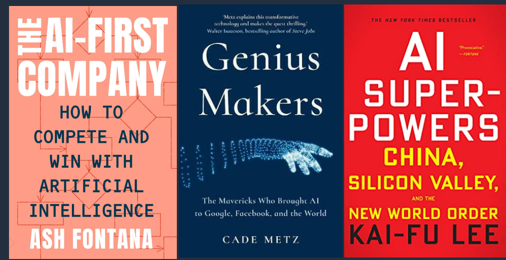
MAI 622: AI Entrepreneurship

This document has been produced with the support of THE EUROPEAN COMMISSION under THE CONNECTING EUROPE FACILITY - TELECOMMUNICATIONS SECTOR AGREEMENT No INEA/CEF/ICT/A2020/2267423. It reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Module 3: AI Companies



Learning Objectives

After attending this module, studying the suggested readings and case studies you should be able to:

- Understand the opportunities arising because of the advances in AI technology.
- Understand and explain the particular characteristics and challenges of AI companies.
- Recognize and describe some key concepts, terminology you need to know and questions you need to ask, to be able to take advantage of machine learning and AI for the benefit of your business.

AI Companies

Section 1: Introduction

Learning Objectives

After attending this section, studying the suggested readings and case studies you should be able to:

- Understand the opportunities arising because of the advances in AI technology.
- Understand and explain the particular characteristics and challenges of AI companies.
- Recognize and describe some key concepts, terminology you need to know and questions you need to ask, to be able to take advantage of machine learning and AI for the benefit of your business.

WHAT IS AN AI-FIRST COMPANY?

IT IS A COMPANY THAT GENERATES INCOME FROM DATA

NAME A COMPANY THAT IS PUTTING AI AT THE CENTER OF ITS BUSINESS

Google
amazon
Microsoft
Tencent
Alibaba.com
Bai du 百度

?

Microsoft + OpenAI

Google + DeepMind

TESLA + xAI

Meta + Meta AI

amazon + ANTHROPIC

www.therundown.ai

Introduction to AI-First Companies

- AI-first companies prioritize AI: These organizations focus on **integrating AI at the core of their business** strategies, ensuring that all products and services are designed with AI capabilities from the outset.
- Data as a core competitive advantage: In these companies, data is not just an asset but the **foundation of their competitive edge**, driving innovation and operational efficiency.
- Leveraging data for innovation: By **harnessing the power of vast data sets**, AI-first companies can innovate more rapidly than their competitors, leading to enhanced products and services.

AI Transformation beyond the Big Platforms

- Leaders of legacy organizations in other industries feel that it's beyond their companies' capabilities to transform themselves using AI.
- At many organizations AI initiatives are too small and too tentative:
  - 7 out of 10 companies reported that their AI efforts had had minimal or no impact (MIT Sloan Management Review and Boston Consulting Group, 2019).
  - Among the 90% of companies that had made some investment in AI, fewer than 40% had achieved business gains over the previous three years (MIT Sloan &BCG, 2019).
  - AI initiatives at many organizations are too small and too tentative.
- Most organizations never get to the only step that can add economic value:

Deploying a model on a large scale.

- Testing the waters may deliver valuable insights, but it's not enough to achieve true transformation.

Stop Tinkering with AI!

Study on 30 companies that have gone all in on AI—and achieved success — identified **10 actions** those companies took to become successful AI adopters:

1. Know what you want to accomplish.
2. Work with an ecosystem of partners.
3. Master analytics.
4. Create a modular, flexible IT architecture.
5. Integrate AI into existing workflows.
6. Build solutions across the organization.
7. Create an AI governance and leadership structure.
8. Develop and staff centers of excellence.
9. Invest continually.
10. Always seek new sources of data.

Know What You Want to Accomplish

- All companies want to apply AI to be more financially successful.
- However: identifying and developing transformational AI requires a clearer objective:
  - Improve process speed;
  - Reduce operating costs;
  - Become better marketers.
- Whatever the goals are: identifying one well-defined, overarching objective and making it a guiding principle for your adoption.

The case of Deloitte's Omnia

OmniaAI

- Omnia is Deloitte's proprietary AI platform for auditing and assurance.
- Omnia's guiding principle: improve service quality **globally**.
- Remarks:
  - Important differences exist in how countries regulate data, including standards for privacy, audit processes, and risk management.
  - Different companies use different data structures to store financial and operational data.
- The goal of making Omnia a global tool created several unique challenges including developing a single data model that would work across clients and regions.
- Envisioning Omnia as a global tool before it had been created allowed Deloitte's developers to focus on standardizing information from different companies in different countries—a huge undertaking that would have been even more challenging later in the development process.

Work with an Ecosystem of Partners

- Building Omnia required Deloitte to monitor technology start-ups around the world to find solutions that fit its audit and assurance practice's needs.
- Developing technologies in-house might have been possible, but at a much higher cost and on a much slower timeline.
- A company needs strong partnerships to succeed with AI.
- Deloitte worked with a number of start-ups:
  - Kira Systems: expertise in NLP software that extracts contract terms from legal documents.
  - Signal AI: built a platform that analyzes publicly available financial data to identify potential risk factors in a client's business.
  - Chatterbox Labs: helped create Trustworthy AI, a module which evaluates AI models for bias.

https://hbr.org/2023/01/stop-tinkering-with-ai

Master Analytics

- Most successful AI adopters had significant analytics initiatives underway before they moved headlong into AI/ML.
- Mastering analytics requires a commitment to using data and analytics for most decisions; this dictates:
  - changing the way you deal with customers by embedding AI in products and services
  - conducting many tasks—even entire business processes—in a more automated and intelligent fashion
  - increasingly have unique or proprietary data.

https://hbr.org/2023/01/stop-tinkering-with-ai

Data is the foundation of ML success: models can't make accurate predictions without large quantities of good data.

The single biggest obstacle for most organizations in scaling up AI systems is acquiring, cleaning, and integrating the right data.

It's also important to actively pursue new sources of data for new AI initiatives.

Seagate Case Study

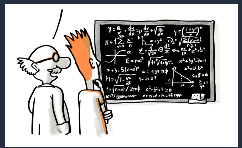
- Seagate Technology, has tremendous amounts of sensor data in its factories and has been using it extensively to improve the quality and efficiency of its manufacturing processes.
- Focus: automating the visual inspection of silicon wafers, from which disk-drive heads are made, and the tools that manufacture them.
- Multiple microscope images taken from various tool sets throughout wafer fabrication.
- Using data provided by the images, Seagate's Minnesota factory created an automated system that allows machines to find and classify wafer defects directly.
- Other image-classification models detect out-of-focus electron microscopes in the monitoring tools to determine whether defects actually exist.
- Since these models were first deployed, in late 2017, their use has grown extensively across the company's wafer factories in the United States and Northern Ireland, saving millions of dollars in inspection labor costs and scrap prevention.
- Visual inspection accuracy, at 50% several years ago, now (2023) exceeds 90%.

Create a Modular, Flexible IT Architecture

- To deploy AI solutions, you'll need a way to easily deploy data, analytics, and automation across your enterprise applications.
- That requires a technology infrastructure that can communicate and understand data from other IT environments, both inside and outside your company.
- A flexible IT architecture makes it easier to automate complex processes.
- Developing and maintaining **in house** such an architecture that can offer instantly data storage and computing power, which is software-driven and massively scalable, can be very **expensive** and **difficult**
- Migrating data and applications to the cloud, can help companies become aggressive AI adopters, that focus on developing software and business capabilities.

Lecture 7/3/2024

Stop Tinkering with AI



Companies need to take **10 actions** to become successful AI adopters:

1. Know what you want to accomplish.
2. Work with an ecosystem of partners.
3. Master analytics.
4. Create a modular, flexible IT architecture.
5. Integrate AI into existing workflows.
6. Build solutions across the organization.
7. Create an AI governance and leadership structure.
8. Develop and staff centers of excellence.
9. Invest continually.
10. Always seek new sources of data.

https://hbr.org/2023/01/stop-tinkering-with-ai

Integrate AI into Existing Workflows

- Determine which of your workflows are ripe for AI speed and intelligence and begin integrating AI into them as soon as possible.
  - Avoid trying to cram AI into workflows that wouldn't benefit from machine speed and scale, such as seldom-used business processes that neither involve nor generate enormous amounts of data and repetition.
- Workflow integration requires a very specific plan of attack:
  - If you have determined that you want to improve customer service, you need acute on-the-ground knowledge of those processes that few executives have.
  - Line employees have an ideal perspective for determining which processes can benefit from AI and how the processes can be specifically improved.

Build Solutions Across the Organization

- Once having **tested internally** and **mastered AI** across **a specific workflow**, an organization needs to become more aggressive in deploying it throughout the organization.
- Rather than designing one algorithmic model for one process, your goal should be to find a unified approach that can be replicated across the company.
- Example: Cleveland Clinic
  - The clinic faces a huge challenge involving data and analytics, as hospitals have much less data than organizations in other industries, and it is less likely to be clean and well structured.
  - Hospital data have quality issues, are captured poorly, are entered in different ways, and involve different definitions across the institution.
  - Knowledge of each practice's data structures is required to interpret the data accurately: Rather than leave data preparation to each practice within the clinic for each individual data set, they make it a part of every AI project and work to provide useful data sets to all AI projects.

Create an AI Governance and Leadership Structure

- Putting someone in charge of determining how AI is deployed throughout the organization makes transformation easier.
- The best leaders are aware of what:
  - AI can do in general
  - AI can do for their companies
  - implications AI might have for strategies, business models, processes, and people.
- But the greatest challenge leaders face is creating a culture that:
  - emphasizes data-driven decisions and actions and
  - makes employees enthusiastic about AI's potential to improve the business.
- In the absence of that culture, even if a few AI advocates are scattered around the organization
  - they won't get the resources they need to build great applications
  - they won't be able to hire great people
  - And if AI applications are built, the business won't make effective use of them.

WHAT KIND OF LEADER CAN FOSTER THE RIGHT CULTURE?

Leader's profile

- It helps to have a CEO or another C-level executive who is familiar with information technology leading the initiative:

1. Someone with no technical knowledge can lead AI efforts at your company, but that person would have to learn a lot, and quickly.
2. It's important that the leader work on multiple fronts: participation by a senior executive is particularly important to signaling interest in the technology, establishing a culture of data-driven decisions, prompting innovation across the business, and motivating employees to adopt new skills.
3. Leaders hold the power of the purse. Exploring, developing, and deploying AI is **expensive**. Leaders must invest—or persuade others to invest—enough to enable all levels of adoption.

- Having a single AI leader helps, but ultimately **commitment to this work must go deep into the organization**.
- If upper, middle, and even frontline managers are only paying lip service to the idea of transforming with AI, things will move slowly, and the organization will most likely revert to old habits.

Develop and Staff Centers of Excellence

- Decision-makers from all business units should ensure that AI projects get sufficient funding and time, and they should also implement AI in their own work.
- It's important to educate them on how AI functions, when it's appropriate, and what a major commitment to it involves.
- For the great majority of companies it's still early days for this **upskilling** and **reskilling** work, and not every employee needs to be trained in AI.
- But some clearly do, and probably the more the better.
- To be successful, a company needs considerable talent and training in AI, data engineering, and data science.

DBS Bank Case Study

- When Piyush Gupta joined DBS Bank as CEO, in 2009, it was Singapore's lowest-rated bank for customer service.
- Gupta has invested heavily in AI experimentation—about $300 million a year over the past few years—and has given business units and functions the flexibility to hire data scientists to see what they can accomplish.
- The bank's head of HR, who had no technical background, created a **small working group** to identify and pilot AI applications, including JIM—the **Job Intelligence Maestro**—a model that predicts personnel attrition and helps the bank recruit the most-qualified employees. DBS used it to hire many of the 1,000 data scientists and data engineers who work at the organization today.
- DBS now has twice as many engineers as bankers: they work on **emerging technologies** such as blockchain and asset-backed tokens as well as on AI projects. The bank's **culture has greatly improved**: Euromoney named DBS the world's best bank for each of the four years from 2018 to 2021, and its capital positions and credit ratings are now among the highest in the Asia-Pacific region.
- In 2019 HBR named Gupta the 89th best-performing CEO in the world.

Invest Continually

- Choosing to be aggressive with AI is a commitment with **significant implications**:
  - It will have a major influence on a company for decades;
  - For large enterprises may ultimately involve hundreds of millions or billions of dollars.
- At first such resource commitments may be scary for organizations.
- After seeing the benefits organizations received from early projects, AI-powered companies found it much easier to spend on AI-oriented data, technologies, and people.

CCC Intelligent Solutions

- CCC Intelligent Solutions has spent and expects to continue spending more than $100 million a year on AI and data.
- The company was founded in 1980 as **Certified Collateral Corporation**.
  - Originally created to provide car valuation information to insurers. If you've had a car accident requiring substantial repair work, you've probably benefited from CCC's data, ecosystem, and AI-based decision-making.
  - Over 40-plus years CCC has evolved to collect and manage more and more data, to establish more and more relationships with parties in the automobile insurance industry, and to make more and more decisions.
- CCC has enjoyed solid growth and is approaching $700 million in annual revenues.
- CCC's **machine-learning models** are based on more than a trillion dollars' worth of historical claims, billions of historical images, and other data on automobile parts, repair shops, collision injuries, and regulations. It also has gathered more than 50 billion miles' worth of historical data through telematics and sensors in vehicles.
- It provides **data** and **decisions** to an extensive ecosystem of some 300 insurers, 26,000 repair facilities, 3,500 parts suppliers, and all major automobile-original-equipment manufacturers.
  - All those transactions take place in the cloud.
  - They connect 30,000 companies and 500,000 individual users and process $100 billion worth of commercial transactions annually.
- CCC's goal is to link those diverse organizations in a seamless ecosystem to **process claims quickly**.

Always Seek New Sources of Data

- Gathering data is typically not a problem for large companies, but AI strategies are driven in large part by whatever data can be assembled:
  - More data is good.
  - More accurate data is great.
  - More accurate, structured data that can be applied to AI models immediately is ideal.
- Integrating data from client systems can be very challenging.
- Data is not just words and numbers. It could be images and videos as well.

Companies with:

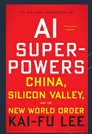
the most aggressive AI adoption

the best integration with strategy and operations, and

the best implementation

will achieve the greatest business value.

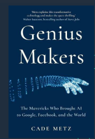
Reading Assignment



Read Chapter 5 (The Four Waves of AI) of the book "AI Super-Powers" by Kai-Fu Lee.

- The chapter identifies "four waves" of AI progress with distinct characteristics:
  - Internet AI
  - Business AI
  - Perception AI
  - Autonomous AI
- Identify the characteristics of and discuss opportunities arising from the four waves of AI.
- What are the necessary means to reap these opportunities in different entrepreneurial scenarios?

Reading Assignment



Read Chapter 14 (Hybris) of the book "Genious Makers" by Cade Metz.

- The chapter discusses Google's failed attempt to enter the AI market of China and Baidu's strategy.
- Identify the characteristics which, according to Baidu's Chief Operations Officer, are necessary to enable the emergence of AI-based innovations with a large impact.

Reading Assignment

- "Strategies for an Accelerating Future." by Ethan Molick, 2/2024
  - https://www.oneusefulthing.org/p/strategies-for-an-accelerating-future
- "Stop Tinkering with AI, It's time to go all in." by Thomas H. Davenport and Nitin Mittal. Harvard Business Review, Jan-Feb 2023.

Module 3: AI Companies

Section 2: Data Learning Effects

WHAT COMPETITIVE ADVANTAGES CAN AI BRING TO BUSINESS?

Competitive Advantage Factors

- Learning effects:
  - Human vs Machine formula
- Scale effects
  - Scale effects with data
- Network effects
  - Entry vs next-level data network effects

Factors of Competitive Advantage

Learning effects

HOW HUMANS LEARN?

THE HUMAN FORMULA..

The Human Formula

- Humans collect information **across** and **between generations**
- Human ability: get information from a collective - a network - and derive new information from it
  - a form of cooperation in space and time
- Allows for compounding growth as we are not always going backwards to relearn things.
- The more you know —> the more you **can** know
- The more information you can access across your network —> the faster you learn.

Learning Effects

- Economists study learning effects:
  - **The process through which information leads to economic benefit**
- Example:
  - Management consulting firms exploit information accumulated across all of their clients to develop strategic frameworks, best practices, and resource allocation models
- Traditional learning effects **accumulate**:
  - … information on individuals or organizations
  - … structured or unstructured information
  - … when information is processed by people or machines
  - … a qualitative or quantitative benefit
- Limits: they grow slowly because information must be processed or structured by a human before it can be processed by a machine
  - Humans can process only certain types of information
  - Organizations generally limit the internal and external flow of information

?

WHAT HAPPENS IN THE MACHINE AGE?

THE MACHINE FORMULA..

The Machine Formula

- Machines can form collectives - networks - to compute information:
  - Capture a critical mass of data
  - Develop capabilities to process that data into information
  - Feed that information into a computer that runs calculations over data to learn something new

Human vs Machine

Human Formula

- Gather data across our senses and process it in parallel
- Process input into useful information
- Get information from a collective – a network – and pass it to the next generation a form of cooperation across time and space

Machine Formula

- Gather data through sensors.
- Form collectives – networks.
- Capture a critical mass of data
- Develop capabilities to process that data into information
- Feed that information into a computer that runs calculations over data to learn something new.

Factors of Competitive Advantage

Scale effects

Scale Effects

- Scale effects refer to competitive advantages gained from increased scale in supply. E.g.
- Accumulation of assets or capabilities can lead to:
  - lower costs
  - reduced prices
  - increased demand and
  - more scale gained.

Source: Ash Fontana (2021) **"The AI-First Company"**

Scale effects with data

- **More data** can offer a competitive advantage: larger datasets can potentially reveal more **insights**, **patterns**, and **correlations**.
- However, there's a point where **accumulating more data** reaches **diminishing returns** because it is effectively duplicating existing data or failing to provide novel insights.
  - Beyond this point, additional data may not significantly enhance the utility of a product or service.

Scale effects with data

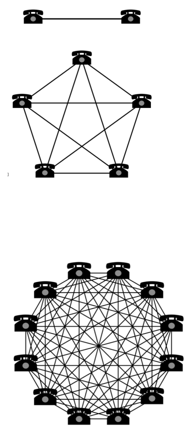
- The distinction between data and Information informs whether data has marginal utility - information is measured in how much uncertainty it resolves to the receiver.
  - **Data** refers to raw, unprocessed facts and figures
  - **Information** is data that has been processed and organized in a way that resolves uncertainty or adds value to the recipient.
- One way that data becomes information is by interacting with other data;
  - Interactions typically happen across a network where different datasets intersect and contribute to each other's meaning or significance.
  - Through these interactions, data can be refined, contextualized, and transformed into actionable insights or knowledge.

Factors of Competitive Advantage

Network effects

Network Effects

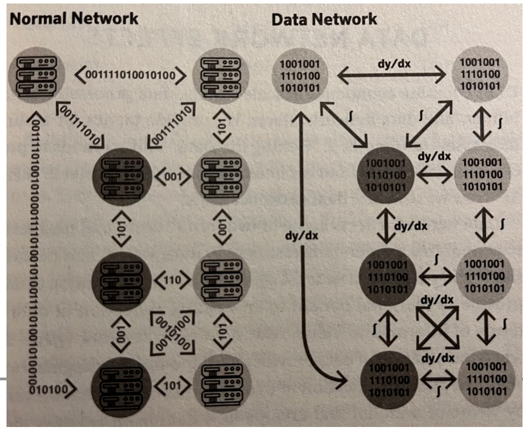
- Network effects means that the value of a network is larger than the sum of the value of its nodes; and
- the value of the network grows faster than the size of its nodes.
- Network effects occur when, from a consumer's perspective, a product becomes more useful as more people use it.



Source: Ash Fontana (2021) **"The AI-First Company"**

Data Network

- Definition: A set of data that is built by a group of otherwise unrelated entities, rather than a single entity.



Source: Ash Fontana (2021) **"The AI-First Company"**

Data Network Effects

- Usefulness of a product/service is enhanced by the addition of data to the network
- **Network edges** are **informational** and **calculate**, delivering information to other nodes on the network
- "Data" networks transmit derivatives of data (information) not just data itself.

Source: Ash Fontana (2021) **"The AI-First Company"**

Data Network Effects Categoies

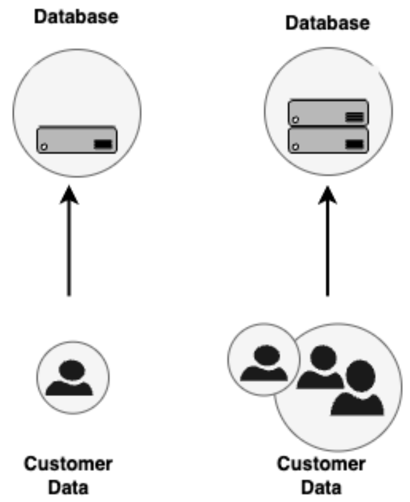
- Two forms: Entry-level and Next-level
  - Each requires a different investment in data, talent, and partnerships.
- **Entry-level**: the addition of data provides a marginal benefit to existing collection of data in terms of information value.
- **Next-level**: the addition of data provides a compounding marginal benefit to an existing collection of data in terms of information value by virtue of a model that creates new data from existing data (e.g. ML).

Entry-level Data Network Effects

- Situations where: a person use her brain to accumulate data, turn it into information, compare it to other information, learn something, make a prediction, make a decision, and then learn more from the effect of that decision.
- Any leveling-up limited by that person's brain.
- The **learning isn't shared**; just make a decision & move on.
- Are a form of collective intelligence: obtaining more information from a collective of individuals helps make better decisions.
- Occur with products/services where companies get data from customers (*give-to-get*): customers contribute data that may make the product more useful, attracting more customers.

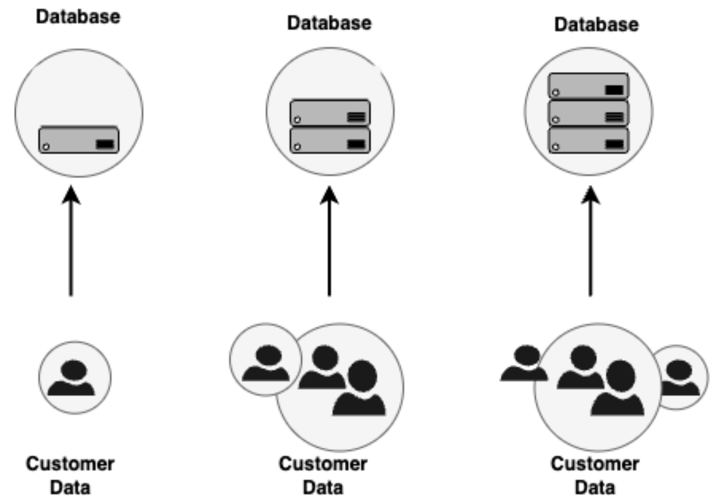
Entry-level Data Netw. Effects

Entry-level Data Netw. Effects



University of Cyprus
Department of Computer Science

Entry-level Data Netw. Effects



University of Cyprus
Department of Computer Science

Entry-level DNE

- **Shopping** becomes better when you have more:
  1. selection and
  2. ability to select.
- **Department stores:** lots of products in one place; store associates help you select:
  - More selection comes from sourcing inventory: **no data network effect.**
  - More ability to select comes from gathering, structuring, and presenting information on those products: a **data network effect**.
- **E-commerce sites** enhance **selection** and **ability to select**.
  - Often, the e-commerce product data takes the form of consumer reviews: each review gives you a broader perspective on product, enhancing your ability to make a purchasing decision.
  - By reading the reviews, you're reaping a benefit from everyone who wrote a review.
  - When you buy a product and then write your own review, **you kick off the data network effect:**
    - the last person who wrote a review affects your purchasing decision
    - if you purchase the product and leave a review, you're both helping to make the sale to the next person.
  - Entry-level data network effects compound with the addition of **exogenous information** to the network.

Factors of Competitive Advantage

“Next-level” Data Network Effects

University of Cyprus
Department of Computer Science

Next-level Data Network Effects

Occur when going **beyond our own brain** to utilize:

- even bigger collective networks computed at a larger scale
- … running on many computers
- … in many places
- … and on more data at the same time.

- The addition of data with something that generates more information (like AI) makes something more useful.

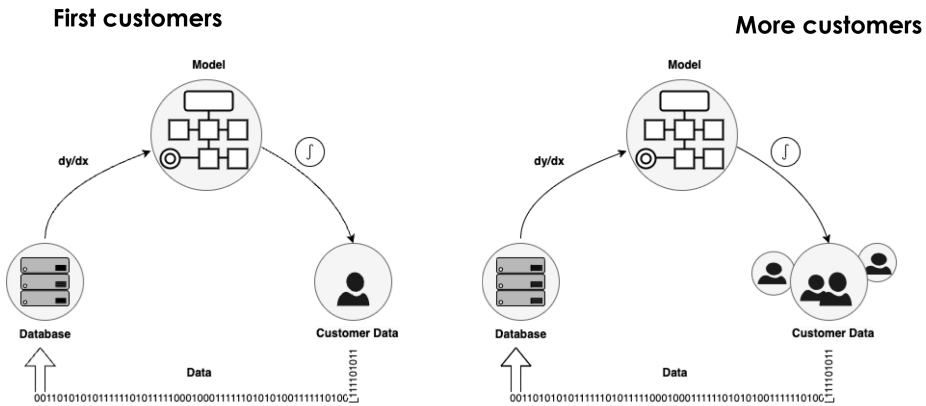
Next-level Data Network Effects

- In practice, these happen when
  - each consumer of a product generates data by providing data as part of using the product (**feedback**) and where data…
  - … feeds a system/model that compounds the value of this data and
  - … turns it into information/predictions.
- The consumers of the product:
  - effectively form a network of data contributors, and
  - benefit from the data added by new contributors
  - because the network generates information.

Next-level DNE

- **Online shopping** often starts with a search
- Shopping experience is better when you find faster what you want to buy (better ability to select).
- Search engines powered by ML models can:
  - use data gathered from what people typed into the search box, clicked on, purchased, and positively reviewed to…
  - … move products up and down on the search results page ideally presenting the most relevant result first.
- The ML models learn as shoppers either click on or ignore the search results.

Next-level Data Netw. Effects



Entry vs Next Data Net. Effects

Entry-level:

- Direct
- Easier to build: just add information to the network by getting users, customers, and partners to contribute information.
- Start when the addition to a network makes the product on top of the network more useful.
- New information is exogenous.
- Think about a simple table of data, then add another row of data to that table; the table is now more useful because it has more data to analyze and turn into information.

Next-level:

- Indirect
- Require building something else
- Automatically multiplies the size of the network (a system that generates its own information).
  - Grow faster because the network has a multiplying factor that compounds the competitive advantage.
- Addition of new data **plus** AI generates new information.
- Key difference with entry-level: feedback data and predictive models.

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

DNE requirements

	ENTRY LEVEL	NEXT LEVEL
Data	High	Low
Technology	Low	High
Talent	Low	High
Customers	High	Low
Partnerships	High	Low

Factors of Competitive Advantage from AI

AI comes to play: Data Learning Effects

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

Data Learning Effects

DLE: the **accumulation of information from data that automatically compounds**.

Possible thanks to:

- Economies of scale to data: data deluge in Internet, captured by sensors on personal, industrial, Internet devices and platforms
- Data processing capabilities: Cloud runs calculations over data at a reasonable cost and people can make connections between disparate datasets
- **Data network effects**: Intelligent systems allow **data to be organized into networks**, wherein calculations run on one part of the network, results are sent to another part for more calculations, and come up with new information

Source: Ash Fontana (2021) **"The AI-First Company"**

Data Learning Effects

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

Data Learning Effects:
economies of scale to data +
data processing capabilities +
data network effects

- How to achieve DLEs:
  - Get lots of data
  - Process it into something useful in terms of making a decision
  - Create a system that **automatically generates more useful data**

DLE Characteristics

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- DLEs can accumulate information:
  - Across single or multiple organizations
  - That is structured
  - When processed by machines
  - That has a quantitative benefit
- DLES have few limits:
  - They grow fast because structured information feeds into machines that calculate faster than humans
  - Modern computers can process multiple types of information fast

Source: Ash Fontana (2021) **"The AI-First Company"**

Data Learning Effects: Remarks

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- Data generates marginal output when combined with data processing capabilities and data network effects.
- Data learning effects articulate the value chain around data:
- Data learning effects:
  - start with a supply side competitive advantage that …
  - … kicks off a demand-side competitive advantage and
  - … combines privileged access to a resource with capabilities to transform that resource into something valuable.

Factors of Competitive Advantage from AI

The Power of Data Learning Effects

Data Learning Effects Power

- “Winner takes all” with DLEs
- DLEs make products more useful
- DLEs compound faster than network effects
- DLEs drive cost leadership
- DLEs and Price Optimization

Winners take all

- **Markets tip to a single technology** that succeeds in offering both:
  - Economies of scale: marginal cost decreases as production increases, and
  - Demand for variety: customers demand for lots of different features or products
- DLEs can:
  - have high economies of scale to data: they need lots of data
  - create a wide variety of predictions: constantly generate different predictions based feedback data.
  - support customers with high-demand for variety: each customer needs a model specific to their business to make predictions.
- DLEs tip markets in favor of one winner.

Make products more useful

- DLEs usually work in the background rather than the foreground.
- Manifesting utility in the **foreground** means that the increasing utility of DLEs is obvious to the end user
  - User sees that adding some data generates a more accurate prediction for them or immediately triggers a new insight.
- Manifesting utility in the **background** means that increasing utility of DLEs is not obvious to the user.
  - The user doesn't see that adding some data generates a better prediction for them, and they don't see any information they don't already have.

SQUARE CAPITAL

DLE example (foreground)

- **Square Capital**, lends money to customers of the company's point-of-sale (POS) systems:
  - A product getting better in the foreground thanks to DLES.
- Merchants, such as restaurant owners, add data from their POS systems and get loans based on how much money they’re making, almost immediately.
- Square is able to offer them that loan by comparing the data uploaded by the merchant to other merchants that previously received loans from Square.
- The user (the merchant) sees an immediate benefit of adding data (the offer of a loan) because the product utilizes a DLE to run a predictive system that delivers that **product**: an interest rate based on the prediction that the merchant will pay back the loan.
- You don't get a loan if you don't add data, and you qualify for a loan only because your data can be compared with other data to generate a prediction.

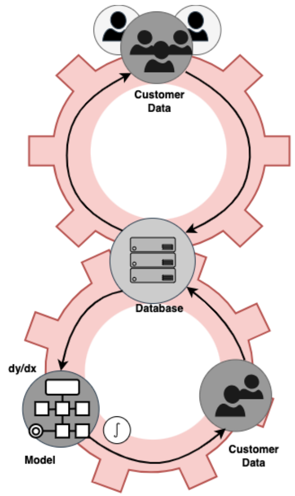
CLOUDFLARE

DLE example (background)

- **Cloudflare**, a website performance and security product, based on DLES, which gets better in the background.
- Customers such as news websites add a Cloudflare data collection mechanism to the network, which serves up page requests to the website's viewers and provides protection against requests from bad actors.
- Cloudflare offers protection by comparing which requests were bad for other Cloudflare customers, performing DoS against them or trying to exploit security holes.
- The customer (the website owner) does not see an alert to deny a particularly dangerous request *immediately after adding the Cloudflare data collection mechanism*, but the product is constantly learning and delivering alerts to potentially bad requests.

DLEs compound faster

- DLEs arise when the process of collecting data and generating information is automated, allowing the compounding with other things learned.
- "Flywheel" effect: one network effect kicks into another, more powerful network effect.



Source: Ash Fontana (2021) **“The AI-First Company”**

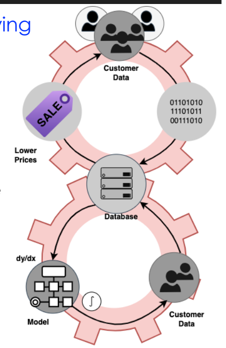
Amazon’s “flywheel”

amazon

- Amazon, first, gathered a great deal of data on products and helped customers make better buying decisions by putting all of that data in the product listings, providing comparison tables with structured product information.
  - More information meant better comparisons and decisions.
- Then Amazon invested in to build ML search and recommendation systems: **A9**.
  - The A9 team got matched product data with purchase data to learn which products customers want to buy so that Amazon could recommend similar products to those customers in listing pages and search results.
  - Gathering a lot of data started the **entry-level network effect**: Amazon was the most useful shopping website to consumers because it had the most product information.
  - Learning over that data kicked off the **next-level network effect.**
- Amazon is the most useful shopping website to consumers because it offers the best recommendations and has the best search experience.

DLEs driving Cost Leadership

- DLES make products more valuable by improving performance of underlying models that generate value for customers, empowering them to make:
  - more accurate predictions,
  - better decisions,
  - and achieve a higher ROI (%) = (Revenues from Investment - Cost of Investment) / Cost of Investment
- Companies need to spend a lot of money before their customers can see value in predictions:
  - Getting the data: Buying, collecting, cleaning, storing
  - Getting the expertise: Hiring data scientists and ML engineers
- **Expenses start falling** once customers start using the product/contributing data. Thus, businesses can:
  - reduce the cost of making products and increase their value for customers.
  - achieve cost leadership by **charging less** or charging the same but **providing more value**.
- Cost leadership is a strategy that may help to quickly build a DLE, **attracting more customers**, who, in turn, **generate more data**.
  - Possible unprofitability for limited period of time to accumulate the critical mass of data required.



Data Learning Effects Power

- "Winner takes all" with DLEs
- DLEs make products more useful
- DLEs compound faster than network effects
- DLEs drive cost leadership
- DLEs & Price Optimization

Price Optimization

- The strategic use of data to determine the willingness of customers to pay, ensuring maximum profitability.
- Price determination based on:
  - Experience, observations, guesses, or
  - predictive systems trained on data from prior pricing experiments and/or
  - personalization, driven by continuous experiments that take into account customer profiles.

Price Optimization Strategy

- Information game seeking to predict what someone will pay for a product:
  - Utilize predictive systems and experiments for price setting.
  - Use historical data for better pricing accuracy.
- E-commerce websites often personalize pricing:
  - Run manual vs. automated pricing experiments
  - Leverage customer data and behavior to implement AI-driven dynamic pricing strategies
  - Implement yield management systems: consider numerous variables (e.g., seat availability, time of year) to set prices(airlines)
  - Extract maximum willingness to pay through promotions that may incur short-term losses for long-term data gains

DLEs & Price optimization

- Using DLEs for pricing leads to a virtuous cycle where:
  - better pricing attracts more customers, providing more data,
  - which in turn is used to refine DLEs and ML models for even better pricing.

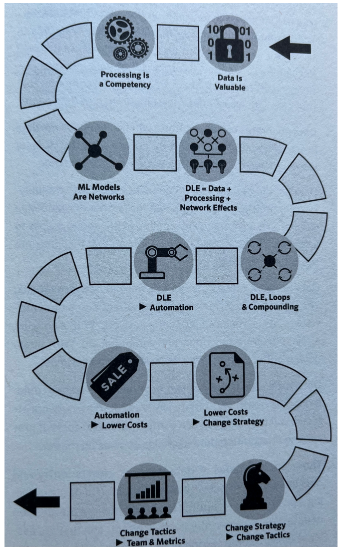
Benefits of Predictive Pricing

- Better pricing:
  - Yields higher profits, allowing reinvestment in ML R&D, enhancing DLEs.
  - Attracts more customers and, thus, more data (at no cost), increasing profits.
  - Reduces expenses in sales and marketing, increasing profits.

Building Data Learning Effects

Steps:

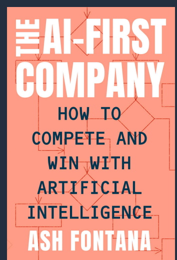
- Capture a critical mass of data;
- Develop capabilities to process that data into information;
- Feed that information into a computer that runs calculations over data, learning from new data points.



Source: Ash Fontana (2021) **"The AI-First Company"**

Module 3: AI Companies

Section 3: Lean AI



WHAT IS THE PRODUCT OF AN AI COMPANY?

PREDICTIONS!

WHAT IS THE KEY METRIC TO EVALUATE PRODUCT PERFORMANCE OF AI COMPANIES?

ACCURACY OF PREDICTIONS!

HOW DO YOU FIGURE OUT WHAT CUSTOMERS NEED?

Figuring out what customers need

- **Induction**: process is focused *loosely* on getting information from a group of **potential** customers to:
  - induce a demand trend and
  - come up with a list of features to meet that demand
  - Process entails: running surveys, coming up with a design to test with different groups of potential customers, iterating on the design.
- **Deduction**: process focused *tightly* on getting information from **one** customer to:
  - deduce a supply trend and
  - formulate a list of product features; often used with existing business offers, by seeing what worked in the past and adding new products.
  - Process entails: calling customers, coming up with a design, having lots of meetings to refine it, and collecting feedback on the product post-implementation.

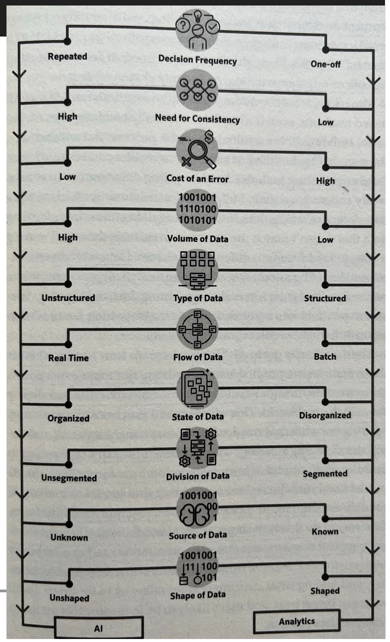
Figuring out what customers need

Does the product prototype need AI to work well?

- Answer depends on whether customers need to:
  - generate an insight
  - make a prediction, or
  - automate a process.
- AI helps make better and faster decisions.
- How much AI depends on decisions a company is making and the data on which it is basing those decisions.

Lean-AI Decision Tree

- To decide whether you need analytics and/or AI:
  - go through a decision tree
  - put the types of decisions and data into two buckets
  - see which scores more.
- If analytics is heavier, then it's likely that customers need features on data: **logging**, **cleaning**, and operating to derive **statistical properties**.
- If AI is heavier, then it's likely that customers need AI features: **classification**, **segmentation**, and **manipulation of data**.



 Source: Ash Fontana (2021) **"The AI-First Company"**

Before developing an AI solution

- **Data Engineering**
  - Instrumenting data sources to consistently collect good data
  - Building infrastructure in which to store the data
  - Extract data from existing data stores
  - Transform the data that does not match the structure of existing data
  - Make it easy to load the data into different databases
- **Data Science**
  - Understand the meaning of data
  - Detect anomalies
  - Setup analytical processes on the data on regular intervals
  - Segment data
  - Aggregated datasets to put data into context
  - Figuring our which features of an algorithm might predict something useful

Source: Ash Fontana (2021) **"The AI-First Company"**

Getting to the AI solution / product

- Test if identified features are predictive of something
- Experiment with more data
- Design new algorithms
- Train models
- Deploy models in the real world
- Joint undertaking with customers:
  - Figure our what they need: **analytics** or **AI**
  - Do the data engineering and data science
  - Do the ML engineering to build a small model
  - Do testing to guide how to package the AI model and build the right team to bring that model to market

HOW DO YOU START?

START SMALL!

Sales, models, products succeed when starting small by answering one question, for one set of stakeholders, using one method!

Expand engagement from there, by picking the best possible problem to tackle for the most motivated customer.

HOW DO YOU START SMALL?

Start small: Statistics

- Use statistics to establish what customers want and what their data are saying:
  - Histograms, scatter plots
  - Clustering to group similar objects
  - Dimensionality reduction, to reduce the measures associated with each data point (PCA)
- Try to pinpoint interesting features to include in a ML model, explore the importance of different features:
  - Variable importance plots
- Focus on one statistical question / one equation to see if it answers customer's questions.

Start small: Data Science

- Starting with data science analysis of a
  - a single dataset that is likely to have the answer, and
  - provide personalized, data-driven answers to a single question of the customer
  - by identifying and using one predictive feature to find and give that answer

to demonstrate potential for *return on investment*:

$$ROI(\%) = \frac{(Revenues \quad from \quad Investment - Cost \quad of \quad Investment)}{Cost \quad of \quad Investment} x100$$

Tradeoffs

- Answering multiple questions requires:
  - collecting
  - combining
  - processing

  multiple datasets
- Significant investments data collection, processing, analysis

	ONE-OFF	SELF-LEARNING
Data acquisition	Manually fetch	Automatically fetch through a direct database connection
Data preparation	None—pick a clean dataset	Clean and label multiple datasets
Storage	Local	Cloud
Data pipeline	One pipeline	Many pipelines
Feature development	Find one feature	Try many features
Training	One calculation	Many calculations
Computation	Local central processing unit (CPU) or graphics processing unit (GPU)	Cloud GPUs
Modeling	One model	Network of models
Deployment	Local	Cloud
Presentation	Print a report	Build an interface

Source: Ash Fontana (2021) **"The AI-First Company"**

Start small: data

During early experimentation it is not the time to build a "fat" data pipeline but rather the time to stay lean.

1. Focus on getting just enough data to build an AI and demonstrate its use:
   - Hopefully, this means just one dataset located in one database and can be retrieved with a single query.
   - Best starting point: customer's first guess at which data might be predictive.
2. Data preparation and formatting: minimal at this stage, as data is taken from one data source.
3. Data cleaning: fill in missing values, delete duplicates, remove errant values.
4. Make sure the data is efficiently computable by the models.

Most of the this isn't a major consideration with small-scale experiments.

Source: Ash Fontana (2021) **"The AI-First Company"**

Start small w Data: What to avoid?

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- DO NOT LABEL EXTENSIVELY. Determining at the outset what data customers have that might be predictive, and can spare the time-consuming and costly task of data labeling.
- DO NOT HARVEST DATA FROM MULTIPLE SOURCES. Doing so requires obtaining extra permissions, building more integrations, and more formatting. Instead: pick one dataset in one data store, run an experiment, then get another only if the dataset doesn't have any predictive power.
- DO NOT WORK WITH SENSITIVE DATA. Anonymizing data is costly and may obfuscate results. However, it may be necessary to avoid being held responsible for a data breach.
- DO NOT BUILD A SEPARATE DATA STORE. Instead, just download the dataset somewhere secure with low latency, such as a local machine.
- DO NOT BUILD A DATA PLATFORM. Decide on all the tools that the entire team will use to explore and manage data. Needs are very likely to change, so consider delaying this choice beyond the initial phase of a project.

Source: Ash Fontana (2021) **"The AI-First Company"**

Benefits of starting small

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- Easier to build trust, gain access to customer data, and learn how they interact with legacy data tools.
- Increases customer engagement because it is easier to:
  - meet expectations about the power (accuracy) of the model;
  - avoid/prevent data privacy and security issues.
- Reduces the need to wrestle with poor data from multiple DBs.
- Makes deployment relatively simple and reduces chances of solution breaks, when starting with one algorithm.
- Delivers predictions to teams in a way that's easy for them to consume. Starting small makes explaining the "why analysis works" easier.

AI Companies

The Lean AI Approach

Lean AI

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- The process of trying to make predictions from a small sample of data, then presenting those to stakeholders to figure out their expectations regarding the product and its features.
  - Approach intended to build a small but complete AI to solve a specific problem.
- Lean AI entails:
  - Implementing and evaluating a Proof of Concept (**POC**) phase for the customer
  - Coming up with a plan on how to reach a higher level of accuracy and reset expectations with the customer based on that level of accuracy.
- The process can help new and established companies to become AI-First companies.

POC Design: The Lean AI Approach

- Accuracy: **Set a benchmark for predictions** based on honest assessments of what's feasible technically.
  - The extrinsic way to set a benchmark is based on what accuracy a customer already achieved through their own efforts.
- Business goal: Define the **metric** that gets closest to what customers **need to hit to make money**.
  - A business generally has a good idea of its goals but sometimes needs help to understand the ways that AI can help it achieve them. Then separate the goals between those to hit during the POC and those to hit in subsequent engagements or phases.
- Data: List the data sources needed and decide if they're accessible.
  - Typically, **80%** of the time dedicated to building AI is spent preparing data and the other **20%** is spent creating the models.
- Dependency: Document dependencies on legacy systems, to mitigate problems.
- Team: Limit the team members, to strike a balance between getting enough stakeholder engagement and getting the work done.
- Timeline: Assess what to build, how long it will take, and how long it will take to hit the accuracy benchmark.
  - Help customers understand that **getting to 80% accuracy** may take **just 20% of the time**, but achieving that **remaining 20%** may consume **80% of the time**.
- Cost: Clarify the total cost after figuring out the time required, external consultants, labeling data, and engineering time.

Lean AI vs Lean Startup Milestones

Lean Start-Up	Lean AI
Minimum Viable Product	Minimum Predictive Accuracy
Product Features	Model Features
Output a Calculation	Output a Prediction
Performant	Accurate
Functional	Reliable
Product Usage	Prediction Acceptance
Launch a Company	Launch an AI-First Product

Building a Lean Startup vs Lean AIs

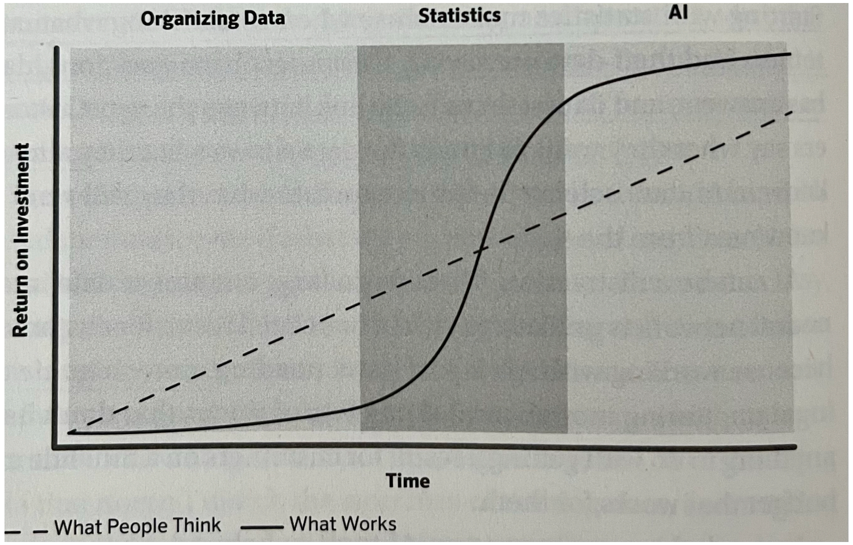
Step	Lean Startup	Lean AI
0	Customer needs & pain points	Data aspects of the problem
1	Determine product features	Determine model features
2	Build a product	Generate a prediction
3	Show a demo	Show a report
4	Receive qualitative feedback	Receive quantitative feedback
5	Build more features	Collect more data
6	Relaunch the product	Retrain the model
7	Measure usage	Measure accuracy
8	Launch a company	Launch an AI-First product

Prediction Usability Threshold

- Instead of focusing on MVP, Lean AI uses the concept of prediction usability threshold (**PUT**) as the target of an AI-first company
- PUT: the point (threshold) at which a prediction becomes useful to a customer.
  - Where the prediction starts getting **better than a human's**.
  - Sometimes a prediction is usable even if it's less accurate than what a human can make because it may be **more consistent** coming from a computer.
- A "prediction" usually means a classification.
  - The reason to define the PUT is to optimize the amount of time spent building and tuning the models that generate predictions before showing and selling them to customers.
  - That is, to not waste time getting data and building model features that don't make the prediction more useful to a customer.
- The PUT is nuanced and specific to a customer. Figuring out when, how, and for what purpose a customer needs a particular prediction helps to nail down the PUT.
- Ideally, at this stage, the PUT, customer ROI, and POC metrics are **linked**.

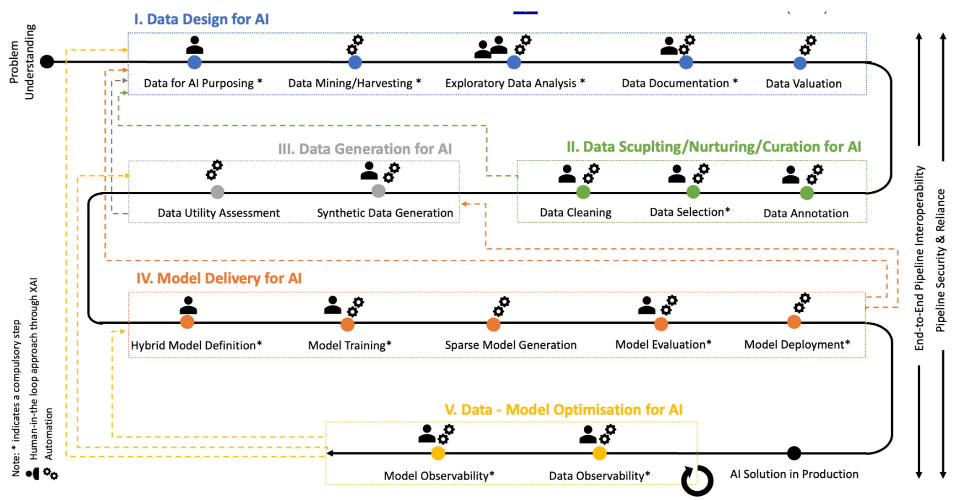
Source: Ash Fontana (2021) **"The AI-First Company"**

ROI: assumptions vs reality



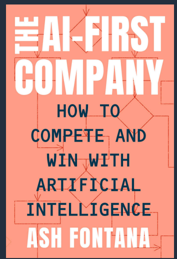
Source: Ash Fontana (2021) **"The AI-First Company"**

Data-AI Pipelines



Module 3: AI Companies

Section 4: Getting the Data



HOW TO VALUE DATA?

Data valuation framework

- Need to have before investing in data acquisition
- Two valuation axes:
  - Discrimination: Is it **hard to get**?
  - Determination: Is it **useful**?

Discrimination: Is it Hard to Get?

- Accessibility
- Availability
- Cost
- Time
- Fungibility

Discrimination: Accessibility

- Acquiring data may require physical efforts and processing, such as visiting locations to photocopy documents and converting them to digital formats using OCR software.
- Assessing future data obtainability is essential, often dependent on contractual or policy terms that govern access restrictions.
- Governments and private vendors may initially offer data for free but later impose charges, impacting its long-term availability.

Discrimination: Availability

- Data availability varies, with some systems imposing slow data harvesting rates to reduce costs or to differentiate products.
- Financial market data providers may restrict stock price data access, offering it only at specific intervals unless additional payment is made for more frequent access.
- Access to more timely data, such as receiving stock prices at shorter intervals, can offer a competitive edge for making crucial decisions.

Discrimination: Cost

- Data vendor costs present a significant barrier to obtaining AI trading data - prices vary from clear dollar amounts to complex revenue-sharing agreements.
- High-quality data (e.g. Bloomberg terminals) can be expensive, requiring thousands of dollars per month plus specialized software.
- Non-monetary costs may include contributing your internal data to the data vendor: hard to assess their exact cost.

Discrimination: Time

- The rate of data collection can provide a competitive advantage.
- Certain types of data are accumulated at a predictable rate, such as weather or employment data, which are controlled by natural phenomena or government bureaus, respectively.
- To obtain a critical mass of such data you simply need to collect them for a long time.

Discrimination: Fungibility

- **Fungibility** / interchangeability:
  - fungible data can be swapped out for different data without negatively affecting the quality of decision made based on that data.

Determination: Is it Useful?

- Perishability
- Veracity
- Dimensionality
- Breadth
- Self-reinforcement

Determination: Perishability

- The perishability of data affects its relevance: outdated data can lead to inaccurate predictions:
  - E.g. stock prices change rapidly, so recent data is vital.
- Data types vary in longevity, e.g.:
  - Mount Everest's height remains relevant over time.
  - Consumer preferences may have an intermediate perishability.
- Perishability is influenced by the updating frequency of data;
  - Clothing sizes may be stable.
  - Fashion trends can be short-lived and require more current data.
- The cost of perishable data is impacted by its need for frequent updates, where vendors might price data based on its freshness, and constant updates or processing of such data can incur additional expenses.

Determination: Veracity

- Determines reliability in the context of making a decision
- Often, requires manually validating data points.

Determination: Self-reinforcement

- **Self-reinforcing** data are those whose attributes remain the same or trend the same way as time progresses.

Determination: Dimensionality

- Dimensions are attributes of a given entity:
  - Typically, the # of columns in a table.
- Dimensionality is a powerful determinant of value:
  - Their intended use is training ML models
  - Each dimension informs the model.

Determination: Breadth

- Determines how closely the data represents reality.
  - More breadth means more examples of the same type, more variations in the attributes of entities and edge cases.
  - Typically, the # of rows in a table.
- Sometimes, more breadth comes through joining datasets from different sources or vendors (must have same attributes).

The Question of Volume

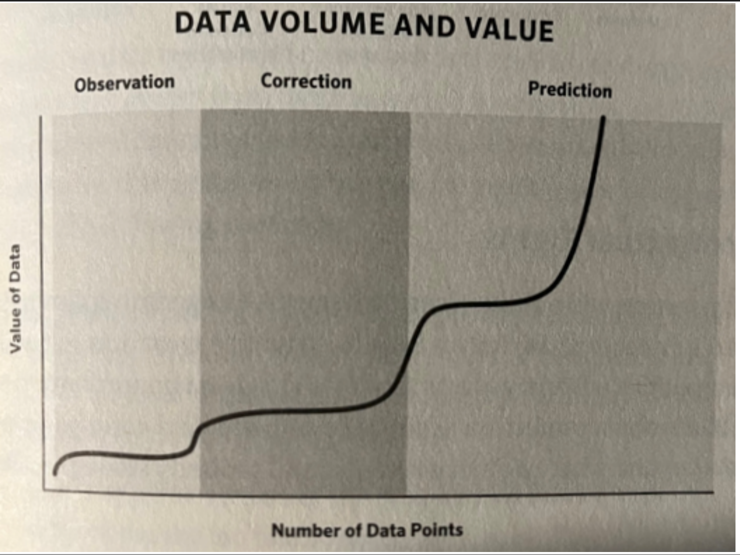
- Acquiring large volumes of data can be a concern:

  Trade-off: more data vs better data.
- Answer depends on:
  - the type of decision to make
  - the models that help make that decision.

Value vs. Volume

- **Observation Phase**: Initial data collection provides fundamental insights and identifies observable patterns or trends.
  - However, the value derived is limited to surface-level understanding without in-depth analysis.
- **Correction Phase**: Increased data volume leads to a focus on improving data quality, involving cleaning and correcting data, which is essential.
  - But it doesn't directly increase its analytical value.
- **Prediction Phase**: With clean and substantial data, advanced analytical techniques are applied to extract predictive insights, greatly enhancing the data's value for informed decision-making.

Value vs. Volume



Source: Ash Fontana (2021) **"The AI-First Company"**

WHAT IS THE MOST SIGNIFICANT SOURCE OF DATA FOR AN AI-FIRST COMPANY?

MAI4CAREU University of Cyprus Department of Computer Science

THEIR CUSTOMERS!

IT MAKES SENSE TO EXPECT THAT PREDICTIVE MODELS BUILT ARE BASED ON THOSE CUSTOMERS' DATA

MAI4CAREU University of Cyprus Department of Computer Science

Getting the data

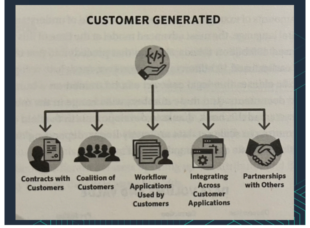
Customer-generated data

University of Cyprus Department of Computer Science

Customer-generated data

- Perhaps the most significant source of data for AI entrepreneurs and their startups.
- Why?
  - Typical products seek to predict something of commercial value or industrial consequence for those customers, exploiting their (and other) data.

Issues to consider



- **Contracts with customers**
- Customer coalitions
- Workflow applications
- Integrating across customer applications
- Partnerships with others

Legacy Agreements Challenge

- In the early 2000s, companies were reluctant to store data offsite.
- Cloud vendors often gave up any rights to customer data to alleviate concerns.
- Legacy agreements from the cloud era now pose challenges for companies trying to build intelligent systems, as they lack the rights to use the data effectively.
- **Getting the data is essential!**

The Clean Start Advantage

- **AI-first organizations** are not burdened by previous technology decisions (legacy), allowing for more flexibility in data strategy:
  - Can negotiate more favorable data rights.
  - Can adapt quickly to technological advancements.
  - With fewer constraints, AI companies can explore innovative data acquisition and usage strategies.

Negotiating for Data

- Chicken-and-egg problem for AI companies:
  - Company's leverage comes from demonstrating good results to potential customers.
  - Good results require data to train models.
- Approach:
  - Negotiate to find alternative sources of data.
  - Establish contracts on data access and use.

Negotiation strategies

- Target SME business customers : they may have more open attitudes towards sharing data, perhaps trading off data rights for reduced pricing.
- Offer free or at-cost products to capture essential data.
- Sell ancillary products at cost as a strategy to obtain valuable data.

Structuring negotiations

- AI-First startups can **achieve partnerships on data** and preempt compliance concerns by:
  - presenting impressive demos to large potential clients
  - stating that their interest is to learn from the data exhaust of their perspective customers:
    - user engagement & interaction data
    - metadata
    - data flow information.

Negotiation goals

Getting rights on customers' data to:

- Make better models for more useful products
- Adapt existing models to customers' changing conditions
- Prevent competitors from reaching similar levels of efficacy
- Own a valuable asset

Concepts in the negotiation

- Models used to make predictions about the customer based on her needs.
- Global, multiuser models, making predictions about something common to all customers, trained on data aggregated across all customers, useful *across* customers.
- Data of the customer, not processed in any way, often proprietary to customers.
- Anonymized and aggregated data: customer data processed so they cannot be referred back to a particular customer or user, anonymized, pseudonymized, randomized or redacted. Can be aggregated across all customers.
- Personally Identifiable Information (PII): may need to be handled in a certain way to comply with regulations.
- Storage in the public cloud or privately owned servers (different data may be stored in different places).

Structuring contracts

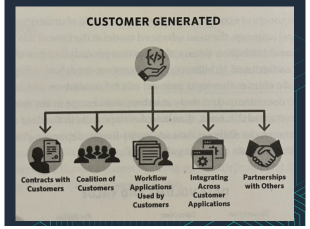
- Access to customer data through contracts: By carefully structuring contracts, companies can ensure ongoing access to vital customer data, enhancing their ability to innovate.
- Improving product functionality: Access to real-time data allows for continuous product improvement, ensuring that offerings remain competitive and relevant.
- Legal considerations in data protection: Contracts must also address **data protection** and **privacy concerns**, safeguarding against potential legal and reputational risks.

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

Contractual terms to avoid

- Customers owning global, multiuser models.
- Restrictions on using anonymized, aggregated data across customers to train those global models.
- Liabilities for managing PII.
- Fracturing data architecture by having to store data in different locations.

Issues to consider



- Contracts with customers
- **Customer coalitions**
- Workflow applications
- Integrating across customer applications
- Partnerships with others

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

Customer data coalitions

- A single company (the vendor) organizing a group of companies (the customers) to share data with one another.
- Often create a unique data asset for both **vendors** and **customers**.
- By forming coalitions, companies can share and access a larger pool of data, enhancing the quality and breadth of insights available.

Benefits for members

- Gain access to greater volumes of data:
  - Smaller companies often do not have enough data to train and run intelligent systems (overfitting).
- Can see greater variations of the same category of data from other members.
- Can validate data points of each other.
- Can gain access to higher frequency data from other members.

Challenges with building coalitions

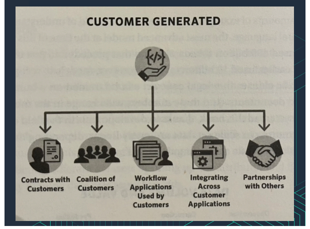
Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- Establishing and maintaining data coalitions involves navigating **technical**, **legal**, and **trust**-based challenges to ensure mutual benefit.
- E.g.
  - Marketing: getting members to team up requires **inspiration** to overcome hesitation in sharing data.
  - Contracts: same contract for all; no special deals.
  - Anonymization: customers may compete against one another, so sharing data should not disclose customers' identities.

Competing with Amazon

- **Amazon's** effective search and recommendations are attributed to their extensive data and engineering efforts since 2003.
- A **data coalition of retailers** could challenge Amazon by combining their data to enhance search and recommendation functions.
- Retailers in the coalition would share resources to create a high-quality product discovery experience.
- Such a collaboration would harness collective intelligence to improve machine learning suggestions and compete with Amazon.

Issues to consider



- Contracts with customers
- Customer coalitions
- **Workflow applications**
- Integrating across customer applications
- Partnerships with others

Workflow Applications

- A workflow application is a piece of software that *takes a sequence of things that someone does in the real world and puts those steps into software*.
- Workflow products gather data.
- Each piece of data entered into a workflow app goes into a database.
- Workflow apps are all around us: **we' re still in the era of building such workflow apps for many industries**.
  - A huge opportunity for software developers willing to learn how specific industries get things done.

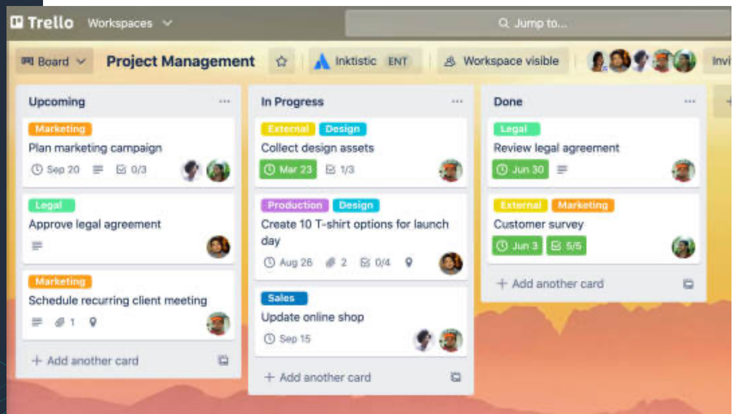
Workflow Applications

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- Business process data goes in and out of workflow apps continuously.
- Companies can use this data to build intelligent systems that go
  - beyond recording work to
  - … automating work by predicting a next step, and
  - … even automatically filling out parts of a form.

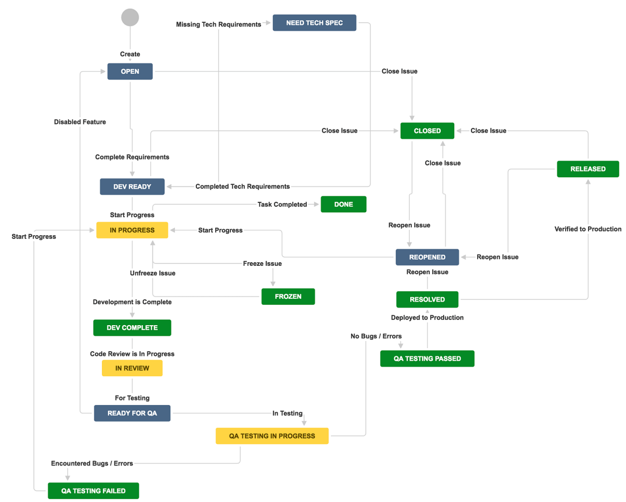
Workflow Tools:

Trello



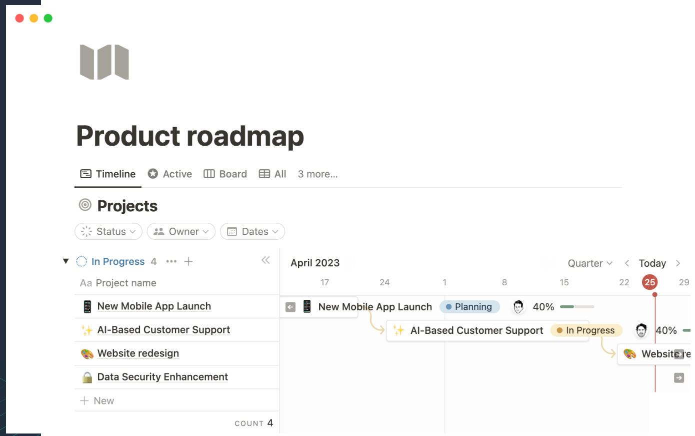
Workflow Tools

Jira by Atlassian



Workflow Tools

Notion



All business software can become "intelligent" thanks to the ability to:

- take real-world workflows
- develop software around them
- add more data and features.

Example: Project Management

- Take a list of tasks from a project manager on a construction site—currently written down on a piece of paper.
- Input them into a mobile app that the manager can track those tasks, assign them, and generate a report.
- Learn from the data and automate tasks or suggest improvements to the business processes.

Example: Car insurance claim processing

1. Assessor *goes out* to *damaged vehicle* sitting in the body shop, *takes some photos* of it and *writes a report*.
2. Report is *sent to a loss adjuster* at the insurance company so that she can *decide* whether to pay for repairing or replacing the car, and for how much.
3. Sometimes, the customer will disagree with the assessment, so the whole *process is repeated*.

These steps currently happen with Pen-and-Paper, clipboards, faxes/ emails, and cameras.

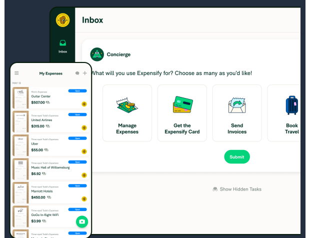
Example: Car insurance claim processing app

- A single app with text fields for the report; automatically sends the report to a queue for the loss adjuster to process.
- Disputed claims can be kicked back to the assessor for reassessment if necessary.
- Data on what type of damage should be repaired or replaced, how much it will cost to repair, what cars are valued in what way, etc, is most valuable and very hard to gather.

Example: Car insurance claim processing with AI

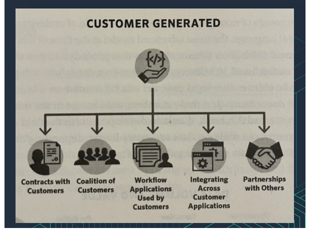
- Using AI to improve workflow app:
  - An app that inserts the car's specifications
  - Takes photos and determines from these the extent of the damage
  - Calculates total cost and presents report to the assessor
- Automation benefits:
  - Cuts down many of the more tedious aspects of the assessor's job
  - Reduces traveling
  - Audits or replaces the work of the loss adjuster

Expensify



- First step: Software for improved management of expenses, offering integration with banks and credit card providers
- Second step: AI-based product to automatically categorize expenses, trained from data from core workflow app

Issues to consider



- Contracts with customers
- Customer coalitions
- Workflow applications used by customers
- **Integrating across customer applications**
- Partnerships with others

Integrators

- Applications increasingly make data available through APIs.
- Data integration software **passively** assembles large volumes of data.
- Facilitates:
  - Linking one data source to another.
  - Normalizing data across sources.
  - Updating integrations as the connections to the sources change.
- **Data integrators** build a valuable data asset by:
  - directly collecting data that flows through their pipeline
  - generating metadata based on usage patterns, derivations or other observations

Examples

MuleSoft

zapier

TWILIO Segment

Segment: Step 1

TWILIO Segment

- Started with simple analytics product that was getting data from:
  - Mobile apps through iOS, Android
  - Web apps through own plugin, Shopify, Word-Press
  - Servers (directly)
  - Many cloud apps, from CRMs, to payment to email apps
- Data piped into email systems, analytics dashboards, Helpdesk apps, marketing attribution tools, and data warehouses.

Segment: Step 2

TWILIO Segment

Using data acquired through customers to build an intelligent system that:

- Automatically unifies user history across data sources into one comprehensive profile with an associated, intelligently generated user persona
- Synthesizes data into traits, audiences, and predictions for each customer
- Uses these enrichments to personalize marketing campaigns and in-app experiences

Challenges in building Integrators Start-ups

- Easy entry: Basic forms of data integration are easy to replicate, and the need for data integration is well known
  - There are lots of companies in this market.
- Competition: Cloud computing companies have a strategic imperative to offer a product that pulls in data from a multitude of sources
  - They wish to be able to charge more for storing more data.
- Unexpected Costs: Data pipelines break for unforeseeable reasons that are specific to sources
  - Fixing them can involve a great deal of manual work.
  - Heavy, unforeseen work can cost a lot, reducing gross margins.
- Low pricing power: purchasing a product to integrate data is typically a decision made after the purchase of another product that's solving the core problem.
  - Customers tend to have less of their budget left when they get around to purchasing the integration product,
  - or they just go with whatever data integration the vendor of the core product recommends.

Workflow vs Integrator Apps

- Integration-first apps gather voluminous, near real-time, structured data directly from machines & software:
  - Help develop more valuable data assets
  - Can be more difficult to develop
  - Gather data that is often non-proprietary
  - Lead to predictions of higher quality
- Workflow apps may collect user-input data that may be inaccurate, unstructured, out of date:
  - Applications are easier to build

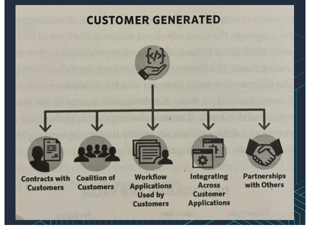
Workflow vs Integrator Apps

- Workflow applications as a data source: These applications can be a rich source of operational data, offering **insights** into **user behavior** and **process efficiency**.
- Advantages of data integration: Integrating data from various sources can significantly enhance product functionality, providing a more **comprehensive view of customer needs** and **opportunities for innovation**.
- Strategies for effective integration: Effective data integration requires careful **planning**, robust technology **infrastructure**, and strategic **partnerships**, ensuring seamless functionality across services.

Workflow-first vs Integrator-first Companies

Feature	**Workflow-First**	**Integrations-First**
Data Collection	Collect data from what humans see	Collect data from what machines do
Data Analysis	Analytics on static data	Machine learning on streaming data
Response Time	Reactive/post hoc	Proactive/real time
Impact on Data Originators	Threatening to data originators/other workflow apps	Neutrally positioned with respect to workflow apps

Issues to consider



- Contracts with customers
- Customer coalitions
- Workflow applications used by customers
- Integrating across customer applications
- **Partnerships with others**

Partnerships for Data Acquisition

- Complementary data through partnerships: Forming strategic partnerships can provide access to **complementary data sets**, enriching the company's analytical capabilities.
- Advantages:
  - Complementary data: increased value of existing data for both partners
  - Complementary business models: if partners make money in different ways it may be easier to collaborate
- Partnerships can be a cost-effective way to enhance data assets, enabling companies to leverage external data without the need for extensive investment.
- Many examples of successful partnerships demonstrate how companies can significantly expand their data capabilities and market reach through collaborative efforts.

Medical Imaging

- Computer vision companies need data on which to train AI models for X-rays analysis.
- Medical facilities need AI models to improve diagnoses and collect data in their imaging devices.

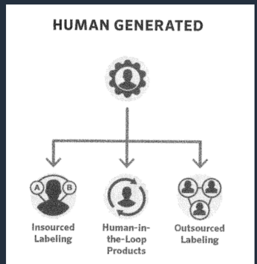
Conclusion and Future Directions

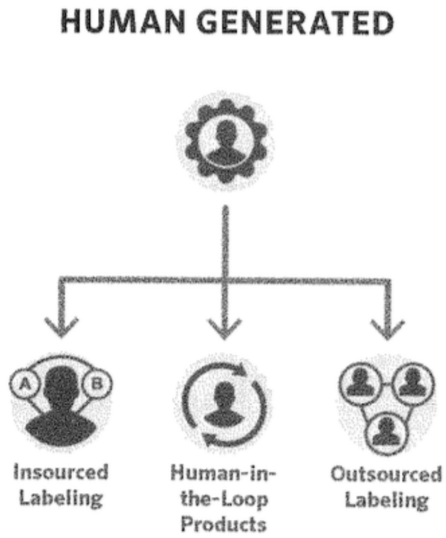
- Key strategies recap: Summarizing the innovative strategies AI-first companies use to acquire, negotiate, and leverage data for competitive advantage.
- The evolving data landscape: As the digital landscape continues to evolve, so too will the strategies for data acquisition and utilization, requiring companies to remain agile.
- Predictions for future trends: Insights into potential future trends in AI and data strategy, emphasizing the need for ongoing innovation and adaptation to maintain a competitive edge in the market.

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

Getting the data

Human Generated Data





Human Generated Data

Data Labeling

Importance of Data Labeling

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- Necessity of Labeled Data: Machine learning models, especially in recognition tasks, are **heavily reliant on labeled data**.
  - This data is crucial for the training and accuracy of algorithms.
  - Accessing a vast amount of labeled data for specific domains remains a significant challenge.
- Labeled data access and ownership of data to feed models can be the single hardest problem in starting a vertical AI company.
- The effectiveness of a labeling initiative is measured not just by the volume of data produced but also by its accuracy.
  - Labeled data must align well with the expert annotations, ensuring high-quality training sets for machine learning models.

Managing Data Labeling

- Data labeling is a measurable and manageable activity that can scale with proper management practices, if clear goals are set.
- Build a data labeling team combining experts and non-experts with necessary tools for labeling large volumes of data.
- Ensure clean data for efficient labeling and calculate ROI to optimize the labeling process for future scalability.
- For example, if goal is to get a model to an expert level of accuracy:
  - Start by having experts label each observation.
  - Then move to having a machine label some observations, with a non-expert correcting those labels.
  - Goal: have the machine plus the non-expert agree with the expert.
  - Over time, economic value of data labeling (ROI) can be quantified

$$ROI\ of\ labeling\ operation = \frac{Money\ saved\ through\ automation}{(Cost\ of\ each\ label\ \times\ \#labels)}$$

Best Practices for Building Data Labeling Operations

- Establishing a data labeling team is a strategic move for AI-first companies.
  - It allows for the rapid scaling of data annotation efforts.
- A blend of expert and nonexpert labelers can **optimize costs** while **maintaining quality**.
  - Decisions on the composition of the team should align with the complexity of the task.
- Effective management of the labeling team is essential, with roles varying from direct oversight to reporting to executive leadership:
  - Facilitate **quick feedback loops** and **quality control**.
- Incentivizing and managing a diverse team of labelers can significantly impact the **quality** of labeled data. The background of labelers can range from operations to specialized fields depending on the data.

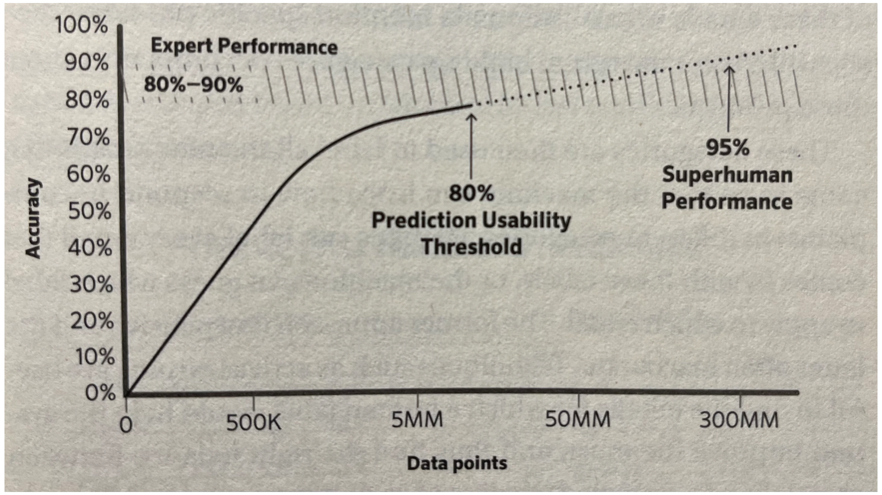
Managing Data Labeling and ROI

- Data labeling is a measurable and manageable activity that can scale with proper management practices.
  - Setting clear goals and ensuring quality are key to successful data labeling.
  - Sourcing Data for AI-first Businesses: Initiatives like crowdsourcing and surveys often fall short when rapid accumulation of domain-specific data is required.
  - Build a data labeling team combining experts and non-experts with necessary tools for labeling large volumes of data: this balance is critical for practical and cost-effective operations.
  - Ensure clean data for efficient labeling and calculate ROI to optimize the labeling process for future scalability.

Measuring Success and Cost-Effectiveness in Data Labeling

- **Success** in data labeling is measured by the improvement in AI classifier accuracy.
  - **Quantity** and **quality** of **labels** both play a critical role.
  - Even if individual labels are incorrect, a large volume of labeled data can enhance the model's accuracy. This is due to the AI's ability to learn from trends in the data.
- The cost of labeling should be weighed against the quality and utility of the data produced. Efficient labeling tools and processes can reduce expenses.
- The ratio of expert to non-expert labelers should be adjusted over time to balance costs and accuracy.
  - Tracking agreement rates with expert labels helps in maintaining quality standards.

Expert Labels and Performance



Source: Ash Fontana (2021) **"The AI-First Company"**

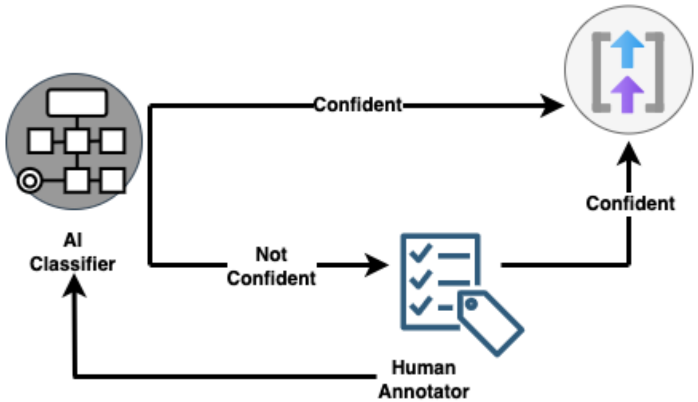
Key Messages

- As the number of data points increases, the accuracy of the ML model improves significantly, especially during the initial stages where data is scarce.
- The model reaches a critical point, labeled as the "Prediction Usability Threshold," at ~80% accuracy. This threshold indicates a level of accuracy that is considered usable for practical applications.
- Beyond this threshold, the rate of accuracy improvement slows down, entering a phase of diminishing returns. However, with continued addition of data points, the accuracy gradually inches towards "Expert Performance," which is denoted by the range of 80%-90%.
- Eventually, the model can achieve what is referred to as "Superhuman Performance," surpassing 95% accuracy. This suggests that the model's performance exceeds the capability of human experts in the task it is designed for.
- The graph shows an asymptotic trend approaching 100% accuracy, implying that there is an upper limit to the accuracy that can be achieved, regardless of how much more data is added.

The Active Learning Process

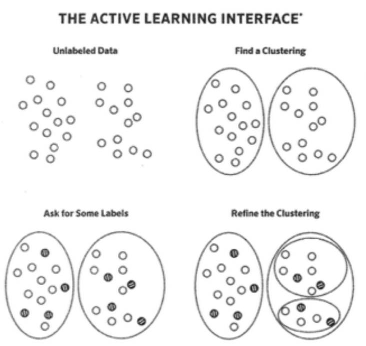
- Engage a human annotator to label data points that an AI classifier is unsure of.
- Active learning: a ML technique where the algorithm selectively queries a human or another source to label data points with uncertain predictions.
- The goal is to improve the learning accuracy with fewer labeled instances, as the algorithm focuses on instances where it is least confident and therefore can learn the most from the additional information.
  - By **iteratively requesting labels for carefully chosen data points** and **focusing on uncertain data**, active learning ensures the AI model is trained on the most informative examples:, an active learning system efficiently improves the model's performance over time by incorporating human insights where they are most needed.
- Particularly effective for tasks like categorizing customer feedback, where nuances in text can be challenging for AI to interpret alone.
- The active learning cycle is critical for models where manual labeling is impractical due to the sheer volume of data.

Active Learning Process



The Active Learning Interface

- The interface displays clusters of unlabeled data and guides labelers to refine the AI's understanding.
  - Labelers enhance the model by teaching it to recognize similarities in data.
- Through clustering and labeling, the system learns to generalize from specific examples.
  - This **reduces the overall effort** required for model training.



How Active Learning helps?

- Labeling often requires engineers to clean data before applying the labels. For example:
  - it is very hard for a machine to learn patterns across the text in millions of customer service emails. In this case, an engineer may use NLP to locate the segments of these emails where customers mention specific products, then she will cluster the text to build up categories of complaints about those products.
- These categories are then used to label all the new emails that come in so that the machine can learn how to respond to complaints in different categories.
- Humans can label every email that comes in with these labels, or the machine can guess which label to apply to which email:
  - The former approach is expensive and
  - the latter often inaccurate.
- Active learning can help in finding the emails for which a human label would help the system improve the most, and thus find the right balance between manual and automated labeling of each incremental email.

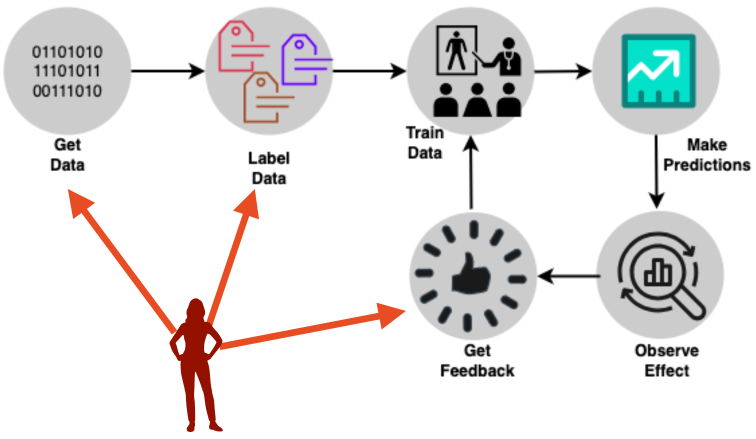
Human Generated Data

Human-in-the-Loop (HITL)

Human-in-the-Loop Systems

- Integrate human input to generate outcomes, not necessarily involving active or interactive learning.
  - These systems form the basis of many machine learning operations by involving humans for tasks like labeling data.
- Typically involve creation of new data, labeling, and providing feedback, which may employ binary scoring or a scalar score to assess the AI's output.

Human-in-the-Loop Systems



Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

Outsourcing in Data Labeling

- Outsourcing data labeling by integrating it with an IML system can accelerate the process and improve results by involving expert human input.
  - The **decision** to outsource **depends on**: expert availability, cost considerations, and the potential for increased automation over time.
- The primary reason for outsourcing is to obtain a large volume of labels, with the understanding that even experienced labelers will incorrectly label objects some of the time.
  - The aggregate correct labels, however, dilute the impact of mistakes, making the process mathematically efficient.

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

Crowdsourced Labeling and Machine-Generated Data

- Crowdsourced labeling involves various methods such as online searches, calling people, and completing other tasks that can be condensed into short labeling tasks.
  - Can be effective for data collection and cleaning, as well as deduplication and correction of data.
- Machine-generated data provides a way to supplement human-generated data, offering consistent, quick outputs that can complement or replace the need for human-labeled data.

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

Getting the data

Machine-Generated Data

Agent-Based Models & Synthetic Data

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- ABMs simulate the behavior of agents within a set of incentives and environmental constraints to predict outcomes.
  - ABMs, which draw from fields like game theory and economics, are used for complex simulation tasks such as forecasting and policy assessment.
- Synthetic data is created by setting rules that generate data points, which can be based on learning from existing datasets or entirely new rules for creation.
  - This method maintains the structure and dependencies of the original data while allowing for scalable, cost-effective data production.

Synthetic data generators

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- Labeling scalability can be challenging when dealing with varied forms of a single object due to the myriad of variations that must be accounted for. Synthetic data generators can mitigate these challenges by creating numerous examples at a low cost.
- Flexibility in labeling is important as objects often need to be recognized from multiple perspectives. Synthetic data generators can provide a diversity of examples that would be impractical to collect in the real world.

Synthetic Data

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

- Accessibility to certain objects for labeling can be limited, making synthetic data generators a valuable tool for creating necessary training data. These generators can replicate objects in various environments, enhancing the breadth of data available for model training.
- Labeling cost and probability are significant factors, with some objects or events being rare or expensive to capture in the real world. Synthetic data generators can create these rare occurrences, providing valuable data points for model training.

Master Programs in Artificial Intelligence for Careers in EU (MAI4CAREU)

Getting the data

Consumer and Public Data

Consumer Data

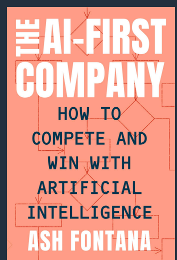
- Customers vs Consumers:
  - Customers: pay for AI-First products
  - Consumers: take the output of the product
- Token-based incentives
  - Blockchains & crypto-tokens to reward contributing data to data networks
- Consumer Apps: Google, FB
- Sensor Networks

Public Data

- Crawling
- Consulting and Competitions (Kaggle)
- Data-driven Media
- Governments
- Buying Data & Brokerage

Module 3: AI Companies

Section 5: AI-First Teams



Key observations

- AI-First companies need a diverse group of people to manage different technologies
  - Competencies required: **data infrastructure engineer**, **data engineer**, **data scientist**, **data analyst**, **ML engineer**, and **ML researcher**.
- AI-First teams need people with different skill sets.
  - Specialization in **databases**, **statistics**, **mathematics**, **physics**, **econometrics**, and **economics** all help to understand the fundamental principles of AI and run it on modern computing infrastructure.
- AI-First teams need AI-First tools.
  - Notebooks, frameworks, cloud services, pre-trained models, data labeling, data prep, and visualization tools created for data scientists and machine learners help these teams get the job done.
- Engineers and data scientists need a different type of manager.
  - Data scientists ask questions like researchers. They **deliver results** as **numbers** in spreadsheets, **graphs** in presentations, **models** in code, or **discussions** in person.
  - They don't get requirements from business users, so they **require regular meetings**. They may need a larger budget and ad hoc approval of purchases--with commensurate oversight.
- AI-First companies put AI everywhere. AI-First companies **distribute AI talent across their organization**, have **AIs working behind all of their products**, use **AI-enabled computing infrastructure**, decentralize data science, and have an executive team that sets data strategy for the whole company.

Module 3: AI Companies

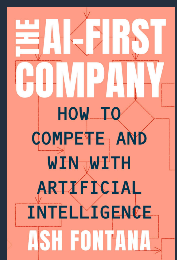
Section 6: Making the Models

Key Insights

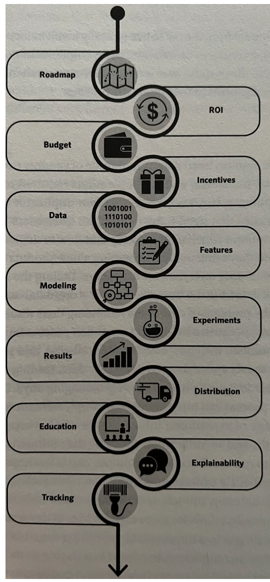
- Pick a machine learning method based on available data:
  - Supervised learning needs training and feedback data, whereas unsupervised ML just requires lots of data.
- Some models need an objective:
  - Reinforcement-learned models need objectives.
  - Other forms of ML generally do not, and they will even surface information without objectives.
- Learn to learn.
  - Some AIs generate data about how they learn, accumulating a valuable asset to leverage when leverage when searching for solutions to new problems.

Module 3: AI Companies

Section 7: Managing the Models

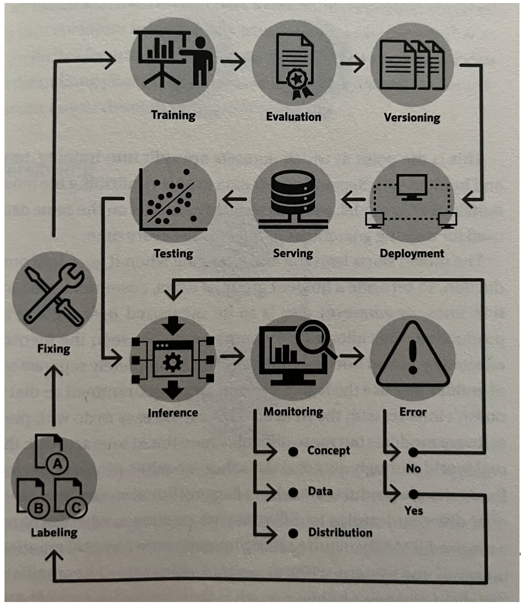


Steps to Acceptance



Source: Ash Fontana (2021)
"The AI-First Company"

The ML Management Loop



Source: Ash Fontana (2021) "The AI-First Company"

Managing Models: Playbook

- **Keep models close to reality:** Intelligent systems are powerful because they constantly adapt, evolve, and spawn new data, but be aware that they can run way from reality. Constantly getting feedback keeps models in check.
- **Strike a balance.** The ideal model management system allows for decentralized experiments, rigorous testing of models, and monitoring in real-world data.
- **Don't drown in a data lake**. Tightly specify the necessary data, lead the teams responsible for accessing it, and actively manage data-related vendor selection to quickly implement AI-First products.
- **Set security parameters for every dataset**. Experimentation, testing, and production require different levels of security. Customers in regulated industries may need to run models on their premises without ever touching their data.
- **Outsource implementations that involve sensors.** Implementing and managing sensors involve significant logistics, industrial design, IT, and environmental challenges. Outsource this to a systems integrator that works with the sensor manufacturer.
- **Communication ensures a smooth implementation.** Data validators and engineers clear up inconsistencies in customers' data. Data translators communicate early results. Both can ensure a smooth implementation and ultimate acceptance of AI-First products.
- **Involve customers in training models.** Demonstrating the output at each training step can yield commonsense feedbackand ideas for new features.

Managing Models: Playbook

- **Customers want models that are accurate in the real world, not just in the lab.** Quickly incorporate real-world data and make models automatically learn from that data.
- **Acceptance of AI is a surmountable challenge.** Get early and broad distribution, make sure the AI works, lower time to value, create a realistic road map, promote engagement with experiments, provide executive education, retrain regularly, build features fast, augment (don't automate), embed explainability, incentivize the right people, ensure accountability, add buffers to budgets, measure usage, set business unit-level ROI, and focus on delivering revenue (not reducing costs).
- **Model management is not code management.** Model management needs to manage both data and code, rather than just code.
- **Version code and manage metadata before trying to version data.** Versioning data is hard and expensive. Focus on versioning model code first.
- **The goal of versioning is reproducibility.** Reproducibility is particularly important in academia and regulated industries. Add software packages, coding tools, and other dependencies that may enhance someone else's ability to replicate results to versioning systems.
- **Split up training, test, and production data.** Testing on training data always gets a perfect score. Keeping a holdout set keeps models honest.

Managing Models: Playbook

- **Deploy predictions in the form that customers prefer.** That could be in a report, spreadsheet, dashboard, or template, integrated into another software product, as a stand-alone application, through an API, or in a piece of hardware.
- **Test.** Use statistical measures for data quality, accuracy measures for model quality, and a correlation matrix for relevance. Don't forget to check that the code runs alongside existing software.
- **Keep an eye on drift.** Whether it's the concept or data, don't let predictions get too removed from reality.
- **Deal with bias by setting hard constraints.** Restrict what the model can output, control access, limit feedback data, and make acceptable uses of the predictions clear to all stakeholders.
- **Give models the data they need.** Constantly monitor data for missing sources, values, and labels. Proactively perturb data to catch quality issues before they break something.
- **The world is always changing, so models will too.** Retrain, refit, reweight, redo, and redeploy. Automate later.