



Social data flows

technological, economic, and strategic challenges

Stéphane Frénot
Université de Lyon

Stéphane Grumbach
INRIA



Online life: from spatial to timeline

Our life online generates tons of data,
direct and indirect, known or unknown

These data are fueling an emerging industry

They are deeply changing our societies

They will lead to new political balances



How does that work?

At the heart of this industry:

the intermediation platforms

Technological challenges

handle huge flows with continuous service

Human challenges

ever changing world

Economic challenges

new models with extremely fast growth



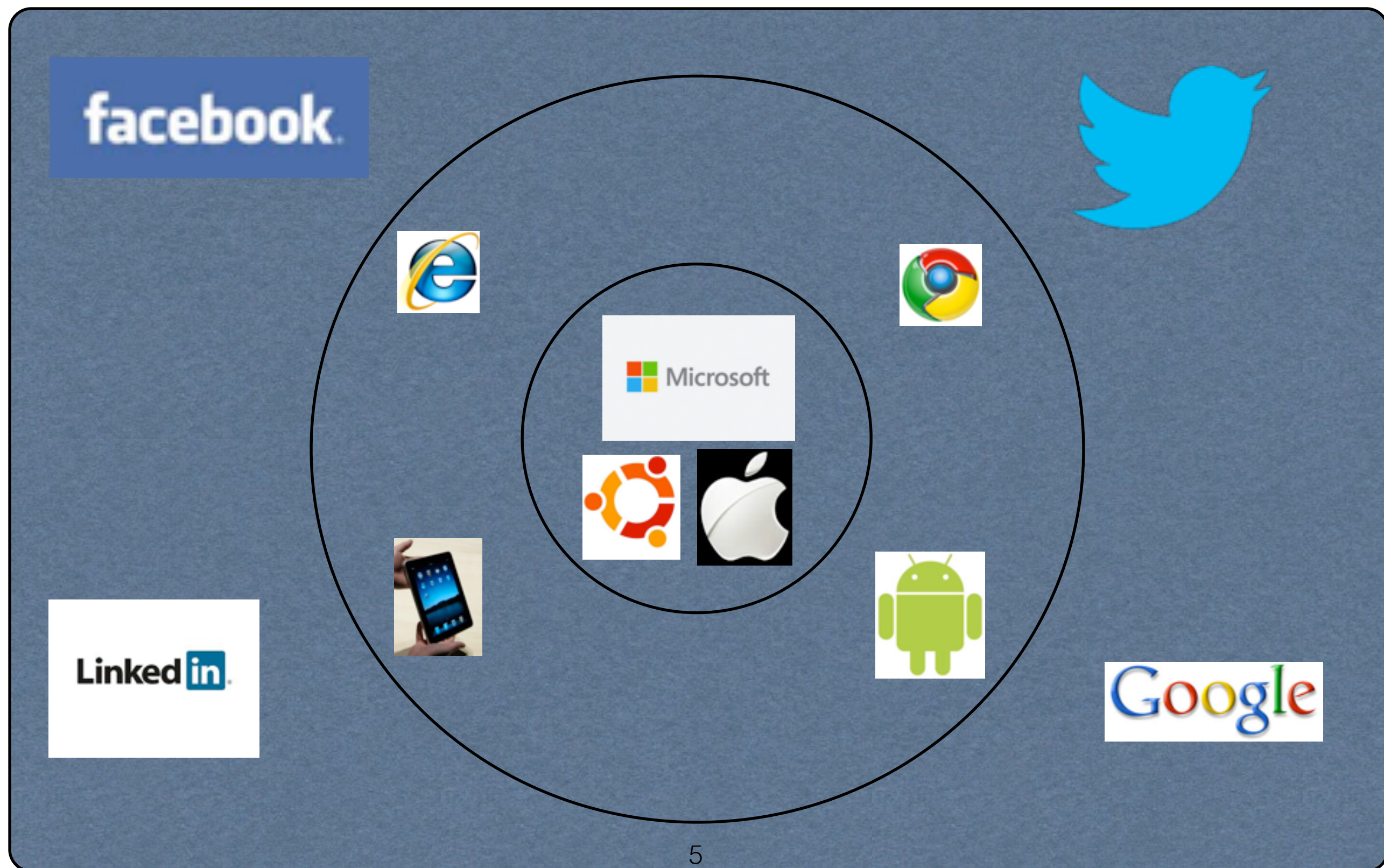
A world of agility



Built to adapt permanently vs. built to last for ever



A matryoshka of systems





Organization

1. Technological challenges
2. Agile organization & digital users
3. The economy of intermediation platforms
4. Towards a new world order



Organization

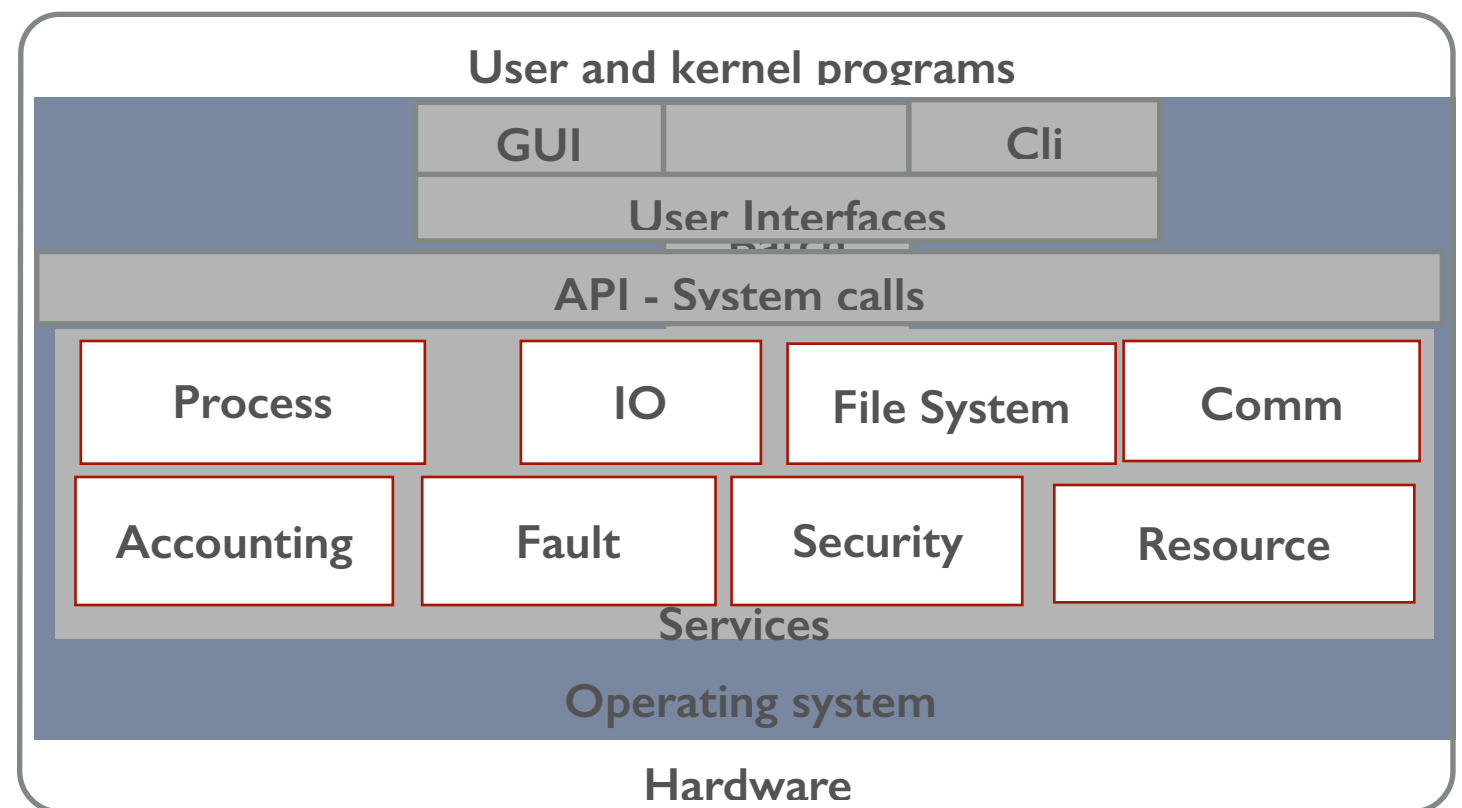
1. Technological challenges
2. Agile organization & digital users
3. The economy of intermediation platforms
4. Towards a new world order



Legacy operating systems

Conceptual paradigms

- Layering
- Programming API
- Hardware abstractions

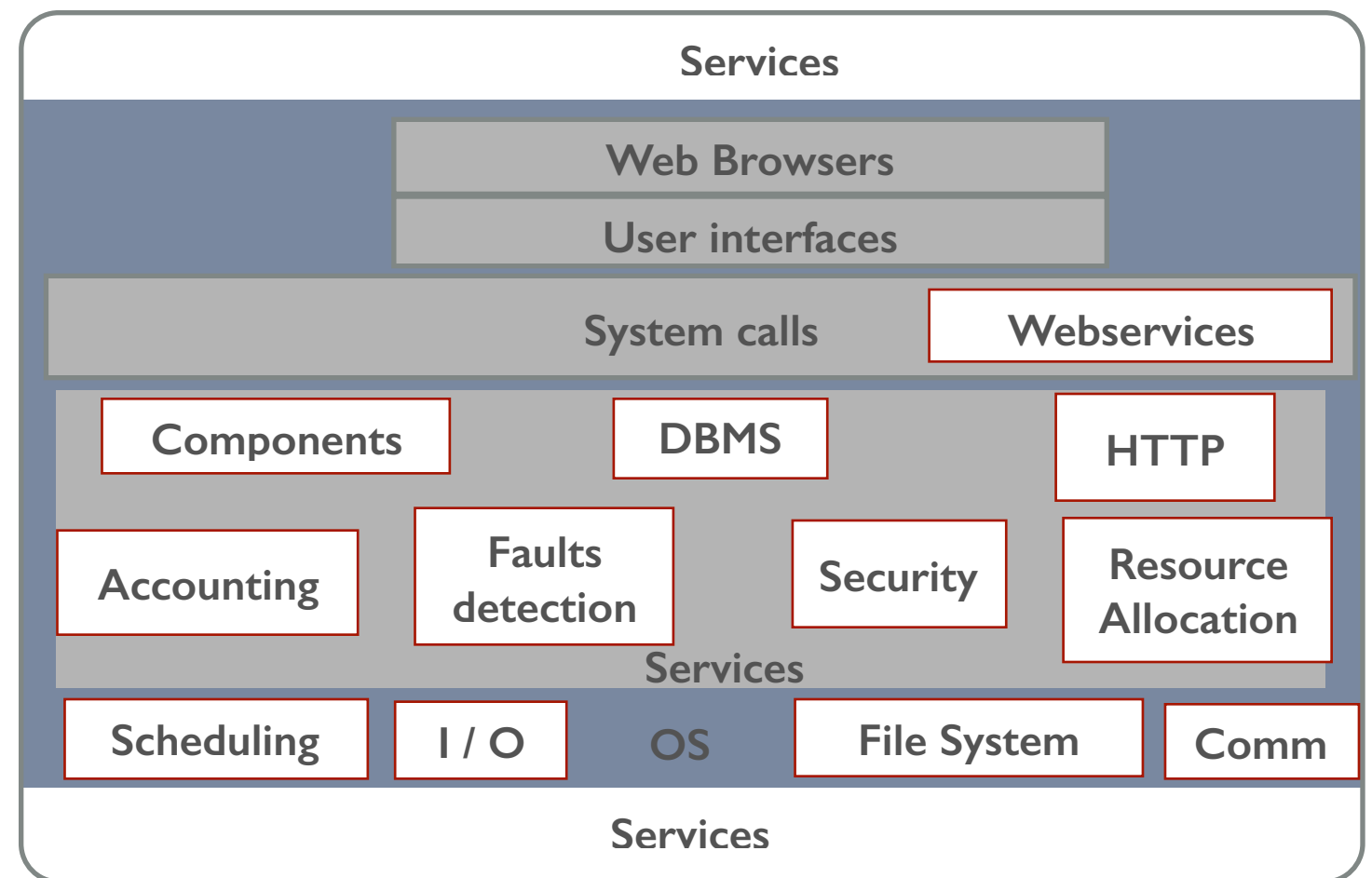




Web platforms from a «spatial» web ...

Conceptual paradigms

- Universal remote access through URI
- Big Data
- Service abstractions

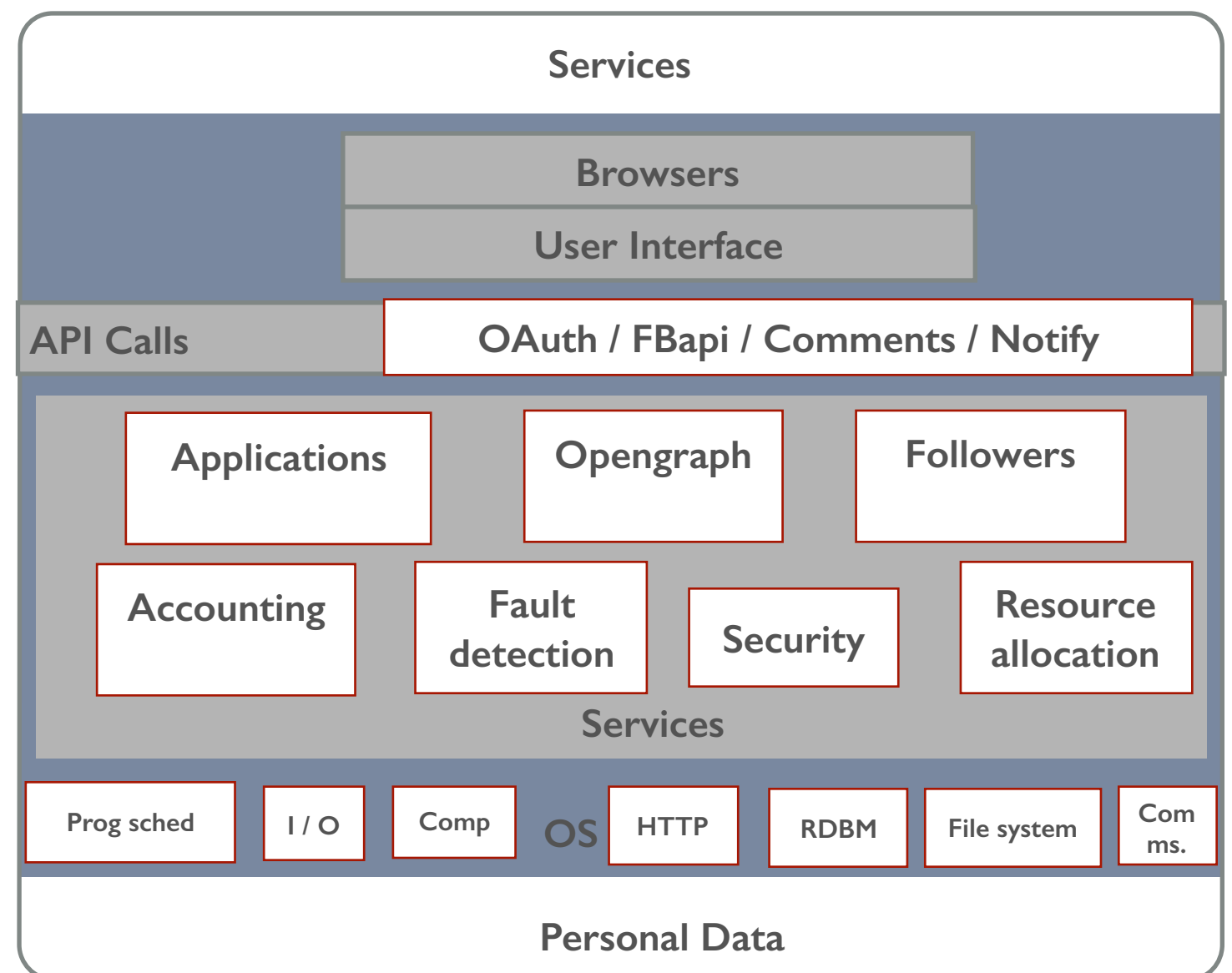




Web platforms ... to a TimeLine Web

Conceptual Paradigms

- Focus user to user interactions
- Continuous Streams
- User abstractions

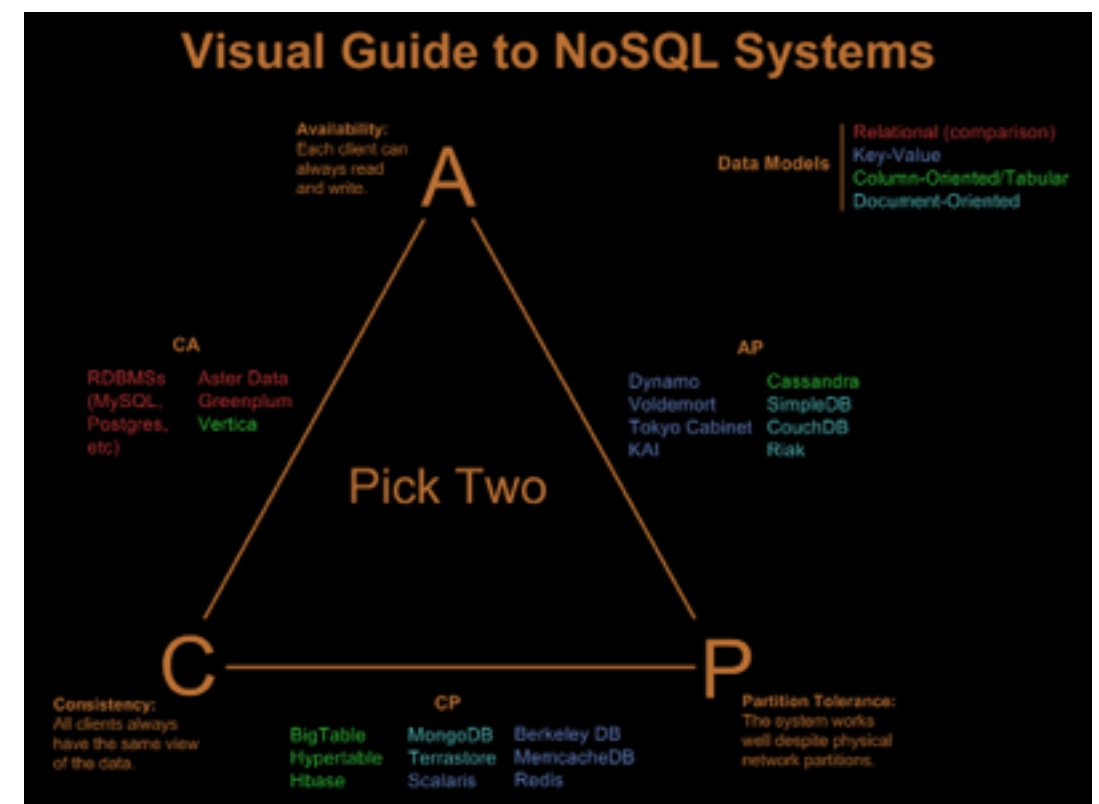


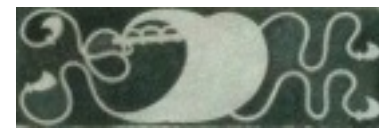


The timeline transition

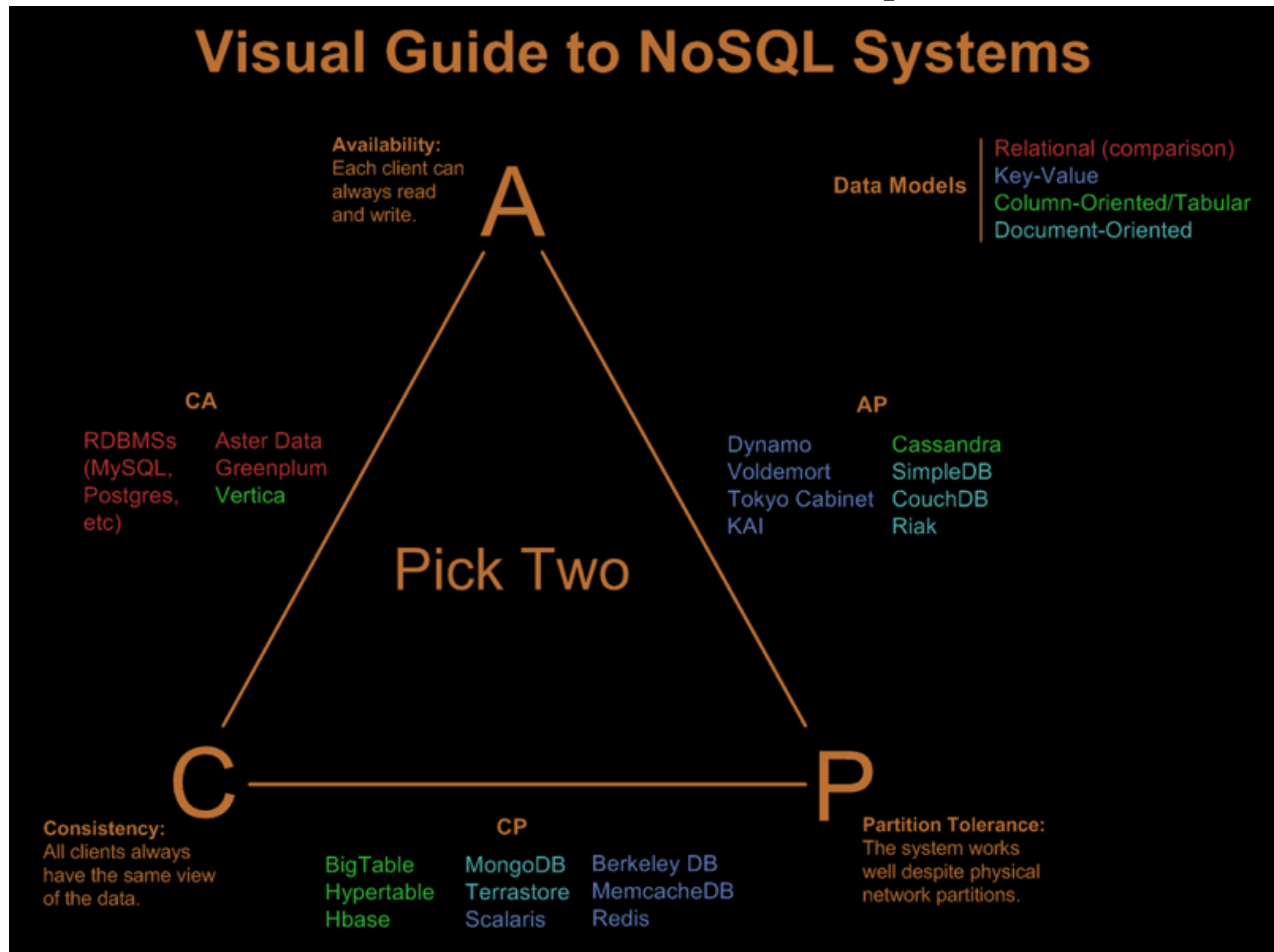
Impacts Technological Systems

- Operating Systems
- Databases
- Programming languages
- Networking





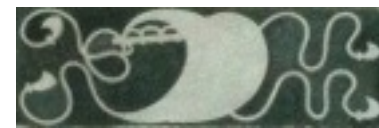
The Brewer's CAP problem



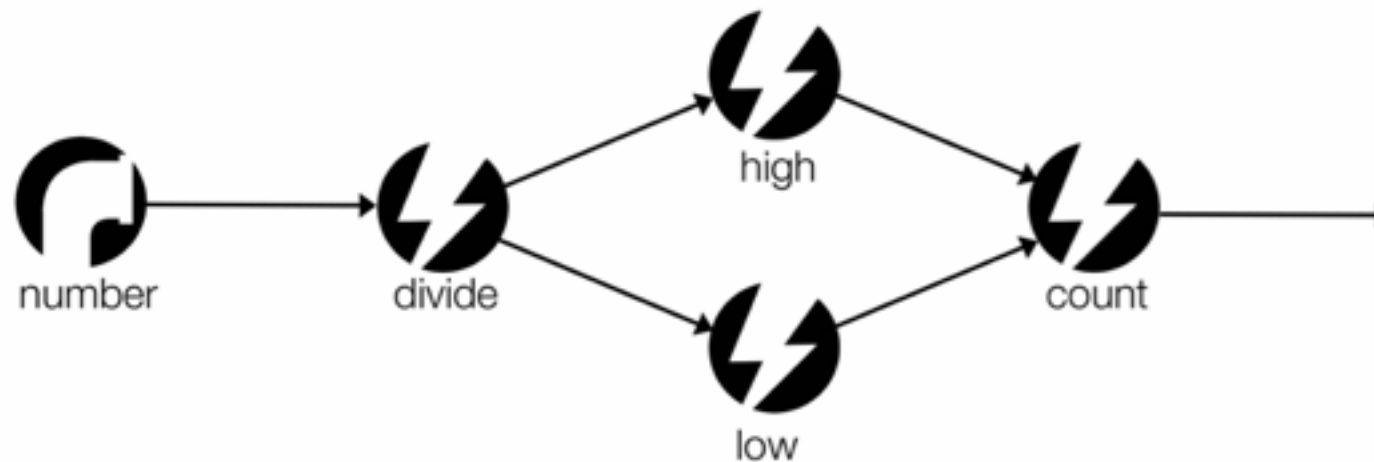


The data stream map

	Declarative	Imperative
Batch	Dryad/LINQ	MapReduce
Streams	TimeStream	Yahoo S4 Google MillWheel Twitter Storm/Trident



Data Flow

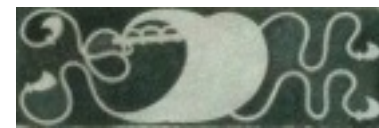


```
// Topologie
builder.setSpout("number", new NumberSpout(), 1);
builder.setBolt("divide", new DivideBolt(50), 2)
    .shuffleGrouping("number", "source");
builder.setBolt("high", new HighBolt(), 5)
    .shuffleGrouping("divide", "high");
builder.setBolt("low", new LowBolt(), 5)
    .shuffleGrouping("divide", "low");
builder.setBolt("count", new CountBolt(), 2)
    .fieldsGrouping("high", "high", new Fields("from"))
    .fieldsGrouping("low", "low", new Fields("from"));
```

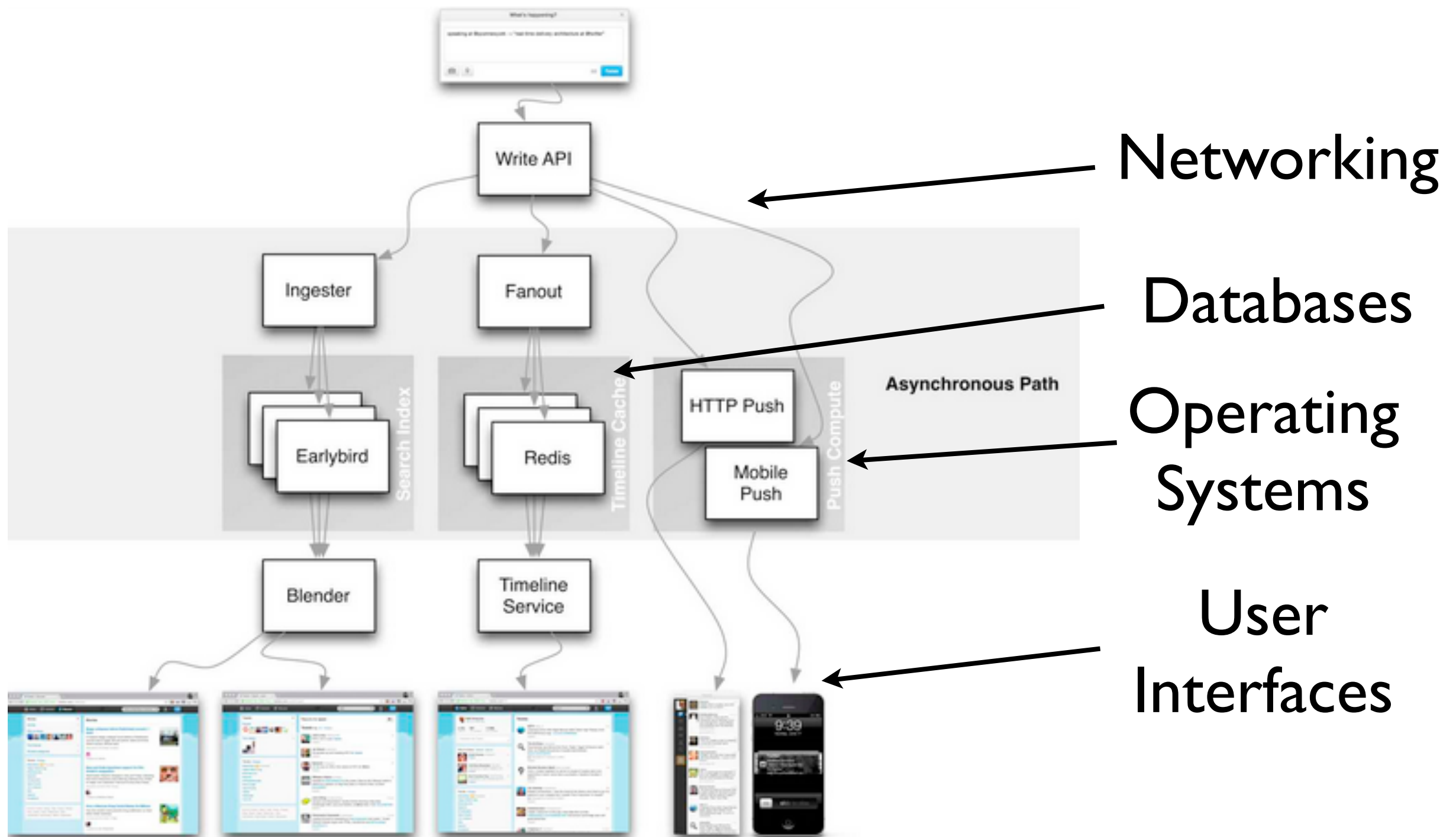
Volume Variety Velocity

+

Continuity => ! Bounded latency !



Twitter's architecture



<http://highscalability.com/blog/2013/7/8/the-architecture-twitter-uses-to-deal-with-150m-active-users.html>



Programing language shift

Flow-based Programing

Reactive-based Programing

AND

StreamIt

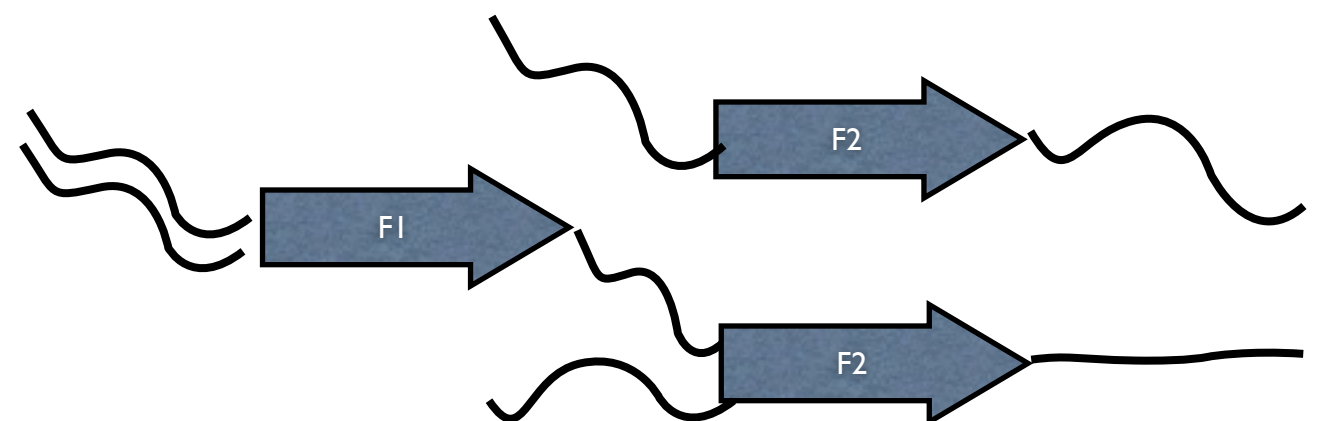
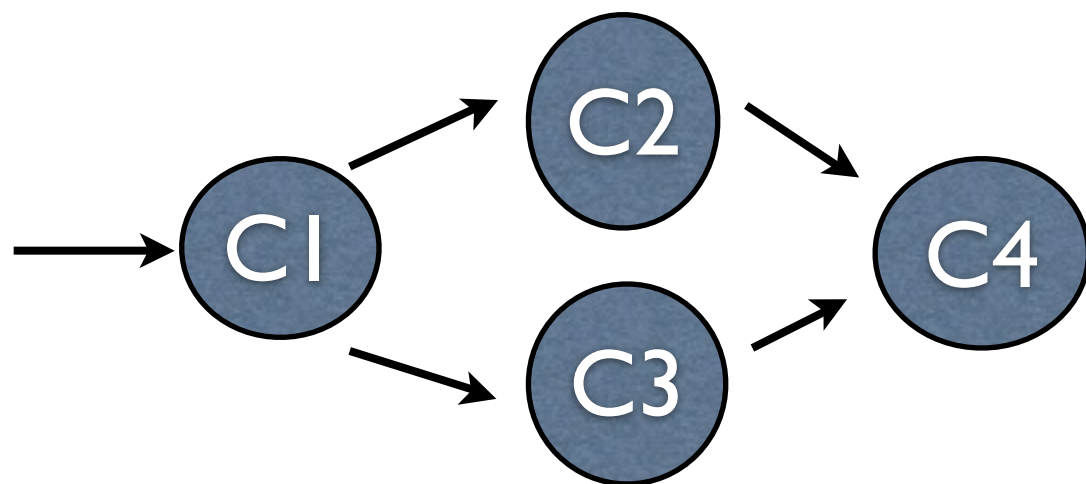


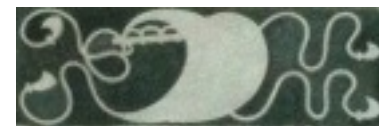
NoFlow

Lucid



Bacon.js





The Reactive Manifesto

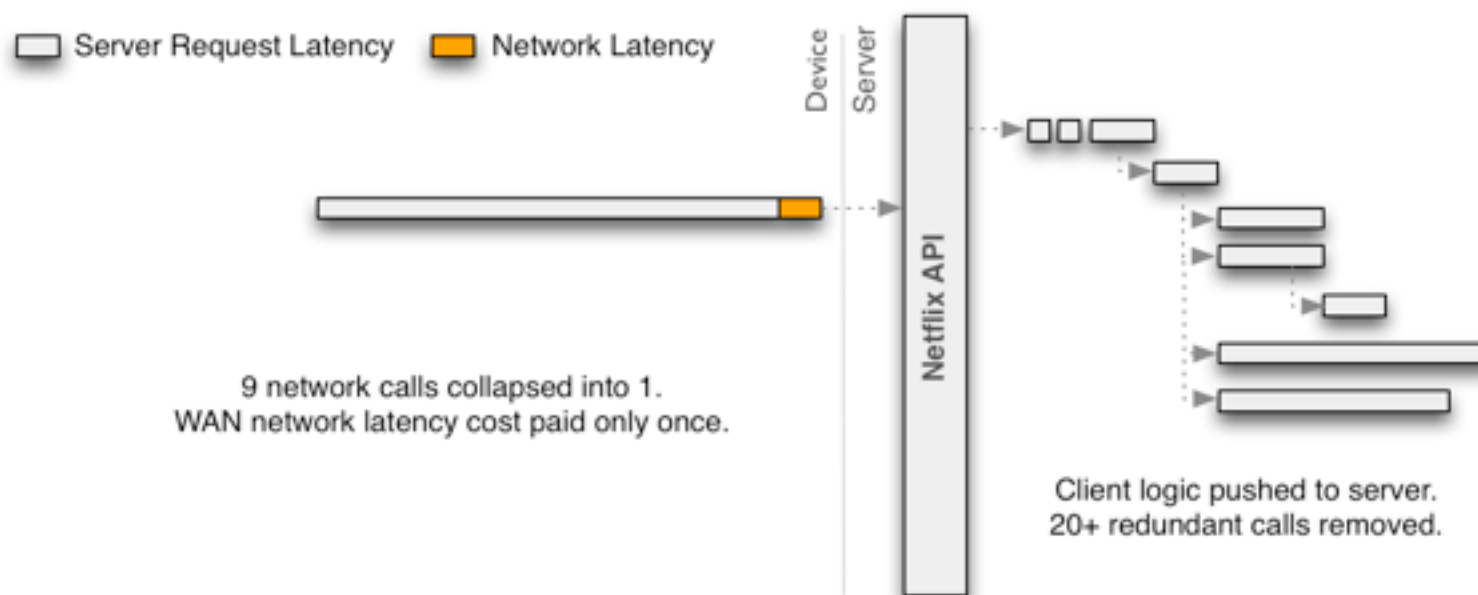
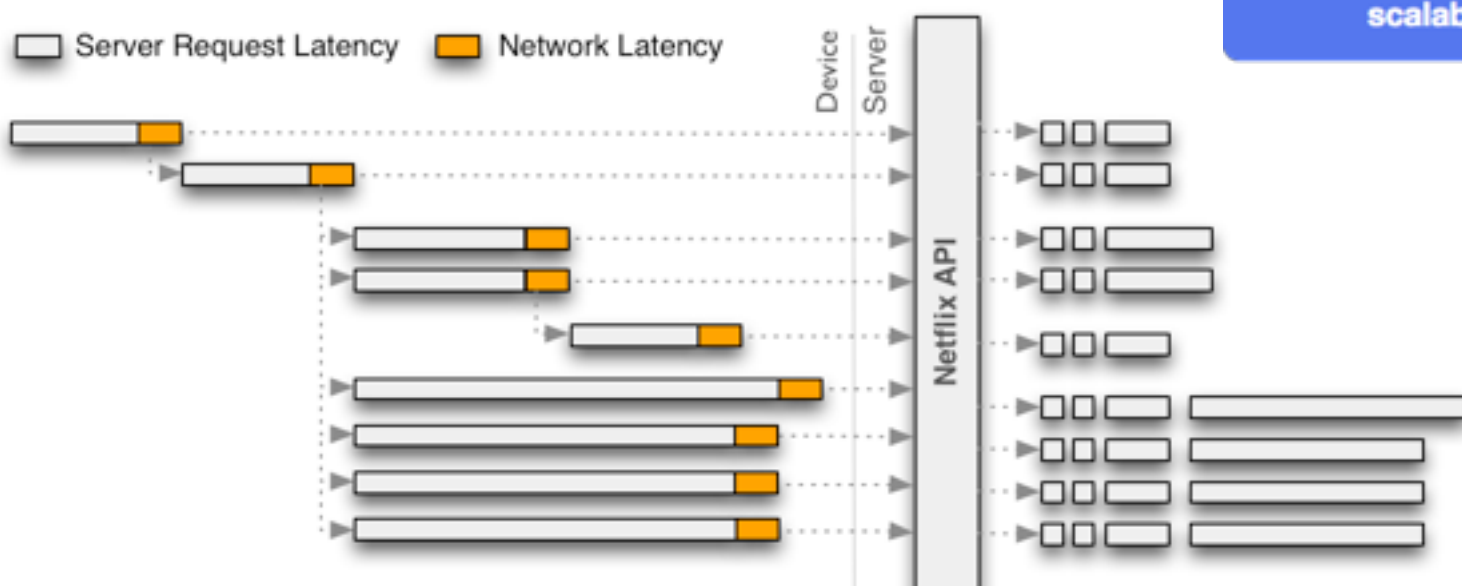
new programming languages must be

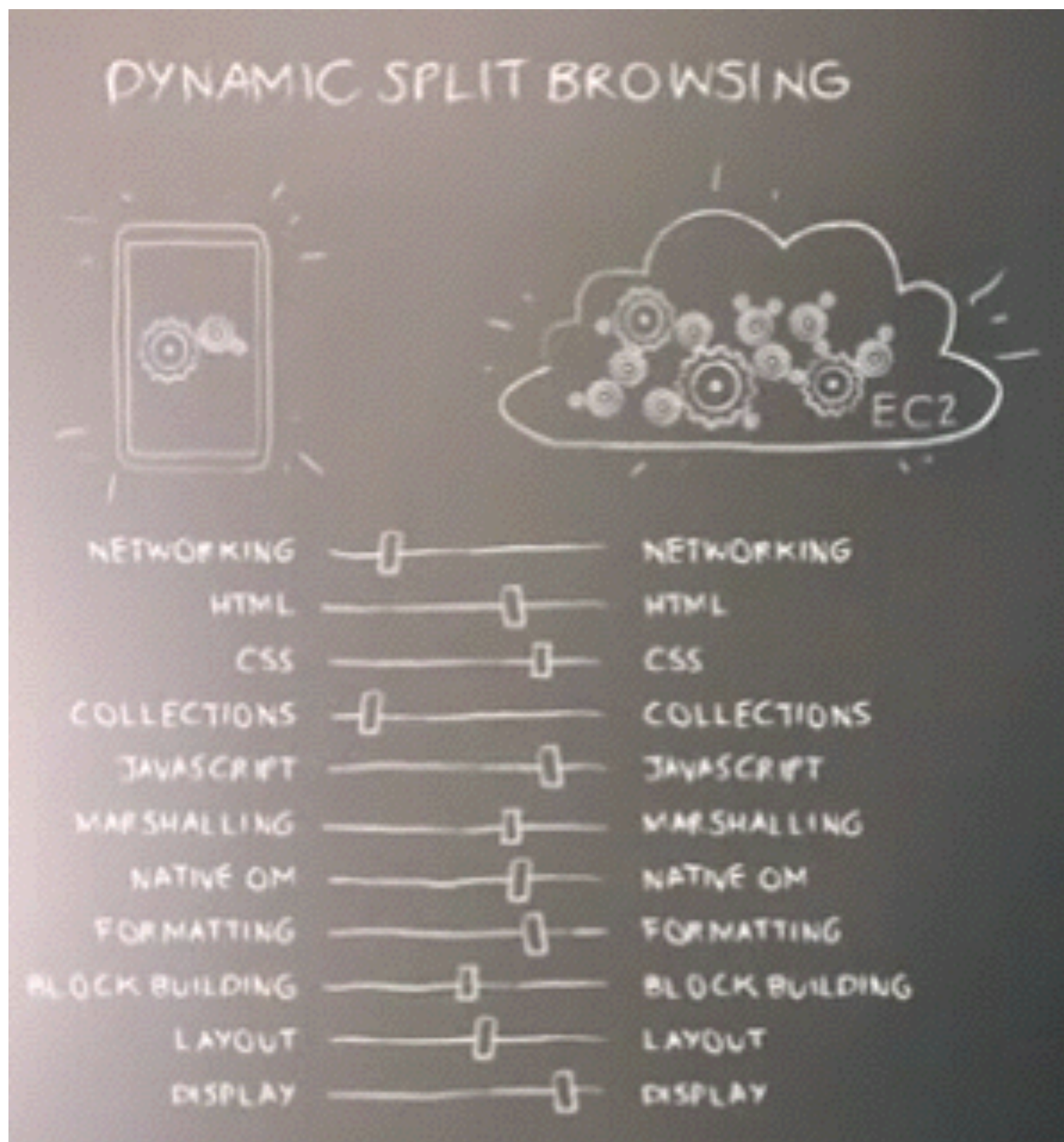
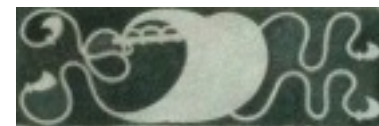
responsive

scalable

resilient

event-driven





Amazon silk web browser

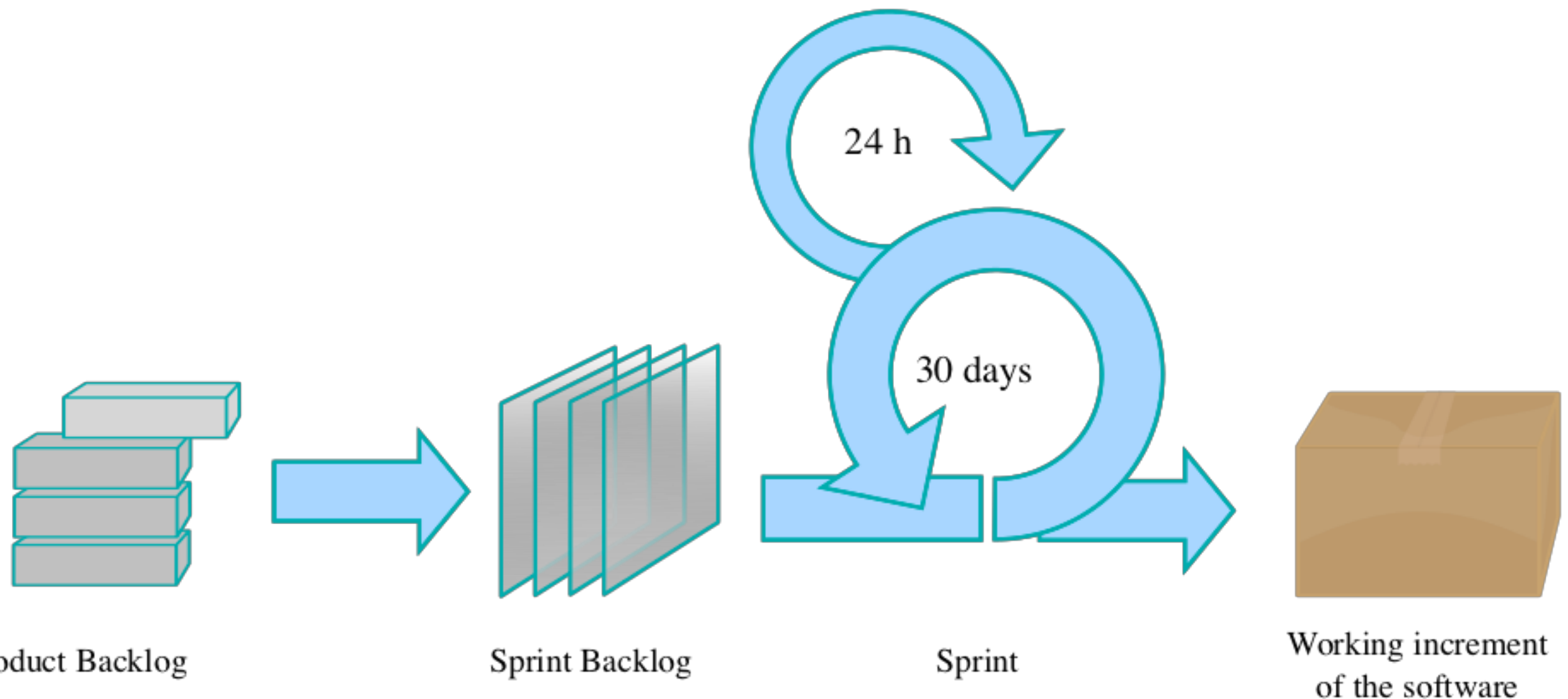


Organization

1. Technological challenges
- 2. Agile organization & digital users**
3. The economy of intermediation platforms
4. Towards a new world order



Scrum: iterative approach



Development Velocity, Sprint and Iterations aim at occupying the software space as fast as possible



Lean Manufacturing

Origin: Toyota from 1948 - 1975

Objective: Eliminate wastes in production

- Over production
- Time on hand
- Transportation
- Processing
- Stock
- Movement
- Making defective products



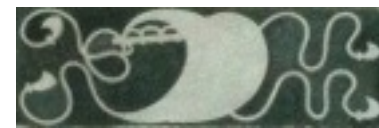
Continuously solving root problems drives organizational learning

- Kaizen (改善): Continuous improvement
- Genchi Genbustu (現地現物): Go and See
- Nemawashi (根回し): Keep all options
- Hansei (反省): Learning organization

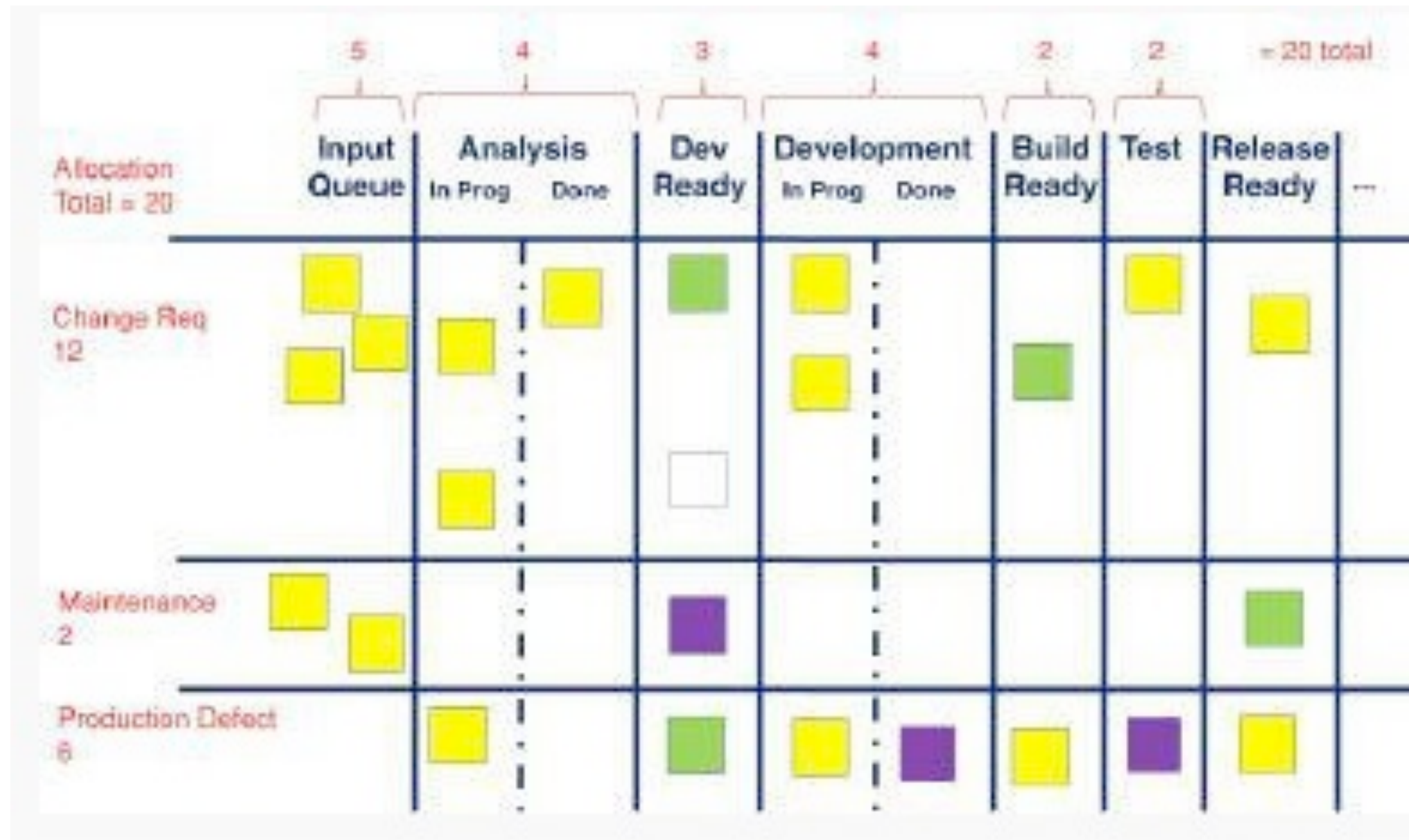


Lean Tools

- heijunka (平準化): Create continuous process flow to bring problems to the surface
- Use the "pull" system to avoid overproduction
- Build a culture of stopping to fix problems, to get quality right from the first (5 whys, Andon cord)
- Kanban (看板) Use visual control so no problems are hidden



Lean Kanban (Microsoft, 2004)

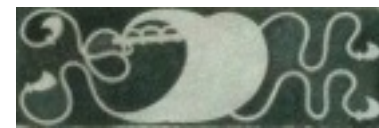


Development Pace, Lead Time and Continuous delivery aim at maintaining software improvement over long period

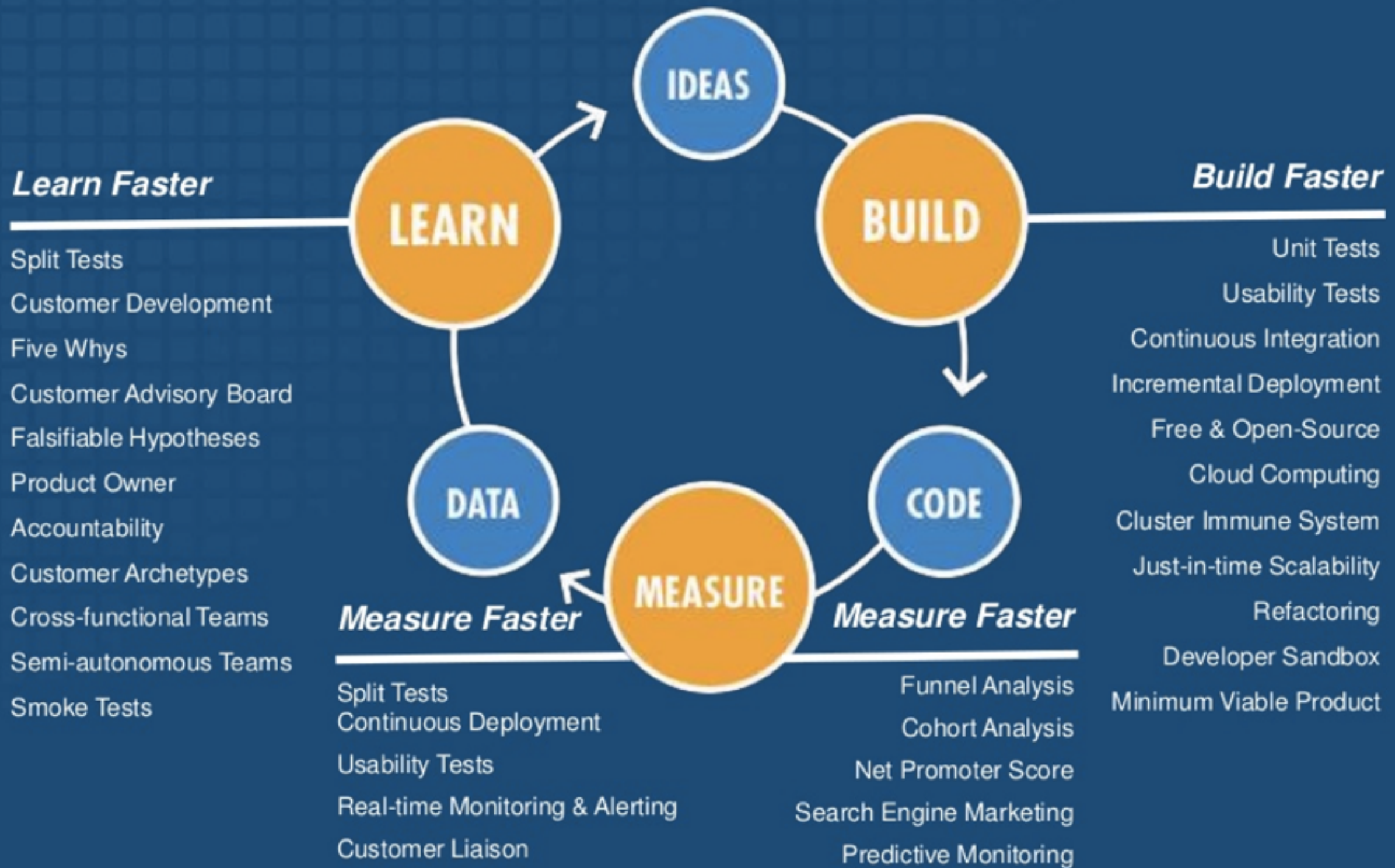


Lean Startup (Ries, 2011)

- Minimum Viable Product (MVP)
- Continuous deployment
- Split Testing
- Actionable, Accessible, Auditable metrics (AAA)
- Pivot



There's much more...





Digital users





Uninterruptible systems

Amazon / eBay / Walmart

- 1 click
- 1 hour delivery

Twitter Firehose:

- 300 k QPS
- 22 MB/s tweets that generates 57 TB / month
- Users get tweets in less than 5s (10s in 2011)
- Continuity, Immutability, Incoherency



Uninterruptible systems





Instantaneous Interactions

- 1/10th sec for Amazon leads to 1% drop in sales
- Google 0,5s latency brings down the traffic 1/5th
- 10 links per page is the best size for google responses to queries

How do they know that ?



Continuous improvement

Continuous

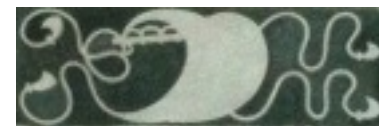
- Upgrades
- Information
- **A/B testing**
- Active user



A/B Testing

(Dan Siroker)





A/B Testing

(Dan Siroker)





A/B Testing

(Dan Siroker)



+ Videos



A/B Testing

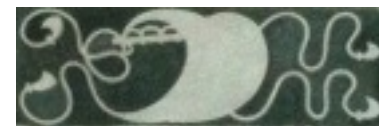
(Dan Siroker)

Combinations (24)

Page Sections (2)

Download: XML CSV TSV | Print

Relevance Rating ?	Variation	Est. conv. rate ?	Chance to Beat Orig. ?	Observed Improvement ?	Conv./Visitors ?
Button 	Original	7.51% ± 0.2%	—	—	5851 / 77858
	Learn More	8.91% ± 0.2%	100%	18.6%	6927 / 77729
	Join Us Now	7.62% ± 0.2%	73.5%	1.37%	5915 / 77644
	Sign Up Now	7.34% ± 0.2%	13.7%	-2.38%	5660 / 77151
Media 	Original	8.54% ± 0.2%	—	—	4425 / 51794
	Family Image	9.66% ± 0.2%	100%	13.1%	4996 / 51696
	Change Image	8.87% ± 0.2%	92.2%	3.85%	4595 / 51790
	Barack's Video	7.76% ± 0.2%	0.04%	-9.14%	3992 / 51427
	Sam's Video	6.29% ± 0.2%	0.00%	-26.4%	3261 / 51864
	Springfield Video	5.95% ± 0.2%	0.00%	-30.3%	3084 / 51811



A/B Testing

(Dan Siroker)

Combinations (24)		Page Sections (2)		Download: XML CSV TSV Print		
Disable	All Combinations (24) ▼		Key: Winner Inconclusive Loser			
<input type="checkbox"/> Combination	Status	Est. conv. rate	Chance to Beat Orig.	Observed Improvement	Conv./Visitors	
Original	Enabled	8.26% ± 0.5%	—	—	1088 / 13167	
★ Top high-confidence winners. Run a follow-up experiment »						
<input type="checkbox"/> Combination 11	Enabled	11.6% ± 0.6%	100%	40.6%	1504 / 12947	
<input type="checkbox"/> Combination 7	Enabled	10.3% ± 0.6%	100%	24.0%	1340 / 13073	
<input type="checkbox"/> Combination 3	Enabled	9.80% ± 0.6%	99.7%	18.7%	1277 / 13025	
<input type="checkbox"/> Combination 10	Enabled	9.23% ± 0.6%	95.9%	11.7%	1203 / 13031	
<input type="checkbox"/> Combination 8	Enabled	9.03% ± 0.6%	91.6%	9.28%	1178 / 13046	
<input type="checkbox"/> Combination 9	Enabled	8.77% ± 0.6%	81.8%	6.10%	1111 / 12672	
<input type="checkbox"/> Combination 6	Enabled	8.64% ± 0.5%	75.3%	4.58%	1108 / 12822	



A/B testing summary

- Obama campaign
 - +4 million of the 13 million e-mail addresses
 - \$75 million money raised
- Google
 - In 2011, google produced 7000 A/B tests on the search algorithm
- Amazon
 - Personalized «Impulsed buy»

No Meeting, No HiPPO



Without user data flow there is no value



Data Storage


streams



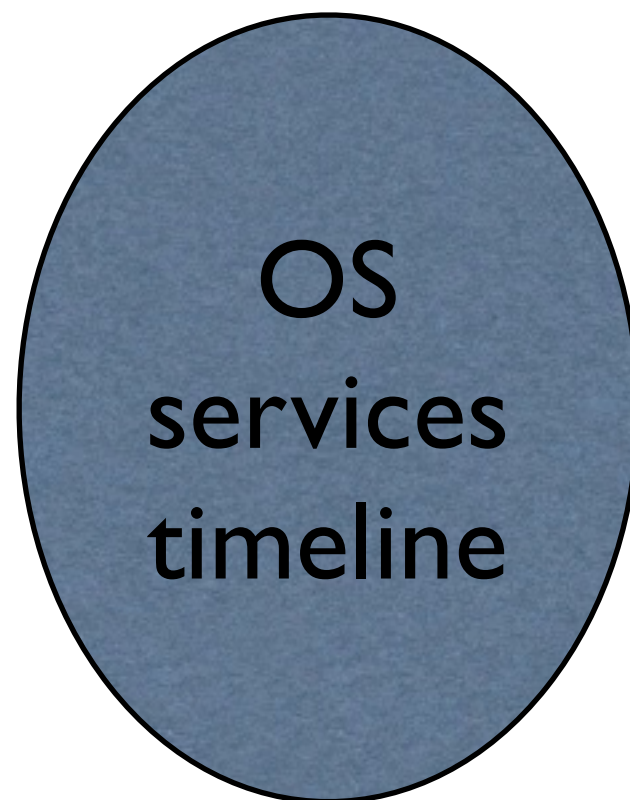
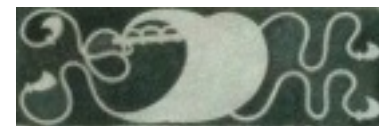
Market Place



Fast prototyping for fast mutations

- Facebook: Php - Mysql / Hop - Memcached
- Twitter: Ruby on rails / Scala
- Netflix:  / Reactive Java

Research sponsors



SOSP

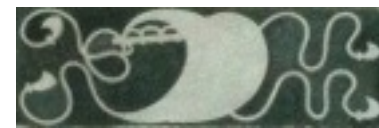
VLDB





Organization

1. Technological challenges
2. Agile organization & digital users
- 3. The economy of intermediation platforms**
4. Towards a new world order



Facebook? a new world!

A social gatekeeper

1 billion users

130 billions friend links

2.45 billion piece of content shared daily

350 millions pictures uploaded daily



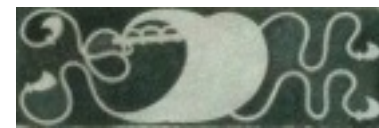
An ecosystem for the industry

storage, authentication, communication

millions of Apps developers on the API

9 million Apps active





Mobile services

Relying on a few mobile operating systems
iOS,
Android



Library of Apps
very knowledgeable
personal assistant
location based
banking, ...



Google now





Towards “appperating systems”

From devices to dematerialized environment

Squeezing between systems

Software platforms between
OSs and apps

Single interface
user / online environment

Complete data/event flow control
gatekeeper



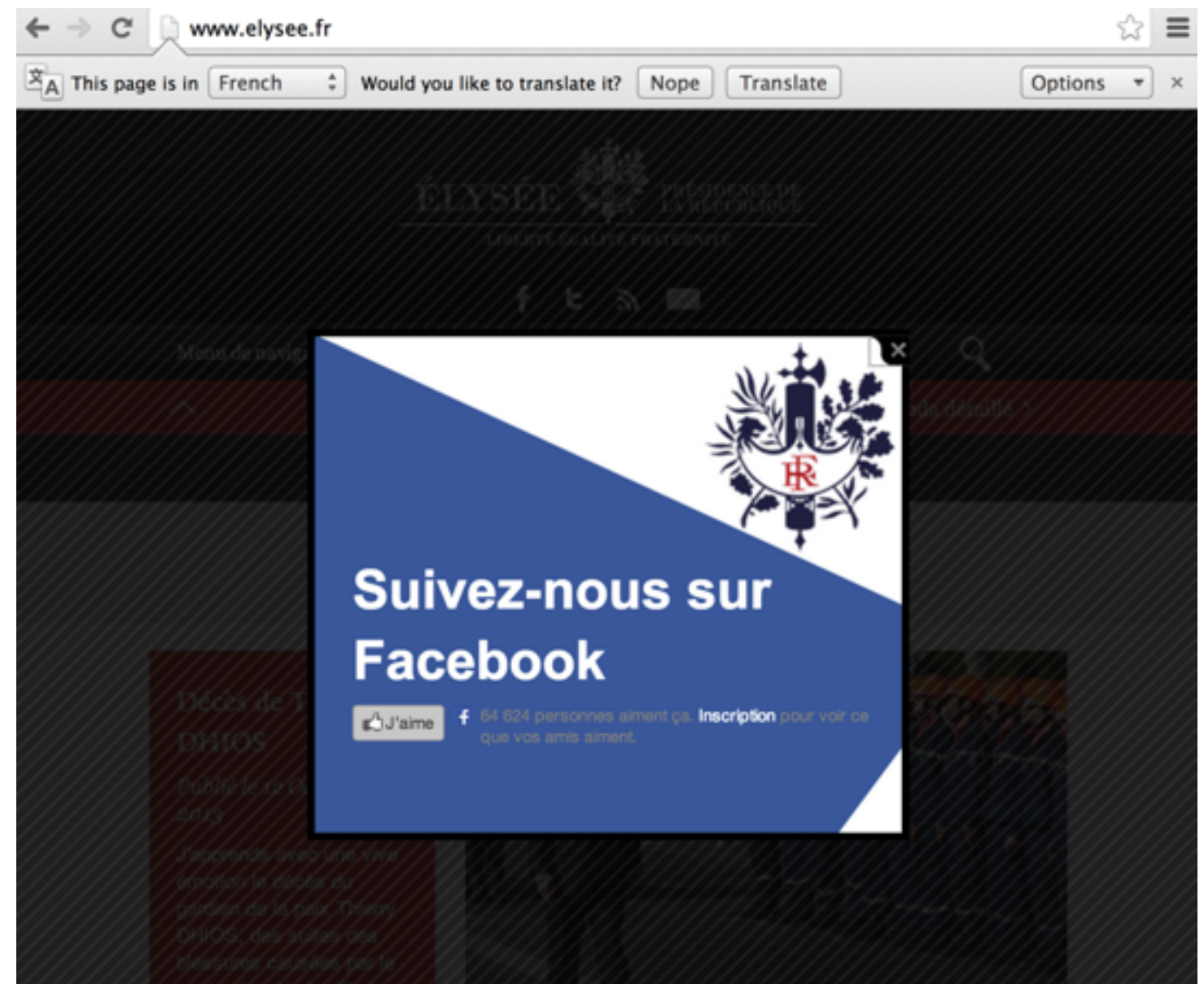
Facebook Home on top of Android



“Facebook is not cool”

Mark Zuckerberg

It is a utility
much like
electricity





Data flow at their heart

Data is a raw material,
to be transformed into value/information

Data is a money
“free” paradigm of the Web 2.0

Data can be duplicated at will
and is to ensure quality of service

Data can be transformed by people everywhere
Crowdsourcing, new open enterprises



Where are the data?

Huge concentration of data

85% of data handled by (large) **corporations**

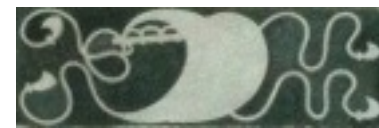
Virtualization/dematerialization of infrastructures

Social networks, Cloud, ...

Most of the prominent corporations based in the **USA**

Google, Facebook, Amazon, Twitter, ...

Storage capacity of Europe = 70% USA [McKinsey 2011]



From primary to secondary data

Primary data:

produced by users and their services
open with tunable restrictions

Secondary data:

derived from traces, activity
mostly exclusive to the platform





The challenges of the industry

First challenge:

capture users and data
scale up as much as possible

Second challenge:

capture developers and Apps
stay as open and adaptive as possible

Third challenge:

generate (undisclosed) secondary data
capture added value (business model)



Size matters exponentially

number of users of a search engine

=> traffic ↗♂

=> interest of advertisers ↗♂

=> word auction prices ↗♂

=> relevance (because of price) ↗♂

=> probability of successful click ↗♂

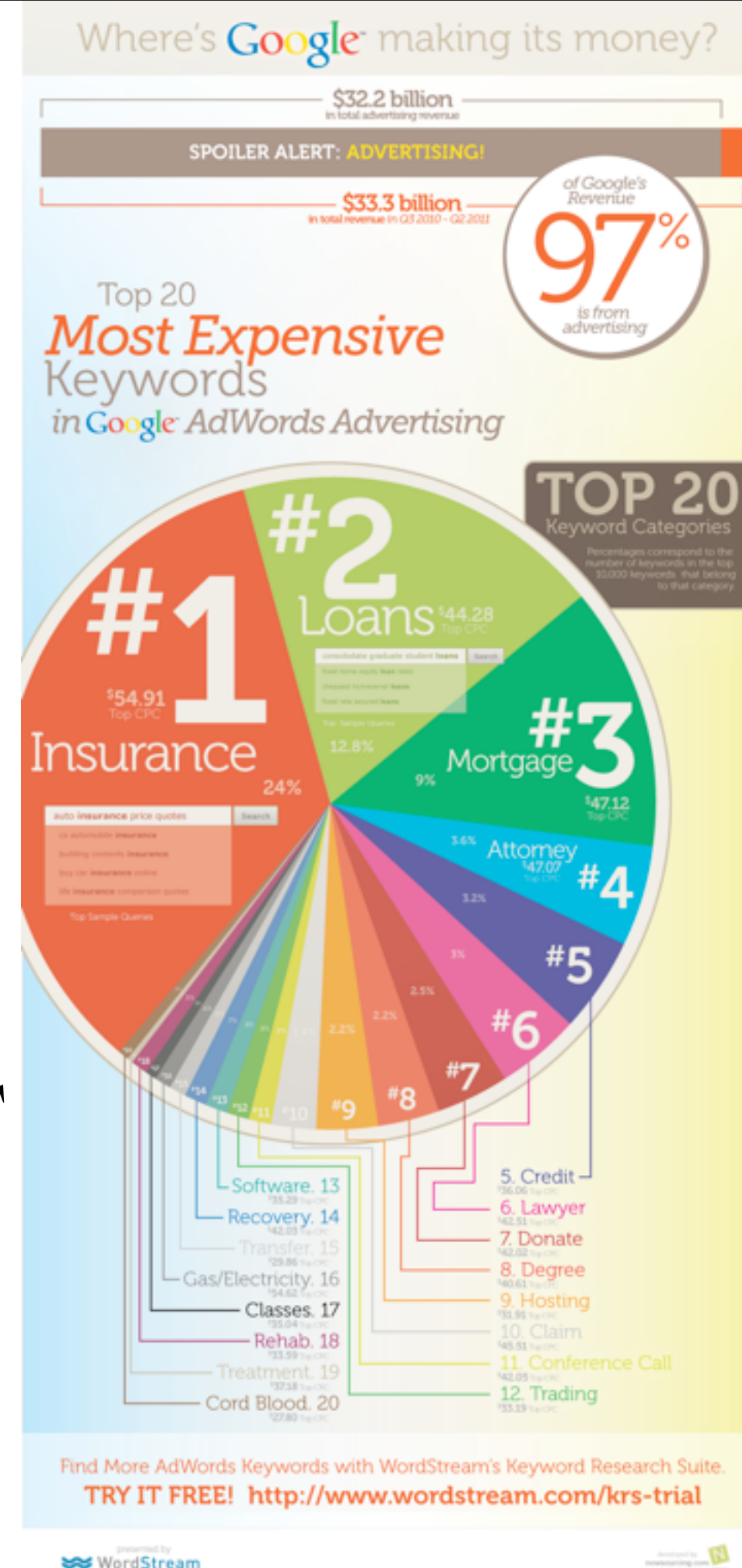
=> word covering ↗♂

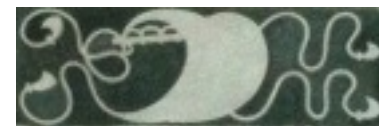
=> monetization covering ↗

Thanks to François Bourdoncle

<http://www.wordstream.com/blog/ws/2011/07/18/most-expensive-google-adwords-keywords>

50





Largest IPO

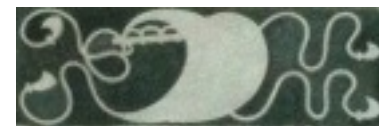
1. Agricultural Bank of China US\$22.1 billion (2010)
2. Indus. and Com. Bank of China US\$21.9 billion (2006)
3. American International Assurance US\$20.5 billion (2010)
4. Visa Inc. US\$19.7 billion (2008)
5. General Motors US\$18.15 billion (2010)
6. Facebook, Inc. US\$16 billion (2012)

Facebook: 421 Million shares x \$38 = \$16 billion

Google: 19 Million shares x \$85 = \$1,6 billion

Twitter: \$1 billion (expected)

Microsoft: 1986 -> 2004 no dividend

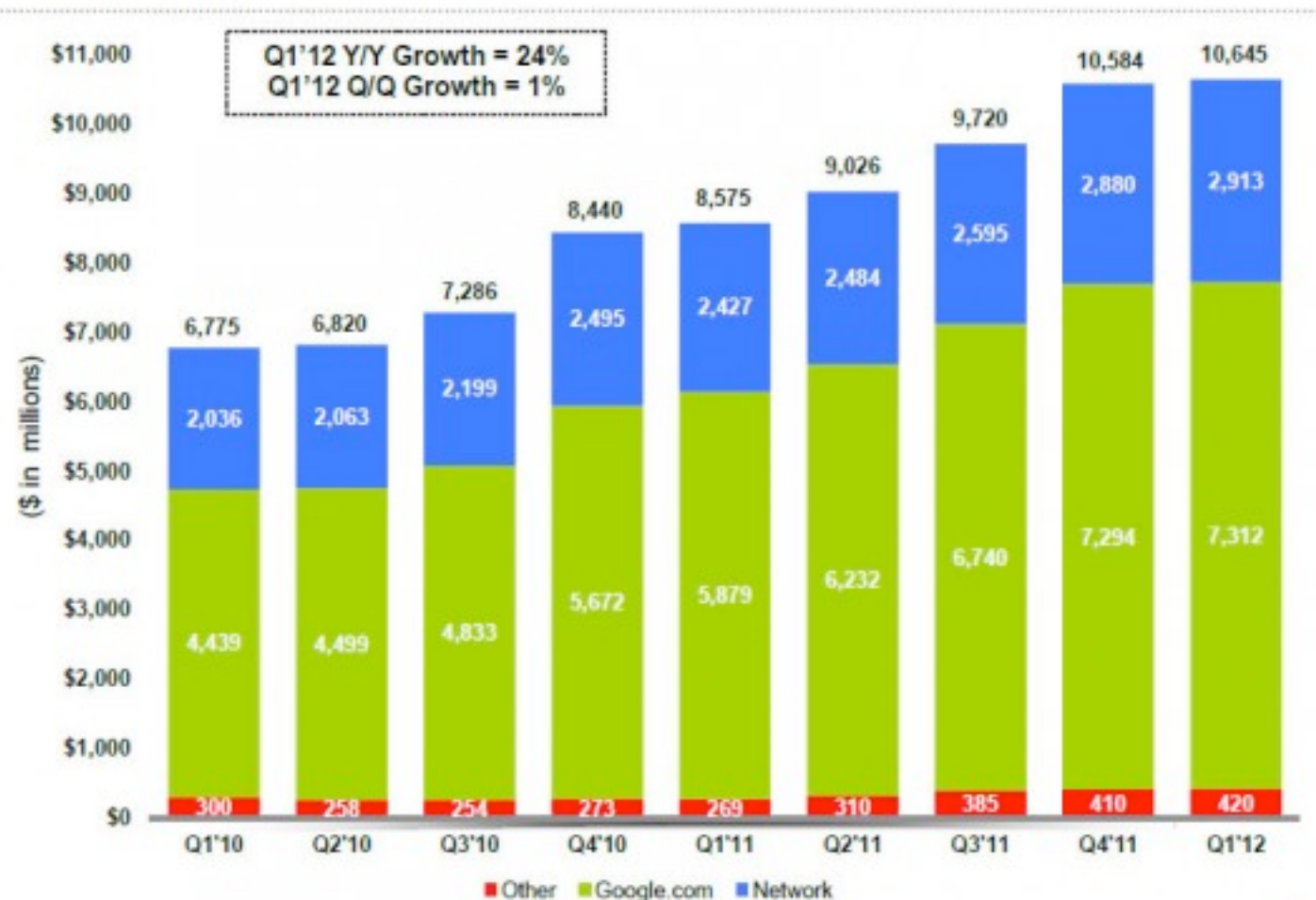


“Google is a Vacuum cleaner for revenue”

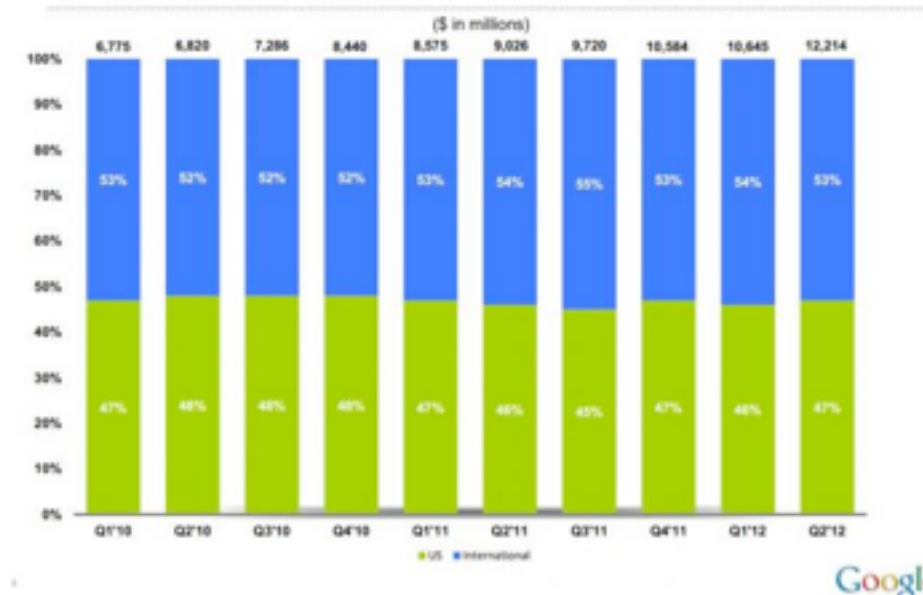
Barry Diller



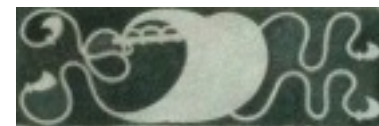
Quarterly Revenues



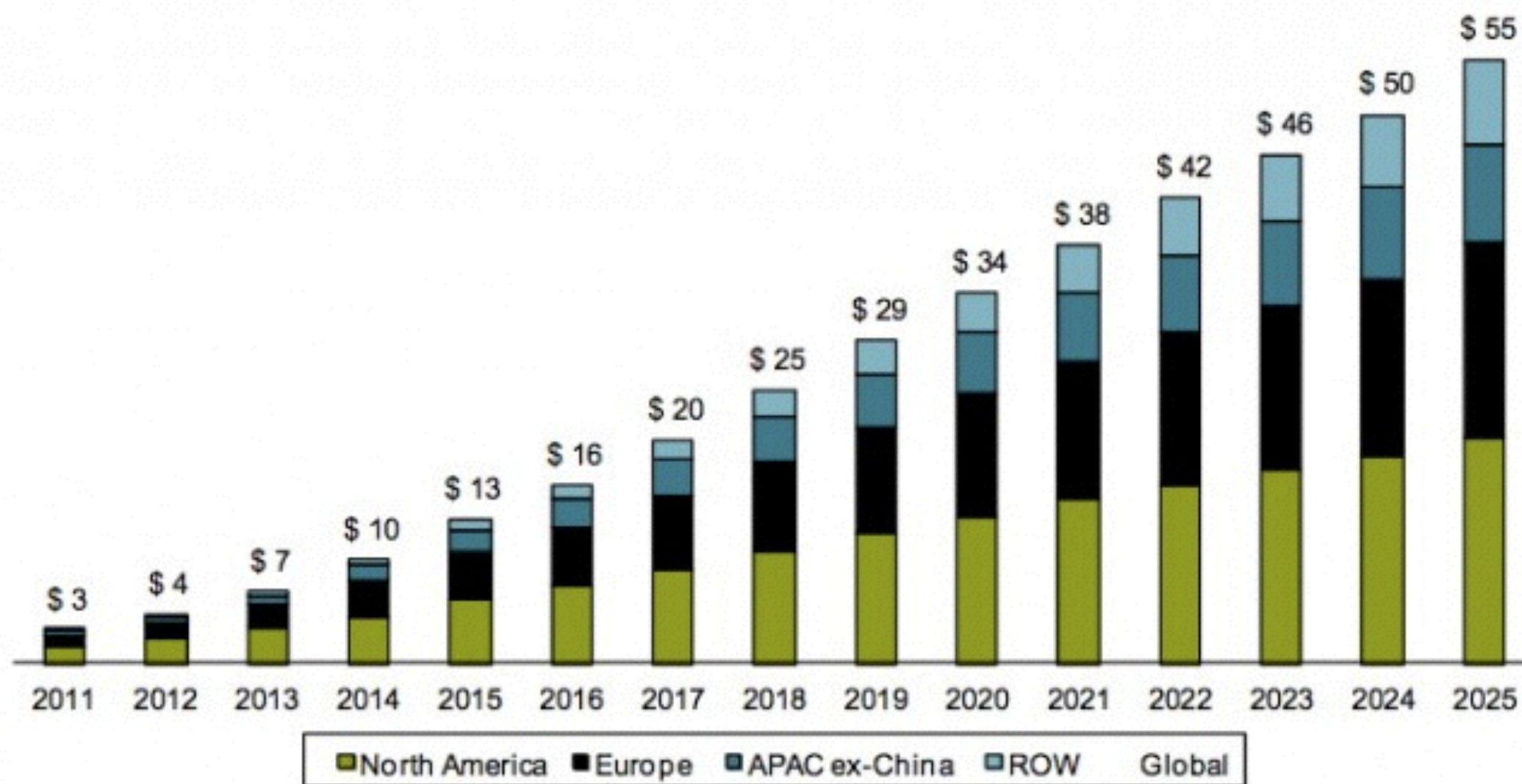
U.S. vs. International Revenues - Consolidated



\$50 billion in 2012



Facebook revenue forecast





Web Giants, Corporations or States?

What is a state?

population

territory

government

sovereignty / diplomacy

money

defense

legal system



Facebook's territory





Internet giants as Extraterritorial powers

No real binding to the place of operation

Regulation, taxation: optimal use of national differences

Own raw material resources and industry
harvested without borders

Own legal systems
contracts users/corporations

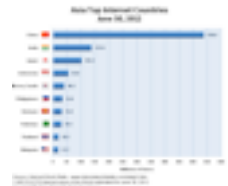
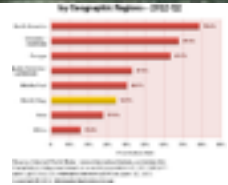
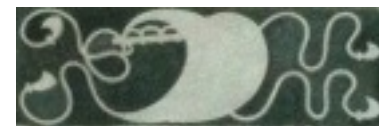
Own monetary systems
emerging virtual currencies



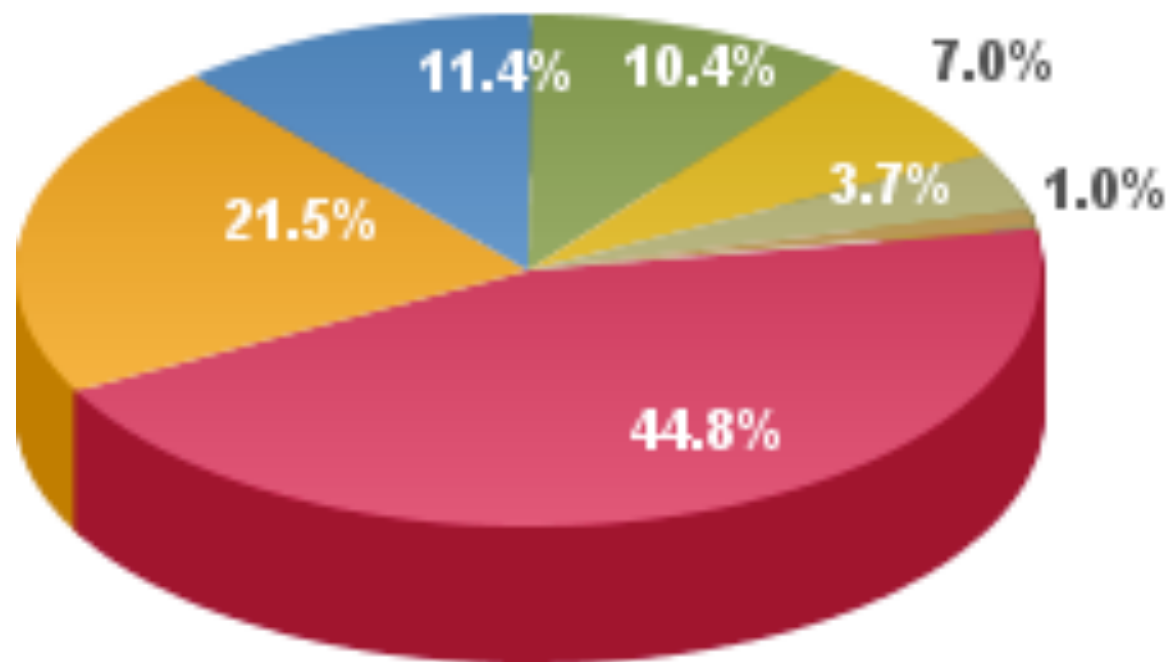


Organization

1. Technological challenges
2. Agile organization & digital users
3. The economy of intermediation platforms
4. Towards a new world order



Internet Users in the World Distribution by World Regions - 2012 Q2

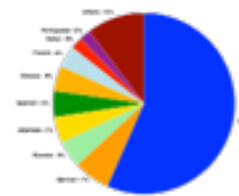
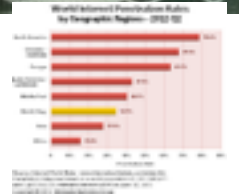
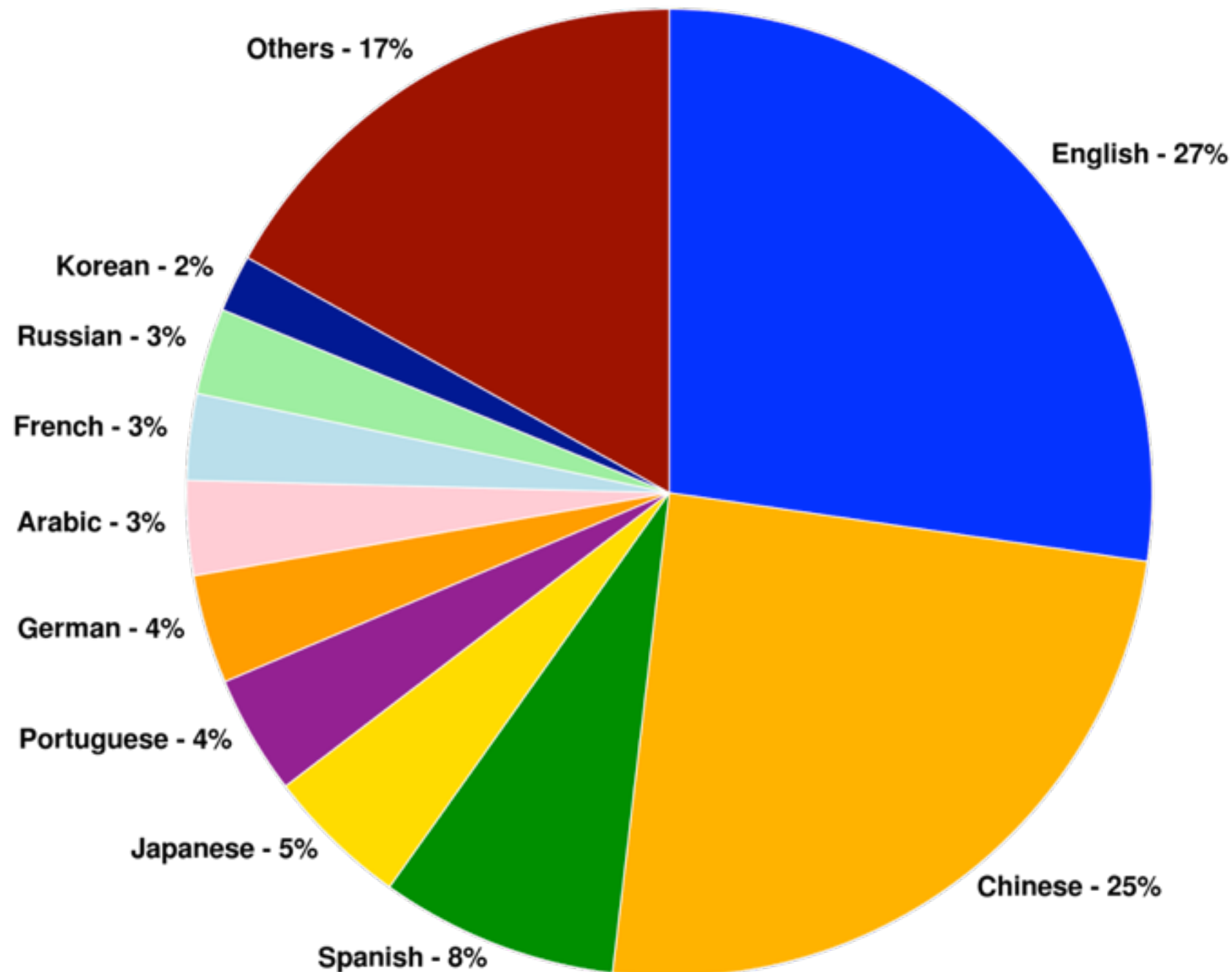


Source: Internet World Stats - www.internetworldstats.com/stats.htm

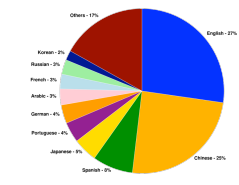
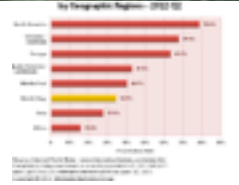
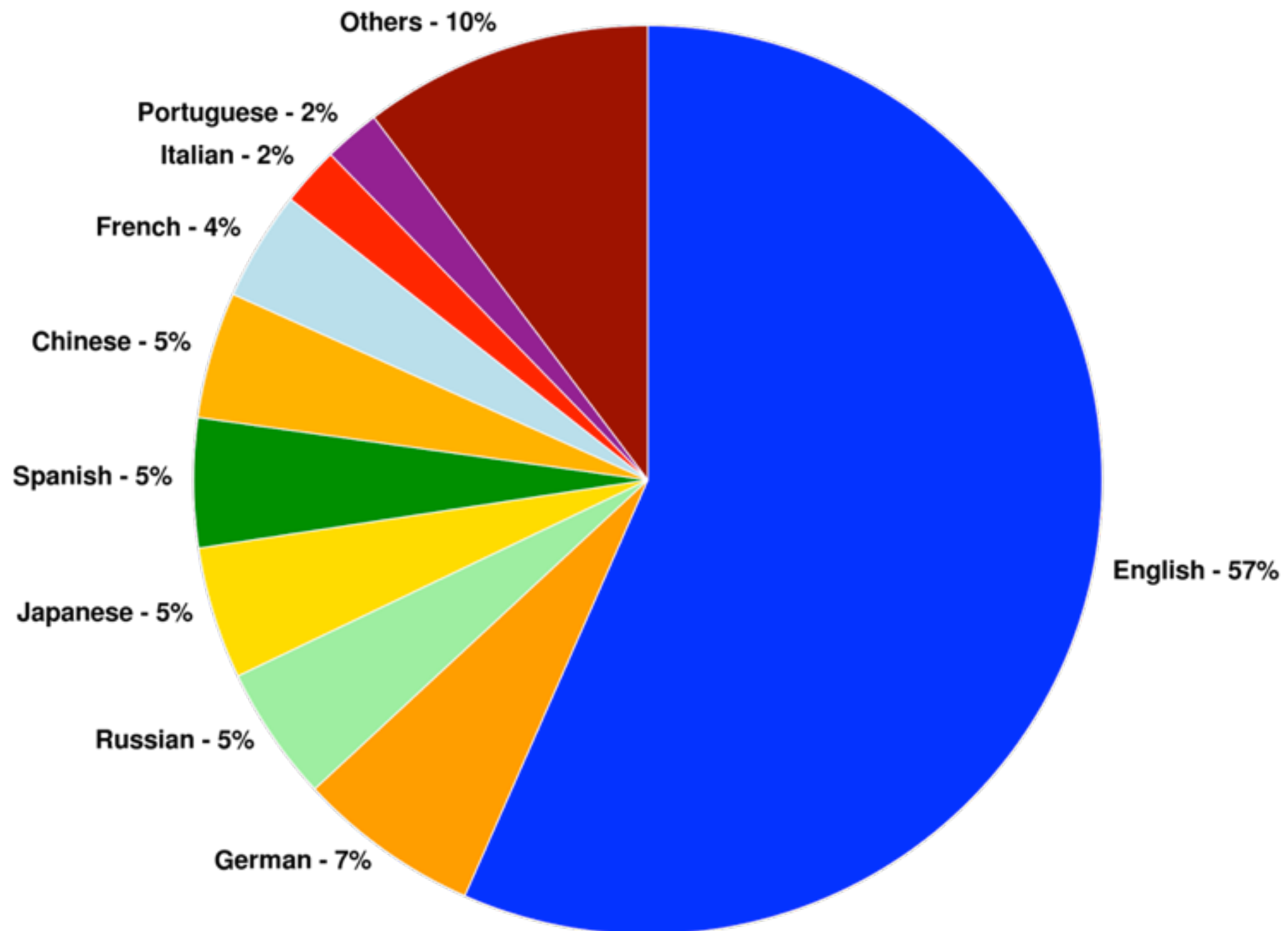
Basis: 2,405,518,376 Internet users on June 30, 2012

Copyright © 2012, Miniwatts Marketing Group

Online Population



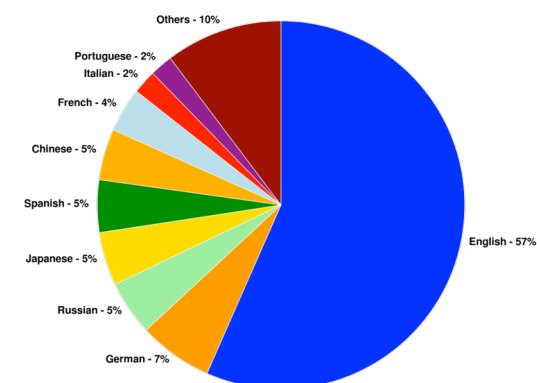
Web content language





High penetration and impact

Sweden (1)
Singapore (2)
USA (8)
Canada (9)
Taiwan (11)
South Korea (12)
Hong Kong (13)
Japan (18)



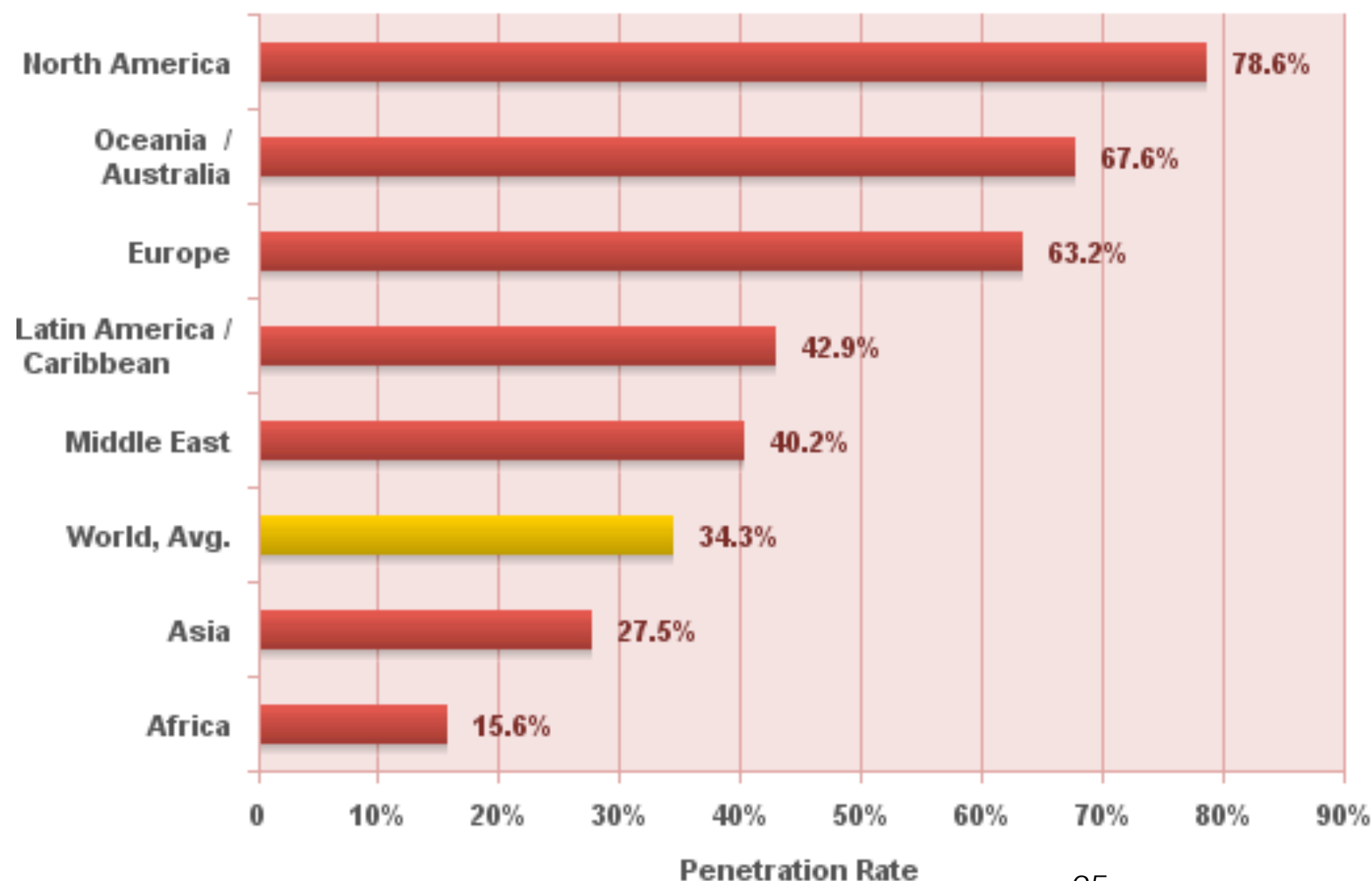
...

China (51)
Russia (56)
Brazil (65)
India (69)



Large penetration in Europe

**World Internet Penetration Rates
by Geographic Regions - 2012 Q2**





Top Sites in France

The top 500 sites in France. 



- 1 **Google France**
google.fr
Version française du moteur de recherche. Propose des outils et des services pour les internautes... [More](#)
- 2 **Google**
google.com
Enables users to search the world's information, including webpages, images, and videos. Offers... [More](#)
- 3 **Facebook**
facebook.com
A social utility that connects people, to keep up with friends, upload photos, share links and ... [More](#)
- 4 **YouTube**
youtube.com
YouTube is a way to get your videos to the people who matter to you. Upload, tag and share your... [More](#)
- 5 **Wikipedia**
wikipedia.org
A free encyclopedia built collaboratively using wiki software. (Creative Commons Attribution-Sh... [More](#)
- 6 **leboncoin.fr**
leboncoin.fr
site de petites annonces gratuit et sans commission (produits d'occasion, annonces immobilières... [More](#)
- 7 **Yahoo!**
yahoo.com
A major internet portal and service provider offering search results, customizable content, cha... [More](#)
- 8 **Amazon.fr**
amazon.fr
Livres en français et en anglais, neufs ou d'occasion, produits culturels.
- 9 **Orange**
orange.fr
Présente les offres de cet opérateur et leurs tarifs, permet de souscrire à certaines d'entre e... [More](#)
- 10 **Freebox, la meilleure offre ADSL : Internet, Téléphone, Télévision**
free.fr
Free: ADSL Jusqu'à 28 Méga, 10Go d'espace disque, WiFi-MiMo, Ligne téléphonique, Appels vers 10...
..



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



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)
- Netherlands: 2 % (AVG: 47)



Europe at the periphery of the information society?

Discrepancy between the

importance of Europe

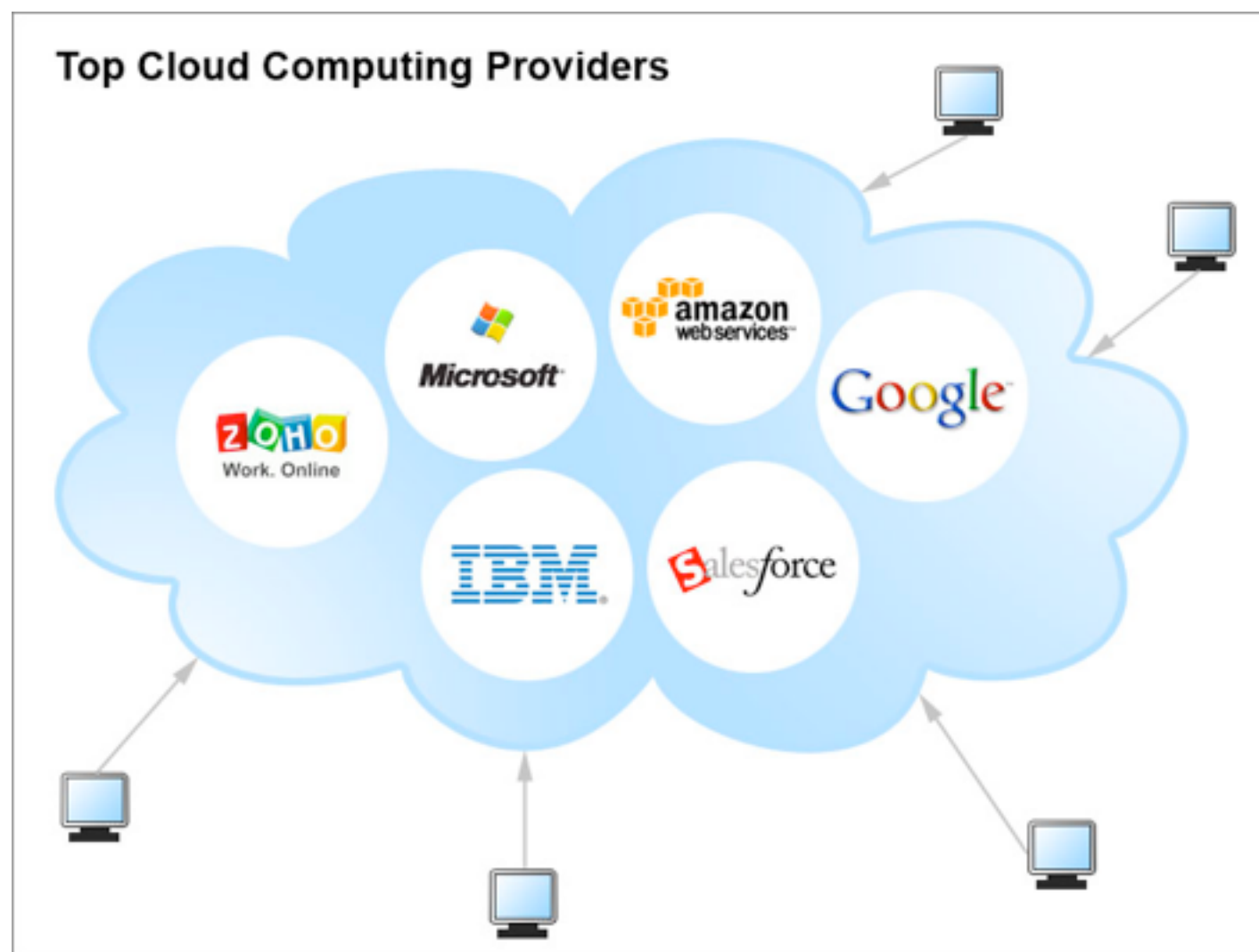
cultural, economical, political, ...

its weak influence in the information society

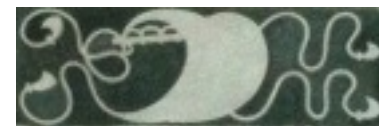
materials, systems, services, ...



supported by non European devices and operating systems



<http://talkcloudcomputing.com/cloud-service-providers-compete-to-capture-the-cloud-market/>



carrying American services



WIKIPEDIA
The Free Encyclopedia





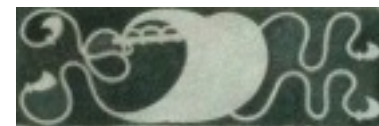
The digital precautionary principle

The European dream: allow systems with
a predefined service
using the minimal amount of data required for that service

The exact opposite of intermediation platforms
open to apps on private data with users consent

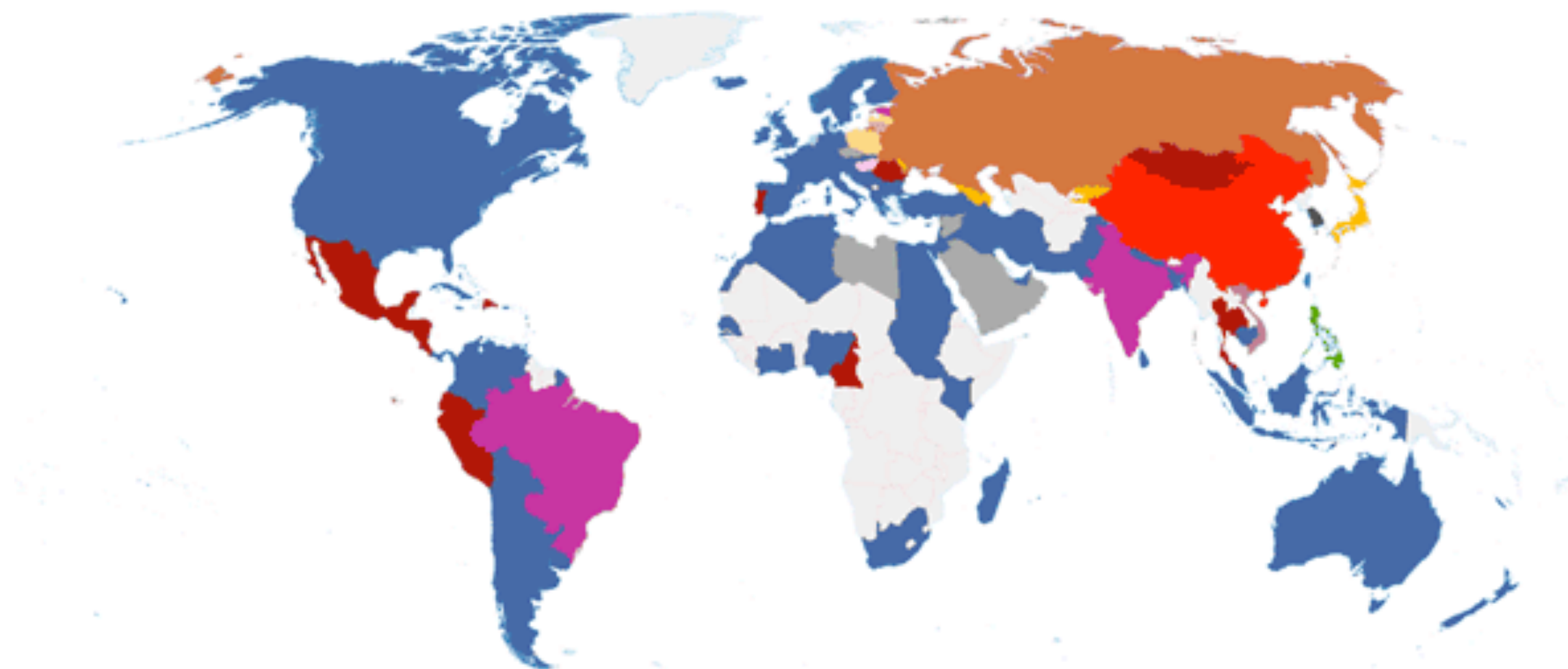
How can personal data be protected ?

How shall systems be restrained ?

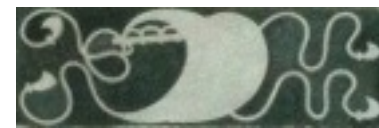


WORLD MAP OF SOCIAL NETWORKS

June 2009



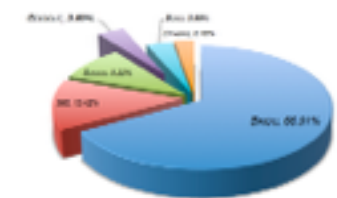
Facebook	V Kontakte	Odnoklassniki	Lidè	Hyves	Zing	Hi5
Orkut	Nasza-Klasa	QQ Zone	Iwiv	Maktoob	One	Mixi
Friendster	Wretch	Cyworld				



Diversity of search engines

Market share in 2012

- 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%

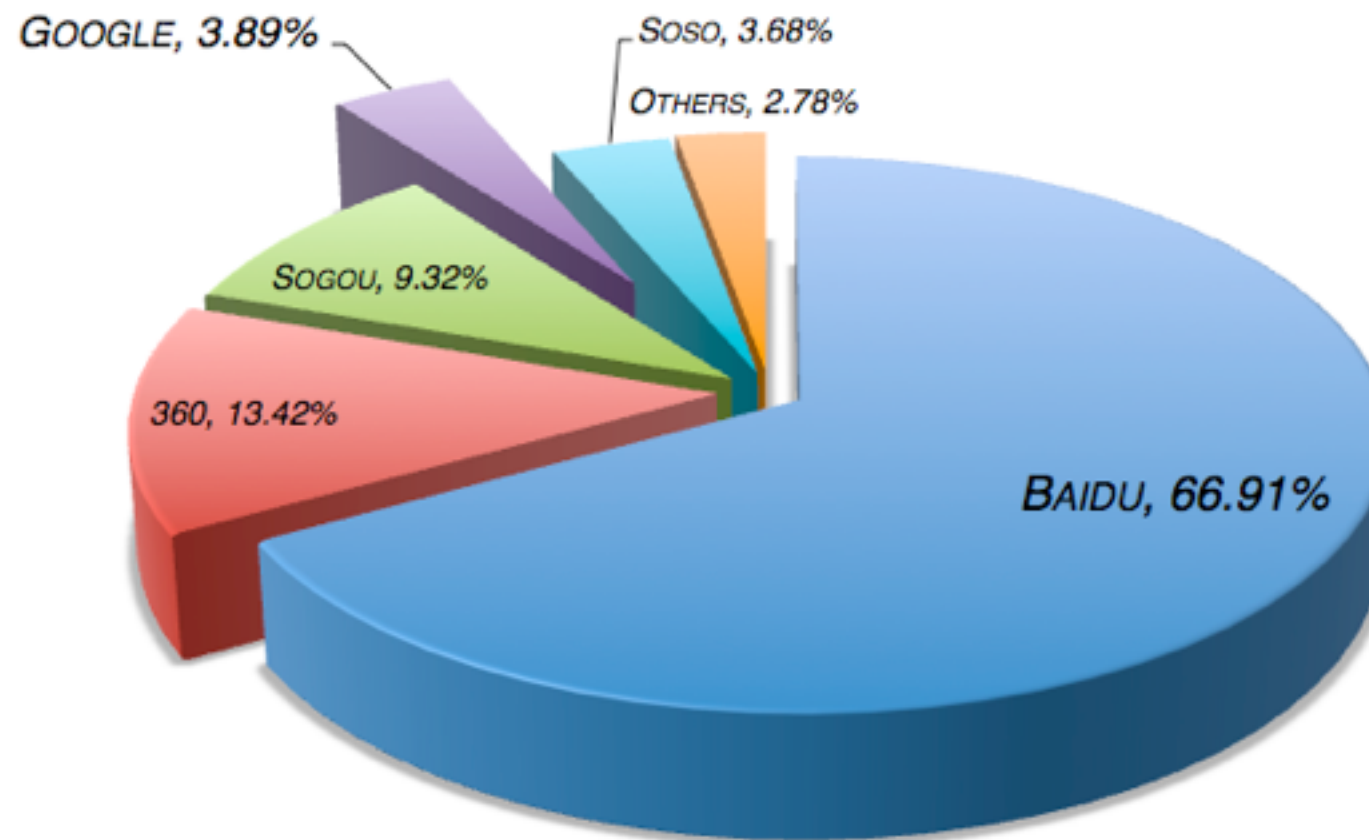


In France,

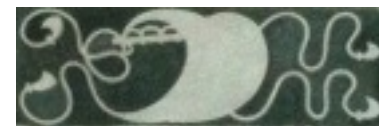
- Google has a de facto monopoly
- Google knows more about France than INSEE



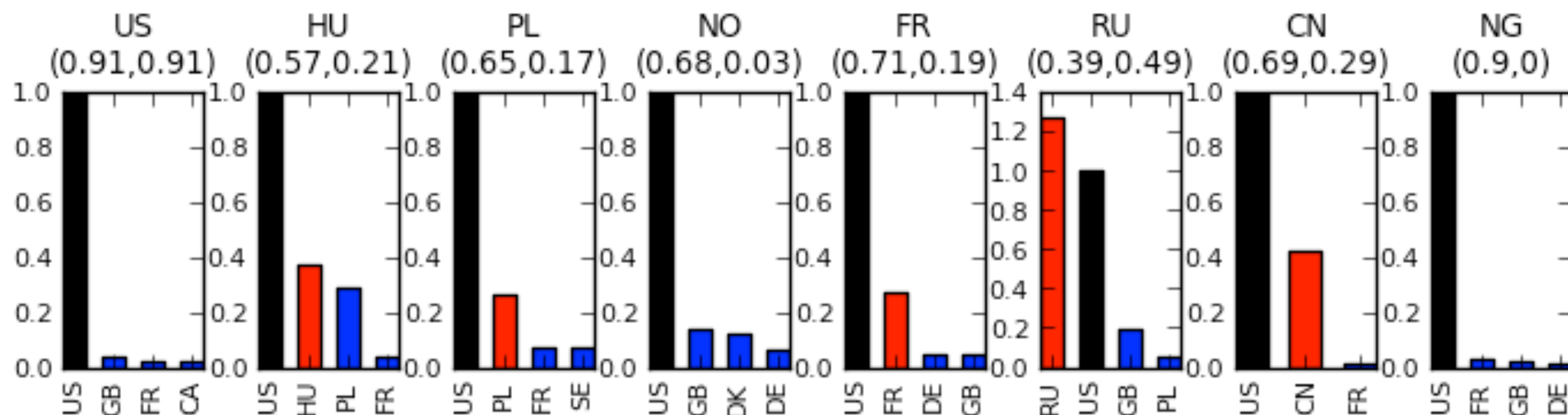
Dynamics of Chinese market



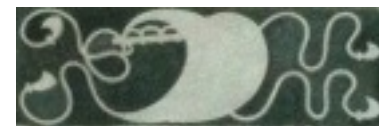
A lesson for Europe?



Global tracking



Proportion of trackers in different countries



Conclusion

Intermediation platforms for social data flows
between users and/or services
capture the secondary data

New form of supremacy that challenge
most of the industries and institutions
as well as the political organization,
diplomacy and defense

