



Amsterdam  
Data Science

# Coffee & Data on Infrastructures

**Henri Bal (VU)**

**Cees de Laat (UvA)**

**COMMIT/**

# Program

- 09:05-09:30:
  - Leon Gommans (KLM): Smart Industry Future Internet Field Lab
- 09:30-09:55:
  - Pieter Hijma (UvA, VU): Preparing for the Future of Computing
- 09:55-10:20:
  - Niels Nes (CWI): The SciLens Infrastructure, Big Data at Work
- 10:20-10:40: Coffee Break
- 10:40-11:05:
  - Ben Kröse (HvA, UvA): Sense in the City
- 11:05-11:30:
  - Rutger Hofman (VU) + Anton Feenstra (VU):  
Discussion on Amsterdam Data Science Infrastructure
- 11:30-12:00: Pitches & Networking
  - Roshan Das (VU): SWAN
  - Ralph Koning (UvA): SARnet



# A Medium-Scale Distributed System for Computer Science Research: Infrastructure for the Long Term

**Henri Bal**, VU University Amsterdam

**Dick Epema**, Delft University of Technology

**Cees de Laat**, University of Amsterdam

 **Job van Nieuwpoort**, Netherlands eScience Center

**John Romeln**, ASTRON

**Frank Selnstra**, Netherlands eScience Center

**Cees Snoek**, University of Amsterdam

**Harry Wijshoff**, Leiden University

**L**ike any science that heavily uses experimentation, computer science depends on quality physical research infrastructures based on supercomputers and clouds through which researchers can access platforms for large-scale collaborative experiments. Although these parallel and distributed production systems have many benefits for computational work, such as rapid results when running complex simulations, they are ill-suited for research that needs detailed hardware and software control. The difficulties of using production systems to conduct controlled, reproducible, distributed experiments on multiple resources are well known.

PlanetLab, a global research network that supports the development of new network services, federates existing resources to enable large-scale experiments on thousands of nodes. However, many applications use

*The Dutch Advanced School for Computing and Imaging (ASCI) has built five generations of a 200-node distributed system over nearly two decades while remaining aligned with the shifting **computer science research** agenda. The system has supported years of award-winning research, underlining the benefits of investing in a smaller-scale, tailored design.*

the nodes simultaneously, making it hard to obtain reliable performance measures. Grid'5000 is a from-scratch large-scale distributed system dedicated to computer