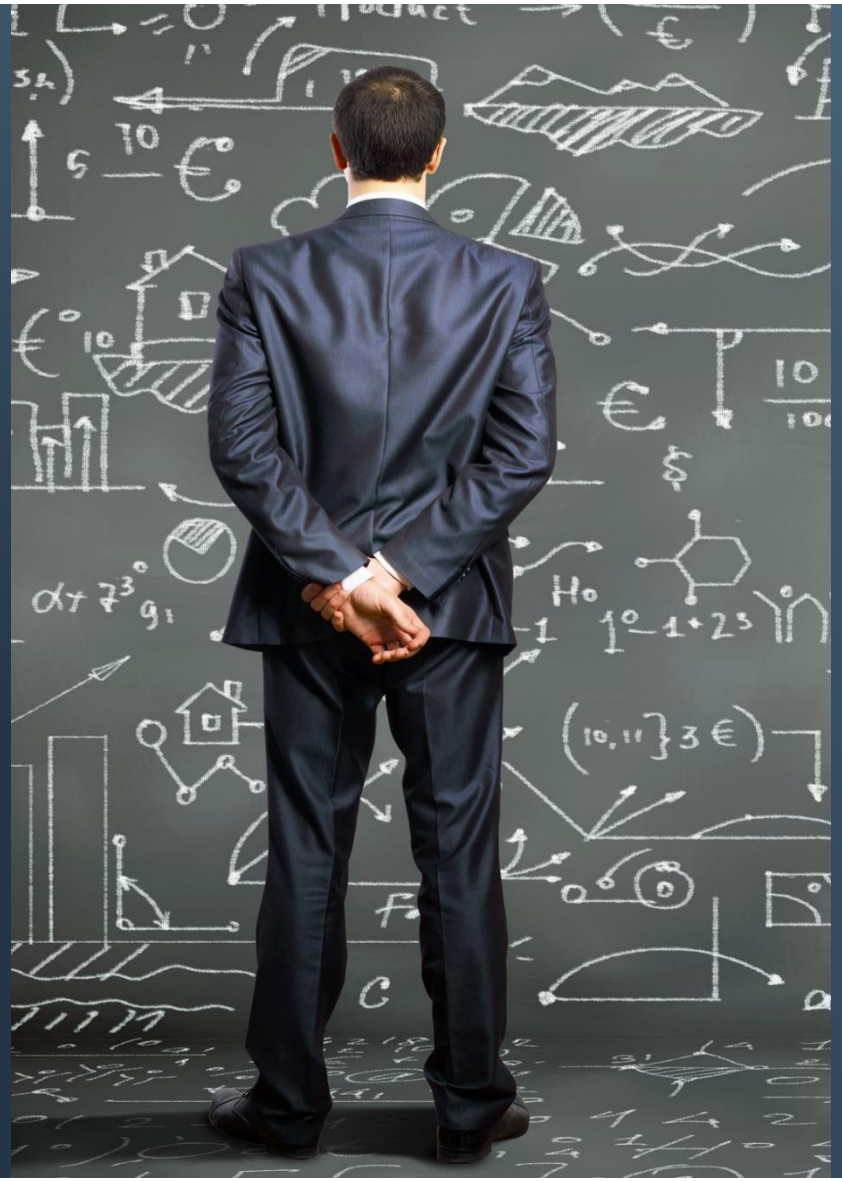


CHALLENGES IN PRACTICAL MACHINE LEARNING AND ADVANCED ANALYTICS

Sean P. McKenna

JULY 2017

OUTLINE

- **CAVEATS AND OBJECTIVE**
- **INTRODUCTION**
- **WHERE DO CHALLENGES STEM FROM**
- **CHALLENGES ENCOUNTERED ALONG THE PROGRESSION OF AN ANALYTICS PROJECT**
  - *Identifying the problem*
  - *Getting the data*
  - *Wrangling*
  - *Analysis*
  - *Evaluation*
  - *Presenting results*
- **QUESTIONS & COMMENTS**

SOME CAVEATS AND THE OBJECTIVE

- Not a quantitative talk; no code will appear in this talk
- A “lessons-learned” theme, but certainly, not exhaustive
- Mostly common sense, *but might not be as common as it should be*
- Practical, day-to-day challenges that I have come across
  - Predominantly, quick-turn projects from the world of tech consulting (“transactional”)
  - Viewed through my own lens (see next section, “Introduction”)
- The data science community has made analytics very accessible these days
  - Get some data, throw something like `caret` (R) or `tpot` (Python) at it, and voilá, you’re doing data science!
  - Just because you can swing a hammer, doesn’t mean you know what you’re doing

GOAL: Bring to the fore, the kinds of “gotchas” and challenges I’ve come across, in the hope that by doing so, you are able to avoid them, or at least recognize them

INTRODUCTION - WHO IS THIS GUY?

Education

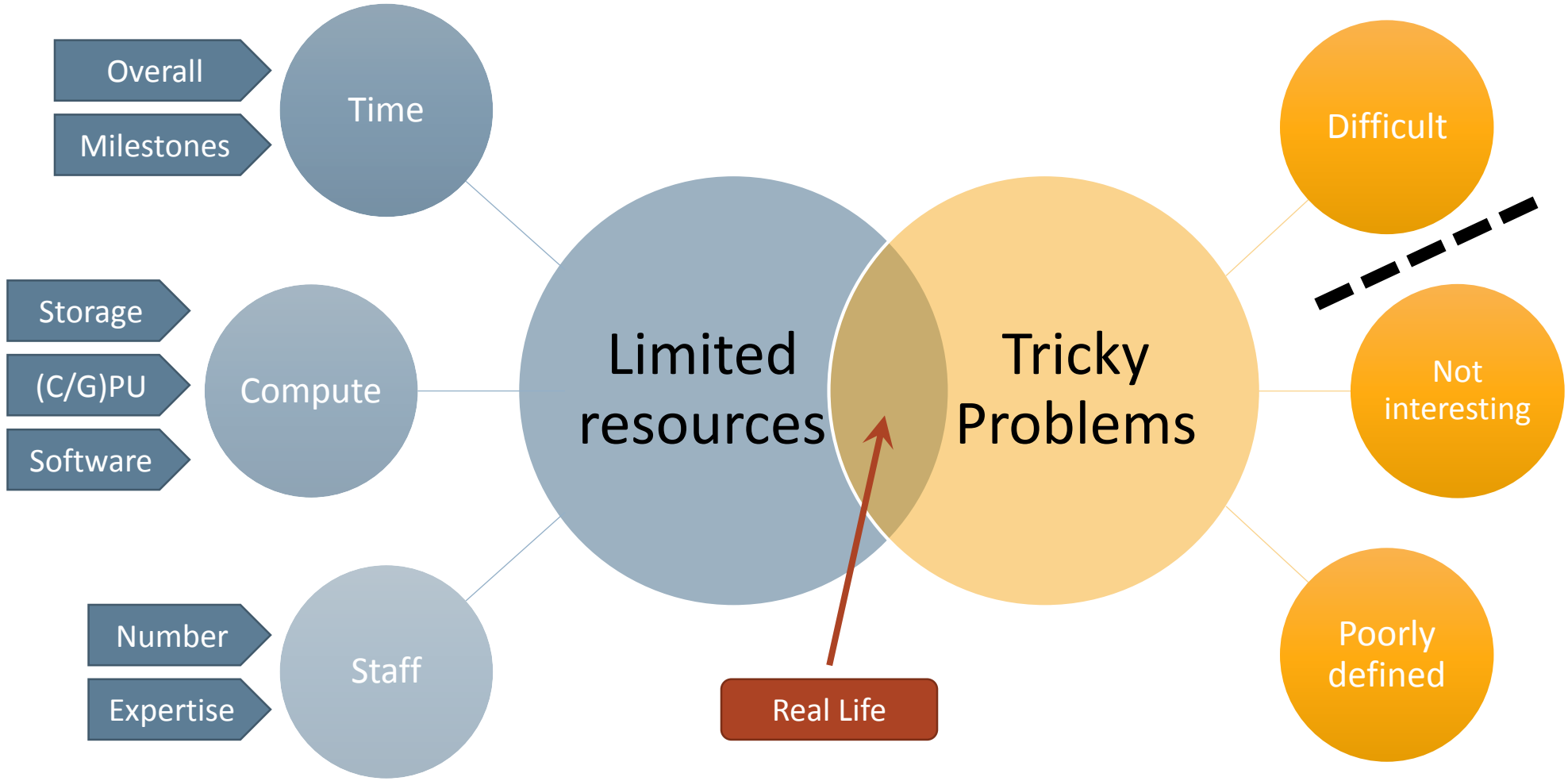
- BS in Mechanical Engineering, Rensselaer
  - Fluids, controls
- MS, PhD in Applied Ocean Physics, MIT/WHOI
  - Experimentation, analysis, hydrodynamics, turbulence, waves, air–water gas transfer, surface chemistry, flow measurement techniques (e.g., acoustic Doppler, laser Doppler, particle image velocimetry), optics/lasers, video, data acquisition, electronics, mechanical fabrication, …
  - C, Perl, MATLAB, HTML, Windows, Unix, Linux

INTRODUCTION - WHO IS THIS GUY?

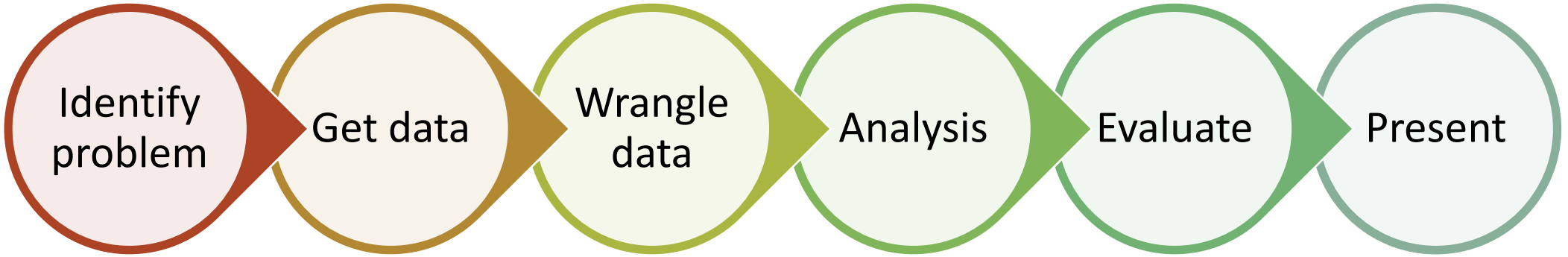
Experience

- DoD contractor/Special Assistant for Military Sensing and Analysis/Research Associate Professor
  - + {Modeling, simulation, aircraft, missiles, gravity, electromagnetics, geophysics, communication networks, software process, ship design/analysis, infrasound, target detection}
  - + {Visual Basic, FORTRAN}
- Coursera
  - Machine Learning (Ng), Introduction to Data Science (UWash)
  - + {Python, Tableau, SQL}
- Senior Analytical Scientist (Booz Allen)
  - + {Unsupervised and supervised machine learning, text analytics, Bayesian belief networks, sport analytics (prediction, clustering, classification, …), RaspberryPi, dynamic visualization}
  - + {R, JavaScript, D3.js, CSS, Hadoop/HDFS/MapReduce, AWS}

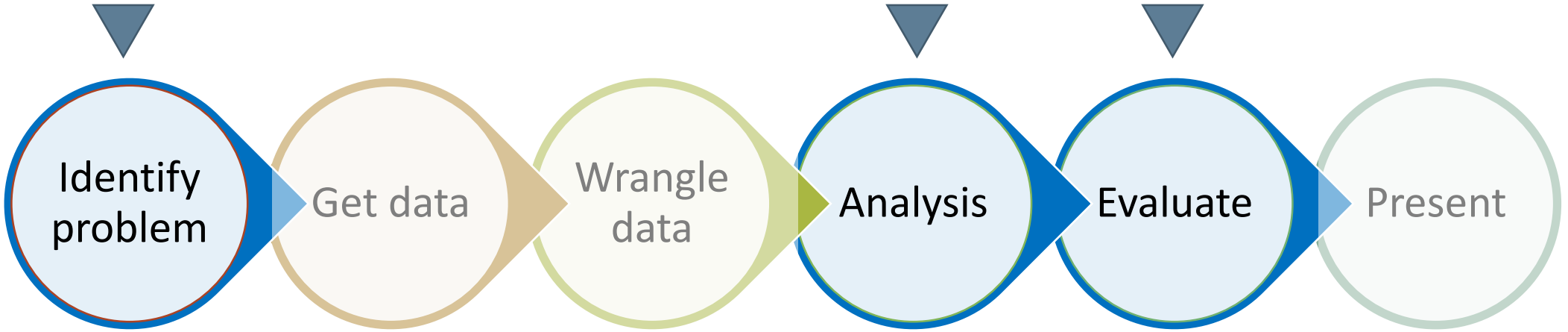
WHERE DO CHALLENGES COME FROM?



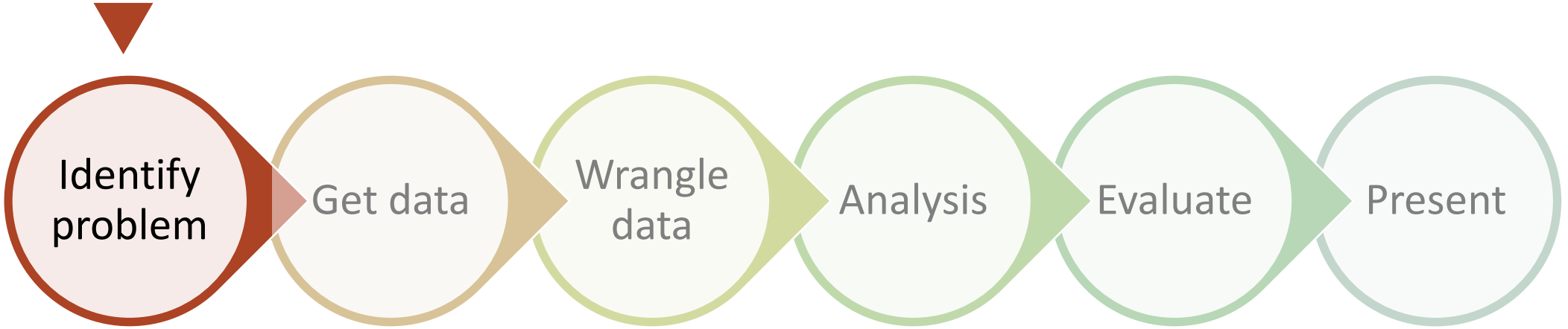
A SIMPLE PROGRESSION OF AN ANALYSIS PROJECT



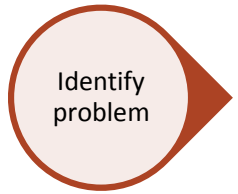
A SIMPLE PROGRESSION OF AN ANALYSIS PROJECT



THE PROGRESSION OF AN ANALYSIS PROJECT: PROBLEM DEFINITION



IT IS NOT UNCOMMON FOR AN ANALYSIS PROJECT TO LACK AN OBJECTIVE



- Consider the "fishing expedition" project:

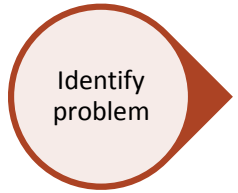
Here's our data – find insights in it!

- **Challenge:** Problem is poorly defined
- **Example:**

A firm is interested in understanding its employees' salary.

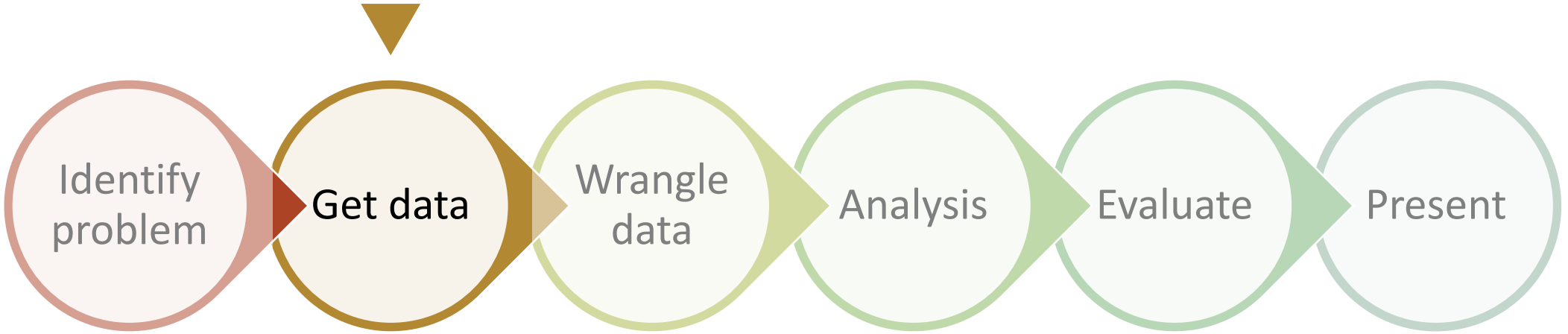
- The "Plan of Action" presented to us:
  1. Identify the scope of the pilot
  2. Assess current state of compensation within the firm and baseline
  3. Identify needed data and resources to execute pilot
  4. Baseline current state of data for the pilot and determine initial- and long-term requirements for increased automation and tracking of data
  5. Execute pilot
  6. Provide recommendation and estimated LOE and time to roll out firm-wide
- **What is the reason for understanding salary and what is the goal of the pilot?**

GET CLARITY ON WHY YOU'RE DOING WHAT YOU'RE DOING



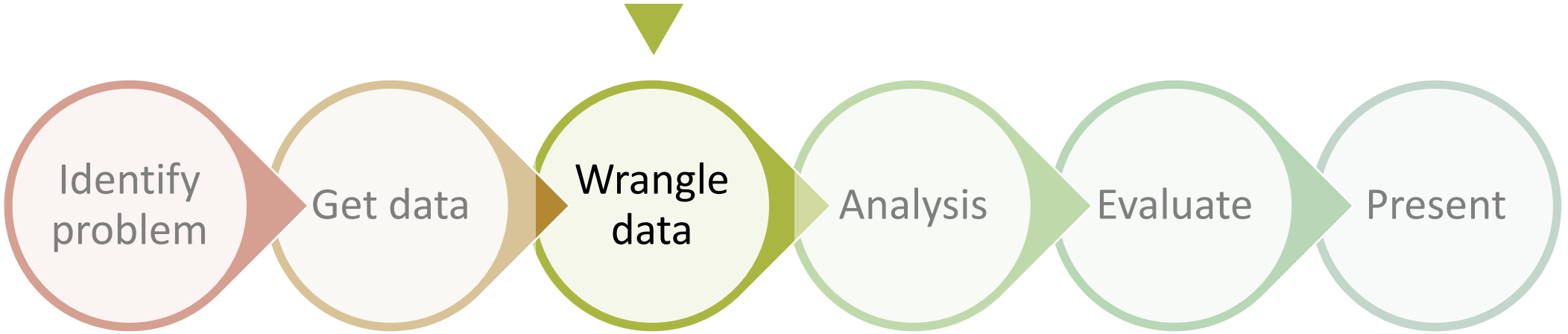
- Try to get stakeholder(s) to articulate what they want – they have to want something, e.g.,
  - More traffic to their site
  - More revenue
  - Lower costs
  - Improved performance of X
  - Better understanding of customers … why?
  - Find "bad things"
- Maintain greater communication with stakeholder(s)
- Manage expectations

THE PROGRESSION OF AN ANALYSIS PROJECT: GETTING THE DATA



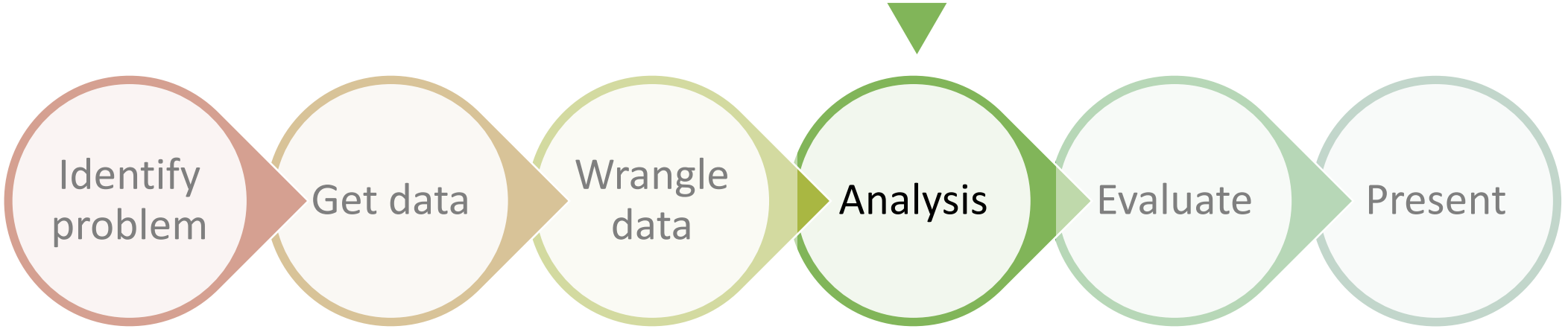
- Make sure the data actually has the power to do what you need it to do
  (Raw or through feature generation)

THE PROGRESSION OF AN ANALYSIS PROJECT: WRANGLING THE DATA



- Expect pathological cases.
- Be careful and verify.

THE PROGRESSION OF AN ANALYSIS PROJECT: ANALYZING THE DATA



SPEND MOST OF YOUR TIME *NOT* DOING MACHINE LEARNING

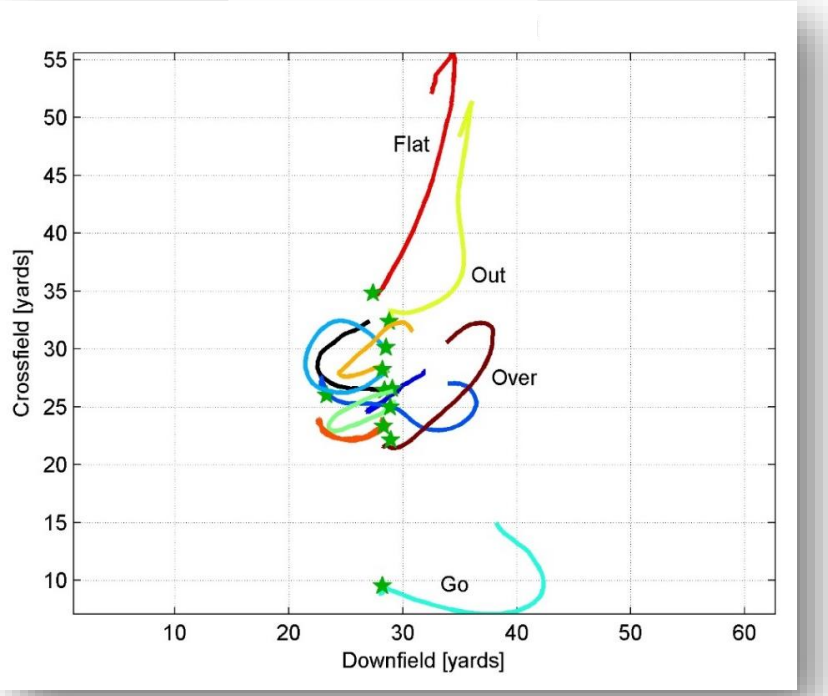
"If I had an hour to solve a problem, I'd spend 55 minutes thinking about the problem and 5 minutes thinking about solutions." – Albert Einstein

- **Challenge:** Avoid building models too soon
- **Example:**

Using Next Gen Stats data from an NFL team, we were to develop an algorithm-based solution that automatically labeled receiver routes

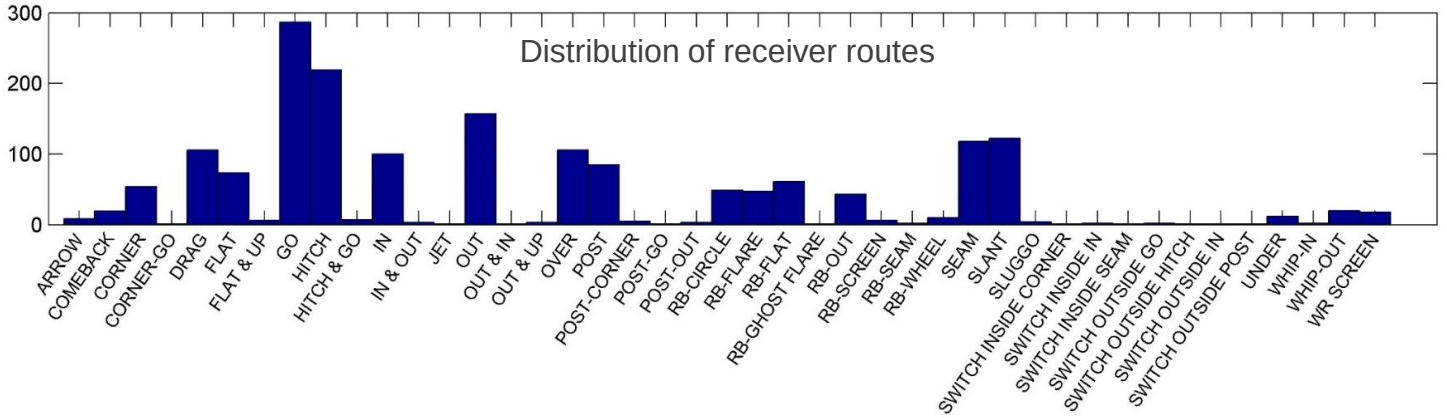
- The data contained 1,769 routes and 43 labels
- Project was planned to run for about 6 weeks
- Within about 2 weeks, we were building models and, in some ways, we never looked back
- **Several weeks' of hammering many different machine learning approaches at the problem did not yield a commensurate increase in accuracy; we had hit a performance limit early on**
- In the end, we developed a set of six different classification schemes that were fused to obtain our final outputs

Individual play example with labels



Analysis

KNOW YOUR DATA AND EXERCISE PATIENCE



- Invest time in exploratory data analysis
  - Identify outliers
  - Understand bounds
  - Identify correlations
- Think hard about ways to tackle the problem
  - Is there a simple way that will suffice
  - Make components modular
- Spend more time thinking about features
  - Not every problem is amenable to deep learning (among other things, you need a lot of data!)
- Once you build a model, test it, and go right back to the beginning
  - In our case of the route-labeling problem, examining where the misclassifications were occurring could have helped identify the kinds of features that would yield improvement and prevent us hitting the accuracy wall
- While most of the time these things are difficult to do,
  - Self-imposed pressure
  - External pressure,

  the more you can adopt these guidelines, the better

TRY TO AVOID OVER-COMPLICATING THINGS

"*Simplicity is achieved by resisting the urge to make things complex.*" – James Gosling

- **Challenge:** Avoid the temptation to use elaborate analysis techniques as a way to show how smart you are
- **Example**

An MLB team wanted to improve their Rule 4 Draft process and outcomes.

- The analysis team decided on two approaches: (1) unsupervised clustering and (2) train a model to predict prospect success in Minor League Baseball (MiLB)
- A key aspect of this project was defining "success in MiLB"
- Team member A explored using at-bats (ABs) in the different MiLB levels as a way to segment all hitters into 2 populations ("favorable" and "unfavorable")
- To do this, he performed a principal component analysis (PCA), followed by clustering step (model-based clustering using finite normal mixture modelling, i.e., `mclust` in R)
- Team member B reproduced A's results with a couple if-then statements looking at the ABs in the last three levels
- Team member A realized a flaw in his thinking – in the end, we used his modified approach

DON'T DISCOUNT THE POWER OF SIMPLE HEURISTICS

- "*Le mieux est l'ennemi du bien (The better is enemy of the good).*" – Voltaire
- **Challenge:** Look for ways to get a decent answer quickly when a great answer is out of reach
- **Example:**

As part of the route-labeling project, a critical task was determining the time when a route ended (end-of-route, EOR)

- Team member A came up with an elegant, data-driven methodology that used an iterative process to develop a predictive model for the best EOR in terms of classifier performance
  - Unfortunately, this approach was highly computationally intensive and not practical given various constraints
- As an alternative, Team member B proposed a heuristic approach
  - This scheme used times that were available in the data (snap, pass, catch) to estimate EOR; in cases where such time were missing, the scheme used a simple model to estimate the EOR
  - **This approach was easily implemented and provided a sufficiently accurate result**

BRING THE ANALYTICS HEAT ONLY WHEN IT'S WARRANTED

- Complexity often adds code, room for error, and confusion
- If you use advanced analytics, make sure there is a good reason
- If getting the "best" answer exceeds available resources, it may not be the best approach

NOT ALL INSIGHTS ARE ACTIONABLE

- *"A large portion of analytical insights are nothing but water-cooler trivia."* – Sean McKenna
- **Challenge:** Make sure the insights you uncover are relevant and address the project objective
- **Example:**

During a project that looked at PGA ShotLink data, and also on the current salary project, we uncovered a lot of insights, none of which were/are particularly useful

- For example, we've found that the average employee's salary is X% below the market median for their role

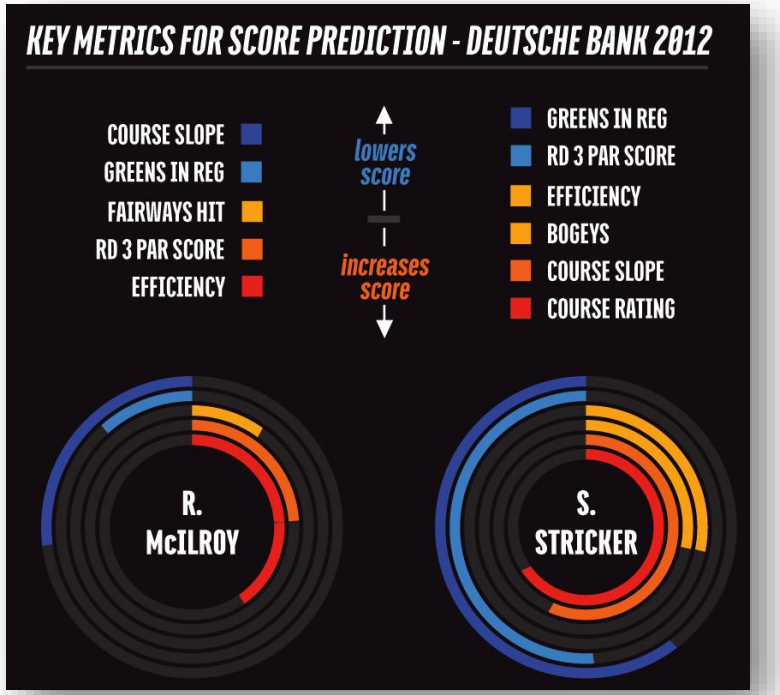
  "OK, now what? Give everyone an X% raise?"

- Similarly, we could tell you how many times Rory McIlroy bogeyed on each day of every 2012 tournament in which he played.

  "Wow, that's fascinating!"

- And so on …

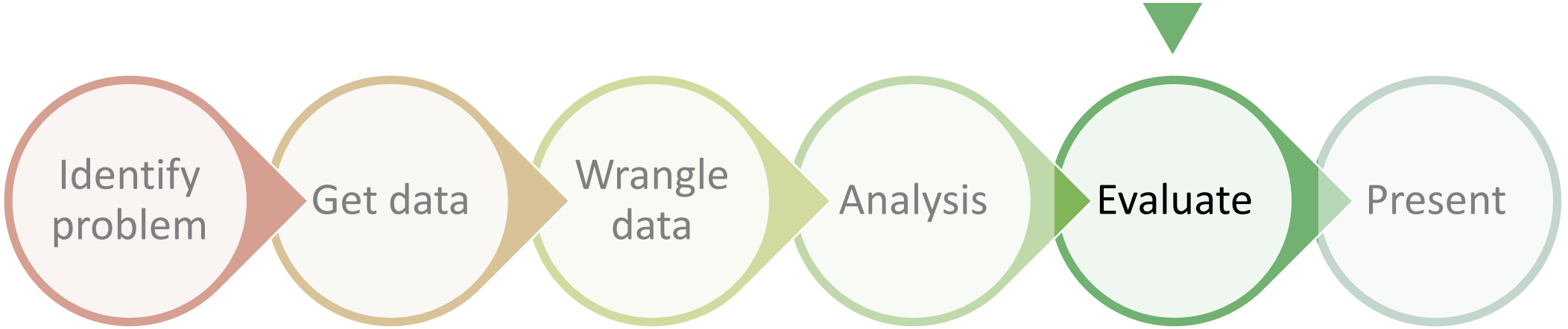
Notional PGA player dashboard



WATER COOLER INSIGHTS ARE FUN BUT NOT VERY USEFUL

- Trivia is great for Jeopardy, not great for data science
- As interesting as a result may be, if someone can’t use it, it may not be of any value
- This is why having a defined problem or objective is so important

THE PROGRESSION OF AN ANALYSIS PROJECT: EVALUATE YOUR RESULTS



IS YOUR MODEL BETTER THAN THE RELEVANT BENCHMARK

- "*Prediction is very difficult, especially about the future.*" – Neils Bohr
- **Challenge:** Coming up with a model that either exceeds a benchmark, or meets a benchmark, but does so at scale or at speed
- **Example:**

One of our initial forays into sports analytics was developing an model to predict whether an NFL offense would pass or run, and then, what type of pass or what direction of run

- Consider a benchmark for pass/run prediction accuracy:
  - Teams tend to pass about 60% of the time – predict pass every time, and you're at 60% accuracy
  - The fan on the couch eating chips can probably increase that by 10 to 20% – now, you're at 70 – 80%
- Can you develop an algorithm that beats that?
- We ended up with a mean accuracy over all games in a single season of 73%, with a best-game accuracy of 83% and a worst-game accuracy of 61%
- Results for type-of-pass or direction-of-run were not as good – about 58%; the data we had simply lacked the power to adequately make those more detailed predictions

ACCURACY IS NOT NECESSARILY A GREAT MEASURE OF PERFORMANCE

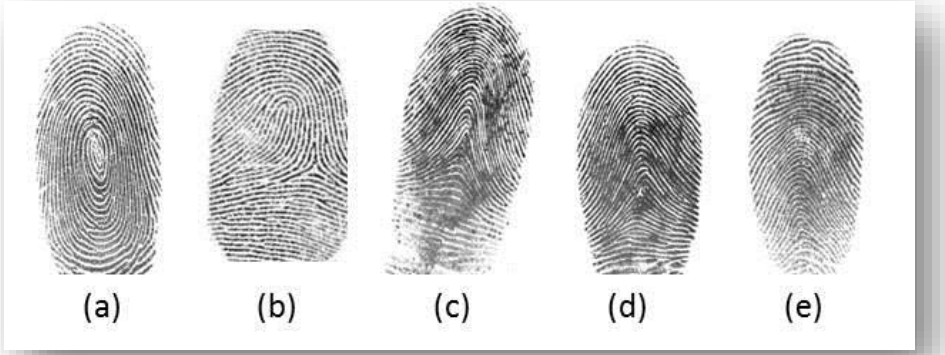
- "*Even a broken clock is right twice a day.*" – Marie von Ebner-Eschenbach
- **Challenge:** Properly quantifying the performance of your model – in particular, classification models against unbalanced data
- **Example:**

Quantifying performance of an anomaly detection scheme

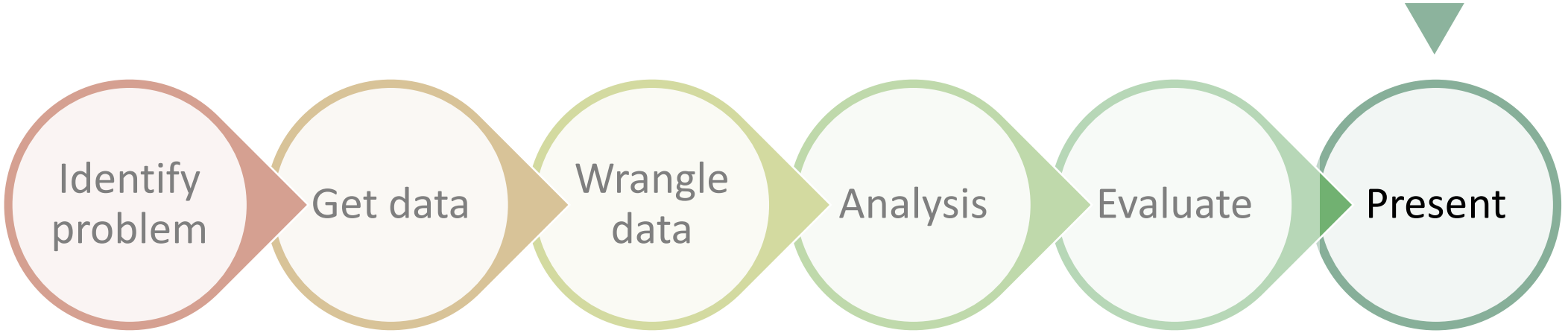
- You're looking to identify spoofed fingerprints in a database of N prints
- Roughly 3% of the prints are believed to be spoofs (high)
- You build your model and measure its accuracy:

  Σ(pred == truth)/N
- You get 94.1% – awesome!
  - It turns out, you predicted the right number of spoofs (i.e., 0.03×N), but your algorithm was as good as random
- This is a trivial example, but the same sort of thing can happen as the problems get more complicated and recognizing this mistake is much harder

Sample fingerprint images – one is synthetic



THE PROGRESSION OF AN ANALYSIS PROJECT: PRESENT YOUR FINDINGS



"If you can't explain it simply, you don't understand it well enough." – Albert Einstein

QUESTIONS? COMMENTS?

THANK YOU

ABSTRACT

This talk will focus on the practical, day-to-day challenges encountered when dealing with a variety of problems involving machine learning and other advanced analytical techniques. Using several real-world examples, we'll talk about the origins of such challenges - why them seem to be unavoidable - and things to consider when facing them. The content should be relevant to a general audience, from those with deep analytical expertise to the casual analyst, since many of the themes span the spectrum of analysis. The aim of the presentation will be to illustrate that, while the recent advances in machine learning have been impressive, successfully tackling today's problems still requires a savvy analyst.

ACCURACY EXAMPLE

10,000

300 true spoofs

Prediction:

9405 not-spoof correct

5 spoof correct (lucky)

9410 correct total

9405	295
295	5