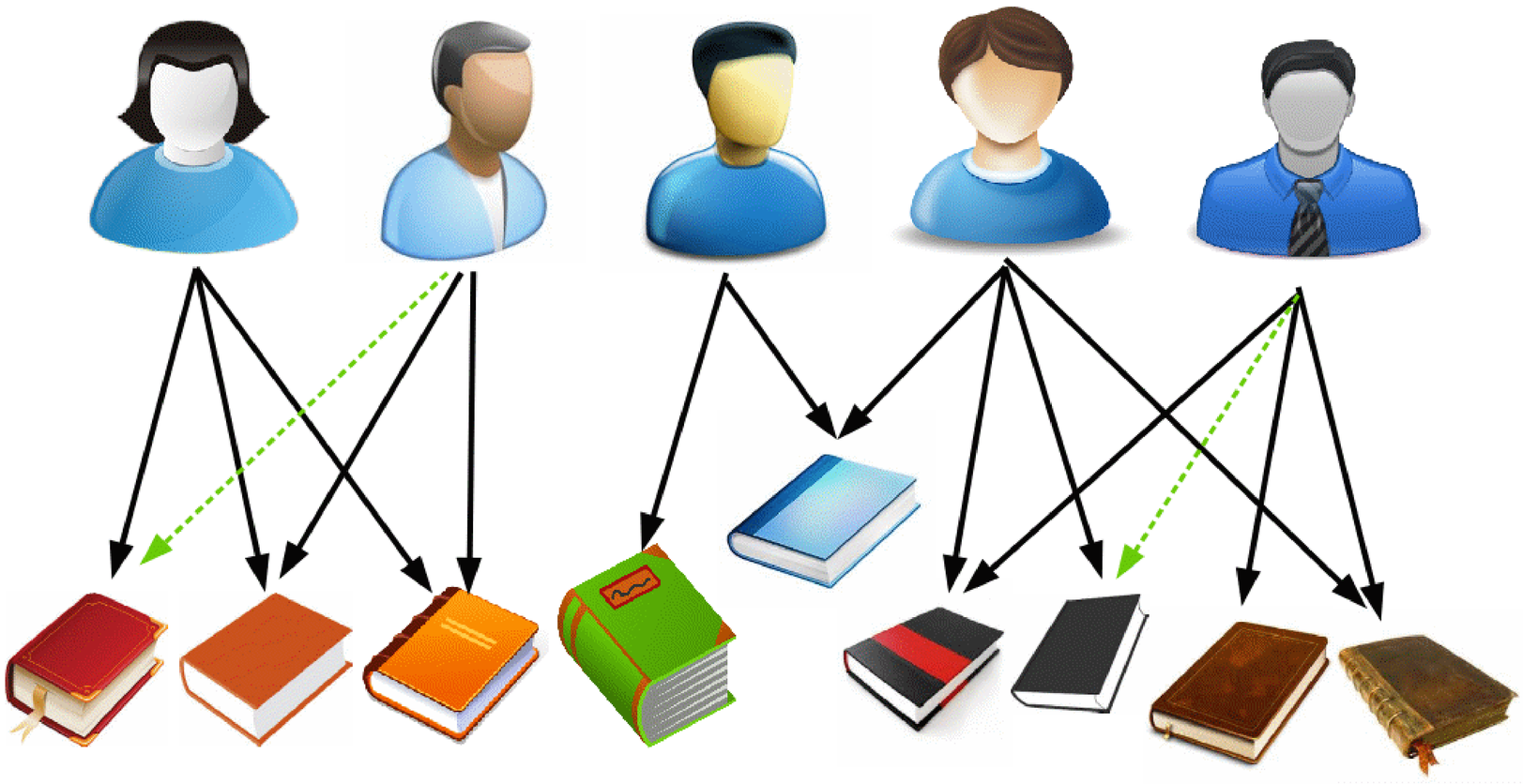
Outlier

Script

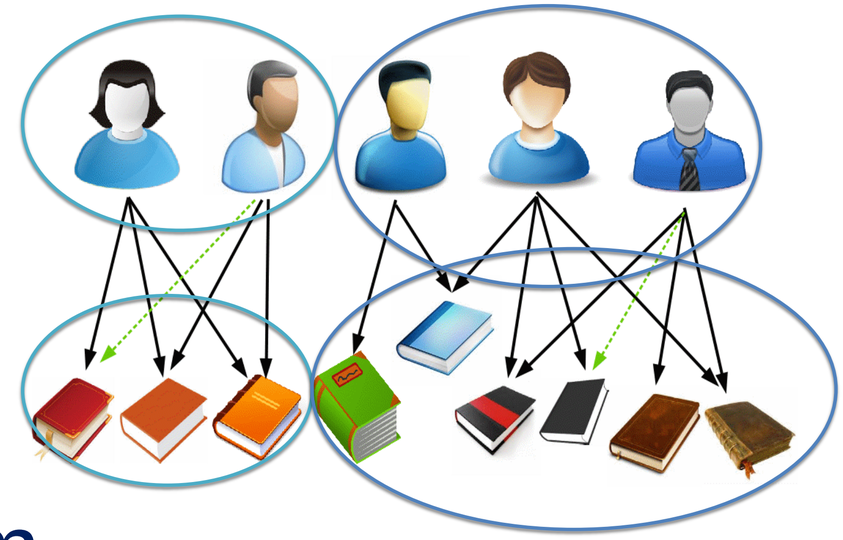


User risk

Recommender systems

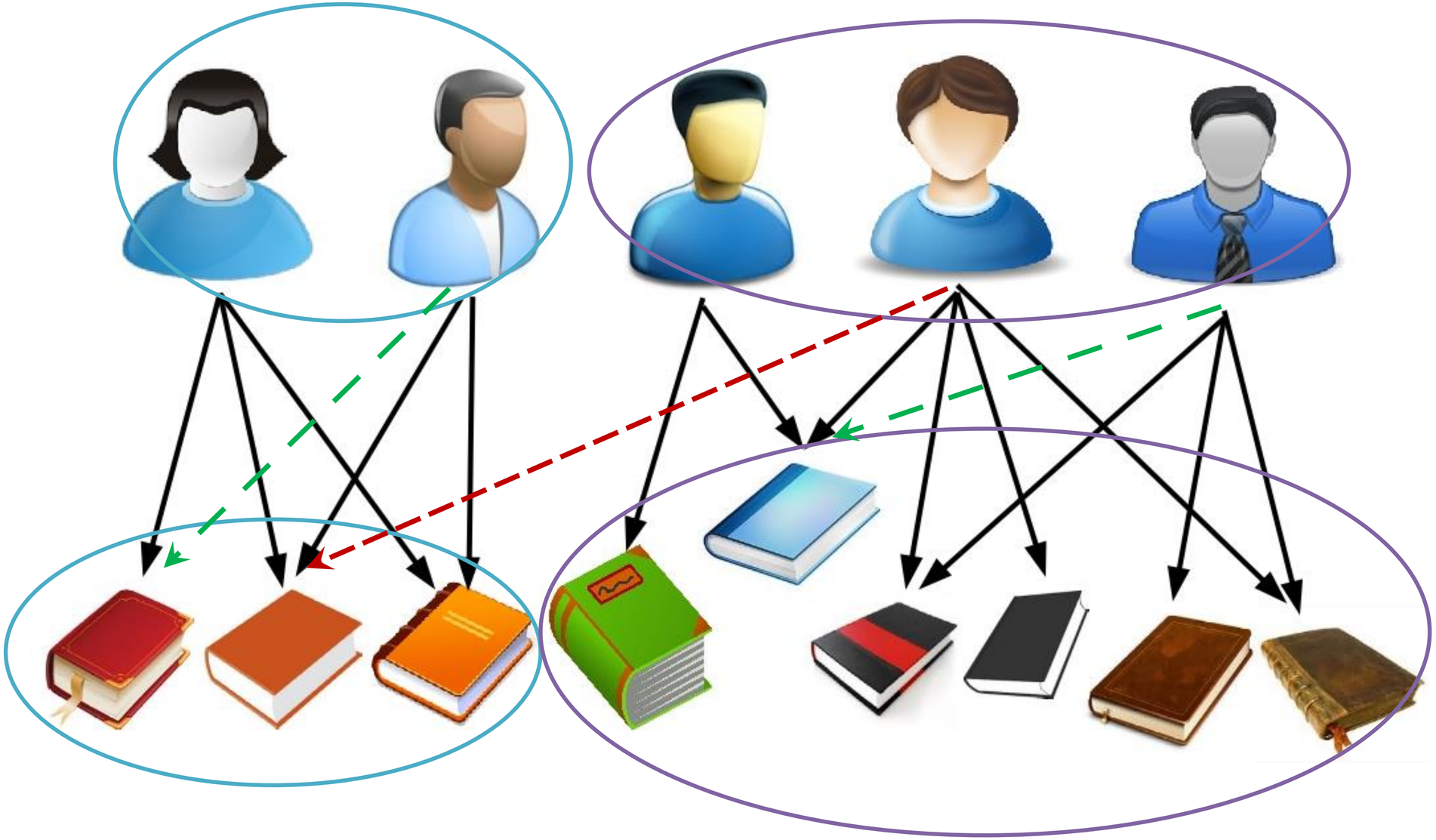


Recommender systems



- Content based approach
- Collaborative filtering
  - nearest neighborhood algorithm (user or item similarity)
  - latent factor model

Anomaly detection



BALABIT

Example







	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	
Jabba	10					1	
Lars Owen	5	1					
Leia Organa		1	3				
~~R2D2~~	~~2~~	~~2~~	~~2~~	~~2~~	~~2~~	~~2~~	
Wilhuff Tarkin					10	3	
Yoda	1			8			
Darth Vader	2				20	10	
Emperor		1			15	8	
Luke Skywalker	8		3	2			
Wuher	6	1					
Obi Wan Kenobi	1		4	4			
Boba Fett	1				5	4	
Lando Calrissian							5

Steps

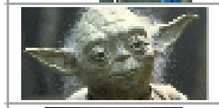
1. **Transform: get relative frequencies**
2. **Calculate item distances**
3. **Convert into item similarities**
4. **Calculate affinities**

Relative frequencies







	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
 Jabba	0.909	0	0	0	0	0.091	0
 Lars Owen	0.833	0.167	0	0	0	0	0
 Leia Organa	0	0.25	0.75	0	0	0	0
 Wilhuff Tarkin	0	0	0	0	0.769	0.231	0
 Yoda	0.111	0	0	0.889	0	0	0
 Darth Vader	0.062	0	0	0	0.625	0.312	0
 Emperor	0	0.042	0	0	0.625	0.333	0
 Luke Skywalker	0.615	0	0.231	0.154	0	0	0
 Wuher	0.857	0.143	0	0	0	0	0
 Obi Wan Kenobi	0.111	0	0.444	0.444	0	0	0
 Boba Fett	0.1	0	0	0	0.5	0.4	0
 Lando Calrissian	0	0	0	0	0	0	1

Item similarities

Bray-Curtis **distance**

	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Tatooine	0	0,853	0,864	0,852	0,947	0,898	1
Alderaan	0,853	0	0,753	1	0,973	0,958	1
Hoth	0,864	0,753	0	0,589	1	1	1
Dagobah	0,852	1	0,589	0	1	1	1
Death Star	0,947	0,973	1	1	0	0,343	1
Kamino	0,898	0,958	1	1	0,343	0	1
Bespin	1	1	1	1	1	1	0

Similarity = 1 - distance

$$d(u,v) = \frac{\sum_i |u_i - v_i|}{\sum_i |u_i + v_i|}$$

	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Tatc				148	0,053	0,102	0
Alde				0	0,027	0,042	0
Hoth	0,136	0,247	1	0,411	0	0	0
Dagobah	0,148	0	0,411	1	0	0	0
Death Star	0,053	0,027	0	0	1	0,657	0
Kamino	0,102	0,042	0	0	0,657	1	0
Bespin	0	0	0	0	0	0	1

Affinity calculation

	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Tatooine	1	0,147	0,136	0,148	0,053	0,102	0
Alderaan	0,147	1	0,247	0	0,027	0,042	0
Hoth	0,136	0,247	1	0,411	0	0	0
Dagobah	0,148	0	0,411	1	0	0	0
Death Star	0,053	0,027	0	0	1	0,657	0
Kamino	0,102	0,042	0	0	0,657	1	0
Bespin	0	0	0	0	0	0	1

0,0033 + 0 + 0 + 0+ 0,625+ 0,313 + 0

Relative frequencies

	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Darth Vader	0,0625	0	0	0	0,625	0,3125	0

	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Darth Vader					?		

Affinity calculation

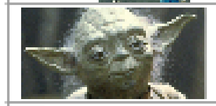
	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Tatooine	1	0,147	0,136	0,148	0,053	0,102	0
Alderaan	0,147	1	0,247	0	0,027	0,042	0
Hoth	0,136	0,247	1	0,411	0	0	0
Dagobah	0,148	0	0,411	1	0	0	0
Death Star	0,053	0,027	0	0	1	0,657	0
Kamino	0,102	0,042	0	0	0,657	1	0
Bespin	0	0	0	0	0	0	1

Relative frequencies

	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Darth Vader	0,0625	0	0	0	0,625	0,3125	0

	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Darth Vader					0,834		

Affinities

	 Tatooine	 Alderaan	 Hoth	 Dagobah	 Death Star	 Kamino	 Bespin
 Jabba	0.918	0.138	0.124	0.134	0.108	0.184	0
 Lars Owen	0.858	0.289	0.155	0.123	0.049	0.092	0
 Leia Organa	0.139	0.435	0.812	0.308	0.007	0.011	0
 Wilhuff Tarkin	0.064	0.03	0	0	0.921	0.736	0
 Yoda	0.243	0.016	0.38	0.905	0.006	0.011	0
 Darth Vader	0.128	0.039	0.009	0.009	0.834	0.729	0
 Emperor	0.073	0.072	0.01	0	0.845	0.746	0
 Luke Skywalker	0.67	0.148	0.378	0.34	0.033	0.063	0
 Wuher	0.878	0.269	0.152	0.127	0.049	0.094	0
 Obi Wan Kenobi	0.237	0.126	0.642	0.643	0.006	0.011	0
 Boba Fett	0.167	0.045	0.014	0.015	0.768	0.739	0
 Lando Calrissian	0	0	0	0	0	0	1

Anomaly scores

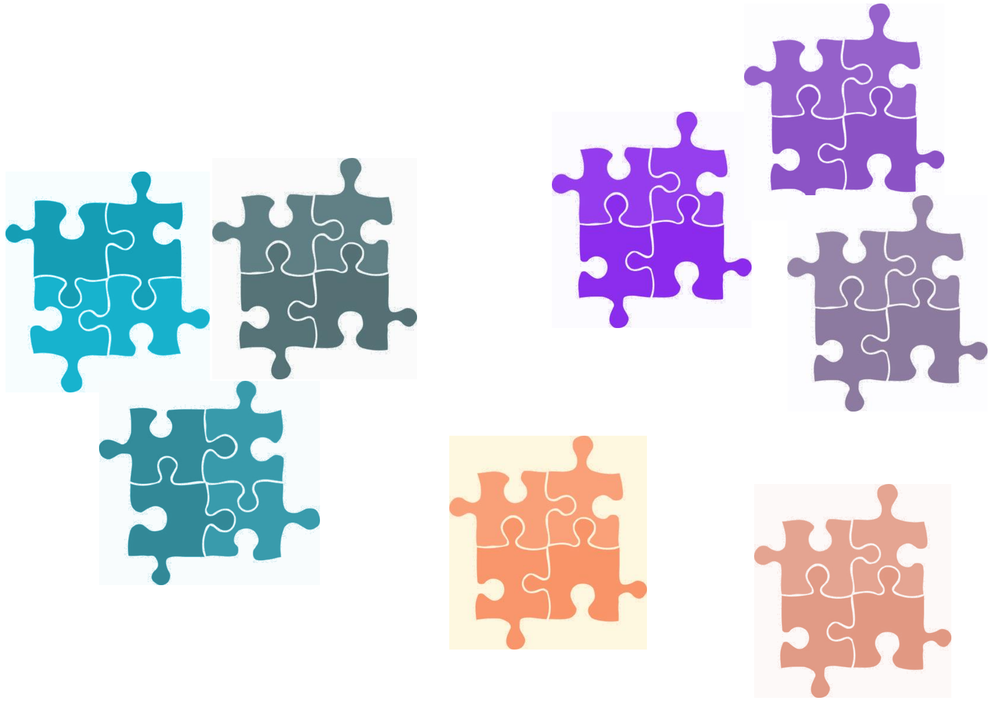
100*(1-x)

	Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Jabba	8	86	88	87	89	82	100
Lars Owen	14	71	85	88	95	91	100
Leia Organa	86	56	19	69	99	99	100
Wilhuff Tarkin	94	97	100	100	8	26	100
Yoda	76	98	62	9	99	99	100
Darth Vader	87	96	99	99	17	27	100
Emperor	93	93	99	100	15	25	100
Luke Skywalker	33	85	62	66	97	94	100
Wuher	12	73	85	87	95	91	100
Obi Wan Kenobi	76	87	36	36	99	99	100
Boba Fett	83	95	99	99	23	26	100
Lando Calrissian	100	100	100	100	100	100	0

Anomaly detection

Compare activities to user patterns

Compare user patterns to each other

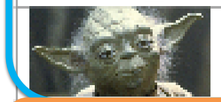


Peer groups



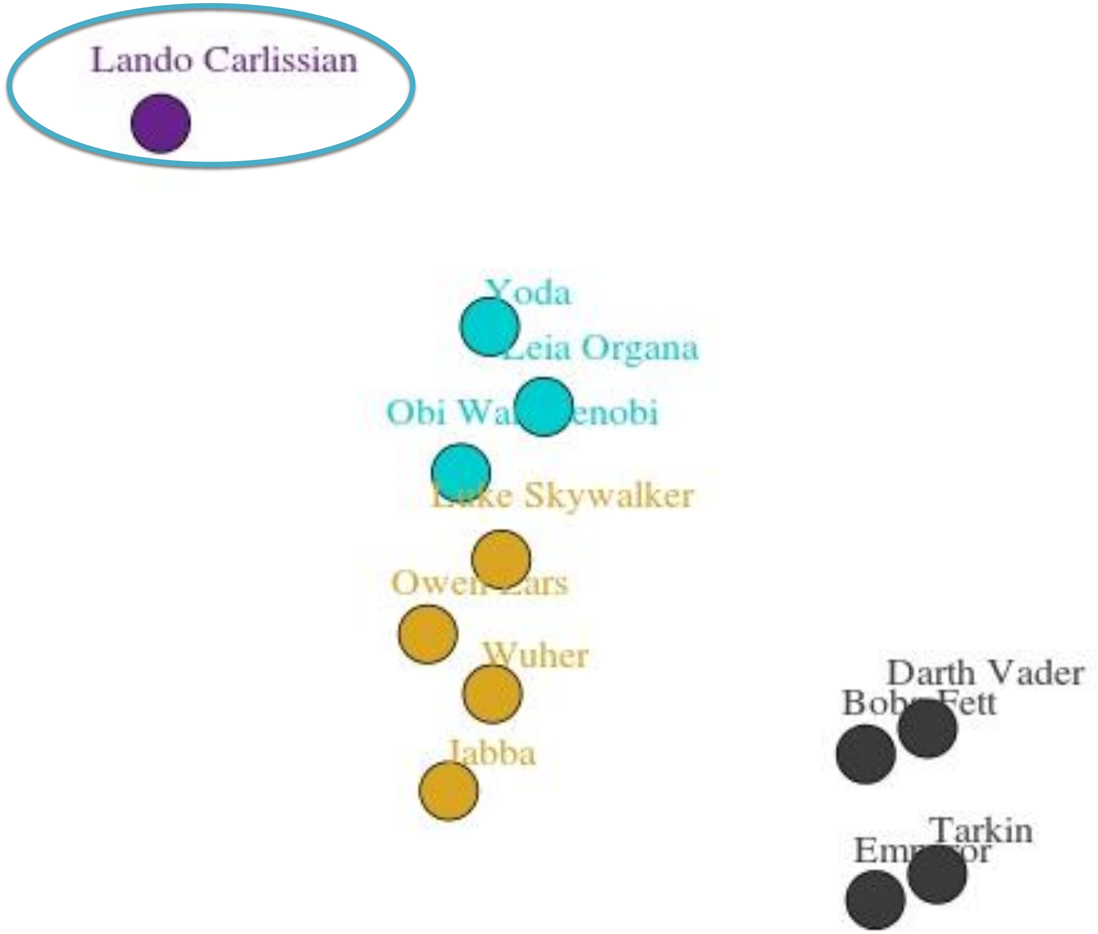




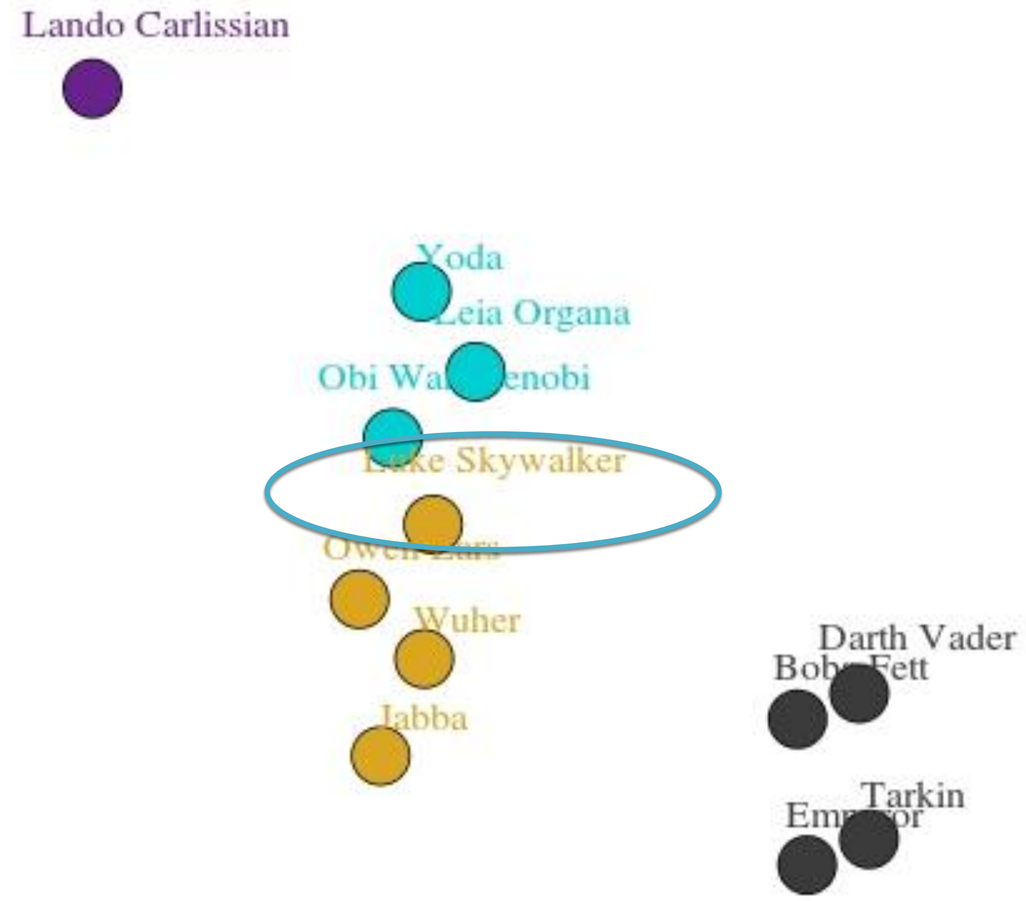
		Tatooine	Alderaan	Hoth	Dagobah	Death Star	Kamino	Bespin
Dark side	 Wilhuff Tarkin	94	97	100	100	8	26	100
	 Emperor	93	93	99	100	15	25	100
	 Darth Vader	87	96	99	99	17	27	100
	 Boba Fett	83	95	99	99	23	26	100
Tatooine	 Jabba	8	86	88	87	89	82	100
	 Wuher	12	73	85	87	95	91	100
	 Lars Owen	14	71	85	88	95	91	100
	 Luke Skywalker	33	85	62	66	97	94	100
Rebels	 Leia Organa	86	56	19	69	99	99	100
	 Obi Wan Kenobi	76	87	36	36	99	99	100
	 Yoda	76	98	62	9	99	99	100
Other	 Lando Calrissian	100	100	100	100	100	100	0

Anomalous user detection

Far from the peer groups



Far from the colleagues



Summary

- Build recommender system
- Detect anomalous **activities**
- Identify peer groups
- Detect anomalous **users**