House of Finance Days 2016
Machine Learning and Finance

Introduction to machine learning, application to Hawkes processes

S. Gaïffas

Stéphane Gaïffas

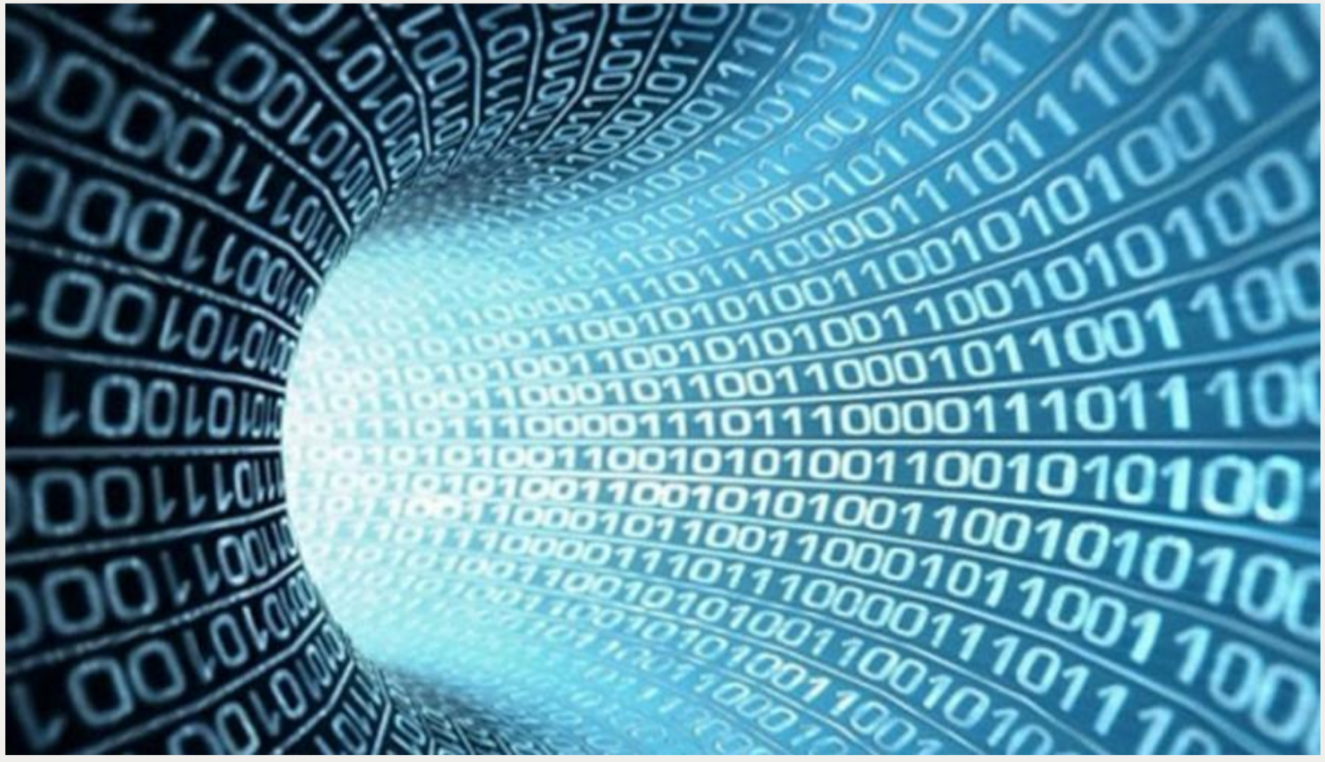
- Professeur Chargé de Cours
- CMAP, Ecole polytechnique
- `http://www.cmap.polytechnique.fr/~gaiffas/`
- `stephane.gaiffas@polytechnique.edu`
- Enseignement + Recherche
- *mots-clés*: Machine learning, Statistique, Big-data

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

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

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

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.

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

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.

Add to Portfolio

International Business Machines Corporation

Go to your Portfolio »

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

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

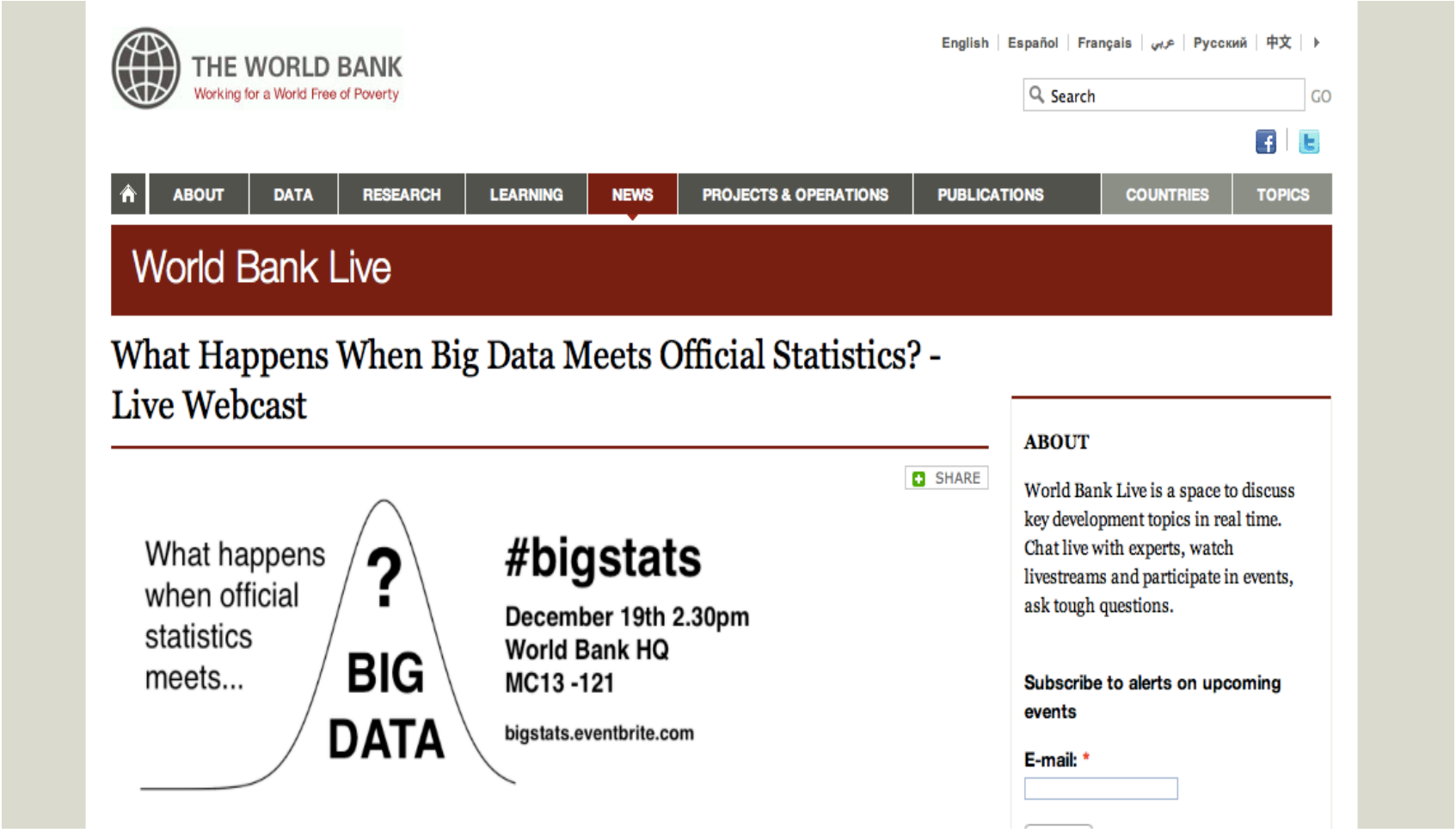
2. CRITIC'S NOTEBOOK
The Rainbow That Follows 'Jersey Shore'

3. TAKING NOTE
Opinion Report: Tax Reform

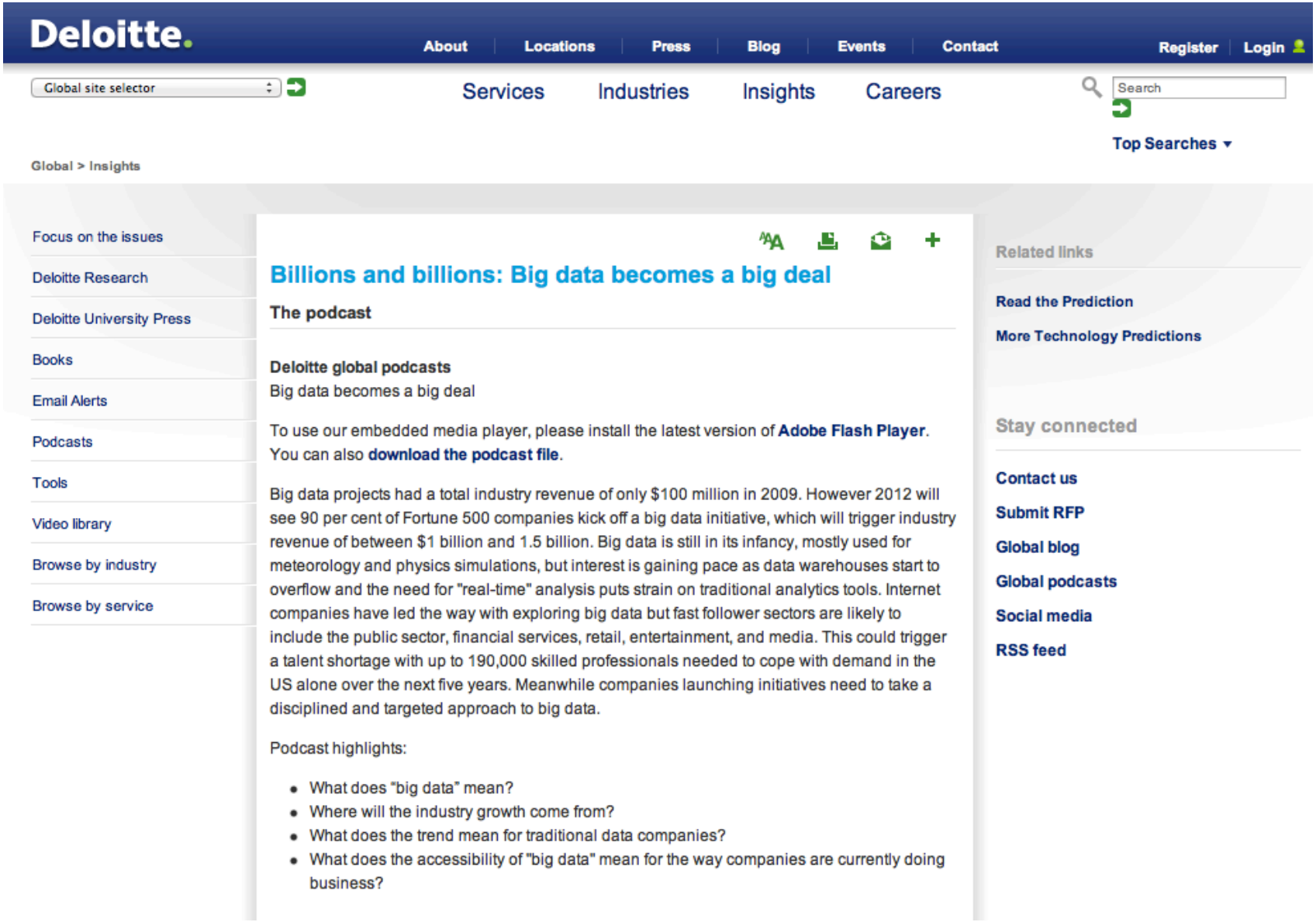
4. THE LEARNING NETWORK
Fill-In | Trendy Spot Urges Tourists to Ride In and Spend, 'Gangnam Style'

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.

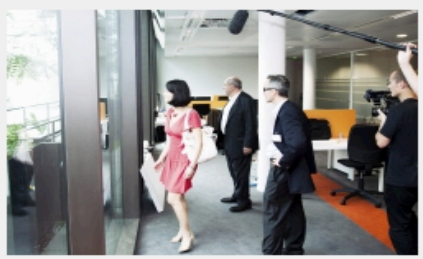
Criteo inaugure à Paris l'un des plus gros pôles européens de R&D dédiés à la publicité prédictive, CriteoLabs. Sur 10.000 m2, ce nouveau centre a vocation à accueillir 300 ingénieurs et à permettre ainsi à Criteo de garantir son avancée technologique sur ses 30 marchés d'exportation, des Etats-Unis, 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 :

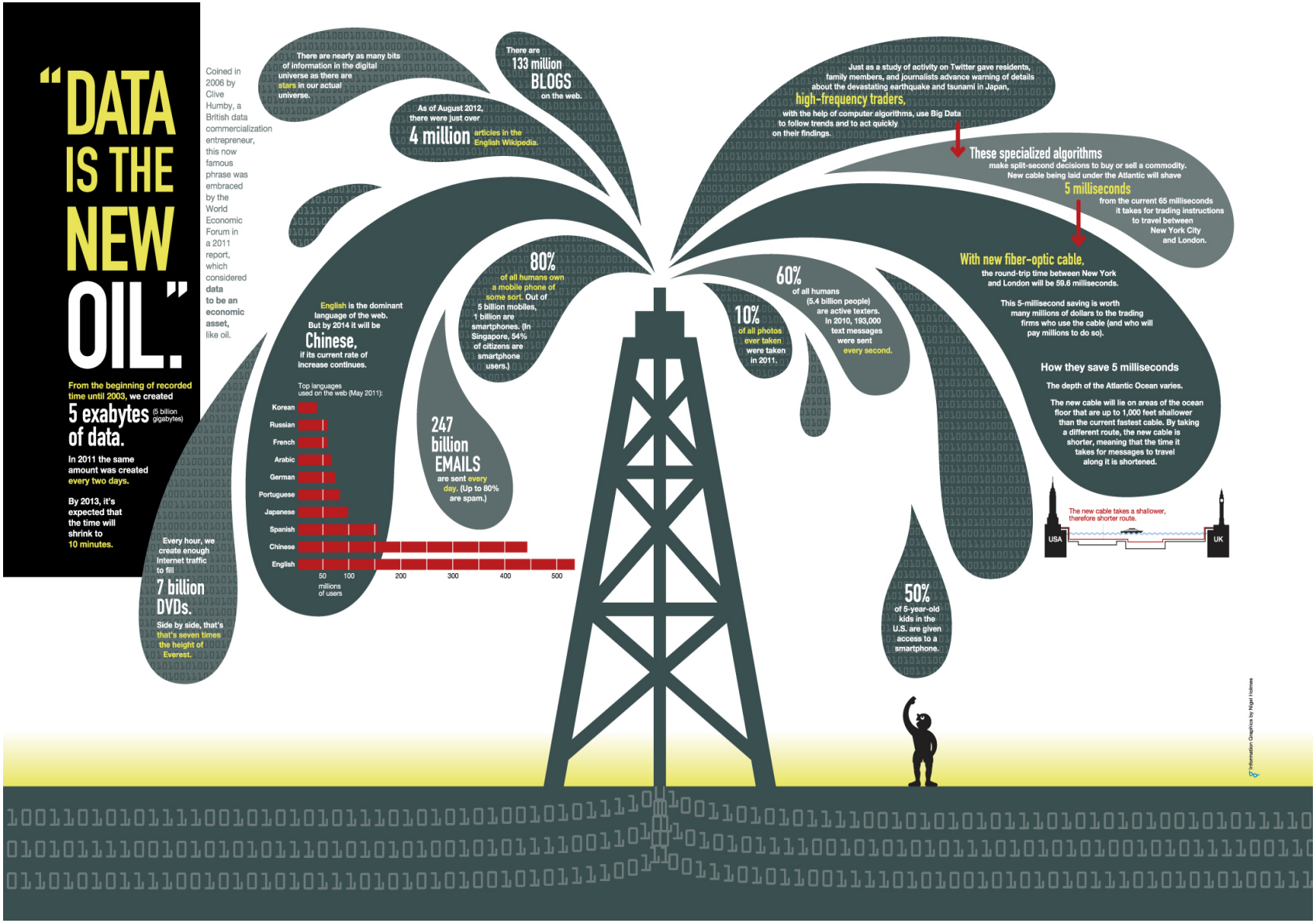
- 600 salariés présents dans 15 bureaux dans le monde
- 2 000 annonceurs, parmi les plus importants e-commerçants mondiaux tels que Dell, Macy's, John Lewis, Marks & Spencers, Zalando, La Redoute, Les 3 Suisses, etc.
- 4 000 éditeurs
- Plus de 200 millions de dollars de CA en 2011

Jean-Baptiste Rudelle, CEO et Pascal Gauthier COO

Arrivée de Fleur Pellerin



Data is the new oil?



Web search

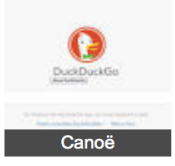
Google moteur de recherche

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**



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**"

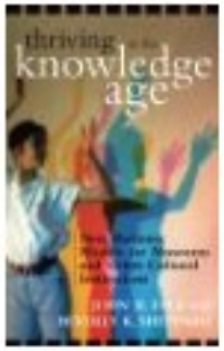
Moteur de recherche — Wikipédia
fr.wikipedia.org/wiki/**Moteur_de_recherche**
Un **moteur de recherche** est une application web permettant de retrouver des ressources (pages web, articles de forums Usenet, images, vidéo, fichiers, etc.) ...

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

Product or friend recommendation

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...

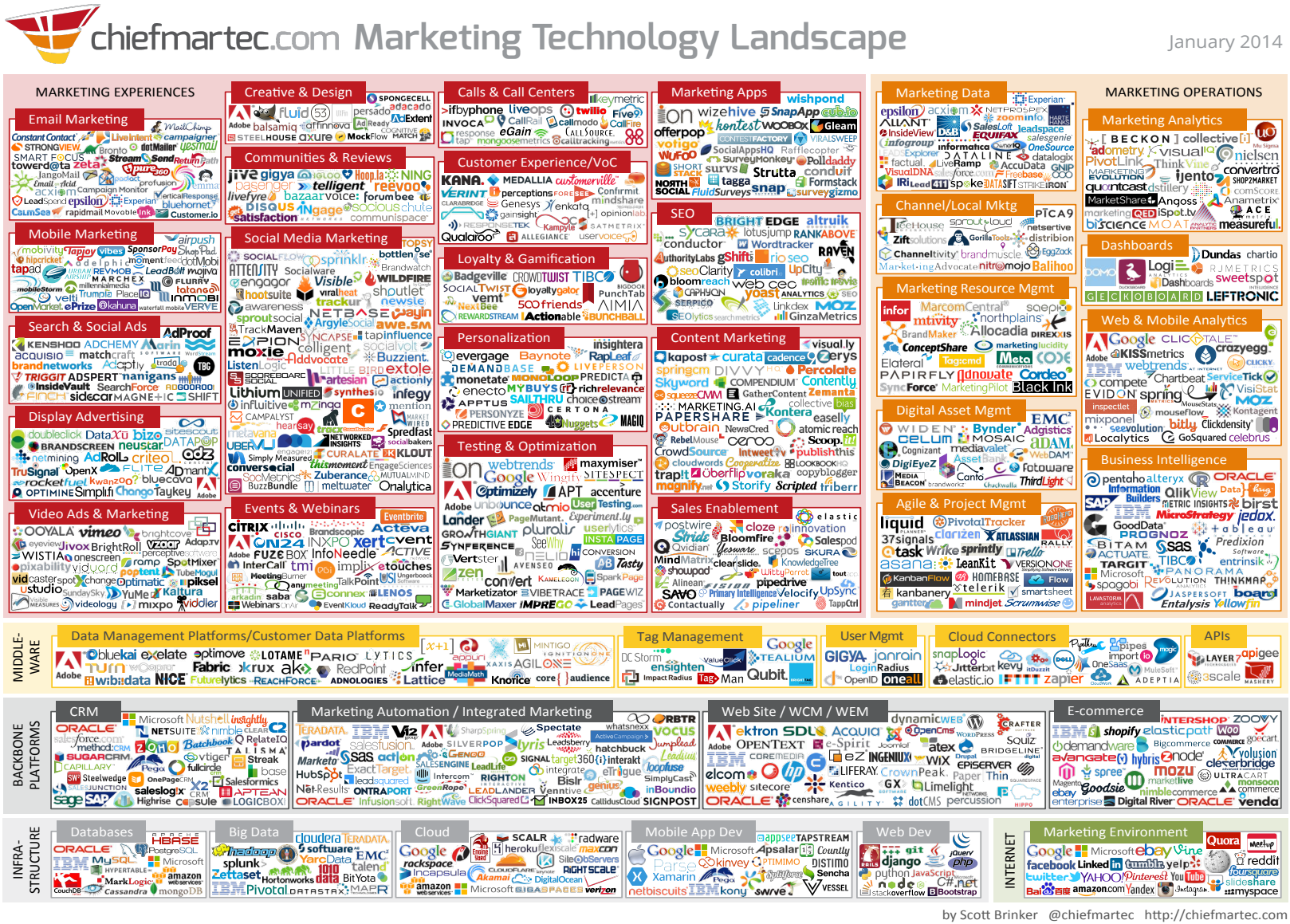
Display advertising

Terrorism


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-ato-machi
Nanjing
Quarry Bay
China
Ranked #8
Total Attacks
2,460
Looped Attacks
77
Unique Attacks
32
Distinct Sources
82

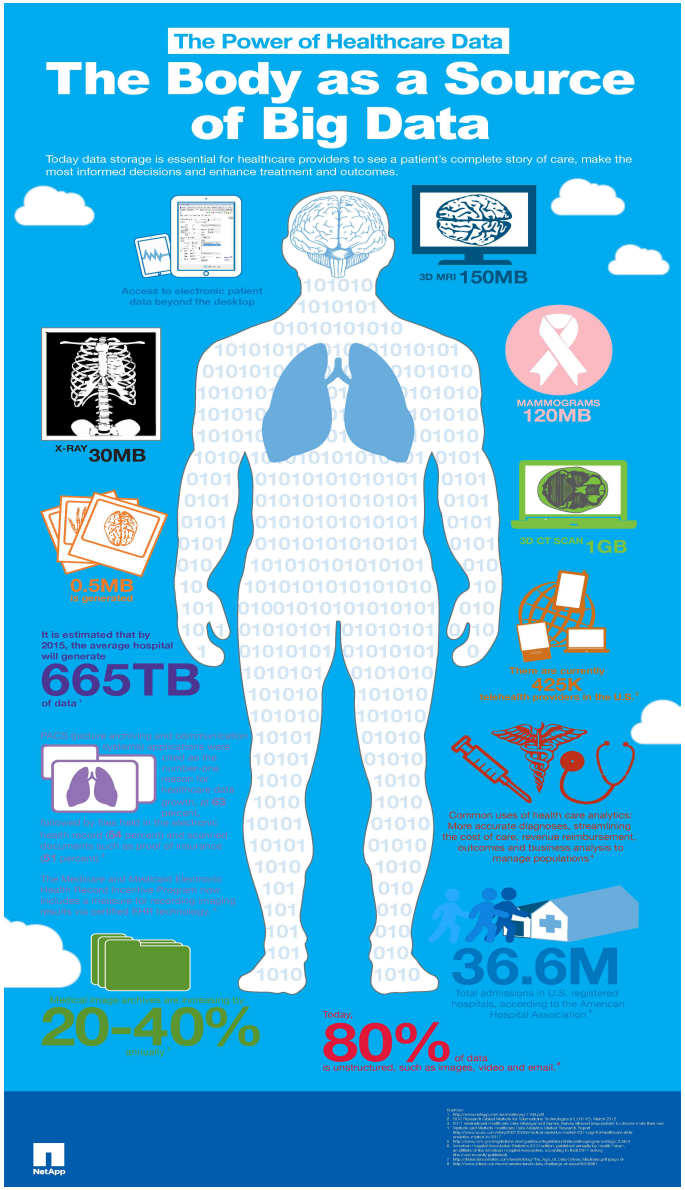
Crime prevention

Marketing

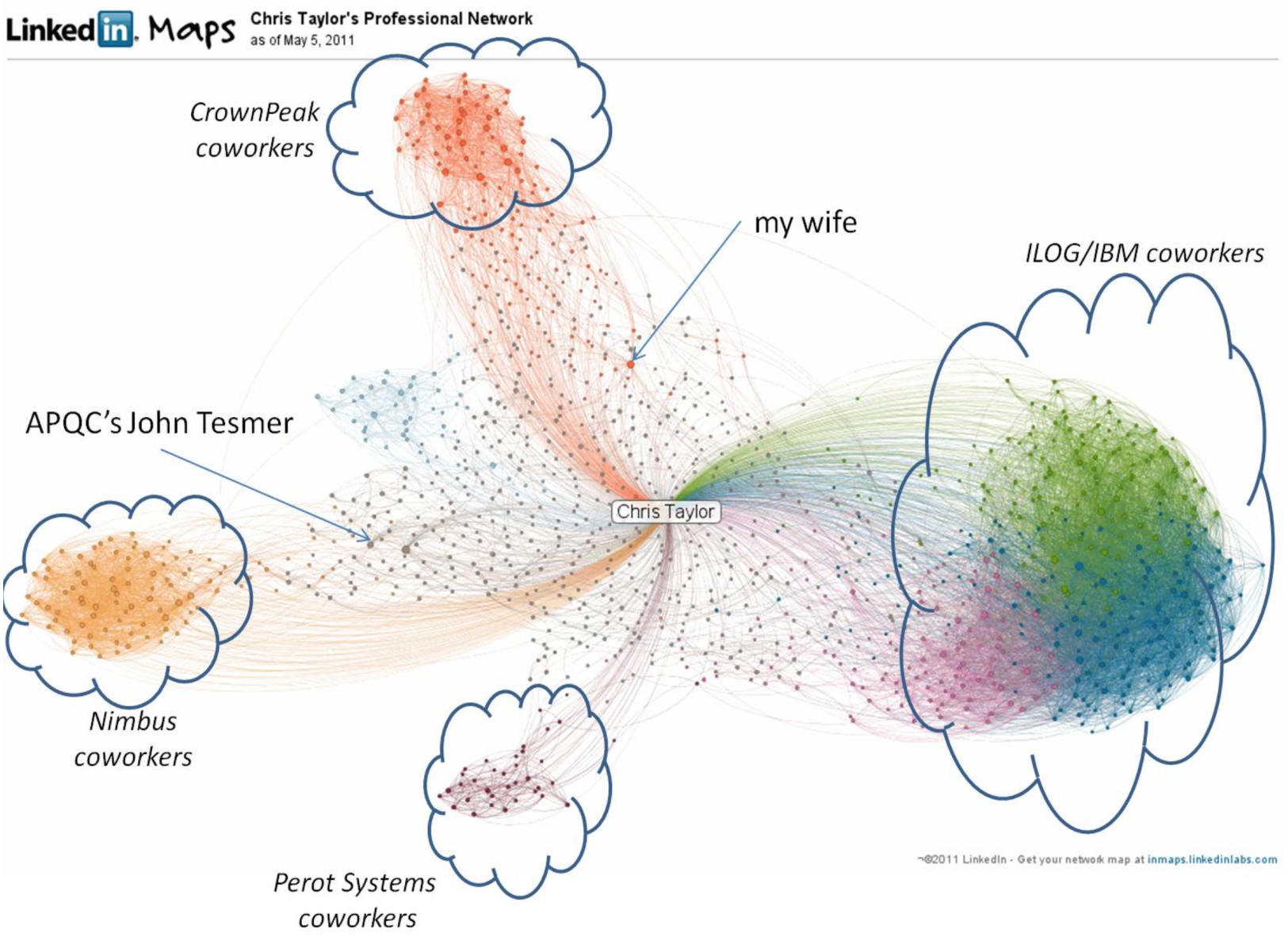


by Scott Brinker @chiefmartec http://chiefmartec.com

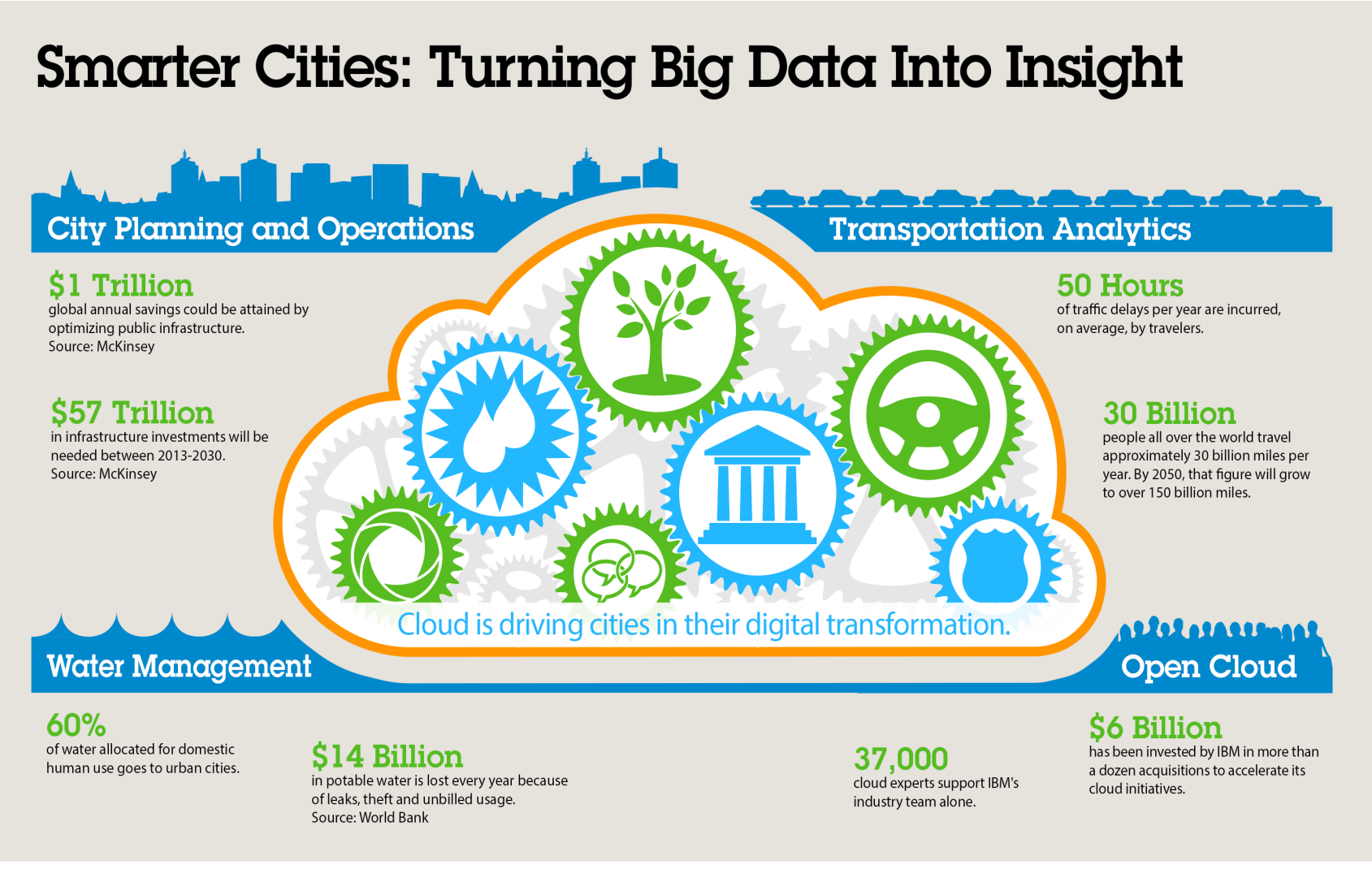
Health


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 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

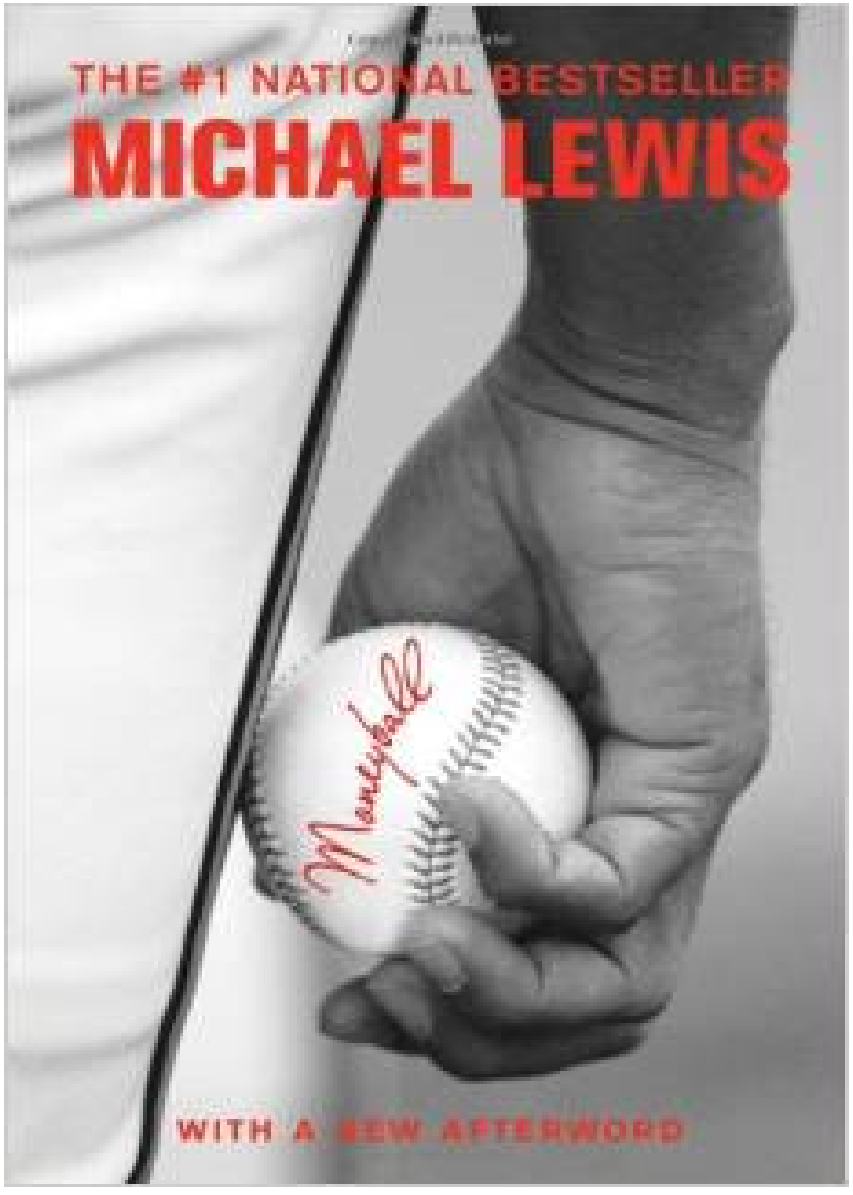
Social networks analytics



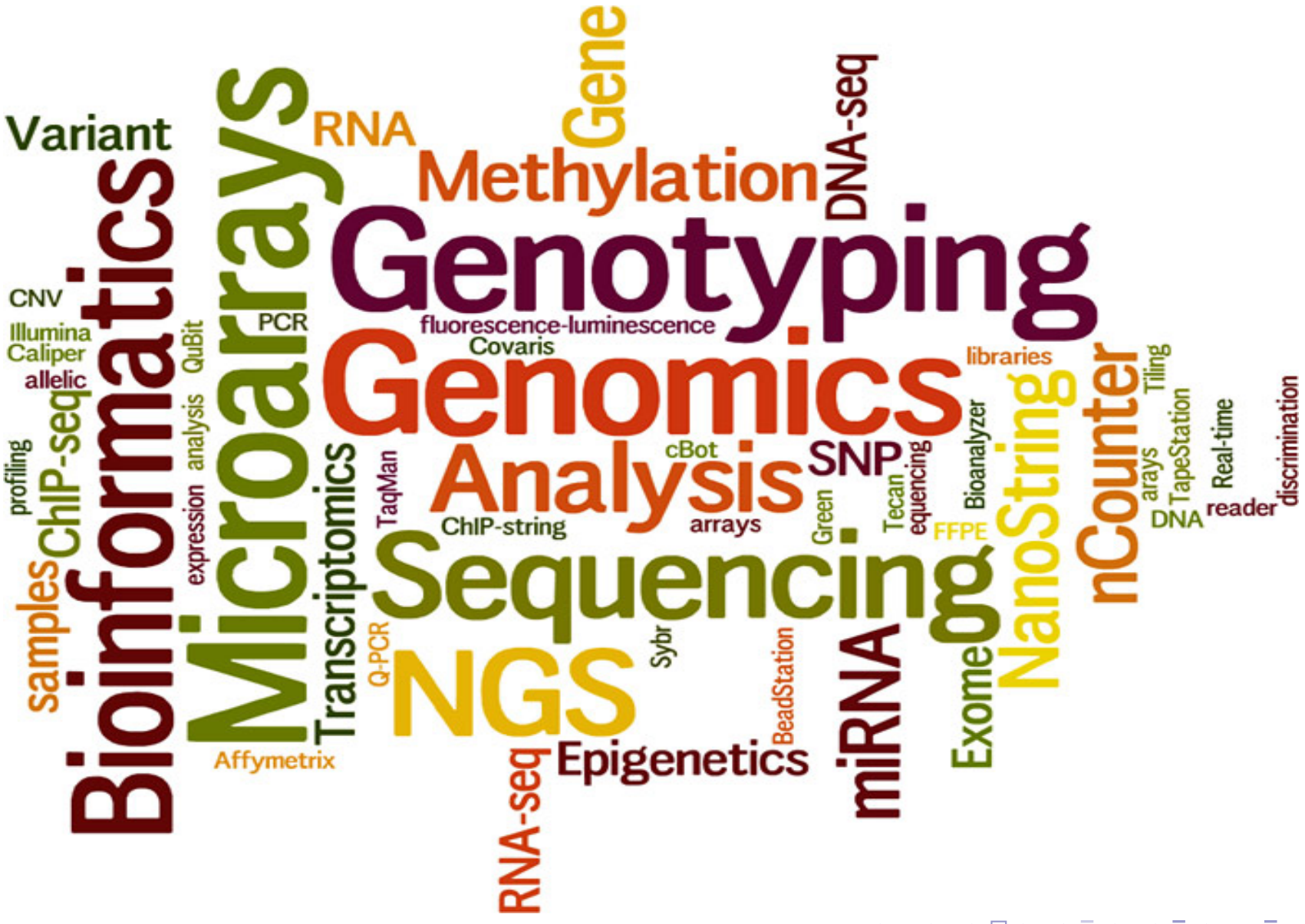
Energy (smart grids)



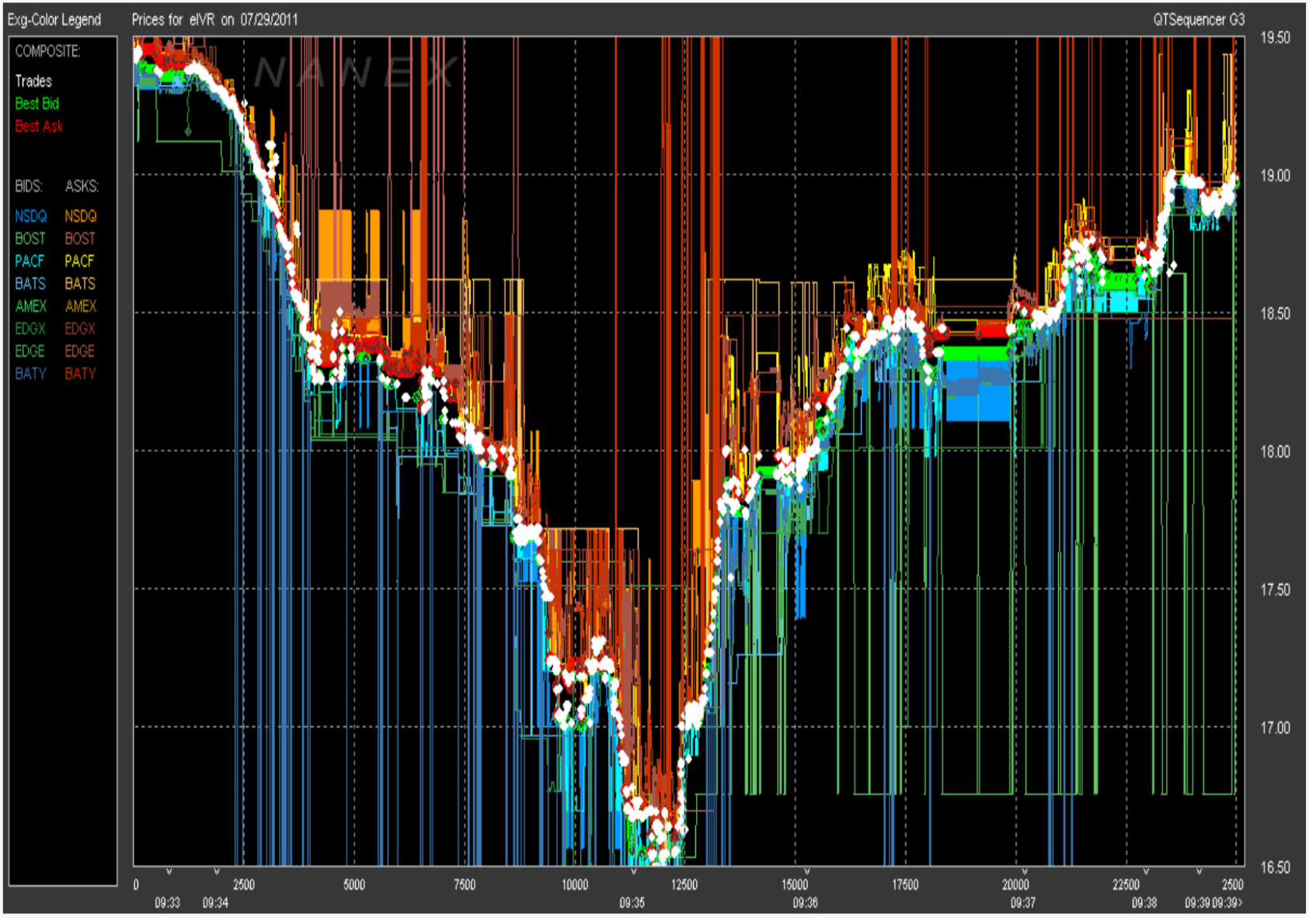
Sports


THE #1 NATIONAL BESTSELLER
MICHAEL LEWIS
Moneyball
WITH A NEW AFTERWORD

Genomics


Variant
RNA
Gene
DNA-seq
Methylation
Genotyping
Microarrays
Bioinformatics
CNV
PCR
fluorescence-luminescence
Illumina
Caliper
QuBit
Covaris
allelic
Genomics
libraries
Tiling
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
nCounter
NanoString
Sequencing
samples
Q-PCR
Sybr
NGS
BeadStation
Exome
miRNA
Affymetrix
RNA-seq
Epigenetics

High frequency trading



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 companies such as Google, Amazon, Facebook, Apple (GAFA) but also banking, marketing, pharmaceuticals, insurance, telecoms, personalized medicine, bioinformatics, etc.

Big Data is (quite) Easy

Example of *off the shelves* solution


```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

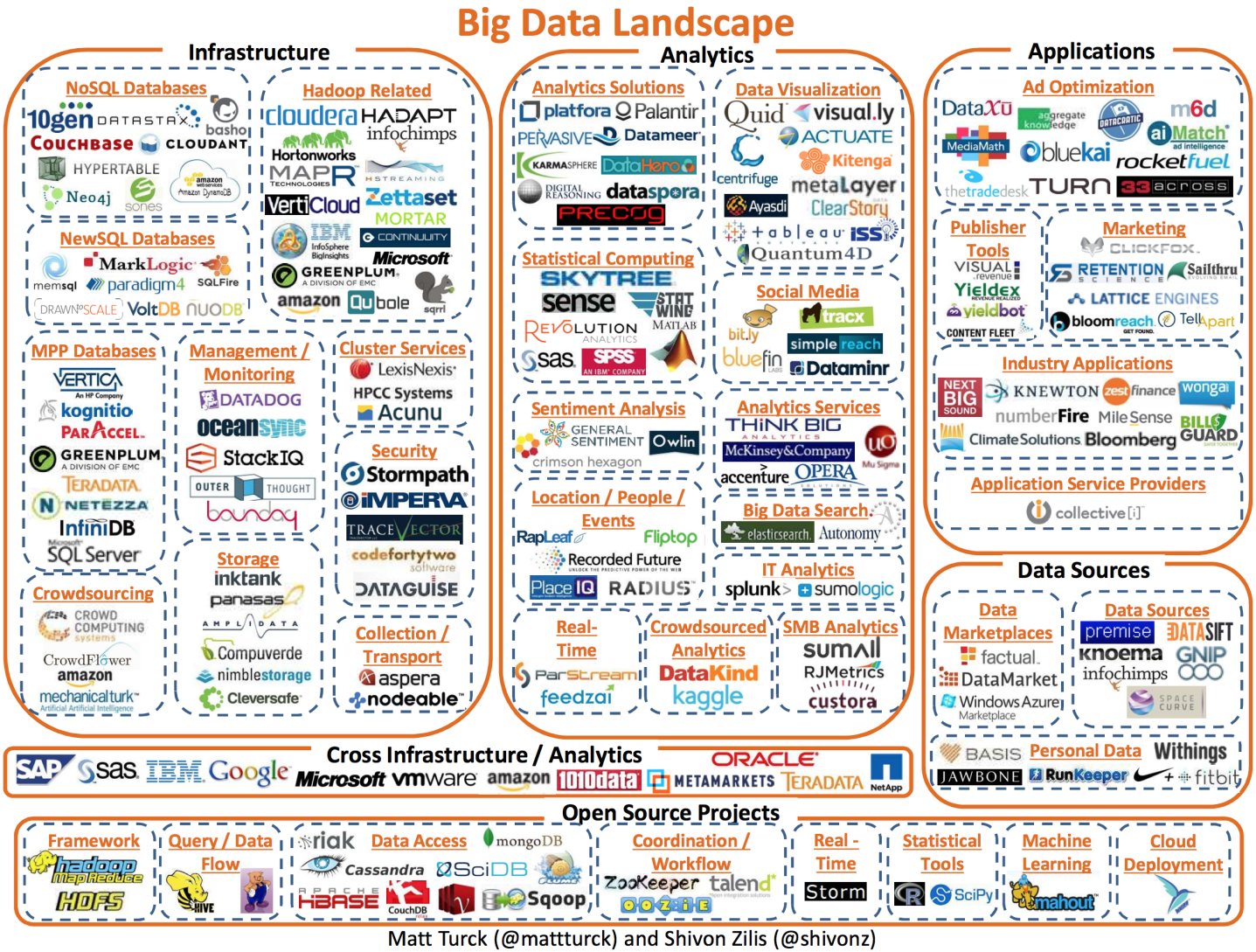

```
export AWS_ACCESS_KEY_ID=<your-access-keyid>
export AWS_SECRET_ACCESS_KEY=<your-access-key-secret>
cellule/spark/ec2/sparl-ec2 -i cellule.pem -k cellule -s <number of machines> launch <cluster-name>
ssh -i cellule.pem root@<your-cluster-master-dns>
spark-ec2/copy-dir ephemeral-hdfs/conf
ephemeral-hdfs/bin/hadoop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



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

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

Supervised learning

Setting

- Data $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ for $i = 1, \ldots, n$
- x_i is the *features* of i
- y_i is the *label* of i
- $y_i \in \mathbb{R}$ (regression) $y_i \in \{-1, 1\}$ (binary classification)
- Usually, assume (x_i, y_i) are i.i.d

Aim

- Based on (x_i, y_i), learn a function that predicts y based on a new x (generalization property)

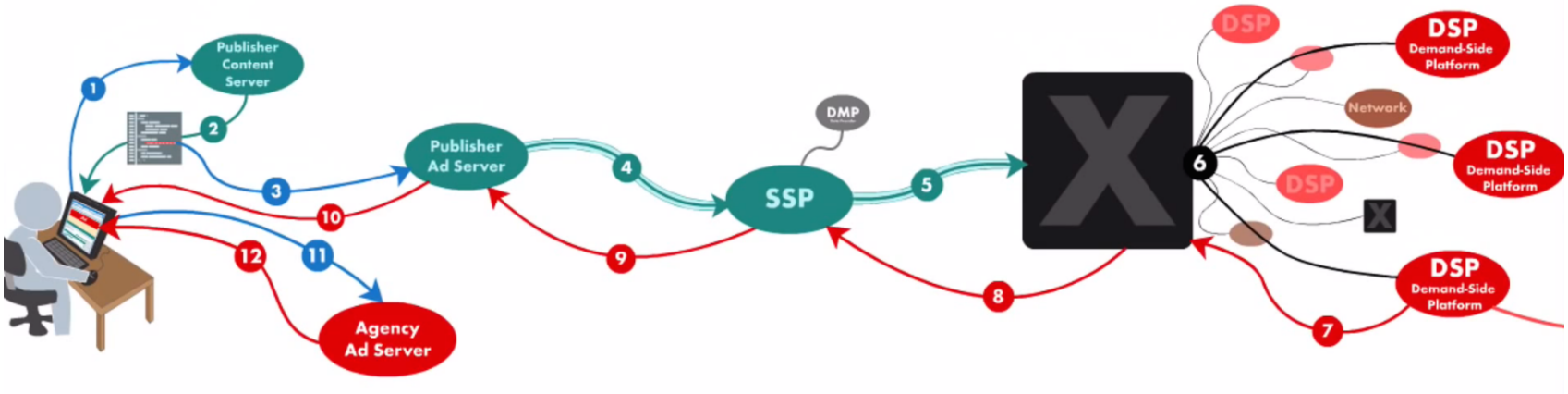
Scaling

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

Scenarios

- 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

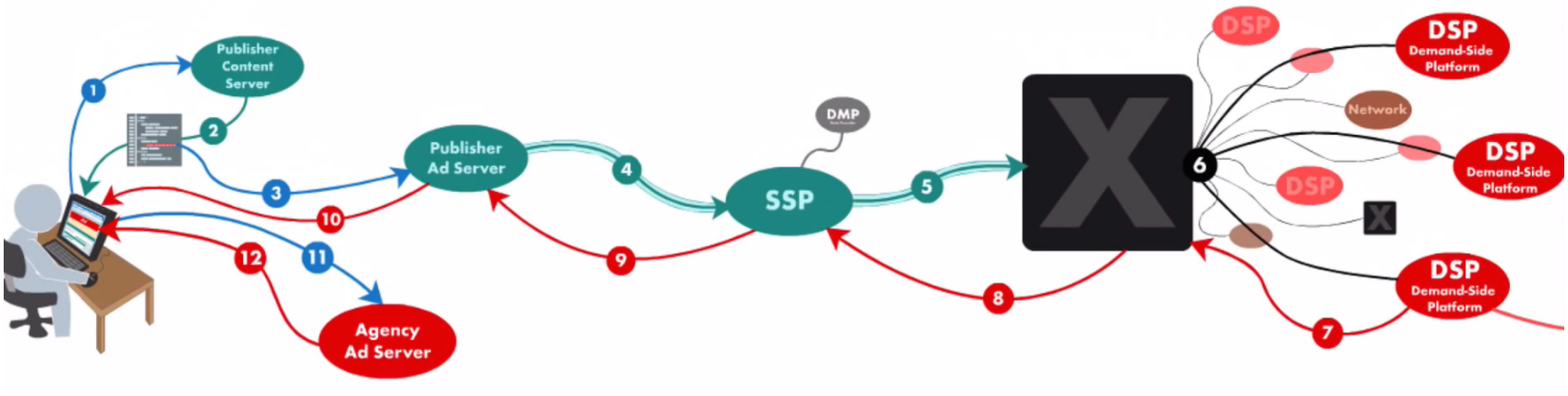
An example: Real Time Bidding



- A **customer** visits a webpage with his browser: a complex process of content selection and delivery beggins.
- 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)

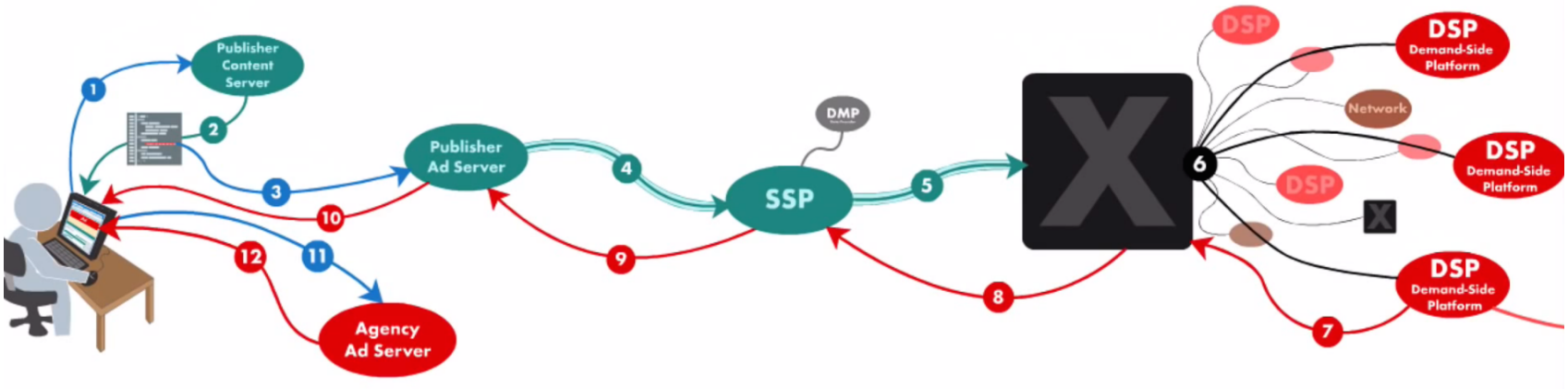
An example: Real Time Bidding



- 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

An example: Real Time Bidding



Some figures

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

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**

A problem

- n and d are large: training is too time-consuming for a complex function f

A simplification

- Choose a **linear** function f :

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

for a parameter $\theta \in \mathbb{R}^d$ to be **trained**

Remark

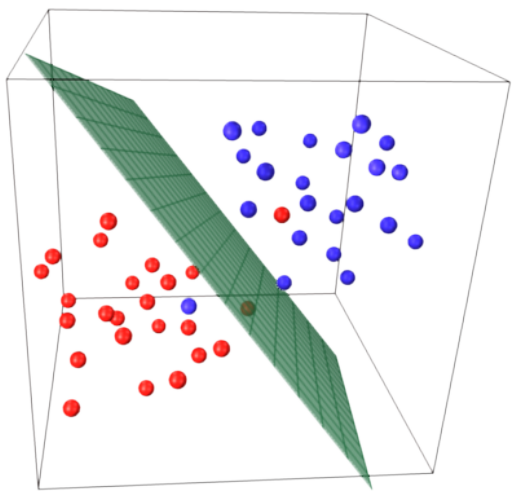
- linear with respect to x_j, but **you** choose the features x_j
- usually not linear w.r.t the raw features: **feature engineering**

Logistic regression

- The most widespread approach
- Assumes that the **log-odd ratio** is linear:

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

- Leads to a **linear** separation between the 1s and $-1s$



Goodness of fit $=$ $-$ log-likelihood, equal in this case to

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

So, let's just find out (more on that later...)

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

Now, given a new x, predict y using

$$\hat{y} = \text{sign}(x^\top \hat{\theta})$$

End of story ?

- No !
- Overfitting problem

Penalization

Minimizing only

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, x_i^\top \theta)$$

is generally a bad idea. Minimize instead

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

where

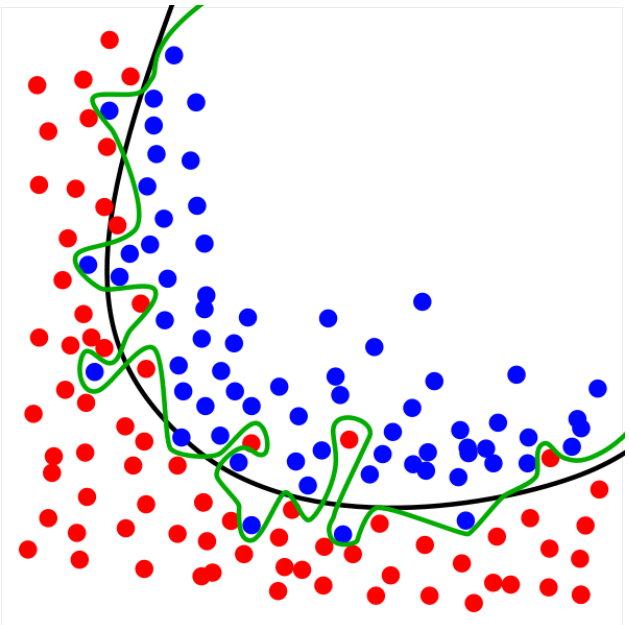
- pen is a **penalization** function, it forbids θ to be "too complex"
- $\lambda > 0$ is a **tuning** or **smoothing** parameter, that **balances** goodness-of-fit and penalization

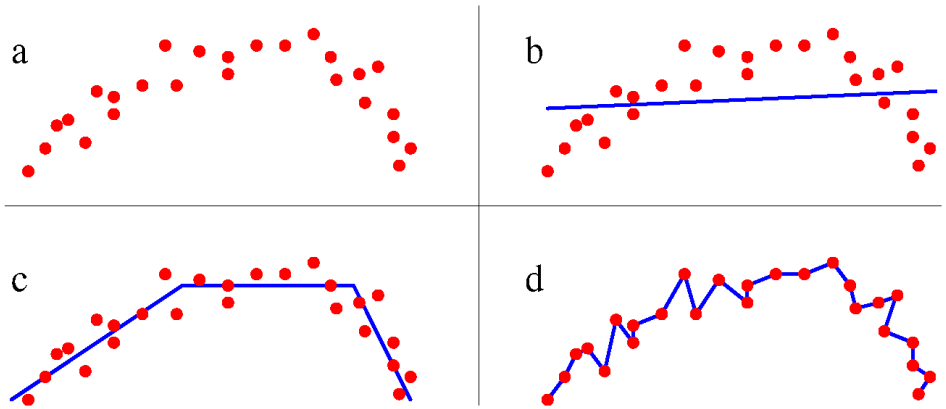
Penalization

Why using penalization?

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

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





Penalization

Penalization most widely used is

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

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, model is simpler, faster prediction (e.g. RTB example)
- We say that $\hat{\theta}$ is **sparse**
- How to do it ?

Tempting to use

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

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: ℓ_1-penalization

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

- Convex relaxation principle. Also called Lasso
- 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?

Solution of

$$\underset{a \in \mathbb{R}}{\text{argmin}} \left\{ \frac{1}{2}(a-b)^2 + \lambda |a| \right\},$$

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

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

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

Penalization – Lasso

A minimizer

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

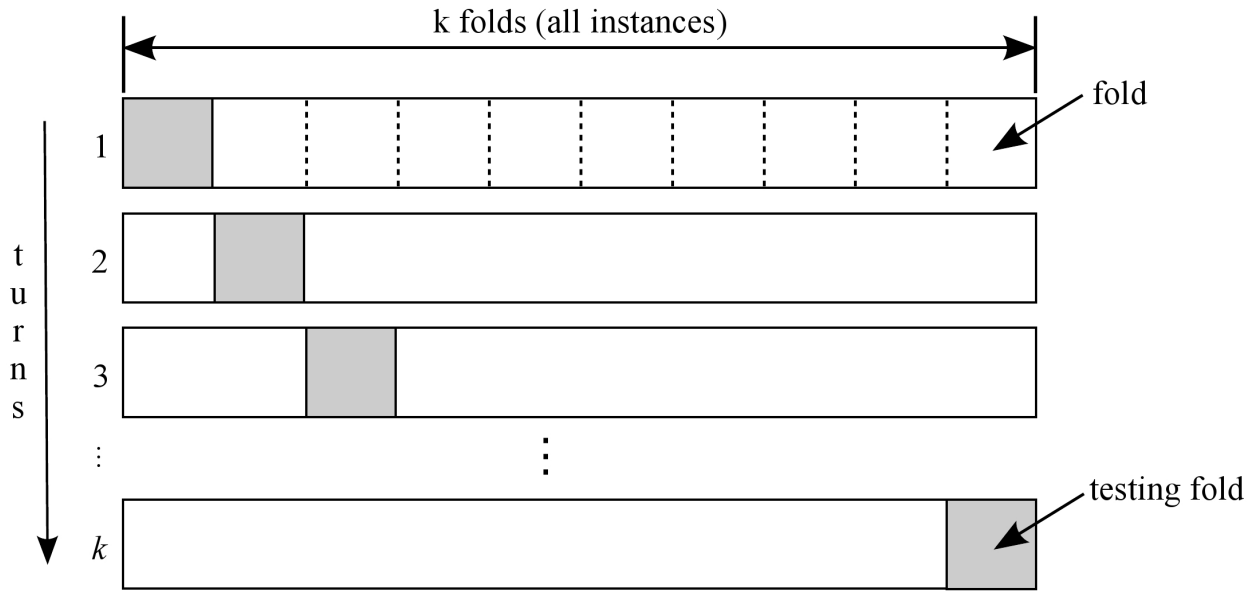
is typically sparse ($\hat{\theta}_j = 0$ for many j).

- 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

How to choose it?

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

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

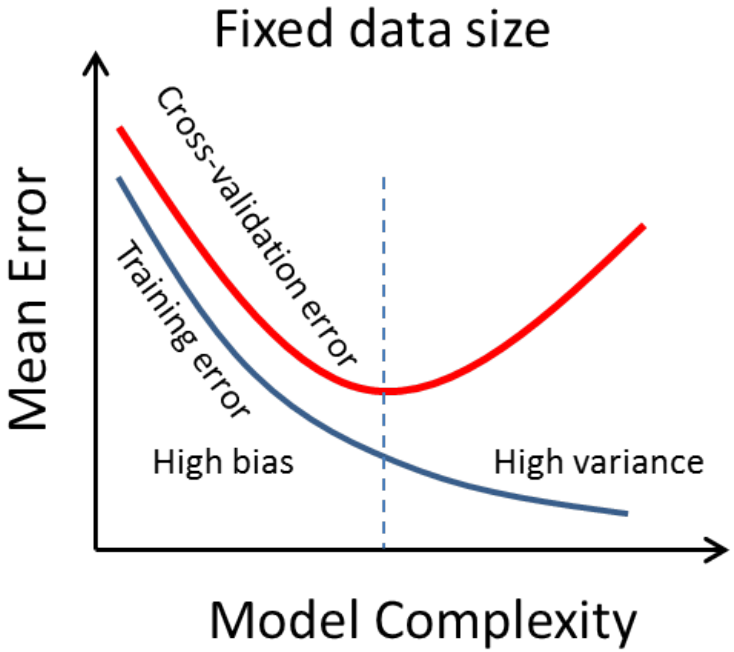
- Put $D_{v,\texttt{train}} = \cup_{v' \neq v} I_{v'}$ and $D_{v,\texttt{test}} = I_v$
- Find

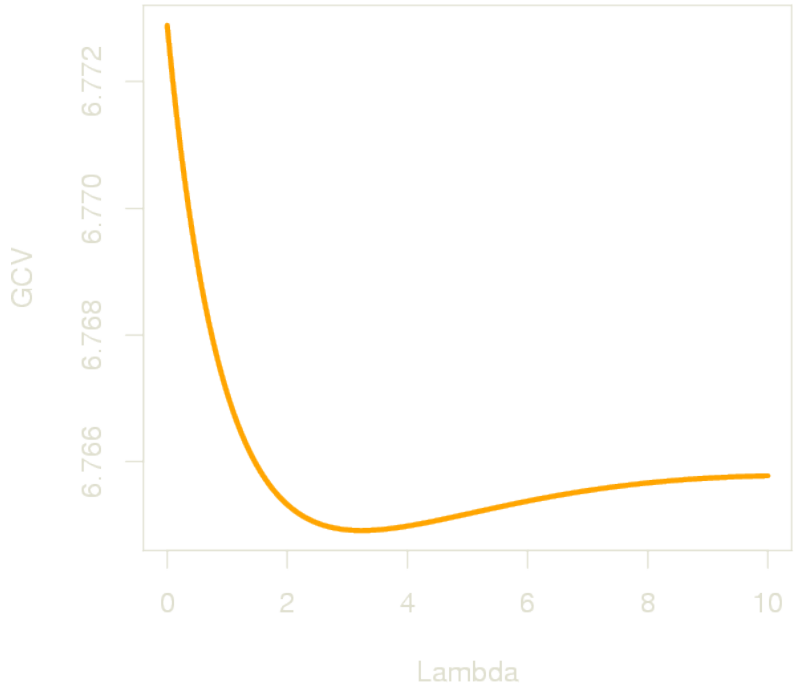
$$\hat{\theta}_{v,\lambda} \in \operatorname*{argmin}_{\theta} \left\{ \frac{1}{|D_{v,\texttt{train}}|} \sum_{i \in D_{v,\texttt{train}}} \ell(y_i, x_i^\top \theta) + \lambda \, \text{pen}(\theta) \right\}$$

Take

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

Cross-validation





- Training error:

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

- Testing, validation or cross-validation error:

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

Introduction

- Optimization is the main problem in practice
- Need efficient optimization for the training step
- Recent advances on this topic lately, aimed at machine learning applications

Let's put for short

$$f(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, x_i^\top \theta) \quad \text{and} \quad g(\theta) = \lambda \|\theta\|_1$$

How to minimize $f + g$?

A key point: the **descent lemma**

- If f convex and ∇f is L-Lipschitz, then for any $\theta, \theta_k \in \mathbb{R}^d$

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

and remark that

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

where $T_{g/L}(\theta)=\operatorname{sign}(\theta)\odot(|\theta|-\lambda/L)_+$

Gradient descent

Proximal gradient descent algorithm

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

Simplest algorithm, a plethora of others:

- Coordinate descent
- L-BFGS-B
- Primal-Dual
- ...

Stochastic gradient descent

What if n (and d) is large?

- Each iteration of a full gradient method has complexity $O(nd)$
- A large dataset makes a modern computer look old: go back to "old" algorithms (Robbins and Monro 1951)

Idea: we want to minimize an average of losses...

- If I choose uniformly at random $I \in \{1, \dots, n\}$, then

$$\mathbb{E}[\nabla f_I(\theta)] = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\theta) = \nabla f(\theta)$$

- $\nabla f_I(\theta)$ is an *unbiased* but very noisy estimate of the full gradient $\nabla f(\theta)$

Robbins and Monro 1951

Stochastic Gradient Descent (SGD)

- **Input**: starting point θ^0, sequence of learning rates $\{\eta_t\}_{t\geq 0}$
- For $t = 1, 2, \ldots$ until *convergence* do
  - Pick at random (uniformly) i in $\{1, \ldots, n\}$
  - Put
$$\theta^t = \theta^{t-1} - \eta_t \nabla f_i(\theta^{t-1})$$
- **Return** last θ^t

- Each iteration has complexity $O(d)$ instead of $O(nd)$ for full gradient methods
- The step size must be decreasing
- Fast in the early iterations
- Very slow convergence to a precise minimizer

Variance reduction

Recent results improve this:

- Bottou and LeCun (2005)
- Shalev-Shwartz et al (2007, 2009)
- Nesterov et al. (2008, 2009)
- Bach et al. (2011, 2012, 2014, 2015)
- T. Zhang et al. (2014, 2015)

Variance reduction

- Put $X = \nabla f_I(\theta)$ with I uniformly chosen at random in $\{1, \ldots, n\}$
- Variance of X as an approximation of $\mathbb{E}X$ is large
- Reduce it by finding C s.t. $\mathbb{E}C$ is easy to compute and such that C is highly correlated with X
- Put $Z = X - C + \mathbb{E}C$, so that $\mathbb{E}Z = \mathbb{E}X$ and

$$\operatorname{Var} Z = \operatorname{Var} X + \operatorname{Var} C - 2\operatorname{cov}(X, C)$$

Stochastic Variance Reduced Gradient (SVRG)

Input: starting point θ^0, learning rate $\eta > 0$

Put $\widetilde{\theta} \leftarrow \theta^0$

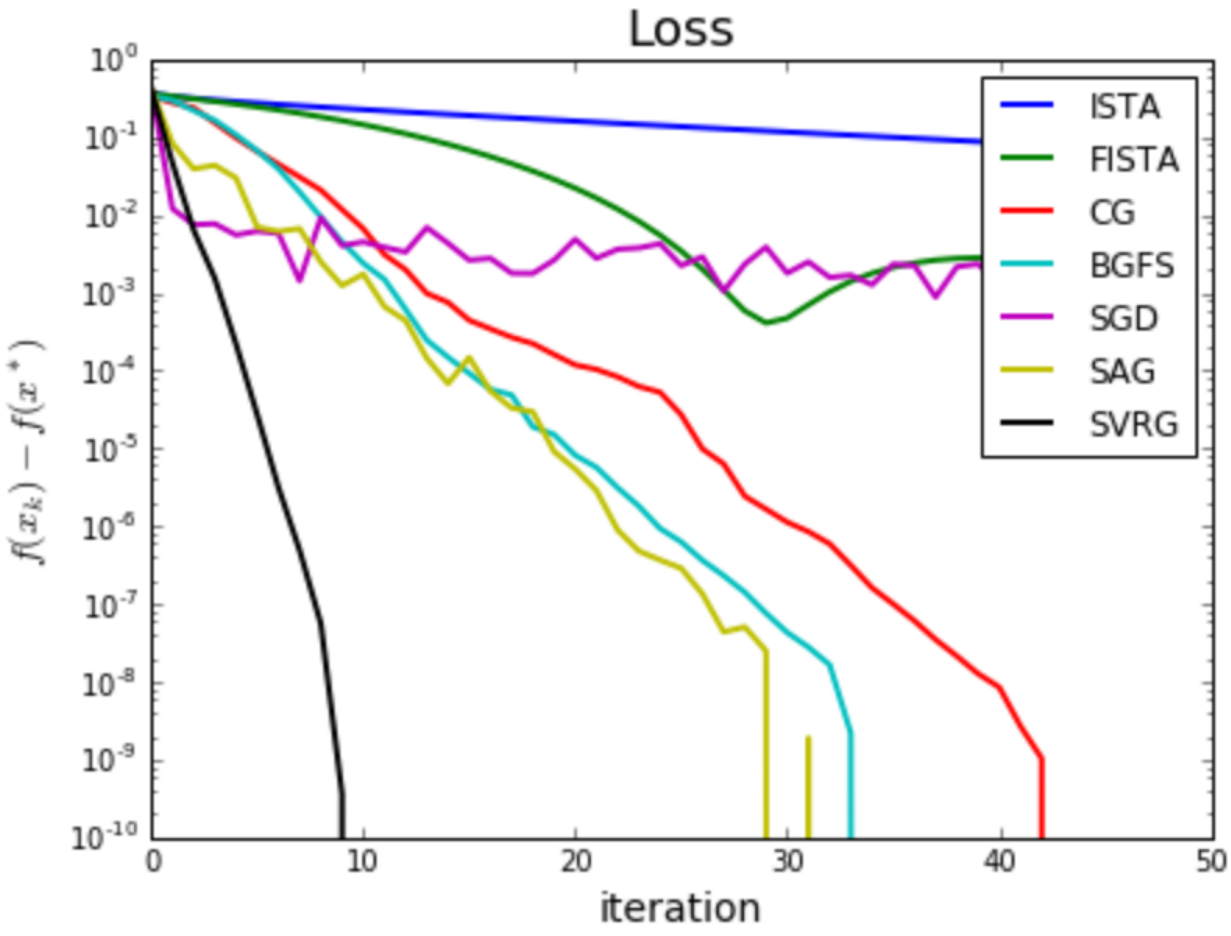
For $k = 1, 2, \ldots$ until *convergence* do

- Put $\theta_1^k \leftarrow \widetilde{\theta}$
- Compute $\nabla f(\widetilde{\theta})$
- For $t = 1, \ldots, n$
  - Pick uniformly at random i in $\{1, \ldots, n\}$
  - Apply the step

$$\theta_{t+1}^k \leftarrow \theta_t^k - \eta\left(\nabla f_i(\theta_t^k) - \nabla f_i(\widetilde{\theta}) + \nabla f(\widetilde{\theta})\right)$$

- Set $\widetilde{\theta} \leftarrow \theta_{n+1}^k$

Numerical comparison: batch VS stochastic



Conclusion

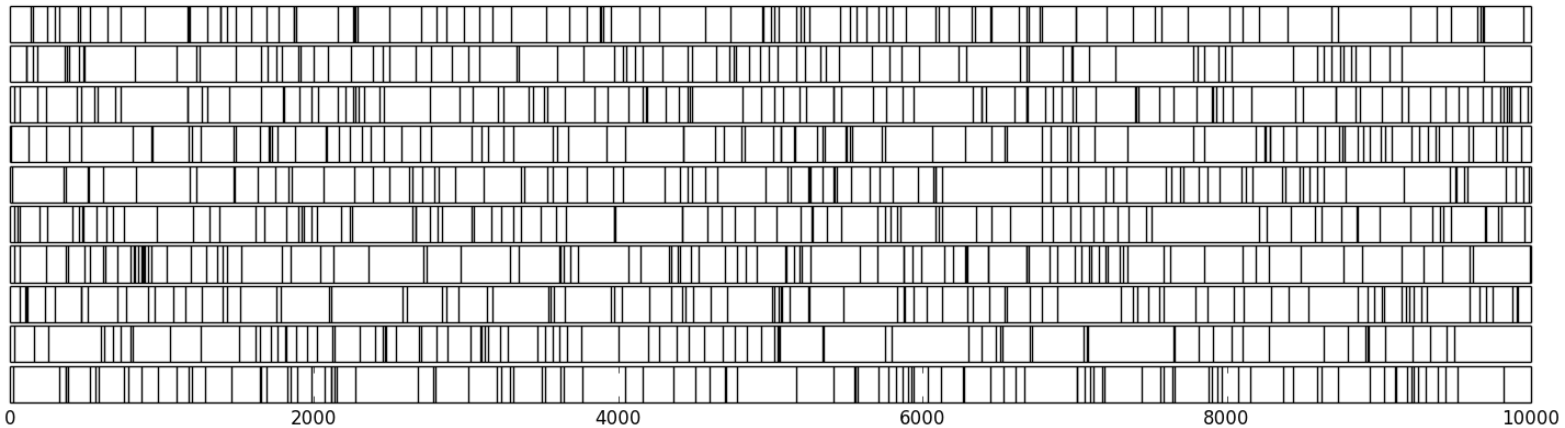
What we've seen in this glimpse of supervised learning:

- Model assumptions, simplifying assumptions
- Overfitting, penalization, sparsity, cross-validation
- Optimization matters!

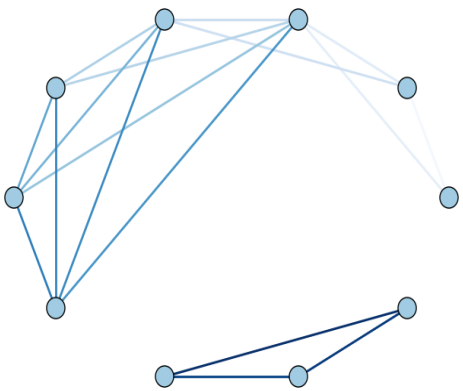
Beyond supervised learning?

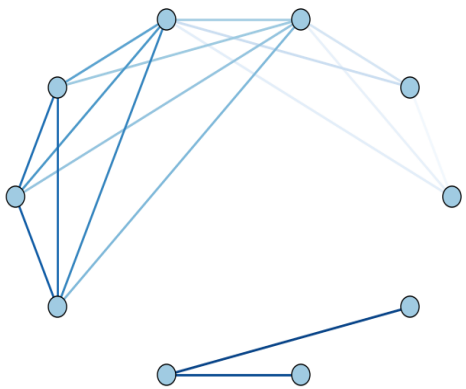
Let us describe an example: Hawkes processes

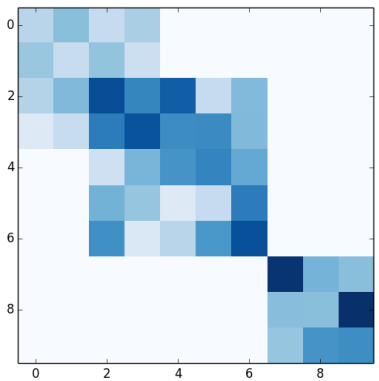
Observation



What we want to infer







Introduction

One Hawkes component can model (on a high frequency, data can be as precise as 10 microseconds)

- an arrival of a particular type of order (with a particular size, any agent)
- a move of the price of a particular size and direction and of a particular asset
- an order of a particular agent

NB:

- possible to combine all that and do some high-dimensional models
- possible to add exogenous events with some impacts (e.g., news)

Introduction

- A two-components model of the high-frequency (mid)price dynamics (one component for upward jumps and one for downward jumps). Basket of assets=high dimension
- Multiple agents models: an event is a particular type of order of a particular size by a particular agent, including components for the moves of the price. Analysis of the interactions of all the agents and their impact on the price

Model: Multivariate Hawkes Process (MHP)

- A d-dimensional counting process $N = [N_1, \ldots, N_d]^\top$
- d is "large"
- Observed on $[0, T]$
- N_j has intensity λ_j, namely

$$\mathbb{P}\big(N_j \text{ has a jump in } [t, t + dt] \mid \mathcal{F}_t\big) = \lambda_j(t)dt$$

for $j = 1, \ldots, d$ where $\mathcal{F}_t$ some filtration

Model: Multivariate Hawkes Process (MHP)

- MHP assumes the following autoregressive structure:

$$\lambda_j(t) = \mu_j(t) + \int_{(0,t)} \sum_{k=1}^{d} \varphi_{j,k}(t-s) dN_k(s),$$

- $\mu_j(t) \geq 0$ baseline intensity of the j-th coordinate
- $\varphi_j : \mathbb{R}^+ \to \mathbb{R}^+$ self-exciting component
- Write this in matrix form

$$\lambda(t) = \boldsymbol{\mu} + \int_{(0,t)} \boldsymbol{\varphi}(t-s) dN(s),$$

with $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_d]^\top$ and $\boldsymbol{\varphi}(t) = [\varphi_{j,k}(t)]_{1 \leq j,k \leq d}$.

- Notation:

$$\int_{(0,t)} \varphi(t-s) dN_k(s) = \sum_{i:0<T_{i,k}<t} \varphi(t - T_{i,k})$$

A brief history of MHP

Introduced by Hawkes in 1971

- **Earthquakes and geophysics** : Kagan and Knopoff (1981), Zhuang, Harte, Werner, Hainzl and Zhou (2012)
- **Genomics** : Reynaud-Bouret and Schbath (2010)
- **High-frequency Finance** : Bacry Delattre Hoffmann and Muzy (2013)
- **Terrorist activity** : Porter and White (2012)
- **Neurobiology** : Hansen, Reynaud-Bouret and Rivoirard (2012)
- **Social networks** : Carne and Sornette (2008), Simma and Jordan (2010), Zhou Song and Zha (2013)
- And even **FPGA-based implementation** : Guo and Luk (2013)

Estimation for MHP: some references

Parametric estimation (Maximum likelihood)

- First work : Ogata 78
- Simma and Jordan (2010), Zhou Song and Zha (2013)
  $\rightarrow$ Expected Maximization (EM) algorithms, with priors

Non parametric estimation

- Marsan Lengliné (2008), generalized by Lewis, Mohler (2010)
  $\rightarrow$ EM for penalized likelihood function
  $\rightarrow$ Monovariate Hawkes processes, Small amount of data, No theoretical results
- Reynaud-Bouret and Schbath (2010)
  $\rightarrow$ Developed for small amount of data (Sparse penalization)
- Bacry and Muzy (2014)
  $\rightarrow$ Larger amount of data

MHP in large dimension

Dimension d is large:

- Need a simple parametric model on μ and φ
- For inference: we want a **tractable** and **scalable** optimization problem

A recent work [Bacry, G., Mastromatteo, Muzy 2016]:

- Focus on optimization (the training task) for parametric Hawkes models
- Exploits a recent mean-field property of this model
- Improves state-of-the-art solvers in this case

A simple parametrization of the MHP

Simple parametrization:

- Constant baselines $\mu_j(\cdot) \equiv \mu_j$
- Take
$$\varphi_{j,k}(t) = a_{j,k}\alpha e^{-\alpha t}$$
- $a_{j,k}$ = level of interaction between agents j and k
- α = lifetime of instantaneous excitation (assumed known here...)

The matrix
$$\boldsymbol{A} = [a_{j,k}]_{1 \le j,k \le d}$$
is understood has a **weighted adjacency matrix** of mutual excitement of agents $\{1, \dots, d\}$
NB: we can consider actually
$$\varphi_{j,k}(t) = \sum_{l=1}^{L} a_{j,k,l}\alpha_l e^{-\alpha_l t}$$

A simple parametrization of the MHP

We end up with intensities

$$\lambda_{j,\theta}(t) = \mu_j + \int_{(0,t)} \sum_{k=1}^{d} a_{j,k} \alpha e^{-\alpha(t-s)} dN_k(s)$$

for $j \in \{1, \ldots, d\}$ where

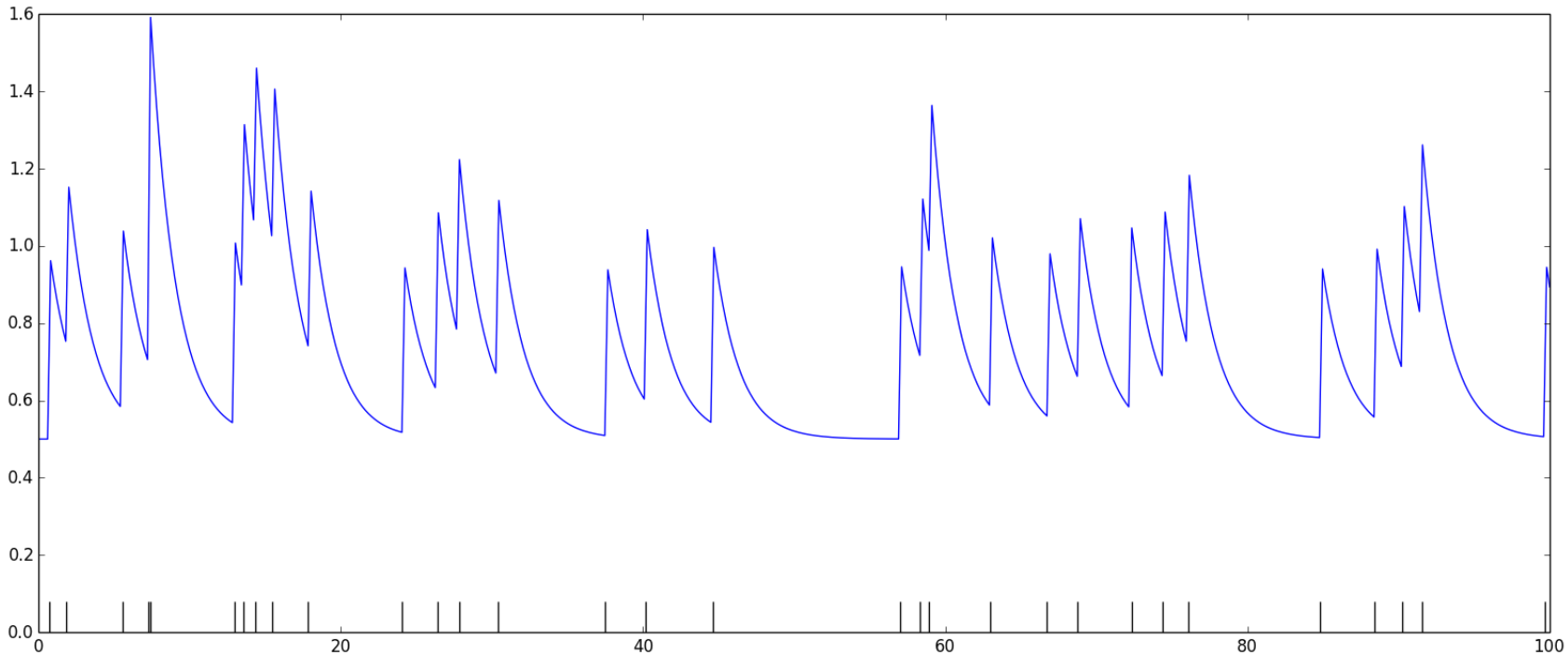
$$\theta = [\mu, \boldsymbol{A}]$$

with

- baselines $\mu = [\mu_1, \ldots, \mu_d]^\top \in \mathbb{R}_+^d$
- interactions $\boldsymbol{A} = [a_{j,k}]_{1 \leq j,k \leq d} \in \mathbb{R}_+^{d \times d}$

A simple parametrization of the MHP

For $d = 1$, intensity λ_θ looks like this:



Maximum likelihood functional

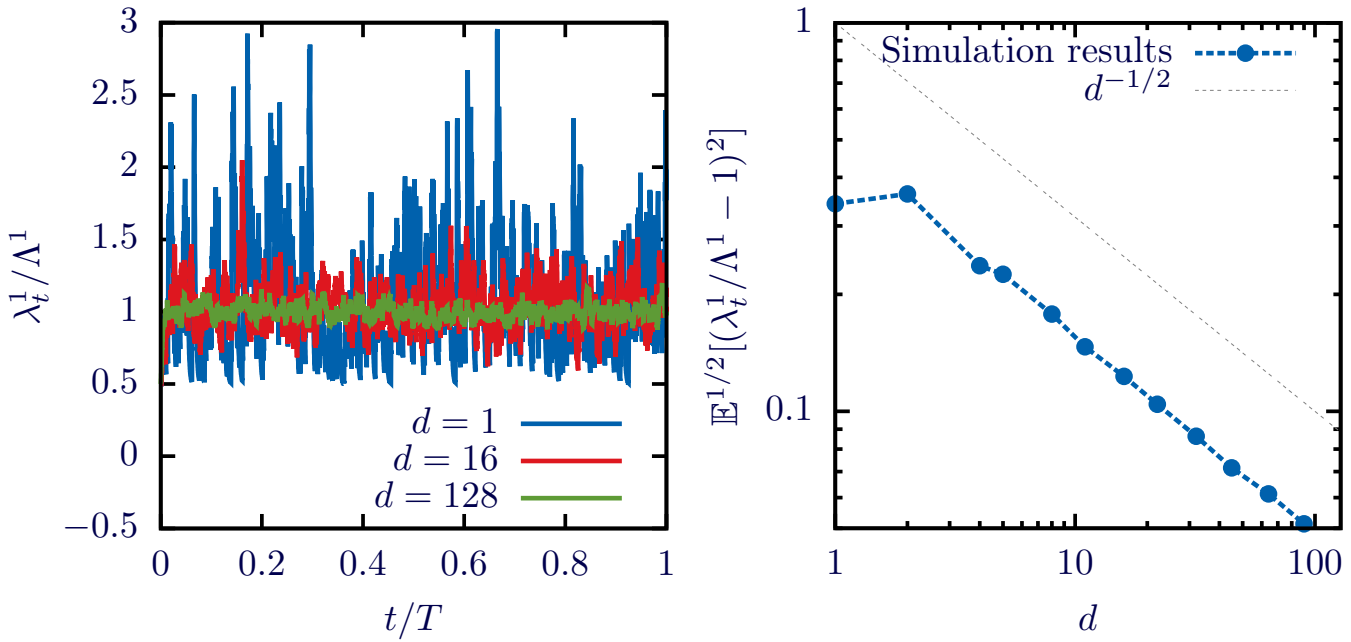
$$-\ell_T(\theta) = \sum_{j=1}^{d} \Big\{ \int_0^T (\lambda_{j,\theta}(t) - 1) dt - \int_0^T \log \lambda_{j,\theta}(t) dN_j(t) \Big\}$$

with

$$\lambda_{j,\theta}(t) = \mu_j + \sum_{k=1}^{d} a_{j,k} \int_{(0,t)} \alpha \exp\big(-\alpha(t-s)\big) dN_k(s)$$

Mean-Field approximation

- For inference, we exploit the fact that d is large
- Using a Mean-Field approximation! [Delattre, Fournier and Hoffmann 2015]



Mean-Field approximation

Idea: when d is large, we have dilute interactions, ie

$$\lambda_{j,\theta}(t) \approx \frac{N_j([0,T])}{T} = \bar{\Lambda}_j$$

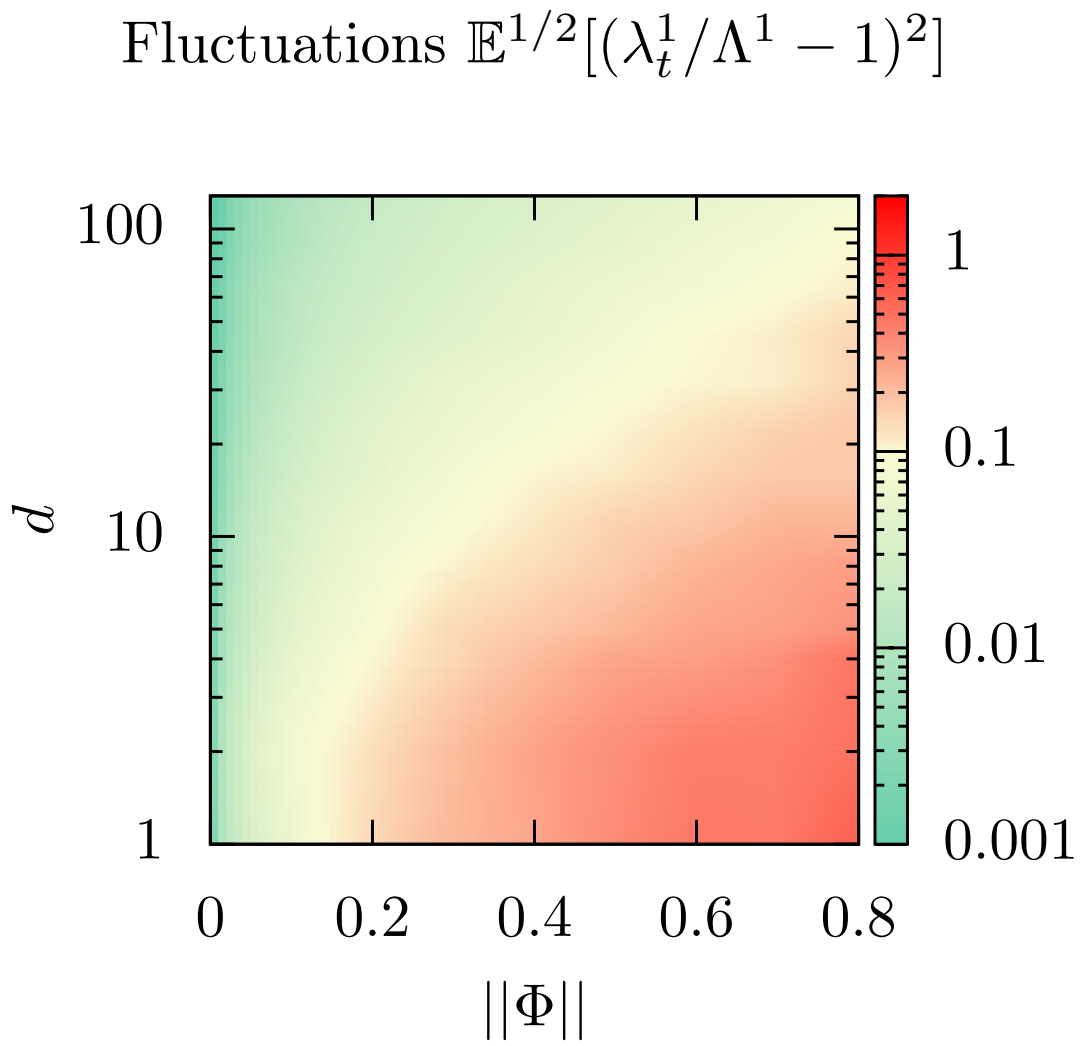
So, we use the following approximation for inference

$$\log \lambda_{j,\theta}(t) \approx \log \bar{\Lambda}_j + \frac{\lambda_{j,\theta}(t) - \bar{\Lambda}_j}{\bar{\Lambda}_j} - \frac{(\lambda_{j,\theta}(t) - \bar{\Lambda}_j)^2}{2(\bar{\Lambda}_j)^2}$$

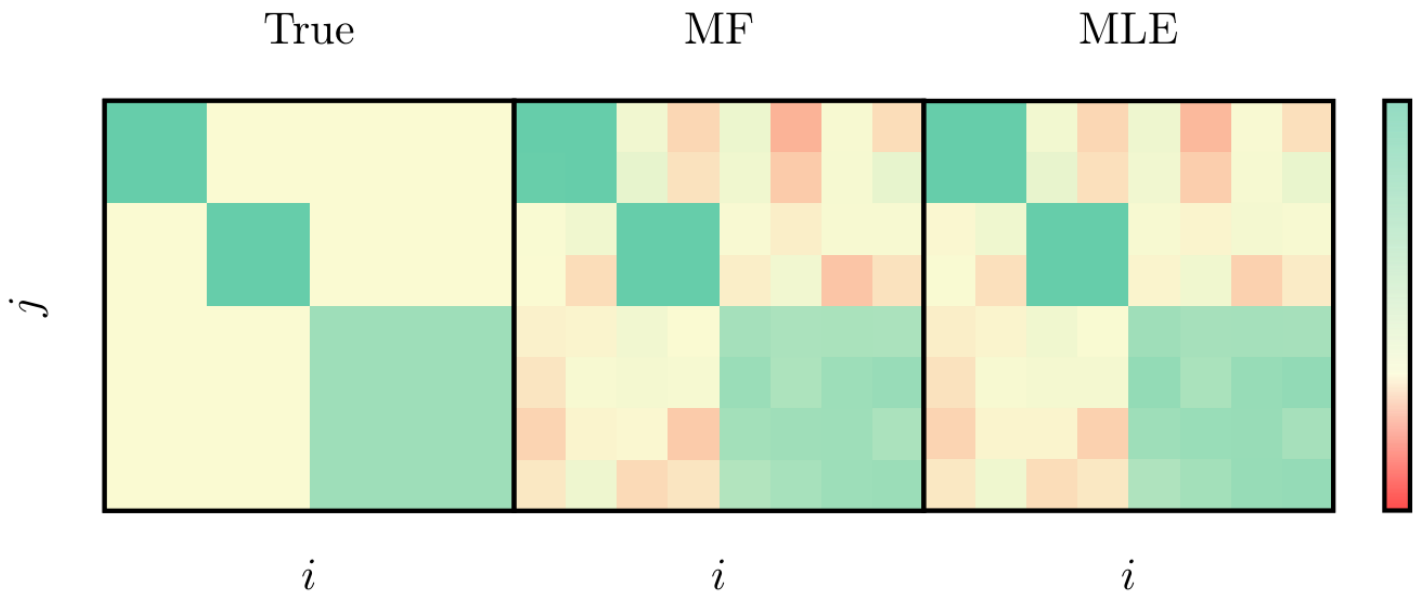
It gives an approximation of the log-likelihood

- Faster to solve
- Surprisingly sharp in large dimension

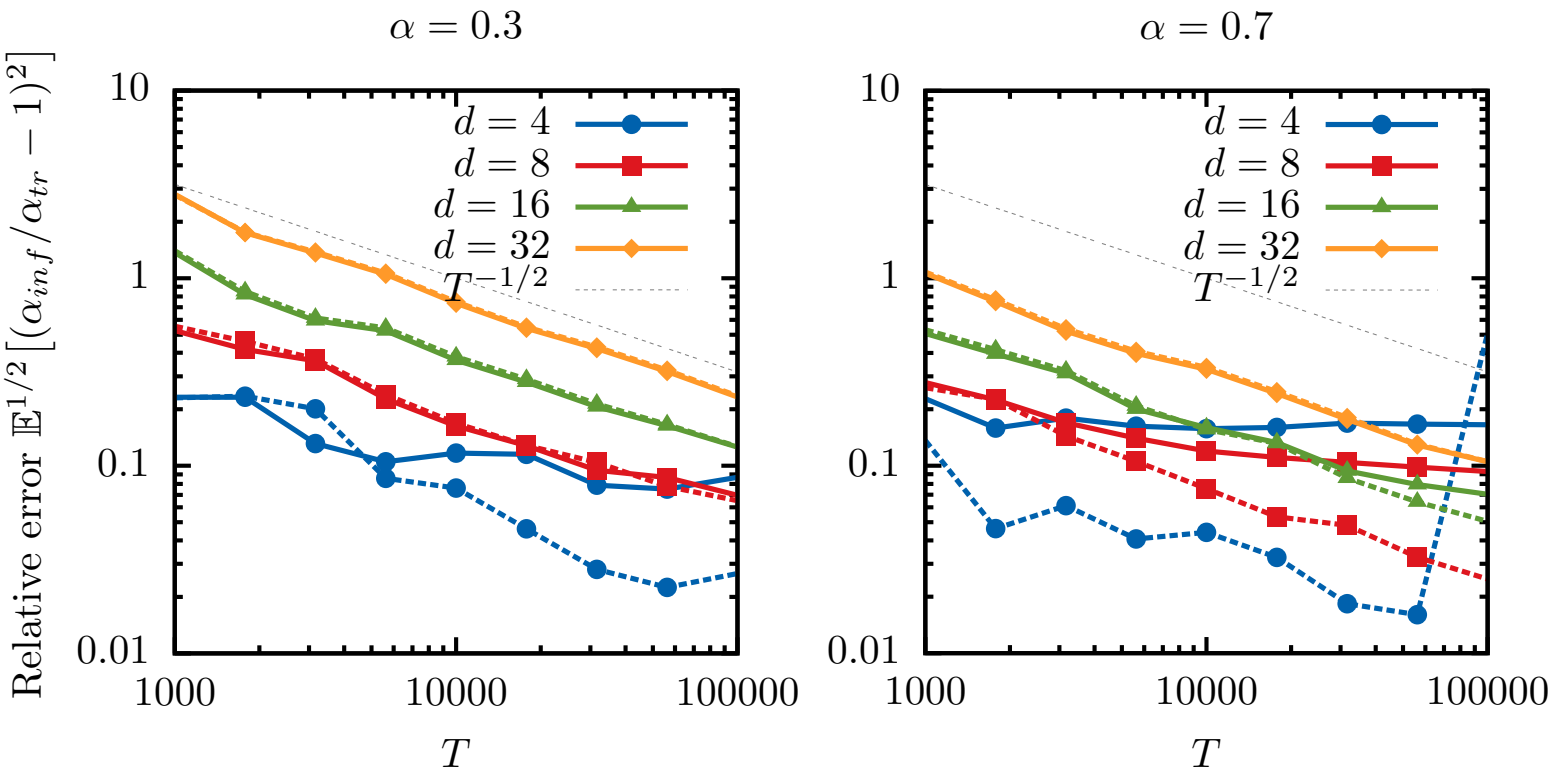
Mean-field inference for Hawkes



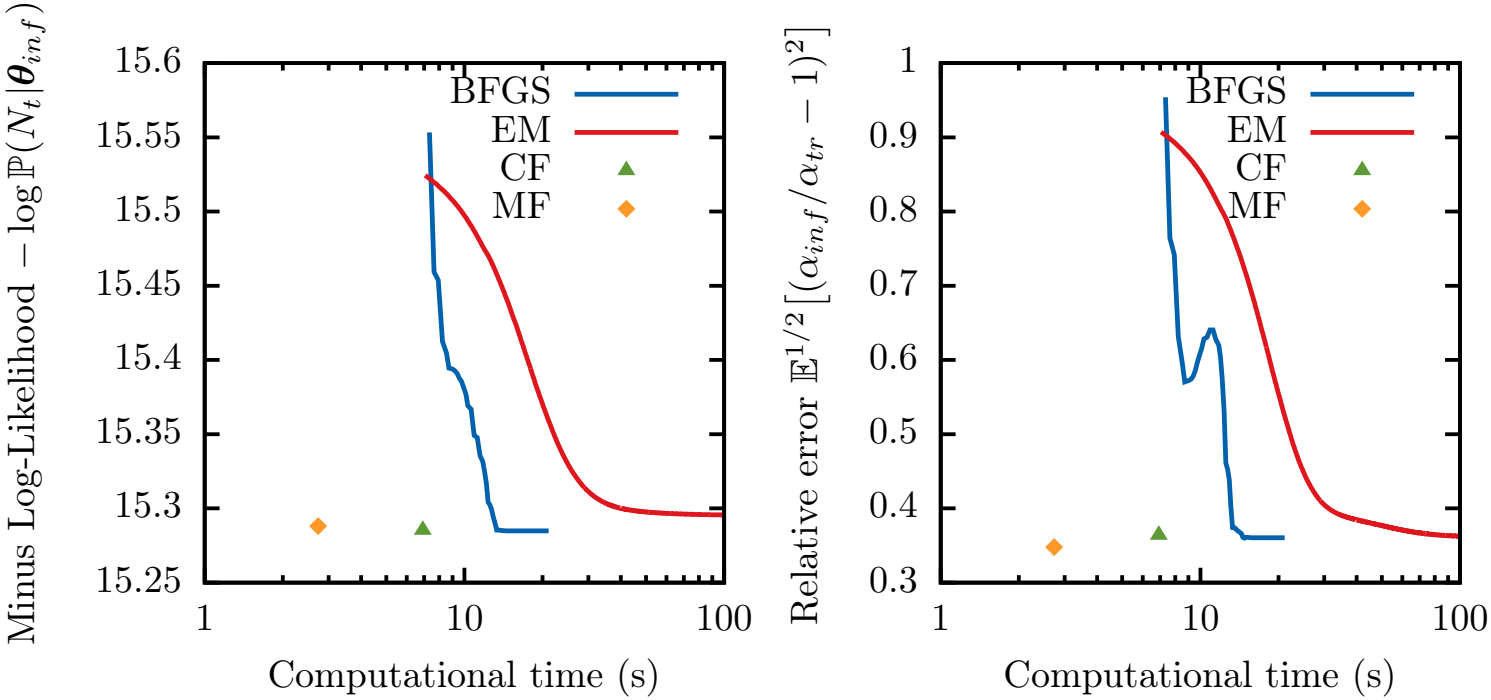
Mean-field inference for Hawkes



Mean-field inference for Hawkes



Training algorithm faster by several order of magnitude than state-of-the-art solvers



Thank you!