Go with the dataflow

Presentation at the meeting of the OECD
Working Party on Indicators for the Information Society

Pim den Hertog (Dialogic)
Robbin te Velde (Dialogic)
Henrik Schulze (Ipoque)

Paris, April 30th 2008

dialogic
innovatie • interactie

Universiteit Utrecht

[Faculty of Science]
Information and
Computing Sciences

Outline

1. Introduction, lessons learned & future implications (Pim den Hertog)

2. Methodological issues (Robbin te Velde)

3. Measuring Internet Usage – The Whole Truth (Henrik Schulze)

Documentation:

- Handouts
- IaD main report
- IaD Annexes
- IaD case studies

dialogic
innovatie • interactie

Introduction (1): policy background

- Policy need to better understand phenomena associated with the Emerging Digital Economy (EDE)
- These phenomena are only partly captured in “established” statistics
- Notion of “digital footprints”

dialogic
innovatie • interactie

Introduction (2): research questions

May 2007 the Internet as a Datasource (IaD) project took off with 2 key questions:

1. **Identify new data and indicators derived directly from the Internet and describe new phenomena associated with EDE**

2. **To explore and assess the usefulness of the various IaD methods for deriving new, extra and substitute data for the EDE**

dialogic
innovatie • interactie

Introduction (3): activities performed

- Conceptualization IaD
- Typology IaD methods: user-, network- and site-centric measurements
- Usability of spiders/web crawlers
- Analysis of international sources using IaD (see Annex 1)
- 8 Case studies using fixed format, including some experiments
- Contribution to CBS publication on the Digital Economy
- Attempt to organize a network-centric measurement
- Overall analysis & reporting
- *International seminar*

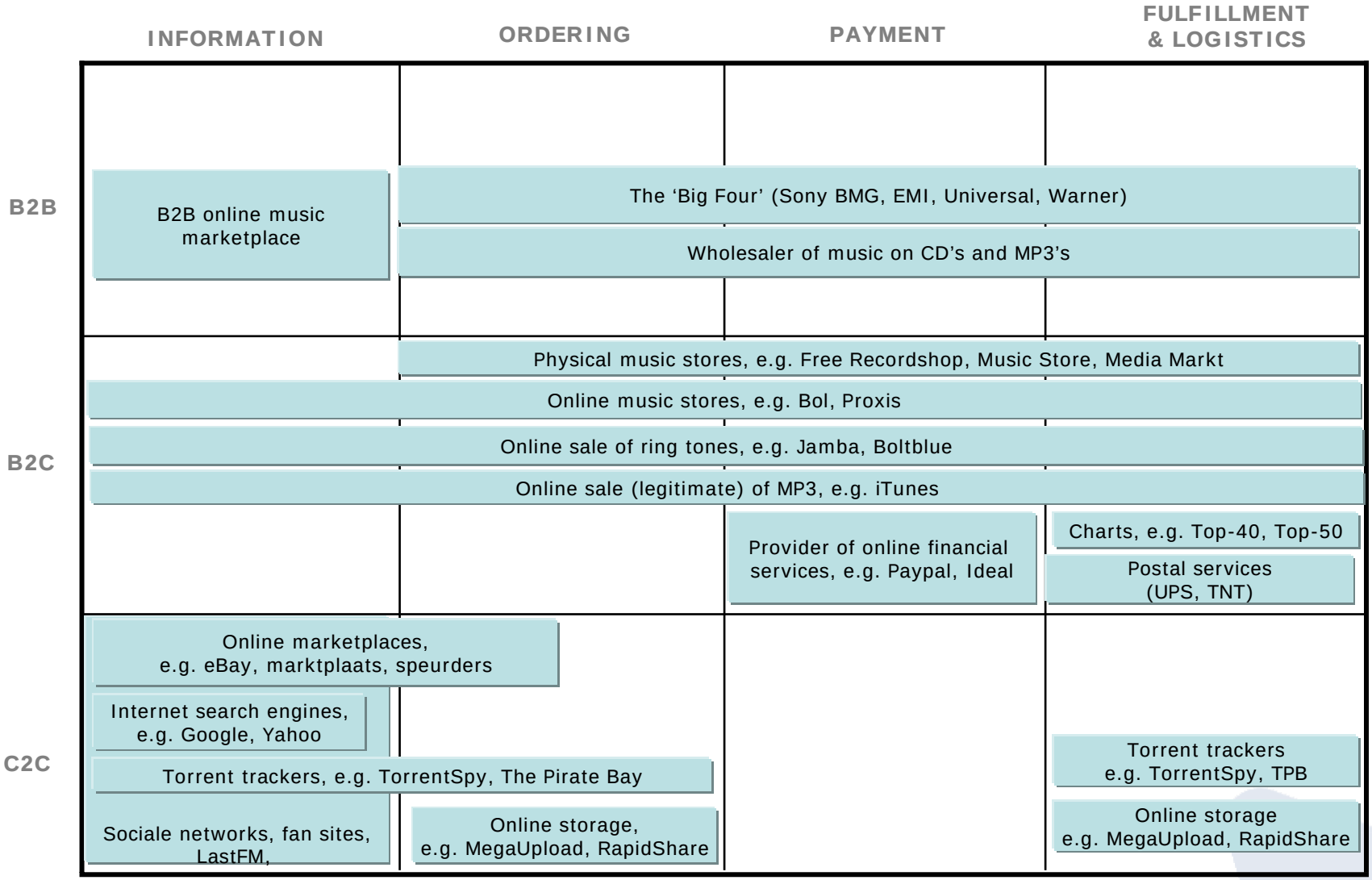
dialogic
innovatie • interactie

Introduction (4): covering “old” & “new” economy

	“Old economy” (established) markets & phenomena	“New economy” (emerging) market & phenomena
Established (analogue, mostly invasive) data collection methods	(1) ICT investments in NACE-measured	(2) New media use by final users through a survey among a panel of households
Internet-based (digital, mostly non-invasive) data collection methods	(3) Price of pigs traded over an electronic market through a site centric measurement	(4) Share of illegal content in P2P traffic as measured through a network centric measurement

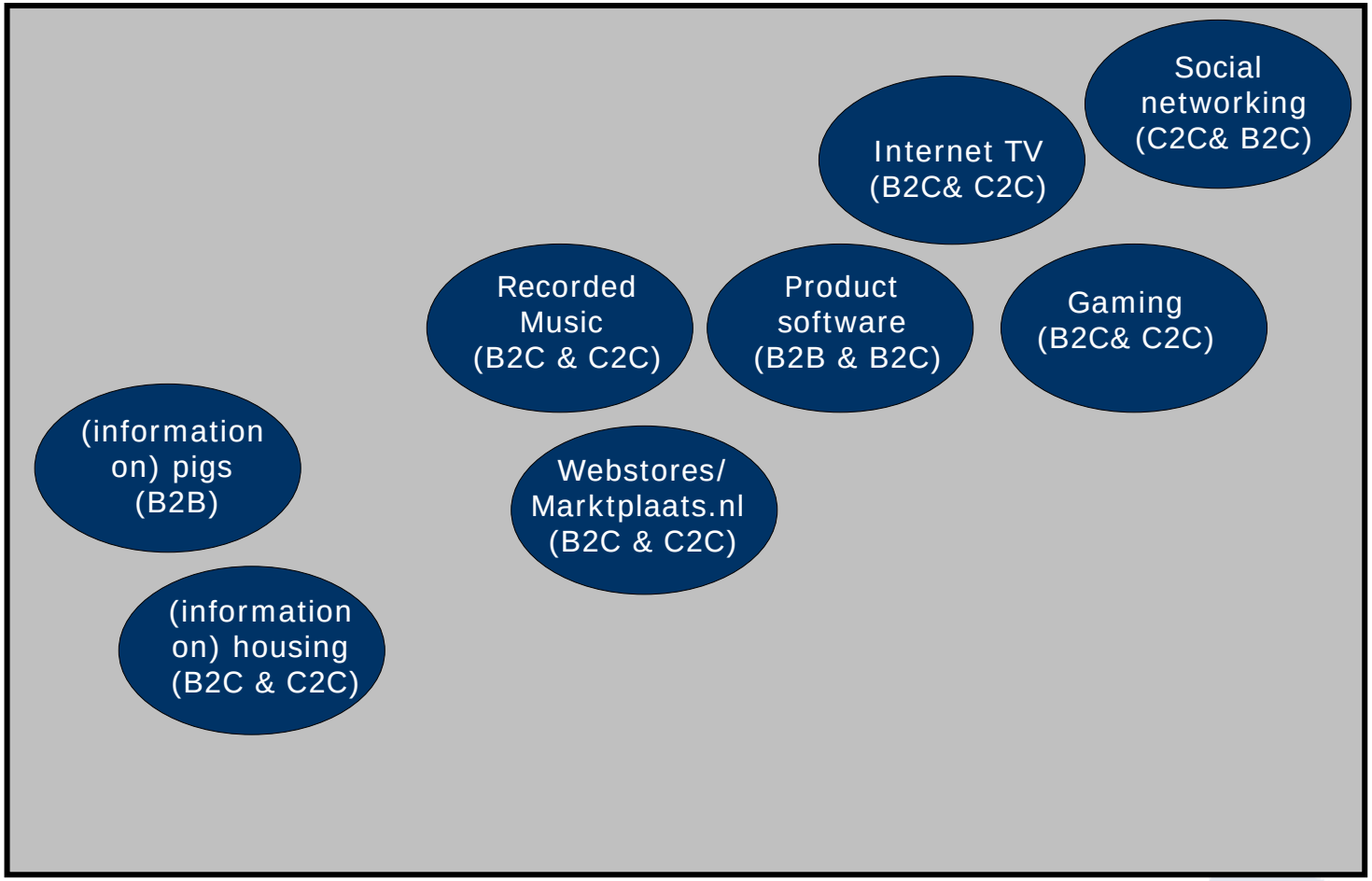
dialogic
innovatie • interactie

Introduction (5): digital concentration points



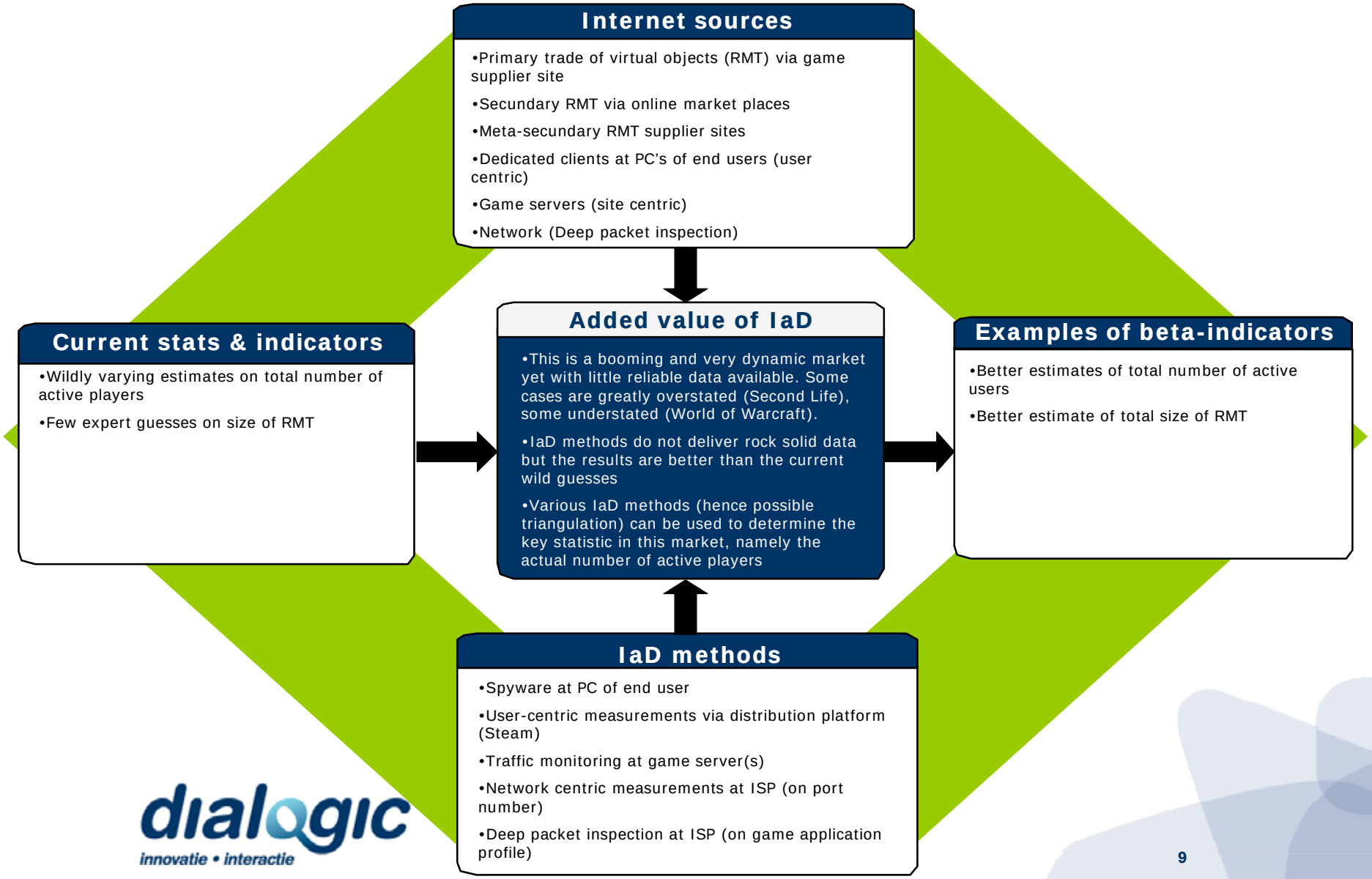
dialogic
innovatie • interactie

Introduction (6): Selection of cases



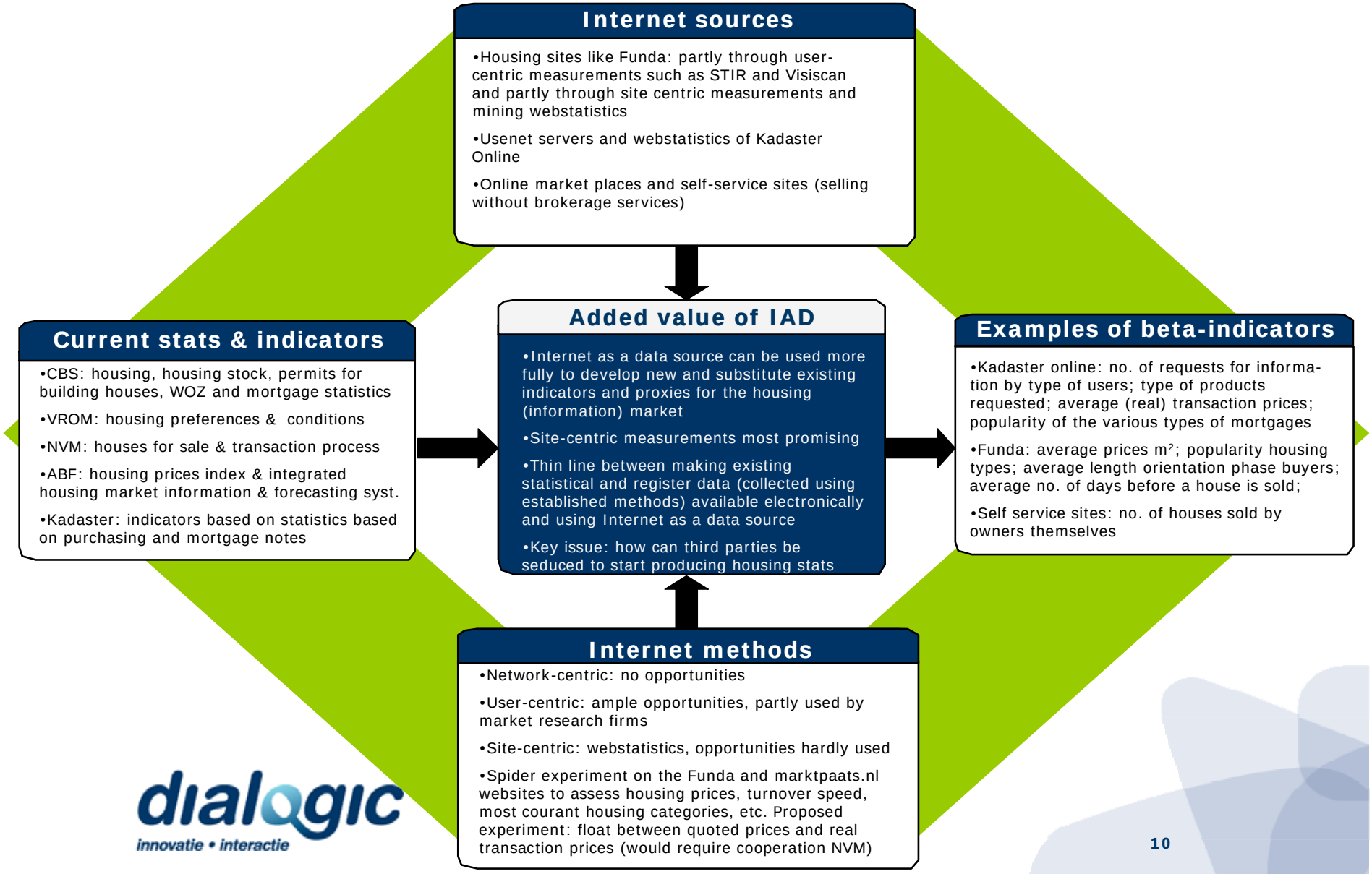
dialogic
innovatie • interactie

Lessons learned (1): format cases (online gaming)



dialogic
innovatie • interactie

Lessons learned (2): format cases (housing market)



dialogic
innovatie • interactie

Lessons learned (3): overall lessons

1. IaD helps in signaling new trends, developments and phenomena

2. Mix of IaD methods applied will vary widely between markets and industries

3. The notion of digital footprints is not limited to digitalized products and services, but applies to a wider set of markets and industries

dialogic
innovatie • interactie

Lessons learned (4): overall lessons

4. Information on goods and services represents an economic value in itself

5. Industry itself has already started to mine digital footprints

6. Markets associated with the emerging digital economy are more fuzzy and diffuse

7. In some markets user generated content (already) starts to mix with traditional economic production (online music, internet TV, housing) ➔ barriers between the social and economic realm are blurring (SNS, C2C marketplaces)

dialogic
innovatie • interactie

Lessons learned (5): overall lessons

8. Technical and practical availability of digital sources for third parties may differ considerably

9. Added value of using IaD may be higher in newly developing markets

10. IaD can shed light on the darker side or grey zone of the emerging digital economy

dialogic
innovatie • interactie

Lessons learned (6) added value IaD

- Relevant data source for policy-makers, researchers & statisticians, market research firms, industrialists and trade organizations
- Provide insight into markets and phenomena in areas where the statistical agencies have no established statistics available
- The potential of IaD methods for substitution of existing indicators and statistics should not be overestimated
- IaD-methods may lead to beta-indicators for the EDE ➔ trade off between “early warning quality” indicators vs. lower statistical quality.

dialogic
innovatie • interactie

Lessons learned (7) added value IaD

Added value of using IaD is highest when:

- value chains are highly digitalized;
- online activities are the subject of research;
- subjects of research are highly dynamic and/or real time information about the subject is required.
- markets are dominated by a few players;
- market players are very transparent;
- markets are highly regulated (➔ e.g. high quality registers)
- administrative tasks are labour intensive (scope for reducing administrative burden).

dialogic
innovatie • interactie

Future implications (1): statistical agencies

- Statistical agencies need to better capture phenomena associated with EDE – if they do not perform IaD methods themselves they should at least guarantee the quality of the statistical data/indicators that are generated by private firms (that are already filling the gaps that are left by public agencies)

- They are well positioned to play a key role in the switch to IaD-methods as they have:
  - scale and expertise for developing and collecting statistical indicators (sunk costs);
  - possibility to validate data and indicators derived from IaD measurements using regular statistics;
  - possibility to guarantee privacy if needed;
  - a judicial status they might want to use to enforce co-operation of data providers;
  - the international network for international benchmarking, exchange of expertise and setting standards and developing international guidelines.

dialogic
innovatie • interactie

Future implications (2): policy-makers

Various options to spur the further experimentation & use of IaD:

- new " beta statistics" publication on the emerging digital economy
- create a network of researchers, market research agencies, policy makers and statisticians
- establish a clearinghouse for Internet statistics.
- support statistical agencies to pro-actively experiment & use IaD-methods
- start exploratory talks with organisations and companies that can contribute to this R&D network
- governments themselves can anticipate on the use of digital sources for statistical purposes when developing or implementing their own registers and ICT projects

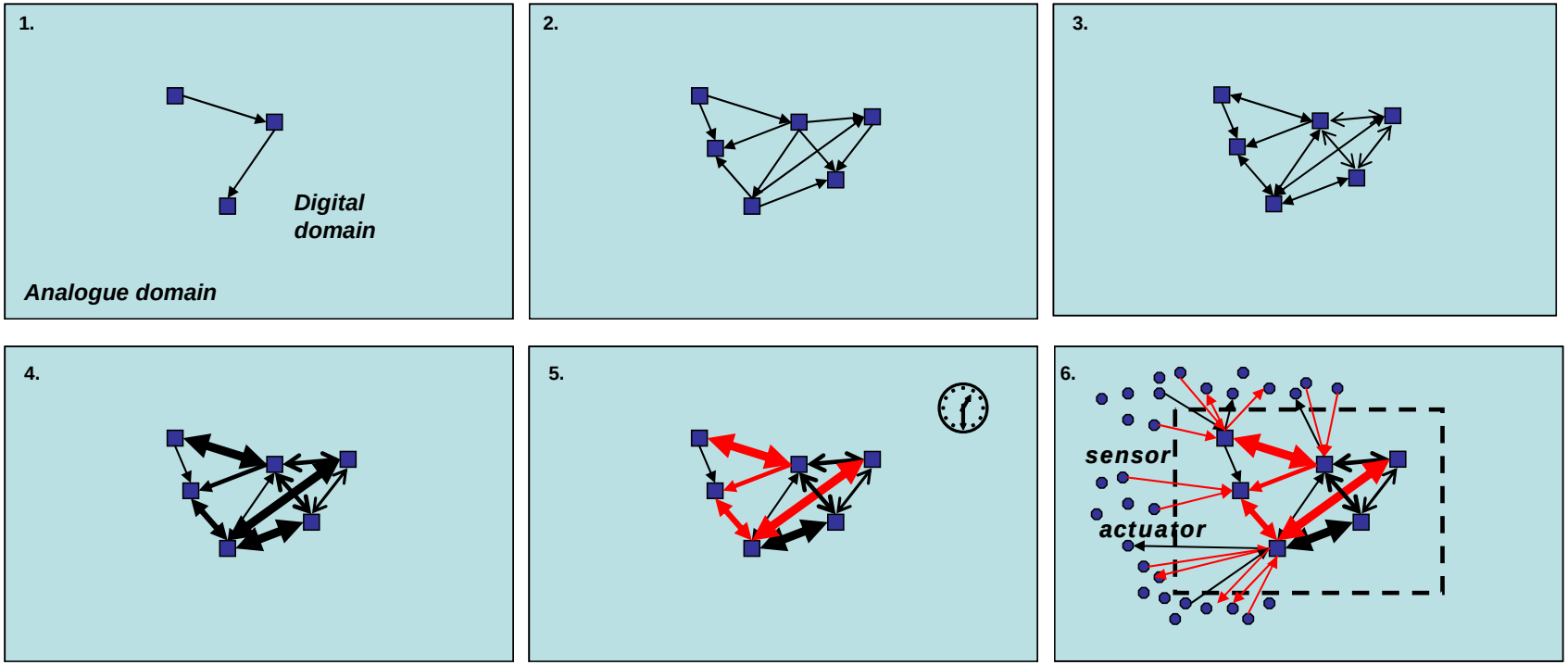
... to be taken up adopting an international perspective

dialogic
innovatie • interactie

but do not get overloaded…

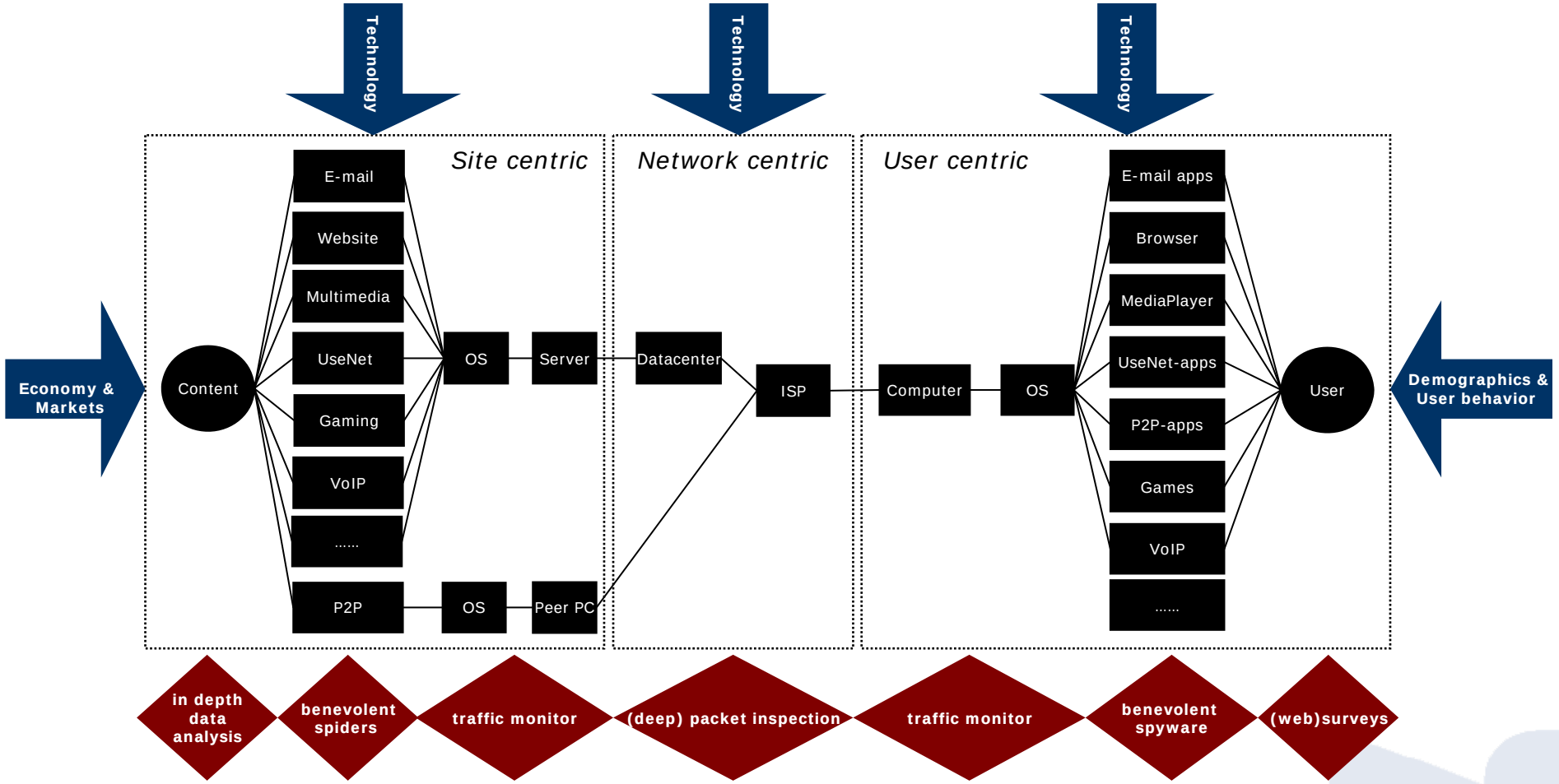
dialogic
innovatie • interactie

With the advance of digitalization the scope of IaD-methods continuously expands



dialogic
innovatie • interactie

A typology of IaD-methods



dialogic
innovatie • interactie

Usability & disadvantages of IaD methods

Method	Disadvantages	When to use?
Spiders	-Custom-made for every application -Feasible if the set of applications is limited and stable -Owner of an application can hinder being spidered	-When insight in content is needed
DPI@ISP	-The data does not allow generalisation -Difficult to obtain insight in content -Privacy of users can be threatened -Very hard to find ISPs willing to cooperate	-When a full scope of internet traffic is needed -When strong 'sympathy effects' are present -When very small effects and / or trends real time have to be identified
Traffic Monitor at OS	-Measurements are relatively shallow -A (costly) panel is needed -Illegal and shameful behaviour not measured correctly -The limited panel size makes is hard to find small effects	-When insight in user behaviour on all applications is needed.
Spyware	-Spyware needs to be custom-made -Measurements are relatively shallow -A (costly) panel is needed -Illegal and shameful behaviour not measured correctly -The limited panel size makes is hard to find small effects	-When insight in user behaviour on a certain application is needed.

Statistical usability (ii)

IaD-method	Robustness (internal validity)	Representativity (external validity)	Transparency	Longitudinal use
Spyware and Traffic Monitor	**High** but underestimates illegal behaviour	**High**. depends on panel.	**Very High.** Like conventional surveys.	**High**. Sometimes changes in software
DPI at ISP	**High**. But advanced users can hinder DPI	**Low**. user characteristics are usually unknown.	**Very low.** Non-disclosure agreements	**Medium**. Small changes in infrastructure have major implications.
Benevolent spiders	**Low-medium**. Differences between 'websites' hinder measurement. structural bias Underestimates illegal content	**Varies**. High in Concentrated markets. Low in fragmented market	**Medium**. OS Spiders	**Low**. Continuous changes 'websites'

dialogic
innovatie • interactie

Suggestions for additional experiments

- User centric
  - Set up panel for use of C2C marketplaces
  - Monitor use of internet-TV (analogue to traditional method)
- Network centric
  - Measure use of Citrix
  - Measure use of DRM on the internet
  - Measure use of instant messaging
- Site centic
  - Real time comparison of prices (Funda) and transaction prices (NVM) at real estate sites
  - Measure occurrence long tail at web shops
  - Spider the market for virtual goods
  - Spider social networking sites on demographic properties

dialogic
innovatie • interactie

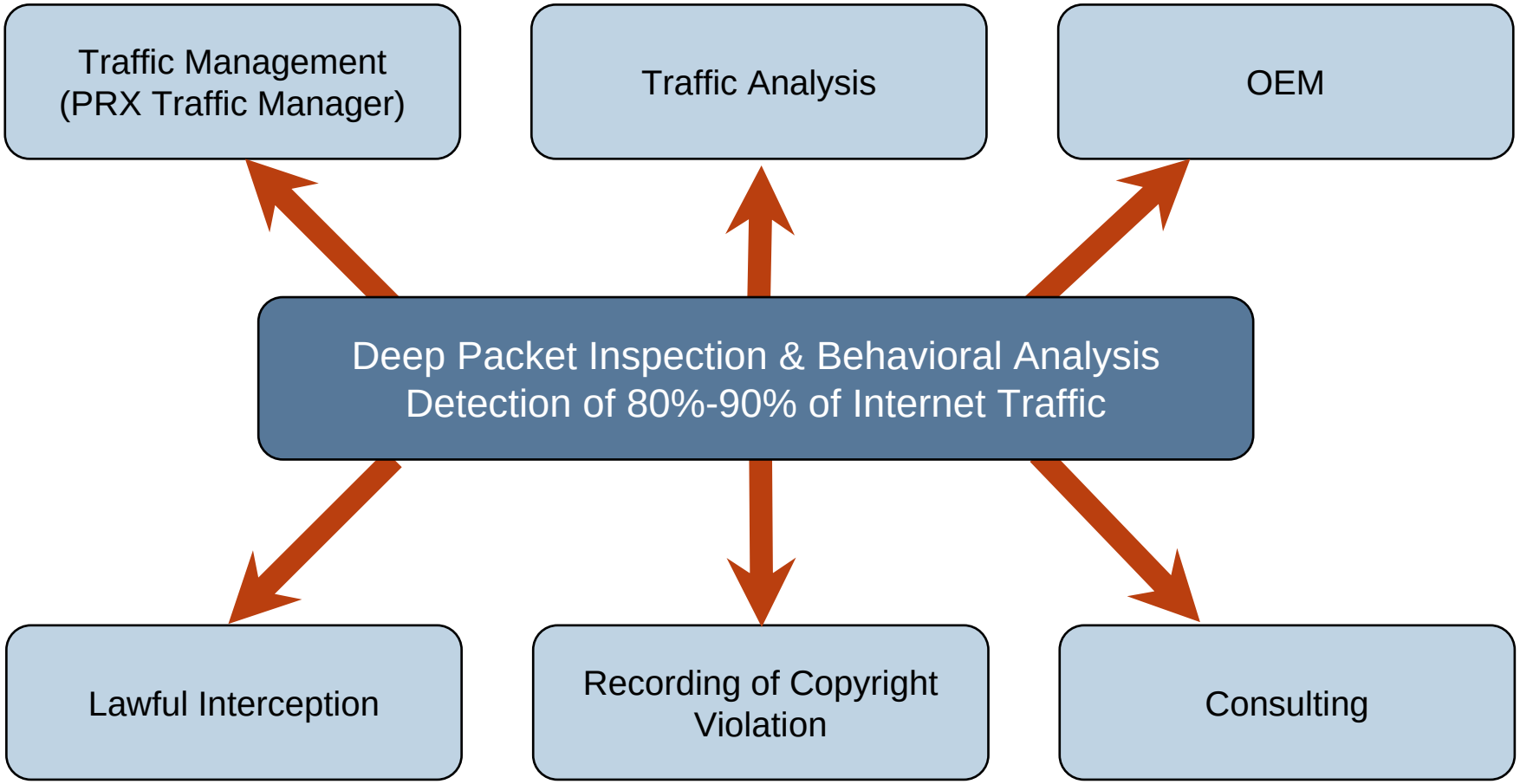
Final considerations to the use of IaD-methods

- Faustian bargain: trade-off between efficiency, objectivity, timeliness and cost-effectiveness on the one hand and validity and privacy on the other hand
- Sometimes there are simply no alternatives to the use of IaD methods
- How we see these beta-indicators:
  - There is a new category of beta-statistics that are especially suitable to pick up early trends in the EDE.
  - We should assess the practical and statistical quality of these statistics :
    - If these statistics do not meet the quality standards of beta-statistics they should be dropped.
    - The quality of the remaining beta-statistics should be improved (e,g., definitions, standardization, more sophisticated indicators).
    - Eventually, some beta-indicators could be promoted to the league of alpha-statistics.

dialogic
innovatie • interactie

Agenda

- About ipoque
- How to measure?
  - DPI: state of the art for Internet traffic classification
  - Passive network measurement: How to do it?
- Some results
  - The ipoque Internet Study
- The whole truth
  - Technical, administrative and procedural issues

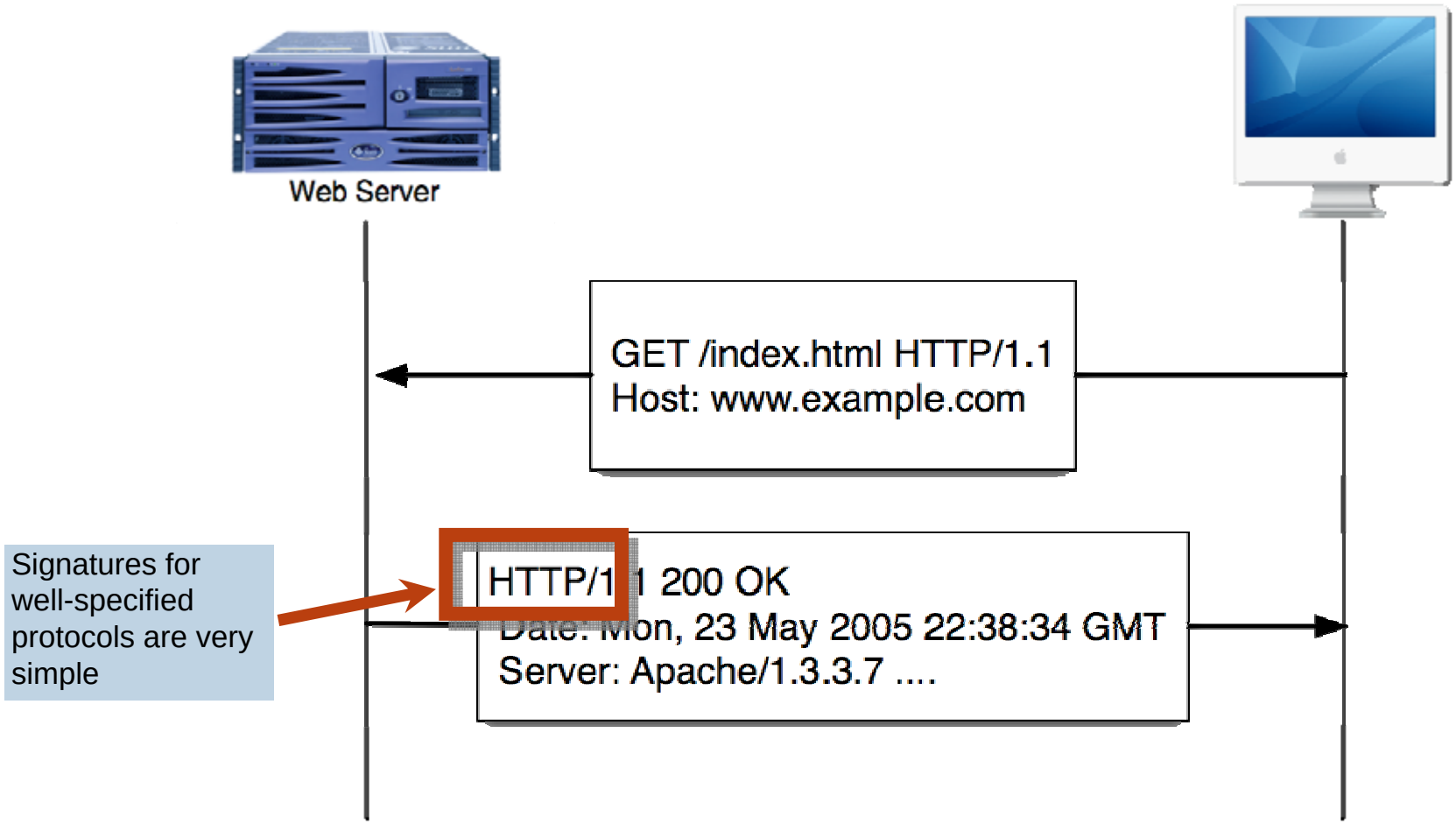

Traffic Management
(PRX Traffic Manager)
Traffic Analysis
OEM
Deep Packet Inspection & Behavioral Analysis
Detection of 80%-90% of Internet Traffic
Lawful Interception
Recording of Copyright
Violation
Consulting

http://www.ipoque.com:80/website.html

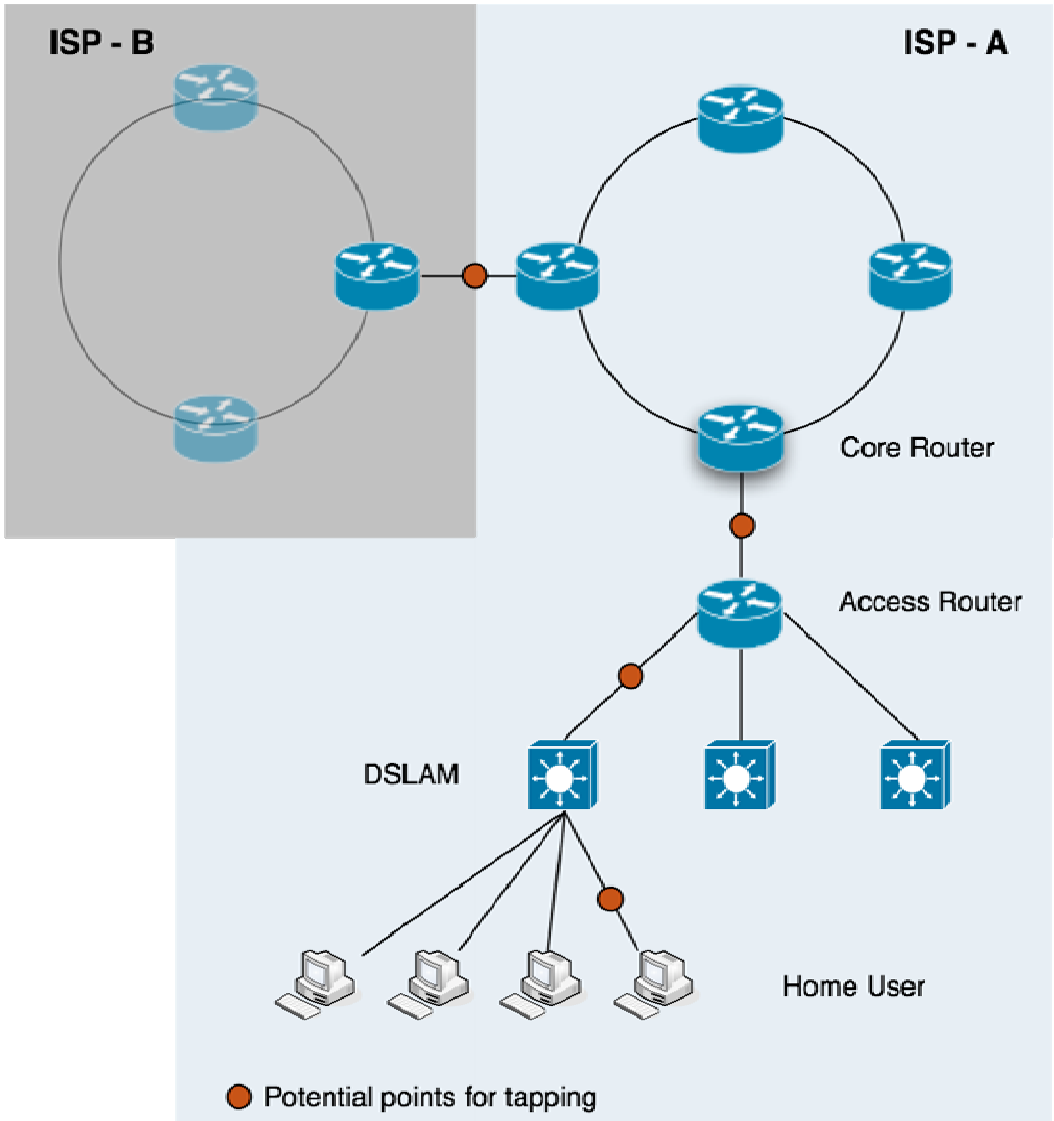
The port is part of a network address and normally hard coded into the application

- Port based traffic classification:
  - does not work any longer
  - modern applications, like Skype or Instant Messengers are not bound to dedicated ports
- Deep Packet Inspection:
  - classification of network traffic based on unique application signatures

Example: http (World Wide Web)

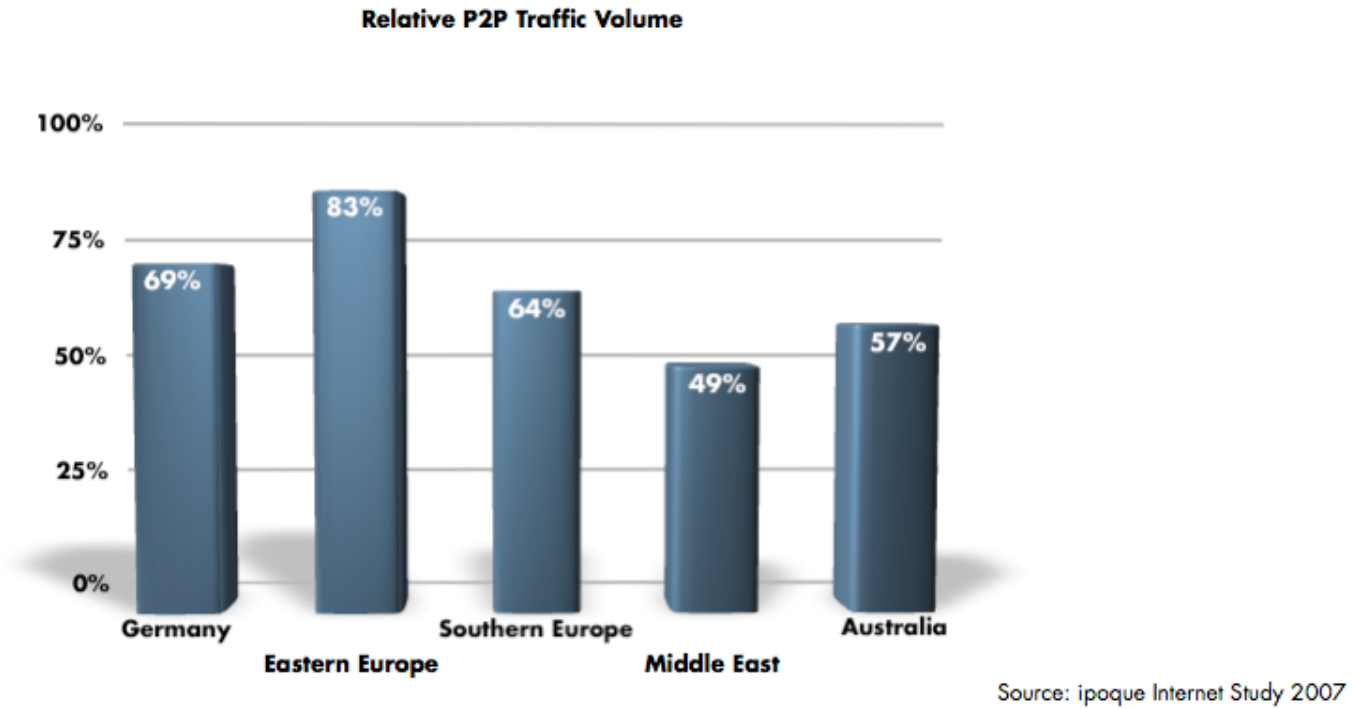


- Tapping a network connection
  - guaranteed: no impact on original traffic
  - working on 1:1 copy of network traffic
  - taps are standard equipment

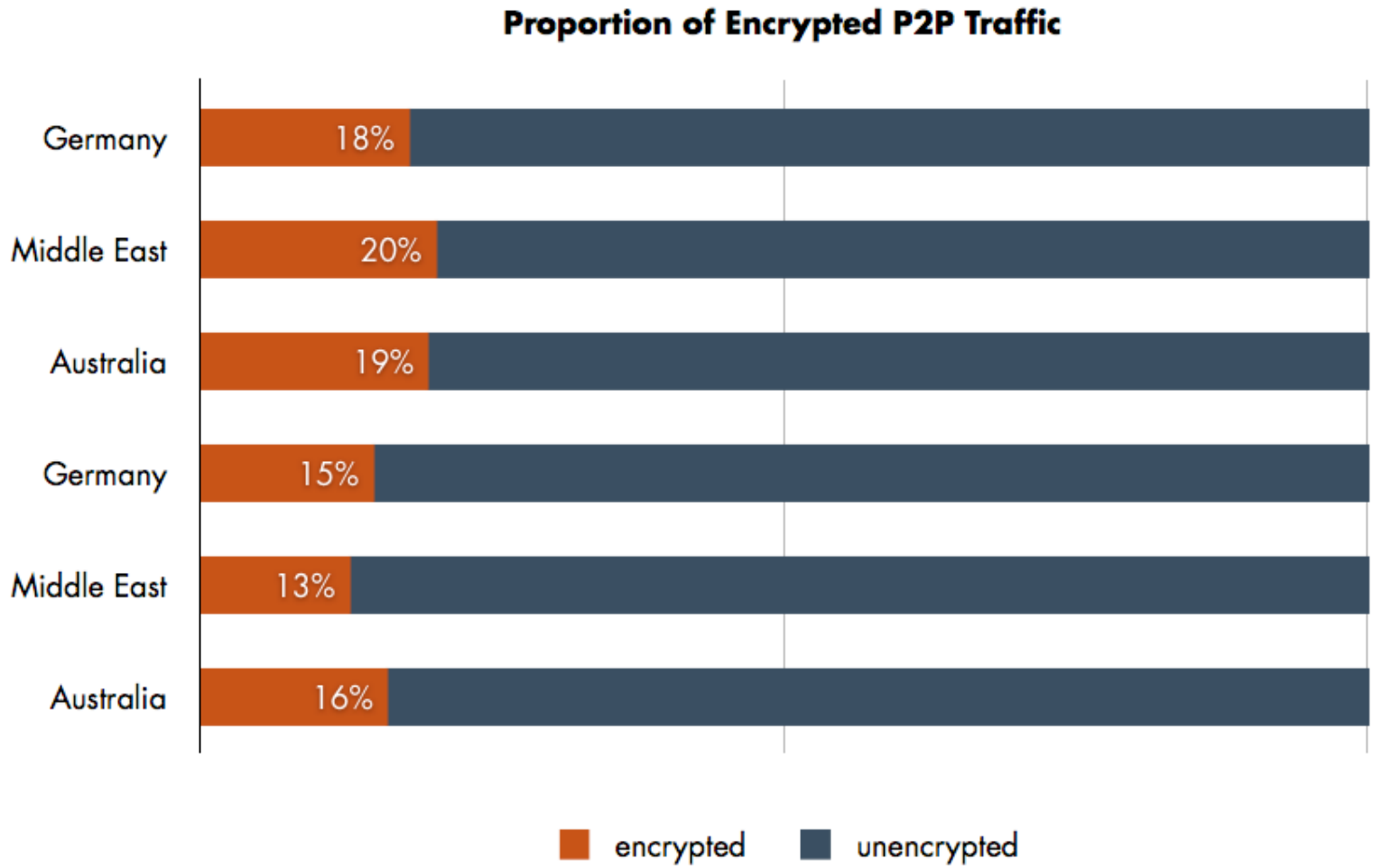

ISP - B
ISP - A
Core Router
Access Router
DSLAM
Home User
Potential points for tapping

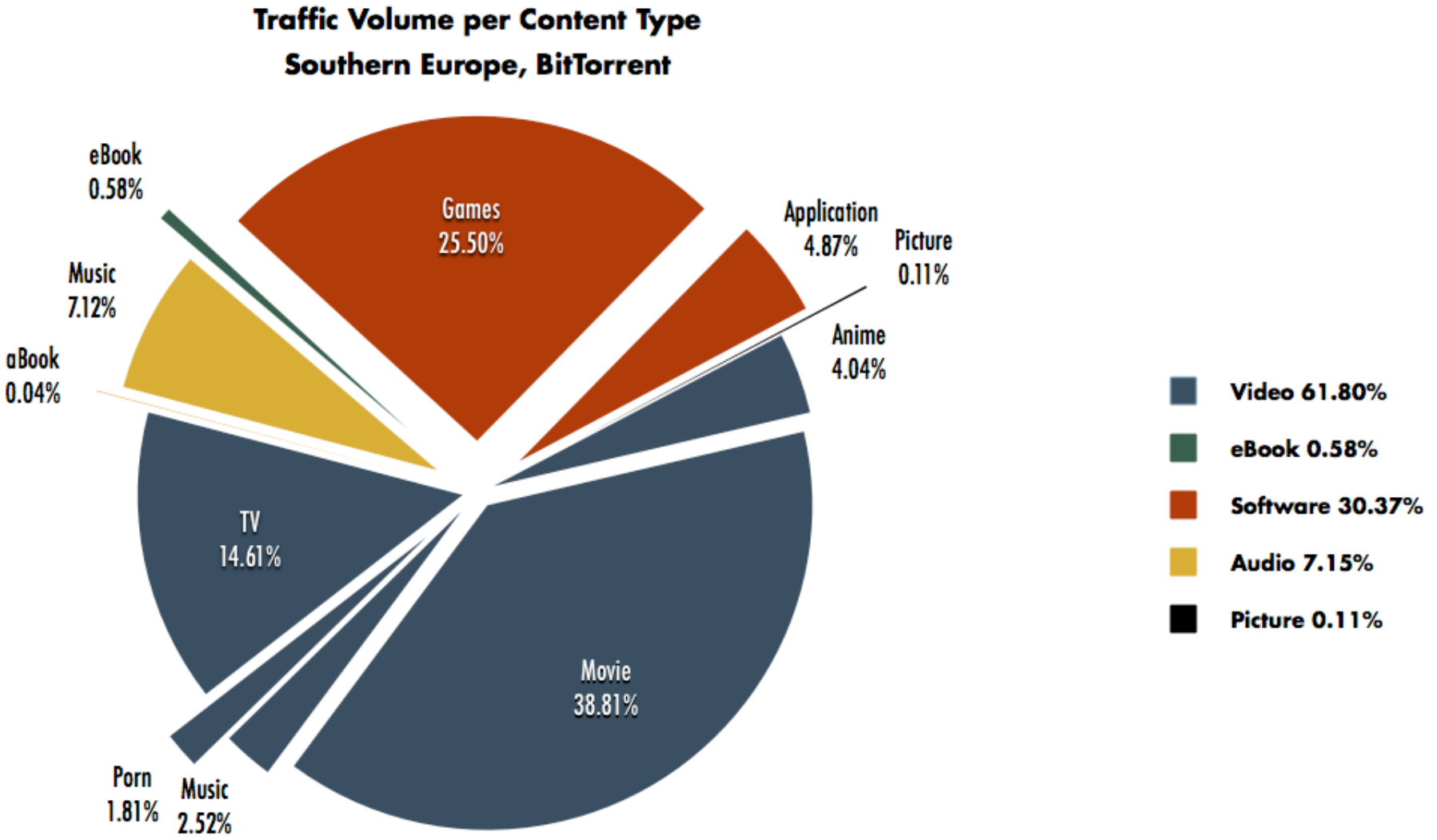
- Snapshot of the current state of the Internet
  - 18 monitoring sites at ISP (13) and Universities(5)
  - 5 regions
    - Southern Europe, Australia, Germany, Eastern Europe, Middle East
  - 3 Petabytes analyzed traffic
  - representing more than 1m people
  - data taken from the PRX Traffic Manager, installed at customers
  - not representative but a good estimation of
  - “What happens in the Internet”
  - Not just P2P, also VoIP, Skype, IM, Video Streaming, DDL

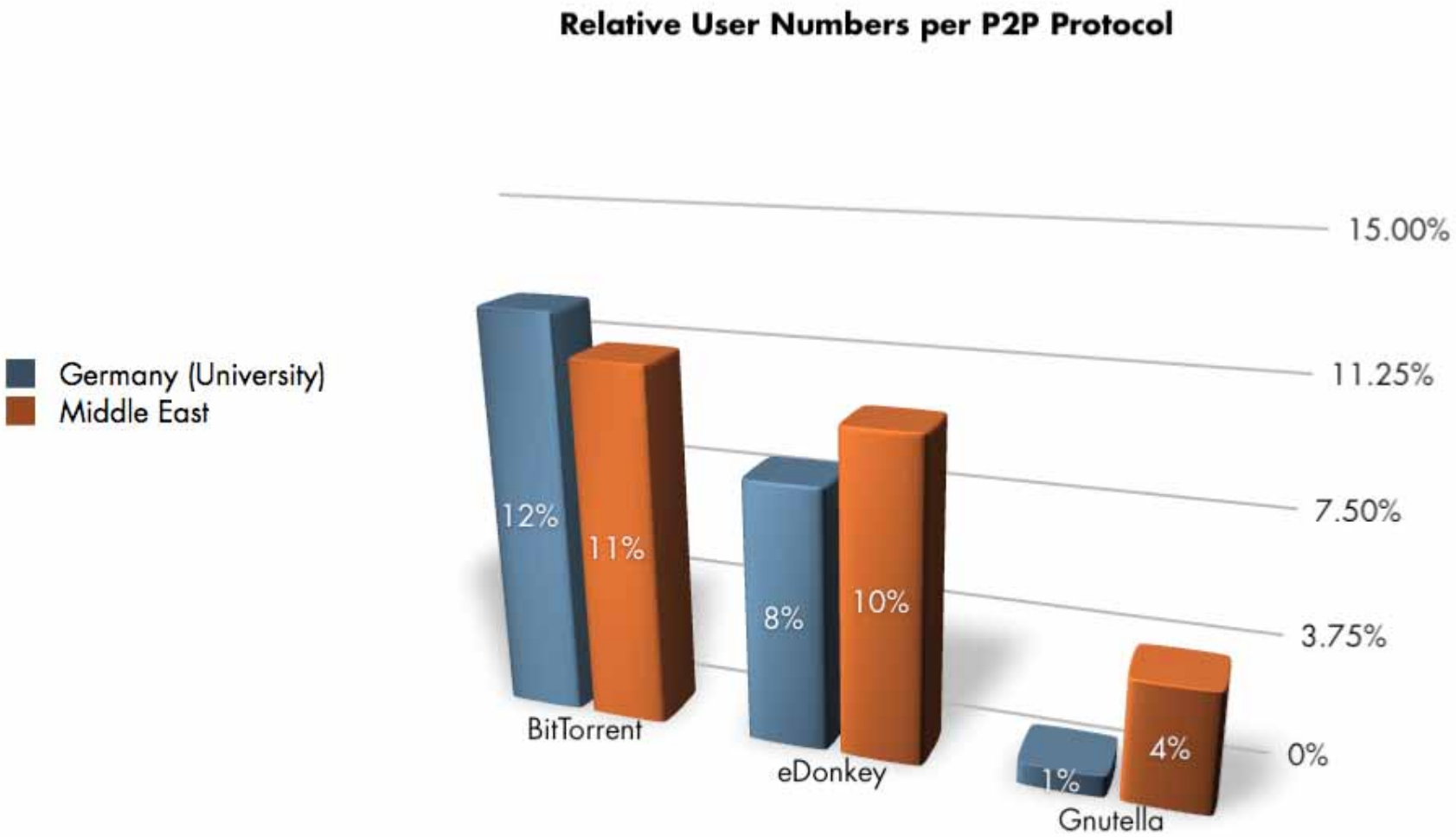
More than 50% of the Internet traffic - worldwide



Source: ipoque Internet Study 2007


Proportion of Encrypted P2P Traffic
Germany
18%
Middle East
20%
Australia
19%
Germany
15%
Middle East
13%
Australia
16%
encrypted
unencrypted


Traffic Volume per Content Type
Southern Europe, BitTorrent
eBook 0.58%
Games 25.50%
Application 4.87%
Picture 0.11%
Music 7.12%
Anime 4.04%
aBook 0.04%
TV 14.61%
Movie 38.81%
Porn 1.81%
Music 2.52%
Video 61.80%
eBook 0.58%
Software 30.37%
Audio 7.15%
Picture 0.11%


Relative User Numbers per P2P Protocol
Germany (University)
Middle East
15.00%
11.25%
7.50%
3.75%
0%
12%
11%
8%
10%
1%
4%
BitTorrent
eDonkey
Gnutella

Video

1.	Movie	Next 2007
2.	Movie	The Simpsons Movie(Spanish)
3.	Movie	Shooter
4.	Movie	Evan Almighty
5.	Movie	Premonition

Music

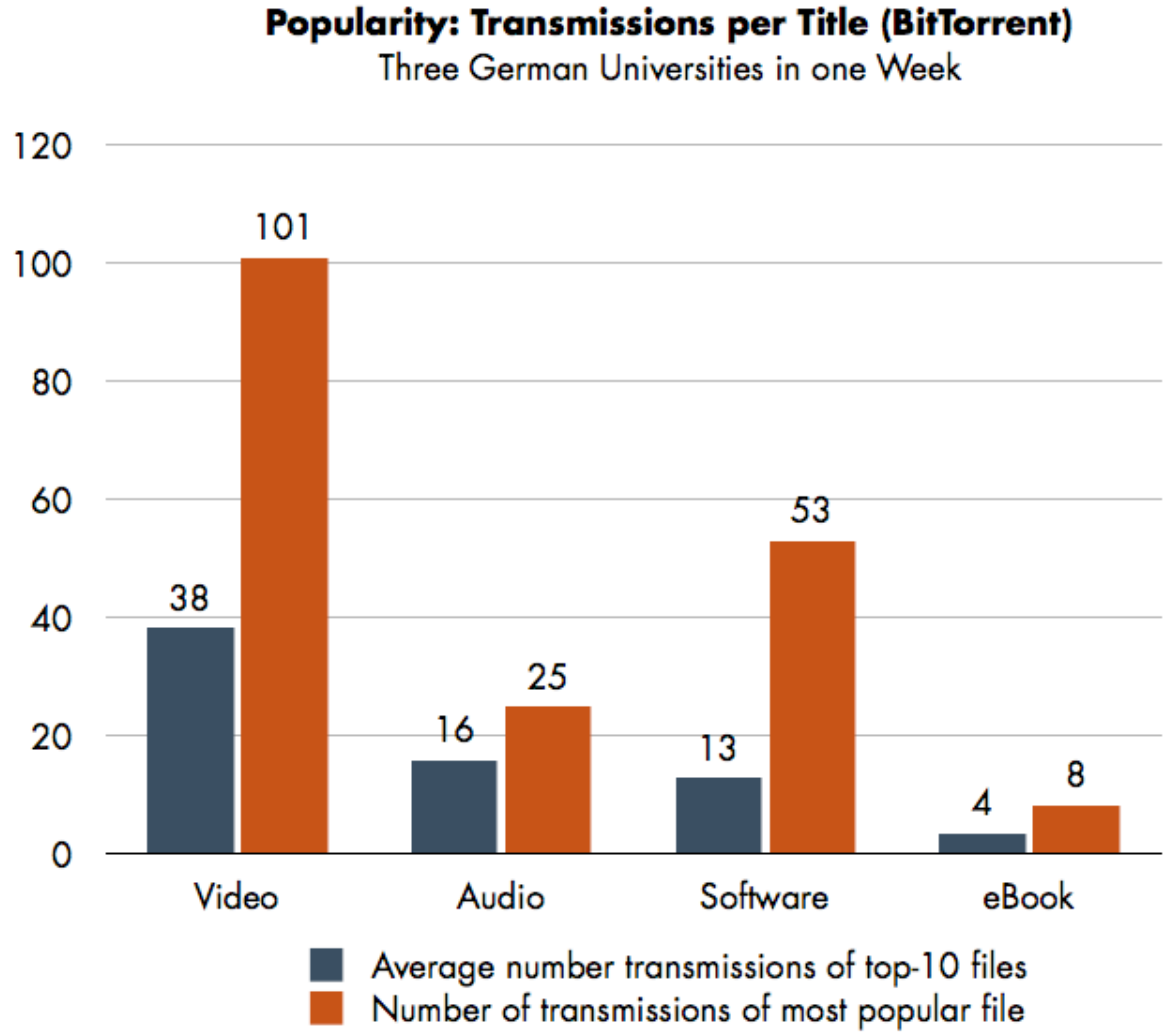
1.	Music	Bob Dylan-Blues-2006-MTD
2.	Music	Da Weasel 2007 Amor Escarnio e ....
3.	Music	Celine Dion 2007 D'elles
4.	Music	Bob Dylan - Live at the Gaslight 1962 [2005]
5.	Music	Maroon 5 -It Won`t be soon bevor long

Software

1.	Application	K-Lite Mega Codec Pack 3.3.5
2.	Games	Football Manager 2007
3.	Application	Nero 7 Ultra Edition
4.	Application	Adobe Photoshop CS3
5.	Games	SilkRoad v1.110 Europe Legend 1

eBooks

1.	eBook	Muscle & Fitness 101 Workouts
3.	eBook	Muay Thai - The Art of Fighting
4.	eBook	All Social Interactions Books
5.	eBook	Get the Dream Job- Cover letter Secrets ...
6.	eBook	tomtom map 6 75 ES and PT

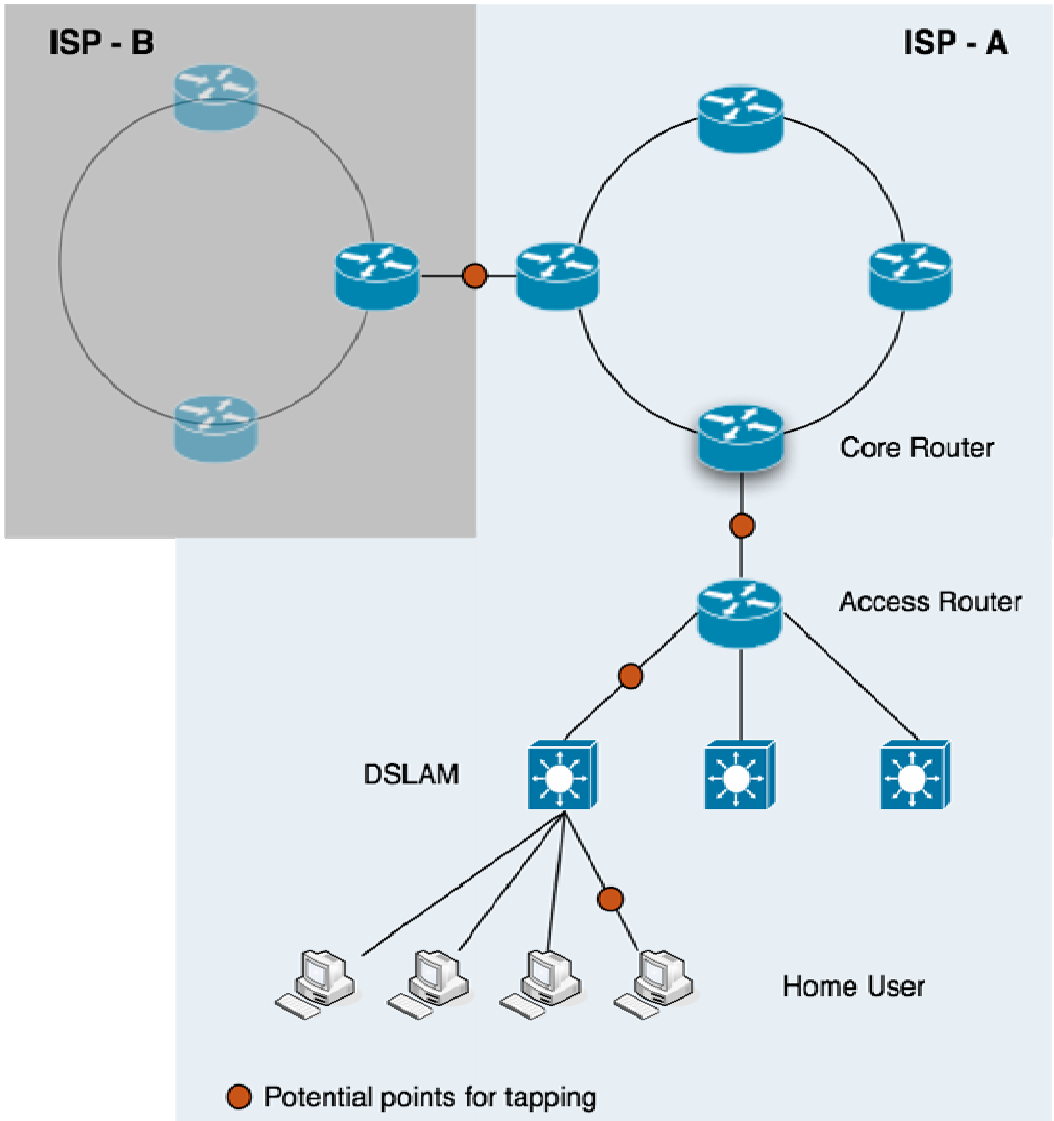

Popularity: Transmissions per Title (BitTorrent)
Three German Universities in one Week
120
100
80
60
40
20
0
101
38
25
16
53
13
8
4
Video
Audio
Software
eBook
Average number transmissions of top-10 files
Number of transmissions of most popular file

- Asynchronous Traffic
  - request and response do not necessarily have the same route through the Internet
- Multiple counting
  - measurements on multiple points might count some traffic twice
- Visibility of all traffic
  - at a peering point I can't see local traffic
- Privacy
  - passive network measurements look at private user data
  - either no storage of user data, or on-the-fly anonymization
- Administrative issues
  - no ISP likes external equipment in its network
    - they don't like the results (e.g. amount of copyright infringements)

    → potential requests for regulation

All these problems can be solved (with some effort)

The Ideal Measurement Infrastructure?



Thank you!

Pim den Hertog
denhertog@dialogic.nl

Robbin te Velde
tevelde@dialogic.nl

http://www.dialogic.nl

Hendrik Schulze
Hendrik.Schulze@ipoque.com

http://www.ipoque.com