

Defining Data Science

© IBM Corporation. All rights reserved.

What Data Science is

Data scientists' role in an organization

What makes a skilled data scientist

Advice on acquiring these skills



What is Data Science?



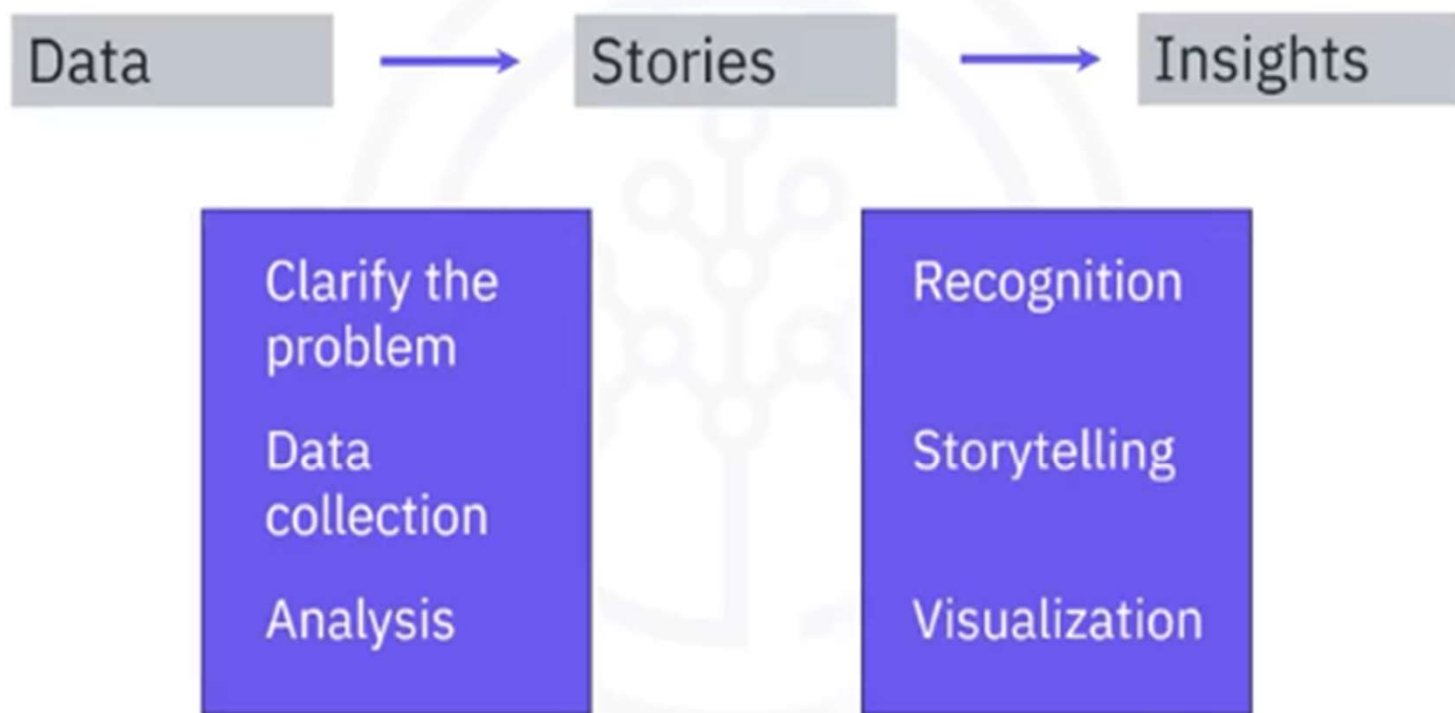
Data Science: The study of data to understand the world

An art of uncovering insights and trends

Data access drives new insights

And the needed computing power to analyze it

Data Scientist's role in an organization



Experts Opinion: Qualities of a data scientist



Curiosity



Sound argumentation



Good judgment

Makings of a skilled data scientist

- Versatility
- Subject area knowledge
- Experience programming and analyzing data

- Comfortable with math
- Curious
- Storyteller

- Diverse background
- Adept at selecting suitable tools
- Apply expertise to problem-solving

An ever-evolving field

The role will continue to evolve

May require certifications

Always will need to:

Think logically, algorithmically, and methodically

Gather data

Carefully analyze



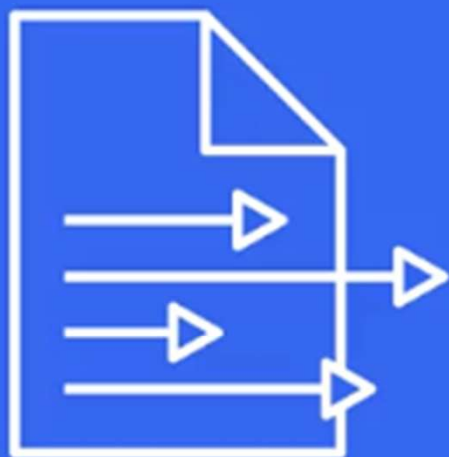
Understanding different types of file formats

You will be working with a variety of data file types and formats. It is important to understand the underlying structure of file formats along with their benefits and limitations. This understanding will support you to make the right decisions on the formats best suited for your data and performance needs.

Standard file formats:

1. Delimited text file formats, or .CSV
2. Microsoft Excel Open .XML Spreadsheet, or .XLSX
3. Extensible Markup Language, or .XML
4. Portable Document Format, or .PDF
5. JavaScript Object Notation, or .JSON

Delimited text files



Files used to store data as text

Each value is separated by a delimiter

Delimiter - A sequence of one or more characters for specifying the boundary between independent entities or values.

Comma, Tab, Colon, Vertical Bar, Space



Comma-separated values



Tab-separated values

Delimited text files

```
Manufacturer, Model, Sales_in_thousands, __year_resale_value, Vehicle_type, Price_in_thousands
Acura, Integra, 16.919, 16.36, Passenger, 21.5
Acura, TL, 39.384, 19.875, Passenger, 28.4
Acura, CL, 14.114, 18.225, Passenger, 14
Acura, RL, 8.588, 29.725, Passenger, 42
Audi, A4, 20.397, 22.255, Passenger, 23.99
Audi, A6, 18.78, 23.555, Passenger, 33.95
Audi, A8, 1.38, 39, Passenger, 62
BMW, 323i, 19.747, Passenger, 26.99
BMW, 328i, 9.231, 28.675, Passenger, 33.4
BMW, 528i, 17.527, 36.125, Passenger, 38.9
Buick, Century, 91.561, 12.475, Passenger, 21.975
```

.CSV

Delimited text files

```
Manufacturer,Model,Sales_in_thousands,___year_resale_value,Vehicle_type,Price_in_thousands
Acura, Integra, 16.919, 16.36, Passenger, 21.5
Acura, TL, 39.384, 19.875, Passenger, 28.4
Acura, CL, 14.114, 18.225, Passenger, 14
Acura, RL, 8.588, 29.725, Passenger, 42
Audi, A4, 20.397, 22.255, Passenger, 23.99
Audi, A6, 18.78, 23.555, Passenger, 33.95
Audi, A8, 1.38, 39, Passenger, 62
BMW, 323i, 19.747, Passenger, 26.99
BMW, 328i, 9.231, 28.675, Passenger, 33.4
BMW, 528i, 17.527, 36.125, Passenger, 38.9
Buick, Century, 91.561, 12.475, Passenger, 21.975
```

.CSV

.TSV

Manufacturer	Model	Sales_in_thousands	___year_resale_value
Acura	Integra	16.919	16.36 Passenger 21.5
Acura	TL	39.384	19.875 Passenger 28.4
Acura	CL	14.114	18.225 Passenger
Acura	RL	8.588	29.725 Passenger 42
Audi	A4	20.397	22.255 Passenger 23.99
Audi	A6	18.78	23.555 Passenger 33.95
Audi	A8	1.38	39 Passenger 62
BMW	323i	19.747	Passenger 26.99
BMW	328i	9.231	28.675 Passenger 33.4
BMW	528i	17.527	36.125 Passenger 38.9
Buick	Century	91.561	12.475 Passenger 21.975

Delimiters also represent one of various means to specify boundaries in a data stream

Microsoft Excel Open XML Spreadsheet, or .XLSX

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK	DL	DM	DN	DO	DP	DQ	DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK	EL	EM	EN	EO	EP	EQ	ER	ES	ET	EU	EV	EW	EX	EY	EZ	FA	FB	FC	FD	FE	FF	FG	FH	FI	FJ	FK	FL	FM	FN	FO	FP	FQ	FR	FS	FT	FU	FV	FW	FX	FY	FZ	GA	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV	GW	GX	GY	GZ	HA	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK	HL	HM	HN	HO	HP	HQ	HR	HS	HT	HU	HV	HW	HX	HY	HZ	IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO	IP	IQ	IR	IS	IT	IU	IV	IW	IX	IY	IZ	JA	JB	JC	JD	JE	JF	JG	JH	JI	IJ	JK	KL	KM	KN	KO	KP	KQ	KR	KS	KT	KU	KV	KW	KX	KY	KZ	LA	LB	LC	LD	LE	LF	LG	LH	LI	LJ	LK	LM	LN	LO	LP	LQ	LR	LS	LT	LU	LV	LW	LX	LY	LZ	MA	MB	MC	MD	ME	MF	MG	MH	MI	MJ	MK	ML	MM	MN	MO	MP	MQ	MR	MS	MT	MU	MV	MW	MX	MY	MZ	NA	NB	NC	ND	NE	NF	NG	NH	NI	NJ	NK	NL	NM	NN	NO	NP	NQ	NR	NS	NT	NU	NV	NW	NX	NY	NZ	OA	OB	OC	OD	OE	OF	OG	OH	OI	OJ	OK	OL	OM	ON	OO	OP	OQ	OR	OS	OT	OU	OV	OW	OX	OY	OZ	PA	PB	PC	PD	PE	PF	PG	PH	PI	PJ	PK	PL	PM	PN	PO	PP	PQ	PR	PS	PT	PU	PV	PW	PX	PY	PZ	QA	QB	QC	QD	QE	QF	QG	QH	QI	QJ	QK	QL	QM	QN	QO	QP	QR	QS	QT	QU	QV	QW	QX	QY	QZ	RA	RB	RC	RD	RE	RF	RG	RH	RI	RJ	RK	RL	RM	RN	RO	RP	RQ	RR	RS	RT	RU	RV	RW	RX	RY	RZ	SA	SB	SC	SD	SE	SF	SG	SH	SI	SJ	SK	SL	SM	SN	SO	SP	SQ	SR	SS	ST	SU	SV	SW	SX	SY	SZ	TA	TB	TC	TD	TE	TF	TG	TH	TI	TJ	TK	TL	TM	TN	TO	TP	TQ	TR	TS	TT	TU	TV	TW	TX	TY	TZ	UA	UB	UC	UD	UE	UF	UG	UH	UI	UJ	UK	UL	UM	UN	UO	UP	UQ	UR	US	UT	UU	UV	UW	UX	UY	UZ	VA	VB	VC	VD	VE	VF	VG	VH	VI	VJ	VK	VL	VM	VN	VO	VP	VQ	VR	VS	VT	VU	VV	VW	VX	VY	VZ	WA	WB	WC	WD	WE	WF	WG	WH	WI	WJ	WK	WL	WM	WN	WO	WP	WQ	WR	WS	WT	WU	WV	WW	WX	WY	WZ	XA	XB	XC	XD	XE	XF	YG	YH	YI	YJ	YK	YL	YM	YN	YO	YP	YQ	YR	YS	YT	YU	YV	YW	YX	YZ	ZA	ZB	ZC	ZD	ZE	ZF	ZG	ZH	ZI	ZJ	ZK	ZL	ZM	ZN	ZO	ZP	ZQ	ZR	ZS	ZT	ZU	ZV	ZW	ZX	ZY	ZZ	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR	BS	BT	BU	BV	BW	BX	BY	BZ	CA	CB	CC	CD	CE	CF	CG	CH	CI	CJ	CK	CL	CM	CN	CO	CP	CQ	CR	CS	CT	CU	CV	CW	CX	CY	CZ	DA	DB	DC	DD	DE	DF	DG	DH	DI	DJ	DK	DL	DM	DN	DO	DP	DQ	DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC	ED	EE	EF	EG	EH	EI	EJ	EK	EL	EM	EN	EO	EP	EQ	ER	ES	ET	EU	EV	EW	EX	EY	EZ	FA	FB	FC	FD	FE	FF	FG	FH	FI	FJ	FK	FL	FM	FN	FO	FP	FQ	FR	FS	FT	FU	FV	FW	FX	FY	FZ	GA	GB	GC	GD	GE	GF	GG	GH	GI	GJ	GK	GL	GM	GN	GO	GP	GQ	GR	GS	GT	GU	GV	GW	GX	GY	GZ	HA	HB	HC	HD	HE	HF	HG	HH	HI	HJ	HK	HL	HM	HN	HO	HP	HQ	HR	HS	HT	HU	HV	HW	HX	HY	HZ	IA	IB	IC	ID	IE	IF	IG	IH	II	IJ	IK	IL	IM	IN	IO	IP	IQ	IR	IS	IT	IU	IV	IW	IX	IY	IZ	JA	JB	JC	JD	JE	JF	JG	JH	JI	IJ	JK	KL	KM	KN	KO	KP	KQ	KR	KS	KT	KU	KV	KW	KX	KY	KZ	LA	LB	LC	LD	LE	LF	LG	LH	LI	LJ	LK	LM	LN	LO	LP	LQ	LR	LS	LT	LU	LV	LW	LX	LY	LZ	MA	MB	MC	MD	ME	MF	MG	MH	MI	MJ	MK	ML	MM	MN	MO	MP	MQ	MR	MS	MT	MU	MV	MW	MX	MY	MZ	NA	NB	NC	ND	NE	NF	NG	NH	NI	NJ	NK	NL	NM	NN	NO	NP	NQ	NR	NS	NT	NU	NV	NW	NX	NY	NZ	OA	OB	OC	OD	OE	OF	OG	OH	OI	OJ	OK	OL	OM	ON	OO	OP	OQ	OR	OS	OT	OU	OV	OW	OX	OY	OZ	PA	PB	PC	PD	PE	PF	PG	PH	PI	PJ	PK	PL	PM	PN	PO	PP	PQ	PR	PS	PT	PU	PV	PW	PX	PY	PZ	QA	QB	QC	QD	QE	QF	QG	QH	QI	QJ	QK	QL	QM	QN	QO	QP	QR	QS	QT	QU	QV	QW	QX	QY	QZ	RA	RB	RC	RD	RE	RF	RG	RH	RI	RJ	RK	RL	RM	RN	RO	RP	RQ	RR	RS	RT	RU	RV	RW	RX	RY	RZ	SA	SB	SC	SD	SE	SF	SG	SH	SI	SJ	SK	SL	SM	SN	SO	SP	SQ	SR	SS	ST	SU	SV	SW	SX	SY	SZ	TA	TB	TC	TD	TE	TF	TG	TH	TI	TJ	TK	TL	TM	TN	TO	TP	TQ	TR	TS	TT	TU	TV	TW	TX	TY	TZ	UA	UB	UC	UD	UE	UF	UG	UH	UI	UJ	UK	UL	UM	UN	UO	UP	UQ	UR	US	UT	UU	UV	UW	UX	UY	UZ	VA	VB	VC	VD	VE	VF	VG	VH	VI	VJ	VK	VL	VM	VN	VO	VP	VQ	VR	VS	VT	VU	VV	VW	VX	VY	VZ	WA	WB	WC	WD	WE	WF	WG	WH	WI	WJ	WK	WL	WM	WN	WO	WP	WQ	WR	WS	WT	WU	WV	WW	WX	WY	WZ	XA	XB	XC	XD	XE	XF	YG	YH	YI	YJ	YK	YL	YM	YN	YO	YP	YQ	YR	YS	YT	YU	YV	YW	YX	YZ	ZA	ZB	ZC	ZD	ZE	ZF	ZG	ZH	ZI	ZJ	ZK	ZL	ZM	ZN	ZO	ZP	ZQ	ZR	ZS	ZT	ZU	ZV	ZW	ZX	ZY	ZZ
Manufacturer	Model	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount	Sales_Amount																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																													

- Open file format, accessible to most other applications
- Can use and save all functions available in excel
- Is a secure file format as it cannot save malicious code

Extensible Markup Language or .XML

```
<?xml version="1.0"?>
<car-specs>

<manufacturer>Acura</manufacturer>
<model>Integra</model>
<sales_in-thousands>16.919</sales_in-thousands>
<year_resale_value>16.36</year_resale_value>
<vehicle_type>passenger</vehicle_type>
</car-specs>
```

Extensible Markup Language, or XML, is a markup language with set rules for encoding data.

- Readable by both humans and machines
- Self-descriptive language
- Similar to .HTML in some respects
- Does not use predefined tags like .HTML does
- Platform independent
- Programming language independent
- Makes it simpler to share data between systems

Portable Document Format or PDF

USAID
U.S. AGENCY FOR INTERNATIONAL DEVELOPMENT

Form No. 100-100
Revision Date: 02/2010

APPLICATION FOR APPROVAL OF COMMODITY ELIGIBILITY
OF 100-100-100

Transaction No. (Assigned by USAID)

COMMODITY INFORMATION

1. Commodity Name
2. Commodity Code
3. Commodity Description
4. Commodity Quantity
5. Commodity Unit of Measure
6. Commodity Origin
7. Commodity Status
8. Commodity Remarks

COMMODITY ORIGIN

9. Country of Origin
10. Country of Production
11. Country of Assembly
12. Country of Origin of Components

COMMODITY ELIGIBILITY

13. Commodity Eligibility
14. Commodity Eligibility Remarks
15. Commodity Eligibility Date

Portable Document Format, or PDF, is a file format developed by Adobe to present documents independent of application software, hardware, and operating systems.

- Can be viewed the same way on any device
- Is frequently used in legal and financial documents
- Can also be used to fill in data for forms

JavaScript Object Notation or JSON

```
{  
  "employee": [  
    {  
      "id": "1",  
      "Manufacturer": "Audi",  
      "Model": "Integra",  
    },  
    {  
      "id": "2",  
      "Manufacturer": "Buick",  
      "Model": "LeSabre",  
    },  
    {  
      "id": "3",  
      "Manufacturer": "Cadillac",  
      "Model": "Escalade",  
    }  
  ]  
}
```

JavaScript Object Notation, or JSON, is a text-based open standard designed for transmitting structured data over the web.

- Language-independent data format
- Can be read in any programming language
- Easy to use
- Compatible with a wide range of browsers
- Considered as one of the best tools for sharing data

What Do Data Scientists Do?

Lesson Review

© IBM Corporation. All rights reserved.

Problem investigation

Toronto public transit customer complaints



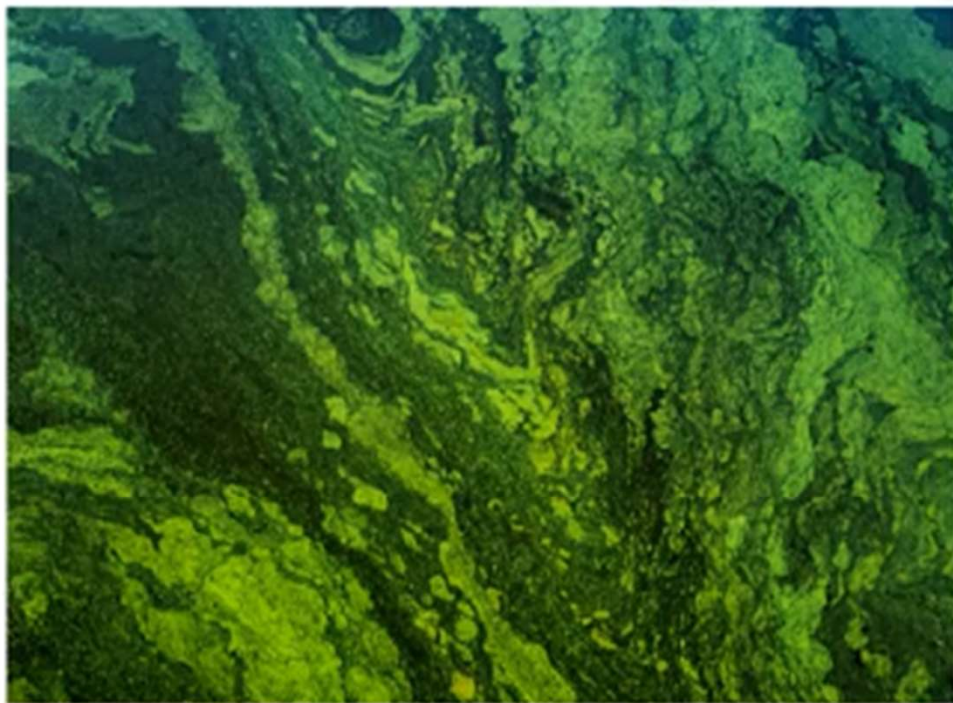
Unexpected bad weather days



More customer complaints

Environmental problem investigation

Predicting algae blooms to prevent water toxicity



Helping water treatment companies safeguard the water

Simplify problems

Dr. Norman White's
recommendation engine

Solving real-world issues
with innovative solutions



Education



Python notebooks

Linux

Databases

Pandas


Mathematical sciences

Algebra

Calculus

Probability

Statistics

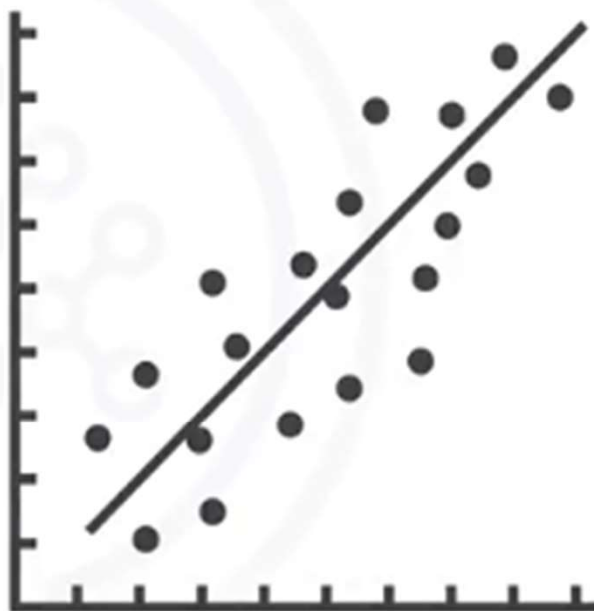


Not only a statistician

Models and algorithms

Regression

Relationship
between two
variables



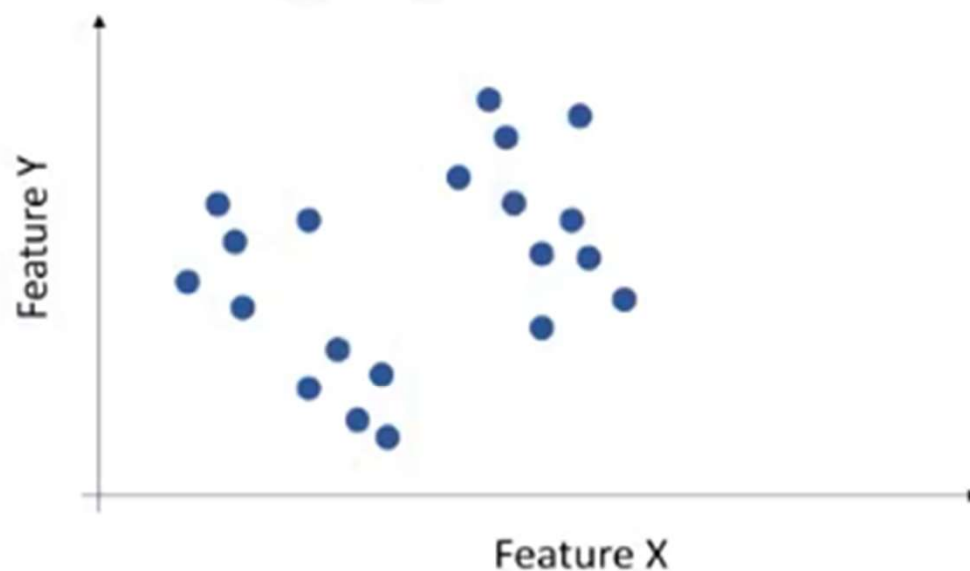
Models and algorithms

Regression

Relationship
between two
variables

Nearest neighbor

Machine learning
algorithm



Ability to analyze data

What is
“Big Data?”



Storytelling with data



Not the tools or the size
of the data set

Tell a compelling story
with data

About the data

Data sources

- Video
- Audio
- Text

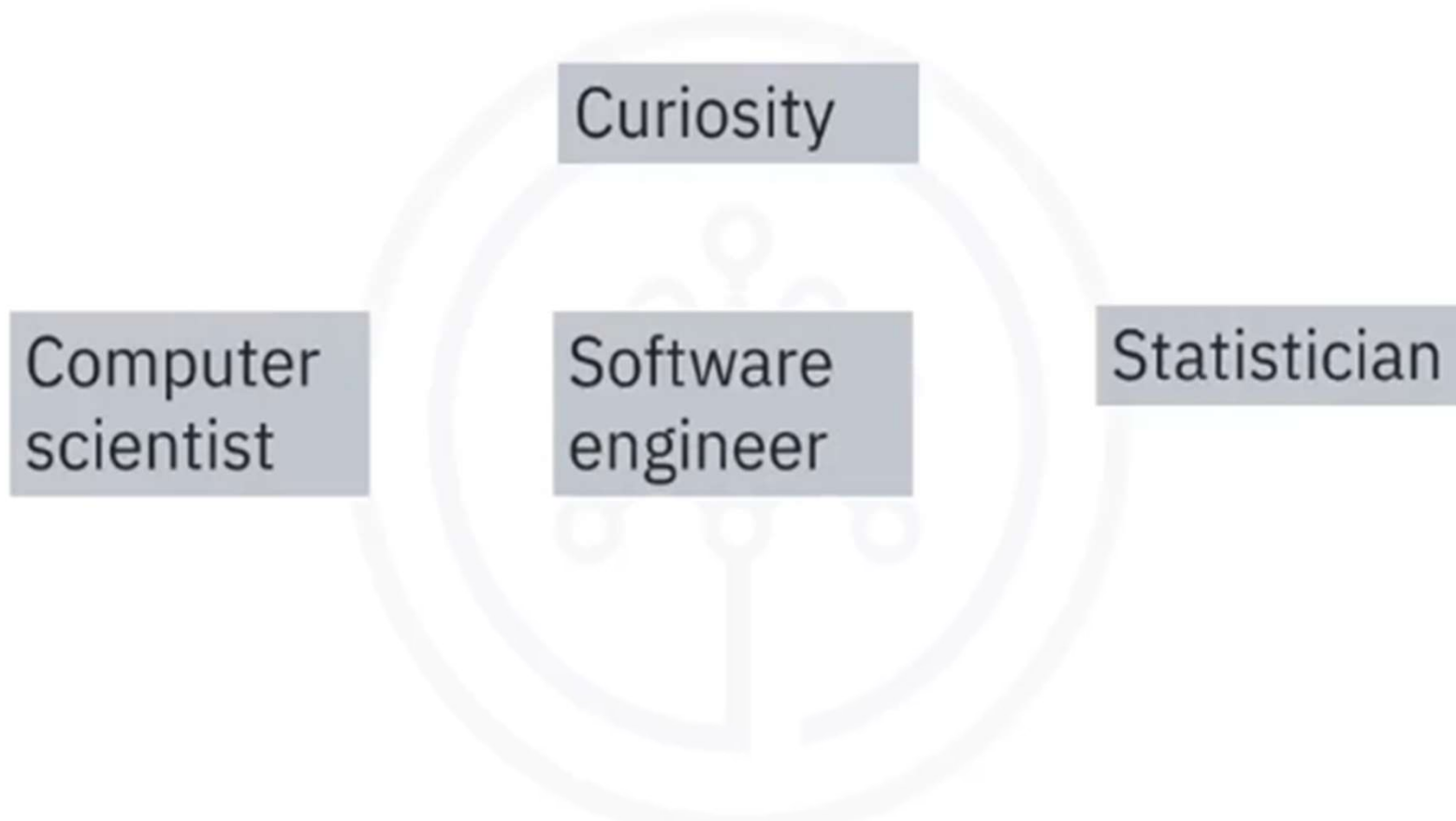
Data structure

- Structured
- Unstructured

Data formats

- Text files
- Spreadsheets
- XML
- PDFs
- JSON

An exceptional data scientist



It's a journey



Exploration

Innovation

Storytelling