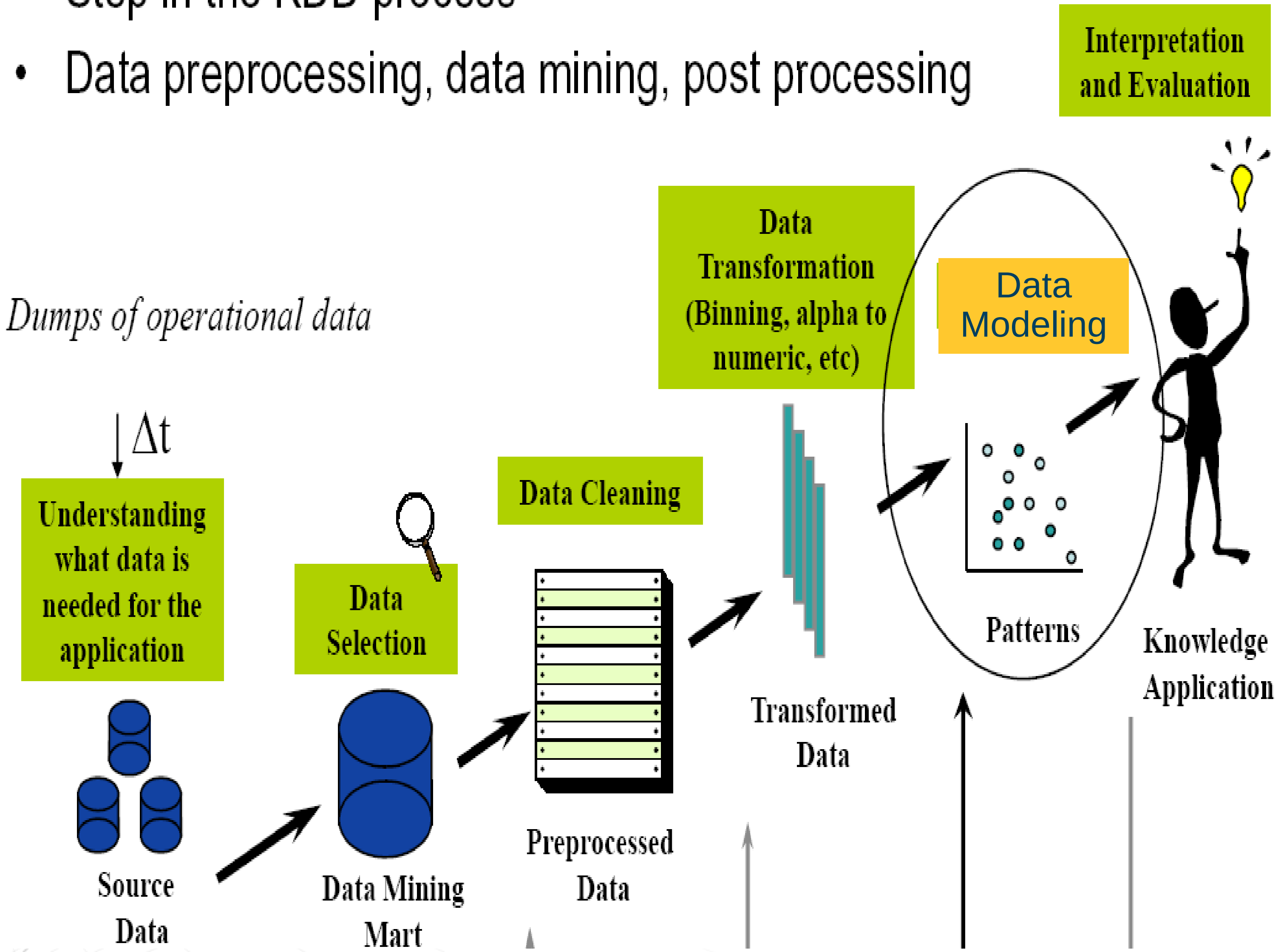
The Traditional˛ Fashionable˛ and Future Modeling Techniques & Application in Financial Companies

December 11, 2015

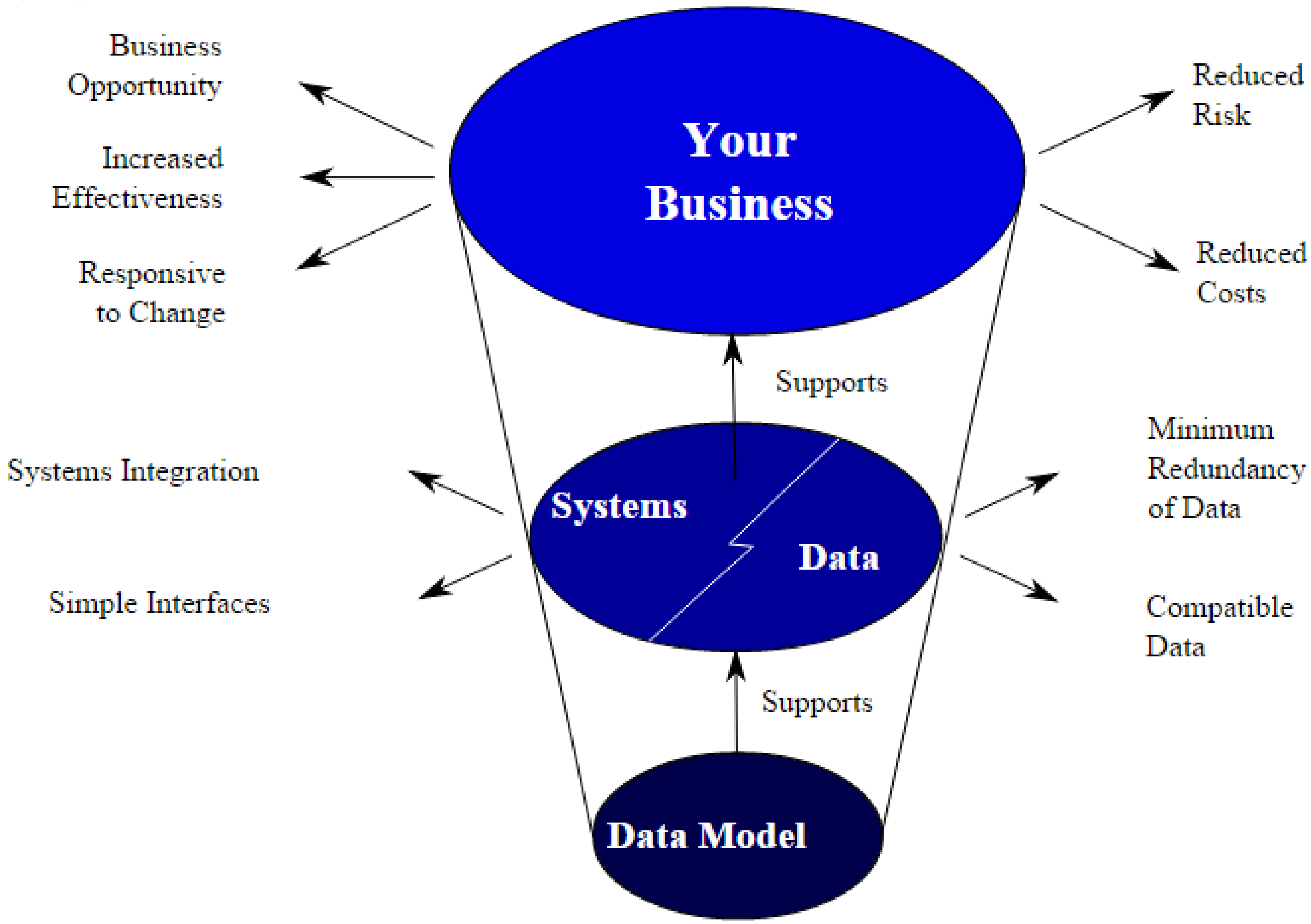
FICO

Why Data Modeling

- Step in the KDD process
- Data preprocessing, data mining, post processing



How Data Models Deliver Benefit



What Modeling Techniques

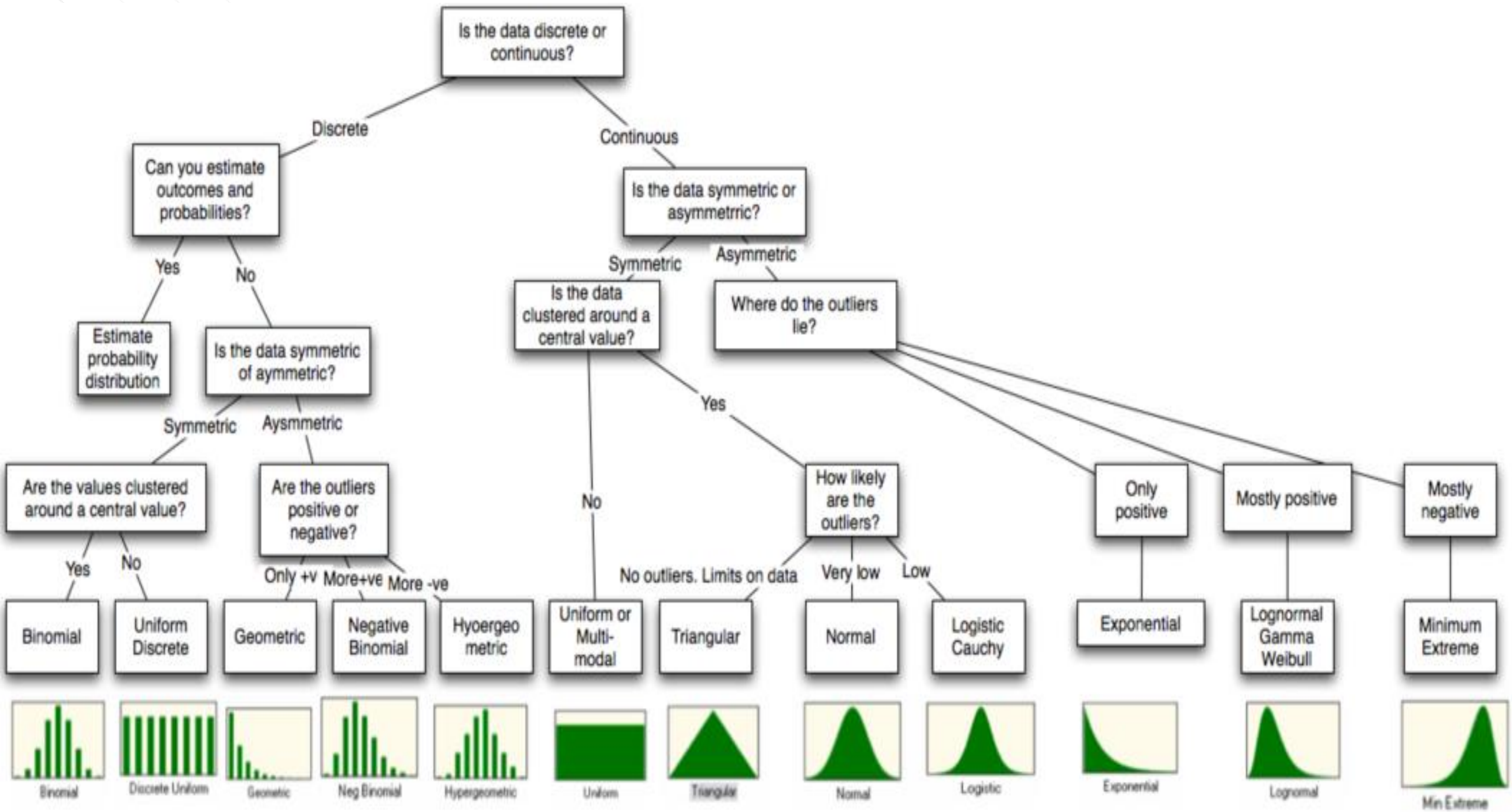
- **The Traditional(传统的) Modeling Techniques**
  - **Regression: continuous target**
  - **Classification: discrete target (binary or multiclass)**

- **The Fashionable(时髦的) Modeling Techniques**
  - **Neural Networks: represent the functionality of the human brain**
  - **Association rules: frequent correlations**
  - **Scorecard: scores directly estimate or rank-order future outcomes**

- **The Future(将来的) Modeling Techniques**
  - **Big Data Modeling: the intersection of artificial intelligence**
  - **Text mining of unstructured data: semantic models**
  - **Machine learning: natural language processing**
  - **Hadoop Data Modeling: extract knowledge from large databases**

How to Select Modeling Techniques

Based upon Distribution of Data



FICO

Logistic Regression

» ***Why use logistic regression?***

- We try to model the discrete responses.
  - Binary responses (for example, Good and Bad, Default and Non_Default, Fraud and Non_Fraud …)
  - Ordinal responses (for example, Tiers of PI: 1, 2, 3, 4, 5)
  - Nominal responses (for example, Marketing channels: agent-led, bank-led, company-led, Internet-led)

FICO

Logistic Regression

» ***Logit Transformation***

- We can not apply the linear regression (the errors will be either 0 or 1)
- We will apply the logistic regression
- Logistic regression models transform probabilities called logits:

$$\log it = \log\left(\frac{p}{1-p}\right) = \sum \beta_i x_i$$

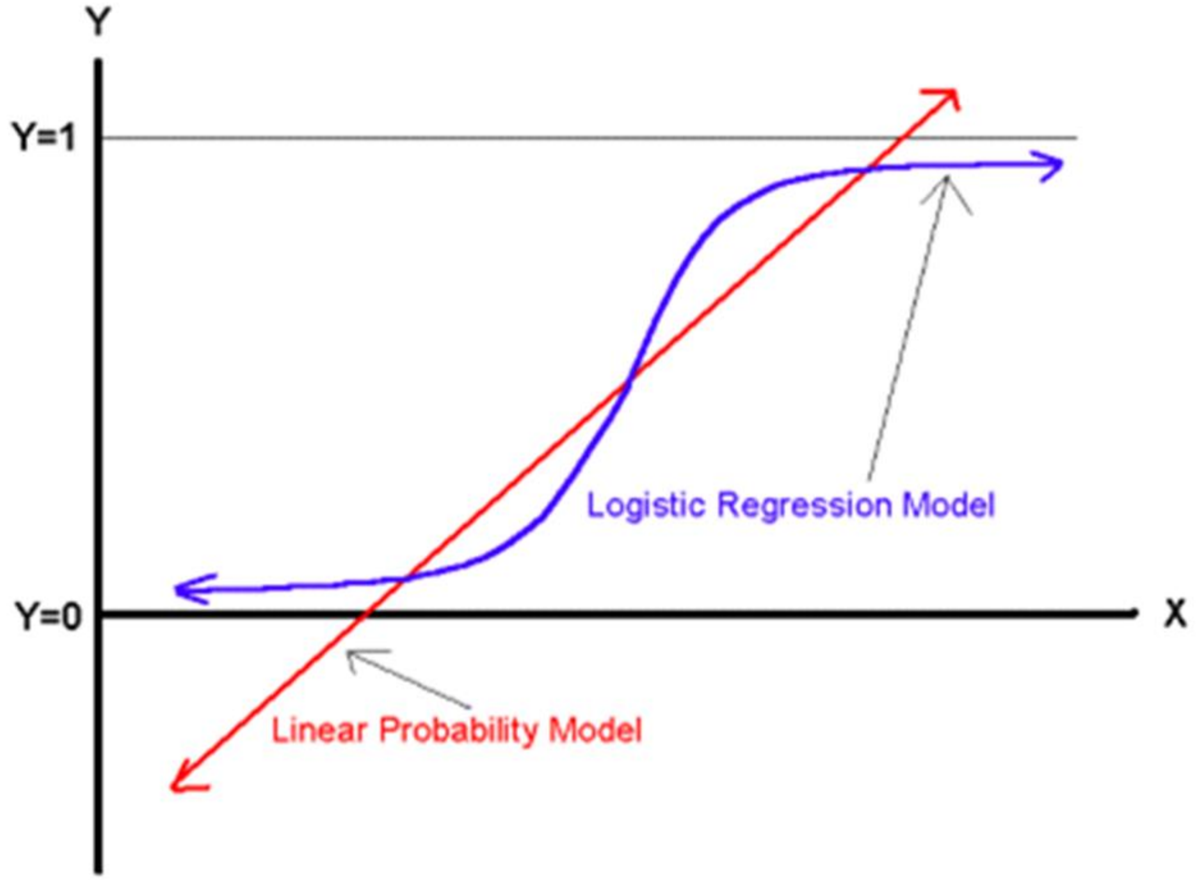
» which yields the probability estimation:

$$p = \frac{\exp\left(\sum \beta_i x_i\right)}{1+\exp\left(\sum \beta_i x_i\right)}$$

The Traditional Modeling Techniques

Logistic Regression

»*Logit Transformation*



FICO

Logistic Regression

»*For Example*

Final model

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Pr > ChiSq	Standardized Estimate
Intercept	1	-5.2168	<.0001	
criticalnum	1	0.7500	<.0001	0.2396
debttoincome	1	0.0999	<.0001	0.4380
delinquentnum	1	0.7087	<.0001	0.3164
inquirynum	1	0.1138	0.0019	0.0972
linesnum	1	-0.0197	0.0094	-0.1018
tradeage	1	-0.00563	<.0001	-0.2571

LOGIT=-5.2168 + 0.7500*criticalnum + 0.0999*debttoincome + 0.7087*delinquentnum + 0.1138*inquirynum - 0.0197*linesnum - 0.00563*tradeage

FICO

Logistic Regression

»*How to Interpret Your Model?*

- Interpretation of Regression Coefficient (β):
  - In linear regression, the slope coefficient is the change in the mean response as *x* increases by 1 unit
  - In logistic regression, we can show that:

$$\frac{odds(x+1)}{odds(x)} = e^{\beta} \qquad \left(odds(x) = \frac{\pi(x)}{1-\pi(x)} \right)$$

Logistic Regression

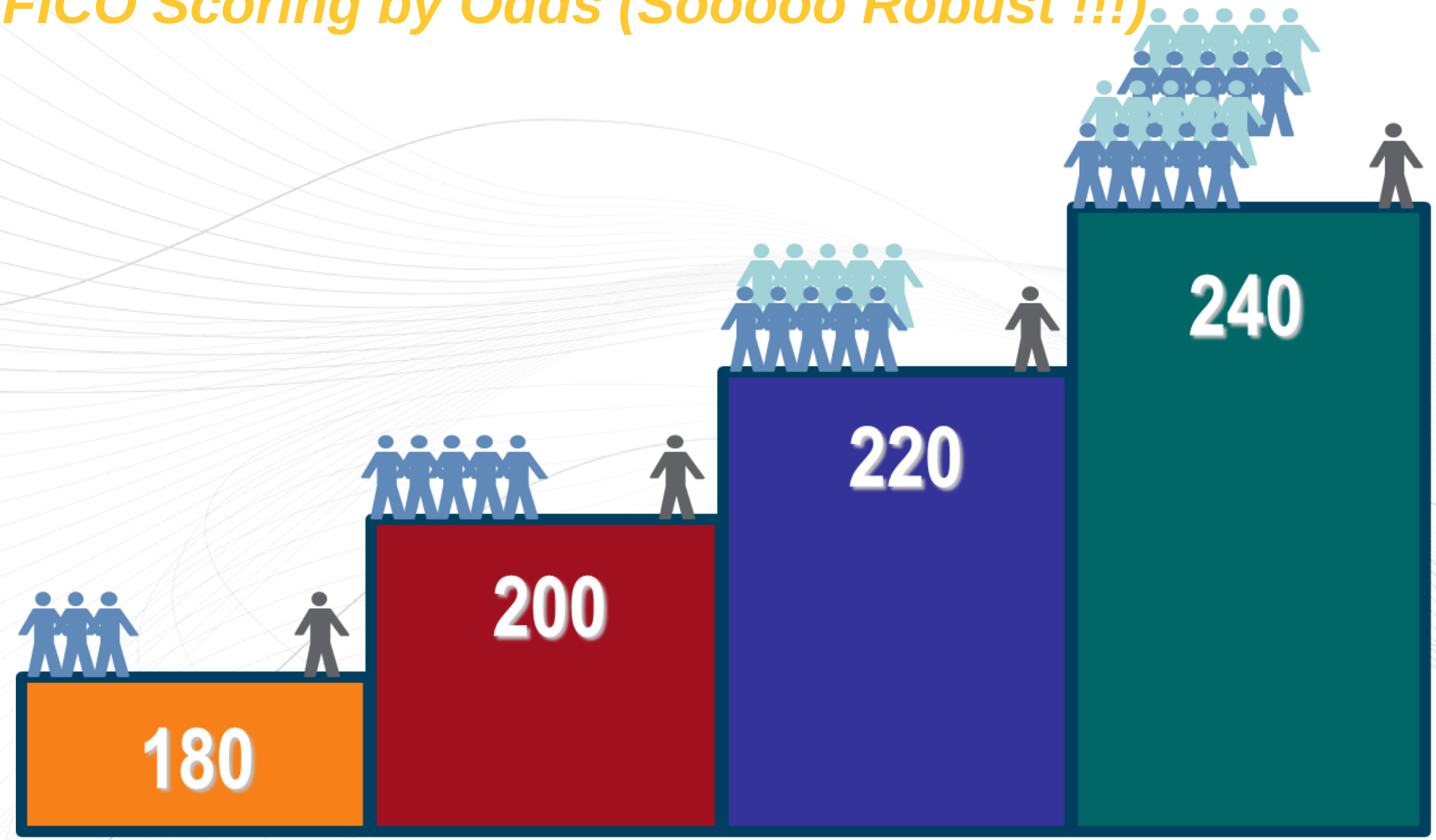
»*How to Interpret Your Model?*

- Interpretation of Regression Coefficient (β):
  - Thus e^{β} represents the change in the odds of the outcome (multiplicatively) by increasing x by 1 unit
  - If $\beta = 0$, the odds and probability are the same at all x levels ($e^{\beta}=1$)
  - If $\beta > 0$, the odds and probability increase as x increases ($e^{\beta}>1$)
  - If $\beta < 0$, the odds and probability decrease as x increases ($e^{\beta}<1$)

The Traditional Modeling Techniques

Logistic Regression

» *FICO Scoring by Odds (Sooooo Robust !!!)*



The Traditional Modeling Techniques

Modeling Value

Data Collection: Historical Sales Revenue, Sales Coverage, Competitor Marketing, TV Advertising, Paid Search, Macroeconomic Data (more drivers if you have them)

Optimization: Use the econometric model as an input to *proc optmodel* that identifies sales force and marketing investment levels that maximize revenue

Modeling: Develop Econometric Models relating Historical Sales with important drivers of business

Results: Compares optimized investment levels with actual, yields insight into opportunity to increase sales revenue through marketing reallocation

The Traditional Modeling Techniques

FICO

Modeling Value

»*For Example*

Model Output

Estimated Elasticities

10% increase in sales force leads to a 6.4% increase in revenue

Variable	Elasticity	Standard Error
Sales Force	+0.64	0.08
TV	+0.25	0.04
Paid Search	+0.15	0.02
Comp TV	-0.13	0.03
Oil	-0.91	0.03
GDP	+3.60	0.60

FICO

Modeling Value

»*For Example*

»Optimization Problem

- Given the results of the econometric model, how can we maximize Sales Revenue?
  - Suppose we have a budget of $3M and initial assumptions of GDP and competitor TV year advertising for next year
  - What is the allocation of the budget across Sales Force, TV, and Paid Search that maximizes sales revenue?

The Traditional Modeling Techniques

FICO

Modeling Value

» *For Example*

»Optimization Problem

```
/* declare variables */
var sales >= 1, tv >= 1,  paid_search >=1 ;

/* declare objective */
max sales_revenue = exp(-19.0856713 +  0.64*log(sales) + 0.25*log(tv) +
                        0.15*log(paid_search) - 0.13*log(&comp_tv) +
                        3.16*log(&gdp) + (-.91)*log(&oil));

/* declare constraints */
con sales + tv + paid_search  <= 3000000;

solve;
```

The Traditional Modeling Techniques

FICO

Modeling Value

» *For Example*

Optimization Results: Invest proportionally to the effectiveness of activities

Variable	Elasticity (Raw)	Elasticity (Normalized)	Optimal Allocation	%
Sales Force	0.64	0.615	$1,846,154	0.615
TV	0.25	0.240	$721,154	0.240
Paid Search	0.15	0.144	$432,692	0.144

Expected to generate $229.6M in sales revenue.

Association Analysis

(Market Basket Analysis)

Rule	Support	Confidence
A ⇒ D	2/5	2/3 (P(A & D) /P(A)
C ⇒ A	2/5	2/4 (P(A & C) /P(C)
A ⇒ C	2/5	2/3
B & C ⇒ D	1/5	1/3

A➔B:

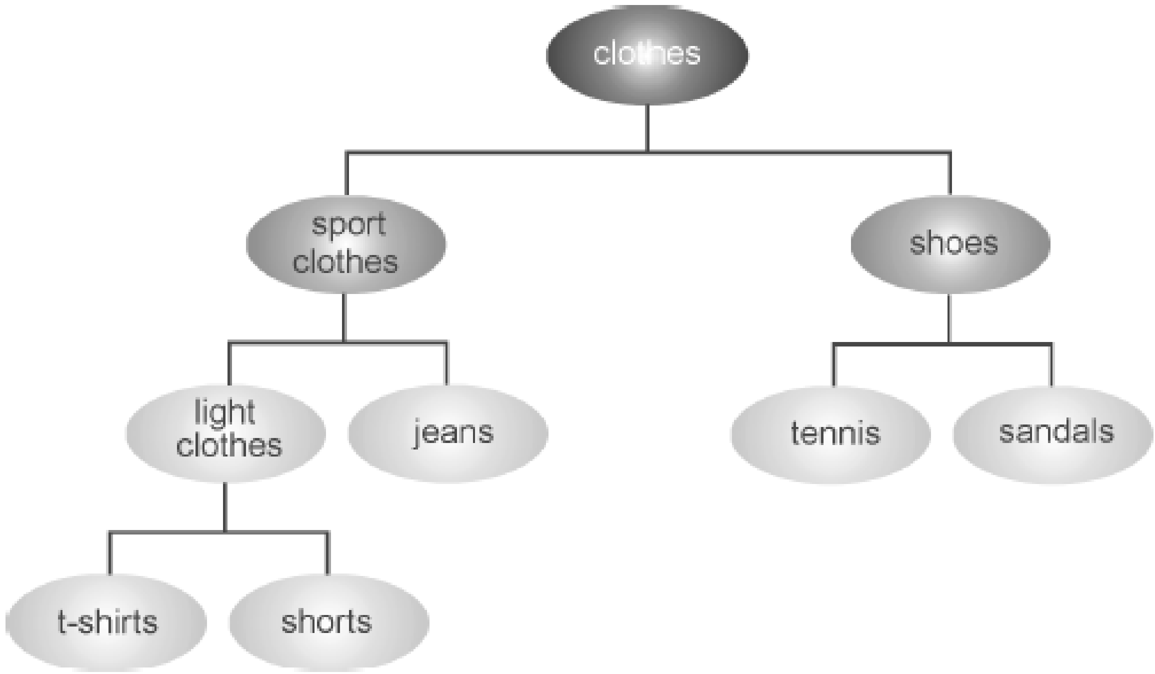
Support = Pr(A & B).

Confidence: if A, then B (A ‘causes’ B) = S (A ➔B) / S (A)

LIFT (A ➔B) = confidence / expected confidence if A and B are statistically independent: S(A➔B) / S(A) S(B).

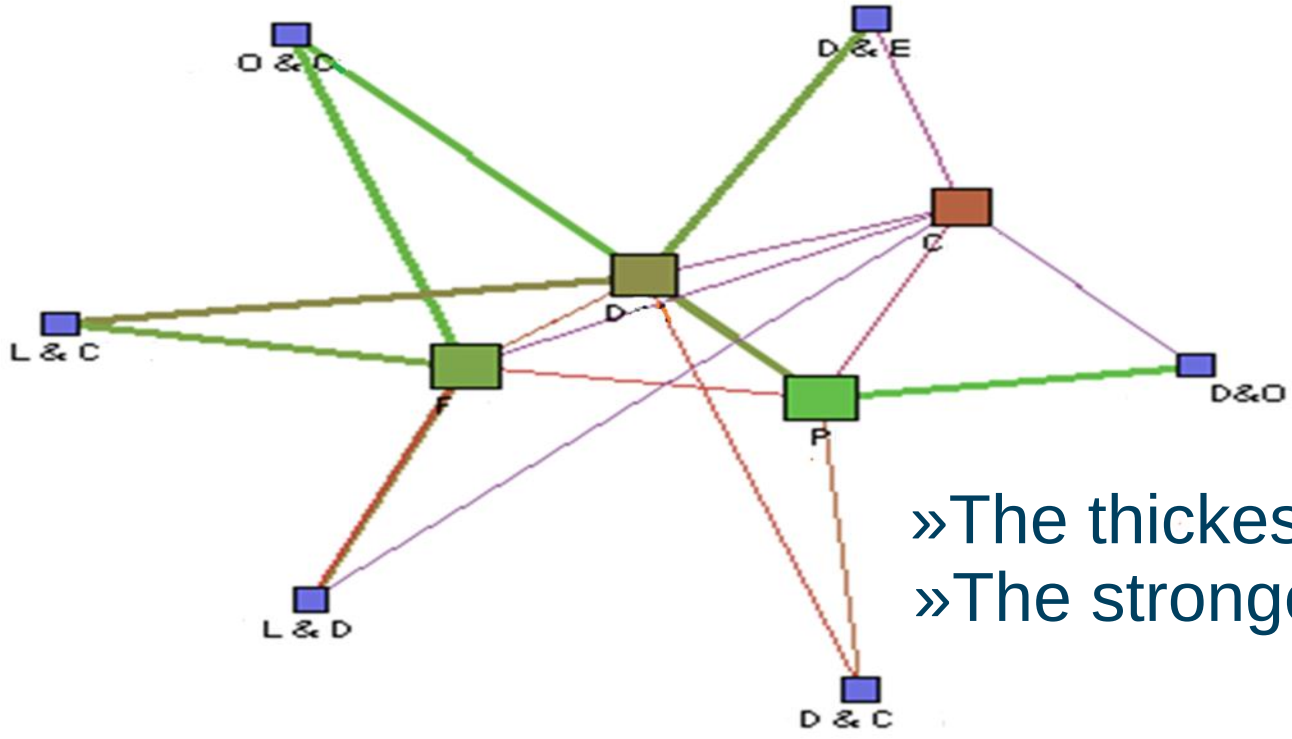
Association Analysis

Taxonomy – Hierarchy of products



Association Analysis

The link graph were used to produce the intelligent algorithm of offers based on the lift, confidence, and level of support



- »The thickest linkline
- »The strongest confidence

Association Analysis

	Relations	Lift	Support(%)	Confidence(%)	Rule
1	2	1.02	54.17	63.15	CKING ==> SVG
2	2	1.02	54.17	87.56	SVG ==> CKING
3	2	1.10	36.19	94.11	ATM ==> CKING
4	2	1.10	36.19	42.19	CKING ==> ATM
5	2	1.08	25.69	66.81	ATM ==> SVG
6	2	1.08	25.69	41.53	SVG ==> ATM
7	2	1.17	16.47	100.00	HMEQLC ==> CKING
8	2	1.17	16.47	19.20	CKING ==> HMEQLC
9	2	1.04	15.72	64.08	CD ==> SVG
10	2	1.04	15.72	25.40	SVG ==> CD

FICO Has a Lot of Powerful Modeling Techniques

FICO

- Model Builder has modeling techniques
  - Linear Regression
  - Logistic Regression
  - Neural Networks
  - Semantic models
  - Scorecard
    - Divergence
    - Bernoulli likelihood
    - Least squares
    - Multi-goal
    - Ranged divergence
    - Segmented Ensemble

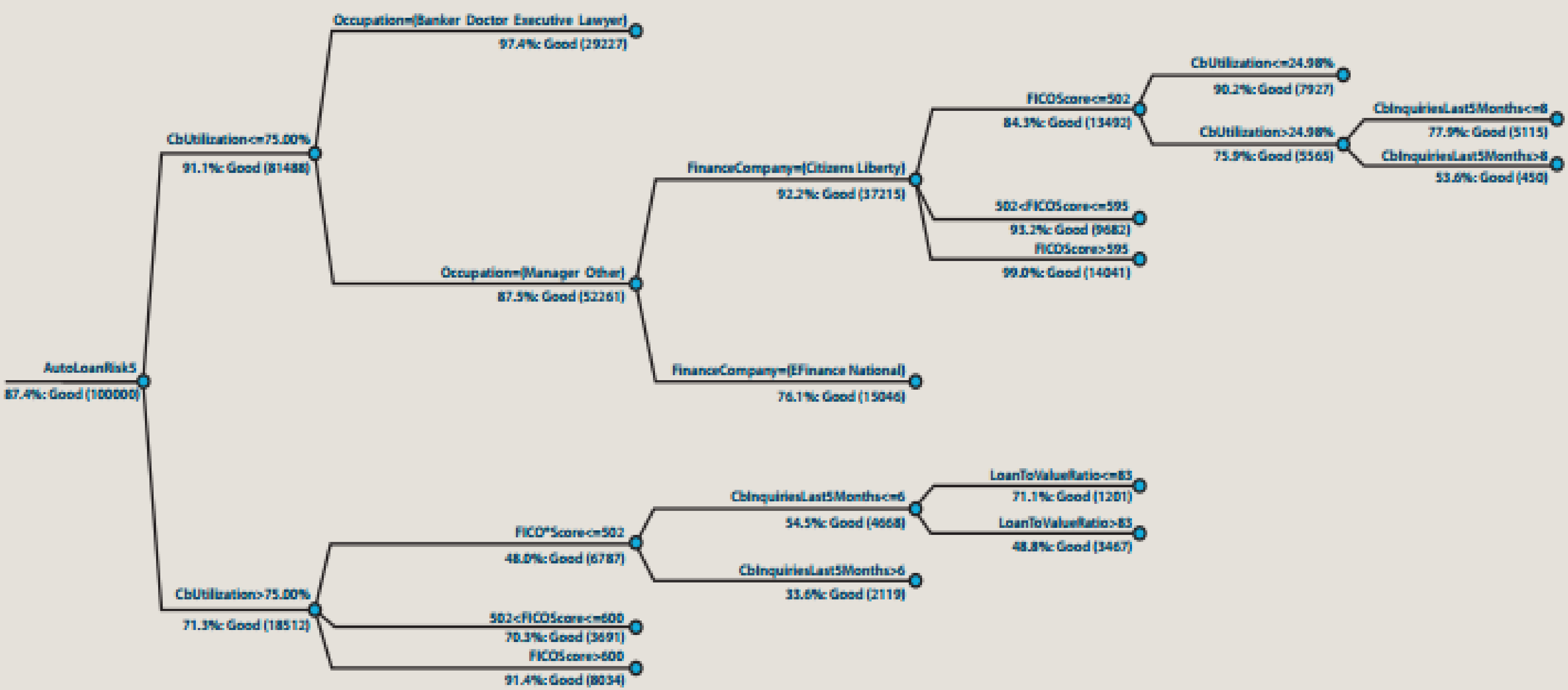
FICO Has Decision Tree

»FICO Has

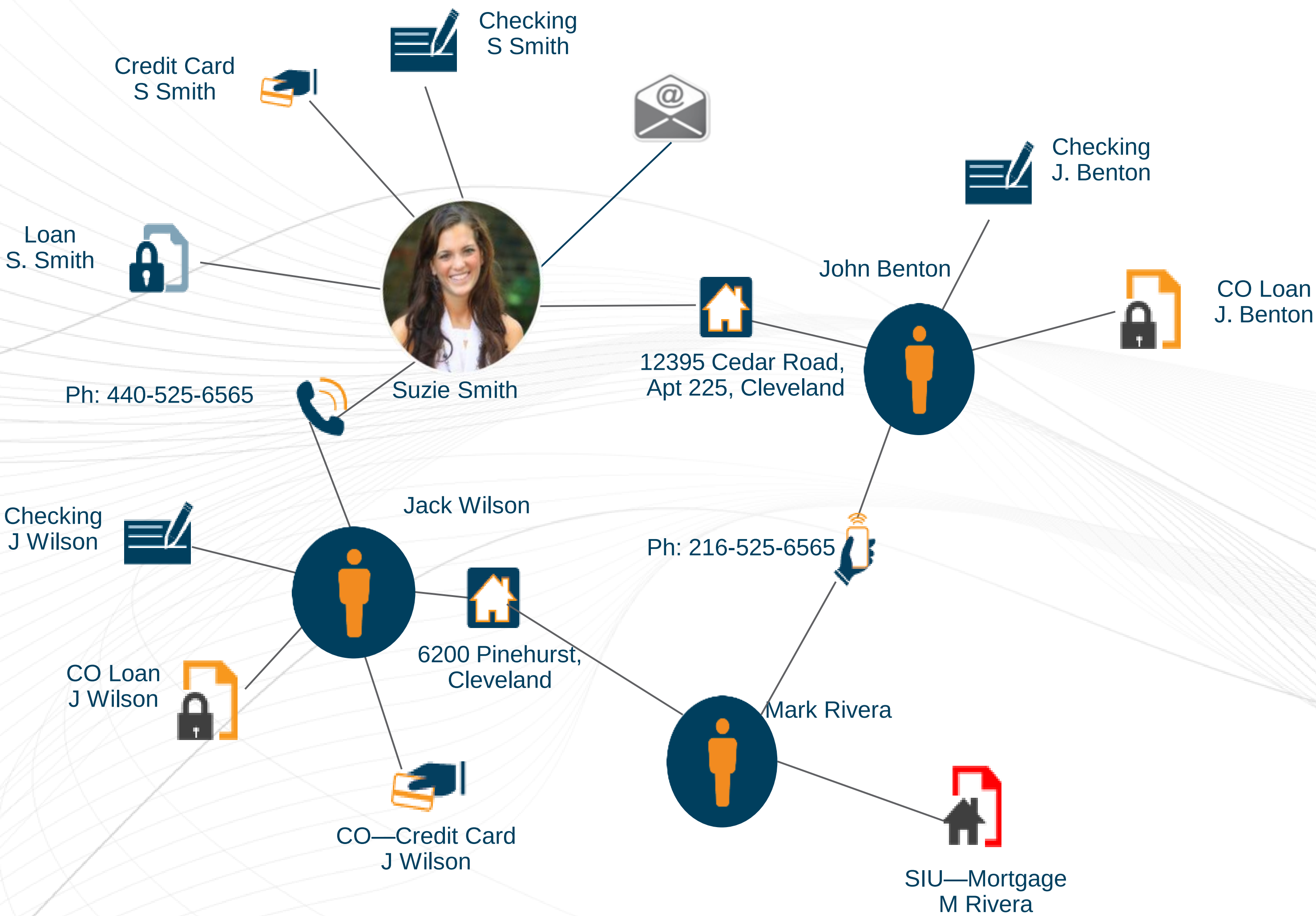
The CRT algorithm

The CHAID algorithm

FIGURE 5: CHAID TREE PREDICTING CREDIT RISK



FICO Has IRE



Credit Card
S Smith
Checking
S Smith
Checking
J. Benton
Loan
S. Smith
John Benton
CO Loan
J. Benton
12395 Cedar Road,
Apt 225, Cleveland
Ph: 440-525-6565
Suzie Smith
Jack Wilson
Checking
J Wilson
Ph: 216-525-6565
6200 Pinehurst,
Cleveland
CO Loan
J Wilson
Mark Rivera
CO—Credit Card
J Wilson
SIU—Mortgage
M Rivera

FICO大数据云评分

打破数据孤岛：整合中国市场上典型的大数据资源

卡交易及账户金融数据

- 全面、真实、实时、具体的跨行交易数据
- 覆盖消费、取现、转账、逾期等信息

小贷行业数据

- 覆盖小贷及P2P行业的黑名单
- 小贷行业信贷查询情况

运营商数据

- 通话信息
- 上网信息
- 定位信息

航空数据

- 订票信息
- 飞行信息
- 其他信息

政府部门数据

- 税务数据
- 法务数据
- 社保数据等

持续丰富的其他各类数据

- 电商数据
- 学历学籍数据
- 等

FICO大数据云评分

丰富珍贵的外部数据将持续接入费埃哲大数据云平台
FICO
大数据
云平台
BIG DATA
政府部门数据
▪ 税务数据
▪ 法务数据
▪ 等
……
更多外部数据将持续接入
运营商数据
▪ 通话信息
▪ 上网信息
▪ 定位信息
黑名单数据
金融交易数据
▪ 交易信息
▪ 商户信息
▪ 逾期信息
▪ 等

FICO大数据云评分

大数据云评分使用方式

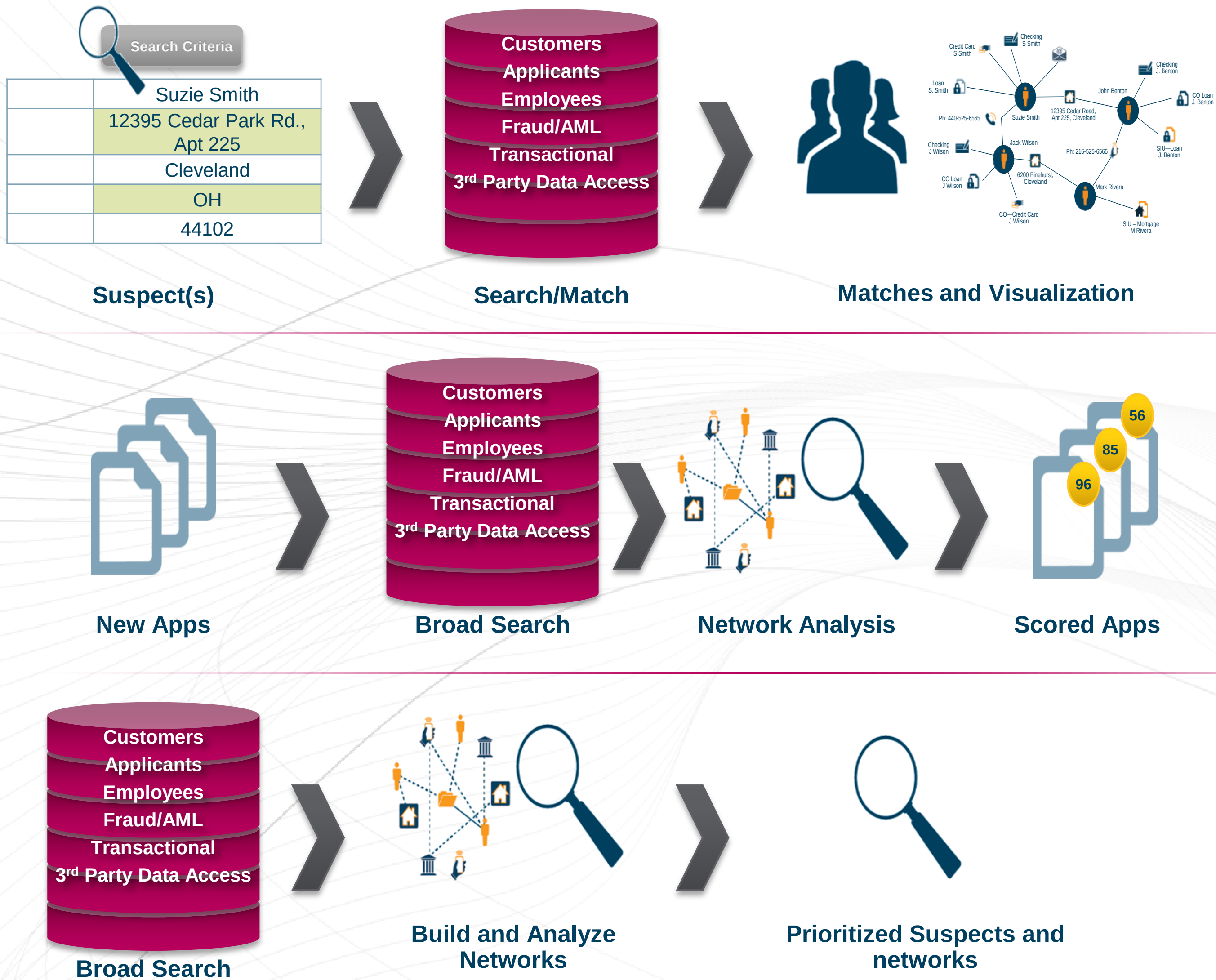
- ✓ FICO – 专注于模型评分，所有数据不落地、不留痕、无存储、实时删除，保护数据合作伙伴数据价值的专属性、保护用户的客户信息保密性
- ✓ 用户的客户信息为电话号码、或银行卡卡号，全部MD5加密、脱敏

FICO Is Introducing FICO® SCORE XD

FICO Is Launching FICO® Big Data Analyzer

»*FICO® Big Data Analyzer is a purpose-built analytics environment for business users, analysts and data scientists to gain valuable insights from the exploration and analysis of any type and size of data on Hadoop. It makes Big Data accessible by masking Hadoop complexity, allowing all users to drive more business value from any data.* **CURRENTLY, BIG DATA ANALYZER USES A RANDOM FOREST MACHINE LEARNING ALGORITHM TO BUILD PREDICTIVE MODEL.**

FICO Has …… LUCK



I WANT YOU

The Traditional、Fashionable、and Future Modeling Techniques & Application in Financial Companies

December 11, 2015