Large Scale Machine Learning

European Summer School in Financial Mathematics, Le Mans

S. Gaïffas

ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

Presentation

Stéphane Gaïffas

A. Professor, CMAP, Ecole polytechnique

http://www.cmap.polytechnique.fr/~gaiffas/

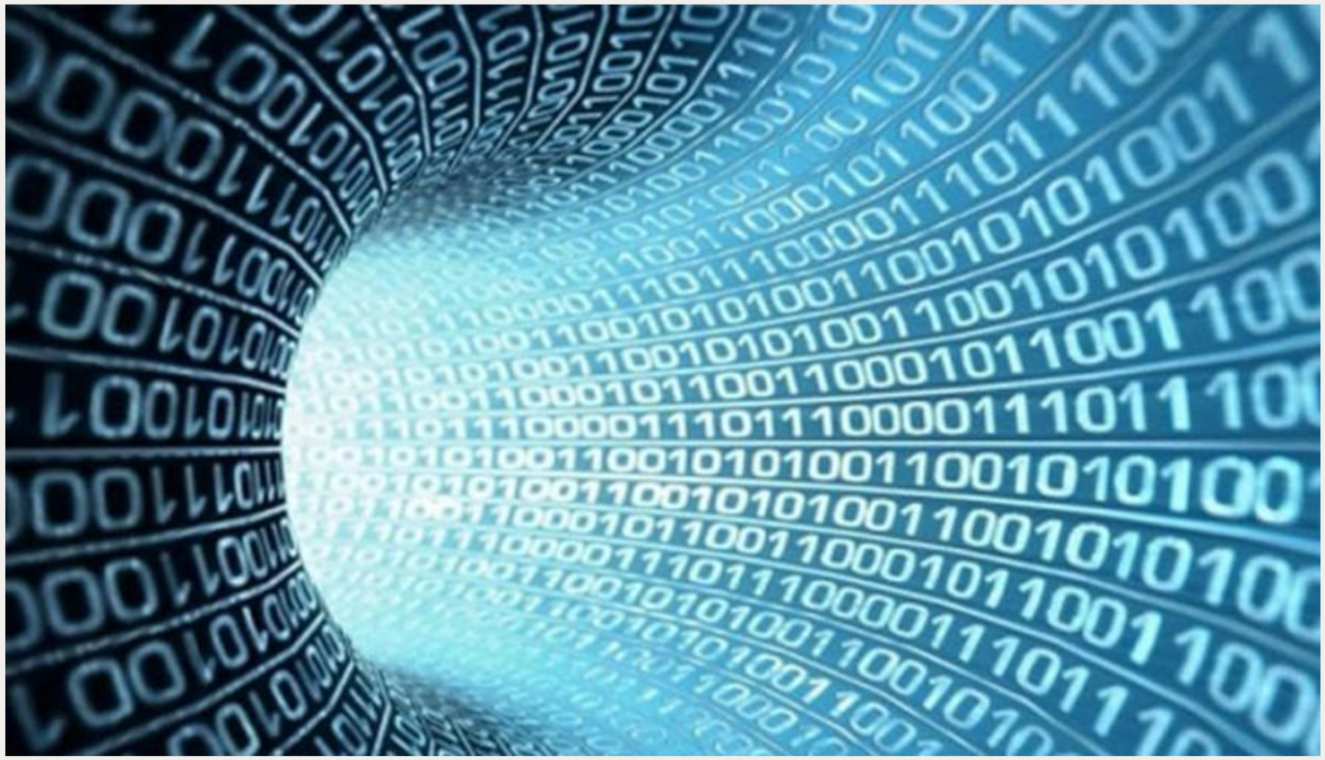
stephane.gaiffas@cmap.polytechnique.fr

Direct Matin

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

Le Monde

INTERNATIONAL | POLITIQUE | SOCIÉTÉ | ÉCONOMIE | CULTURE | IDÉES | SPORT | SCIENCES | TECHNO | STYLE | VOUS | ÉDITION ABONNÉS

M Idées

IDÉES | Les débats | Think tanks | Points de vue | Editoriaux | Opinions du Monde | Analyses | Idées chroniques | Chats | Blogs | Forums

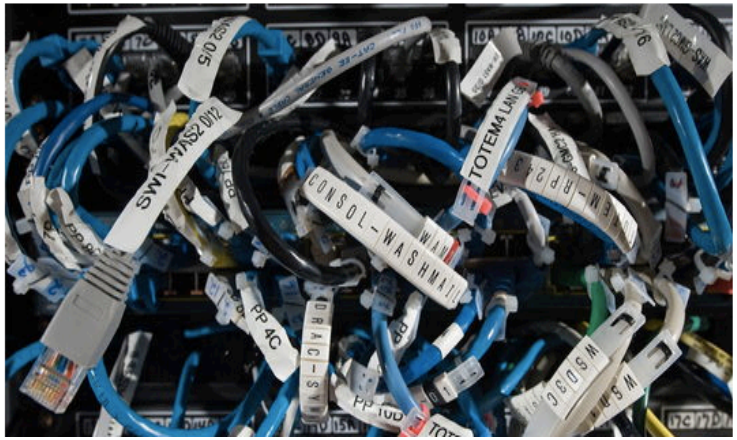
Les données, puissance du futur

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

Par Stéphane Grumbach, Stéphane Frénot

Abonnez-vous à partir de 1 € | Réagir | Classer | Imprimer | Envoyer | Partager

Recommander | Envoyer | 467 personnes le recommandent.



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

NY Times I

The New York Times Technology | Personal Tech | Business Day

Log In | Register Now

Bits

Search Bits Go

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

Big Data in More Hands

By QUENTIN HARDY

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT

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

AROUND THE WEB »

THE NEXT WEB
Google says Maps redirect on Windows Phone was a product decision, and will be removed

BLOOMBERG
HTC Posts Lowest Net Income in Eight Years After Revenue Drops

SCUTTLEBOT *News from the Web, annotated by our staff*

Google's Schmidt arrive in North Korea
REUTERS | From Mountain View to...errr, Pyongyang? - *Somini Sengupta*

AP provides sponsored tweets during electronics show
AP.ORG | The Associated Press is renting out its Twitter feed, with 1.5 million followers, to advertisers during C.E.S. - *Joshua Brustein*

A history of griefing
EDGE-ONLINE.COM | Meet the cult of gamers who want to ruin your day -- just for kicks. - *Jenna Wortham*

A Million First Dates
THE ATLANTIC | Is online romance threatening monogamy? - *Jenna Wortham*

SEE MORE »

NY Times II

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | TIMES TOPICS

Log In | Register Now | Help

The New York Times
SundayReview | The Opinion Pages

Search All NYTimes.com
Go

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION ARTS STYLE TRAVEL JOBS REAL ESTATE AUTOS

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

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

RECOMMEND
TWITTER
LINKEDIN
COMMENTS (82)
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Multimedia

Log in to see what your friends are sharing on nytimes.com.
Privacy Policy | What's This?

Log In With Facebook

What's Popular Now

Despite New Health Law, Some See Sharp Rise in Premiums

The Big Fail

MOST E-MAILED | RECOMMENDED FOR YOU

1. OFF THE DRIBBLE
Stoudemire Commemorates Brother's Death

2. CRITIC'S NOTEBOOK
The Rainbow That Follows 'Jersey Shore'

3. TAKING NOTE
Opinion Report: Tax Reform

4. THE LEARNING NETWORK
Fill-In | Trendy Spot Urges Tourists to Ride In and Spend, 'Gangnam Style'

NY Times III

HOME PAGE | TODAY'S PAPER | VIDEO | MOST POPULAR | U.S. Edition

Log In | Register Now | Help

The New York Times

Business Day

Search All NYTimes.com

Go

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | ARTS | STYLE | TRAVEL | JOBS | REAL ESTATE | AUTOS

Search | Global | DealBook | Markets | Economy | Energy | Media | Personal Tech | Small Business | Your Money

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.

FACEBOOK
TWITTER
GOOGLE+
E-MAIL
SHARE
PRINT
REPRINTS

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

Add to Portfolio

International Business Machines Corporation

Go to your Portfolio »

Log in to see what your friends are sharing on nytimes.com.
Privacy Policy | What's This?

Log In With Facebook

What's Popular Now

Despite New Health Law, Some See Sharp Rise in Premiums

The Big Fail

MOST E-MAILED | RECOMMENDED FOR YOU

1. OFF THE DRIBBLE
Stoudemire Commemorates Brother's Death

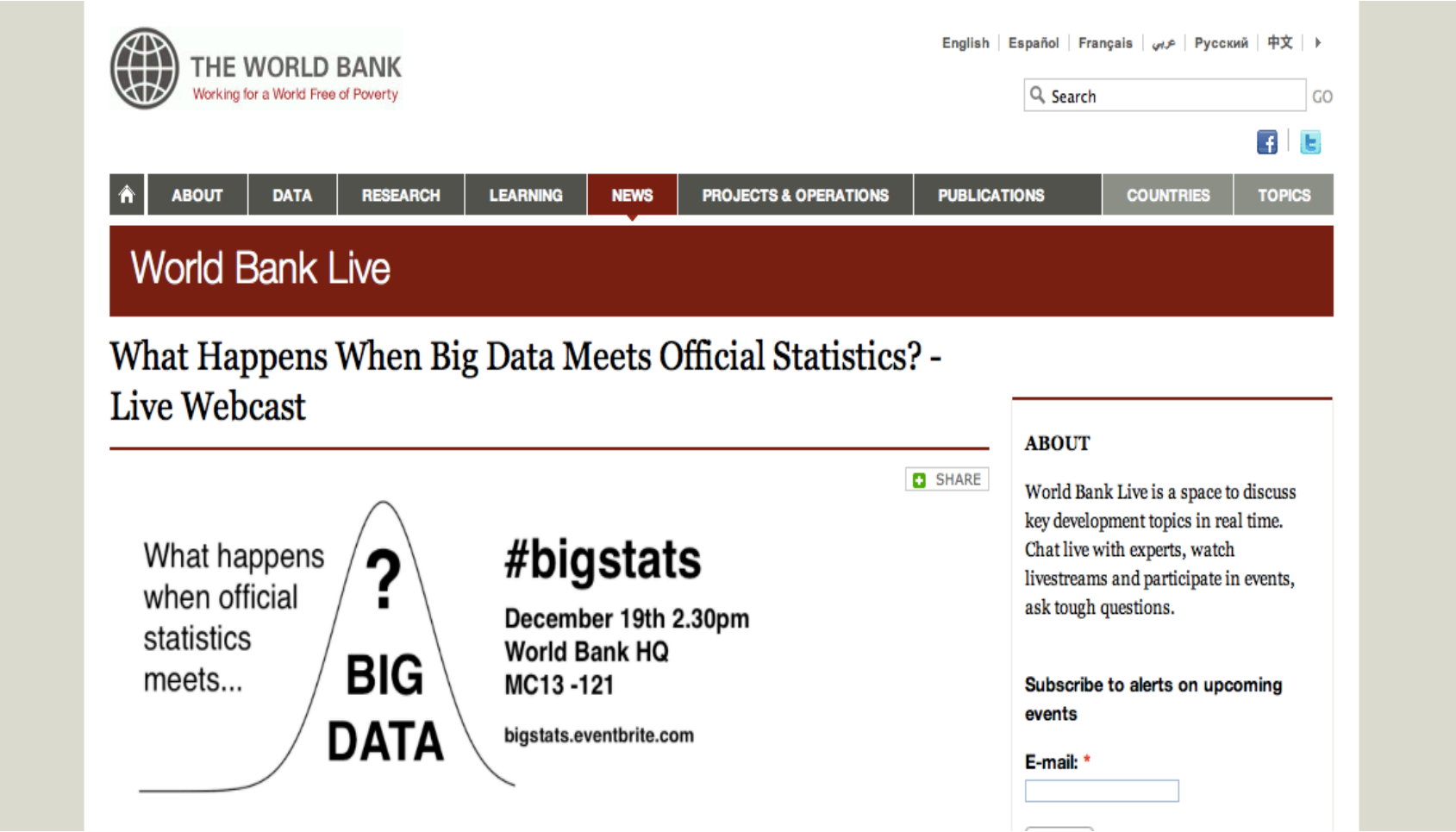
2. CRITIC'S NOTEBOOK
The Rainbow That Follows 'Jersey Shore'

3. TAKING NOTE
Opinion Report: Tax Reform

4. THE LEARNING NETWORK
Fill-In | Trendy Spot Urges Tourists to Ride In and Spend, 'Gangnam Style'

5. Major Companies Push the Limits of a Tax Break

World Bank


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
What happens when official statistics meets...
?
BIG DATA
#bigstats
December 19th 2.30pm
World Bank HQ
MC13 -121
bigstats.eventbrite.com
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

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

Criteo

criteo

Solutions CPOP Plateforme Editeurs Success Stories Actualités Carrières Contact Login annonceur | FR

Accueil > Actualités & Évènements

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.

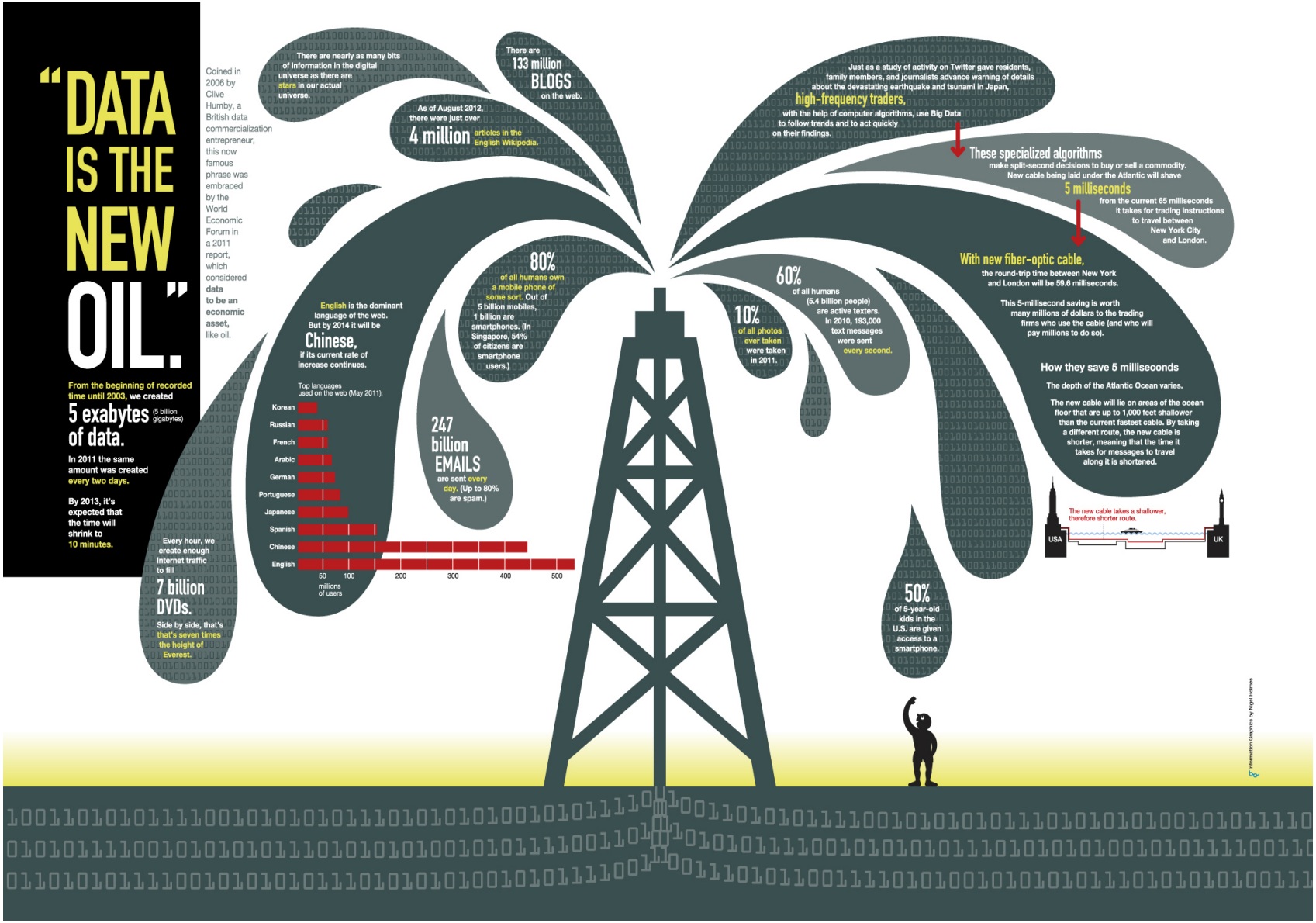
Ce nouveau siège, que Criteo a choisi délibérément de situer à Paris, vient ponctuer un développement continu, qui a permis à l'entreprise d'atteindre des résultats remarquables, 3 ans seulement après son lancement commercial :

- 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?



Web search

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**

Canoë

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**"

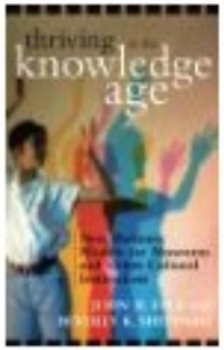
Moteur de recherche — Wikipédia
fr.wikipedia.org/wiki/**Moteur_de_recherche**
Un **moteur de recherche** est une application web permettant de retrouver des ressources (pages web, articles de forums Usenet, images, vidéo, fichiers, etc.) ...

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

Recommendation system

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

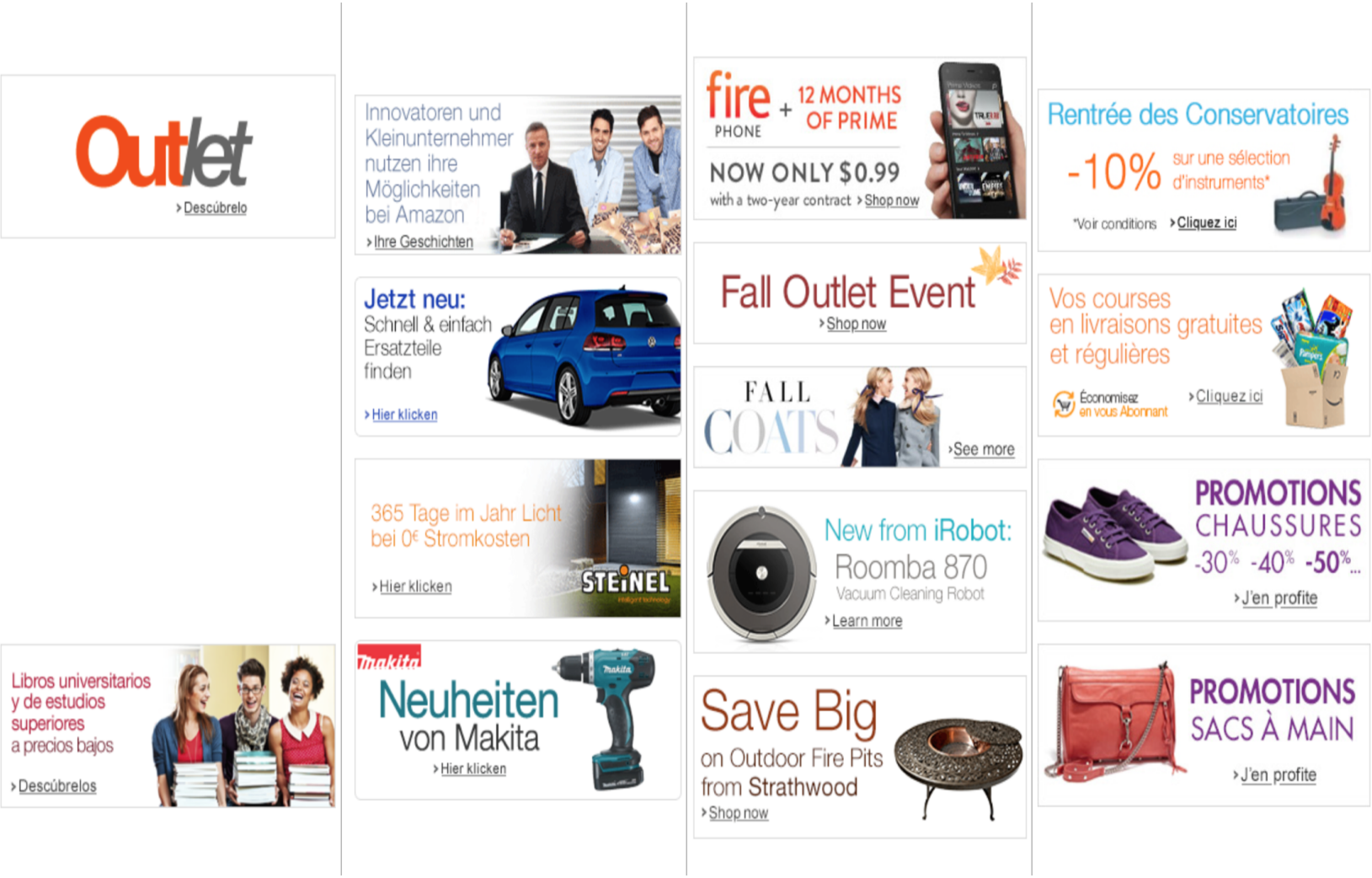
Museum Administration: An Introduction Paperback by Hugh H. Genoways
~~$31.95~~ **$28.75**

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...

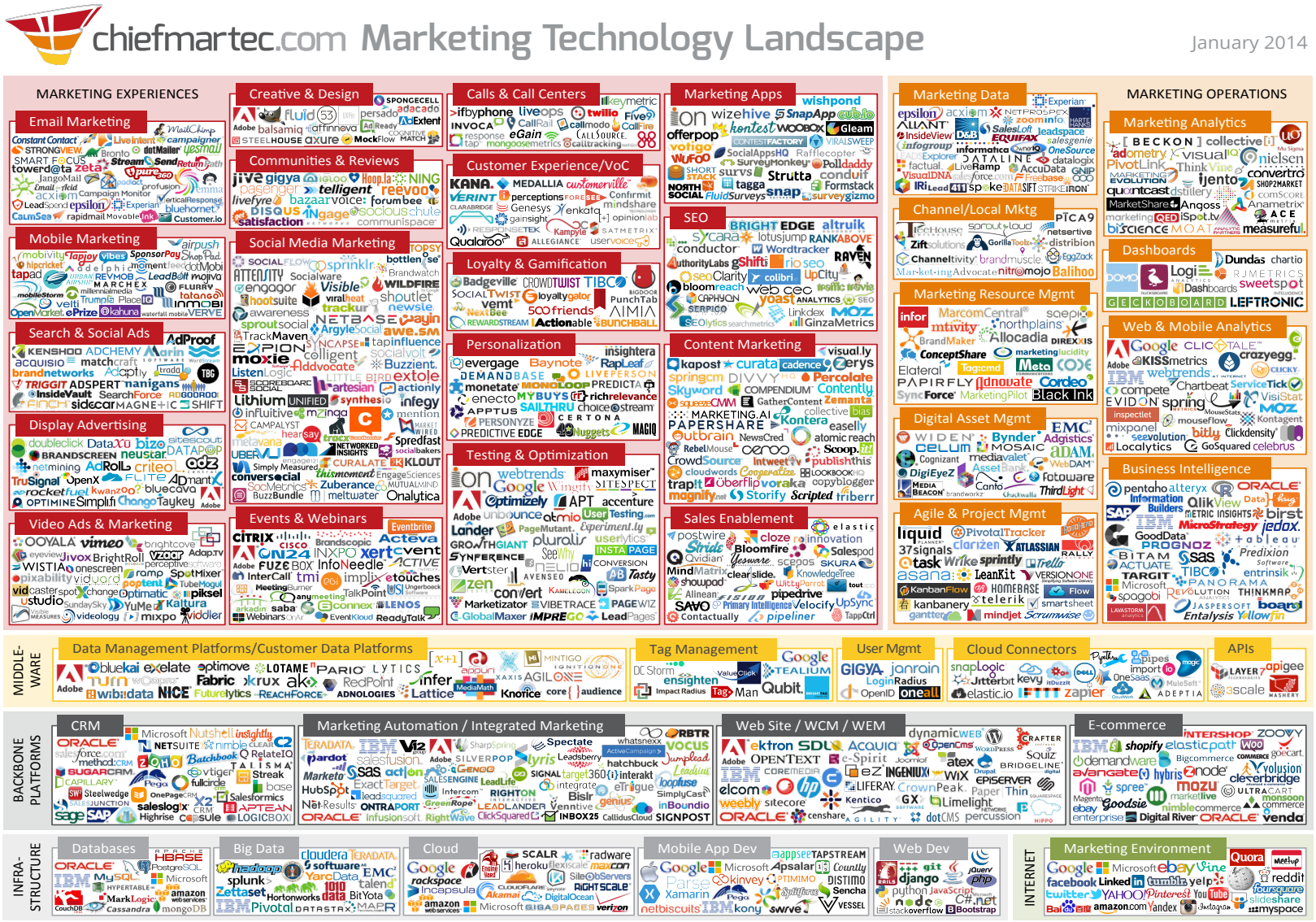
Advertisement



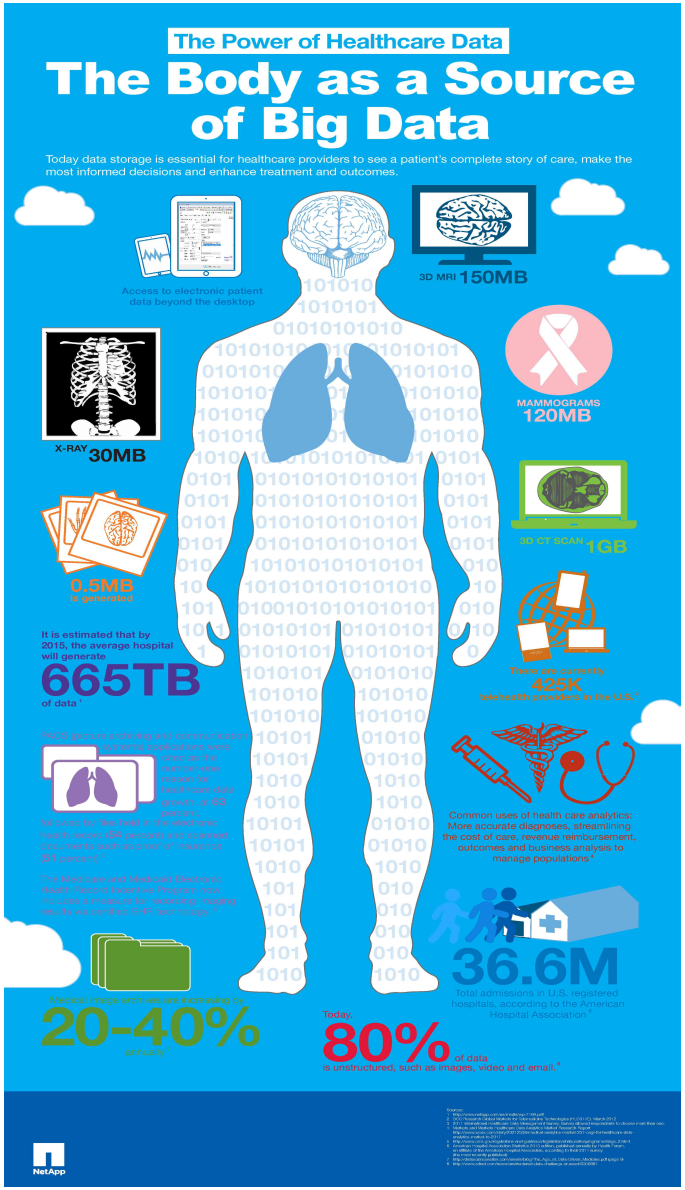
Intrusion detection

Crime prevention

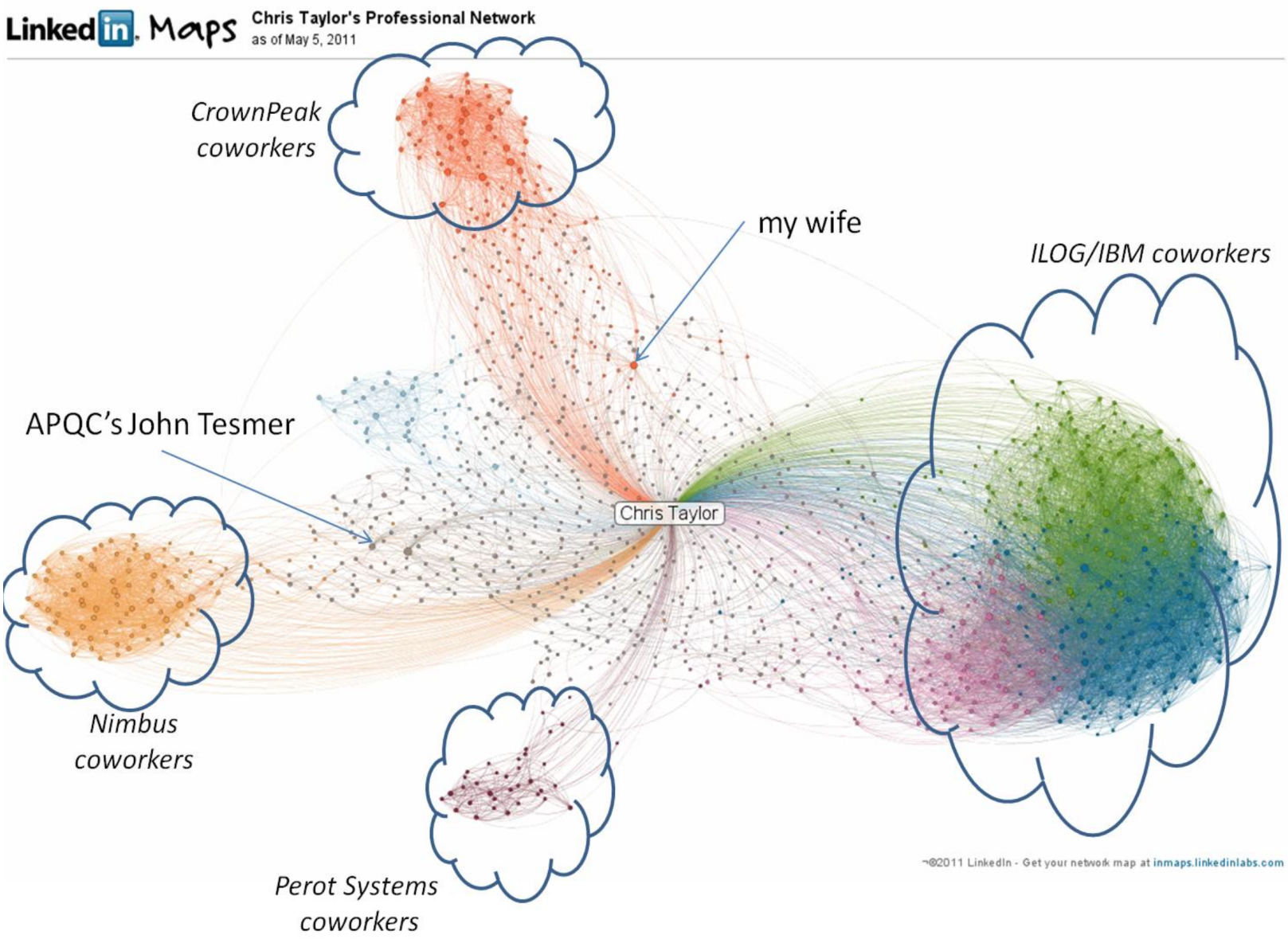
Marketing


chiefmartec.com Marketing Technology Landscape
January 2014
MARKETING EXPERIENCES
Email Marketing
Mobile Marketing
Search & Social Ads
Display Advertising
Video Ads & Marketing
Creative & Design
Communities & Reviews
Social Media Marketing
Events & Webinars
Calls & Call Centers
Customer Experience/VoC
Loyalty & Gamification
Personalization
Testing & Optimization
Marketing Apps
SEO
Content Marketing
Sales Enablement
Marketing Data
Channel/Local Mktg
Marketing Resource Mgmt
Digital Asset Mgmt
Agile & Project Mgmt
MARKETING OPERATIONS
Marketing Analytics
Dashboards
Web & Mobile Analytics
Business Intelligence
MIDDLE-WARE
Data Management Platforms/Customer Data Platforms
Tag Management
User Mgmt
Cloud Connectors
APIs
BACKBONE PLATFORMS
CRM
Marketing Automation / Integrated Marketing
Web Site / WCM / WEM
E-commerce
INFRA-STRUCTURE
Databases
Big Data
Cloud
Mobile App Dev
Web Dev
INTERNET
Marketing Environment
by Scott Brinker @chiefmartec http://chiefmartec.com

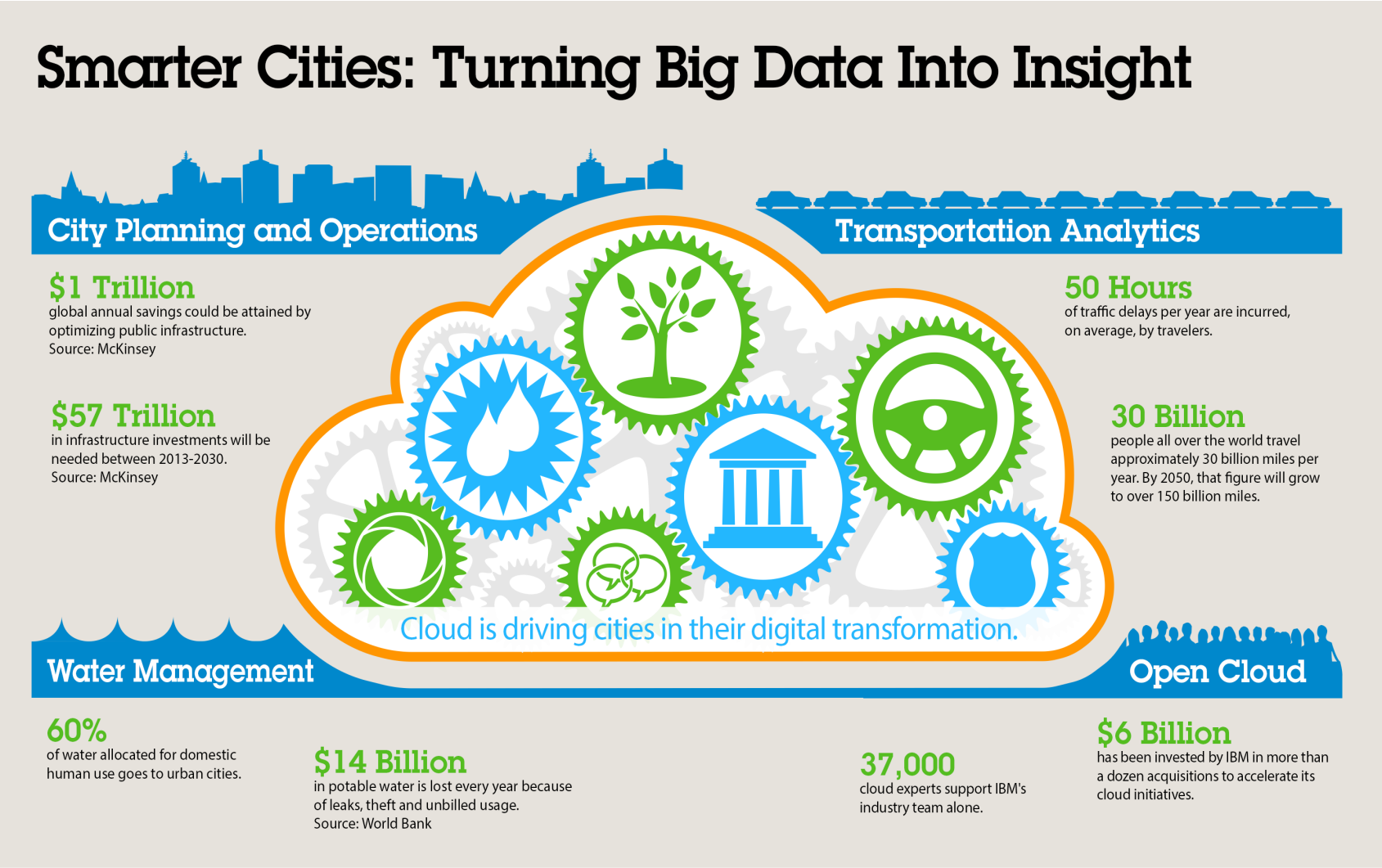
Health


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 potential
It is estimated that by 2015, the average hospital will generate
665TB
of data
425K
36.6M
Total admissions in U.S. registered hospitals, according to the American Hospital Association
Common uses of health care analytics: More accurate diagnoses, streamlining the cost of care, revenue reimbursement, outcomes and business analysis to manage populations
Medical image archives are increasing by
20-40%
annually
Today,
80%
of data
is unstructured, such as images, video and email.
NetApp

Linkedin

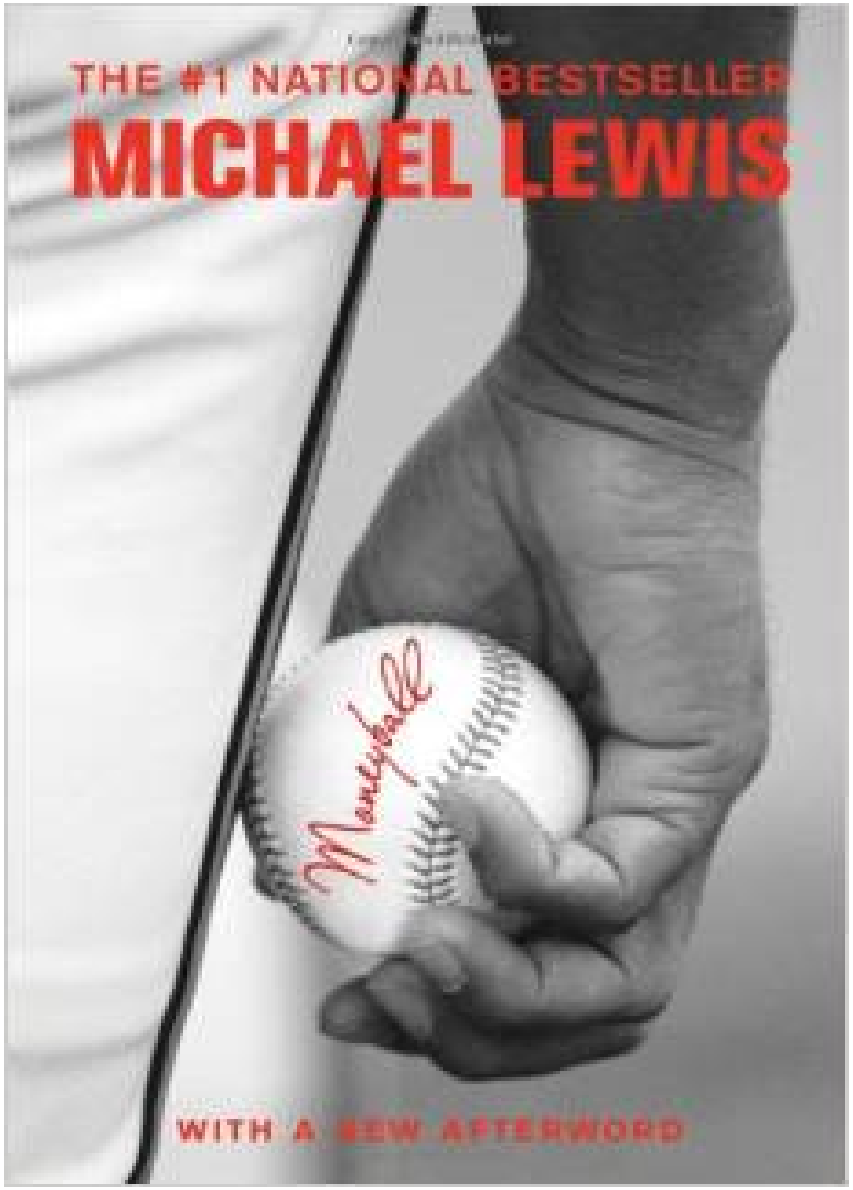

LinkedIn Maps
Chris Taylor's Professional Network
as of May 5, 2011
CrownPeak coworkers
my wife
ILOG/IBM coworkers
APQC's John Tesmer
Chris Taylor
Nimbus coworkers
Perot Systems coworkers
©2011 LinkedIn - Get your network map at inmaps.linkedinlabs.com

Smart city

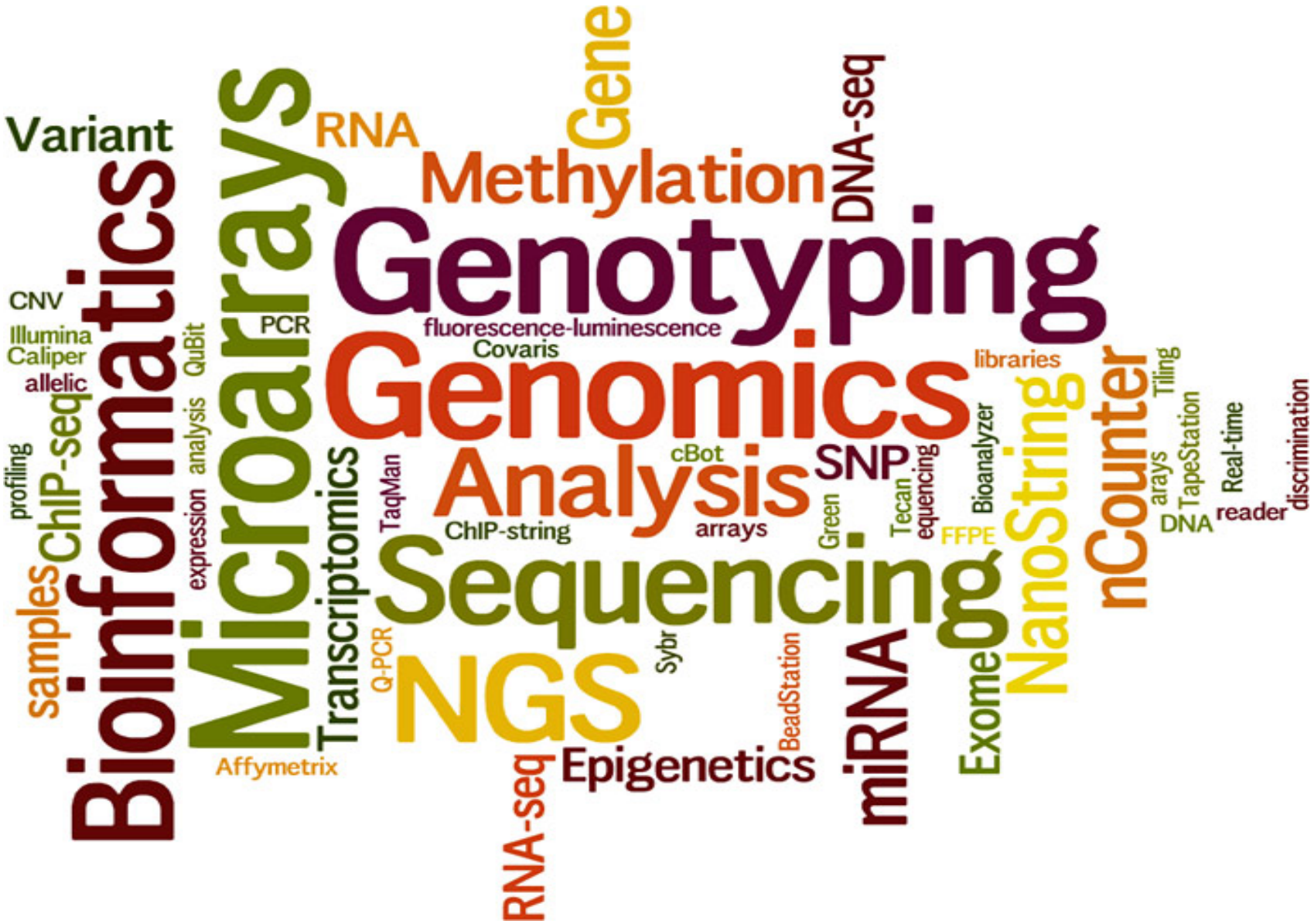


IBM Intelligent Operations software is designed with cities, for cities, to provide the tools to monitor, visualize and analyze vital city services such as water and wastewater systems, transportation, infrastructure planning, permit management and emergency response.

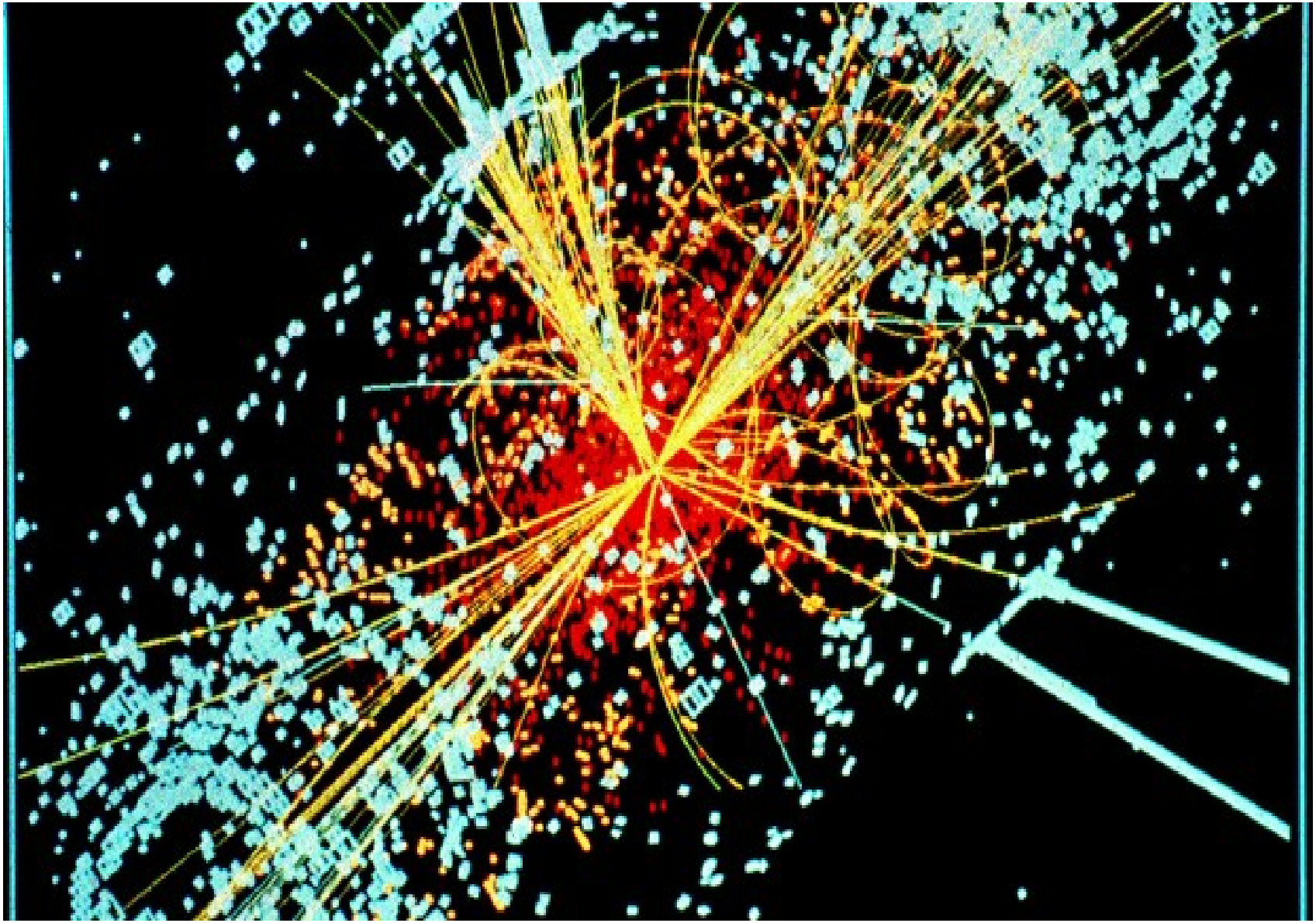
Sports


THE #1 NATIONAL BESTSELLER
MICHAEL LEWIS
WITH A NEW AFTERWORD

Genomics


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

Physics



Wikipedia I

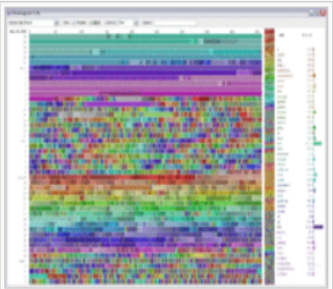
Big data

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]



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 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]

Wikipedia II

- **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

A new Context

Data everywhere

- Huge volume,
- Huge variety...

Affordable computation units

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

- Growing academic and industrial interest!

Big Data is (quite) Easy

Example of *off the shelves* solution

Spark — *Lightning-fast cluster computing*

amazon web services

```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()
}
```

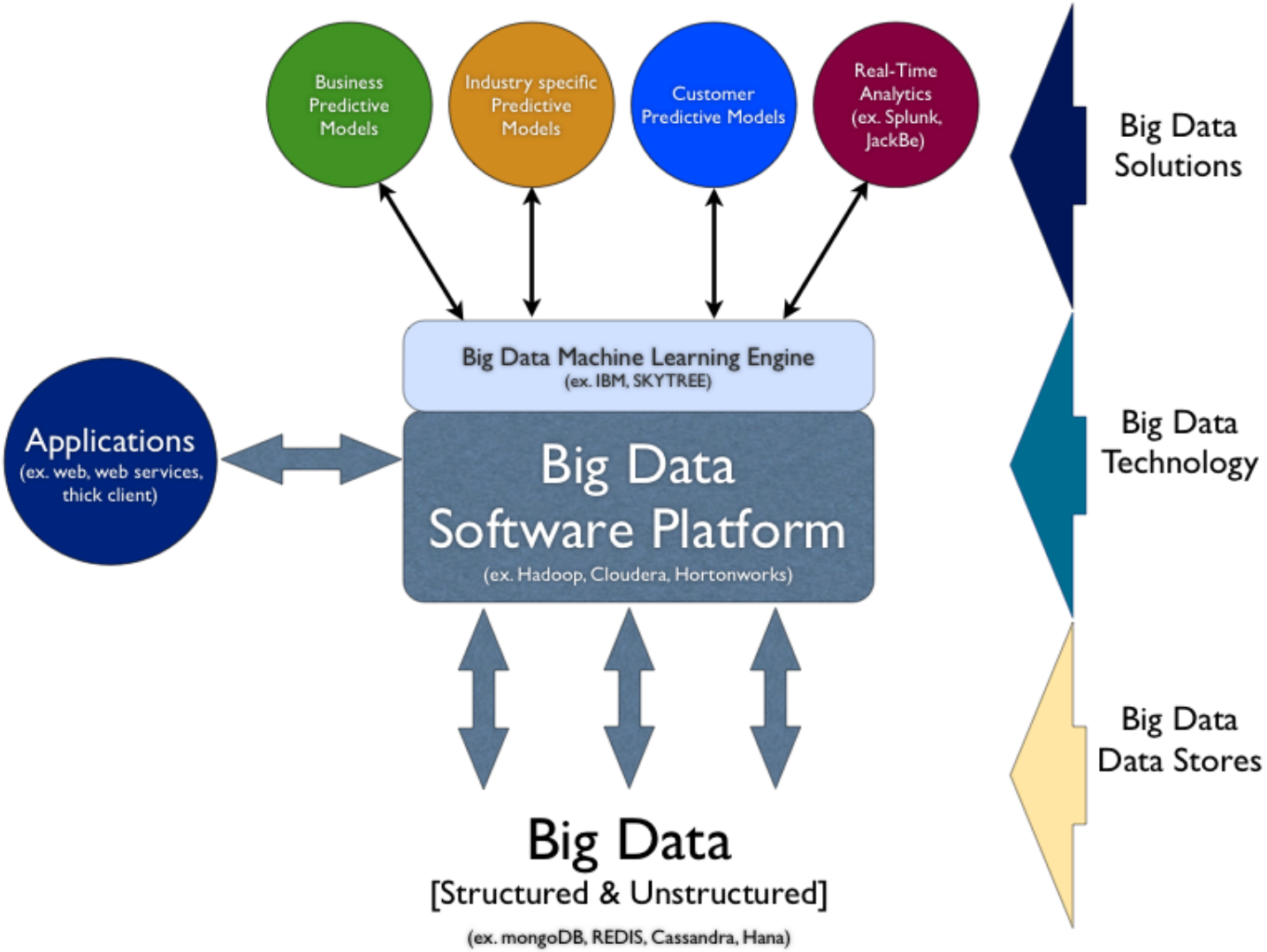
Big Data is (quite) Easy

Example of *off the shelves* solution

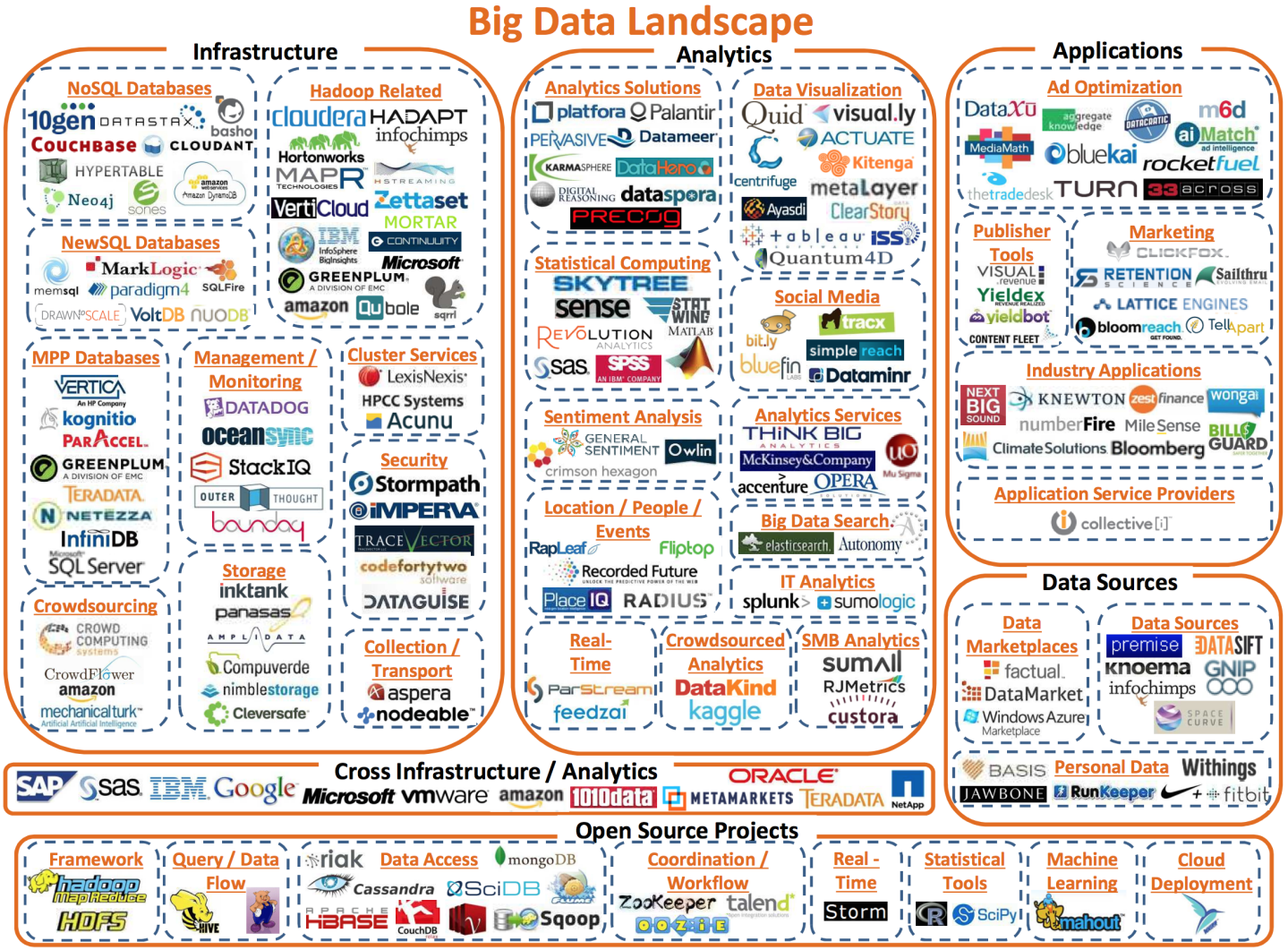

```
export AWS_ACCESS_KEY_ID=<your-access-keyid>
export AWS_SECRET_ACCESS_KEY=<your-access-key-secret>
cellule/spark/ec2/sparl-ec2 -i cellule.pem -k cellule -s <number of machines> launch <cluster-name>
ssh -i cellule.pem root@<your-cluster-master-dns>
spark-ec2/copy-dir ephemeral-hdfs/conf
ephemeral-hdfs/bin/hadoop 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! I



A Complex Ecosystem! II



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

Data science or statistics? I

A vocabulary problem:

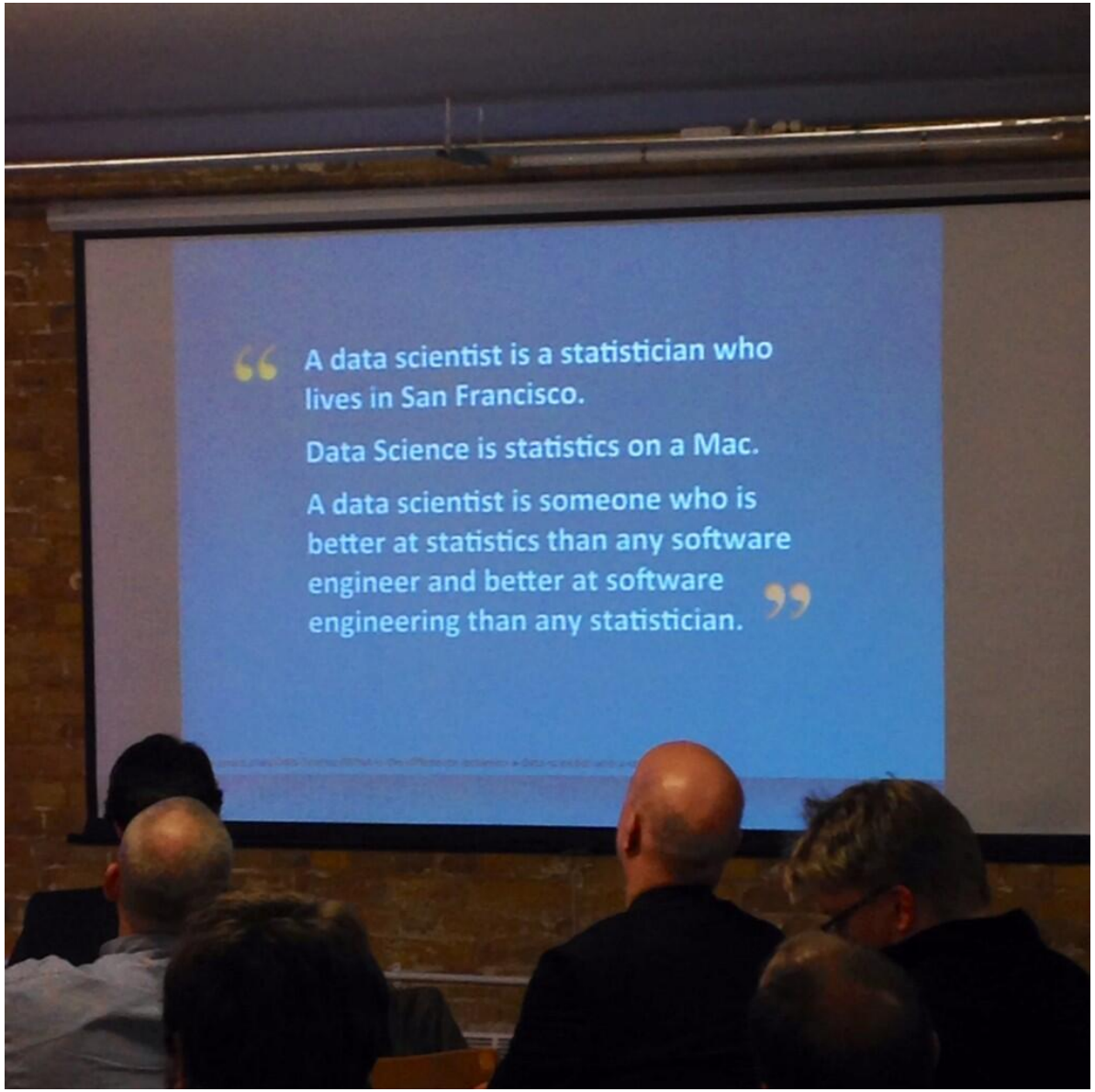
data scientist or statistician?

statistics or data science?

Data science or statistics? II

A possible answer:

Data science or statistics? III



Setting

- Data $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$ for $i = 1, \ldots, n$
- x_i is an input and y_i is an output
- x_i are called **features** and $x_i \in \mathcal{X} = \mathbb{R}^d$
- y_i are called **labels**
  $\mathcal{Y} = \{-1, 1\}$ or $\mathcal{Y} = \{0, 1\}$ for binary classification
  $\mathcal{Y} = \{1, \ldots, K\}$ for multiclass classification
  $\mathcal{Y} = \mathbb{R}$ for regression
- Usually, assume (x_i, y_i) are i.i.d
- **Goal:** given x, predict y.

- *High-dimension*: d is large, say $d \geq 10^4$
- *Big data*: n is large, say $n \geq 10^6$

Scenarios where:

- d is large, n is small: computational biology
- d is small, n is large: marketing
- d is large, n is large: web-advertisement, ad display

What to do

Minimize with respect to $f : \mathbb{R}^d \to \mathbb{R}$

$$R_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$$

where

- ℓ is a **loss** function. $\ell(y_i, f(x_i))$ small means y_i is close to $f(x_i)$
- $R_n(f)$ is called **goodness-of-fit** or **empirical risk**

Computation of f is called **training** or **estimation** step

- When d is large, impossible to fit a complex functions f on the data
- When n is large, training is too time-consuming for a complex function f

Hence:

- Choose a **linear** function f :

$$f(x) = \langle x, \theta \rangle = \sum_{j=1}^{d} x_j \theta_j,$$

for some parameter vector $\theta \in \mathbb{R}^d$ to be trained

Remark: linear with respect to x_i, but **you** can choose the x_i based on the data. Hence, not linear w.r.t the original features: "**feature engineering**"

- **Least-squares** loss (linear regression): $\ell(y,z) = \frac{1}{2}(y-z)^2$ for $y \in \mathbb{R}$, namely

$$R_n(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle x_i, \theta\rangle)^2 = \frac{1}{2n}\|Y - X\theta\|_2^2$$

where $X = [X_1, \cdots, X_n]^\top \in \mathbb{R}^{n\times d}$ and $y = [y_1, \cdots, y_n] \in \mathbb{R}^d$

- **Logistic regression** loss (logit, log-linear regression): $\ell(y,z) = \log(1+e^{-yz})$ for $y \in \{-1,1\}$, namely

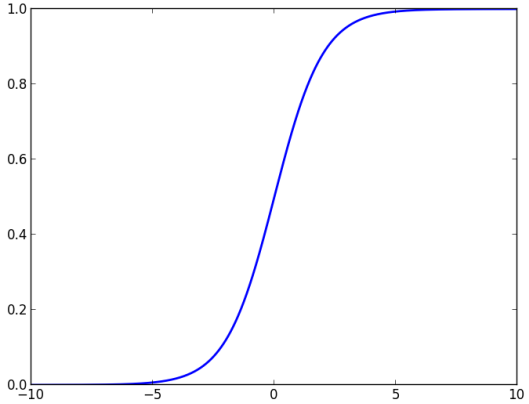
$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-y_i\langle x_i,\theta\rangle})$$

Binary classification: label $y \in \{0, 1\}$. Assume that

$$\mathbb{P}(Y = y | X = x) = \mathrm{Bernoulli}(\sigma_\theta(x))$$

with $\sigma_\theta(x) = \sigma(\langle \theta, x \rangle)$ where σ is the **sigmoid** function

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



Binary classification: label $y \in \{0, 1\}$. Assume that

$$\mathbb{P}(Y = y | X = x) = \mathrm{Bernoulli}(\sigma_\theta(x))$$

with $\sigma_\theta(x) = \sigma(\langle \theta, x \rangle)$ where σ is the **sigmoid** function

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

Hence for $y \in \{0, 1\}$:

$$\mathbb{P}(Y = y | X = x) = \sigma_\theta(x)^y (1 - \sigma_\theta(x))^{1-y} = \sigma_\theta(x)^y \sigma_\theta(-x)^{1-y}$$

and the log-likelihood is given by (if we replace label 0 by -1 for convenience)

$$\sum_{i=1}^{n} \log \mathbb{P}[Y = y_i | X = x_i] = -\sum_{i=1}^{n} \log(1 + e^{-y_i \langle x_i, \theta \rangle})$$

Goodness of fit $= -$ log-likelihood, so this leads to

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

Equivalent to assuming that the **log-odd ratio** is linear:

$$\log \left(\frac{\mathbb{P}[Y = 1 | X = x]}{\mathbb{P}[Y = 0 | X = x]} \right) = \langle x, \theta \rangle$$

This leads to a **linear** separation between the 1s and 0s. Logistic regression is a **linear** classifier

Now I've trained a logistic classifier: I have an estimation $\hat{\theta}$ of the parameters based on data $(x_1, y_1), \ldots, (x_n, y_n)$

I have a new point x_{n+1} but no label y_{n+1} for him. I want to have a prediction $\hat{y}_{n+1} \in \{-1, 1\}$ of its label

How do I proceed?

- I compute probability scores of 1 and -1:

$$\hat{p}_{n+1}^{(1)} = \frac{1}{1 + e^{-\langle x_{n+1}, \hat{\theta} \rangle}} \quad \text{and} \quad \hat{p}_{n+1}^{(-1)} = 1 - \hat{p}_{n+1}^{(1)}$$

- Now I predict the label using the MAP rule (Maximum A Posteriori)

$$\hat{y}_{n+1} = \begin{cases} 1 & \text{if } \hat{p}_{n+1}^{(1)} \geq t \\ -1 & \text{otherwise} \end{cases}$$

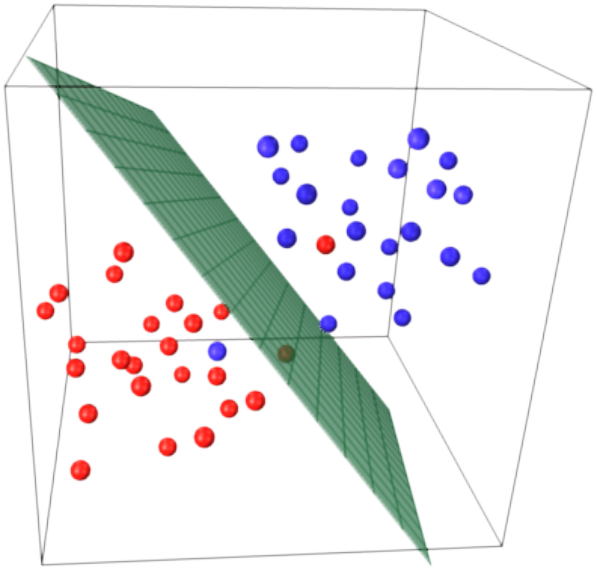
for a threshold $t \in (0, 1)$ (usually $t = 1/2$)

Remark:

$$\hat{p}_{n+1}^{(1)} \geq t \Leftrightarrow \langle x_{n+1}, \hat{\theta} \rangle \geq \log\left(\frac{t}{1-t}\right) \qquad (\langle x_{n+1}, \hat{\theta} \rangle \geq 0 \text{ if } t = 1/2)$$

This means that the logistic classifier separates data points into 1 and -1 with a hyperplane

We say that it is a **linear** classifier



Training the model: compute

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} R_n(\theta)$$

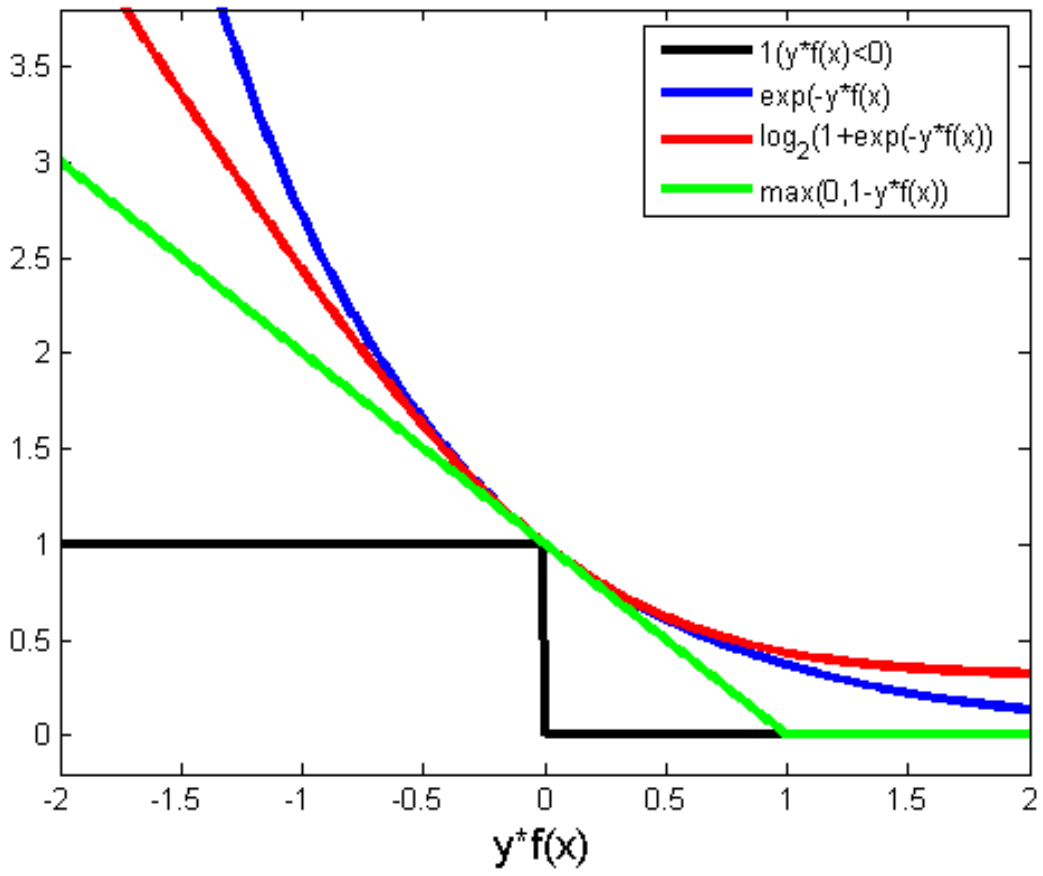
where

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, \theta \rangle).$$

Classical losses

- $\ell(y, z) = \frac{1}{2}(y - z)^2$: least-squares loss, linear regression (label $y \in \mathbb{R}$)
- $\ell(y, z) = (1 - yz)_+$ hinge loss, or SVM loss (binary classification, label $y \in \{-1, 1\}$)
- $\ell(y, z) = \log(1 + e^{-yz})$ logistic loss (binary classification, label $y \in \{-1, 1\}$)

Supervised learning – Loss functions, linearity



$$\ell_{\texttt{least-sq}}(y,z) = \frac{1}{2}(y-z)^2 \quad \ell_{\text{hinge}}(y,z) = (1-yz)_+$$

$$\ell_{\texttt{logistic}}(y,z) = \log(1+e^{-yz})$$

Penalization – Introduction

You should never actually fit a model by minimizing only

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

You should minimize instead

$$\hat{\theta}_n \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, \theta \rangle) + \lambda \, \mathrm{pen}(\theta) \right\}$$

where

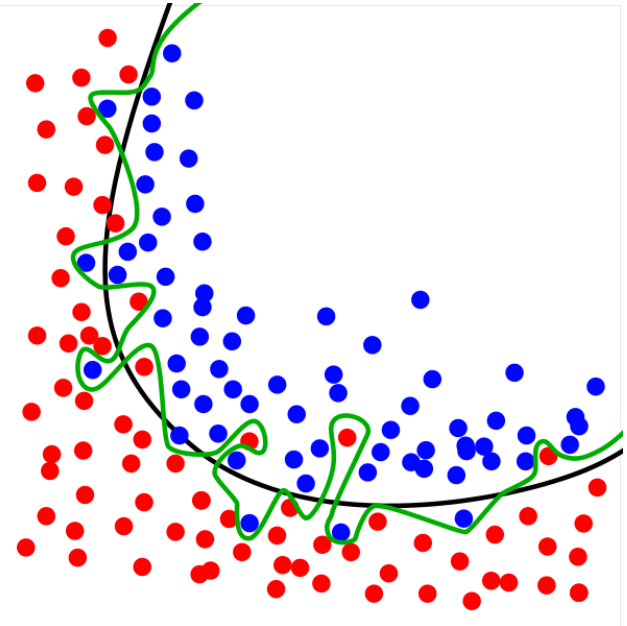
- pen is a **penalization** function, that encodes a prior assumption on θ. It forbids θ to be "too complex"
- $\lambda > 0$ is a **tuning** or **smoothing** parameter, that **balances** goodness-of-fit and penalization

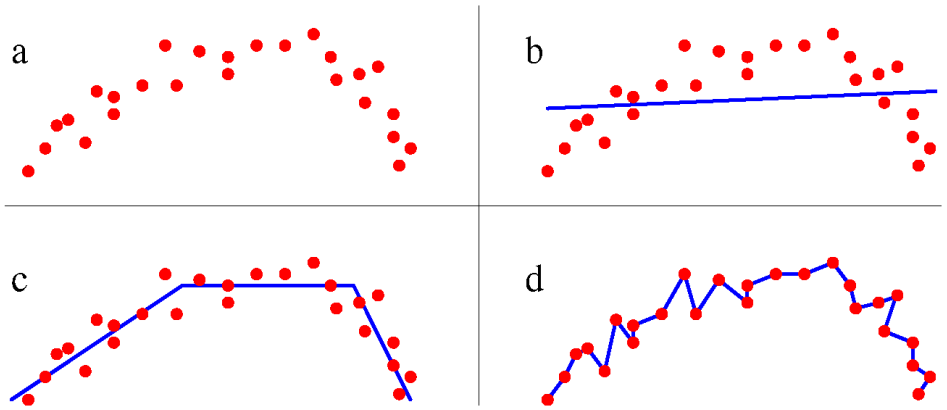
Penalization – Introduction

Why using penalization?

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

Penalization, for a well-chosen $\lambda > 0$, allows to avoid **overfitting**





Penalization – Ridge

Most classical penalization is the **Ridge** penalization

$$\text{pen}(\theta) = \|\theta\|_2^2 = \sum_{j=1}^{d} \theta_j^2.$$

It penalizes the energy of θ, measured by squared ℓ_2-norm

Sparsity inducing penalization.

- It would be nice to find a model where $\hat{\theta}_j = 0$ for many coordinates j
- few features are useful for prediction, the model is simpler, with a smaller dimension
- We say that $\hat{\theta}$ is **sparse**
- How to do it ?

Penalization – Sparsity

It is tempting to use

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, \theta \rangle) + \lambda \|\theta\|_0 \right\},$$

where

$$\|\theta\|_0 = \#\{j : \theta_j \neq 0\}.$$

But, to do it exactly, you need to try **all** possible subsets of non-zero coordinates of θ: 2^d possibilities. Impossible!

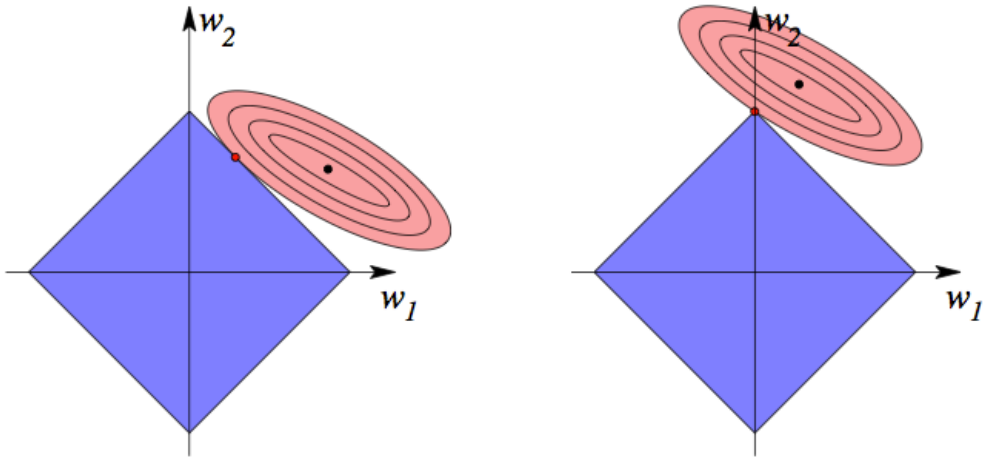
Penalization – Lasso

A solution: **Lasso** penalization (least absolute shrinkage and selection operator)

$$\text{pen}(\theta) = \|\theta\|_1 = \sum_{j=1}^{d} |\theta_j|.$$

This is penalization based on the ℓ_1-norm $\|\cdot\|_1$.

- In a noiseless setting, in a certain regime, ℓ_1-minimization gives the "same solution" as $\|\cdot\|_0$
- Why do ℓ_1-penalization leads to sparsity?



Why ℓ_2 (ridge) does not induce sparsity?

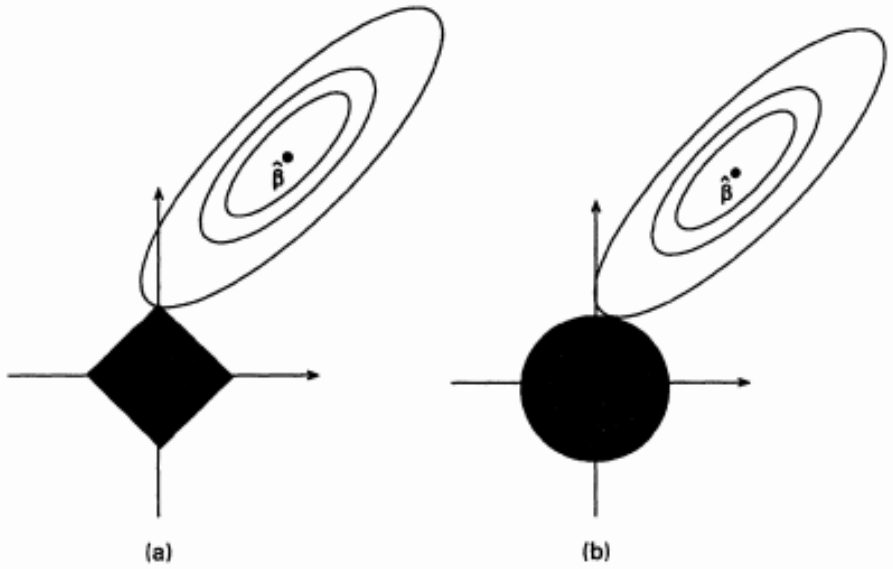


Fig. 2. Estimation picture for (a) the lasso and (b) ridge regression

Penalization – Lasso

Hence, a minimizer

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, \theta \rangle) + \lambda \|\theta\|_1 \right\}$$

is typically sparse ($\hat{\theta}_j = 0$ for many j).

- for λ large (larger than some constant) $\hat{\theta}_j = 0$ for all j
- for $\lambda = 0$ then there is no penalization
- Between the two, the "sparsity" depends on the value of λ: once again, it is a regularization or penalization parameter

Penalization – Lasso

For the least squares loss

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n} \| Y - X\theta \|_2^2 + \frac{\lambda}{2} \| \theta \|_2^2 \right\}$$

is called **ridge** linear regression, and

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n} \| Y - X\theta \|_2^2 + \lambda \| \theta \|_1 \right\}$$

is called **Lasso** linear regression.

Penalization – Lasso

Consider the minimization problem

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

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

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

Let a_* be the solution

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

Hence

$$a_* = \text{sign}(b)(|b| - \lambda)_+$$

where $a_+ = \max(0, a)$

Penalization – Lasso

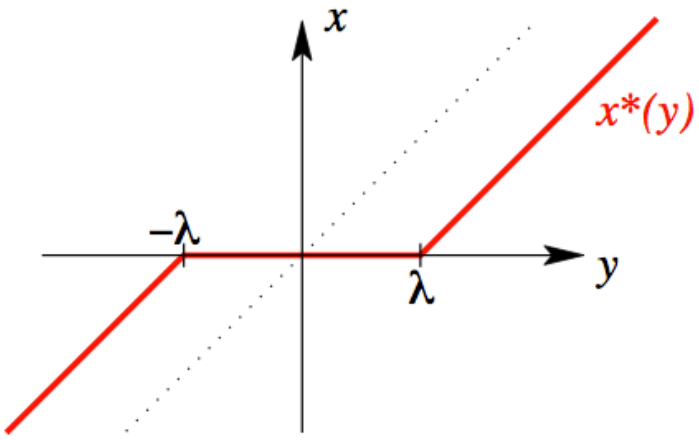
As a consequence, we have

$$a_* = \underset{a \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{2}\|a - b\|_2^2 + \lambda \|a\|_1 = S_\lambda(b)$$

where

$$S_\lambda(b) = \operatorname{sign}(b) \odot (|b| - \lambda)_+$$

is the **soft-thresholding** operator



Proximal operator

- For any g convex [lower semi-continuous] and any $y \in \mathbb{R}^d$, we define the **proximal operator**

$$\mathrm{prox}_g(y) = \underset{x \in \mathbb{R}^d}{\mathrm{argmin}} \left\{ \frac{1}{2} \|x - y\|_2^2 + g(x) \right\}$$

  (strongly convex problem $\Rightarrow$ unique minimum)
- We already proved that soft-thresholding is the proximal operator of the ℓ_1-norm

$$\mathrm{prox}_{\lambda \|\cdot\|_1}(y) = S_\lambda(y) = \mathrm{sign}(y) \odot (|y| - \lambda)+$$

Proximal operators and proximal algorithms are now **fundamental tools** for optimization in machine learning

Examples of proximal operators

- $g(x) = c$ for a constant c, $\text{prox}_g = Id$
- If C convex set, and

$$g(x) = \delta_C(x) = \begin{cases} 0 & \text{if } x \in C \\ +\infty & \text{if } x \notin C \end{cases}$$

then

$$\text{prox}_g = \text{proj}_C = \text{projection onto } C.$$

- If $g(x) = \langle b, x \rangle + c$, then

$$\text{prox}_{\lambda g}(x) = x - \lambda b$$

- If $g(x) = \frac{1}{2} x^\top A x + \langle b, x \rangle + c$ with A symetric positive, then

$$\text{prox}_{\lambda g}(x) = (I + \lambda A)^{-1}(x - \lambda b)$$

Examples of proximal operators

- If $g(x) = \frac{1}{2}\|x\|_2^2$ then

$$\mathrm{prox}_{\lambda g}(x) = \frac{1}{1+\lambda}x = \text{shrinkage operator}$$

- If $g(x) = -\log x$ then

$$\mathrm{prox}_{\lambda g}(x) = \frac{x + \sqrt{x^2 + 4\lambda}}{2}$$

- If $g(x) = \|x\|_2$ then

$$\mathrm{prox}_{\lambda g}(x) = \left(1 - \frac{\lambda}{\|x\|_2}\right)_+ x,$$

the block soft-thresholding operator

Examples of proximal operators

- If $g(x) = \|x\|_1 + \frac{\gamma}{2}\|x\|_2^2$ (elastic-net) where $\gamma > 0$, then

$$\mathrm{prox}_{\lambda g}(x) = \frac{1}{1+\lambda\gamma}\,\mathrm{prox}_{\lambda\|\cdot\|_1}(x)$$

- If $g(x) = \sum_{g\in G}\|x_g\|_2$ where G partition of $\{1,\ldots,d\}$,

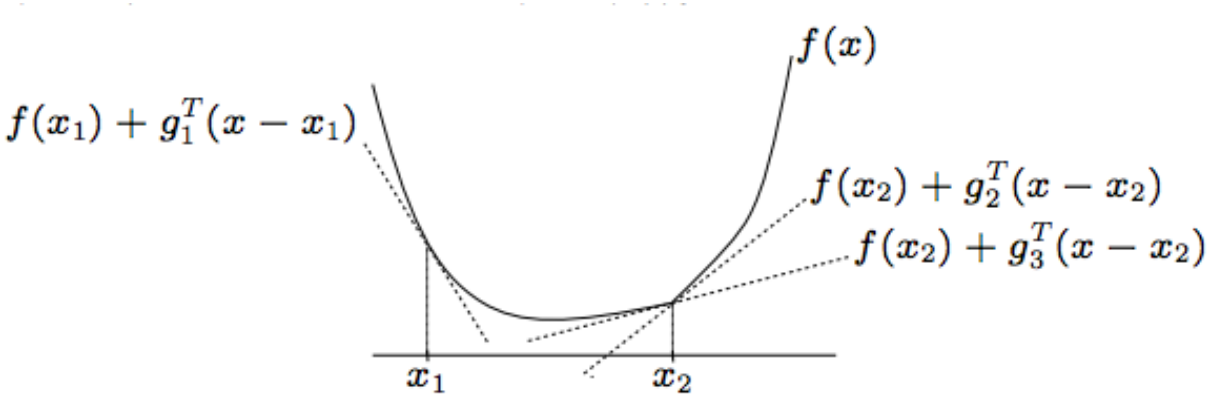
$$(\mathrm{prox}_{\lambda g}(x))_g = \left(1 - \frac{\lambda}{\|x_g\|_2}\right)_+ x_g,$$

for $g \in G$. Block soft-thresholding, used for group-Lasso

Subdifferential

The **subdifferential** of $f \in \Gamma^0$ at x is the set

$$\partial f(x) = \left\{ g \in \mathbb{R}^d : f(y) \geq \langle g, y - x \rangle + f(x) \text{ for all } y \in \mathbb{R}^d \right\}$$



- Each element is called a **subgradient**
- **Optimality criterion**

$$0 \in \partial f(x) \;\text{ iff }\; f(x) \leq f(y)\ \forall y$$

- If f is differentiable at x, then $\partial f(x) = \{\nabla f(x)\}$
- Example: $\partial |0| = [-1, 1]$

Smoothness

$f : \mathbb{R}^d \to [-\infty, +\infty]$ is

- f is L-**smooth** if it is continuously differentiable and if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \quad \text{for any } x, y \in \mathbb{R}^d.$$

  Equivalent to $H_f(x) \preceq LI_d$ for all x, where $H_f(x)$ Hessian at x when twice continuously differentiable [i.e. $LI_d - H_f(x)$ positive semi-definite]

- f is μ-strongly convex if $f(\cdot) - \frac{\mu}{2}\|\cdot\|_2^2$ is convex. Equivalent to

$$f(y) \geq f(x) + \langle g, y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2$$

  for $g \in \partial f(x)$. Equivalent to $H_f(x) \succeq \mu I_d$ when twice differentiable.

Optimality criterion

Optimality criterion: $\theta_* \in \text{argmin}_{\theta \in \mathbb{R}^d} \{f(\theta) + g(\theta)\}$ iff

$$-\nabla f(\theta_*) \in \partial g(\theta_*)$$

namely

$$-\frac{1}{\lambda} \nabla R_n(\theta_*) \in \partial g(\theta_*)$$

For the Lasso

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2n} \| Y - X\theta \|_2^2 + \lambda \|\theta\|_1 \right\}$$

this optimality criterion is

$$\begin{cases} \frac{1}{n} |X_j^\top (Y - X\hat{\theta})| \leq \lambda & \text{if } \hat{\theta}_j = 0 \\ \frac{1}{n} X_j^\top (Y - X\hat{\theta}) = \lambda \, \text{sign}(\hat{\theta}_j) & \text{if } \hat{\theta}_j \neq 0 \end{cases}$$

for any $j = 1, \ldots, d$, where X_j is the j-th conumn of X.

The general problem we want to solve

How to solve

$$\hat{\theta} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, \theta \rangle) + \lambda \, \mathrm{pen}(\theta) \right\} \quad ???$$

Put for short

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \langle x_i, \theta \rangle) \;\text{ and }\; g(\theta) = \lambda \, \mathrm{pen}(\theta)$$

Assume that

- f is convex and L-smooth
- g is convex and continuous, but possibly non-smooth (for instance ℓ_1 penalization)
- g is **prox-capable**: not hard to compute its proximal operator

Examples

Smoothness of f:

- Least-squares:

$$\nabla f(\theta) = \frac{1}{n} X^\top (X\theta - Y), \quad L = \frac{\|X^\top X\|_{\mathrm{op}}}{n}$$

- Logistic loss:

$$\nabla f(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{1 + e^{y_i \langle x_i, \theta \rangle}} x_i, \quad L = \frac{\max_{i=1,\dots,n} \|X_i\|_2^2}{4n}$$

Prox-capability of g:

- we gave the explicit prox for many penalizations above

Gradient descent

Now how do I minimize $f + g$?

- Key point: the **descent lemma**. If f convex and L-smooth, then for any $L' \geq L$:

$$f(\theta') \leq f(\theta) + \langle \nabla f(\theta), \theta' - \theta \rangle + \frac{L'}{2} \|\theta' - \theta\|_2^2$$

  for any $\theta, \theta' \in \mathbb{R}^d$

- At iteration k, the current point is θ^k. I use the descent lemma:

$$f(\theta) \leq f(\theta^k) + \langle \nabla f(\theta^k), \theta - \theta^k \rangle + \frac{L'}{2} \|\theta - \theta^k\|_2^2.$$

Gradient descent

- Remark that

$$\begin{aligned}&\operatorname*{argmin}_{\theta\in\mathbb{R}^d}\left\{f(\theta^k)+\langle\nabla f(\theta^k),\theta-\theta^k\rangle+\frac{L'}{2}\|\theta-\theta^k\|_2^2\right\}\\&=\operatorname*{argmin}_{\theta\in\mathbb{R}^d}\left\|\theta-\left(\theta^k-\frac{1}{L'}\nabla f(\theta^k)\right)\right\|_2^2\end{aligned}$$

- Hence, choose

$$\theta^{k+1}=\theta^k-\frac{1}{L'}\nabla f(\theta^k)$$

  This is the basic **gradient descent** algorithm [cf previous lecture]

- Gradient descent is based on a **majoration-minimization** principle, with a quadratic majorant given by the descent lemma
- But we forgot about g...

ISTA

Let's put back g:

$$f(\theta) + g(\theta) \leq f(\theta^k) + \langle \nabla f(\theta^k), \theta - \theta^k \rangle + \frac{L'}{2}\|\theta - \theta^k\|_2^2 + g(\theta)$$

and again

$$\begin{aligned}
&\operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ f(\theta^k) + \langle \nabla f(\theta^k), \theta - \theta^k \rangle + \frac{L'}{2}\|\theta - \theta^k\|_2^2 + g(\theta) \right\} \\
&= \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{L'}{2} \left\| \theta - \left(\theta^k - \frac{1}{L'} \nabla f(\theta^k) \right) \right\|_2^2 + g(\theta) \right\} \\
&= \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{2} \left\| \theta - \left(\theta^k - \frac{1}{L'} \nabla f(\theta^k) \right) \right\|_2^2 + \frac{1}{L'} g(\theta) \right\} \\
&= \operatorname{prox}_{g/L'} \left(\theta^k - \frac{1}{L'} \nabla f(\theta^k) \right)
\end{aligned}$$

The prox operator naturally appears because of the descent lemma

ISTA

Proximal gradient descent algorithm [also called ISTA]

- **Input**: starting point θ^0, Lipschitz constant $L > 0$ for ∇f
- For $k = 1, 2, \ldots$ until *converged* do
  - $\theta^k = \text{prox}_{g/L}\left(\theta^{k-1} - \frac{1}{L}\nabla f(\theta^{k-1})\right)$
- **Return** last θ^k

Also called **Forward-Backward splitting**. For Lasso with least-squares loss, iteration is

$$\theta^k = S_{\lambda/L}\Big(\theta^{k-1} - \frac{1}{L}(X^\top X\theta^{k-1} - X^\top Y)\Big),$$

where S_λ is the soft-thresholding operator

ISTA

- Put for short $F = f + g$,
- Take any $\theta^* \in \text{argmin}_{\theta \in \mathbb{R}^d} F(\theta)$

Theorem (Beck Teboulle (2009))

If the sequence $\{\theta^k\}$ is generated by ISTA, then

$$F(\theta^k) - F(\theta^*) \leq \frac{L\|\theta^0 - \theta^*\|_2^2}{2k}$$

- Convergence rate is $O(1/k)$
- Is it possible to improve the $O(1/k)$ rate?

FISTA

Yes! Using **Accelerated proximal gradient descent** (called FISTA, Nesterov 83, 04, Beck Teboule 09)

- Idea: to find θ^{k+1}, use an interpolation between θ^k and θ^{k-1}

Accelerated proximal gradient descent algorithm [FISTA]

- **Input**: starting points $z^1 = \theta^0$, Lipschitz constant $L > 0$ for ∇f, $t_1 = 1$
- For $k = 1, 2, \ldots$ until *converged* do
  - $\theta^k = \text{prox}_{g/L}(z^k - \frac{1}{L}\nabla f(z^k))$
  - $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
  - $z^{k+1} = \theta_k + \frac{t_k - 1}{t_{k+1}}(\theta^k - \theta^{k-1})$
- **Return** last θ^k

FISTA

Theorem (Beck Teboulle (2009))

If the sequence $\{\theta^k\}$ is generated by FISTA, then

$$F(\theta^k) - F(\theta^*) \leq \frac{2L\|\theta^0 - \theta^*\|_2^2}{(k+1)^2}$$

- Convergence rate is $O(1/k^2)$
- Is $O(1/k^2)$ the optimal rate in general?

FISTA

Yes. Put $g = 0$

Theorem (Nesterov)

For any optimization procedure satisfying

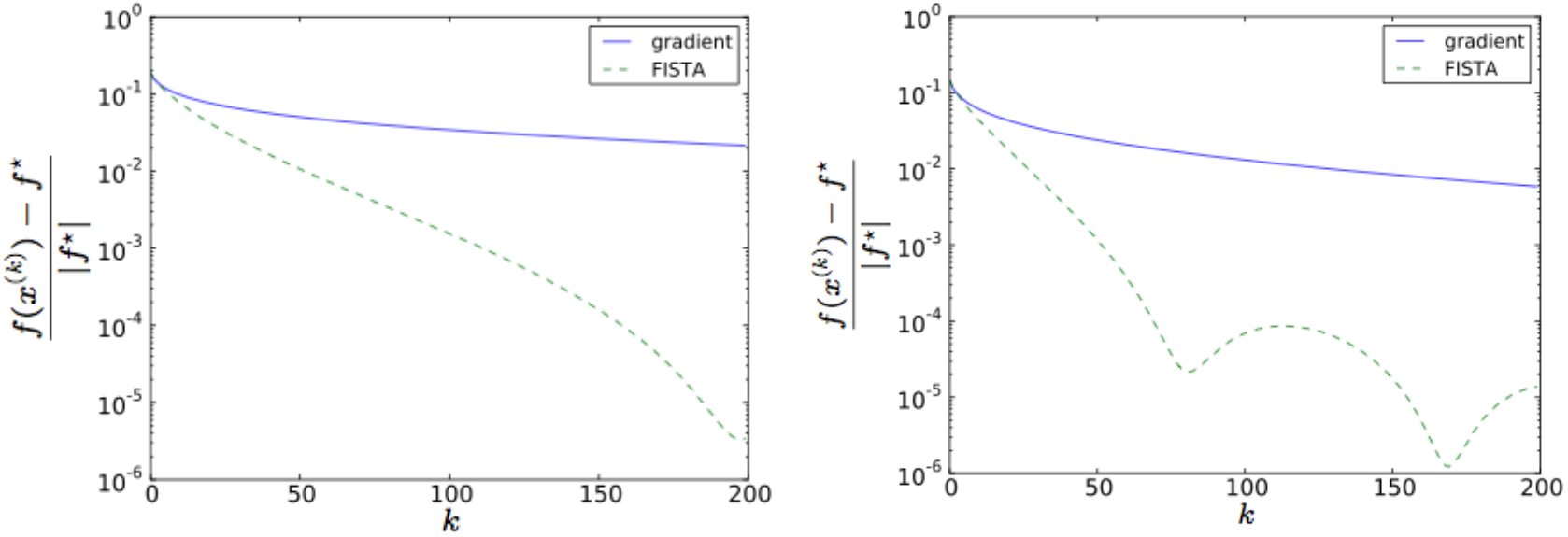
$$\theta^{k+1} \in \theta^1 + \mathrm{span}(\nabla f(\theta^1), \ldots, \nabla f(\theta^k)),$$

there is a function f on $\mathbb{R}^d$ convex and L-smooth such that

$$\min_{1 \le j \le k} f(\theta^j) - f(\theta^*) \ge \frac{3L}{32} \frac{\|\theta^1 - \theta^*\|_2^2}{(k+1)^2}$$

for any $1 \le k \le (d-1)/2$.

Comparison of ISTA and FISTA



FISTA is **not** a descent algorithm, while ISTA is

Backtracking linesearch

What if I don't know $L > 0$?

- $\|X^\top X\|_{\mathrm{op}}$ can be long to compute
- Letting L evolve along iterations k generally improve convergence speed

Backtracking linesearch. Idea:

- Start from a very small lipschitz constant L
- Between iteration k and $k+1$, choose the smallest L satisfying the lemma descent at z^k

Backtracking linesearch

At iteration k of FISTA, we have z^k and a constant L_k

1. Put $L \leftarrow L_k$
2. Do an iteration

$$\theta \leftarrow \text{prox}_{g/L}\left(z^k - \frac{1}{L}\nabla f(z^k)\right)$$

3. Check it this step satisfies the descent lemma at z^k:

$$f(\theta) + g(\theta) \leq f(z^k) + \langle \nabla f(z^k), \theta - z^k \rangle + \frac{L}{2}\|\theta - z^k\|_2^2 + g(\theta)$$

4. If yes, then $\theta^{k+1} \leftarrow \theta$ and $L_{k+1} \leftarrow L$ and continue FISTA
5. It not, then put $L \leftarrow 2L$ (say), and go back to point 2

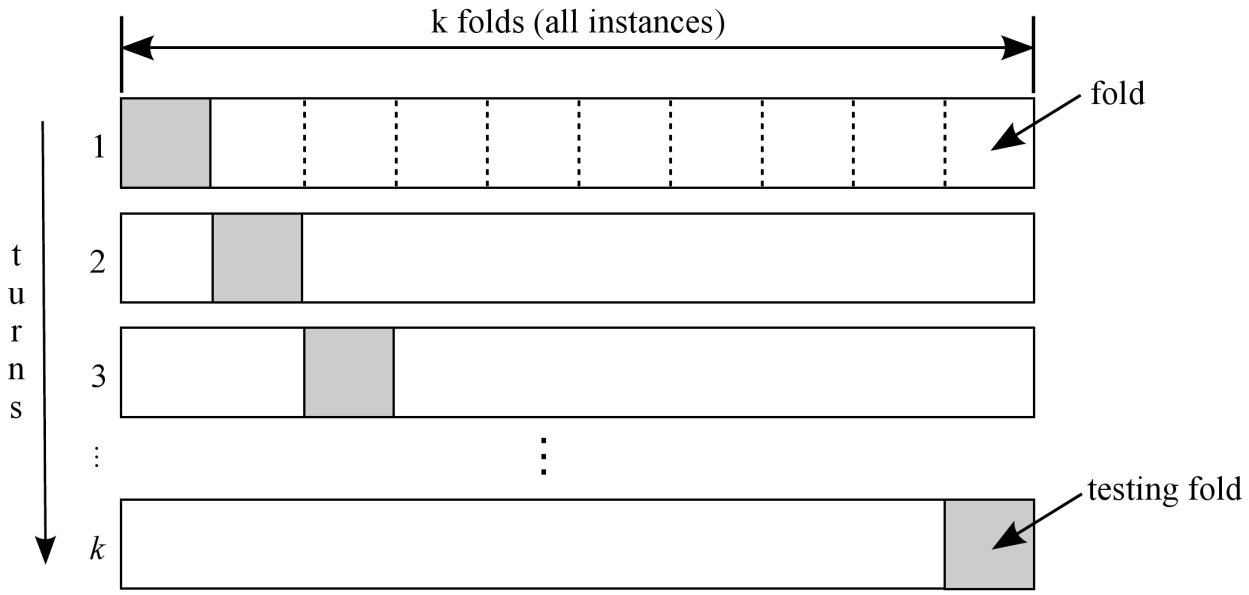
Sequence L_k is non-decreasing: between iteration k and $k+1$, a tweak is to *decrease* it a little bit to have (much) faster convergence

Cross-validation

- **Generalization** is one the most important goal of machine learning. A trained classifier has to be "generalizable", namely it can be applied in another context than the one of the training dataset, without **overfitting**
- This can be achieved using **cross-validation**
- There is **no machine learning** without cross-validation at some point!
- **We have to choose a penalization parameter λ 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 $v = 1, \ldots, V$



Cross-validation

Put

For each $v = 1, \ldots, V$

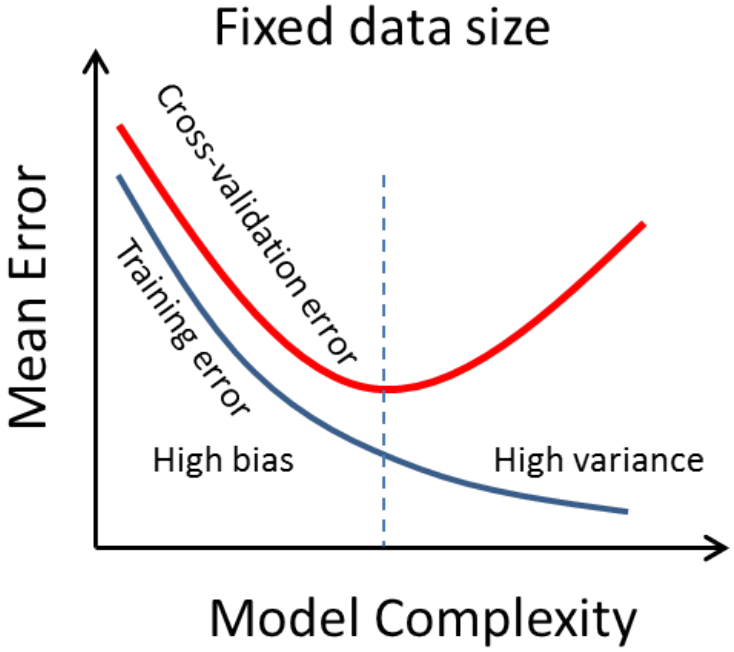
- Put $D_{v,\texttt{train}} = \cup_{v' \neq v} I_{v'}$ and $D_{v,\texttt{test}} = I_v$
- Find

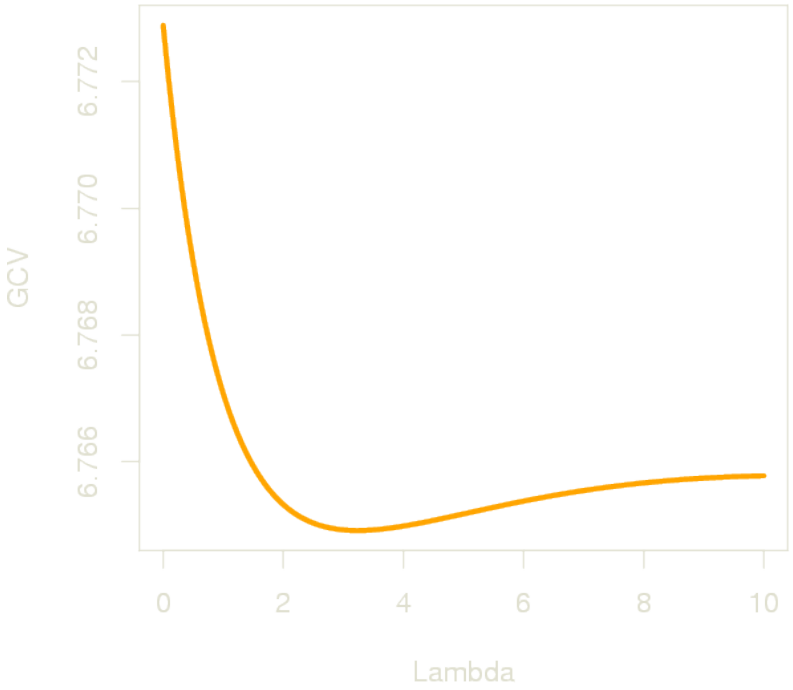
$$\hat{\theta}_{v,\lambda} \in \operatorname*{argmin}_{\theta} \left\{ \frac{1}{|D_{v,\texttt{train}}|} \sum_{i \in D_{v,\texttt{train}}} \ell(y_i, \langle X_i, \theta \rangle) + \lambda \, \mathrm{pen}(\theta) \right\}$$

Take

$$\hat{\lambda} \in \operatorname*{argmin}_{\lambda} \sum_{v=1}^{V} \sum_{i \in D_{v,\texttt{test}}} \ell(y_i, \langle X_i, \hat{\theta}_{v,\lambda} \rangle)$$

Cross-validation





- Training error:

$$\lambda \mapsto \sum_{v=1}^{V} \sum_{i \in D_{v,\texttt{train}}} \ell(y_i, \langle X_i, \hat{\theta}_{v,\lambda} \rangle)$$

- Testing, validation or cross-validation error:

$$\lambda \mapsto \sum_{v=1}^{V} \sum_{i \in D_{v,\texttt{test}}} \ell(y_i, \langle X_i, \hat{\theta}_{v,\lambda} \rangle)$$

Classification scores

- Now I've trained a logistic classifier (or any other classifier), I have an estimation $\hat{\theta}$ of θ
- Or I'm training it but I want to test it as well along my cross-validation loop
- On testing samples (x_i, y_i), compute (if using logistic classifier)

$$\hat{p}_{i,0} = \mathbb{P}[Y = 0 | X = x_i] = \frac{1}{1 + e^{\langle x_i, \hat{\theta} \rangle}},$$
$$\hat{p}_{i,1} = \mathbb{P}[Y = 1 | X = x_i] = \frac{1}{1 + e^{-\langle x_i, \hat{\theta} \rangle}}$$

- Predict the label using the MAP rule (Maximum A Posteriori)

$$\hat{y}_i = \arg\max_{y=0,1} \hat{p}_{i,y}$$

- Test it by comparing prediction $\hat{y}_i$ and ground truth y_i

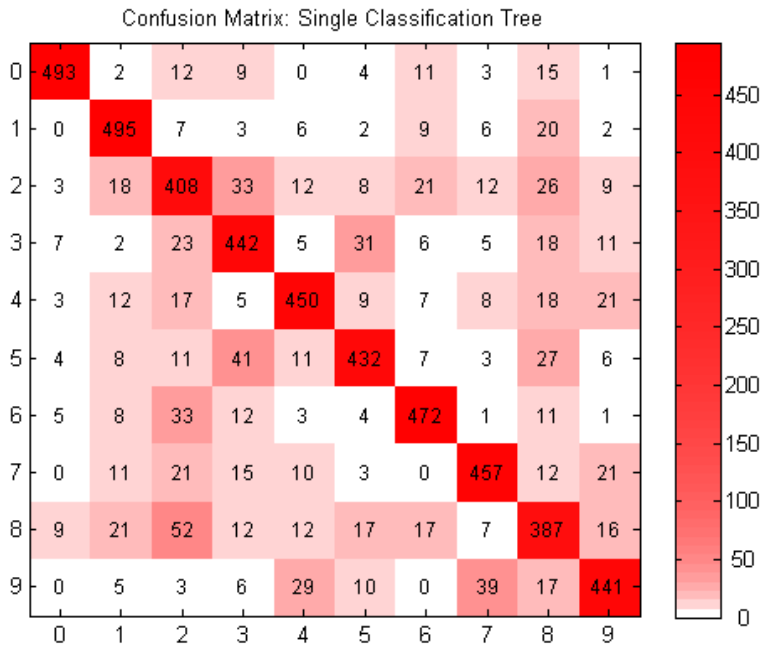
Classification scores

Standard error metrics in classification

- Precision, Recall, F-Score, AUC

For each i: true label y_i, predicted label $\hat{y}_i$

Confusion matrix



Classification scores

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

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

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

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \qquad \text{F-Score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Classification scores

- In multiclass classification (more than 2 labels), can compute precision and recall for each label
- Recall = Sensitivity
- False-Discovery Rate FDR $= 1 -$ Precision
- Many other metrics...

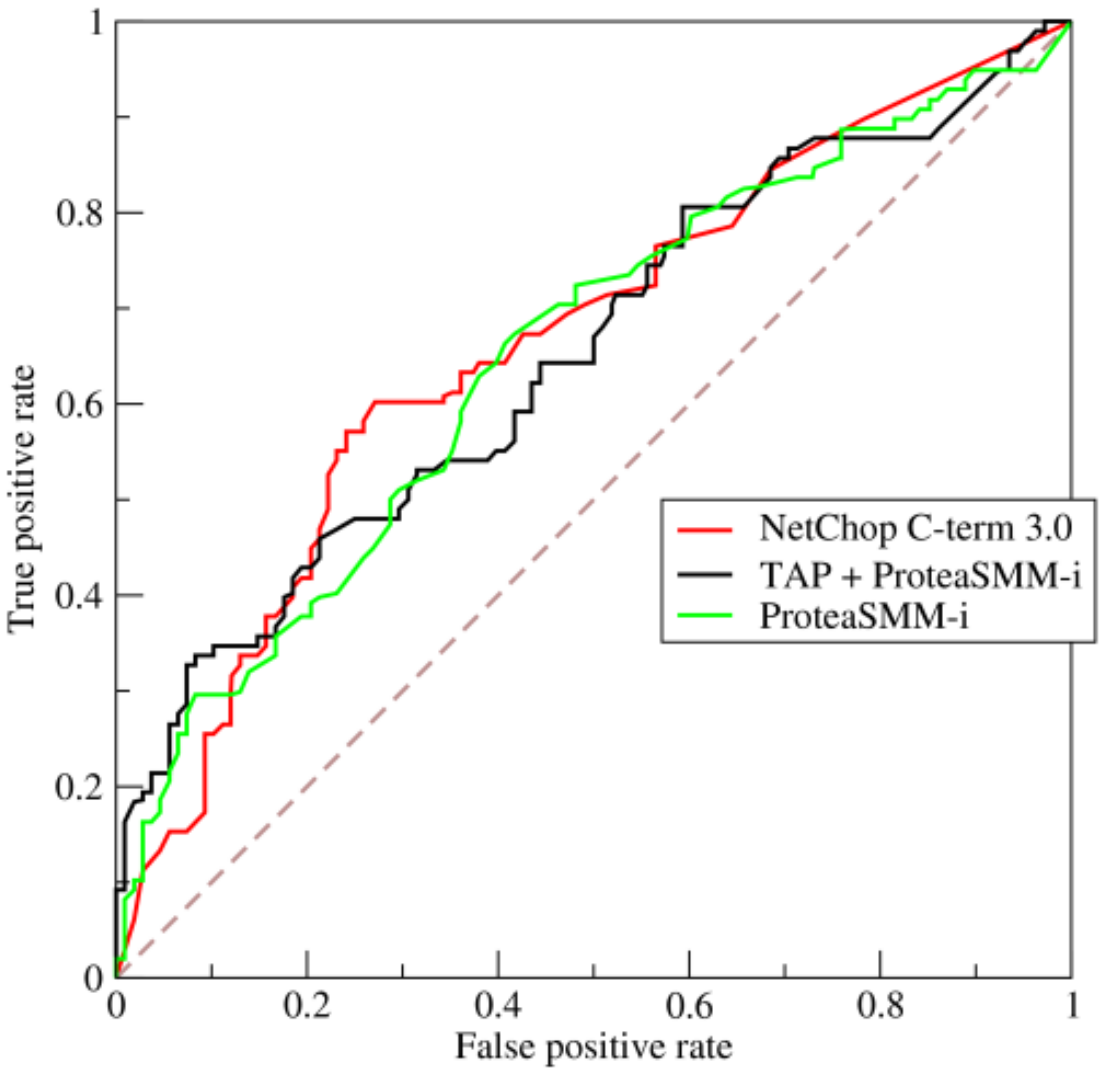
ROC Curve (Receiver Operating Characteristic)

- For binary classification
- True positive Rate $\text{TPR} = \frac{\text{TP}}{\#(\text{real P})} = \frac{\text{TP}}{\text{TP}+\text{FN}}$
- False positive Rate $\text{FPR} = \frac{\text{FP}}{\#(\text{real N})} = \frac{\text{FP}}{\text{FP}+\text{TN}}$
- x-axis: FPR, y-axis: TPR
- 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

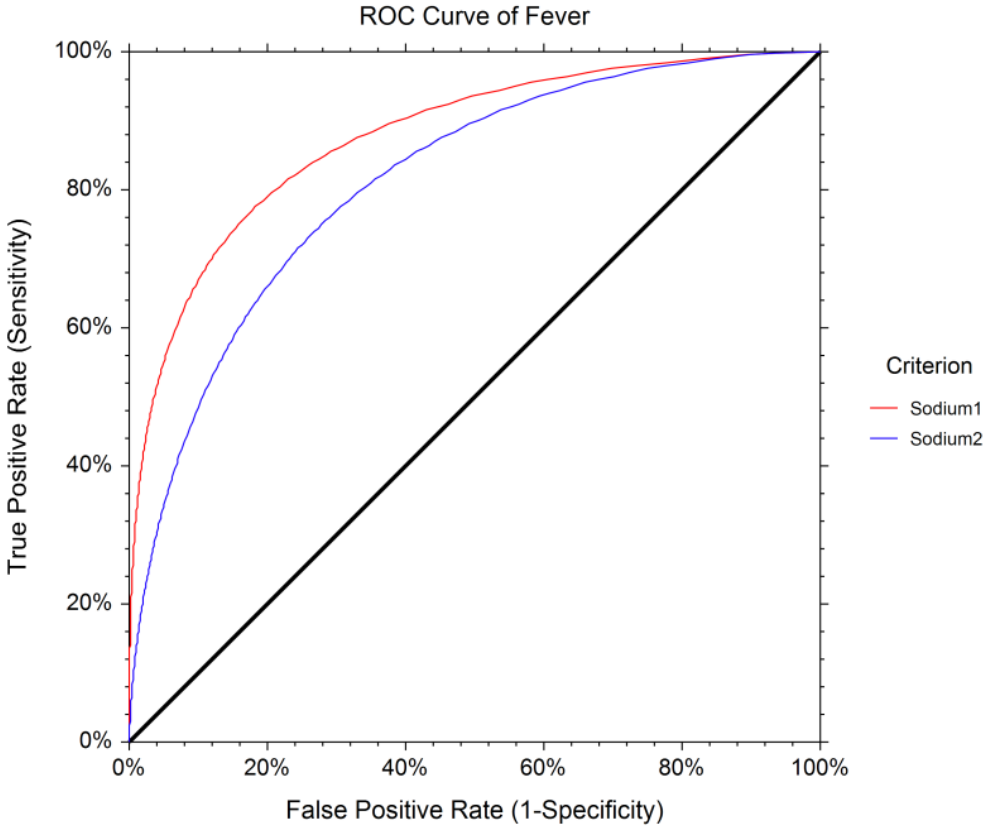
$$\hat{y}_i = \mathbf{1}_{\hat{p}_{i,0} \geq t}$$

- AUC score is the Area Under the ROC Curve

Classification scores



Classification scores



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

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

Class unbalancing

The logistic goodness-of-fit is

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

Just replace it by

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

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

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 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

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

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

- Then

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

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.  ]]
```

Agenda for tomorrow

- Stochastic Gradient Descent and beyond
- Collaborative Filtering - Matrix Completion