

SOFTWARE MATTERS:

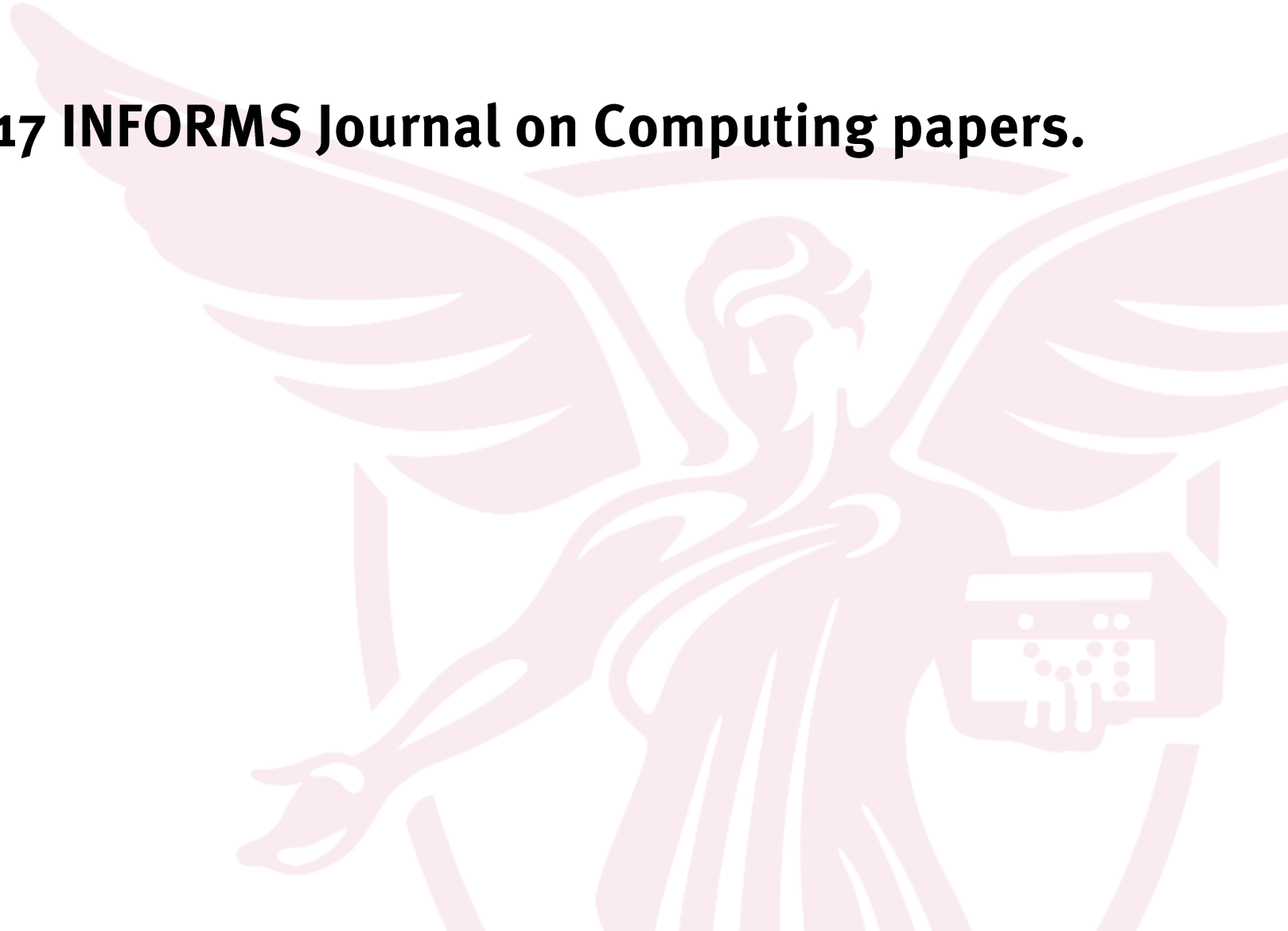
A Snapshot of Software Development Practice in Operations Research



Huseyin Ergin & Mesut Yavuz

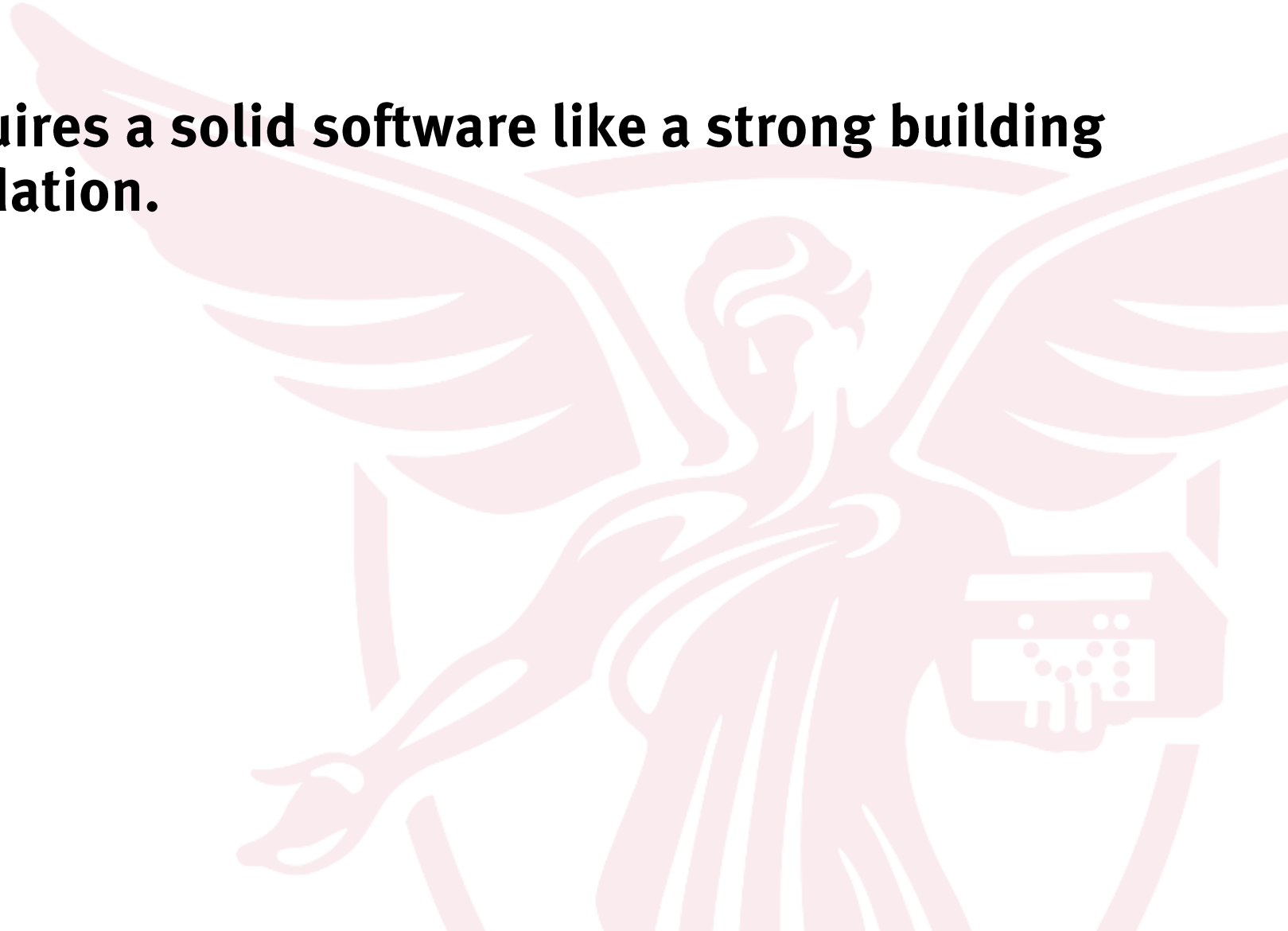
Summary

- **We have analyzed 2017 INFORMS Journal on Computing papers.**



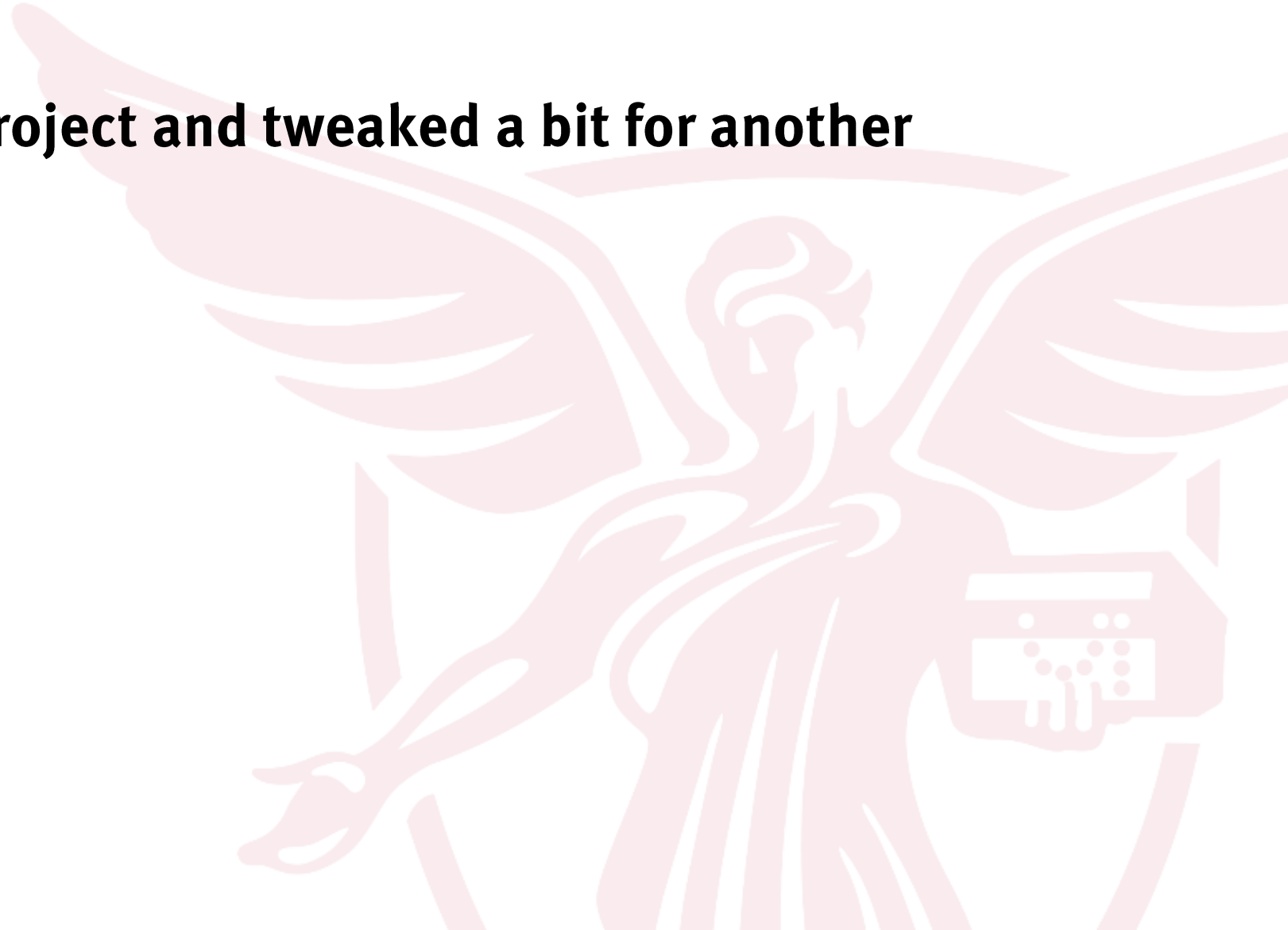
Why?

- **Serious research requires a solid software like a strong building requires a solid foundation.**



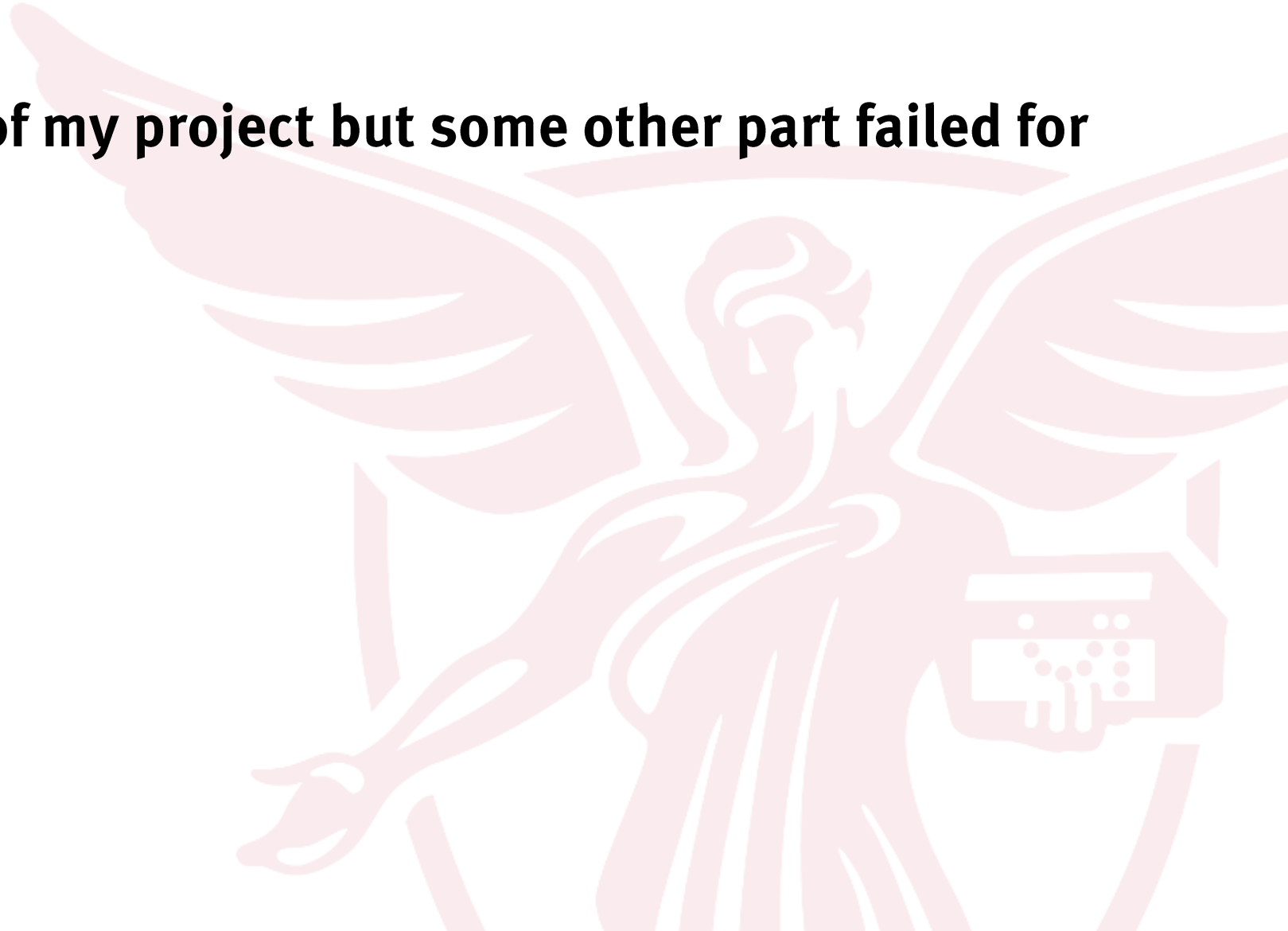
We have been there and done that!

- **“I copied my whole project and tweaked a bit for another research!”**



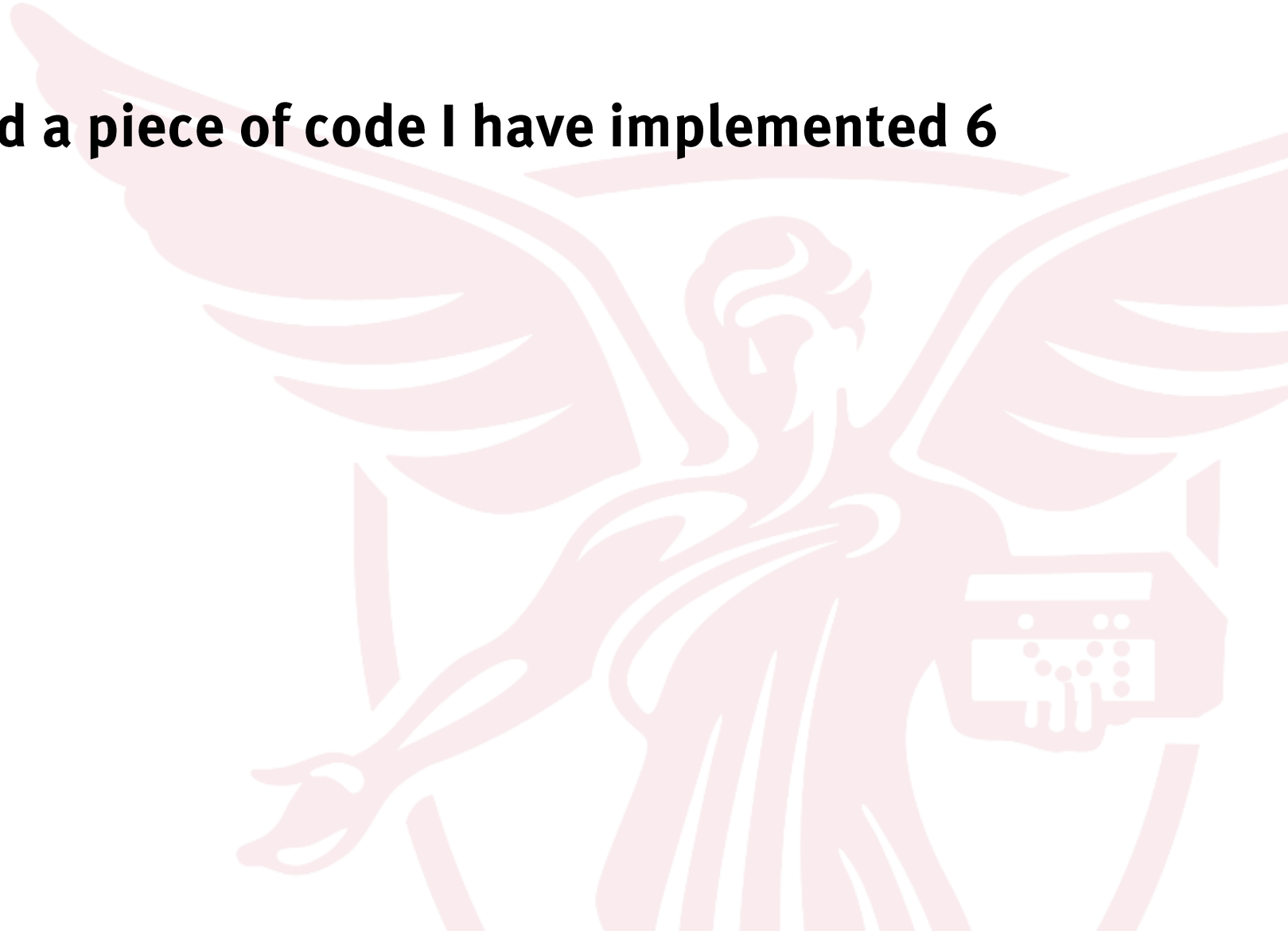
We have been there and done that!

- **“I modified one part of my project but some other part failed for this!”**



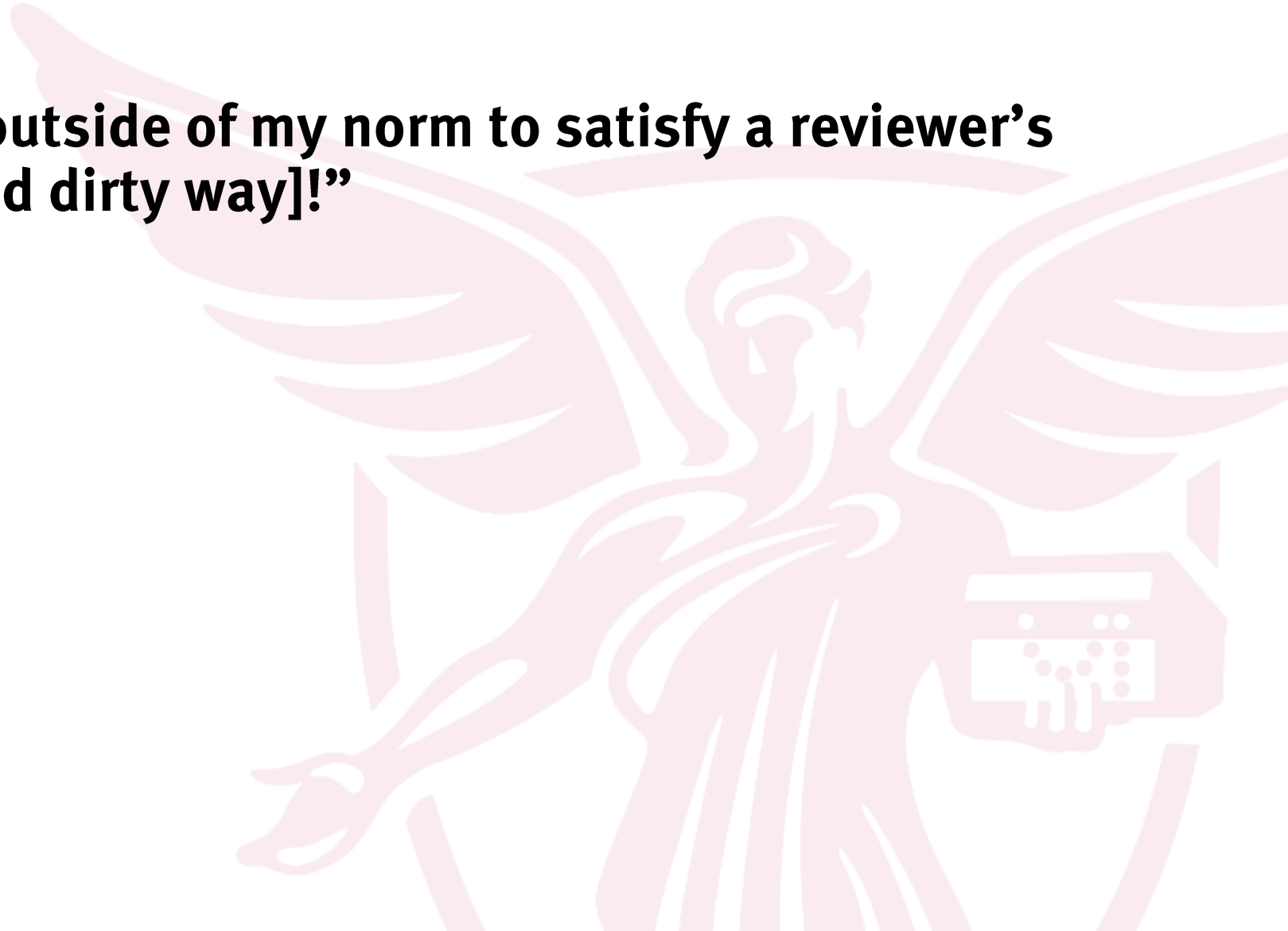
We have been there and done that!

- **“I couldn’t understand a piece of code I have implemented 6 months ago!”**



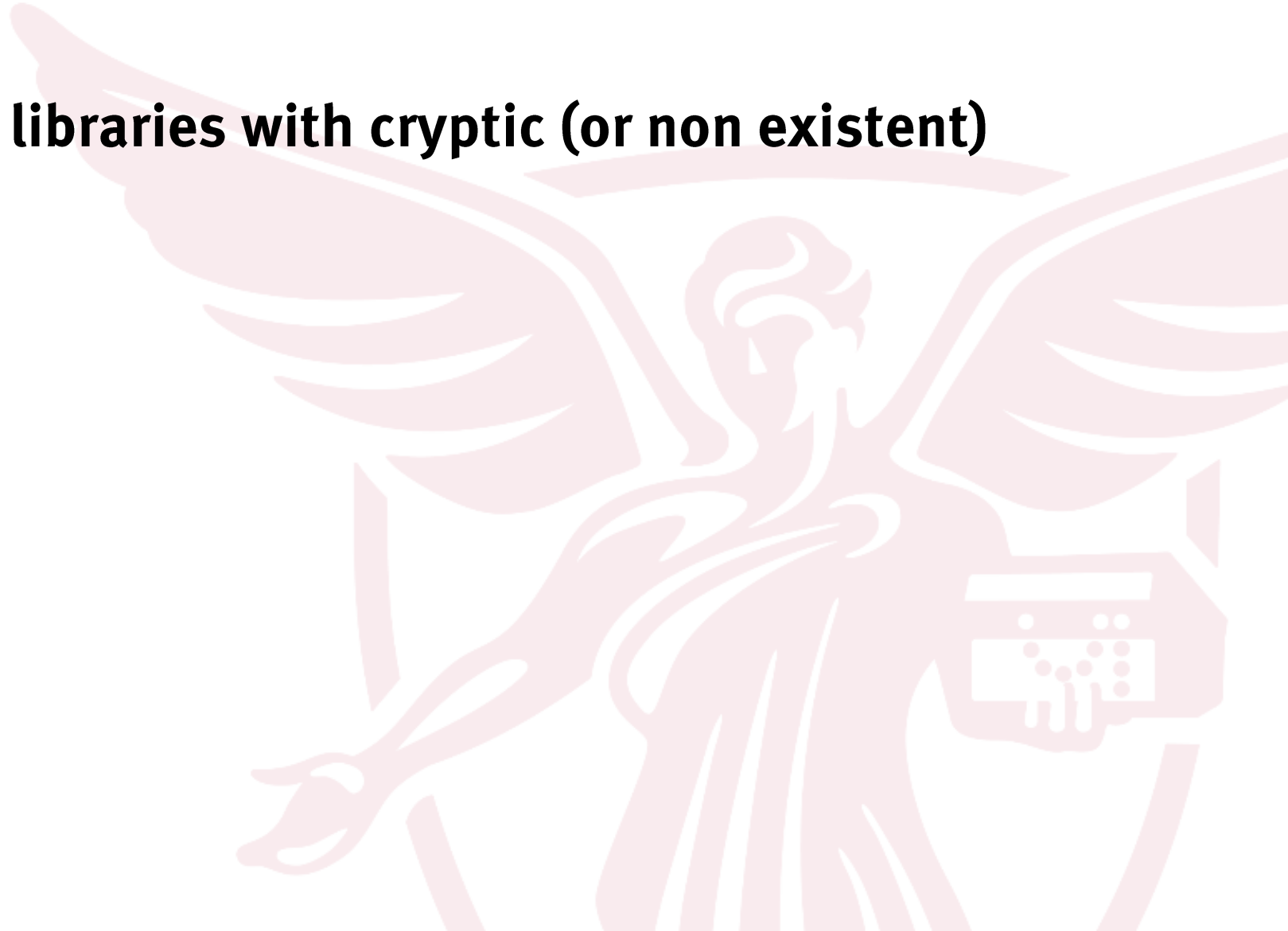
We have been there and done that!

- **“I modified my code outside of my norm to satisfy a reviewer’s request [in a quick and dirty way]!”**



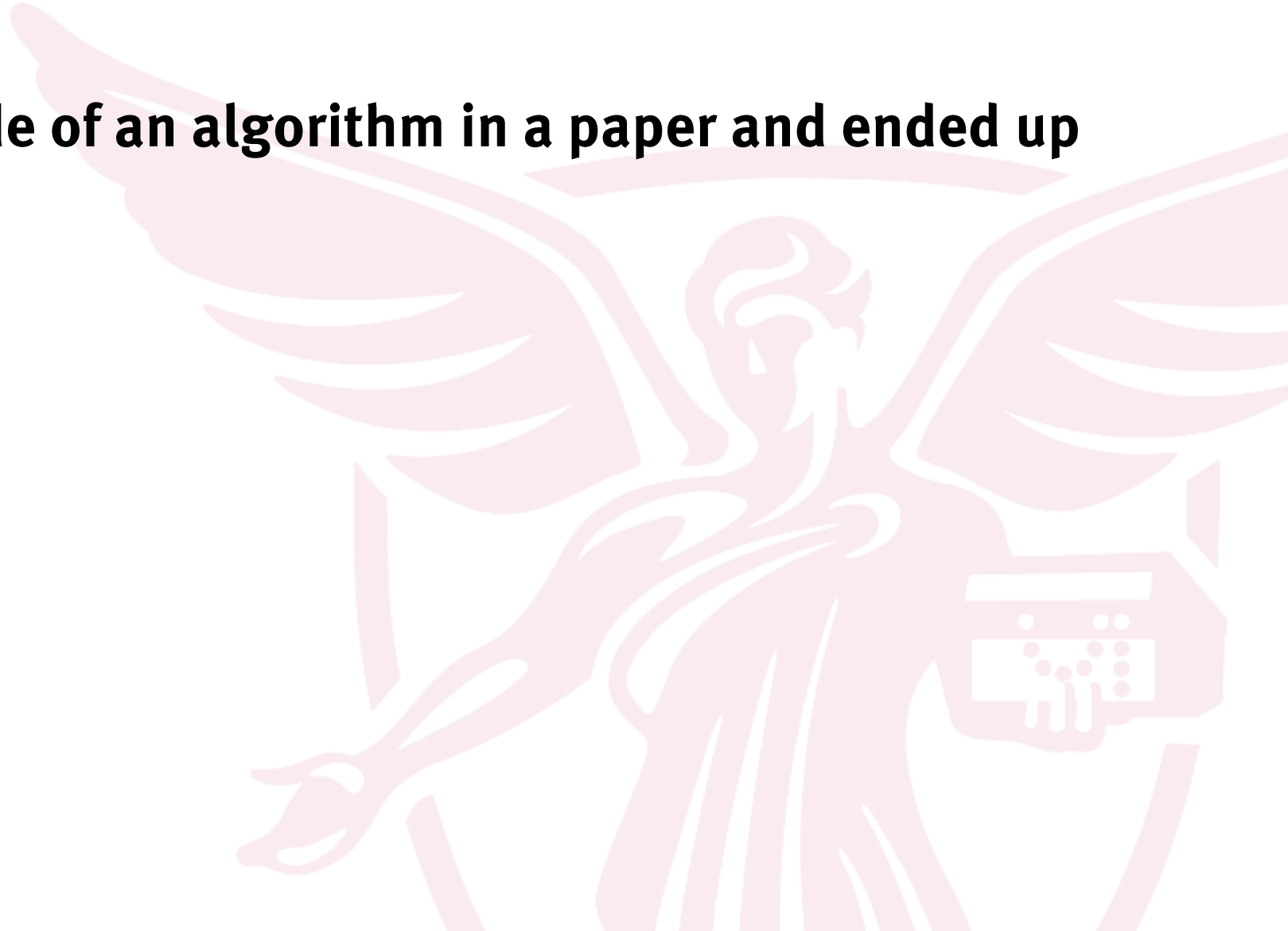
We have been there and done that!

- **“I have used external libraries with cryptic (or non existent) documentation!”**



We have been there and done that!

- **“I tracked down a code of an algorithm in a paper and ended up in a desert!”**



Software is important!

- But why?

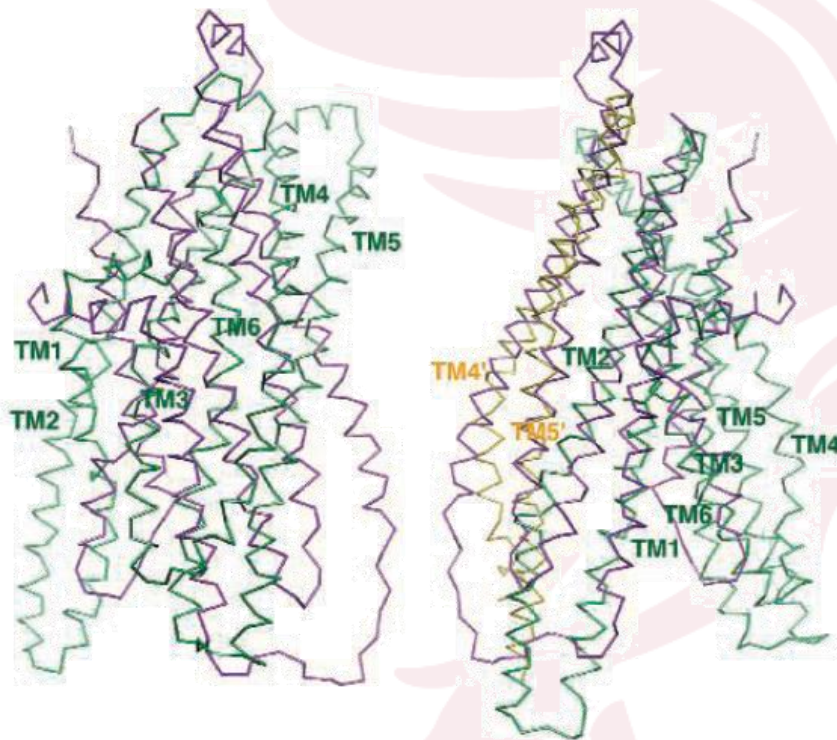


A Scientist's Nightmare: Software Problem Leads to Five Retractions

Until recently, Geoffrey Chang's career was on a trajectory most young scientists only dream about. In 1999, at the age of 28, the protein crystallographer landed a faculty position at the prestigious Scripps Research Institute in San Diego, California. The next year, in a ceremony at the White House, Chang received a Presidential Early Career Award for Scientists and Engineers, the country's highest honor for young researchers. His lab generated a stream of high-profile papers detailing the molecular structures of important proteins embedded in cell membranes.

Then the dream turned into a nightmare. In September, Swiss researchers published a paper in *Nature* that cast serious doubt on a protein structure Chang's group had described in a 2001 *Science* paper. When he investigated, Chang was horrified to discover that a homemade data-analysis program had flipped two columns of data, inverting the electron-density map from which his team had derived the final protein structure. Unfortunately, his group had used the program to analyze data for

2001 *Science* paper, which described the structure of a protein called MsbA, isolated from the bacterium *Escherichia coli*. MsbA belongs to a huge and ancient family of molecules that use energy from adenosine triphosphate to transport molecules across cell membranes. These so-called ABC transporters perform many



Flipping fiasco. The structures of MsbA (purple) and Sav1866 (green) overlap little (left) until MsbA is inverted (right).

Sciences and a 2005 *Science* paper, described EmrE, a different type of transporter protein.

Crystallizing and obtaining structures of five membrane proteins in just over 5 years was an incredible feat, says Chang's former postdoc adviser Douglas Rees of the California Institute of Technology in Pasadena. Such proteins are a challenge for crystallographers because they are large, unwieldy, and notoriously difficult to coax into the crystals needed for x-ray crystallography. Rees says determination was at the root of Chang's success: "He has an incredible drive and work ethic. He really pushed the field in the sense of getting things to crystallize that no one else had been able to do." Chang's data are good, Rees says, but the faulty software threw everything off.

Ironically, another former postdoc in Rees's lab, Kaspar Locher, exposed the mistake. In the 14 September issue of *Nature*, Locher, now at the Swiss Federal Institute of Technology in Zurich, described the structure of an ABC transporter called Sav1866 from *Staphylococcus aureus*. The structure was dramatically—and unexpectedly—different from that of MsbA. After pulling up Sav1866 and Chang's MsbA from *S. typhimurium* on a computer screen, Locher says he realized in minutes that the MsbA structure was inverted. Interpreting the "hand" of a molecule is always a challenge for crystallographers,

A Scientist's Nightmare: Software Problem Leads to Five Retractions

For several years, Chang's research team has been working on a project called "Protein Structure Prediction" (PSP). In 2015, at the age of 35, the young researcher published his first paper in the journal *Science*. The paper, titled "A New Approach to Protein Structure Prediction," was widely cited and earned Chang a reputation as a rising star in the field.

San Diego, California. The next year, in a ceremony at the White House, Chang received a Presidential Early Career Award for Scientists and Engineers, the country's highest honor for young researchers. His lab generated a

series of high-quality protein structure predictions that were used by other researchers in the field. Chang's work was widely praised, and he was invited to give the keynote address at a major conference in 2017.

Then, in 2018, a major software bug was discovered in the code that generated the protein structure predictions. The bug was traced back to a specific line of code that had been introduced in 2015. The error was subtle but significant, leading to incorrect predictions for a large number of proteins. The discovery was a major setback for Chang's lab, and it led to a series of retractions of his papers. The first retraction was in 2018, followed by four more in the following years. The retractions were a major blow to Chang's reputation, and he was forced to publicly acknowledge the error. The software problem was eventually fixed, but the damage to Chang's career was significant.

2018. The software, which was used to generate the protein structure predictions, contained a major bug that led to incorrect predictions for a large number of proteins. The error was traced back to a specific line of code that had been introduced in 2015. The bug was subtle but significant, leading to incorrect predictions for a large number of proteins. The discovery was a major setback for Chang's lab, and it led to a series of retractions of his papers. The first retraction was in 2018, followed by four more in the following years. The retractions were a major blow to Chang's reputation, and he was forced to publicly acknowledge the error. The software problem was eventually fixed, but the damage to Chang's career was significant.

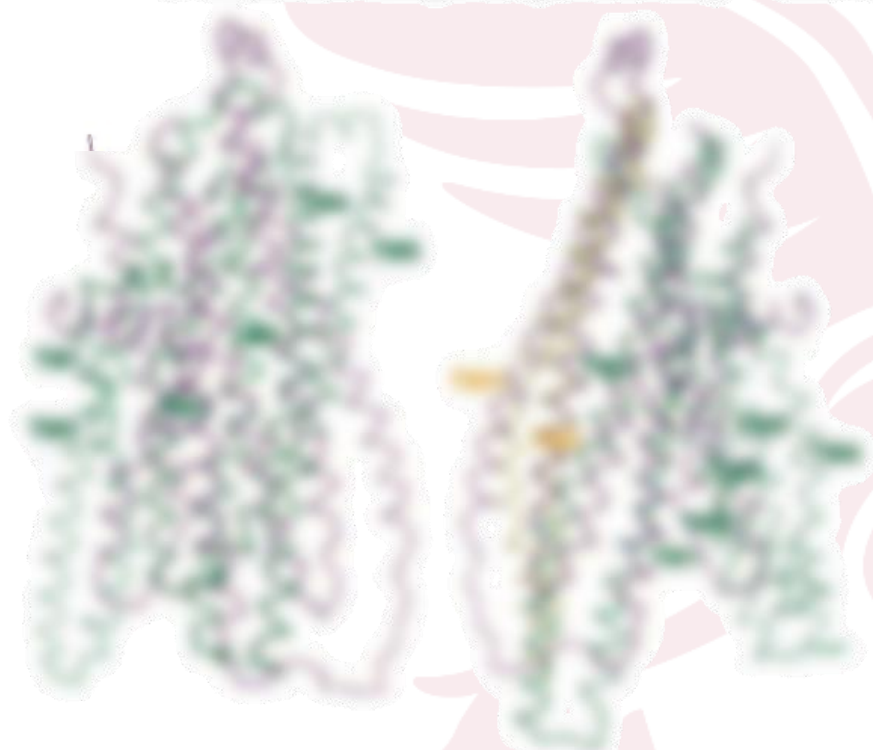


Figure 1: Two protein structures generated by the software. The left structure is a dimeric protein, and the right structure is a monomeric protein.

Chang's work was widely praised, and he was invited to give the keynote address at a major conference in 2017.

Then, in 2018, a major software bug was discovered in the code that generated the protein structure predictions. The bug was traced back to a specific line of code that had been introduced in 2015. The error was subtle but significant, leading to incorrect predictions for a large number of proteins. The discovery was a major setback for Chang's lab, and it led to a series of retractions of his papers. The first retraction was in 2018, followed by four more in the following years. The retractions were a major blow to Chang's reputation, and he was forced to publicly acknowledge the error. The software problem was eventually fixed, but the damage to Chang's career was significant.

Chang's work was widely praised, and he was invited to give the keynote address at a major conference in 2017.

Then, in 2018, a major software bug was discovered in the code that generated the protein structure predictions. The bug was traced back to a specific line of code that had been introduced in 2015. The error was subtle but significant, leading to incorrect predictions for a large number of proteins. The discovery was a major setback for Chang's lab, and it led to a series of retractions of his papers. The first retraction was in 2018, followed by four more in the following years. The retractions were a major blow to Chang's reputation, and he was forced to publicly acknowledge the error. The software problem was eventually fixed, but the damage to Chang's career was significant.

A Scientist's Nightmare: Software Problem Leads to Five Retractions

For nearly a decade, Chang's team has been working on a project that is one of the most important in the field of structural biology. The project is to determine the structure of a protein called the ribosome. The ribosome is a complex of proteins and RNA that is responsible for synthesizing proteins in all living cells. Chang's team has been working on this project since 2005, and they have made significant progress. They have determined the structure of several parts of the ribosome, and they have published their findings in several high-profile journals. However, in 2010, Chang's team published a paper in the journal *Nature* that described the structure of the entire ribosome. The paper was widely cited and was considered a major breakthrough in the field. However, in 2011, Chang's team discovered a software error in their data-analysis program that had flipped two columns of data, inverting the electron-density map from which they had derived the final protein structure.

The error was discovered by a postdoctoral fellow named David Baker, who was working on a different project in Chang's lab. Baker noticed that the structure of the ribosome that Chang's team had published in *Nature* was very different from the structure that he had determined using a different software program. He investigated the error and discovered that the software program had flipped two columns of data, inverting the electron-density map. Baker reported the error to Chang, and Chang's team investigated the error. They discovered that the error was caused by a software bug in their data-analysis program. The bug had flipped two columns of data, inverting the electron-density map. Chang's team had to re-analyze all of their data, and they had to re-publish their paper in *Nature*.

The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*.

paper. When he investigated, Chang was horrified to discover that a homemade data-analysis program had flipped two columns of data, inverting the electron-density map from which his team had derived the final protein structure.

The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*.

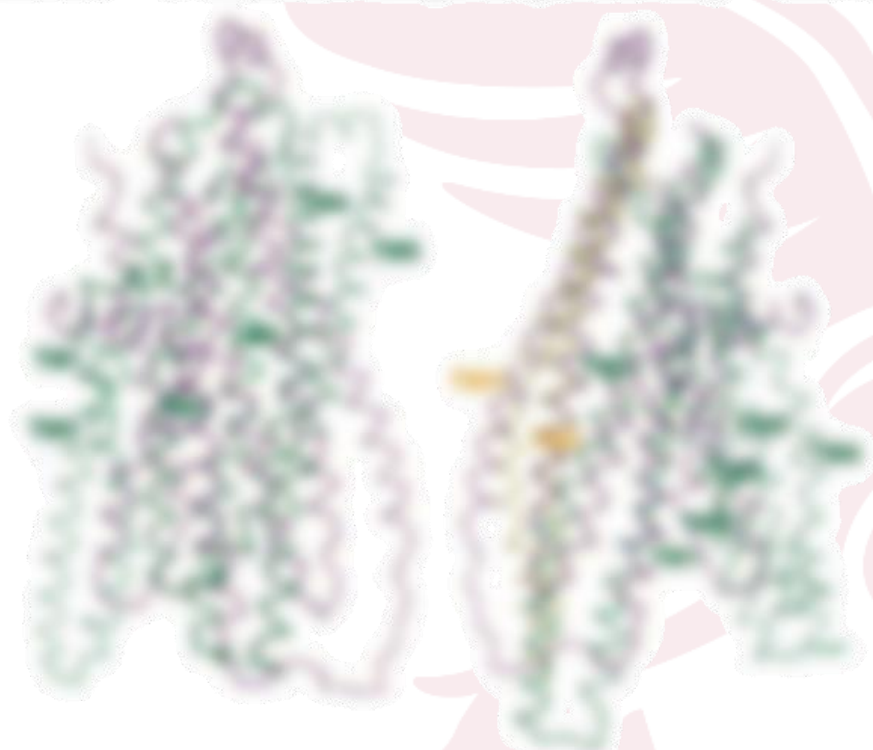


Figure 1. The structure of the ribosome, showing the large and small subunits. The large subunit is shown in green and purple, and the small subunit is shown in yellow.

The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*.

The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*.

The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*. The error was a major setback for Chang's team, and it was a major embarrassment for the journal *Nature*.

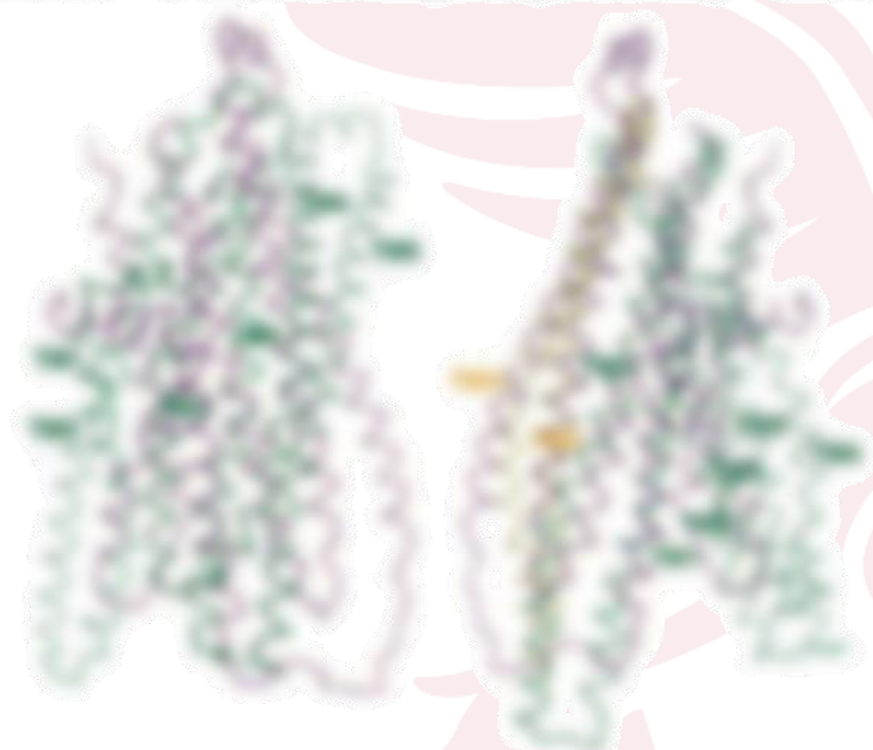
A Scientist's Nightmare: Software Problem Leads to Five Retractions

For decades, scientists have used computers to analyze complex data sets. In 1998, at the age of 28, the young crystallographer Chang began his work at the Princeton Biophysics Research Institute at the Chevy Chase campus. He was one of the first to use the new software, which was developed by the Princeton-based company, Bruker.

Chang's first job was to determine the structure of a protein called myoglobin. He was given a small amount of the protein and a small amount of heavy metal ions to help with the structure determination. He was given a small amount of the protein and a small amount of heavy metal ions to help with the structure determination.

When he started using the software, he noticed that the results were different from what he expected. He was given a small amount of the protein and a small amount of heavy metal ions to help with the structure determination. He was given a small amount of the protein and a small amount of heavy metal ions to help with the structure determination.

His first paper, which described the structure of myoglobin, was published in the journal *Nature*. It was one of the first papers to use the software. He was given a small amount of the protein and a small amount of heavy metal ions to help with the structure determination.



Chang's first paper, which described the structure of myoglobin, was published in the journal Nature. It was one of the first papers to use the software.

Chang's first paper, which described the structure of myoglobin, was published in the journal Nature. It was one of the first papers to use the software.

Chang's first paper, which described the structure of myoglobin, was published in the journal Nature. It was one of the first papers to use the software. He was given a small amount of the protein and a small amount of heavy metal ions to help with the structure determination.

Chang's data are good, Rees says, but the faulty software threw everything off.

Chang's data are good, Rees says, but the faulty software threw everything off. He was given a small amount of the protein and a small amount of heavy metal ions to help with the structure determination.

How good can we write software?

- **“Scientists typically spend 30% or more of their time developing software.”**
- **“90% or more of them are primarily self-taught.”**
- **“Most scientists learn most of what they know about developing software on their own or informally from their peers, rather than through formal training.”**

Replication crisis

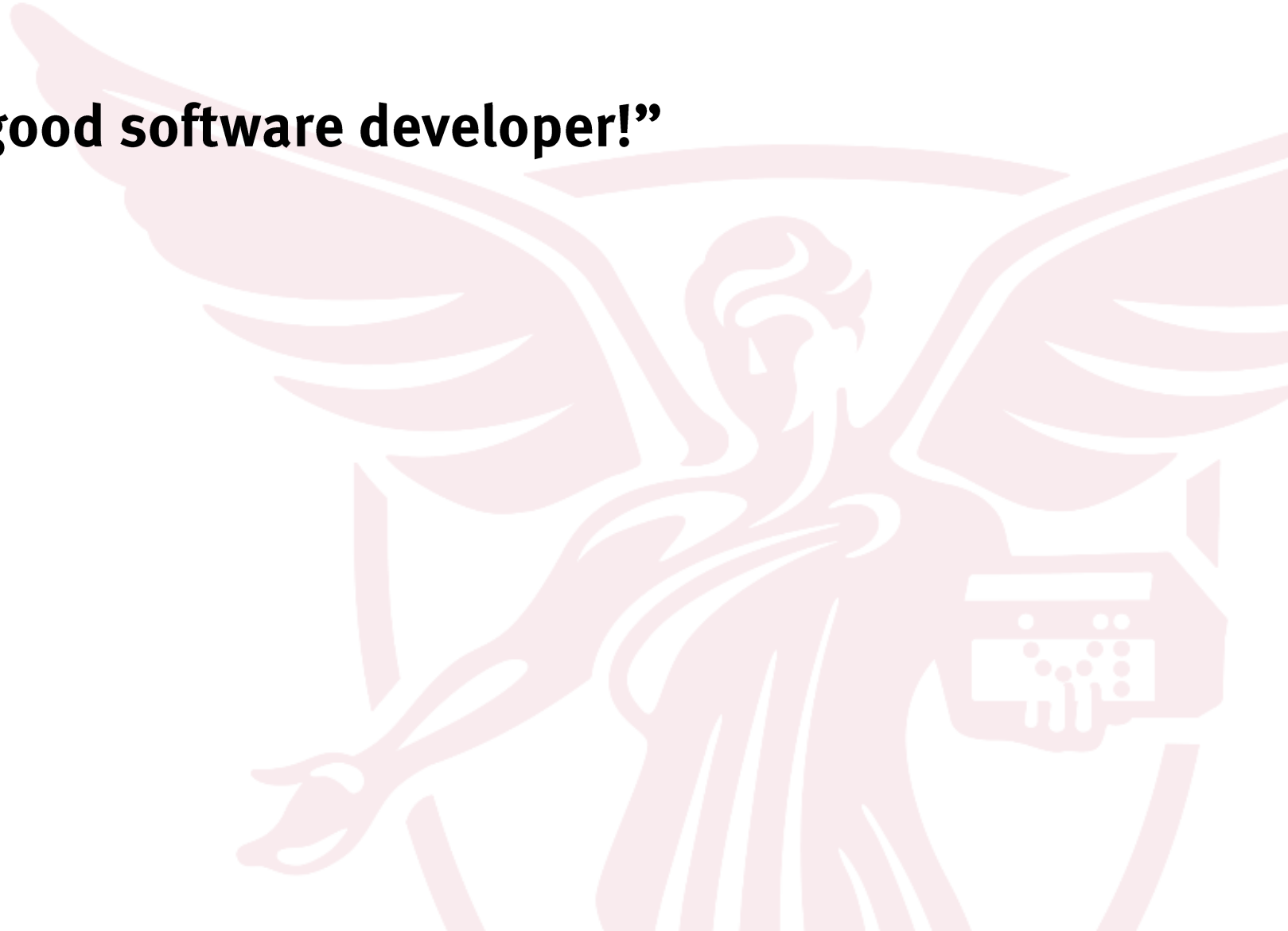
- **Chang and Li tried to replicate 67 economics papers from 13 well-regarded economics journal and ended up replicating only 49% of the studies with the extended help of the authors.**

Replication crisis

- Ioannidis et al. evaluated the replicability of analyses from 18 genetics articles and reproduced only 2 of them entirely and 6 of them with discrepancies.

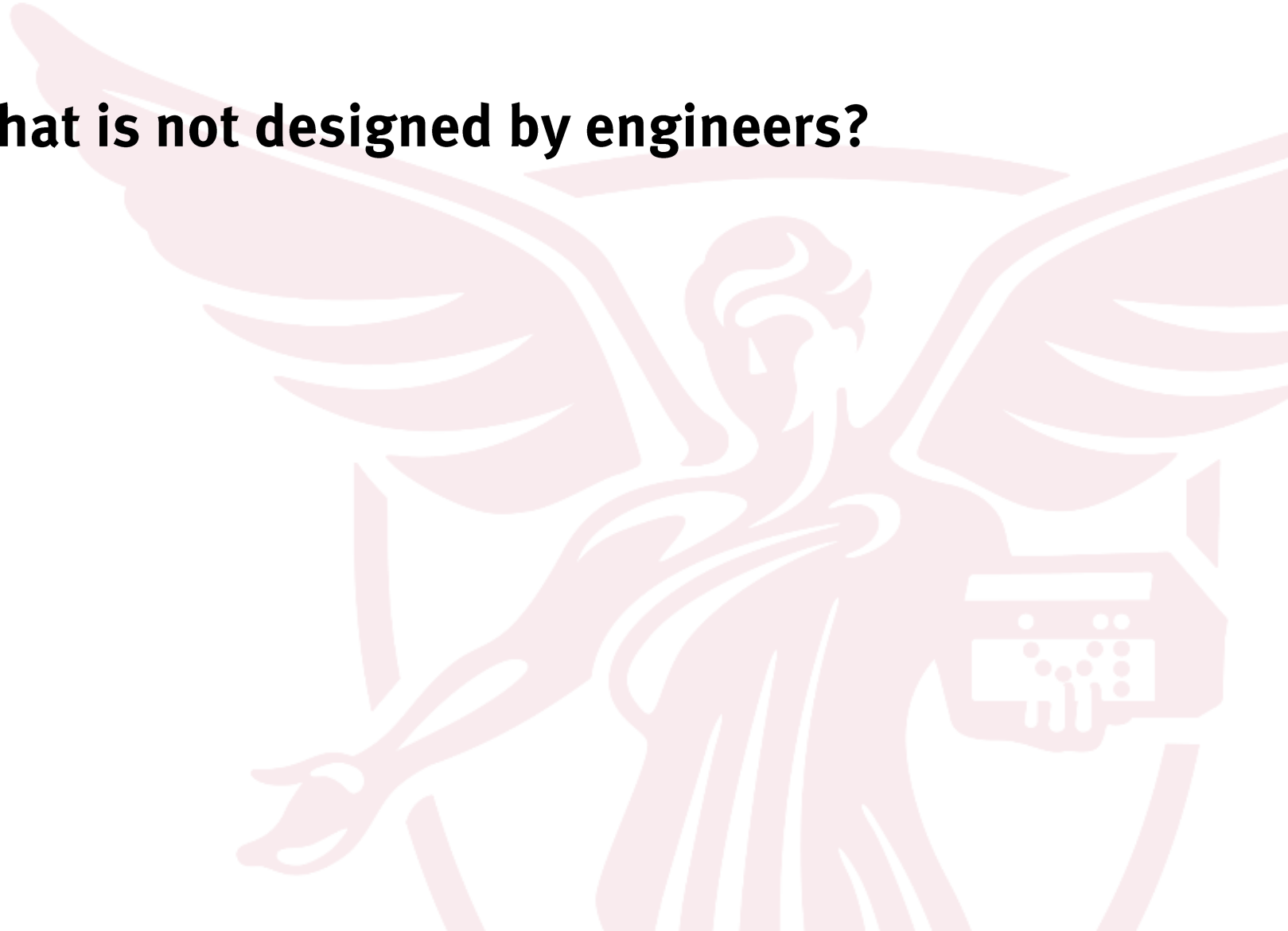
It is your turn:

- **“I consider myself a good software developer!”**



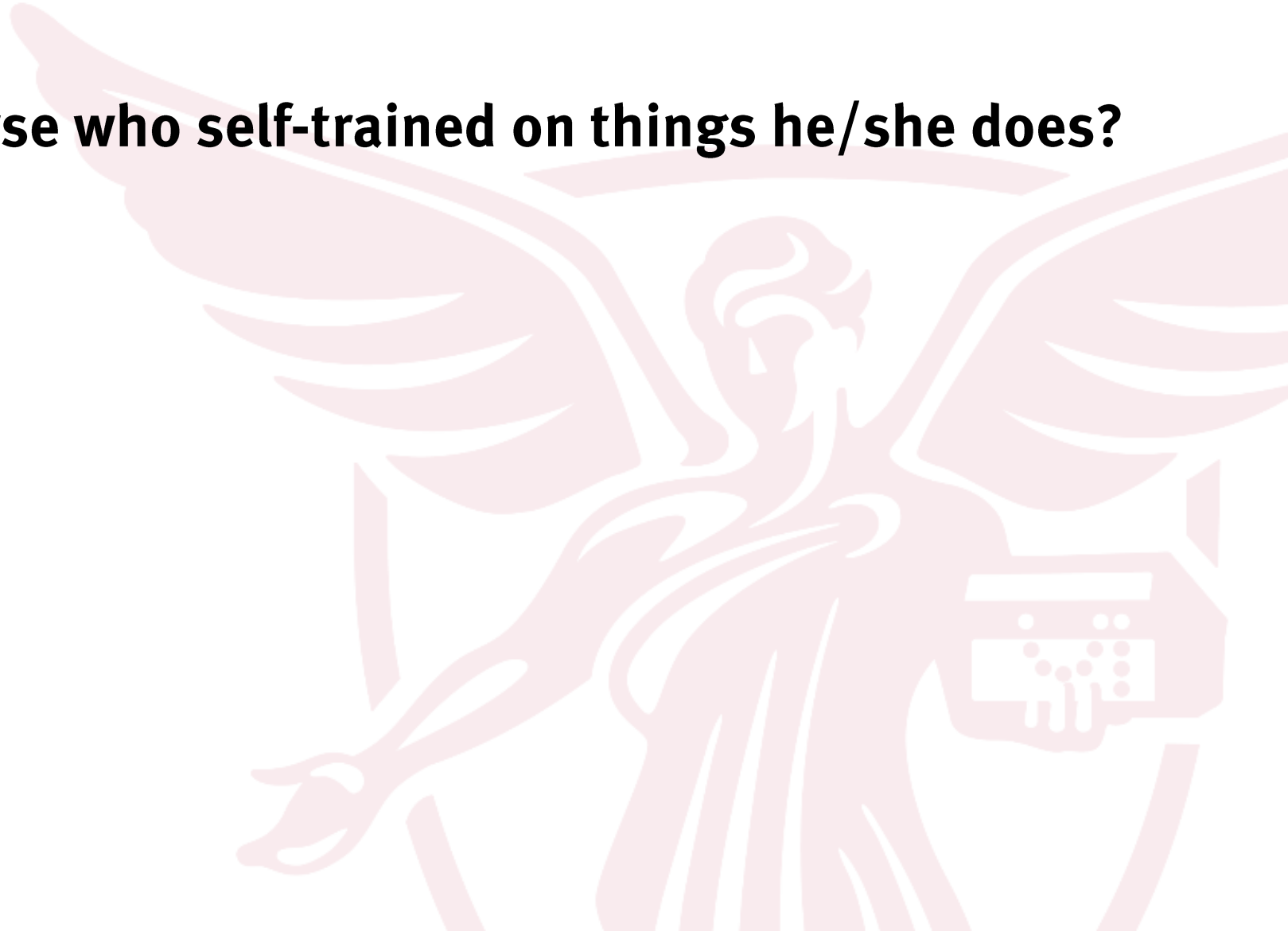
It is your turn:

- **Would you use a car that is not designed by engineers?**



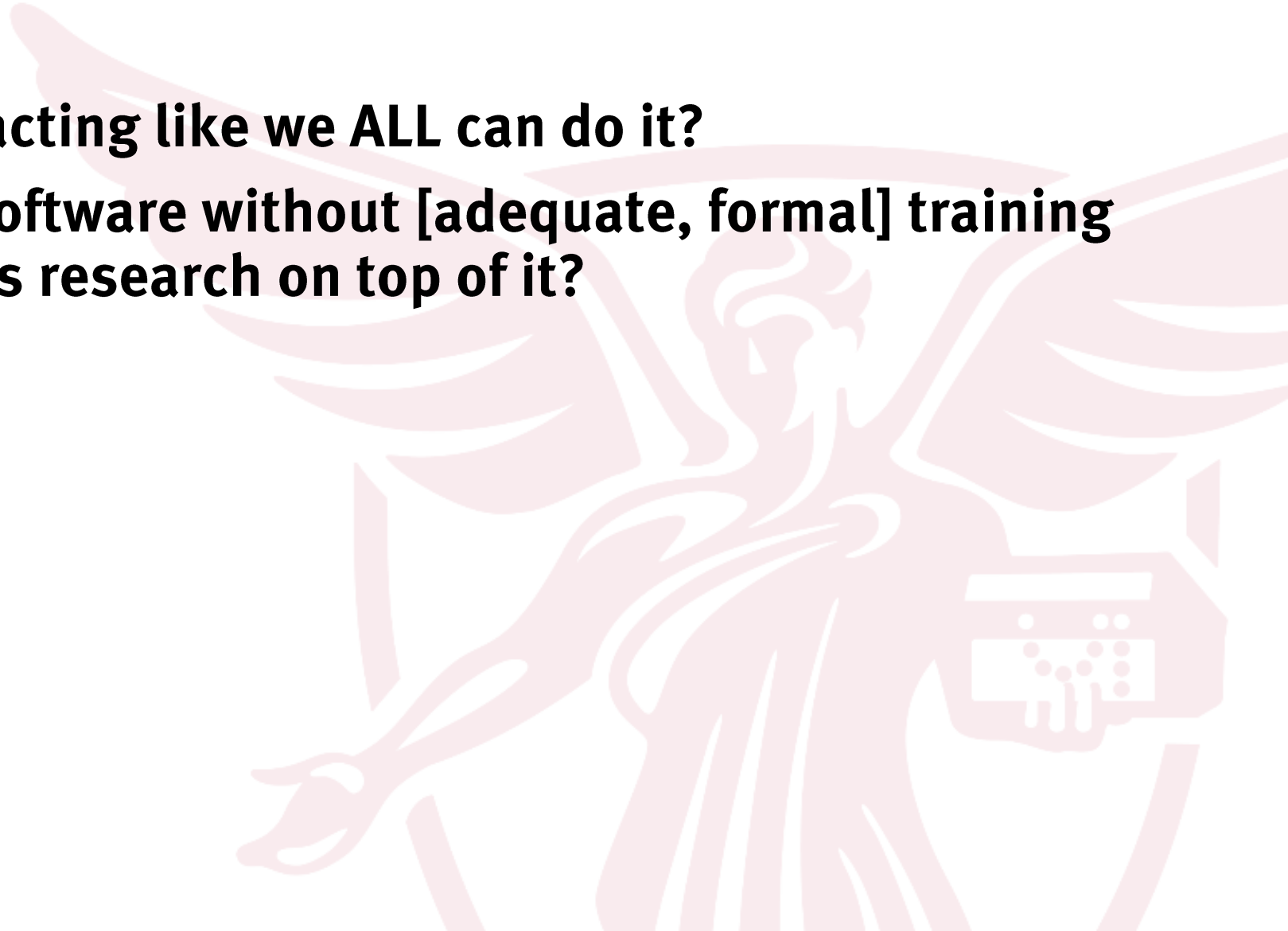
It is your turn:

- **Would you trust a nurse who self-trained on things he/she does?**



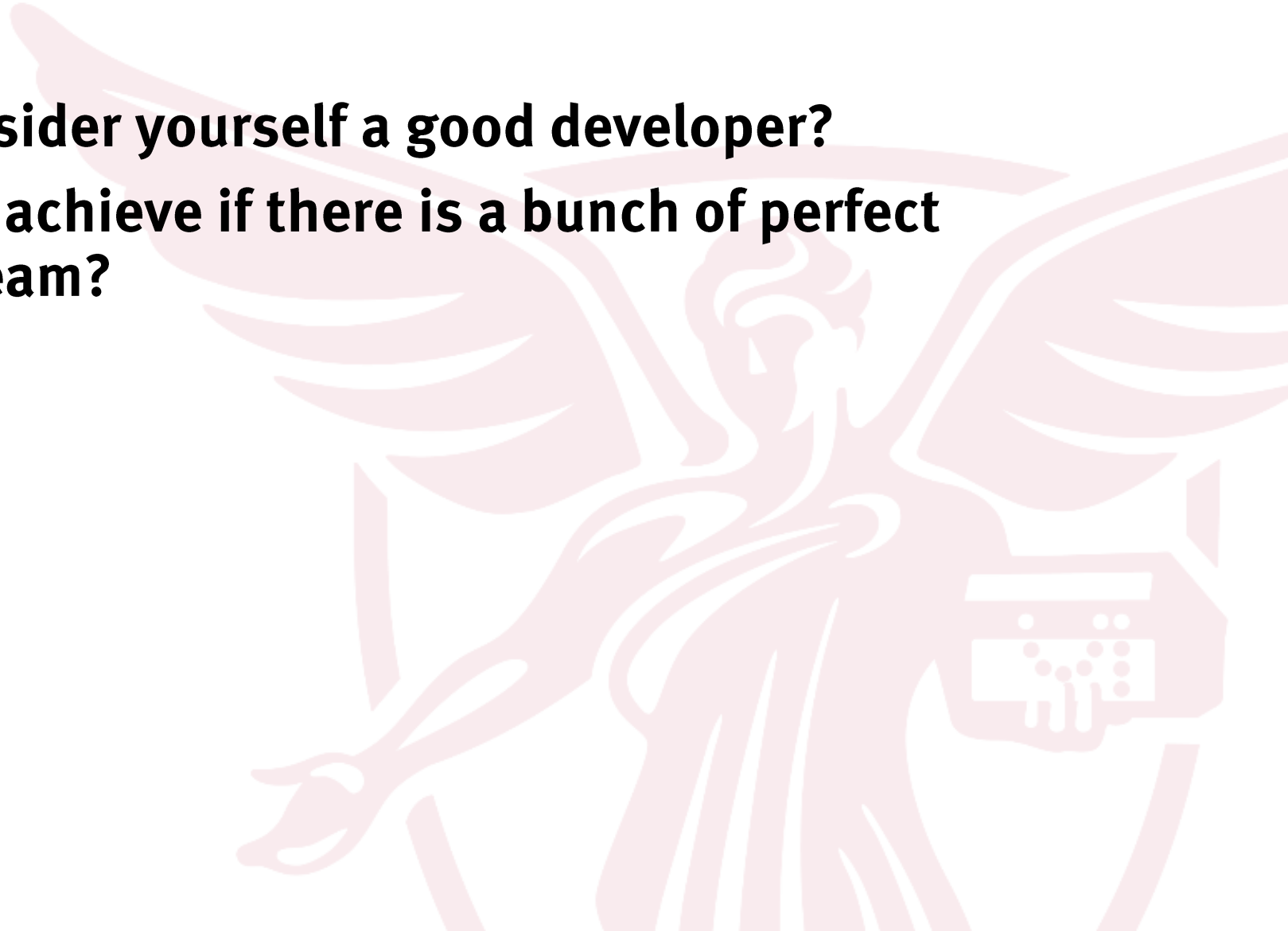
Why?

- **Why do we continue acting like we ALL can do it?**
- **Why do we produce software without [adequate, formal] training and build our precious research on top of it?**



Consider

- **How many of you consider yourself a good developer?**
- **What more could you achieve if there is a bunch of perfect developers on your team?**



Code policies

- **Stodden et al. analyzed 170 journals regarding code sharing and supplemental material policies and found out that during the consecutive years they did the study, the code adoption policy increased 30% by the journals.**

INFORMS JoC Software Policy

- How many of you have known there was a policy?
 - <https://pubsonline.informs.org/page/ijoc/softwarepolicy>

INFORMS Journal on Computing Software Policy

Like the **Data Policy**, the Software Policy does not apply to every paper and there will be numerous exceptions for good reasons.

For papers whose primary contribution is computational experiments, as a condition of final acceptance of the paper, software must be released to the research community to provide support for researchers to reproduce results given in the paper. It is generally expected that relevant source code will be published as an online supplement to the journal article.

Statistics

- Now, let's look at some statistics!



Meta: year

2017

A faint, light pink watermark of a winged figure, possibly an angel or a personification of a concept, is visible in the background. The figure has large, spread wings and is holding a staff or scepter. The watermark is centered behind the large text '2017'.

Meta: number of papers

37

A faint, light pink watermark of a winged figure, possibly an angel or a personification of knowledge, is visible in the background. The figure has large, feathered wings and is holding a scroll or tablet. The watermark is centered and partially obscured by the large number '37'.

Meta: collected data

Data availability

Code availability

Programming language

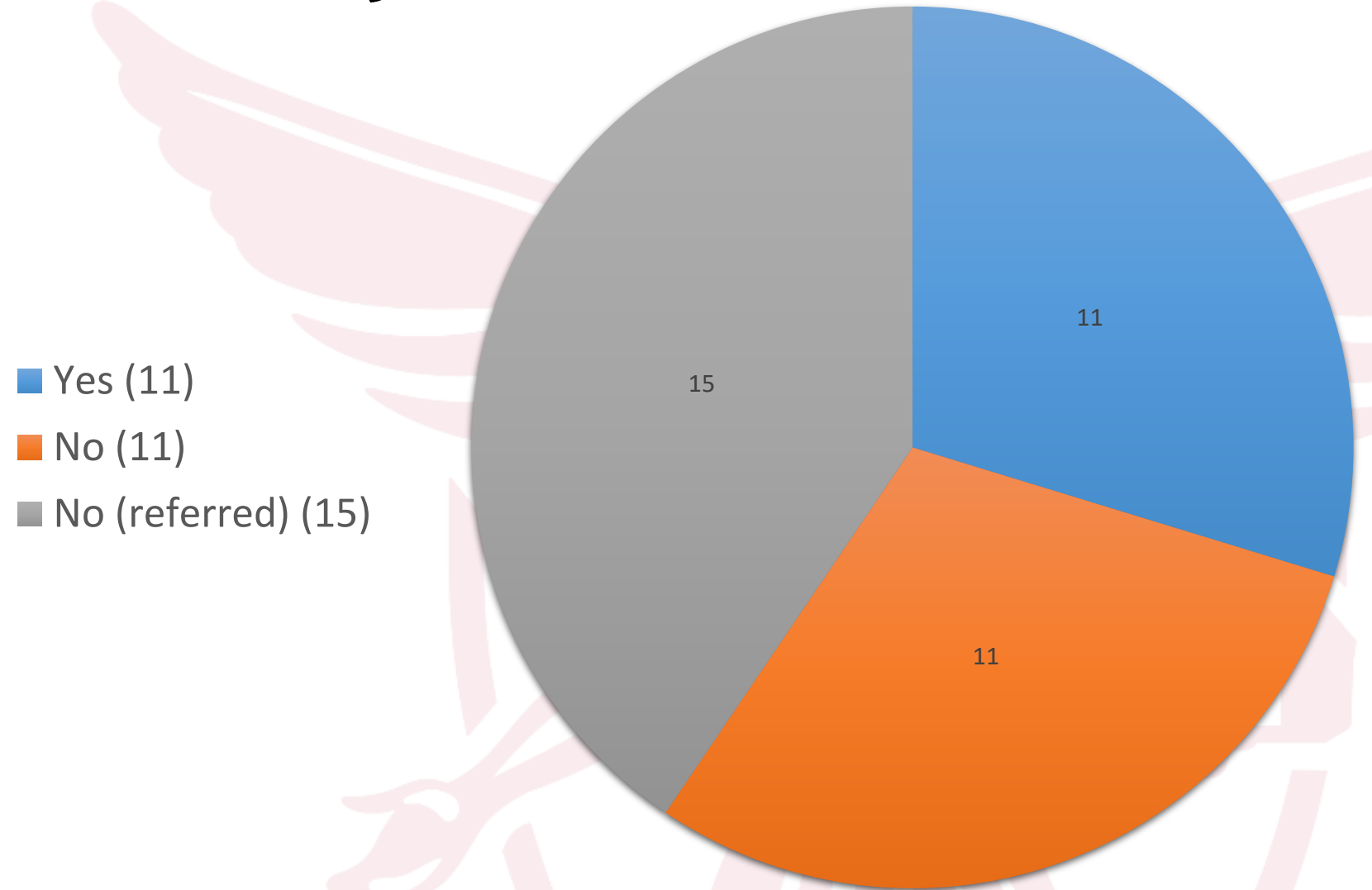
Pseudo-code

Comparison

Software practices

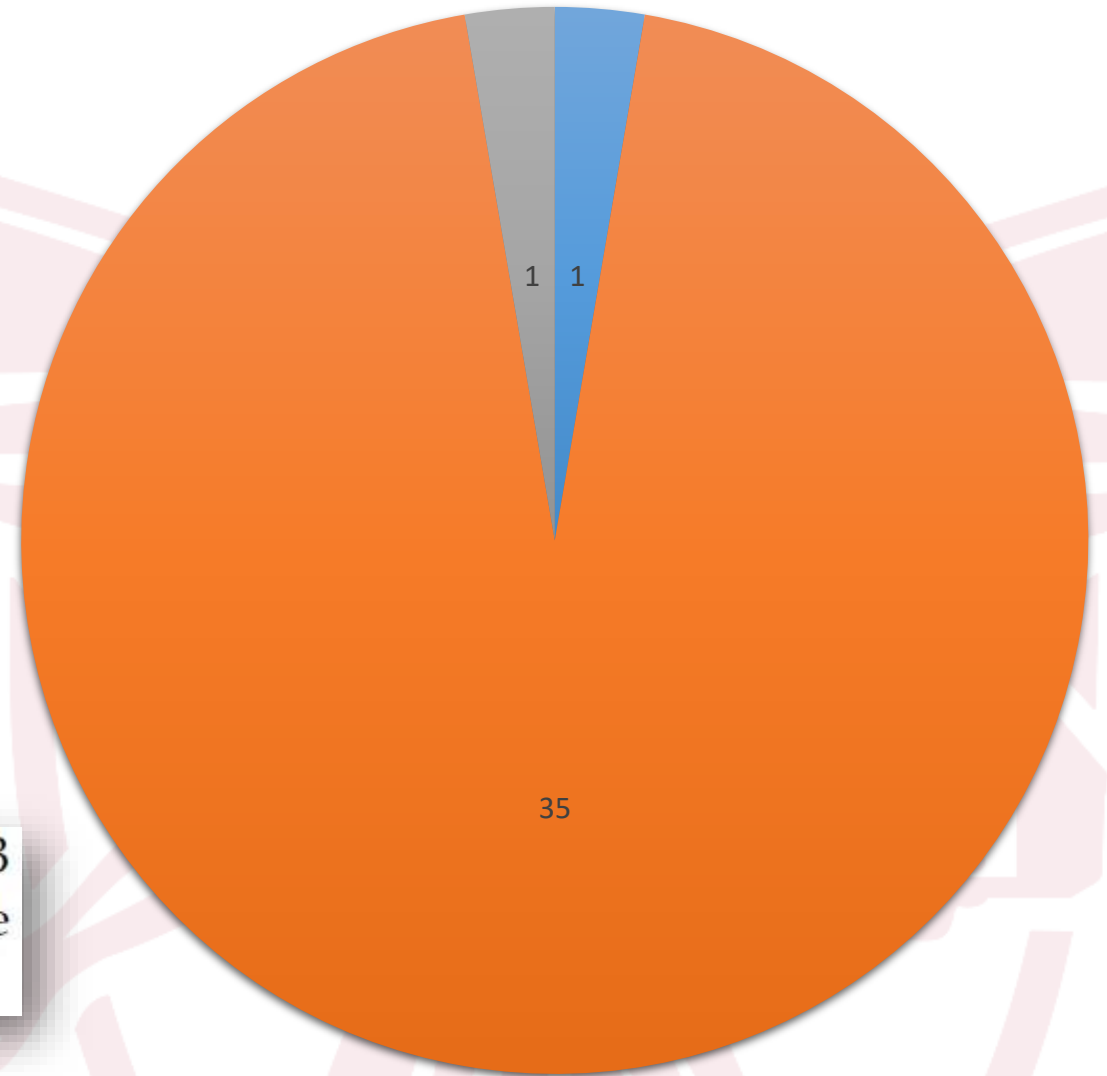
Technical information

Stat: data availability



Stat: code availability

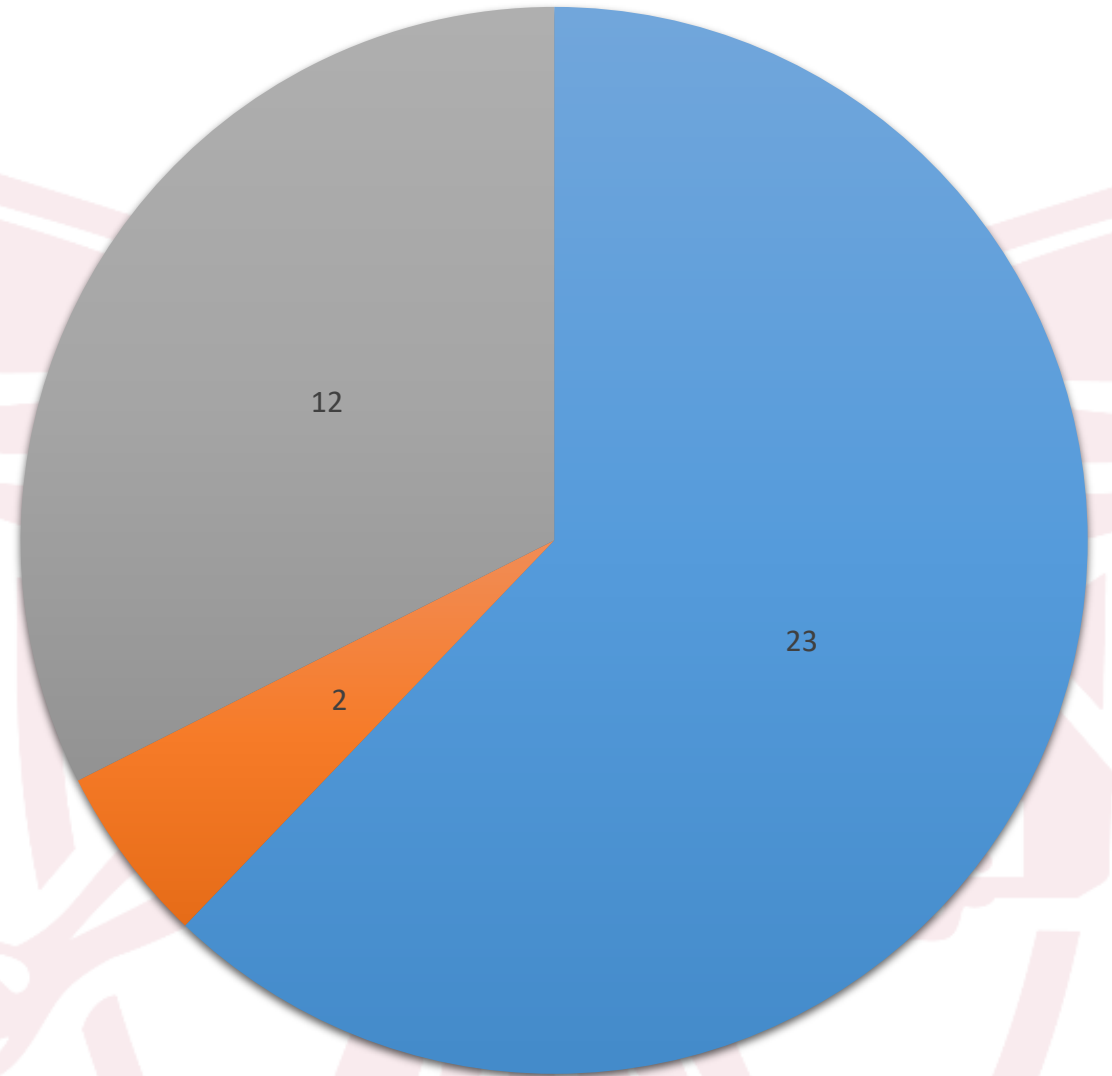
- Yes (1)
- No (35)
- No (promised) (1)



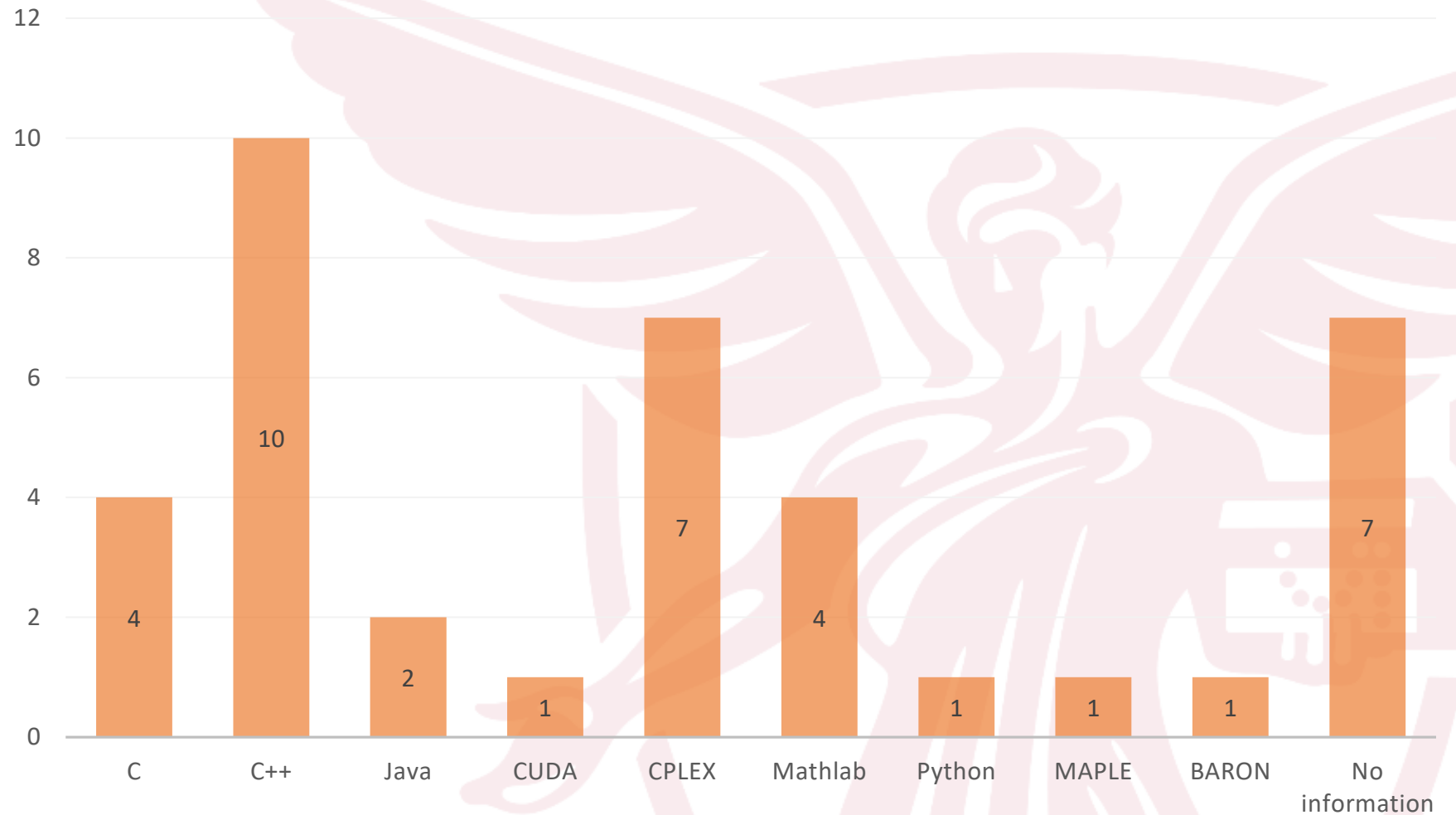
problems and further numerical results. A MATLAB implementation of the algorithm will be made available by the author.

Stat: pseudo-code

- Yes (23)
- Yes (explained in text) (2)
- No (12)



Stat: programming languages



Stat: papers mentioning software practices

0/37



Stat: computers

Intel Xeon E3 1,240 V2
processor at 3.40GHz and 24 GB
Ram (win7)

Intel Xeon E3-1220
3.1 GHz and 16 GB Ram

Intel Xeon X5650 2.67 GHz and
24 GB Ram (openSUSE)

3.1 GHz iMAC with 16 GB ram

Intel Core i5-2430M CPU, 16GB
Ram, (win7 64bit)

Intel Core2duo @1.65 GHz
with 8GB DDR2 RAM

Dell Precision T1650
workstation with 3.3. GHz intel
core i3-212- CPU, 3.7 GB Ram

64-bit Xeon X5650 2.66 GHz

Intel Xeon ES-2637 3.5 GHz with
128 GB Ram, Linux oracle server

Intel core 2.8 GHZ and 16 GB
Ram

Stat: papers with not computer info

7/37

A faint, light pink watermark of a winged figure, possibly an angel or a personification of knowledge, is visible in the background. The figure has large, feathered wings and is holding a scroll or tablet. The watermark is centered behind the large numbers.

You are not alone!

- **These problems exist in a lot of other fields!**



Some suggestions

- **More collaboration with domain experts**



Some suggestions

- **Modularity and unit testing**



Some suggestions

- Version control



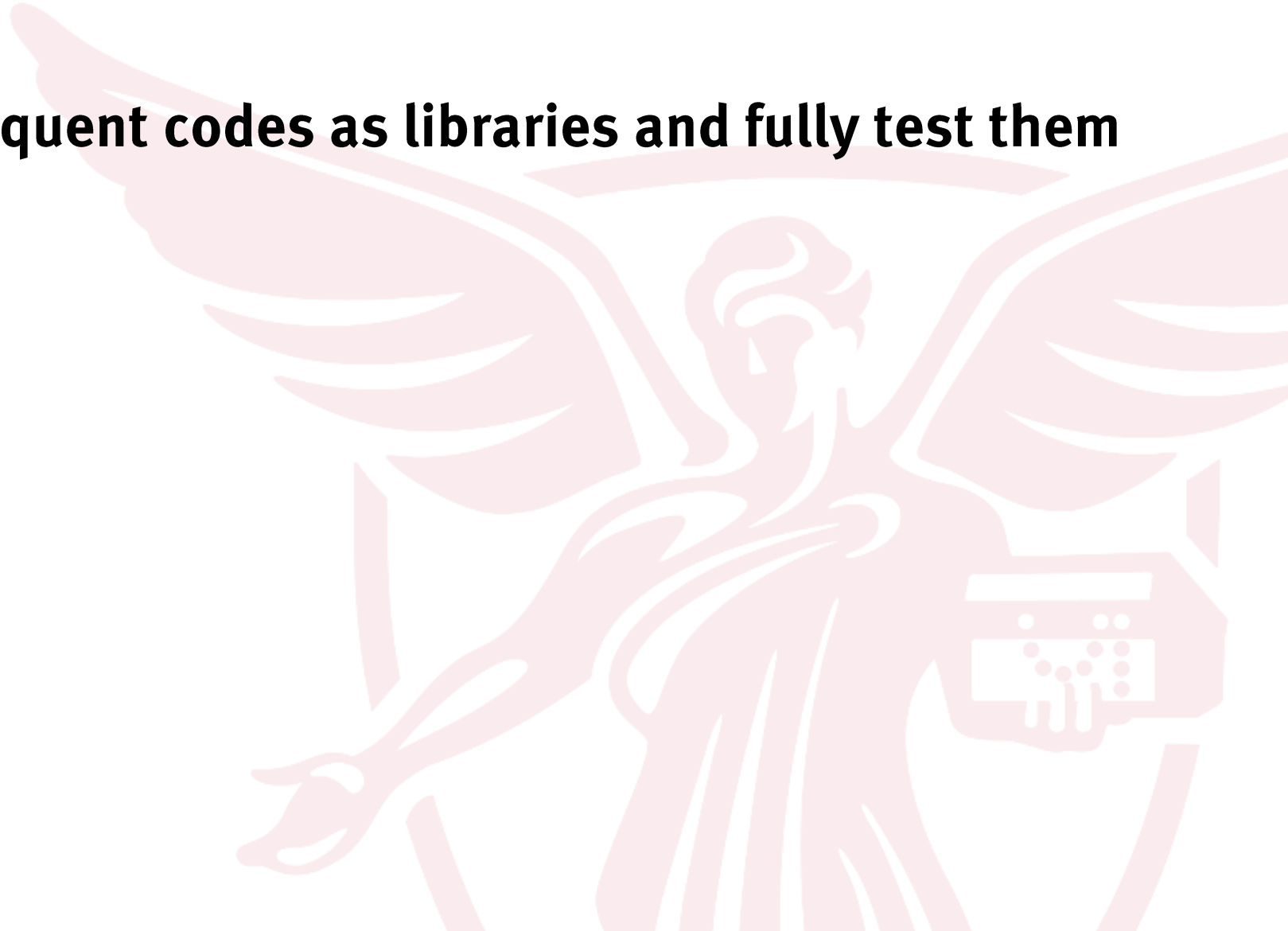
Some suggestions

- **Documentation while coding**



Some suggestions

- **Publish your most frequent codes as libraries and fully test them**



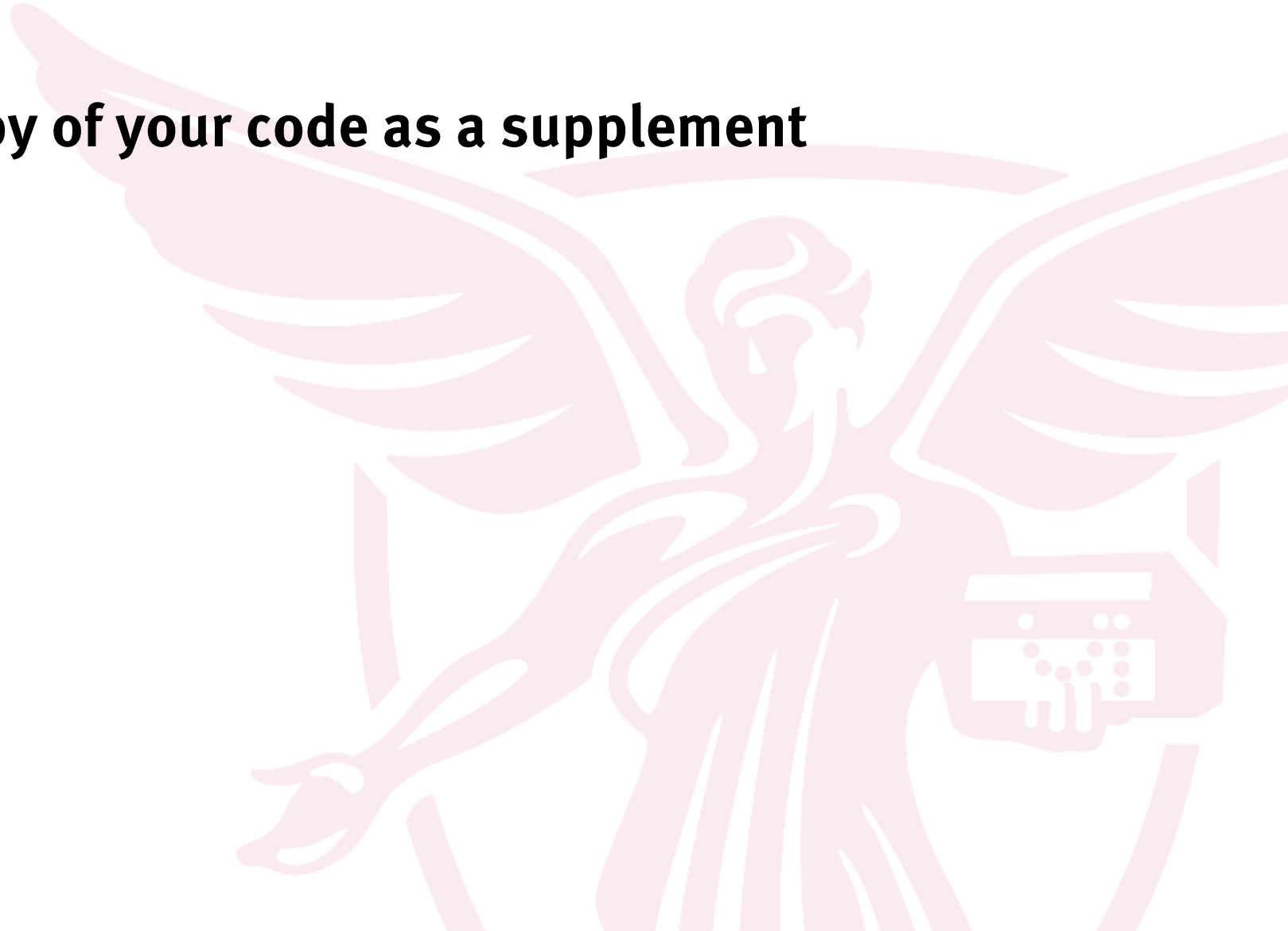
Some suggestions

- Include examples



Some suggestions

- **Put an immutable copy of your code as a supplement**



Some suggestions

- Code reviews



What's in it for you?

- Shorter development times



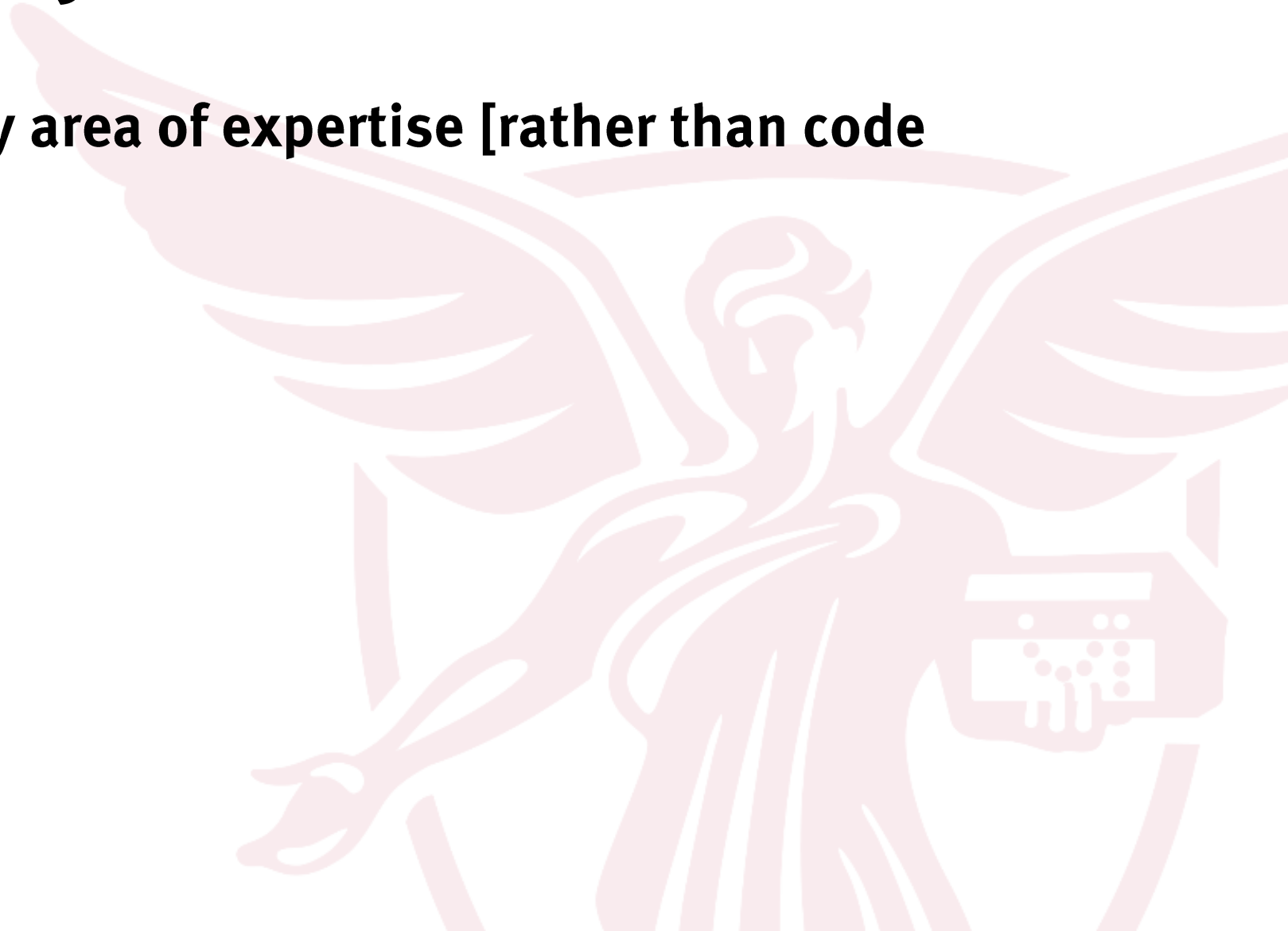
What's in it for you?

- Easily switch between algorithms on the same data



What's in it for you?

- **Focus on your primary area of expertise [rather than code development]**



What's next?

- We are planning to extend this study with active participation of authors.
- A survey in the upcoming weeks that wants to learn some insider information about YOU!
- We will distribute it through various channels.
 - Email us directly:
 - Huseyin Ergin: hergin@bsu.edu
 - Mesut Yavuz: myavuz@culverhouse.ua.edu