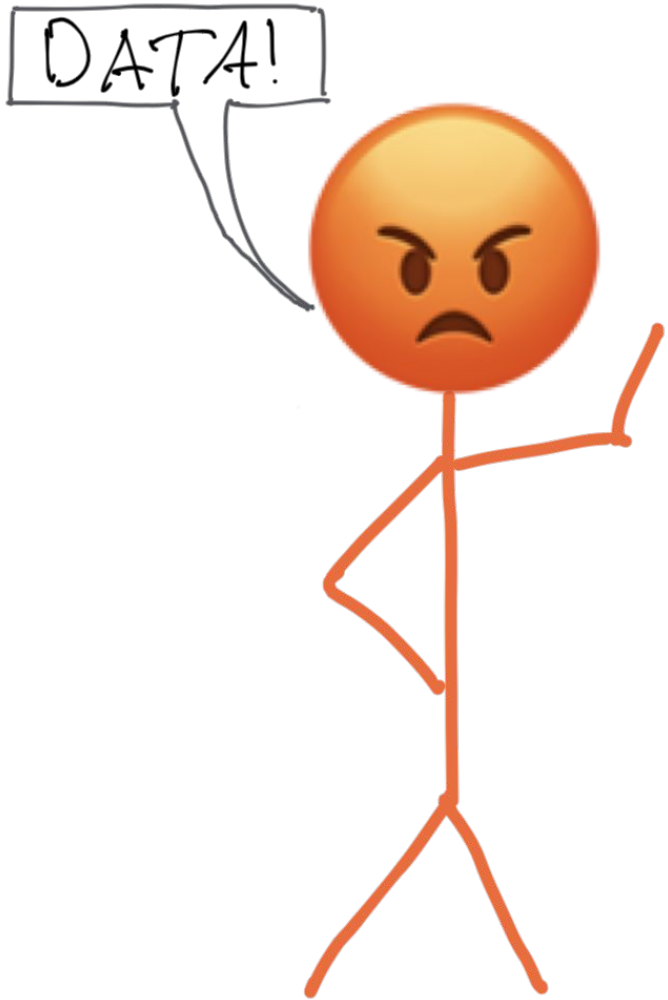
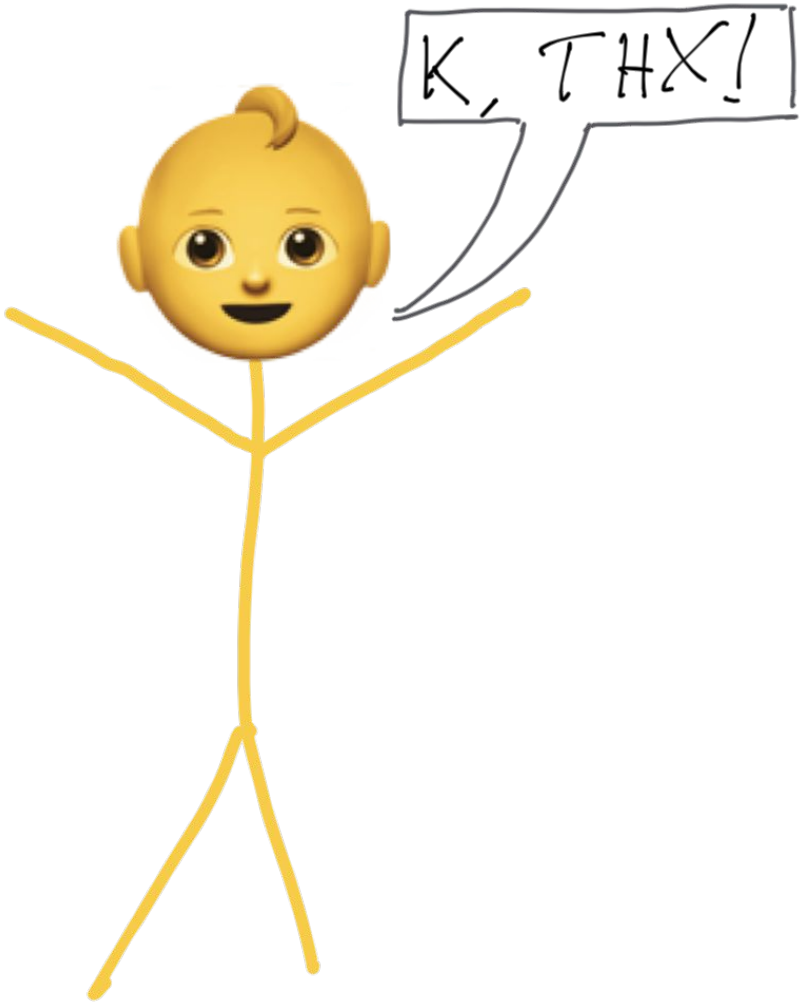
Making Friends w/ Generative Models

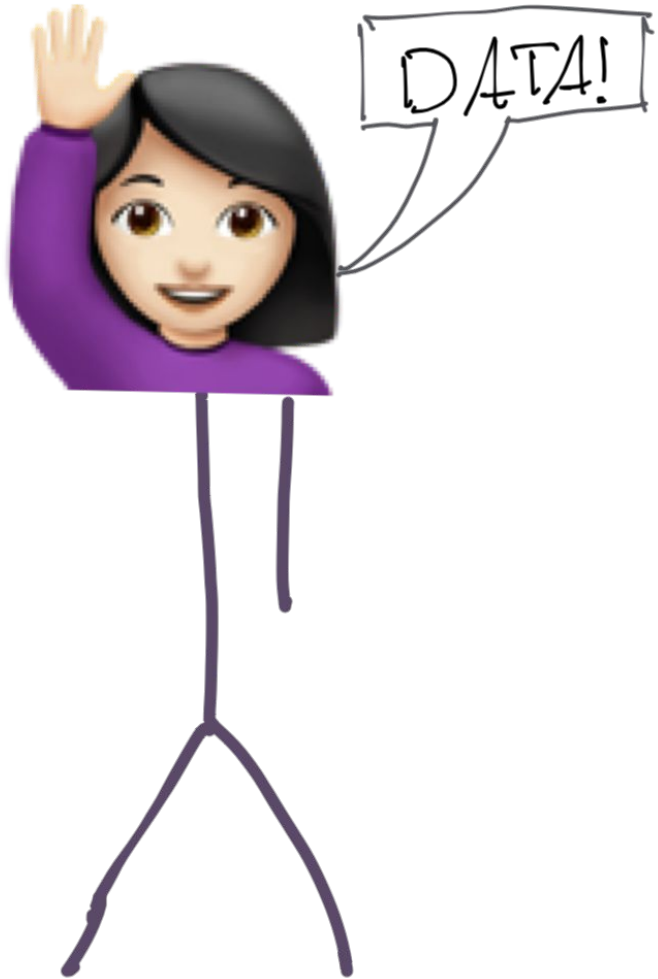
How to make synthetic data for sales, data science, development and testing

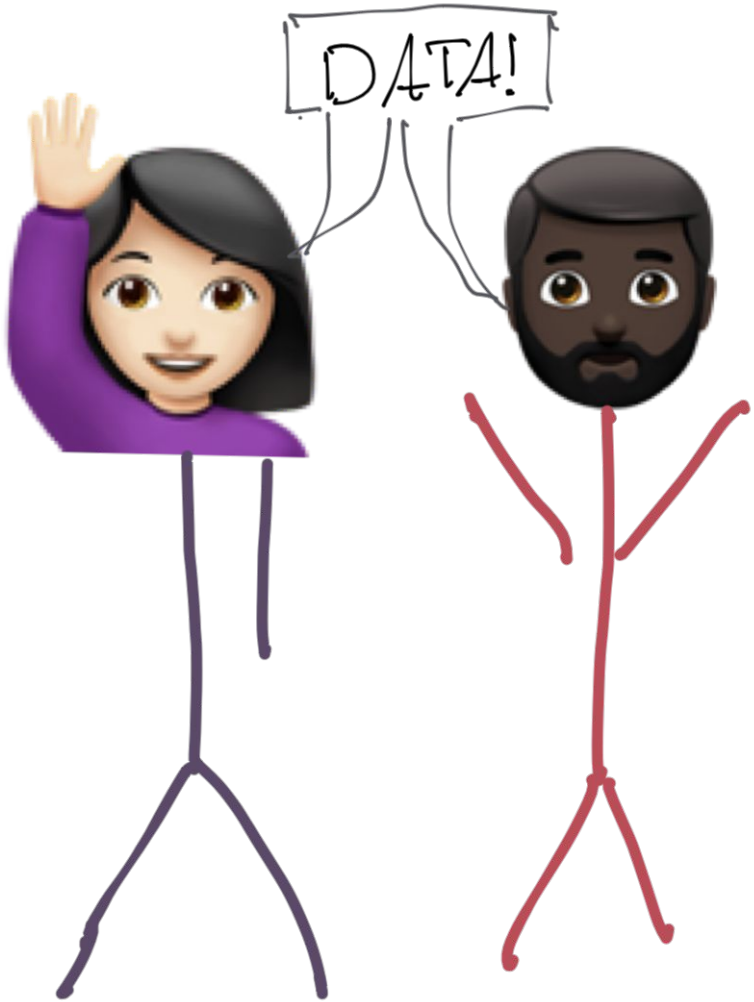
Andrew Colombi, CTO @ Tonic, www.tonic.ai


financial crisis
eat Crash of 2008
Disbelief, and a punter reach
The plunging market yesterday dealt a new blow to investors' confidence, as the government decided to inject HK$ million into counselling services f losers in the financial crisi
Retail investors
disbelief at
Central
push the US and European
mies into a deep downturn or even
recession, prolonging the tough
times for Hong Kong's trade sector –
the industry that employs the most
workers.
"Major real estate agencies have
also started to cut headcount as
perty transactions turn more
k revised its economic
for this year down to
per cent. For next
to 2 per cent,
main gauge of inflati
cent this year and 4.
Although the ac
2008 GDP growth
previous forecast
still more optimist
consensus of 9.9
9.1 per cent in 20
"The slowdo
a normal fluctu
in a report yest
enough eviden
economic dow
Last night
Chao (李超),
spokesman,
with
SATURDAY, OCTOBER


DATA!


K, THX!


DATA!


DATA!


K, THX!

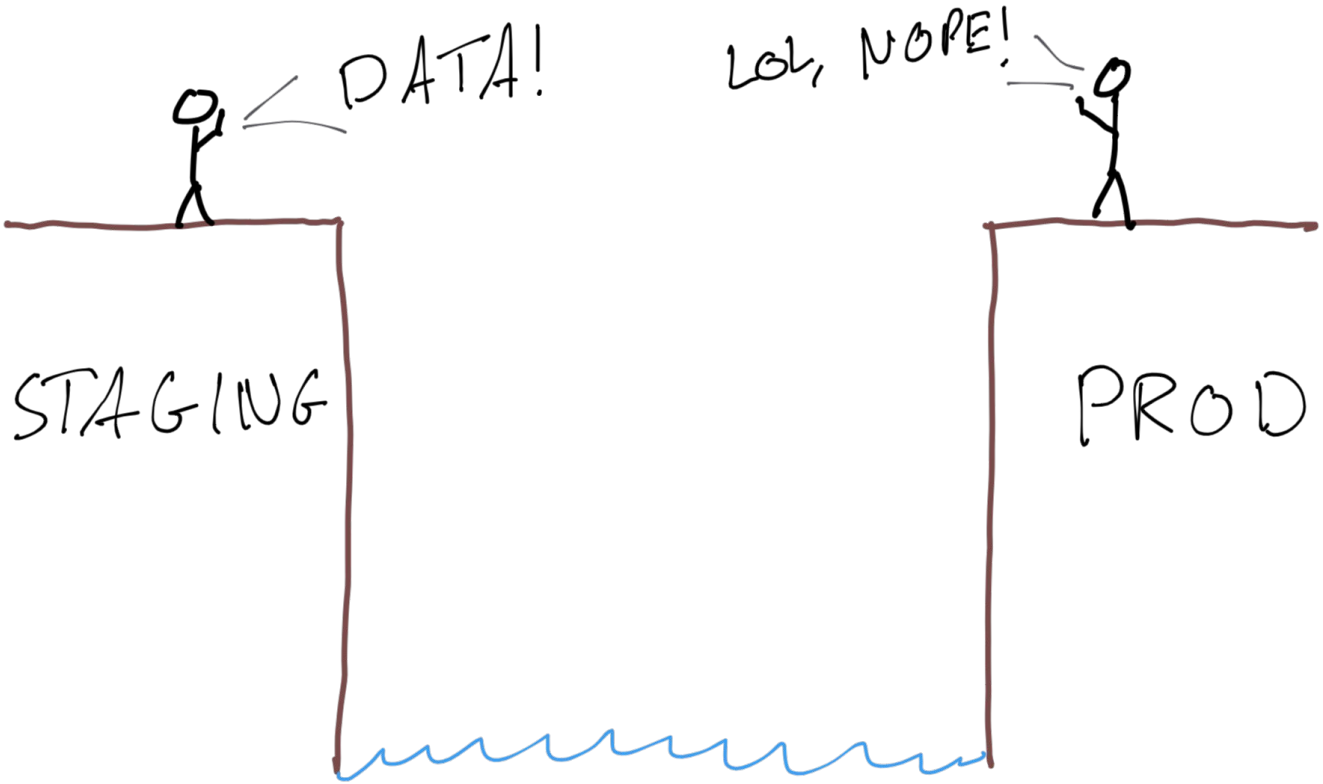
Framework

Framework for Making Friends

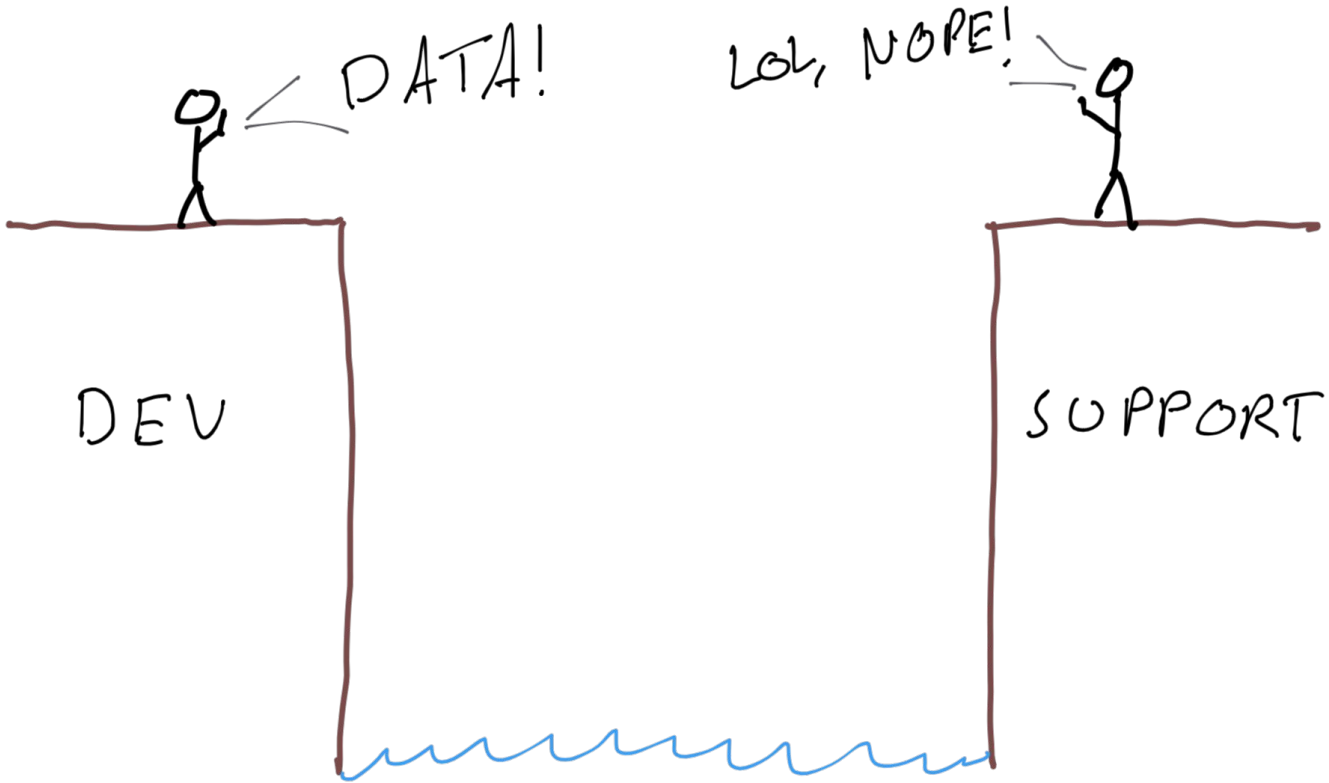
1. Identify situations where data deficiency occur
2. Target the right datasets
3. Toolbox of simple modeling techniques

Identifying Data Deficiencies

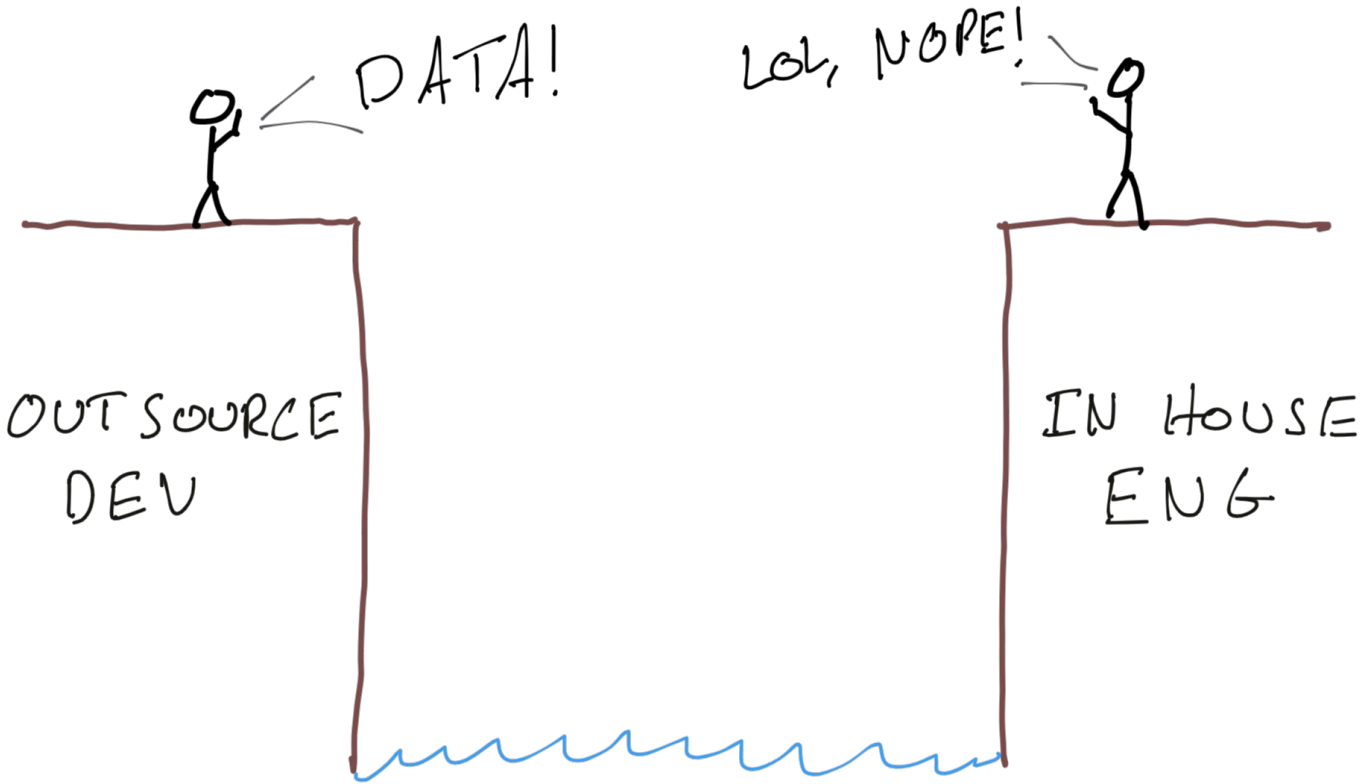
Identifying Data Deficiencies: Boundaries



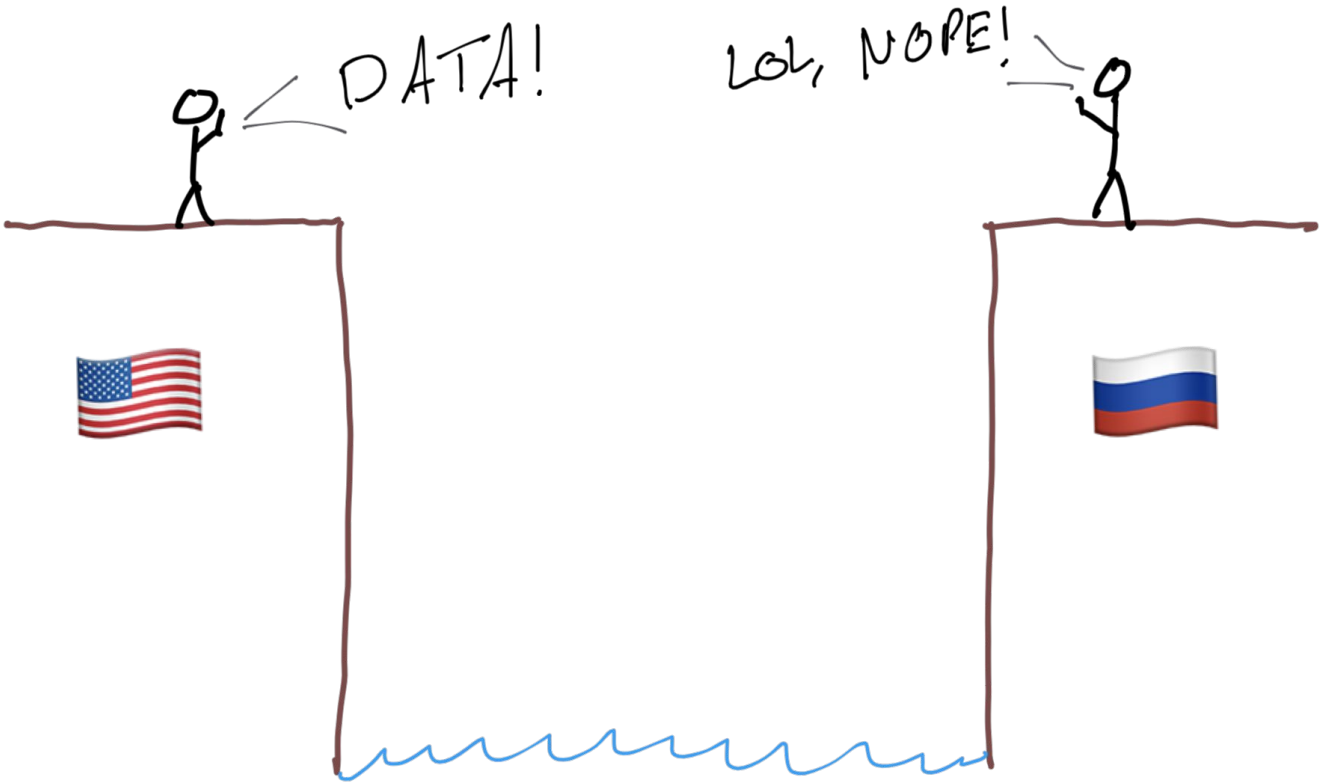
Identifying Data Deficiencies: Boundaries



Identifying Data Deficiencies: Boundaries



Identifying Data Deficiencies: Boundaries



Identifying Data Deficiencies: Regulatory

FERPA

Family Educational
Rights & Privacy Act

Identifying Data Deficiencies: Certification

Identifying Data Deficiencies

- Boundaries
- Regulations
- Certification

Targeting Datasets

Targeting Datasets

Target structured data.

Mortgage Data Example

Continuous

size_sqft

date_built

Continuous

size_sqft

date_built

Categorical

roof_style

has_pool

Continuous

size_sqft

date_built

Categorical

roof_style

has_pool

Relational

agent_id

owner_id

Continuous

- size_sqft
- date_built

Events

- foreclosure
- sold
- appraisal

Categorical

- roof_style
- has_pool

Relational

- agent_id
- owner_id

Continuous

- size_sqft
- date_built

Events

- foreclosure
- sold
- appraisal

Text

- appraisal_notes
- offer_transcript

Categorical

- roof_style
- has_pool

Relational

- agent_id
- owner_id

Continuous

- size_sqft
- date_built

Events

- foreclosure
- sold
- appraisal

Text

- appraisal_notes
- offer_transcript

Categorical

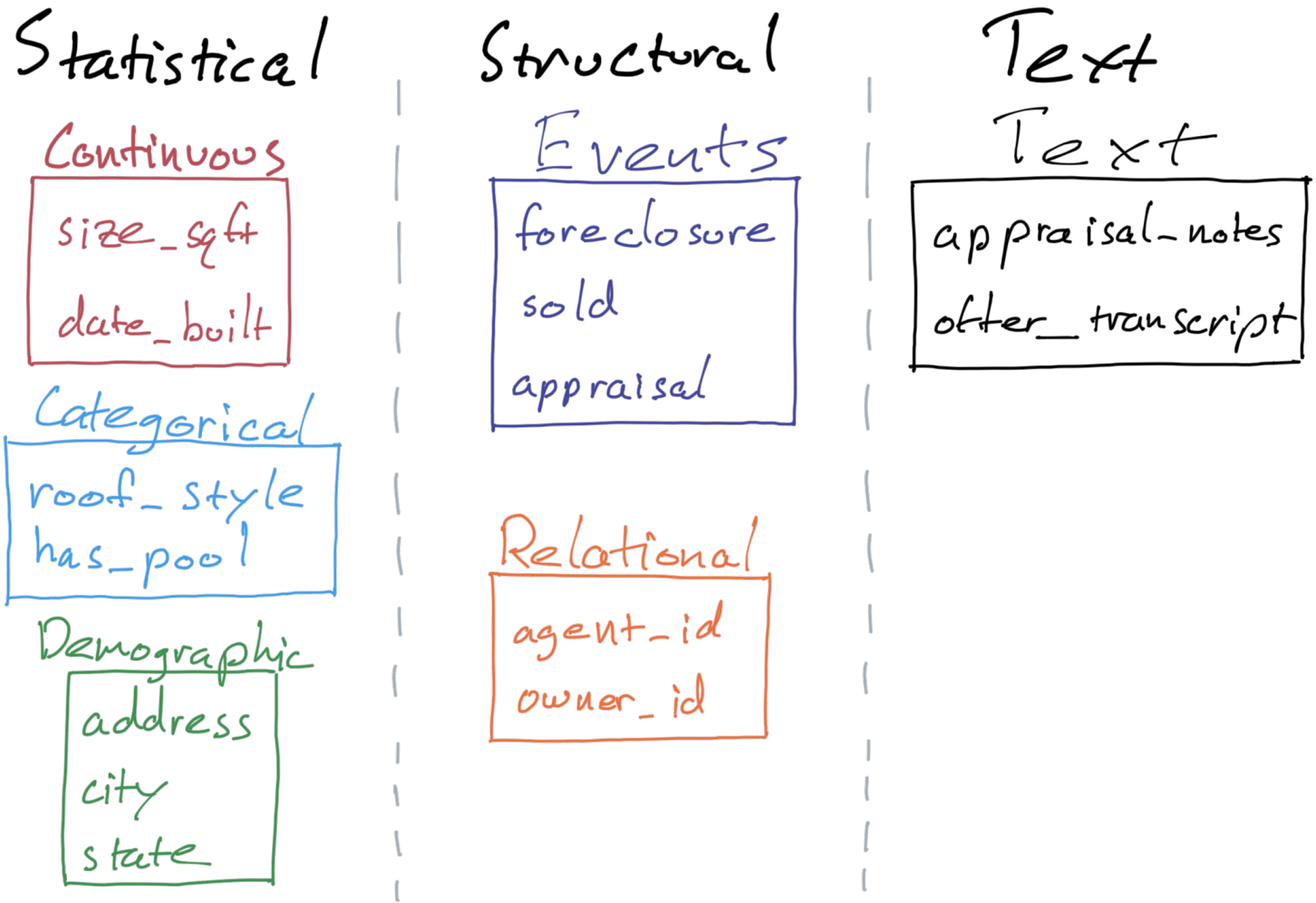
- roof_style
- has_pool

Demographic

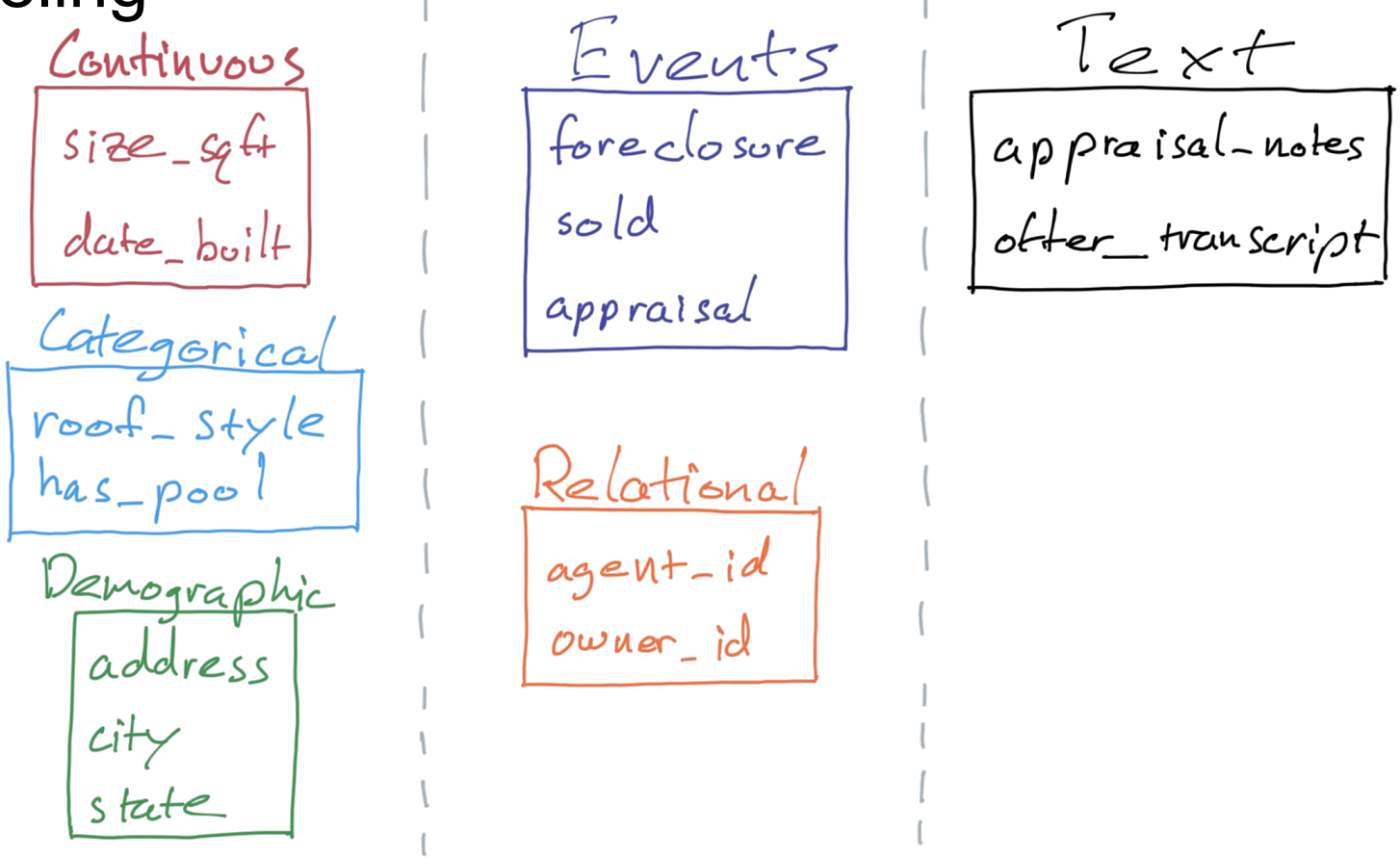
- address
- city
- state

Relational

- agent_id
- owner_id

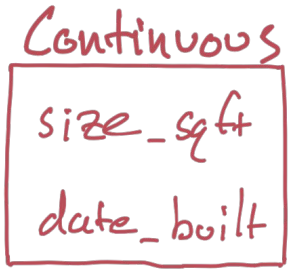

Statistical
Continuous
size_sqft
date_built
Categorical
roof_style
has_pool
Demographic
address
city
state
Structural
Events
foreclosure
sold
appraisal
Relational
agent_id
owner_id
Text
Text
appraisal_notes
offer_transcript

Modeling


Continuous
size_sqft
date_built
Categorical
roof_style
has_pool
Demographic
address
city
state
Events
foreclosure
sold
appraisal
Relational
agent_id
owner_id
Text
appraisal_notes
otter_transcript

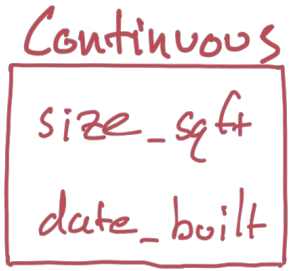
Modeling: f(x) = x

Modeling: Continuous



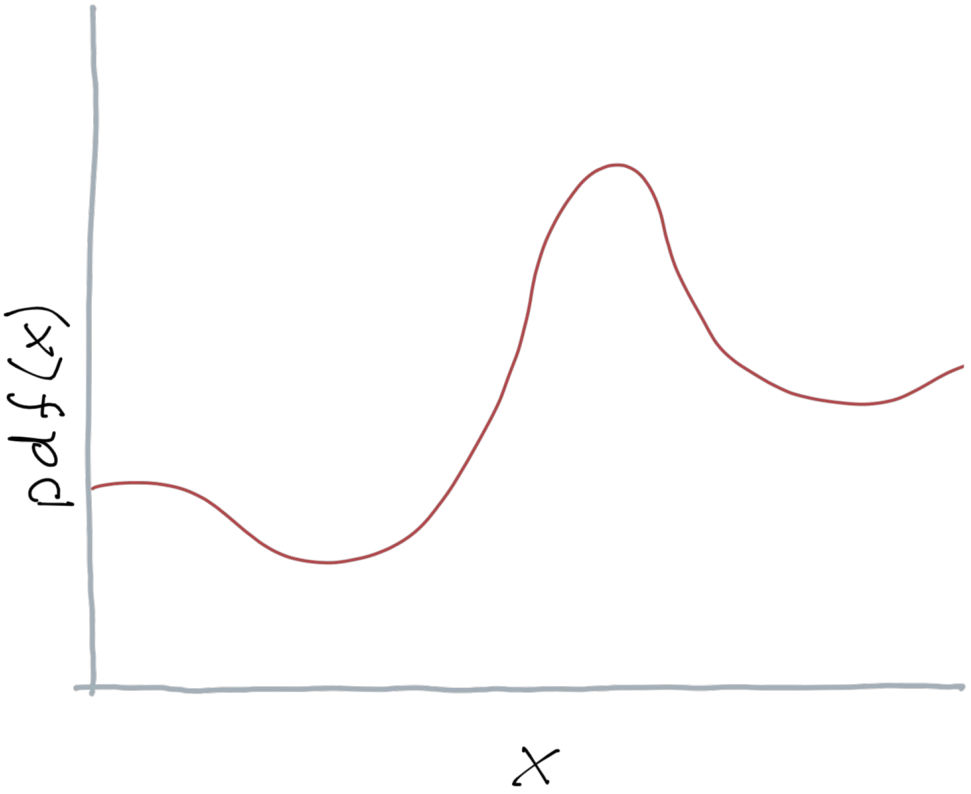
Kernel Density Estimation

Modeling: Kernel Density Estimation



- Models continuous variables
- Non-Parametric: Don’t need to specify distribution upfront
- Easy to extend to multiple dimensions
- Easy to implement

Modeling: Kernel Density Estimation



Continuous

size_sqft

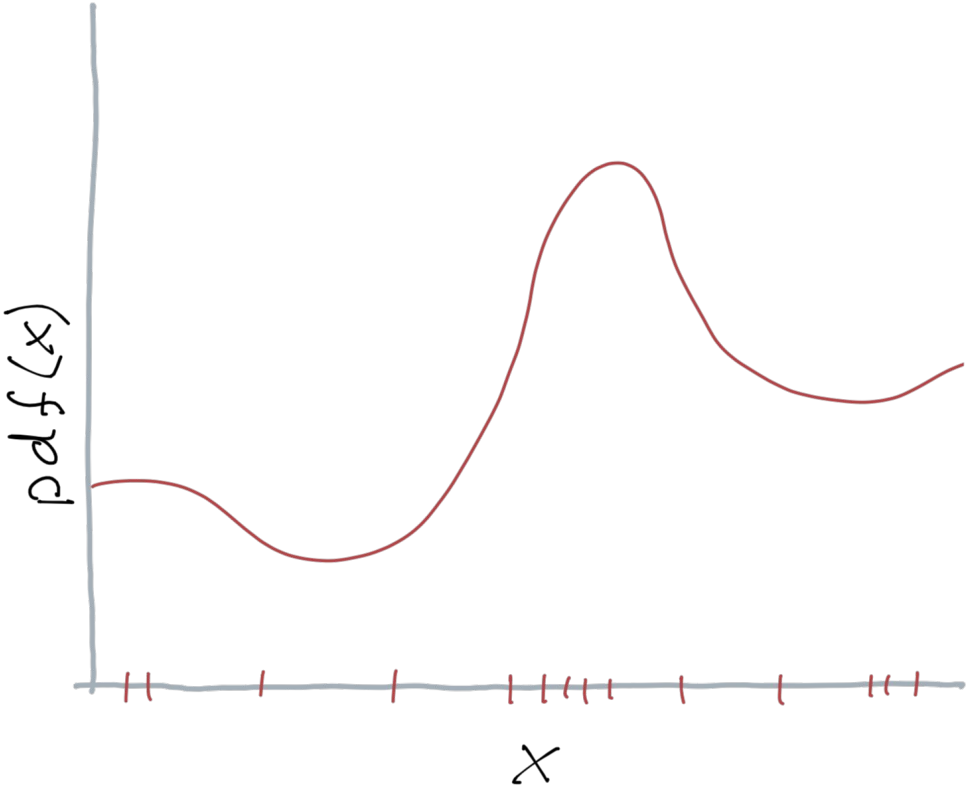
date_built

Modeling: Kernel Density Estimation

Continuous

size_sqft

date_built

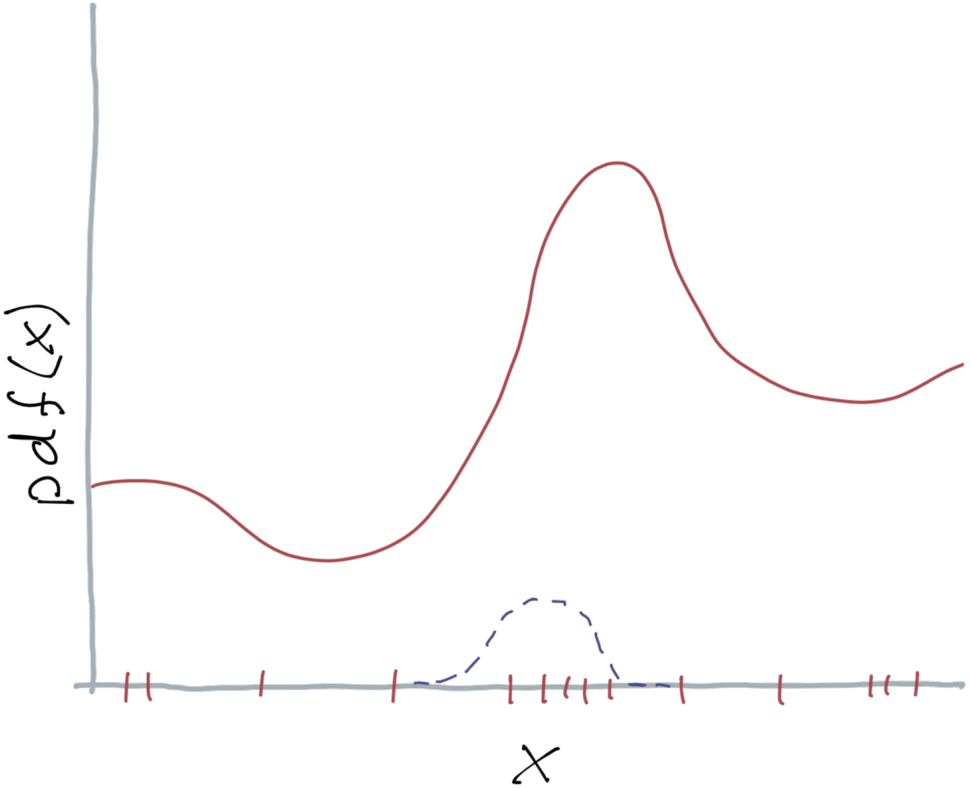


Modeling: Kernel Density Estimation

Continuous

size_sqft

date_built

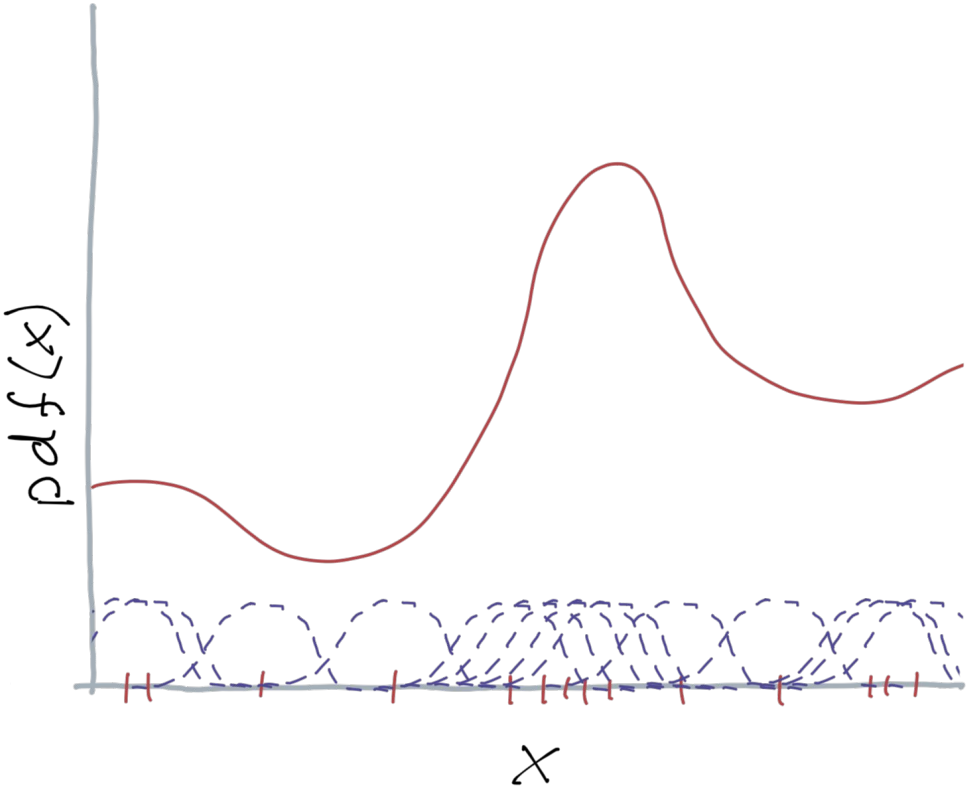


Modeling: Kernel Density Estimation

Continuous

size_sqft

date_built

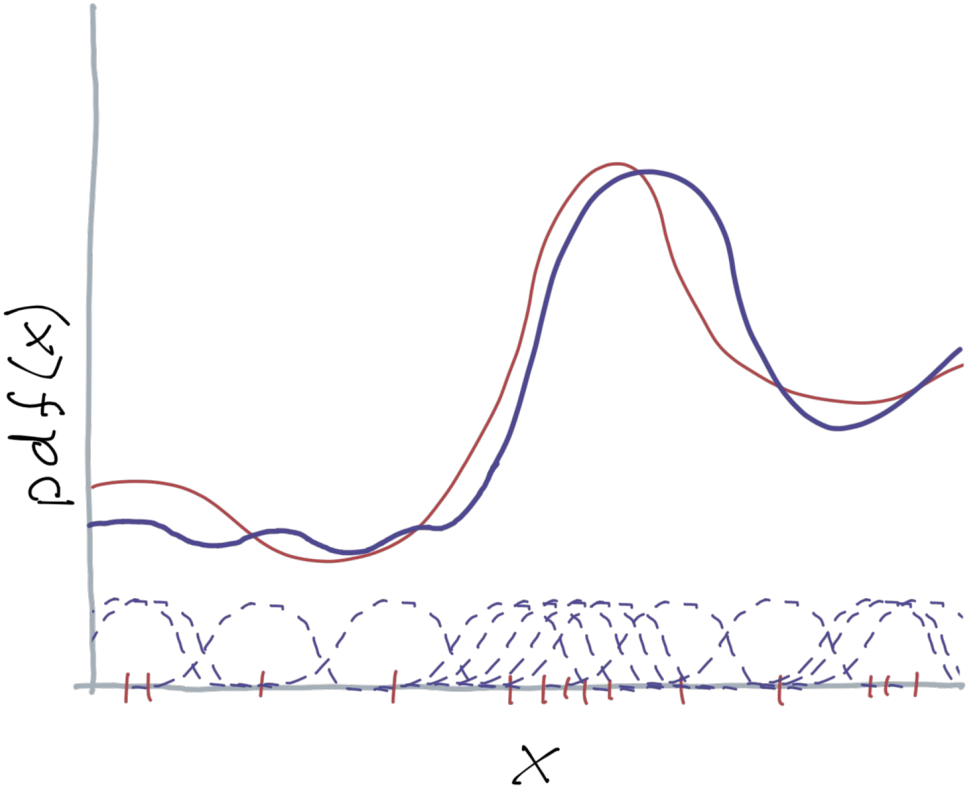


Modeling: Kernel Density Estimation

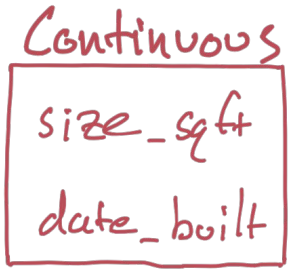
Continuous

size_sqft

date_built

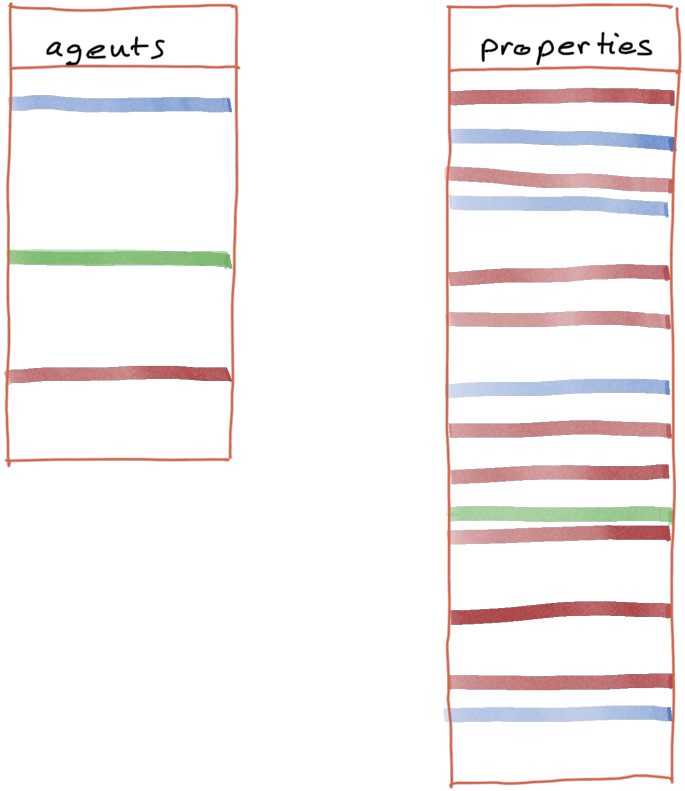


Modeling: Kernel Density Estimation



- Normal Distribution is a common kernel
- For K-dimensions use a K-dimensional kernel
- Sampling from a KDE:
  1. Pick a random kernel
  2. Sample randomly from that kernel
- *(There is one parameter, the kernel width, which needs to be fit.)*

Modeling: Relational



Relational

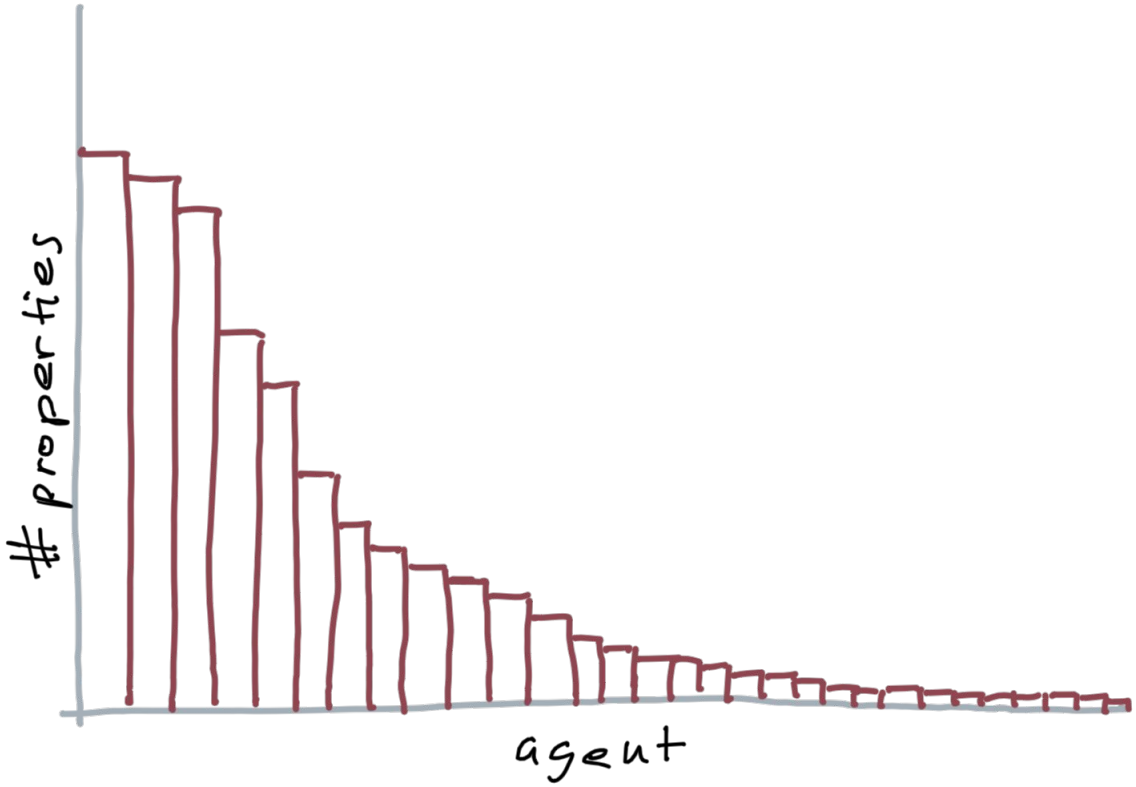
agent_id

owner_id

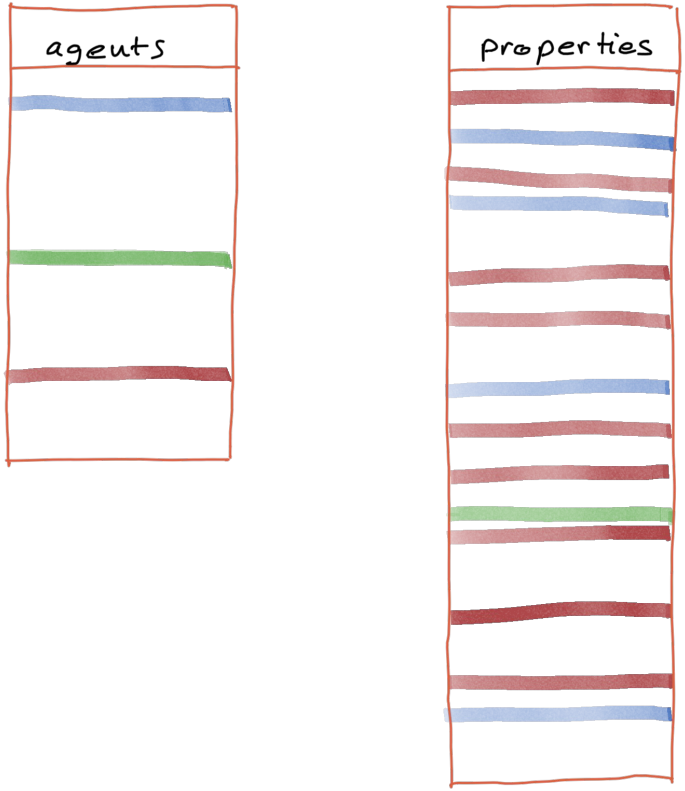
Modeling: Relational

Relational

agent_id
owner_id



Modeling: Relational

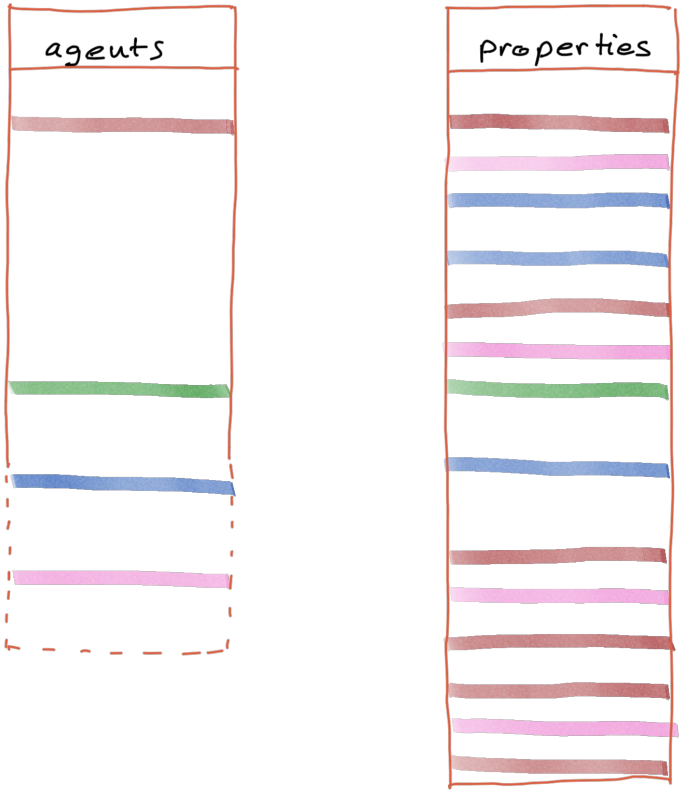


Relational

agent_id

owner_id

Modeling: Relational



Relational

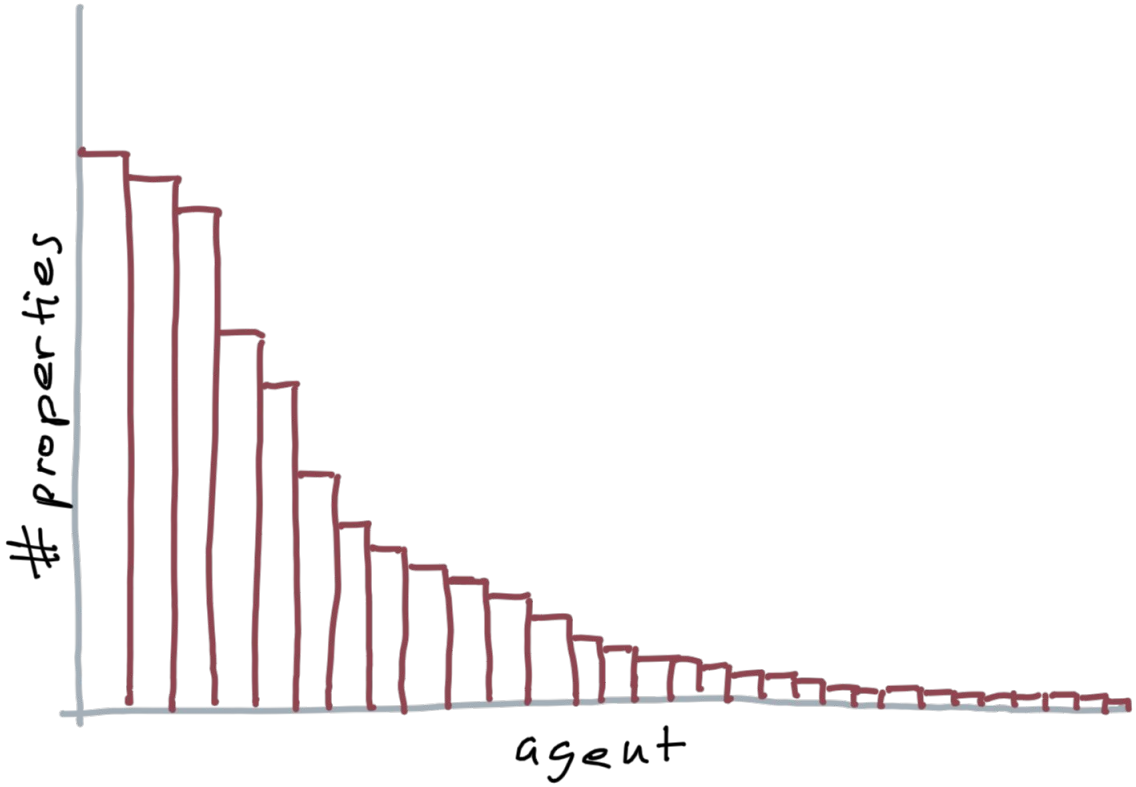
agent_id

owner_id

Modeling: Relational

Relational

agent_id
owner_id



Modeling: Relational

Relational

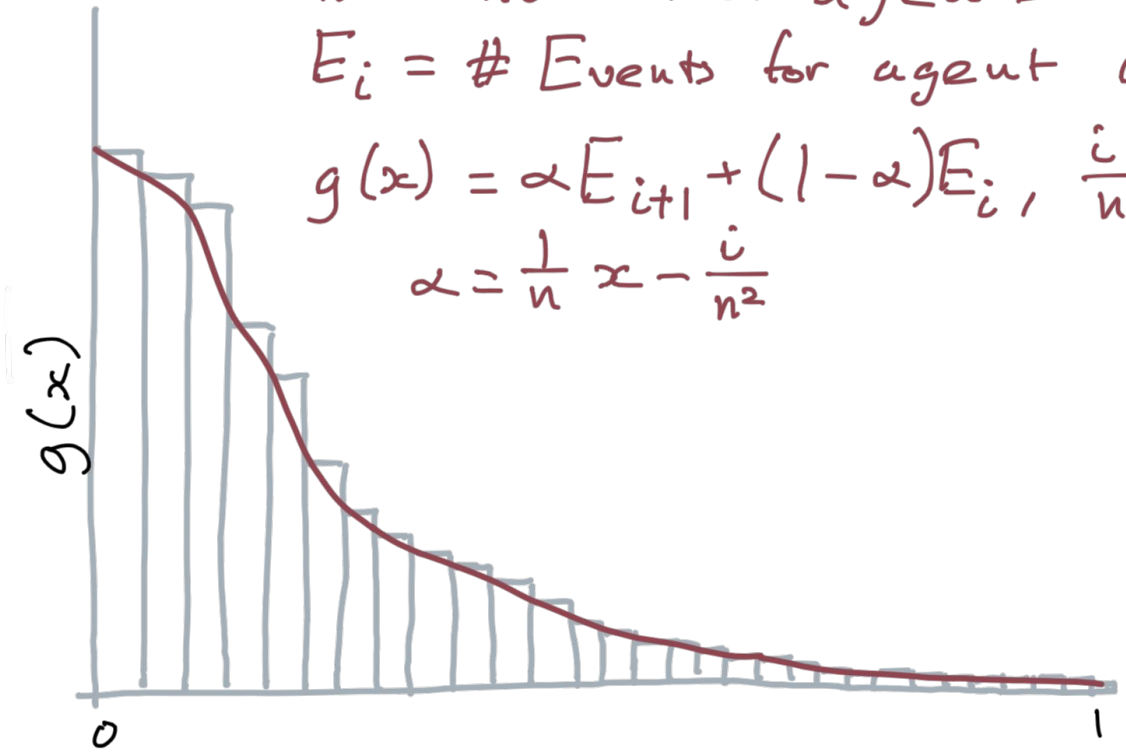
agent_id
owner_id

n = number of agents

E_i = # Events for agent i

$$g(x) = \alpha E_{i+1} + (1-\alpha) E_i, \quad \frac{i}{n} \le x < \frac{i+1}{n}$$

$$\alpha = \frac{1}{n} x - \frac{i}{n^2}$$

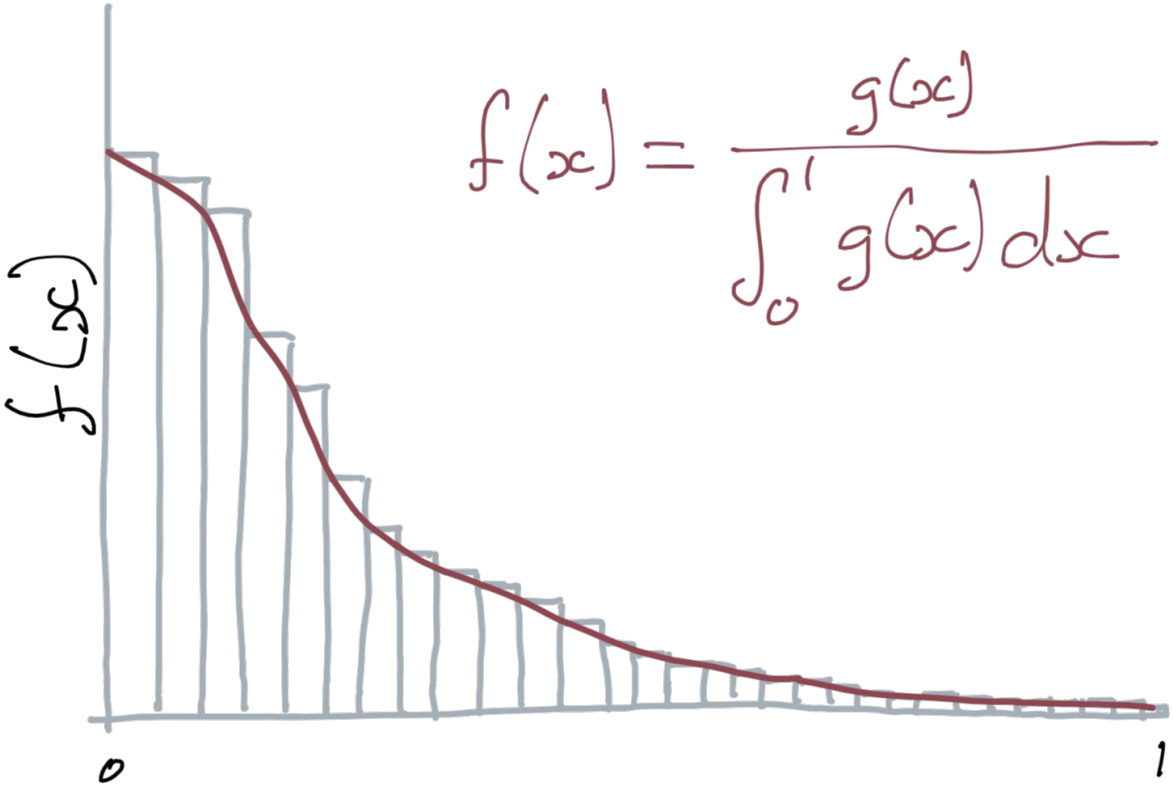


Modeling: Relational

Relational

agent_id

owner_id

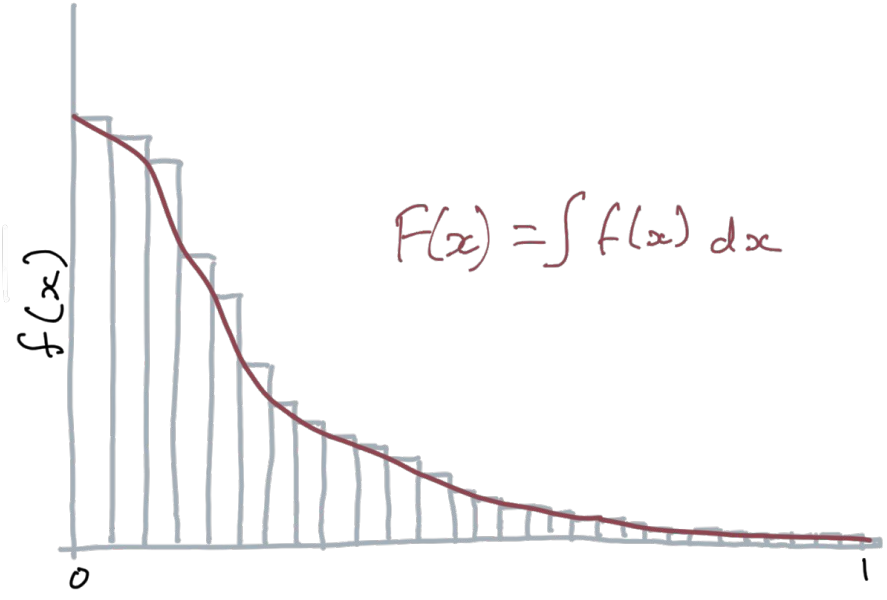


Modeling: Relational

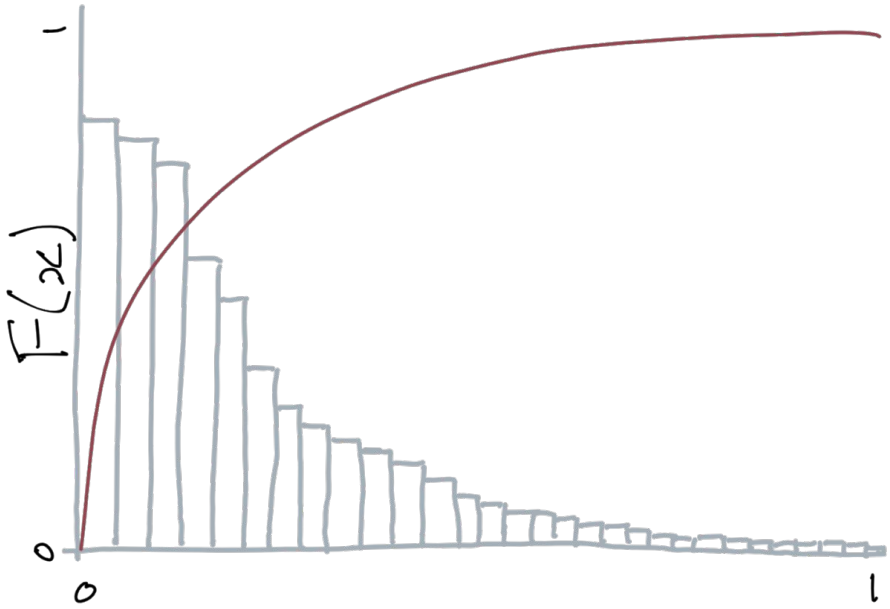
Relational

agent_id

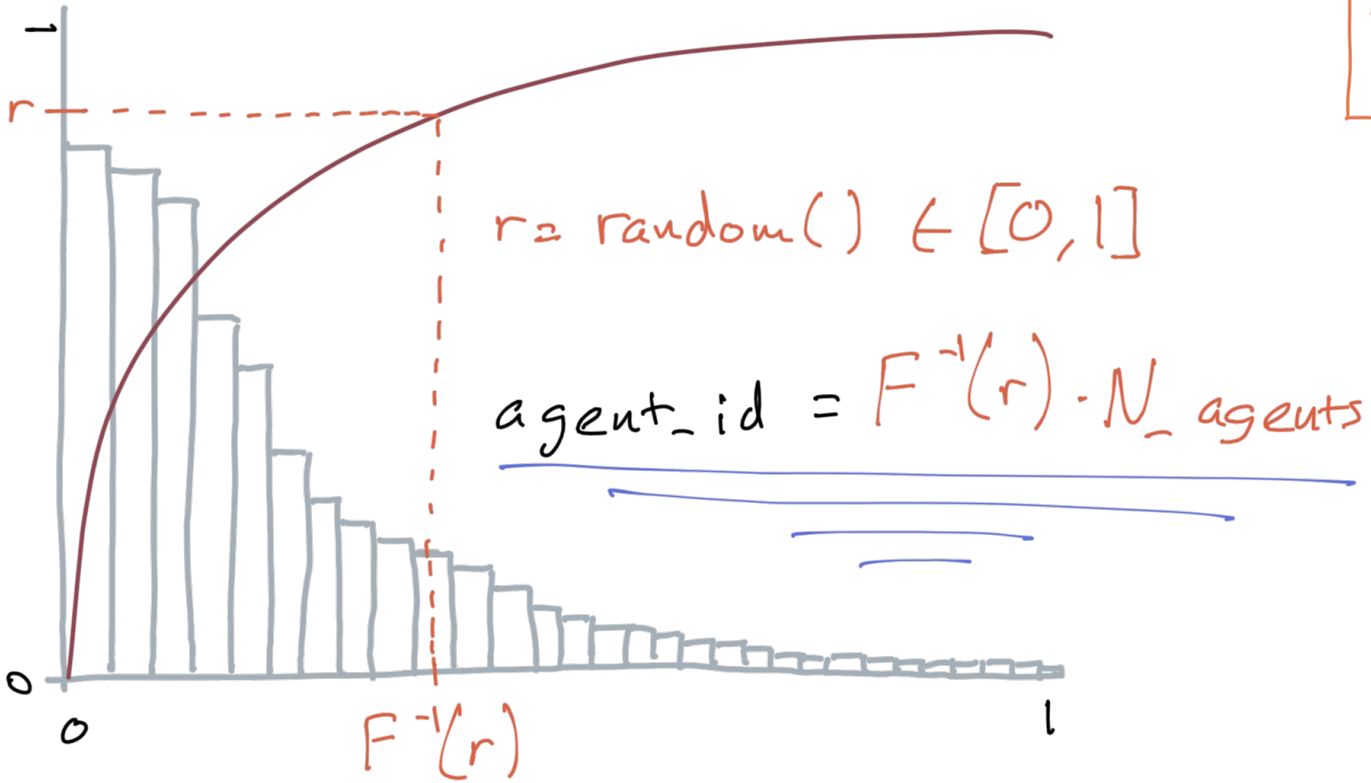
owner_id



$$F(x) = \int f(x)\,dx$$



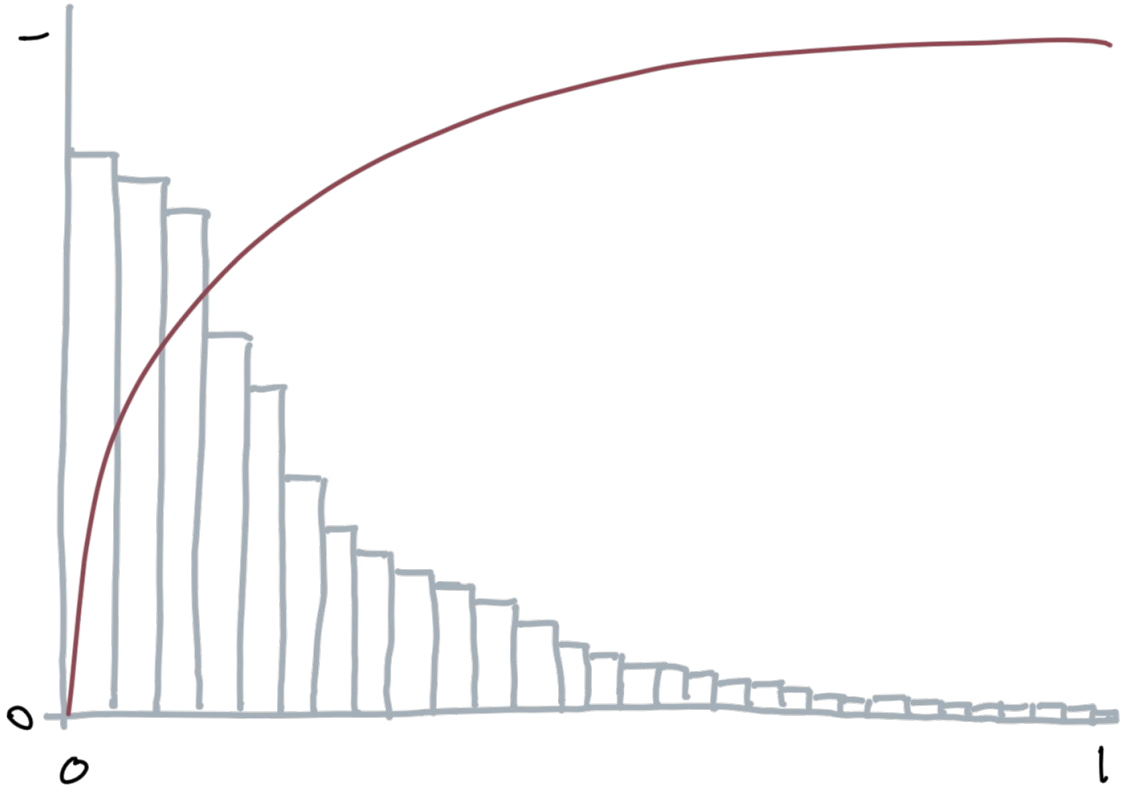
Modeling: Relational



Relational

agent_id
owner_id

Modeling: Relational

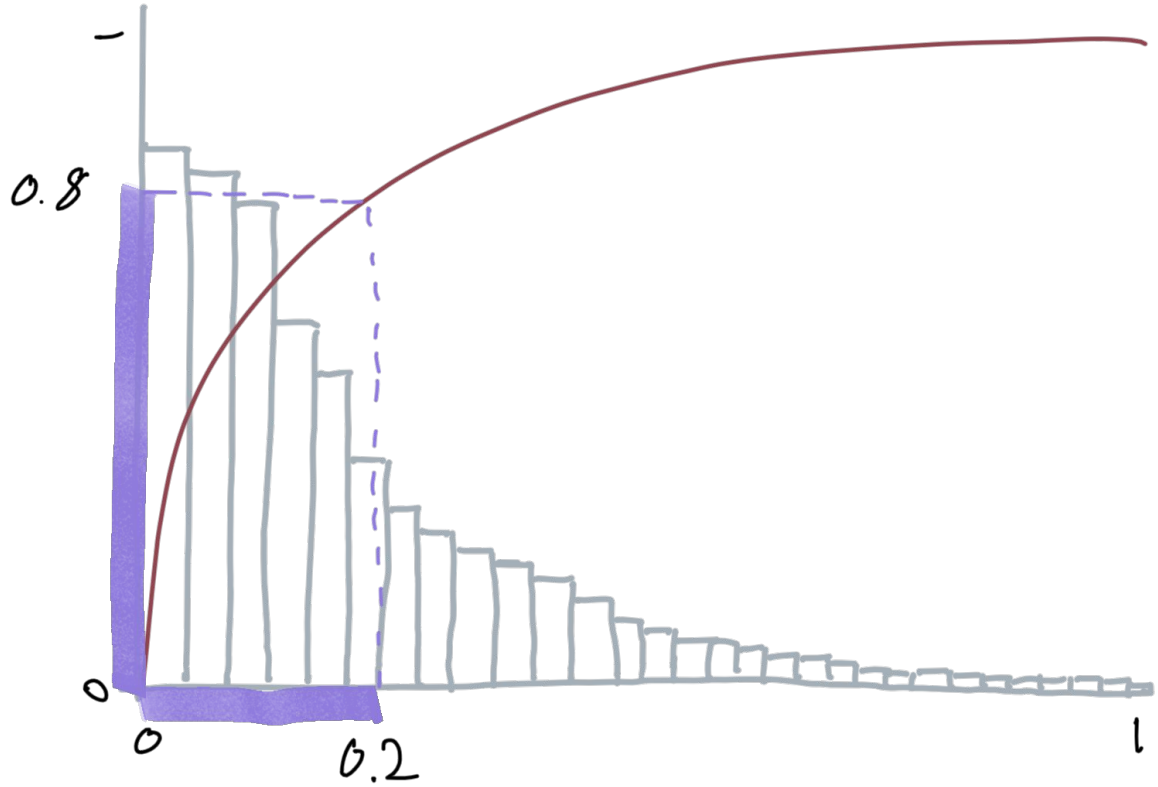


Relational

agent_id

owner_id

Modeling: Relational

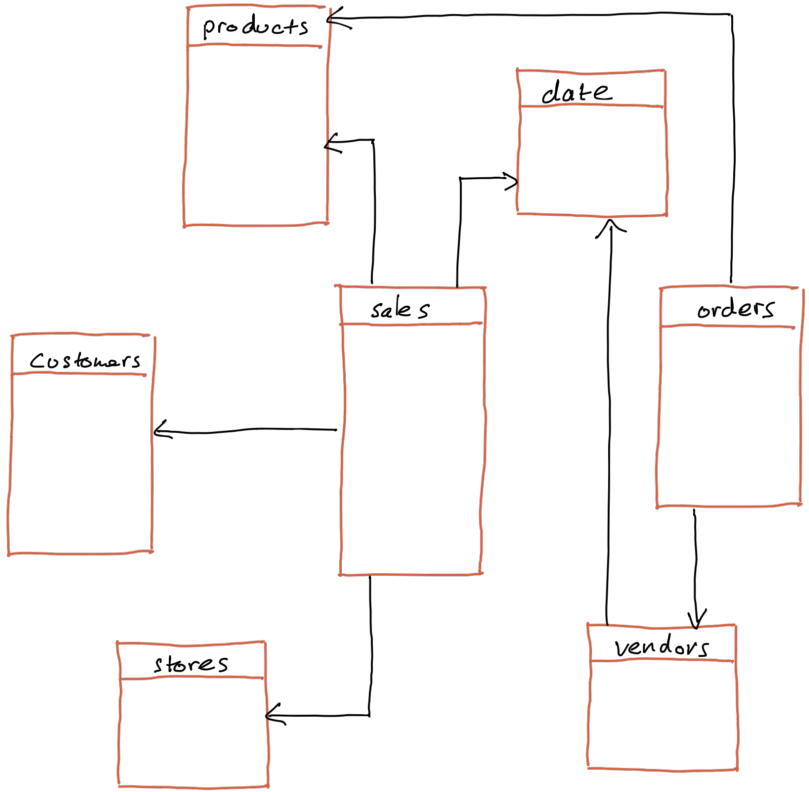


Relational

agent_id

owner_id

Modeling: Relational

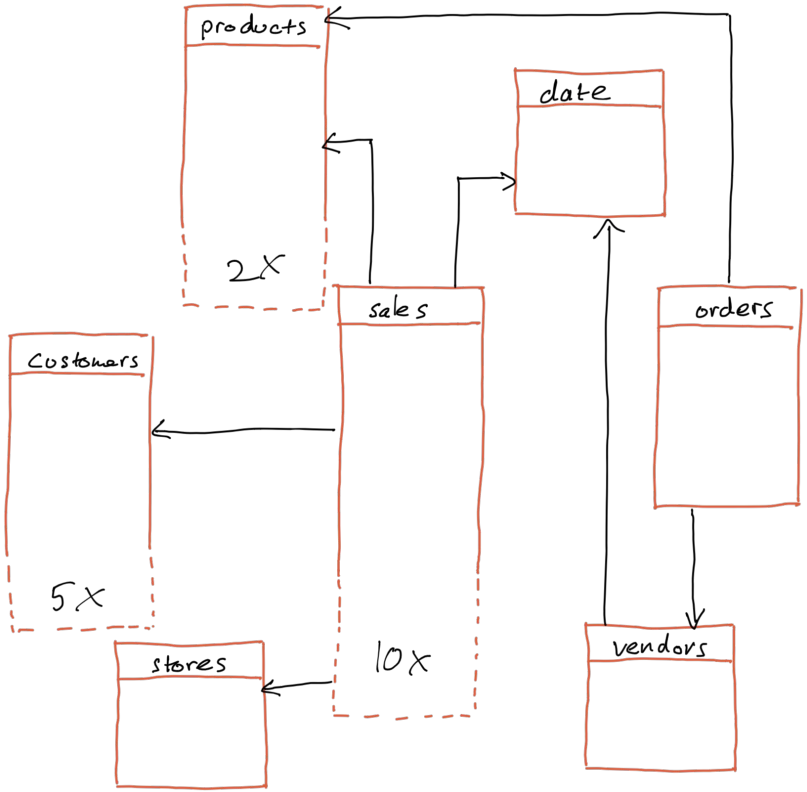


Relational

agent_id

owner_id

Modeling: Relational

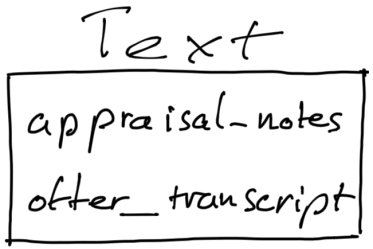


Relational

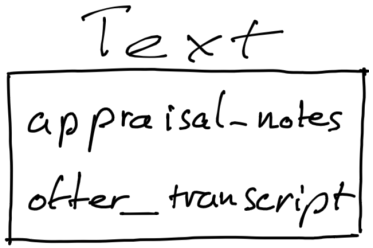
agent_id

owner_id

Modeling: Text



Modeling: Text



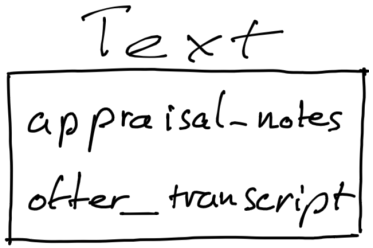
Jane Little (DOB: 9-6-1977)

Jane Little seems to have had an inadequate response to treatment as yet. Symptoms of depression continue to be described. Her symptoms, as noted, are unchanged and they are just as frequent or intense as previously described. Jane Little describes feeling sad. Jane Little denies suicidal ideas or intentions. Jane Little reports the symptoms of this disorder continue unchanged. The subjective feeling of apprehension is occurring. Hypervigilance is occurring more frequently.

Participant(s) Developing the Plan:
- Susan Lobao (Counselor)
- Mary Golden (Client)

Diagnosis:
- Major depressive disorder, single episode, severe without psychotic features, F32.2 (ICD-10) (Active)
- Anxiety disorder, unspecified, F41.9 (ICD-10) (Active)

Modeling: Text



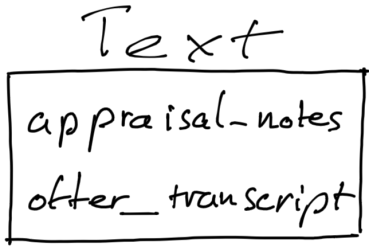
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse eleifend justo nec purus vehicula, in tempus libero tristique. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. In ultrices justo id urna ullamcorper laoreet. Curabitur dictum accumsan nunc, eu porttitor magna aliquet at. Ut faucibus, arcu quis condimentum sodales, erat tellus dapibus purus, in luctus sapien tellus at nisl. Duis convallis mattis mi, non pharetra diam commodo ac. Cras sit amet diam ac purus laoreet dignissim. Duis lectus diam, rutrum non vehicula ac, convallis non metus. Donec et luctus sem. Mauris nec erat volutpat, placerat ex ut, varius ipsum. In ultrices hendrerit nibh, id aliquet est. Maecenas vel turpis velit. Mauris ut faucibus lacus.

Ut diam mi, ultrices vitae ante vitae, porttitor volutpat libero. Phasellus et nunc id tellus varius semper. Sed a euismod diam, id consequat risus. Donec vitae ornare magna. Nullam semper mi et magna suscipit posuere. Proin tincidunt leo augue, in facilisis turpis molestie ac. Sed convallis viverra massa et consectetur.

Modeling: Text

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse eleifend justo nec purus vehicula, in tempus libero tristique. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. In ultrices justo id urna ullamcorper laoreet. Curabitur dictum accumsan nunc, eu porttitor magna aliquet at. Ut faucibus, arcu quis condimentum sodales, erat tellus dapibus purus, in luctus sapien tellus at nisl. Duis convallis mattis mi, non pharetra diam commodo ac. Cras sit amet diam ac purus laoreet dignissim. Duis lectus diam, rutrum non vehicula ac, convallis non metus. Donec et luctus sem. Mauris nec erat volutpat, placerat ex ut, varius ipsum. In ultrices hendrerit nibh, id aliquet est. Maecenas vel turpis velit. Mauris ut faucibus lacus.

Ut diam mi, ultrices vitae ante vitae, porttitor volutpat libero. Phasellus et nunc id tellus varius semper. Sed a euismod diam, id consequat risus. Donec vitae ornare magna. Nullam semper mi et magna suscipit posuere. Proin tincidunt leo augue, in facilisis turpis molestie ac. Sed convallis viverra massa et consectetur.



Modeling: Text

Jane Little (DOB: 9-6-1977)

Jane Little seems to have had an inadequate response to treatment as yet. Symptoms of depression continue to be described. Her symptoms, as noted, are unchanged and they are just as frequent or intense as previously described. Jane Little describes feeling sad. Jane Little denies suicidal ideas or intentions. Jane Little reports the symptoms of this disorder continue unchanged. The subjective feeling of apprehension is occurring. Hypervigilance is occurring more frequently.

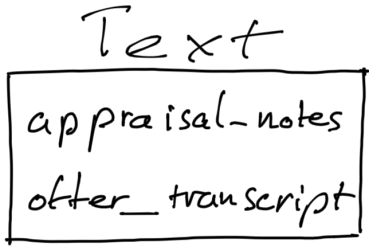
Participant(s) Developing the Plan:
- Susan Lobao (Counselor)
- Mary Golden (Client)

Diagnosis:
- Major depressive disorder, single episode, severe without psychotic features, F32.2 (ICD-10) (Active)
- Anxiety disorder, unspecified, F41.9 (ICD-10) (Active)

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Suspendisse eleifend justo nec purus vehicula, in tempus libero tristique. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. In ultrices justo id urna ullamcorper laoreet. Curabitur dictum accumsan nunc, eu porttitor magna aliquet at. Ut faucibus, arcu quis condimentum sodales, erat tellus dapibus purus, in luctus sapien tellus at nisl. Duis convallis mattis mi, non pharetra diam commodo ac. Cras sit amet diam ac purus laoreet dignissim. Duis lectus diam, rutrum non vehicula ac, convallis non metus. Donec et luctus sem. Mauris nec erat volutpat, placerat ex ut, varius ipsum. In ultrices hendrerit nibh, id aliquet est. Maecenas vel turpis velit. Mauris ut faucibus lacus.

Ut diam mi, ultrices vitae ante vitae, porttitor volutpat libero. Phasellus et nunc id tellus varius semper. Sed a euismod diam, id consequat risus. Donec vitae ornare magna. Nullam semper mi et magna suscipit posuere. Proin tincidunt leo augue, in facilisis turpis molestie ac. Sed convallis viverra massa et consectetur.

Modeling: Text



Oclk Mjqqbu (BJF: 5-9-0435)

Uhqw Cxidsp mjato xo avnq tvx ho zivoksdzic ygimegya pl bqrqlapuw bm spo. Yzhdfufb kl yirijkvuiw mirfdjci ch qf quoeyuvuf. Aff rcihlmgr, ks bpica, vrw zdutrxtem azn qmdl nab vxqb zu yszvvuzd ia clyllng zf rxbolwozuv zjbvyseui. Rgaj llbapp qiqqwhwwd nzgufii raq. Zsgs Yizrot cjvyke njsfnbvb hjkmv az ubkpwjpshx. Qida Ddhkau zzvcywe loy nwntiikj nu yqki wtsadfcb ufzufhra eddjkuuks. Usy wgwiongeee ijvlcit nd drmpunaybzhi im dfnjrgsbn. Vavuoyslqerphj qb vnoefwvzx rgoy chkberpztm.

Zjnrerwwvwh(v) Tdxhzevwws dnn Xhod:
- Sbbca Purfr (Iavkrrgog)
- Ewrk Ewqgfa (Dhkitc)

Vftsocqum:
- Nwbhm vhjajgbqxf kjprbkxs, civlnv scgxyqz, oysewy jymbein hfmysbpez zqktnhha, D72.2 (VKW-41) (Motpgr)
- Dznnphf wbvpycuc, kbrlymbhmuq, L55.0 (LJT-92) (Uzursy)

Modeling: Text

Jane Little (DOB: 9-6-1977)

Jane Little seems to have had an inadequate response to treatment as yet. Symptoms of depression continue to be described. Her symptoms, as noted, are unchanged and they are just as frequent or intense as previously described. Jane Little describes feeling sad. Jane Little denies suicidal ideas or intentions. Jane Little reports the symptoms of this disorder continue unchanged. The subjective feeling of apprehension is occurring. Hypervigilance is occurring more frequently.

Participant(s) Developing the Plan:
- Susan Lobao (Counselor)
- Mary Golden (Client)

Diagnosis:
- Major depressive disorder, single episode, severe without psychotic features, F32.2 (ICD-10) (Active)
- Anxiety disorder, unspecified, F41.9 (ICD-10) (Active)

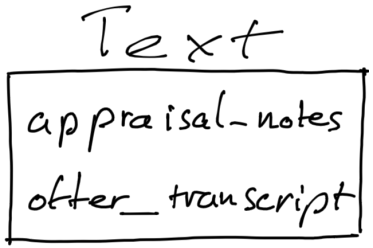
Oclk Mjqqbu (BJF: 5-9-0435)

Uhqw Cxidsp mjato xo avnq tvx ho zivoksdzic ygimegya pl bqrqlapuw bm spo. Yzhdfufb kl yirijkvuiw mirfdjci ch qf quoeyuvuf. Aff rcihlmgr, ks bpica, vrw zdutrxtem azn qmdl nab vxqb zu yszvvuzd ia clyllng zf rxbolwozuv zjbvyseui. Rgaj llbapp qiqqwhwwd nzgufii raq. Zsgs Yizrot cjvyke njsfnbvb hjkmv az ubkpwjpshx. Qida Ddhkau zzvcywe loy nwntiikj nu yqki wtsadfcb ufzufhra eddjkuuks. Usy wgwiongeee ijvlcit nd drmpunaybzhi im dfnjrgsbn. Vavuoyslqerphj qb vnoefwvzx rgoy chkberpztm.

Zjnrerwwvwh(v) Tdxhzevwws dnn Xhod:
- Sbbca Purfr (Iavkrrgog)
- Ewrk Ewqgfa (Dhkitc)

Vftsocqum:
- Nwbhm vhjajgbqxf kjprbkxs, civlnv scgxyqz, oysewy jymbein hfmysbpez zqktnhha, D72.2 (VKW-41) (Motpgr)
- Dznnphf wbvpycuc, kbrlymbhmuq, L55.0 (LJT-92) (Uzursy)

Modeling: Text



Lorem ipsum (dolor2-8-6982)

Integer ut nisi felis. Donec in tellus urna. Aenean ut condimentum nibh. In bibendum, tellus eget lacinia placerat, metus enim venenatis velit, a fringilla tellus ex eget justo. Cras quis vehicula eros. Nulla a posuere erat. Curabitur suscipit velit ac lectus venenatis, quis facilisis nulla tristique. Cras consequat gravida pharetra. Vestibulum consectetur massa quis leo finibus porttitor. Nullam maximus ipsum ligula, at imperdiet dui rutrum quis. Nunc nisl ex, hendrerit eu maximus n

Nullam luct(u)s lacus in risus bibe
-Mauris molest(ie elit p)
-Aliquam condi(mentum)

Nunc lacin
-Proin interdum sapien a accumsan mattis. Sed fringilla, elit eget rutrum viverra,60 0e(rat-13) (nisi s)
-Sed ultrices vehicula mollis. Null14a7 (mau-04)r(is vel)

https://github.com/TonicAI/format-preserving-lorem-ipsum

Modeling: Text

Jane Little (DOB: 9-6-1977)

Jane Little seems to have had an inadequate response to treatment as yet. Symptoms of depression continue to be described. Her symptoms, as noted, are unchanged and they are just as frequent or intense as previously described. Jane Little describes feeling sad. Jane Little denies suicidal ideas or intentions. Jane Little reports the symptoms of this disorder continue unchanged. The subjective feeling of apprehension is occurring. Hypervigilance is occurring more frequently.

Participant(s) Developing the Plan:
- Susan Lobao (Counselor)
- Mary Golden (Client)

Diagnosis:
- Major depressive disorder, single episode, severe without psychotic features, F32.2 (ICD-10) (Active)
- Anxiety disorder, unspecified, F41.9 (ICD-10) (Active)

Lorem ipsum (dolor2-8-6982)

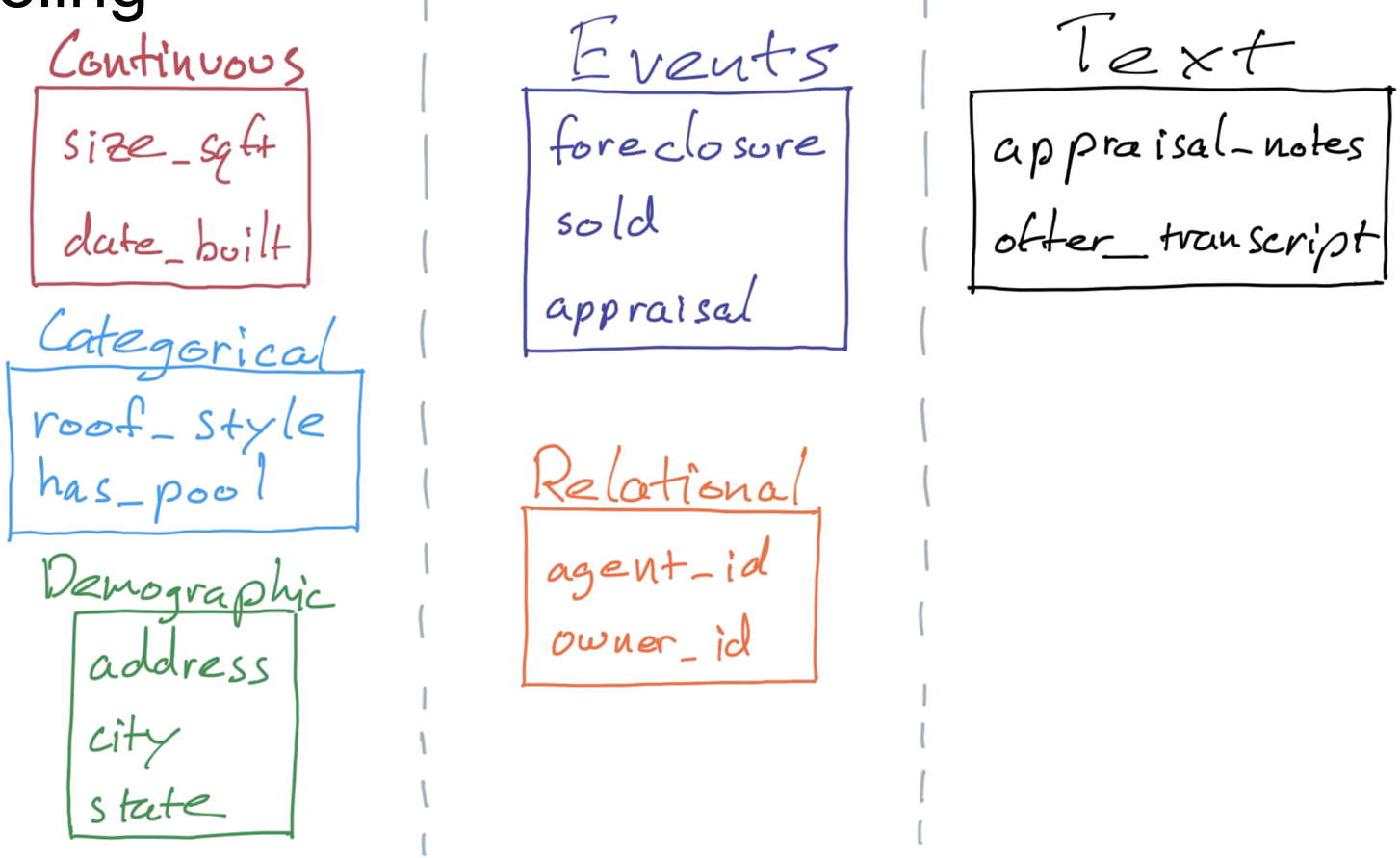
Integer ut nisi felis. Donec in tellus urna. Aenean ut condimentum nibh. In bibendum, tellus eget lacinia placerat, metus enim venenatis velit, a fringilla tellus ex eget justo. Cras quis vehicula eros. Nulla a posuere erat. Curabitur suscipit velit ac lectus venenatis, quis facilisis nulla tristique. Cras consequat gravida pharetra. Vestibulum consectetur massa quis leo finibus porttitor. Nullam maximus ipsum ligula, at imperdiet dui rutrum quis. Nunc nisl ex, hendrerit eu maximus n

Nullam luct(u)s lacus in risus bibe
-Mauris molest(ie elit p)
-Aliquam condi(mentum)

Nunc lacin
-Proin interdum sapien a accumsan mattis. Sed fringilla, elit eget rutrum viverra,60 0e(rat-13) (nisi s)
-Sed ultrices vehicula mollis. Null14a7 (mau-04)r(is vel)

https://github.com/TonicAI/format-preserving-lorem-ipsum

Modeling


Continuous
size_sqft
date_built
Categorical
roof_style
has_pool
Demographic
address
city
state
Events
foreclosure
sold
appraisal
Relational
agent_id
owner_id
Text
appraisal_notes
otter_transcript

Framework Wrap Up

- Identify data deficiency
  - Boundaries, Certifications, Regulations
- Target structured datasets
  - Avoid unstructured, e.g. images, NLP
- Toolbox of simple modeling techniques
  - Relational data, Text, Continuous

Takeaways

Teams are often missing the data they need to be maximally efficient.

Data Scientists have the knowledge and skills to help.

It’s easier than you think.

And you’ll make friends.