Lecture 1

Introduction

Find the Next Number of the Sequence…

1, 3, 5, 7, ?

Correct Solution:

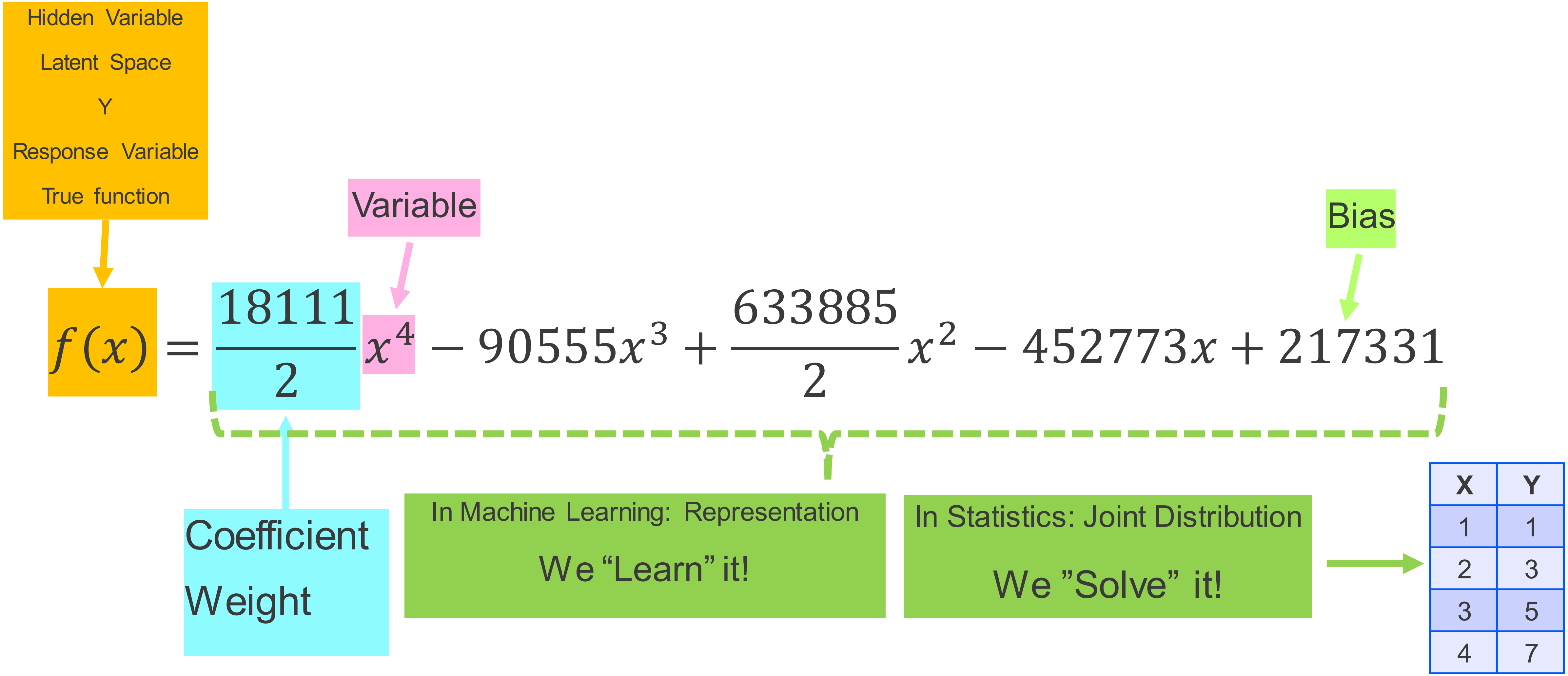
217341

Because when

$$f(x) = \frac{18111}{2}x^4 - 90555x^3 + \frac{633885}{2}x^2 - 452773x + 217331$$

$$f(1) = 1$$
$$f(2) = 3$$
$$f(3) = 5$$
$$f(4) = 7$$
$$f(5) = 217341$$

Find the Next Number of the Sequence…



Hidden Variable

Latent Space

Y

Response Variable

True function

Variable

Bias

$$f(x) = \frac{18111}{2}x^4 - 90555x^3 + \frac{633885}{2}x^2 - 452773x + 217331$$

Coefficient

Weight

In Machine Learning: Representation

We “Learn” it!

In Statistics: Joint Distribution

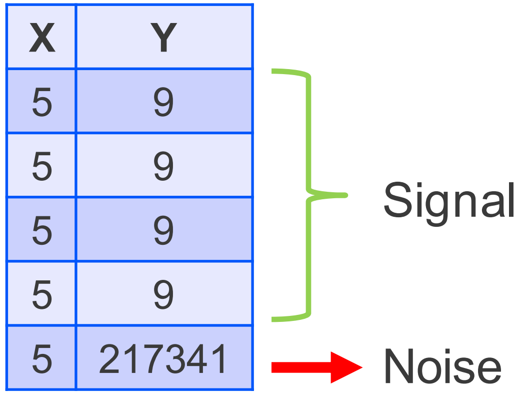
We ”Solve” it!

X	Y
1	1
2	3
3	5
4	7

Data Scientist vs. Oracle

Data Scientist

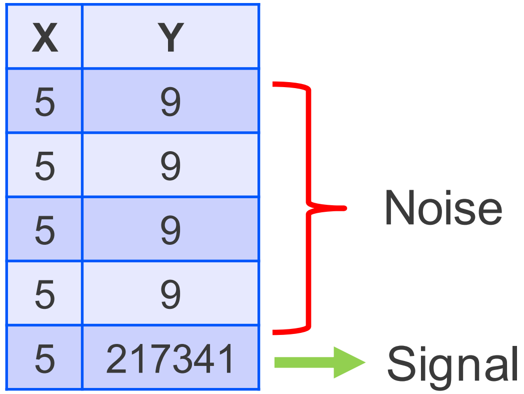
- The Data Scientist estimates $f(x)$
- The Data Scientist may know the “Domain Knowledge”



X	Y
5	9
5	9
5	9
5	9
5	217341

Oracle

- The Oracle gives us $f(x)$
- The Oracle may be replaced by a “Knowledge Expert” in a field.



X	Y
5	9
5	9
5	9
5	9
5	217341

ML Research

offline datasets
annotated a long time ago
simulated environments
abstract domains
restart experiments at will
...

Reality

horns
nose
tail
...
also more cute

Image credit: Keenan Crane & Nepluno CC BY-SA

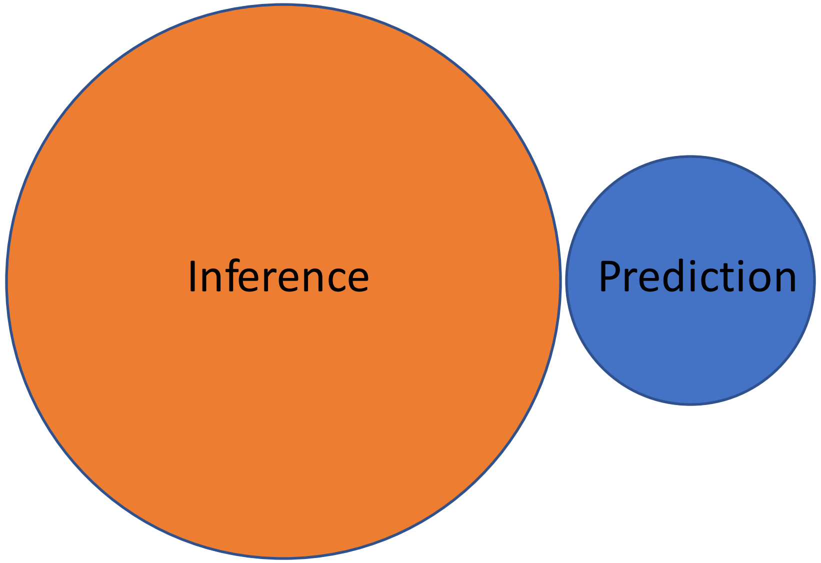
Predictive Vs. Inference Models

Inference

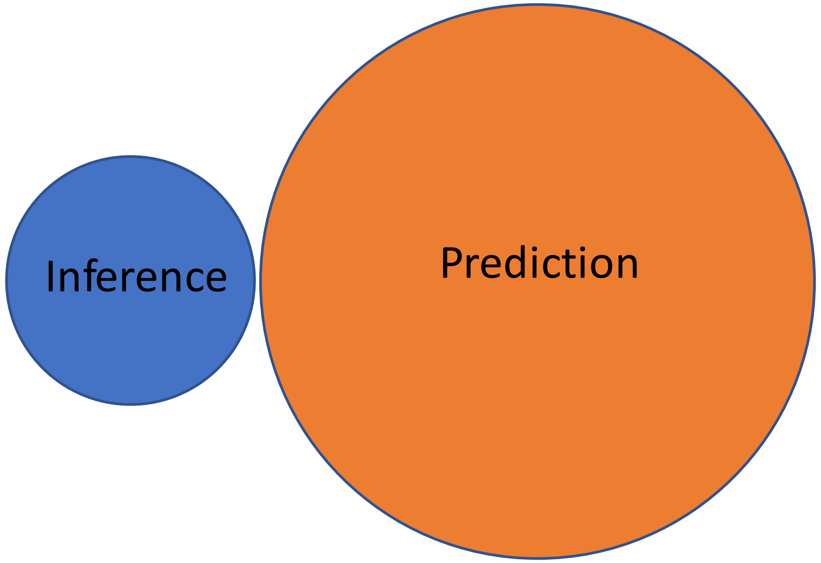
Predictive

Predictive Vs. Inference Models

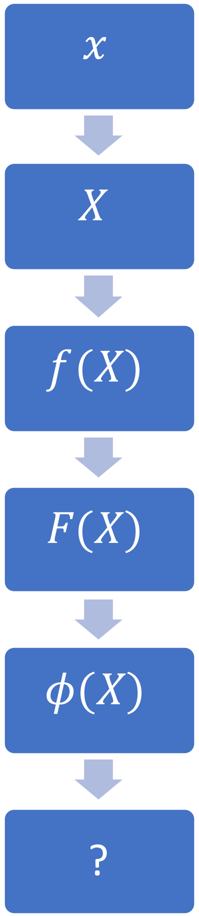
Statistics



Machine Learning



Model Hierarchy



x

X

$f(X)$

$F(X)$

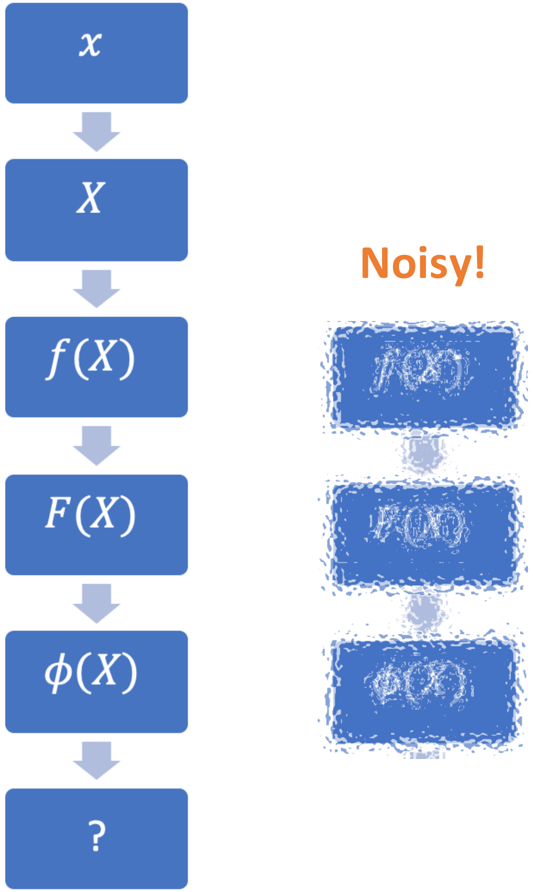
$\phi(X)$

?

Predictive Vs. Inference Models

Inference by Statistics:

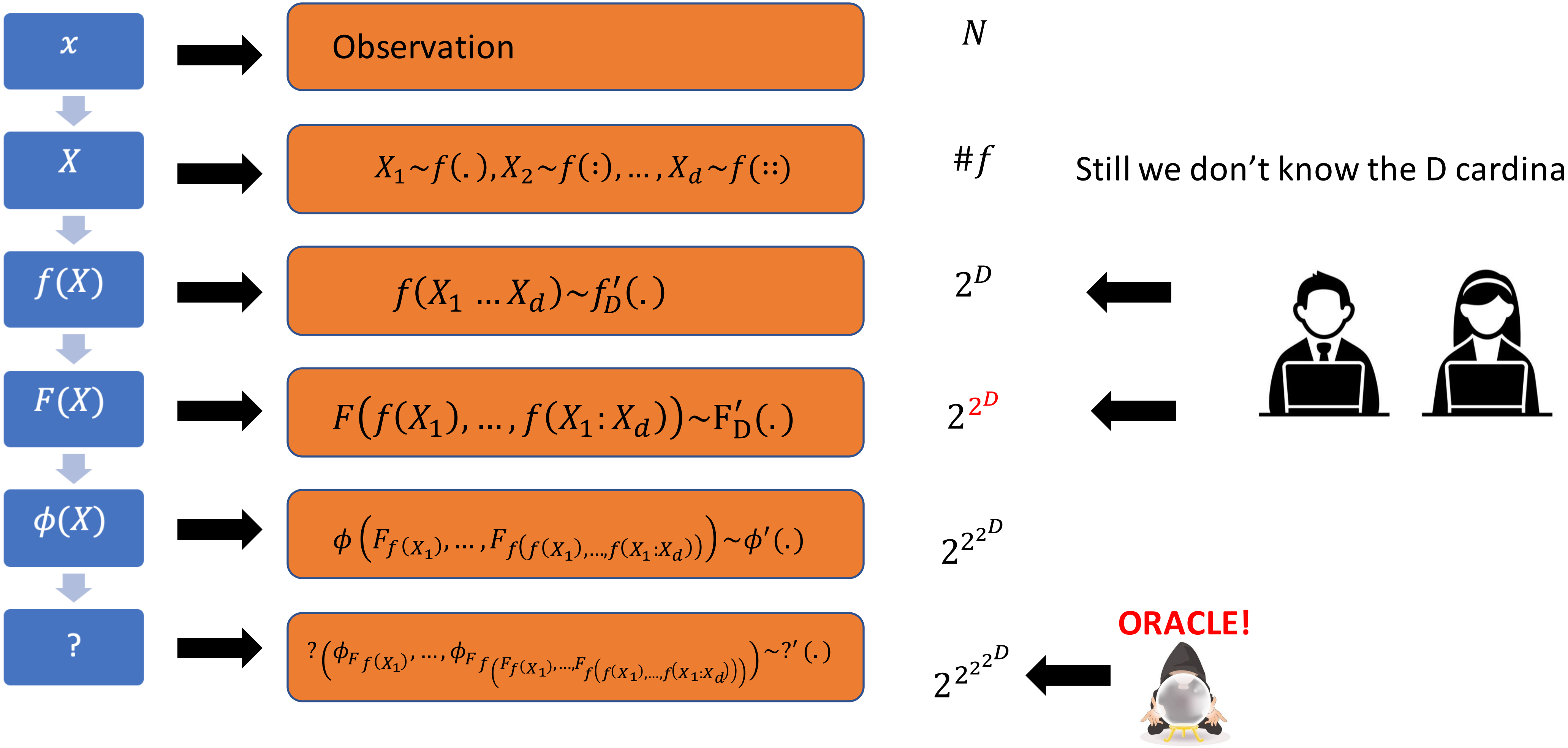
- **More assumptions**
- Closed form solution
- Inference is first
- Expert selects features
- Convex models
- Low # parameters
- …



Prediction by Machine Learning:

- **Less assumptions**
- Fast evolving
- Generalization is first
- Machine extracts features
- Optimized solutions
- Super high # parameters
- …

Model Complexity



x → Observation — N

X → $X_1 \sim f(.), X_2 \sim f(:), \ldots, X_d \sim f(::)$ — $\#f$

Still we don't know the D cardinality!

$f(X)$ → $f(X_1 \ldots X_d) \sim f'_D(.)$ — 2^D

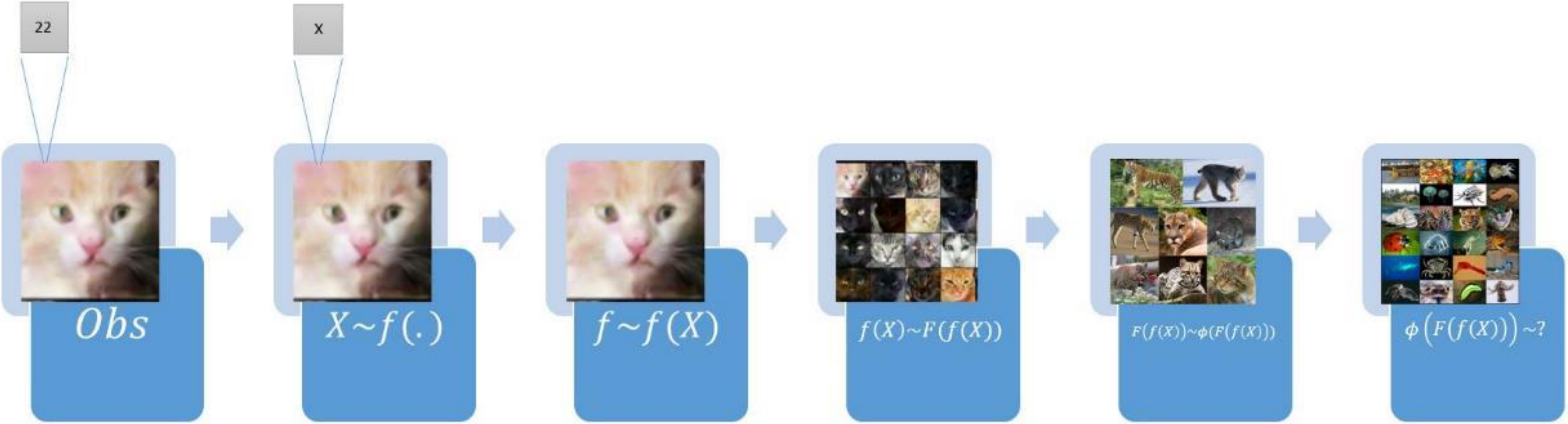
$F(X)$ → $F\big(f(X_1), \ldots, f(X_1 : X_d)\big) \sim \mathrm{F}'_{\mathrm{D}}(.)$ — 2^{2^D}

$\phi(X)$ → $\phi\left(F_{f(X_1)}, \ldots, F_{f(f(X_1),\ldots,f(X_1:X_d))}\right) \sim \phi'(.)$ — $2^{2^{2^D}}$

$?$ → $?\left(\phi_{F_{f(X_1)}}, \ldots, \phi_{F_{f\left(F_{f(X_1)},\ldots,F_{f\left(f(X_1),\ldots,f(X_1:X_d)\right)}\right)}}\right) \sim ?'(.)$ — $2^{2^{2^{2^D}}}$

ORACLE!

Model Hierarchy



22

x

Obs

$X \sim f(.)$

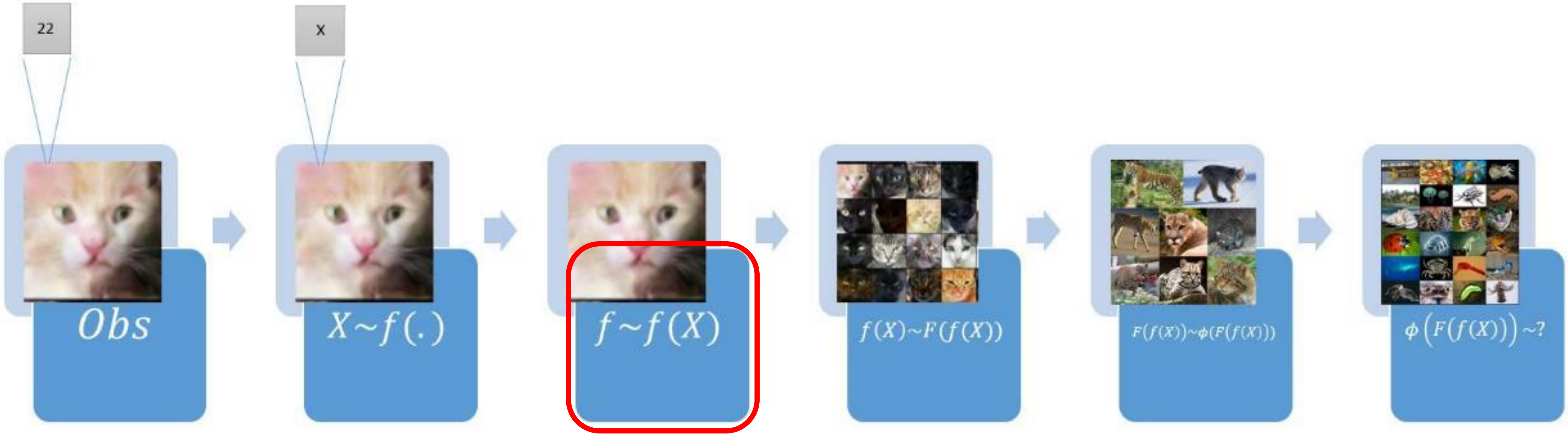
$f \sim f(X)$

$f(X) \sim F(f(X))$

$F(f(X)) \sim \phi(F(f(X)))$

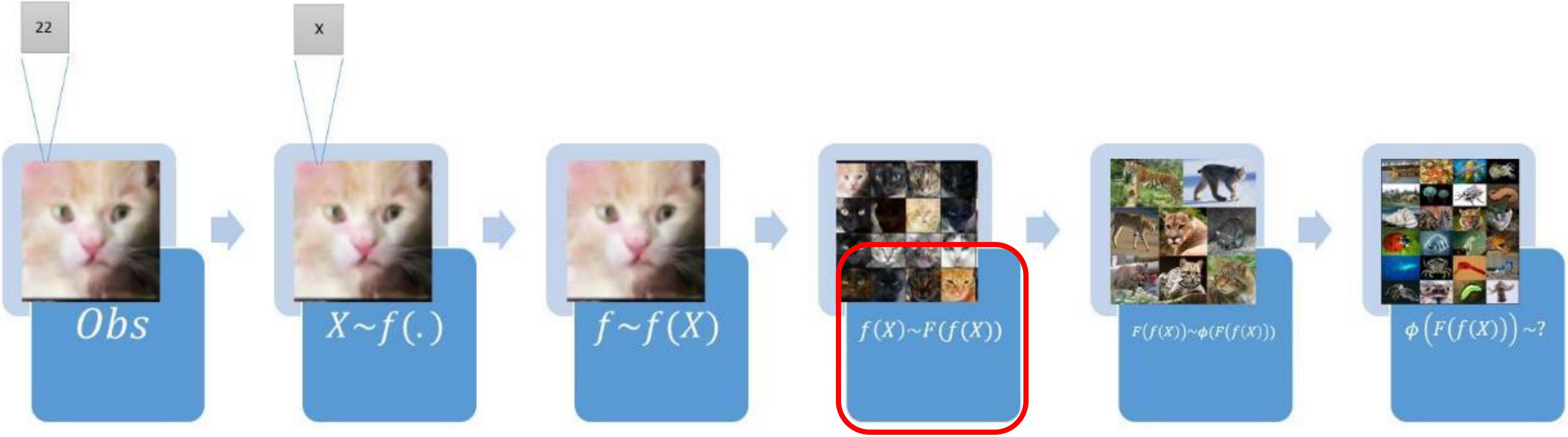
$\phi\left(F(f(X))\right) \sim ?$

Model Hierarchy


22
x
Obs
$X \sim f(.)$
$f \sim f(X)$
$f(X) \sim F(f(X))$
$F(f(X)) \sim \phi(F(f(X)))$
$\phi\left(F(f(X))\right) \sim ?$

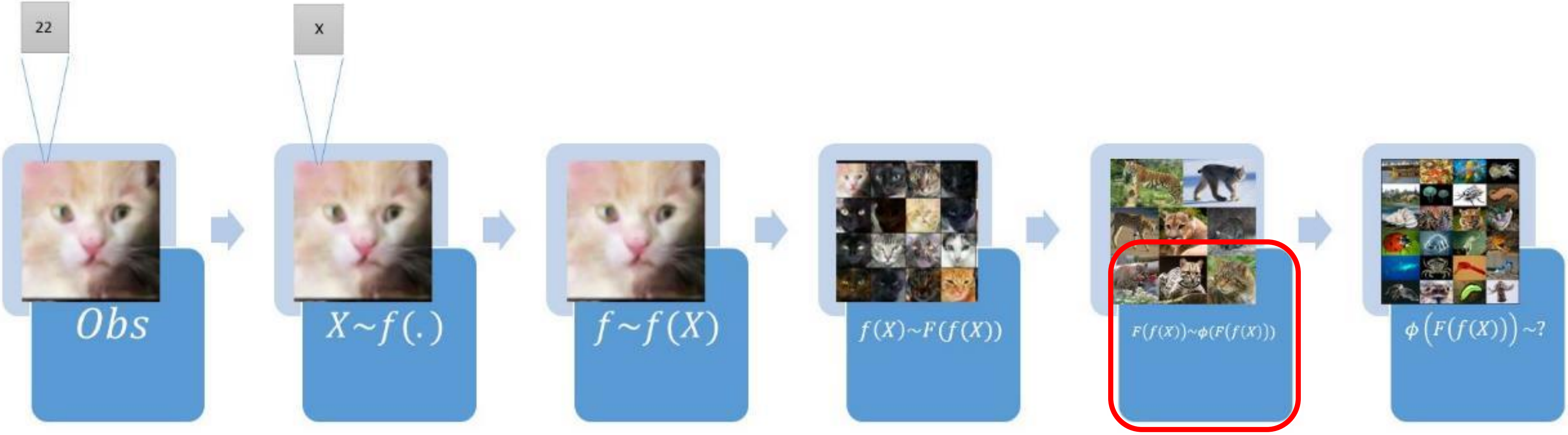
Most Machine Learning models

Model Hierarchy


22
x
Obs
$X \sim f(.)$
$f \sim f(X)$
$f(X) \sim F(f(X))$
$F(f(X)) \sim \phi(F(f(X)))$
$\phi\left(F(f(X))\right) \sim ?$

Generative Adversarial Network (GAN)
Latent Dirichlet Allocation (LDA)
Variational Auto-Encoder (VAE)

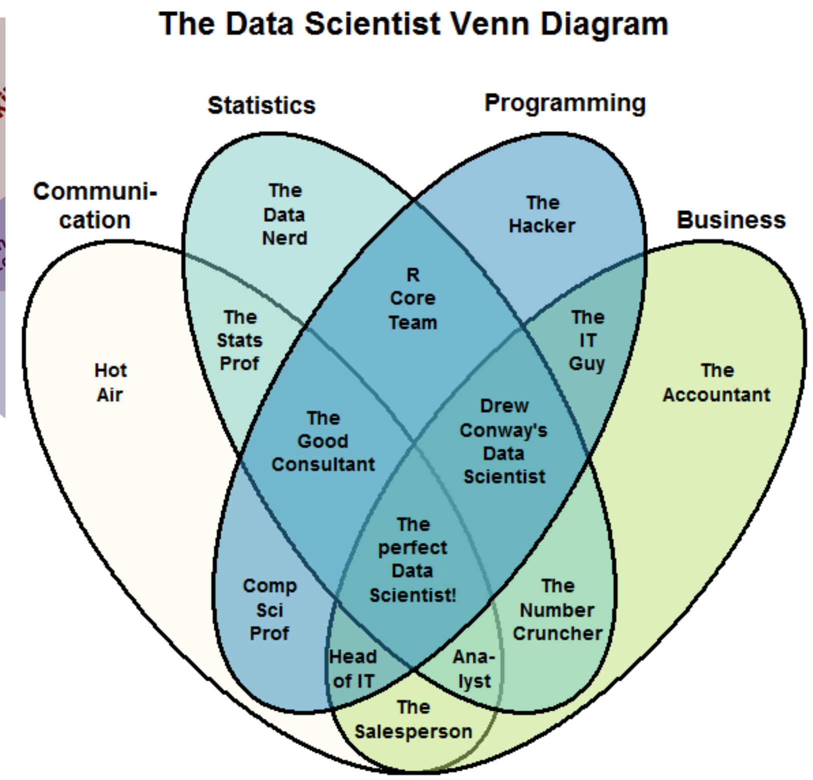
Model Hierarchy

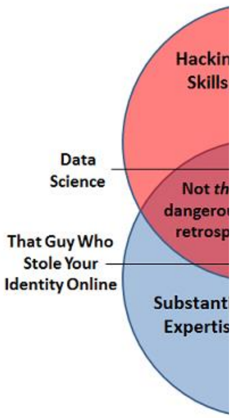


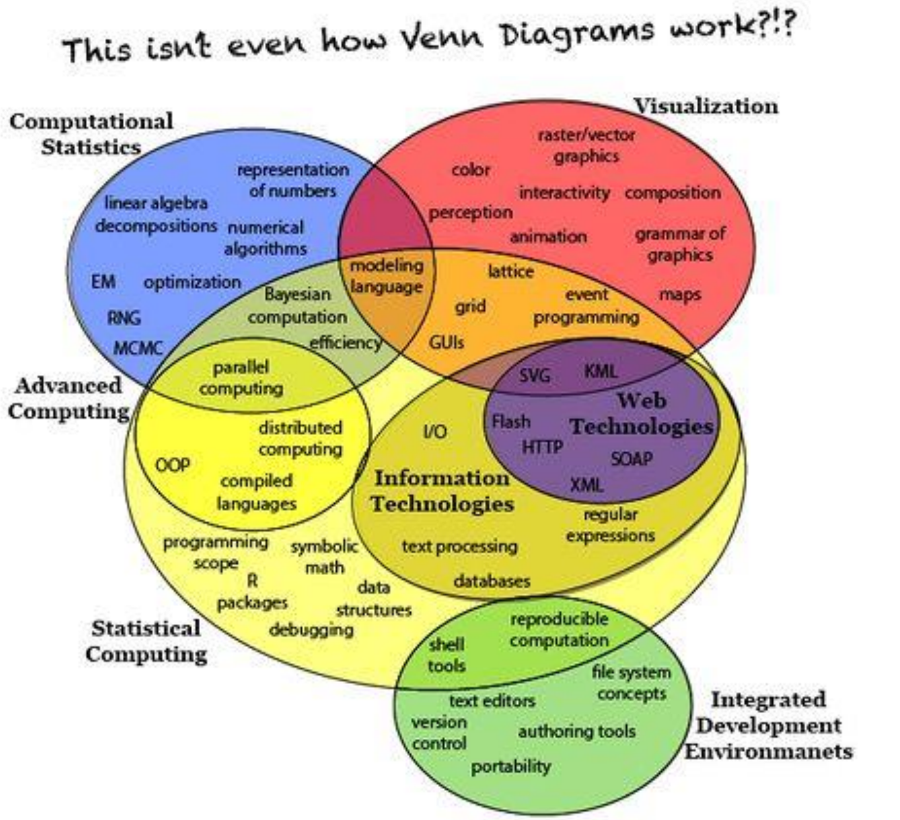
Diffusion Models

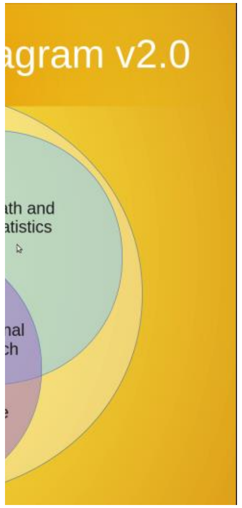
What is Data Science?

- It’s a field singularly devoted to bringing back the Venn Diagram









What is Data Science?

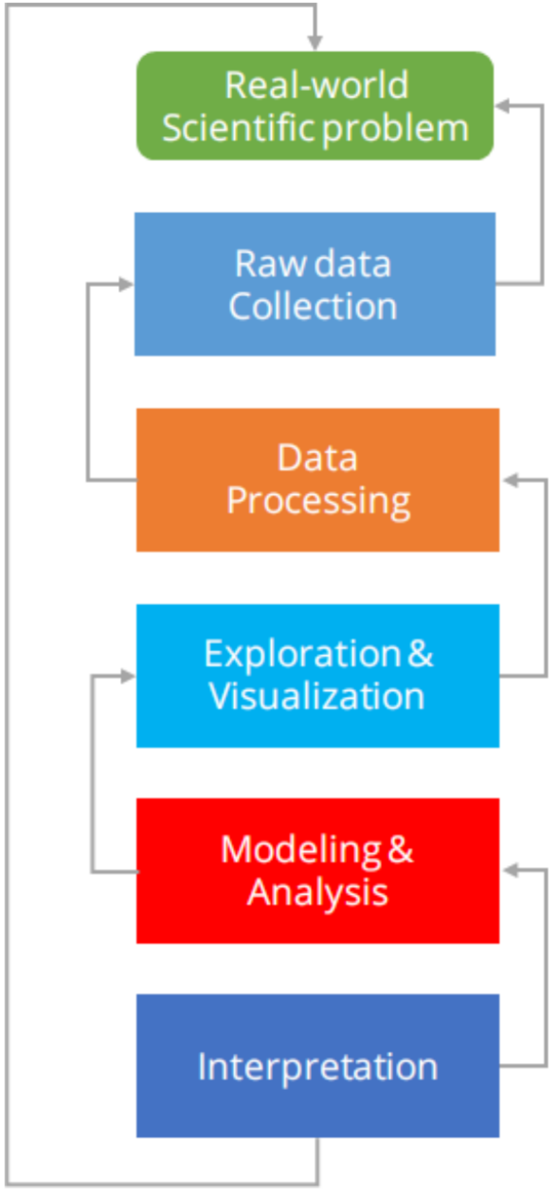
From Wikipedia:

It is an **interdisciplinary field** about scientific methods, processes, and systems to **extract knowledge or insights from data in various forms, either structured or unstructured.**

What is Data Science?

Data Science encompasses **the entire problem stack:**

- Problem definition
- Data collection & cleaning
- Exploration
- Modeling
- Interpretation & insights



When is Data Science Dangerous?

Data science can be possibly dangerous:

- When hacking skills are applied without statistics
- When statistics is used without domain knowledge.

Who is a Data Scientist?

An expert who knows **<u>Statistics</u>** more than a computer scientist and knows **<u>Computer Science</u>** more than a Statistician.

Machine Learning Engineer vs. Data Scientist

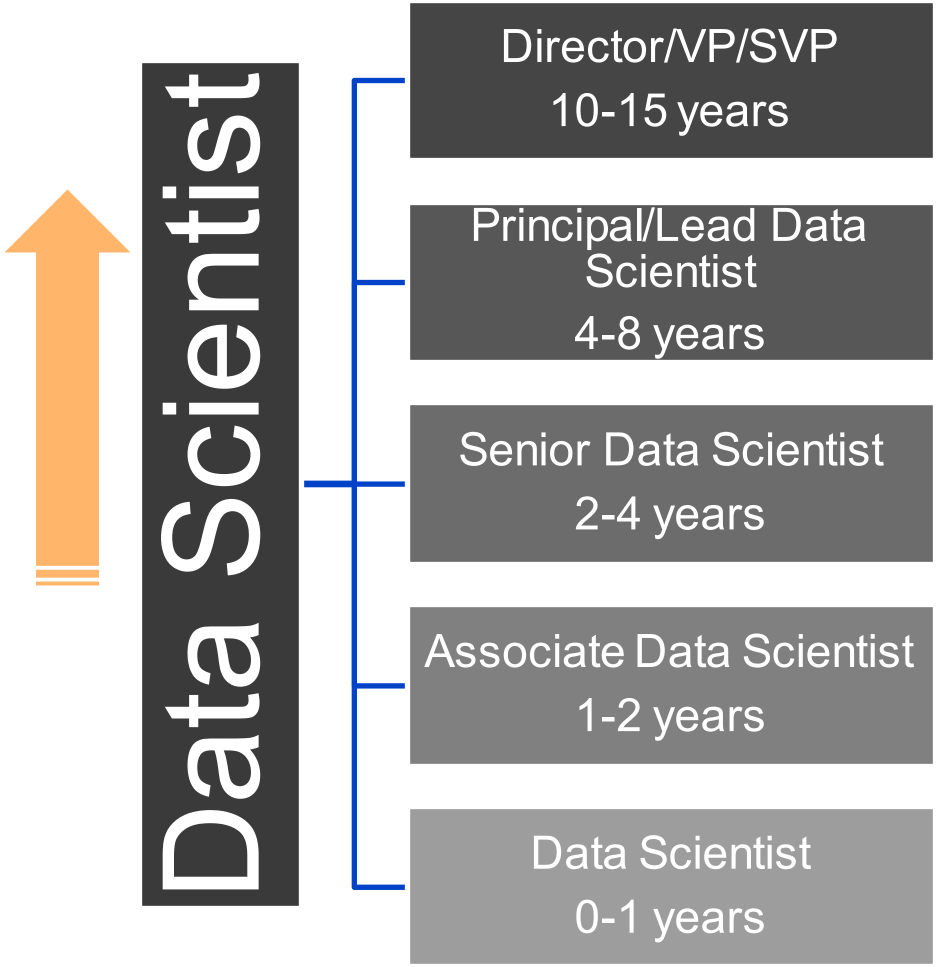
Data Scientist

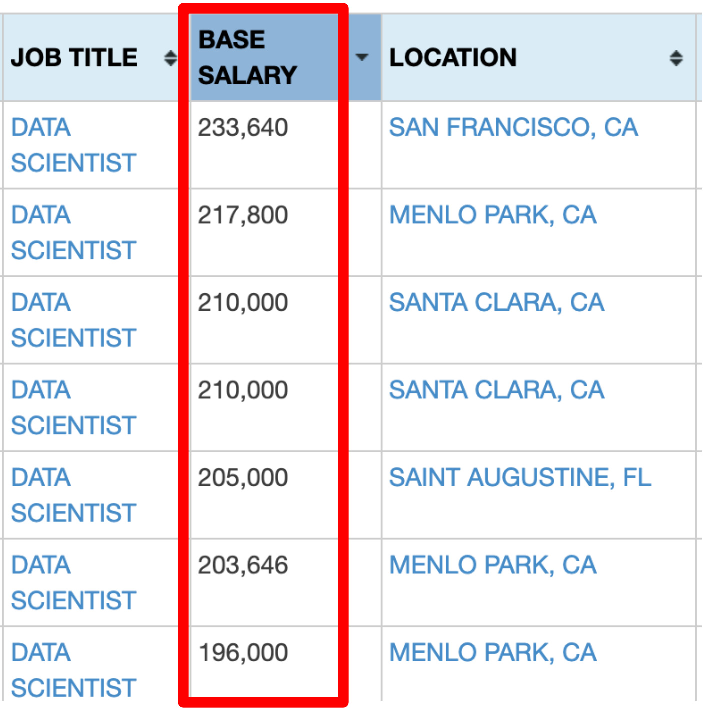
- Product Sense
- Experiment Design
- A/B Test
- Business Strategy
- Less Structured
- More Uncertainties
- Mostly R, SAS or Python
- Interpretability

Machine Learning Engineer

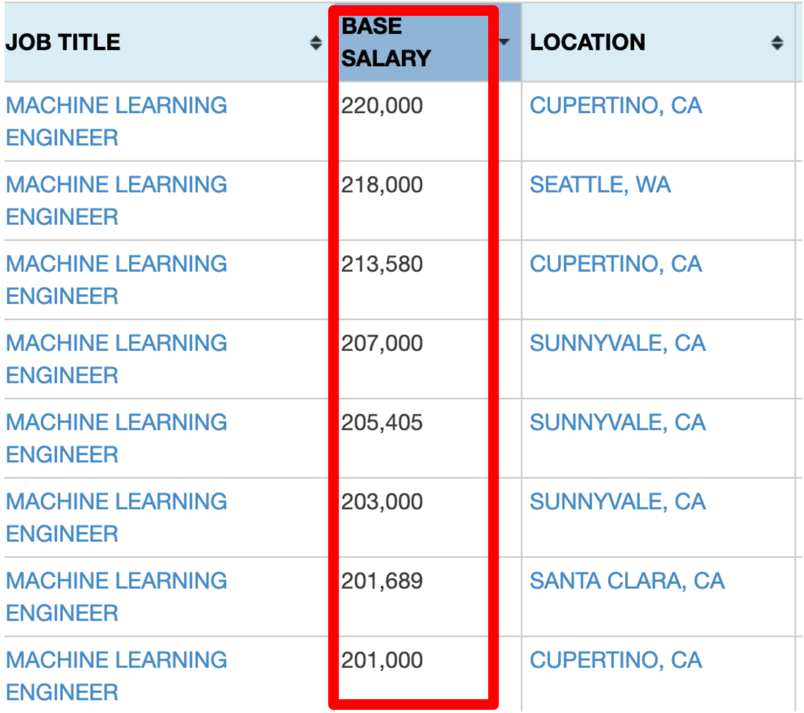
- Model Building
- SWE Role
- State-of-the-art ML Models
- More Structured
- Less Uncertainties
- Mostly Python (OOP)
- Scalability and Accuracy

Data Scientist Career Path





JOB TITLE	BASE SALARY	LOCATION
DATA SCIENTIST	233,640	SAN FRANCISCO, CA
DATA SCIENTIST	217,800	MENLO PARK, CA
DATA SCIENTIST	210,000	SANTA CLARA, CA
DATA SCIENTIST	210,000	SANTA CLARA, CA
DATA SCIENTIST	205,000	SAINT AUGUSTINE, FL
DATA SCIENTIST	203,646	MENLO PARK, CA
DATA SCIENTIST	196,000	MENLO PARK, CA



JOB TITLE	BASE SALARY	LOCATION
MACHINE LEARNING ENGINEER	220,000	CUPERTINO, CA
MACHINE LEARNING ENGINEER	218,000	SEATTLE, WA
MACHINE LEARNING ENGINEER	213,580	CUPERTINO, CA
MACHINE LEARNING ENGINEER	207,000	SUNNYVALE, CA
MACHINE LEARNING ENGINEER	205,405	SUNNYVALE, CA
MACHINE LEARNING ENGINEER	203,000	SUNNYVALE, CA
MACHINE LEARNING ENGINEER	201,689	SANTA CLARA, CA
MACHINE LEARNING ENGINEER	201,000	CUPERTINO, CA

Source: h1bdata.info

Which Programming Language?

Which Programming Language?

```
model = lm(y~., data=df)
score = predict(model, newdf)
```

Which Programming Language?

```
proc Reg data=mydata;
title ‘Example of Linear Regression in SAS’
model y = x;
run;
```

Which Programming Language?

python™

```
From sklearn.linear_model import LinearRegression
model=LinearRegression().fit(x,y)
r_sq=model.score(x,y)
```

Which Programming Language?

Which Programming Language?

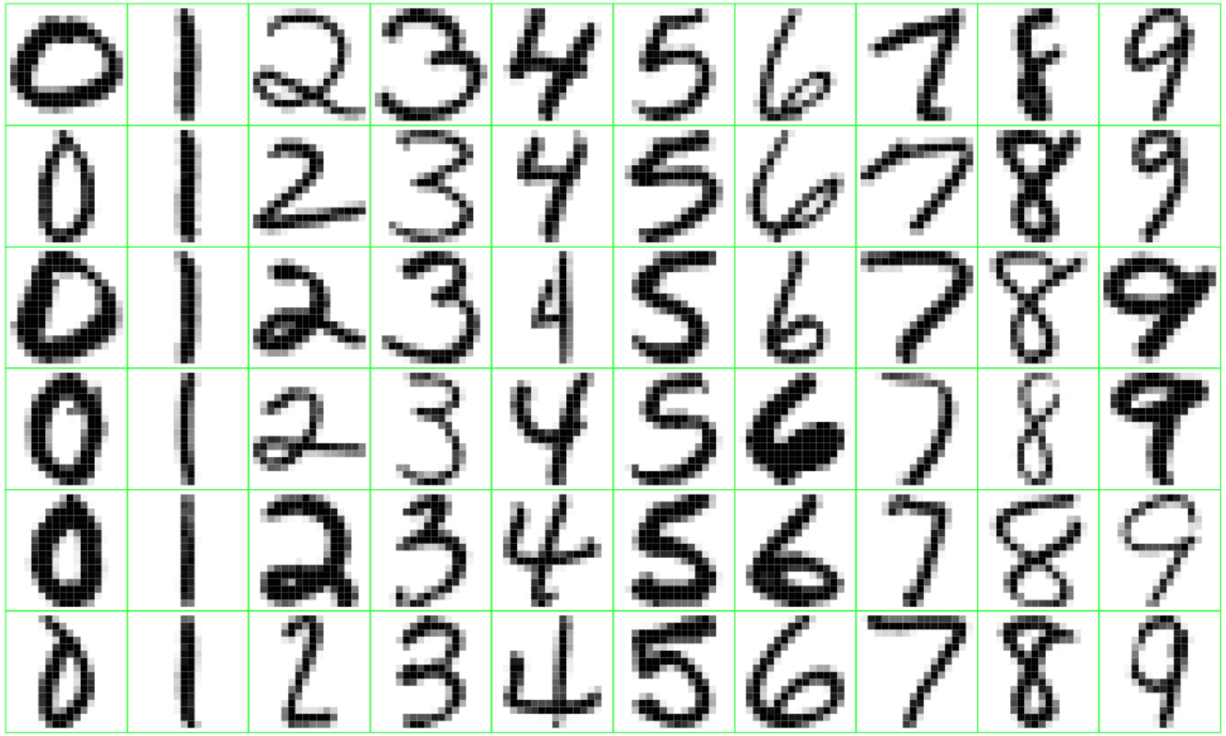
	R	sas	python™	Java
COST	🟩	🟥	🟩	🟩
Ease of Learning	🟩	🟨	🟨	🟥
Scalability	🟨	🟨	🟩	🟩
Visualization	🟩	🟨	🟩	🟥
Advancements in Tool	🟩	🟨	🟩	🟥
Reporting	🟩	🟩	🟨	🟥
Customer Service	🟥	🟩	🟥	🟥
Deep Learning Support	🟩	🟥	🟩	🟨
Online Resources	🟩	🟨	🟩	🟨
Reliability	🟥	🟩	🟥	🟥

Data Science in Practice

- Banking & Insurance
- Media & Communications
- Healthcare
- Education
- Manufacturing
- Retail & CPG
- Energy & Utilities

Data Science in Practice

- Can we automatically sort mail based on ZIP code?

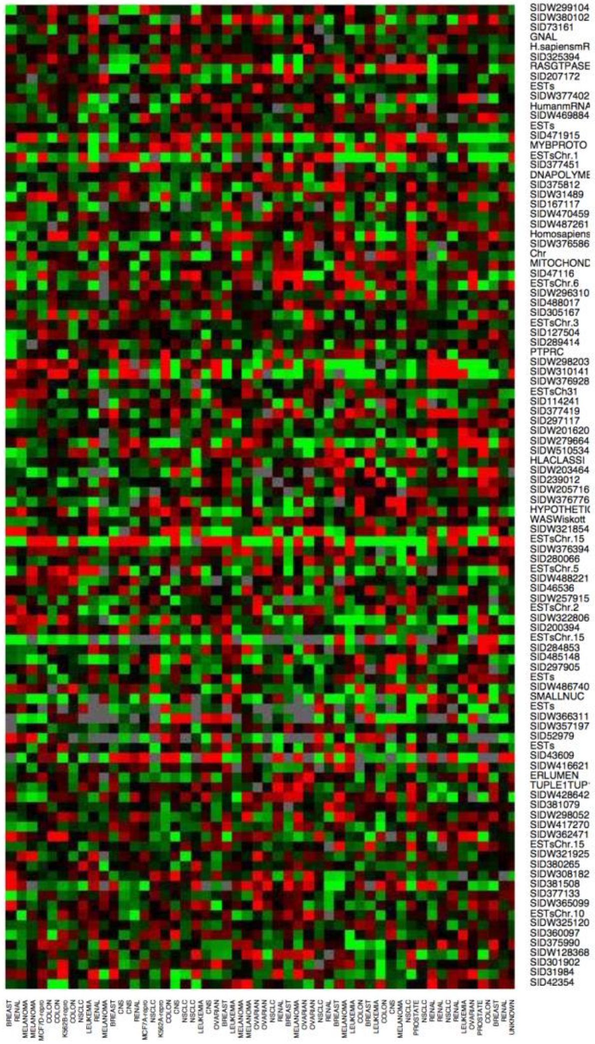


Data Science in Practice

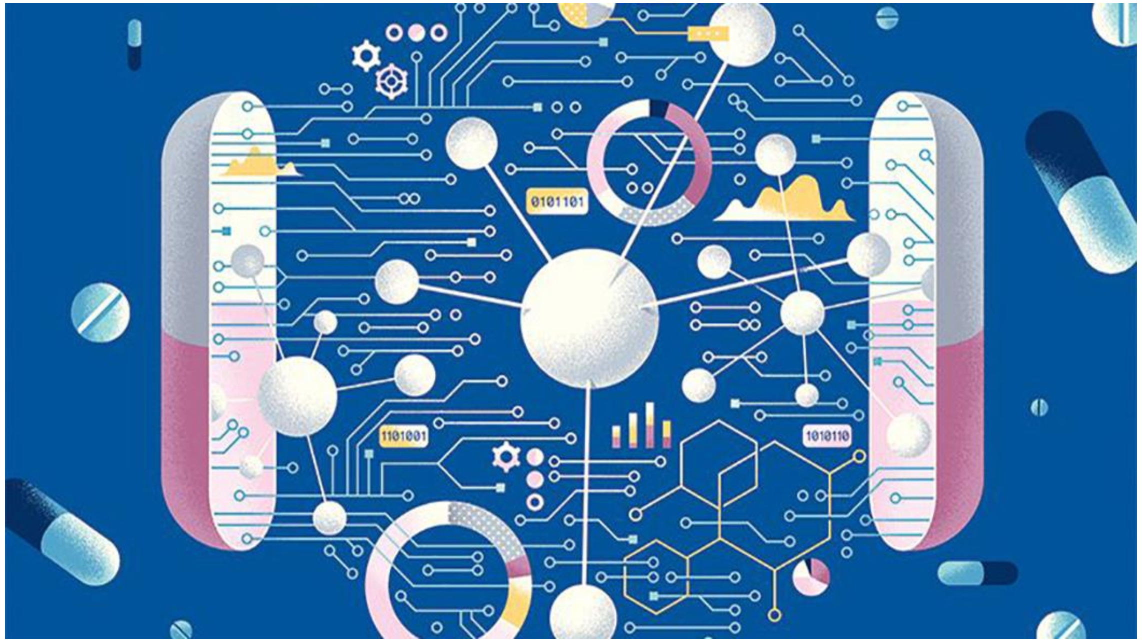
- Can we detect online fraudulent transactions?

Data Science in Practice

Which genes are overactive or underactive in cancer patients?



Data Science in Practice

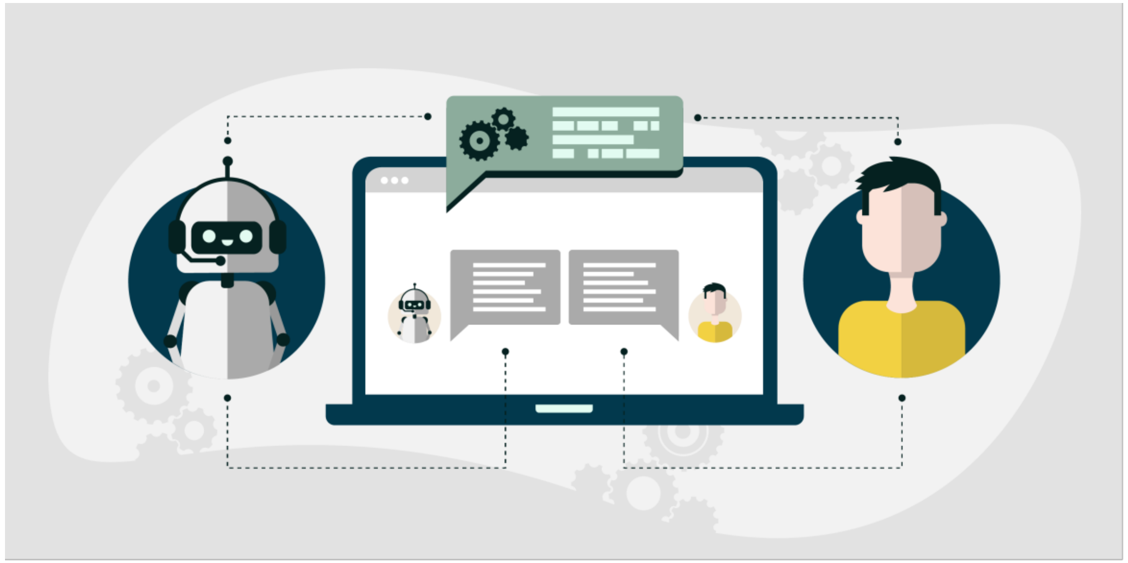


- Can we shorten the process of drug discovery in the treatment of diseases?

Data Science in Practice

Can data science algorithms make video games responsive and adaptive?

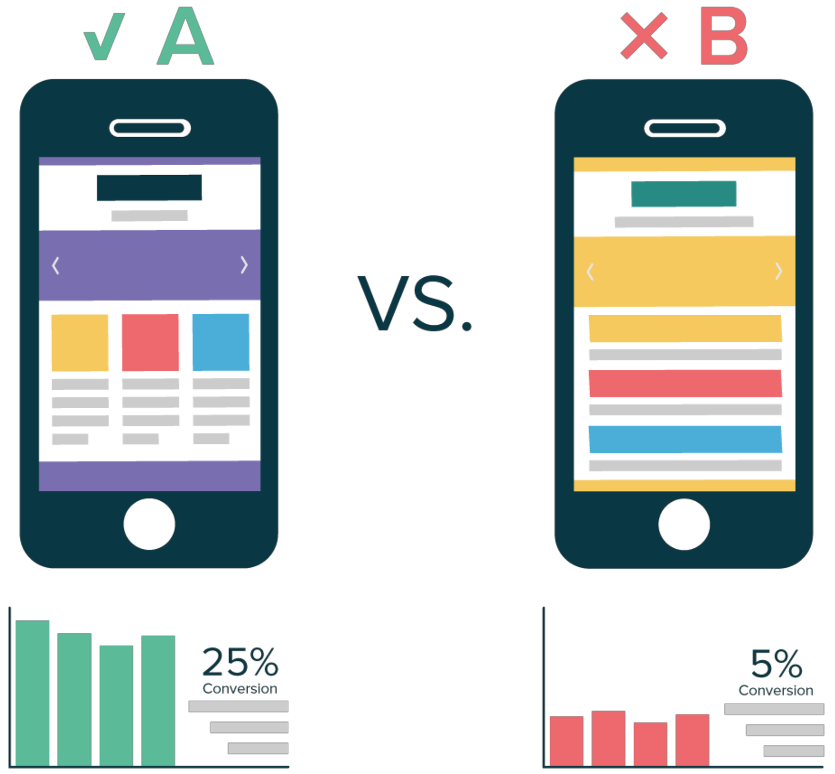
Data Science in Practice



- Virtual assistance for patients and customer support.

Data Science in Practice

Can we shorten A/B test process while maximizing user satisfaction?



Trend in Data Science

- Responsible Data Science: addresses the issues of fairness, diversity, accountability, transparency, privacy, quality, legal compliance and ethics of data and algorithms.

- Knowledge Graph: refers to integration, unification, analytics and sharing of data via linking and semantic metadata of entities, objects, events or concepts.

- Causal-inference: helps predictive models to be more reliable by reasoning what might happen if we change a system or take an action.
- …

Case Study

Netflix

Case Study: Netflix

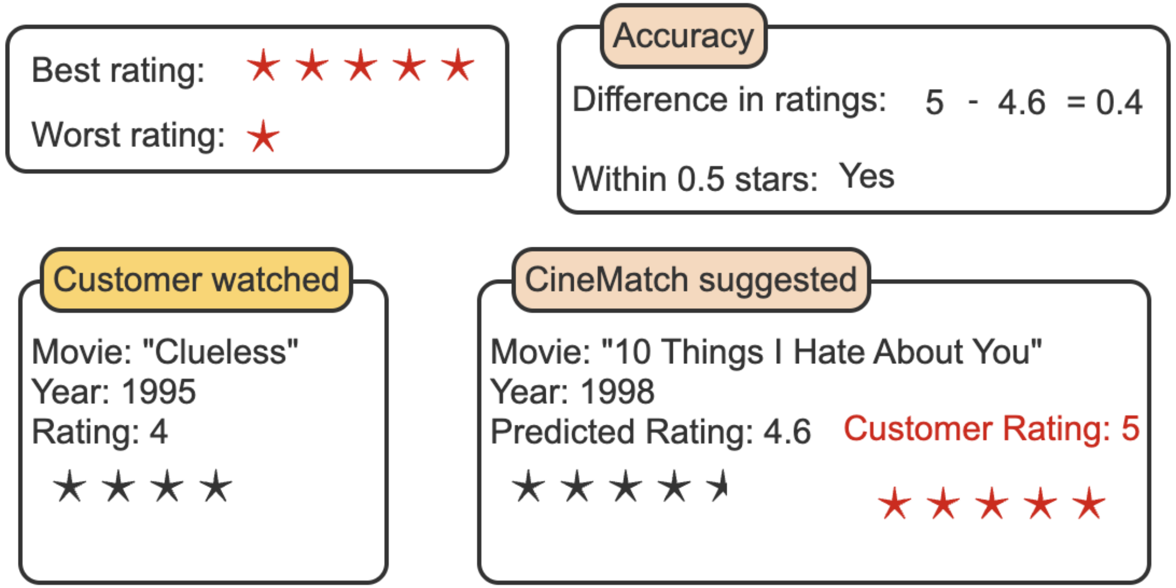
- Netflix began in 1997 as a DVD rental-by-mail service.
- In 2000, Netflix switched to a monthly subscription model while still emphasizing DVD rentals.
- They used an algorithm named CineMatch to suggest movies based on customer ratings.
- CineMatch used past ratings to predict movies that customers might enjoy.

Case Study: Netflix

- Netflix users rated movies from 1 to 5 stars.

- Using past ratings, CineMatch predicted ratings for new movies, making recommendations.

- For example, if a user liked a 1990s comedy, CineMatch suggested similar comedies like "10 Things I Hate About You," accurately predicting ratings about 75% of the time.



Best rating: ★★★★★

Worst rating: ★

Accuracy

Difference in ratings: 5 - 4.6 = 0.4

Within 0.5 stars: Yes

Customer watched

Movie: "Clueless"
Year: 1995
Rating: 4
★★★★

CineMatch suggested

Movie: "10 Things I Hate About You"
Year: 1998
Predicted Rating: 4.6
★★★★⯪

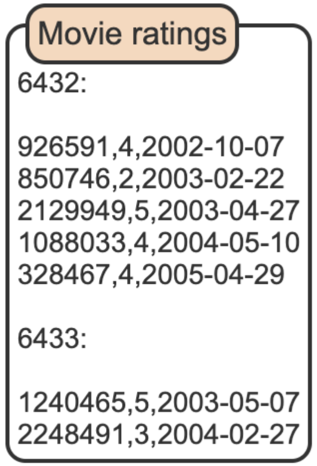
Customer Rating: 5
★★★★★

Netflix Prize

- In 2006, Netflix launched the Netflix Prize contest to enhance CineMatch ratings.
- A $1 million reward was promised to those boosting CineMatch's performance by 10%.
- The competition shared a dataset with 17,770 movies, 480,189 users, ratings (1-5 stars), and watch dates.
- Participants aimed to create better models for predicting user preferences.

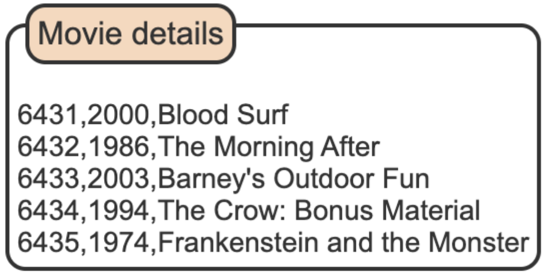
Netflix Prize

- Netflix provided dataset for Netflix Prize as two text files: one with movie ratings and another with movie names.
- Ratings were structured as movie ID, followed by customer ID, rating, and date.
- A separate file had movie ID, release year, and title.
- Teams often organized the data into a structured format before analysis.



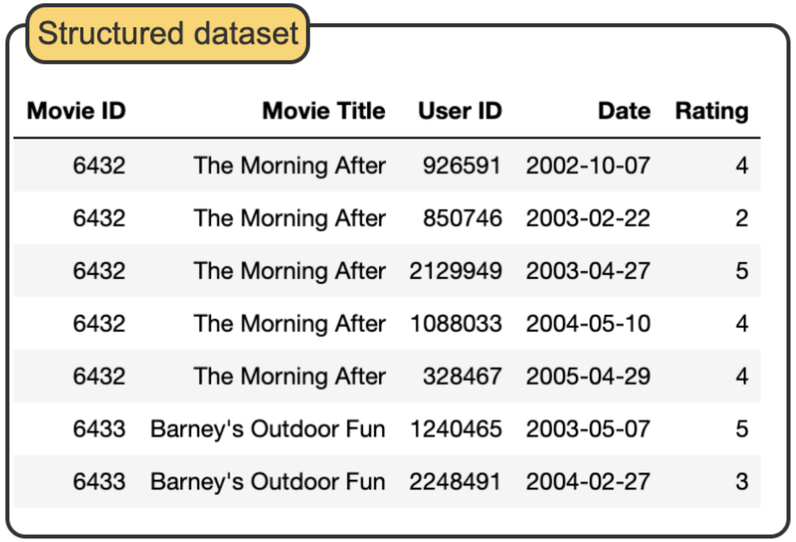
Movie ratings

```
6432:

926591,4,2002-10-07
850746,2,2003-02-22
2129949,5,2003-04-27
1088033,4,2004-05-10
328467,4,2005-04-29

6433:

1240465,5,2003-05-07
2248491,3,2004-02-27
```



Movie details

```
6431,2000,Blood Surf
6432,1986,The Morning After
6433,2003,Barney's Outdoor Fun
6434,1994,The Crow: Bonus Material
6435,1974,Frankenstein and the Monster
```



Structured dataset

Movie ID	Movie Title	User ID	Date	Rating
6432	The Morning After	926591	2002-10-07	4
6432	The Morning After	850746	2003-02-22	2
6432	The Morning After	2129949	2003-04-27	5
6432	The Morning After	1088033	2004-05-10	4
6432	The Morning After	328467	2005-04-29	4
6433	Barney's Outdoor Fun	1240465	2003-05-07	5
6433	Barney's Outdoor Fun	2248491	2004-02-27	3

The Netflix Dataset

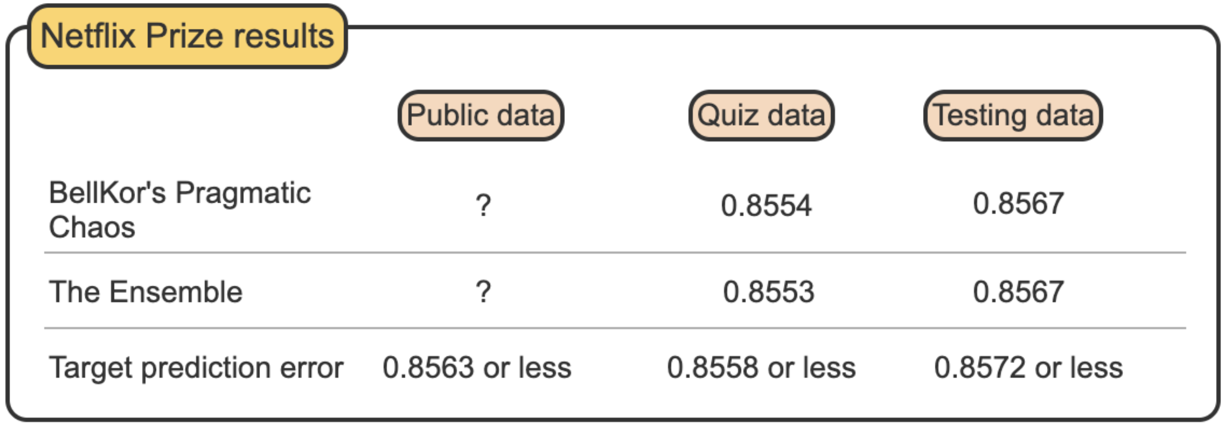
Challenges in the Netflix dataset included:

- Varying numbers of ratings per customer, from few to many.
- Uneven movie ratings, with some having numerous and others few.
- Limited movie details, only names and release years.

In 2009, BellKor's Pragmatic Chaos achieved the 10% goal, starting a 30-day submission period. The Ensemble's late submission led to BellKor's victory.

Netflix Prize Results

- BellKor and The Ensemble built models with public datasets. Netflix used larger customer datasets for evaluation.
- Two datasets, quiz and test, had distinct customer ratings, influencing target prediction errors.
- Quiz scores were displayed on the leaderboard. The Ensemble had a slightly better quiz score.
- The test dataset determined the ultimate winner. Final teams had near-identical performance.



Netflix Prize results

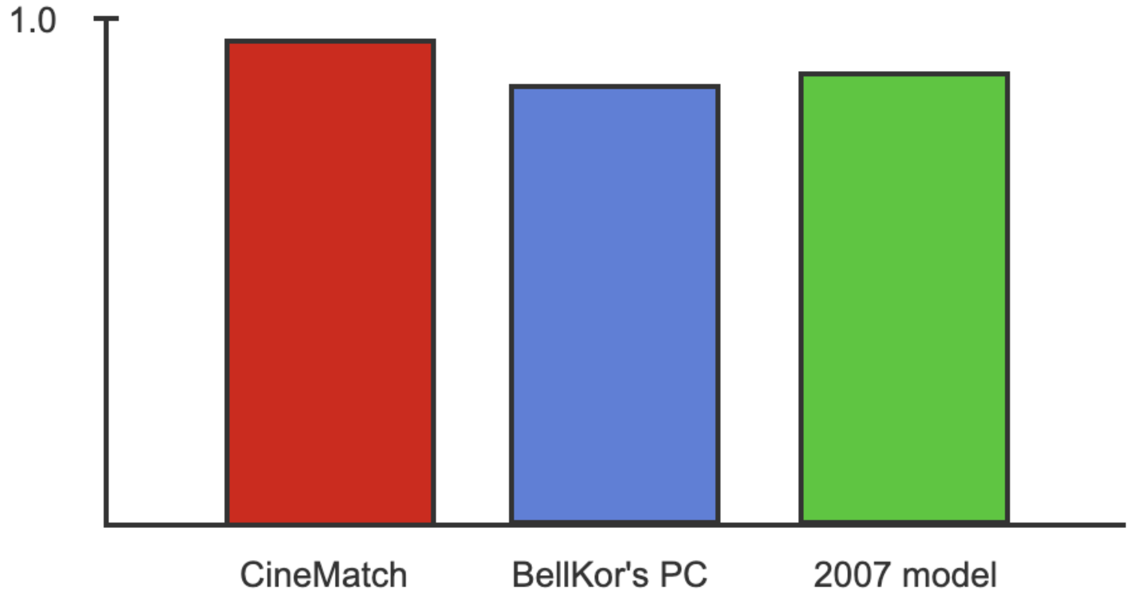
	Public data	Quiz data	Testing data
BellKor's Pragmatic Chaos	?	0.8554	0.8567
The Ensemble	?	0.8553	0.8567
Target prediction error	0.8563 or less	0.8558 or less	0.8572 or less

The Winning Algorithm

- Netflix assessed teams' predictions using RMSE, favoring lower values.
- BellKor's Pragmatic Chaos complex algorithm wasn't practical due to data volume.
- With 5 billion user ratings, it couldn't run effectively.
- A simpler model with slightly less improvement was adopted.

The Winning Algorithm

- Netflix's CineMatch had an RMSE of 0.9525.
- BellKor's Pragmatic Chaos cut RMSE to 0.8554, a 10.10% drop.
- Netflix chose a simpler 2007 model for its site despite higher error.

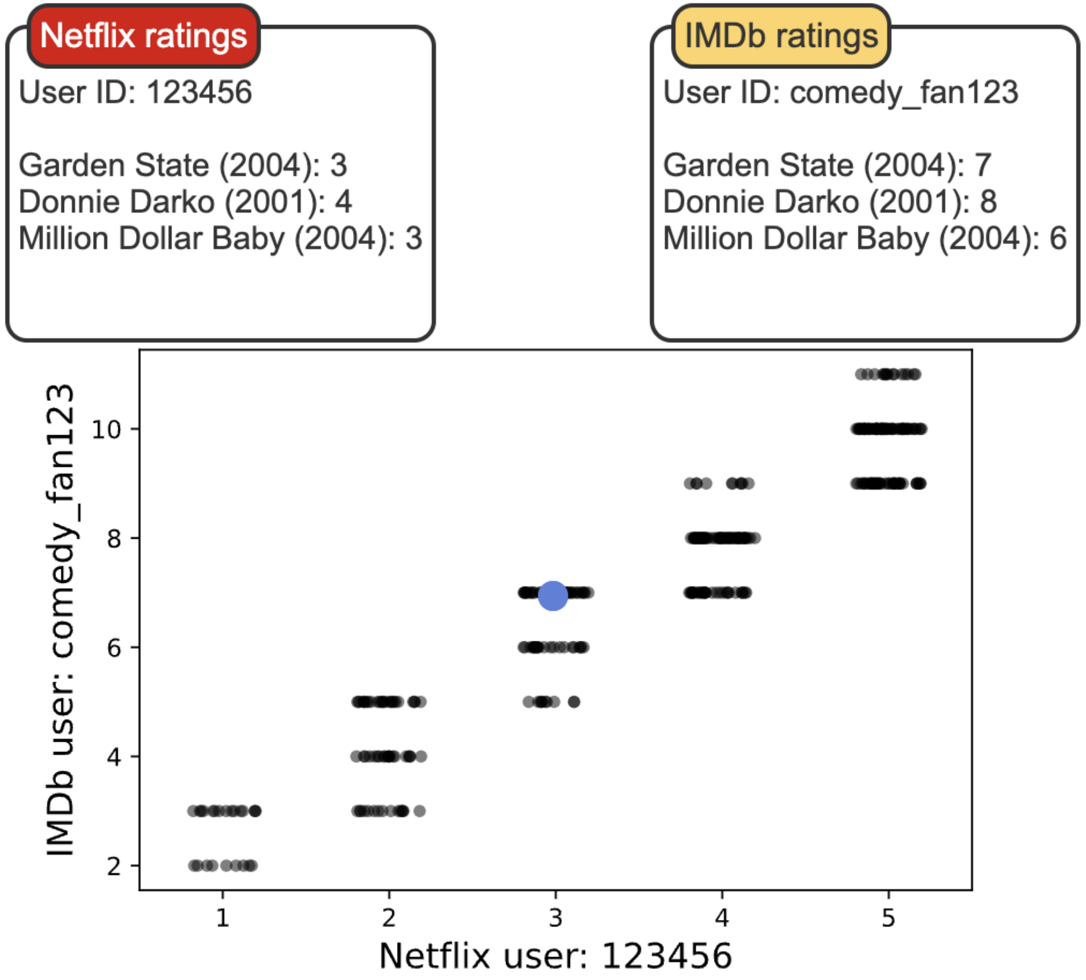


Legacy of The Netflix Prize

- The Netflix Prize spurred recommender system advancements, now widely used by online platforms.
- Netflix employed this to learn user preferences, leading to original content creation.
- Privacy issues arose from the Prize, with a lawsuit over potential outing of users.
- Concerns and consent issues halted a planned follow-up competition.

Privacy Issues

- Netflix shared ratings anonymously using random ID numbers, not personal info.
- Privacy worries persisted, like matching Netflix and IMDb ratings.
- Matching ratings could indicate shared users, raising identity concerns.



Netflix ratings

User ID: 123456

Garden State (2004): 3
Donnie Darko (2001): 4
Million Dollar Baby (2004): 3

IMDb ratings

User ID: comedy_fan123

Garden State (2004): 7
Donnie Darko (2001): 8
Million Dollar Baby (2004): 6

Questions

1. CineMatch predicted that a customer would rate "Titanic" (1997) 3.2 stars. The customer watched "Titanic" and rated the movie 3 stars. Find the prediction error.

2. CineMatch accurately predicted a customer's movie rating within 0.5 stars 75% of the time. How often did CineMatch fail to predict a customer's rating within 0.5 stars?

3. The datasets Netflix provided to competing teams were ______.

Next Lecture

Programming with Python and R