Ivan Miller. Summary of Sections 1 and 2 of "Recent Advances in High-Dimensional Robust Statistics" by Diakonikolas and Kane

1.Describe why robust mean estimation in high dimensions is challenging.

Outliers could contaminate the statistical analysis, and even a small number of perturbed data could significantly reduce the performance of a model and increase the error. The error tends to grow with the increase in dimension size of the data.

The simplest idea to estimate the mean of a distribution would be to identify the outliers and output the empirical mean of the remaining points. The difficulty when operating in high dimensions, the outliers cannot be identified at an individual level even when they move the mean significantly.

2.What's one of the main ideas behind the recent advances discussed in this paper?
The paper presents a basic filtering method along with an algorithm for its implementation that yields efficient robust mean estimators with optimal error bounds (see pseudocode below):

1. Compute $\mathbf{Cov}[T]$ and its largest eigenvalue ν.
2. If $\nu \leq 1 + \lambda$, return μ_T.
3. Else
    - Compute f as guaranteed in the theorem statement.
    - Remove each $x \in T$ with probability $f(x)/\max_{x \in T} f(x)$ and return to Step 1 with the new set T.

It also covers several practical methods of implementing the filtering algorithm that have have different practical performance on real datasets:

- randomized thresholding - the easiest method for implementing, only produce bad results when the size of the corrupt sample is very large
- independent removal - leads to less variance but has a higher probability of failure
- deterministic reweighting - is somewhat slower in practice as its worst-case runtime and its typical runtime are comparable

3.Do you think the adversarial model here of getting to corrupt a fixed fraction of the data is a reasonable one?

I think the most reasonable approach would be to focus on a combination of getting to corrupt a fixed fraction of data together with expanding the bounds of perturbed values. Since increase in both metrics gives more control to the adversary, our ability to control the error in those conditions could yield highest results in detecting adversary behavior.

4. What's a question you would like to see answered in the rest of the paper or in follow-up work?

I would like to see how the performance of the presented algorithm changes as a function of the number of dimensions.