“On Granular High-Resolution Analytics in Big Data Applications”

Irad Ben-Gal

Paris, April 9, 2014

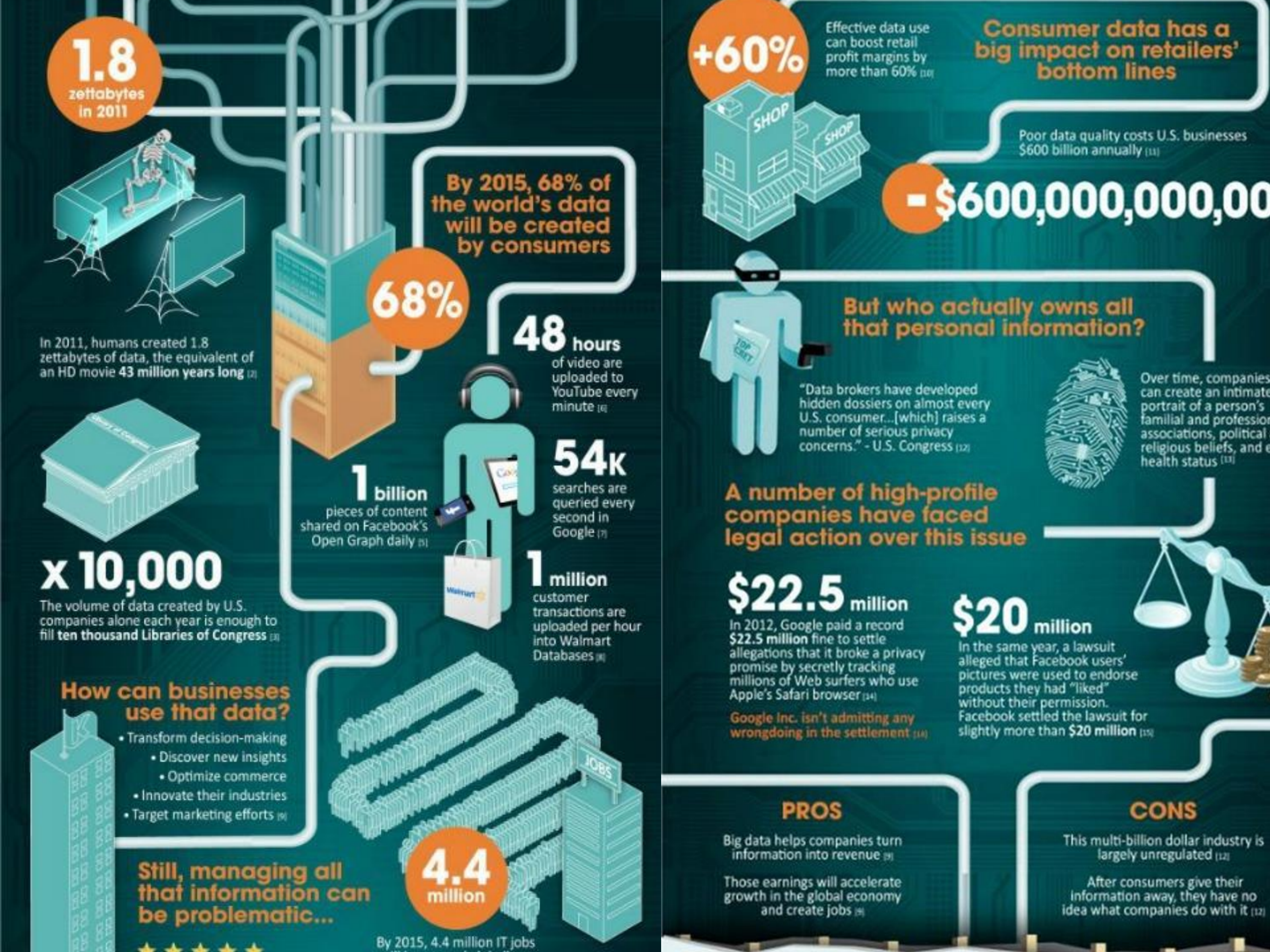
http://www.eng.tau.ac.il/~bengal/

http://www.linkedin.com/pub/irad-ben-gal/4/582/4b4

bengal@tau.ac.il

Agenda

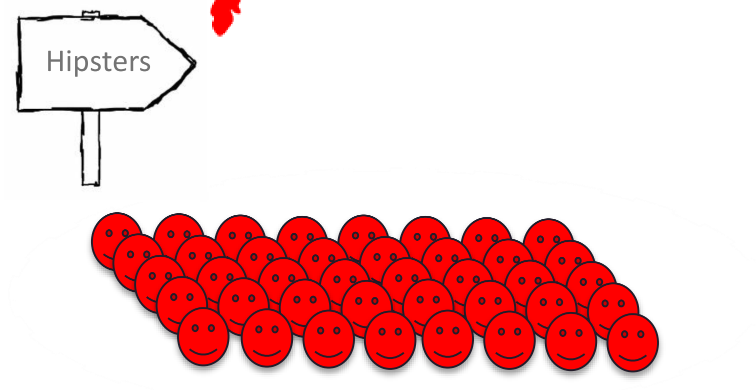
- Granular Big Data Analytics Opportunity
- Pattern-Based Graphical Modeling
- Academic Example & Real-world Applications
  - **Anomaly Detection in Cellular Nets**
  - **Learning Mobility Patterns**
  - **Predictive maintenance (vehicle battery)**
  - **Suspect Tracking**
  - User-Oriented Cyber Security
  - Demand Sensing & Inventory Control
  - e-Commerce analysis
- Q&A

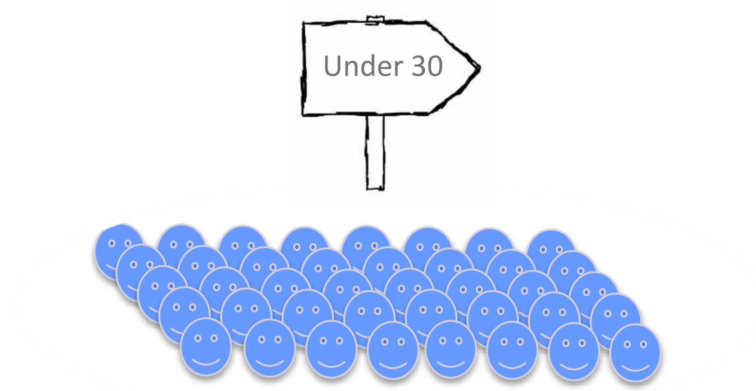

1.8
zettabytes
in 2011
In 2011, humans created 1.8 zettabytes of data, the equivalent of an HD movie 43 million years long [2]
By 2015, 68% of the world's data will be created by consumers
68%
48 hours
of video are uploaded to YouTube every minute [6]
54K
searches are queried every second in Google [7]
1 billion
pieces of content shared on Facebook's Open Graph daily [5]
1 million
customer transactions are uploaded per hour into Walmart Databases [8]
x 10,000
The volume of data created by U.S. companies alone each year is enough to fill ten thousand Libraries of Congress [3]
How can businesses use that data?
• Transform decision-making
• Discover new insights
• Optimize commerce
• Innovate their industries
• Target marketing efforts [9]
Still, managing all that information can be problematic...
4.4
million
JOBS
By 2015, 4.4 million IT jobs
+60%
Effective data use can boost retail profit margins by more than 60% [10]
Consumer data has a big impact on retailers' bottom lines
SHOP
SHOP
Poor data quality costs U.S. businesses $600 billion annually [11]
-$600,000,000,00
But who actually owns all that personal information?
TOP SECRET
"Data brokers have developed hidden dossiers on almost every U.S. consumer...[which] raises a number of serious privacy concerns." - U.S. Congress [12]
Over time, companies
can create an intimate
portrait of a person's
familial and professio
associations, political
religious beliefs, and e
health status [13]
A number of high-profile companies have faced legal action over this issue
$22.5 million
In 2012, Google paid a record $22.5 million fine to settle allegations that it broke a privacy promise by secretly tracking millions of Web surfers who use Apple's Safari browser [14]
Google Inc. isn't admitting any wrongdoing in the settlement [14]
$20 million
In the same year, a lawsuit alleged that Facebook users' pictures were used to endorse products they had "liked" without their permission. Facebook settled the lawsuit for slightly more than $20 million [15]
PROS
Big data helps companies turn information into revenue [9]
Those earnings will accelerate growth in the global economy and create jobs [9]
CONS
This multi-billion dollar industry is largely unregulated [12]
After consumers give their information away, they have no idea what companies do with it [12]

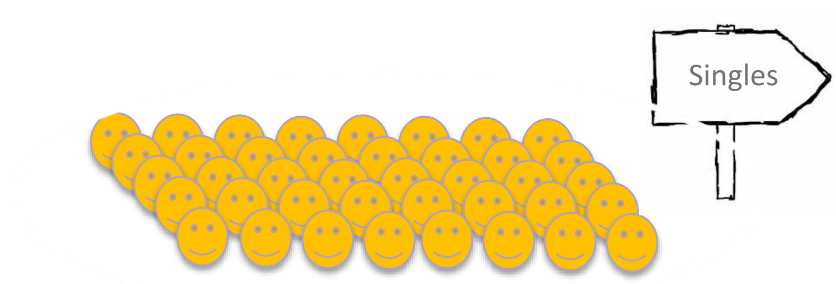
Granular “High Resolution” Analytics

Granularity vs. Segmentation

~~Granularity~~







GAP=

Opportunity!

Hidden Patterns Example: Virtual & Smart TV

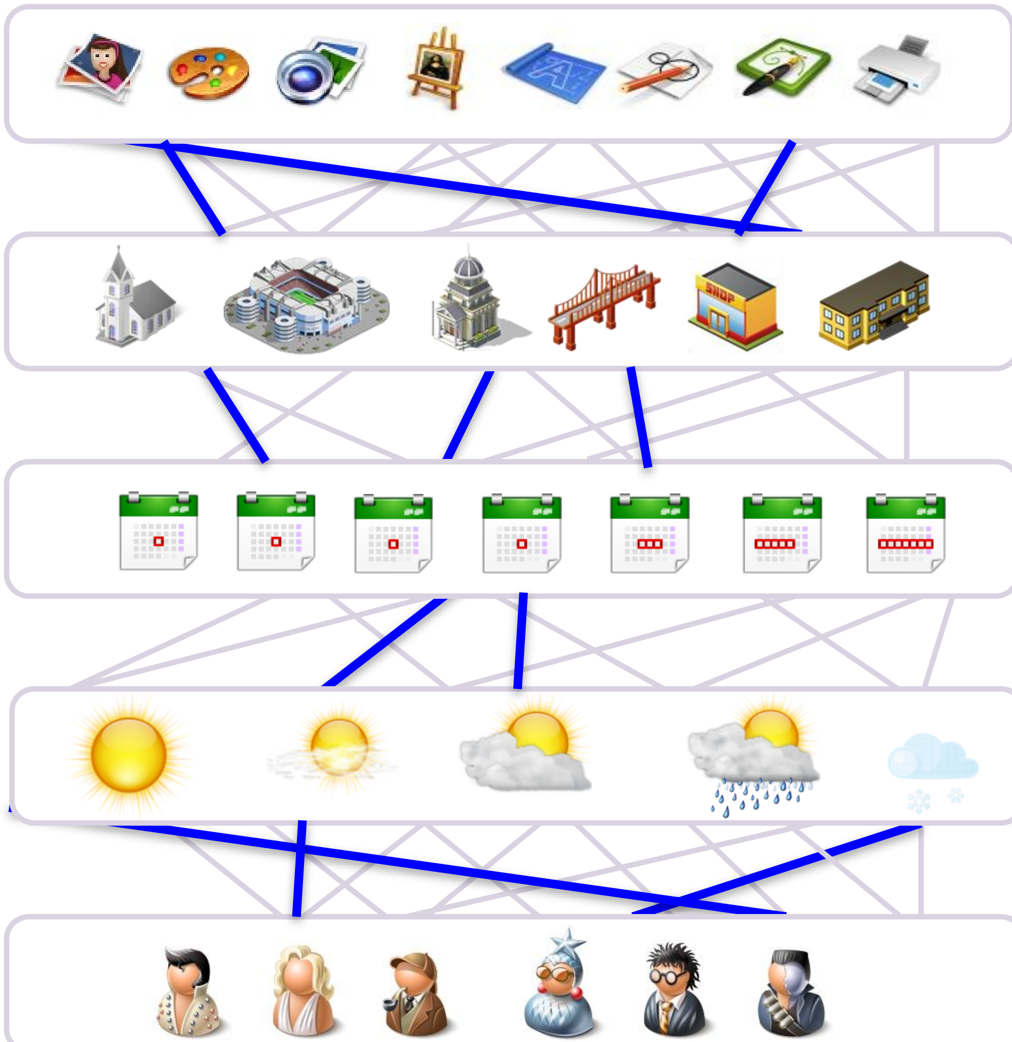
Products (8)

(e)Locations (6)

Week day (7)

Weather (5)

Customer types (6)



In this example:
4,294,967,296
potential patterns!!!

Sample pattern:
“On rainy Tuesdays customer A look for service B via smartphone if he is in downtown”

C-B4 identifies the significant patterns and maps them into business opportunities.

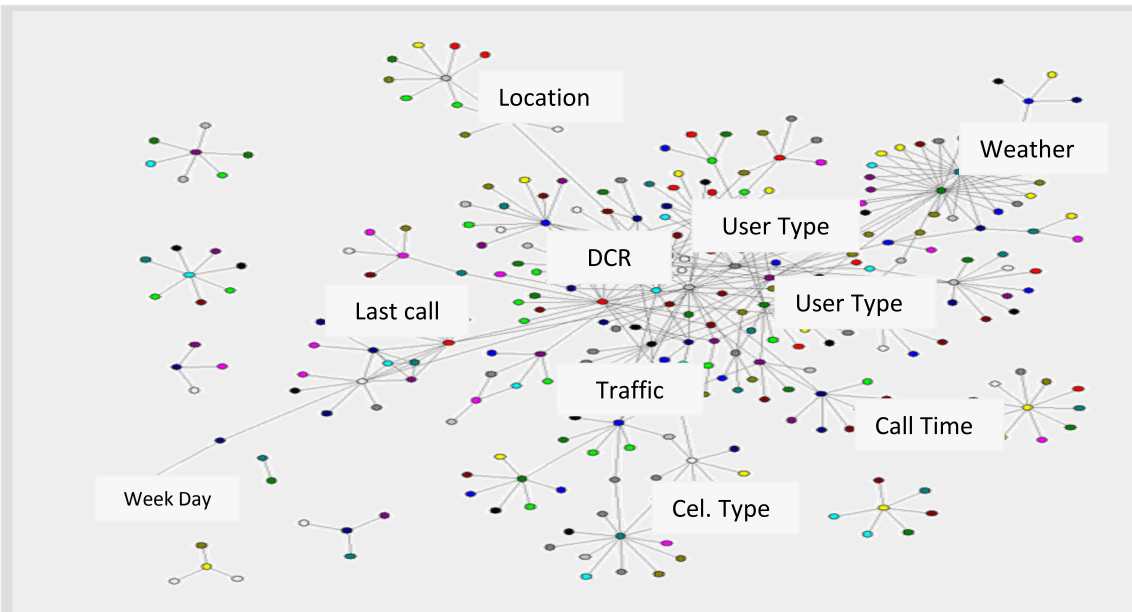
Granular High-Resolution Analytics

- Volume
- Velocity
- Variability
- …… on Information Quality (Kenett and Shmueli 2013)

Apply to the principles to BD modeling!

Data Compression Graphical models

Instated of library of statistical models



Data

h	l	l	m
l	h	m	h
m	m	h	m
h	l	m	l
m	h	l	h
l	m	h	m

Pattern-Based Monitoring & Predictions per entity

Cellular {ID}+Location {A}+Time slot {Noon} Service request {A} ➔ Anomaly {True} Priority {High}

IP {#136.55.65.02}+Time {Sunday noon}+View {Page A} ➔ Alert {True}, Probability{>0.8}

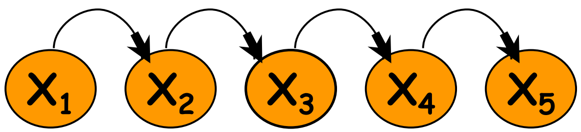
SP user {Type A} + TV Program B {Low rate} + Promotion {True} + Weather {Rain} ➔ Program C {High} > 0.8

BN Generalization **[Ben-Gal I., et al, ~ 5 publications & patents]**

$P(X_1\, X_2\, X_3\, X_4\, X_5) = ?$ Joint distribution of a sequence of RV

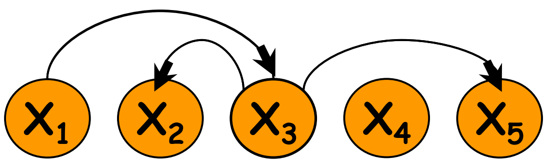
Independent: marginal distributions

$$P(X_1 \cdots X_5) = P(X_1)P(X_2)P(X_3)P(X_4)P(X_5)$$



Markov Model: fixed-order dependency

$$P(X_1 \cdots X_5) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)P(X_4 \mid X_3)P(X_5 \mid X_4)$$



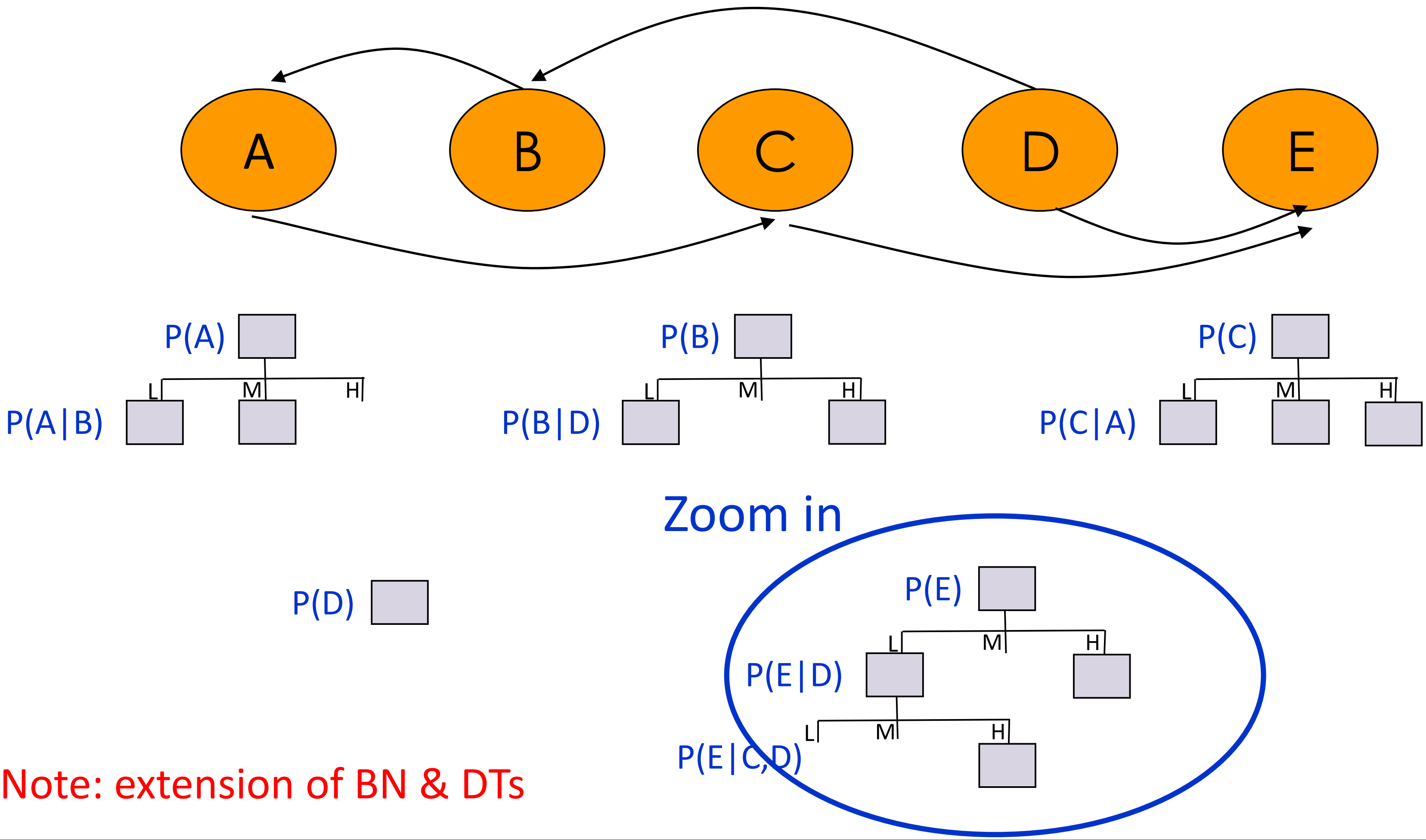
BN Tree: Direct dependencies

$$P(X_1 \cdots X_5) = P(X_1)P(X_2 \mid X_3)P(X_3 \mid X_1)P(X_4)P(X_5 \mid X_3)$$

The model can be further generalized to VOBN

VOBN Model (inhomogeneous & state-dependent net of variable-order trees)

Input data: 5 Products, 3 levels of sales {Low, Medium, High}

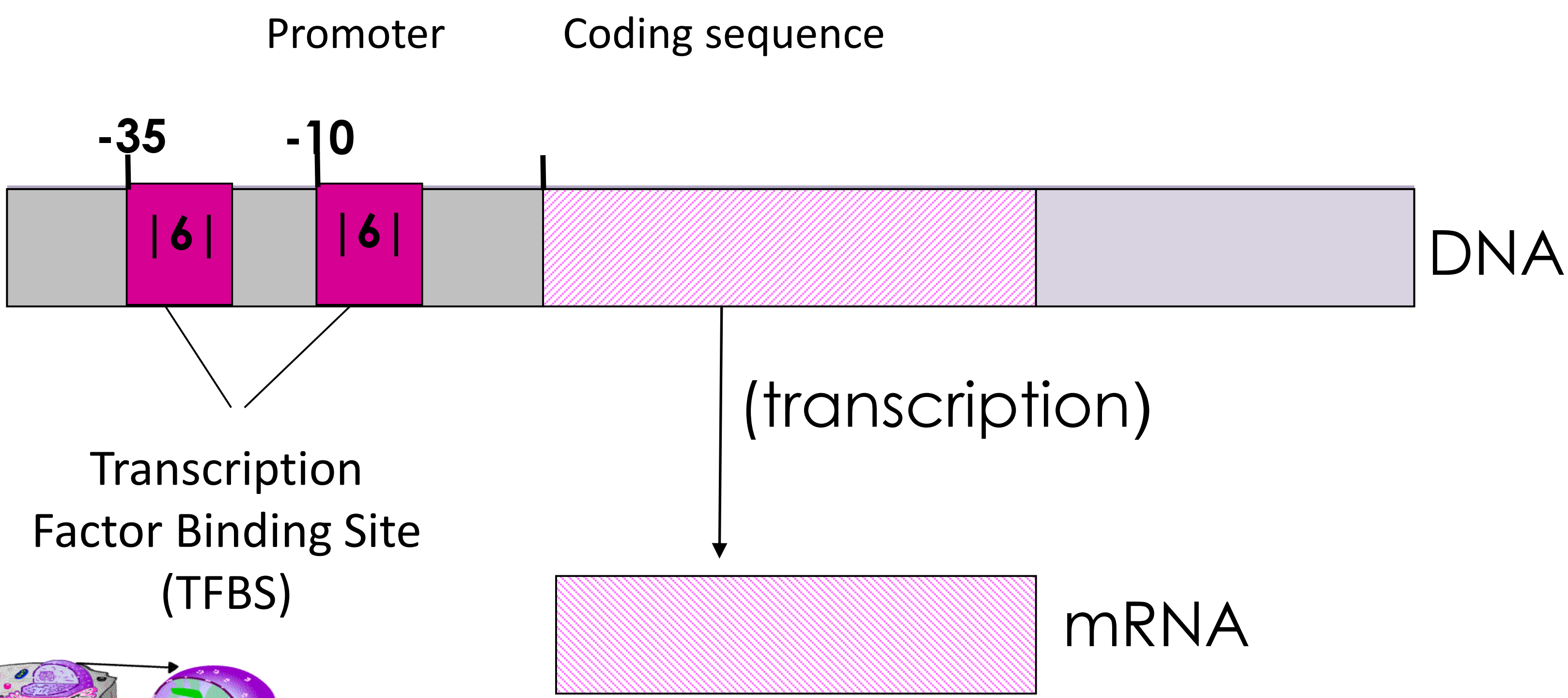


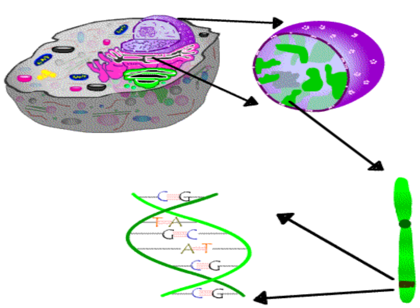
Note: extension of BN & DTs

Agenda

- Big Data Analytics Opportunity
- Pattern-Based Graphical modeling
- **Academic Example**
- Real-world applications
  - **Anomaly Detection in Cellular Nets**
  - **Learning Mobility Patterns**
  - **Predictive maintenance (vehicle battery)**
  - **Suspect Tracking**
  - User-Oriented Cyber Security
  - Demand Sensing & Inventory Control
  - e-Commerce analysis
- Q&A

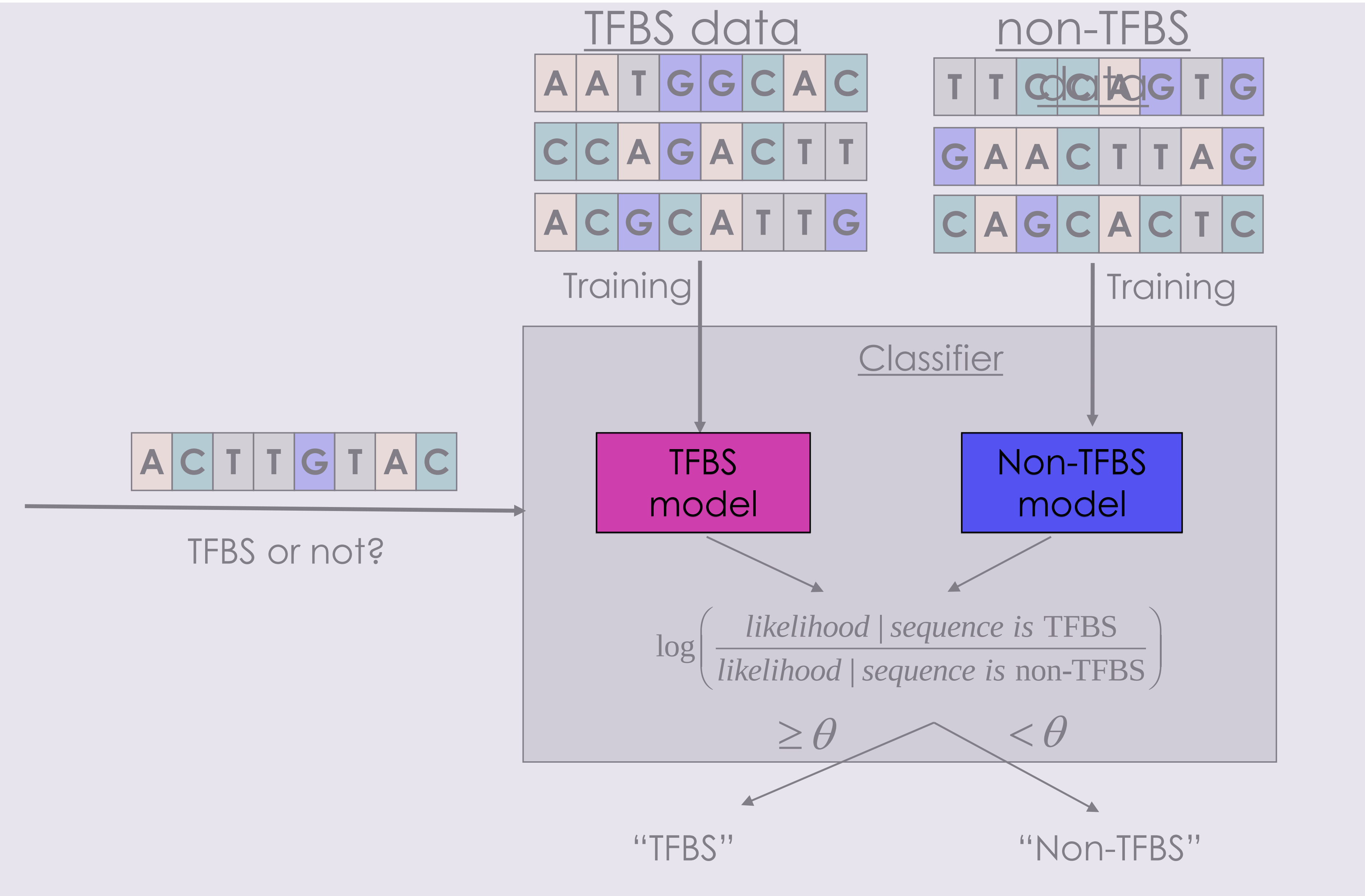
TFBS Identification (Ben-Gal et al., Bioinformatics 2005)



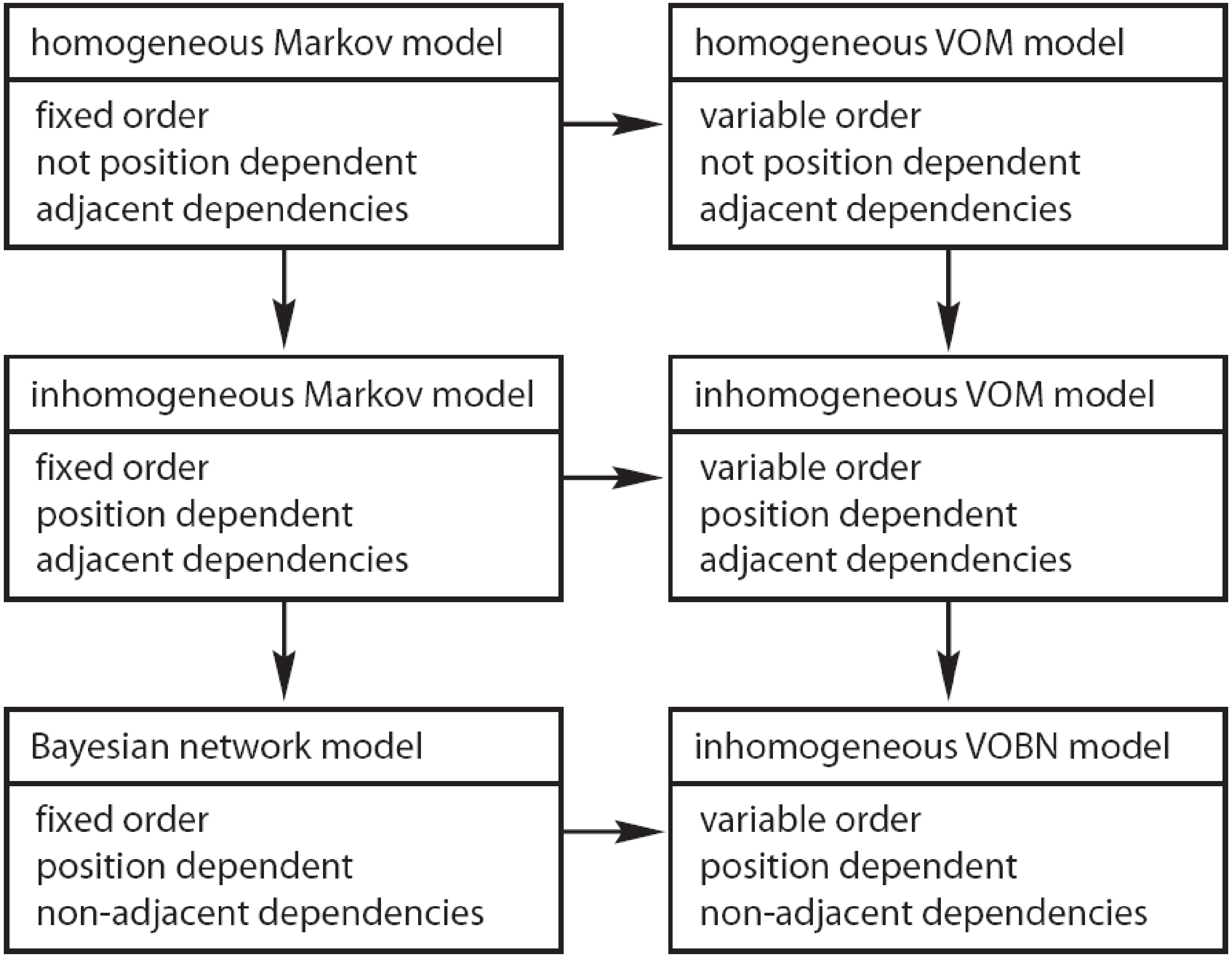


The goal (#1): identify TFBS in E.coli

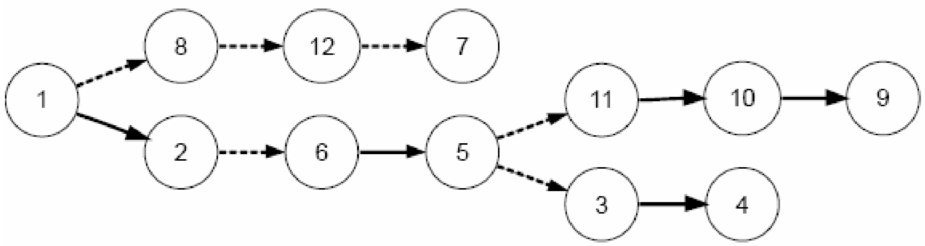
TFBS Identification

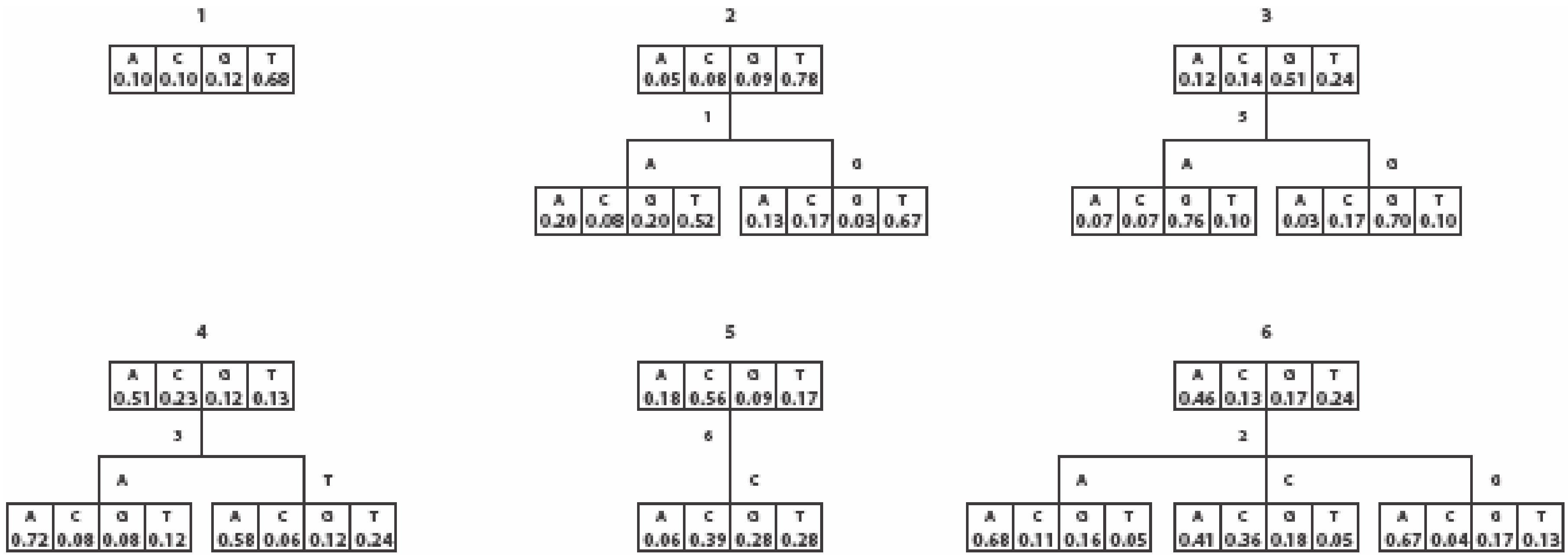


TFBS Identification (Ben-Gal et al., Bioinformatics 2005)



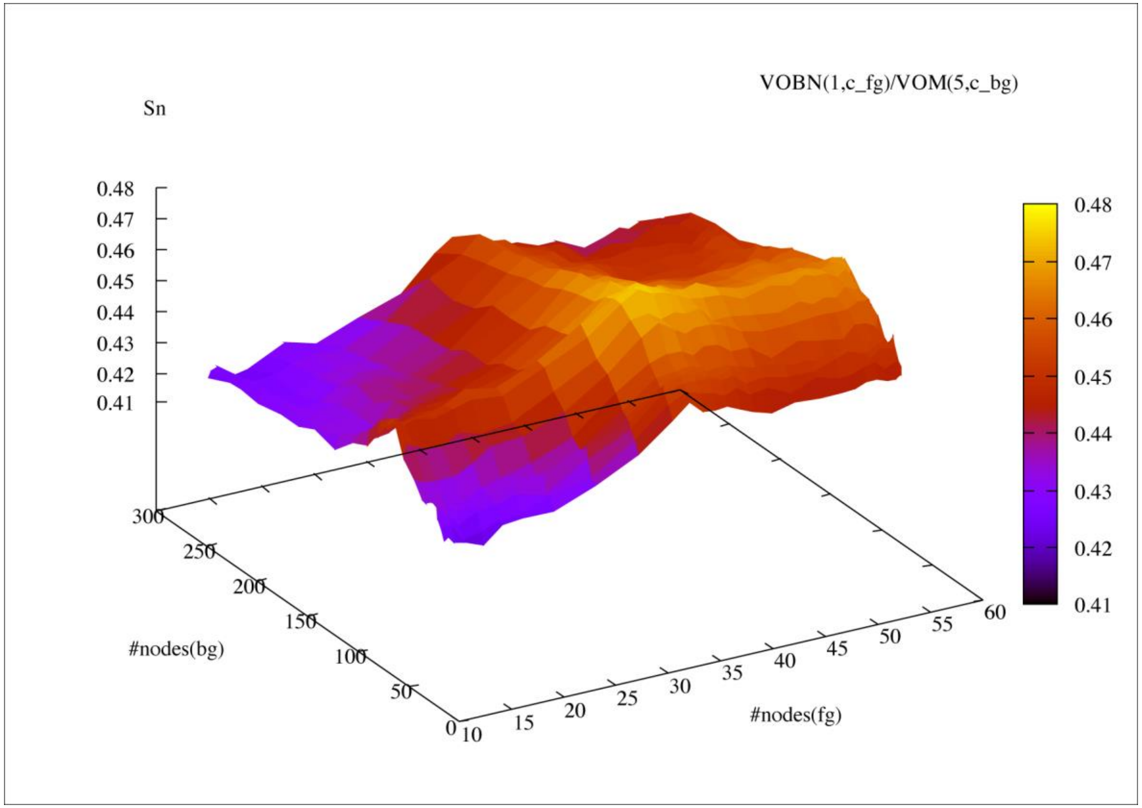
Academic Example (Ben-Gal et al., 2005)





Academic Example 2 (Ben-Gal et al., Bioinformatics 2005)

Sensitivity: 47.8%



Fixed specificity (TN rate) = 99.9%, Std. error = 0.01%

Academic Example 2 (Ben-Gal et al., Bioinformatics 2005)

	TP (for TN=99.9%)
PWM/Markov(2)	43.5%
VOBN models	47.8%
Overall Improvement	+4.3%

E. Coli Genome (~ 4000 genes) $\Rightarrow$ + 172 potential TFBS

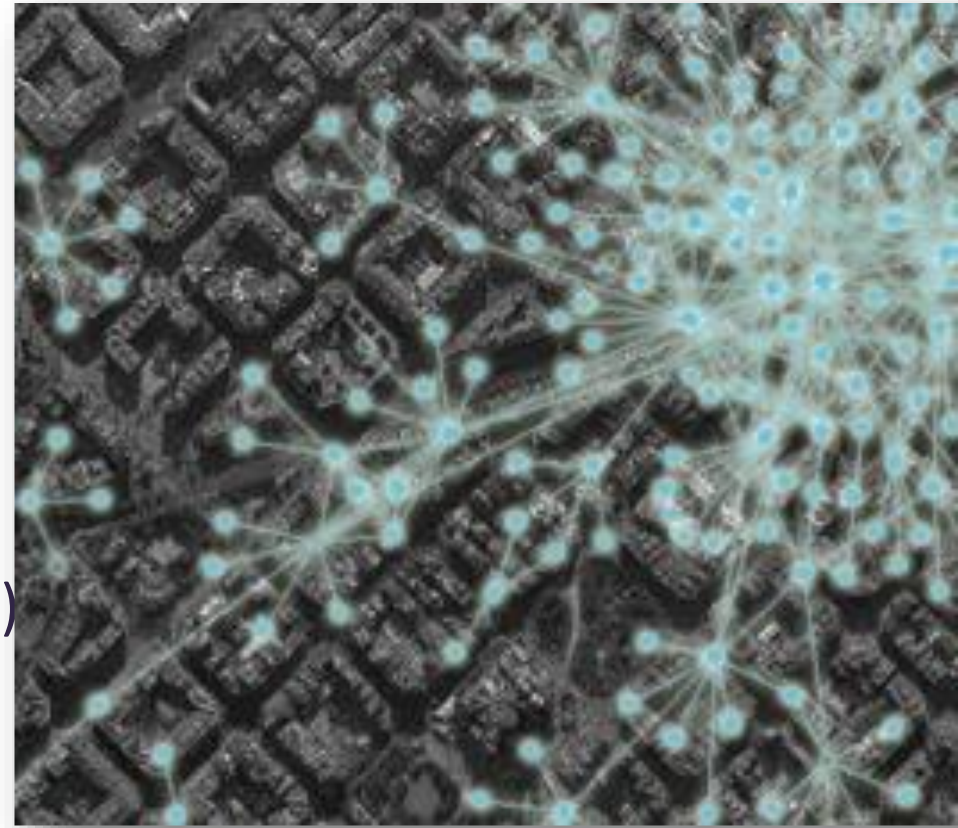
Human Genome (~25,000 genes) $\Rightarrow$ + 1075 suspected TFBS

Agenda

- Big Data Analytics Opportunity
- Granular Pattern-Based Graphical modeling
- Real-world Applications
  - **Anomaly Detection in Cellular Nets**
  - **Learning Mobility Patterns**
  - **Predictive maintenance (vehicle battery)**
  - **Suspect Tracking**
  - User-Oriented Cyber Security
  - Demand Sensing & Inventory Control
  - e-Commerce analysis
- Q&A

Learning Mobility Patterns (LMP)

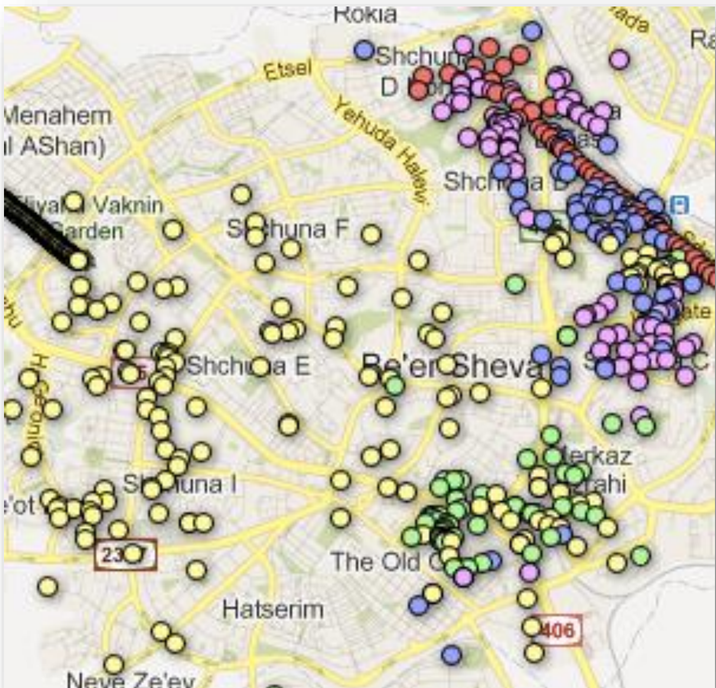
- Analyzing Mobility Patterns in city areas
  - Using massive data sets
  - Privacy-preserving methods
- Funded by the Ministry of Science Infrastructure grant.
- Join work with Dr. Eran Toch (TAU) Prof. Boaz Lerner (BGU).

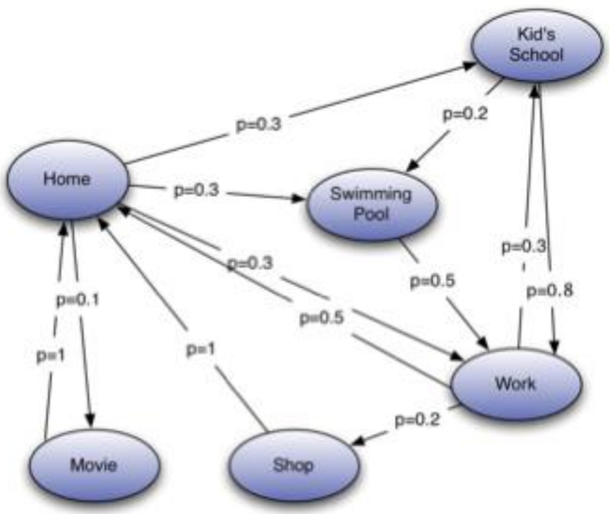


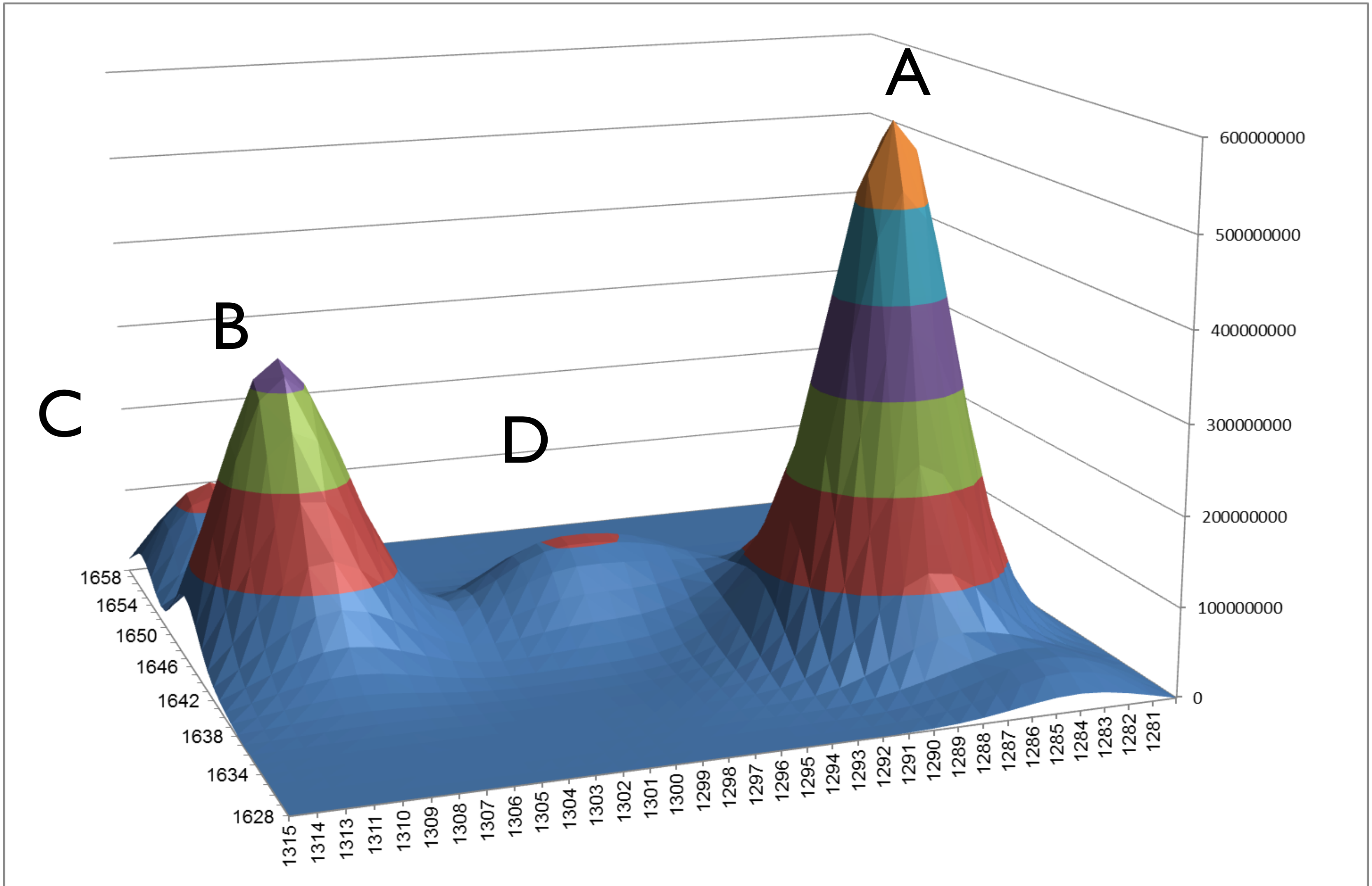
31

LMP: Modeling People & Places

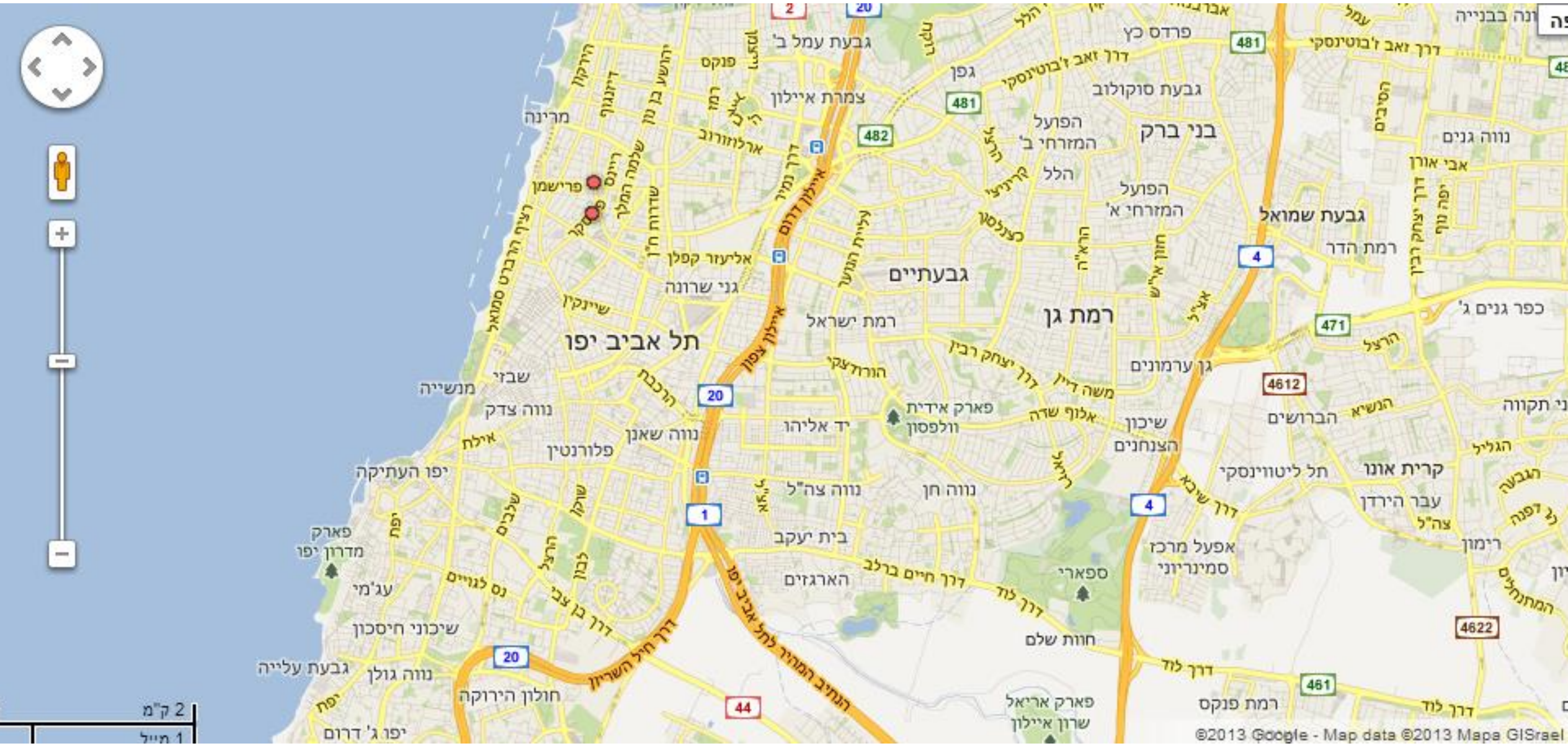
Recognizing meaningful places, analyzing their use, and using probabilistic models to predict people behavior.



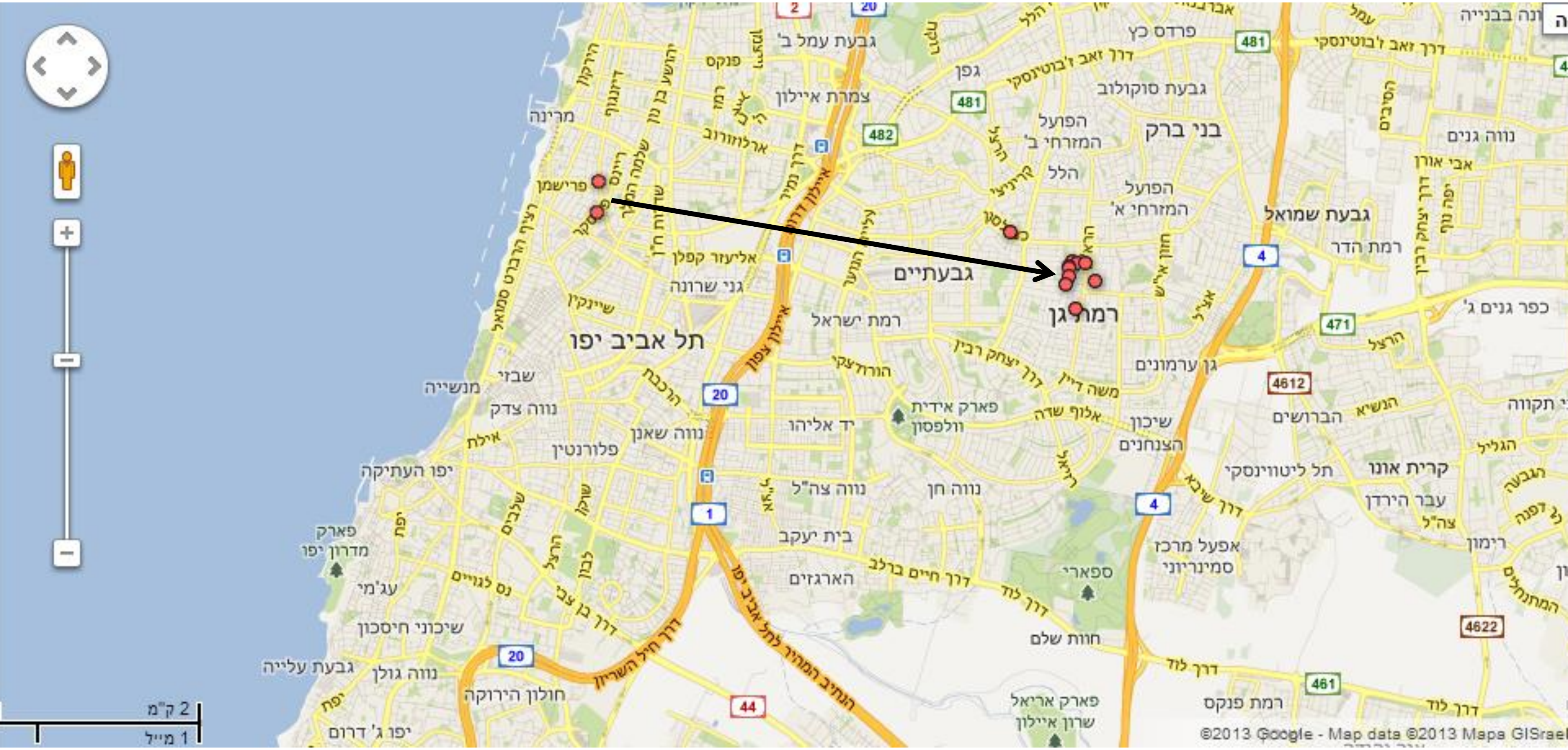



A
B
C
D
600000000
500000000
400000000
300000000
200000000
100000000
0
1658
1654
1650
1646
1642
1638
1634
1628
1315
1314
1313
1311
1310
1309
1308
1307
1306
1305
1304
1303
1302
1301
1300
1299
1298
1297
1296
1295
1294
1293
1292
1291
1290
1289
1288
1287
1286
1285
1284
1283
1282
1281

Friday, 22:00

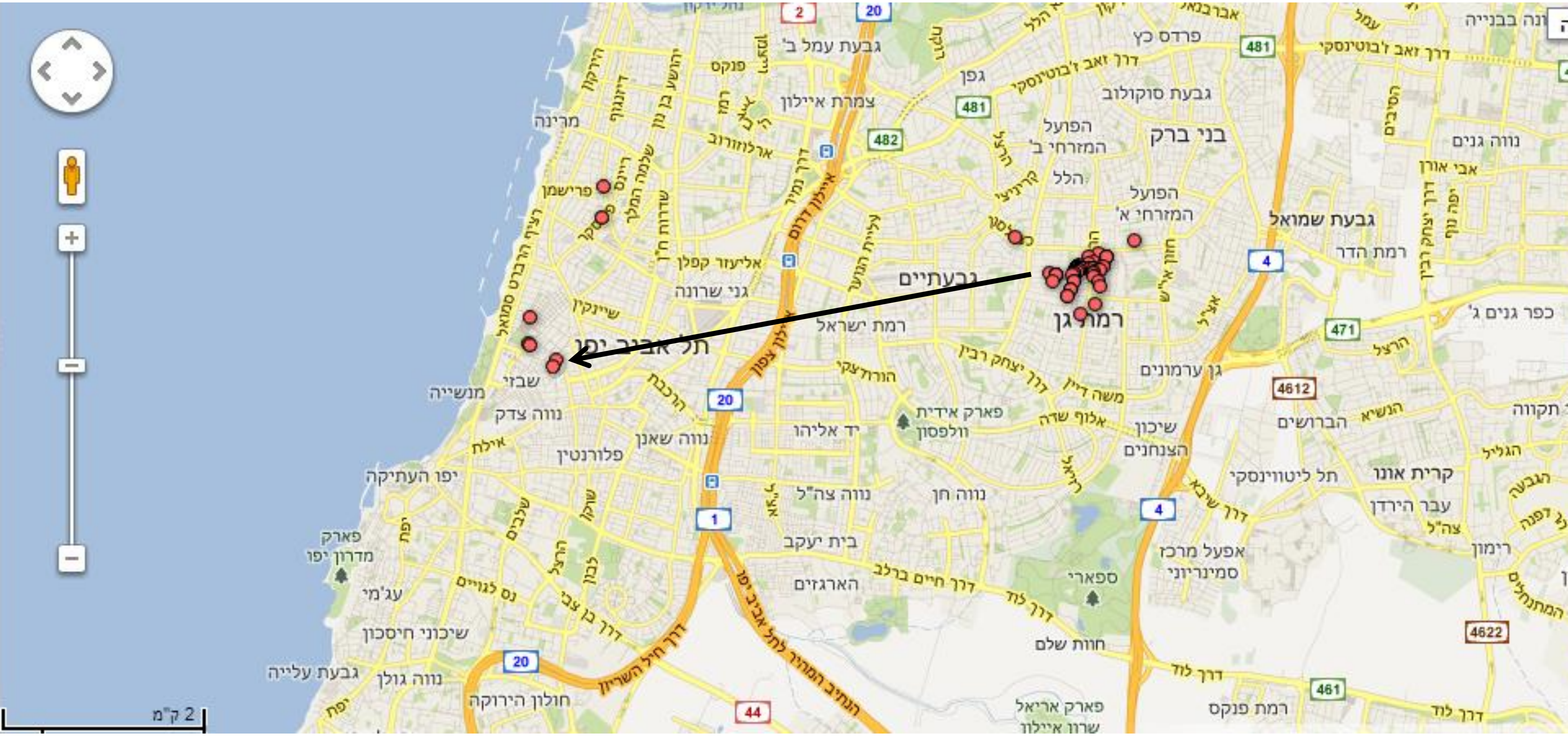

תל אביב יפו
רמת גן
גבעתיים
בני ברק
גבעת שמואל
קרית אונו
גבעת עמל ב'
צמרת איילון
פרדס כץ
גבעת סוקולוב
מרינה
גני שרונה
רמת ישראל
שבזי
מנשייה
נווה צדק
נווה שאנן
פלורנטין
יפו העתיקה
יד אליהו
פארק אידית וולפסון
נווה צה"ל
נווה חן
בית יעקב
הארגזים
ספארי
חוות שלם
רמת פנקס
גבעת עלייה
נווה גולן
חולון הירוקה
יפו ג' דרום
ע'ג'מי
שיכוני חיסכון
כפר גנים ג'
נווה גנים
רמת הדר
תל ליטווינסקי
עבר הירדן
גן ערמונים
הברושים
דרך חיים בר־לב
דרך לוד
איילון צפון
איילון דרום
אליעזר קפלן
2 ק"מ
1 מייל
©2013 Google - Map data ©2013 Mapa GISrael

Saturday, 03:00



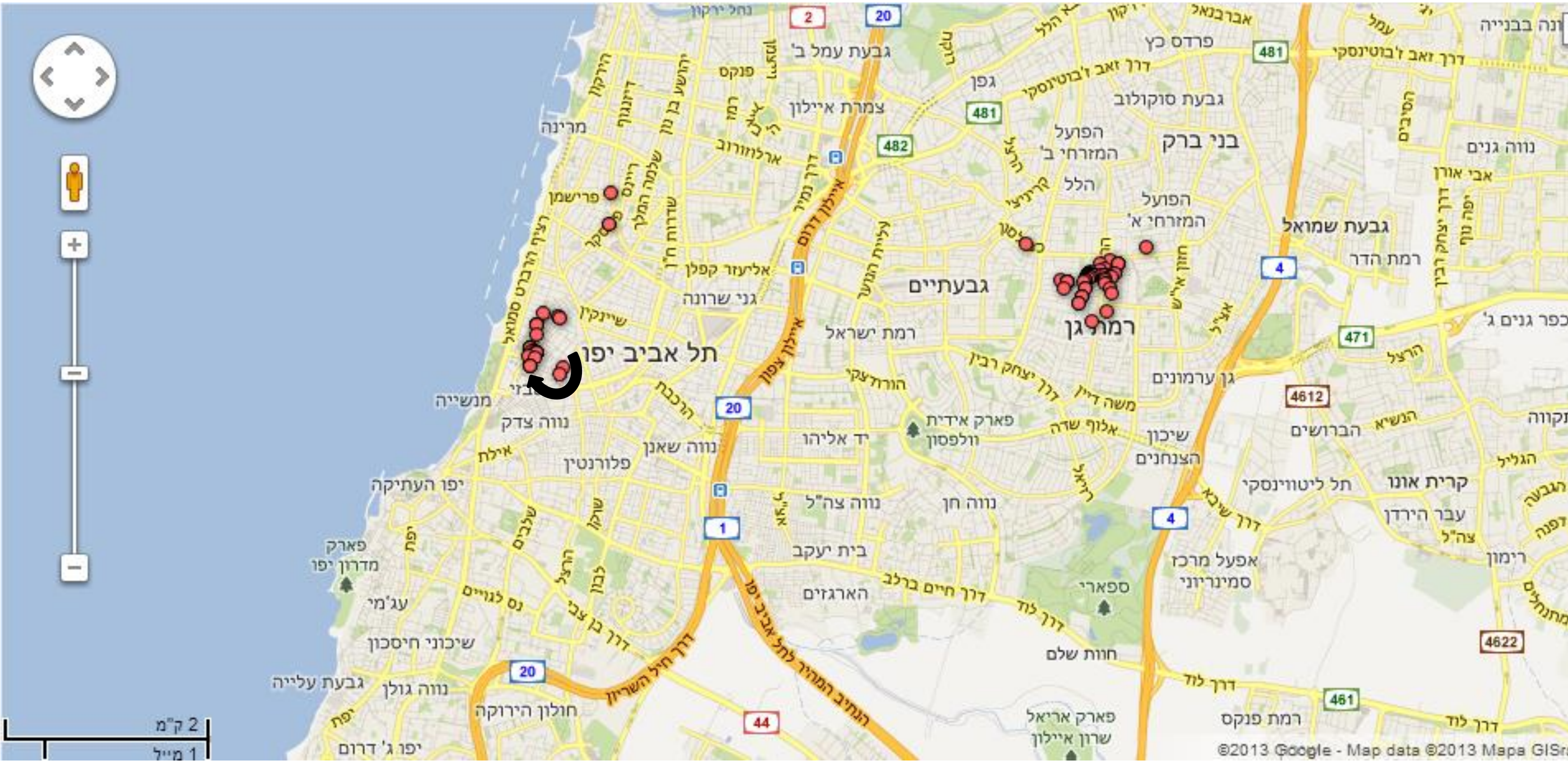
Saturday, 16:00

Sunday, 09:00



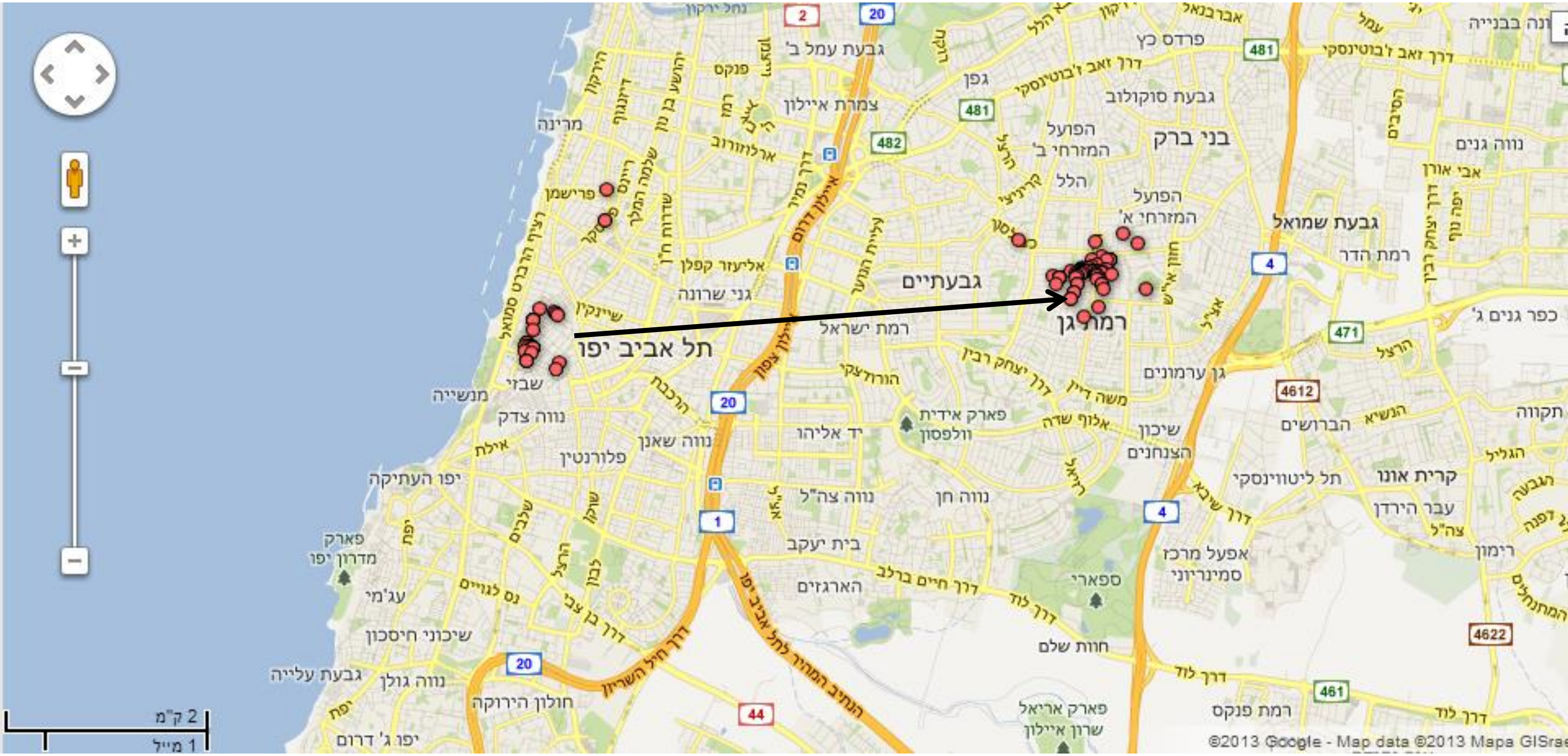
תל אביב יפו
גבעתיים
רמת גן
בני ברק
גבעת שמואל
קרית אונו
רמת ישראל
גני שרונה
יפו העתיקה
נווה צדק
פלורנטין
נווה שאנן
יד אליהו
פארק אידית וולפסון
בית יעקב
הארגזים
חוות שלם
ספארי
רמת פנקס
כפר גנים ג'
רמת הדר
נווה גנים
פרדס כץ
גבעת סוקולוב
גבעת עמל ב'
צמרת איילון
נווה חן
נווה צה"ל
גן ערמונים
רמת אביב
חולון הירוקה
נווה גולן
גבעת עלייה
שיכוני חיסכון
עג'מי
2 ק"מ

Sunday, 19:00

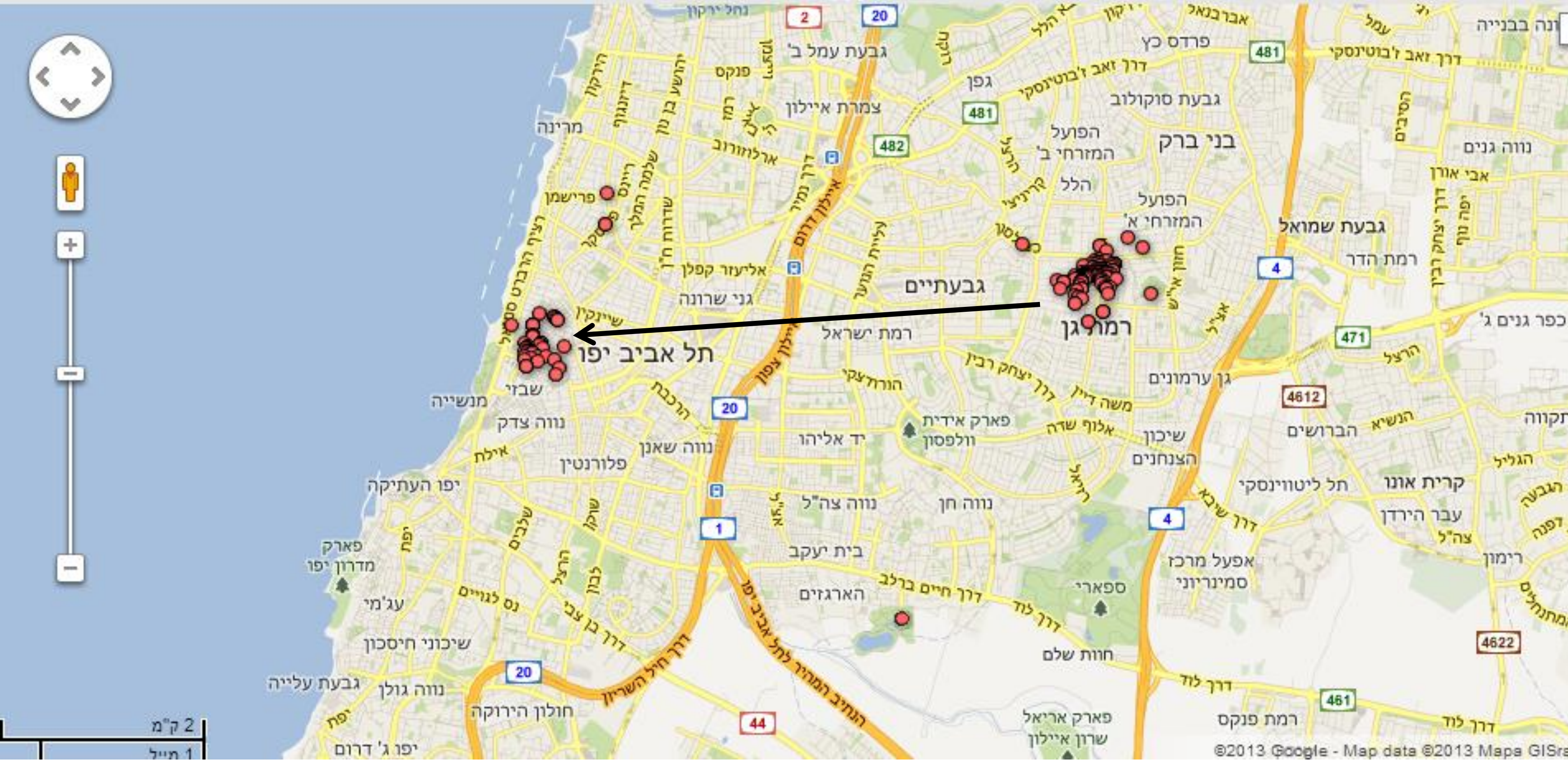


תל אביב יפו
רמת גן
גבעתיים
בני ברק
גבעת שמואל
קרית אונו
רמת ישראל
גבעת עמל ב'
צמרת איילון
פרדס כץ
גבעת סוקולוב
רמת הדר
כפר גנים ג'
נווה צדק
פלורנטין
יפו העתיקה
מנשייה
נווה שאנן
גני שרונה
מרינה
יד אליהו
ביצרון
רמת פנקס
חולון הירוקה
גבעת עלייה
שיכוני חיסכון
עג'מי
יפו ד' דרום
נווה גולן
נווה צה"ל
נווה חן
בית יעקב
הארגזים
חוות שלם
פארק אריאל שרון איילון
ספארי
אפעל מרכז סמינריוני
עבר הירדן
רימון
נווה גנים
אבי אורן
ג'ן ערמונים
שיכון הצנחנים
תל ליטווינסקי
הברושים
פארק אידית וולפסון
פארק מדרון יפו
©2013 Google - Map data ©2013 Mapa GISrael
2 ק"מ
1 מייל

Sunday, 22:00

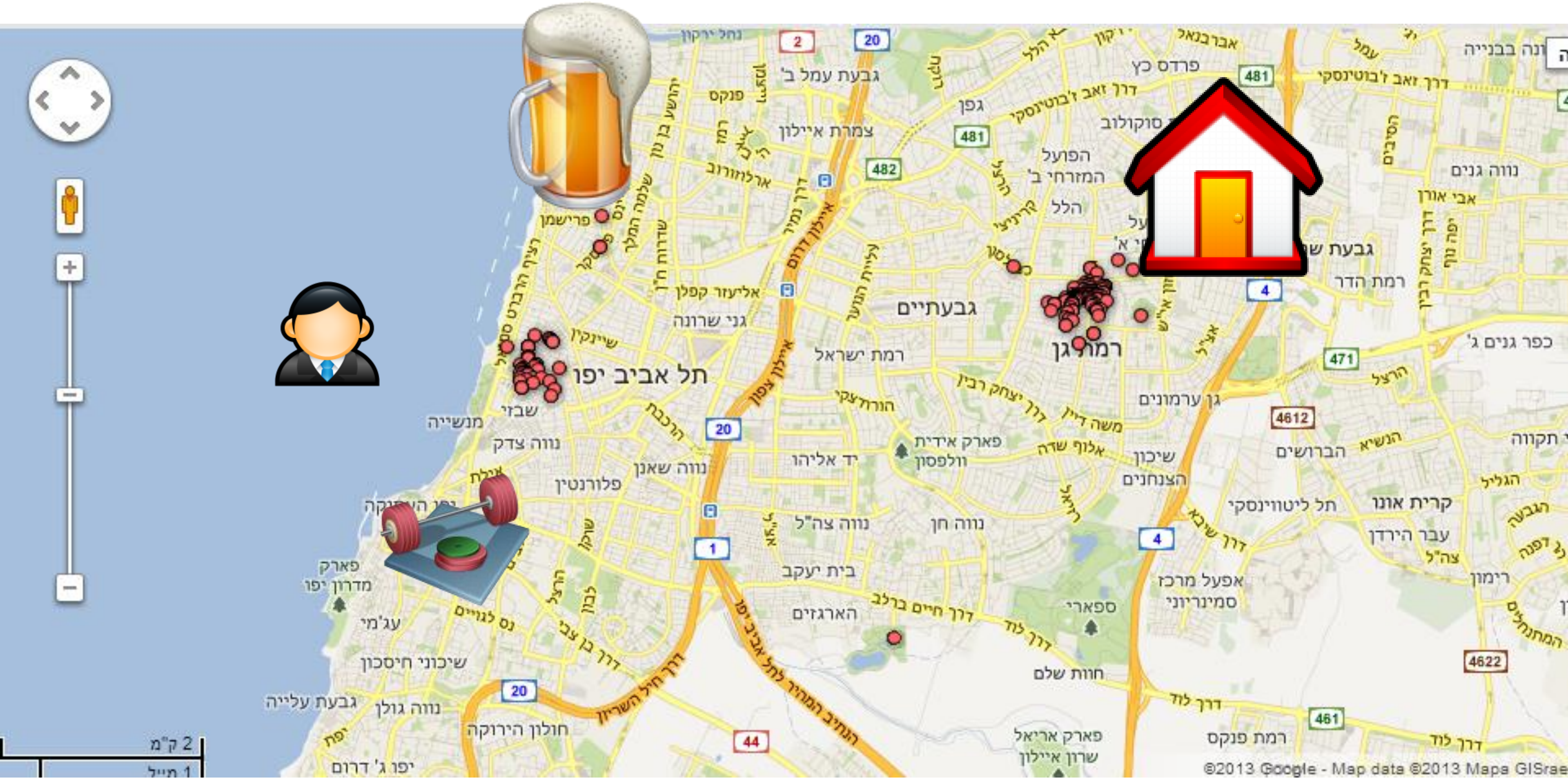

תל אביב יפו
רמת גן
גבעתיים
בני ברק
גבעת שמואל
קרית אונו
רמת ישראל
פרדס כץ
גבעת סוקולוב
רמת הדר
יפו העתיקה
מנשייה
נווה צדק
פלורנטין
נווה שאנן
יד אליהו
נווה חן
ביית יעקב
הארגזים
חוות שלם
רמת פנקס
2 ק"מ
1 מייל
©2013 Google - Map data ©2013 Mapa GISrae

Monday, 08:00

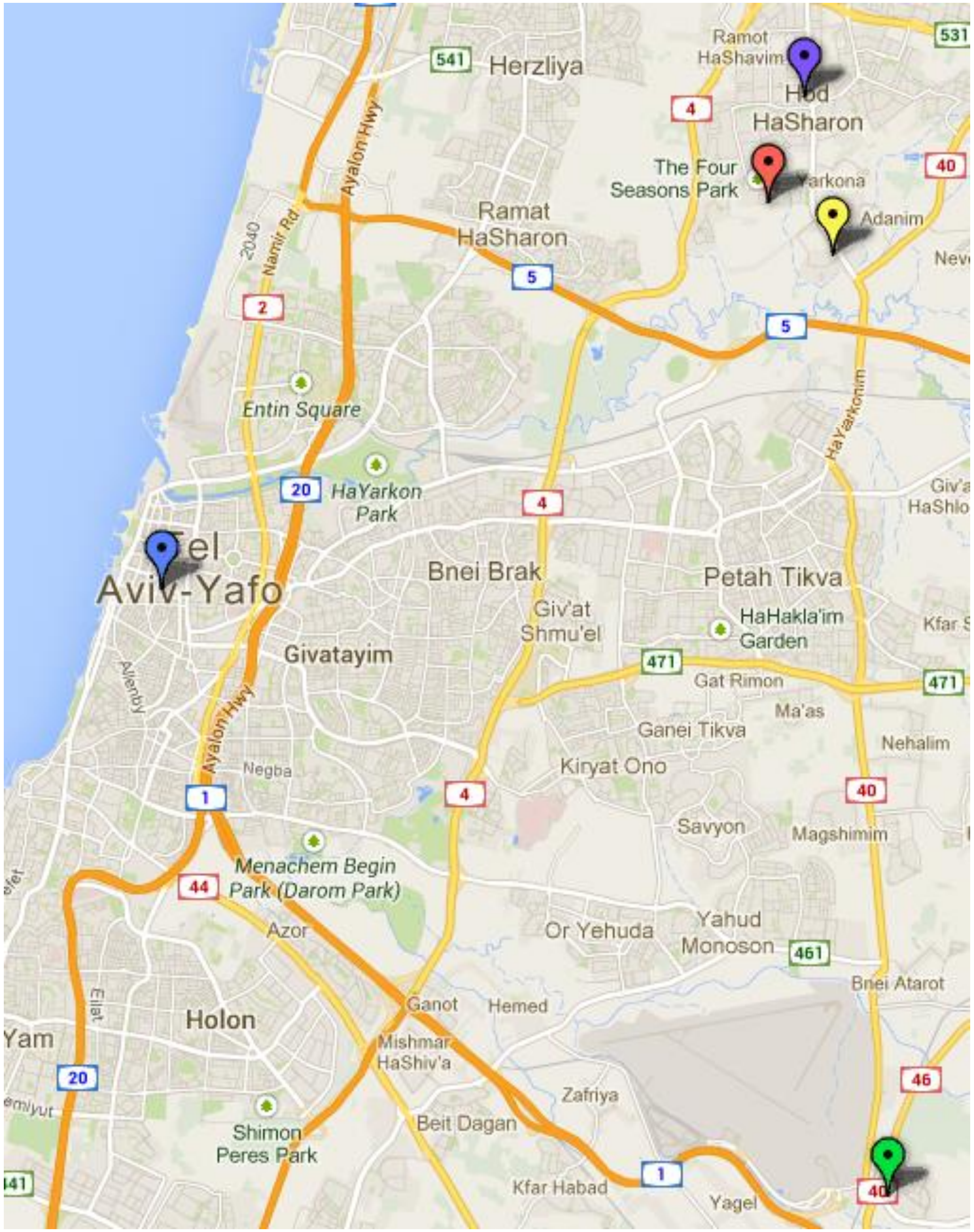


תל אביב יפו
רמת גן
גבעתיים
בני ברק
גבעת שמואל
קרית אונו
רמת ישראל
יפו העתיקה
נווה צדק
פלורנטין
שבזי
מנשייה
בית יעקב
הארגזים
חוות שלם
רמת פנקס
פרדס כץ
גבעת סוקולוב
צמרת איילון
גבעת עמל ב'
מרינה
גני שרונה
רמת הדר
כפר גנים ג'
גן ערמונים
נווה צה"ל
נווה חן
יד אליהו
נווה שאנן
עג'מי
שיכוני חיסכון
גבעת עלייה
נווה גולן
חולון הירוקה
יפו ג' דרום
ספארי
פארק אריאל שרון איילון
2 ק"מ
1 מייל
©2013 Google - Map data ©2013 Mapa GISrael

Conclusions up to now?



Semantic Information



E

A

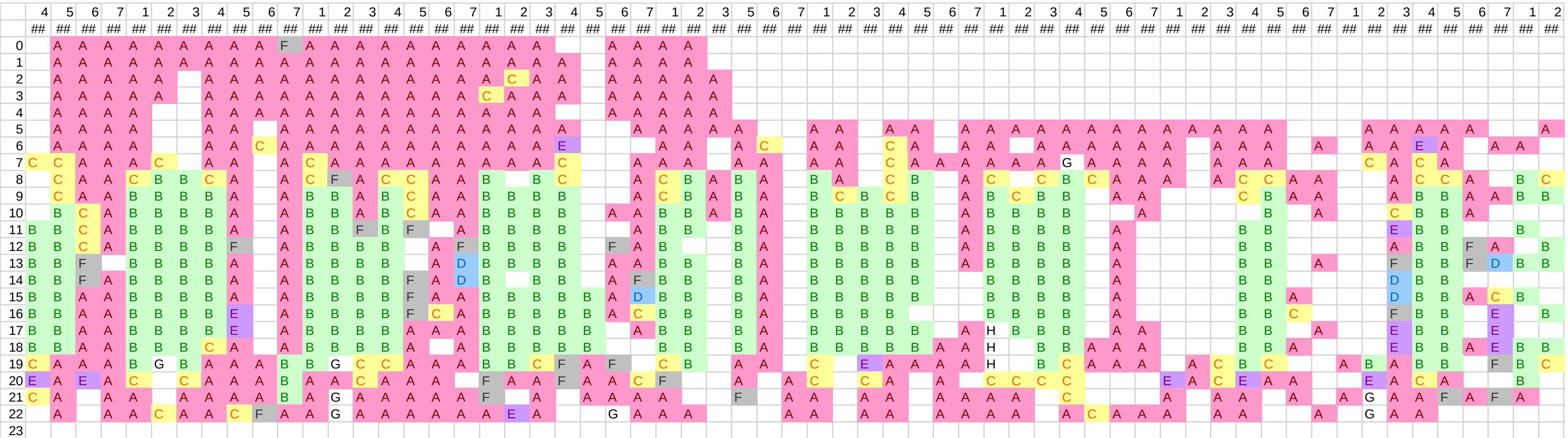
C

D

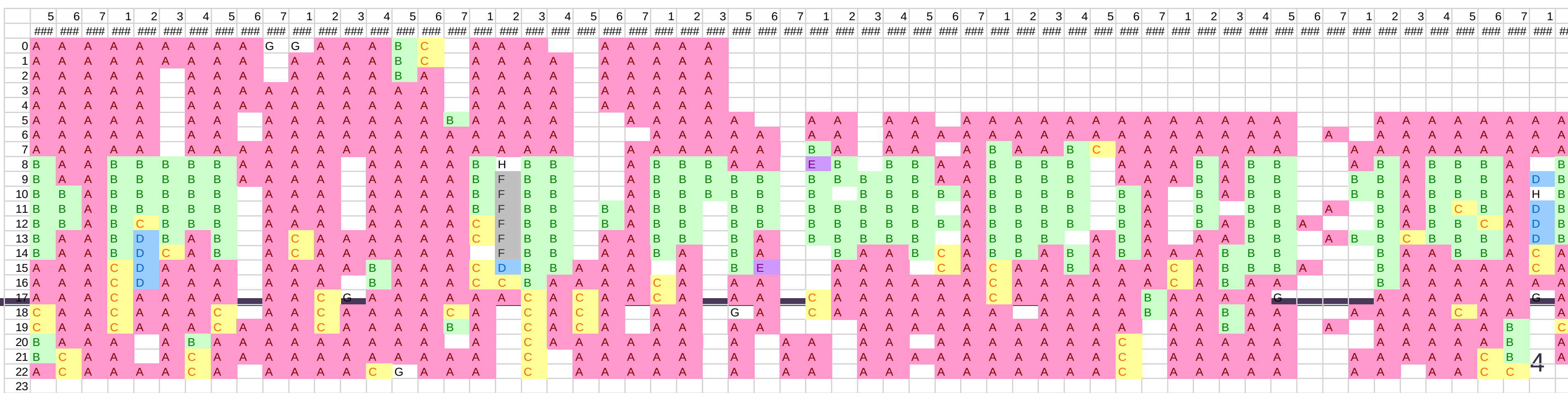
B

LMP

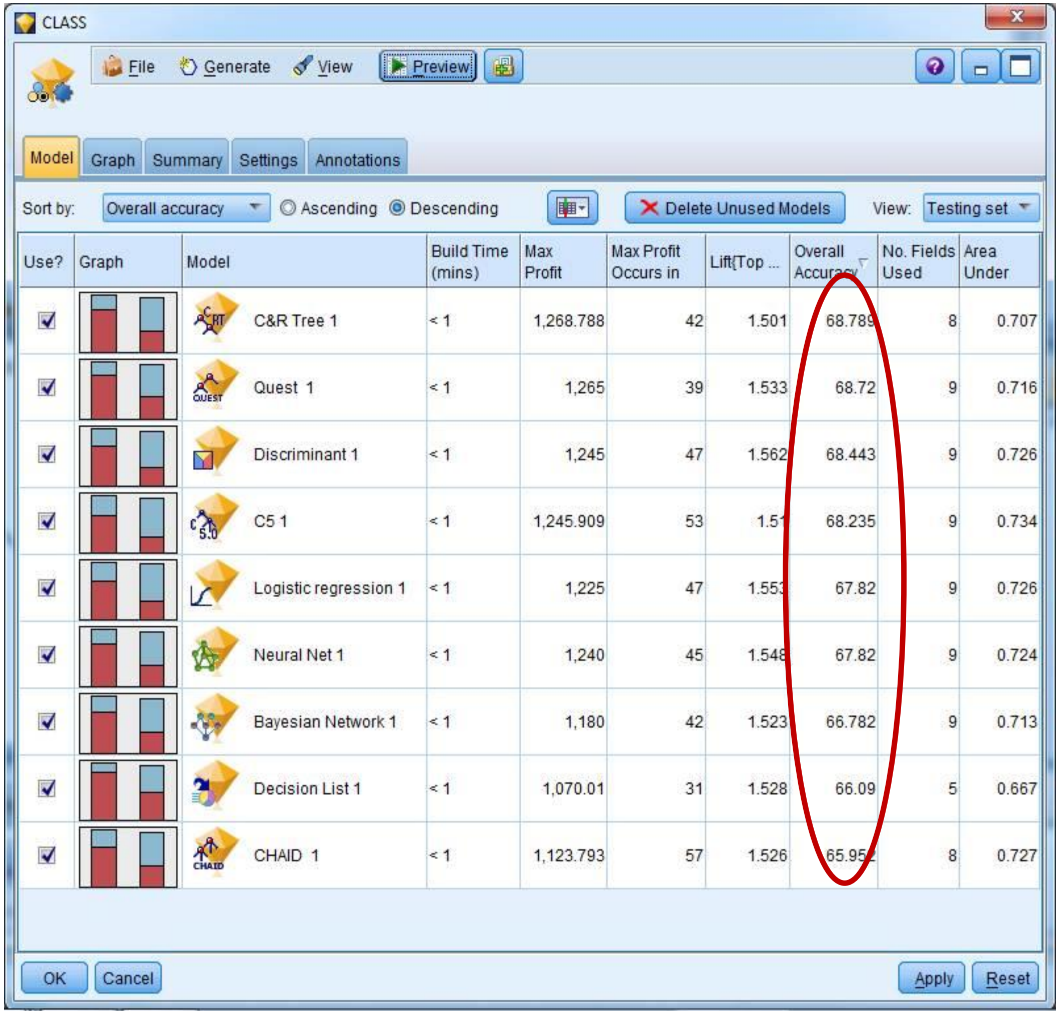
- User 3470



- User 3465



Customers Churn (benchmarked vs. SPSS/R)



CLASS

File Generate View Preview

Model | Graph | Summary | Settings | Annotations

Sort by: Overall accuracy ○ Ascending ◉ Descending | Delete Unused Models | View: Testing set

Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in	Lift{Top ...	Overall Accuracy	No. Fields Used	Area Under
☑		C&R Tree 1	< 1	1,268.788	42	1.501	68.789	8	0.707
☑		Quest 1	< 1	1,265	39	1.533	68.72	9	0.716
☑		Discriminant 1	< 1	1,245	47	1.562	68.443	9	0.726
☑		C5 1	< 1	1,245.909	53	1.51	68.235	9	0.734
☑		Logistic regression 1	< 1	1,225	47	1.553	67.82	9	0.726
☑		Neural Net 1	< 1	1,240	45	1.548	67.82	9	0.724
☑		Bayesian Network 1	< 1	1,180	42	1.523	66.782	9	0.713
☑		Decision List 1	< 1	1,070.01	31	1.528	66.09	5	0.667
☑		CHAID 1	< 1	1,123.793	57	1.526	65.95	8	0.727

OK Cancel Apply Reset

	R / SPSS	C-B4
Total Accuracy:	69.0%	77.8%
False Alarm:	30.3%	20.2%
False Positive	28.8%	18.5%
Churn Prediction	66.9%	74.1%
Engaged Prediction:	71.2%	81.5%

Gaining accuracy by using personalized granular modeling (per each user!)

LMP Algorithmic Infrastructure

1. Predicting personal and general mobility.

Predicting traffic jams before they start

2. Automatically recognizing places, their function and activities.

Understanding urban land-use, eg creating context-aware mobile apps

3. Modeling large-scale social patterns through their mobility footprint.

Mobility for disease spread

Mobility patterns in wars and crisis

Research: Suspects Tracking (leading Intl' Security Company)

Requirement >50% Recall with 1% False Alarm at most

TBNL: Bayesian Network

User Data

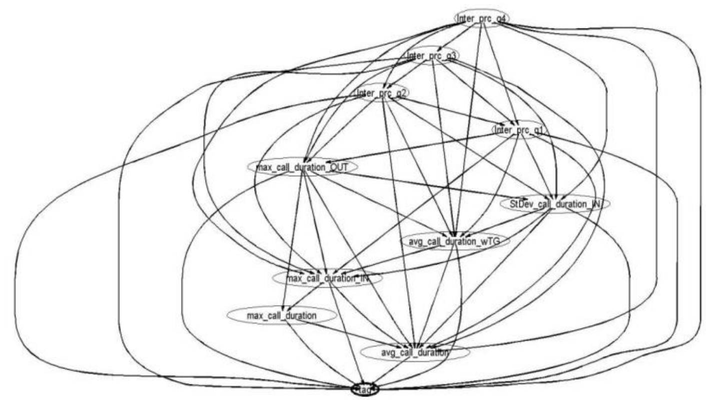
CDRs

Interactions

Others:

- Call/SMS ratio
- Time/Holidays
- Co-parties

Output



Network Data

(No text /content Mining)

Research: Suspects Tracking (leading Intl' Security Company)

User Data

CDRs

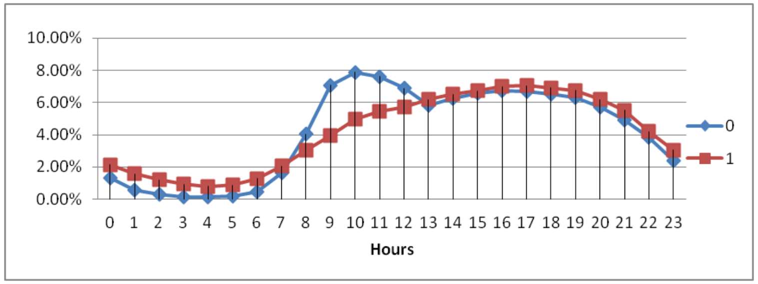
Interactions

Others:

- Call/SMS ratio
- Time/Holidays
- Co-parties

Output

Pattern Example:



Call/SMS ratio during the day

Suspects vs. Non-suspects

Research: Suspects Activities (using C-B4 SW)

CBC performance: >90% Recall with 1% False Alarm at most
(significantly better than all other ML models 60%)

User Data

CDRs

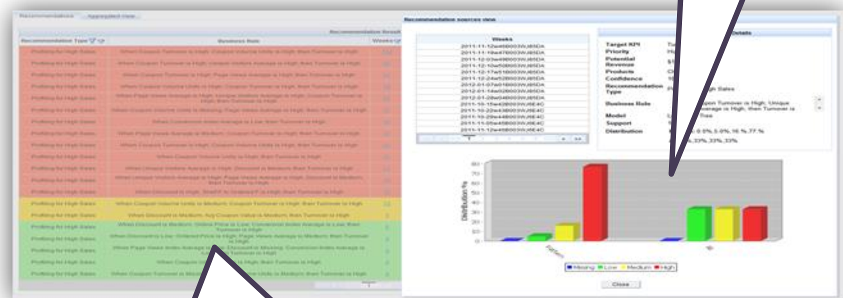
Interactions

Others:

- Call/SMS ratio
- Time/Holidays
- Co-parties

Output

Suspects' Distribution



Automated profiling & classification (in natural language)– e.g.:

- "When In-coming call duration is low (0-29sec) and sms to call ratio is high (>23) and fixed, mail activity hour (23:00-02:30); then likelihood of being suspect is high (>87%)"

Cellular Net Usage Monitoring

Cellcom Largest cellular company in Israel

Improved service Levels by 15% and saving millions of dollars

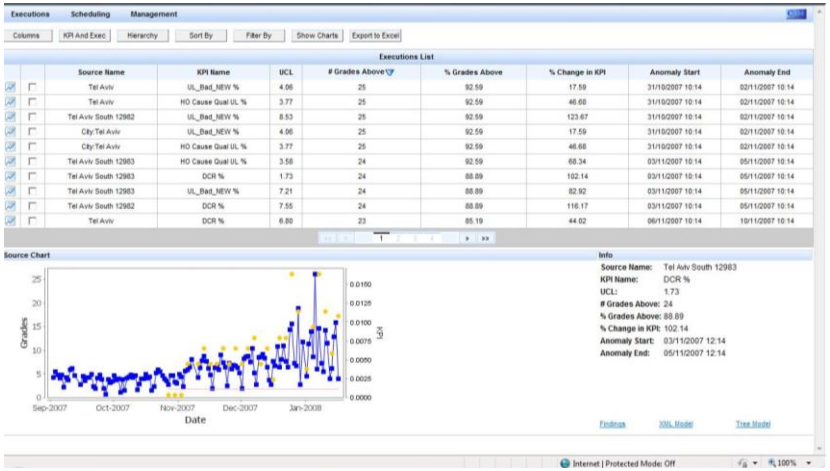
Network data

User Data

Other Data:

- Weather
- Dates
- Promotions
- Location …..

Output



Source Name	KPI Name	UCL	# Grades Above	% Grades Above	% Change in KPI	Anomaly Start	Anomaly End
Tel Aviv	UL_Bad_NEW %	4.06	25	92.59	17.59	31/10/2007 10:14	02/11/2007 10:14
Tel Aviv	HO Cause Qual UL %	3.77	25	92.59	46.68	31/10/2007 10:14	02/11/2007 10:14
Tel Aviv South 12982	UL_Bad_NEW %	8.53	25	92.59	123.67	31/10/2007 10:14	02/11/2007 10:14
City:Tel Aviv	UL_Bad_NEW %	4.06	25	92.59	17.59	31/10/2007 10:14	02/11/2007 10:14
City:Tel Aviv	HO Cause Qual UL %	3.77	25	92.59	46.68	31/10/2007 10:14	02/11/2007 10:14
Tel Aviv South 12983	HO Cause Qual UL %	3.58	24	92.59	68.34	03/11/2007 10:14	05/11/2007 10:14
Tel Aviv South 12983	DCR %	1.73	24	88.89	102.14	03/11/2007 10:14	05/11/2007 10:14
Tel Aviv South 12983	UL_Bad_NEW %	7.21	24	88.89	82.92	03/11/2007 10:14	05/11/2007 10:14
Tel Aviv South 12982	DCR %	7.55	24	88.89	116.17	03/11/2007 10:14	05/11/2007 10:14
Tel Aviv	DCR %	6.80	23	85.19	44.02	06/11/2007 10:14	10/11/2007 10:14

- Early alerts
- Anomaly detection
- New demand/usage patterns
- Root cause analysis

Predictive Maintenance and Usage Monitoring

Cellcom Largest Cellular company in Israel

Executions Scheduling Management

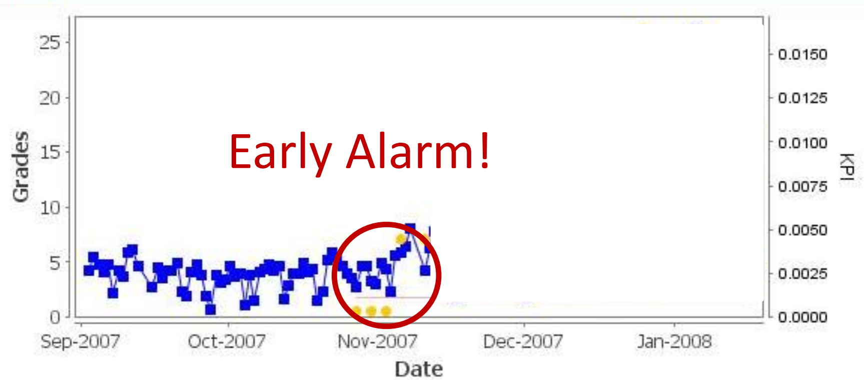
Columns | KPI And Exec | Hierarchy | Sort By | Filter By | Show Charts | Export to Excel

Executions List

		Source Name	KPI Name	UCL	# Grades Above	% Grades Above	% Change in KPI	Anomaly Start	Anomaly End
		Tel Aviv	UL_Bad_NEW %	4.06	25	92.59	17.59	31/10/2007 10:14	02/11/2007 10:14
		Tel Aviv	HO Cause Qual UL %	3.77	25	92.59	46.68	31/10/2007 10:14	02/11/2007 10:14
		Tel Aviv South 12982	UL_Bad_NEW %	8.53	25	92.59	123.67	31/10/2007 10:14	02/11/2007 10:14
		City:Tel Aviv	UL_Bad_NEW %	4.06	25	92.59	17.59	31/10/2007 10:14	02/11/2007 10:14
		City:Tel Aviv	HO Cause Qual UL %	3.77	25	92.59	46.68	31/10/2007 10:14	02/11/2007 10:14
		Tel Aviv South 12983	HO Cause Qual UL %	3.58	24	92.59	68.34	03/11/2007 10:14	05/11/2007 10:14
		Tel Aviv South 12983	DCR %	1.73	24	88.89	102.14	03/11/2007 10:14	05/11/2007 10:14
		Tel Aviv South 12983	UL_Bad_NEW %	7.21	24	88.89	82.92	03/11/2007 10:14	05/11/2007 10:14
		Tel Aviv South 12982	DCR %	7.55	24	88.89	116.17	03/11/2007 10:14	05/11/2007 10:14
		Tel Aviv	DCR %	6.80	23	85.19	44.02	06/11/2007 10:14	10/11/2007 10:14

« ‹ 1 2 3 4 › »

Source Chart



Info

Source Name: Tel Aviv South 12983
KPI Name: DCR %
UCL: 1.73
 # Grades Above: 24
% Grades Above: 88.89
% Change in KPI: 102.14
Anomaly Start: 03/11/2007 12:14
Anomaly End: 05/11/2007 12:14

Findings XML Model Tree Model

Predictive Maintenance and Usage Monitoring

Cellcom Largest Cellular company in Israel

Executions Scheduling Management

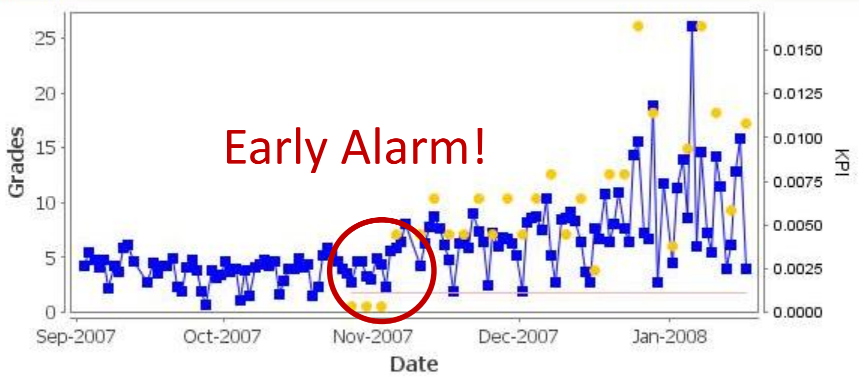
Columns | KPI And Exec | Hierarchy | Sort By | Filter By | Show Charts | Export to Excel

Executions List

Source Name	KPI Name	UCL	# Grades Above	% Grades Above	% Change in KPI	Anomaly Start	Anomaly End
Tel Aviv	UL_Bad_NEW %	4.06	25	92.59	17.59	31/10/2007 10:14	02/11/2007 10:14
Tel Aviv	HO Cause Qual UL %	3.77	25	92.59	46.68	31/10/2007 10:14	02/11/2007 10:14
Tel Aviv South 12982	UL_Bad_NEW %	8.53	25	92.59	123.67	31/10/2007 10:14	02/11/2007 10:14
City:Tel Aviv	UL_Bad_NEW %	4.06	25	92.59	17.59	31/10/2007 10:14	02/11/2007 10:14
City:Tel Aviv	HO Cause Qual UL %	3.77	25	92.59	46.68	31/10/2007 10:14	02/11/2007 10:14
Tel Aviv South 12983	HO Cause Qual UL %	3.58	24	92.59	68.34	03/11/2007 10:14	05/11/2007 10:14
Tel Aviv South 12983	DCR %	1.73	24	88.89	102.14	03/11/2007 10:14	05/11/2007 10:14
Tel Aviv South 12983	UL_Bad_NEW %	7.21	24	88.89	82.92	03/11/2007 10:14	05/11/2007 10:14
Tel Aviv South 12982	DCR %	7.55	24	88.89	116.17	03/11/2007 10:14	05/11/2007 10:14
Tel Aviv	DCR %	6.80	23	85.19	44.02	06/11/2007 10:14	10/11/2007 10:14

« ‹ 1 2 3 4 › »

Source Chart



Info

- **Source Name:** Tel Aviv South 12983
- **KPI Name:** DCR %
- **UCL:** 1.73
- **# Grades Above:** 24
- **% Grades Above:** 88.89
- **% Change in KPI:** 102.14
- **Anomaly Start:** 03/11/2007 12:14
- **Anomaly End:** 05/11/2007 12:14

Findings XML Model Tree Model

Thank You