CSX415
Data Science Principals and Practice

Christopher Brown
U.C. Berkeley / Decision Patterns LLC
Spring 2018

Berkeley
UNIVERSITY OF CALIFORNIA

Christopher Brown

Founder, Decision Patterns

Adjunct Professor of Computer Science
University of California Berkeley

Courses Taught

- *Practical Machine Learning*
- *Data Science Principles and Practice*

Christopher Brown has spent the last 18 years as a consultant in a variety of industries: Financial Services, Health Care, Retail, Defense, etc. **Chris** also teaches statistics and computer science at the University of California, Berkeley.

Chris and his teams are frequent contributors to Open Source Software in a variety of projects and programming languages.

DECISION PATTERNS

We help our client

use data to make (better) strategic and operational decisions.

- Strategy
- Organizational Structure
- Talent
- Execution
  - Data acquisition
  - Data organization (ETL & data warehousing)
  - Data consumption (data science/analysis/ML)
  - Analytical application development

christopher.brown@berkeley.edu

Goals

best kept secrets of …

Data science and machine learning

Organization

Roles and responsibility

Project lifecycle

Tech stack


MachineLearning
DataMining
Statistics
AdvancedAnalytics
DataScience
ArtificialIntelligence
SupervisedLearning
DeepLearning
ReinforcementLearning
UnsupervisedLearning
MachineIntelligence

Data Science

Oh Really …

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

THE WORLD NEEDS DATA SCIENTISTS

IF YOU ARE A MATH- OR DATA-DRIVEN INDIVIDUAL LOOKING FOR THE PERFECT CAREER FIT, look no further than data science. Due to the ongoing explosion of big data, companies have more information at their fingertips than ever—and not enough people who can make sense of it all. This reality has created a big market for quantitative analysts and individuals who can put massive amounts of data into perspective. Take a look.

Source: http://venturebeat.com/2013/11/11/data-scientists-needed/

CAREERS IN DEMAND

15,000%

ACCORDING TO FICO, THERE WAS A 15,000% INCREASE IN **JOB POSTINGS** FOR DATA SCIENTISTS

Currently the job market seeks

140,000–190,000

DATA SCIENTISTS TO FILL OPEN POSITIONS.

IN ADDITION,

1.5 million

data literate managers will need to be retrained or hired to meet needs.

Salary

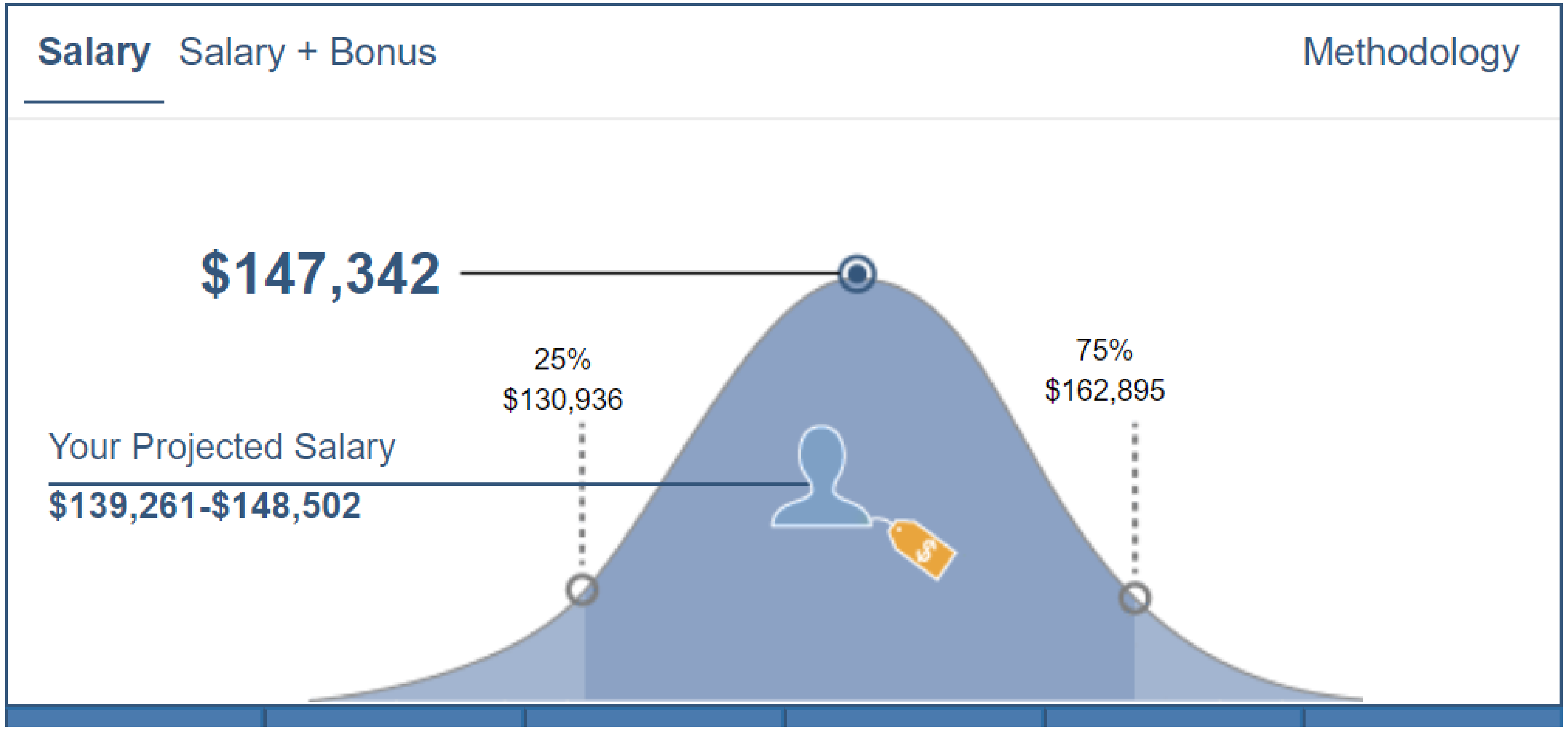
Mid-level Data Scientist

Des Moines, IA	$114,784
Redmond, WA	$129,875
San Francisco, CA	$147,342

Data Scientist IV Des Moines, IA

Data Scientist IV Redmond, WA

Data Scientist IV San Francisco, CA



Source: Salary.com

Additional Expenses

Bonus +
Overhead +
Technical Resources +

$150,000+
per data scientist

Data science is expensive

… great impact on the viability of projects

… how can we make it *cheaper* or *more efficient*?

Same as anything …
develop *efficient* and *repeatable* processes automating those parts that make sense.

What is Data Science?

An interdisciplinary **field** about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured,[1][2] similar to Knowledge Discovery in Databases (KDD).

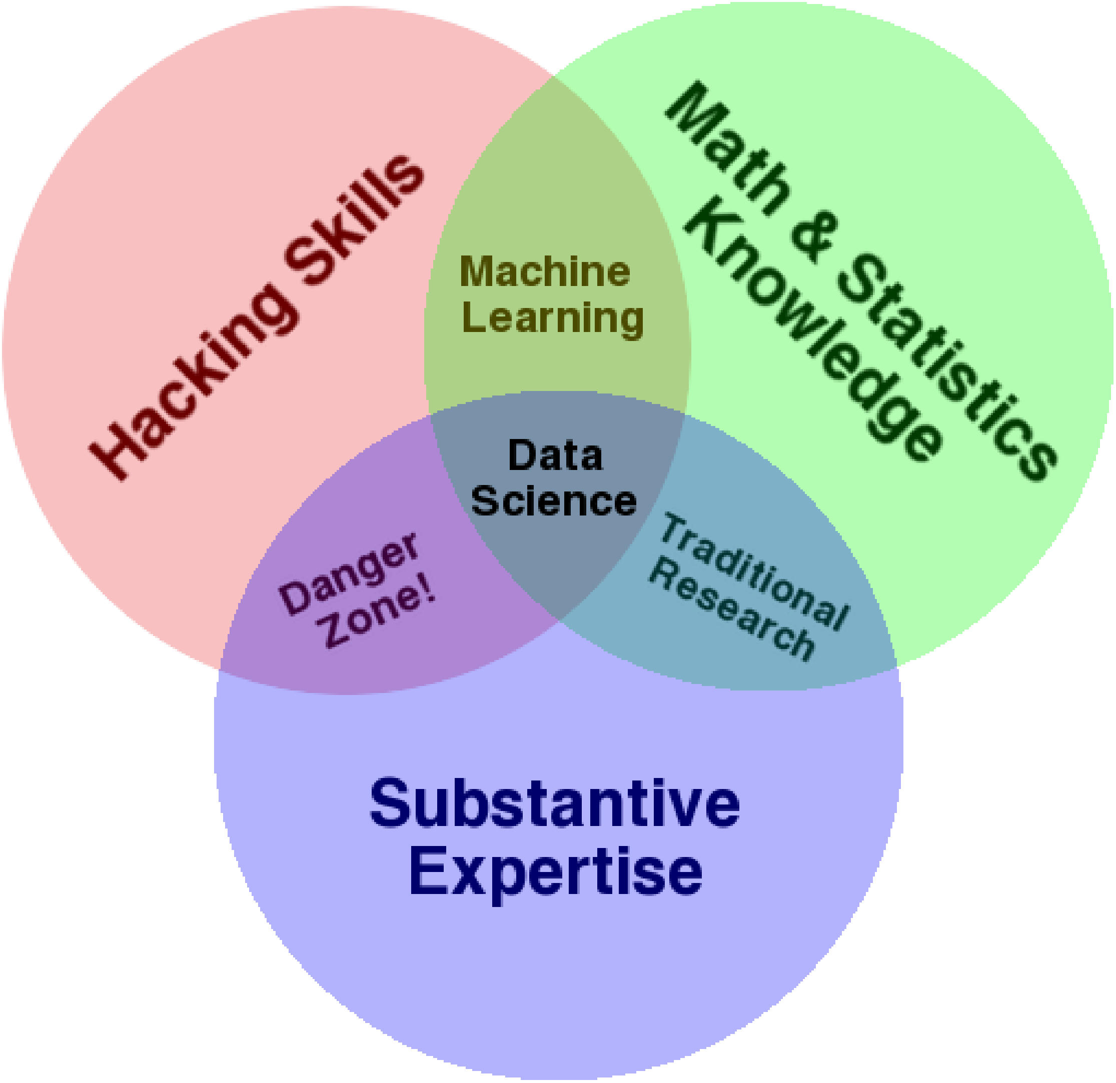
https://en.wikipedia.org/wiki/Data_science

There is no consensus …

definition of **data science**.

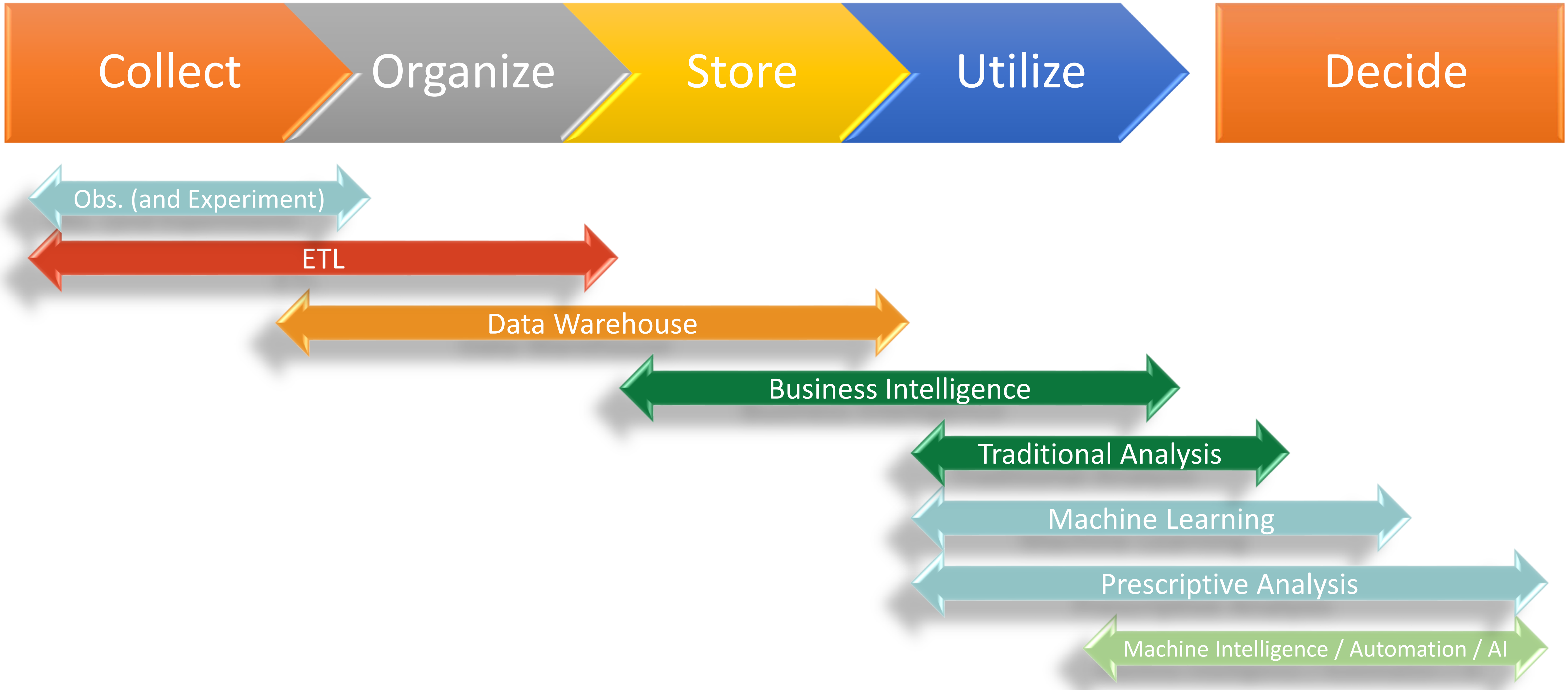
Data Science by Skill Set

Data Science Venn Diagram



Ref. http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science by Responsibility



Data Science ≠ Machine Learning

Data Science by Function

Functional

The **process/practice** of using **data**

Collect >> Organize >> Store >> Utilize

In order to

- **Understand or explain (traditional statistical analysis)**
- aid or automate ***<u>decision making</u>***

Data Scientists only do 7 things

Forecasting

Determining the state of your business, application or process

- at a future time?
- and/or in a diff. environment?

Forecasting

Examples

Macro

- Profit/ROI Forecasts
  - Revenues
  - Cost
- Capacity and Resource Planning (ERP)

Micro

- Product Demand (PD)
- Member Lifetime Value (MLV)

Rare Event Detection

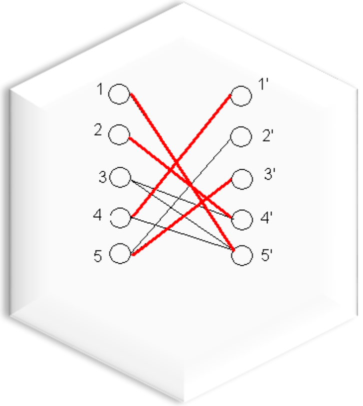
Identify outliers or observations with interesting or aberrant characteristics?

(before those characteristics are known/observed)

Rare Event Detection

Examples

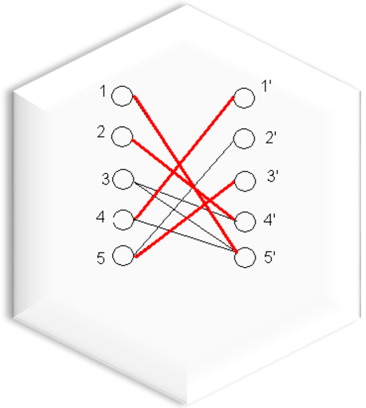
- Member Attrition
- Risk events (Fraud/Default)
- Security (anomaly detection, network intrusion)



Relationships

Determine

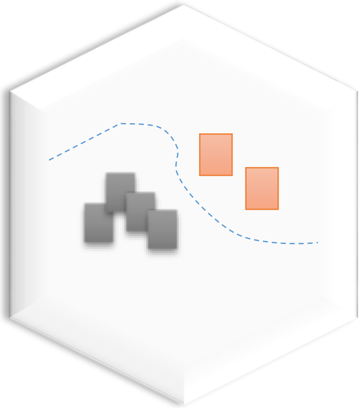
- the type and/or strength of relationship between two things?
- Do you have anything to recommend?



Relationships

Examples

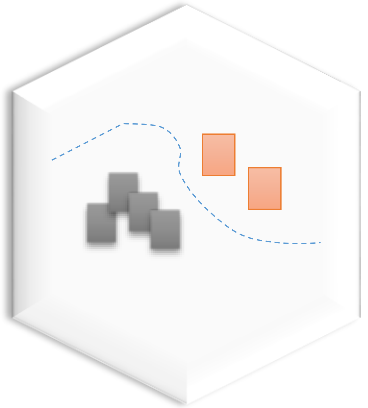
- Recommendation (product – member)
  - Product recommendations
- Affinity or similarity (member – member)
  - CU Evangelists
- Co-occurrence / Market Basket Analysis (product – product)



Segmentation / ID

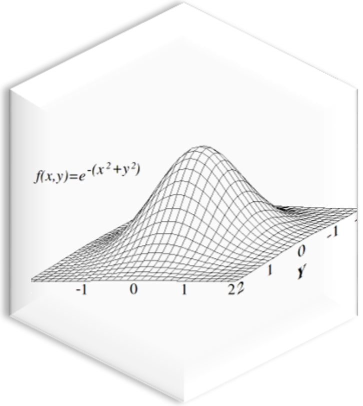
Determine how people or things are group together for :

- promoting understanding of a system?
- Implementation of a (often complex) strategy?



Segmentation / ID

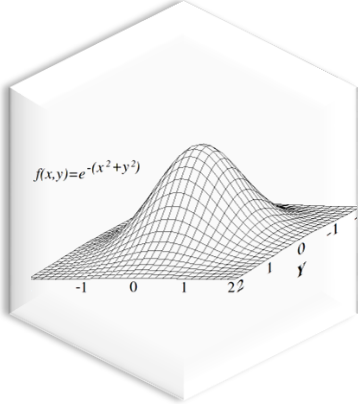
- What are my member segments (usually related to a need or outcome)
- What do I market to existing or potential members?
- How are member support reps alike or different?



Optimization

Determine the best choice

- among a (possibly infinite) number of possibilities and/or
- with a one or more constraints?



Optimization

Resource allocation

- Which sales and marketing opportunities should be pursued?
  How do I deploy a fixed sales staff?
- How do I apportion budget(s)?
- How do I most effectively partition my (limited) set of efforts?

Causal Analysis

Identify the *causes* that bring about various *outcomes*?

Causal Analysis

- Key Driver Discovery
- What are profits key drivers?
- What are product features driving adoption? (stickiness)

Data Collection

Do you need more data and want to

- ensure that the right information is collected
- efficiently as possible?

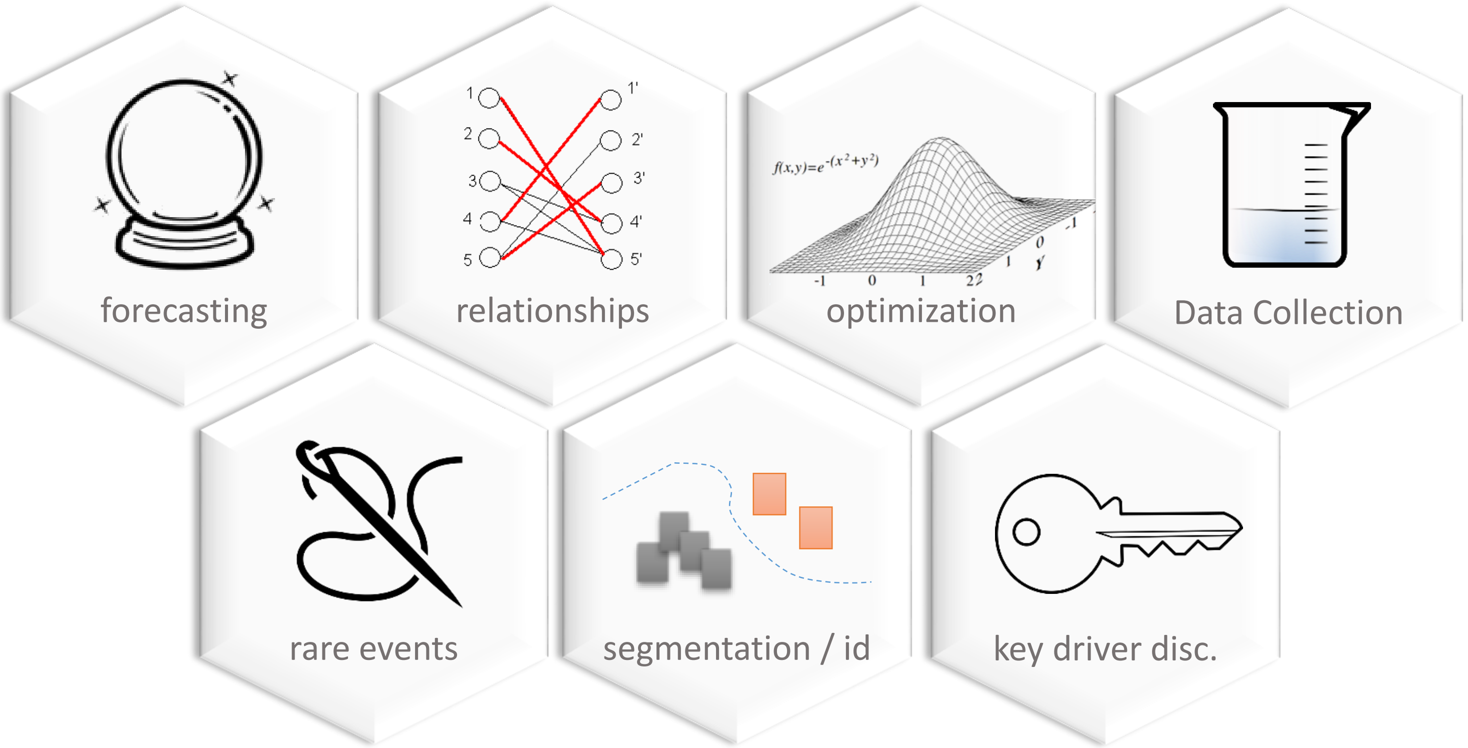
Data Collection

Surveys : how to collect data concerning a topic of interest.

- What future application features are most desirable?

Design of experiments : how to expend the least effort to get the data in order to answer the best question.

Levers – what levers need to be pulled to enhance/maximize member acquisition?


forecasting
relationships
optimization
Data Collection
rare events
segmentation / id
key driver disc.

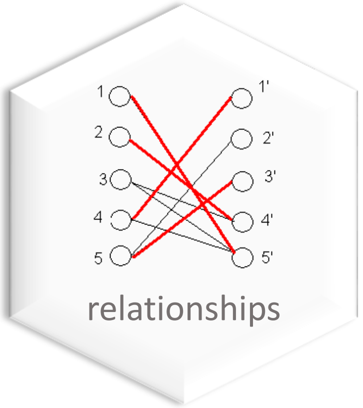
No so fast

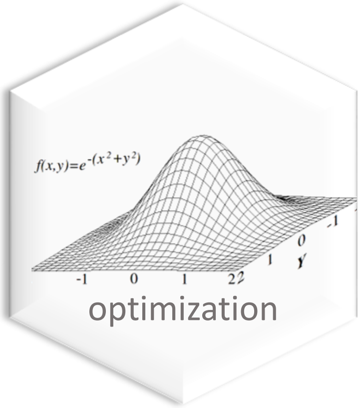
Solutions are not that simple

Solutions often involve more than one techniques

Recommend a product …

*that **increases stickiness**!*



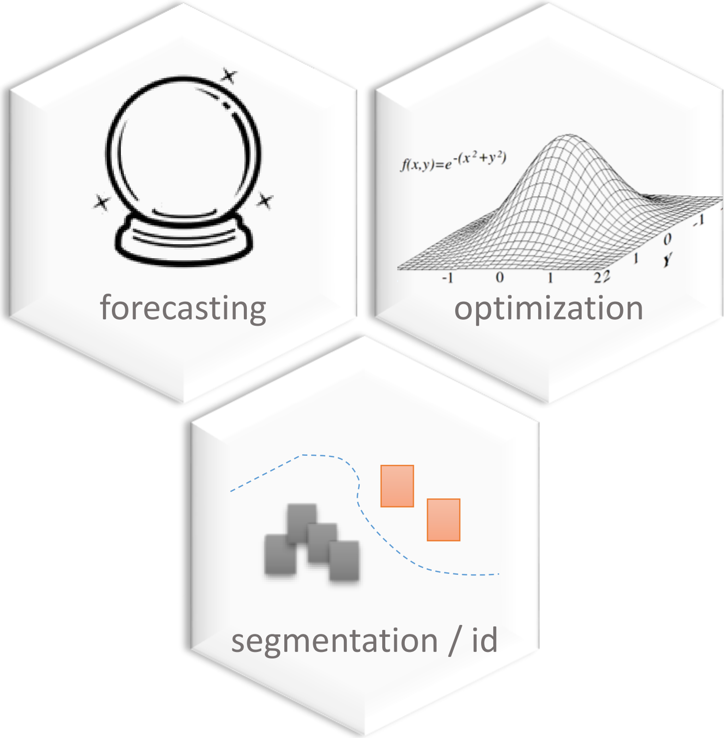


Provide member-tailored experiences that …

lead to ***deeper adoption***

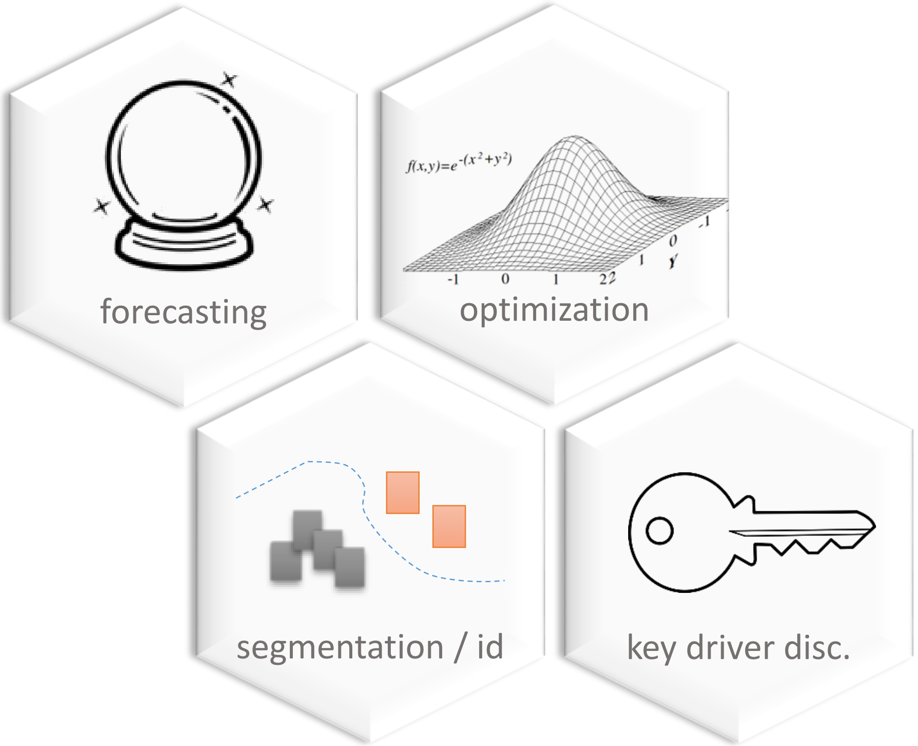
Which opportunities/leads should be pursued first…

to increase revenues

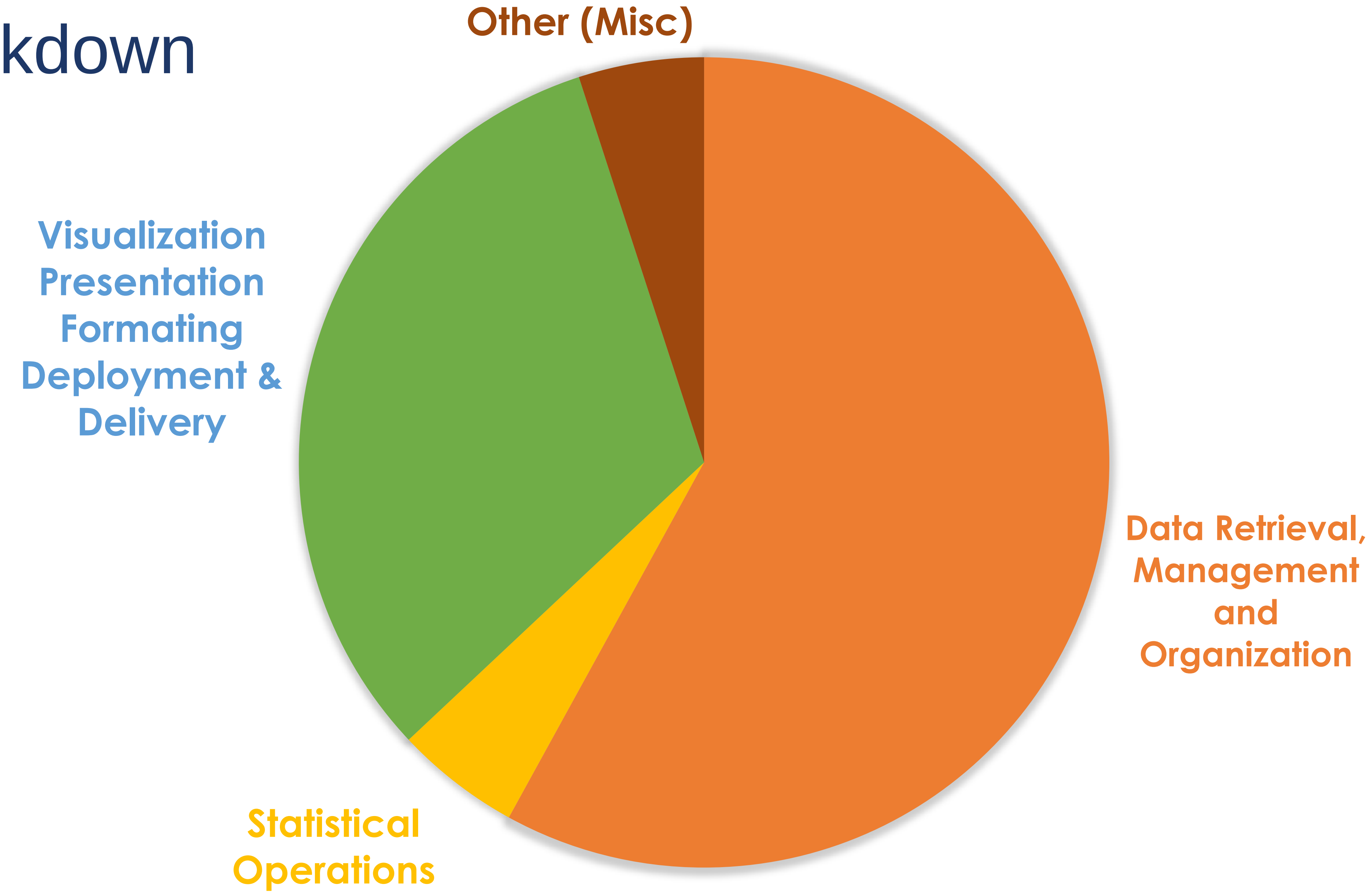


Improve member support outcomes by ….

setting ***optimal staffing levels, skills coverage and schedules***

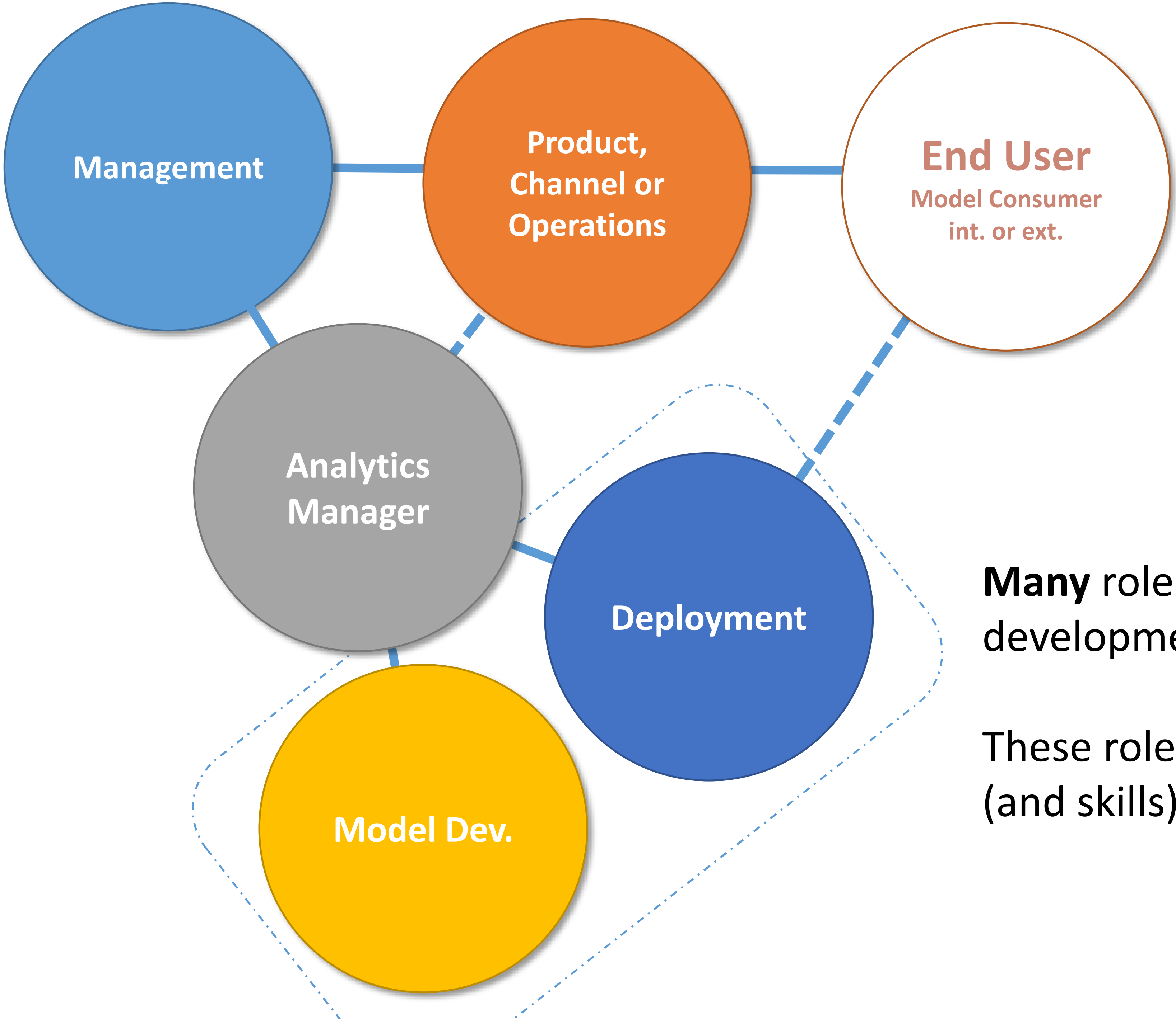


Task Breakdown



Data Science in the Organization

as Technical Discipline	**as Business Discipline**
Reports to technical organization usually closely associated with DW or BI team.	Reports to channel, product or operational department



Many roles involved in model development.

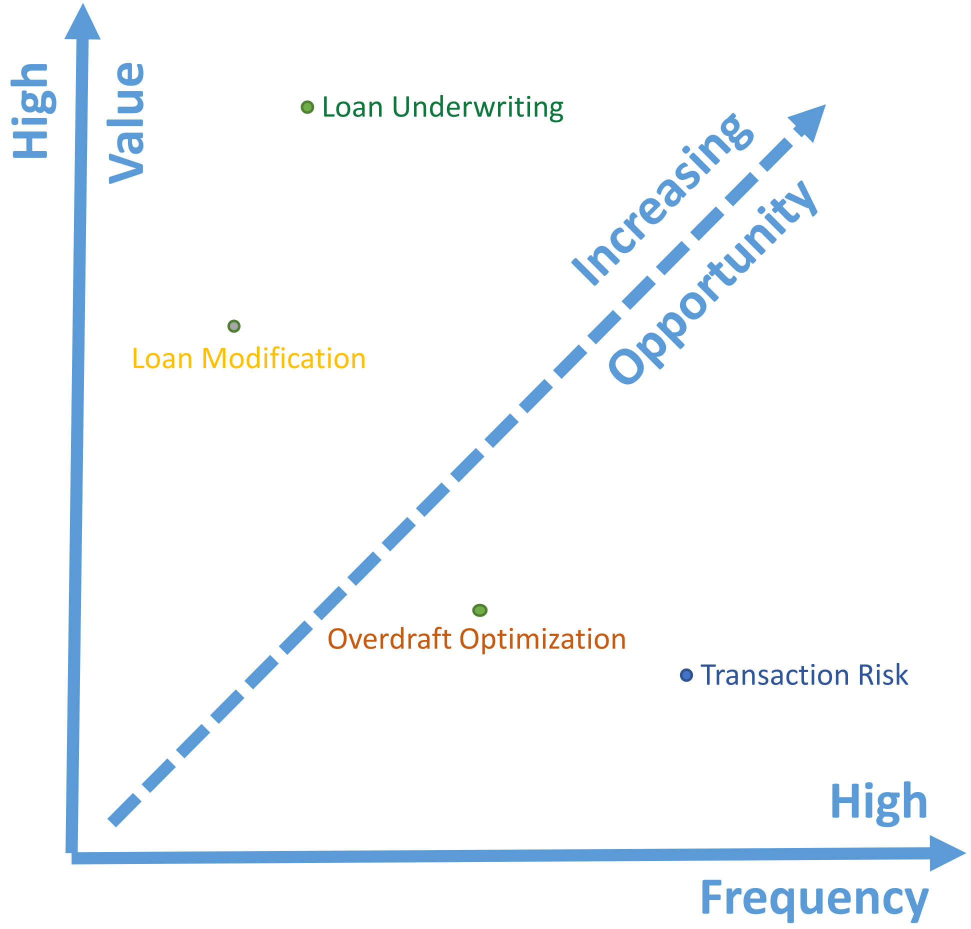
These roles align with responsibilities (and skills) not necessarily individuals.

Management

- Ensure proposed project aligns with organization strategic objectives
- Ensure benefits outweigh cost
  - Positive ROI
- Ensure project is prioritized
- Track ongoing performance

Types of Decisions

- High Frequency
- High Value*



* relative to existing operations or systems

Management Success is …

having a model
addressing the right opportunity
with the expected impact

Product, Channel or Operations

- Estimate benefits
- Identify deployment endpoints
- User testing
- Change management
- Identify business risk

Analytics Manager

- Consult with internal client (Product, Client or Operations)
  - Gather Requirements
- Manage Model Development and Deployment
- Enforce standards
- Testing
- Track performance metrics

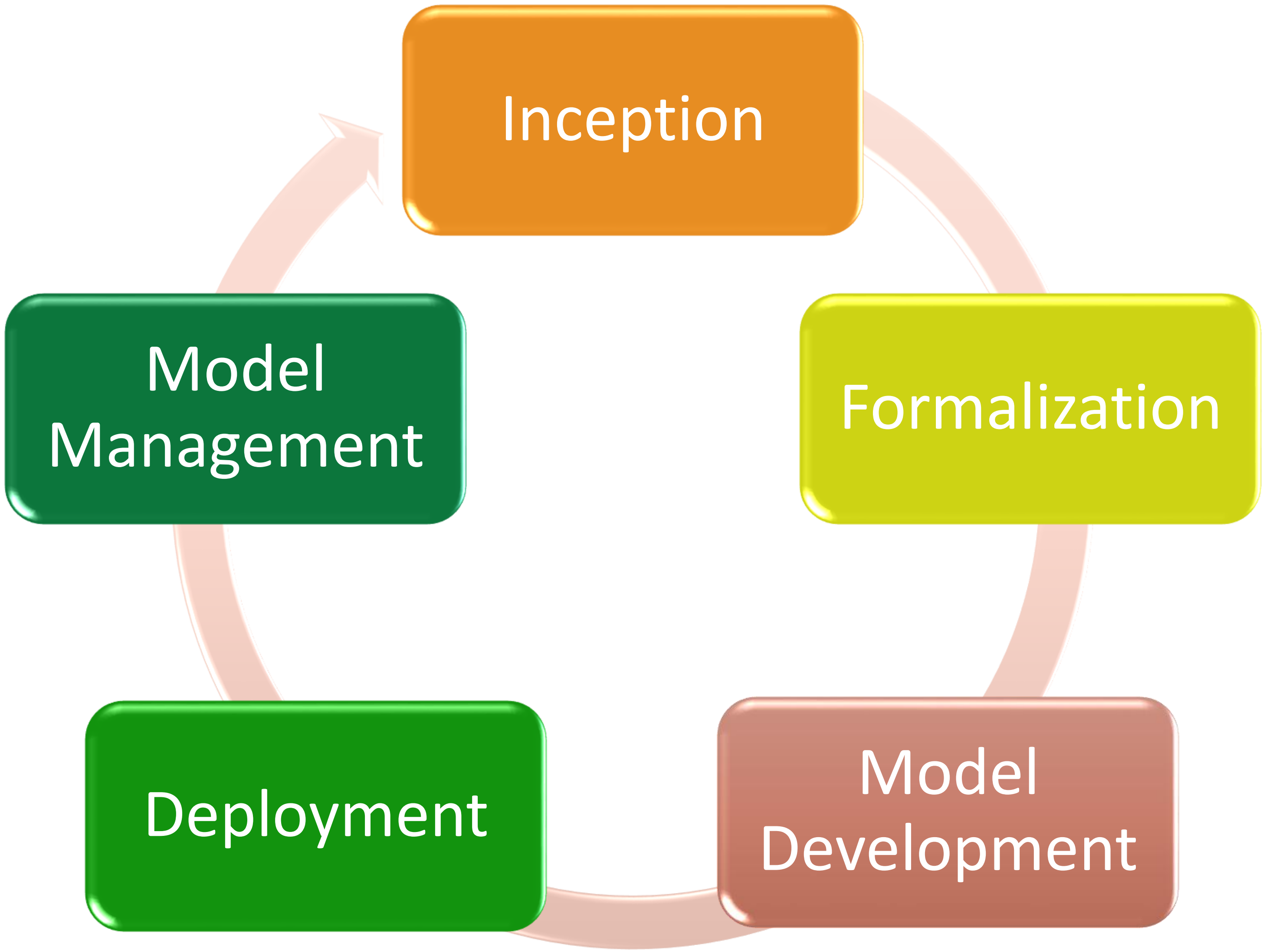
Model Development

- Build Model
  - Model specification
  - Feature Specification
  - Calibration
- Document model
- Ongoing model evaluation and maintenance model

Deployment

- Use feature specification and model specification to operationalize model
- Deliver scores to deployment endpoints
- Test and maintenance deployment

5 Phases


Inception
Formalization
Model Development
Deployment
Model Management

1

Inception

"I want help deciding …"

Determine Goal(s)

- State benefits
- Success criteria (near/long term)
- Timeline

Identify Data Sources (especially responses)

- **Deployment (if applicable)**
  - Use / delivery of model(s)

2

Formalization

"You said …
which means"

Formalize goals

- Identify success metric
- Quantify success criteria

Review Data Resources

Identified SME

Plan

- Estimate effort/time line
- Identify resources needed

3

Model Development

"this is what … can be done…"

Develop Code for

- Retrieving and cleansing data
- Transforming raw data into features
- Training model
- Scoring observations

Output

- **Model** Code/Specification
- **Report** on Model Performance

<code>

4

Model Deployment

"getting you what you want…where and when you want it"

Determine how and how often models are consumed:

- One time report
- Repeated reporting (BI)
- Applications
  - Embedded
  - Stand-alone Decision App

- Version controlled code for deployment

5

Model Management

“ensuring you
get what you want…”

Monitoring

Model review and re-training (if necessary)

- Based on schedule
- Based on performance

Continued development / Model improvements

Output: Performance Report