System Security Group | Lancaster University

Economics of Security

Aims

- Provide a new tool set to understand security issues
- Help us to understand broader ranges of risks to security outside of the technical

Why Are We Not More Secure?

System Security Group | Lancaster University

- We know how to build secure systems!
- Wrong incentives
  - Guards don't suffer → If system fails (not the end of the world)
  - Security shift liability → shifting blame on someone else
- The Internet, millions of independent principles interacting
  - Reasonable global outcomes from selfish local actions
- Incentives drive security design and policy

Is Network insecurity the Same as Air Pollution?

- Insecure machines connected to the Internet have costs for all
  - Who should bear all the cost?
    - Individuals, vendors, regulators, authorities?
- Security Economics can be used to help understand
  - Security issues: Privacy, Spam, Phishing etc
  - System Dependability: optimum ratio of dev to test
  - Analysis of Policy Problems: DRM

Public Good

- Same quantity of good regardless of desire
  - Air Quality
- Properties:
  - Non-rivalrous: my use does not deplete yours
  - Non-excludable: inefficient to stop people from using them, lighthouse not good efficent of one entity to stop another from using that
- Public good supply
  - Directly from governments: national defence
  - Patents and Copyright: temporary monopoly

Security and Public Good

- Many aspects of security are public goods
  - Air defence is not an individual action → Public Taxpayers → National Forces
- Strong externalities
  - Cost borne by others
  - One insecure system connected to the Internet affects all
    - Air pollution, toxic dumping
- Is IT security air defence or air pollution
  - Spam used to be a large number of small groups
  - Spam now a small group of powerful teams
  - Is it a national defence issue?

Should defence be provided by national forces or individuals?

The Price of a Good

- Jerons and Menger: the price of a good in equilibrium is the marginal cost of production
- A good cost £10 to produce, not every producer sells at £10, only marginal ones
  - Those producers just stay in business
  - If price goes down marginal producers close
  - If price goes up marginal producers open

Supply/Demand → Profit

The Price of Information

- In a competitive market price should be its marginal cost
  - Information has high fixed costs
  - Information re-production is free
  - Reason for so much free info, zero is a fair price
- If you can produce at 0 cost then the incentive is to cut without limit to undercut competitors
- Encyclopaedias
  - Britannica $1600, Encarta $49.95, Wikipedia $0

Business Models

- Linux is free, support is not
- Snort is free, rules are not
- Open source devs contribute for free, but gain CV experience
- Information Goods and Services Characteristics
  - High fixed costs, low production = service or advertising model
  - Dominated by network effects
  - Technical lock in → Expenses to switch to a computing technology or supplier
  - Tend to lead to dominate firms and monopolies

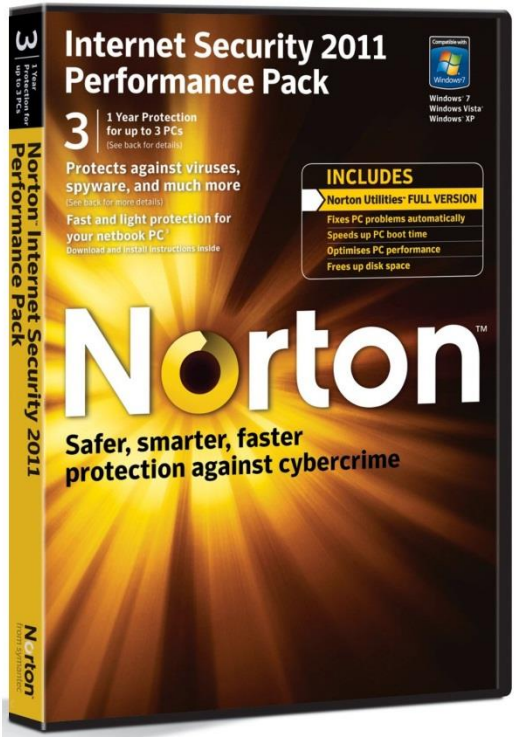
The Value of Lock In

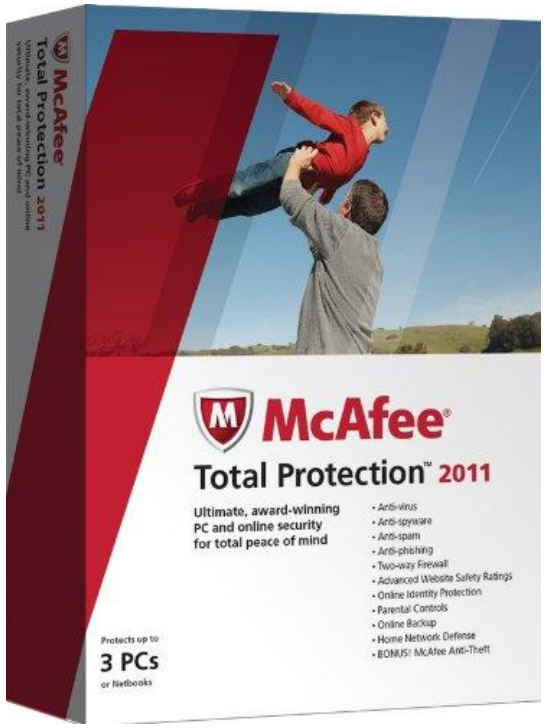
- Shapiro and Varian: The value of a company is the total lock in cost
- Consider a company with 100 staff with Office @ £500 a pop
  - Company switch to Open Office save £50000
  - If costs of change were less, they would switch
  - If they were more MS would put up price
- Consider Apple and Itunes

Information Asymmetry

- George Akerlof - “Market for Lemons” – 1970
  - Some know more than others
- Example
  - 100 used cars, 50 good £2000, 50 bad £1000
  - Sellers know which is which, buyers don’t
  - What is the market price of the used car?
  - At £1500 no good cars will be offered, so price will be closer to £1000.

Can You Decide?





- Poor security products dominate when users can't tell the difference
  - Race to the bottom on price

What about you? Why do you get insurance?

- Hidden information – adverse selection
- Hidden action – moral hazard
- Volvos are safe cars but have higher accident rates
  - Do bad drivers buy them? – AS
  - Do you drive badly because you think you are safer? – MH
- Consider AV products?
  - Do they make you feel safer – act riskier
  - Get the best AV because you are risky
- What about in private browsing?

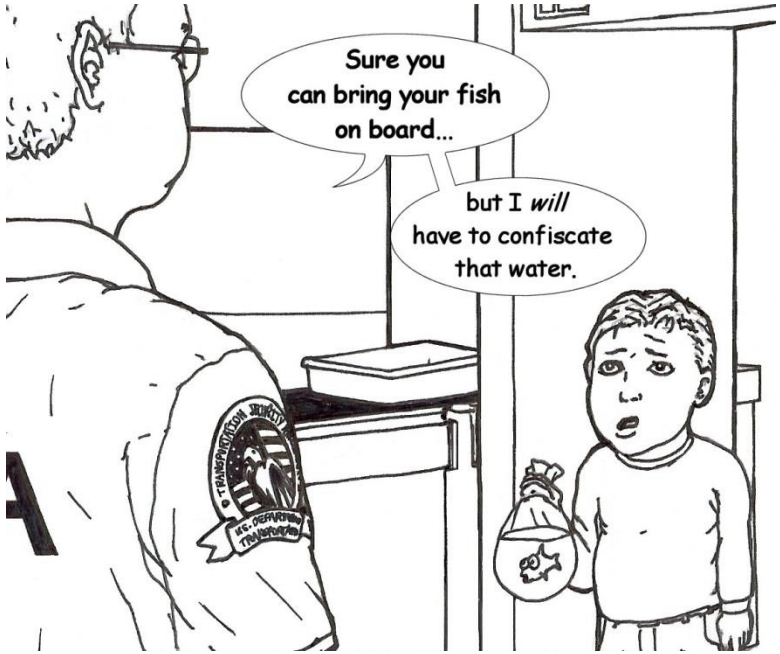
legimate reasons? Want more protection?

Why does security fail?

System Security Group | Lancaster University

- Those guarding have no incentives to protect what we think is important.
  - Guards don't suffer a point of failure → Not directly
  - Risks are dumped on others → Lack of liability
- Security is a power relationship
  - Principles control security meaning to advance power

Have to increase power to give more protection



What is the best Strategy?

- Jack Hirshleifer founded conflict theory
- Consider the country of Anarchia
  - Flood defence managed by everyone on the coast
    - As good as the weakest link
    - The more defenders the greater the number of weaknesses
  - Missile defence is based on best shot
    - Best effort

System Reliability and Freeriding

- Hal Varian work applying previous theory to effort applied in securing systems.
- **Total effort.** Reliability depends on the sum of the efforts exerted by the individuals.
- **Weakest link.** Reliability depends on the minimum effort.
- **Best shot.** Reliability depends on the maximum effort.

How should you structure your dev team?

How should you structure your dev team?

System Security Group | Lancaster University

- Program correctness can depend on minimum effort
  - Most careless programmer
- Software vulnerability testing may depend on sum of all testers efforts
- Security depends on best effort
  - Actions taken by individual champion, architect/designer
- More agents
  - Less reliability in min. effort case
  - More reliability in total effort case

Whys is Windows insecure?

System Security Group | Lancaster University

- Why are there still so many bugs when Windows is so dominant?
- Why no comparable effort in commodity platforms compared to defence or healthcare?
- Technically we know how to build good systems, so why don't we?
- Product insecure at first then improve, why?
  - Symbian, IBM
  - Win95->Win98->WinXP->Vista->Win7->Win10

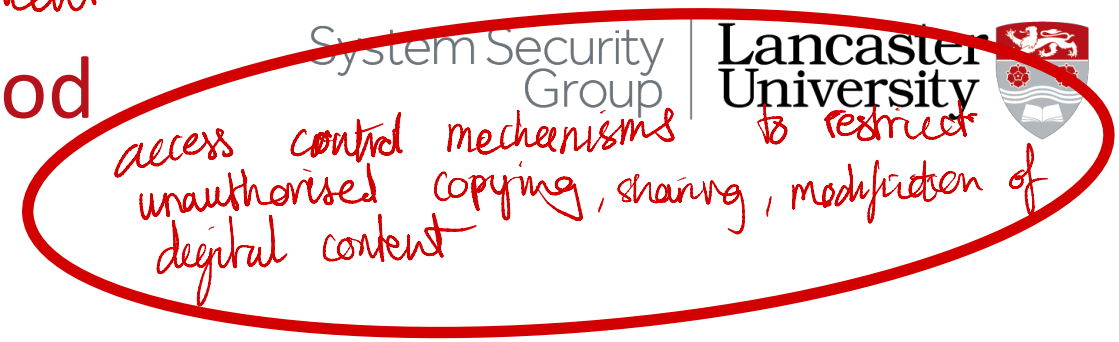
What is the software market like?

System Security Group | Lancaster University

- Low marginal but high fixed costs → low for reproduction
- Network effects → more ppl use product → the software spreads
- Technical lock-in → very difficult to change to other services
- Race to dominate, the dominant firm gets all the money → It has become very difficult to enter market
- MS 1990's philosophy *"ship it Tuesday and get it right by V3"* is rational → Publish whatever yo
- You must appeal to complementers
  - Security gets in the way
  - Add security later, but make sure it helps lock in

Digital Rights Management

DRM, is it a good thing?



- Varian, DRM is about tying, bundling and price discrimination
- Transfer of control from owner of contain to owner of file → to Platform
  - Potential for lock in increases
  - Amazon Kindle 1984, Itunes DRM
- Oberholzer & Strumpf showed music shared was not bad backed up by Canadian government
  - Varian in early 2005 showed DRM helps system manufacturers not music industry
  - End of the year publishers protesting against Apple

Questions?

System Security Group | Lancaster University

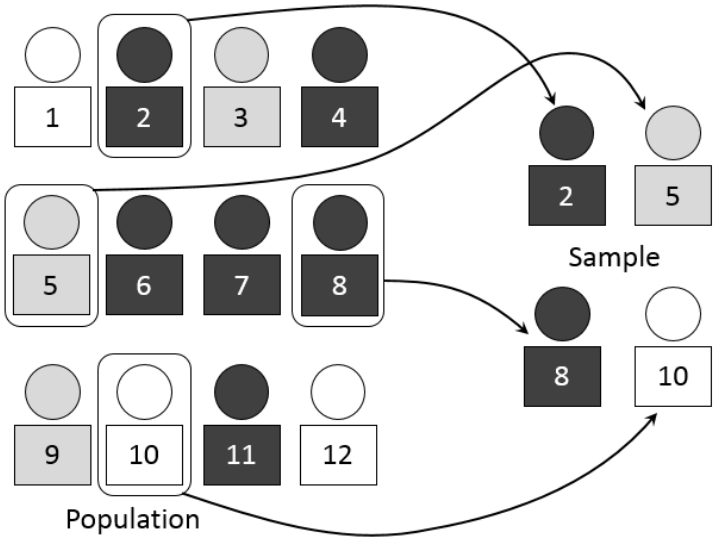
Some Numerical Methods

Topics

- Random sampling
- Linear programming
- Linear regression

Random sampling

- Random sampling is the selection of a subset of a population to estimate characteristics of the population.
- Selected samples should be representative.
- Lower costs and faster data collection than measuring the entire population and can provide insights where it is infeasible to sample an entire population.

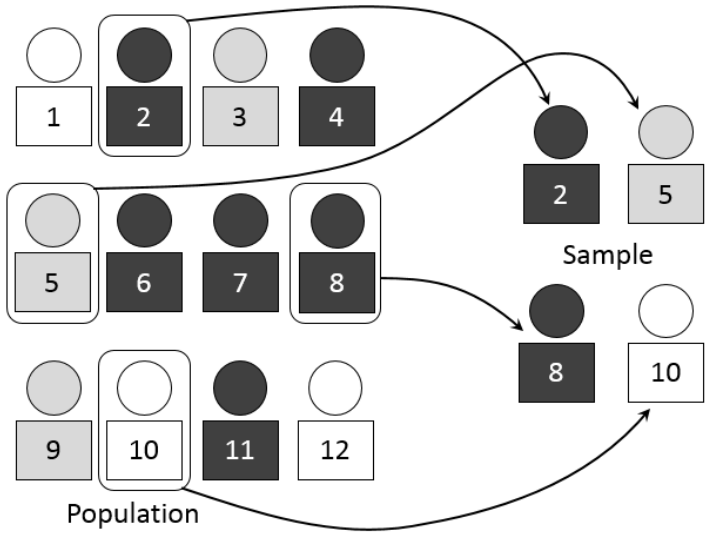


Population

- A population can be defined as including all items with the characteristics one wishes to understand.
- Sometimes necessary to sample over time, space, or some combination of these dimensions.
- The examined ‘population’ may be less tangible—often arises when seeking knowledge about the cause system of which the observed population is an outcome.
- The population from which the sample is drawn may not be the same as the population from which information is desired.

Simple random sampling

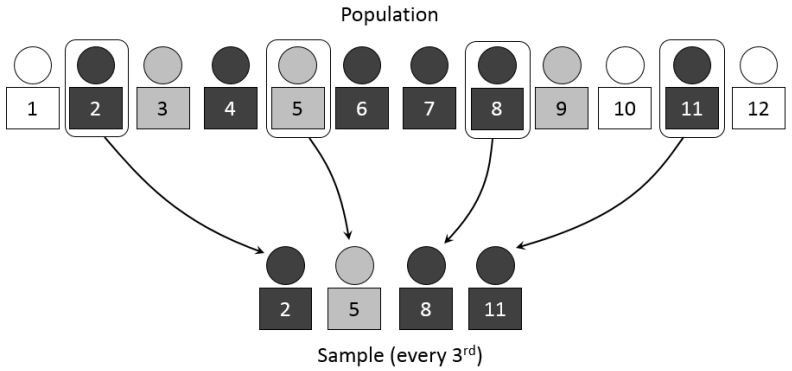
- All subsets of the same size have the same probability of being selected.
- This minimizes bias and simplifies analysis.
- Vulnerable to sampling error—selection randomness may result in a sample that doesn't reflect the makeup of the population. Skewed?
- Cannot accommodate to cases where we are interested in questions specific to subgroups of the population.



Systematic sampling

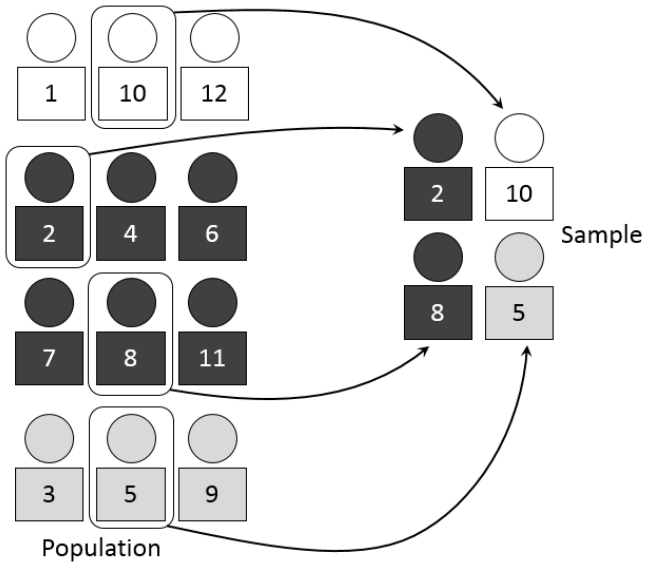
- Arrange the population by some ordering scheme. Start from a random position and proceed with selecting every *k*th element.
- It ensures that the sample is spread evenly along the list.
- Vulnerable to periodicities—unrepresentative if period is a multiple or factor of *k*.
- Difficult to quantify sampling accuracy.

removes skewness



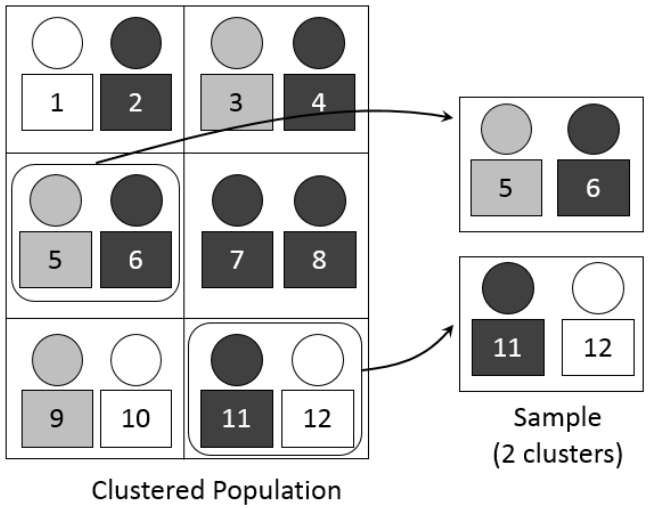
Stratified sampling

- Organize distinct categories of the population into separate 'strata', and each stratum is sampled as an independent sub-population.
- Most effective when:
  - Variability within strata are minimized;
  - Variability between strata are maximized;
  - The variables upon which the population is stratified are strongly correlated with the desired variable.
- Focus on subpopulations and ignores irrelevant ones.
- Selection of relevant stratification variables can be difficult.



Cluster sampling

- Separate the population into different clusters by geography or time, and do cluster-level sampling.
- Reduce travel and administrative costs.
- Require a larger sample than simple random sampling.



Monte Carlo Method

- Monte Carlo methods are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results.
- The underlying concept is to use randomness to solve problems that might be deterministic.
- A general pattern:
  - Define a domain of possible inputs;
  - Generate inputs randomly from a probability distribution over the domain; (min, max) in domain; normal, pareto,
  - Perform a deterministic computation on the inputs;
  - Aggregate the results.

mean, variance, etc

If iterations > 1, gather the results to get a theorical answer

Linear programming

- Linear programming (LP), also called linear optimization, is a method to achieve the best outcome in a mathematical model whose requirements are represented by linear relationships.

- More specifically, LP is a technique for the optimization of a linear objective function, subject to linear equality and inequality constraints.

LP formulation

- Canonical form:

$$\max_{x} \; c^T x$$
$$\text{s.t. } Ax \leq b$$

- max=maximize, s.t.=subject to, x and c are both n-dimensional column vectors, A is an $m \times n$ matrix, and b is an m-dimensional column vector.
- c, A and b are problem parameters, which are given and fixed, x is called the decision variable, $c^T x$ is the objective function, and $Ax \leq b$ are the constraints.
- The purpose is to find a vector x^* such that $Ax^* \leq b$ (feasibility), and $c^T x^* \geq c^T x$ for all x such that $Ax \leq b$ (optimality).

Example

A company makes two products X and Y using two machines P and Q. Each unit of X needs 50 minutes on P and 30 minutes on Q. Each unit of Y needs 24 minutes on P and 33 minutes on Q.

At the start of the current week, there are 30 units of X and 90 units of Y in stock. Available processing time on P is 40 hours and on Q is 35 hours.

The demand for X in the current week is 75 units and for Y is 95 units. The company aims to maximize the combined sum of the units of X and Y in stock at the end of the week.

Question: Formulate the problem of deciding how many of each product to make in the current week as an LP.

Solution

- Let x and y be the number of units of X and Y to be produced in the current week, respectively.
- Constraints:
  - Machine P time: $50x + 24y \leq 40 \times 60$
  - Machine Q time: $30x + 33y \leq 35 \times 60$
  - Product X demand: $x + 30 \geq 75$
  - Product Y demand: $y + 90 \geq 95$
- Objective: maximize $(x + 30 - 75) + (y + 90 - 95)$
  - Effectively, maximize $(x + y)$

In the unit of minutes

Solution

Direct formulation:

$$\max_{x,\,y} \; (x + y)$$

$$\text{s.t. } 50x + 24y \leq 2400$$

$$30x + 33y \leq 2100$$

$$x \geq 45$$

$$y \geq 5$$

difference of current vs demand

Canonical form:

$$\max_{x,\,y} \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

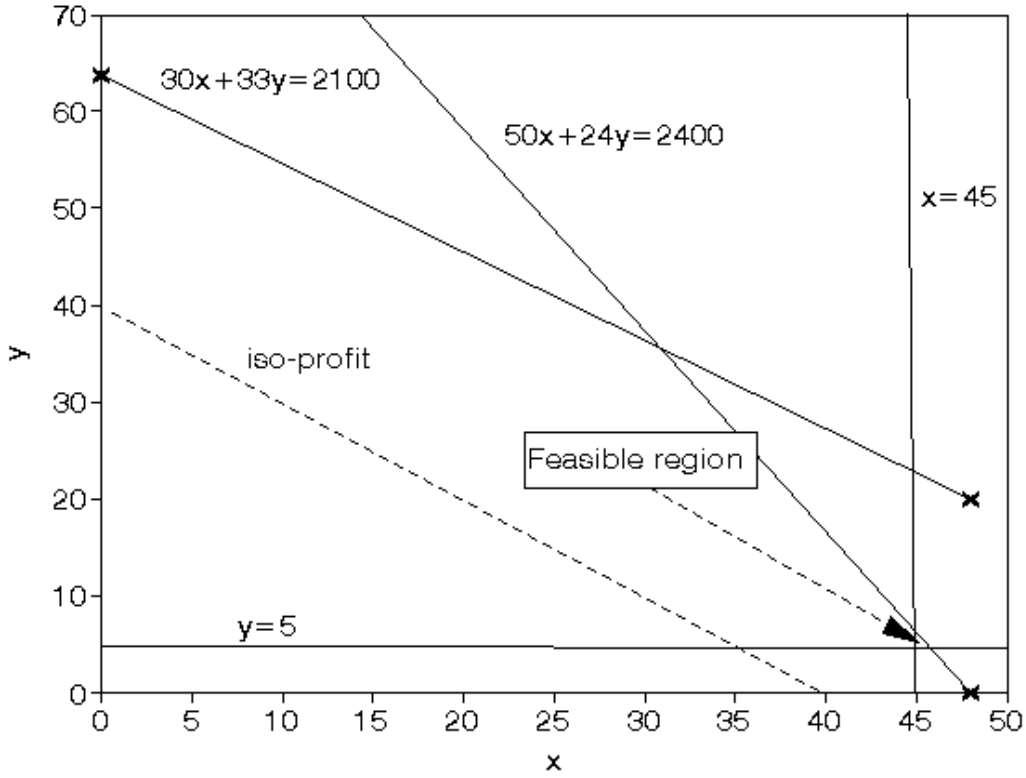
$$\text{s.t. } \begin{bmatrix} 50 & 24 \\ 30 & 33 \\ -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \leq \begin{bmatrix} 2400 \\ 2100 \\ -45 \\ -5 \end{bmatrix}$$

Need for high variety in terms of less than/equal to direction

$$c = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, A = \begin{bmatrix} 50 & 24 \\ 30 & 33 \\ -1 & 0 \\ 0 & -1 \end{bmatrix}, b = \begin{bmatrix} 2400 \\ 2100 \\ -45 \\ -5 \end{bmatrix}.$$

Solving the LP graphically

- The maximum occurs at the intersection of $x = 45$ and $50x + 24y = 2400$.
- $x = 45$ and $y = 6.25$.
- If x and y must take integer values, then we take $x = 45$ and $y = 6$.



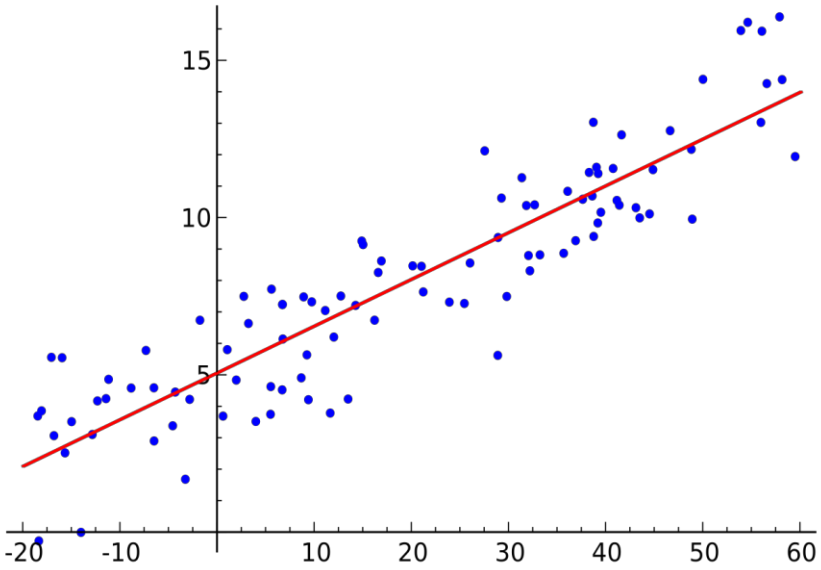
Solving methods for LP

- Simplex algorithms
- Interior point methods
- Details of these methods are out of our scope
- In Python, these methods have been implemented in the library *scipy.optimize*, see the function *linprog*.
- Be careful with the syntax of the function!

Linear regression

- Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables.

- In linear regression, the relationships are modelled using linear predictor functions whose unknown model parameters are estimated from the data.



Formulation

- Given a data set $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^{n}$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors $\mathbf{x}$ is linear.
- The relationship is modelled via a disturbance term ε that adds noise to the linear relationship between the dependent variable and regressors.
- The model takes the form $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$ for each $i = 1, \ldots, n$.

not an ideal relationship

Solving for the weights

- Linear regression models are often fitted using the least squares approach. That is, we aim to find the weights $\beta_0, \beta_1, \dots, \beta_p$ that minimize certain norm of the absolute deviations, e.g., $\sum_{i=1}^{n} (\varepsilon_i)^2$.
- Can be formulated as an optimization problem:

$$\min_{\beta_0,\beta_1,\dots,\beta_n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2$$

- In Python, methods for solving the fitting problem have been implemented in the library *scipy.optimize*, see the function *curve_fit*.
- Again, be careful with the syntax of the function!

Applications

- Businesses often use LR to understand the relationship between advertising spending and revenue:
  $revenue = \beta_0 + \beta_1(£ad1) + \cdots + \beta_p(£ad1).$
  - What does $\beta_i > 0$, $\beta_i < 0$ or $\beta_i \approx 0$ indicate?
    making money losing money no profit
- Medical researchers often use LR to understand the relationship between drug dosage and blood pressure:
  $blood\ pressure = \beta_0 + \beta_1(dosage).$
- Agricultural scientists often use LR to measure the effect of fertilizer and water on crop yields:
  $crop\ yield = \beta_0 + \beta_1(amount\ of\ fertilizer) + \beta_2(amount\ of\ water).$

Questions?