Statistical learning, high dimension and big data

Stéphane Gaïffas

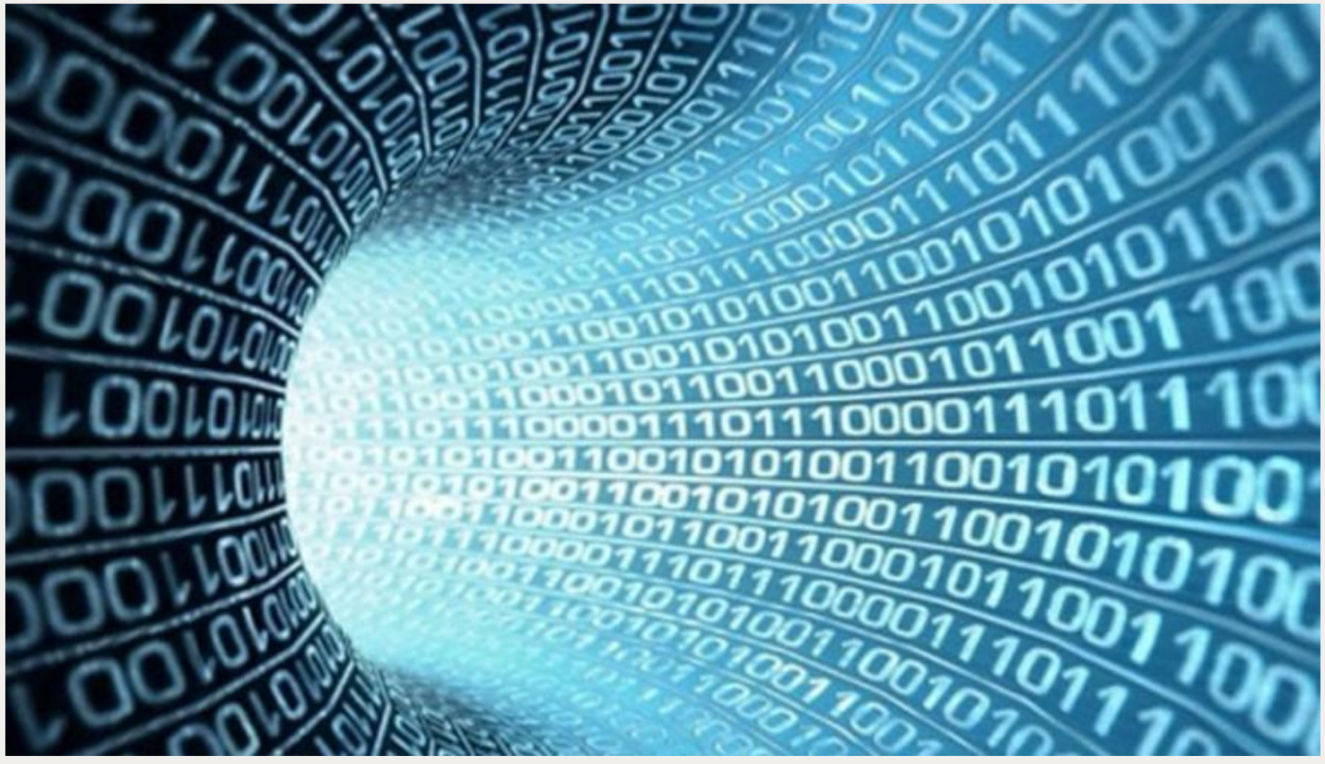
Who I am ?

- Stéphane Gaïffas
- Professor
- LPSM - Univ. Paris Diderot and CMAP - Ecole polytechnique
- `http://www.cmap.polytechnique.fr/~gaiffas/`
- `stephane.gaiffas@polytechnique.edu`

Le Big Data, nouvel eldorado des entreprises

Par Direct Matin, publié le 26 Septembre 2014 à 08:32

GOOGLE+ FACEBOOK TWITTER PINTEREST LINKEDIN



Les mégadonnées représentent un marché de plusieurs milliards d'euros[© infocux technologies]

Considéré comme le "nouveau pétrole du XXIe siècle", le Big data attise toutes les convoitises.

EN COMPLÉMENT

Les données, puissance du futur

LE MONDE | 07.01.2013 à 15h10 • Mis à jour le 07.01.2013 à 18h03

Par Stéphane Grumbach, Stéphane Frénot

Les plus partagés

1 Une équipe de scientifiques filme un calamar géant par 900 mètres de fond dans le Pacifique — 2212

2 Infirmiers et aides-soignants refusent d'être des "pigeons" — 1664

3 Messi remporte son 4e Ballon d'or consécutif — 987

4 Mariage homosexuel : Wauquiez veut "forcer" le débat sur un référendum — 629

5 La première Eglise athéiste ouvre à Londres — 603

Nous suivre

Retrouvez le meilleur de notre communauté

OCTOBER 24, 2012, 9:00 AM | 4 Comments

Big Data in More Hands

By QUENTIN HARDY

Business people, Big Data is coming for you.

Software that captures lots of data and uses it to make predictions has mostly been the province of engineers skilled in arcane databases and statisticians capable of developing complex algorithms. As the business gets bigger, however, software makers are domesticating their products in the hope they will prove attractive to a broader population.

Cloudera, which offers a popular version of the open source database called Hadoop, released software on Wednesday that makes it possible to run queries from a more mainstream SQL programming language interface. SQL, thanks to its adoption by Oracle, Microsoft and others, is known to millions of business analysts.

"This enables us to talk to a whole other class of customer," said Mike Olson, the chief executive of Cloudera. "The knock against Hadoop was that it is too complex."

There is a reason for that. Hadoop is one of several so-called unstructured databases that were created at Yahoo and Google, after those two companies found they had previously unimaginable amounts of data about activities like people's Web-surfing habits. Put into databases designed to handle this unstructured behavior, then analyzed, this information was

PREVIOUS POST
Google Shifts Pitch for Its New Chromebooks

NEXT POST
In Contest for Rescue Robots, Darpa Offers $2 Million Prize

NEWS ANALYSIS

The Age of Big Data

By STEVE LOHR
Published: February 11, 2012 | 82 Comments

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

Enlarge This Image

Chad Hagen

Multimedia

Mo Zhou was snapped up by I.B.M. last summer, as a freshly minted Yale M.B.A., to join the technology company's fast-growing ranks of data consultants. They help businesses make sense of an explosion of data — Web traffic and social network comments, as well as software and sensors that monitor shipments, suppliers and customers — to guide decisions, trim costs and lift sales. "I've always had a love of numbers," says Ms. Zhou, whose job as a data analyst suits her skills.

To exploit the data flood, America will need many more like her. A report last year by the McKinsey Global Institute, the

UNBOXED

How Big Data Became So Big

By STEVE LOHR

Published: August 11, 2012

THIS has been the crossover year for Big Data — as a concept, as a term and, yes, as a marketing tool. Big Data has sprung from the confines of technology circles into the mainstream.

Enlarge This Image

Lloyd Miller for The New York Times

First, here are a few, well, data points: Big Data was a featured topic this year at the World Economic Forum in Davos, Switzerland, with a report titled "Big Data, Big Impact." In March, the federal government announced $200 million in research programs for Big Data computing.

Rick Smolan, creator of the "Day in the Life" photography series, has a new project in the works, called "The Human Face of Big Data." The New York Times has adopted the term in headlines like "The Age of Big Data" and "Big Data on Campus." And a sure sign that Big Data has arrived came just last month, when it became grist for satire in the "Dilbert" comic strip by Scott Adams. "It comes from everywhere. It knows all," one frame reads, and the next concludes

THE WORLD BANK
Working for a World Free of Poverty

English | Español | Français | عربي | Русский | 中文

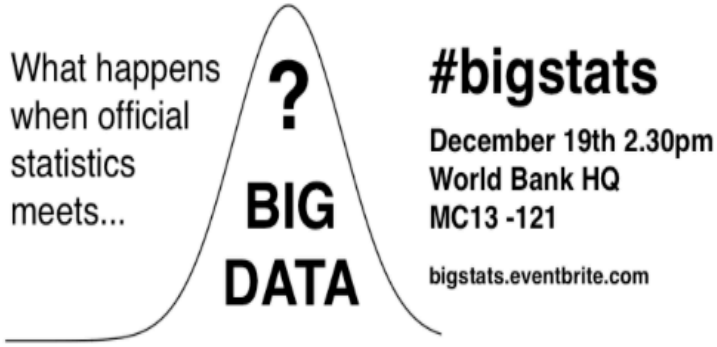
Search GO

ABOUT | DATA | RESEARCH | LEARNING | NEWS | PROJECTS & OPERATIONS | PUBLICATIONS | COUNTRIES | TOPICS

World Bank Live

What Happens When Big Data Meets Official Statistics? - Live Webcast

SHARE

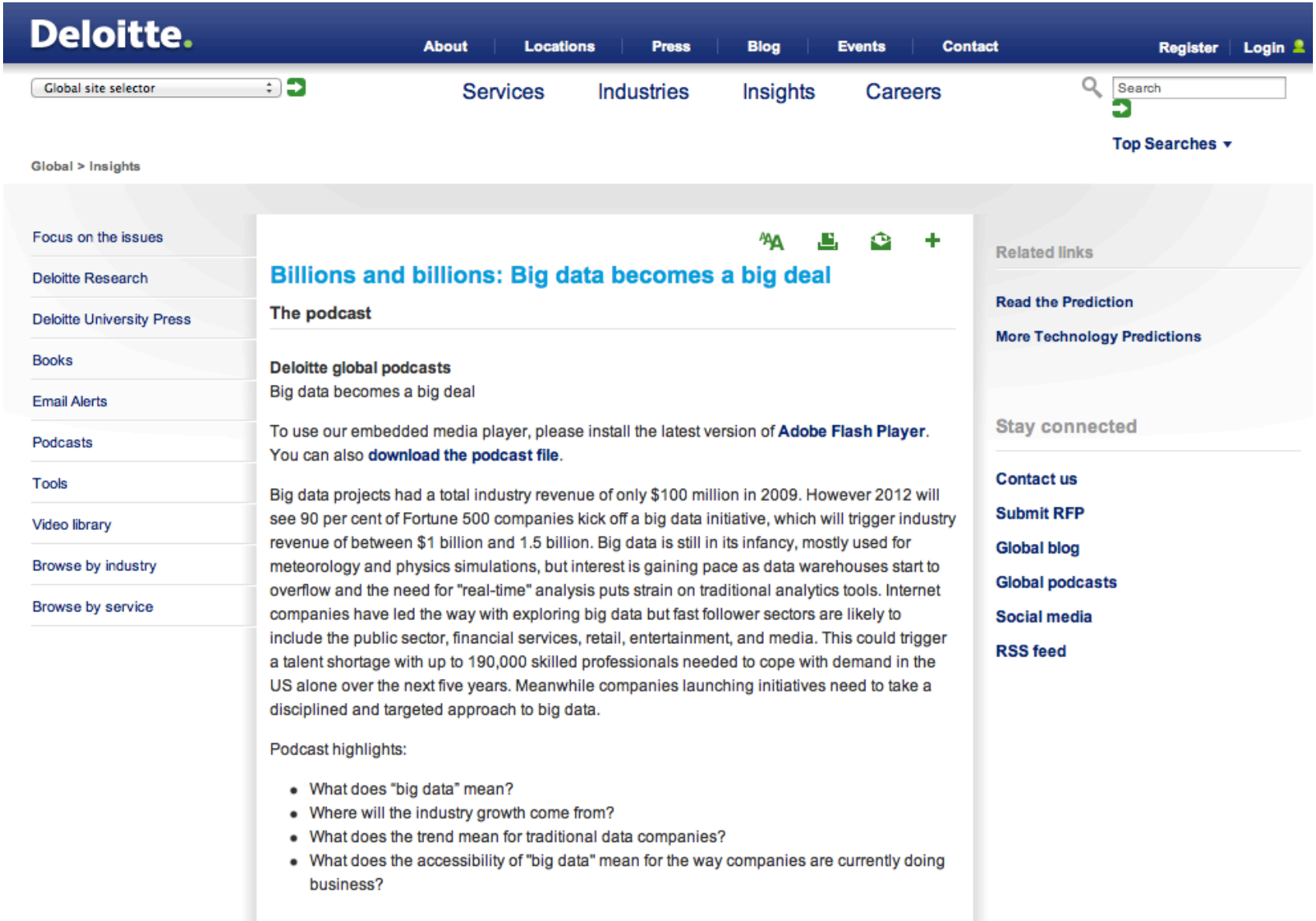


ABOUT

World Bank Live is a space to discuss key development topics in real time. Chat live with experts, watch livestreams and participate in events, ask tough questions.

Subscribe to alerts on upcoming events

E-mail: *


Deloitte.
About
Locations
Press
Blog
Events
Contact
Register
Login
Global site selector
Services
Industries
Insights
Careers
Search
Top Searches
Global > Insights
Focus on the issues
Deloitte Research
Deloitte University Press
Books
Email Alerts
Podcasts
Tools
Video library
Browse by industry
Browse by service
Billions and billions: Big data becomes a big deal
The podcast
Deloitte global podcasts
Big data becomes a big deal
To use our embedded media player, please install the latest version of Adobe Flash Player.
You can also download the podcast file.
Big data projects had a total industry revenue of only $100 million in 2009. However 2012 will see 90 per cent of Fortune 500 companies kick off a big data initiative, which will trigger industry revenue of between $1 billion and 1.5 billion. Big data is still in its infancy, mostly used for meteorology and physics simulations, but interest is gaining pace as data warehouses start to overflow and the need for "real-time" analysis puts strain on traditional analytics tools. Internet companies have led the way with exploring big data but fast follower sectors are likely to include the public sector, financial services, retail, entertainment, and media. This could trigger a talent shortage with up to 190,000 skilled professionals needed to cope with demand in the US alone over the next five years. Meanwhile companies launching initiatives need to take a disciplined and targeted approach to big data.
Podcast highlights:
What does "big data" mean?
Where will the industry growth come from?
What does the trend mean for traditional data companies?
What does the accessibility of "big data" mean for the way companies are currently doing business?
Related links
Read the Prediction
More Technology Predictions
Stay connected
Contact us
Submit RFP
Global blog
Global podcasts
Social media
RSS feed

CriteoLabs : soirée d'inauguration

Criteo inaugure à Paris l'un des premiers centres de R&D en publicité prédictive d'Europe

- Fleur Pellerin, Ministre déléguée chargée des PME, de l'innovation et de l'Economie Numérique, apporte son soutien à cette entreprise innovante du secteur numérique, véritable « success story » à la française.
- Criteo inaugure CriteoLabs, son nouveau centre de R&D de 10.000 m2 au cœur de Paris.
- Avec à terme 300 ingénieurs, ce site est déjà l'un des premiers centres européens de R&D en algorithmes appliqués à la publicité en ligne. Pour accompagner sa forte croissance, Criteo recrute cette année 250 nouveaux collaborateurs.

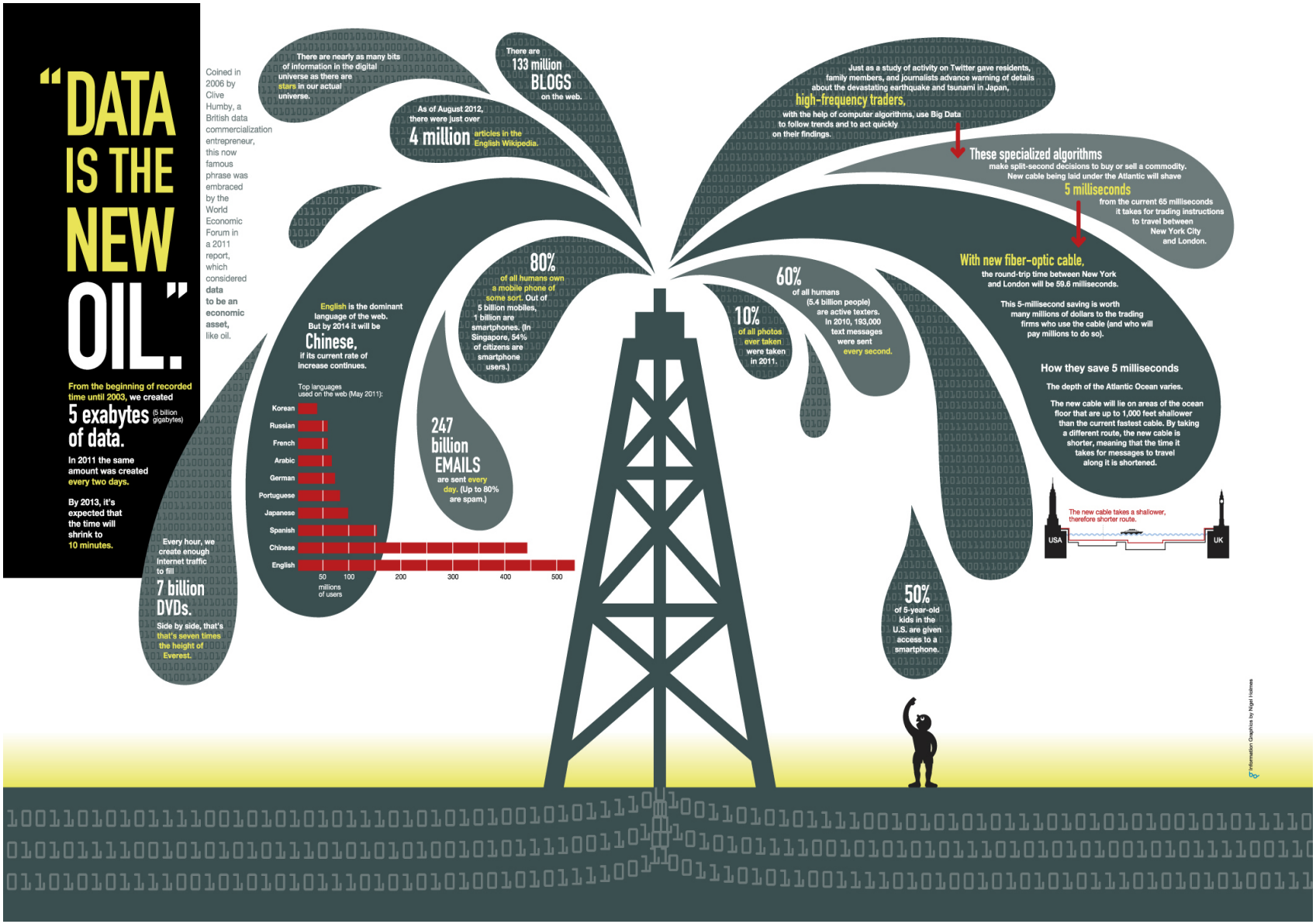
Criteo inaugure à Paris l'un des plus gros pôles européens de R&D dédiés à la publicité prédictive, CriteoLabs. Sur 10.000 m2, ce nouveau centre a vocation à accueillir 300 ingénieurs et à permettre ainsi à Criteo de garantir son avancée technologique sur ses 30 marchés d'exportation, des Etats-Unis, à l'Europe, en passant par l'Asie. Cette année, l'entreprise compte ainsi recruter 250 nouveaux collaborateurs, dont une centaine d'ingénieurs.

Ce nouveau siège, que Criteo a choisi délibérément de situer à Paris, vient ponctuer un développement continu, qui a permis à l'entreprise d'atteindre des résultats remarquables, 3 ans seulement après son lancement commercial :

- 600 salariés présents dans 15 bureaux dans le monde
- 2 000 annonceurs, parmi les plus importants e-commerçants mondiaux tels que Dell, Macy's, John Lewis, Marks & Spencers, Zalando, La Redoute, Les 3 Suisses, etc.
- 4 000 éditeurs
- Plus de 200 millions de dollars de CA en 2011

Jean-Baptiste Rudelle, CEO et Pascal Gauthier COO

Arrivée de Fleur Pellerin


"DATA IS THE NEW OIL."
Coined in 2006 by Clive Humby, a British data commercialization entrepreneur, this now famous phrase was embraced by the World Economic Forum in a 2011 report, which considered data to be an economic asset, like oil.
From the beginning of recorded time until 2003, we created 5 exabytes (5 billion gigabytes) of data.
In 2011 the same amount was created every two days.
By 2013, it's expected that the time will shrink to 10 minutes.
There are nearly as many bits of information in the digital universe as there are stars in our actual universe.
As of August 2012, there were just over 4 million articles in the English Wikipedia.
There are 133 million BLOGS on the web.
Just as a study of activity on Twitter gave residents, family members, and journalists advance warning of details about the devastating earthquake and tsunami in Japan, high-frequency traders, with the help of computer algorithms, use Big Data to follow trends and to act quickly on their findings.
These specialized algorithms make split-second decisions to buy or sell a commodity. New cable being laid under the Atlantic will shave 5 milliseconds from the current 65 milliseconds it takes for trading instructions to travel between New York City and London.
With new fiber-optic cable, the round-trip time between New York and London will be 59.6 milliseconds.
This 5-millisecond saving is worth many millions of dollars to the trading firms who use the cable (and who will pay millions to do so).
How they save 5 milliseconds
The depth of the Atlantic Ocean varies.
The new cable will lie on areas of the ocean floor that are up to 1,000 feet shallower than the current fastest cable. By taking a different route, the new cable is shorter, meaning that the time it takes for messages to travel along it is shortened.
The new cable takes a shallower, therefore shorter route.
USA
UK
80% of all humans own a mobile phone of some sort. Out of 5 billion mobiles, 1 billion are smartphones. (In Singapore, 54% of citizens are smartphone users.)
60% of all humans (5.4 billion people) are active texters. In 2010, 193,000 text messages were sent every second.
10% of all photos ever taken were taken in 2011.
247 billion EMAILS are sent every day. (Up to 80% are spam.)
English is the dominant language of the web. But by 2014 it will be Chinese, if its current rate of increase continues.
Top languages used on the web (May 2011):
Korean
Russian
French
Arabic
German
Portuguese
Japanese
Spanish
Chinese
English
50
100
200
300
400
500
millions of users
Every hour, we create enough Internet traffic to fill 7 billion DVDs. Side by side, that's that's seven times the height of Everest.
50% of 5-year-old kids in the U.S. are given access to a smartphone.
IT Information Graphics by Nigel Holmes

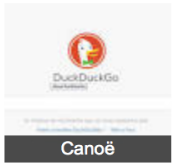
Google moteur de recherche

Web Actualités Images Vidéos Maps Plus Outils de recherche

Environ 10 100 000 résultats (0,24 secondes)

Moteur de recherche - Mozbot France - La recherche facile ...
www.mozbot.fr/
Moteur de recherche Mozbot en partenariat avec Brioude-Internet, Abondance et Google : résultats, synonymes, expressions connexes, statistiques mots clés, ...

Actualités correspondant à **moteur de recherche**



Le **moteur de recherche** DuckDuckGo bloqué en Chine
Le Monde - il y a 3 heures
Selon le site spécialisé TechInAsia, le **moteur de recherche** serait bloqué depuis le 4 septembre dans le pays. DuckDuckGo, qui se présente ...

L'Allemagne souhaite que Google dévoile les algorithmes ...
Clubic.com - il y a 5 jours

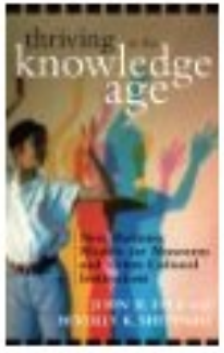
Plus d'actualités pour "**moteur de recherche**"

Moteur de recherche — Wikipédia
fr.wikipedia.org/wiki/**Moteur_de_recherche**
Un **moteur de recherche** est une application web permettant de retrouver des ressources (pages web, articles de forums Usenet, images, vidéo, fichiers, etc.) ...

Moteur de Recherche SEEK.fr ™
www.seek.fr/
Moteur de recherche alternatif français respectant la vie privée via un métamoteur utilisant les principaux **moteurs de recherche** ainsi qu'un annuaire ...
Metamoteur Web SEEK.fr - A Propos de Seek - Horoscope - Seek annuaire

More Ideas Based on Your Browsing History

You looked at



Thriving in the Knowledge Age: New... Paperback by John H. Falk
$29.95

You might also consider

Museum Administration: An Introduction Paperback by Hugh H. Genoways
~~$31.95~~ **$28.75**

Exhibit Labels: An Interpretive Approach Paperback by Beverly Serrell
~~$34.95~~ **$27.85**

› Find similar items

Recommendations don't have to be about showing you more of the same...


Threats Location
171,904 attacks by 27 countries on 33 destinations
Covering 3 months
at a 1 day resolution
with a total of 64.0 million packets
Time period:
3 months
Jun 2, 1998 - Aug 2, 1998
Confidence Level:
High
Low
Insignificant
Beijing
Nishi-ku-machi
Nanjing
Quarry Bay
China
Ranked #8
Total Attacks
2,460
Looped Attacks
77
Unique Attacks
32
Distinct Sources
82

PREDPOL®

ABOUT HOW PREDPOL WORKS PROVEN RESULTS TECHNOLOGY PRESS CONTACT US BLOG

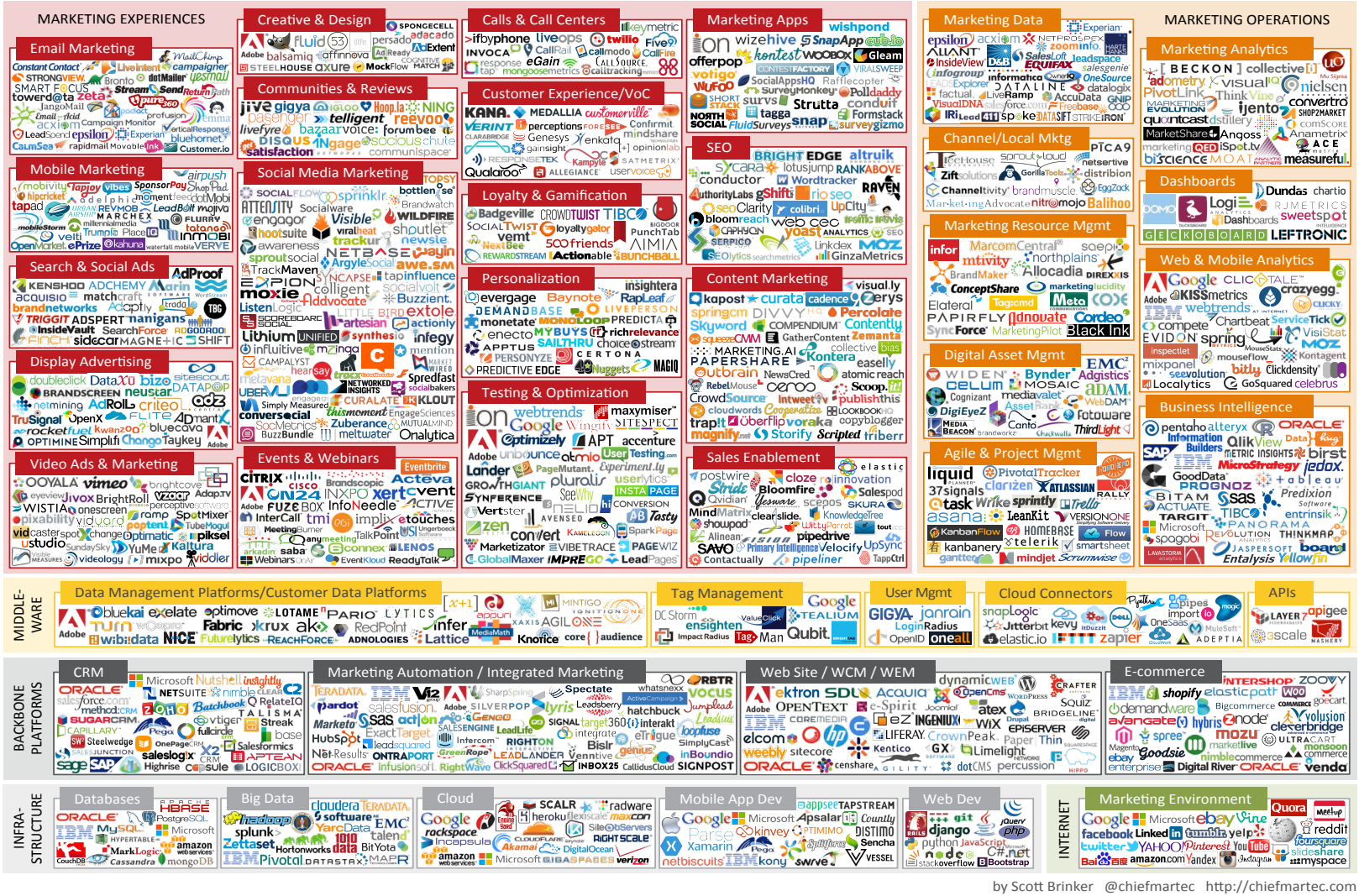
PREDICTIVE POLICING®

The Predictive Policing Company.

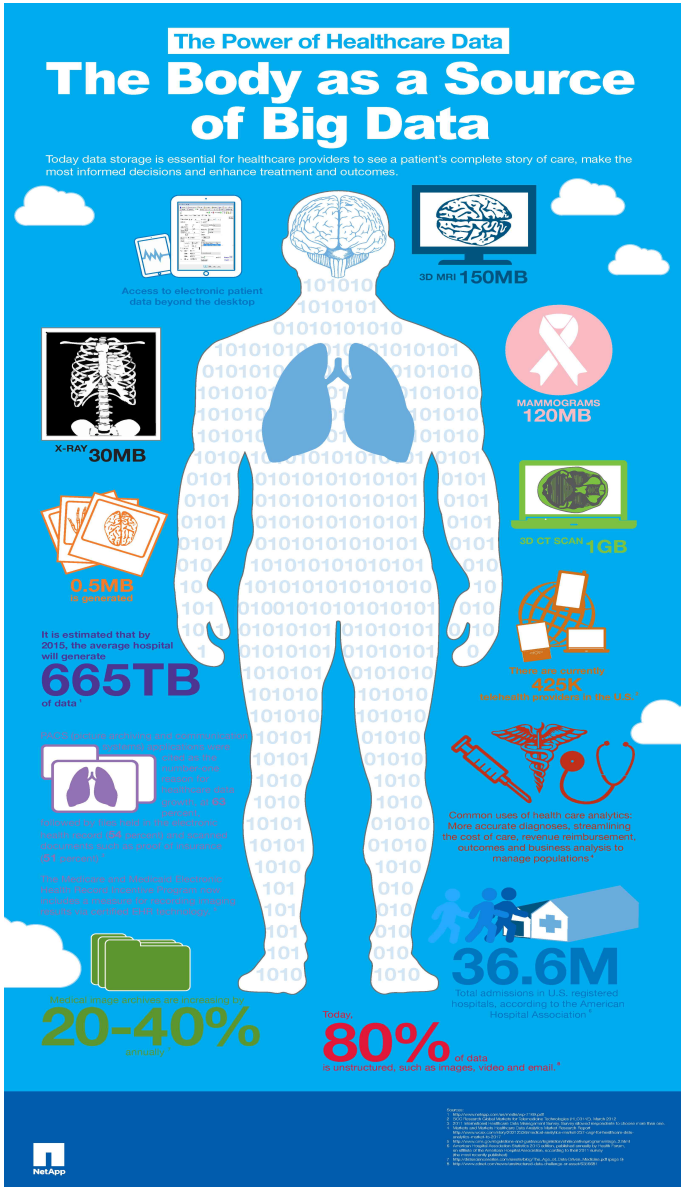
PredPol's cloud-based software enables law enforcement agencies to better prevent crime in their communities by generating predictions on the places and times that future crimes are most likely to occur.

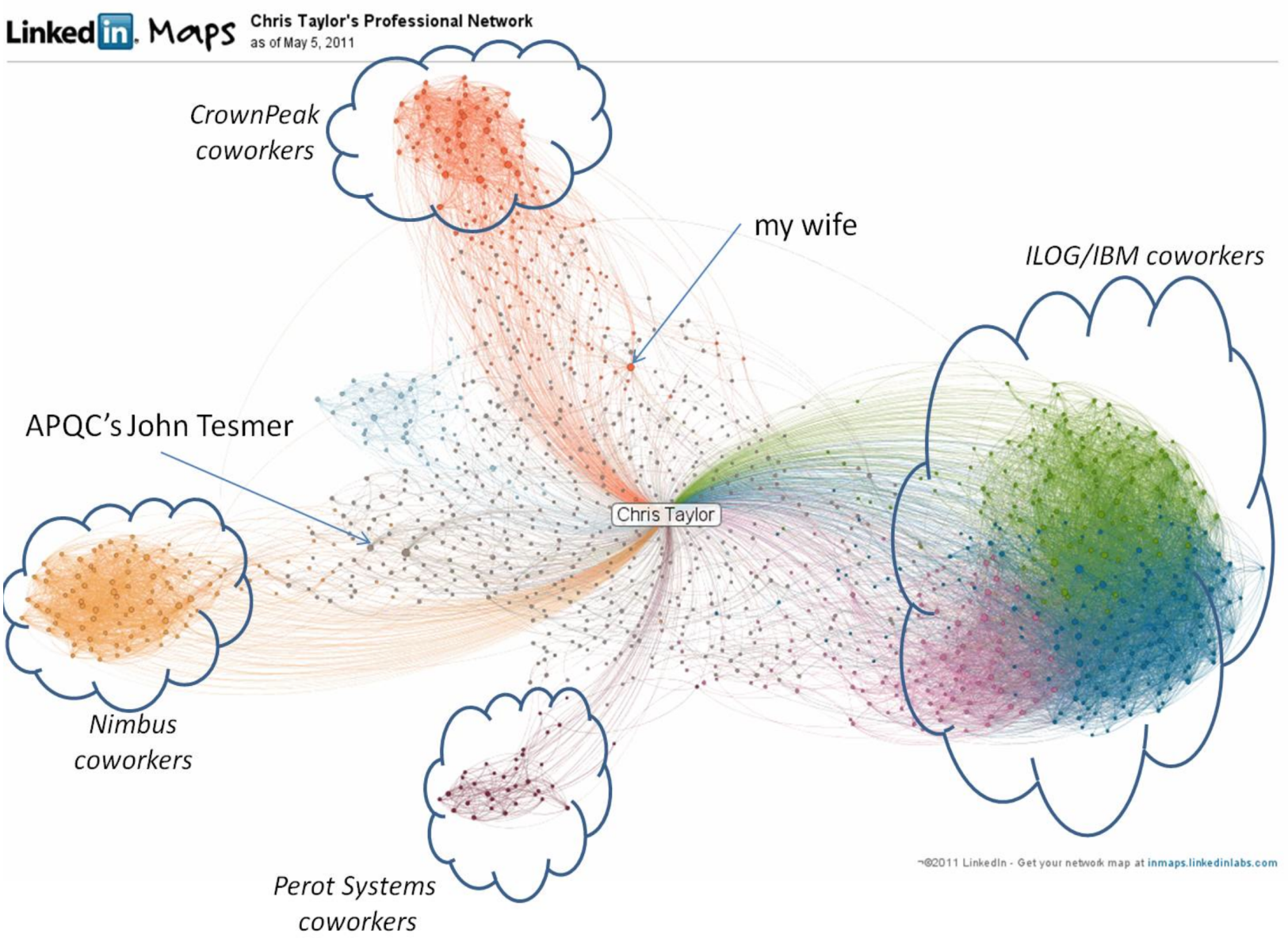
chiefmartec.com Marketing Technology Landscape

January 2014

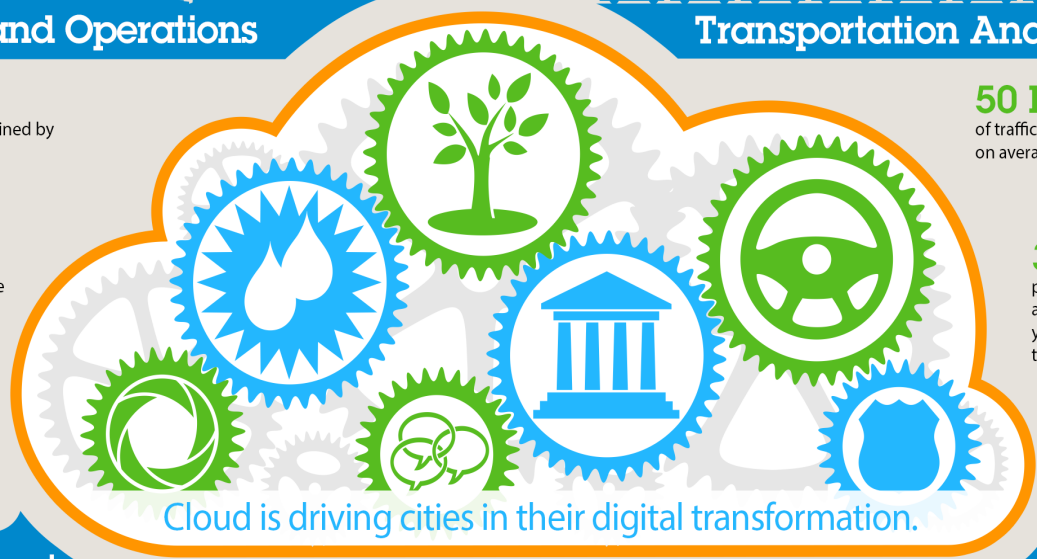


by Scott Brinker @chiefmartec http://chiefmartec.com


The Power of Healthcare Data
The Body as a Source of Big Data
Today data storage is essential for healthcare providers to see a patient's complete story of care, make the most informed decisions and enhance treatment and outcomes.
Access to electronic patient data beyond the desktop
3D MRI 150MB
X-RAY 30MB
MAMMOGRAMS 120MB
3D CT SCAN 1GB
0.5MB is generated
It is estimated that by 2015, the average hospital will generate
665TB
of data
There are currently 425K telehealth providers in the U.S.
Common uses of health care analytics: More accurate diagnoses, streamlining the cost of care, revenue reimbursement, outcomes and business analysis to manage populations
36.6M
Total admissions in U.S. registered hospitals, according to the American Hospital Association
Medical image archives are increasing by
20-40%
annually
Today,
80%
of data,
is unstructured, such as images, video and email.
NetApp


LinkedIn Maps
Chris Taylor's Professional Network
as of May 5, 2011
CrownPeak coworkers
my wife
ILOG/IBM coworkers
APQC's John Tesmer
Chris Taylor
Nimbus coworkers
Perot Systems coworkers
©2011 LinkedIn - Get your network map at inmaps.linkedinlabs.com

Smarter Cities: Turning Big Data Into Insight



City Planning and Operations

$1 Trillion
global annual savings could be attained by optimizing public infrastructure.
Source: McKinsey

$57 Trillion
in infrastructure investments will be needed between 2013-2030.
Source: McKinsey

Transportation Analytics

50 Hours
of traffic delays per year are incurred, on average, by travelers.

30 Billion
people all over the world travel approximately 30 billion miles per year. By 2050, that figure will grow to over 150 billion miles.

Water Management

60%
of water allocated for domestic human use goes to urban cities.

$14 Billion
in potable water is lost every year because of leaks, theft and unbilled usage.
Source: World Bank

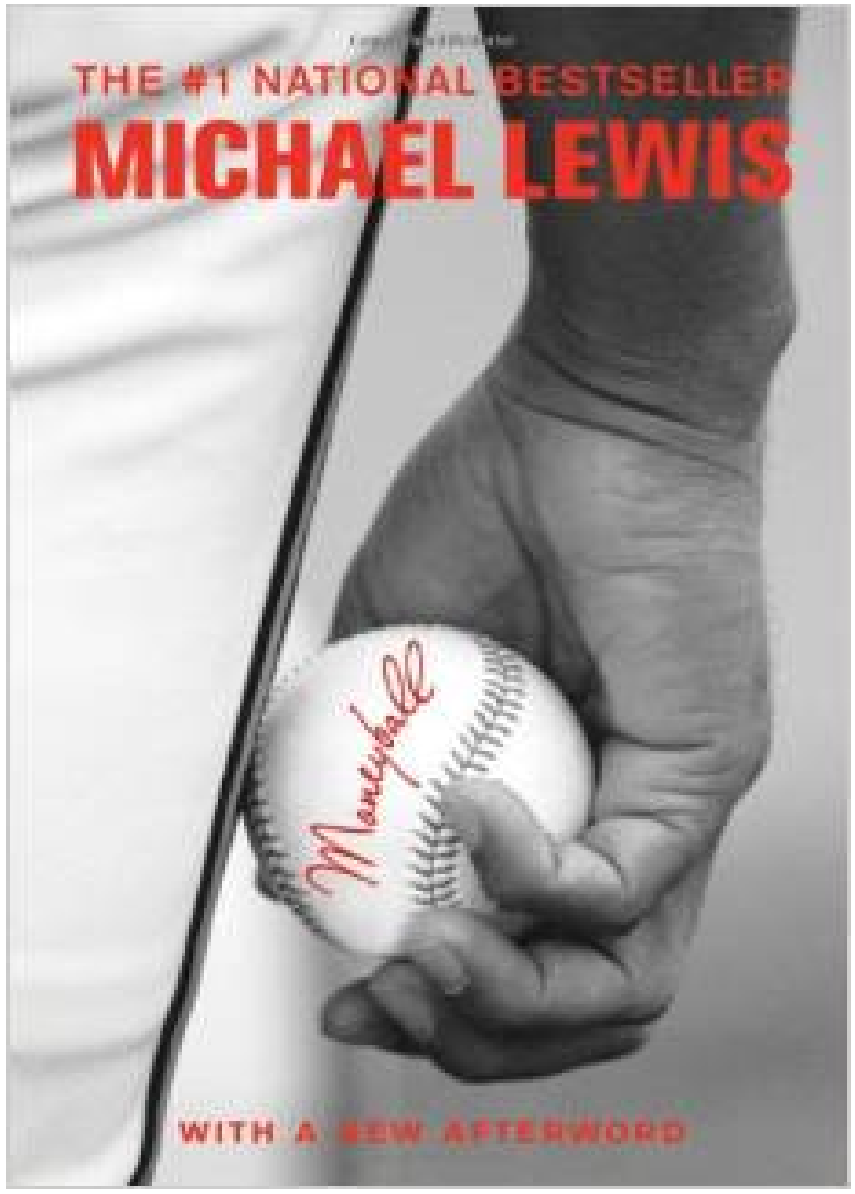
Open Cloud

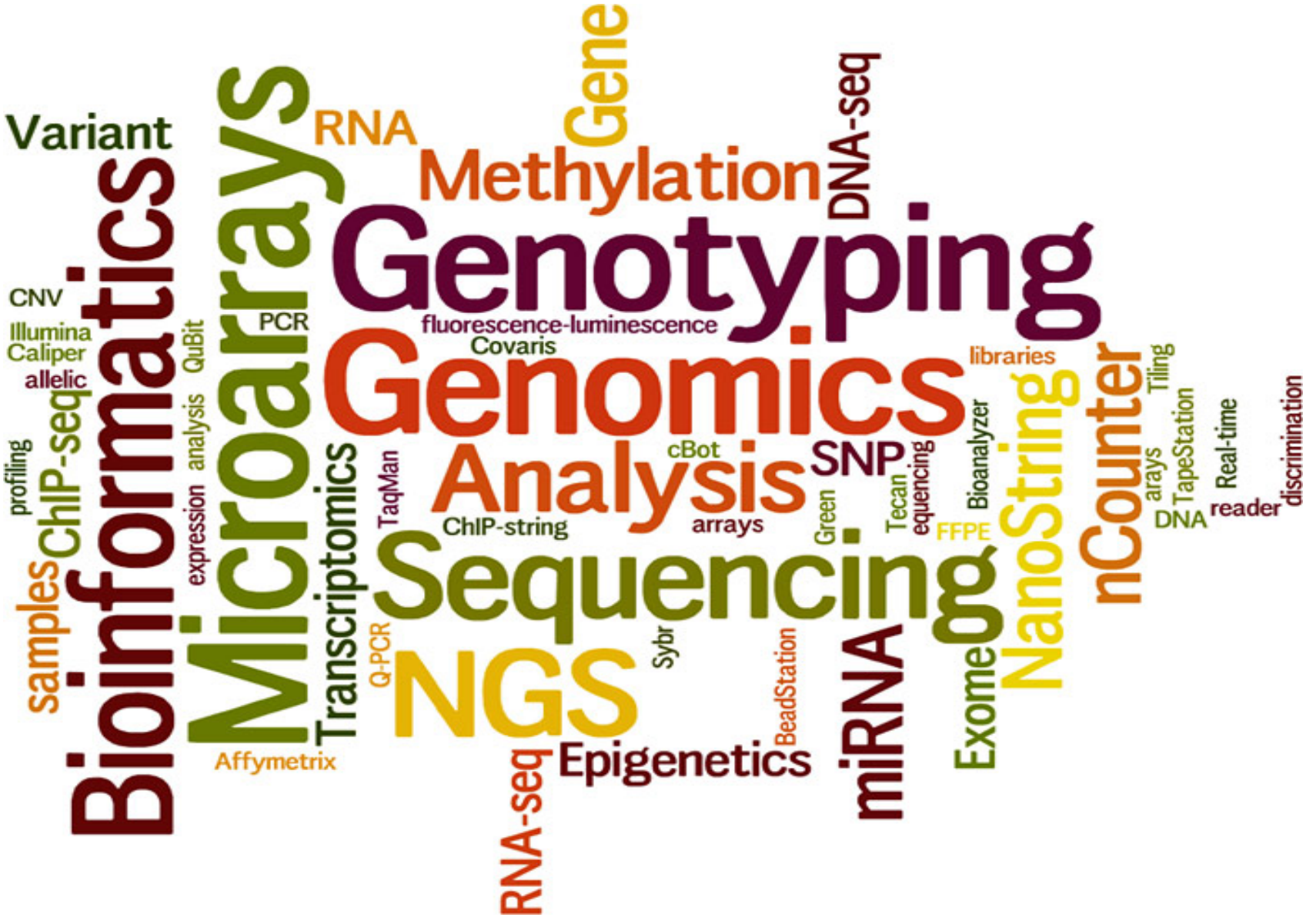
37,000
cloud experts support IBM's industry team alone.

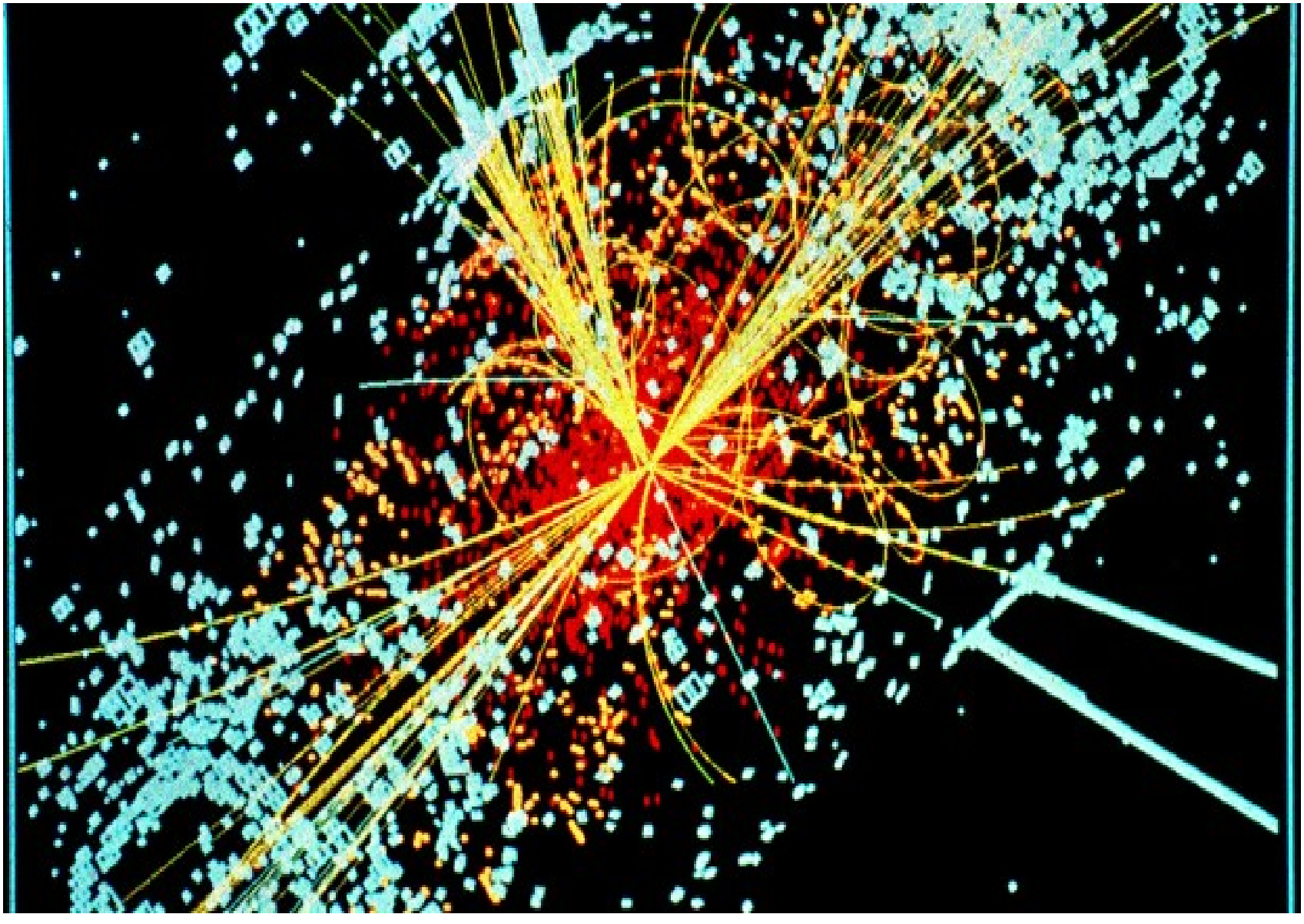
$6 Billion
has been invested by IBM in more than a dozen acquisitions to accelerate its cloud initiatives.

IBM Intelligent Operations software is designed with cities, for cities, to provide the tools to monitor, visualize and analyze vital city services such as water and wastewater systems, transportation, infrastructure planning, permit management and emergency response.

IBM


THE #1 NATIONAL BESTSELLER
MICHAEL LEWIS
WITH A NEW AFTERWORD


Variant
RNA
Gene
DNA-seq
Methylation
Genotyping
Microarrays
Bioinformatics
CNV
PCR
fluorescence-luminescence
Illumina
Caliper
Covaris
QuBit
allelic
Genomics
libraries
Tiling
nCounter
ChIP-seq
expression analysis
profiling
TaqMan
cBot
SNP
Analysis
Transcriptomics
Bioanalyzer
TapeStation
Real-time
arays
DNA
reader
discrimination
ChIP-string
arrays
Green
Tecan
equencing
FFPE
NanoString
Sequencing
samples
Q-PCR
Sybr
NGS
BeadStation
miRNA
Exome
Affymetrix
RNA-seq
Epigenetics



Big data

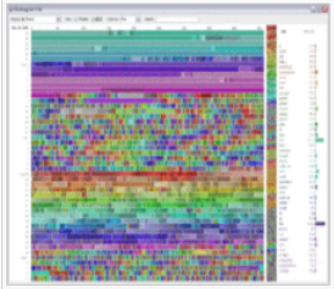
From Wikipedia, the free encyclopedia

This article is about large collections of data. For the band, see Big Data (band).

Big data[1][2] is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curation, storage,[3] search, sharing, transfer, analysis[4] and visualization. The trend to larger data sets is due to the additional information derivable from analysis of a single large set of related data, as compared to separate smaller sets with the same total amount of data, allowing correlations to be found to "spot business trends, determine quality of research, prevent diseases, link legal citations, combat crime, and determine real-time roadway traffic conditions."[5][6][7]

As of 2012, limits on the size of data sets that are feasible to process in a reasonable amount of time were on the order of exabytes of data.[8] Scientists regularly encounter limitations due to large data sets in many areas, including meteorology, genomics,[9] connectomics, complex physics simulations,[10] and biological and environmental research.[11] The limitations also affect Internet search, finance and business informatics. Data sets grow in size in part because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, radio-frequency identification readers, and wireless sensor networks.[12][13] The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s;[14] as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data were created.[15] The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.[16]

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers".[17] What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."[18]



A visualization created by IBM of Wikipedia edits. At multiple terabytes in size, the text and images of Wikipedia are a classic example of big data.

- **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
- **Data science** is the study of the generalizable extraction of knowledge from data, yet the key word is science.
- **Statistics** is the study of the collection, analysis, interpretation, presentation and organization of data.

Big data

- Capacity to store information has doubled every 40 months since the 1980s
- In 2012, 2.5 exabytes (2.5×10^{18}) created per **day**
- Big internet companies such as Google, Amazon, Facebook, but also industries from pharmaceuticals, insurance, banks, telecoms, personalized medicine, marketing, bioinformatics

Data everywhere

- Huge volume,
- Huge variety...

Affordable computation units

- Cloud computing
- Graphical Processor Units (GPU)...

Big Data is (quite) easy?

Example of "off the shelf" solution

Lightning-fast cluster computing


```scala
def run(params: Params) {
  val conf = new SparkConf()
    .setAppName(s"BinaryClassification with $params")
  val sc = new SparkContext(conf)

  Logger.getRootLogger.setLevel(Level.WARN)

  val examples = MLUtils.loadLibSVMFile(sc, params.input).cache()

  val splits = examples.randomSplit(Array(0.8, 0.2))
  val training = splits(0).cache()
  val test = splits(1).cache()
  val numTraining = training.count()
  val numTest = test.count()
  println(s"Training: $numTraining, test: $numTest.")
  examples.unpersist(blocking = false)

  val updater = params.regType match {
    case L1 => new L1Updater()
    case L2 => new SquaredL2Updater()
  }

  val algorithm = new LogisticRegressionWithSGD()
    algorithm.optimizer
      .setNumIterations(params.numIterations)
      .setStepSize(params.stepSize)
      .setUpdater(updater)
      .setRegParam(params.regParam)
  val model = algorithm.run(training).clearThreshold()

  val prediction = model.predict(test.map(_.features))
  val predictionAndLabel = prediction.zip(test.map(_.label))

  val metrics = new BinaryClassificationMetrics(predictionAndLabel)
  val myMetrics = new MyBinaryClassificationMetrics(predictionAndLabel)

  println(s"Empirical CrossEntropy = ${myMetrics.crossEntropy()}.")
  println(s"Test areaUnderPR = ${metrics.areaUnderPR()}.")
  println(s"Test areaUnderROC = ${metrics.areaUnderROC()}.")

  sc.stop()
}
```

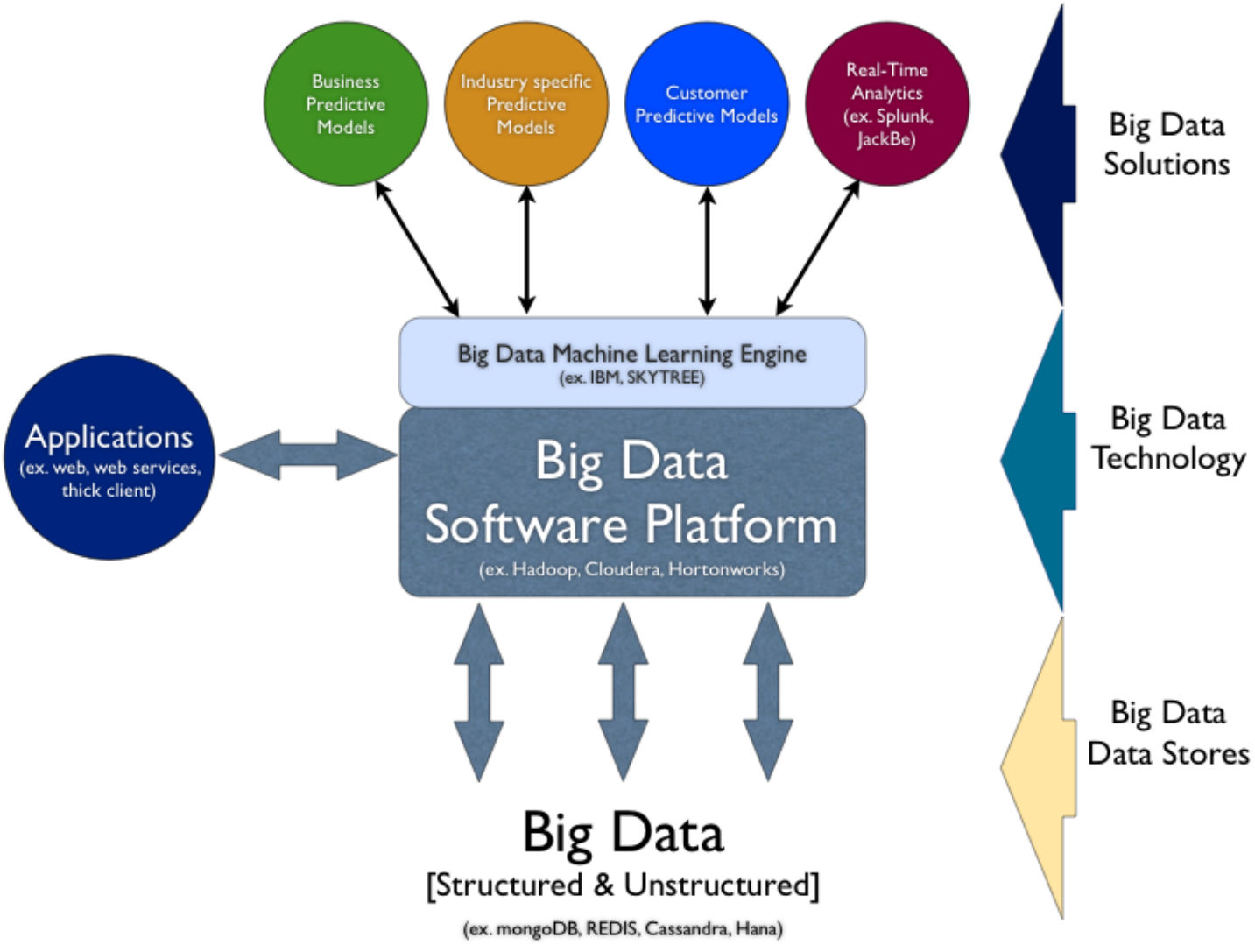
Big Data is (quite) easy?

Example of "off the shelf" solution

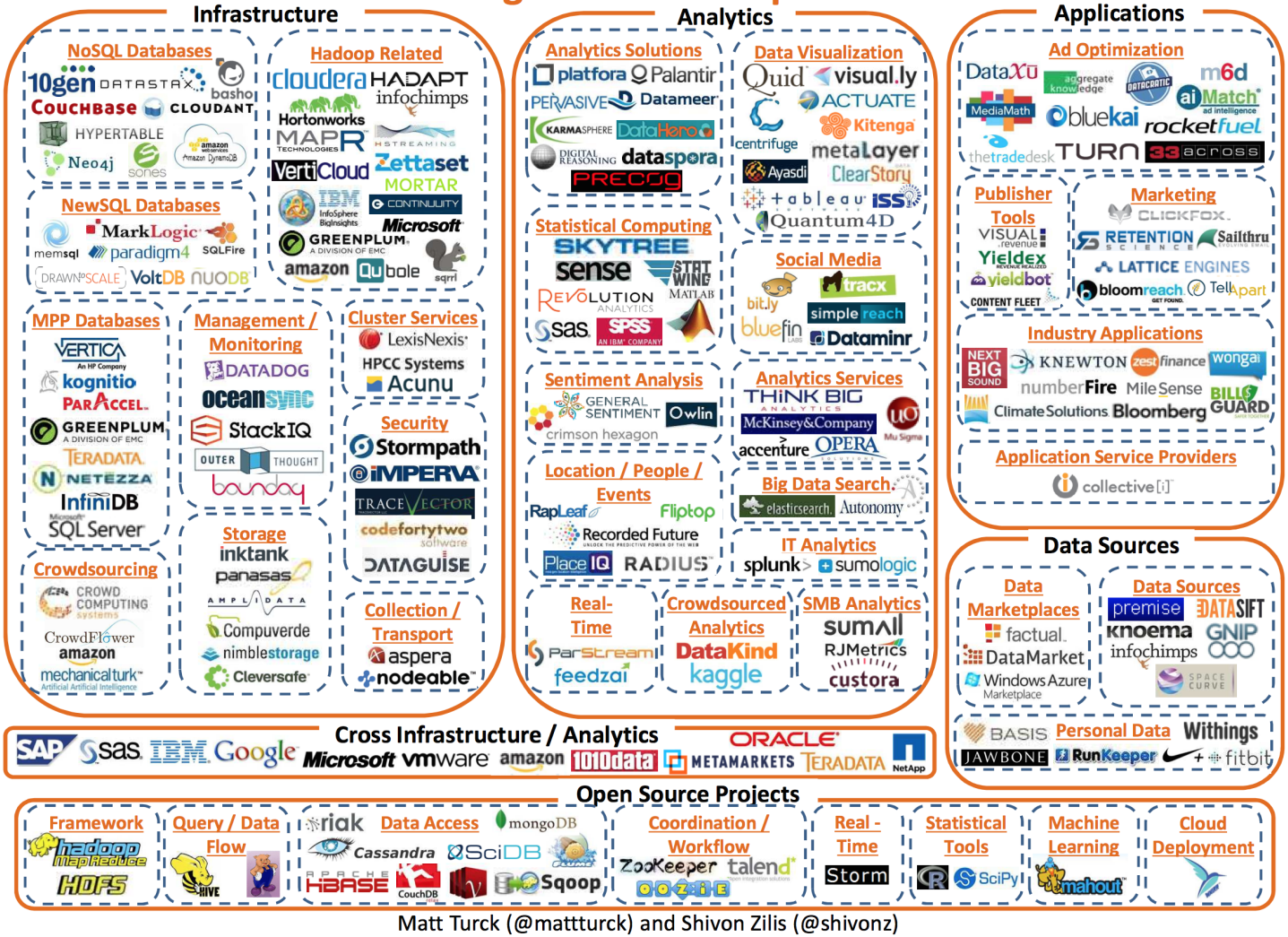

```
export AWS_ACCESS_KEY_ID=<your-access-keyid>
export AWS_SECRET_ACCESS_KEY=<your-access-key-secret>
cellule/spark/ec2/sparl-ec2 -i cellule.pem -k cellule -s <number of machines> launch <cluster-name>
ssh -i cellule.pem root@<your-cluster-master-dns>
spark-ec2/copy-dir ephemeral-hdfs/conf
ephemeral-hdfs/bin/hadoop distcp s3n://celluledecalcul/dataset/raw/train.csv /data/train.csv
scp -i cellule.pem cellule/challenge/target/scala-2.10/target/scala-2.10/challenges_2.10-0.0.jar

cellule/spark/bin/spark-submit \
        --class fr.cc.challenge.Preprocess \
        challenges_2.10-0.0.jar \
        /data/train.csv \
        /data/train2.csv

cellule/spark/bin/spark-submit \
       --class fr.cc.sparktest.LogisticRegression \
       challenges_2.10-0.0.jar \
       /data/train2.csv
```

$\Rightarrow$ Logistic regression for arbitrary large dataset!

A Complex Ecosystem!



Big Data Landscape



Matt Turck (@mattturck) and Shivon Zilis (@shivonz)

Teasers: data science or statistics?

Jeremy Jarvis
@jeremyjarvis

Follow

"A data scientist is a statistician who lives in San Fransisco" #monkigras pic.twitter.com/HypLL3Cnye

12:13 PM - 30 Jan 2014

1,475 841

Big Data Borat
@BigDataBorat

Follow

Data Science is statistics on a Mac.

3:32 PM - 27 Aug 2013

611 273

Josh Wills
@josh_wills

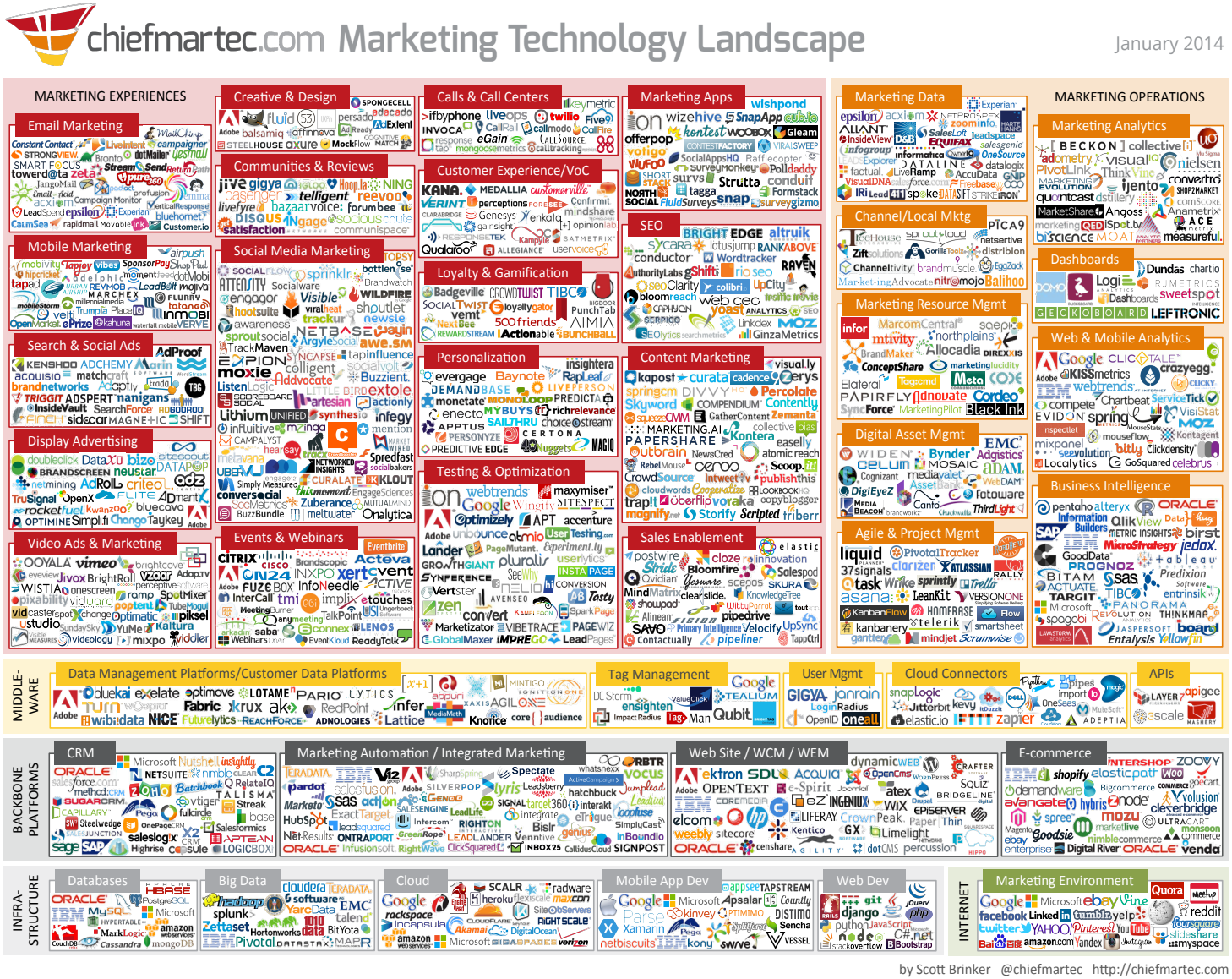
Follow

Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician.

6:55 PM - 3 May 2012

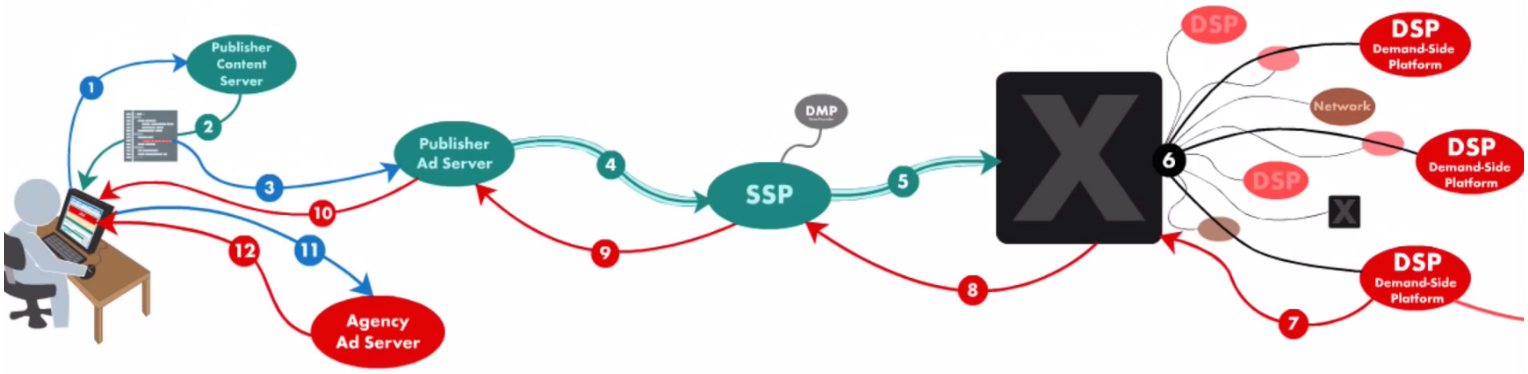
1,360 821

Teasers: an application in marketing



by Scott Brinker @chiefmartec http://chiefmartec.com

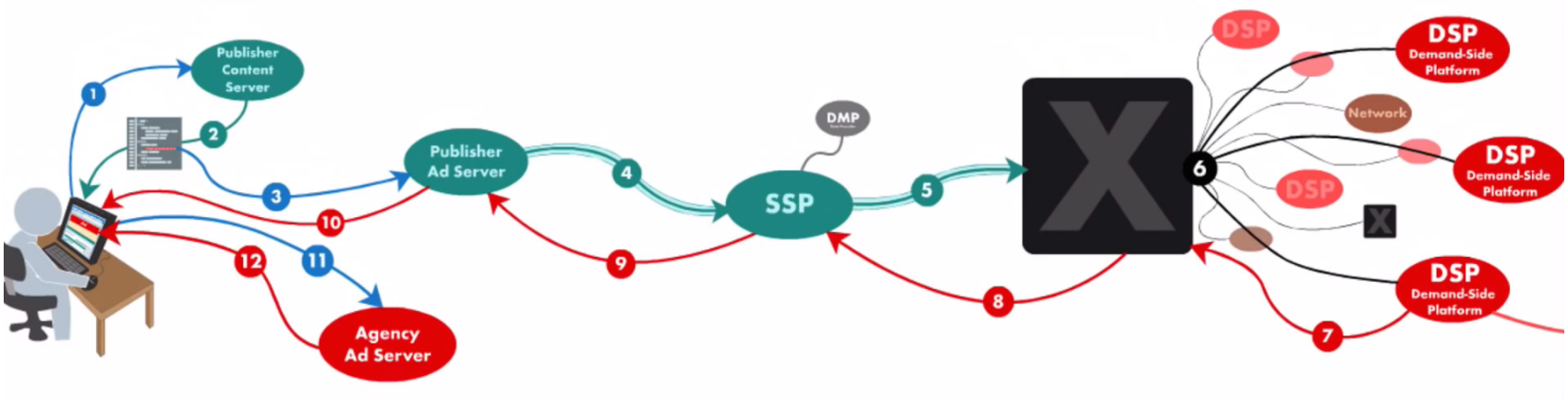
Teasers: an application in marketing (Real Time Bidding)



- A **customer** visits a webpage with his browser: a complex process of content selection and delivery begins.
- An **advertiser** might want to display an ad on the webpage where the user is going. The webpage belongs to a **publisher**.
- The publisher sells ad space to advertisers who want to reach customers

In some cases, an auction starts: **RTB** (Real Time Bidding)

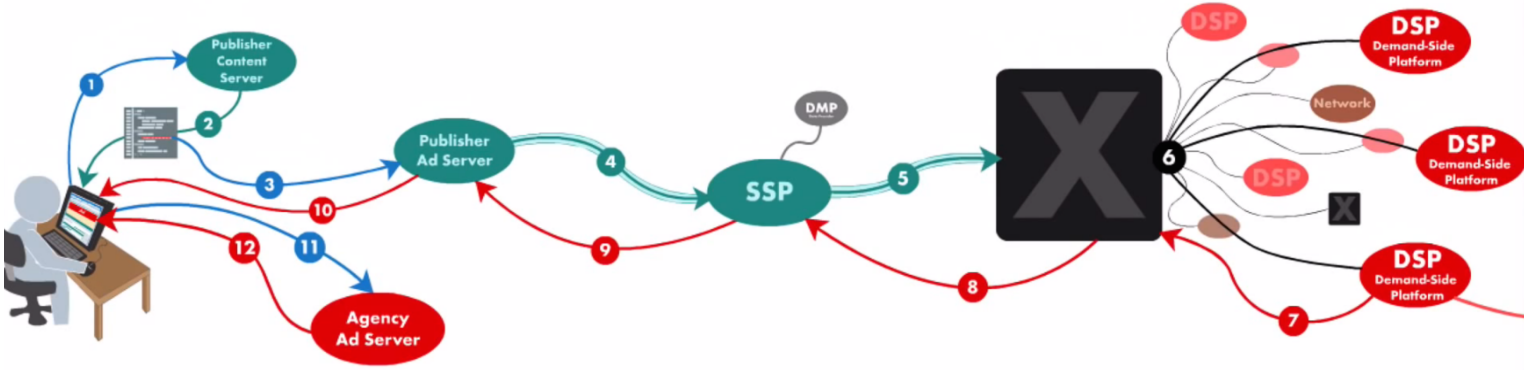
Teasers: an application in marketing (Real Time Bidding)



- Advertisers have **10ms** (!) to give a price: they need to assess quickly how willing they are to display the ad to this customer
- Machine learning is used here to **predict the probability of click on the ad**. Time constraint: few model parameters to answer quickly
- Feature selection / dimension reduction is crucial here

Full process takes < 100ms

Teasers: an application in marketing (Real Time Bidding)



Some figures:

- 10 million prediction of click probability per second
- answers within 10ms
- stores 20Terabytes of data daily

Aim

- Based on past data, you want to find users that will click on some ads

This problem can be formulated as a **binary classification problem**

Classification = **supervised learning** with a binary label

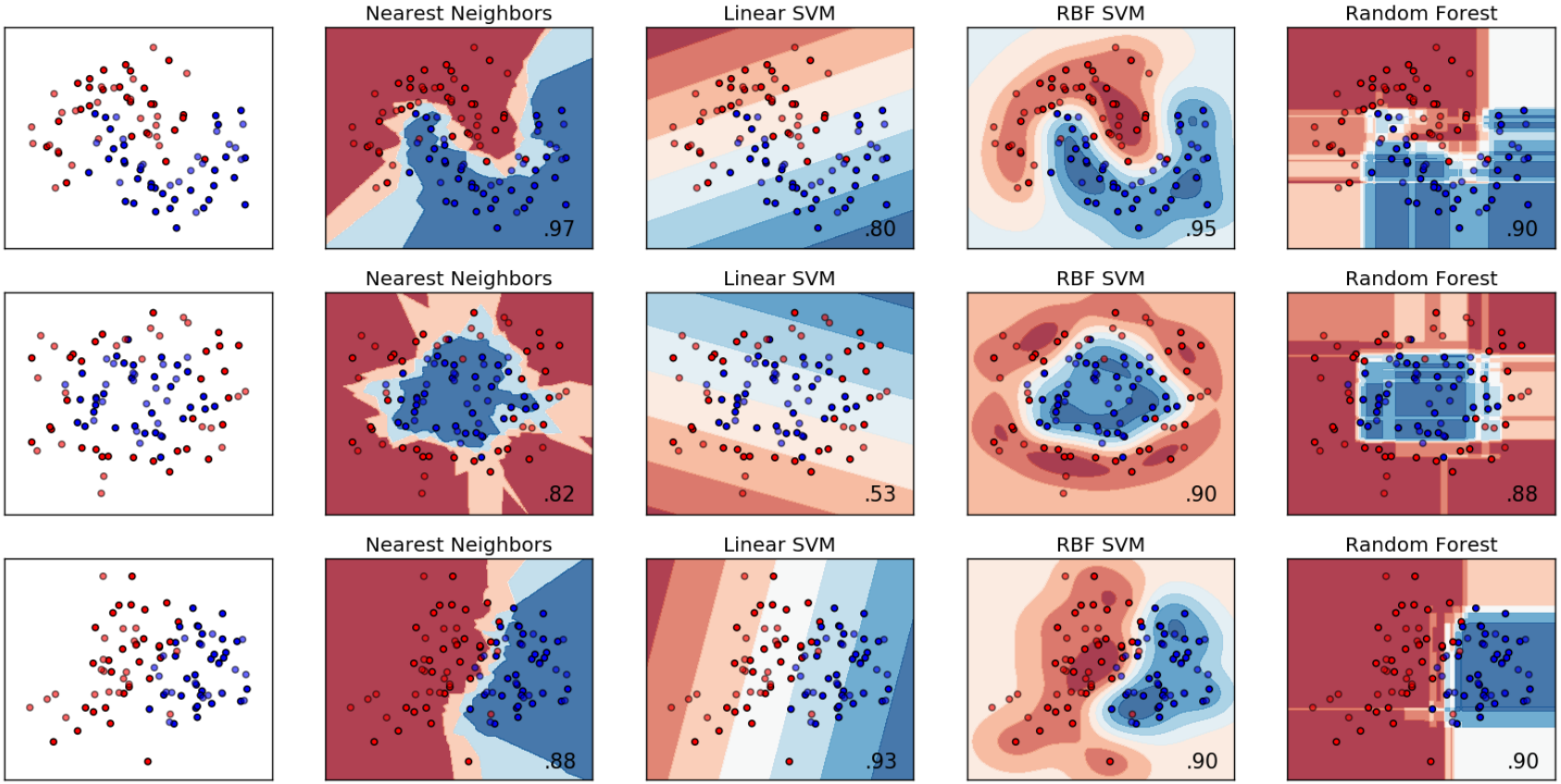
Setting

- You have past/historical data, containing data about individuals $i = 1, \ldots, n$
- You have a **features** vector $x_i \in \mathbb{R}^d$ for each individual i
- For each i, you know if he/she clicked ($y_i = 1$) or not ($y_i = -1$)
- We call $y_i \in \{-1, 1\}$ the **label** of i
- (x_i, y_i) are i.i.d realizations of (X, Y)

Aim

- Given a features vector x (with no corresponding label), predict a label $\hat{y} \in \{-1, 1\}$
- Use data $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ to construct a **classifier**

Many ways to separate points!



Today: model-based classification

- Naive Bayes
- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)
- Logistic regression
- Penalization
- Cross-validation

Probabilistic / statistical approach

- Model the distribution of $Y|X$
- Construct estimators $\hat{p}_1(x)$ and $\hat{p}_{-1}(x)$ of

$$p_1(x) = \mathbb{P}(Y = 1|X = x) \quad \text{and} \quad p_{-1}(x) = 1 - p_1(x)$$

- Given x, classify using

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{p}_1(x) \geq t \\ -1 & \text{otherwise} \end{cases}$$

for some threshold $t \in (0, 1)$

Bayes formula. We know that

$$p_y(x) = \mathbb{P}(Y = y|X = x) = \frac{\mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y)}{\mathbb{P}(X = x)}$$
$$= \frac{\mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y)}{\sum_{y'=-1,1} \mathbb{P}(X = x|Y = y')\mathbb{P}(Y = y')}$$

If we know the distribution of $X|Y$ and the distribution of Y, we know the distribution of $Y|X$

Bayes classifier. Classify using Bayes formula, given that:

- We model $\mathbb{P}(X|Y)$
- We are able to estimate $\mathbb{P}(X|Y)$ based on data

Maximum a posteriori. Classify using the *discriminant* functions

$$\delta_y(x) = \log \mathbb{P}(X = x|Y = y) + \log \mathbb{P}(Y = y)$$

for $y = 1, -1$ and decide (largest, beyond a threshold, etc.)

Remark.

- Different models on the distribution of $X|Y$ leads to different classifiers
- The simplest one is the Naive Bayes
- Then, the most standard are Linear Discriminant Analysis (LDA) and Quadratic discriminant Analysis (QDA)

Naive Bayes. A crude modeling for $\mathbb{P}(X|Y)$: assume features X^j are independent conditionally on Y:

$$\mathbb{P}(X = x|Y = y) = \prod_{j=1}^{d} \mathbb{P}(X^j = x^j|Y = y)$$

Model the univariate distribution $X^j|Y$: for instance, assume that

$$\mathbb{P}(X^j|Y = y) = \text{Normal}(\mu_{j,y}, \sigma^2_{j,y}),$$

parameters $\mu_{j,y}$ and $\sigma^2_{j,y}$ easily estimated by MLE

- If the feature X^j is discrete, use a Bernoulli or multinomial distribution
- Leads to a classifier which is very easy to compute
- Requires only the computation of some averages (MLE)

Discriminant Analysis. Assume that

$$\mathbb{P}(X|Y=y) = \text{Normal}(\mu_y, \Sigma_y),$$

where we recall that the density of $\text{Normal}(\mu, \Sigma)$ is given by

$$f(x) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} \exp\Big(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\Big)$$

In this case, discriminant functions are

$$\begin{aligned}\delta_y(x) &= \log \mathbb{P}(X=x|Y=y) + \log \mathbb{P}(Y=y) \\ &= -\frac{1}{2}(x-\mu_y)^\top \Sigma_y^{-1}(x-\mu_y) - \frac{d}{2}\ln(2\pi) \\ &\quad - \frac{1}{2}\log\det\Sigma_y + \log \mathbb{P}(Y=y)\end{aligned}$$

Estimation. Use "natural" estimators, obtained by maximum likelihood estimation. Define for $y \in \{-1, 1\}$

$$I_y = \{i = 1, \ldots, n \ : \ y_i = y\} \quad \text{and} \quad n_y = |I_y|$$

MLE estimators are given by

$$\hat{\mathbb{P}}(Y = y) = \frac{n_y}{n}, \quad \hat{\mu}_y = \frac{1}{n_y} \sum_{i \in I_y} x_i,$$

$$\hat{\Sigma}_y = \frac{1}{n_y} \sum_{i \in I_y} (x_i - \hat{\mu}_y)(x_i - \hat{\mu}_y)^\top$$

for $y \in \{-1, 1\}$. These are simply the proportion, sample mean and sample covariance within each group of labels

Linear Discriminant Analysis (LDA)

- Assumes that $\Sigma = \Sigma_1 = \Sigma_{-1}$
- All groups have the same correlation structure
- In this case decision function is linear $\langle x, w \rangle \geq c$ with

$$w = \Sigma^{-1}(\mu_1 - \mu_{-1})$$

$$c = \frac{1}{2}(\langle \mu_1, \Sigma^{-1}\mu_1 \rangle - \langle \mu_{-1}, \Sigma^{-1}\mu_{-1} \rangle) + \log\left(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = -1|X = x)}\right)$$

Quadratic Discriminant Analysis (QDA)

- Assumes that $\Sigma_1 \neq \Sigma_{-1}$
- Decision function is quadratic

Logistic regression

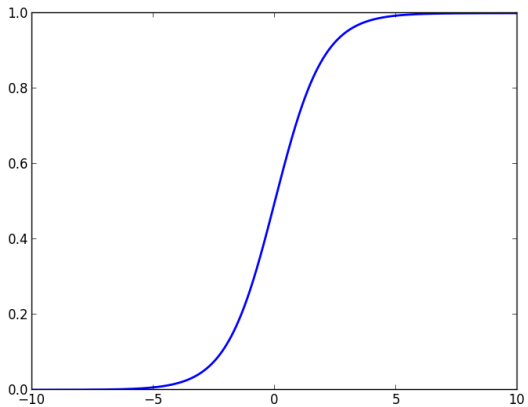
- By far the most widely used classification algorithm
- We want to explain the label y based on x, we want to "regress" y on x
- Models the distribution of $Y|X$

For $y \in \{-1, 1\}$, we consider the model

$$\mathbb{P}(Y = 1|X = x) = \sigma(x^\top w + b)$$

where $w \in \mathbb{R}^d$ is a vector of model **weights** and $b \in \mathbb{R}$ is the **intercept**, and where σ is the **sigmoid** function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



- The sigmoid choice really is a **choice**. It is a **modelling choice**.
- It's a way to map $\mathbb{R} \to [0, 1]$ (we want to model a probability)
- We could also consider

$$\mathbb{P}(Y = 1|X = x) = F(\langle x, w \rangle + b)$$

for any distribution function F. Another popular choice is the Gaussian distribution

$$F(z) = \mathbb{P}(N(0, 1) \leq z),$$

which leads to another loss called **probit**

- However, the sigmoid choice has the following nice interpretation: an easy computation leads to

$$\log\left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=-1|X=x)}\right) = \langle x, w\rangle + b$$

  This quantity is called the **log-odd ratio**
- Note that

$$\mathbb{P}(Y=1|X=x) \geq \mathbb{P}(Y=-1|X=x)$$

  iff

$$\langle x, w\rangle + b \geq 0.$$

- This is a **linear classification** rule
- Linear with respect to the considered features x
- But, **you** choose the features: **features engineering** (more on that later)

Estimation of w and b

- We have a model for $Y|X$
- Data (x_i, y_i) is assumed i.i.d with the same distribution as (X, Y)
- Compute estimators $\hat{w}$ and $\hat{b}$ by **maximum likelihood estimation**
- Or equivalently, minimize the minus log-likelihood
- More generally, when a model is used

$$\text{Goodness-of-fit} = -\log \text{ likelihood}$$

- log is used mainly since averages are easier to study (and compute) than products

Likelihood is given by

$$\prod_{i=1}^{n} \mathbb{P}(Y = y_i | X = x_i)$$

$$= \prod_{i=1}^{n} \sigma(\langle x_i, w \rangle + b)^{\frac{1+y_i}{2}} \left(1 - \sigma(\langle x_i, w \rangle + b)\right)^{\frac{1-y_i}{2}}$$

$$= \prod_{i=1}^{n} \sigma(\langle x_i, w \rangle + b)^{\frac{1+y_i}{2}} \sigma(-\langle x_i, w \rangle - b)^{\frac{1-y_i}{2}}$$

and the minus log-likelihood is given by

$$\sum_{i=1}^{n} \log(1 + e^{-y_i(\langle x_i, w \rangle + b)})$$

Compute $\hat{w}$ and $\hat{b}$ as follows:

$$(\hat{w}, \hat{b}) \in \operatorname*{argmin}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \log(1 + e^{-y_i(\langle x_i, w \rangle + b)})$$

- It is a convex and smooth problem
- Many ways to find an approximate minimizer
- Convex optimization algorithms (more on that later)

If we introduce the **logistic loss** function

$$\ell(y, y') = \log(1 + e^{-yy'})$$

then

$$(\hat{w}, \hat{b}) \in \operatorname*{argmin}_{w \in \mathbb{R}^d, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, w \rangle + b)$$

A goodness-of-fit

$$(\hat{w}, \hat{b}) \in \underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, w \rangle + b)$$

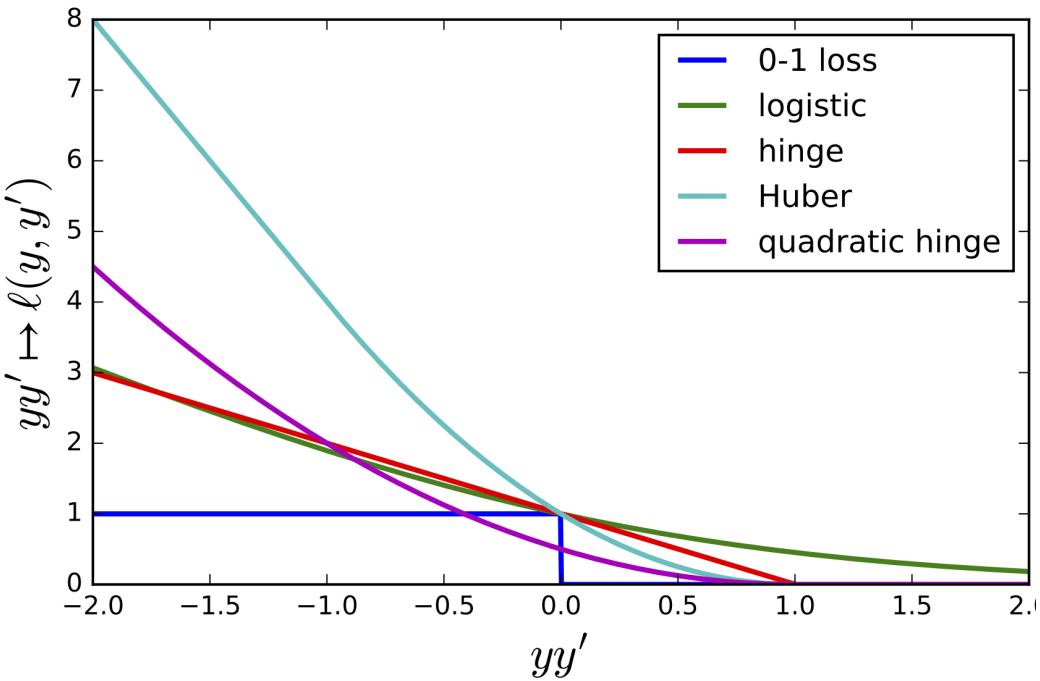
is natural: it is an **average of losses**, one for each sample point

Note that

- $\ell(y, y') = \log(1 + e^{-yy'})$ for logistic regression
- $\ell(y, y') = \frac{1}{2}(y - y')^2$ for least-squares linear regression

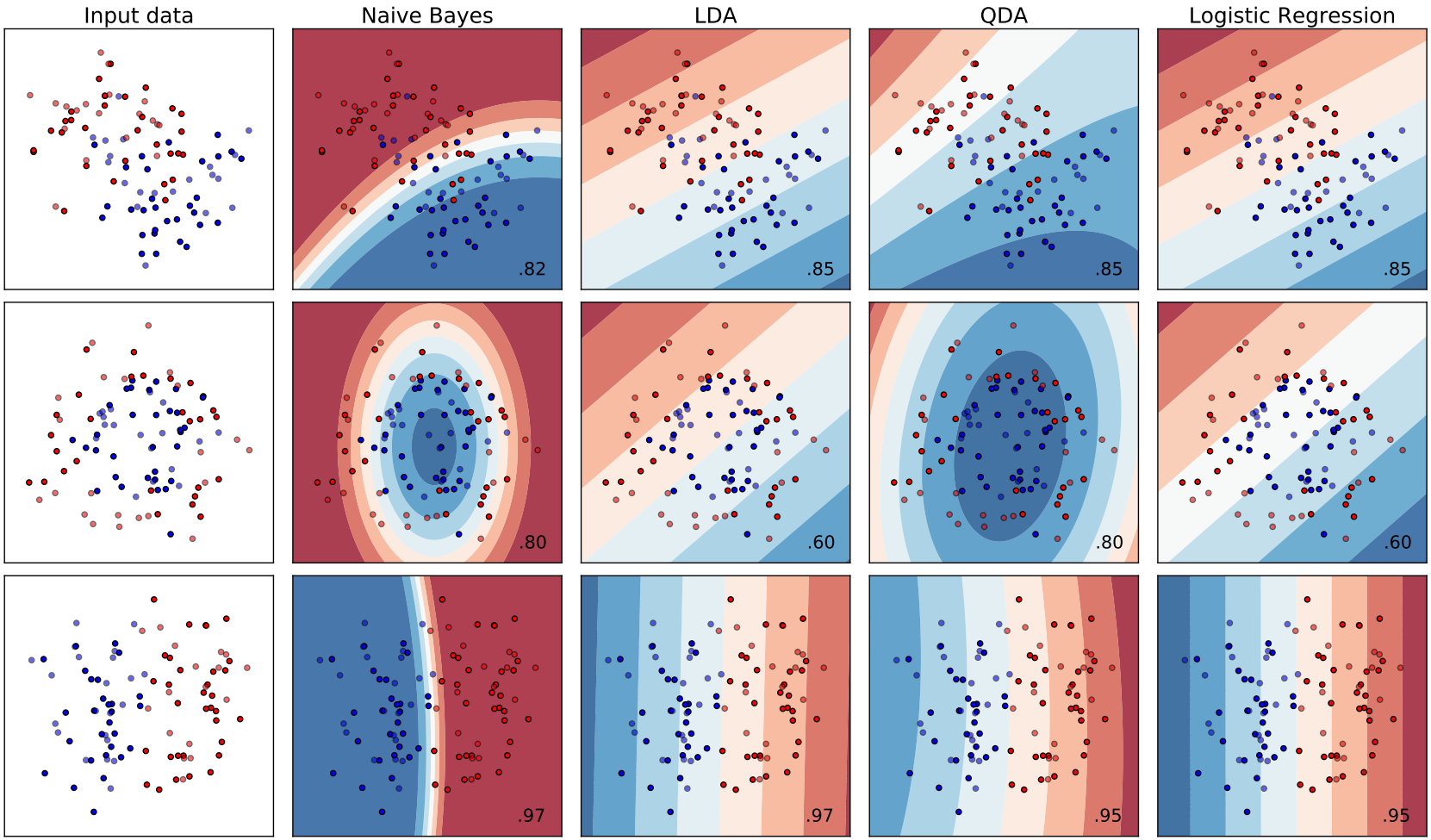
Classical loss functions for binary classication

- Hinge loss (SVM), $\ell(y, y') = (1 - yy')_+$
- Quadratic hinge loss (SVM), $\ell(y, y') = \frac{1}{2}(1 - yy')_+^2$
- Huber loss $\ell(y, y') = -4yy'\mathbf{1}_{yy'<-1} + (1 - yy')_+^2 \mathbf{1}_{yy'\geq -1}$



- These losses can be understood as a convex approximation of the 0-1 loss $\ell(y, y') = \mathbf{1}_{yy'\leq 0}$

A comparison of classifiers on toy datasets



Standard error metrics in classification

- Precision, Recall, F-Score, AUC

For each sample i we have

- an actual label y_i
- a predicted label $\hat{y}_i$

We can construct the **confusion matrix**

Actual Class \ Predicted Class	Yes	No
Yes	TP	FN
No	FP	TN

with

$$\begin{aligned} \text{TP} &= \sum_{i=1}^{n} \mathbf{1}_{y_i=1,\hat{y}_i=1} \\ \text{TN} &= \sum_{i=1}^{n} \mathbf{1}_{y_i=-1,\hat{y}_i=-1} \\ \text{FN} &= \sum_{i=1}^{n} \mathbf{1}_{y_i=1,\hat{y}_i=-1} \\ \text{FP} &= \sum_{i=1}^{n} \mathbf{1}_{y_i=-1,\hat{y}_i=1} \end{aligned}$$

with yes $= 1$ and no $= -1$

Standard error metrics in classification

$$\text{Precision} = \frac{\text{TP}}{\#(\text{predicted P})} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\#(\text{real P})} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{F-Score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Some vocabulary

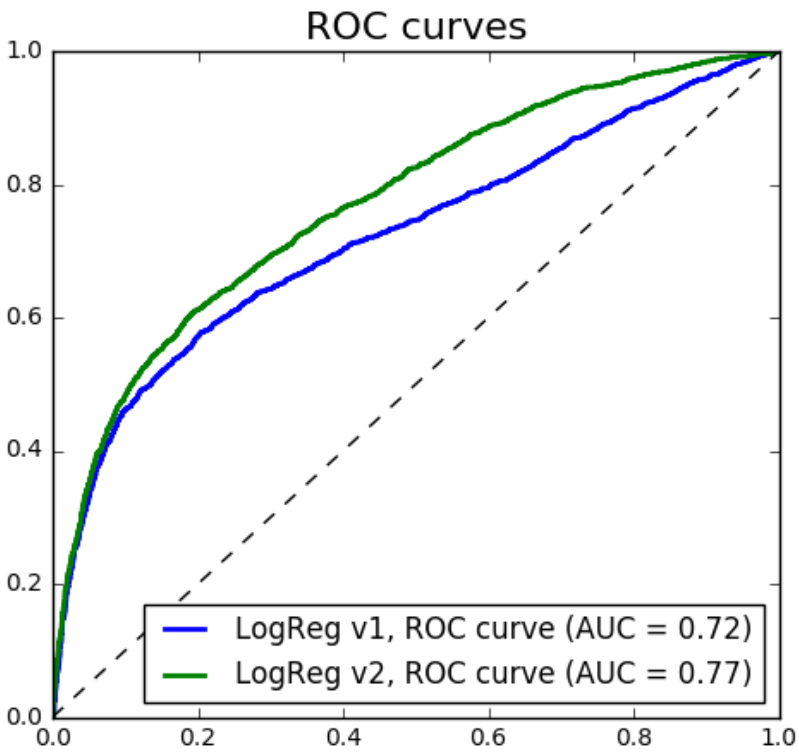
- Recall = Sensitivity
- False-Discovery Rate FDR $= 1 -$ Precision

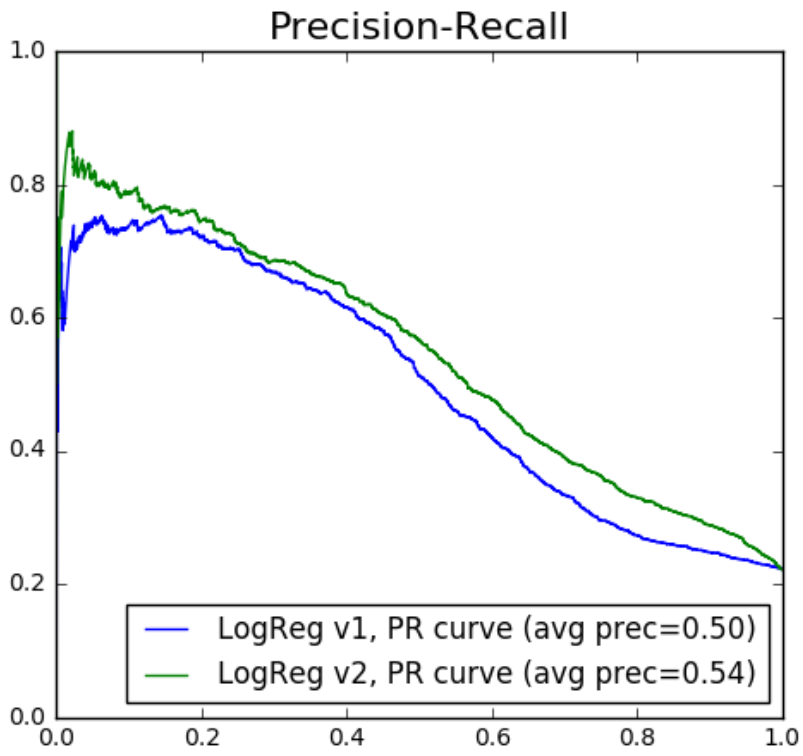
ROC Curve (Receiver Operating Characteristic)

- Based on the estimated probabilities $\hat{p}_{i,1} = \hat{\mathbb{P}}(Y = 1|X = x_i)$
- Each point A_t of the curve has coordinates $(\text{FPR}_t, \text{TPR}_t)$, where FPR_t and TPR_t are FPR and TPR of the confusion matrix obtained by the classification rule

$$\hat{y}_i = \begin{cases} 1 & \text{if } \hat{p}_{i,1} \geq t \\ -1 & \text{otherwise} \end{cases}$$

  for a threshold t varying in $[0, 1]$
- AUC score is the Area Under the ROC Curve


ROC curves
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
LogReg v1, ROC curve (AUC = 0.72)
LogReg v2, ROC curve (AUC = 0.77)


Precision-Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
LogReg v1, PR curve (avg prec=0.50)
LogReg v2, PR curve (avg prec=0.54)

Penalization to avoid overfitting

Computing

$$\hat{w}, \hat{b} \in \underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, w \rangle + b)$$

generally leads to a bad classifier. Minimize instead

$$\hat{w}, \hat{b} \in \underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, w \rangle + b) + \frac{1}{C} \operatorname{pen}(w) \right\}$$

where

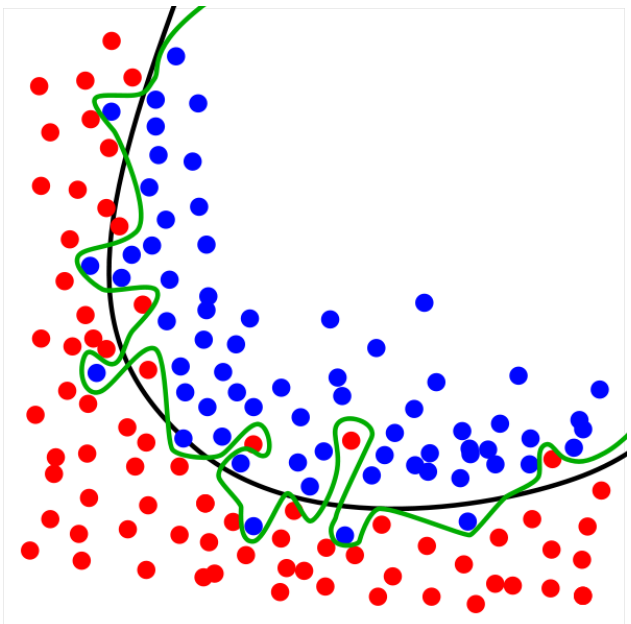
- pen is a **penalization** function, it forbids w to be "too complex"
- $C > 0$ is a **tuning** or **smoothing** parameter, that **balances** goodness-of-fit and penalization

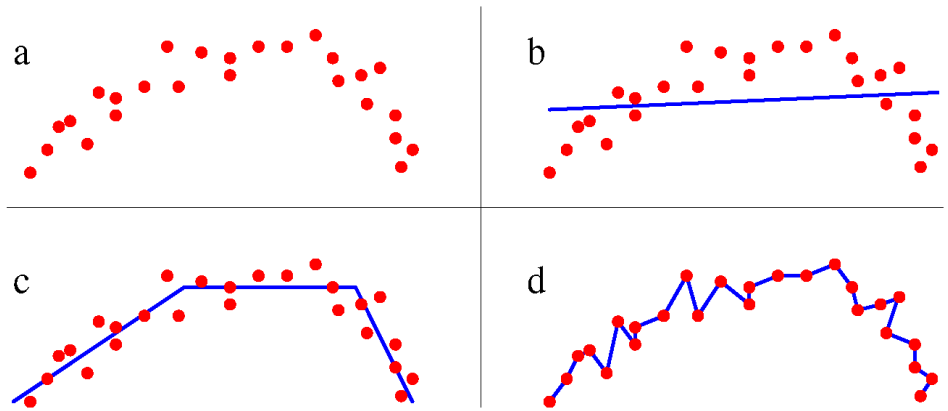
Penalization to avoid overfitting

In the problem

$$\hat{w}, \hat{b} \in \underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, w \rangle + b) + \frac{1}{C} \operatorname{pen}(w) \right\},$$

a well-chosen $C > 0$, allows to avoid **overfitting**





Overfitting is what you want to avoid

Which penalization? The **ridge** penalization considers

$$\text{pen}(w) = \frac{1}{2}\|w\|_2^2 = \frac{1}{2}\sum_{j=1}^{d} w_j^2$$

It penalizes the "size" of w

In the case of the SVM (hinge loss) it has a nice interpretation: **corresponds to the margin** (more on that later)

This is the most widely used penalization

- It's nice and easy
- It allows to "deal" with correlated features (more on that later)
- It actually helps training! With a ridge penalization, the optimization problem is easier (more on that later)

There is another desirable property on $\hat{w}$

If $\hat{w}_j = 0$, then feature j has no impact on the prediction:

$$\hat{y} = \text{sign}(\langle x, \hat{w} \rangle + \hat{b})$$

If we have many features (d is large), it would be nice if $\hat{w}$ contained **zeros**, and many of them

- Means that only **few** features are statistically relevant.
- Means that only **few** features are useful to predict the label

Leads to a simpler model, with a "reduced" dimension

How to do it ?

Tempting to use

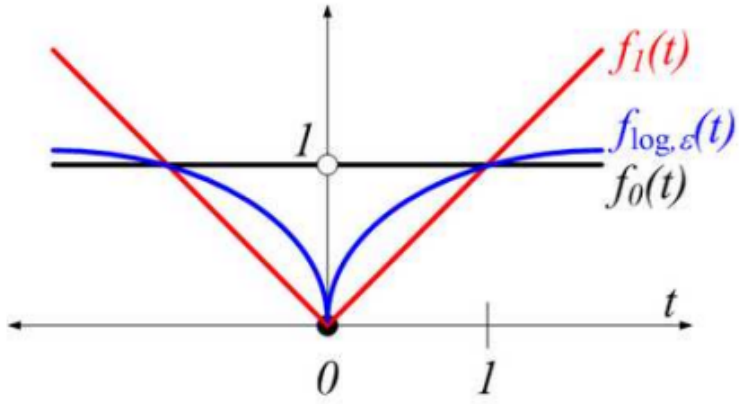
$$\hat{w}, \hat{b} \in \underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, w \rangle + b) + \frac{1}{C} \|w\|_0 \right\},$$

where

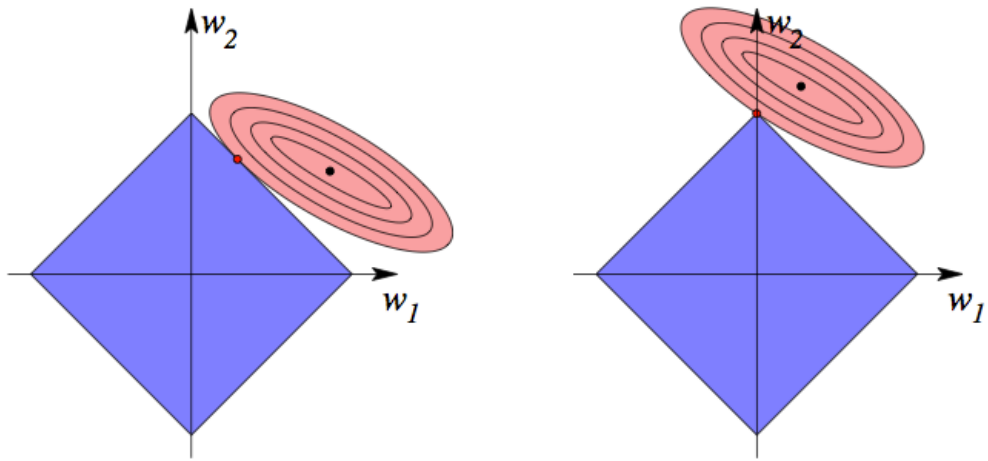
$$\|w\|_0 = \#\{j \in \{1, \dots, d\} : w_j \neq 0\}.$$

To solve this, explore **all** possible supports of w. Too long! (NP-hard)

Find a convex proxy of $\|\cdot\|_0$: the ℓ_1-**norm** $\|w\|_1 = \sum_{j=1}^{d} |w_j|$



Why does it induce sparsity?



Why ℓ_2 (ridge) does not induce sparsity?

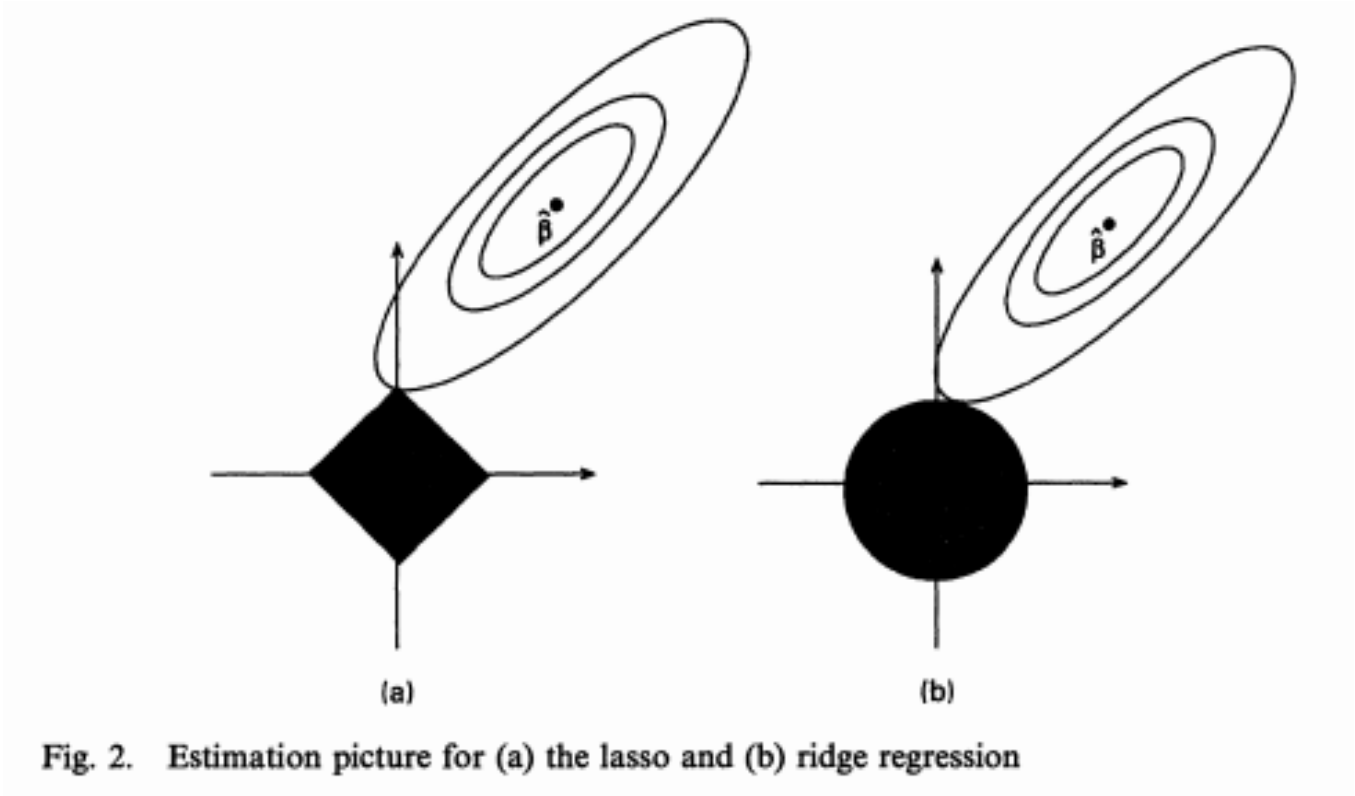


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

A direct computation

Consider the minimization problem

$$\min_{z' \in \mathbb{R}} \frac{1}{2}(z' - z)^2 + \lambda |z'|$$

for $\lambda > 0$ and $z \in \mathbb{R}$

- Derivative at 0_+: $d_+ = \lambda - z$
- Derivative at 0_-: $d_- = -\lambda - z$

Let z_* be the solution

- $z_* = 0$ iff $d_+ \geq 0$ and $d_- \leq 0$, namely $|z| \leq \lambda$
- $z_* \geq 0$ iff $d_+ \leq 0$, namely $z \geq \lambda$ and $z_* = z - \lambda$
- $z_* \leq 0$ iff $d_- \geq 0$, namely $z \leq -\lambda$ and $z_* = z + \lambda$

Hence

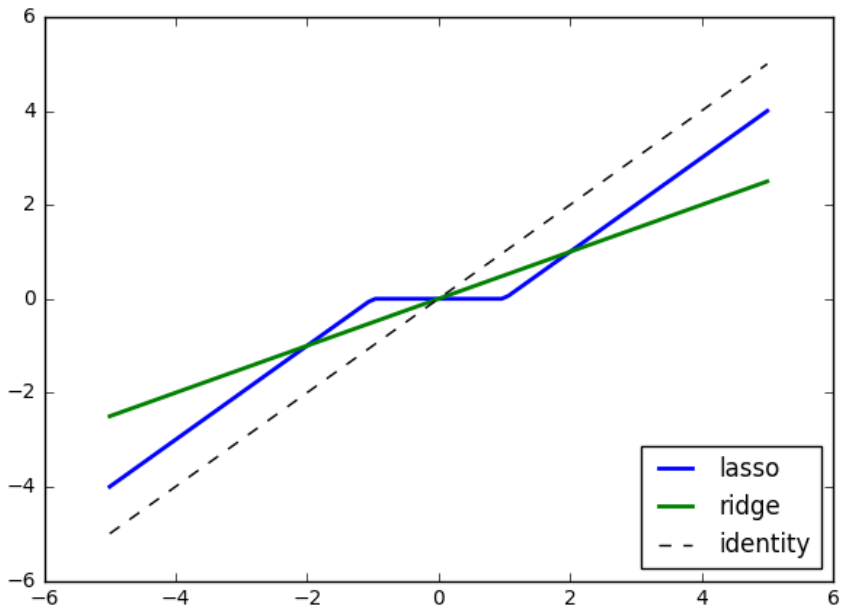
$$z_* = \text{sign}(z)(|z| - \lambda)_+.$$

$$\operatorname*{argmin}_{z' \in \mathbb{R}} \frac{1}{2}(z' - z)^2 + \frac{1}{C}|z'| = \operatorname{sign}(z)\left(|z| - \frac{1}{C}\right)_+$$

so that

$$\operatorname*{argmin}_{w' \in \mathbb{R}^d} \frac{1}{2}\|w' - w\|_2^2 + \frac{1}{C}\|w'\|_1 = \operatorname{sign}(w) \odot \left(|w| - \frac{1}{C}\right)_+.$$

Example with $C = 1$



Particular instances of problem

$$\hat{w}, \hat{b} \in \underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, w \rangle + b) + \frac{1}{C} \operatorname{pen}(w) \right\},$$

For $\ell(y, y') = \frac{1}{2}(y - y')^2$ and $\operatorname{pen}(w) = \frac{1}{2}\|w\|_2^2$, the problem is called **ridge regression**

For $\ell(y, y') = \frac{1}{2}(y - y')^2$ and $\operatorname{pen}(w) = \|w\|_1$, the problem is called **Lasso** (Least absolute shrinkage and selection operator)

For $\ell(y, y') = \log(1 + e^{-yy'})$ and $\operatorname{pen}(w) = \|w\|_1$, the problem is called ℓ_1-**penalized logistic regression**

Many combinations possible...

The combinations

$$\text{(linear regression or logistic)} + \text{(ridge or } \ell_1\text{)}$$

are the most wildely used

Another penalization is

$$\text{pen}(w) = \frac{1}{2}\|w\|_2^2 + \alpha\|w\|_1$$

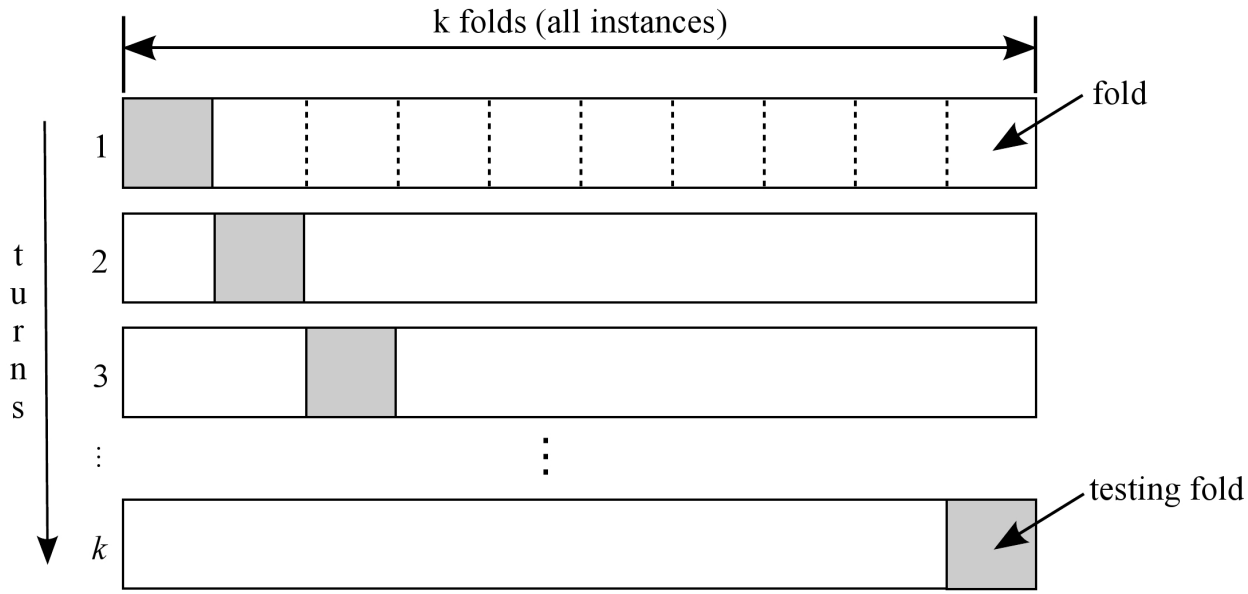
called **elastic-net**, benefits from both the advantages of ridge and ℓ_1 penalization (where $\alpha \geq 0$ balances the two)

Cross-validation

- **Generalization** is the goal of supervised learning
- A trained classifier has to be **generalizable**. It must be able to work on other data than the training dataset
- Generalizable means "works without **overfitting**"
- This can be achieved using **cross-validation**
- There is **no machine learning without cross-validation** at some point!
- In the case of penalization, we need to choose a penalization parameter C that generalizes

V-Fold cross-validation

- Most standard cross-validation technique
- Take $V = 5$ or $V = 10$. Pick a random partition $I_1, \ldots, I_V$ of $\{1, \ldots, n\}$, where $|I_v| \approx \frac{n}{V}$ for any $v = 1, \ldots, V$



Consider a set

$$\mathcal{C} = \{C_1, \dots C_K\}$$

of possible values for C. For each $v = 1, \dots, V$

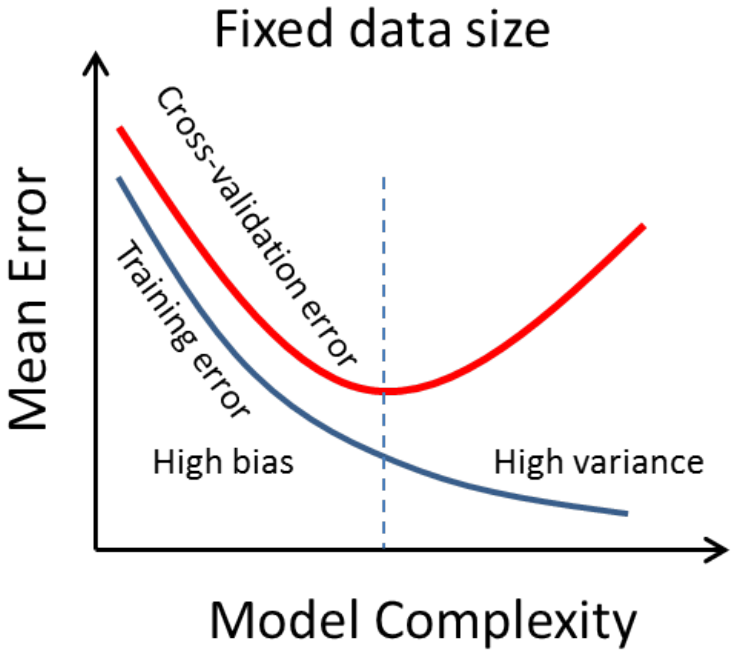
- Put $I_{v,\texttt{train}} = \cup_{v' \neq v} I_{v'}$ and $I_{v,\texttt{test}} = I_v$
- For each $C \in \mathcal{C}$, find

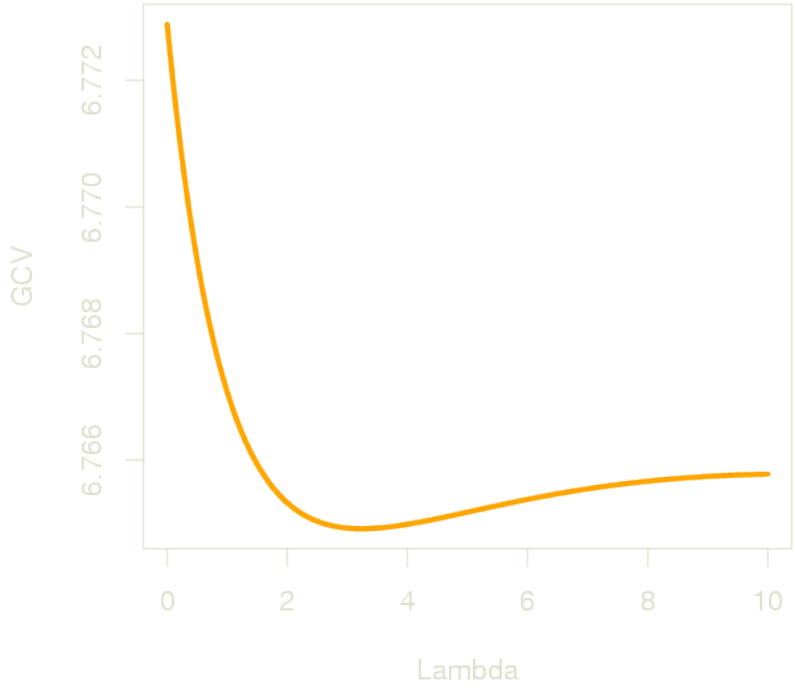
$$\hat{w}_{v,C} \in \operatorname*{argmin}_w \left\{ \frac{1}{|I_{v,\texttt{train}}|} \sum_{i \in I_{v,\texttt{train}}} \ell(y_i, \langle x_i, w \rangle) + \frac{1}{C} \operatorname{pen}(w) \right\}$$

Take

$$\hat{C} \in \operatorname*{argmin}_{C \in \mathcal{C}} \sum_{v=1}^{V} \sum_{i \in I_{v,\texttt{test}}} \ell(y_i, \langle x_i, \hat{w}_{v,C} \rangle)$$

Remark: depending on the problem, we might use a different loss (or score) for choosing $\hat{C}$





- Training error:

$$C \mapsto \sum_{v=1}^{V} \sum_{i \in I_{v,\texttt{train}}} \ell(y_i, \langle x_i, \hat{w}_{v,C} \rangle)$$

- Testing, validation or cross-validation error:

$$C \mapsto \sum_{v=1}^{V} \sum_{i \in I_{v,\texttt{test}}} \ell(y_i, \langle x_i, \hat{w}_{v,C} \rangle)$$

Class unbalancing

- In my supervised dataset there are 90% labels 0 and 10% labels 1, but I want to detect 1s
- What if I train without including this in my training rule?
- You'll only predict 0s!

In logistic regression, just correct the likelihood using the class balancing: put

$$\hat{q}_0 = \frac{n}{\{\#i : y_i = 0\}} \quad \text{and} \quad \hat{q}_1 = \frac{n}{\{\#i : y_i = 1\}}$$

Class unbalancing

The logistic goodness-of-fit is

$$R_n(w) = \frac{1}{n}\sum_{i=1}^{n}\Big(\log(1+e^{\langle w,x_i\rangle})\mathbf{1}_{y_i=1} + \log(1+e^{-\langle w,x_i\rangle})\mathbf{1}_{y_i=0}\Big)$$

Just replace it by

$$R_n(w) = \frac{1}{n}\sum_{i=1}^{n}\Big(\hat{q}_1\log(1+e^{\langle w,x_i\rangle})\mathbf{1}_{y_i=1} + \hat{q}_0\log(1+e^{-\langle w,x_i\rangle})\mathbf{1}_{y_i=0}\Big)$$

- This changes the gradient you use in a solver
- Gradient steps for 1s are larger then the ones for 0s, when $\#1 \ll \#0$ (more on optimization later)

Class unbalancing

- In an unbalanced dataset, when using V-Fold cross-validation, I'm likely to end up with a fold without 1s!

Use "stratified" V-Fold cross-validation:

- if there is $p_1\%$ of label 1s in the dataset
- proportion of of 1s must be $p_1\%$ inside each fold
- easy: put 1s in the dataset first, and find fold number of a line using the modulo with the number of folds (see above)

Features scaling

- Features matrix X with n-lines and d-columns
- $X_{\bullet,j} = j$-th column of X and $X_{j,\bullet} = j$-th row.

Scale of features vector $X_{\bullet,1}, \ldots, X_{\bullet,d}$ is important at the training step

- when using penalization, the coefficients of the classifier won't be penalized the same
- Lipschitz constant of the loss often depend on $\|X_{\bullet,j}\|_2$ (e.g. logistic): features with large scale slow down convergence

Often need to scale the features:

- center, include an intercept, standardize
- min-max scaling
- binarize

Features scaling

On continuous features (continuous is discrete with many modalities...)

- Centering and standardization (or "whitening") of j-th feature: replace $X_{\bullet,j}$ by

$$\frac{X_{\bullet,j} - \bar{X}_{\bullet,j}}{\|X_{\bullet,j} - \bar{X}_{\bullet,j}\|_2}$$

  where $\bar{X}_{\bullet,j} = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}$

- Min-max scaling of j-th feature: replace $X_{\bullet,j}$ by

$$\frac{X_{\bullet,j} - \min_i X_{i,j}}{\max_i X_{i,j} - \min_i X_{i,j}}$$

  (better for sparse features: keep the zeros)

- Include an intercept: include a constant feature $X_{\bullet,0} = \mathbf{1}$

Features scaling

Feature binarization of j-th feature

If $X_{\bullet,j}$ is discrete

- If $X_{i,j} \in \{1, \dots M_j\}$, $M_j =$ number of modalities (small) create $M_j - 1$ new "dummy" binary features: replace

$$\begin{bmatrix} 1 \\ 1 \\ 2 \\ 1 \\ 3 \\ 3 \end{bmatrix} \text{ by } \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

Features scaling

If $X_{\bullet,j}$ is continuous

- Choose number of bins M
- Compute the quantiles $q_{m/M}$ for $m = 0, \ldots, M$ of $X_{\bullet,j}$, put $I_m = [q_{(m-1)/M}, q_{m/M}]$ for $m = 1, \ldots, M$
- Create $M - 1$ dummy binary features $\widetilde{X}_{\bullet,j,1}, \ldots, \widetilde{X}_{\bullet,j,M-1}$ such that

$$\widetilde{X}_{i,j,m} = 1 \quad \text{if} \quad X_{i,j} \in I_m$$

for $m = 1, \ldots, M - 1$

Featuring for text data

Corpus:

```
[ "The lecture about machine learning is really awesome",
  "The teacher is nice and funny. The teacher is a nerd",
  "I'm wondering what I'm going to do with all of this",
  "Maybe create a startup or maybe use these ideas in finance",
  "Maybe I'm just curious about learning things" ]
```

Features:

```
['about', 'all', 'and', 'awesome', 'create', 'curious', 'do',
 'finance', 'funny', 'going', 'ideas', 'in', 'is', 'just', 'learning',
 'lecture', 'machine', 'maybe', 'nerd', 'nice', 'of', 'or', 'really',
 'startup', 'teacher', 'the', 'these', 'things', 'this', 'to', 'use',
 'what', 'with', 'wondering']
```

Binarized features:

```
[[1 0 0 1 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0]
 [0 0 1 0 0 0 0 0 1 0 0 0 2 0 0 0 0 0 1 1 0 0 0 0 2 2 0 0 0 0 0 0 0 0]
 [0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 1 1 1]
 [0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 2 0 0 0 1 0 1 0 0 1 0 0 0 1 0 0 0]
 [1 0 0 0 0 1 0 0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0]]
```

Featuring for text data

With many documents and many words, use hashing

Hash function:

$$\text{set of all possible words} \to \{1, \ldots, M\}$$

as much injective as possible. It gives the position of each word in a vector

```
{'and': 26, 'all': 28, 'just': 46, 'awesome': 14, 'startup': 12,
  'learning': 6, 'in': 25, 'curious': 41, 'nerd': 49, 'really': 3,
  'funny': 5, 'use': 10, 'things': 27, 'create': 0, 'ideas': 49,
  'machine': 0, 'to': 37, 'going': 33, 'wondering': 6, 'lecture': 9,
  'is': 12, 'nice': 47, 'do': 21, 'finance': 43, 'what': 20, 'with':
  8, 'teacher': 41, 'about': 12, 'these': 44, 'maybe': 49, 'this': 22,
  'of': 47, 'the': 34, 'or': 17}
```

Standard algorithm: `MurmurHash`

Featuring for text data

For scaling word counts ("bag of words"), standard scaling is given by TF-IDF (Time Frequenty - Inverse Document Frequency)

- Reflect how important a word is in a document, relatively to all documents in corpus
- Words $w_1, \ldots, w_J$, corpus of documents $\mathcal{D} = \{D_1, \ldots, D_I\}$
- Put

$$\text{TF}(w, D) = \#\text{ times } w \text{ occurs in } D$$

$$\text{IDF}(w, \mathcal{D}) = \log\left(\frac{\#\mathcal{D}}{\#\{D \in \mathcal{D} : w \in D\}}\right)$$

- Then

$$\text{TF-IDF}(w, D, \mathcal{D}) = \text{TF}(w, D) \times \text{IDF}(w, \mathcal{D})$$

Featuring for text data

Corpus:

```
[ "I like machine learning",
  "I like machine learning a lot",
  "I hate machine learning",
  "I don't understand machine learning",
  "I am an expert of machine learning",
  "My cousin is an expert of machine learning"]
```

Words:

```
['am', 'an', 'cousin', 'don', 'expert', 'hate', 'is', 'learning',
'like', 'lot', 'machine', 'my', 'of', 'understand']
```

TF-IDF:

```
[[ 0.    0.    0.    0.    0.    0.    0.    0.43  0.79  0.    0.43  0.    0.    0.  ]
 [ 0.    0.    0.    0.    0.    0.    0.    0.31  0.57  0.7   0.31  0.    0.    0.  ]
 [ 0.    0.    0.    0.    0.    0.85  0.    0.38  0.    0.    0.38  0.    0.    0.  ]
 [ 0.    0.    0.    0.65  0.    0.    0.    0.29  0.    0.    0.29  0.    0.    0.65]
 [ 0.54  0.44  0.    0.    0.44  0.    0.    0.24  0.    0.    0.24  0.    0.44  0.  ]
 [ 0.    0.35  0.43  0.    0.35  0.    0.43  0.19  0.    0.    0.19  0.43  0.35  0.  ]]
```

Next week

- The linear SVM: the hinge loss
- Kernels methods
- And some jokes too...

Thank you!