



Analyse de Données

Master 1 IEF, FE & CB, Rennes

Sylvain BARTHELEMY

Qui suis-je?

gwen**lake**.

Build next-gen apps using AI

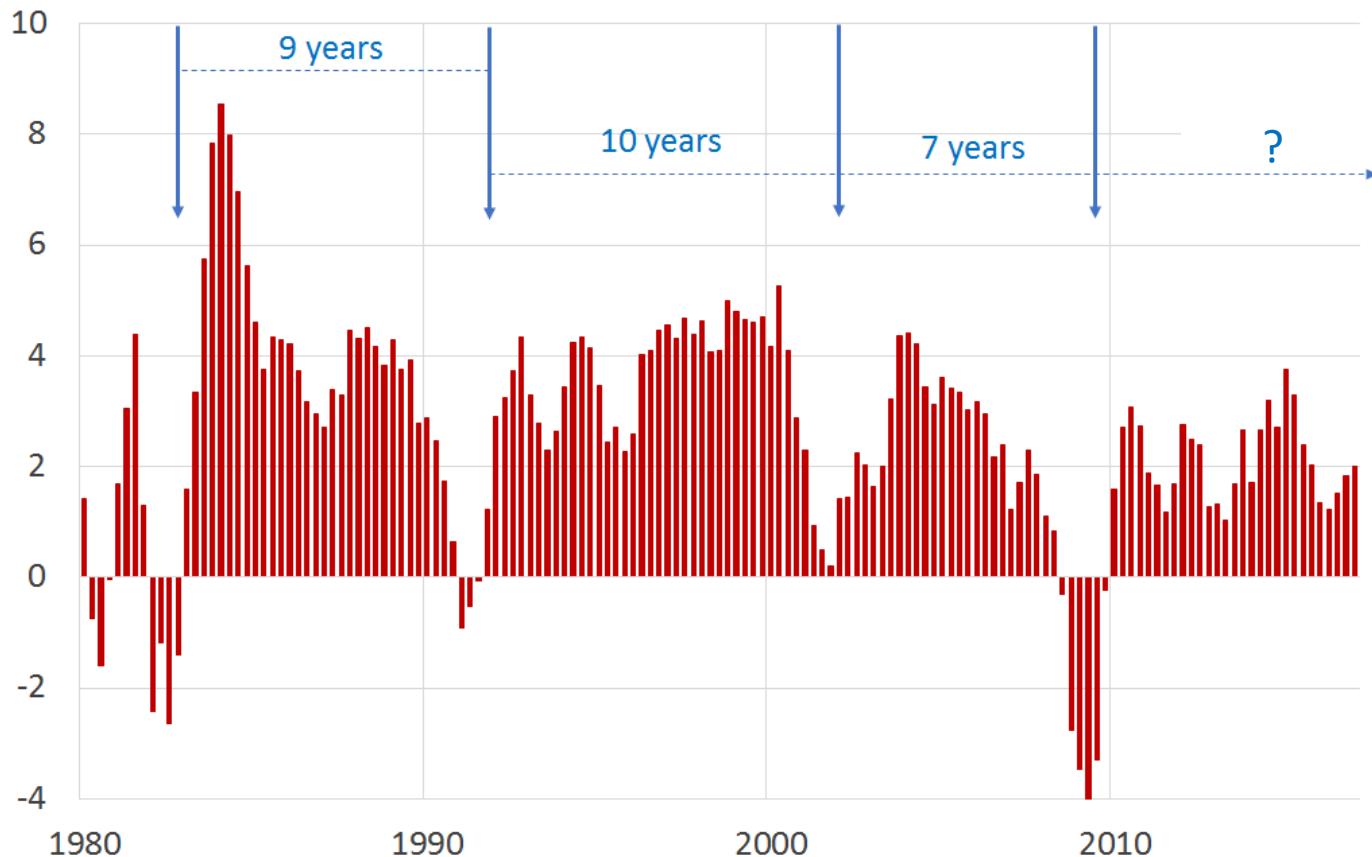
We help organizations by operationalizing AI, machine learning and data analytics

IA, Big Data & Data-Science

Retournement du cycle économique ?

US GDP growth

Y/Y, in %



Source: Datastream, TAC ECONOMICS

Bulle spéculative sur les marchés financiers

Indice Shiller PE



Source: Shiller

Les crises économiques et financières dans le monde depuis 2017...

ACTIVITE

Moyenne croissance du PIB des marchés émergents proche de 4.5
Crise Covid-19

Venezuela (-13.2%), Congo-Brazzaville (-4.6%), Kuwait (-2.5%)... et la Chine fait presque 6%, contre 14% en 2007 !

CHANGE

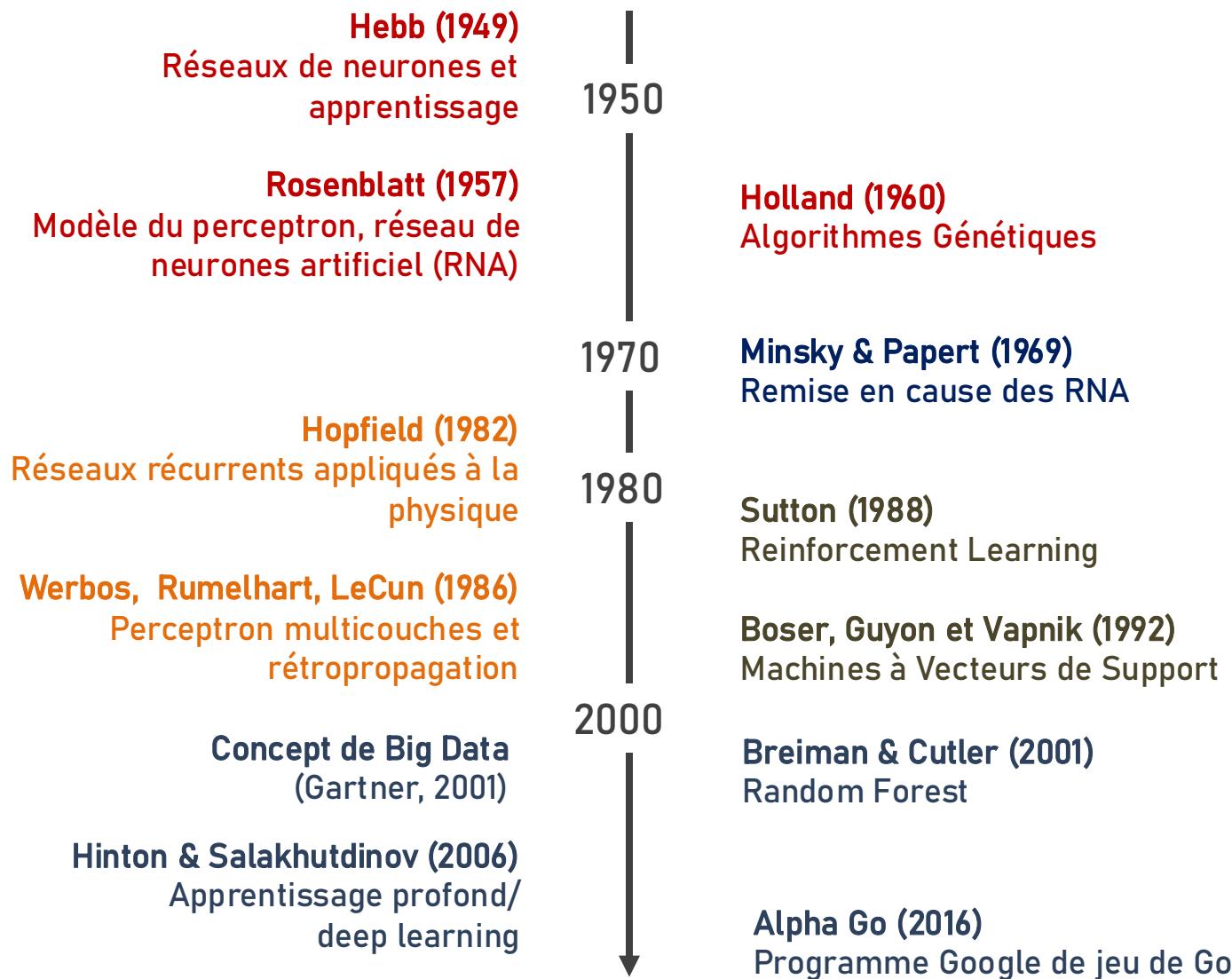
Argentine -20%
Turquie -15%
Brésil -10% Russie -8%
En 2017:
Egypte -44%
Ouzbékistan -43%
Congo-Kinshasa -31%

INFLATION

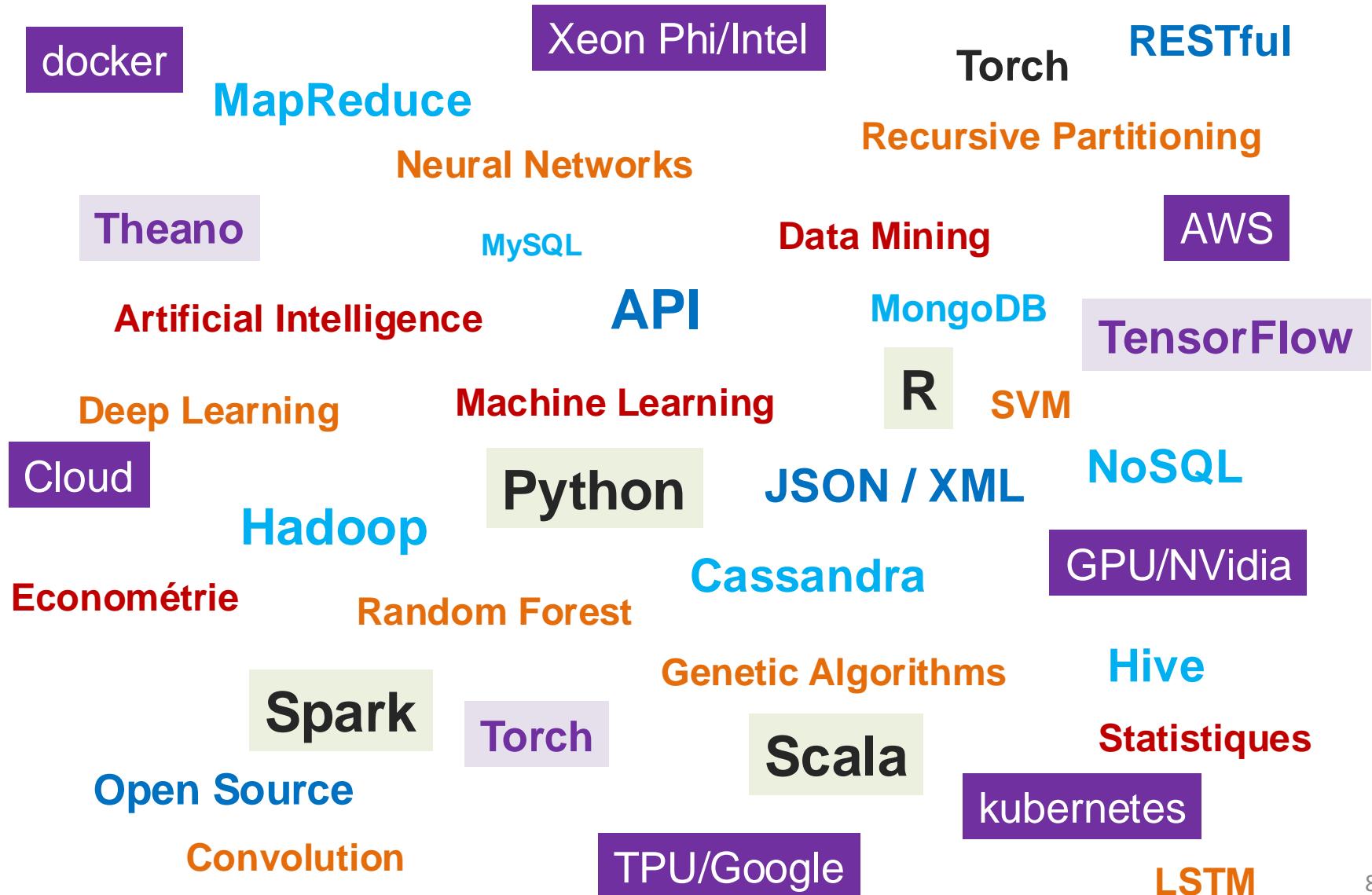
En moyenne proche de 3%

Argentine >30%, Nigéria 17%, Iran 10%, Turquie 15%, Ukraine 14%
Congo-Kinshasa 42% Libye 28%, Egypte 24%
Venezuela >1000%

Une brève histoire du machine learning et de l'IA

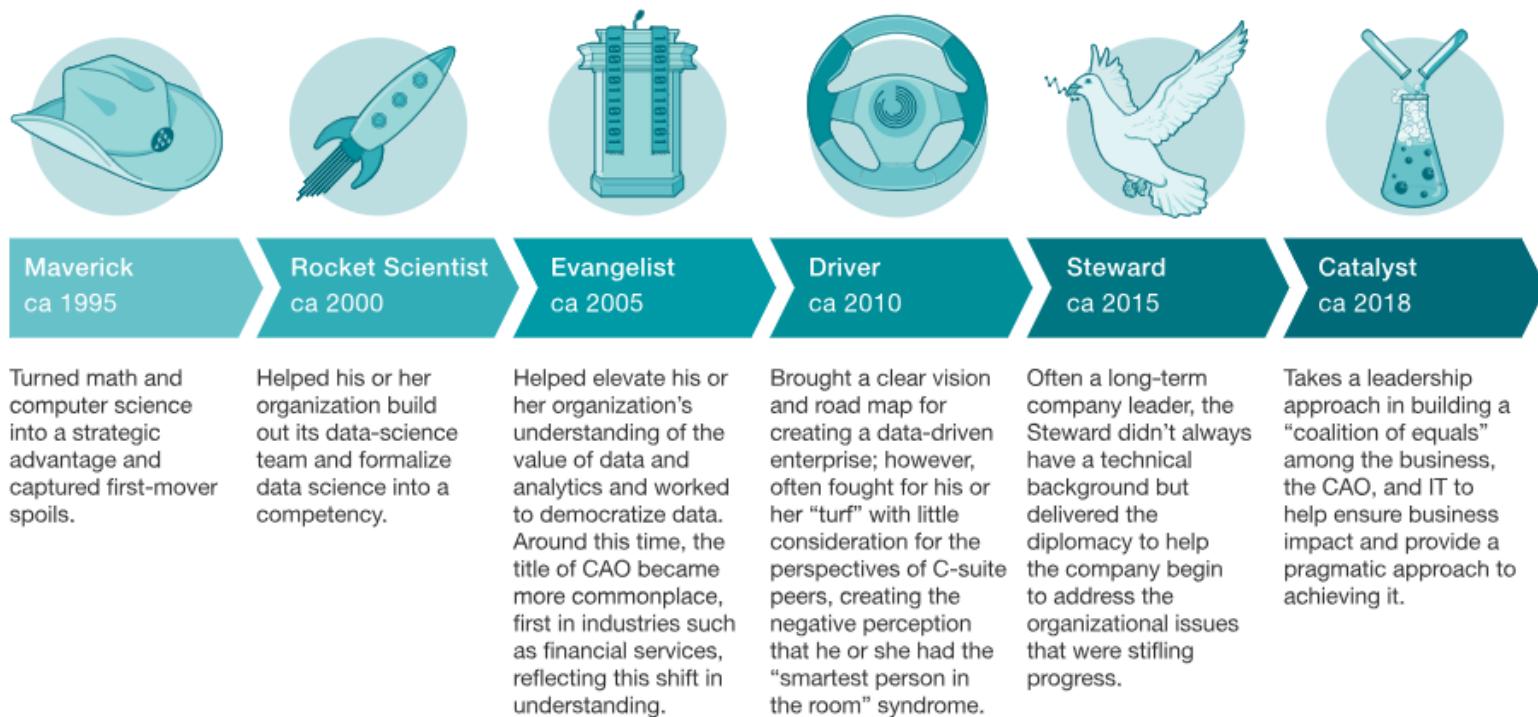


Big data and AI

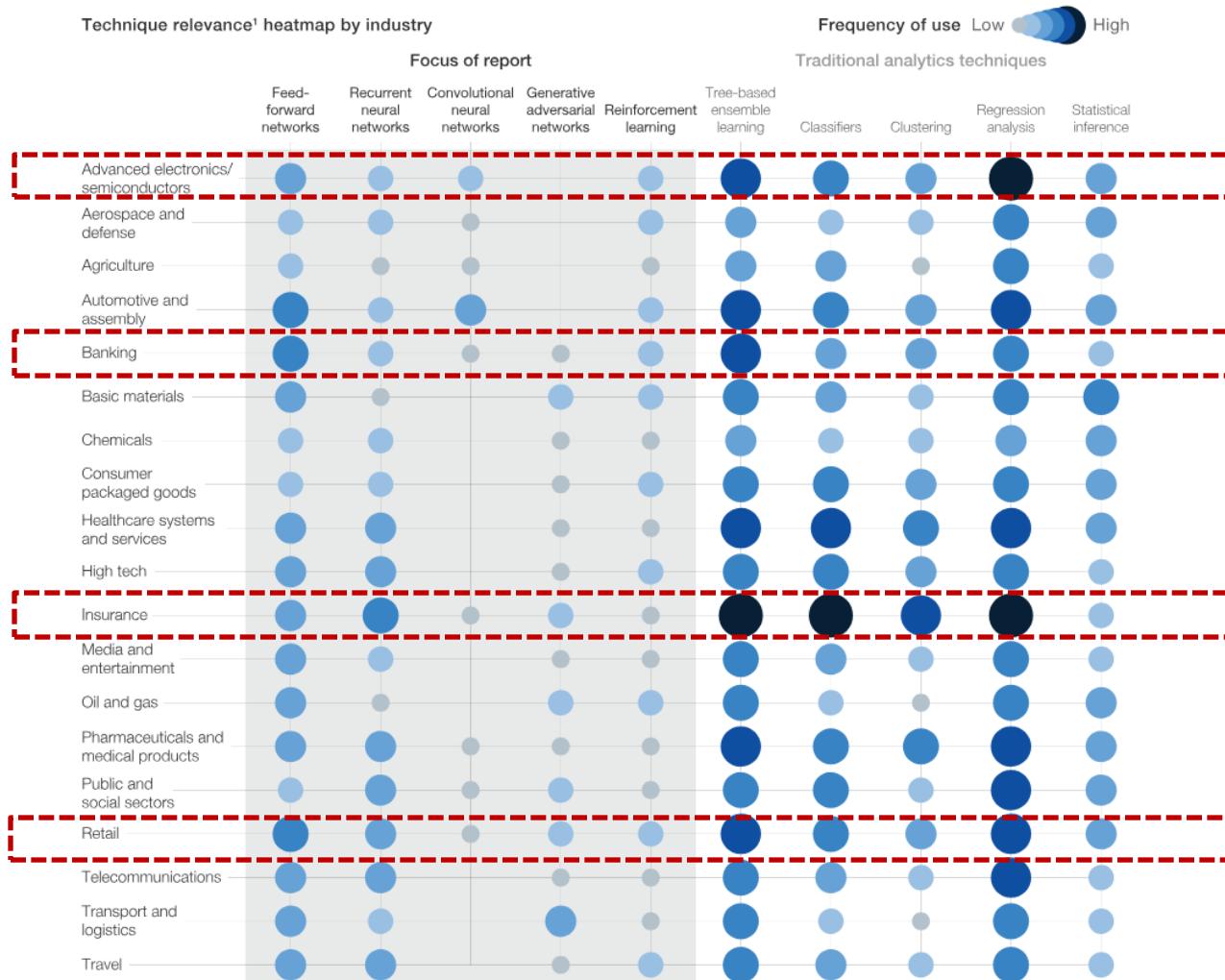


The evolving role of the Chief Analyst Officer

Dominant CAO personas emerged as data and analytics advanced over the past quarter century.

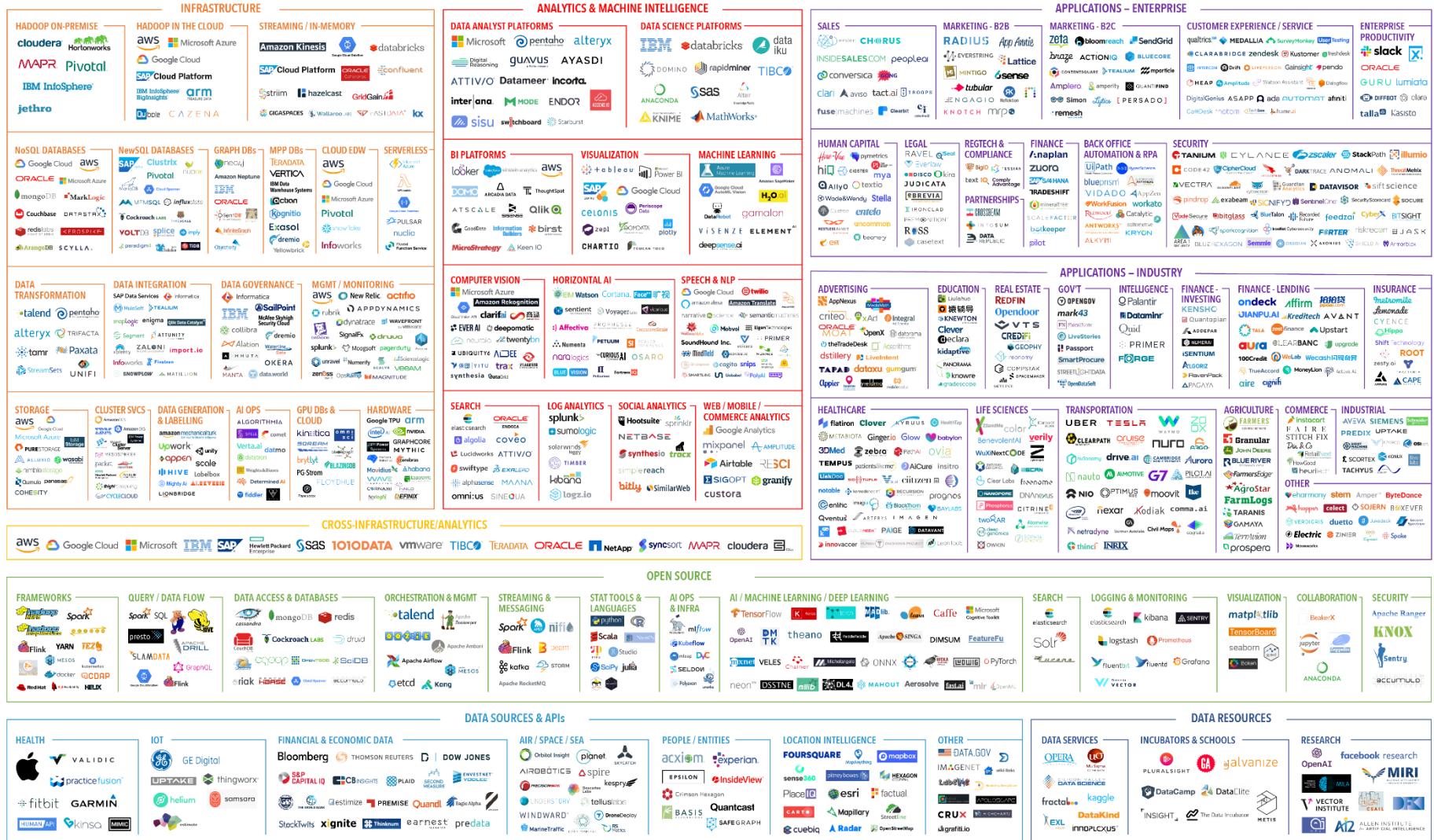


Traditional and advanced analytics techniques



L'écosystème du « datascience » et du « big data »

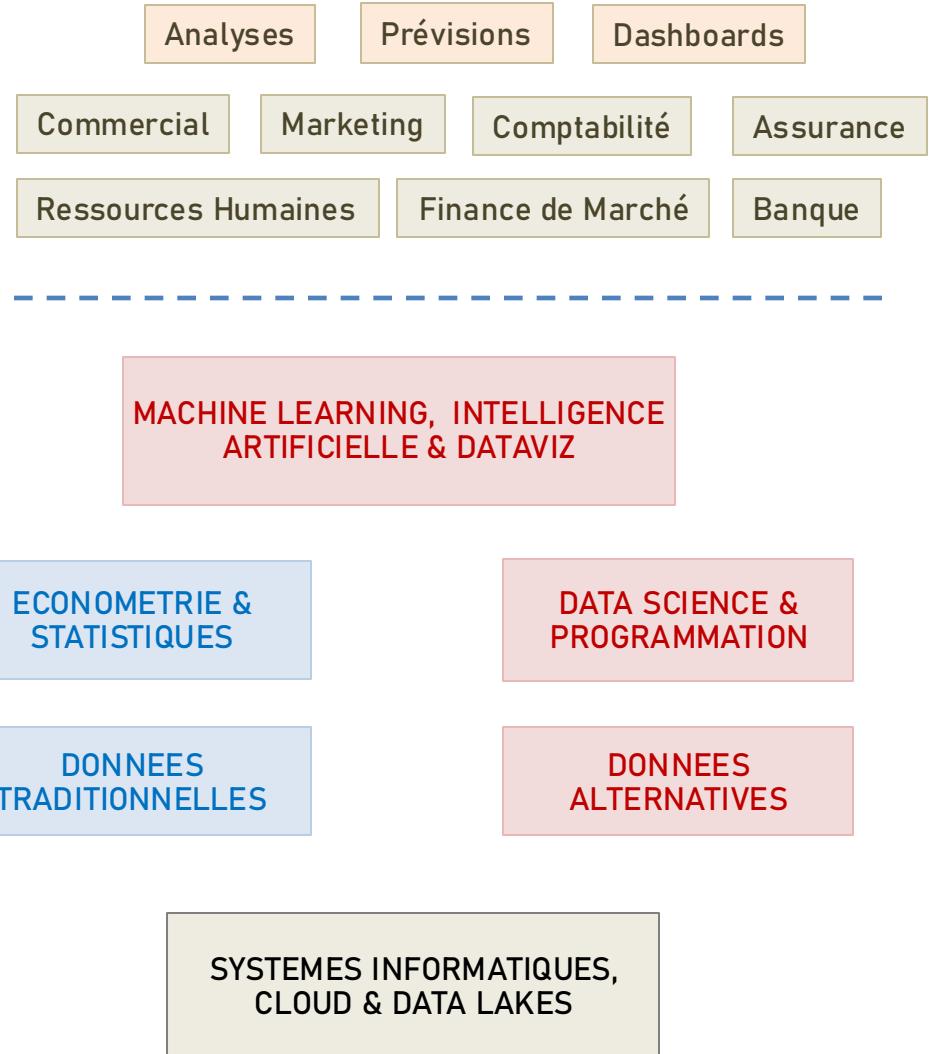
DATA & AI LANDSCAPE 2019



Is it useful for me?

**Even if you are not data-scientists, one day or another,
you will have to work with data-scientists**

Le syndrome de l'iceberg

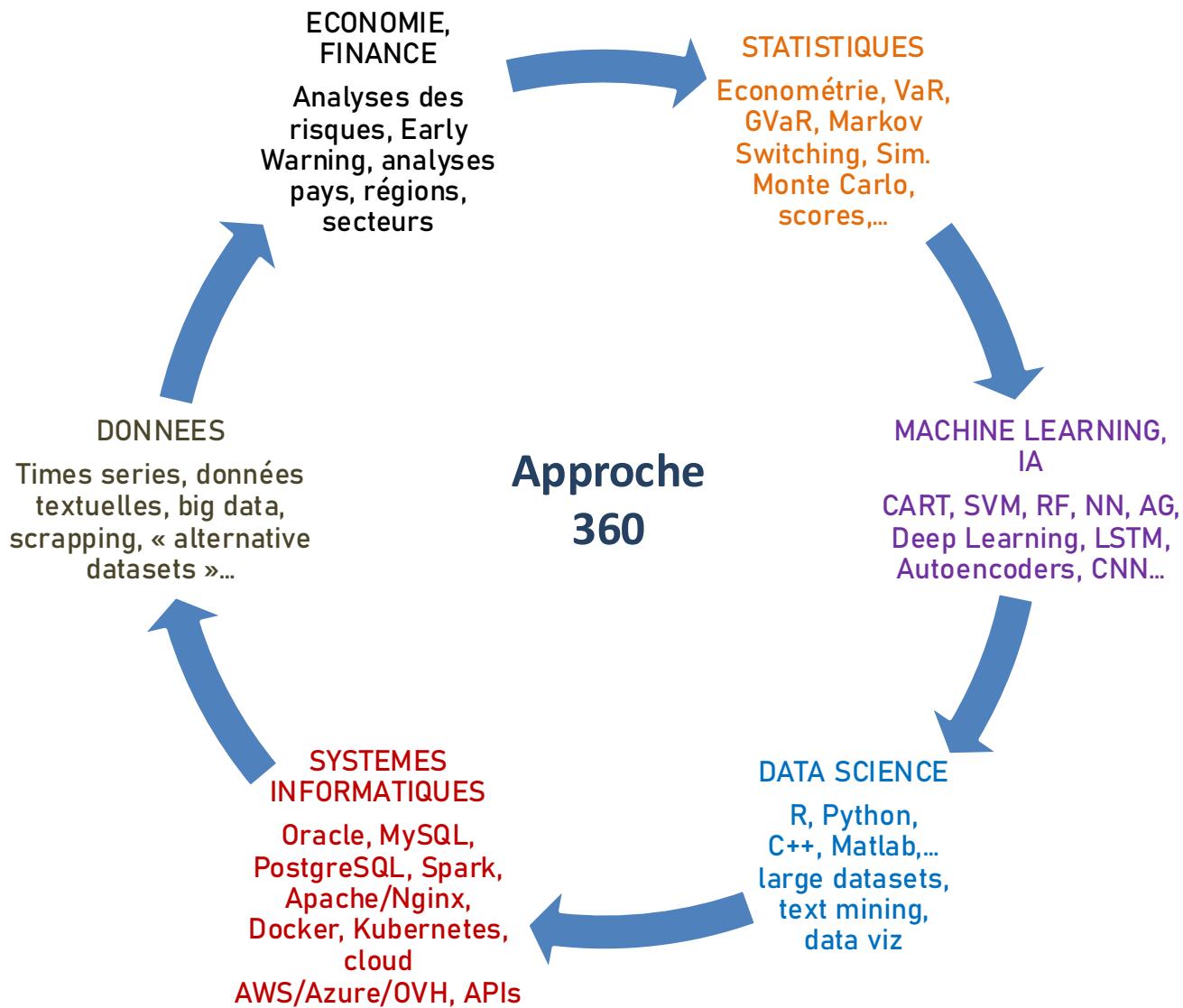


Les langages

Classement IEEE Spectrum Top Programming Languages 2021

Rank		Language	Type	Score
1	Python	🌐💻⚙️	100.0	
2	Java	🌐📱💻	95.4	
3	C	📱💻⚙️	94.7	
4	C++	📱💻⚙️	92.4	
5	JavaScript	🌐	88.1	
6	C#	🌐📱💻⚙️	82.4	
7	R	💻	81.7	
8	Go	🌐💻	77.7	
9	HTML	🌐	75.4	
10	Swift	📱💻	70.4	
11	Arduino	⚙️	68.4	

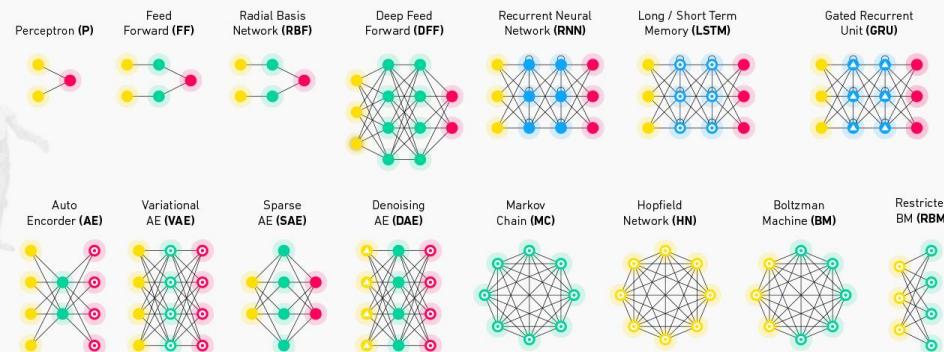
Approche / Méthodologie



L'écosystème du « deep learning »

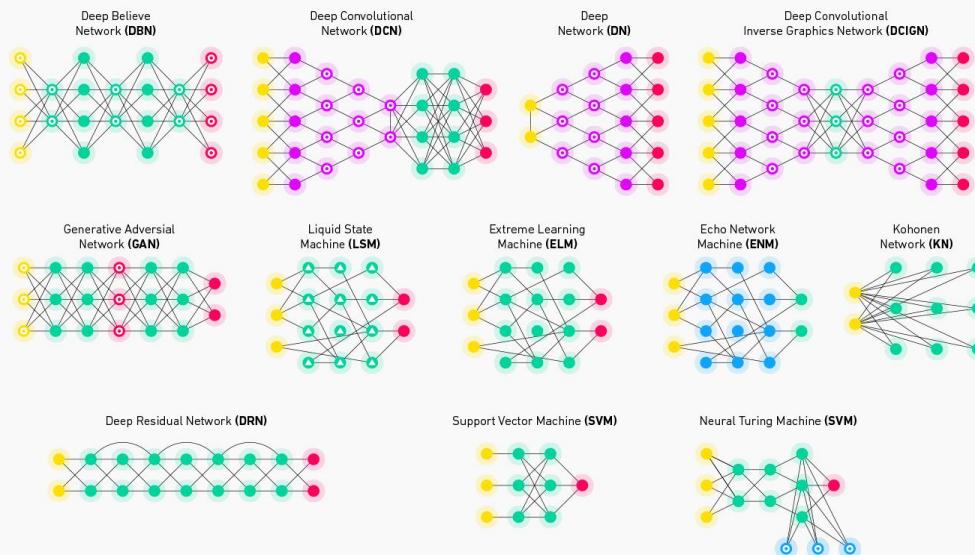
Neural Networks Basic Cheat Sheet

BecomingHuman.AI



Index

- Backfed Input Cell
- Input Cell
- Noisy Input Cell
- Hidden Cell
- Probabilistic Hidden Cell
- Spiking Hidden Cell
- Output Cell
- Match Input Output Cell
- Recurrent Cell
- Memory Cell
- Different Memory Cell
- Kernel
- Convolutional or Pool



Original Copyright by AsimovInstitute.org [See original here](#)

Mieux anticiper les crises économiques & financières sur les marchés émergents

Prévoir les crises économiques et financières (EWS)

- De nombreuses études empiriques sur le risque pays et les indicateurs avancés de crises économiques et financières: Krugman (1979), Obstfeld (1994), Cantor and Packer (1996), Eichengreen et al. (1996), Frankel and Rose (1996), Goldstein (1996), Goldstein and Turner (1996), Kaminsky and Reinhart (1999), Komulainen and Lukkarila (2003) , ...
- Mais malgré les classifications existantes, un très grand nombre de crises économiques et financières ont laissé les observateurs perplexes.
- Le manque de relations causales homogènes et des interactions complexes rendent l'identification ex-ante des facteurs de risque extrêmement difficile.

A quoi servent ces algorithmes ?

- Très performants pour identifier des patterns, des combinaisons critiques (non linéaires): cas du skieur, modèles à deux variables, signaux faibles et apprentissage en temps réel.
- Beaucoup d'entre eux sont presque entièrement « automatisables » (à la différence des approches économétriques traditionnelles)
- Attention au risque de sur-apprentissage, au « machine learning sauvage », aux risques de bases de données « poubelles » (données erronées, pas de notion d'intégration/stationnarité, liens économiques), phénomènes de « taches solaires », attention au sampling, etc...
- Certains algorithmes sont des « boites noires »... mais pas tous !

From econometry to data-science

Term in Statistics/Econometry	Equivalent in Machine Learning
Independent Variable, X	Input Feature, attribute, pattern
Dependent Variable, Y	Output Feature, response, label
In-Sample	Training Set
Out-of-Sample	Test Set
Estimate a Model	Learn a Model
Model Parameters	Model Weights
Regression	Supervised Learning
Clustering	Unsupervised Learning
Accuracy, R2	Sensitivity, Specificity, ROC, Likelihood

Supervised & Unsupervised Methods

SUPERVISED

Explain relationships between inputs and targets

Examples: economic crises, inflationary episodes, corporate defaults, etc....

Tools: econometry, linear discriminant analysis, neural networks/perceptrons, SVM, Random Forests, deep learning, etc...

UNSUPERVISED

Analyze a dataset that has not been « labelled »

Examples: analogies between countries, classifying companies (beyond sector & size), discovering macroeconomic patterns,

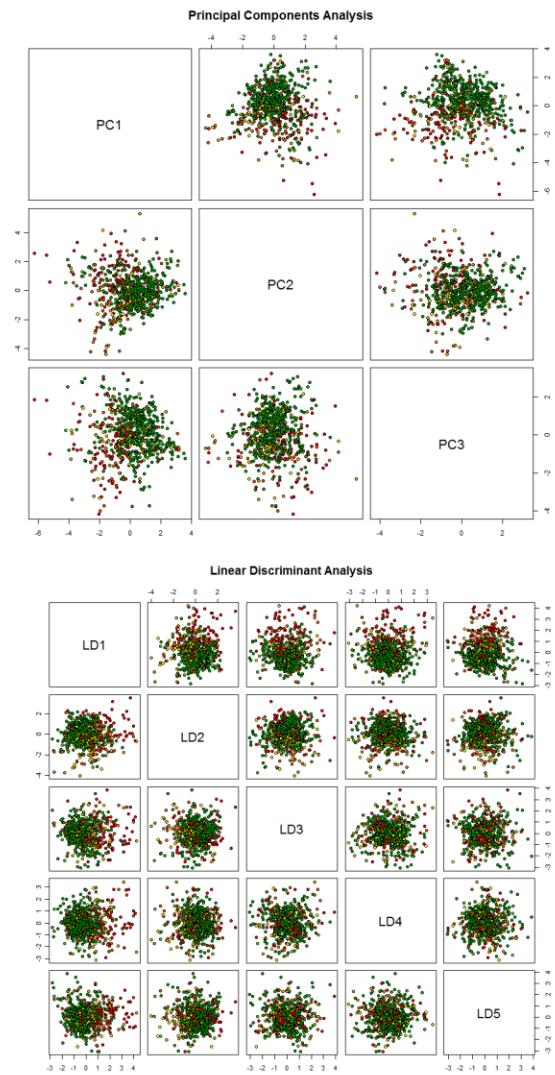
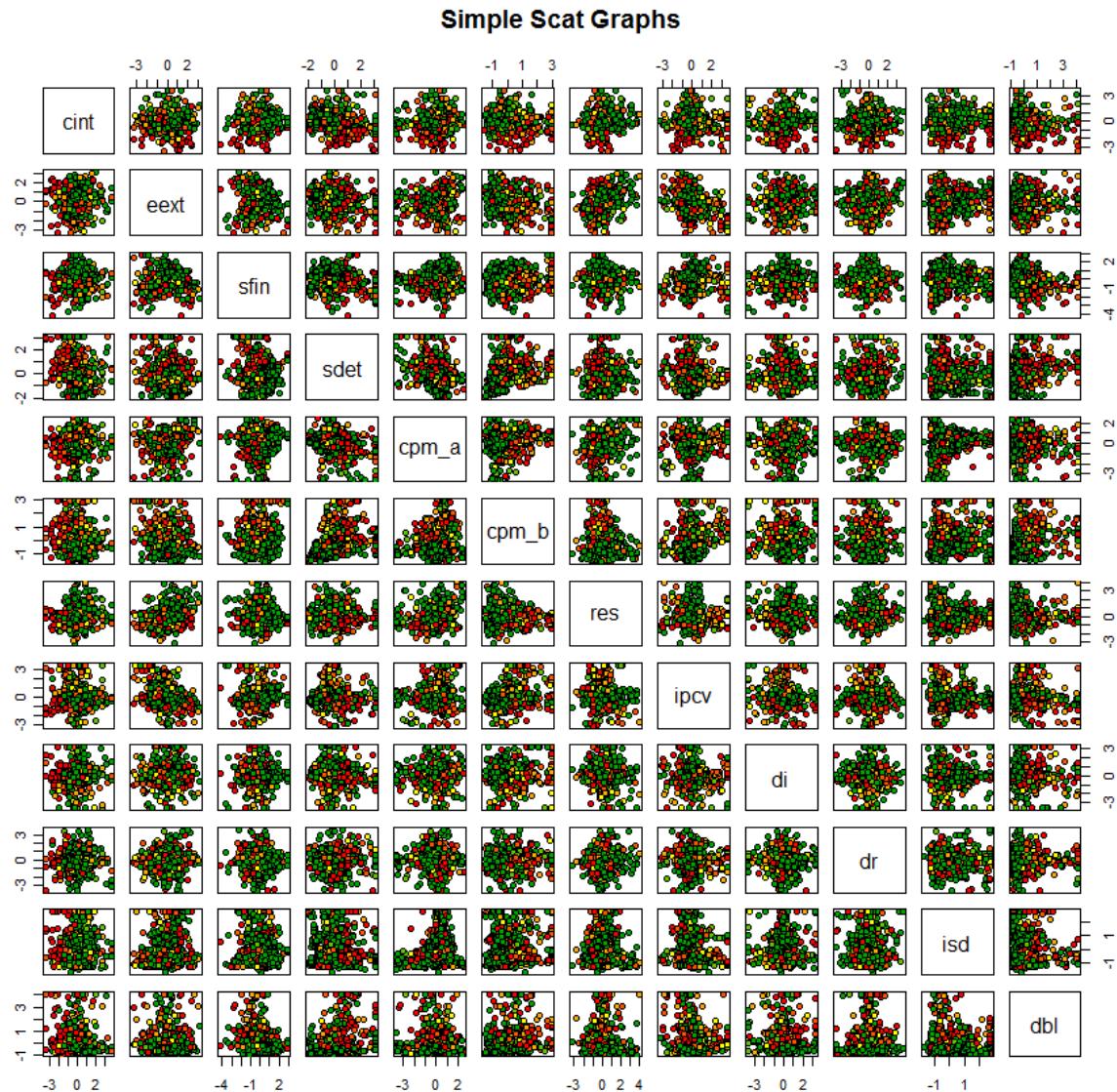
Tools: principal components analysis, k-means, clustering algorithms, self organizing maps, etc...

TAC ECONOMICS/RiskMonitor Fundamental Balances

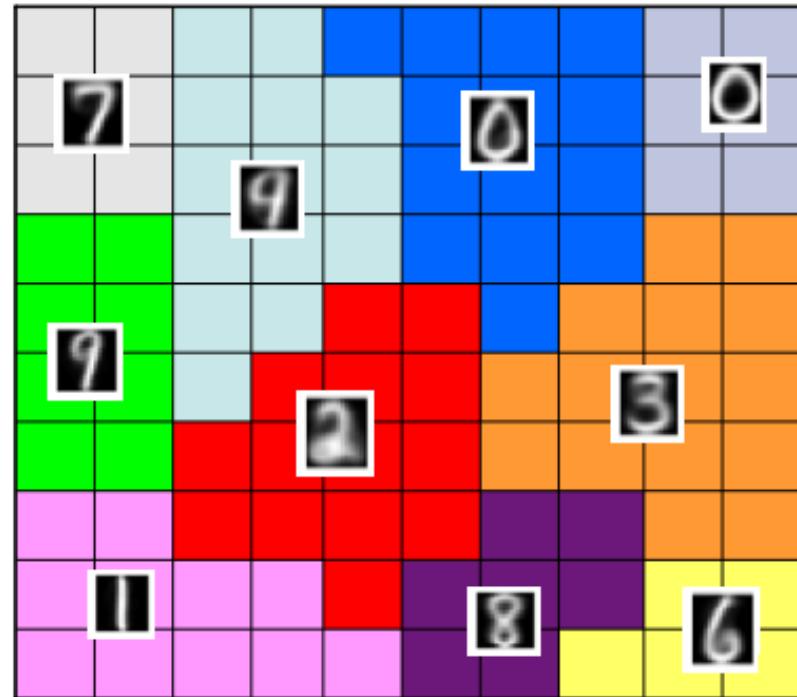
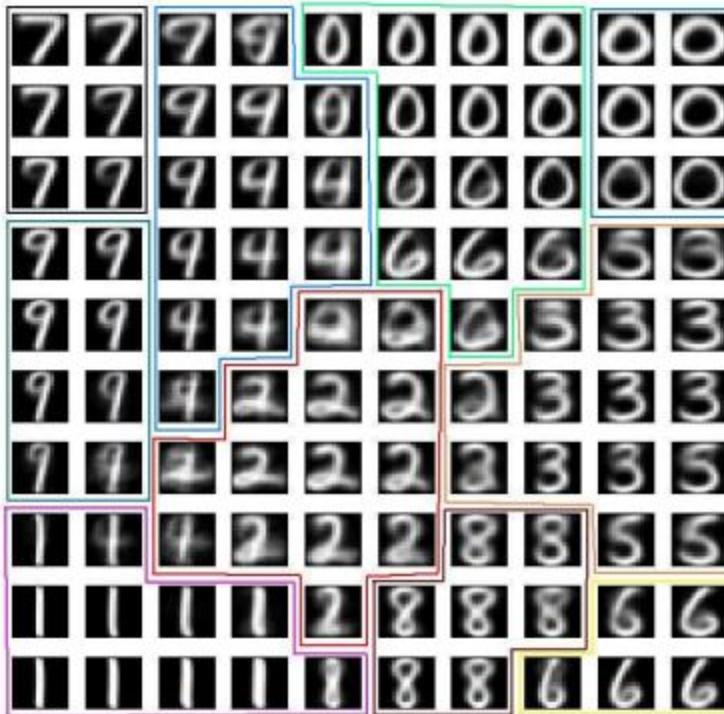
Macroeconomic indicators of the RiskMonitor fundamental balances

Indicator	Periodicity	Description
Economic growth	Annual	GDP growth
External balance	Annual	External balance sustainability
Financing stability	Annual	Stability of FDI inflows
Debt service	Annual	External financing
Forex liquidity	Quarterly	Foreign currency situation
Maximum potential service	Quarterly	Short-term foreign currency liabilities
Forex reserves quality	Quarterly	Dynamics in forex reserves
Exch. rate competitiveness	Quarterly	International competitiveness of exchange rate
Monetary pressure	Quarterly	Quality of monetary policy
Real economic pressure	Quarterly	Momentum of domestic activity
Domestic leverage	Quarterly	Activity and banks' health
Foreign financing	Quarterly	Dependence on foreign financing

Les méthodes traditionnelles



Self Organizing Maps... et Tessellations de Voronoi



Source

Data Analysis using Self-Organizing Maps

Marie Cottrell and Patrick Letrémy

https://samos.univ-paris1.fr/IMG/pdf_Porvoo_Kohonen_Data_Analysis_V3-2.pdf

World Poverty Map using Self Organizing Maps

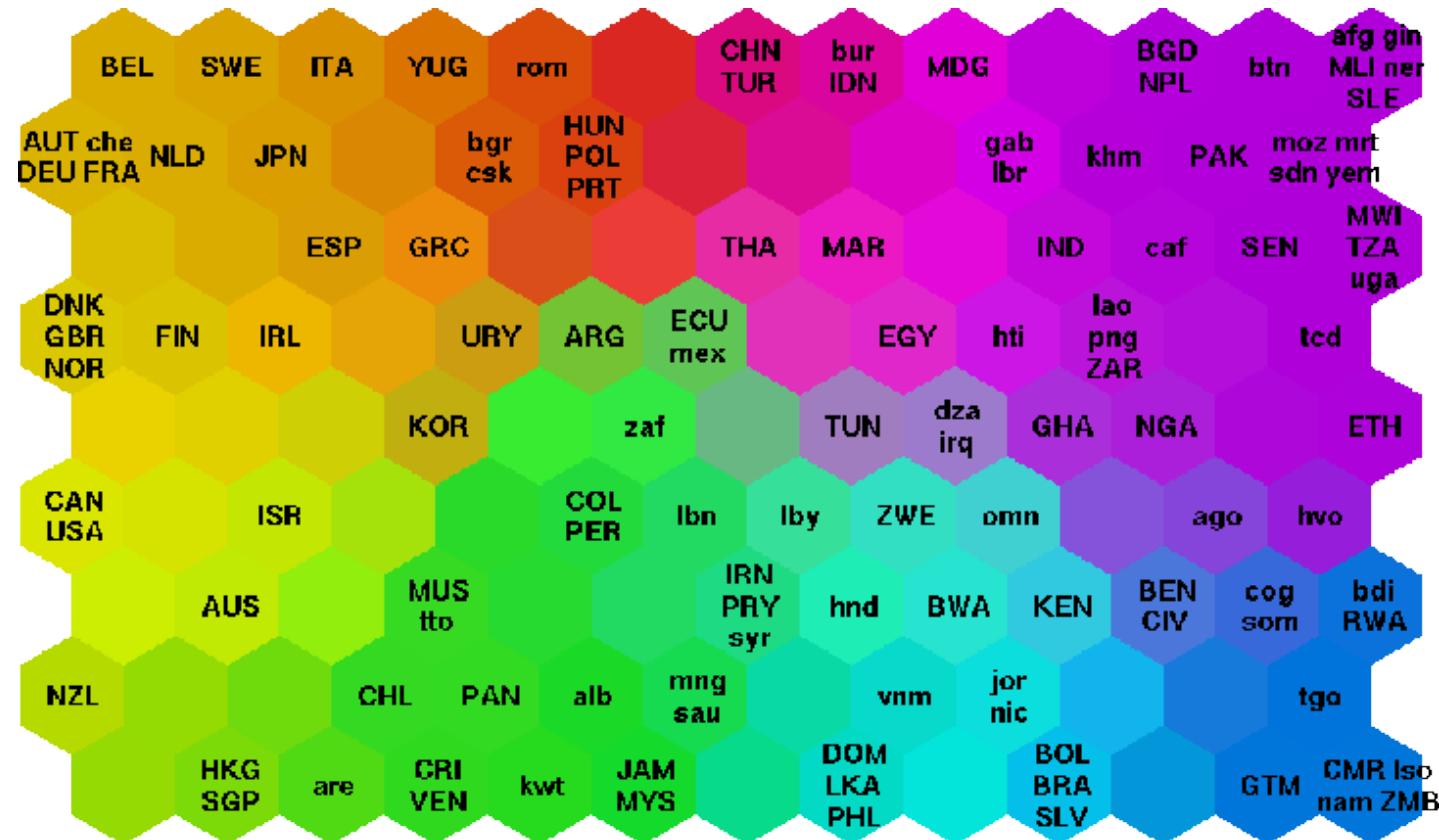
An example on World Bank data, by Samuel Kaski

- 39 World Bank country indicators describing various quality-of-life factors, such as state of health, nutrition, educational services, etc, were used.
- The complex joint effect of these factors can be visualized by organizing the countries using the self-organizing map.
- The map consists of a regular grid of processing units, "neurons". A model of some multidimensional observation, eventually a vector consisting of features, is associated with each unit.
- The map attempts to represent all the available observations with optimal accuracy using a restricted set of models. At the same time the models become ordered on the grid so that similar models are close to each other and dissimilar models far from each other.

World Poverty Map using Self Organizing Maps

An example on World Bank data, by Samuel Kaski

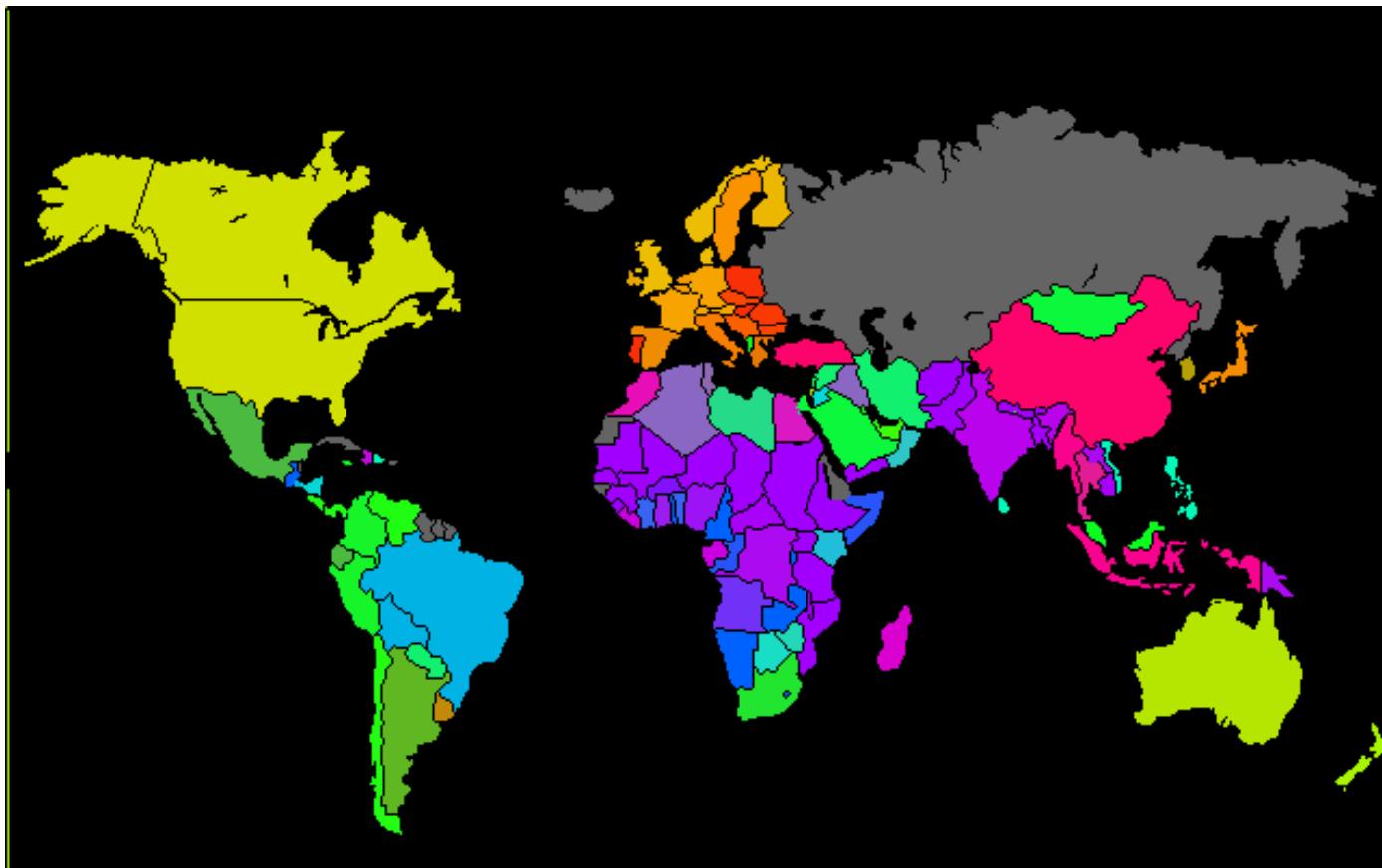
each country is automatically assigned a color describing its poverty type in relation to other countries



Source: <http://www.cis.hut.fi/research/som-research/worldmap.html>

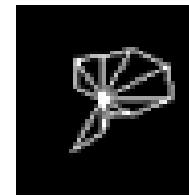
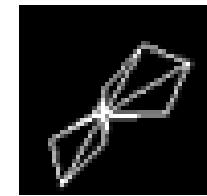
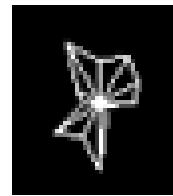
World Poverty Map using Self Organizing Maps

An example on World Bank data, by Samuel Kaski



Source: <http://www.cis.hut.fi/research/som-research/worldmap.html>

Des pays, des indicateurs, des années... et des papillons

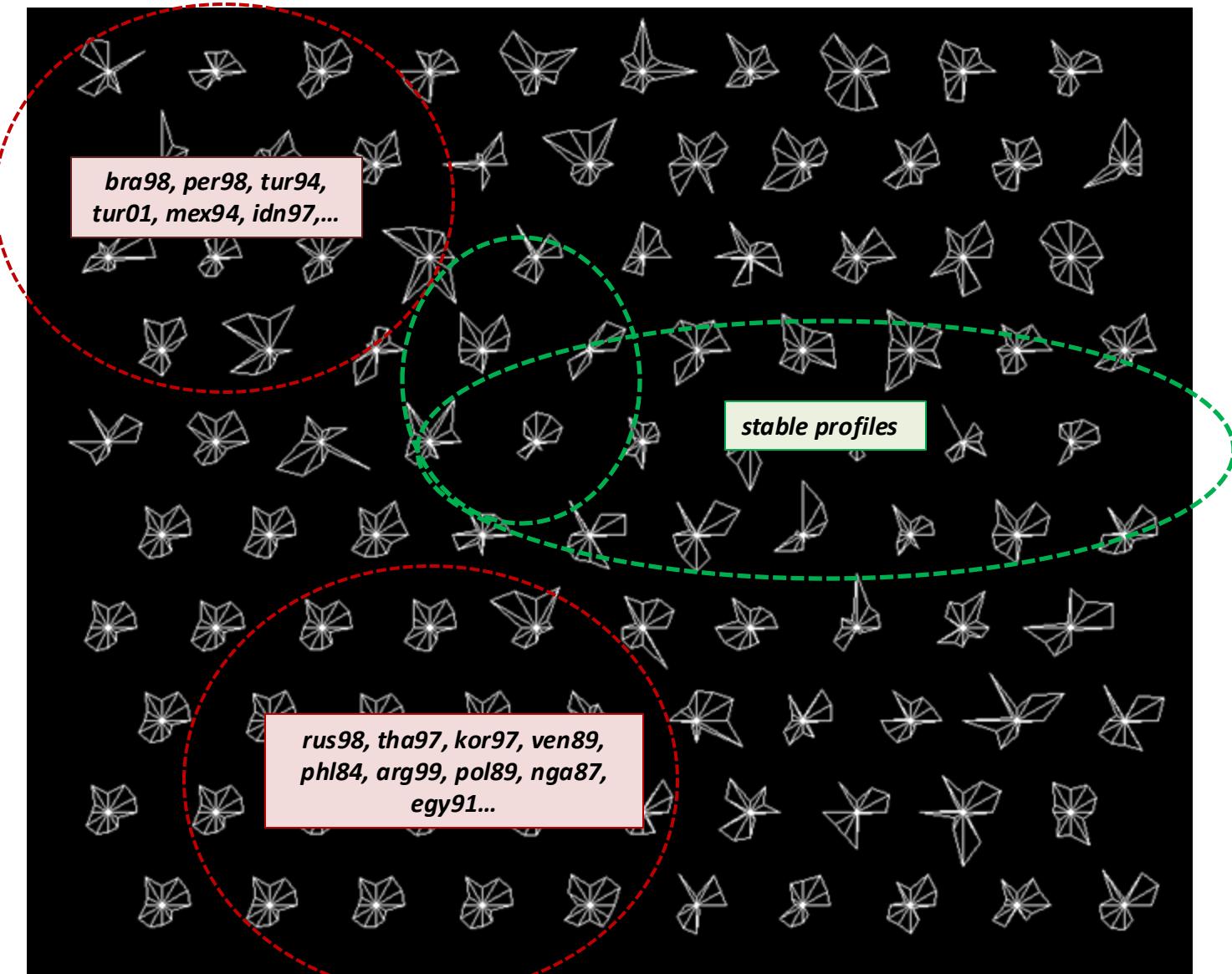


on more than 100 countries since the 80s

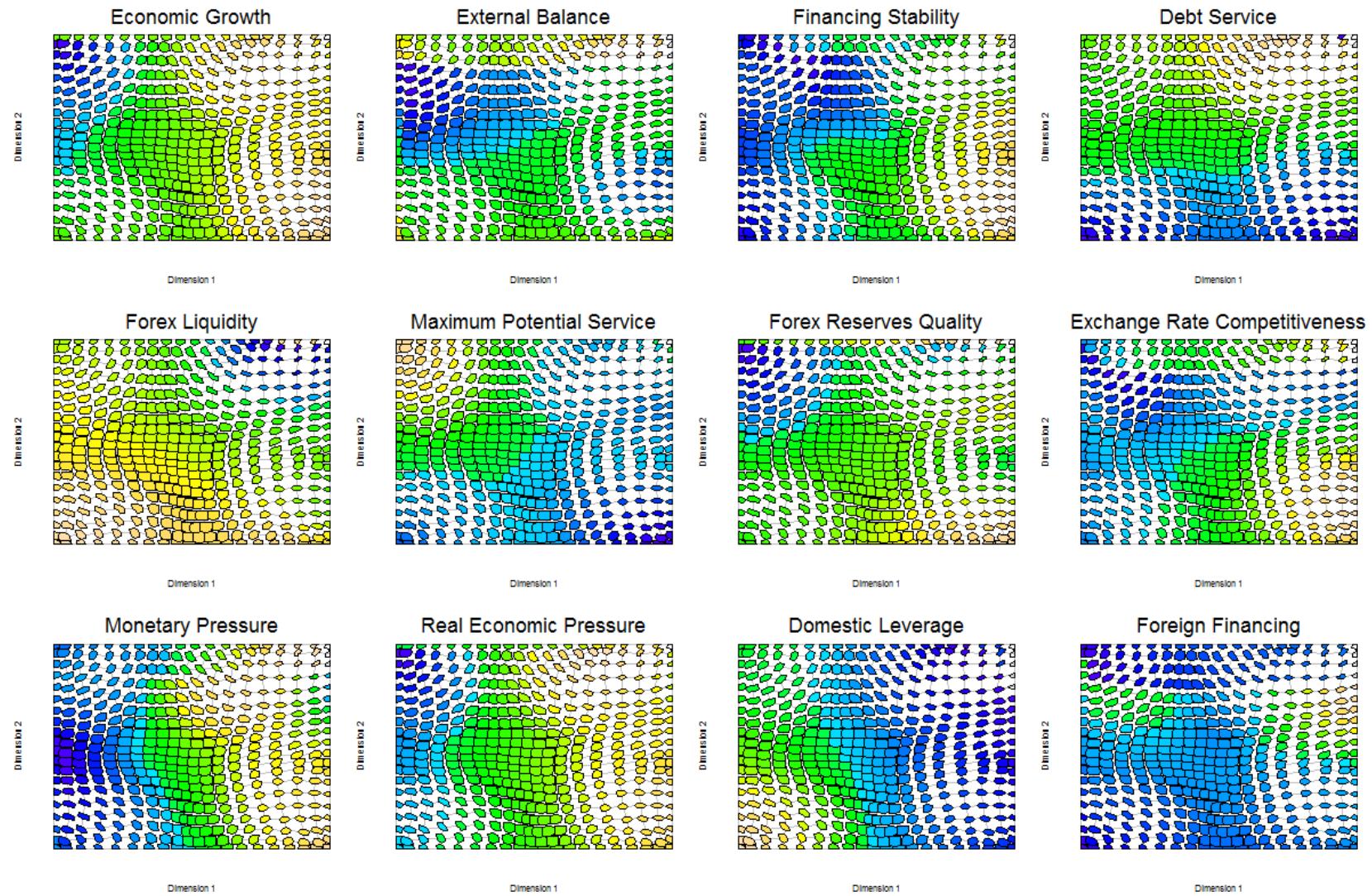
100 countries x 40 years x 4 quarters per year

16,000 country web charts to analyse!

Similarités macroéconomiques et réseaux de neurones



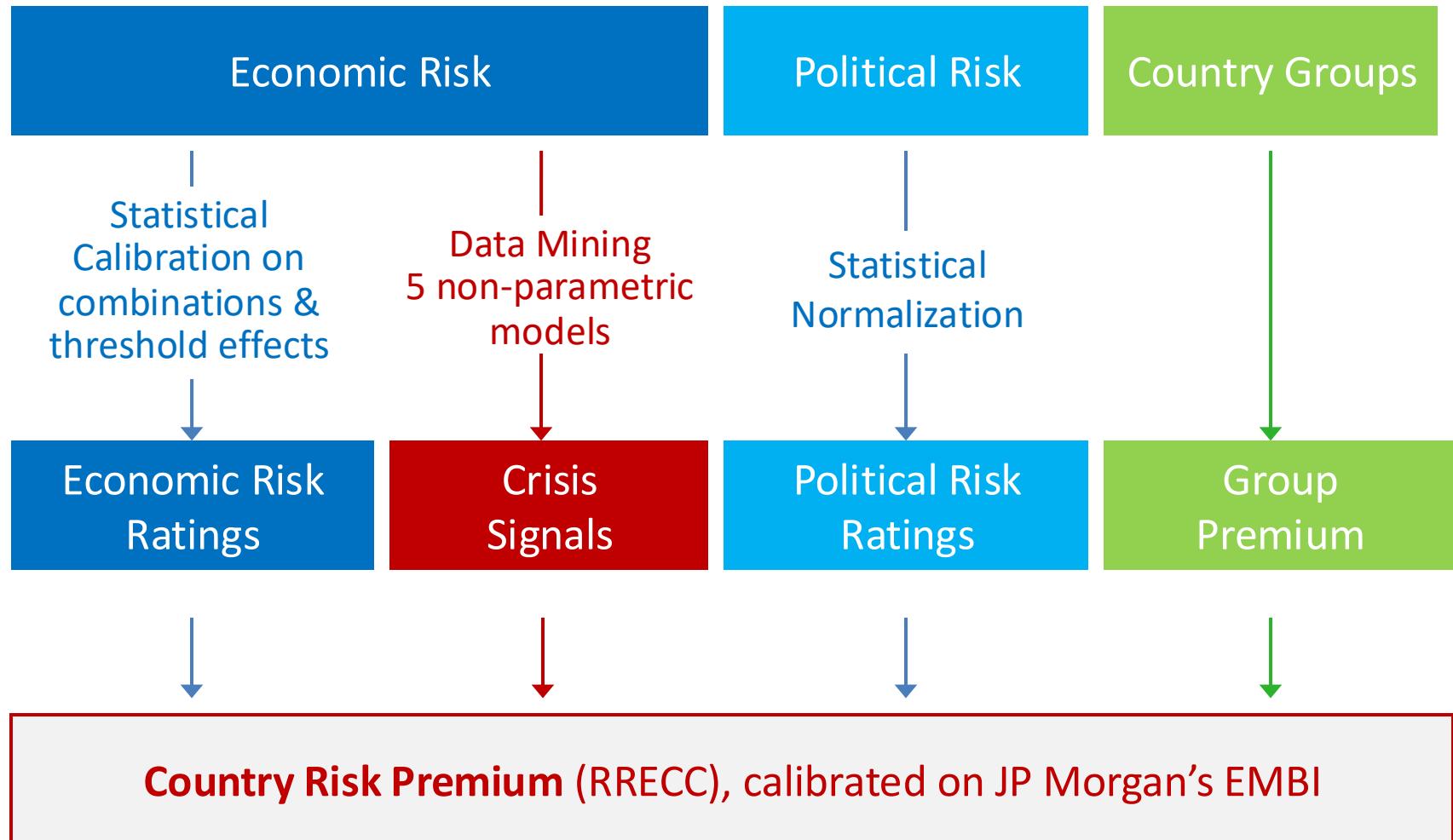
Influence des indicateurs différente selon le groupe



Performances des outils de EWS/IA développés

Avec moins d'une vingtaine d'indicateurs macroéconomiques « bien choisis » et une séquence de modèles d'IA relativement sophistiquée, nous arrivons ainsi prévoir près de 90% des crises économiques et financières, jusqu'à 2 ans à l'avance

Méthodologie RiskMonitor



ClientGate

The ClientGate dashboard provides a comprehensive view of Brazil's risk profile and recent economic data.

Risk Premium: 563bp (basis points)

Economic Risk: 47 (Average, Region: 45, World: 45)

Political Risk: 51 (High, Region: 50, World: 54)

Early Warning Signals: Currency (Watch List)

Recent Data:

GDP GROWTH (Percentage Y/Y)	INFLATION RATE (Percentage Y/Y)	EXCHANGE RATE (LC per USD)	FOREX RESERVES (USD)	POLICY (Percentage)
1.1% As of 2018 1.1% in 2017	4.7% As of May 2019 2.9% in May 2018	3.85 As of Jul 02, 2019 3.91 in Jul 2018	372bn As of May 2019 365bn in May 2018	6.5% As of Jun 2019

IN THE NEWS:

- Thousands Of Anti-corruption Protesters Rally In Brazil In Support Of Justice Minister Sergio Moro Wake Of 'car Wash' Probe (Jul 1st, 2019 - SCMP)
- Innovative Time-share Model In São Paulo Allows Apartment To Have Several Owners (Jun 29th, 2019 - Rio Times)

LATEST PUBLICATIONS:

- COUNTRY SNAPSHOT BRAZIL** A Monthly
- Exchange Rate Index (average for 10 emerging markets, weighted by GDP [LC per USD])
- Exchange Rate (y/y growth of LC per USD (positive + depreciation), Last Data Available)
- Turkey: Adjustment Spoiled By Politics Leading To Further Troubles For The Corporate Sector And Currency Risks, Introducing RiskMonitor's Monthly Dashboard & Heatmap, A Visual Check On All Key Changes On Our Quantitative Outputs
- Flash Comment ARGENTINA: Argentina: Economic Pressures On Tight Political Agenda (Jun 14th, 2019) PDF
- Flash Comment INDIA: India: PM Modi Reelected Amidst Economic Slowdown (Jun 5th, 2019) PDF
- Flash Comment ARGENTINA: Argentina: Economic Pressures On Tight Political Agenda (May 29th, 2019) PDF
- Flash Comment NAMIBIA: Namibia: A High Power Of Attraction For A Small And Almost Peaceful Nation | Haiti: Half Of Hispaniola Suffering | India-Pakistan-Afghanistan: Haunting Borders (May 16th, 2019) PDF

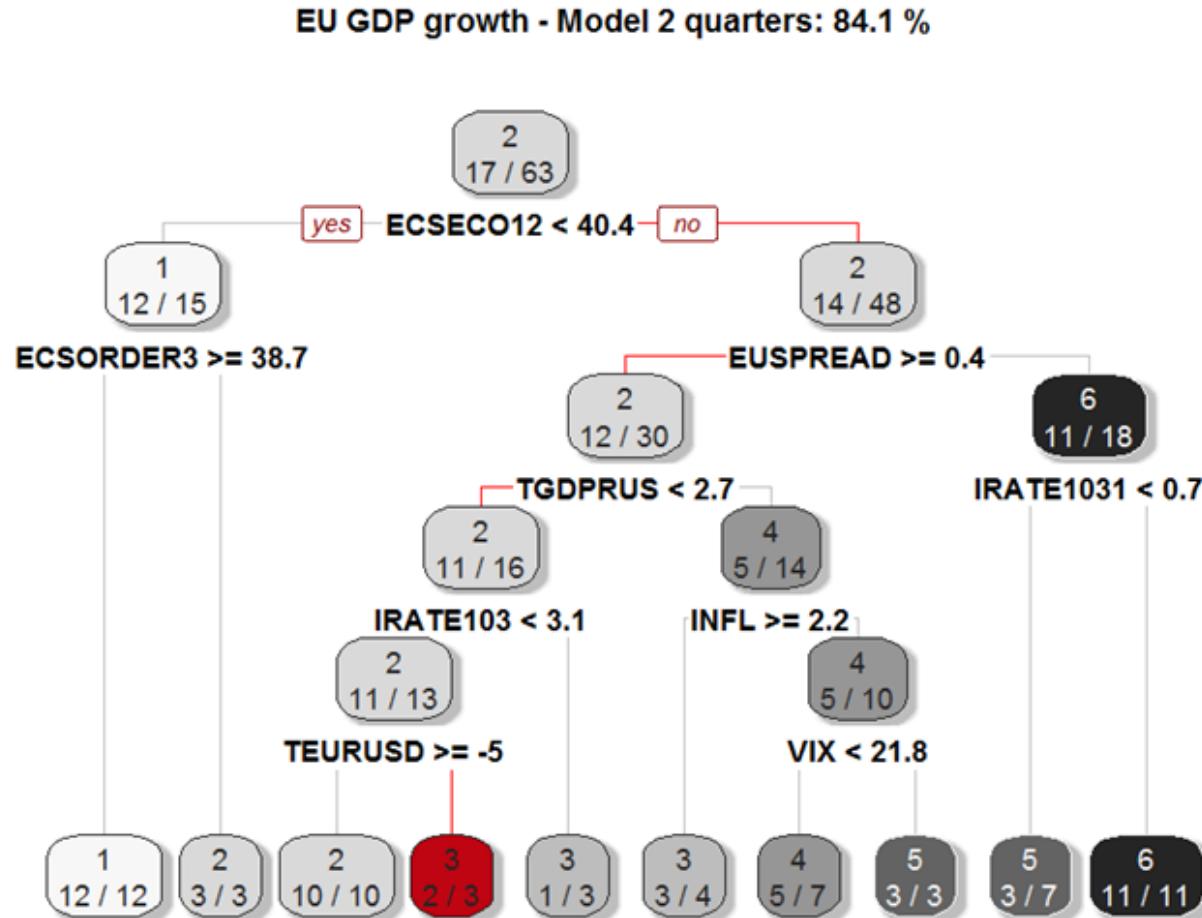
Publications: Publications, Countries, Data & Charts, Search..

Countries: Brazil

Data & Charts: Activity, Prices, Rates, Trade

Search: Search..

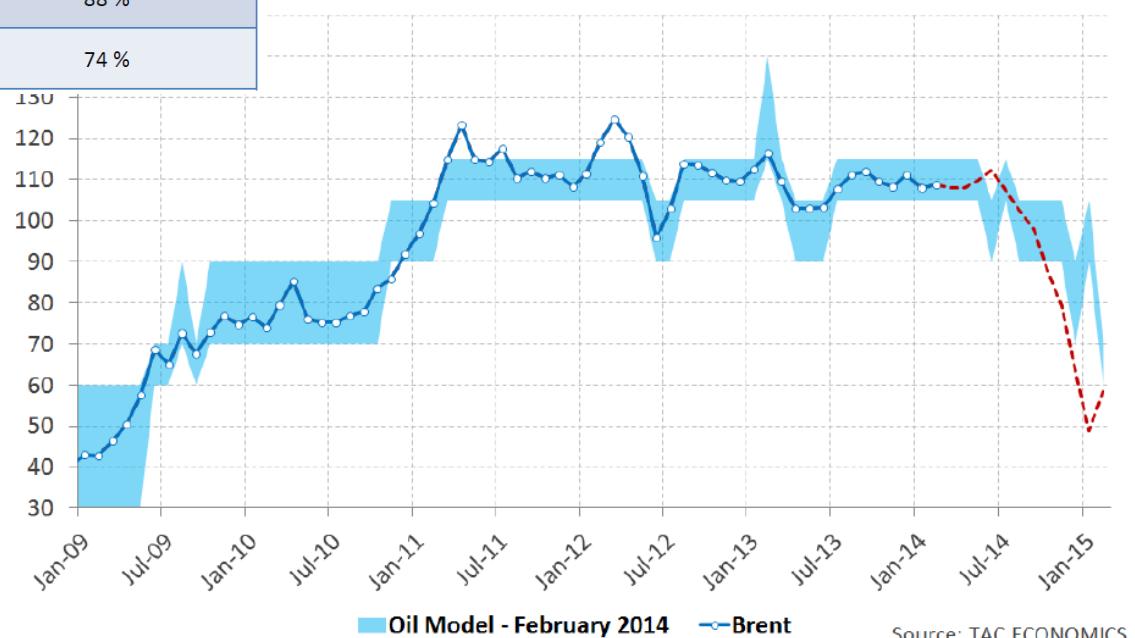
CART Applied to Real GDP Growth Forecast



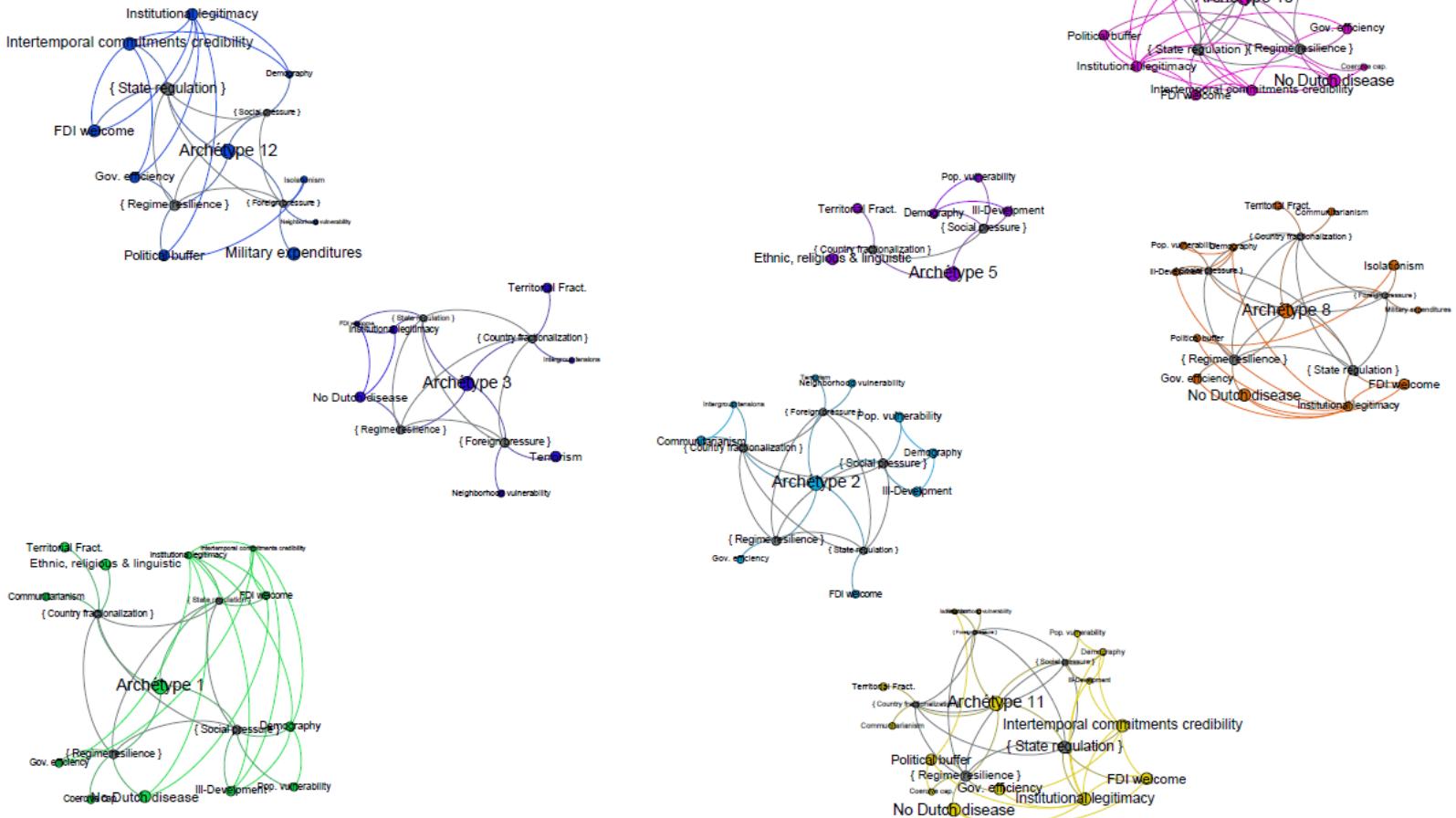
Machine Learning on Oil Price Forecasts

	Average accuracy (training dataset)	Average accuracy (testing, dataset 1)
Naive Bayes	96 %	88 %
Tree Bagging	99.9 %	90 %
Gradient Boosted Machine	100 %	90 %
Supervised SOM	86 %	75 %
Neural Network multilayer perceptron	82 %	74 %
Random Forest	100 %	90 %
Support Vector Machine	96 %	88 %
k-nearest neighbors	84 %	74 %

TAC Brent short-term Projections (\$/bl)



Archetypal of Political Risk



A dark blue background featuring a complex network of light blue dots connected by thin lines, forming a mesh-like structure.

Prévoir les risques de ruptures sur les marchés financiers

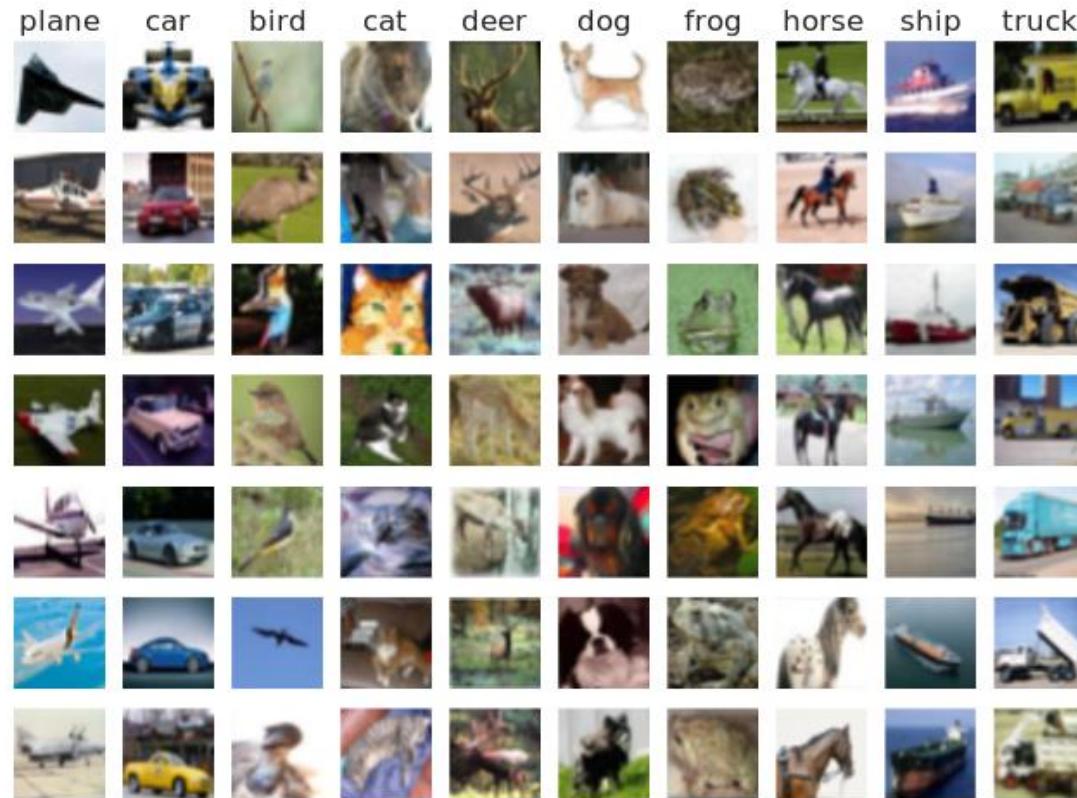
Petit retour sur les travaux historique

- Réseaux de neurones en finance d'abord comme substitut aux systèmes experts, dans les années 90.
- Prévision des cours boursiers dès Kimoto et Yoda (1993), sur le Tokyo Stock Index avec... 5 variables d'entrée ! Mais aussi sur l'or ou le S&P500 (Grudnitski et Obsburn, 1993, Quang Do 1995).
- Mais aussi des prévisions de taux de change, dès le milieu des années 90 (Rawani 1993, Azoff 1994, Avouyi Dovi 1995).
- Depuis, de très nombreux travaux sur l'utilisation de méthodes de machine learning en finance.

Les plus gros « hedge funds » du monde

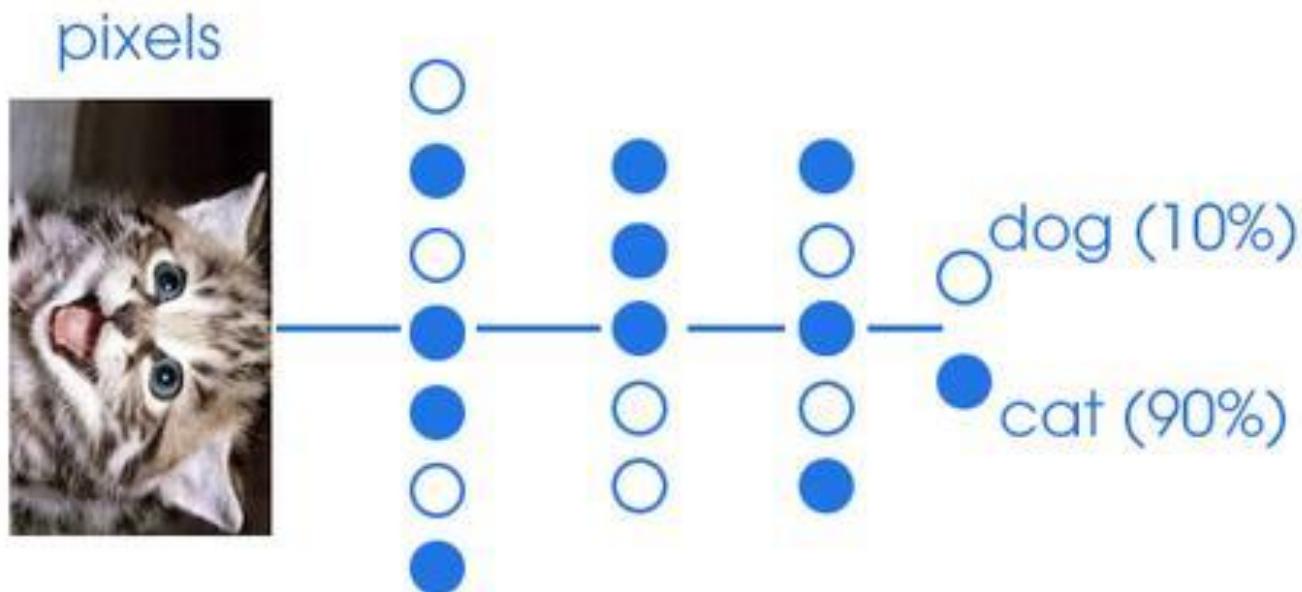
	Actifs sous Gestion en mds de \$ en 2017	Quantitatif ?
Bridgewater Associates	122.2	(Non)
AQR Capital Management	69.6	Oui
JPMorgan AM	45.0	(Oui)
Renaissance Technologies	42.0	Oui
Two Sigma	38.9	Oui
De Shaw & Co	34.7	Oui
Man Group	33.9	Oui
Millennium Management	33.9	Oui
Och-Ziff Capital Management	33.5	(Oui)
Winton Group	32.0	(Oui)

Computer Vision

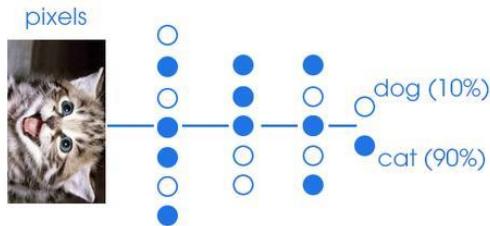


Source: dominodatalab

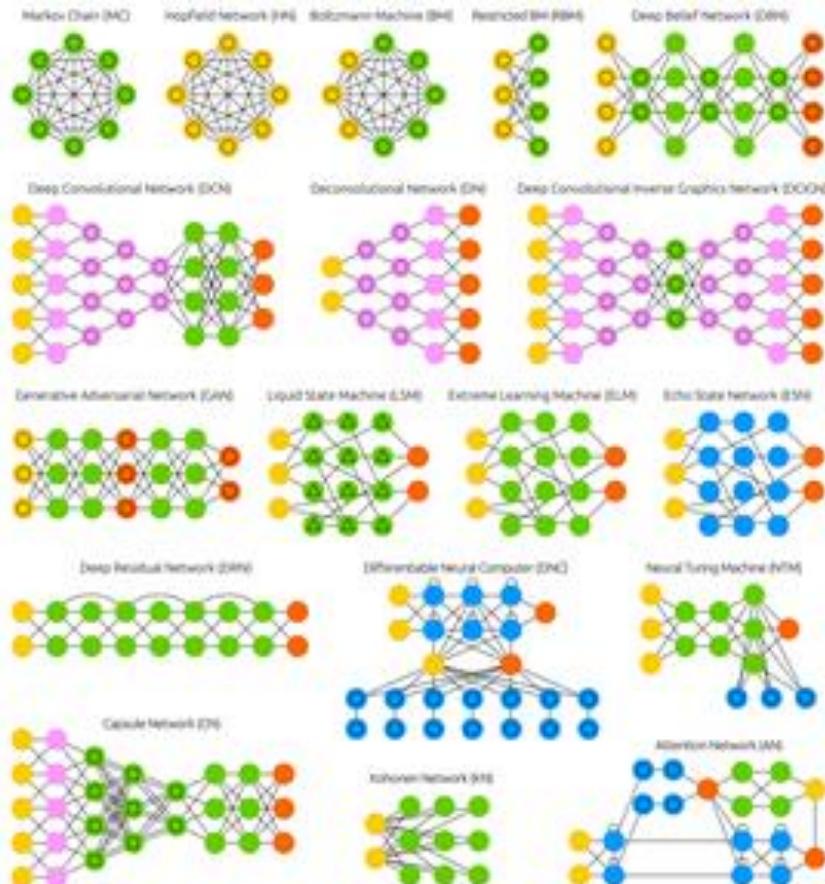
Réseaux de neurones « traditionnels »



Du perceptron au « deep learning »



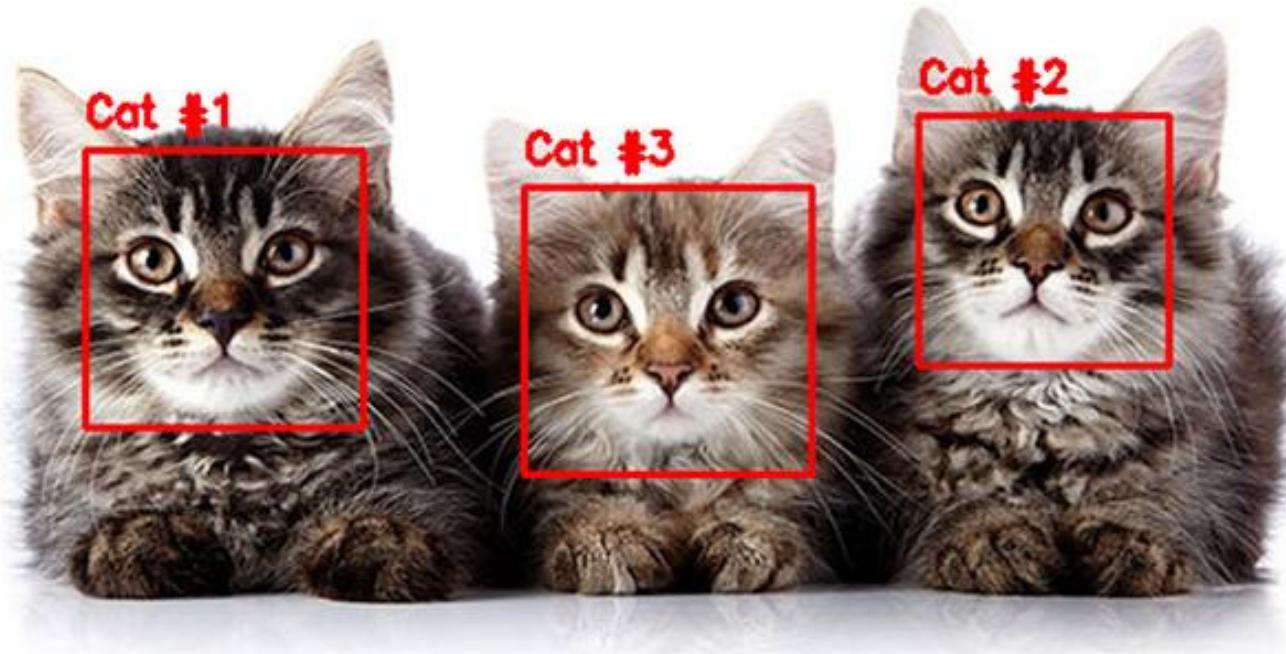
Le Neural Network Zoo du Asimov Institute



...parfois compliqué: distinguer les chiens des muffins



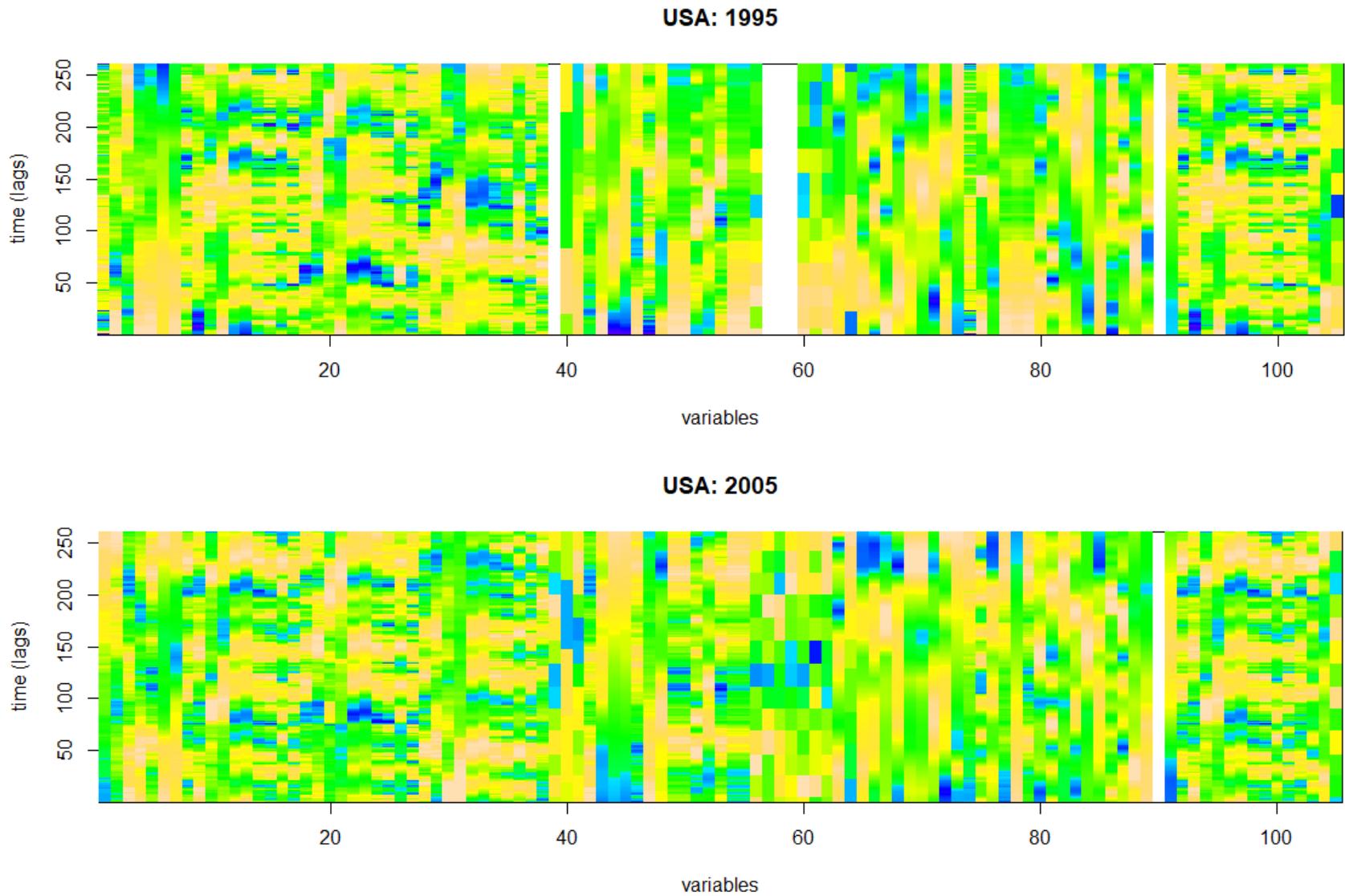
La convolution ou la « révolution du petit chat »...



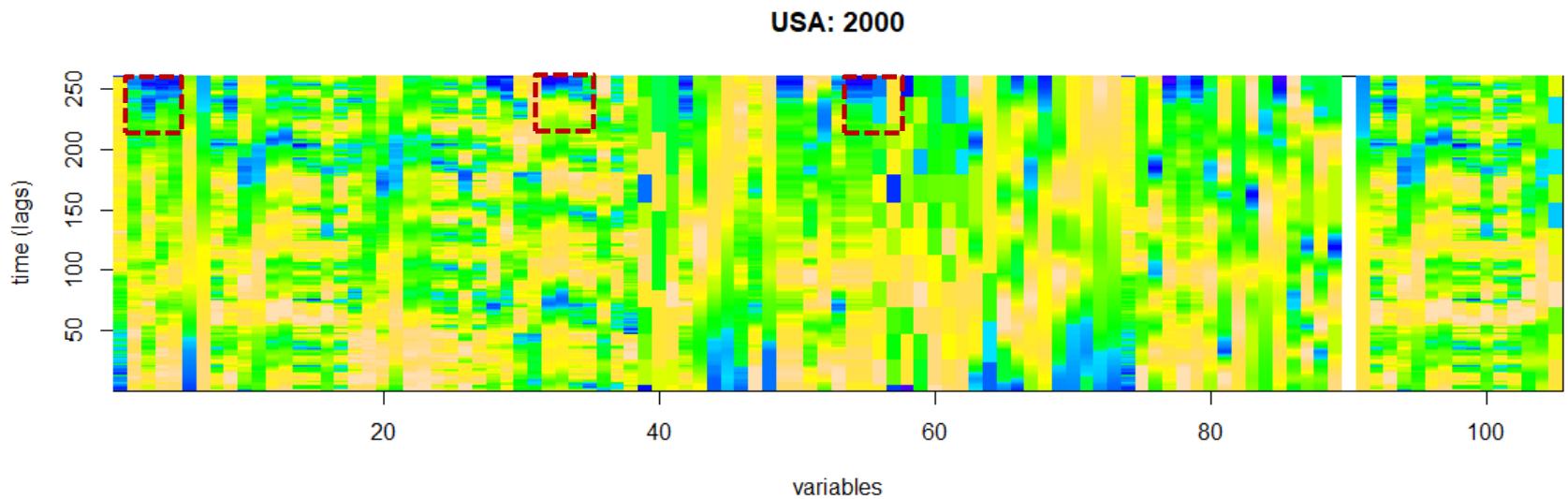
Quantitative Market Alert (QMA)

- Outil de détection des tendances et des ruptures sur plus de 20 marchés internationaux (actions, obligations, spreads corporates).
- Signaux sur les tendances, et calibrage automatique des poids des différents actifs dans un portefeuille.
- Possibilité d'ajuster les fonctions objectifs sur la base de cibles « combinées ».
- Outil mixant à la fois des outils de machine learning « simples », à des outils plus puissants de deep learning appliqués à des échantillons de plus de 200 données macroéconomiques et financières sur des périodes de 40 jours (identification de « patterns »).
- Difficulté majeure: bien gérer les échantillons, et l'instabilité des performances historiques.

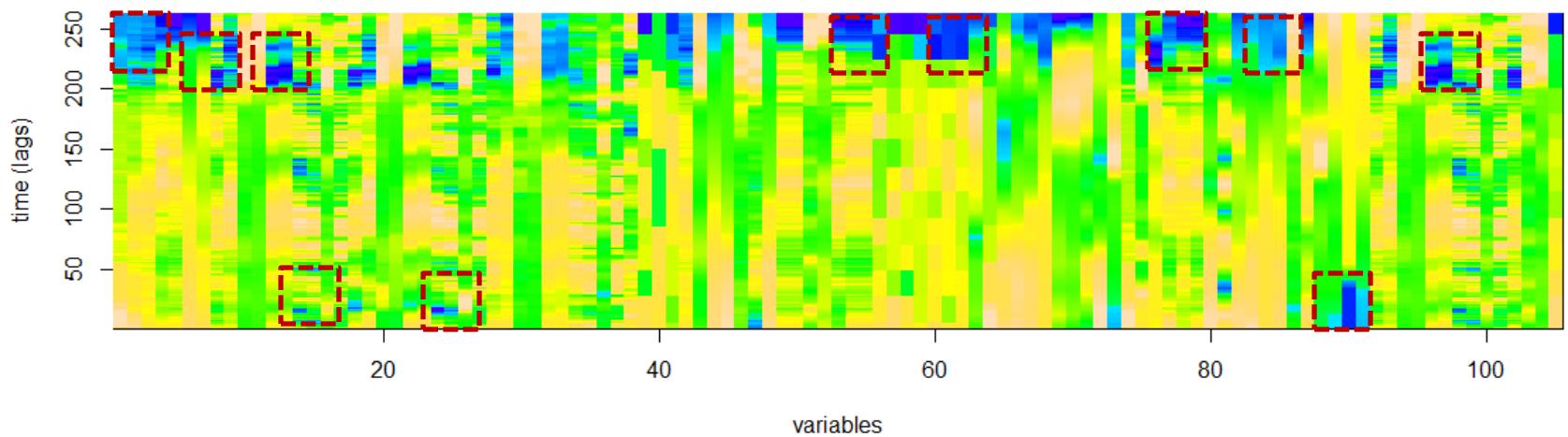
« Patterns » du QMA et convolution



« Patterns » du QMA et convolution



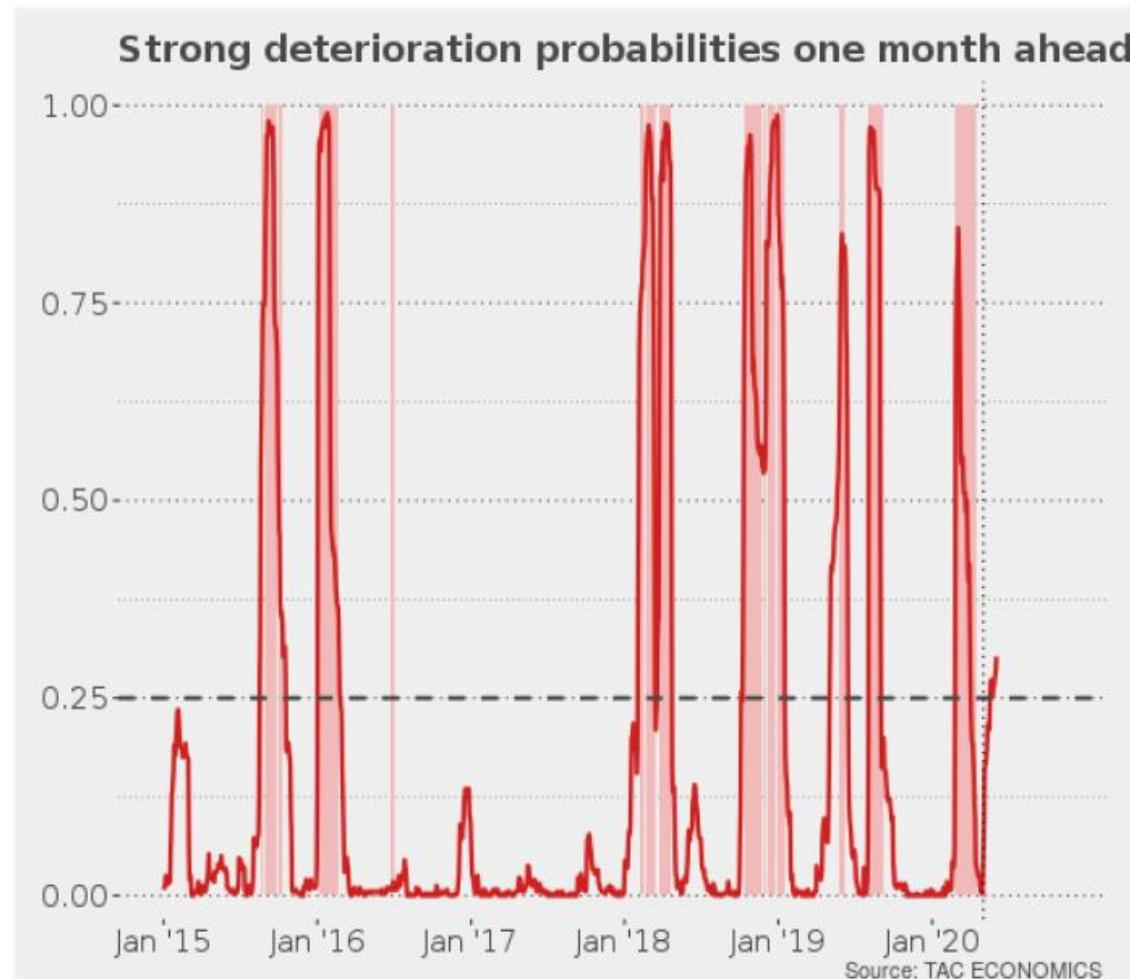
USA: 2008



Performances récente du QMA : les Fair Values en Jan. 2020

	Market Level (Jan. 03)	Fair Value November 19	Gap to Fair Value	Expected short-term direction
Equity Indices				
S&P 500	3 235	2 849	386	Decrease
CAC 40	6 044	5 344	700	Decrease
DAX 30	13 219	12 103	1116	Decrease
FTSE 100	7 622	7 268	354	Decrease
Nikkei 225	23 657	21 578	2079	Decrease

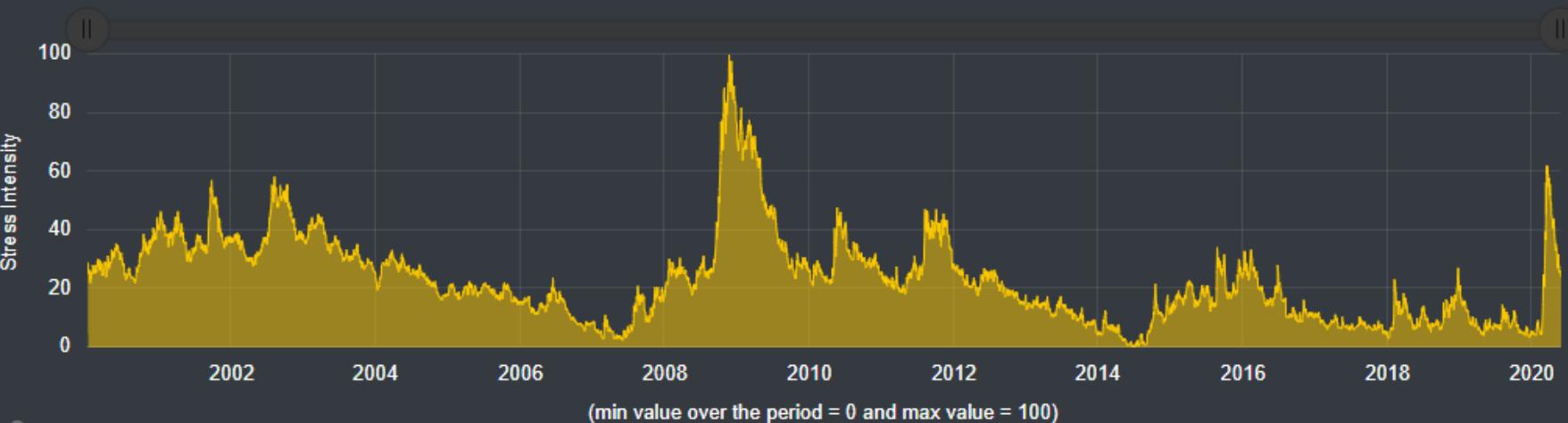
Performances et signaux du QMA sur le S&P500



Financial Index « Covid-19 »

Financial Index

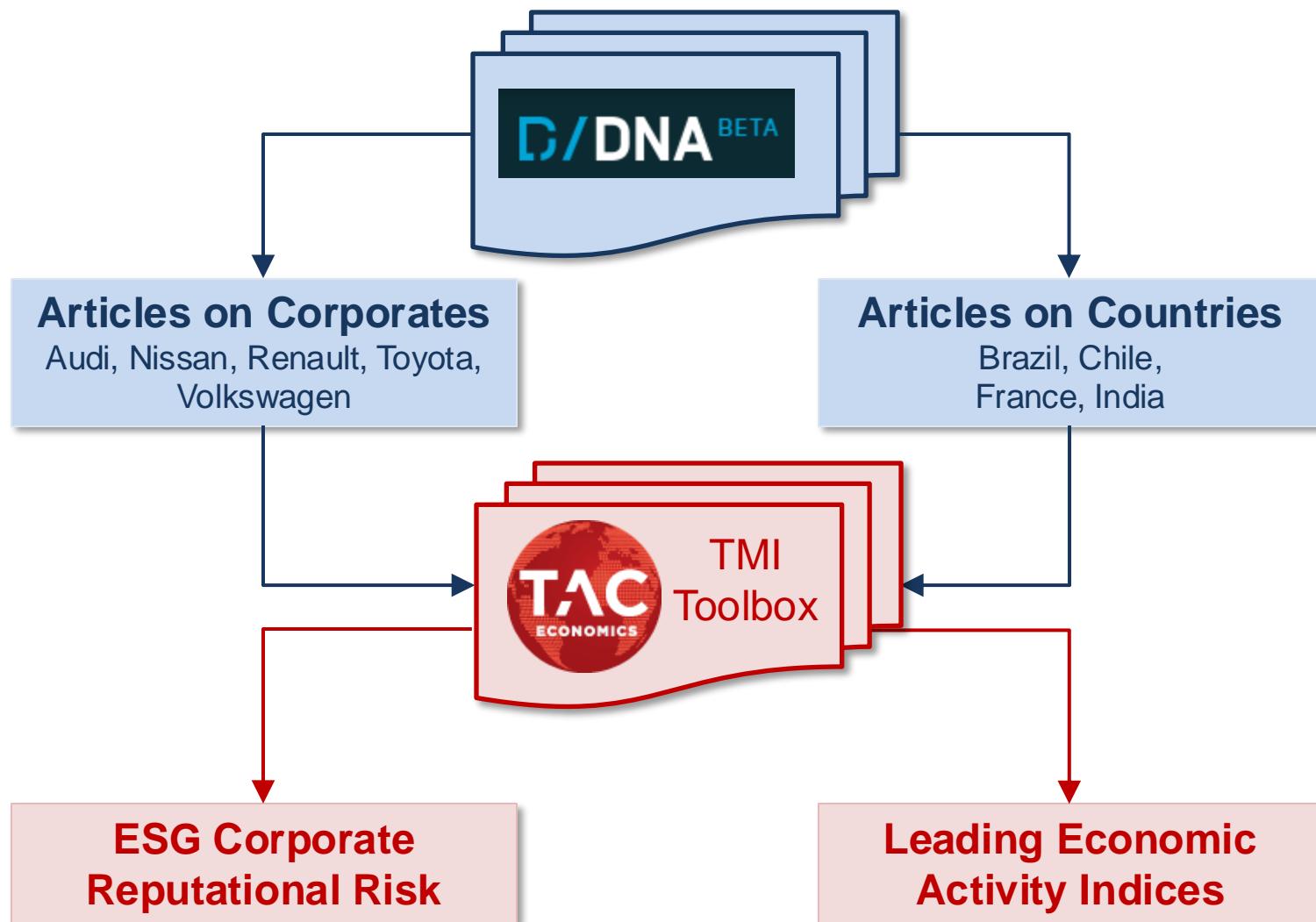
The Financial Index is calculated using long-run measures on historical drawdowns on financial markets, corporate and sovereign spreads, the VIX index and a volatility of the EUR/USD parity.



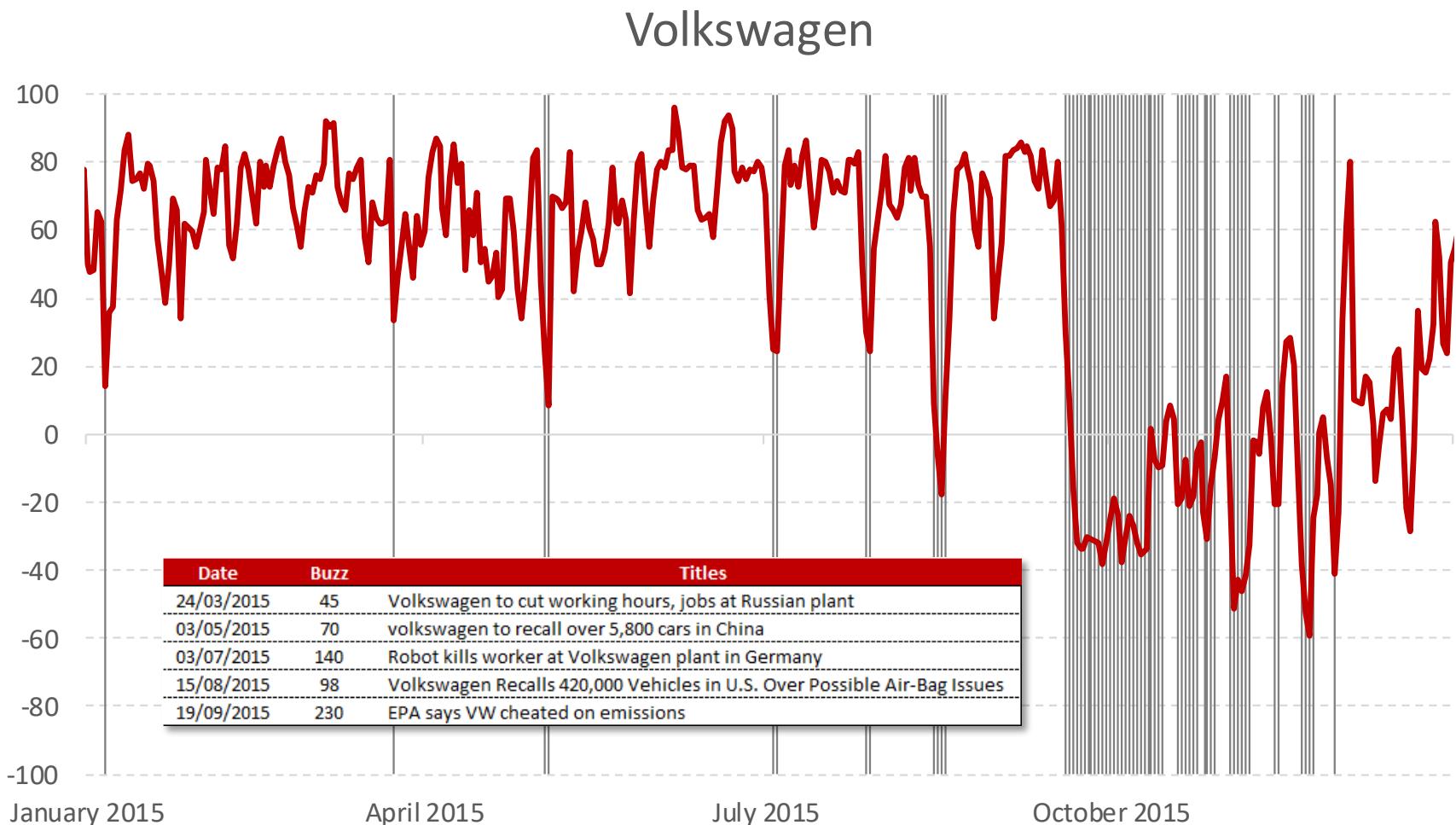
<https://www.taceconomics.com/covid19/>

Text mining, indicateurs de sentiment et identification de thématiques

Text mining, topics, sentiments et risque réputationnel

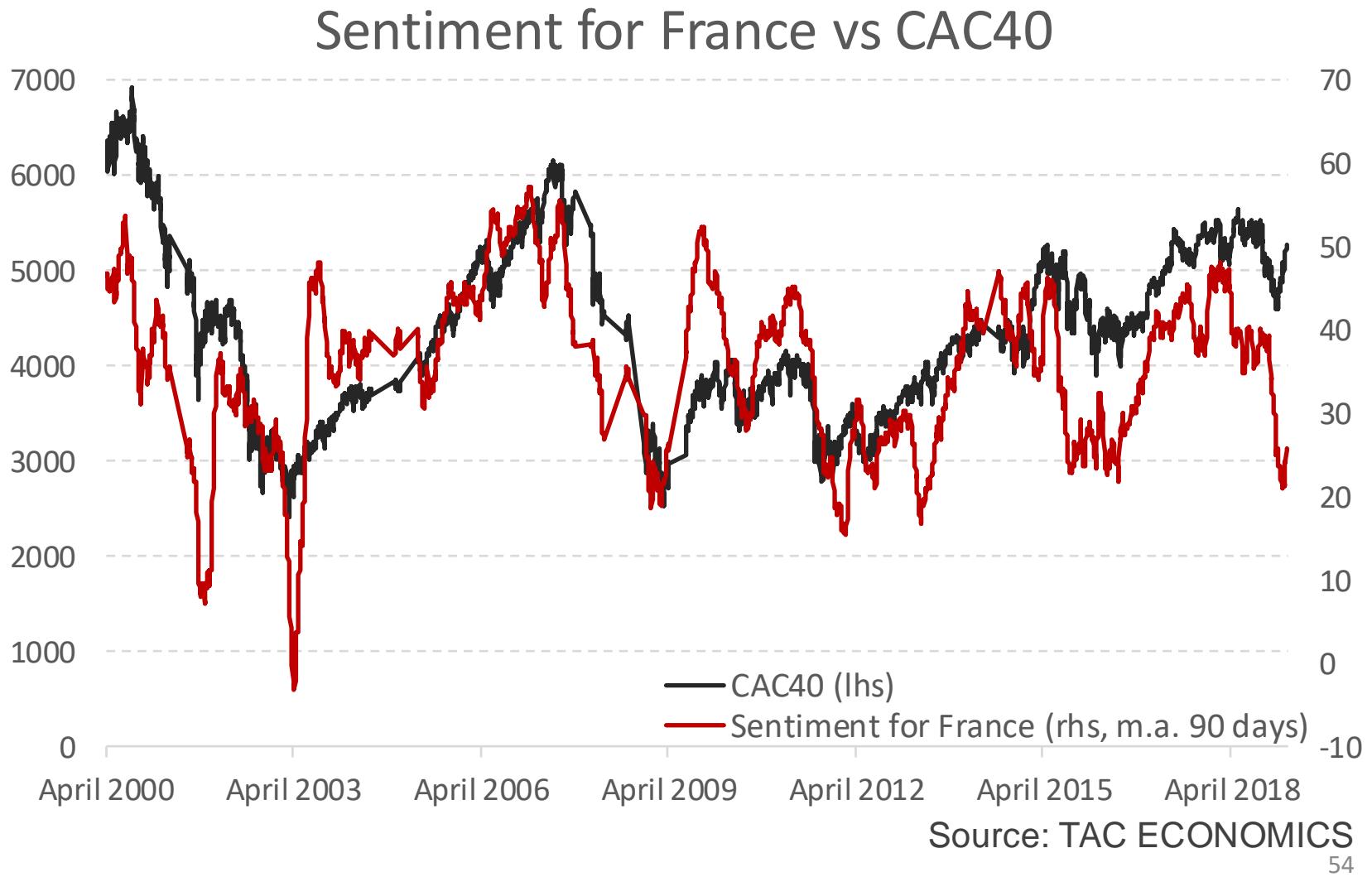


Risque réputationnel: Volkswagen & the Diesel Gate

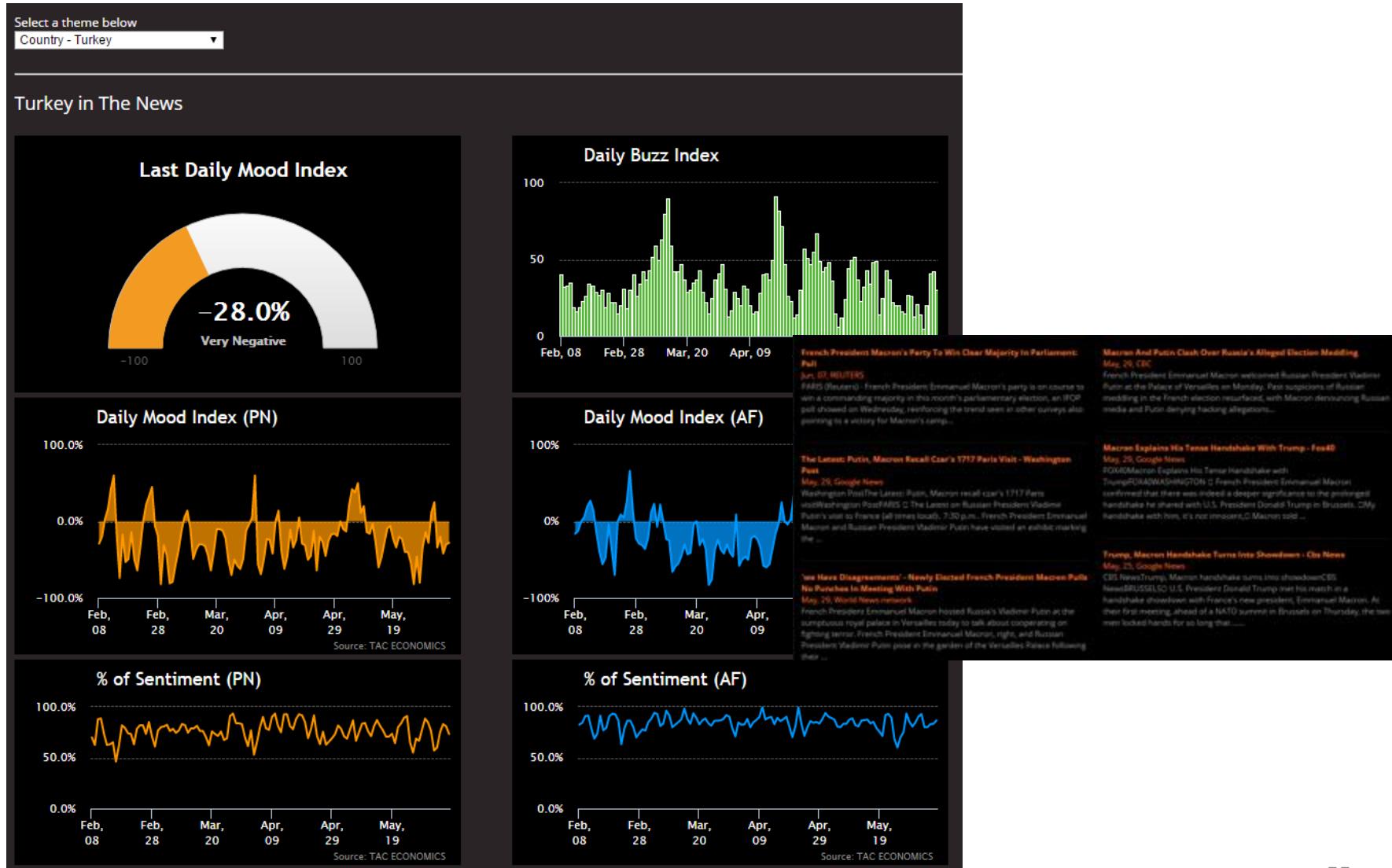


Source: TAC ECONOMICS

Indicateurs de sentiment et indices boursiers



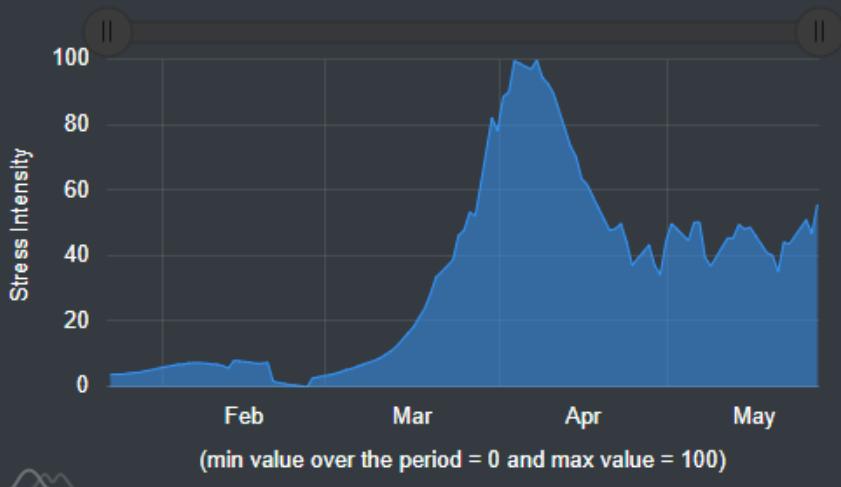
Sentiment Analysis, Topics & Finance



Sentiment Index « Covid-19 »

Epidemiologic Index

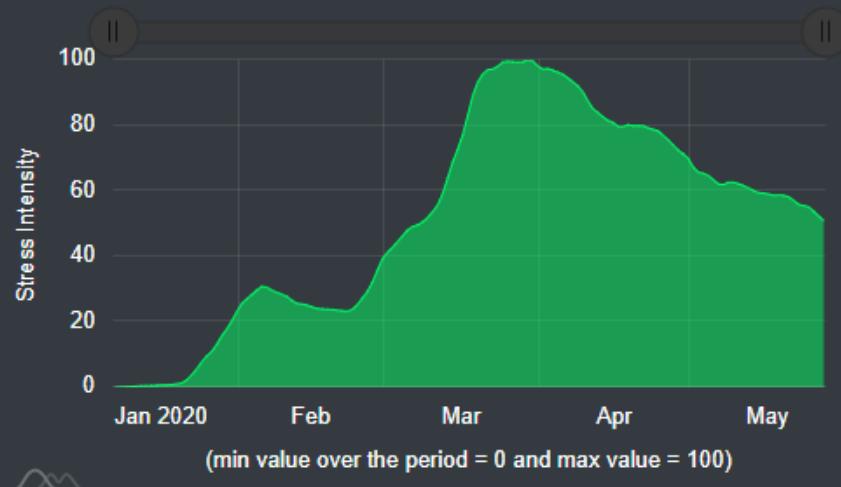
The Epidemiologic Index is calculated using indicators based on speed and acceleration of cases and deaths due to the Covid-19.



Last update: May 27, 2020

Sentiment Index

The Sentiment Index is calculated using text-mining techniques applied to the analysis of thousands press articles published on a daily basis.



Last update: May 27, 2020

<https://www.taceconomics.com/covid19/>

Introduction à R

Savoir programmer aujourd’hui

- On n’attend plus la même chose d’un programmeur qu’il y a 20 ans.
- Il doit avant tout savoir quoi chercher et quel langage est le plus adapté à la résolution de son problème.
- Connaitre la syntaxe « par cœur » est un moins important.
- En revanche, savoir assembler habilement des bouts de codes, éventuellement issus de plusieurs langages est essentiel.

Les langages de programmation

Langage de programmation dit de « haut niveau »... et interpréte

- Langage de « bas niveau ». Exemples : C, C++,...
- Langage de « haut niveau ». Exemples : R, Python, SAS, Java, PHP,....

Langage puissant pour applications mathématiques et statistiques car développé uniquement dans ce but

- Basé sur des notions matricielles/vectorielles, ce qui simplifie calculs et réduit recours aux procédures itératives (boucles)
- Langage peu typé, et pas besoin de déclarer les variables (avantage et inconvénient)

Présentation de R

- A la fois un langage de programmation, un logiciel de statistiques descriptives et un outil de data mining/machine learning avancé.
- R est un logiciel libre et gratuit, multiplateformes (Windows, MacOs, Linux), mais avec des versions payantes (version Microsoft de R)
- Large collection d'outils statistiques et graphiques sous forme de packages
- Toujours très utilisé dans le domaine académique, un peu moins dans le domaine privé ces dernières années.

Présentation de R

- **Avantages**
 - Gratuit
 - Logiciel Open-Source (libre) : chacun peut contribuer
 - Beaucoup de pack (librairies) déjà disponibles
 - Un logiciel qui suit l'évolution des méthodes statistiques
 - Graphiques plus complets que la plupart des logiciels de statistiques
- **Inconvénients**
 - L'interface graphique est pauvre (utilisation de RStudio)
 - Peu optimisé pour des calculs complexes.
 - Risque de temps de calcul important avec des bases de données très lourdes (> 1 To)

Présentation de R

- R est finalement composé de trois parties
 - La base R, disponible avec le logiciel, qui contient toutes les fonctions de base
 - Les packs « de base » (zoo, stats, ...)
- Les packages issus de monde « libre »:
 - ACP avec « FactoMineR »
 - datamining avec « caret »
 - économétrique avec « dynlm »
 - graphiques avec « ggplot2 »
 - big data avec « Rhadoop»,...

Commandes de base

- Exécution et résultats dans la console
- Programmation des commandes dans un éditeur de texte (R script)
- « > » l'utilisateur peut entrer une commande
- « + » la commande précédente est en cours d'exécution. On peut stopper à tout moment la commande
- « Ctrl + R » ou « Ctrl + Enter » pour exécuter une commande
- On navigue dans les anciennes commandes avec les flèches du clavier « haut » et « bas »
- R est « case sensitive », il fait la différence entre les majuscules et les minuscules

Commandes de base

- A chaque nouvelle session, commencer par définir le dossier de travail ou « working directory » (lecture et écriture) :
`getwd()` donne le chemin du dossier de travail de R
Pour le modifier `setwd` (chemin du dossier) ou utiliser l'onglet « Session » dans RStudio
- `help.start()` lance l'aide en ligne de R
- `help(nom de la fonction)` ouvre la documentation de la fonction
- `example(nom de la fonction)` exécute l'exemple de la fonction
- `install.packages(« nom du package »)` installe le package sur l'ordinateur
- `library (nom du package)` charge le package et permet de l'utiliser

Où trouver R ?

- Télécharger R
<https://cran.r-project.org/>



- Télécharger RStudio
<http://www.rstudio.com/>



Un peu de documentation

Aide en ligne

- CRAN
<http://cran.r-project.org>
- Google : mots clés + « r » ou copier/coller le message d'erreur
- Communauté R
<http://www.r-bloggers.com/>

Documentation

- Vincent GOULET – Introduction à la programmation en R
https://cran.r-project.org/doc/contrib/Goulet_introduction_programmation_R.pdf
- CRAN – An Introduction to R
<https://cran.r-project.org/manuals.html>
- Autres documents CRAN
<https://cran.r-project.org/other-docs.html>
- Quelques commandes R
http://www.math.unicaen.fr/~chesneau/RCarte_Commandes-R.pdf
- Aide RStudio
<https://www.rstudio.com/wp-content/uploads/2016/01/rstudio-IDE-cheatsheet.pdf>