



# Official statistics and privately held data: challenges of today pave the way to the best practices of the future

Kaisa Vent

Pisa 2022



# Synopsis

1. Why business to government for official statistics (B2G4S) data sharing?
2. Mobile positioning data (MPD) as an example of privately held data (PHD) that is being used for official statistics.
3. Access and processing constellations.
4. Building trust and finding balance.
5. Skills needed for operating in B2G4S domain.

## Why include alternative data sources?

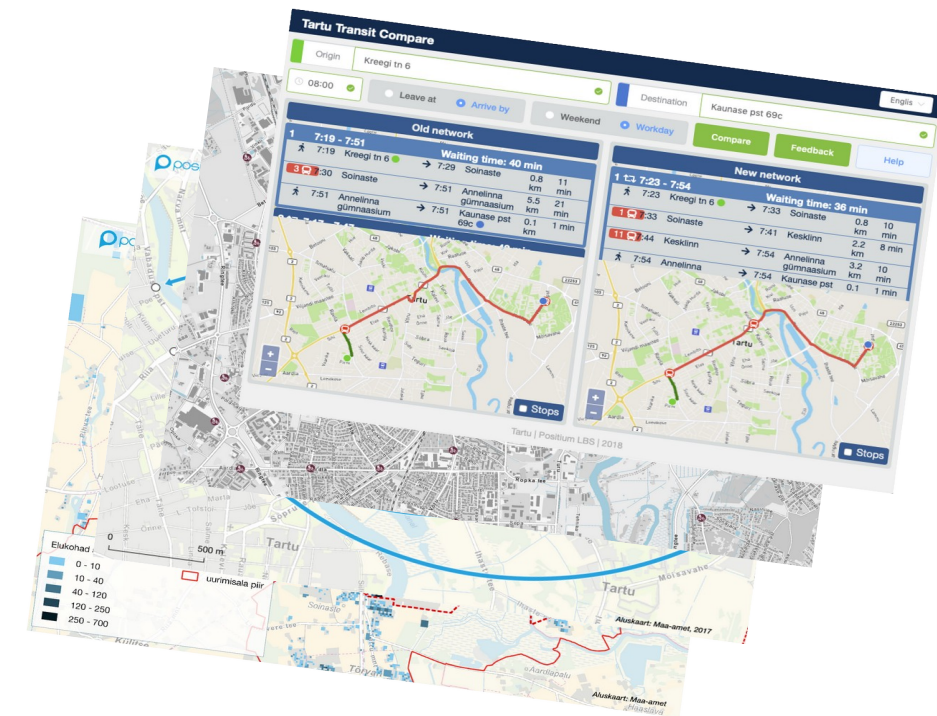
- **Changes:** the European Green Deal, NextGenerationEU.
- Need for **timely, relevant** and **accurate** measures for impact and proactive approach.
- **Cross-validation:** multiple sources to measure a phenomena.
- More information can lead to **stronger foundation**.
- **Single** data source vs **combination** of data sources



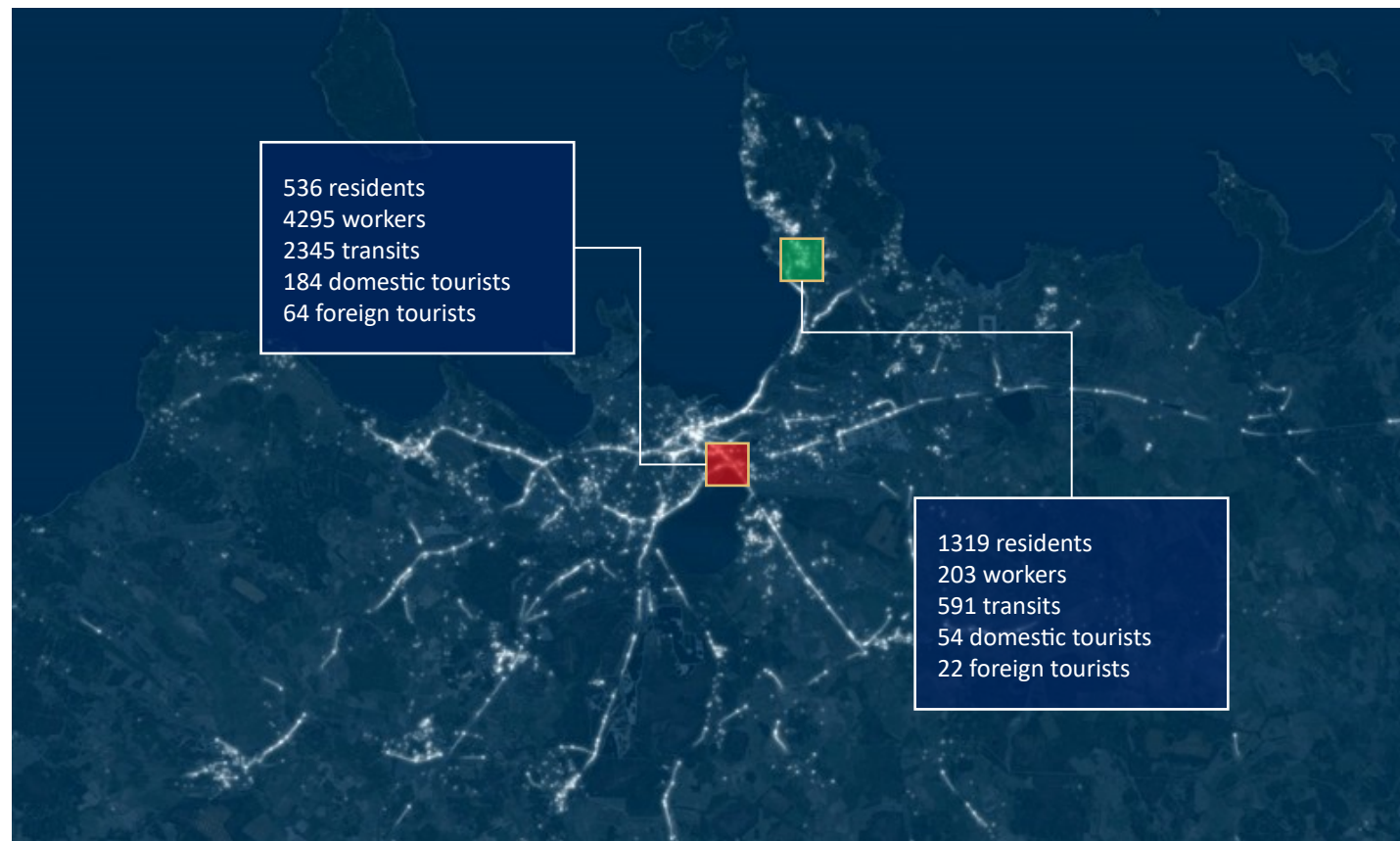
**Increased bus ridership** from month 1, results continued to improve during COVID-19 (with the exception of the first wave of lockdown)

**Bicycles** were so popular that just within the first two days, they travelled over **80 000 km**

Source: Positium



# Mobile Positioning Data (MPD)



Source: Positium

Information about people's **presence**:

- **Temporal** variations
- **Spatial** variations



# Official statistics and MPD in Estonia

- Estonia was the first country in the world to start producing official statistics based on mobile positioning data (MPD):
  - Inbound and outbound tourism statistics for Bank of Estonia
  - The **longest running official statistics** time series based on MPD in the world – 14 years.
- Collaboration model involves three main parties:
  - **Statistics producer (Bank of Estonia)** – responsible for dissemination and quality assurance according to the statistics code of practice. MPD collection through mandate of the Statistics Act.
  - **Data provider (MNO)** – on-time data provision.
  - **Data processor and trusted intermediary (Positium)** – developing methodology fit for purpose and processing raw data into tourism domain indicators.



CENTRAL BANK OF ESTONIA

**4x**  
faster

**200x**  
sample size

**12x**  
countries  
breakdown

**2.5x**  
more cost-  
efficient

**100%**  
less burden  
on tourists

Source: Positium



## Official statistics and MPD in Indonesia<sup>1</sup>

### Use case and collaboration model

- Statistics Indonesia has used MPD since 2016 for **measuring cross-border tourism**.
- Priorly used survey method was too **expensive** (lots of remote areas) and introduced **under-coverage issues**.
- **Collaboration model** involved 4 main parties:
  - Statistics Indonesia
  - Ministry of Tourism
  - Data provider (MNO)
  - Trusted intermediary (Positium)

### Lessons learned

- More **accurate** compared to cross-border survey.
- MPD did introduce noise that needed **additional processing** steps to remove it from final results.
- **Surveys** can help **validate** MPD based results and provide input to **counter** other **coverage issues**.
- There is no data source that is superior, data source should **complete** not **compete** each other!

<sup>1</sup>Lestari, T. K., Esko, S., Sarpono, Saluveer, E., Rufiadi, R. (2018). Indonesia's Experience of using Signaling Mobile Positioning Data for Official Tourism Statistics

## Census in Estonia<sup>2</sup>

### Challenge

One of the biggest challenges on the way to register based census in Estonia is the **difference between registered and actual places of residence**.

**20–25%** of all permanent residents of Estonia have **not registered their actual place of residence in PR** and live at a different address.


***Anchor point model** for identification of country of residence, **place of residence**, work-time, second home, usual environment, and other regular meaningful locations*



### Solution

**Mobile positioning data for validating registries.** Cooperation between NSO, university and private company.

The pilot study consists of following steps:

- ✗ **Panel study** with volunteers
- Set of potential addresses was created for each volunteer based on registers.
- **Home locations (HL)** were also calculated based on the **mobile positioning data**. 
- MPD-based HL and other auxiliary information was used to build a **model** for selecting the **most probable place of residence** from the set of addresses.

<sup>2</sup>Sõstra, K., Lehto, K. (2018). Using mobile positioning for improving the quality of register data.

PANEL	NSO ACCESS TO PSEUDONYMIZED RAW DATA	PROCESSING IN MNO SERVERS
<p>Volunteer panel opts in to allow access to their data from registries, census and telecom operators</p>	<p>All data is de-identified through pseudonymization of subscriber ID and transferred to NSO</p>	<p>Data is de-identified and (1) MNO processes the data in their own servers according to set criteria or (2) MNO creates a sandbox for NSO data scientists</p>
<ul style="list-style-type: none"> <li>+ Can link directly link MPD data with other data</li> <li>+ Proof of concept</li> <li>+ Good for methodological development and validation</li> </ul>	<ul style="list-style-type: none"> <li>+ Full control of data processing</li> <li>+ Enables more experimentation and use cases</li> </ul>	<ul style="list-style-type: none"> <li>+ Data is not transferred</li> <li>+ Fewer approvals needed</li> </ul>
<ul style="list-style-type: none"> <li>- Usually small sample size</li> <li>- Difficult to scale</li> </ul>	<ul style="list-style-type: none"> <li>- Requires strong commitment from NSO and TRA</li> <li>- Batch upload, lower frequency</li> </ul>	<ul style="list-style-type: none"> <li>- Difficult to check if method are applied correctly (unless done in sandbox)</li> </ul>





## Operational modalities of data reuse<sup>3</sup>

- Most **suitable operational modality** is highly **dependent** on the **stakeholders** working together for the common goal.
- The task of **establishing fit-for-purpose operational modalities** that minimise the total costs, risks and burden should be left to mutual agreement between the statistical authorities and the data holder concerned.
- There are **four main dimensions** that should be considered:
  - access
  - quality
  - roles and responsibilities
  - cost.

<sup>3</sup>Empowering society by reusing privately held data for official statistics - A European approach.

Final report prepared by the high-level expert group on facilitating the use of new data sources for official statistics. 2022 edition.

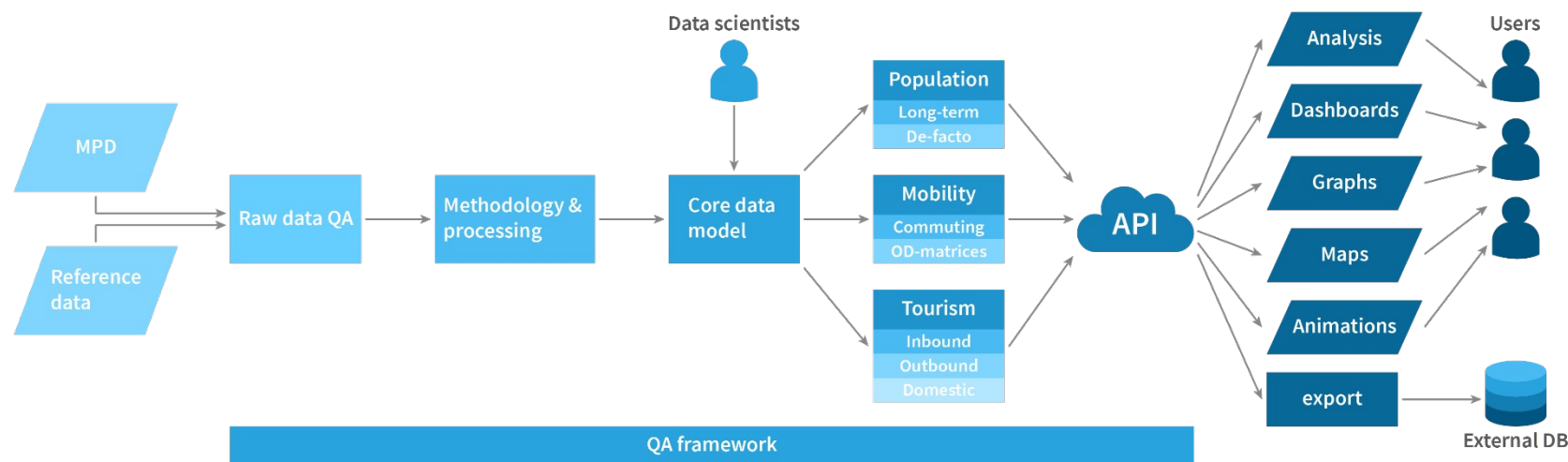


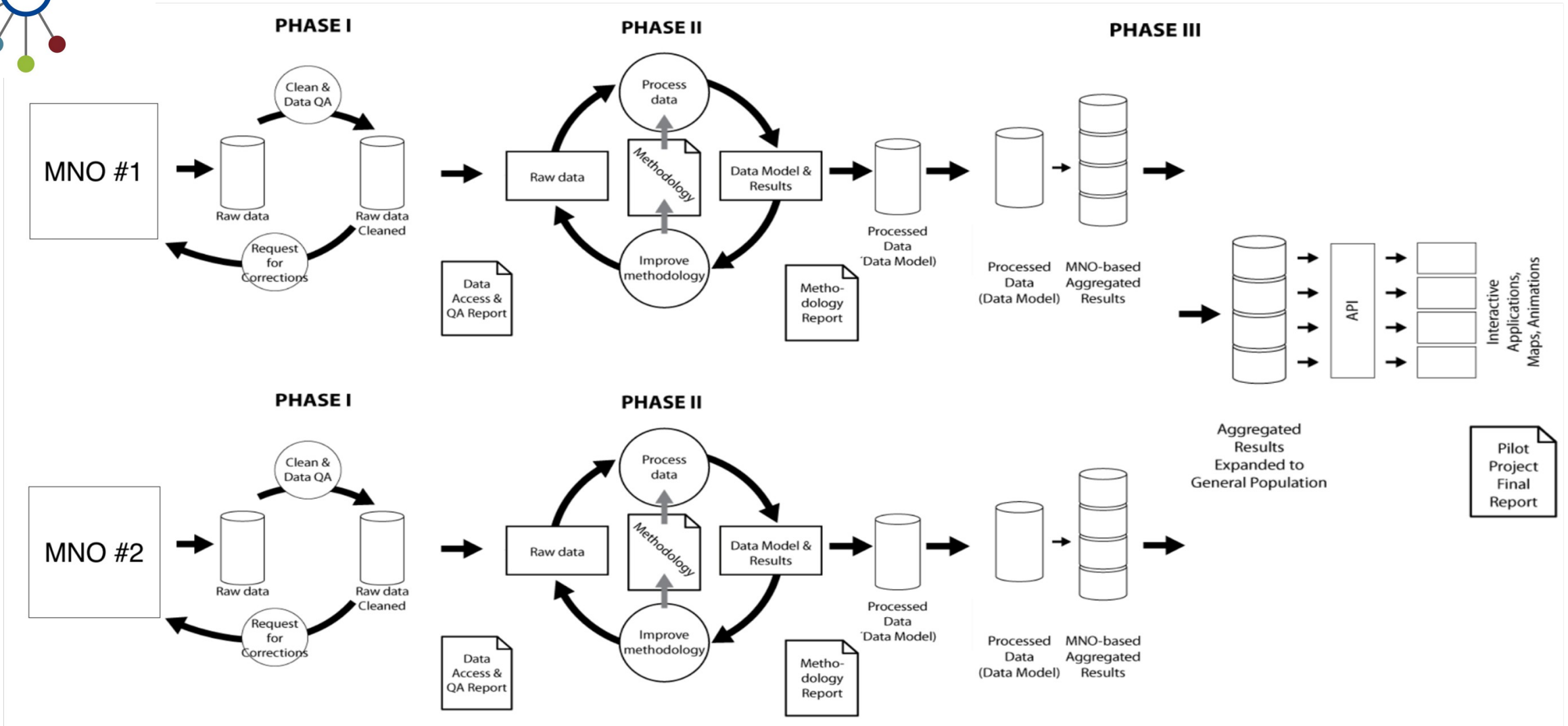
## Data access

- Time extension:
  - **continuous production** of statistical products
  - **pilot project** - exploring potential for new statistical products.
- Purpose:
  - **new statistical product** - research, exploration and design process
  - **regular statistical product** – production process.
- Data processing:
  - **raw** – more **flexibility** and bigger role in processing (NSO), **reduced burden** (data holder)
  - **processed** – **increased importance** for transparency and documentation (data holder).
- Location:
  - **On-site** – on the premises of data holder
  - **Off-site** – on the premises of NSO, trusted intermediary, other government body.

# Process quality

- Methods - methodological decisions should be always **considered against fit-for purpose**.
- Input – significant effect on output quality, rigorous **quality assurance procedures** should be in place.
- Throughput – discover **methodological inconsistencies** and **inherent changes** to the data.
- Output – **comparability over time**, consistency checks with other similar statistical product (**cross-validation**).





RESPONSIBILITY	ROLE
Privacy protection	Any of the involved stakeholders dependent on their role
Methodology development	Setting out an end-to-end methodology may involve <b>co-development</b> between <b>statistical authorities</b> and <b>data holders</b> , and possibly <b>specialised third parties</b> .
Methodology implementation and execution	<b>Statistical authorities</b> , the <b>private data holder</b> can be required to do this or it can be outsourced to a <b>third party</b> .
Auditing and approval	<b>Statistical authorities</b> should <b>retain responsibility</b> for auditing and approving the methods and solutions for the methodological components. In this way the authorities can <b>guarantee compliance</b> with the overall methodological framework and end-to-end quality.
Integration	<b>Statistical authorities</b> should arrange for adequate <b>protection of business secrets</b> and <b>data confidentiality</b> . PET technologies, trusted intermediary.



# Building trust and creating balance

## Maintain trust<sup>3</sup>

- Fit for purpose
- Professional independence
- Privacy protection
- Quality assurance
- International comparability

## Create balance

- Choose **methodological model** that is the best approximation of reality (model).
- Understand and communicate **differences between the model and the reality.**
- Conduct **data privacy impact assessment** - is proposed method proportional to privacy risks?
- Use the above **throughout the project** and come back to these assessments when needed.

<sup>3</sup>Jansen, R., Kovacs, K., Esko, S., Saluveer, E., Sõstra, K., Bengtsson, L., Li, T., Adewole, W. A., Nester, J., Arai, A., Magpantay, E. (2021). Guiding principles to maintain public trust in the use of mobile operator data for policy purposes. *Data & Policy*, 3, E24. doi:10.1017/dap.2021.21



## PHD related skills

- PHD domain entails **myriad of different technologies** and **systems** with every new data source bringing in something new.
- No need to become an expert in all of those aspects. However, you need to be able to **differentiate** between good and bad quality solution.
- **No single tech stack** or skillset to cover all those aspects PHD will bring.
- Instead **willingness to:**
  - communicate
  - co-develop
  - be agile
  - learn a common language
  - explain.



# Questions