# Microdata and remote access
## - Experiences in Finland

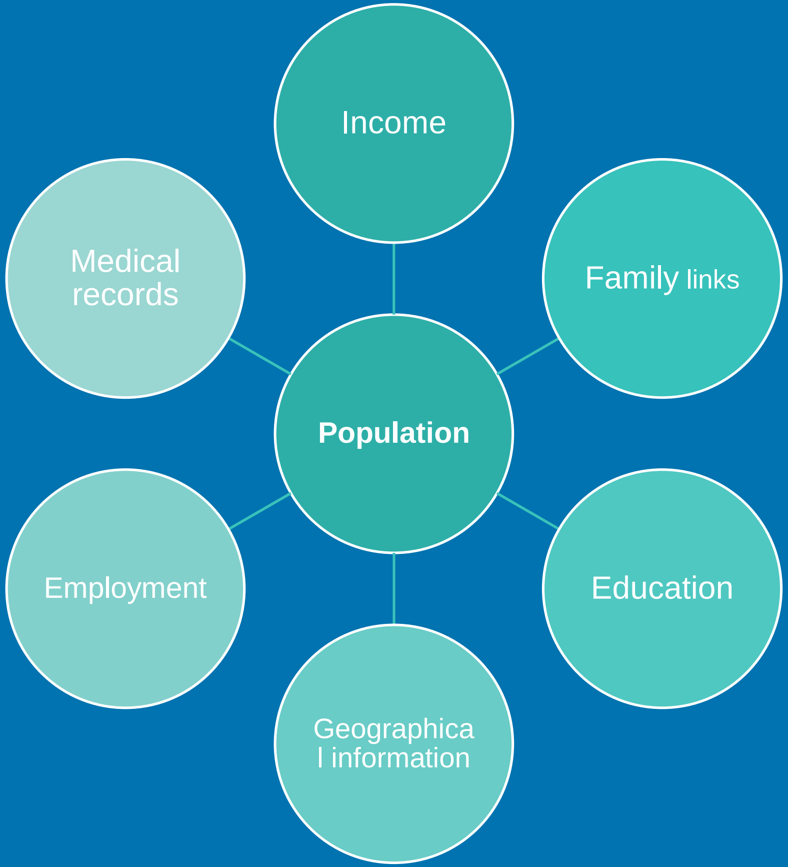
Markus Wanamo, Statistics Finland
3rd March 2020, KSWD

# Research services at Statistics Finland

- Provide researchers and other government institutions with access to register and survey data
  - Handel license requests
  - Produce data sets
  - Manage and develop the remote access system FIONA

- Maintain and develop a tax-benefit microsimulation model

# Microdata at Statistics Finalnd



- Wide range of data since moving to registerbased census in 1987
  - Individuals
  - Households
  - Firms
- Tailor-made and ready made datasets
  - Taika - research data catalogue [https://taika.stat.fi/](https://taika.stat.fi/)

# Modes of distribution and access

- Delivered datasets
  - Must be totally anonymized (samples, rounding, top-coding etc.)
  - Still common for survey data or ’service sets’ (small sample sizes)
- Remote access
  - Pseudonymized (only direct identifiers need be deleted)
  - Total data available for research if the need is justified
  - Also more sensitive data can be accessed (e.g. medical and criminal records)
- ’Research laboratory’ (on site)
  - Decreased demand

# Criteria for gainig license and access to data

- Legal basis in Statistics Act (reformed 2013)
  - License can be granted for specified research pourposes and statistical surveys

- Can be obtained from abroad
  - EU member states
  - Other countries determined by the European Commission to have adequate data protection

- Affiliation with a Finnish institution required for remote access (geo-restriction on IP-adresses)

# FIONA remote access system

- Secure closed environment for handling of micro-level data

- First pilot at Statistics Finland in 2010
  - Modelled after similar systems in the Netherlands, Denmark, Sweden
  - Geographical equality
  - Increased efficiency
  - Data security

Research projects 217

557 Users

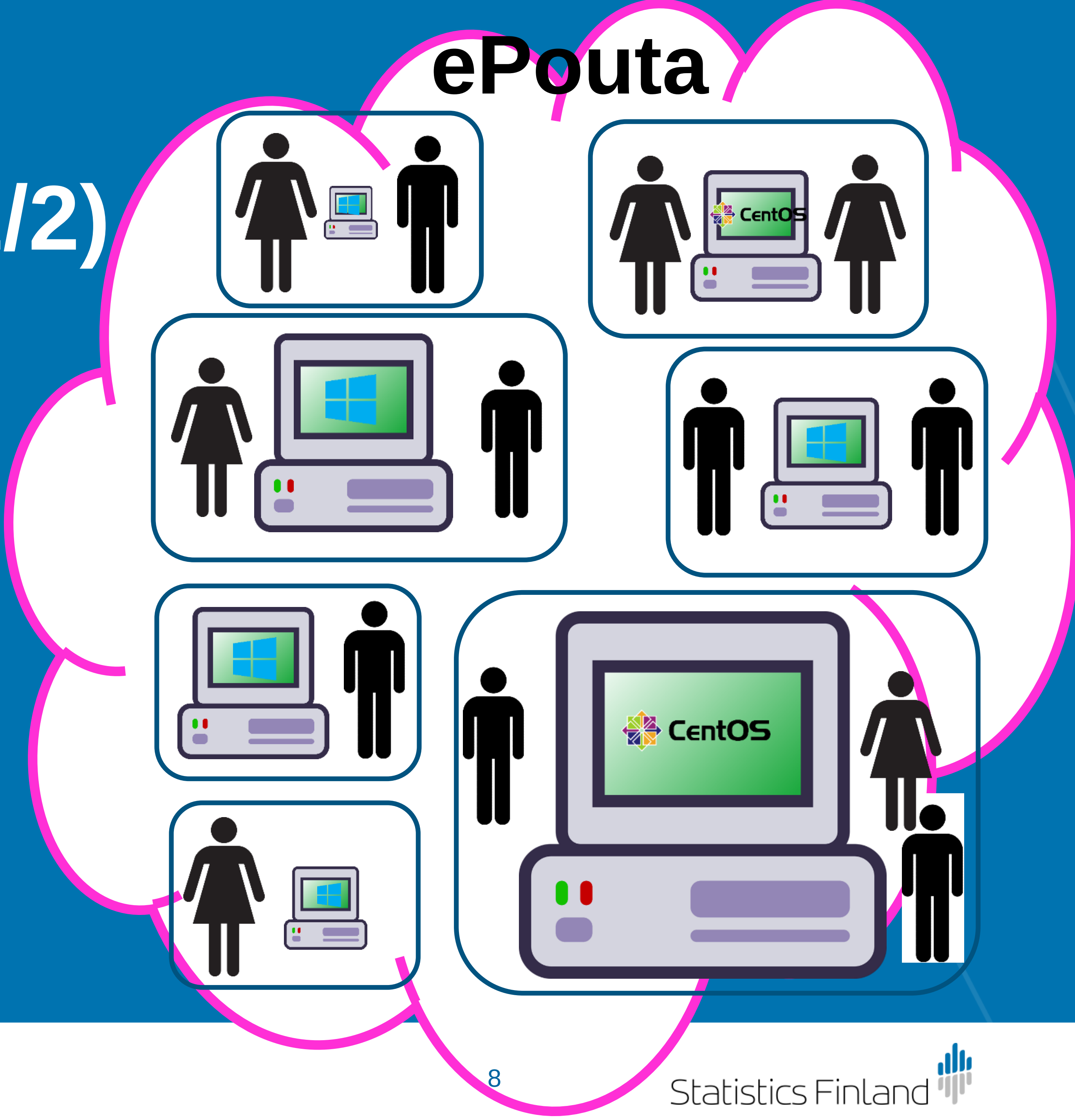
Active, november 123/223

# Data security

- Administrative
  - License to use data
  - Pledge of secrecy
  - Remote access commitment
  - Our rules and regulations
- Technical
  - IP-restriction
  - MFA
  - DMZ
  - Interface policies

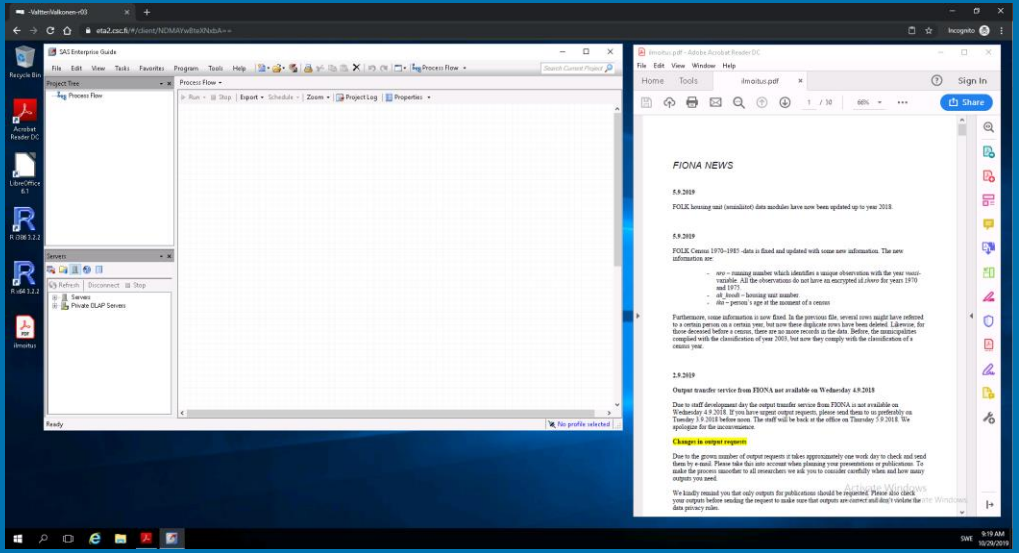
# Technical implementation (1/2)

- The ePouta cloud
  - Virtual private cloud designed to meet the security requirements for handling sensitive data.
  - Maintained by CSC - IT Center for Science (non-profit state enterprise)
- Individual virtual machines for each reasearch project
  - Researchers can choose size of dedicated resource



# Technical implementation (2/2)

- Windows desktop accessed via browser
  - Apache Guacamole
- Applications
  - SAS
  - Stata
  - R
  - Libre office etc.



*FIONA NEWS*

**5.9.2019**

FOLK housing unit (asuinliitot) data modules have now been updated up to year 2018.

**5.9.2019**

FOLK Census 1970–1985 -data is fixed and updated with some new information. The new information are:

- *nro* – running number which identifies a unique observation with the year vuosi-variable. All the observations do not have an encrypted id *shnro* for years 1970 and 1975.
- *ak_koodi* – housing unit number.
- *ika* – person's age at the moment of a census

Furthermore, some information is now fixed. In the previous file, several rows might have referred to a certain person on a certain year, but now these duplicate rows have been deleted. Likewise, for those deceased before a census, there are no more records in the data. Before, the municipalities complied with the classification of year 2003, but now they comply with the classification of a census year.

**2.9.2019**

**Output transfer service from FIONA not available on Wednesday 4.9.2018**

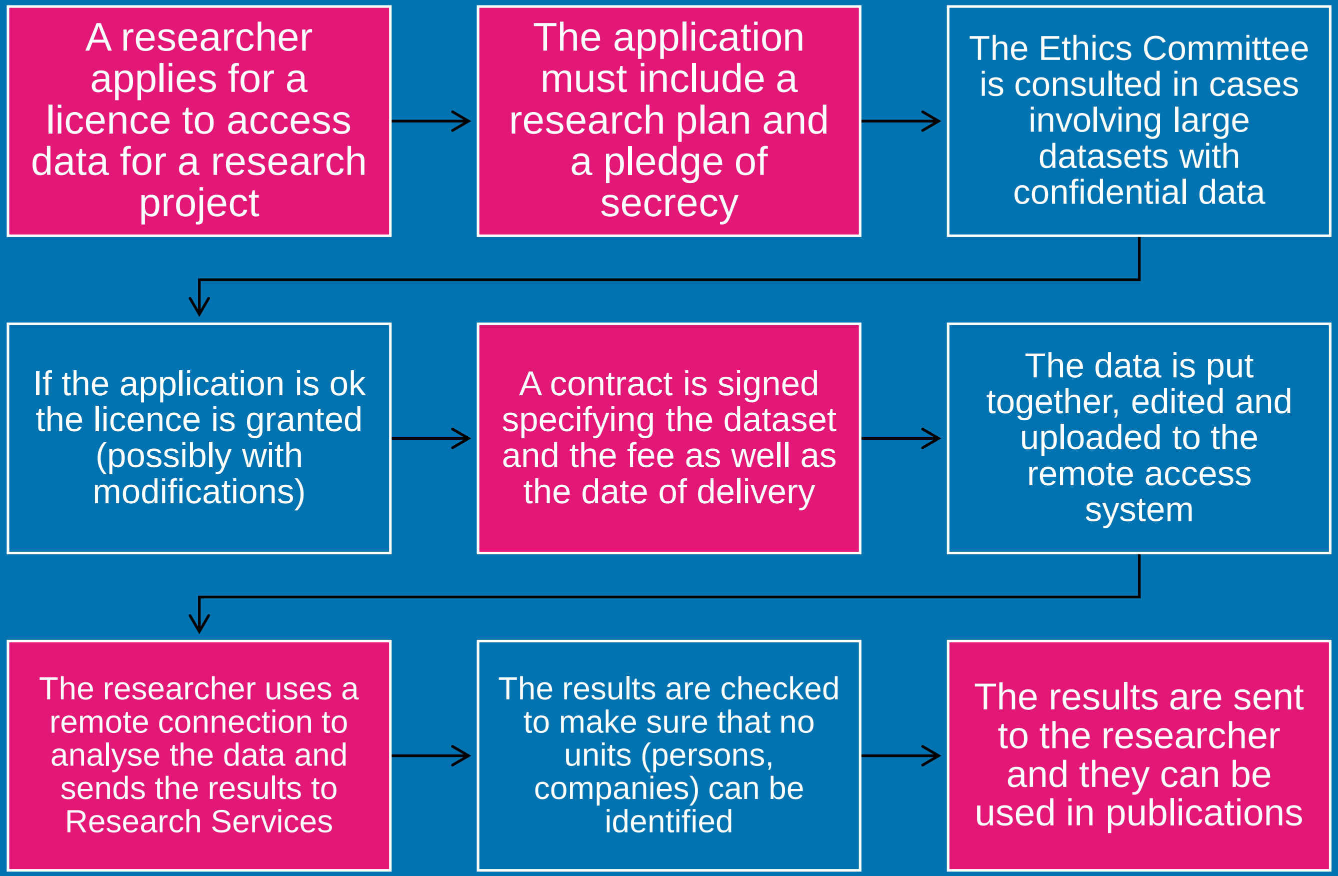
Due to staff development day the output transfer service from FIONA is not available on Wednesday 4.9.2018. If you have urgent output requests, please send them to us preferably on Tuesday 3.9.2018 before noon. The staff will be back at the office on Thursday 5.9.2018. We apologize for the inconvenience.

**Changes in output requests**

Due to the grown number of output requests it takes approximately one work day to check and send them by e-mail. Please take this into account when planning your presentations or publications. To make the process smoother to all researchers we ask you to consider carefully when and how many outputs you need.

We kindly remind you that only outputs for publications should be requested. Please also check your outputs before sending the request to make sure that outputs are correct and don't violate the data privacy rules.

# A typical process in applying for sensitive research data



A researcher applies for a licence to access data for a research project

The application must include a research plan and a pledge of secrecy

The Ethics Committee is consulted in cases involving large datasets with confidential data

If the application is ok the licence is granted (possibly with modifications)

A contract is signed specifying the dataset and the fee as well as the date of delivery

The data is put together, edited and uploaded to the remote access system

The researcher uses a remote connection to analyse the data and sends the results to Research Services

The results are checked to make sure that no units (persons, companies) can be identified

The results are sent to the researcher and they can be used in publications

# Pricing for research data and services

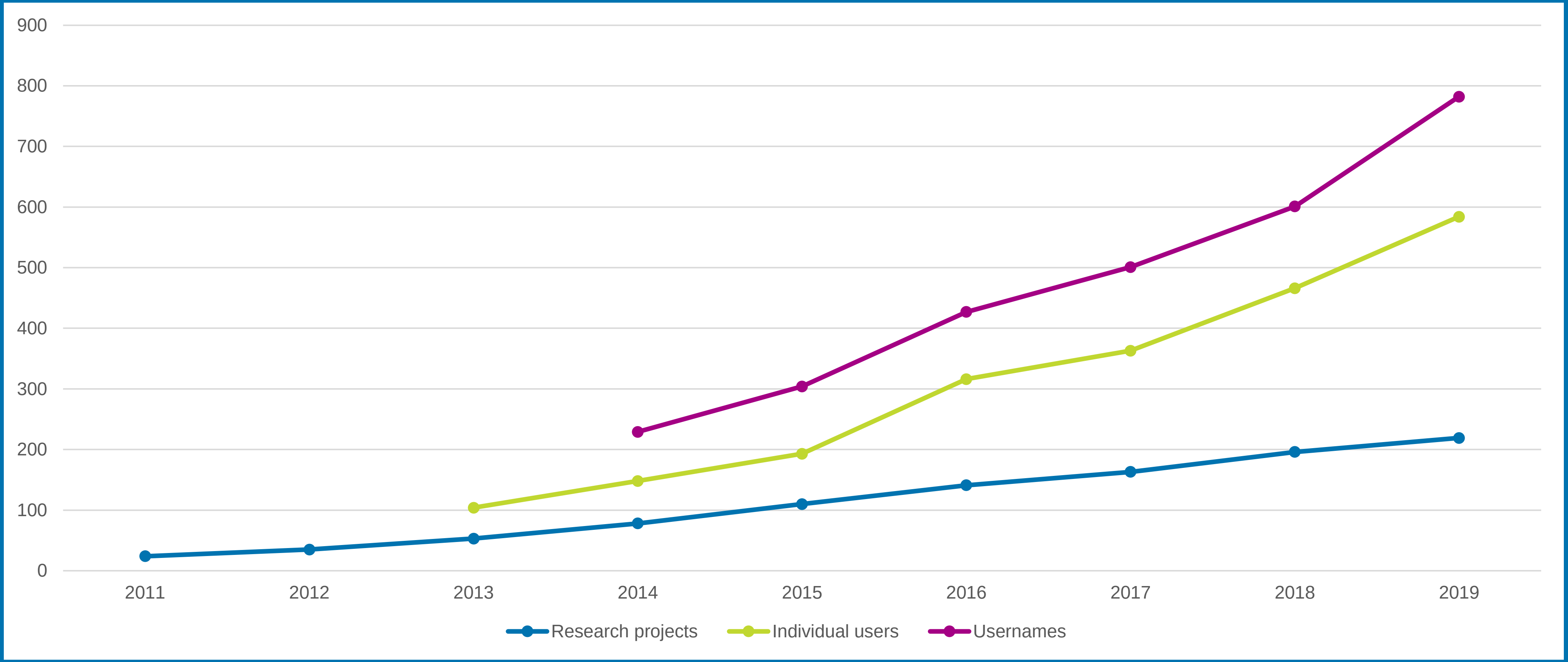
- Ready-made datasets have a fixed price
  - FOLK-modules 300 euro per module

*E.g. FOLK Basic data + Income module + Family module = 900 euros*

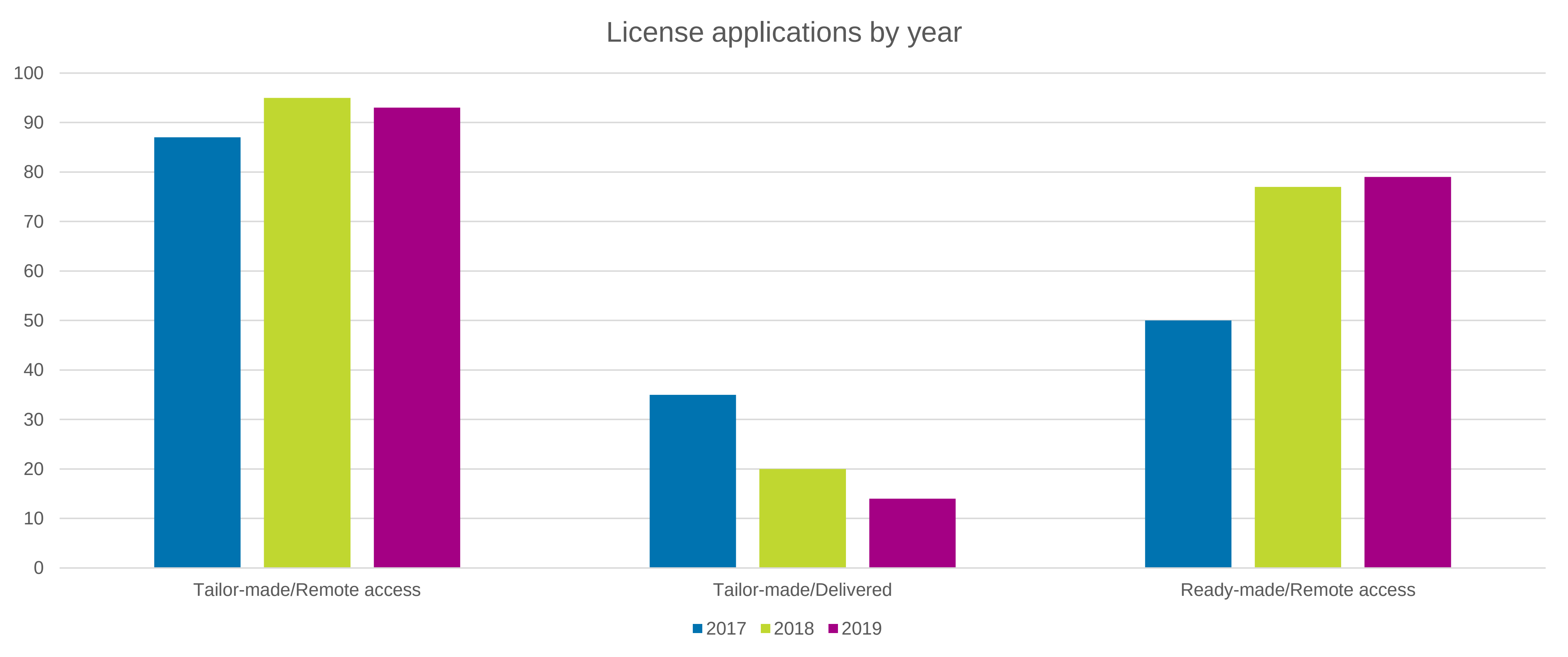
- Tailor-made data priced according to working hours required

- Cost of remote access determined by size of virtual machine
  - 2250-3500 EUR/year

# Rise in use of remote access



# Changes in type of access



# Development

- From tailor-made data to ready-made datasets
  - Increased efficiency (tailor-made data is labour-intensive)
  - More easily comparable results and replication researchers use identical datasets
- From delivered data to remote access
  - Proper control over distribution
  - Researchers can access better data (pseudonymized vs. anonymized, total population vs. sample)
  - Growing demand increases workload (e.g. more inquiries, output-checking)

Thank you!

Contact:
tutkijapalvelut@stat.fi

Homepage:
www.stat.fi

Taika - research data catalogue:
https://taika.stat.fi/