Best Practices For Scaling Data Science Across the Organization

Peter Wang – CTO - Anaconda
Guest Speaker: Kjell Carlsson, PHD – Senior Analyst - Forrester

May 2018

Data scientists & business executives are frustrated

At ~1/3 of firms models are deployed only sometimes, rarely or never*

"I've built 10 prototypes, none of them have gone live and I'm not confident any of them ever will"

We should drive so much **more business value** using data science

*Rexer Analytics Data Science Survey 2015

Typical challenges go beyond the data science team

Data Science

- Solves the **wrong problem** or insights are **not actionable**
- **Can not communicate** results

Data Engineering

- Stalls on **data access**
- Doesn't support key **data management technologies**

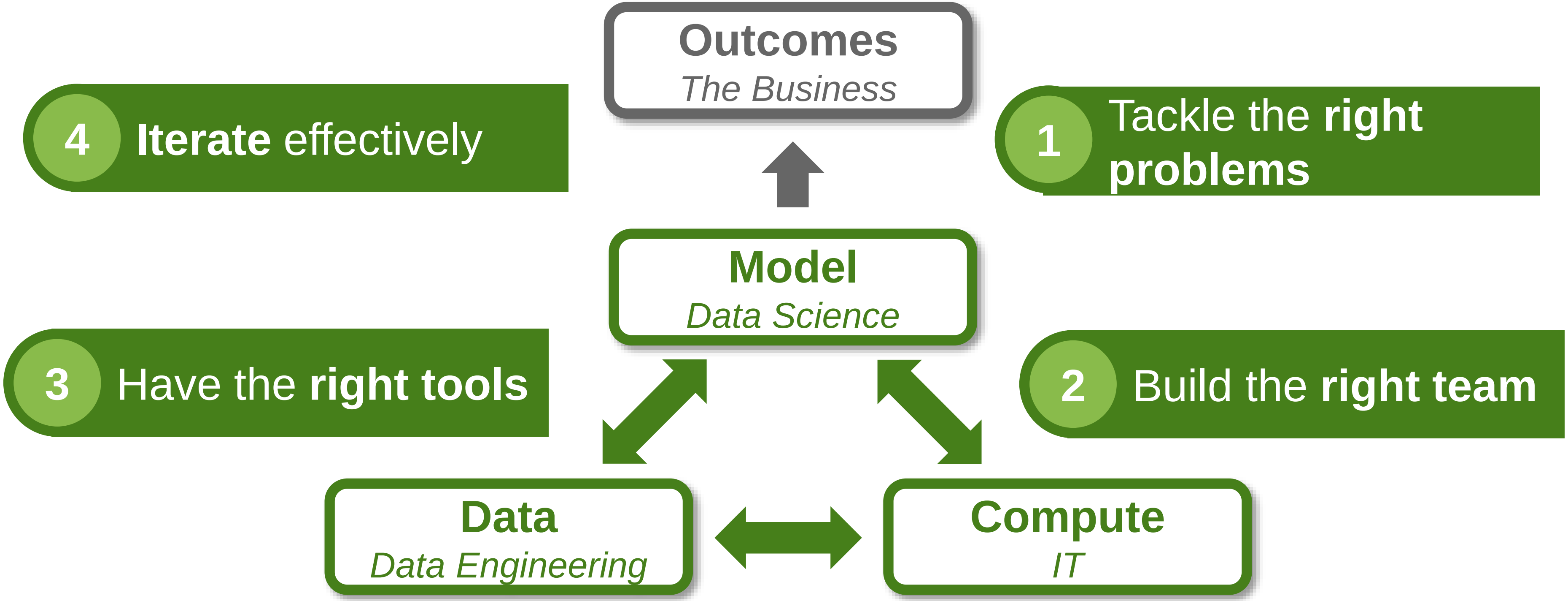
IT

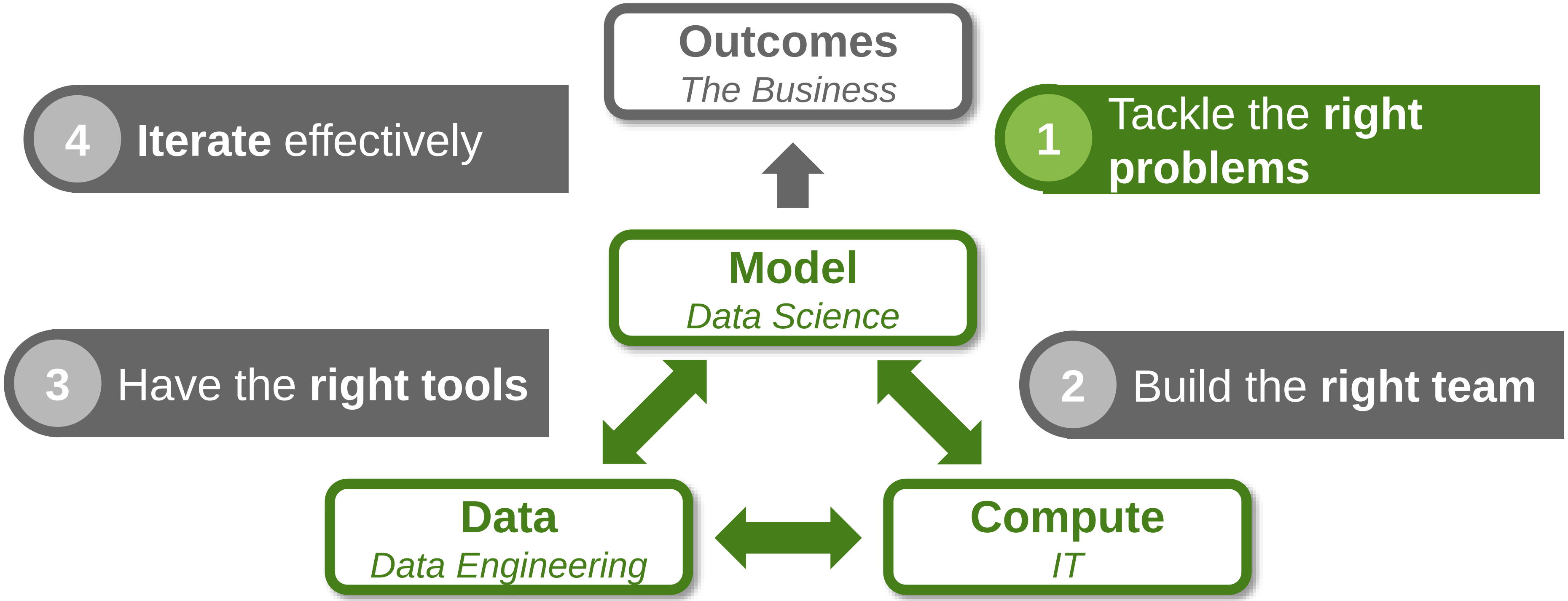
- Does not support the **infrastructure and tools**
- Abdicates **operationalization** and **app development**

The Business

- Does **not actively participate** or **advocate for resources**
- Does **not implement** the results

Effective data science is about aligning the right model, data and infrastructure with the right outcomes




Outcomes
The Business
4 Iterate effectively
1 Tackle the right problems
Model
Data Science
3 Have the right tools
2 Build the right team
Data
Data Engineering
Compute
IT

1 Tackle projects with large, clearly defined business value

The issue:

- Stakeholders do **not actively participate**
- Stakeholders do **not advocate for resources**
- Stakeholders do **not implement** the results

*"Identifying the **objective function** is key to getting everyone aligned"*

Benefits

- Drives **senior executive support**
- Minimizes **churn on goals & requirements**
- Ensures active, ongoing **stakeholder participation**
- Provides **commitment to take action** on results

How do you identify valuable projects?

1

Christensen's Jobs To Be Done Framework (Adapted)

What **"job"** would someone **"hire"** your solution to do?

~~Customer insight~~
Increase conversion, Reduce attrition

Who is the **"customer"**?

~~The organization~~
Sales, Account Mgmt.

How **else** could you do the job?

Marketing campaigns
Sales coaching

How **much** is it worth to them?

X% * $XM

Tackling the Right Problem

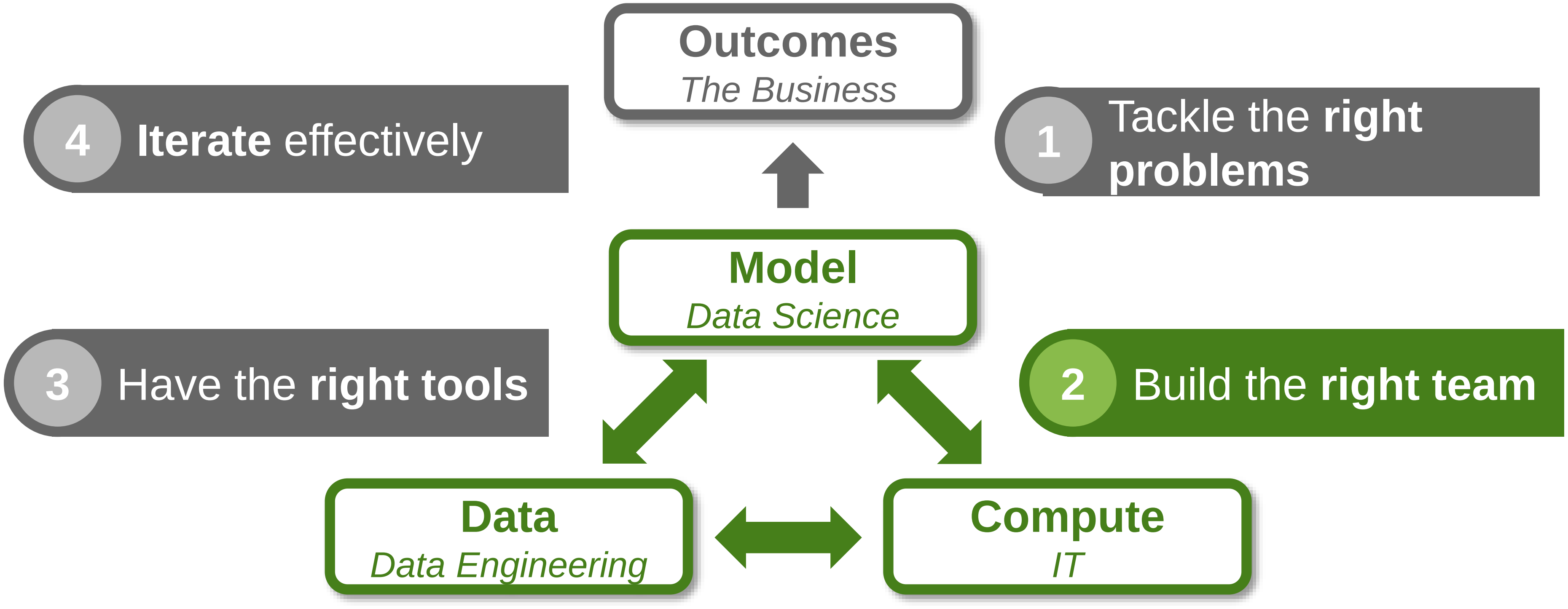
Business value	• Clear understanding of value to the business • Often counterintuitive. Be careful what you wish for: “You can’t handle the truth!” • Know your ~~customer~~ sponsor

Tackling the Right Problem

Business value	• Clear understanding of value to the business • Often counterintuitive. Be careful what you wish for: “You can’t handle the truth!” • Know your ~~customer~~ sponsor
Data availability	• Can we get timely access to the data we need? • Goldilocks problem

Tackling the Right Problem

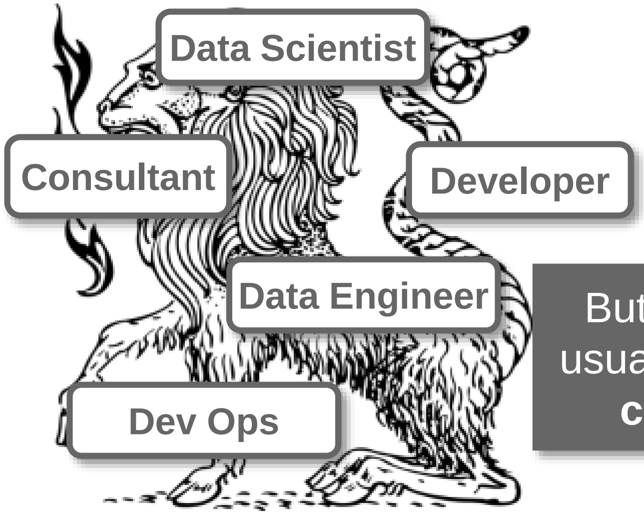
Business value	• Clear understanding of value to the business • Often counterintuitive. Be careful what you wish for: “You can’t handle the truth!” • Know your ~~customer~~ sponsor
Data availability	• Can we get timely access to the data we need? • Goldilocks problem
IT Deployability & Maintainability	• Realistic assessment of tech/skills • Select tools and approaches that fit within the IT capability envelope



Outcomes
The Business
Model
Data Science
Data
Data Engineering
Compute
IT
4 Iterate effectively
1 Tackle the right problems
3 Have the right tools
2 Build the right team

2 Build the right team

Companies complain that good data scientists are **unicorns**



But what they usually want are **chimeras**

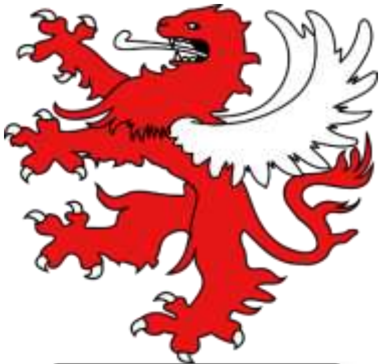
Chimeras are too **few**, too **expensive**, hard to **retain** and **inefficient**

Image Source: [Lilla Frerichs](), [OpenClipart-Vectors]()

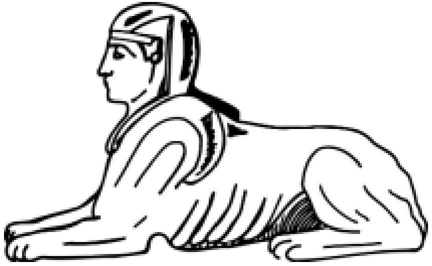
2 Build hybrid teams, not unicorns (continued)

Data Scientist + a bit of Data Engineer

Data Engineer + a bit of Data Scientist



Consultant + a bit of Data Scientist



Business Analyst + a bit of Developer

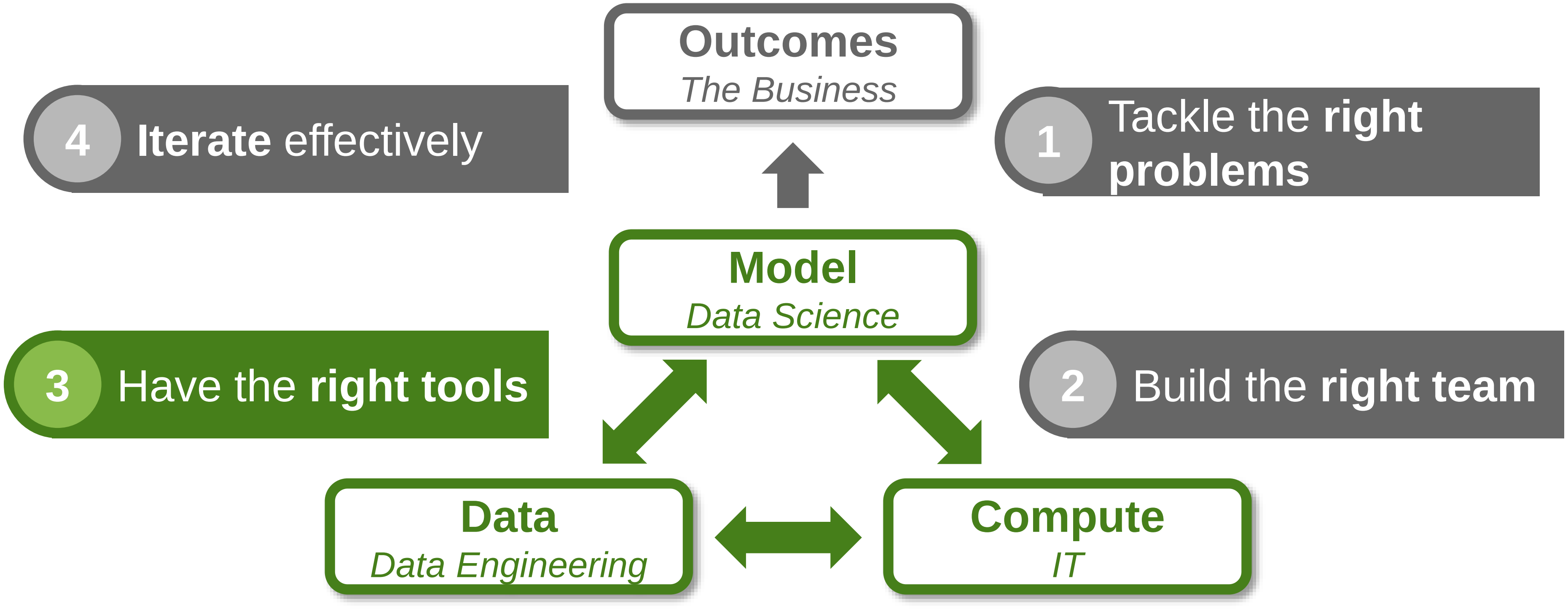
Teams of fantastic beasts are easier to **find**, **cheaper** and achieve **synergistic** results

Image Source: LadyofHats, Clker-Free-Vector-Images, Clker-Free-Vector-Images, OpenClipart-Vectors,

Building Data Science Teams

Because data science is *interdisciplinary*, the single biggest mistake is to pattern-match against existing tasks or technology footprint

Data Science Task	Existing Role
Query data, build reports	Analyst
Create datasets, define schemas and data reqs	Database specialist, Data engineer
Write code	Software engineer
Explore data, build models	Advanced Analyst, Predictive Modeler


4 Iterate effectively
Outcomes
The Business
1 Tackle the right problems
Model
Data Science
3 Have the right tools
2 Build the right team
Data
Data Engineering
Compute
IT

3 Deploy platforms for efficiency

Data Engineering Platform(s)	Model Development & Operationalization platform(s)	Visualization Platform(s)
› **Shared, re-usable data pipelines** accelerates data discovery and improves data quality	› Reduces need to **rebuild models** for deployment › Shares data science **knowledge** › Leverages **common infrastructure**	› **Faster development** of end-user apps › Broadens **access to insights**

Have the Right Tools

- Data science is code-heavy. Tools of choice are Python and R, but Matlab, SAS, and a few others are also used
  - Scala is mostly used with the Spark framework
  - Java, in a data science context, is most frequently used at the deployment stage by some teams
- Data Science != Software Development

Analyst

- Uses graphical tools
- Can call functions, cut & paste code
- Can change some variables

Gets paid for:
Insight

Excel, VB, Tableau,

Python

Analyst / Data Developer

- Builds simple apps & workflows
- Used to be "just an analyst"
- Likes coding to solve problems
- Doesn't want to be a "full-time programmer"

Gets paid (like a rock star) for:
Code that produces insight

SAS, R, Matlab,

Python

Programmer

- Creates frameworks & compilers
- Uses IDEs
- Degree in CompSci
- Knows multiple languages

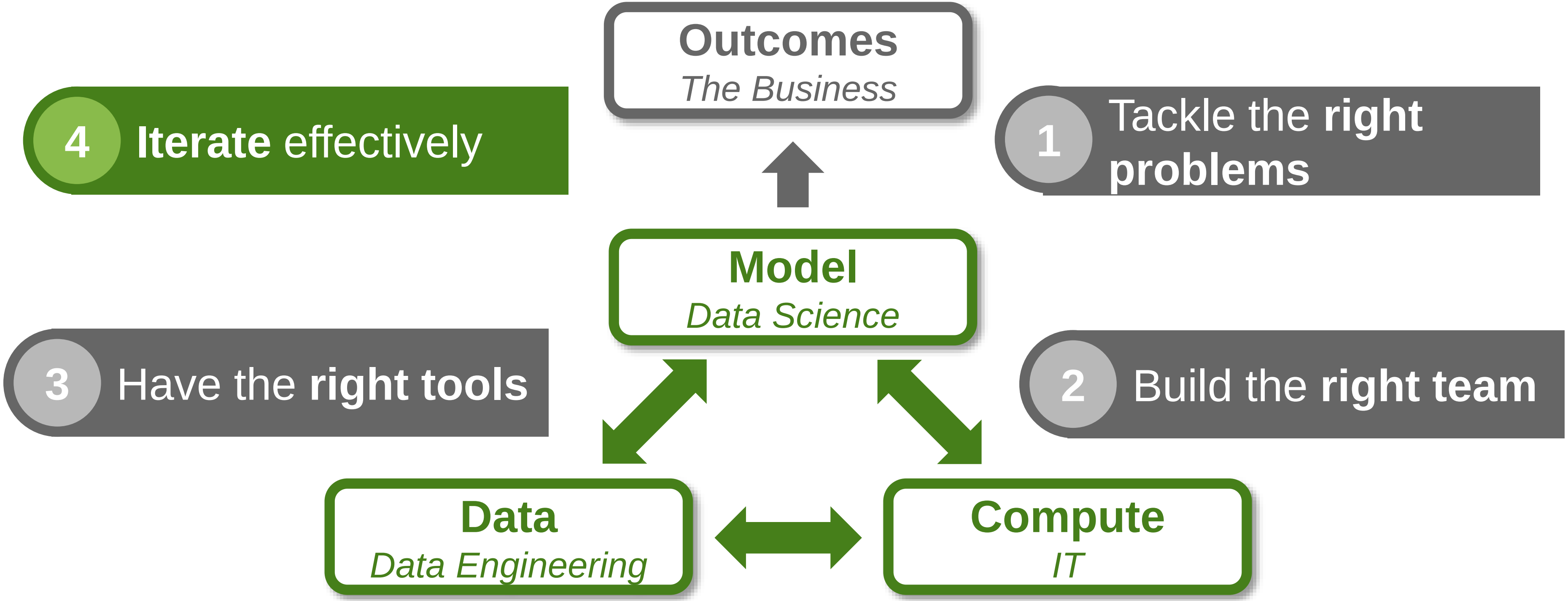
Gets paid for:
Code

C, C++, Java, JS,

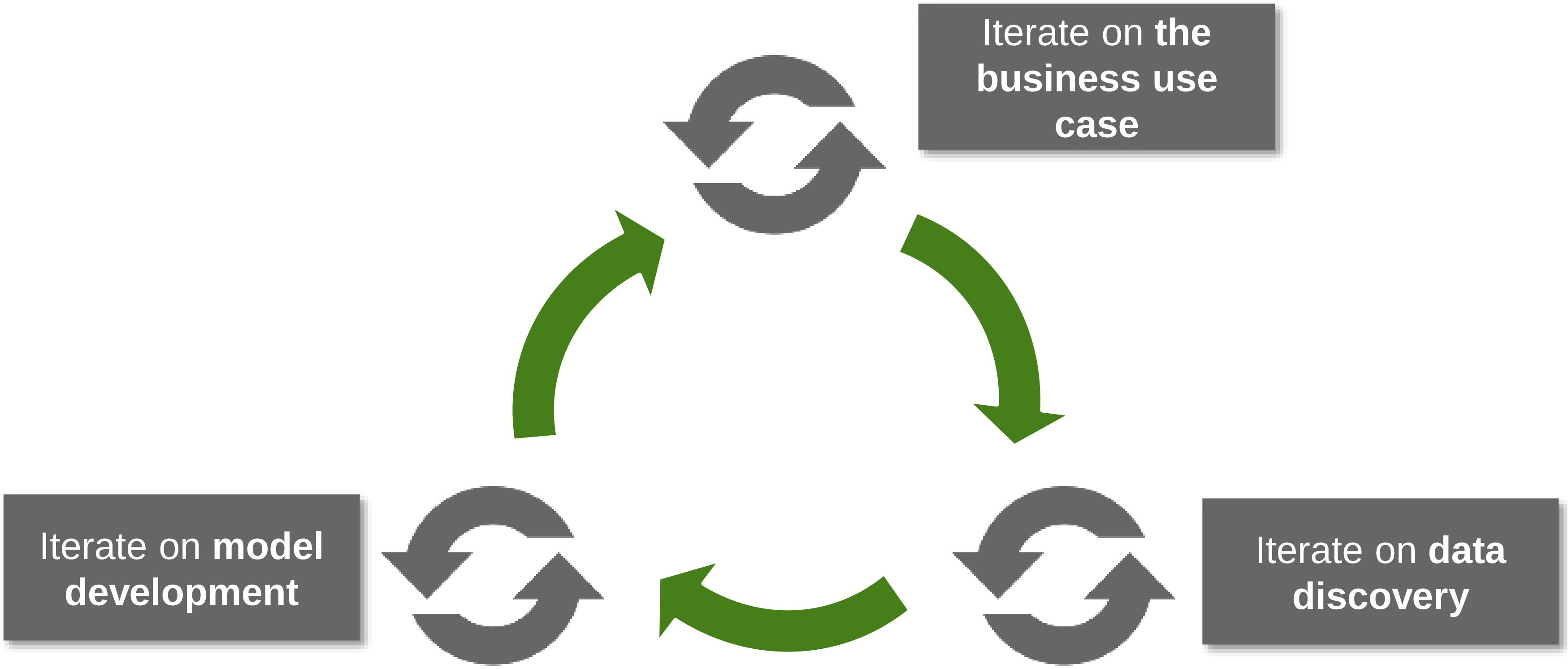
Python

Python - Most Misunderstood Language

- Python is probably the most misunderstood language
  - There are “tribes” and ecosystems in Python: web dev, scipy, pydata, embedded, scripting, 3D graphics, etc.
- But businesses tend to pigeonhole it:
  - IT/software/data engineering view: competes with Java, C#, Ruby…
  - Analyst, statistician view: competes with R, SAS, Matlab, SPSS, BI systems


Outcomes
The Business
Model
Data Science
Data
Data Engineering
Compute
IT
1 Tackle the right problems
2 Build the right team
3 Have the right tools
4 Iterate effectively

4 Successful Data Science projects are more agile than “Agile”



Iterate Effectively

- Exploration part of data science clearly requires agility
- But need to design for agile iteration of models once they are deployed
- Biggest hurdle to agile iteration is: How to break data science out of the sandbox in a way that is repeatable and maintainable?

Sandboxing Data Science

- Data Science Sandbox is on isolated network, outside of “GRC reservation”
  - Provides freedom to data scientists
  - Protects production ETL, DW, event processing
  - …but moving anything from Sandbox to Production is a huge pain
- Multiple orgs / LOBs interface with Data Science team in the mixed sandbox environment
- Compliance, audit, & risk control?

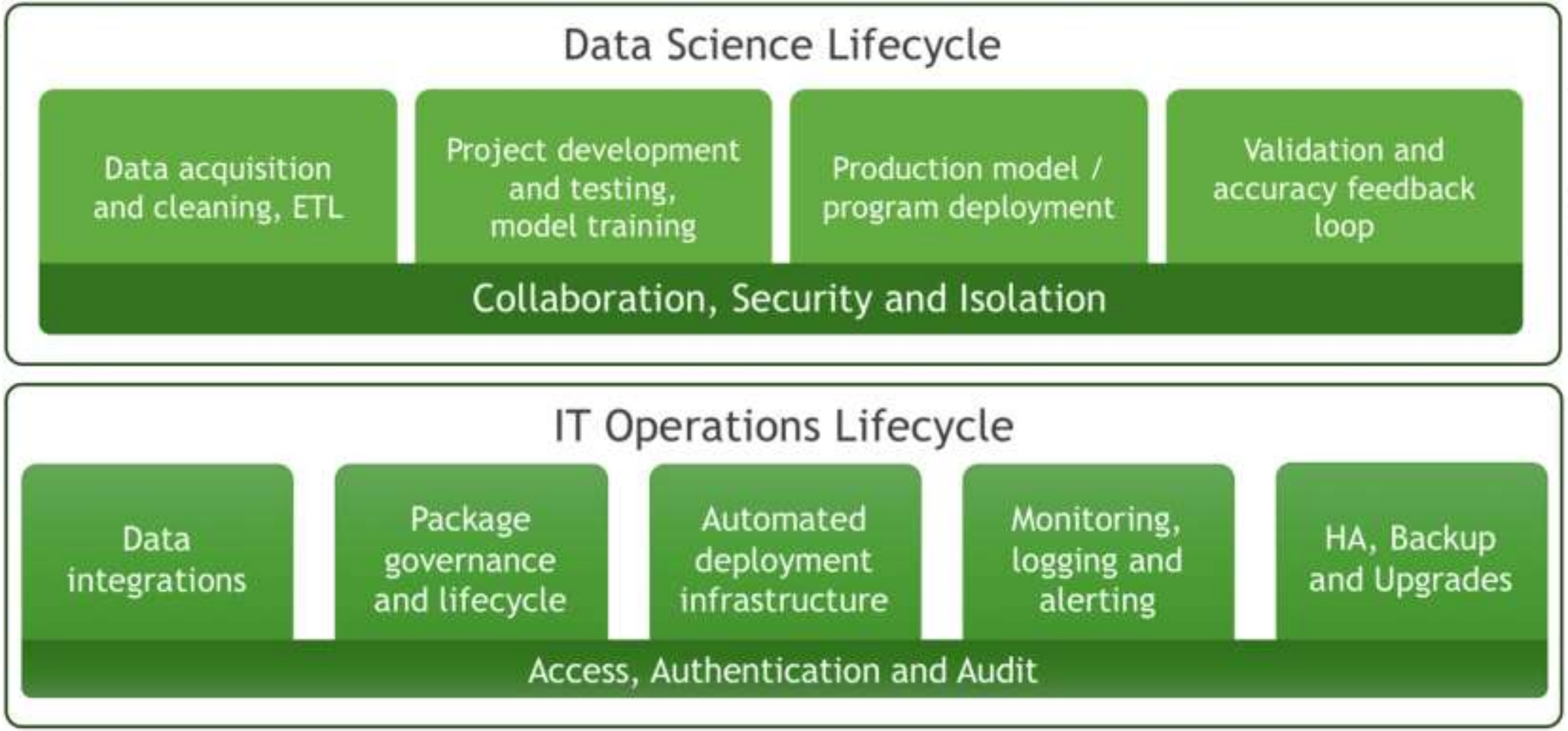
Contrasting Concerns

	Exploration	Production
Data	• Fast, unfettered access • Ease of introducing new, varied, messy datasets • Reproducibility	• Strict, governed access • Well-defined schema • Provenance & auditability
Compute Infrastructure	• High performance • Low latency, interactive • Individualized & specialized	• Scalable, high-availability • Manageable at scale • Cost amortization over many machines and users
Organization	• Individual high-achievers with lots of context & capability • Agile, able to quickly learn new skills and approaches	• Sustain operations at lowest possible cost • Robustness against unintended change

Looking forward: SDLC and ALM Have to Evolve

- DSDLC emerged in a time when software was (mostly) independent of hardware and data itself
- New applications are highly data-dependent
- New applications are blends of code, services, and specialized hardware
- Very different and broader set of change management, risk, governance concerns
- Cannot insist on tech monoculture (language, architecture, DBs): *The future is heterogeneous*

Data Science & IT Lifecycles



Anaconda Enterprise

Where to go next

Download Anaconda

https://www.anaconda.com/distribution/

Test Drive Anaconda Enterprise

ambassador@anaconda.com

Learn about consulting, training, and support

ambassador@anaconda.com

FORRESTER®

Kjell Carlsson, PhD
kcarlsson@forrester.com
Twitter: @kjellkeli

Peter Wang
pwang@anaconda.com
Twitter: @pwang

Thank you

FORRESTER.COM