Hiring (and being hired) for a Successful First Data Science Project

MinneBOS 2018

Terran Melconian

terran@terran.us

twitter @terranmelconian

What You Will Learn

- Executives and Managers:
  - How to tell what type of projects you are ready to do
  - How to hire a person or team, or contract out
- Students and Job Seekers:
  - How to tell what the company actually needs (regardless of what they put in the job description)
  - How to train for the type of role you want

Should I listen to this guy?

- MIT Aero and Operations Research
- Google Image Search
- TripAdvisor data warehouse; in-house Hadoop
- Data science from zero at Jobcase, local startup
- Market-neutral microcaps
- Teaching data science to engineers and analysts

- but really, listen to me to the extent that I can offer a useful framework for organizing your experience.

Outline

- Motivating Examples
- Data science life cycle
    1. Questions to know if you (they) have done this step
    2. Appropriate data sizes
    3. Titles associated with people who do this
    4. Terms describing this work
    5. Examples of tools (sorry if I missed your favorite)
    6. How to hire or contract
    7. What distinguishes mediocre from good work
    8. Learning resources
- Occasional digressions for key concepts

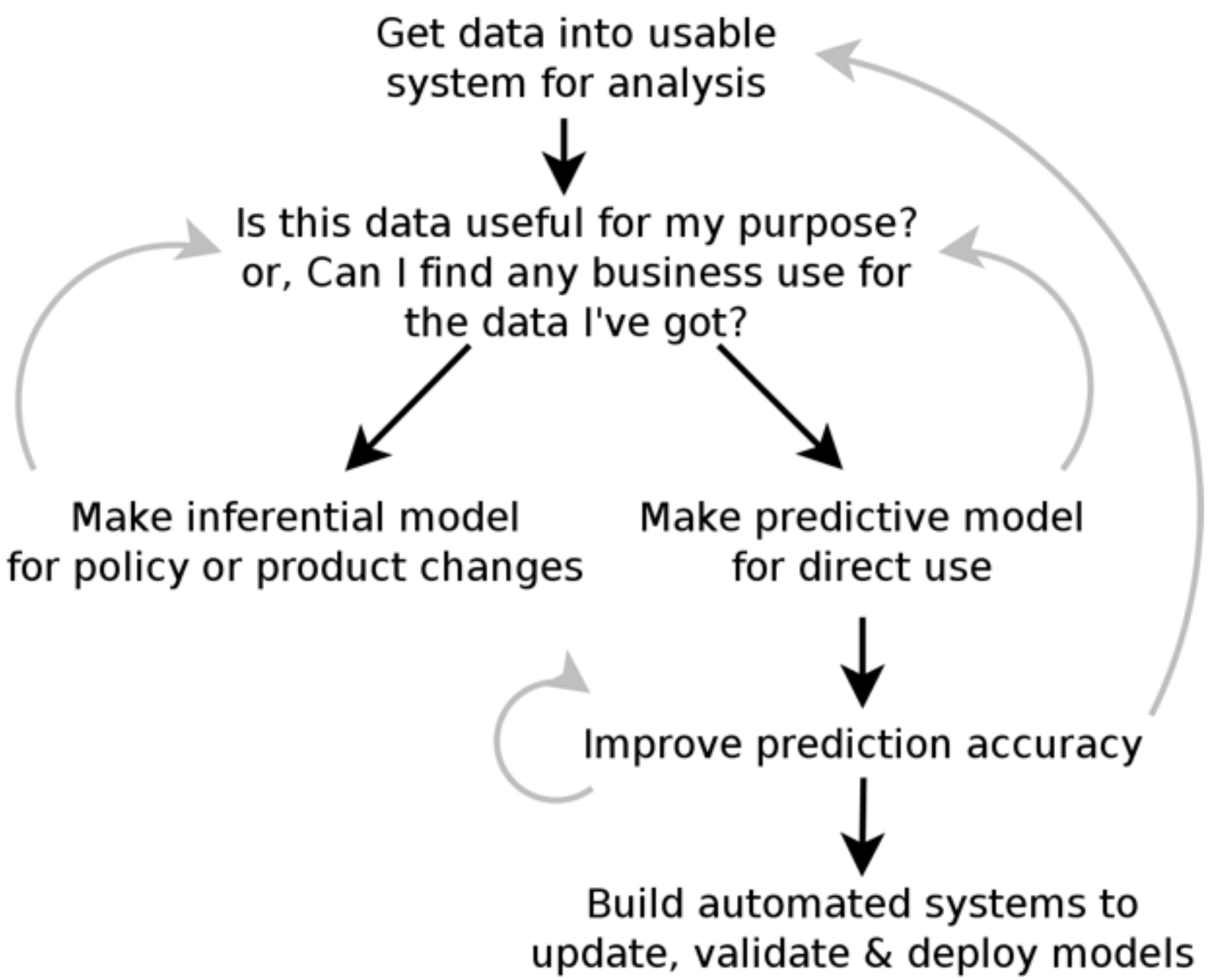
Motivating Examples

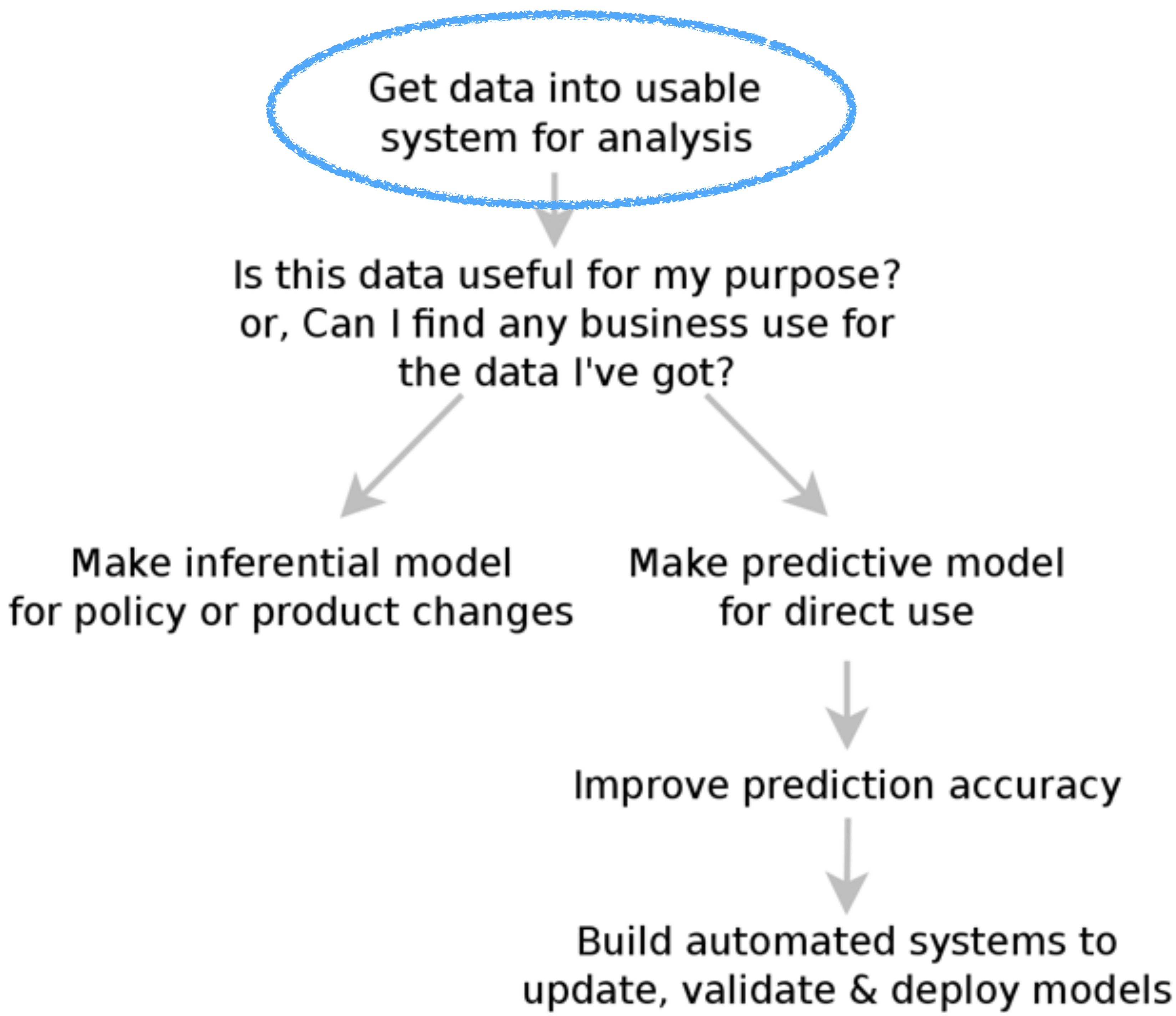
- Data science means almost everything to everyone.
  - *and that's not a good thing.*
- "I thought you wanted data science, but you seem to want a business analyst." "I wanted something I could use, but you just gave me a lot of complicated models."
- "We don't have enough data to actually learn anything about this."
- "Why are you asking me what our goals are? I hired a data scientist to optimize my business."

Motivating Examples

- "The ML revolution is new PC revolution. Get with the program or get left in the dust."
- "I'm going to rebrand my company as Machine Learning so I can get venture capital."
- "What do you mean 'open source'? I've always used my employer's custom in-house tools before."
- or maybe you read the blog post from Lyft about how they're intentionally inflating their titles, and you wondered what was going on

Data Science Life Cycle




Get data into usable system for analysis
Is this data useful for my purpose? or, Can I find any business use for the data I've got?
Make inferential model for policy or product changes
Make predictive model for direct use
Improve prediction accuracy
Build automated systems to update, validate & deploy models

Get data into usable system for analysis

- Q: "I can identify the data I need to answer most business questions, fetch it in a reasonable amount of time, and understand the results."
- Size: Usually big.
  - If small, it's not a separate team or function.

Big Data

- What is “big” data?
  - “Big” means it won’t fit on one computer, so now everything I want to do with it is more difficult and more expensive.
  - One computer in 2018 ~ 5 TB disk, 0.5 TB memory
- Is “big data” an asset?
  - Depends on whether it’s useful data!
  - Cost is superlinear; value is sublinear
  - Even huge cloud companies have *some* limits on what they are willing to store forever

Get data into usable system for analysis

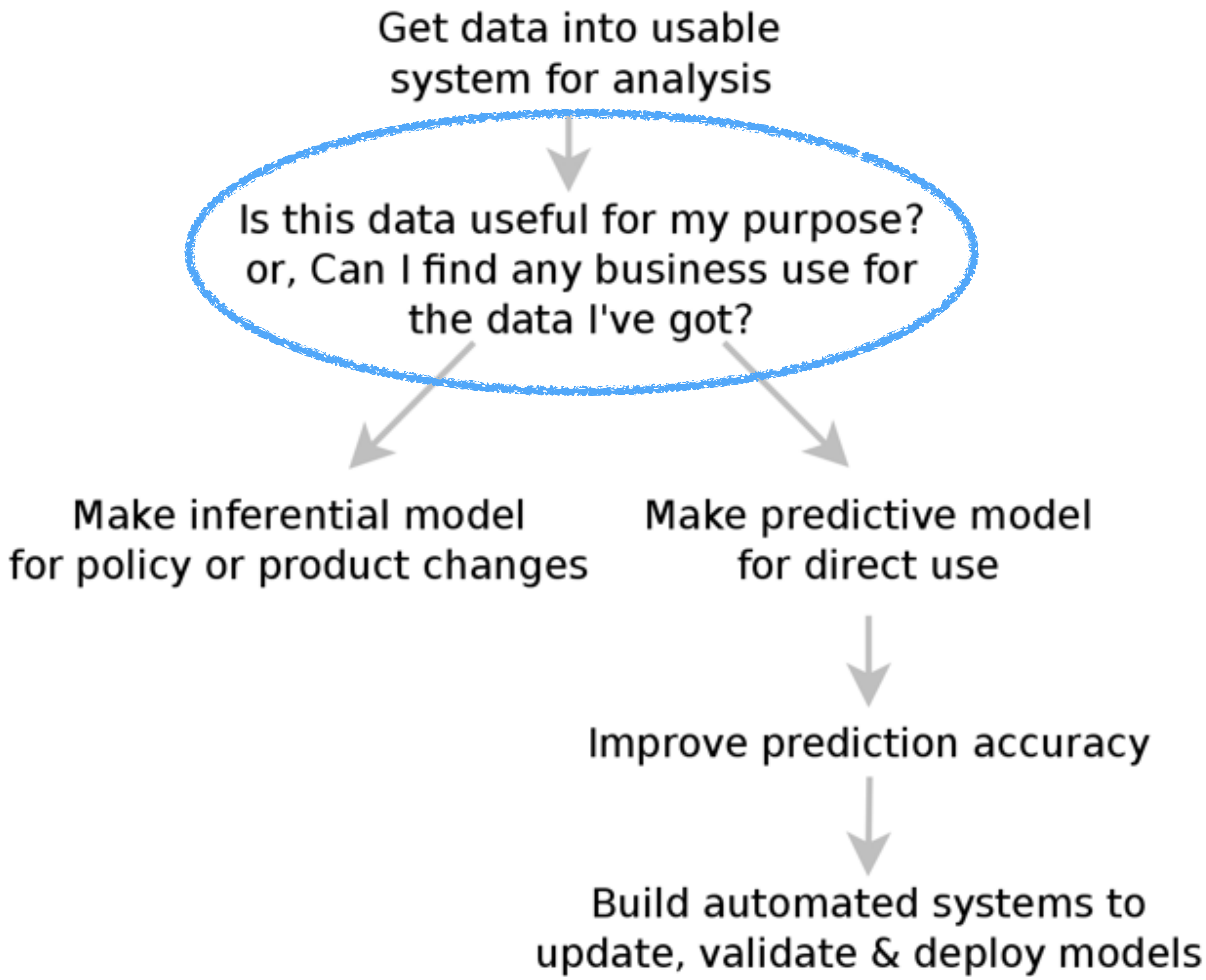
- Titles: Director of Business Intelligence, Data Warehouse Engineer, Data Engineer
- Terms: SQL, OLAP.
- Tech: Redshift, Vertica, Hadoop (Presto, Hive)

Get data into usable system for analysis

- Team: Substantial
  - Amenable to outsourcing if your source data schemata are not constantly changing.
  - First hire: based on knowledge of modeling, ability to communicate with business, and past experience in warehousing.
  - Blended team with consultant to who knows specific data warehousing technology and analytics data models.
  - Plan for ongoing maintenance.

Get data into usable system for analysis

- Differentiators: Extent of data cleaning and documentation; deep knowledge of specific warehouse technology.
- Learning: Kimball for model (30%). Apprenticeship for specific technologies (70%). You can’t afford your own data warehouse to learn on.


Get data into usable
system for analysis
Is this data useful for my purpose?
or, Can I find any business use for
the data I've got?
Make inferential model
for policy or product changes
Make predictive model
for direct use
Improve prediction accuracy
Build automated systems to
update, validate & deploy models

Is this data useful?

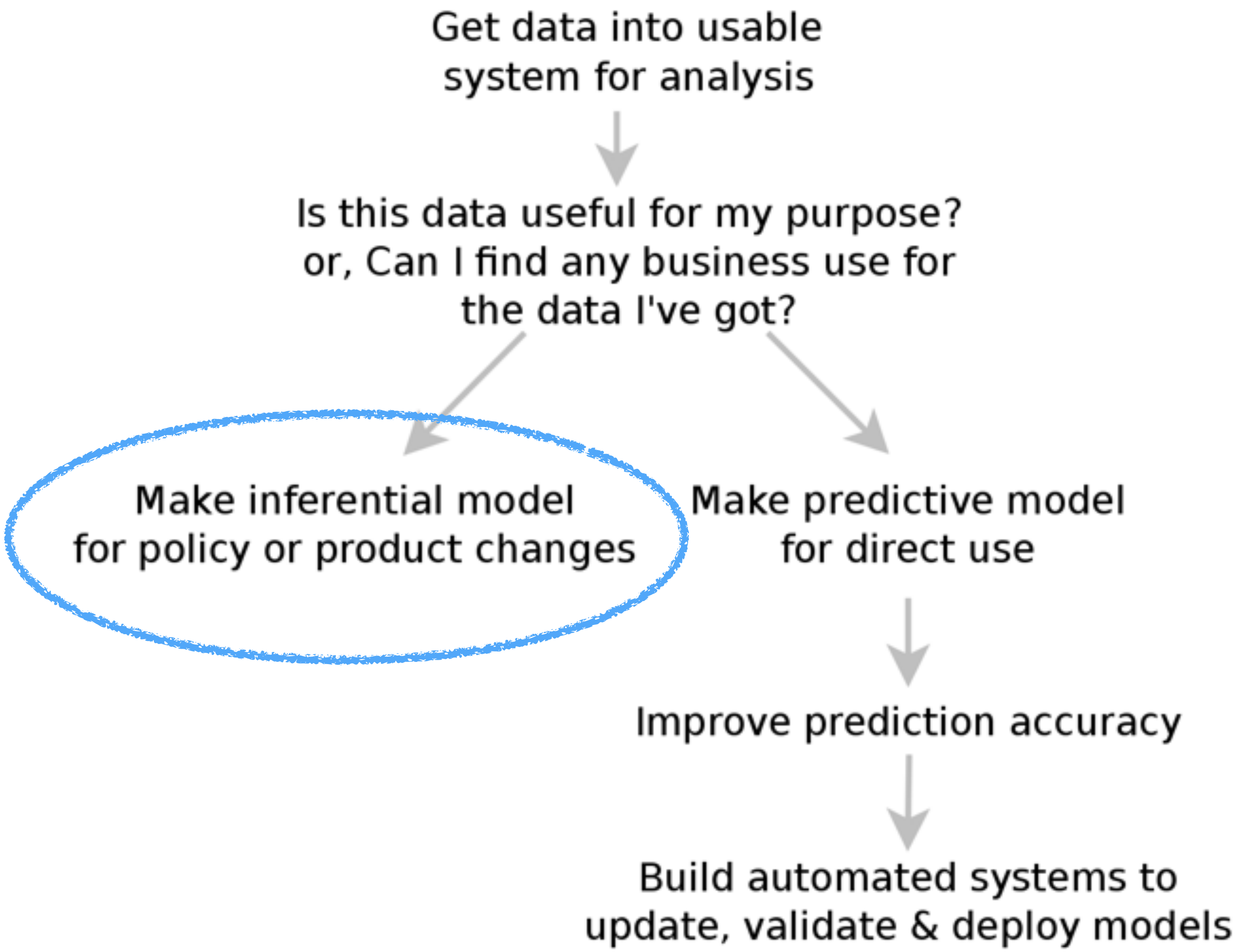
- Q: "Our data shows clear and reasonable relationships to our key metrics"
- Size: Small is better.
  - You may *have* big data, but you won't use all of it for this – you'll use one small subset or aggregation at a time for analysis.
- Titles: Business Analyst, Data Scientist.
- Terms: Exploratory Data Analysis, Data Mining, Knowledge Discovery
- Tools: R, Excel, Tableau, etc. Python?

Is this data useful?

- Team: Start with 1 person
  - Contract or consultant
  - Full time if successful
  - Substantial domain knowledge investment
  - Hiring: Give them some data and have them come back and make a presentation.
- Differentiators: Domain knowledge. Causal reasoning.

Is this data useful?

- Learning
  - R for Data Science, Grolemund and Wickham
  - Stephen Few for visualization
  - I haven't found any books yet which teach the reasoning part. Maybe I'll have to write one…


Get data into usable system for analysis
Is this data useful for my purpose? or, Can I find any business use for the data I've got?
Make inferential model for policy or product changes
Make predictive model for direct use
Improve prediction accuracy
Build automated systems to update, validate & deploy models

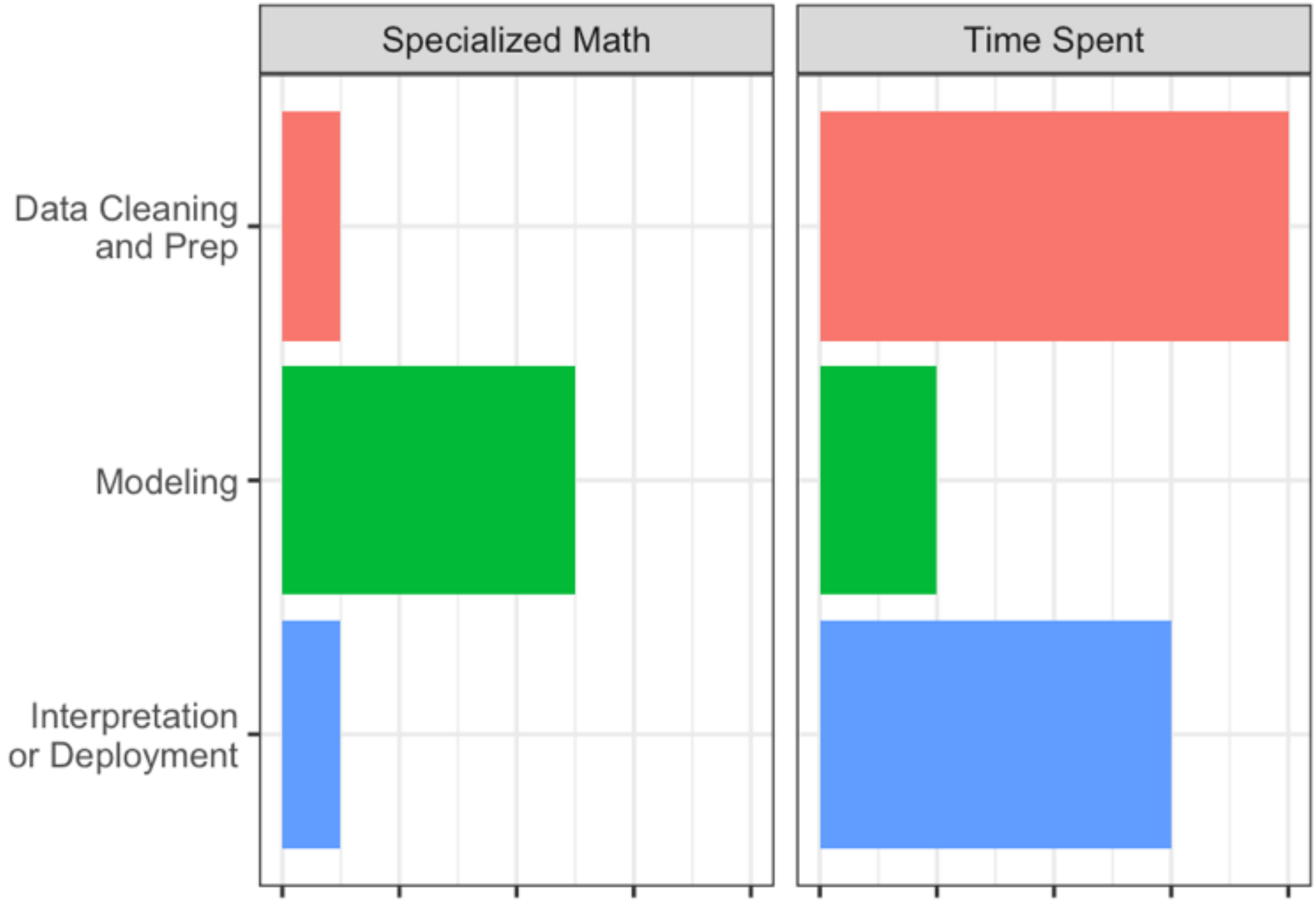
Inferential Model

- Q: "I know, with evidence, how the factors under my control impact key business outcomes."
- Size: Small is better. Large is slow, and you can't understand highly complex models anyway.
- Titles: Data Scientist, Statistician, Econometrician
- Terms: Regression, Statistical Modeling, Predictive Analytics
- Tools: R, SAS, SPSS. Excel? Python?

Inferential Model

- Team: Start with 1 person
  - Good place to bring in consultant for the modeling part
  - Not great for “throw it over the wall”, Kaggle-style. Models usually expose data problems; better to fix than work-around.
  - Do you want your staff to learn? It’s an investment. Decide before you select the consultant.

How much math?
How much time?

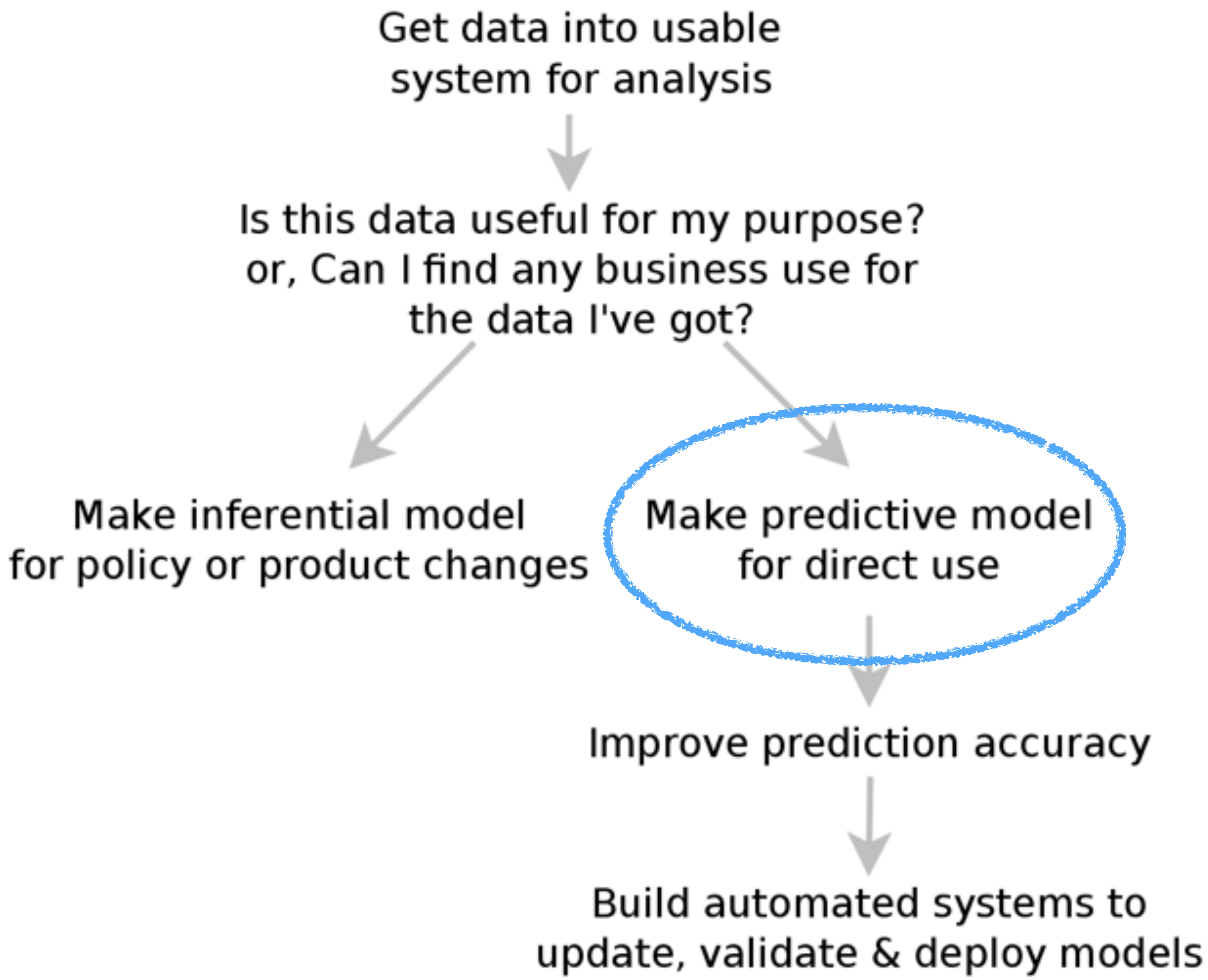


Hiring with an Exercise

- Standard practice for perm data science roles. Candidates usually willing after in-person interviews.
- Realistic data size, realistic data type
- Realistic deliverable! Report, model, or presentation?
- Correct stakeholders present for evaluation
- (I have a hiring talk which expands on this; find it on my web site)

Inferential Model

- Differentiators
  - Stronger inferences about causality.
  - Better diagnosis of data problems.
  - Clearer presentation.
- Learning
  - OpenIntro Statistics, Diez et al
  - Data Analysis Using…, Gelman and Hill
  - Practice on real data that is meaningful to you. Have “skin in the game” to stay engaged.


Get data into usable system for analysis
Is this data useful for my purpose? or, Can I find any business use for the data I've got?
Make inferential model for policy or product changes
Make predictive model for direct use
Improve prediction accuracy
Build automated systems to update, validate & deploy models

Predictive Model

- Q: “We have achieved statistically significant lift in our business metrics using our predicted values.”
- Data size: Small to medium. Large data and complex models are slow to iterate on; not justified until after first POC.
- Titles: Data Scientist, Machine Learning Engineer
- Terms: Machine Learning, Predictive Analytics, Data Science
- Tools: Python (sklearn), Weka. R?

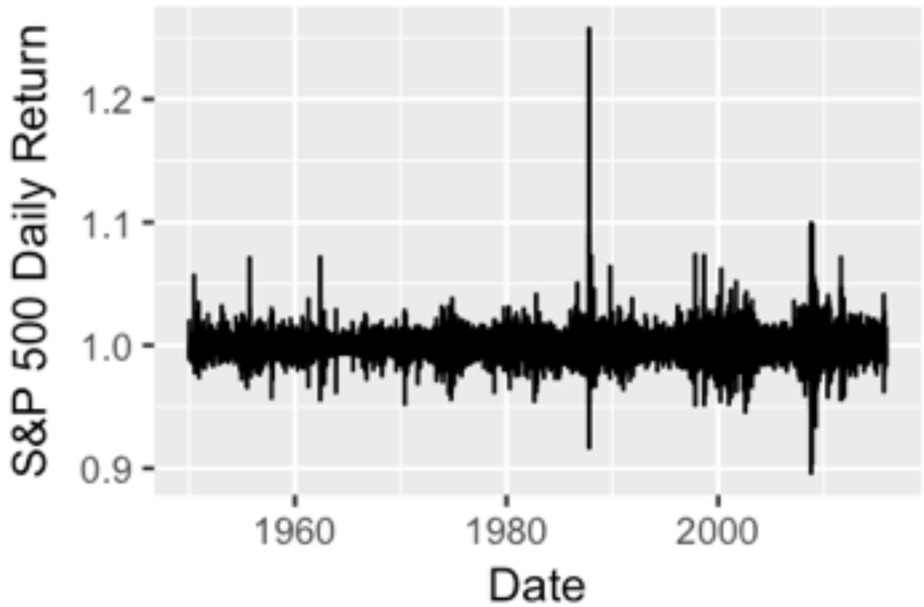
Predictive Model

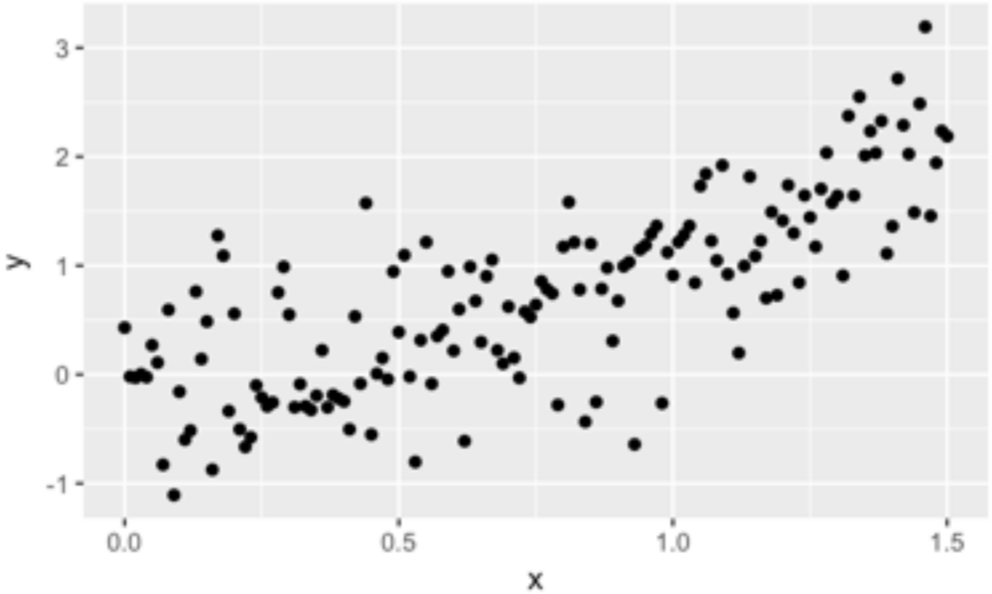
- Team: Start with 1 person
  - Reasonable to outsource or contract (remember, you have already ascertained that your data is relevant!)
  - Hire with exercise, as for inferential models, or competition-style
- Differentiators: accuracy, speed, required data
- Learning
  - Introduction to Statistical Learning (with Applications in R), James et al
  - Introduction to Machine Learning in Python, Müller and Guido

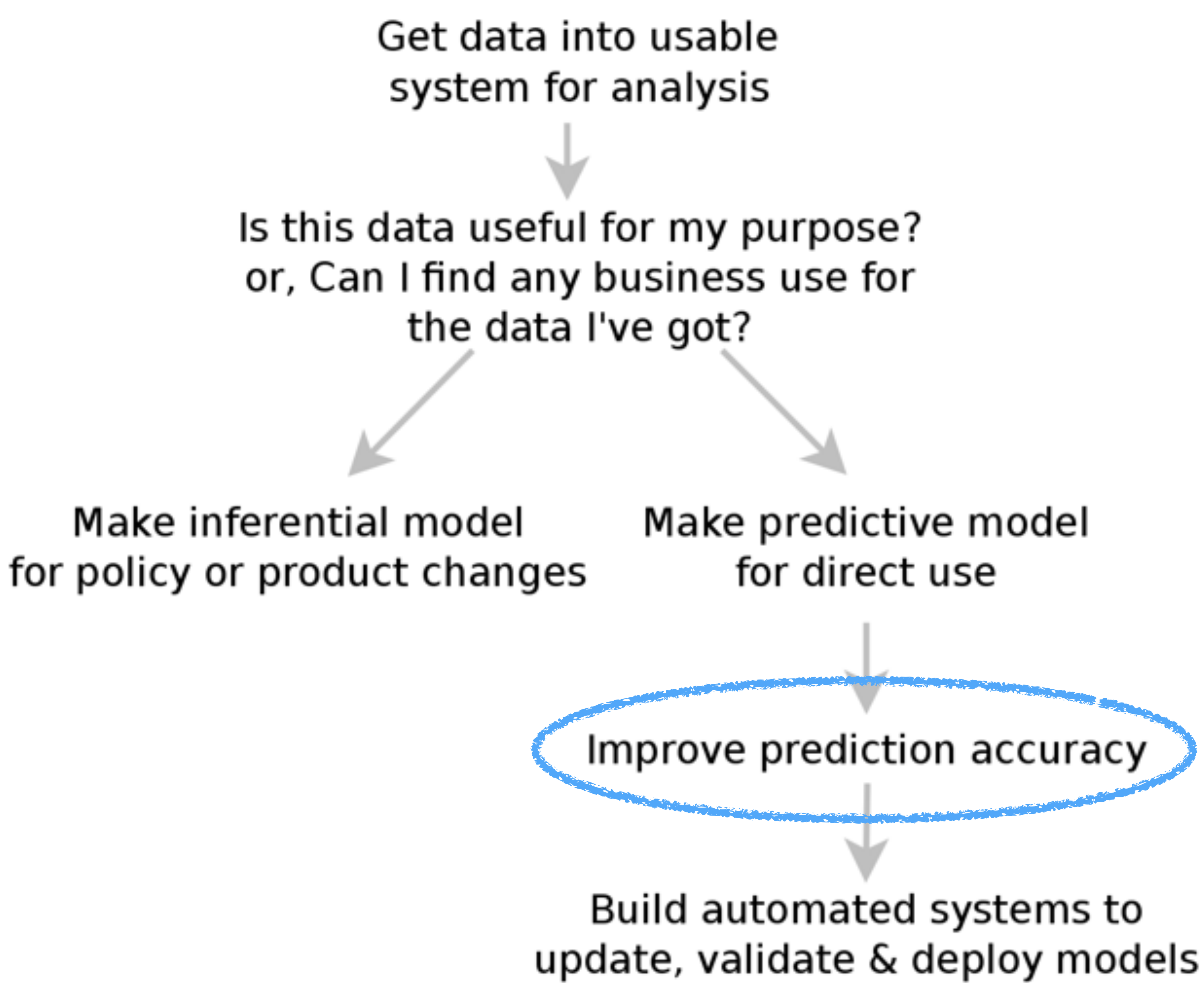
Signal to Noise

- High:

- Low:



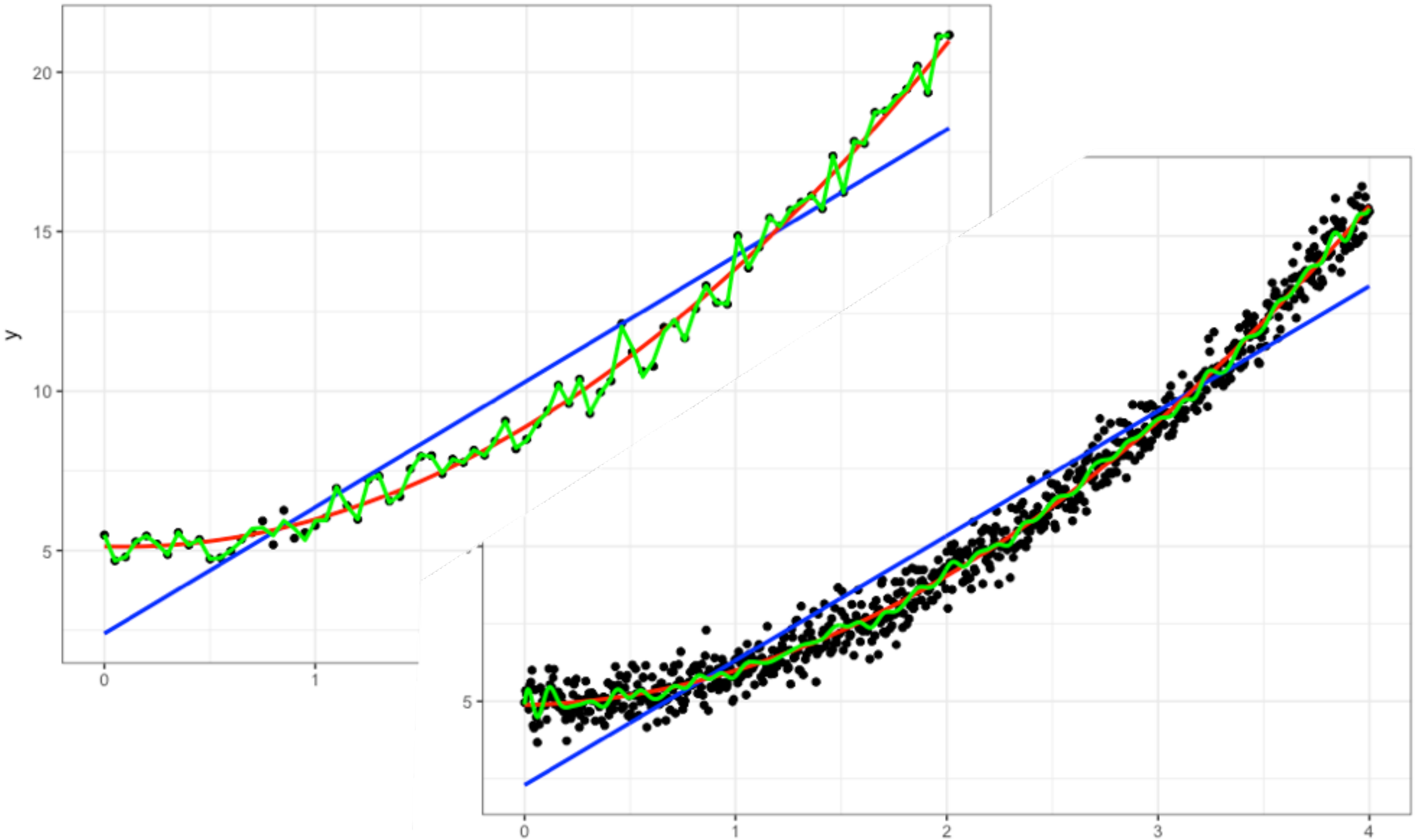



Get data into usable
system for analysis
Is this data useful for my purpose?
or, Can I find any business use for
the data I've got?
Make inferential model
for policy or product changes
Make predictive model
for direct use
Improve prediction accuracy
Build automated systems to
update, validate & deploy models

Improve Accuracy

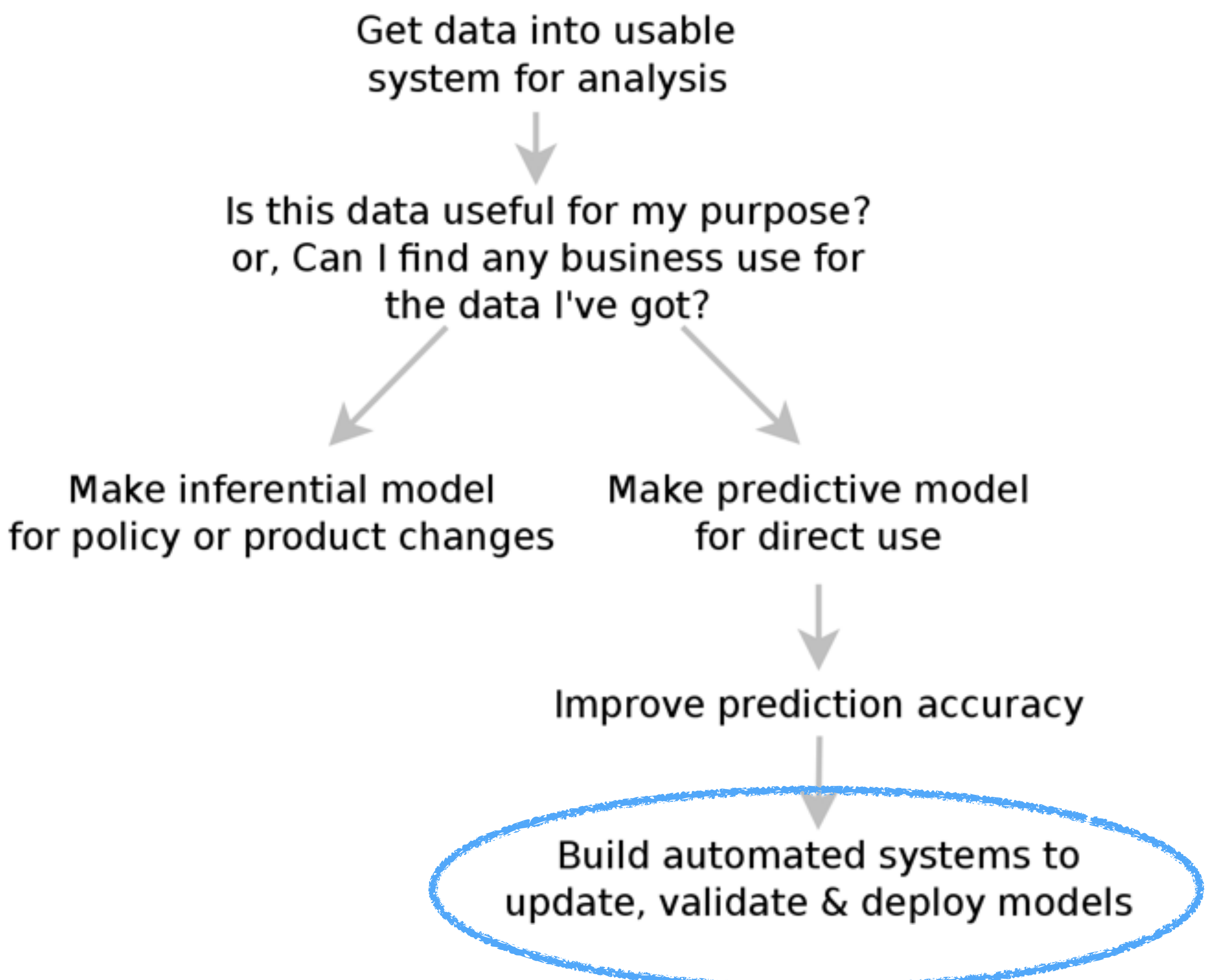
- Q: "Further plausible improvements to this accuracy would not materially help the business."
- Size: Medium to large. Large data enables more complex and potentially more accurate models.
- Titles: Machine Learning Engineer or Scientist
- Terms: Random Forests, Gradient Boosting, Deep Learning
- Tools: XGBoost, Keras, TensorFlow

Overfitting



Improve Accuracy

- Team: Substantial
  - considerable work just storing, moving, and processing data
  - Outsource to an API: DataRobot, Microsoft, Google, H2O
- Differentiator: Accuracy
- Learning:
  - Deep Learning, Ian Goodfellow
  - Deep Learning with Python, François Chollet
  - Kaggle Competitions
  - Also learn data pipeline tools – entry level people don’t get to do model architecture.


Get data into usable system for analysis
Is this data useful for my purpose? or, Can I find any business use for the data I've got?
Make inferential model for policy or product changes
Make predictive model for direct use
Improve prediction accuracy
Build automated systems to update, validate & deploy models

Model Training Pipelines

- Q: “I can automatically refit models and be confident they do not have serious errors.”
- Size: Large
- Titles: Machine Learning Engineer, ML Operations Engineer, Data Engineer
- Terms: Data Engineering, Anomaly Detection, DevOps
- Tools: Spark? Mostly custom, in-house tools.

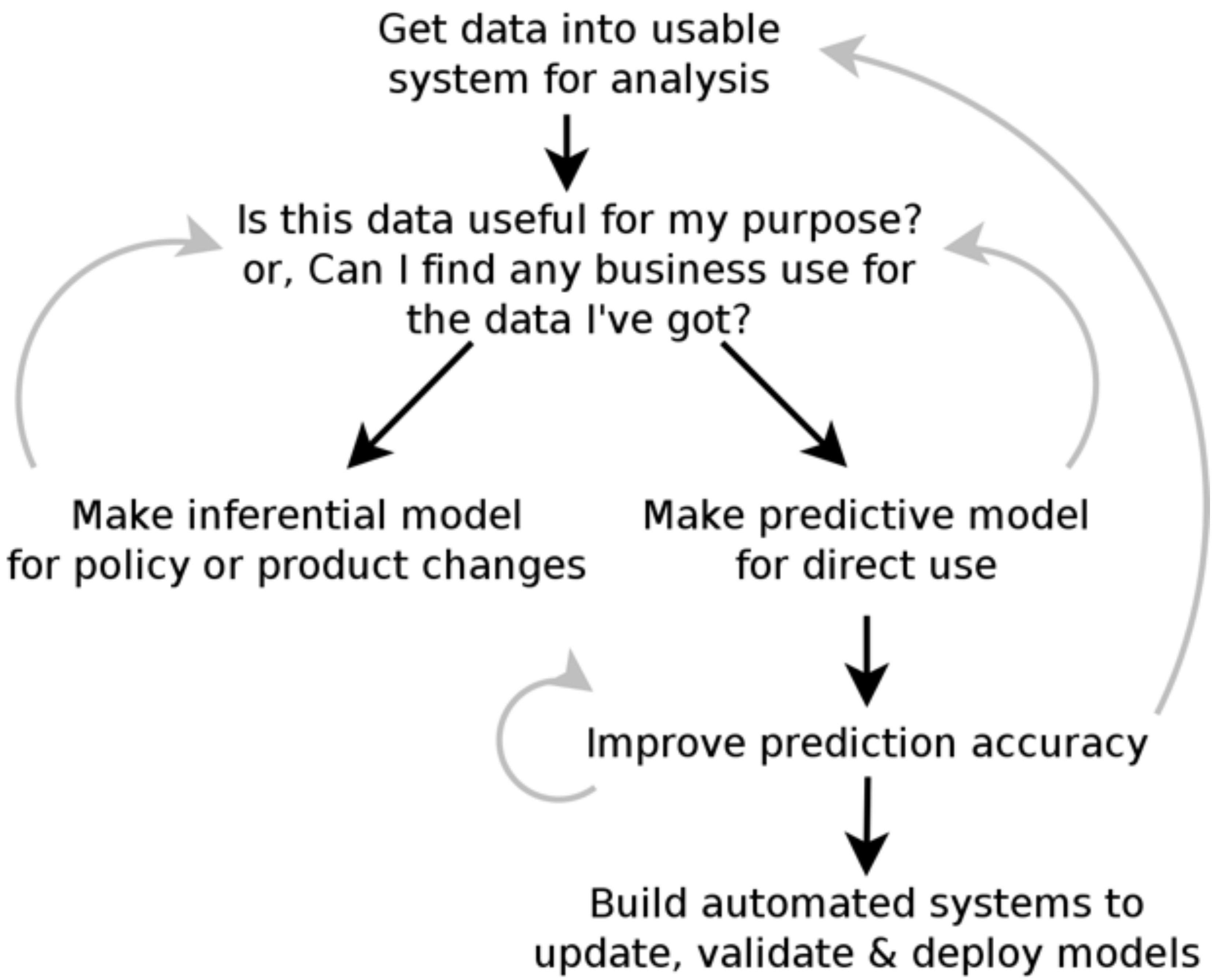
Science and Engineering

- These words still have meanings
- Science
  - ➡ “What” and “why” knowledge
  - ➡ Reports, papers, models, and algorithms
- Engineering
  - ➡ “How” knowledge
  - ➡ Working systems and reliable tools
- Please don’t say “our data scientists do all their own production engineering.”

Model Training Pipelines

- Team: Substantial. At this point you're investing as a core capability.
  - Hiring is similar to engineering hiring – skills interviews.
- Differentiators: Systematic improvement instead of firefighting and emergency response
- Learning: System engineering and understanding emergent behavior. Outlier Analysis by Aggarwal.

Life Cycle Recap



Culture and Politics

- Data is meritocratic. Is your culture?
- Power in an organization is zero-sum
  - If the data team is gaining influence, somebody is losing
  - Primates do not generally give up power voluntarily
  - 💰 Cajole
  - 🔫 Coerce
  - 🚪 Cashier

Deep Hype

- Deep neural networks *do* deliver the best accuracy on high signal-to-noise problems
  - Image classification
  - Natural language processing
  - Surveillance (or voluntary user action) data
- The hype is in the breadth of problems to which these techniques are applicable
- Most problems are not high signal to noise problems
- The “shortage” of deep neural network labor is a myth

Software Developer Educational Antipatterns

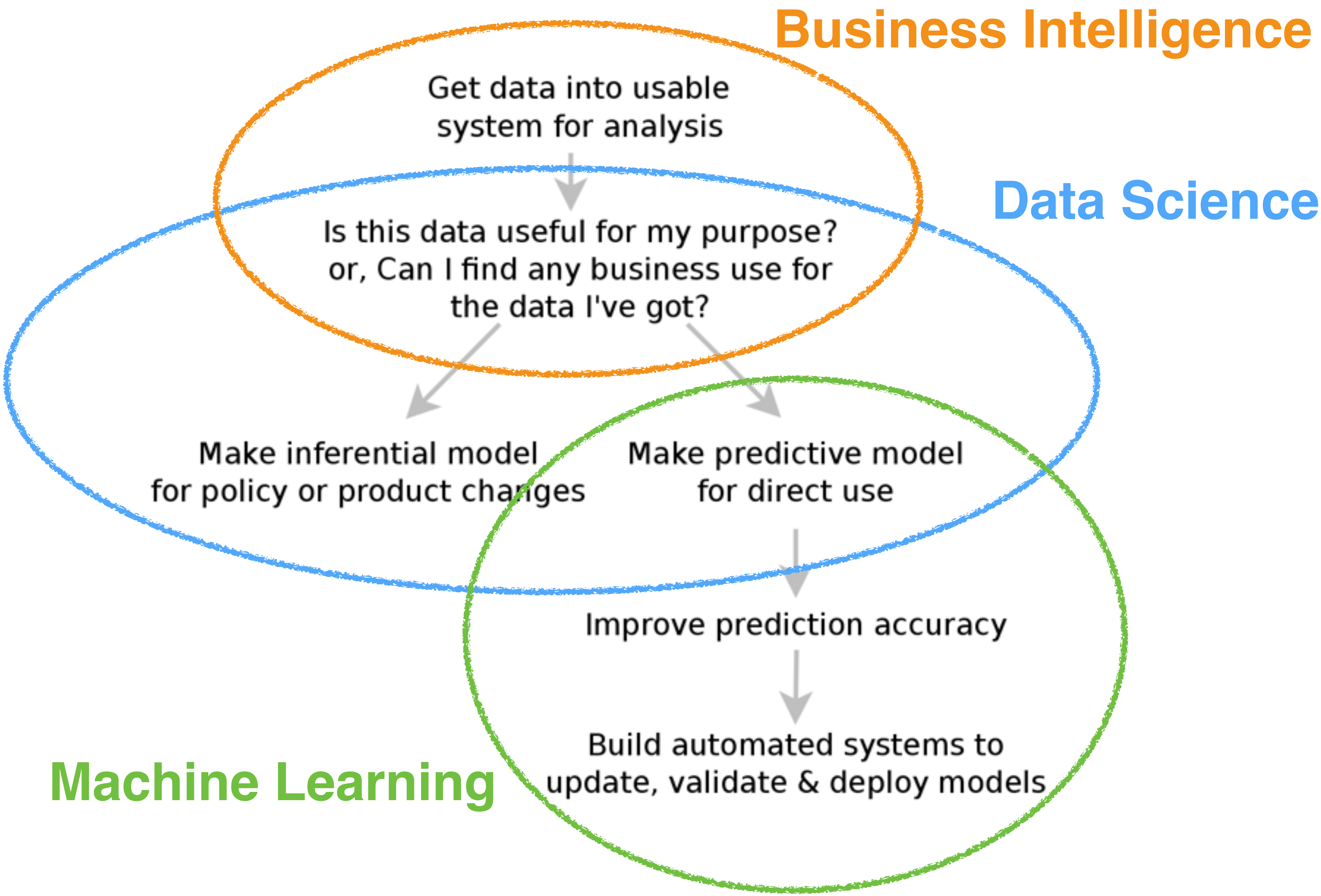
- Wrong
  - “I’m a software engineer and I love algorithms. Machine learning looks like a great way to learn some cool new algorithms, but not to change anything else about how I think or approach problems.”
- Also Wrong
  - “I got a master’s degree in machine learning, so now I can go do cutting edge algorithm work in industry or implement awesome chatbots from scratch!”
- Right
  - “I’m a software engineer who studied ML. Now I can go build data pipelines to bring new features into models that somebody else designed, and make their models more operationally reliable!”

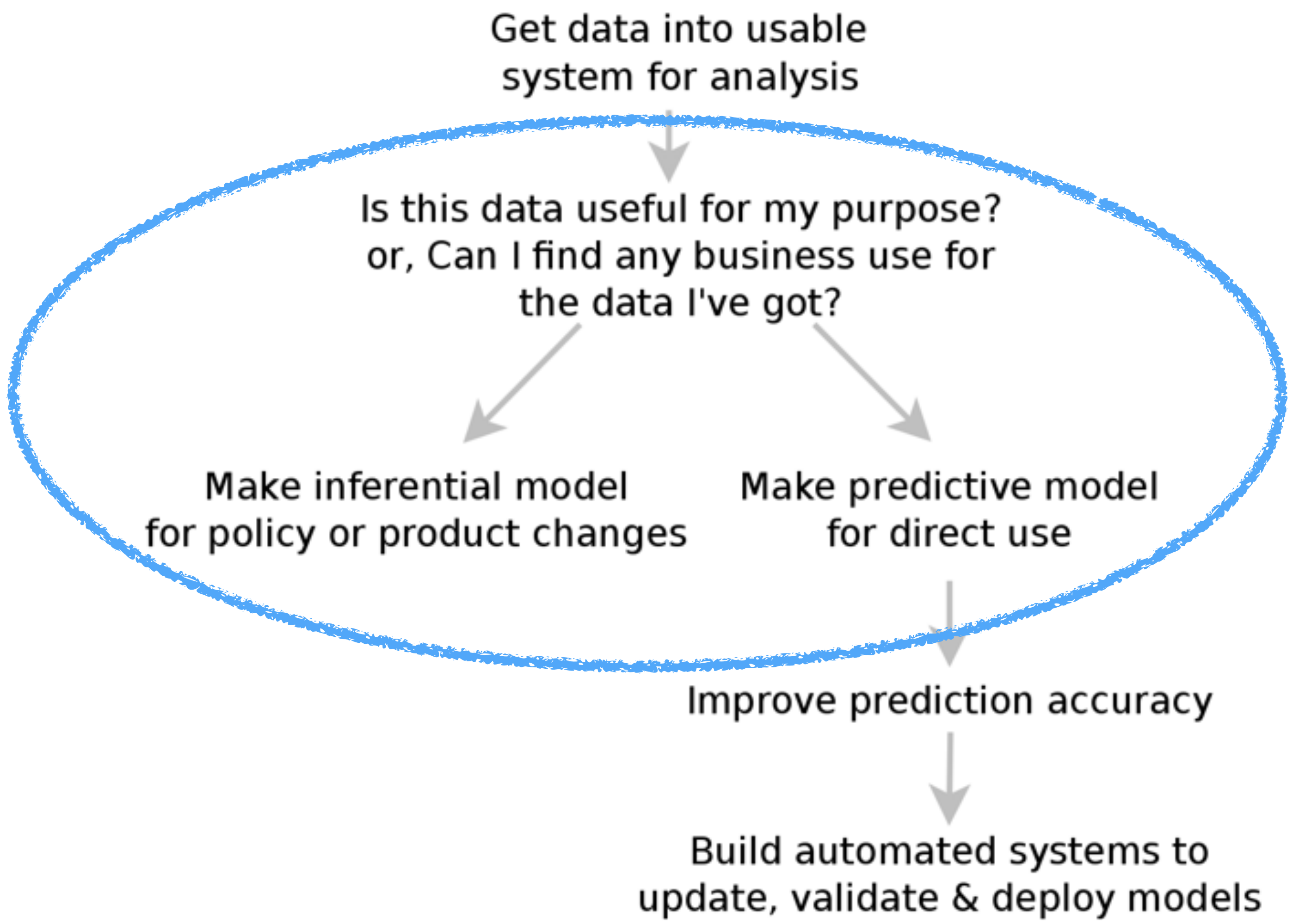
Managing Data Science

- You don’t need the math or programming
- You still need the qualitative understanding
- Top challenge is that managers are too busy executing day to day to have time to learn.

- Data Science for Business, Provost and Fawcett

Best Guess 2018 Taxonomy





Terran Melconian

terran@terran.us