BIG DATA?

THE GLOBAL IMBALANCE!

Stéphane Grumbach
INRIA

The digital universe 2.7 Zettabytes

Data deluge in all sectors of activity

U.S. Library of Congress: 235 Terabytes of data

Walmart: 2.5 petabytes of data, 1 million customer transactions / hour

Facebook: 30 Petabytes of user data

Google: processing 20 petabytes a day (2008)

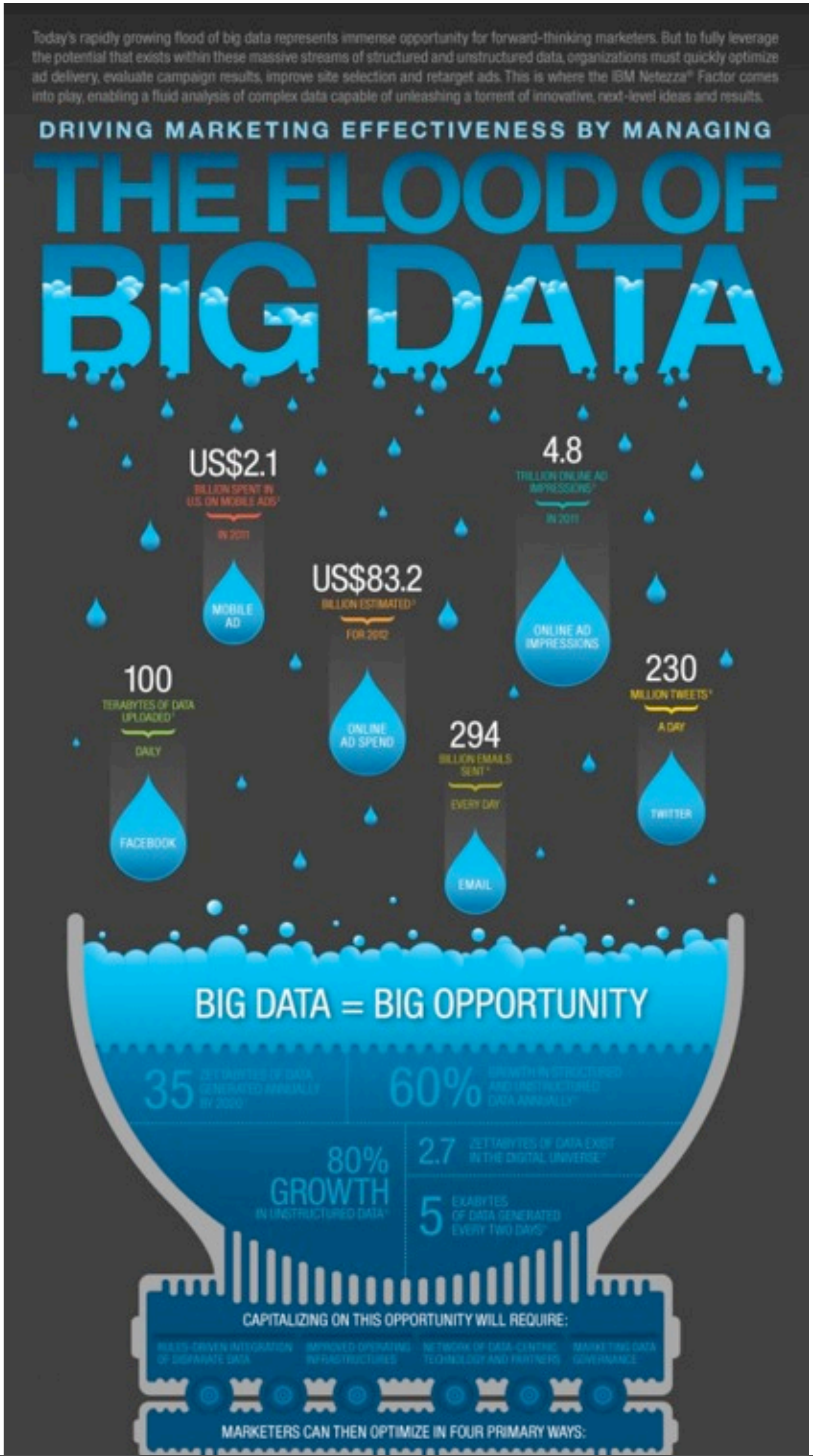
World: 5 billion people calling, tweeting, browsing on mobile phones

Exponential increase

doubles every two years **35 zettabytes in 2020**

followed by the capacity to store, compute, and communicate

kilo	10^{3}
mega	10^{6}
giga	10^{9}
tera	10^{12}
peta	10^{15}
exa	10^{18}
zetta	10^{21}
yotta	10^{24}


Today's rapidly growing flood of big data represents immense opportunity for forward-thinking marketers. But to fully leverage the potential that exists within these massive streams of structured and unstructured data, organizations must quickly optimize ad delivery, evaluate campaign results, improve site selection and retarget ads. This is where the IBM Netezza® Factor comes into play, enabling a fluid analysis of complex data capable of unleashing a torrent of innovative, next-level ideas and results.
DRIVING MARKETING EFFECTIVENESS BY MANAGING
THE FLOOD OF BIG DATA
US$2.1
BILLION SPENT IN U.S. ON MOBILE ADS
IN 2011
MOBILE AD
4.8
TRILLION ONLINE AD IMPRESSIONS
IN 2011
ONLINE AD IMPRESSIONS
US$83.2
BILLION ESTIMATED
FOR 2012
ONLINE AD SPEND
100
TERABYTES OF DATA UPLOADED
DAILY
FACEBOOK
294
BILLION EMAILS SENT
EVERY DAY
EMAIL
230
MILLION TWEETS
A DAY
TWITTER
BIG DATA = BIG OPPORTUNITY
35
60%
80% GROWTH
IN UNSTRUCTURED DATA
2.7
ZETTABYTES OF DATA EXIST IN THE DIGITAL UNIVERSE
5
EXABYTES OF DATA GENERATED EVERY TWO DAYS
CAPITALIZING ON THIS OPPORTUNITY WILL REQUIRE:
MARKETERS CAN THEN OPTIMIZE IN FOUR PRIMARY WAYS:

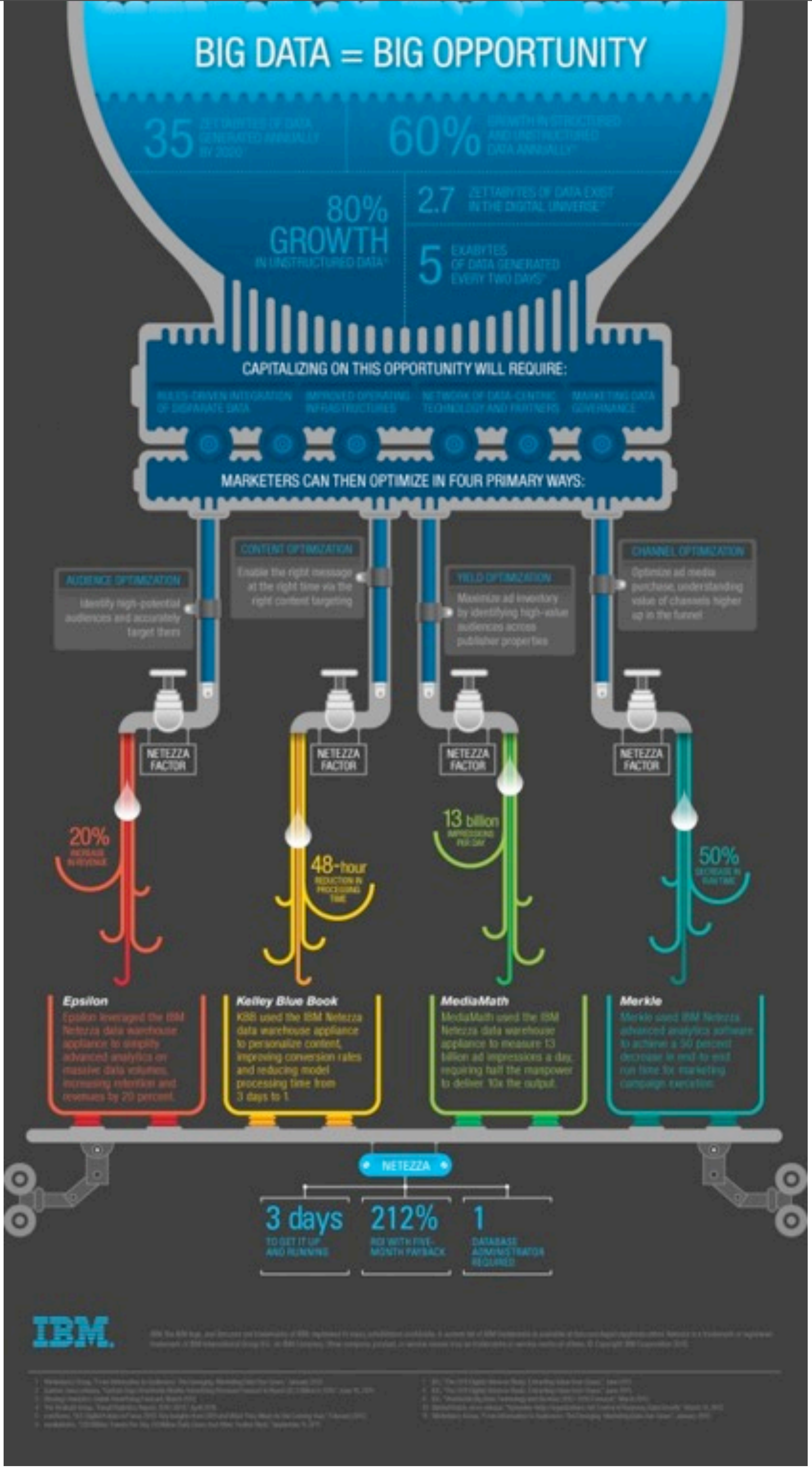
The Big Data Industry

Advertising

Capture users data

Generate users profiles

Target ads


BIG DATA = BIG OPPORTUNITY
35 ZETTABYTES OF DATA GENERATED ANNUALLY BY 2020
60% GROWTH IN STRUCTURED AND UNSTRUCTURED DATA ANNUALLY
80% GROWTH IN UNSTRUCTURED DATA
2.7 ZETTABYTES OF DATA EXIST IN THE DIGITAL UNIVERSE
5 EXABYTES OF DATA GENERATED EVERY TWO DAYS
CAPITALIZING ON THIS OPPORTUNITY WILL REQUIRE:
RULES-DRIVEN INTEGRATION OF DISPARATE DATA
IMPROVED OPERATING INFRASTRUCTURES
NETWORK OF DATA-CENTRIC TECHNOLOGY AND PARTNERS
MARKETING DATA GOVERNANCE
MARKETERS CAN THEN OPTIMIZE IN FOUR PRIMARY WAYS:
AUDIENCE OPTIMIZATION
Identify high-potential audiences and accurately target them
CONTENT OPTIMIZATION
Enable the right message at the right time via the right content targeting
YIELD OPTIMIZATION
Maximize ad inventory by identifying high-value audiences across publisher properties
CHANNEL OPTIMIZATION
Optimize ad media purchase, understanding value of channels higher up in the funnel
NETEZZA FACTOR
NETEZZA FACTOR
NETEZZA FACTOR
NETEZZA FACTOR
20% INCREASE IN REVENUE
48-hour REDUCTION IN PROCESSING TIME
13 billion IMPRESSIONS PER DAY
50% DECREASE IN RUNTIME
Epsilon
Epsilon leveraged the IBM Netezza data warehouse appliance to simplify advanced analytics on massive data volumes, increasing retention and revenues by 20 percent.
Kelley Blue Book
KBB used the IBM Netezza data warehouse appliance to personalize content, improving conversion rates and reducing model processing time from 3 days to 1.
MediaMath
MediaMath used the IBM Netezza data warehouse appliance to measure 13 billion ad impressions a day, requiring half the manpower to deliver 10x the output.
Merkle
Merkle used IBM Netezza advanced analytics software to achieve a 50 percent decrease in end to end run time for marketing campaign execution.
NETEZZA
3 days TO GET IT UP AND RUNNING
212% ROI WITH FIVE-MONTH PAYBACK
1 DATABASE ADMINISTRATOR REQUIRED
IBM.

The Big Data Industry

beyond advertising

- $300 billion/year

  US health care

- €250 billion/year

  Europe public administration

[McKinsey 2011]

Tremendous economic impact

Teraeuros (thousands billions)

First challenge: Data Harvesting

- 70% of the data produced by individuals
  - directly produced by users:
    - email, photos, blogs, etc. (less than half)
  - indirectly digital shadow/footprint:
    - surveillance, web usage, transactions
- The free paradigm of the 2.0
  - Free services traded for private user data
  - Free exploitation of the accumulated data

Second challenge: knowledge extraction

User profiles (business)

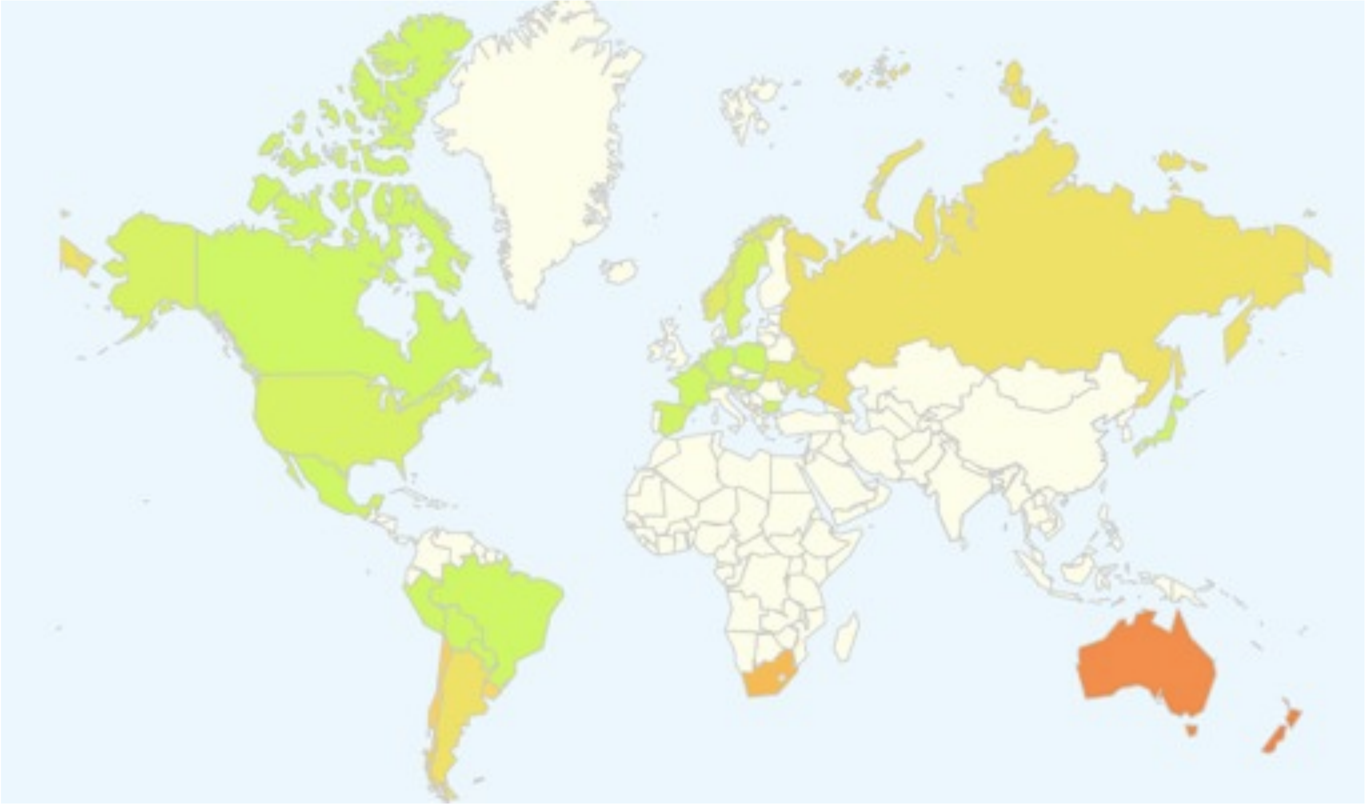
=> Ads target

Automatic discovery (science)

=> Google Flu
monitoring of flu related queries

a search engine company knows everything

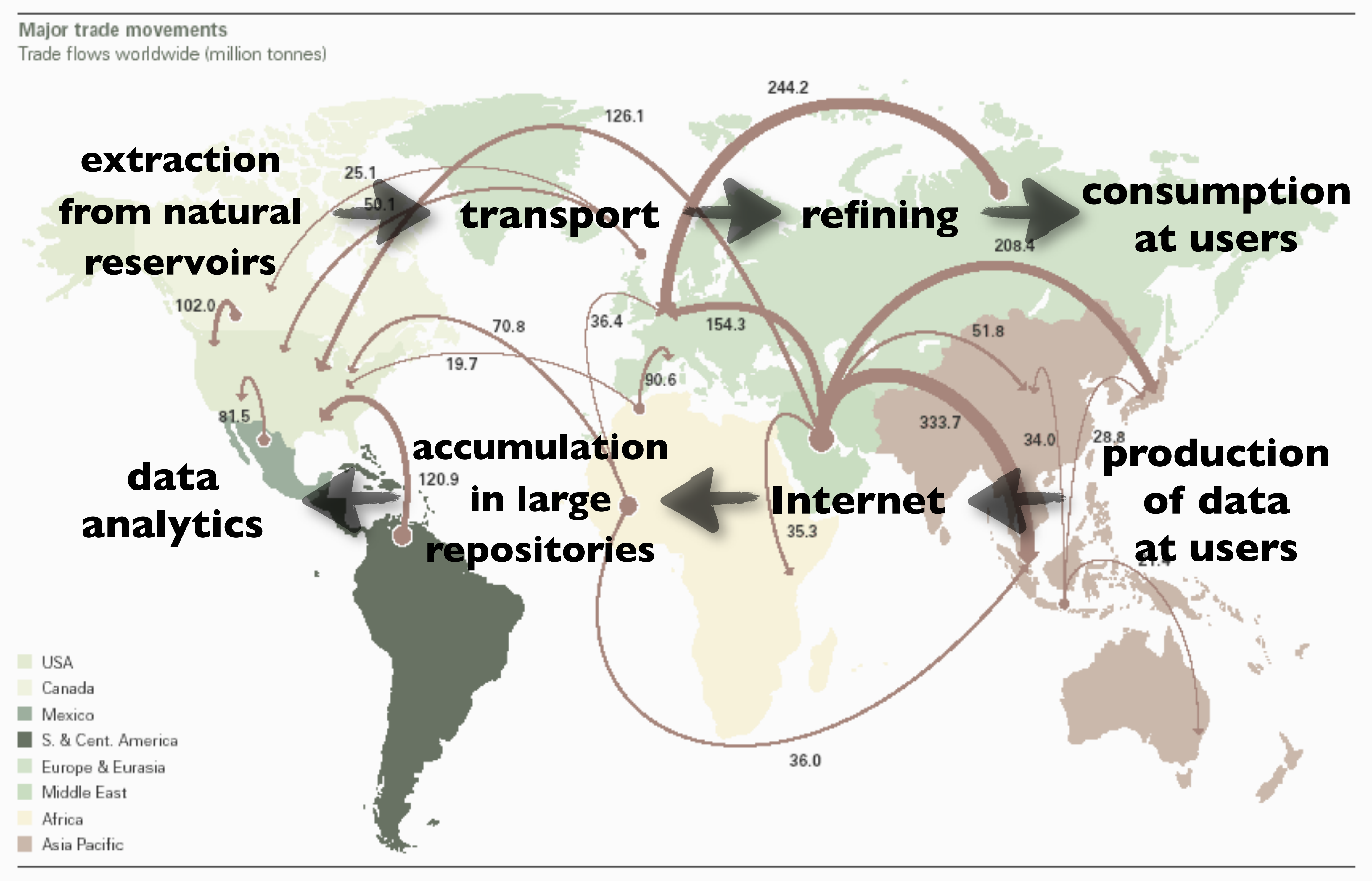
=> Biological, sociological data...



NSA (security)

=> Ambition to handle yottabytes (10^{24}) !!!

Data: raw material of the 21st century (much like crude oil)



extraction from natural reservoirs → transport → refining → consumption at users

production of data at users → Internet → accumulation in large repositories → data analytics

Where are these data?

Huge concentration of data

85% of data handled by (large) corporations

- Virtualization/dematerialization of infrastructures
- Social networks, Cloud, ...

Most of the prominent corporations based in the USA

- Google, Facebook, Amazon, Twitter, ...
- Storage capacity of Europe = 70% USA [McKinsey 2011]

1/3 of world data stored in the cloud by 2020

Geopolitics of big data

Data from the Web 2.0

produced by users everywhere in the world

but accumulated by corporations most often abroad

Percentage of national web corporations among top 25 by country

- USA: 100%
- China: 92% (only Google makes it in the top 25)
- France: 36% (but mostly marginal sites, not data intensive)

  leboncoin, Orange, Free, commentcamarche, lemonde, lequipe, lefigaro, pagesjaunes, sfr

Alexa.com

Geopolitics of big data

The Top 50 websites worldwide

- USA: 72 %
- China: 16 % (Baidu: 5; QQ: 8; Taobao: 13; Sina: 17; 163: 28; Soso:29; Sina weibo:31; Sohu:43)
- Russia: 6 % (Yandex: 21; kontakte:30; Mail: 33;)
- Israel: 2 % (Babylon: 22)
- UK: 2 % (BBC: 46)
- Netherland: 2 % (AVG: 47)

Geopolitics of big data

Diversity of search engines

- USA: Google: 65 % ; Bing: 15% ;Yahoo: 15%
- China: Baidu: 78% ; Google: 16%
- Russia: Yandex: 60% ; Google: 25%
- UK: Google: 91 % ; Bing: 5%
- France: Google: 92 % ; Bing: 3%

In France,

- Google has a de facto monopoly
- Google knows more about France than INSEE

The global imbalance

→ Information asymmetry

“Since asymmetries of information give rise to market power, and perfect competition is required if markets are to be efficient, it is perhaps not surprising that markets with information asymmetries and other information imperfections are far from efficient.”

JOSEPH E. STIGLITZ

Impact of the global imbalance

Regulation

- What legislations over a dematerialized global industry?
- Aren’t the rules defined by those who have the control?

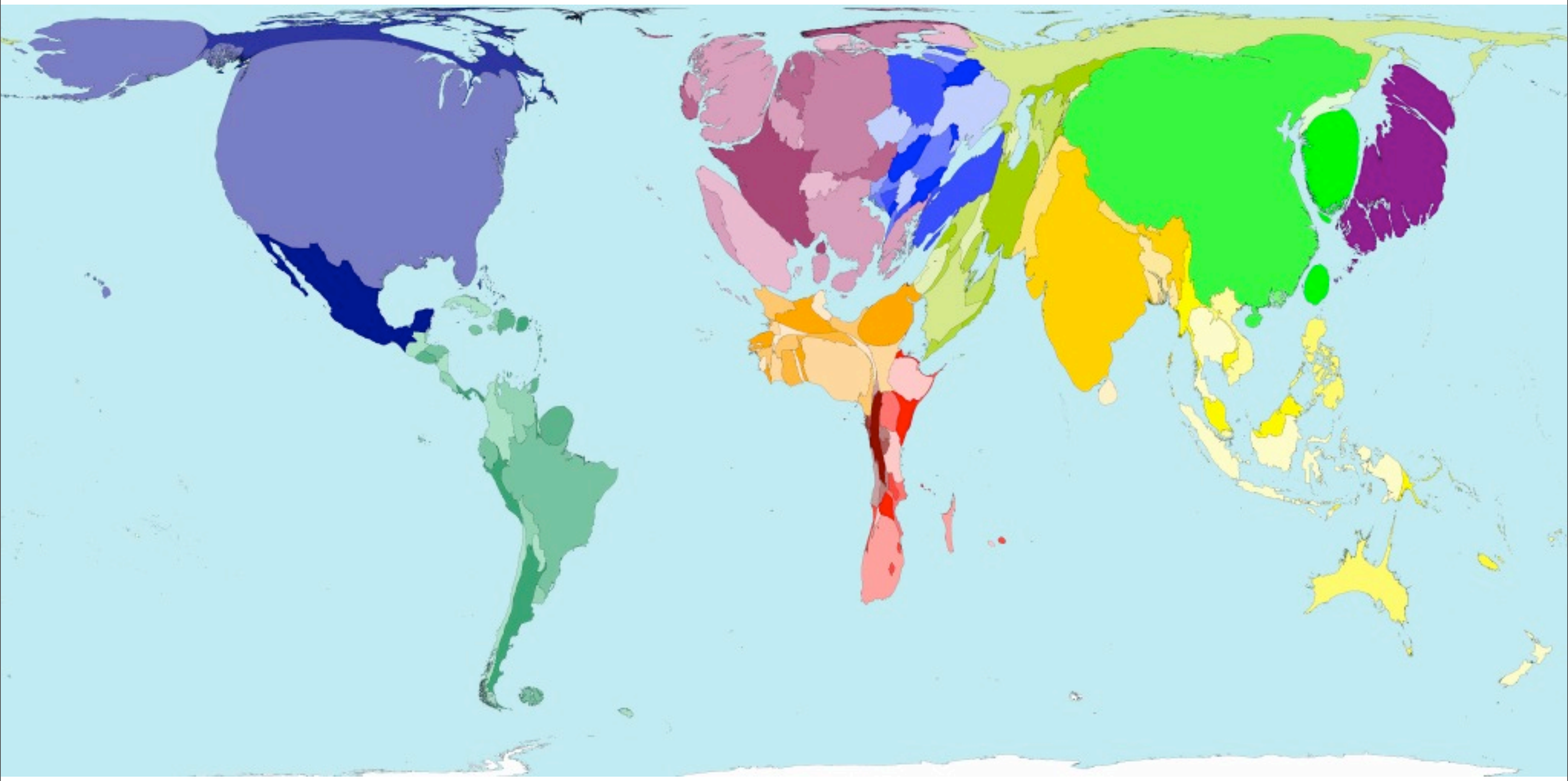
Business

- How to face monopolistic positions?
- How to handle the information asymmetry?

Security

- Data at the core of nations independence

The power of data



Map Ecological Footprint
http://www.csa.com/discoveryguides/china/review.php

What’s at stake in Europe?

Suspicion (fear?) regarding data

- concern for privacy protection high in Europe
- active legislative work
- historical reasons?

Weak industrial/innovation environment

- no strong corporation emerging

But essential dependence on foreign systems

Are there alternatives?

dominant (centralized) model

unclear privacy

lost property

active (centralized) business

little share of business capacity

Google f

decentralized 'utopian' model

high privacy

real ownership

little business

Faroo, Yacy

Diaspora

dominant (centralized) model

unclear privacy

lost property

active (centralized) business

little share of business capacity

an alternative path ?

active (competitive) business

symmetry of information

ownership & privacy

anti monopoly

decentralized ‘utopian’ model

high privacy

real ownership

little business

Faroo, Yacy

Diaspora

An alternative path for Europe?

The information society

it is only emerging

it will continue to evolve

it will impact political systems

new business models, new equilibrium will appear

Europe should embrace the future

谢谢