LIGHTWEIGHT MULTI-DIRECTION-OF-ARRIVAL ESTIMATION ON A MOBILE ROBOTIC PLATFORM

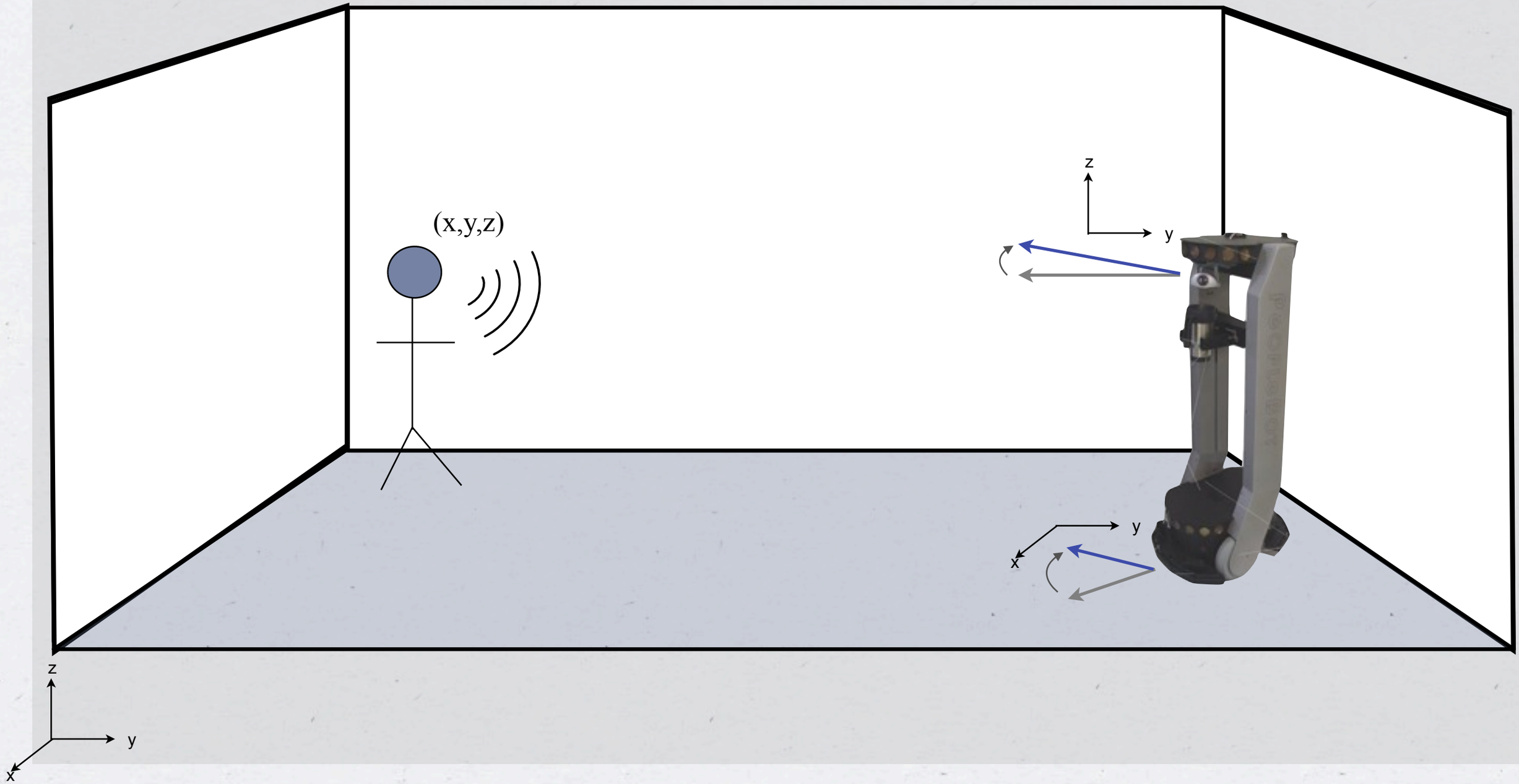
Dr. Caleb Rascon, Dr. Luis Pineda

Oct. 26-29, 2012

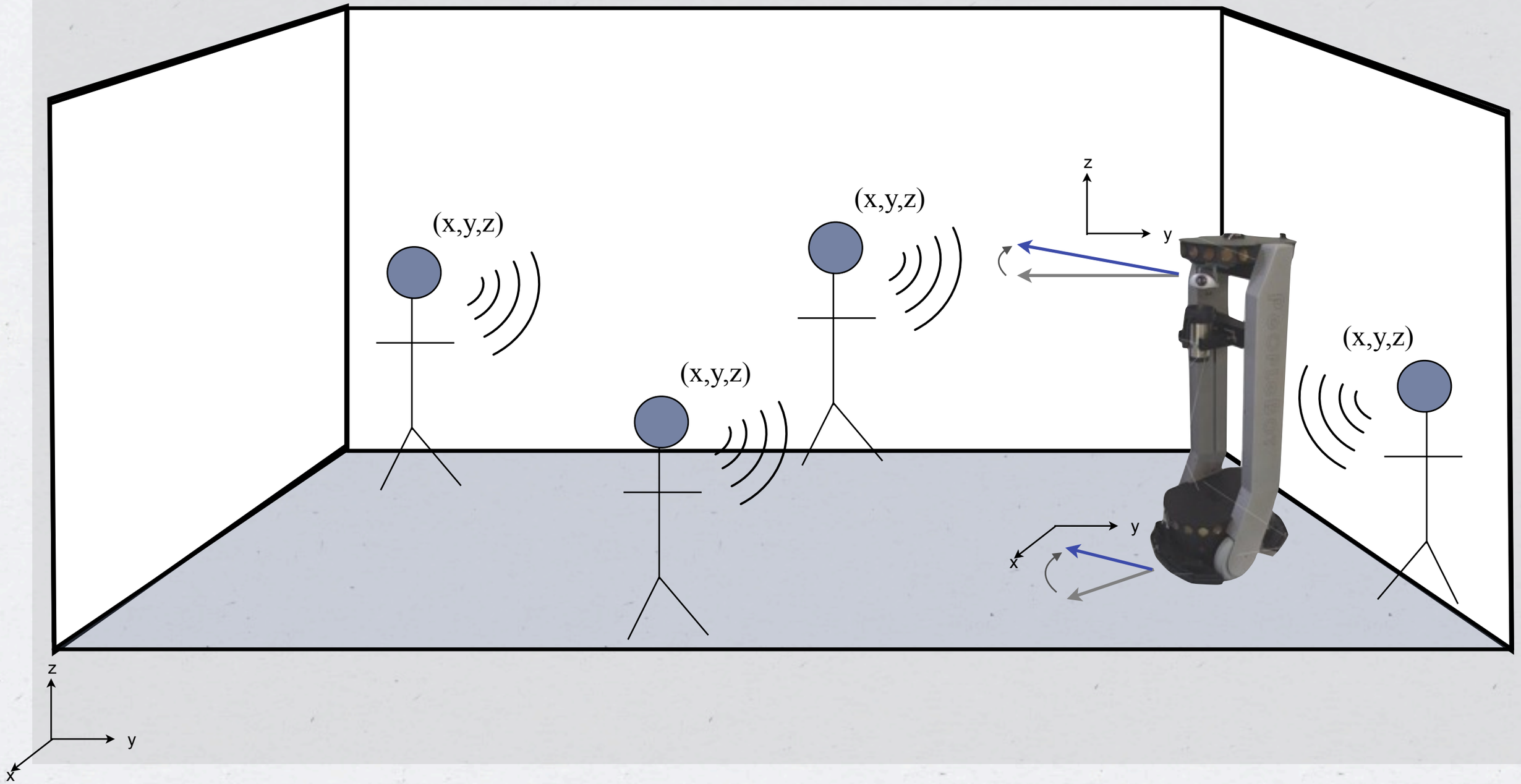
Outline

* Why Multi-DOA (or Single-DOA, for that matter)?
* Challenges in a Mobile Robotic Platform
* Proposed Algorithm
* Evaluation
* Conclusions

Direction of Arrival (DOA)



Multiple Directions of Arrival (Multi-DOA)



Motivation

* **From the user point-of-view:**

* ‘Facing’ the user enhances the “naturality” of the conversation.

    * The users feels as though the robot is “putting attention”.

Motivation

- **From the point of view of the robot (and its developers):**
  - Pointing a directional microphone or using directional noise cancellation can enhance ASR.
  - It removes the limitation of the camera's visual range when employed for face detection/recognition.

Doing it with a Robot

* **Limitations:**

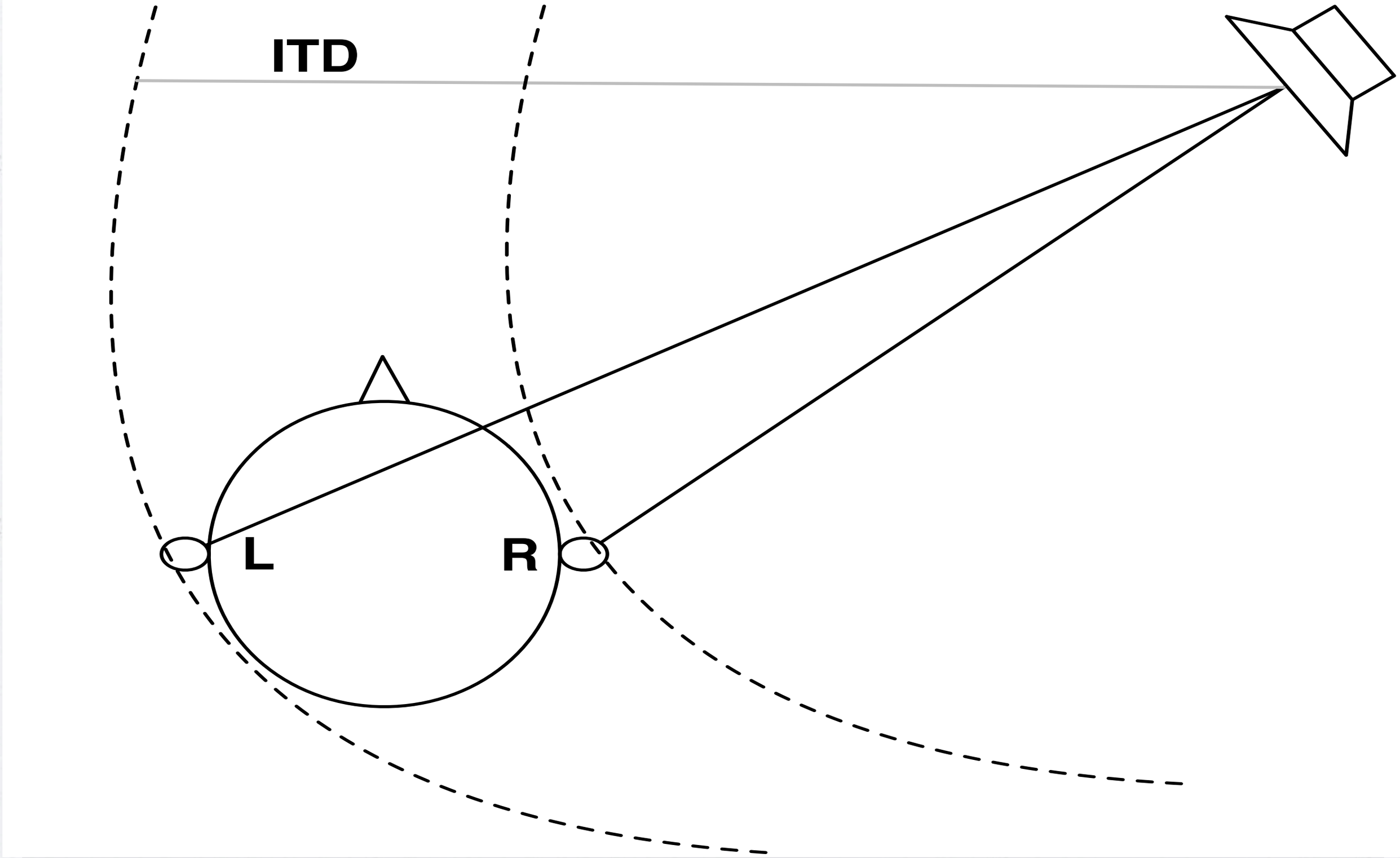
* The robot needs to be able to carry the audio hardware setup.
* Navigation should not be affected.
* Microphone positioning should not hinder the robot's appearance.
    * It is directly correlated to the robot's "usability" by the user.

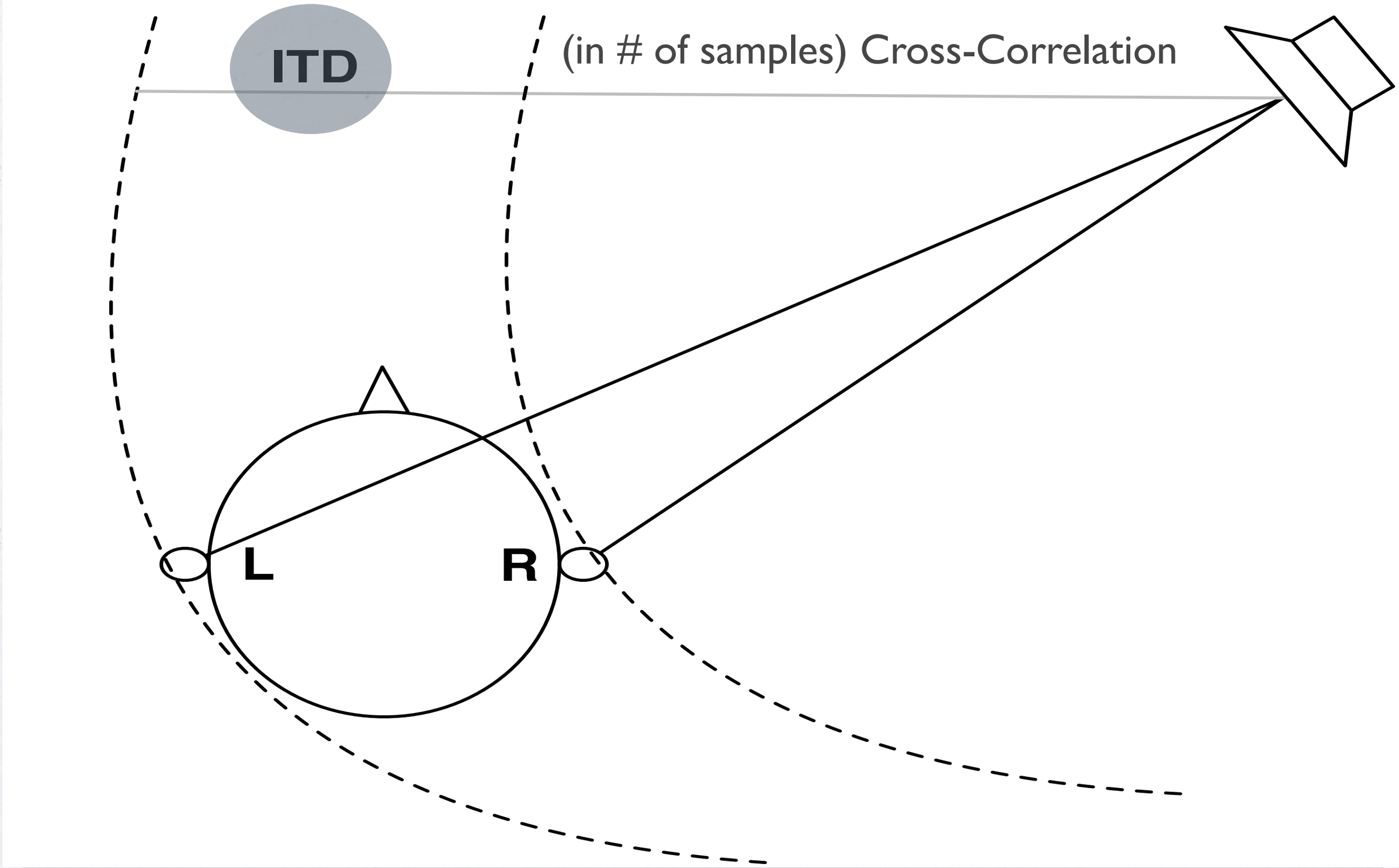
Doing it with a Robot

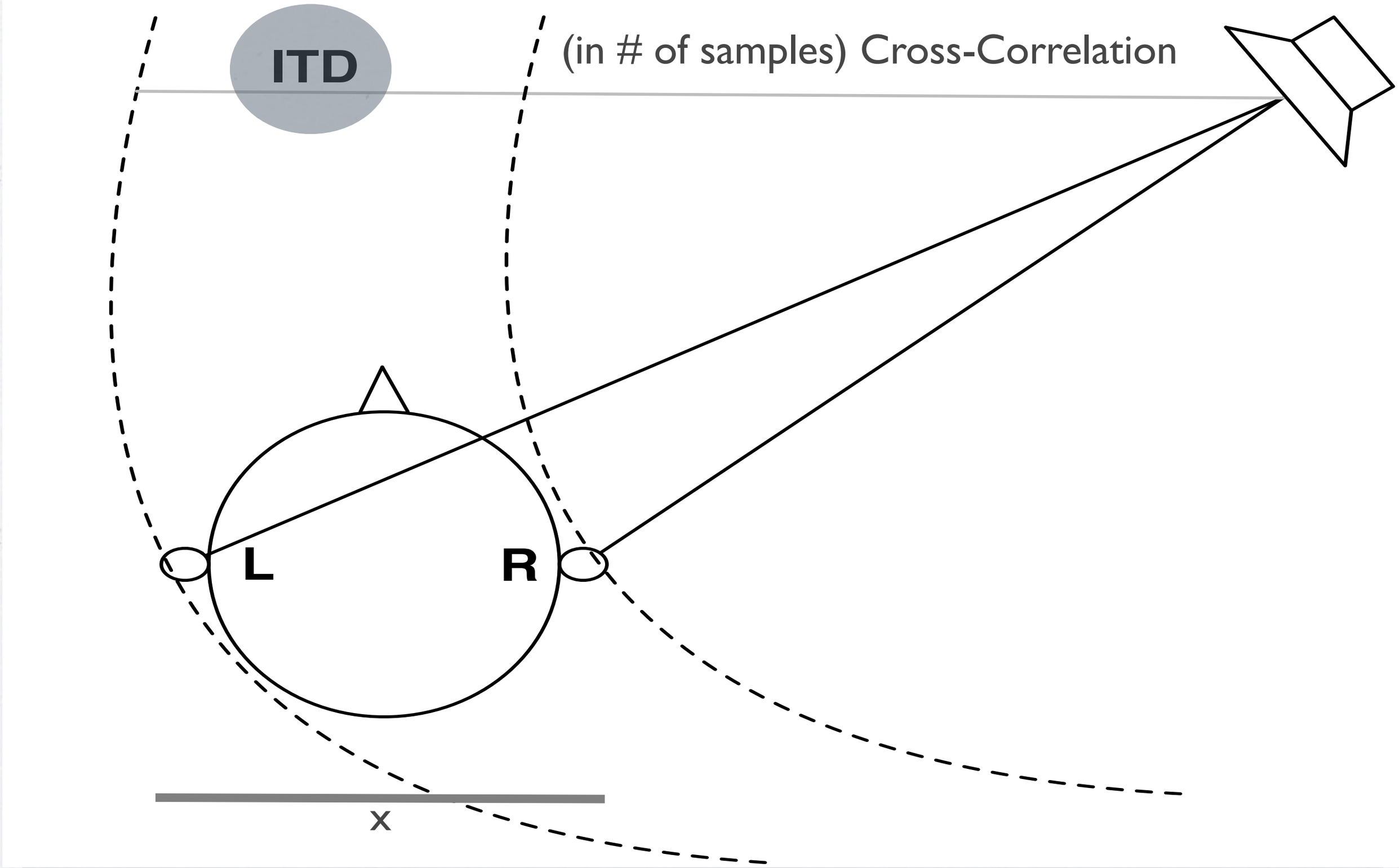
* **Requirements:**
  * The amount of users and their location are unknown and can change throughout.
  * The “curse” of the mobile.
    * Background noise and room characteristics are unknown and can change throughout.

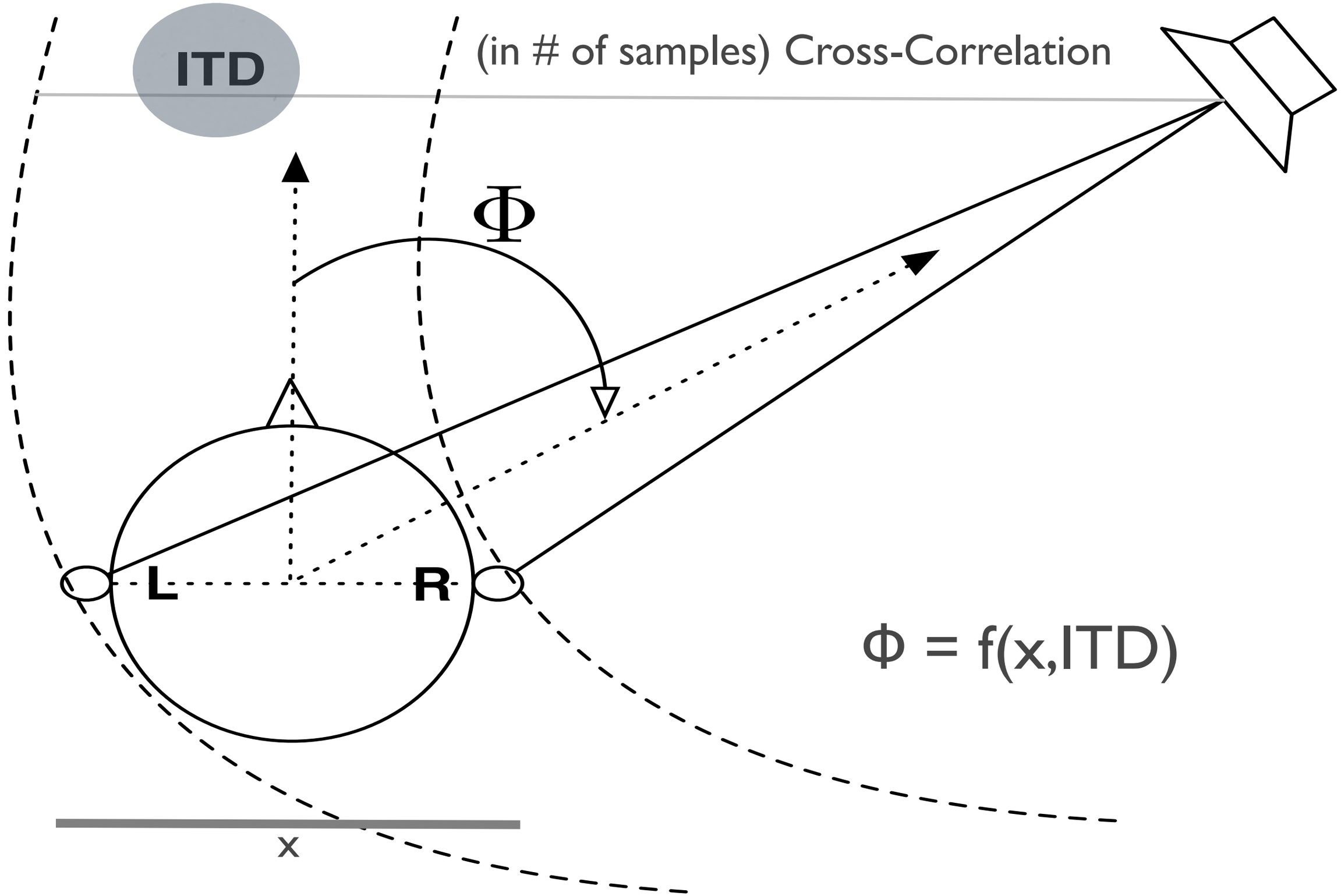
The Basics

* A popular feature used for estimating the DOA of a source is the *Inter-Aural Time Difference* (ITD).
    * The amount of time it takes a signal to reach one microphone once it reached another.


ITD
L
R


ITD
(in # of samples) Cross-Correlation
L
R


ITD
(in # of samples) Cross-Correlation
L
R
x


ITD
(in # of samples) Cross-Correlation
Φ
L
R
x
Φ = f(x,ITD)

The Basics

* Microphones can be set in 1-, 2- or 3-dimensional arrays for DOA estimation.
* Each have their pros and cons.
* There's a **big** con in most of them…

ITD Calculation and Reverberation and Noise

* Usually based on Cross-Correlation, calculating the ITD is prone to have errors when in presence of reverberation and ambient noise.

* However, this can be solved by adding **redundancy** measures.

    * One of the main reasons to *have lots of microphones*.

D. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, Algorithms, and Applications.* IEEE Press/Wiley- Interscience, 2006.

So, the More Microphones, the Better?

* With many microphones, several concurrent ITD's can be calculated and be compared to each other: **redundancy**.

* And, current Multi-DOA estimators (e.g. MUSIC), welcomes many-microphone arrays:

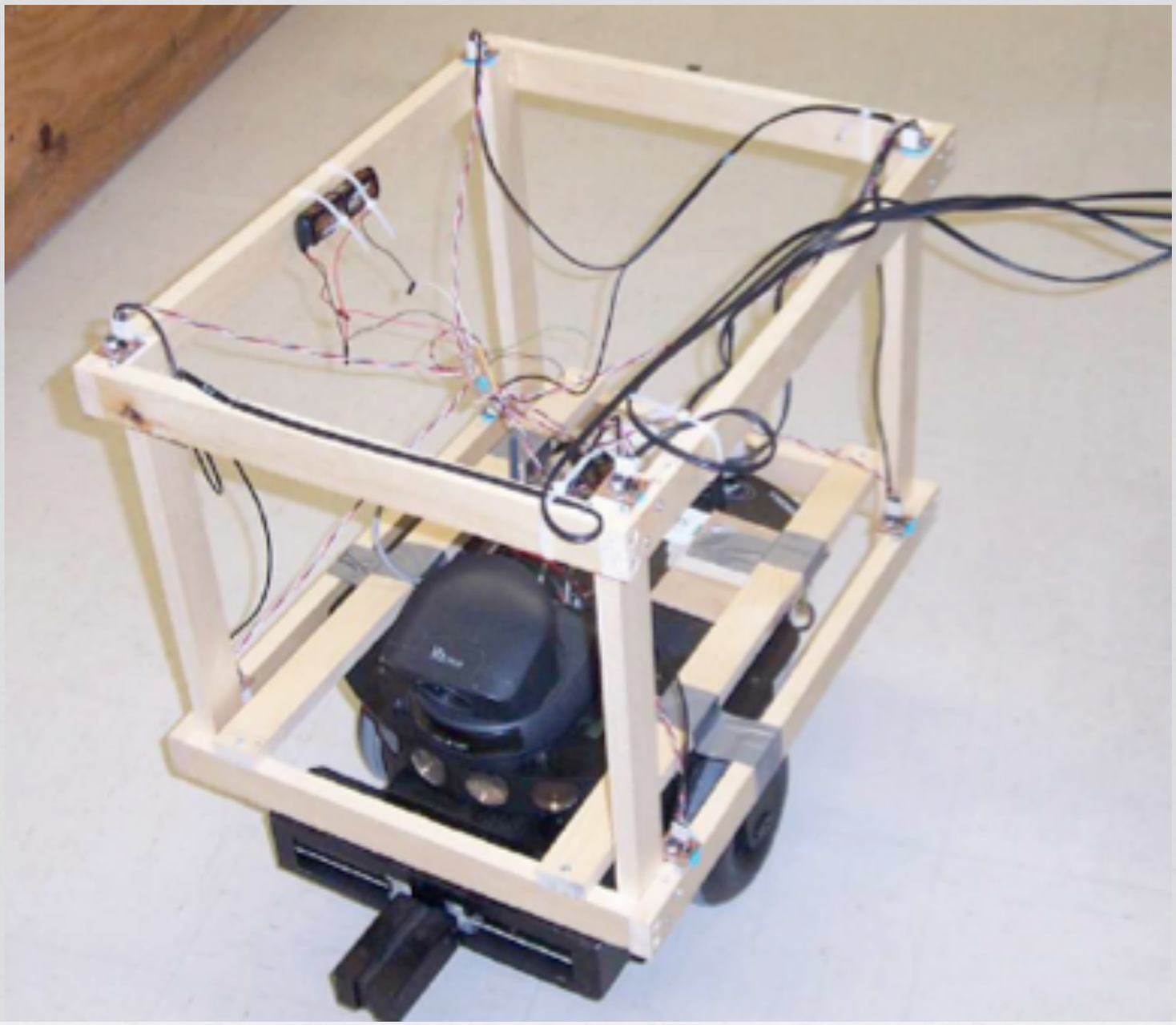
  * More microphones, more concurrent DOA's it can estimate.

  * Number of DOA's = Number of microphones - 1

* However...

R. Schmidt, "Multiple emitter location and signal parameter estimation," Antennas and Propagation, IEEE Transactions on, vol. 34, no. 3, pp. 276–280, 1986.

Space is a Luxury



J. Valin, J. Rouat, and F. Michaud, “Enhanced robot audition based on microphone array source separation with post-filter,” in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems,* 2004, pp. 2123– 2128.

What to do?

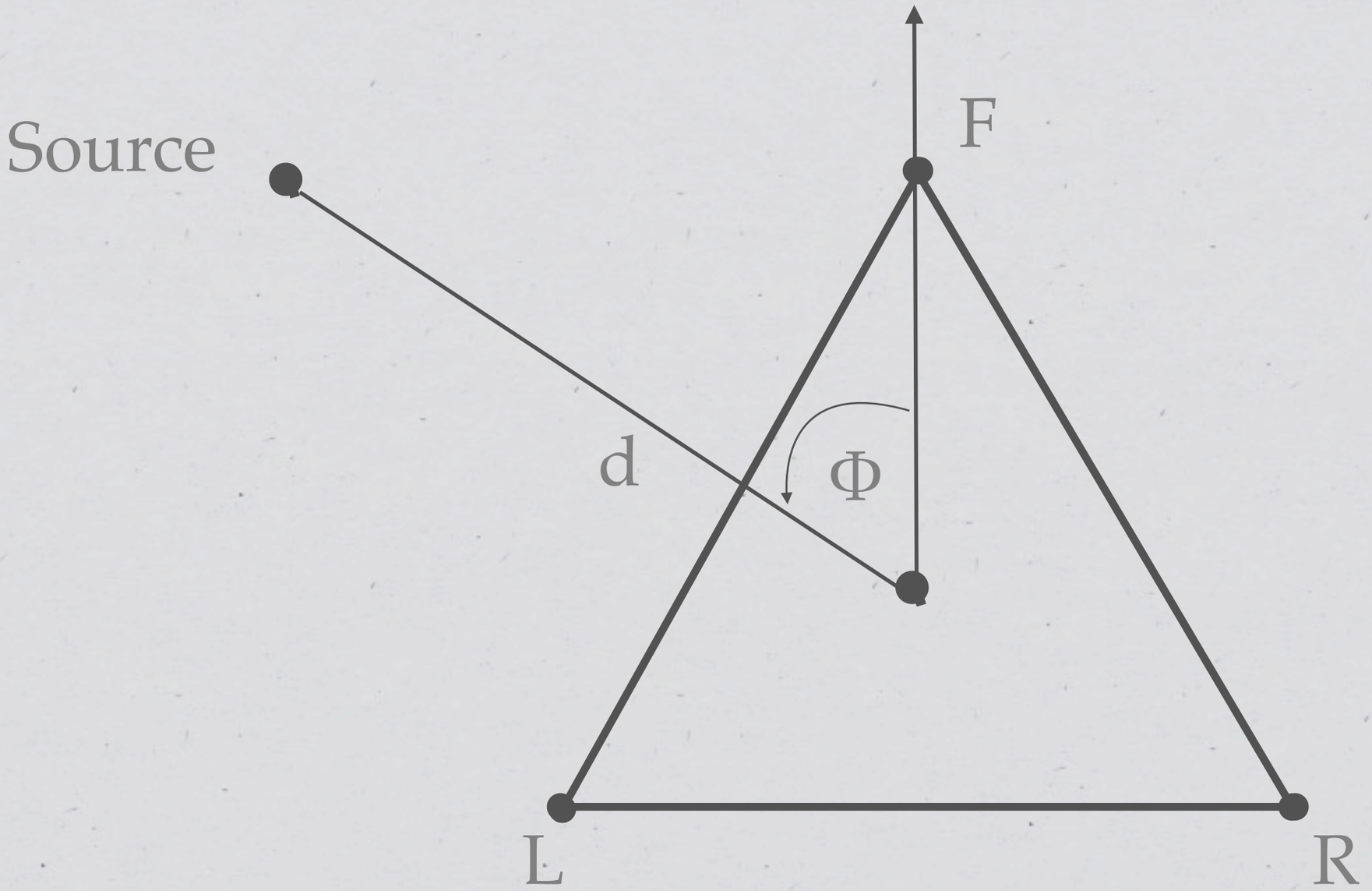
* Need to find balance between redundancy (many microphones) and mobility (few microphones).

* With only **two** microphones, there is little opportunity for redundancy.

* Well, lets go with **three** microphones…

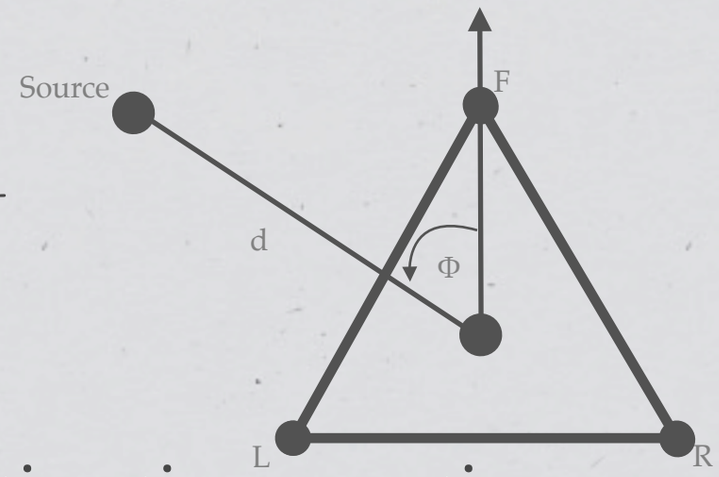
Proposed Arrangement

… from a while ago



C. Rascon, H. Aviles, and L. A. Pineda, “Robotic orientation towards speaker for human-robot interaction,” *Advances in Artificial Intelligence - IBERAMIA 2010,* vol. 6433, pp. 10–19, 2010.

Algorithm Summary



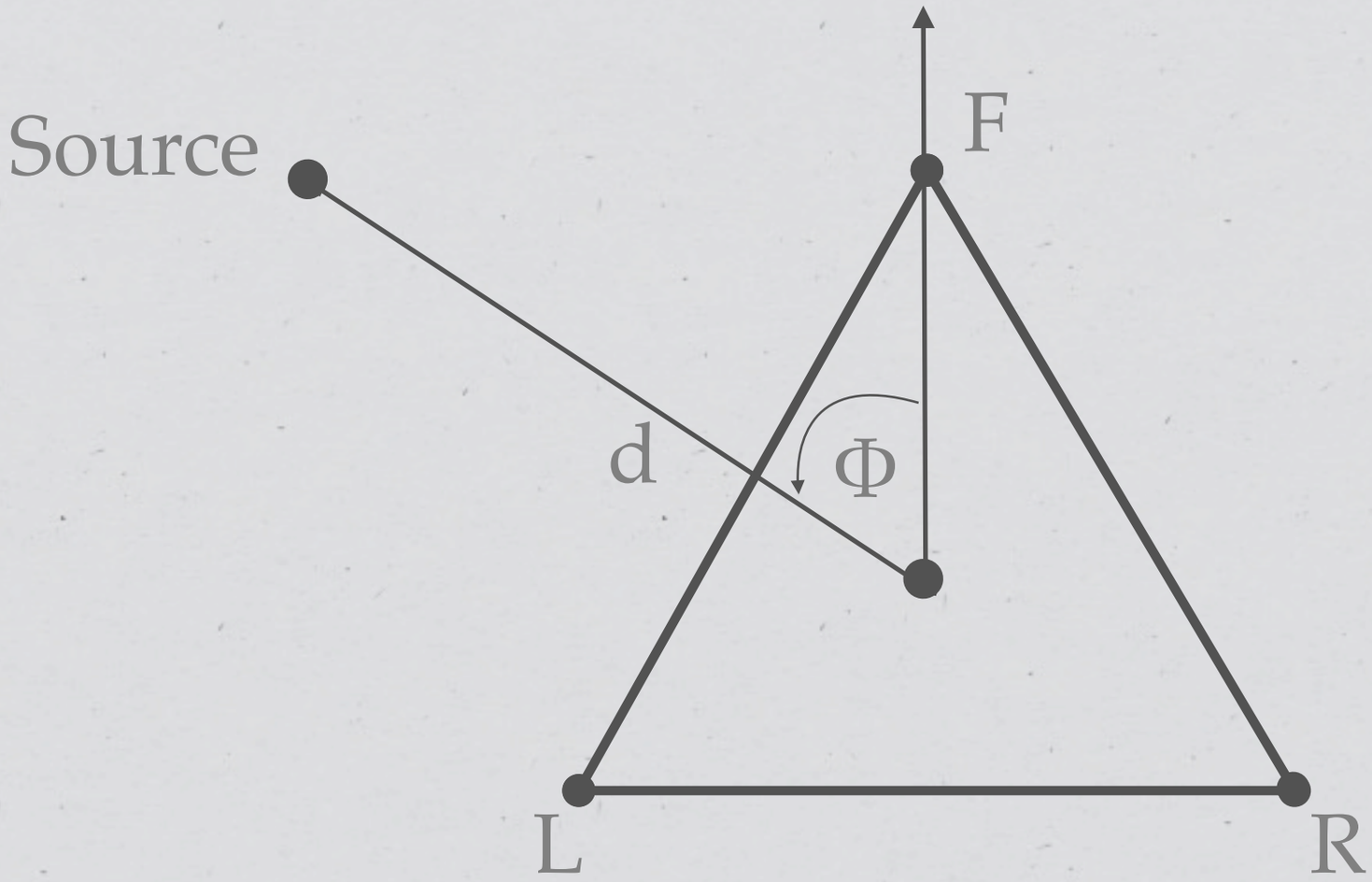
* Every microphone pair provides an ITD estimation, creating a set of 3 ITD's per sampling window.

* An *incoherence* value is obtained from the ITD set. It serves as a redundancy measure: highly incoherent ITD sets are ignored.

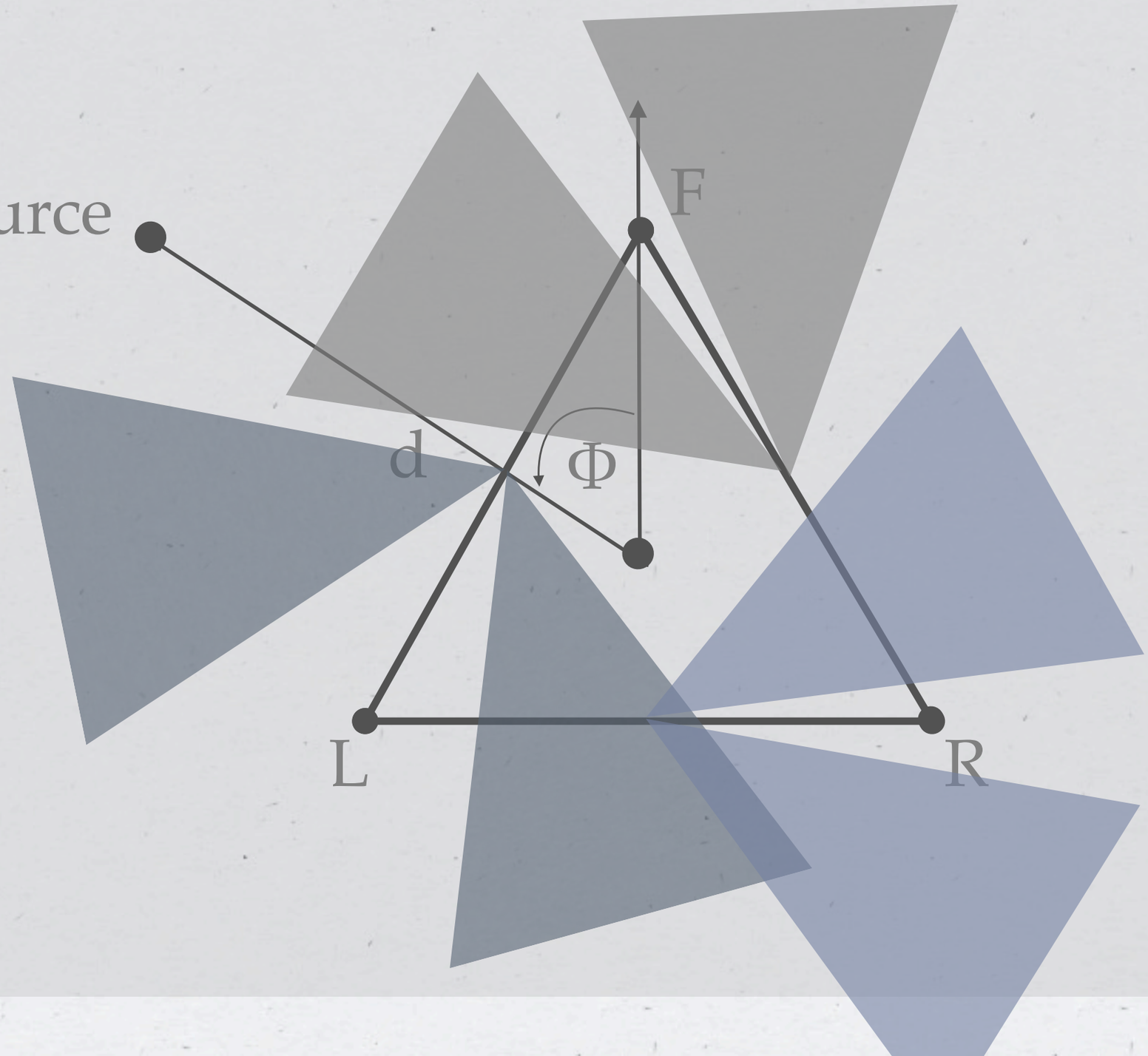
* If the incoherence value is low, a DOA is estimated with the ITD from the microphone pair that is most perpendicular to the source.

    * Forcing the DOA to be estimated with an ITD value in the -30° -- 30° range (well within the -50° -- 50° linear range).

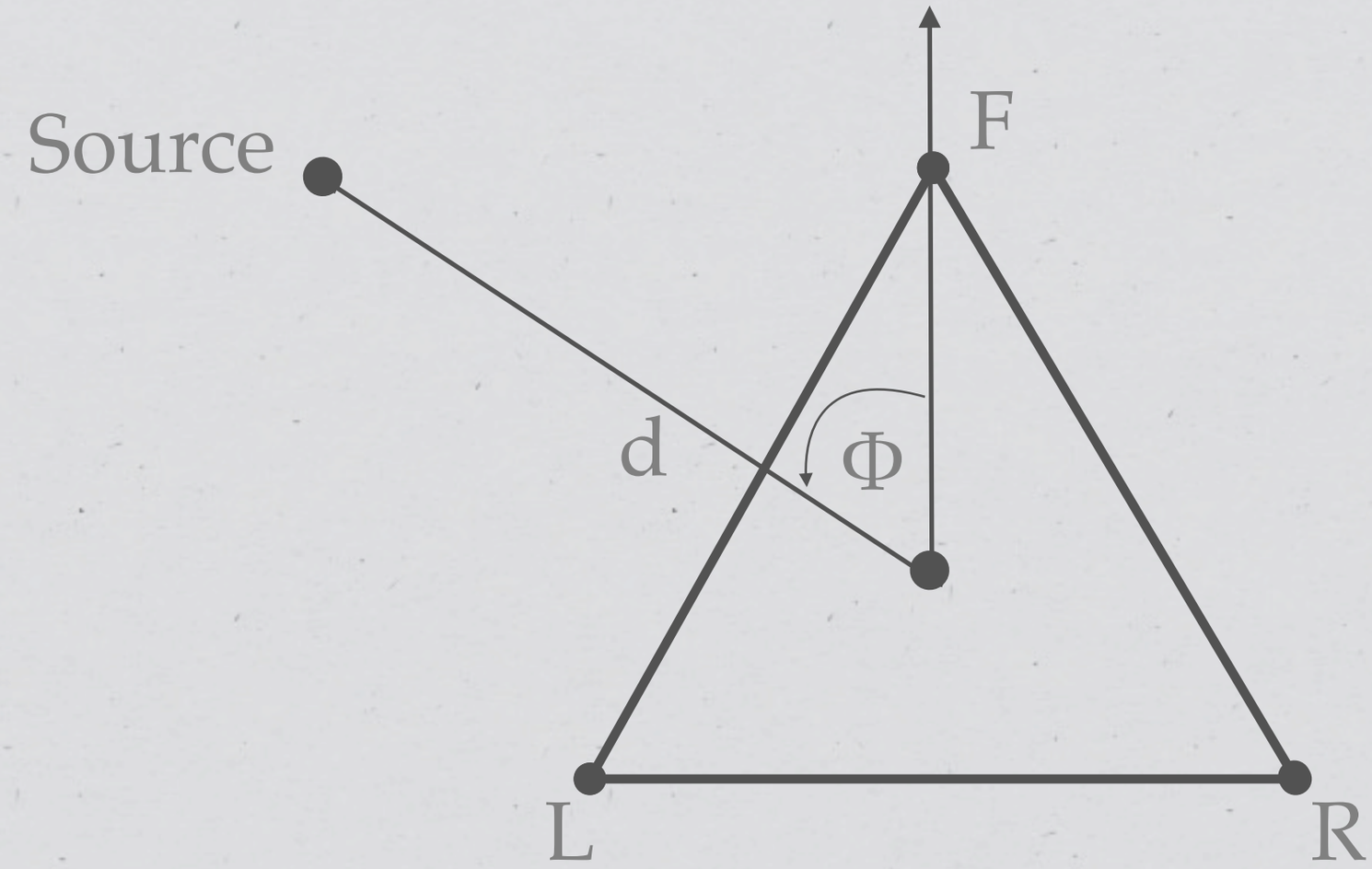
High-Incoherence ITD Set



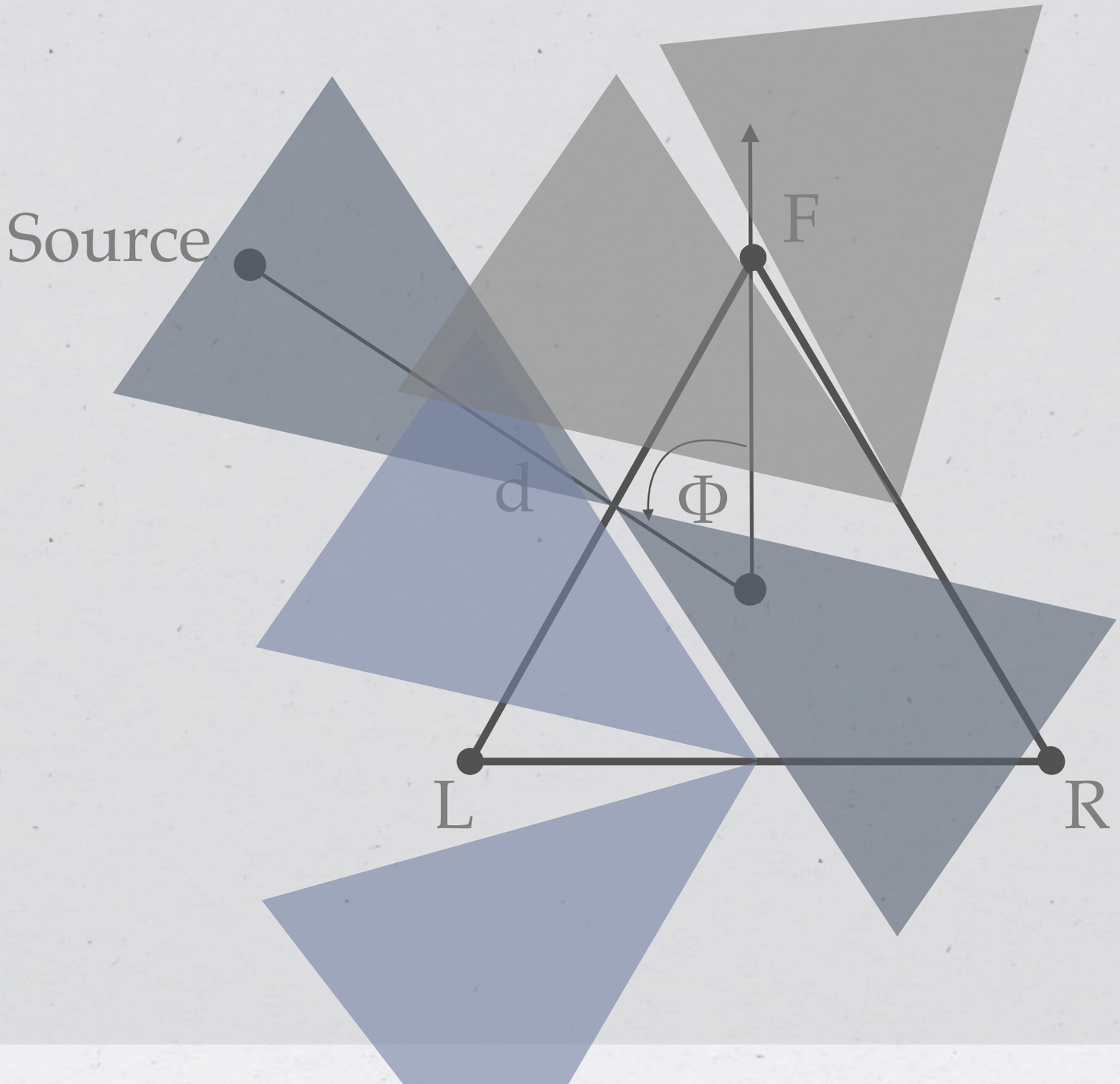
High-Incoherence ITD Set



Low-Incoherence ITD Set



Low-Incoherence ITD Set



Benefits

* Complete angle range [-180° -- 180°]
* Almost-linear ITD-to-DOA resolution throughout
* ITD estimation redundancy in every sample
* High confidence of the DOA estimation of **one source** in multiple-source environments.

ONE SOURCE?!

ISN'T IT "MULTI"?

That's our contribution in this paper.

Overlap between Human Speech

* It isn't 100%.
* In natural conversations, it doesn't even reach 10%.
* When forced to overlap via artificially superimposing pre-recorded sources, single-source windows of up to 500 ms have been observed.

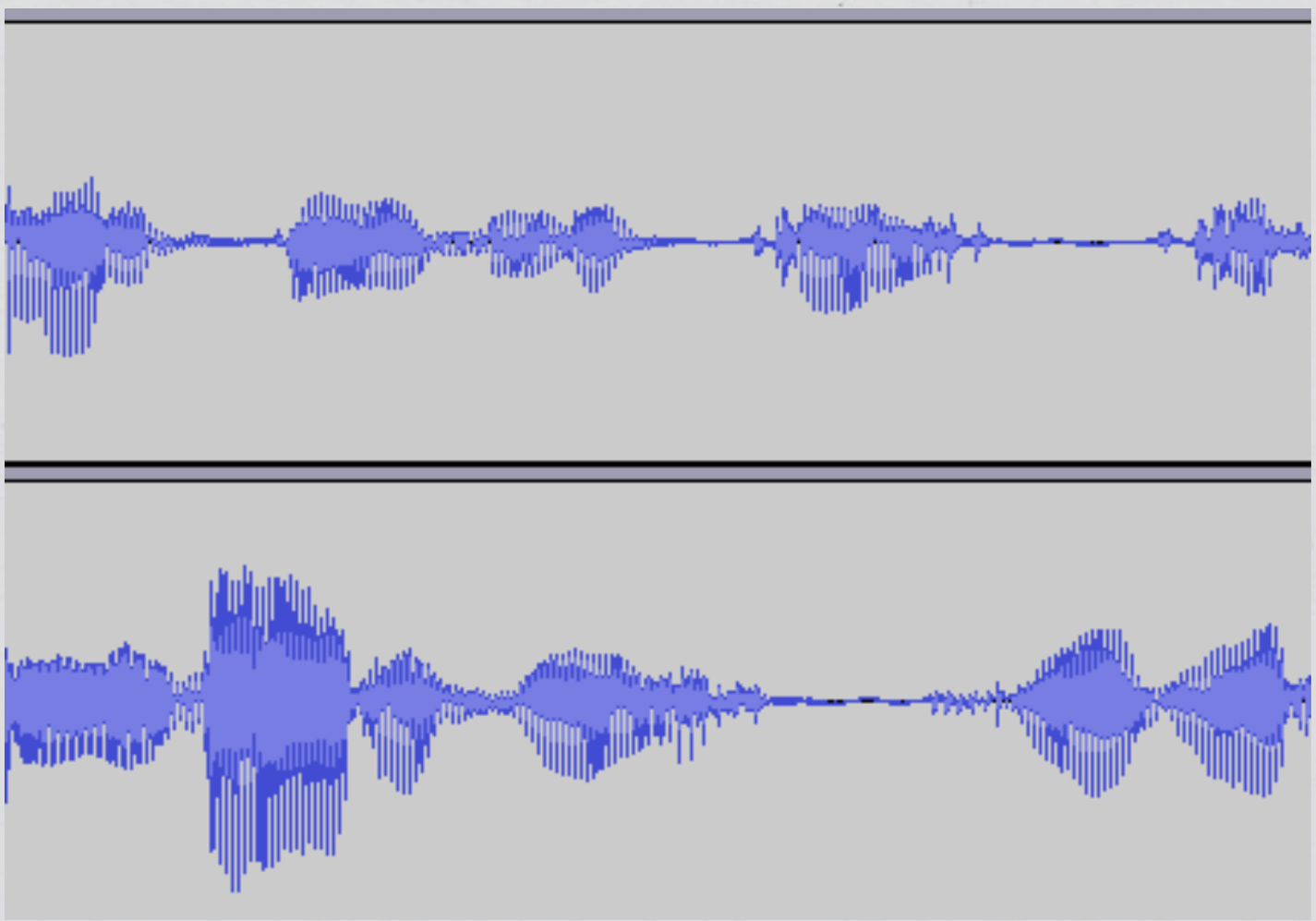
T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. de Ruiter, K. E. Yoon and S. C. Levinson, "Universals and Cultural Variation in Turn-Taking in Conversation," *Proceedings of the National Academy of Sciencies of the United States of America*, vol 106-26, pp 10587-10592, June 30, 2009.

E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in in Proceedings of Eurospeech 2001, 2001, pp. 1359– 1362.

Single DOA Estimator with Multiple Simultaneous Sources

User 1

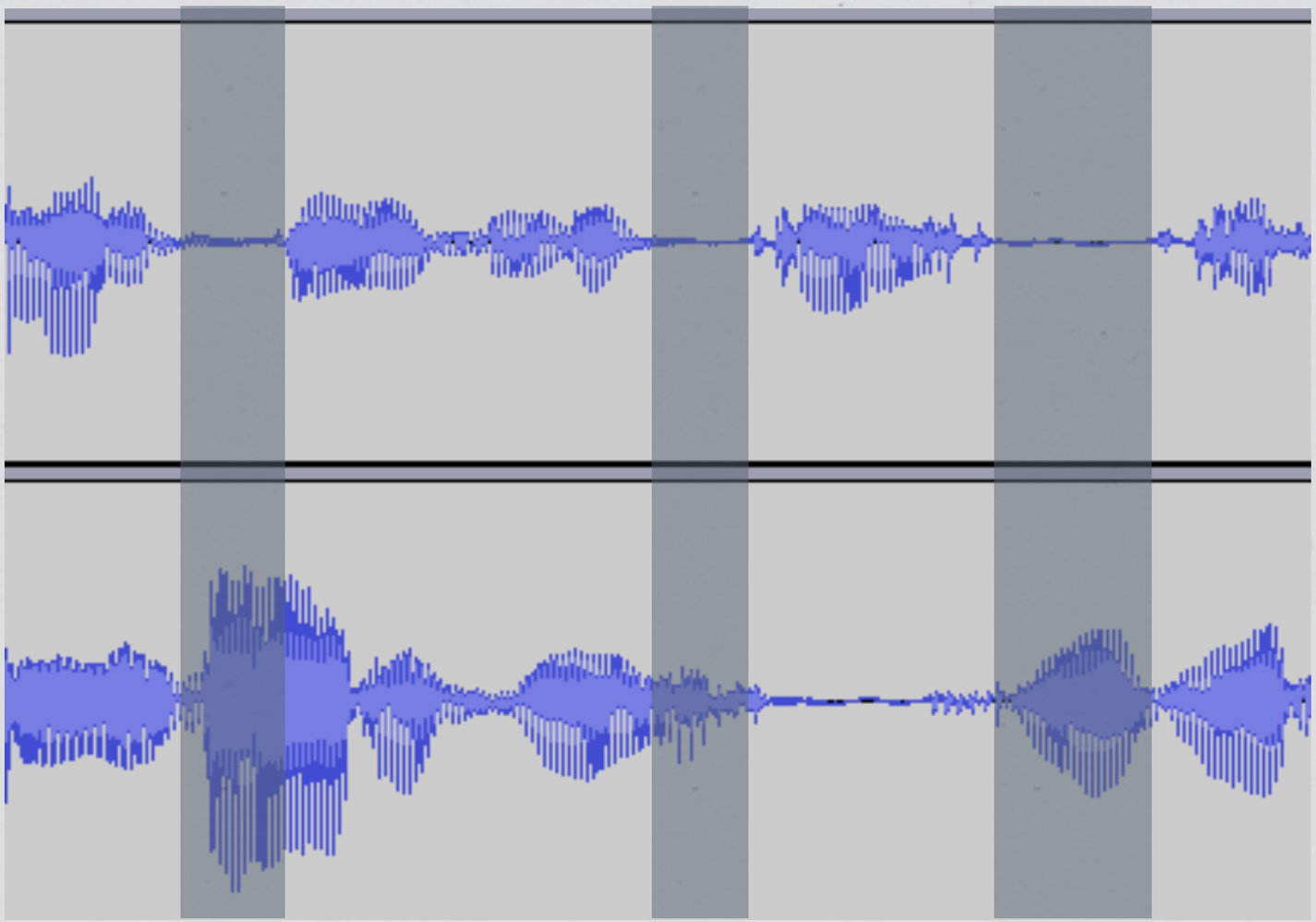
User 2



Single DOA Estimator with Multiple Simultaneous Sources

User 1

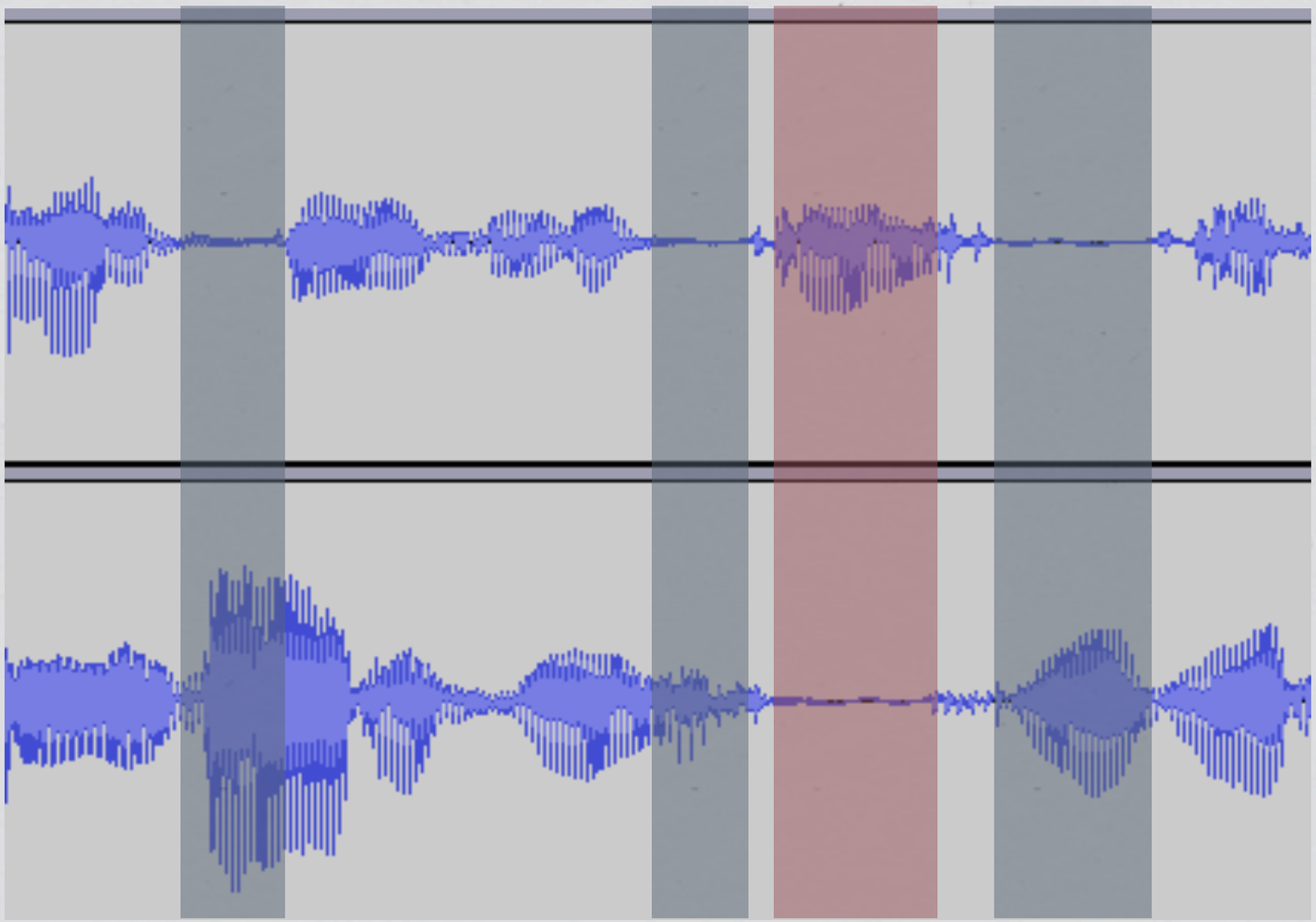
User 2



Single DOA Estimator with Multiple Simultaneous Sources

User 1

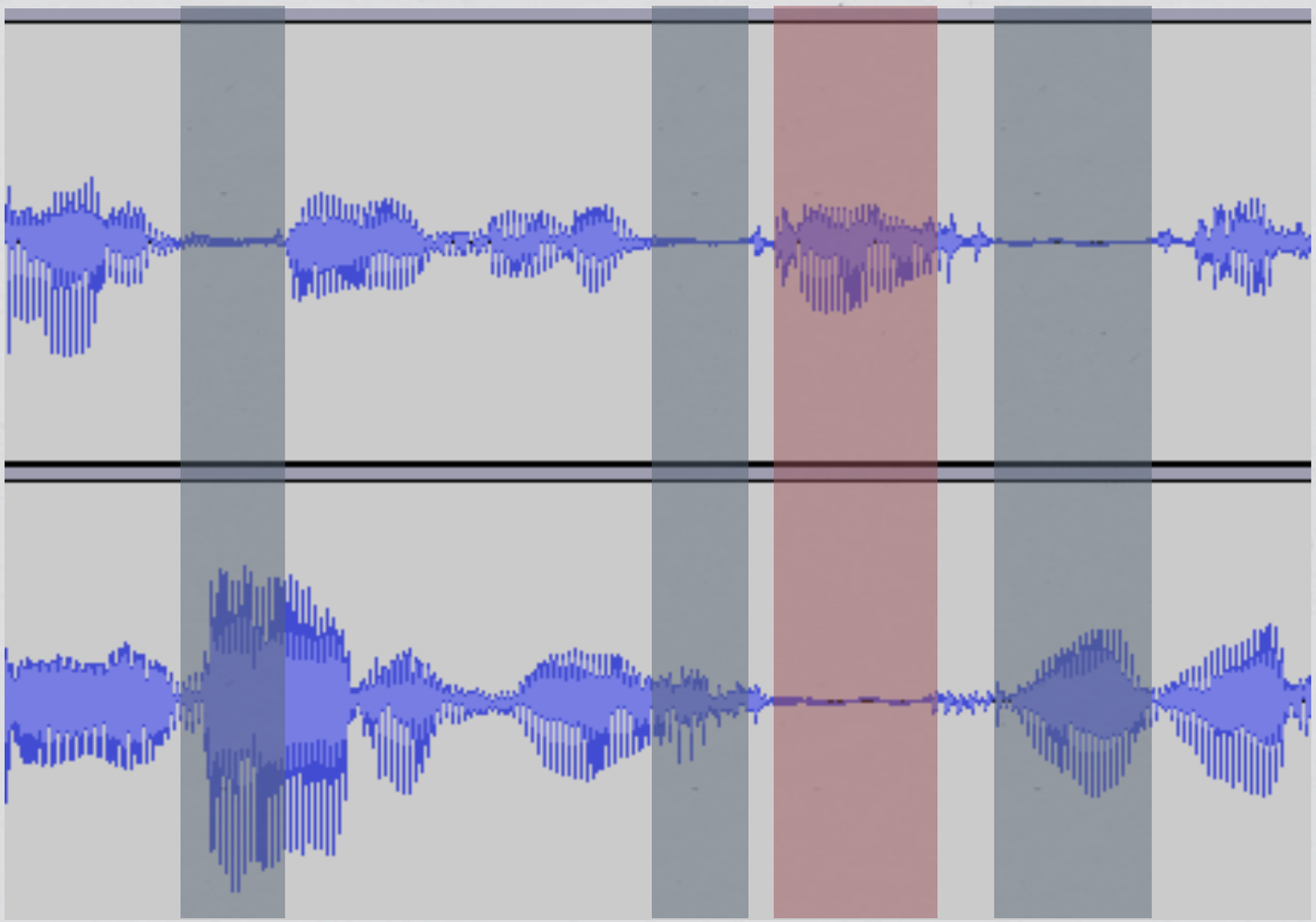
User 2



Single DOA Estimator with Multiple Simultaneous Sources

User 1

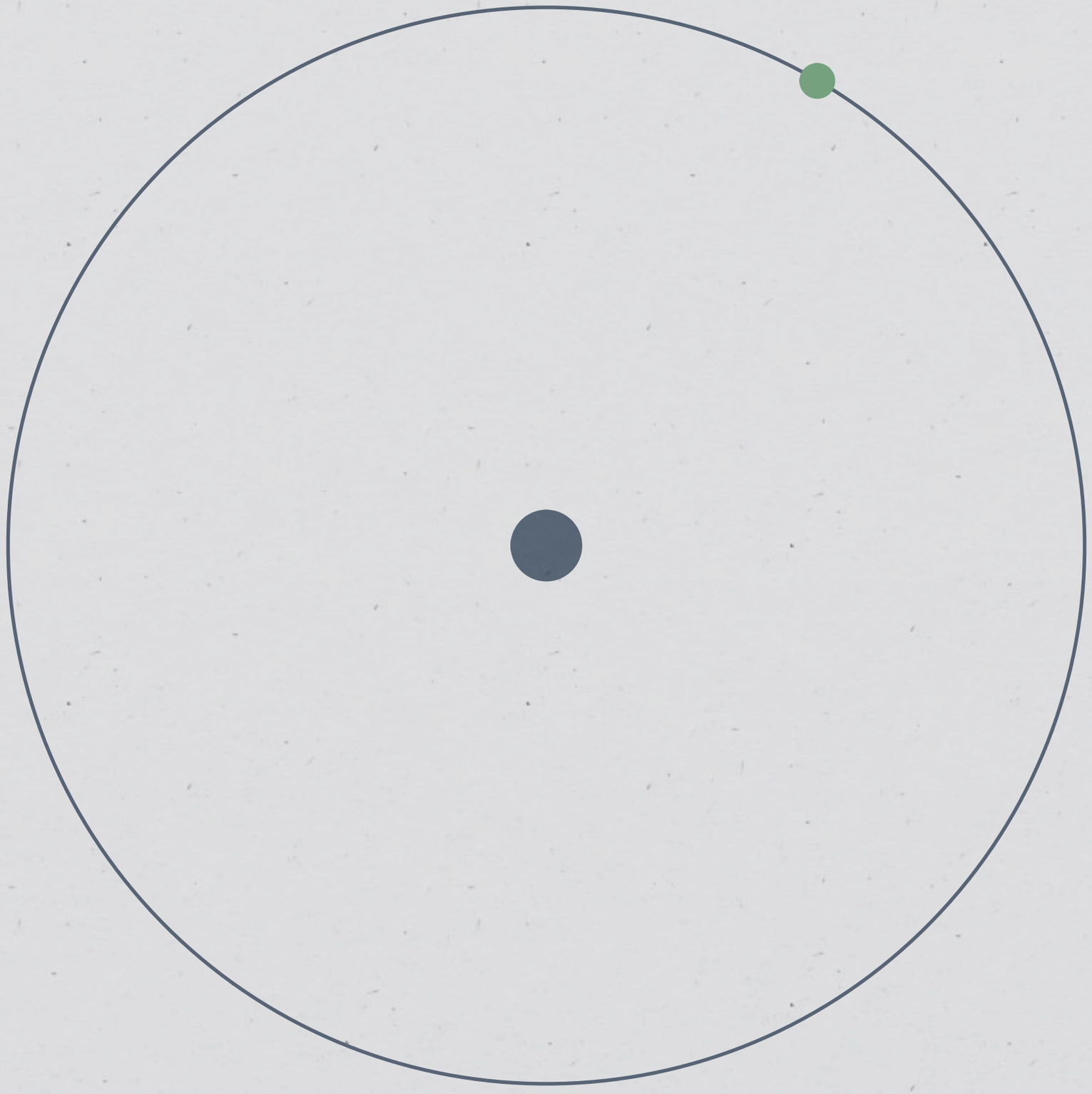
User 2

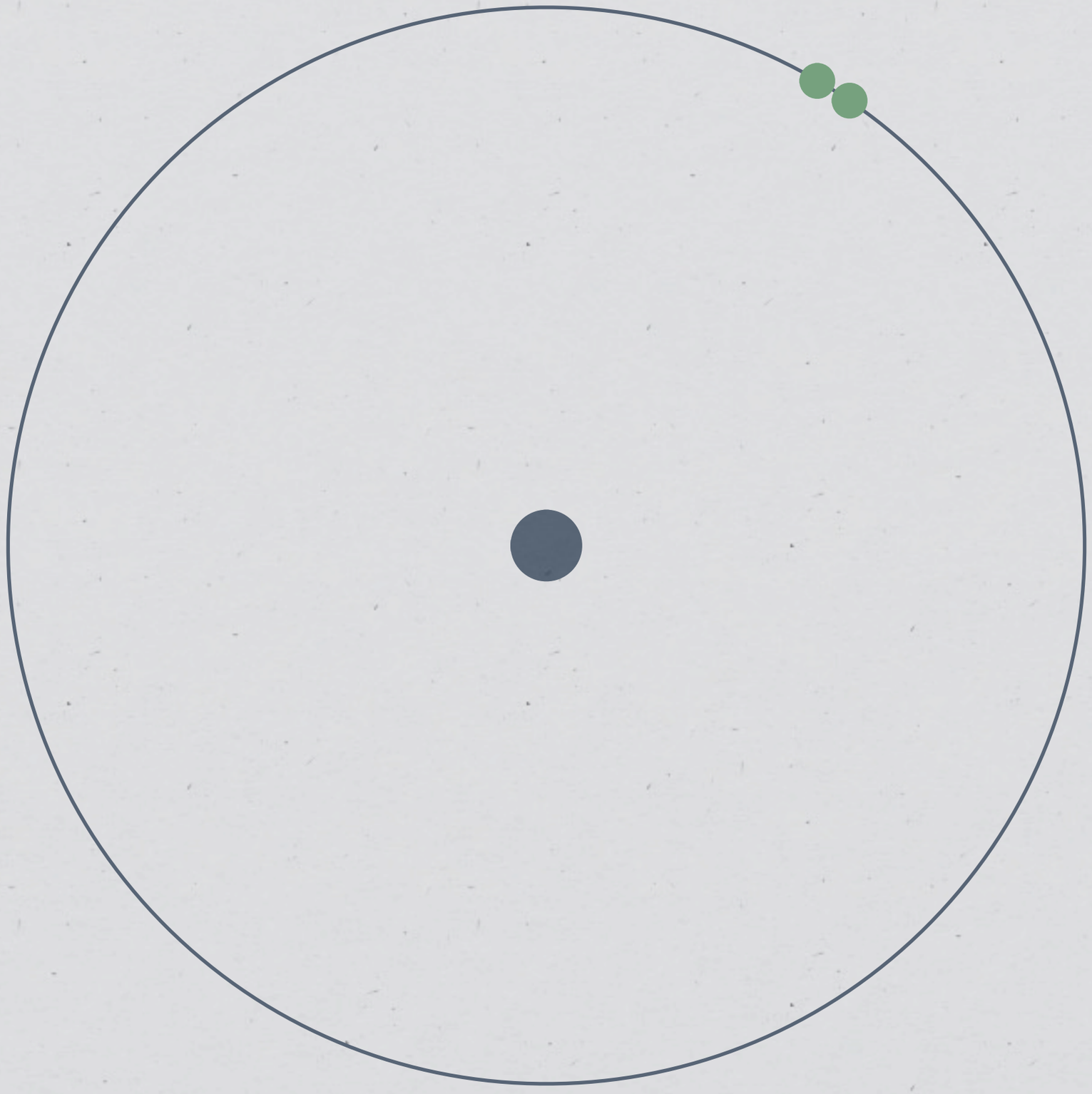


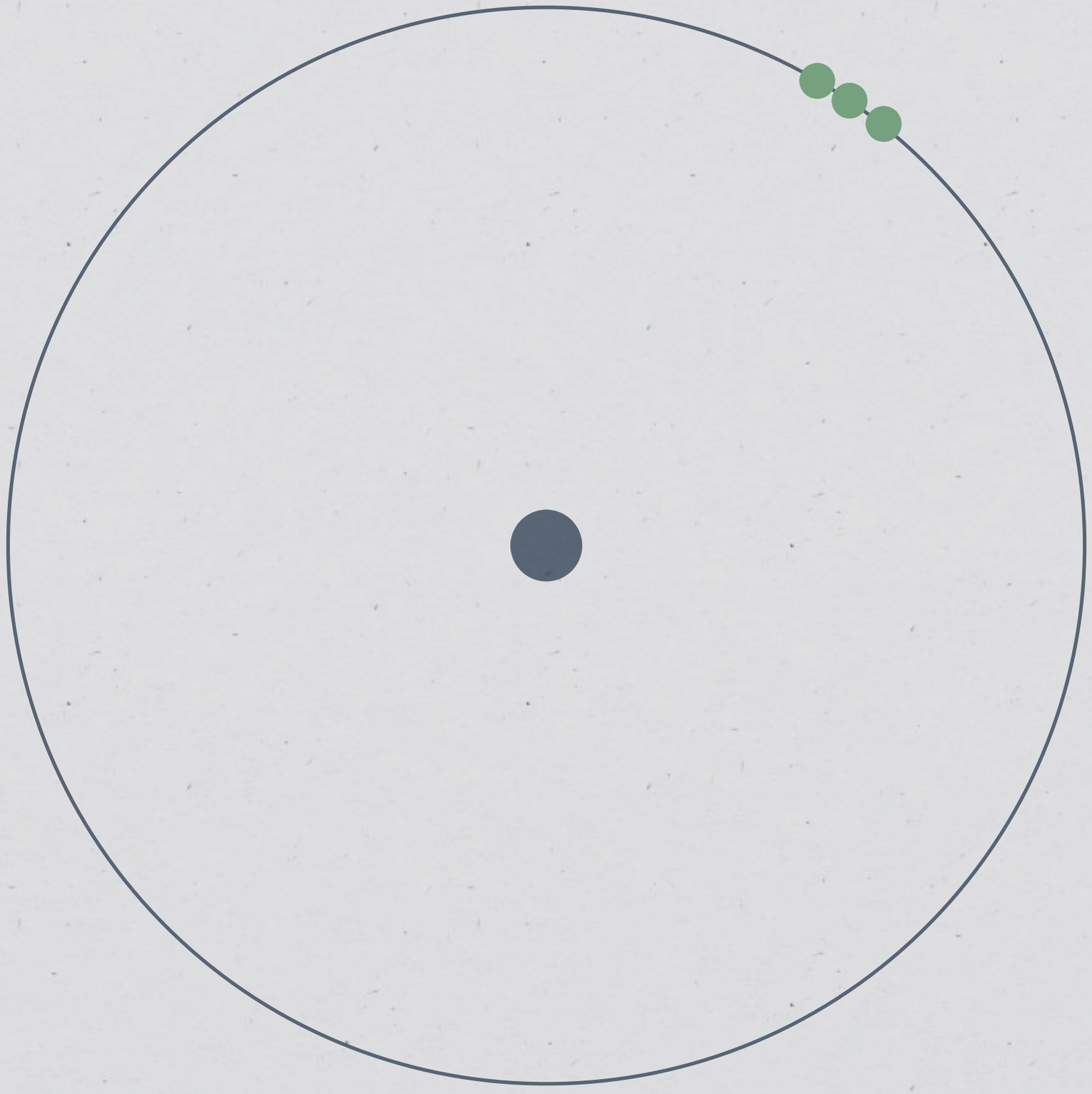
Occurrence of single-source windows is stochastic in nature.

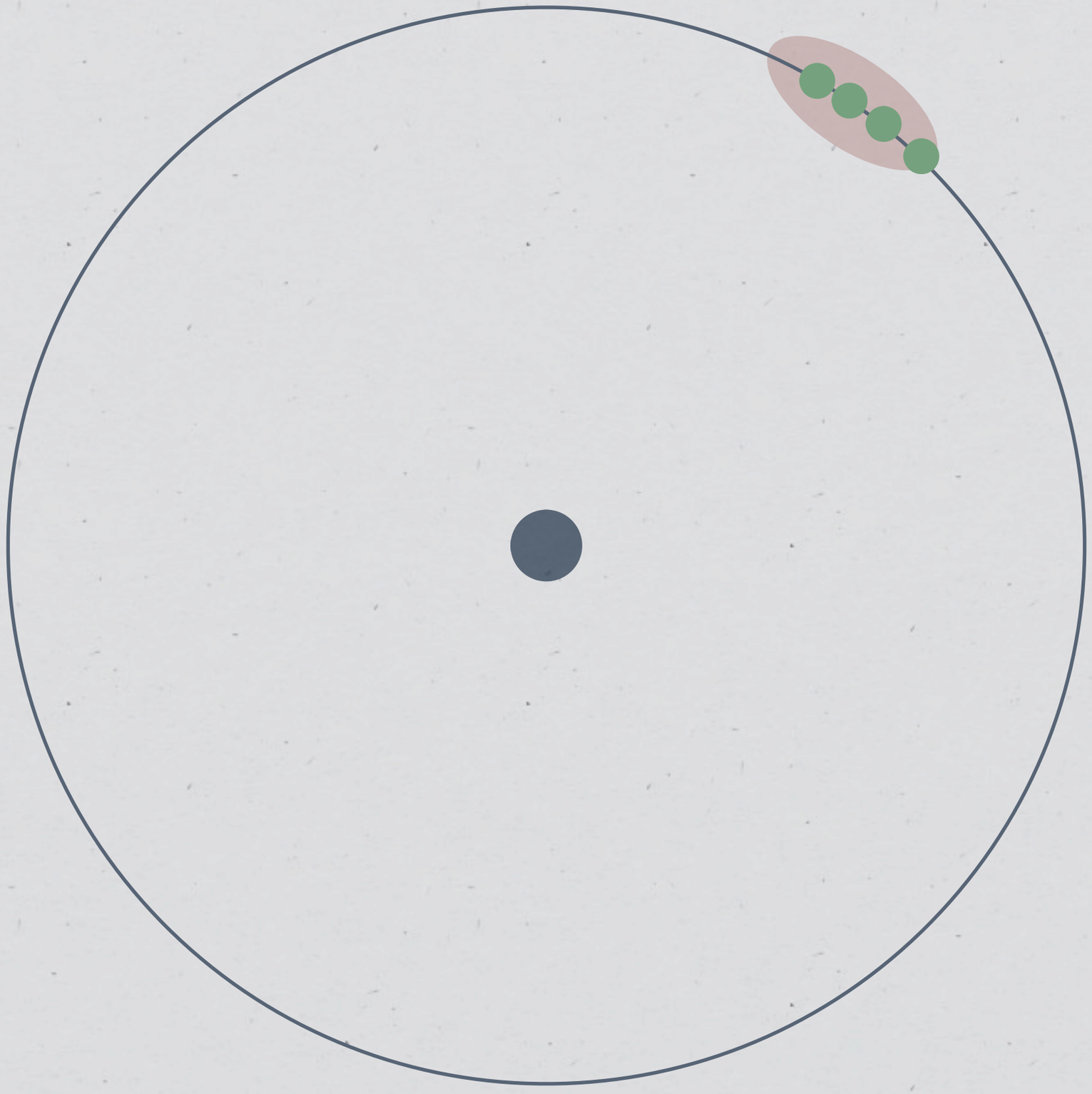
Multi-DOA Estimation: Tracking Problem

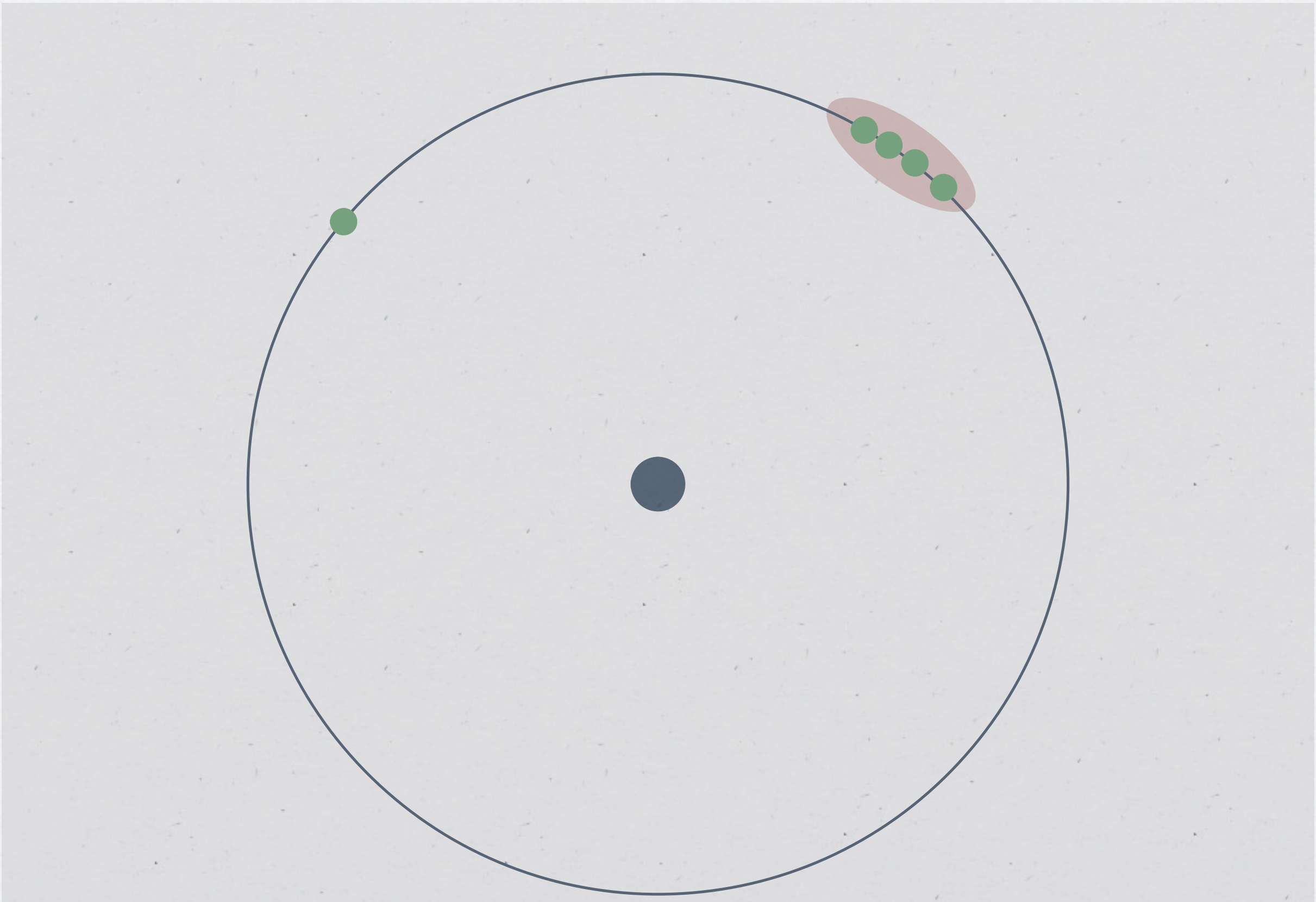
* The job of the tracker is to "gather" DOA's, provided by the Single DOA Estimator, into clusters.
* A DOA belongs in a cluster if it is close to its average DOA; if not, it is the beginning of a new cluster.
* A cluster becomes a **source** when it is composed by more than a pre-specified number of DOA's. The DOA of the source is then the average DOA of the cluster.
* "Old" DOA's are forgotten, which provides movement tracking.

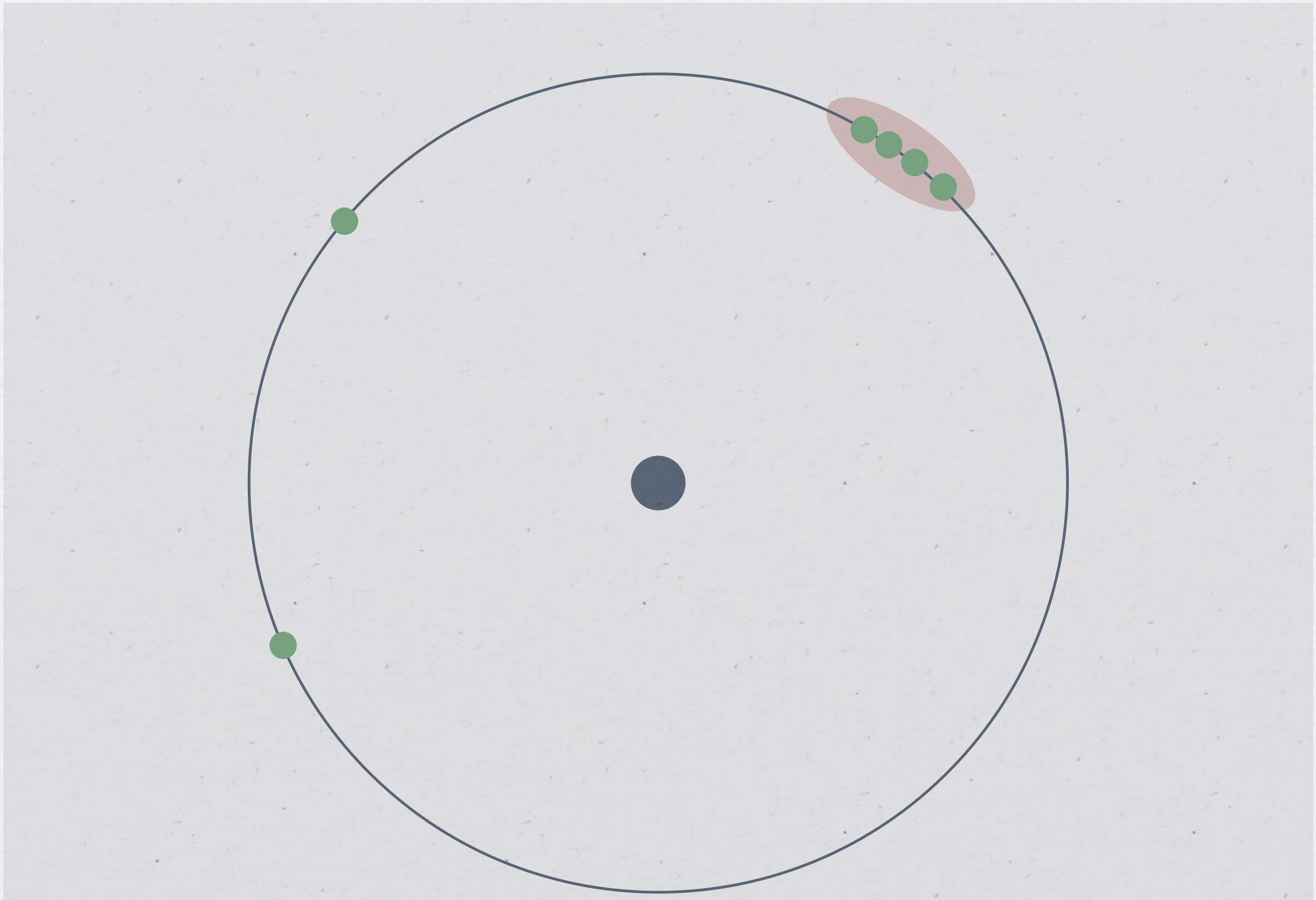


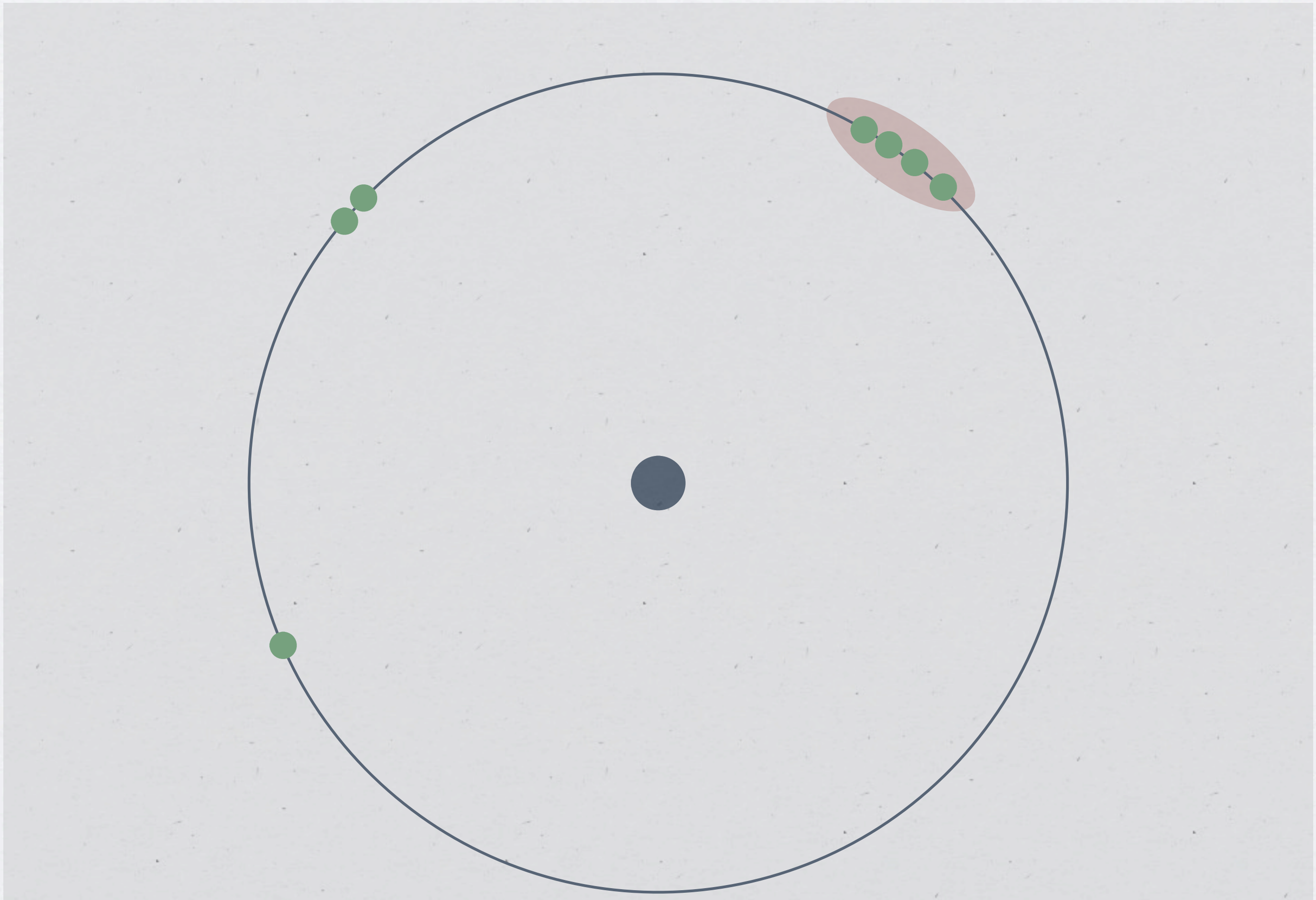


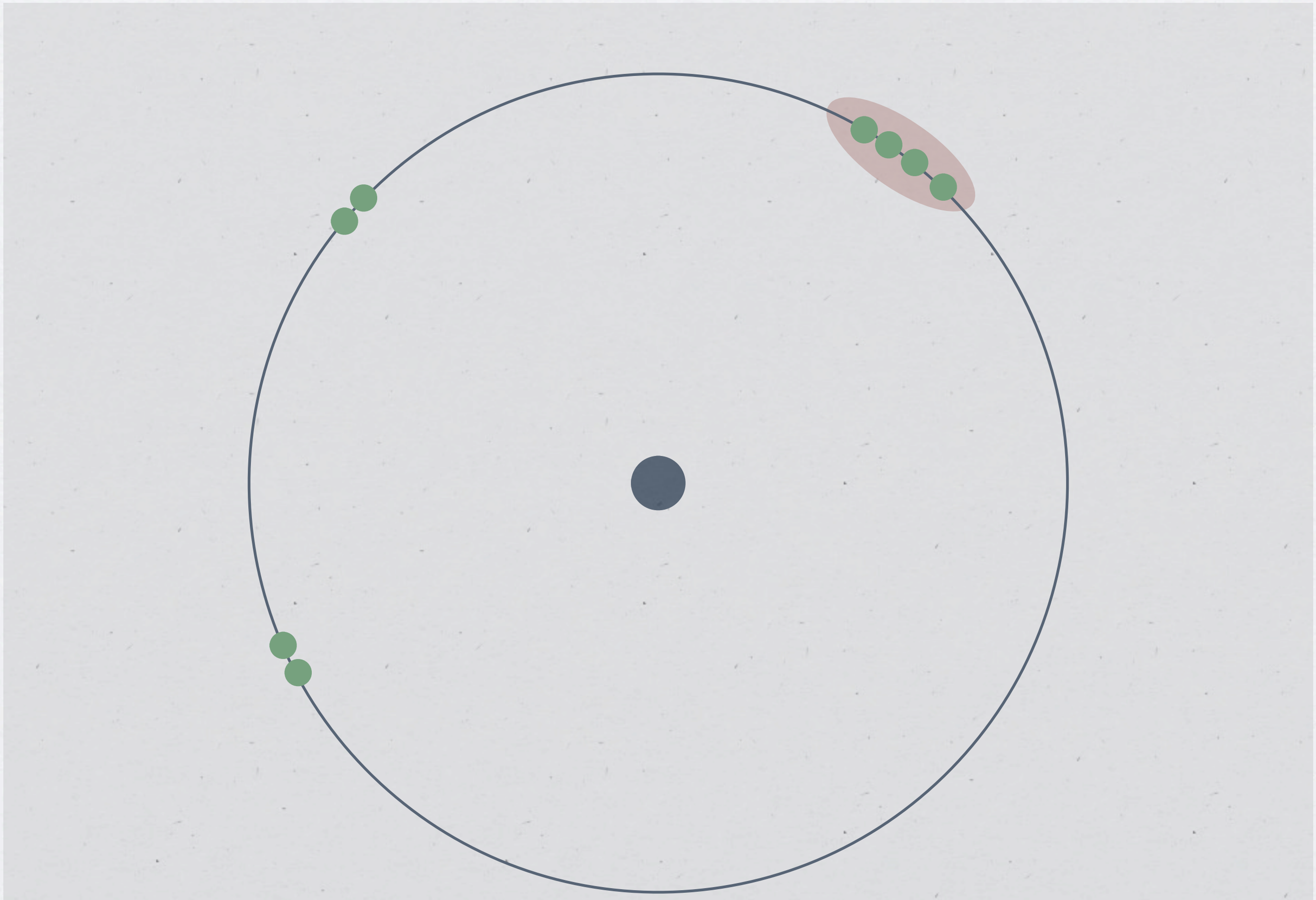


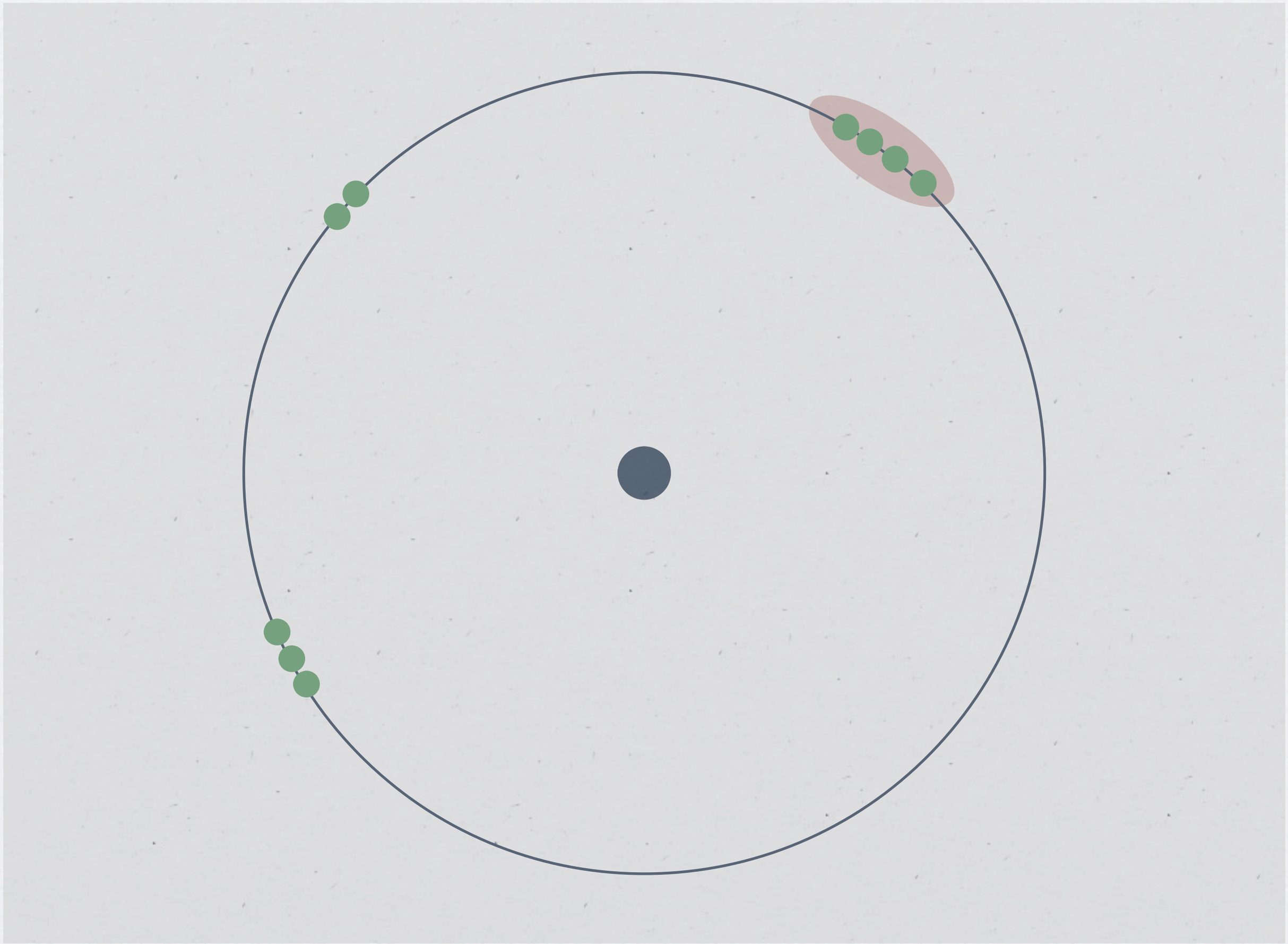


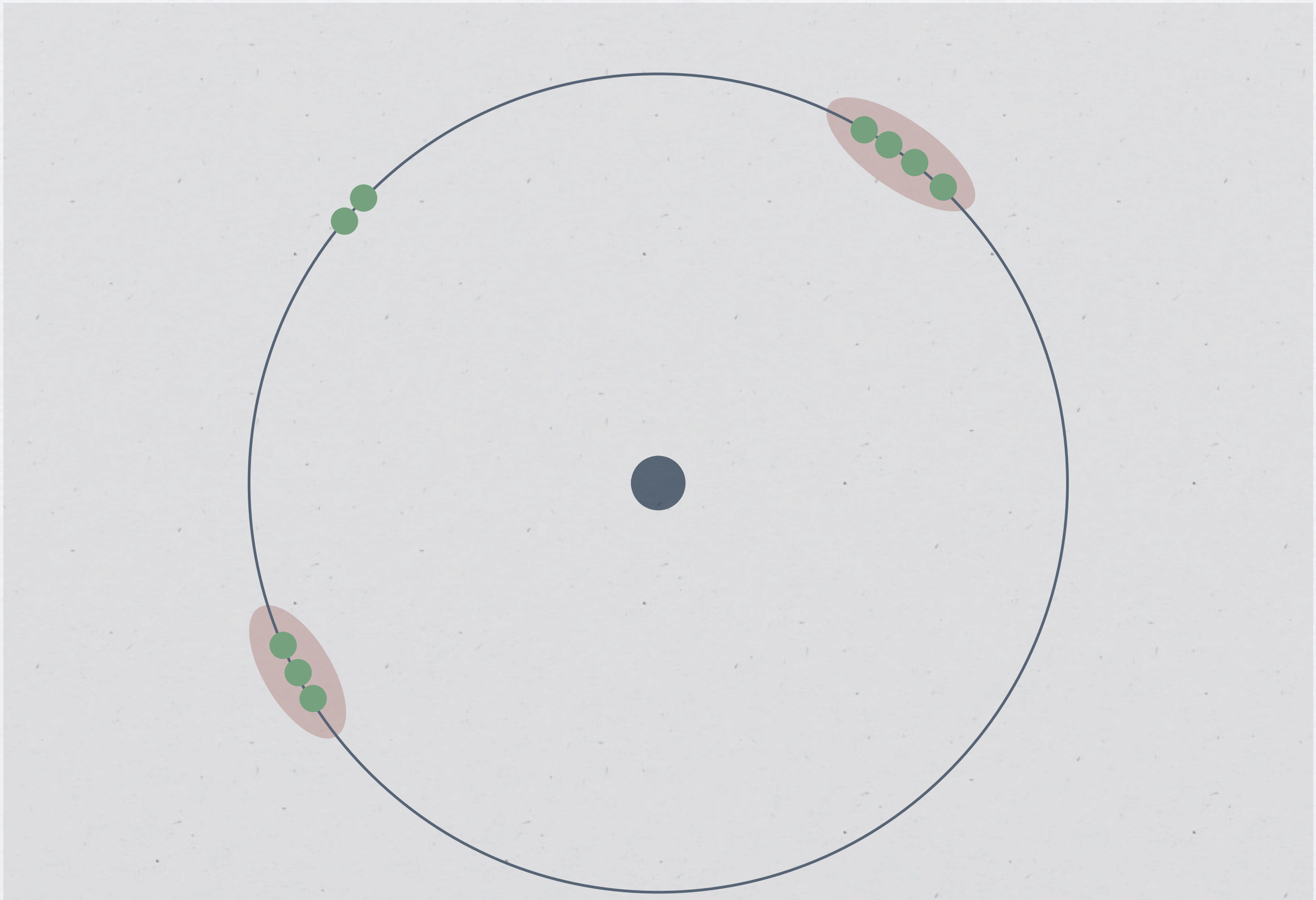


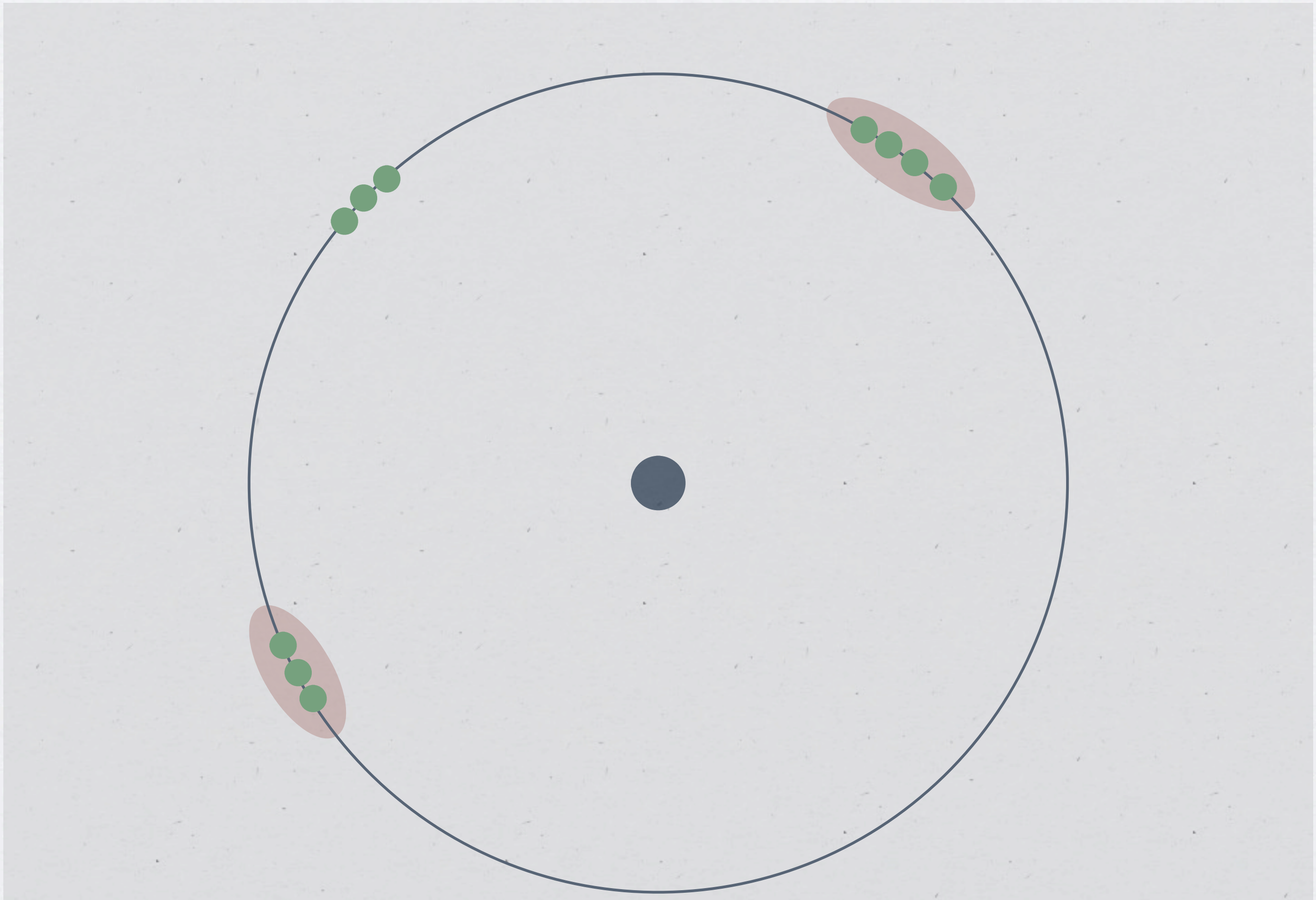


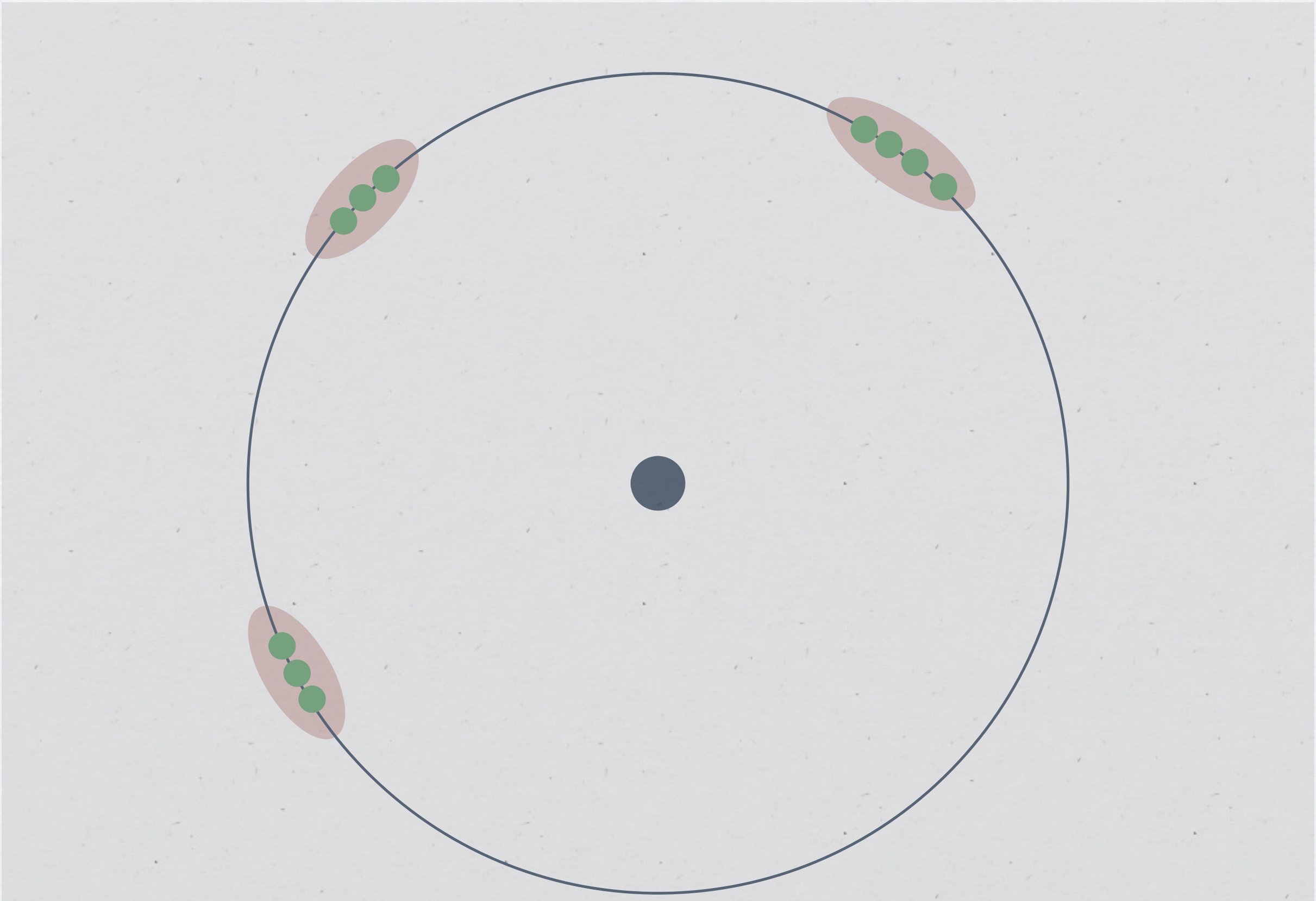


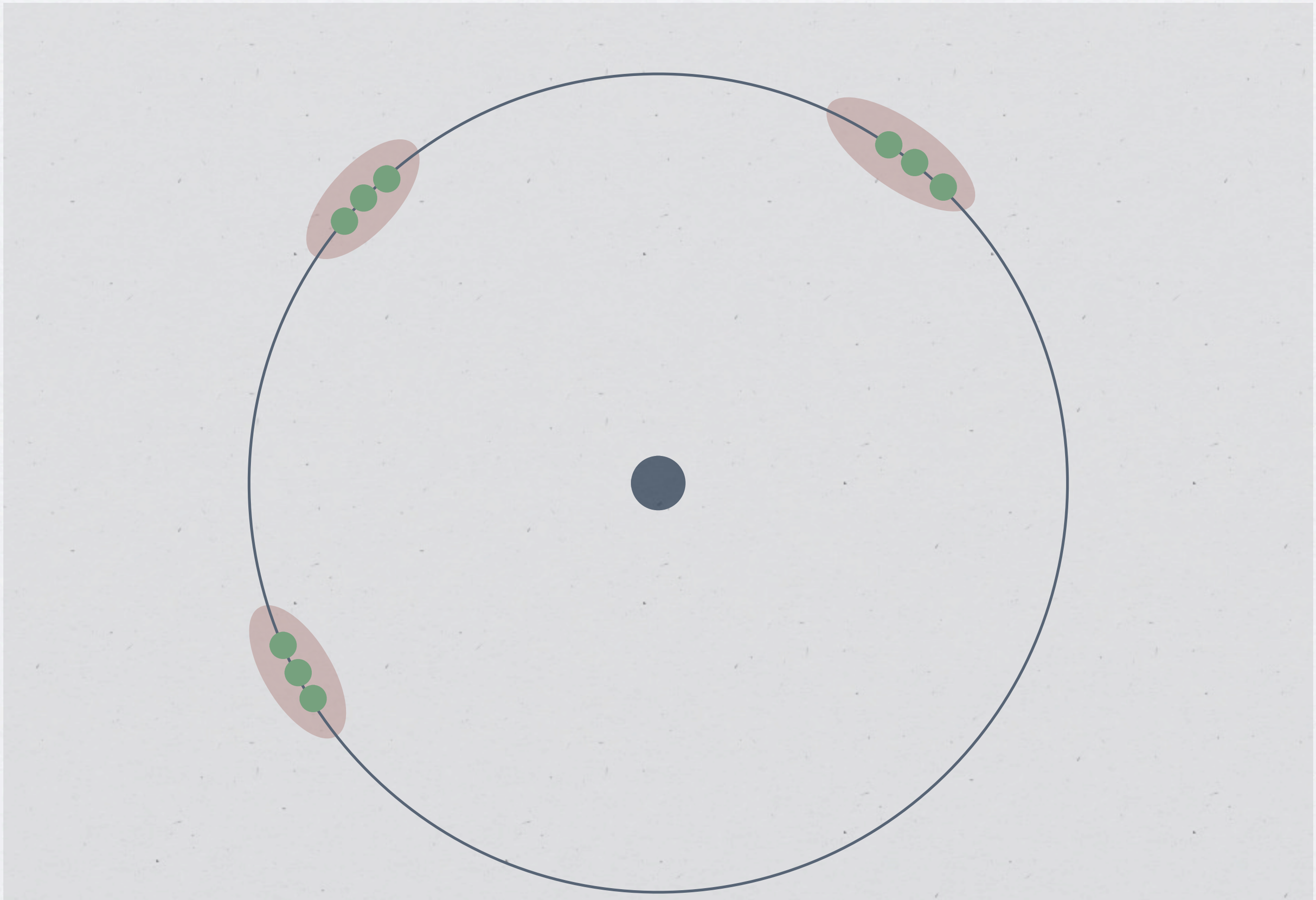


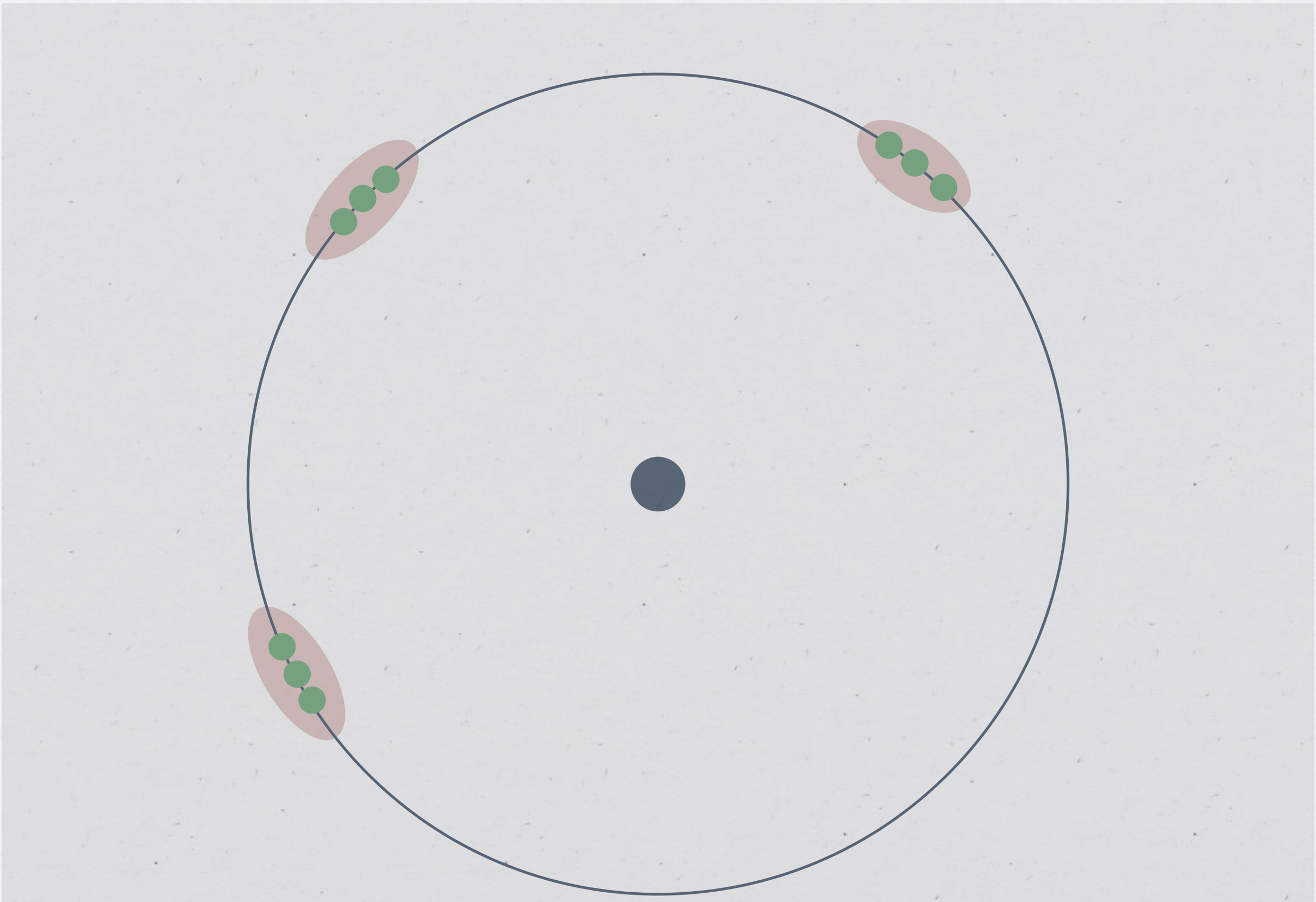


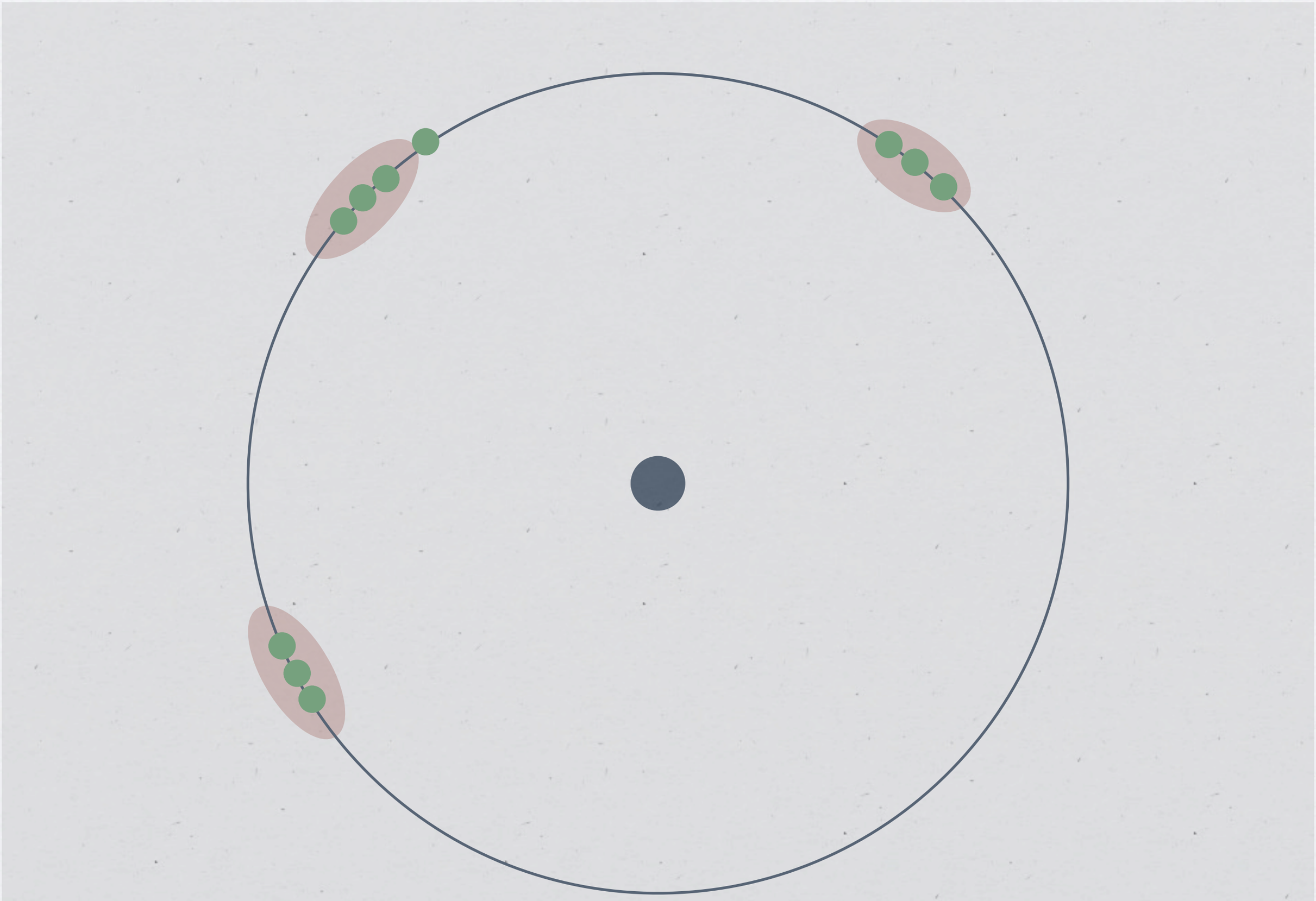


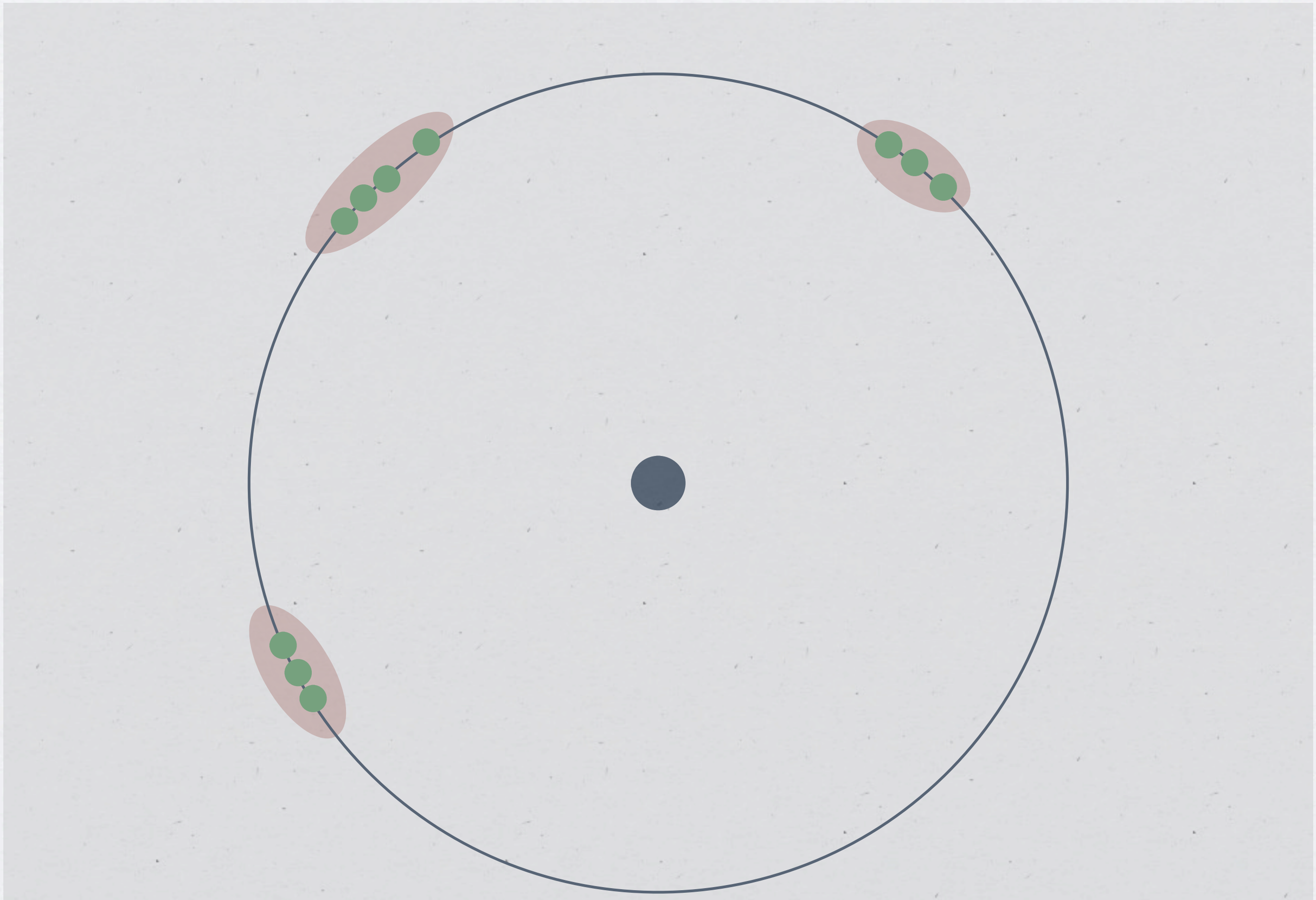


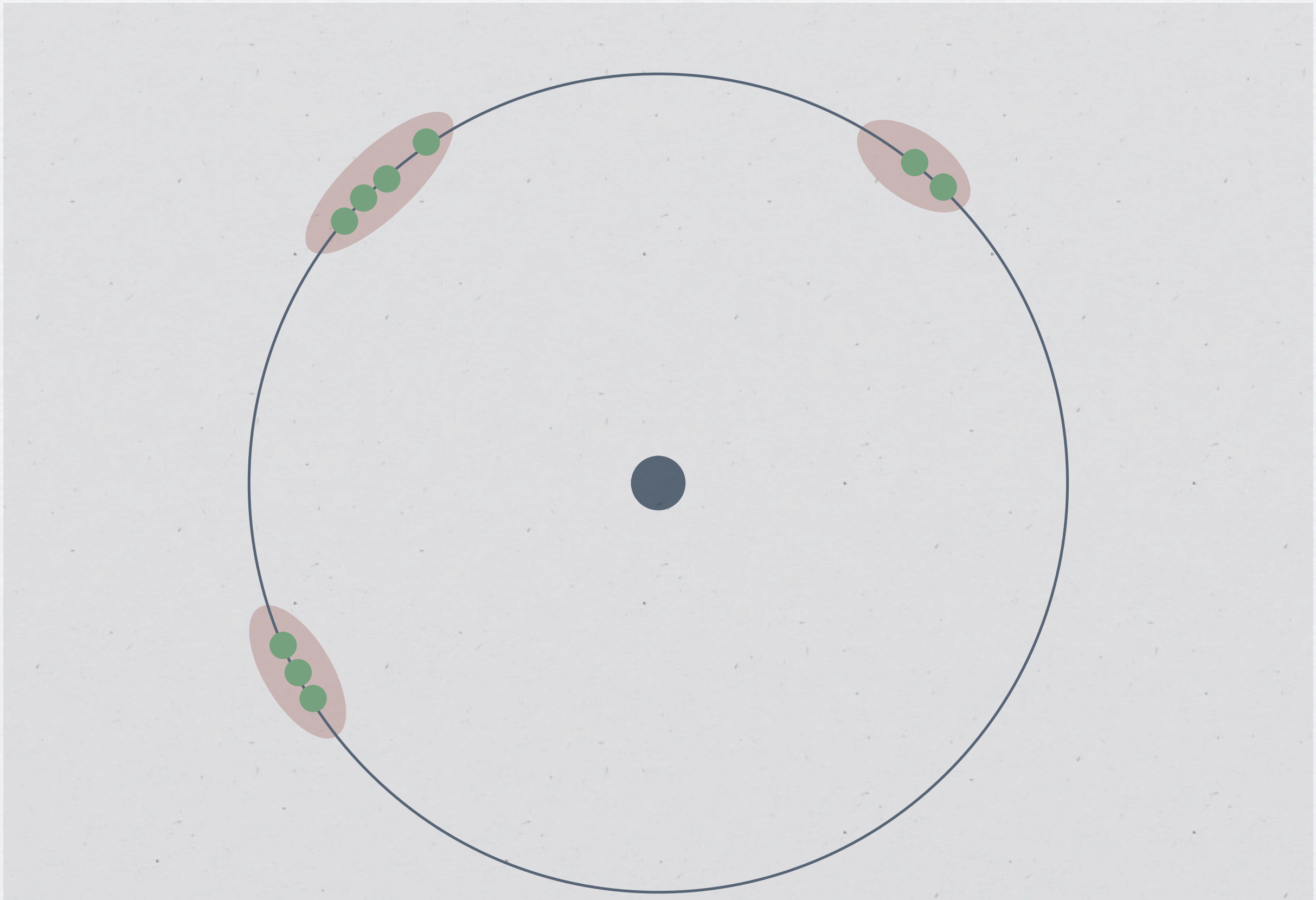


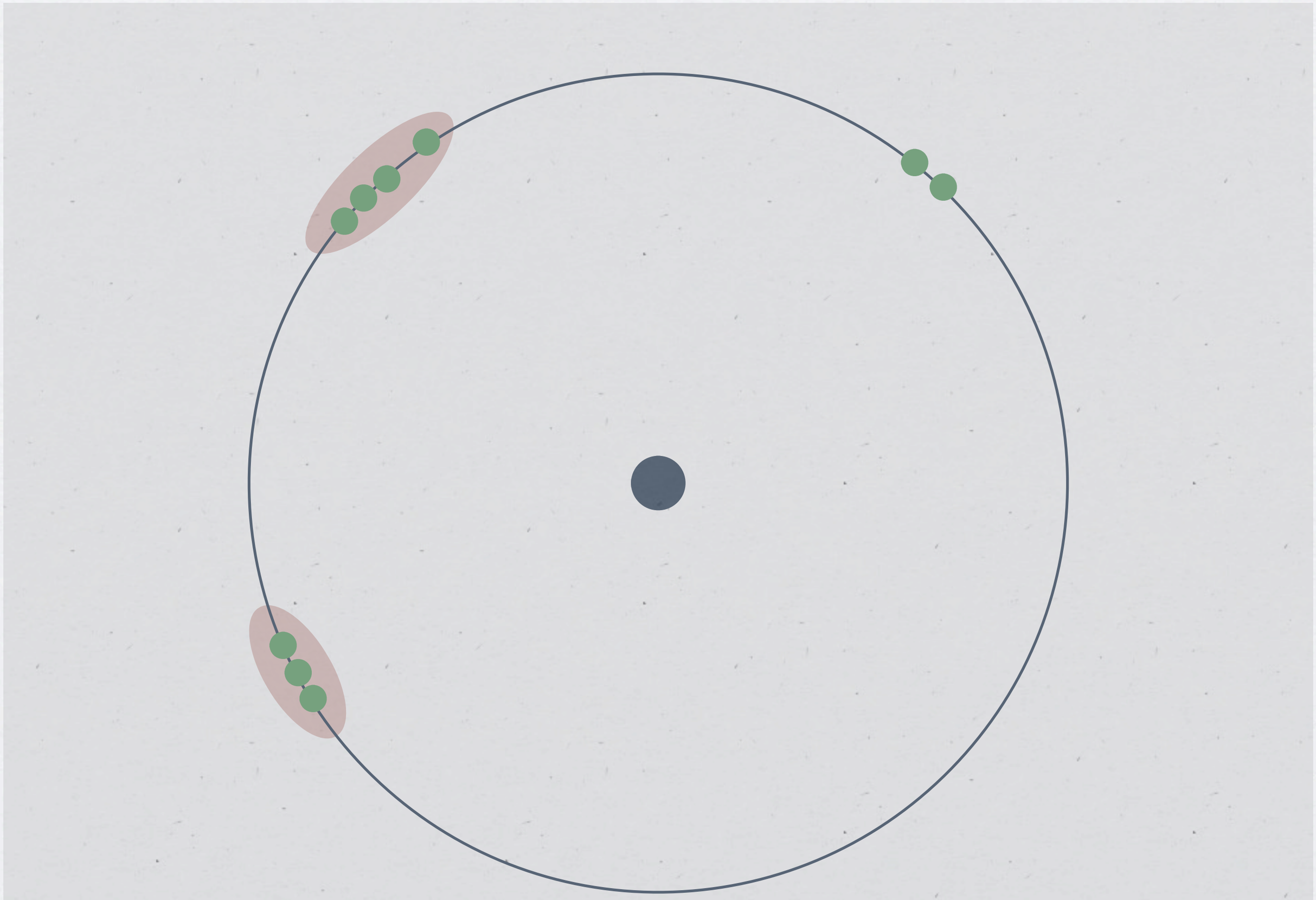




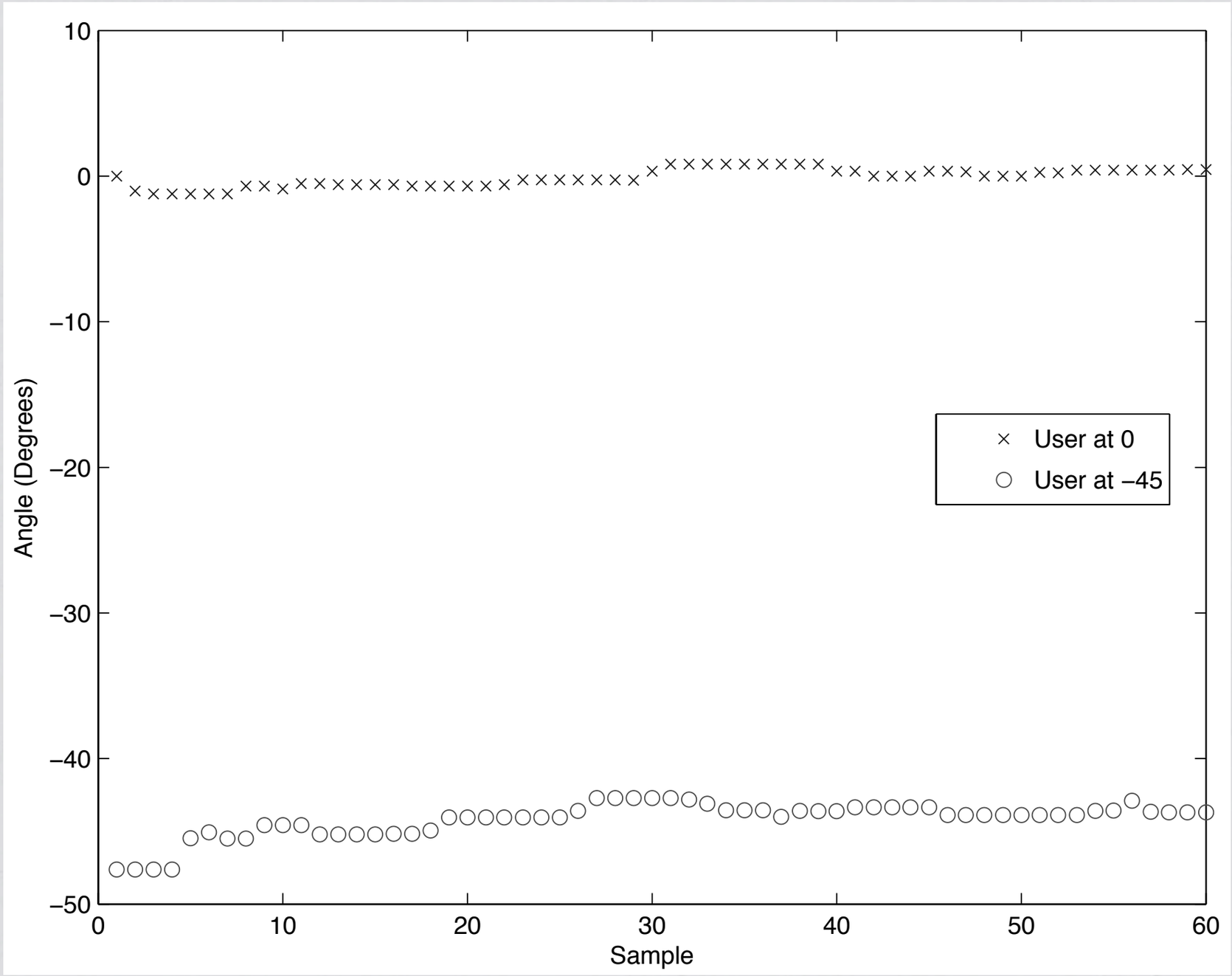




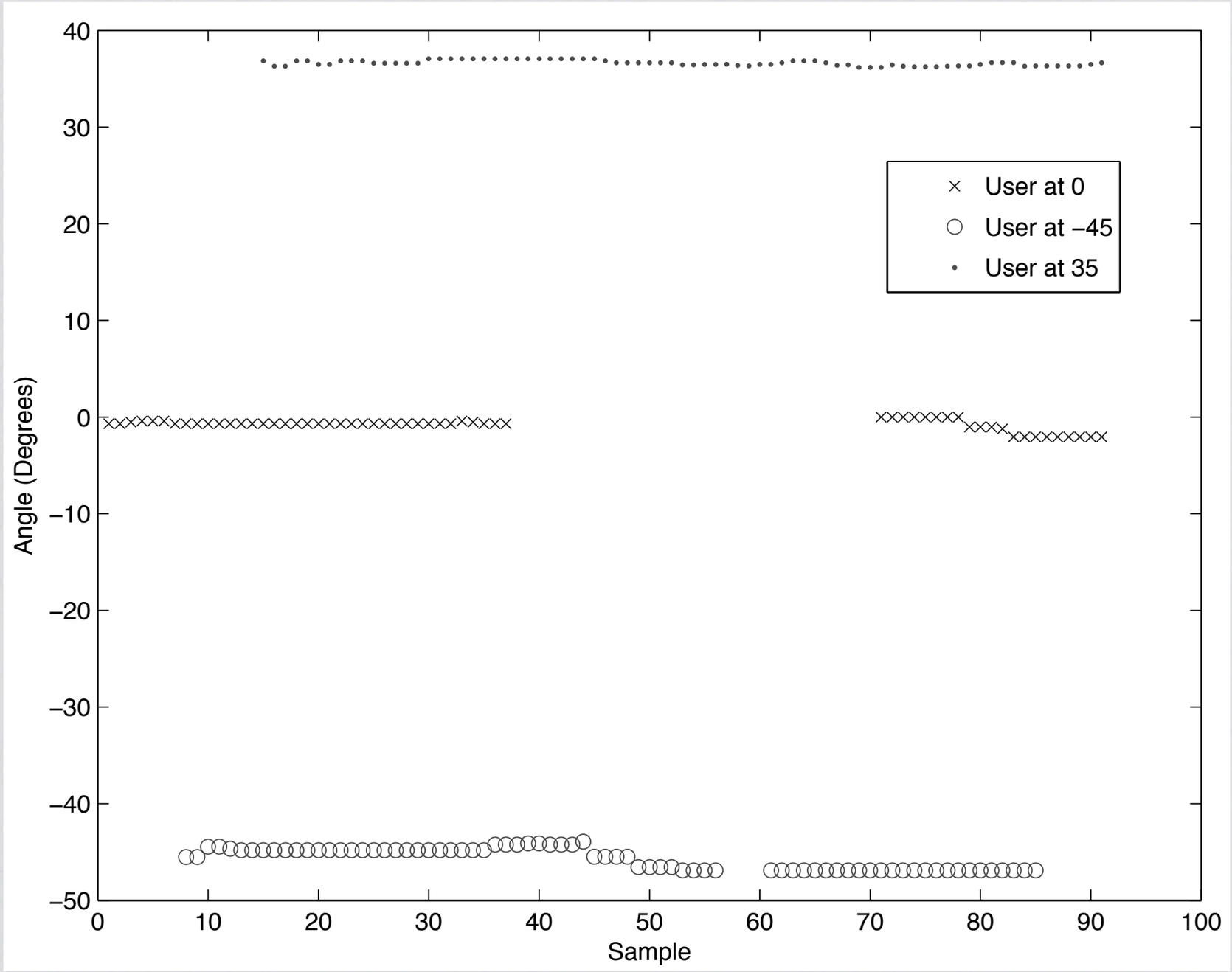




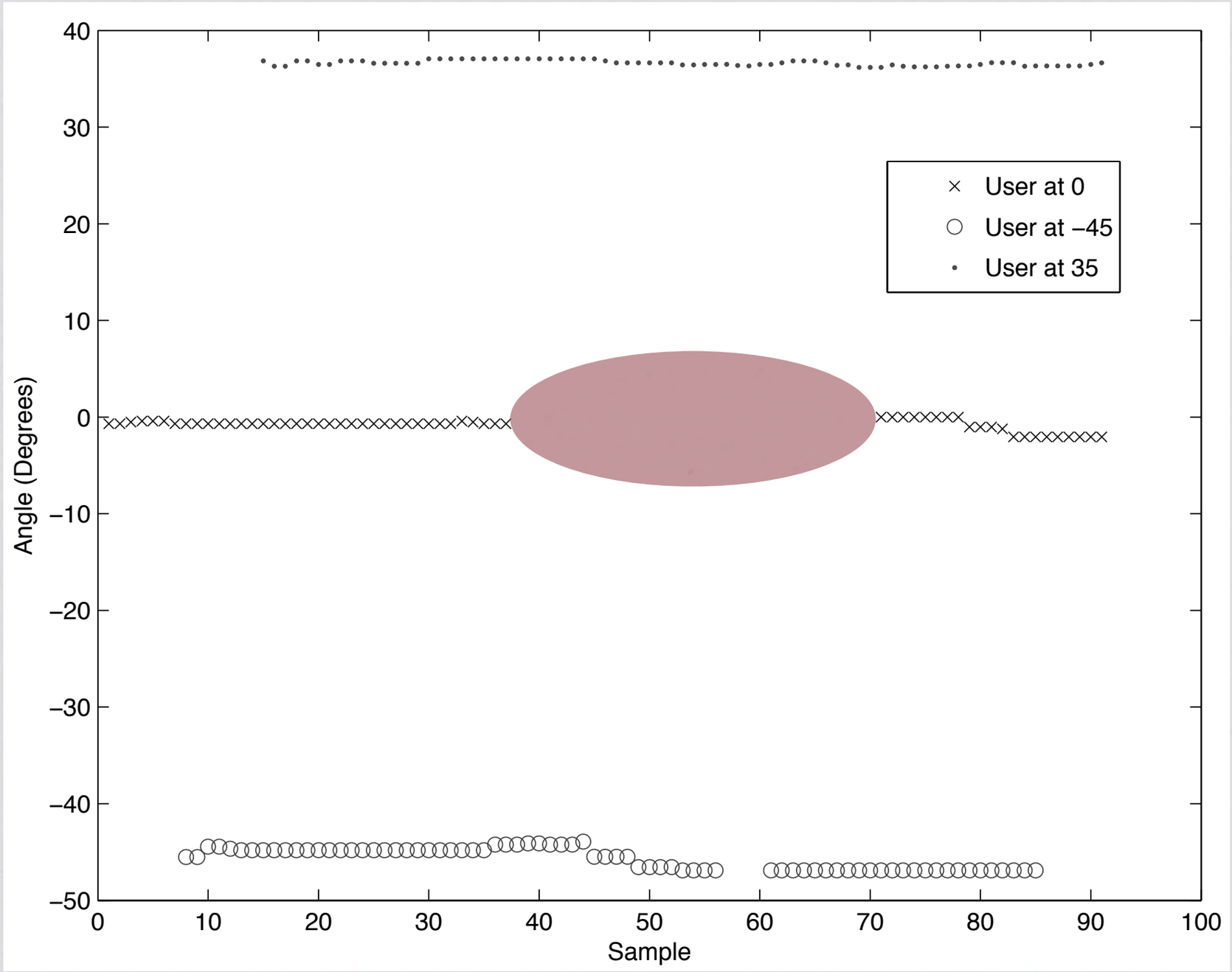
Evaluation: 2 Users



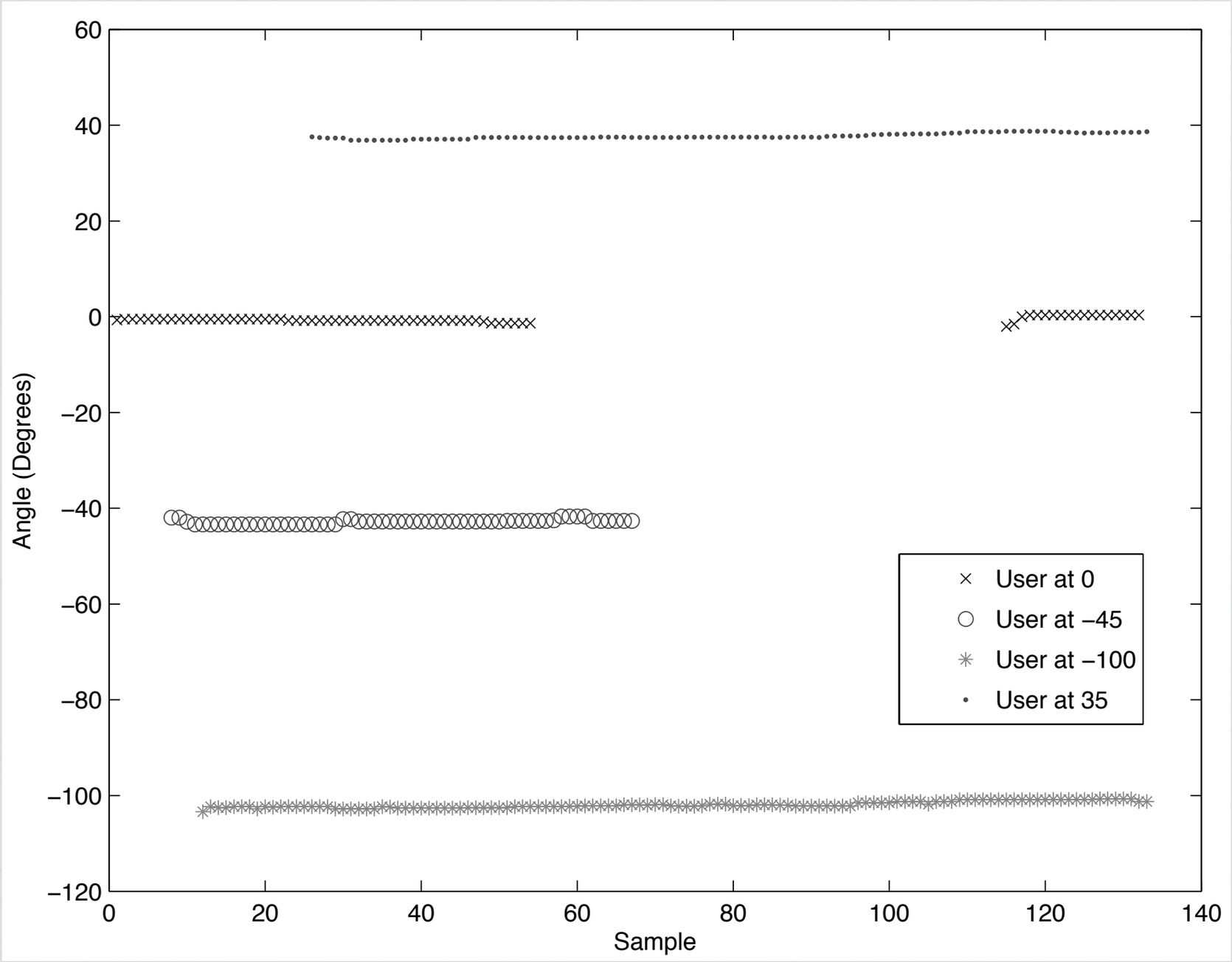
Evaluation: 3 Users



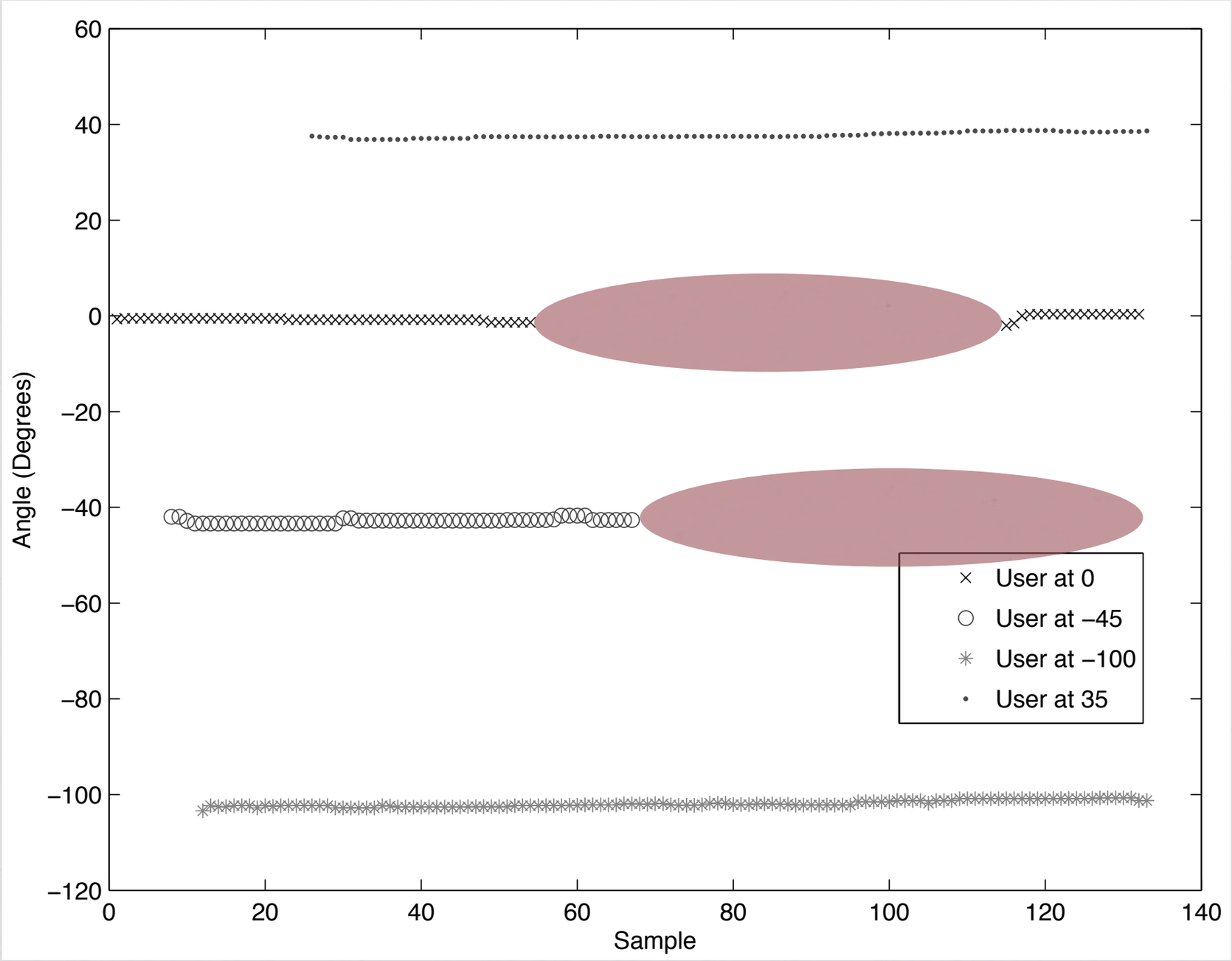
Evaluation: 3 Users



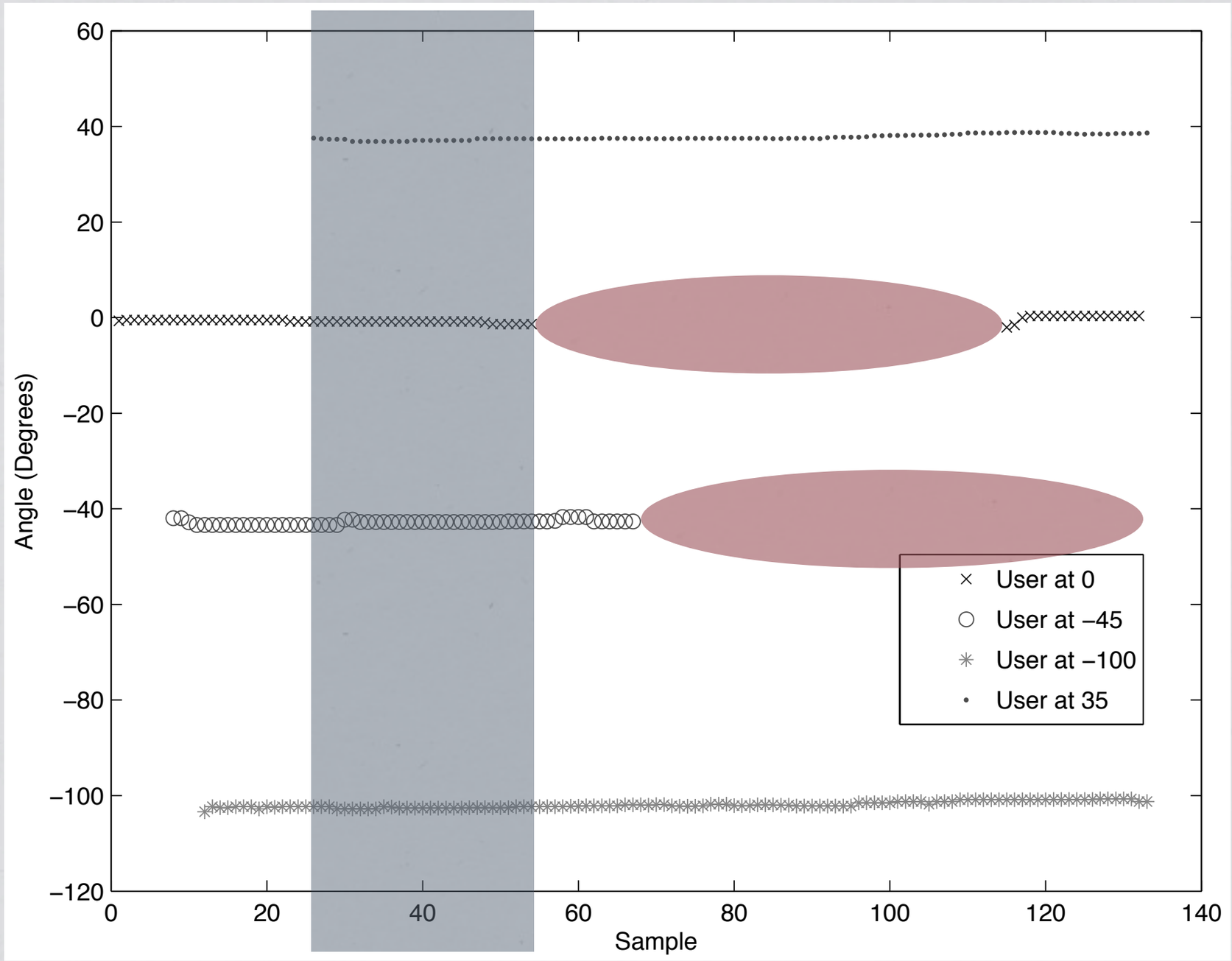
Evaluation: 4 Users



Evaluation: 4 Users



Evaluation: 4 Users



Evaluation

* Human speech overcame the electronic speaker audio when being tracked, but this is something desirable in the algorithm.

* ⇒ The evaluation was carried out in an acoustically-complex setting: highly reflective walls, low ceiling, computer fan noises, and moderate reverberation.

Evaluation

* Although, further testing needs to be carried out, these results report that with only **3 microphones**, the algorithm was able to track **4 users**.
    * MUSIC is not able to accomplish this.

Conclusion

* Multi-DOA Estimation can play an important part of Human-Robot Interaction.
* Carrying it out in a mobile robotic platform provides unique challenges in terms of hardware setup as well dynamic scenarios.
* The proposed algorithm/tracker, while being lightweight (a 3-mic setup), was able to perform adequately in a highly complex environment, tracking more users than microphones employed.

THANK YOU

Questions?