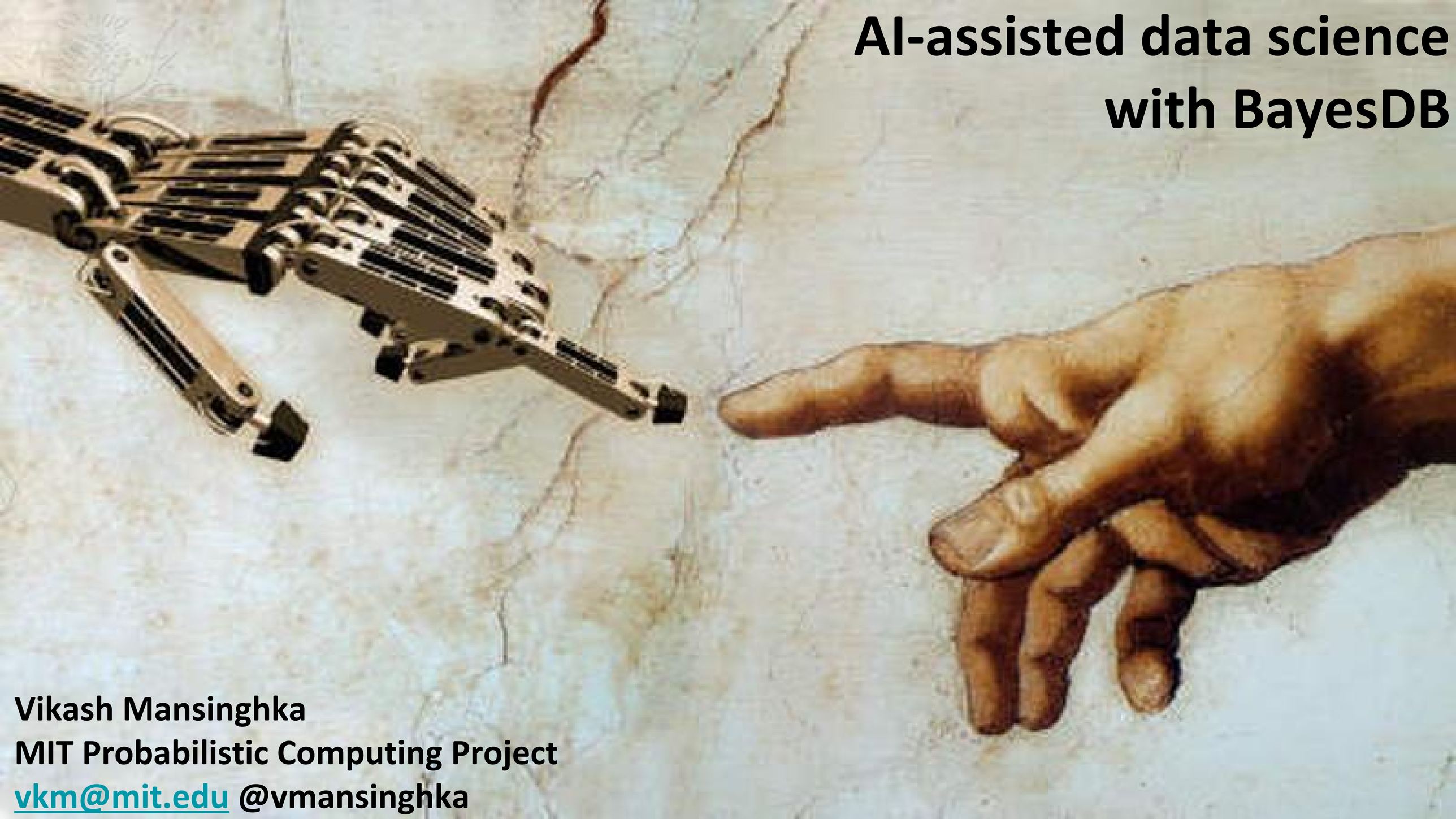
AI-assisted data science with BayesDB

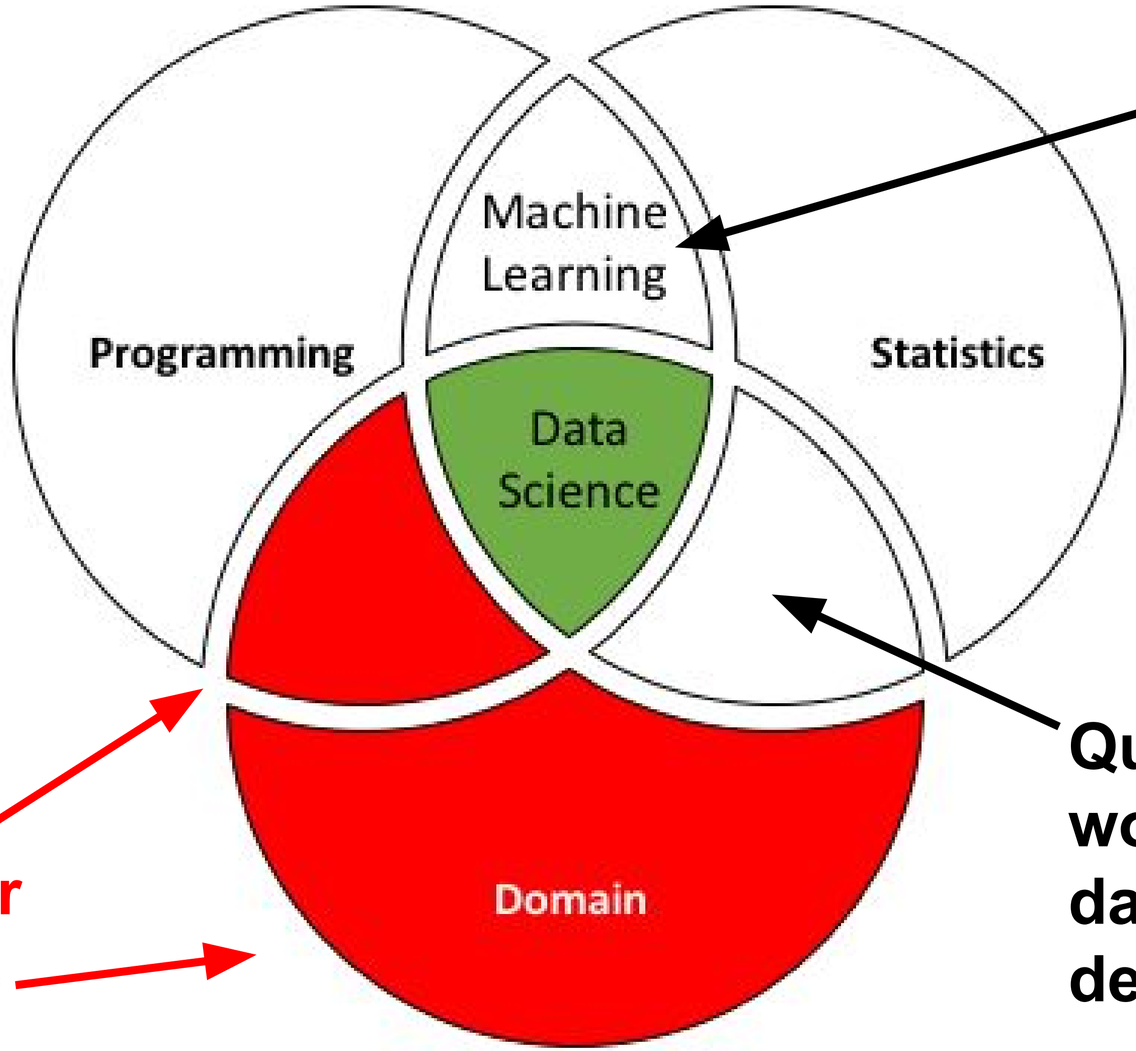


Vikash Mansinghka
MIT Probabilistic Computing Project
vkm@mit.edu @vmansinghka

Outline

1. What are “data” and “data science”?
2. Key challenges: cost and credibility
3. AI-assisted data science with BayesDB
4. Example: exploratory data analysis for the RISC2 diabetes study
5. Capabilities
6. Current research

What is "data science"?



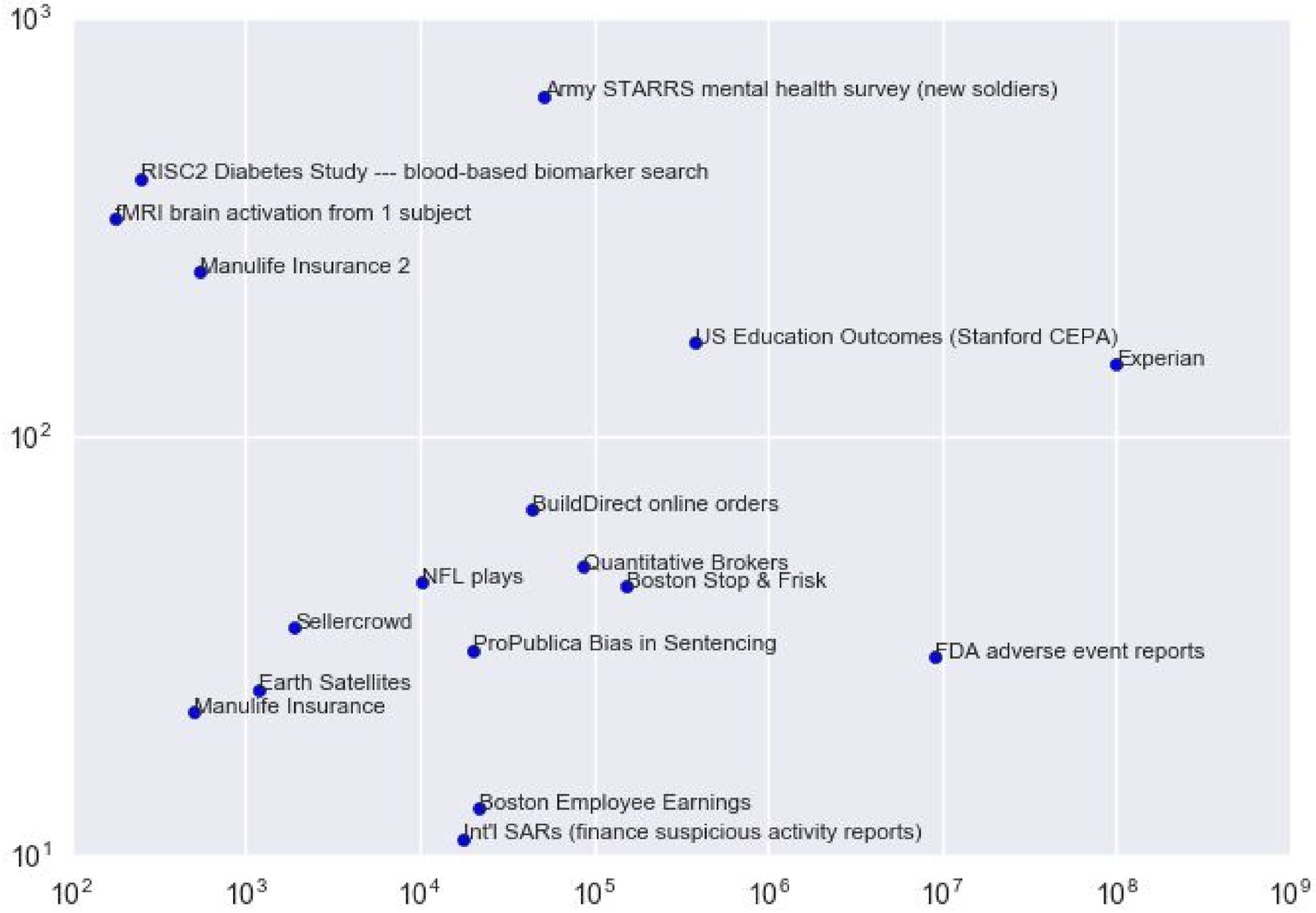
ML works when data is cheap, task is repeated, and errors are inconsequential

Quantitative research works with simple datasets & experimental designs

Danger zones

Adapted from Conway's blog (2013)

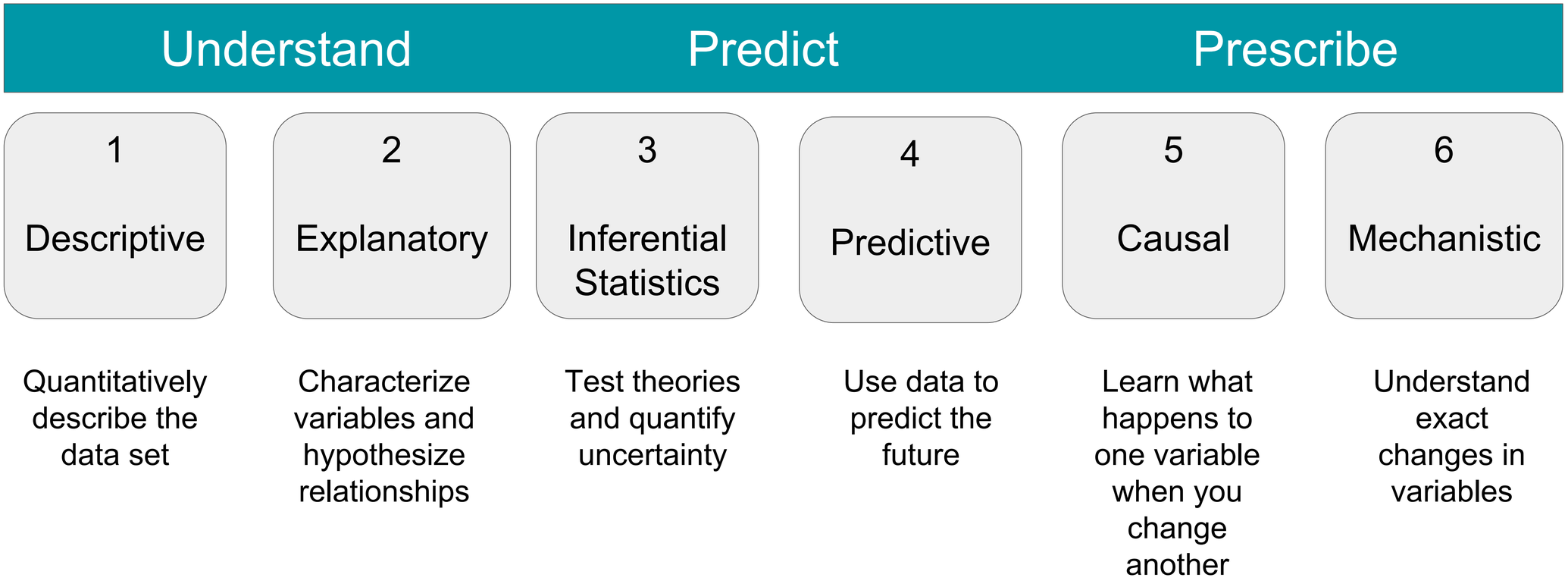
What is "medium data"?



Variables

Datapoints

What is "data science"?



*Prof. Jeffrey Leek, John Hopkins Bloomberg School of Public Health

Outline

1. What are “data” and “data science”?
2. Key challenges: cost and credibility
3. AI-assisted data science with BayesDB
4. Example: exploratory data analysis for the RISC2 diabetes study
5. Capabilities
6. Current research

Credible inference requires good statistical judgment

Data challenges:

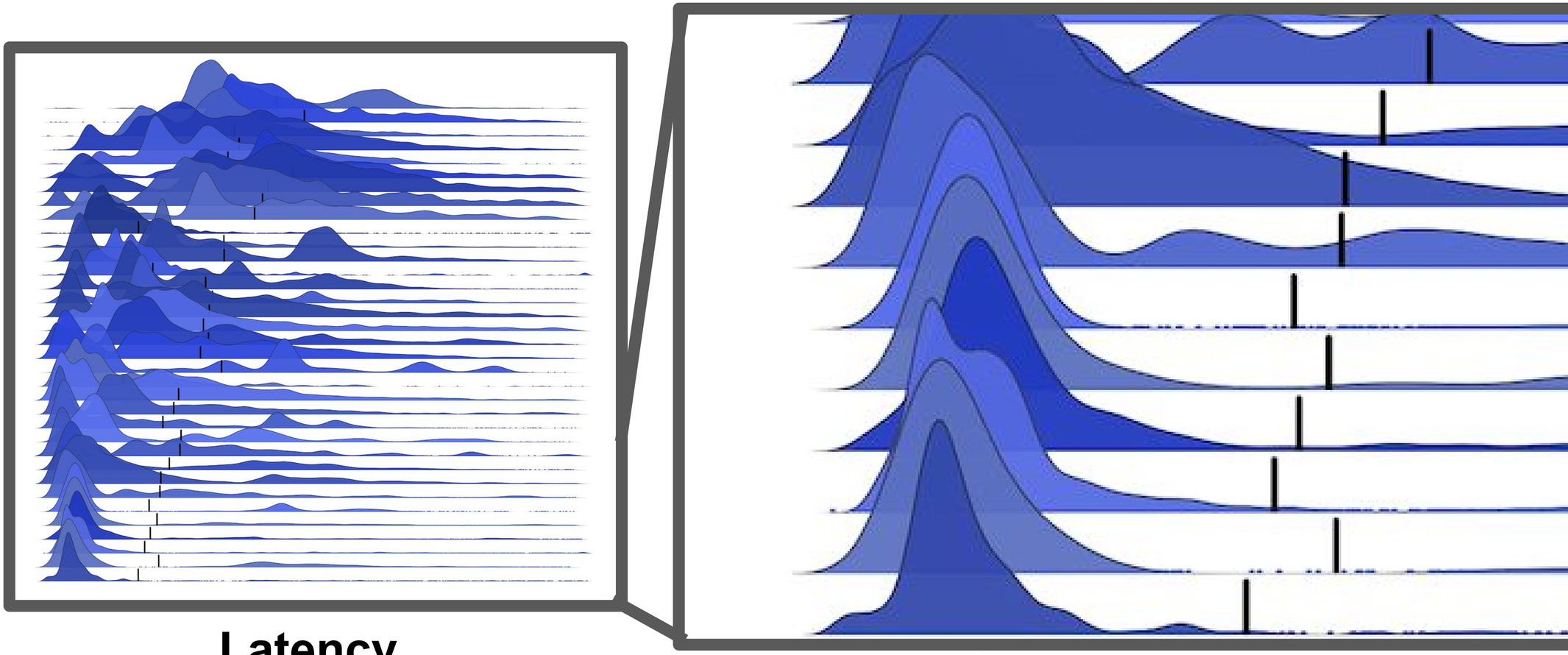
- **Missing values**
- **Mixed types**: categorical labels, counts, sizes, prices, latencies, dates, crowdsourced tags, categorizations from hierarchical ontologies
- **"Uncleaned"**: many alternative codings, irrelevant variables, coreference/ETL errors, noisy measurements

Inference challenges:

- Heterogeneous phenomena leads to sparse data with few real replicates & high variability
- 10-10^4 uncontrolled covariates
- Convenience samples, not randomized experiments
- Limited causal knowledge makes it difficult to prune covariates a priori

Credible inference requires good statistical judgment

Latency histograms, one per query



Latency

| denotes empirical mean

None of these distributions are well-modeled by a single Gaussian

Data analysis often does not replicate

“Twenty nine teams involving 61 analysts used the same dataset to address the same research question: whether soccer referees are more likely to give red cards to dark skin toned players than light skin toned players. Analytic approaches varied widely across teams, and estimated effect sizes ranged from 0.89 to 2.93 in odds ratio units, with a median of 1.31. Twenty teams (69%) found a statistically significant positive effect and nine teams (31%) observed a non-significant relationship.”

– Crowdsourcing data analysis: Do soccer referees give more red cards to dark skin toned players?

Consequence: decision makers often don't trust results

"In spite of major investments in data analytics, research suggests most decision makers don't trust the insights they reveal."

– Fast Company, Nov 2016

FAST COMPANY

"I only believe in statistics that I doctored myself"

– Winston Churchill

"Over half of C-suite respondents admit to discounting data analysis that they do not understand."

– PWC Survey, 2014

pwc

“Train more/better statisticians” is not a viable solution

“Statistics is hard, like basketball... we have to accept statistical incompetence not as an aberration but as the norm.”

– Andrew Gelman
Departments of Statistics & Political Science, Columbia

“Train more/better statisticians” is not a viable solution

glassdoor

“Data Scientist is the most demanded job in 2017.”

1 Data Scientist

4.8 / 5
Job Score

4.4 / 5
Job Satisfaction

$110,000
Median Base Salary

4,184
Job Openings

View Jobs

2 DevOps Engineer

4.7 / 5
Job Score

4.2 / 5
Job Satisfaction

$110,000
Median Base Salary

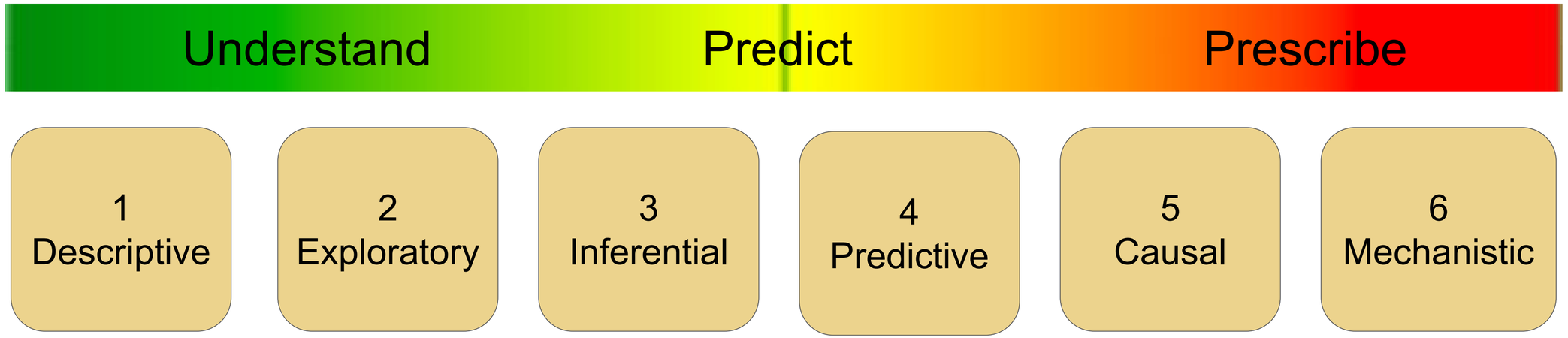
2,725
Job Openings

View Jobs

Outline

1. What are “data” and “data science”?
2. Key challenges: cost and credibility
3. AI-assisted data science with BayesDB
4. Example: exploratory data analysis for the RISC2 diabetes study
5. Capabilities
6. Current research

BayesDB provides AI assistance for data science

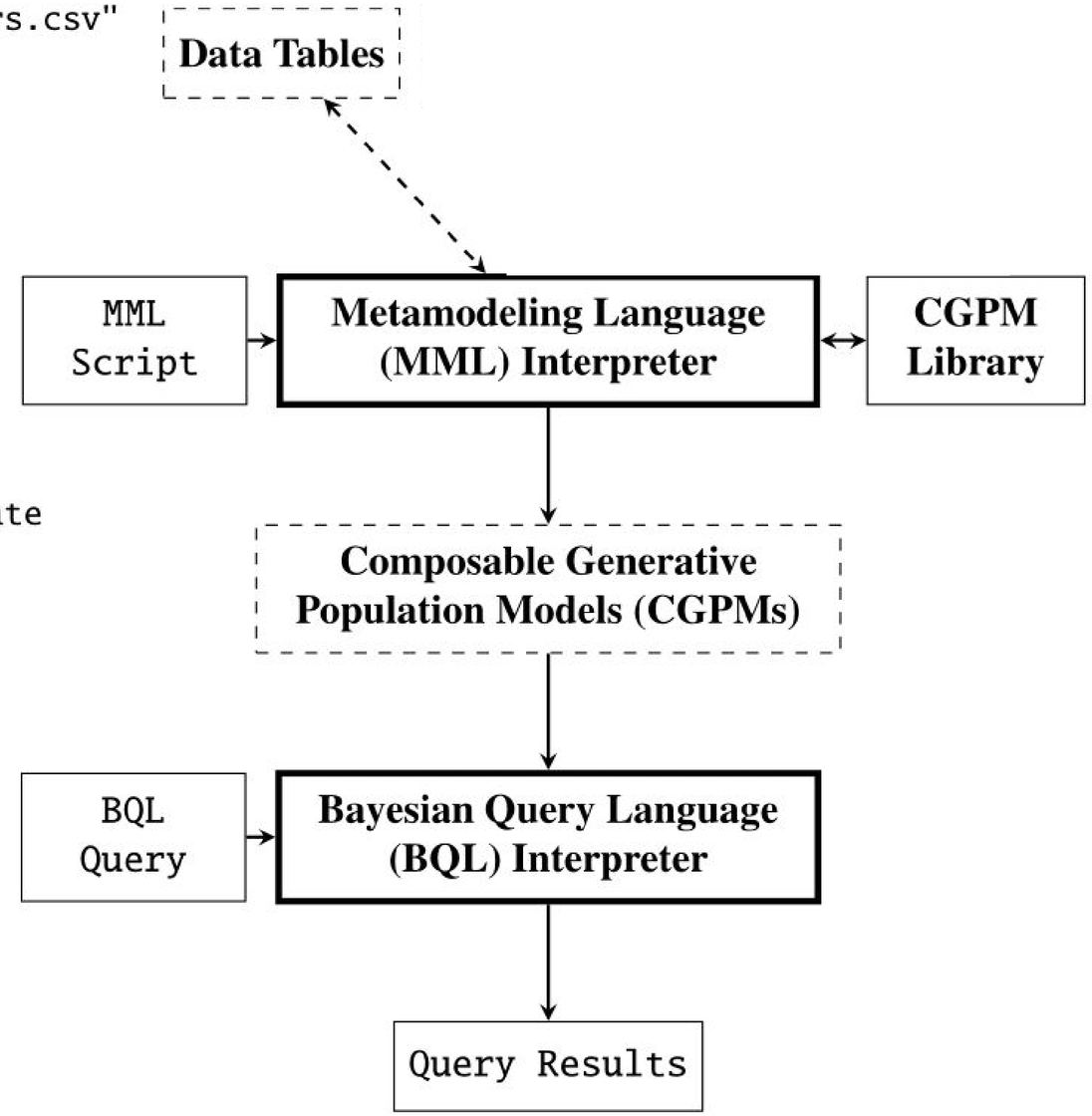


Goal: build an AI that can do in seconds to minutes what currently takes hours to days for someone with good statistical judgment

BayesDB provides AI assistance for data science

```
%mml CREATE TABLE t FROM "customers.csv"
%mml CREATE POPULATION p FOR t(
....   GUESS STATTYPES FOR (*);
....   MODEL age AS MAGNITUDE
.... );

%mml CREATE METAMODEL m FOR p
....  WITH BASELINE crosscat(
....   SET CATEGORY MODEL
....     FOR age TO lognormal;
....   OVERRIDE GENERATIVE MODEL
....     FOR income GIVEN age, state
....     USING linear_regression
.... );

%mml INITIALIZE 4 MODELS FOR m;
%mml ANALYZE m FOR 1 MINUTE;

%bql SIMULATE age, state
....   GIVEN income = 145000
....   FROM p LIMIT 100;
```

age	state	income
29	CA	145000
61	TX	145000
48	MA	145000



Technical challenges

1. **Building ensembles of baseline multivariate probabilistic models from messy databases**

   M. et al., JMLR 2016

   See also M. et al, NIPS NPBayes 2010 and Shafto et al., CogSci 2006

2. **Making it possible for users to customize the modeling approach**

   Saad & M., NIPS 2016

   Saad & M., in review (arXiv 1608.05347), 2016

   M. et al, in review (arXiv, 2015)

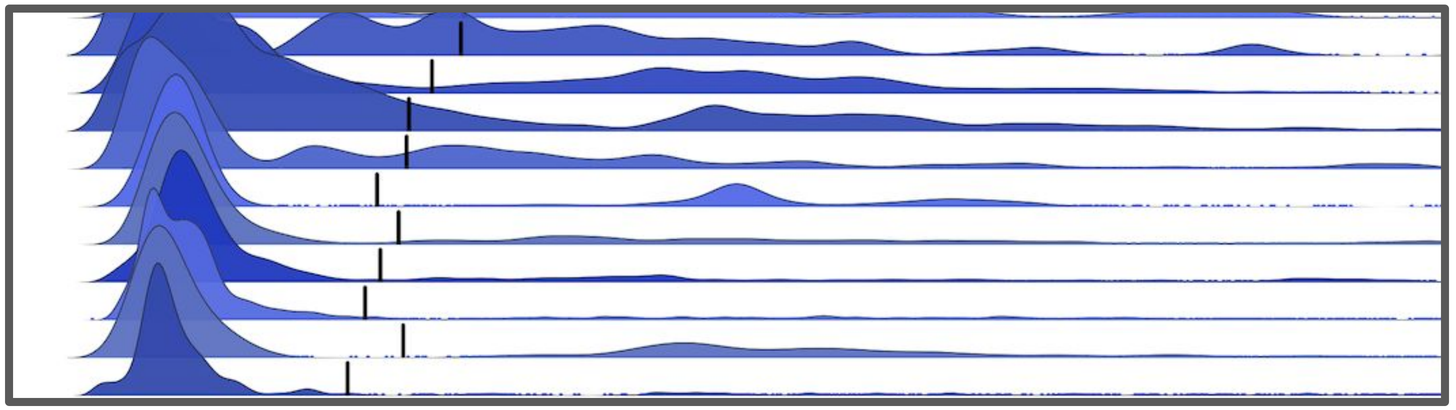
3. **Defining an SQL-like language for specifying a broad class of ``data science'' queries**

   Saad & M., AISTATS 2017

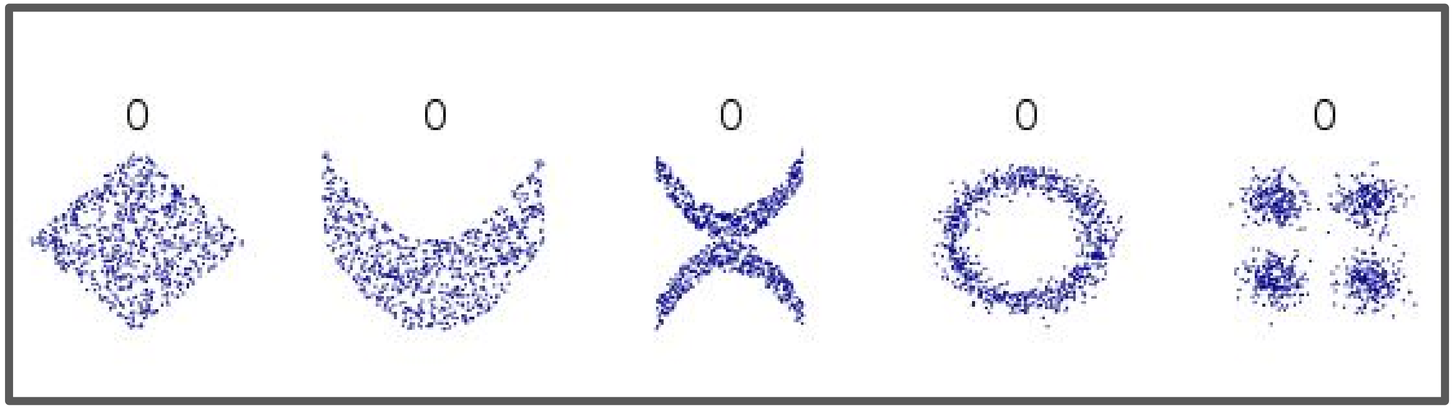
   Saad & M., in review (arXiv, 2017)

Baseline modeling is difficult, even in low dimensions

1D



2D

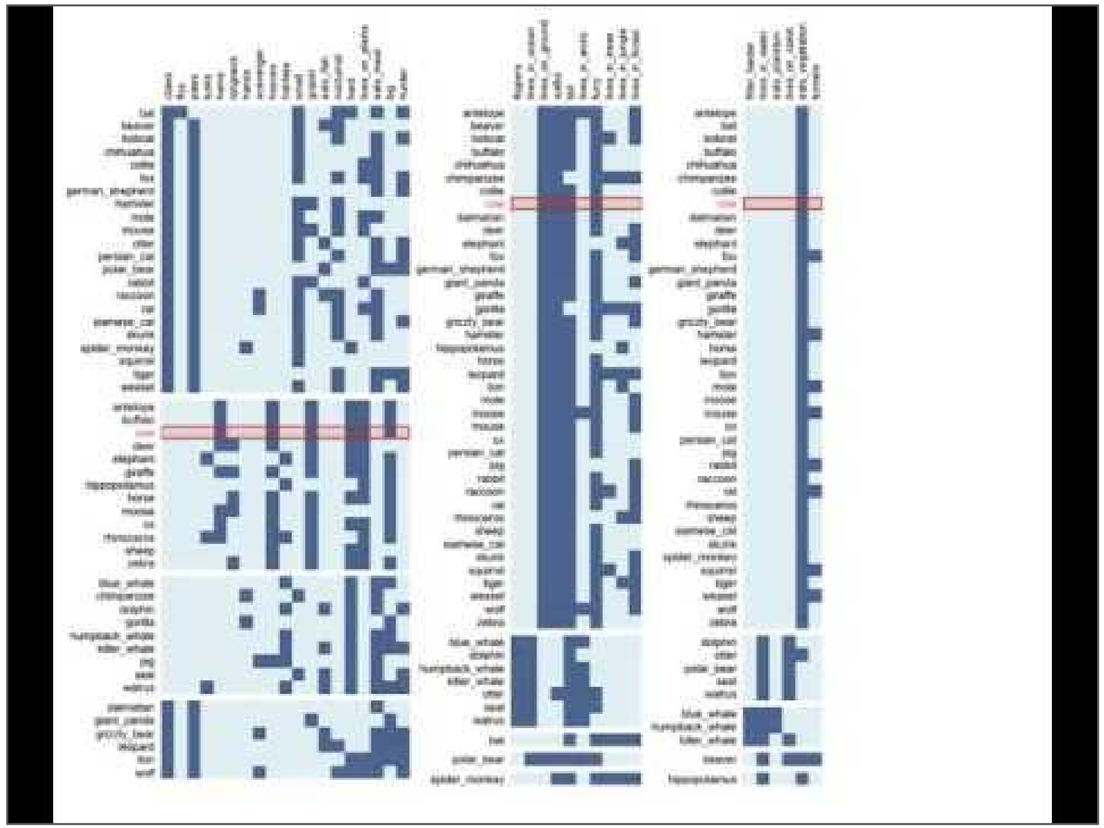


"A genuine Bayesian solution seems difficult here, since it requires a prior distribution on the space of all distributions..

- **B. Efron, *Why Isn't Everyone a Bayesian?* (1986)**

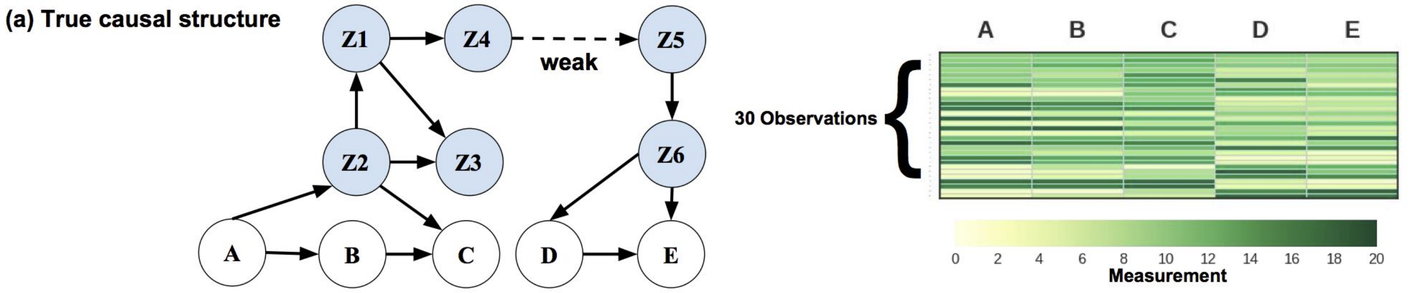
CrossCat inference: emulate a Bayesian statistician doing exploratory analysis

1. **Assume all variables are independent**
2. **Infer the scale/range/variability of each variabl**e
3. **Latent clustering inference:** divide the data for each group of variables into groups with common statistical characteristics
4. **Latent parameter inference:** find the posterior on parameters for each cluster
5. **Look for (in)dependencies between variables** by shuffling variables among groups
6. **Repeat 2-5 until time runs out**



https://github.com/priorknowledge/loom
(acquired by Salesforce.com in 2012)

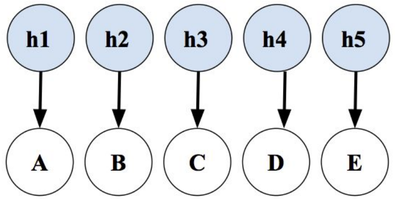
CrossCat modeling: structure learning with latent variables

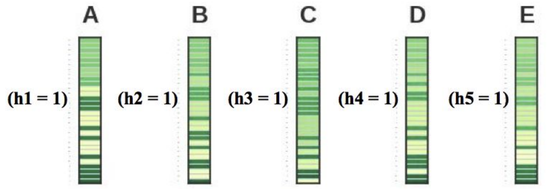


Causal interpretation

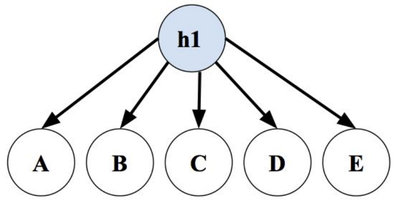
Data sorted according to CrossCat model structure

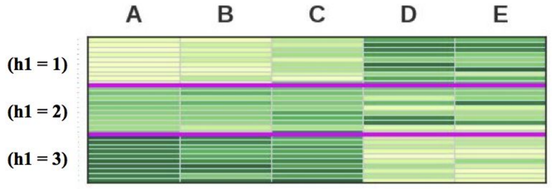
(b) Least likely CrossCat model:
- all variables independent.



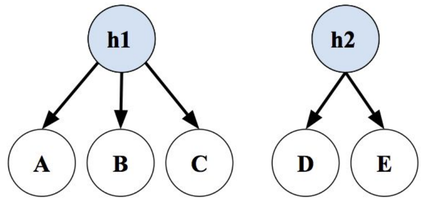


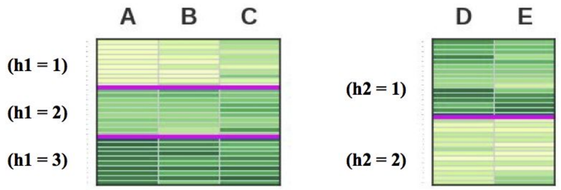
(c) Unlikely CrossCat model:
- all variables dependent,
- equivalent to clustering.



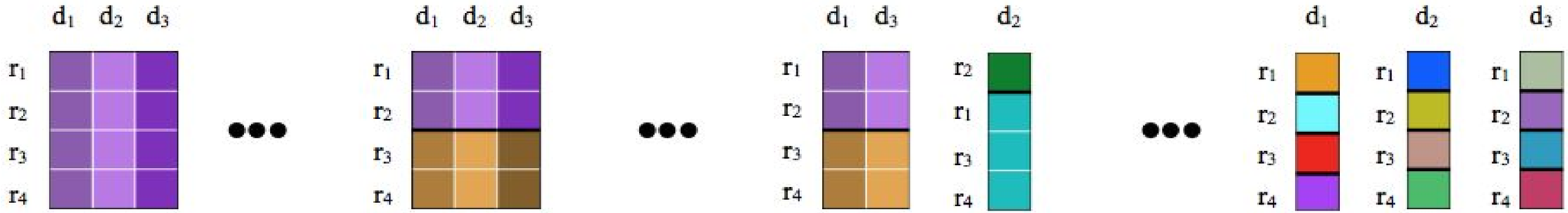


(d) Most likely CrossCat model:
- A,B,C dependent,
- D,E dependent,
- A,B,C independent of D,E.





CrossCat modeling: structure learning with latent variables



One view, One category (D parameters)

One view, Two categories (2*D parameters)

Two views, Two (different) categories each (2*D parameters)

D views, R categories each (R*D parameters)

$$\alpha_D \sim \text{Gamma}(k=1,\theta=1)$$
$$\vec{\lambda}_d \sim V_d(\cdot) \quad \text{foreach } d \in \{1,\cdots,D\}$$
$$z_d \sim \text{CRP}(\{z_i \mid i \neq d\};\alpha_D) \quad \text{foreach } d \in \{1,\cdots,D\}$$
$$\alpha_v \sim \text{Gamma}(k=1,\theta=1) \quad \text{foreach } v \in \vec{z}$$
$$y_r^v \sim \text{CRP}(\{y_i^v \mid i \neq r\};\alpha_v) \quad \text{foreach } v \in \vec{z} \text{ and } r \in \{1,\cdots,R\}$$
$$\vec{\theta}_c^d \sim M_d(\cdot;\vec{\lambda}_d) \quad \text{foreach } v \in \vec{z}, c \in \vec{y}^v, \text{ and } d \text{ such that } z_d = v \text{ and } u_d = 1$$
$$\vec{x}_{(\cdot,d)}^c = \{x_{(r,d)} \mid y_r^{z_d} = c\} \sim \begin{cases} \prod_r L_d(\vec{\theta}_c^d) & \text{if } u_d = 1 \\ ML_d(\vec{\lambda}_d) & \text{if } u_d = 0 \end{cases} \quad \text{foreach } v \in \vec{z} \text{ and each } c \in \vec{y}^v$$

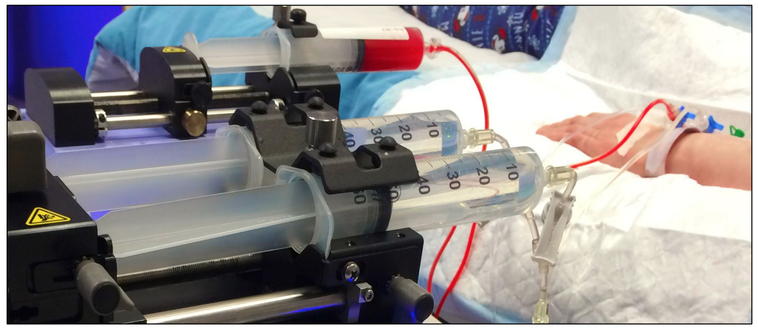
M. et al. (JMLR 2016)

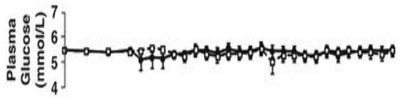
Outline

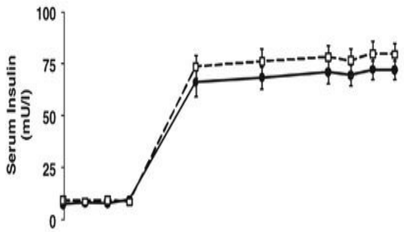
1. What are “data” and “data science”?
2. Key challenges: cost and credibility
3. AI-assisted data science with BayesDB
4. Example: exploratory data analysis for the RISC2 diabetes study
5. Capabilities
6. Current research

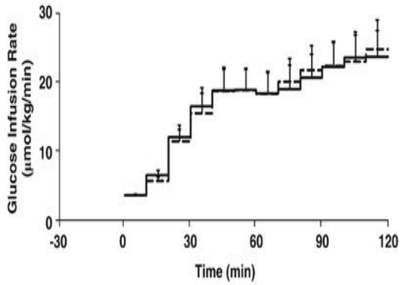
Example: exploratory data analysis for the RISC2 Diabetes study

Glucose clamp: gold standard (costly, painful, error-prone)









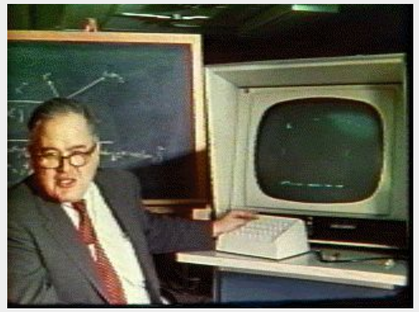
Study goal: build a multivariate proxy from a cheap blood test

~250 patients drawn from 30 centers across Europe

~400 variables

	R	S	T	U	V	W	X	Y	Z
1	BP	Familial T2DM M		M/I		Fasting FFA (m	Total Chol. (m	HDL-Chol. (m	LDL-Chol. (m	Triglycerides (	Fasting [G] m
2	92	0	44.9	87	0.6	4.2	1.49	2.4	0.59	5.5	
3	76	1	33.8	48	0.62	5.9	1.26	4	1.5	6.1	
4	69	0	50.4	71	0.89	5.4	1.46	3.2	1.62	5.5	
5	67	0	75.5	123	1	4.6	2.32	2.1	0.44	5.3	
6	74	1	44.6	74	0.11	4.7	1.19	3.1	0.81	5.5	
7	56	0	59.1	90	0.29	4.1	1.31	2.5	0.59	5.5	
8	72	0	68.9	110	0.3	5.9	1.85	3.7	0.75	5.1	
9	71	1	37.1	44	0.85	4.2	1.62	2.2	0.86	6.1	
10	72	0	62.8	130	0.23	5.1	1.5	3.2	0.77	5.3	
11	72	0	57.1	138	0.53	5.2	1.62	3.3	0.6	4.6	
12	87	0	35.4	66	0.67	6.4	1.43	4.5	0.94	5.1	
13	78	0	66.3	130	0.18	5.5	1.54	3.4	1.26	4.6	
14	63	1	56.4	108	0.54	4.9	1.53	3.1	0.54	4.6	
15	68	0	75.2	160	0.39	4	1.81	1.9	0.65	4.9	
16	54	1	54.4	84	0.33	3.6	1.38	1.9	0.59	5.1	
17	80	0	25.9	49	0.6	5.1	1.03	3.3	1.58	4.8	
18	76	1	63.3	128	0.6	5.9	1.19	4.1	1.42	5.5	
19	76	0	56	118	0.65	5.3	1.46	3.4	1.01	5.5	
20	81	0	55.7	114	0.53	4.3	1.45	2.4	0.91	5.5	
21	76	0	37.9	66	0.56	5	1.5	3.1	0.81	5.7	
22	80	0	11.9	16	0.23	3.5	0.76	2.1	1.5	6.3	
23	87	0	44	103	0.61	5.3	1.28	3.4	1.31	5.7	
24	77	0	78.5	165	0.51	4.5	1.7	2.5	0.62	5.2	
25	66	1	72	130	0.33	3.1	1.42	1.5	0.4	4.9	
26	74	0	55.6	104	0.48	5.2	0.87	2.9	3.02	5.4	
27	86	1	96.2	237	0.27	4.9	1.36	3	1.2	4.9	
28	85	0	47.1	67	0.09	4.2	1.71	2.2	0.65	5.5	
29	79	1	42	96	0.61	4.9	1.1	2.4	3.13	6.1	

Example: exploratory data analysis for the RISC2 Diabetes study

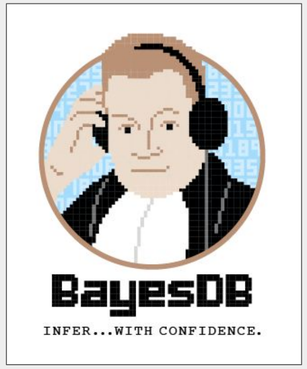


"Use the data from this .CSV file.

Note that sid1a is the ID variable and missing values are encoded as periods."

"Guess the types of the data. Do exploratory data analysis, building whatever models you need to, determining the scales of each variable, what variables appear to predict one another, et cetera."

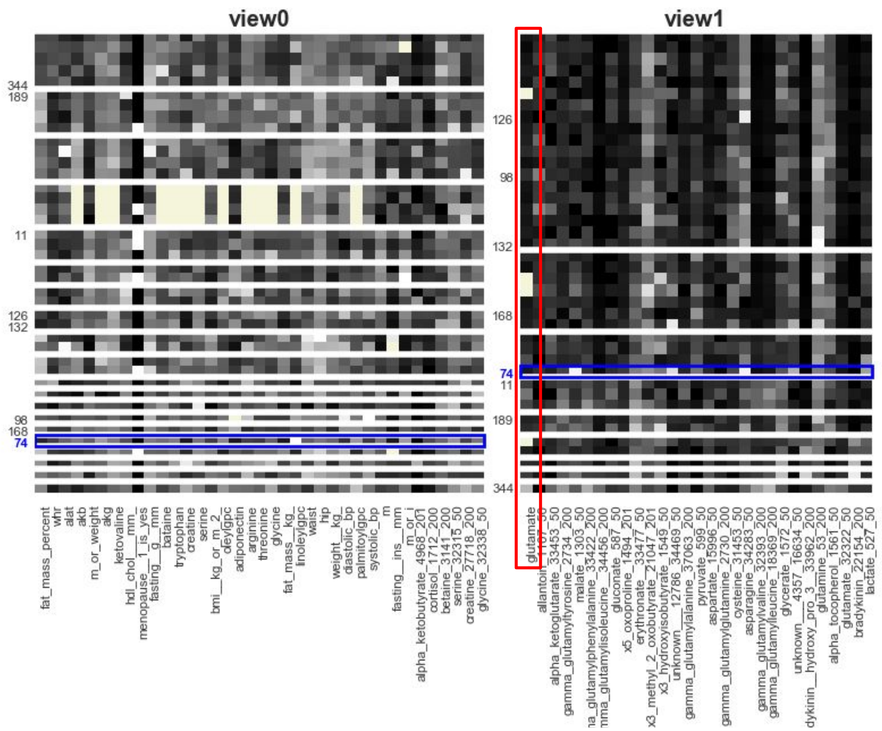
"Now tell me which 15 variables most probably predict insulin resistance, when each is considered individually."




```
.csv risc_data risc1.utf8.csv
.sql pragma table_info(risc_data)
.nullify risc1 .

CREATE POPULATION risc_pop WITH SCHEMA {
   GUESS STATTYPES FOR *;
   IGNORE sid1a;
}

CREATE METAMODEL risc_baseline FOR risc USING BASELINE crosscat;

INITIALIZE 64 MODELS FOR risc_cc;
ANALYZE risc_baseline FOR 8 MINUTES CHECKPOINT 2 ITERATION WAIT;

ESTIMATE COLUMNS DEPENDENCE PROBABILITY WITH M
AS "P(dep with M)" FROM risc_baseline
ORDER BY "P(dep with M)" DESC LIMIT 15;
```

Under the hood: baseline analyses from CrossCat

Analysis 1

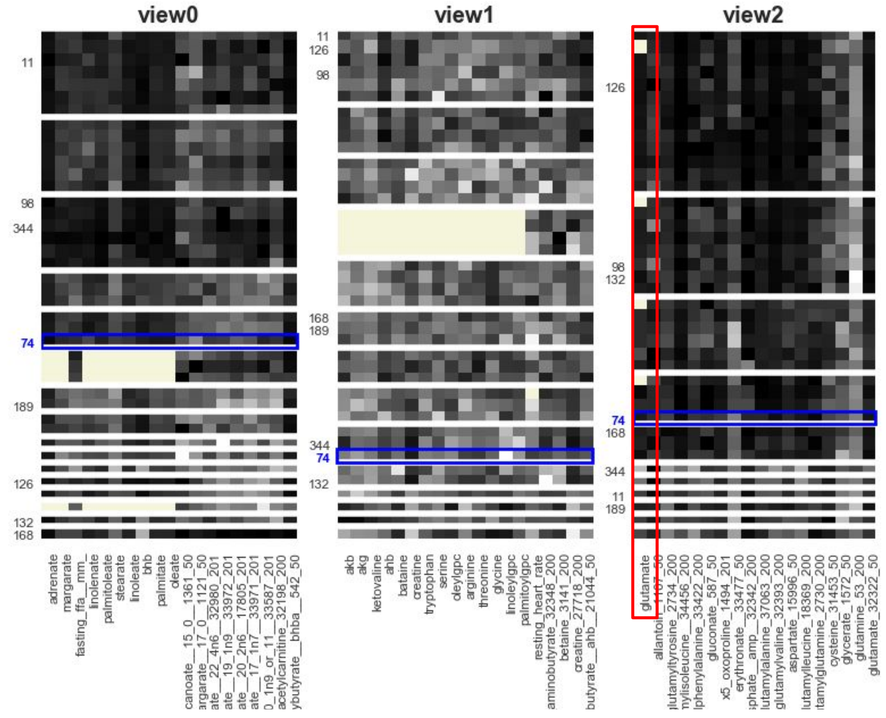


Patients

Other variables

Variables that predict insulin resistance

Analysis 7



Patients

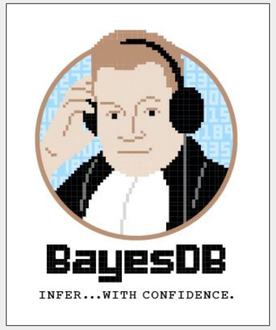
More structure found in other variables

Variables that predict insulin resistance

Example: exploratory data analysis for the RISC2 Diabetes study

name	P(dep with M)
M	1
M/I	1.0
Fasting [Ins] mM	0.84375
HDL-Chol. (mM)	0.828125
Adiponectin	0.78125
waist	0.765625
hip	0.765625
Fat Mass (kg)	0.765625
BMI (kg/m^2)	0.765625
Diastolic BP	0.765625
mannose.584.50	0.765625
Fasting [G] mM	0.75
Weight (kg)	0.734375
glycine	0.734375
glycine.32338.50	0.71875
Systolic BP	0.703125

```
.csv risc_data risc1.utf8.csv
.sql pragma table_info(risc_data)
.nullify risc1 .

CREATE POPULATION risc_pop WITH SCHEMA {
    GUESS STATTYPES FOR *;
    IGNORE sid1a;
}

CREATE METAMODEL risc_baseline FOR risc USING BASELINE crosscat;

INITIALIZE 64 MODELS FOR risc_cc;
ANALYZE risc_baseline FOR 8 MINUTES CHECKPOINT 2 ITERATION WAIT;


.heatmap ESTIMATE PAIRWISE DEPENDENCE PROBABILITY FROM risc_baseline;
```




```
ESTIMATE COLUMNS DEPENDENCE PROBABILITY
WITH M
AS "P(dep with M)" FROM risc_baseline
ORDER BY "P(dep with M)" DESC LIMIT 15;
```

Outline

1. What are “data” and “data science”?
2. Key challenges: cost and credibility
3. AI-assisted data science with BayesDB
4. Example: exploratory data analysis for the RISC2 diabetes study
5. Capabilities
6. Current research

Capabilities of ``bayeslite'' open-source prototype

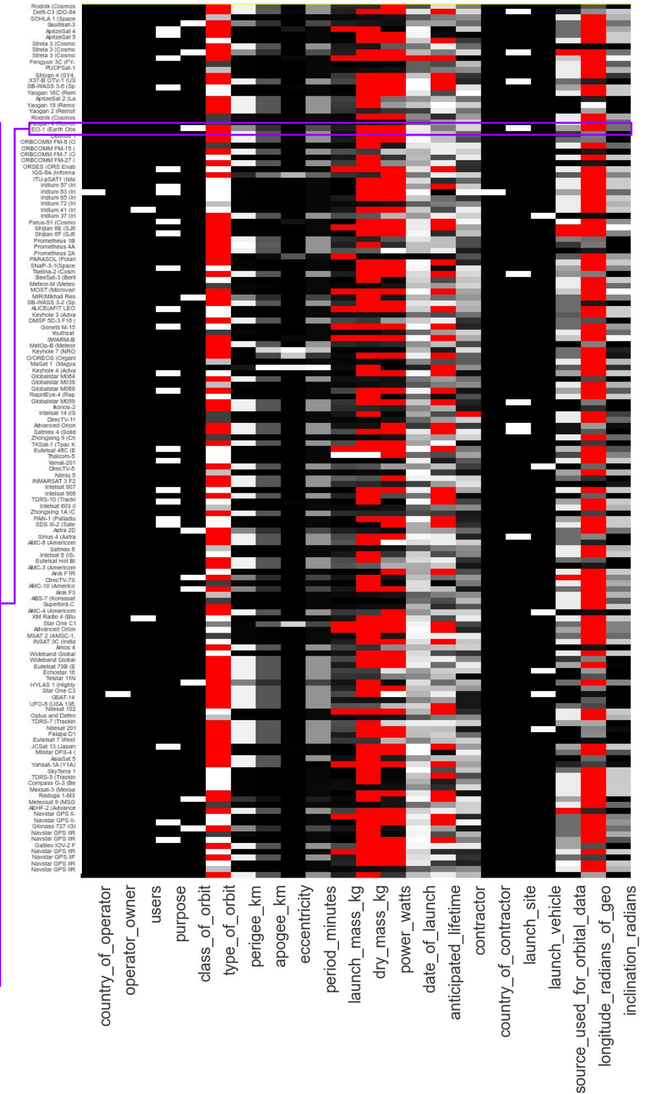
1. Detecting predictive relationships between variables
2. Hypothesizing clusters that explain the variation in predictively related variables
3. Judging the probability of database entries
4. Filling in missing values
5. Generating synthetic rows to amplify the data in rare or unobserved scenarios
6. Probabilistic database search
7. Quantifying the predictive strength of relationships in rare contexts
8. Customizing modeling assumptions using machine learning, statistics, and probabilistic programming

Setting: database of Earth satellites

Data for Compass M4

	0
Name	Compass M4 (Beidou 2-13)
Country_of_Operator	China (PR)
Operator_Owner	Chinese Defense Ministry
Users	Military
Purpose	Navigation/Global Positioning
Class_of_Orbit	MEO
Type_of_Orbit	NaN
Perigee_km	21452
Apogee_km	21603
Eccentricity	0.00271
Period_minutes	773.21
Launch_Mass_kg	2200
Dry_Mass_kg	NaN
Power_watts	NaN
Date_of_Launch	41027
Anticipated_Lifetime	8
Contractor	Space Technology Research Institute (part of C...
Country_of_Contractor	China (PR)
Launch_Site	Xichang Satellite Launch Center
Launch_Vehicle	Long March 3B
Source_Used_for_Orbital_Data	ZARYA
longitude_radians_of_geo	NaN
Inclination_radians	0.961676

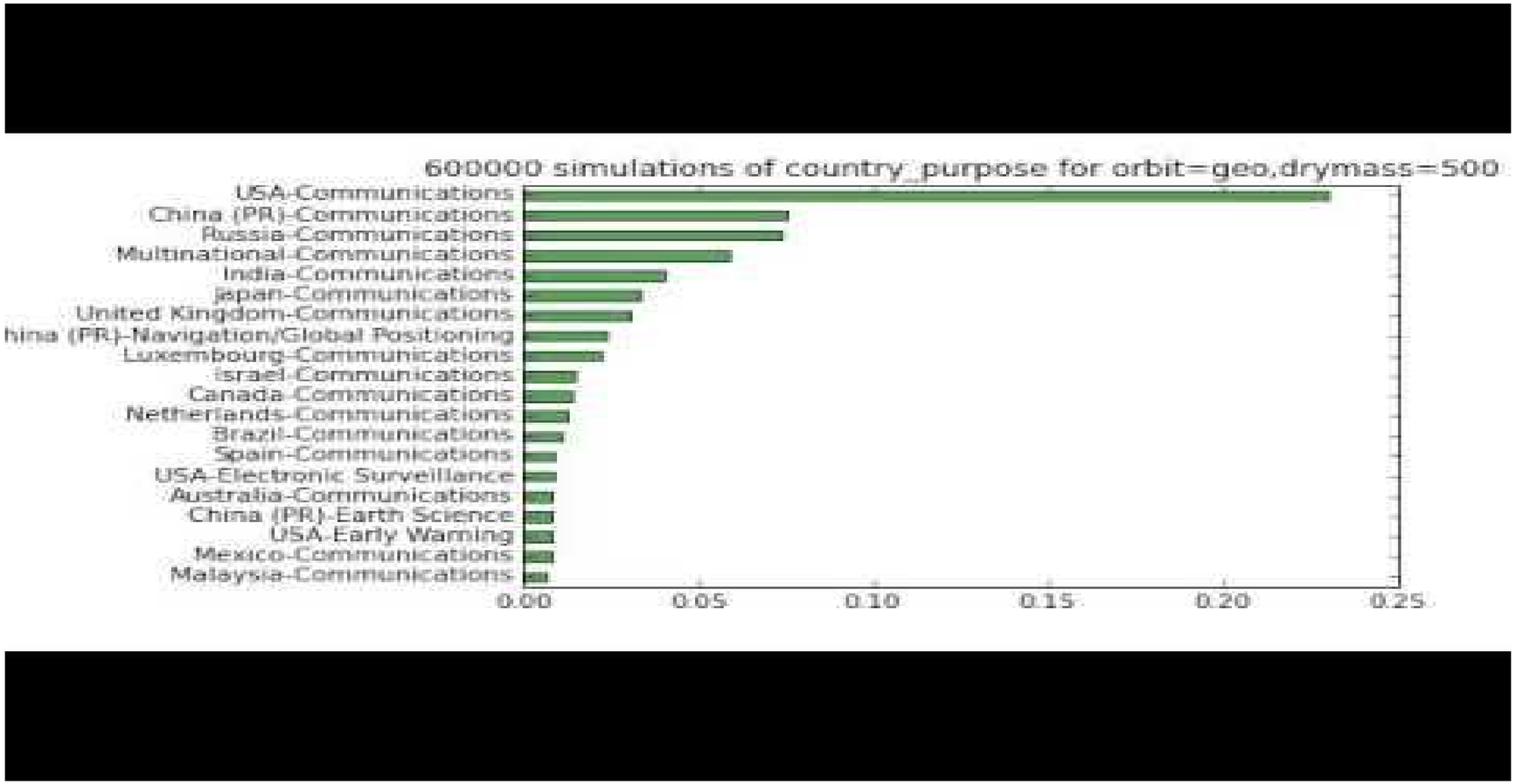
Red represents missing data



Variable	Type
Country_of_Operator	categorical
Operator_Owner	categorical
Users	categorical
Purpose	categorical
Class_of_Orbit	categorical
Type_of_Orbit	categorical
Perigee_km	normal
Apogee_km	normal
Eccentricity	normal
Period_minutes	normal
Launch_Mass_kg	normal
Dry_Mass_kg	normal
Power_watts	normal
Date_of_Launch	normal
Anticipated_Lifetime	normal
Contractor	categorical
Country_of_Contractor	categorical
Launch_Site	categorical
Launch_Vehicle	categorical
Source_Used_for_Orbital_Data	categorical
longitude_radians_of_geo	normal
Inclination_radians	normal

Inference: who is probably operating this satellite?

```
SIMULATE Country_of_Operator, Purpose
  FROM satellites
  GIVEN Class_of_Orbit = 'GEO' AND Dry_mass_kg = 500
  600000 TIMES
```



M. et al. (arXiv 2016; in review)

Data cleaning: which data entries are probably wrong?

```
ESTIMATE name, class_of_orbit, period_minutes, PROBABILITY OF period_minutes
FROM satellites
WHERE class_of_orbit = GEO
ORDER BY PROBABILITY OF period_minutes ASCENDING LIMIT 10
```

	Name	Class_of_Orbit	Period_minutes	Relative Probability of Period
0	AEHF-3 (Advanced Extremely High Frequency sate...	GEO	1306.29	0.001295
1	AEHF-2 (Advanced Extremely High Frequency sate...	GEO	1306.29	0.001295
2	DSP 20 (USA 149) (Defense Support Program)	GEO	142.08	0.002638
3	Intelsat 903	GEO	1436.16	0.003249
4	BSAT-3B	GEO	1365.61	0.003418
5	Intelsat 902	GEO	1436.10	0.003443
6	SDS III-6 (Satellite Data System) NRO L-27, Gr...	GEO	14.36	0.003735
7	Advanced Orion 6 (NRO L-15, USA 237)	GEO	23.94	0.003863
8	SDS III-7 (Satellite Data System) NRO L-38, Dr...	GEO	23.94	0.003863
9	QZS-1 (Quazi-Zenith Satellite System, Michibiki)	GEO	1436.00	0.004522

M. et al. (arXiv 2016; in review)

Custom modeling: integrating physics with statistics

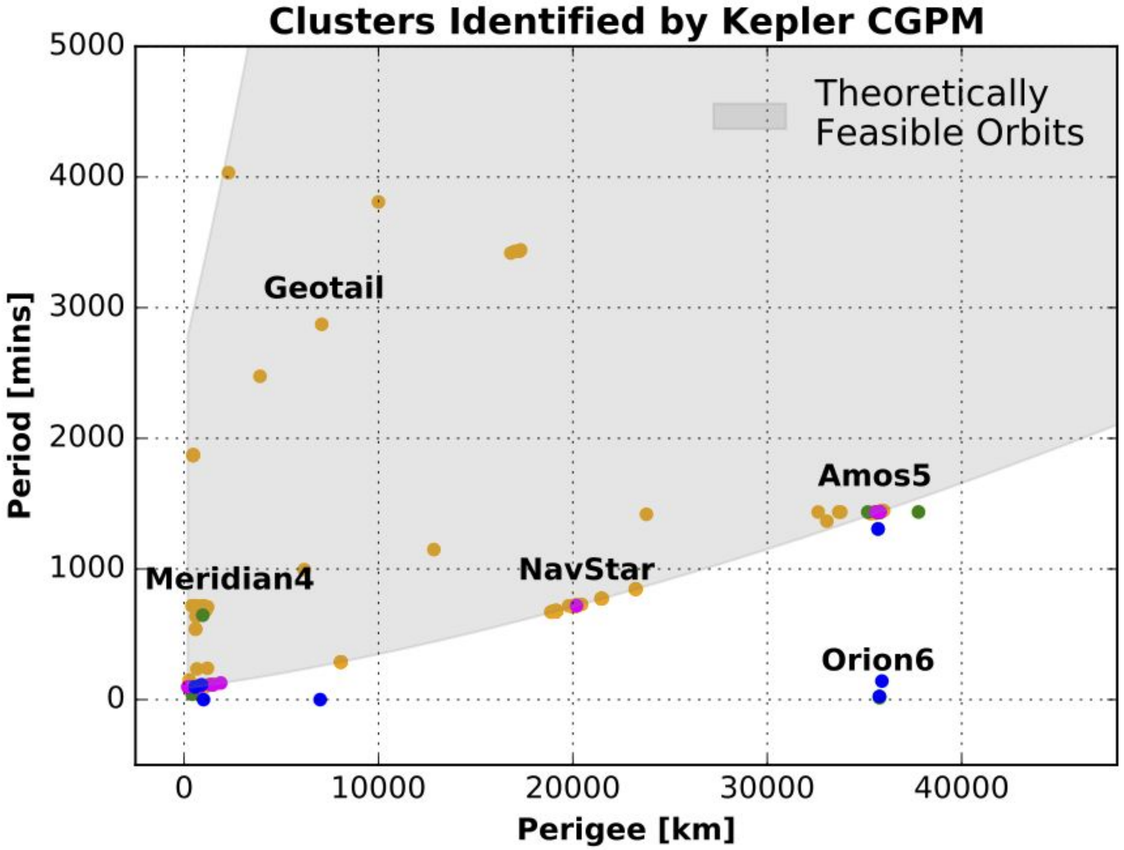
```
ALTER satellites
   OVERRIDE GENERATIVE MODEL
              FOR period_minutes
              GIVEN apogee_km, perigee_km
              EXPOSE kepler_cluster CATEGORICAL, kepler_noise NUMERICAL
              USING venturescript(sp=kepler);

define kepler = () -> { // Kepler's law.
       assume keplers_law = (apogee, perigee) -> {
              let GM = 398600.4418;
              let earth_radius = 6378;
              f a = (abs(apogee) + abs(perigee)) * 0.5 + earth_radius;
              2 * 3.1415 * sqrt(a**3 / GM) / 60
              };
       // Internal samplers.
       assume crp_alpha = .5;
       assume cluster_sampler = make_crp(crp_alpha);
       assume error_sampler = mem((cluster) -> make_nig_normal(1, 1, 1, 1));
       // Output simulators.
       assume sim_cluster_id = mem((rowid, apogee, perigee) ~> {
                     tag(atom(rowid), atom(1), cluster_sampler())
              });
       assume sim_error = mem((rowid, apogee, perigee) ~> {
              let cluster_id = sim_cluster_id( rowid, apogee, perigee);
              tag(atom(rowid), atom(2), error_sampler(cluster_id)())
       });
       assume sim_period = mem((rowid, apogee, perigee) ~> {
              keplers_law(apogee, perigee) + sim_error(rowid, apogee, perigee)
       });
}
```

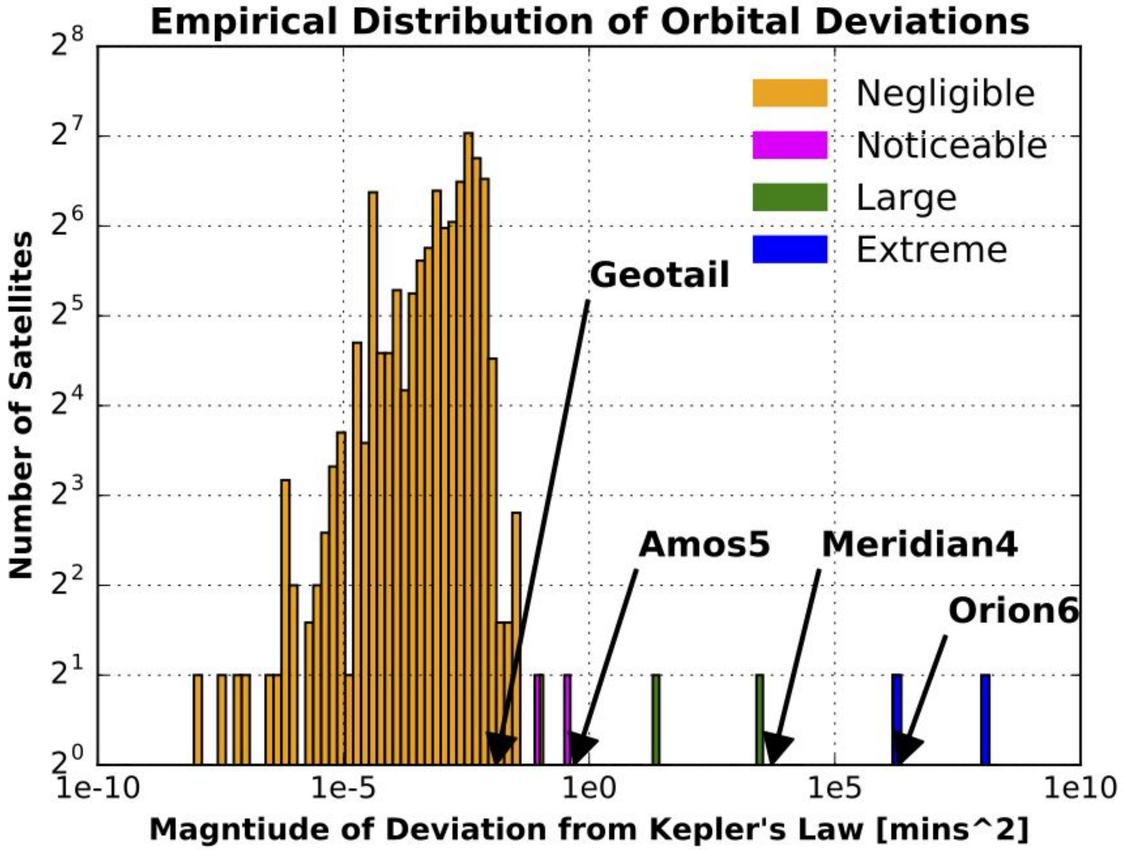
Kepler's 3rd Law

Cluster the residuals via a Dirichlet Process Mixture of Gaussians

Saad and M., NIPS 2016

Custom modeling: integrating physics with statistics

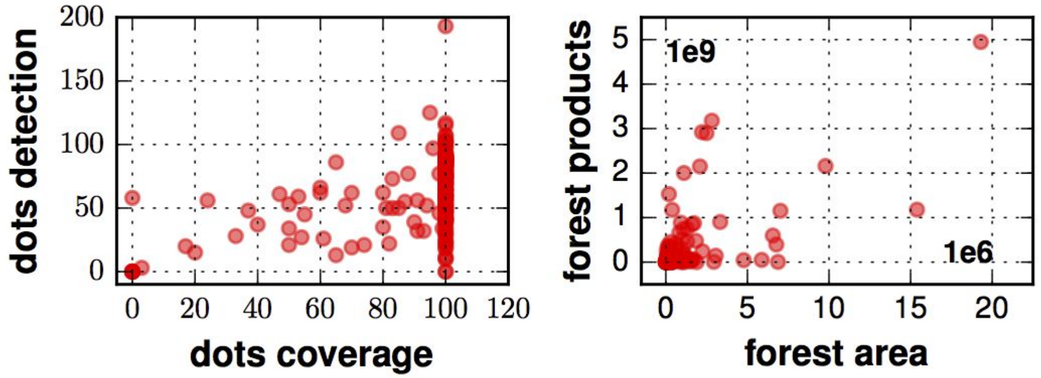




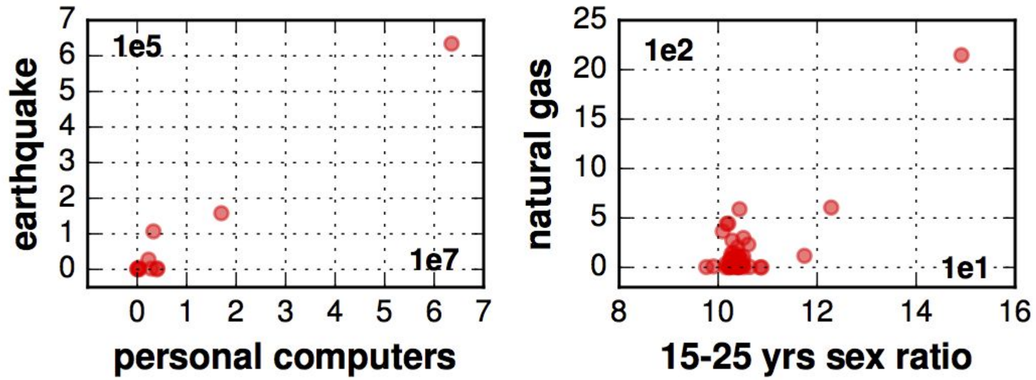

```
%bql  INFER period_minutes, perigee, name, keper_cluster, kepler_noise
...   FROM satellites
```

Saad and M., NIPS 2016

Detecting predictive relationships between variables



BayeDB finds nonlinear / heteroscedastic relationships that Pearson correlation misses



BayesDB suppresses spurious relationships that Pearson correlation reports

Saad & M., *Detecting dependencies in sparse, multivariate databases using probabilistic programming and non-parametric Bayes, AISTATS 2017*

Probabilistic data search: find urban colleges with >10% admit rates but comparable in instructional investment to Harvard, MIT, Yale, and Duke

```
%bql SELECT
...    "institute",
...    "admit_rate",
...    "median_sat_math",
...    "tuition",
...    "median_student_debt",
...    "instructional_invest",
...    "locale"
... FROM college_scorecard
... WHERE
...    "admit_rate" > 0.10
...    AND "locale" LIKE '%City%'
... ORDER BY
...    RELEVANCE PROBABILITY
...    TO EXISTING ROWS IN (
...     'Duke University',
...     'Harvard University',
...     'Mass Inst Technology',
...     'Yale University',
...    )
...   IN THE CONTEXT OF
...    "instructional_invest"
... DESC
... LIMIT 10
```

BQL search query

institute	admit	sat	tuition	debt	investment	locale
Duke University	11%	745	47,243	7,500	50,756	Midsize City
Georgetown Univ	17%	710	46,744	17,000	31,102	Midsize City
Johns Hopkins Univ	16%	730	47,060	16,250	77,339	Midsize City
Vanderbilt Univ	13%	760	43,838	13,000	79,372	Large City
University of Penn.	10%	735	47,668	21,500	49,018	Large City
Carnegie Mellon	24%	750	49,022	25,250	31,807	Midsize City
Rice University	15%	750	40,566	9,642	40,056	Midsize City
Univ Southern Calif	18%	710	48,280	21,500	43,170	Midsize City
Cooper Union	15%	710	41,400	18,250	21,635	Large City
New York University	35%	685	46,170	23,300	30,237	Large City

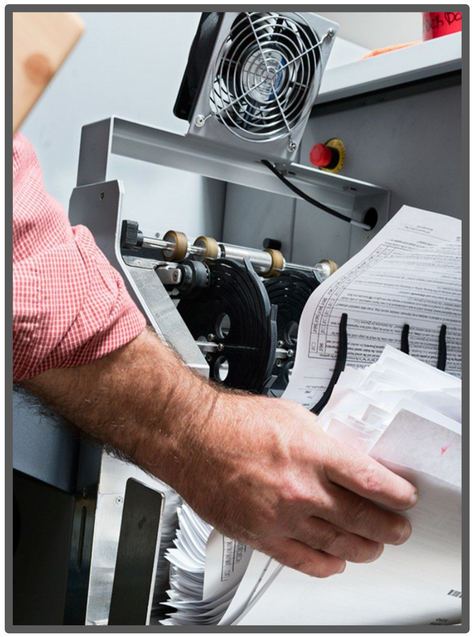
Result set

Saad & M., *Probabilistic Data Search via Probabilistic Programming and Non-parametric Bayes,* in review (arXiv, 2017)

Outline

1. What are “data” and “data science”?
2. Key challenges: cost and credibility
3. AI-assisted data science with BayesDB
4. Example: exploratory data analysis for the RISC2 diabetes study
5. Capabilities
6. Current research

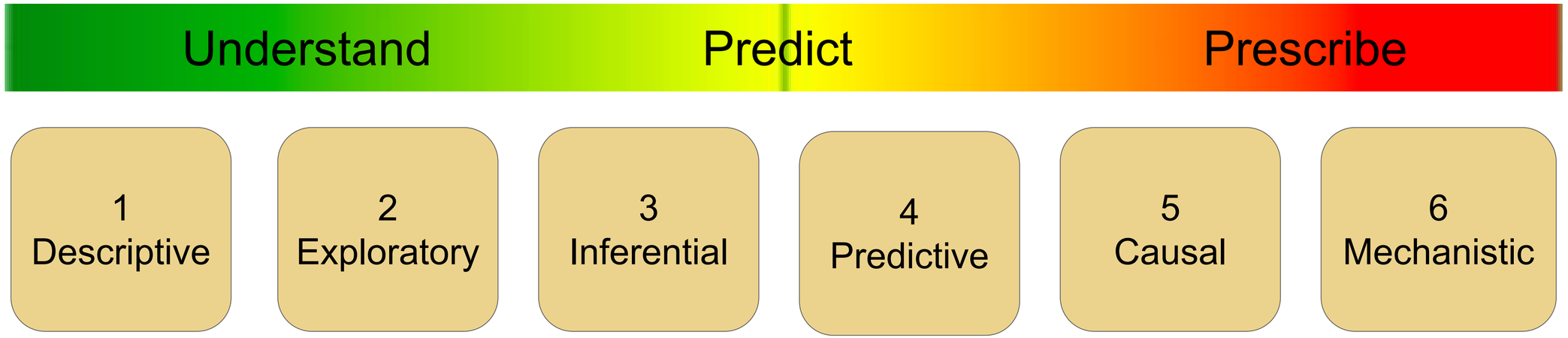
Data science today is like data storage without databases



Searching for guns by id, USA, 21st century, using microfilm

GQ, August 30, 2016
http://www.gq.com/story/inside-federal-bureau-of-way-too-many-guns

BayesDB provides AI assistance for data science



Goal: build an AI that can do in seconds to minutes what currently takes hours to days for someone with good statistical judgment

Research area #1: statistical foundations

1. How can we quantify the quality of inferences from BayesDB?

2. How can we relax the assumptions of IID sampling?

3. How can we formalize imperfect randomization and control?

4. Can we use posterior distributions on mutual information to perform causal inference?

5. Can we automate the transition from multivariate statistics to discriminative ML, as query workloads grow more predictable?

The public and non-profit sectors can’t afford data scientists

1 Data Scientist

4.8 / 5
Job Score

4.4 / 5
Job Satisfaction

$110,000
Median Base Salary

4,184
Job Openings

View Jobs

Research area #2: applications in the public interest

Can we turn data sources into intelligent data assets?

1. **Bangladeshi children's study, administered by the Bill & Melinda Gates Foundation & Boston Children's Hospital**

2. **USC Children's Health Study** --- ~10,000 children followed over 10 years

3. **MIT Digital Currency Initiative & Deloitte** --- AI-assistance for private & public audits

4. **Stanford database of US school performance** --- ~380,000 records, ~150 variables

5. **US Army STARRS mental health database** --- ~50,000 soldiers, ~650 variables

Research area #3: new data science capabilities

1. **Probabilistic record linkage** via STOCHASTIC PRIMARY KEY and FOREIGN KEY
2. **SparkBayesDB** rather than current sqlite3 embedding
3. **AI-assisted data visualization**
4. **Adaptive data acquistition via expected value of information**
5. **Integrated action selection:** OPTIMIZE not just ESTIMATE

Thank you!

Contact vkm@mit.edu if you are interested in:

- contributing to the open-source effort

- collaborating on applications or foundations

- testing a commercial version from Empirical Systems, Inc.

- helping us run 4-hour “AI-assisted data science” workshops this summer