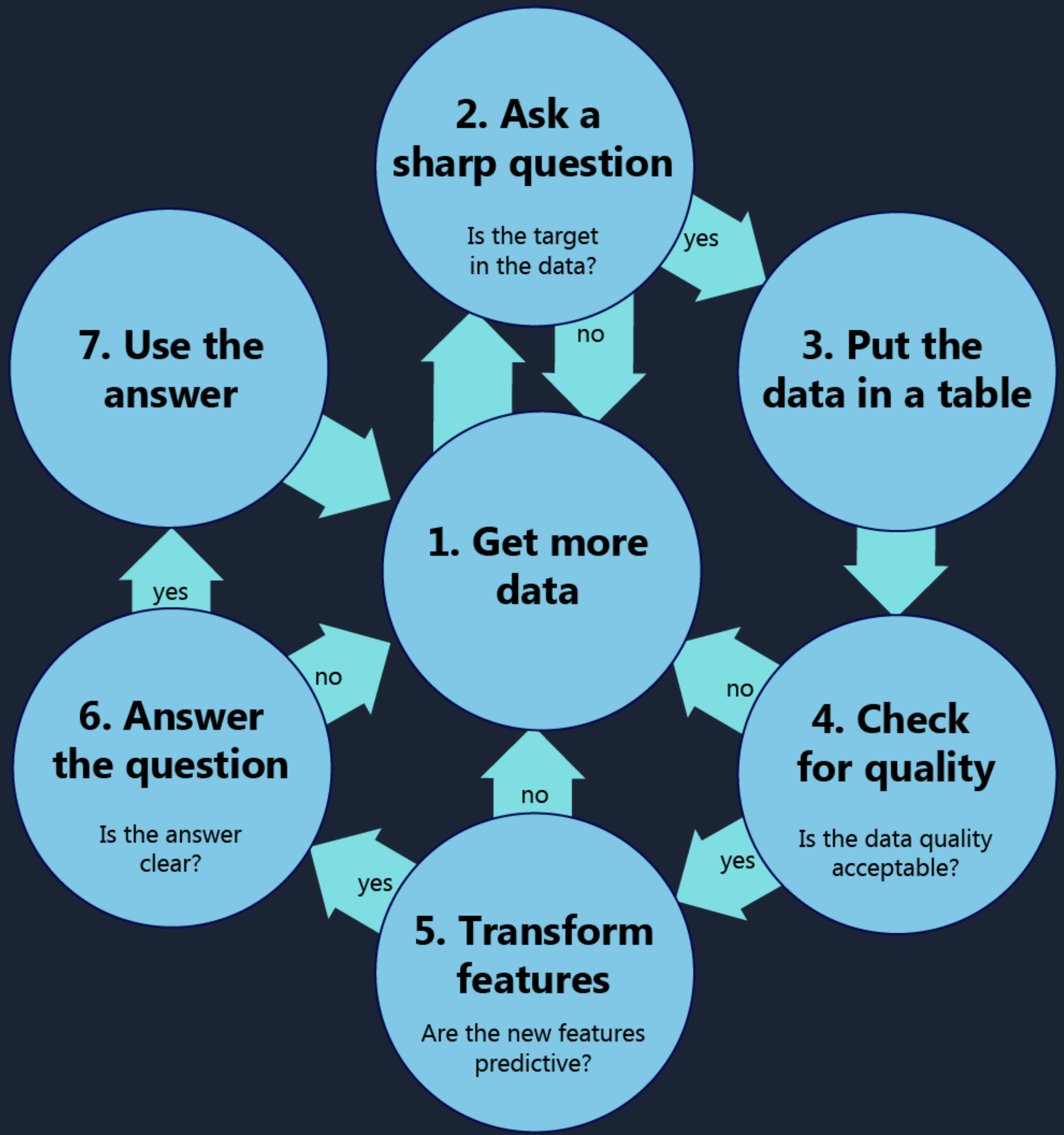
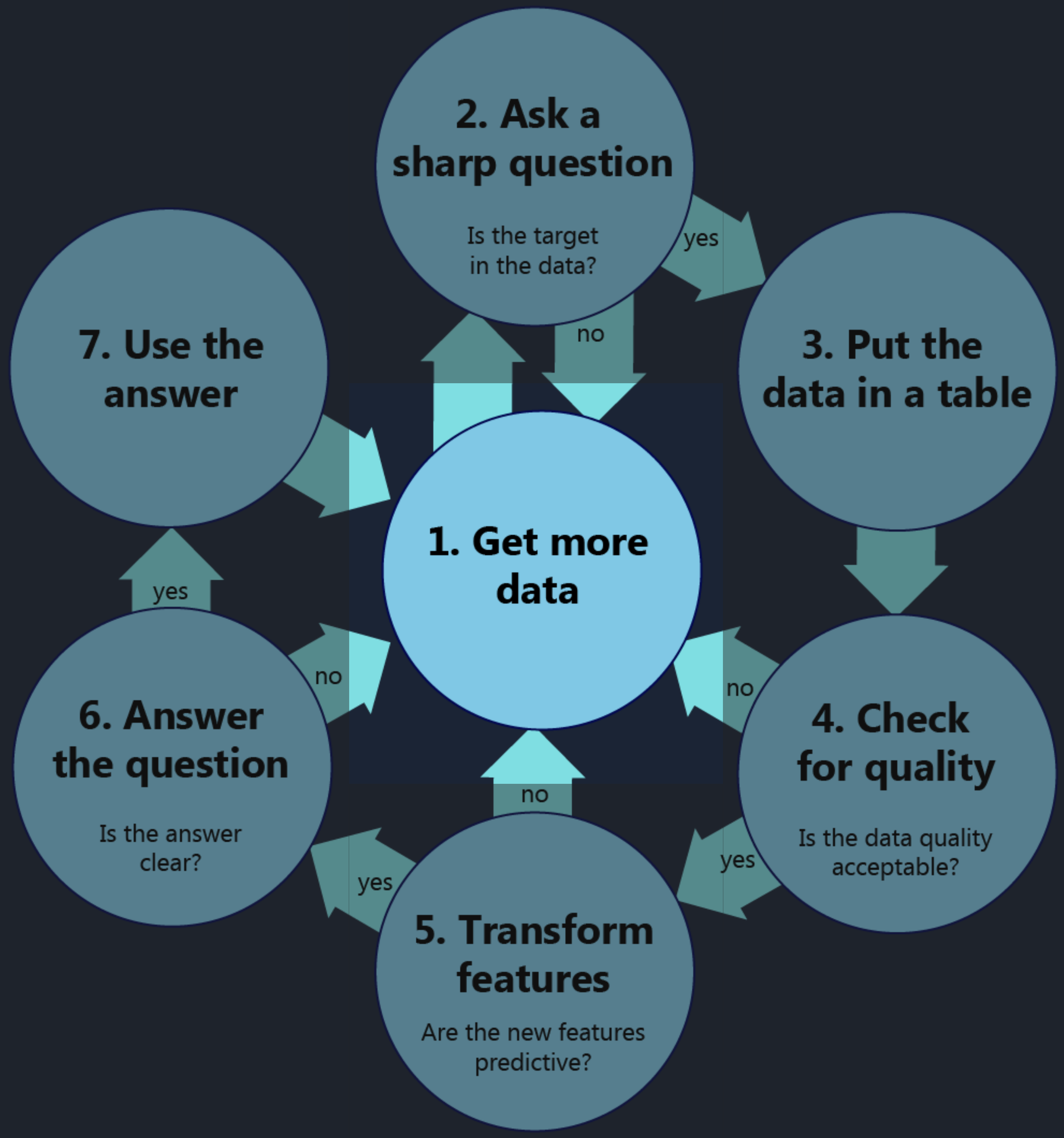
Data Science for Absolutely Everyone

Brandon Rohrer
Senior Data Scientist
Microsoft



1. Get more data
2. Ask a sharp question
Is the target in the data?
yes
no
3. Put the data in a table
4. Check for quality
Is the data quality acceptable?
no
yes
5. Transform features
Are the new features predictive?
no
yes
6. Answer the question
Is the answer clear?
no
yes
7. Use the answer



1. Get more data
2. Ask a sharp question
Is the target in the data?
yes
no
3. Put the data in a table
4. Check for quality
Is the data quality acceptable?
yes
no
5. Transform features
Are the new features predictive?
yes
no
6. Answer the question
Is the answer clear?
yes
no
7. Use the answer

Numbers and Names (Numerical and Categorical)

Numbers

Amount : 38.3 degrees

Count : 39 pizzas

Money : $1,387

Pixel brightness : 232/255

Sound intensity : .64

Names

Type : Shih Tzu

Variety : Caramel latte

ID : Air Force One

Model number : R2-D2

Category : Chocolate

Text : "Best. Show. Ever. <3"

Names that look like numbers

Phone number : 847-5609

Zip code : 90210

ID number : 007

Serial number : 100000184573

Credit card number : 5738-7539-9898-0023

Social security number : 627-42-0932

Numbers that look like names (ordinals) and names that can be turned into numbers

Place : first, second, third

Size : small, medium, large

Side : left, middle, right

Time zone : Pacific, Mountain, Central, Eastern

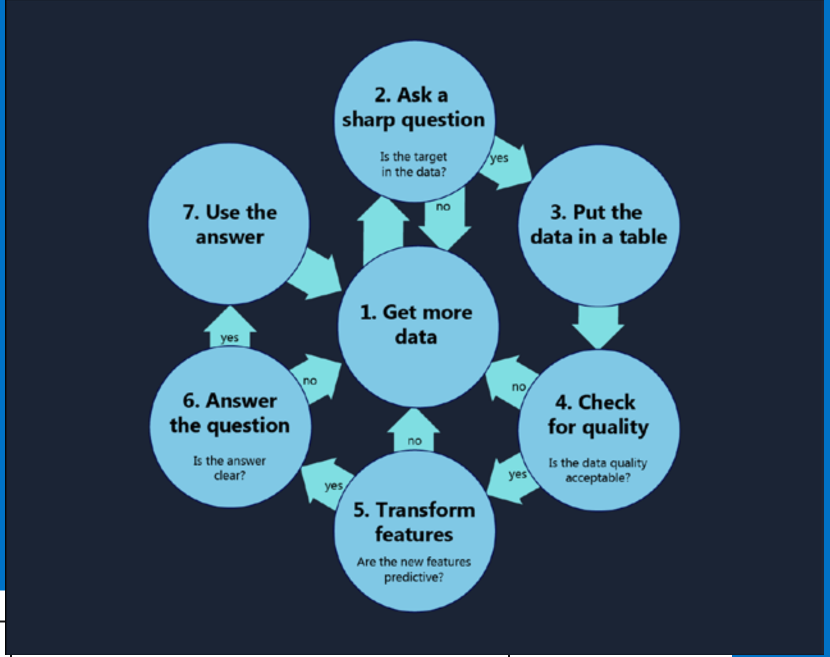
Train stops : Kendall, Central, Harvard, Porter

Data Engineering

Measure
Collect
Store
Search
Move
Transform

Azure Event Hub
Azure Stream Analytics
Azure Data Factory
Hadoop and Spark on Azure HDInsight
Azure Search
Azure DocumentDB
Azure Data Lake
Azure Data Catalog

Cortana Analytics Process



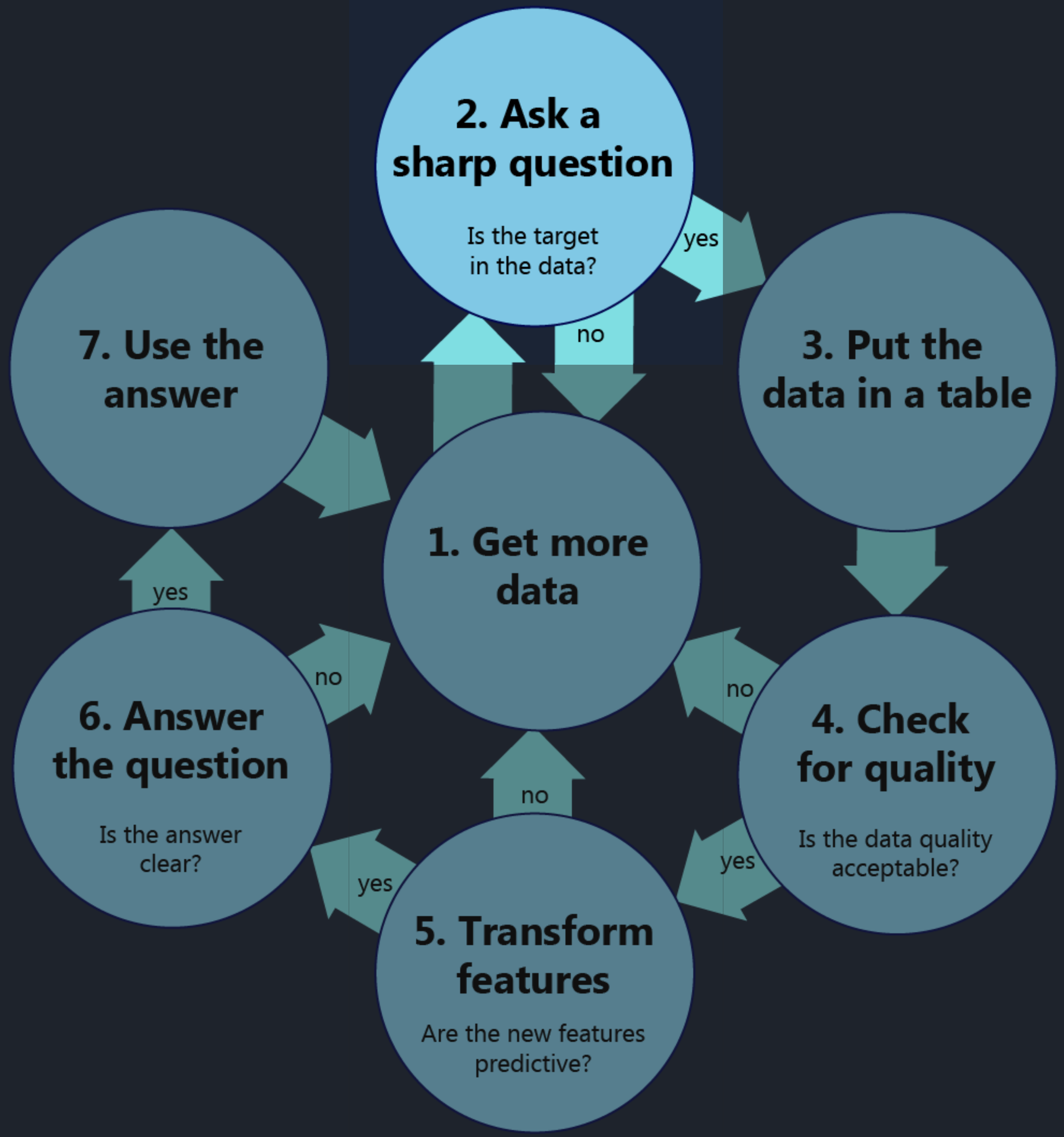
1. Ingest data
- Load data into storage environments
- Import data into Azure Machine Learning Studio

2. Explore and pre-process data
- Prepare data
- Explore data
- Sample data

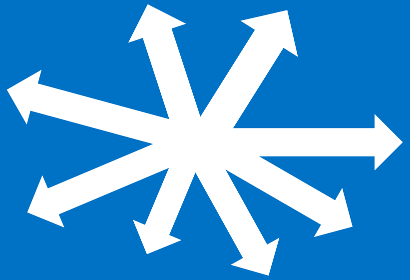
3. Create features
- Engineer features
- Select features
- Learn with counts

4. Create model
- Train the model
- Evaluate the model
- Tune the model

5. Deploy & consume model
- Publish model as a Web service
- Consume a model programmatically
- Consume a model in Excel


2. Ask a sharp question
Is the target in the data?
yes
no
3. Put the data in a table
7. Use the answer
1. Get more data
yes
no
no
4. Check for quality
Is the data quality acceptable?
yes
6. Answer the question
Is the answer clear?
no
yes
5. Transform features
Are the new features predictive?

Vague questions vs. Sharp questions



Doesn't have to be answered with a name or a number

What can my data tell me about my business?

What should I do?

How can I increase my profits?

Must be answered with a name or a number.

How many times will the feature I built get used by a new user?

Which route through downtown will get me to work the fastest?

Target

What will my stock price be next week?

Date	Americas sales	Europe and Africa sales	Asia sales

Competitor	Product	Market share

Product	First month users	First quarter users	First year users

Date	Dow Jones	Nikkei

Target

What will my stock price be next week?

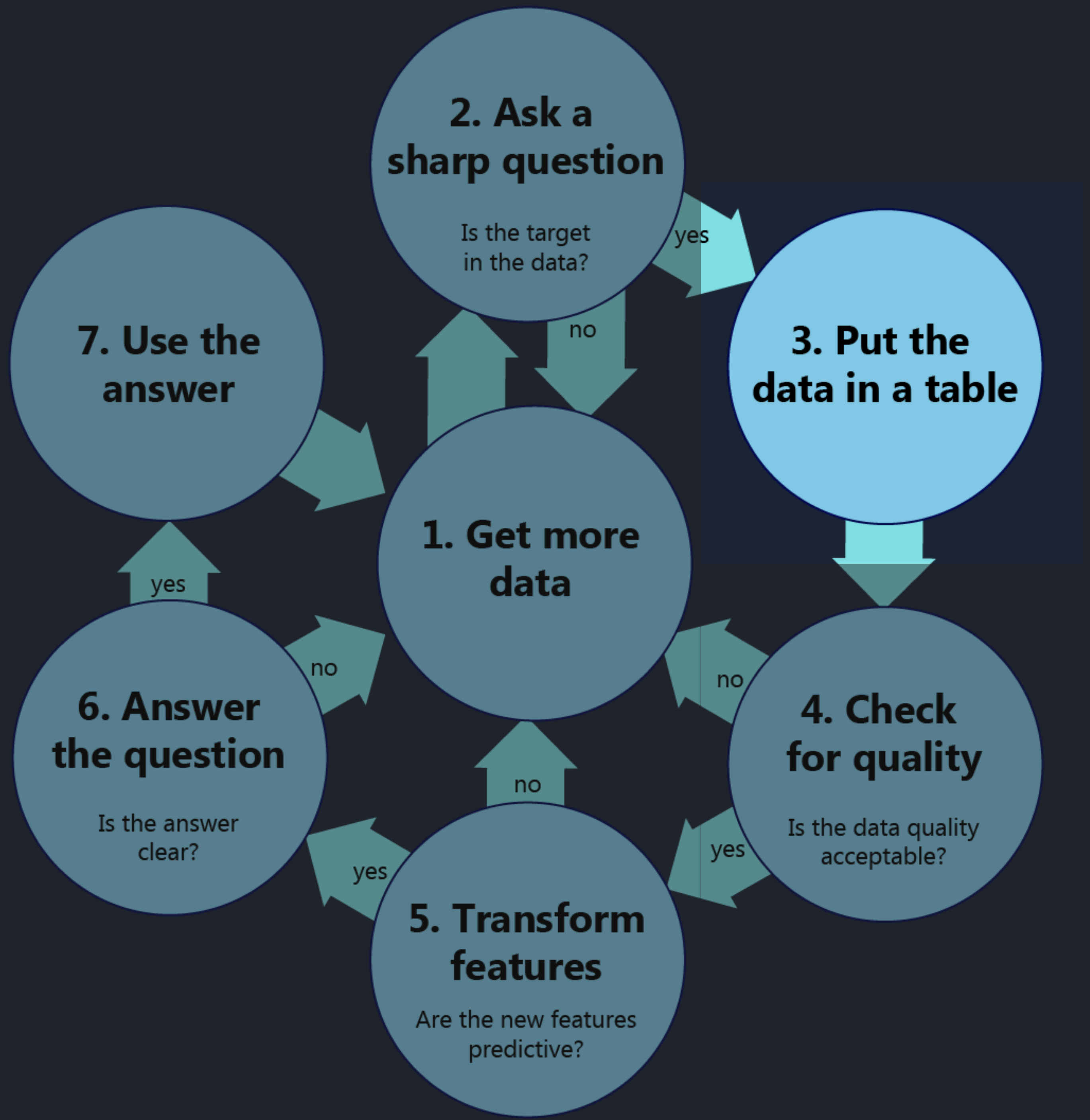
Date	My stock price

Date	Americas sales	Europe and Afri sales

Competitor

Product	First month users	First quarter users

Date


2. Ask a sharp question
Is the target in the data?
yes
no
3. Put the data in a table
7. Use the answer
1. Get more data
yes
no
no
6. Answer the question
Is the answer clear?
no
4. Check for quality
Is the data quality acceptable?
yes
yes
5. Transform features
Are the new features predictive?

One target per row

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

One target per row

Aggregate

User name	Date joined
little_lil	Jan 27, 2014
popoverGuy	Jan 27, 2014
Red_Red	Jan 28, 2014
David_G_53	Jan 30, 2014
randll	Jan 30, 2014
…	…

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

One target per row

Aggregate

Distribute

Month	Total sales
2016/01	43.0M
2016/02	60.1M
2016/03	55.5M
2016/04	41.7M
2016/05	68.8M
…	…

Quarter	Total sales
2015Q4	119.2M
2016Q1	221.0M
2016Q2	215.9M
2016Q3	189.3M
2016Q4	211.2M
…	…

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.493M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.494M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.494M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.495M

One target per row

Aggregate

Distribute

Compute

Press release date	Subject
2016/03/24	Mega amazing whizbang
2016/05/03	Super widget upgrade
2016/05/18	New gizmos on the flimflam
…	…

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

One target per row

Aggregate

Distribute

Compute

Measure

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

One target per row

Aggregate
Distribute
Compute

Measure
Estimate

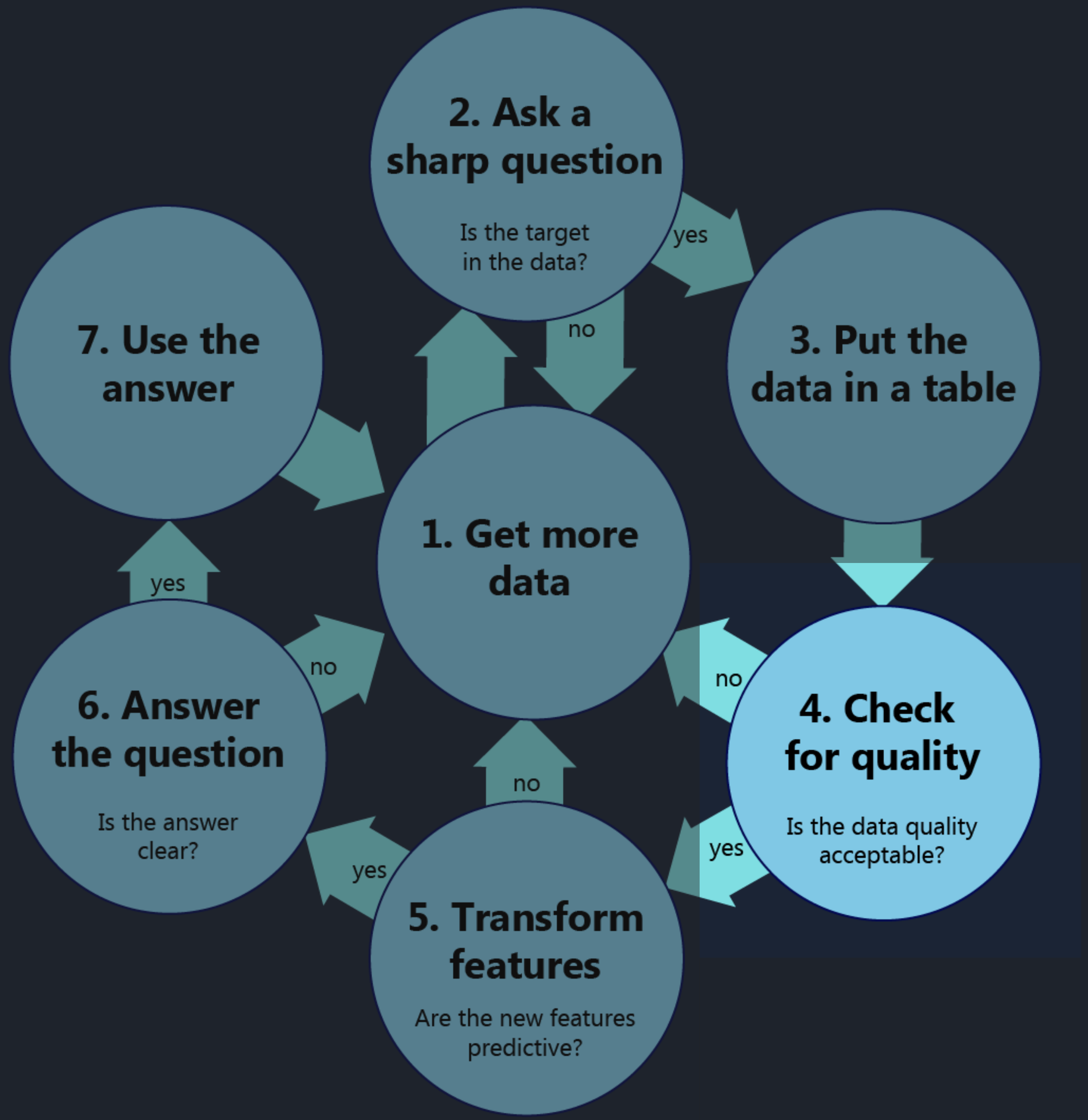
Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

One target per row

Aggregate
Distribute
Compute

Measure
Estimate
Leave blanks

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M


2. Ask a sharp question
Is the target in the data?
yes
no
3. Put the data in a table
7. Use the answer
1. Get more data
yes
no
no
4. Check for quality
Is the data quality acceptable?
6. Answer the question
Is the answer clear?
no
yes
yes
5. Transform features
Are the new features predictive?

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969*	6′ 2″	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	--1979--	5′ 11″	Manhattan		9	good	long
9471	Diana	Trevor	1618	5′ 8″	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	19.42	5′ 4″	Cresskill		tiny	Good	Not really
0696	Peter	Parker	1111983	5′ 10″	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	5′ 2″	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1-9-3-2	6′ 0″	Hamburg		Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	5′ 7″	St. Petersburg		jet	depends	No way
0323	Jean	Grey	“1977”	5′ 6″	Annandale		No	good	Mostly not
3980	Clark	Kent	“1954”	6′ 4″	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	“1943”	6′ 2″	Latveria		1	Bad	yes
0573	Stephen	Strange	1968	6′ 2″	Philidelphia		not	light	Y
7452	Thor	Odinson	2287 BC	6′ 6″	Norway		10	Good	Of course
1437	Selina	Kyle	1998	5′ 7″	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	..1911..	5′ 10″	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	5′ 7″	Krypton	Y	fast	G	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969*	6′ 2″	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	--1979--	5′ 11″	Manhattan		9	good	long
9471	Diana	Trevor	1618	5′ 8″	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	19.42	5′ 4″	Cresskill		tiny	Good	Not really
0696	Peter	Parker	1111983	5′ 10″	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	5′ 2″	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1-9-3-2	6′ 0″	Hamburg		Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	5′ 7″	St. Petersburg		jet	depends	No way
0323	Jean	Grey	“1977”	5′ 6″	Annandale		No	good	Mostly not
3980	Clark	Kent	“1954”	6′ 4″	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	“1943”	6′ 2″	Latveria		1	Bad	yes
0573	Stephen	Strange	1968	6′ 2″	Philidelphia		not	light	Y
7452	Thor	Odinson	2287 BC	6′ 6″	Norway		10	Good	Of course
1437	Selina	Kyle	1998	5′ 7″	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	..1911..	5′ 10″	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	5′ 7″	Krypton	Y	fast	G	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	6′ 2″	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	1979	5′ 11″	Manhattan		9	good	long
9471	Diana	Trevor	1618	5′ 8″	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	1942	5′ 4″	Cresskill		tiny	Good	Not really
0696	Peter	Parker	1983	5′ 10″	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	5′ 2″	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1932	6′ 0″	Hamburg		Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	5′ 7″	St. Petersburg		jet	depends	No way
0323	Jean	Grey	1977	5′ 6″	Annandale		No	good	Mostly not
3980	Clark	Kent	1954	6′ 4″	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	1943	6′ 2″	Latveria		1	Bad	yes
0573	Stephen	Strange	1968	6′ 2″	Philidelphia		not	light	Y
7452	Thor	Odinson	-2287	6′ 6″	Norway		10	Good	Of course
1437	Selina	Kyle	1998	5′ 7″	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	1911	5′ 10″	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	5′ 7″	Krypton	Y	fast	G	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	6′ 2″	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	1979	5′ 11″	Manhattan		9	good	long
9471	Diana	Trevor	1618	5′ 8″	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	1942	5′ 4″	Cresskill		tiny	Good	Not really
0696	Peter	Parker	1983	5′ 10″	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	5′ 2″	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1932	6′ 0″	Hamburg		Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	5′ 7″	St. Petersburg		jet	depends	No way
0323	Jean	Grey	1977	5′ 6″	Annandale		No	good	Mostly not
3980	Clark	Kent	1954	6′ 4″	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	1943	6′ 2″	Latveria		1	Bad	yes
0573	Stephen	Strange	1968	6′ 2″	Philidelphia		not	light	Y
7452	Thor	Odinson	-2287	6′ 6″	Norway		10	Good	Of course
1437	Selina	Kyle	1998	5′ 7″	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	1911	5′ 10″	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	5′ 7″	Krypton	Y	fast	G	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	74	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	1979	71	Manhattan		9	good	long
9471	Diana	Trevor	1618	68	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	1942	64	Cresskill		tiny	Good	Not really
0696	Peter	Parker	1983	70	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	62	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1932	72	Hamburg		Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	67	St. Petersburg		jet	depends	No way
0323	Jean	Grey	1977	66	Annandale		No	good	Mostly not
3980	Clark	Kent	1954	76	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	1943	74	Latveria		1	Bad	yes
0573	Stephen	Strange	1968	74	Philidelphia		not	light	Y
7452	Thor	Odinson	-2287	78	Norway		10	Good	Of course
1437	Selina	Kyle	1998	67	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	1911	70	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	67	Krypton	Y	fast	G	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	74	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	1979	71	Manhattan	NA	9	good	long
9471	Diana	Trevor	1618	68	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	1942	64	Cresskill		tiny	Good	Not really
0696	Peter	Parker	1983	70	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	62	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1932	72	Hamburg	NA	Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	67	St. Petersburg	NA	jet	depends	No way
0323	Jean	Grey	1977	66	Annandale		No	good	Mostly not
3980	Clark	Kent	1954	76	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	1943	74	Latveria	Missing	1	Bad	yes
0573	Stephen	Strange	1968	74	Philidelphia		not	light	Y
7452	Thor	Odinson	-2287	78	Norway		10	Good	Of course
1437	Selina	Kyle	1998	67	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	1911	70	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	67	Krypton	Y	fast	G	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	74	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	1979	71	Manhattan	N	9	good	long
9471	Diana	Trevor	1618	68	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	1942	64	Cresskill	N	tiny	Good	Not really
0696	Peter	Parker	1983	70	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	62	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1932	72	Hamburg	N	Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	67	St. Petersburg	N	jet	depends	No way
0323	Jean	Grey	1977	66	Annandale	N	No	good	Mostly not
3980	Clark	Kent	1954	76	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	1943	74	Latveria	N	1	Bad	yes
0573	Stephen	Strange	1968	74	Philidelphia	N	not	light	Y
7452	Thor	Odinson	-2287	78	Norway	N	10	Good	Of course
1437	Selina	Kyle	1998	67	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	1911	70	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	67	Krypton	Y	fast	G	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	74	Gotham	Y	3	anti-villain	black
0958	Ororo	Munroe	1979	71	Manhattan	N	9	good	long
9471	Diana	Trevor	1618	68	Paradise Island	Y	Jet	truth	rarely
9483	Janet	Van Dyne	1942	64	Cresskill	N	tiny	Good	Not really
0696	Peter	Parker	1983	70	Queens	Y	Fall	right	never
5531	Harleen	Quinzell	1981	62	Gotham	Y	-	evil	no
4734	Erik	Lehnsherr	1932	72	Hamburg	N	Lev.	mutants	Absolutely
7757	Natasha	Romanova	1983	67	St. Petersburg	N	jet	depends	No way
0323	Jean	Grey	1977	66	Annandale	N	No	good	Mostly not
3980	Clark	Kent	1954	76	Krypton	Y	12	Truth	always
3057	Victor	Von Doom	1943	74	Latveria	N	1	Bad	yes
0573	Stephen	Strange	1968	74	Philidelphia	N	not	light	Y
7452	Thor	Odinson	-2287	78	Norway	N	10	Good	Of course
1437	Selina	Kyle	1998	67	Gotham	Y	NA	Neutral	It clashes
1883	Raven	Darkholme	1911	70	unknown	Y	no	mostly bad	Not really
5830	Kara	Zor-el	1961	67	Krypton	Y	fast	G	Yes

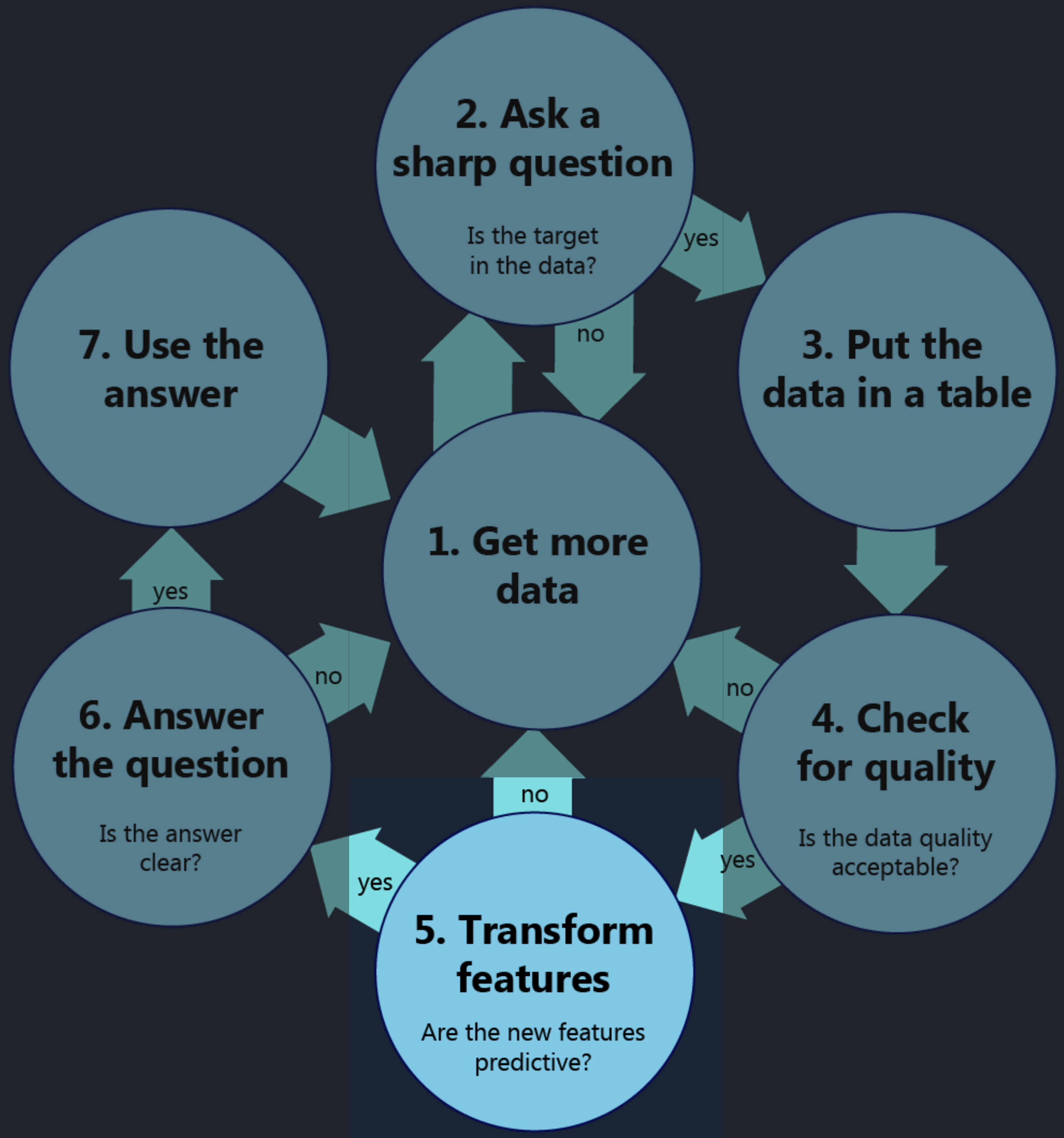
ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	74	Gotham	Y	N	anti-villain	black
0958	Ororo	Munroe	1979	71	Manhattan	N	Y	good	long
9471	Diana	Trevor	1618	68	Paradise Island	Y	N	truth	rarely
9483	Janet	Van Dyne	1942	64	Cresskill	N	Y	Good	Not really
0696	Peter	Parker	1983	70	Queens	Y	N	right	never
5531	Harleen	Quinzell	1981	62	Gotham	Y	N	evil	no
4734	Erik	Lehnsherr	1932	72	Hamburg	N	N	mutants	Absolutely
7757	Natasha	Romanova	1983	67	St. Petersburg	N	N	depends	No way
0323	Jean	Grey	1977	66	Annandale	N	N	good	Mostly not
3980	Clark	Kent	1954	76	Krypton	Y	Y	Truth	always
3057	Victor	Von Doom	1943	74	Latveria	N	N	Bad	yes
0573	Stephen	Strange	1968	74	Philidelphia	N	N	light	Y
7452	Thor	Odinson	-2287	78	Norway	N	Y	Good	Of course
1437	Selina	Kyle	1998	67	Gotham	Y	N	Neutral	It clashes
1883	Raven	Darkholme	1911	70	unknown	Y	N	mostly bad	Not really
5830	Kara	Zor-el	1961	67	Krypton	Y	Y	G	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	74	Gotham	Y	N	anti-villain	black
0958	Ororo	Munroe	1979	71	Manhattan	N	Y	good	long
9471	Diana	Trevor	1618	68	Paradise Island	Y	N	truth	rarely
9483	Janet	Van Dyne	1942	64	Cresskill	N	Y	Good	Not really
0696	Peter	Parker	1983	70	Queens	Y	N	right	never
5531	Harleen	Quinzell	1981	62	Gotham	Y	N	evil	no
4734	Erik	Lehnsherr	1932	72	Hamburg	N	N	mutants	Absolutely
7757	Natasha	Romanova	1983	67	St. Petersburg	N	N	depends	No way
0323	Jean	Grey	1977	66	Annandale	N	N	good	Mostly not
3980	Clark	Kent	1954	76	Krypton	Y	Y	Truth	always
3057	Victor	Von Doom	1943	74	Latveria	N	N	Bad	yes
0573	Stephen	Strange	1968	74	Philidelphia	N	N	light	Y
7452	Thor	Odinson	-2287	78	Norway	N	Y	Good	Of course
1437	Selina	Kyle	1998	67	Gotham	Y	N	Neutral	It clashes
1883	Raven	Darkholme	1911	70	unknown	Y	N	mostly bad	Not really
5830	Kara	Zor-el	1961	67	Krypton	Y	Y	G	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	74	Gotham	Y	N	Good	black
0958	Ororo	Munroe	1979	71	Manhattan	N	Y	Good	long
9471	Diana	Trevor	1618	68	Paradise Island	Y	N	Good	rarely
9483	Janet	Van Dyne	1942	64	Cresskill	N	Y	Good	Not really
0696	Peter	Parker	1983	70	Queens	Y	N	Good	never
5531	Harleen	Quinzell	1981	62	Gotham	Y	N	Bad	no
4734	Erik	Lehnsherr	1932	72	Hamburg	N	N	Bad	Absolutely
7757	Natasha	Romanova	1983	67	St. Petersburg	N	N	Good	No way
0323	Jean	Grey	1977	66	Annandale	N	N	Good	Mostly not
3980	Clark	Kent	1954	76	Krypton	Y	Y	Good	always
3057	Victor	Von Doom	1943	74	Latveria	N	N	Bad	yes
0573	Stephen	Strange	1968	74	Philidelphia	N	N	Good	Y
7452	Thor	Odinson	-2287	78	Norway	N	Y	Good	Of course
1437	Selina	Kyle	1998	67	Gotham	Y	N	Neutral	It clashes
1883	Raven	Darkholme	1911	70	unknown	Y	N	Bad	Not really
5830	Kara	Zor-el	1961	67	Krypton	Y	Y	Good	Yes

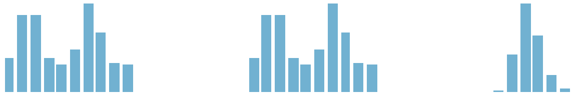
ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	74	Gotham	Y	N	Good	black
0958	Ororo	Munroe	1979	71	Manhattan	N	Y	Good	long
9471	Diana	Trevor	1618	68	Paradise Island	Y	N	Good	rarely
9483	Janet	Van Dyne	1942	64	Cresskill	N	Y	Good	Not really
0696	Peter	Parker	1983	70	Queens	Y	N	Good	never
5531	Harleen	Quinzell	1981	62	Gotham	Y	N	Bad	no
4734	Erik	Lehnsherr	1932	72	Hamburg	N	N	Bad	Absolutely
7757	Natasha	Romanova	1983	67	St. Petersburg	N	N	Good	No way
0323	Jean	Grey	1977	66	Annandale	N	N	Good	Mostly not
3980	Clark	Kent	1954	76	Krypton	Y	Y	Good	always
3057	Victor	Von Doom	1943	74	Latveria	N	N	Bad	yes
0573	Stephen	Strange	1968	74	Philidelphia	N	N	Good	Y
7452	Thor	Odinson	-2287	78	Norway	N	Y	Good	Of course
1437	Selina	Kyle	1998	67	Gotham	Y	N	Neutral	It clashes
1883	Raven	Darkholme	1911	70	unknown	Y	N	Bad	Not really
5830	Kara	Zor-el	1961	67	Krypton	Y	Y	Good	Yes

ID	First name	Last name	Birth year	Height	Birthplace	Identity is secret	Can fly	Alignment	Wears cape
7435	Bruce	Wayne	1969	74	Gotham	Y	N	Good	Y
0958	Ororo	Munroe	1979	71	Manhattan	N	Y	Good	Y
9471	Diana	Trevor	1618	68	Paradise Island	Y	N	Good	N
9483	Janet	Van Dyne	1942	64	Cresskill	N	Y	Good	N
0696	Peter	Parker	1983	70	Queens	Y	N	Good	N
5531	Harleen	Quinzell	1981	62	Gotham	Y	N	Bad	N
4734	Erik	Lehnsherr	1932	72	Hamburg	N	N	Bad	Y
7757	Natasha	Romanova	1983	67	St. Petersburg	N	N	Good	N
0323	Jean	Grey	1977	66	Annandale	N	N	Good	N
3980	Clark	Kent	1954	76	Krypton	Y	Y	Good	Y
3057	Victor	Von Doom	1943	74	Latveria	N	N	Bad	Y
0573	Stephen	Strange	1968	74	Philidelphia	N	N	Good	Y
7452	Thor	Odinson	-2287	78	Norway	N	Y	Good	Y
1437	Selina	Kyle	1998	67	Gotham	Y	N	Neutral	N
1883	Raven	Darkholme	1911	70	unknown	Y	N	Bad	N
5830	Kara	Zor-el	1961	67	Krypton	Y	Y	Good	Y


2. Ask a sharp question
Is the target in the data?
yes
no
3. Put the data in a table
7. Use the answer
1. Get more data
yes
no
no
6. Answer the question
Is the answer clear?
4. Check for quality
Is the data quality acceptable?
no
yes
yes
5. Transform features
Are the new features predictive?

rows	columns
65670	3

view as



0	1	2
5.107477	5.135881	60.479023
5.113939	5.141432	61.419001
5.117143	5.13772	82.774271
5.118805	5.145063	62.552338
5.119299	5.144294	66.799533
5.11949	5.140815	77.870507
5.120502	5.147892	64.326006
5.121868	5.14889	61.743756
5.121949	5.149292	64.493967
5.123392	5.148504	69.140338
5.124216	5.148921	69.449809

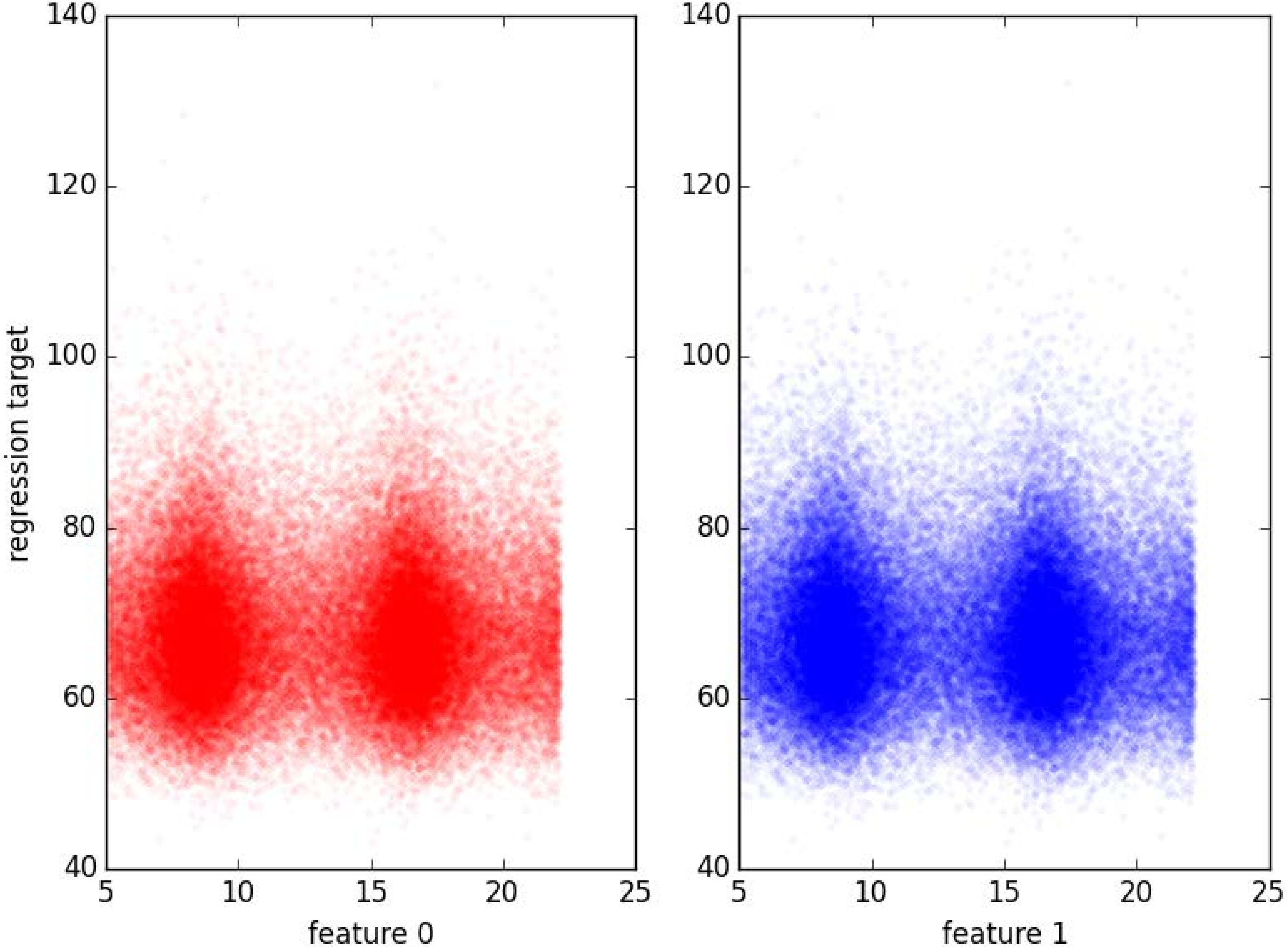
Feature engineering

Sometimes you have to massage the data before it becomes useful in answering your question.

rows	columns
65670	3

view as

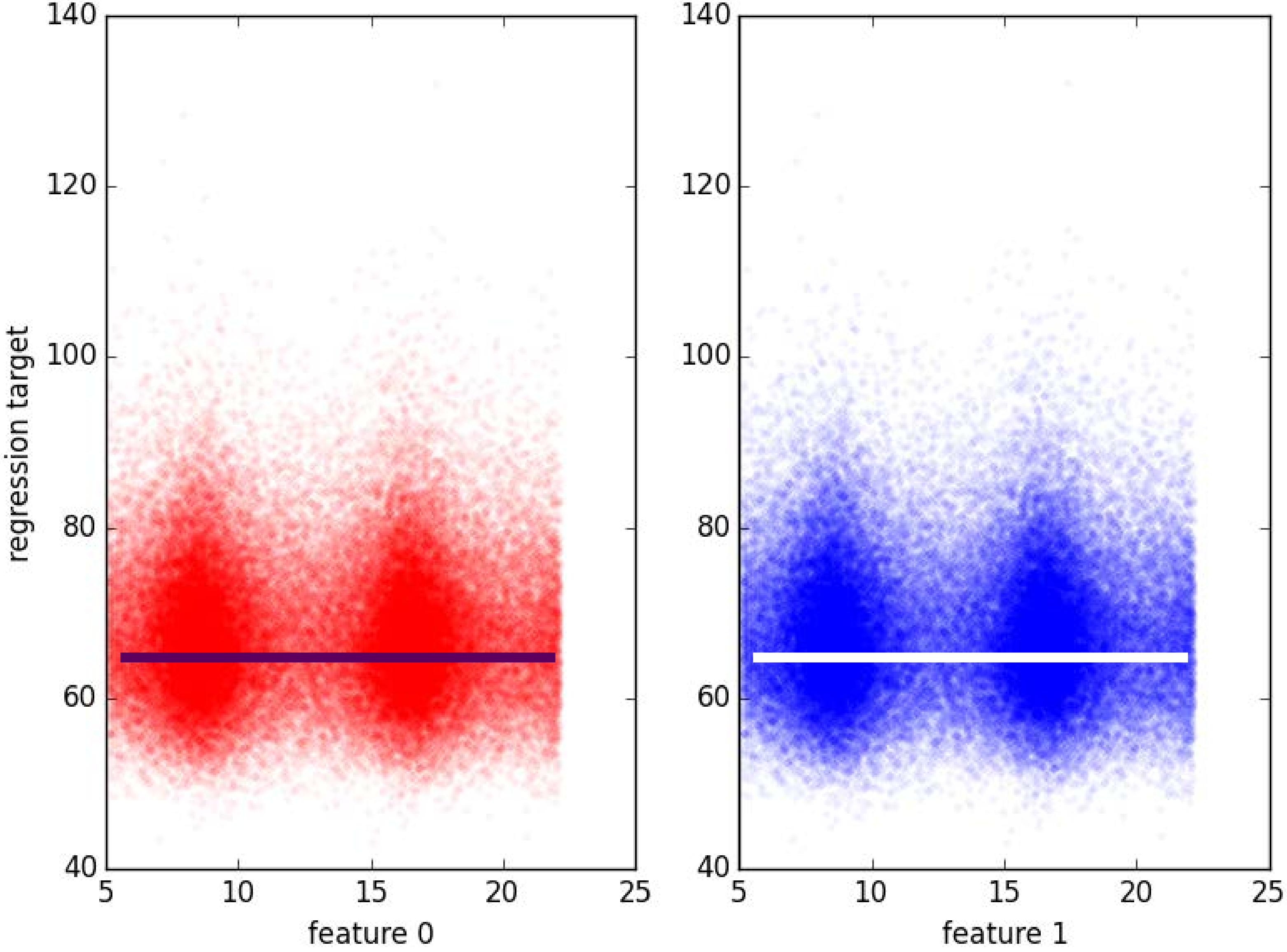
0	1	2
5.107477	5.135881	60.479023
5.113939	5.141432	61.419001
5.117143	5.13772	82.774271
5.118805	5.145063	62.552338
5.119299	5.144294	66.799533
5.11949	5.140815	77.870507
5.120502	5.147892	64.326006
5.121868	5.14889	61.743756
5.121949	5.149292	64.493967
5.123392	5.148504	69.140338
5.124216	5.148921	69.449809



rows	columns
65670	3

view as

0	1	2
5.107477	5.135881	60.479023
5.113939	5.141432	61.419001
5.117143	5.13772	82.774271
5.118805	5.145063	62.552338
5.119299	5.144294	66.799533
5.11949	5.140815	77.870507
5.120502	5.147892	64.326006
5.121868	5.14889	61.743756
5.121949	5.149292	64.493967
5.123392	5.148504	69.140338
5.124216	5.148921	69.449809



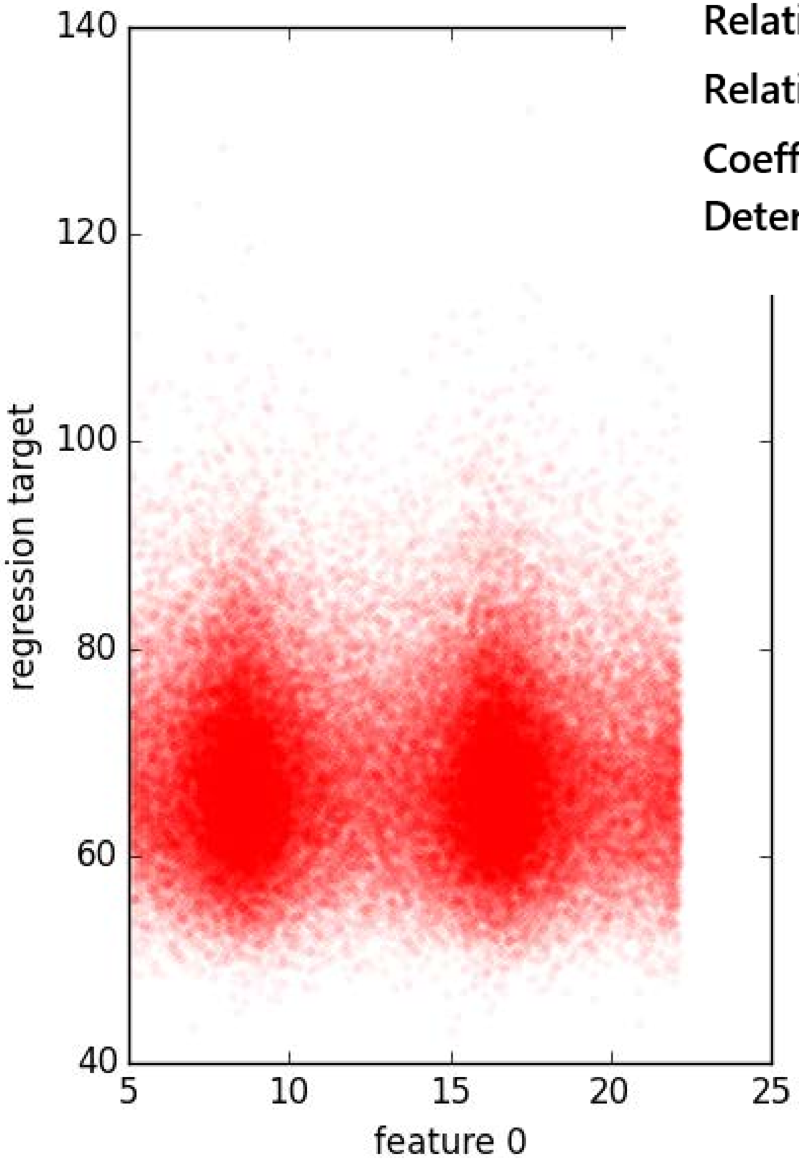
rows	columns
65670	3

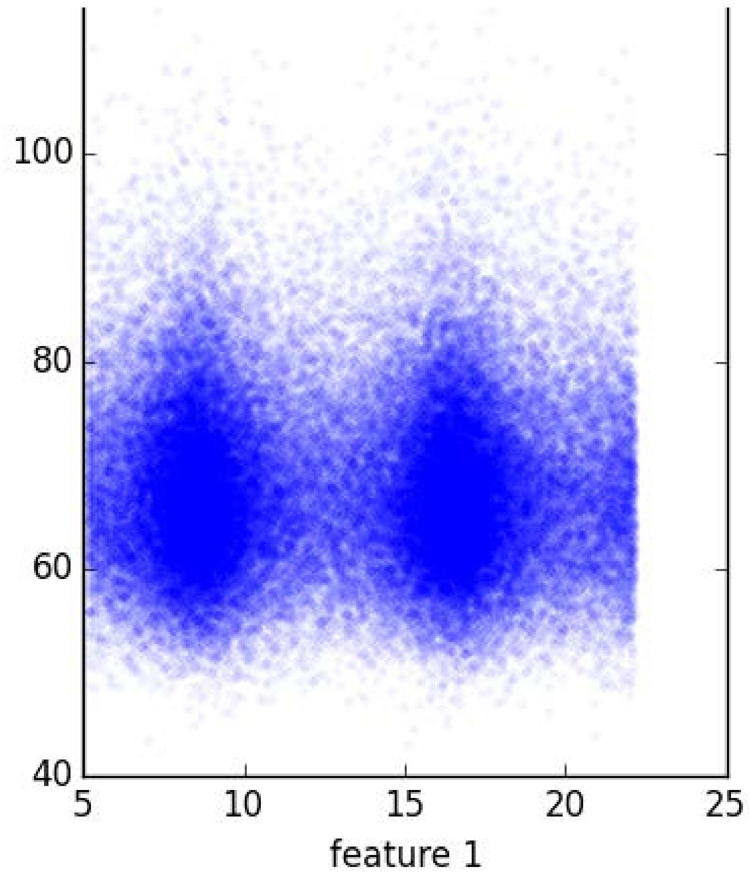
view as

0	1	2
5.107477	5.135881	60.479023
5.113939	5.141432	61.419001
5.117143	5.13772	82.774271
5.118805	5.145063	62.552338
5.119299	5.144294	66.799533
5.11949	5.140815	77.870507
5.120502	5.147892	64.326006
5.121868	5.14889	61.743756
5.121949	5.149292	64.493967
5.123392	5.148504	69.140338
5.124216	5.148921	69.449809

Metrics

Mean Absolute Error	6.485434
Root Mean Squared Error	8.280206
Relative Absolute Error	0.991422
Relative Squared Error	0.983903
Coefficient of Determination	0.016097

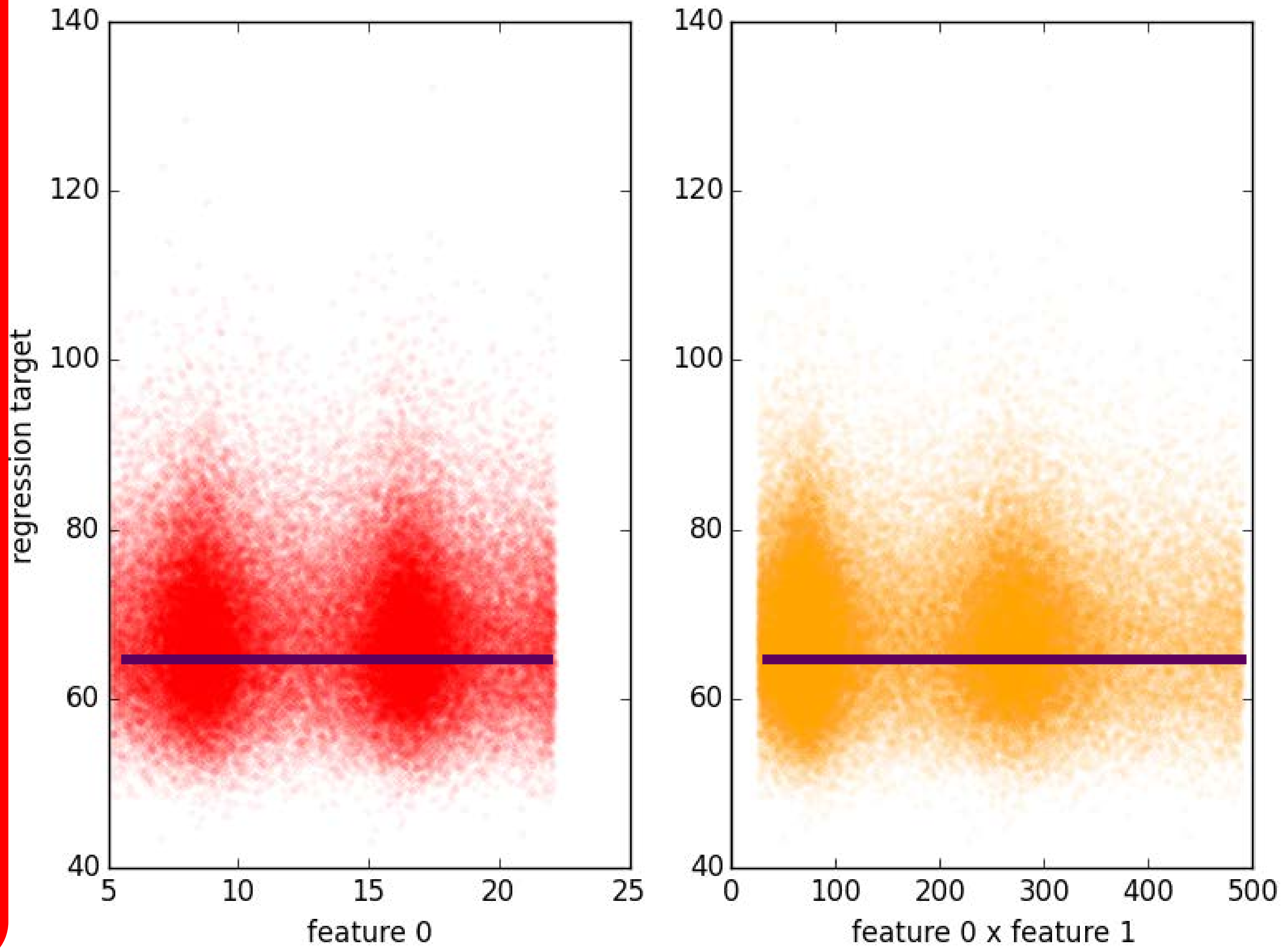




columns

4

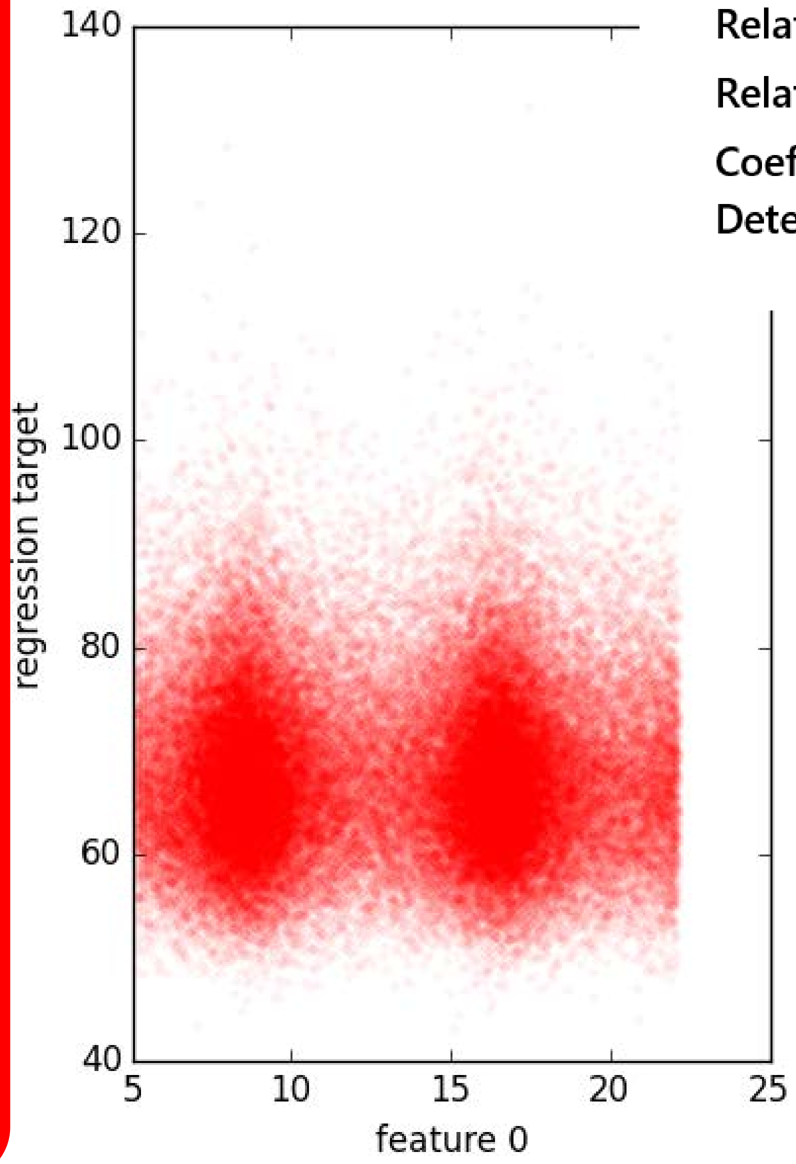
0	1	2	Multiply(1_0)
5.107477	5.135881	60.479023	26.231395
5.113939	5.141432	61.419001	26.292971
5.117143	5.13772	82.774271	26.290449
5.118805	5.145063	62.552338	26.336574
5.119299	5.144294	66.799533	26.335178
5.11949	5.140815	77.870507	26.318351
5.120502	5.147892	64.326006	26.359789
5.121868	5.14889	61.743756	26.371937
5.121949	5.149292	64.493967	26.374413
5.123392	5.148504	69.140338	26.3778
5.124216	5.148921	69.449809	26.384186
5.126409	5.154655	62.028089	26.42487



columns

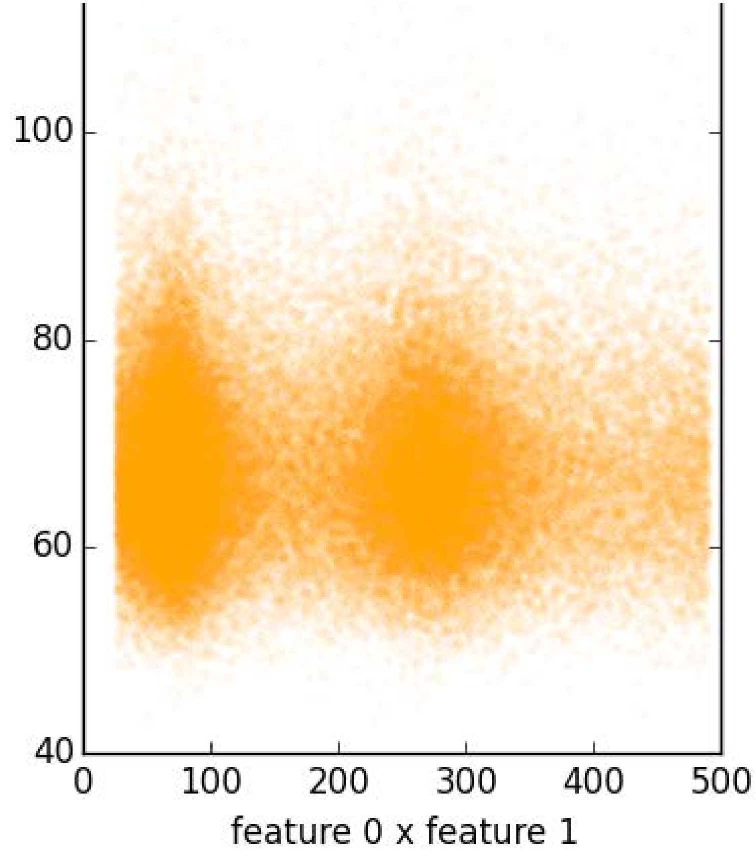
4

0	1	2	Multiply(1_0)
5.107477	5.135881	60.479023	26.231395
5.113939	5.141432	61.419001	26.292971
5.117143	5.13772	82.774271	26.290449
5.118805	5.145063	62.552338	26.336574
5.119299	5.144294	66.799533	26.335178
5.11949	5.140815	77.870507	26.318351
5.120502	5.147892	64.326006	26.359789
5.121868	5.14889	61.743756	26.371937
5.121949	5.149292	64.493967	26.374413
5.123392	5.148504	69.140338	26.3778
5.124216	5.148921	69.449809	26.384186
5.126409	5.154655	62.028089	26.42487



Metrics

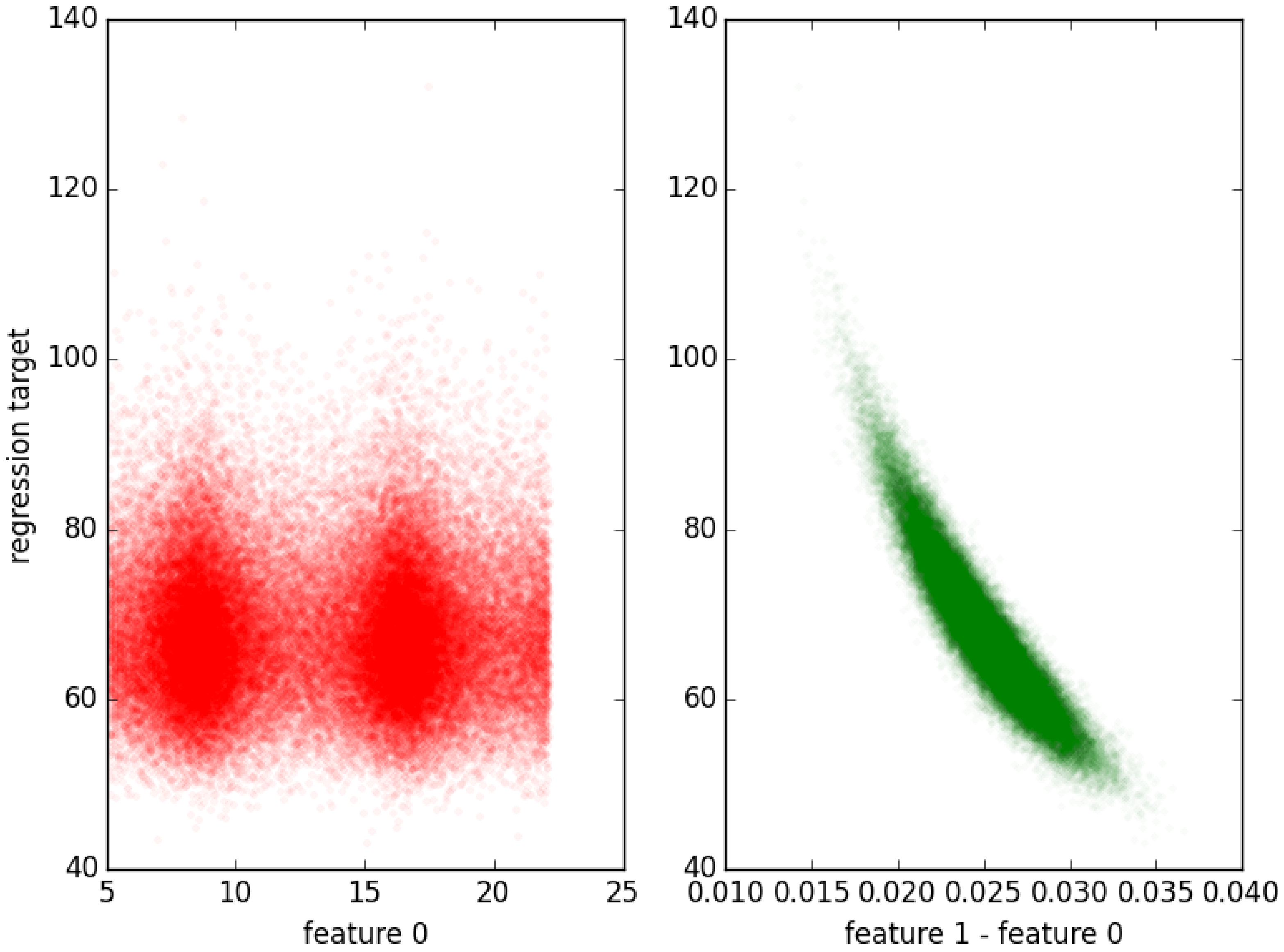
Mean Absolute Error	6.491614
Root Mean Squared Error	8.285875
Relative Absolute Error	0.992366
Relative Squared Error	0.98525
Coefficient of Determination	0.01475



columns

4

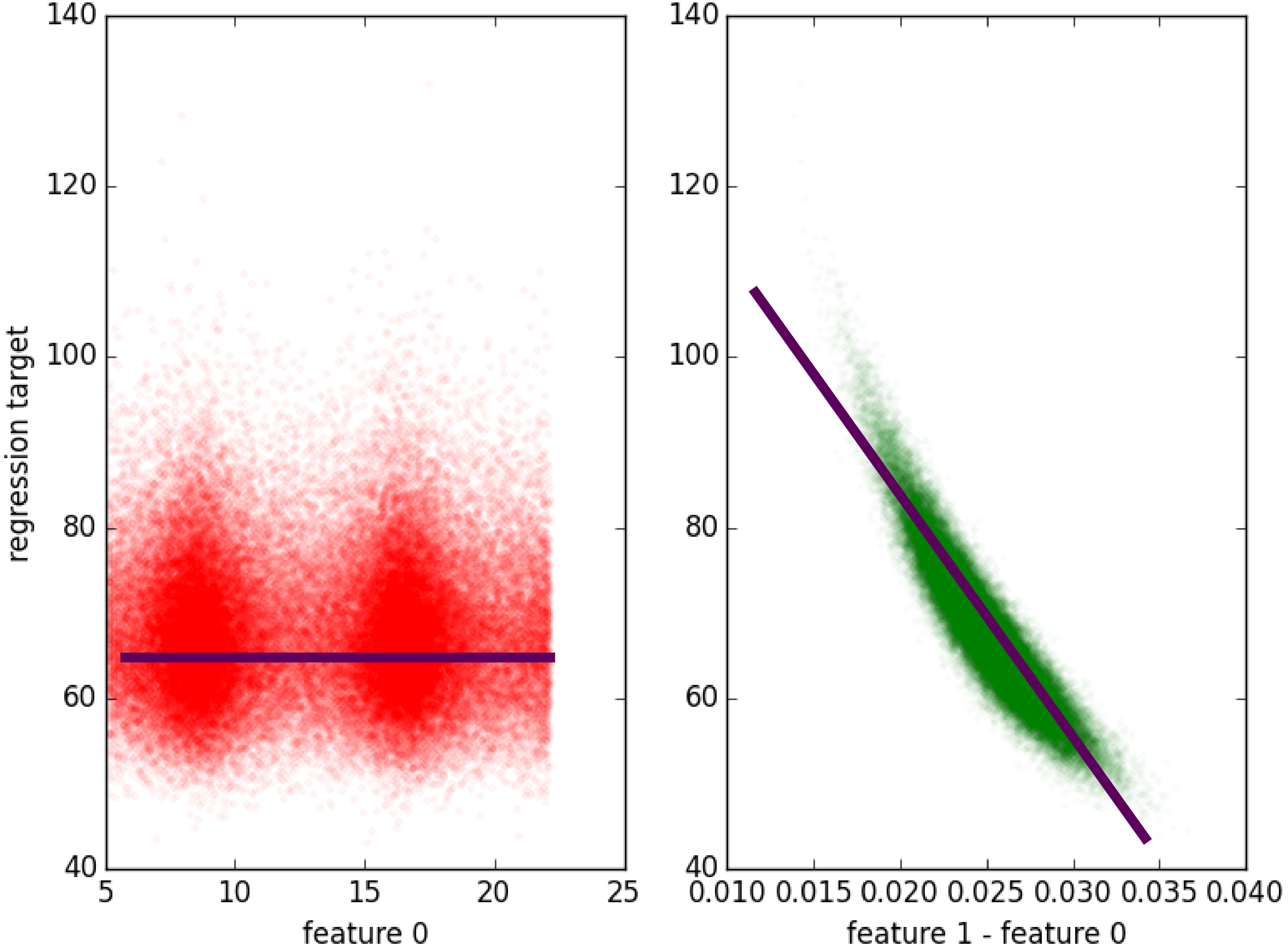
0	1	2	Subtract(1_0)
5.107477	5.135881	60.479023	0.028404
5.113939	5.141432	61.419001	0.027493
5.117143	5.13772	82.774271	0.020578
5.118805	5.145063	62.552338	0.026258
5.119299	5.144294	66.799533	0.024995
5.11949	5.140815	77.870507	0.021325
5.120502	5.147892	64.326006	0.02739
5.121868	5.14889	61.743756	0.027022
5.121949	5.149292	64.493967	0.027343
5.123392	5.148504	69.140338	0.025112
5.124216	5.148921	69.449809	0.024705
5.126409	5.154655	62.028089	0.028246



columns

4

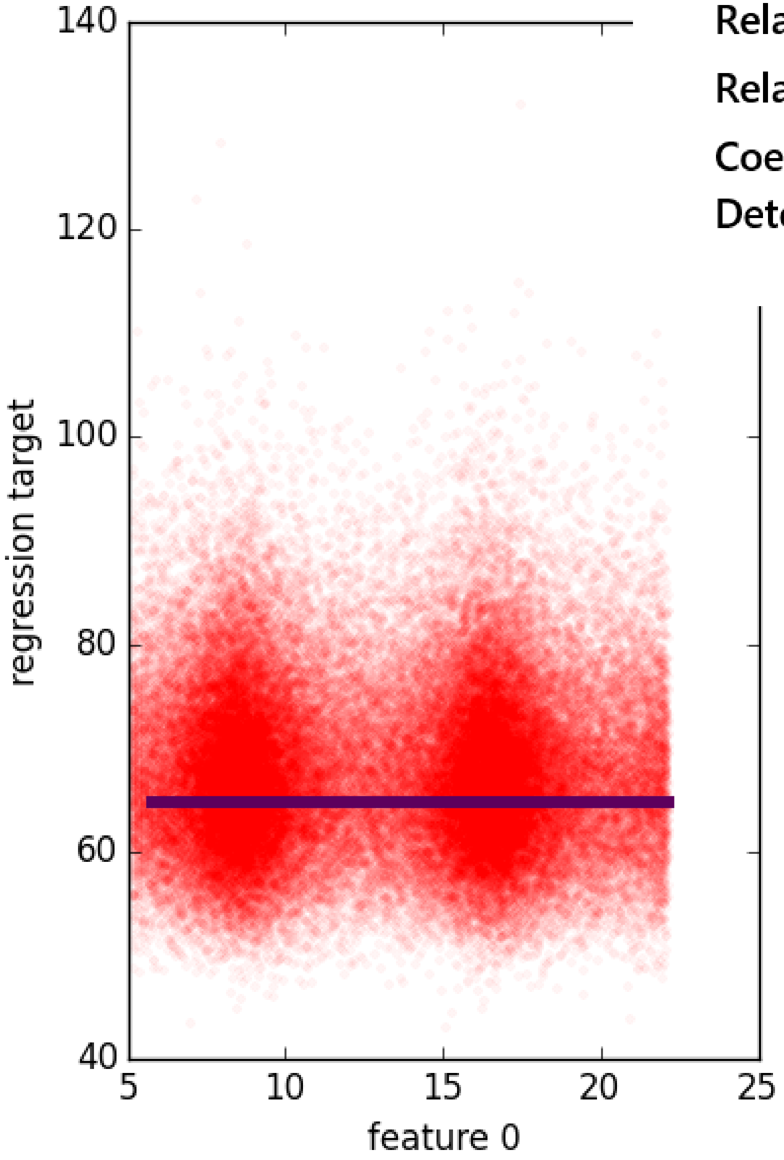
0	1	2	Subtract(1_0)
5.107477	5.135881	60.479023	0.028404
5.113939	5.141432	61.419001	0.027493
5.117143	5.13772	82.774271	0.020578
5.118805	5.145063	62.552338	0.026258
5.119299	5.144294	66.799533	0.024995
5.11949	5.140815	77.870507	0.021325
5.120502	5.147892	64.326006	0.02739
5.121868	5.14889	61.743756	0.027022
5.121949	5.149292	64.493967	0.027343
5.123392	5.148504	69.140338	0.025112
5.124216	5.148921	69.449809	0.024705
5.126409	5.154655	62.028089	0.028246



columns

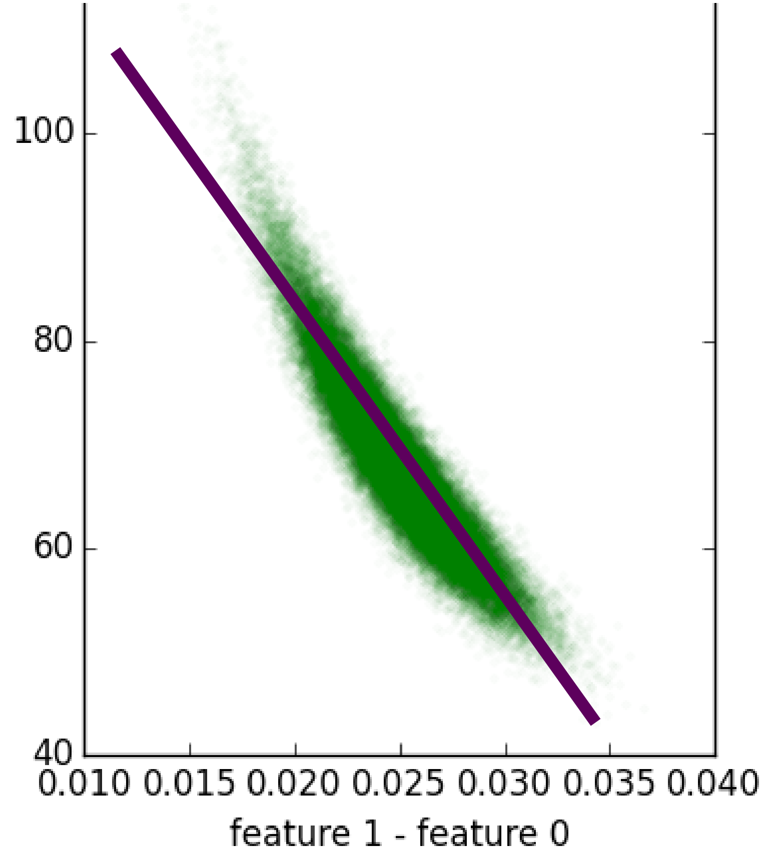
4

0	1	2	Subtract(1_0)
5.107477	5.135881	60.479023	0.028404
5.113939	5.141432	61.419001	0.027493
5.117143	5.13772	82.774271	0.020578
5.118805	5.145063	62.552338	0.026258
5.119299	5.144294	66.799533	0.024995
5.11949	5.140815	77.870507	0.021325
5.120502	5.147892	64.326006	0.02739
5.121868	5.14889	61.743756	0.027022
5.121949	5.149292	64.493967	0.027343
5.123392	5.148504	69.140338	0.025112
5.124216	5.148921	69.449809	0.024705
5.126409	5.154655	62.028089	0.028246



Metrics

Mean Absolute Error	2.243981
Root Mean Squared Error	2.834526
Relative Absolute Error	0.343035
Relative Squared Error	0.1153
Coefficient of Determination	0.8847



Other feature engineering tricks

Data-specific

- Images (SIFT)
- Text (TF-IDF)

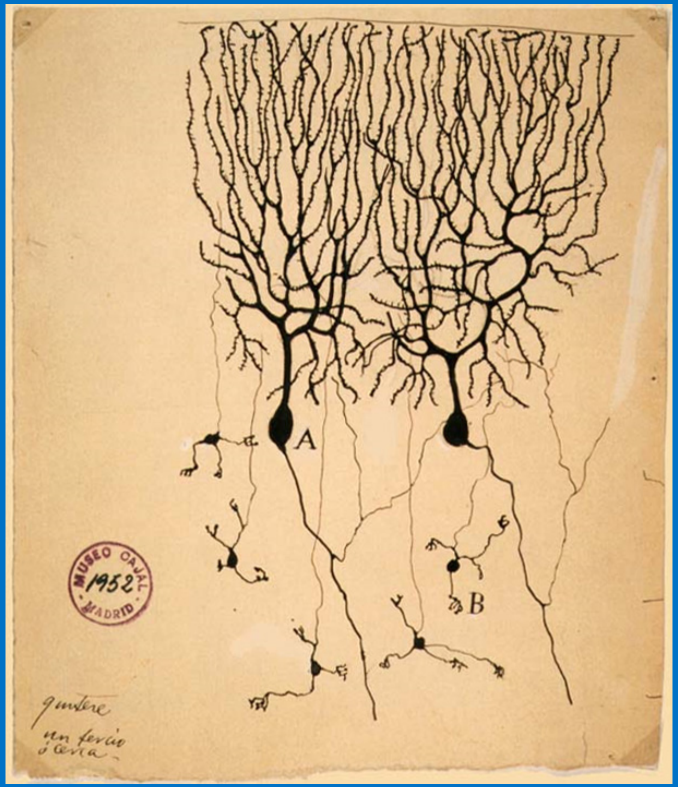
Domain specific

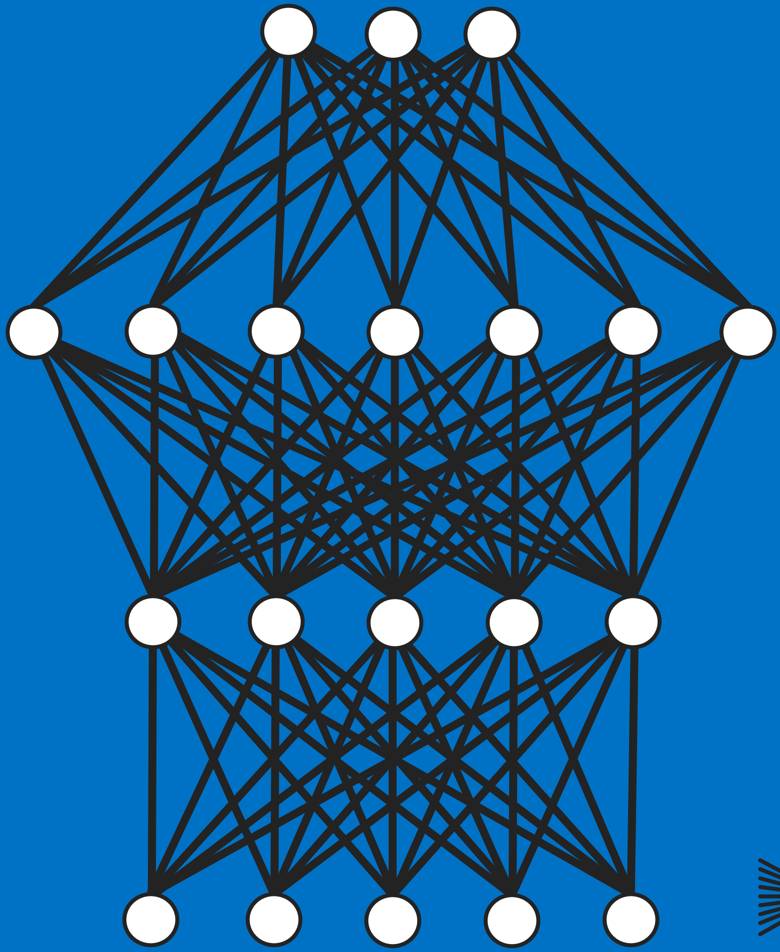
- Econometric, agricultural, sociological, …

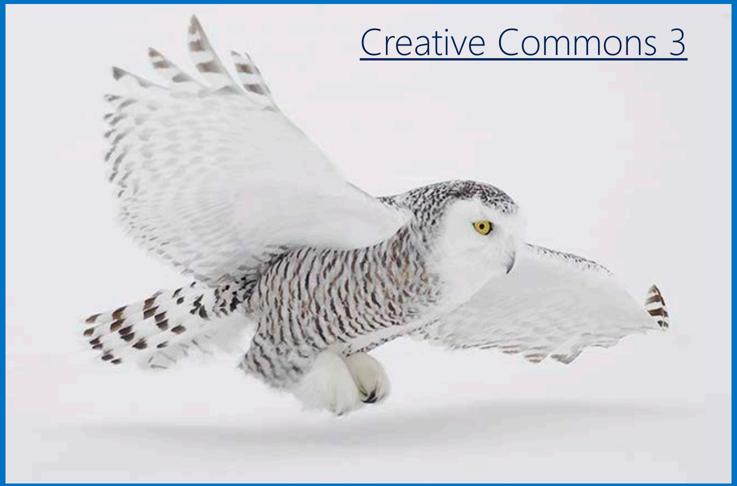
Deep learning

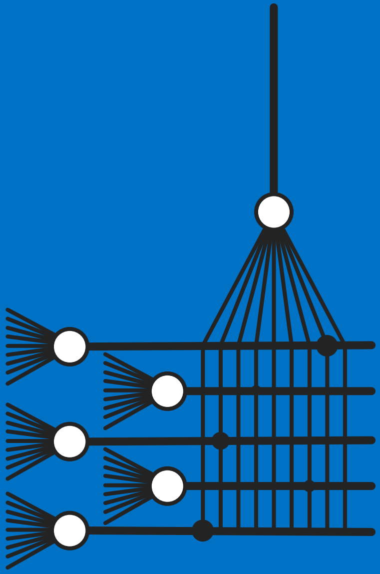
- Images, text, audio

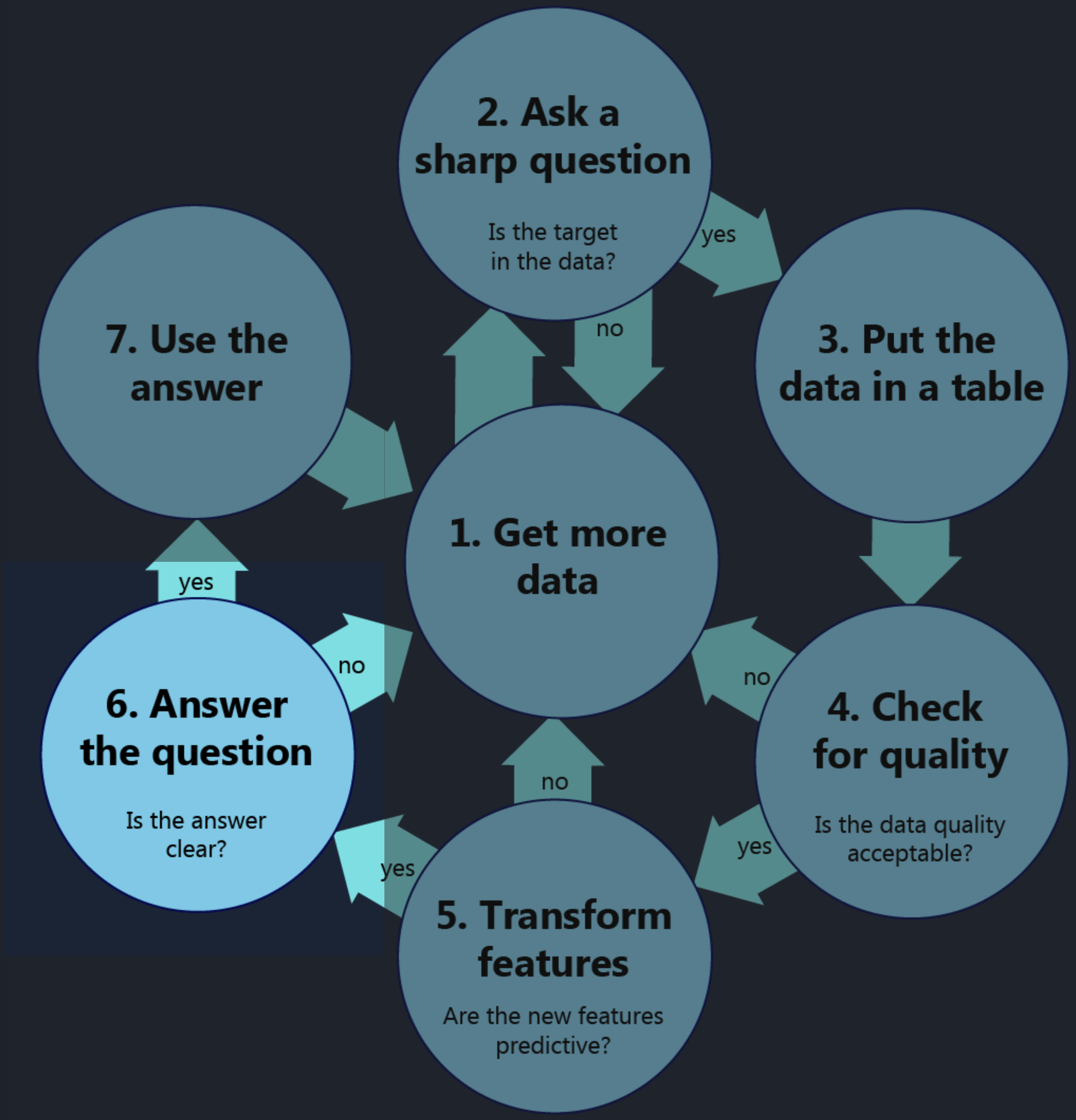
Deep Learning Demystified










2. Ask a sharp question
Is the target in the data?
yes
no
3. Put the data in a table
7. Use the answer
1. Get more data
yes
no
no
4. Check for quality
Is the data quality acceptable?
6. Answer the question
Is the answer clear?
no
yes
yes
5. Transform features
Are the new features predictive?

1. How much / how many?
2. Which category?
3. Which groups?
4. Is it weird?
5. Which action?

[algorithm]

How much / how many?

What will the temperature be next Tuesday?

What will my fourth quarter sales in Portugal be?

How many new followers will I get next week?

[regression]

Which category?

Is this an image of a cat or a dog?

Which aircraft is causing this radar signature?

What is the topic of this news article?

[classification]

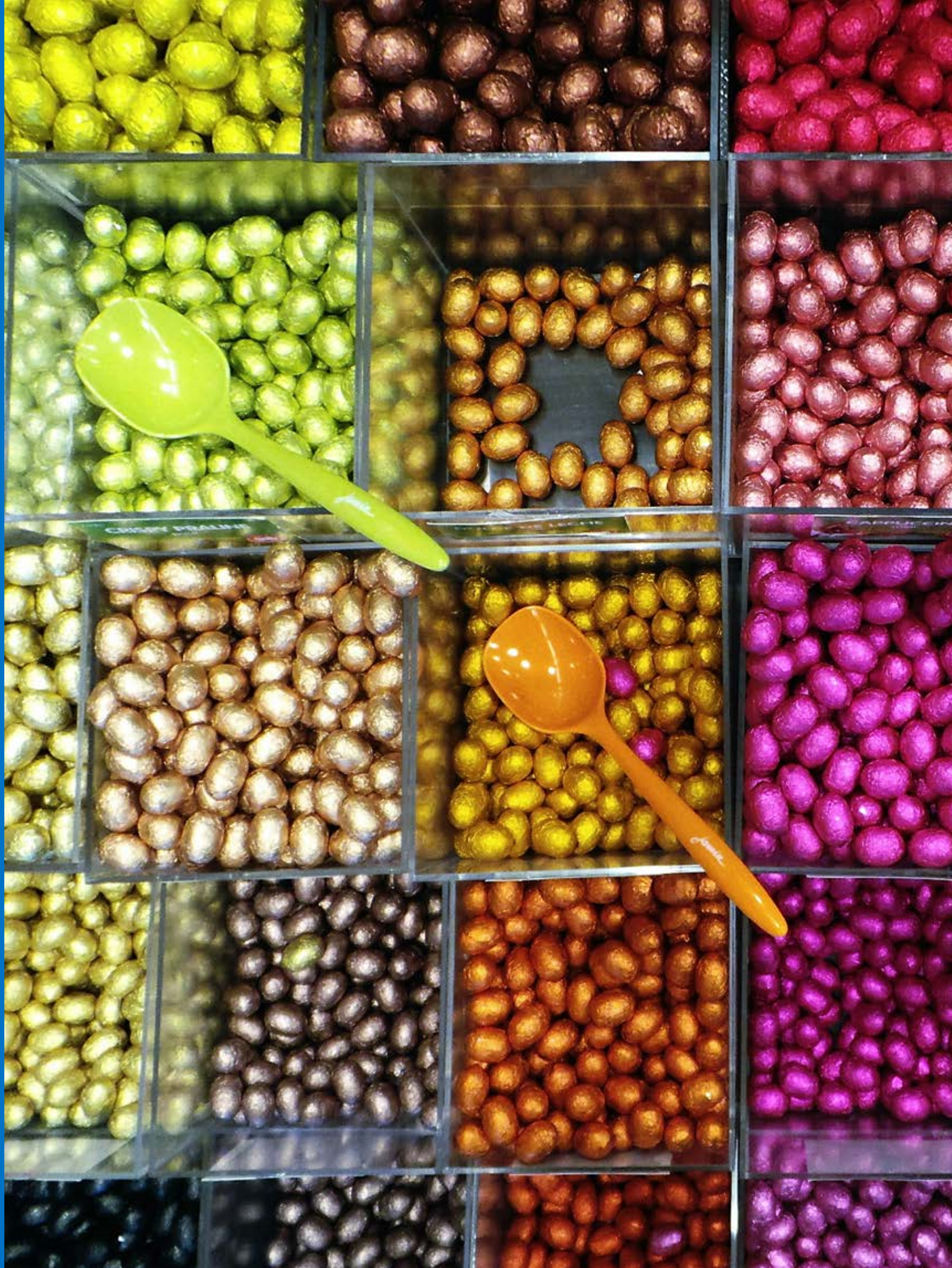
Which groups?

Which shoppers have similar tastes in produce?

Which viewers like the same kind of movies?

What is a natural way to break these documents into five topic groups?

[clustering] [recommendation]



Is this weird?

Is this pressure reading unusual?

Is this internet message typical?

Is this combination of purchases very different from what this customer has made in the past?

[anomaly detection]

Which action?

Should I raise or lower the temperature?

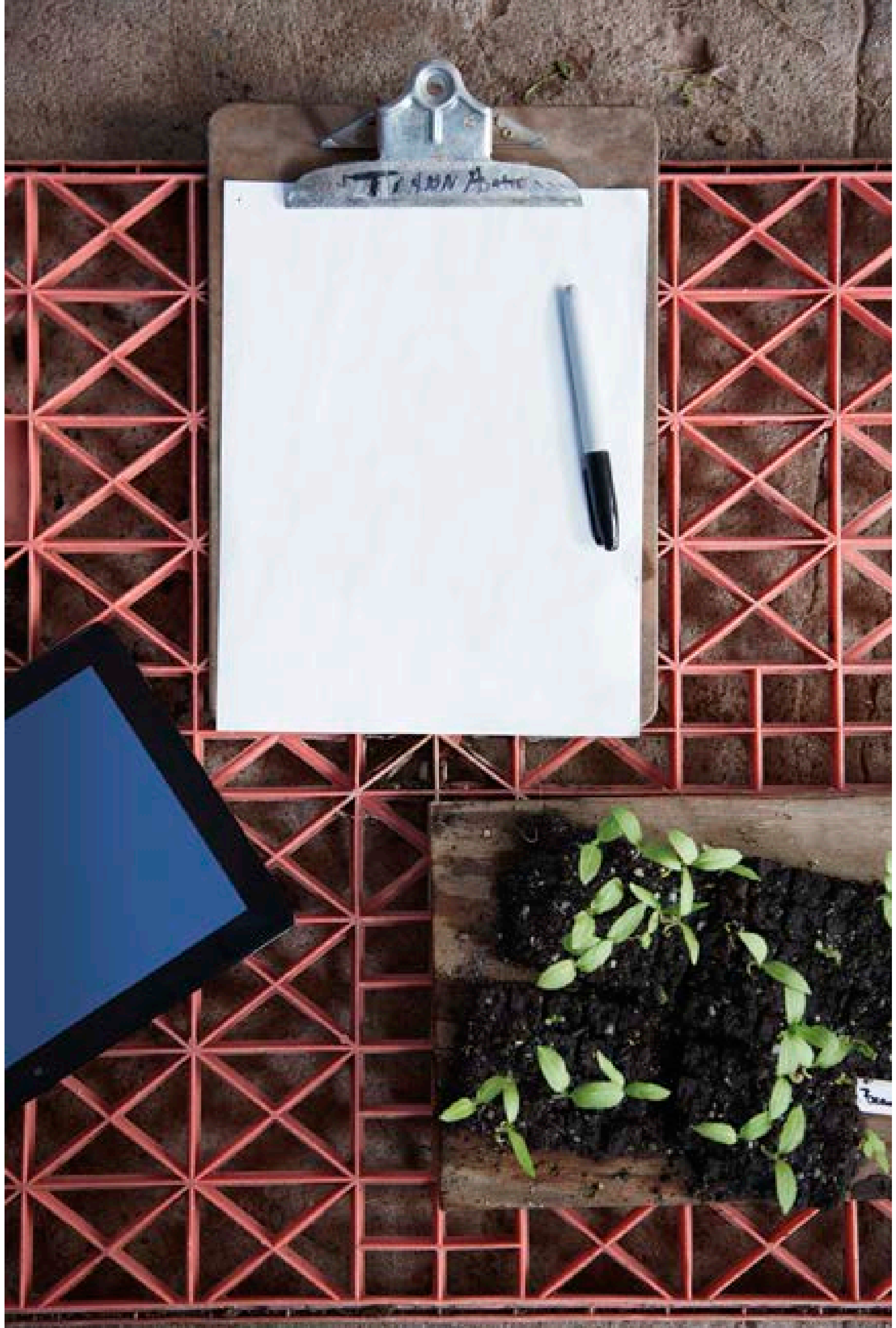
Should I vacuum the living room again or stay plugged in to my charging station?

Should I brake or accelerate in response to that yellow light?

[reinforcement learning]

Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



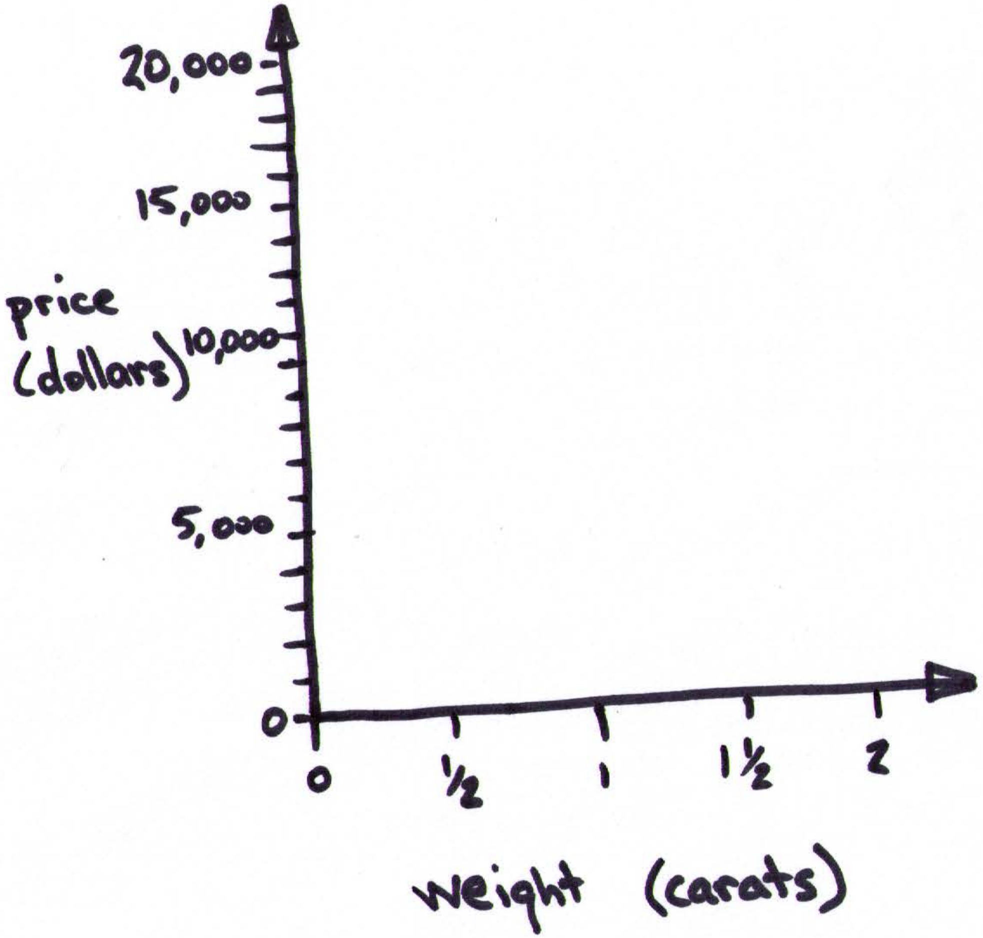
Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695

[number line]
[axis]

Diamonds

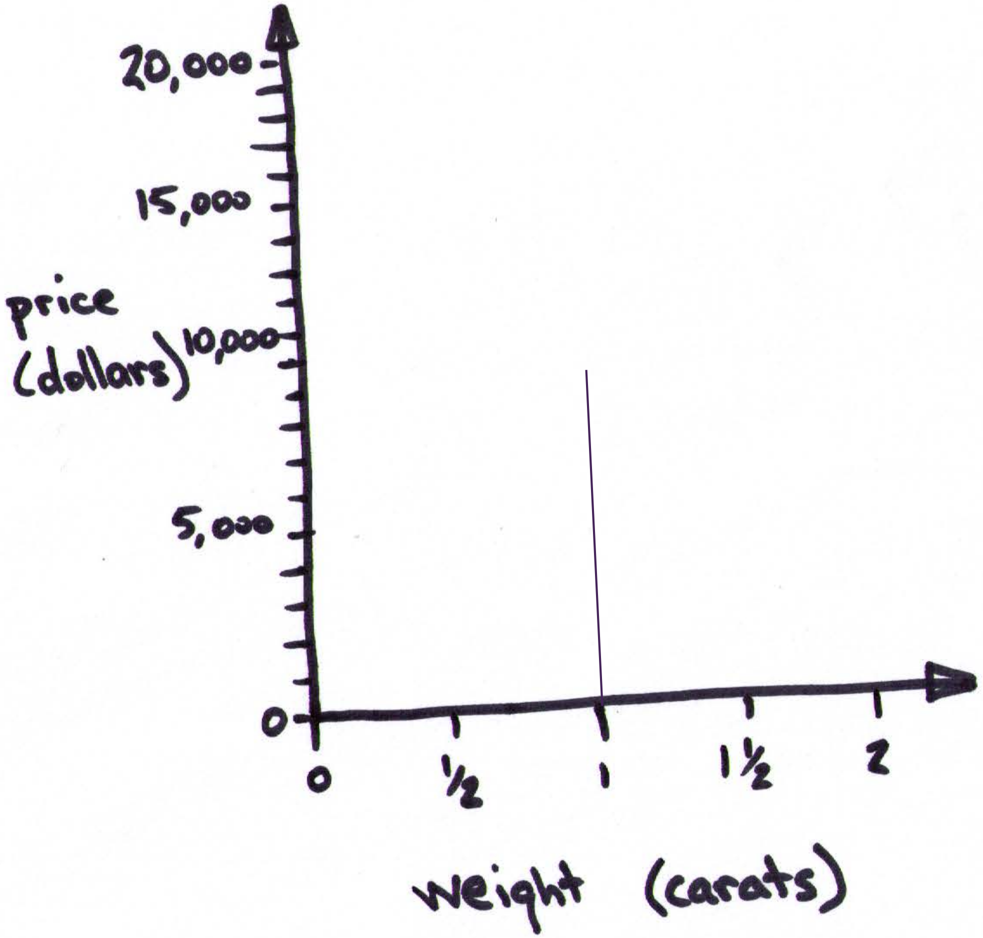
carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



[axes]
[units]

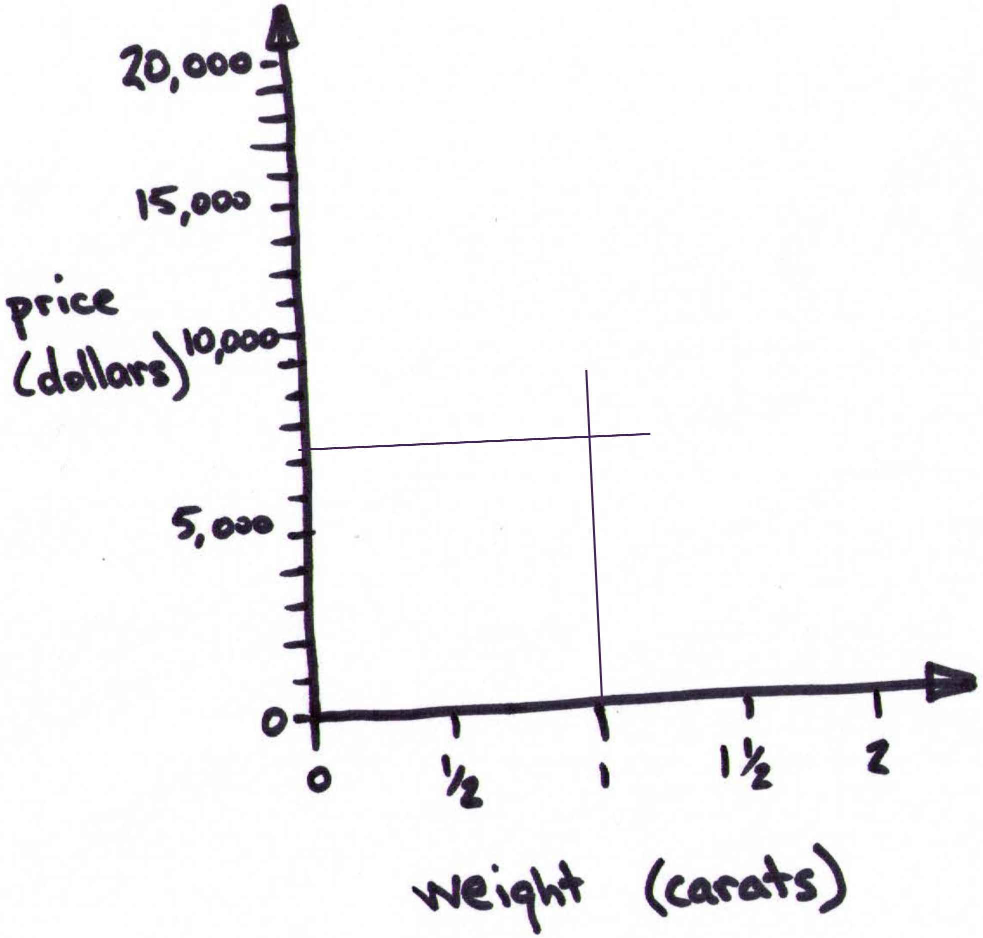
Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



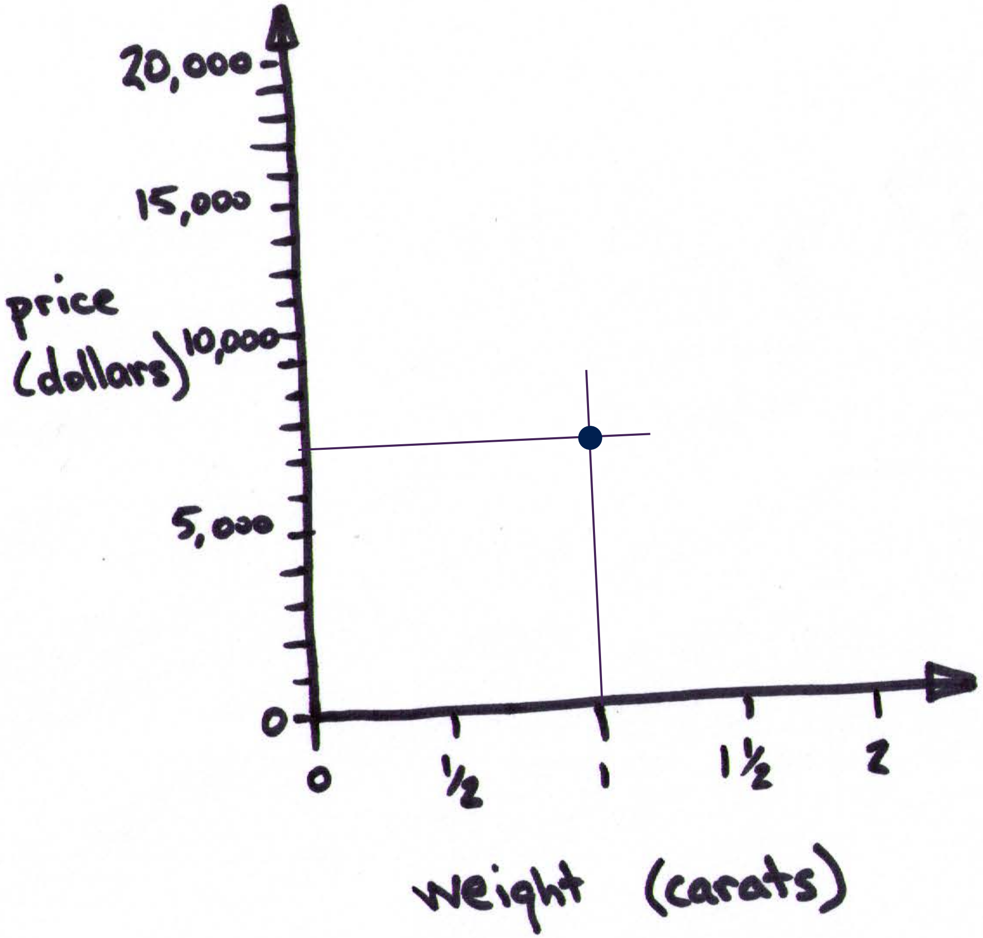
Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



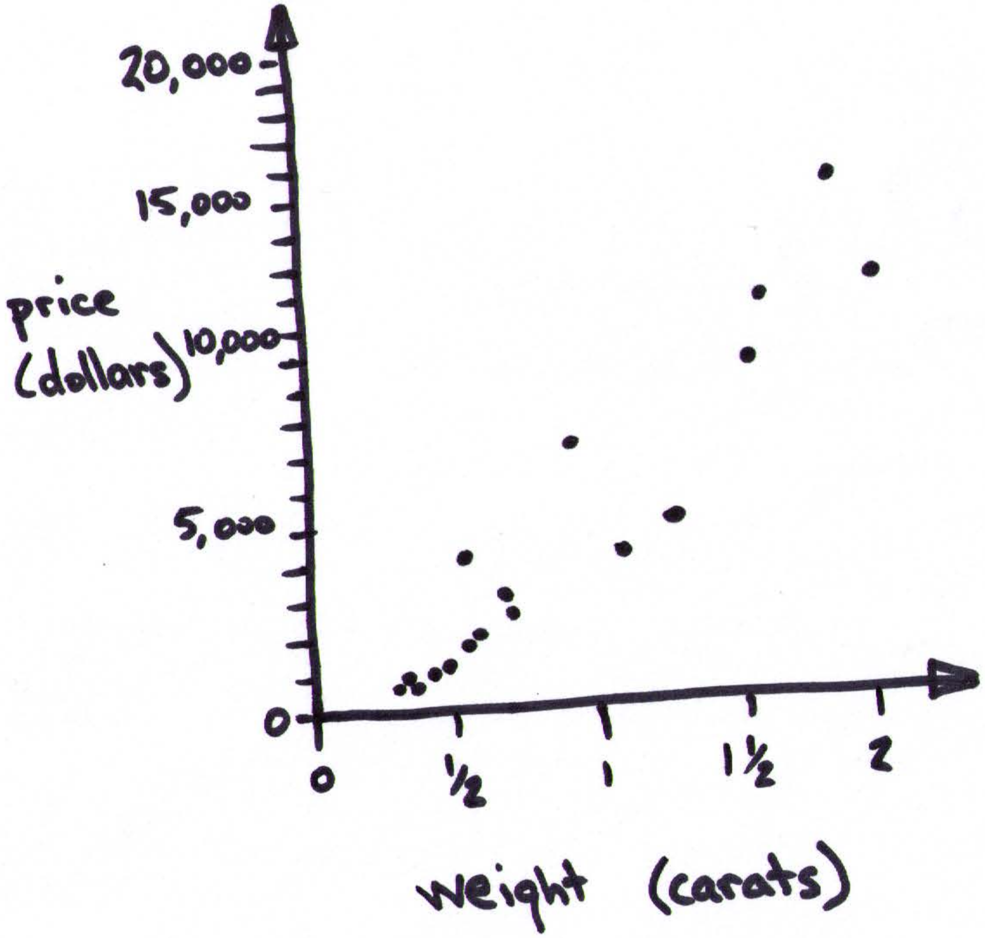
Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

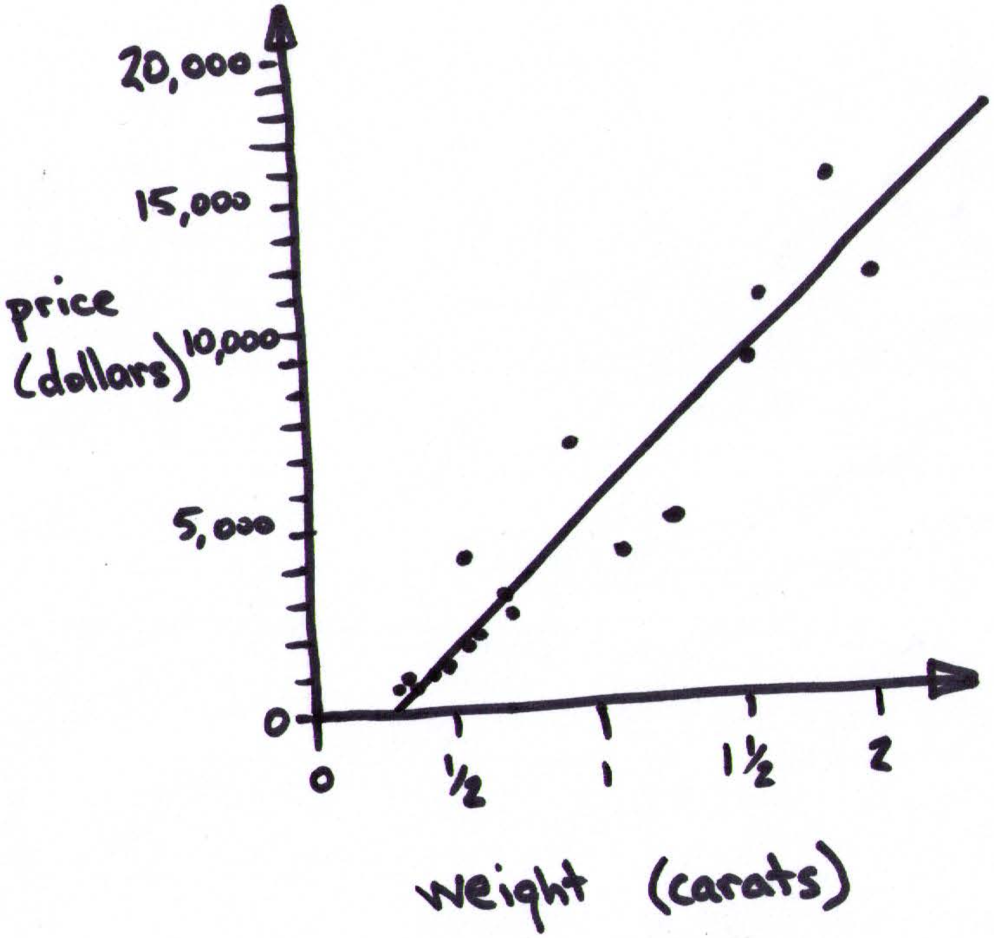
carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



[plot]
[scatter plot]

Diamonds

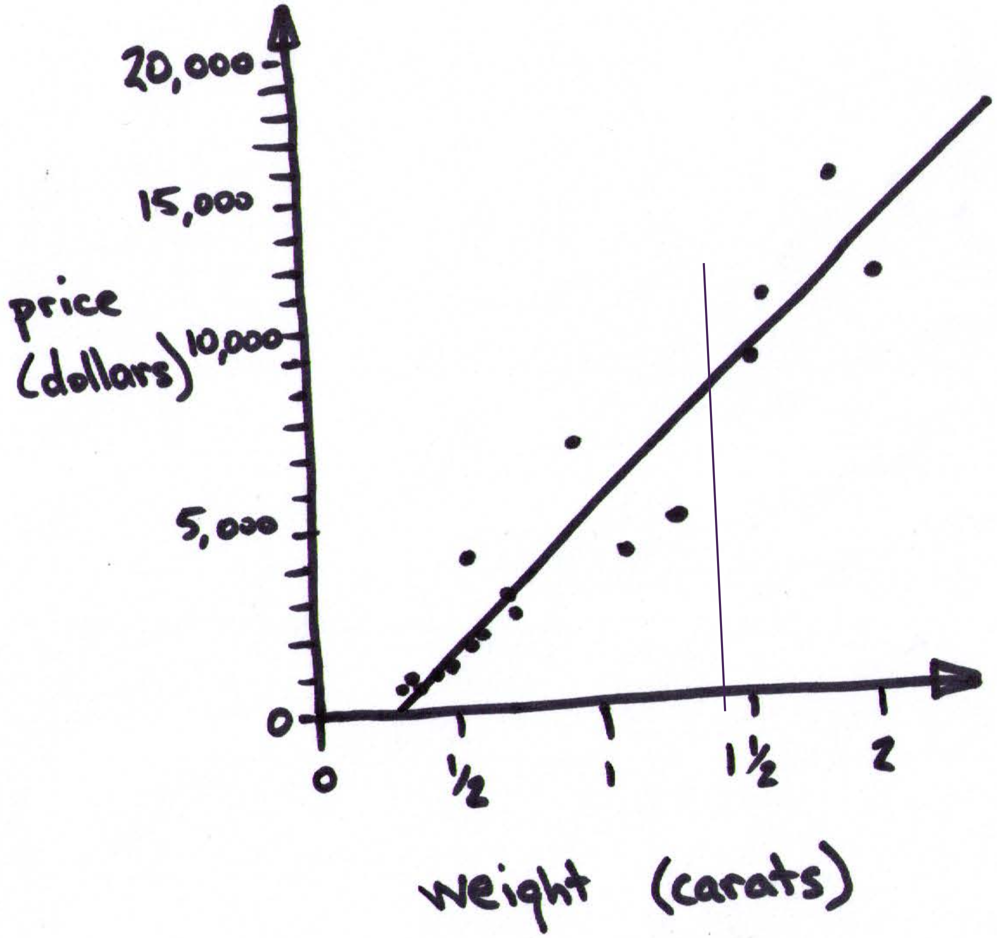
carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



[modeling]
[linear regression]
[variance]
[noise]

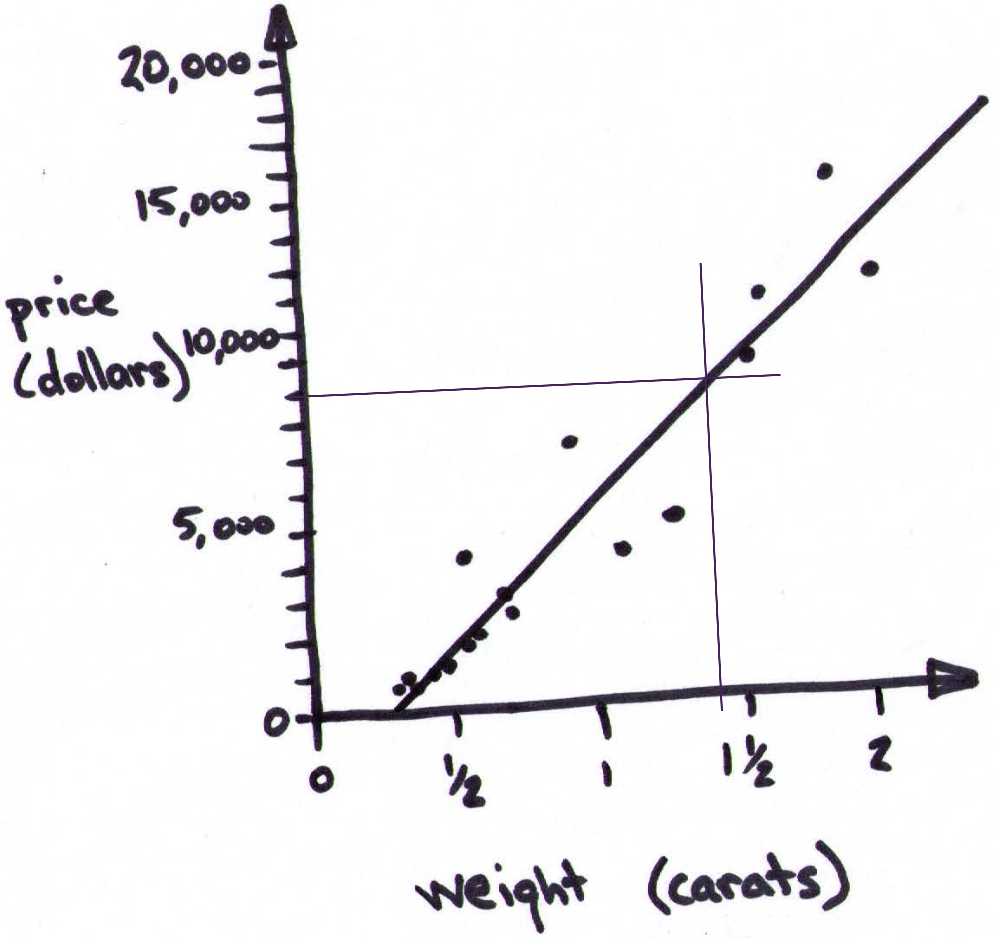
Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

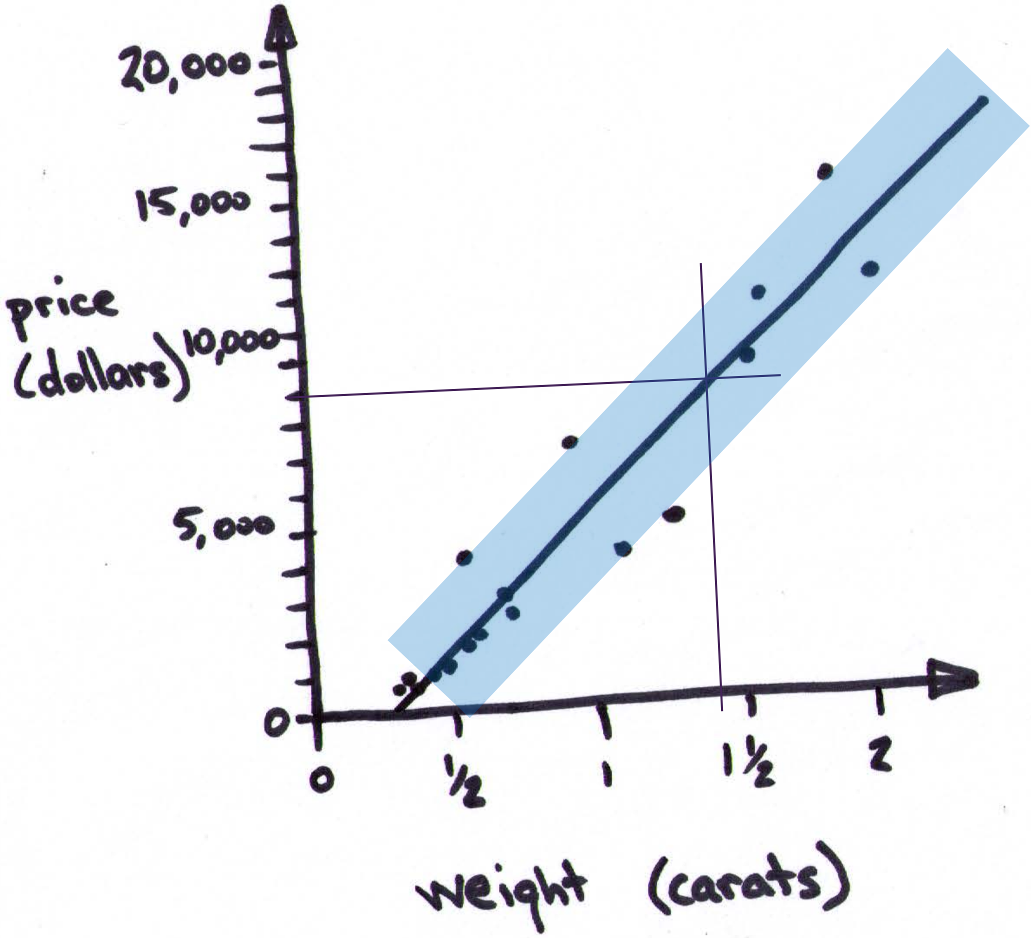
carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



[prediction]

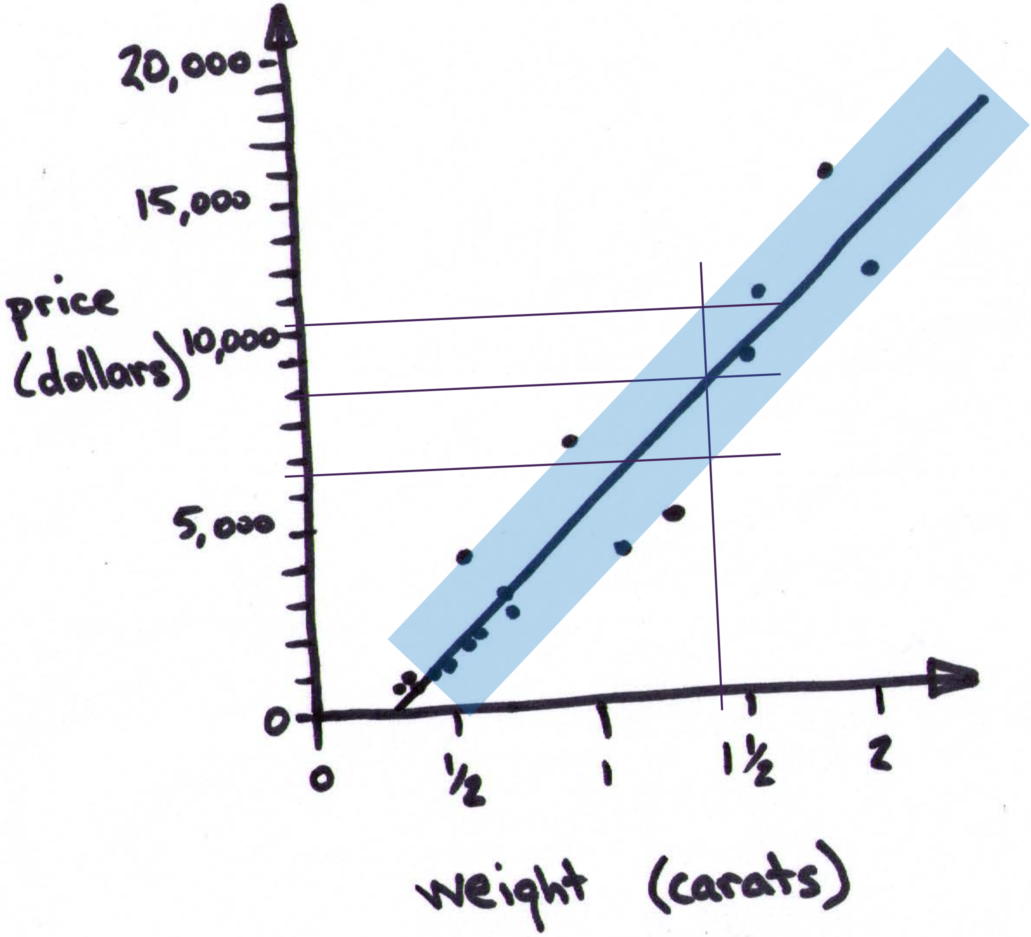
Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



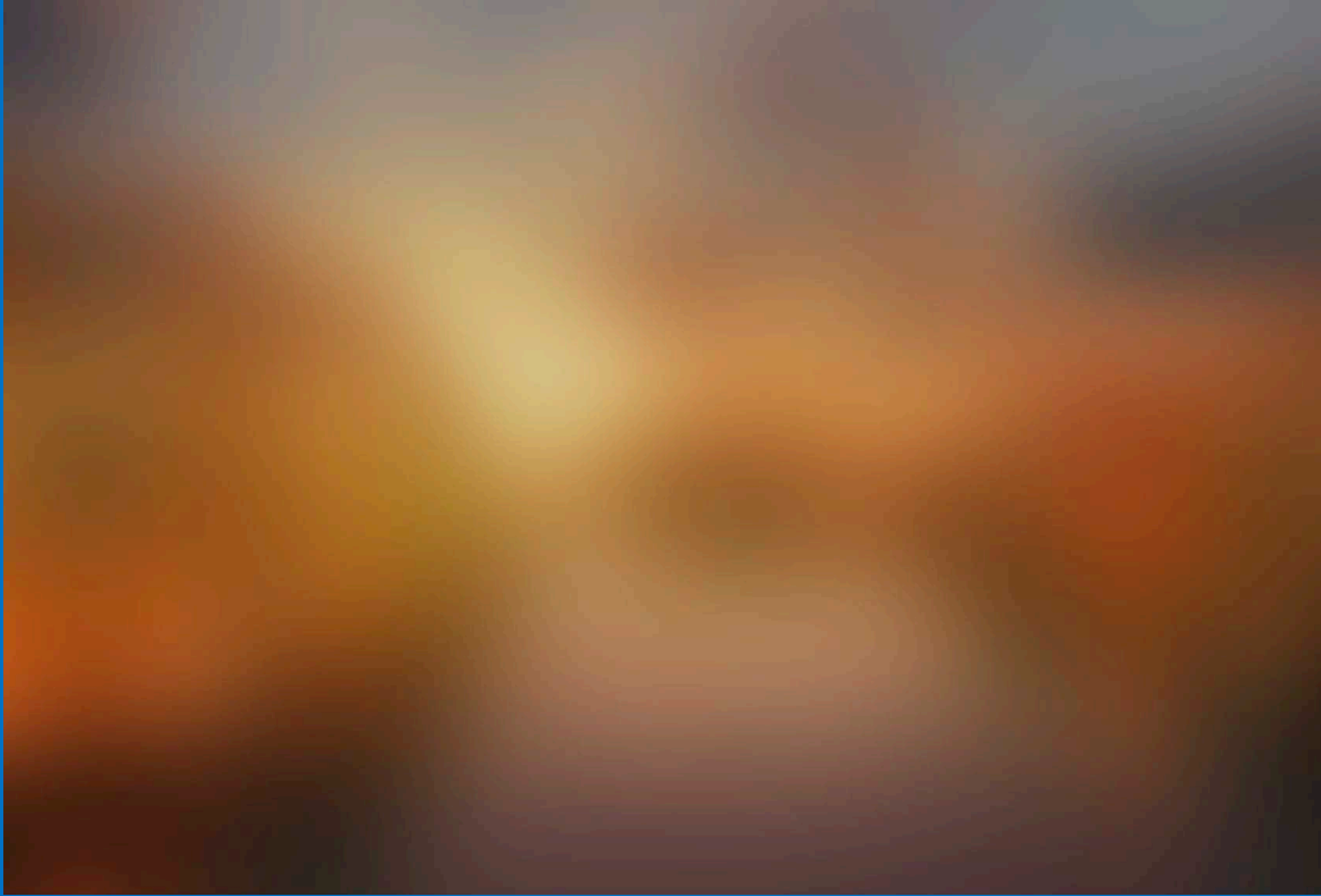
Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695

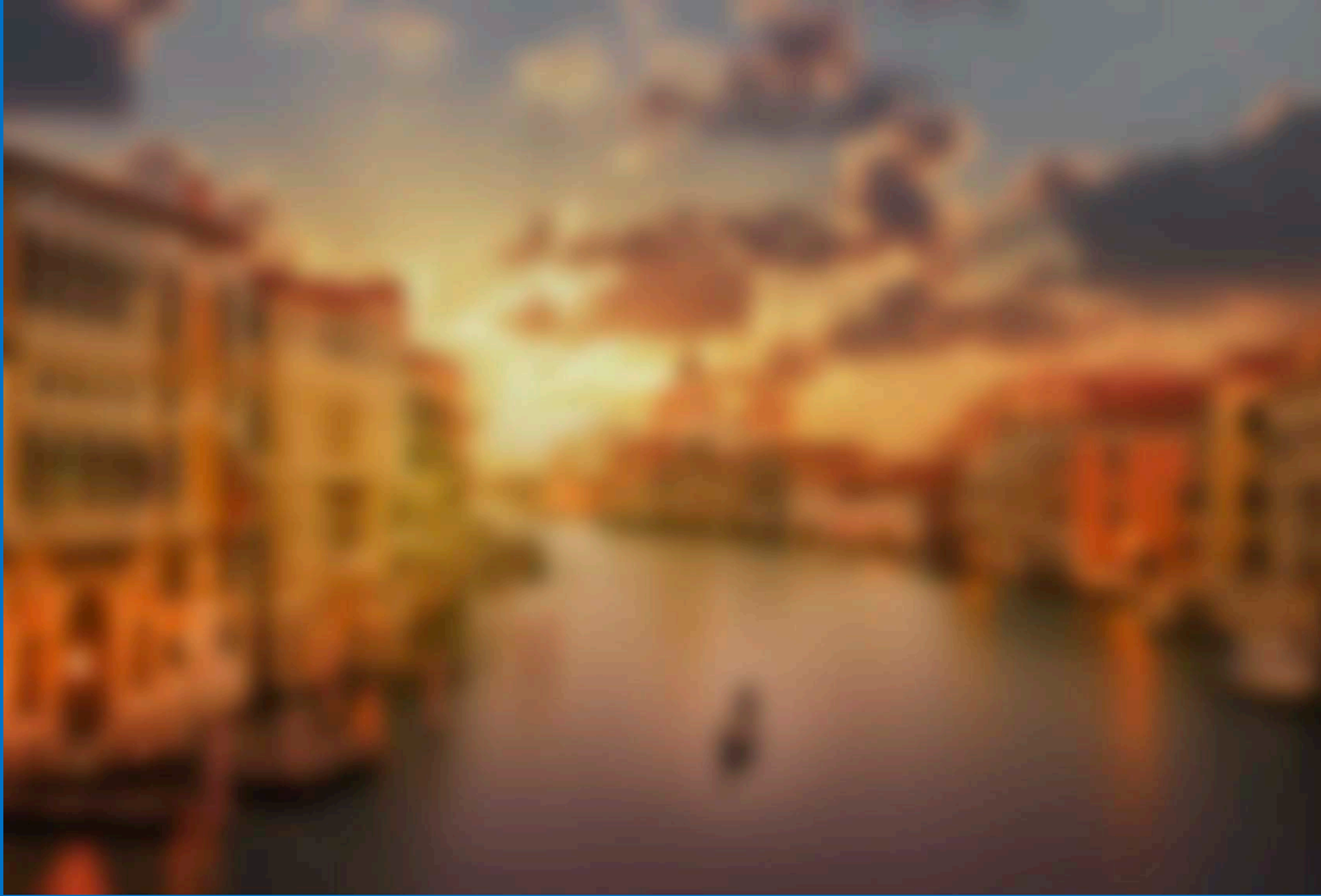


[confidence interval]

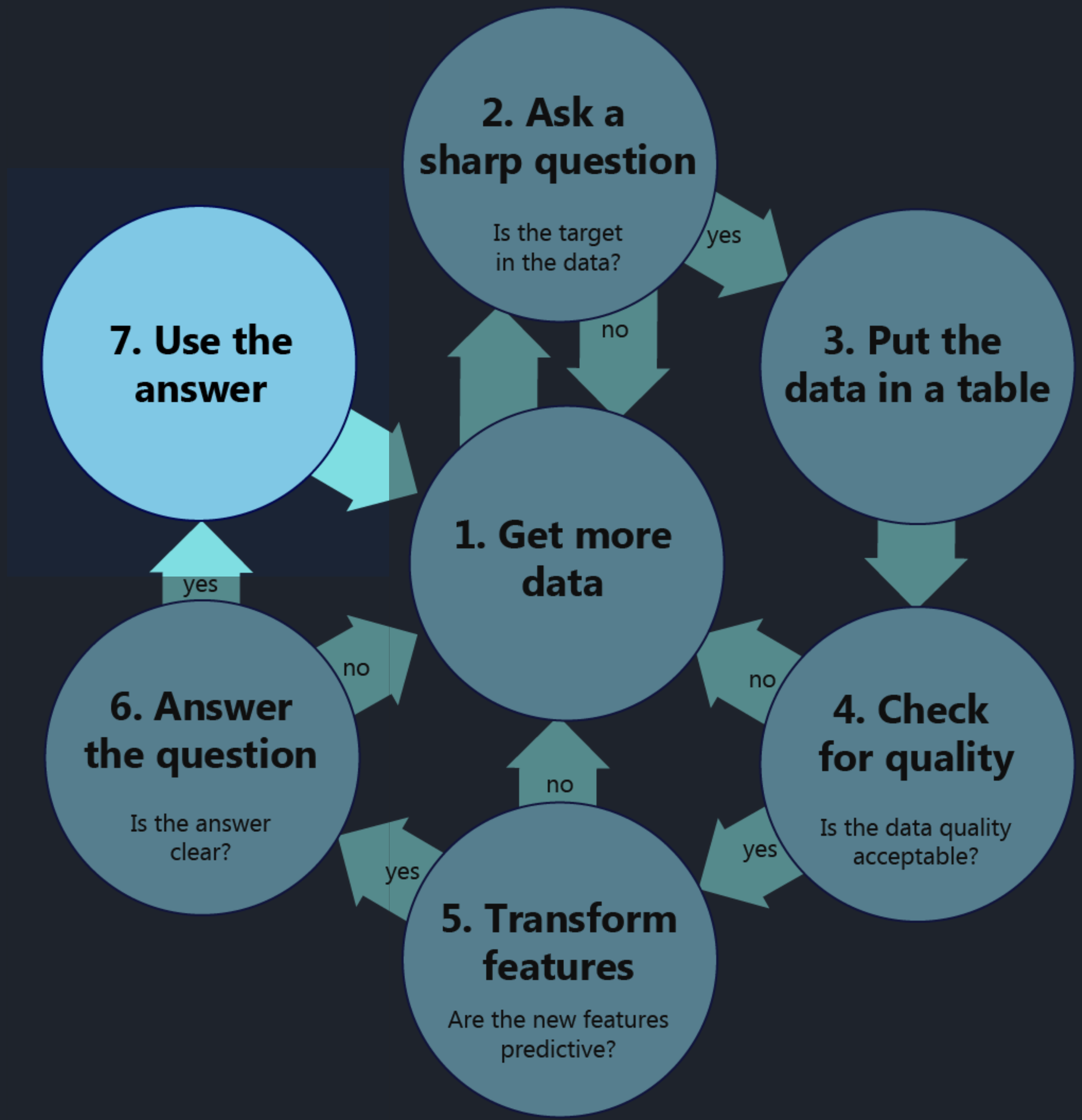
Not enough data



Barely enough data



Enough data


2. Ask a sharp question
Is the target in the data?
yes
no
3. Put the data in a table
7. Use the answer
1. Get more data
no
no
4. Check for quality
Is the data quality acceptable?
yes
yes
6. Answer the question
Is the answer clear?
no
yes
5. Transform features
Are the new features predictive?

If a tree falls in the forest and no one is around to hear it, it might still make a sound, but if you build a brilliant model and no one sees it, it will certainly not get you a raise.

Ways to use your answer

Make a web service (Azure Machine Learning)

Make a decision

Set a price

Publish your code on GitHub

Write a PDF showing your results

Build a dash board (Power BI)

Gap 1

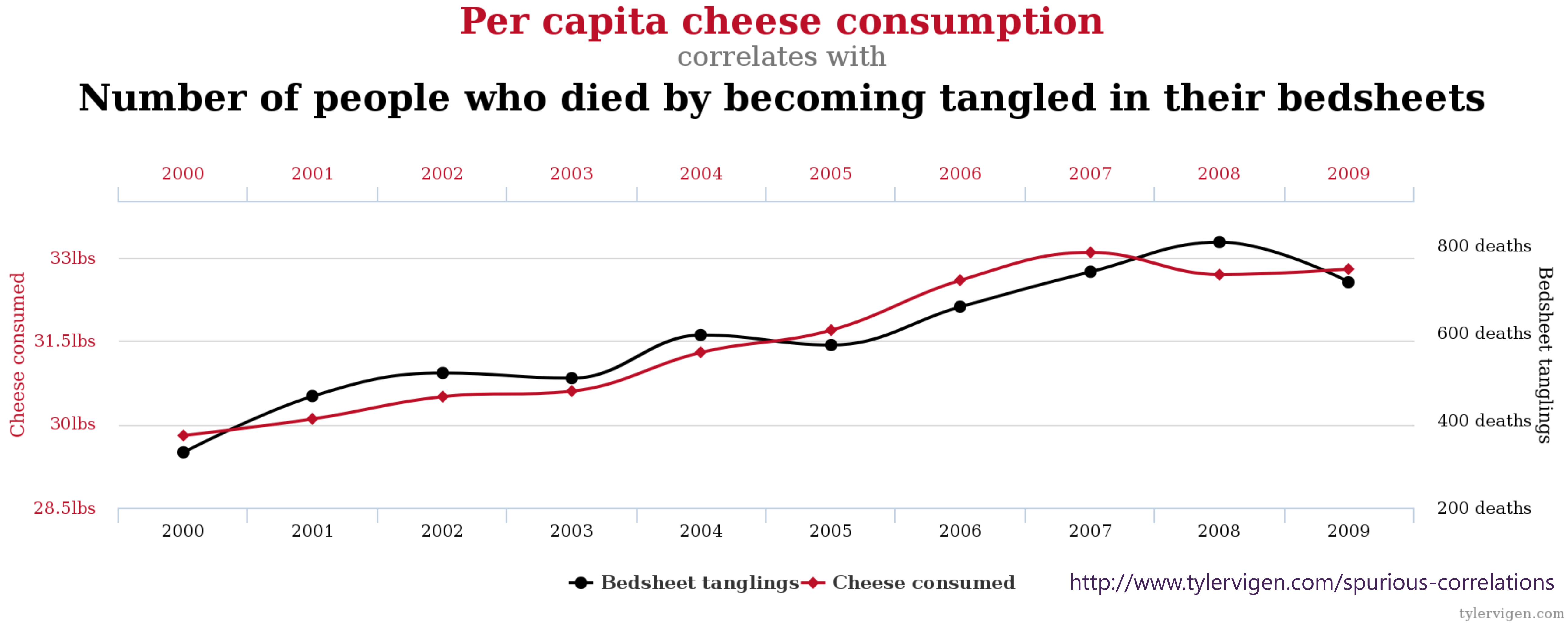
Nearly all machine learning algorithms assume that the world does not change.

Gap 2

Most machine learning algorithms take a lot of examples to learn.

Gap 3

Machine learning can't tell what caused what.

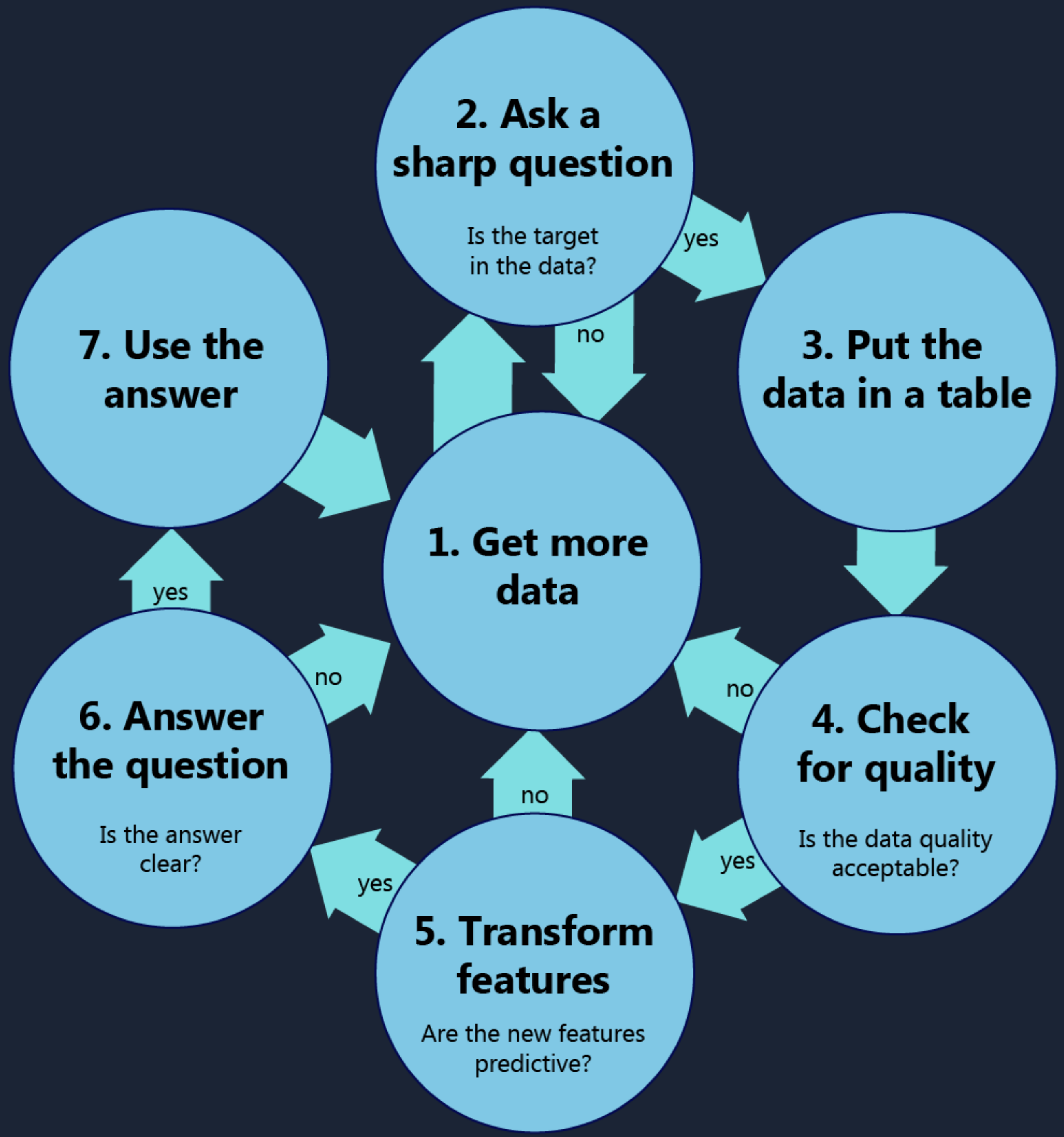


http://www.tylervigen.com/spurious-correlations

Human insight and judgment close the gap

We're good at making reasonable guesses without enough information

http://150mph.planetrambler.com/090914-18_GreatWesternDivide/090914_0046.jpg


2. Ask a sharp question
Is the target in the data?
yes
no
3. Put the data in a table
7. Use the answer
1. Get more data
yes
no
no
4. Check for quality
Is the data quality acceptable?
yes
6. Answer the question
Is the answer clear?
no
yes
5. Transform features
Are the new features predictive?

Resources

1. Get more data.	
2. Ask a sharp question.	Asking a question
3. Put the data in a table.	
4. Check for quality.	Methods for handling missing values
5. Transform features.	Feature engineering example Demystifying neural networks
6. Answer the question.	Turn your data into a picture Questions machine learning can answer Algorithms for business use cases Machine learning algorithm cheat sheet Choosing a machine learning algorithm
7. Use the answer.	Cortana Intelligence Gallery
Presentations	Data Science for Absolutely Everyone (slides) Data Science 101 (slides) The Other Stuff (slides)

Thanks!

Questions? Want to chat about data?

LinkedIn Brandon Rohrer

@_brohrer_

brohrer.github.io

BRohrer@microsoft.com

Microsoft

© 2015 Microsoft Corporation. All rights reserved.