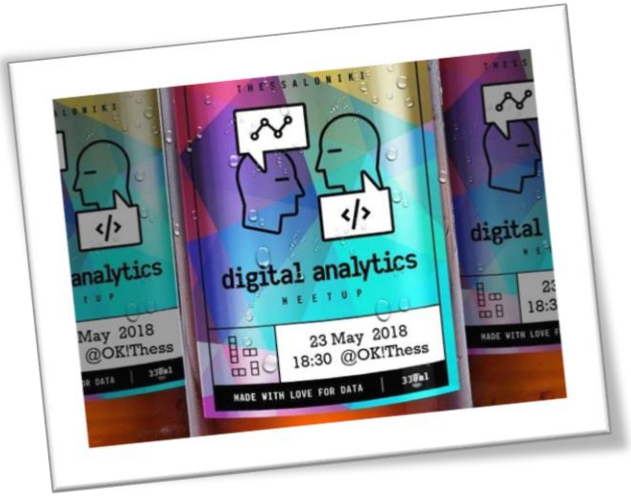
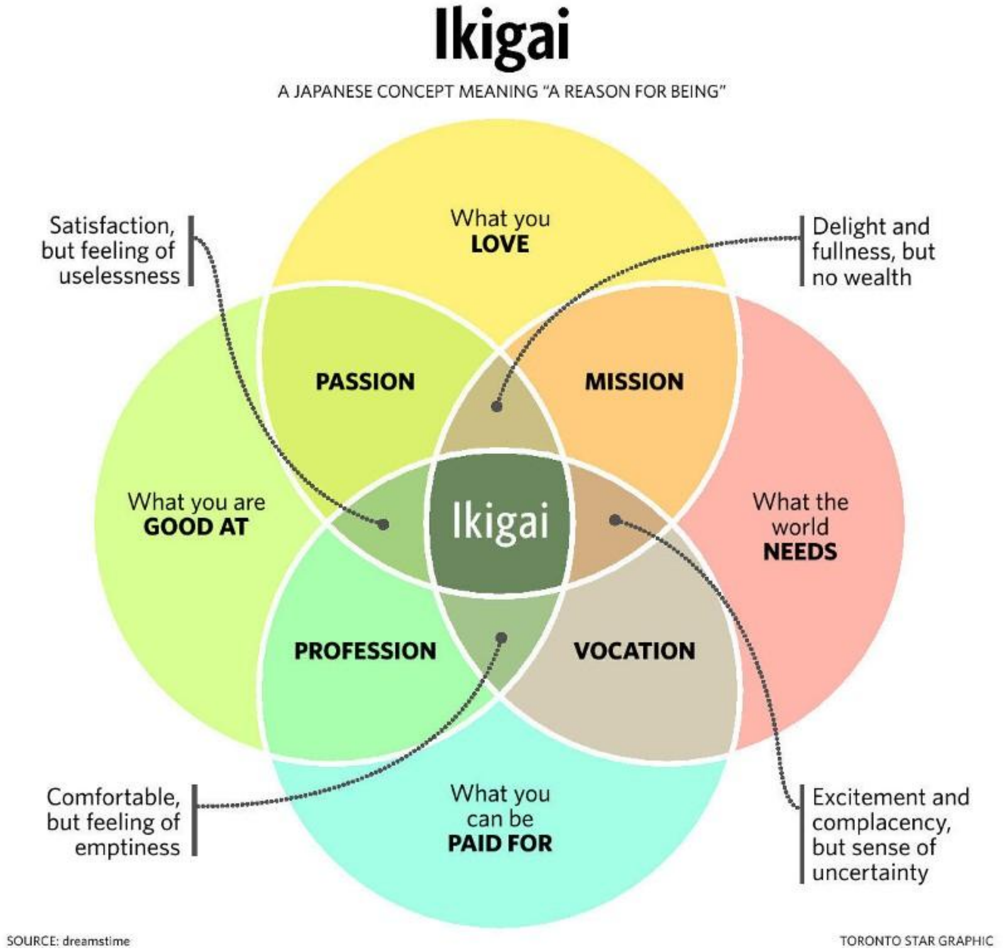
From Science to Data

Following a principled path to
Data Science

Dr. Anestis Fachantidis
Lead Data Scientist
Intelligent Systems Group
Dept. of Informatics - AUTH



A Data Scientist's Ikigai


Ikigai
A JAPANESE CONCEPT MEANING "A REASON FOR BEING"
What you LOVE
What you are GOOD AT
What the world NEEDS
What you can be PAID FOR
PASSION
MISSION
PROFESSION
VOCATION
Ikigai
Satisfaction, but feeling of uselessness
Delight and fullness, but no wealth
Comfortable, but feeling of emptiness
Excitement and complacency, but sense of uncertainty
SOURCE: dreamstime
TORONTO STAR GRAPHIC

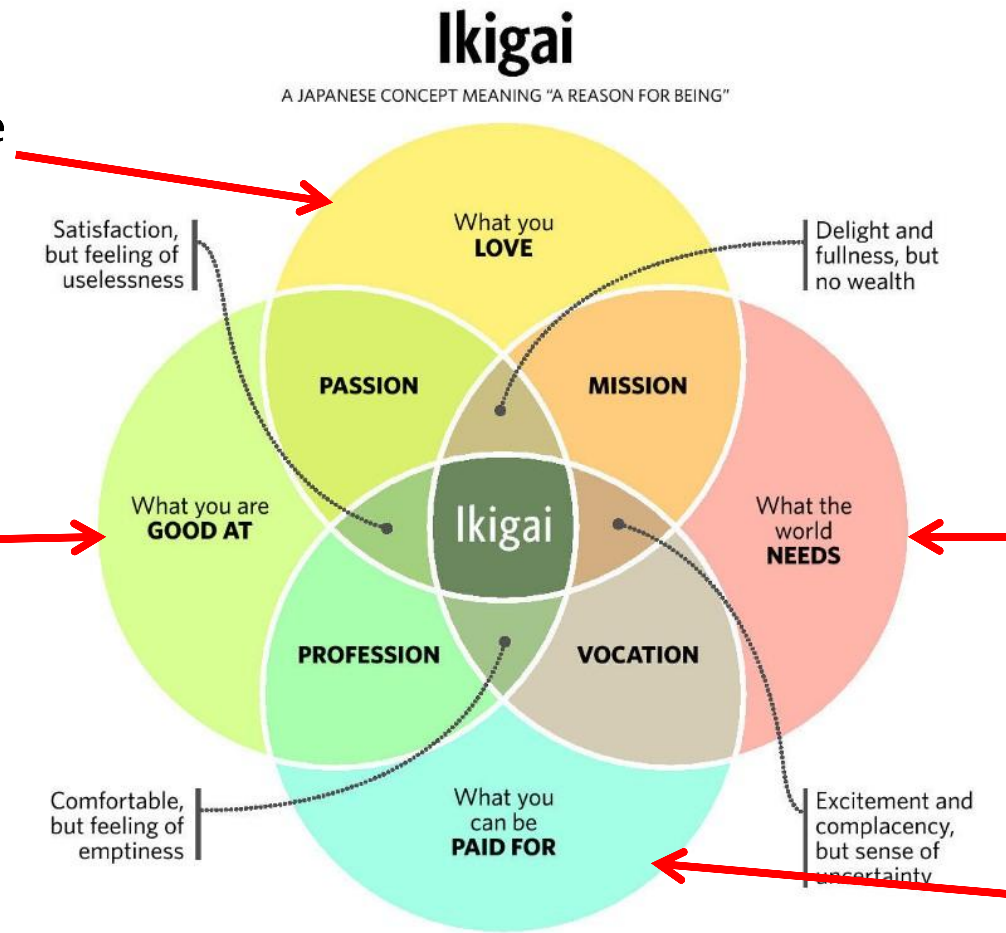
A Data Scientist's Ikigai

What to love in DS?

How to become good at

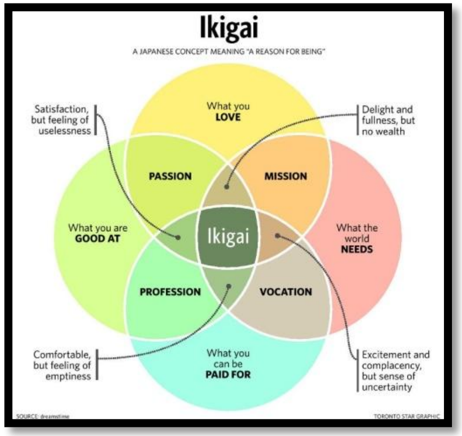
What kind of DS does the world need?

Which Industries pay for DS now?


Ikigai
A JAPANESE CONCEPT MEANING "A REASON FOR BEING"
What you LOVE
What you are GOOD AT
What the world NEEDS
What you can be PAID FOR
PASSION
MISSION
PROFESSION
VOCATION
Ikigai
Satisfaction, but feeling of uselessness
Delight and fullness, but no wealth
Comfortable, but feeling of emptiness
Excitement and complacency, but sense of uncertainty

SOURCE: dreamstime

TORONTO STAR GRAPHIC


Ikigai
A JAPANESE CONCEPT MEANING "A REASON FOR BEING"
What you LOVE
What you are GOOD AT
What the world NEEDS
What you can be PAID FOR
PASSION
MISSION
PROFESSION
VOCATION
Ikigai
Satisfaction, but feeling of uselessness
Delight and fullness, but no wealth
Comfortable, but feeling of emptiness
Excitement and complacency, but sense of uncertainty

SOURCE: dreamstime

TORONTO STAR GRAPHIC

Become good at some of the tools used in real projects

WHAT YOU ARE GOOD AT

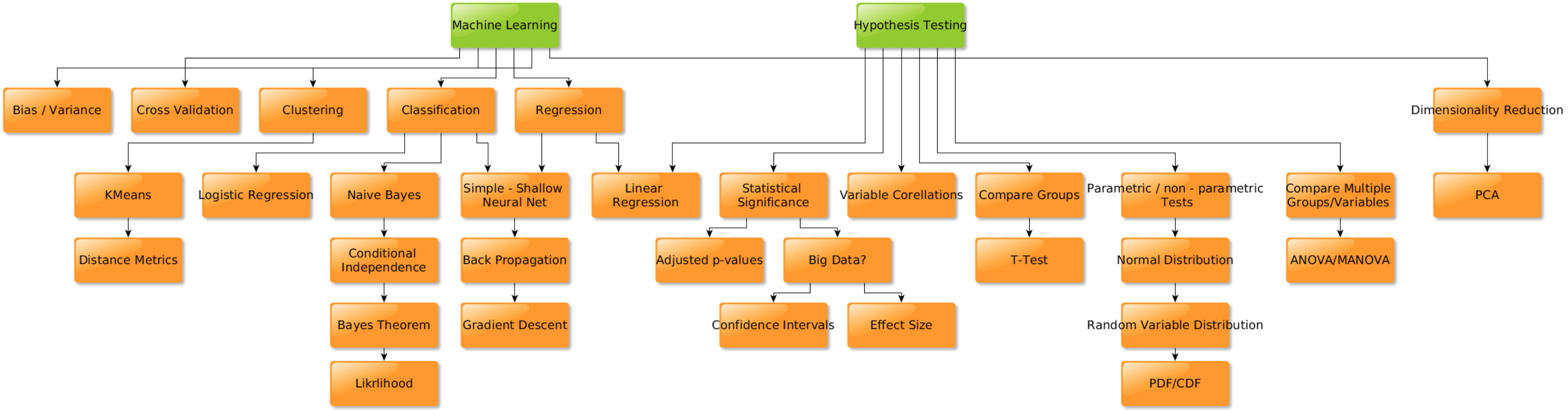
Become good at

- Next slides:

Theoretical Background	Web apps & services for DS
Accessing Data Sources	Cluster Computation
Data Preparation	Git & Documentation
Learning Models	Business Understanding
Data Visualization & Reporting	Team Work & Project Management

- Reasons that the theoretical background is **mandatory**
- No one will ever master all these skills ! This is a **maximal set**, we are just looking for your entry point as a beginner.
- This could be the **pair** of your weakest skill with the strongest one, one to make you a better DS and one to give you early gradification.
- The field is leaning more and more towards **specialization**.

Theoretical Background
Minimum set of keywords for beginners



Machine Learning
Hypothesis Testing
Bias / Variance
Cross Validation
Clustering
Classification
Regression
Dimensionality Reduction
KMeans
Logistic Regression
Naive Bayes
Simple - Shallow Neural Net
Linear Regression
Statistical Significance
Variable Corellations
Compare Groups
Parametric / non - parametric Tests
Compare Multiple Groups/Variables
PCA
Distance Metrics
Conditional Independence
Back Propagation
Adjusted p-values
Big Data?
T-Test
Normal Distribution
ANOVA/MANOVA
Bayes Theorem
Gradient Descent
Confidence Intervals
Effect Size
Random Variable Distribution
Likrlihood
PDF/CDF

**External .png file available*

Accessing Data

- Reading Data 101: read .csv and excel files (read.xlsx)
- Connect to a database
- Set up MariaDB or SQL server express locally and download a sample DB (e.g., Employers DB)
- Connect and read through R using frameworks like DBI and dbplyr
- You will still need basic knowledge of SQL!

Accessing Data

dbplyr Example:

```
con <- DBI::dbConnect(RSQLite::SQLite())
flights <- tbl(con,”flights”)
flights %>%
  select(distance, air_time) %>%
  mutate(speed = distance / (air_time / 60))%>%
  show_query()

#> <SQL>
#> SELECT `distance`, `air_time`, `distance` /
(`air_time` / 60.0) AS `speed`
#> FROM (SELECT `distance`, `air_time`
#> FROM `nycflights13::flights`)
```

Data Preparation

- Data preparation takes 60 to 80 percent of the whole analytical pipeline in a real DS project
- In R, frameworks like dplyr and data.table make data handling and processing easier
- Numerous packages and libraries for data cleaning
- Learning magrittr pipelines will make your code readable and will clear up the data manipulation process
- Understanding feature extraction, feature selection and their interconnection with the business context at hand.

Learning Models

- Frameworks like scikit-learn (Python) and caret (R) will make your first ML experimentation steps much easier.
- They provide a standardized interface to training, testing and hyper-parameter tuning.
- Try them on a Kaggle dataset!

GIT & Documentation

GIT, The most popular version control system today. Among other reasons you need that in DS too:

- Results are paired with {parameters, features selected, code} which comprise an (almost) deterministic state. Capture that state for all your results!
- Make a Bitbucket or GitHub account to:
  - Create a small DS "portfolio" of personal projects
  - Collaborate on open source projects
- Comment your code and always consider packaging for reusability
- Comment your objects:

```
comment(object) <- "…"
```

Don't name their files like:

```
Results23-5withoutsumofmoney2.rds
```

Data Visualization & Reporting

- Significant DS task on their own, the most powerful communication tools of a DS
- Learn at least one plotting “language”
- In R, the most well known plotting grammar is that of ggplot
- For interactive charts you can also use platforms like plot.ly and rCharts
- The fully reproducible paradigm of a compiled report: *“Code and text in one document side by side”*
- Learn tools like rMarkdown (R) and Jupiter (Python) which use an easy markdown syntax

Data Visualization & Reporting

Web Apps & Services for DS

Analytics APIs and ML web services:

- For small teams a Platform-as-a-service solution like Heroku makes it easy to deploy a data product
- Basic understanding of the HTTP protocol and data formats such as JSON

Data Products as Web Apps:

- Web app frameworks like Shiny (R) or Django (Python) can help deliver analytics or ML results having limited knowledge of web development

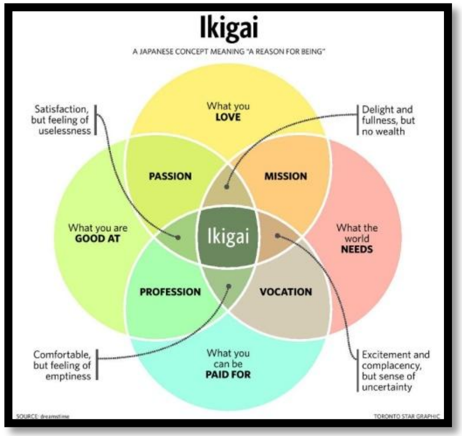
Web Apps & Services for DS

Cluster Computation

- SparkR and Sparklyr will let you easily setup a distributed computation cluster in R
- Access to MLlib library
- Package H2O for access to H20 open source engine for analytics and ML
- Set them up locally and try them on sample data!

Business Understanding

- More general: Domain understanding
- Requirements Analysis = following a structured analytical process + communication skills
- Beyond understanding the meaning of each business variable, understanding:
  - If the variable is directly controlled or not
  - How does this influences other variables


Ikigai
A JAPANESE CONCEPT MEANING "A REASON FOR BEING"
What you LOVE
What you are GOOD AT
What the world NEEDS
What you can be PAID FOR
PASSION
MISSION
PROFESSION
VOCATION
Ikigai
Satisfaction, but feeling of uselessness
Delight and fullness, but no wealth
Comfortable, but feeling of emptiness
Excitement and complacency, but sense of uncertainty
SOURCE: dreamstime
TORONTO STAR GRAPHIC

It should not be only about summing up money

WHAT THE WORLD NEEDS

Insights

- Open Insights
  - People need to make sense of data
  - Think of any NGO organization you love its purpose or even the shop around the corner you just love its products:
    - Wouldn't you want to give them relevant insights about their business environment?
    - Insights that will make them act accordingly and become sustainable
  - Data Scientists also have the responsibility to educate people on the **interpretation of results** and on how they could **identify bad data journalism**

…Openness…

Open Source

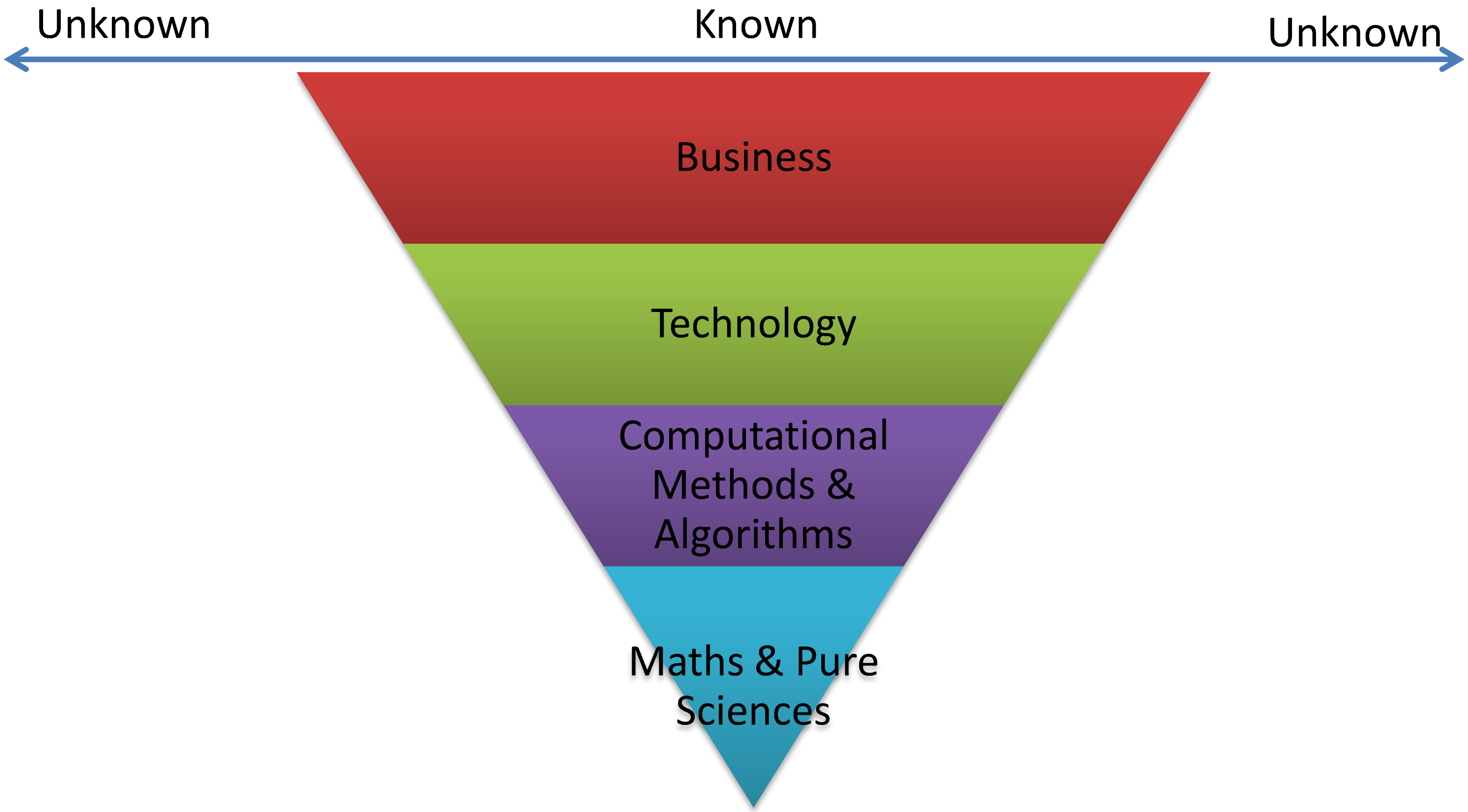
- The open source communities need support and people that care about their projects too
  - Don’t just use them, think of ways to contribute
- Write your own R package or Python library
  - Share it on a Git web platform and it might be the next big thing in open source!

Open Data, which are of critical value for the following reasons:

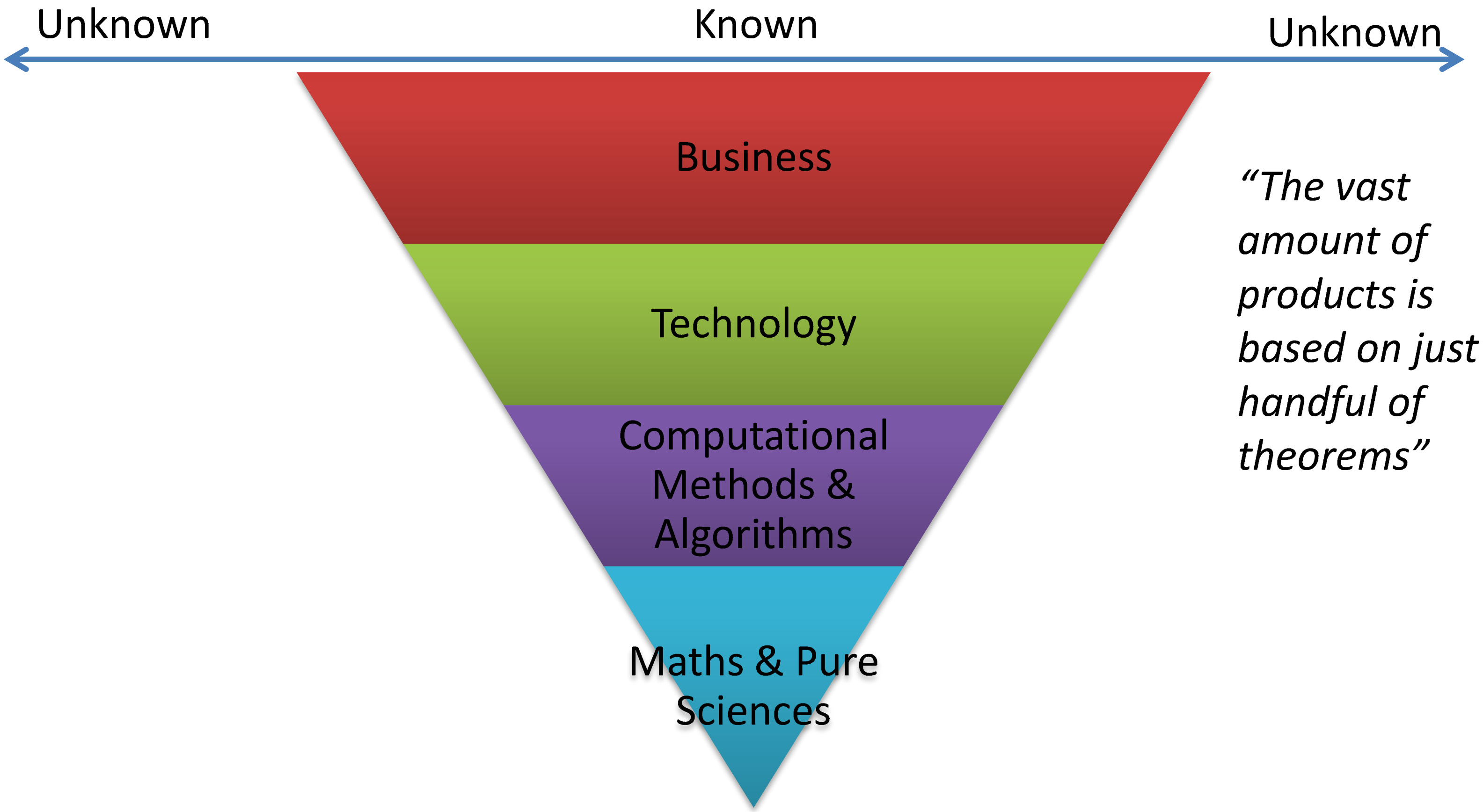
  - Transparency and democratic control
  - Improved or new private products and services
  - Improved efficiency and effectiveness of government services
  - New knowledge from combined data sources and patterns in large data volumes

…which all need a Data Scientist!

Innovation

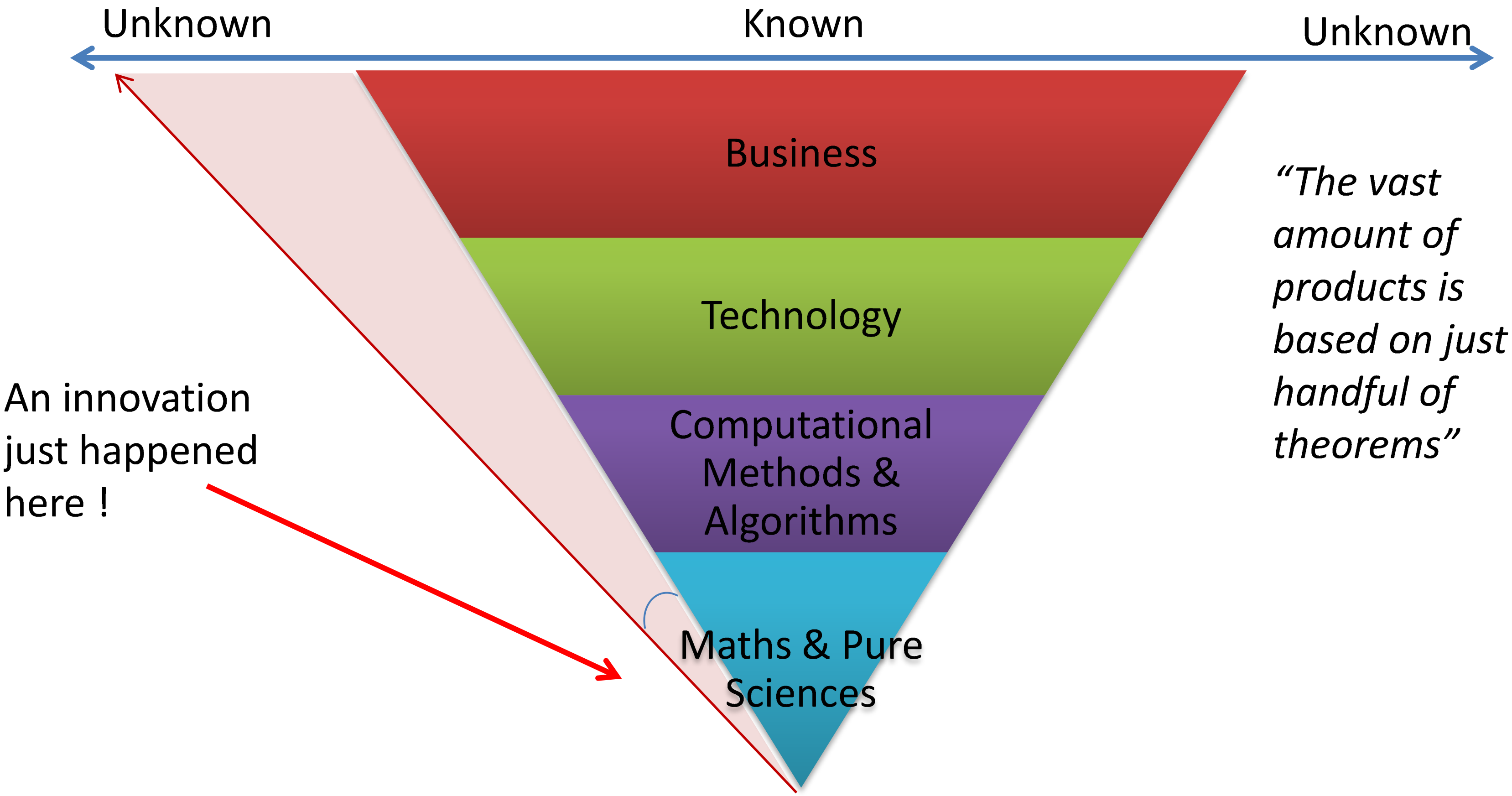

Unknown
Known
Unknown
Business
Technology
Computational Methods & Algorithms
Maths & Pure Sciences

The reverse pyramid of Technology Innovation



“The vast amount of products is based on just a handful of theorems”

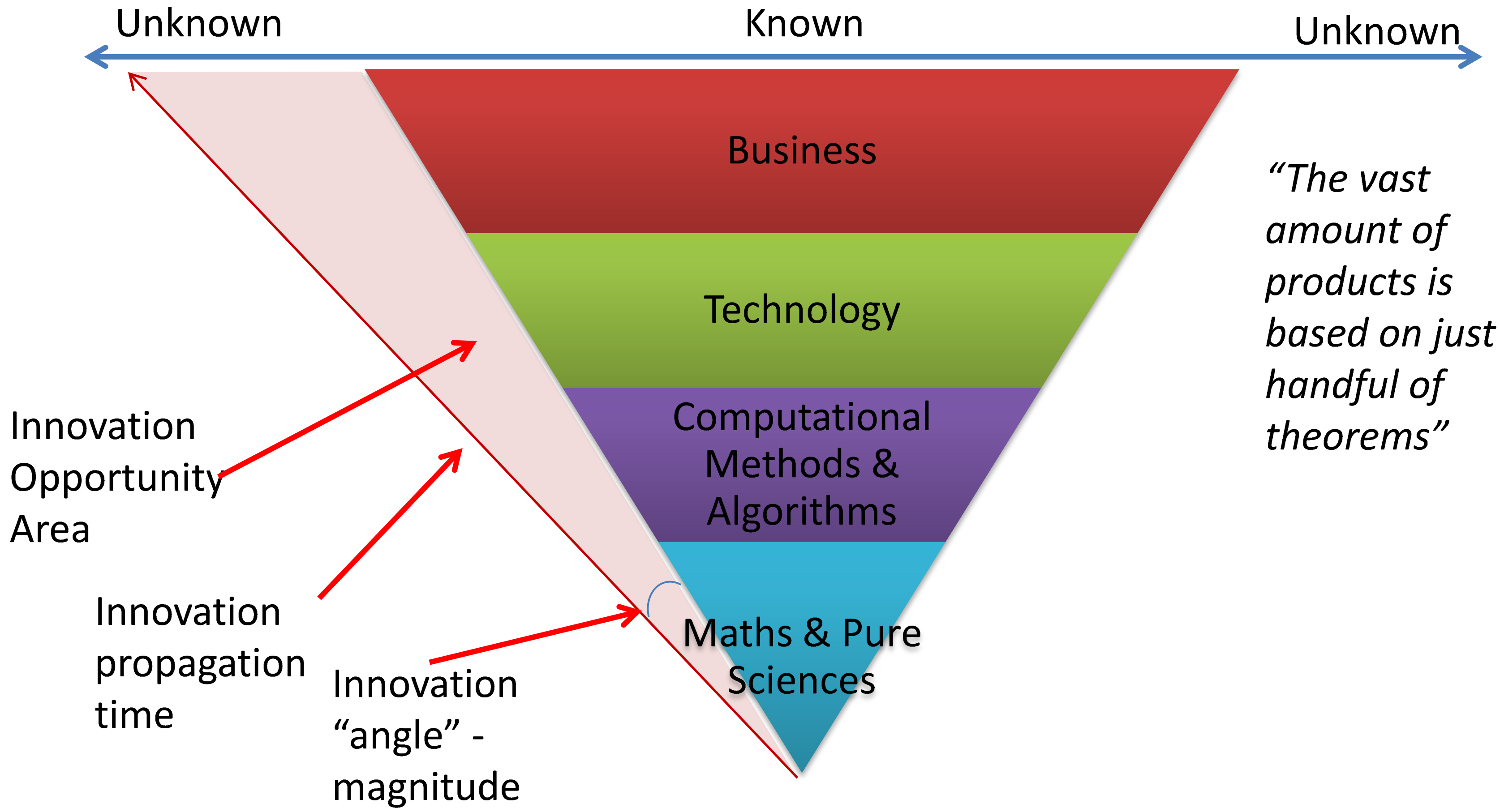
The reverse pyramid of Technology Innovation



An innovation just happened here !

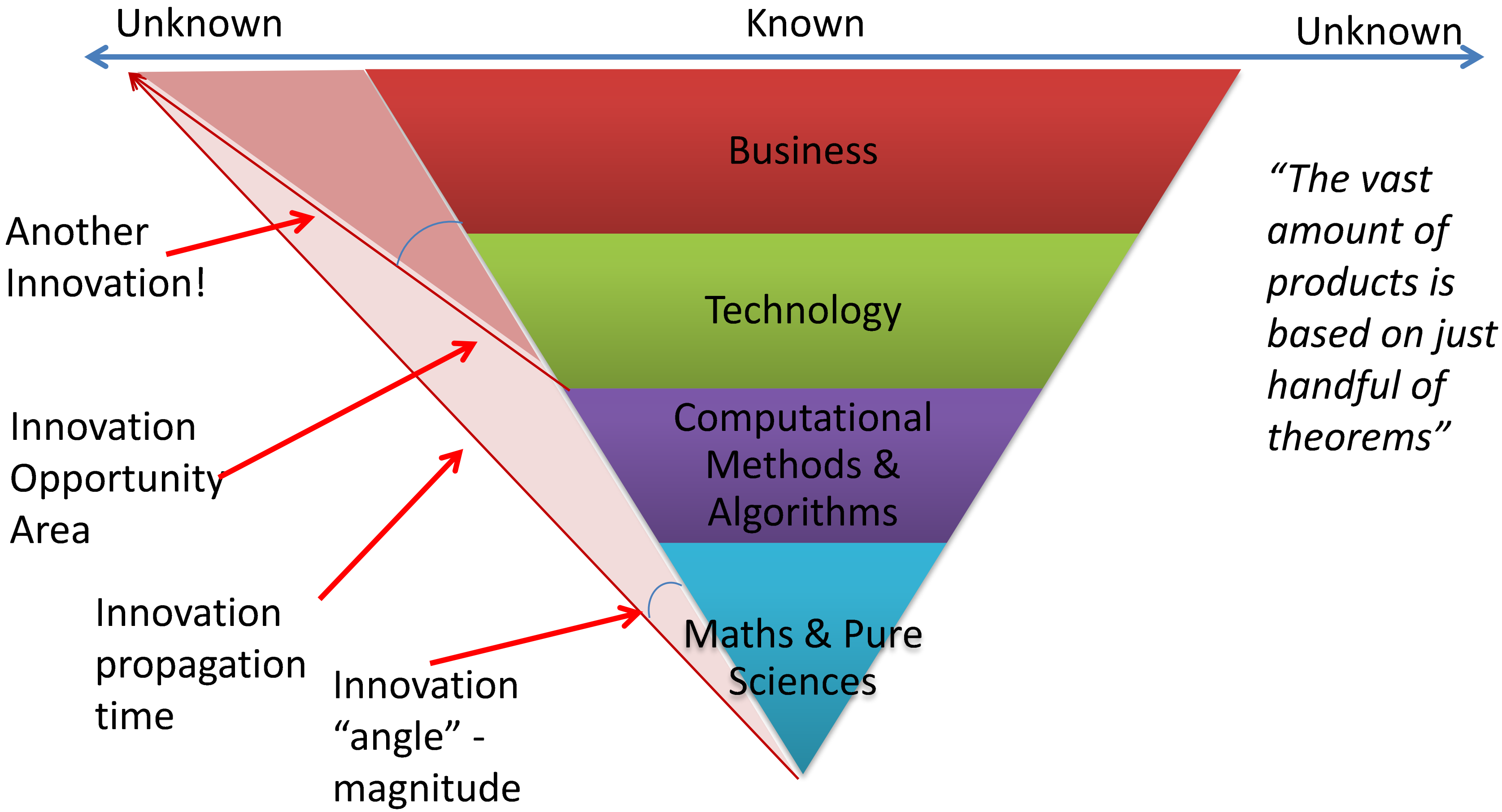
“The vast amount of products is based on just a handful of theorems”

The reverse pyramid of Technology Innovation


Unknown
Known
Unknown
Business
Technology
Computational Methods & Algorithms
Maths & Pure Sciences
Innovation Opportunity Area
Innovation propagation time
Innovation “angle” - magnitude

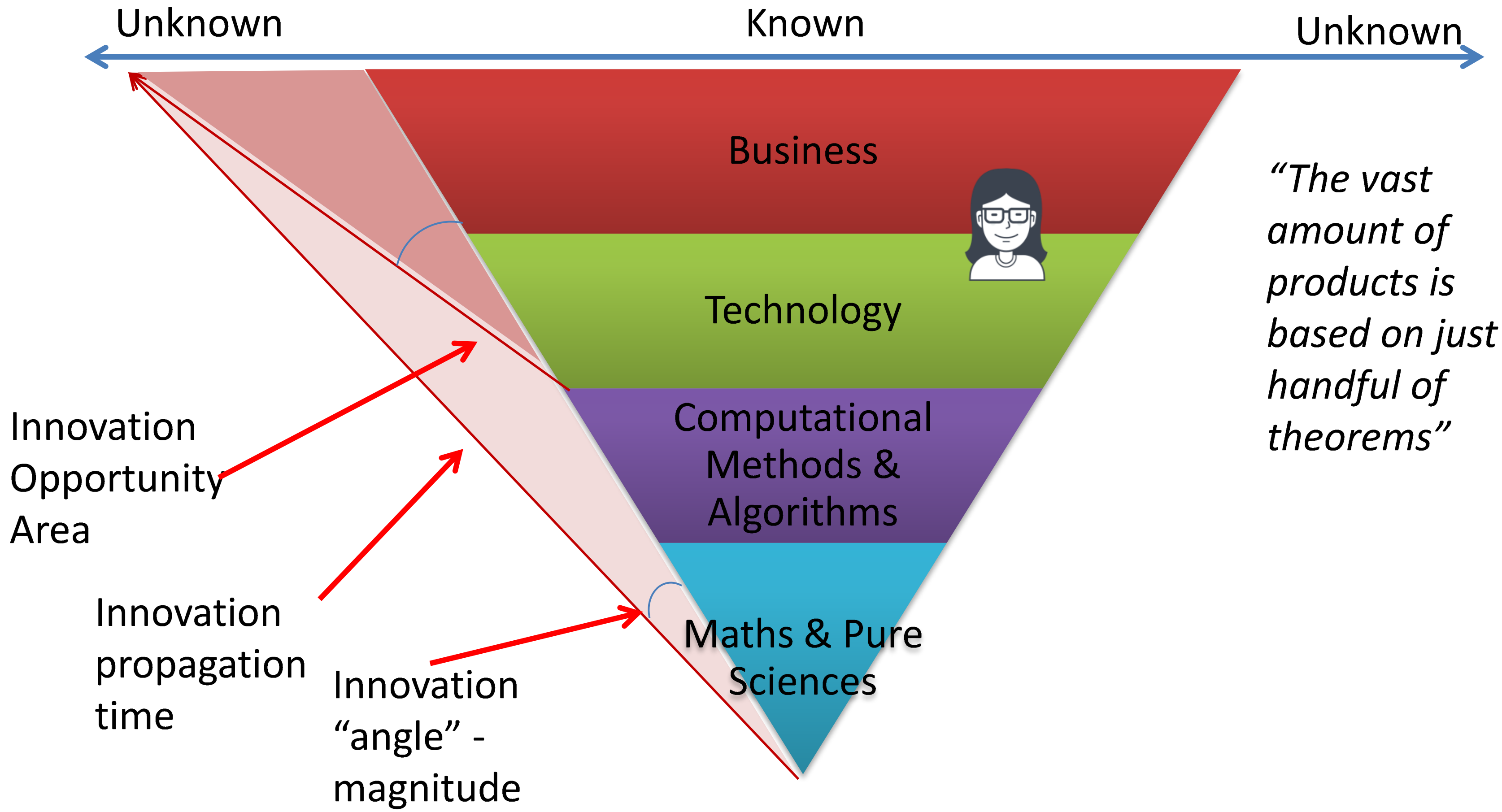
“The vast amount of products is based on just a handful of theorems”

The reverse pyramid of Technology Innovation


Unknown
Known
Unknown
Business
Technology
Computational Methods & Algorithms
Maths & Pure Sciences
Another Innovation!
Innovation Opportunity Area
Innovation propagation time
Innovation "angle" - magnitude

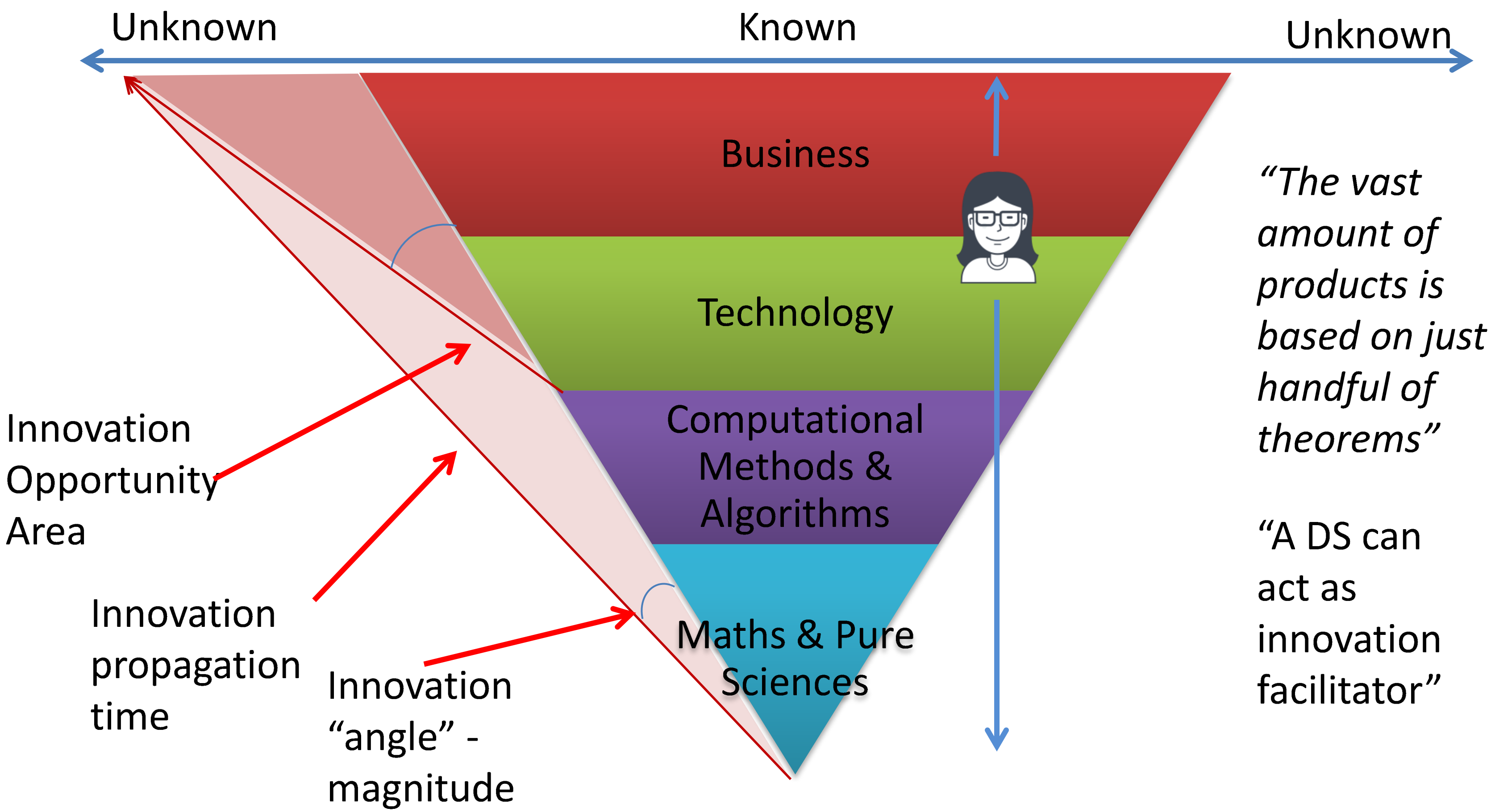
"The vast amount of products is based on just a handful of theorems"

The reverse pyramid of Technology Innovation


Unknown
Known
Unknown
Business
Technology
Computational Methods & Algorithms
Maths & Pure Sciences
Innovation Opportunity Area
Innovation propagation time
Innovation “angle” - magnitude

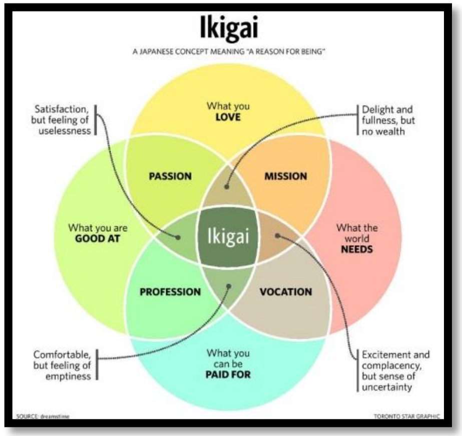
“The vast amount of products is based on just a handful of theorems”

The reverse pyramid of Technology Innovation


Unknown
Known
Unknown
Business
Technology
Computational Methods & Algorithms
Maths & Pure Sciences
Innovation Opportunity Area
Innovation propagation time
Innovation "angle" - magnitude

"The vast amount of products is based on just a handful of theorems"

"A DS can act as innovation facilitator"

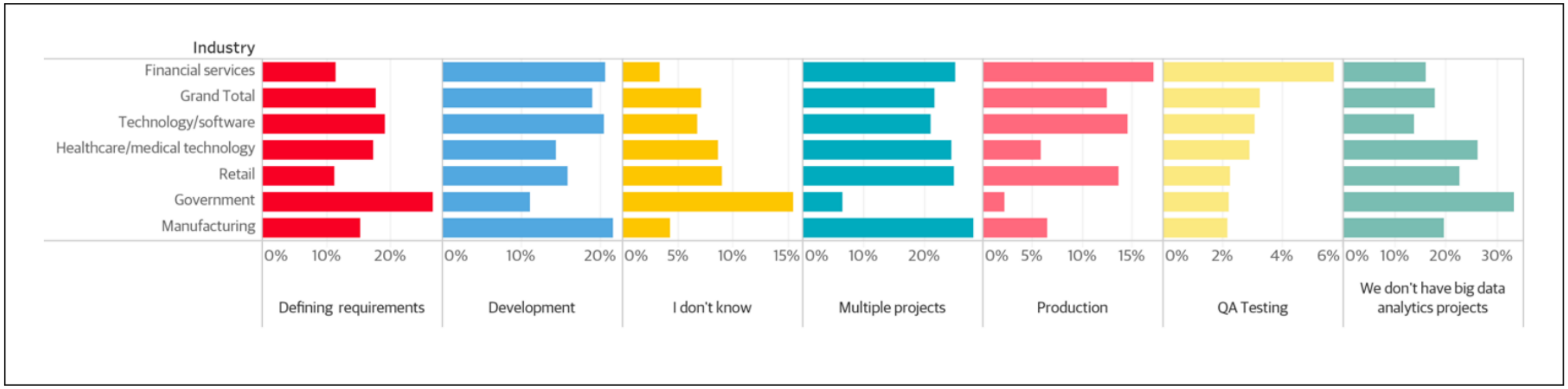

Ikigai
A JAPANESE CONCEPT MEANING "A REASON FOR BEING"
What you LOVE
What you are GOOD AT
What the world NEEDS
What you can be PAID FOR
PASSION
MISSION
PROFESSION
VOCATION
Ikigai
Satisfaction, but feeling of uselessness
Delight and fullness, but no wealth
Comfortable, but feeling of emptiness
Excitement and complacency, but sense of uncertainty
SOURCE: dreamstime
TORONTO STAR GRAPHIC

Get hired in industries that have already adopted DS and ML

WHAT YOU CAN BE PAID FOR

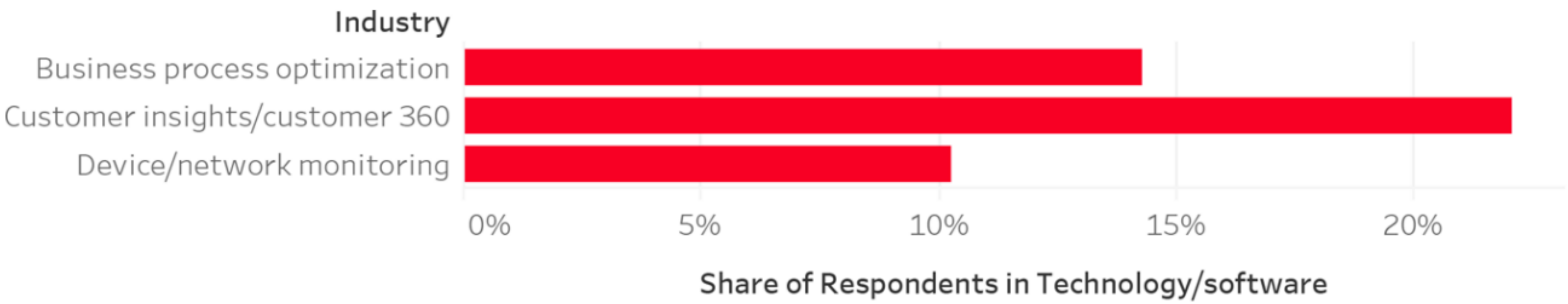
Industries’ adoption of Data Science

Stage of data analytics projects by specific industry:



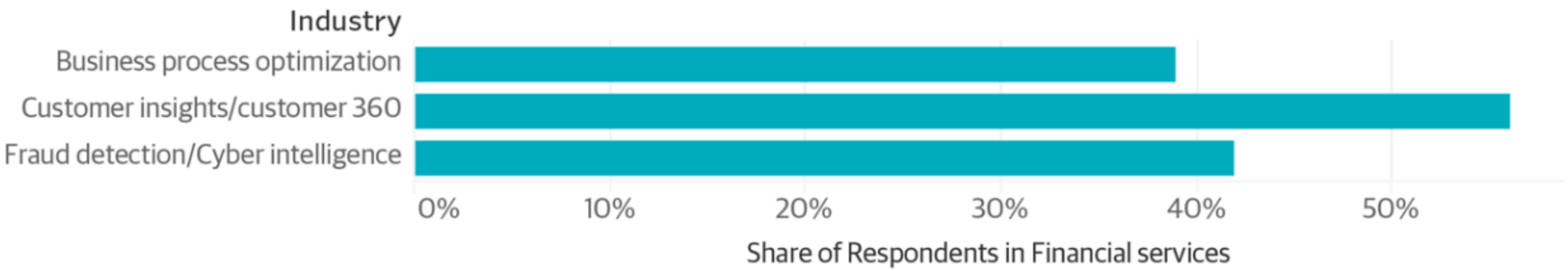
source: O’Reilly Media - spring 2017 - 875 respondents

Technology/software

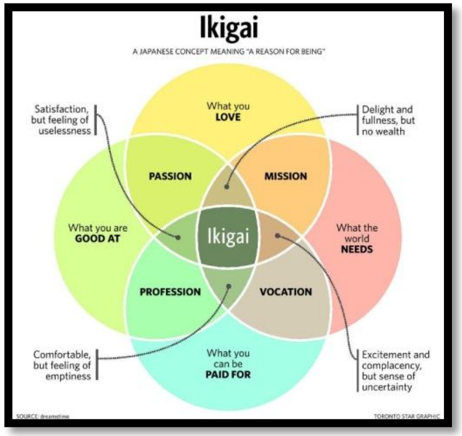


source: O'Reilly Media - spring 2017 - 875 respondents

Financial Services



source: O'Reilly Media - spring 2017 - 875 respondents


Ikigai
A JAPANESE CONCEPT MEANING "A REASON FOR BEING"
What you LOVE
What you are GOOD AT
What the world NEEDS
What you can be PAID FOR
PASSION
MISSION
PROFESSION
VOCATION
Ikigai
Satisfaction, but feeling of uselessness
Delight and fullness, but no wealth
Comfortable, but feeling of emptiness
Excitement and complacency, but sense of uncertainty

SOURCE: dreamstime

TORONTO STAR GRAPHIC

A reason to wake up happily in the morning

WHAT YOU LOVE

“Love at first result”

You will probably love DS when:

- You produce your first *insight* on *actual business or real-life* data
- You see a learning process reducing its error on test data
- Your result dazzled a business stakeholder and…

  …actually made him/her take a decision…

  …and a successful one, as measured later on based on some KPI.

- Check for the **new MSc program** on Data Science in the School of Informatics - AUTH (to be announced)
- **We are hiring Data Scientists!** Contact Intelligent Systems Group and you may start working on new innovative projects!
- Contact Details and Talk Notes in:

  http://bit.ly/science2data

Thank You!

- Check for the **new MSc program** on Data Science in the school of Informatics - AUTH (to be announced)
- **We are hiring Data Scientists!** Contact Intelligent Systems Group and you may start working on new innovative projects!
- Contact Details and Talk Notes in:

  http://bit.ly/science2data