



Büyük Veri Analitiği İnovasyon ve Karar Almada Yeni Hudutlar

Y. Doç. Dr. Şeyda Ertekin

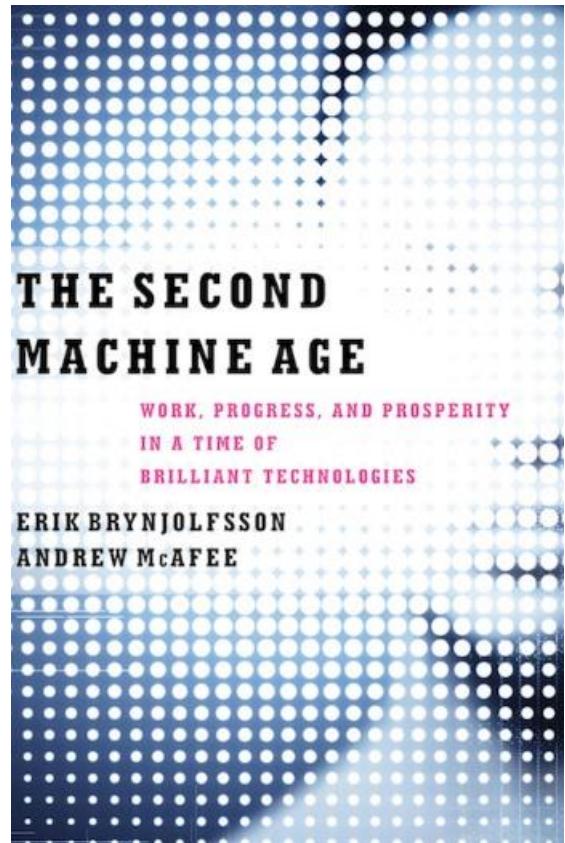


Bilgisayar Mühendisliği
Orta Doğu Teknik Üniversitesi (ODTÜ)
{ seyda@ceng.metu.edu.tr }



Massachusetts Institute of Technology (MIT)
MIT Sloan School of Management
Operations Research and Statistics Group
{ seyda@mit.edu }

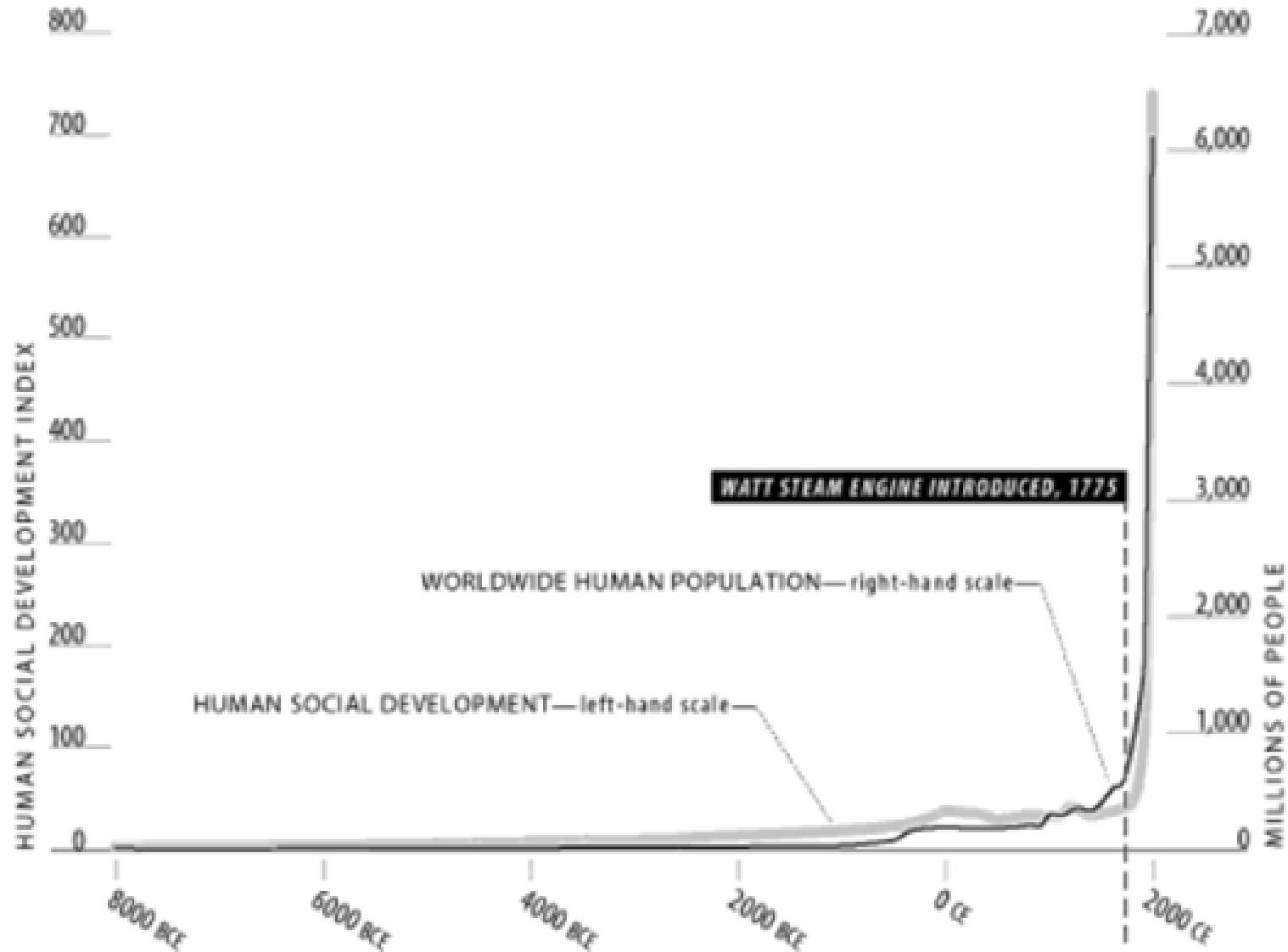
 : seydaertekin



Erik Brynjolfsson

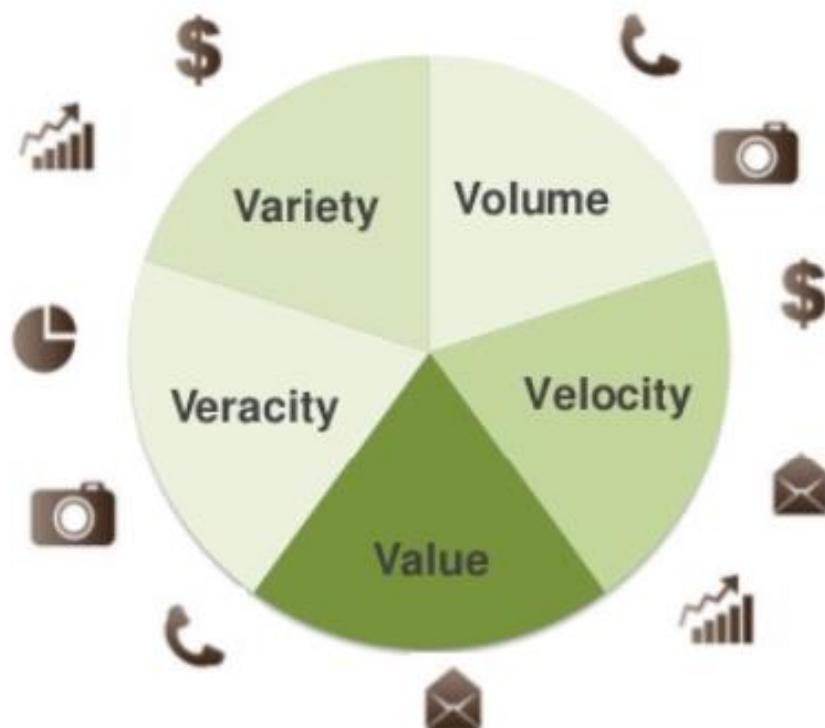
Andrew McAfee

FIGURE 1.2 What Bent the Curve of Human History? The Industrial Revolution.



5 V's of Big Data

To get a better understanding of what Big Data is, it is often described using Five Vs:



Volume

We see increasing volumes of data, that grow at exponential rates:



Volume refers to the vast amounts of data generated every second. We are not talking Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute. This makes most data sets too large to store and analyse using traditional database technology. New big data tools use distributed systems so that we can store and analyse data across databases that are dotted around anywhere in the world.

Velocity

We see increasing velocity (or speed) at which data changes, travels or increases



Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology allows us now to analyse the data while it is being generated (sometimes referred to as in-memory analytics), without ever putting it into databases.

Variety

We see increasing variety of data types:



Variety refers to the different types of data we can now use. In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of the world's data is unstructured (text, images, video, voice, etc.) With big data technology we can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.

Veracity

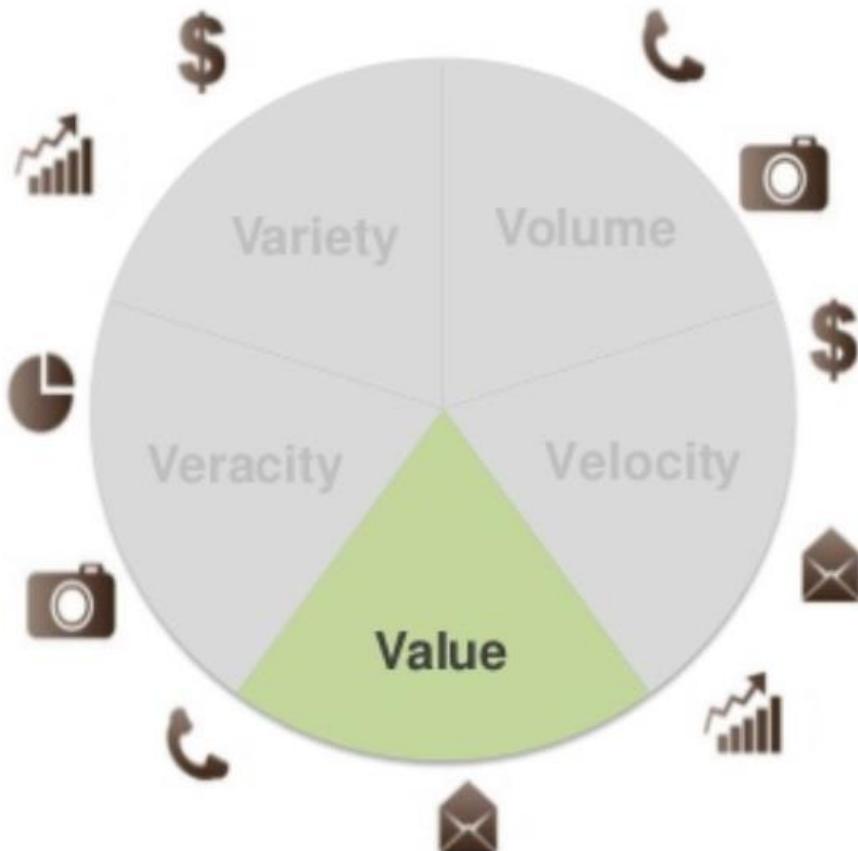
We see increasing veracity (or accuracy) of data



Veracity refers to the messiness or trustworthiness of the data. With many forms of big data quality and accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but technology now allows us to work with this type of data.

Value

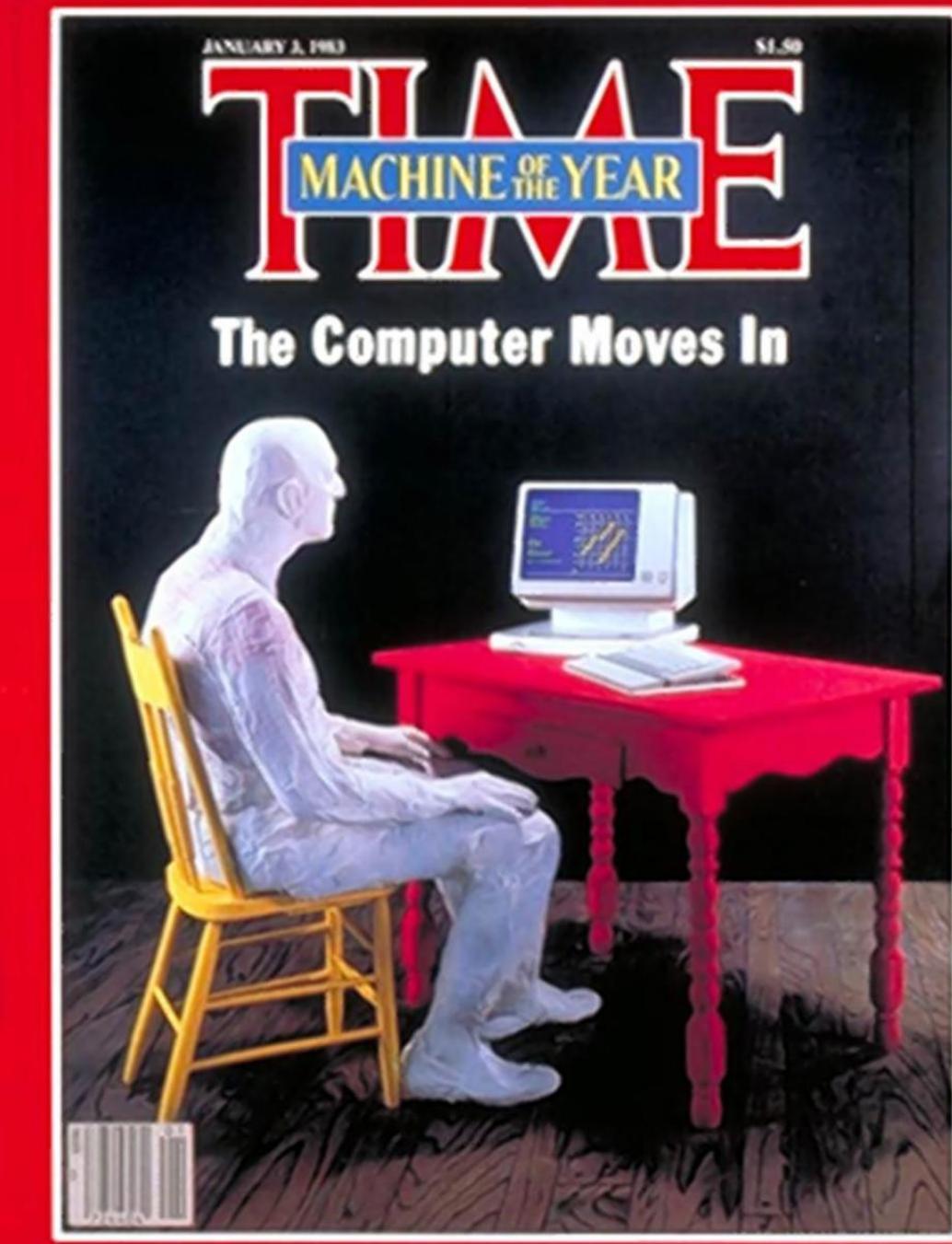
Value – The most important V of all!

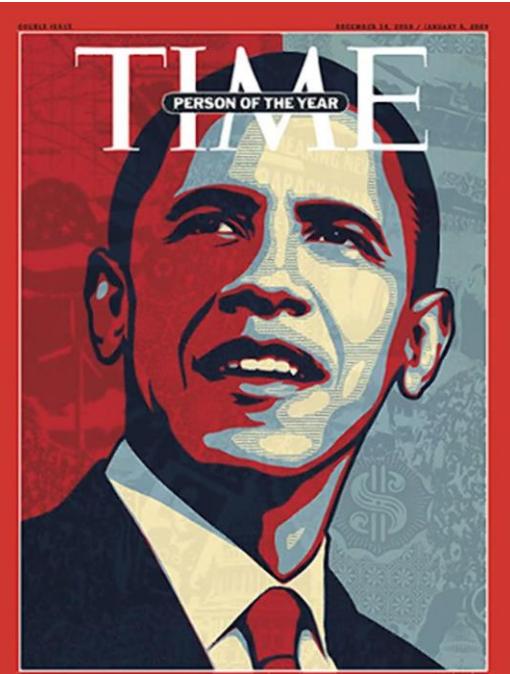
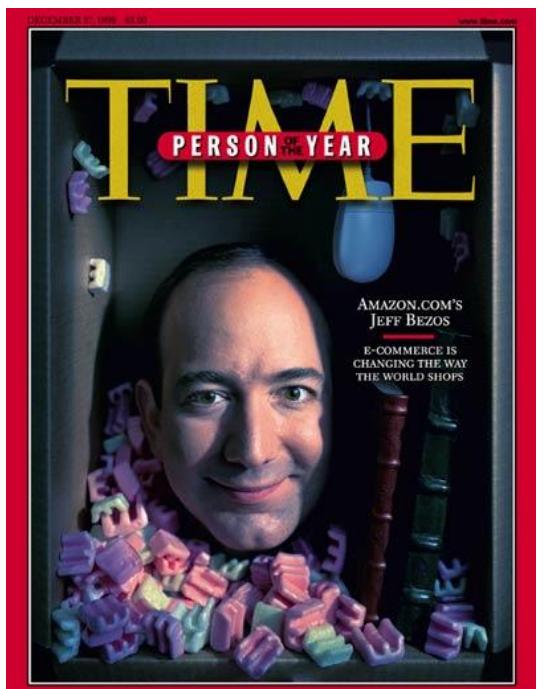
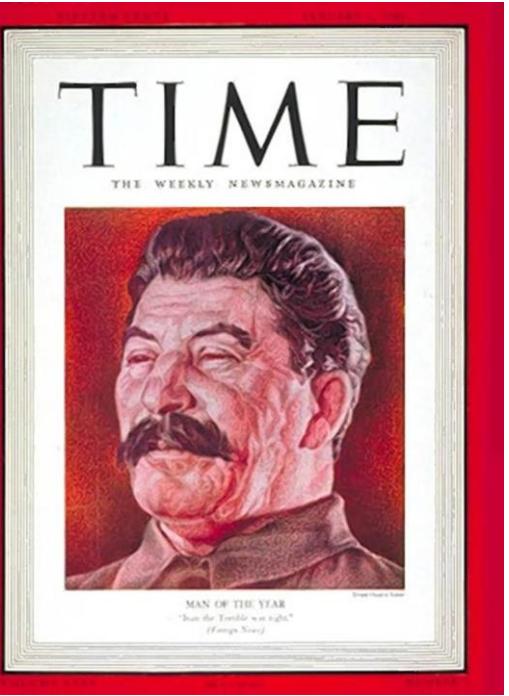
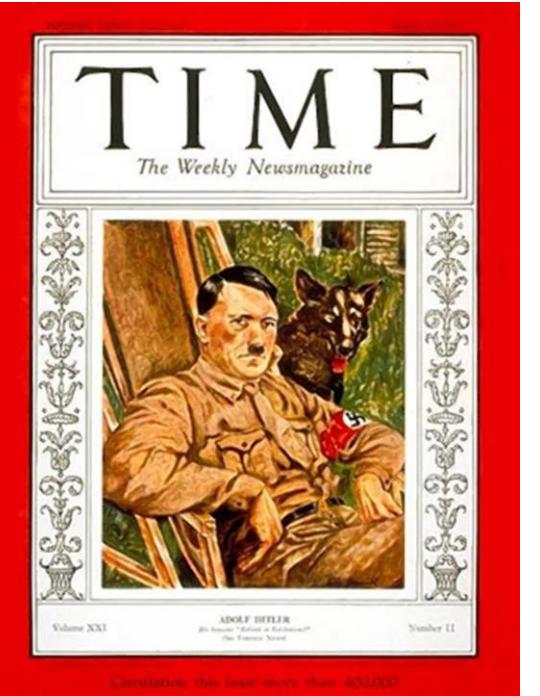
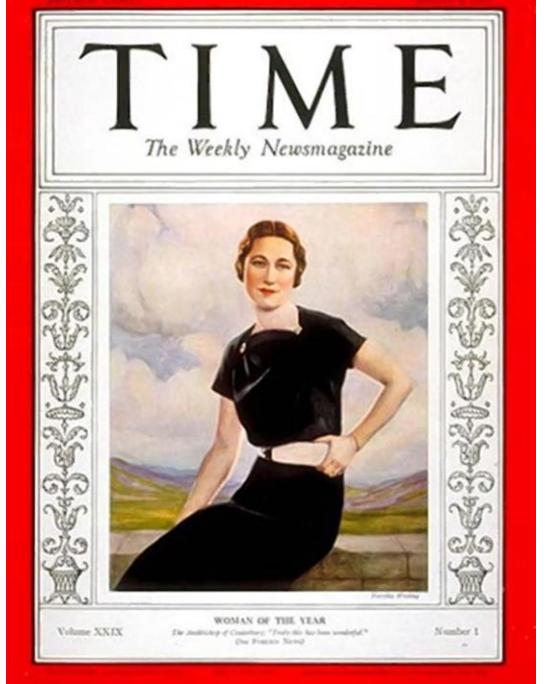


Then there is another V to take into account when looking at Big Data: Value!

Having access to big data is no good unless we can turn it into value.

Companies are starting to generate amazing value from their big data.



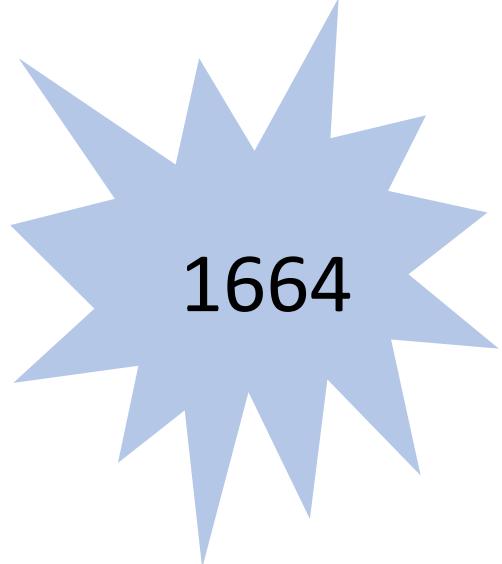


2006



Content is created by

- Employers
- Users
- Machines



1664

Bills of Mortality

- First known document that shows the spread of a human disease
 - 7165 deaths during a single week in late August by plague.
 - For the first time it became clear that public health could be
 - charted
 - mathematical
 - predicted

A generall Bill for this present year, ending the 19 of December 1665, according to the Report made to the KINGs most Excellent Majestie.					
By the Company of Parish Clerks of London, &c.					
Item	No.	Item	No.	Item	No.
1st Alkales Woodlark	163	1' Chamer Saltchape	16	1" Margaret Mole	15
2d Alkales Earling	124	2' Dame Saltchape	17	2' Margaret Newell	14
3d Alkales Rould	125	3' Distillor Calmado	18	3' Margaret Parson	15
4d Alkales Grouse	116	4' Edmonds Lumbard	16	4' Margaret Parson	17
5d Alkales Hen	151	5' Elkhalfebrugh	19	5' Margaret Alecham	9
6d Alkales Lark	10	6' Falc	10	6' Margaret Alchamber	12
7d Alkales Lett	119	7' Falster	11	7' Margaret Alchamber	10
8d Alkale Lambeth	51	8' Falster	12	8' Margaret Alchamber	11
9d Alkales Stoen	131	9' Gabriel Fren balyng	13	9' Maple Bow	6
10d Alkales Wren	210	10' George Boulphol	14	10' May Bawle	13
11d My Alphage	123	11' Gergonne by Pauls	25	11' May Bawle	14
12d Andrew Holles	71	12' Holles	18	12' May Bawle	15
13d Andrew Wadenshal	174	13' Janet Dales place	16	13' May Bawle	16
14d Andrew Wadenshal	270	14' James Galdehouse	19	14' May Bawle	17
15d Anne Blakke Toone	131	15' John Bayly	11	15' May Bawle	18
16d Anthonie Farre	51	16' John Bourgall	9	16' May Bawle	19
17d Andis Fauell	41	17' John Eathorne	11	17' May Bawle	20
18d Bawle Ranchane	71	18' Katherin Colman	19	18' May Bawle	21
19d Bawle Ranchane	71	19' Katherin Clarke	15	19' May Bawle	22
20d Bawle Ranchane	71	20' Lawrence Jervis	24	20' May Bawle	23
21d Bawle Ranchane	71	21' Lawrence Prount	14	21' May Bawle	24
22d Bawle Ranchane	71	22' Leonard Edelstone	13	22' May Bawle	25
23d Bawle Ranchane	71	23' Leoued Tidderstone	15	23' May Bawle	26
24d Bawle Ranchane	71	24' Margaret Parcell	19	24' May Bawle	27
25d Bawle Ranchane	71	25' Margaret Parcell	16	25' May Bawle	28
26d Bawle Ranchane	71	26' Margaret Lushbury	10	26' May Bawle	29
27d Bawle Ranchane	71	27' Margaret Lushbury	10	27' May Bawle	30
28d Bawle Ranchane	71	28' Michael Baillie	15	28' May Bawle	31
29d Bawle Ranchane	71	29' Michael Baillie	15	29' May Bawle	32
30d Bawle Ranchane	71	30' Michael Baillie	15	30' May Bawle	33
31d Bawle Ranchane	71	31' Michael Baillie	15	31' May Bawle	34
32d Bawle Ranchane	71	32' Michael Baillie	15	32' May Bawle	35
33d Bawle Ranchane	71	33' Michael Baillie	15	33' May Bawle	36
34d Bawle Ranchane	71	34' Michael Baillie	15	34' May Bawle	37
35d Bawle Ranchane	71	35' Michael Baillie	15	35' May Bawle	38
36d Bawle Ranchane	71	36' Michael Baillie	15	36' May Bawle	39
37d Bawle Ranchane	71	37' Michael Baillie	15	37' May Bawle	40
38d Bawle Ranchane	71	38' Michael Baillie	15	38' May Bawle	41
39d Bawle Ranchane	71	39' Michael Baillie	15	39' May Bawle	42
40d Bawle Ranchane	71	40' Michael Baillie	15	40' May Bawle	43
41d Bawle Ranchane	71	41' Michael Baillie	15	41' May Bawle	44
42d Bawle Ranchane	71	42' Michael Baillie	15	42' May Bawle	45
43d Bawle Ranchane	71	43' Michael Baillie	15	43' May Bawle	46
44d Bawle Ranchane	71	44' Michael Baillie	15	44' May Bawle	47
45d Bawle Ranchane	71	45' Michael Baillie	15	45' May Bawle	48
46d Bawle Ranchane	71	46' Michael Baillie	15	46' May Bawle	49
47d Bawle Ranchane	71	47' Michael Baillie	15	47' May Bawle	50
48d Bawle Ranchane	71	48' Michael Baillie	15	48' May Bawle	51
49d Bawle Ranchane	71	49' Michael Baillie	15	49' May Bawle	52
50d Bawle Ranchane	71	50' Michael Baillie	15	50' May Bawle	53
51d Bawle Ranchane	71	51' Michael Baillie	15	51' May Bawle	54
52d Bawle Ranchane	71	52' Michael Baillie	15	52' May Bawle	55
53d Bawle Ranchane	71	53' Michael Baillie	15	53' May Bawle	56
54d Bawle Ranchane	71	54' Michael Baillie	15	54' May Bawle	57
55d Bawle Ranchane	71	55' Michael Baillie	15	55' May Bawle	58
56d Bawle Ranchane	71	56' Michael Baillie	15	56' May Bawle	59
57d Bawle Ranchane	71	57' Michael Baillie	15	57' May Bawle	60
58d Bawle Ranchane	71	58' Michael Baillie	15	58' May Bawle	61
59d Bawle Ranchane	71	59' Michael Baillie	15	59' May Bawle	62
60d Bawle Ranchane	71	60' Michael Baillie	15	60' May Bawle	63
61d Bawle Ranchane	71	61' Michael Baillie	15	61' May Bawle	64
62d Bawle Ranchane	71	62' Michael Baillie	15	62' May Bawle	65
63d Bawle Ranchane	71	63' Michael Baillie	15	63' May Bawle	66
64d Bawle Ranchane	71	64' Michael Baillie	15	64' May Bawle	67
65d Bawle Ranchane	71	65' Michael Baillie	15	65' May Bawle	68
66d Bawle Ranchane	71	66' Michael Baillie	15	66' May Bawle	69
67d Bawle Ranchane	71	67' Michael Baillie	15	67' May Bawle	70
68d Bawle Ranchane	71	68' Michael Baillie	15	68' May Bawle	71
69d Bawle Ranchane	71	69' Michael Baillie	15	69' May Bawle	72
70d Bawle Ranchane	71	70' Michael Baillie	15	70' May Bawle	73
71d Bawle Ranchane	71	71' Michael Baillie	15	71' May Bawle	74
72d Bawle Ranchane	71	72' Michael Baillie	15	72' May Bawle	75
73d Bawle Ranchane	71	73' Michael Baillie	15	73' May Bawle	76
74d Bawle Ranchane	71	74' Michael Baillie	15	74' May Bawle	77
75d Bawle Ranchane	71	75' Michael Baillie	15	75' May Bawle	78
76d Bawle Ranchane	71	76' Michael Baillie	15	76' May Bawle	79
77d Bawle Ranchane	71	77' Michael Baillie	15	77' May Bawle	80
78d Bawle Ranchane	71	78' Michael Baillie	15	78' May Bawle	81
79d Bawle Ranchane	71	79' Michael Baillie	15	79' May Bawle	82
80d Bawle Ranchane	71	80' Michael Baillie	15	80' May Bawle	83
81d Bawle Ranchane	71	81' Michael Baillie	15	81' May Bawle	84
82d Bawle Ranchane	71	82' Michael Baillie	15	82' May Bawle	85
83d Bawle Ranchane	71	83' Michael Baillie	15	83' May Bawle	86
84d Bawle Ranchane	71	84' Michael Baillie	15	84' May Bawle	87
85d Bawle Ranchane	71	85' Michael Baillie	15	85' May Bawle	88
86d Bawle Ranchane	71	86' Michael Baillie	15	86' May Bawle	89
87d Bawle Ranchane	71	87' Michael Baillie	15	87' May Bawle	90
88d Bawle Ranchane	71	88' Michael Baillie	15	88' May Bawle	91
89d Bawle Ranchane	71	89' Michael Baillie	15	89' May Bawle	92
90d Bawle Ranchane	71	90' Michael Baillie	15	90' May Bawle	93
91d Bawle Ranchane	71	91' Michael Baillie	15	91' May Bawle	94
92d Bawle Ranchane	71	93' Michael Baillie	15	93' May Bawle	95
93d Bawle Ranchane	71	94' Michael Baillie	15	94' May Bawle	96
94d Bawle Ranchane	71	95' Michael Baillie	15	95' May Bawle	97
95d Bawle Ranchane	71	96' Michael Baillie	15	96' May Bawle	98
96d Bawle Ranchane	71	97' Michael Baillie	15	97' May Bawle	99
97d Bawle Ranchane	71	98' Michael Baillie	15	98' May Bawle	100
98d Bawle Ranchane	71	99' Michael Baillie	15	99' May Bawle	101
99d Bawle Ranchane	71	100' Michael Baillie	15	100' May Bawle	102
100d Bawle Ranchane	71	101' Michael Baillie	15	101' May Bawle	103
101d Bawle Ranchane	71	102' Michael Baillie	15	102' May Bawle	104
102d Bawle Ranchane	71	103' Michael Baillie	15	103' May Bawle	105
103d Bawle Ranchane	71	104' Michael Baillie	15	104' May Bawle	106
104d Bawle Ranchane	71	105' Michael Baillie	15	105' May Bawle	107
105d Bawle Ranchane	71	106' Michael Baillie	15	106' May Bawle	108
106d Bawle Ranchane	71	107' Michael Baillie	15	107' May Bawle	109
107d Bawle Ranchane	71	108' Michael Baillie	15	108' May Bawle	110
108d Bawle Ranchane	71	109' Michael Baillie	15	109' May Bawle	111
109d Bawle Ranchane	71	110' Michael Baillie	15	110' May Bawle	112
110d Bawle Ranchane	71	111' Michael Baillie	15	111' May Bawle	113
111d Bawle Ranchane	71	112' Michael Baillie	15	112' May Bawle	114
112d Bawle Ranchane	71	113' Michael Baillie	15	113' May Bawle	115
113d Bawle Ranchane	71	114' Michael Baillie	15	114' May Bawle	116
114d Bawle Ranchane	71	115' Michael Baillie	15	115' May Bawle	117
115d Bawle Ranchane	71	116' Michael Baillie	15	116' May Bawle	118
116d Bawle Ranchane	71	117' Michael Baillie	15	117' May Bawle	119
117d Bawle Ranchane	71	118' Michael Baillie	15	118' May Bawle	120
118d Bawle Ranchane	71	119' Michael Baillie	15	119' May Bawle	121
119d Bawle Ranchane	71	120' Michael Baillie	15	120' May Bawle	122
120d Bawle Ranchane	71	121' Michael Baillie	15	121' May Bawle	123
121d Bawle Ranchane	71	122' Michael Baillie	15	122' May Bawle	124
122d Bawle Ranchane	71	123' Michael Baillie	15	123' May Bawle	125
123d Bawle Ranchane	71	124' Michael Baillie	15	124' May Bawle	126
124d Bawle Ranchane	71	125' Michael Baillie	15	125' May Bawle	127
125d Bawle Ranchane	71	126' Michael Baillie	15	126' May Bawle	128
126d Bawle Ranchane	71	127' Michael Baillie	15	127' May Bawle	129
127d Bawle Ranchane	71	128' Michael Baillie	15	128' May Bawle	130
128d Bawle Ranchane	71	129' Michael Baillie	15	129' May Bawle	131
129d Bawle Ranchane	71	130' Michael Baillie	15	130' May Bawle	132
130d Bawle Ranchane	71	131' Michael Baillie	15	131' May Bawle	133
131d Bawle Ranchane	71	132' Michael Baillie	15	132' May Bawle	134
132d Bawle Ranchane	71	133' Michael Baillie	15	133' May Bawle	135
133d Bawle Ranchane	71	134' Michael Baillie	15	134' May Bawle	136
134d Bawle Ranchane	71	135' Michael Baillie	15	135' May Bawle	137
135d Bawle Ranchane	71	136' Michael Baillie	15	136' May Bawle	138
136d Bawle Ranchane	71	137' Michael Baillie	15	137' May Bawle	139
137d Bawle Ranchane	71	138' Michael Baillie	15	138' May Bawle	140
138d Bawle Ranchane	71	139' Michael Baillie	15	139' May Bawle	141
139d Bawle Ranchane	71	140' Michael Baillie	15	140' May Bawle	142
140d Bawle Ranchane	71	141' Michael Baillie	15	141' May Bawle	143
141d Bawle Ranchane	71	142' Michael Baillie	15	142' May Bawle	144
142d Bawle Ranchane	71	143' Michael Baillie	15	143' May Bawle	145
143d Bawle Ranchane	71	144' Michael Baillie	15	144' May Bawle	146
144d Bawle Ranchane	71	145' Michael Baillie	15	145' May Bawle	147
145d Bawle Ranchane	71	146' Michael Baillie	15	146' May Bawle	148
146d Bawle Ranchane	71	147' Michael Baillie	15	147' May Bawle	149
147d Bawle Ranchane	71	148' Michael Baillie	15	148' May Bawle	150
148d Bawle Ranchane	71	149' Michael Baillie	15	149' May Bawle	151
149d Bawle Ranchane	71	150' Michael Baillie	15	150' May Bawle	152
150d Bawle Ranchane	71	151' Michael Baillie	15	151' May Bawle	153
151d Bawle Ranchane	71	152' Michael Baillie	15	152' May Bawle	154
152d Bawle Ranchane	71	153' Michael Baillie	15	153' May Bawle	155
153d Bawle Ranchane	71	154' Michael Baillie	15	154' May Bawle	156
154d Bawle Ranchane	71	155' Michael Baillie	15	155' May Bawle	157
155d Bawle Ranchane	71	156' Michael Baillie	15	156' May Bawle	158
156d Bawle Ranchane	71	157' Michael Baillie	15	157' May Bawle	159
157d Bawle Ranchane	71	158' Michael Baillie	15	158' May Bawle	160
158d Bawle Ranchane	71	159' Michael Baillie	15	159' May Bawle	161
159d Bawle Ranchane	71	160' Michael Baillie	15	160' May Bawle	162
160d Bawle Ranchane	71	161' Michael Baillie	15	161' May Bawle	163
161d Bawle Ranchane	71	162' Michael Baillie	15	162' May Bawle	164
162d Bawle Ranchane	71	163' Michael Baillie	15	163' May Bawle	165
163d Bawle Ranchane	71	164' Michael Baillie	15	164' May Bawle	166
164d Bawle Ranchane	71	165' Michael Baillie	15	165' May Bawle	167
165d Bawle Ranchane	71	166' Michael Baillie	15	166' May Bawle	168
166d Bawle Ranchane	71	167' Michael Baillie	15	167' May Bawle	169
167d Bawle Ranchane	71	168' Michael Baillie	15	168' May Bawle	170
168d Bawle Ranchane	71	169' Michael Baillie	15	169' May Bawle	171
169d Bawle Ranchane	71	170' Michael Baillie	15	170' May Bawle	172
170d Bawle Ranchane	71	171' Michael Baillie	15	171' May Bawle	173
171d Bawle Ranchane	71	172' Michael Baillie	15	172' May Bawle	174
172d Bawle Ranchane	71	173' Michael Baillie	15	173' May Bawle	175
173d Bawle Ranchane	71	174' Michael Baillie	15	174' May Bawle	176
174d Bawle Ranchane	71	175' Michael Baillie	15	175' May Bawle	177
175d Bawle Ranchane	71	176' Michael Baillie	15	176' May Bawle	178
176d Bawle Ranchane	71	177' Michael Baillie	15	177' May Bawle	179
177d Bawle Ranchane	71	178' Michael Baillie	15	178' May Bawle	180
178d Bawle Ranchane	71	179' Michael Baillie	15	179' May Bawle	181
179d Bawle Ranchane	71	180' Michael Baillie	15	180' May Bawle	182
180d Bawle Ranchane	71	181' Michael Baillie	15	181' May Bawle	183
181d Bawle Ranchane	71	182' Michael Baillie	15	182' May Bawle	184
182d Bawle Ranchane	71	183' Michael Baillie	15	183' May Bawle	185
183d Bawle Ranchane	71	184' Michael Baillie	15	184' May Bawle	186
184d Bawle Ranchane	71	185' Michael Baillie	15	185' May Bawle	187
185d Bawle Ranchane	71	186' Michael Baillie	15	186' May Bawle	188
186d Bawle Ranchane	71	187' Michael Baillie	15		

Reserve Army Strength without the USA		15107	Strength of the Regiments	9337	
S Andrew Hussars	1013 110 1	Baldwin's Hussars	1200 179	S' Dundas' Light Dragoons	4516 1446
S Lancashire Guards	1444 144	S' Royal Ulster	997 5 175	S' George's Own	1811 1120
S Lancashire Lancers	91 147	S' Royal Alfonso	4916 4015	S' Gloucestershire	4700 1746
2 Dragoon Guards	311 1447	S' Royal Dragoon Guards	1424 1200	S' Gloucestershire	3059 1212
Artillery	16	Regimental Artillery	41511	Artillery Reserve	1519 156

S' Glastonbury Fields	145 7'	16° 5' Eastward Tower-1956	801 2° Magdalen Beacon	194 1'	16° 5' West Victoria Chapel 1956	805
Maypole Fields	113	81 Lubenham Parish	273 5 17 2° May Newington	123	800 16° 5' Lubenham Parish	204 220
S' Tawny Cockle	155	127 2° Luton Stowthorpe	165 2 1 249 2° May Bishopstone	165	191 16° 5' Bishopstone	204 221

<i>Parishes in the Non-Parishion, in Middlesex and Surrey.—1854.</i>	<i>Murder, of the Poor.—2240.</i>
<i>St. Clement Danes.—169-11197</i>	<i>Murder, Slaves.—303 1987.</i>
<i>St. Paul Covent Garden.—163</i>	<i>Murders, Wives.—1470 1743.</i>
<i>St. Martin in the Fields.—704 255.</i>	<i>Murders, of the Poor.—156.</i>
<i>Buried in the Cemeteries in the City and Liberties of Westminster.—11894.</i>	<i>The Total of all the Christnings.—9967.</i>
<i>Murder, of the Poor.—2240.</i>	<i>The Total of all the Burials this year.—27306.</i>
	<i>Whereof, of the Poor.—6316.</i>

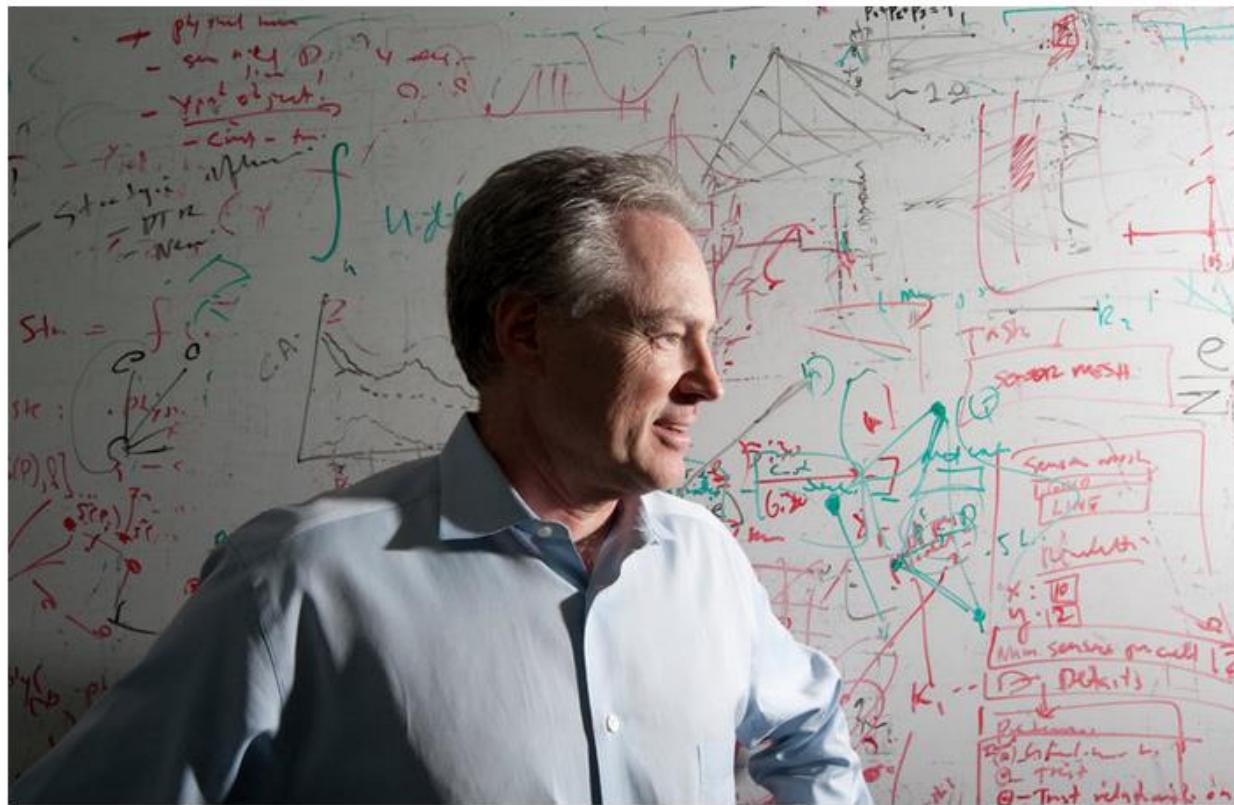
The Diseases and Casualties this year.					
A Bonite and Silbore	617	Executed	12	Pallie	50
Aged	1545	Fox and Small Pox	655	Plague	68196
Ague and Fever	5157	Found dead in streets, fields, &c. &c.	20	Planner	6
Appoplex and Suddenly	216	French Pox	86	Plurise	25
Bedrid	10	Frighted	13	Poisoned	1
Blaffed	5	Gout and Sciatica	17	Quassie	35
Bleeding	16	Grief	46	Rickets	557
Bloody Flux, Scouring & Flux	185	Gripping in the Guts	128	Rifing of the Lighes	397
Borst and Scalded	8	Hanged & made away themselves	7	Rupture	34
Calenture	3	Headsmoldhoit & Mouldfallen	16	Scurvy	105
Cancer, Gangrene and Fistula	56	Jaudies	116	Stungles and Swine pot	8
Canker, and Thrush	113	Impofhamme	237	Sores, Ulcers, broken and bruised	
Childbed	643	Kild by severall accidents	35	Limbs	82
Chorfones and Infants	1258	Kungs Evill	36	Spleen	14
Cold and Cough	68	Leprofic	1	Spotted Fever and Purples	1929
Collick and Winde	134	Lethargy	14	Scopping of the flonack	332
Consumption and Tidlick	4808	Livergrown	31	Stone and Stunary	98
Cowslion and Mother	2034	Megatom and Headach	13	Suflet	112
Dilatred	1	Males	7	Tech and Wonns	2614
Drephic and Timpany	147	Murthered and Shoe	9	Vomiting	52
Drowned	50	Overlef and Scarfed	45	VVash	
Males	51542	Males	48369		
Chairled Femroles	4851	Buried Femroles	48737	Of the Plague	68596

(In all—9967) (In all—7303)

Increased in the Burials in the 130 Parishes under the Poor-Laws this year—
Increase of the Births in the 500 Parishes and at the Poor-Houses this year—

Microsoft Finds Cancer Clues in Search Queries

By JOHN MARKOFF JUNE 7, 2016



Dr. Eric Horvitz is one of the Microsoft researchers who conducted the study.

Kevin P. Casey for The New York Times

[Microsoft](#) scientists have demonstrated that by analyzing large samples of search engine queries they may in some cases be able to identify internet users who are suffering from pancreatic cancer, even before they have

THE WALL STREET JOURNAL.

Home World U.S. Politics Economy Business **Tech** Markets Opinion Arts |



Apple's New
iPhones Arrive as
Glow Fades



Samsung's
World-Wide Recall
Brings Battery-Maker
Into Focus



Patrick Drahi
Makes \$2 Billion
Share Offer for Full
Control of SFR



TECH



Can a Smartphone Tell if You're Depressed?

Apps, Other Tools Help Doctors, Insurers Measure Psychological Well-Being

By JOSEPH WALKER

Updated Jan. 5, 2015 7:03 p.m. ET

HUNTERSVILLE, N.C.—Toward the end of Janisse Flowers's pregnancy, a nurse at her gynecologist's office asked her to download an iPhone app that would track how often she text messaged with friends, how long she talked on the phone and how far she traveled each day.

The app was part of an effort by Ms. Flowers's health-care provider to test whether smartphone data could help detect symptoms of postpartum depression, an

PREDICTIVE POLICING: USING MACHINE LEARNING TO DETECT PATTERNS OF CRIME



Image: [lydia_shiningbrightly/Flickr](#)

TRYING TO DETECT specific patterns of crime and criminal behavior is extremely challenging. Crime analysts can spend countless hours sifting through data to determine whether a crime fits into a known pattern and to discover new patterns. Once a pattern is detected, the information can be used to predict, anticipate and prevent crime.



GET WIRED

Don't Let The
Behind. Get 6

SUBSCRIBE NOW

MOST PO



HOV
Kid
Mai
Gar
06.:



PHO
Bey
Pap
Wo
04.1



TV
Sta
Imp
Rar
15 |

→ MORE STORI

IBM detects skin cancer more quickly with visual machine learning



MORE



Credit: Thinkstock

Cognitive computing could improve the ability to spot melanoma early on



By Joab Jackson

FOLLOW

IDG News Service | Dec 17, 2014 11:27 AM PT

RELATED TOPICS

Data Analytics

Big Data

Healthcare IT

COMMENTS

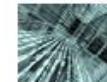
Skin cancer can be detected more quickly and accurately by using cognitive computing-based visual analytics, researchers at IBM Research have found, in collaboration with New York's Memorial Sloan Kettering Cancer Center.

In a scan of 3,000 images, IBM technology was able to spot melanoma with an accuracy of about 95 percent, much better than the 75 percent to 84 percent average of today's largely manual methods.

MORE LIKE THIS



IBM expands Watson's presence to multiple industries



IBM looks to replace the spreadsheet with Watson Analytics



10 IBM Watson-powered apps that are changing our world

on IDG Answers

How does Simply Secure improve cloud security?

M • Technology • Game of Thrones

Meet the artificial intelligence that knows who's going to die next in Game of Thrones season 6

19:39, 21 APR 2016

UPDATED 19:59, 21 APR 2016

BY ALPHR.COM

Want to know who's going to die in Game of Thrones season six? This algorithm can tell you



SHARES



COMMENTS

Enter your e-mail for our daily newsletter

Subscribe



★ Recommended In Technology



IPHONE 7

iPhone 7 could be completely waterproof, new reports claim ahead of big launch this week



IPHONE 7

iPhone 7 last-minute leak claims Apple's new iPhone will come in five colours and include one killer feature for photo fanatics



How Big Data Is
Changing Healthcare

+145,375 views in the last 24 hours



Apple Leak Reveals
Massive New iPhone 7



Active on Twitter

Apple Watch Sales
Plunge 55% As Samsung
Doubles Market Share:
Report



India Just Planted 50
Million Trees In 24
Hours

APR 21, 2015 @ 10:50 AM

87,359 VIEWS

How Big Data Is Changing Healthcare

**Bernard Marr**

CONTRIBUTOR

I write about big data,
analytics and enterprise
performance

[FULL BIO >](#)

Opinions expressed by Forbes
Contributors are their own.

If you want to find out how Big Data is helping to make the world a better place, there's no better example than the uses being found for it in healthcare.

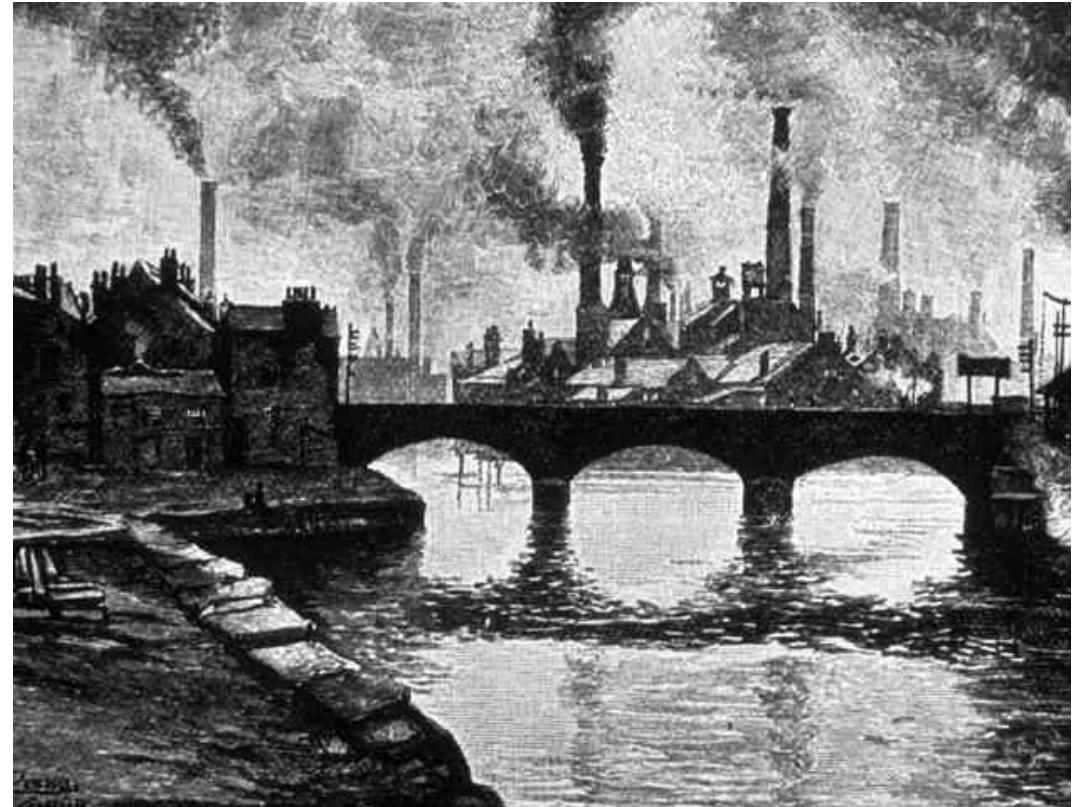
The last decade has seen huge advances in the amount of data we routinely generate and collect in pretty much everything we do, as well as our ability to use technology to analyze and understand it. The intersection of these trends is what we call "Big Data" and it is helping businesses in every industry to become more efficient and productive.

Healthcare is no different. Beyond improving profits and cutting down on wasted overhead, Big Data in healthcare is being used to predict epidemics, cure disease, improve quality of life and avoid preventable deaths. With the

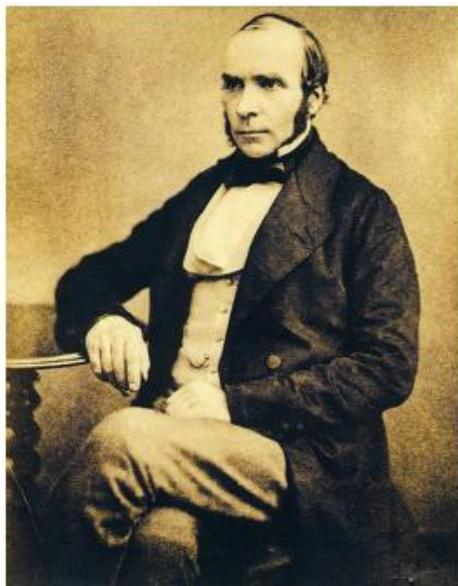
Story from 19th century

London, 1854 :

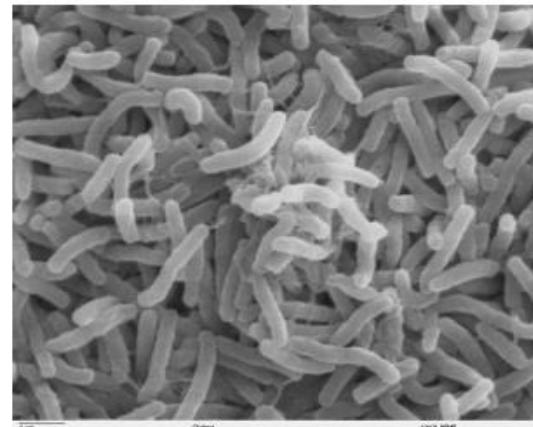
- Industrial revolution.
- Two millions people.
- Insufficient sewage.
- Garbage removal problems.
- Little clean water.



Characters:



John Snow
1813-1858



Vibrio Cholerae
still around

John Snow was a strong advocate of *hygiene* and *anesthesia*.

Against the Miasma Theory

Outbreak of cholera

The most terrible outbreak of cholera which ever occurred in this kingdom, is probably that which took place in Broad Street, Golden Square, and the adjoining streets, a few weeks ago. Within two hundred and fifty yards of the spot where Cambridge Street joins Broad Street, there were upwards of five hundred fatal attacks of cholera in ten days. The mortality in this limited area probably equals any that was ever caused in this country, even by the plague, and it was much more sudden, as the greater number of cases terminated in a few hours.

John Snow, *On the mode of communication of cholera*, 1854.

Snow uses **simple statistics** to infirm competing hypotheses.

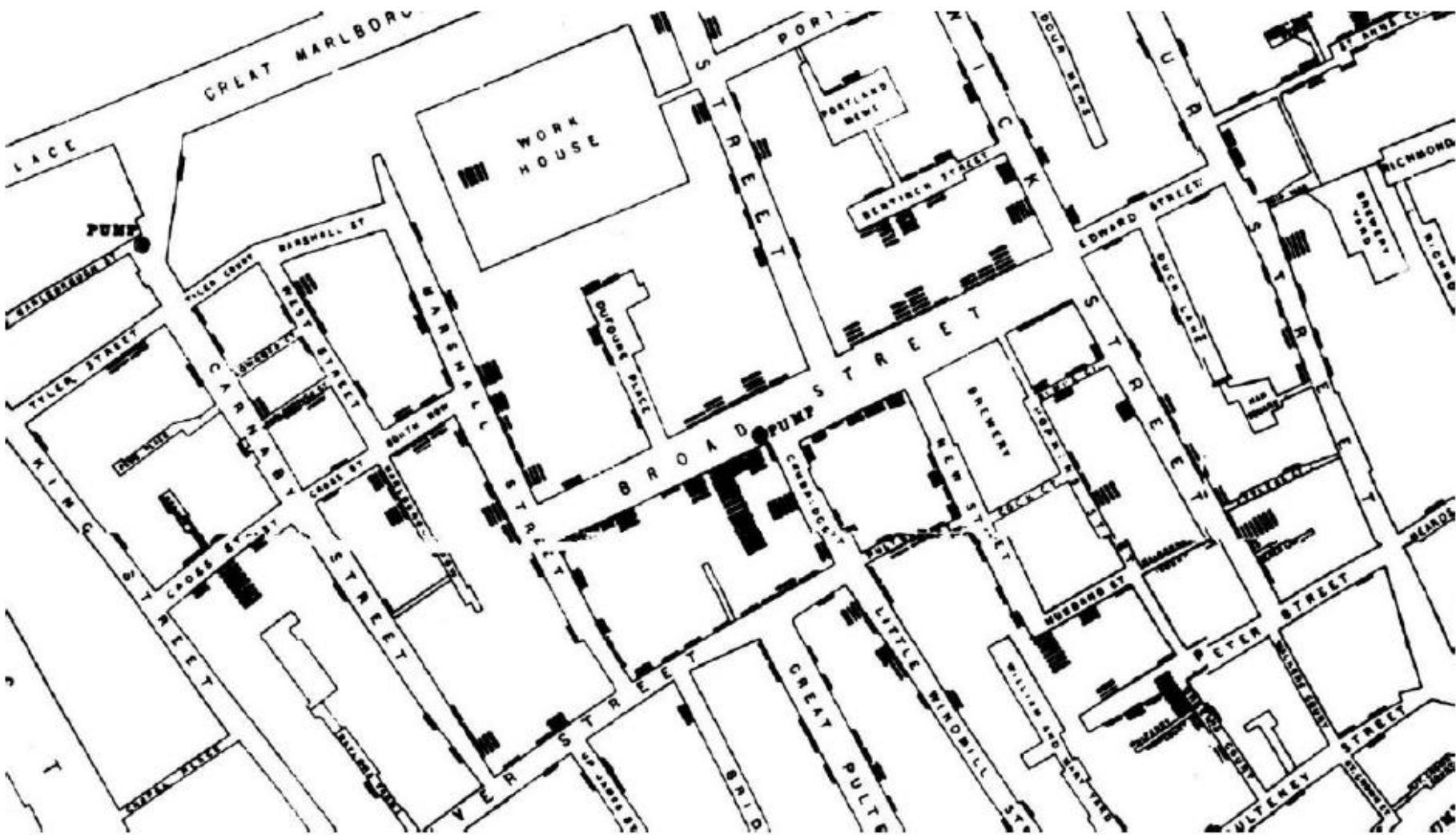
Table XIV (partial)

	No of Deaths	Ratio
Agents	12	1 in 40
Bricklayers and builders	14	1 in 39
Physicians, surgeons, ...	16	1 in 265
Magistrates, barristers, ...	13	1 in 375
Merchants	11	1 in 348
Footmen and men servants	25	1 in 1572

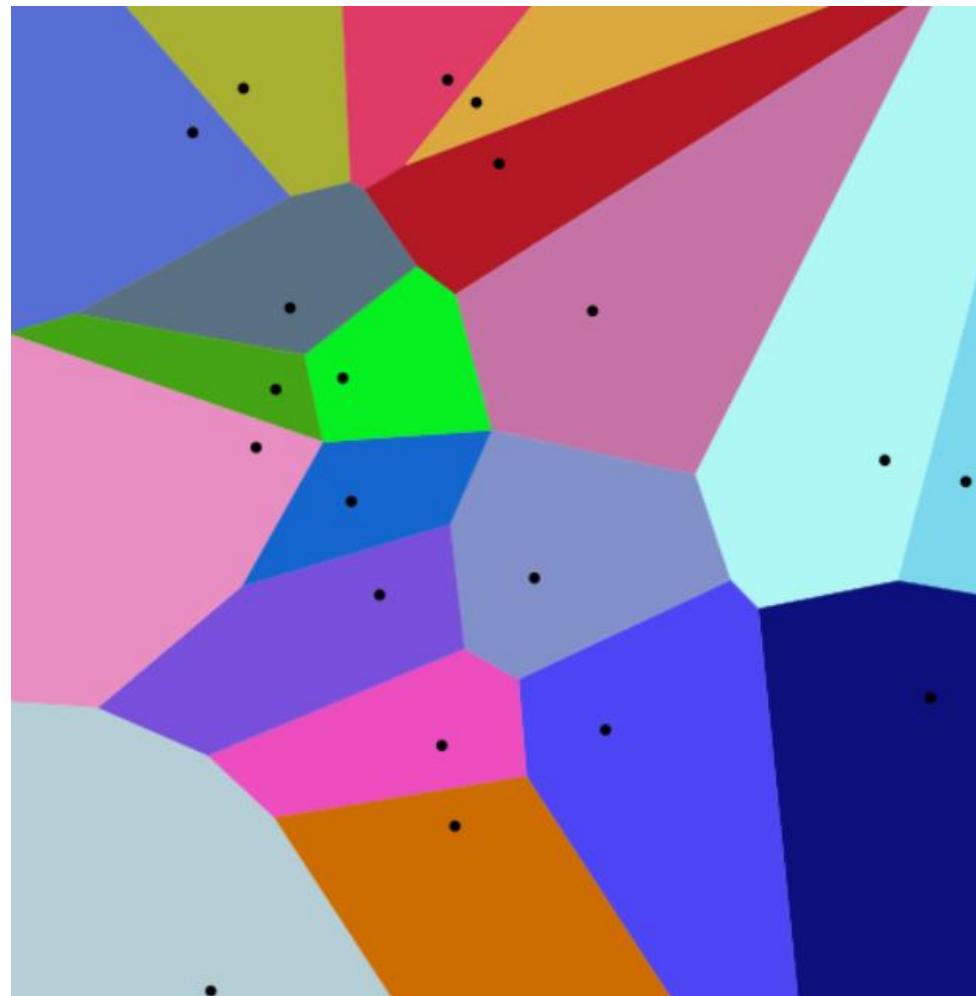
If cholera was propagated by effluvia from sick people,

- why are physicians less affected than their patients?
- why are men servants less affected than their masters?
- why are master brewers virtually immune?

The map



Voronoi Diagram



Epidemiology and statistics

Snow uses **simple statistics** to confirm the role of impure water.

TABLE VIII (MORTALITY FROM CHOLERA IN 7 WKS ENDING 26TH AUGUST)

Sub-Districts	Pop.	Deaths by Cholera in the four wks. ending 5th August	Water Supply				
			Southwark & Vauxhall	Lambeth	Pump-wells	River Thames, ditches, etc.	Unascertained
St. Saviour, Southwark	19,709	125	115	-	-	10	-
St. Olave, Southwark	8,015	53	43	-	-	5	5
St. John, Horsleydown	11,360	51	48	-	-	3	-
St. James, Bermondsey	18,899	123	102	-	-	21	-
St. Mary Magdal'en	13,934	87	83	-	-	4	-
Leather Market	15,295	81	81	-	-	-	-
Rotherhithe	17,805	103	68	-	-	35	-

The Broad Street water pump



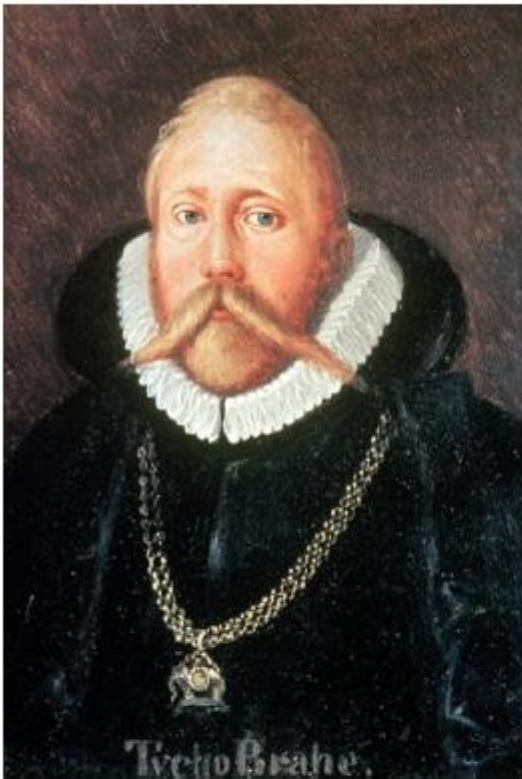
On proceeding to the spot, I found that nearly all the deaths had taken place within a short distance of the [Broad Street] pump. There were only ten deaths in houses situated decidedly nearer to another street-pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pumps which were nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street...

John Snow, *On the mode of communication of cholera*, 1854.

Big Data Story from 17th century

- Astronomy laws
- Planetary motion
- The orbits of the planets are ellipses

The characters:

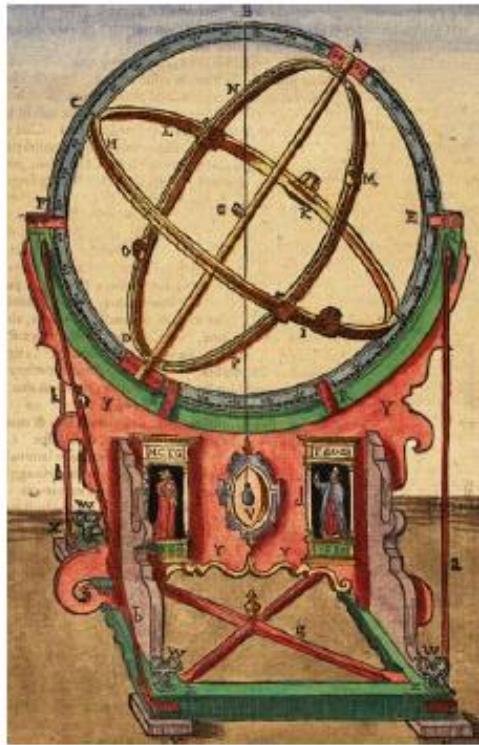
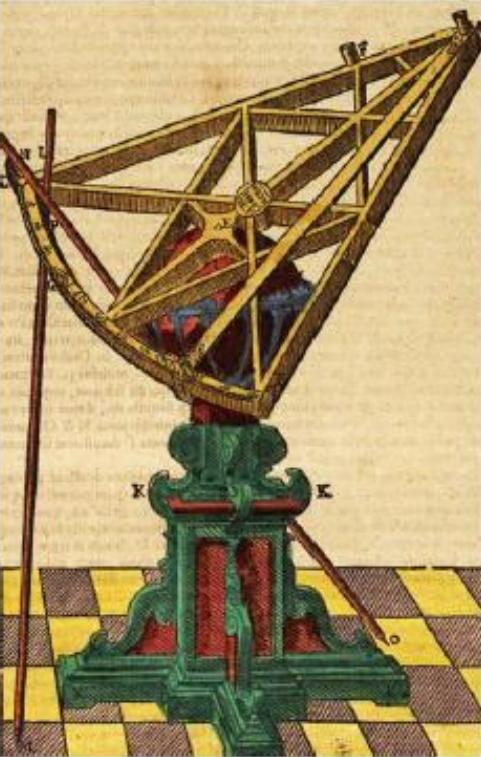


Tycho Brahe
1546-1601



Johannes Kepler
1571-1630

Tycho's Observatories



First in Uraniborg.

Then near Prague,
thanks to a “grant”
from emperor Rudolf II.

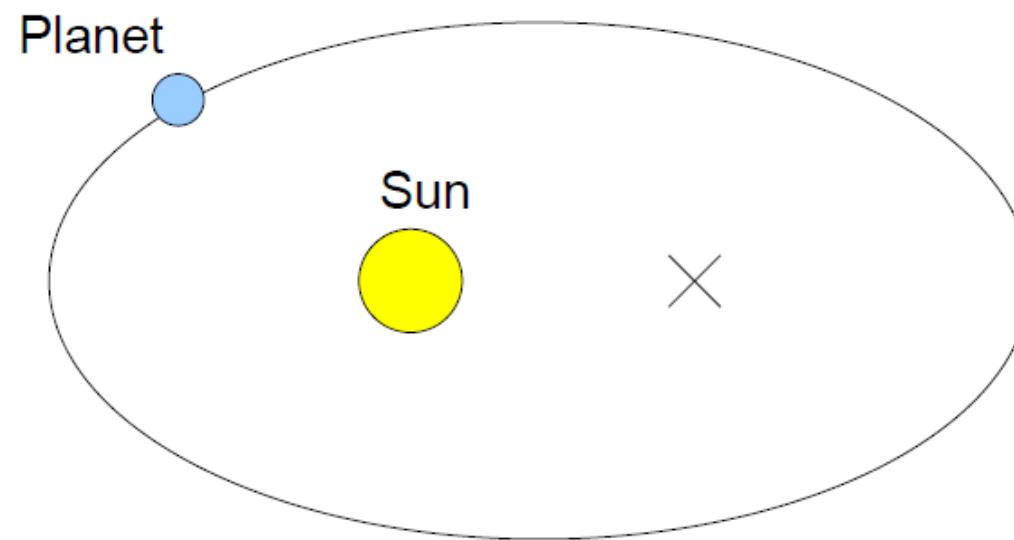
There he hires a bright
young assistant named
Johannes Kepler.

Without telescope, but with a modern approach to data collection:

- daily observation of 1000 stars and 7 planets,
- record positions $\pm 1'$ arc.

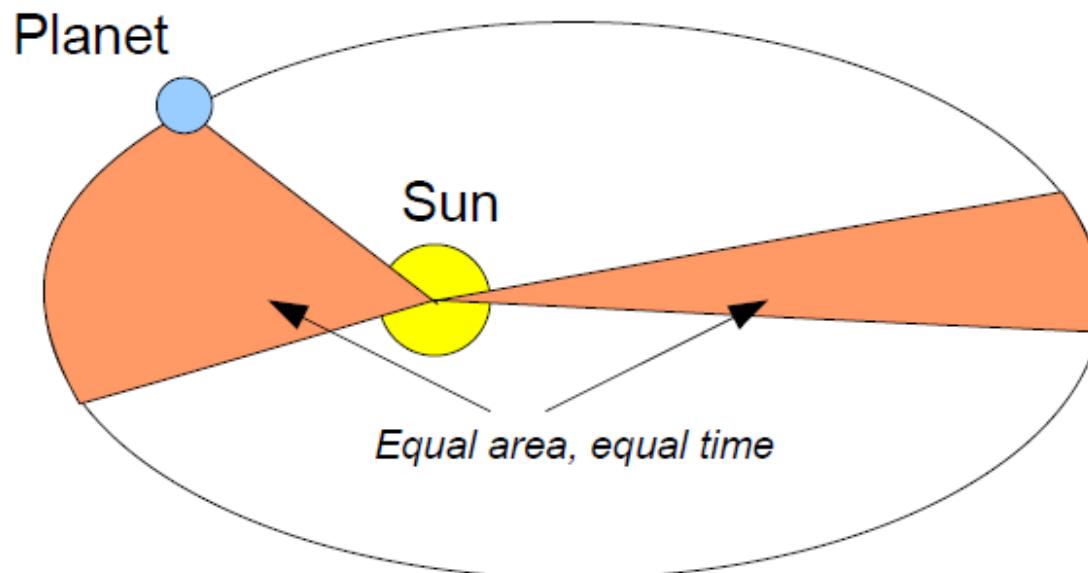
The 1st law of Kepler

The orbits of the planets are ellipses
with the Sun at a focal point.



The 2nd law of Kepler

The line joining the planet to the Sun sweeps out equal areas in equal times as the planet travels around the ellipse.



The 3rd law of Kepler

The ratio of the squares of the revolutionary periods for two planets is equal to the ratio of the cubes of the length of their major axes.

$$\frac{P_a^2}{P_b^2} = \frac{R_a^3}{R_b^3}$$

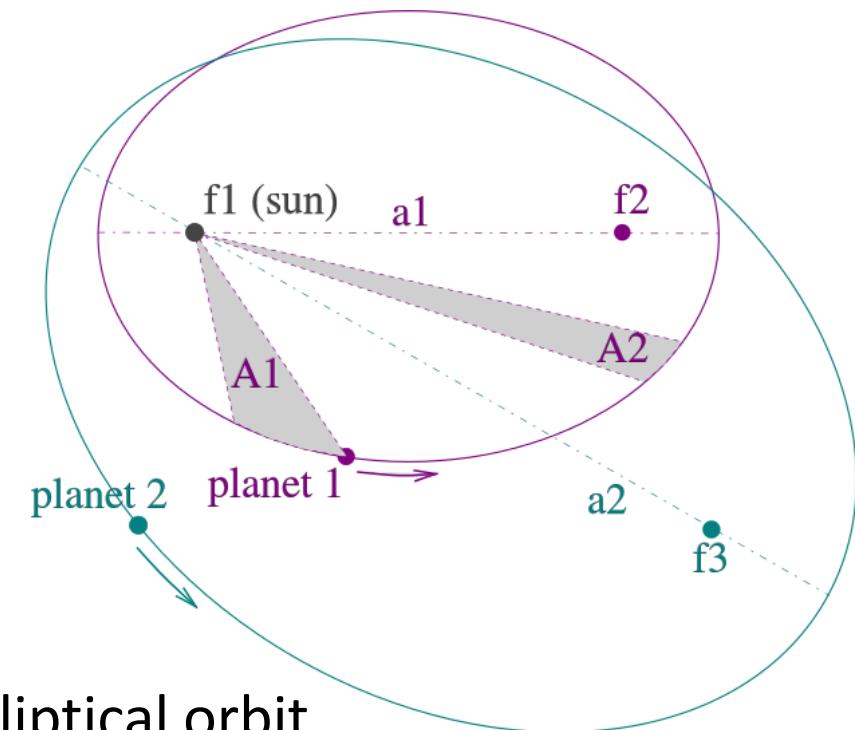
Kepler's laws improve the model of Copernicus

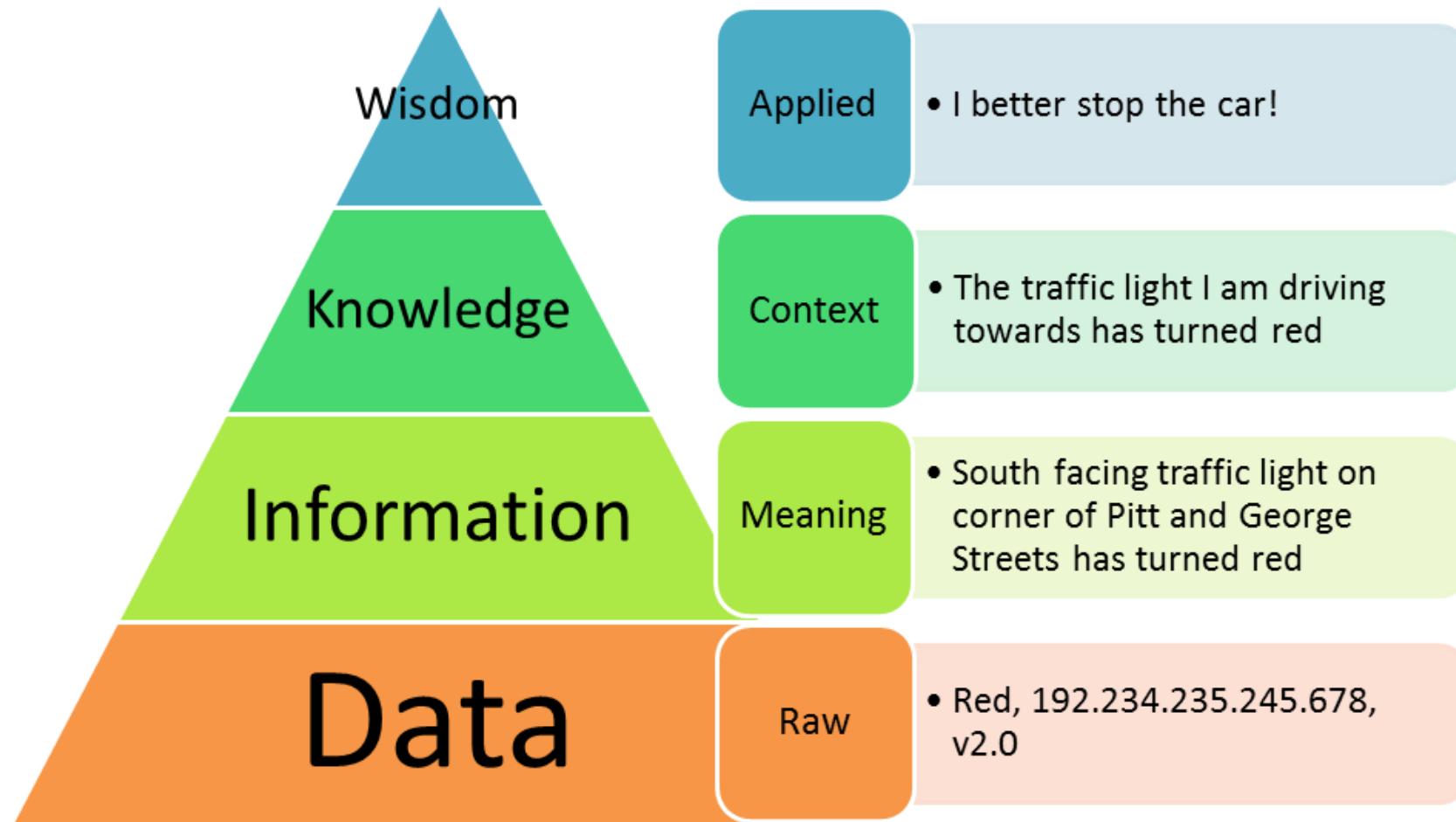
Copernicus's model say:

- The planetary orbit is a circle
- The Sun at the center of the orbit
- The speed of the planet in the orbit is constant

Kepler's model say:

- The planetary orbit is *not* a circle, but an *ellipse*.
- The Sun is *not* at the center but at a *focal point* of the elliptical orbit.
- Neither the linear speed nor the angular speed of the planet in the orbit is constant, but the *area speed* is constant.





Yapay Öğrenme (Machine Learning)

Why “learn”?

Learning from the data is useful when:

- Human expertise is absent
(navigating on Mars, autonomous driving)
- Humans are unable to explain their expertise
(speech recognition, vision, language)
- Solution changes in time, self customizing programs
(portfolio management, news/books recommendations)
- Using historical data to improve decisions
(understanding customer behavior from transactions, predicting which electrical structures will explode)



My research on limitations of Big Data

- Dataset sizes are outgrowing the computing power of computers
get rid of redundant data.
- The size of the model grows which affects the memory requirements.
Streaming data.
online algorithms.
- Data can be imbalanced!
make it balanced by smart methods.
- Who is going to label all those data?
crowdsource it.
- Data can be misleading,
mislabeled, untrusted!
find ways to exclude them.

Projects authored by seyda ertekin.

Search

 Go

Manage

- [Login](#) to submit a new project

Sort by

- [Last Update](#)
- [Publication Date](#)
- [Project Title](#)
- [Rating](#)
- [Number of Views](#)
- [Number of Downloads](#)



LASVM 1.1

by [leonbottou](#) - August 3, 2009, 15:50:30 CET [[Home](#) [Edit](#) [View](#)] 11263 views, 2089 downloads, 0 subscriptions

About: Reference implementation of the LASVM online and active SVM algorithms as described in the JMLR paper. The interesting bit is a small C library that implements the LASVM process and reprocess [...]

Changes:

Minor bug fix

- **Authors:** [Leon Bottou](#), [Antoine Bordes](#), [Seyda Ertekin](#)
- **License:** [Gpl Version 2 Or Later](#)
- **Programming Language:** [C](#)

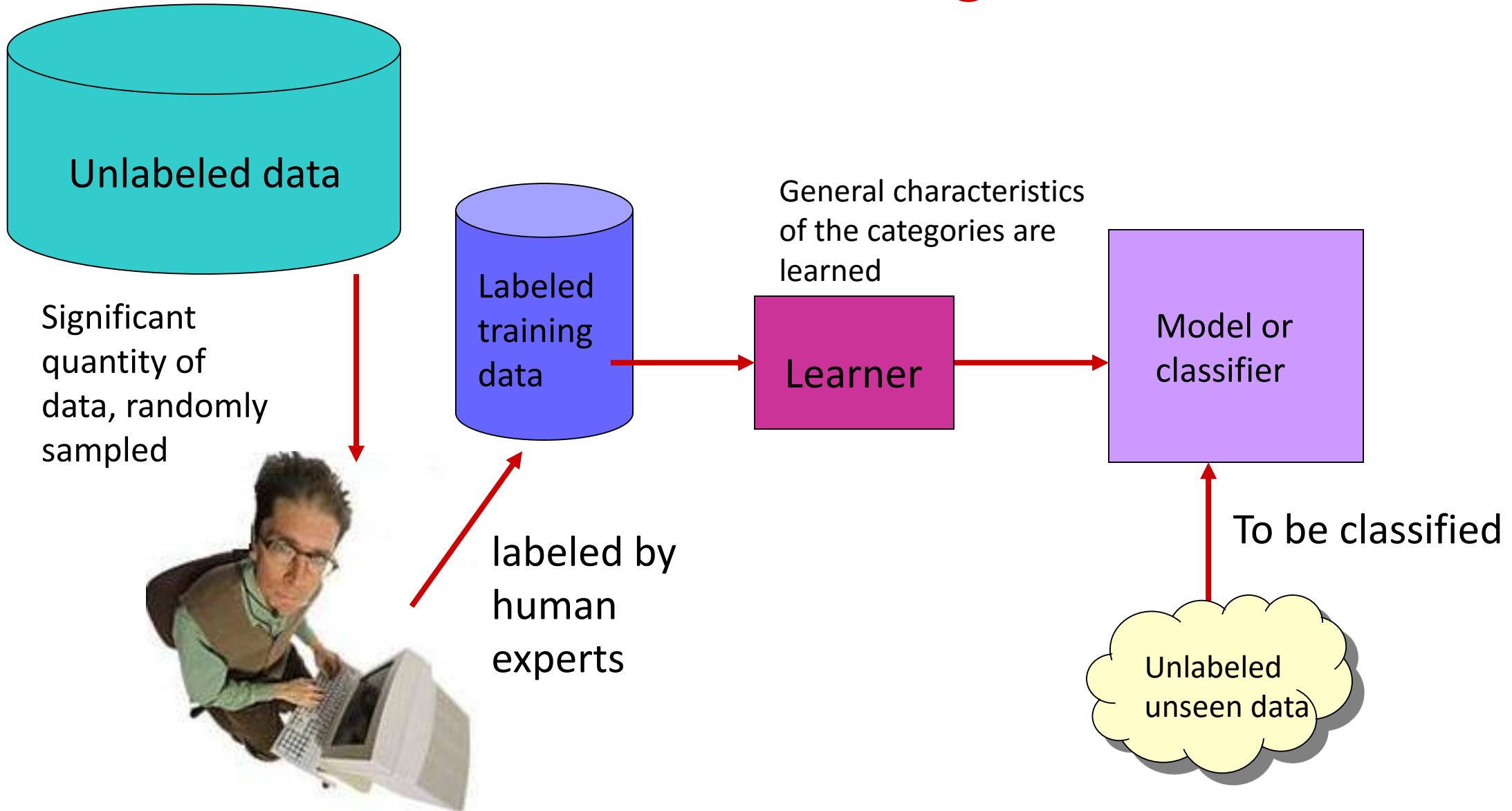
- **Operating System:** [Agnostic](#)
- **Data Formats:** [None](#)
- **Tags:** [Large Scale](#), [Support Vector Machine](#), [Active Learning](#), [Kernels](#), [Online Learning](#)

Fast Kernel Classifiers with Online and Active Learning.

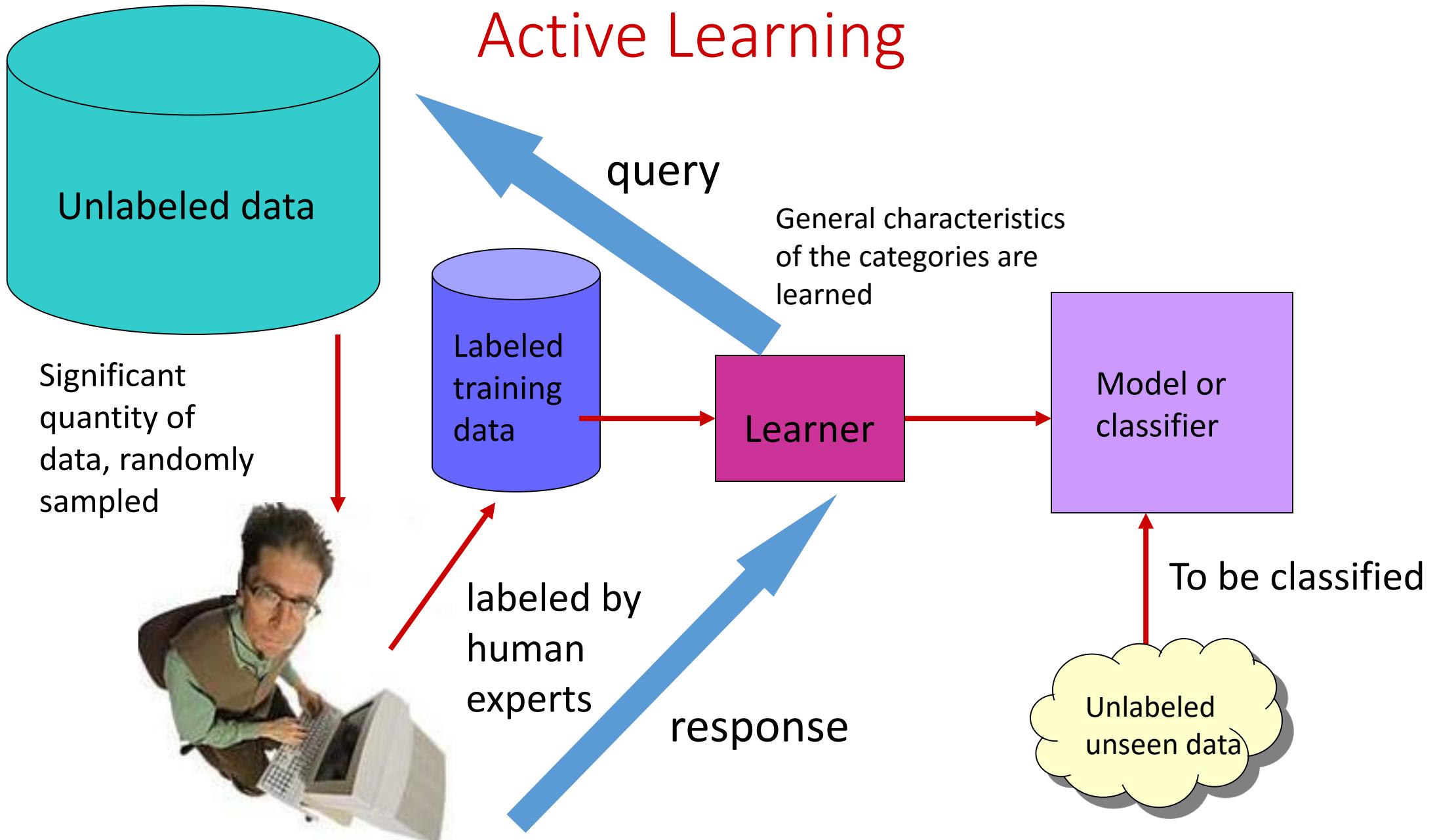
Antoine Bordes, Seyda Ertekin, Jason Weston, Leon Bottou.

Journal of Machine Learning Research (JMLR)

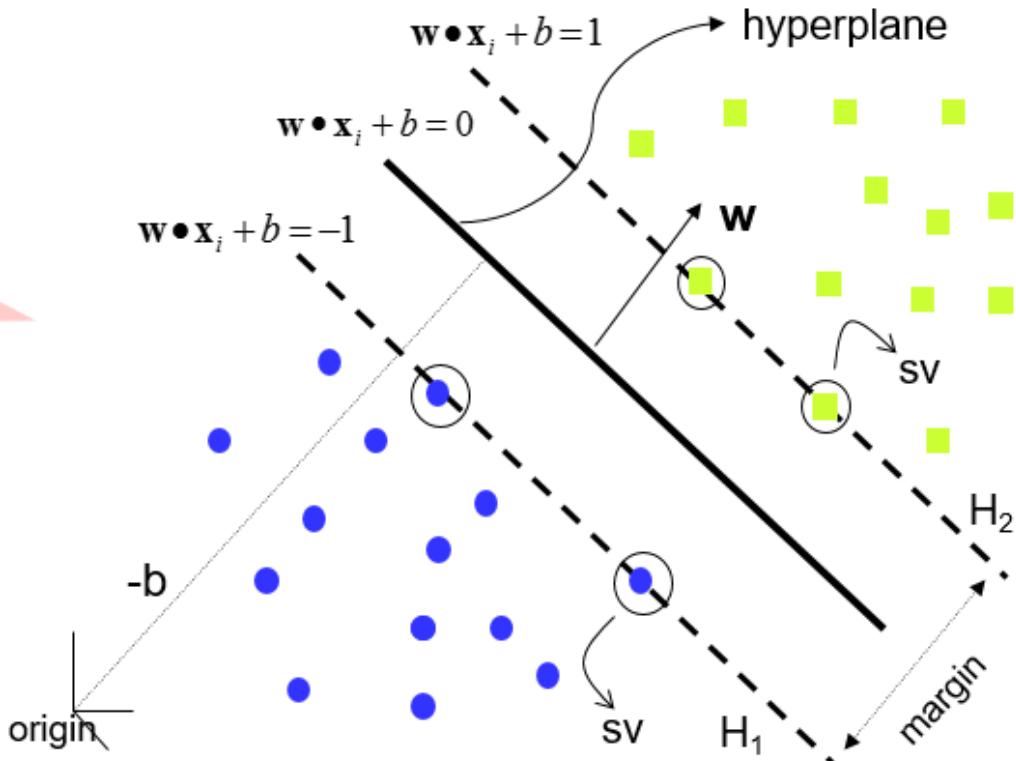
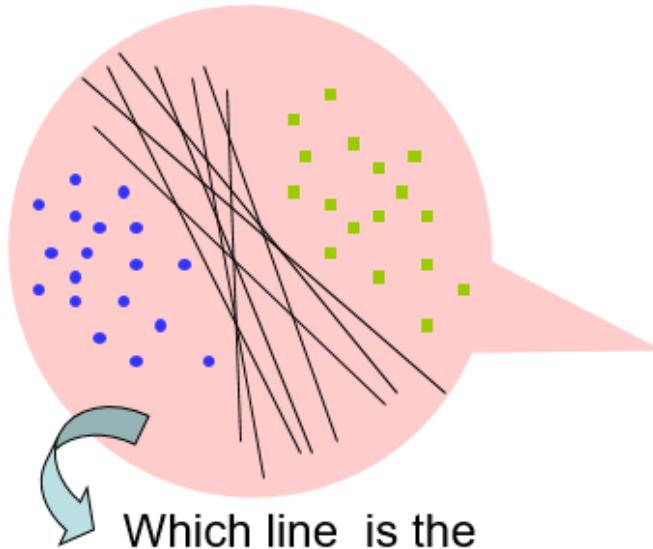
Passive Learning



Active Learning



Support Vector Machines (SVMs) I



Separable case (Boser B., Guyon I., Vapnik V. '92)

$$\min_{w,b} \frac{1}{2} \|w\|^2 \text{ with the constraints } y_i f_\theta(x_i) \geq 1$$

Support Vector Machines (SVMs) II

Inseparable Case (C. Cortes & V. Vapnik '95)

Primal SVM formulation

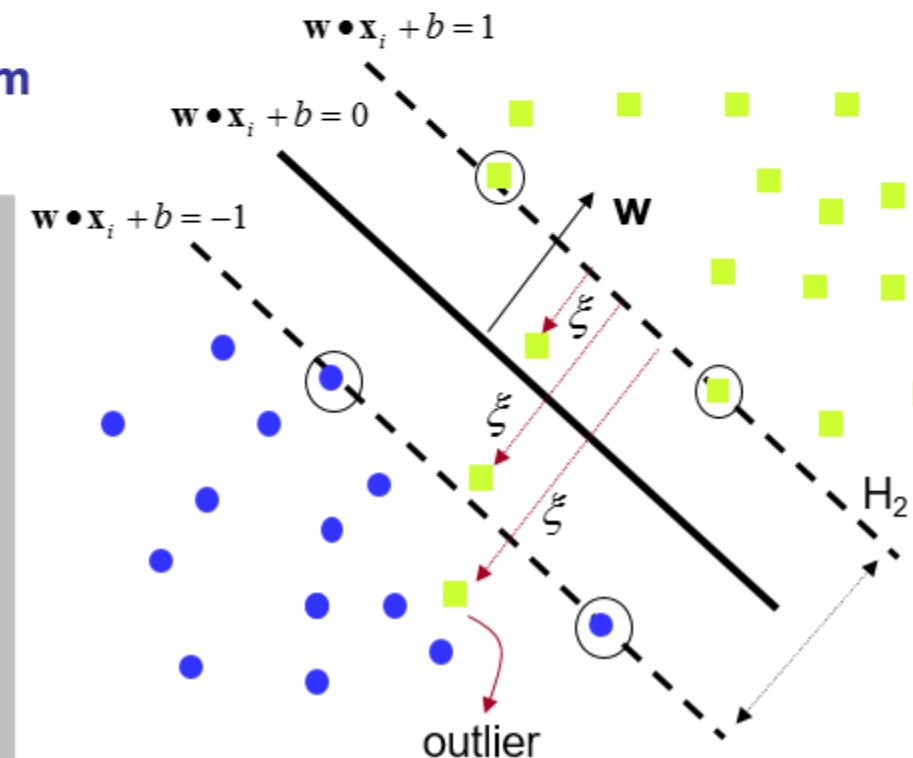
$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \text{ with the constraints } \forall i \ y_i f(x_i) \geq 1 - \xi_i \\ \forall i \ \xi_i \geq 0$$



Dual convex optimization problem
needs Quadratic Programming (QP)

$$\max_{\alpha} W(\alpha) = \sum_i \alpha_i y_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j K(x_i x_j)$$

$$\text{with the constraints} \quad \begin{cases} \sum_i \alpha_i = 0 \\ A_i \leq \alpha_i \leq B_i \\ A_i = \min(0, Cy_i) \\ B_i = \max(0, Cy_i) \end{cases}$$



SVM Quadratic Programming

QP problem is solved when
(KKT conditions)



$$\begin{aligned}\alpha_i = 0 &\Leftrightarrow y_i f(x_i) \geq 1 \\ 0 < \alpha_i < C &\Leftrightarrow y_i f(x_i) = 1 \\ \alpha_i = C &\Leftrightarrow y_i f(x_i) < 1\end{aligned}$$

After solving QP, we get

$$f_{\theta}(x) = \sum_i \alpha_i \Phi(x_i)^T \Phi(x) + b$$

$$\omega = \sum_{i=1}^n \alpha_i \Phi(x_i)$$

$\alpha_i \neq 0$ for SV's

$\alpha_i = 0$ for non SV's

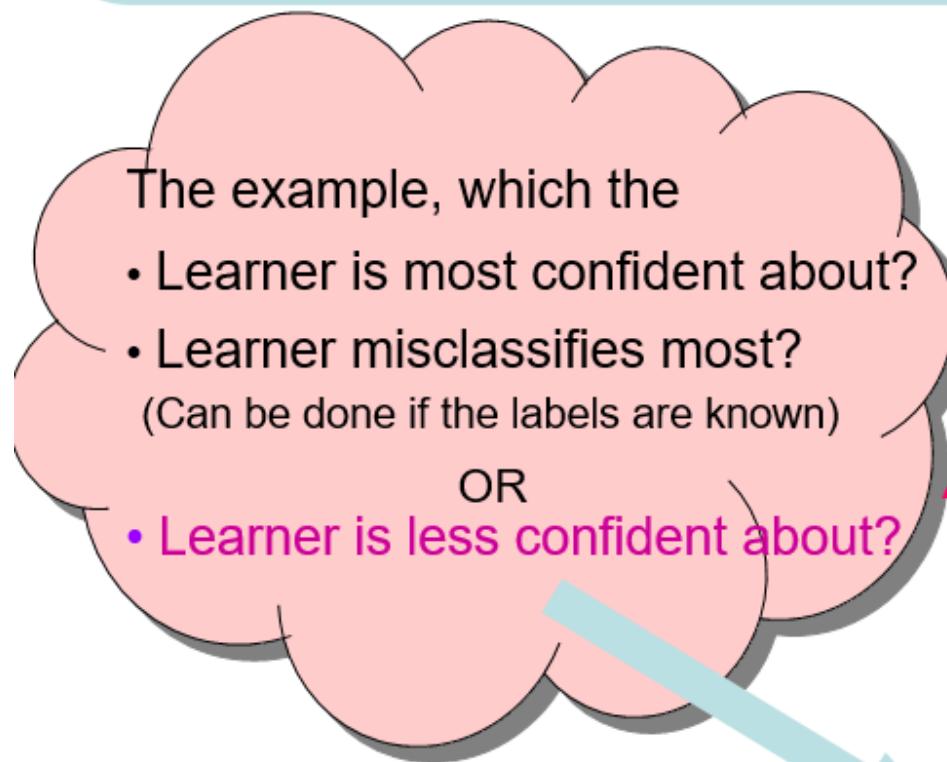
of SVs are important!

Rate of convergence is slow when there are many support vectors.

SVMs may be quite costly with large datasets

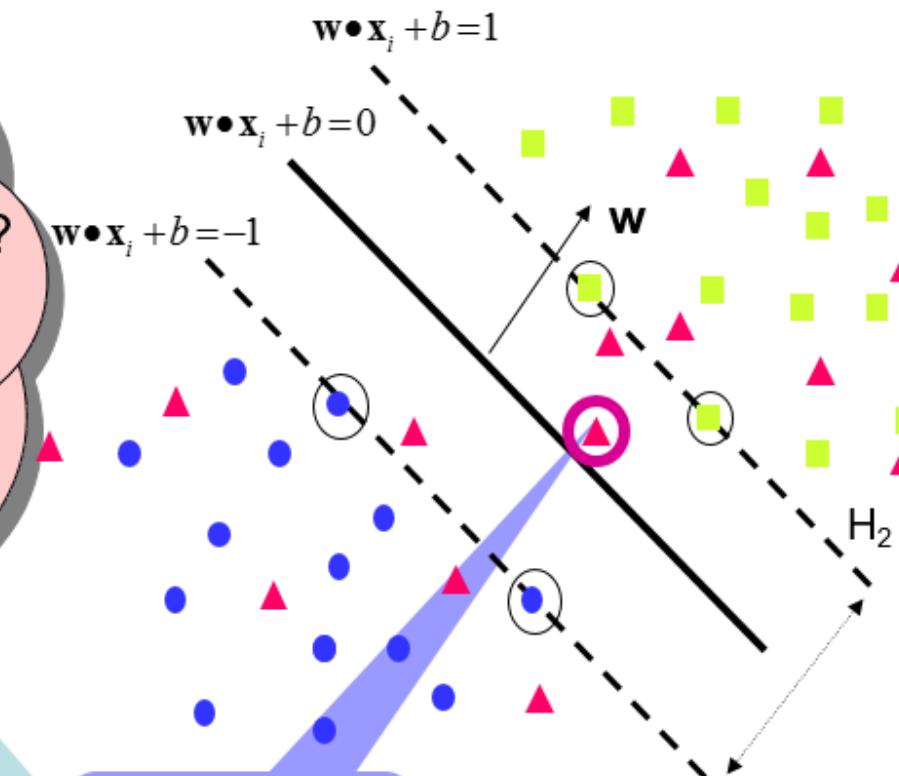
$$\frac{\# \text{ SV}}{\# \text{ examples}} \rightarrow 2B_{\Phi}$$

Which example is the most informative?



Do we really have to search the
entire data?

Not really!



Most
informative

Small Pool Active Learning

Pick a random subset X_L , $L \ll \#$ of training instances

Select x_i from X_L { x_i is among the top $p\%$ closest in X_N with prob. $(1-q)$ }

Probability that at least one of the L instances is among the closest $p\%$

$$1 - (1 - p\%)^L$$

Due to the requirement of $(1-q)$ probability, we have

$$1 - (1 - p\%)^L = 1 - q$$

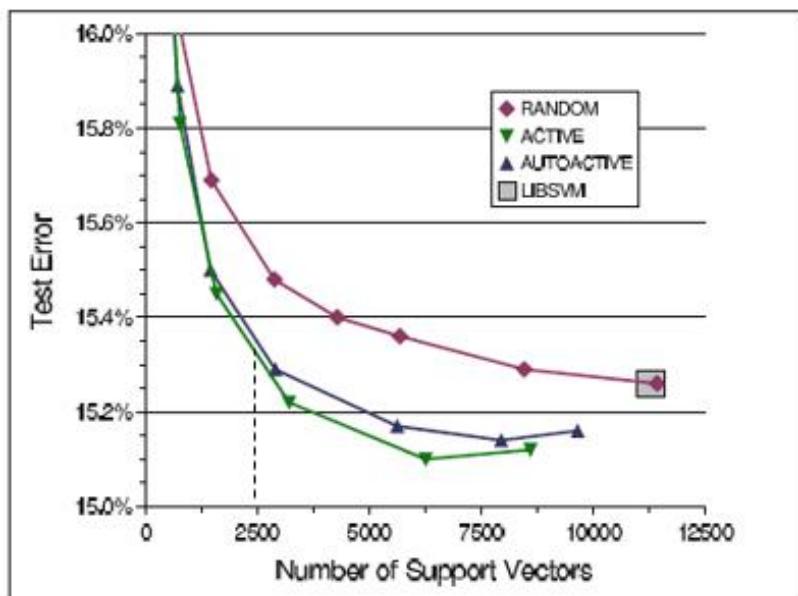
L in terms of p and q

$$L = \log q / \log (1 - p\%)$$

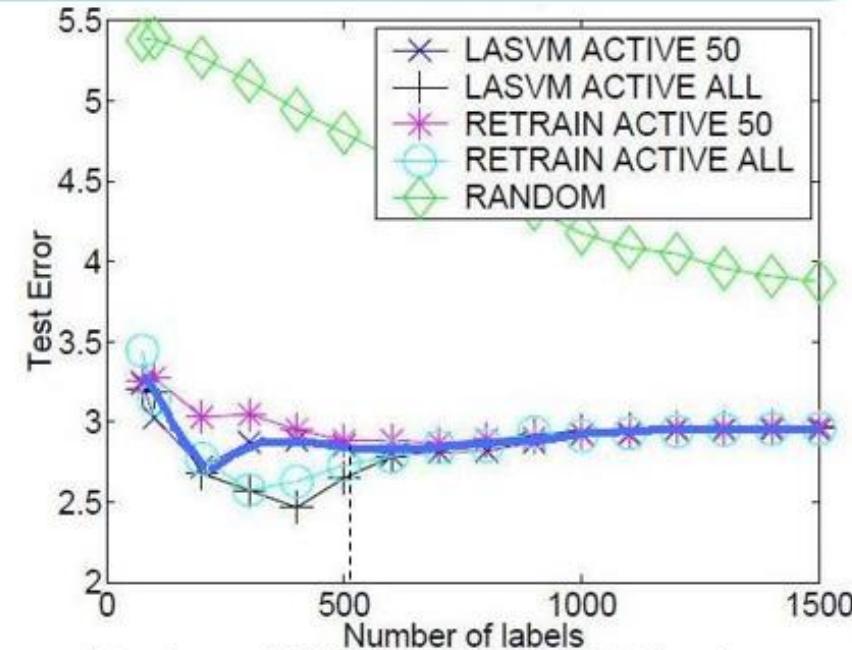
For example, the active learner will pick one instance, with 95% probability, that is among the top 5% closest instances to the hyperplane, by randomly sampling only $\log(.05)/\log(.95)=59$ instances

regardless of the training set size!

Online Learning (LASVM) with Active Learning



Adult: 32562 training 16282 test examples



Reuters: 7770 training, 3299 test examples

AL reaches the LibSVM solution with 2500 SVs rather than 11250 in
1/4th of time (Adult dataset)

For 500 queried
labels (Reuters)



LibSVM Active All: 920 seconds

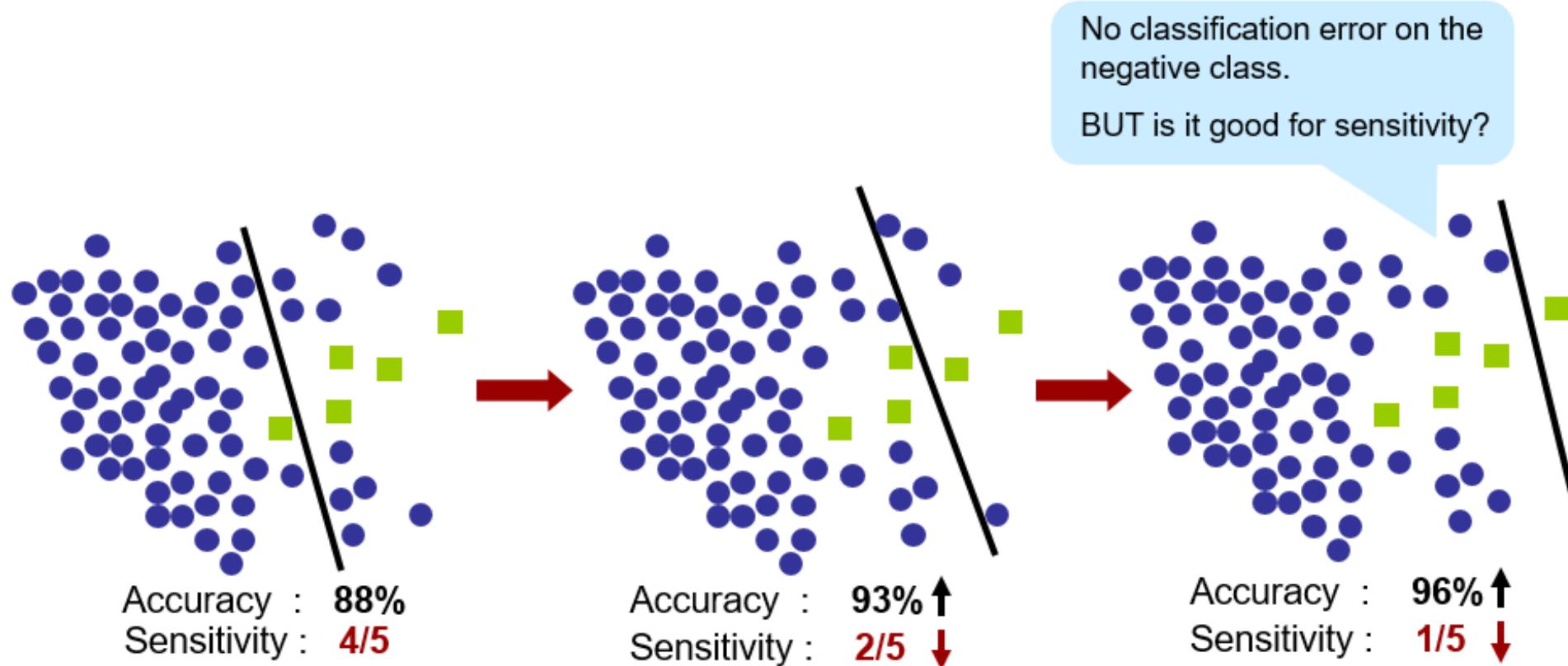
LASVM Active All: 99 seconds

LASVM Active 50: **9 seconds**



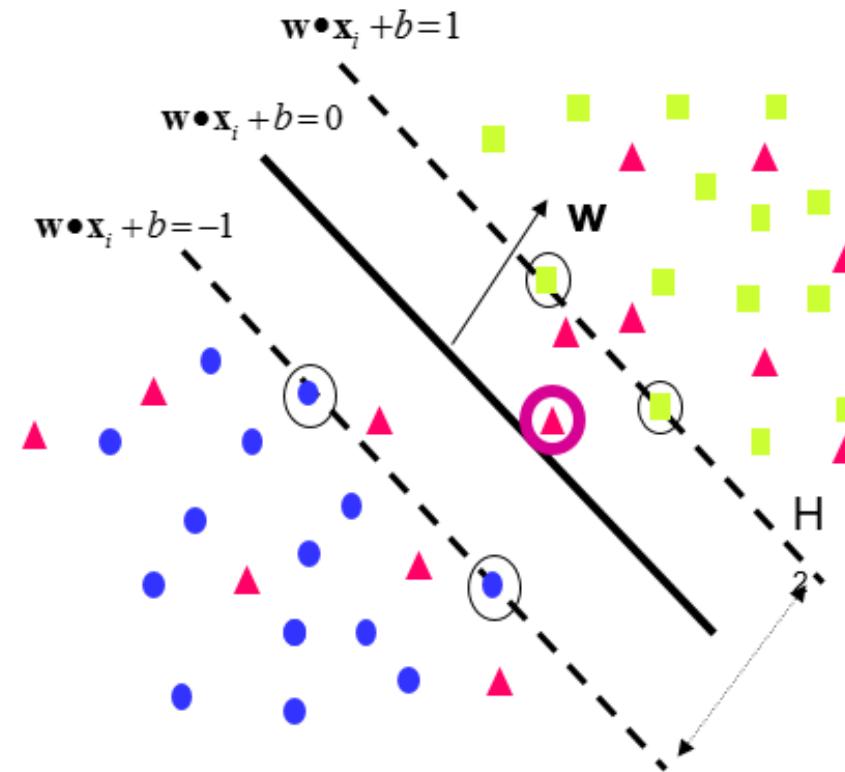
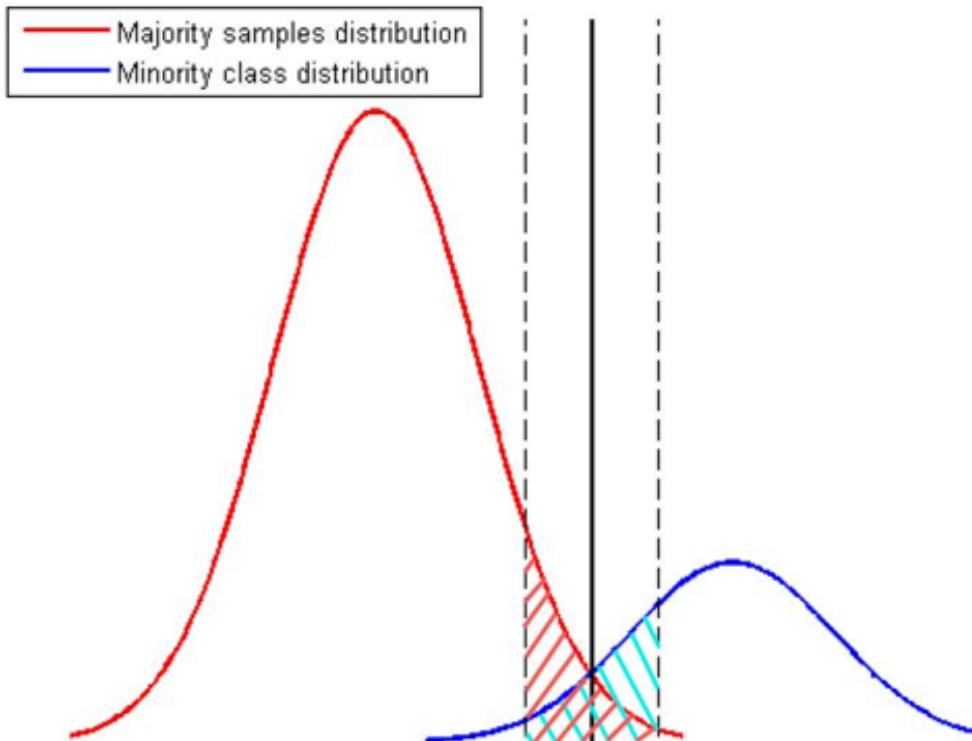
Class Imbalance Problem

The class boundary learned by standard machine learning algorithms can be severely skewed toward the positive class.



Machine learning algorithms try to maximize the classification accuracy.
But **accuracy on positive class (sensitivity)** may decrease!

Why AL works for Class Imbalance Problem



The imbalance ratio of the classes within the margin is much smaller than the class imbalance ratio of the entire dataset

(Ertekin et. al. CIKM)

g-means Comparison of Different Methods for Class Imbalance Problem

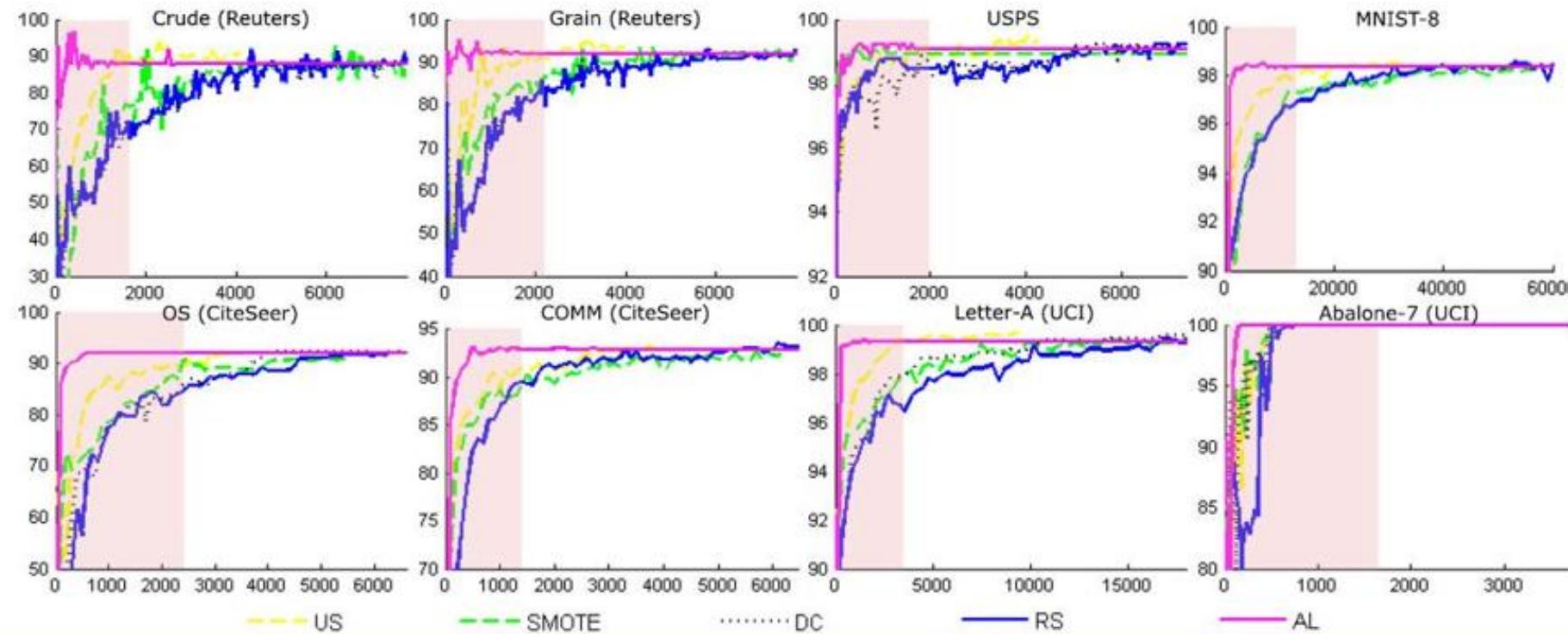
US → UnderSampling

SMOTE → OverSampling

DC → Distinct Costs to the classification errors

RS → Random Sampling

AL → Active Learning

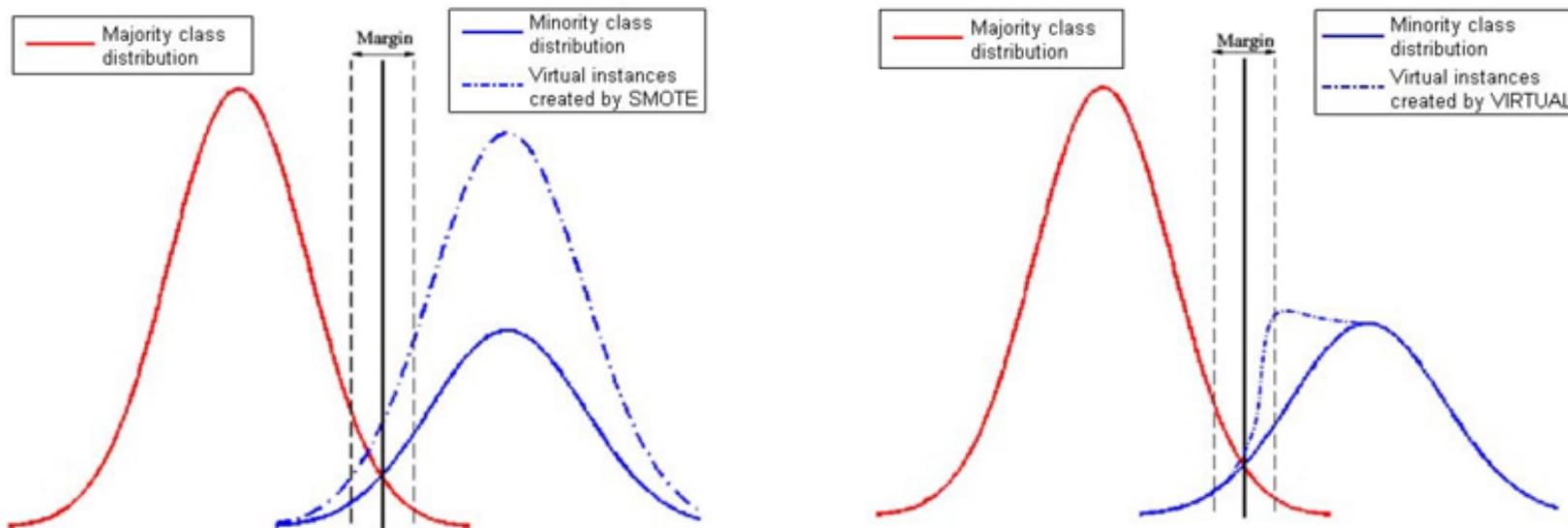


AL converges to the similar level of g-means in much earlier steps of the learning process.

Adaptive OverSampling with Active Learning

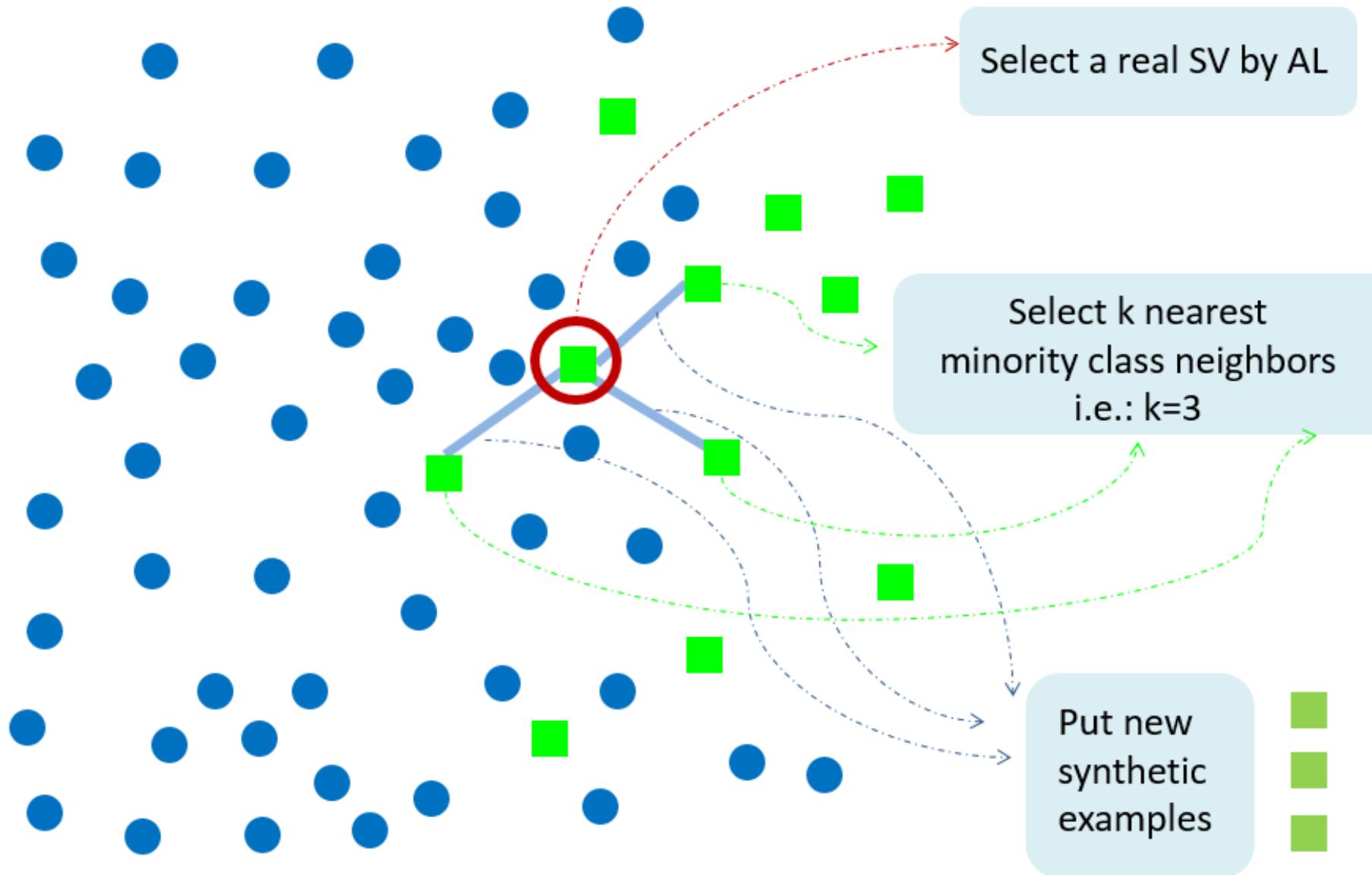
VIRTUAL

Virtual Instances Resampling Technique Using Active Learning

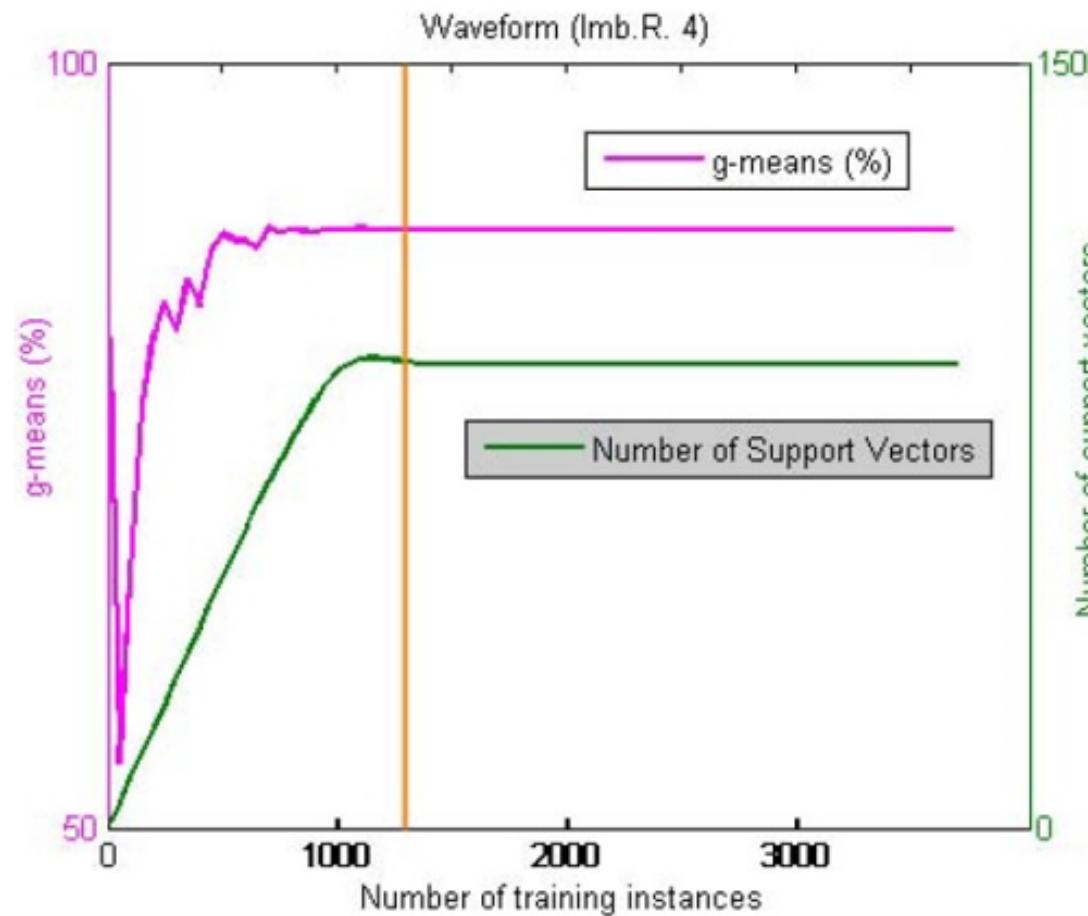


- Generates virtual (synthetic) examples for the minority class support vectors.
- No preprocessing is required. Synthetic examples can be created on the fly.
- Faster runtime than oversampling methods
- Better prediction performance than Active Learning (and oversampling)

VIRTUAL Example Generation

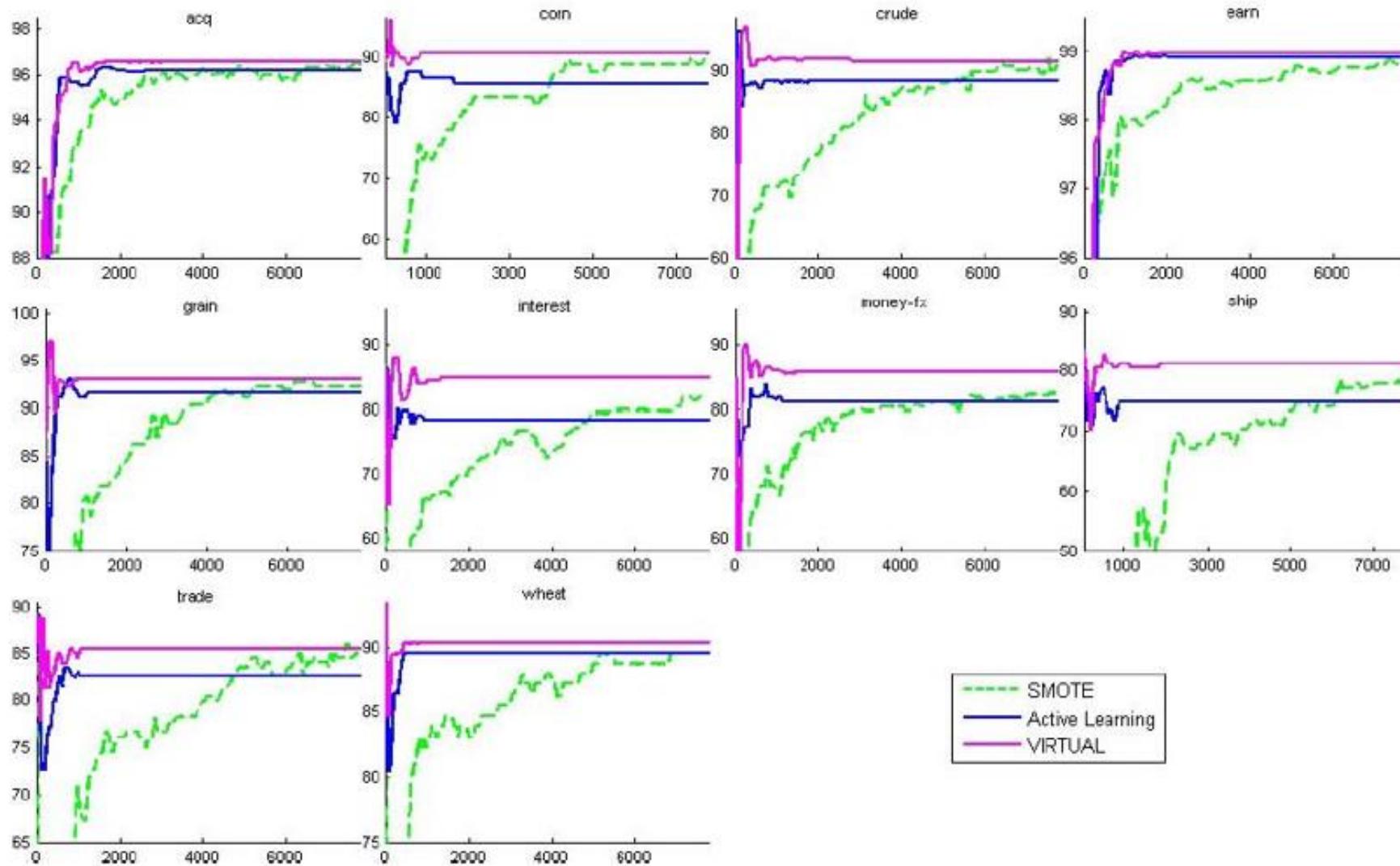


Saturation of g-means and Support Vectors



We can do early stopping!

Prediction Performance of SMOTE, AL and VIRTUAL



Noise in the Training Data

Noise and SVMs?

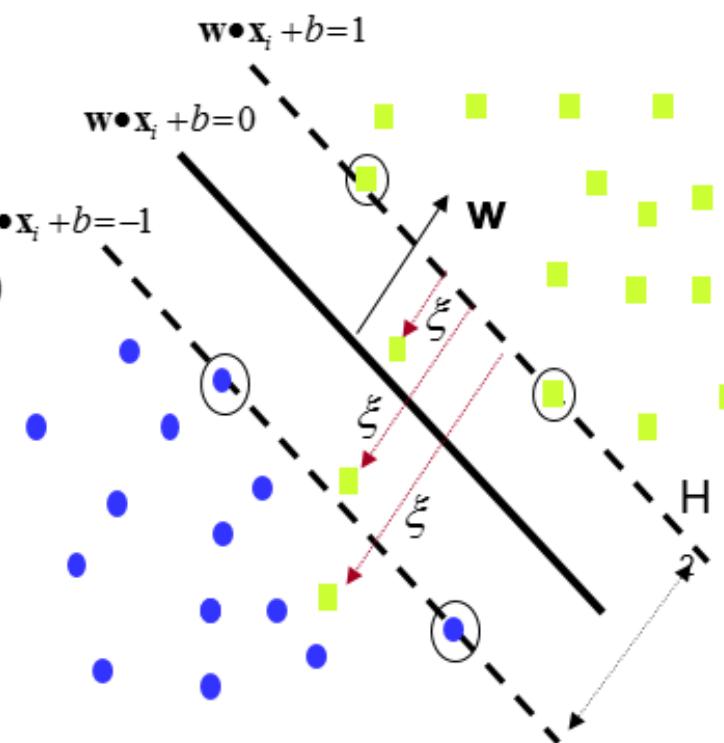
- In the presence of noise standard maximum margin algorithm can be subject to overfitting.

- Soft margin SVMs are introduced to cope with noise.

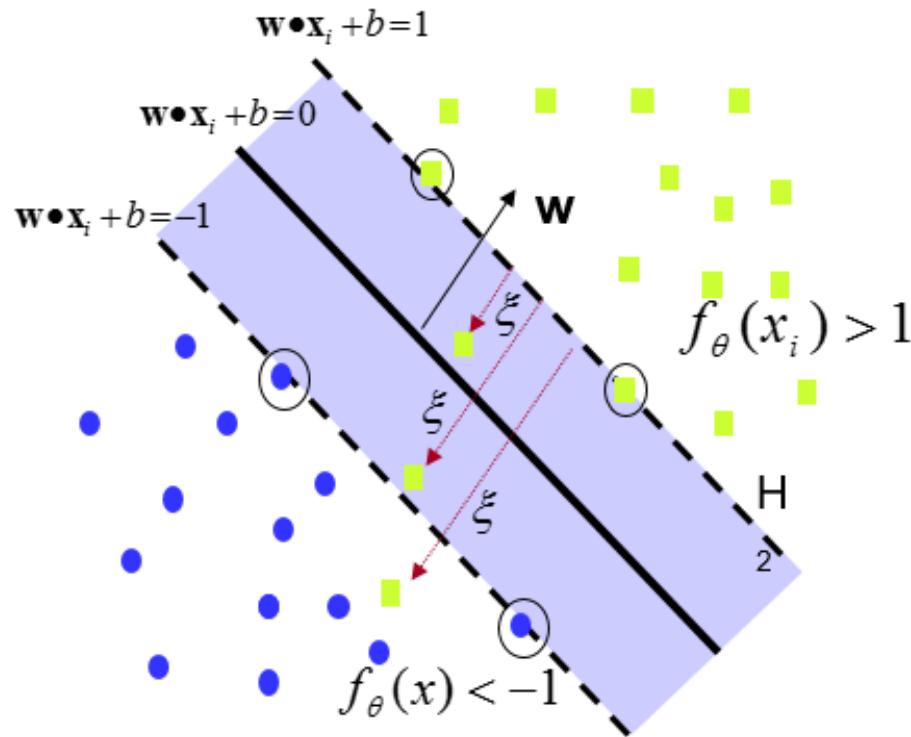
Inseparable case $\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l H(y_i f_\theta(x_i))$

Problem!

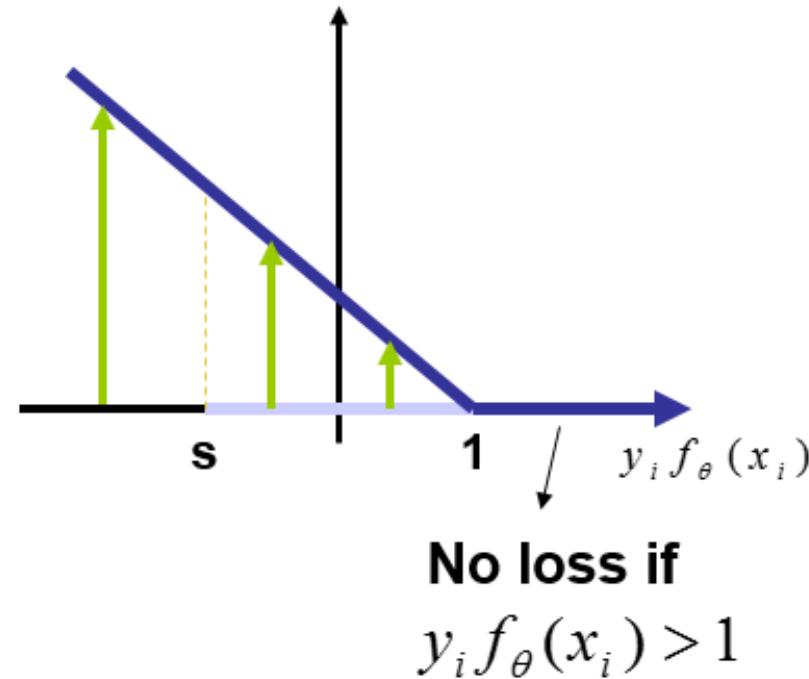
Misclassified examples (outliers) start to become more important than they should due to the Hinge Loss.



Inseparable case - Hinge Loss



$$H(y_i f_\theta(x_i)) = \max(0, 1 - y_i f_\theta(x_i))$$



No loss if
 $y_i f_\theta(x_i) > 1$

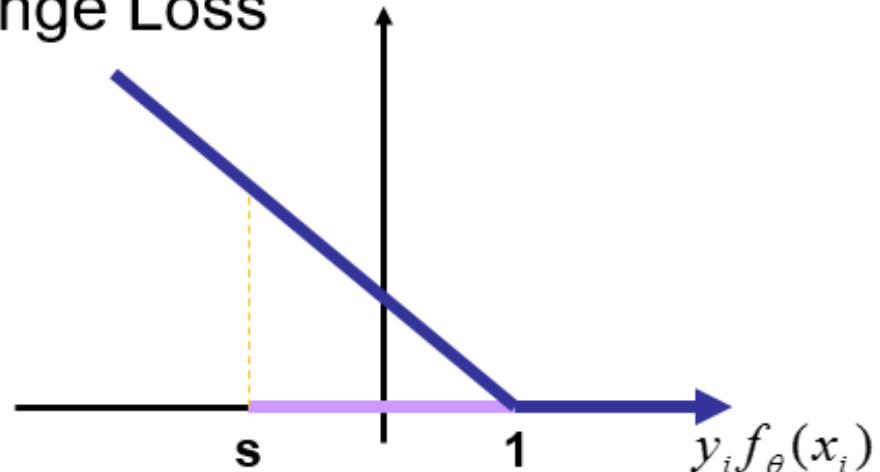
Misclassified examples become support vectors

Curse of dual variables:

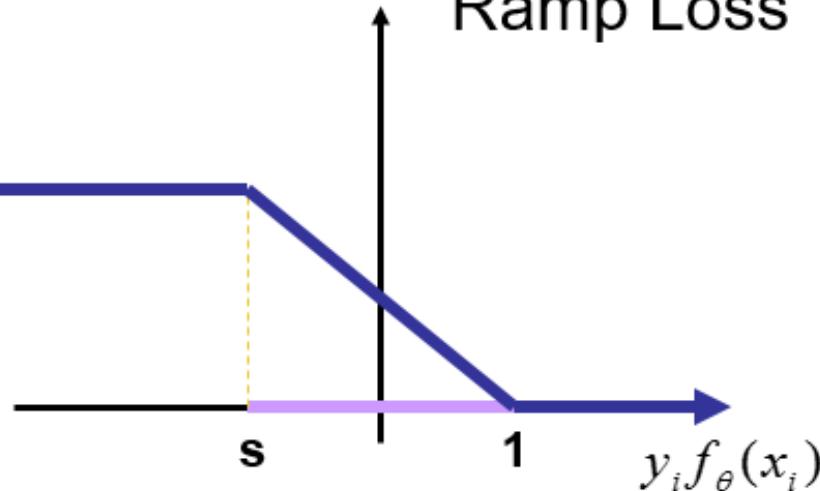
of SVs increases linearly with the # of training examples

Non-Convex Loss Function for SVMs

Hinge Loss



Ramp Loss



Cost Function

$$\rightarrow J^s(\theta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^L H_1(y_i f_\theta(x_i))$$

minimum \mathbf{w} must satisfy $\mathbf{w} = \sum_{i=1}^L -C y_i H_1'(y_i f_\theta(x_i)) \Phi(x_i)$

Examples located in the flat area can not become SVs because $H_1'(z)$ is zero.

Ramp Loss

Non-convex

Convex or Non-convex

Convexity is good

Can be mathematically analyzed & bounds can be produced

Guarantees to reach the global optimum

However,

Increase in the size of the model

Increase in the learning time

We show,

Non-convexity is good

Sparser models

Scalable solutions

Handles label noise well

3 variants of
LASVM

(S.Ertekin, L. Bottou, C.
L. Giles 'TPAMI 2011)

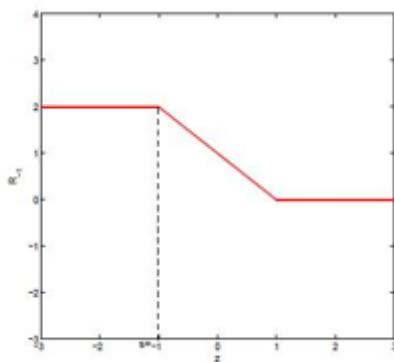
LASVM-G

LASVM-NC

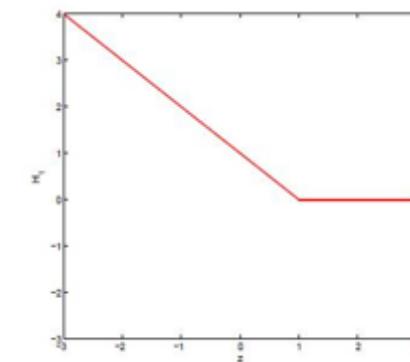
LASVM-I

Concave-Convex Procedure (CCCP)

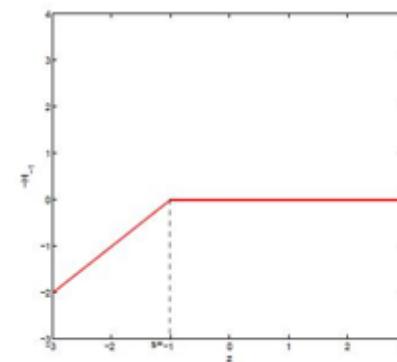
The Ramp Loss can be decomposed into
a Convex Hinge Loss and a Concave Loss



(a) Ramp Loss



(b) Convex Hinge Loss



(c) Concave Loss

Algorithm 1 The Concave-Convex Procedure (CCCP)

Initialize θ^0 with a best guess

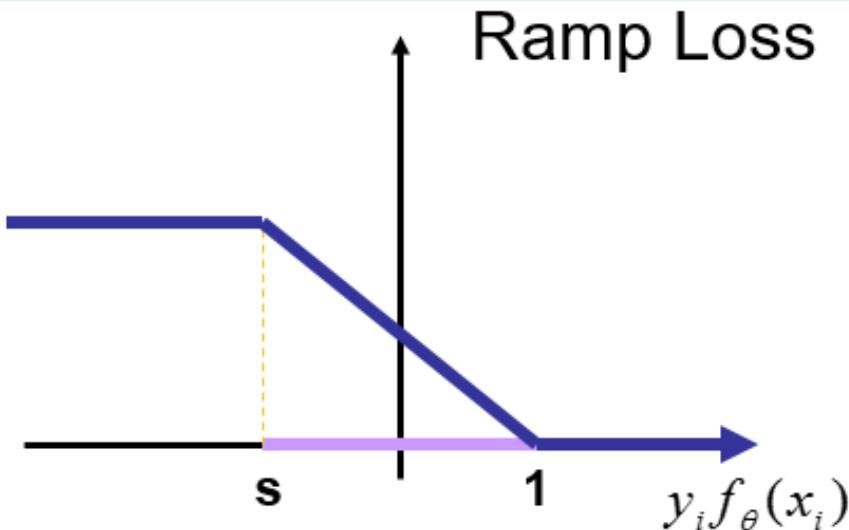
repeat

$$\theta^{t+1} = \arg \min_{\theta} (J_{vex}(\theta) + J'_{cav}(\theta^t) \cdot \theta)$$

until convergence of θ^t

Each iteration of CCCP approximates the concave part by its tangent and minimizes the resulting convex function.

Discussion

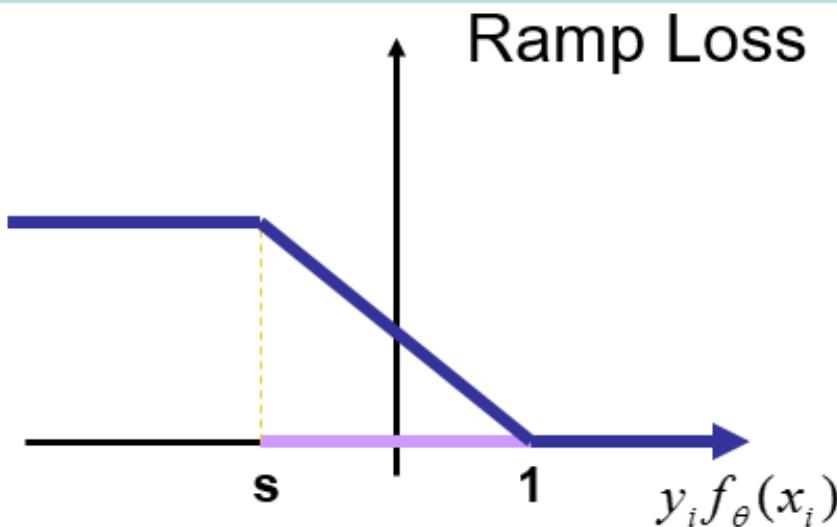


Two important values for s at noisy datasets:

$s=0$: Misclassified documents can not be support vectors.

$s=-1$: Only examples in the margin are processed.

Discussion



Two important values for s at noisy datasets:

$s=0$: Misclassified documents can not be support vectors.

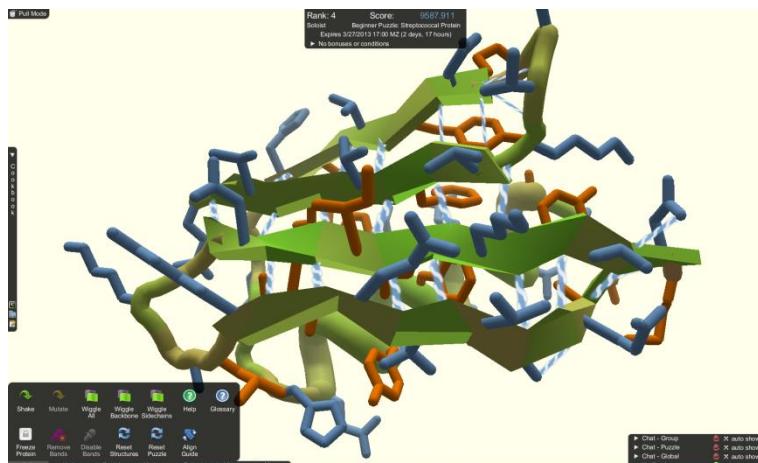
$s=-1$: Only examples in the margin are processed.

$s=-1$

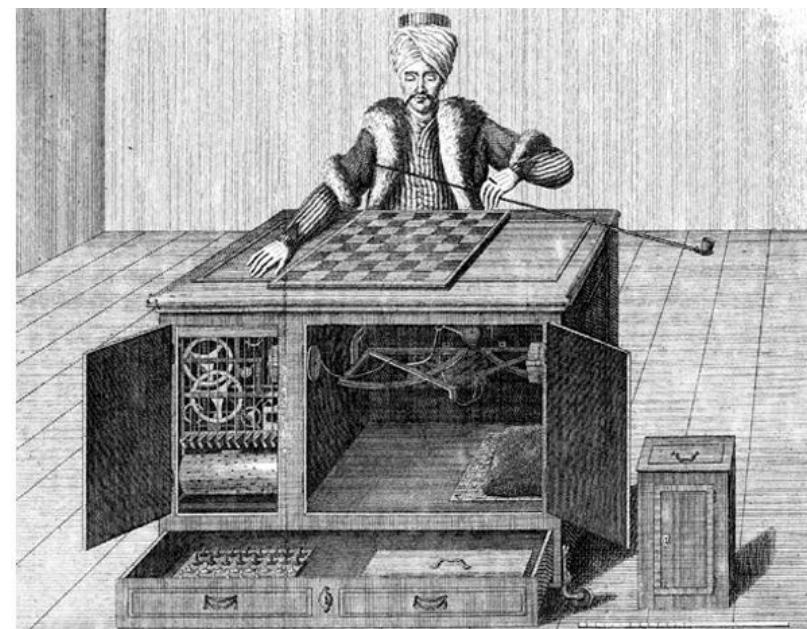
Works like Active Learning!

Machine Learning with Humans in the Loop

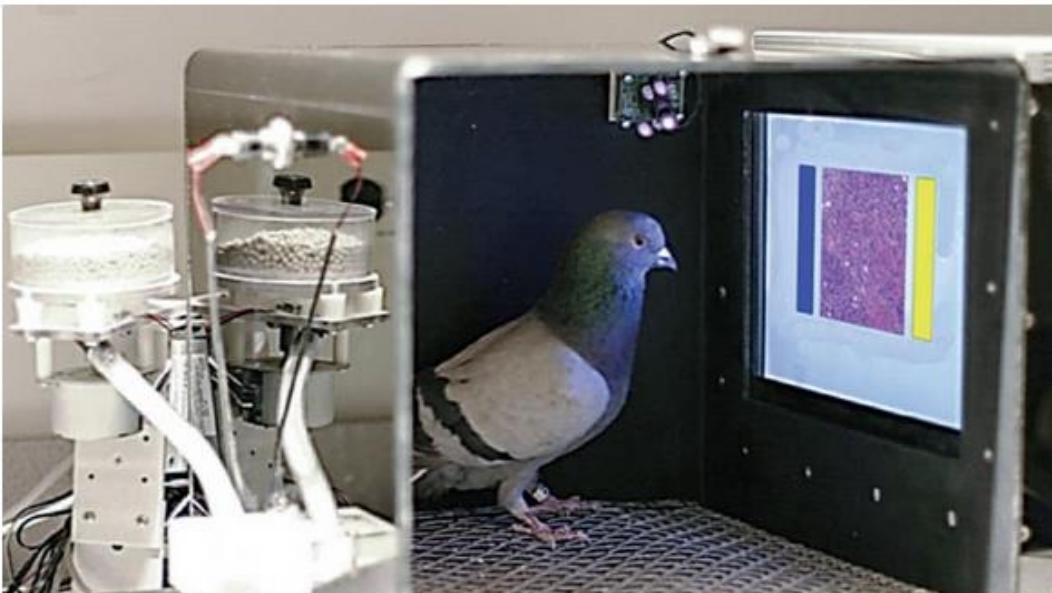
Ex: analyzing
microscopy images



Ex: FoldIt Game



SHARE



A flock of pigeons was able to correctly spot cancer in breast tissue biopsy images with 99% accuracy, on par with human experts.

R. M. Levenson et al., *PLOS ONE* (18 November 2015)

Pigeons spot cancer as well as human experts

By John Bohannon | Nov. 18, 2015, 2:00 PM

It may sound like a bird-brained idea, but scientists have trained pigeons to spot cancer in images of biopsied tissue. Individually, the avian analysts can't quite match the accuracy of professional pathologists. But as a flock, they did as well as trained humans, according to a new study appearing this week in *PLOS ONE*.

FOLLOW NEWS FROM SCIENCE



Advertisemen

Crowdsourcing

Main Challenges:

- Limited budget (each vote may have a cost)
- Labelers with range of capabilities, motives, knowledge and views

Approximating the Crowd

Determining the majority vote on a budget

- Approximating the Crowd**

Seyda Ertekin, Cynthia Rudin, Haym Hirsh.

Data Mining and Knowledge Discovery, 2014.

Approximating the Crowd

Determining the majority vote on a budget



Approximating the Crowd

Determining the majority vote on a budget



Approximating the Crowd

Determining the majority vote on a budget



**Can we guess the majority vote of the crowd without asking everyone?
Can we do this “online” when the items arrive one-by-one ?**



CrowdSense



Key Ideas of CrowdSense

I. Balance of exploration and exploitation

II. Weighted vote over labelers. Higher quality
labelers have higher weights

III. Build up the set of labelers
dynamically for each item.

Key Ideas of CrowdSense

I. Balance of exploration and exploitation.

- Exploitation: Trust the “best” labelers.
- Exploration: Give a chance to each labeler to be selected.

Key Ideas of CrowdSense

II. Weighted vote over labelers. Higher quality
labelers get higher weights

- Weight = labeler accuracy

Q_{it} : The weight of labeler i after seeing item t

c_{it} : How many times we asked i about items

a_{it} : How many times i was “right”

$$Q_{it} = \frac{a_{it} + K}{c_{it} + 2K}$$

Key Ideas of CrowdSense

III. Build up the set of labelers
dynamically for each item

Key Ideas of CrowdSense

III. Build up the set of labelers
dynamically for each item



- Keep adding more labelers until we are certain.
- Start with 3 labelers



Outline of CrowdSense

For $t = 1, \dots, T$

end

Outline of CrowdSense

For $t = 1, \dots, T$

-Start with 3 labelers

- Exploitation: Pick 2 based on quality estimates
- Exploration: Pick 1 uniformly at random

end

Outline of CrowdSense

For $t = 1, \dots, T$

-Start with 3 labelers

- Exploitation: Pick 2 based on quality estimates
- Exploration: Pick 1 uniformly at random

- Get their votes

end

Outline of CrowdSense

For $t = 1, \dots, T$

-Start with 3 labelers

- Exploitation: Pick 2 based on quality estimates
- Exploration: Pick 1 uniformly at random

- Get their votes

- Keep adding labelers and getting their votes until we are confident with the prediction

end

Outline of CrowdSense

For $t = 1, \dots, T$

-Start with 3 labelers

- Exploitation: Pick 2 based on quality estimates
- Exploration: Pick 1 uniformly at random

- Get their votes

- Keep adding labelers and getting their votes until we are confident with the prediction

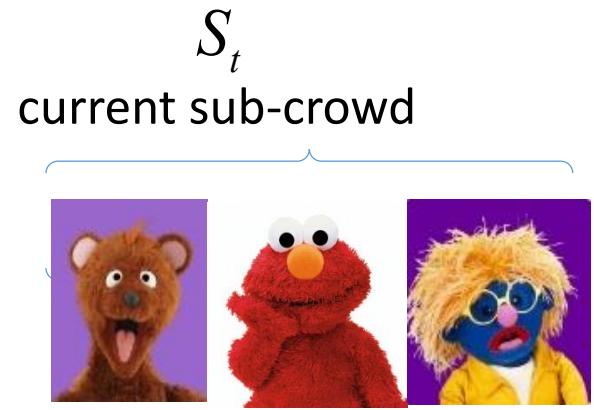
– Combine votes $\left(\text{sign} \sum_{i \in S_t} V_{it} Q_{it}\right)$ where V_{it} is the vote of labeler i for item t , $V_{it} \in \{-1, +1\}$.

end

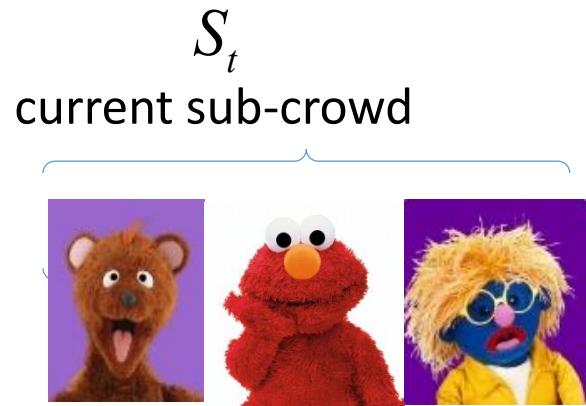


... until we are
confident with the
prediction ?

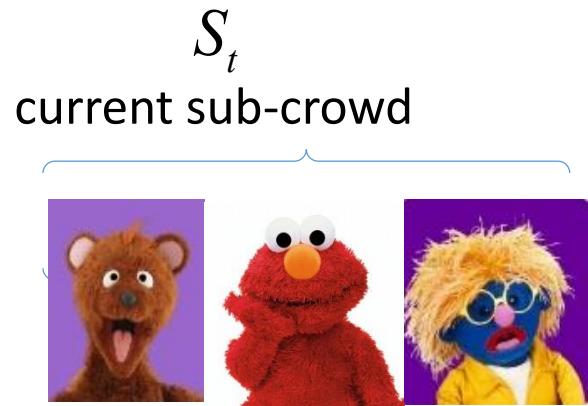
- Keep adding labelers and getting their votes until we are
confident with the prediction.



If the next best labeler has enough weight to change the vote (or come close), add it in.

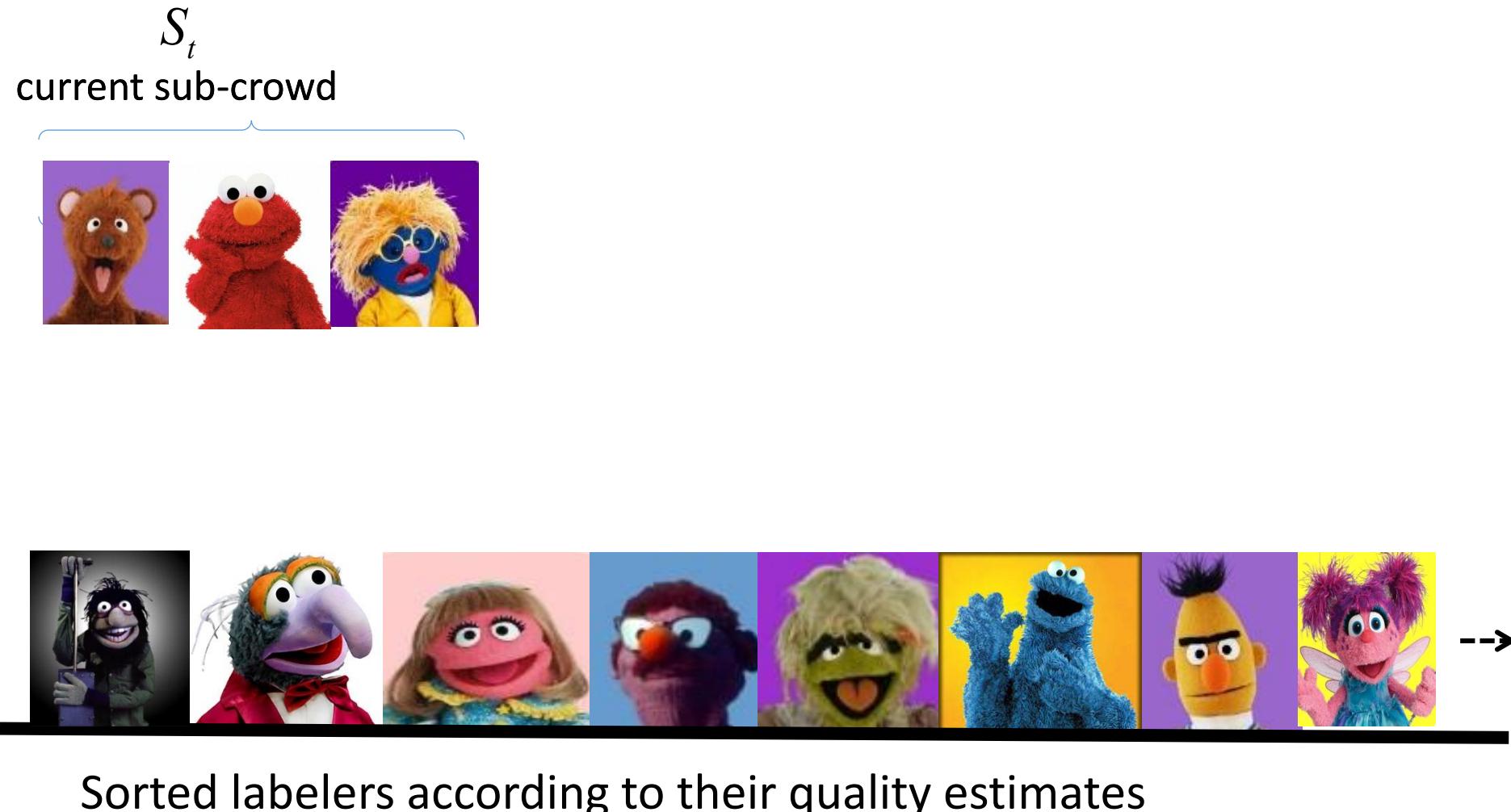


If the next best labeler has enough weight to change the vote (or come close), add it in.

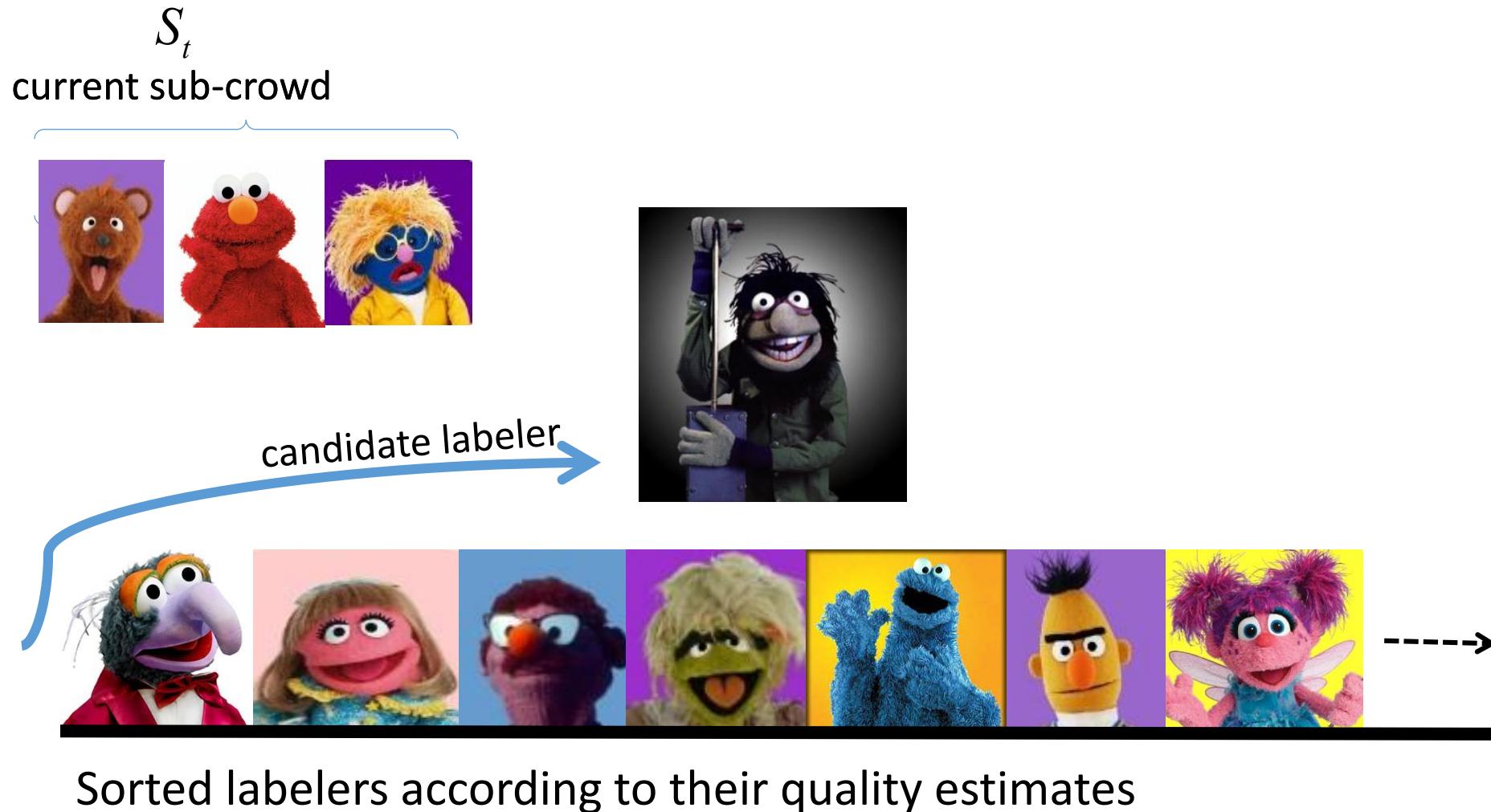


$$\text{Score}(S_t) = \sum_{i \in S_t} V_{it} Q_{it}$$

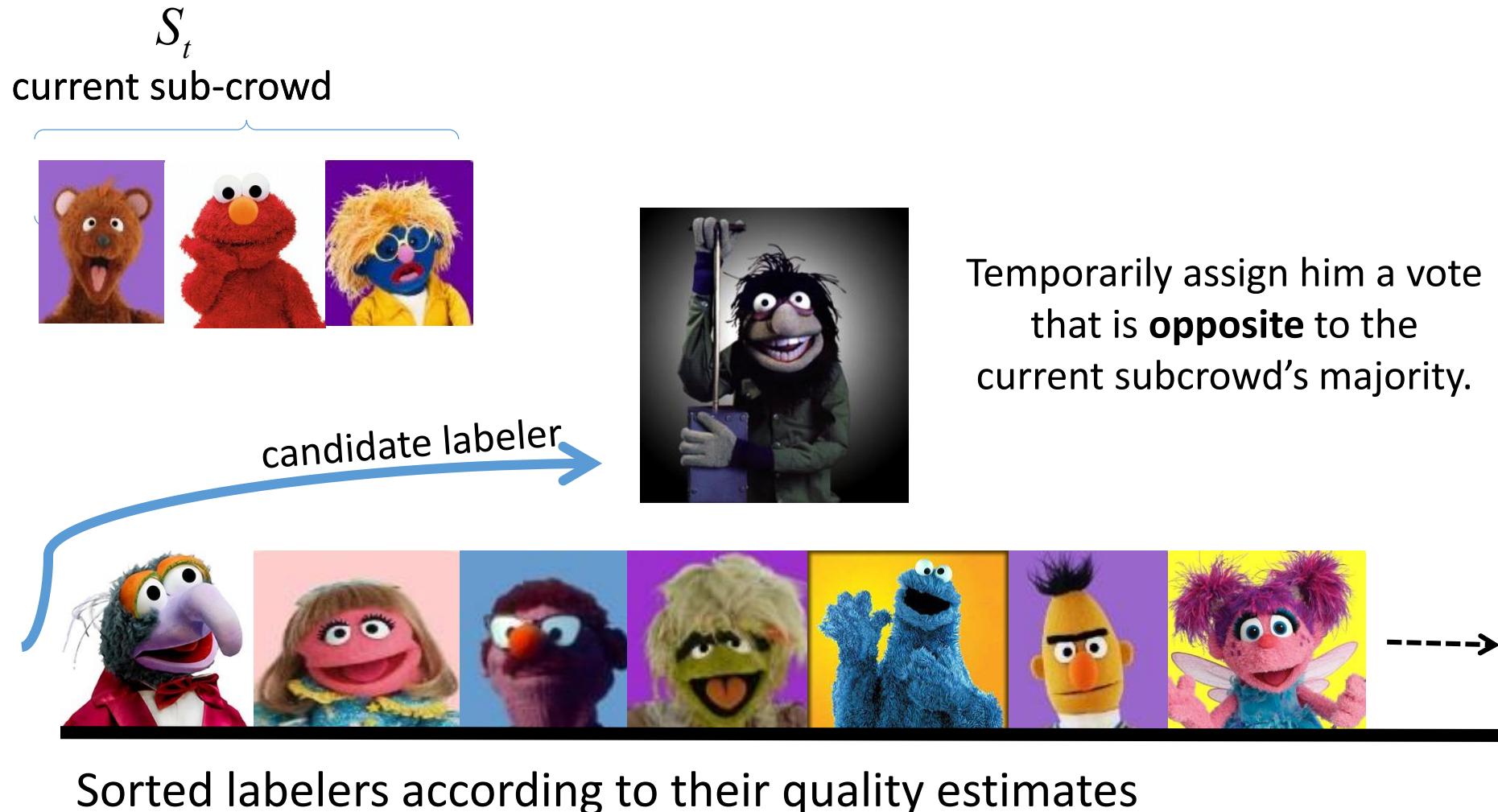
If the next best labeler has enough weight to change the vote (or come close), add it in.



If the next best labeler has enough weight to change the vote (or come close), add it in.



If the next best labeler has enough weight to change the vote (or come close), add it in.



If the next best labeler has enough weight to change the vote (or come close), add it in.

$$\frac{|\text{Score}(S_t)| - Q_{l_{\text{candidate}}, t}}{|S_t| + 1} < \varepsilon$$



$$\text{Score}(S_t) = \sum_{i \in S_t} V_{it} Q_{it}$$



Temporarily assign him a vote
that is **opposite** to the
current subcrowd's majority.

$$\frac{|\text{Score}(S_t)| - Q_{l_{\text{candidate}}, t}}{|S_t| + 1} < \varepsilon$$

ε is a tradeoff between the total cost we are willing to spend and the desired accuracy.

High ε : more exploration
Low ε : less exploration

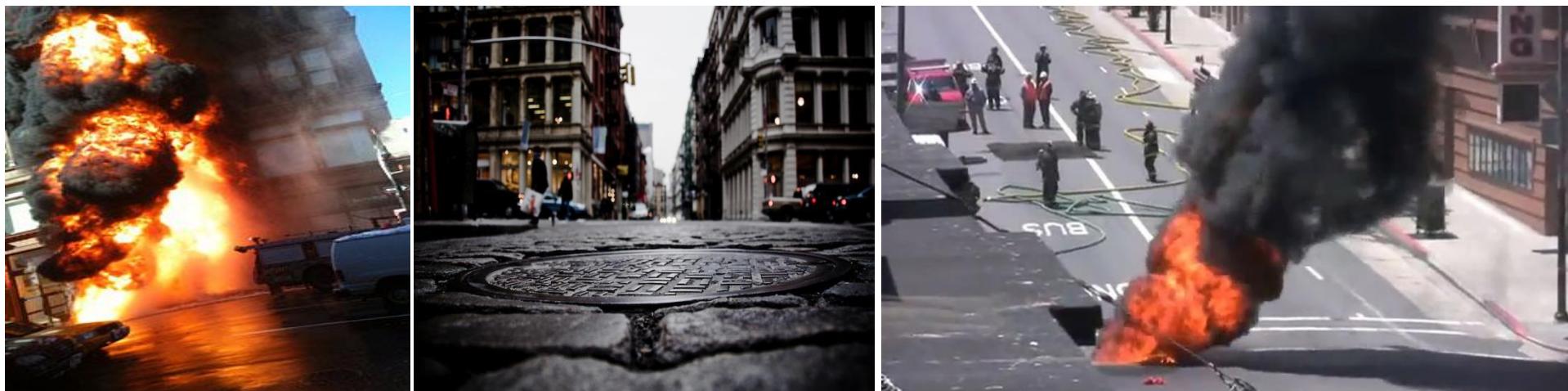
A Big Data Project: Predicting Serious Events in Energy Grids

**Reactive Point Processes: A New Approach to Predicting Power Failures
in Underground Electrical Systems.**

Seyda Ertekin, Cynthia Rudin, Tyler McCormick.
Annals of Applied Statistics, 2015.



First prize winner of the INFORM's
Innovative Applications in Analytics Award



Reactive Point Processes (RPP)



February 11th, 2010 Manhattan, NY
<http://www.inquisitr.com>



January 27th, 2013 Omaha, NE
<http://www.theblaze.com>

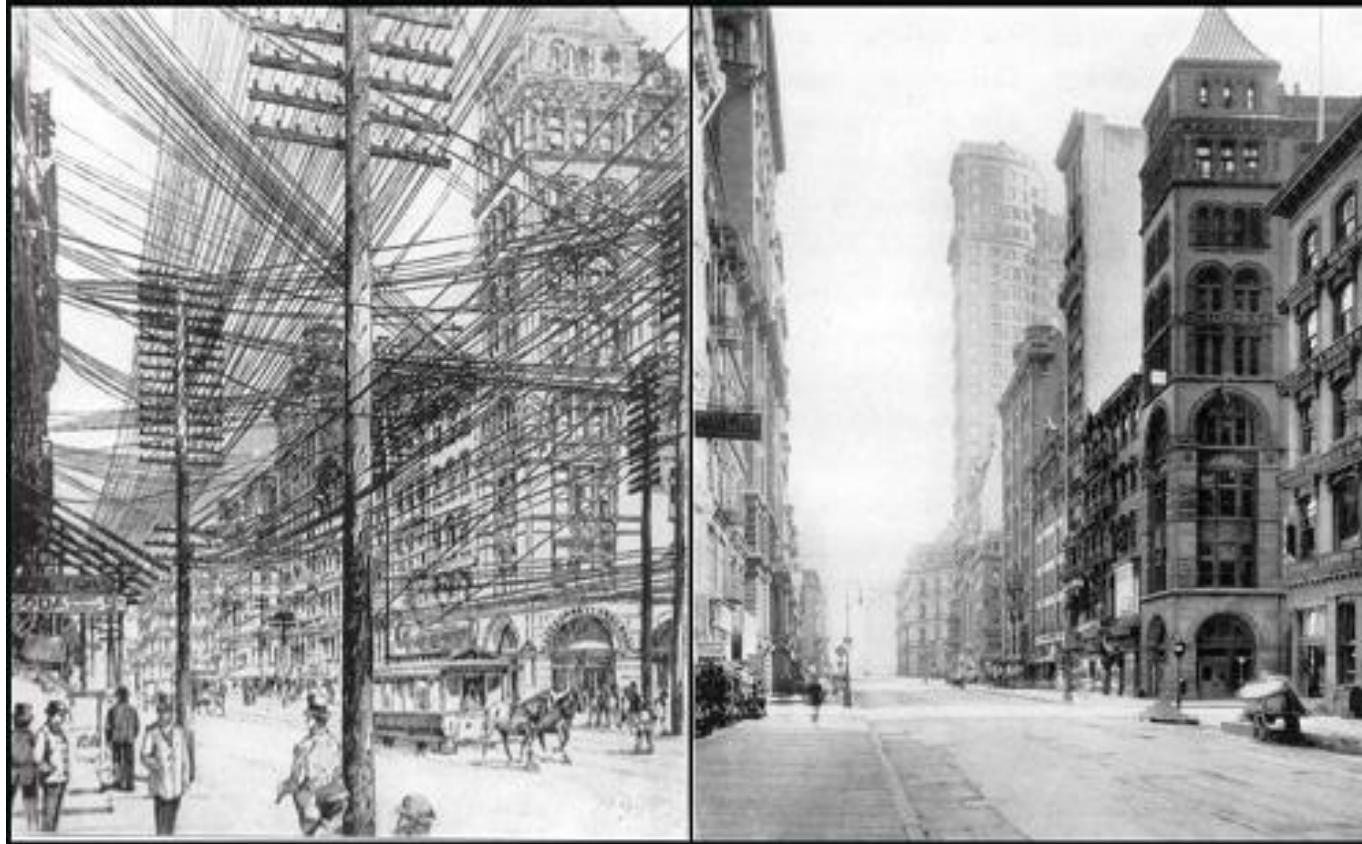


June 5th, 2009 San Francisco, CA
<http://www.sfgate.com>

A new statistical model designed for predicting discrete events* in time based on the past history

*fires, explosions & power failures

Underground Electric Grid

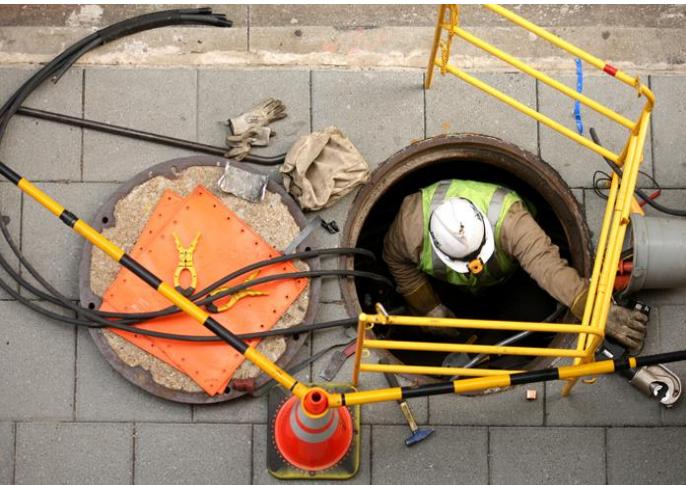


Operational Challenges



More than 53K manholes in Manhattan

Power companies are switching:
from **reactive** maintenance plans
to **proactive** maintenance plans.



Operational challenges for power companies:

- i) Can we predict failures continuously in time?
- ii) Can we model how quickly the influence of past events and inspections fade over time?
- iii) Can we develop a cost/benefit analysis for proactive maintenance policies?

Features of the Manhole

Manhole is represented as: [5 3 120 12 1 0]

Number of events last year
Number of serious events last year
Number of electrical cables
Number of pre-1930 electrical cables
Vented cover?
Inspected?

Each observation is represented by a set of numbers (features).

Manhole is represented as:

[5	3	120	12	1	0]	-1
[0	0	89	5	1	1]	1
[1	0	20	0	0	1]	-1

:

:



Features, called X



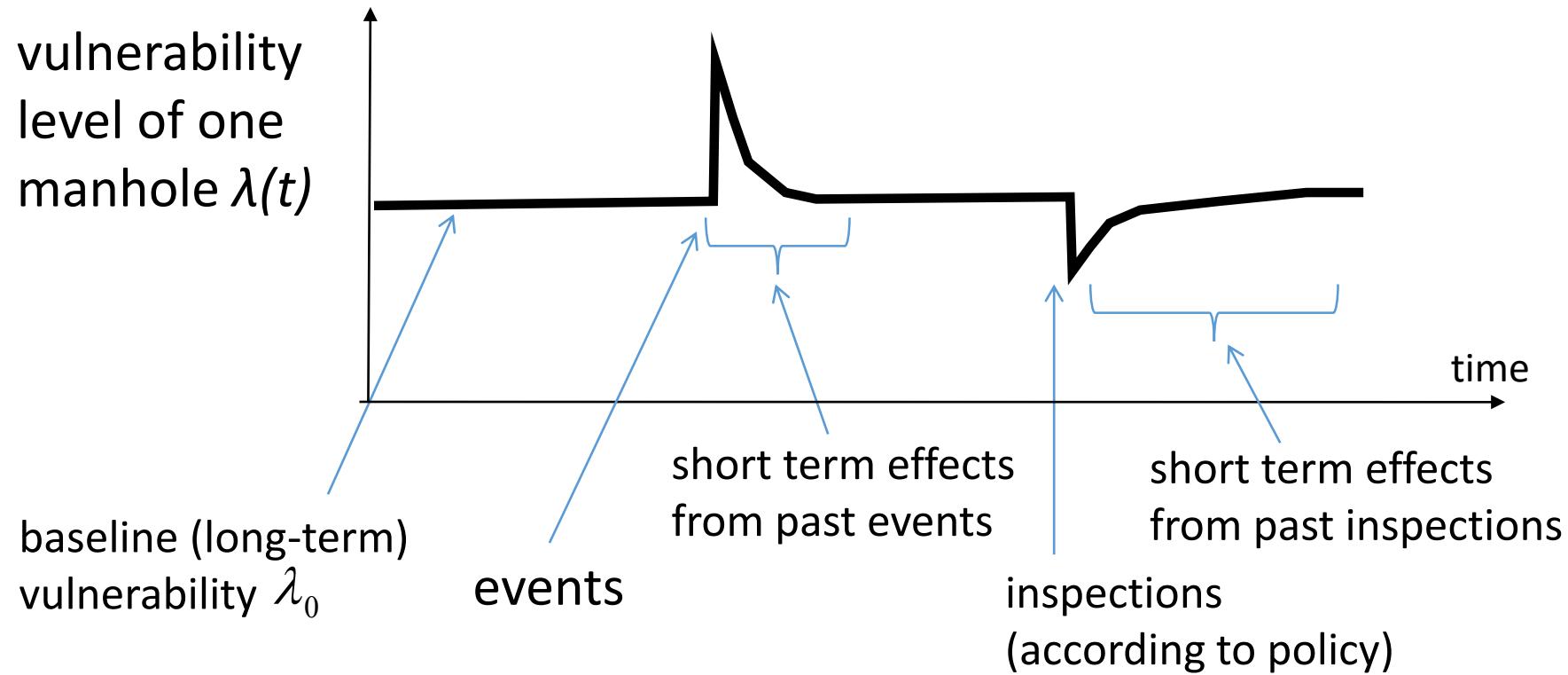
Labels, called Y

(Predictors, Covariates,
Explanatory Variables,
Independent Variables)

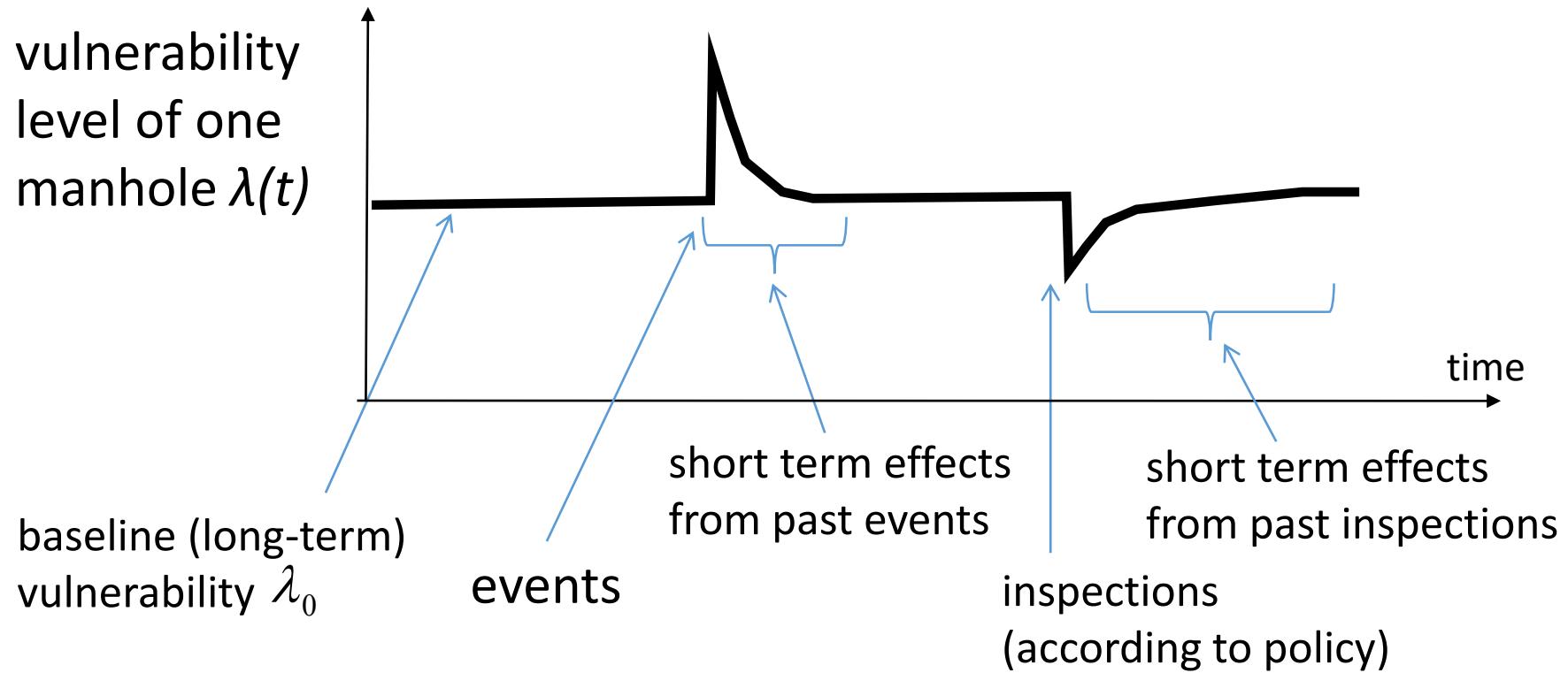
Properties of Reactive Point Processes (RPPs)

- Self-exciting property (captures “**events**”)
- Self-regulating property (captures “**inspections**”)
- Diminishing returns (captures **cumulative effect** of events and inspections)
- Baseline vulnerability can change if there is at least one past event.
- Vulnerability between similar entities should be similar (Adaptive excitation and regulation)

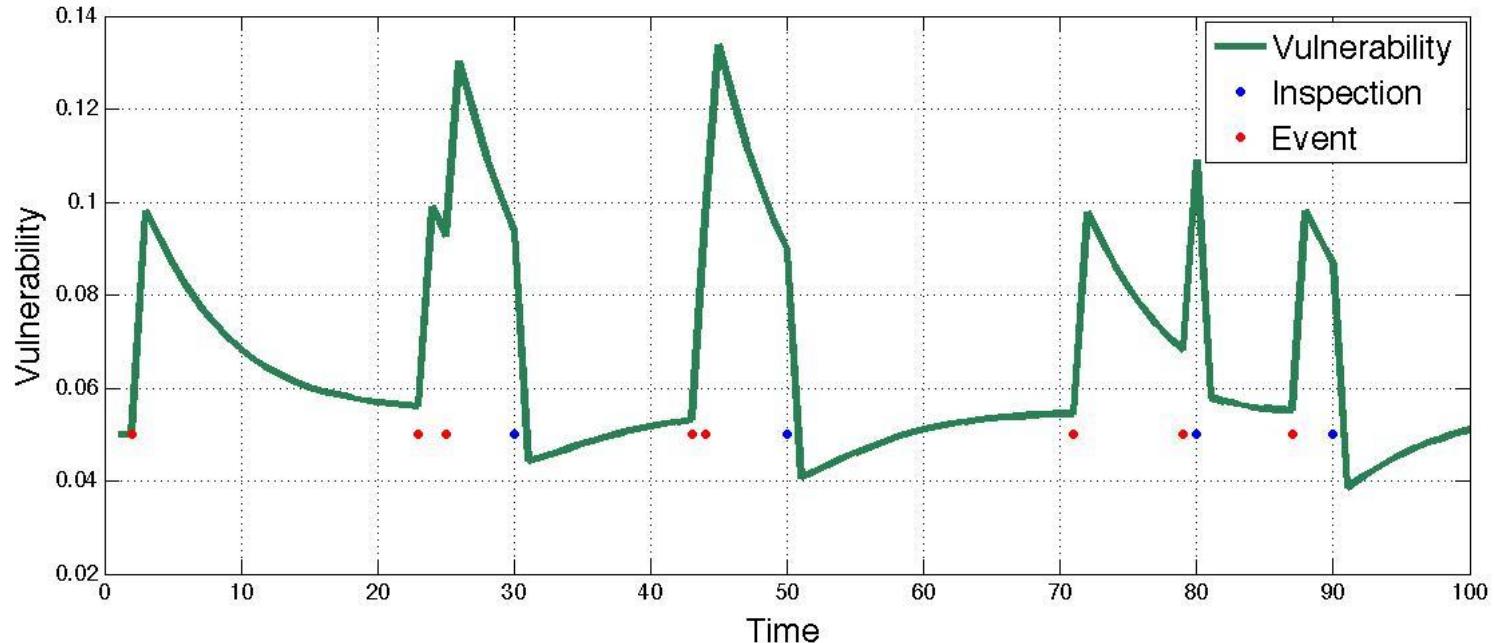
Short-term Vulnerability Model



Short-term Vulnerability Model



$$\lambda(t) = \lambda_0 \left[1 + g_1 \left(\sum_{\forall t_e < t} g_2(t - t_e) \right) - g_3 \left(\sum_{\forall \bar{t}_i < t} g_4(t - \bar{t}_i) \right) + C_1 \mathbf{1}_{[N_E \geq 1]} \right]$$



$$\lambda(t) = \lambda_0 \left[1 + g_1 \left(\sum_{\forall t_e < t} g_2(t - t_e) \right) - g_3 \left(\sum_{\forall \bar{t}_i < t} g_4(t - \bar{t}_i) \right) + C_1 \mathbf{1}_{[N_E \geq 1]} \right]$$


 baseline (long-term) event probability
 short term effects from past events
 short term effects from past inspections
 long-term effects from at least one past event

RPP statistical models

Nonparametric

Likelihood-based

Bayesian

Nonparametric Model: Conditional Frequency (CF) Estimator

The idea is:

- Estimate the parameters (λ_0 and C1) and the functions (g2, g4 and g1, g3) in a case-controlled way.

To do this:

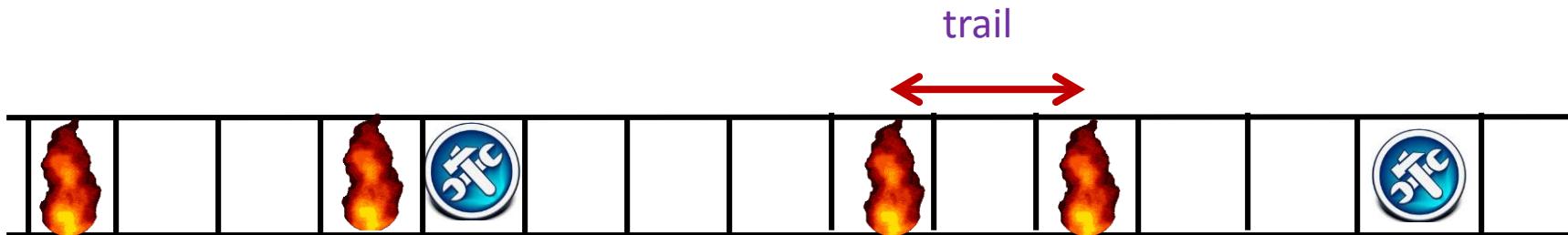
- Locate event/inspection “**trails**” in the time series data for each manhole.

Trail: A sequence of time steps for a specific manhole where a predefined pattern of inspection/event record exists.

Tracing out g2 function

$$\lambda(t) = \lambda_0 \left[1 + \cancel{g_1} \left(\sum_{\forall t_e < t} g_2(t - t_e) \right) - \cancel{g_3} \left(\sum_{\forall \bar{t}_i < t} g_4(t - \bar{t}_i) \right) + C_1 \mathbf{1}_{[N_E \geq 1]} \right]$$

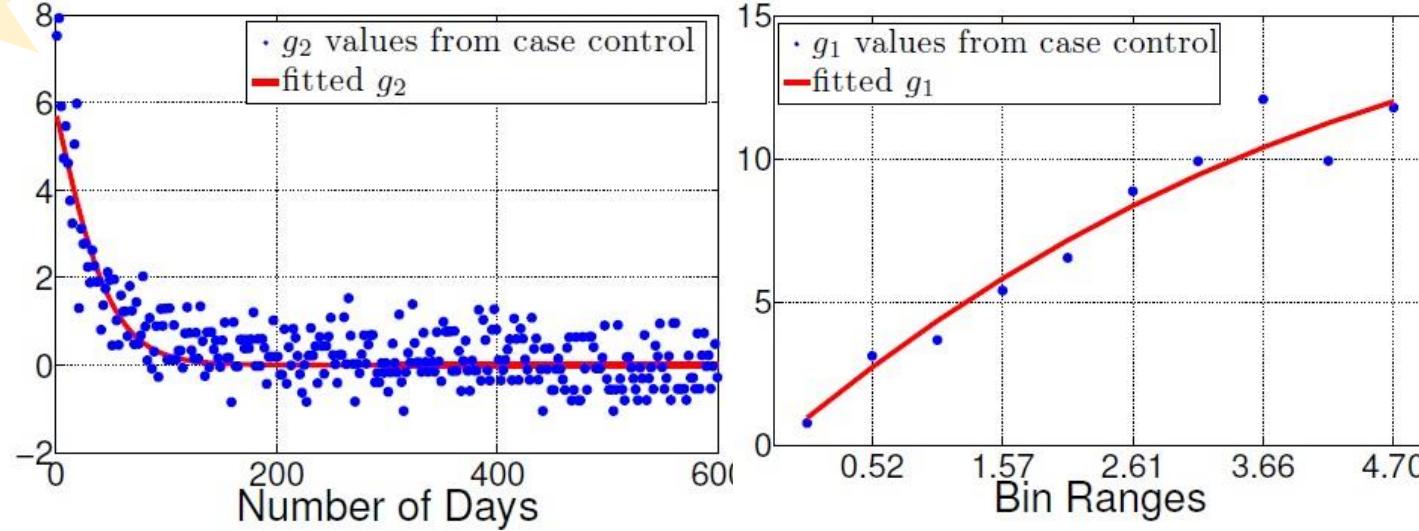
- Choose trails starting at events that occurred at baseline vulnerability levels (no effect of previous event or inspection)
- Consider only the trails where there was already a previous serious event in that manhole (to use C1)



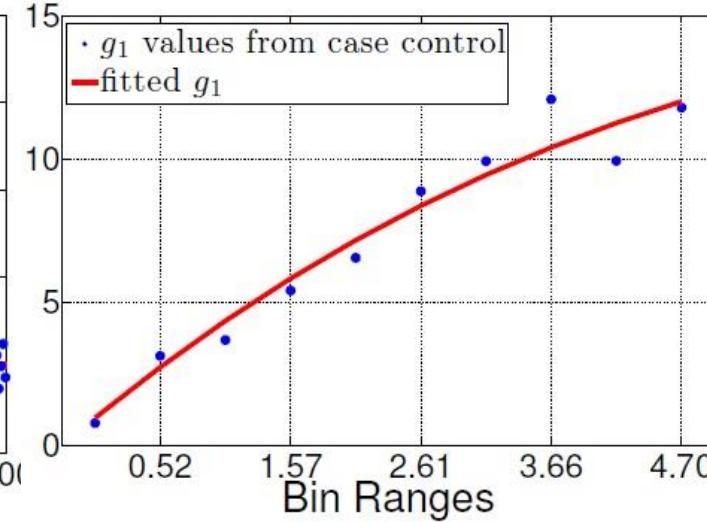
$$g_2^{\text{nonparametric}}(\Delta t) = \frac{1}{\lambda_0} \frac{E_{\Delta t}}{E_{\Delta t} + \bar{E}_{\Delta t}} - C_1 - 1$$

Fitted g_2 and g_1 functions (Nonparametric Model)

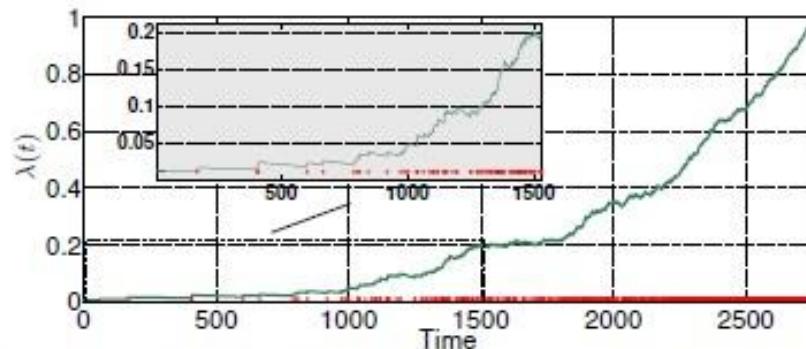
g_2



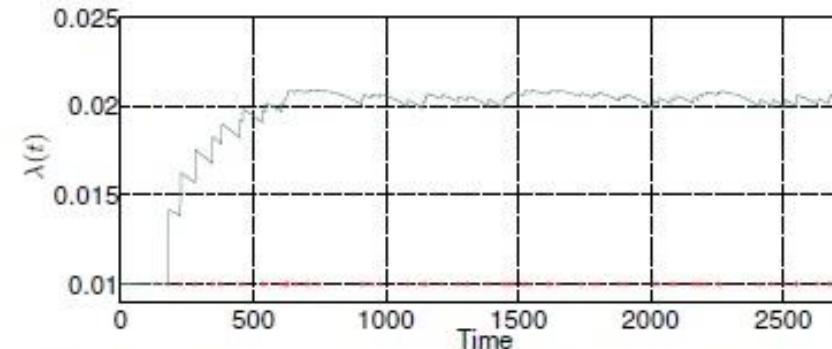
g_1



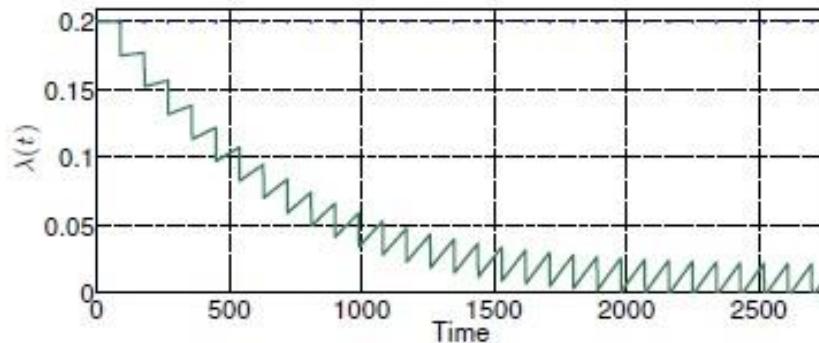
The effect of Saturation Functions, g_1 and g_3



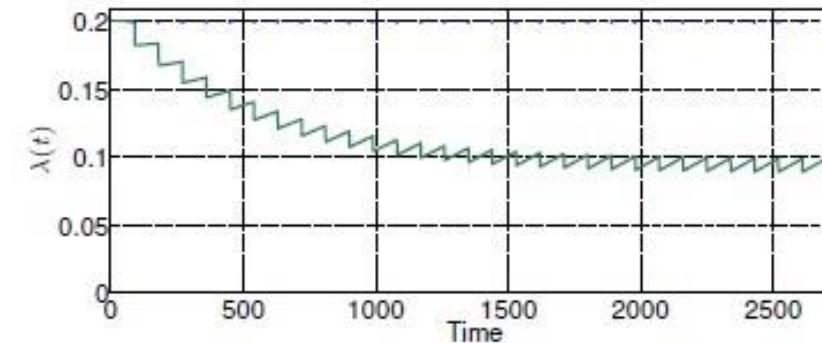
(a) Model with self-excitation function g_2 (without saturation function g_1 and inspections) and a zoomed view of the first 1,500 days.



(b) Model with self-excitation function g_2 and saturation function g_1 (no inspections).



(c) Model with self-regulation function g_4 (without saturation function g_3 and events).



(d) Model with self-regulation function g_4 and saturation function g_3 (no events).

Incorporating covariates

$$g_2^{(p)}(x) = \frac{1}{1 + e^{\beta^{(p)}x}}$$

$$g_4^{(p)}(x) = \frac{-1}{1 + e^{-\gamma^{(p)}x}}$$

$$\boldsymbol{\beta} = \log(1 + e^{-\mathbf{M}\boldsymbol{v}})$$

$$\boldsymbol{\gamma} = \log(1 + e^{-\mathbf{M}\boldsymbol{\omega}})$$

This permits slower or faster decay of the self-exciting and self-regulating components based on the characteristics of the manhole.

Incorporating covariates

$$g_2^{(p)}(x) = \frac{1}{1 + e^{\beta^{(p)}x}}$$

$$g_4^{(p)}(x) = \frac{-1}{1 + e^{-\gamma^{(p)}x}}$$

$$\boldsymbol{\beta} = \log(1 + e^{-\mathbf{M}\boldsymbol{v}})$$

$$\boldsymbol{\gamma} = \log(1 + e^{-\mathbf{M}\boldsymbol{\omega}})$$

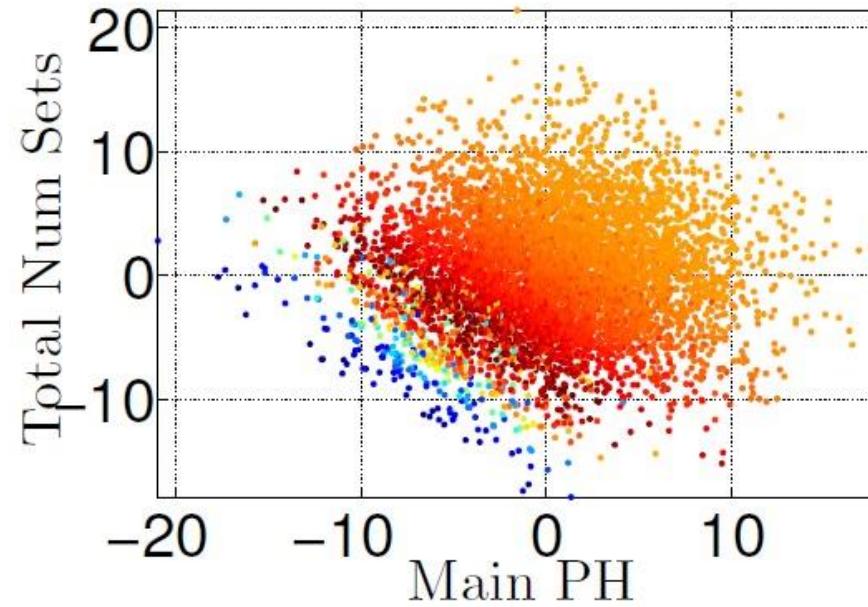
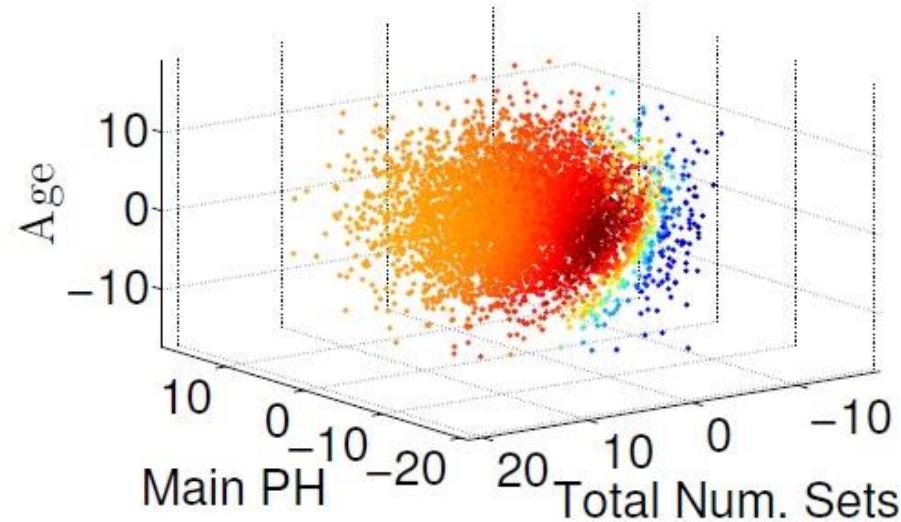
This permits slower or faster decay of the self-exciting and self-regulating components based on the characteristics of the manhole.

Covariates
we used:

- Number of main phase cables
- Age of the oldest cable set
- Total number of cable sets

Likelihood-based RPP

$$\log \mathcal{L} \left(\left\{ t_1^{(p)}, \dots, t_{N_E^{(p)}}^{(p)} \right\}_p ; \boldsymbol{v}, a_1, \mathbf{M} \right) = \sum_{p=1}^P \left[\sum_{e=1}^{N_E^{(p)}} \log(\lambda_p(t_e^{(p)})) - \int_0^{T_{\max}} \lambda_p(u) du \right]$$



Bayesian RPP

For Approximate Bayesian Computation (ABC),
we need two things:

- To be able to simulate from the model
- A summary statistic

How ABC works:

- Randomly choose proposed parameter values
- Use those values to generate simulated data using the model
- Then compare the generated and the observed data using a summary statistic.

Bayesian RPP

For Approximate Bayesian Computation (ABC),
we need two things:

- To be able to simulate from the model
- A summary statistic



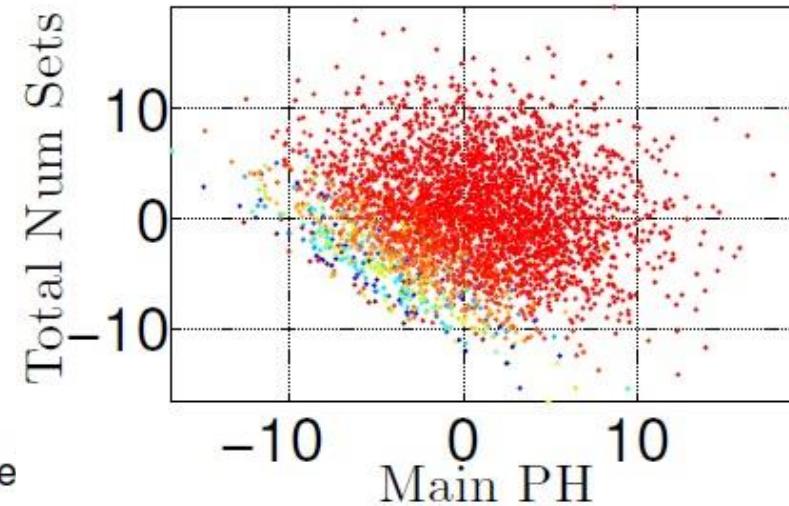
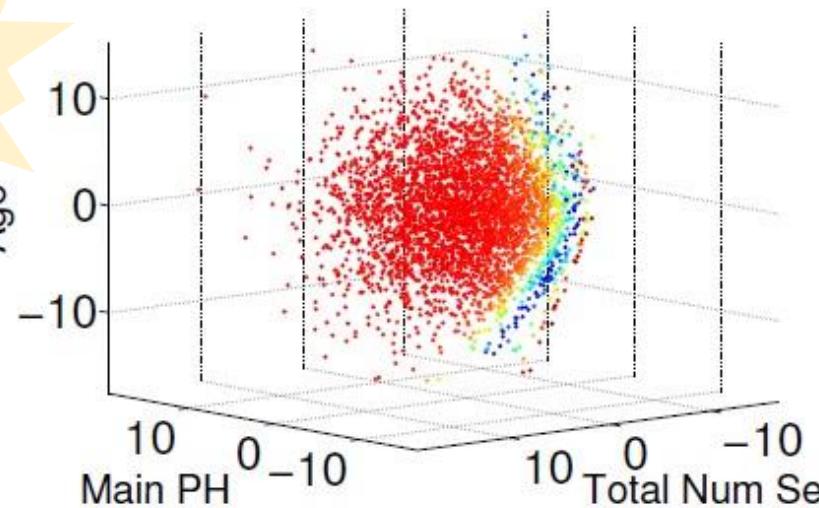
DNE: The difference in the number of events in the simulated and observed data

KL: The Kullback-Leibler divergence between two histograms, one from the observed data, and one from the real data (Histograms of time differences between events).

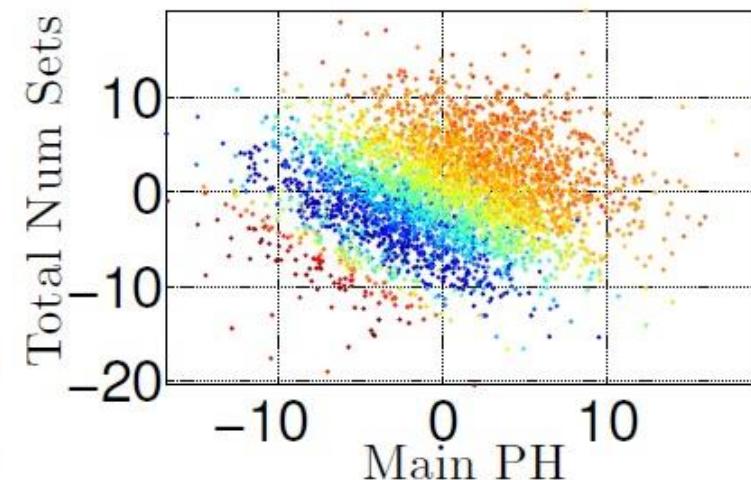
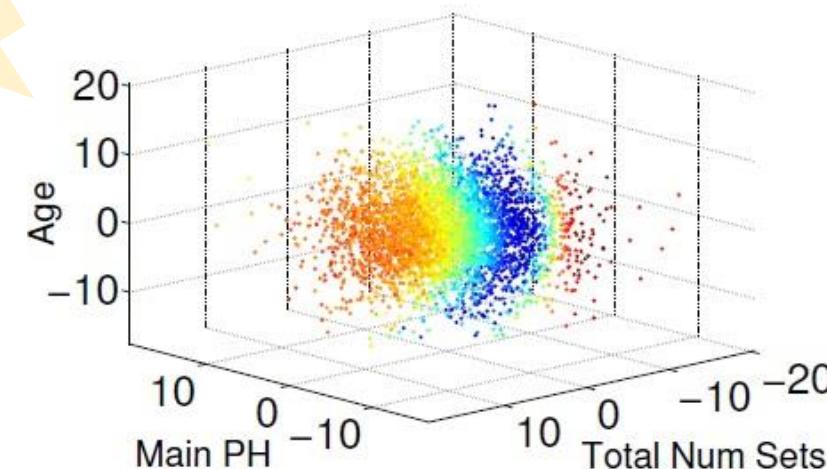
$$\text{KL}(P||Q_v) = \sum_{\text{bin}} \ln \left(\frac{P(\text{bin})}{Q_v(\text{bin})} \right) P(\text{bin})$$

DD & KL for Manhattan Dataset

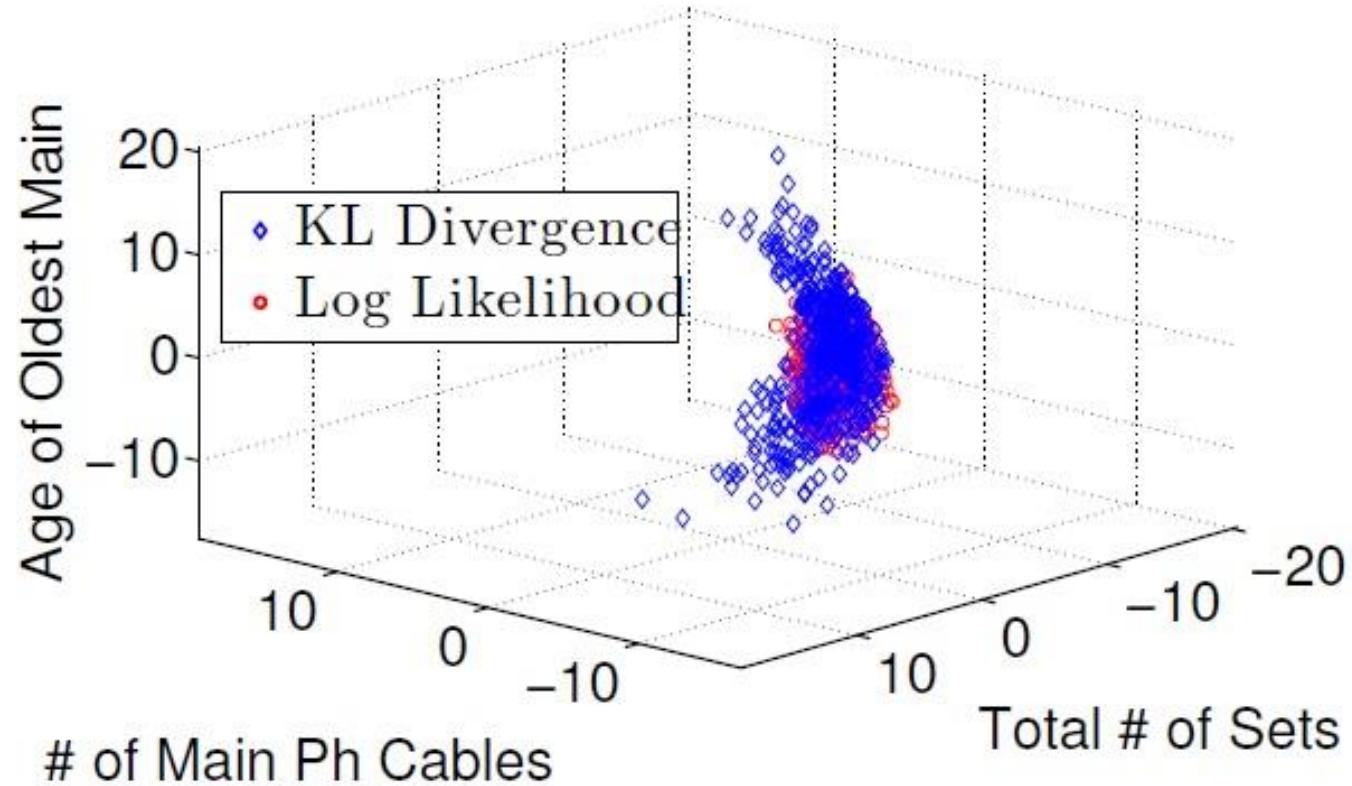
DD



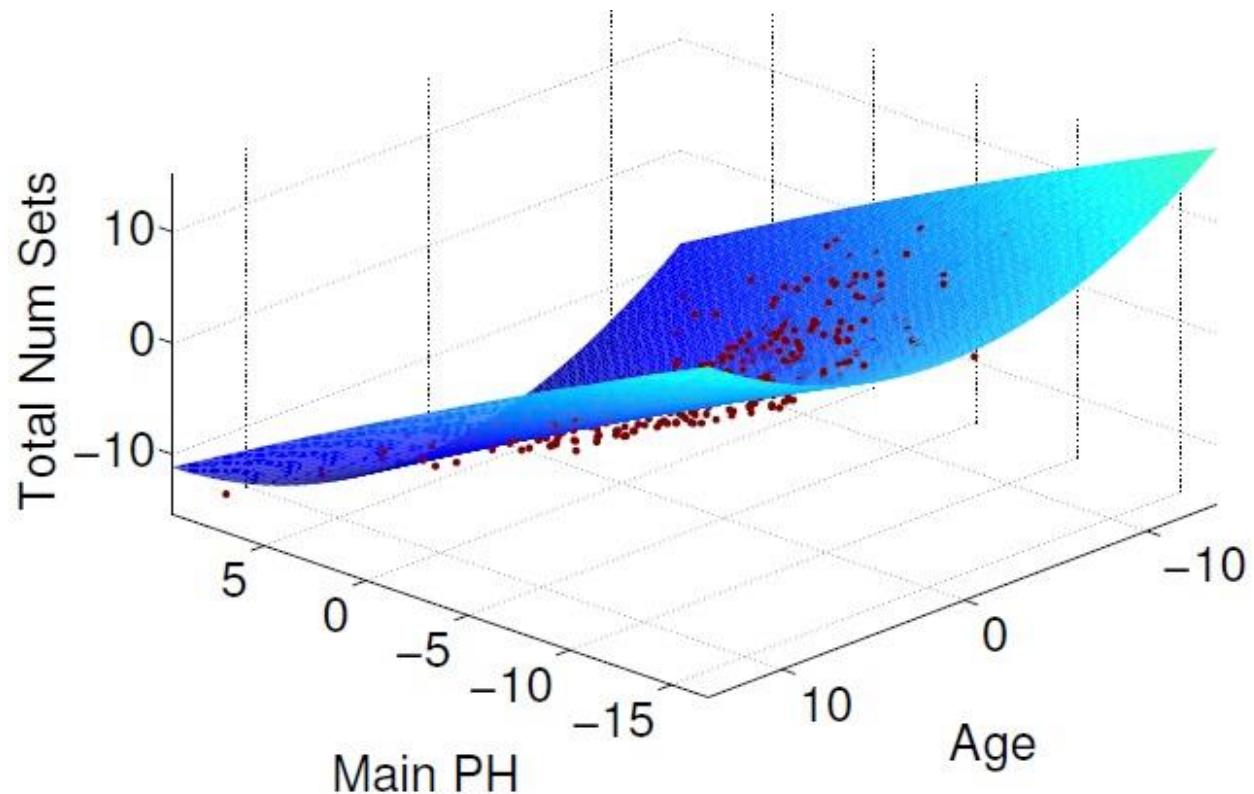
KL



Choosing Parameter Values



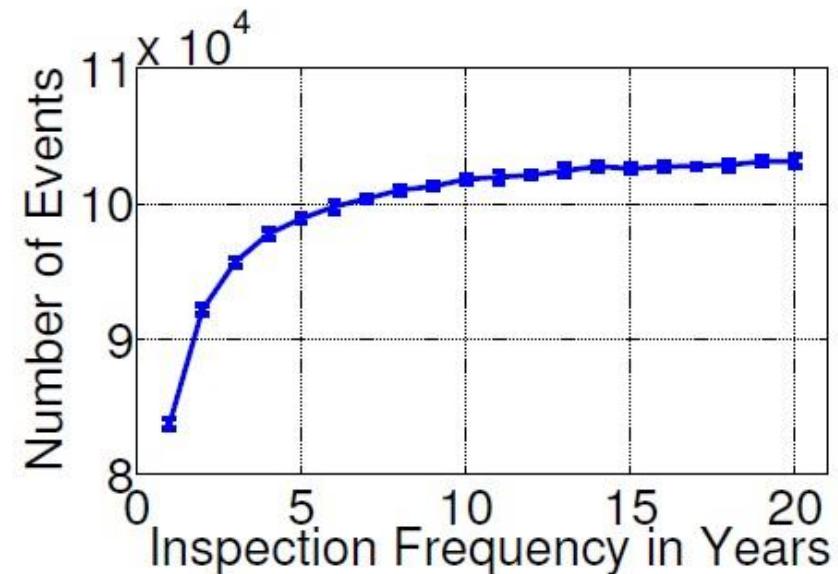
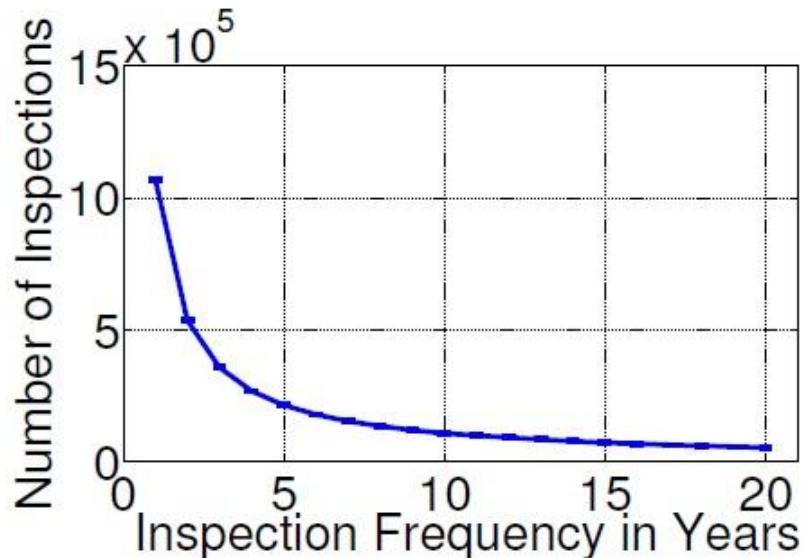
Choosing Parameter Values



Making Policy Decisions with RPP

ConEd's inspection policy:

- Targeted inspections
(each manhole is inspected in every 5 years)
- Adhoc inspections



Ruzgar Enerji Santrallerinde Turbinlerde Cikacak olan Arizalari Tahmin Etme



Preemptive maintenance rather than reactive maintenance



Maintenance Prioritization



A large, red, textured character resembling Elmo from Sesame Street is positioned in the upper left corner of the image. It has a black mouth, two white eyes with black pupils, and a small orange nose.

Thank you!

