Smart Cyber Infrastructure for Big Data Processing PIRE and RDA

Cees de Laat

What Happens in an **Internet Minute?**

639,800 GB of global IP data transferred

20 New victims of identity theft

47,000 App downloads

61,141 Hours of music

204 million Emails sent

$83,000 In sales

20 million Photo views

3,000 Photo uploads

320+ New Twitter accounts

100,000 New tweets

135 Botnet infections

6 New Wikipedia articles published

1,300 New mobile users

100+ New Linkedin accounts

277,000 Logins

6 million Facebook views

2+ million Search queries

30 Hours of video uploaded

1.3 million Video views

And **Future Growth** is **Staggering**

Today, the number of **networked devices** = the global population

By **2015**, the number of **networked devices** = **2x** the global population

In **2015**, it would take you **5 years** — **IP** — to view all video crossing IP networks each **second**

(intel)
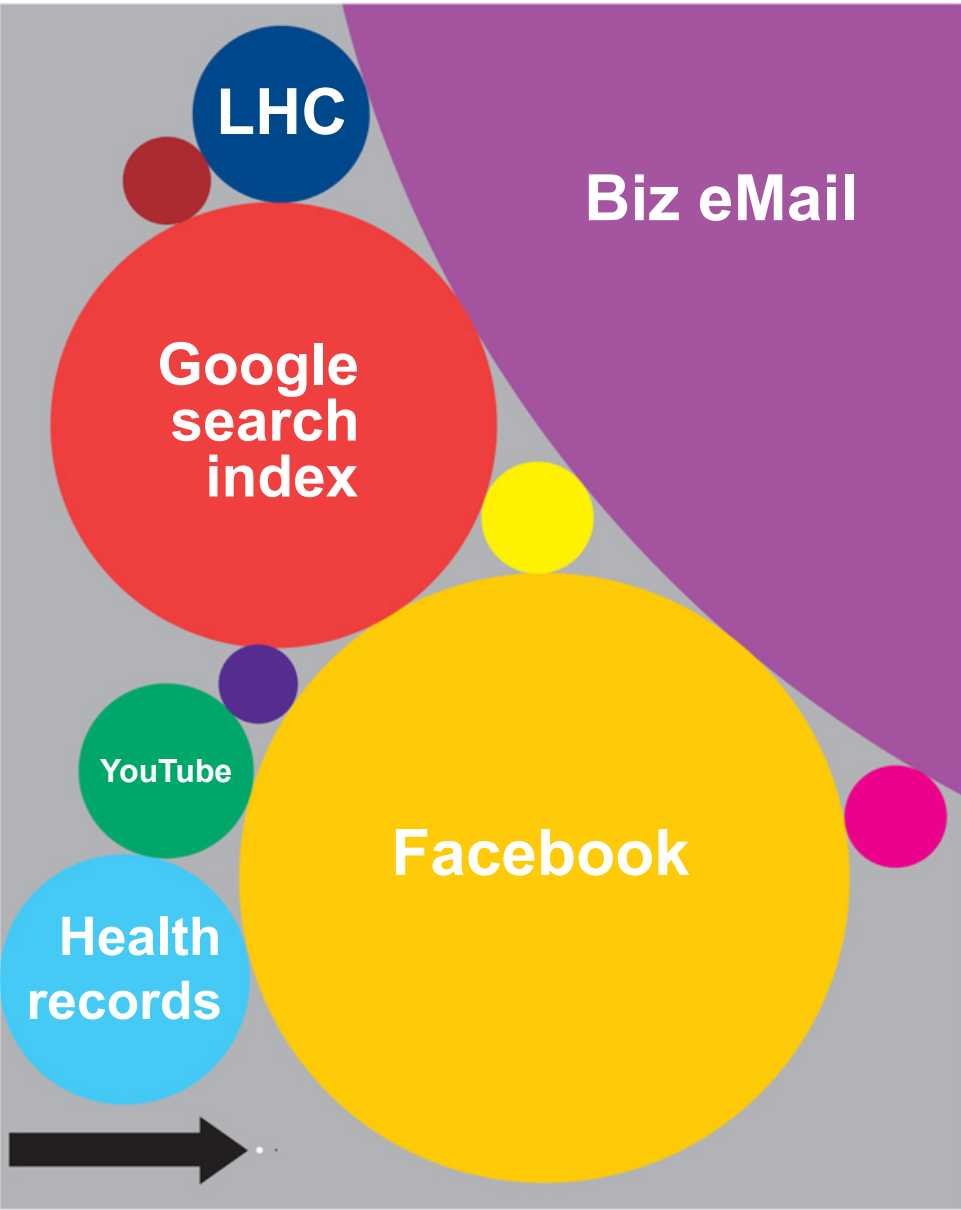
There
is
always
a
bigger
fish

Size of data sets in terabytes

Business email sent per year	2,986,100	National Climactic Data Center database	6,144
Content uploaded to Facebook each year	182,500	Library of Congress' digital collection	5,120
Google's search index	97,656	US Census Bureau data	3,789
Kaiser Permanente's digital health records	30,720	Nasdaq stock market database	3,072
Large Hadron Collider's annual data output	15,360	Tweets sent in 2012	19
Videos uploaded to YouTube per year	15,000	Contents of every print issue of WIRED	1.26

Computing vs Data

Computing per unit cost has doubled roughly every 18 months.

Hard Drive Cost per Gigabyte
1980 - 2009
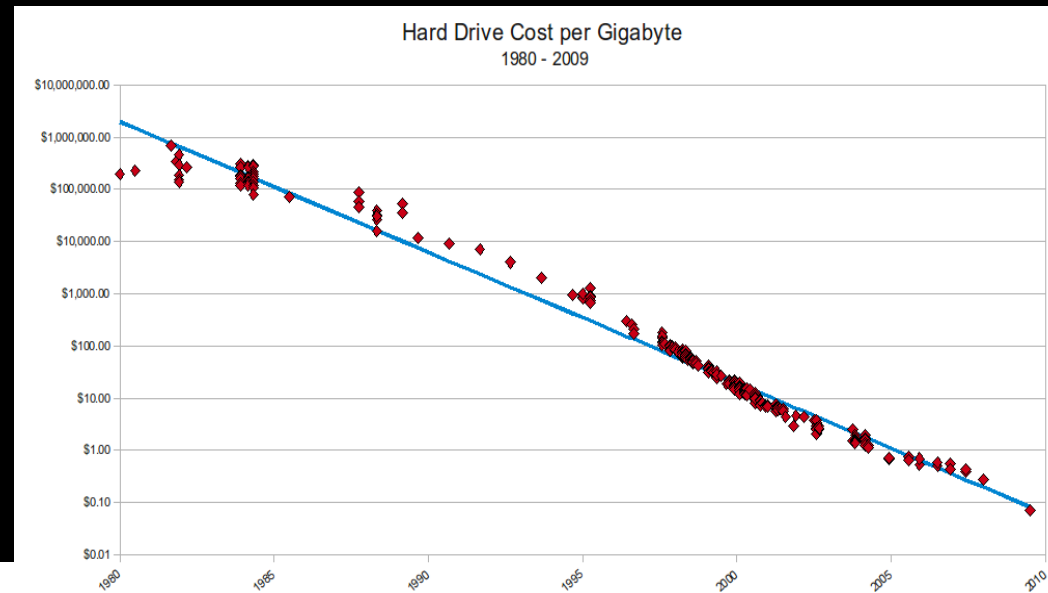
Performance Development
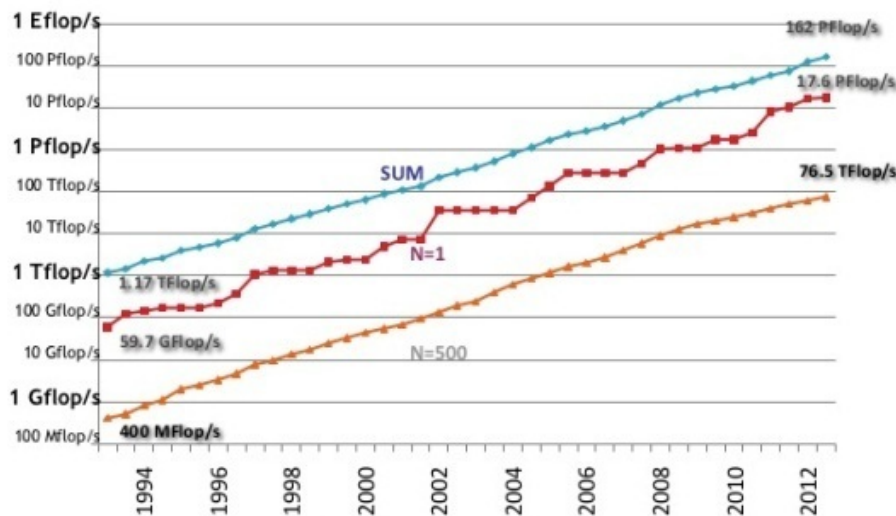
Space per unit cost has doubled roughly every 14 months.

So: data becomes exponentially uncomputable.

http://www.mkomo.com/cost-per-gigabyte

Mission

Can we create smart and safe data processing systems that can be tailored to diverse application needs?

- *Capacity*
  - *Bandwidth on demand, QoS, architectures, photonics, GPU, performance*
- *Capability*
  - *Programmability, virtualization, complexity, semantics, workflows*
- *Security*
  - *Anonymity, integrity of data in distributed data processing*
- *Sustainability*
  - *Greening infrastructure, awareness*
- *Resilience*
  - *Systems under attack, failures, disasters*

Research direction

- Information on Infrastructure
- Control & programmability of Infrastructure
- Virtualization
- Networked data processing
- Sustainability & Complexity

Events on the horizon

- I4DW & DSRC
  - Launch Nov 13
- PIRE & OpenScienceDataCloud.org
  - Workshop June 2014 @ UvA
- Research Data Alliance
  - Conference in Amsterdam Sept 2014

Announcement June 14-21 2014
PIRE Workshop Amsterdam

- OpenScienceDataCloud.org
- PIRE Fellowship Application (+/- 15)
- The OSDC PIRE Program is six to eight week fully funded fellowship for US graduate student researchers with an information technology background.
- Format:
  - 1 week tutorials and hands on training
  - 2 months research at a participating institute
  - Results in science/tools and papers/posters/

PIRE - OpenScienceDataCloud.org

1000 Genomes Project
Human sequence data from populations around the world with the goal of cataloging human genetic variation.
Total Size: 383.5TB   Categories: genomics, biology

ASTER
ASTER Level-1B Registered Radiance at the Sensor
Total Size: 12.7TB   Categories: earth science

Complete Genomics Public Data
Whole human genome sequence data sets provided by Complete Genomics, containing 69 standard, non-diseased samples as well as two matched tumor and normal sample pairs.
Total Size: 47.1TB   Categories: genomics, biology

Earth Observing-1 Mission
Data gathered by the Advanced Land Imager (ALI) Hyperspectral Imager (Hyperion) instruments on NASA's Earth Observing-1 Mission (EO-1) satellite.
Total Size: 45.2TB   Categories: earth science, satellite imagery

City of Chicago Public Datasets
Data set from the City of Chicago Data Portal in JSON format for tabular data and the raw files for "blob" data.
Total Size: 9.7GB   Categories: social science

EMDataBank
Unified Data Resource for 3-Dimensional Electron Microscopy
Total Size: 91.3GB   Categories: biology

Enron Emails
Data sets based on the original Enron emails released to the public by the Federal Energy Regulatory Commission as part of their investigation.
Total Size: 155.9GB   Categories: social science

FlyBase
FlyBase is the leading database and web portal for genetic and genomic information on the fruit fly Drosophila melanogaster and related fly species.
Total Size: 614.8GB   Categories: biology, genomics
....

Research Data Alliance

- https://rd-alliance.org
- The Research Data Alliance implements the technology, practice, and connections that make Data Work across barriers.
- The Research Data Alliance aims to accelerate and facilitate research data sharing and exchange.
  - Working groups and interest groups
  - Joining groups and attendance at the twice-yearly plenary meetings is open.
- Plenary Sep 2014 hosted by the Netherlands - Amsterdam
  - Conference Management Team (CMT) Chair: Peter Doorn (DANS)
  - Program Committee (PC): co-chairs Cees de Laat & Wouter Los (UvA)
  - Satellite Events Committee (SEC): Jeroen Rombouts (TUD)

…the DUTCH
Hosting of the 4th RDA Plenary Meeting in 2014

The RANDSTAD Area:
an excellent location with an excellent infrastructure

We combine:
50 years of experience in data curation

with

a strong commitment to and participation
in the RDA initiative

and

back up from RDA/Europe

RDA Core Programme

Satellite programme to attract young talent

Satellite programme interaction with private partners

Our offer:
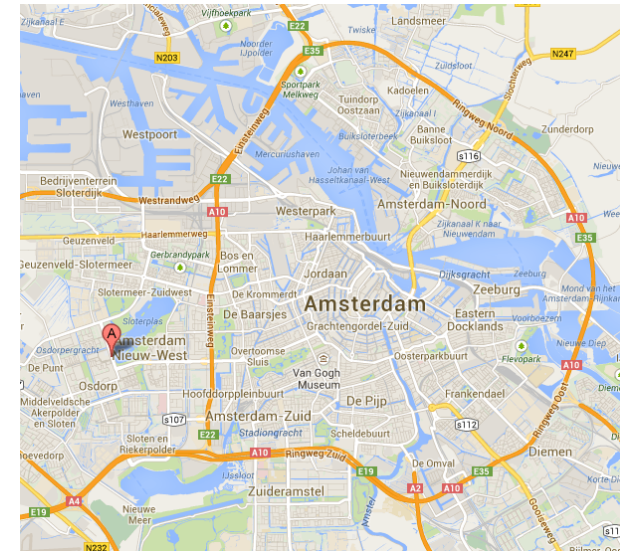
Location: **Amsterdam**

Goal: **600 people**
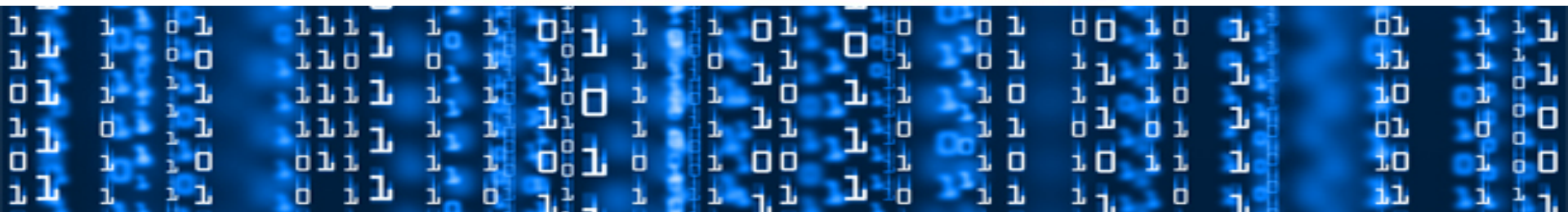
Date: **September/October 2014**

Dates & Location

- RDA:
  - Monday Sep 22 – Wednesday noon Sep 24
- Satellite events:
  - Sunday Sep 21 (optional)
  - Wednesday noon Sep 24– Friday Sep 26
- De Meervaart <http://www.meervaart.nl/congres>
- Conference center
  - Large auditorium (800)
  - Auditorium (285)
  - 10 parallel session rooms (12-120)
- Located in the southwest of Amsterdam

Program week, very very preliminary!

22-sep		23-sep		24-sep		25-sep		26-sep	
morning	afternoon	morning	afternoon	morning	afternoon	morning	afternoon	morning	afternoon
RDA 4P	RDA 4P	RDA 4P	RDA 4P	RDA 4P	EUDAT	EUDAT	EUDAT	EUDAT	EUDAT
						BioDiversity	BioDiversity	BioDiversity	BioDiversity
					Dutch Dataprize	DataCite	DataCite		
						ASTRON	ASTRON		
						DSRC-A'dam	DSRC-A'dam	DSRC/UvA-Bits	DSRC/UvA-Bits
						DSA workshop	DSA workshop		
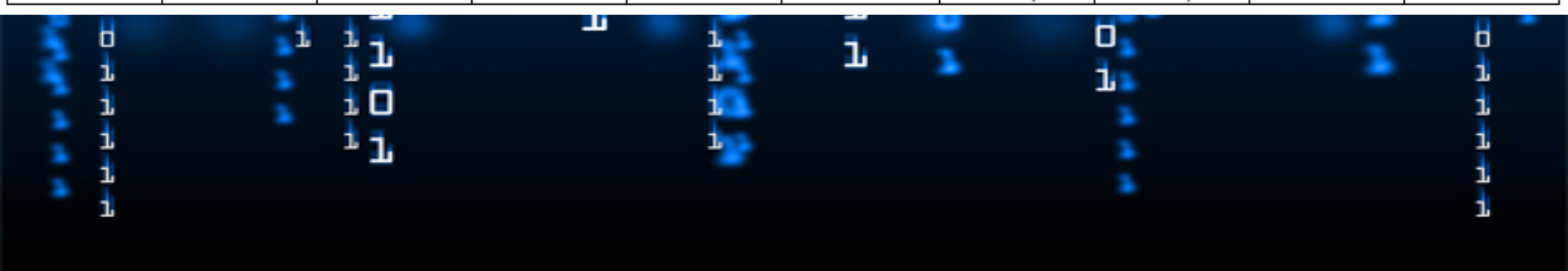
Program RDA, very very preliminary!

	Day 1		Day 2		Day 3
08:00 08:30	Registration		Registration		Registration
09:00 09:30 10:00	Welcome & Keynotes		Keynotes		Parallel 7
10:30	Coffee		Coffee		Coffee
11:00 11:30 12:00	Parallel 1		Parallel 4		Keynote/Wrap up
12:30 13:00	Lunch		Lunch		Lunch
13:30 14:00 14:30	Parallel 2		Parallel 5		
15:00	Tea		Tea		
15:30 16:00 16:30	Parallel 3		Parallel 6		
17:00 17:30 18:00	Posters + reception		Posters + reception		
18:30 19:00 19:30 20:00 20:30 21:00 21:30 22:00	Social Event?		Social Event & Conference Dinner		

2 – 3 Keynotes
9 sessions parallel
- **56 RG/WG**
- **7 PC**

RDA
RESEARCH DATA ALLIANCE

Committee's

- **Conference Management Team (CMT):**
    - Chair: Peter Doorn (DANS)
    - Ingrid Dillo (DANS)
    - Elly Dijk (DANS)
    - Cees de Laat (chair PC)
    - Jeroen Rombouts (chair SEC)
    - Herman Stehouwer: as liaison RDA
    - Andrew Treloar: as consultant

- **Program Committee (PC):**
    - Cees de Laat (Chair)   UvA
    - Wouter Los (Co-Chair)   UvA
    - Herman Stehouwer   MPI
    - Patrick Aerts   NL – eScience
    - Tony Hey   USA, Microsoft
    - Sander Klous   KPMG
    - Satoshi Sekiguchi   AIST - Japan
    - Leif Laaksonen   RDA Europe
    - Liu Chuang   Chinese partnership
    - Dinesh Katre   India
    - Wim Hugo   South Africa
    - Andrew Treloar   Australia

    - **Satellite Events Committee (SEC):**
    - Chair: Jeroen Rombouts (TUD)
    - Jacco Konijn (UvA)

We look forward to 2014…

..and hope to welcome you all in the Netherlands!

Thank you, also on behalf of all our Dutch partners,
Cees de Laat (UvA)
Peter van Doorn (DANS)

Questions?

DSRC

SNE

http://sne.science.uva.nl

http://www.os3.nl/

http://i4dw.nl/

http://dsrc.nl/

http://sne.science.uva.nl/openlab/

http://pire.opensciencedatacloud.org

http://staff.science.uva.nl/~delaat/pire/

https://rd-alliance.org

Arie Taal
Paola Grosso Ana Oprescu
Cees de Laat Marc Makkes Ralph Koning
Bas Terwijn Leon Gommans Fahimeh Alizadeh
Pieter Adriaans Cosmin Dumitru Karst Koymans
Yuri Demchenko Rob Meijer Karel van der Veldt
Rudolf Strijkers Miroslav Zivkovic Reggie Cushing
Spiros Koulouzis Hao Zhu Jan Sipke van der Veen
Naod Duga Jebessa Sander Klous
Guido van 't Noordende
Jaap van Ginkel Jeroen van der Ham
Mikolaj Baranowski Steven de Rooij
Ngo Tong Canh Souley Madougou Paul Klint
Adianto Wibisono
Zhiming Zhao Magiel Bruntink
Anna Varbanescu Marijke Kaat
Niels Sijm Hans Dijkman Gerben de Vries
Adam Belloum Arno Bakker Marian Bubak
Daniel Romao Erik-Jan Bos
Peter Bloem