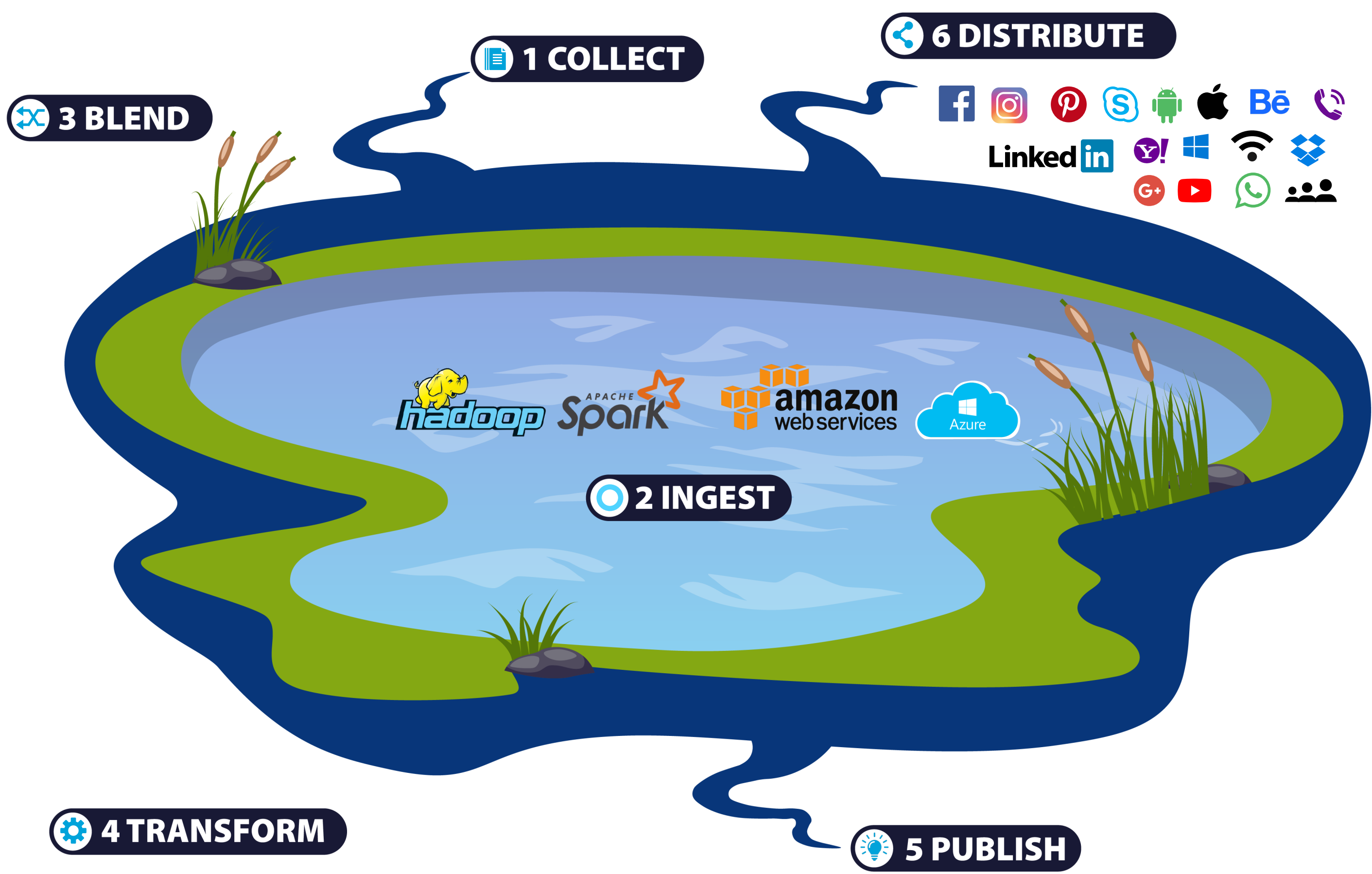


Cathal Gallagher
cathal.gallagher@instem.com

THE DATA LAKE

The first step of working on a data lake is finding information with which you can hydrate a lake. You cannot simply populate your lake with the clinical trial data stored on your servers. Patient data belongs to the patients and should only be used for the purpose it was collected. When working with data lakes and machine learning we are usually looking for new information.

Anonymisation will transform patient data into anonymous data. Anonymous data falls out of scope of GDPR. Best practice dictates filling data lakes with anonymous clinical data only.



EXAMPLE CODE: PROC DISCRIM

Hypothesis: Can we predict which patients are likely to have an adverse event based on demographic data and a subset of the adverse event data.

The code below "trained on 500" patients from an overall dataset of 1253.

Demographic data is coded previously into numeric variables for algorithm consumption.

Conclusion: Overall accuracy was 68.87% . There were 50 patients on this trial who actually had Serious adverse events. The algorithm correctly predicted 39 of these. This represents 78% correct predictions. If you can predict the likelihood of an adverse event, then precautions can be put in place to allow for better responses to these events.

```
2=proc surveyselect data=dm_ae_ser
3   method=srs n=500 out=train;
4 run;
5
6=
7=
8=
9=
10
11=proc discrim data=train
12      test=AE_Predict
13      testout=Predict_out
14      method=npair k=10
15      testlist;
16      class aesern;
17      var sexn age countrysn racen ethnicsn actarmcdn siteidn;
18 run;
19
```

MACHINE LEARNING



There are many free & licensed languages and tools to choose when writing and running Machine Learning algorithms on data. However, you need people, "Data Scientists" , who can wield the tools and understand the data mining objectives. These resources are in high demand and command large salaries which represents a significant investment for any company building out this capability.

Our industry already has a plethora of SAS programmers with vast experience in clinical trials. They may not have experience in Machine Learning algorithms, but they do know clinical data and are familiar with a powerful tool that can leverage Machine learning Algorithms. This presents an easy starting point where one can learn about machine learning and develop techniques and knowledge whilst using a familiar tool with familiar data structures. Offering new opportunities to existing employees increases employee retention and broadens their expertise.

NEXT STEPS

- As a first attempt these results are encouraging, but steps can be taken to increase the accuracy.
- Divide the data into training and (held-out)tests set(the test set should only be used once to evaluate the final model). The proportion should be around 80% train and 20% test.
 - Now you can divide the training data into a further dev and validation set to try out different optimisations (e.g., with/without scaling, feature selection etc.).
 - You would also want to present result on both your validation and test – to show that the model/algorithm generalises and is not over fitted.

CONCLUSION

The example above is a small experiment with a small set of data. Machine learning algorithms get better when they are exposed to large datasets of realistic data that allows the smoothing out of outliers ultimately resulting in more accurate predictions. The use of machine learning algorithms has not been broadly adopted in our space but could be instrumental to topics such as trial design and patient safety as demonstrated above. Many large sponsors are building data lakes today, have processes to anonymize the data, and have tools available to explore these concepts, but the real roadblock to leveraging machine learning is the investment in Data Scientists and commitment to this approach. I would encourage Sponsors to begin looking internally to visualise the potential value their data holds.