Büyük Veri Analitiği İnovasyon ve Karar Almada Yeni Hudutlar

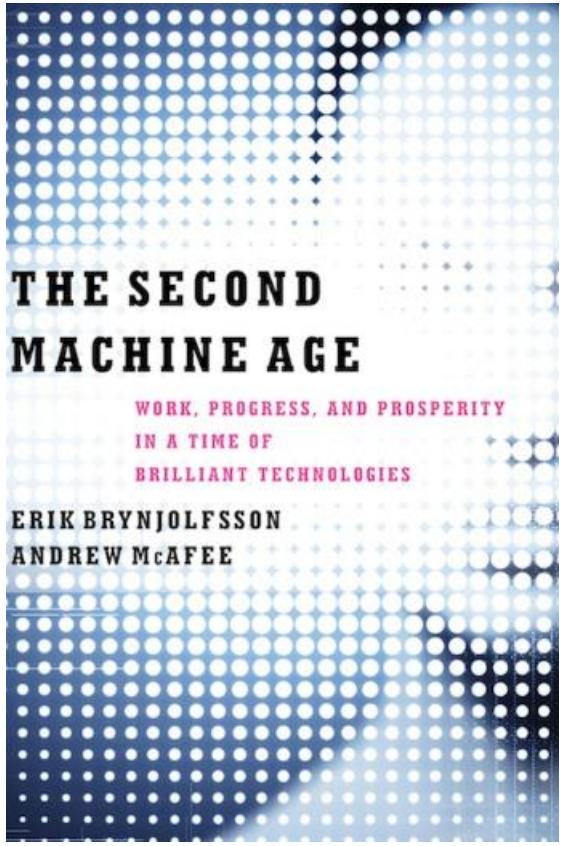
Y. Doç. Dr. Şeyda Ertekin

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}

twitter : seydaertekin

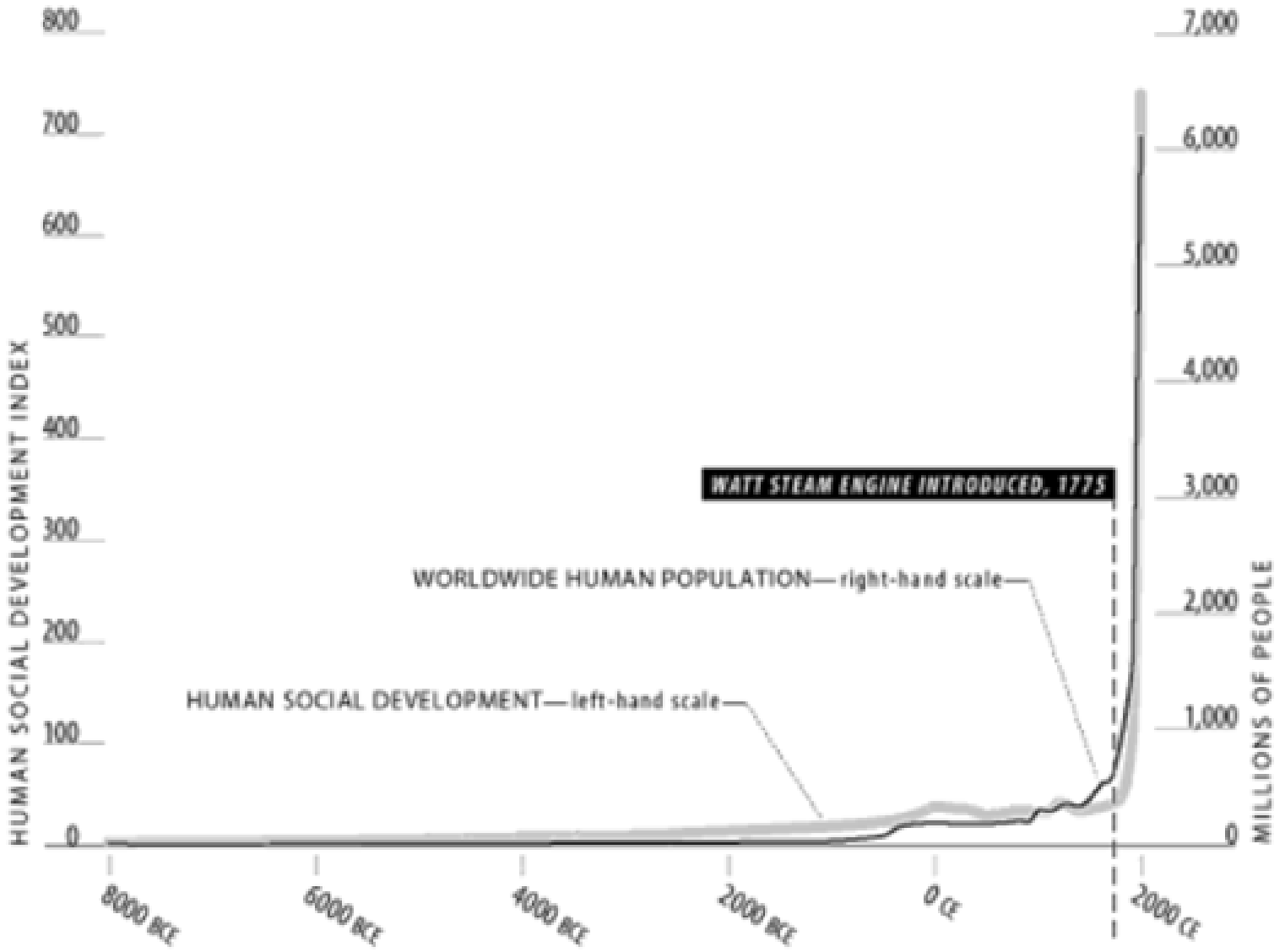
Yapay Öğrenme ve Bilgi İşlemede Yeni Teknikler Yaz Okulu (2016)



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.

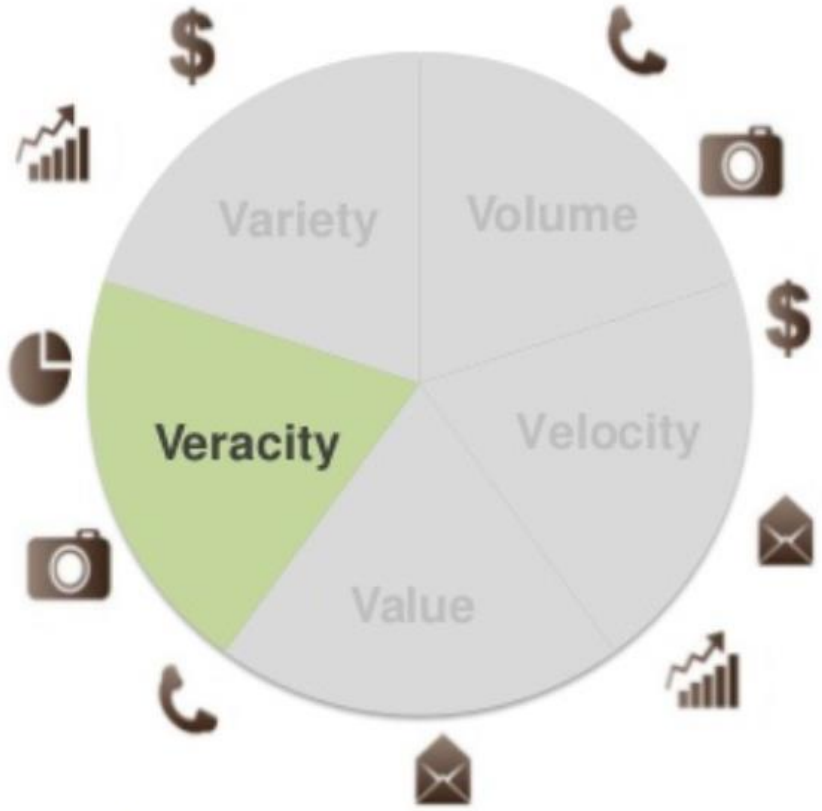
Variety

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

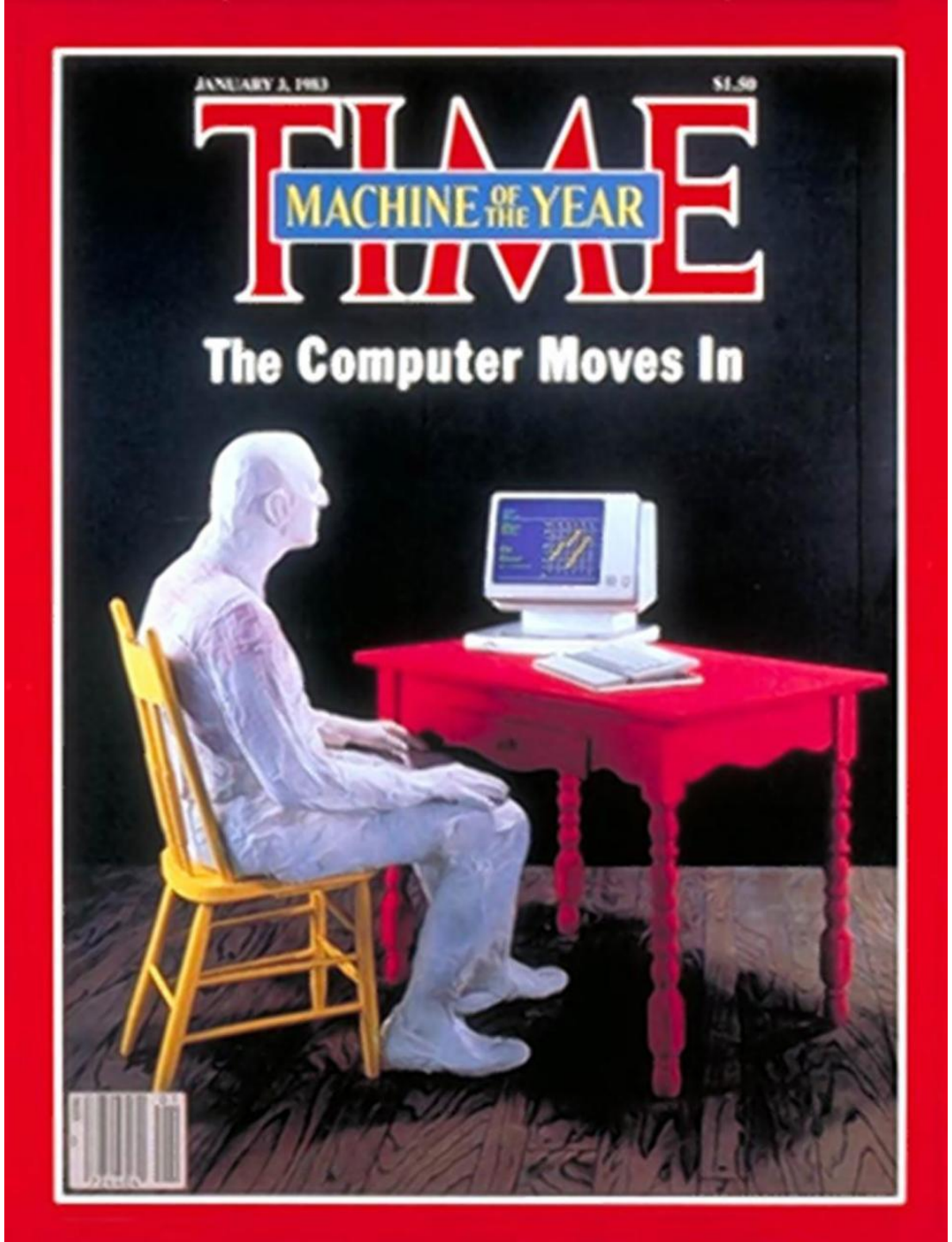
Value – The most important V of all!

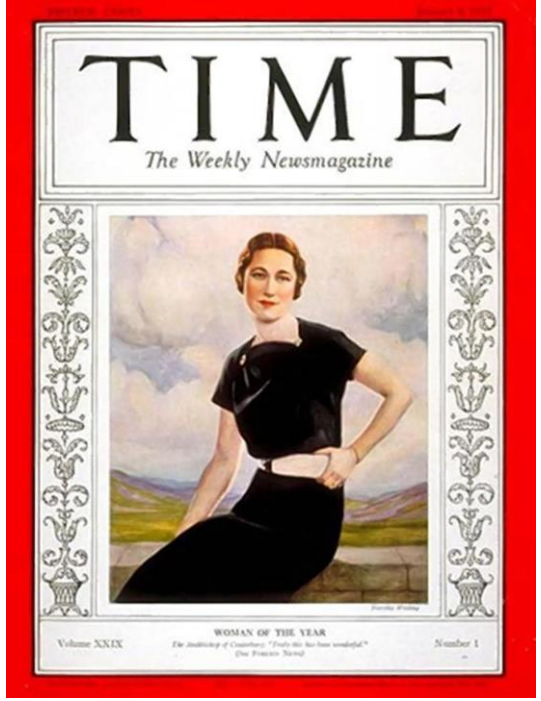
Then there is another V to take into account when looking at Big Data: Value!

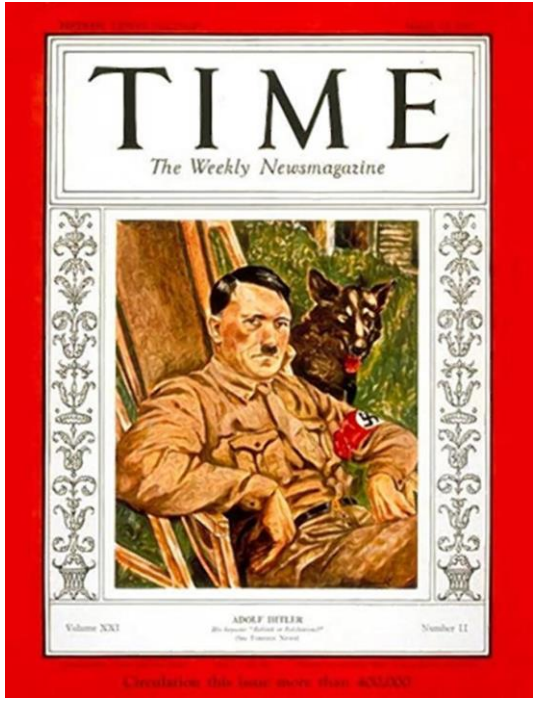
Having access to big data is no good unless we can turn it into value.

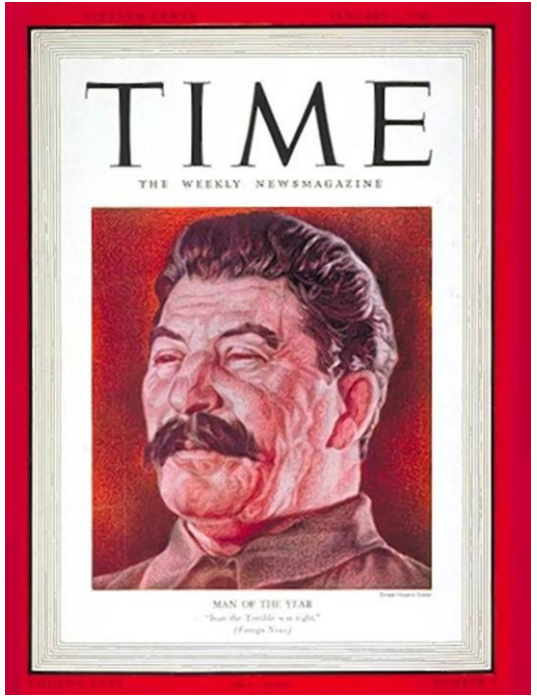
Companies are starting to generate amazing value from their big data.

1982

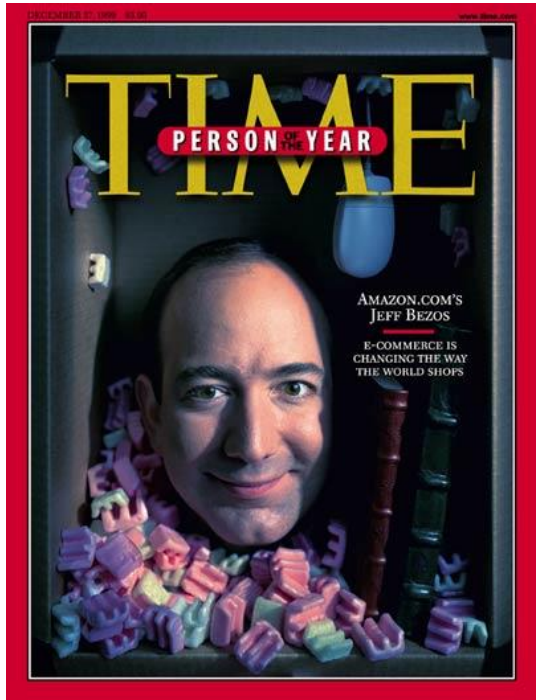

JANUARY 3, 1983
$1.50
TIME
MACHINE OF THE YEAR
The Computer Moves In


TIME
The Weekly Newsmagazine
WOMAN OF THE YEAR
Volume XXIX
Number 1


TIME
The Weekly Newsmagazine
ADOLF HITLER
Volume XXI
Number 11


TIME
THE WEEKLY NEWSMAGAZINE
MAN OF THE YEAR


MAN of the YEAR
TIME
THE WEEKLY NEWSMAGAZINE
MARTIN
LUTHER
KING JR.

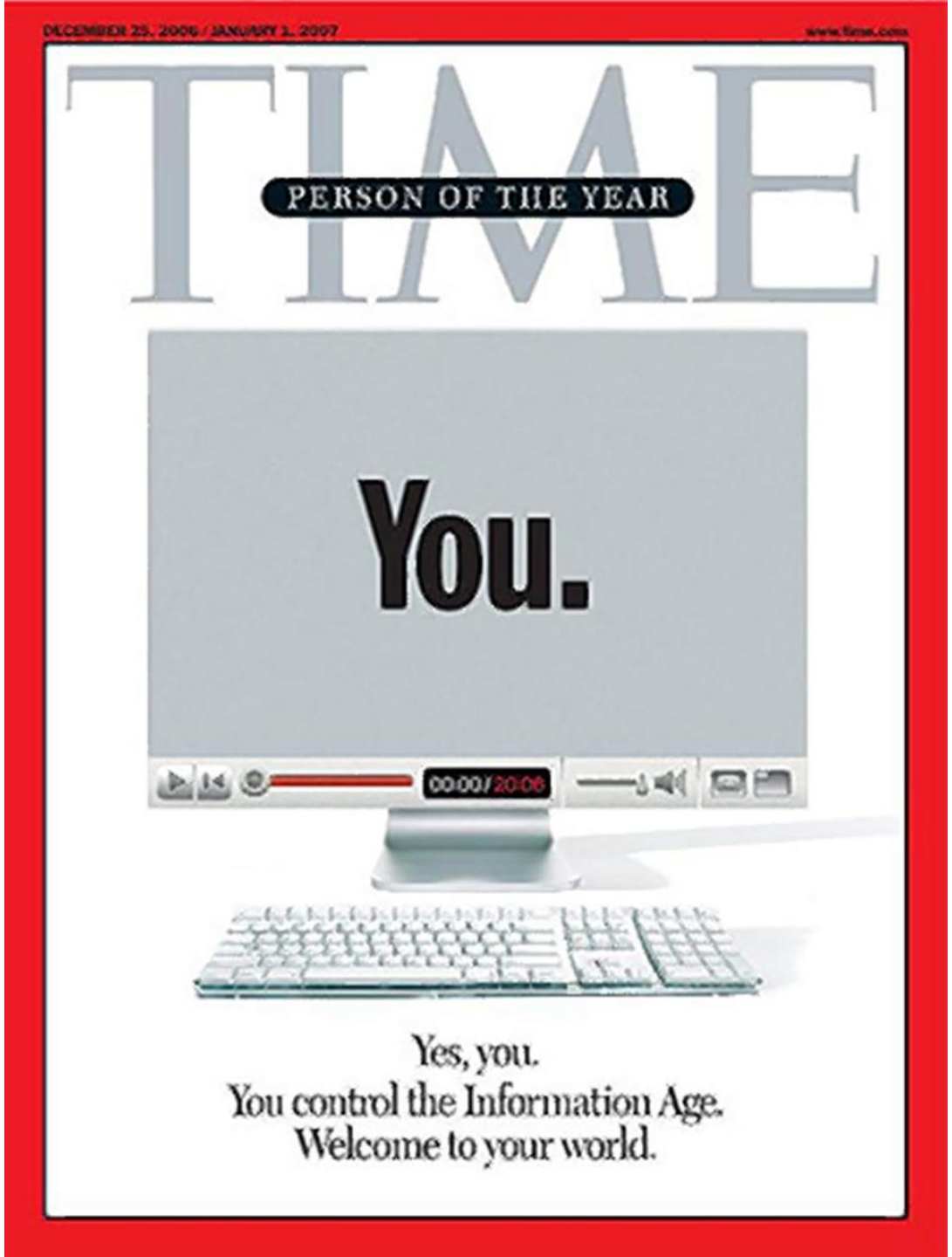

TIME
PERSON OF THE YEAR
AMAZON.COM'S
JEFF BEZOS
E-COMMERCE IS
CHANGING THE WAY
THE WORLD SHOPS


Person of the Year
And the Runners-Up
TIME
Vladimir
Putin
Tsar of
The New
Russia


Person of the Year
TIME
Facebook's
Mark
Zuckerberg
THE CONNECTOR


TIME
PERSON OF THE YEAR

2006


DECEMBER 25, 2006 / JANUARY 1, 2007
www.time.com
TIME
PERSON OF THE YEAR
You.
00:00/2006
Yes, you.
You control the Information Age.
Welcome to your world.

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 Majesty.

By the Company of Parish Clerks of London, &c.

The Diseases and Casualties this year.

Abortive and Stillborne	617	Executed	21	Palsie	50
Aged	1545	Flox and Small Pox	655	Plague	68596
Ague and Feaver	5257	Found dead in streets, fields, &c.	20	Planets	6
Appoplex and Suddenly	116	French Pox	86	Plurisie	25
Bedrid	10	Frighted	13	Poysoned	1
Blasted	5	Gout and Sciatica	27	Quinsie	35
Bleeding	16	Grief	46	Rickets	557
Bloody Flux, Scowring & Flux	185	Griping in the Guts	1288	Rising of the Lights	397
Burst and Scalded	8	Hanged & made away themselves	7	Rupture	34
Calenture	3	Headmouldshot & Mouldfallen	16	Scurvy	105
Cancer, Gangrene and Fistula	56	Jaundice	110	Shingles and Swine pox	3
Canker, and Thrush	111	Impostume	227	Sores, Ulcers, broken and bruised Limbs	82
Childbed	625	Kild by severall accidents	46	Spleen	14
Chrisomes and Infants	1258	Kings Evill	86	Spotted Feaver and Purples	1929
Cold and Cough	68	Leprosie	1	Stopping of the Stomack	138
Collick and Winde	134	Lethargy	14	Stone and Strangury	98
Consumption and Tissick	4808	Livergrown	31	Surfet	2352
Convulsion and Mother	2036	Meagrom and Headach	13	Teeth and Worms	2614
Distracted	5	Measles	7	Vomiting	5
Dropsie and Timpany	1478	Murthered and Shot	9	Wenn	1
Drowned	50	Overlaid & Starved	45		

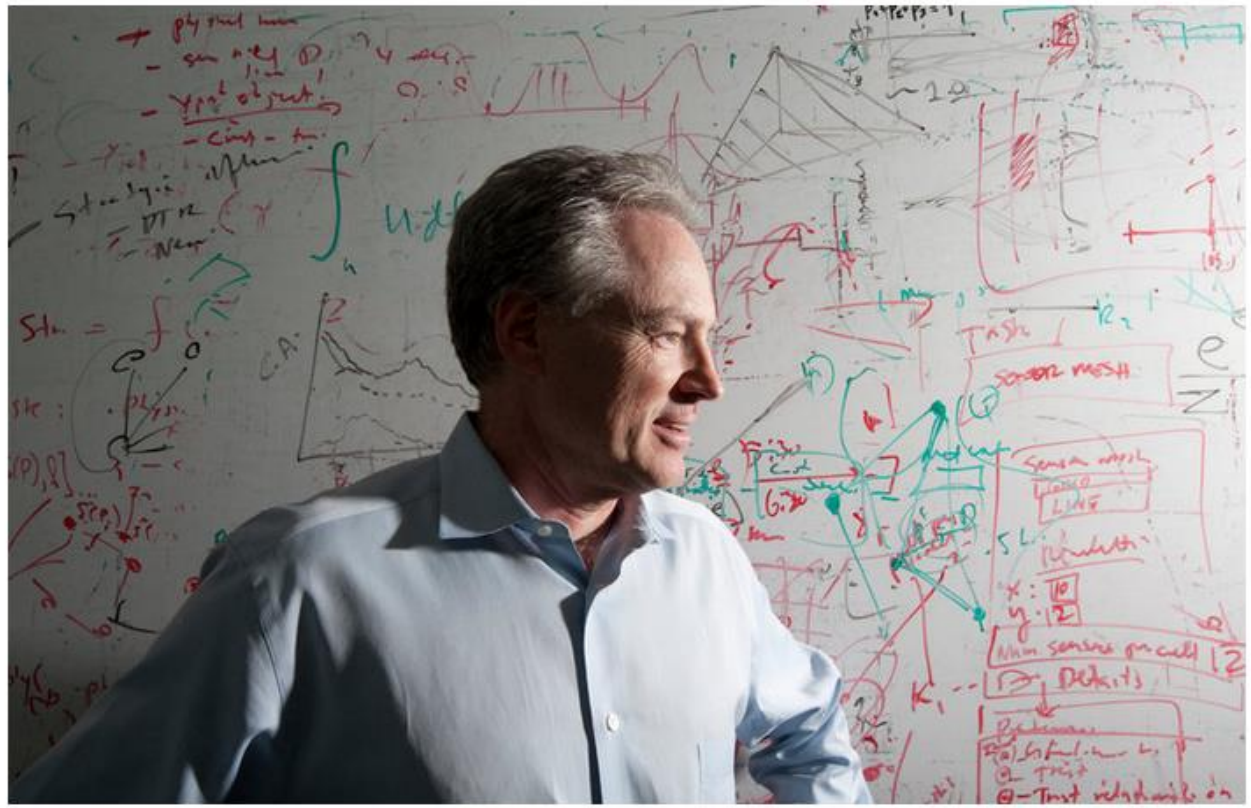
Christned	Males 5114	Buried	Males 48568	Of the Plague	68596
	Females 4853		Females 48737		
	In all 9967		In all 97306		

Increased in the Burials in the 130 Parishes and at the Pest-house this year 77009

Increased of the Plague in the 120 Parishes and at the Pest-house this year 68590

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.

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

PARTNER CONTENT CYNTHIA RUDIN, MIT SLOAN

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.

MOST PO


HOV
Kid
Mar
Gar
06.


PHO
Bey
Pap
Wo
04.


TV
Sta
Imp
Ran
15

MORE STORI

IBM detects skin cancer more quickly with visual machine learning

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

MORE LIKE THIS


IBM expands Watson's presence to multiple industries


IBM looks to replace the spreadsheet with Watson Analytics

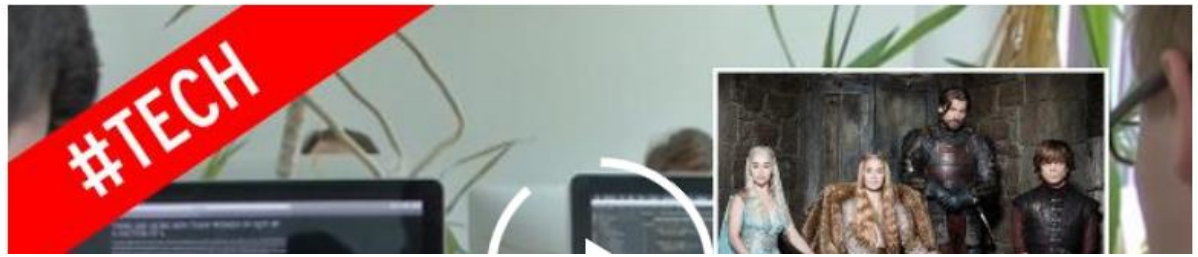
10 IBM Watson-powered apps that are changing our world

on IDG Answers
How does Simply Secure improve cloud security?

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



★ 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

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

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.

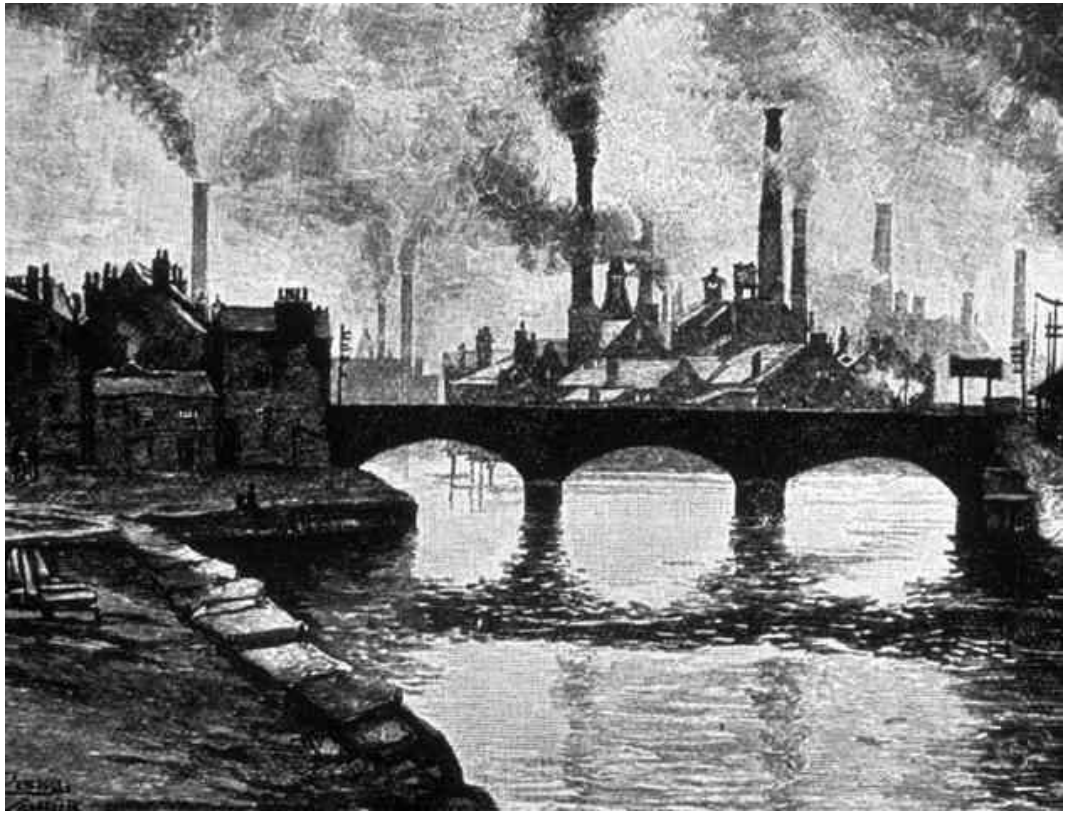
The last decade has seen huge advances in the amount of data we routinely generate and collect in pretty much everything we do, as well as our ability to use technology to analyze and understand it. The intersection of these trends is what we call "Big Data" and it is helping businesses in every industry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

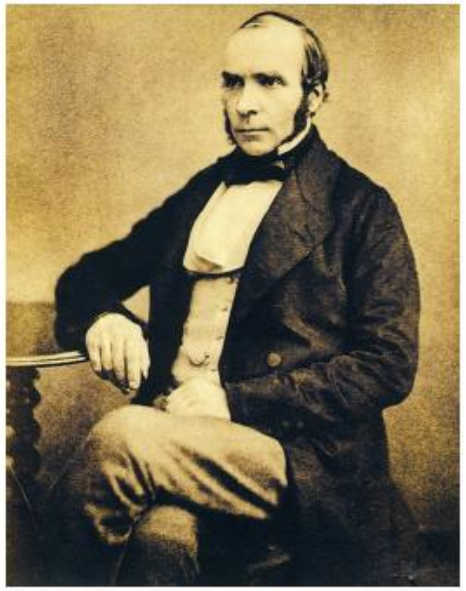
Story from 19th century

London, 1854 :

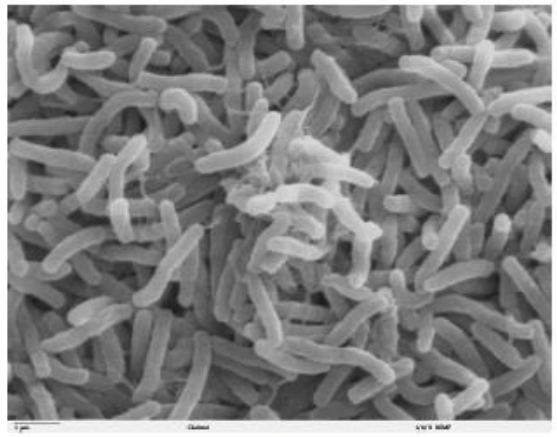
- Industrial revolution.
- Two millions people.
- Insufficient sewage.
- Garbage removal problems.
- Little clean water.



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

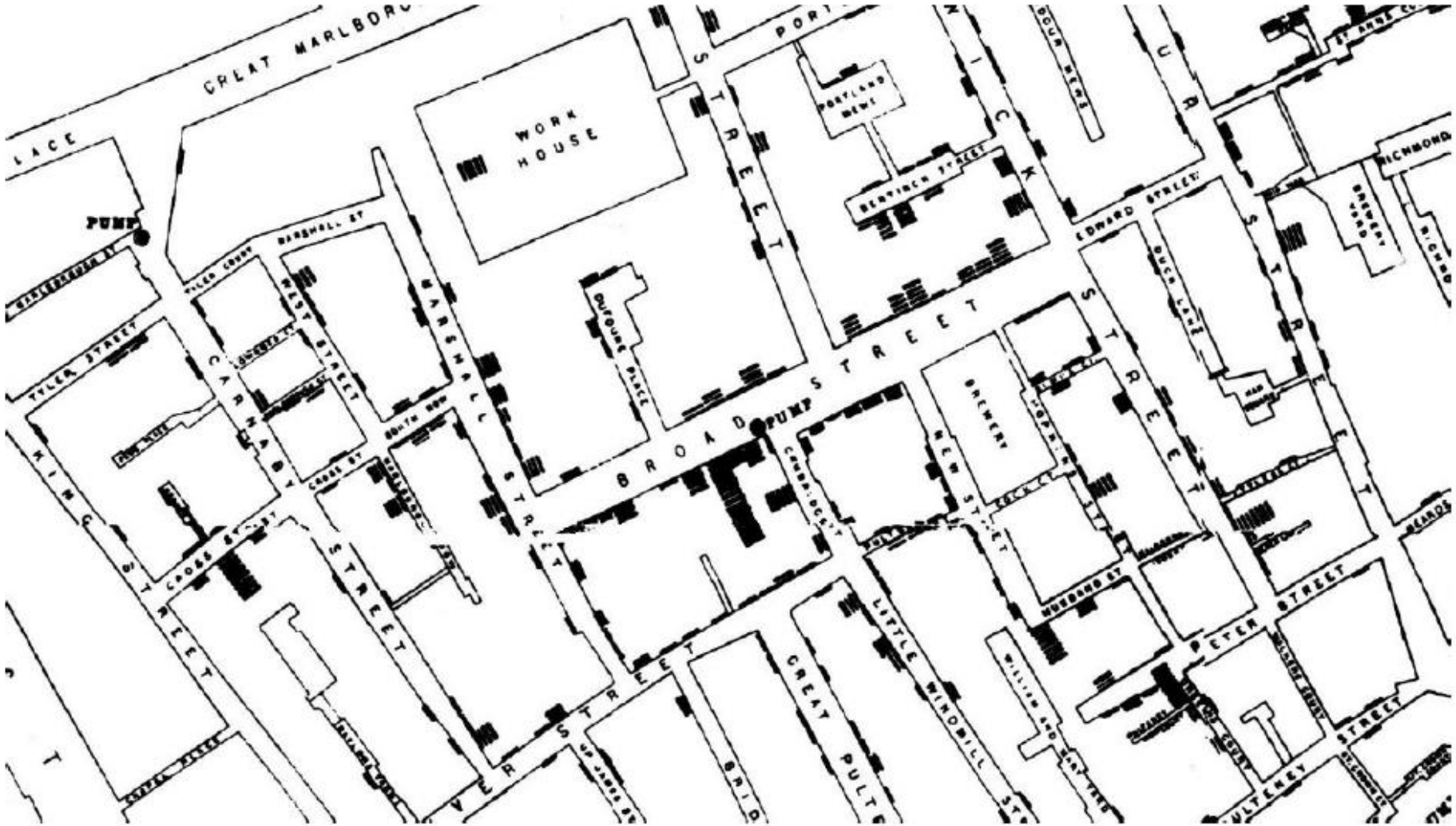
If cholera was propagated by effluvia from sick people,

– why are physicians less affected than their patients?

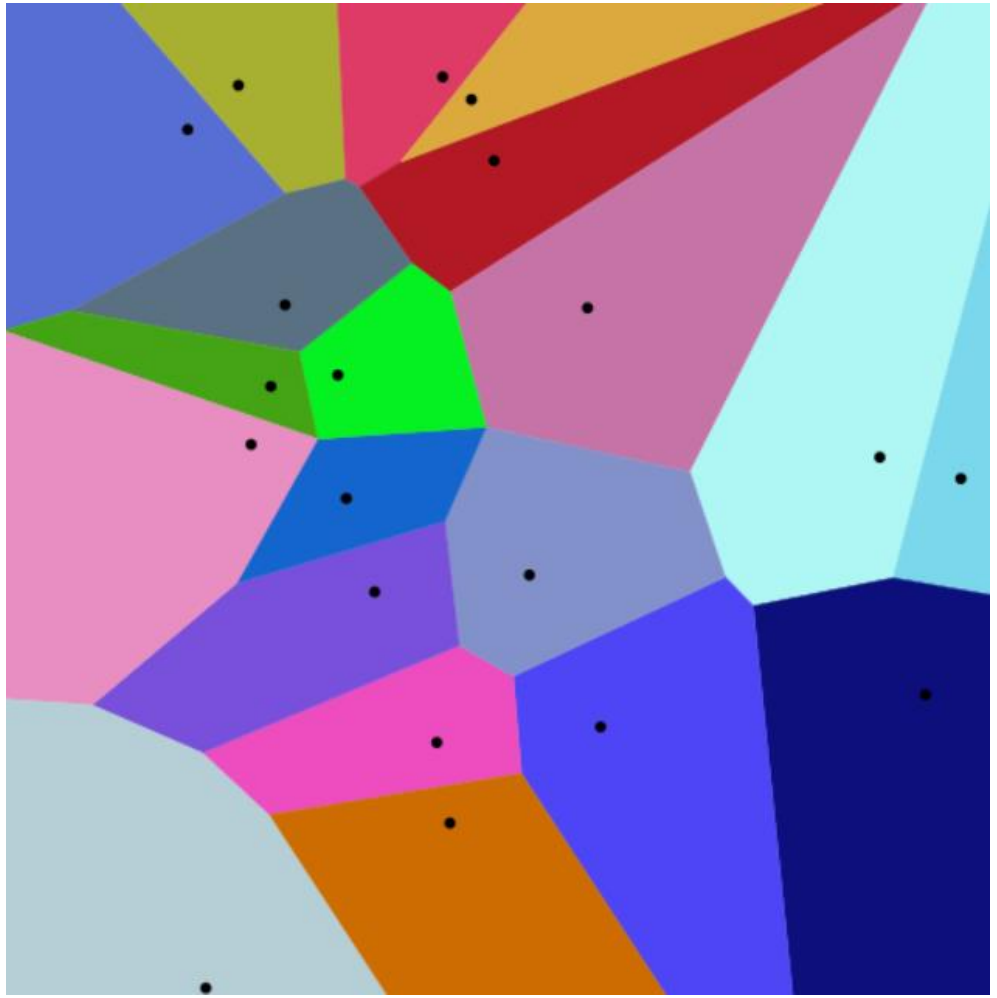
– why are men servants less affected than their masters?

– why are master brewers virtually immune?

The map


WORK HOUSE
BROAD STREET
PUMP
BREWERY
PETER STREET
GREAT PULTENEY STREET
LITTLE WINDMILL STREET
CROSS STREET
BERWICK STREET
EDWARD STREET
RICHMOND
TYLER STREET
CARNABY STREET
MARSHALL STREET
POLAND STREET
NEW STREET
HUDSON ST
DUCK LANE

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 Magdalen	13,934	87	83	-	-	4	-
Leather Market	15,295	81	81	-	-	-	-
Rotherhithe	17,805	103	68	-	-	35	-

The Broad Street water pump

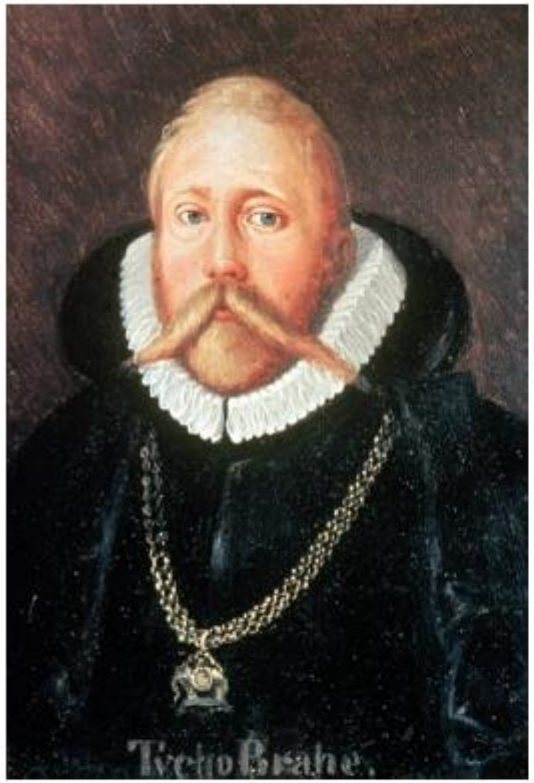
On proceeding to the spot, I found that nearly all the deaths had taken place within a short distance of the [Broad Street] pump. There were only ten deaths in houses situated decidedly nearer to another street-pump. In five of these cases the families of the deceased persons informed me that they always 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

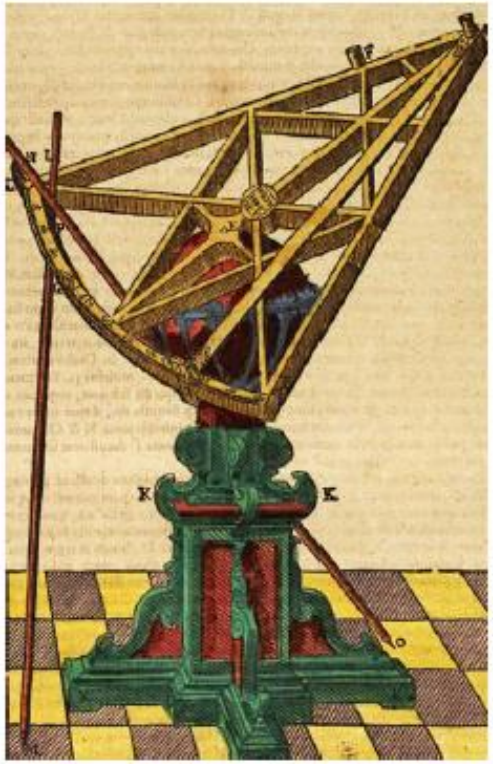
The characters:



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.

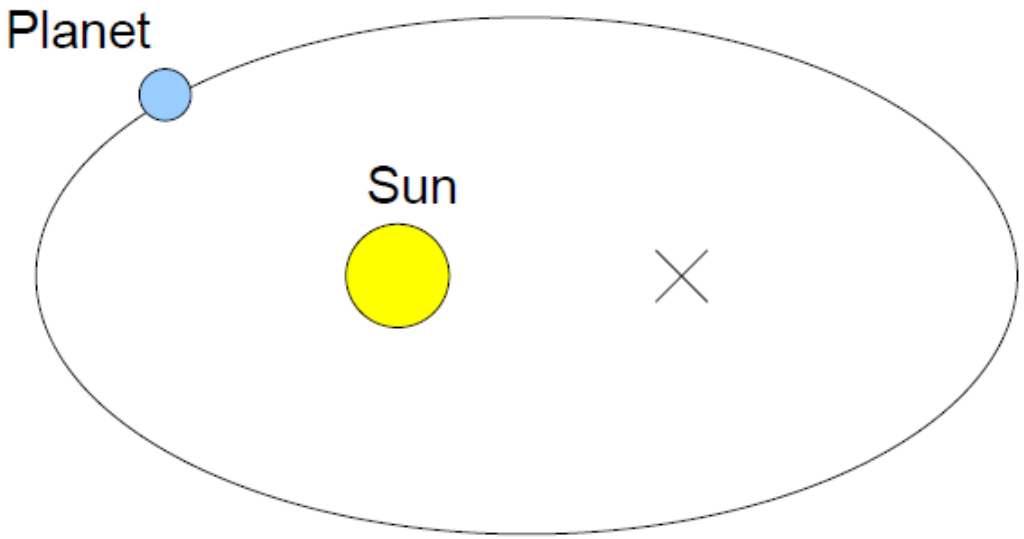
There he hires a bright young assistant named Johannes Kepler.

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

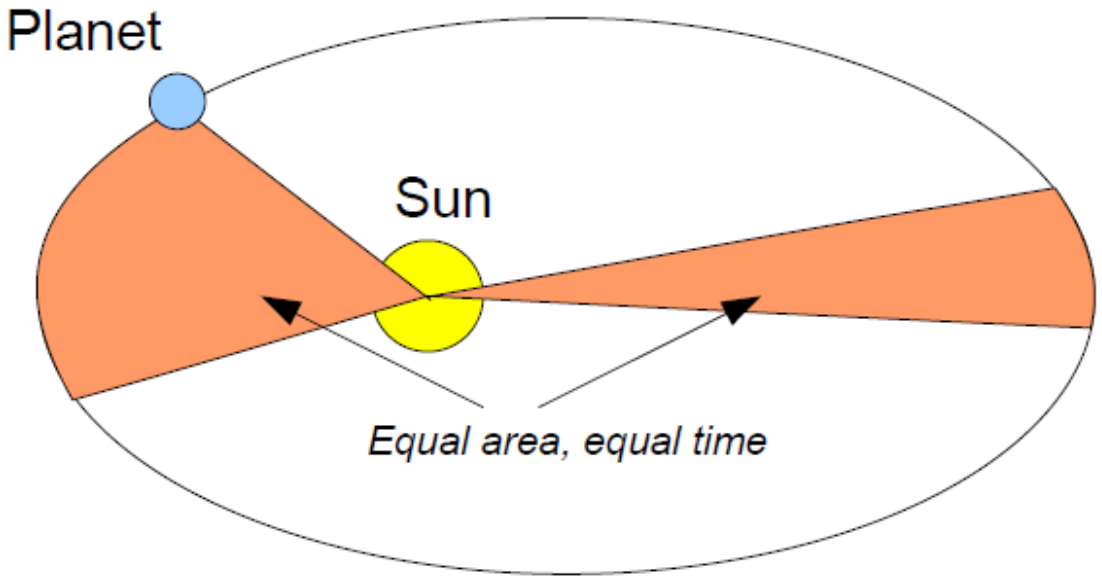
The orbits of the planets are ellipses with the Sun at a focal point.



J. Kepler, *Astronomia nova*, 1609

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.



J. Kepler, *Astronomia nova*, 1609

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.

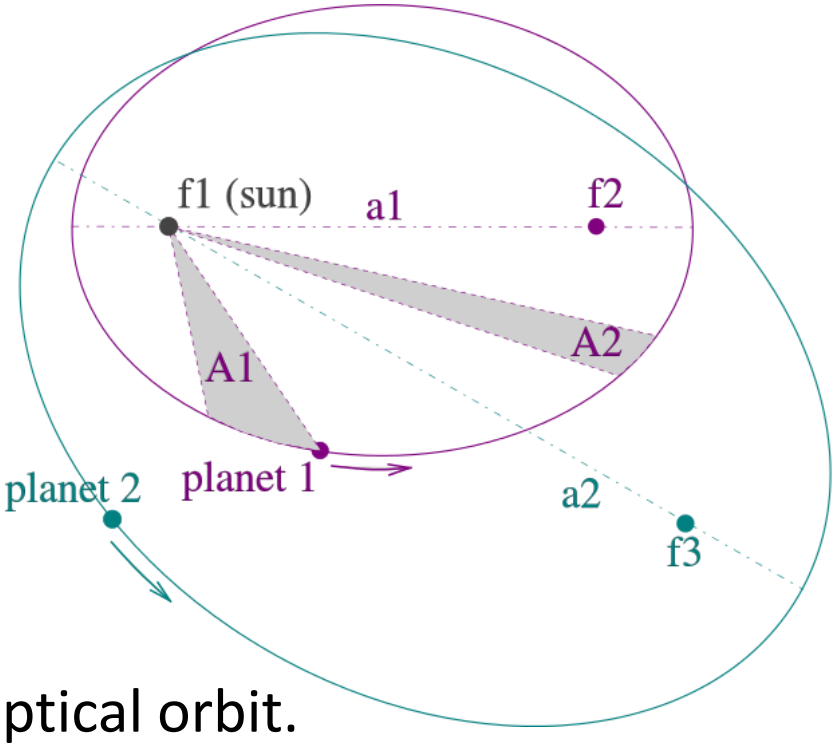
$$\frac{P_{\mathrm{a}}^2}{P_{\mathrm{b}}^2} = \frac{R_{\mathrm{a}}^3}{R_{\mathrm{b}}^3}$$

J. Kepler, *Harmonices Mundi*, 1619

Kepler's laws improve the model of Copernicus

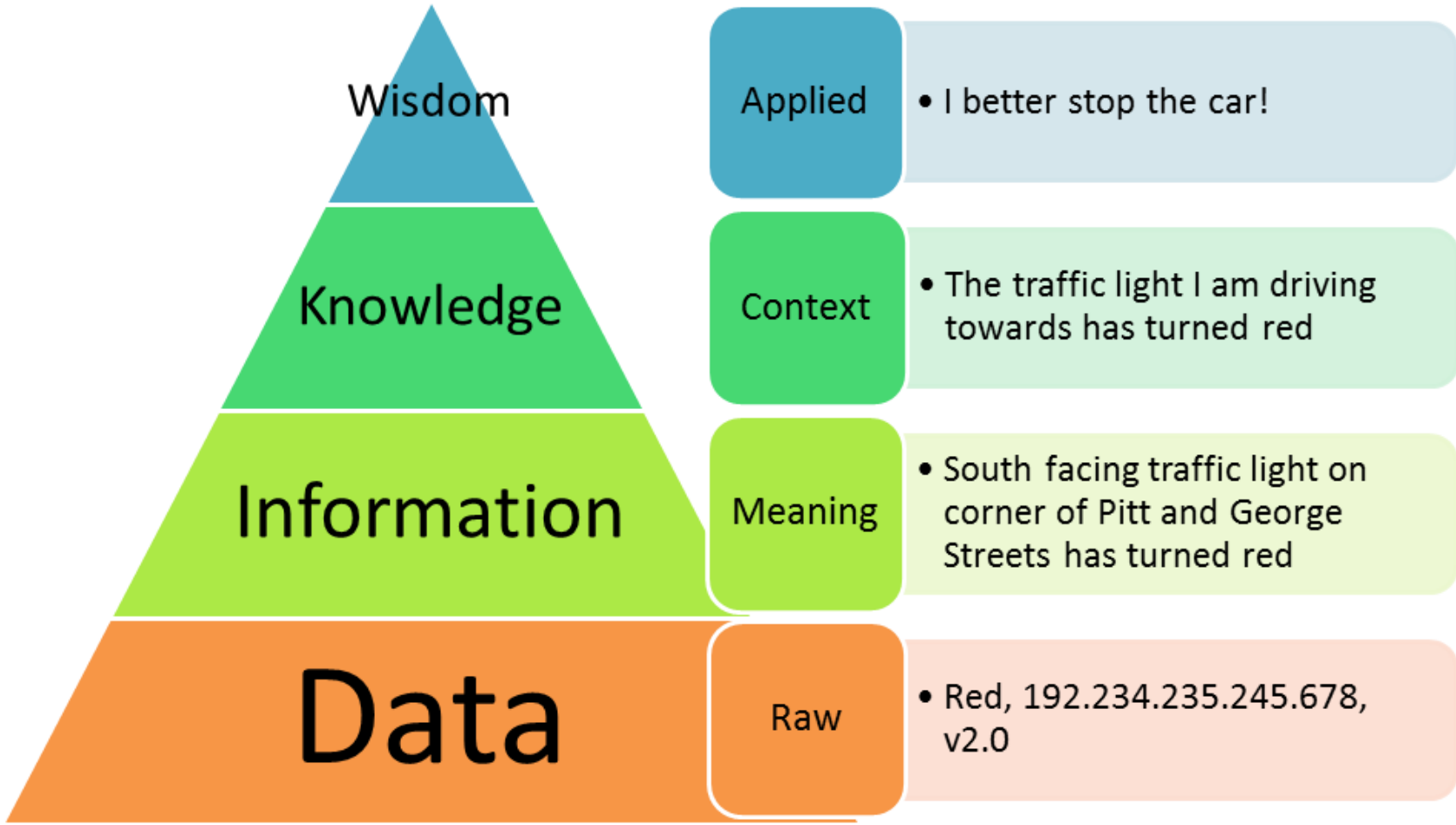
Copernicus's model say:

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

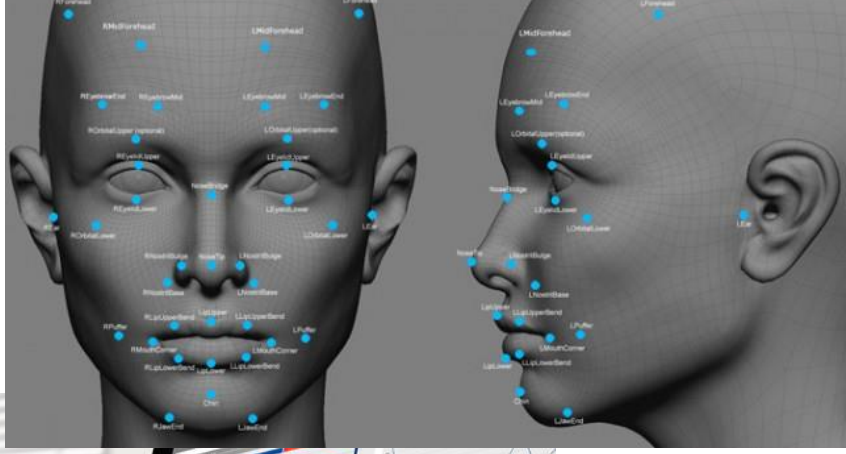


Kepler's model say:

- 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 Ogrenme (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.

mloss.org
machine learning open source software

About | Software | Forum | Blog | Workshop

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 [ ] 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

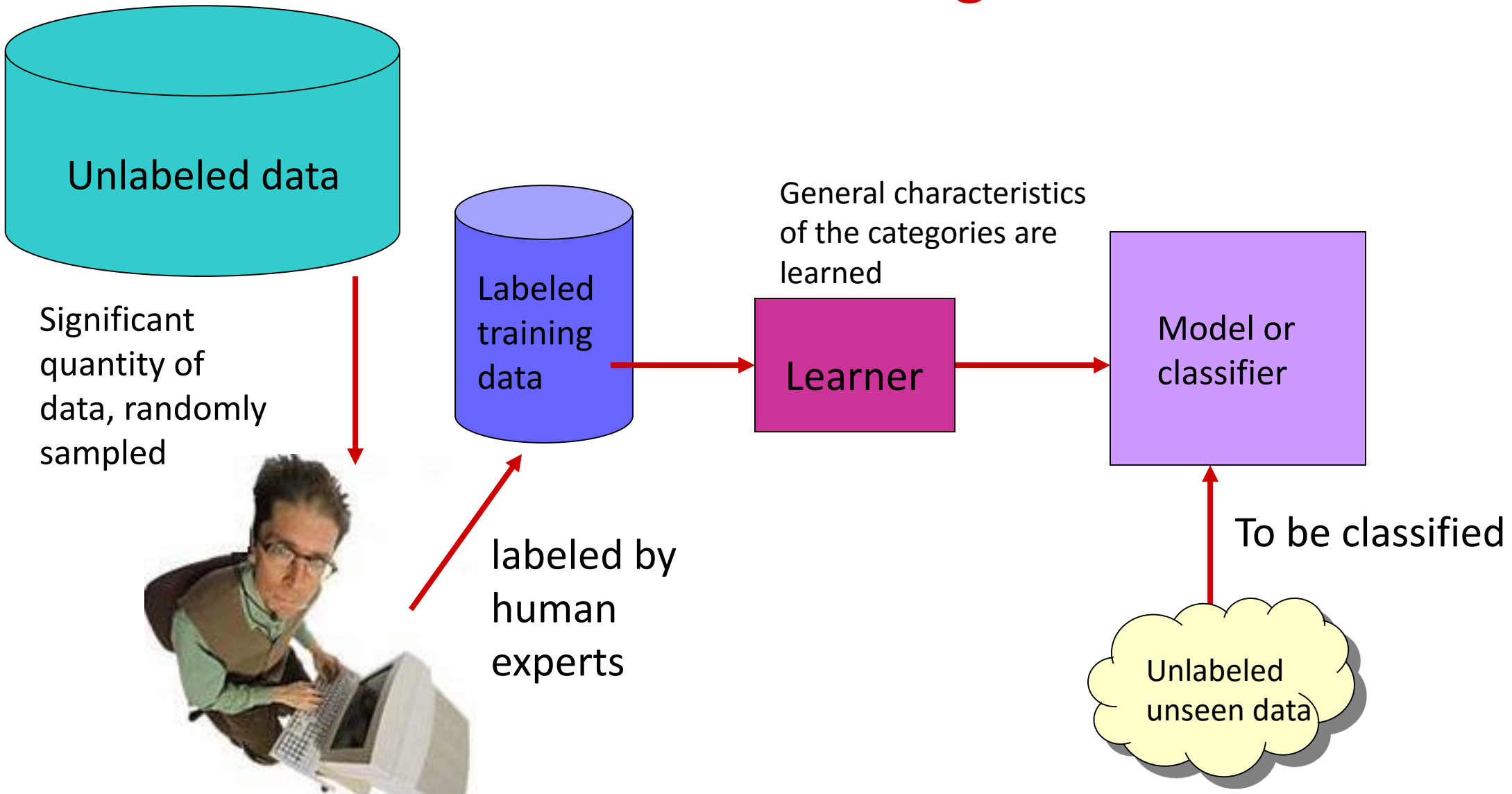
- **Authors:** Leon Bottou, Antoine Bordes, Seyda Ertekin
- **License:** Gpl Version 2 Or Later
- **Programming Language:** C
- **Operating System:** Agnostic
- **Data Formats:** None
- **Tags:** Large Scale, Support Vector Machine, Active Learning, Kernels, Online Learning

Fast Kernel Classifiers with Online and Active Learning.

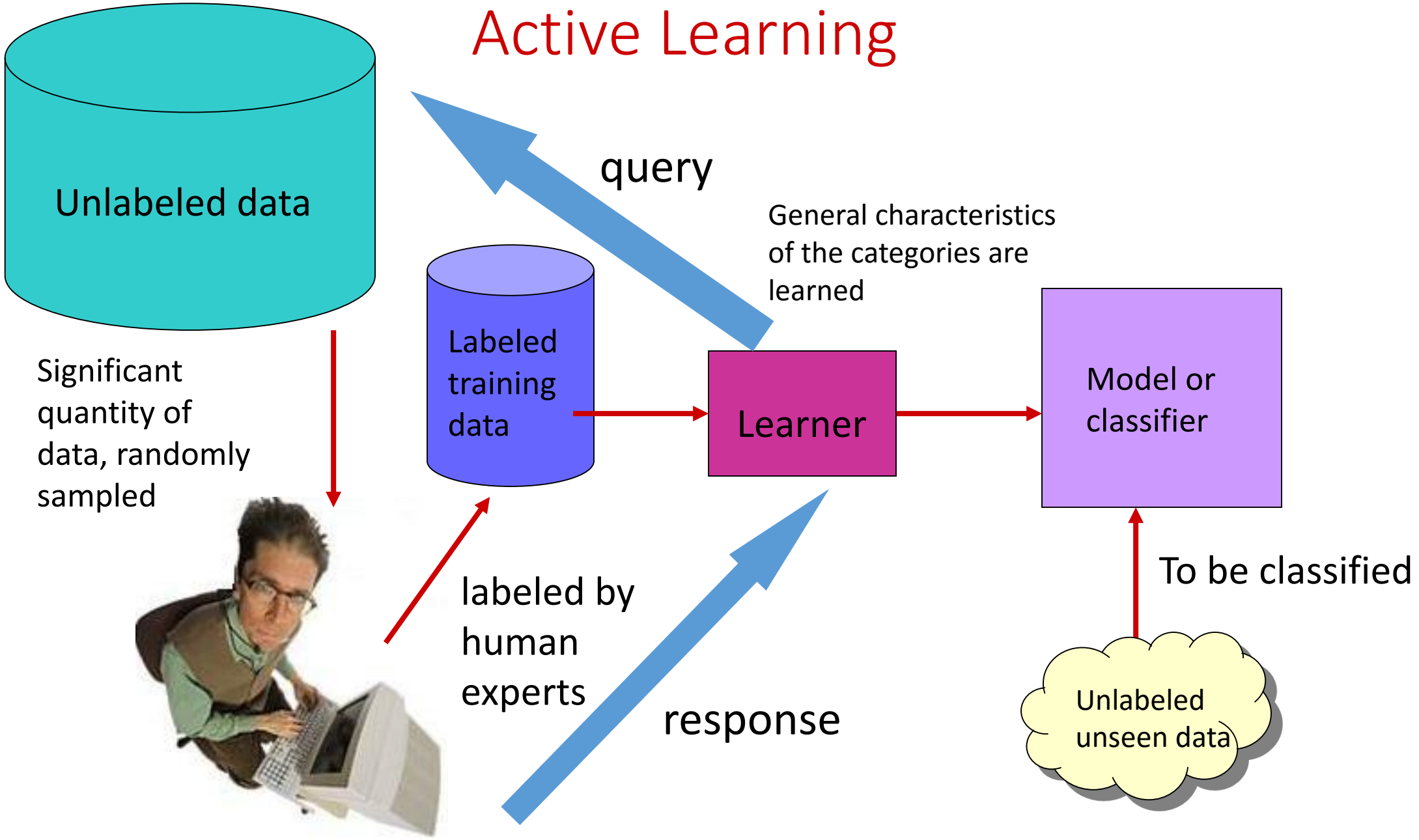
Antoine Bordes, Seyda Ertekin, Jason Weston, Leon Bottou.

Journal of Machine Learning Research (JMLR)

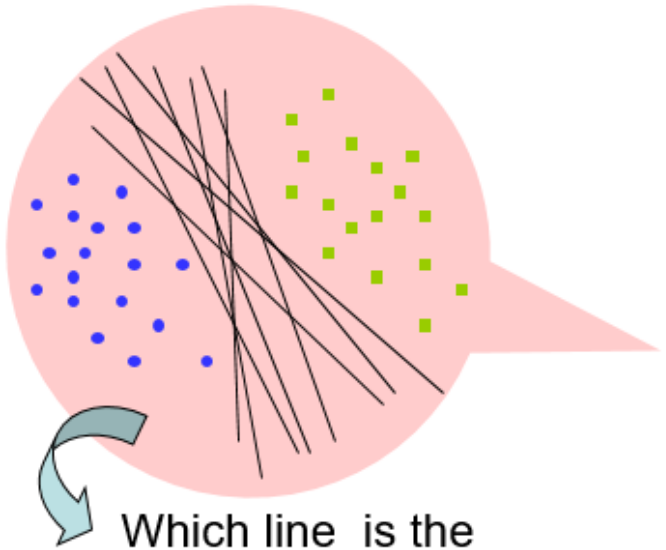
Passive Learning

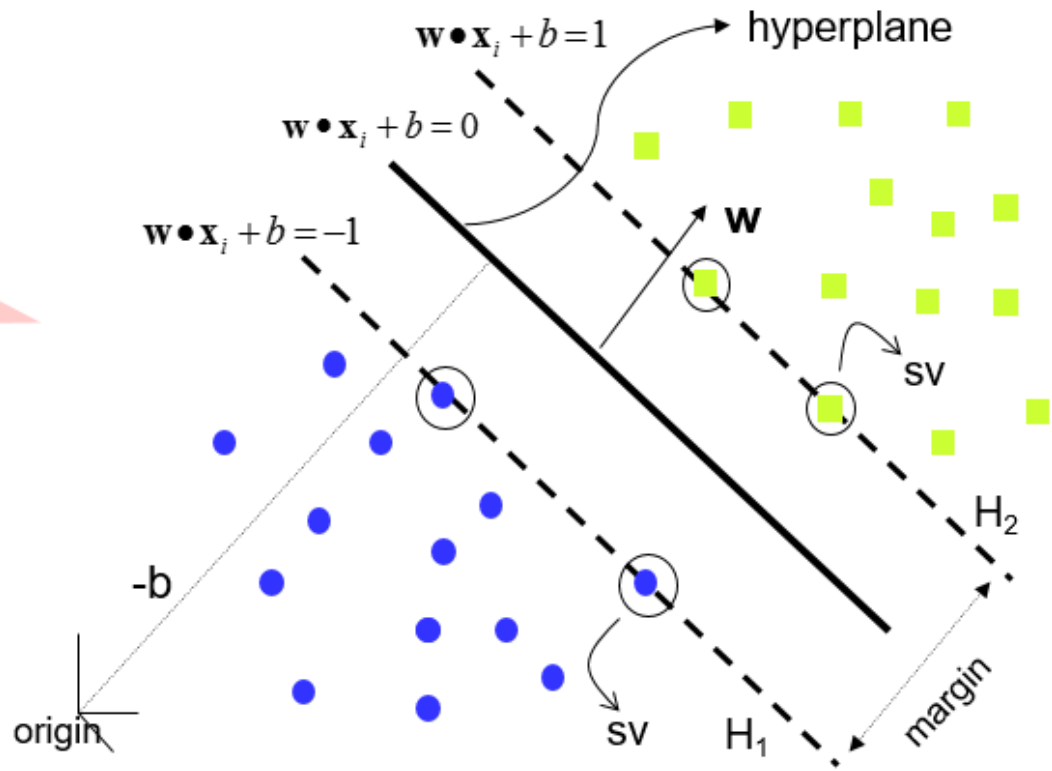


Active Learning



Support Vector Machines (SVMs) I





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

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

Support Vector Machines (SVMs) II

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

Primal SVM formulation

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

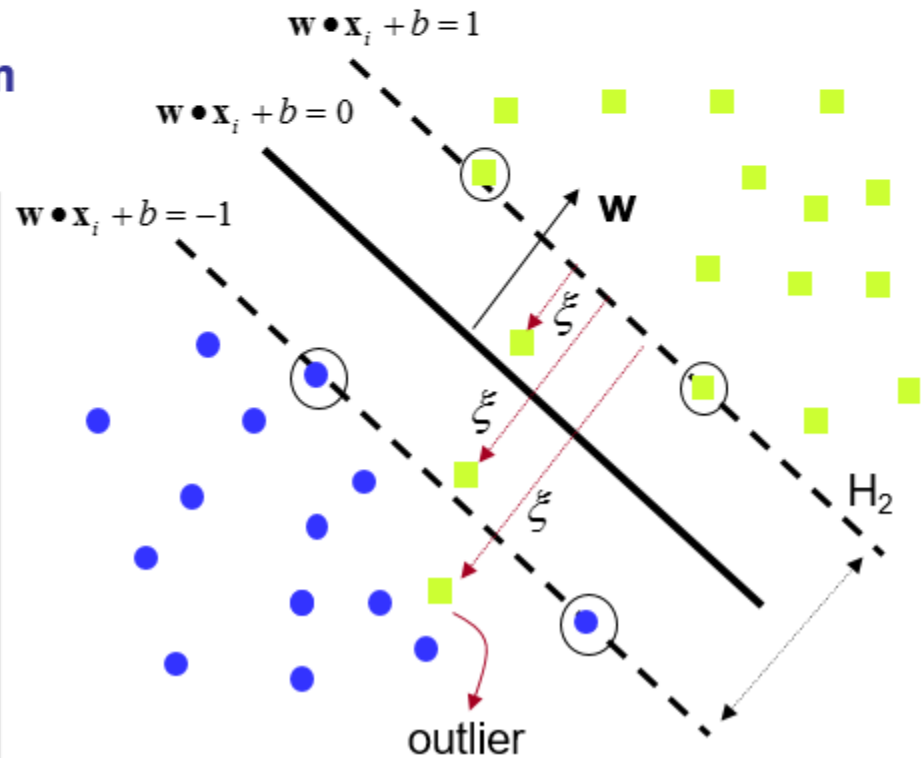
$$\forall i \; \xi_i \geq 0$$

Dual convex optimization problem

needs **Q**uadratic **P**rogramming (QP)

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

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



SVM Quadratic Programming

QP problem is solved when
(KKT conditions)

$$\alpha_i = 0 \Leftrightarrow y_i f(x_i) \geq 1$$
$$0 < \alpha_i < C \Leftrightarrow y_i f(x_i) = 1$$
$$\alpha_i = C \Leftrightarrow y_i f(x_i) < 1$$

After solving QP, we get

$$f_\theta(x) = \sum_i \alpha_i \Phi(x_i) \Phi(x) + b$$

$\alpha_i \neq 0$ for SV's

$\alpha_i = 0$ for non SV's

# of SVs are important!

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

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

SVMs may be quite costly with large datasets

$$\frac{\#\,\text{SV}}{\#\,\text{examples}} \rightarrow 2\text{B}_\Phi$$

Which example is the most informative?

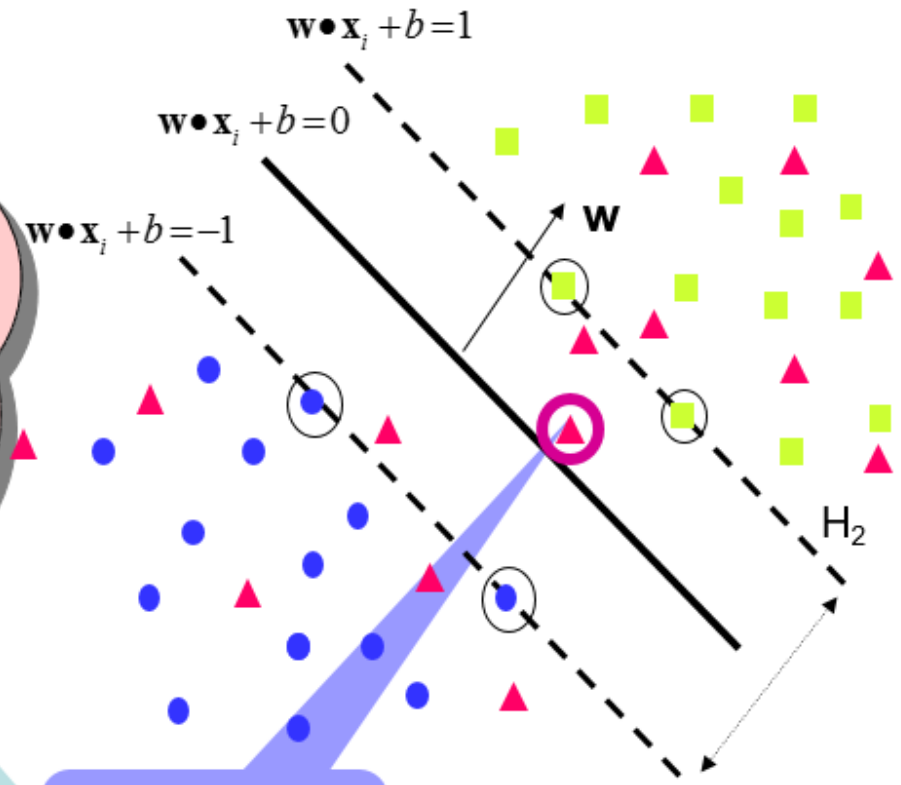
The example, which the

- Learner is most confident about?
- Learner misclassifies most?

(Can be done if the labels are known)

OR

- Learner is less confident about?



Do we really have to search the entire data?

Not really!

Small Pool Active Learning

Pick a random subset X_L, L<< # 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%

$$\mathbf{1 - (1 - p\%)^L}$$

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

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

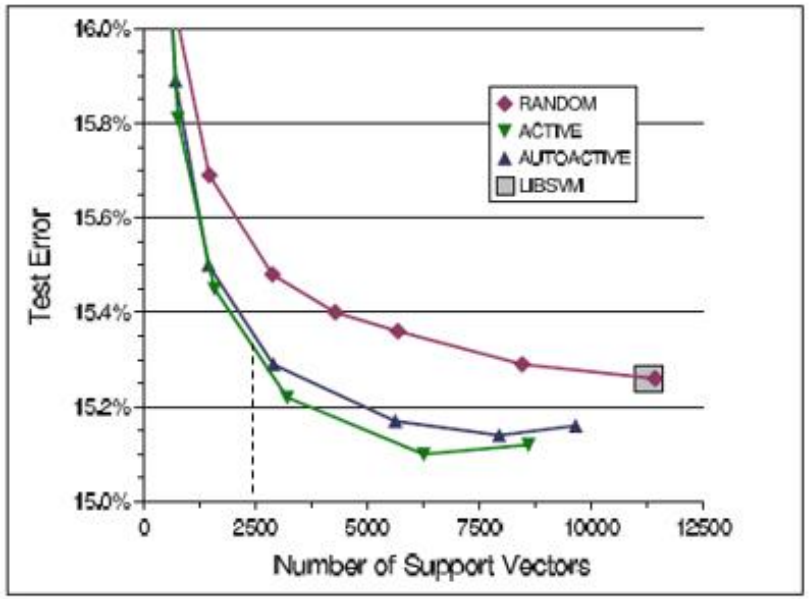
L in terms of p and q

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

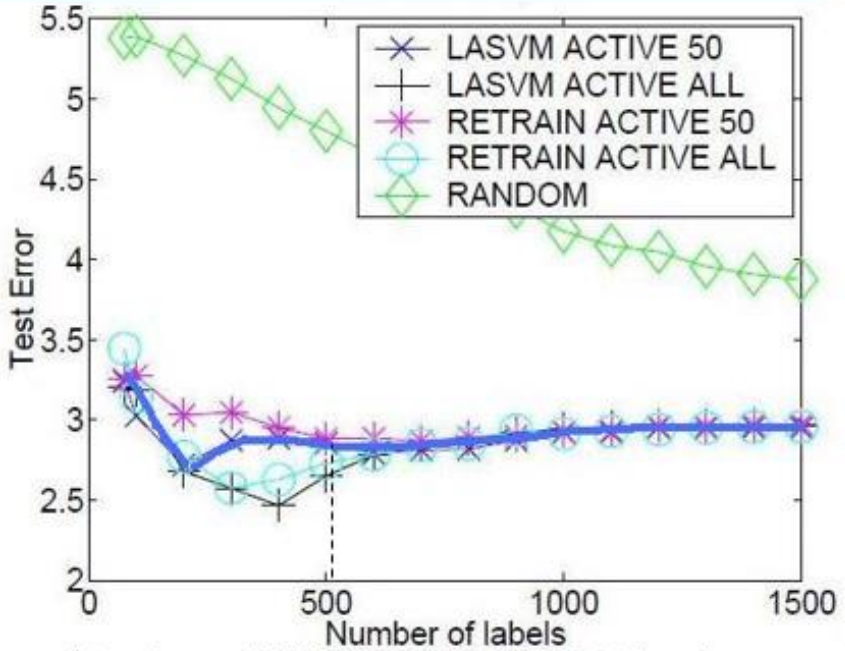
For example, the active learner will pick one instance, with 95% probability, that is among the top 5% closest instances to the hyperplane, by randomly sampling only **log(.05)/log(.95)=59** *instances*

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)

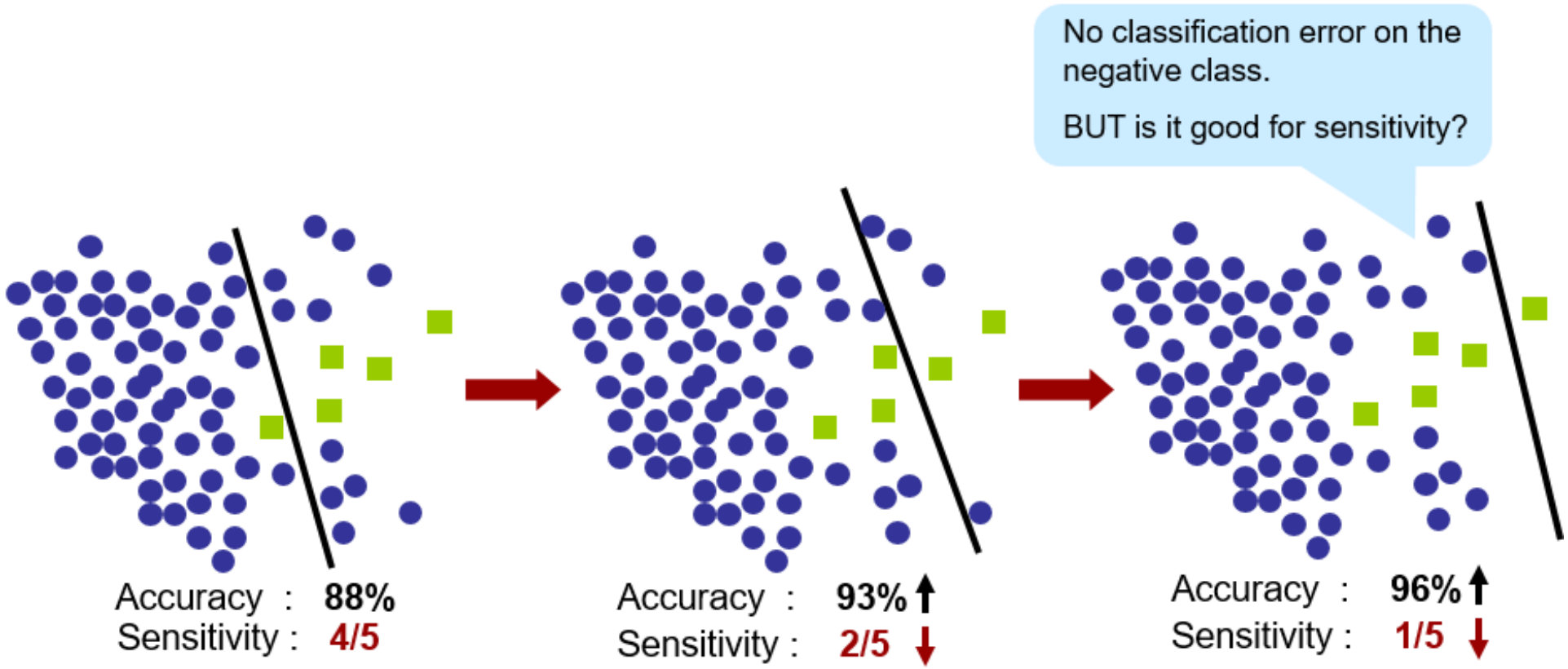
LibSVM Active All: 920 seconds

LASVM Active All: 99 seconds

LASVM Active 50: **9 seconds**

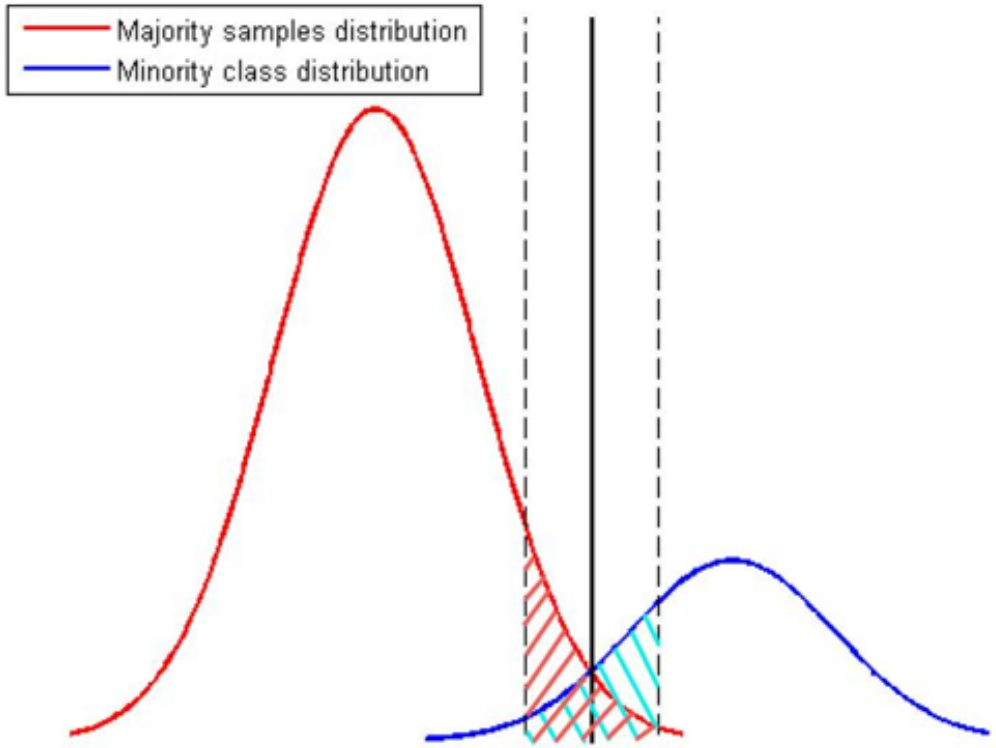
Class Imbalance Problem

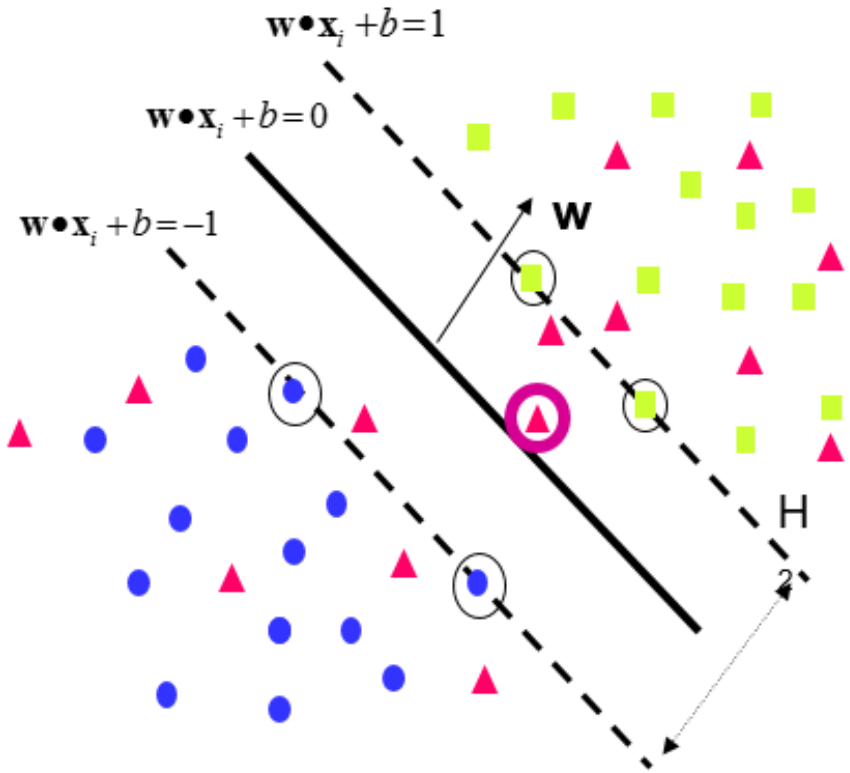
The class boundary learned by standard machine learning algorithms can be severely skewed toward the positive class.



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

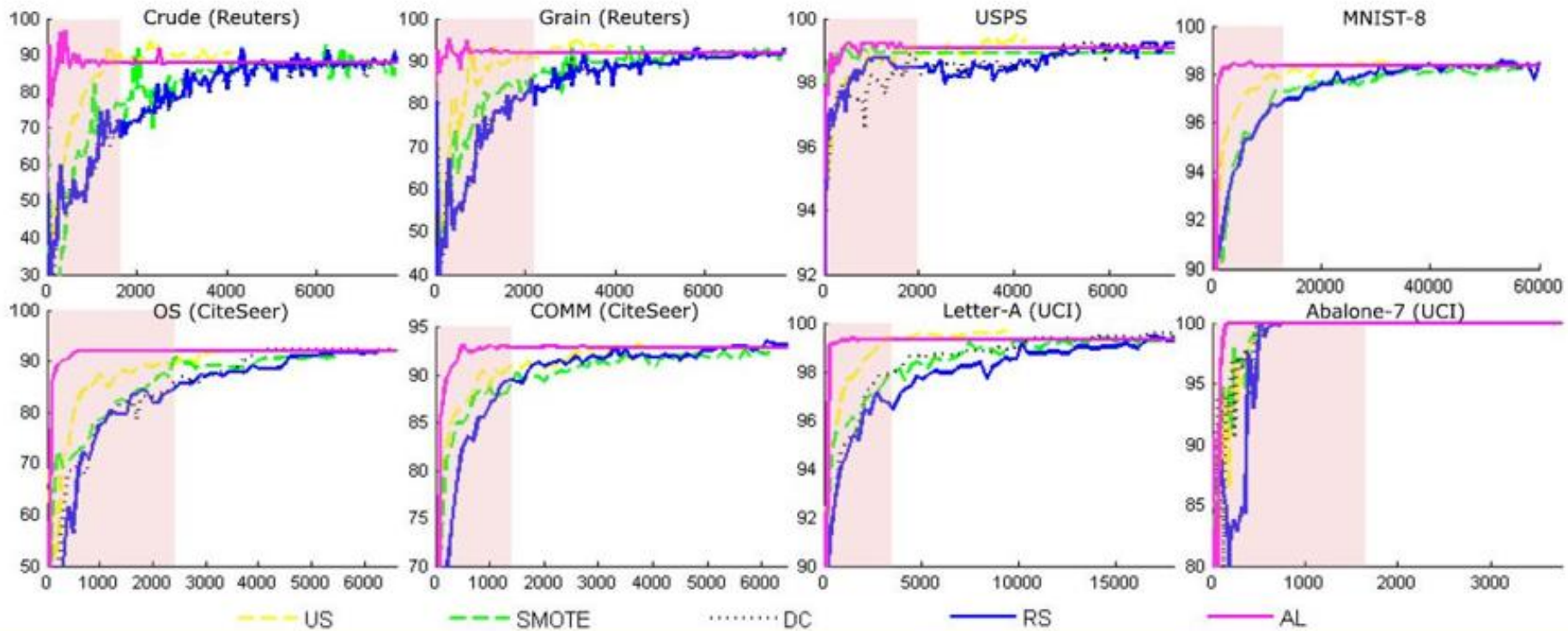
US → UnderSampling

SMOTE → OverSampling

DC → Distinct Costs to the classification errors

RS → Random Sampling

AL → Active Learning

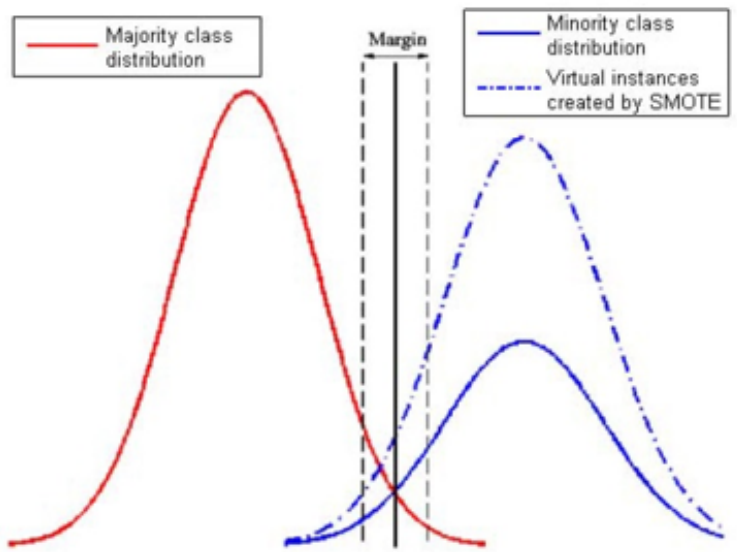


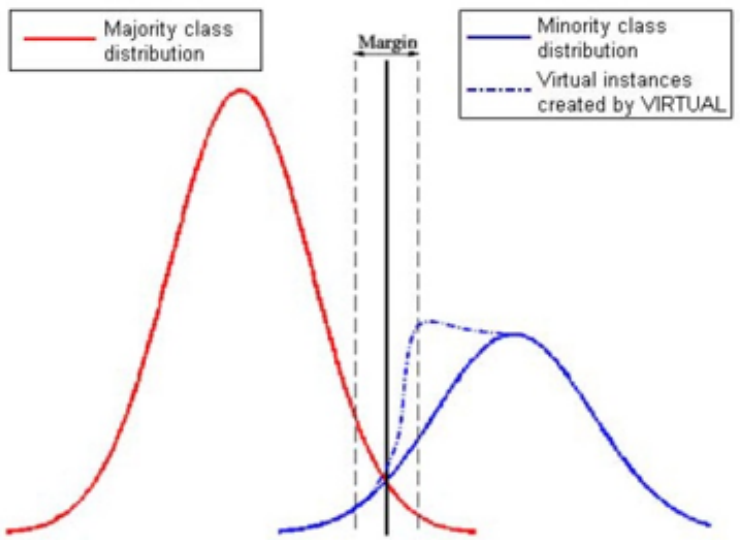
AL converges to the similar level of g-means in much earlier steps of the learning process.

Adaptive OverSampling with Active Learning

VIRTUAL

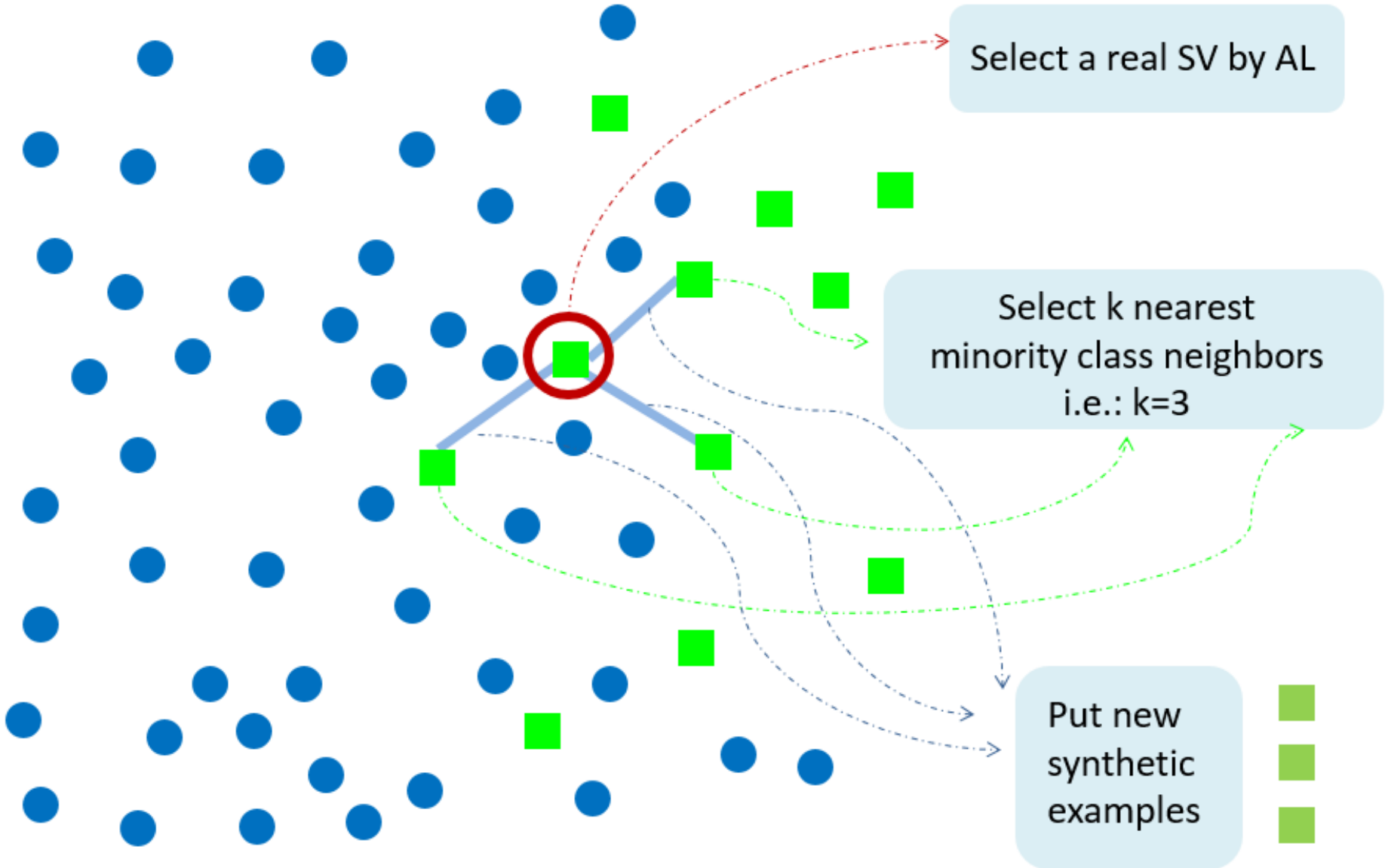
Virtual **I**nstances **R**esampling **T**echnique **U**sing **A**ctive **L**earning



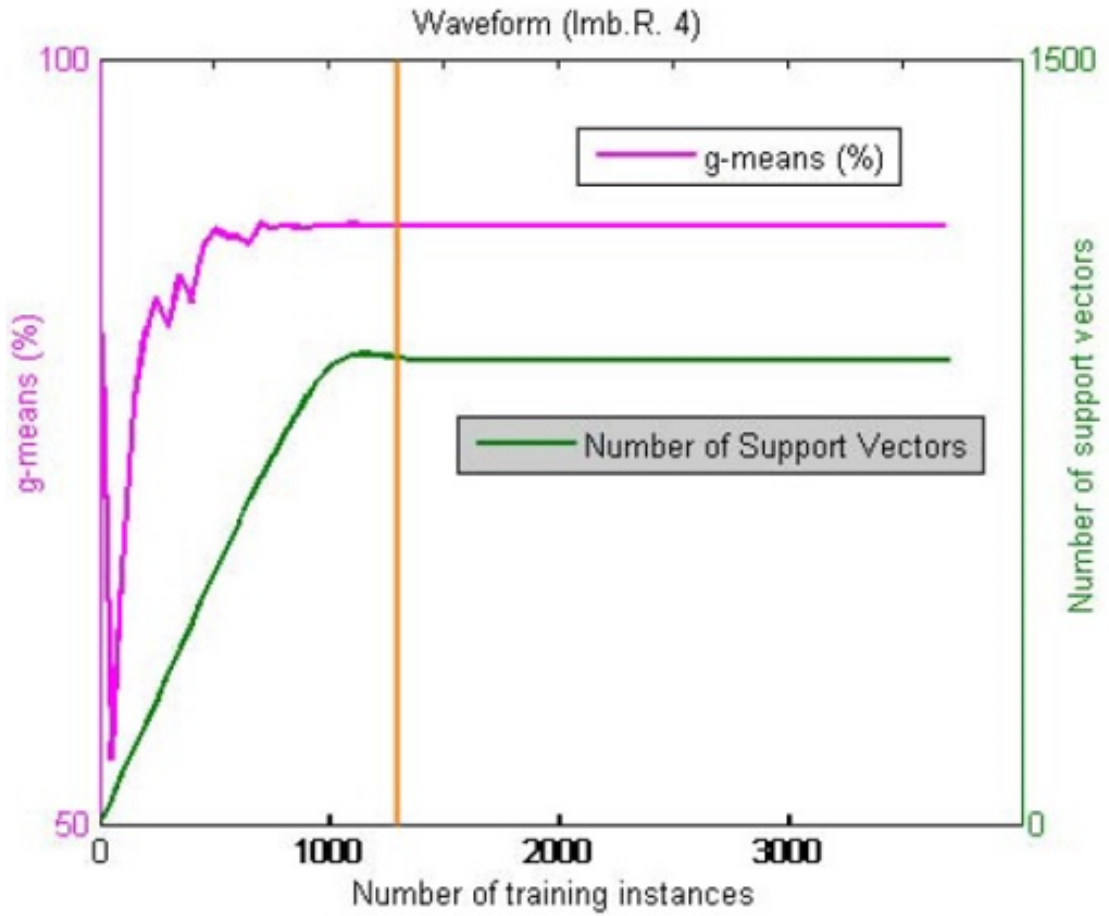


- Generates virtual (synthetic) examples for the minority class support vectors.
- No preprocessing is required. Synthetic examples can be created on the fly.
- Faster runtime than oversampling methods
- Better prediction performance than Active Learning (and oversampling)

VIRTUAL Example Generation

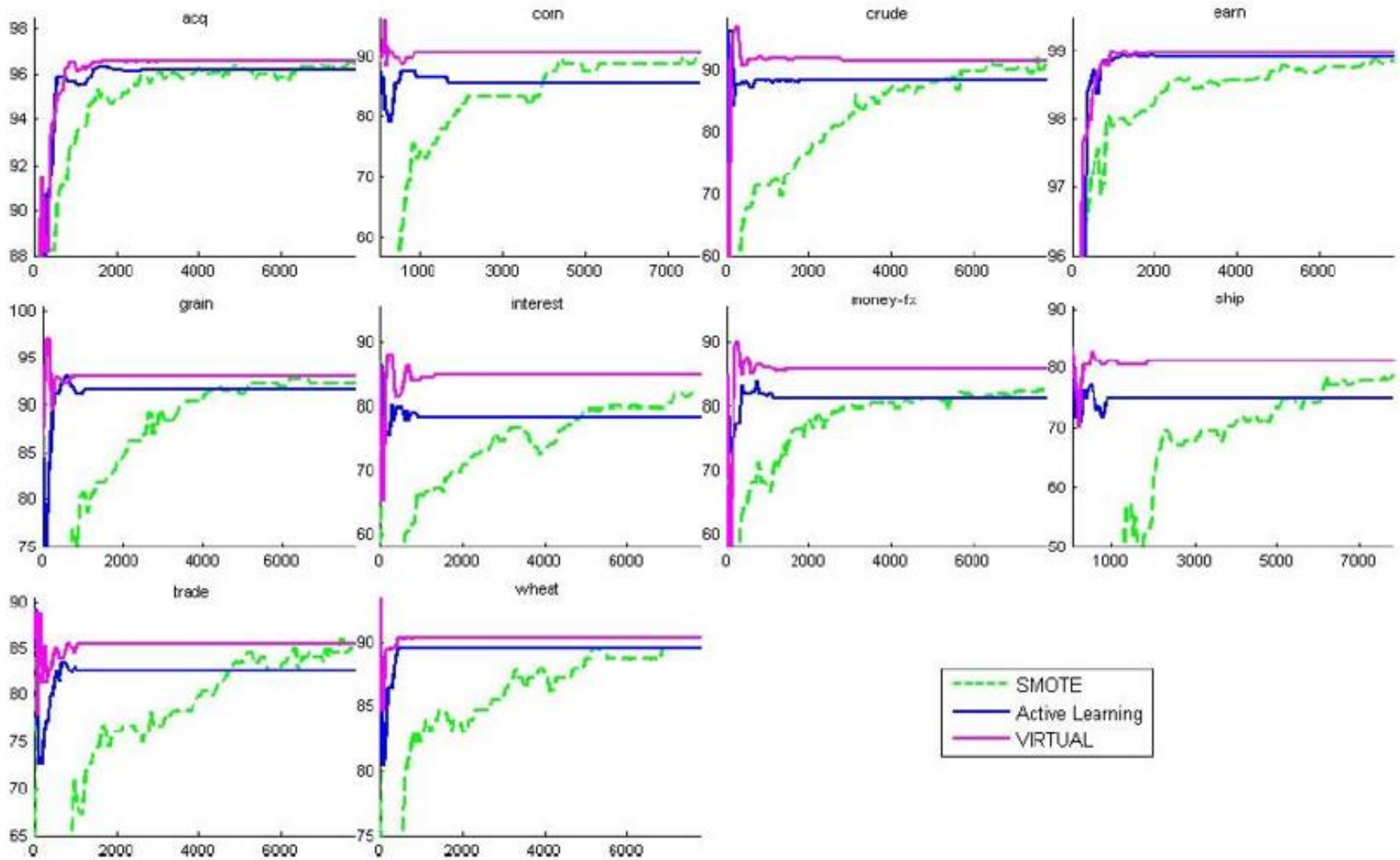


Saturation of g-means and Support Vectors



We can do early stopping!

Prediction Performance of SMOTE, AL and VIRTUAL

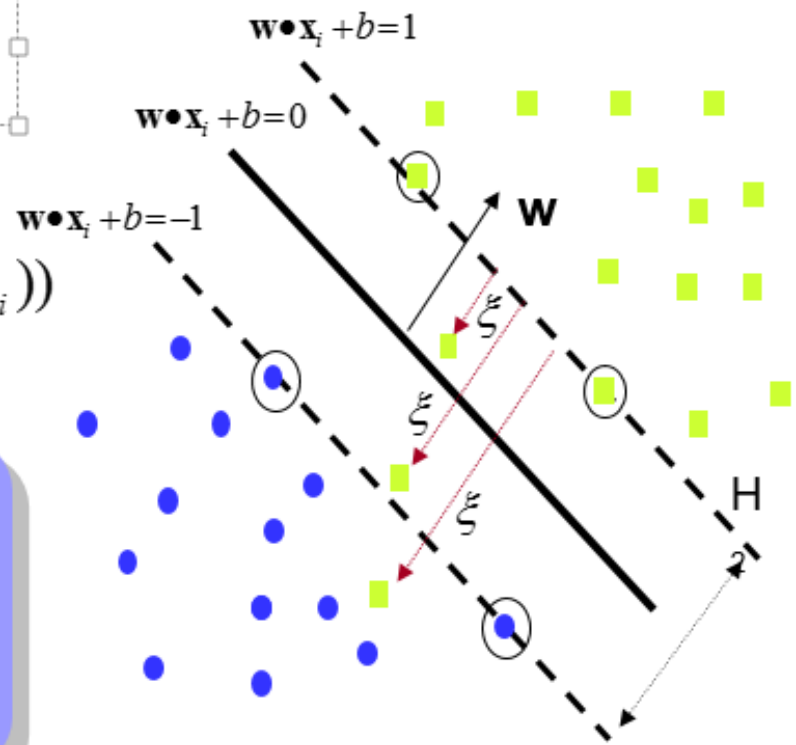


Noise in the Training Data

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.

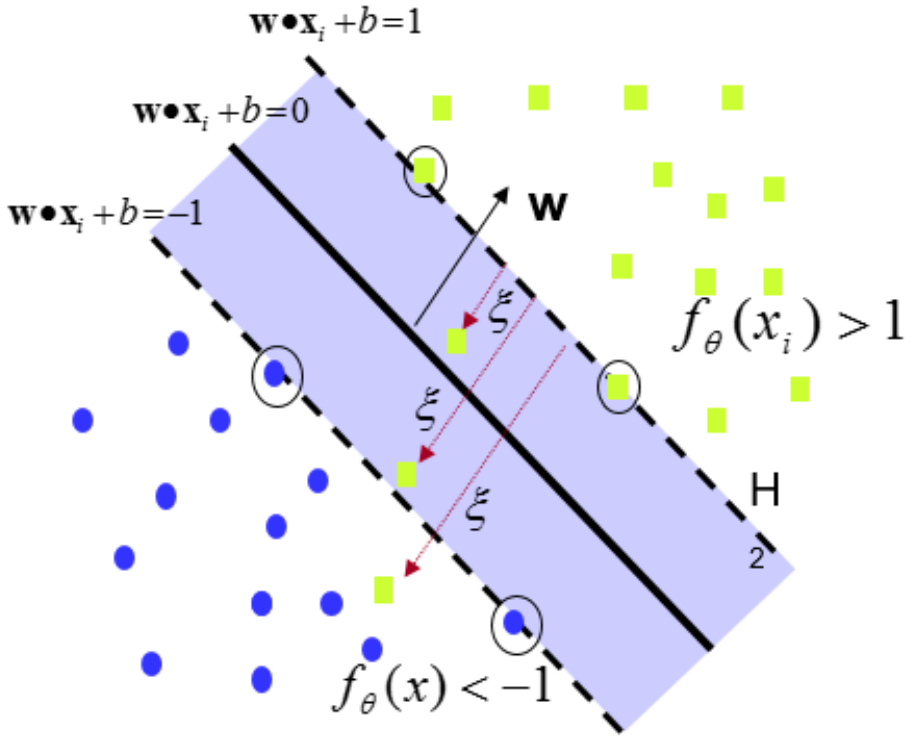
Inseparable case $\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} H(y_i f_\theta(x_i))$



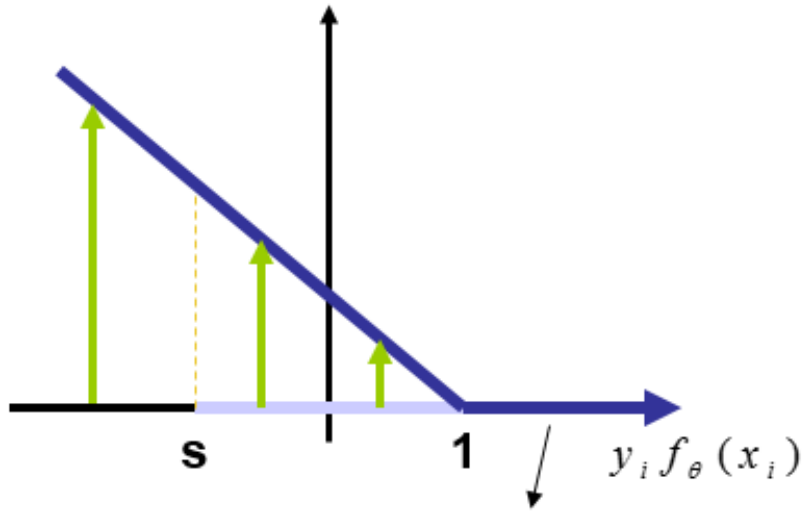
Problem!

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

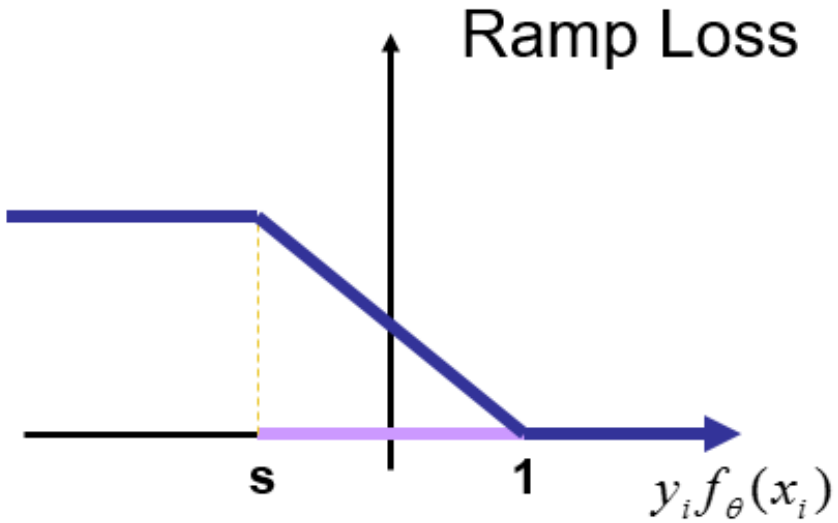
$$y_i f_\theta(x_i) > 1$$

Misclassified examples become support vectors

Curse of dual variables:

# of SVs increases linearly with the # of training examples

Non-Convex Loss Function for SVMs



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

α_i'**s**

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

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)

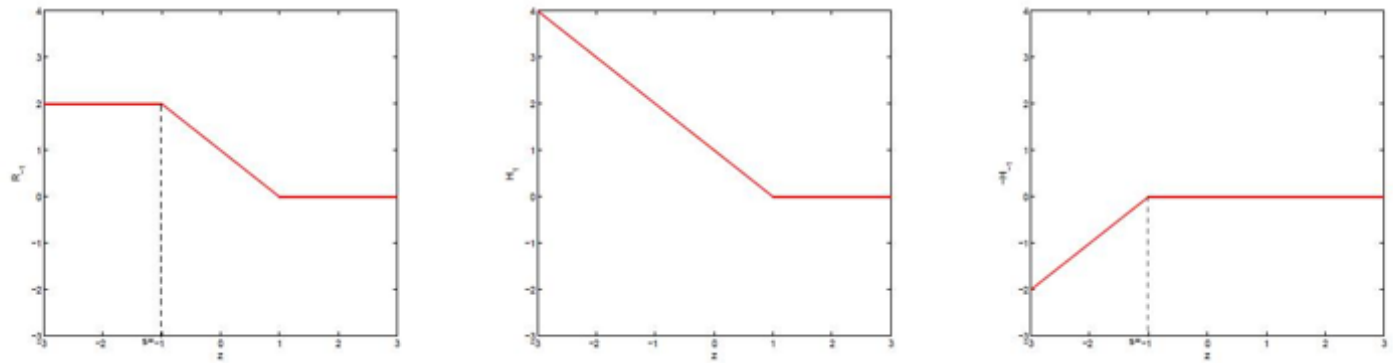
LASVM-G

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 $\boldsymbol{\theta}^0$ with a best guess

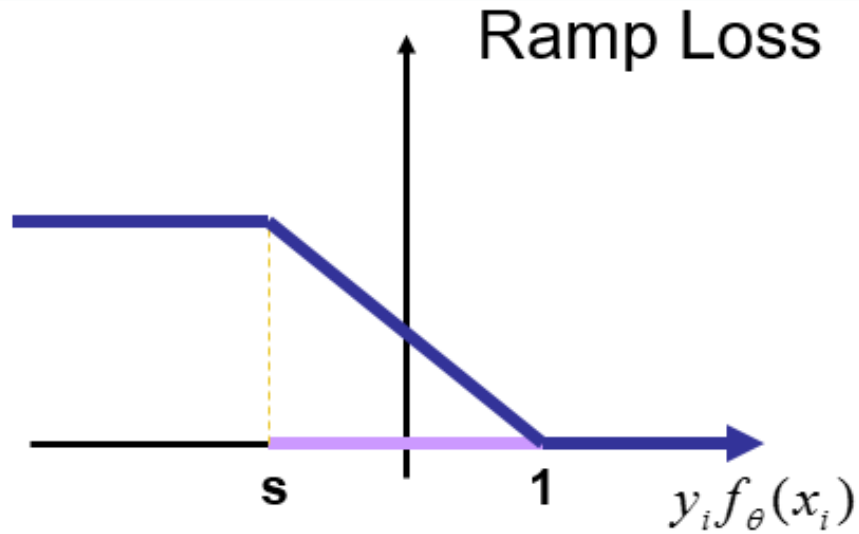
repeat

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

until convergence of $\boldsymbol{\theta}^t$

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

Discussion

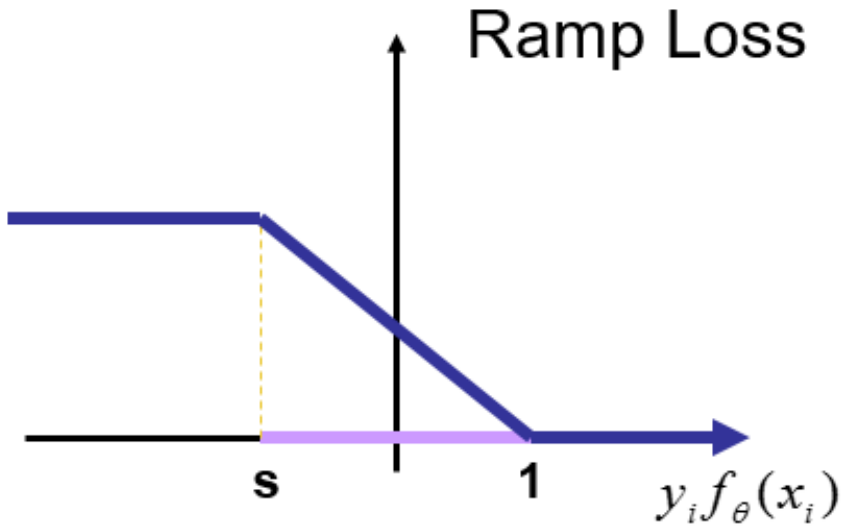


Two important values for *s* at noisy datasets:

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.

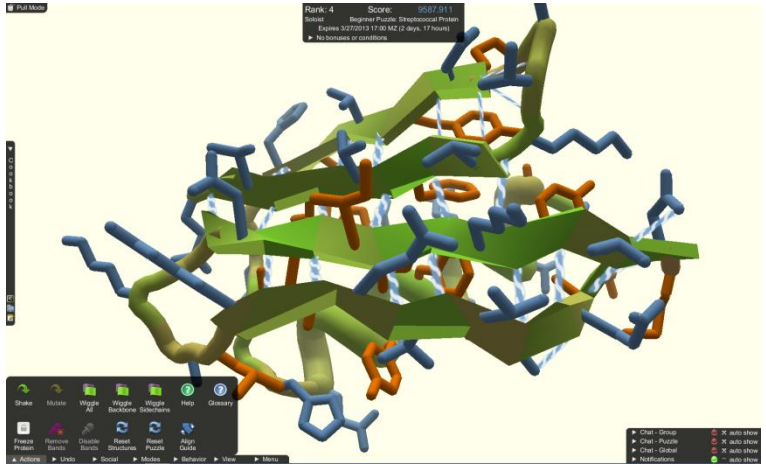
s=-1: Only examples in the margin are processed.

s=-1

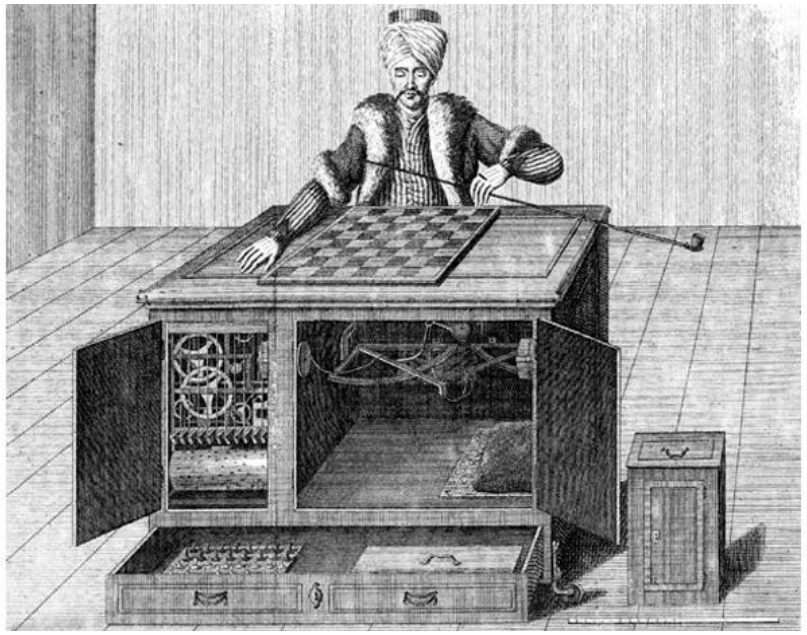
Works like Active Learning!

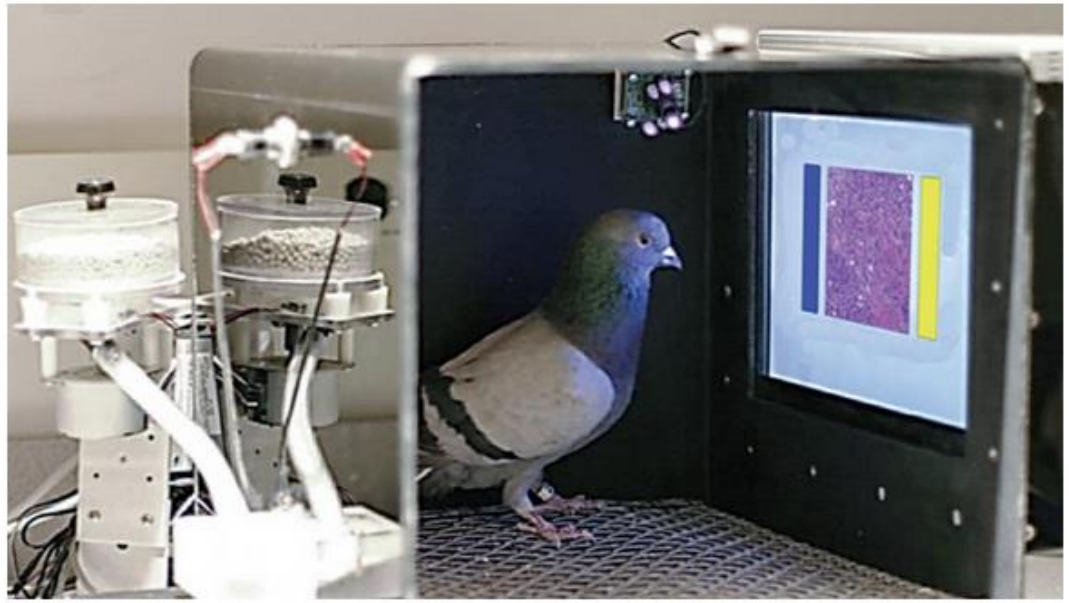
Machine Learning with Humans in the Loop

Ex: analyzing microscopy images



Ex: FoldIt Game





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*.

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

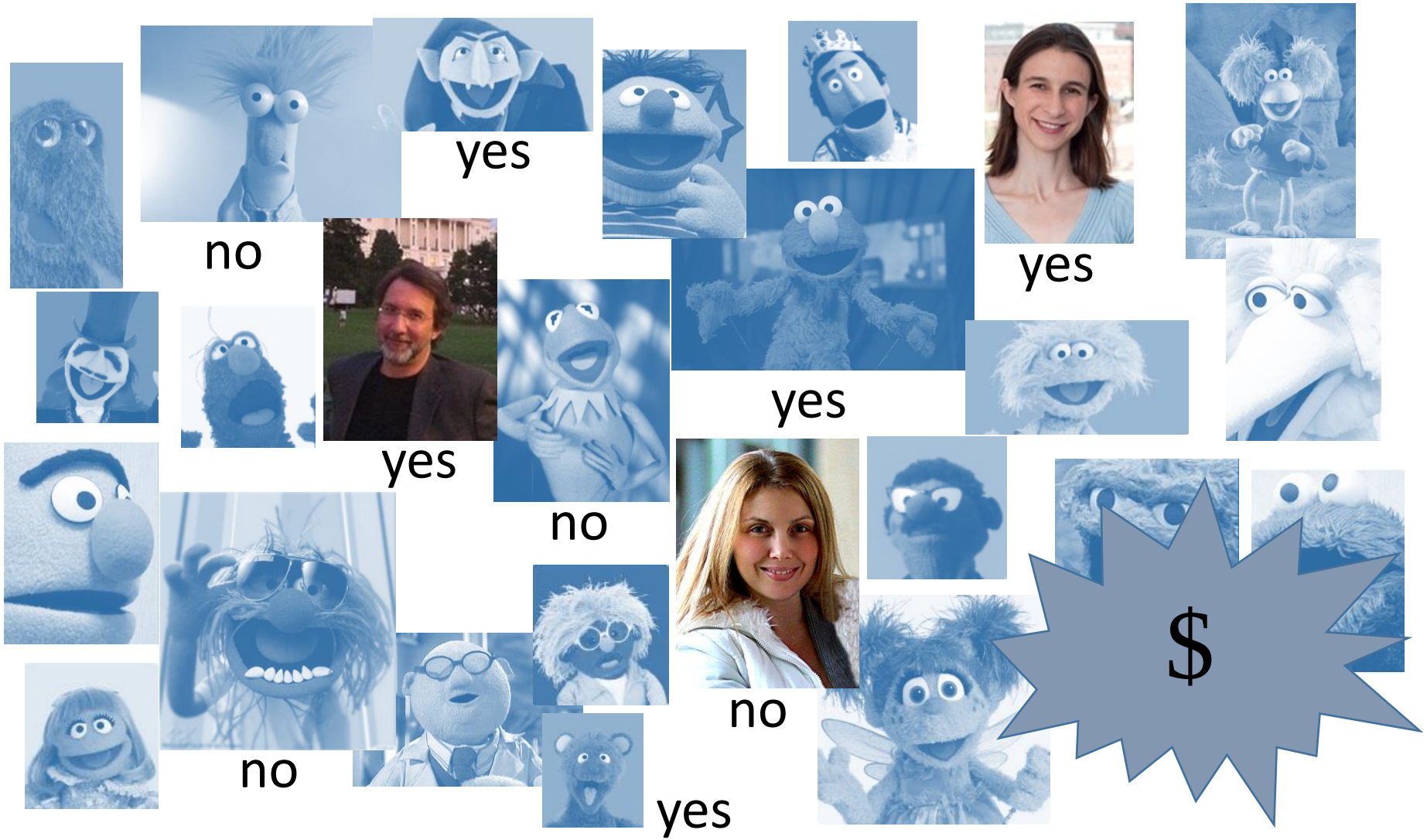
Approximating the Crowd

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

- Start with 3 labelers

- Keep adding more labelers until we are certain.

Outline of CrowdSense

For $t = 1,...,T$

end

Outline of CrowdSense

For $t = 1,...,T$

-Start with 3 labelers

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

end

Outline of CrowdSense

For $t = 1,...,T$

-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 $t = 1,...,T$

-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 $t = 1,\ldots,T$

-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.

S_t

current sub-crowd

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

S_t

current sub-crowd

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

S_t

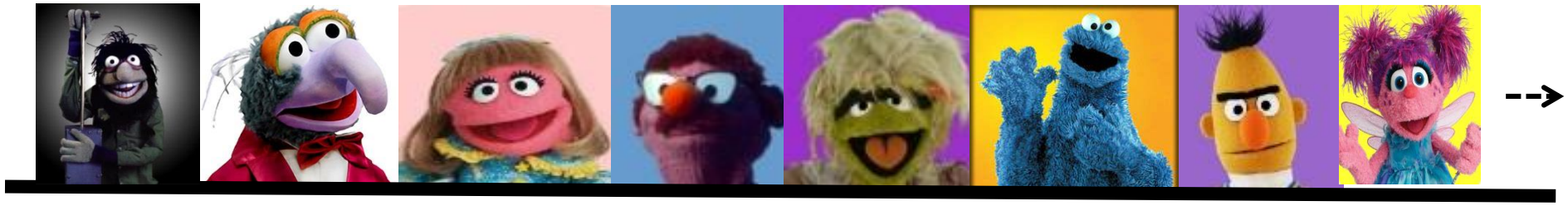
current sub-crowd

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

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

S_t

current sub-crowd



Sorted labelers according to their quality estimates

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

S_t

current sub-crowd

candidate labeler

Sorted labelers according to their quality estimates

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

S_t

current sub-crowd

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

candidate labeler

Sorted labelers according to their quality estimates

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

S_t

current sub-crowd

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

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

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

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

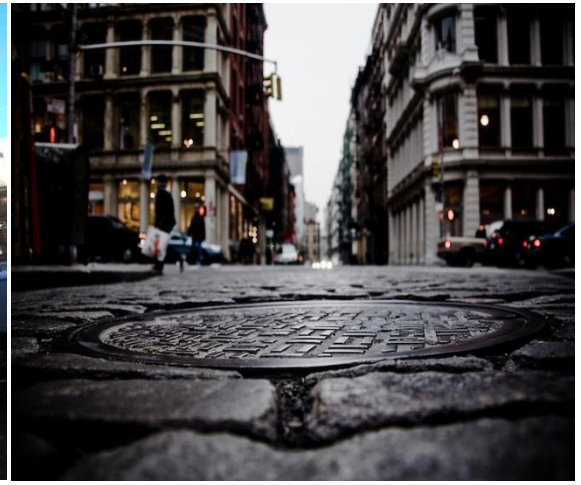
ε is a tradeoff between the total cost we are willing to spend and the desired accuracy.

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

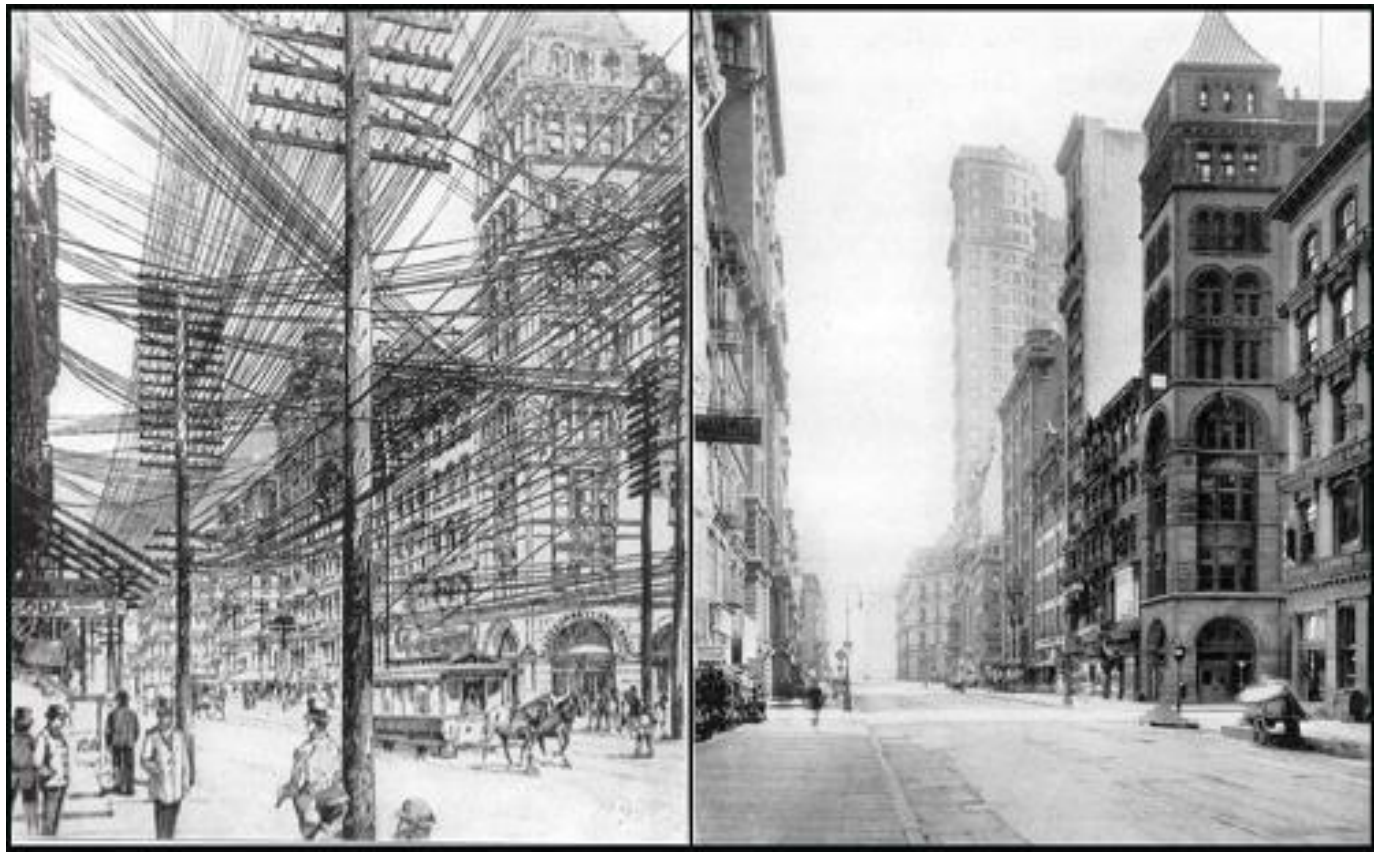
January 27th, 2013 Omaha, NE
http://www.theblaze.com

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:

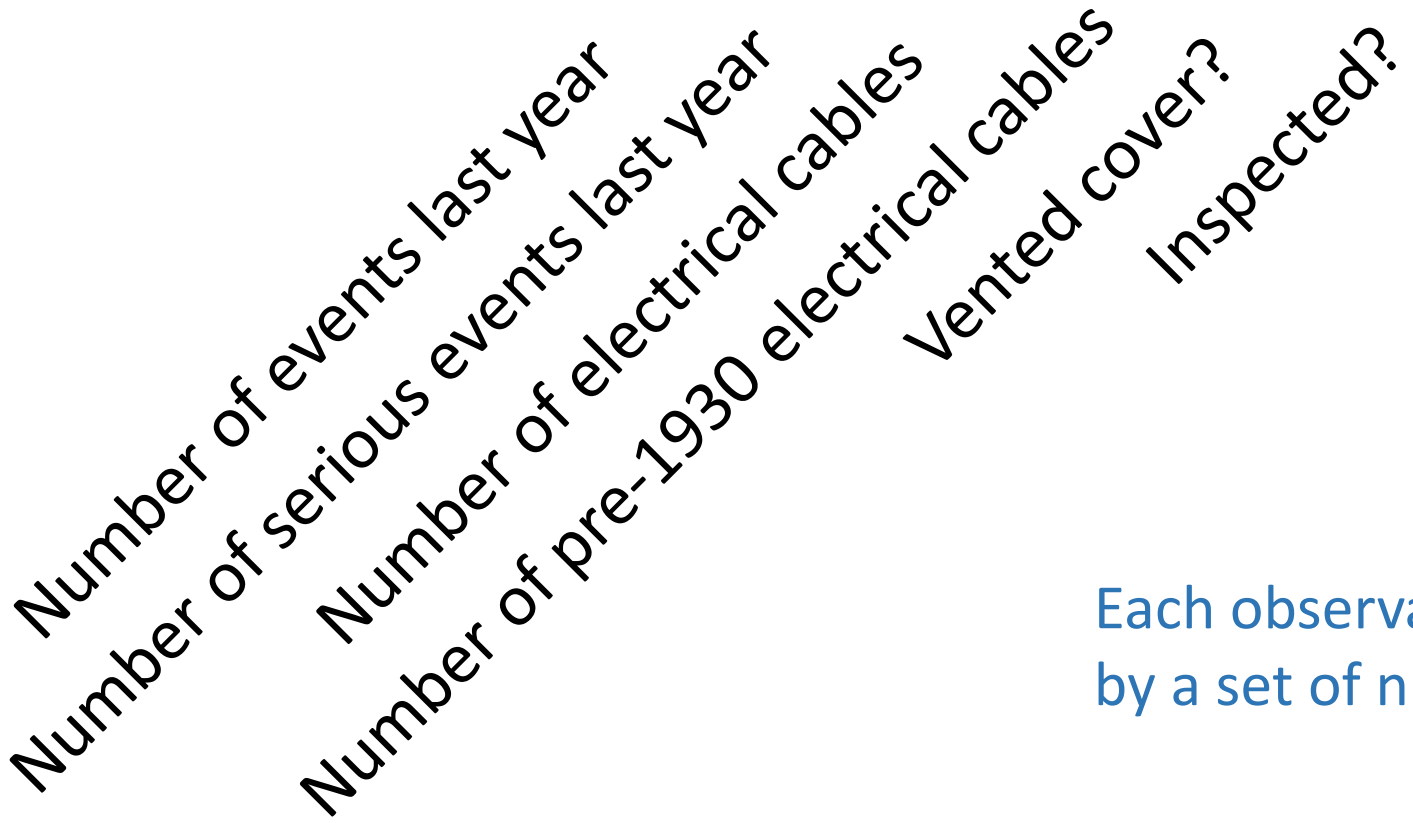
i)Can we predict failures continuously in time?

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 …..]



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

Manhole is represented as:

[	5	3	120	12	1	0	…..	]
[	0	0	89	5	1	1	…..	]
[	1	0	20	0	0	1	…..	]

:

:

Features, called X

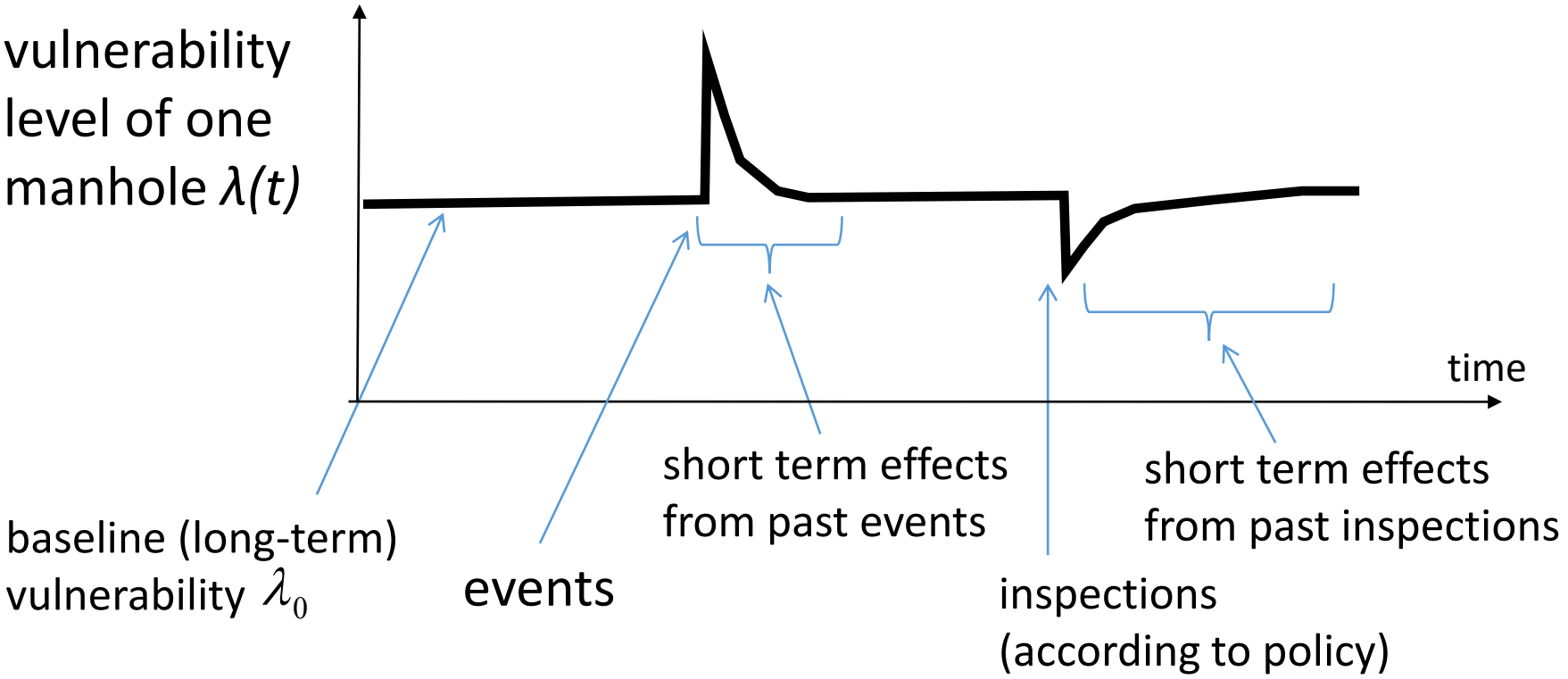
Labels, called Y

(Predictors, Covariates,
Explanatory Variables,
Independent Variables)

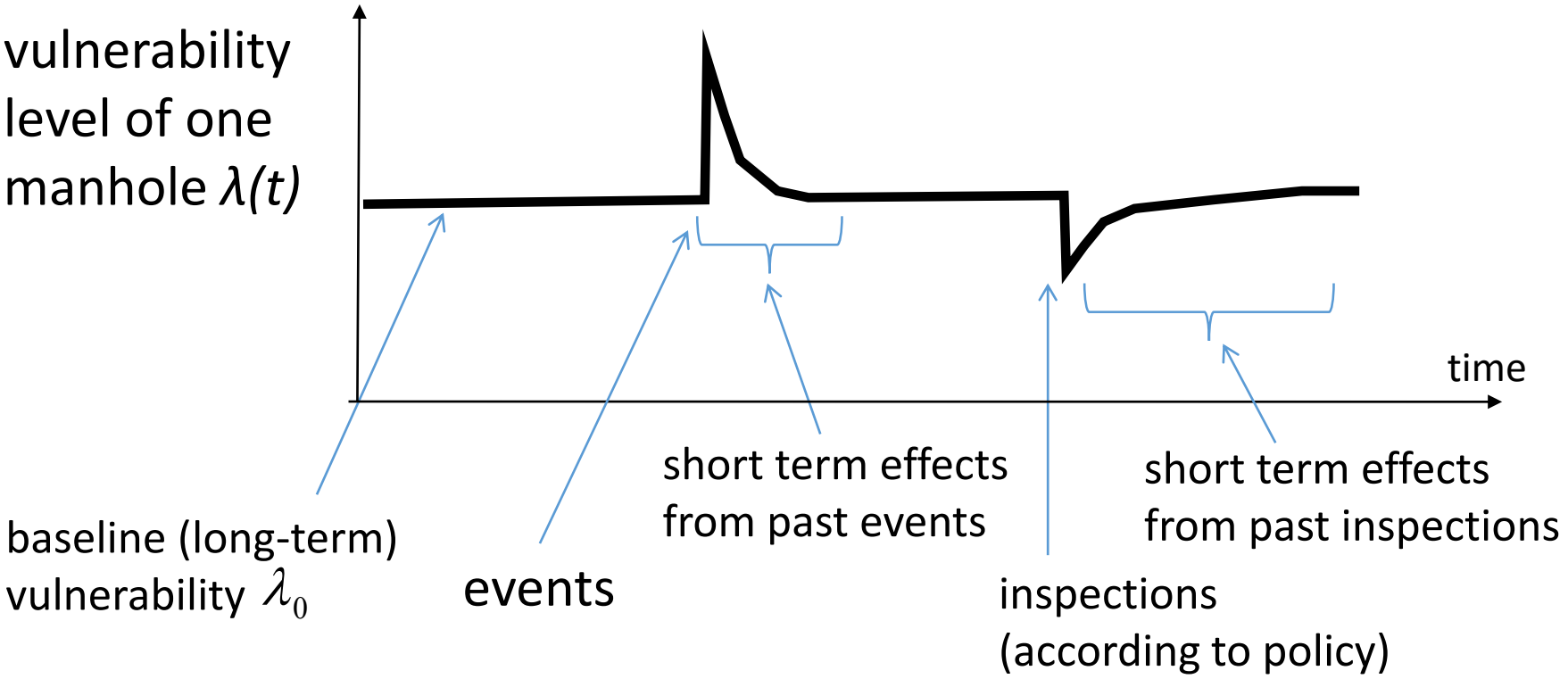
Properties of Reactive Point Processes (RPPs)

- 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)

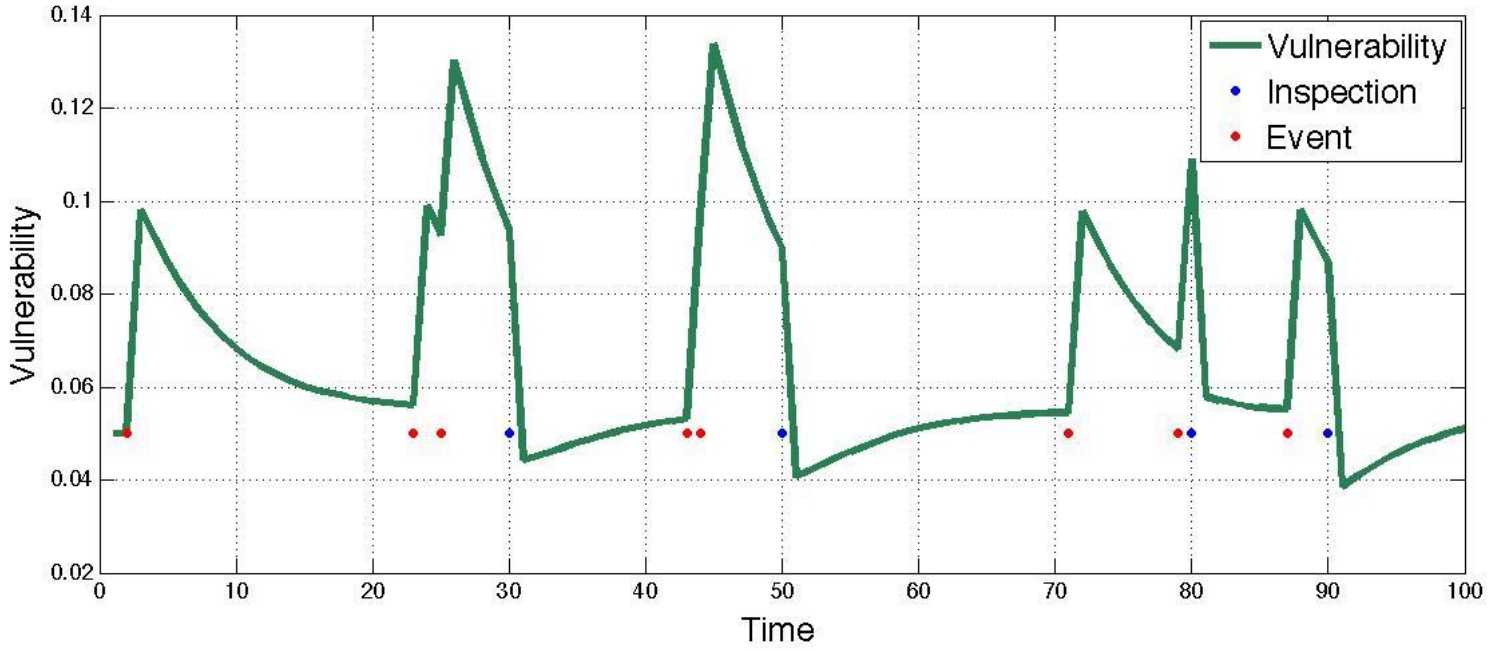
Short-term Vulnerability Model



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:

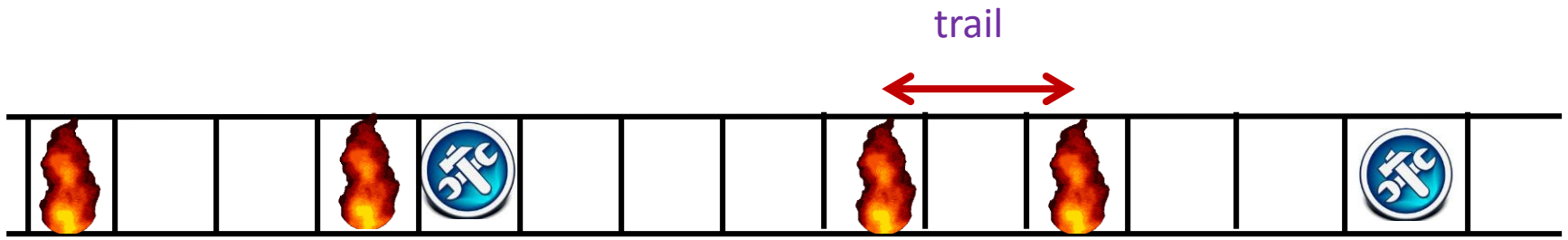
- Locate event/inspection **"trails"** in the time series data for each manhole.

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

Tracing out g2 function

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

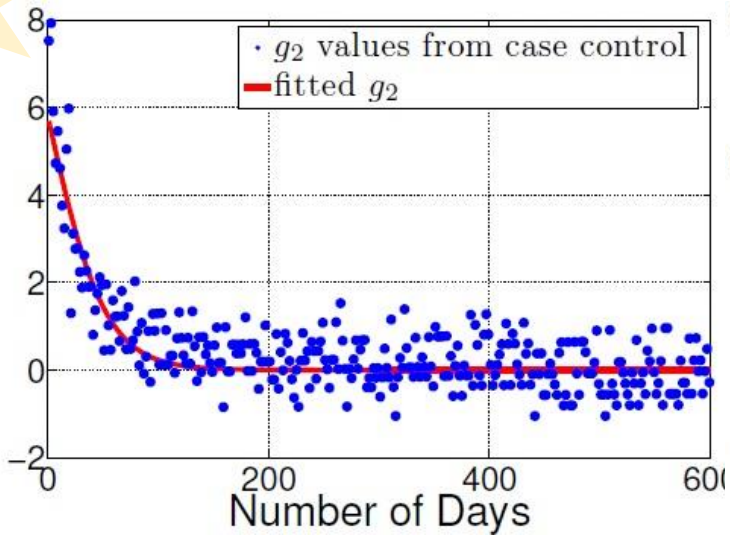
- Choose trails starting at events that occurred at baseline vulnerability levels (no effect of previous event or inspection)
- Consider only the trails where there was already a previous serious event in that manhole (to use C1)



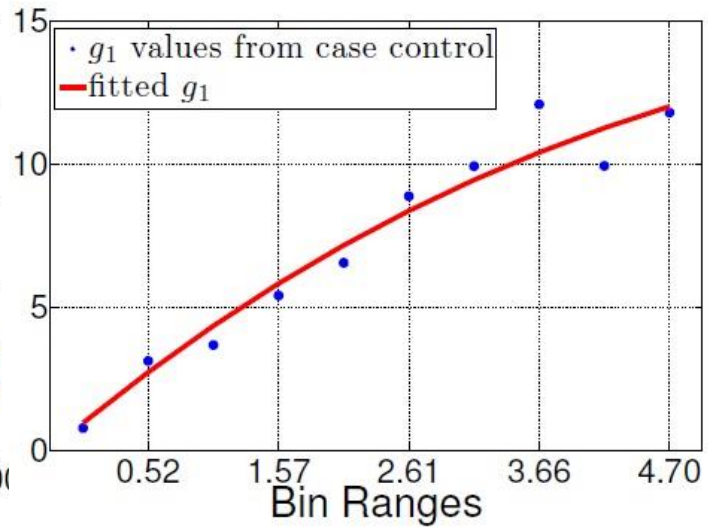
$$g_2^{\text{nonparametric}}(\Delta t) = \frac{1}{\lambda_0} \frac{E_{\Delta t}}{E_{\Delta t} + \bar{E}_{\Delta t}} - C_1 - 1$$

Fitted g2 and g1 functions
(Nonparametric Model)

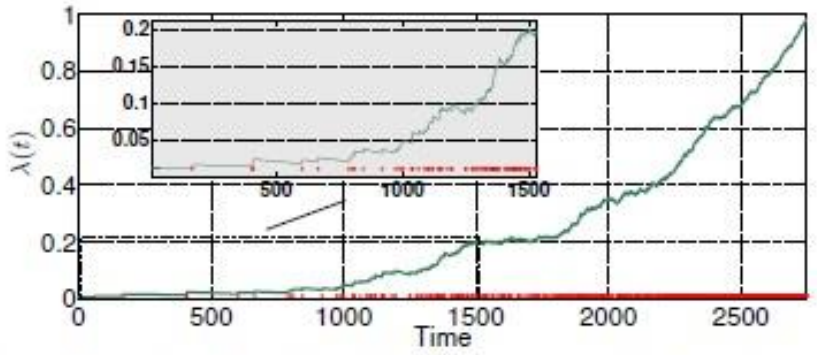
g2



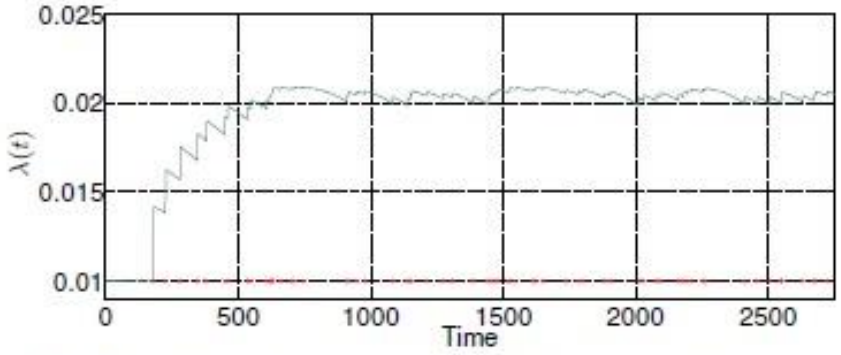
g1



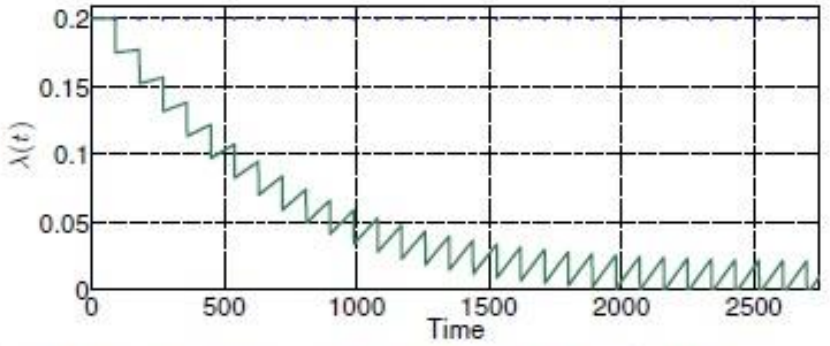
The effect of Saturation Functions, g1 and g3



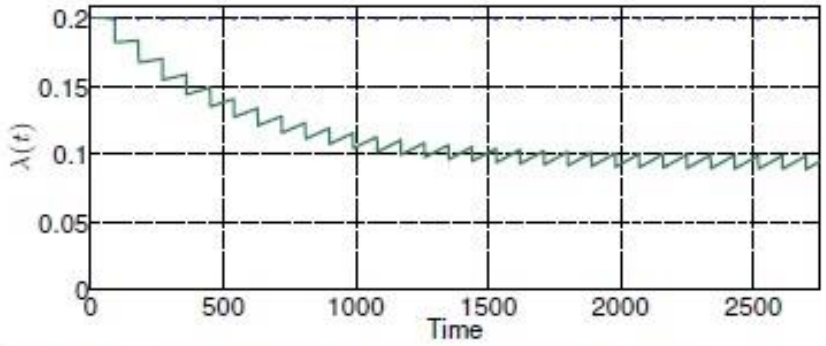
(a) Model with self-excitation function g_2 (without saturation function g_1 and inspections) and a zoomed view of the first 1,500 days.



(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

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

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

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

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

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

Incorporating covariates

$$g_2^{(p)}(x) = \frac{1}{1 + e^{\beta^{(p)} x}} \qquad g_4^{(p)}(x) = \frac{-1}{1 + e^{-\gamma^{(p)} x}}$$

$$\boldsymbol{\beta} = \log\left(1 + e^{-\mathbf{M}\boldsymbol{\upsilon}}\right) \qquad \boldsymbol{\gamma} = \log\left(1 + e^{-\mathbf{M}\boldsymbol{\omega}}\right)$$

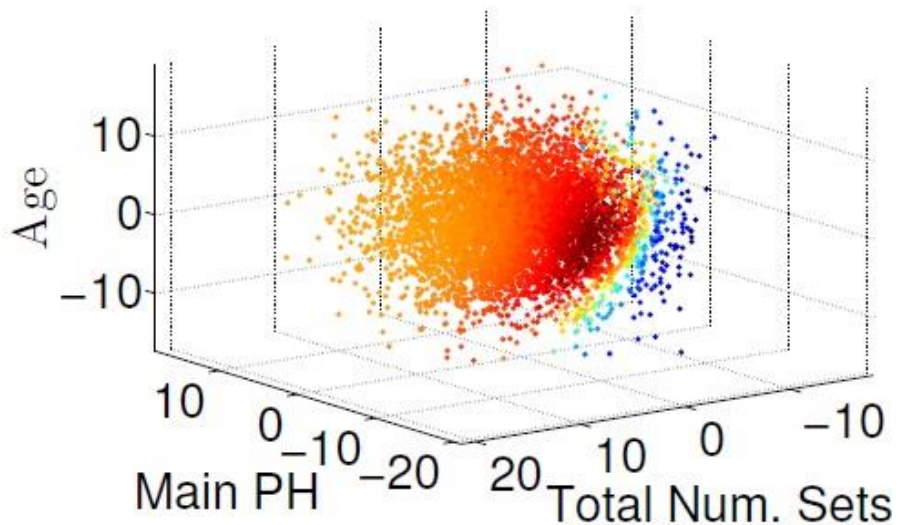
This permits slower or faster decay of the self-exciting and self-regulating components based on the characteristics of the manhole.

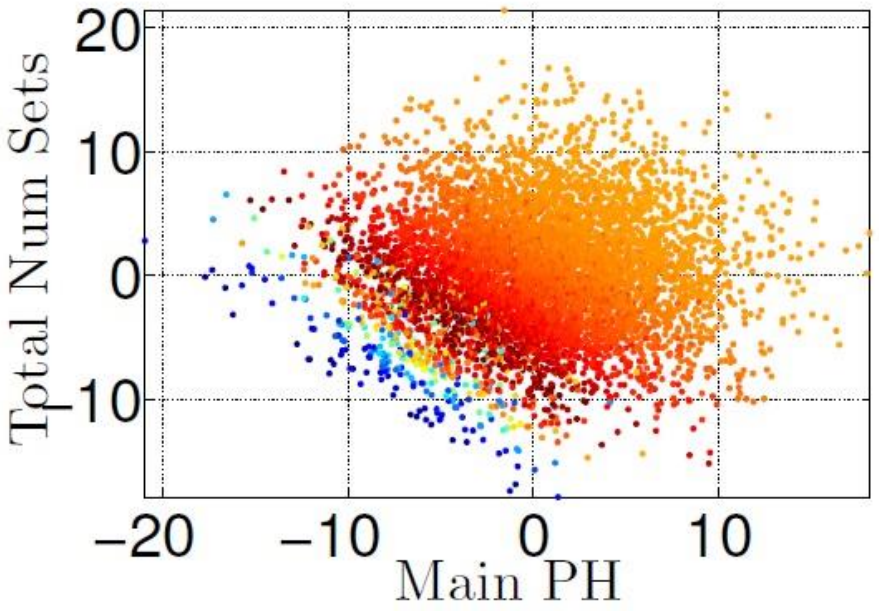
Covariates we used:

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

Likelihood-based RPP

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





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

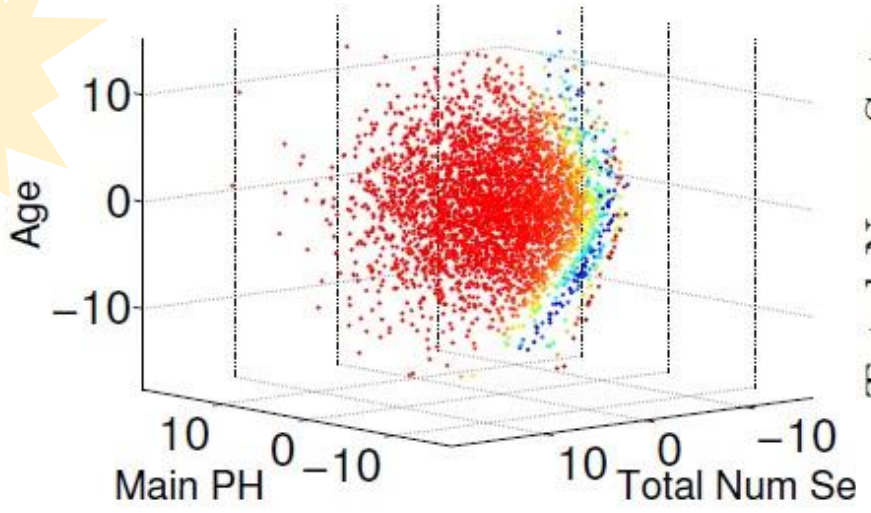
DNE: The difference in the number of events in the simulated and observed data

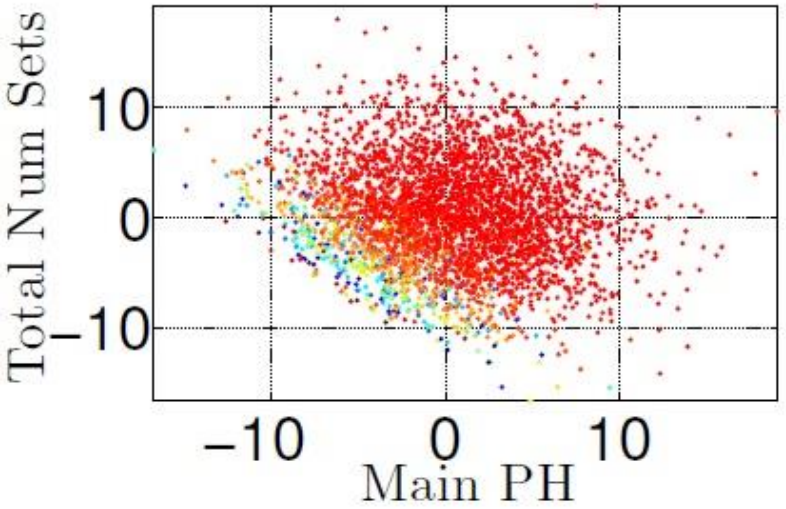
KL: The Kullback-Leibler divergence between two histograms, one from the observed data, and one from the real data (Histograms of time differences between events).

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

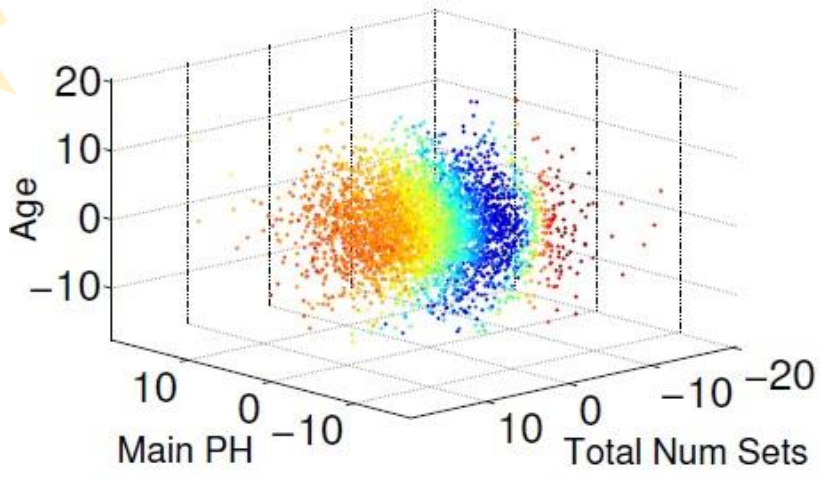
DD & KL for Manhattan Dataset

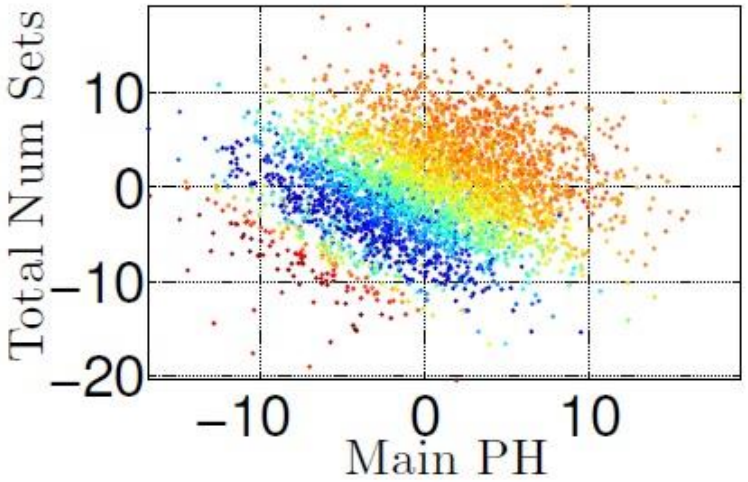
DD



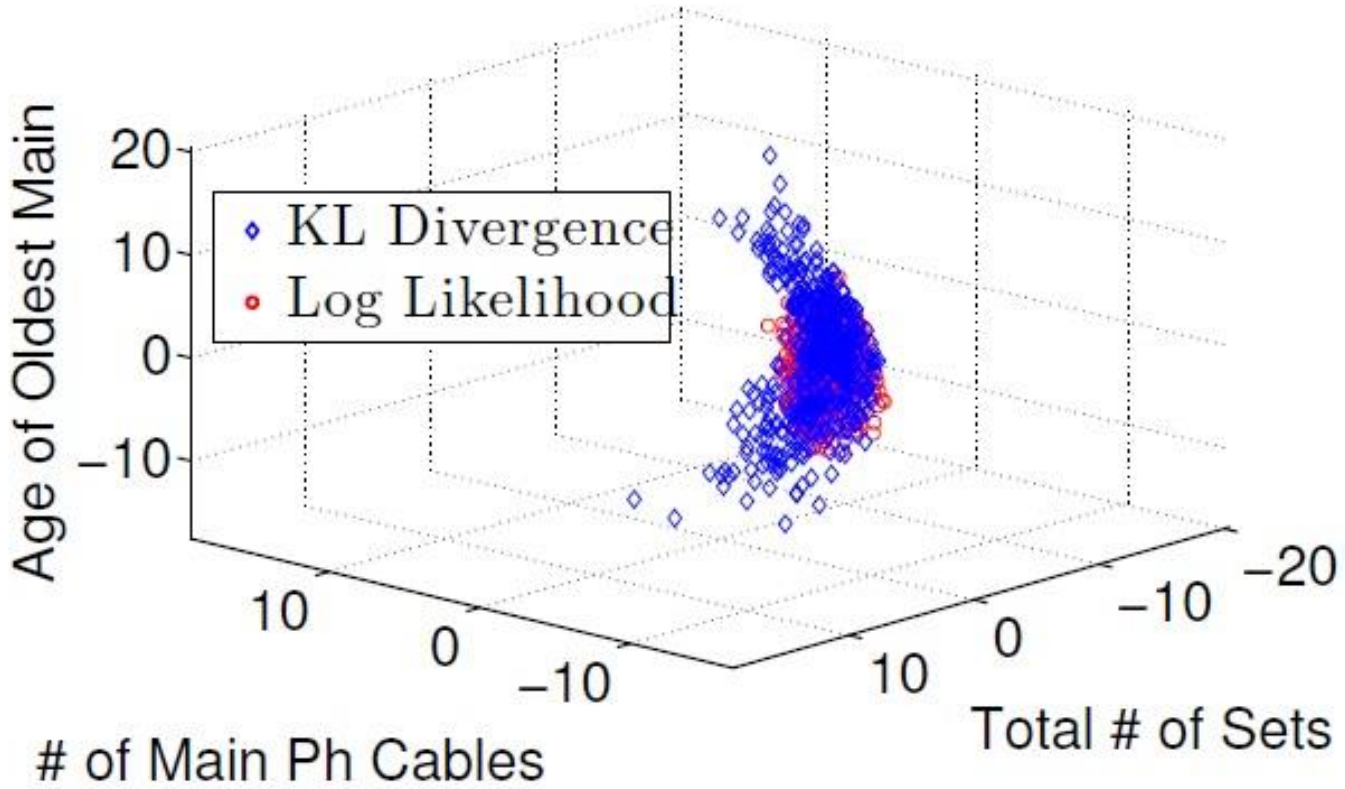


KL

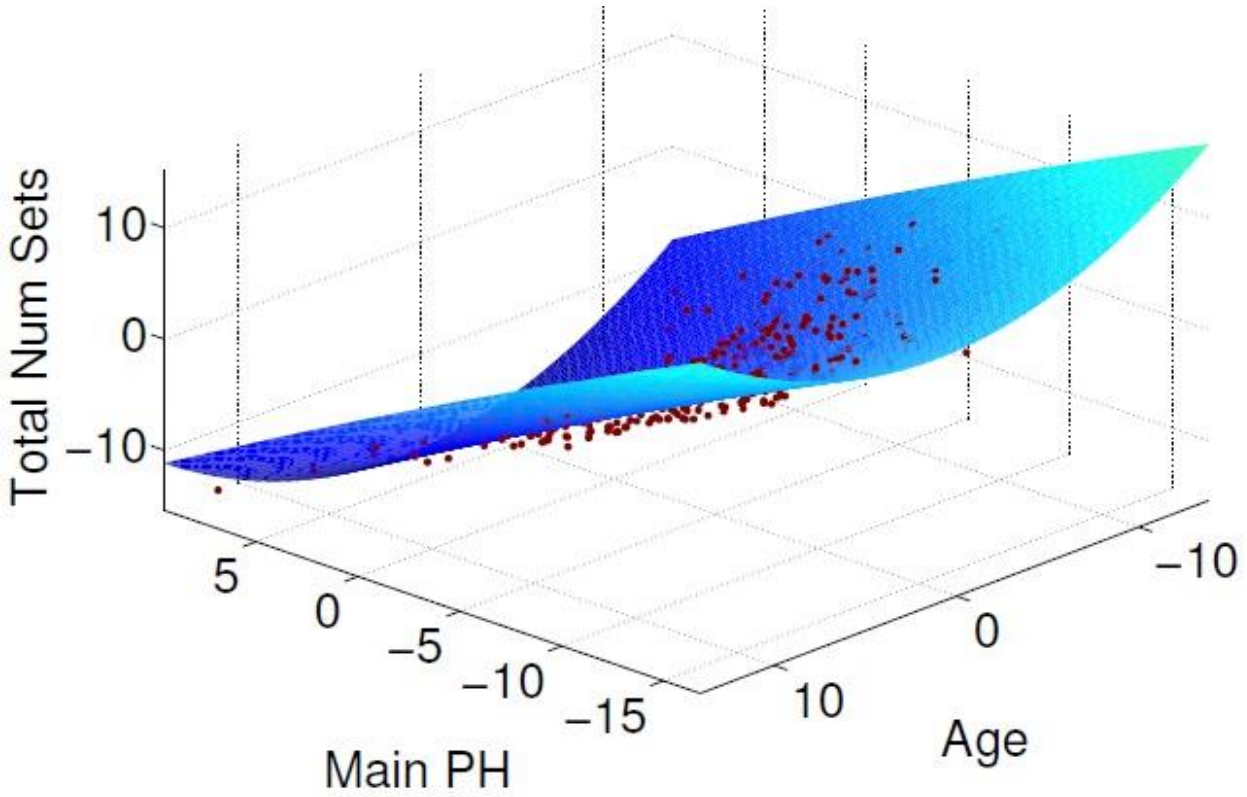




Choosing Parameter Values



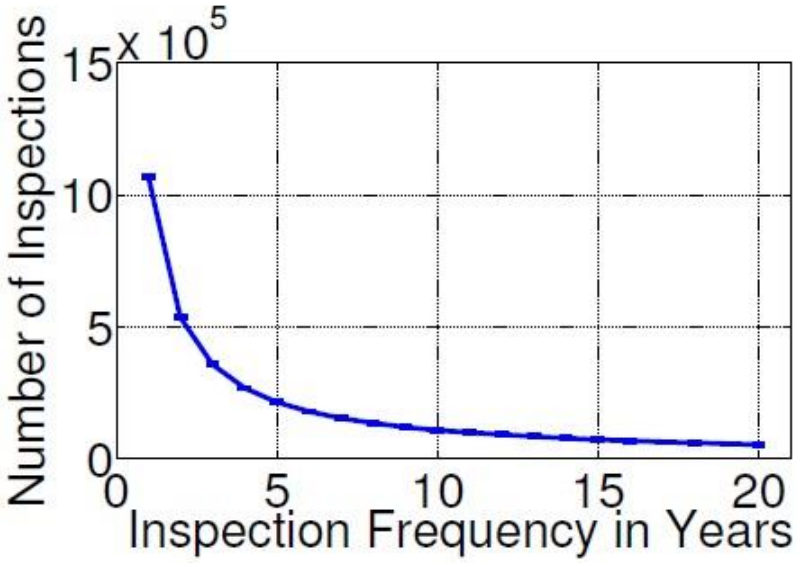
Choosing Parameter Values

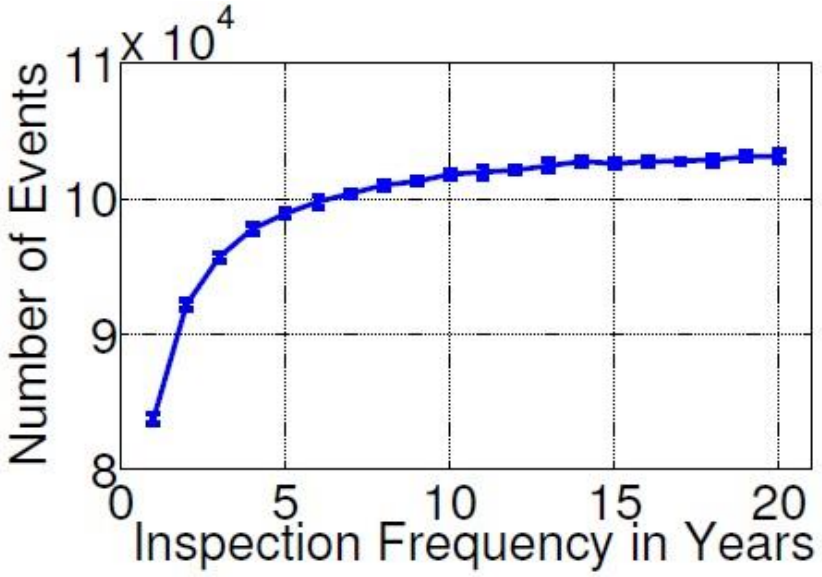


Making Policy Decisions with RPP

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

Thank you!


ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY