

Big Data** , **GeoAnalytics** and **Systems Analysis

Eduardo de Rezende Francisco
eduardo.francisco@fgv.br

 **FGV EAESP**
ESCOLA DE
ADMINISTRAÇÃO
DE EMPRESAS
DE SÃO PAULO





Professor of the TDS (Technology and Data Science) Department at **FGV EAESP**. He teaches disciplines related to GeoAnalytics, Spatial Statistics and Data Science. Bachelor in Computer Science from IME-USP and Master and PhD in Business Administration from FGV EAESP. Deputy Coordinator of the Graduate Course in Business Administration at FGV EAESP. He is a visiting researcher at the Spatial Information Research Center at the University of Otago, New Zealand.

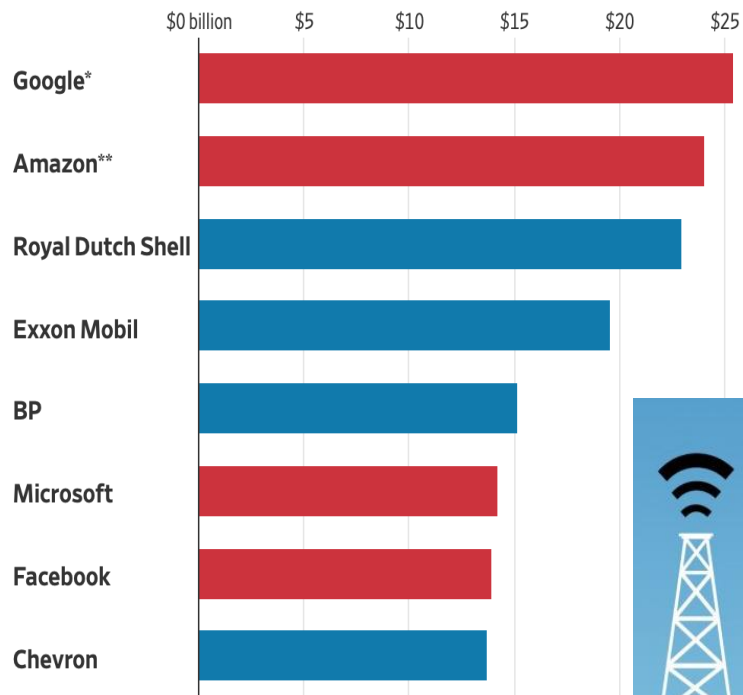
He is CKO (Chief Knowledge Officer) of startup **Meia Bandeirada** and founding partner of **GisBI**, a study and promotion group for Geotechnologies and Business Intelligence.



He has been working in the GIS, Geomarketing and Business Intelligence markets since 1994. He was technology manager of the GIS project and Market Planning project and responsible for the development of Marketing Strategies and Market Research Planning at AES Eletropaulo for 13 years. Participates as a speaker in various Geomarketing, BI, GIS and Utilities events. He is a columnist for GV Executivo and InfoGEO magazines, and a consultant in Spatial Statistics and Predictive Models for Credit and Real Estate. Member of the **Fundação SEADE** Board of Trustees.

Greasing the Wheels

Capital expenditures for 2018



*excludes Other Bets **includes capital leases

Source: company data, FactSet

You are now logged in

Regulating the internet giants

The world's most valuable resource is no longer oil, but data

The data economy demands a new approach to antitrust rules



David Parkins

Print edition | Leaders >

May 6th 2017



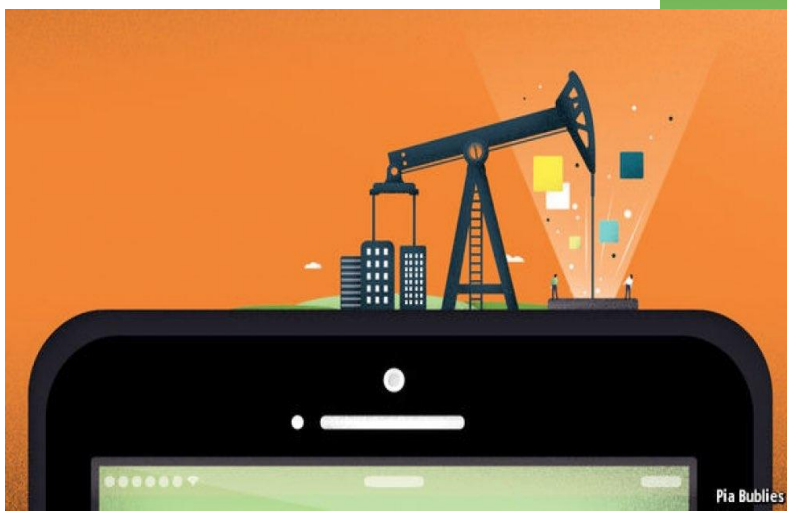
A NEW commodity spawns a lucrative, fast-growing industry, prompting antitrust regulators to step in to restrain those who control its flow. A century ago, the resource in question was oil. Now similar concerns are being raised by the giants that deal in data, the oil of the digital era. These

Data is the new oil.

We see in data the same transformative, wealth-creating power that 19th-century visionaries once sensed in the crude black ooze trapped underground.

If "crude" data can be extracted, refined, and piped to where it can impact decisions in real time, its value will soar. And if data can be properly shared across an entire ecosystem and made accessible in the places where analytics are most useful, then it will become a true game changer, altering the way we live, work, learn, and play.

Source: Cisco IBSG, 2012. #DataInMotion



DATA & AI LANDSCAPE 2019

INFRASTRUCTURE

HADOOP ON-PREMISE
 cloudera Hortonworks
 MAPR Pivotal
 IBM InfoSphere
 jethro

HADOOP IN THE CLOUD
 AWS Microsoft Azure
 Google Cloud
 SAP Cloud Platform
 IBM InfoSphere BigInsights
 arm
 blue CAZENA

STREAMING / IN-MEMORY
 Amazon Kinesis
 databricks
 SAP Cloud Platform
 ORACLE
 confluent
 strimzi hazelcast
 GridGain
 GIGASPACEs Wallarooes FASTDATA Ix

ANALYTICS & MACHINE INTELLIGENCE

DATA ANALYST PLATFORMS
 Microsoft pentaho alteryx
 Digital Reasoning
 GUAVUS AYASDI
 ATTIVO Datameer incorta.
 inter|ana MODE ENDOR
 sisu switchboard Starburst

DATA SCIENCE PLATFORMS
 IBM databricks dataiku
 DOWING rapidminer TIBCO
 SAS
 ANAconda
 KNIME MathWorks

APPLICATIONS - ENTERPRISE

SALES
 CHORUS
 INSIDESALES.COM peopleai
 conversica
 clar|avisio tact.ai
 fusesmachines Clearbit

MARKETING - B2B
 RADIUS App Annie
 EVERSTRING
 MINTIGO
 tubular
 JE N G A G I O
 KNOTCH mrp

MARKETING - B2C
 ZETA bloomreach SendGrid
 Inzage ACTIONIQ BLUECORE
 CONTENT SQUARE TEALUM mparticle
 Amperio amperity QUANTIFIND
 Simon [PERSADO]
 -remesh

CUSTOMER EXPERIENCE / SERVICE
 qualtrics MEDALLIA SurveyMonkey User Testing
 CLARABRIDGE zendesk Customer Freshdesk
 INTERCOM Drift LIVEPERSON Gainight pandao
 HEAP Amplify Warcon Assistants
 DigitalGenius ASAPP ads automatt ahmiti
 CallDesk notonix clintec frame.ai

ENTERPRISE PRODUCTIVITY
 slack
 ORACLE
 GURU lumiaio
 DIFFBOT clara
 talla Kasisto

NoSQL DATABASES
 Google Cloud AWS
 ORACLE Microsoft Azure
 mongoDB MarkLogic
 Couchbase DTRBSTAR
 redshift ELASTICSEARCH
 ArangoDB SCYLLA

NewSQL DATABASES
 SAP clustrix
 MICROSOFT PIVOTAL
 MEMSQL
 Cockroach LABS
 VOLTDB splice
 paradedigm

GRAPH DBs
 Amazon Neptune
 IBM
 ORACLE
 Neo4j
 GraphDB
 Dgraph

MPP DBs
 TERADATA
 IBM Data Warehouse Systems
 ORACLE
 Cognitio
 EXASOL
 Yellowbrick

CLOUD EDW
 AWS
 Google Cloud
 Microsoft Azure
 Pivotal
 Snowflake
 nuclio
 Infoworks

SERVERLESS
 AWS
 Google Cloud
 Microsoft Azure
 Pivotal
 nuclio
 Infoworks

BI PLATFORMS
 looker
 DOWING
 ATSCALE
 Qlik
 MicroStrategy Keen IO

VISUALIZATION
 +tableau
 Power BI
 SAP
 Google Cloud
 CELONIS
 zepi
 CHARTIO

MACHINE LEARNING
 AWS
 Google Cloud
 H2O
 DataRobot gamalon
 VISENZE ELEMENT
 deepsensei

HUMAN CAPITAL
 HireVue
 hiQ
 Allyo
 Wades&Wendy
 entelo
 ONCOMAN
 RASS
 casetext

LEGAL
 RAVEL
 Everlaw
 JUDICATA
 BREVIATA
 IPRONCI AD
 PREEMPTION
 RASS

REGTECH & COMPLIANCE
 TextIQ
 Comply Advantage
 PARTNERSHIPS
 DATA REPUBLIC

FINANCE
 Anaplan
 ZUOFO
 SAHANA
 TRADESHIFF
 SCALES FACTOR
 burkkeeper
 pilot

BACK OFFICE - AUTOMATION & RPA
 UiPath
 blue Prism
 VIDADO
 Workfusion
 REMOS
 ANTWORKS
 ALKYEM

SECURITY
 CYLANCE
 zscaler
 StackPath
 illumio
 CODE42
 CipherCloud
 DARKTRACE ANOMALI
 Vectra
 pindrop
 exabeam
 SINCERYD
 SentinelOne
 SecurityScorecard
 Socrate
 VadeSecure
 bitglass
 BlueTalon
 Secured
 feedzai
 Cyber
 BIT SIGHT
 AKATI
 BLUEHEXAGON
 Semmle
 SASSON
 XENYU
 SHENJI
 APTOR

DATA TRANSFORMATION
 talend pentaho
 alteryx TRIFACTA
 tamr
 StreamSets UNIFI

DATA INTEGRATION
 SAP Data Services Informatica
 Mulesoft TEALUM
 anaplogic enigma
 Segment ATTUNITY
 ZALONI
 Infoworks
 FIVEPAK
 SNOWFLOW MATLION

DATA GOVERNANCE
 Informatica
 ScalPoint
 IBM
 Alation
 PHMUTA
 OKERA
 MANTA dataworld

MGMT / MONITORING
 AWS New Relic octrifio
 rubrik
 APPDYNAMICS
 dynatrace
 Signalix
 druva
 splunk
 Microsoft paperduty
 Unravel Numentry
 ZEPHYRUS
 Opstrace
 MAGNITUDE

COMPUTER VISION
 Microsoft Azure
 Amazon Rekognition
 clarifai
 EVERAI
 deepomatic
 twentybn
 USBIQUITYS
 ADEE
 YITU
 TRAC
 synthesis

HORIZONTAL AI
 IBM Watson Cortana
 sentient
 Afffectiva
 Numenta
 nalogics
 BLUE VISION
 FORTRESS

SPEECH & NLP
 Google Cloud
 Amazon Alexa
 Amazon Transcribe
 narrative
 Movel
 SoundHound Inc.
 PRIMER
 cogito snips
 SHRETTLE
 Ustund
 Piplai

ADVERTISING
 Appexus
 critico
 ORACLE
 MOAT
 theTradeDesk
 distillery
 TAPAD
 dataxu
 gungumti
 Appier

EDUCATION
 Lulliship
 KNEWTON
 Clever
 Clever
 kidapdf
 PANORAMA
 knowrope

REAL ESTATE
 REDFIN
 Opendoor
 VTS
 CREDIFI
 GEMPHY
 reonomy
 COMPSTAK
 STREET SCHEIDATA
 STAGE MAKER

GOVT
 OPENGOV
 mark43
 LiveStories
 Passport
 SmartProcure

INTELLIGENCE
 Palantir
 Dataminr
 Quid
 PRIMER
 FORGE

FINANCE - INVESTING
 KENSHC
 Quantopian
 MIBRASA
 ISENTIUM
 ALGORIZ
 TrueAccord
 PACAYA

FINANCE - LENDING
 ondeck
 Affirm
 JIANPUJAI
 KREDITECH
 AVANT
 aurea
 BLEARBANC
 upgrade
 100Credit
 WeLab
 wecasian
 aire
 ognifi

INSURANCE
 Thronomix
 Anomix
 CYFENCE
 Hippo
 Shift Technology
 ROOT
 resty.ai
 TRACABLE
 CAPE

STORAGE
 AWS
 Microsoft Azure
 PuresStorage
 ALLUXIO
 wasabi
 nimbustorage
 Omundo
 parasec
 COHESITY

CLUSTER SVCS
 Amazon EMR
 Databricks
 EMRFS
 GigaScale
 GigaScale
 GigaScale
 GigaScale
 GigaScale

DATA GENERATION & LABELLING
 Amazon Mechanical Turk
 Upwork
 appen
 unity
 scale
 HIVE
 Labelbox
 Magnifi AI
 ALSEVERE
 LIONBRIDGE

AI OPS
 ALGORITHMIA
 Vertaai
 datmo
 Riiidoo
 Determined AI
 fiddler

GPU DBs & CLOUD
 Kinetica
 ORACLE
 BYTET
 PG-Stream
 HAYDRIE

HARDWARE
 Google TPU
 ARM
 intel
 NVIDIA
 GRAPHCORE
 MYTHIC
 Graphcore
 VIVACE
 CRYNAMI
 VIVACE
 DEFNIX

SEARCH
 ORACLE
 ELASTICSEARCH
 algolia
 COVEO
 Lucidworks
 ATTIVO
 swifttype
 EXPANDED
 alpha-search
 MAANA
 omni:us
 SINEOUIA
 logz.io

LOG ANALYTICS
 splunk
 sumologic
 solarwinds
 loggly
 TIMBER
 logz.io

SOCIAL ANALYTICS
 Hootsuite
 sprinklr
 NETBASE
 synthesio
 trackr
 simpereach
 bitly
 SimilarWeb

WEB / MOBILE / COMMERCE ANALYTICS
 Google Analytics
 mixpanel
 AMPITUDE
 Airtable
 RESCI
 SIGOPT
 granify
 CUSTORA

HEALTHCARE
 flatiron
 CLOVER
 KYRUS
 HealthTap
 METABOTA
 Gingerio
 Glow
 babyzen
 3D Med
 zebra
 POCANI
 ovia
 TEMPUS
 patientsilome
 AICure
 insitro
 LINKDOO
 meffrulo
 citizen
 notable
 @enitic
 imago
 Biochthon
 BAYLABS
 Qventus
 AXELERS
 I M A G E N
 innovaccer
 PAIGE
 DATAMANN
 LUMINA
 LEAP THERAPEUTICS

LIFE SCIENCES
 BenevolentAI
 WuXiNextGen
 Clear Labs
 InnovaMetric
 Phosphorix
 CITRINE
 twoSTAR
 Atomize
 DOWIN

TRANSPORTATION
 CLEARPATH
 CRUISE
 Nuro
 Auro
 Nauto
 AMOTIVE
 G7
 PILOTAI
 NIO
 OPTIMUS
 moovit
 nexar
 xodiak
 comma.ai
 netradyno
 Sentinix
 Civil Maps
 think
 INRIX

AGRICULTURE
 FARMERS
 Instacart
 STITCH FIX
 Dis & Co
 BLUEVERVE
 FarmersEdge
 AgroStar
 FarmLogs
 TARANIS
 GAMAYA
 ierivation
 prospero

COMMERCE
 Instacart
 STITCH FIX
 Dis & Co
 BLUEVERVE
 FarmersEdge
 TACHYUS

INDUSTRIAL
 AVEVA
 SIEMENS
 PREDIX
 UPTAKE
 SCORTEX
 KODIX
 TACHYUS

OTHER
 eHarmony
 stem
 Amper
 ByteDance
 happes
 celec
 SOJERN
 BBOXEE
 VERDIGIS
 duetto
 GAMAYA
 Electric
 ZINBER
 Spike

CROSS-INFRASTRUCTURE/ANALYTICS

aws Google Cloud Microsoft IBM SAP Oracle NetApp syncsort MAPR cloudera

OPEN SOURCE

FRAMEWORKS
 Spark
 Flink
 YARN
 TEZ
 MESOS
 chcker
 CDAP
 Red Hat
 HELIX

QUERY / DATA FLOW
 Spark SQL
 presto
 SLAMDATA
 GraahQL
 Flink

DATA ACCESS & DATABASES
 cassandra
 mongoDB
 redis
 Cockroach LABS
 druid
 SciDB
 sriak
 HBASE
 HIVE
 HADOOP

ORCHESTRATION & MGMT
 talend
 Apache Ambari
 Apache Airflow
 MESOS
 etcd
 Kong

STREAMING & MESSAGING
 Spark
 nifi
 Flink
 beam
 kafka
 STORM
 Apache RocketMQ

STAT TOOLS & LANGUAGES
 jupyter
 Scala
 Studio
 SciPy
 julia

AI OPS & INFRA
 miflow
 Kubeflow
 mlops
 DC
 SELDON
 Polysyn

AI / MACHINE LEARNING / DEEP LEARNING
 TensorFlow
 Keras
 PyTorch
 OpenAI
 DM
 TK
 theano
 mxnet
 VELES
 Chainer
 Microsoft Cognitive Toolkit
 DIMSUM
 FeatureFu

SEARCH
 elasticsearch
 Solr

LOGGING & MONITORING
 elasticsearch
 kibana
 SENTRY
 logstash
 Prometheus
 fluentd
 fluentbit
 Grafana
 VICTOR

VISUALIZATION
 matplotlib
 TensorBoard
 seaborn
 D3.js

COLLABORATION
 BeakerX
 Jupyter
 Anaconda

SECURITY
 Apache Ranger
 KNOX
 Sentry
 SCOUT24
 CCUTTUDDO

DATA SOURCES & APIs

HEALTH
 Apple
 VALIDIC
 practice fusion
 fitbit
 GARMIN
 HUMAN API
 kinsa
 LAMIC

IOT
 GE Digital
 UPTAKE
 thingworx
 helium
 samsara
 estimate

FINANCIAL & ECONOMIC DATA
 Bloomberg
 THOMSON REUTERS
 DOW JONES
 S&P CAPITAL IQ
 CBINSIGHTS
 FLAID
 SECOND MEASURE
 INVESTNET
 YOCOLITE
 THEWORKS DATA
 Gestimize
 PREMISE
 Quandl
 Engo Alpha
 StockTweets
 xignite
 Thinknum
 earnest
 predata

AIR / SPACE / SEA
 Orbital Insight
 planet
 SKYCATCH
 AIRBOTICS
 aspire
 PROCESSION
 KESPRY
 UNCESTORY
 tellusdata
 WINDWARD
 DroneDeploy
 MarineTraffic
 LIFT ORBITAL
 PERSYS

PEOPLE / ENTITIES
 acxiom
 Experian
 EPSILON
 InsideView
 Crismson Hexagon
 BASIS
 Quantcast
 SAFEGRAPH

LOCATION INTELLIGENCE
 FOURSQUARE
 mapbox
 sense360
 PONYOVES
 HEXAGON
 PlaceIQ
 esri
 factual
 CARTA
 Mapillary
 Streetline
 cuebiq
 Radar
 OpenStreetMap

OTHER
 DATA.GOV
 IMAGENET
 LEONARDO
 KAGGLE
 ANNADEGAN
 CRUX
 Mapillary
 Streetline
 Ugraffiti

DATA RESOURCES

DATA SERVICES
 OPERA
 DATA SCIENCE
 fractal
 kaggle
 EXL
 DataKind
 INNOPLCUS

INCUBATORS & SCHOOLS
 PLURAL SIGHT
 GA
 galvanize
 DataCamp
 DataElite
 INSIGHT
 The Data Incubator
 METIS

RESEARCH
 facebook research
 OpenAI
 MIRI
 VECTOR INSTITUTE
 META
 CSAIL
 AI2
 ALLIAN INSTI TUTE
 ARTIFICIAL INTELLIGENCE

“Big Data” are data and processes whose scale, distribution, diversity and / or creation speed requires the use of new storage and analysis technologies to allow the capture of the value inserted in them

(FRANCISCO, 2016, adapted from EMC, 2013)

1. Data Volume

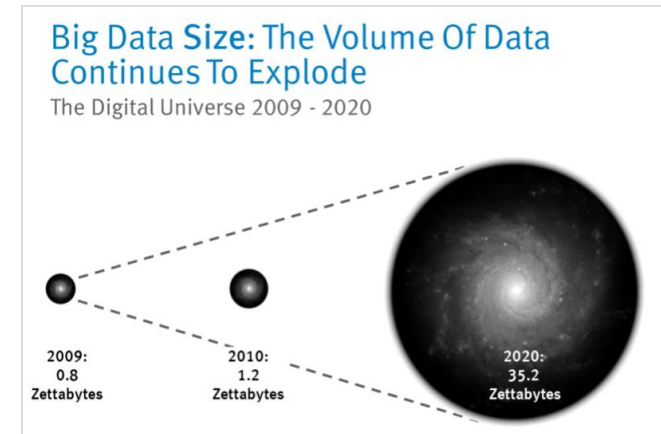
- ▶ Billions of lines x billions of columns
- ▶ Increase of 44 times from 2009 to 2020 (0,9ZB to 35ZB)

2. Processing Complexity

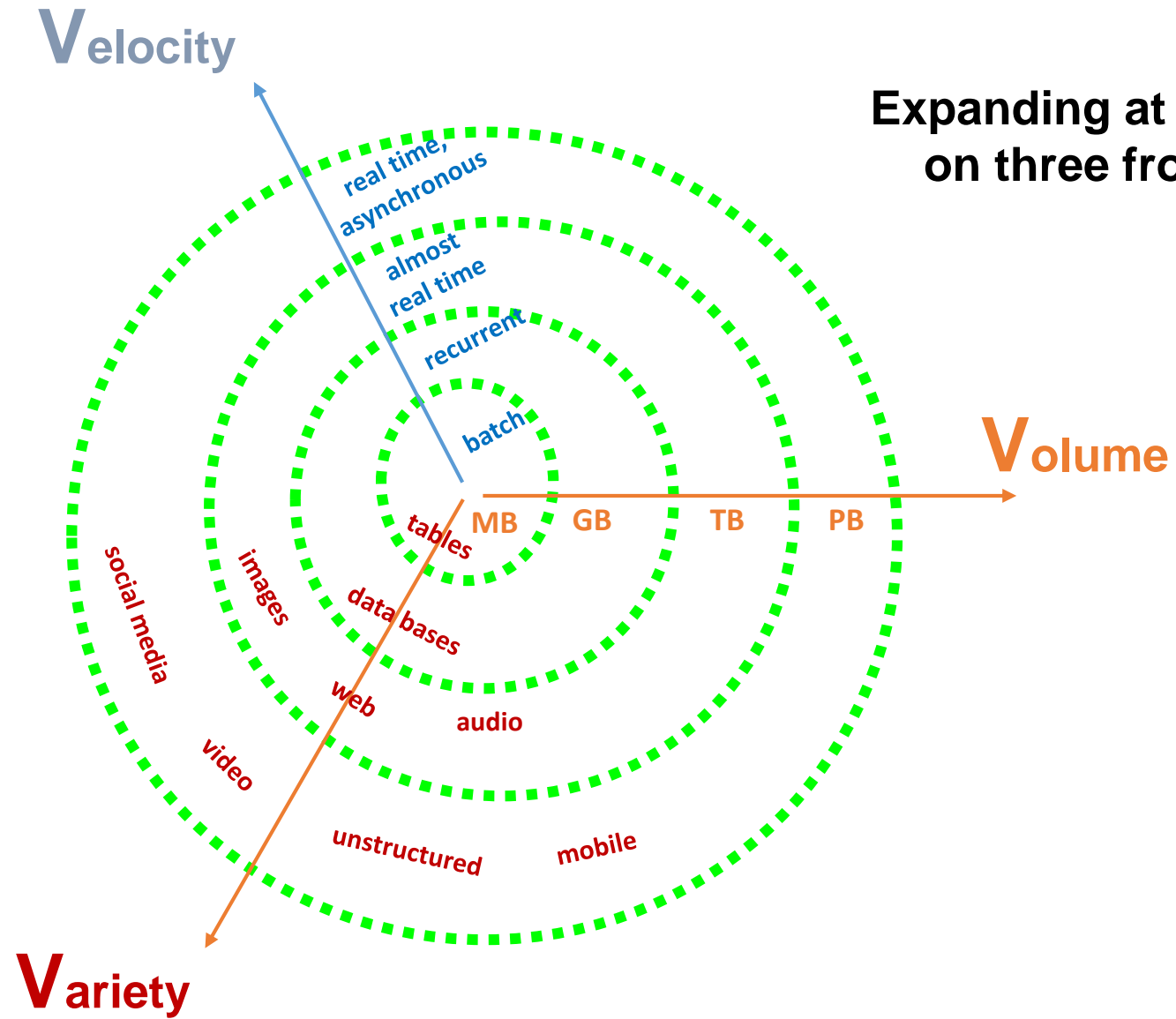
- ▶ Data structures are constantly changing
- ▶ Need to analyze such data in real time

3. Data Structure

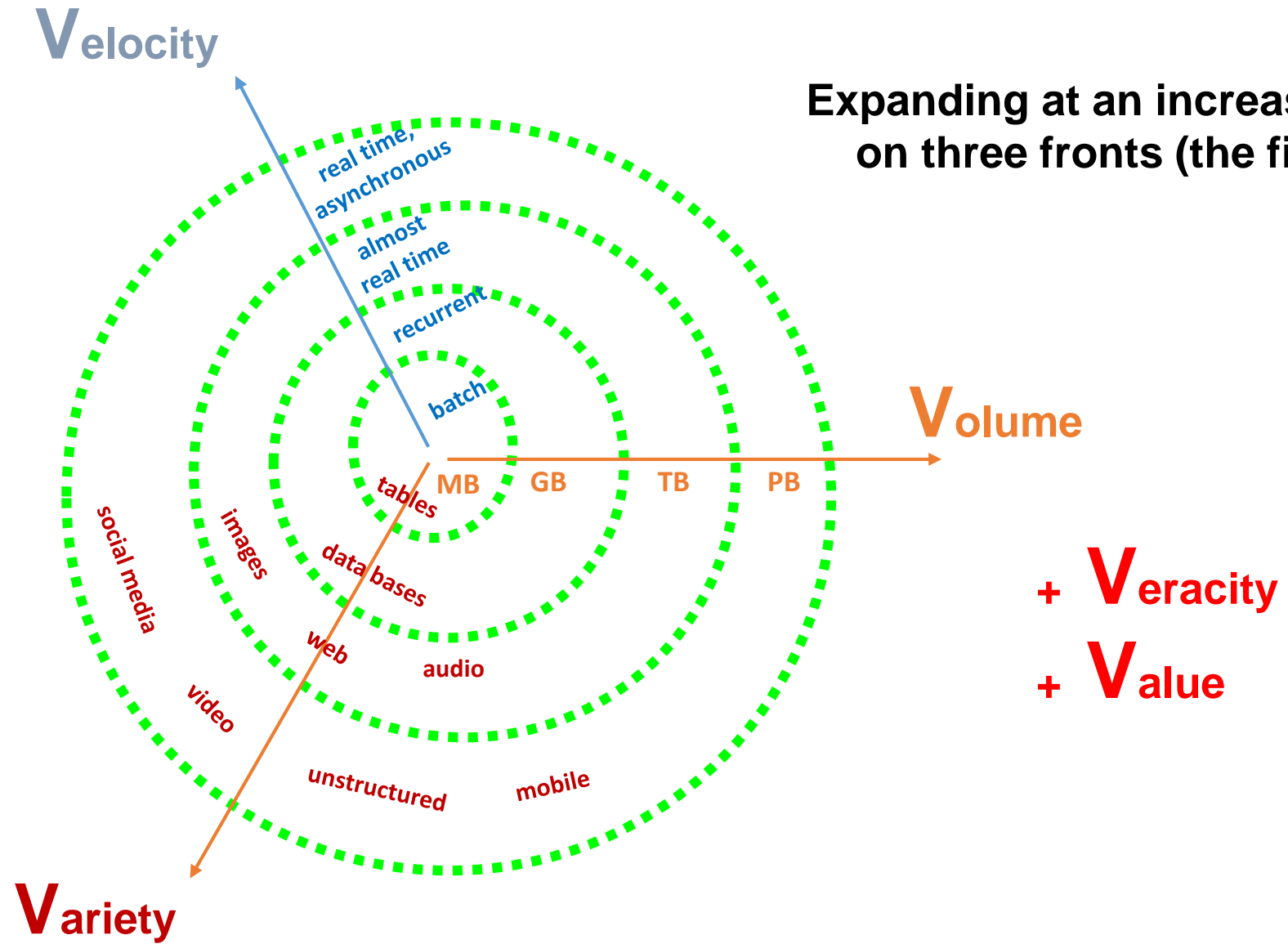
- ▶ Large variety (80-90% unstructured) to be analyzed
- ▶ These characteristics make it necessary to use parallel and parallel mass computing (MPP) systems



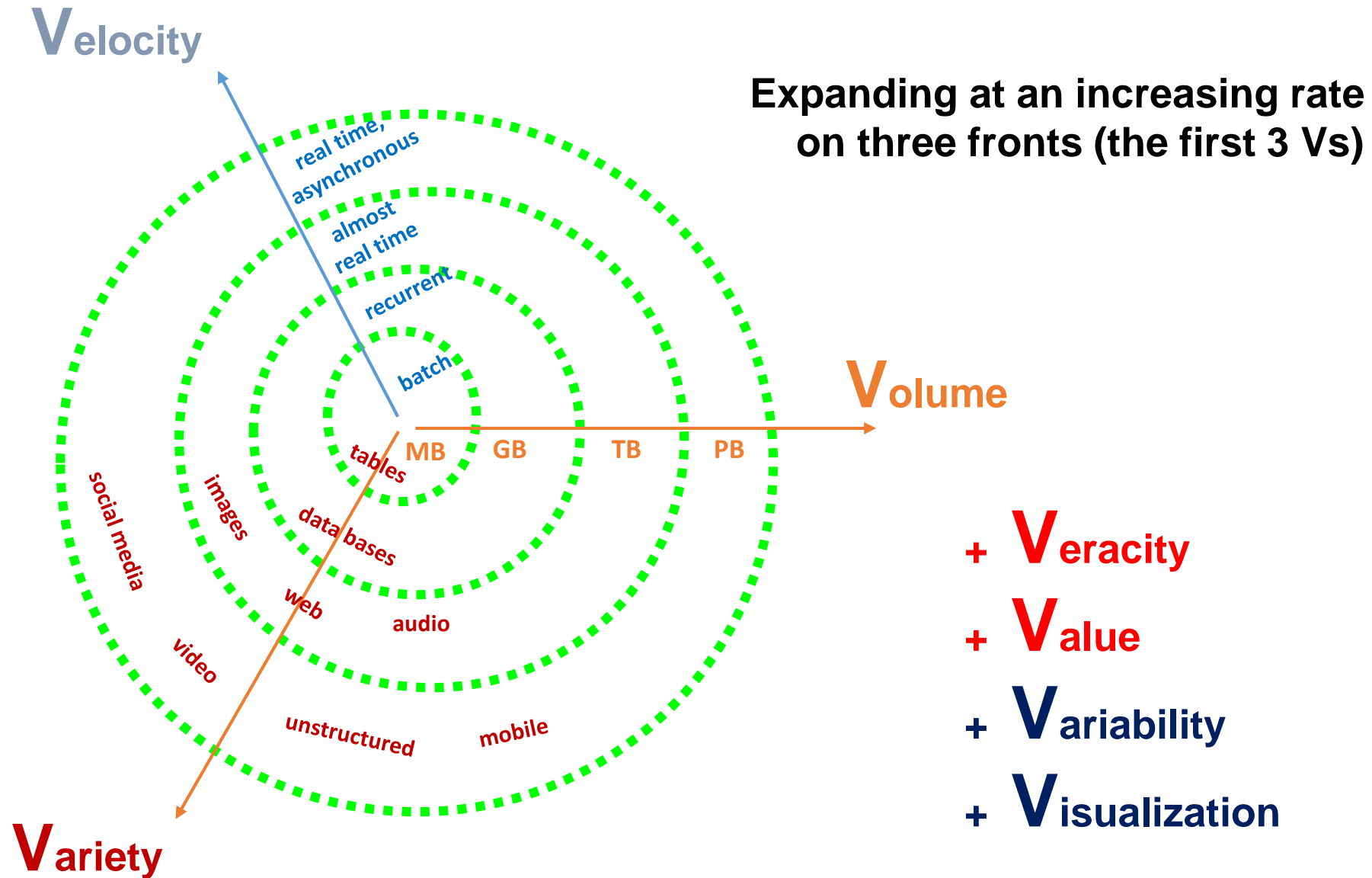
Source: EMC (2013)



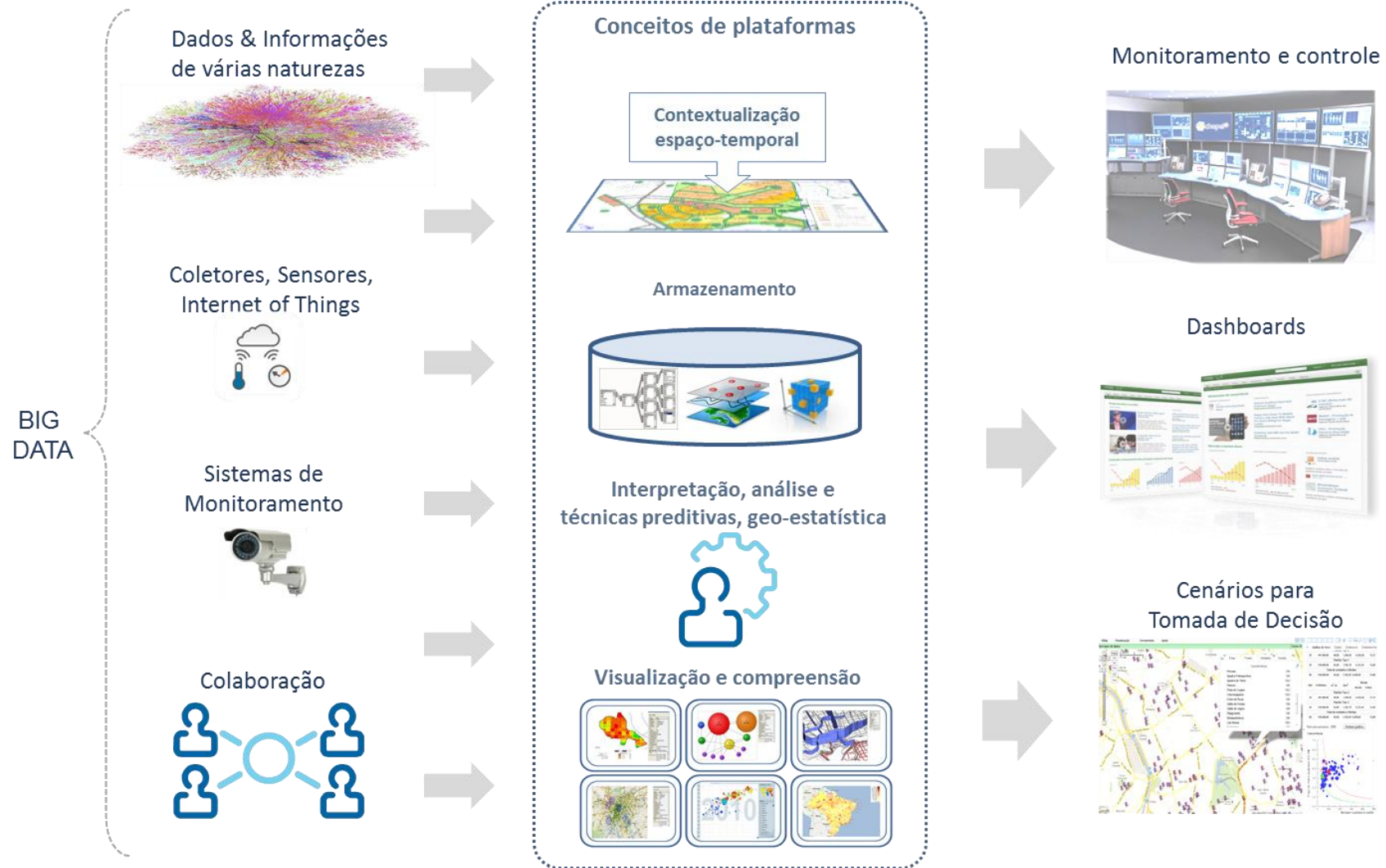
Expanding at an increasing rate on three fronts (the first 3 Vs)



Expanding at an increasing rate on three fronts (the first 3 Vs)



Big Picture



Challenges for Systems Analysis in the Big Data Era

- Big Data, Data Science, Analytics ...
- Challenge in the integration of analytical techniques
 - AI, Neuroscience applied to marketing and business, spatial statistics, behavioral models
 - Models to handle stakeholders' relationships
- Computational Challenge
- Challenge in the adoption of AI and Data Science by companies and public organizations
 - Adoption of forceps, apart from core management
 - Analytical Sandbox vs. IT Policy
- **Cultural challenge - new skills of analytical teams and managers**

Use of alternative information (non-structured "Big" data) in the generation of indicators

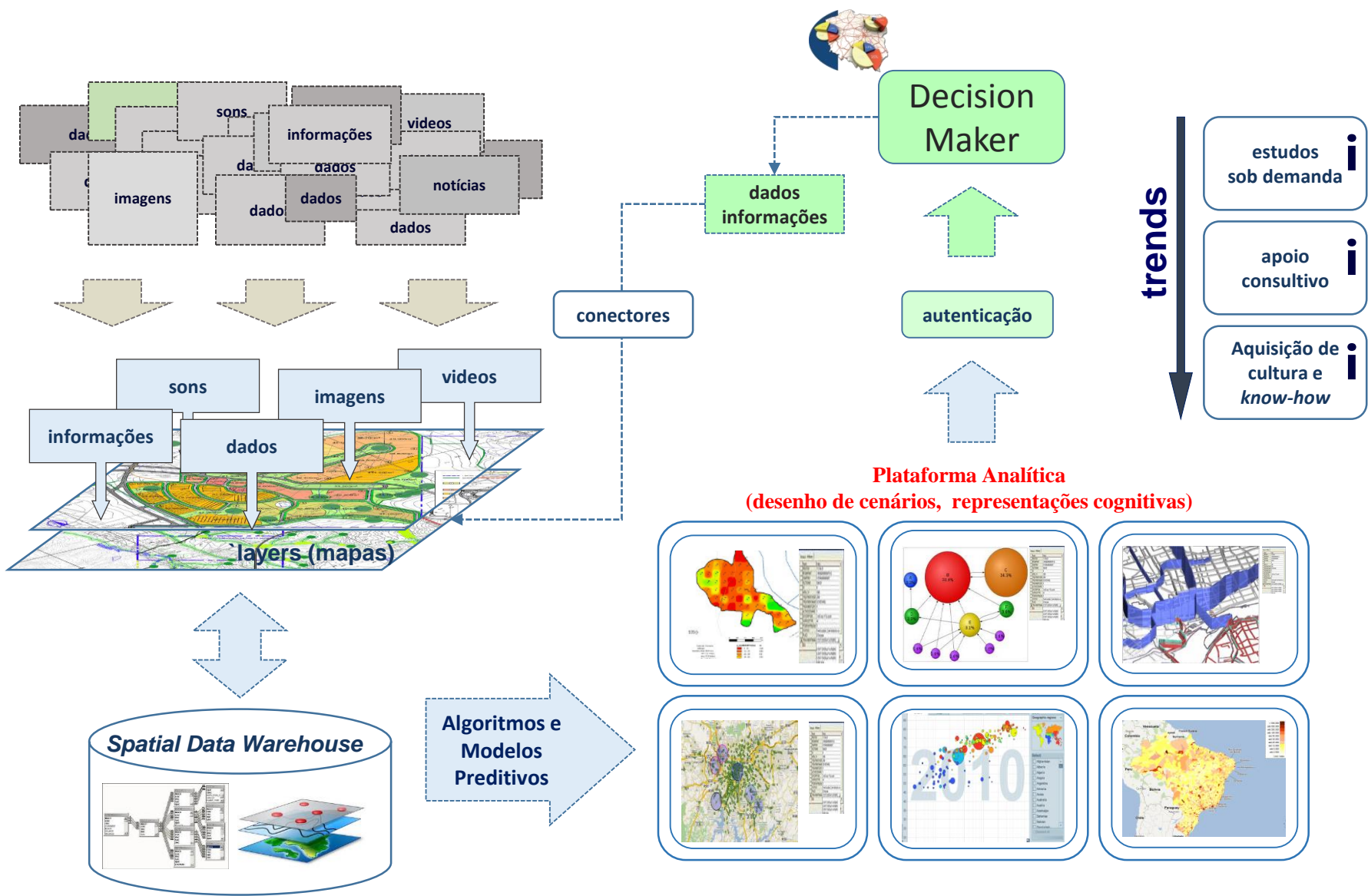
- Open Data - API or Web Scrapping
- Sectoral Reports, Management Reports, Integrated Reports - Web Scrapping
- Interpretation of Images
- Data Enrichment by Geo-Analysis and Spatial Statistics

Serious Implications for Scientific Research in Applied Social Sciences

Classic (Inferential) Statistics	Machine Learning
Strong hypotheses	Flexible approach
Theoretical justification	Empirical Efficacy
Exact solution	Approximate solution
Explanation/Interpretation	Prediction

Source: OLLION, 2018

Framework: Adaptive Business Intelligence and Big Data



Source: LETOUZÉ, 2017;
 MICHALEWICZ et al., 2006 ;
 GisBI, 2015



Inferring socio-demographic indicators



Scientific Prize and Ethics Mention: Construction of socio-demographic indicators with digital breadcrumbs

F. Bruckschen ⁽¹⁾, T. Schmid ⁽²⁾, T. Zbiranski ⁽¹⁾

We show that socio-demographic indicators such as population, age, literacy, poverty, religion, ethnicity, electricity supply and others can be estimated in unprecedented detail and virtually ad-hoc using antenna-to-antenna traffic data only. We offer a uniform approach that can be easily extended to other variables. Results are tested for spatio-temporal robustness and visualized as heat maps.

(1) Humboldt Universität Berlin, Germany - (2) Freie Universität Berlin, Germany

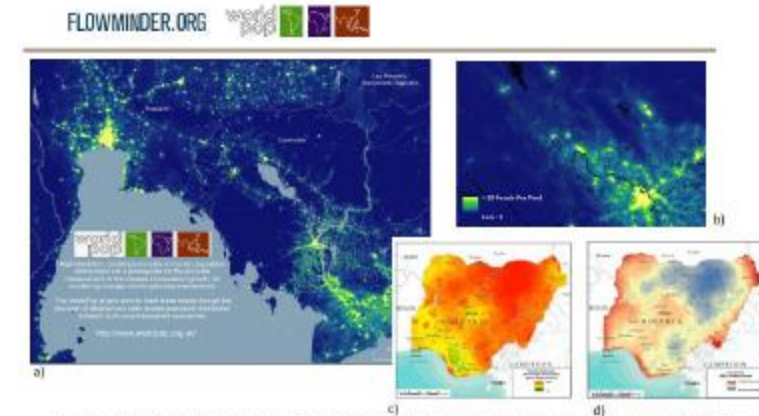
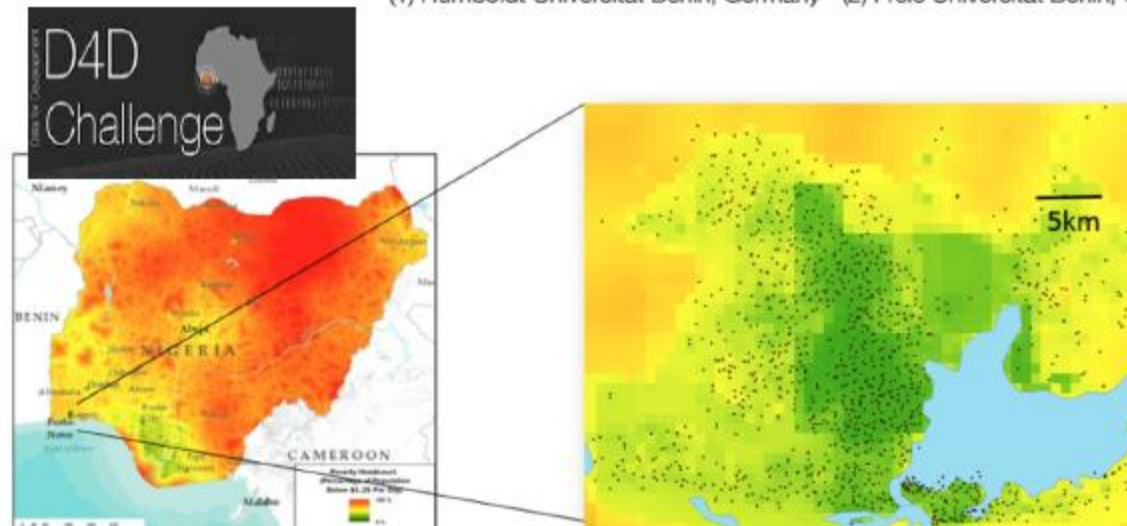


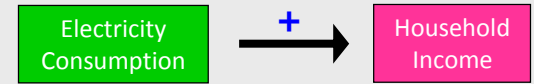
Figure 1: a) WorldPop population density distributions for the the Anzov region; b) close-up picture of population distributions (100x100m) for the Anzov region; c) Poverty headcount for Nigeria (<math>< 1.25\text{ USD/day}</math>) per 1 km²; d) Uncertainty in poverty headcount estimates per 1 km² area.

Source: LETOUZÉ, 2017

OBJ: Analyze the relationship between household income and electricity consumption

Create an income indicator based on electricity

Prediction Model:



Traditional OLS Regression:

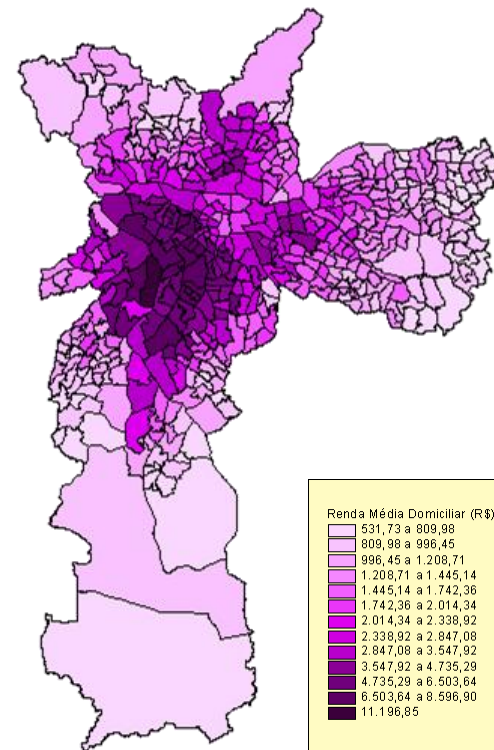
$$\hat{y} = \beta_0 + \beta_1 x \quad R^2 = 86,6\%$$

SAR (Spatial Auto-Regressive):

$$\hat{y} = \beta_0 + \beta_1 x + \rho W y \quad R^2 = 94,4\%$$

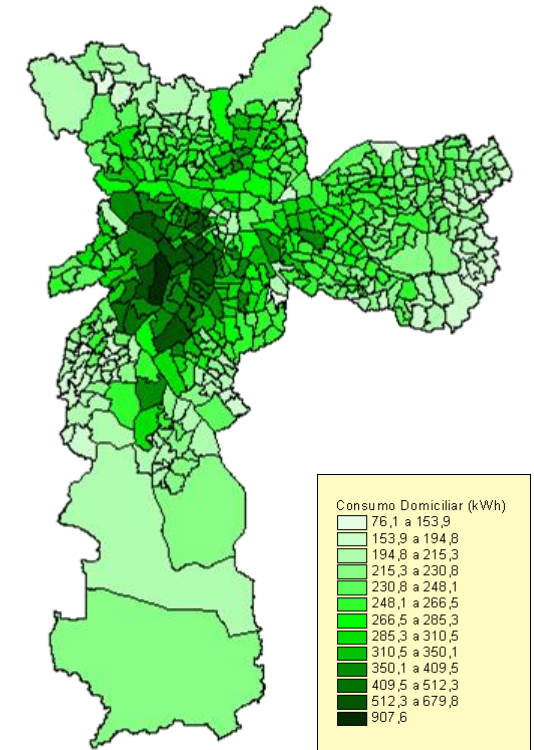
GWR (Geographically Weighted Regression):

$$\hat{y}_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i) x_i \quad R^2 = 96,8\%$$



Household Income

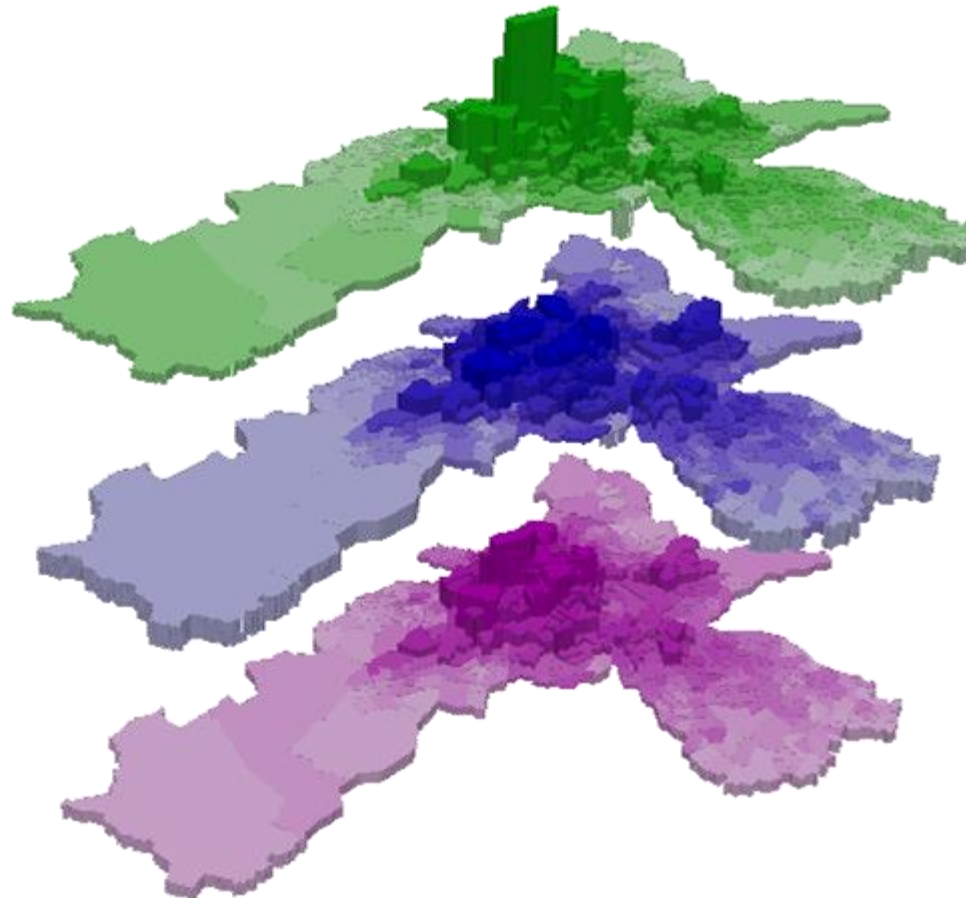
(IBGE)



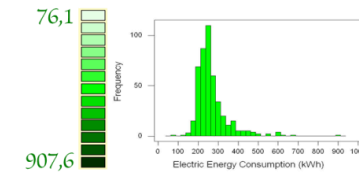
Electricity Consumption

(AES Eletropaulo)

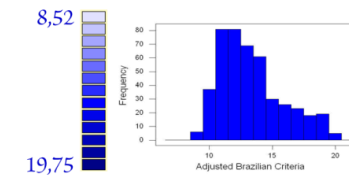
Microcredit Score and Socio-Economic Indicators based on Electric Energy



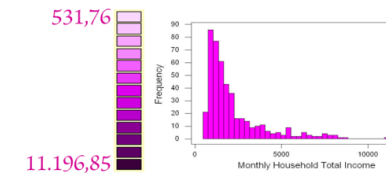
Consumo Residencial de Energia Elétrica (kWh)



Microcredit Score



Renda Domiciliar (R\$)

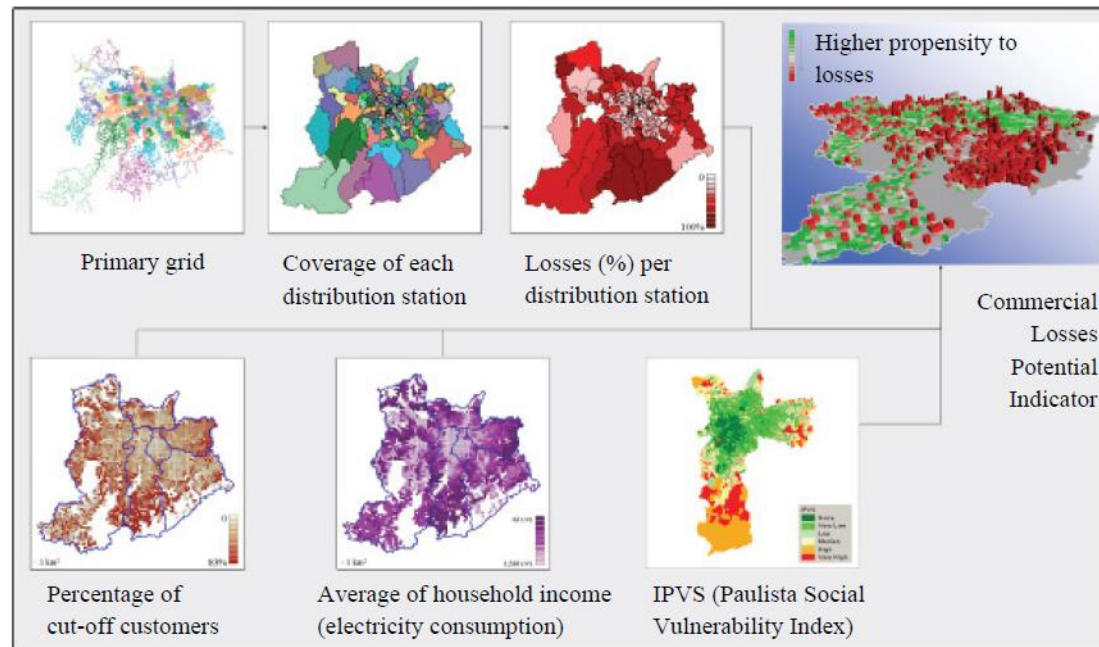


Source: FRANCISCO, 2011

- Development of an Indicator of Propensity to Energy Commercial Losses using Geospatial Statistical Techniques and Socio-Economic Data: the Case of AES Eletropaulo**

FRANCISCO, E., FAGUNDES, E., PONCHIO, M., ZAMBALDI, F. - EnANPAD 2009

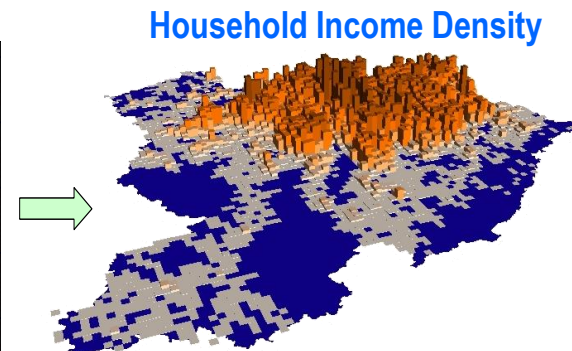
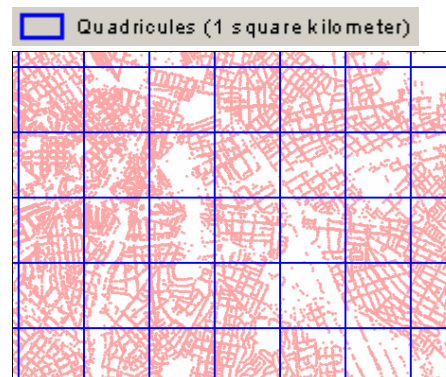
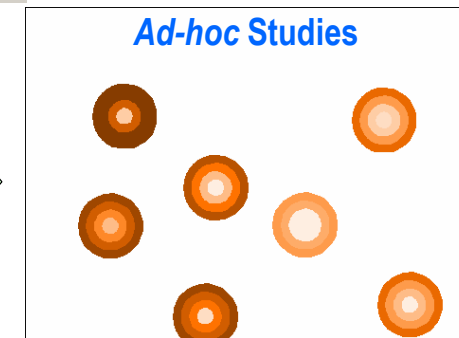
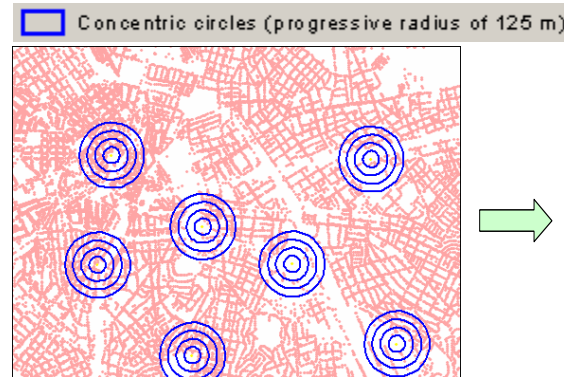
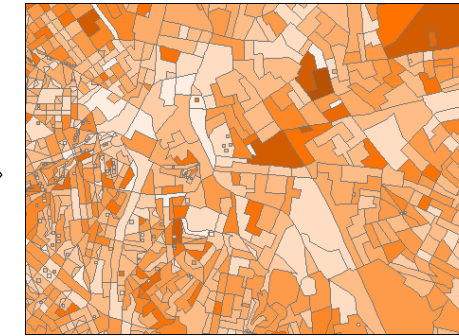
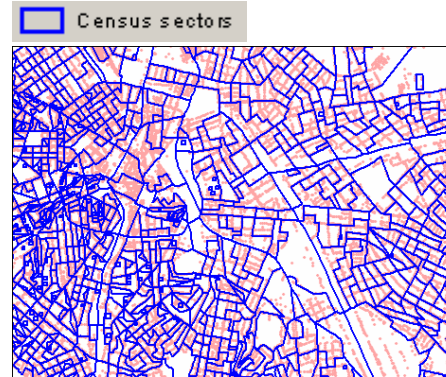
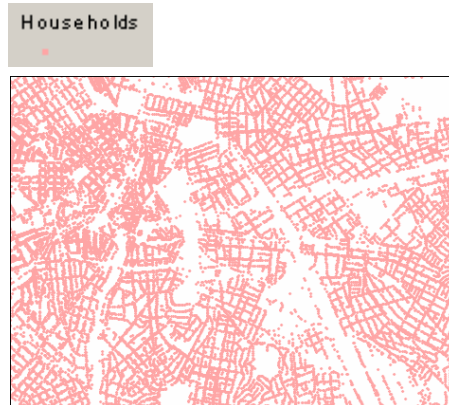
- Use of geospatial techniques for the design and operation of a fraud-prone indicator, and analysis (through GWR) of the association between this indicator and customer satisfaction



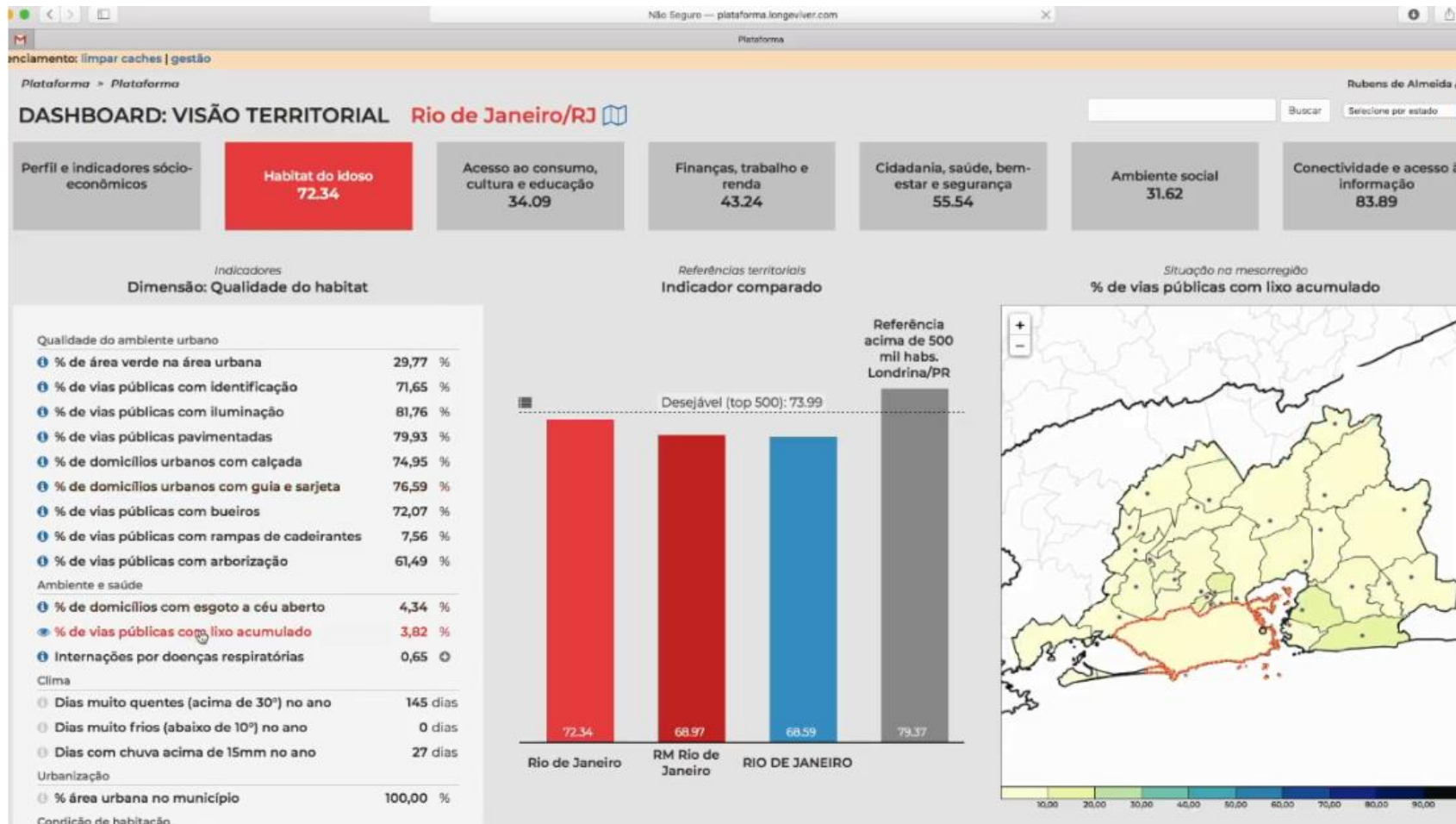
Source: FRANCISCO *et al.*, 2009

Socio-Economic Indicators based on Electricity Consumption

- Should be published widely by the electricity distributors
- Useful for strategy formulation and decision support
- *Support for Customer Relationship Characterization and Management*
- *Support for Public Policies and Systems Analyses*



Health and Longevity Research and Study Platform



<https://youtu.be/1ET4glwLAe0>



Thank you!
¡Gracias!
Obrigado!

Eduardo de Rezende Francisco
eduardo.francisco@fgv.br

 **FGV EAESP**
ESCOLA DE
ADMINISTRAÇÃO
DE EMPRESAS
DE SÃO PAULO

