Laboratorio di web scraping
AA. 2023-2024

docente: Laura Ricci

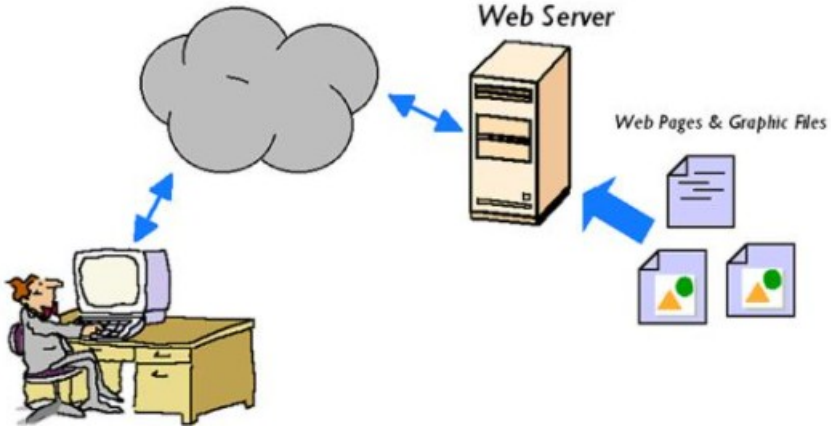
laura.ricci@unipi.it

Lezione 16
The Bitcoin ecosystsm: scraping and analysis

5/04/2024

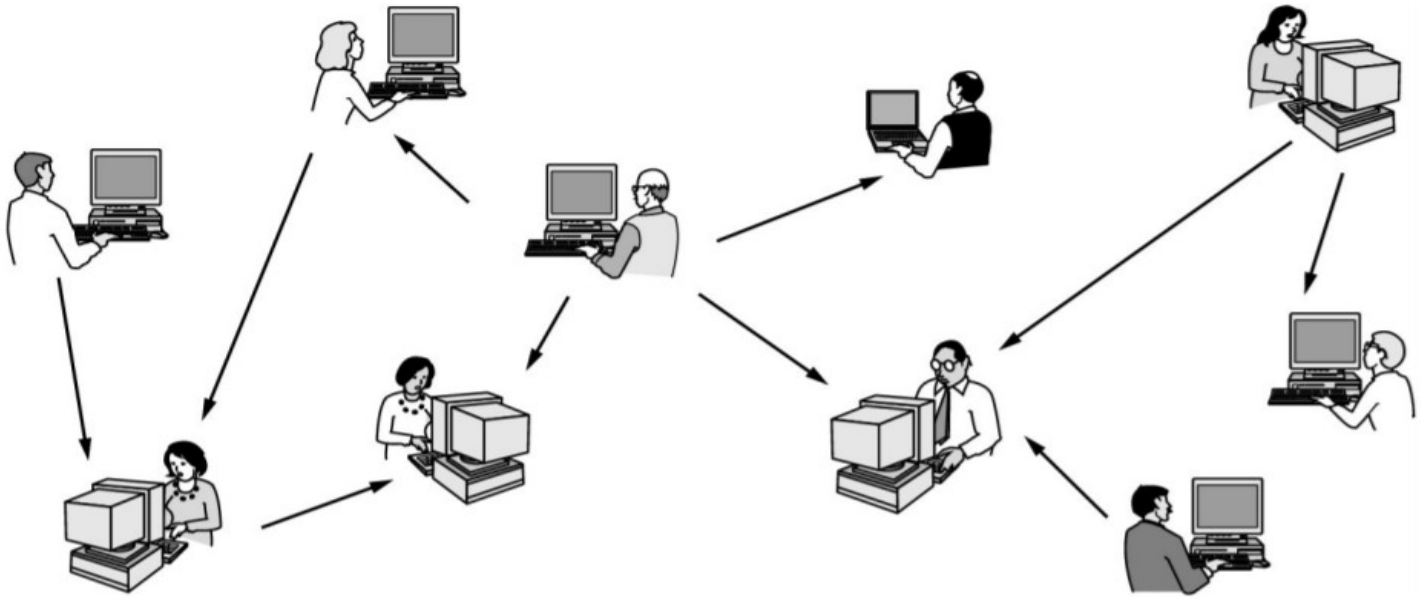
THE CLIENT SERVER MODEL FOR THE WEB

- the web client is your browser
  - makes an `HTTP` request to a specific web server
- the web server receives the request
  - sends back the requested document to the client
  - an `HTML` page, possibly with CCS and `JavaScript`
- the web client interprets the information returned by the server and displays it appropriately

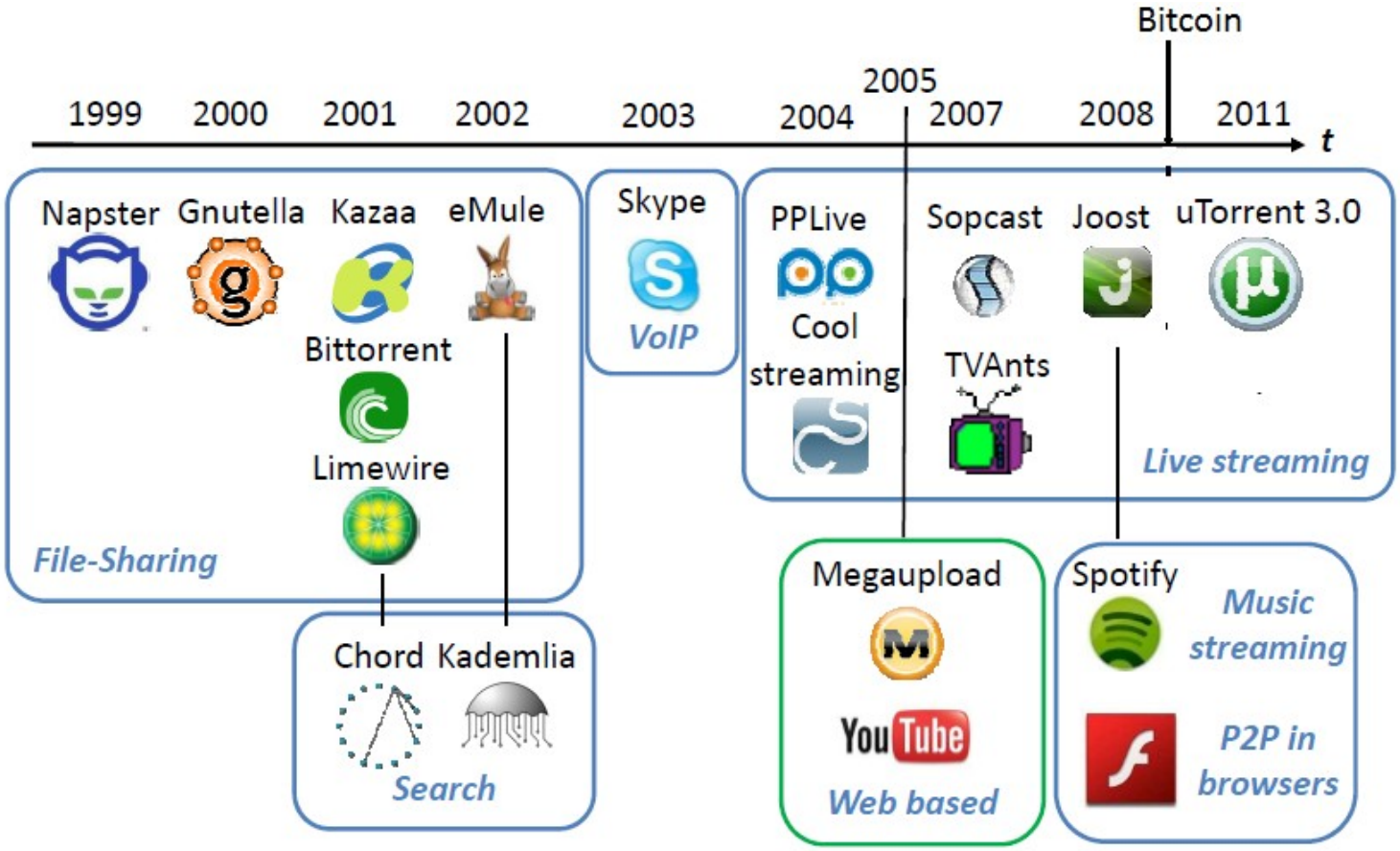


A CHANGE OF PARADIGM: PEER TO PEER

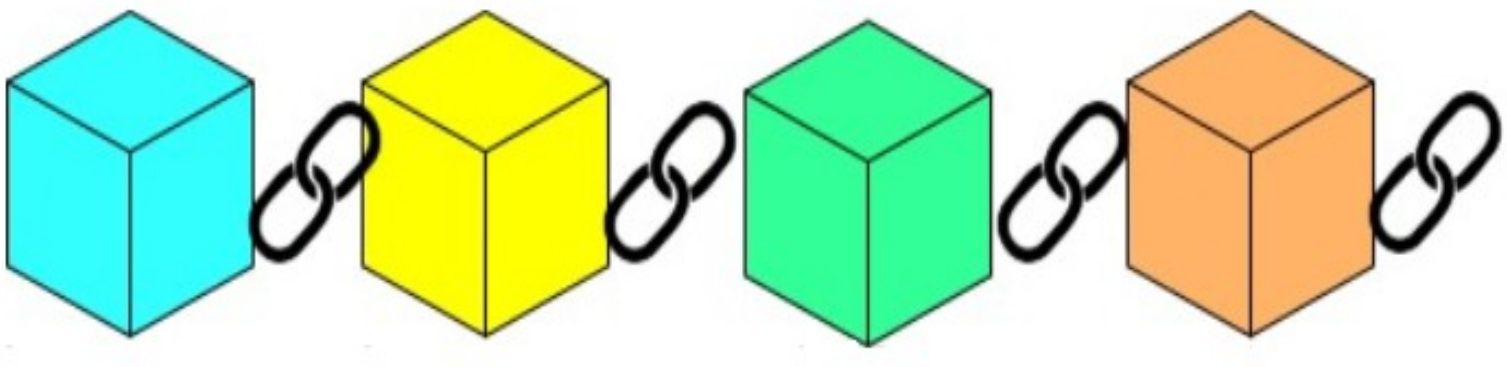
- by the turn of the twenty-first century, new models for delivering online services emerged.
- instead of relying on a centralized server, parties began experimenting with peer-to- peer (P2P) networks
  - a networks consisting of nodes (peers) working together based on equals rights/functionalities



PEER TO PEER: FROM FILE SHARING TO BITCOIN

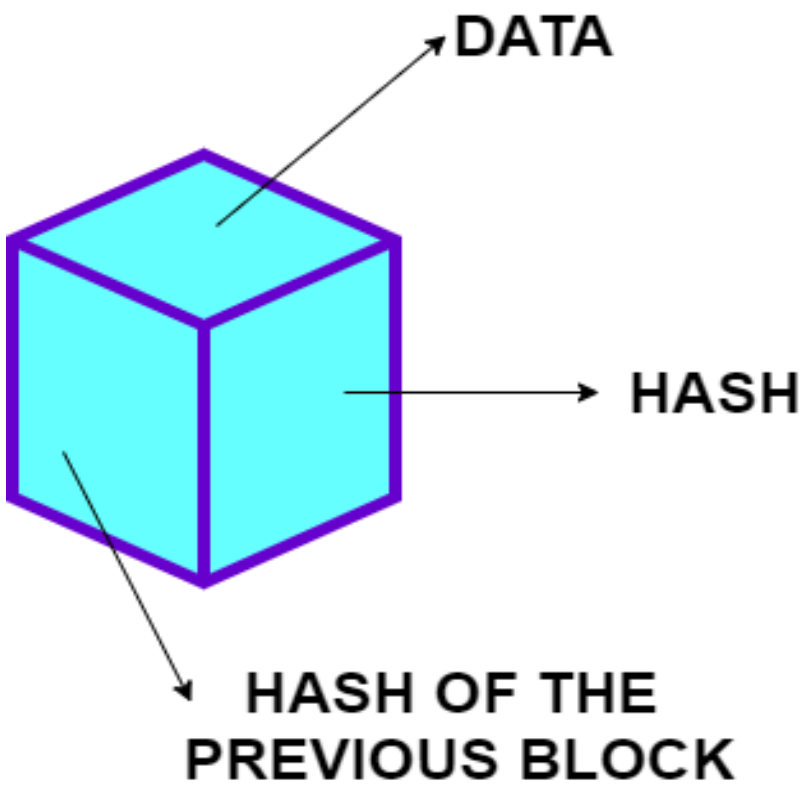


BLOCKCHAIN "AT A GLANCE"

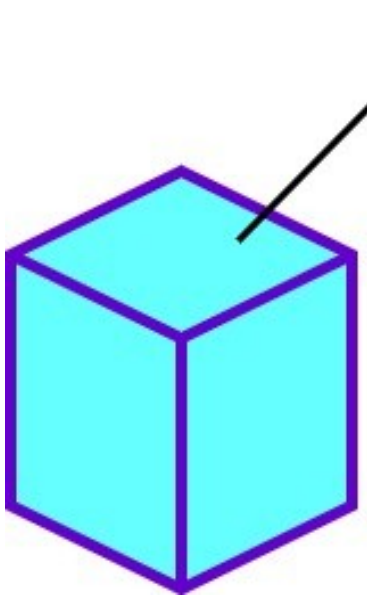


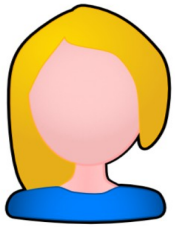
- the 'killer P2P application'
- a distributed ledger which is replicated between the nodes of a peer-to-peer network
  - like a "notary"
  - characterized by the tamper freeness property

LOOKING INSIDE A BLOCK



WHICH KIND OF DATA CAN WE FIND IN A BLOCK?



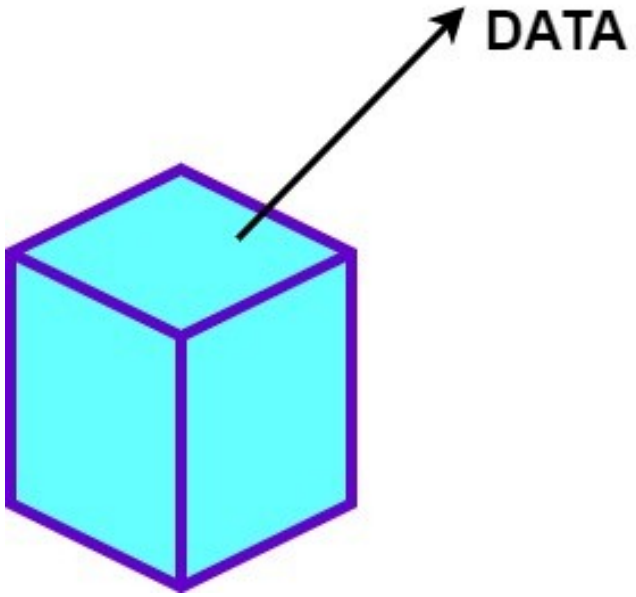


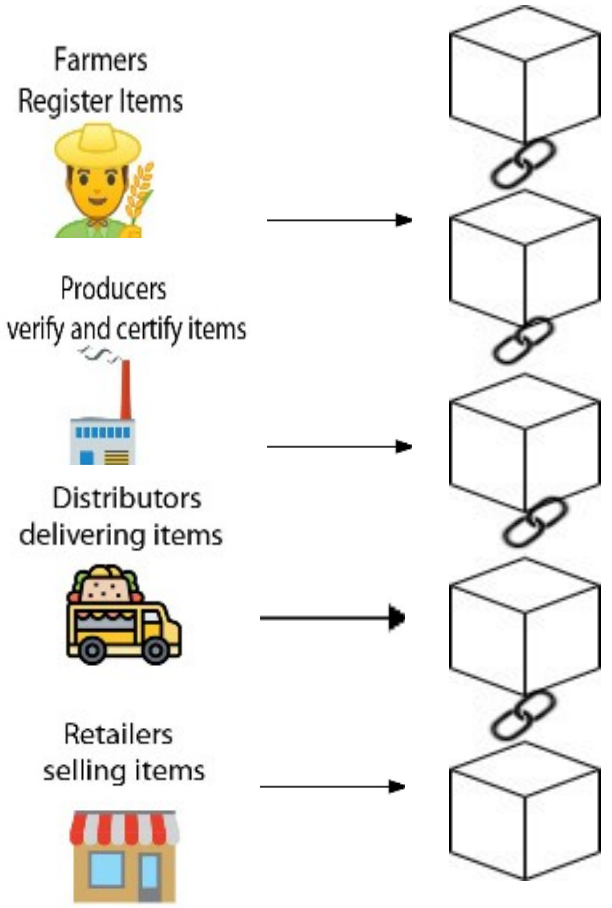
FROM

TO

AMOUNT

WHICH KIND OF DATA CAN WE FIND IN A BLOCK?

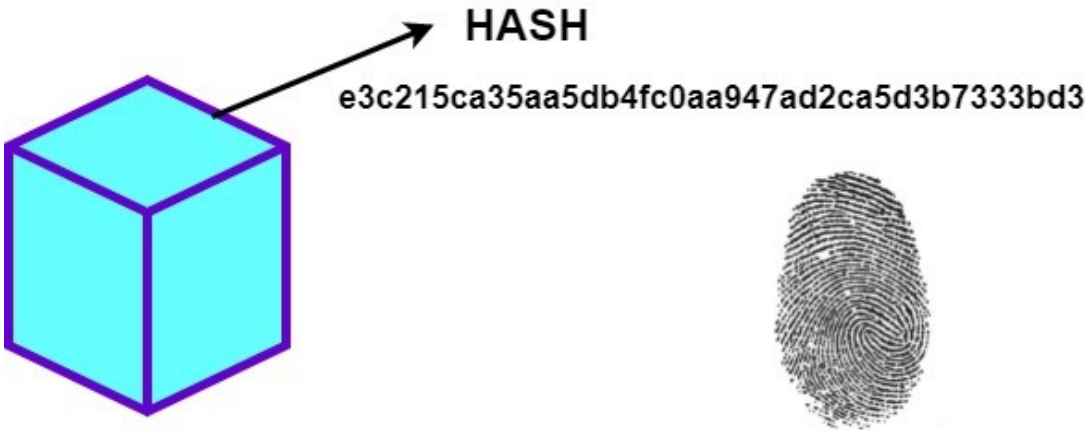


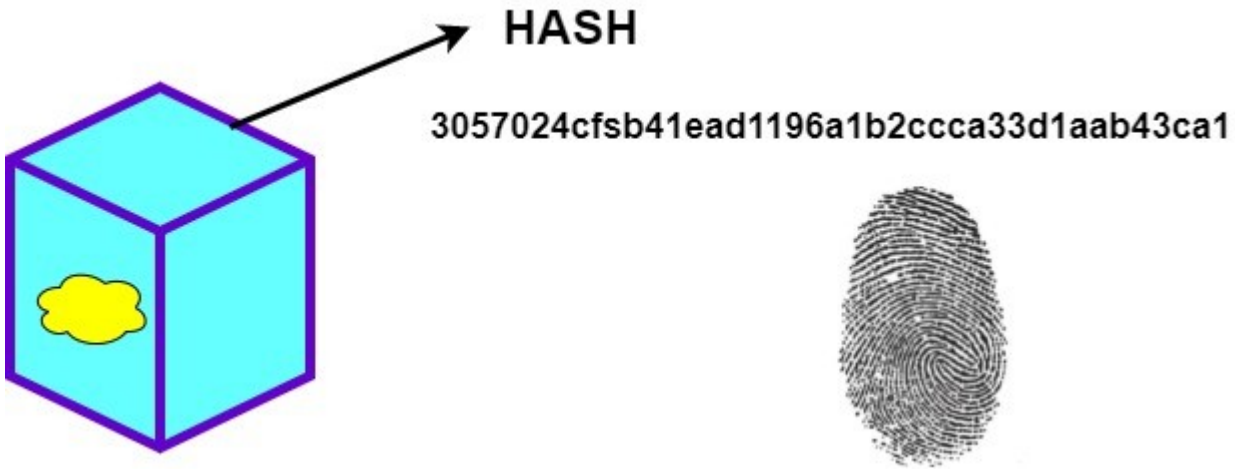


WHICH KIND OF DATA CAN WE FIND IN A BLOCK?

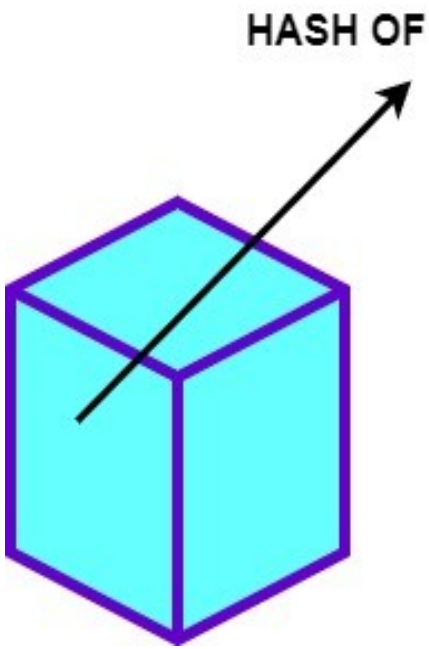
- in general, transactions
  - cryptocurrencies transactions (`Bitcoin`, `Ethereum`,...) or finantial assets
- but not only!
  - sensor measurements `IoT`, `ECG`.,...
  - supply chain (e.g.: diamonds,....)
  - asset certification (`NFT`)
  - intellectual property (audio/video content)
  - health-care contents
- all data are public
  - ideal scenario for this course: lots of free data to analyse!

WHICH KIND OF DATA CAN WE FIND IN A BLOCK?



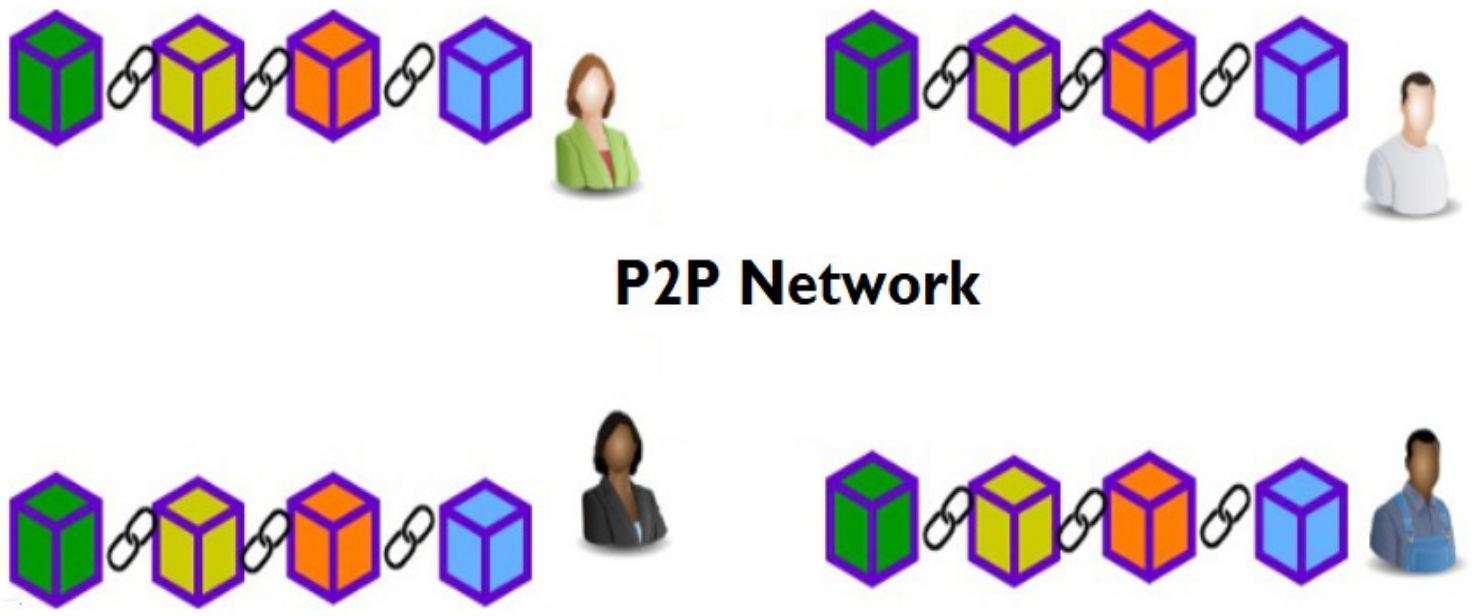


WHICH KIND OF DATA CAN WE FIND IN A BLOCK?



Hash pointers create the chain

THE DISTIBUTE LEDGER

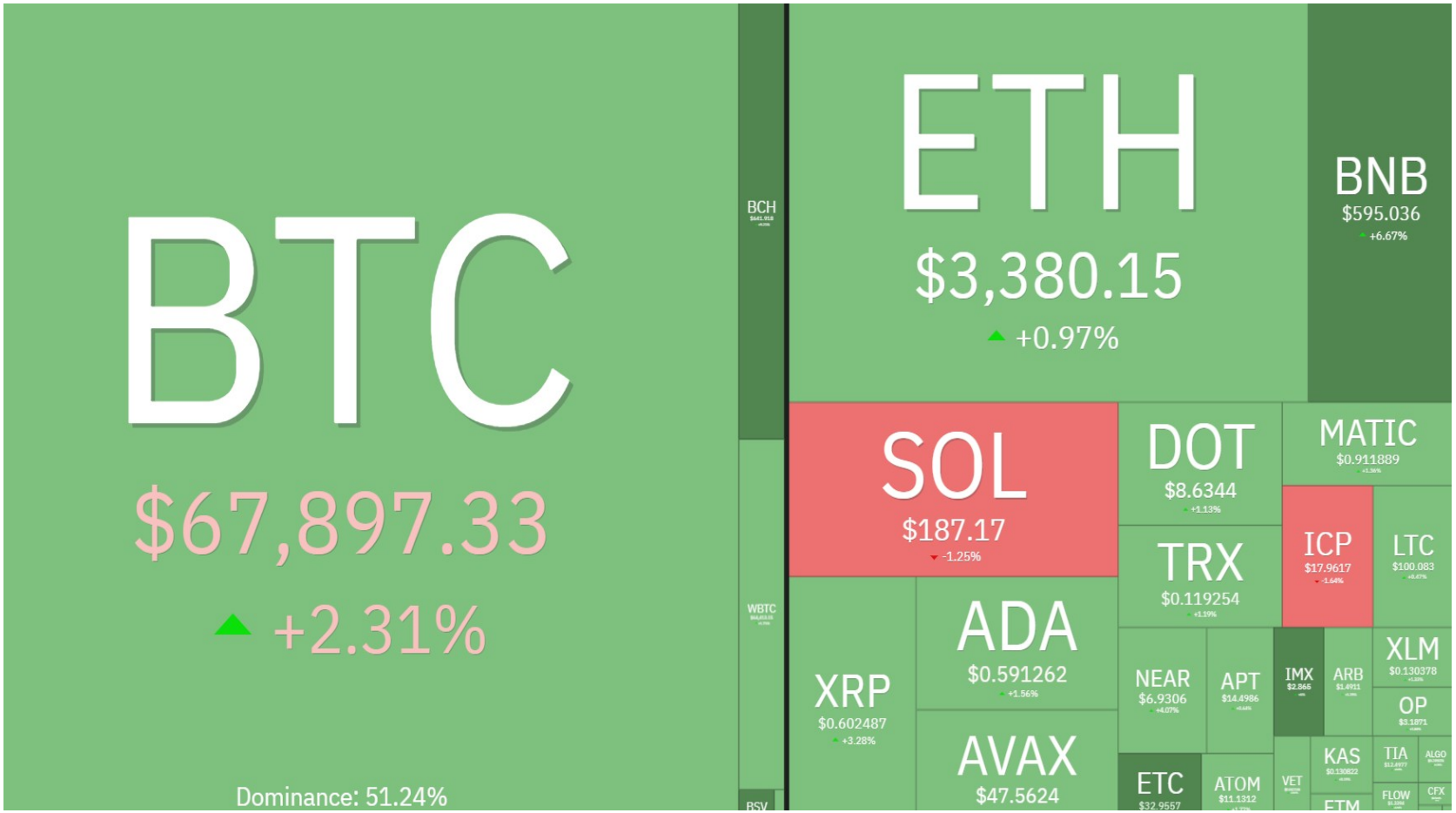


- distributed consensus to decide which block has to be added to the blockchain
- not a topic of this course

BITCOIN: WHAT IS IT?

- the "killer application" for blockchain
- a cryptocurrency invented by `Satoshi Nakamoto`, started in `2009`
- a decentralized digital currency
  - no central authority (no central bank or state issue or guarantee the currency)
  - cryptographic methods guarantee that no-one is cheating
    - issuing their own coins
    - stealing coins
    - double spending
    - etc….
- the protocol is still evolving, the official `Bitcoin` core is a `GitHub` repository, on which anyone can propose contributions (https://github.com/bitcoin/bitcoin)
  - goal: more efficient, faster, more secure, more anonymous,…

TONS OF BLOCKCHAINS: WHY FOCUS ON BITCOIN?



BITCOIN: AN UTXO BASED BLOCKCHAIN

- why data analysis mainly from the Bitcoin blockchain?
  - has by far the highest capitalization
  - many other cryptocurrencies are based on similar models
  - a class of blockchains: UTXO based blockchain
- the structure of Bitcoin transactions is based on a model called UXTO
  - UTXO: Unspent Transaction Output
  - why focusing on this model?
    - because it is adopted by many other cryptocurrencies (Litecoin, Solana,Cardano, ...)
- Ethereum uses a different model: account-based model
  - different transaction format, we will see it, if there will be time

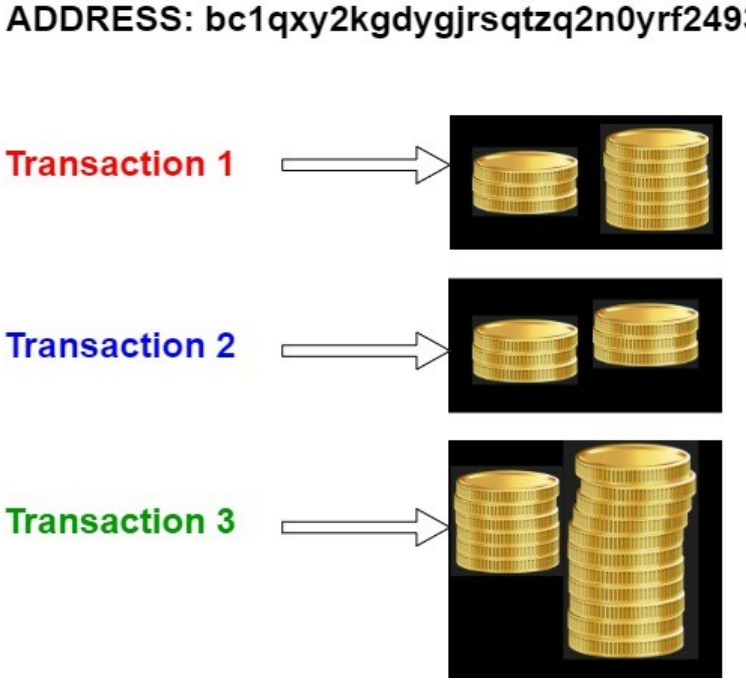
THE STRUCTURE OF A BITCOIN

- a bank transaction

FROM	TO	AMOUNT
account1	account2	10

- a transfer of money from an account to another account
- Bitcoin address is, in some way, is similar to an account number that holds bitcoin, but there is an important difference
  - it keeps separate the bitcoin received by different transactions
  - like having a different money box for each pile of bitcoin received
  - an address is a container for several bitcoin piles
  - users may have, in their wallet, several addresses

KEEP EACH PILE OF BITCOIN SEPARATE

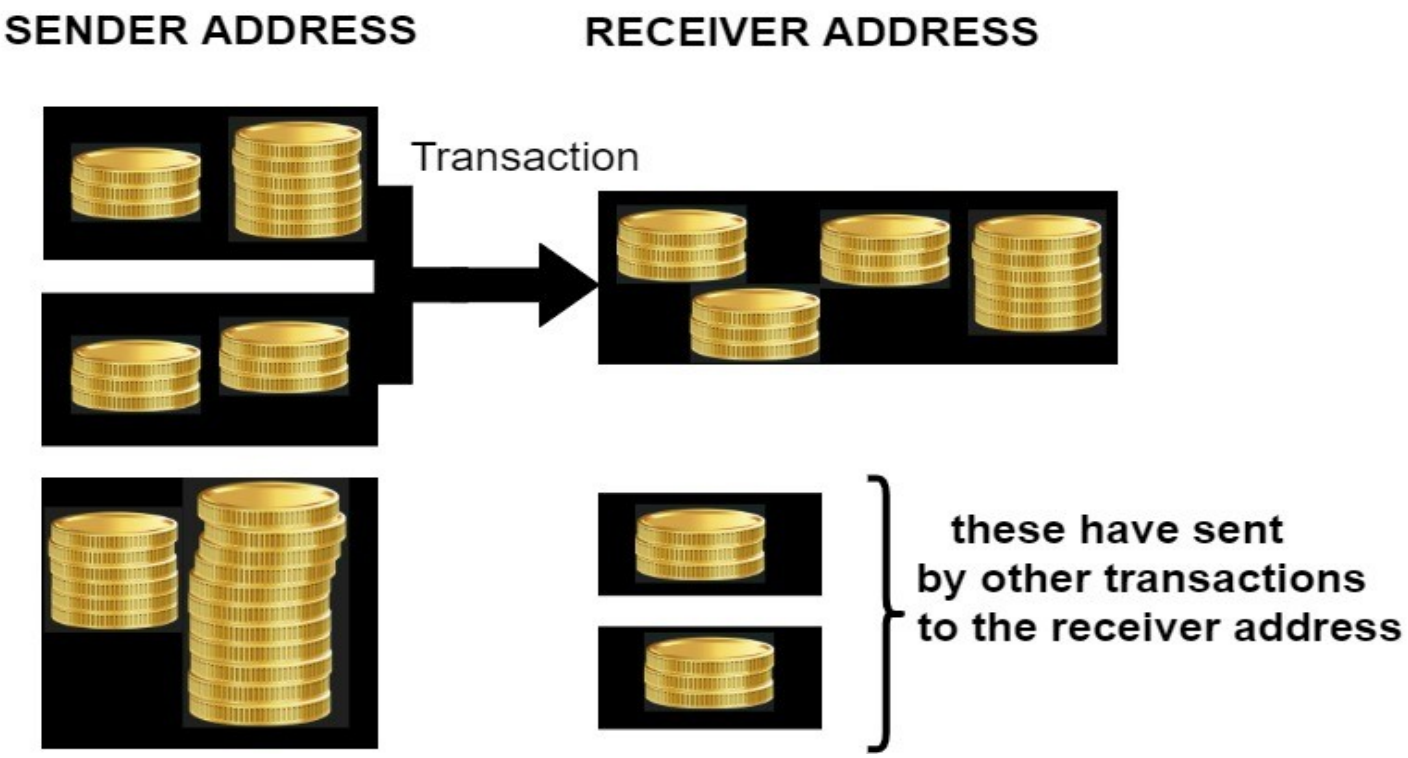


- different coin piles for `bitcoin` received from different transactions
- in a shop, each pile corresponds to the `bitcoin` received from a different sale
- many piles for the same address or also different addresses with different piles

THE UXTO MODEL

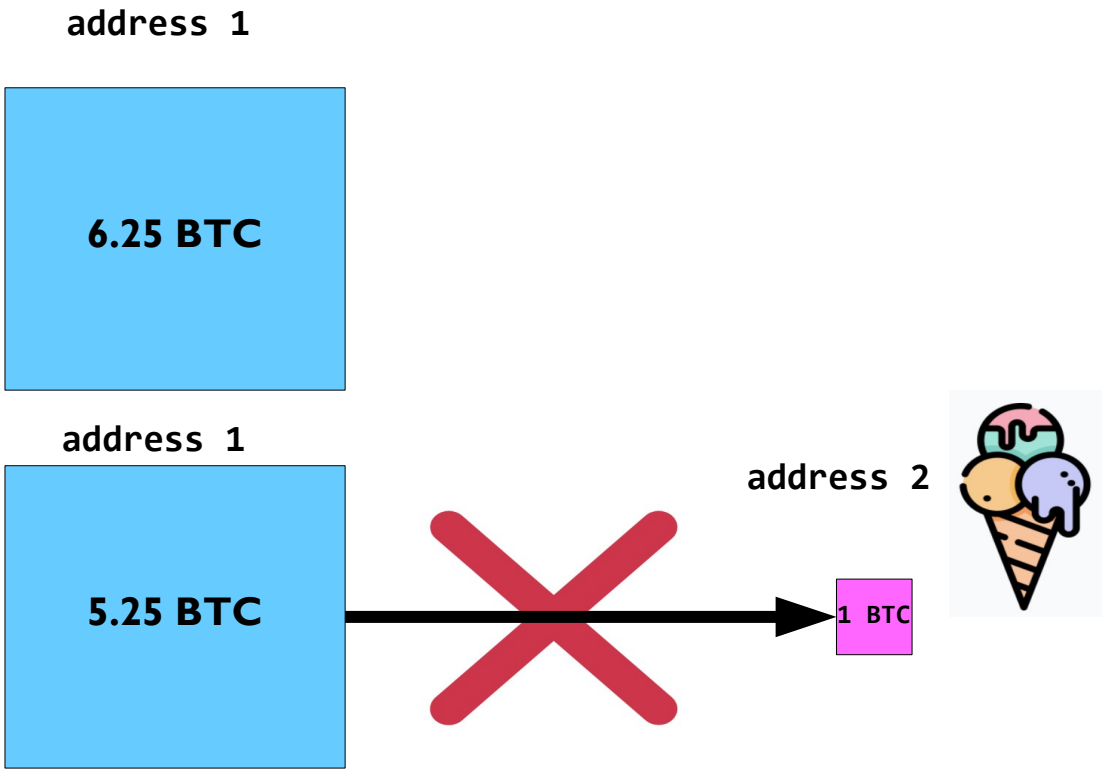
- when the user want to pay something, creates a transaction
  - uses a set of Bitcoin piles to generate a new pile to send to the receiver.
  - takes the whole amount from one or more of the stacks and sends it to the address of the receiver
  - no coin can remain in a pile
  - the receiver maintains the received bitcoin pile separated from the other one in his/her addresses
- this model is called UTXO model (Unspent Transaction Output)
  - unspent bitcoin piles are spread in Bitcoin transaction which are spread on the blockchain
  - it seems a bit complex, but it has several technological advantages (anonymity, parallelism,.....)
  - adopted by several blockchains: Bitcoin, Litecoin, Cardano,....

WHAT IS A BITCOIN TRANSACTION?



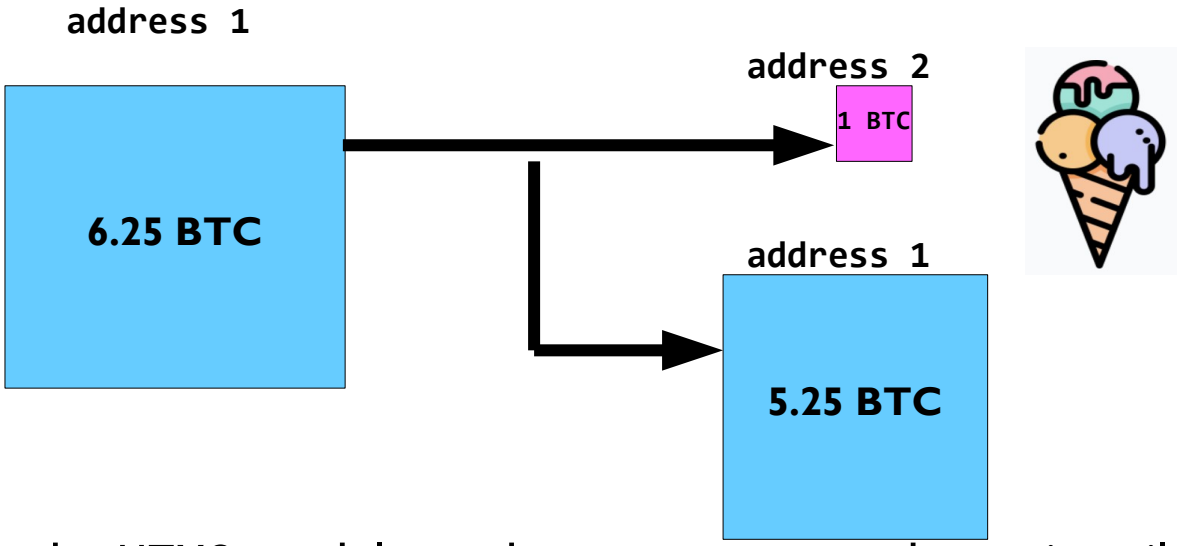
- when the user wants to pay something, creates a transaction
  - takes 2 batches of `Bitcoin` from the same address or from different addresses and send to the receiver
- the received `Bitcoin` are a separate pile in the receiver address

TRANSACTION EXAMPLES



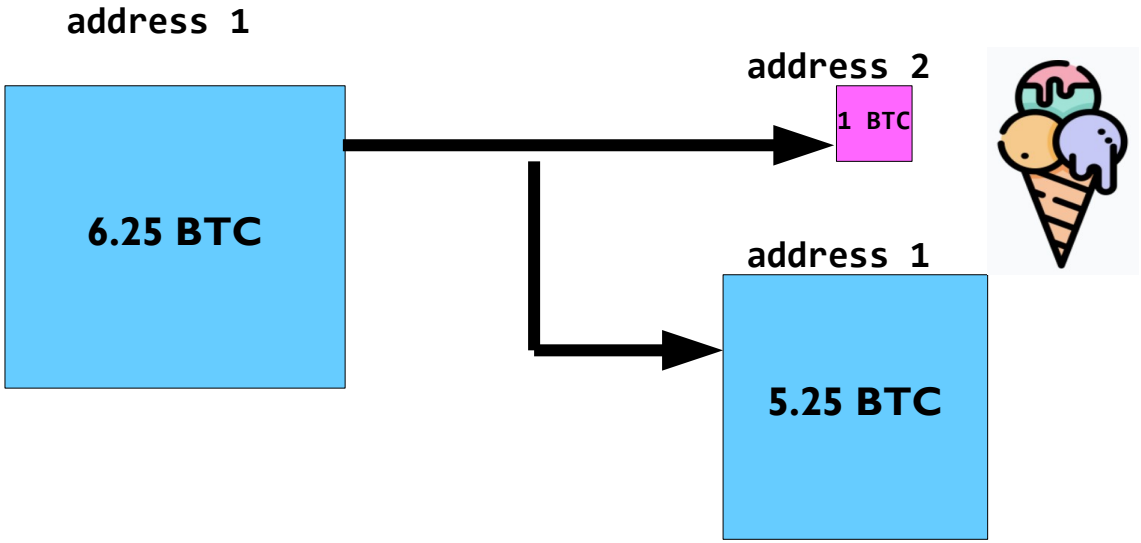
- what can I do with these `bitcoin`? Buy an ice cream cone!
- use one of these bitcoin to pay the ice cream (it's a joke, it would be a very expensive beer!)
- but this does not work with the UTXO model, because all the bitcoin in the pile must be spent

TRANSACTION EXAMPLES



- in the UTXO model, you have to consume the entire pile of `6.25 BTC`:
- can leave a change in the same address but this is a new pile in a new transaction output (but in general you use a different address)
- spit the pile of `bitcoin` up and send it to two destinations
  - the ice cream shop (the payment, in pink)
  - back to our own address (the change, a new pile, in blue)
- the original batch of `6.25` bitcoin has now been "used up", does not exist any more

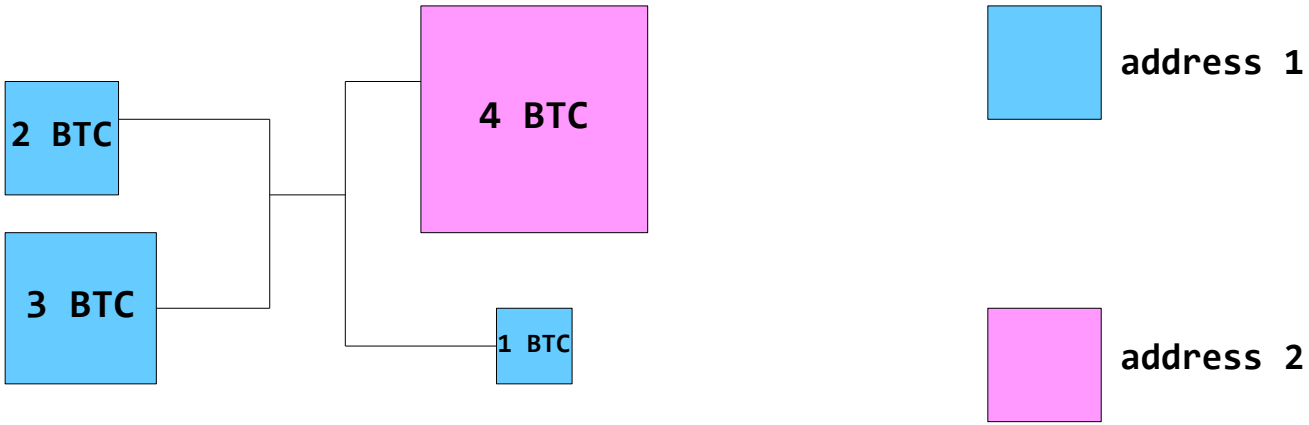
TRANSACTION EXAMPLES



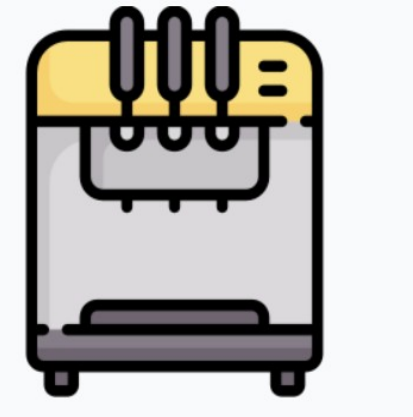
- this is what the `Bitcoin` transaction system is designed to do
  - take an existing transaction output (a pile of bitcoins)
  - this is a transaction through which you have received bictoin
  - create newly-sized outputs (batches) from it
  - send those outputs to different addresses

GENERIC STRUCTURE OF A TRANSACTION

- generally the total of the bitcoin's pile may add up to more than the user want to spend
  - in this case just add another output to the transaction and send the difference back to this output
    - like a change
  - generate a new money box for the change
  - the procecedure is performed automatically by your wallet software



TRANSACTION EXAMPLES



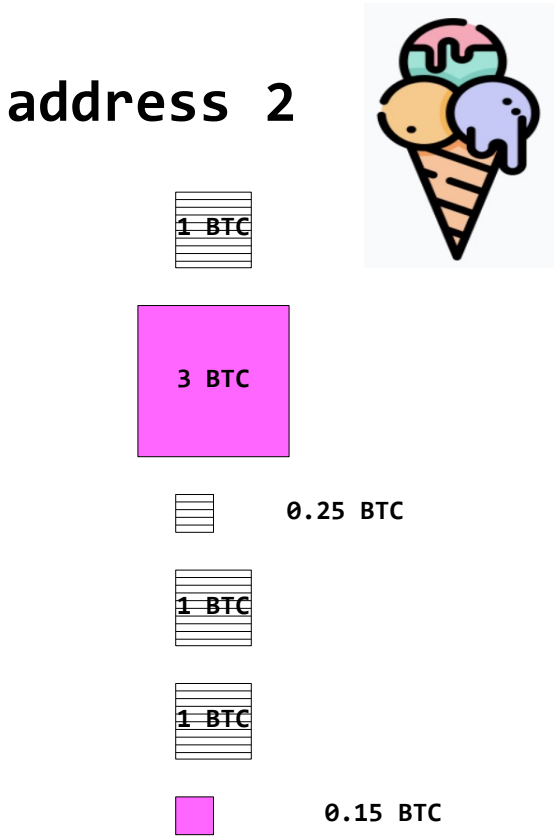
- after a while, the ice cream shop has received a lot of payments
  - the ice cream business is booming!
- the shop decides to buy a new ice cream machine for `3.10 BTC`

TRANSACTION EXAMPLES

the shop decides to buy a new ice cream machine for `3.10 BTC`

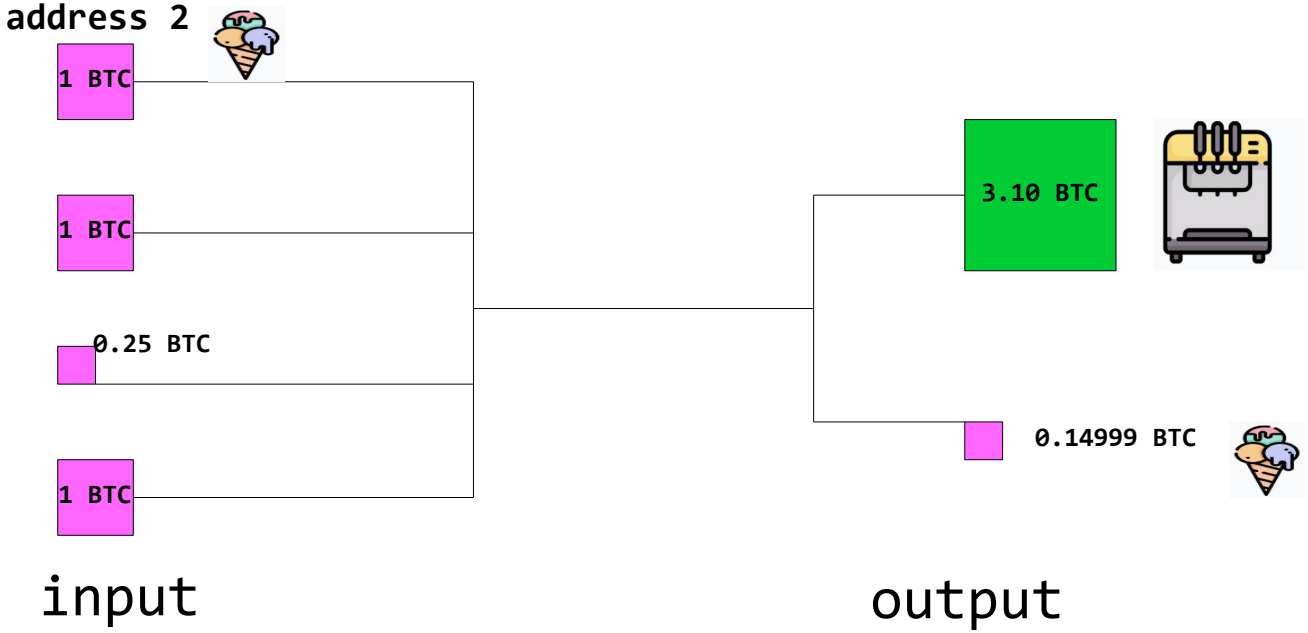
- the ice cream shop does not have a single pile at its address to cover the cost of the machine
- it gathers a handful of outputs to have a sum > `3.10`
- the output gathered from previous batches (transaction outputs) are the inputs for this transaction

STATE OF THE BATCHES AFTER THE TRANSACTION



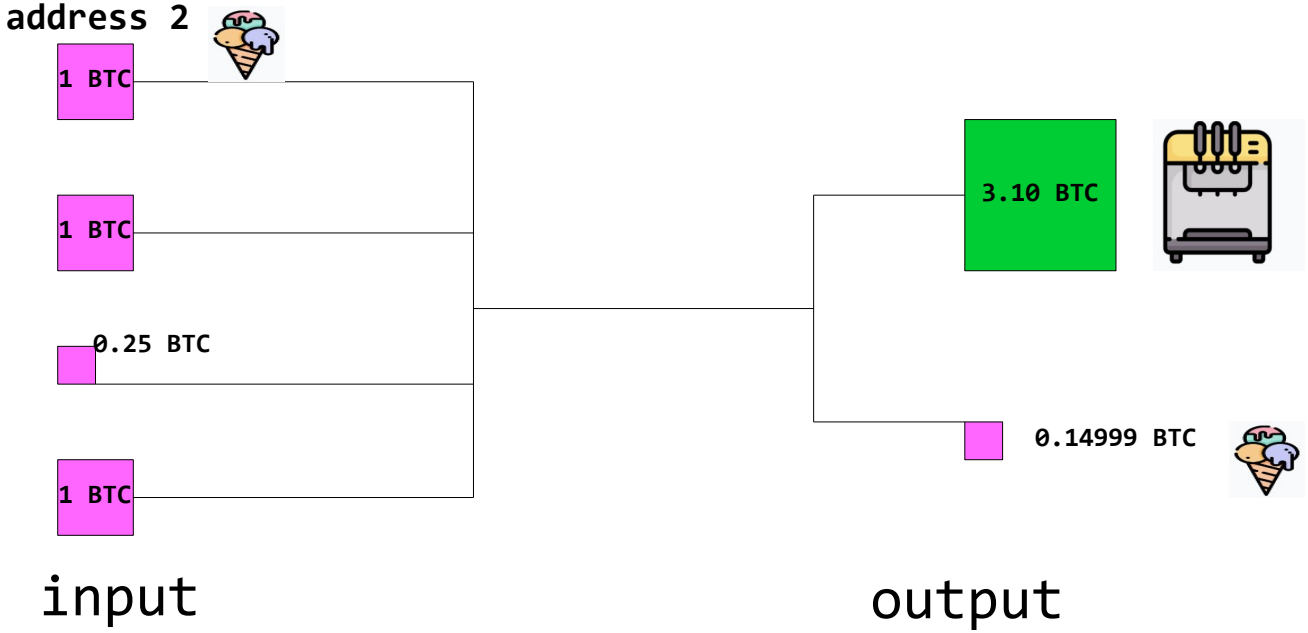
- the "unspent output" (purple squares) are still good for spending
  - these are called Unspent Transaction Outputs (UTXO)
- the total number of bitcoins of an address is the sum of the address's UTXO

TRANSACTION FEES



- observe that in this transaction the total of the outputs is less than the total of the inputs
  - 3.25 bitcoin as input 3.249999 the output
- there are some remaining bitcoins that aren't being used up.
  - this "left over" amount is the transaction fee.

TRANSACTION FEES



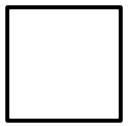
- the transaction fees are picked up by miners when they mine a block
  - adding a transaction fee basically is an incentive for miners to include your transaction in a block: gives your transaction priority.
- without a transaction fee, a transactions will probably take a while to get included in to a block.

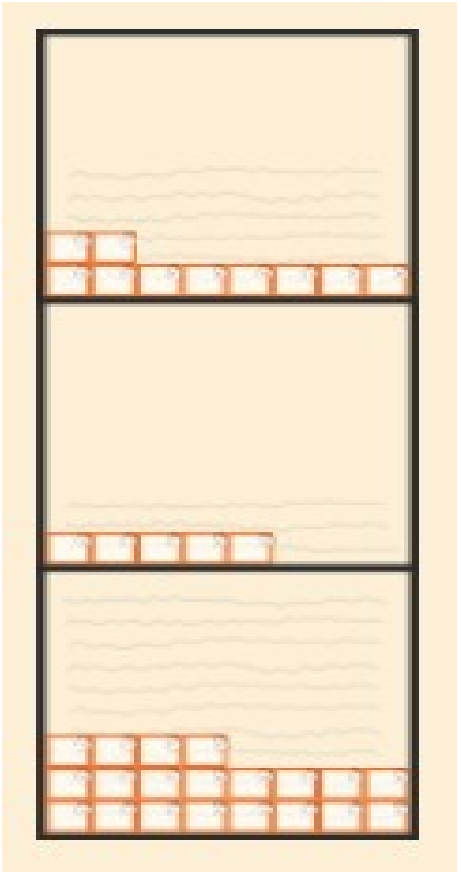
TRANSACTION EXAMPLES: COINBSE

- Coinbase transactions
  - transactions minting bitcoins
  - send "new bitcoin" to miners solving the Proof of Work, as a reward

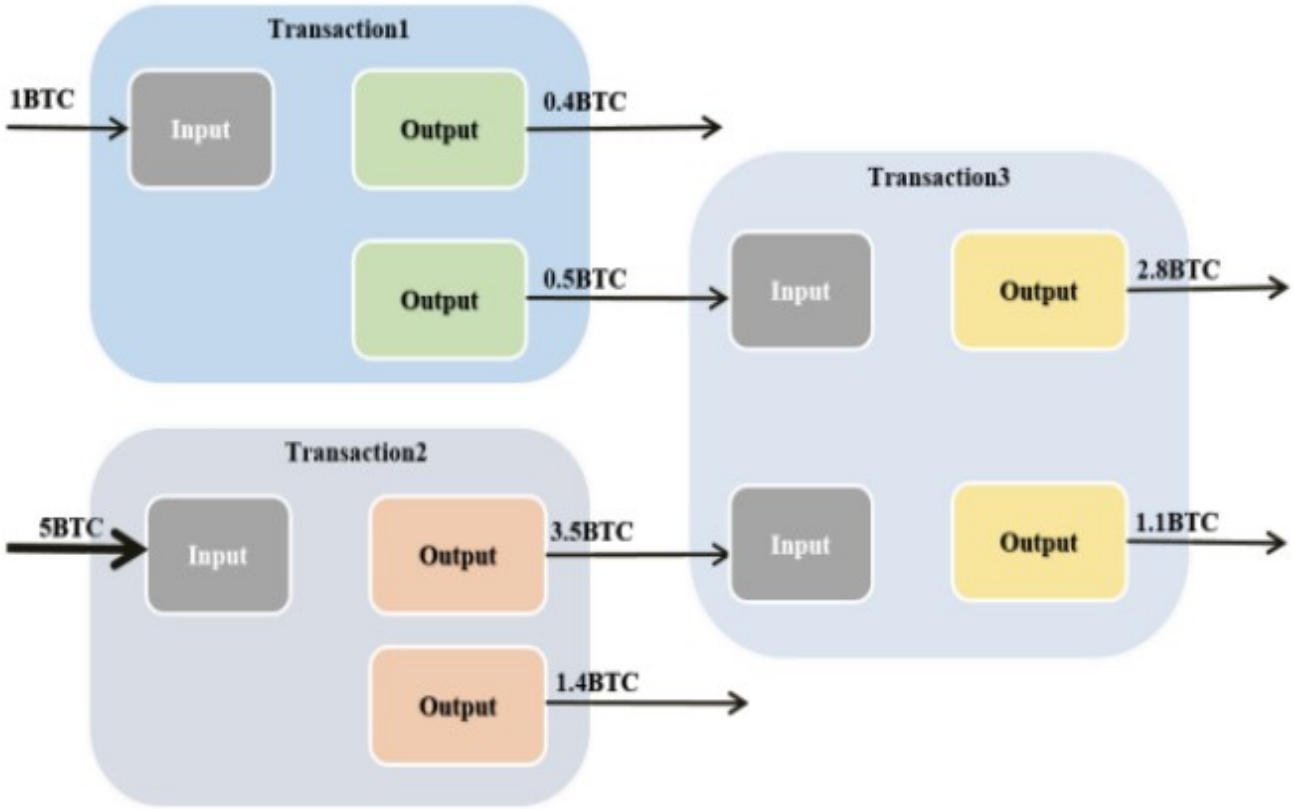
- no input to this transaction
- Bitcoin are not taken from a previous pile, they are generated by the system!

THE UXTO MODEL

- transactions are stored on the blockchain
- output of transaction are like slot on the blockchain storing "pile of bitcoin" to be spent
- from now on, we will represent a transaction output with 

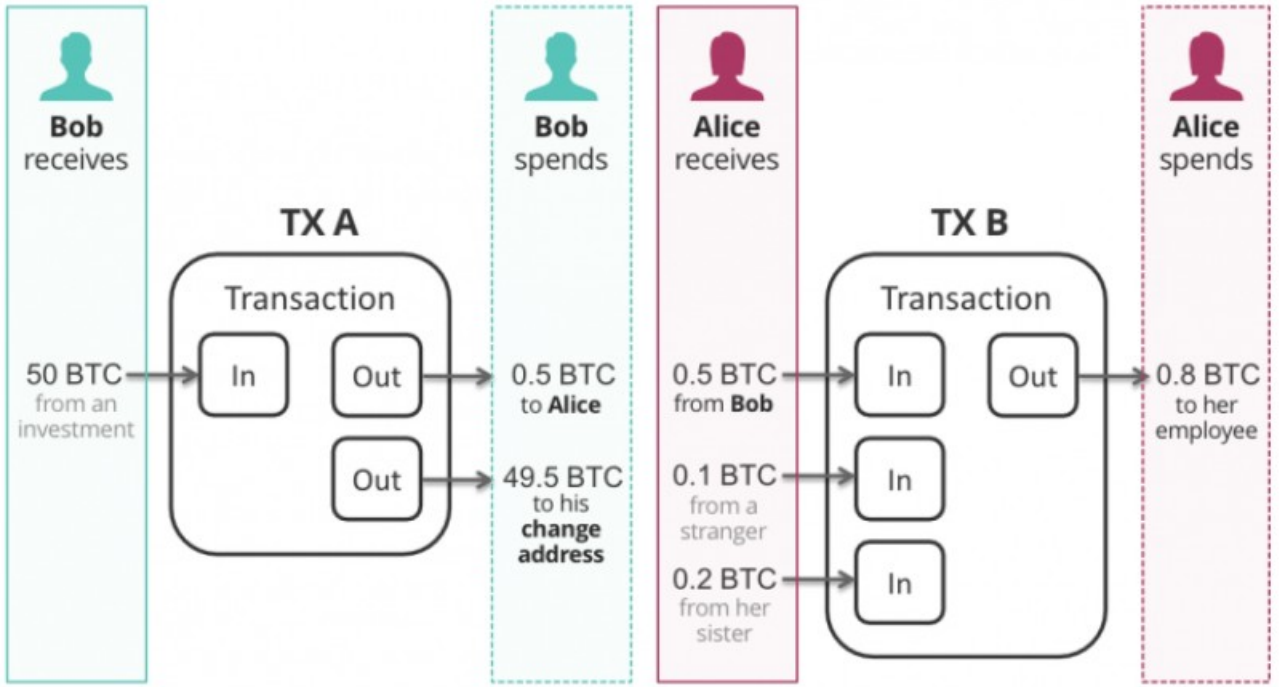


THE FORMAT OF THE TRANSACTIONS



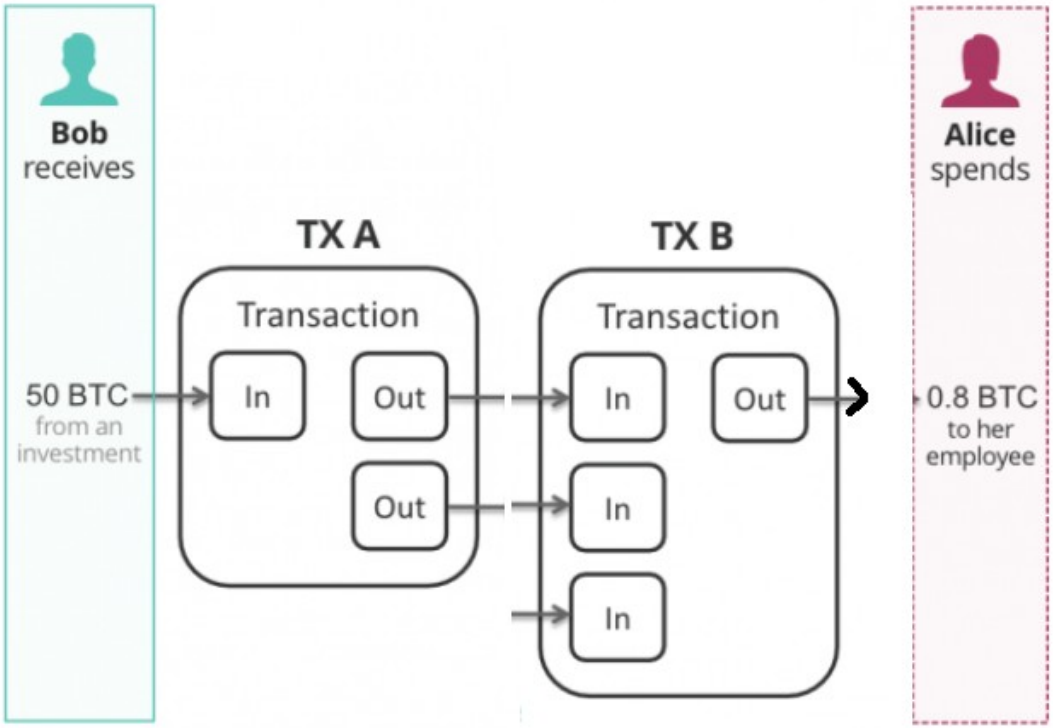
- but where are really these stack of `bitcoins`?
- actually, they are stored in the output of the transactions, which are registered on the blocks of the blockchain

BITCOIN TRANSACTIONS CHAIN



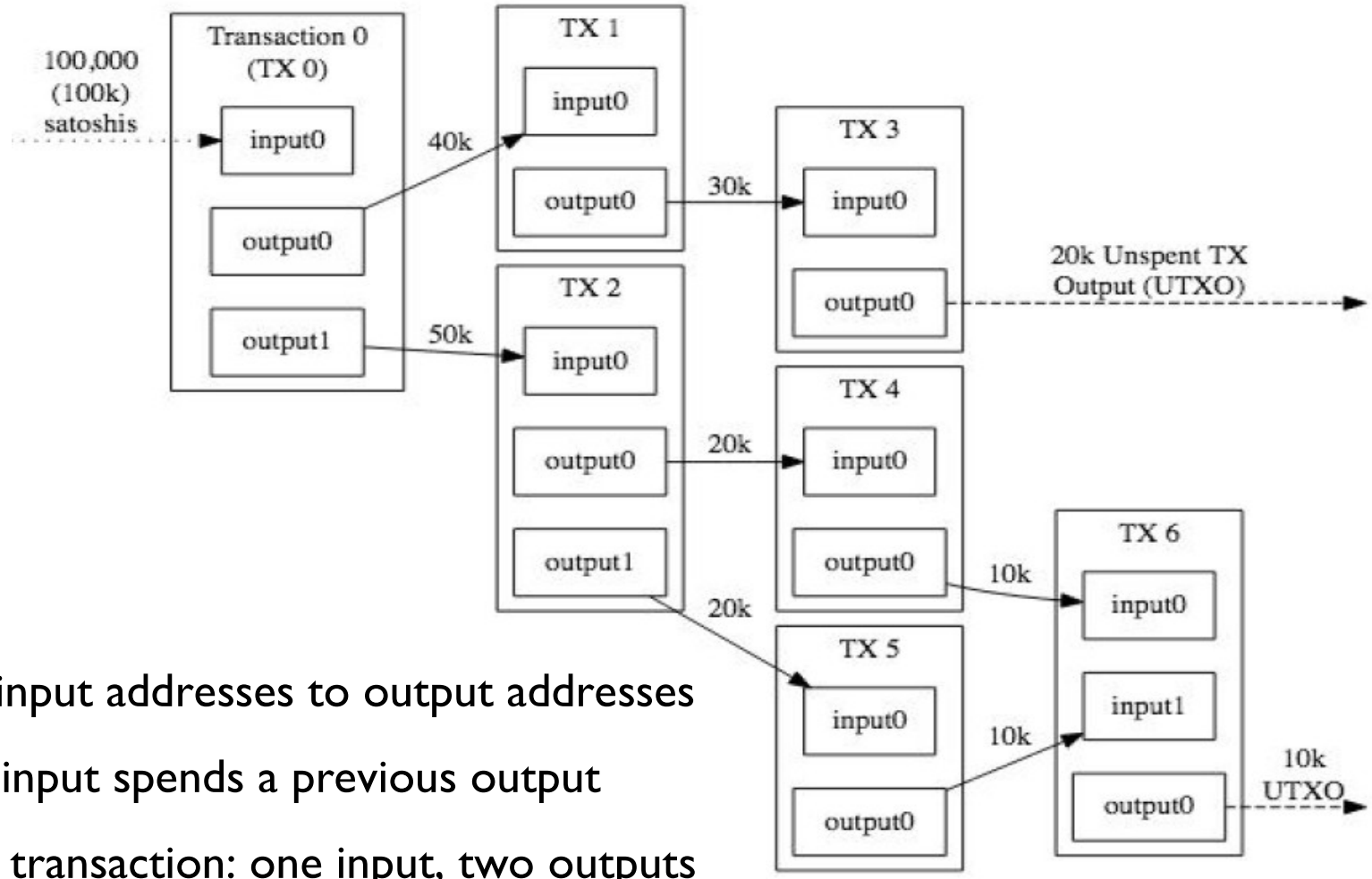
- a chain of ownership is registered on the blockchain
- value is moved from address to address…

BITCOIN TRANSACTIONS CHAIN



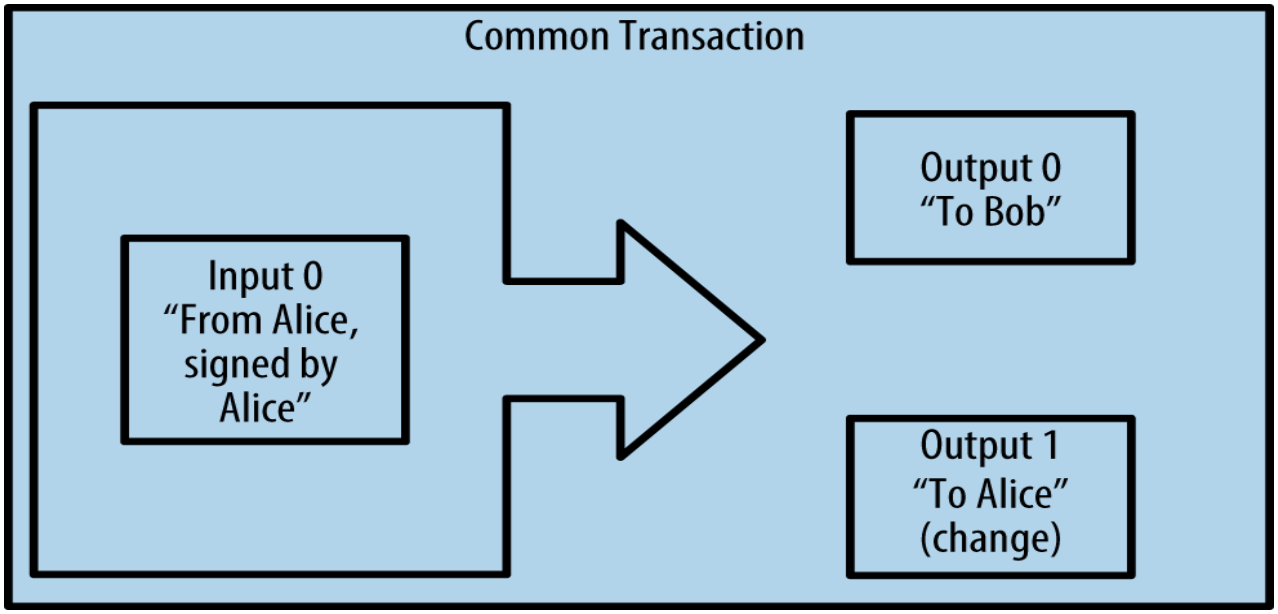
- a chain of ownership is registered on the blockchain
- value is moved from address to address…

THE UXTO MODEL



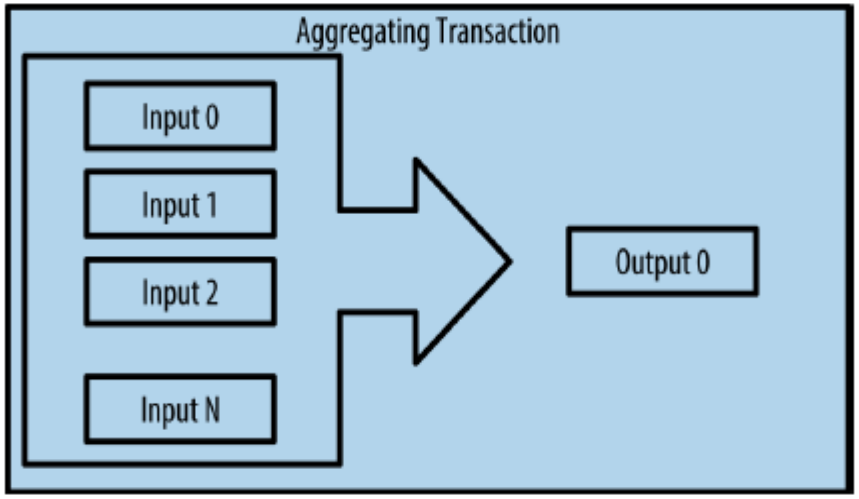
- map input addresses to output addresses
- each input spends a previous output
- usual transaction: one input, two outputs
- only unspent output are significative: mainteined in the UXTO

COMMON TYPE OF TRANSACTIONS



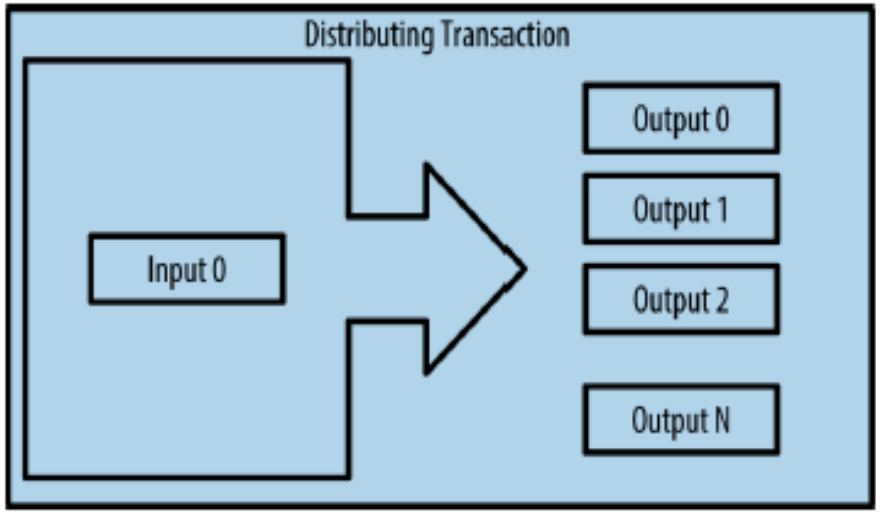
- the most common form of transaction: a simple payment from one address to another
- often includes some "change" returned to the original address.
- this type of transaction has one input and two outputs

TYPE OF TRANSACRIONS: AGGREGATING FUNDS



- a transaction aggegating several inputs into a single output
  - the equivalent of echanging a pile of coins for a single larger note
- may be generated to clean up lots of smaller amounts that were received as change for payments (generated by wallet applications)
- merging funds belonging to the same user in the output of the transaction, but exploited also for joint payments (multisignature transactions)

TYPE OF TRANSACTIONS: DISTRIBUTING FUNDS



- transactions distributing one input to multiple outputs representing multiple recipients
- used to distribute funds, for instance processing payroll payments to multiple employees

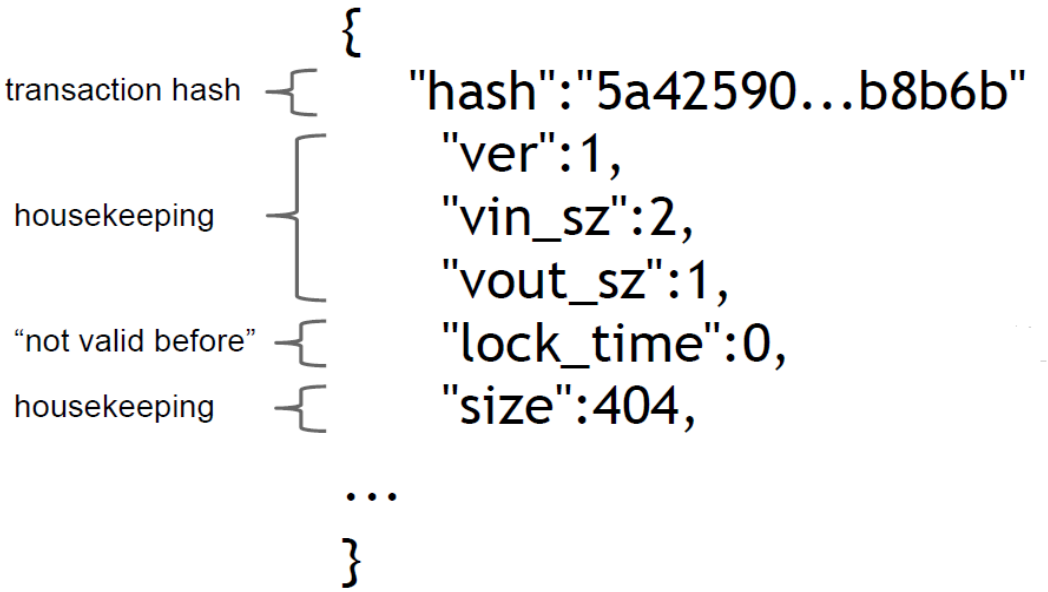
BITCOINs AND SATOSHIS

1 Satoshi	= 0.00000001 BTC
10 Satoshi	= 0.00000010 BTC
100 Satoshi	= 0.00000100 BTC
1,000 Satoshi	= 0.00001000 BTC
10,000 Satoshi	= 0.00010000 BTC
100,000 Satoshi	= 0.00100000 BTC
1 million Satoshi	= 0.01000000 BTC
10 m Satoshi	= 0.10000000 BTC
100m Satoshi	= 1.00000000 BTC

THE TRANSACTION'S AMOUNT

BTC / EUR	Date : 2024.04.05
1 BTC to EUR =	61,988.970
10 BTC to EUR =	619,889.700
50 BTC to EUR =	3 099 449
100 BTC to EUR =	6 198 897
500 BTC to EUR =	30 994 485
1000 BTC to EUR =	61 988 970
1500 BTC to EUR =	92 983 455
2000 BTC to EUR =	123 977 940
3000 BTC to EUR =	185 966 910
5000 BTC to EUR =	309 944 850

A REAL BITCOIN TRANSACTION (JSON)



- `hash` of the entire transaction, an unique identifier
- `version` allow different interpretation of some fields
- `locktime` defines the earliest time that a transaction can be added to the blockchain
  - set to zero in most transactions to indicate immediate execution
  - used in escrow and for the lightning network (a payment channel)

A REAL BITCOIN TRANSACTION (JSON)

previous transaction

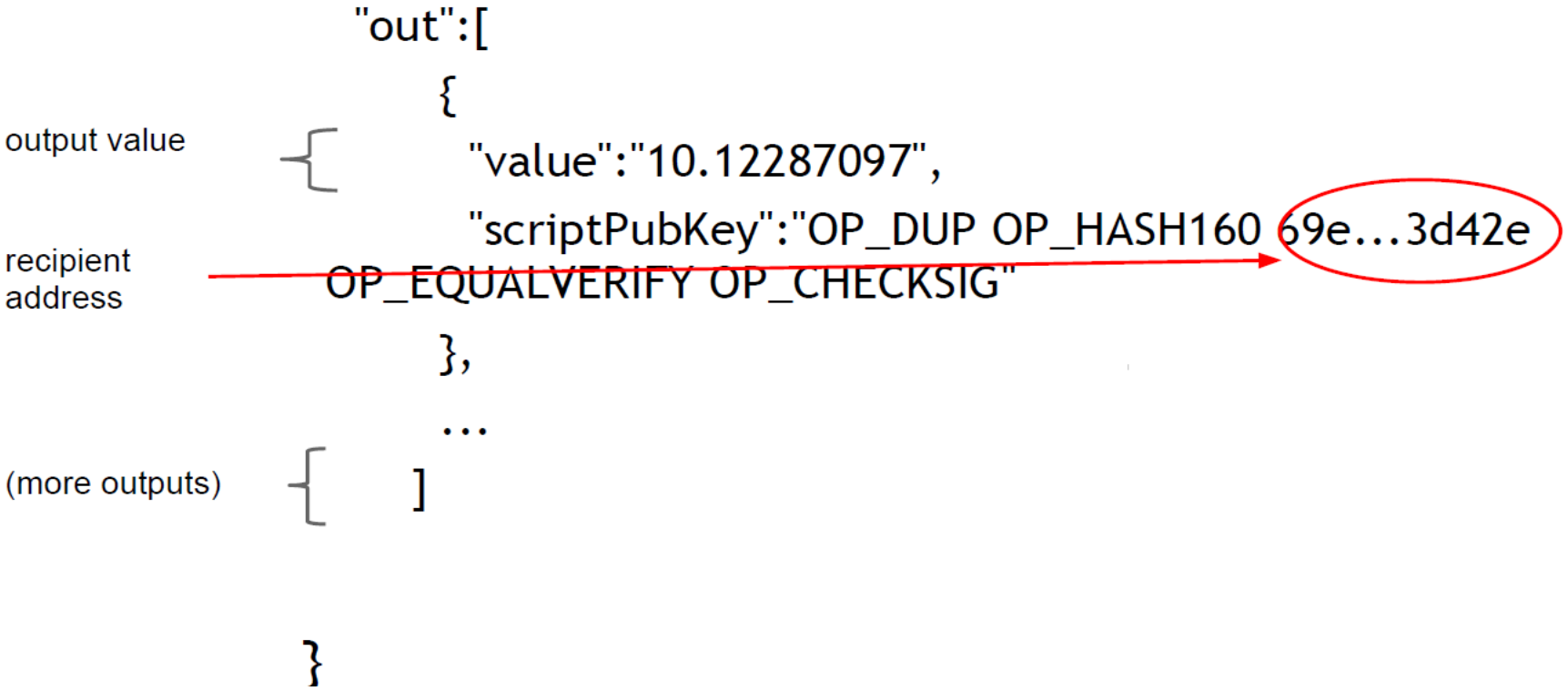
signature

(more inputs)

```
"in":[
 {
  "prev_out":{
   "hash":"3be4…80260",
   "n":0
  },
"scriptSig":"30440….3f3a4ce81"
 },
…
],
```

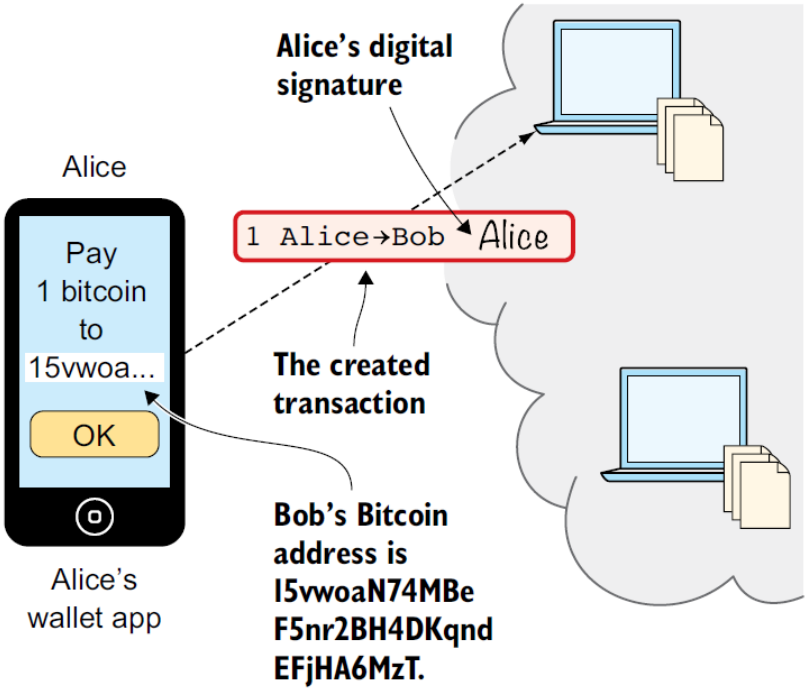
- a JSON array: each element contains a key value pair
  - hash pointer to a previous transaction and index of the previous transaction's output to be spent
  - an unlocking script

A REAL BITCOIN TRANSACTION (JSON)



- a JSON array where each element contains a pair
  - value to be transferred in that output
  - a locking script containing the address where that value has to be transferred

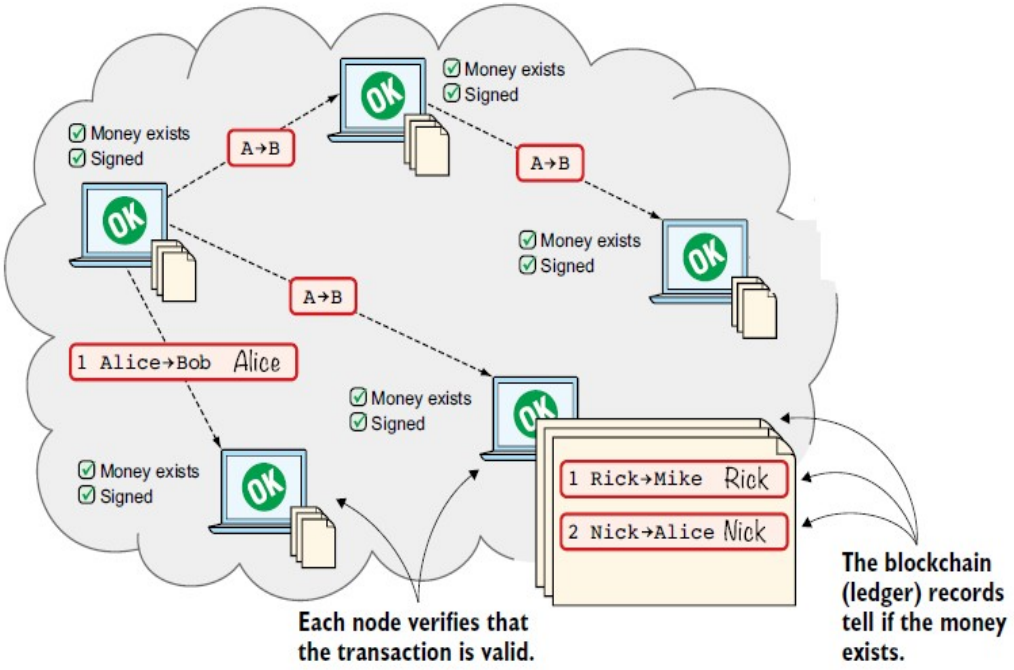
BITCOIN IN A NUTSHELL: RECAP



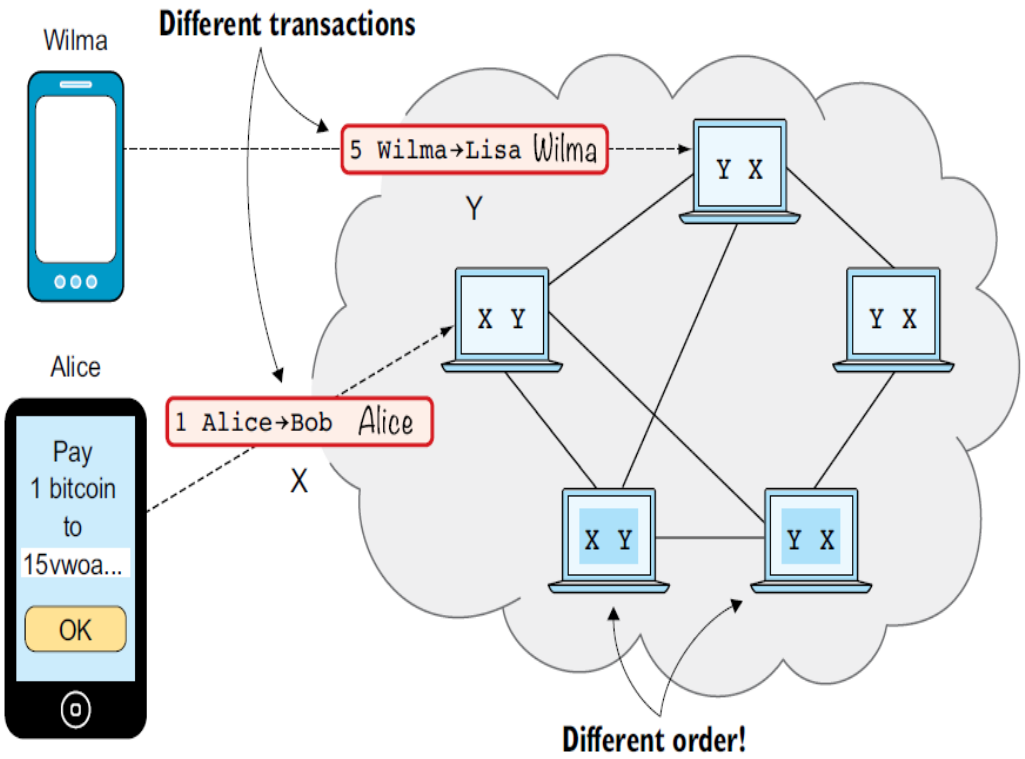
- what do you find inside a transaction?
  - amount to move (ex: `1bitcoin`)
  - the address of the receiver, something like
    15vwoaN74MBeF5nr2BH4DKqndEFjHA6MzT
  - a digital signature
    - created through Alice's private key
- in this case, the transaction starts from Alice's mobile wallet app

BITCOIN IN A NUTSHELL: RECAP

- the transaction is propagated on the P2P network
- each node checks that the transaction is valid
  - the bitcoin `Alice` is spending exists in a transaction output
  - `Alice`'s digital signature is valid
- to decide if `Alice` can spend the bitcoin, check the blockchain



BITCOIN IN A NUTSHELL: RECAP



- the ledger stores the history of all transactions

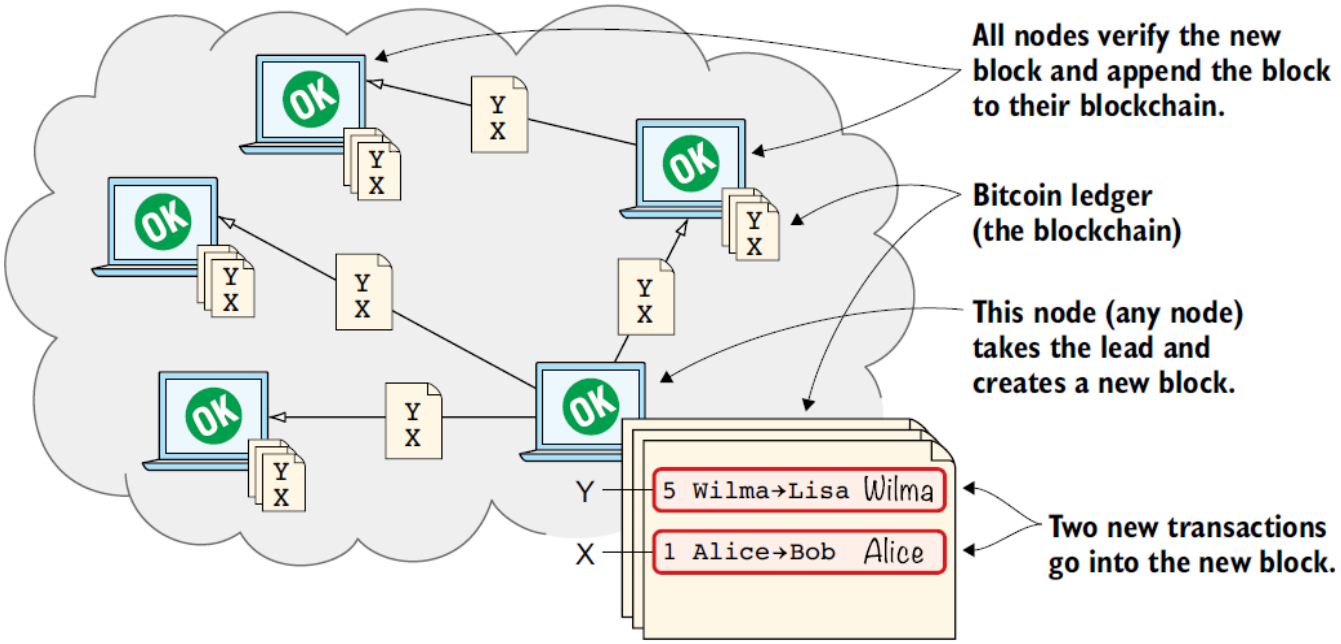
transactions

- may arrive in different orders on different nodes
- due to different network delays

what is needed?

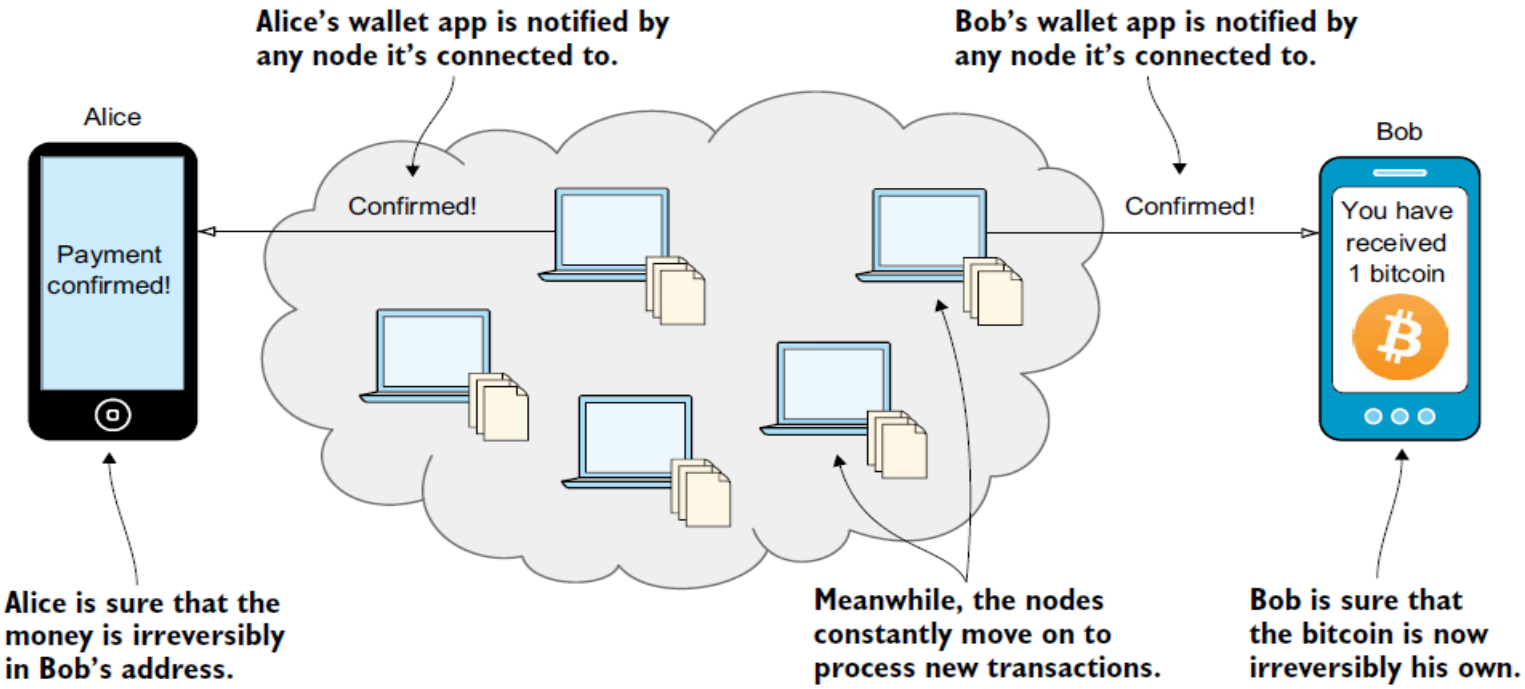
- one node taking the lead and deciding the next transaction to add
- all the other nodes agree

BITCOIN IN A NUTSHELL: RECAP



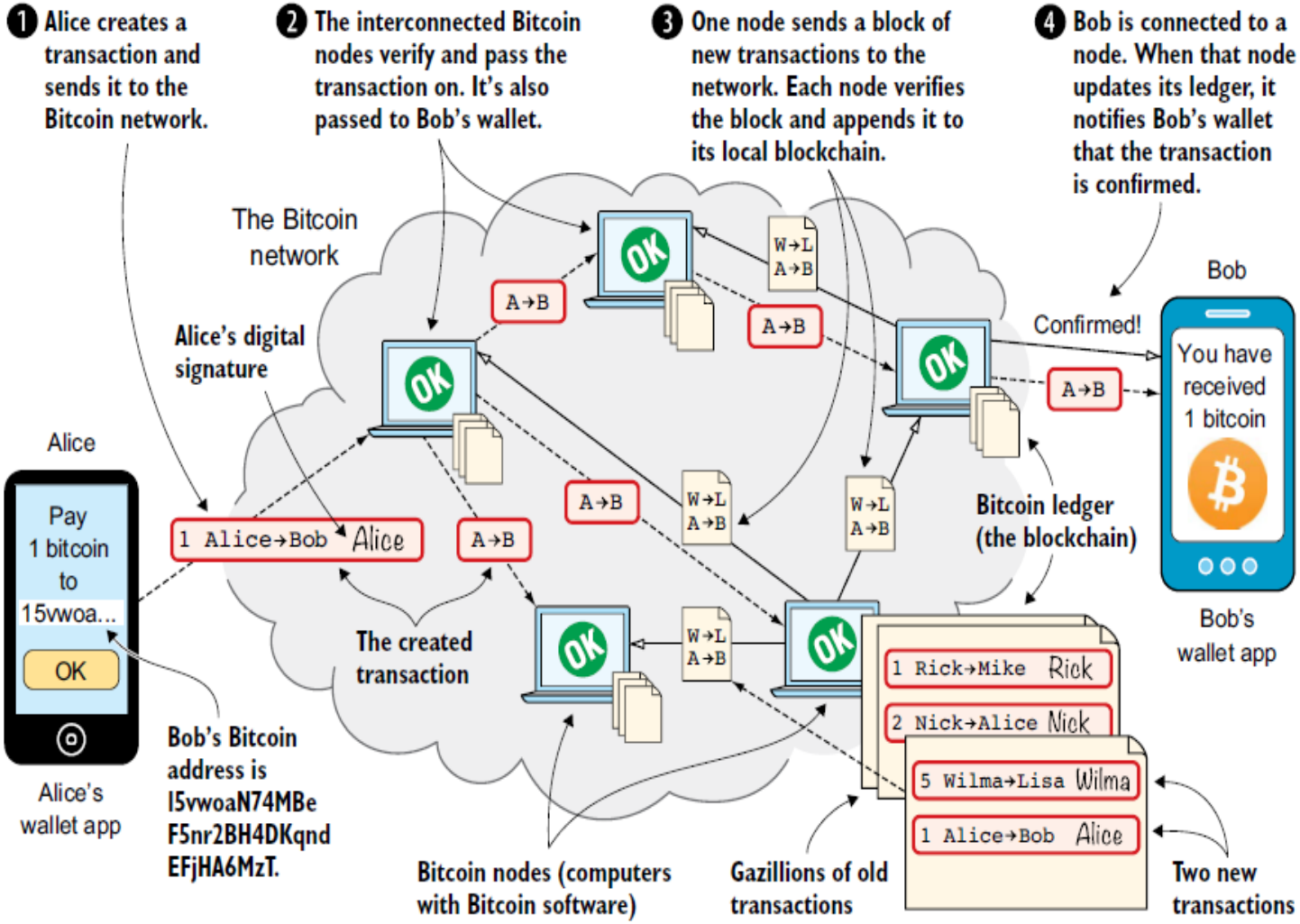
- one node takes the lead and tells the others in what order to add transactions.
  - yhis is implemented through consensus
  - sends the block on the network, as the transactions in the previous step
- the other nodes verify the block and update their blockchain copies accordingly.

BITCOIN IN A NUTSHELL: RECAP

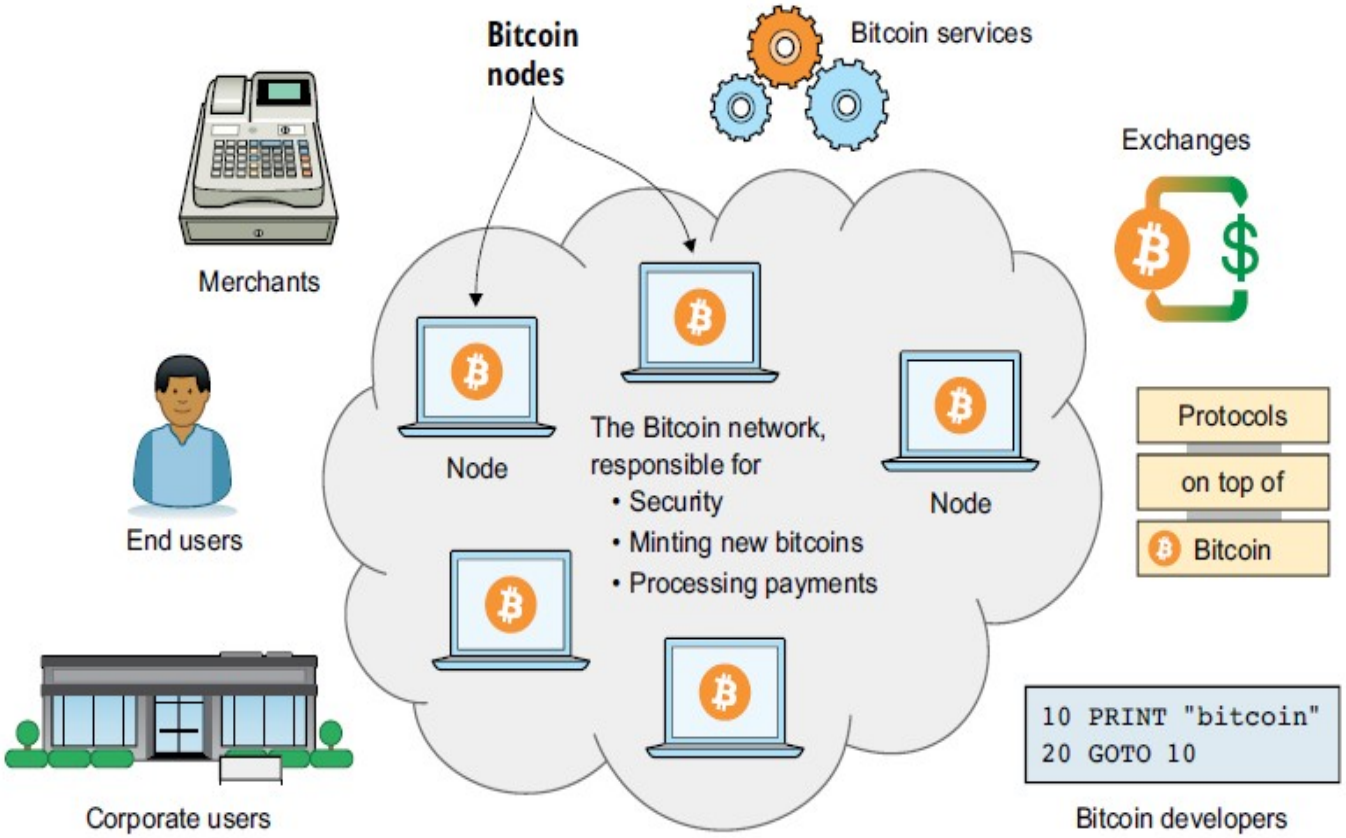


- when all the nodes have updated their local copy of the blockchain, the transaction of `Alice` has been accepted
- the node that the `Bob`'s wallet is connected to will notify `Bob`'s wallet
- the wallet display a message to `Bob` notifying that `Bob` has received 1 bitcoin

THE BITCOIN SYSTEM: ALL TOGETHER

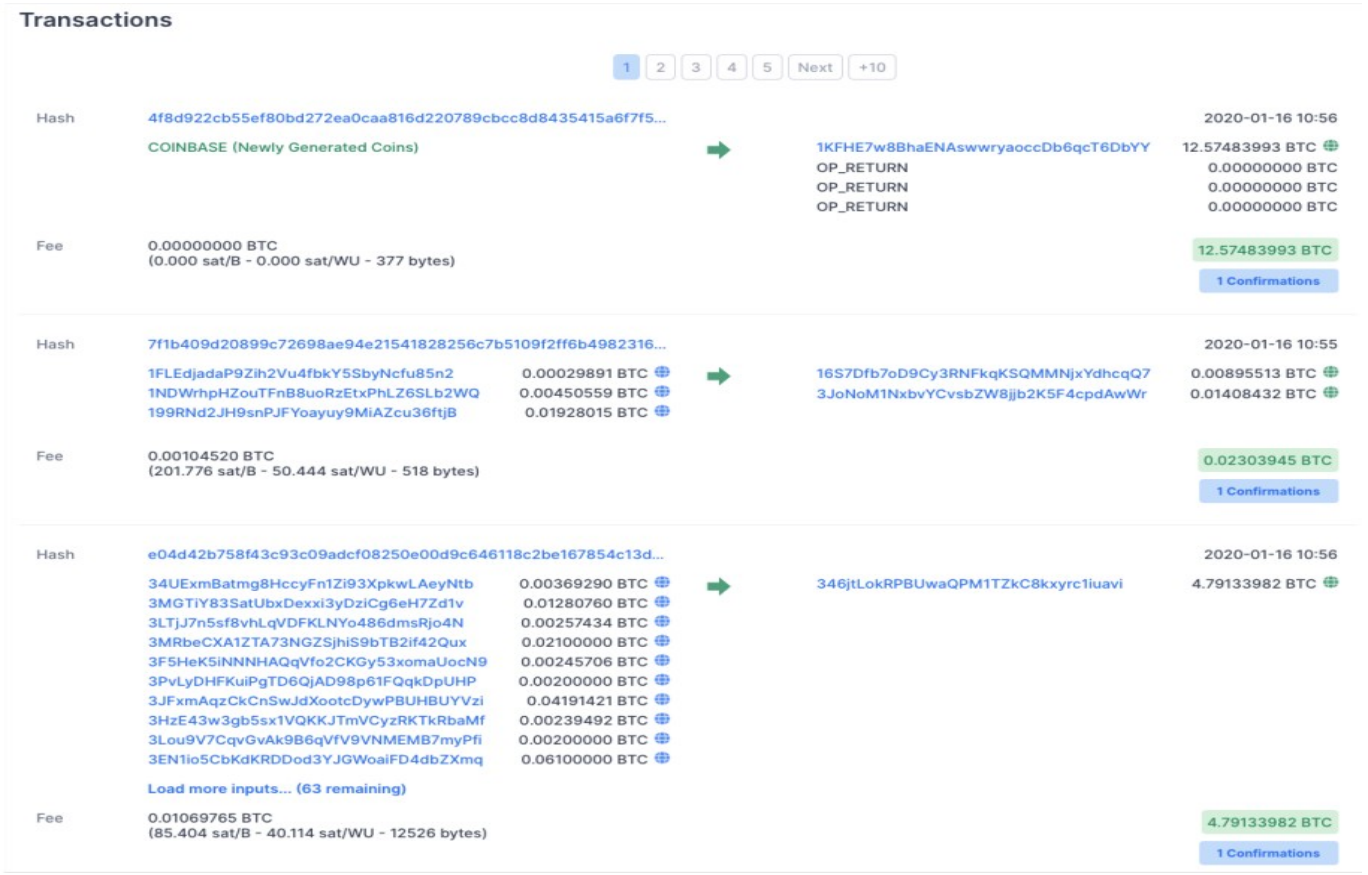


THE BITOCOIN ECOSYSTEM



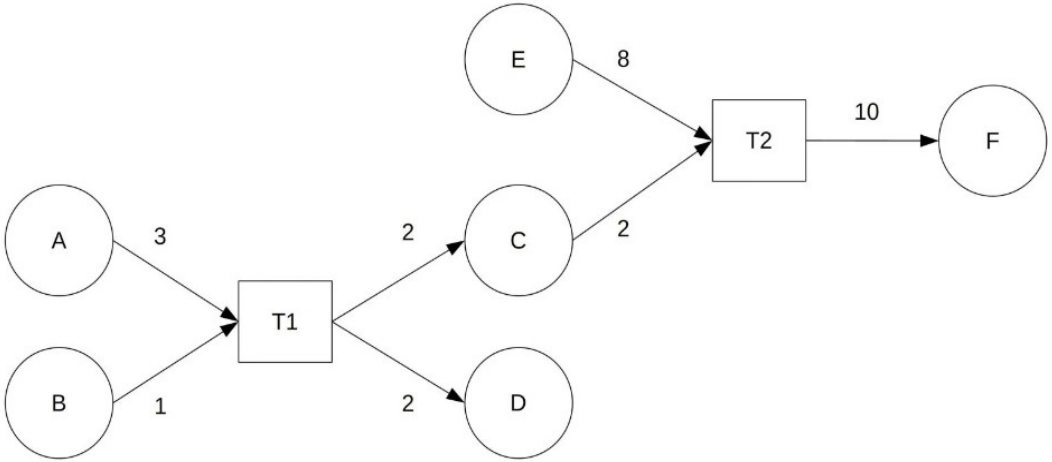
SCRAPING TO RETRIEVE BITCOIN TRANSACTIONS

- tools like `Bitcoin explorer` (`https://www.blockchain.com/explorer`)



CRYPTO ANALYSIS: TAINT ANALYSIS

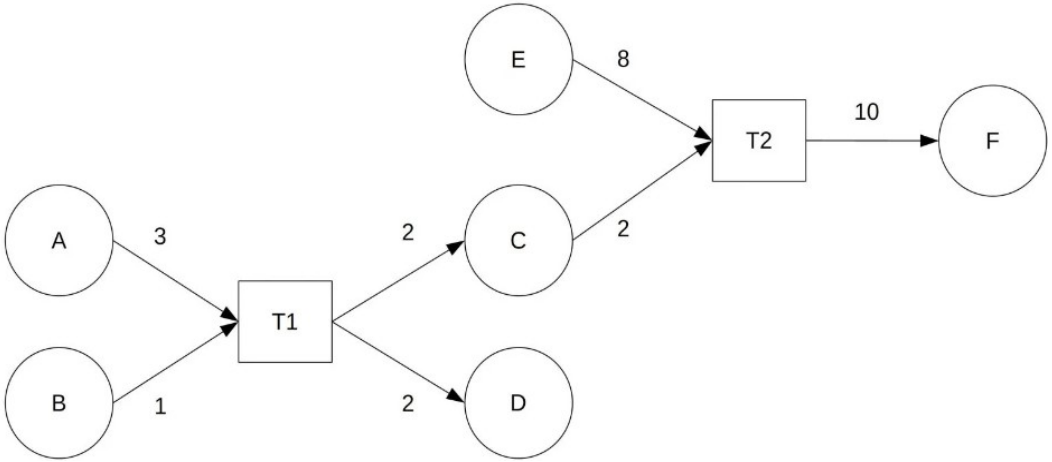
- shows how large is the percentage of a coin in one address from another address



- taint = contaminazione
- address A and B contribute to a transaction T1 with 3 coins and 1 coin, respectively.
- since transactions do not explicitly say which input goes to which output, we assume that they are split evenly.
- so, address C contains ¾ of the coin from address A and ¼ of the coin from address B.

CRYPTO ANALYSIS: TAINT ANALYSIS

- shows how large is the percentage of a coin in one address from another address



- now consider transaction T2, where address E combines its UTXO with address C.
- there are eight out of ten parts that come from address E, and two out of ten from address C. So
- Taint of A in F = 0.2*0.75=0.15
- Taint of B in F = 0.2*0.25=0.05

CRYPTO ANALYSIS: TAINT ANALYSIS

- Taint analysis is utilised as a tracking method in cryptocurrency
- tracks targeted cryptocurrency coins using transaction information in the blockchain.
- determine the association between two addresses in a transaction
  - classify the targeted cryptocurrency coins (e.g., stolen Bitcoins resulting from a known theft transaction) as tainted (or "dirty")
  - any address that uses or transfers them will be considered a tainted address.
  - coins that are unrelated to tainted coins are considered clean coins.
  - different taint analysis strategy with specific rule-set to estimate how the targeted cryptocurrency coins are distributed in the subsequent transactions.

EXCHANGES

- key players in the Bitcoin ecosystem
- offer exchanging services to cryptocurrency users, allow cryptocurrency users
  - to sell their coins for fiat currency
  - to buy new coins with fiat currency
  - to exchange crypto for another crypto, for instance bitcoin for ether
- usually require payment of a fee
- may offer also a service of wallet custodian provider (for example Bitfinex)

BUYING FROM AN EXCHANGER

- several place where you can buy
  - `Coinbase` (require ID)
  - `Binance` (require ID)
  - `Local Bitcoin` (help you other people to trade with)
  - `Bisq` (decentralized exchanger, no ID other mechanism to ensure trust)
  - `Hodl` (decentralized exchanger, no ID other mechanism to ensure trust)

BUYING FROM AN EXCHANGER: COINBASE

- started in June 2012
- first regulated exchange: 2015, highly secure
- may be accessed from 102 countries, 0,1% transaction fee
- can buy and sell several cryptocurrencies, among them
  - Bitcoin (B T C)
  - Basic Attention Token (B A T)
  - Bitcoin Cash (B C H)
  - EOS (E O S)
  - Ethereum (E T H)
  - Ethereum Classic (E T C)
  - Chainlink (L I N K)
  - Litecoin (L T C)
  - Augur (R E P)
  - Stellar Lumens (X L M)
  - Ripple (X R P)
  - Zcash (Z EC)
  - ...

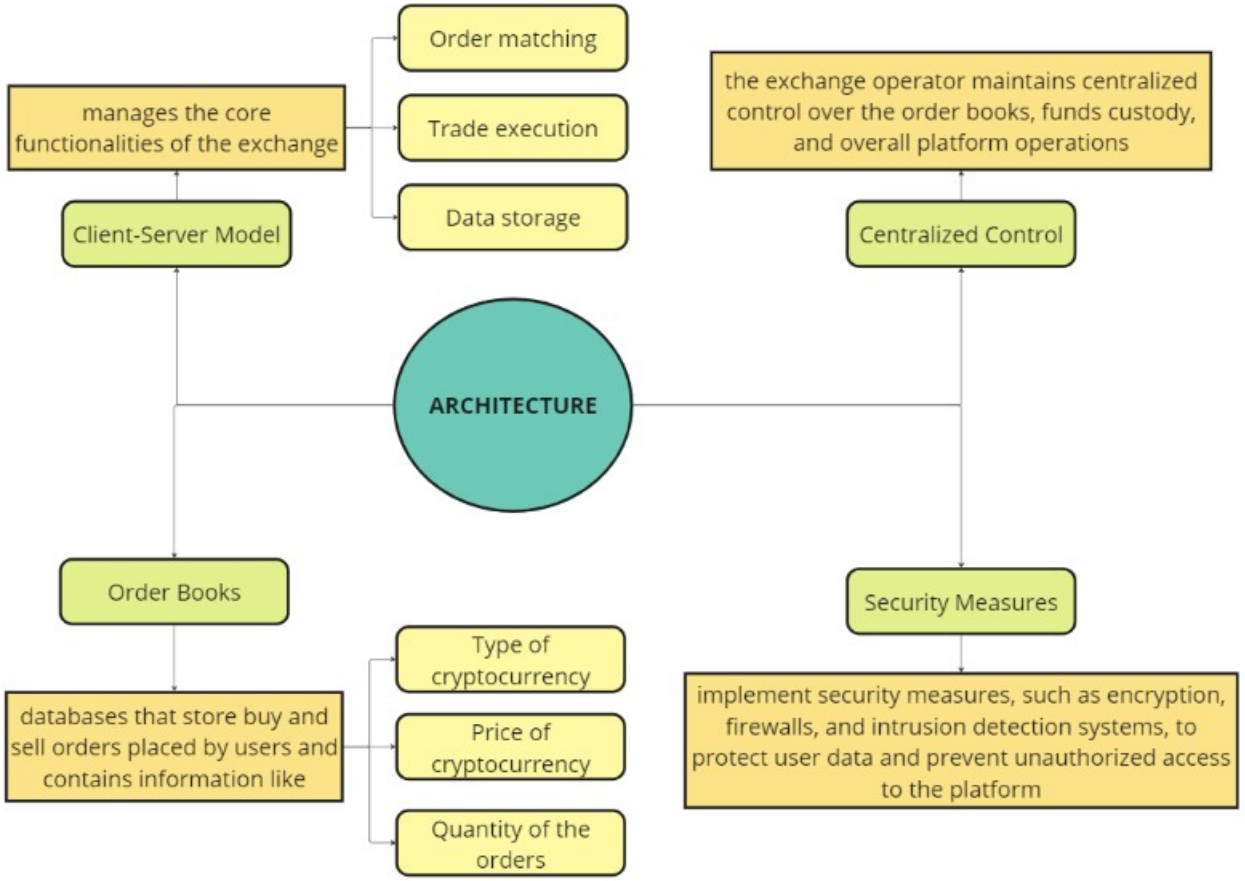
BUYING FROM AN EXCHANGER: BINANCE

- younger than Coinbase, started in 2017
- started in China
- may be accessed from 120 countries
- offers different services
  - centralized exchange
  - also decentralized exchange

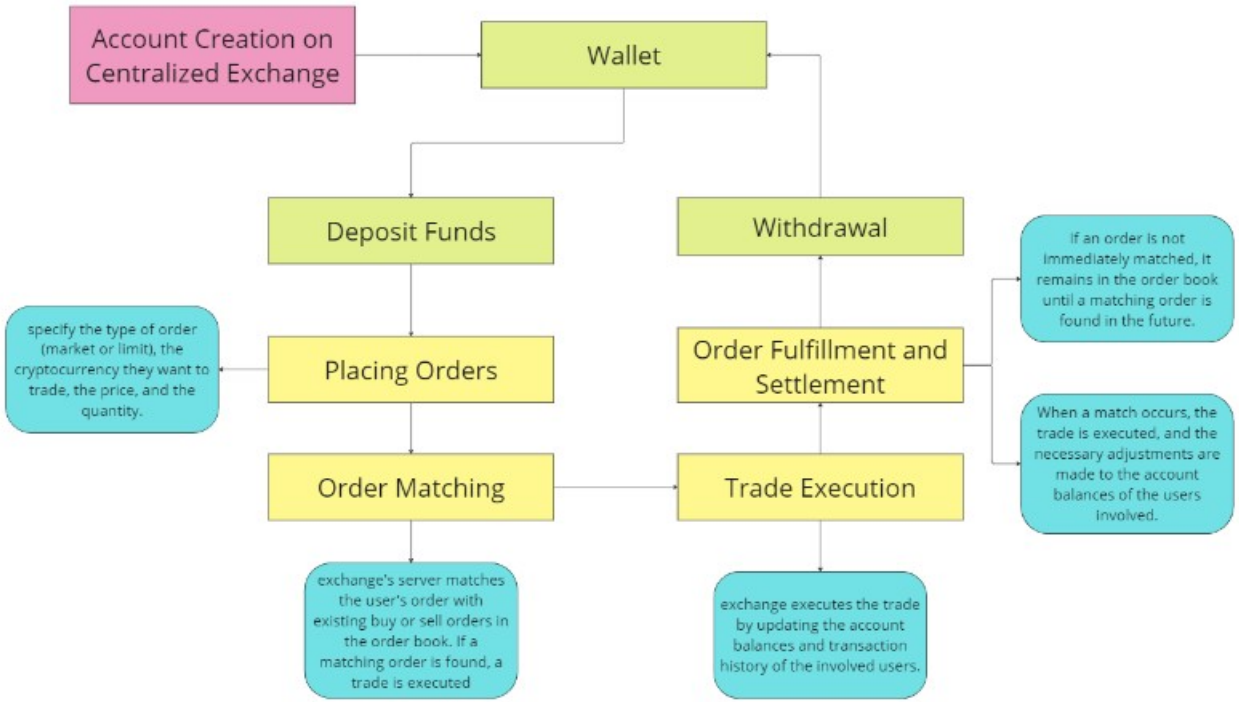
UNDERSTANDING CENTRALIZED EXCHANGERS

- blockchain exchanges are platforms where users can buy, sell, and trade cryptocurrencies and digital assets.
- act as intermediaries, facilitating transactions between buyers and sellers.
- how they work?
  - when a user signs up on a blockchain exchange, they create an account and obtain a digital wallet.
  - they can deposit funds into their wallet and use them to buy cryptocurrencies.
  - exchanges maintain order books, which display the current buy and sell orders for various cryptocurrencies.
  - when a user places an order, the exchange matches it with an existing order or adds it to the order book.

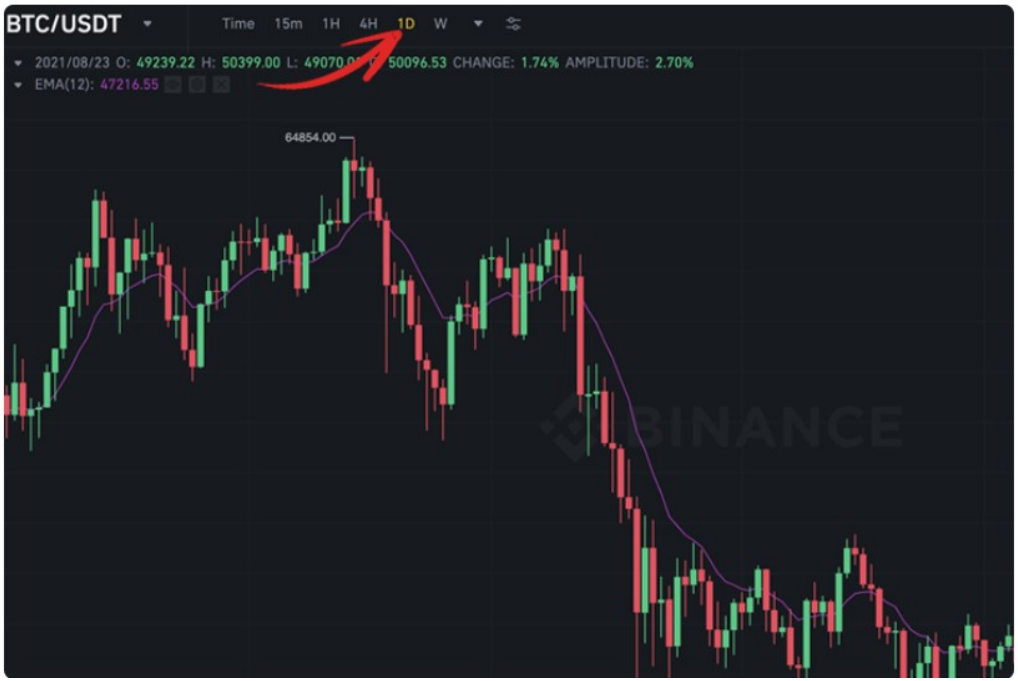
UNDERSTANDING CENTRALIZED EXCHANGERS

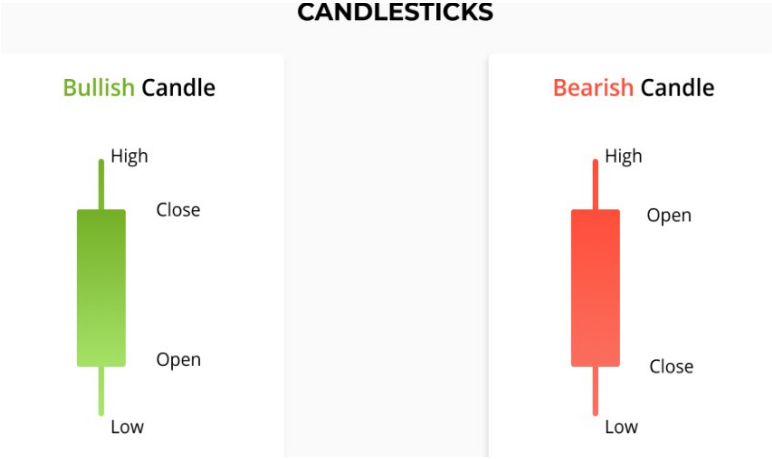


UNDERSTANDING CENTRALIZED EXCHANGERS



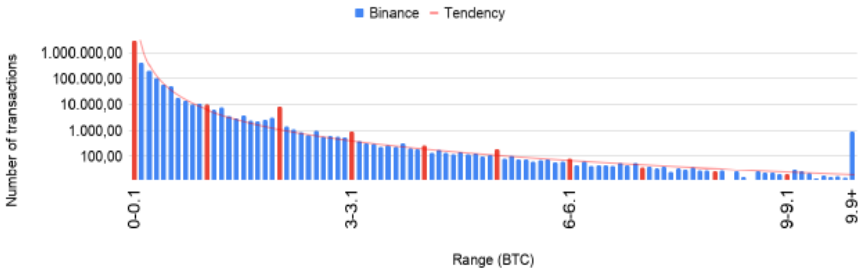
ANALYSING CENTRALIZED EXCHANGERS



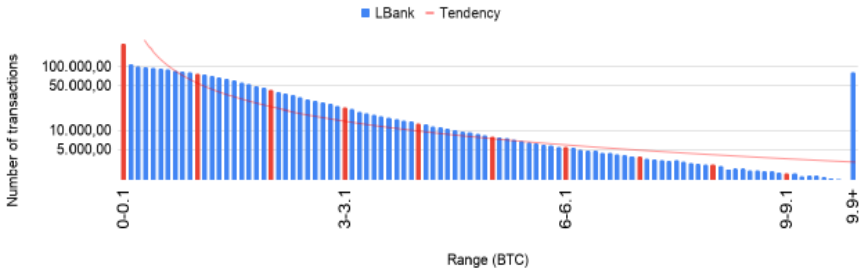


ANALYSING CENTRALIZED EXCHANGERS

- scape data or use the API of exchangers to monitor time series of order books exchanges
- goal: find fraudolent exchangers
- automatically generated volume of transactions



(a) Distribution of Binance.



(b) Distribution of LBank.

Figure 5.8: Comparison between the distributions of the number of BTC-USD transactions for ranges of 0.1 BTC (Binance and LBank).