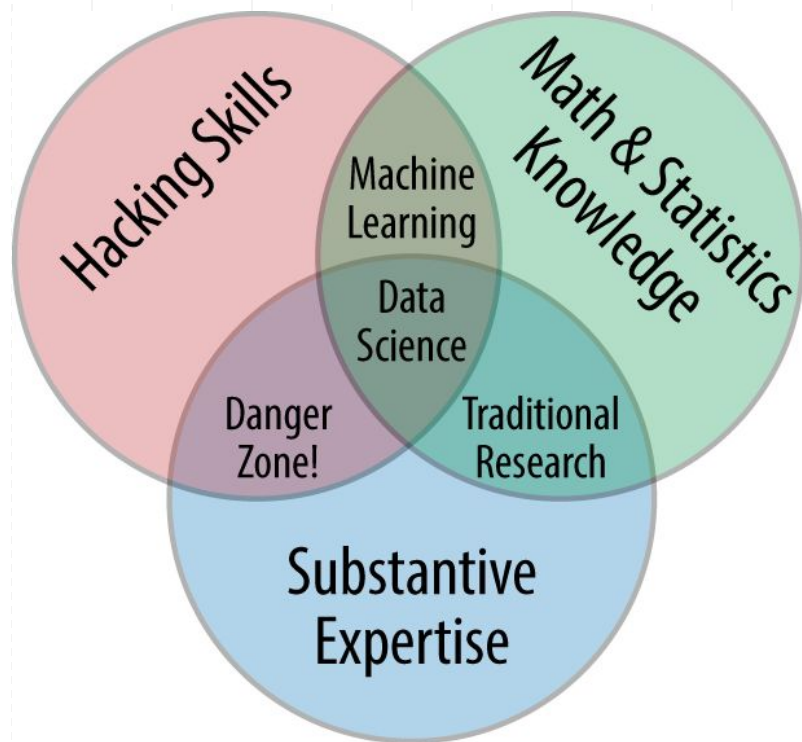# Business Skills for Data Scientists

Liz Sander, Tech Lead, Data Science
Civis Analytics

Drew Conway's Data Science Venn Diagram

Working efficiently?

Project planning?

Prioritization?

Collaboration?

Drew Conway's Data Science Venn Diagram

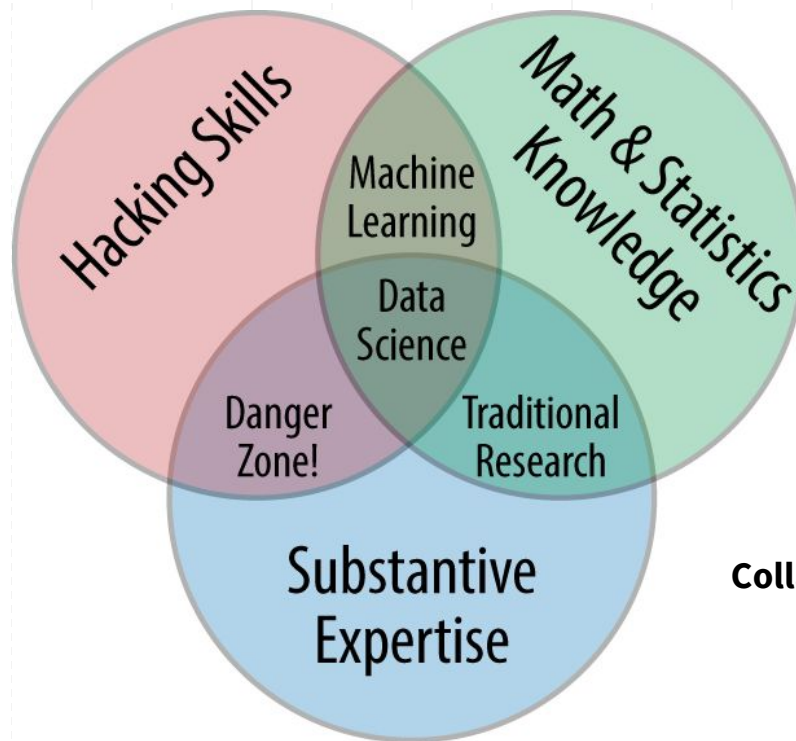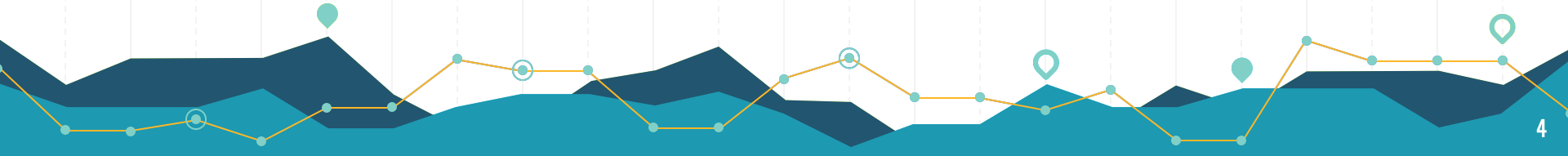*Technical skills help you build a model. Business skills ensure it gets used.*
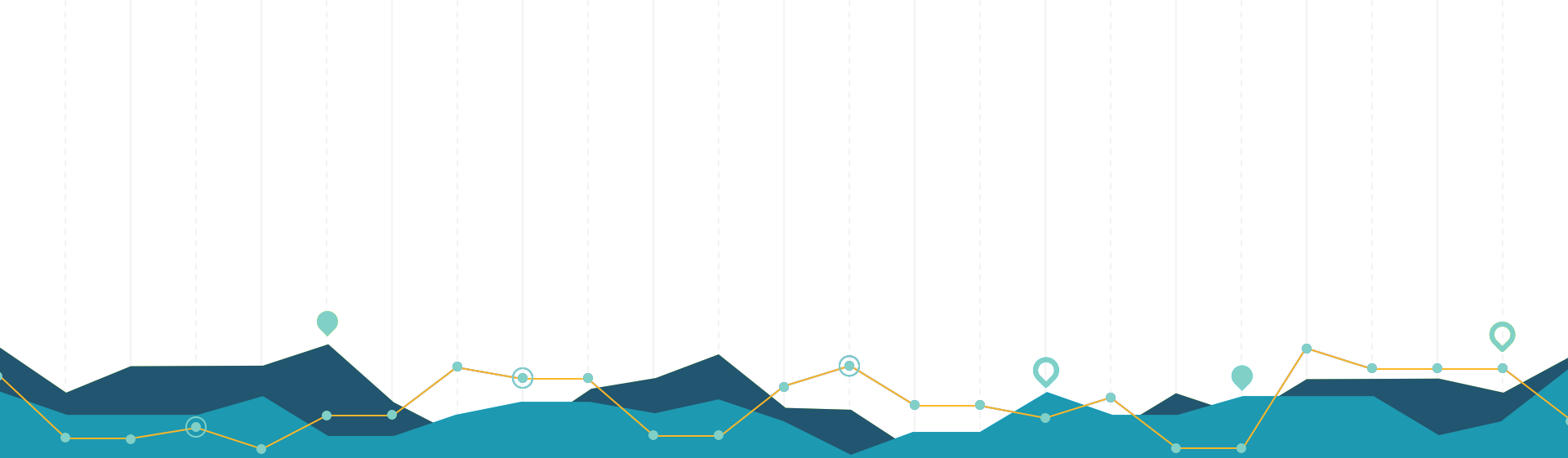
BUSINESS SKILLS FOR DATA SCIENTISTS

Project planning: Solve the right problem

Prioritization: Estimate impact upfront

Working Efficiently: Find the Minimum Viable Method

Collaboration: Be pseudo-agile
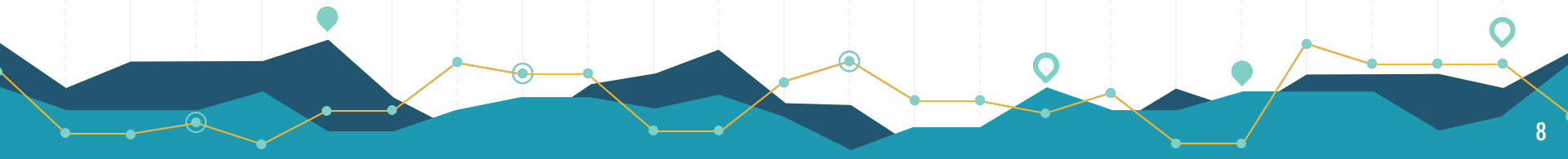
Project Planning

Solve the right problem

1

"

If I had asked people what they wanted, they would have said faster horses.

Henry Ford, maybe?

BAD NEWS
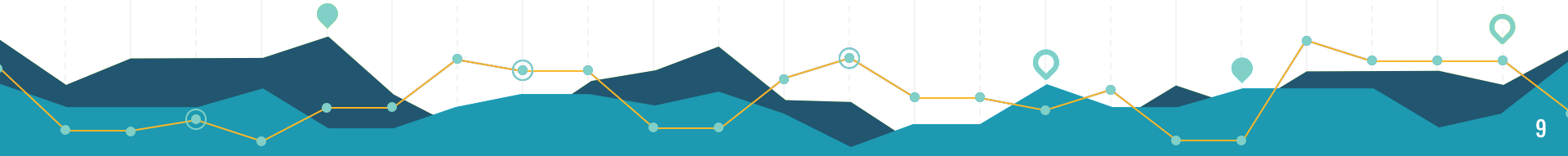
No one understands your model.

BAD NEWS 2

When people ask for solutions, it is their best guess based on limited knowledge.

DON'T BUILD A MODEL, SOLVE A PROBLEM

- Be curious about your user
- Learn about the big picture
- Think about the problem without the current solution

CASE STUDY: SEGMENTATION

- Segmentation app for a client
  - hierarchical model
  - Provided diagnostics and clusters
- Client felt that certain features were underweighted and wanted flexibility to upweight them
- Worked through the math and found that this would have **little impact** on the clusters

CASE STUDY: SEGMENTATION
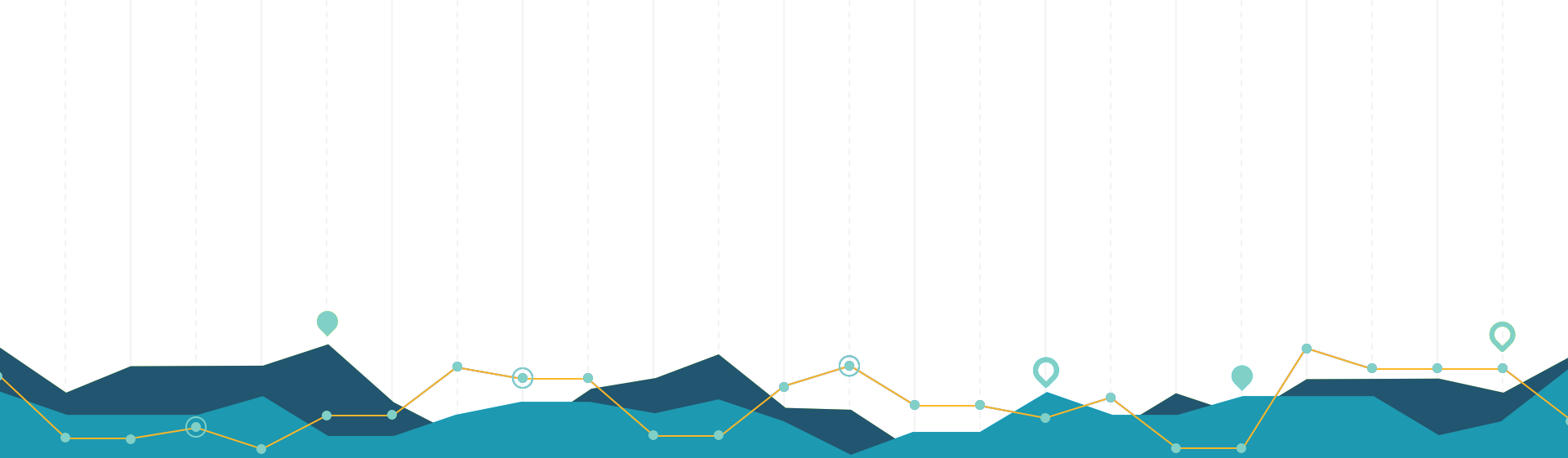
So we talked to the users:

- ◉   How do you use the app?
- ◉   How do you decide if a cluster is good or bad?
- ◉   How do you decide if a feature is "underweighted"?

It turned out that the client used entirely different diagnostics outside of the app! We were solving the **wrong problem**.

CASE STUDY: SEGMENTATION

We rebuilt the app, using a supervised model that optimized for the client's definition of a good cluster.

✓ Positive response from the client

🎯 No requests for methodological changes since release

Prioritization

Estimate Impact Upfront

2

WHY ESTIMATE IMPACT?

Clarify your **priorities**

Spend your **time** well

Explain your choices to stakeholders

Estimating impact requires **context.** It should be a conversation that includes **data scientists, product managers, and stakeholders.**

15

STRATEGY 1: WRITE A PITCH

Problem: Bad survey takers cost money and make estimates worse.

Solution: Prototype model based on attention checks in the survey.

Value: $X saved in survey sample, evidence of bias reduction

Effort (2, 6, or 20 weeks?): M (6 weeks)

STRATEGY 2: STACK RANK

1. Gather a backlog
2. Estimate effort
3. Rate 1-1000
   a. **Marginal improvement**
4. Stack rank
5. Refresh regularly

Project	Effort	Priority
Model bad survey takers	M	850
Document survey design standards	S	700
Update Census data	M	600
Research: new survey weighting method	L	300

STRATEGY 3: RICE

| Project	Reach	Impact	Certainty	Effort	Score
Model bad survey takers	9	8	7	6	84
Document survey design standards	7	7	9	2	220.5
Update Census data	7	5	10	5	70
Research: new survey weighting method	5	6	3	8	11.25

THE LIMITS OF RICE

- Favors low effort, high certainty work
- Transformative ideas in data science are always speculative
- Keep a diverse portfolio of projects in your roadmap

THE OTHER SIDE OF IMPACT
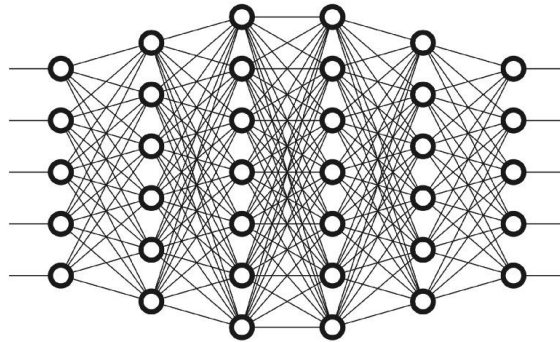
- Security
- Fairness
- Reducing risk of being wrong
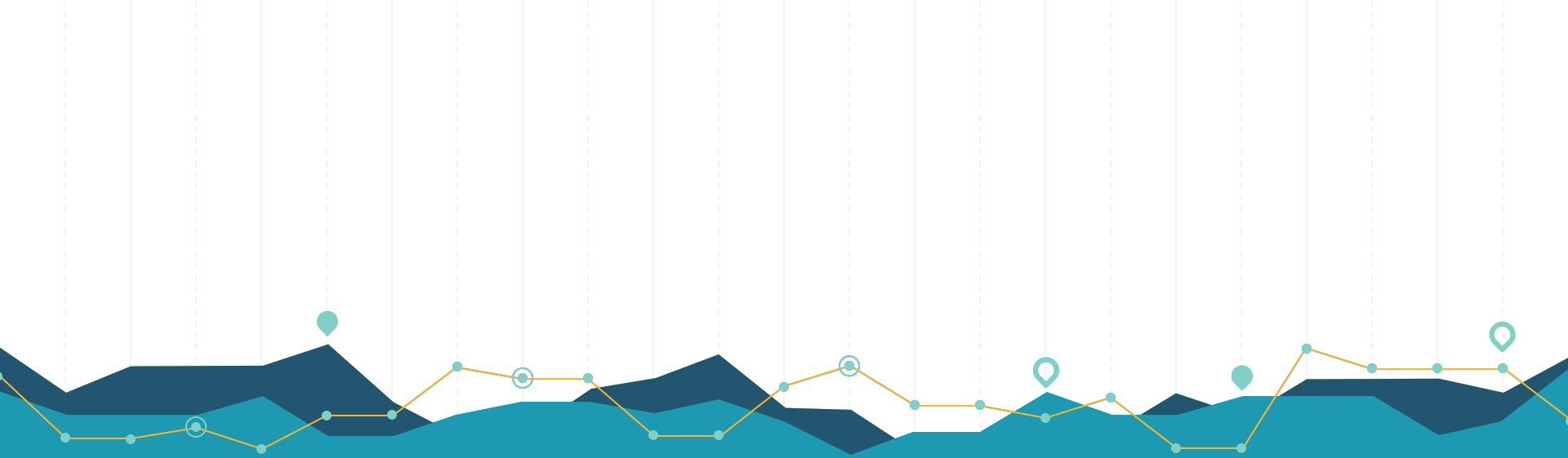
Working Efficiently

Find the Minimum Viable Method

3

PUT DOWN THE DEEP LEARNING

It's tempting to use cutting edge methods for the most predictive model...

...But every project comes with opportunity cost.

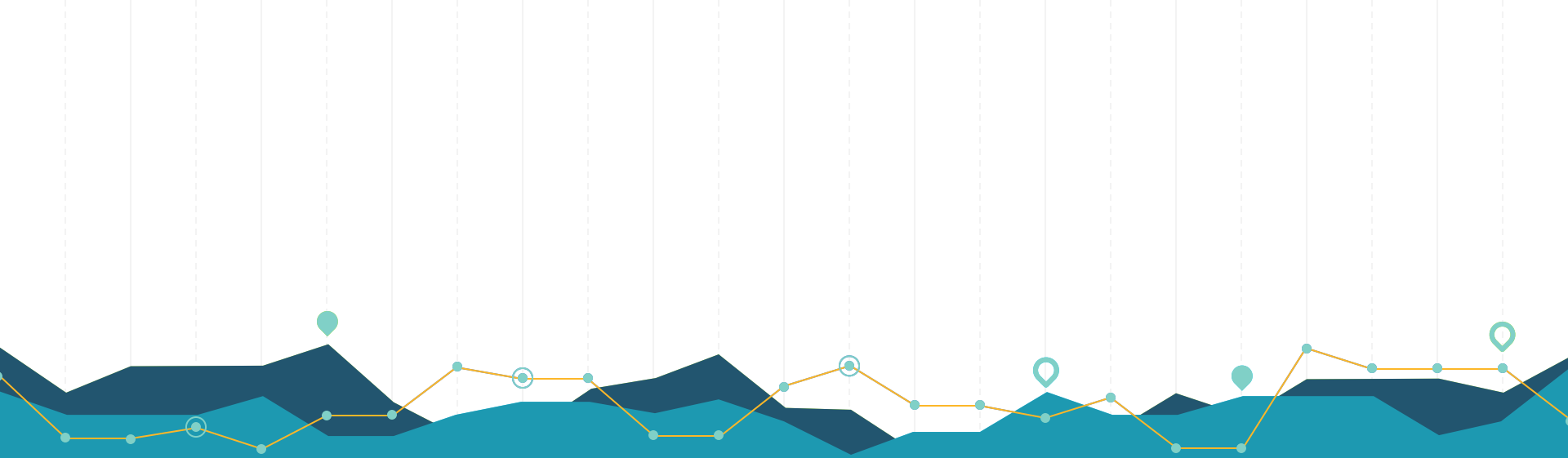# Ask yourself two questions:

"

Could I solve this with a regression?

"

Do I need a model at all?

PERKS OF A SIMPLER APPROACH

◉ Build intuition
- ◉ What features stand out?
- ◉ What are the limitations of the regression?

◉ Easier to explain and maintain
- ◉ "Why did the model do that"?

Collaboration

Be Pseudo-Agile

4

AGILE ENGINEERING

- 2 week sprints
- Daily standups
- Small tickets with clear outcomes

AGILE DATA SCIENCE?

- **2 week sprints**
- Daily standups
- Small tickets with clear outcomes

- **Long timelines with unclear boundaries between steps**

AGILE DATA SCIENCE?

- 2 week sprints
- **Daily standups**
- Small tickets with clear outcomes

- Long timelines with unclear boundaries between steps
- **"Yesterday I worked on <technical gibberish that only one other person understands>. Today I'm going to keep doing that."**
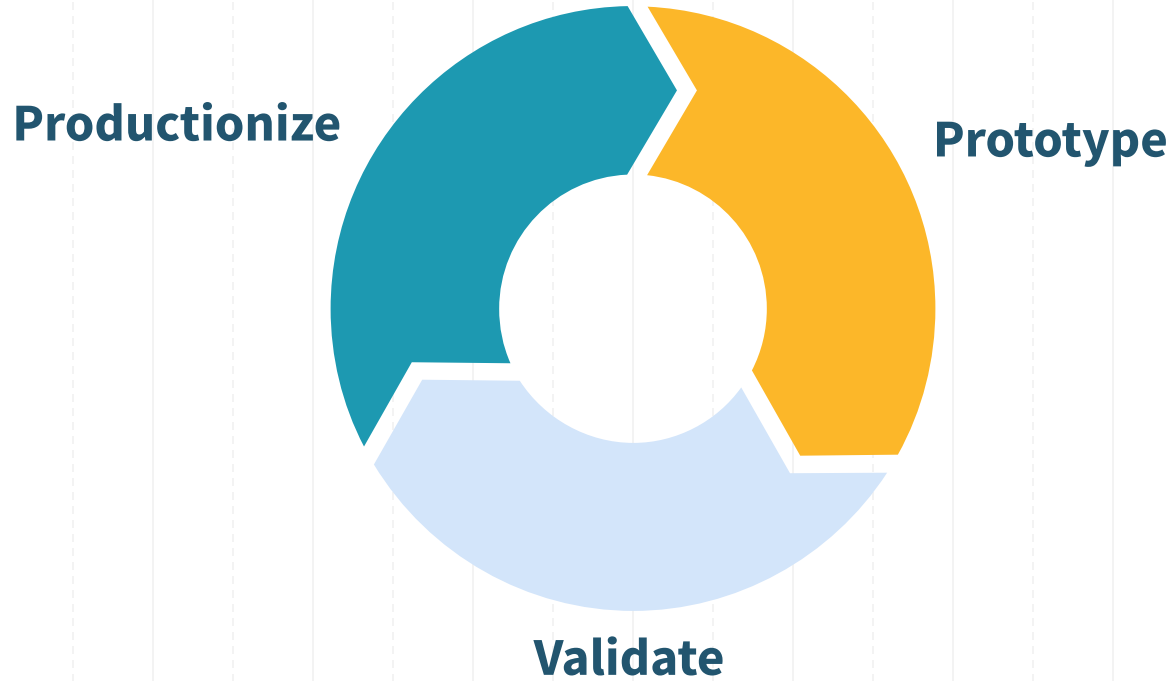
AGILE DATA SCIENCE?

- 2 week sprints
- Daily standups
- **Small tickets with clear outcomes**

- Long timelines with unclear boundaries between steps
- "Yesterday I worked on <technical gibberish>…"
- **Large tickets like "validate the model", that don't help the data scientist or product manager**
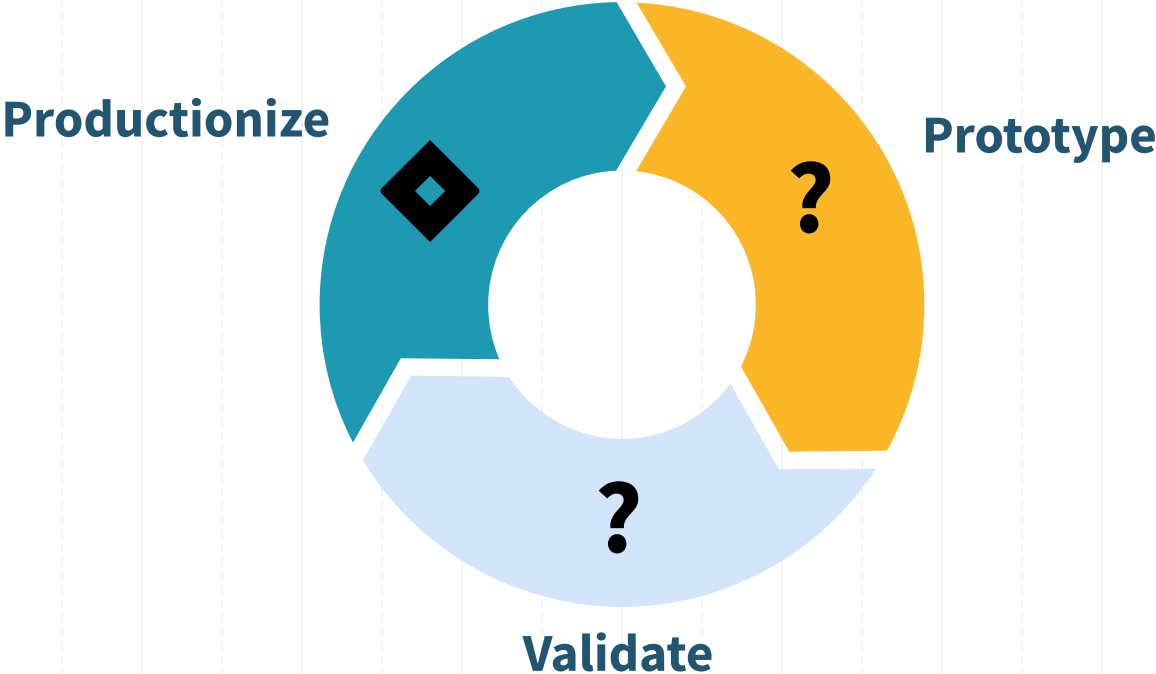
HOW CAN WE...

- ◉ Plan cross-functional work?
- ◉ Give PMs visibility into our progress?
- ◉ Stay focused and avoid rabbit holes?

THE DATA SCIENCE LIFE CYCLE

Productionize

Prototype

Validate

THE DATA SCIENCE LIFE CYCLE

Productionize

Prototype

?

?

Validate

STRATEGY 1: TIMEBOXING

◉ Divide research into questions that can be answered in a week or two

◉ Each question gets a ticket

◉ Outcome: research artifacts (notebook), decision, more tickets

◉ Works well for:

    ◉ go/no-go decisions

    ◉ Projects where individual questions take a while to answer

STRATEGY 2: RESEARCH LOG

◉ Track completed, in progress, planned work
◉ Frame as hypotheses
◉ Keep technical details elsewhere
◉ Works well when:
  ◉ The outcome is **knowledge**
  ◉ Research questions are small and interconnected
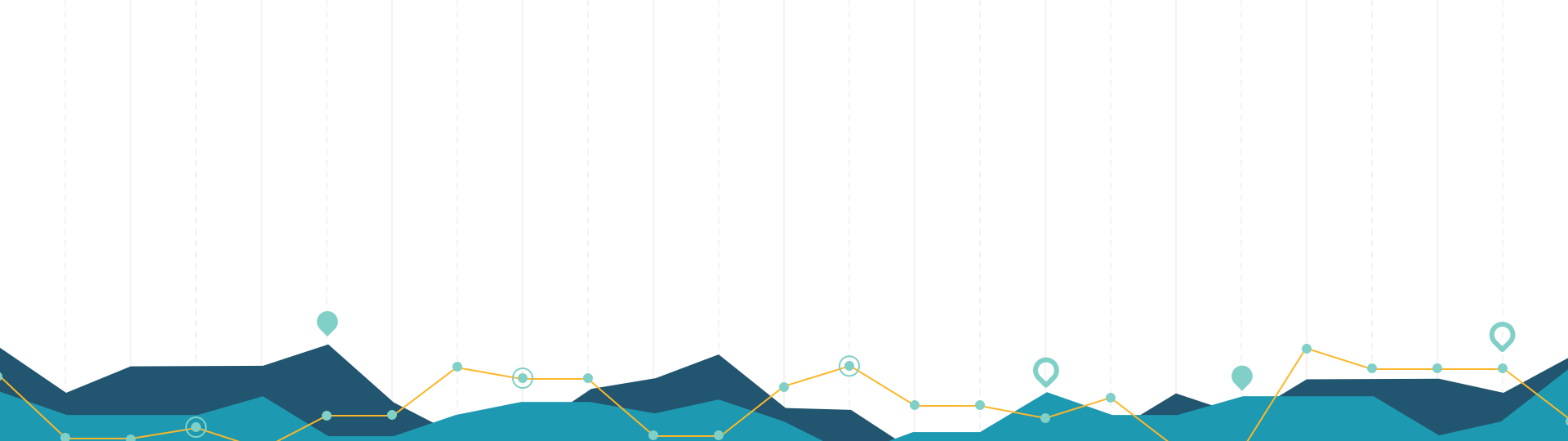
Closing Thoughts

SUMMARY

- Be motivated by problems, not solutions
- Find out what work is most impactful, then do it
- Build the simplest thing that works
- Fit the planning to the project
    - No planning/visibility is never the best option

DATA SCIENCE HOT TAKE

The coolest things to build are ones that:

- ◉ Get built
- ◉ Get used
- ◉ Solve the problem
- ◉ Can be understood
- ◉ Can be maintained

RESOURCES

◉ The Effective Engineer - Edmond Lau
◉ Project Management for the Unofficial Project Manager - Kagan, Blakemore, Wood
◉ "Put Down the Deep Learning" - Rachel Tatman, PyCon 2019

THANKS!

Any questions?

lizsander.com

https://www.linkedin.com/in/lizsander/