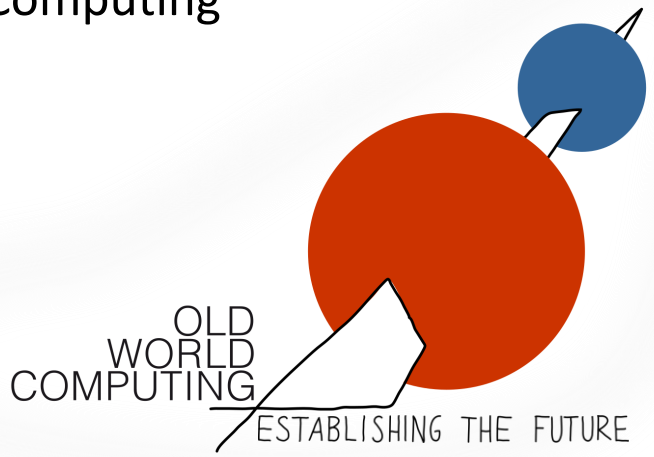
How to establish Data Science as a strategic capability in a company

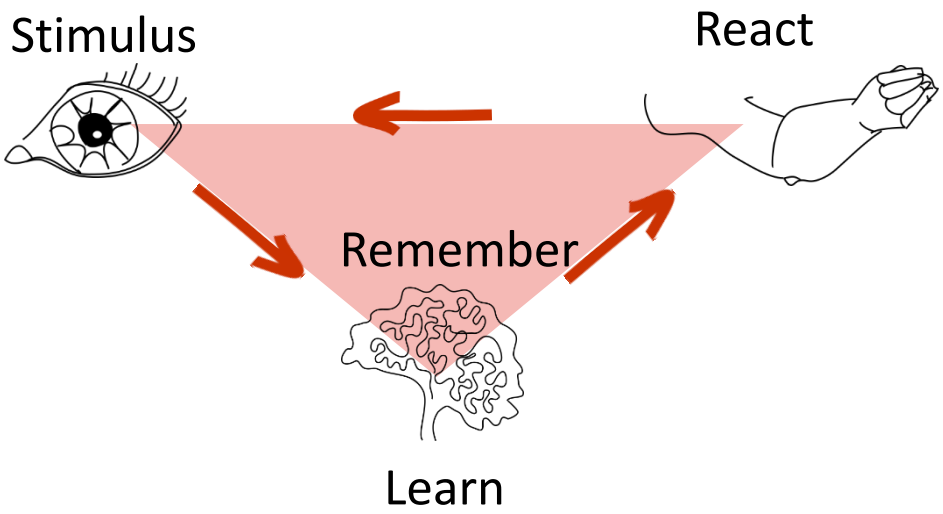
About Costs, Chances, Probabilities and Pitfalls

Sebastian Land - Old World Computing



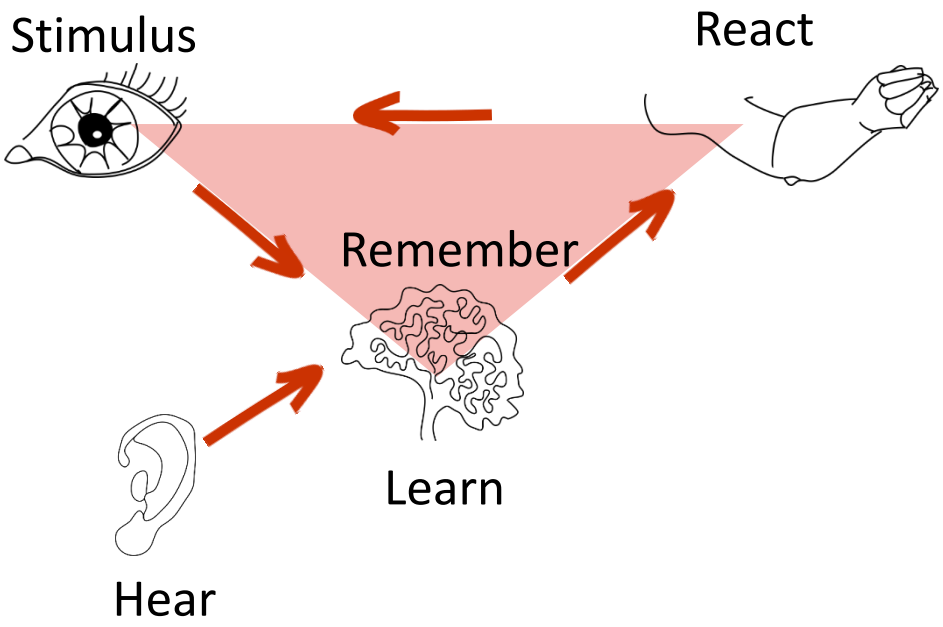
A short history lesson

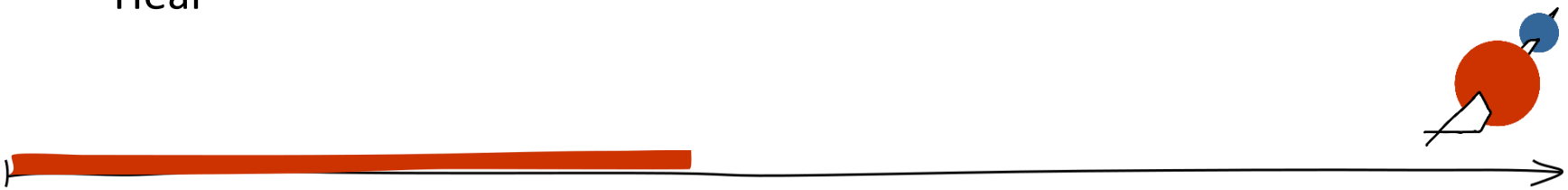
66 Million years ago: Impact helps mammals to become dominant species



A short history lesson

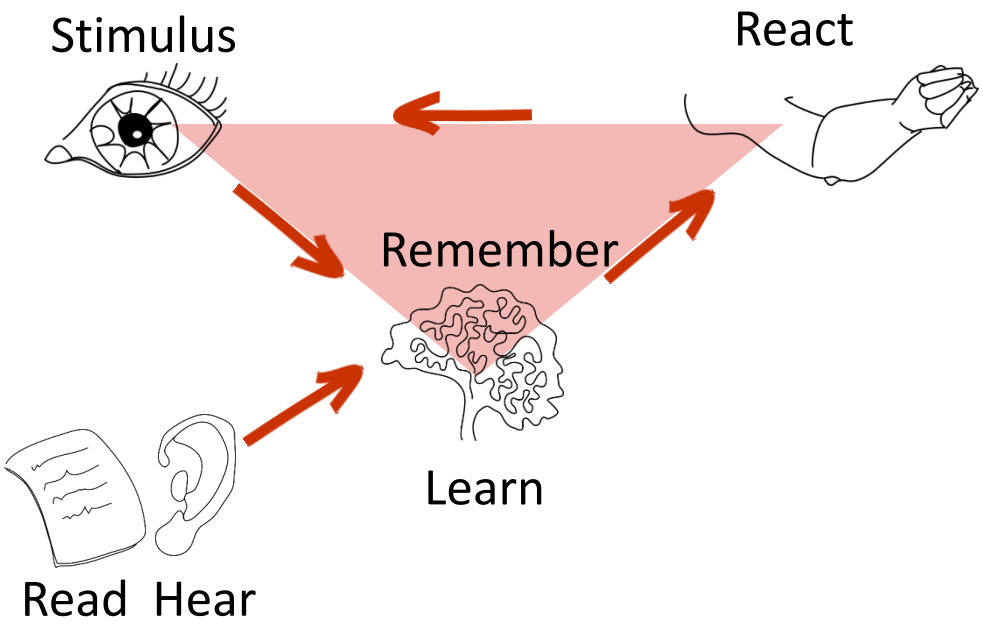
400.000 years ago: Language allows to abstract from reality and learn without experiencing





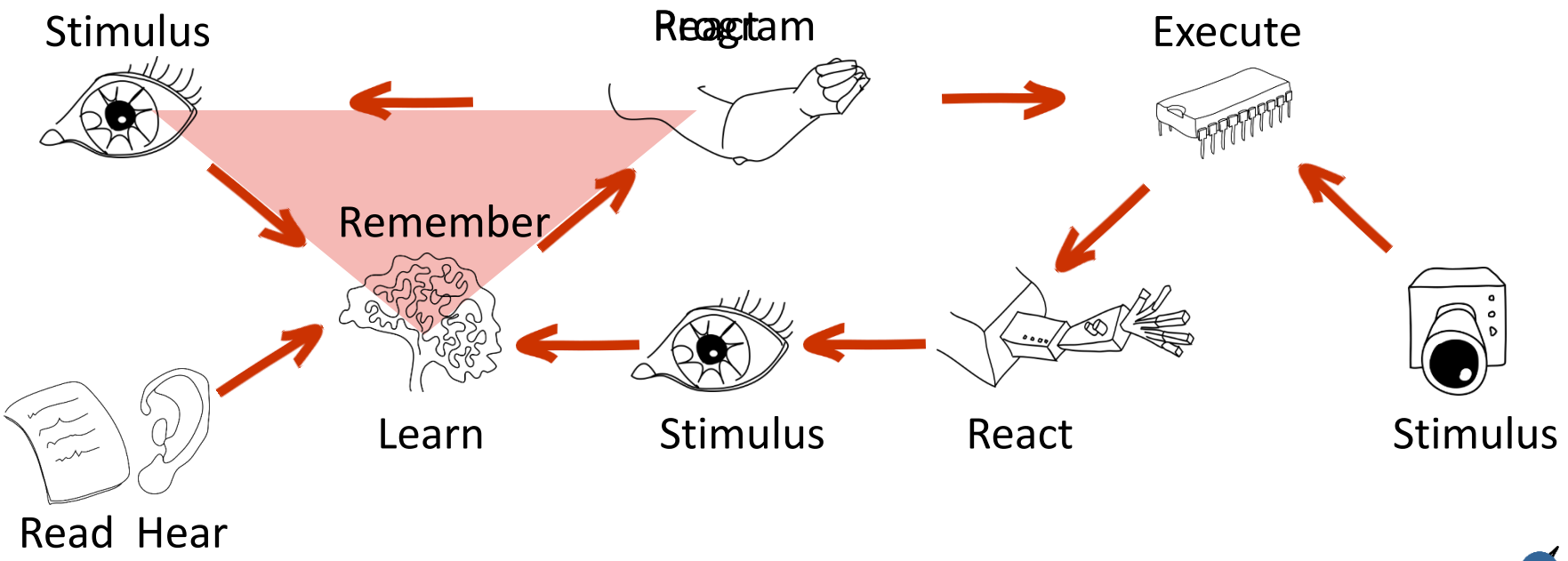
A short history lesson

A few thousand years ago: Writing fixes knowledge and allows learning without teacher



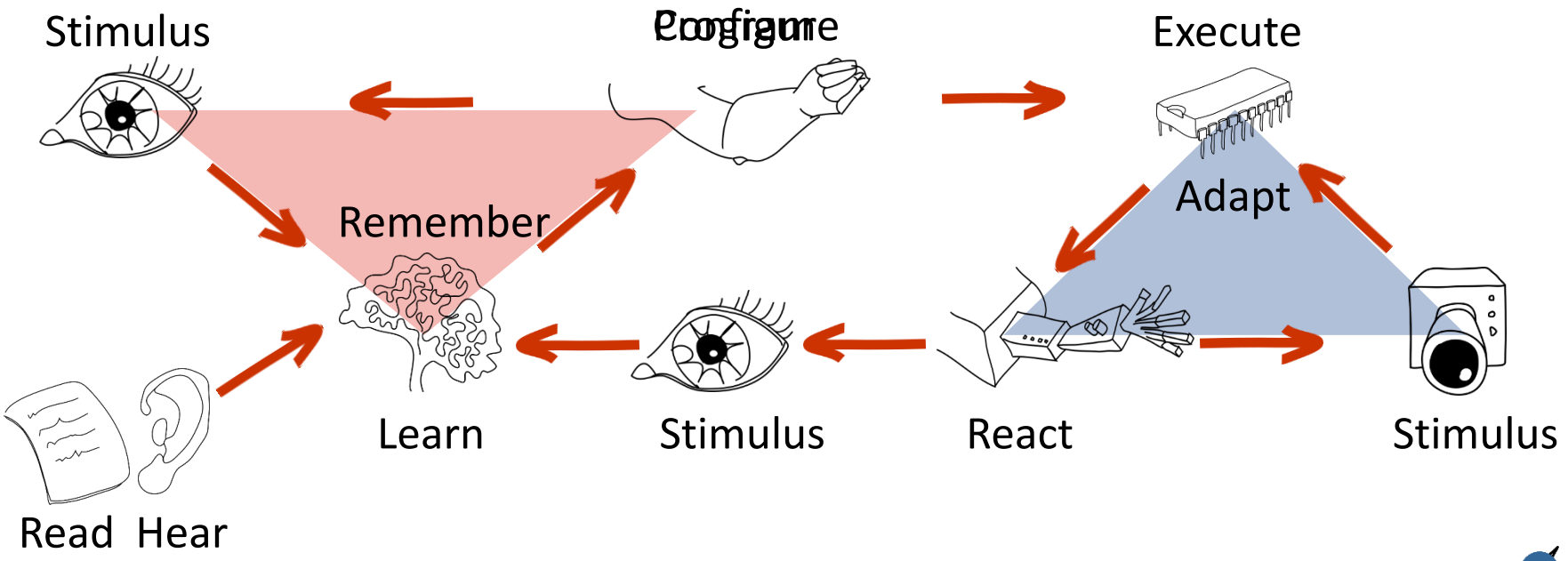
A short history lesson

1960: Software allows to store knowledge in active form that can execute itself in a static context



A short history lesson

2000: Machine Learning allows computers to adapt on changing conditions in the same context



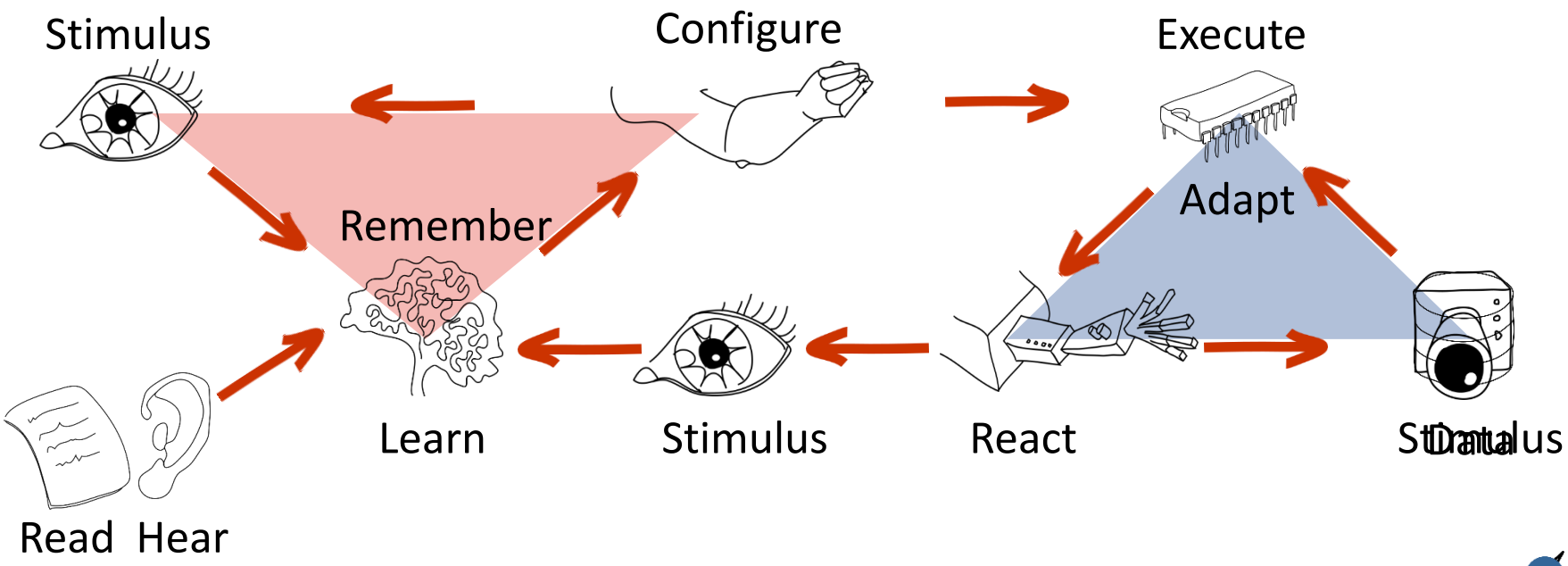
What is Data Science

First: What is data?

- Data is a description of reality like measurement values, customer‘s actions, server activity…
- Data by itself is entirely useless until interpreted by a brain to learn or act upon it
- Hence data consists of recorded stimuli!

What is Data Science

Changing the chart to something more familiar



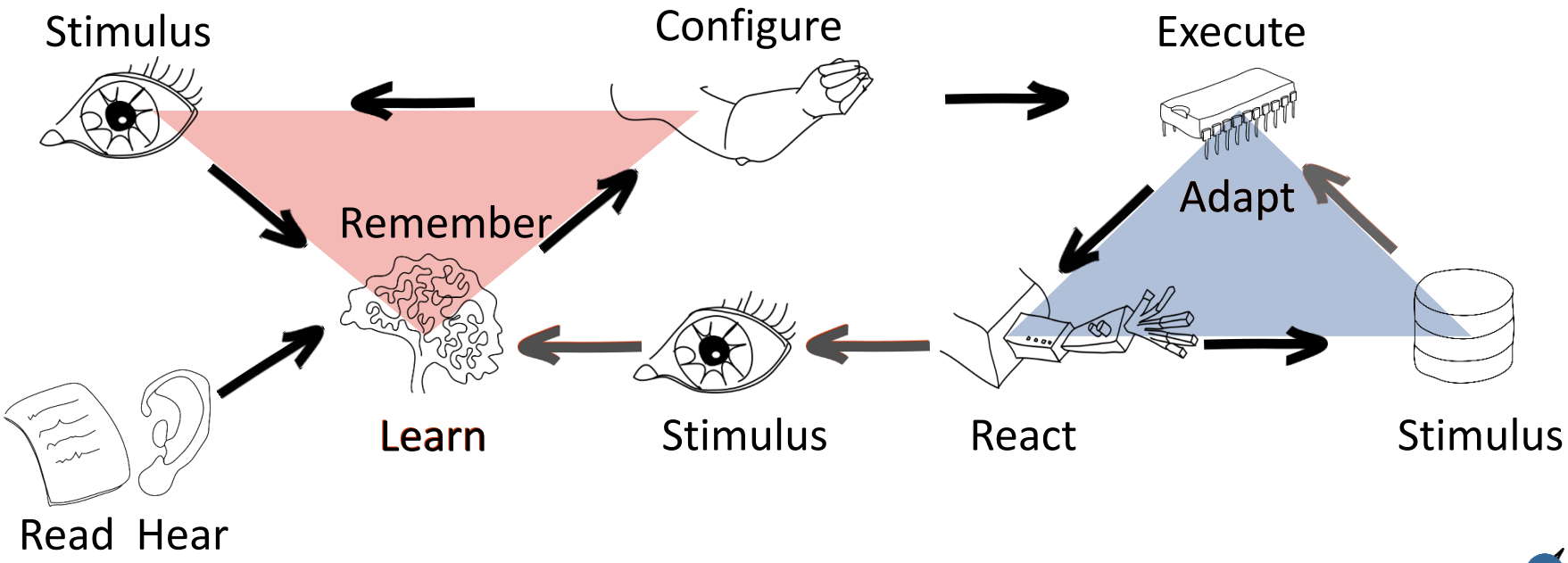
What is Data Science

That makes Data Science cover two major topics:

- How can we humans use data to learn to improve our reactions?
- How can we humans make machines use data to learn and react?

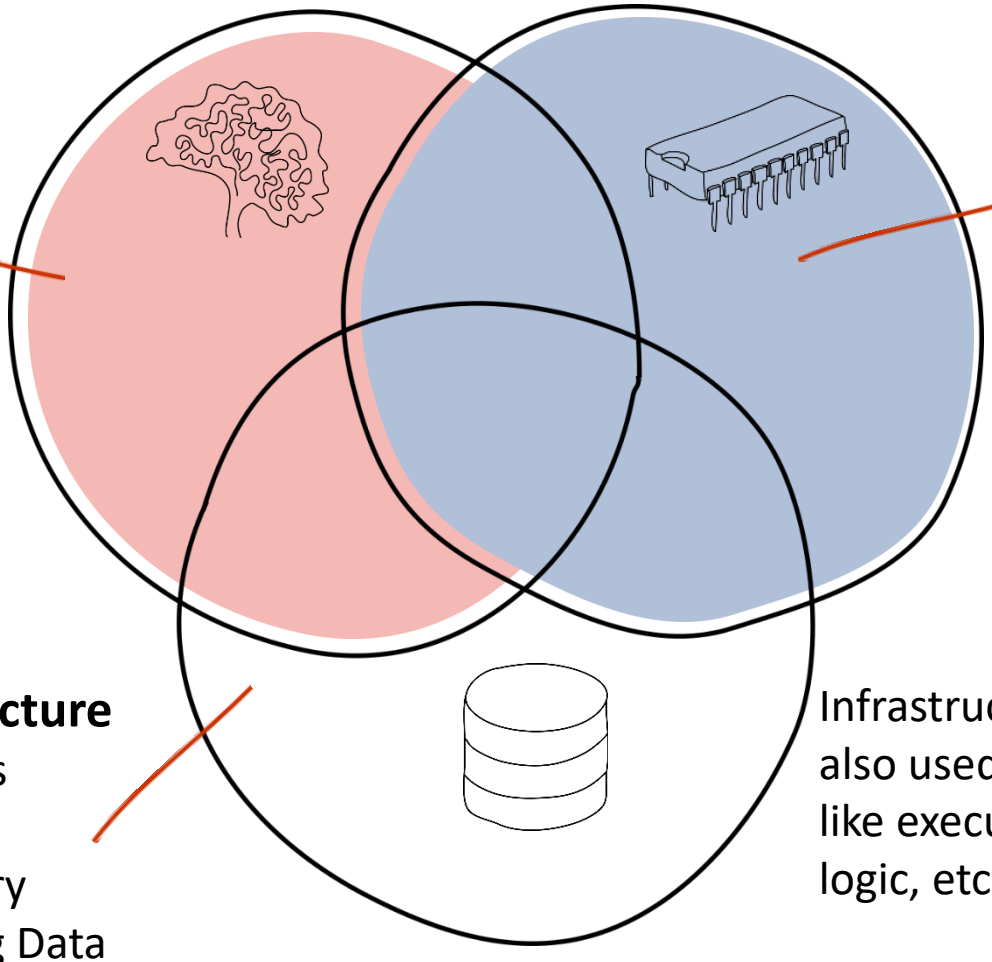
Complemented by several supporting infrastructural areas

What is Data Science



What is Data Science

Human Learning
BI
Reporting
Dashboards
Segmentation
Search
…



Machine Learning
Data Mining
Predictive Analytics
Deeplearning
…

Infrastructure
Databases
Big Data
In-Memory
Streaming Data
ETL
…

Infrastructure
also used for other tasks
like executing business
logic, etc…

Why Data Science?

- Nobody will argue that learning and improving our reactions is worthwhile
- But why should we use machines for learning?
  - Machines can process more stimuli -> smaller patterns can be found
  - Machines can process wider stimuli -> more influences can be taken into account
  - Machines are unbiased, statistically guaranteed correctness

Why Data Science?

- Machines can make more decisions -> individual tactical decisions can be made instead of strategical decisions
- Machines work 24h
- You don‘t need to pay them

How to introduce Data Science?

- If a company is at all in the digital age, human learning is already supported by data
  - Introduction is only technical problem
  - Improving to state of the art will not yield big gains
- Machine Learning may allow completely new approaches
  - Introduction of a new principle is social and technical problem
  - Depending on business model may yield extreme gain

How to introduce Data Science?

- We will focus entirely on machine learning here
- To learn how to do it, let‘s use our last advantage over machines:

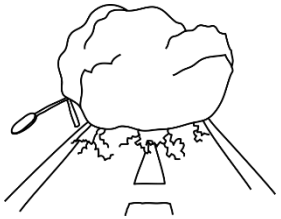
Learning from experience of someone else

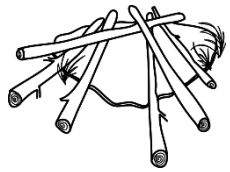
How was Data Science introduced?

- Three different ways how Data Science finds its way into a company:
  - Strategic decision by Management
  - Introduced by Head of Department
  - Applied by Domain Experts
- All three ways differ in the success-rate, efficiency and costs
- Let‘s start with the most common way…

Introduction by Domain Experts

- Domain Experts tend to be curious by nature
  - Already using Human Learning part
  - They look for new methods and tools
- At some point they will apply machine learning
  - Don‘t have the infrastructure to run it
  - And find it to be too complex
  - And find it to be easy and get it wrong
  - And find it to be complex so they get professional help
  - Become Data Scientists





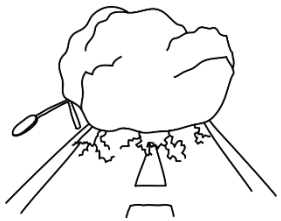
Introduction by Domain Experts

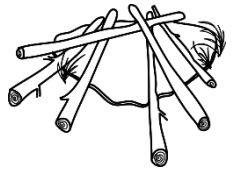
- Independent of how problem was solved, deployment is next hurdle
  - Scalability
  - Infrastructure
  - Reliability
  - Maintenance
- At this point the domain expertise becomes a disadvantage

Introduction by Domain Experts

Because

- Domain Experts are assigned other domain specific projects
- IT Department blocks any attempts as initiative is received as intrusion
- Domain Expert simply does not have knowledge and experience to get infrastructure right
- Has a good standing with IT and get‘s it running





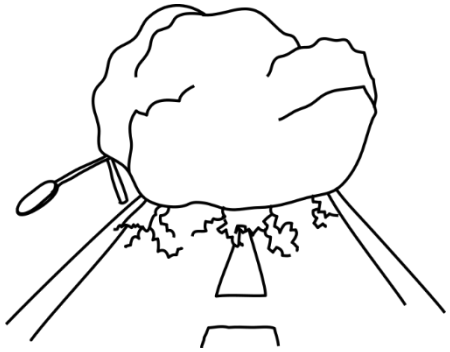
Introduction by Domain Experts

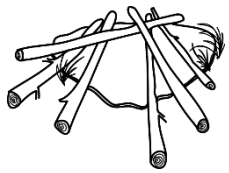
- Best case result is one solved problem
- But a lot of time and resources have been invested in this single problem
- Gained experience would make next problem much cheaper
- Data Science would need to become widely used to justify investment

Introduction by Domain Experts

But

- Domain Experts are assigned other domain specific projects
- IT Department blocks any further attempt as this would be intrusion
- No funding for establishing Data Science Team due to missing support from higher Management
- Work was carried out solely by externals. No Knowledge remains in house





Summary

- Domain Experts can solve single, one-time projects
- Domain Experts may even get it into productive deployment
- But introduction as strategic, company-wide capability will not succeed!

Introduction per Department

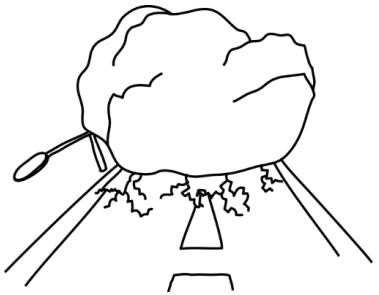
- Head of Departments can be under considerable pressure to run their business more effective
  - Always looking for new tools improving productivity
- At some point they may decide to invest in Data Science
- Someone or a small team is assigned to find out more about it

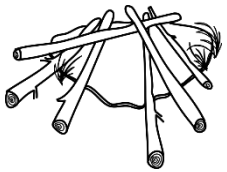
Introduction per Department

- They will likely face the same problems as the single Domain Expert
- But are in a better situation due to more resources and a considerable budget
  - Getting external support for a kick start
  - Having infrastructure covered
- But they miss domain expertise!

Introduction per Department

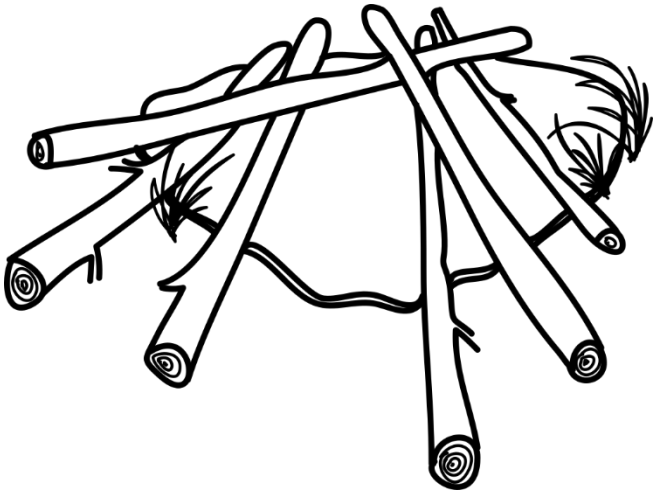
- Without support from Domain Experts introduction is likely to fail
  - Domain knowledge is needed to find use case
  - Domain knowledge is often required to find a working solution in the first place
  - A solution is only useful if actually used by the Domain Experts





Introduction per Department

- This creates a new and unexpected social problem
  - Learning machines make many people fear for their jobs
  - If they are not driving the project Domain Expert may belong to them
  - And refuse any support or actively sabotage the project



Introduction per Department

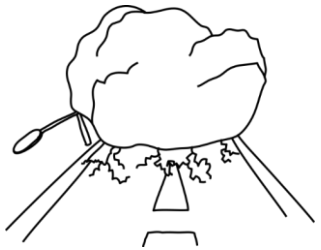
- So a dedicated team needs to have a full set of skills
  - Data Science skills to get results right and don‘t make false promises
  - Infrastructure skill to ensure usability, security and availability
  - Teaching skill to excite end users and domain experts to identify use cases and access domain knowledge
  - Diplomacy skills to avoid fear and resistance within the company to access domain knowledge

Introduction per Department

- Missing skills reduce success probability drastically
- In best case they can solve several problems within the department
- But soon they will need to access data that belongs to another department

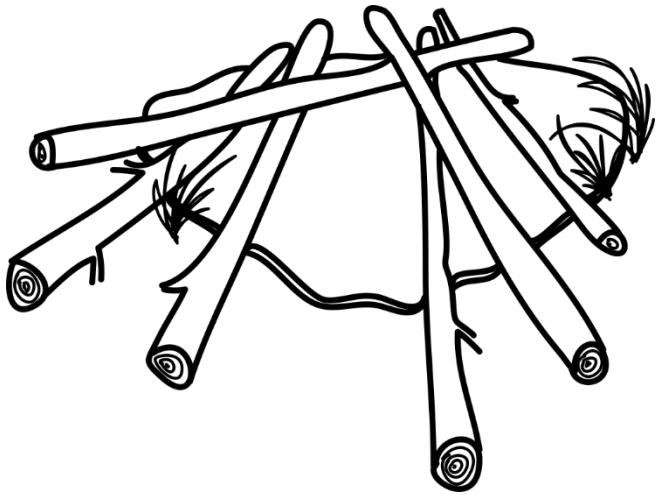
Introduction per Department

- Depending on Company Culture this can cause two problems
  - Other department is not willing to cooperate as it receives it as an intrusion
  - Other department may simply lack resources to do so
  - Other department agrees but that is bought by delivering services



Introduction per Department

- Delivered Services may either
  - Distribute the usage of Data Science over the company
  - Or being only data centered and just consuming up the team‘s resources



Summary

- Many problems can be solved in the department
- Approach much more efficient due to synergetic effects
- Department-wise introduction may even establish Data Science in the entire company

Summary

- But more likely than complete introduction
  - The resources and budget are drained
  - The team is blocked by other departments
  - The project was successful and triggered higher management to start initiative on it‘s own

Strategical Introduction

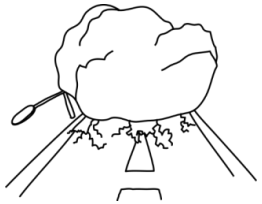
- The Top Management may have heard from this Data Science thing
- At some point they may believe the buzz and decide to invest in Data Science
- Chicken-Egg Problem:
  - How to invest if you don‘t have first hand experience
  - How to get first hand experience if you don‘t invest

Strategical Introduction

- Several solutions present itself:
  - Trust the word of the sales guys or CEOs from the usual IT vendors
  - Hire the usual management consulting company to make an investment plan
  - Hire a huge team from a large consulting company to get done whatever needs to be done using Open Source
- All solutions promise to transform a company within a limited time

Strategical Introduction

- IT Vendors have their own interests
  - A large infrastructure not able to process the real use cases
  - A fitting, but very large infrastructure



- Most management consulting companies missed Data Science completely until recently
  - Will organize vendor selection and use case identification process
  - Recommend infrastructure not optimal for real use cases
  - Recommend optimal infrastructure

Strategical Introduction

- Consulting companies always have mixed interests
  - Infrastructure may solve requirements but with large maintenance effort
  - Infrastructure may be only addressing requirements but lack flexibility beyond
  - Infrastructure may be perfect but knowledge remains with consulting company

Strategical Introduction

- Infrastructure can be bought with money
  - With the risk being locked-in by software
  - With the risk being locked-in by knowledge
  - Higher costs than necessary
- But infrastructure only provides basis
  - Data Science projects need to be conducted
  - And results put into productive use

Strategical Introduction

- Central Data Science Team needs to be established
- Requirements and problems are the same as for department team
- Chief Data Officer may solve problem of getting access to data

Summary

- Strategical introduction of a central data science team by management is most promising way
- But without experience path will contain expensive loop ways
- Fast and forced introduction will likely result in failure due to restrictions in infrastructure and social rejection

What we learn

- Embracing Data Science involves various challenges
  - Missing understanding of the principles
  - Missing Data Science skills and experience
  - Social problems
  - Missing resources

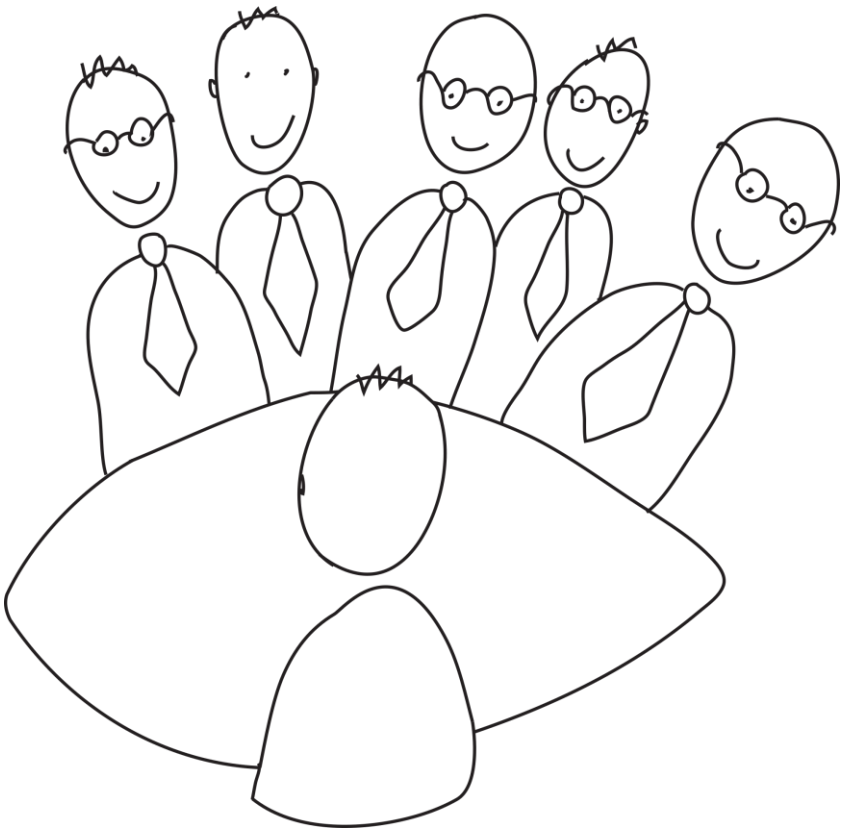
- All of these points have to be addressed to make the transformation successful!

What we can deduce

- Approach this top-down
- Start small and grow exponentially
- Give it some time
- Make sure everybody participates
- Take care of the social factor

- How can we do that?

Phase 1: Strategic Consulting



Targets:

- Bridge the knowledge gap
- Identify areas of application
- Setup corporate structure

Bridge the knowledge gap

- Data Science is a new technology that's fundamentally different
- The principle needs to be understood by decision makers
  - Otherwise costly investments will be wasted and precious time is lost
  - Social problems will rise in the company
- Access experience to look behind marketing buzz words

Identify areas of application

- Data Science is very flexible, it can be used
  - In many different ways
  - In different sectors
  - For different use cases
- Not all are equally complex and likely to succeed
- Some require more or less investments and more or less time
- Some will have more or less social impact

Identify areas of application

- Access experience to identify use cases that are
  - Low hanging
  - That contribute real value
  - That will promote Data Science
  - Without frighten anybody
- This use case should also provide a good foundation for further use cases in terms of
  - Experience
  - Infrastructure

Setup corporate structure

- Applying Machine Learning will impact the social structure of a corporation heavily
  - Doers will become controllers
  - Experience will be taken over by computers
  - Entire jobs may be taken over by computers
- This change needs to be taken into account carefully
- Acceptance must be grown to avoid failure
- Data Science depends on collaboration

Setup corporate structure

- Make someone responsible for company wide data based collaboration: CDO
- Create central Data Science team
  - Should be independent from IT and manage infrastructure themselves
  - Should have a long term budget and mid term targets
  - Team members should be chosen carefully to combine the full skill set
  - Gather interested, curious and intelligent employees
  - Prefer intelligence and a flexible brain over experience

Setup corporate structure

- Make sure Domain Experts have the resources to collaborate!
- Make sure Data Science team has resources to train and excite Domain Experts
- Avoid impression that jobs will be lost
  - Job guarantee
  - New business area
  - Internal trainings

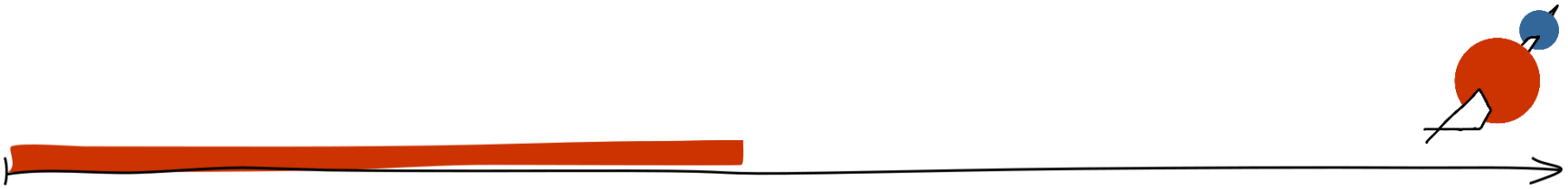
Phase 2: Training

Targets:

- Establish right mind set and grow in house expertise using external experience
  - In-Depth for Data Science Team
  - For interested sponsors throughout the company
- Learn a common language for Machine Learning

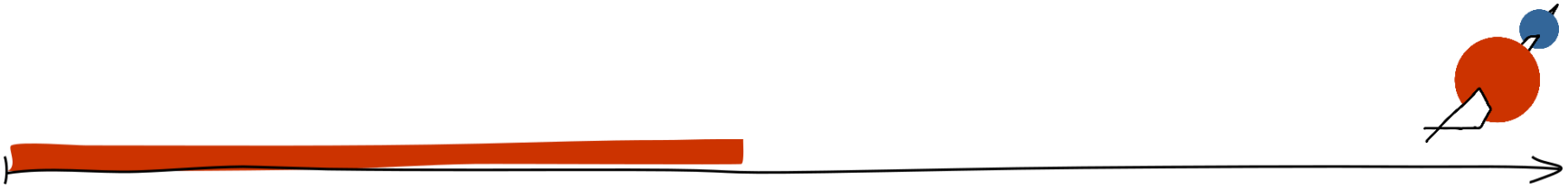
Shift of mind set

- Machine Learning requires to „let go“:
  - We humans don’t need to understand a problem
  - We guide the computer to understand it
  - And let it make the decisions for us
- This contradicts what humanity has done ever before!
- Trust is necessary for that



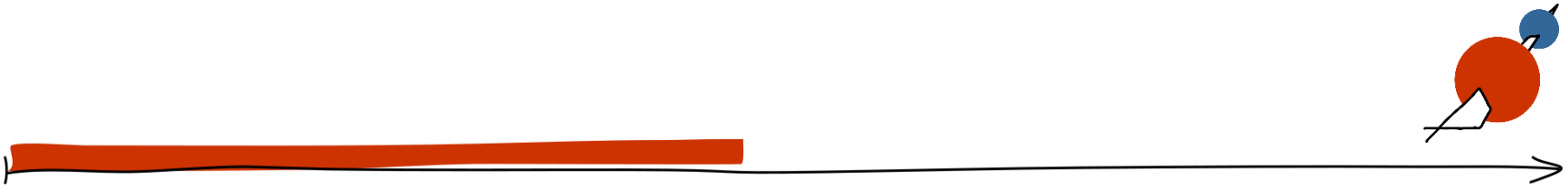
Shift of mind set

- To trust a person, you have to get to know her
  - Her way of reasoning
  - Her experiences
- Same for computers: You need to understand how they work
- Understanding also necessary to identify possible use cases throughout the company

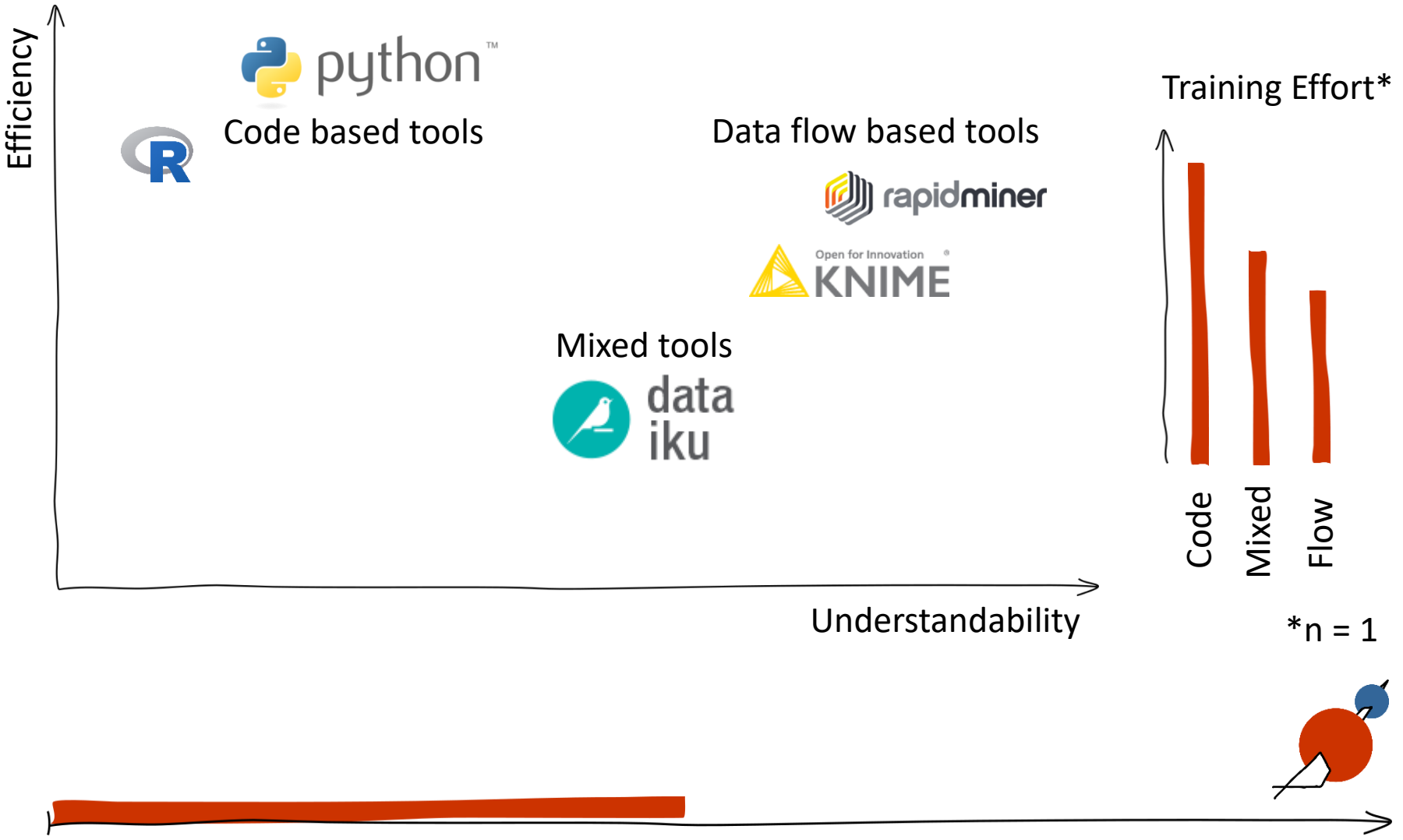


Learn a common language

- Data Science requires collaboration:
  - Data Scientist doesn't understand problem
  - Domain expert doesn't have Data Science experience
- Tool selection needs to reflect a trade off:
  - Domain Expert has to understand Data Scientist
  - Data Scientists need efficient tool



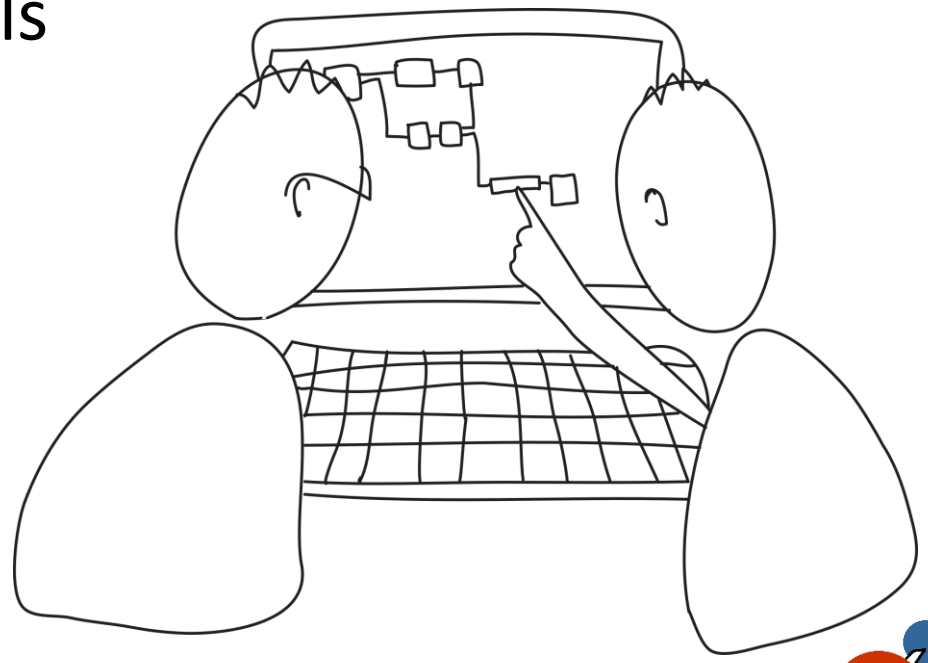
Learn a common language



Phase 3: Learning by Doing

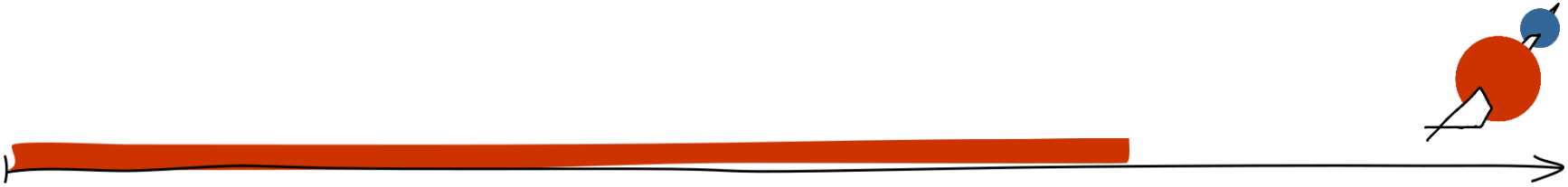
Targets:

- Establish steady knowledge transfer from externals
- Benefit from best practices
- Distinguish methods and infrastructure



Establish knowledge transfer

- Becoming expert in Data Science is a full time job
  - Needs to gain a lot of experience
  - Needs overview over all methods to creatively combine them
- Classroom training is only a preparation

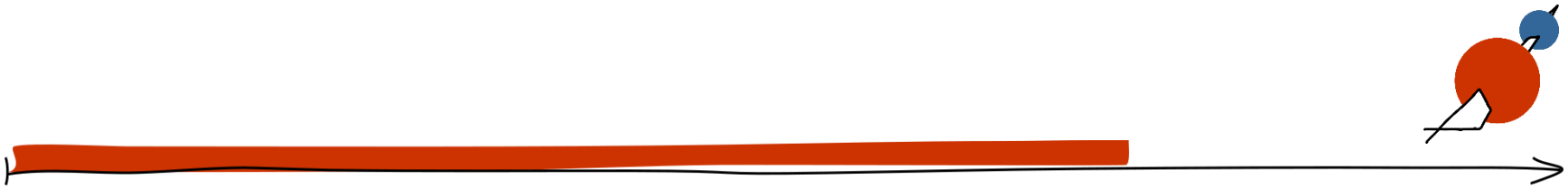


Establish knowledge transfer

- In real world projects diverse, unexpected problems will show up
  - That are relevant in the specific domain, your infrastructure and for communication
  - That will teach to solve such problems collaboratively in general
  - Responsibility should shift more and more towards in house experts over the time of a project
- During knowledge transfer real use cases are solved

Benefit from best practices

- Learn how to organize Data Science projects
  - Many people already were trapped in common pitfalls
  - By using their knowledge how to not do it, you avoid a lot of effort later
- Sticking to common standards will avoid incompatibilities later

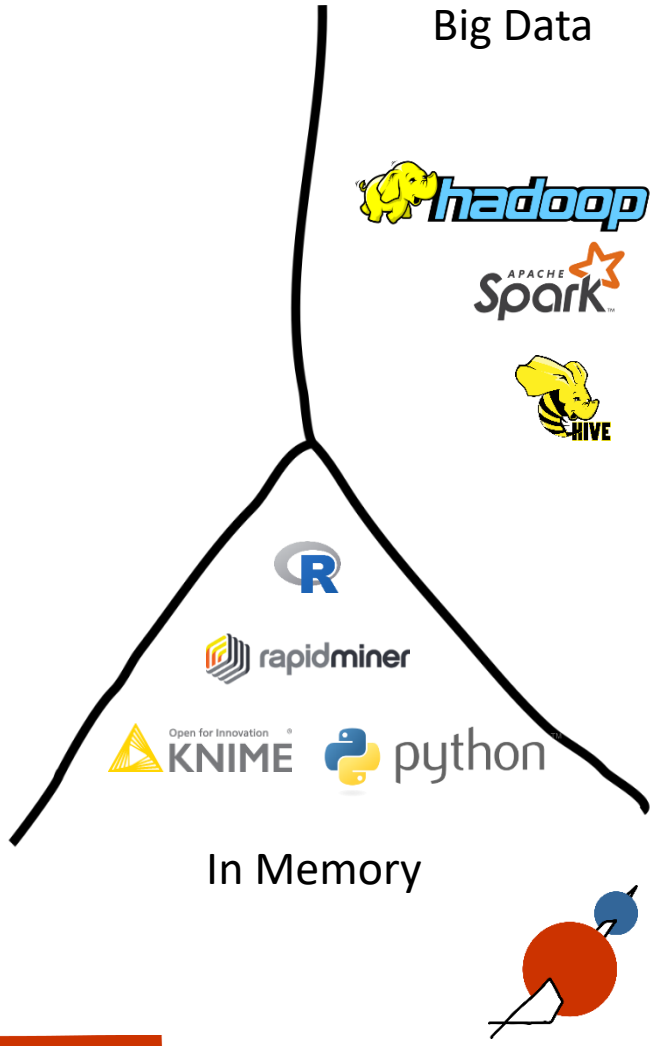


Methods and Infrastructure

- Knowledge about methods and infrastructure is something different
- Methods are generic, but need to be performed on a specific infrastructure
- Projects should start on simplest infrastructure possible
  - Creating least confusion while learning
  - Cheapest
  - Fastest to get access on

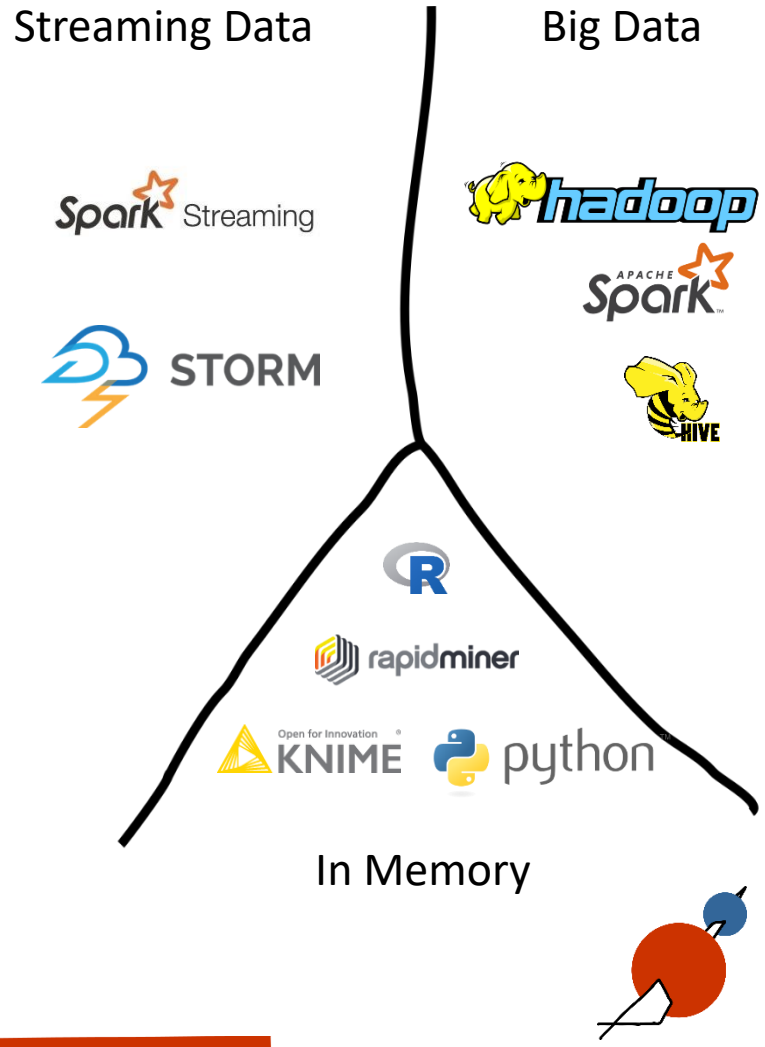
Lambda Stack

- In Memory
  - Cheap, easy, fast
  - Flexible methods but limited data
- Big Data
  - Expensive, complex, sluggish
  - Limited methods but unlimited data



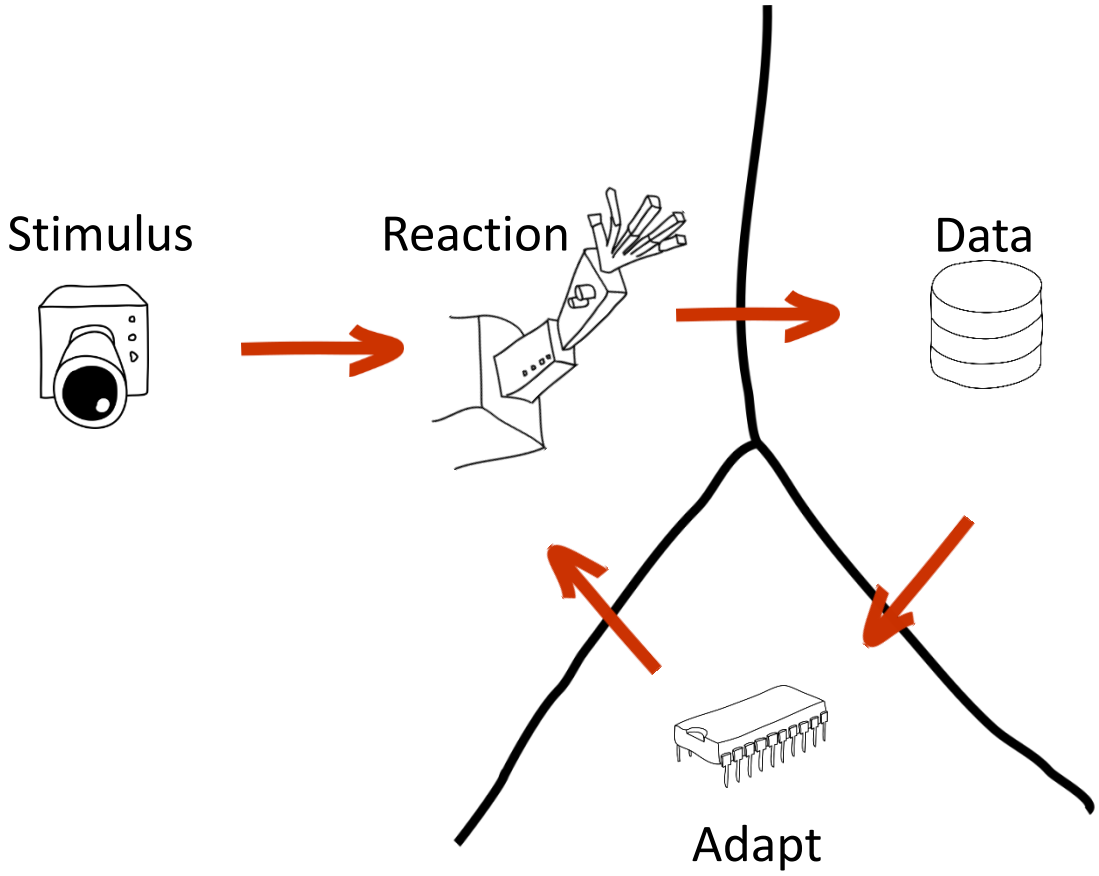
Lambda Stack

- Streaming Data
  - Cheap, hard, near-real time
  - Only deployment, very high throughput
- Necessary in cases when latency of decisions matter!



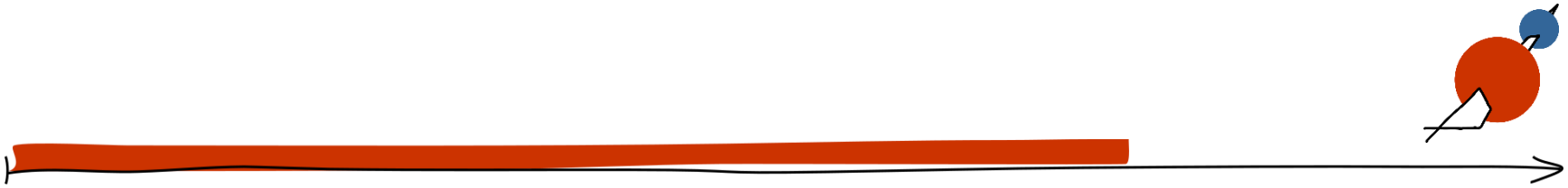
Lambda Stack

- Big Data and In Memory batch oriented
- Streaming event oriented
- Batch orientation necessary for adaption



Methods and Infrastructure

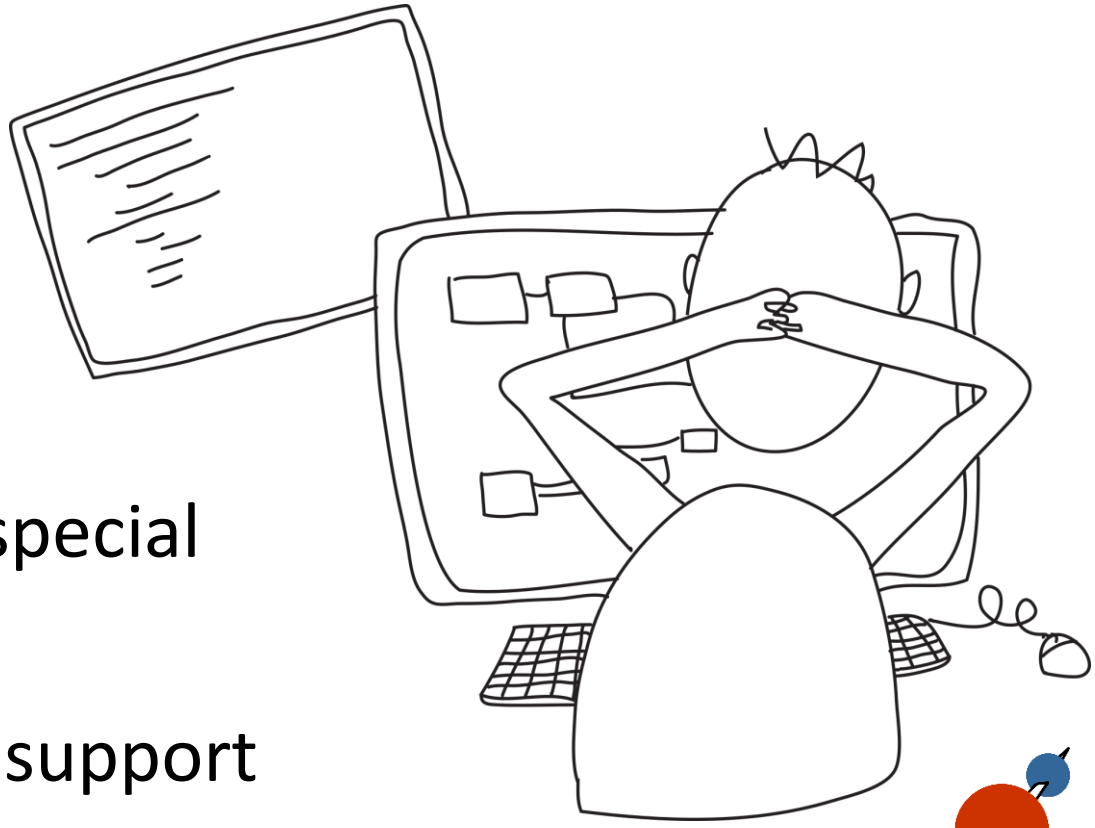
- Just start with In Memory infrastructure
  - No ramp up time
  - Less distracting infrastructural problems
  - Will be sufficient for most use cases
- Combine with big data later
  - For data storage
  - For preprocessing
  - And few problems that are really needing it



Phase 4: Long Term Evolution

Targets:

- Avoid breaking the tool chain
- Custom code for special situations
- Long term expert support



Avoid breaking the tool chain

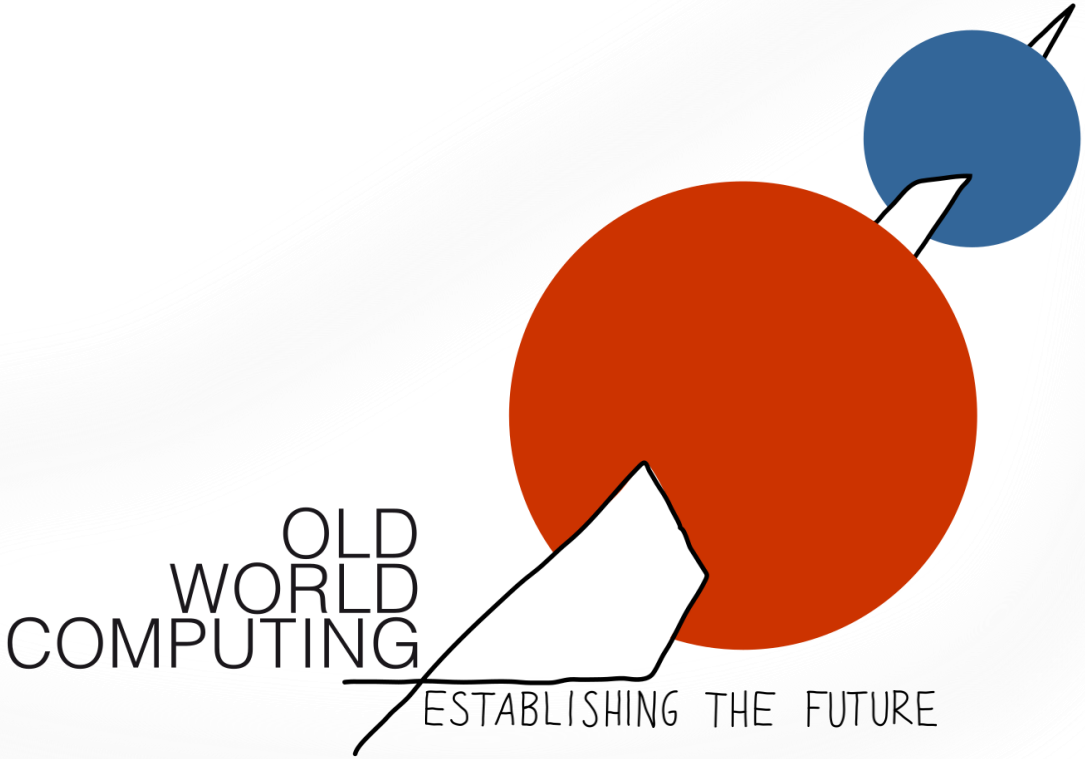
- Maintenance is crucial in long term projects and deployments
- Having to combine multiple tools exponentially grows the risk:
  - Each update in each tool may break chain
  - Dependencies of each tool need to be maintained
  - Expertise for each tool needs to be available
- Better combine in one platform

Custom Code

- Extend the data flow based platforms with custom code
- This ensures understandability in common language
- Reduces costs for maintenance
- Avoids additional training requirements for different tools

Long term expert support

- Even with years of experience one is faced with new, unknown problems
  - New infrastructure
  - Completely different data science problem
  - New technology
- Support from other experts can avoid spending a long time with experiments



contact@oldworldcomputing.com
www.oldworldcomputing.com