2022

Introduction & overview of Ethics in Machine Learning

Christina Hitrova
Digital Ethics and Compliance Consultant, PwC CZ

CorrelAid ML Workshop Series
9 March 2022

Christina Hitrova

Contact me on:
Slack
Email
christina.hitrova@gmail.com
christina.hitrova@pwc.com

Education

International and European Law Bachelor and 2 Master's from:

- Groningen, Netherlands
- Leuven, Belgium
- Zurich, Switzerland

Career

TRILATERAL RESEARCH

Research analyst - work on data protection and privacy-by-design

The Alan Turing Institute

Senior research assistant - ethical digital innovation in the UK public sector

Research associate - values in the creation and use of science and technology

pwc

Consultant - digital ethics and legal compliance of digital transformation and AI projects

What do you think “technology for good” means in practice?

Do you think you can create “technology for good” by yourself?

What do you think you would need for “technology for good”?

Are good intentions enough for “technology for good”?

So you want to use technology and data for good.

When the Ukraine invasion started, many people had the same idea

Why do you think this might not be a good idea?

Source: Alis Dunn's Twitter thread
Source 2: Andrew Therriault thread

Alix Dunn @alixtrot · Feb 28

Replying to @alixtrot

"Cool ideas" coming from a place of limited contextual knowledge can be dangerous.

1 48 419

Alix Dunn @alixtrot · Feb 28

Spinning up tech that is unmoored from institutions and infrastructure is superficial and unsustainable.

1 35 377

Alix Dunn @alixtrot · Feb 28

In hard times, technology is intoxicating. It can feel like a shortcut to impact in a moment of impossible hardship.

It's not.

If you are getting high on your own supply, take a beat and think critically.

3 37 339

A more extreme view….

Alix Dunn @alixtrot · Feb 28

Replying to @alixtrot

"Cool ideas" coming from a place of limited
dangerous.

1 48

Alix Dunn @alixtrot · Feb 28

Spinning up tech that is unmoored from ins
superficial and unsustainable.

1 35

Alix Dunn @alixtrot · Feb 28

In hard times, technology is intoxicating. It
in a moment of impossible hardship.

It's not.

If you are getting high on your own supply,

3 37

Andrew Therriault @therriaultphd · Feb 27

Whether it's politics, public health, humanitarian relief, or war, there are thousands and thousands of professionals who spend their careers building specialized skills. The thing they lack isn't knowledge or skill, it's almost always the resources to act. (3/10)

3 204 1,968

Andrew Therriault @therriaultphd · Feb 27

So amateur volunteers, however well-intentioned and skilled in their own fields, are often solving the wrong problem. And even if they can deliver something valuable, handing it off and walking away just creates technical debt for the people you think you're helping. (4/10)

4 113 1,634

Andrew Therriault @therriaultphd · Feb 27

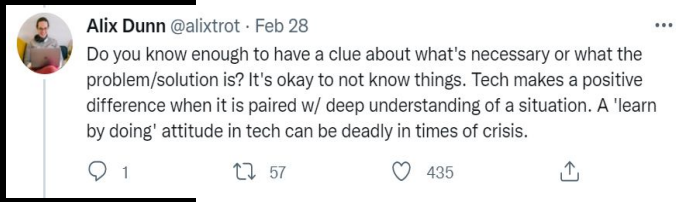
This is why I hate hackathons with a passion - they let volunteers do the fun part of the work and feel good about themselves, while sticking someone else with the burden of trying to implement and maintain something that's only half-done at best. (5/10)

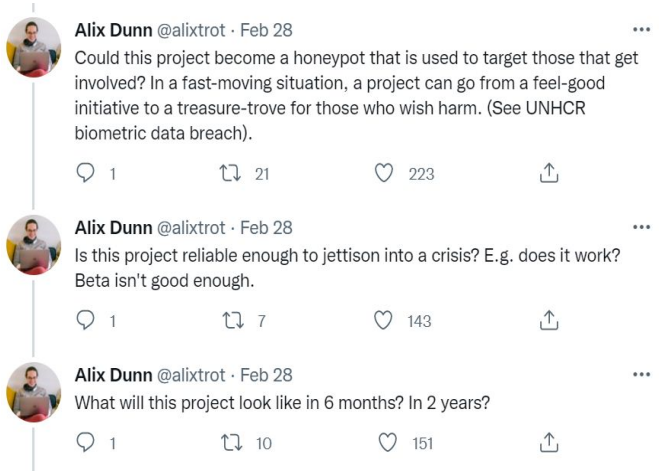
15 210 2,370

Don’t be discouraged if you still want to help on important issues!

What do Twitter experts recommend you do?

1. Make money with your tech skills and donate to actual experts
2. Think about your plan, skills, context critically
3. Consider the limitations of your solution
4. Consider risks of abuse or misuse of your work
5. Plan how the project will be maintained long term





This is how a typical CorrelAid project might look like

Partners might be:

- Working with vulnerable people (and having their data)
- Working on important issues
- Don't have technical skills
- Don't have enough resources

Volunteers might lack knowledge about:

- The subject matter, the stakeholders
- Where the data originated from, the reality on the ground

Volunteers make a lot of decisions around how exactly the technical solution would function and look like in practice.

Volunteers introduce the solution to the partner with relevant information and instructions on how to use and maintain it.

There is a period of testing the tool in practice.

Volunteers leave

The partner now has:

- A tool to inform decisions, possibly affecting vulnerable people
- A tool that needs to be maintained, updated

How to design and use science and technology in a way that it achieves our goals to improve the world for all?

Is this a difficult question to answer? Why (not)?

Science and technology are special:

An authority to support decision-making (evidence-based, data-driven..)
Aura of reliability despite limitations of scientific and technological methods
Beyond criticism by lay people
Imperfect: Assumes the lab > nature and reality, scientists are objective, everything can be measured

How to design and use science and technology in a way that it achieves our goals to improve the world for all?

Is this a difficult question to answer? Why (not)?

ML can have high stakes

- ML is imperfect
- It brings an “aura of authority” and seems autonomous, like magic
- Requires regular review and maintenance
- CorrelAid’s partners might be working with vulnerable people or on difficult topics without a simple solution
- ML tools can affect the welfare of people, e.g.:
  - Automatically classify and categorise people
  - Predict “risk” or “need” scores to prioritise them
  - Target resources

The point of today is **to give you a toolkit to think and act critically.** This can improve your ML development work or how you plan and design future projects

Frameworks to help: Responsible Innovation and The CorrelAid Code of Conduct +

A basic risk assessment and management framework

What is special about machine learning in particular?

How to act with humility and in service of others when you are not an expert but where you are creating tools that bring authority

You can use this exercise to put everything else we cover today in your project workflow

Have an overview of the risks of ML specifically and some tools to counter them
Most likely won't have enough time for everything but you will have the reading!

So how can you innovate responsibly when:

- You aren’t an expert in the area
- Your work will affect what decisions others take and how they understand the world around them
- We are talking about socially impactful fields
- You won’t be around for long to maintain whatever you create

Responsible innovation manages the limitations of science and technology by prescribing humility and inclusion

Anticipation

Anticipate the impact of innovation and the changing circumstances.

Ask “what if” questions to shape the innovation to be resilient.

Reflexivity

Be mindful and aware of your biases and knowledge limitations that inform your actions and assumptions.

Always have justifiable reasons for your actions and reasoning.

Inclusion

Include diverse members of the public, affected stakeholders, and domain experts in your work.

Consult them about the context, ask them about their expertise, their experience, their views of your work.

Responsiveness

Organise your work in a way that allows you to respond and adapt to what you learn from stakeholders or changing circumstances.

You might have to reconsider the problem you try to solve, your use of technology, or your design assumptions.

Based on: Stilgoe et al (2013) - Developing a framework for responsible innovation

CorrelAid Code of Conduct highlights the ideal of data science with humility and in service of others (selected parts)

Empowering others for a sustainable impact

Pragmatism - We create practical and sustainable solutions that have real **value to our partners**. We do so by striving to use technologies that **can be implemented, used, maintained and updated by our partners**, even if that means that we have to integrate less advanced tools. (...)

Transparency - We are **transparent in how we derive conclusions** and try to **explain methods and technologies** used to our partners to empower them on their individual data journey. We also **present differing views and are honest about the limitations** of our knowledge and work

Knowledge sharing - (...) We believe in **sharing our work and knowledge with a broad community.** We do so by making **code and resources publicly available** whenever possible and attending conferences and talks to share our expertise.

CorrelAid Code of Conduct highlights the ideal of data science with humility and in service of others (selected parts)

Empowering others for a sustainable impact

Pragmatism - We create practical and sustainable solutions that have real **value to our partners**. We do so by striving to use technologies that **can be implemented, used, maintained and updated by our partners**, even if that means that we have to integrate less advanced tools. (...)

Transparency -
conclusions and
our partners to er
also **present diff**
of our knowledge

NB: Sharing knowledge means also sharing information about your methodology and its limitations ! Not all solutions are well-suited for all contexts.

Knowledge sha (...) **knowledge with a broad community.** We do so by making **code and resources publicly available** whenever possible and attending conferences and talks to share our expertise.

CorrelAid Code of Conduct highlights the ideal of data science with humility and in service of others (selected parts)

Empowering others for a sustainable impact

Pragmatism - We create practical and sustainable solutions that have real **value to our partners**. We do so by striving to use technologies that **can be implemented, used, maintained and updated by our partners**, even if that means that we have to integrate less advanced tools. (...)

Transparency - We are **transparent in how we derive conclusions** and try to **explain methods and technologies** used to our partners to empower them on their individual data journey. We also **present differing views and are honest about the limitations** of our knowledge and work

Knowledge sharing - (...) We believe in **sharing our work and knowledge with a broad community.** We do so by making **code and resources publicly available** whenever possible and attending conferences and talks to share our expertise.

High standards of working

Professionalism - We dedicate ourselves to have **high working standards** and ensuring that insights derived from our analysis are profound. **Data protection and security** are our key concerns. We especially make sure that confidential data is treated carefully and **deleted after project closure**.

Self-actualization - (...), we do not prefer specific **coding languages or technologies** but rather encourage our partners and members to **choose the right path** for themselves (...)

CorrelAid Code of Conduct highlights the ideal of data science with humility and in service of others (selected parts)

Empowering others for a sustainable impact

Pragmatism - We create practical and sustainable solutions that have real **value to our partners**. We do so by striving to use technologies that **can be implemented, used, maintained and updated by our partners**, even if that means that we have to integrate less advanced tools. (...)

Transparency - We are **transparent in how we derive conclusions** and try to **explain methods and technologies** used to our partners to empower them on their individual data journey. We also **present differing views and are honest about the limitations** of our knowledge and work

Knowledge sharing - (...) We believe in **sharing our work and knowledge with a broad community.** We do so by making **code and resources publicly available** whenever possible and attending conferences and talks to share our expertise.

High standards of working

Professionalism - We dedicate ourselves to have **high working standards** and ensuring that insights derived from our analysis are profound. **Data protection and security** are our key concerns. We especially make sure that confidential data is treated carefully and **deleted after project closure**.

Self-actualization - (...), we do not prefer specific **coding languages or technologies** but rather encourage our partners and members to **choose the right path** for themselves (...)

Diversity

Appreciation - We **respect and value other people and their views,** especially if they are different from ours. We do so by **learning from each other**. (...)

Diversity - We embrace **diversity** in all its forms and take care of fostering **tolerance** and **acceptance** inside our community and beyond.

Think about the responsible innovation framework and the CorrelAid Code of Conduct. Discuss

Responsible Innovation framework

Anticipation
Reflexibility
Inclusiveness
Responsiveness

CorrelAid Code of Conduct

Pragmatism
Transparency
Knowledge sharing
Professionalism & Self-actualization
Appreciation & Diversity

Do you understand the logic of these frameworks?

Do you understand what you could do in practice for each of them?

Exercise: Conceptboard link

19

How can you put this in practice in an organised manner?

Carry out a risk assessment and management of your proposed technical solution. Plan how to mitigate the risks you identified in order to improve the way you design the technical solution

Carry out a risk assessment and management of your proposed technical solution. Plan how to mitigate the risks you identified in order to improve the way you design the technical solution

A risk assessment and risk management process can be used for anything.

You can use it for any data project with CorrelAid.

Carry out a risk assessment and management of your proposed technical solution. Plan how to mitigate the risks you identified in order to improve the way you design the technical solution

Step 1: Identify what categories of risk to cover

- Ethical risks - fairness, explainability, transparency, stability, accuracy….
- Technical risks - Are there risks that the technology won't function as intended?
- Use risks - Could the tech be misused or abused by people?

Step 2: Collaborate in your team to identify risks and impacts

- Map out the technical tool you are planning (+ data!)
- What risks can you identify?
- What is the likelihood of risks materialising? What is the potential impact? Who would be affected and how?

Step 3: Identify the sources of the risks

- What are the origins of the risk? What determines the likelihood or the impact of a risk?

Step 4: Make a plan to manage the risks

Can you:

- prevent the risk from existing in the 1st place ?
- minimise the risk?
- shift the risk on someone else, e.g. train a user to make a final decision down the line?

Step 5: Be open and engage stakeholders to improve your risk assessment and mitigation

- Who can you talk to to learn about the context, the intended users, desirable impact?
- Publish your risk assessment or at least share it with the NGO partner. This will clarify the limitations of your work and what they need to be careful of

Carry out a risk assessment and management of your proposed technical solution. Plan how to mitigate the risks you identified in order to improve the way you design the technical solution

This is a cross-cutting action. You should engage and learn from stakeholders and domain experts throughout steps 1-4!

Step 1: Identify what categories of risk to cover

- Ethical risks - fairness, explainability, transparency, stability, accuracy….
- Technical risks - Are there risks that the technology won't function as intended?
- Use risks - Could the tech be misused or abused by people?

Step 2: Collaborate in your team to identify risks and impacts

- Map out the technical tool you are planning (+ data!)
- What risks can you identify?
- What is the likelihood of risks materialising? What is the potential impact? Who would be affected and how?

Step 3: Identify the sources of the risks

- What are the origins of the risk? What determines the likelihood or the impact of a risk?

Step 5: Be open and engage stakeholders to improve your risk assessment and mitigation

- Who can you talk to to learn about the context, the intended users, desirable impact?
- Publish your risk assessment or at least share it with the NGO partner. This will clarify the limitations of your work and what they need to be careful of

Let’s discuss: At what point in time would you start the risk assessment process? Do you think if you do it once, it’s done?

Risk assessment steps:

1. Identify categories of risk to cover
2. Collaborate with team to identify risks in the technical solution you propose
3. Identify the sources of the risks
4. Make a plan to manage the risks
5. Be open and engage with stakeholders

Let’s discuss: At what point in time would you start the risk assessment process? Do you think if you do it once, it’s done?

Risk assessment steps:

1. Identify categories of risk to cover
2. Collaborate with team to identify risks in the technical solution you propose
3. Identify the sources of the risks
4. Make a plan to manage the risks
5. Be open and engage with stakeholders

! A risk assessment has to be repeated / adapted whenever your solution changes. Otherwise it isn’t an assessment of your current plan.

Take a break! 10 minutes

Okay, so you want to identify and manage risks to be responsible.

But what are actually the risks that can arise in the context of ML? What what can we do about it?

What is special about machine learning?

Do you have a positive, negative or a neutral view of ML in general?

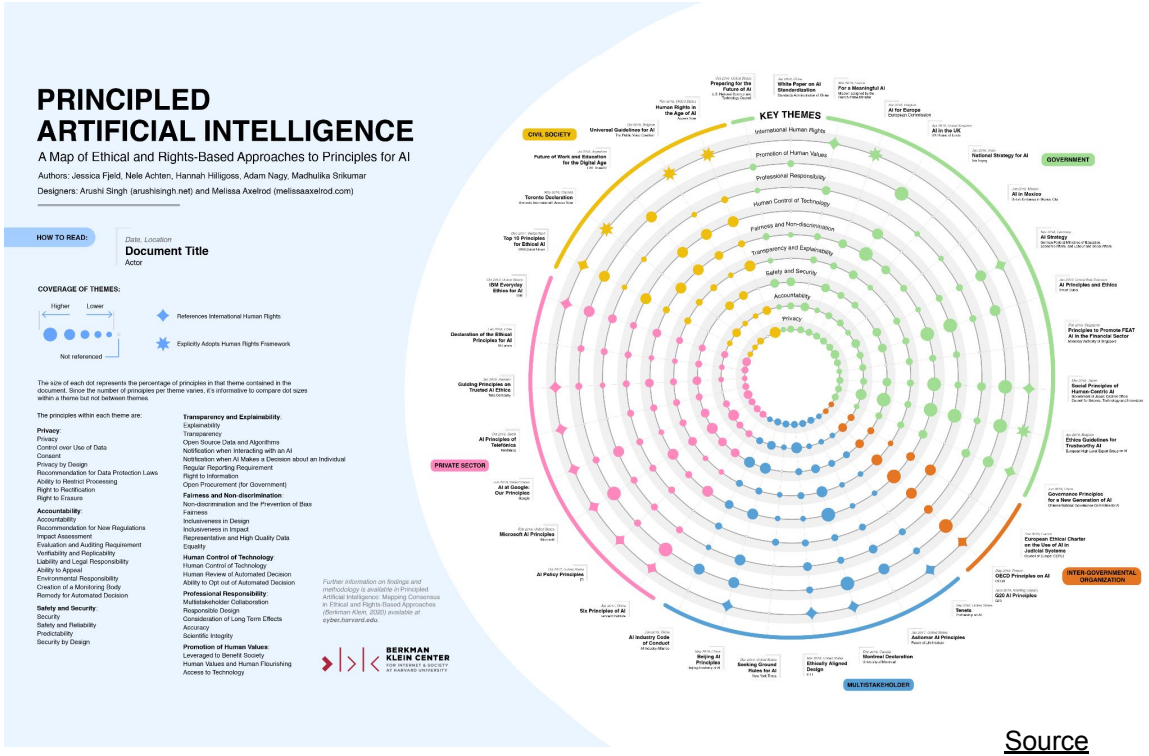
What do you see as the opportunities and benefits of ML?

Have you heard of risks or limitations of ML? Which ones?

Do you think the benefits of ML outweigh its risks? Does that depend? On what?

There are many frameworks and guidelines on AI ethics but there are a number of common issues we keep hearing about…

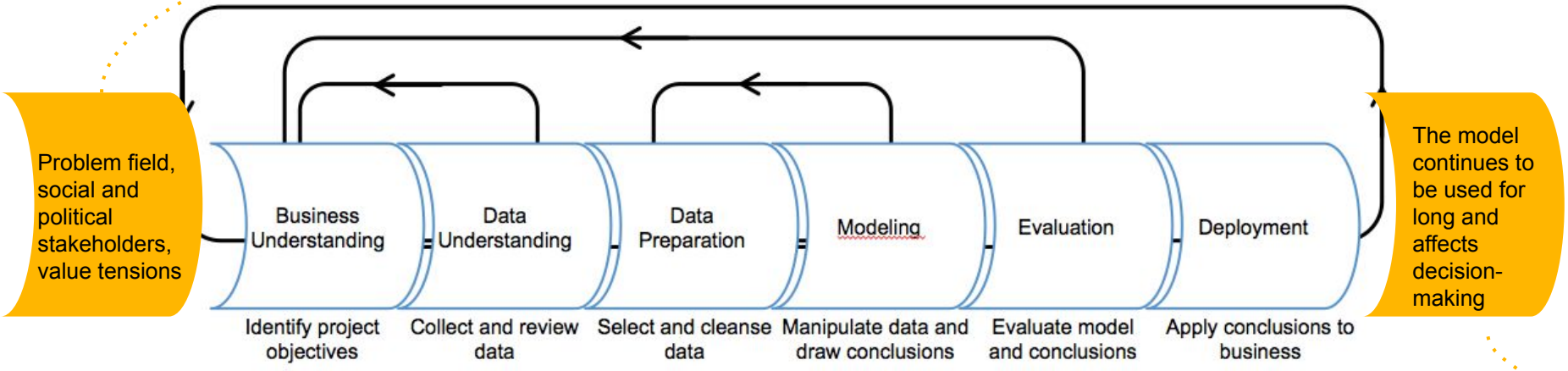
- International human rights
- Promotion of human values
- Professional responsibility
- Human control of technology
- Fairness and non-discrimination
- Transparency and explainability
- Safety and security
- Accountability
- Privacy



Source

What does this mean in practice? How can you take well-reasoned and justifiable decision at every step?

We might often deal with very complex questions that hold many values, affect many people, and could have many competing proposed solutions. We need to understand the context to be able to understand and evaluate the partner's needs, the data, and the model.



Charities and NGOs could easily lack the technical expertise to maintain and test ML models to ensure optimal performance long-term.

Knowledge Discovery in Database (1996)

Model drift

User under-reliance on AI

Explainability and interpretability

No 100% accuracy

Black box AI

Poor data quality

Transparency and auditability of design process

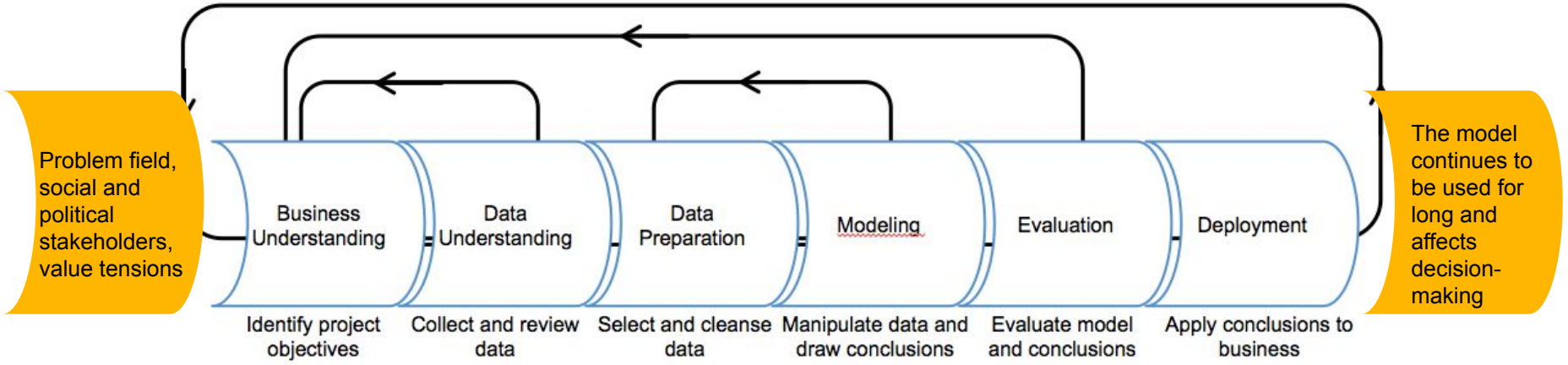
User over-reliance on AI

Feedback loops

Technology solutionism

Model bias

Data bias



Conceptboard - exercise link

Knowledge Discovery in Database (1996)

In the next slides we have (selected) AI risks and limitations and their proposed solutions

We probably won’t cover it all. Please use the slides for reference to read at home

What is special about machine learning in particular? What limitations does ML present? What risks can we anticipate?

1

ML has limitations to its capabilities *even* when it is perfect. That requires us to be careful about what we use ML for.

2

ML is rarely perfect due to human design decisions and imperfect data. This requires quality assessments, documentation, engagement with stakeholders….

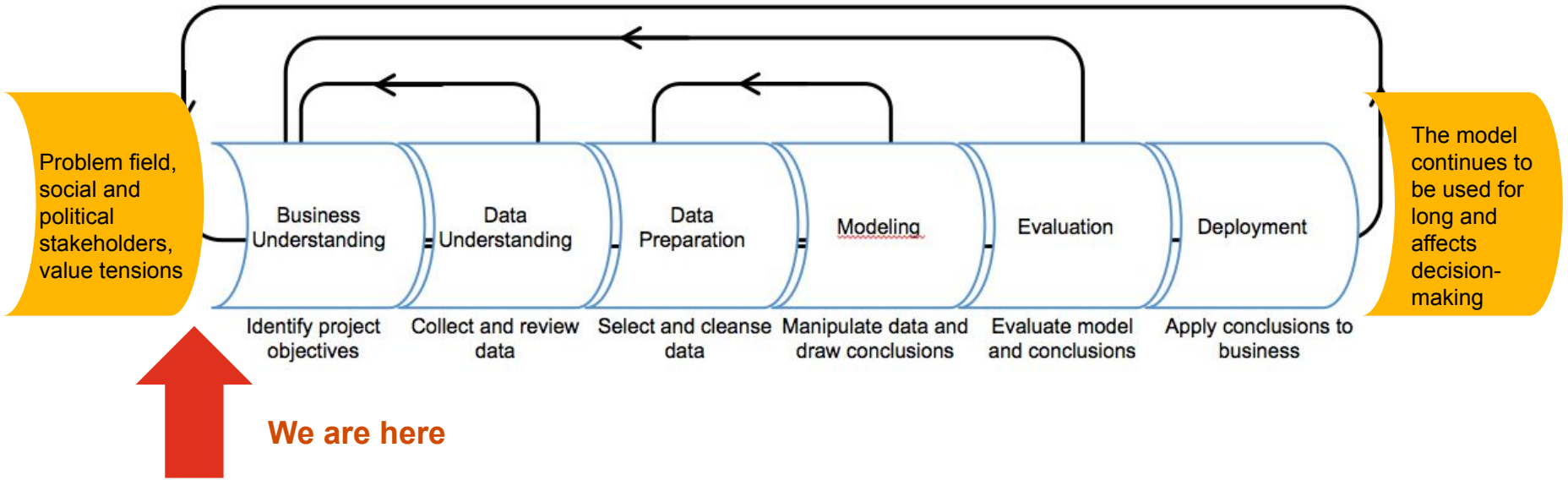
3

The way ML interacts with the real world (*sans humans)* requires continuous review and maintenance of the tools, as well as options for human oversight.

4

The way humans interact with ML requires special instructions, user training and good UI design.

1. ML has limitations to its capabilities *even* when it is perfect



Knowledge Discovery in Database (1996)

1. ML has limitations to its capabilities *even* when it is perfect

Be aware of technology solutionism. Be careful about using ML to treat symptoms rather than underlying issues.

Example:

A public authorities wants to use an ML model to screen calls by people to a child protection services telephone line and give them risk scores in order to help people prioritise them. In theory, this should help target resources better.

The problem here is that there isn't sufficient resource to respond to all calls. Prioritising will still leave some children in need without help.

Further reading: Allegheny Family Screening Tool
Further reading: AI solutionism

1. ML has limitations to its capabilities *even* when it is perfect

Not all problems can be expressed in a neat and quantified way, required by ML models.

ML needs some things to function well, e.g.

- High quality and sufficient amounts of data
- A well-defined target variable that is concrete, measurable and actually represents what you are interested in

\+ **ML solves future problems by using past examples. This can limit creative solutions and negatively affect outlier cases.**

Example:

Content moderation of social media currently can be a difficult and distressing job for humans to do.

Facebook wants to use ML tools to automate and speed up content moderation. ML models should be able to recognise verbal abuse, harassment, offensive words, or hate speech.

It's probably impossible to do this with high accuracy. Can you think of reasons why?

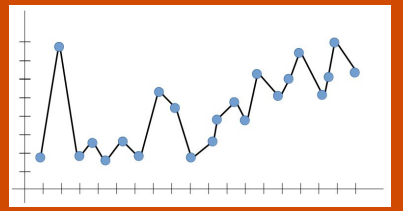
1. ML has limitations to its capabilities *even* when it is perfect

Complete accuracy is impossible and a degree of randomness (stochastic or probabilistic nature) is inherent in any ML model

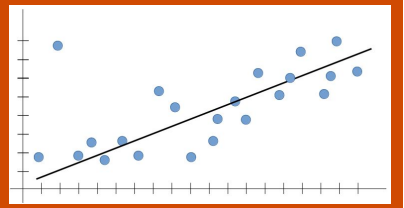
- The data used to train an ML will in any case be an incomplete sample of the general population (the real world). **To be able to generalise to the entire population, models need to have stochastic nature.**
- Randomness is a feature of the process of learning that allows the model now to “get stuck”. Randomness could also be used in optimise ML models. **This means that the same model trained on the same data will be slightly different every time.**

Example:

Model overfitting training data will have 100% accuracy during testing but it won’t be generalisable (usable on new, unseen data)



Ideal fit would be generalisable to new data, but it won’t have 100% testing accuracy



Further reading: Embrace the randomness of ML
Source: What does stochastic mean in ML?
Source: Overfitting and underfitting in ML

1. ML has limitations to its capabilities *even* when it is perfect - WHAT TO DO

Scope your project carefully with partners. Be open about what they can expect. Manage expectations.

Choose problems that actually help learn about or solve the underlying issue.

Think of what is technically feasible given the requirements and limitations of ML and choose appropriate use cases and techniques.

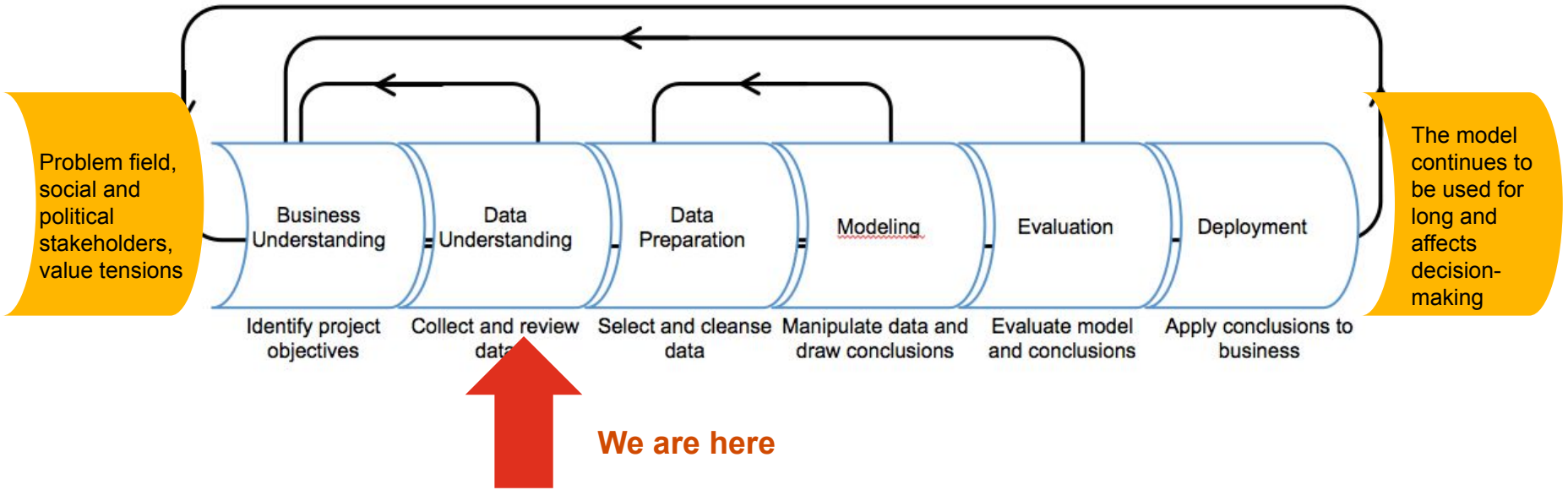
Discuss with the partner organisation what they can expect from the end solution to manage expectations

Be mindful of the probabilistic nature of ML. Think about how the ML will be used in practice and what is at stake.

Avoid use cases that focus on predictive analytics at an individual-level (e.g. this individual is 70% likely to start smoking), as opposed to population-level analytics (e.g. 20% of people will start smoking).

Test and record the performance of a model every time you train it. Exercise version control. Share the information with partners

2. ML is rarely perfect due to imperfect data



Knowledge Discovery in Database (1996)

2. ML is rarely perfect due to imperfect data

“Garbage in, garbage out”. The data that we have is usually imperfect**. The perfect data should be**:

- **Representative** - it should represent the populations you are modelling, it should not over- or under-represent certain groups
- **Relevant** - it should provide relevant and comprehensive information about the phenomenon you are modelling
- **Recent** - it should be up-to-date and still relevant enough to adequately represent reality on the ground (relevant for predictive analytics, less for descriptive)
- **Accurately measured** - is the data measured and recorded accurately and reliably? Does it oversimplify complex situations into a single data point?

Thanks to machine-learning algorithms, the robot apocalypse was short-lived.

2. ML is rarely perfect due to imperfect data - WHAT TO DO

Assess the quality of the data critically.
You can use this overview for data requirements to help. >>> Work with your team and the project partner to understand the context.

Record the data quality assessment. This can help you order your observations in one place and make a decision on what to do with the data. You can use your own document or the list in the paper Datasheets for datasets.

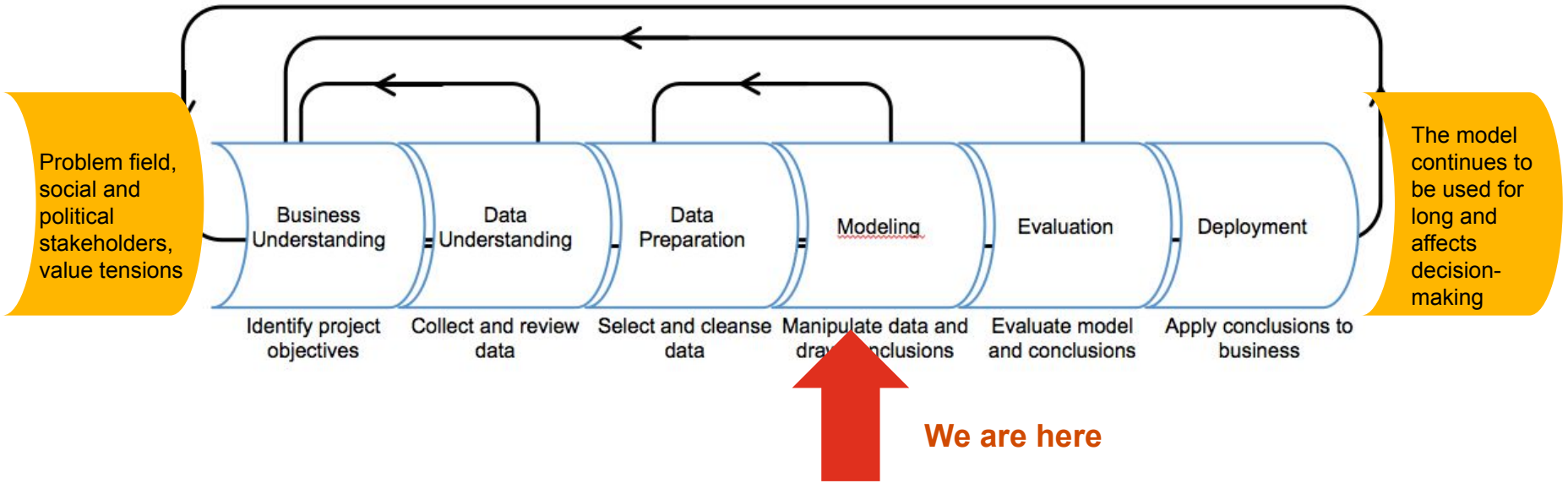
If you can, enhance the quality of your data. Reading

If you can't, adjust the model you plan on creating to fit the data quality you have.

Data requirements	
Representativeness	• Is the training data representative of the population under consideration? • Have you planned how to balance the dataset you will use, so that it appropriately and equitably reflects sub-populations? • Have you thoroughly considered risks of under- or over-representation in the dataset?
Relevance	• Are the chosen data sources relevant to and capable of providing a reasonably comprehensive and balanced view of the phenomenon to be modelled? • Where data to provide a reasonably comprehensive and balanced view of a phenomenon is lacking, have you considered how to amend the purpose of your model to appropriately utilise the data available?
Recency	• Is the data you plan to use an up-to-date reflection of the phenomenon and populations you are trying to model? • Have any large-scale reforms, policy changes, or changes in methods of data recording taken place that affect whether the data you want to use accurately portrays phenomena, populations, or related factors in an accurate and up-to-date manner? • Is the timeliness of the data you plan to use sensitive to small or minor shifts that may take place within neighbourhoods, cultures, or operational policies? If so, have you properly established that your use of such data meets the challenges of these shift sensitivities?
Measurement accuracy	• Are elements of subjective bias or human error potentially involved in any aspects of data collection across your dataset? If so, have you diligently established that such risks have been addressed and mitigated, so that your dataset is sufficiently sound and reliable? • Have appropriate methods for recording data been used? • What information has been lost in the data recording, how valuable is it, and what are the implications of not having it?

Source: Ethics Review of ML in Children's Social Care

2. ML is rarely perfect due to human design decisions



Knowledge Discovery in Database (1996)

2. ML is rarely perfect due to human design decisions

There are many design decisions that go into making an ML model, all of which may be imperfect.

- You could choose a **target variable** that does not actually represent what you are interested in.
- You could include **irrelevant data** in your calculations or fail to take into account **relevant data** that could perpetuate biases.
- Your decisions how to handle **data imperfections**, e.g. missing variables, could affect different demographic groups differently. If you decide to exclude all instances of data gaps you exclude those people from your analysis. If you infer the data values, you include them in an altered way. Which is better?

Millions of black people affected by racial bias in health-care algorithms

An algorithm used to allocate health care to patients based on need was **consistently biased against black people.**

This was because the creators of the model chose the wrong target variable. They **used expected healthcare costs as a proxy for healthcare needs.**

But they didn't consider that:

- They are not the same
- Historically, black population in the US had less access to healthcare and, therefore, lower costs.

Ledford, Nature 574, 608-609 (2019)

2. ML is rarely perfect due to human design decisions - WHAT TO DO

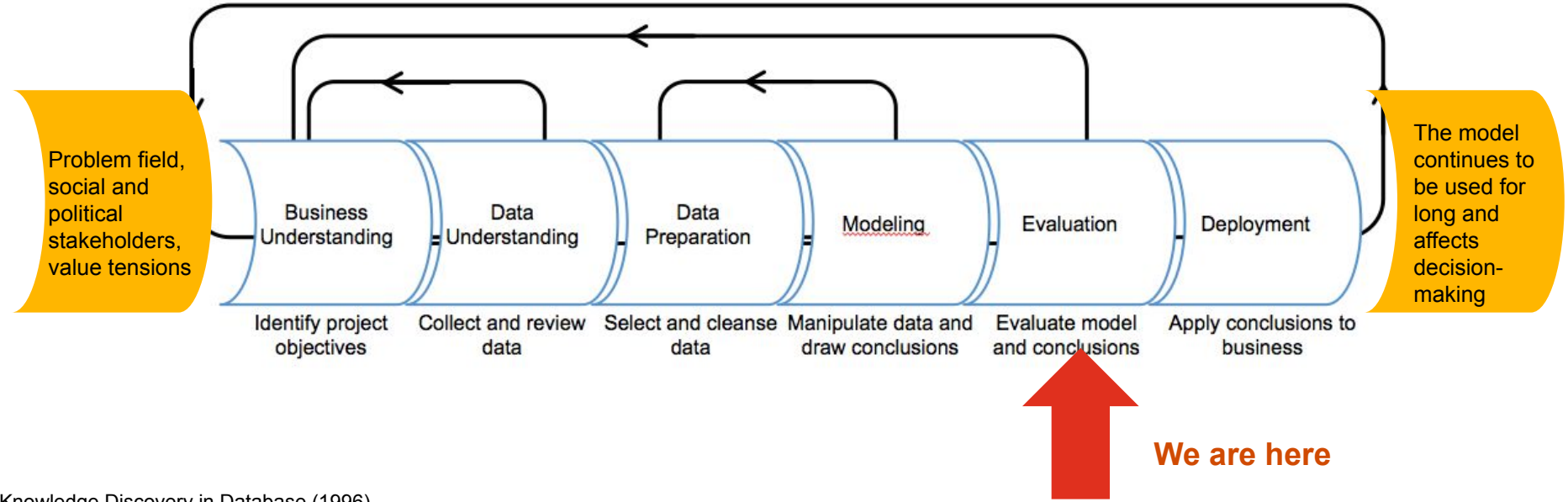
Work closely with your partner organisation and/or domain experts to get the knowledge you lack and make design decisions in a justifiable and reasoned way.

Think about the impact design choices have on the model and on the real world.

You could provide alternative options and engage with partners to learn about their views with regard to:

- Choosing a target variable
- Identifying relevant and irrelevant data
- Handling data quality issues

2. ML is rarely perfect and that could lead to unfairness, bias and discrimination



Knowledge Discovery in Database (1996)

NB: Further reading on the many ways we can evaluate ML models

2. ML is rarely perfect and that could lead to unfairness, bias and discrimination

Issues of data with poor quality, especially biased representativeness, can lead to unfairness and discrimination. Your model will 'learn' and perpetuate whatever patterns are in your dataset:

- Women getting rejected for loans more than men
- Black people being arrested more than white
- Men being hired for tech positions more than women

! NB **non-sensitive variables could act as proxy variables for sensitive characteristics**, e.g. wearing a burka is not a sensitive data, but it does correlate most likely with being Muslim. Religion is sensitive data.

Amazon scraps secret AI recruiting tool that showed bias against women

Apple's 'sexist' credit card investigated by US regulator

Is facial recognition too biased to be let loose?

The technology is improving – but the bigger issue is how it's used.

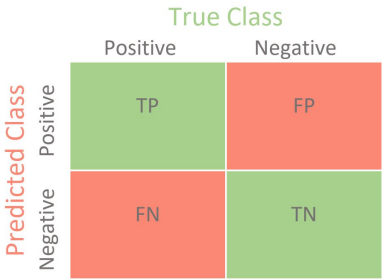
Further reading on ML fairness

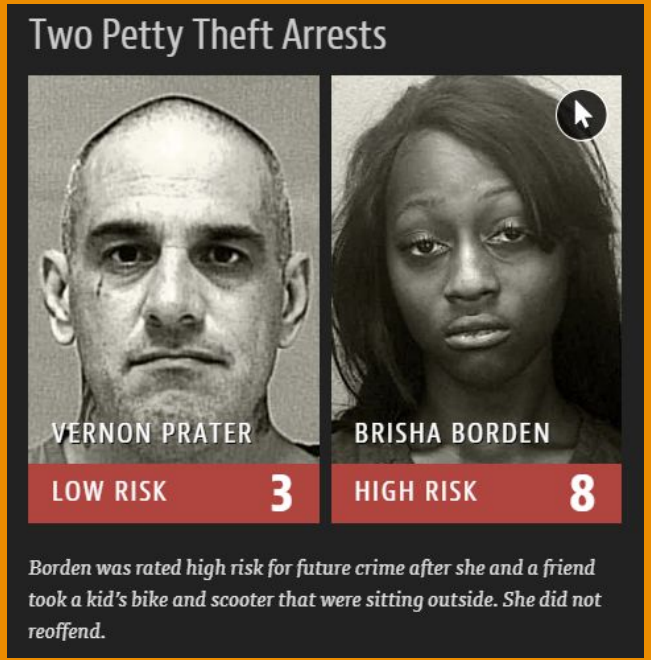
2. ML is rarely perfect and that could lead to unfairness and discrimination

A ML model can be considered **fair if its results don't correlate to certain variables that may be considered sensitive or unjust**, e.g. race, gender, age, nationality, etc.

Assess the model's results by **looking at the distribution of errors across different groups**.

There are many different "fairness" definitions. You will have to choose one that is appropriate for your use case.





2. ML is rarely perfect and that could lead to unfairness, bias, and discrimination - WHAT TO DO

Understand what bias risks are relevant to you - which features are sensitive in the context of your model?

Assess potential bias in your data. Does the data represent enough and equally high quality examples of a diverse population? The model needs good examples to learn to accurately operate vis-a-vis everyone?

Enhance the data

If you need to and can, add more high-quality representative examples of protected groups. It is possible that new data might have to be collected for this.

Identify fairness metrics that are appropriate for the intended model and its use. **Consult with stakeholders on this.**

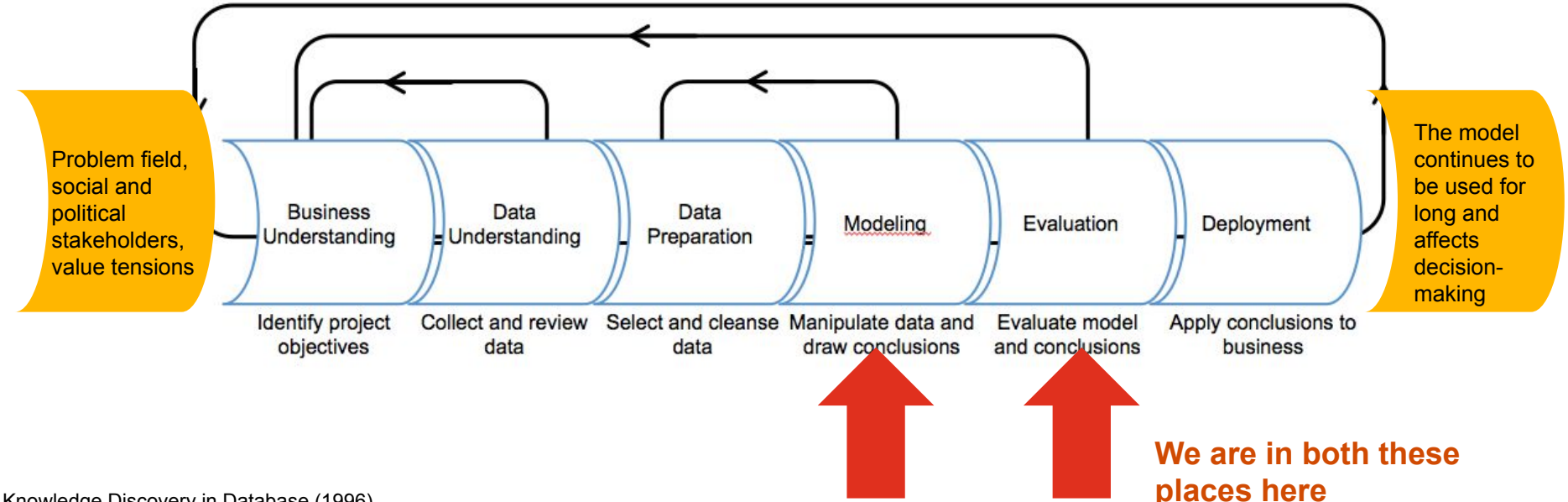
Understand the difference between the different metrics.

Assess your model **on the fairness metric** just as you would assess its accuracy. You can also assess its performance **on benchmark datasets** if available (e.g. Facebook's Casual Conversations).

Mitigate bias in your model by optimising it accordingly or by fine-tuning decision boundaries by hand.

Further reading: Data bias identification and mitigation
Further reading: Understanding bias and fairness in ML
Tools: IBM Fairness 360 toolset for Python

2. ML is rarely perfect and that makes it very important for us to understand why and how it works the way it does



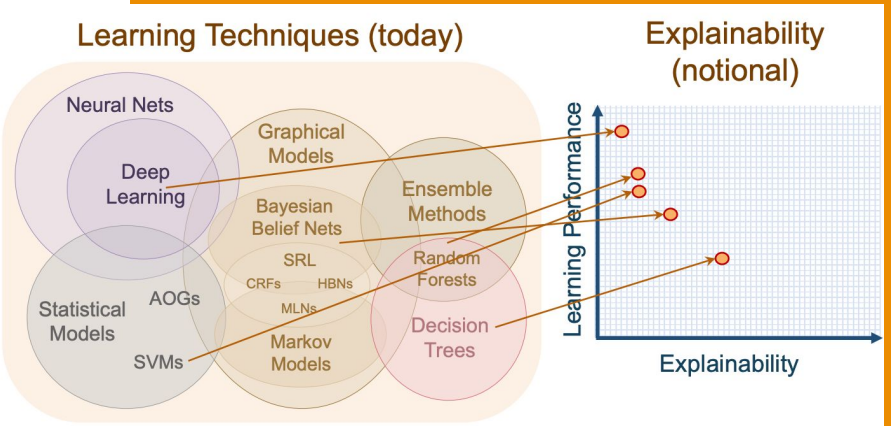
Knowledge Discovery in Database (1996)

2. ML is rarely perfect and that makes it very important for us to understand why and how it works the way it does

The more complex an ML technique is, the better it might be at predictions, but **the less explainable it wi be.** (tradeoff between accuracy and explainability)

“Black box” AI are AI systems for which noone, not even ML experts, can understand how they operate and why a certain result was calculated.

Explainability is necessary for verifying the ‘rules’ that ML ‘learns’ are reasonable and for humans to trust AI. Does the model consid.. certain variables to have predictive power where they shouldn’t (e.g. race, age, gender)?



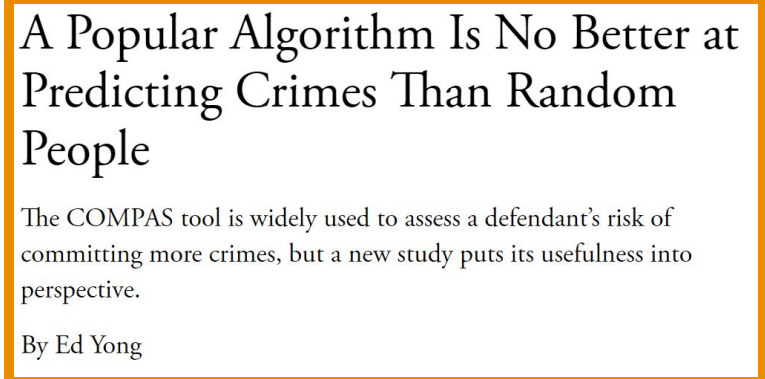
Further reading: Explainable AI
Further reading: Explainable AI - a data scientist’s new challenge

2. ML is rarely perfect and that makes it very important for us to understand why and how it works the way it does

Some tried to use *post hoc* explainability tools, e.g. LAIM, SHAP. These offer some insights, but not always sufficient to understand how the model works.

Using complex AI systems that are unexplainable doesn't always provide superior accuracy. Sometimes, a simpler AI system that focuses on the most relevant and predictive variables and is interpretable can have comparable accuracy levels.

To fully understand and trust an AI model, we also need transparency of the *process of its creation*.



A Popular Algorithm Is No Better at Predicting Crimes Than Random People

The COMPAS tool is widely used to assess a defendant's risk of committing more crimes, but a new study puts its usefulness into perspective.

By Ed Yong

Further reading: Stop explaining black box AI for high stakes decisions and use interpretable models instead (free here)
Further reading: Post modelling explainability

2. ML is rarely perfect and that makes it very important for us to understand why and how it works - WHAT TO DO

Always opt for more explainable models. Especially if your models might be used to affect individual lives.

We already know ML models are statistical and probabilistic tools with no 100% accuracy. Do you really want to have a model that makes impactful calculations for individuals that cannot be understood by anyone?

Further reading: Principles and practice of explainable ML

Review your models to ensure the correlations and 'rules' they learn are reasonable and appropriate.

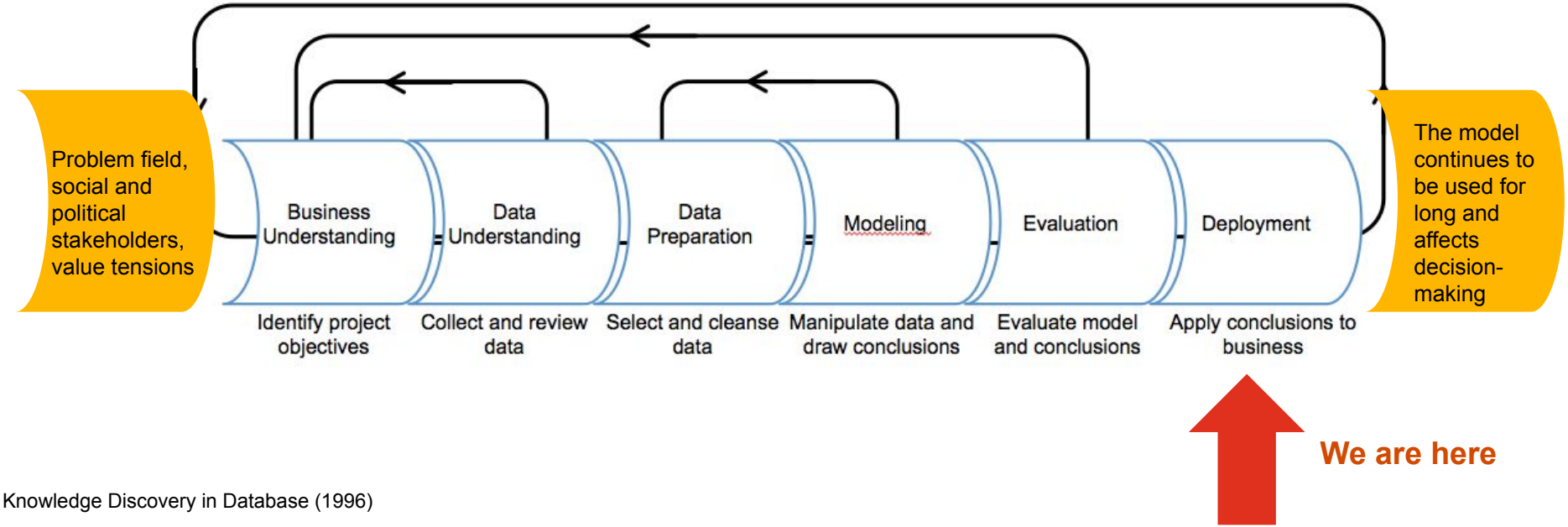
Pay attention that the variables that have predictive power and are influential for the functioning of the model are actually relevant to the outcome and not sensitive, e.g. race, age, gender, etc.

In addition to explainability of the model, think about the transparency of your design process. Keep an auditable trail of your work process.

This allows others to understand how and why a model was created and what assumptions and design decisions were made in the process.

Further reading: Closing the AI accountability gap

3. The way ML interacts with the real world requires continuous review, maintenance and human oversight



Knowledge Discovery in Database (1996)

3. The way ML interacts with the real world

Once you create and deploy an ML model, its performance can change over time and needs to be monitored and, if needed, retrained on newer, more up-to-date data.

- **Data drift -** The type or quality of data your model receives changes over time, e.g. there are new measurement methods that change the measurement unit or what data is measured.

- **Concept drift** - The target variable and its statistical properties could change over time. E.g. when the model is launched, your age might be relevant for your likelihood of starting a business, but at the end, society has changed and age might no longer be a predictive variable.

Example:

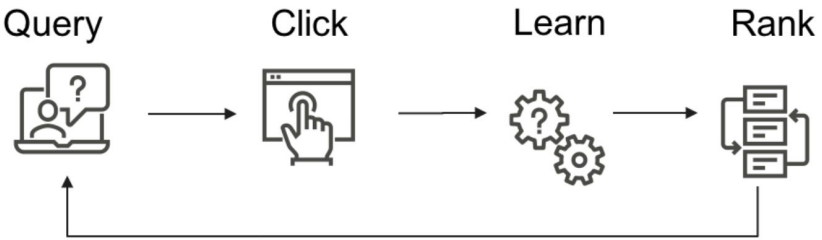
If you train a model to predict the expected rent costs / sq m in Munich for 2022, then having the rent costs for 2010-2021 would be helpful.

The same model would be less useful to predict the rent costs in Munich for 2050.

Further reading: Monitoring and retraining your model
Further reading: Data drift
Further reading: Concept drift

3. The way ML interacts with the real world

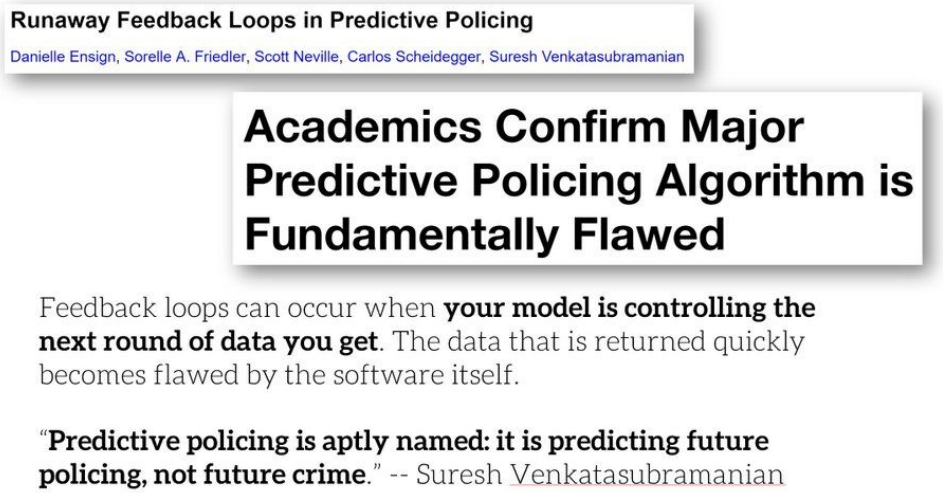
Feedback loops can be valuable tools to help improve a ML model by providing it with feedback from which to learn, e.g. like recommender systems >>>>



3. The way ML interacts with the real world

But feedback loops can also become a problem. They can end up reinforcing a pattern of model predictions or behaviour that has **nothing to do with reality, but that is caused by the model's outputs themselves**.

For example, if a model says "this neighbourhood is risky" more police officers might be sent there. If there are more police officers there, there are likely going to be more arrests, stops and searchers, etc. This gives the model more data to reaffirm the bias that that neighbourhood is risky.



Source: Rachel Thomas twitter thread on AI bias
Further reading: Runaway feedback loops in predictive policing

3. The way ML interacts with the real world - WHAT TO DO

At the stage of defining and scoping your project, define the precise use case for ML models and AVOID feedback loops.

Include this in any instructions and guidance on the intended use of the ML model you give the partner.

You want to avoid feedback loops from occurring as much as possible. This means the results of the ML and how they are used should not be directly linked to inputs in the ML model.

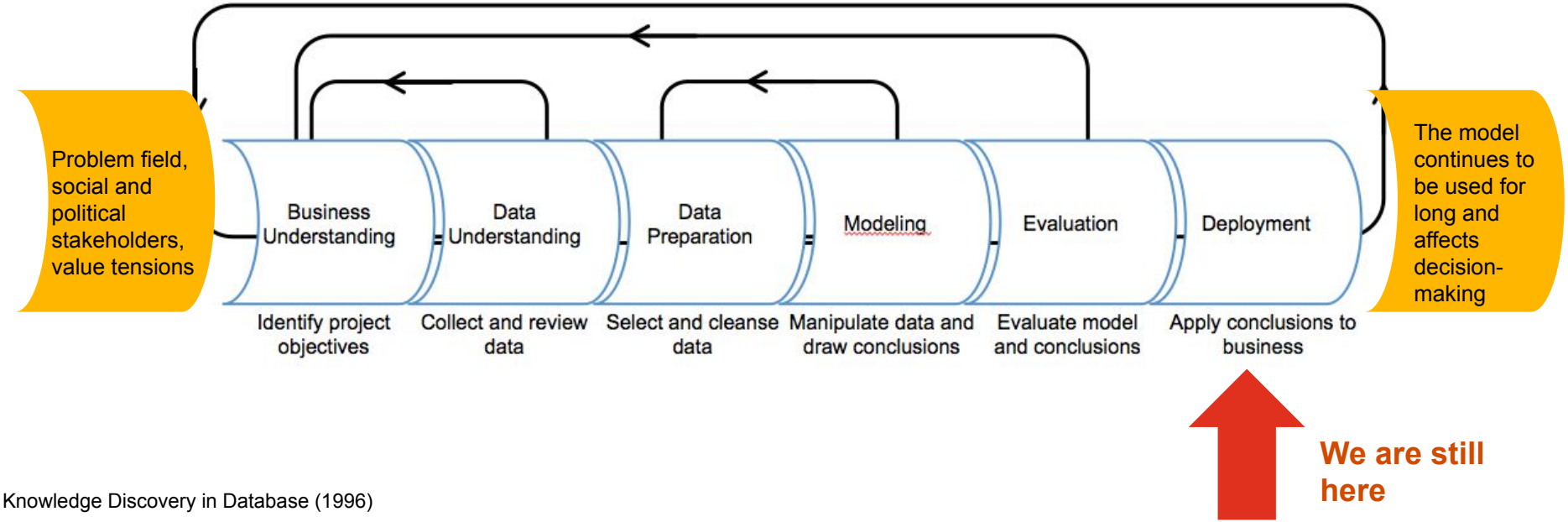
Empower the partner organisation to review and maintain over any ML model you create for them by providing them instructions for use and maintenance.

There are many options how the partners might do that, e.g. having a responsible person to periodically review the performance of ML models.

You should provide them at least with instructions on how to review and maintain the ML, further resources, and things to be careful for.

Further reading: How positive feedback loops are hurting AI applications
Further reading: A guide on when to retrain your ML model
Further reading: How to keep your ML models up-to-date
Further reading: A practical guide to maintaining ML models in production

4. The way humans interact with ML requires special instructions, user training, and UX design



Knowledge Discovery in Database (1996)

4. The way humans interact with ML requires special instructions, user training, and UX design

Over-reliance / Automation bias: people are more likely to trust technology, even when they make a mistake. This can lead to blindly trusting ML models and not identifying errors or wrong predictions. Busy users, overwhelmed with information or tasks are more likely to over-rely on technology. Risk averse work cultures also incentivise users to over-rely on ML and distrust themselves, “better be safe than sorry”.

Under-reliance / Automation aversion: people are more likely to dislike and avoid technology. This could mean people distrust and don’t use technology, even when it would be helpful.

Low digital literacy and not understanding the tools fully could lead to either over- or under-reliance.

UI designs can affect how users interact with models.
Colours, alerts, sounds can create a sense of urgency and need to act. Providing objective visualisations and reminders and disclaimers about the tool’s limitations, accuracy levels and confidence intervals can help keep users in control.

Source

4. The way humans interact with ML requires special instructions, user training, and UX design - WHAT TO DO

Involve stakeholders in the design of the tool. Explain to them how you made it and how it works to build trust. Otherwise people won't use it or won't use it as intended.

Take on their feedback to improve the work and build their trust in and understanding of the tool.

Avoid UI design that "nudges" people to action through colours, dark patterns, etc.

Highlight the model's limitations and intended scope - accuracy, intended use, error rates or confidence levels for model performance visibly.

Further reading on Design ethics

Create information sheets that highlight the capabilities and limitations of the tool, what it is intended for.

To inform people of your model's intended use and limited capabilities, fill in (a version of) Model Cards for model reporting.

Create instruction manuals for the intended users of the model.

There refer to information about the model, how it was created and why, it performance, intended use, and limitations.

! Don't forget the instructions on maintaining the model!

Are you confused yet?

How do you feel about everything from today?

How do you feel about ML now?

Do you think you can help use ML for good? What would you need to do that?

62

If it all sounds confusing, what to take away from today as practical tips:

Take responsibility. There is a great responsibility that comes with writing code because you make decisions that very few people can assess and question afterwards. You should always strive to make the right decisions.

The burden is on you to ensure your actions do not cause harm to others. Risk assessment and risk management exercises can help you do that in practice.

Be mindful of what might go wrong and what you don't know. Be open and engage with domain experts. Learn from them. Consult them on important matters. Ask for their feedback and advice. Especially when you might have to choose between technical trade-offs.

Be honest with stakeholders and partners about what you can and cannot do. Scope the projects so that you can deliver ethical and reliable solutions.

Deliver sustainable solutions that partners can understand, use, and maintain. Opt for simpler solutions that the partners can understand and maintain if need be.